Single-model quality assessment using protein structural and contact information with machine learning techniques

Speaker： Renzhi Cao

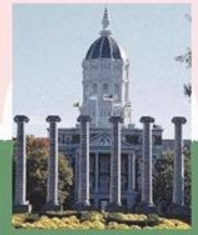
Major： Computer Science

Fifth year Ph.D

Mizzou
University of Missouri

Outline

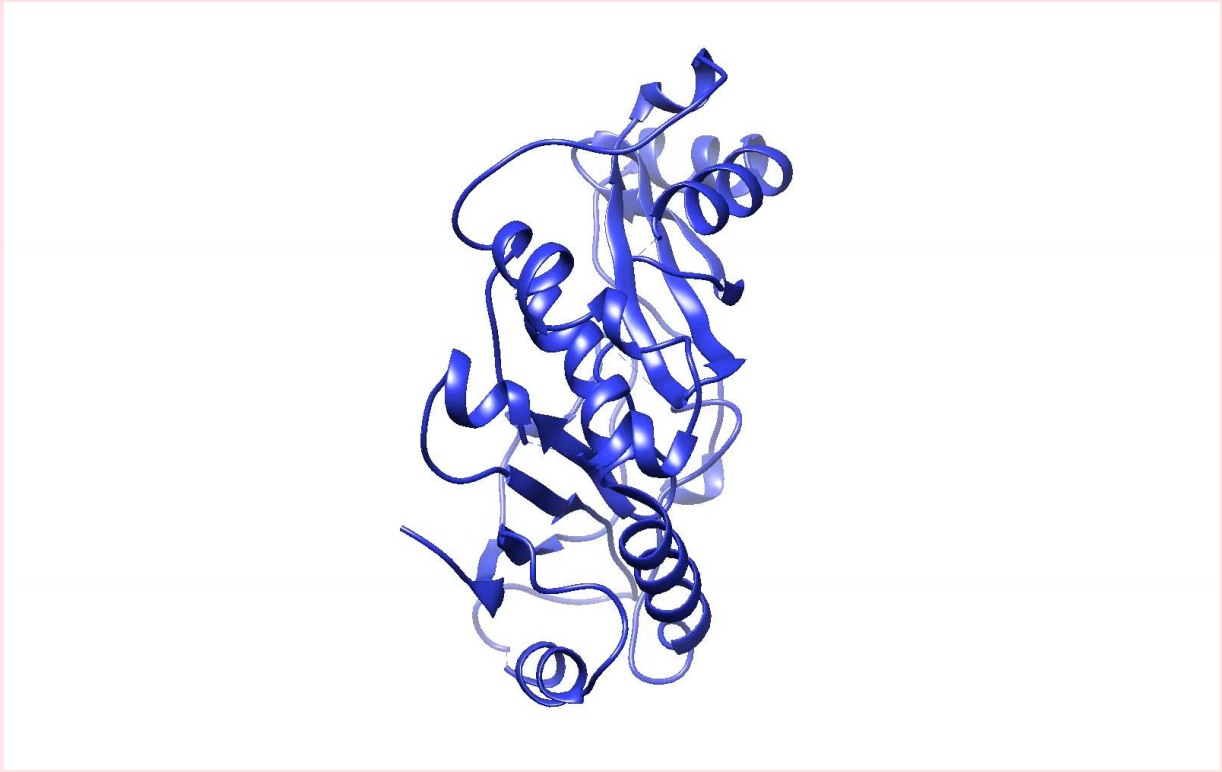
- Part I: Introduction
  - Protein quality assessment
  - CASP competition
- Part II: QAcon method
- Part III: Result
- Part IV: Conclusion

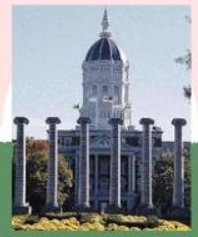


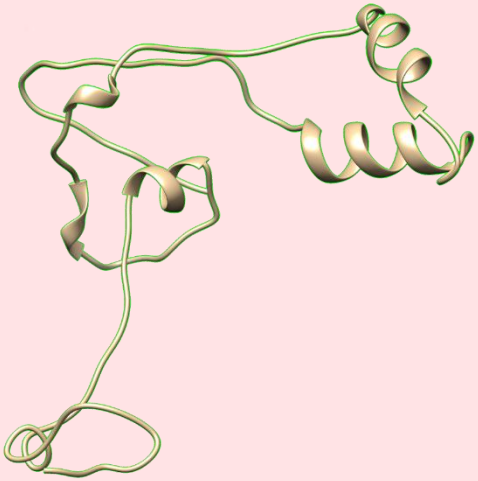
Mizzou
University of Missouri

Part I: Introduction

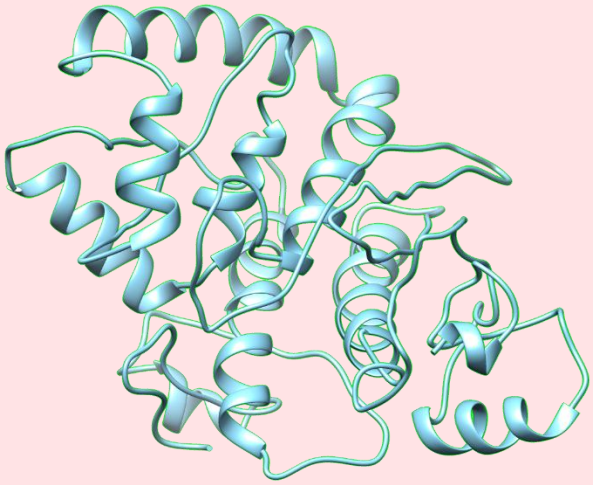
>T0759 HR9083A, Human, 109 residues
MGHHHHHHSHMVVIHPDPGRELSPEEAHRAGLIDWNMFVKLRSQECDWEEISVKGPNGES
SVIHDRKSGKKFSIEEALQSGRLTPAHYDRYVNKDMSIQELAVLVSGQK



