Protein quality assessment

Speaker: Renzhi Cao

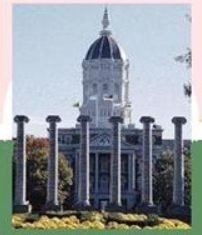
Advisor: Dr. Jianlin Cheng

Major: Computer Science

May 17th, 2013

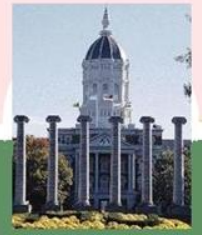
Outline

- Introduction
- Paper1
- Paper2
- Paper3
- Discussion and research plan
- Acknowledgement and references



Outline

- Introduction
- Paper1
- Paper2
- Paper3
- Discussion and research plan
- Acknowledgement and references

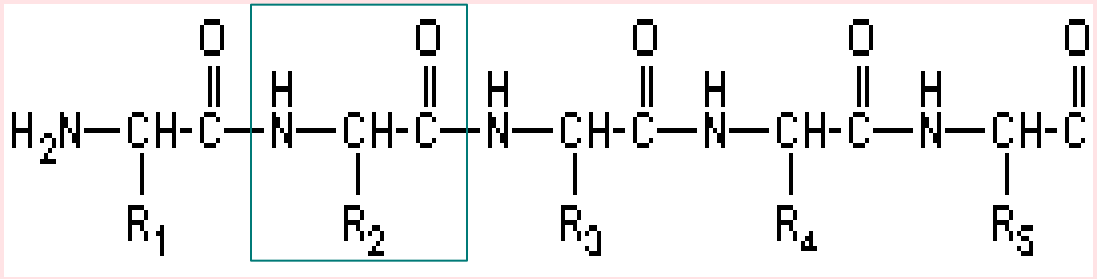


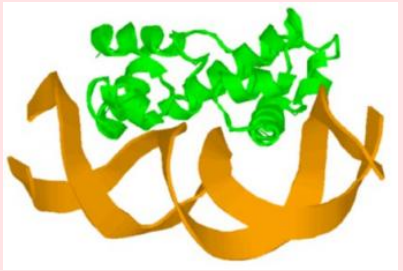
Introduction

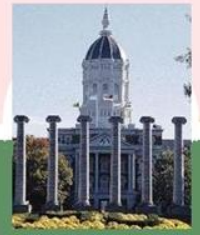
- What is protein?
- Food?

- Protein are composed of small units (amino acid) and can fold into 3D structure.

EIGNIISDAM KKVGRKGVIT



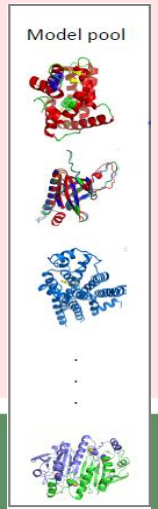




Introduction

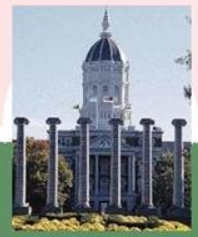
- **What is CASP ?**
- CASP is Critical Assessment of Techniques of Protein Structure Prediction.

- **What is protein quality assessment?**
- Evaluating the quality of protein structure prediction without knowing the native structure.



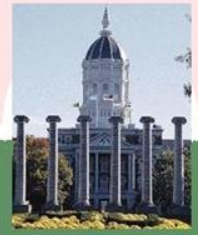
How good is this model?

http://en.wikipedia.org/wiki/CASP



Outline

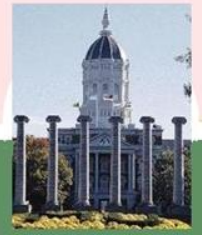
- Introduction
- Paper1
- Paper2
- Paper3
- Discussion and research plan
- Acknowledgement and references



Paper 1

- **A simple and efficient statistical potential for scoring ensembles of protein structures**
- Pilar Cossio, Daniele Granata, Alessandro Laio, Flavio Seno & Antonio Trovato.

- Basic idea: develop a new statistical knowledge based potential (KBP) and apply it to protein quality assessment.
- KBPs are energy functions derived from databases of known protein conformations.



Tanaka, S. & Scheraga,*Macromolecules*,1976

Paper 1 - method

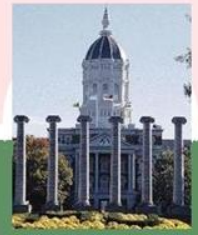
- **Method:**
- The BACH energy function:

$$E_{\text{Bach}} = pE_{\text{PAIR}} + E_{\text{SOLV}}$$

The pairwise statistical potential E_{PAIR} is based on classifying all residue pairs within a protein structure in five different structural classes.

The solvation statistical potential E_{SOLV} is based on classifying all residues in two different environmental classes.

P is a parameter to adjust the weight.

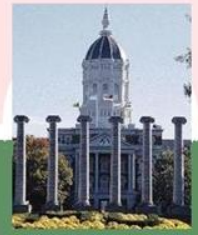


Paper 1 - method

- E_{PAIR}. (Modified DSSP)
- Class 1 : two residues form a α–helical bridge
- Class 2 : two residues form an anti-parallel β-bridge
- Class 3 : two residues form a parallel β-bridge
- Class 4 : two residues in contact(4.5 Ă) through side chain
- Class 5 : other cases

- The pairwise statistical potential E_{PAIR} requires five distinct symmetric matrices $\varepsilon_{ab}{}^{x}$, where a and b vary among the 20 amino acid types, x is the class, for overall 1050 parameters.

$$E_{\mathrm{PAIR}} = \sum_{i<j} \epsilon_{a_i a_j}^{x_{ij}}$$

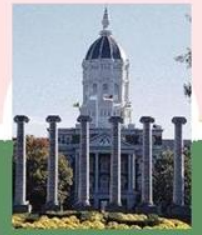


Kabsch, W.&Sander. *Biopolymers*, 1983

Paper 1 - method

$$\epsilon_{ab}^{x} = -\ln\left[\frac{\frac{n_{ab}^{x}}{\sum_{x} n_{ab}^{x}}}{\frac{\sum_{ab} n_{ab}^{x}}{\sum_{x}\sum_{ab} n_{ab}^{x}}}\right]$$

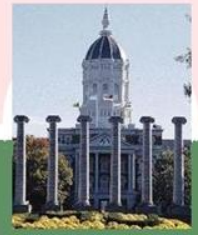
- $n_{ab}{}^{x}$ is the total number of residue pairs of type a and b found in the structural class x within the dataset.



Paper 1 - method

- E_{SOLV}. (SURF tool of VMD graphic software)
- Class 1 : buried
- Class 2 : solvent exposed
- The single residue statistical potential E_{SOLV} requires two separate parameter sets $\lambda_a{}^e$, for overall 40 parameters. e_i=b or s is the environmental class of residue at position i.

$$E_{SOLV} = \sum_i \lambda_{a_i}^{e_i}$$

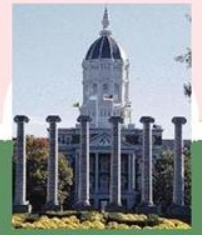


Varshney, and etc. *IEEE computer graphycs and application*. 1994

Paper 1 - method

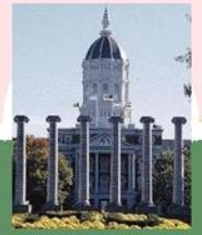
$$\lambda_a^e = -\ln\left[\frac{\frac{m_a^e}{\sum_e m_a^e}}{\frac{\sum_a m_a^e}{\sum_e \sum_a m_a^e}}\right]$$

- $m_a{}^e$, is the total number of residues of type a found in the environment class e within the dataset.



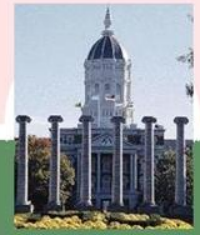
Paper 1 - method

- An alternative implementation of BACH was derived using a reduced amino acid alphabet consisting of 9 classes:
- small hydrophobic (ALA,VAL,ILE,LEU,MET),
- large hydrophobic (TYR,TRP,PHE)
- small polar (SER,THR)
- large polar (ASN,GLN,HIS)
- positively charged (ARG,LYS)
- negatively charged (ASP,GLU)
- GLY, PRO, CYS separately on their own



Paper 1 - method

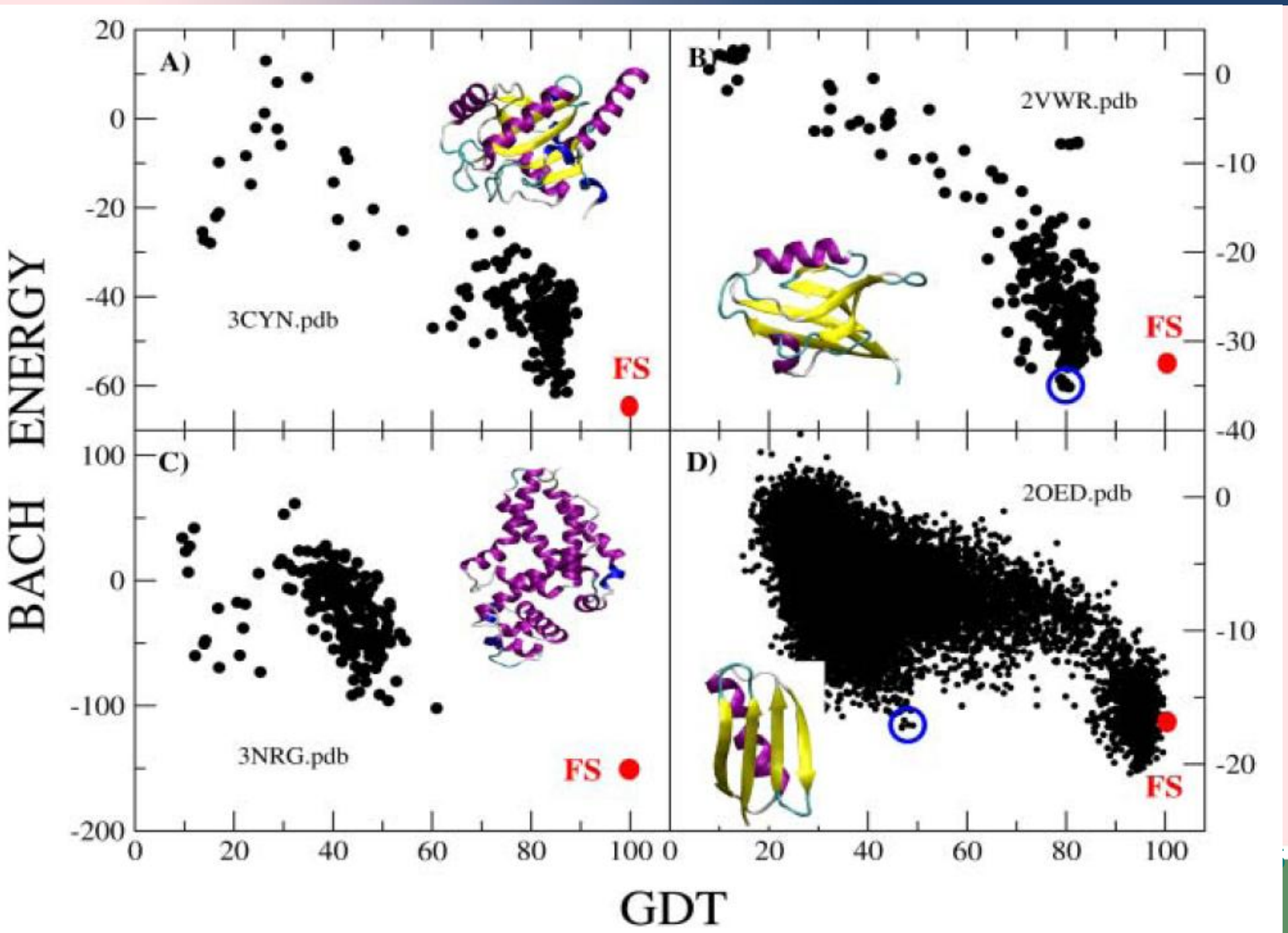
- The parameter p is chosen in such a way that the energy per residue of the two terms has approximately the same standard deviation over the dataset. This criterion gives p = 0.6.
- PDB dataset is the TOP500 database with resolution better than 1.8 Ā by X-ray crystallography (no NMR).
- 33 CASP decoy sets come from CASP8-9. The structures in each decoy set were used if they had the same length and sequence as the native structure, and had all the side-chain and backbone atoms.
- MD simulations were performed using the GROMACS 4.5.3 package.



