UNIVERSITY OF WARSAW
MIM
FACULTY
OF MATHEMATICS, INFORMATICS
AND MECHANICS

Architecture of large projects in bioinformatics
(ADP)

Lecture 08

Łukasz P. Kozłowski

Warsaw, 2024

lukaskoz@mimuw.edu.pl

Tables vs. plots

When you present the data use **both**:

Plots

Tables

Tables vs. plots

When you present the data use **both**:

Plots

the best solution, very natural and easy to interpretation

Tables

Tables vs. plots

When you present the data use **both**:

Plots

the best solution, very natural and easy to interpretation
(but also prone for miss-interpretation)

Tables

Tables vs. plots

When you present the data use **both**:

Plots

the best solution, very natural and easy to interpretation
(but also prone for miss-interpretation)

Tables

harder to interpret in short time, but higher information content

Tables vs. plots

When you present the data use **both**:

Plots

the best solution, very natural and easy to interpretation
(but also prone for miss-interpretation)

Tables

harder to interpret in short time, but higher information content

*** Raw data**
for the sake of completeness if you can add them

*** The scripts**
for the sake of reproducibility if you can add them

Tables vs. plots

When you present the data use **both**:

Plots

the best solution, very natural and easy to interpretation
(but also prone for miss-interpretation)

Tables

harder to interpret in short time, but higher information content

*** Raw data**
for the sake of completeness if you can add them

*** The scripts**
for the sake of reproducibility if you can add them

*** The text**

Tables vs. plots

Some proofs:

Some proofs:

93% of human communication is non-verbal

People remember:

80% of what they **see**
and
20% what they **read**

Albert Mahrabian (1971) „Silent Messages”

Some proofs:

93% of human communication is non-verbal

People remember:

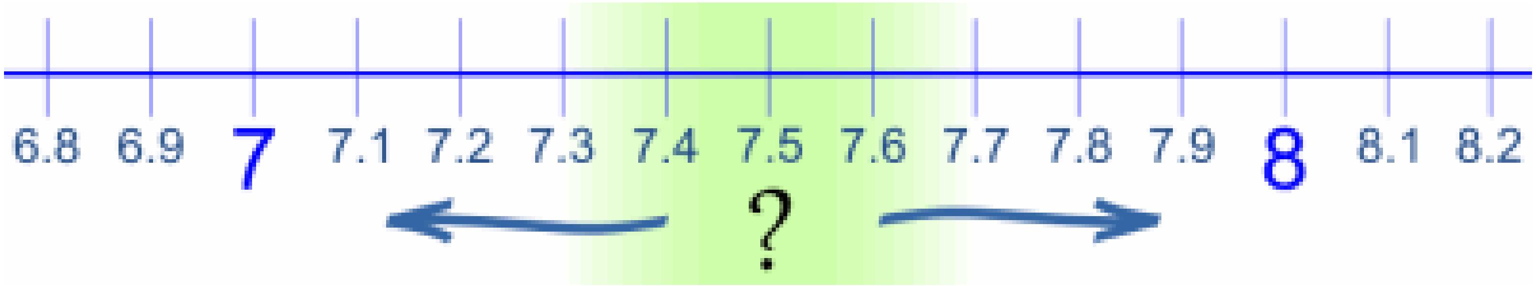
80% of what they **see**
and
20% what they **read**

Albert Mahrabian (1971) „Silent Messages”

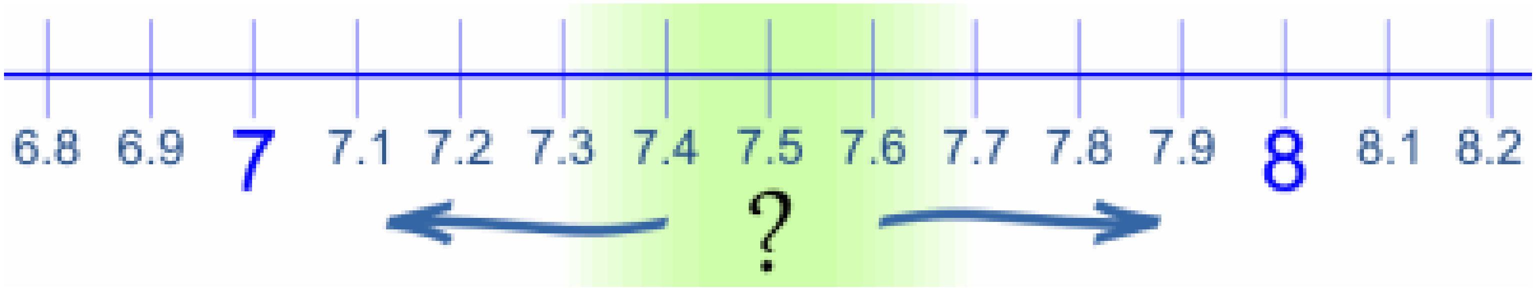
using visuals will make a presentation 43% more persuasive

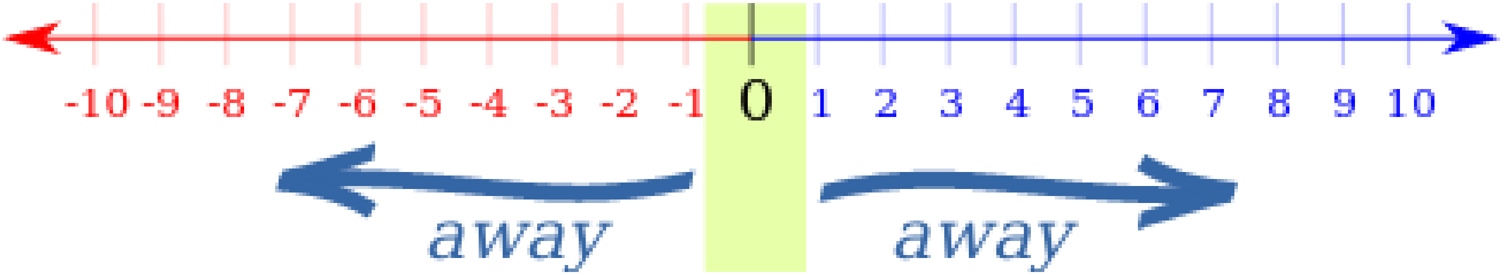
Vogel, D. R., Dickson, G. W., & Lehman, J. A. (1986). Persuasion and the role of visual presentation support: The UM/3M study.

Rounding

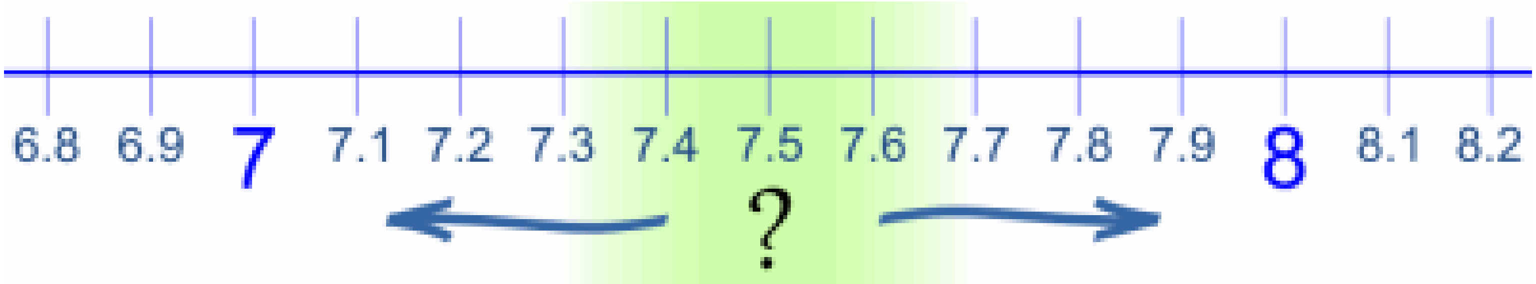


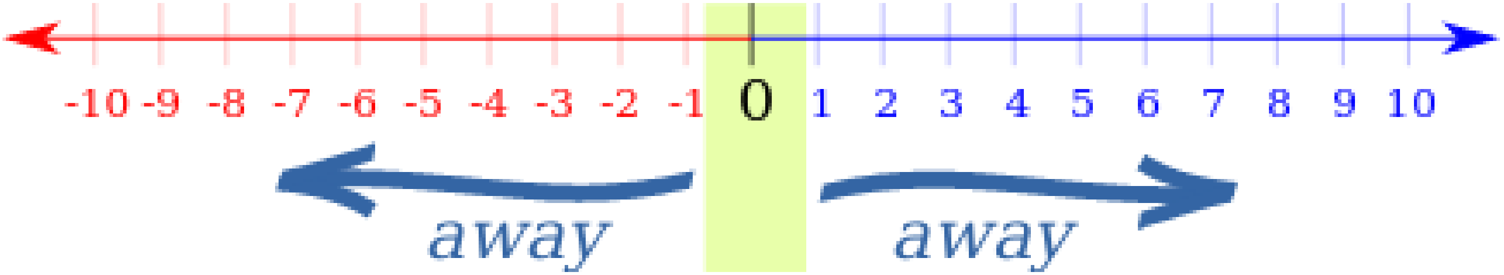
Rounding

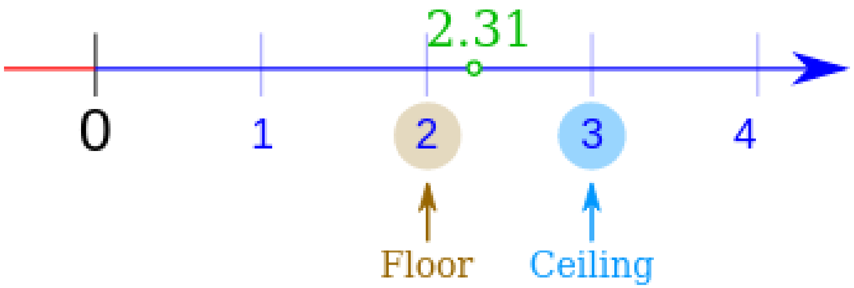

6.8
6.9
7
7.1
7.2
7.3
7.4
7.5
7.6
7.7
7.8
7.9
8
8.1
8.2
?


-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
away
away

Rounding


6.8
6.9
7
7.1
7.2
7.3
7.4
7.5
7.6
7.7
7.8
7.9
8
8.1
8.2
?


-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
away
away


2.31
0
1
2
3
4
Floor
Ceiling

Rounding

1443736684.0   **Population of China in Year 2021**

Rounding

1443736684.0	**Population of China in Year 2021**
1443736684	**Population of China in Year 2021**

Rounding

1443736684.0	**Population of China in Year 2021**
1443736684	**Population of China in Year 2021**
1,443,736,684	**Population of China in Year 2021**

Rounding

1443736684.0	**Population of China in Year 2021**
1443736684	**Population of China in Year 2021**
1,443,736,684	**Population of China in Year 2021**
1 443 736 684	**Population of China in Year 2021**

Rounding

1443736684.0	**Population of China in Year 2021**
1443736684	**Population of China in Year 2021**
1,443,736,684	**Population of China in Year 2021**
1 443 736 684	**Population of China in Year 2021**

Country specific
(comas, dots, spaces, ...)

Rounding

1443736684.0	**Population of China in Year 2021**
1443736684	**Population of China in Year 2021**
1,443,736,684	**Population of China in Year 2021**
1 443 736 684	**Population of China in Year 2021**
1,443 Million	**Population of China in Year 2021**

Rounding

1443736684.0	**Population of China in Year 2021**
1443736684	**Population of China in Year 2021**
1,443,736,684	**Population of China in Year 2021**
1 443 736 684	**Population of China in Year 2021**
1,443 Million	**Population of China in Year 2021**
1,44 Billion	**Population of China in Year 2021**

Rounding

1443736684.0	**Population of China in Year 2021**
1443736684	**Population of China in Year 2021**
1,443,736,684	**Population of China in Year 2021**
1 443 736 684	**Population of China in Year 2021**
1,443 Million	**Population of China in Year 2021**
1,44 Billion	**Population of China in Year 2021**
1,44 Bi	**Population of China in Year 2021**

Rounding

1.578454545454348412211111

Rounding

1.578454545454348412211111

2.344847398437943894794243

2.784353534543

3

1.432328948593543

Rounding

1.578454545454348412211111	**1.58**
2.344847398437943894794243	**2.34**
2.784353534543	**2.78**
3	**3.00**
1.432328948593543	**1.43**

Rounding

1.578454545454348412211111	1.58	1.6
2.344847398437943894794243	2.34	2.3
2.784353534543	2.78	2.8
3	3.00	3.0
1.432328948593543	1.43	1.4

Rounding

1.578454545454348412211111	1.57
1.574847398437943894794243	1.57
1.784353534543	1.78
1.7842328948593543	1.78

Rounding

1.578454545454348412211111	~~1.57~~
1.574847398437943894794243	~~1.57~~
1.784353534543	~~1.78~~
1.7842328948593543	~~1.78~~

Rounding

1.578454545454348412211111

1.574847398437943894794243

1.784353534543

1.7842328948593543

1.578454545454348412211111

1.574847398437943894794243

1.784353534543

1.7842328948593543

Rounding

1.578454545454348412211111

1.574847398437943894794243

1.784353534543

1.7842328948593543

1.578454545454348412211111

1.574847398437943894794243

1.784353534543

1.7842328948593543

Rounding

1.578454545454348412211111	1.5784545454543484122111111
1.574847398437943894794243	1.574847398437943894794243
1.784353534543	1.784353534543
1.7842328948593543	1.7842328948593543

Rounding

1.578454545454348412211111	**1.5784**54545454348412211111
1.574847398437943894794243	**1.5748**47398437943894794243
1.784353534543	**1.7843**53534543
1.7842328948593543	**1.7842**328948593543

Key question: does this make sense to use the given precision?

Rounding

1.5784545454543484122111111 ± 0.2334322324323

1.574847398437943894794243 ± 0.2734322324323

1.784353534543 ± 0.1934322324323

1.7842328948593543 ± 0.4134322324323

Rounding

1.5784 ± 0.2334322324323

1.574847398437943894794243 ± 0.2734322324323

1.784353534543 ± 0.1934322324323

1.7842328948593543 ± 0.4134322324323

1.5784 ± 0.2334322324323

1.5748 ± 0.2734322324323

1.7843 ± 0.1934322324323

1.7842 ± 0.4134322324323

Rounding

1.5784 ± 0.2334322324323

1.574847398437943894794243 ± 0.2734322324323

1.784353534543 ± 0.1934322324323

1.7842328948593543 ± 0.4134322324323

1.5784 ± 0.2334

1.5748 ± 0.2734

1.7843 ± 0.1934

1.7842 ± 0.4134

Rounding

1.5784 ± 0.2334322324323

1.574847398437943894794243 ± 0.2734322324323

1.784353534543 ± 0.1934322324323

1.7842328948593543 ± 0.4134322324323

1.5784 ± 0.2334

1.5748 ± 0.2734

1.7843 ± 0.1934

1.7842 ± 0.4134

1.57 ± 0.23

1.57 ± 0.27

1.78 ± 0.19

1.78 ± 0.41

Rounding

1.5784 ± 0.2334322324323

1.574847398437943894794243 ± 0.2734322324323

1.784353534543 ± 0.1934322324323

1.7842328948593543 ± 0.4134322324323

1.5784 ± 0.2334

1.5748 ± 0.2734

1.7843 ± 0.1934

1.7842 ± 0.4134

1.57 ± 0.23

1.57 ± 0.27

1.78 ± 0.19

1.78 ± 0.41

1.6 ± 0.2

1.6 ± 0.3

1.8 ± 0.2

1.8 ± 0.4

Rounding

1.5784 ± 0.2334322324323
1.574847398437943894794243 ± 0.2734322324323
1.784353534543 ± 0.1934322324323
1.7842328948593543 ± 0.4134322324323

1.5784 ± 0.2334
1.5748 ± 0.2734
1.7843 ± 0.1934
1.7842 ± 0.4134

1.57 ± 0.23
1.57 ± 0.27
1.78 ± 0.19
1.78 ± 0.41

1.6 ± 0.2
1.6 ± 0.3
1.8 ± 0.2
1.8 ± 0.4

Rounding

1.578454545454348412211111	**1.5784**54545454348412211111
1.574847398437943894794243	**1.5748**47398437943894794243
1.784353534543	**1.7843**53534543
1.7842328948593543	**1.7842**328948593543

Key question: does this make sense to use the given precision?

Rounding

1.578454545454348412211111
1.574847398437943894794243
1.784353534543
1.7842328948593543

1.578454545454348412211111
1.574847398437943894794243
1.784353534543
1.7842328948593543

Key question: does this make sense to use the given precision?

Consider example: 1.578454545454348412211111 (that should be presented as 1.5784) is theoretical calculation (e.g. prediction based deep learning model), but … this is only prediction of some natural phenomen that due to the technique we use we can measure with 0.2 precision.

Rounding

1.578454545454348412211111	**1.5784**5454545434841221111 1
1.574847398437943894794243	**1.5748**47398437943894794243
1.784353534543	**1.7843**53534543
1.7842328948593543	**1.7842**328948593543

1.6 ± 0.2

1.6 ± 0.2

1.8 ± 0.2

1.8 ± 0.2

Key question: does this make sense to use the given precision?

Consider example: 1.578454545454348412211111 (that should be presented as 1.5784) is theoretical calculation (e.g. prediction based deep learning model), but … this is only prediction of some natural phenomen that due to the technique we use we can measure with 0.2 precision.

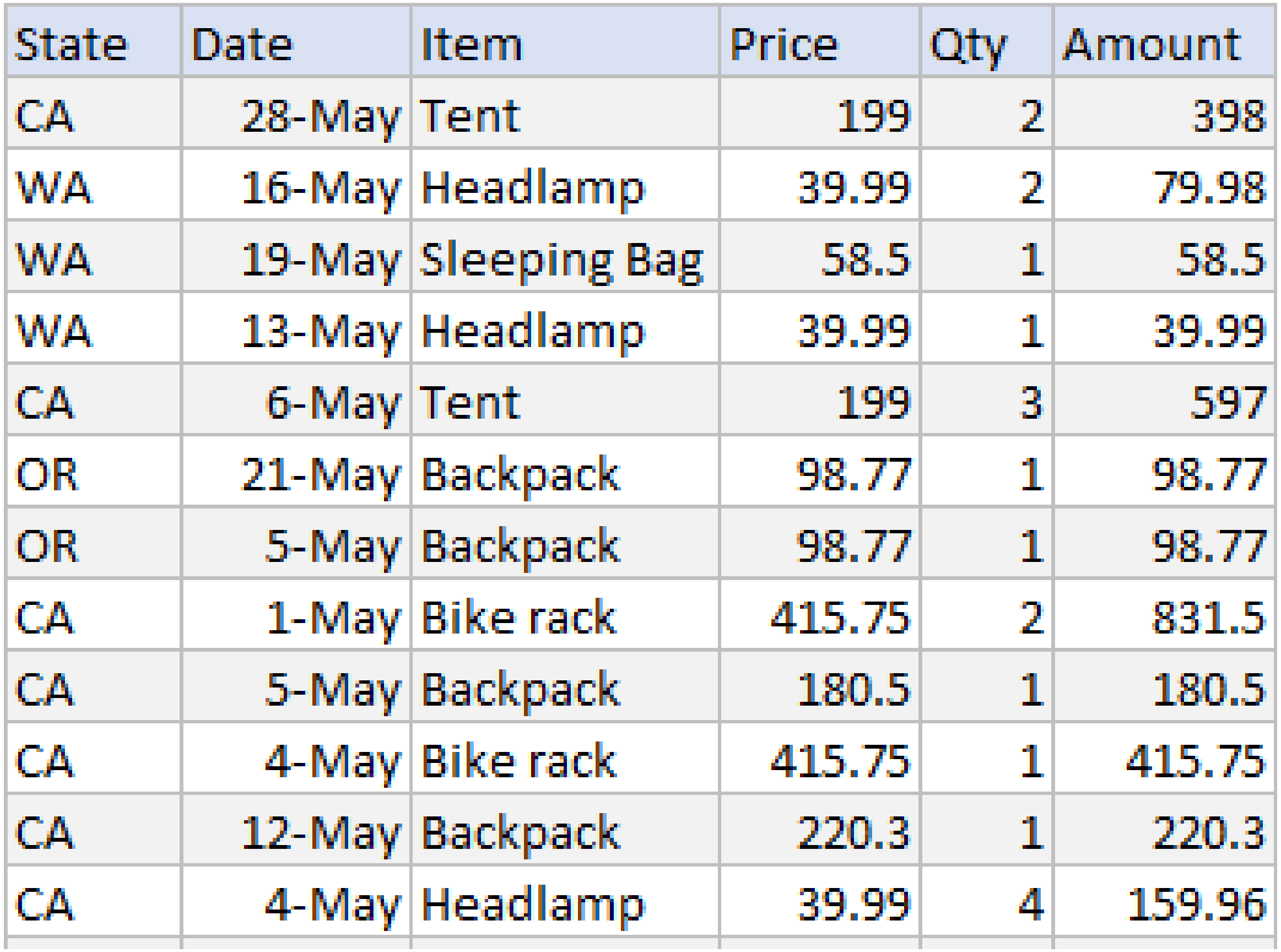
Tables vs. plots

Tables

1) Use clean template

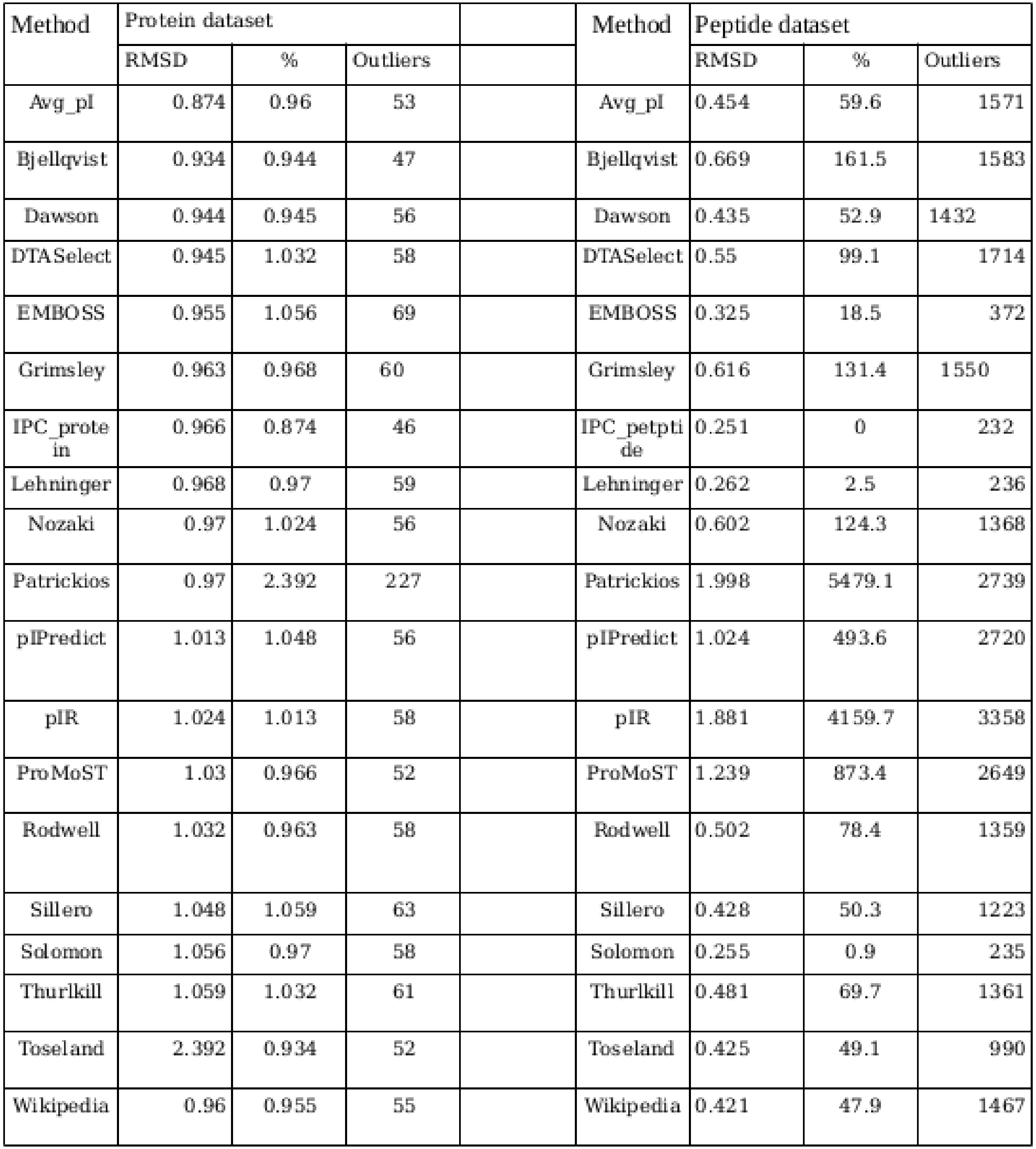
2) Make it interactive if possible (html)

Tables (for print)



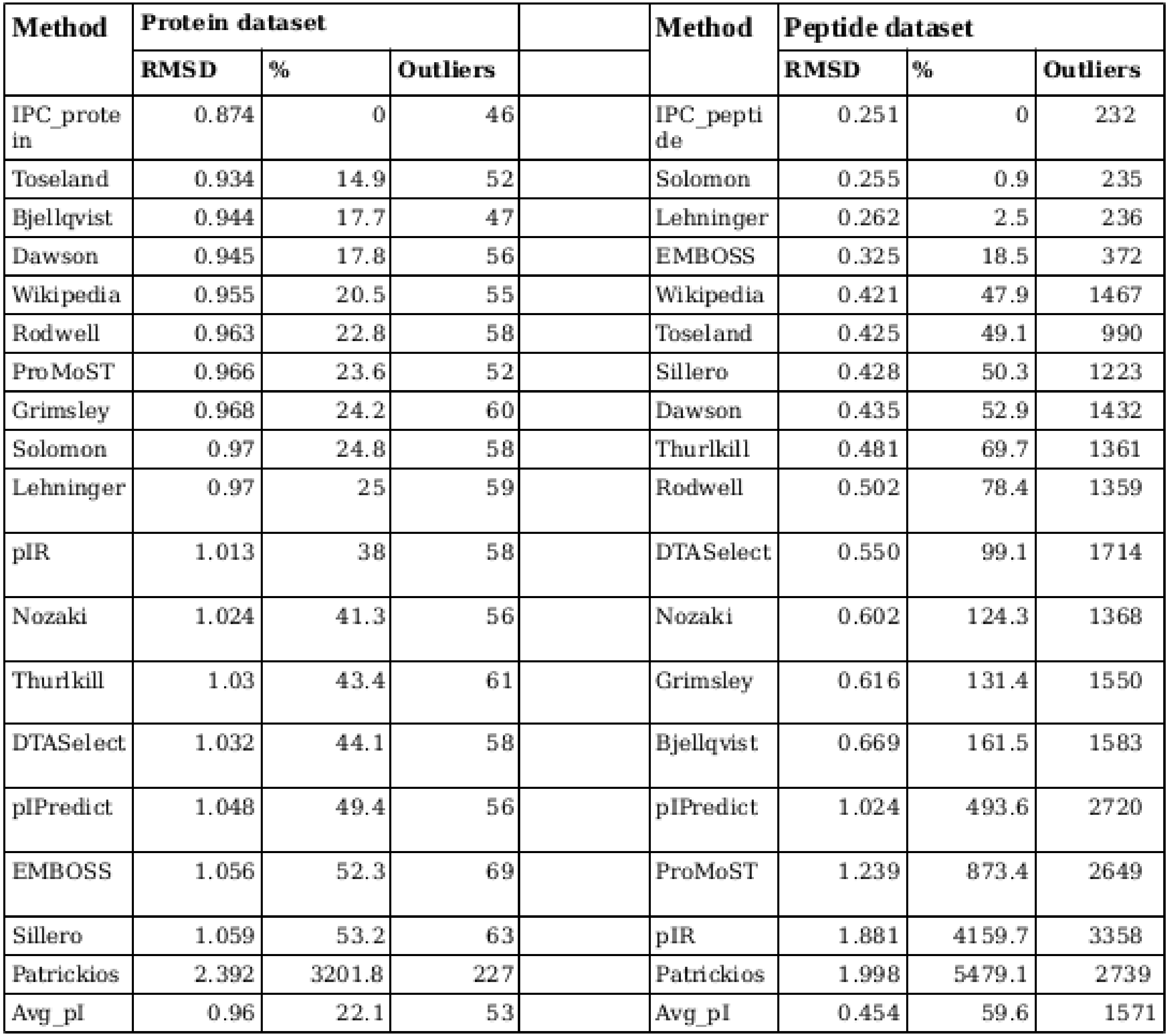
State	Date	Item	Price	Qty	Amount
CA	28-May	Tent	199	2	398
WA	16-May	Headlamp	39.99	2	79.98
WA	19-May	Sleeping Bag	58.5	1	58.5
WA	13-May	Headlamp	39.99	1	39.99
CA	6-May	Tent	199	3	597
OR	21-May	Backpack	98.77	1	98.77
OR	5-May	Backpack	98.77	1	98.77
CA	1-May	Bike rack	415.75	2	831.5
CA	5-May	Backpack	180.5	1	180.5
CA	4-May	Bike rack	415.75	1	415.75
CA	12-May	Backpack	220.3	1	220.3
CA	4-May	Headlamp	39.99	4	159.96

Tables



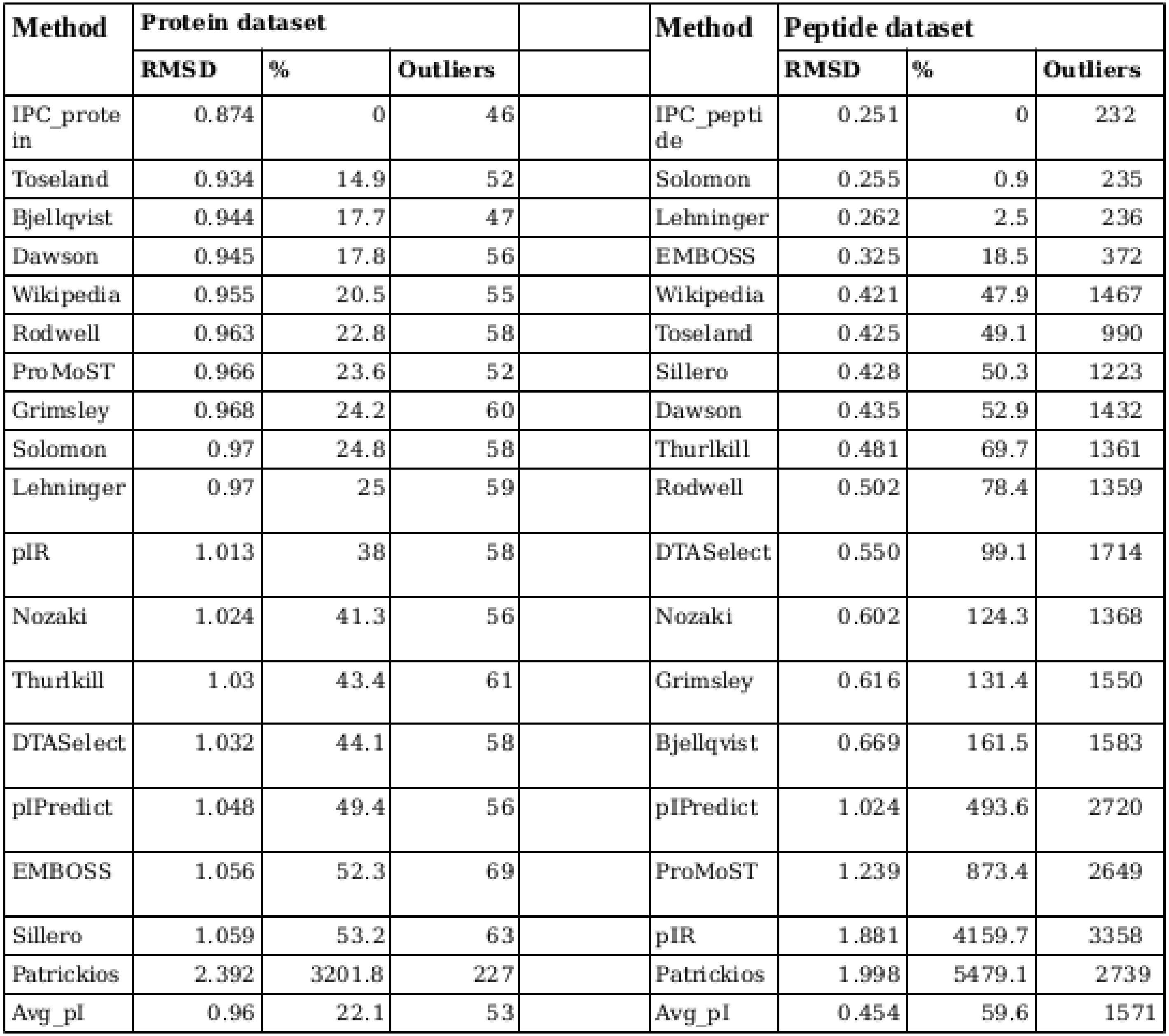
Method	Protein dataset				Method	Peptide dataset		
	RMSD	%	Outliers			RMSD	%	Outliers
Avg_pI	0.874	0.96	53		Avg_pI	0.454	59.6	1571
Bjellqvist	0.934	0.944	47		Bjellqvist	0.669	161.5	1583
Dawson	0.944	0.945	56		Dawson	0.435	52.9	1432
DTASelect	0.945	1.032	58		DTASelect	0.55	99.1	1714
EMBOSS	0.955	1.056	69		EMBOSS	0.325	18.5	372
Grimsley	0.963	0.968	60		Grimsley	0.616	131.4	1550
IPC_protein	0.966	0.874	46		IPC_petptide	0.251	0	232
Lehninger	0.968	0.97	59		Lehninger	0.262	2.5	236
Nozaki	0.97	1.024	56		Nozaki	0.602	124.3	1368
Patrickios	0.97	2.392	227		Patrickios	1.998	5479.1	2739
pIPredict	1.013	1.048	56		pIPredict	1.024	493.6	2720
pIR	1.024	1.013	58		pIR	1.881	4159.7	3358
ProMoST	1.03	0.966	52		ProMoST	1.239	873.4	2649
Rodwell	1.032	0.963	58		Rodwell	0.502	78.4	1359
Sillero	1.048	1.059	63		Sillero	0.428	50.3	1223
Solomon	1.056	0.97	58		Solomon	0.255	0.9	235
Thurlkill	1.059	1.032	61		Thurlkill	0.481	69.7	1361
Toseland	2.392	0.934	52		Toseland	0.425	49.1	990
Wikipedia	0.96	0.955	55		Wikipedia	0.421	47.9	1467

Tables



Method	Protein dataset				Method	Peptide dataset		
	RMSD	%	Outliers			RMSD	%	Outliers
IPC_protein	0.874	0	46		IPC_peptide	0.251	0	232
Toseland	0.934	14.9	52		Solomon	0.255	0.9	235
Bjellqvist	0.944	17.7	47		Lehninger	0.262	2.5	236
Dawson	0.945	17.8	56		EMBOSS	0.325	18.5	372
Wikipedia	0.955	20.5	55		Wikipedia	0.421	47.9	1467
Rodwell	0.963	22.8	58		Toseland	0.425	49.1	990
ProMoST	0.966	23.6	52		Sillero	0.428	50.3	1223
Grimsley	0.968	24.2	60		Dawson	0.435	52.9	1432
Solomon	0.97	24.8	58		Thurlkill	0.481	69.7	1361
Lehninger	0.97	25	59		Rodwell	0.502	78.4	1359
pIR	1.013	38	58		DTASelect	0.550	99.1	1714
Nozaki	1.024	41.3	56		Nozaki	0.602	124.3	1368
Thurlkill	1.03	43.4	61		Grimsley	0.616	131.4	1550
DTASelect	1.032	44.1	58		Bjellqvist	0.669	161.5	1583
pIPredict	1.048	49.4	56		pIPredict	1.024	493.6	2720
EMBOSS	1.056	52.3	69		ProMoST	1.239	873.4	2649
Sillero	1.059	53.2	63		pIR	1.881	4159.7	3358
Patrickios	2.392	3201.8	227		Patrickios	1.998	5479.1	2739
Avg_pI	0.96	22.1	53		Avg_pI	0.454	59.6	1571

