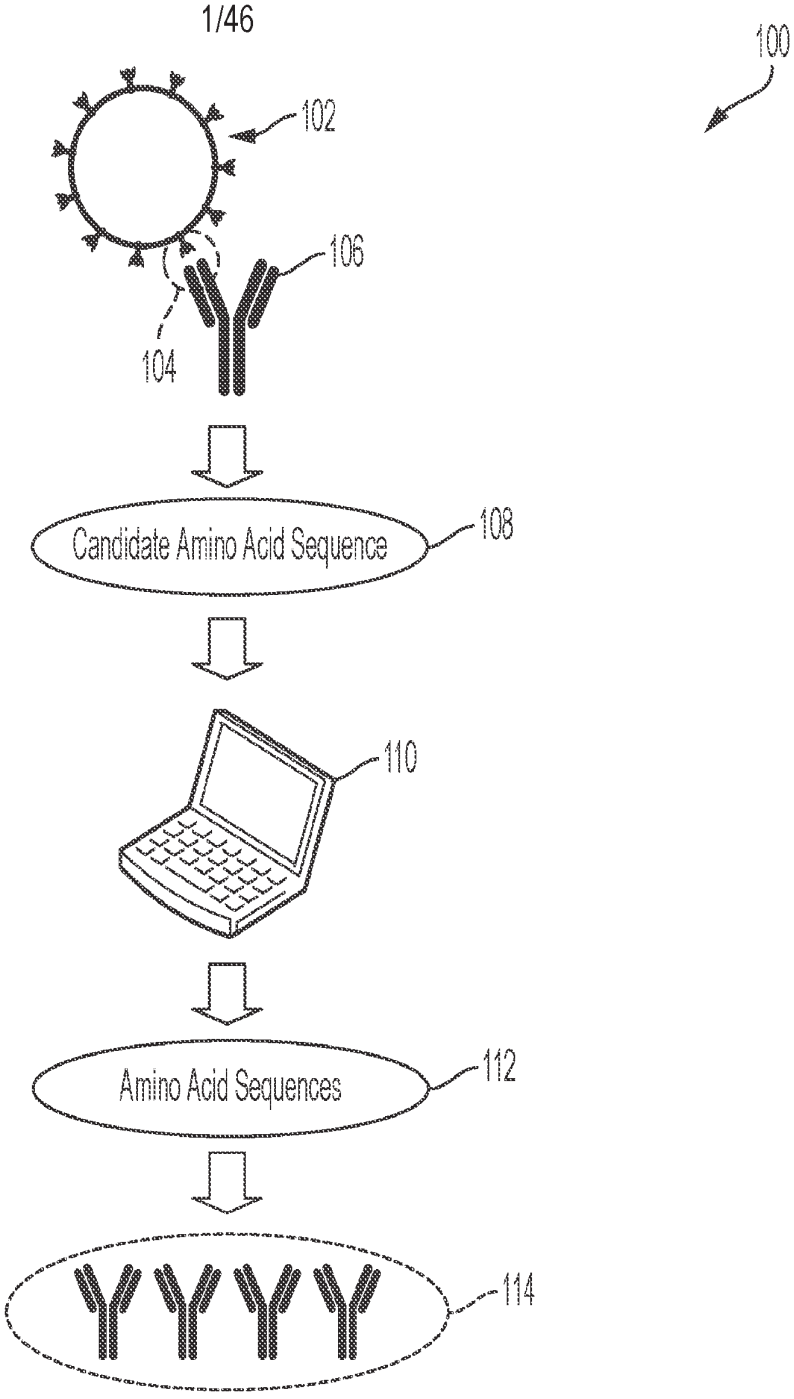

100
102
106
104
Candidate Amino Acid Sequence
108
110
Amino Acid Sequences
112
114

FIG. 1A

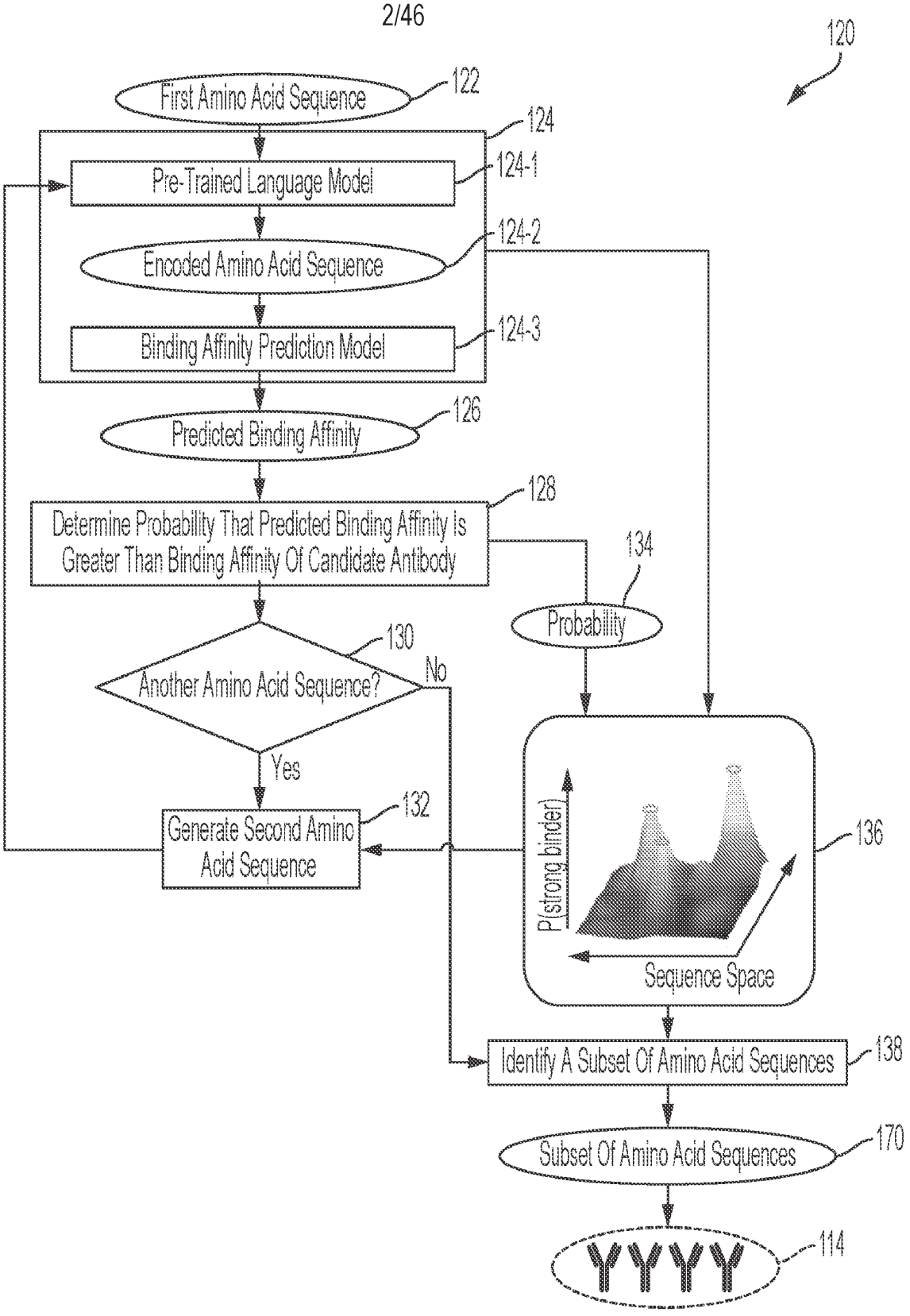

120
First Amino Acid Sequence
122
124
Pre-Trained Language Model
124-1
Encoded Amino Acid Sequence
124-2
Binding Affinity Prediction Model
124-3
Predicted Binding Affinity
126
Determine Probability That Predicted Binding Affinity Is Greater Than Binding Affinity Of Candidate Antibody
128
134
Probability
Another Amino Acid Sequence?
130
No
Yes
Generate Second Amino Acid Sequence
132
P(strong binder)
Sequence Space
136
Identify A Subset Of Amino Acid Sequences
138
Subset Of Amino Acid Sequences
170
114

FIG. 1B

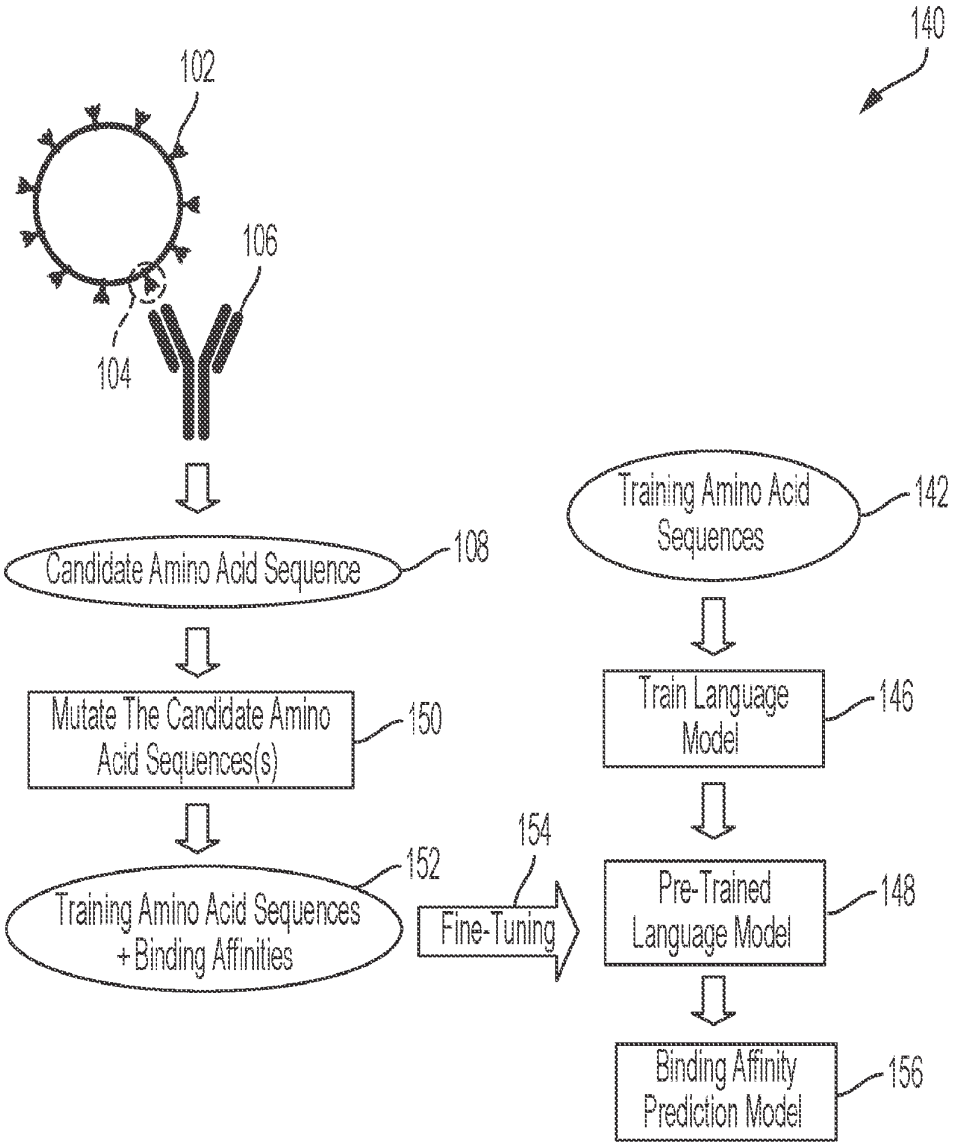

140
102
106
104
Candidate Amino Acid Sequence
108
Mutate The Candidate Amino Acid Sequences(s)
150
Training Amino Acid Sequences + Binding Affinities
152
154
Fine-Tuning
Training Amino Acid Sequences
142
Train Language Model
146
Pre-Trained Language Model
148
Binding Affinity Prediction Model
156

FIG. 1C

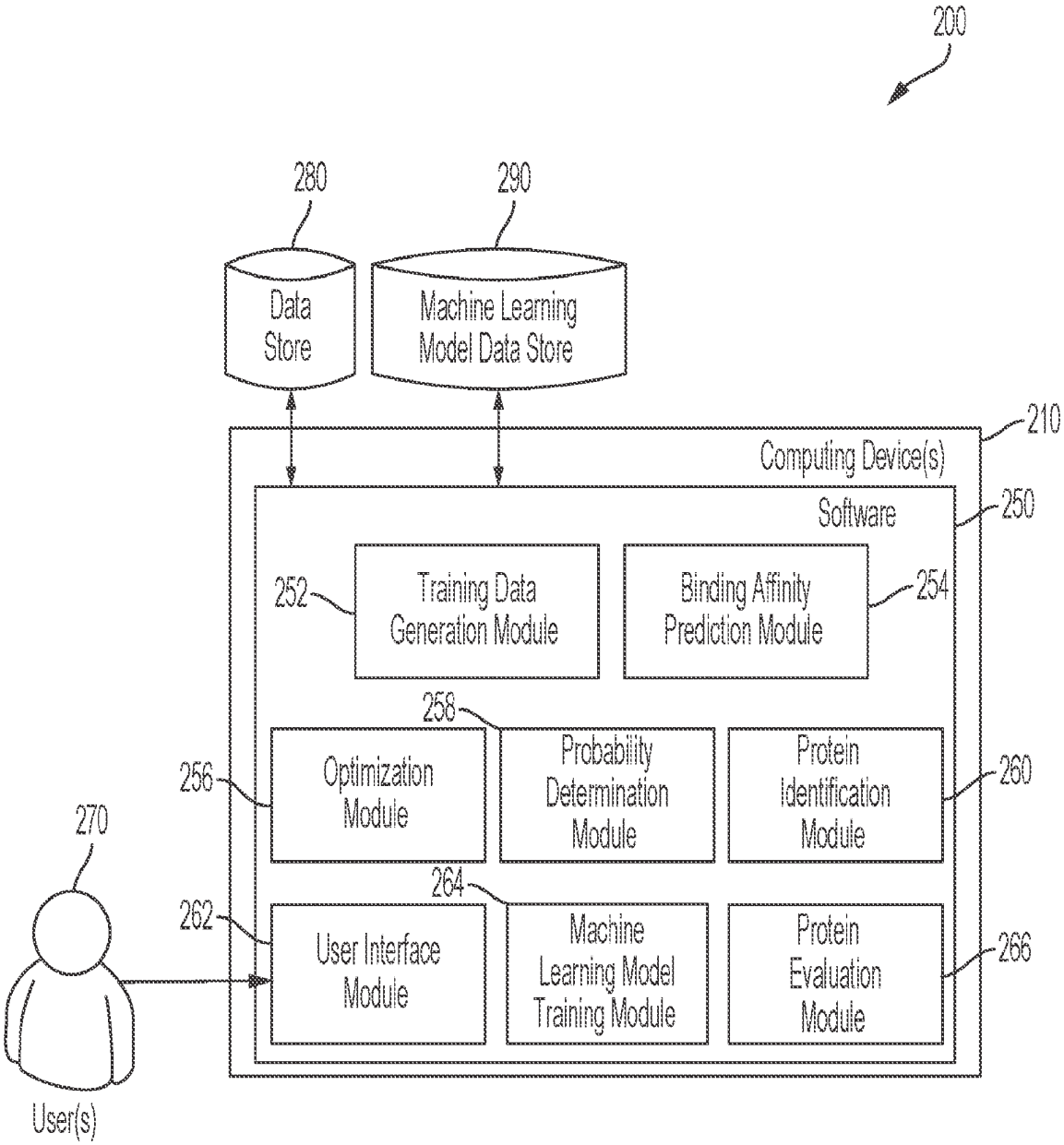

200
280
290
Data Store
Machine Learning Model Data Store
210
Computing Device(s)
250
Software
252
Training Data Generation Module
254
Binding Affinity Prediction Module
258
256
Optimization Module
Probability Determination Module
Protein Identification Module
260
270
264
262
User Interface Module
Machine Learning Model Training Module
Protein Evaluation Module
266
User(s)

FIG. 2

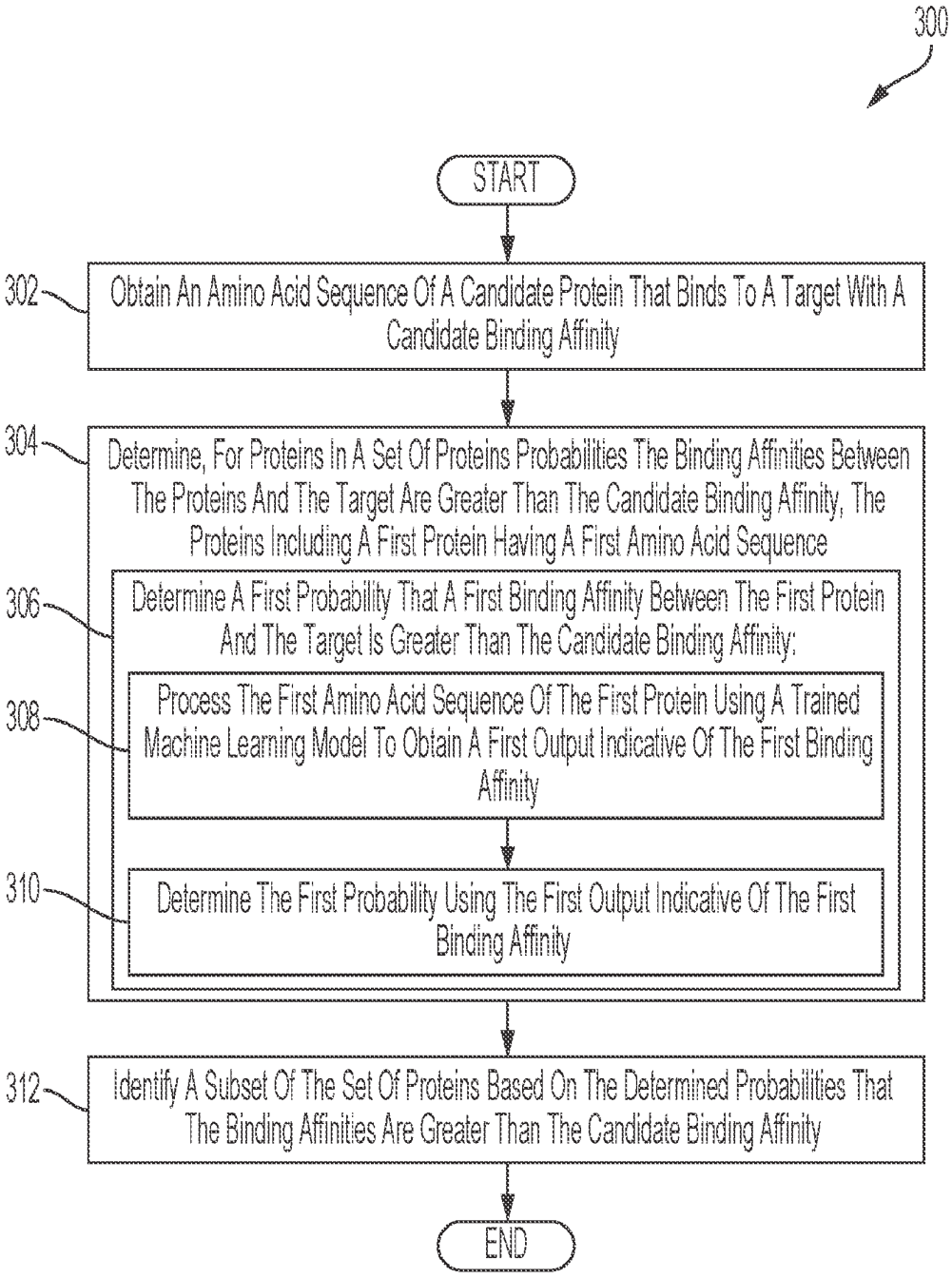

300
START
302
Obtain An Amino Acid Sequence Of A Candidate Protein That Binds To A Target With A Candidate Binding Affinity
304
Determine, For Proteins In A Set Of Proteins Probabilities The Binding Affinities Between The Proteins And The Target Are Greater Than The Candidate Binding Affinity, The Proteins Including A First Protein Having A First Amino Acid Sequence
306
Determine A First Probability That A First Binding Affinity Between The First Protein And The Target Is Greater Than The Candidate Binding Affinity:
308
Process The First Amino Acid Sequence Of The First Protein Using A Trained Machine Learning Model To Obtain A First Output Indicative Of The First Binding Affinity
310
Determine The First Probability Using The First Output Indicative Of The First Binding Affinity
312
Identify A Subset Of The Set Of Proteins Based On The Determined Probabilities That The Binding Affinities Are Greater Than The Candidate Binding Affinity
END