Lovell, et al. *Proteins*, 2003. Lindahl, et al, *J. Mol. Mod*. 2001

Mizzou
University of Missouri

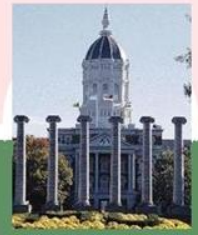
Paper 1 - result



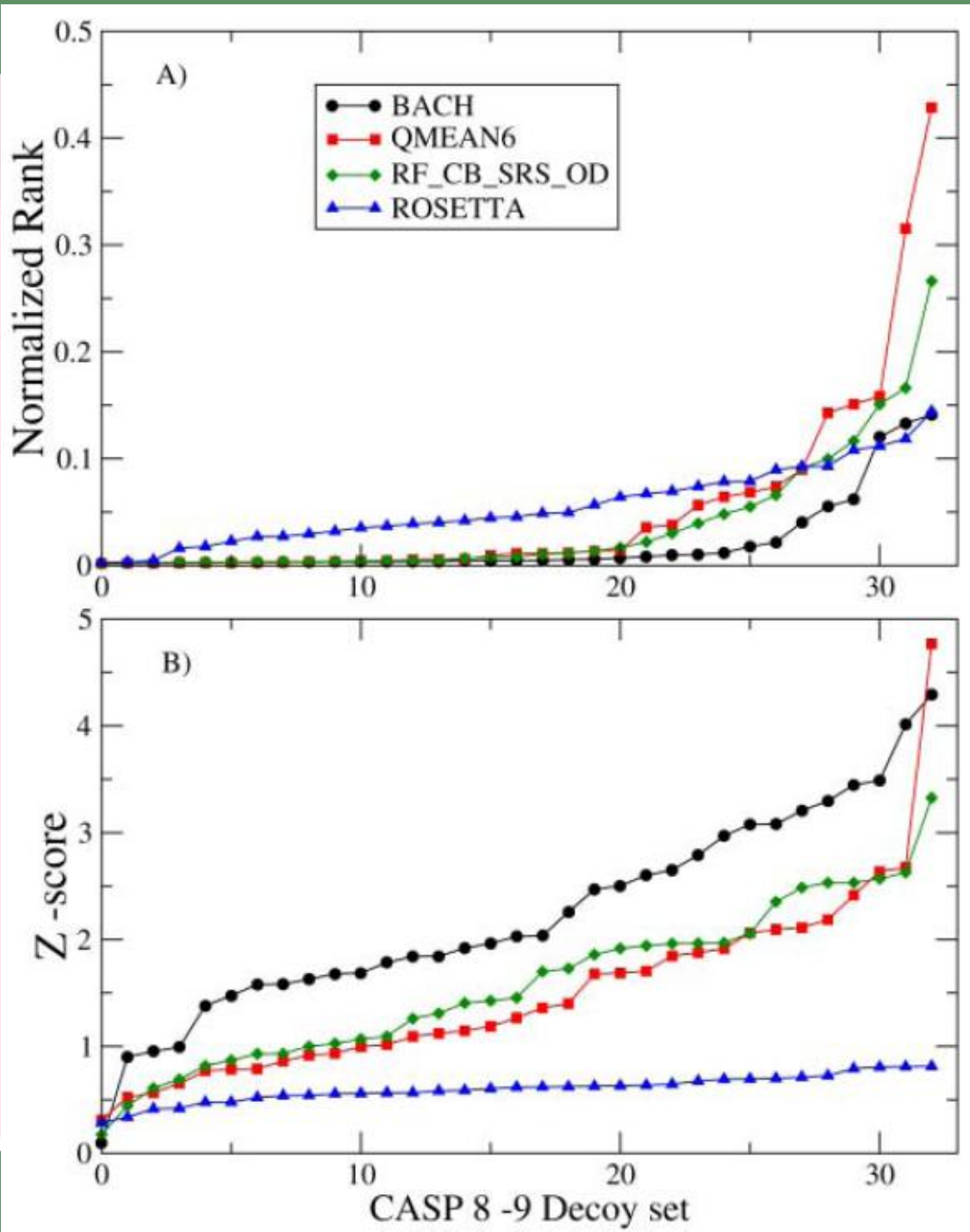
Paper 1 - result

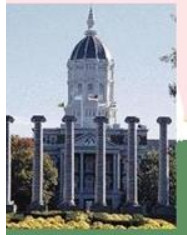
- **Comparison with other knowledge-based potentials.**
- We compare the performance of BACH with QMEAN, ROSETTA and RF_CB_SRS_OD from two aspects:

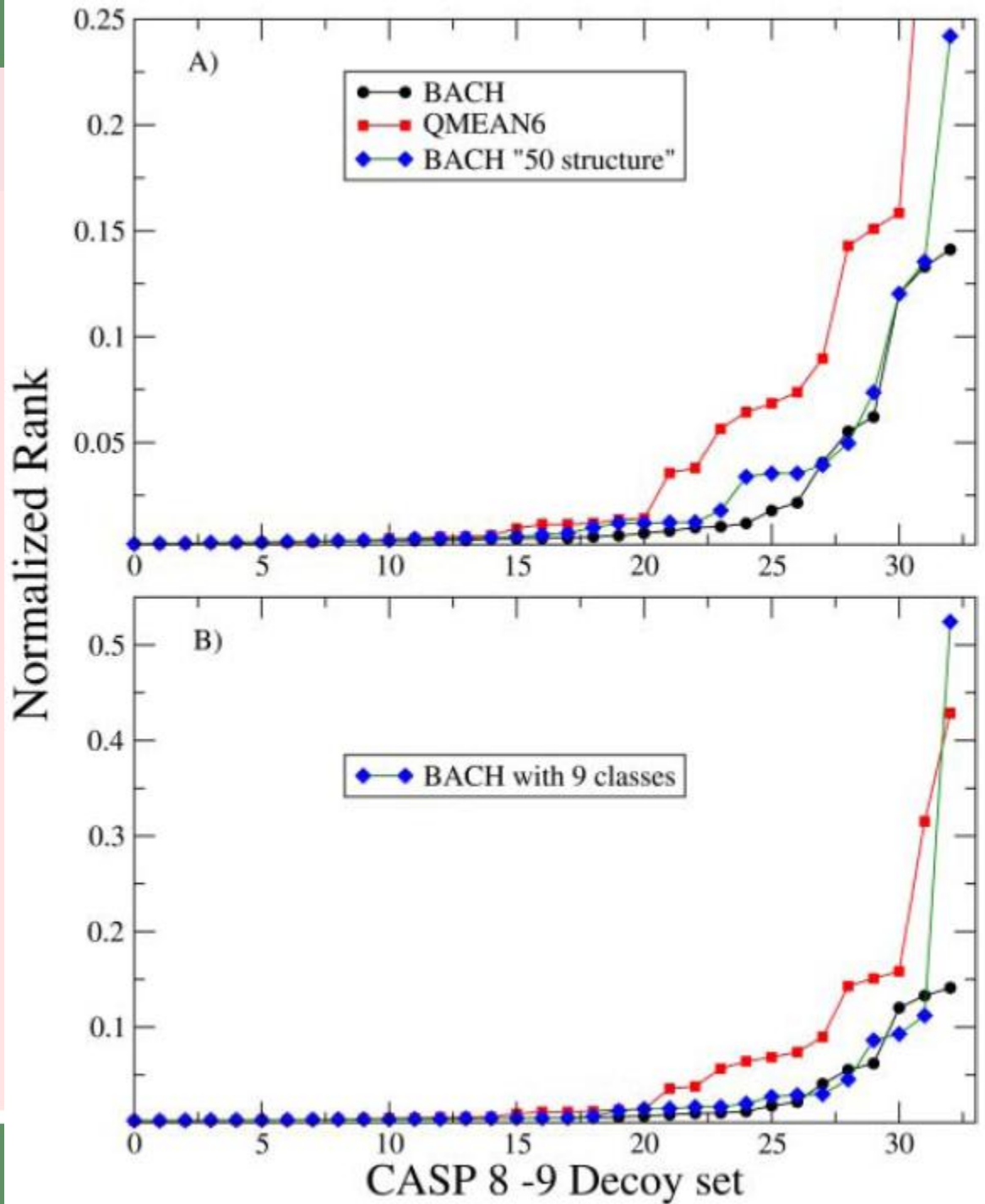
- 1. Normalized rank, defined as the rank of the native structure divided by the total number of structures in the decoy set.
- 2. Z-score, defined as the distance, measured in standard deviations, of the energy of the native state from the mean energy of the set.

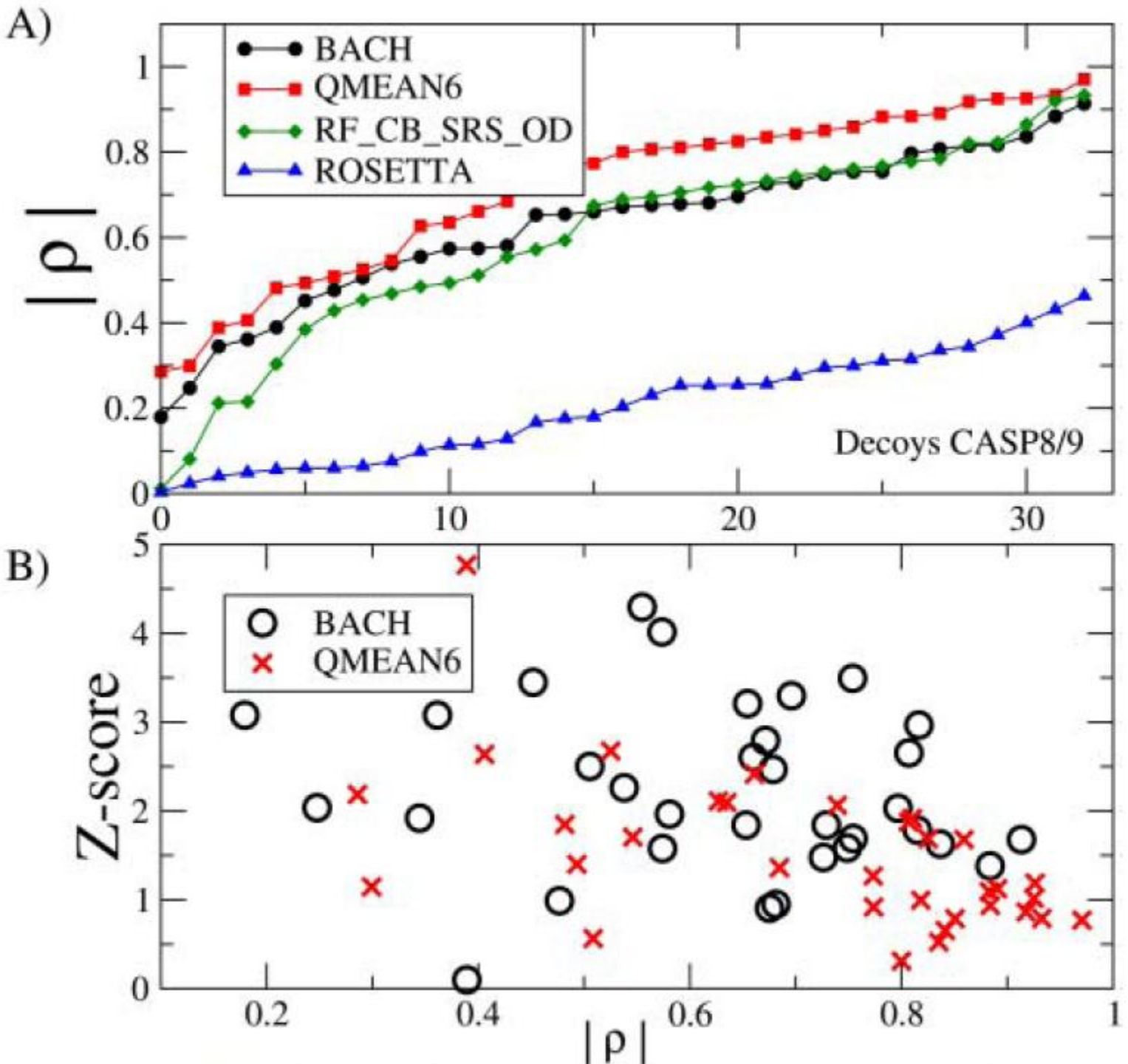


Rykunov et al, *BMC bioinformatics*. 2010


A)
BACH
QMEAN6
RF_CB_SRS_OD
ROSETTA
Normalized Rank
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
B)
Z -score
5
4
3
2
1
0
0
10
20
30
CASP 8 -9 Decoy set

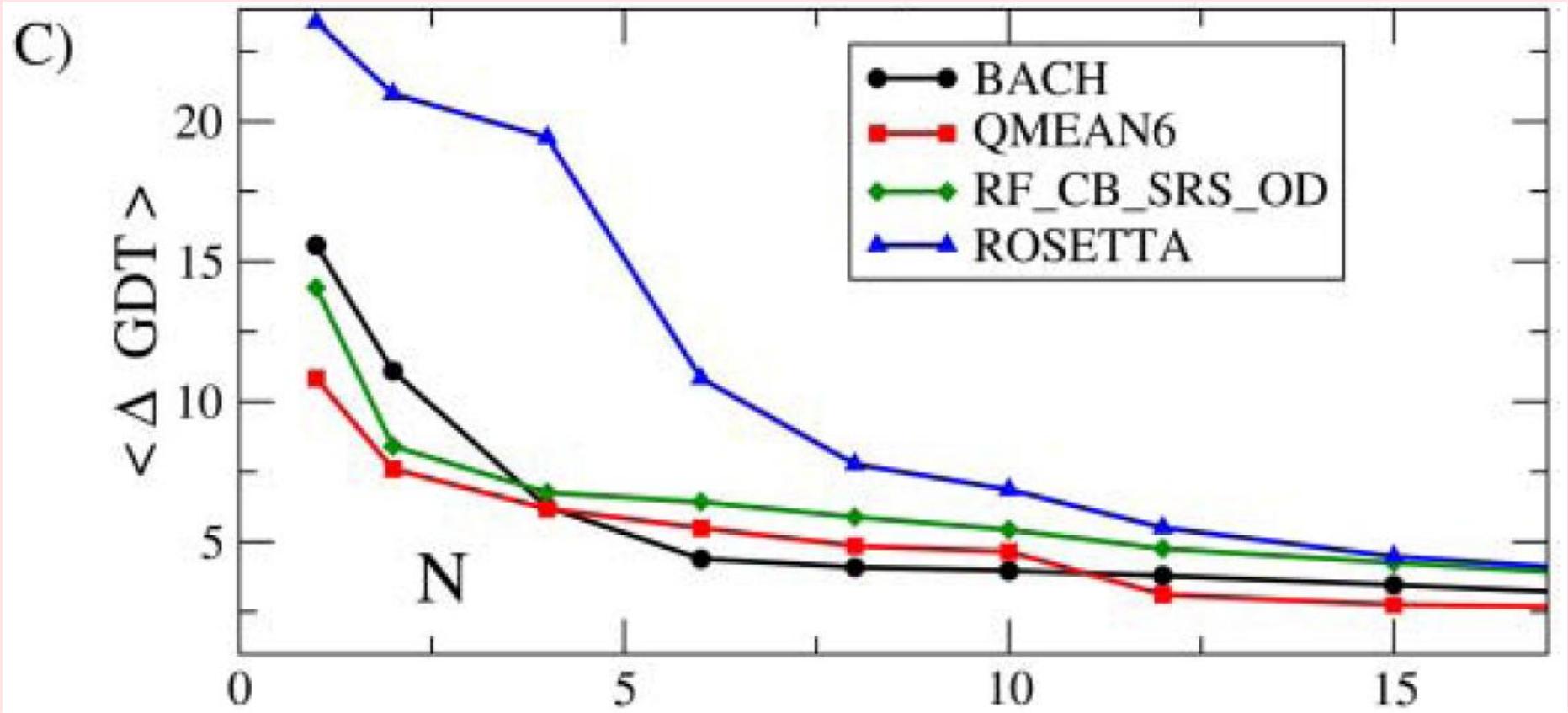



A)
BACH
QMEAN6
BACH "50 structure"
B)
BACH with 9 classes
Normalized Rank
CASP 8 -9 Decoy set

