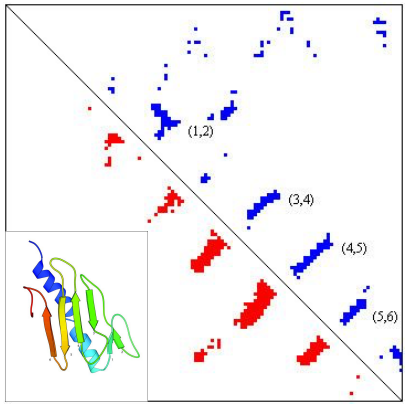
# Protein Structure Modeling Guided by Deep Learning and Contact Prediction


## The MULTICOM Group

## Jianlin Cheng

**Department of Electrical Engineering and Computer Science**
**University of Missouri - Columbia**
**U.S.A.**


**CASP13 Meeting, 2018**



# Outline

- Overview of MULTICOM system
- Three key new methods
- Analysis of examples
- Summary

# I. MULTICOM System

Server

Sequence (MAAKKGMTTVLVSAVICAGVII…)

Sequence Alignment

1D Structural Feature Prediction (SS, SA, Disorder)

Co-evolution

Template Identification (23 Align. + Deep Learn.)

Domain Info

2D Contact Prediction by Deep Learning

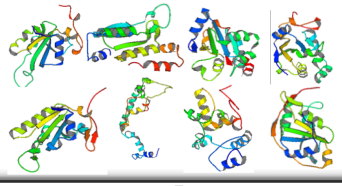
Template-based Modeling

Contact-driven Free Modeling

3D MULTICOM Server Models (full/domain)

Human

CASP13 Server Models



Model Filtering & Side-chain Repacking

Domain Analysis Based on Templates and Align.

Clean 3D Models (full/domain)

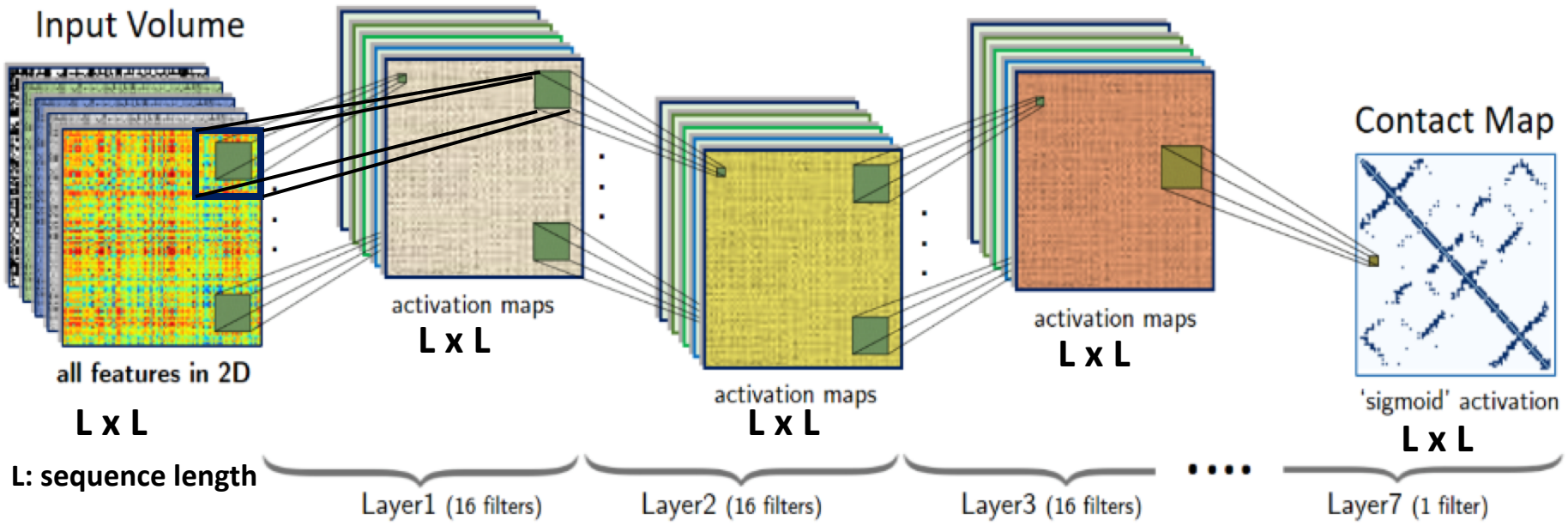
1D, 2D, 3D Features

Deep Learning Model Ranking (full/domain)

Model Combination || Domain Combination || 3DRefine ( )