FIG. 3

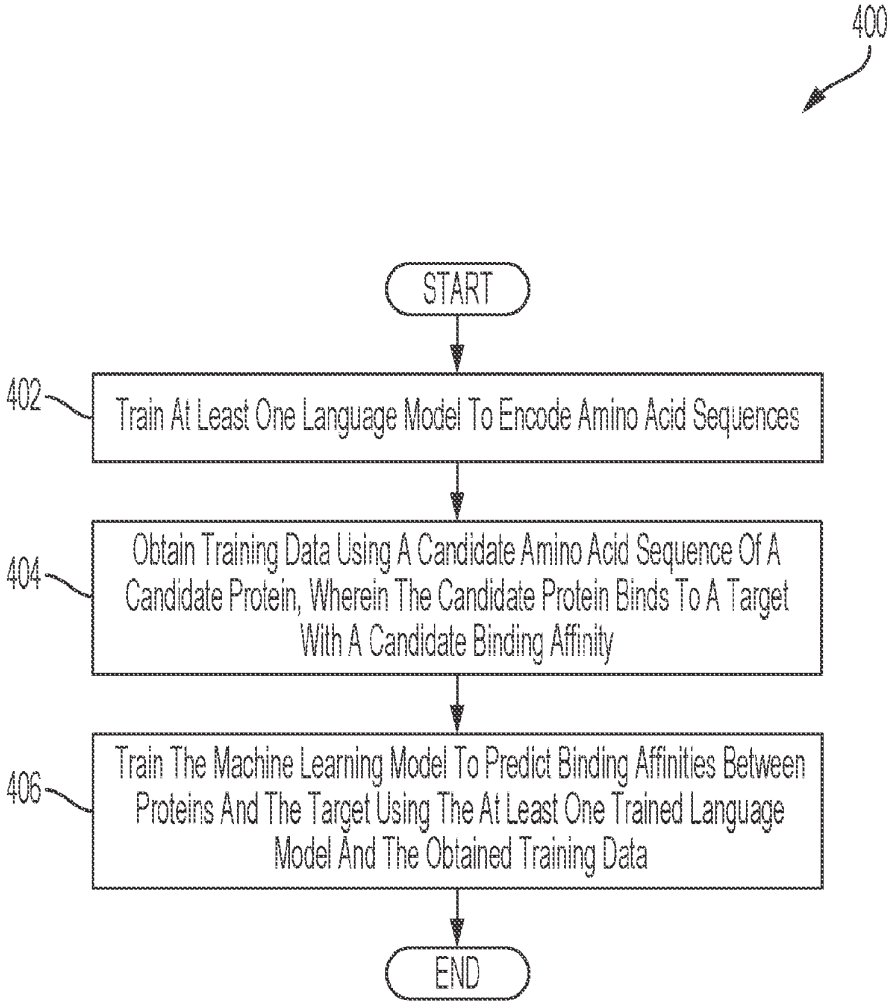

400
START
402
Train At Least One Language Model To Encode Amino Acid Sequences
404
Obtain Training Data Using A Candidate Amino Acid Sequence Of A Candidate Protein, Wherein The Candidate Protein Binds To A Target With A Candidate Binding Affinity
406
Train The Machine Learning Model To Predict Binding Affinities Between Proteins And The Target Using The At Least One Trained Language Model And The Obtained Training Data
END

FIG. 4

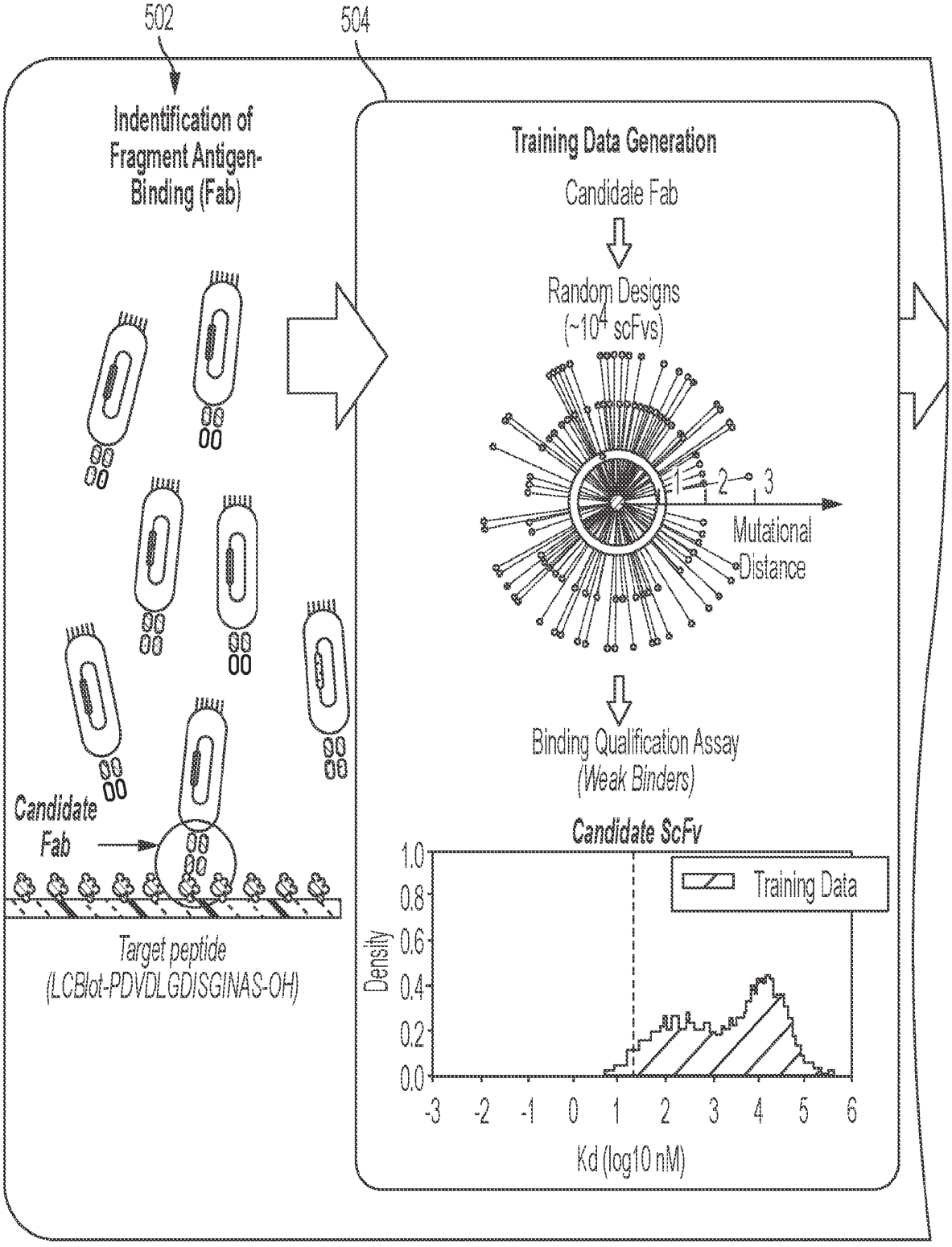

502
504
Indentification of Fragment Antigen-Binding (Fab)
Training Data Generation
Candidate Fab
Random Designs (~10^4 scFvs)
1
2
3
Mutational Distance
Binding Qualification Assay (Weak Binders)
Candidate ScFv
Training Data
Density
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
4
5
6
Kd (log10 nM)
Candidate Fab
Target peptide (LCBlot-PDVDLGDISGINAS-OH)

FIG. 5A

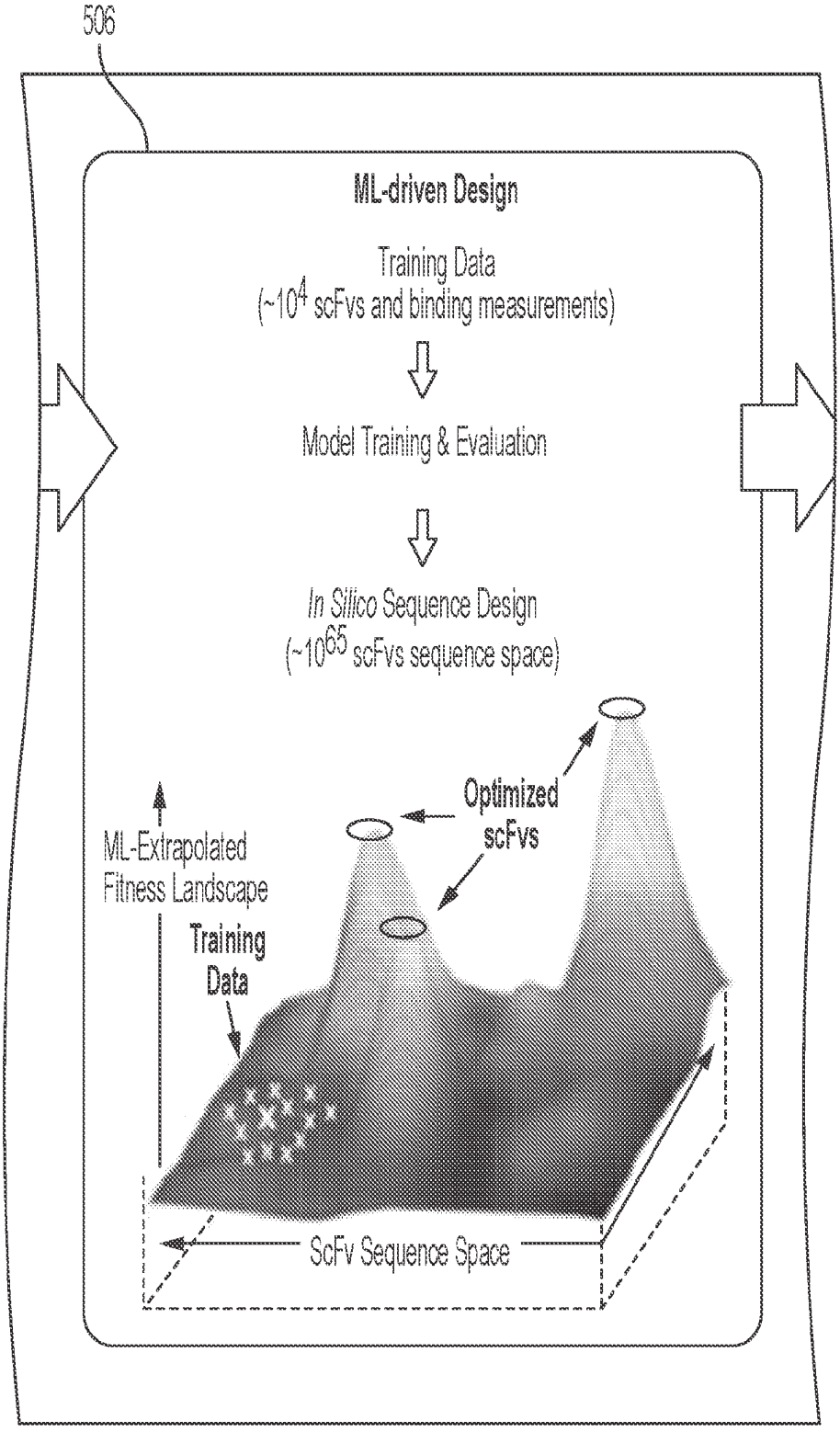


FIG. 5A
CONTINUED

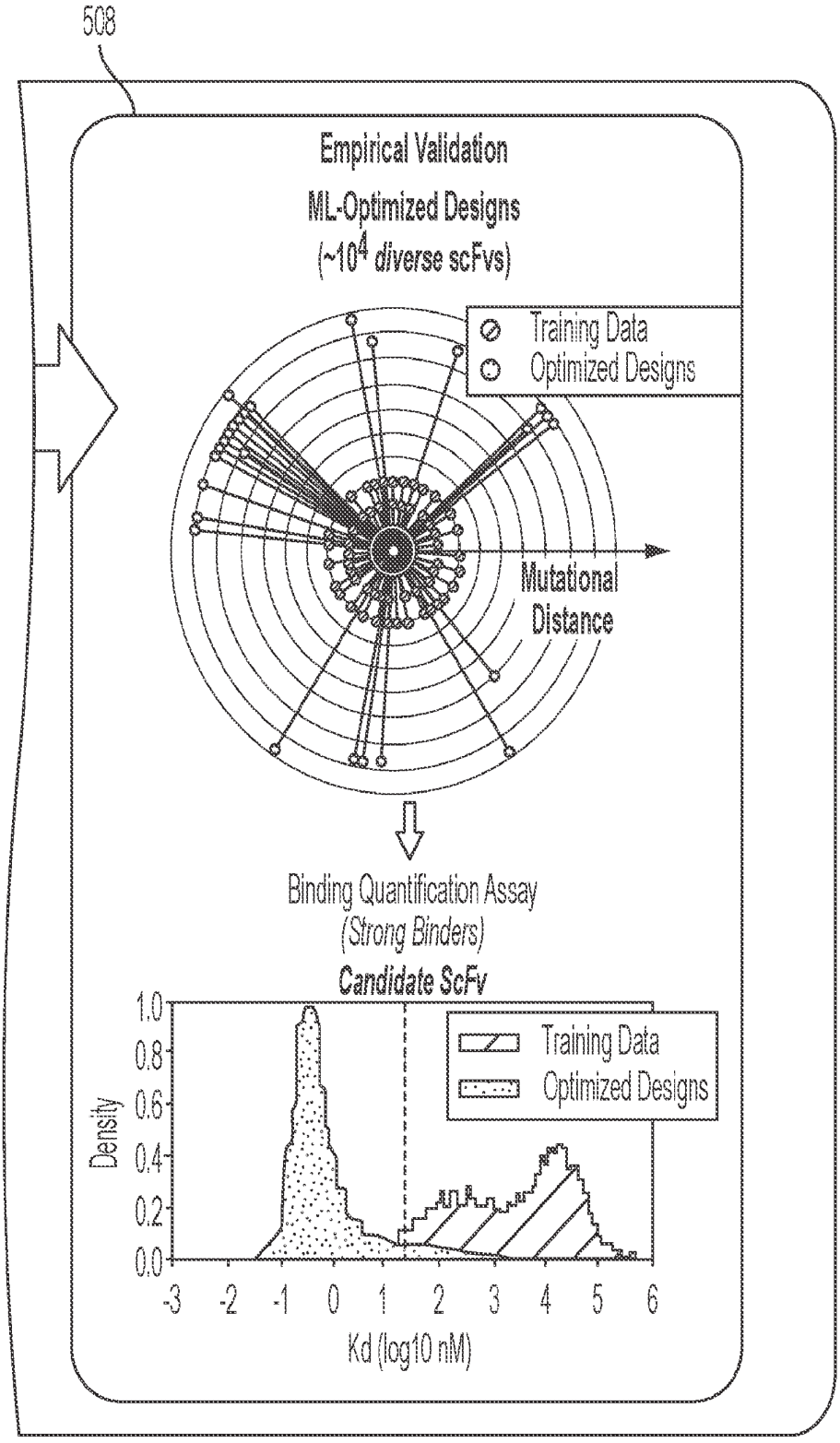


FIG. 5A
CONTINUED

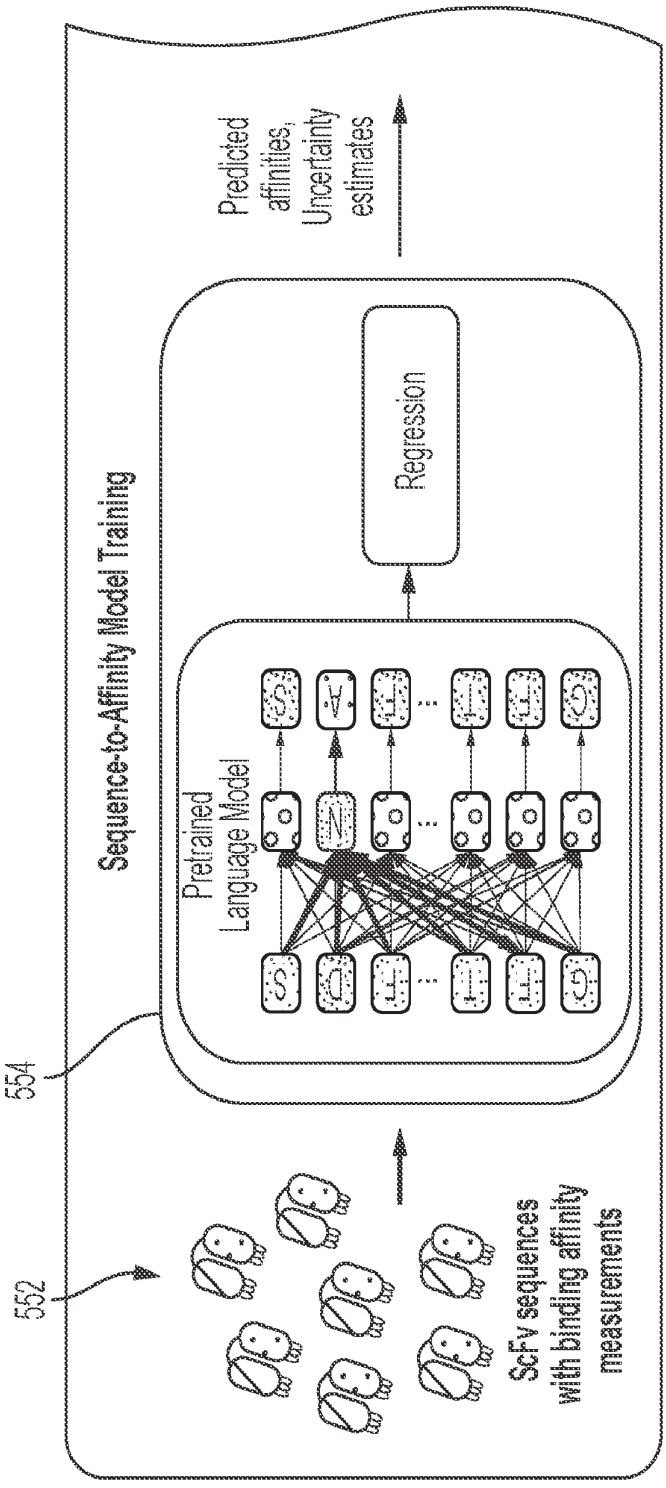

552
554
Sequence-to-Affinity Model Training
ScFv sequences with binding affinity measurements
Pretrained Language Model
Regression
Predicted affinities, Uncertainty estimates

FIG. 5B

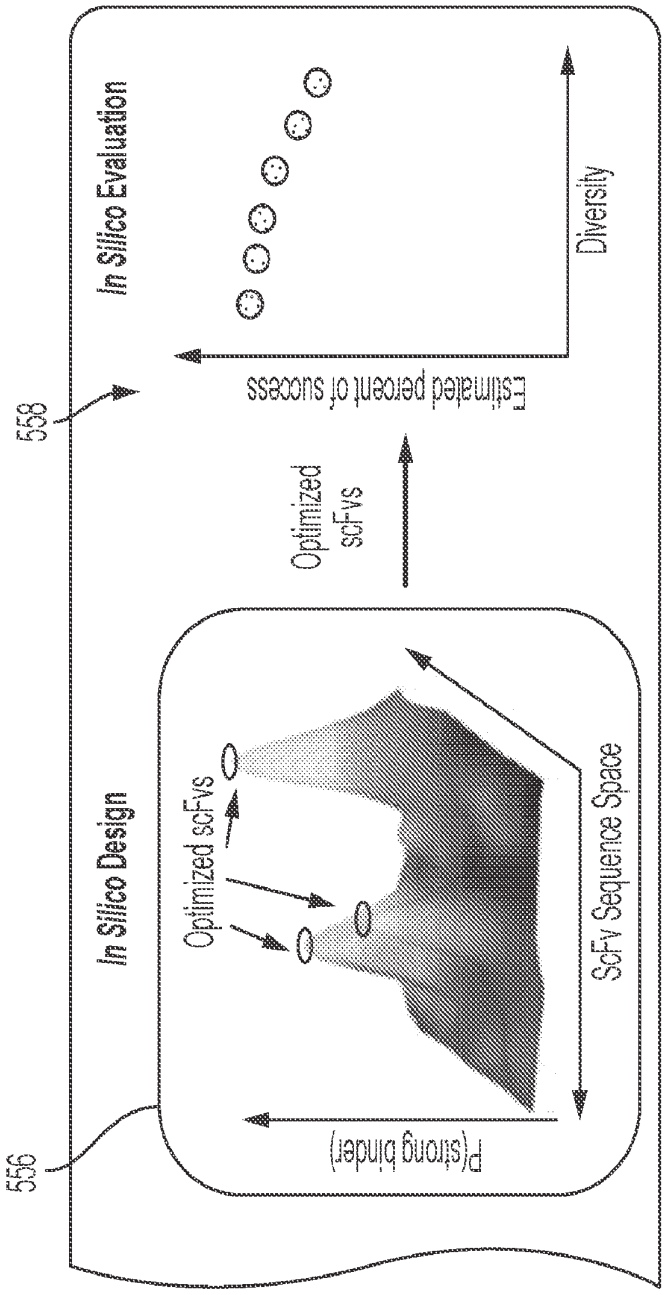


FIG. 5B
CONTINUED

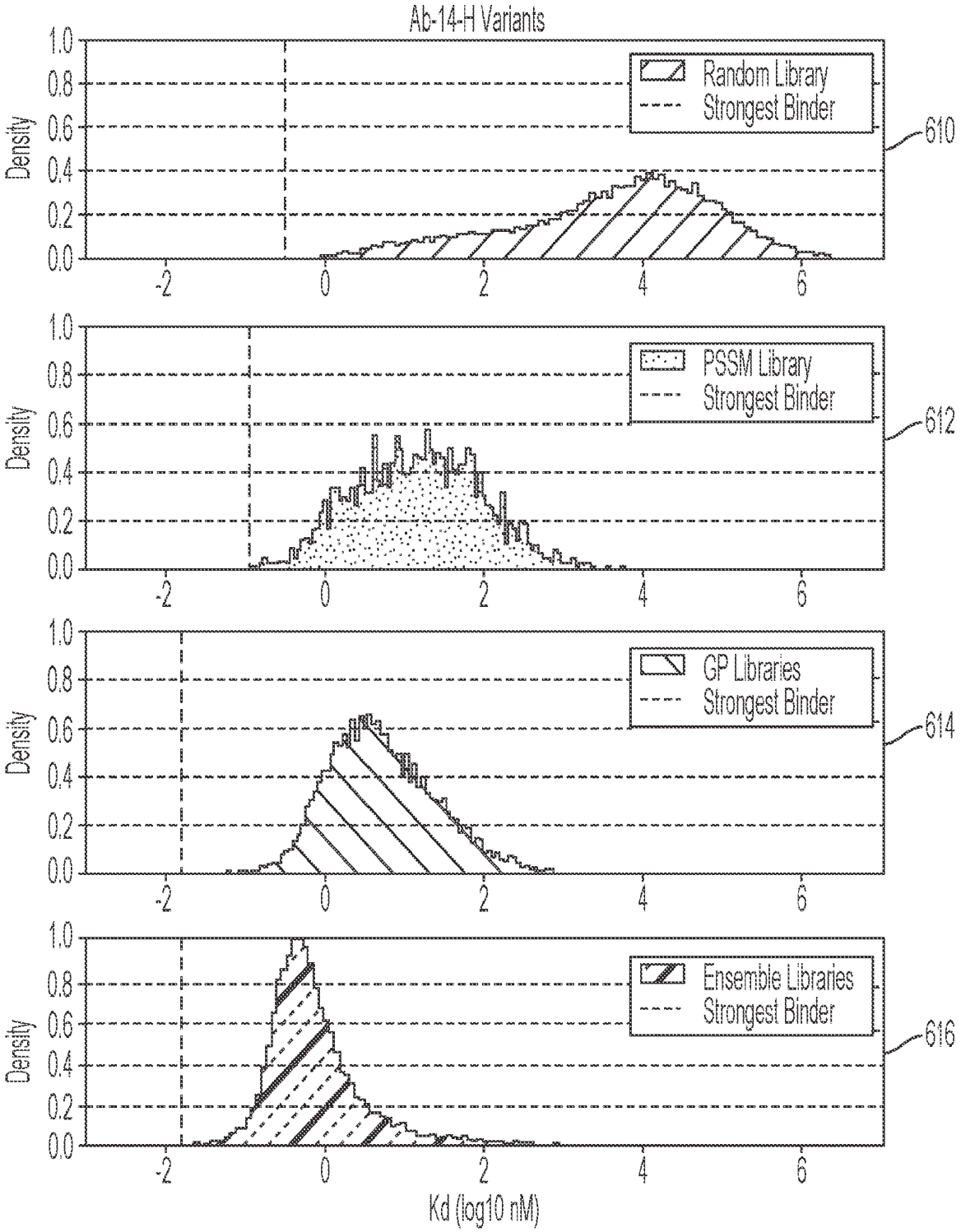

Ab-14-H Variants
Density
Random Library
Strongest Binder
610
PSSM Library
Strongest Binder
612
GP Libraries
Strongest Binder
614
Ensemble Libraries
Strongest Binder
616
Kd (log10 nM)