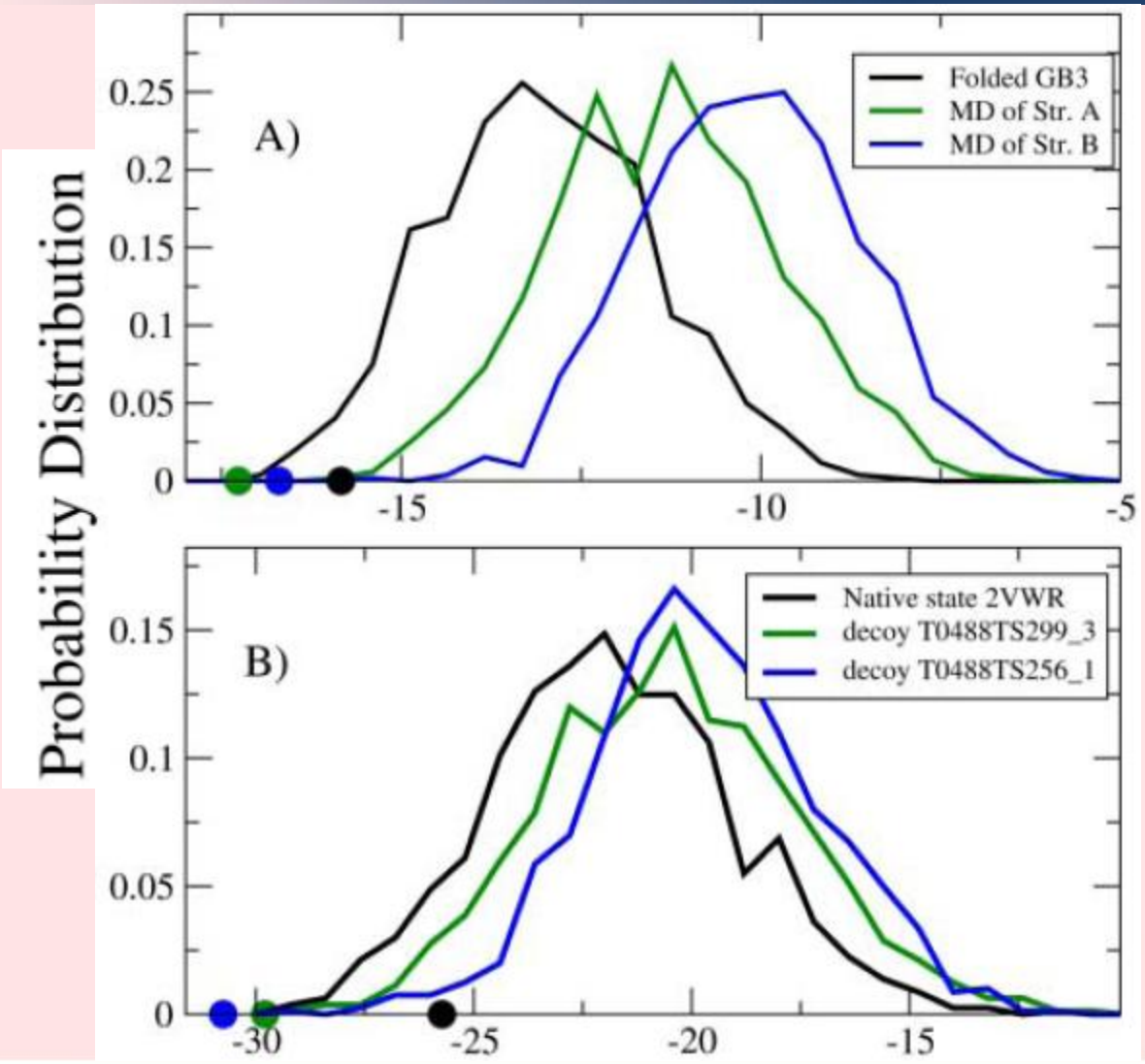

A)
BACH
QMEAN6
RF_CB_SRS_OD
ROSETTA
|ρ|
Decoys CASP8/9
B)
BACH
QMEAN6
Z-score
|ρ|

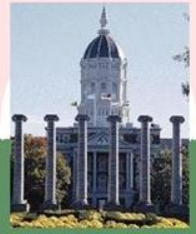
Paper 1 - result

- Δ_NGDT is the GDT score of the best model of N lowest energy structures against best model in the whole dataset.

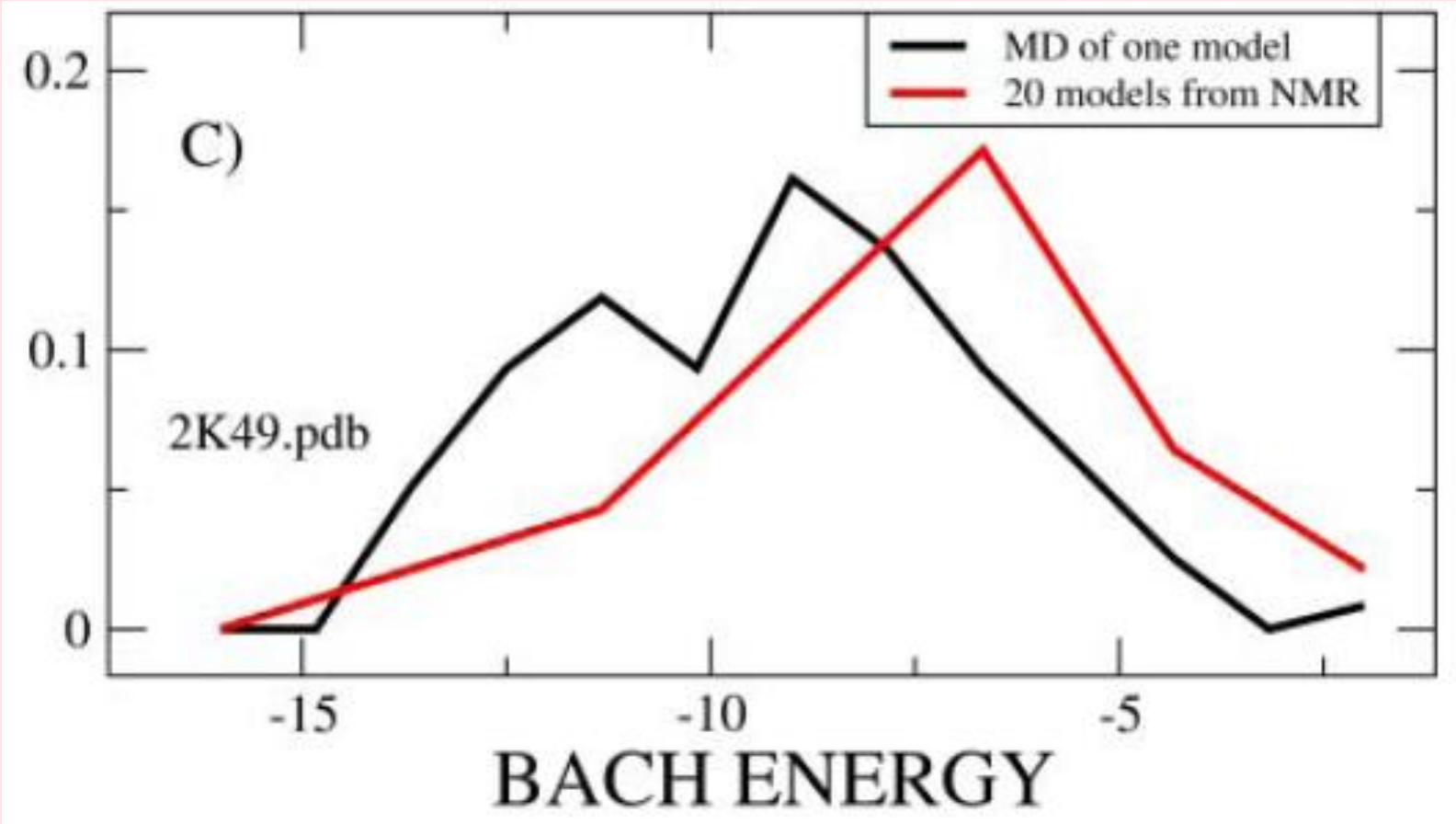


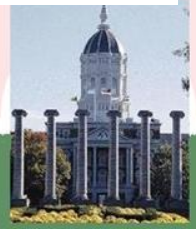
Paper 1 - result





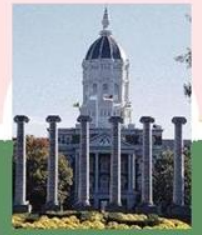
Paper 1 - result





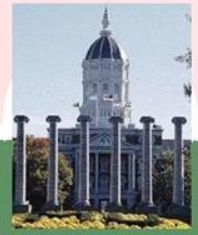
Paper 1 - result

- Discussion:
- This paper developed a knowledge based potential, named BACH, by splitting the residue-residue contact in those present within α–helices or β–sheets, and the evaluation of the propensities of single-residue to be buried or exposed.
- Compared with other state-of-art methods, this one has fewer parameter and perform better in discriminating the native structure, and it's very robust.
- Thermal fluctuation is important to rank two structures.



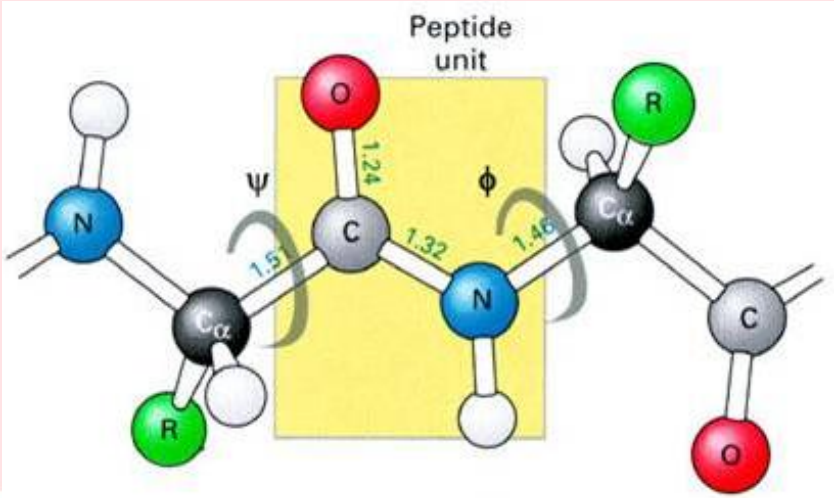
Outline

- Introduction
- Paper1
- Paper2
- Paper3
- Discussion and research plan
- Acknowledgement and references

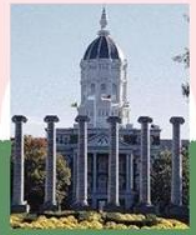


Paper 2

- **A method for evaluating the structural quality of protein models by using higher-order φ–ψ pairs scoring**
- Gregory E. Sims and Sung-Hou Kim.
- Basic idea: evaluating the quality of protein model by higher-order φ–ψ angles.



Φ(phi, involving backbone atoms C'-N-Ca-C')
Ψ(psi, involving backbone atoms N-Ca-C'-N)



Google images.