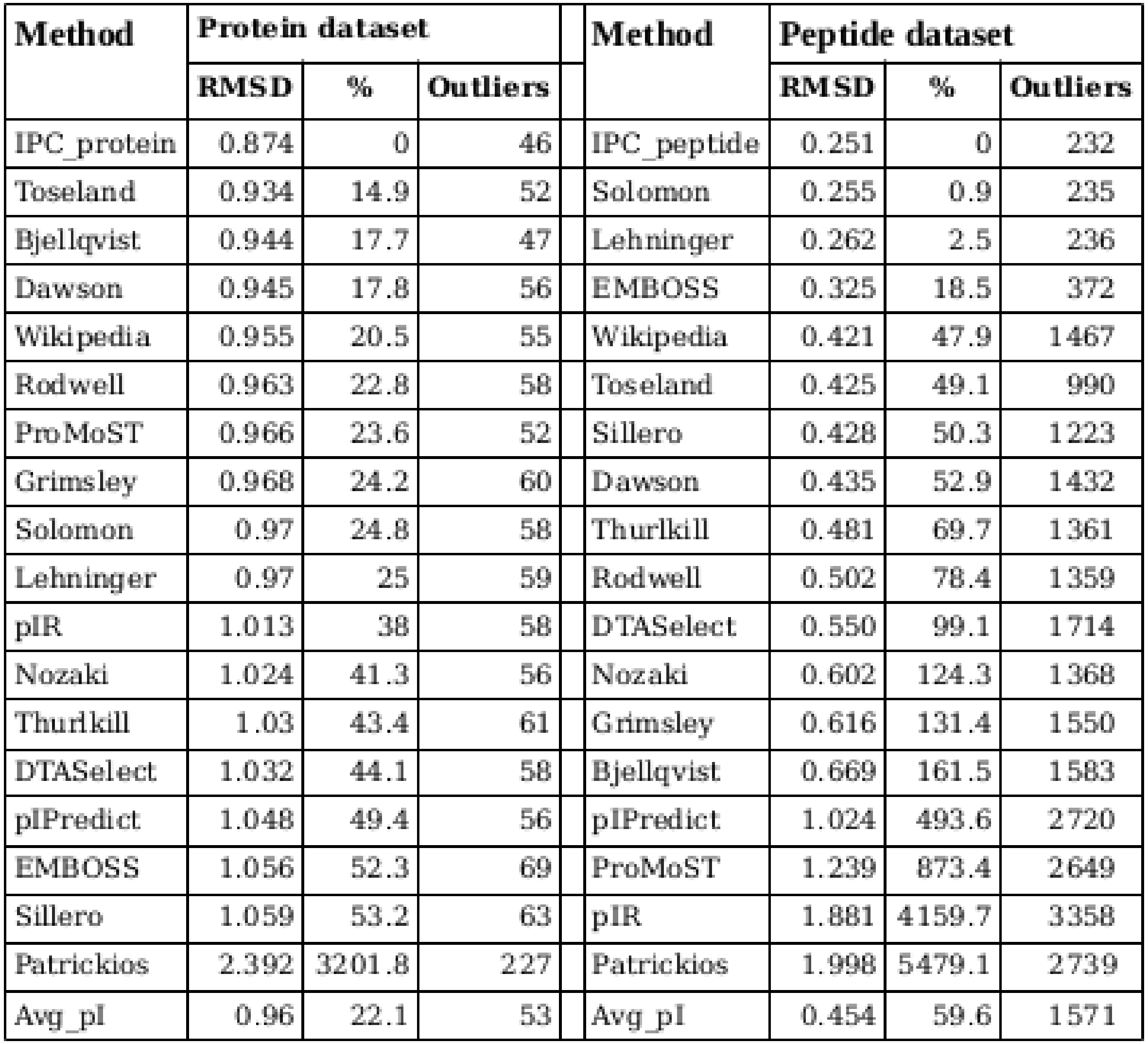
Tables



Method	Protein dataset				Method	Peptide dataset		
	RMSD	%	Outliers			RMSD	%	Outliers
IPC_protein	0.874	0	46		IPC_peptide	0.251	0	232
Toseland	0.934	14.9	52		Solomon	0.255	0.9	235
Bjellqvist	0.944	17.7	47		Lehninger	0.262	2.5	236
Dawson	0.945	17.8	56		EMBOSS	0.325	18.5	372
Wikipedia	0.955	20.5	55		Wikipedia	0.421	47.9	1467
Rodwell	0.963	22.8	58		Toseland	0.425	49.1	990
ProMoST	0.966	23.6	52		Sillero	0.428	50.3	1223
Grimsley	0.968	24.2	60		Dawson	0.435	52.9	1432
Solomon	0.97	24.8	58		Thurlkill	0.481	69.7	1361
Lehninger	0.97	25	59		Rodwell	0.502	78.4	1359
pIR	1.013	38	58		DTASelect	0.550	99.1	1714
Nozaki	1.024	41.3	56		Nozaki	0.602	124.3	1368
Thurlkill	1.03	43.4	61		Grimsley	0.616	131.4	1550
DTASelect	1.032	44.1	58		Bjellqvist	0.669	161.5	1583
pIPredict	1.048	49.4	56		pIPredict	1.024	493.6	2720
EMBOSS	1.056	52.3	69		ProMoST	1.239	873.4	2649
Sillero	1.059	53.2	63		pIR	1.881	4159.7	3358
Patrickios	2.392	3201.8	227		Patrickios	1.998	5479.1	2739
Avg_pI	0.96	22.1	53		Avg_pI	0.454	59.6	1571

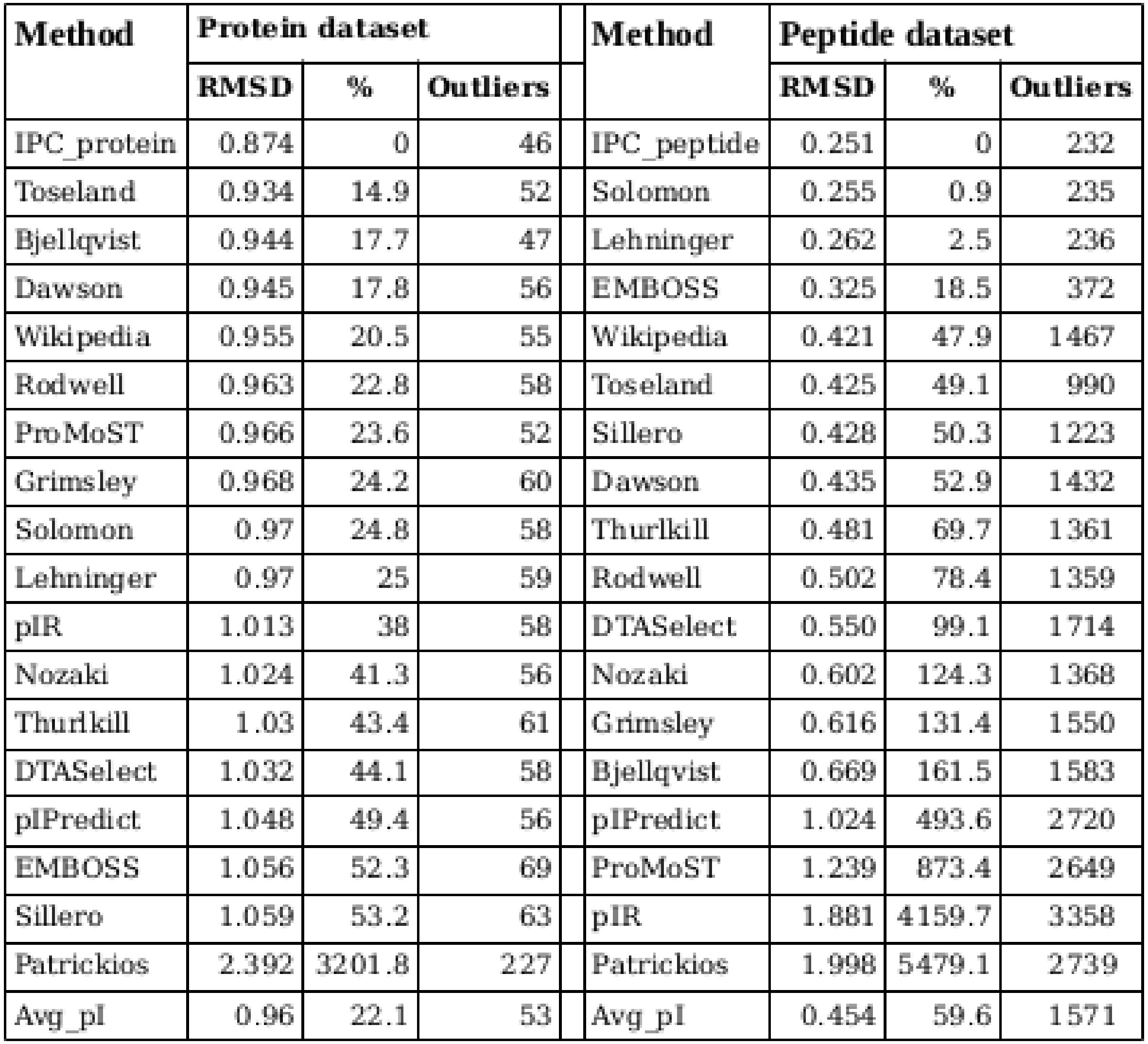
Sort (decide how, use html if possible)

Tables



Method	Protein dataset			Method	Peptide dataset		
	RMSD	%	Outliers		RMSD	%	Outliers
IPC_protein	0.874	0	46	IPC_peptide	0.251	0	232
Toseland	0.934	14.9	52	Solomon	0.255	0.9	235
Bjellqvist	0.944	17.7	47	Lehninger	0.262	2.5	236
Dawson	0.945	17.8	56	EMBOSS	0.325	18.5	372
Wikipedia	0.955	20.5	55	Wikipedia	0.421	47.9	1467
Rodwell	0.963	22.8	58	Toseland	0.425	49.1	990
ProMoST	0.966	23.6	52	Sillero	0.428	50.3	1223
Grimsley	0.968	24.2	60	Dawson	0.435	52.9	1432
Solomon	0.97	24.8	58	Thurlkill	0.481	69.7	1361
Lehninger	0.97	25	59	Rodwell	0.502	78.4	1359
pIR	1.013	38	58	DTASelect	0.550	99.1	1714
Nozaki	1.024	41.3	56	Nozaki	0.602	124.3	1368
Thurlkill	1.03	43.4	61	Grimsley	0.616	131.4	1550
DTASelect	1.032	44.1	58	Bjellqvist	0.669	161.5	1583
pIPredict	1.048	49.4	56	pIPredict	1.024	493.6	2720
EMBOSS	1.056	52.3	69	ProMoST	1.239	873.4	2649
Sillero	1.059	53.2	63	pIR	1.881	4159.7	3358
Patrickios	2.392	3201.8	227	Patrickios	1.998	5479.1	2739
Avg_pI	0.96	22.1	53	Avg_pI	0.454	59.6	1571

Tables

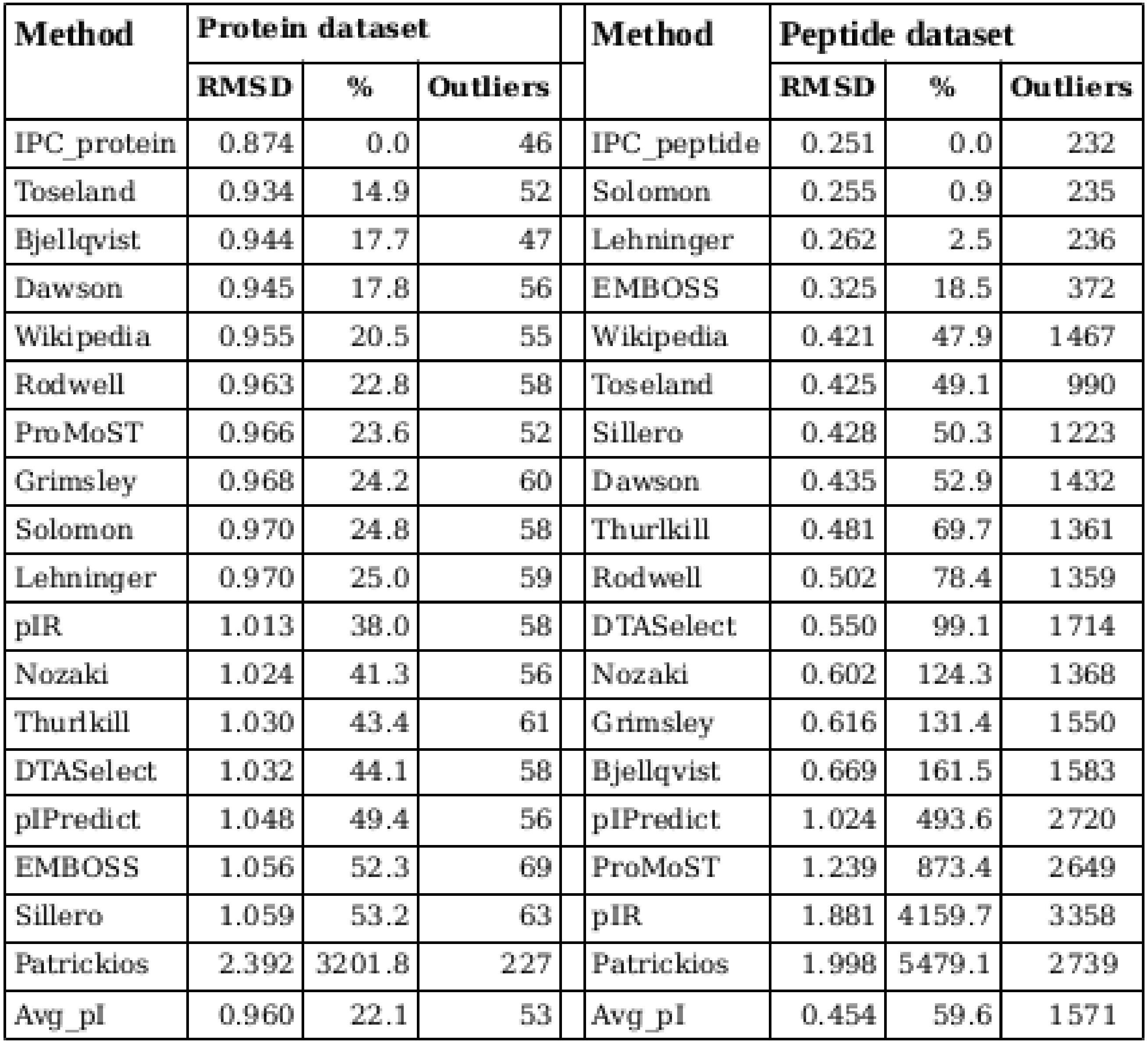


Method	Protein dataset				Method	Peptide dataset		
	RMSD	%	Outliers			RMSD	%	Outliers
IPC_protein	0.874	0	46		IPC_peptide	0.251	0	232
Toseland	0.934	14.9	52		Solomon	0.255	0.9	235
Bjellqvist	0.944	17.7	47		Lehninger	0.262	2.5	236
Dawson	0.945	17.8	56		EMBOSS	0.325	18.5	372
Wikipedia	0.955	20.5	55		Wikipedia	0.421	47.9	1467
Rodwell	0.963	22.8	58		Toseland	0.425	49.1	990
ProMoST	0.966	23.6	52		Sillero	0.428	50.3	1223
Grimsley	0.968	24.2	60		Dawson	0.435	52.9	1432
Solomon	0.97	24.8	58		Thurlkill	0.481	69.7	1361
Lehninger	0.97	25	59		Rodwell	0.502	78.4	1359
pIR	1.013	38	58		DTASelect	0.550	99.1	1714
Nozaki	1.024	41.3	56		Nozaki	0.602	124.3	1368
Thurlkill	1.03	43.4	61		Grimsley	0.616	131.4	1550
DTASelect	1.032	44.1	58		Bjellqvist	0.669	161.5	1583
pIPredict	1.048	49.4	56		pIPredict	1.024	493.6	2720
EMBOSS	1.056	52.3	69		ProMoST	1.239	873.4	2649
Sillero	1.059	53.2	63		pIR	1.881	4159.7	3358
Patrickios	2.392	3201.8	227		Patrickios	1.998	5479.1	2739
Avg_pI	0.96	22.1	53		Avg_pI	0.454	59.6	1571

Bold & Align, the same font (size, type, consider using monotype font for better alignment)

Optimal width of columns and vertical and horizontal alignment, avoid blank spaces

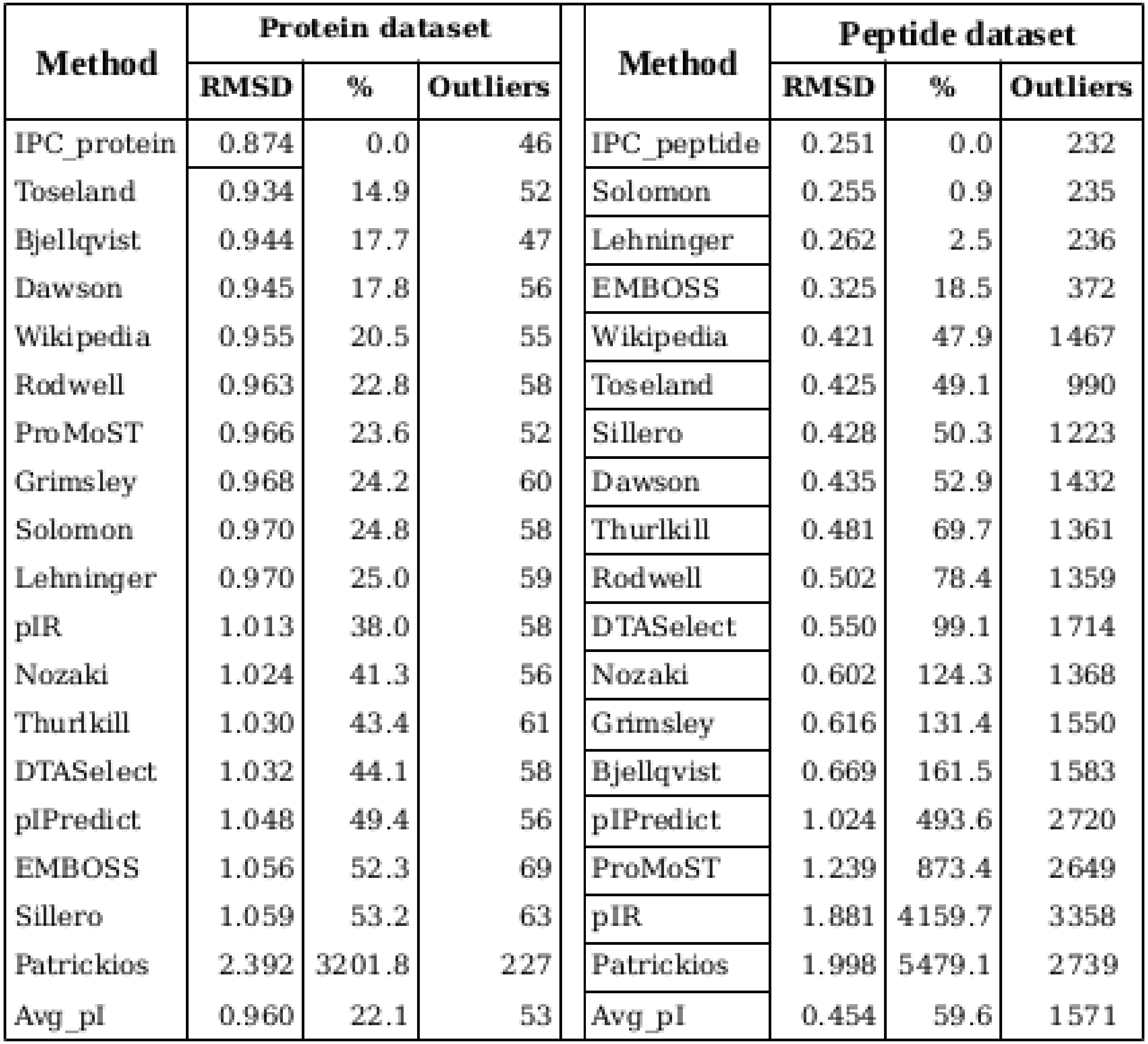
Tables



Method	Protein dataset			Method	Peptide dataset		
	RMSD	%	Outliers		RMSD	%	Outliers
IPC_protein	0.874	0.0	46	IPC_peptide	0.251	0.0	232
Toseland	0.934	14.9	52	Solomon	0.255	0.9	235
Bjellqvist	0.944	17.7	47	Lehninger	0.262	2.5	236
Dawson	0.945	17.8	56	EMBOSS	0.325	18.5	372
Wikipedia	0.955	20.5	55	Wikipedia	0.421	47.9	1467
Rodwell	0.963	22.8	58	Toseland	0.425	49.1	990
ProMoST	0.966	23.6	52	Sillero	0.428	50.3	1223
Grimsley	0.968	24.2	60	Dawson	0.435	52.9	1432
Solomon	0.970	24.8	58	Thurlkill	0.481	69.7	1361
Lehninger	0.970	25.0	59	Rodwell	0.502	78.4	1359
pIR	1.013	38.0	58	DTASelect	0.550	99.1	1714
Nozaki	1.024	41.3	56	Nozaki	0.602	124.3	1368
Thurlkill	1.030	43.4	61	Grimsley	0.616	131.4	1550
DTASelect	1.032	44.1	58	Bjellqvist	0.669	161.5	1583
pIPredict	1.048	49.4	56	pIPredict	1.024	493.6	2720
EMBOSS	1.056	52.3	69	ProMoST	1.239	873.4	2649
Sillero	1.059	53.2	63	pIR	1.881	4159.7	3358
Patrickios	2.392	3201.8	227	Patrickios	1.998	5479.1	2739
Avg_pI	0.960	22.1	53	Avg_pI	0.454	59.6	1571

Use the same decimal point (do not round at different levels)

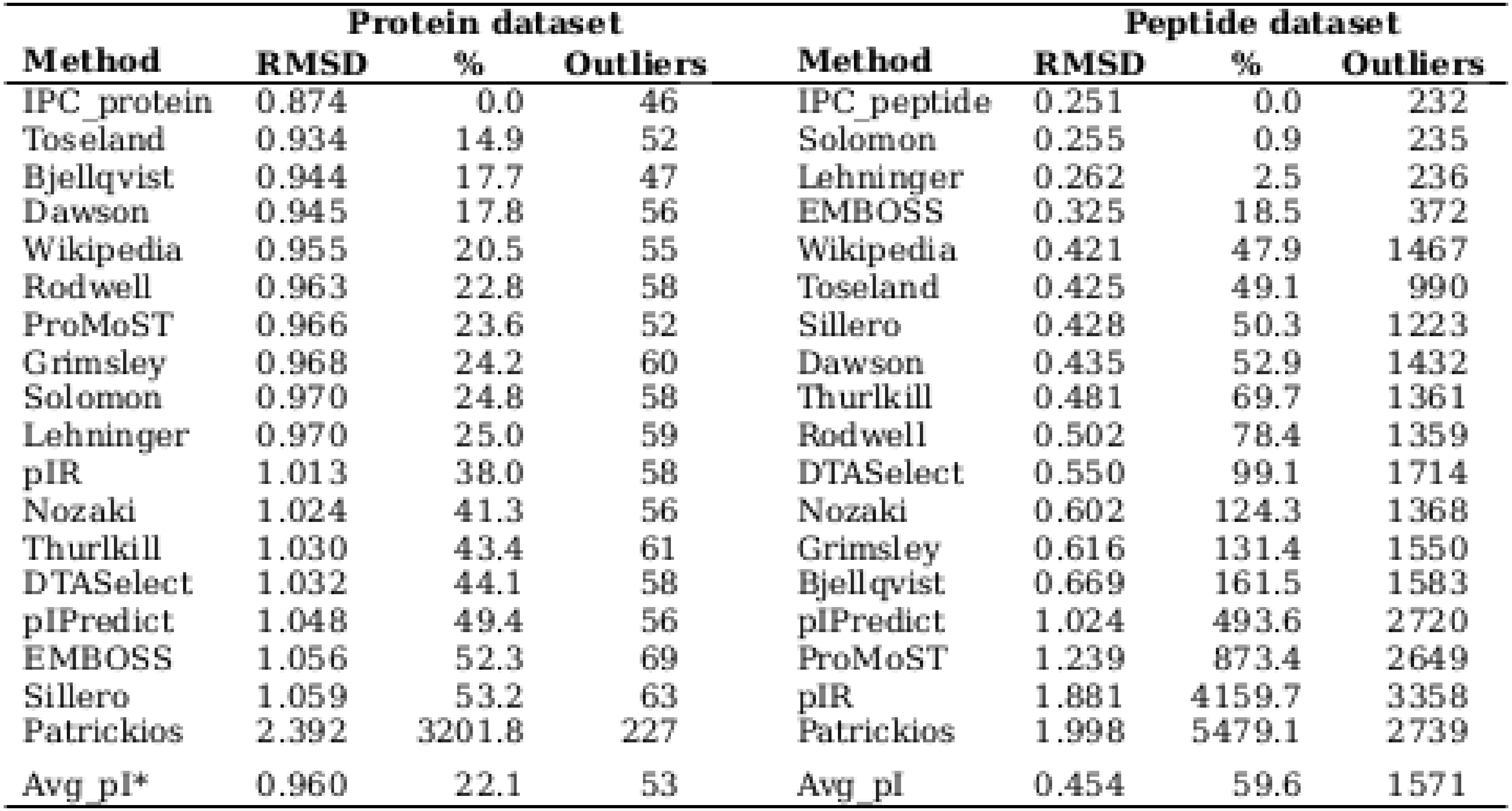
Tables



Method	Protein dataset			Method	Peptide dataset		
	RMSD	%	Outliers		RMSD	%	Outliers
IPC_protein	0.874	0.0	46	IPC_peptide	0.251	0.0	232
Toseland	0.934	14.9	52	Solomon	0.255	0.9	235
Bjellqvist	0.944	17.7	47	Lehninger	0.262	2.5	236
Dawson	0.945	17.8	56	EMBOSS	0.325	18.5	372
Wikipedia	0.955	20.5	55	Wikipedia	0.421	47.9	1467
Rodwell	0.963	22.8	58	Toseland	0.425	49.1	990
ProMoST	0.966	23.6	52	Sillero	0.428	50.3	1223
Grimsley	0.968	24.2	60	Dawson	0.435	52.9	1432
Solomon	0.970	24.8	58	Thurlkill	0.481	69.7	1361
Lehninger	0.970	25.0	59	Rodwell	0.502	78.4	1359
pIR	1.013	38.0	58	DTASelect	0.550	99.1	1714
Nozaki	1.024	41.3	56	Nozaki	0.602	124.3	1368
Thurlkill	1.030	43.4	61	Grimsley	0.616	131.4	1550
DTASelect	1.032	44.1	58	Bjellqvist	0.669	161.5	1583
pIPredict	1.048	49.4	56	pIPredict	1.024	493.6	2720
EMBOSS	1.056	52.3	69	ProMoST	1.239	873.4	2649
Sillero	1.059	53.2	63	pIR	1.881	4159.7	3358
Patrickios	2.392	3201.8	227	Patrickios	1.998	5479.1	2739
Avg_pI	0.960	22.1	53	Avg_pI	0.454	59.6	1571

Less is more (hide some of the borders and make them ticker)

Tables

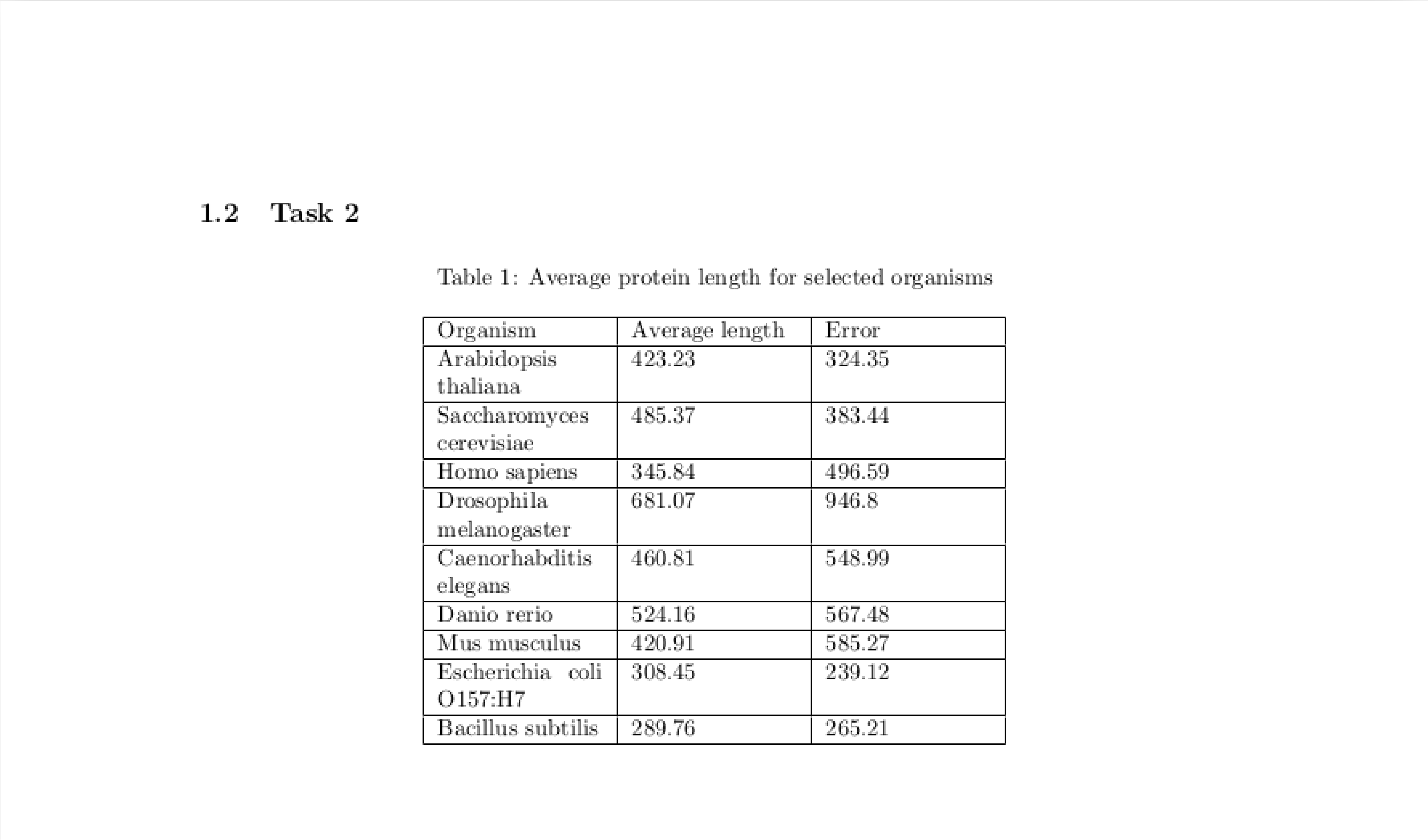


	Protein dataset				Peptide dataset		
Method	RMSD	%	Outliers	Method	RMSD	%	Outliers
IPC_protein	0.874	0.0	46	IPC_peptide	0.251	0.0	232
Toseland	0.934	14.9	52	Solomon	0.255	0.9	235
Bjellqvist	0.944	17.7	47	Lehninger	0.262	2.5	236
Dawson	0.945	17.8	56	EMBOSS	0.325	18.5	372
Wikipedia	0.955	20.5	55	Wikipedia	0.421	47.9	1467
Rodwell	0.963	22.8	58	Toseland	0.425	49.1	990
ProMoST	0.966	23.6	52	Sillero	0.428	50.3	1223
Grimsley	0.968	24.2	60	Dawson	0.435	52.9	1432
Solomon	0.970	24.8	58	Thurlkill	0.481	69.7	1361
Lehninger	0.970	25.0	59	Rodwell	0.502	78.4	1359
pIR	1.013	38.0	58	DTASelect	0.550	99.1	1714
Nozaki	1.024	41.3	56	Nozaki	0.602	124.3	1368
Thurlkill	1.030	43.4	61	Grimsley	0.616	131.4	1550
DTASelect	1.032	44.1	58	Bjellqvist	0.669	161.5	1583
pIPredict	1.048	49.4	56	pIPredict	1.024	493.6	2720
EMBOSS	1.056	52.3	69	ProMoST	1.239	873.4	2649
Sillero	1.059	53.2	63	pIR	1.881	4159.7	3358
Patrickios	2.392	3201.8	227	Patrickios	1.998	5479.1	2739
Avg_pI*	0.960	22.1	53	Avg_pI	0.454	59.6	1571

Less is more

avoid as many blank space as possible, correct the width of columns

Tables



1.2 Task 2

Table 1: Average protein length for selected organisms

Organism	Average length	Error
Arabidopsis thaliana	423.23	324.35
Saccharomyces cerevisiae	485.37	383.44
Homo sapiens	345.84	496.59
Drosophila melanogaster	681.07	946.8
Caenorhabditis elegans	460.81	548.99
Danio rerio	524.16	567.48
Mus musculus	420.91	585.27
Escherichia coli O157:H7	308.45	239.12
Bacillus subtilis	289.76	265.21