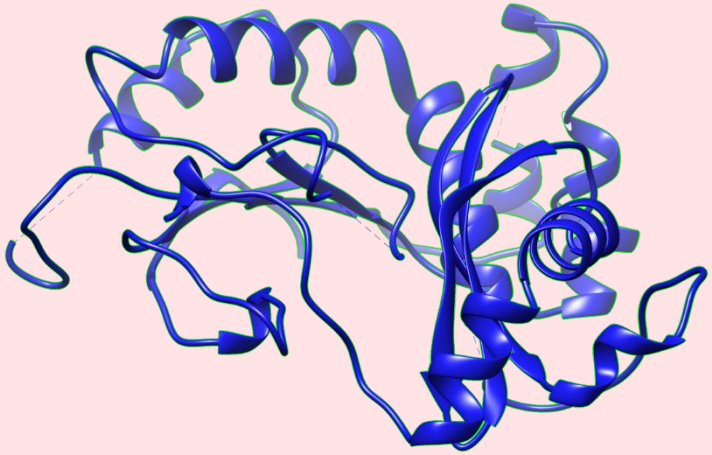




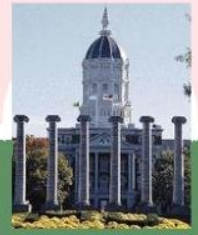
Predicted model1

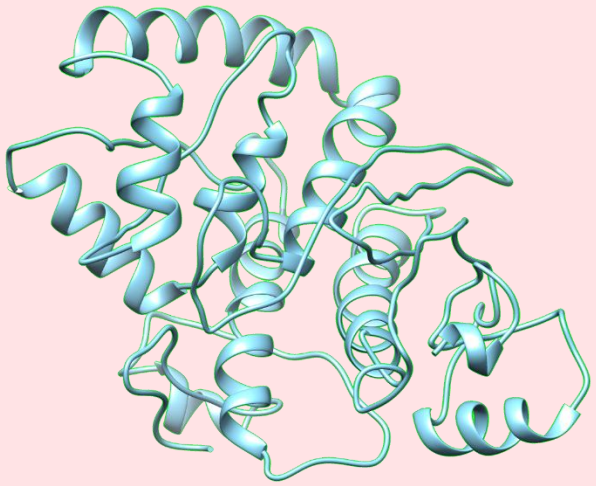


Predicted model2



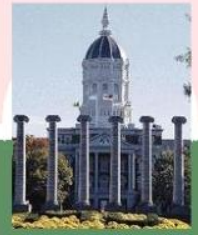
Native



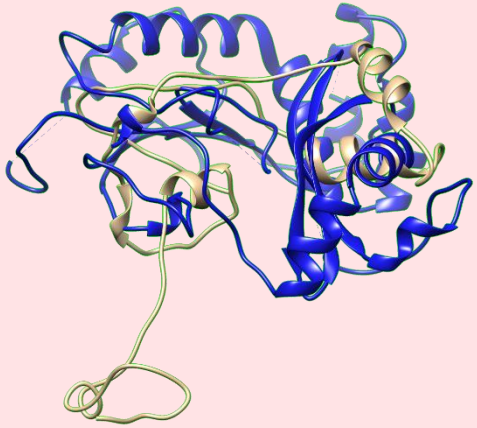


Predicted model2

Predicted model3

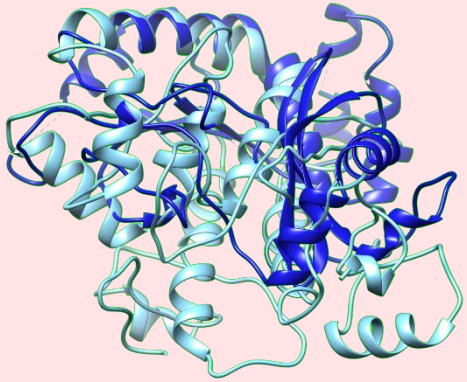


GDT-TS: 0.07



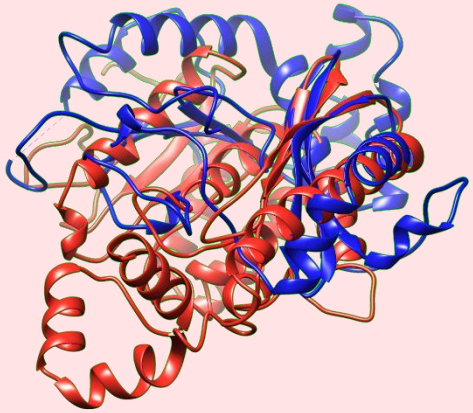
Predicted model1

GDT-TS: 0.21

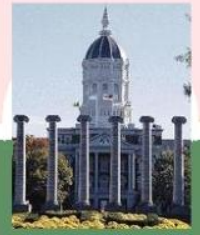


Predicted model2

GDT-TS: 0.33



Predicted model3



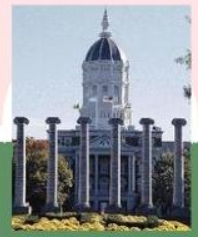
Evaluating metrics

1. Loss

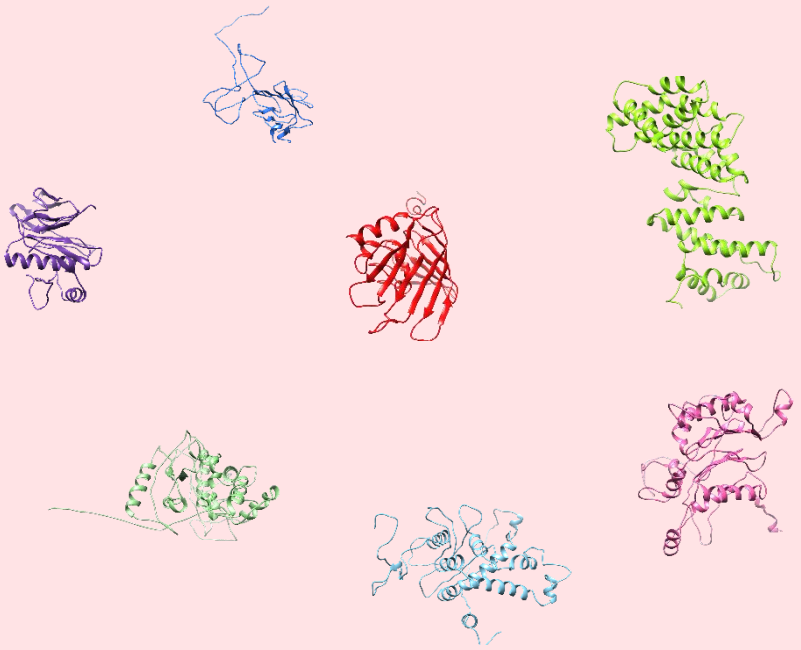
$$GDT_{best\ model} - GDT_{Predicted\ top\ 1}$$

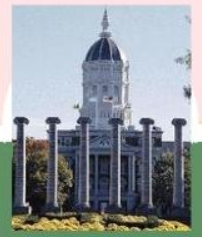
2. Correlation

$$\frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$



- CASP (Critical Assessment of Techniques for Protein Structure Prediction).
- Sel20 (Stage1)
- Top150 (Stage2)

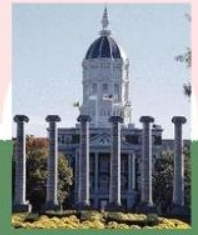


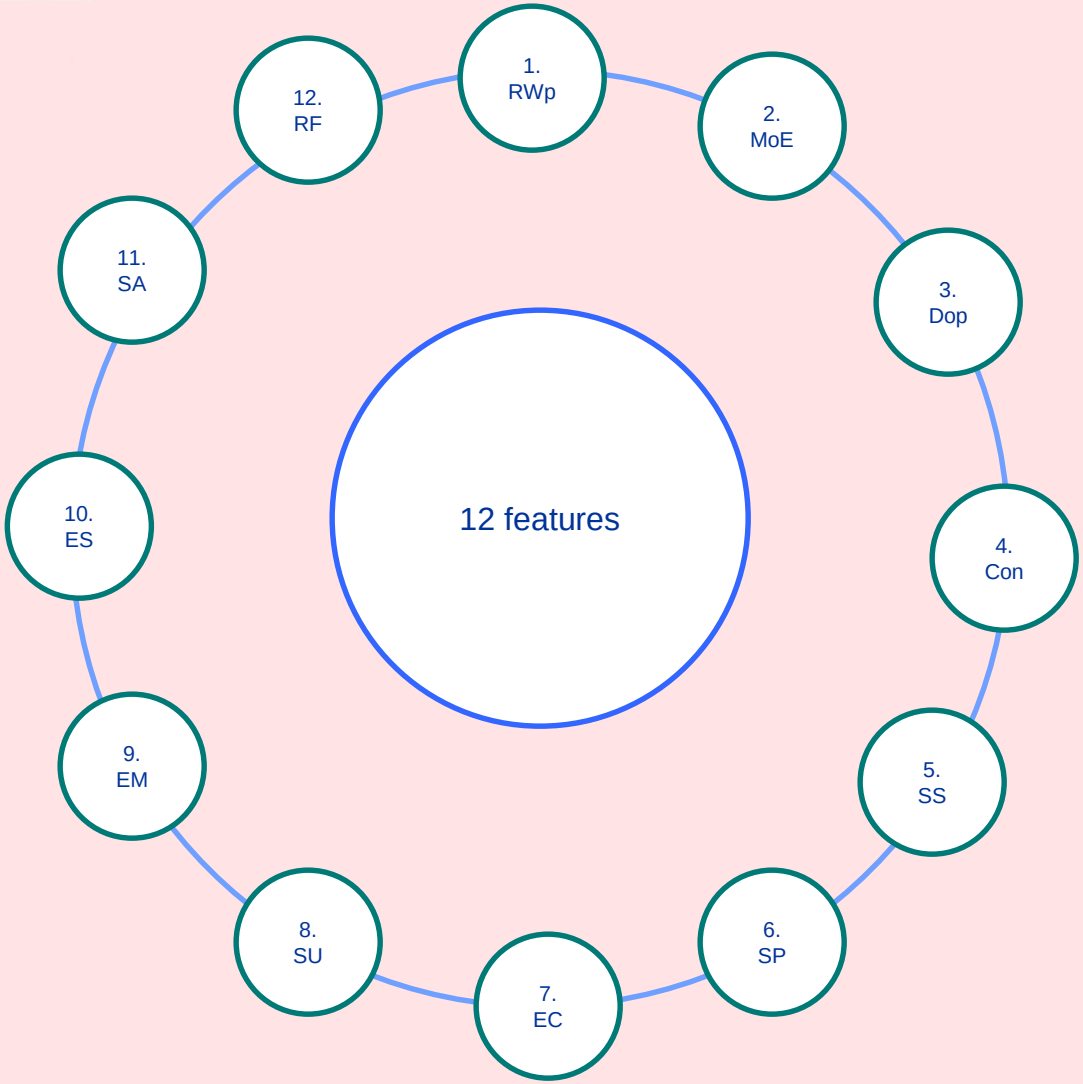


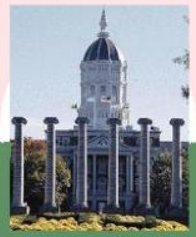
Mizzou
University of Missouri

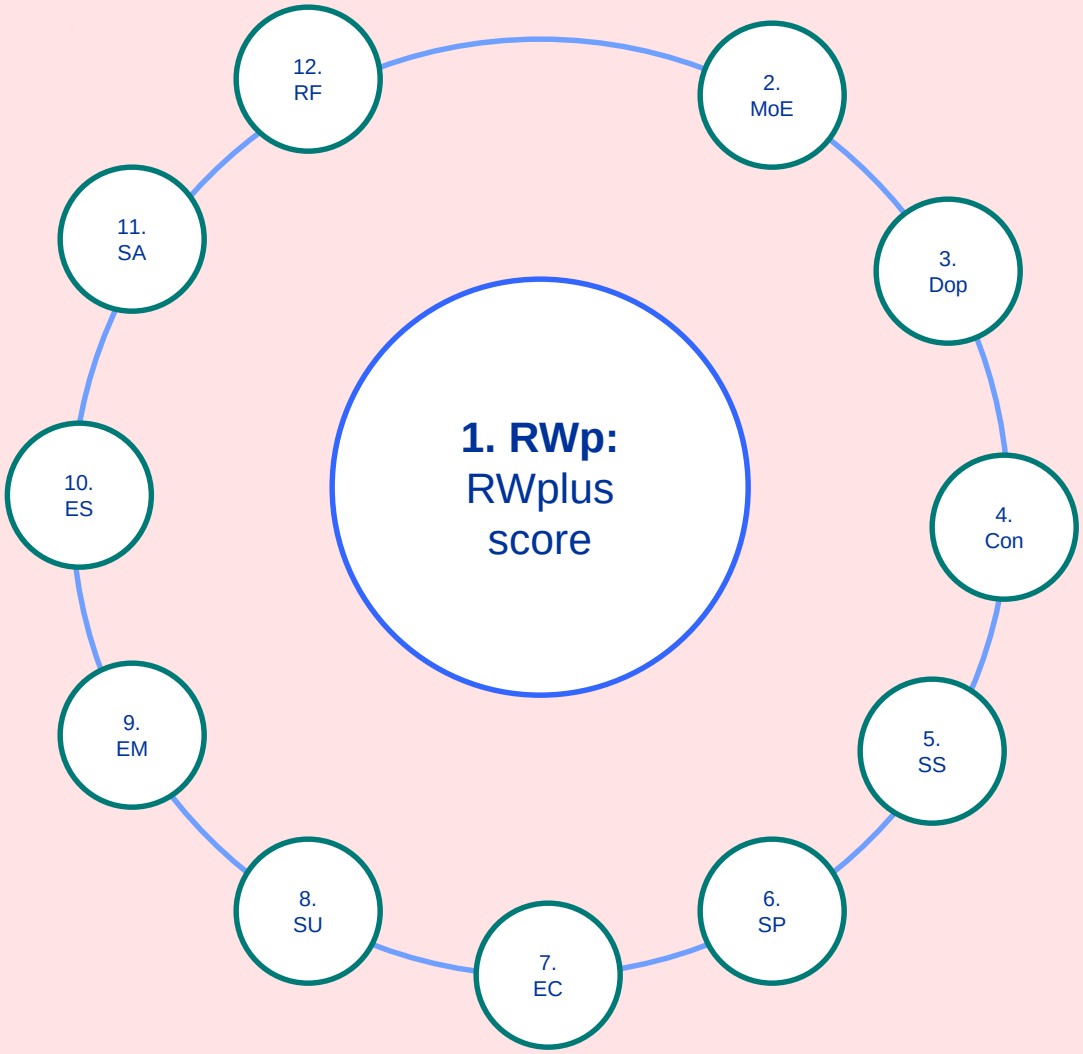
Outline

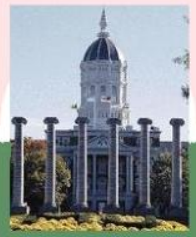
- Part I: Introduction
  - Protein quality assessment
  - CASP competition
- **Part II: QAcon method**
- Part III: Result
- Part IV: Conclusion