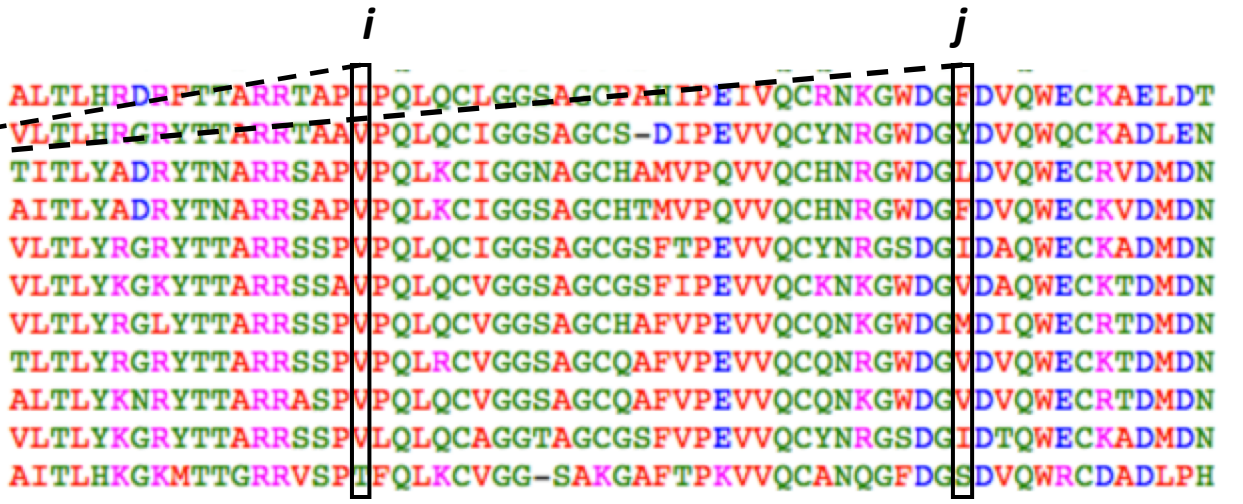
# II. Three Key New Methods in CASP13

# (A) 2D Convolutional Neural Network for Contact Prediction (DNCON2)



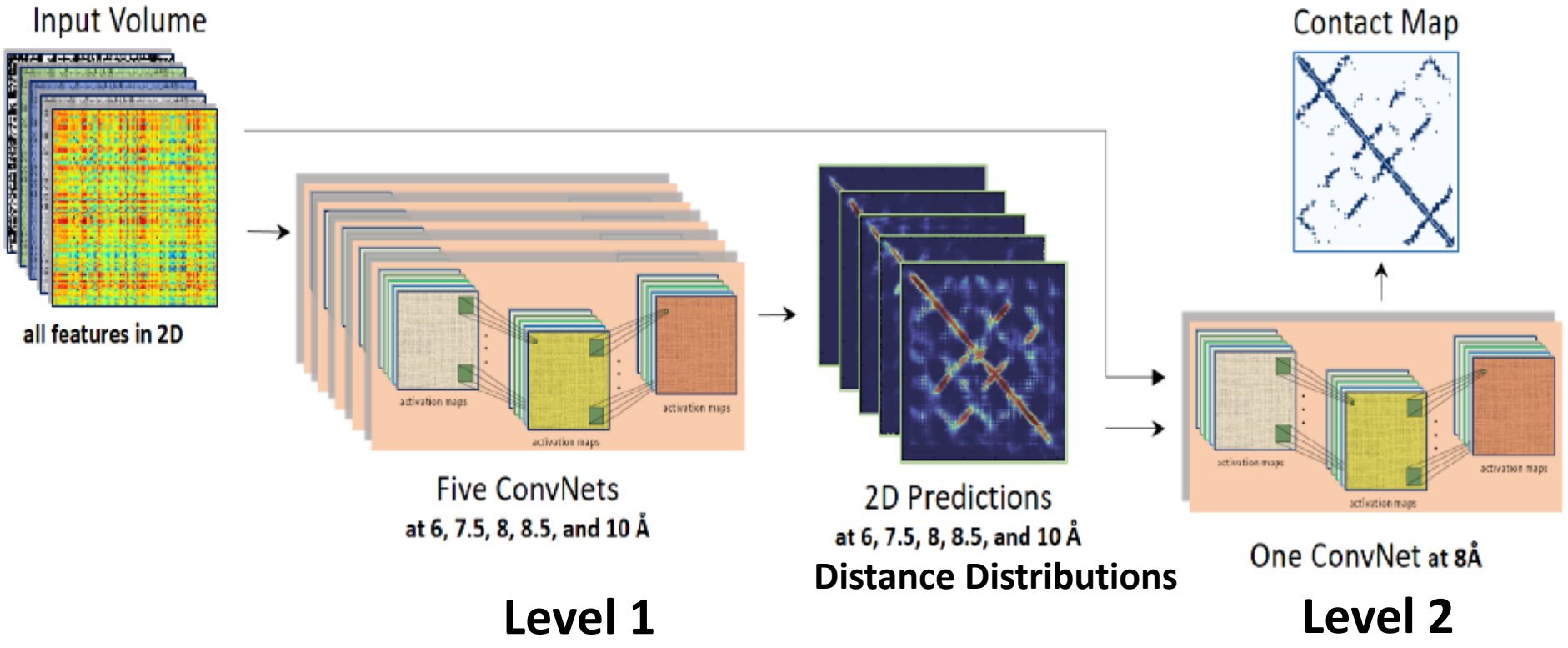

**2D Input Matrices**

- **Co-evolution**
- Secondary structure
- Solvent accessibility
- Mutual information
- Contact potentials
- …



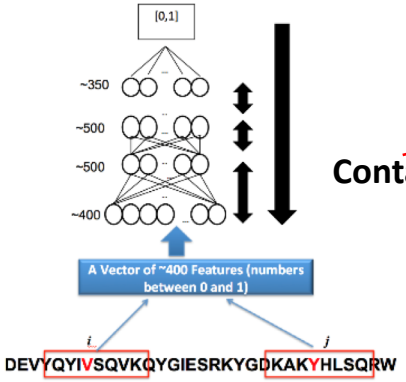

Adhikari et al., 2017

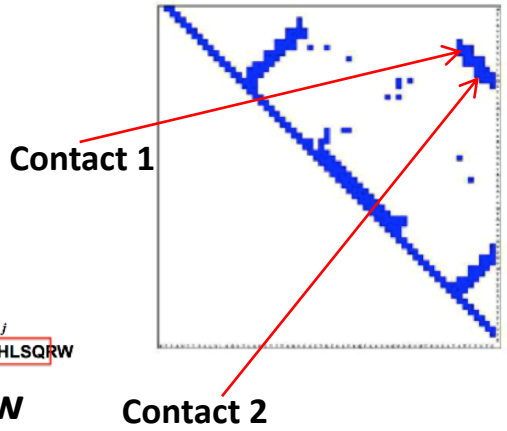
# Two-Level Deep Convolutional Neural Networks



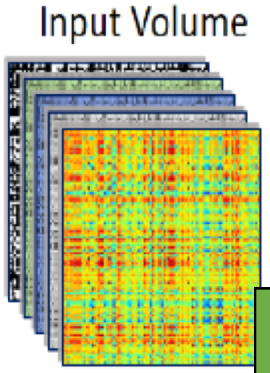

## Key advantages:

- Use global information
- Capture correlation between contacts (high-level contact patterns / clusters)



# Two-Level Deep Convolutional Neural Networks

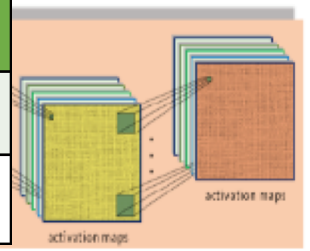



| Method | Accuracy of top L/5 contacts on 115 CASP13 domains |
|---|---|
| DNCON2 (deep learning) | 75% |
| CCMpred (co-evolution) | 45% |