Some examples from students

Tables

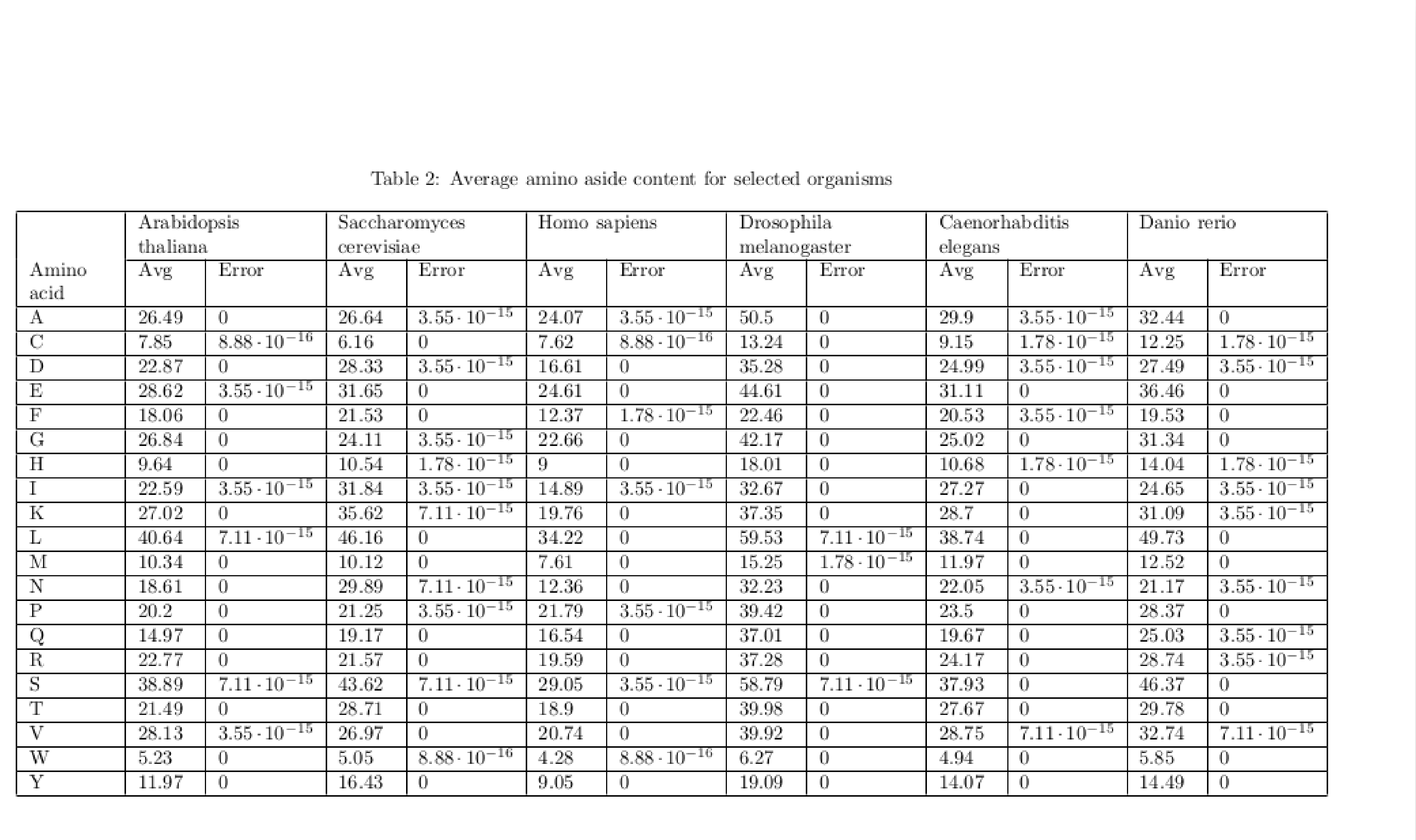


Table 2: Average amino aside content for selected organisms

	Arabidopsis thaliana		Saccharomyces cerevisiae		Homo sapiens		Drosophila melanogaster		Caenorhabditis elegans		Danio rerio	
Amino acid	Avg	Error	Avg	Error	Avg	Error	Avg	Error	Avg	Error	Avg	Error
A	26.49	0	26.64	$3.55 \cdot 10^{-15}$	24.07	$3.55 \cdot 10^{-15}$	50.5	0	29.9	$3.55 \cdot 10^{-15}$	32.44	0
C	7.85	$8.88 \cdot 10^{-16}$	6.16	0	7.62	$8.88 \cdot 10^{-16}$	13.24	0	9.15	$1.78 \cdot 10^{-15}$	12.25	$1.78 \cdot 10^{-15}$
D	22.87	0	28.33	$3.55 \cdot 10^{-15}$	16.61	0	35.28	0	24.99	$3.55 \cdot 10^{-15}$	27.49	$3.55 \cdot 10^{-15}$
E	28.62	$3.55 \cdot 10^{-15}$	31.65	0	24.61	0	44.61	0	31.11	0	36.46	0
F	18.06	0	21.53	0	12.37	$1.78 \cdot 10^{-15}$	22.46	0	20.53	$3.55 \cdot 10^{-15}$	19.53	0
G	26.84	0	24.11	$3.55 \cdot 10^{-15}$	22.66	0	42.17	0	25.02	0	31.34	0
H	9.64	0	10.54	$1.78 \cdot 10^{-15}$	9	0	18.01	0	10.68	$1.78 \cdot 10^{-15}$	14.04	$1.78 \cdot 10^{-15}$
I	22.59	$3.55 \cdot 10^{-15}$	31.84	$3.55 \cdot 10^{-15}$	14.89	$3.55 \cdot 10^{-15}$	32.67	0	27.27	0	24.65	$3.55 \cdot 10^{-15}$
K	27.02	0	35.62	$7.11 \cdot 10^{-15}$	19.76	0	37.35	0	28.7	0	31.09	$3.55 \cdot 10^{-15}$
L	40.64	$7.11 \cdot 10^{-15}$	46.16	0	34.22	0	59.53	$7.11 \cdot 10^{-15}$	38.74	0	49.73	0
M	10.34	0	10.12	0	7.61	0	15.25	$1.78 \cdot 10^{-15}$	11.97	0	12.52	0
N	18.61	0	29.89	$7.11 \cdot 10^{-15}$	12.36	0	32.23	0	22.05	$3.55 \cdot 10^{-15}$	21.17	$3.55 \cdot 10^{-15}$
P	20.2	0	21.25	$3.55 \cdot 10^{-15}$	21.79	$3.55 \cdot 10^{-15}$	39.42	0	23.5	0	28.37	0
Q	14.97	0	19.17	0	16.54	0	37.01	0	19.67	0	25.03	$3.55 \cdot 10^{-15}$
R	22.77	0	21.57	0	19.59	0	37.28	0	24.17	0	28.74	$3.55 \cdot 10^{-15}$
S	38.89	$7.11 \cdot 10^{-15}$	43.62	$7.11 \cdot 10^{-15}$	29.05	$3.55 \cdot 10^{-15}$	58.79	$7.11 \cdot 10^{-15}$	37.93	0	46.37	0
T	21.49	0	28.71	0	18.9	0	39.98	0	27.67	0	29.78	0
V	28.13	$3.55 \cdot 10^{-15}$	26.97	0	20.74	0	39.92	0	28.75	$7.11 \cdot 10^{-15}$	32.74	$7.11 \cdot 10^{-15}$
W	5.23	0	5.05	$8.88 \cdot 10^{-16}$	4.28	$8.88 \cdot 10^{-16}$	6.27	0	4.94	0	5.85	0
Y	11.97	0	16.43	0	9.05	0	19.09	0	14.07	0	14.49	0

Some examples from students

Tables

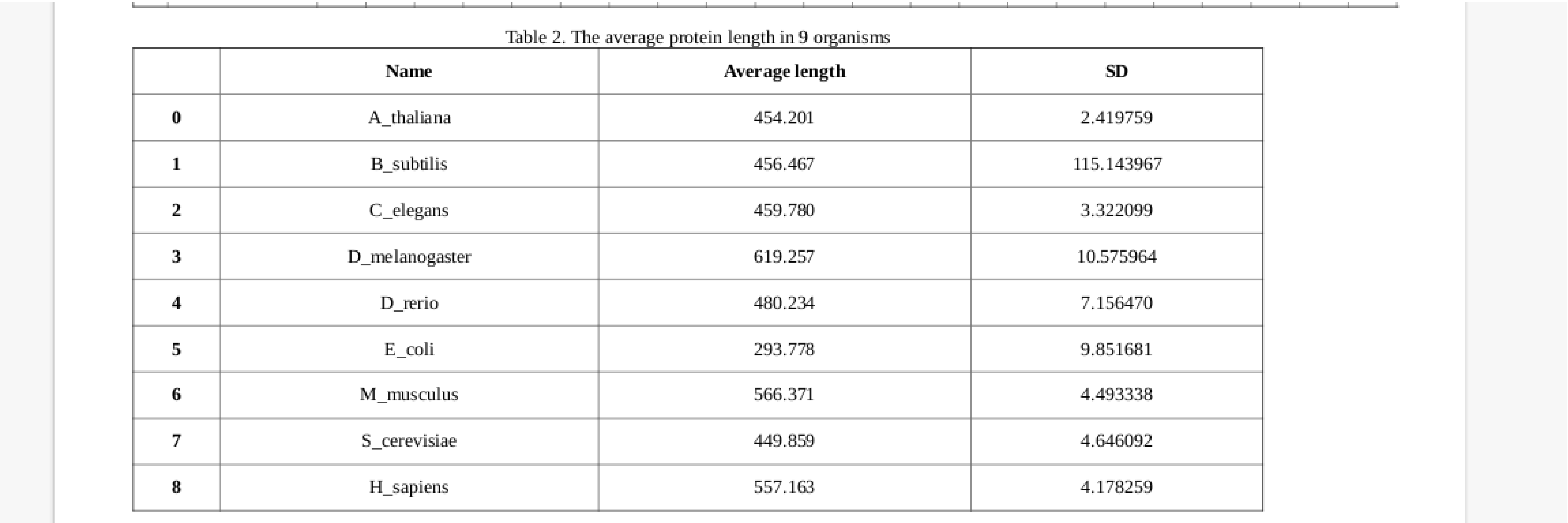
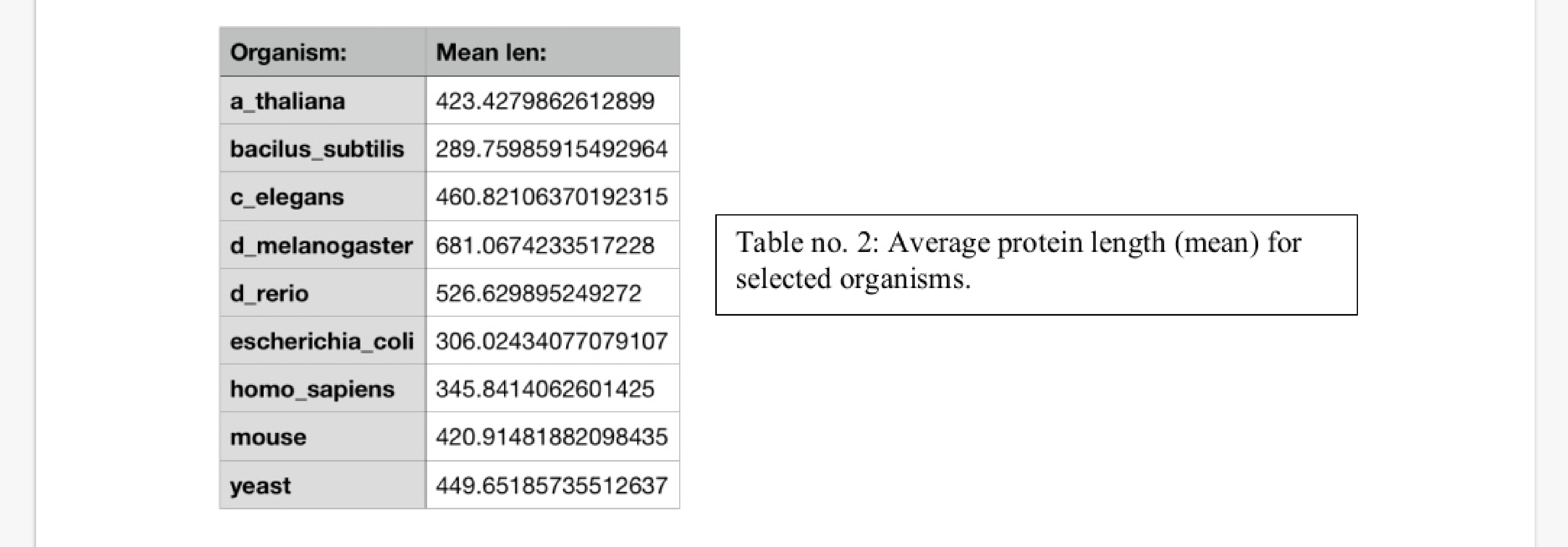


Table 2. The average protein length in 9 organisms

	Name	Average length	SD
0	A_thaliana	454.201	2.419759
1	B_subtilis	456.467	115.143967
2	C_elegans	459.780	3.322099
3	D_melanogaster	619.257	10.575964
4	D_rerio	480.234	7.156470
5	E_coli	293.778	9.851681
6	M_musculus	566.371	4.493338
7	S_cerevisiae	449.859	4.646092
8	H_sapiens	557.163	4.178259

Some examples from students

Tables

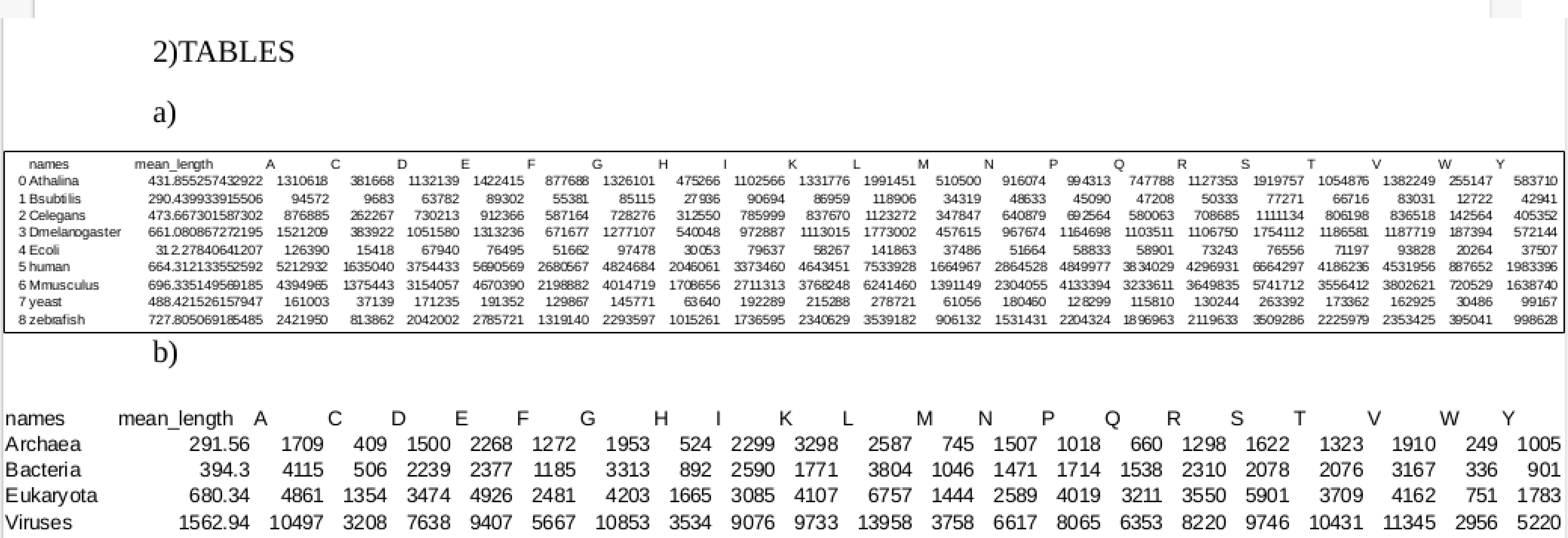


Organism:	Mean len:
a_thaliana	423.4279862612899
bacilus_subtilis	289.75985915492964
c_elegans	460.82106370192315
d_melanogaster	681.0674233517228
d_rerio	526.629895249272
escherichia_coli	306.02434077079107
homo_sapiens	345.8414062601425
mouse	420.91481882098435
yeast	449.65185735512637

Table no. 2: Average protein length (mean) for selected organisms.

Some examples from students

Tables



2)TABLES

a)

	names	mean_length	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	Athalina	431.855257432922	1310618	381668	1132139	1422415	877688	1326101	475266	1102566	1331776	1991451	510500	916074	994313	747788	1127353	1919757	1054876	1382249	255147	583710
1	Bsubtilis	290.439933915506	94572	9683	63782	89302	55381	85115	27936	90694	86959	118906	34319	48633	45090	47208	50333	77271	66716	83031	12722	42941
2	Celegans	473.667301587302	876885	262267	730213	912366	587164	728276	312550	785999	837670	1123272	347847	640879	692564	580063	708685	1111134	806198	836518	142564	405352
3	Dmelanogaster	661.080867272195	1521209	383922	1051580	1313236	671677	1277107	540048	972887	1113015	1773002	457615	967674	1164698	1103511	1106750	1754112	1186581	1187719	187394	572144
4	Ecoli	312.27840641207	126390	15418	67940	76495	51662	97478	30053	79637	58267	141863	37486	51664	58833	58901	73243	76556	71197	93828	20264	37507
5	human	664.312133552592	5212932	1635040	3754433	5690569	2680567	4824684	2046061	3373460	4643451	7533928	1664967	2864528	4849977	3834029	4296931	6664297	4186236	4531956	887652	1983396
6	Mmusculus	696.335149569185	4394965	1375443	3154057	4670390	2198882	4014719	1708656	2711313	3768248	6241460	1391149	2304055	4133394	3233611	3649835	5741712	3556412	3802621	720529	1638740
7	yeast	488.421526157947	161003	37139	171235	191352	129867	145771	63640	192289	215288	278721	61056	180460	128299	115810	130244	263392	173362	162925	30486	99167
8	zebrafish	727.805069185485	2421950	813862	2042002	2785721	1319140	2293597	1015261	1736595	2340629	3539182	906132	1531431	2204324	1896963	2119633	3509286	2225979	2353425	395041	998628

b)

names	mean_length	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Archaea	291.56	1709	409	1500	2268	1272	1953	524	2299	3298	2587	745	1507	1018	660	1298	1622	1323	1910	249	1005
Bacteria	394.3	4115	506	2239	2377	1185	3313	892	2590	1771	3804	1046	1471	1714	1538	2310	2078	2076	3167	336	901
Eukaryota	680.34	4861	1354	3474	4926	2481	4203	1665	3085	4107	6757	1444	2589	4019	3211	3550	5901	3709	4162	751	1783
Viruses	1562.94	10497	3208	7638	9407	5667	10853	3534	9076	9733	13958	3758	6617	8065	6353	8220	9746	10431	11345	2956	5220

Some examples from students

Tables

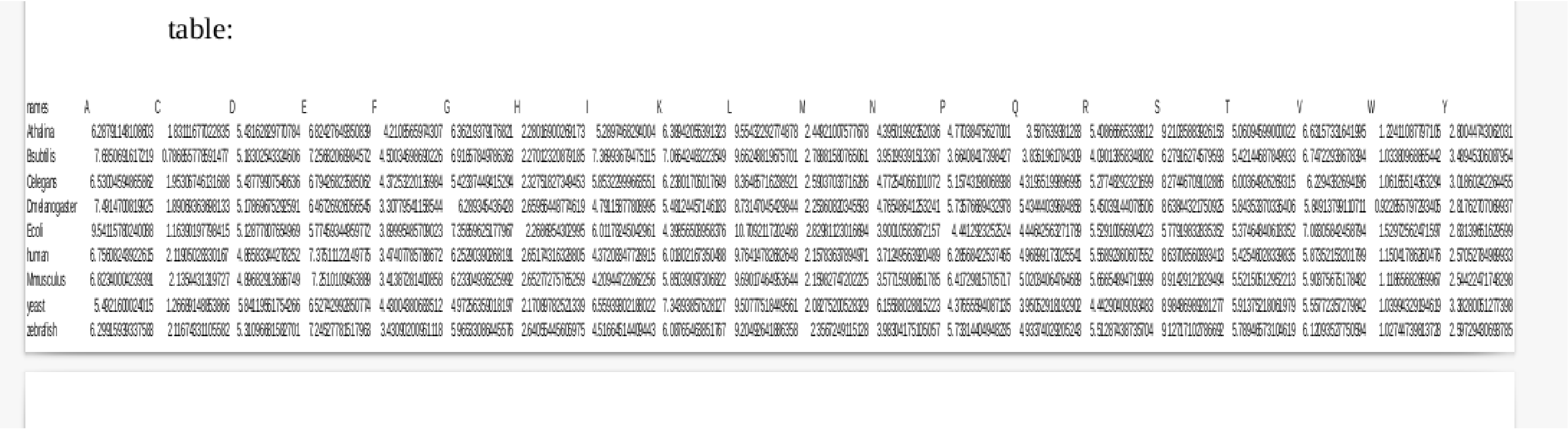
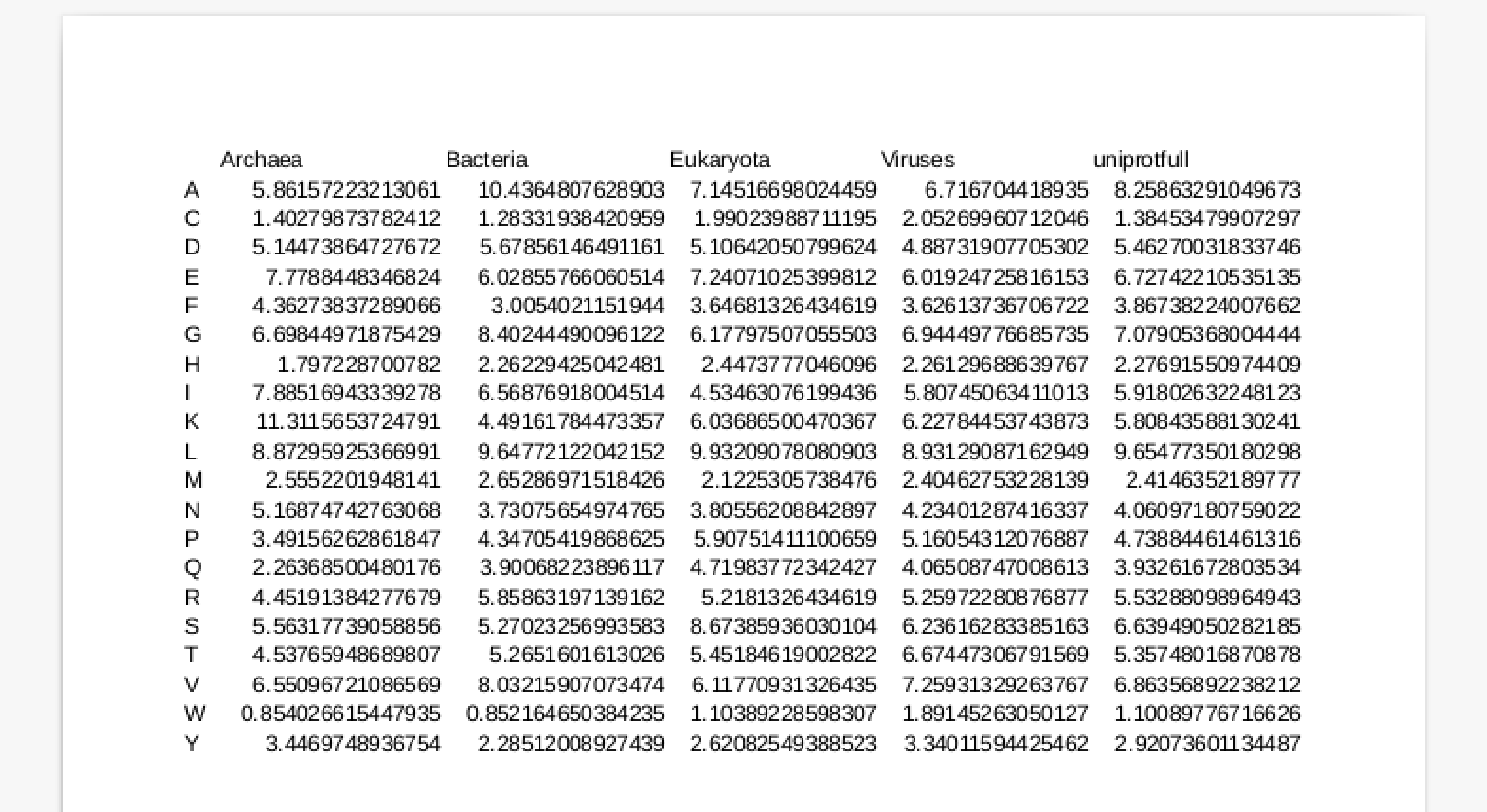


table:

Some examples from students

Tables



	Archaea	Bacteria	Eukaryota	Viruses	uniprotfull
A	5.86157223213061	10.4364807628903	7.14516698024459	6.716704418935	8.25863291049673
C	1.40279873782412	1.28331938420959	1.99023988711195	2.05269960712046	1.38453479907297
D	5.14473864727672	5.67856146491161	5.10642050799624	4.88731907705302	5.46270031833746
E	7.7788448346824	6.02855766060514	7.24071025399812	6.01924725816153	6.72742210535135
F	4.36273837289066	3.0054021151944	3.64681326434619	3.62613736706722	3.86738224007662
G	6.69844971875429	8.40244490096122	6.17797507055503	6.94449776685735	7.07905368004444
H	1.797228700782	2.26229425042481	2.4473777046096	2.26129688639767	2.27691550974409
I	7.88516943339278	6.56876918004514	4.53463076199436	5.80745063411013	5.91802632248123
K	11.3115653724791	4.49161784473357	6.03686500470367	6.22784453743873	5.80843588130241
L	8.87295925366991	9.64772122042152	9.93209078080903	8.93129087162949	9.65477350180298
M	2.5552201948141	2.65286971518426	2.1225305738476	2.40462753228139	2.4146352189777
N	5.16874742763068	3.73075654974765	3.80556208842897	4.23401287416337	4.06097180759022
P	3.49156262861847	4.34705419868625	5.90751411100659	5.16054312076887	4.73884461461316
Q	2.26368500480176	3.90068223896117	4.71983772342427	4.06508747008613	3.93261672803534
R	4.45191384277679	5.85863197139162	5.2181326434619	5.25972280876877	5.53288098964943
S	5.56317739058856	5.27023256993583	8.67385936030104	6.23616283385163	6.63949050282185
T	4.53765948689807	5.2651601613026	5.45184619002822	6.67447306791569	5.35748016870878
V	6.55096721086569	8.03215907073474	6.11770931326435	7.25931329263767	6.86356892238212
W	0.854026615447935	0.852164650384235	1.10389228598307	1.89145263050127	1.10089776716626
Y	3.4469748936754	2.28512008927439	2.62082549388523	3.34011594425462	2.92073601134487

Some examples from students

Tables

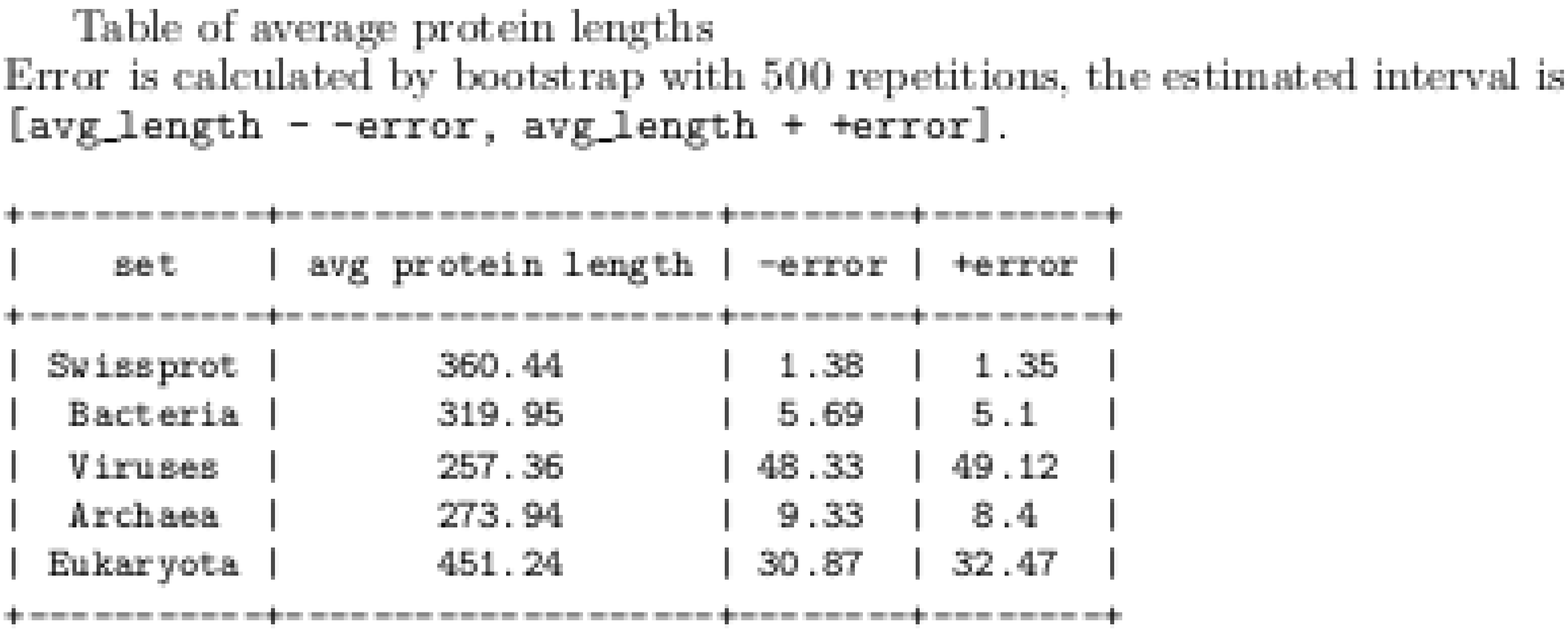


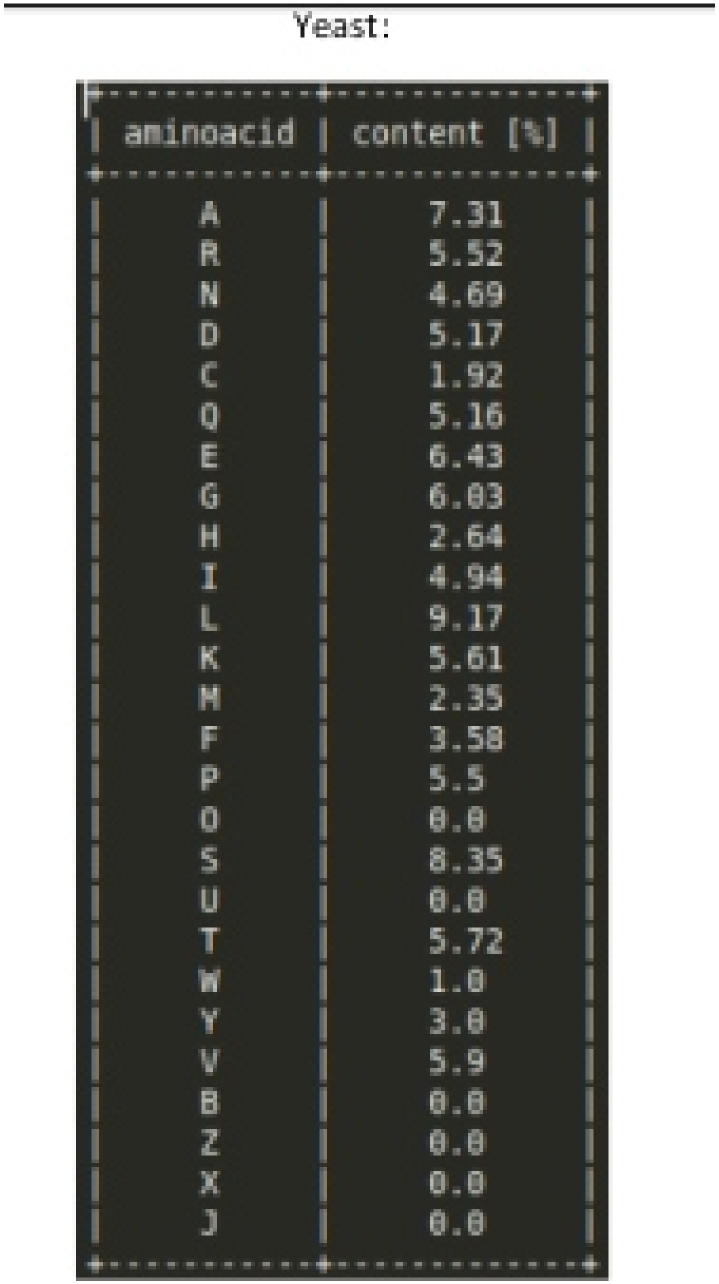
Table of average protein lengths

Error is calculated by bootstrap with 500 repetitions, the estimated interval is `[avg_length - -error, avg_length + +error]`.

```
+-----------+--------------------+--------+--------+
|    set    | avg protein length | -error | +error |
+-----------+--------------------+--------+--------+
| Swissprot |       360.44       |  1.38  |  1.35  |
|  Bacteria |       319.95       |  5.69  |  5.1   |
|  Viruses  |       257.36       | 48.33  | 49.12  |
|  Archaea  |       273.94       |  9.33  |  8.4   |
| Eukaryota |       451.24       | 30.87  | 32.47  |
+-----------+--------------------+--------+--------+
```

Some examples from students

Tables



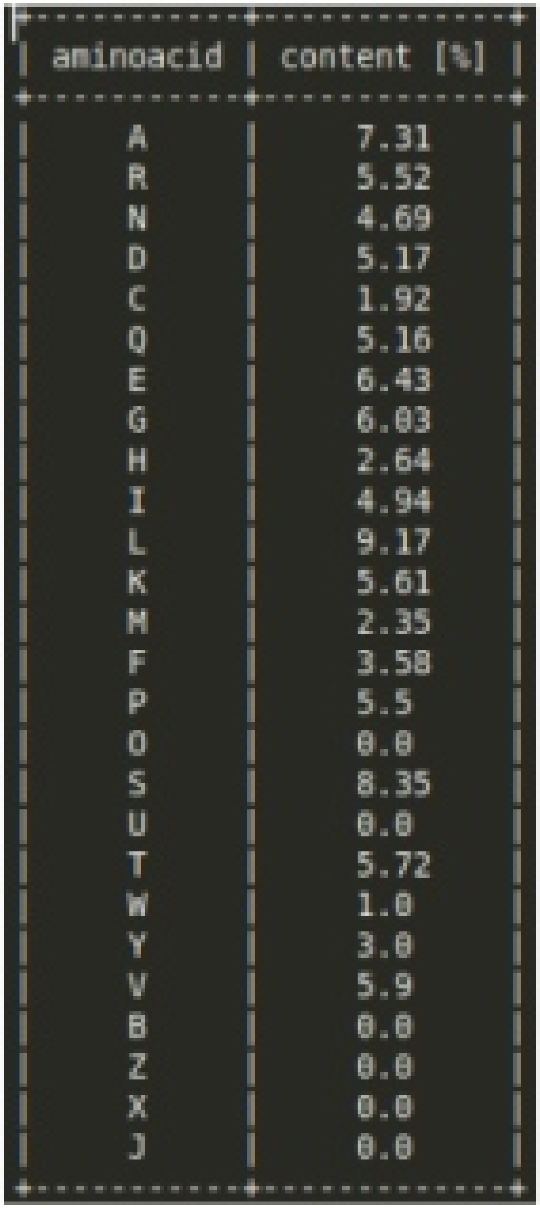
Yeast:

```
+------------+--------------+
| aminoacid  | content [%]  |
+------------+--------------+
|     A      |     7.31     |
|     R      |     5.52     |
|     N      |     4.69     |
|     D      |     5.17     |
|     C      |     1.92     |
|     Q      |     5.16     |
|     E      |     6.43     |
|     G      |     6.03     |
|     H      |     2.64     |
|     I      |     4.94     |
|     L      |     9.17     |
|     K      |     5.61     |
|     M      |     2.35     |
|     F      |     3.58     |
|     P      |     5.5      |
|     O      |     0.0      |
|     S      |     8.35     |
|     U      |     0.0      |
|     T      |     5.72     |
|     W      |     1.0      |
|     Y      |     3.0      |
|     V      |     5.9      |
|     B      |     0.0      |
|     Z      |     0.0      |
|     X      |     0.0      |
|     J      |     0.0      |
+------------+--------------+
```

Some examples from students

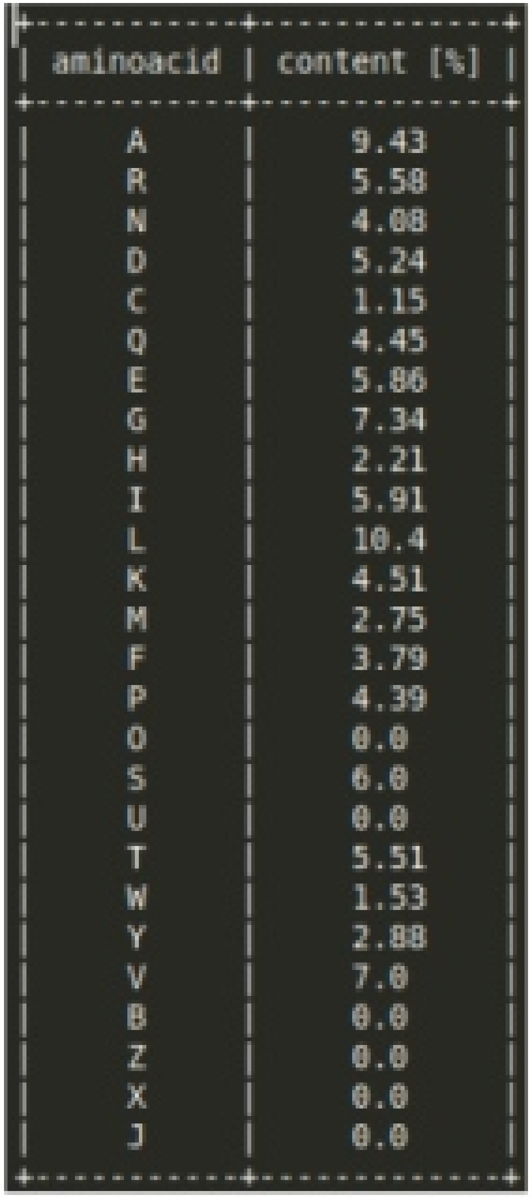
Tables

Yeast:



```
+------------+-------------+
| aminoacid  | content [%] |
+------------+-------------+
|     A      |    7.31     |
|     R      |    5.52     |
|     N      |    4.69     |
|     D      |    5.17     |
|     C      |    1.92     |
|     Q      |    5.16     |
|     E      |    6.43     |
|     G      |    6.03     |
|     H      |    2.64     |
|     I      |    4.94     |
|     L      |    9.17     |
|     K      |    5.61     |
|     M      |    2.35     |
|     F      |    3.58     |
|     P      |    5.5      |
|     O      |    0.0      |
|     S      |    8.35     |
|     U      |    0.0      |
|     T      |    5.72     |
|     W      |    1.0      |
|     Y      |    3.0      |
|     V      |    5.9      |
|     B      |    0.0      |
|     Z      |    0.0      |
|     X      |    0.0      |
|     J      |    0.0      |
+------------+-------------+
```