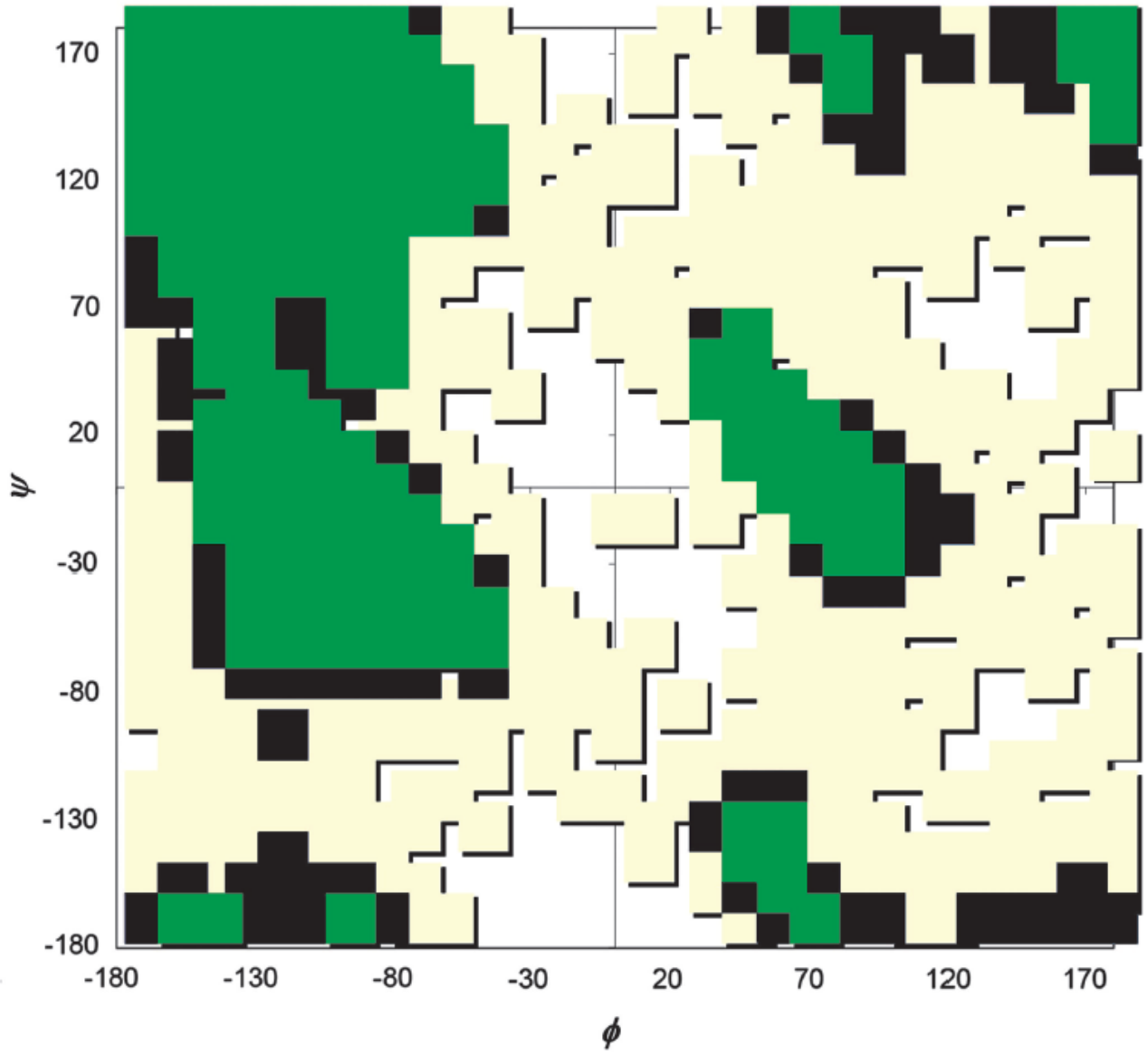
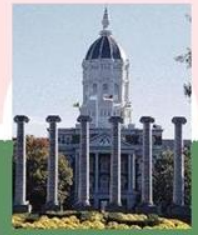


Fig. 1. Ramachandran ϕ–ψ plot. Regions of the ϕ–ψ space are divided into ''core'' favorable regions (green), allowed regions (blue), unfavored regions (tan), and disallowed regions (white). Overall, the plot shows four conformational clusters with their centers around the (ϕ,ψ) values of (−100, −30), (−100, 120), (60, 0), and (60, 180) degrees.

Paper 2 - method

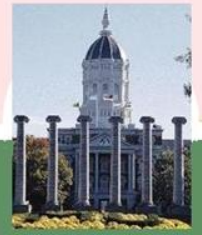
- Problems about using ramachandran plot for protein quality assessment:

- A predicted structure may fit the ramachandran plot very well at single residue level, however, it may composed of very unnatural building blocks consisting of multiple residues.



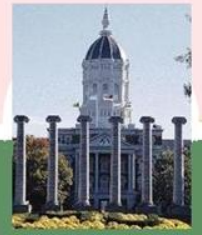
Paper 2 - method

- In this paper, the authors investigate the angular conformation spaces of longer peptide fragment
- 1-10 φ–ψ pairs (3-12 residues).



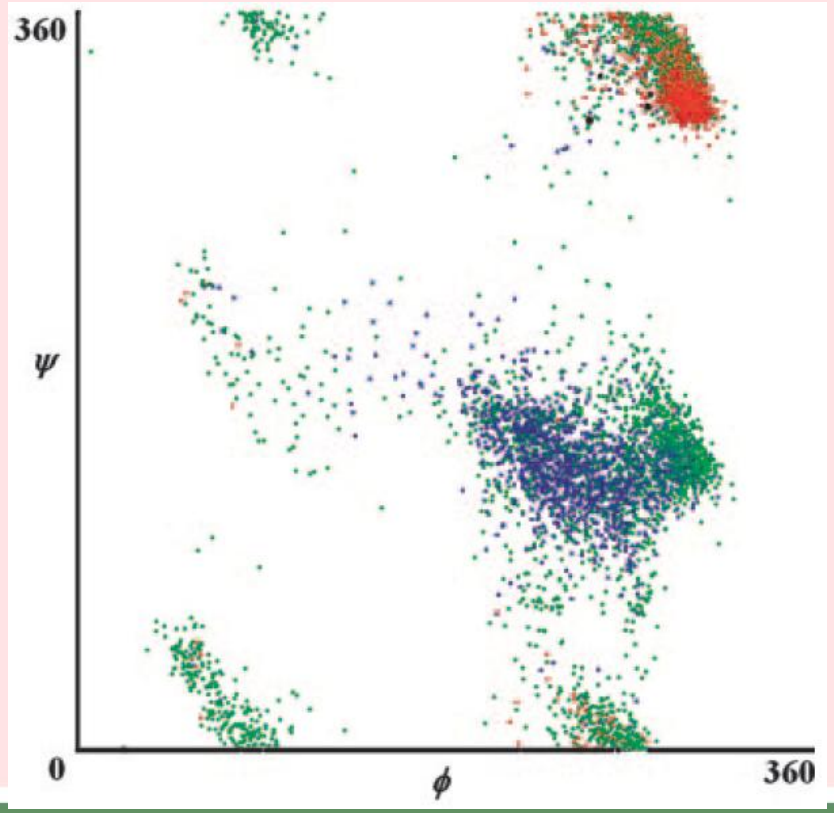
Paper 2 - method

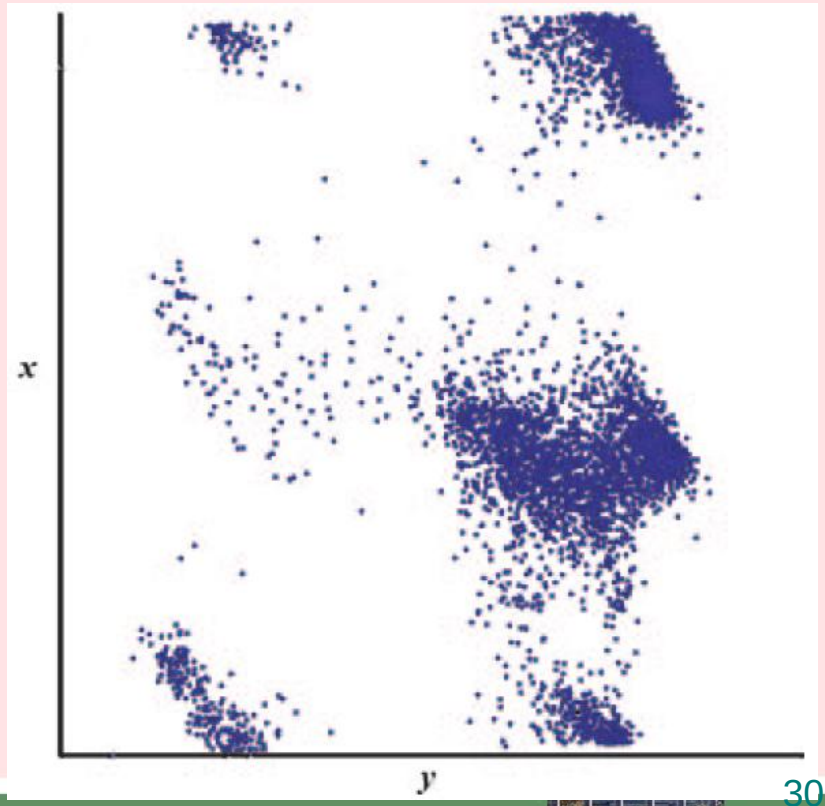
- The observation suggests:
- (1). Protein structure might best be represented as blocks of fragments with designated accessible φ–ψ values
- (2). It maybe possible to construct and delineate a conformational space into a finite number of conformational clusters for a given number of φ–ψ pairs.



Paper 2 - method

- The $(\varphi–\psi)_n$ pairs are mapped to lower dimension using multidimensional scaling(MDS) method.
- Equivalence of φ–ψ map and 2D MDS map.

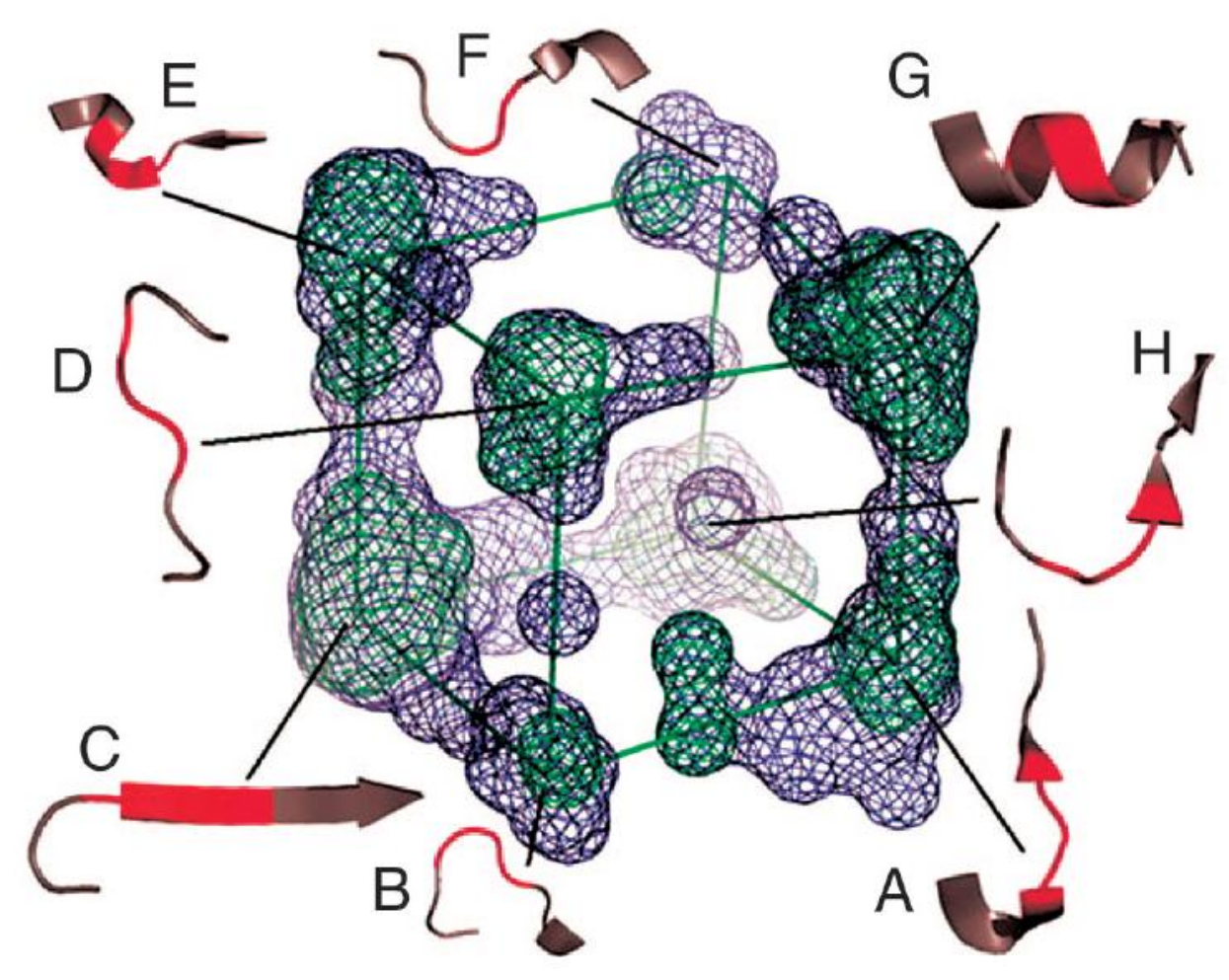




Sims et al, *P.N.A.S.* 2005

Paper 2 - method

- 3D map of conformational space for $(\varphi–\psi)_3$ and representative conformations.



Paper 2 - method

- This paper present a method HOPP score, for defining the conformational space of multiple φ–ψ pairs and testing the fit of queried protein structural models to each of those conformational spaces.

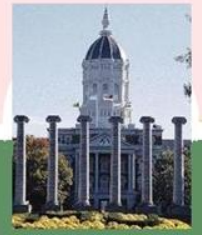
Table 1. HOPPscore allowed regions

Category	Frequency, f	Symbol	Score
Favored	$f >= x + 0.5\sigma$	F	+2
Allowed	$x + 0.5\sigma > f >= x$	A	+1
Unfavored	$x > f$	U	+0.5
Disallowed	$f = 0$	D	−4

x, average frequency; σ, SD.