at 6, 7.5, 8, 8.5, and 10 Å

at 6, 7.5, 8, 8.5, and 10 Å

Distance Distributions

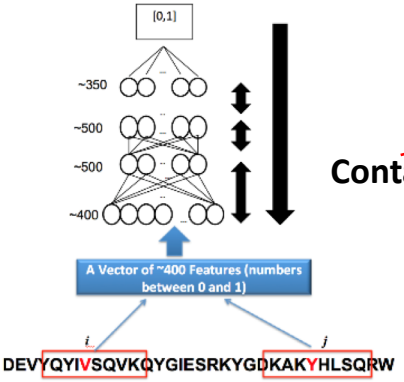

One ConvNet at 8Å

## Level 1

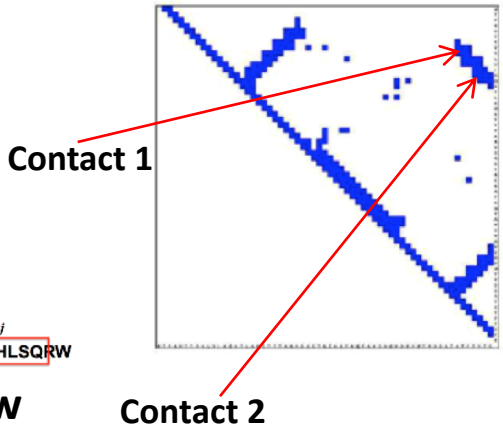
## Level 2

## Key advantages:

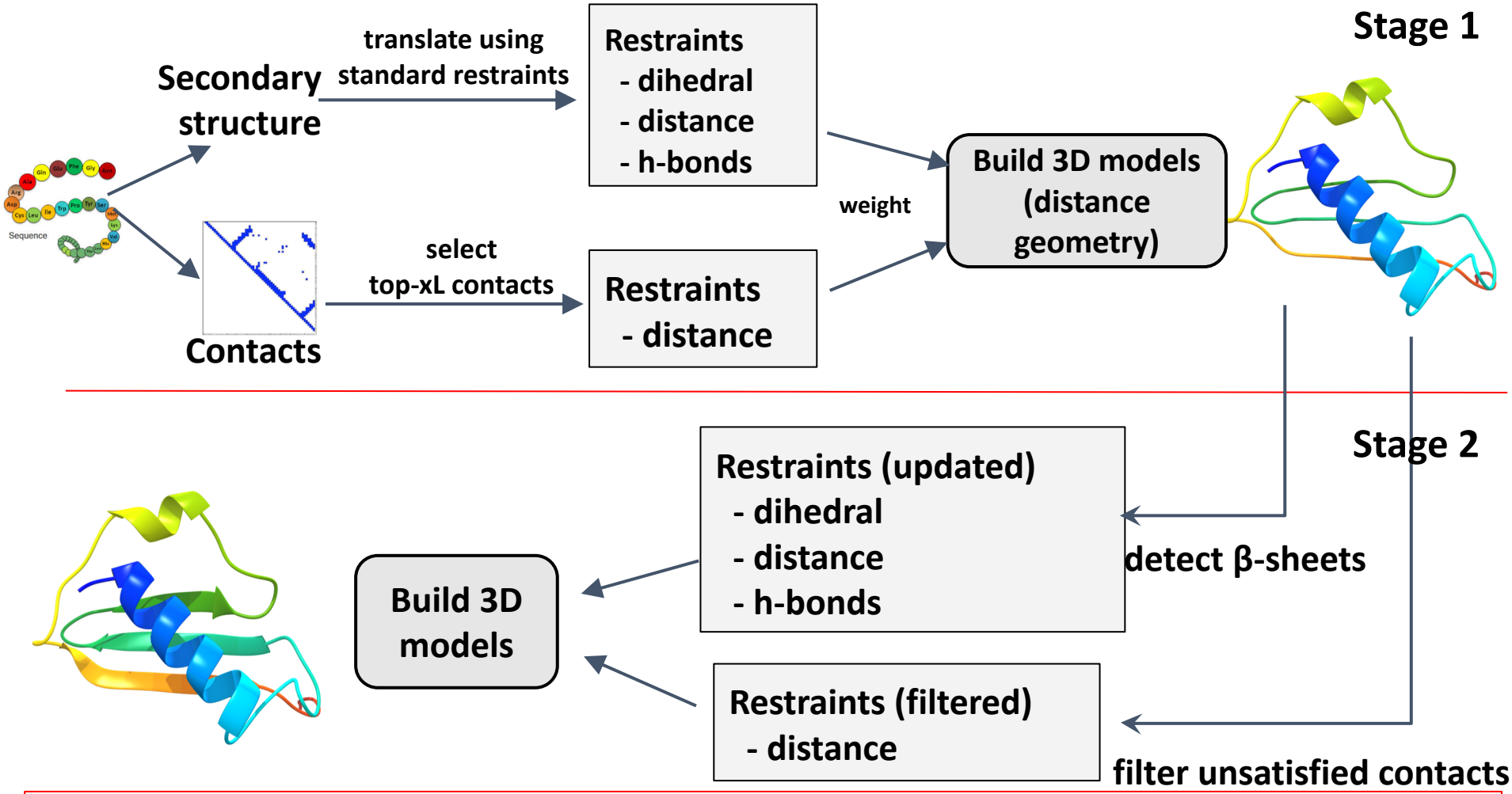
- Use global information
- Capture correlation between contacts (high-level contact patterns / clusters)




Local Window



# (B) Free Modeling by Translating Contact Distances into 3D Models (CONFOLD2)




- Key feature: contacts play a *dominant* and *direct* role in modeling
- Good : Build complicated structures well if there is a sufficient amount of accurate distances; Bad: otherwise may fail

Adhikari, Cheng, 2018

# Free Modeling by Fragment Assembly with Contacts as Energy Terms

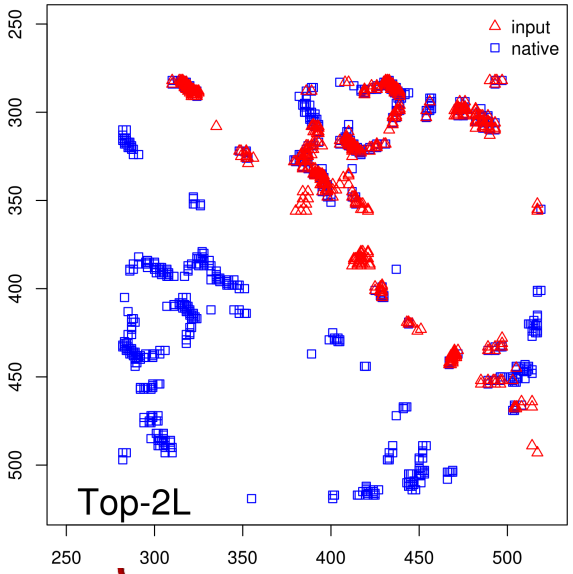
- Rosetta + Contacts
- UniCon3D + Contacts
- FUSION + Contacts

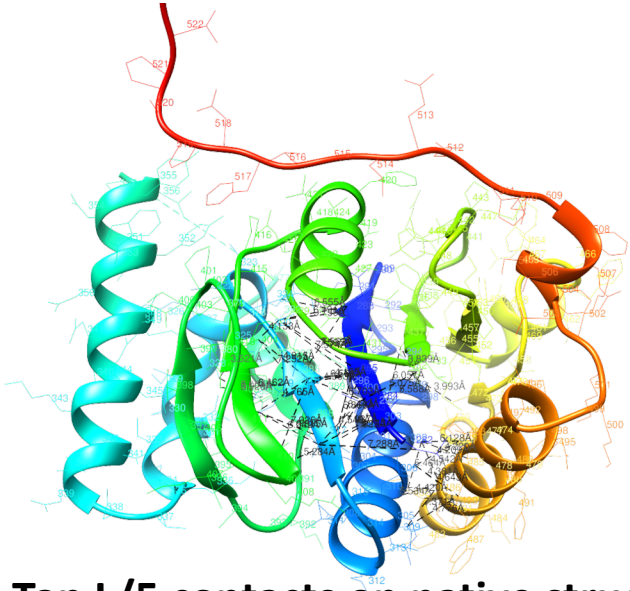
**Key feature**: contacts are used as a part of an energy function to ***indirectly*** guide fragment assembly