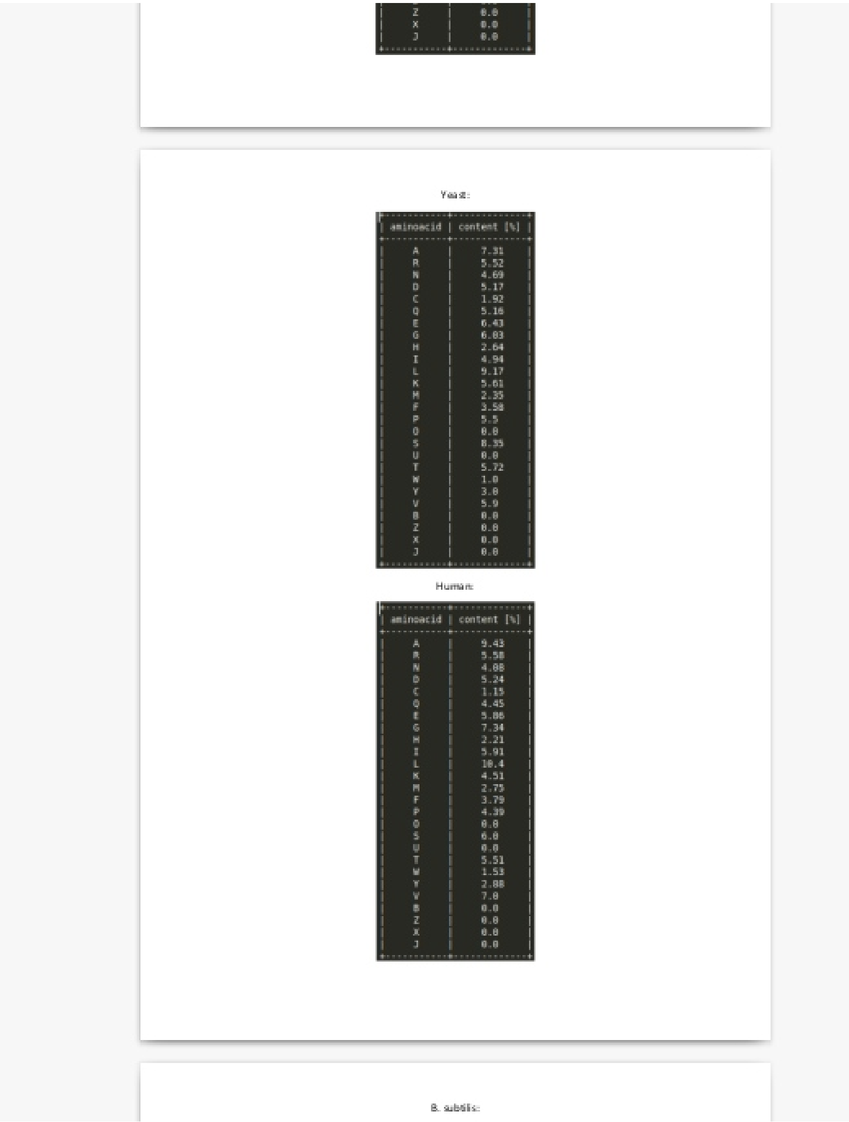
Human:



```
+------------+-------------+
| aminoacid  | content [%] |
+------------+-------------+
|     A      |    9.43     |
|     R      |    5.58     |
|     N      |    4.08     |
|     D      |    5.24     |
|     C      |    1.15     |
|     Q      |    4.45     |
|     E      |    5.86     |
|     G      |    7.34     |
|     H      |    2.21     |
|     I      |    5.91     |
|     L      |    10.4     |
|     K      |    4.51     |
|     M      |    2.75     |
|     F      |    3.79     |
|     P      |    4.39     |
|     O      |    0.0      |
|     S      |    6.0      |
|     U      |    0.0      |
|     T      |    5.51     |
|     W      |    1.53     |
|     Y      |    2.88     |
|     V      |    7.0      |
|     B      |    0.0      |
|     Z      |    0.0      |
|     X      |    0.0      |
|     J      |    0.0      |
+------------+-------------+
```

Some examples from students

Tables



```
|     Z     |     0.0     |
|     X     |     0.0     |
|     J     |     0.0     |
+-----------+-------------+
```

Yeast:

```
+-----------+-------------+
| aminoacid | content [%] |
+-----------+-------------+
|     A     |     7.31    |
|     R     |     5.52    |
|     N     |     4.69    |
|     D     |     5.17    |
|     C     |     1.92    |
|     Q     |     5.16    |
|     E     |     6.43    |
|     G     |     6.03    |
|     H     |     2.64    |
|     I     |     4.94    |
|     L     |     9.17    |
|     K     |     5.61    |
|     M     |     2.35    |
|     F     |     3.58    |
|     P     |     5.5     |
|     O     |     0.0     |
|     S     |     8.35    |
|     U     |     0.0     |
|     T     |     5.72    |
|     W     |     1.0     |
|     Y     |     3.0     |
|     V     |     5.9     |
|     B     |     0.0     |
|     Z     |     0.0     |
|     X     |     0.0     |
|     J     |     0.0     |
+-----------+-------------+
```

Human:

```
+-----------+-------------+
| aminoacid | content [%] |
+-----------+-------------+
|     A     |     9.43    |
|     R     |     5.58    |
|     N     |     4.08    |
|     D     |     5.24    |
|     C     |     1.15    |
|     Q     |     4.45    |
|     E     |     5.06    |
|     G     |     7.34    |
|     H     |     2.21    |
|     I     |     5.91    |
|     L     |     10.4    |
|     K     |     4.51    |
|     M     |     2.75    |
|     F     |     3.79    |
|     P     |     4.39    |
|     O     |     0.0     |
|     S     |     6.0     |
|     U     |     0.0     |
|     T     |     5.51    |
|     W     |     1.53    |
|     Y     |     2.88    |
|     V     |     7.0     |
|     B     |     0.0     |
|     Z     |     0.0     |
|     X     |     0.0     |
|     J     |     0.0     |
+-----------+-------------+
```

B. subtilis:

Some examples from students

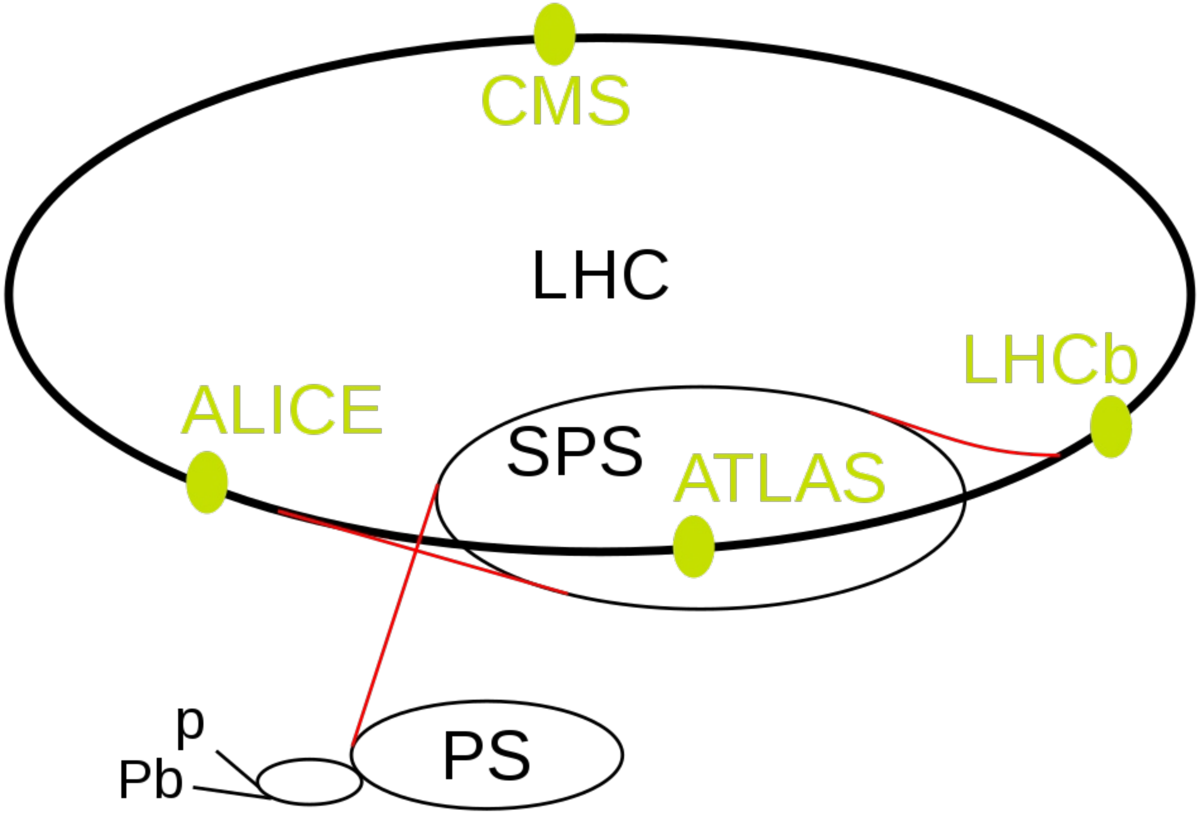
Extra big consortia/initiatives

Consortia

Scientific Competitions

ATLAS experiment

ATLAS (A Toroidal LHC ApparatuS)

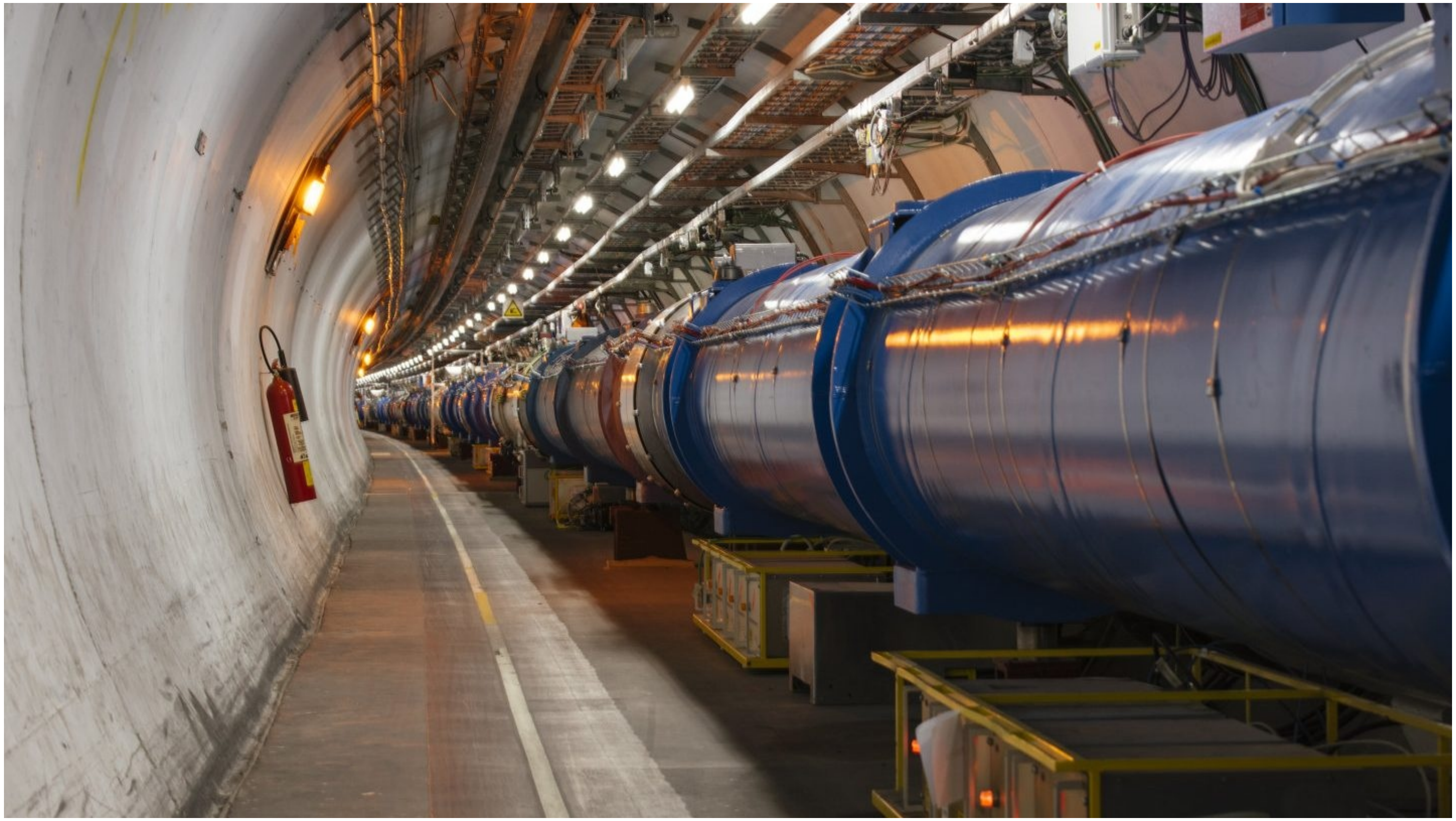


ATLAS experiment

ATLAS (A Toroidal LHC ApparatuS)

ATLAS experiment

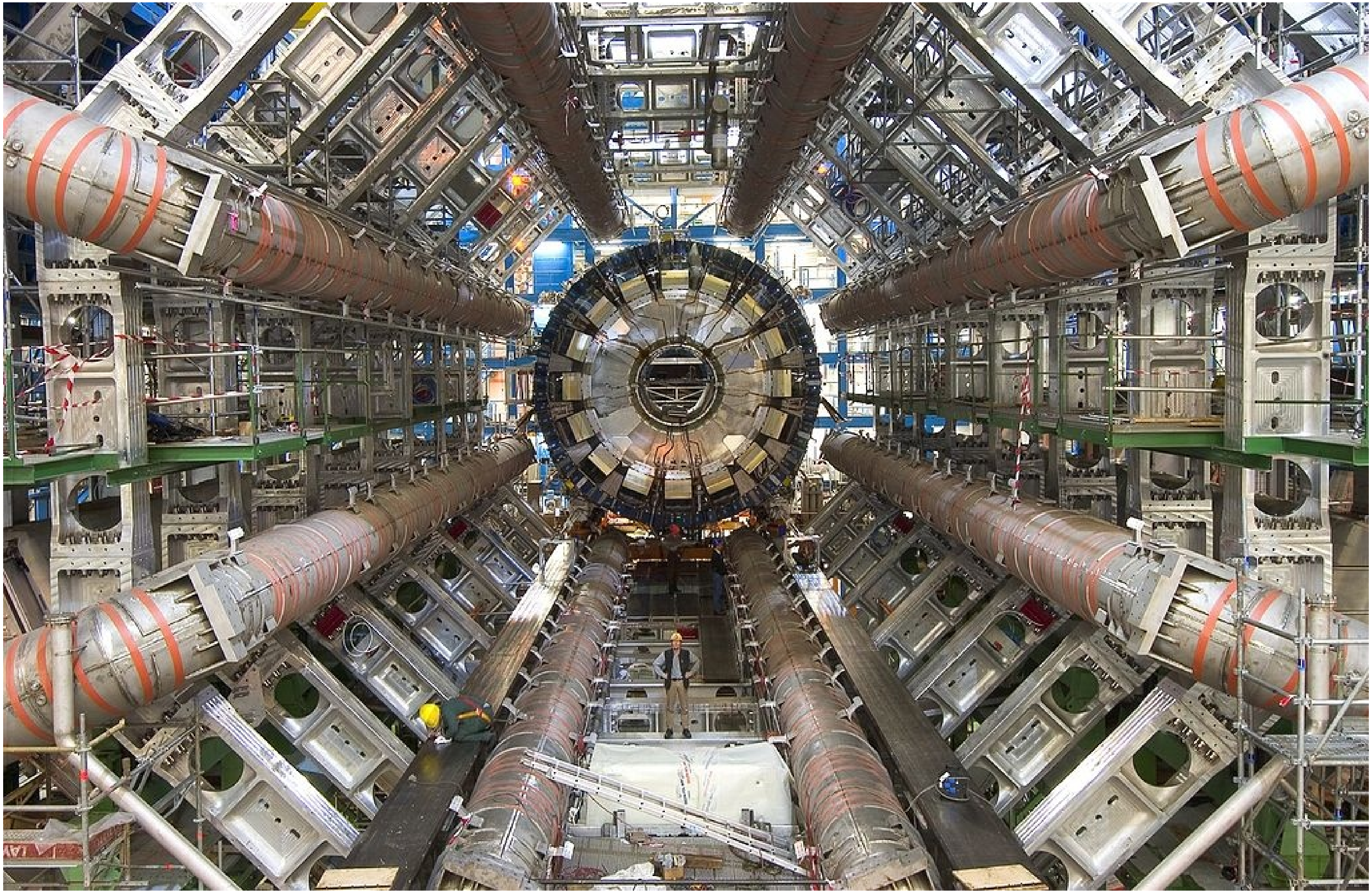
ATLAS (A Toroidal LHC ApparatuS)



ATLAS experiment

ATLAS (A Toroidal LHC ApparatuS)

ATLAS experiment



ATLAS detector is 46 metres long, 25 metres in diameter, and weighs about 7,000 tonnes; it contains some 3000 km of cable

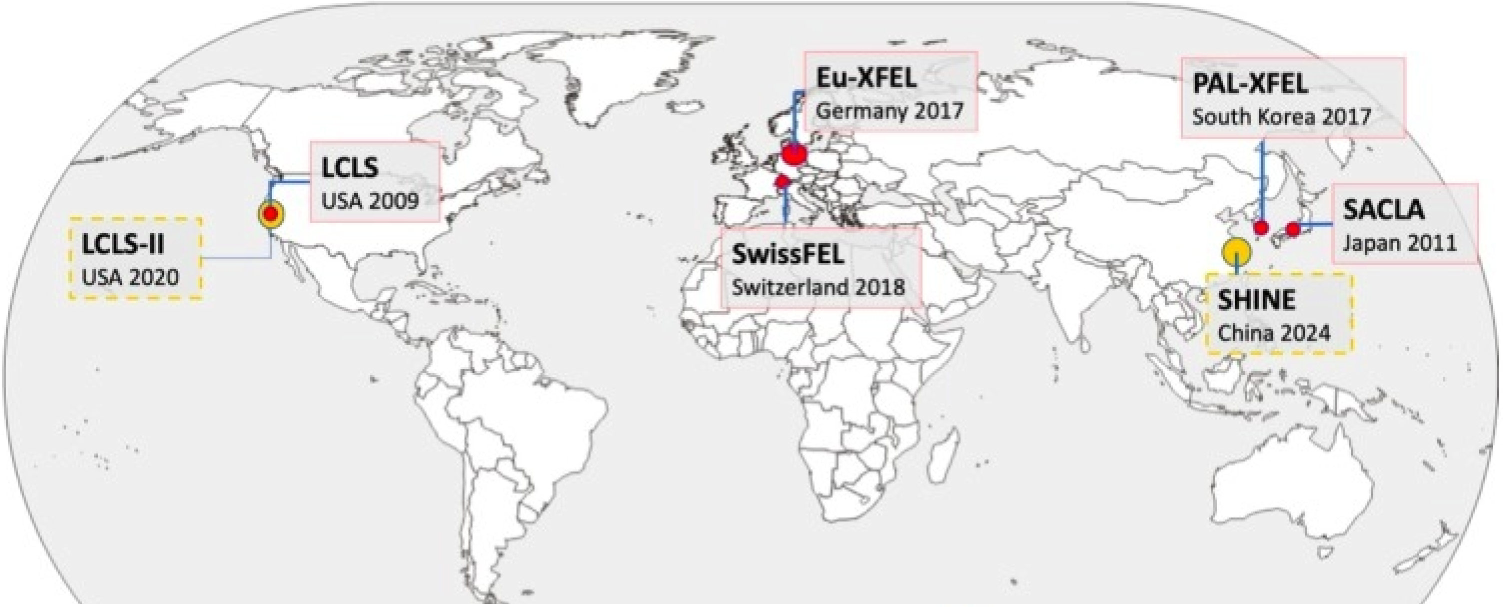
ATLAS experiment

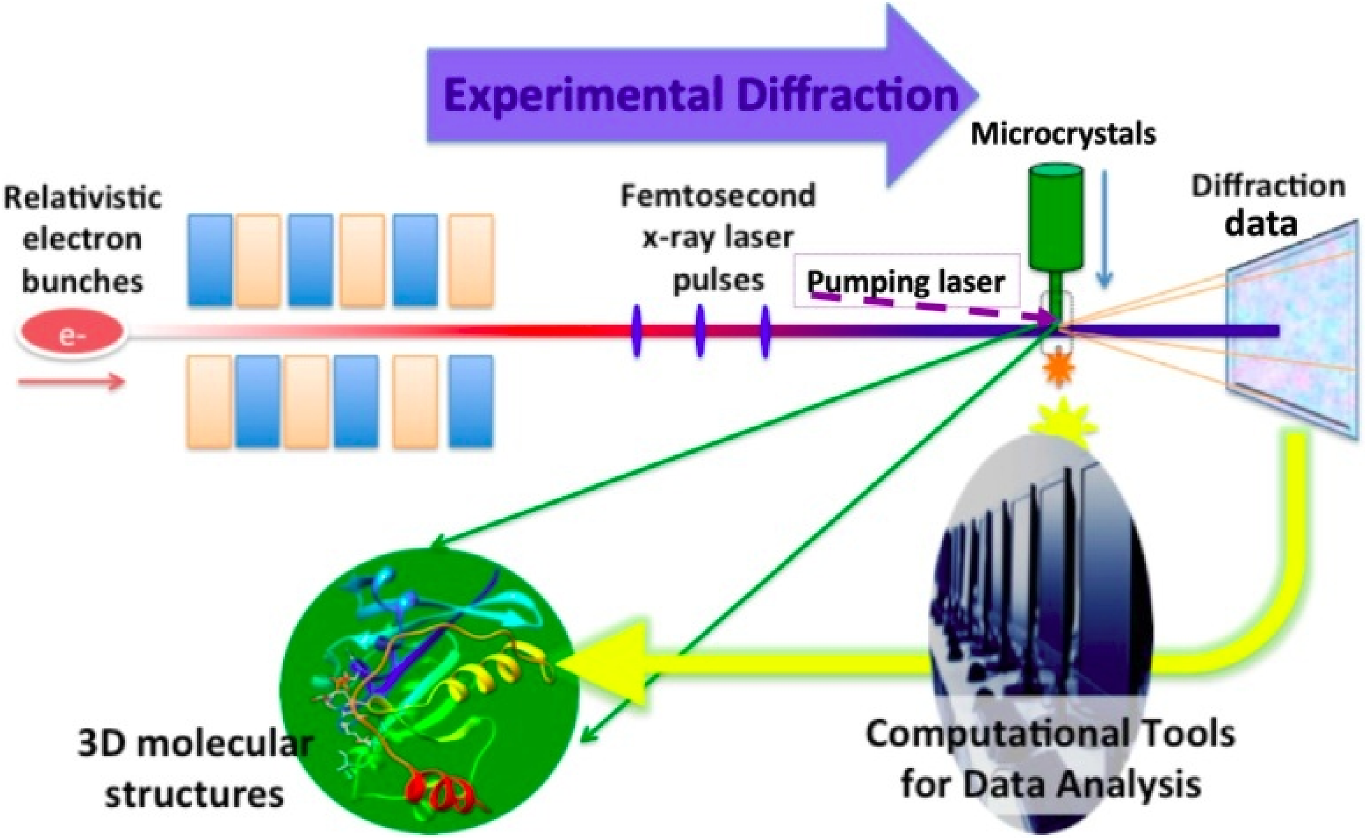
The experiment is a collaboration involving roughly 10,000 physicists from hundreds institutions in >100 countries

Budget of €7.5 billion

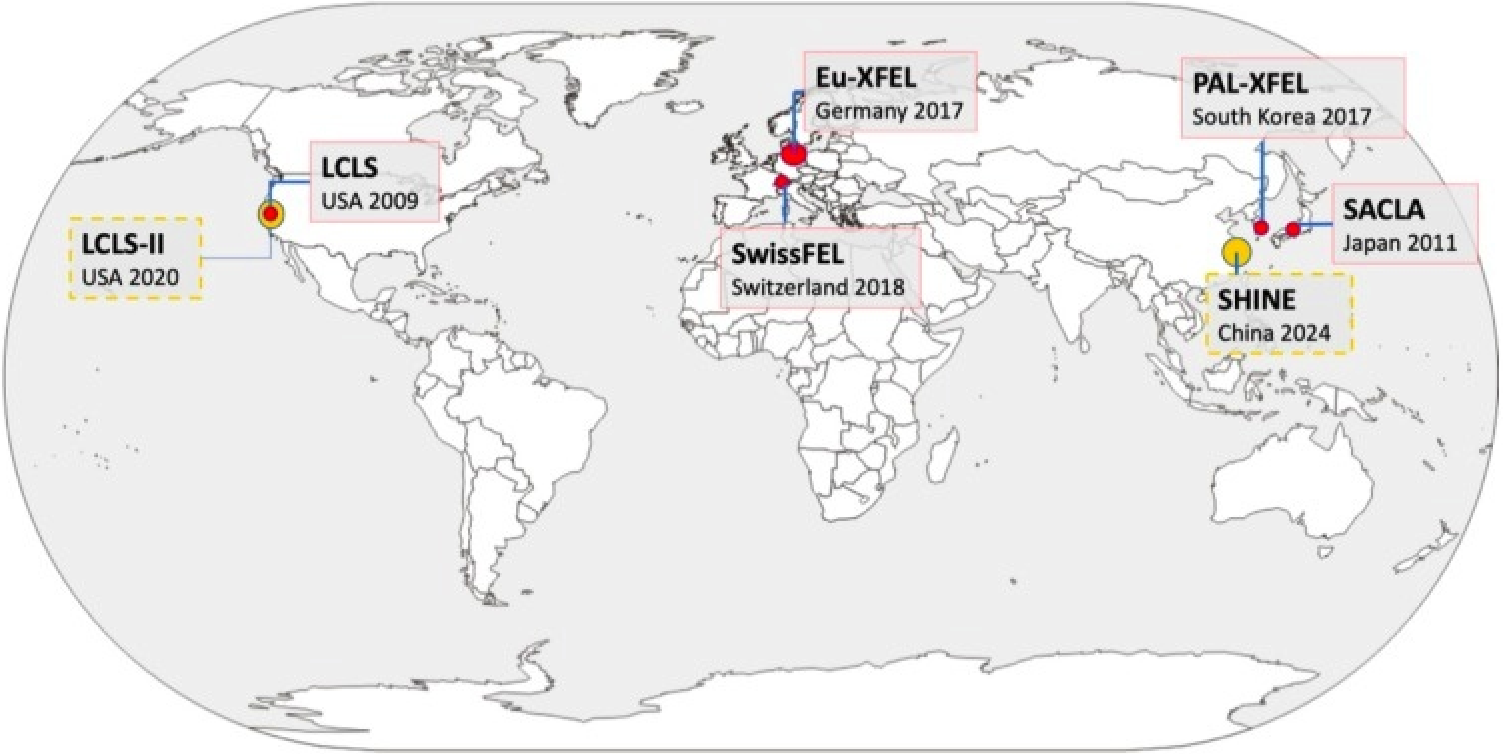
First ring build in 1971–1984

Extra big consortia/initiatives





Extra big consortia/initiatives



European XFEL

3.4-kilometre (2.1 mi) long tunnel

Cost for the construction and commissioning of the facility is as of 2017 estimated at €1.22 billion

Human Brain Project

Future Emerging Technologies (FET) Flagships from EU

121 partners from universities, research institutes and companies in 20 countries

Launched in 2013

Budget of €1 billion (until 2019)

Brain, graphene and quantum technologies

https://www.humanbrainproject.eu

Human Brain Project

Future Emerging Technologies (FET) Flagships from EU

121 partners from universities, research institutes and companies in 20 countries

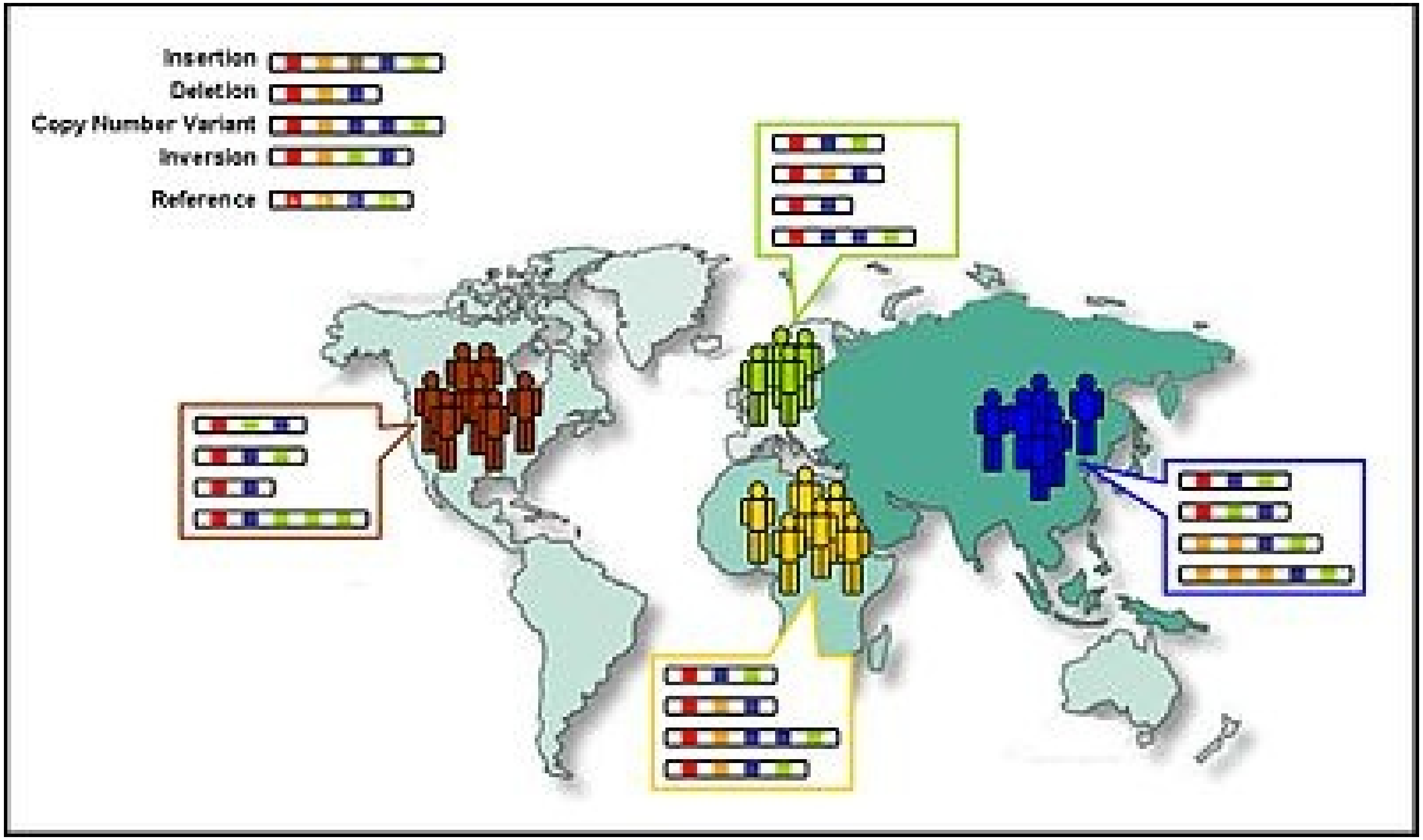
Launched in 2013

Budget of €1 billion (until 2019)

Brain, graphene and quantum technologies

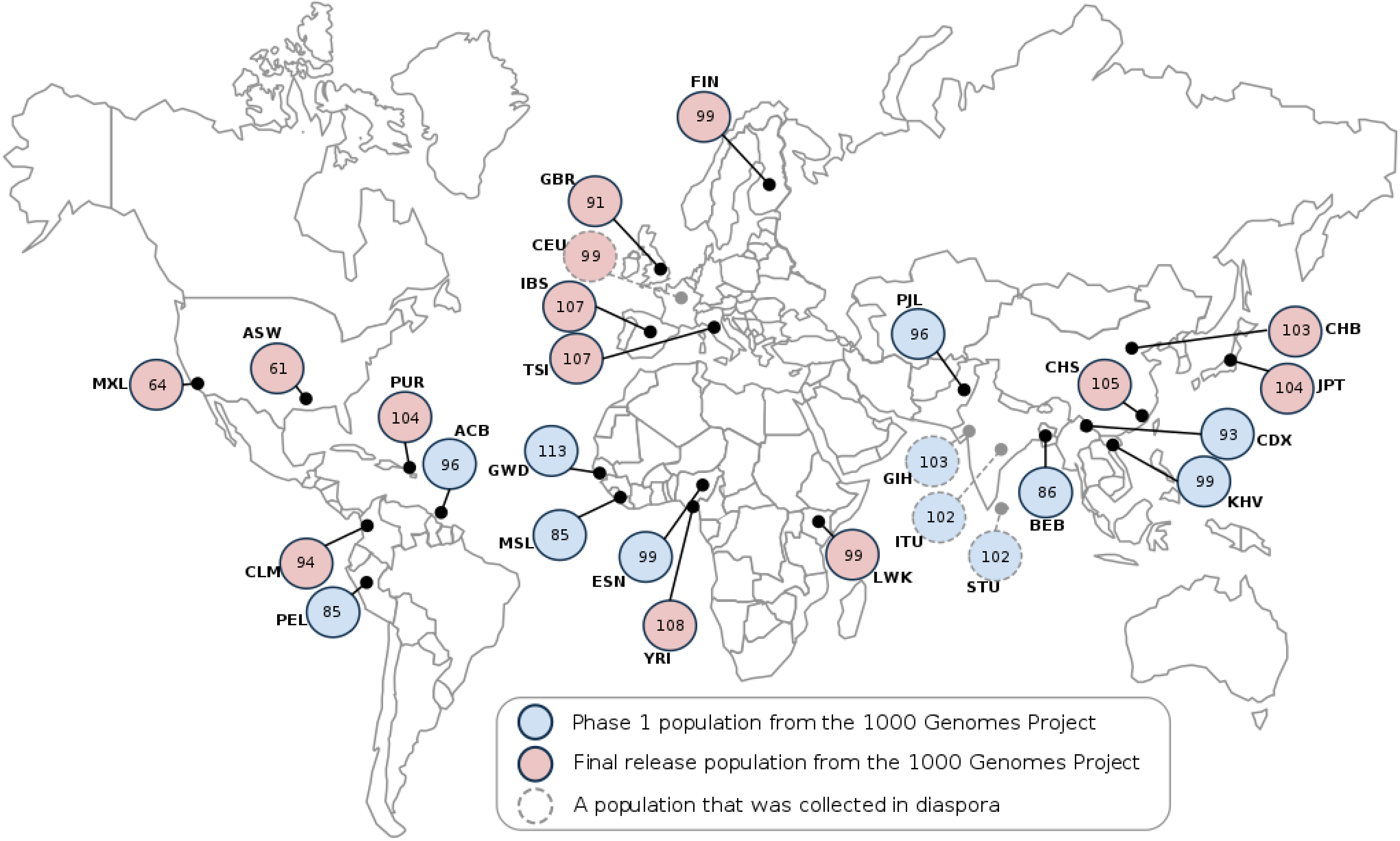
16 of these projects will collaborate with the **Graphene** Flagship and nine with the **Human Brain Project**. These projects will be funded by a total budget of € 16.4 million and are expected to start between December 2019 and March 2020