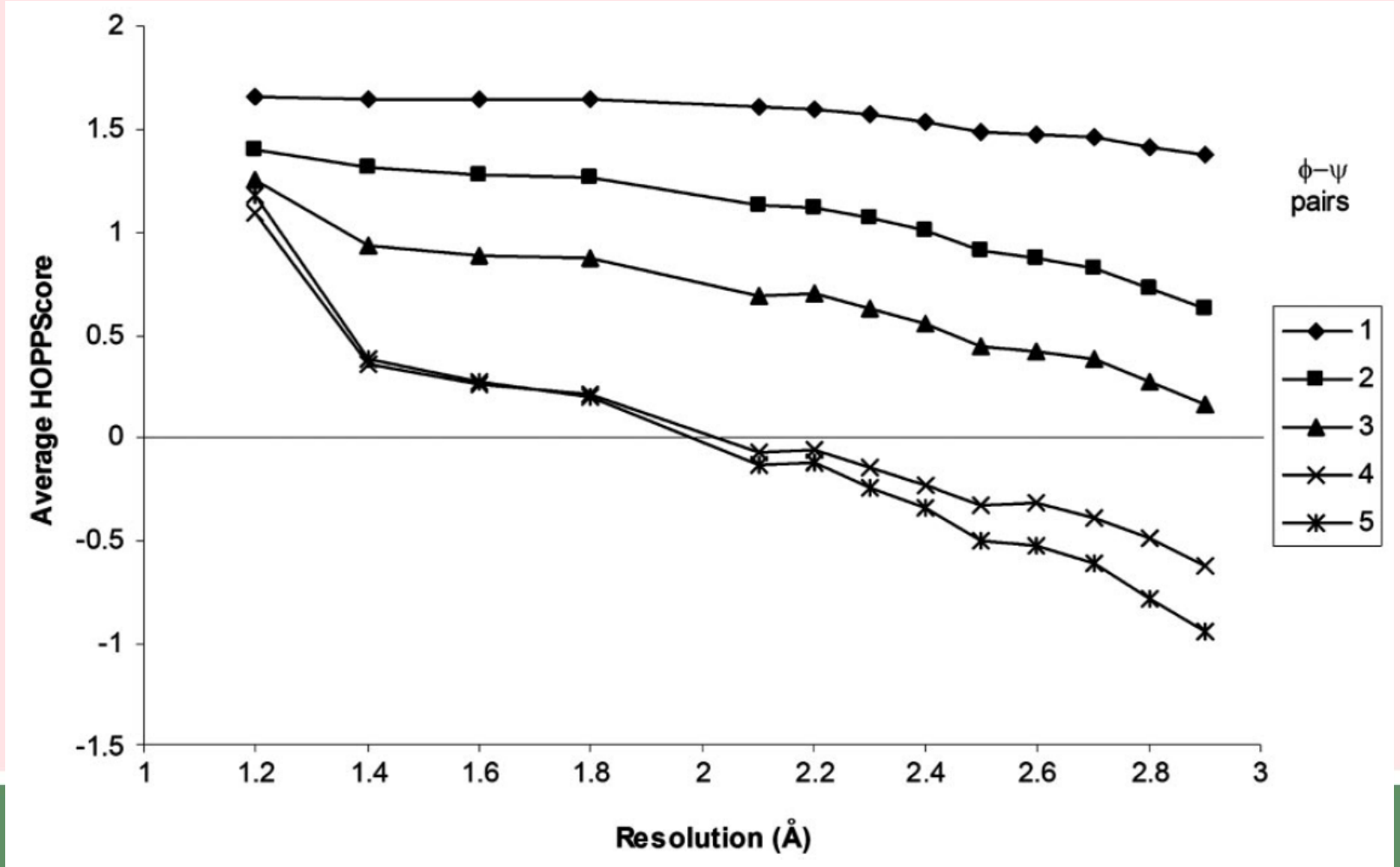
Paper 2 - result

- The HOPPscore database is constructed by all native X-ray structures divided into bins by resolution 0.2 Ã intervals from 0.5 to 3.0 Ã.
- The CASP model database is created from the CASP website.



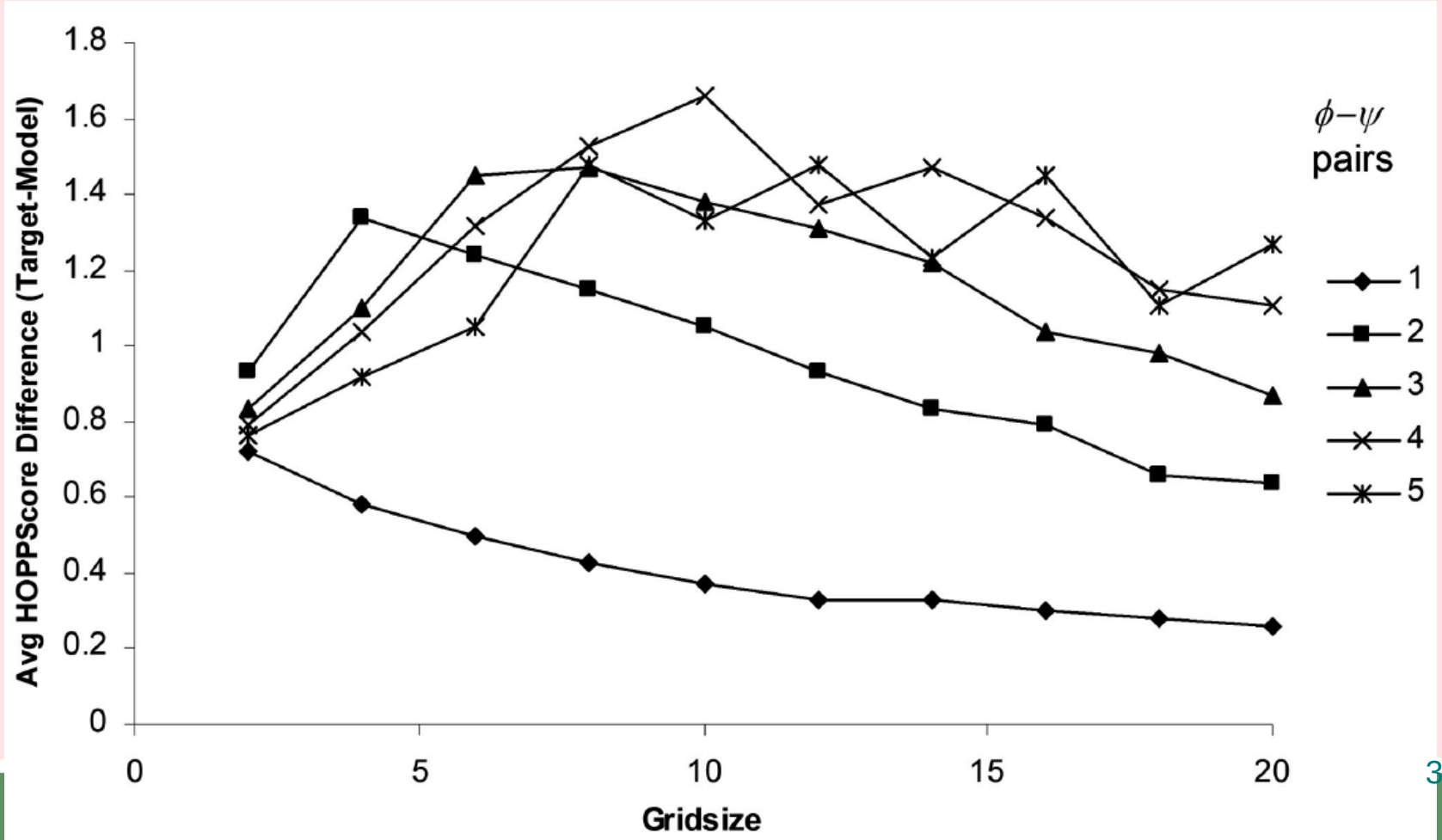
Paper 2 - result

- HOPPscore values correlate with resolution. (gridsize is 12 °)



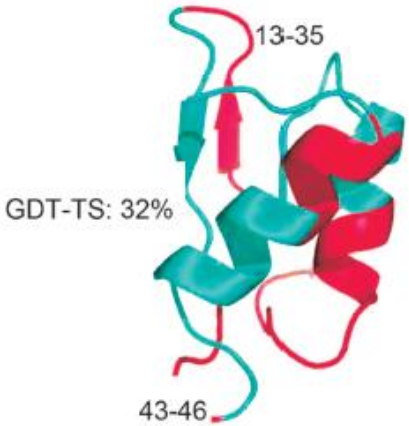
Paper 2 - result

- Best grid size for binning conformational space.

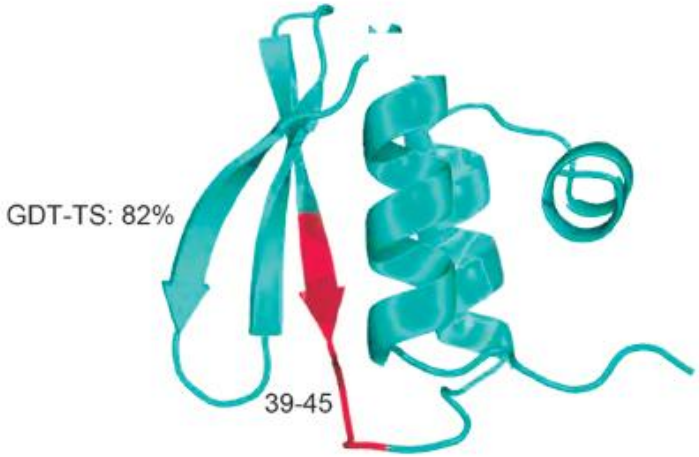


Paper 2 - result

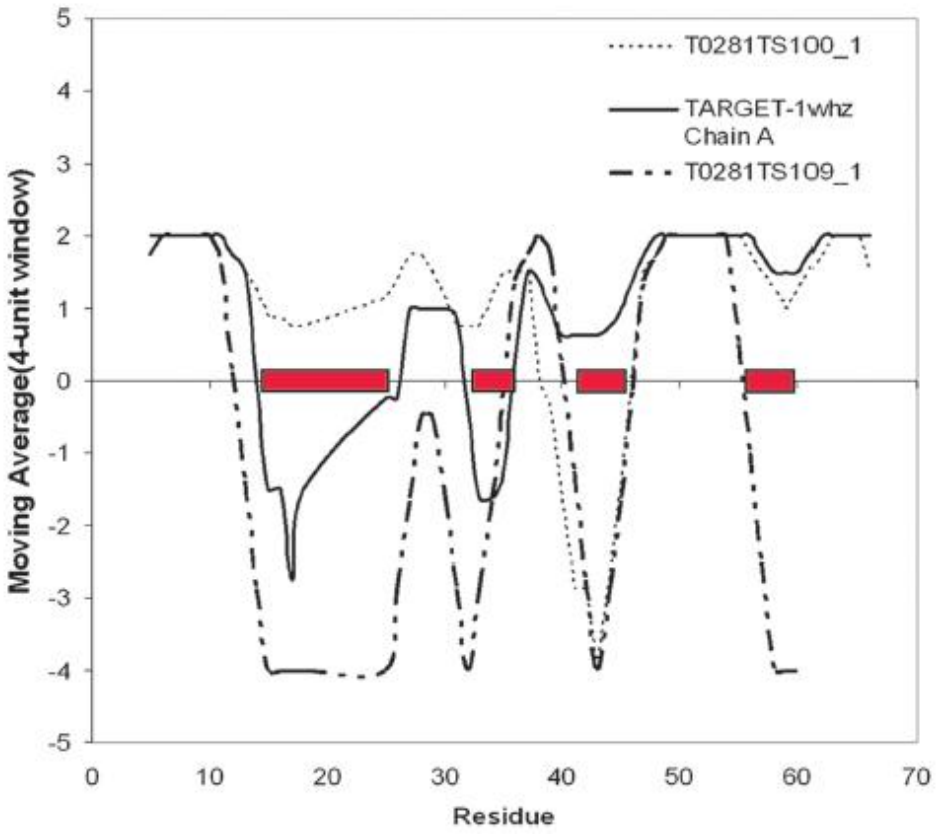
Target: 1whz

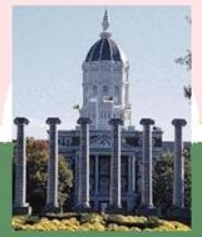


T0281TS109_1

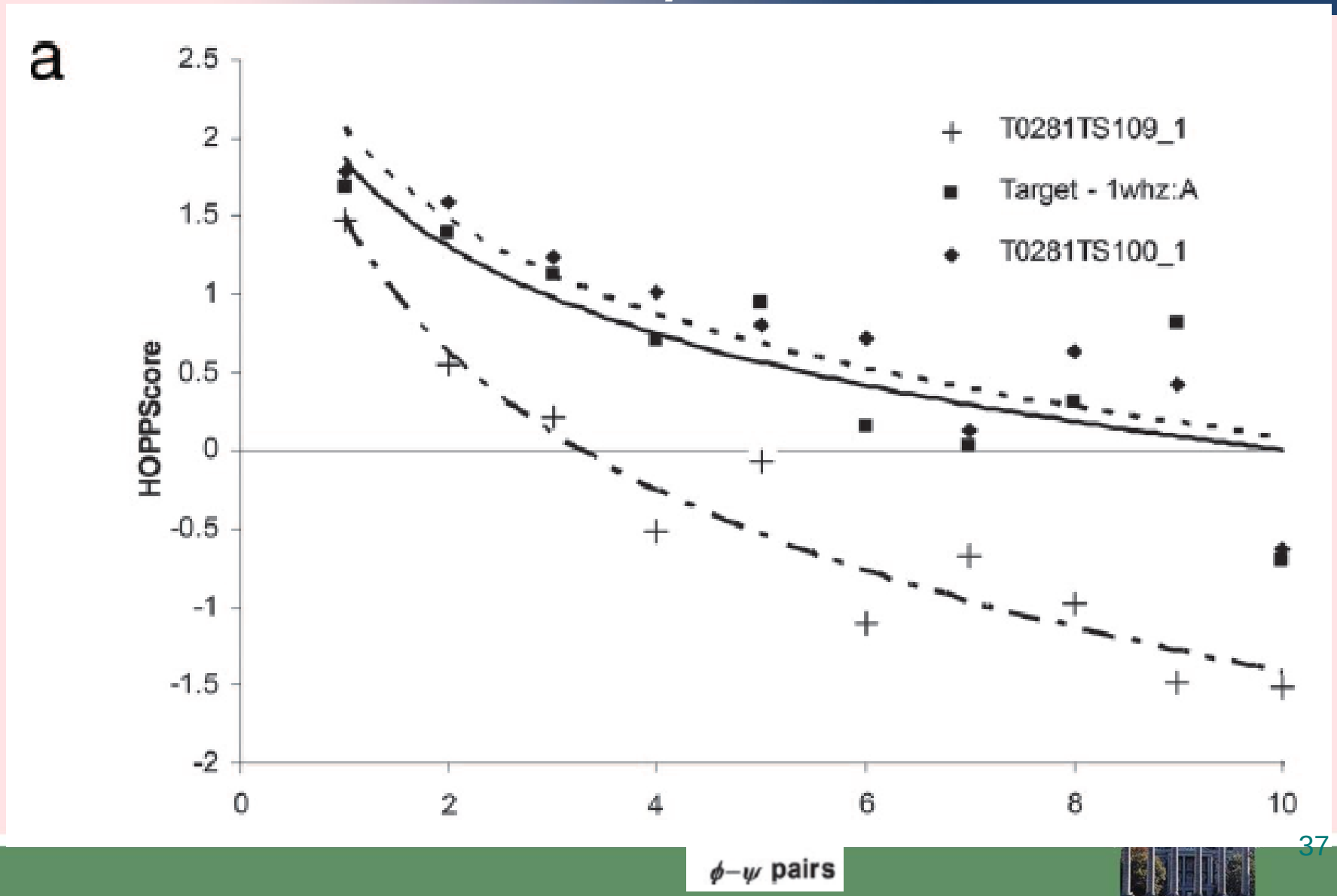


T0281TS100_1



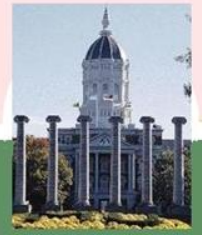


Paper 2 - result



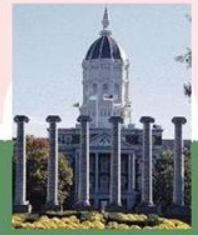
Paper 2 - result

- Discussion:
- This paper developed a tool for protein structure analysis by comparing the higher-order φ–ψ pairs of the experiment and predictions.