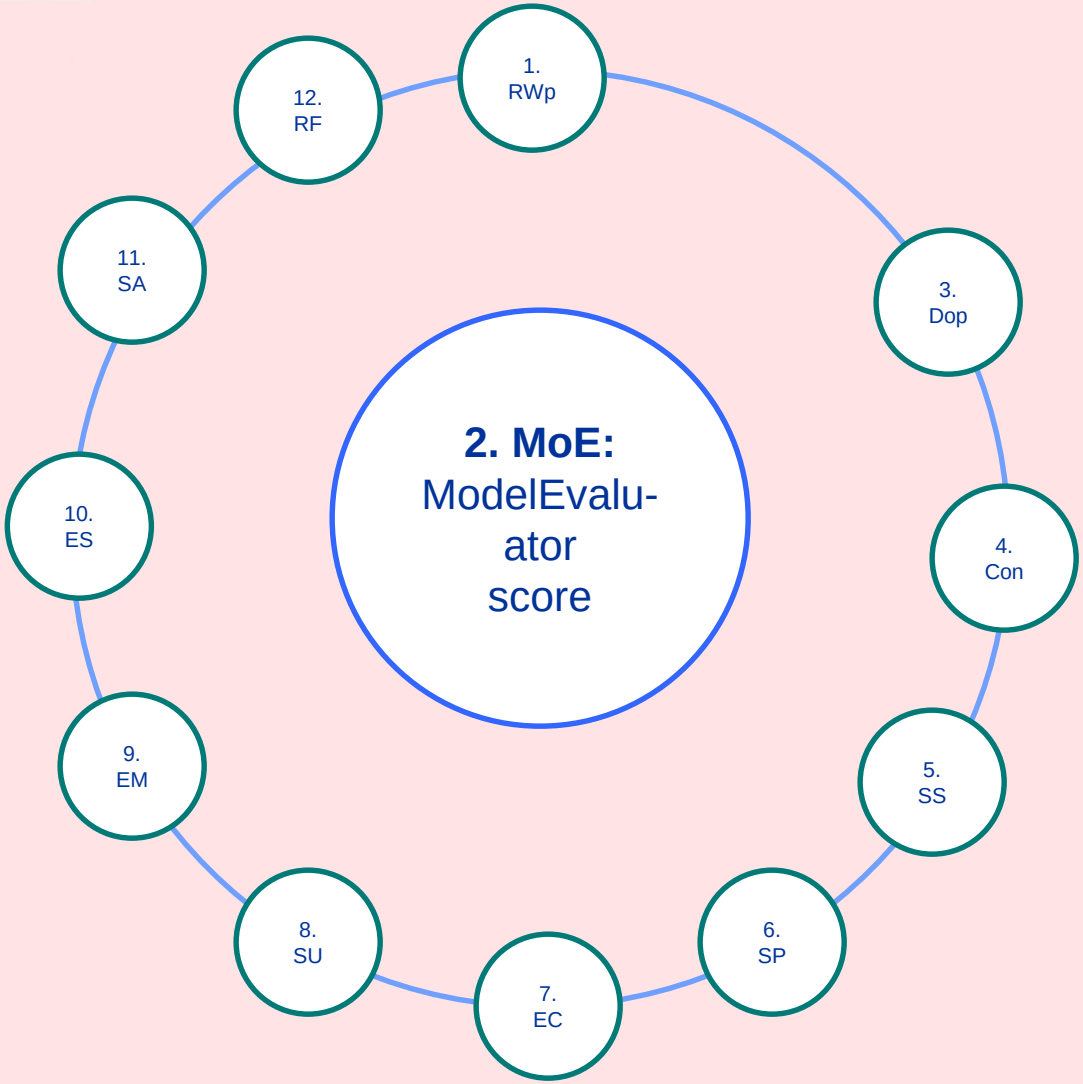


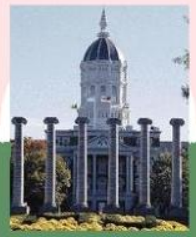

12 features
1. RWp
2. MoE
3. Dop
4. Con
5. SS
6. SP
7. EC
8. SU
9. EM
10. ES
11. SA
12. RF