1000 Genomes Project



https://www.internationalgenome.org

1000 Genomes Project



1000 Plant Genomes Project (1KP)

Followed by 10,000 Plant Genome Project

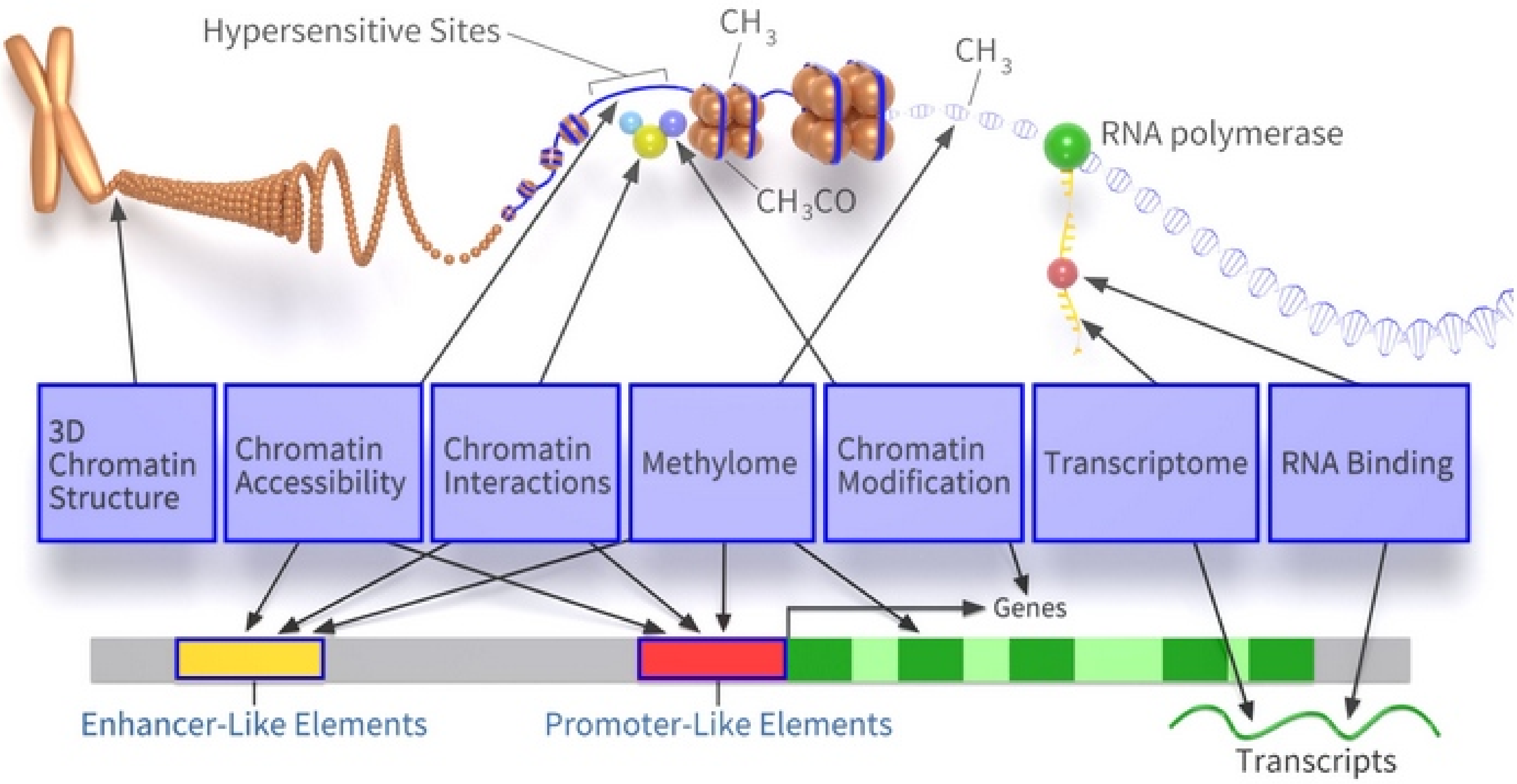
1000 Plant Genomes Project

Funding agency	Alberta Innovates Technology Futures Alberta Agricultural Research Institute (AARI) Genome Alberta University of Alberta BGI China National GeneBank (CNGB) Musea Ventures (Somekh Family Foundation)
Duration	2008 – 2019
Website	www.onekp.com

https://sites.google.com/a/ualberta.ca/onekp/

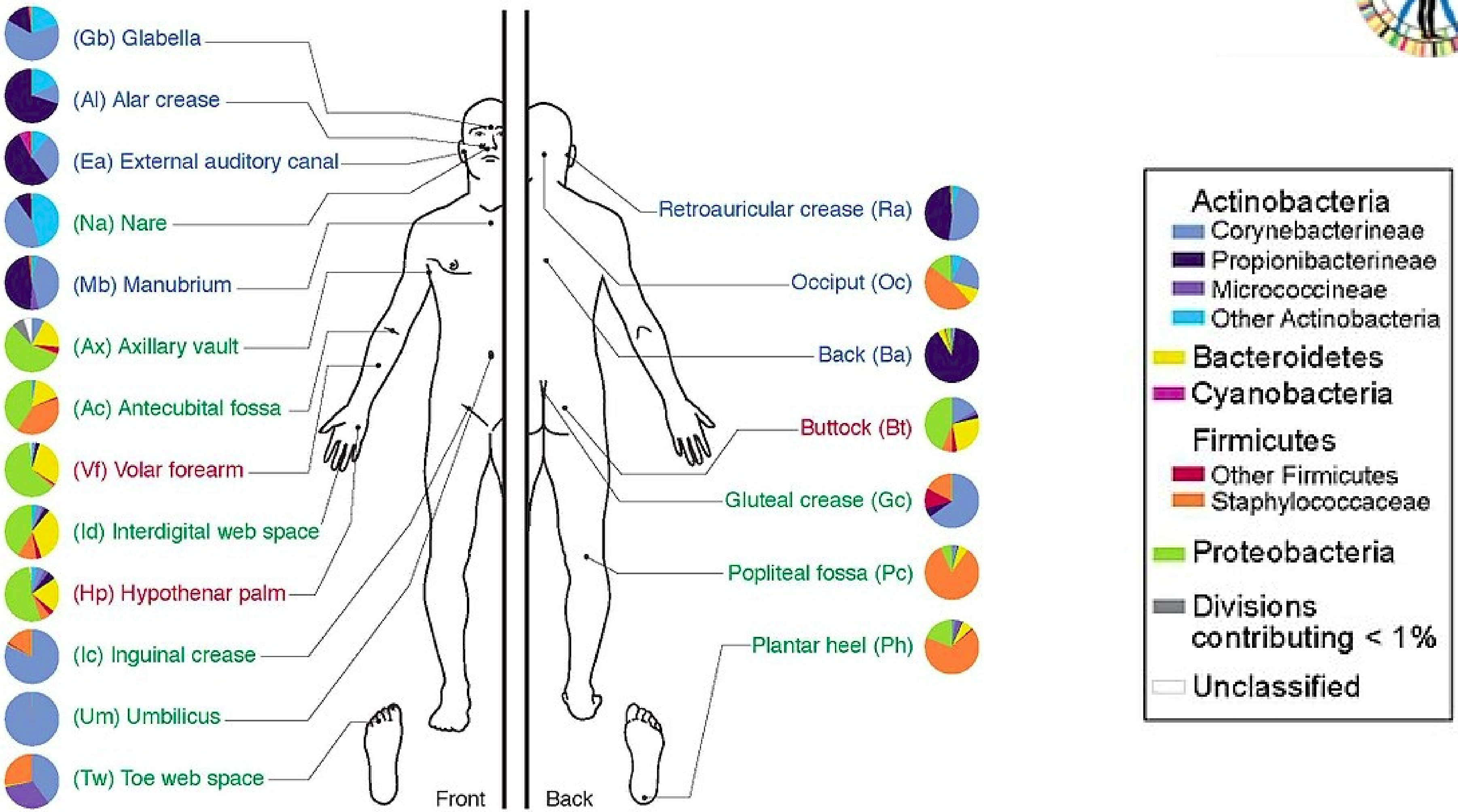
Extra big consortia/initiatives

Encyclopedia of DNA Elements (ENCODE) is a public research project which aims to identify functional elements in the human genome



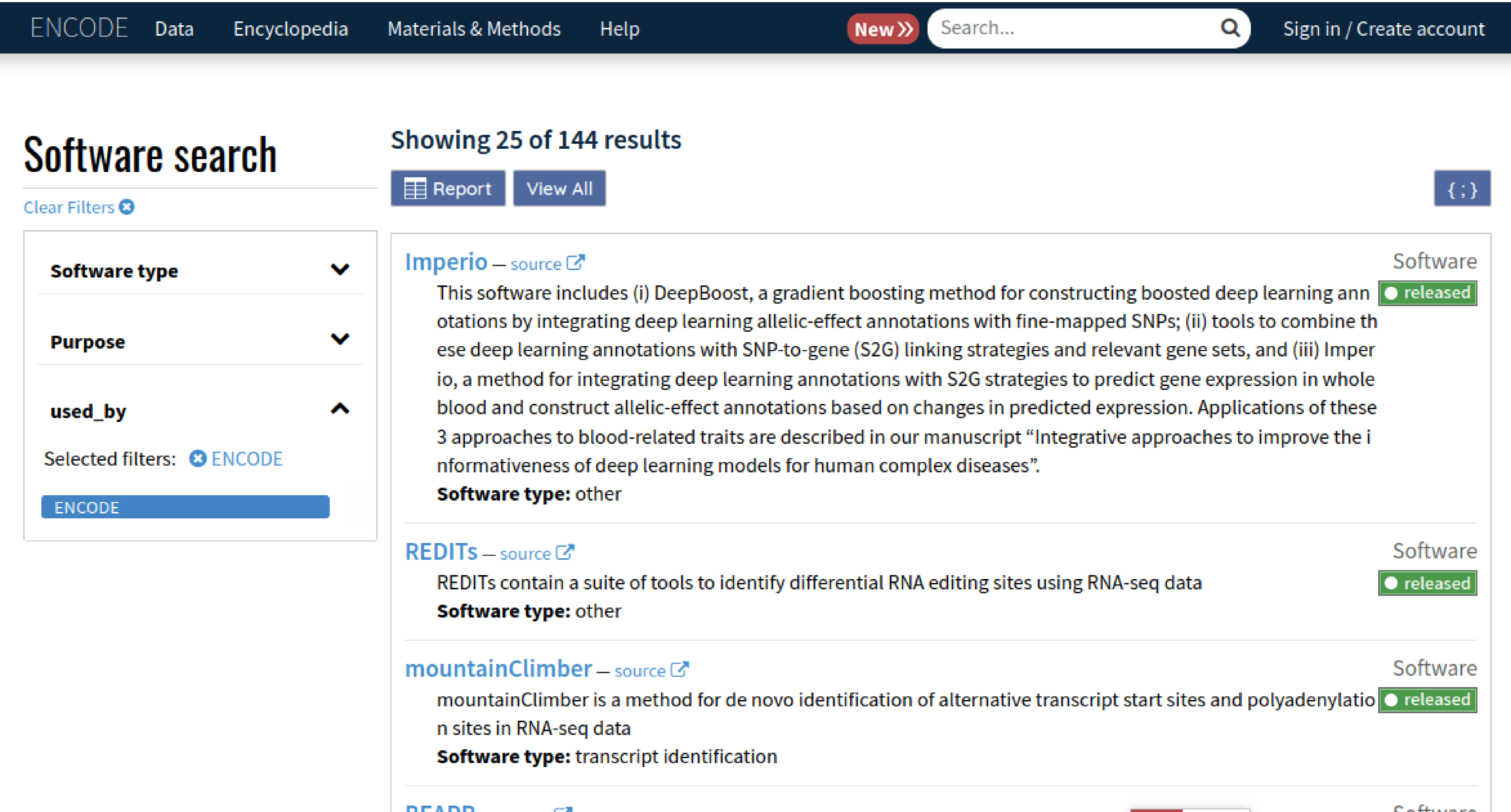
http://encodeproject.org/

Human Microbiome Project



https://hmpdacc.org

Human Microbiome Project



Software search

Clear Filters ✖

Software type ⌄

Purpose ⌄

used_by ⌃

Selected filters: ✖ ENCODE

ENCODE

Showing 25 of 144 results

Imperio — source ↗ Software ● released

This software includes (i) DeepBoost, a gradient boosting method for constructing boosted deep learning annotations by integrating deep learning allelic-effect annotations with fine-mapped SNPs; (ii) tools to combine these deep learning annotations with SNP-to-gene (S2G) linking strategies and relevant gene sets, and (iii) Imperio, a method for integrating deep learning annotations with S2G strategies to predict gene expression in whole blood and construct allelic-effect annotations based on changes in predicted expression. Applications of these 3 approaches to blood-related traits are described in our manuscript “Integrative approaches to improve the informativeness of deep learning models for human complex diseases”.

Software type: other

REDITs — source ↗ Software ● released

REDITs contain a suite of tools to identify differential RNA editing sites using RNA-seq data

Software type: other

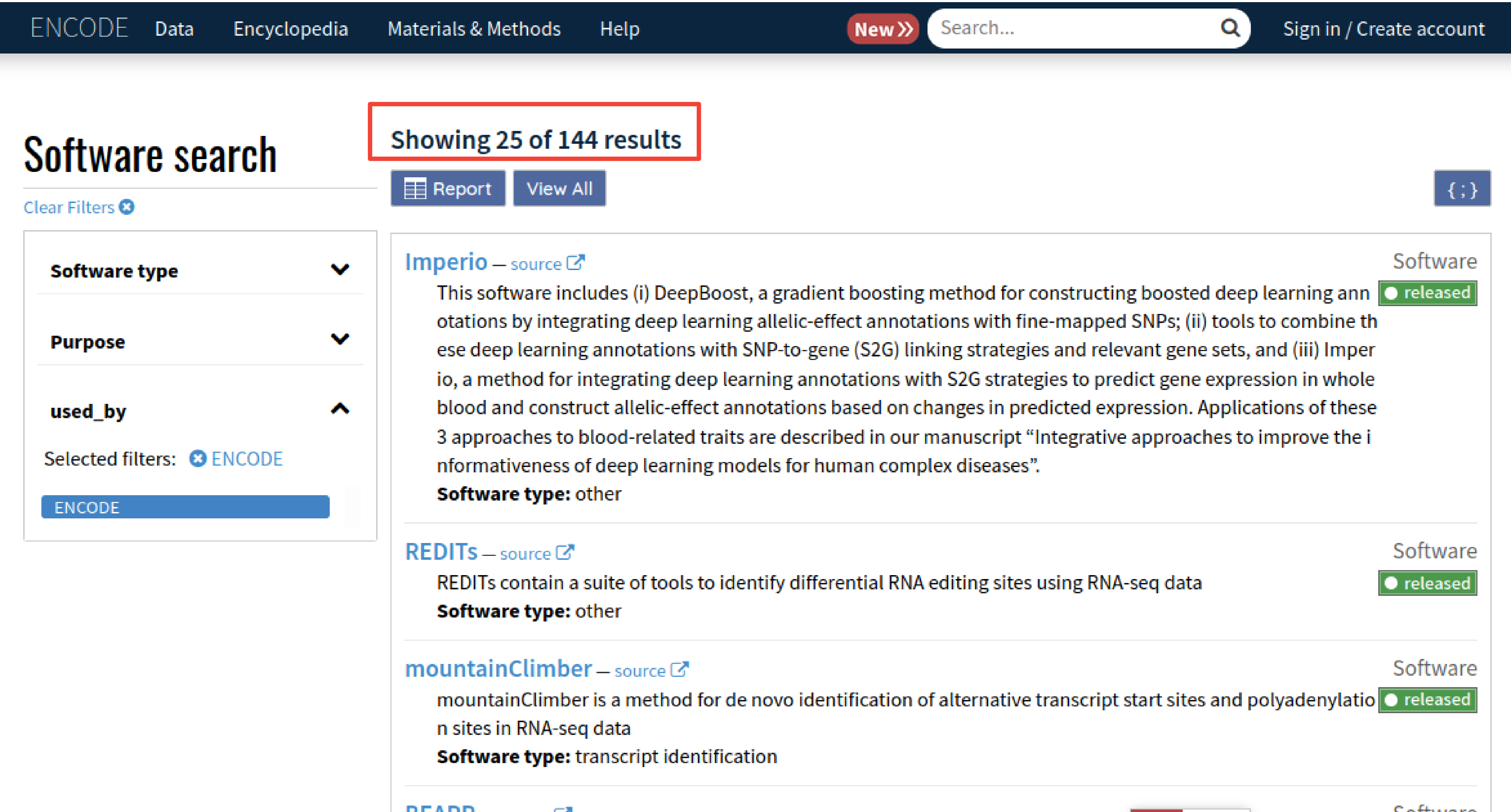
mountainClimber — source ↗ Software ● released

mountainClimber is a method for de novo identification of alternative transcript start sites and polyadenylation sites in RNA-seq data

Software type: transcript identification

Extra big consortia/initiatives

Human Microbiome Project



Software search

Showing 25 of 144 results

Selected filters: ENCODE

ENCODE

Imperio — source

This software includes (i) DeepBoost, a gradient boosting method for constructing boosted deep learning annotations by integrating deep learning allelic-effect annotations with fine-mapped SNPs; (ii) tools to combine these deep learning annotations with SNP-to-gene (S2G) linking strategies and relevant gene sets, and (iii) Imperio, a method for integrating deep learning annotations with S2G strategies to predict gene expression in whole blood and construct allelic-effect annotations based on changes in predicted expression. Applications of these 3 approaches to blood-related traits are described in our manuscript "Integrative approaches to improve the informativeness of deep learning models for human complex diseases".

Software type: other

Software — released

REDITs — source

REDITs contain a suite of tools to identify differential RNA editing sites using RNA-seq data

Software type: other

Software — released

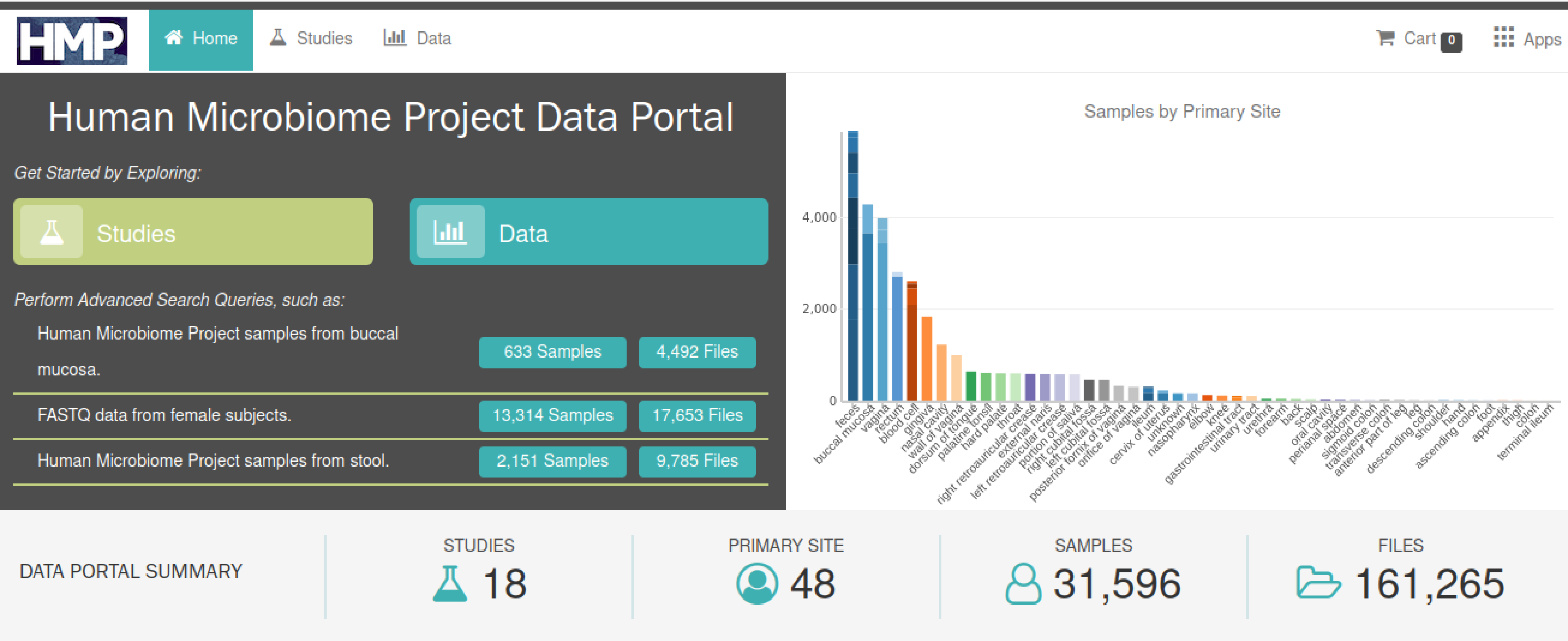
mountainClimber — source

mountainClimber is a method for de novo identification of alternative transcript start sites and polyadenylation sites in RNA-seq data

Software type: transcript identification

Software — released

Human Microbiome Project



Human Microbiome Project Data Portal

Get Started by Exploring:

Perform Advanced Search Queries, such as:

Query	Samples	Files
Human Microbiome Project samples from buccal mucosa.	633 Samples	4,492 Files
FASTQ data from female subjects.	13,314 Samples	17,653 Files
Human Microbiome Project samples from stool.	2,151 Samples	9,785 Files

DATA PORTAL SUMMARY	STUDIES	PRIMARY SITE	SAMPLES	FILES
	18	48	31,596	161,265

https://hmpdacc.org

biobank uk

Enabling scientific discoveries that improve human health

https://www.ukbiobank.ac.uk/

UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from **half a million UK participants**

- NGS data
- Magnetic Resonance Imaging (MRI) data from the brain, heart and abdomen (>60k)
...

Association of alcohol types, coffee and tea intake with mortality

Adiposity, diabetes, lifestyle factors and risk of gastroesophageal reflux disease

Association between household size and COVID-19: A UK Biobank observational study

Association of alcohol types, coffee and tea intake with mortality

Adiposity, diabetes, lifestyle factors and risk of gastroesophageal reflux disease

Association between household size and COVID-19: A UK Biobank observational study

Frontiers in Genetics, November 29th 2022
Association of insomnia and daytime sleepiness with low back pain: A bidirectional mendelian randomization analysis
Peng Shu, Lixian Ji, Zichuan Ping, Zhibo Sun, Wei Liu

Science of The Total Environment, November 1st 2022
Exposure to various ambient air pollutants increases the risk of venous thromboembolism: A cohort study in UK Biobank
J Li et al

Sleep Medicine, October 1st 2022
Gender-specific association between obstructive sleep apnea and cognitive impairment among adults
K Qiu et al

Ecotoxicology and Environmental Safety, September 1st 2022
Long-term exposure to air pollution and risk of incident inflammatory bowel disease among middle and old aged adults
F Li et al

Search Publications:

Enter search term

Year

2022 (584) 2021 (931) 2020 (664)
2019 (429) 2018 (310) 2017 (173)
2016 (92) 2015 (30) 2014 (16)
2013 (6) 2012 (1) 2008 (1)

Enabling scientific discoveries that improve human health

https://www.ukbiobank.ac.uk/

Description	Tier 1	Tier 2	Tier 3
Core data • Questionnaires and physical measurements • Health Outcome phenotypes • Linked health data • Web-based questionnaires	✓	✓	✓
Assay data and enhanced measures • Biochemical and haematological assays • Other platform based assays • Measured and imputed genotypes • Other enhancements		✓	✓
Very large datasets • Imaging data * • Other large-scale assay data • Whole genome sequence data • Whole exome sequence data			✓ Via platform only
First 3 years - access to data with scheduled updates	£3,000	£6,000 (+£3,000 vs Tier 1)	£9,000 (+£3,000 vs Tier 2)
Additional Institution fee - each additional institution added to an application	£1,000 for first 3 years (£500 p.a. extension)		
Low & Middle Income Countries and Student Researchers ** - access to all datasets via the Research Analysis Platform (full fees apply to downloaded data)	£500 for first 3 years (£175 p.a. extension)		

Scientific Competitions

Scientific Competitions

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Collective experiment for blind RNA structure prediction (RNA-Puzzles)

Critical Assessment of Prediction of Interactions (CAPRI)

Critical Assessment of Functional Annotation (CAFA)

Critical Assessment of Microarray Data Analysis (CAMDA)

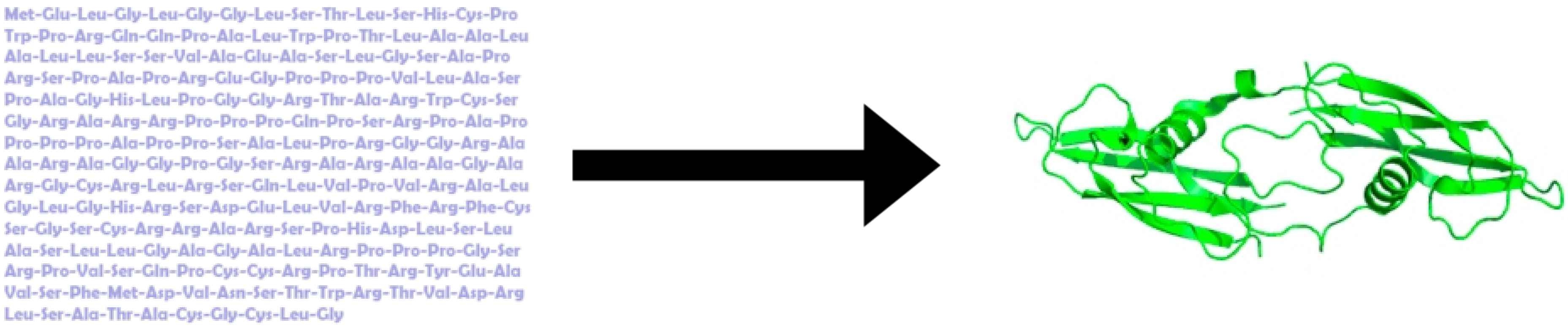
Genome Annotation Assessment Project (GASP)

Bone X-Ray Deep Learning Competition

LUng Nodule Analysis 2016

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994



Met-Glu-Leu-Gly-Leu-Gly-Gly-Leu-Ser-Thr-Leu-Ser-His-Cys-Pro
Trp-Pro-Arg-Gln-Gln-Pro-Ala-Leu-Trp-Pro-Thr-Leu-Ala-Ala-Leu
Ala-Leu-Leu-Ser-Ser-Val-Ala-Glu-Ala-Ser-Leu-Gly-Ser-Ala-Pro
Arg-Ser-Pro-Ala-Pro-Arg-Glu-Gly-Pro-Pro-Pro-Val-Leu-Ala-Ser
Pro-Ala-Gly-His-Leu-Pro-Gly-Gly-Arg-Thr-Ala-Arg-Trp-Cys-Ser
Gly-Arg-Ala-Arg-Arg-Pro-Pro-Pro-Gln-Pro-Ser-Arg-Pro-Ala-Pro
Pro-Pro-Pro-Ala-Pro-Pro-Ser-Ala-Leu-Pro-Arg-Gly-Gly-Arg-Ala
Ala-Arg-Ala-Gly-Gly-Pro-Gly-Ser-Arg-Ala-Arg-Ala-Ala-Gly-Ala
Arg-Gly-Cys-Arg-Leu-Arg-Ser-Gln-Leu-Val-Pro-Val-Arg-Ala-Leu
Gly-Leu-Gly-His-Arg-Ser-Asp-Glu-Leu-Val-Arg-Phe-Arg-Phe-Cys
Ser-Gly-Ser-Cys-Arg-Arg-Ala-Arg-Ser-Pro-His-Asp-Leu-Ser-Leu
Ala-Ser-Leu-Leu-Gly-Ala-Gly-Ala-Leu-Arg-Pro-Pro-Pro-Gly-Ser
Arg-Pro-Val-Ser-Gln-Pro-Cys-Cys-Arg-Pro-Thr-Arg-Tyr-Glu-Ala
Val-Ser-Phe-Met-Asp-Val-Asn-Ser-Thr-Trp-Arg-Thr-Val-Asp-Arg
Leu-Ser-Ala-Thr-Ala-Cys-Gly-Cys-Leu-Gly

Anfinsen

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994

Every second spring-summer around 100 targets* are released

Targets – protein sequences for which the structure has been solved recently (not Available publicly e.g. not in PDB)

Blind benchmark

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994

Every second spring-summer around 100 targets* are released

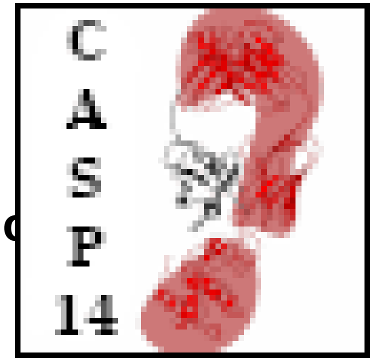
Targets – protein sequences for which the structure has been solved recently (not Available publicly e.g. not in PDB)

Blind benchmark

Categories: servers (72 h) and humans (3 weeks)
homology modeling & Free Modeling

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

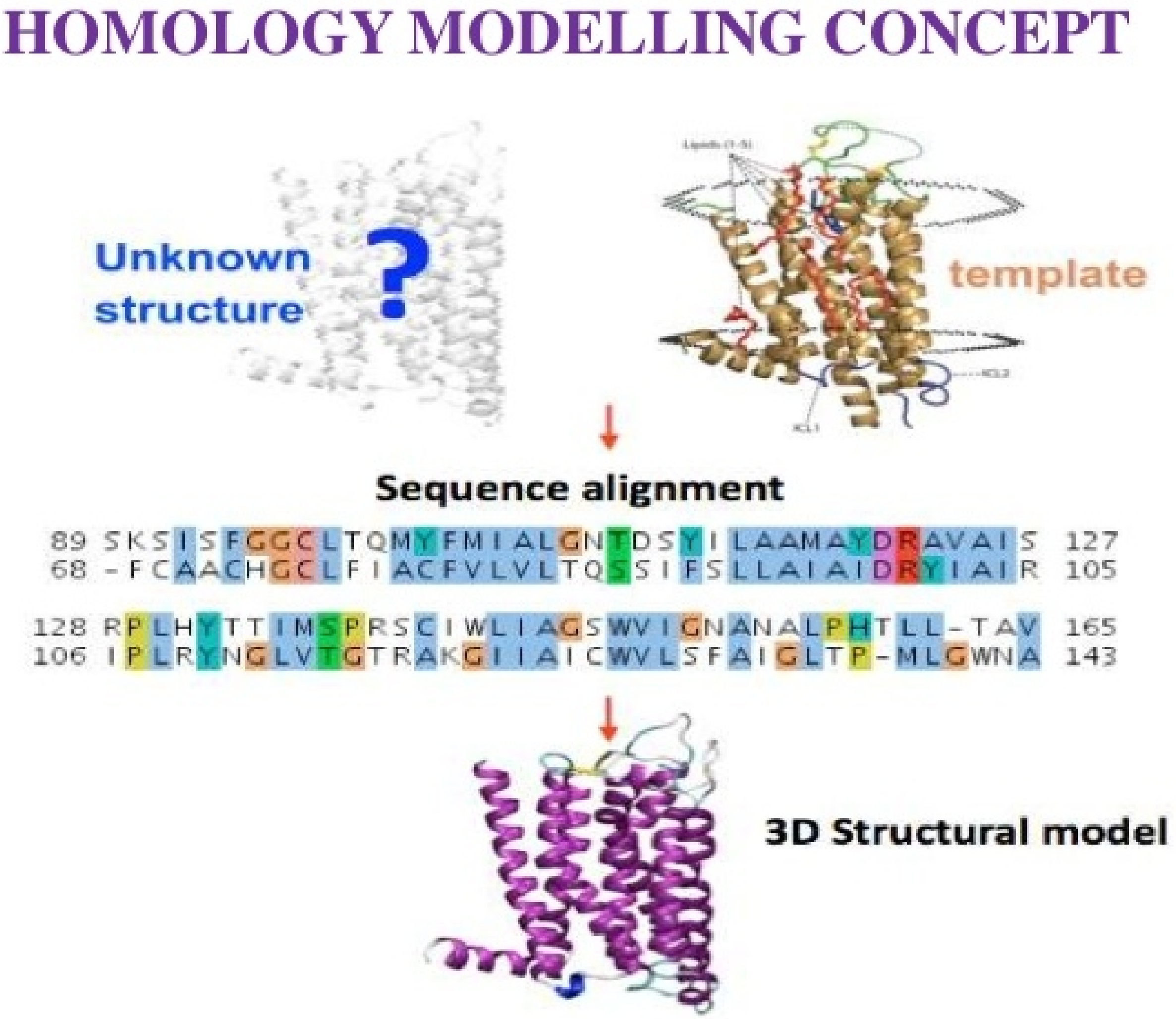
Evaluation of the results is carried out in the following prediction catego



- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only;
  a separate experiment - CAPRI—carries on this subject)
- residue-residue contact prediction (starting CASP4)
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- high-accuracy template-based prediction (starting CASP7)

Tertiary structure prediction (all CASPs)