FIG. 6A

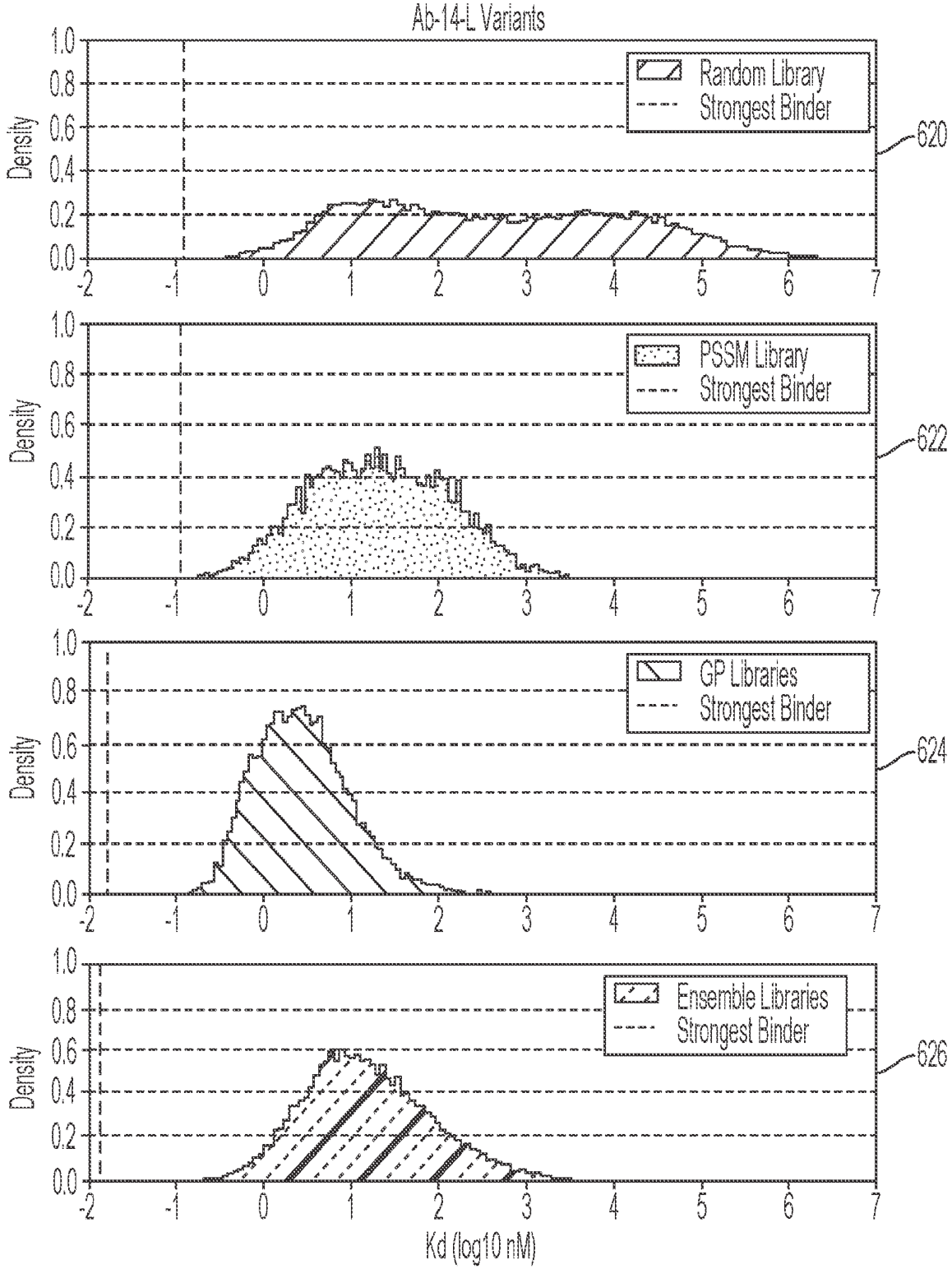

Ab-14-L Variants
Random Library
Strongest Binder
620
PSSM Library
Strongest Binder
622
GP Libraries
Strongest Binder
624
Ensemble Libraries
Strongest Binder
626
Density
Kd (log10 nM)

FIG. 6B

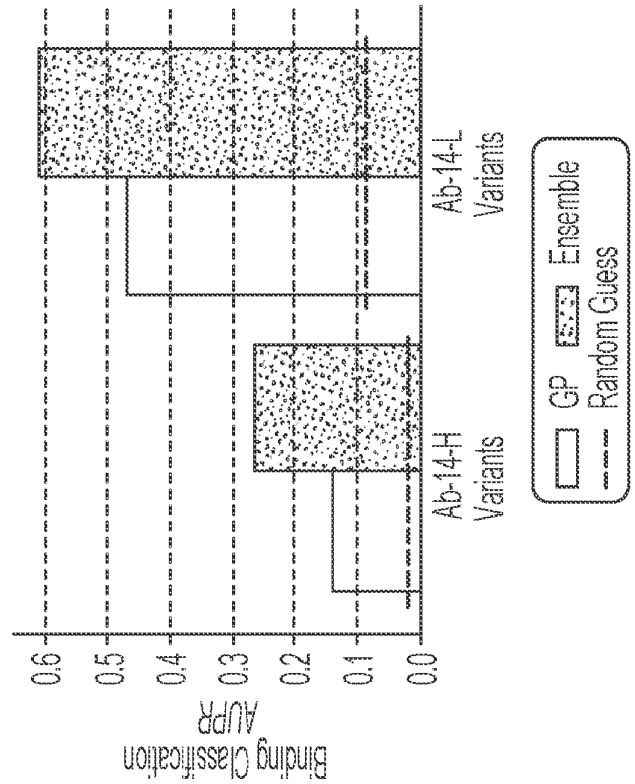

Binding Classification
AUPR
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Ab-14-H Variants
Ab-14-L Variants
GP
Ensemble
Random Guess

FIG. 7B

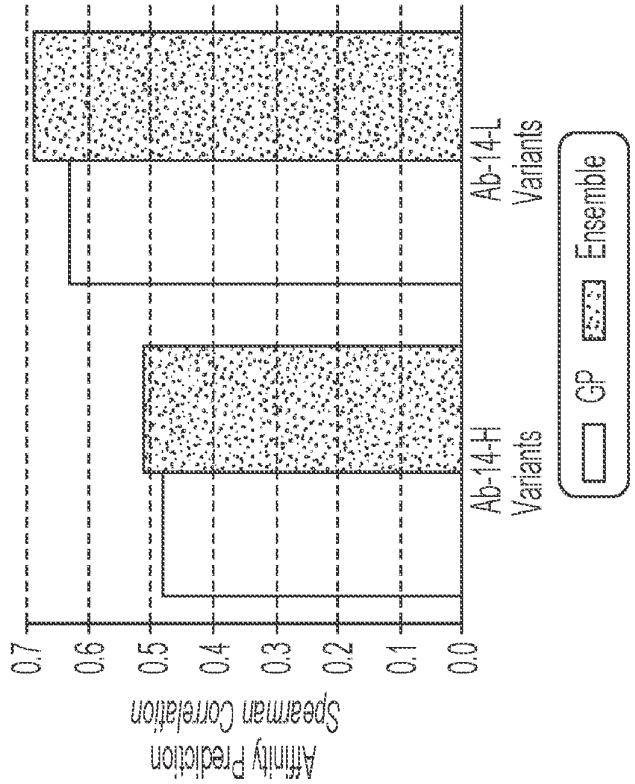

Affinity Prediction
Spearman Correlation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Ab-14-H Variants
Ab-14-L Variants
GP
Ensemble

FIG. 7A

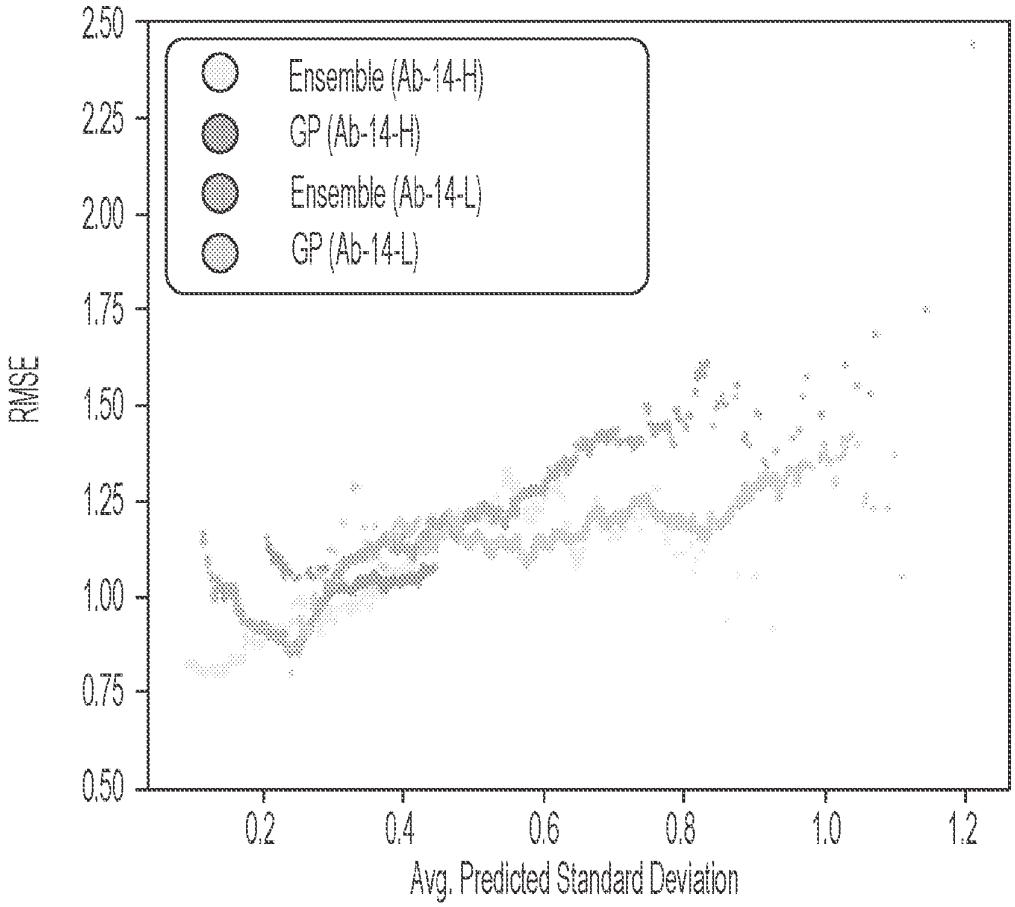

Ensemble (Ab-14-H)
GP (Ab-14-H)
Ensemble (Ab-14-L)
GP (Ab-14-L)
RMSE
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.2
0.4
0.6
0.8
1.0
1.2
Avg. Predicted Standard Deviation

FIG. 7C

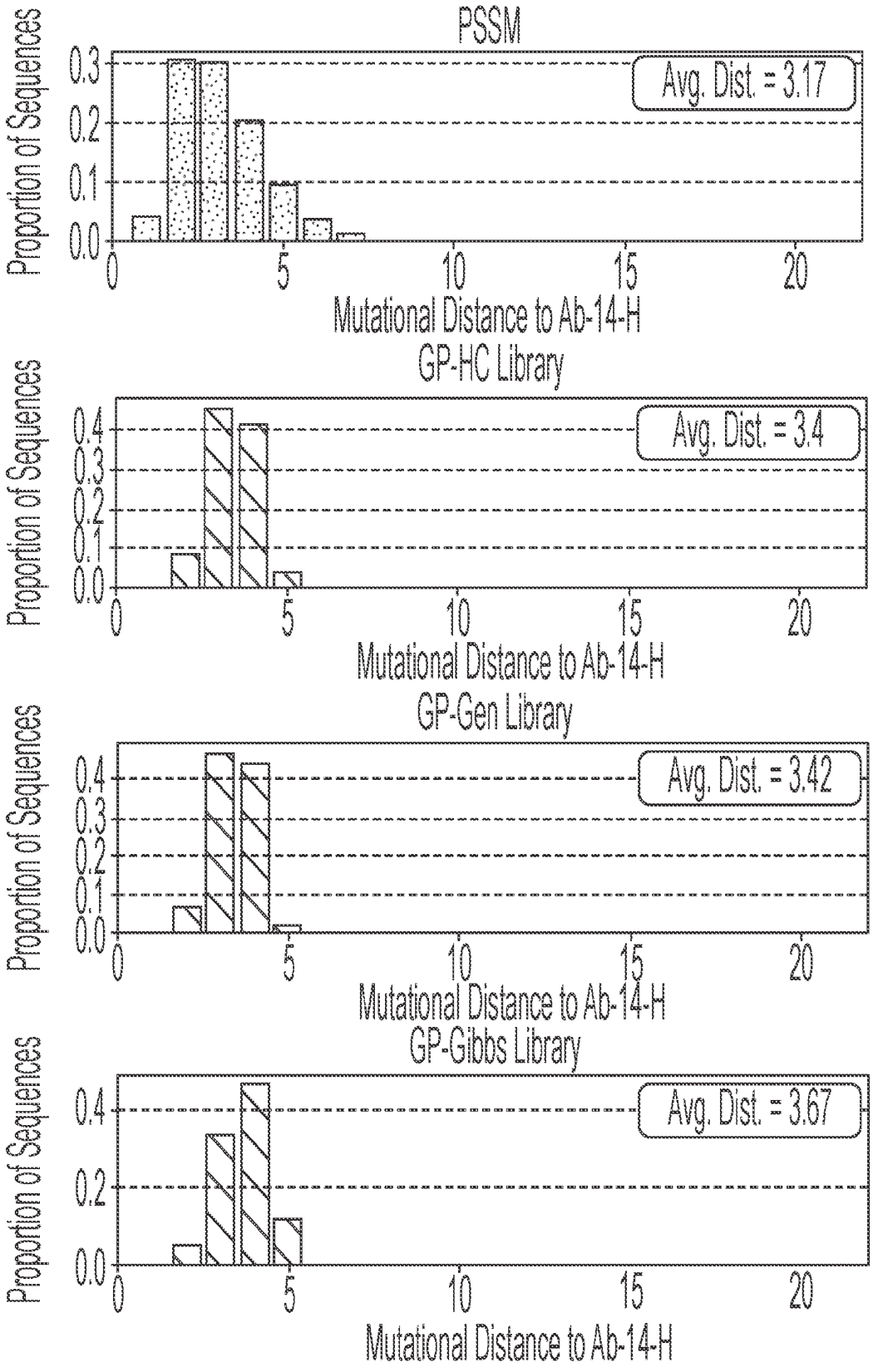

PSSM
Avg. Dist. = 3.17
Proportion of Sequences
0.3
0.2
0.1
0.0
0
5
10
15
20
Mutational Distance to Ab-14-H
GP-HC Library
Avg. Dist. = 3.4
Proportion of Sequences
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
Mutational Distance to Ab-14-H
GP-Gen Library
Avg. Dist. = 3.42
Proportion of Sequences
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
Mutational Distance to Ab-14-H
GP-Gibbs Library
Avg. Dist. = 3.67
Proportion of Sequences
0.4
0.2
0.0
0
5
10
15
20
Mutational Distance to Ab-14-H