- **Good**: use extra fragment information and energy, may work for small proteins with less complex topology
- **Bad**: fail if good models are not sampled, particularly for complicated structures

# Comparison on T1000 – FM Domain (residues: 282-523)

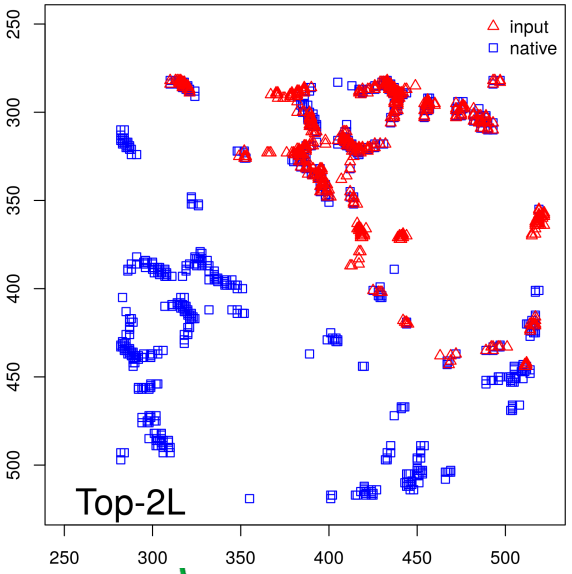
**DNCON2 (red) VS Native (blue)**
**(L/5: 100%, L: 79%, 2L: 50%)**

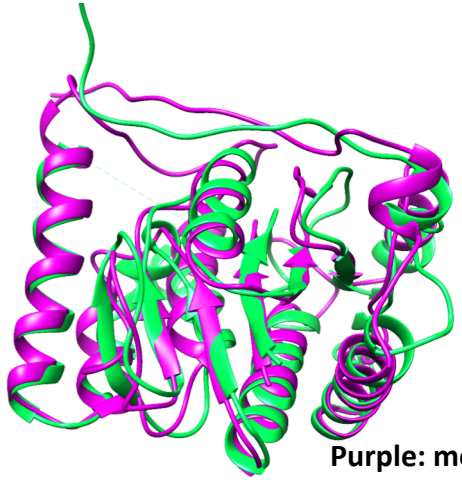


**Top L/5 contacts on native structure**

**CONFOLD (red) VS Native**
**(L/5: 67%, L: 65%, 2L: 55%)**

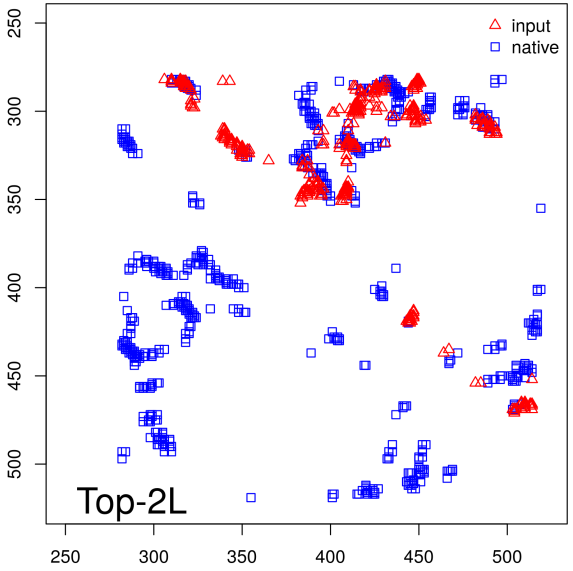


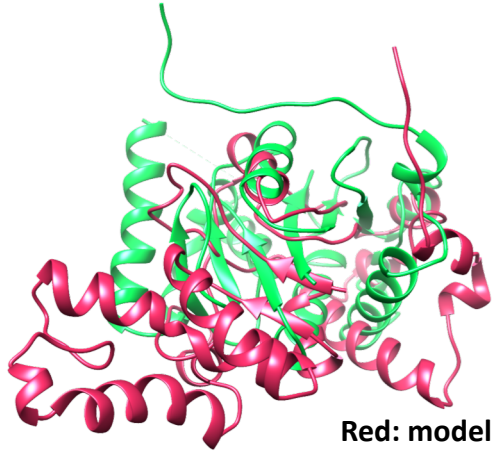



**Purple: model**
**Green: native**

**TM-score: 0.80**
**GDT-TS-score: 0.64**

**Rosetta-Con (red) VS Native**
**(L/5: 20%, L: 18%, 2L: 17%)**



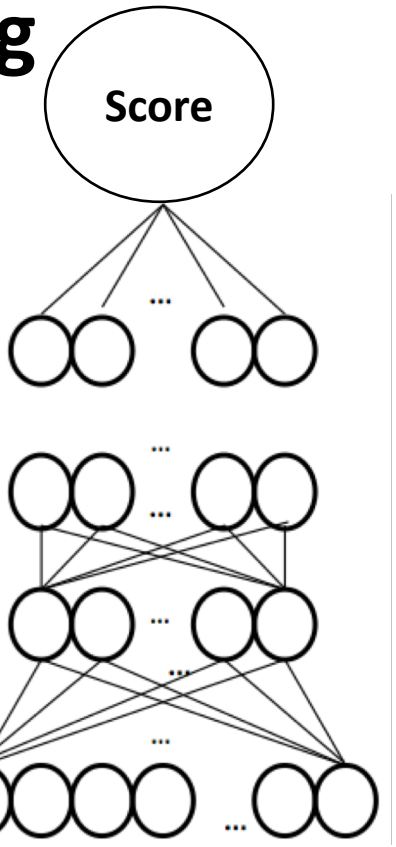



**Red: model**
**Green: native**

**TM-score: 0.33**
**GDT-TS-score: 0.23**

# (C) Deep Learning Model Ranking (DeepRank)

| Method | Loss in CASP13 |
|---|---|
| Deep Learning with Contact | 0.051 |
| Deep Learning without Contact | 0.059 |
| Averaging Feature | 0.088 |




14 tools: OPUS-PSP, RF_SRS, Rwplus, SBROD, QMEAN, Voronota, ModelEvaluator, Dope, DeepQA,ProQ2, ProQ3, Pcons, Apollo, ModFoldclust2

1D Match Scores (SS, SA)

2D Contact Match Scores (short, medium and long)