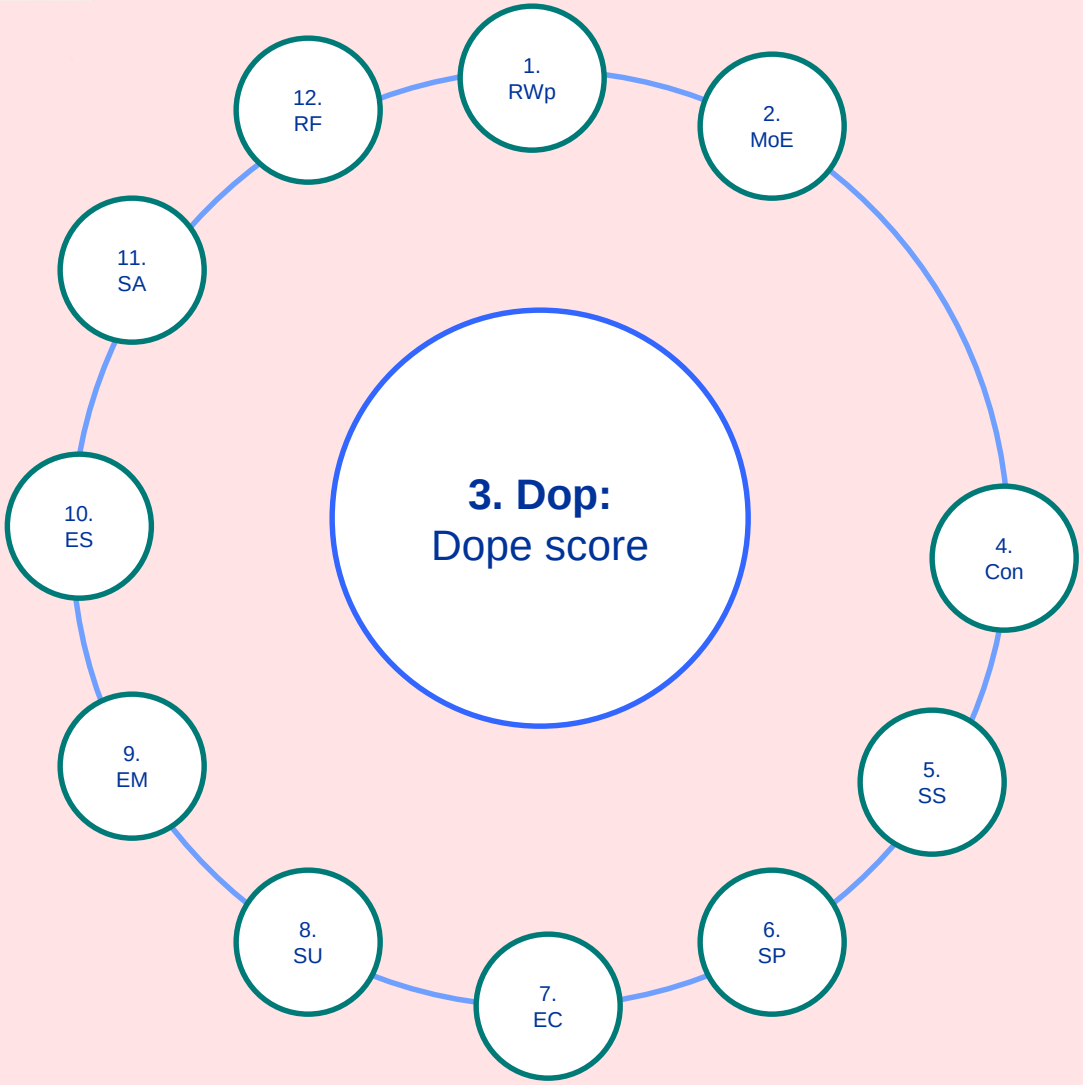


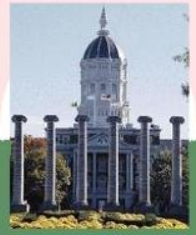

1. RWp:
RWplus
score
2.
MoE
3.
Dop
4.
Con
5.
SS
6.
SP
7.
EC
8.
SU
9.
EM
10.
ES
11.
SA
12.
RF

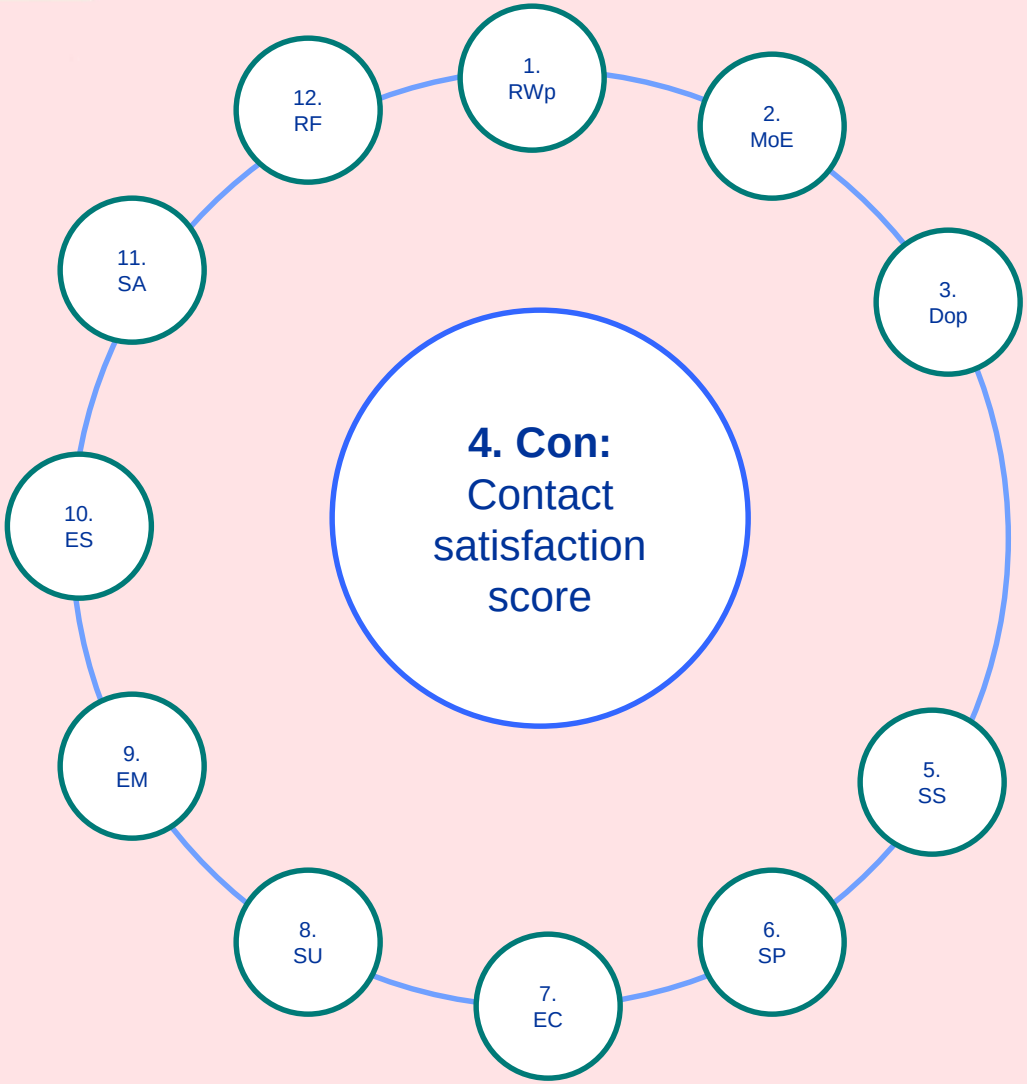


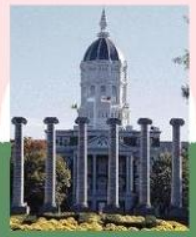

1.
RWp
3.
Dop
4.
Con
5.
SS
6.
SP
7.
EC
8.
SU
9.
EM
10.
ES
11.
SA
12.
RF
2. MoE:
ModelEvalu-
ator
score



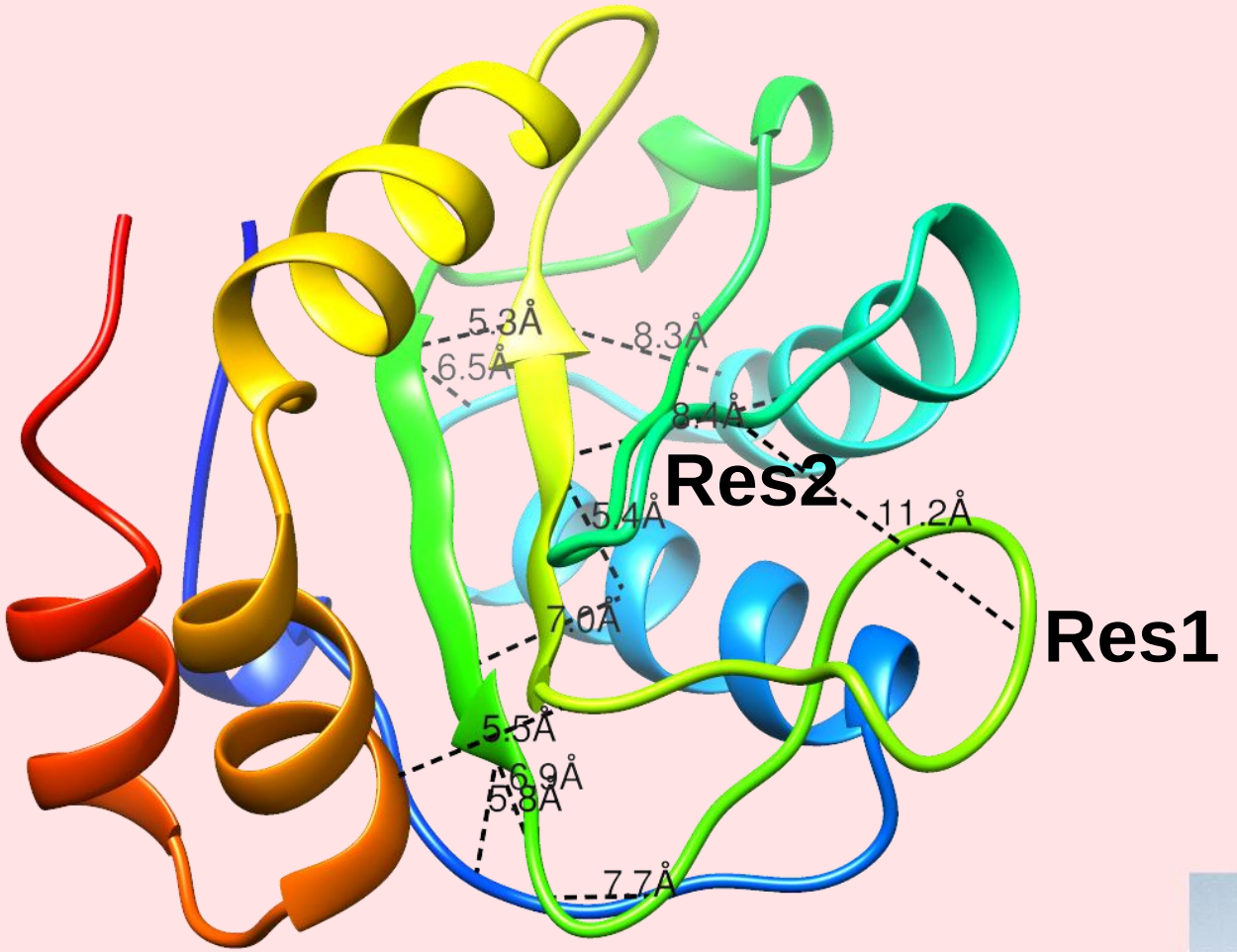

1.
RWp
2.
MoE
3. Dop:
Dope score
4.
Con
5.
SS
6.
SP
7.
EC
8.
SU
9.
EM
10.
ES
11.
SA
12.
RF

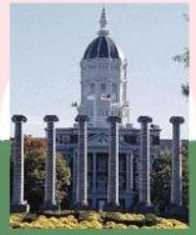


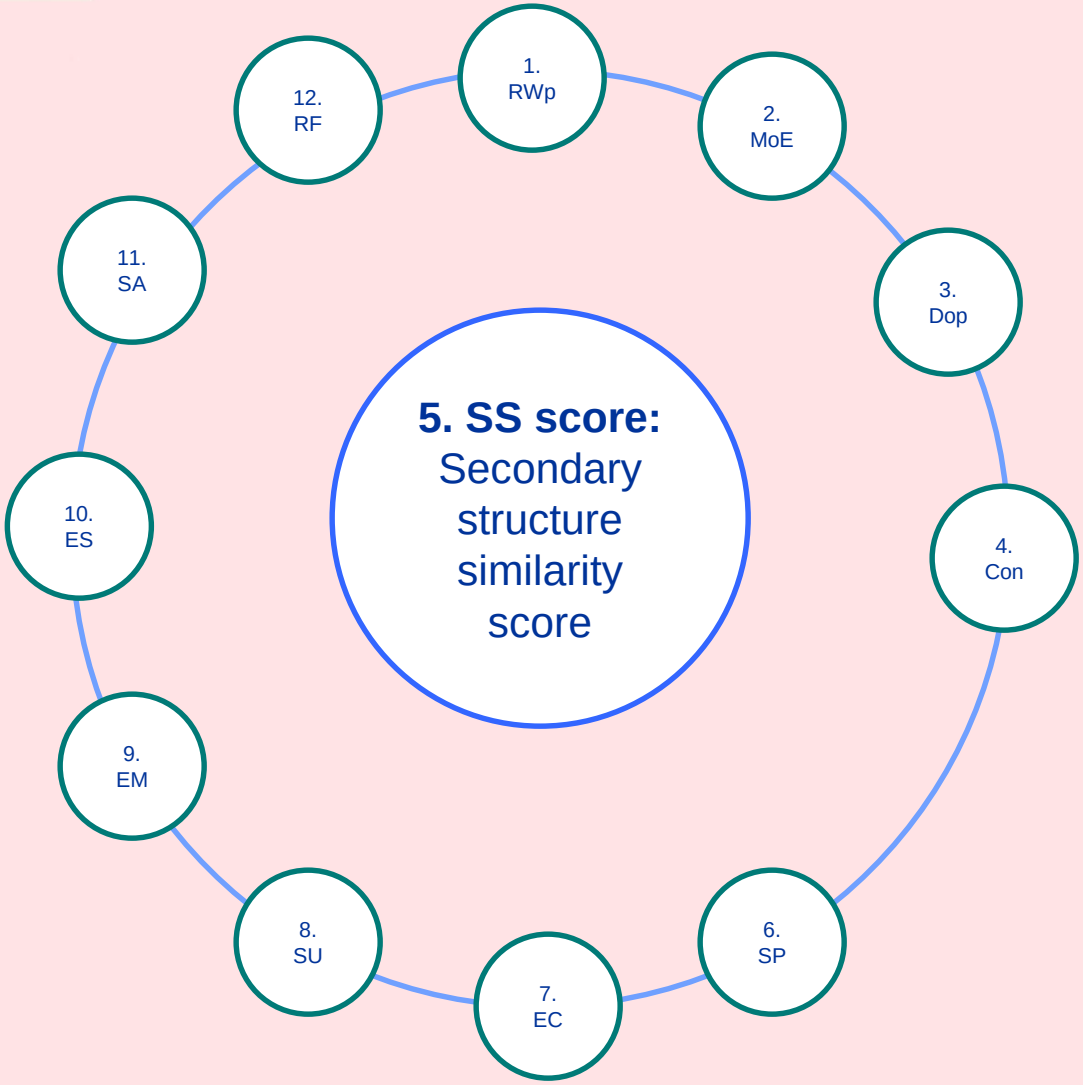

1.
RWp
2.
MoE
3.
Dop
4. Con:
Contact
satisfaction