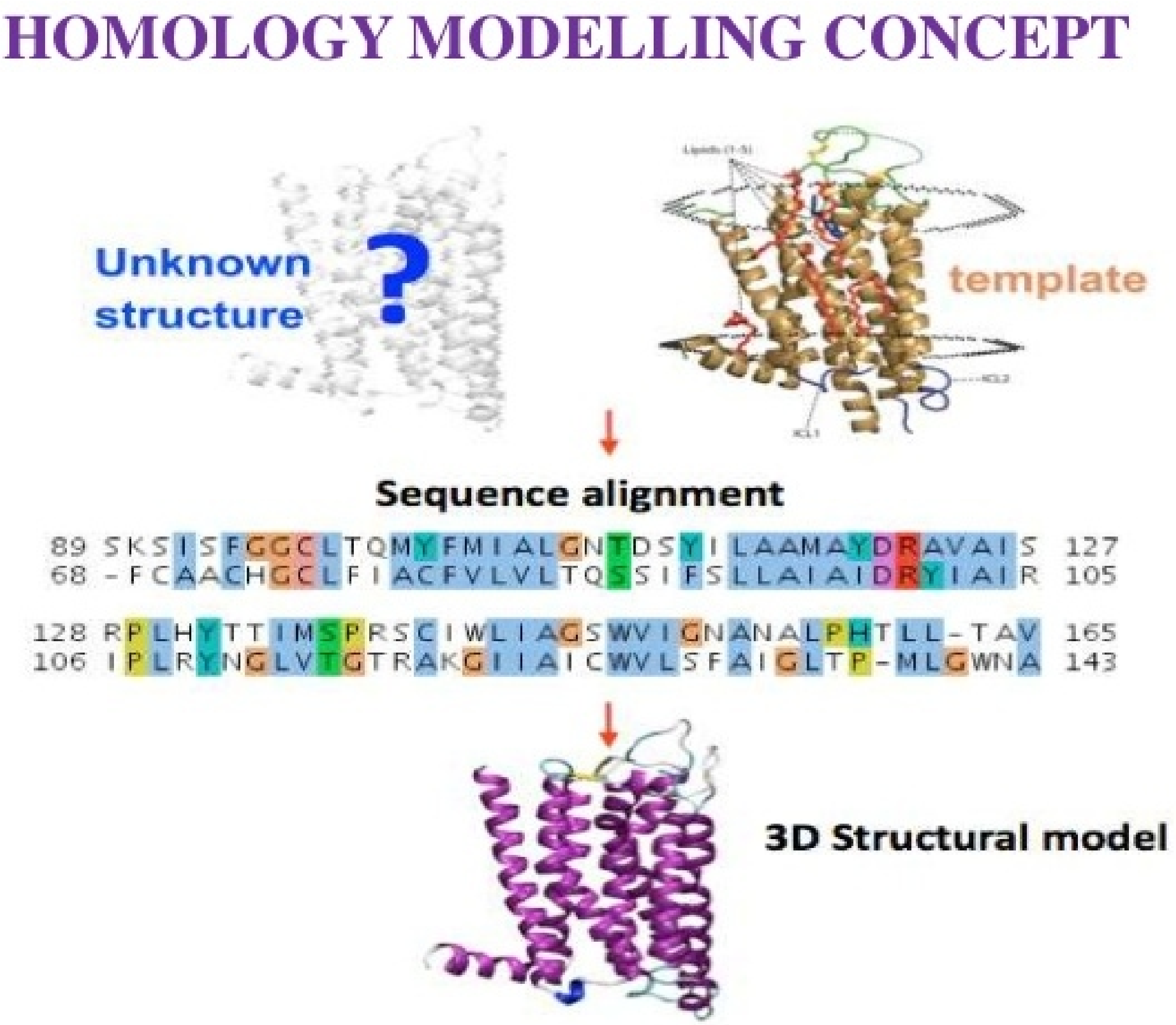
CASP 14

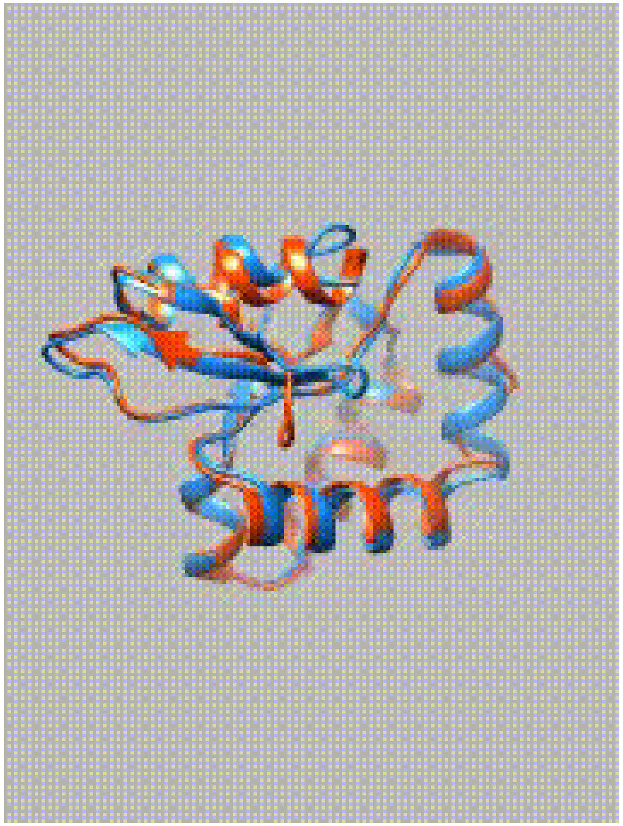


```
 89 SKSISFGGCLTQMYFMIALGNTDSYILAAMAYDRAVAIS 127
 68 -FCAACHGCLFIACFVLVLTQSSIFSLLAIAIDRYIAIR 105

128 RPLHYTTIMSPRSCIWLIAGSWVIGNANALPHTLL-TAV 165
106 IPLRYNGLVTGTRAKGIIAICWVLSFAIGLTP-MLGWNA 143
```

Tertiary structure prediction (all CASPs)



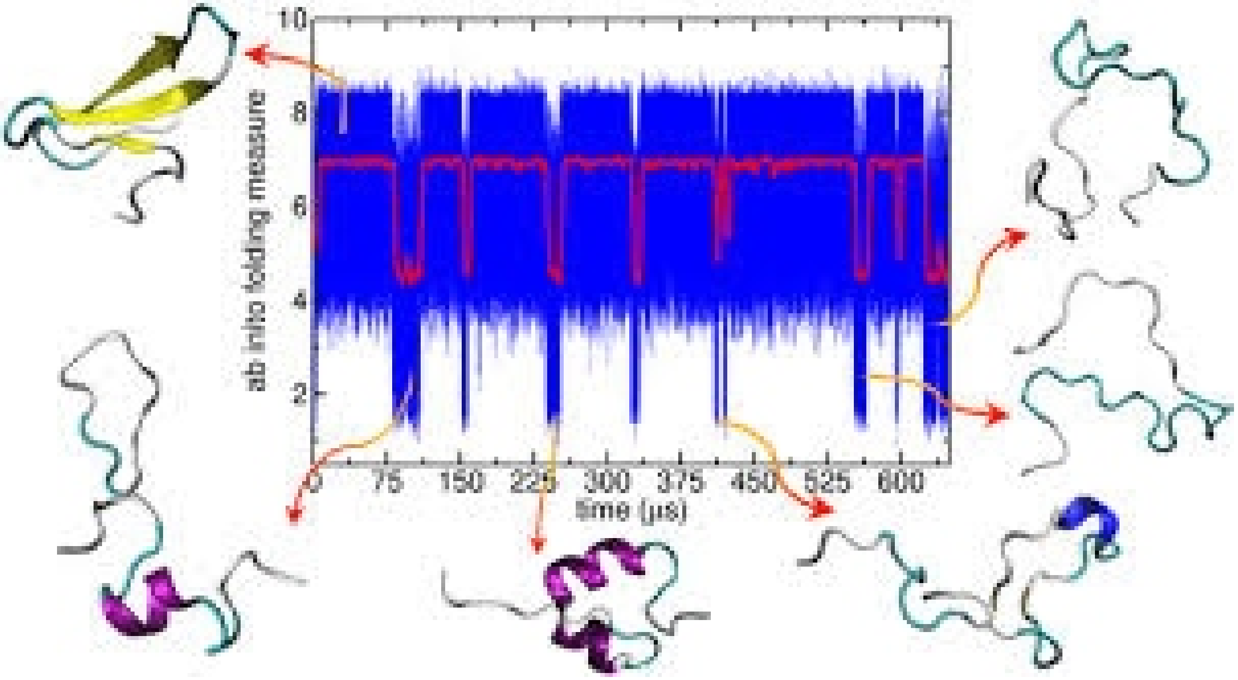


target T0868-D1 (orange)
model 330_2 (blue): GDT_TS=87
best template: 2cw6 (seq.id= 4.2%)

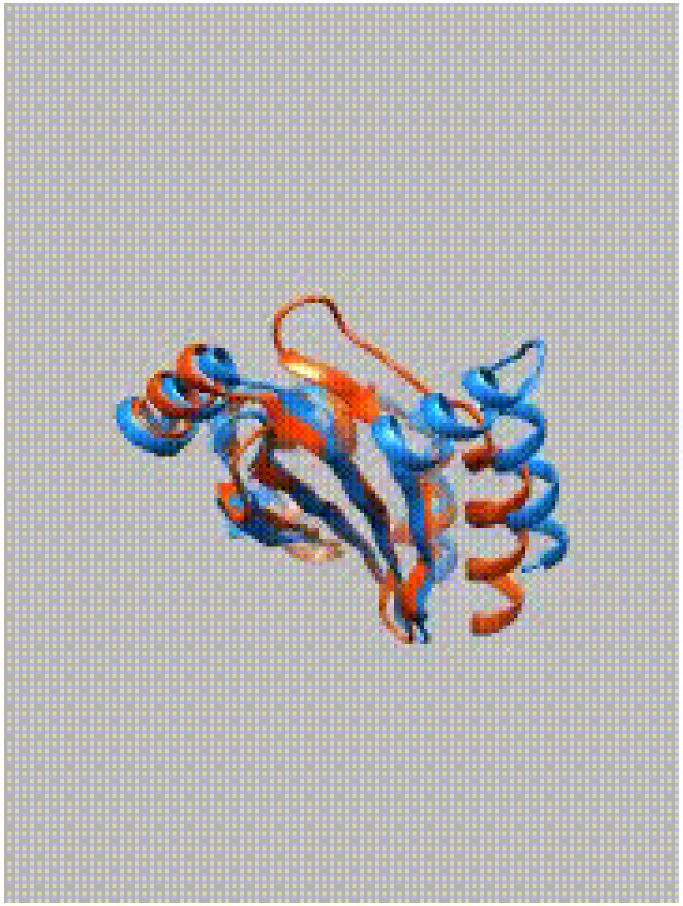
Scientific Competitions

Tertiary structure prediction (all CASPs)

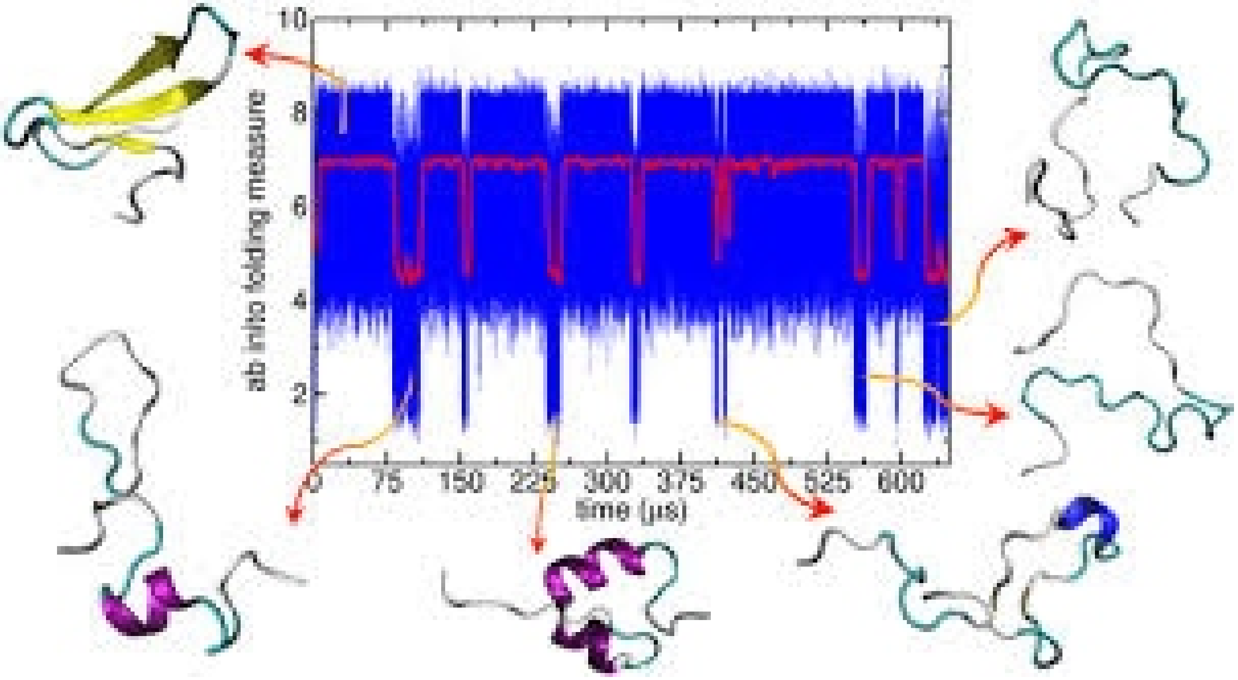
CASP 14



Tertiary structure prediction (all CASPs)



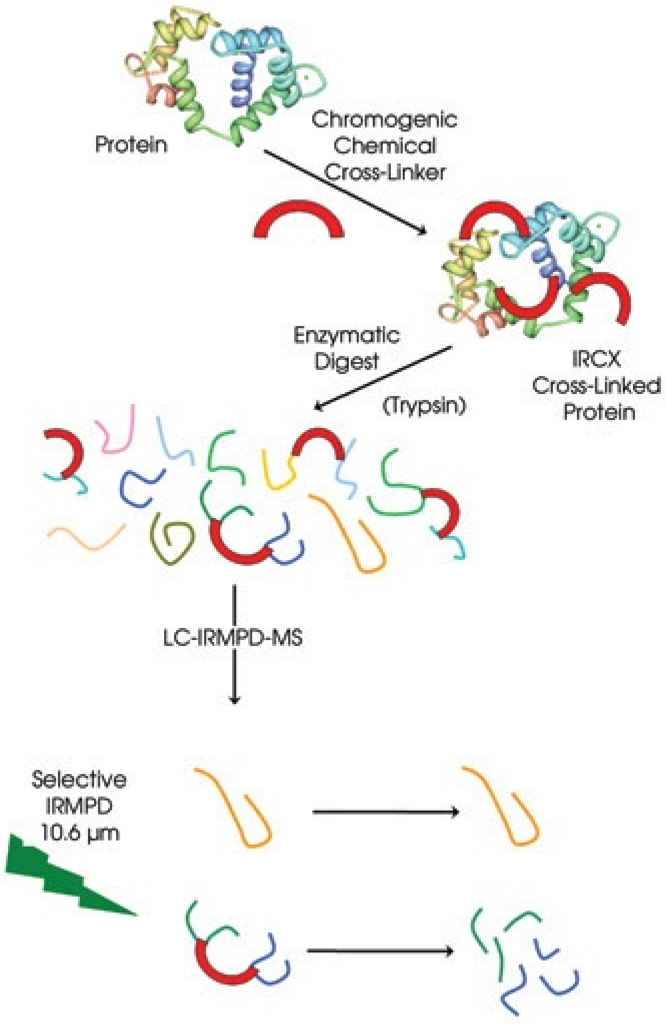
CASP9: T0581-D1
model 170_1: GDT_TS=71

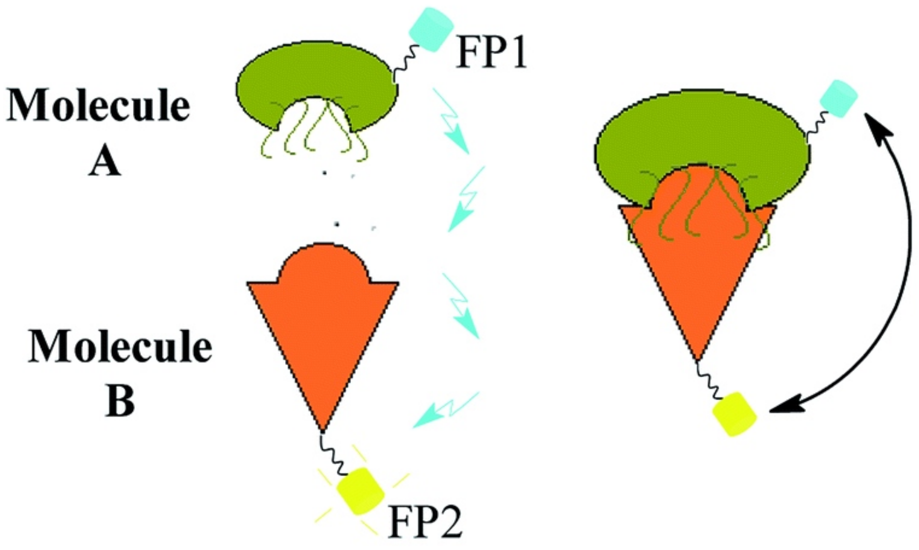


Scientific Competitions

CASP 14

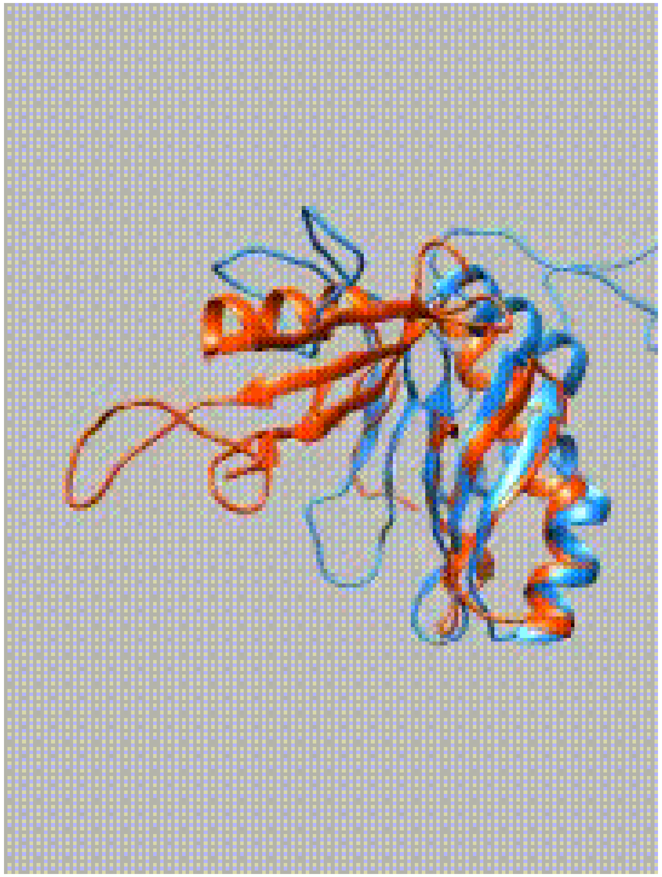
Data-assisted or hybrid modeling, in which low-resolution experimental data are combined with computational methods, is becoming increasing important for a range of experimental data, including NMR, chemical cross-linking and surface labeling, X-ray and neutron scattering, electron microscopy and FRET.



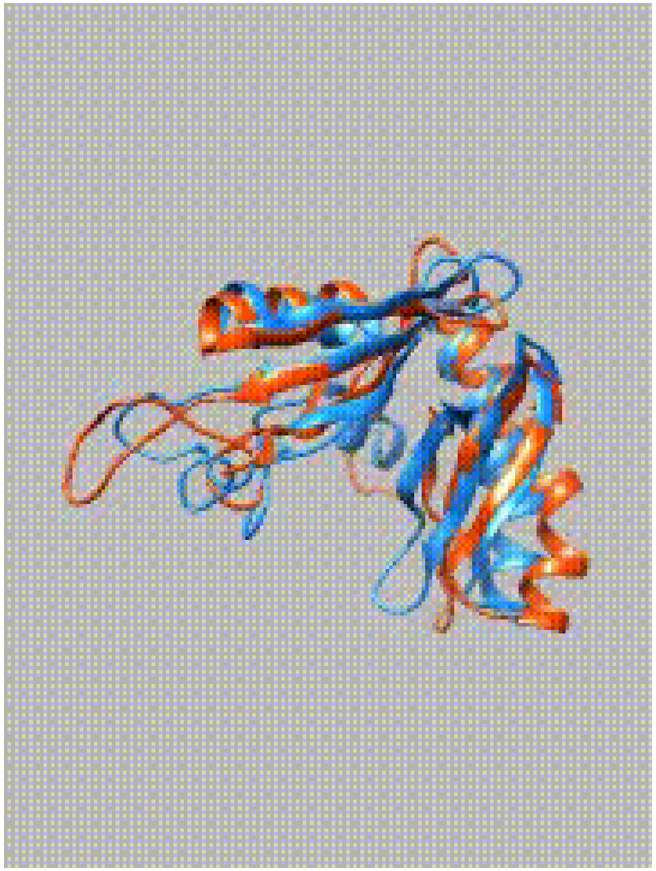


Scientific Competitions

Data-assisted or hybrid modeling, in which low-resolution experimental data are combined with computational methods, is becoming increasing important for a range of experimental data, including NMR, chemical cross-linking and surface labeling, X-ray and neutron scattering, electron microscopy and FRET

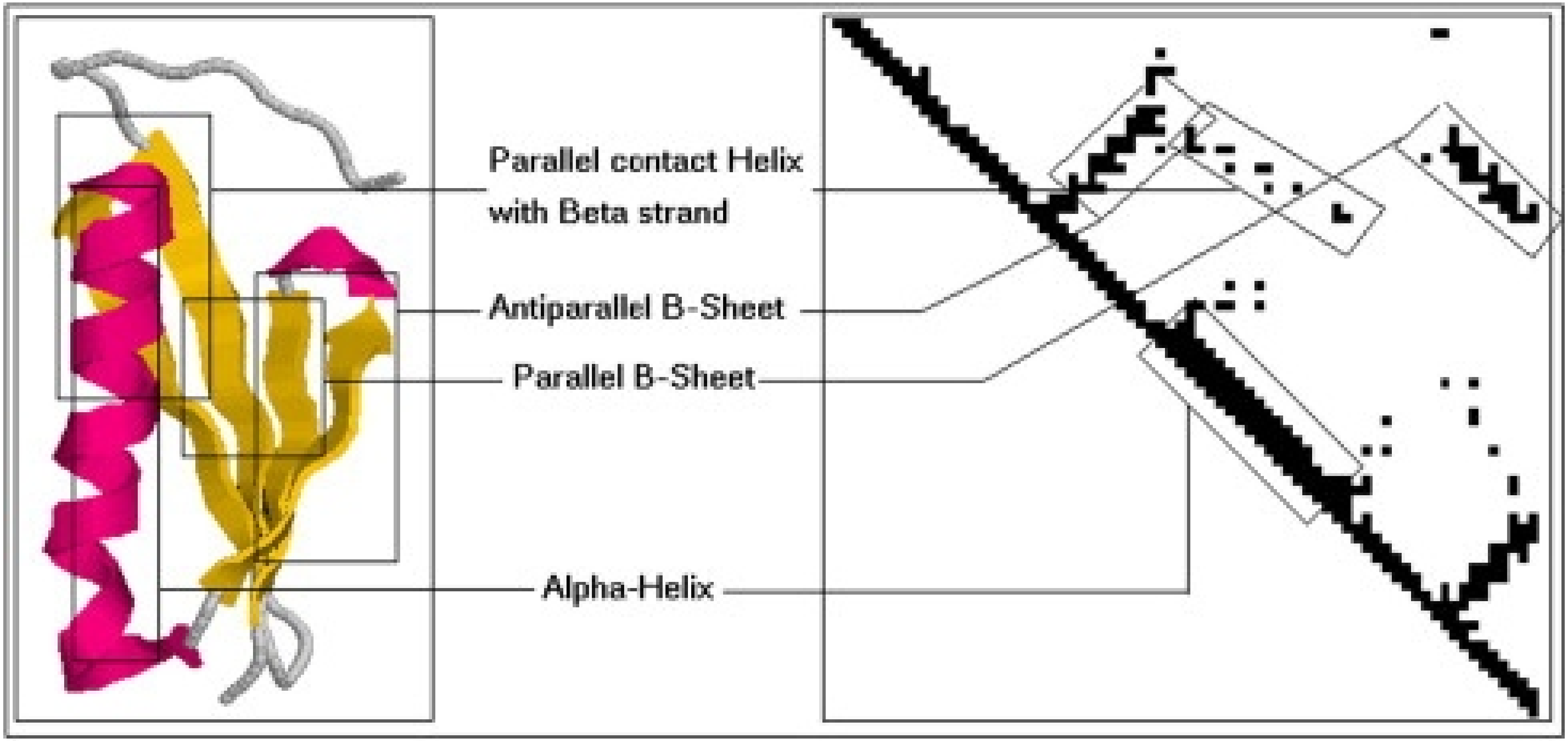


without restrains

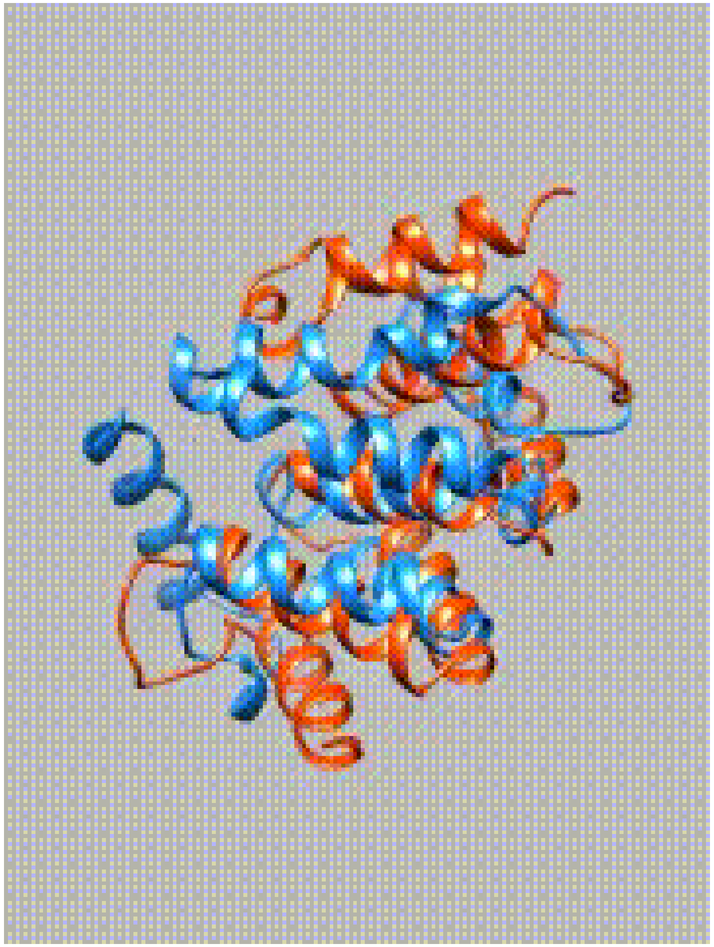


with restrains

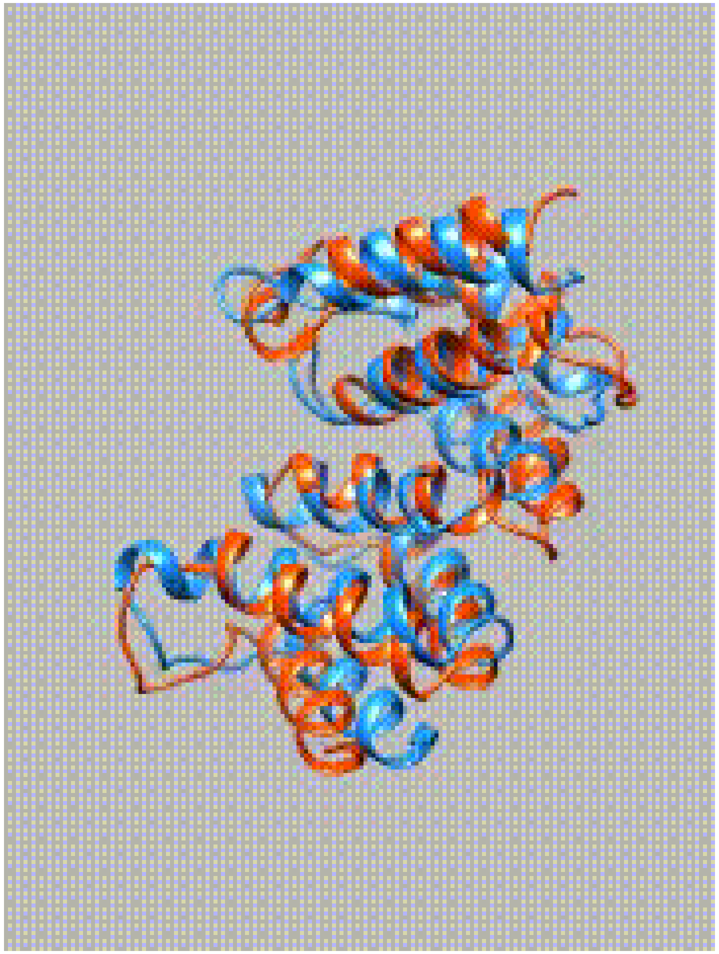
Residue-residue contact prediction



Residue-residue contact prediction

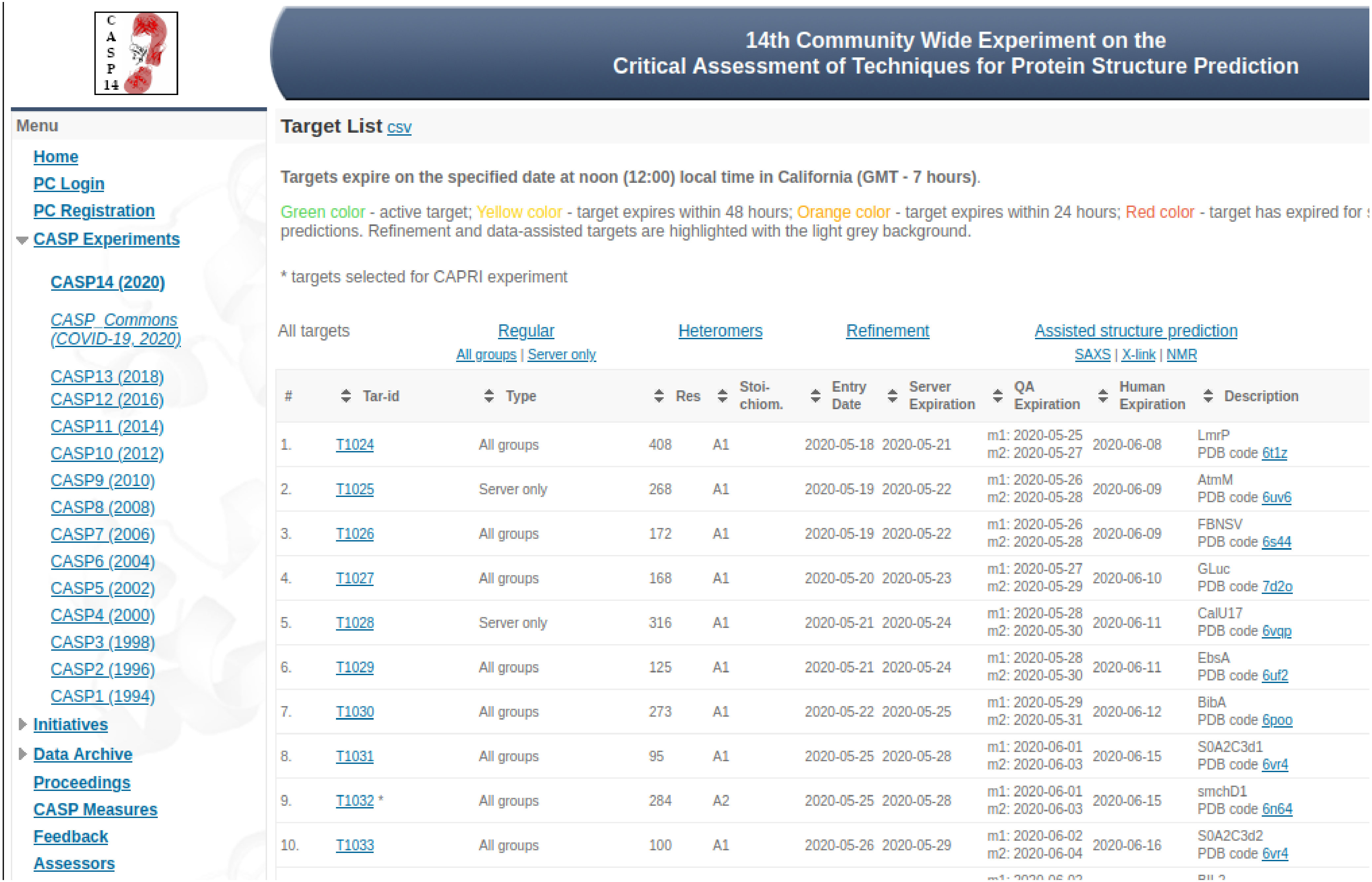


without restrains



with restrains

Scientific Competitions



14th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

Target List csv

Targets expire on the specified date at noon (12:00) local time in California (GMT - 7 hours).

Green color - active target; Yellow color - target expires within 48 hours; Orange color - target expires within 24 hours; Red color - target has expired for predictions. Refinement and data-assisted targets are highlighted with the light grey background.