3D Quality/Energy Scores

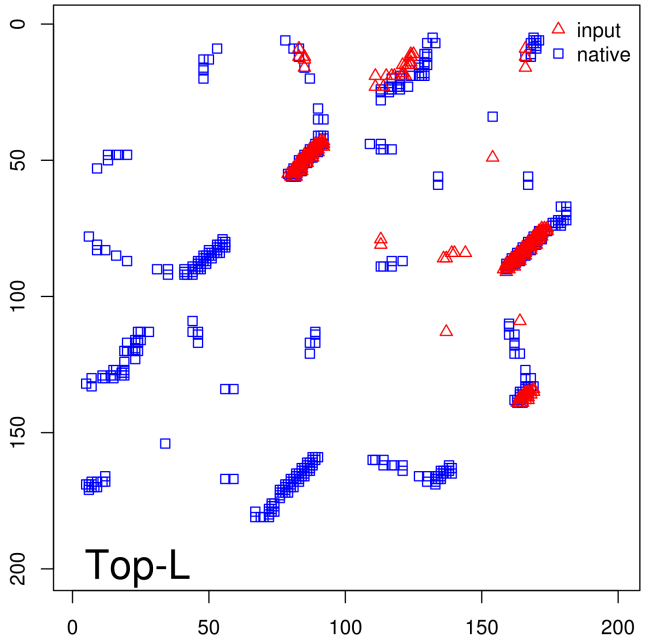
Feature Extraction

CASP13 Models (full length or domain-based)

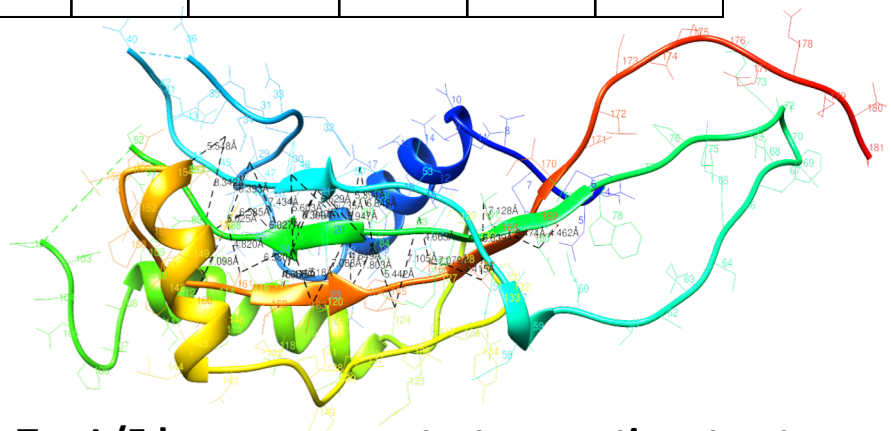
# III. Analysis of Examples (Success and Failure)

# (1) Success of Building Models for T1021s3-D1 (FM) by CONFOLD

## DNCON2 (red) VS Native (blue)

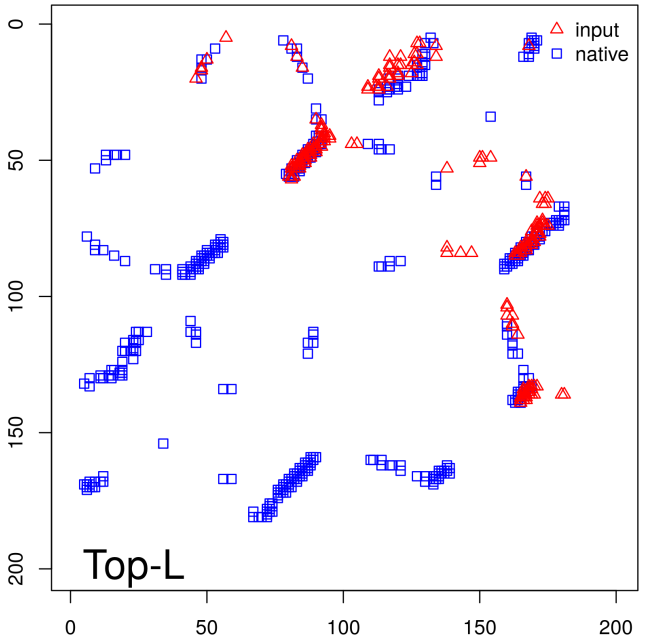



| | Top 5 | Top L/10 | Top L/5 | Top L/2 | Top L |
|---|---|---|---|---|---|
| Acc. | 100% | 94% | 97% | 88% | 61% |

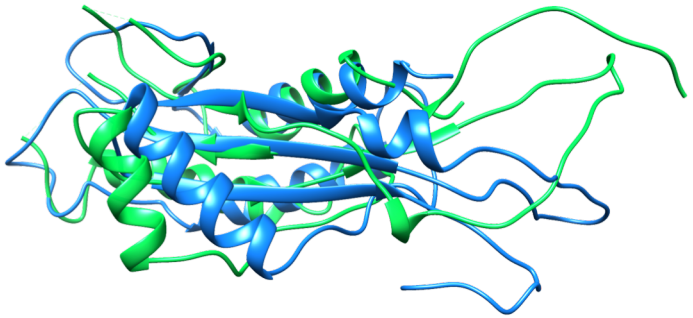


Top L/5 long-range contacts on native structure

## CONFOLD (red) VS Native (blue)




| | Top 5 | Top L/10 | Top L/5 | Top L/2 | Top L |
|---|---|---|---|---|---|
| Acc. | 80% | 47% | 52% | 51% | 46% |

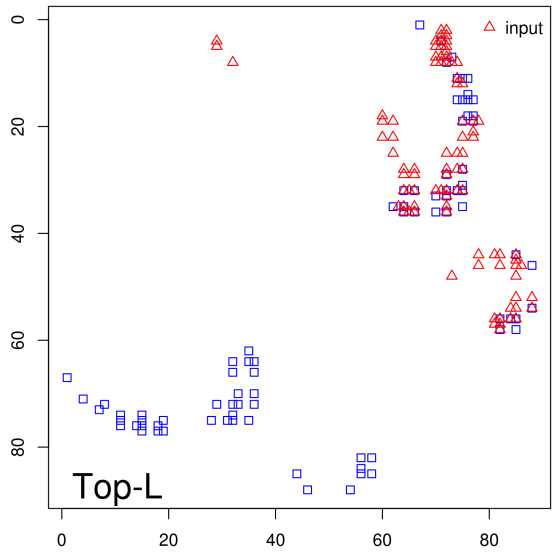


Blue: predicted; Green: native

TM-score: 0.50 GDT-TS-score: 0.41

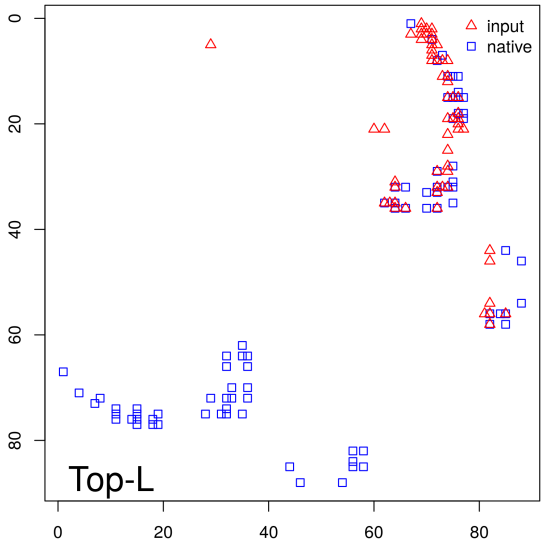
# (2) Success of Building Models from Contacts with Rosetta When Failing to Identify Templates for T1019s2 (TBM)