score
5.
SS
6.
SP
7.
EC
8.
SU
9.
EM
10.
ES
11.
SA
12.
RF

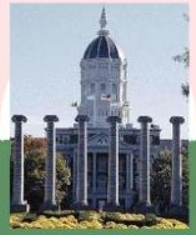


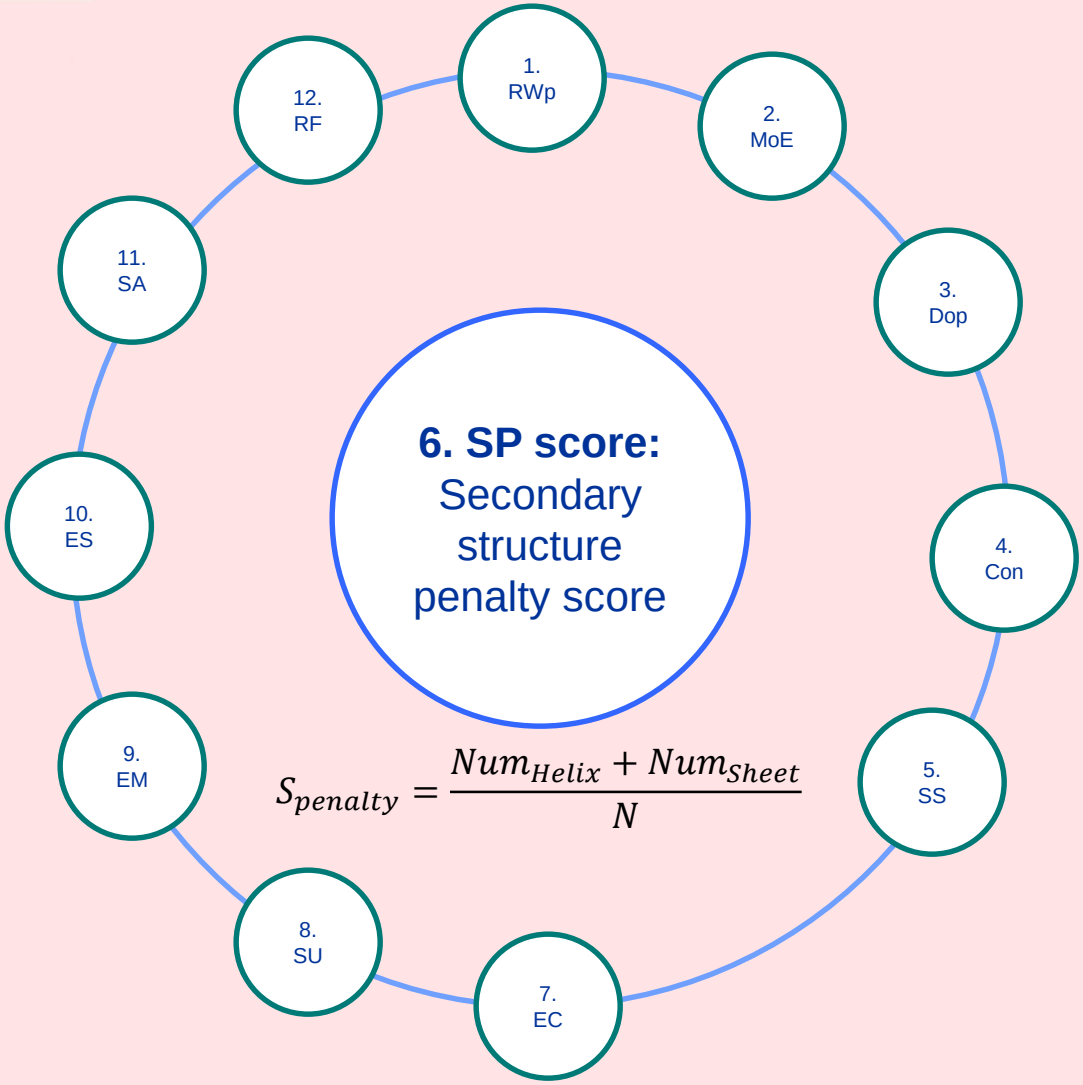
- Contact threshold is set to 8

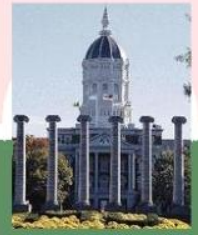


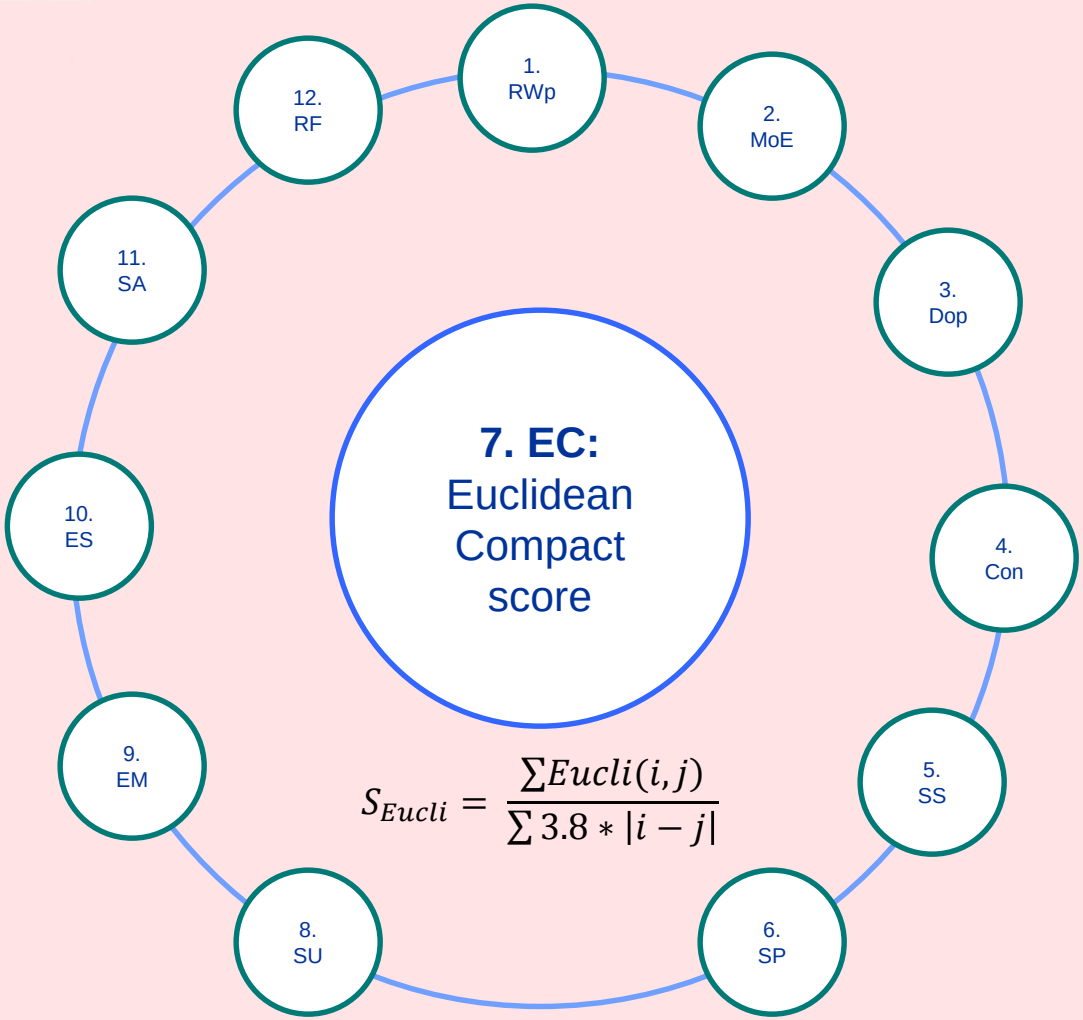



1.
RWp
2.
MoE
3.
Dop
4.
Con
6.
SP
7.
EC
8.
SU
9.
EM
10.
ES
11.
SA
12.
RF
5. SS score:
Secondary structure similarity score

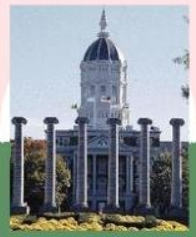


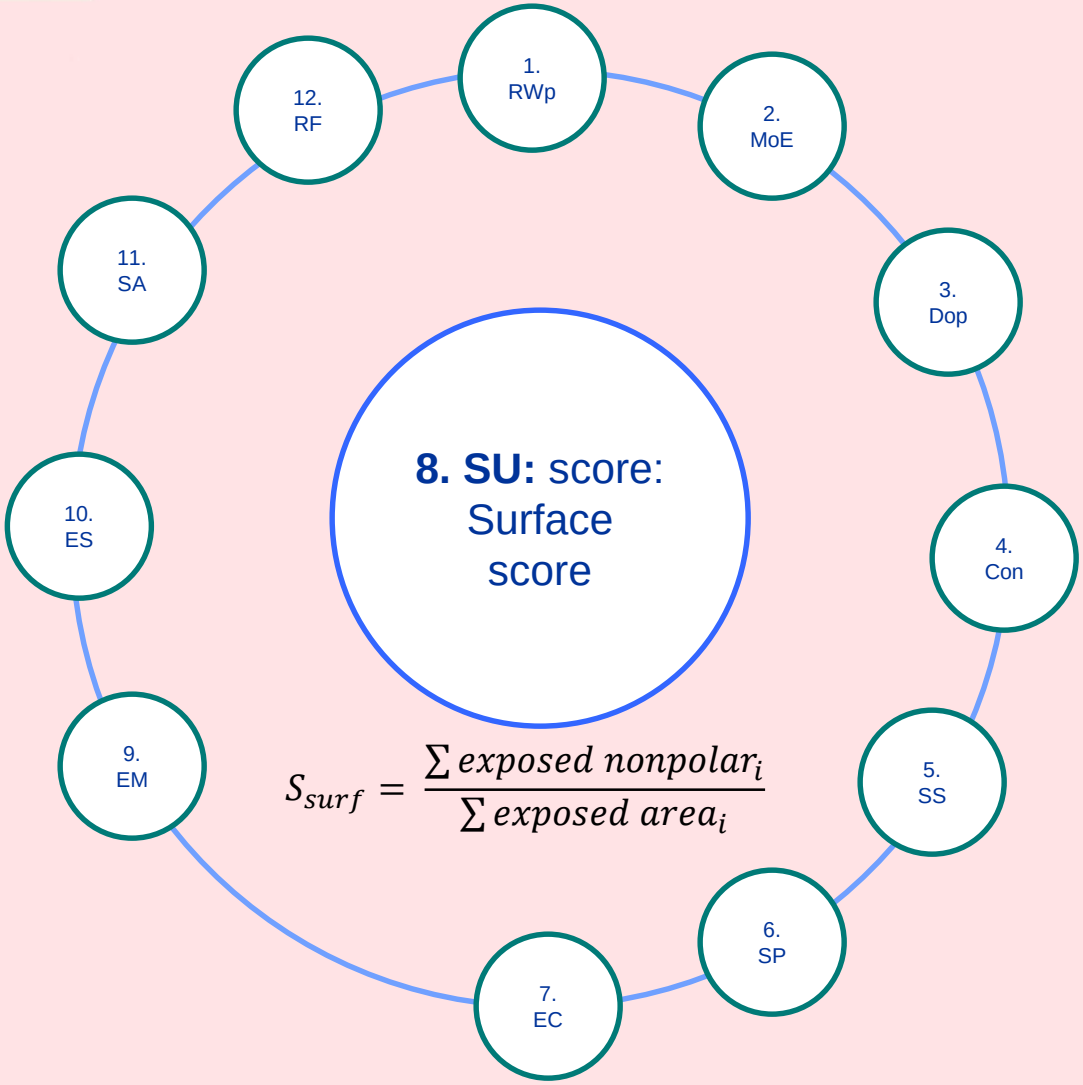

1. RWp
2. MoE
3. Dop
4. Con
5. SS
7. EC
8. SU
9. EM
10. ES
11. SA
12. RF
6. SP score: Secondary structure penalty score
$S_{penalty} = \frac{Num_{Helix} + Num_{Sheet}}{N}$