FIG. 8A

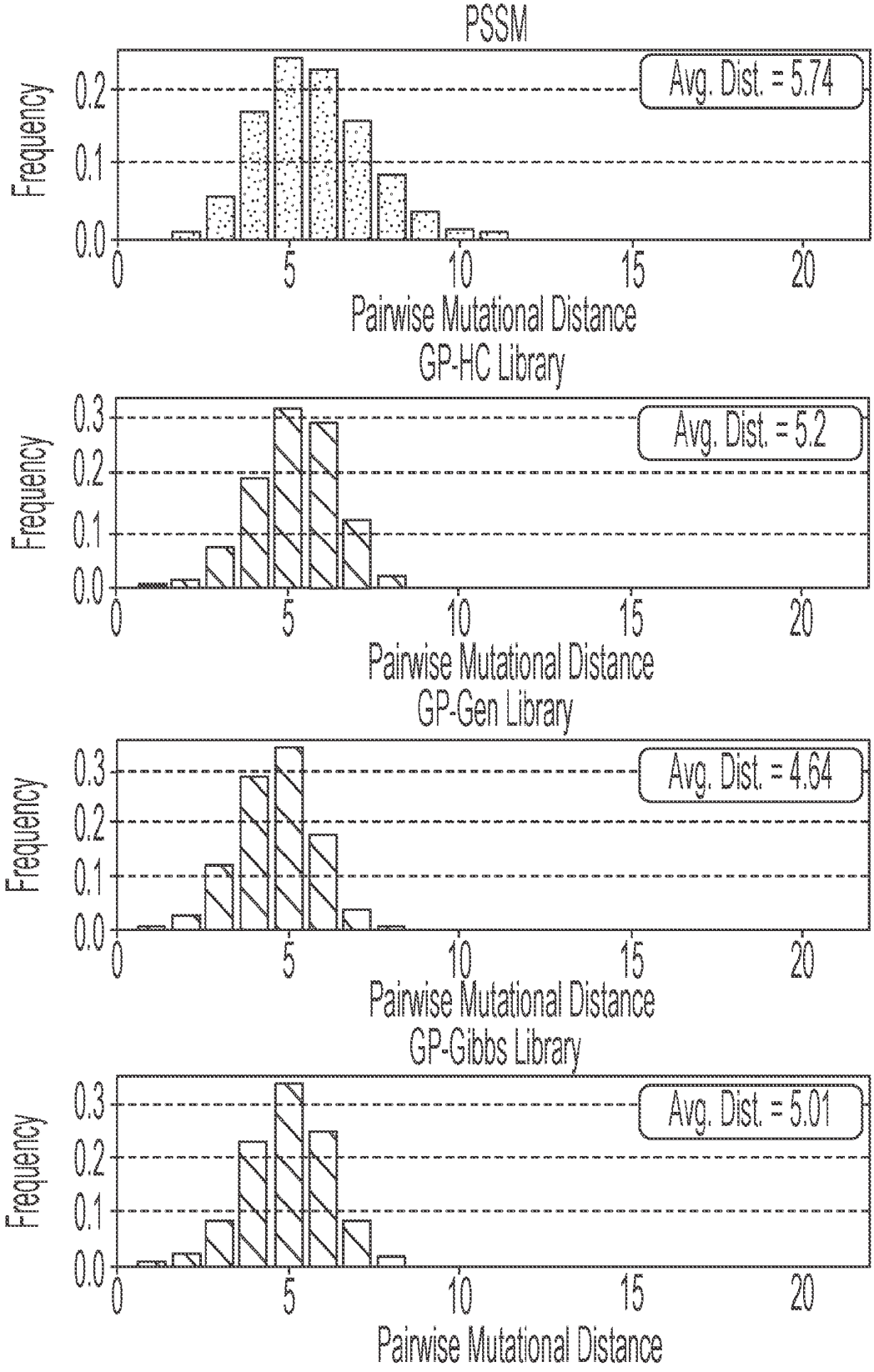


FIG. 8A
CONTINUED

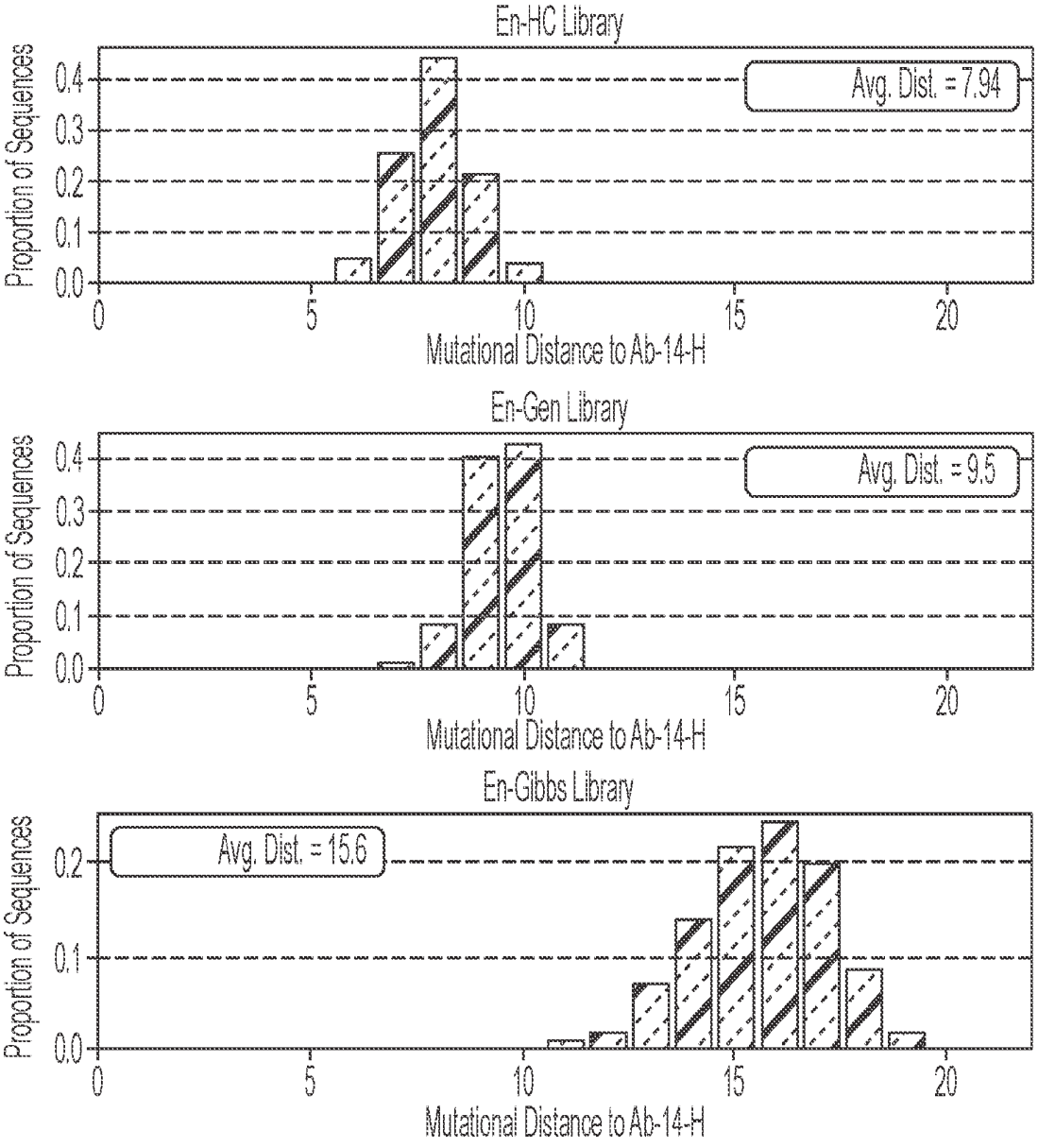

En-HC Library
Avg. Dist. = 7.94
Proportion of Sequences
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
Mutational Distance to Ab-14-H
En-Gen Library
Avg. Dist. = 9.5
Proportion of Sequences
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
Mutational Distance to Ab-14-H
En-Gibbs Library
Avg. Dist. = 15.6
Proportion of Sequences
0.0
0.1
0.2
0
5
10
15
20
Mutational Distance to Ab-14-H

FIG. 8B

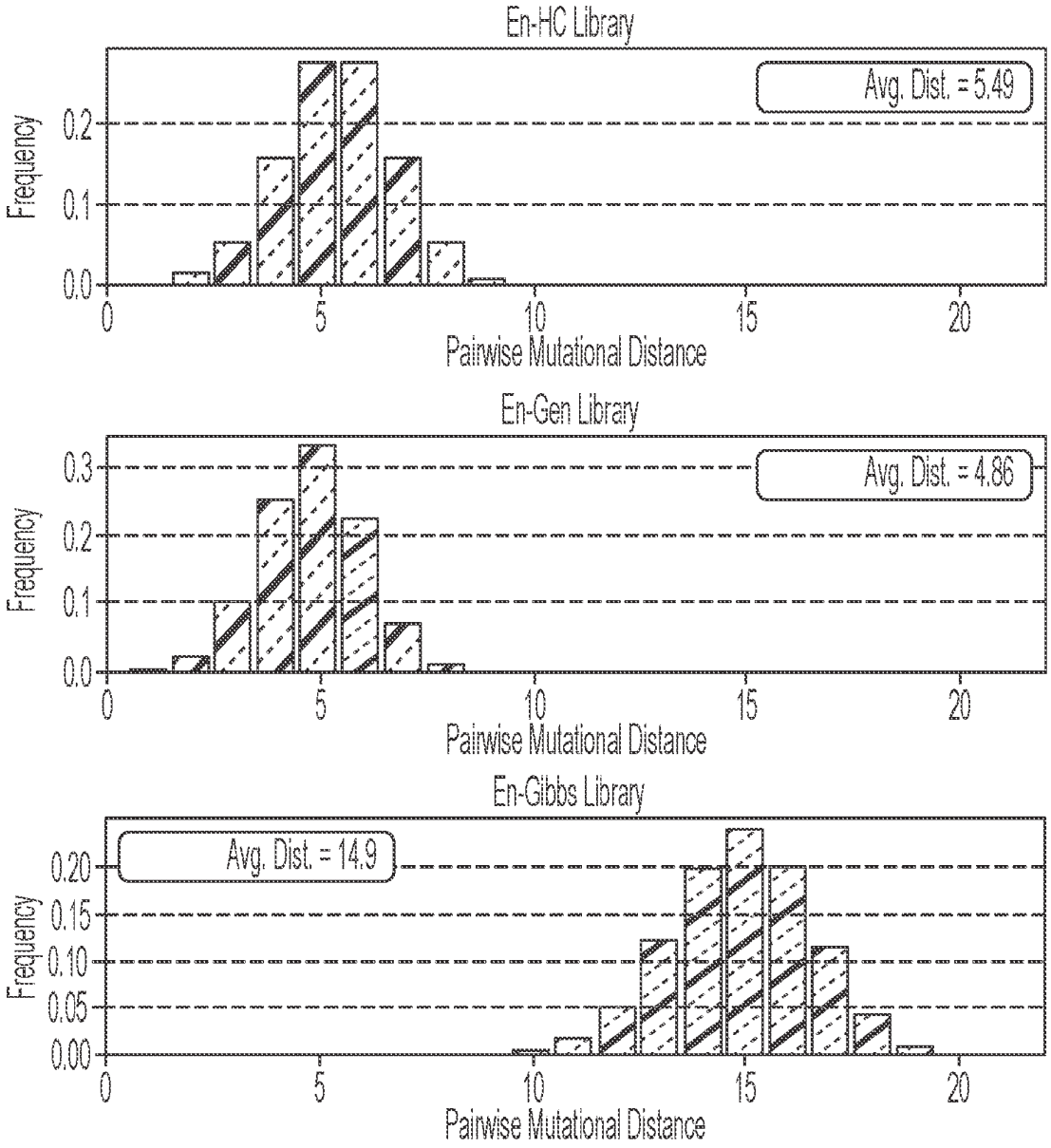


FIG. 8B
CONTINUED

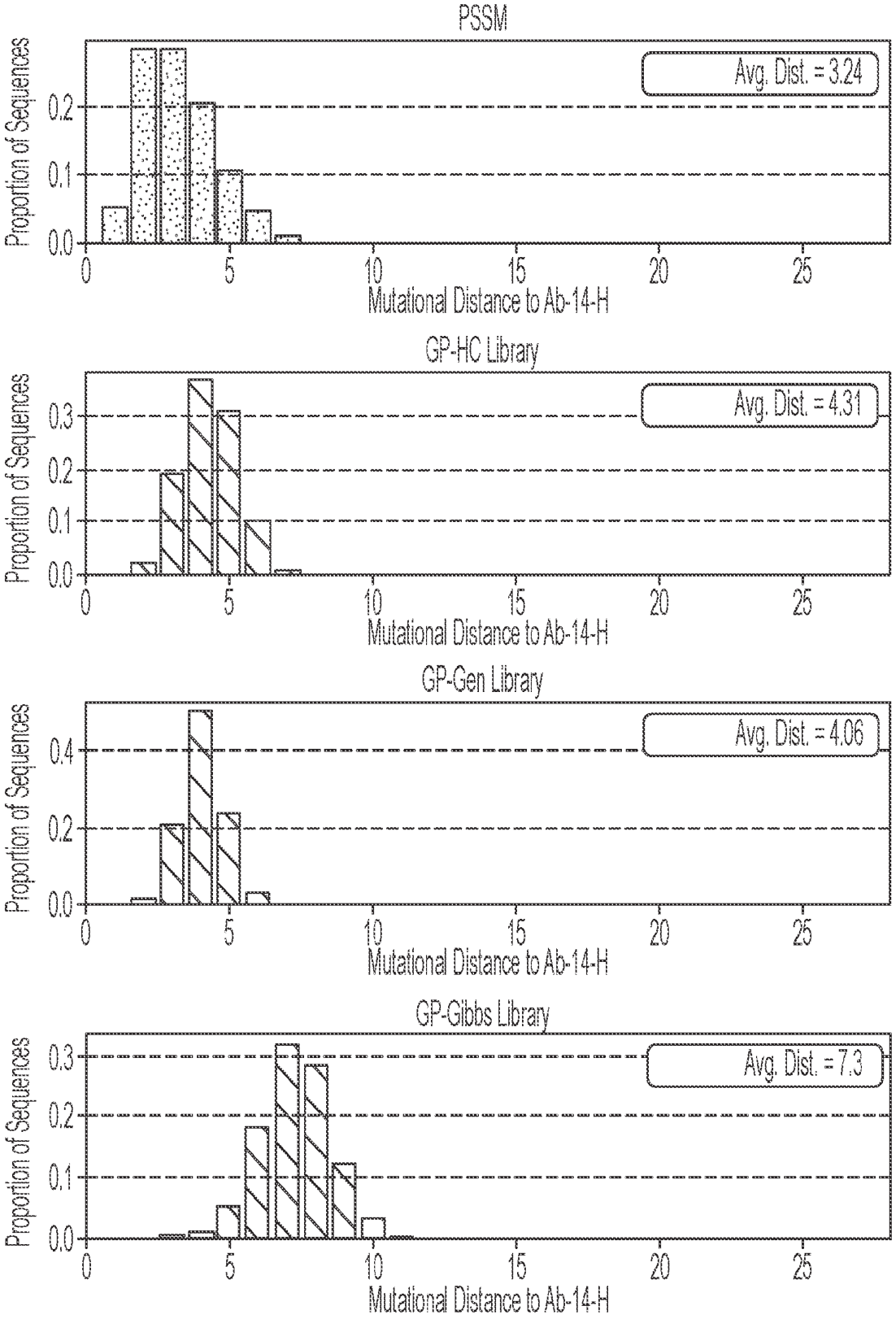

PSSM
Avg. Dist. = 3.24
Proportion of Sequences
0.0
0.1
0.2
0
5
10
15
20
25
Mutational Distance to Ab-14-H
GP-HC Library
Avg. Dist. = 4.31
Proportion of Sequences
0.0
0.1
0.2
0.3
0
5
10
15
20
25
Mutational Distance to Ab-14-H
GP-Gen Library
Avg. Dist. = 4.06
Proportion of Sequences
0.0
0.2
0.4
0
5
10
15
20
25
Mutational Distance to Ab-14-H
GP-Gibbs Library
Avg. Dist. = 7.3
Proportion of Sequences
0.0
0.1
0.2
0.3
0
5
10
15
20
25
Mutational Distance to Ab-14-H

FIG. 9A

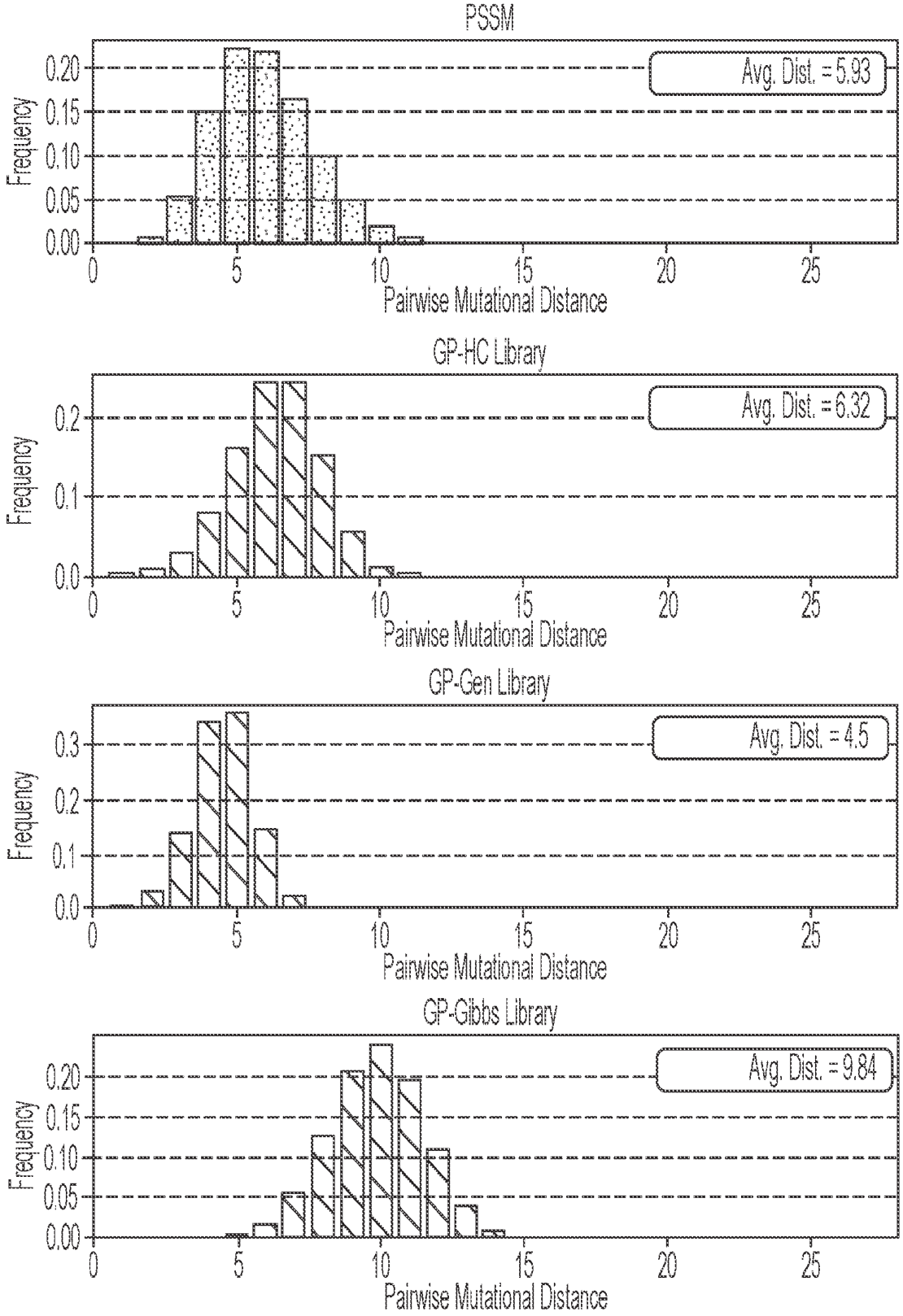


FIG. 9A
CONTINUED

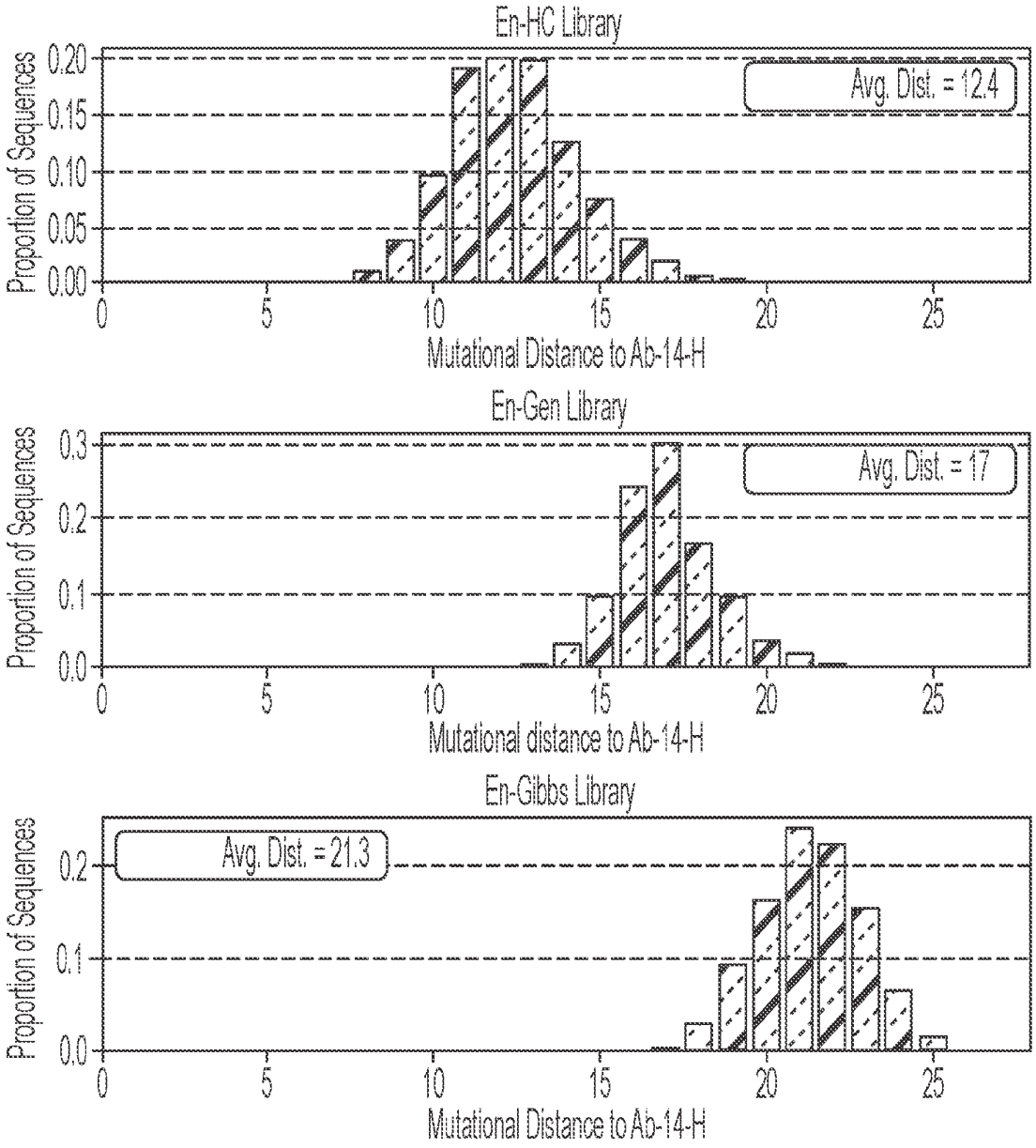

En-HC Library
Avg. Dist. = 12.4
Proportion of Sequences
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
25
Mutational Distance to Ab-14-H
En-Gen Library
Avg. Dist. = 17
Proportion of Sequences
0.3
0.2
0.1
0.0
0
5
10
15
20
25
Mutational distance to Ab-14-H
En-Gibbs Library
Avg. Dist. = 21.3
Proportion of Sequences
0.2
0.1
0.0
0
5
10
15
20
25
Mutational Distance to Ab-14-H