## DNCON2 (red) VS Native (blue)




| | Top L/10 | Top L/5 | Top L/2 | Top L |
|---|---|---|---|---|
| Acc. | 78% | 61% | 39% | 26% |

## Rosetta-Con (red) VS Native (blue)




| | Top L/10 | Top L/5 | Top L/2 | Top L |
|---|---|---|---|---|
| Acc. | 56% | 56% | 39% | 36% |

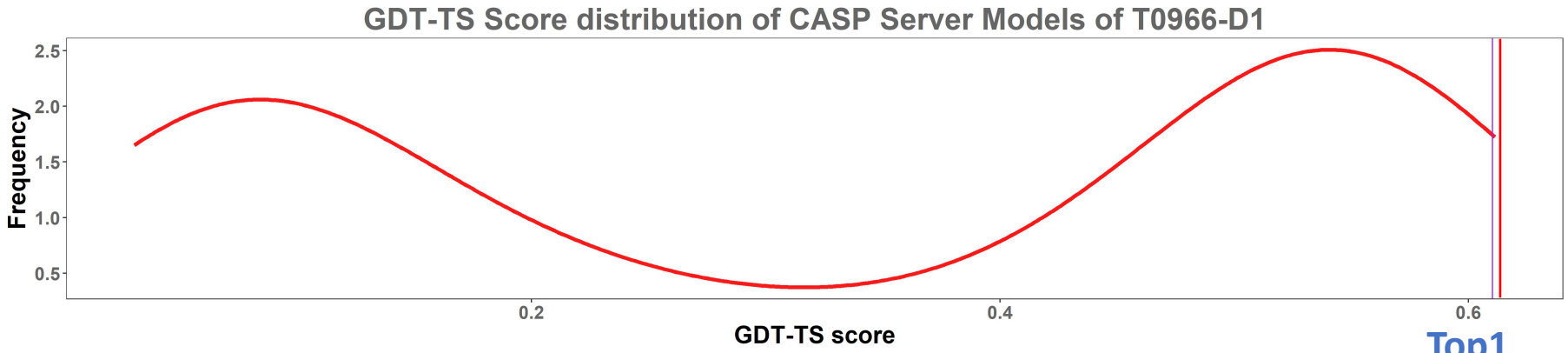
CASP:

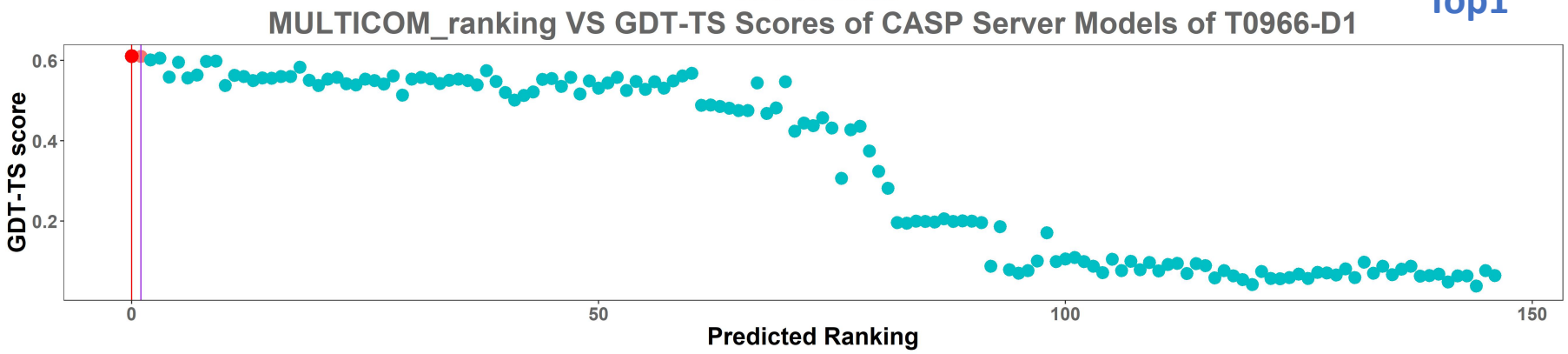
Images of protein structures redacted

Top L/5 long-range contacts on native structure

TM-score: 0.68 GDT-TS-score: 0.67

# (3) Success of Model Ranking and Combination with Deep Learning

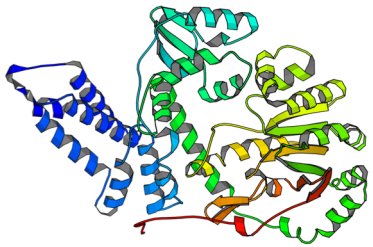


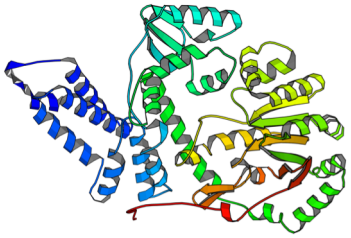

**T0966 (TBM-Hard)**

CASP:
Images redacted

**Top1 Selected & Best:**
**BAKER-ROSETTASERVER_TS1**
**GDT-TS: 0.6103**

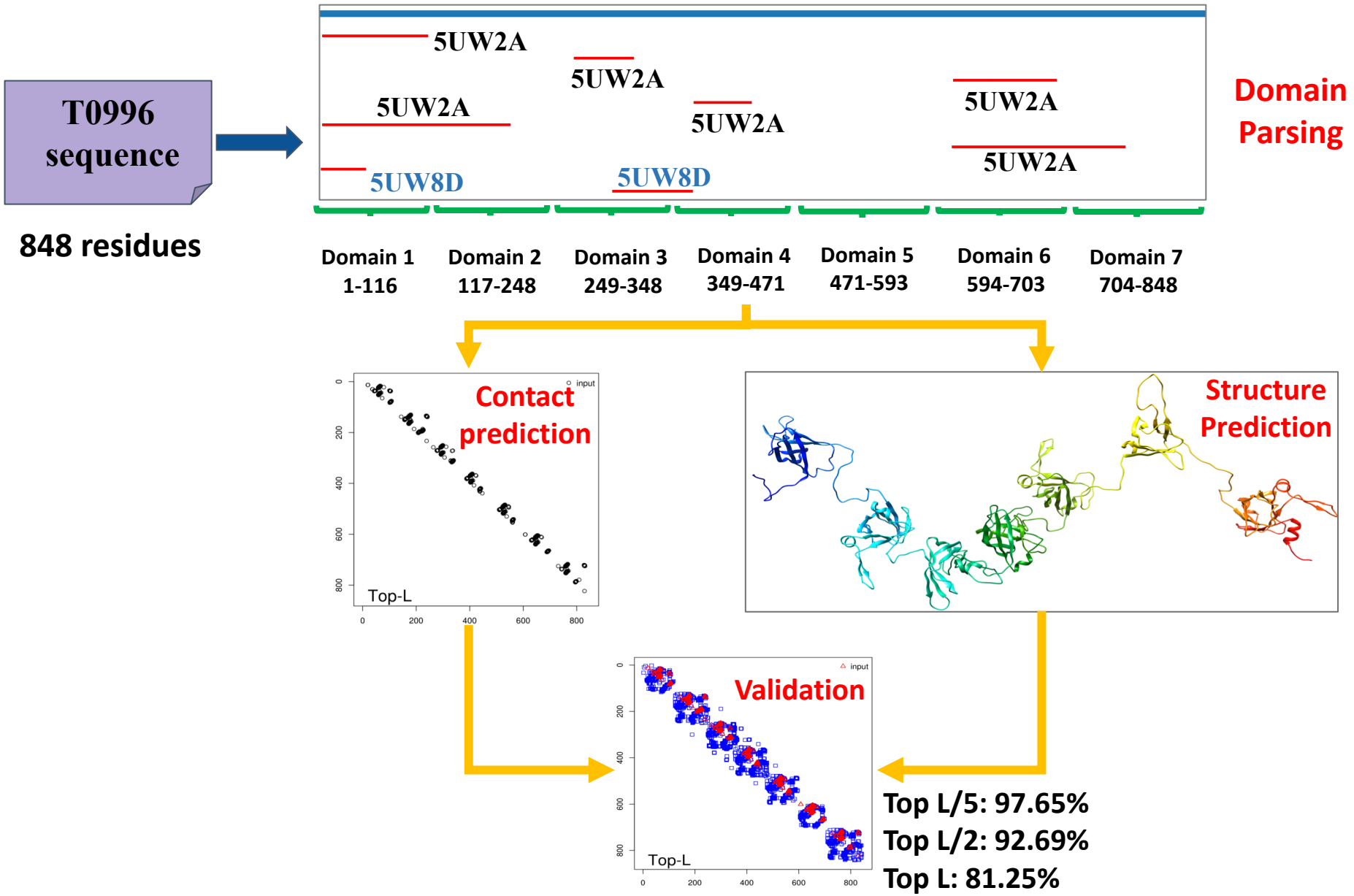


**MULTICOM Model**
**GDT-TS: 0.6113**

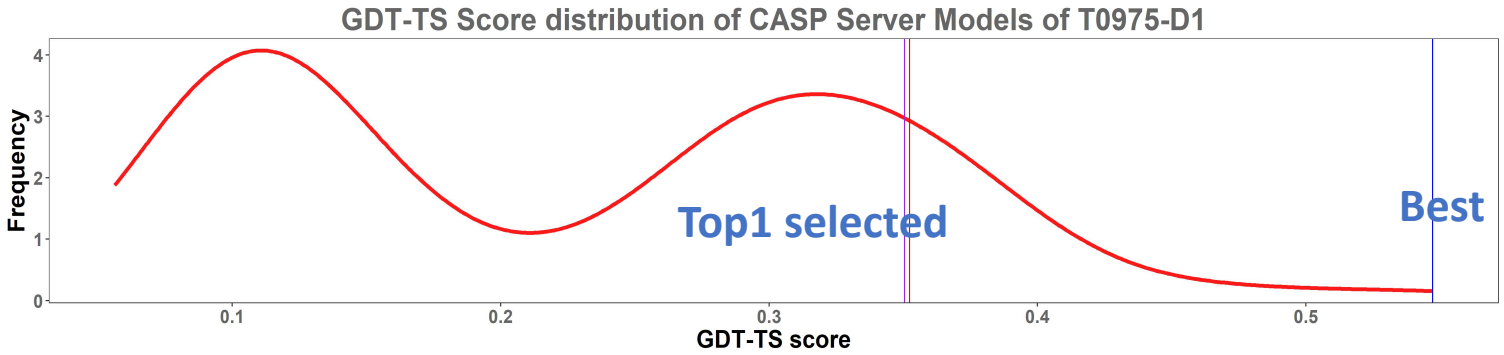


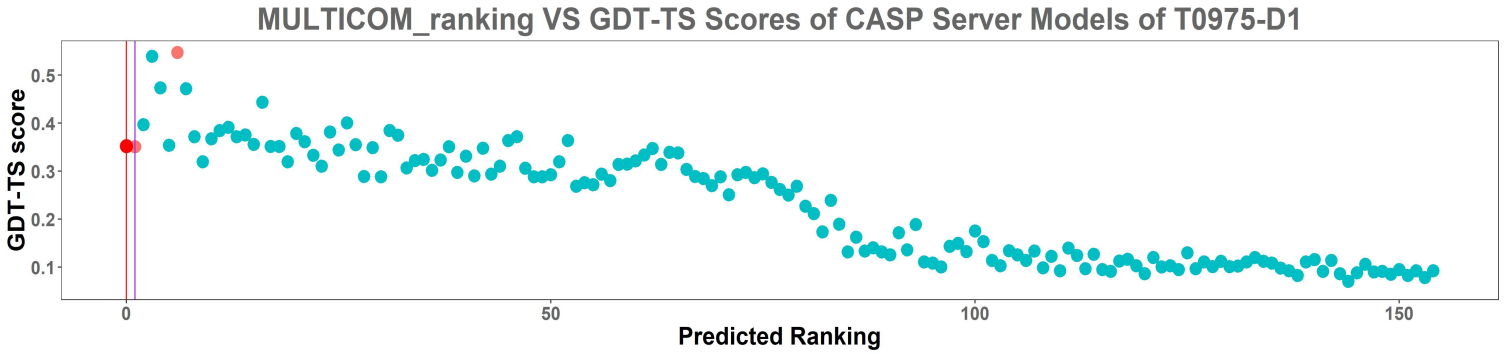
**Other cases: T0954, T0965, T0966, T0980s1, T0982**