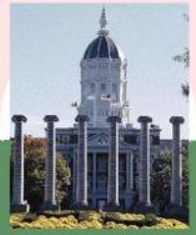


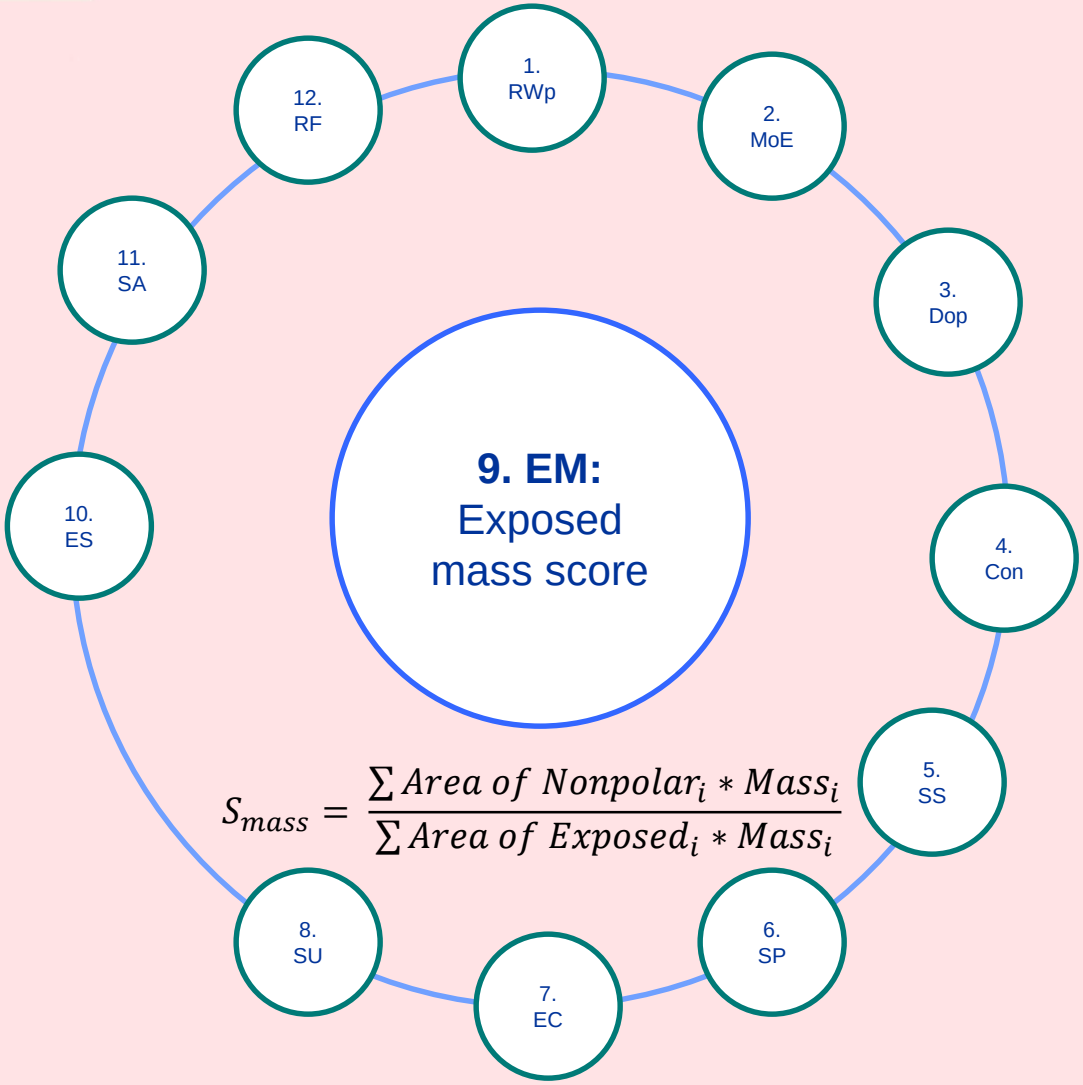


$$S_{Eucli} = \frac{\sum Eucli(i,j)}{\sum 3.8 * |i-j|}$$

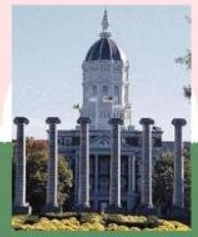


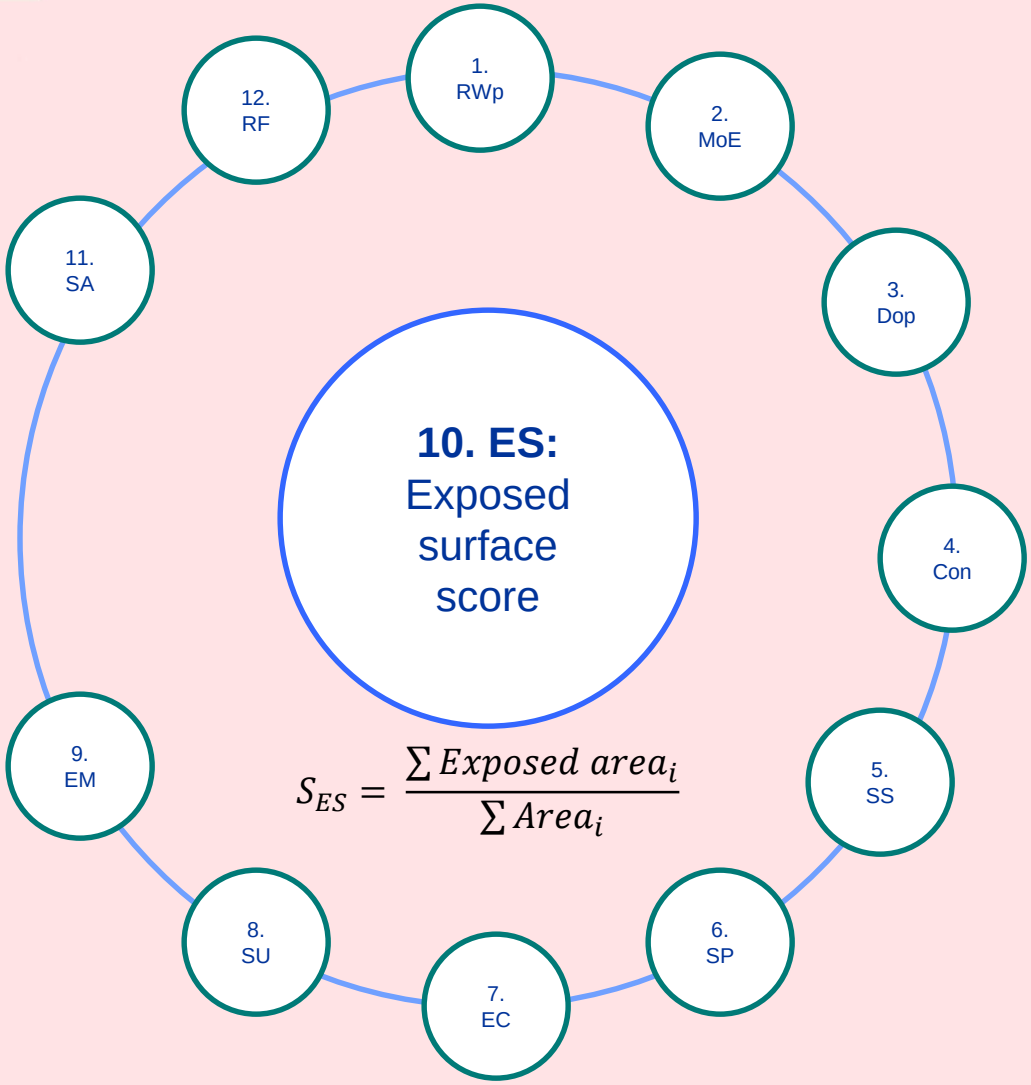

1. RWp
2. MoE
3. Dop
4. Con
5. SS
6. SP
7. EC
9. EM
10. ES
11. SA
12. RF
8. SU: score: Surface score
$S_{surf} = \frac{\sum exposed\ nonpolar_i}{\sum exposed\ area_i}$