* targets selected for CAPRI experiment

#	Tar-id	Type	Res	Stoi-chiom.	Entry Date	Server Expiration	QA Expiration	Human Expiration	Description
1.	T1024	All groups	408	A1	2020-05-18	2020-05-21	m1: 2020-05-25 m2: 2020-05-27	2020-06-08	LmrP PDB code 6t1z
2.	T1025	Server only	268	A1	2020-05-19	2020-05-22	m1: 2020-05-26 m2: 2020-05-28	2020-06-09	AtmM PDB code 6uv6
3.	T1026	All groups	172	A1	2020-05-19	2020-05-22	m1: 2020-05-26 m2: 2020-05-28	2020-06-09	FBNSV PDB code 6s44
4.	T1027	All groups	168	A1	2020-05-20	2020-05-23	m1: 2020-05-27 m2: 2020-05-29	2020-06-10	GLuc PDB code 7d2o
5.	T1028	Server only	316	A1	2020-05-21	2020-05-24	m1: 2020-05-28 m2: 2020-05-30	2020-06-11	CalU17 PDB code 6vqp
6.	T1029	All groups	125	A1	2020-05-21	2020-05-24	m1: 2020-05-28 m2: 2020-05-30	2020-06-11	EbsA PDB code 6uf2
7.	T1030	All groups	273	A1	2020-05-22	2020-05-25	m1: 2020-05-29 m2: 2020-05-31	2020-06-12	BibA PDB code 6poo
8.	T1031	All groups	95	A1	2020-05-25	2020-05-28	m1: 2020-06-01 m2: 2020-06-03	2020-06-15	S0A2C3d1 PDB code 6vr4
9.	T1032 *	All groups	284	A2	2020-05-25	2020-05-28	m1: 2020-06-01 m2: 2020-06-03	2020-06-15	smchD1 PDB code 6n64
10.	T1033	All groups	100	A1	2020-05-26	2020-05-29	m1: 2020-06-02 m2: 2020-06-04	2020-06-16	S0A2C3d2 PDB code 6vr4

Scientific Competitions

ADP

14th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

Groups List

⬍Group Name	⬍Group #	⬍Type	Predictors	Submitted predictions
191227	061	Human	Xi Cheng wenjun he Denghui Liu Dingyan Wang Chi Xu Meng Xu lei zhang Mingyue Zheng	TS(regular targets): 390 models for 78 targets RR(regular targets): 78 models for 78 targets
3DCNN_prof	074	Human	Takashi Ishida	QA(regular targets): 166 models for 83 targets
3D-JIGSAW-SwarmLoop	169	Server	Paul Bates Raphael Chaleil	TS(regular targets): 83 models for 83 targets
A2I2Prot	431	Human	Thin Nguyen Tri Nguyen Minh	RR(regular targets): 76 models for 76 targets
ACOMPMOD	063	Server	Ricardo Nunez Miguel	TS(regular targets): 410 models for 83 targets
AILON	192	Human	kyungmin cho Hyoje Cho Kyeongtak Han Wonjun Lee	TS(regular targets): 402 models for 81 targets TS(refinement targets): 247 models for 50 targets RR(regular targets): 78 models for 78 targets
AIR	100	Human	Hongbin shen Di wang Chengpeng Zhou	TS(refinement targets): 250 models for 50 targets
AlphaFold2	427	Human	Russ Bates Alex Bridgland Timothy Green John Jumper Kathryn Tunyasuvunakool Augustin Zidek	TS(regular targets): 390 models for 78 targets
AmoebaContact	286	Server-E	Yaoguang Xing Yunxin Xu	RR(regular targets): 83 models for 83 targets
angleQA	391	Server-E	Jianzhao Gao Boling Wang	QA(regular targets): 166 models for 83 targets

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Evaluation of the results is carried out in the following prediction catego

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only; a separate experiment - CAPRI—carries on this subject)
- residue-residue contact prediction (starting CASP4)
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- high-accuracy template-based prediction (starting CASP7)

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Evaluation of the results is carried out in the following prediction catego

- **tertiary structure prediction (all CASPs)**
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only;
  a separate experiment - CAPRI—carries on this subject)
- re due contact prediction (st ASP4)
- d regions prediction (startin)
- d ndary prediction (CASP6-
- fu diction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- late-based pr g CASP7)

Janusz Bujnicki

Andrzej Koliński

International Institute of Molecular and Cell Biology in Warsaw

Wydział Chemii
Uniwersytet Warszawski

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Evaluation of the results is carried out in the following prediction catego

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after C
- prediction of structure complexes (CASP2 only;
  a separate experiment - CAPRI—carries on this su
- **residue-residue contact prediction**
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP
- model refinement (starting CASP7)
- high-accuracy template-based prediction

Michał Piętal

International Institute of Molecular and Cell Biology in Warsaw

POLITECHNIKA RZESZOWSKA
im. IGNACEGO ŁUKASIEWICZA

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Evaluation of the results is carried out in the following prediction catego

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after
- prediction of structure complexes (CASP2 onl
  a separate experiment - CAPRI—carries on this
- residue-residue contact prediction
- **disordered regions prediction** (starting CASP5
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP
- model refinement (starting CASP7)
- high-accuracy template-based prediction (

Łukasz P. Kozłowski

International Institute of Molecular and Cell Biology in Warsaw

WYDZIAŁ MATEMATYKI, INFORMATYKI I MECHANIKI

C
A
S
P
14

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Evaluation of the results is carried out in the following prediction catego

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped a
- prediction of structure complexes (CASP2
  a separate experiment - CAPRI—carries on t
- residue-residue contact prediction
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6–CASP8)
- function prediction (starting CASP6)
- **model quality assessment**
- model refinement (starting CASP7)
- high-accuracy template-based predi

Marcin Pawłowski

International Institute of Molecular and Cell Biology in Warsaw

precisionlife
driving precision medicine

Scientific Competitions

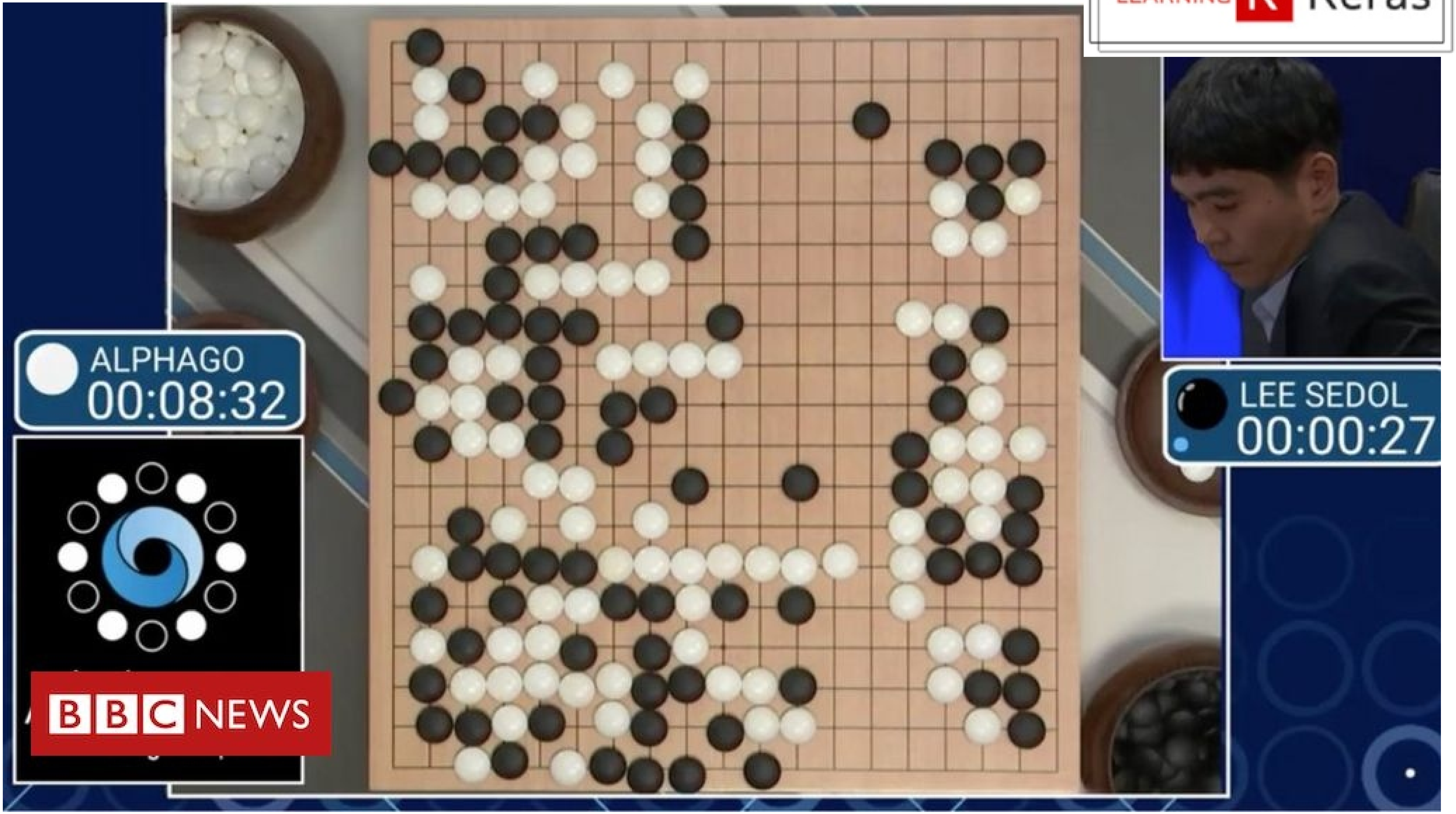
Scientific Competitions

AlphaGo

DEEP LEARNING

TensorFlow

Keras



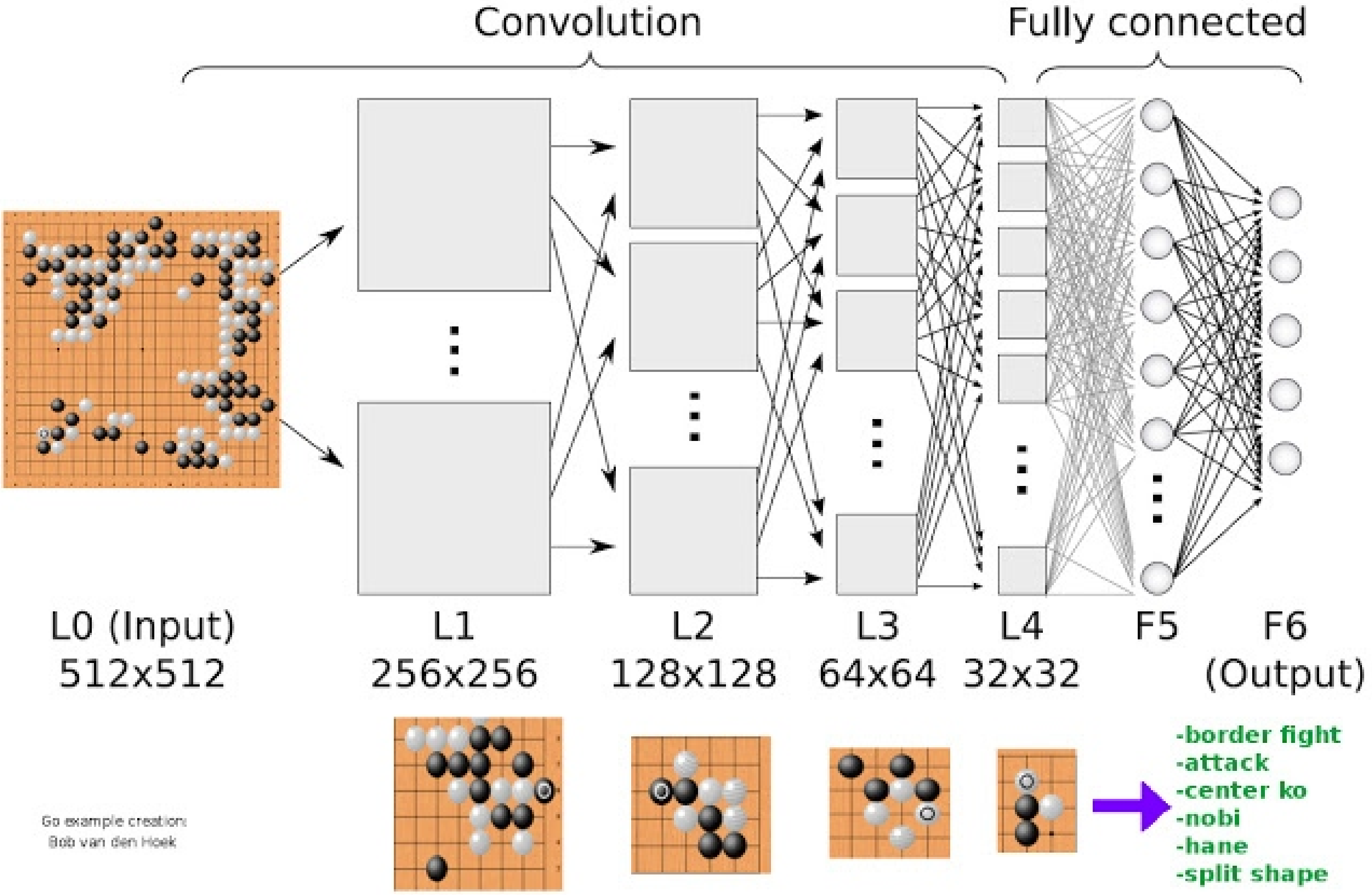
Scientific Competitions

AlphaGo

DEEP LEARNING

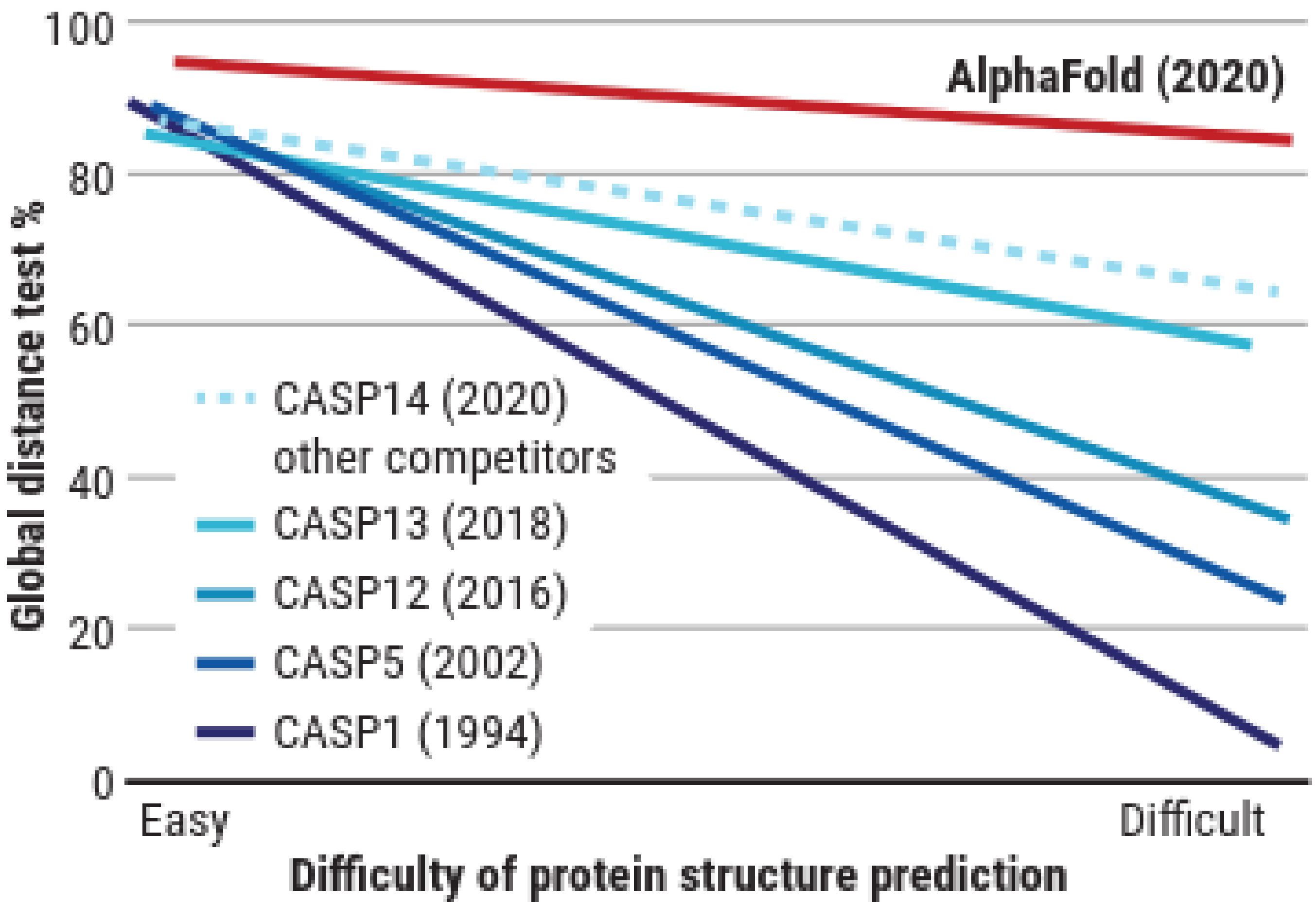
TensorFlow

Keras



Go example creation:
Bob van den Hoek

AlphaFold

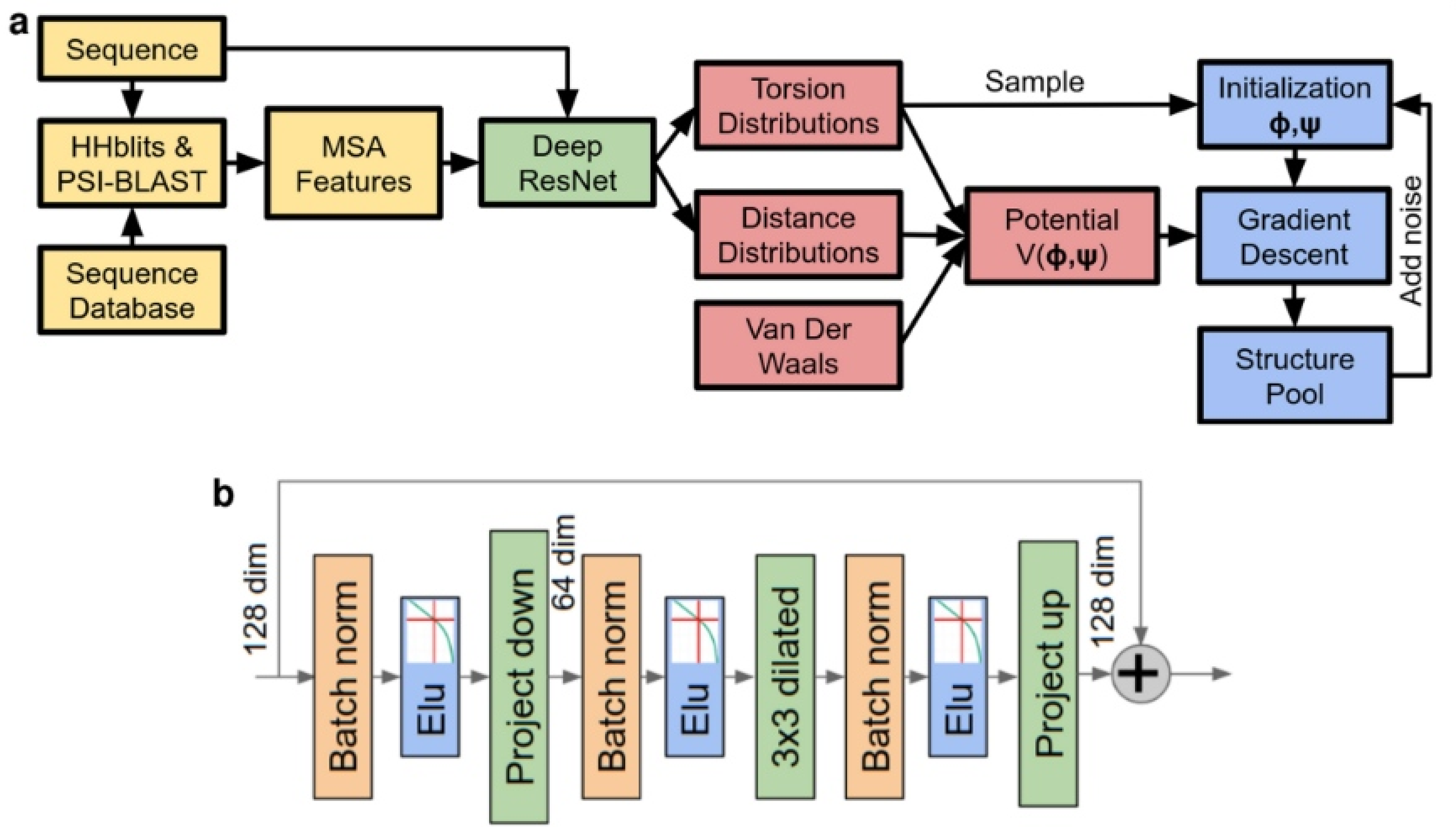

100
80
60
40
20
0
Global distance test %
AlphaFold (2020)
CASP14 (2020) other competitors
CASP13 (2018)
CASP12 (2016)
CASP5 (2002)
CASP1 (1994)
Easy
Difficult
Difficulty of protein structure prediction

Presentations & Videos from CASP15

https://predictioncenter.org/casp15/doc/presentations/

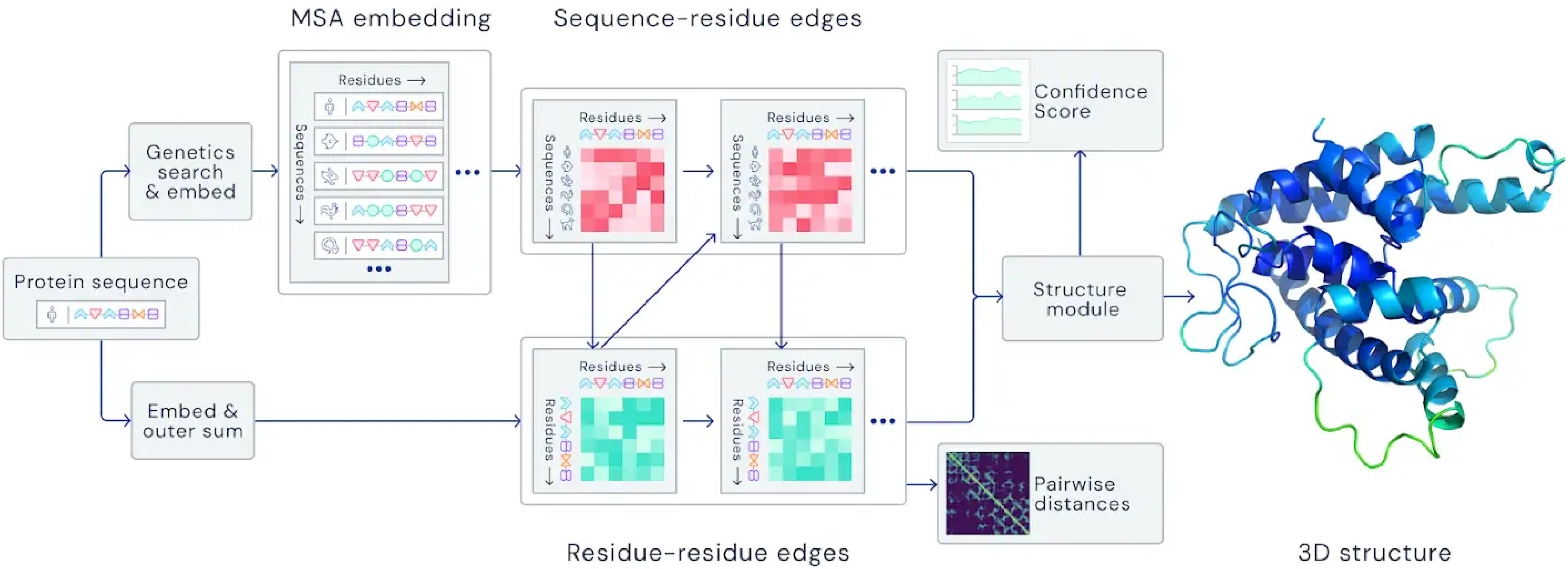
https://www.youtube.com/@CASP-Prediction-Center/videos

AlphaFold

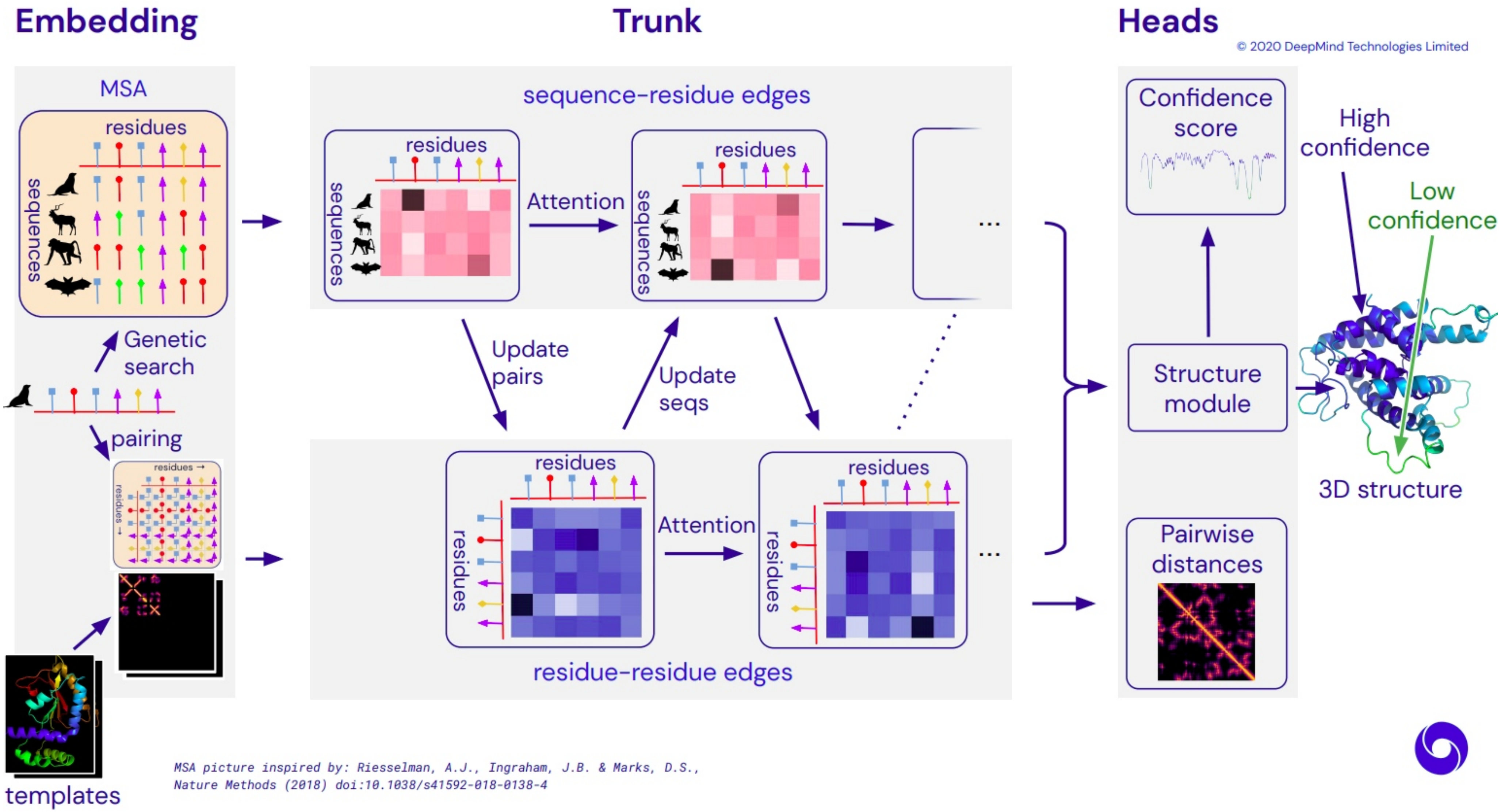


Extended Data Fig. 1 | Schematics of the folding system and neural network. **a**, The overall folding system. Feature extraction stages (constructing the MSA using sequence database search and computing MSA-based features) are shown in yellow; the structure-prediction neural network in green; potential construction in red; and structure realization in blue. **b**, The layers used in one block of the deep residual convolutional network. The dilated convolution is applied to activations of reduced dimension. The output of the block is added to the representation from the previous layer. The bypass connections of the residual network enable gradients to pass back through the network undiminished, permitting the training of very deep networks.

AlphaFold


MSA embedding
Sequence-residue edges
Genetics search & embed
Residues
Sequences
Protein sequence
Embed & outer sum
Confidence Score
Structure module
Pairwise distances
Residue-residue edges
3D structure

AlphaFold2

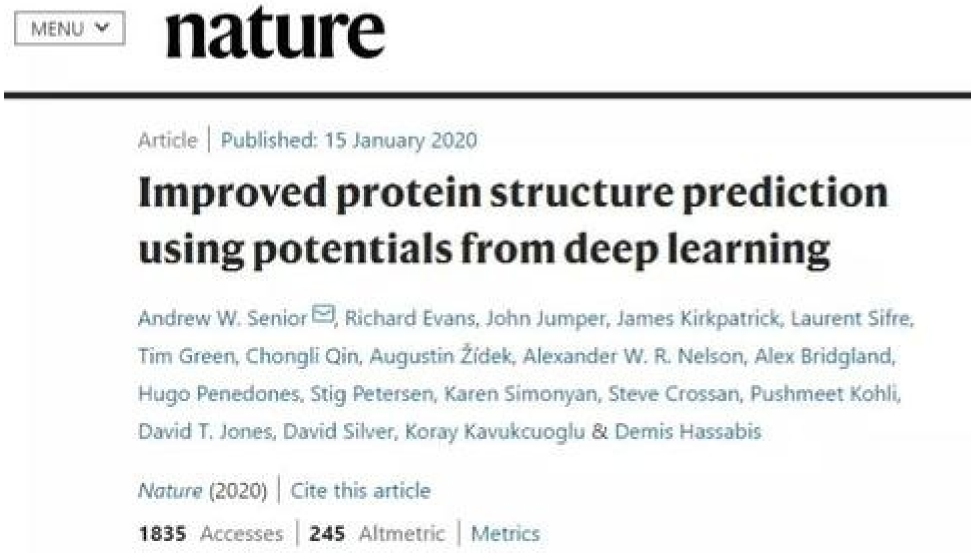


© 2020 DeepMind Technologies Limited

MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

https://predictioncenter.org/casp14/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf

AlphaFold



Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

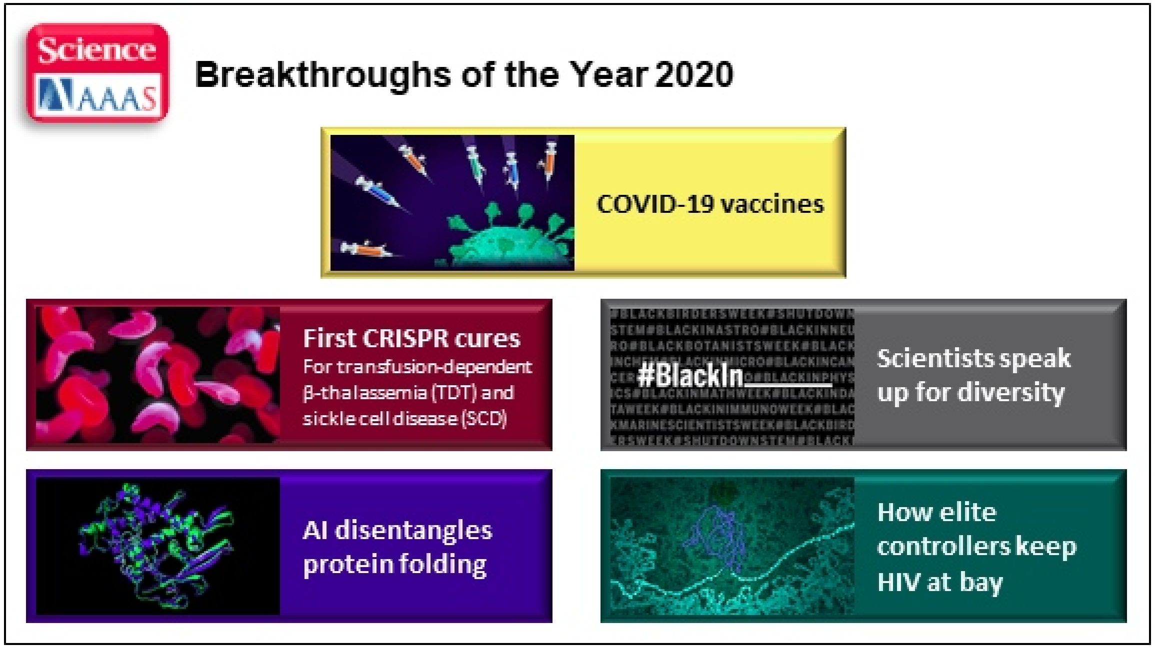
Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis

Nature (2020) | Cite this article

1835 Accesses | **245** Altmetric | Metrics

Scientific Competitions

AlphaFold



Science AAAS

Breakthroughs of the Year 2020

COVID-19 vaccines

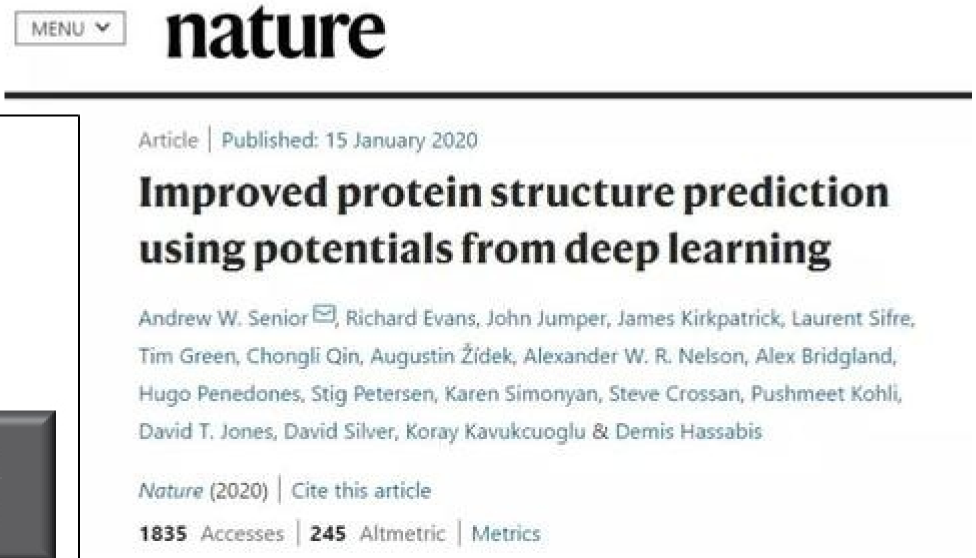
First CRISPR cures
For transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)

#BlackIn

Scientists speak up for diversity

AI disentangles protein folding

How elite controllers keep HIV at bay



nature

Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

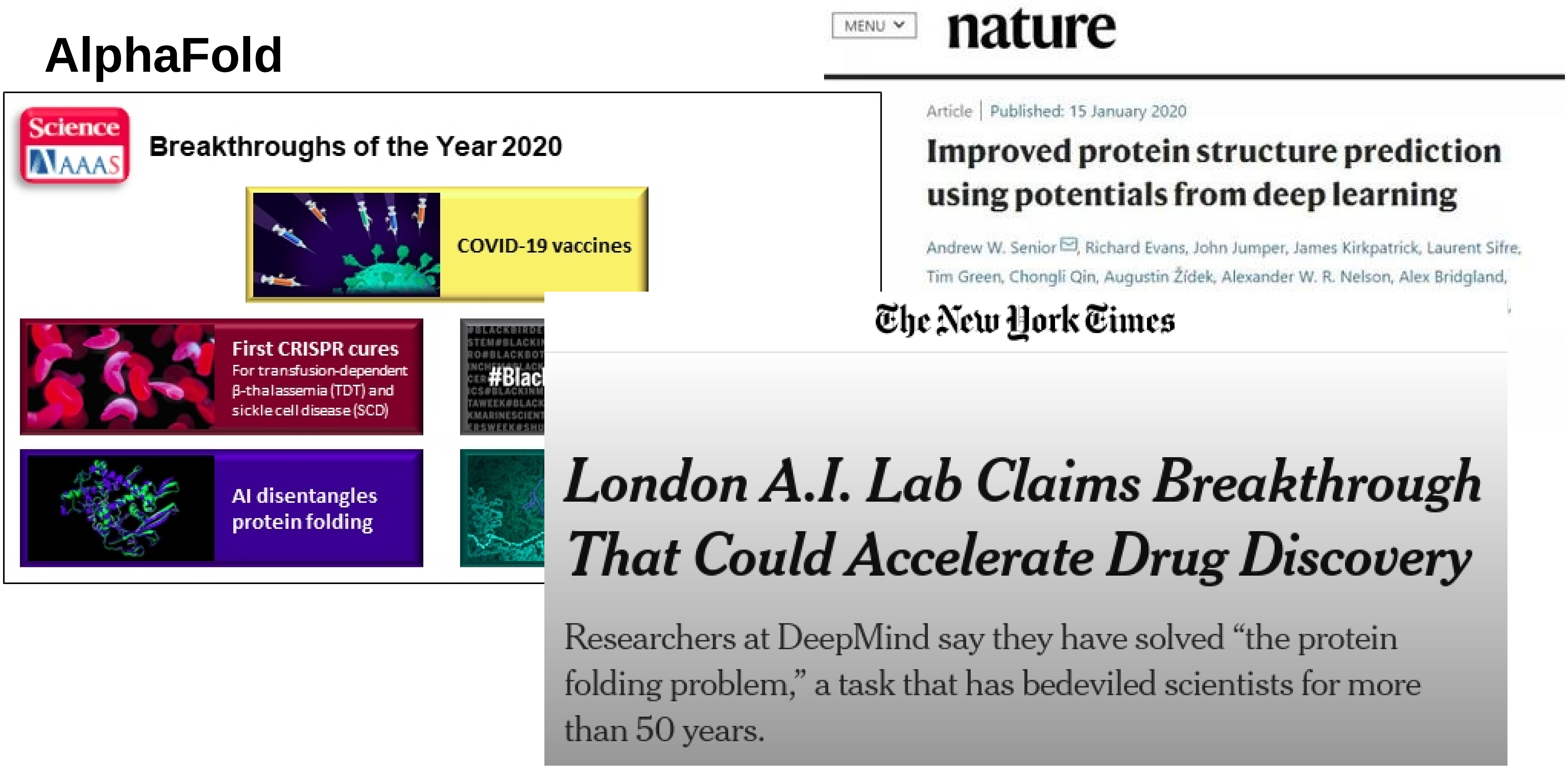
Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis

Nature (2020) | Cite this article

1835 Accesses | **245** Altmetric | Metrics

Scientific Competitions

AlphaFold



Breakthroughs of the Year 2020

COVID-19 vaccines

First CRISPR cures
For transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)

AI disentangles protein folding

nature

Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland,

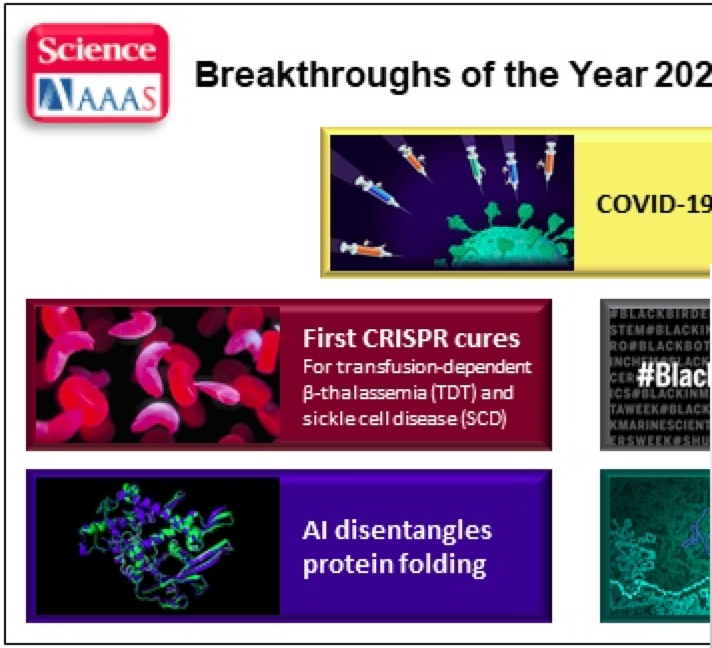
The New York Times

London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery

Researchers at DeepMind say they have solved “the protein folding problem,” a task that has bedeviled scientists for more than 50 years.