FIG. 9B

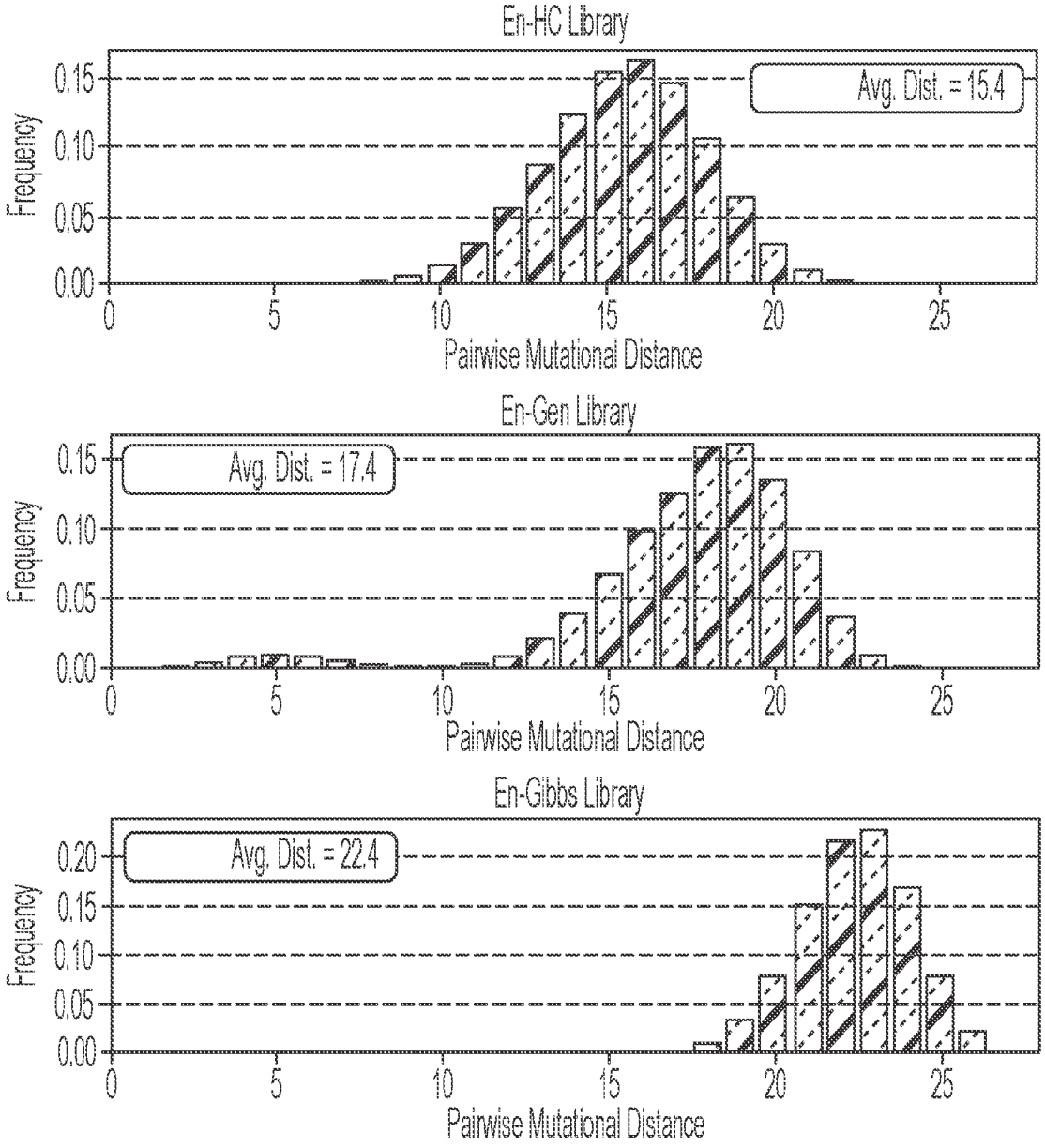


FIG. 9B
CONTINUED

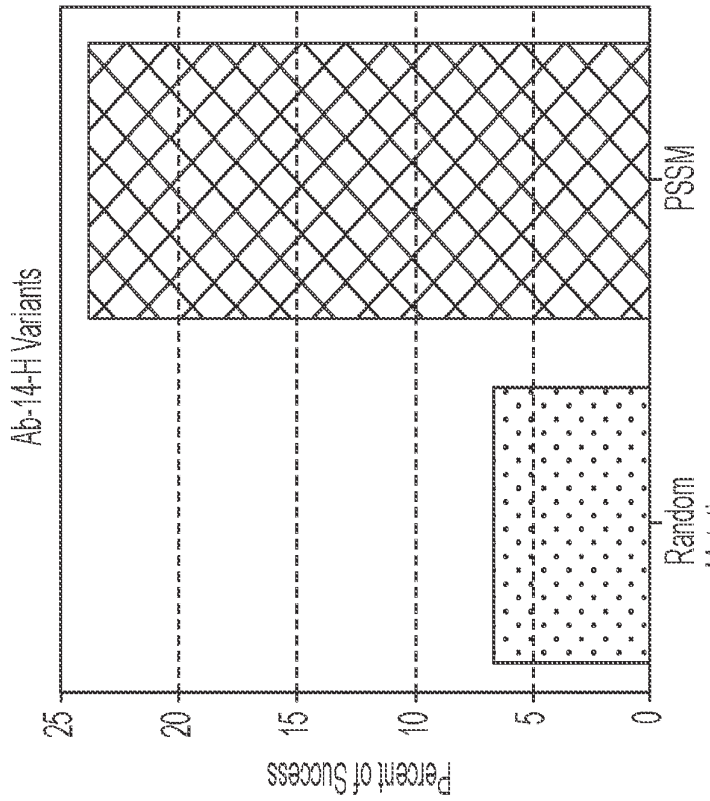


FIG. 10B

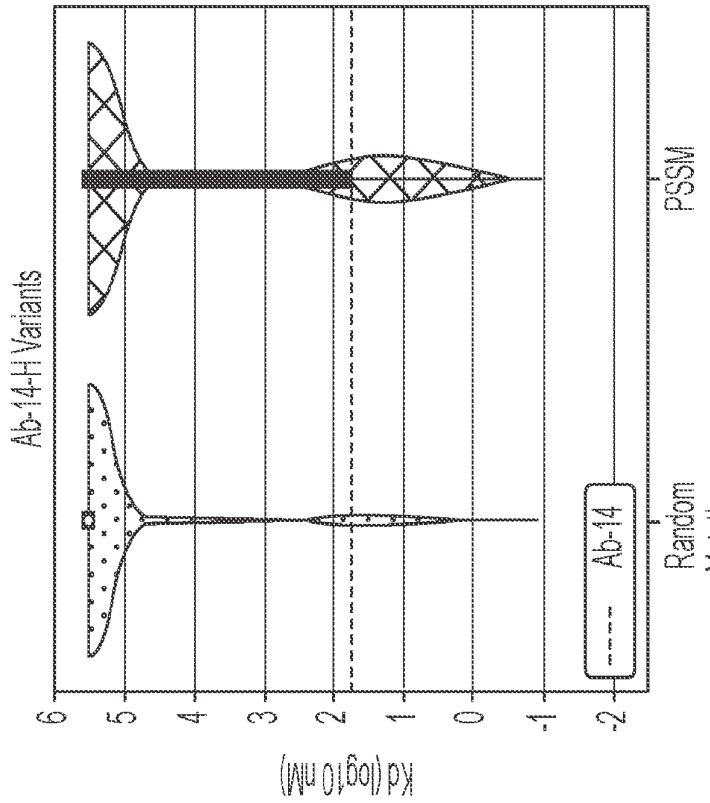


FIG. 10A

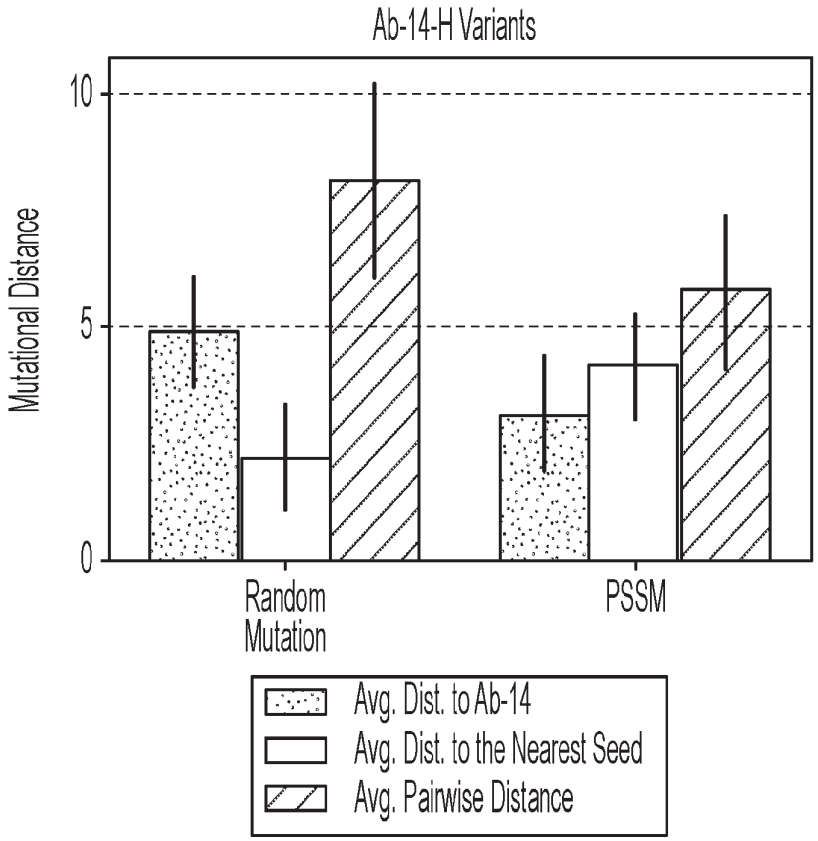

Ab-14-H Variants
Mutational Distance
10
5
0
Random Mutation
PSSM
Avg. Dist. to Ab-14
Avg. Dist. to the Nearest Seed
Avg. Pairwise Distance

FIG. 10C

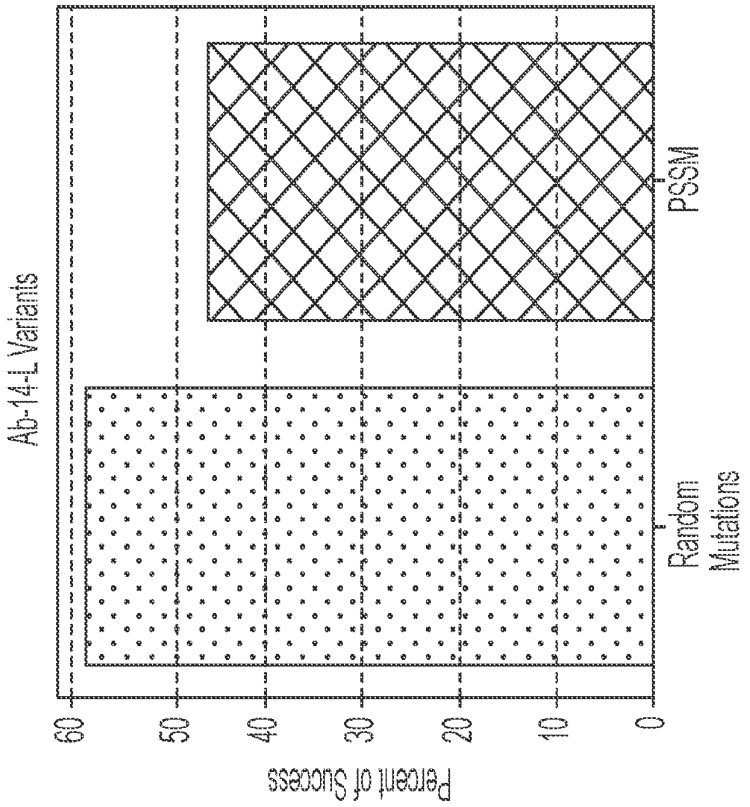


FIG. 10E

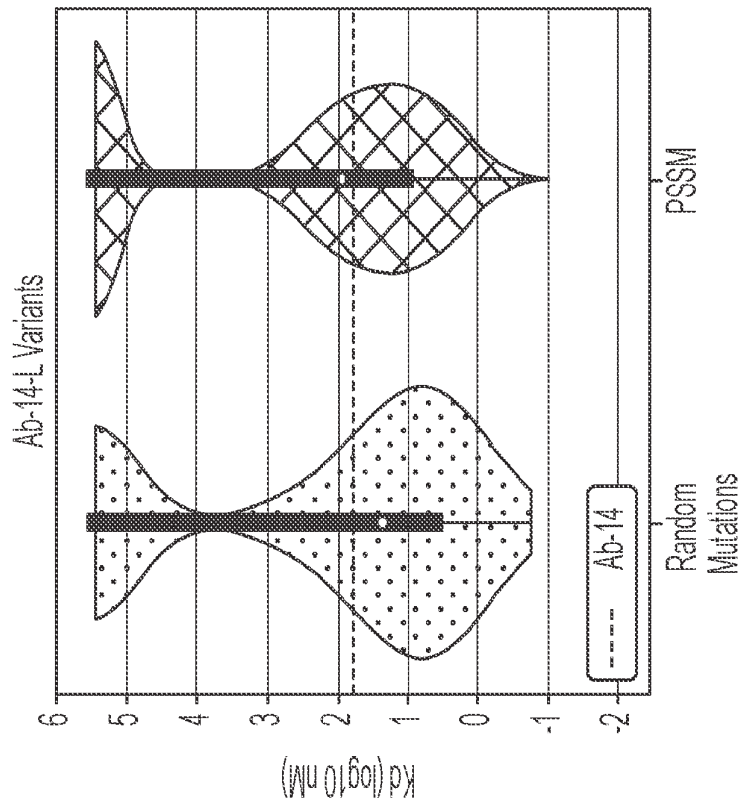


FIG. 10D

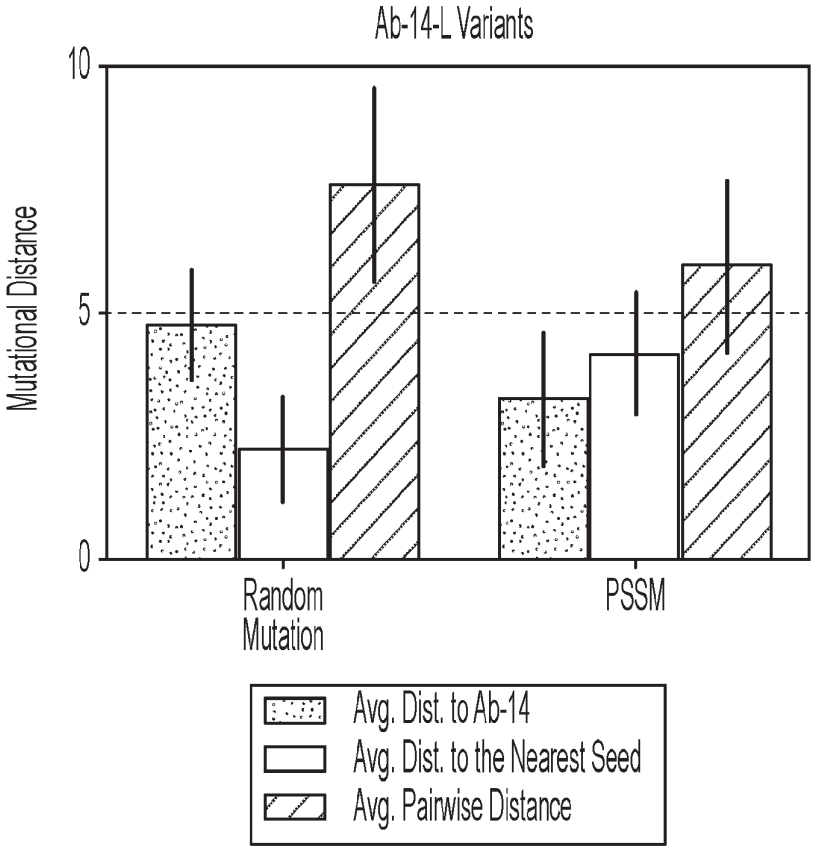

Ab-14-L Variants
Mutational Distance
10
5
0
Random Mutation
PSSM
Avg. Dist. to Ab-14
Avg. Dist. to the Nearest Seed
Avg. Pairwise Distance

FIG. 10F

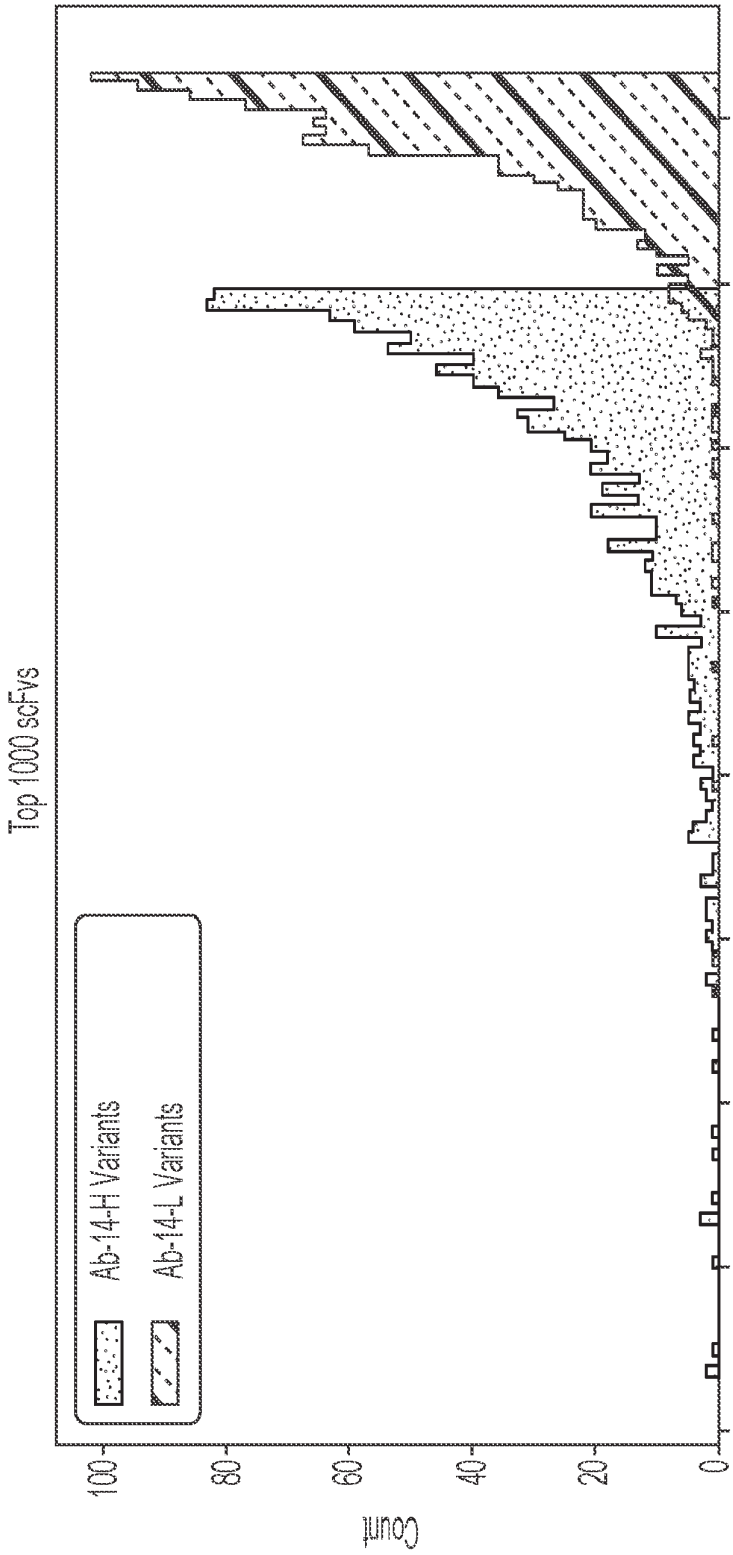

Top 1000 scFvs
Ab-14-H Variants
Ab-14-L Variants
Count
100
80
60
40
20
0
-2.50
-2.25
-2.00
-1.75
-1.50
-1.25
-1.00
-0.75
-0.50
Kd (log10 nM)