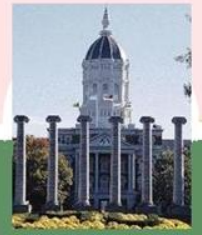
Outline

- Introduction
- Paper1
- Paper2
- Paper3
- Discussion and research plan
- Acknowledgement and references



Paper 3

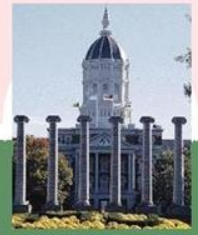
- **Evaluating the absolute quality of a single protein model using structural features and support vector machines**
- Zheng Wang, Allison N. Tegge, and Jianlin Cheng
- Basic idea: apply machine learning method to evaluate the protein quality.



Paper 3

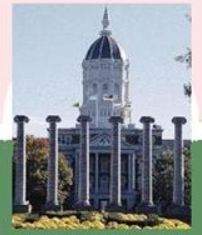
- CASP 6 protein models predicted by Sparks, Robetta and FOLDpro are used as training dataset (64 cross-fold validation are used), CASP 7 protein models are used as testing dataset.

- Support vector machine are used to train a model for predicting the model quality.



Paper 3

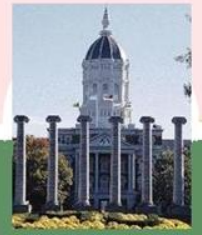
- 1D and 2D structural features include:
- Secondary structure (alpha helix, beta sheet, and loop)
- Relative solvent accessibility (exposed or buried at 25% threshold)
- Contact probability map
- Probability map of beta-strand residue pairs



Paper 3

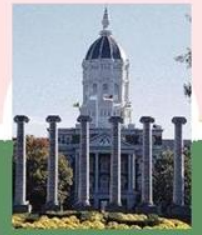
- 1D Features:
- The predicted secondary structure (SS) and relative solvent accessibility (RSA) of each residue are compared with those of the model parsed by DSSP.

- The fraction of identical matches for both SS and RSA.
- Four similarity score by cosine, correlation, Gausian kernal, and dot product of the two composition vectors.



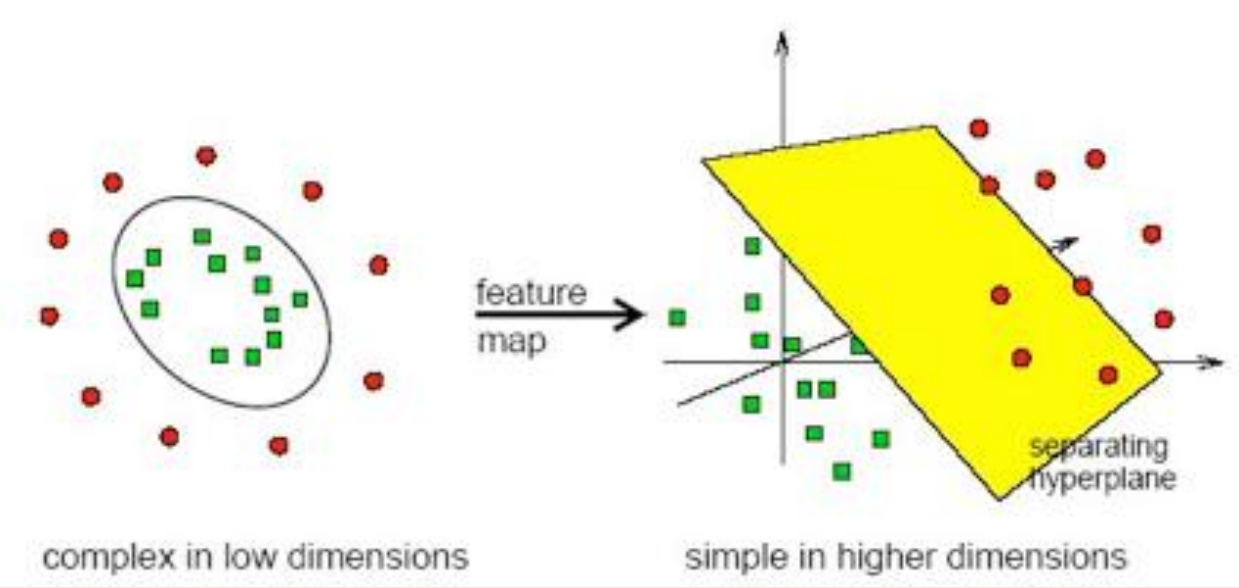
Paper 3

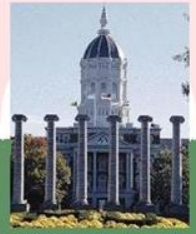
- 2D Features:
- Residue pairs in the model which have sequence separation >= 6, and in contact at a threshold, we use the predicted average contact probability for them as one feature.
- Similarly, for beta-strand pairing probability.
- The contact order (the sum of sequence separation of contacts) and contact number (the number of contacts) for each residue from a 3D model and the predicted contact map are used to calculate the pairwise similarity scores using cosine and correlation functions.



Paper 3

- Support vector machine (SVM-light) are used to train a model for predicting the model quality.

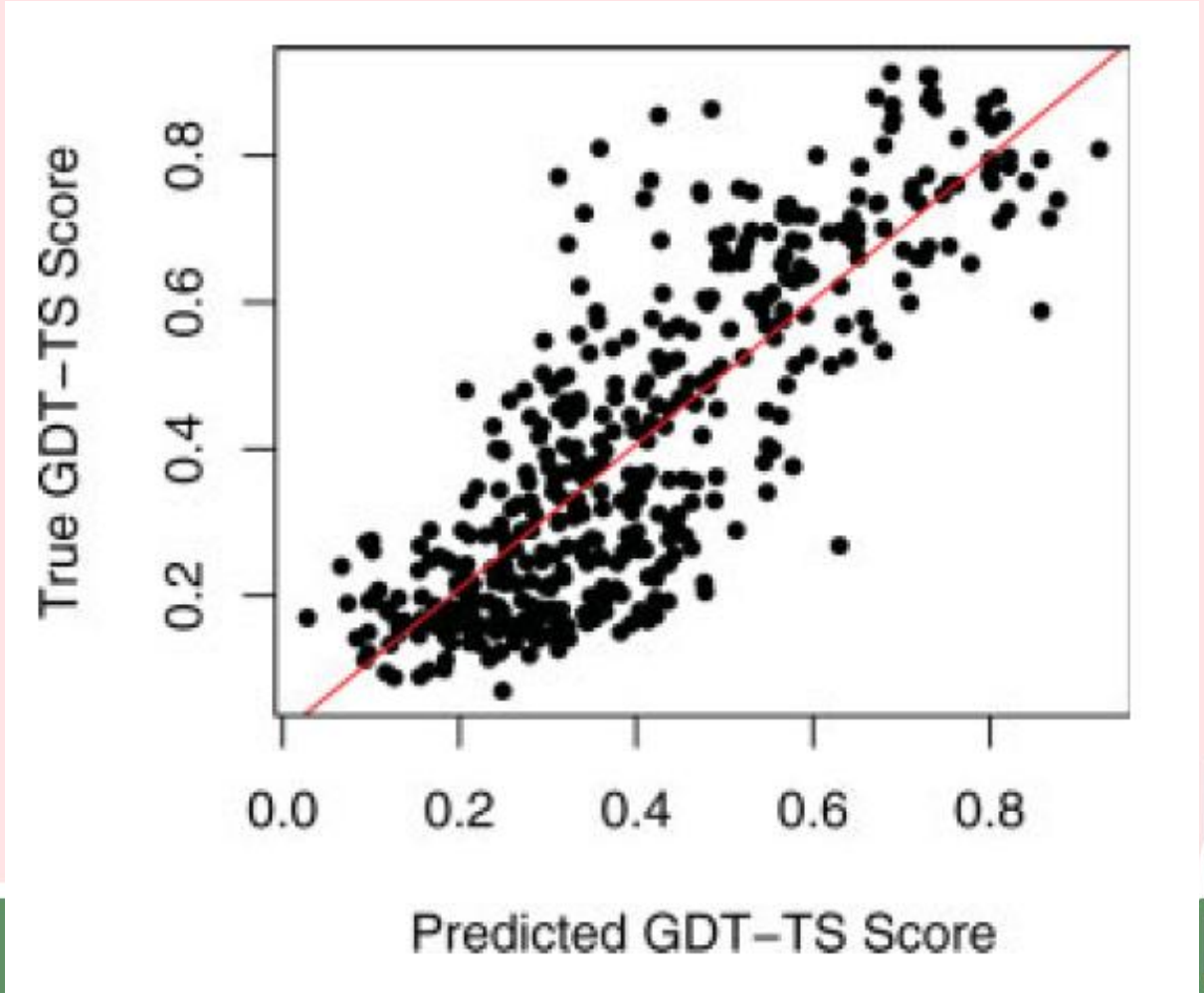


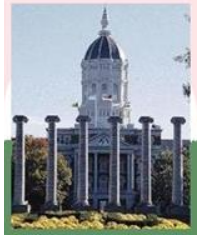


SVM-light : http://svmlight.joachims.org

Paper 3

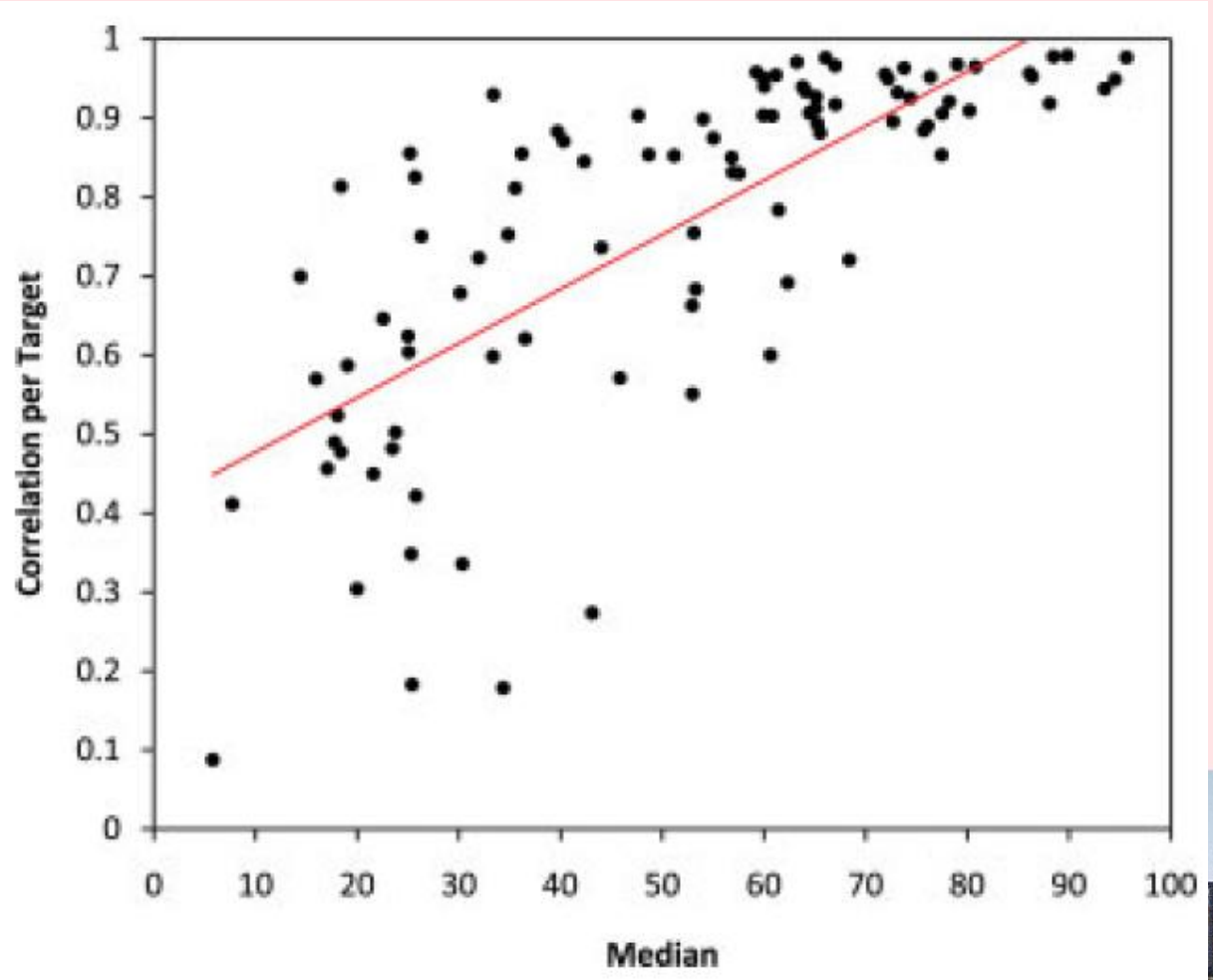
- Predicted GDT-TS score versus real GDT-TS score on CASP6 models using cross-validation.