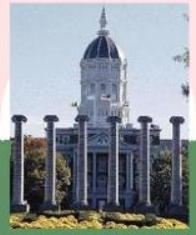


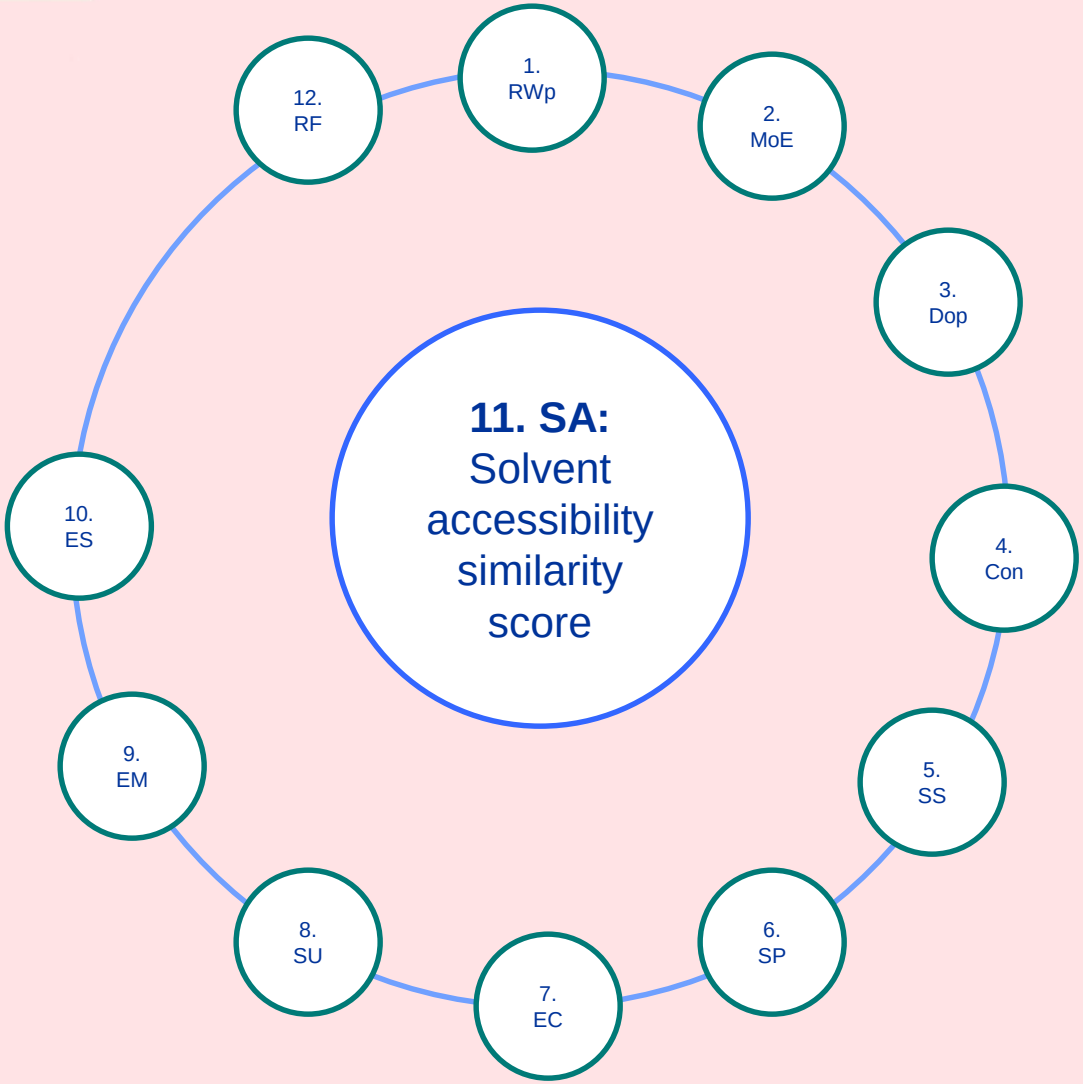


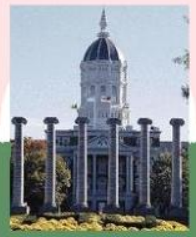
$$S_{mass} = \frac{\sum Area\ of\ Nonpolar_i * Mass_i}{\sum Area\ of\ Exposed_i * Mass_i}$$

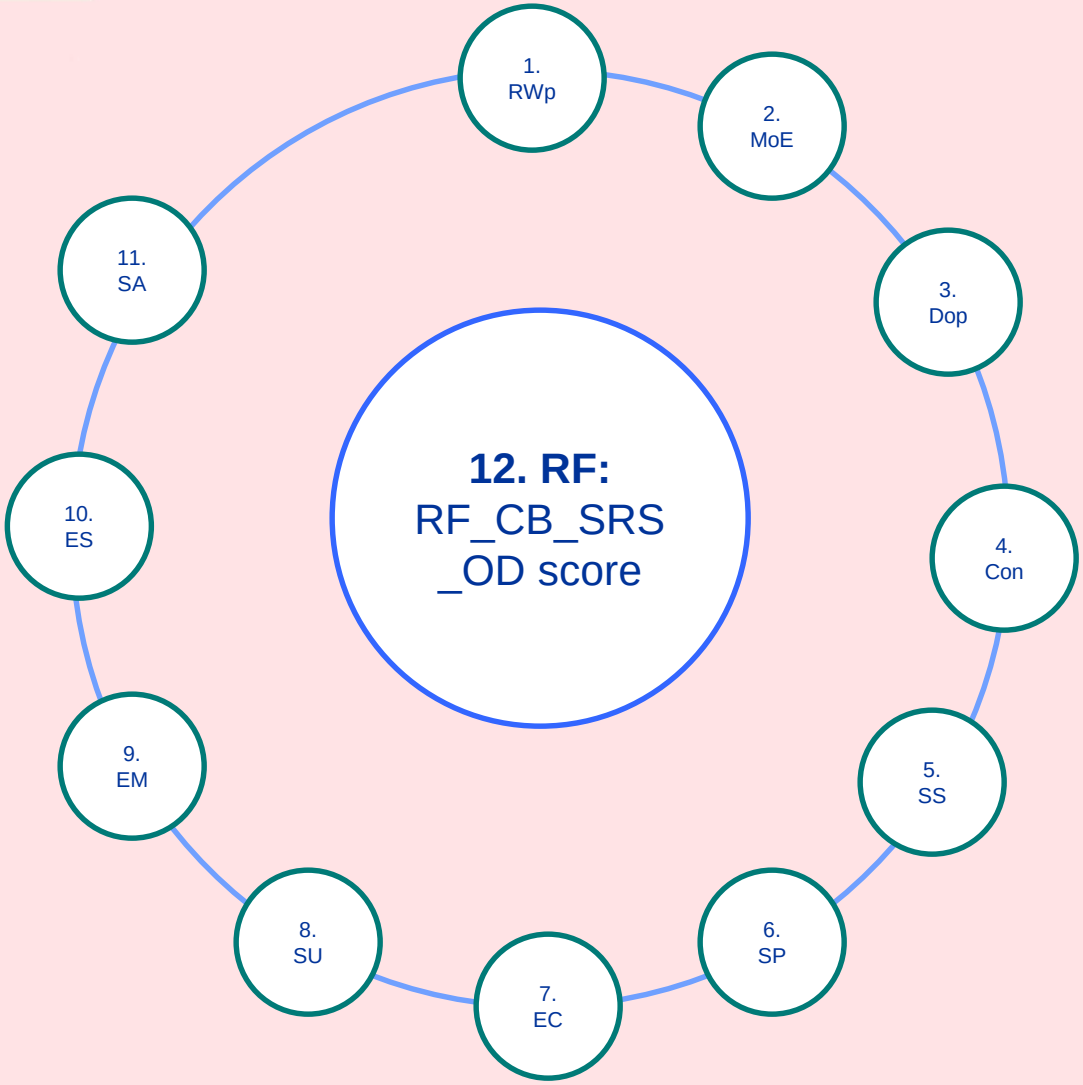


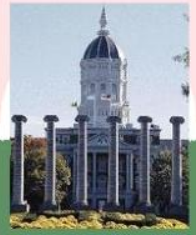

1. RWp
2. MoE
3. Dop
4. Con
5. SS
6. SP
7. EC
8. SU
9. EM
11. SA
12. RF
10. ES: Exposed surface score
$S_{ES} = \frac{\sum Exposed\ area_i}{\sum Area_i}$

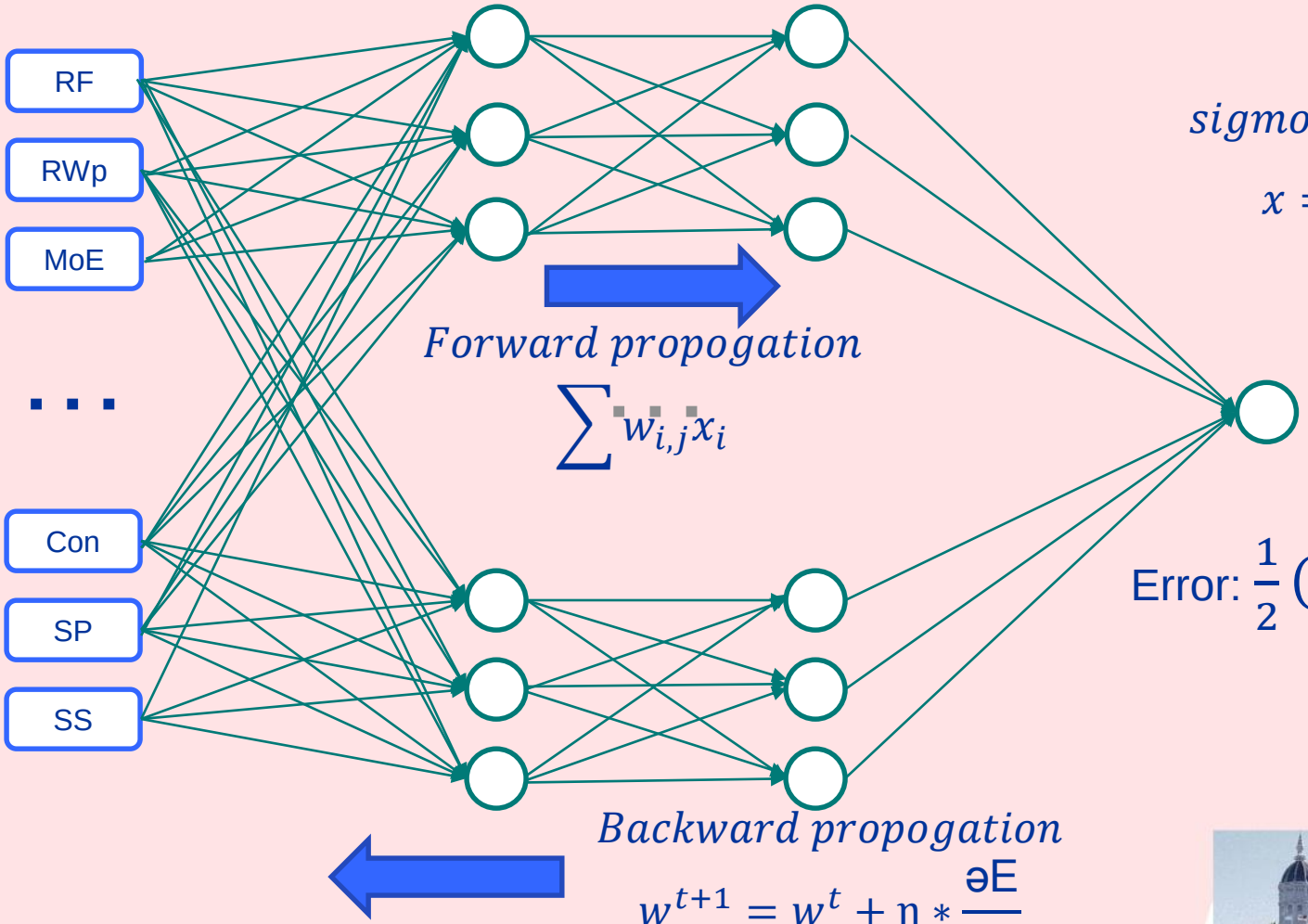



1.
RWp
2.
MoE
3.
Dop
4.
Con
5.
SS
6.
SP
7.
EC
8.
SU
9.
EM
10.
ES
12.
RF
11. SA:
Solvent
accessibility
similarity
score




1.
RWp
2.
MoE
3.
Dop
4.
Con
5.
SS
6.
SP
7.
EC
8.
SU
9.
EM
10.
ES
11.
SA
12. RF:
RF_CB_SRS
_OD score

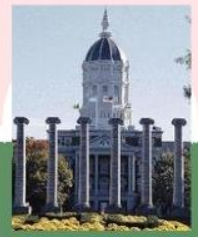




$sigmoid\ function$

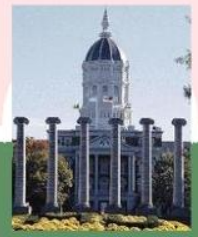
$$x = \frac{1}{1+e^x}$$

Error: $\frac{1}{2}(y-t)^2$



Outline

- Part I: Introduction
  - Protein quality assessment
  - CASP competition
- Part II: QAcon method
- **Part III: Result**
- Part IV: Conclusion

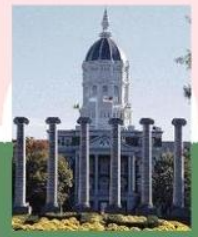


Mizzou
University of Missouri

Part III: Result

Table 1. The per-target average correlation, average loss for QAcon and other methods on sel20 of CASP11.