FIG. 11

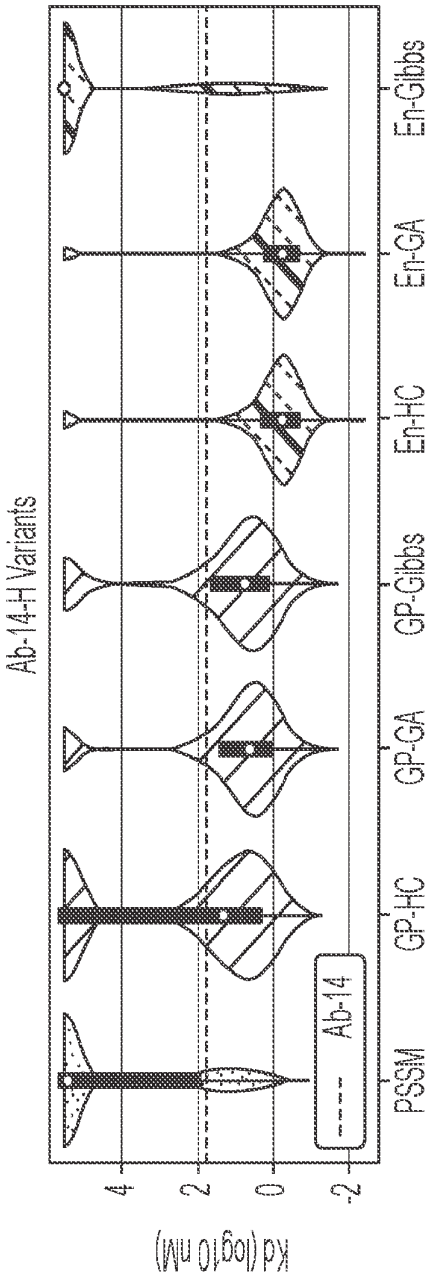


FIG. 12A

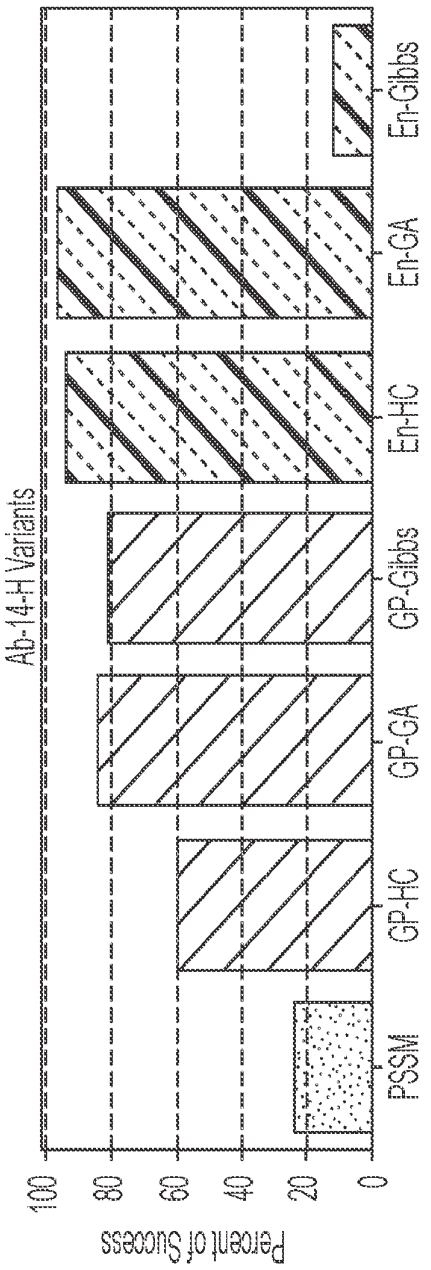


FIG. 12B

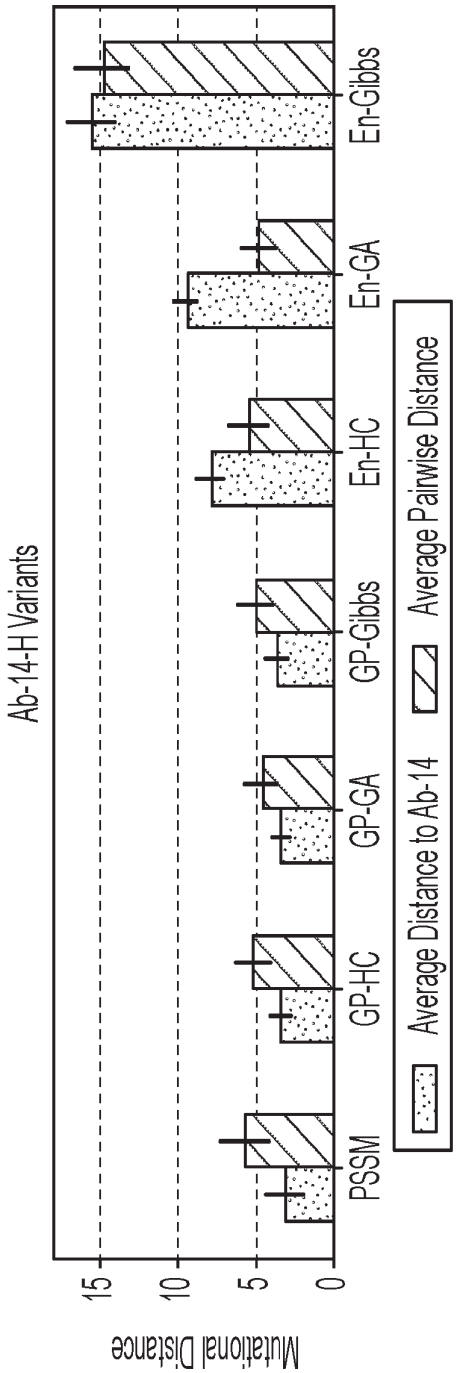

Mutational Distance
0
5
10
15
Ab-14-H Variants
PSSM
GP-HC
GP-GA
GP-Gibbs
En-HC
En-GA
En-Gibbs
Average Distance to Ab-14
Average Pairwise Distance

FIG. 12C

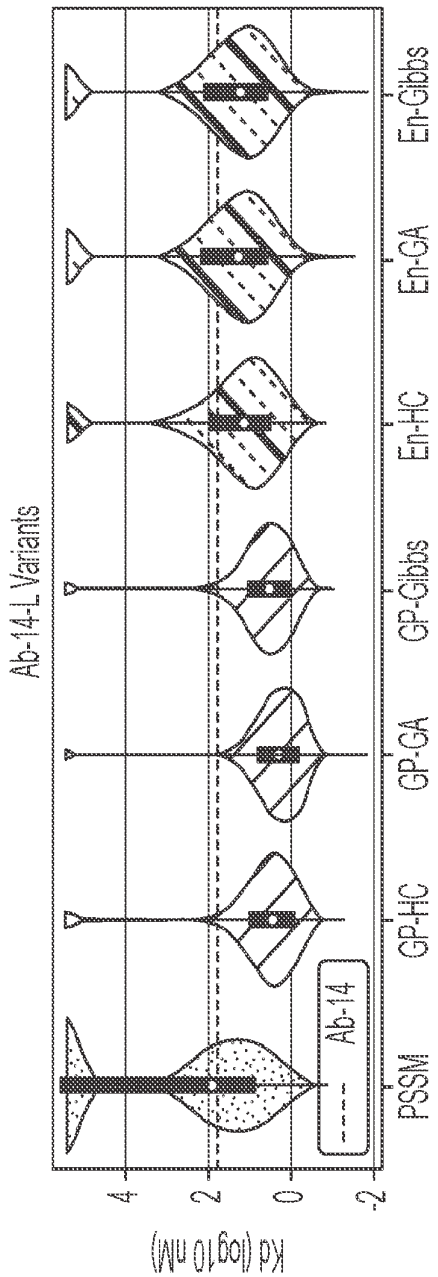


FIG. 13A

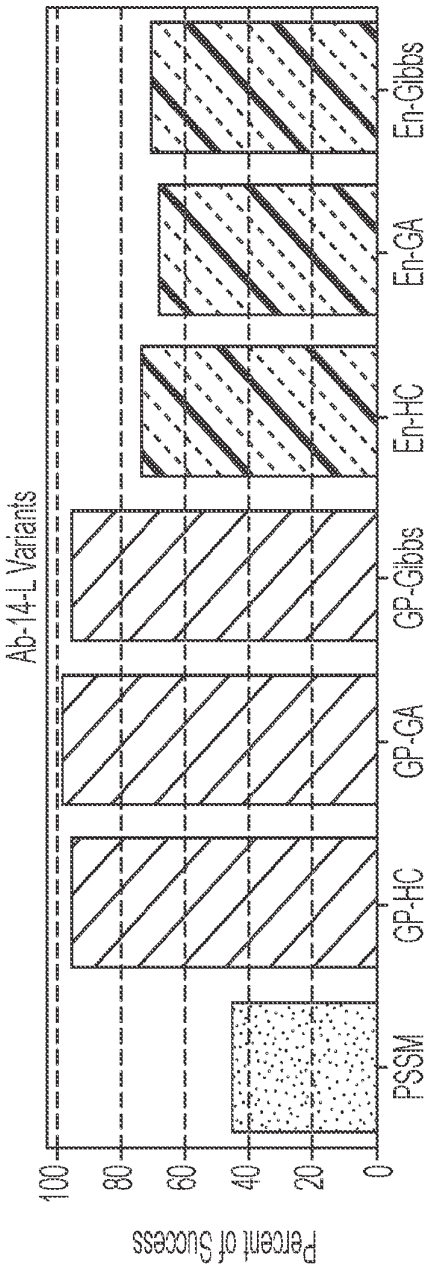


FIG. 13B

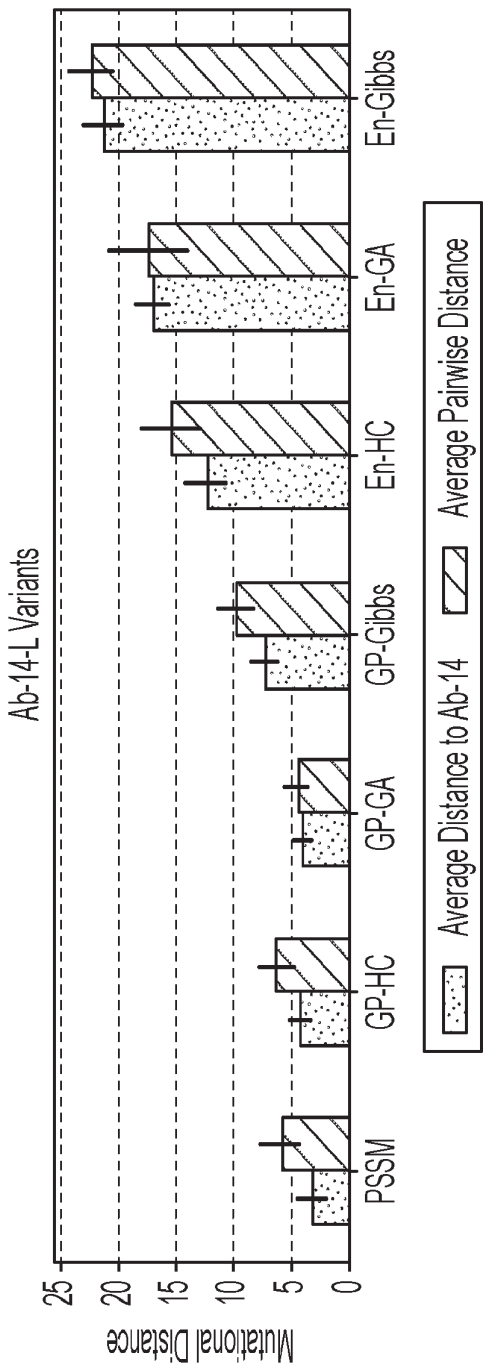

Ab-14-L Variants
Mutational Distance
25
20
15
10
5
0
PSSM
GP-HC
GP-GA
GP-Gibbs
En-HC
En-GA
En-Gibbs
Average Distance to Ab-14
Average Pairwise Distance

FIG. 13C

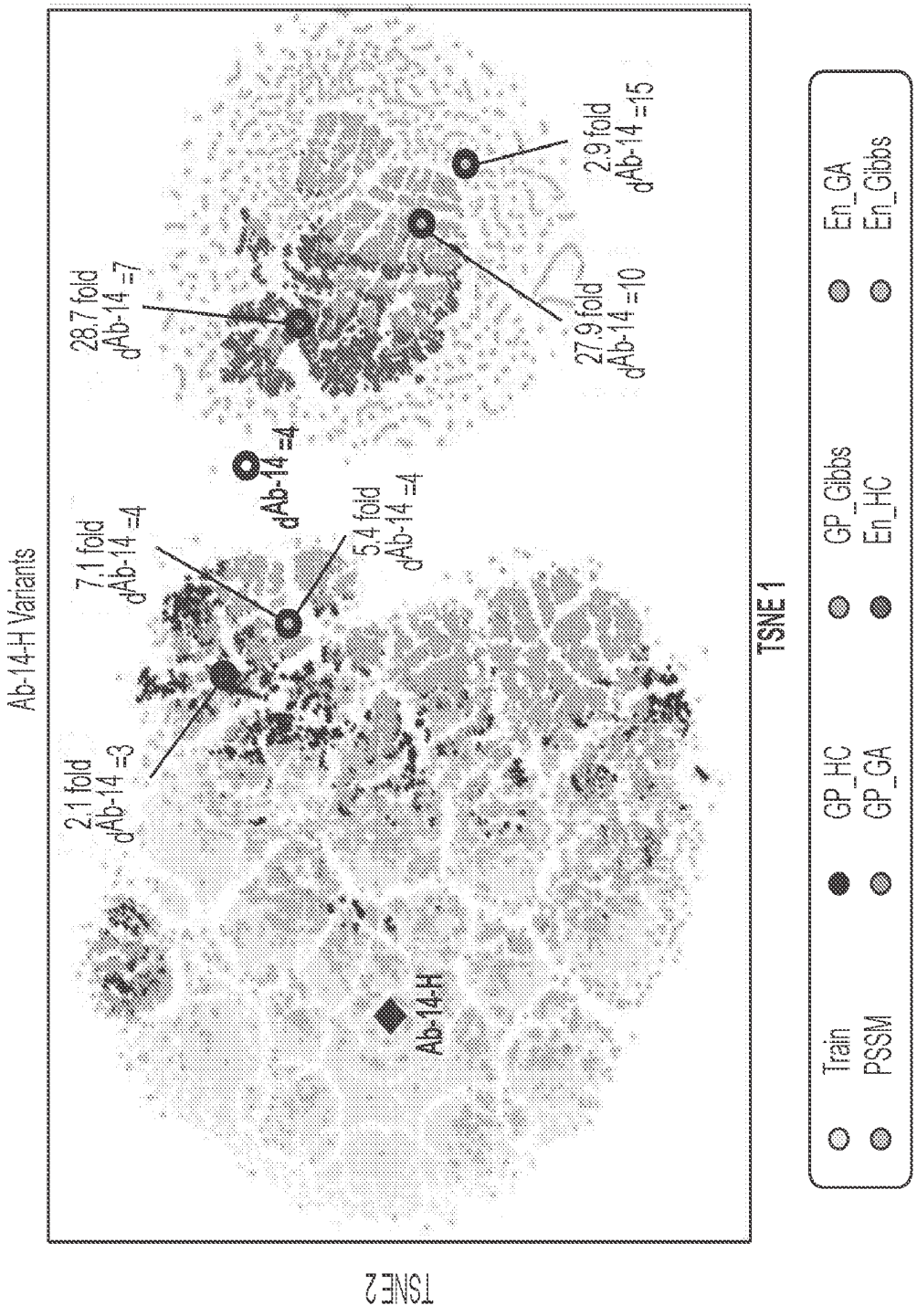

Ab-14-H Variants
2.1 fold
dAb-14 =3
7.1 fold
dAb-14 =4
dAb-14 =4
5.4 fold
dAb-14 =4
28.7 fold
dAb-14 =7
27.9 fold
dAb-14 =10
2.9 fold
dAb-14 =15
Ab-14-H
TSNE 1
TSNE 2
Train
PSSM
GP_HC
GP_GA
GP_Gibbs
En_HC
En_GA
En_Gibbs

FIG. 14A

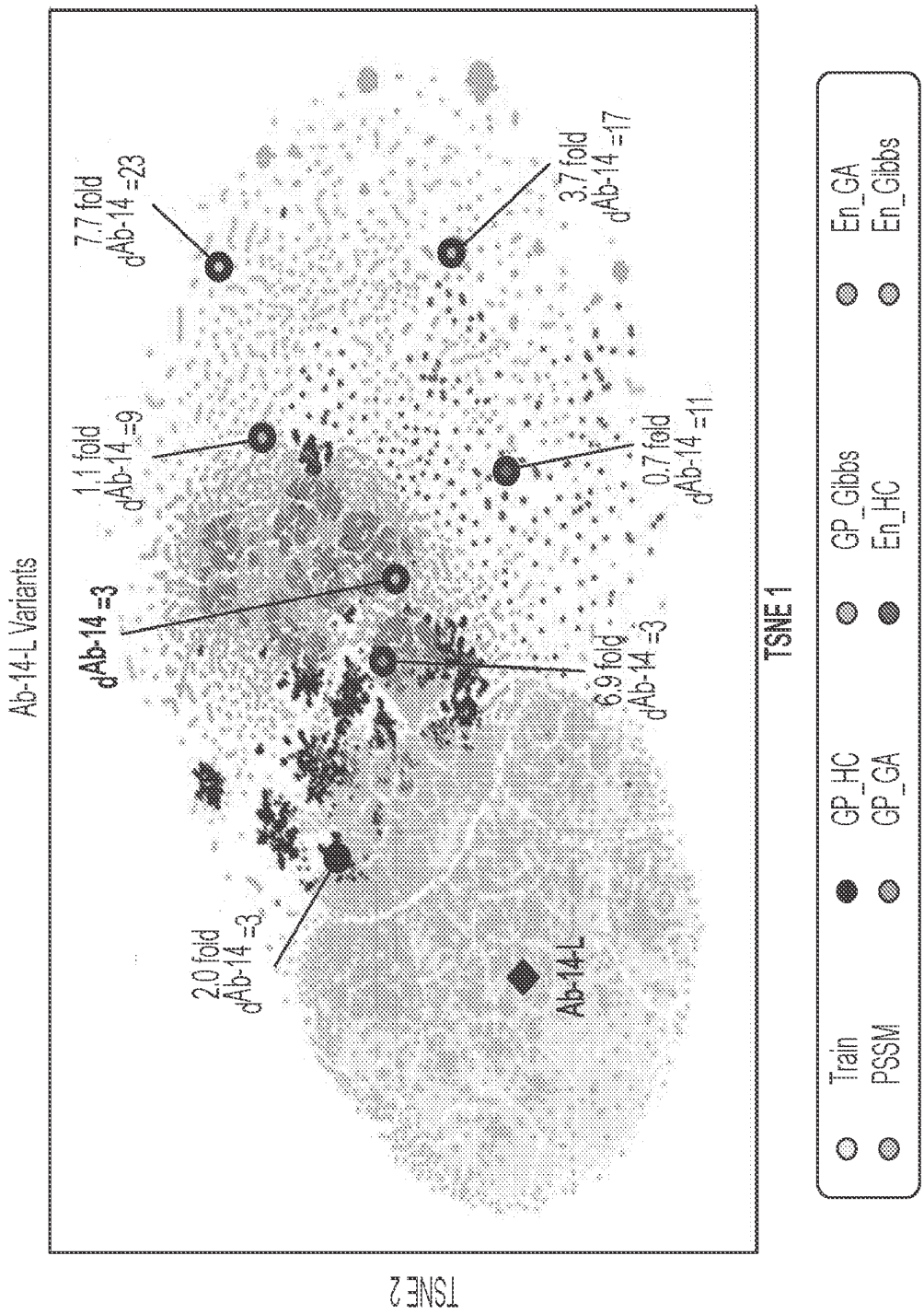

Ab-14-L Variants
TSNE 2
TSNE 1
2.0 fold
dAb-14 =3
dAb-14 =3
1.1 fold
dAb-14 =9
7.7 fold
dAb-14 =23
6.9 fold
dAb-14 =3
0.7 fold
dAb-14 =11
3.7 fold
dAb-14 =17
Ab-14-L
Train
PSSM
GP_HC
GP_GA
GP_Gibbs
En_HC
En_GA
En_Gibbs

FIG. 14B

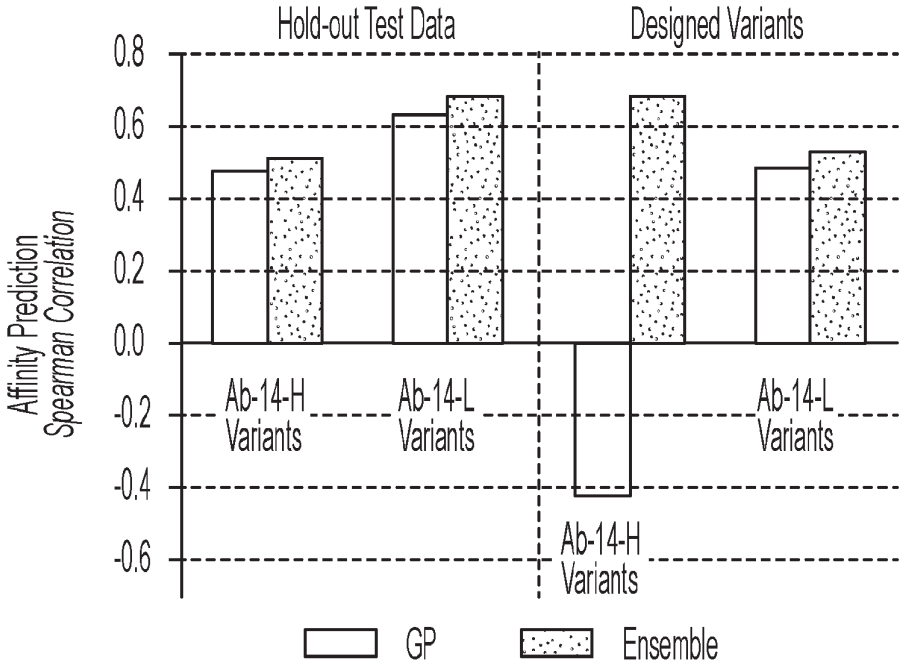

Hold-out Test Data
Designed Variants
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
Affinity Prediction
Spearman Correlation
Ab-14-H Variants
Ab-14-L Variants
Ab-14-H Variants
Ab-14-L Variants
GP
Ensemble