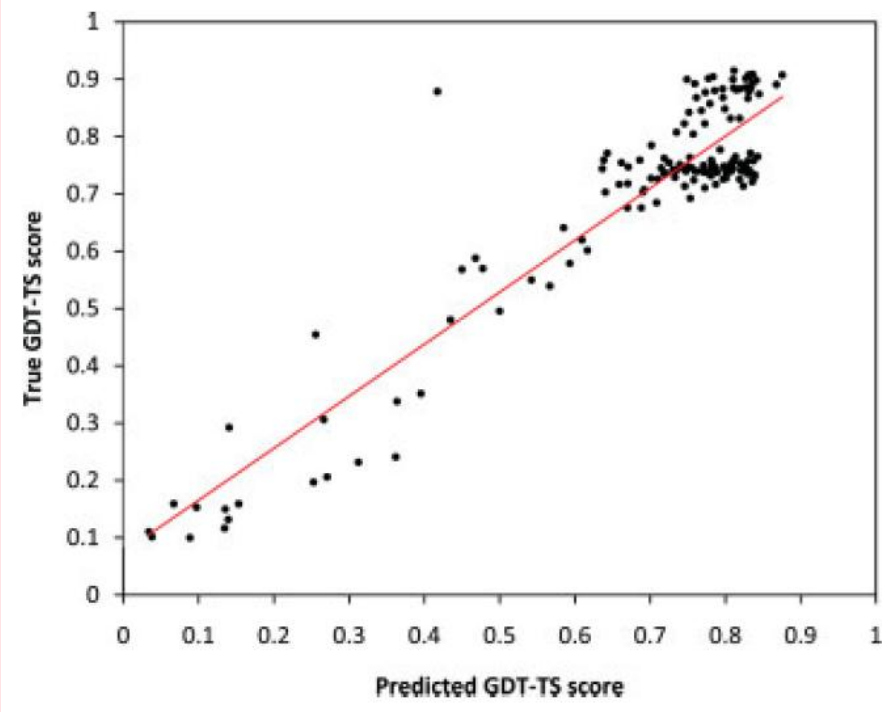
Paper 3

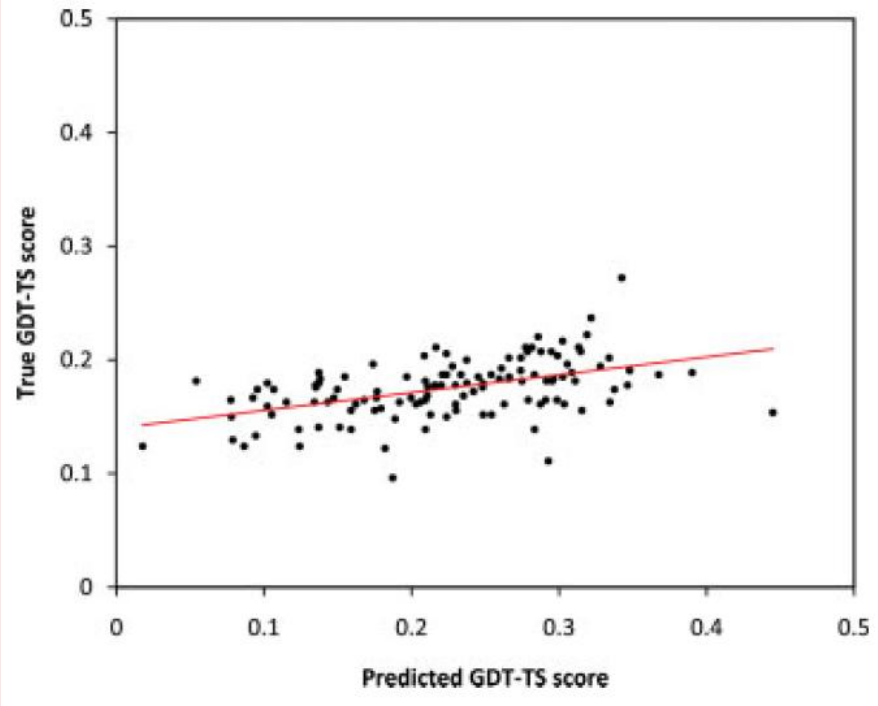
- Correlation against median true GDT-TS score per target.

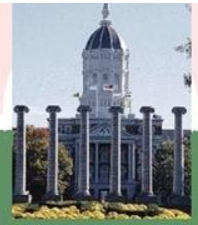


Paper 3

- Predicted GDT-TS score versus true GDT-TS score of easy target T0308 and hard target T0319.

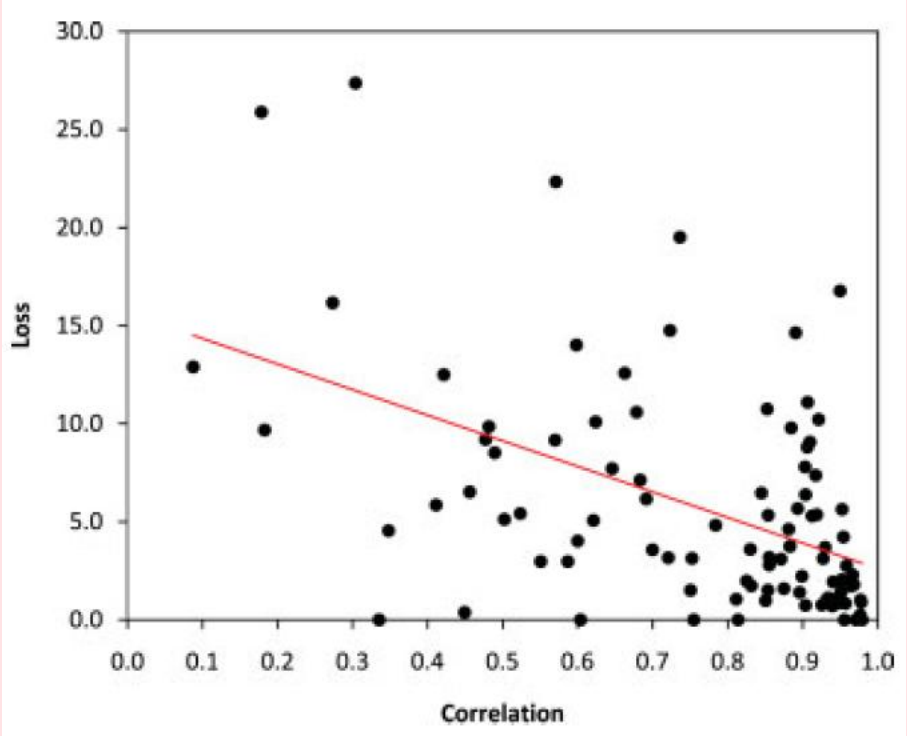


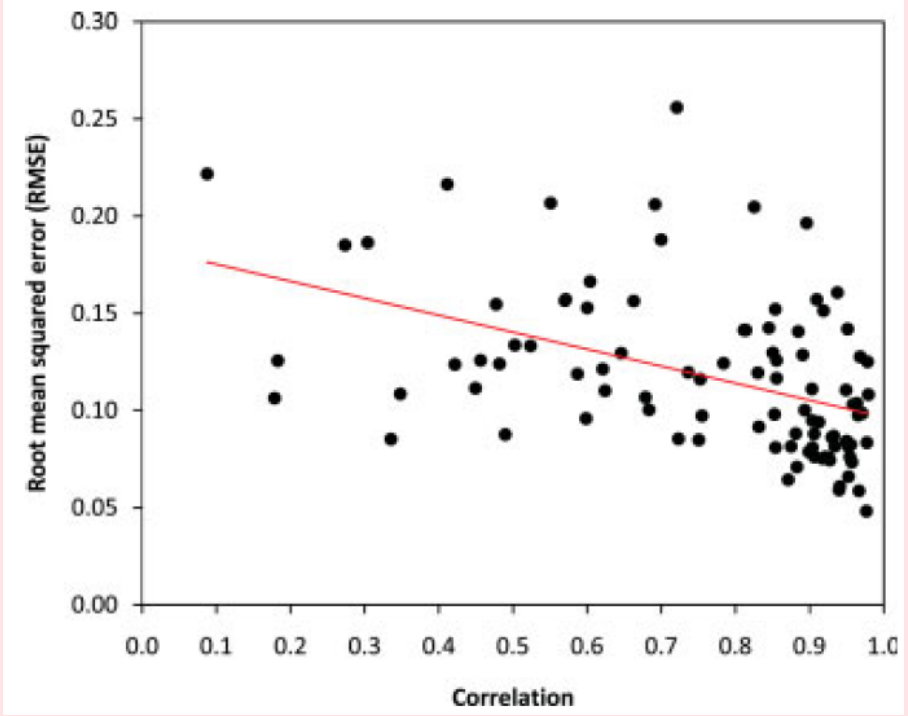


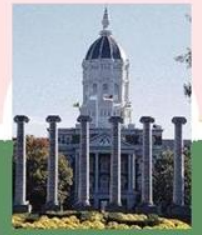


Paper 3

- Correlation versus loss and RMSE of 95 CASP7 targets.