# (4) Success of Domain Parsing, Template Identification, Domain-Based Model Ranking with Deep Learning for T0996 (TBM)

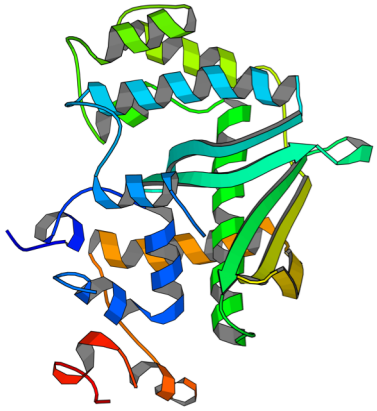


# (1) Failure of Ranking Models (Loss >=10)

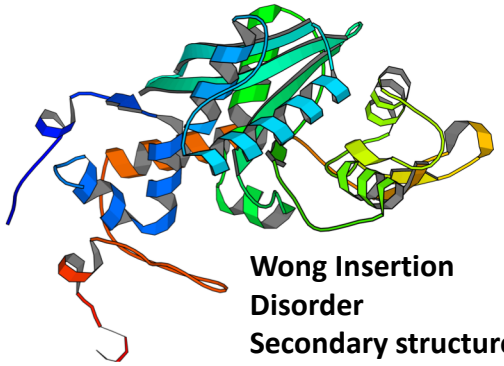



**T0975 (FM)**

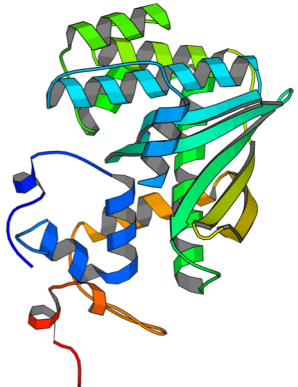


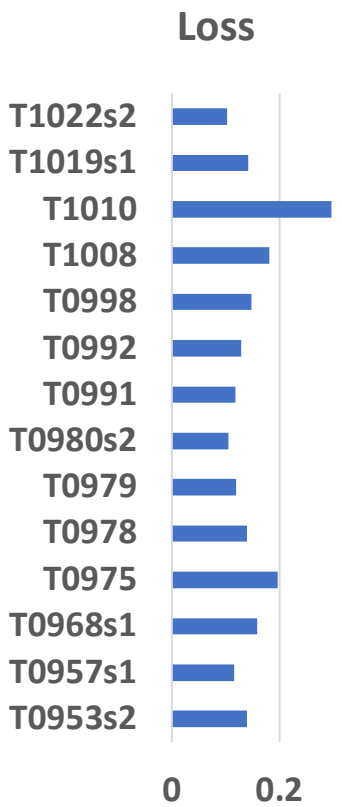
**Top 1 Selected**
**GDT-TS: 0.35**
**TM-score: 0.49**



**Wong Insertion**
**Disorder**
**Secondary structure**

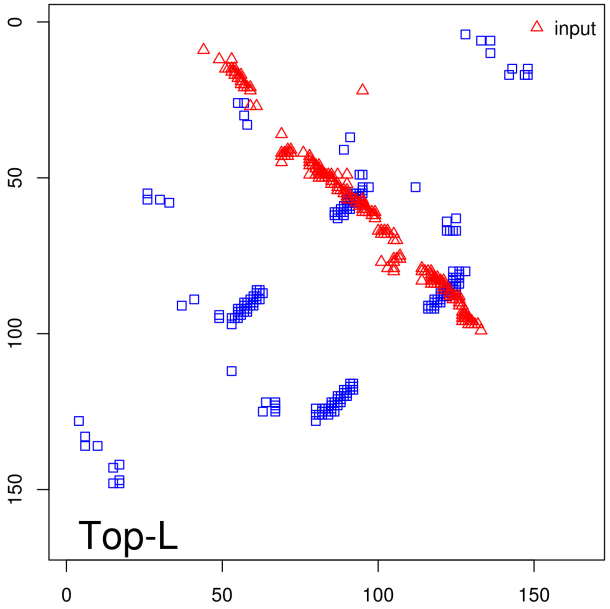
**Best Model (Zhang-Server_TS5)**
**GDT-TS: 0.55**
**TM-score: 0.75**





# (2) Failure of predicting / using contacts (T0998 FM)

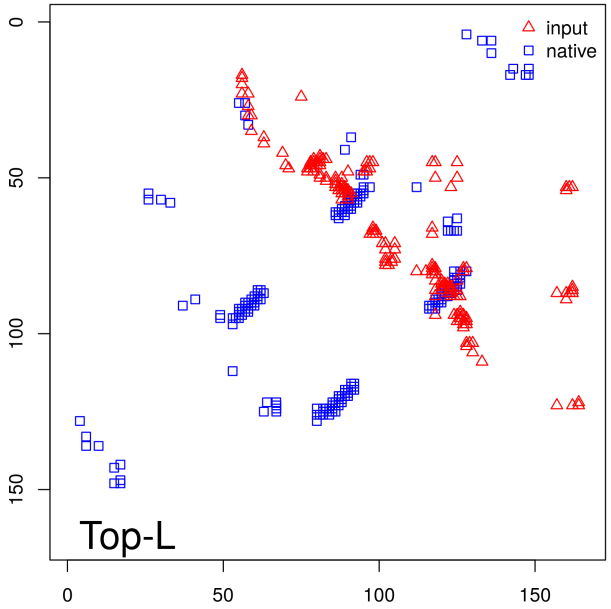
**DNCON2 (red) VS Native (blue)**




| | Top L/10 | Top L/5 | Top L/2 | Top L |
|---|---|---|---|---|
| Acc. | 6% | 6% | 5% | 5% |

**# of effective sequences = 2**

**Model (red) VS Native (blue)**




| | Top L/10 | Top L/5 | Top L/2 | Top L |
|---|---|---|---|---|
| Acc. | 6% | 6% | 6% | 4% |

CASP:

Images of protein structures redacted

**Top L/5 medium-range contacts on native structure**

**TM-score: 0.21 GDT-TS-score: 0.15**

# IV. Summary

- **Contact/distance prediction is the light in the dark world of free modeling.**
- **Contact prediction is valuable for ranking models and templates.**
- **Deep learning holds the key of protein structure prediction.**
- **Contact/distance-based free modeling and fragment-based free modeling are complementary.**
- **There are significant challenges in model ranking and contact/distance-based modeling.**

# Acknowledgements

## Current PhD students:

Jie Hou

Tianqi Wu

## Former PhD Students:

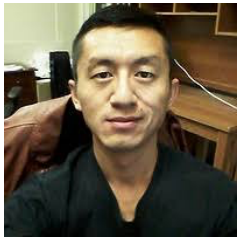
Badri Adhari

Deb Bhattacharya

Renzhi Cao

Xin Deng

Jesse Eickholt



Jilong Li

Zheng Wang