FIG. 15A

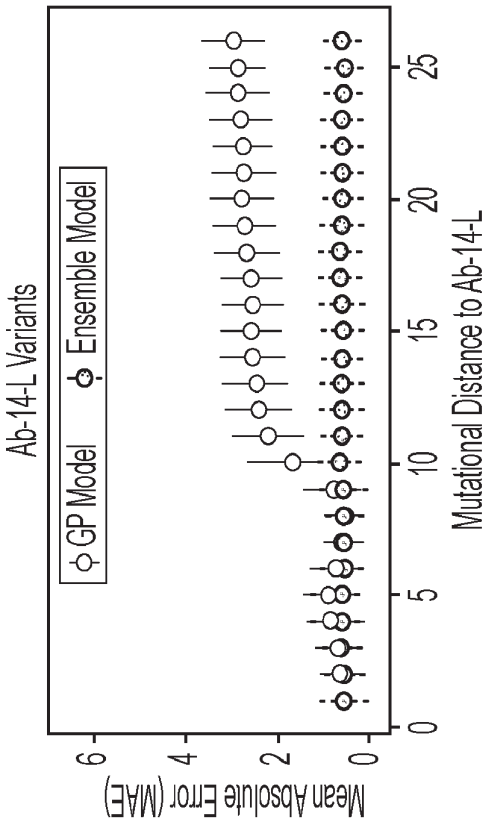


FIG. 15C

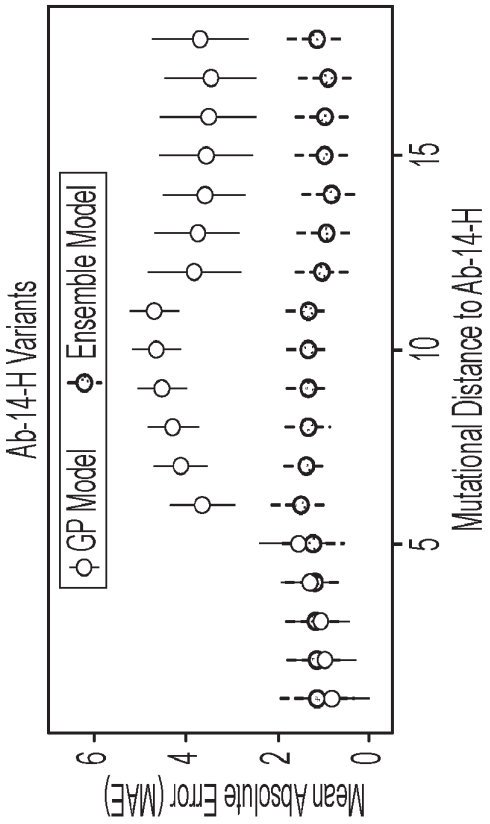


FIG. 15B

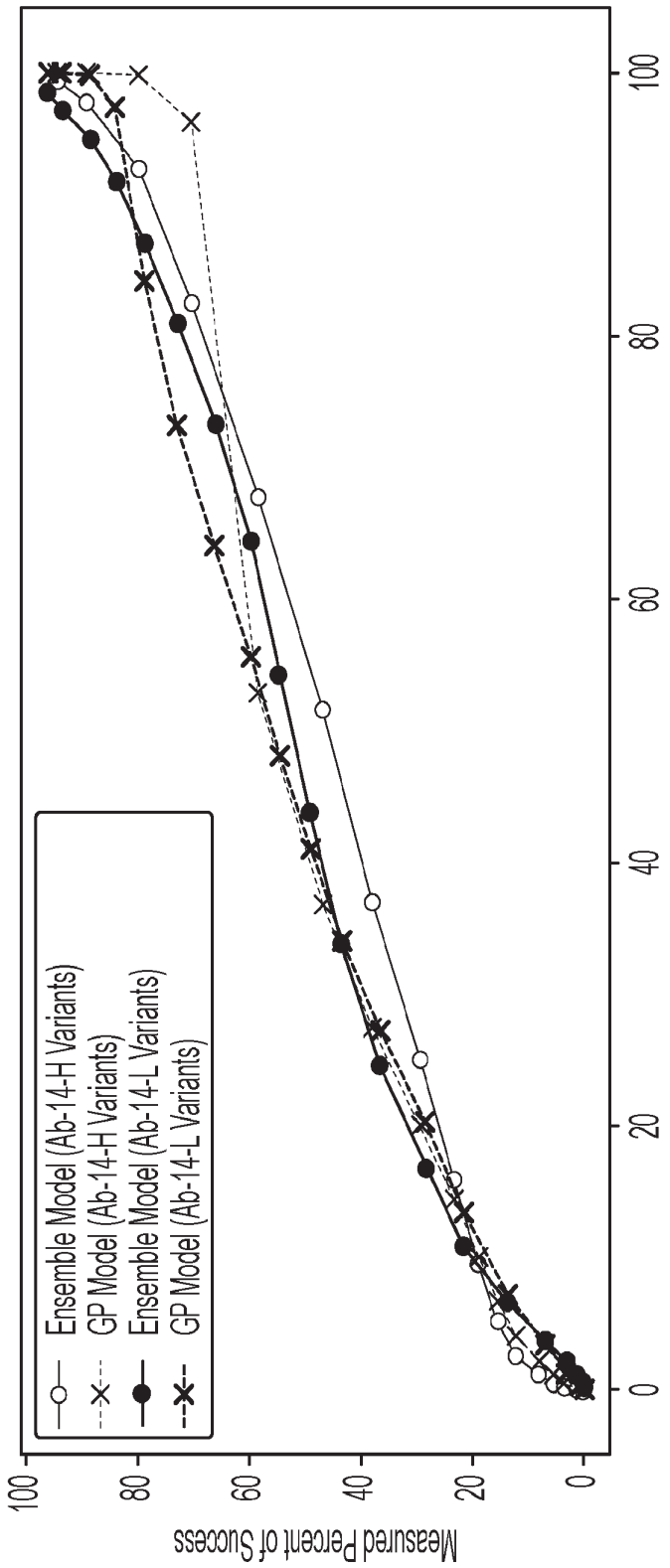

Ensemble Model (Ab-14-H Variants)
GP Model (Ab-14-H Variants)
Ensemble Model (Ab-14-L Variants)
GP Model (Ab-14-L Variants)
Measured Percent of Success
Estimated Percent of Success
0
20
40
60
80
100

FIG. 16

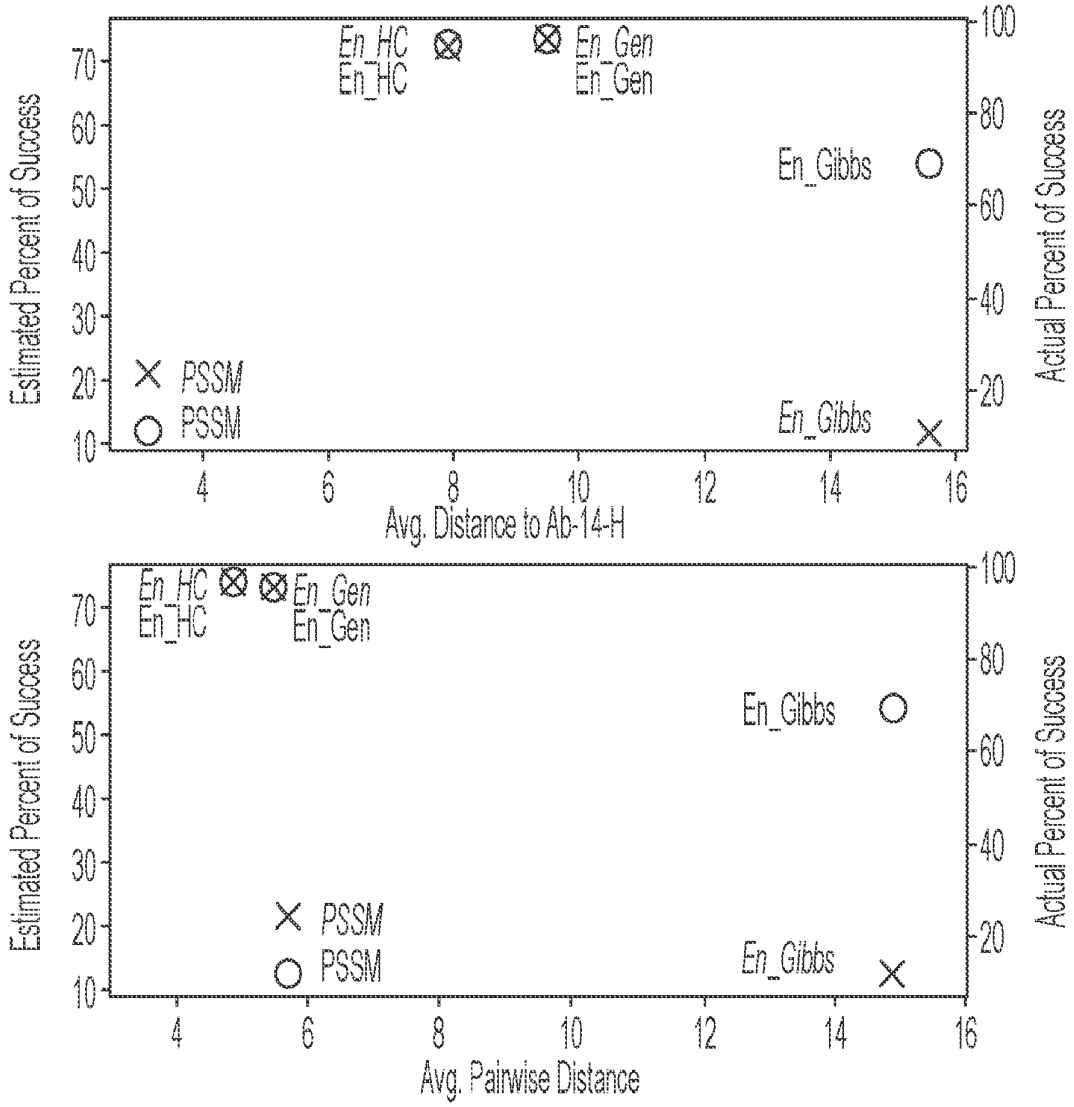

Estimated Percent of Success
Actual Percent of Success
Avg. Distance to Ab-14-H
En_HC
En_HC
En_Gen
En_Gen
En_Gibbs
PSSM
PSSM
En_Gibbs
Avg. Pairwise Distance

FIG. 17A

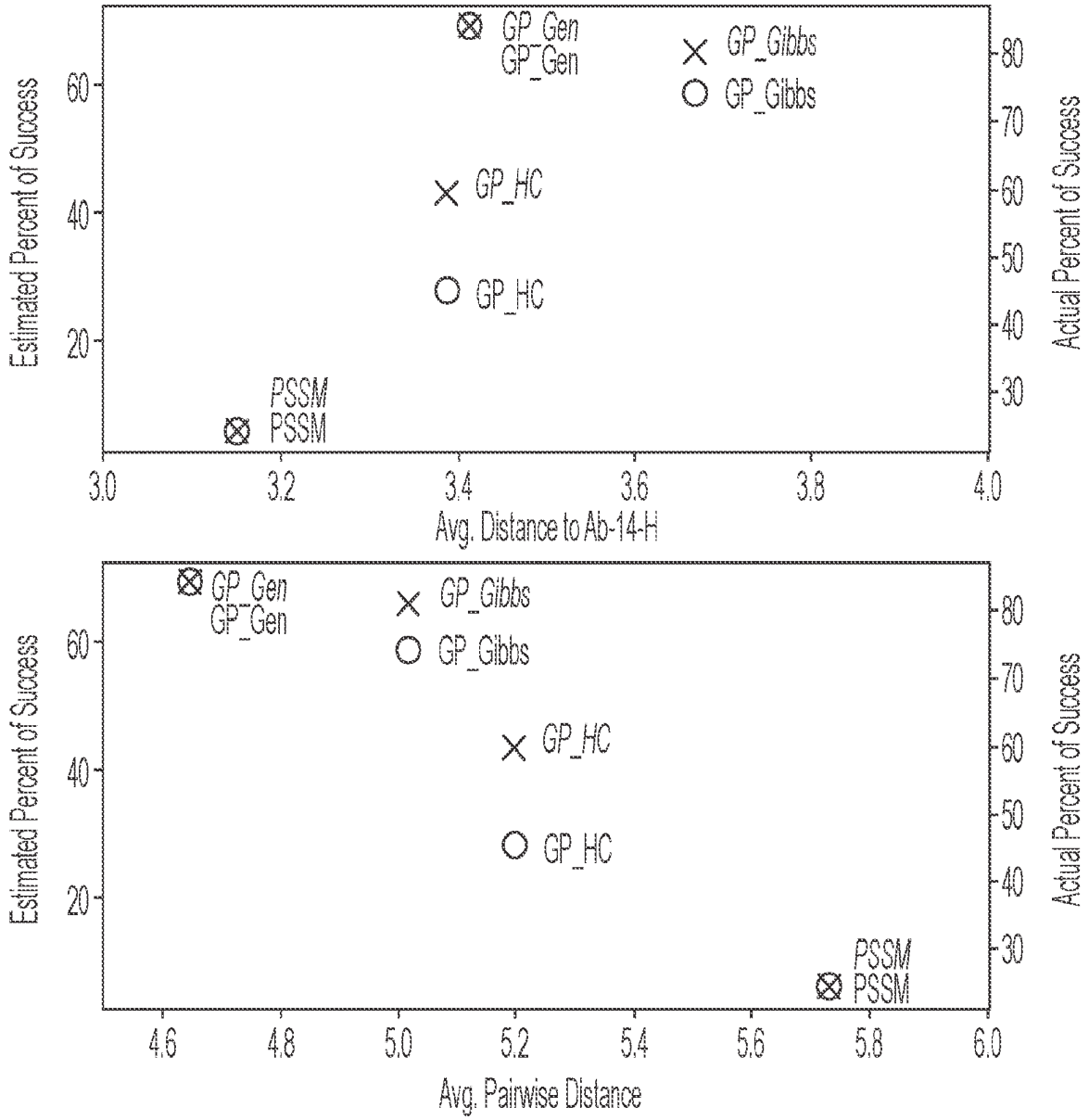

Estimated Percent of Success
Actual Percent of Success
Avg. Distance to Ab-14-H
GP_Gen
GP_Gen
GP_Gibbs
GP_Gibbs
GP_HC
GP_HC
PSSM
PSSM
3.0
3.2
3.4
3.6
3.8
4.0
20
40
60
30
40
50
60
70
80
Estimated Percent of Success
Actual Percent of Success
Avg. Pairwise Distance
GP_Gen
GP_Gen
GP_Gibbs
GP_Gibbs
GP_HC
GP_HC
PSSM
PSSM
4.6
4.8
5.0
5.2
5.4
5.6
5.8
6.0
20
40
60
30
40
50
60
70
80

FIG. 17B

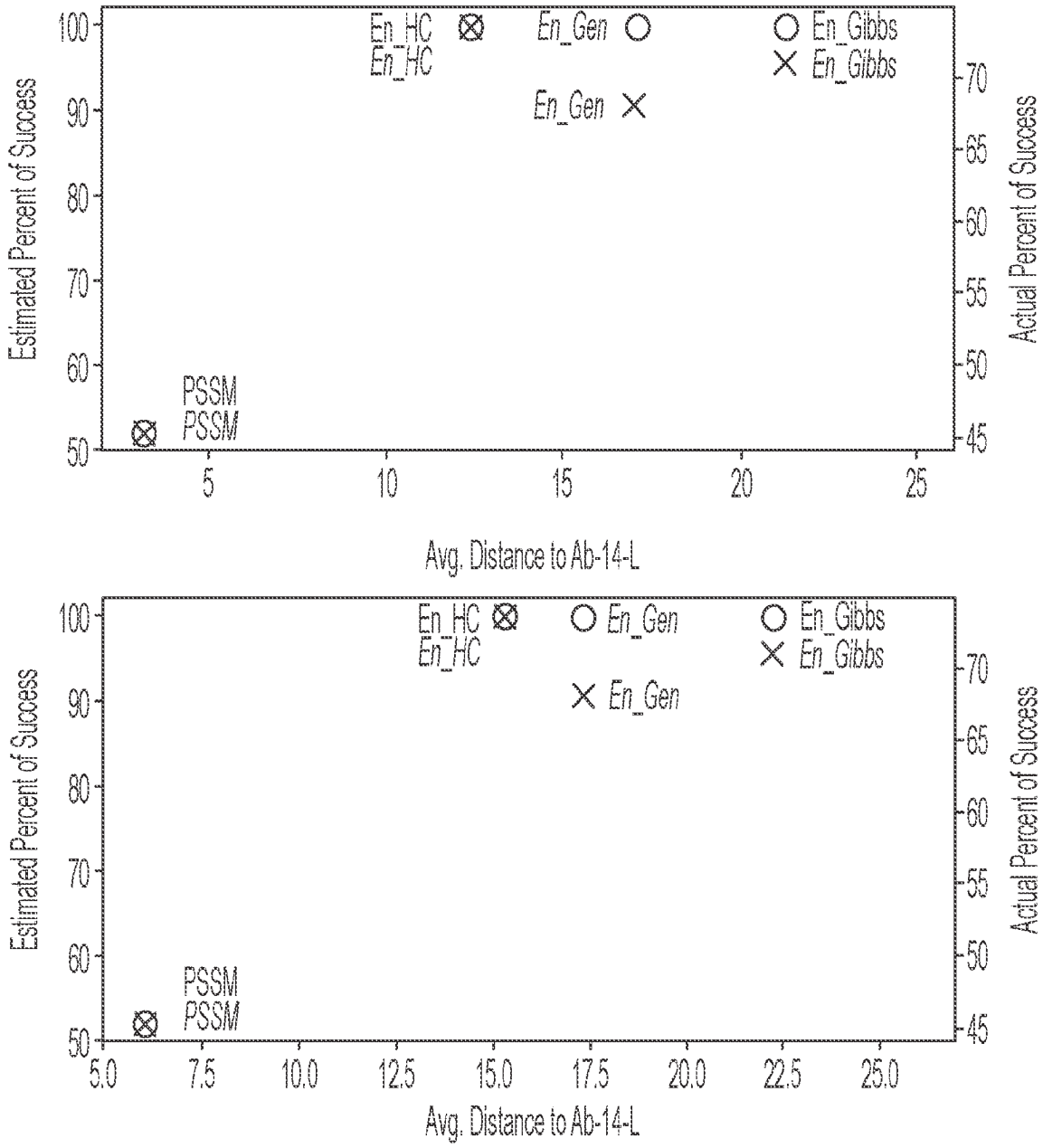

Estimated Percent of Success
Actual Percent of Success
Avg. Distance to Ab-14-L
En_HC
En_HC
En_Gen
En_Gen
En_Gibbs
En_Gibbs
PSSM
PSSM
100
90
80
70
60
50
70
65
60
55
50
45
5
10
15
20
25
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0

FIG. 18A

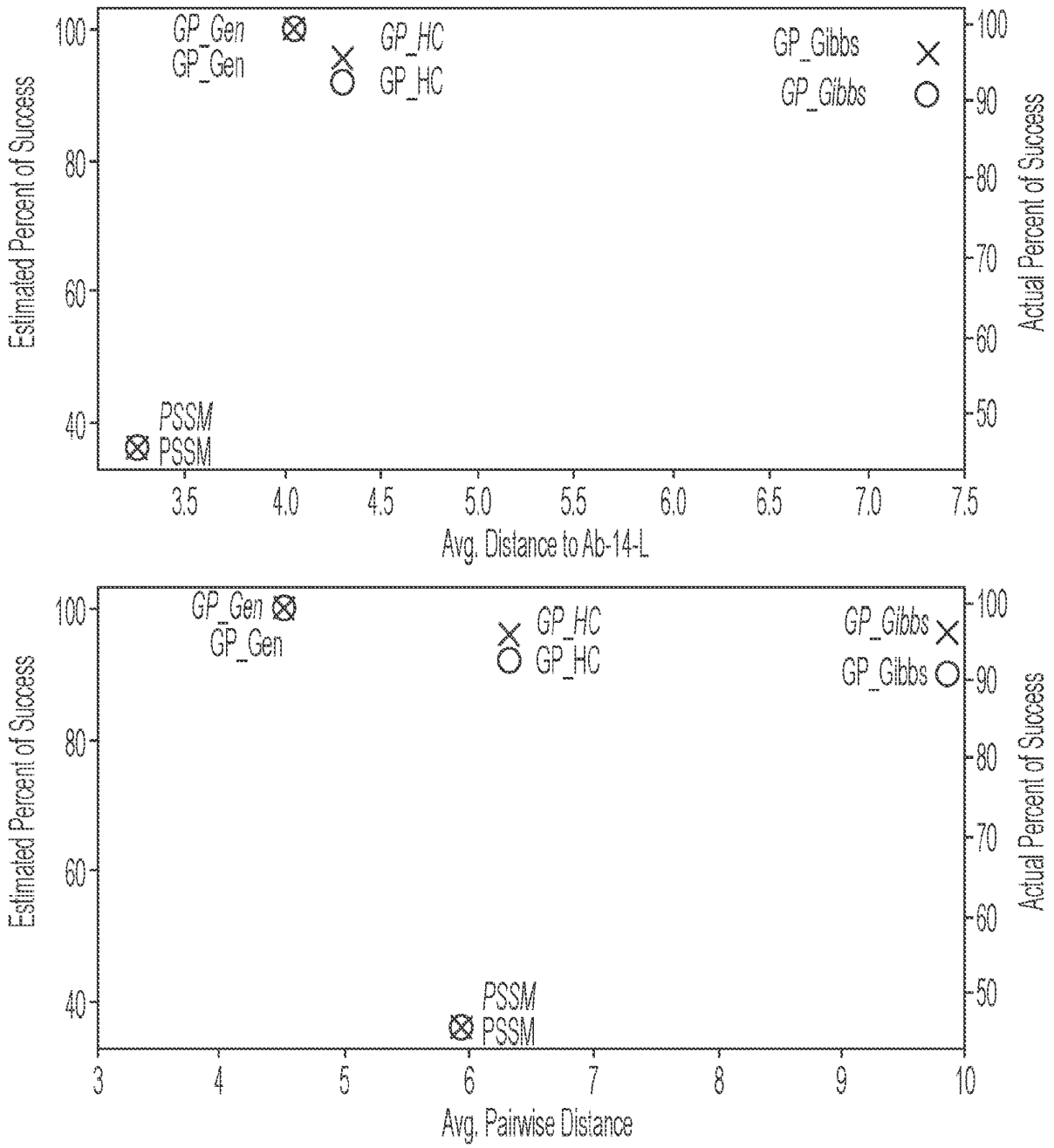

GP_Gen
GP_Gen
GP_HC
GP_HC
GP_Gibbs
GP_Gibbs
PSSM
PSSM
Estimated Percent of Success
Actual Percent of Success
Avg. Distance to Ab-14-L
Avg. Pairwise Distance