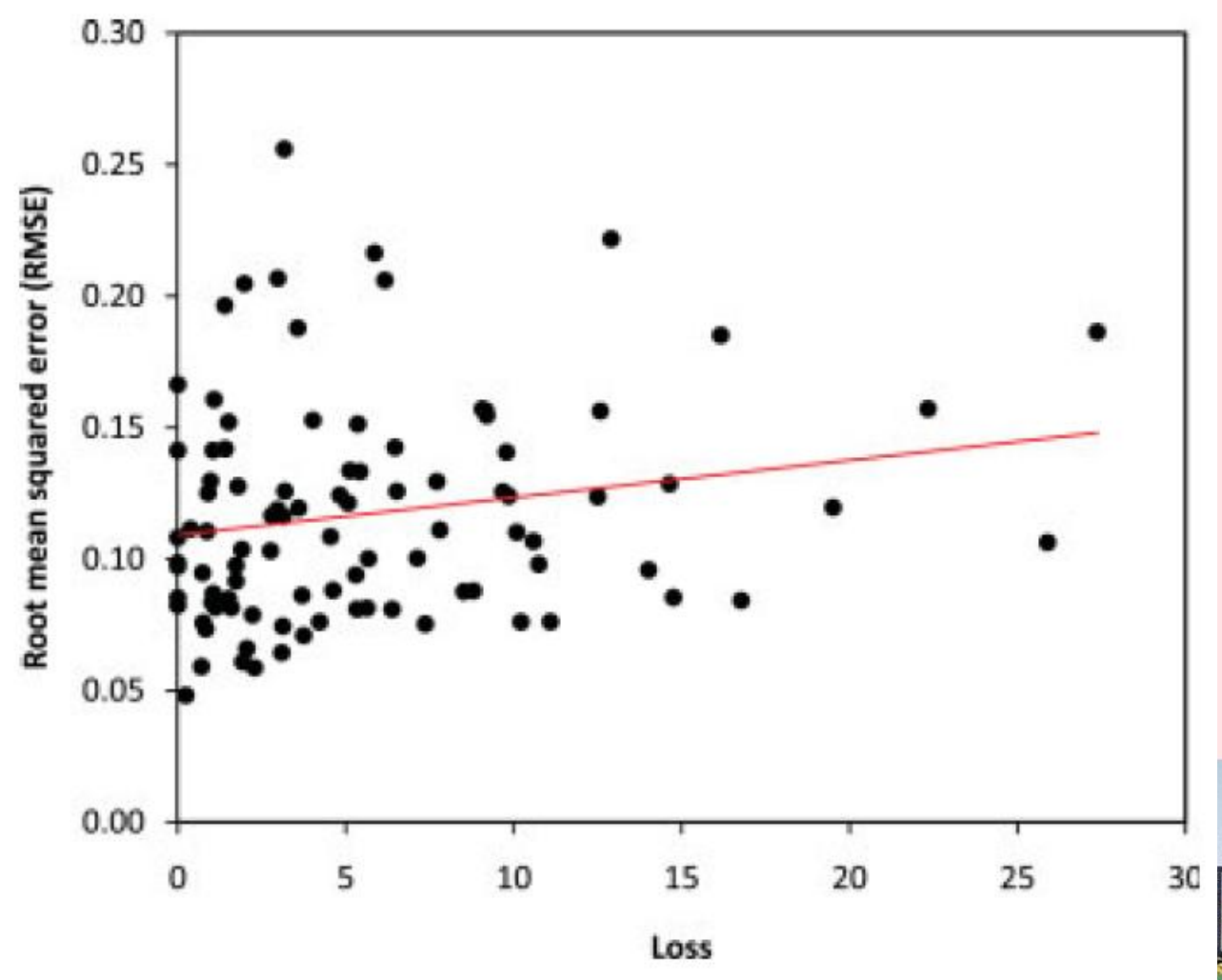


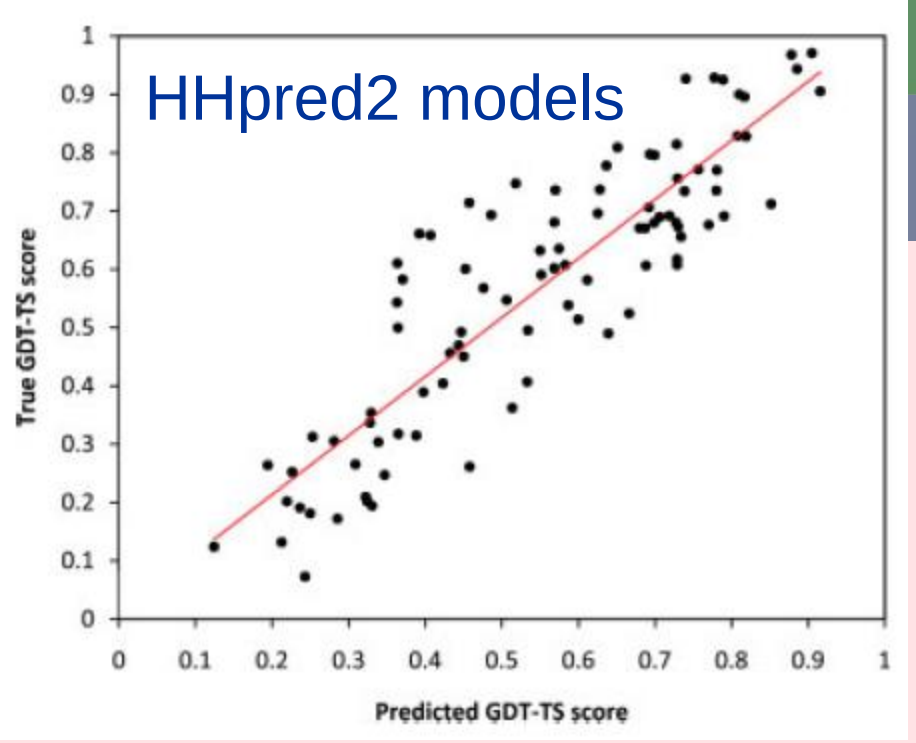


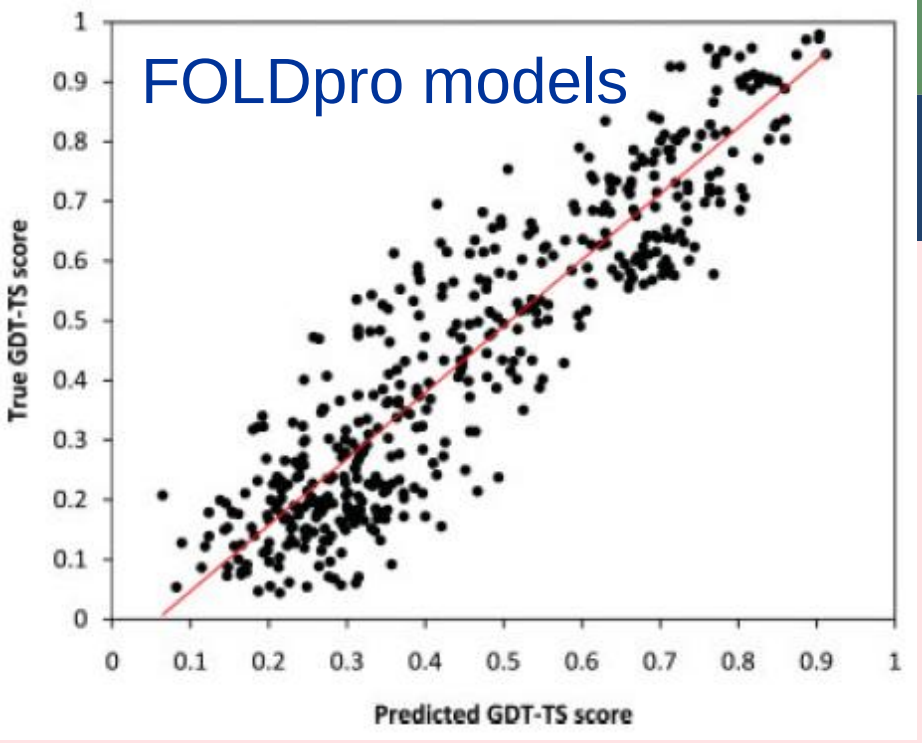


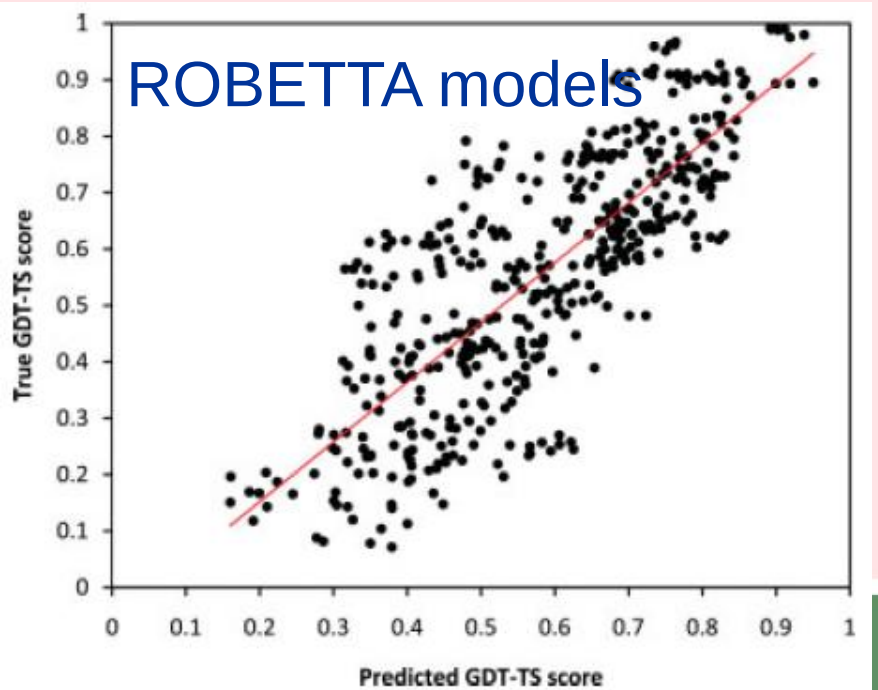
Paper 3

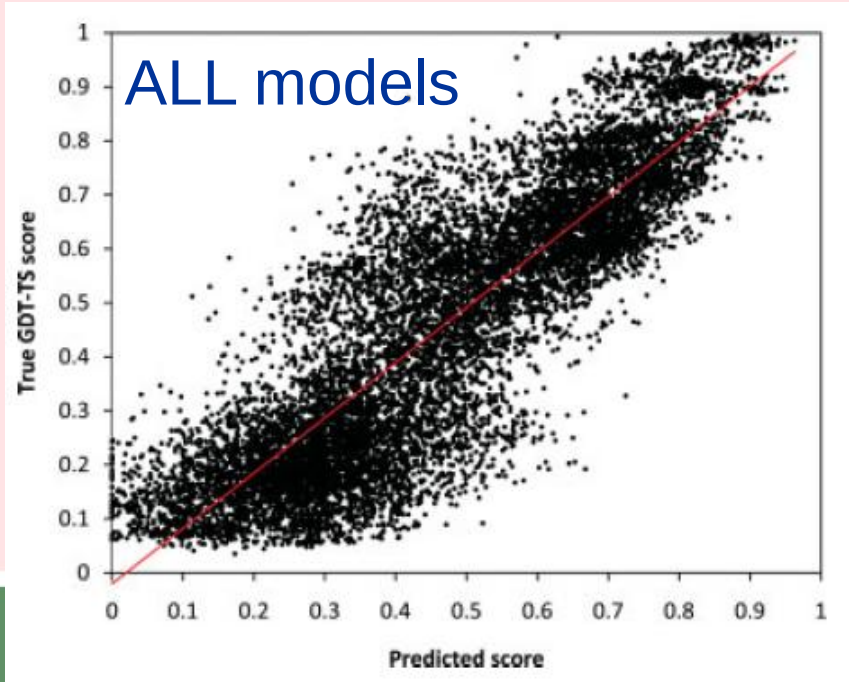
- RMSE versus loss of 95 CASP7 targets.




HHpred2 models
True GDT-TS score
Predicted GDT-TS score


FOLDpro models
True GDT-TS score
Predicted GDT-TS score

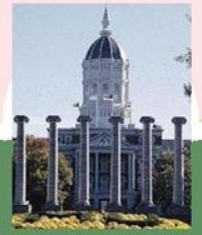

ROBETTA models
True GDT-TS score
Predicted GDT-TS score


ALL models
True GDT-TS score
Predicted score

Paper 3

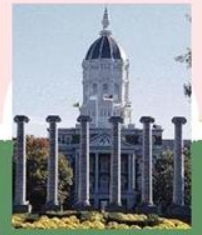
The Results of Three Model Evaluation Methods on CASP7 Models

Method	Ave corr	Corr (TM)	Corr (FM)	Loss	Loss (TM)	Loss (FM)	Over corr
ModelEvaluator	**0.76**	**0.82**	0.50	**5.70**	**5.48**	**6.63**	**0.87**
Circle-QA	0.75	0.79	**0.57**	6.07	5.83	7.09	0.70
ProQ	0.72	0.76	0.53	9.04	9.12	8.69	0.78



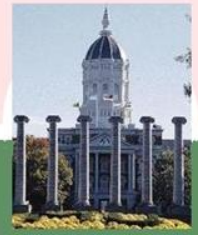
Paper 3 - result

- Conclusion:
- This paper described a quality evaluation model that can predict absolute model quality of a single model. The machine learning method is used to train the model for the prediction.



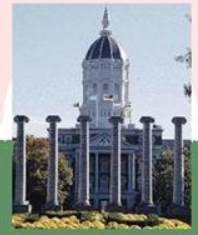
Outline

- Introduction
- Paper1
- Paper2
- Paper3
- Discussion and research plan
- Acknowledgement and references



Discussion

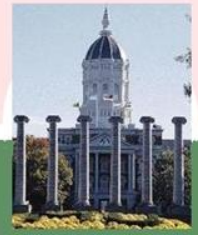
- Discussion:
  - 1. A new statistical knowledge based potential, and apply molecular dynamics for model quality assessment.
  - 2. Apply higher-order φ–ψ pairs scoring for quality assessment.
  - 3. Support vector machine for model quality assessment.
- Limitations:
  - 1. MD takes time. Pearson correlation.
  - 2. Parameters to choose.
  - 3. Accuracy and ability to choose the best model.



Research plan

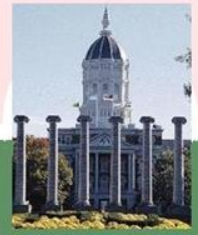
- Research plan:

- Find good features for machine learning method.
- Applying machine learning method (Such as neural network, deep network, support vector machine) to find the patterns for quality assessment.



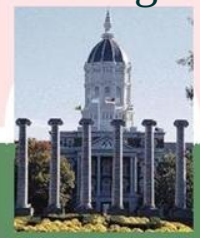
Acknowledgement

- Dr. Jianlin Cheng
- Dr. Ye Duan
- Dr. William L. Harrison
- All members in Bioinformatics, Data Mining and Machine Learning Laboratory (BDML)
- Google images



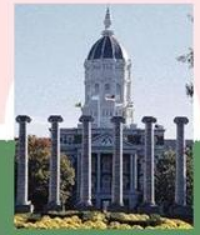
References

- Tanaka, S. & Scheraga, H.A. Mediumand long range interaction parameters between amino acids for predicting three dimensional structures of proteins. Macromolecules. 1976
- Lazaridis, T.&Kooperberg,C.,Huang,E.&Baker,D. Effective energy functions for protein structure predictions. Curr. Opin, Struct, Biol, 1996
- Simons, K.T. et al. Improved recognition of native like protein structures using a combination of sequence dependent and sequence independent features of proteins. 1999.
- Rykunov, D.&Fiser,A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC bioinformatics, 2010.
- Tsai, J.,Bonneau, R.,Morozov, A. V.,Kuhlman,R.,Rohl,C.A.&Baker,D.An. Improved protein decoy set for testing energy functions for protein structure prediction. Proteins, 2003.
- Benkert, P., Tosatto,S.C.E&Schomburg,D.QMEAN: A comprehensive scoring function for model quality assessment. Protein, 2008



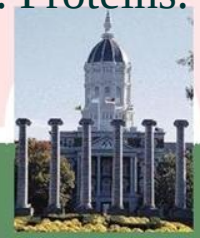
References

- Benkert,P.,Kunzli,M.&Schwede,T. QMEAN server for protein quality estimation. Nucleic Acids Res, 2009.
- Kabsch,W.&Sander,C.Dictionary of protein secondary structure pattern recognition of hydrogen bonded and geometrical features. Biopolymers, 1983.
- Varshney,A.,Brooks,F.P.&Wright,W.V. Computing smooth molecular surfaces. IEEE computer Graphycs and applications. 1994.
- Humphrey,W.,Dalke,A&Schulten,K.VMD. Visual molecular dynamics. Jour.Mol,Gra, 1996
- Lovell,S.C. et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins. 2003
- Lindahl, E.,Hess,B.&van der Spoel,D.GROMACS 3.0:a package for molecular simulation and trajectory analysis. J.Mol.Mod, 2001
- Zemla, A. LGA: a method for finding 3d similarities in protein structures. Nucl,Ac,Res. 2003



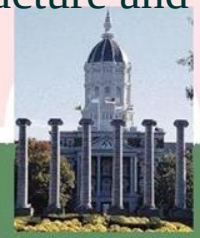
References

- Sims, G.E.&Kim,S-.H. Proc. Natl. Acad. Sci. 2005
- Moult J, Fidelis K,Kryshtafovych A,Rost B,Hubbard T,Tramontano A. Critical assessment of methods of protein structure prediction – round VII. Proteins. 2006
- Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Protein. 2007
- Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci, 2002.
- Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics. 2006
- Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. Proteins. 2004
- Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins. 2005.



References

- Zhou H, Zhou Y. SPARKS 2 and SP3 servers in CASP6. Proteins. 2005.
- Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 1997.
- Chivian D, Kim D, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss C, Bonneau R, Rohl C, Baker D. Automated prediction of CASP 5 structures using the Robetta server. Proteins 2003.
- Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim D, Meiler J, Misura K, Baker D. Free modeling with Rosetta in casp6. Proteins 2005.
- Chivian D, Kim D, Malmstrom L, Schonbrun J, Rohl C, Baker D. Prediction of CASP6 structures using automated robetta protocols. Proteins. 2005
- Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002
- Cheng J, Randall A, Sweredoski M, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005.



Thank you!

Q & A

Email: rcrg4@mail.missouri.edu