Scientific Competitions

AlphaFold



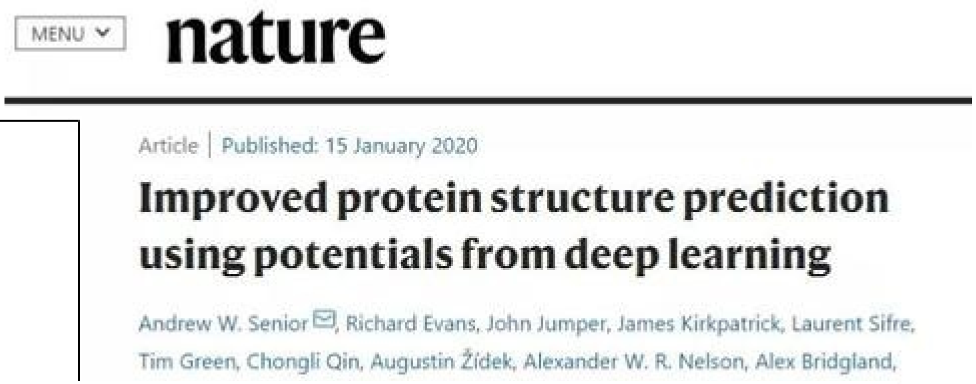
Science AAAS

Breakthroughs of the Year 2020

COVID-19 vaccines

First CRISPR cures
For transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)

AI disentangles protein folding



nature

Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre,
Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland,

The New York Times

London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery

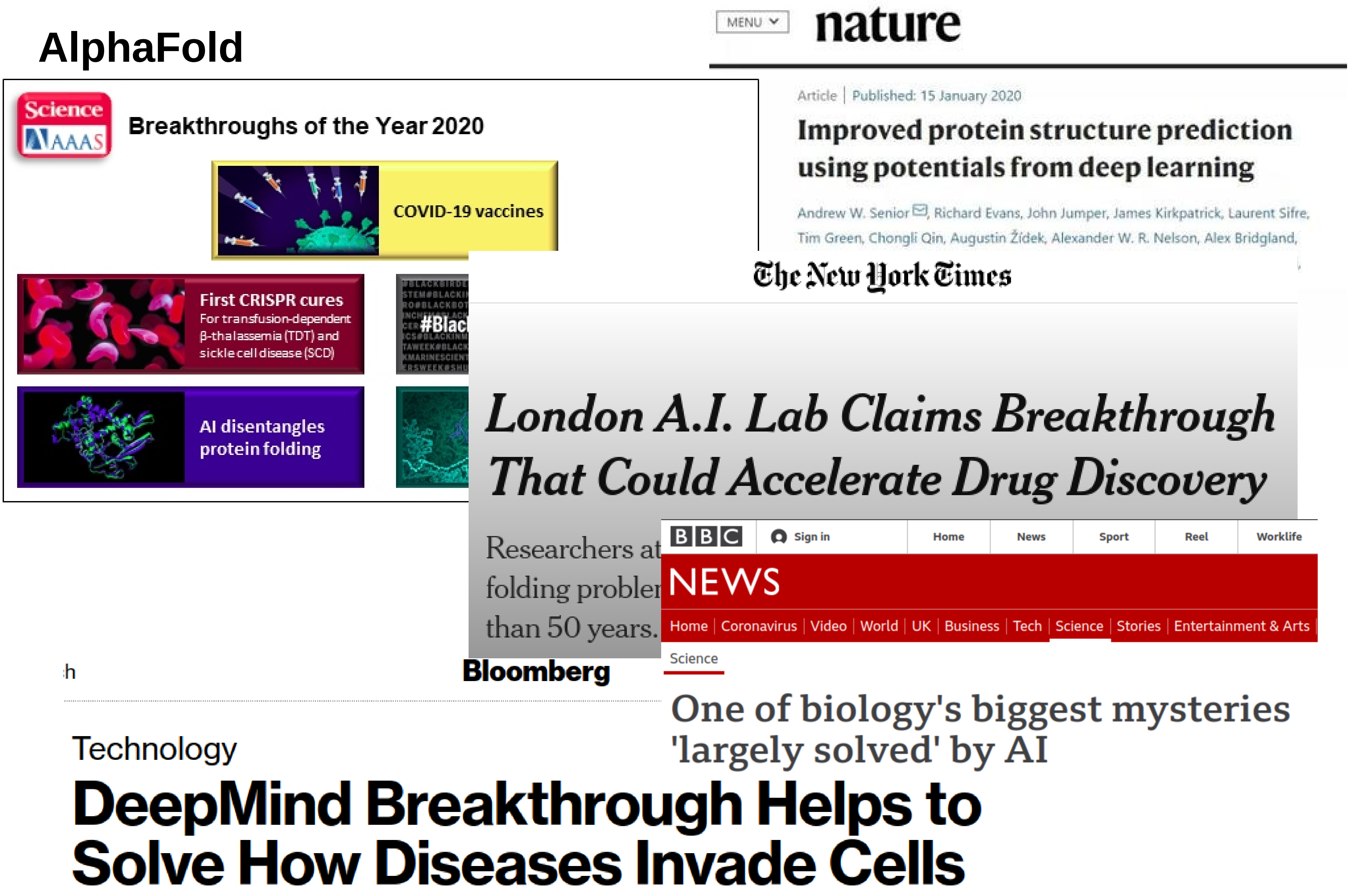
Researchers at DeepMind say they have solved “the protein folding problem,” a task that has bedeviled scientists for more than 50 years.

Bloomberg

Technology

DeepMind Breakthrough Helps to Solve How Diseases Invade Cells

AlphaFold



Science AAAS

Breakthroughs of the Year 2020

COVID-19 vaccines

First CRISPR cures
For transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD)

#Blac

AI disentangles protein folding

nature

Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland,

The New York Times

London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery

Researchers at
folding probler
than 50 years.

BBC

NEWS

Science

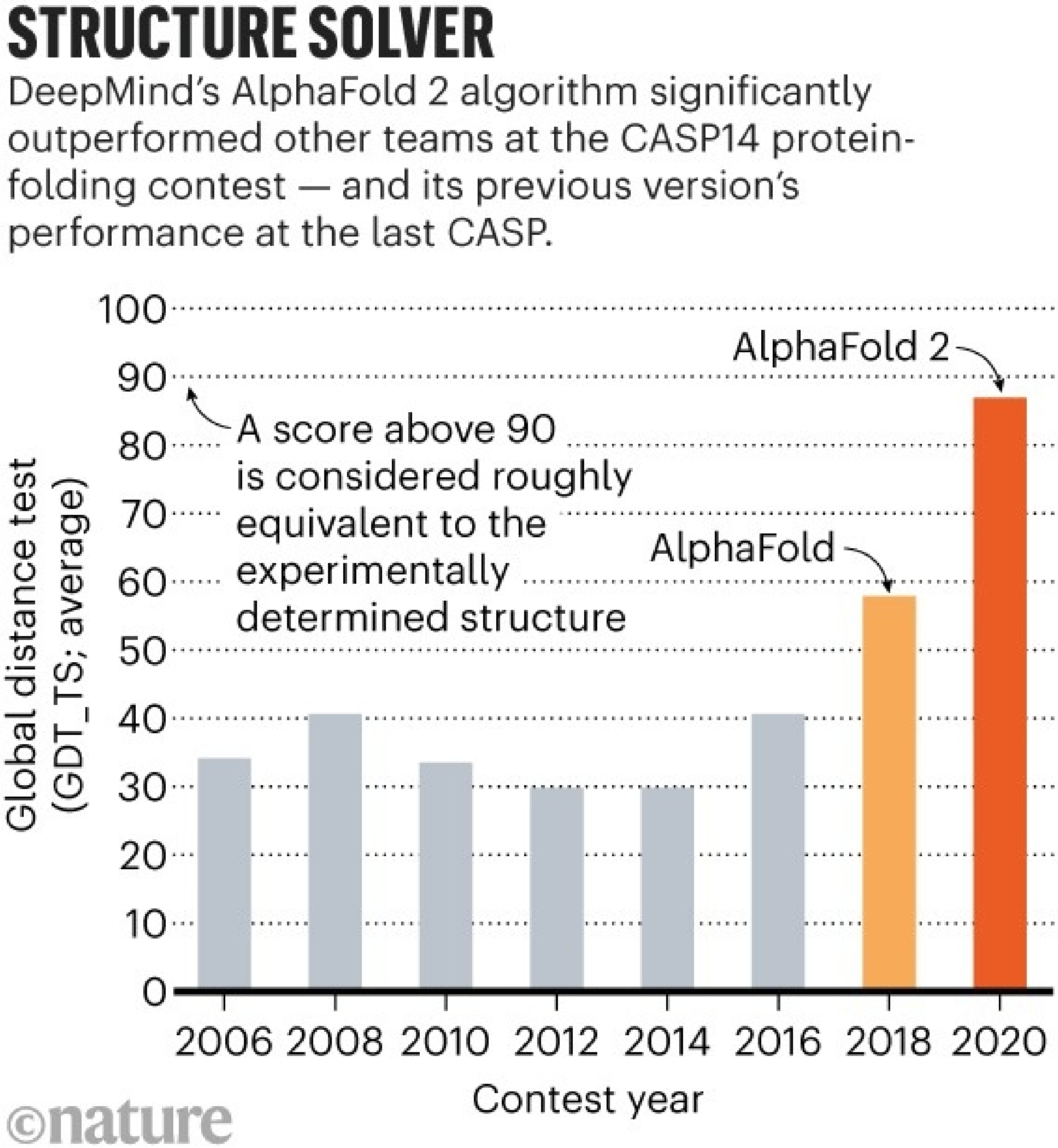
One of biology's biggest mysteries 'largely solved' by AI

Bloomberg

Technology

DeepMind Breakthrough Helps to Solve How Diseases Invade Cells

AlphaFold


STRUCTURE SOLVER
DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.
AlphaFold 2
AlphaFold
A score above 90 is considered roughly equivalent to the experimentally determined structure
Global distance test (GDT_TS; average)
100
90
80
70
60
50
40
30
20
10
0
2006
2008
2010
2012
2014
2016
2018
2020
Contest year
©nature

AlphaFold – like programs

DeepFold https://pubmed.ncbi.nlm.nih.gov/36112717/
RGN2 https://www.nature.com/articles/s41587-022-01432-w
ProtGPT2 https://www.nature.com/articles/s41467-022-32007-7

https://github.com/RosettaCommons**/RoseTTAFold**

equifold
https://www.biorxiv.org/content/10.1101/2022.10.07.511322v1

DMPfold
https://github.com/psipred/DMPfold2
https://www.pnas.org/doi/10.1073/pnas.2113348119

ESMFold
https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1.abstract
https://github.com/facebookresearch/esm
https://www.nature.com/articles/d41586-022-03539-1
esmatlas.com

omegafold
https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1.abstract

HelixFold
https://arxiv.org/pdf/2207.05477.pdf

ProteinBERT
https://www.biorxiv.org/content/10.1101/2021.05.24.445464v1
http://dx.doi.org/10.1093/bioinformatics/btac020

trRosettaX-Single https://doi.org/10.1038/s43588-022-00373-3
https://yanglab.nankai.edu.cn/trRosetta/benchmark_single/

https://analyticsindiamag.com/protein-wars-its-esmfold-vs-alphafold/

AlphaFold – like programs

#protein seq from backbone
ProteinMPNN paper: https://t.co/BLPg2XdmYE
https://colab.research.google.com/github/sokrypton/ColabDesign/blob/v1.1.0/mpnn/examples/proteinmpnn_in_jax.ipynb#scrollTo=GjdIxO4j-Hnn

ProGen2: Exploring the Boundaries of Protein Language Models https://arxiv.org/pdf/2206.13517.pdf

RITA: a Study on Scaling Up Generative Protein Sequence Models https://arxiv.org/pdf/2205.05789.pdf
https://github.com/lightonai/RITA

ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design
https://www.biorxiv.org/content/10.1101/2023.01.23.525232v1

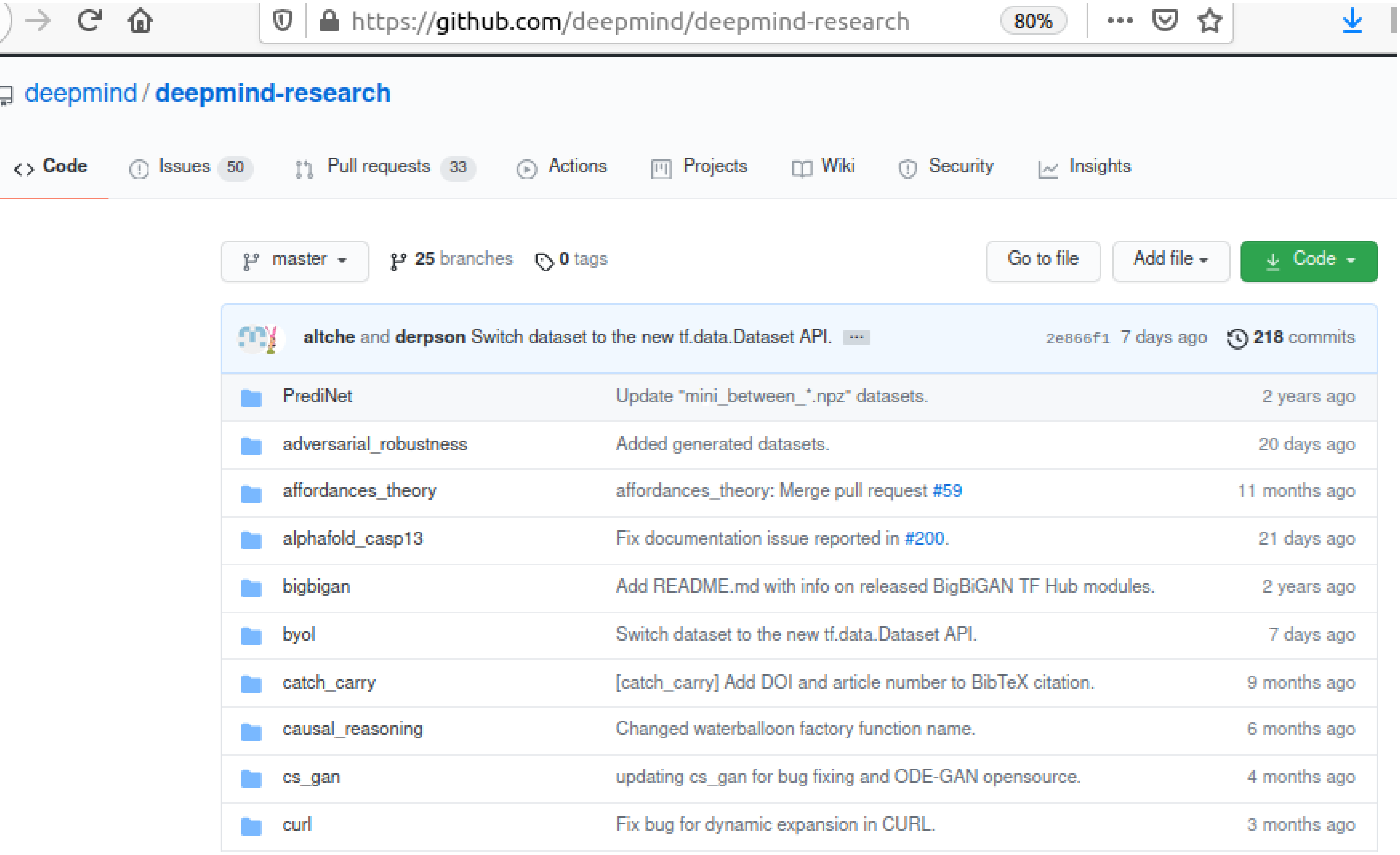
RSA
Retrieved Sequence Augmentation for Protein Representation Learning
https://www.biorxiv.org/content/10.1101/2023.02.22.529597v2.abstract
https://github.com/HKUNLP/RSA

Uni-Fold
https://github.com/dptech-corp/Uni-Fold#download-from-volcengine
https://colab.research.google.com/github/dptech-corp/Uni-Fold/blob/main/notebooks/unifold.ipynb

AlphaLink
https://www.nature.com/articles/s41587-023-01704-zProtein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning

EigenFold Generative Protein Structure Prediction with Diffusion Models
https://arxiv.org/abs/2304.02198
https://github.com/bjing2016/EigenFold

Scientific Competitions

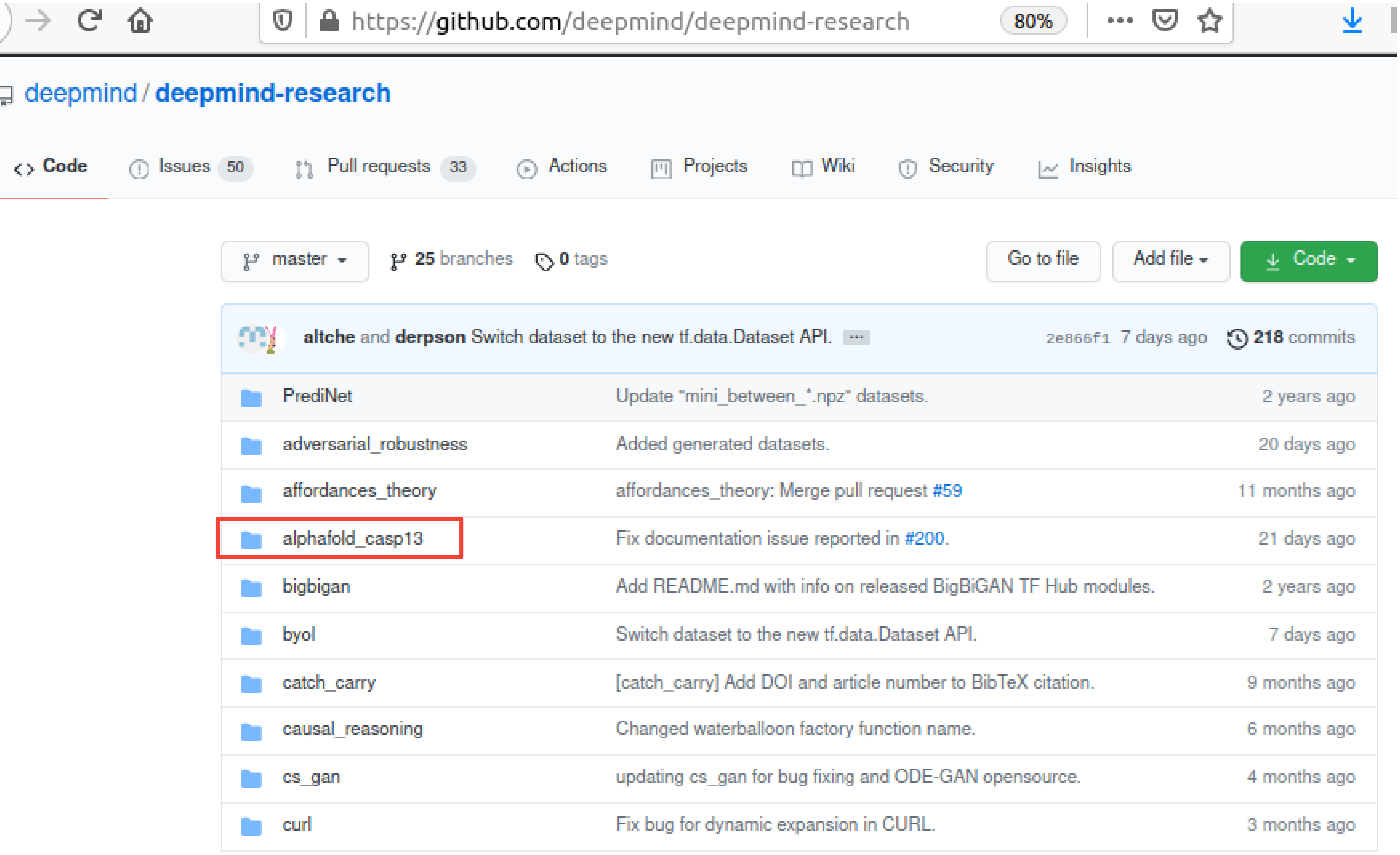


deepmind / **deepmind-research**

altche and **derpson** Switch dataset to the new tf.data.Dataset API. ··· `2e866f1` 7 days ago **218** commits

PrediNet	Update "mini_between_*.npz" datasets.	2 years ago
adversarial_robustness	Added generated datasets.	20 days ago
affordances_theory	affordances_theory: Merge pull request #59	11 months ago
alphafold_casp13	Fix documentation issue reported in #200.	21 days ago
bigbigan	Add README.md with info on released BigBiGAN TF Hub modules.	2 years ago
byol	Switch dataset to the new tf.data.Dataset API.	7 days ago
catch_carry	[catch_carry] Add DOI and article number to BibTeX citation.	9 months ago
causal_reasoning	Changed waterballoon factory function name.	6 months ago
cs_gan	updating cs_gan for bug fixing and ODE-GAN opensource.	4 months ago
curl	Fix bug for dynamic expansion in CURL.	3 months ago

Scientific Competitions



altche and **derpson** Switch dataset to the new tf.data.Dataset API. ··· 2e866f1 7 days ago **218** commits

PrediNet	Update "mini_between_*.npz" datasets.	2 years ago
adversarial_robustness	Added generated datasets.	20 days ago
affordances_theory	affordances_theory: Merge pull request #59	11 months ago
alphafold_casp13	Fix documentation issue reported in #200.	21 days ago
bigbigan	Add README.md with info on released BigBiGAN TF Hub modules.	2 years ago
byol	Switch dataset to the new tf.data.Dataset API.	7 days ago
catch_carry	[catch_carry] Add DOI and article number to BibTeX citation.	9 months ago
causal_reasoning	Changed waterballoon factory function name.	6 months ago
cs_gan	updating cs_gan for bug fixing and ODE-GAN opensource.	4 months ago
curl	Fix bug for dynamic expansion in CURL.	3 months ago

Scientific Competitions

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

Collective experiment for blind RNA structure prediction (RNA-Puzzles)

Critical Assessment of Prediction of Interactions (CAPRI)

Critical Assessment of Functional Annotation (CAFA)

Critical Assessment of Microarray Data Analysis (CAMDA)

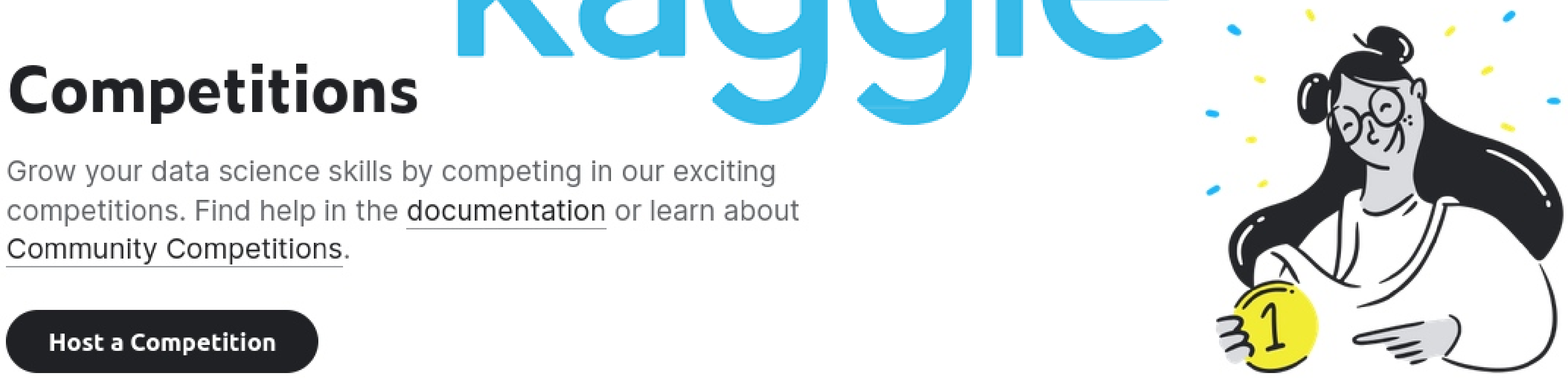
Genome Annotation Assessment Project (GASP)

Bone X-Ray Deep Learning Competition

LUng Nodule Analysis 2016

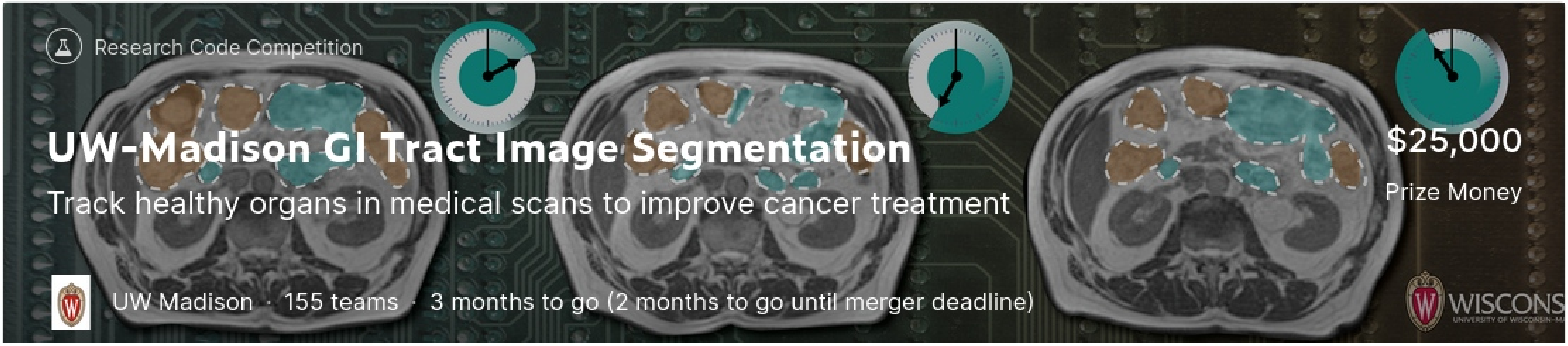
Scientific Competitions

kaggle



Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the documentation or learn about Community Competitions.



Research Code Competition

UW-Madison GI Tract Image Segmentation

Track healthy organs in medical scans to improve cancer treatment

UW Madison · 155 teams · 3 months to go (2 months to go until merger deadline)

$25,000

Prize Money

Research Code Competition

Mechanisms of Action (MoA) Prediction

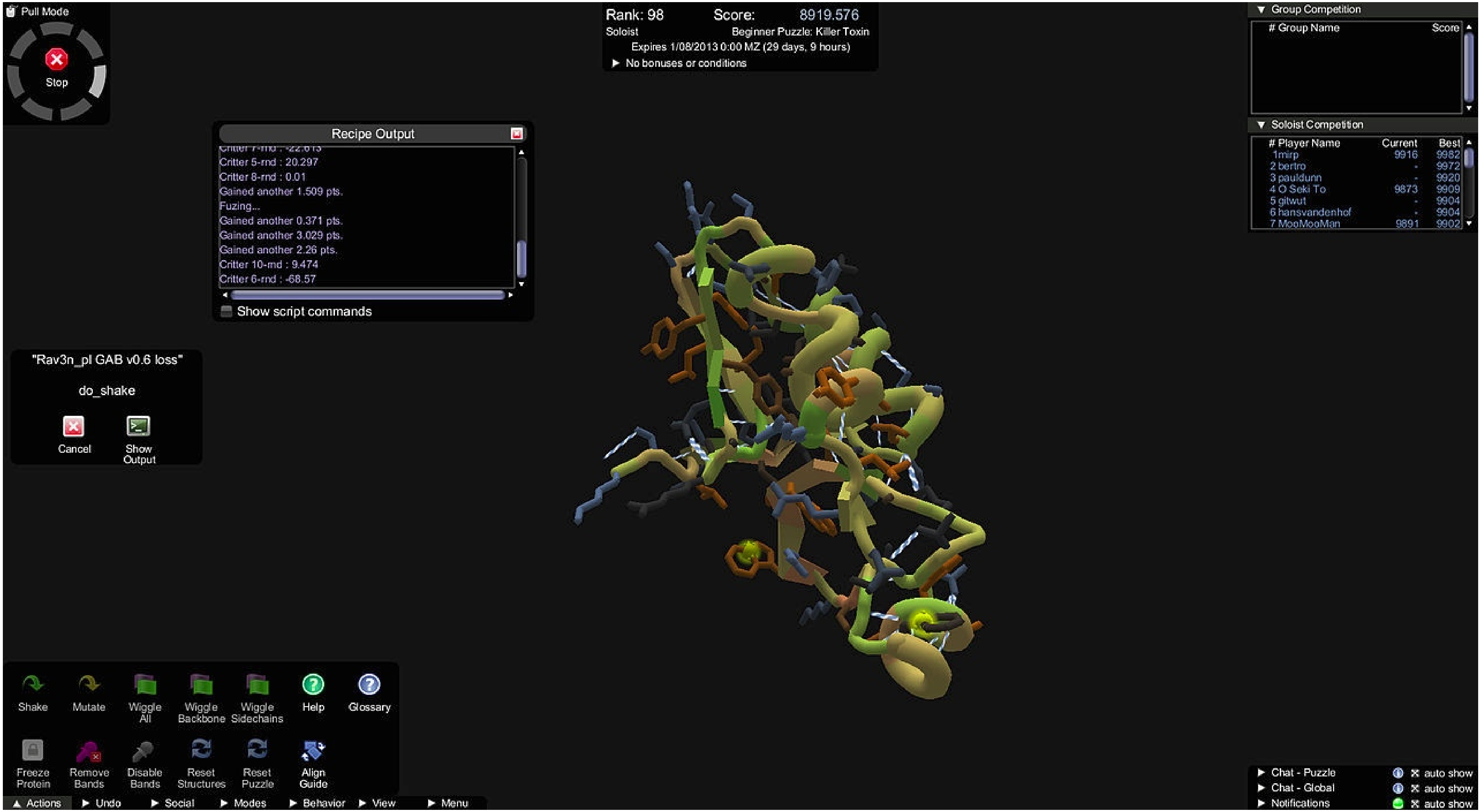
Can you improve the algorithm that classifies drugs based on their biological activity?

Laboratory for Innovation Science at Harvard · 4,373 teams · a year ago

$30,000

Prize Money

FoldIt - online puzzle video game about protein folding



Rank: 98 Score: 8919.576
Soloist Beginner Puzzle: Killer Toxin
Expires 1/08/2013 0:00 MZ (29 days, 9 hours)
▶ No bonuses or conditions

Recipe Output

```
Critter 5-rnd : 20.297
Critter 8-rnd : 0.01
Gained another 1.509 pts.
Fuzing...
Gained another 0.371 pts.
Gained another 3.029 pts.
Gained another 2.26 pts.
Critter 10-rnd : 9.474
Critter 6-rnd : -68.57
```

☐ Show script commands

"Rav3n_pl GAB v0.6 loss"

do_shake

▼ Group Competition

# Group Name	Score

▼ Soloist Competition

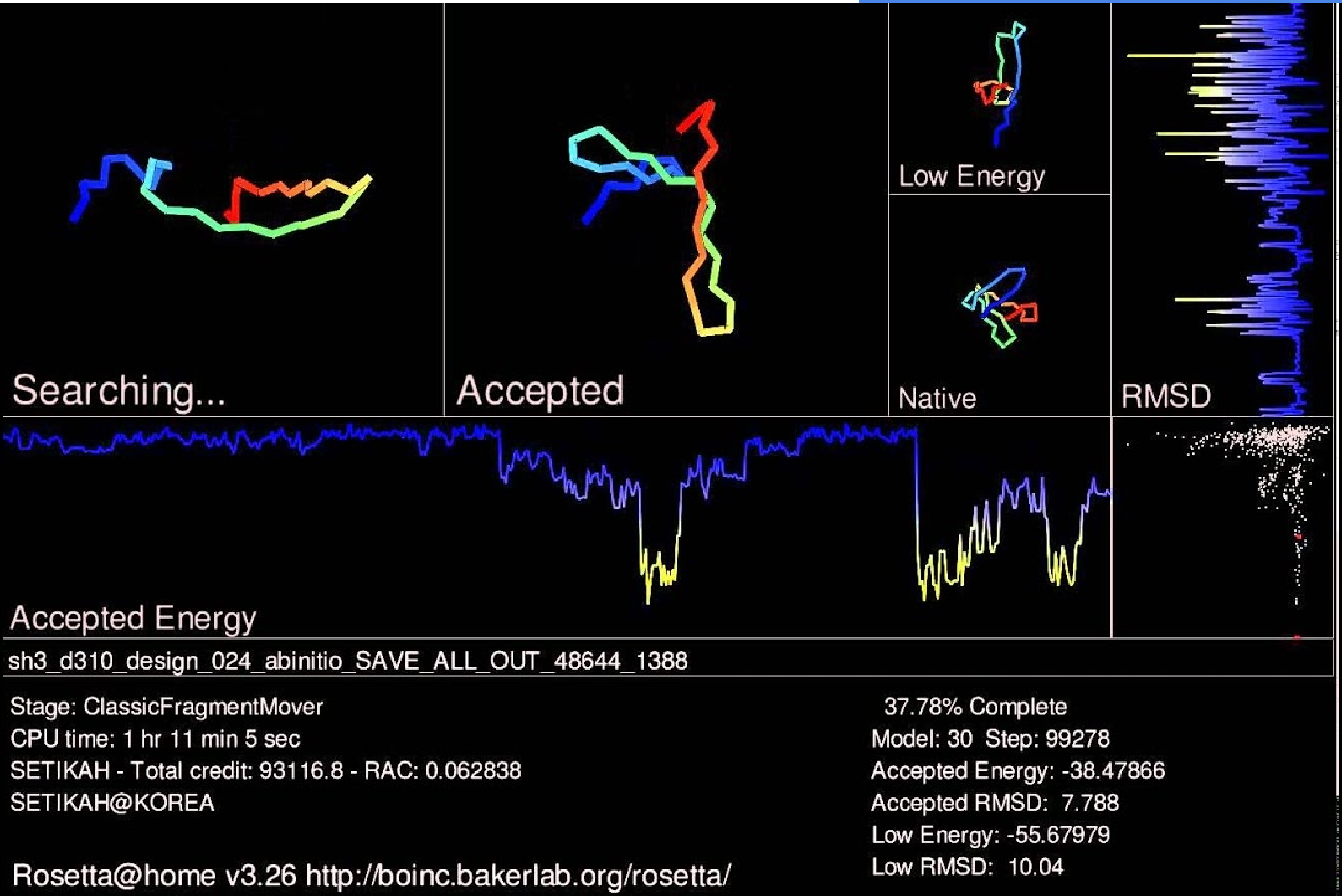
# Player Name	Current	Best
1mirp	9916	9982
2 bertro	-	9972
3 pauldunn	-	9920
4 O Seki To	9873	9909
5 gitwut	-	9904
6 hansvandenhof	-	9904
7 MooMooMan	9891	9902

Scientific Competitions

Rosetta@home
Protein Folding, Design, and Docking

You don't have to be a scientist to do science.

By simply running a free program, you can help advance research in medicine, clean energy, and materials science.



```
sh3_d310_design_024_abinitio_SAVE_ALL_OUT_48644_1388

Stage: ClassicFragmentMover                              37.78% Complete
CPU time: 1 hr 11 min 5 sec                             Model: 30  Step: 99278
SETIKAH - Total credit: 93116.8 - RAC: 0.062838         Accepted Energy: -38.47866
SETIKAH@KOREA                                           Accepted RMSD:  7.788
                                                        Low Energy: -55.67979
Rosetta@home v3.26 http://boinc.bakerlab.org/rosetta/   Low RMSD:  10.04
```

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

lukaskoz@mimuw.edu.pl