FIG. 18B

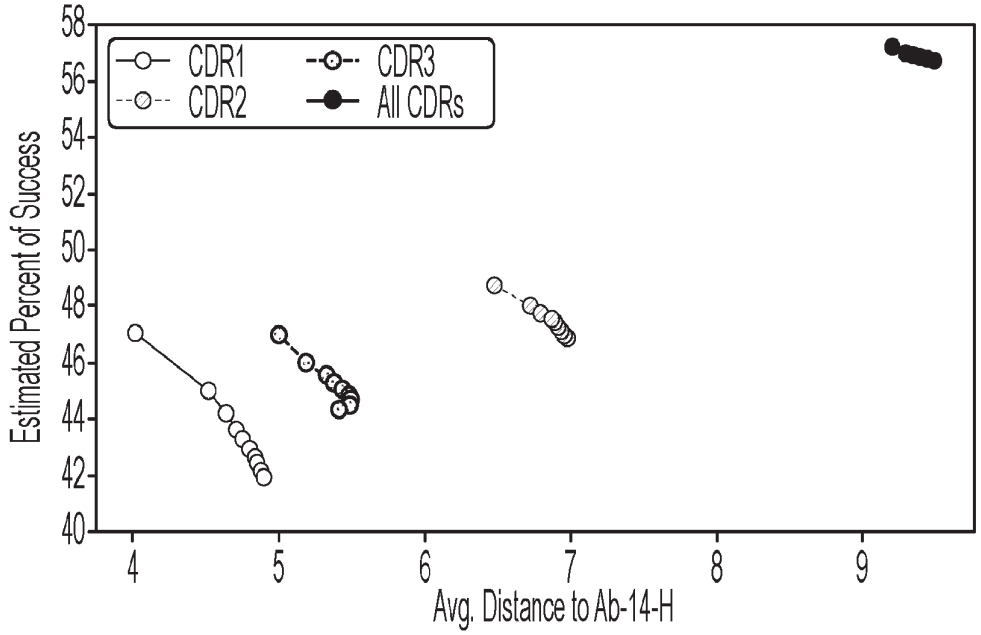


FIG. 19A

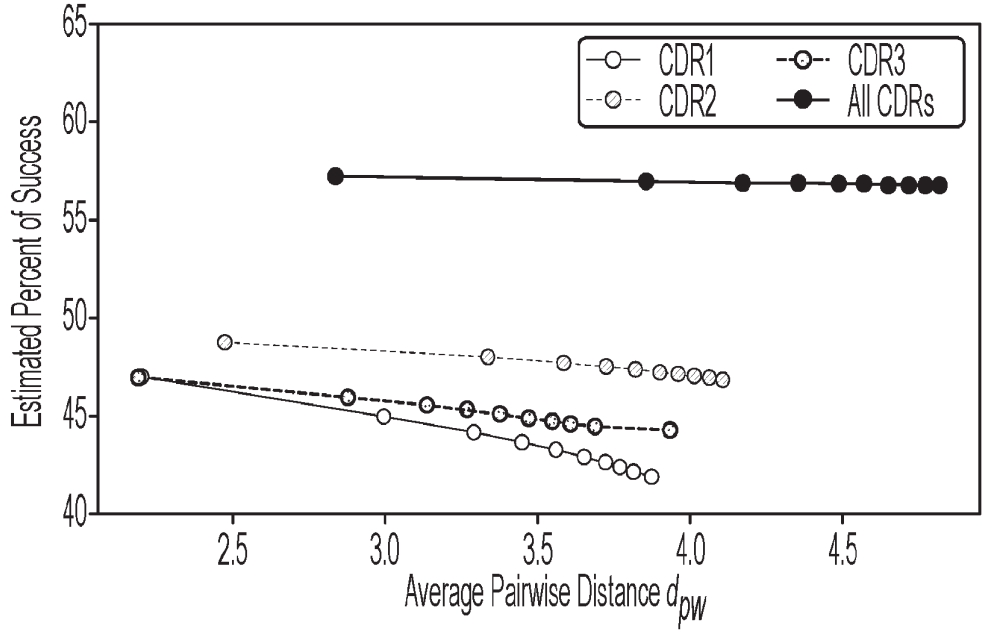


FIG. 19B

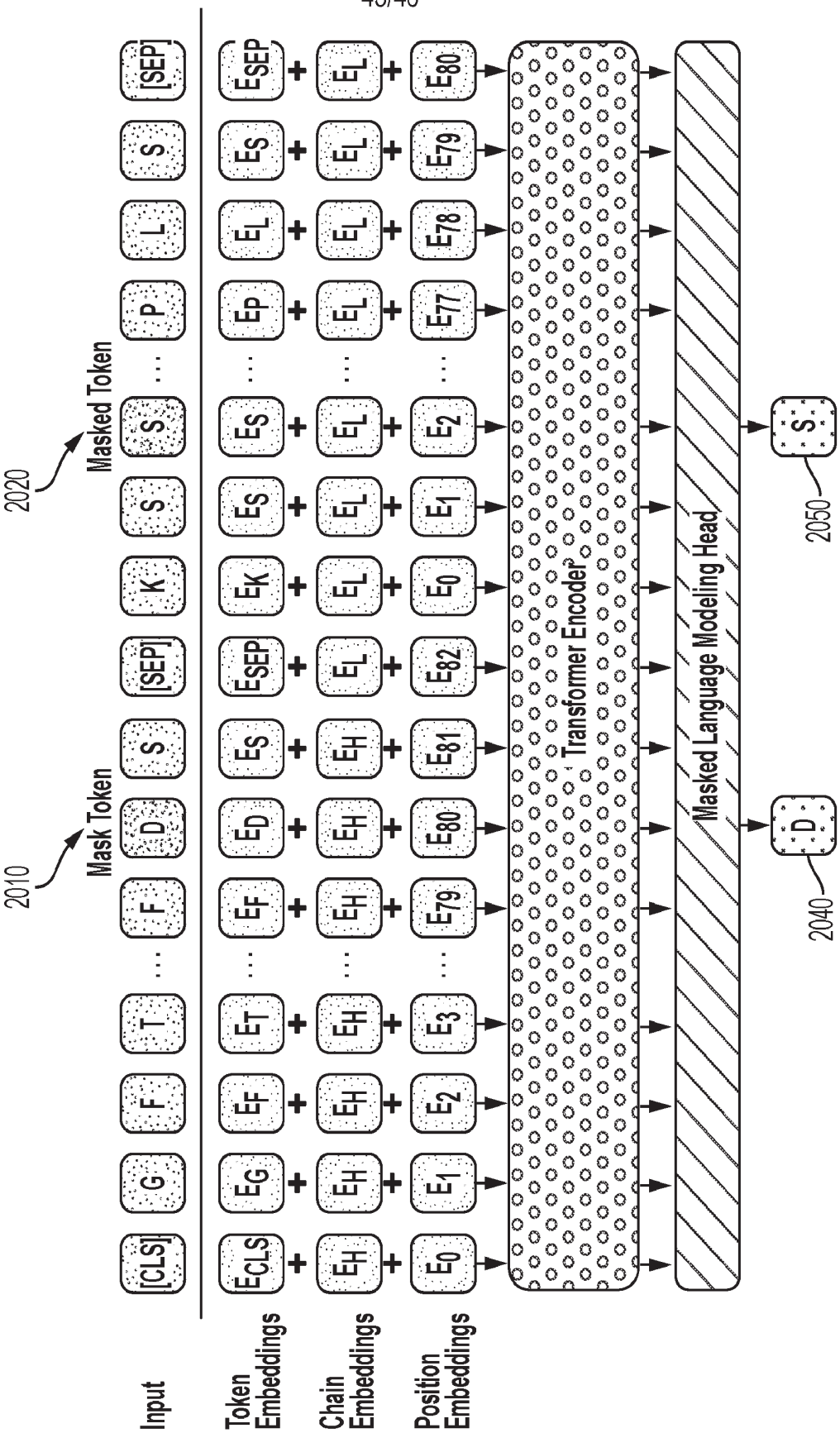

2010
Mask Token
2020
Masked Token
Input
[CLS] G F T ... F D S [SEP] K S S ... P L S [SEP]
Token Embeddings
ECLS EG EF ET ... EF ED ES ESEP EK ES ES ... EP EL ES ESEP
Chain Embeddings
EH EH EH EH ... EH EH EH EL EL EL EL ... EL EL EL EL
Position Embeddings
E0 E1 E2 E3 ... E79 E80 E81 E82 E0 E1 E2 ... E77 E78 E79 E80
Transformer Encoder
Masked Language Modeling Head
2040
D
2050
S

FIG. 20

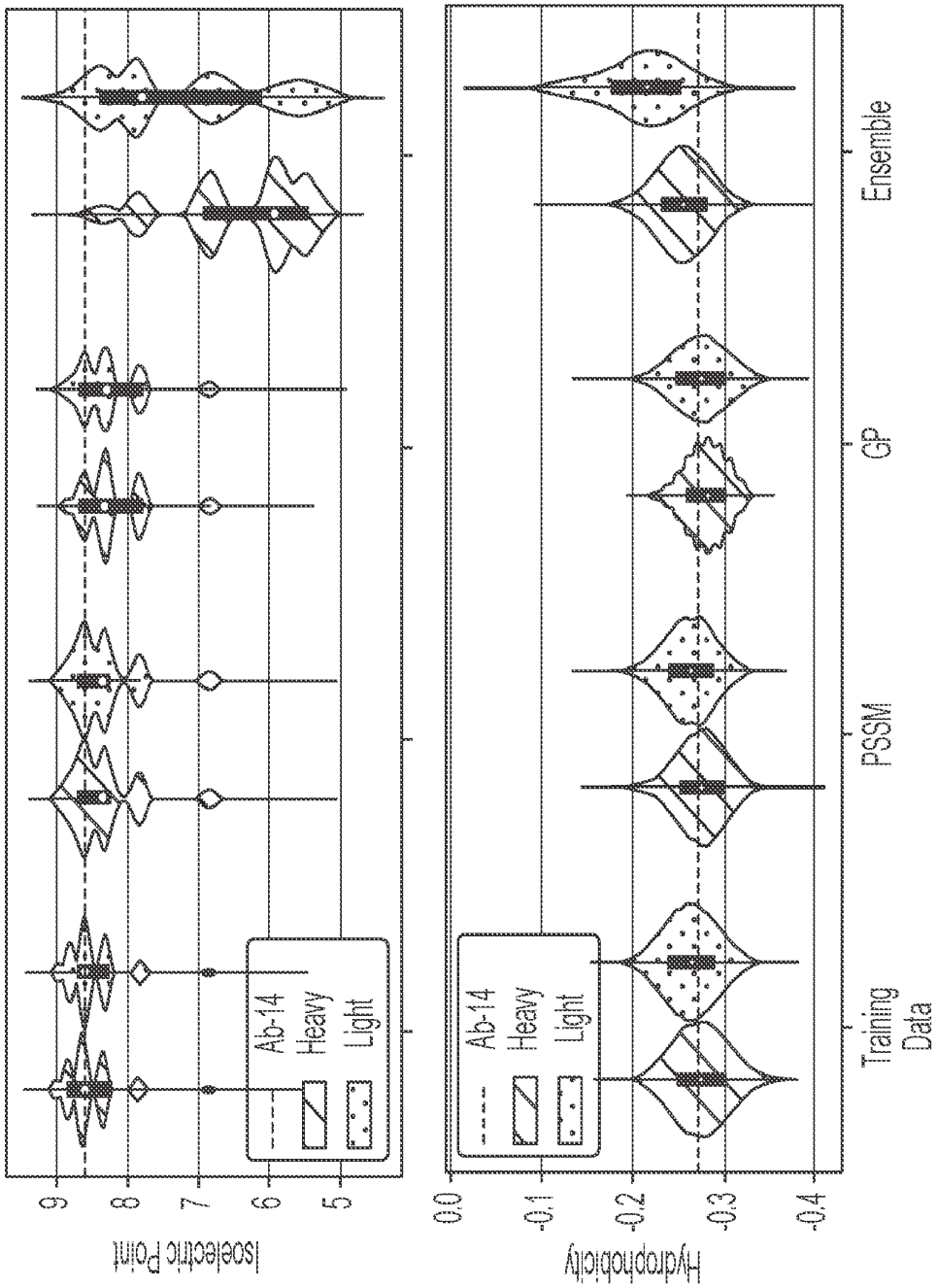

Isoelectric Point
9
8
7
6
5
Ab-14
Heavy
Light
Hydrophobicity
0.0
-0.1
-0.2
-0.3
-0.4
Ab-14
Heavy
Light
Training Data
PSSM
GP
Ensemble

FIG. 21A

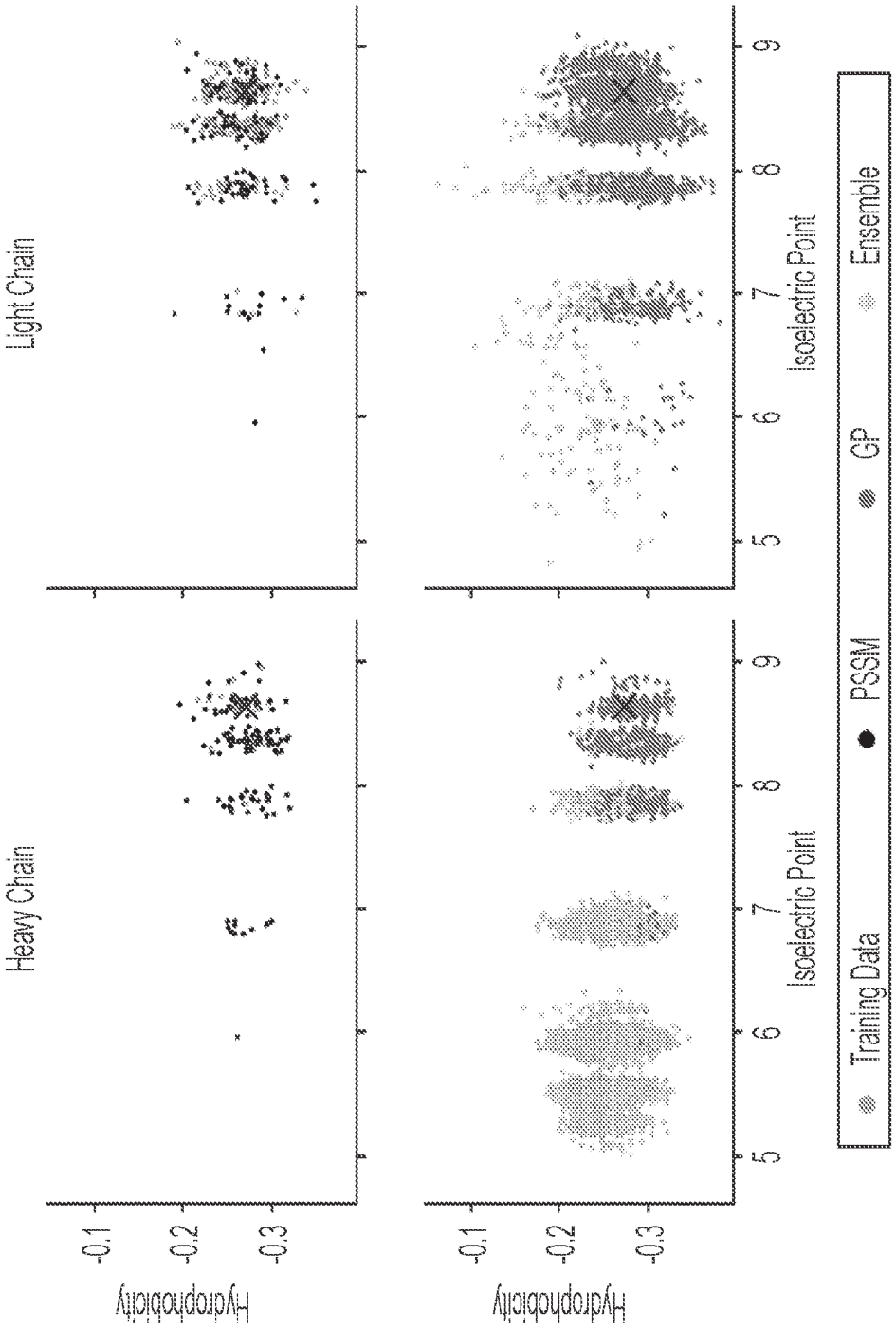

Heavy Chain
Light Chain
Hydrophobicity
-0.1
-0.2
-0.3
5
6
7
8
9
Isoelectric Point
Training Data
PSSM
GP
Ensemble

FIG. 21B

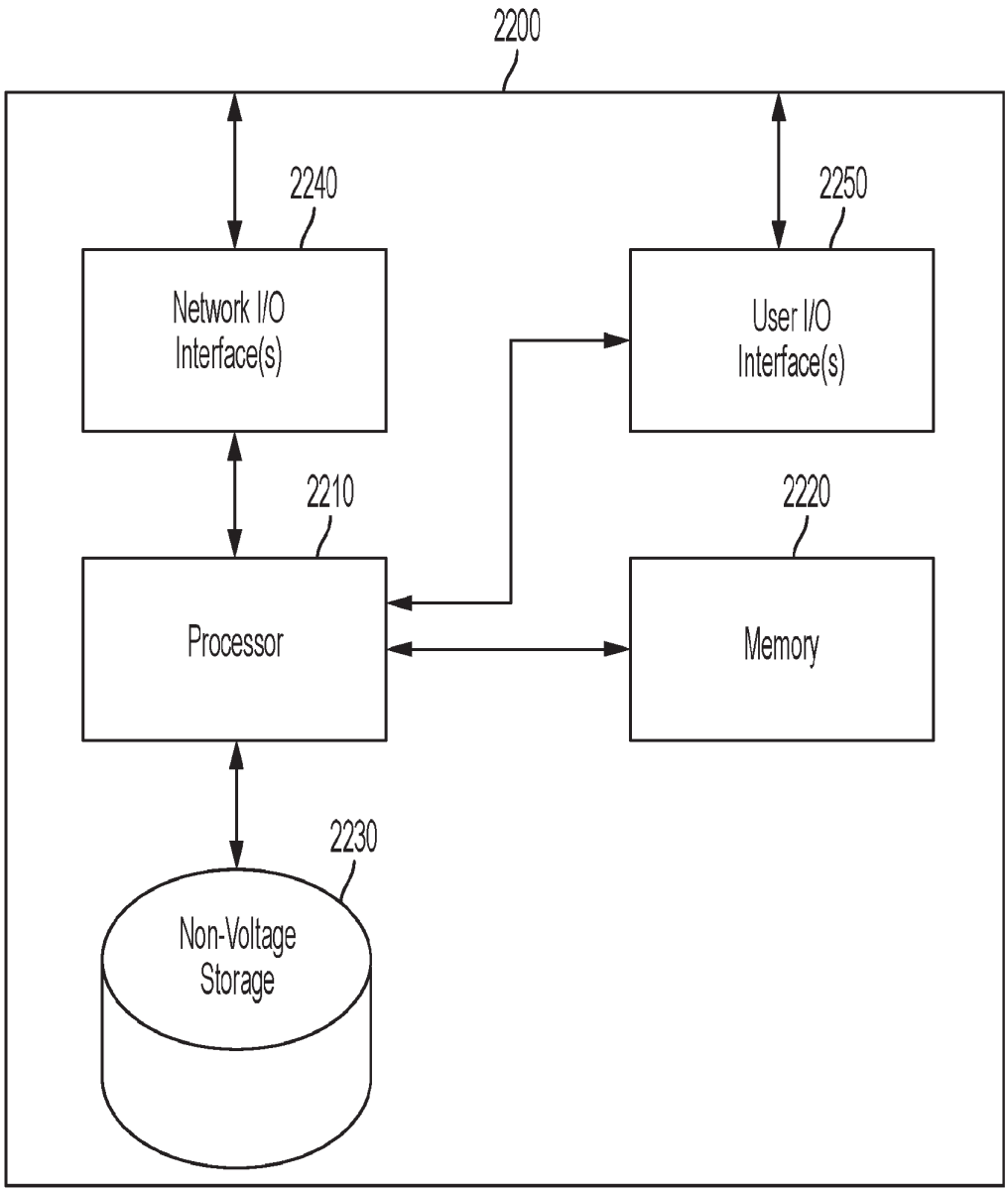

2200
2240
2250
Network I/O Interface(s)
User I/O Interface(s)
2210
2220
Processor
Memory
2230
Non-Voltage Storage

FIG. 22