Server name	Ave. corr.	Ave. loss
ProQ2	**0.643**	**0.090**
QAcon	0.639	0.100
VoroMQA	0.561	0.108
Wang_SVM	0.655	0.109
Wang_deep_1	0.613	0.128
RWplus	0.536	0.135
raghavagps-qaspro	0.35	0.156

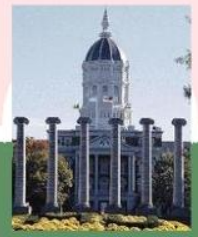


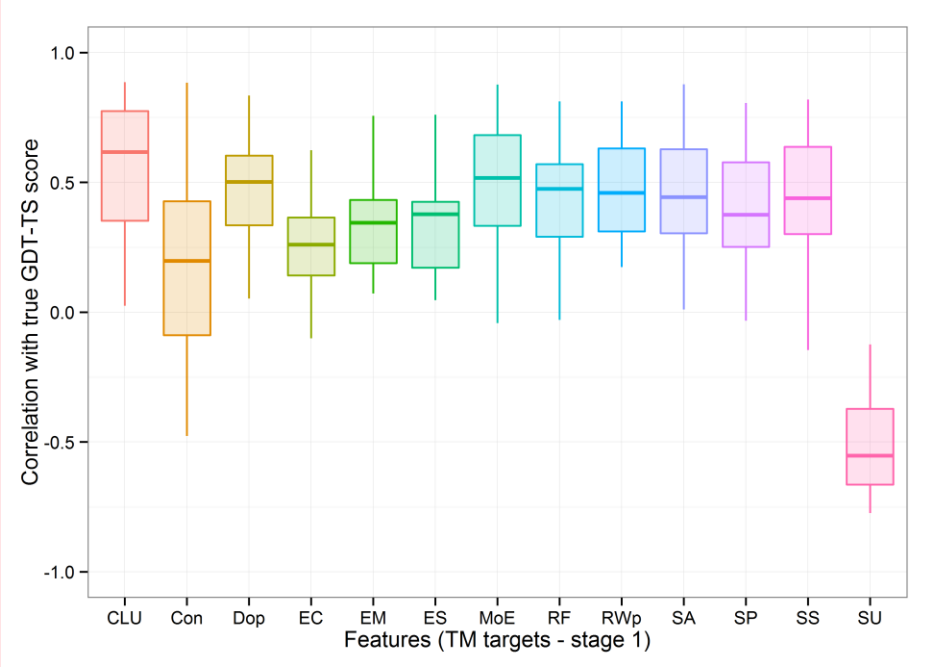
Mizzou
University of Missouri

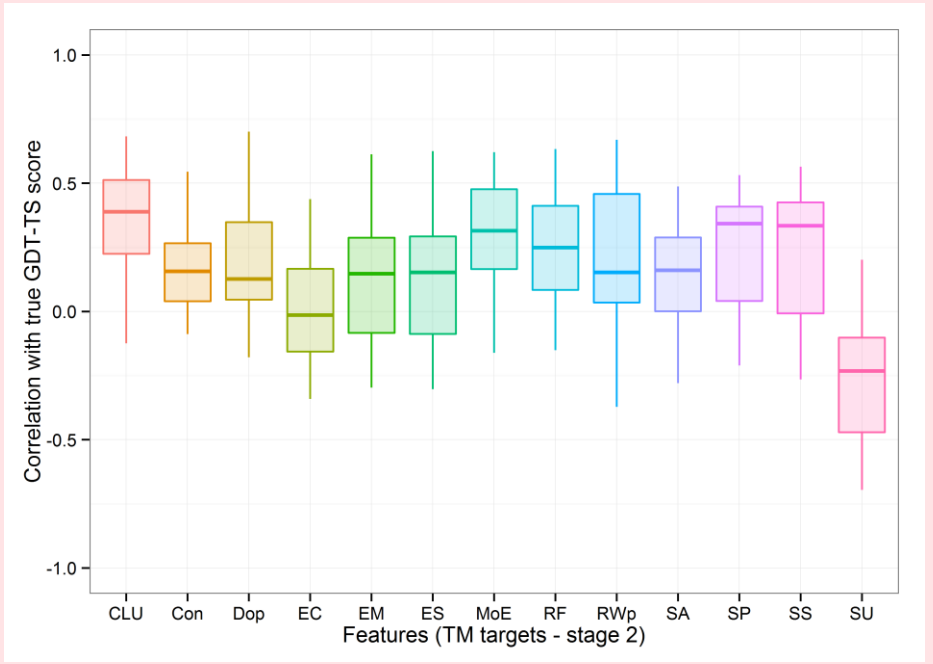
Part III: Result

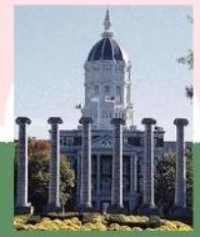
Table 2. The per-target average correlation, average loss for QAcon and other methods on top150 of CASP11.

Server name	Ave. corr.	Ave. loss
ProQ2	0.372	**0.058**
QAcon	**0.395**	0.067
VoroMQA	0.401	0.069
RWplus	0.295	0.084
Wang_SVM	0.362	0.085
raghavagps-qaspro	0.222	0.085
Wang_deep_1	0.302	0.089




1.0
0.5
0.0
-0.5
-1.0
Correlation with true GDT-TS score
CLU
Con
Dop
EC
EM
ES
MoE
RF
RWp
SA
SP
SS
SU
Features (TM targets - stage 1)


1.0
0.5
0.0
-0.5
-1.0
Correlation with true GDT-TS score
CLU
Con
Dop
EC
EM
ES
MoE
RF
RWp
SA
SP
SS
SU
Features (TM targets - stage 2)



Mizzou
University of Missouri

Part III: Result

Table 3.Contact satisfaction score of all CASP11 native structures (top15)

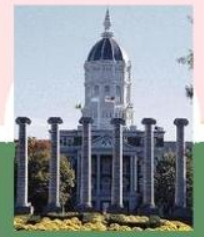
Target name	Contact satisfaction
T0778	0.6142
T0825	0.6049
T0807	0.5387
T0815	0.5189
T0817	0.5181
T0811	0.5176
T0854	0.4953
T0762	0.4607
T0819	0.4531
T0768	0.4529
T0776	0.4492
T0798	0.4343
T0805	0.4252
T0801	0.3936
T0847	0.3864

Table 4. The average correlation and loss for CASP11 sel20 targets

Contact satisfaction	Ave. Corr	Ave. Loss
Con (Top 25)	0.682	0.156
Con (Bottom 25)	-0.016	0.233

Table 5. The average correlation and loss for CASP11 top150 targets

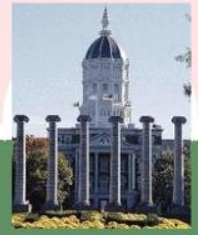
Contact satisfaction	Ave. Corr	Ave. Loss
Con (Top 25)	0.221	0.146
Con (Bottom 25)	0.080	0.134



Mizzou
University of Missouri

Outline

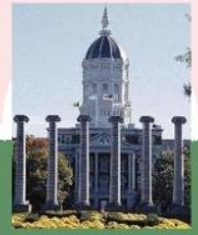
- Part I: Introduction
  - Protein quality assessment
  - CASP competition
- Part II: QAcon method
- Part III: Result
- **Part IV: Conclusion**



Mizzou
University of Missouri

Part IV: Conclusion

- QAcon
- Contact as a potential feature for QA



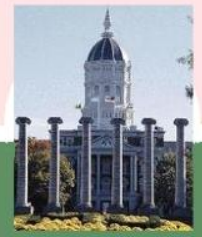
Mizzou
University of Missouri

Acknowledgements

- Badri Adhikari
- Debswapna Bhattacharya
- Miao Sun
- Jie Hou
- All other lab members
- Jianlin Cheng

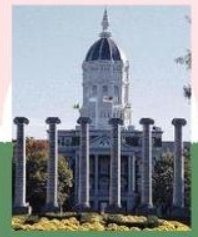
Q & A

Email: rcrg4@mail.missouri.edu



Supplementary

1. The RF_CB_SRS_OD score(Rykunov and Fiser, 2007)	energy score for evaluating the protein structure based on statistical distance dependent pairwise potentials
2. RWplus score(Zhang and Zhang, 2010)	energy score evaluating protein models based on distance-dependent atomic potential
3. ModelEvaluator score(Wang, et al., 2009)	score evaluating protein models based on structural features and support vector machines.



Supplementary

4. Dope score(Shen and Sali, 2006)	energy score evaluating protein models based on the reference state of non-interacting atoms in homogeneous sphere
5. Con score	The contact score is calculated by the satisfaction of contact predicted from the sequence and the one parsed from the model. PSI-COV is used for contact prediction, and the NNcon is used when PSI-COV fails to make predictions.
6. SS score	This score is calculated by the difference between secondary structure predicted by Spine X (Faraggi, et al., 2012) from the protein sequence and those of a model parsed by DSSP (Kabsch and Sander, 1983).

Supplementary

7. SP score

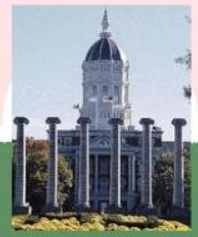
This score is calculated by the percentage of helix and sheet matching between secondary structure predicted and the on parsed from the model

8. EC score

The Euclidian compact score is calculated by summation of pairwise Euclidean distance between amino acids divided by (N*N-1)*3.8, N is the total number of amino acids in the sequence

9. SU score

This surface score is calculated by the total area of exposed nonpolar residues divided by the total area of all residues



Supplementary

10. EM score

The exposed mass score is calculated as the total mass of nonpolar residues area divided by the total mass of exposed residue area

11. ES score

The exposed surface score is calculated as the total exposed residue area divided by the total residue area.

12. SA score

The solvent accessibility score is calculated by the percentage of difference between the predicted solvent accessibility and the one parsed from the model.