Critical Assessment of Techniques for Protein Structure Prediction CASP13 Targets

Lisa Kinch, Andriy Kryshtafovych, Assessor Teams and the Prediction Center

CASP13 Targets: Overview of the Task

1 Domain-Based Definition of Evaluation Units

- Split targets manually (domain parsers and templates)
- Evaluate server performance on splits with Grishin plots
- Remove some extended regions and domains (tags, etc.)
- Difficult examples

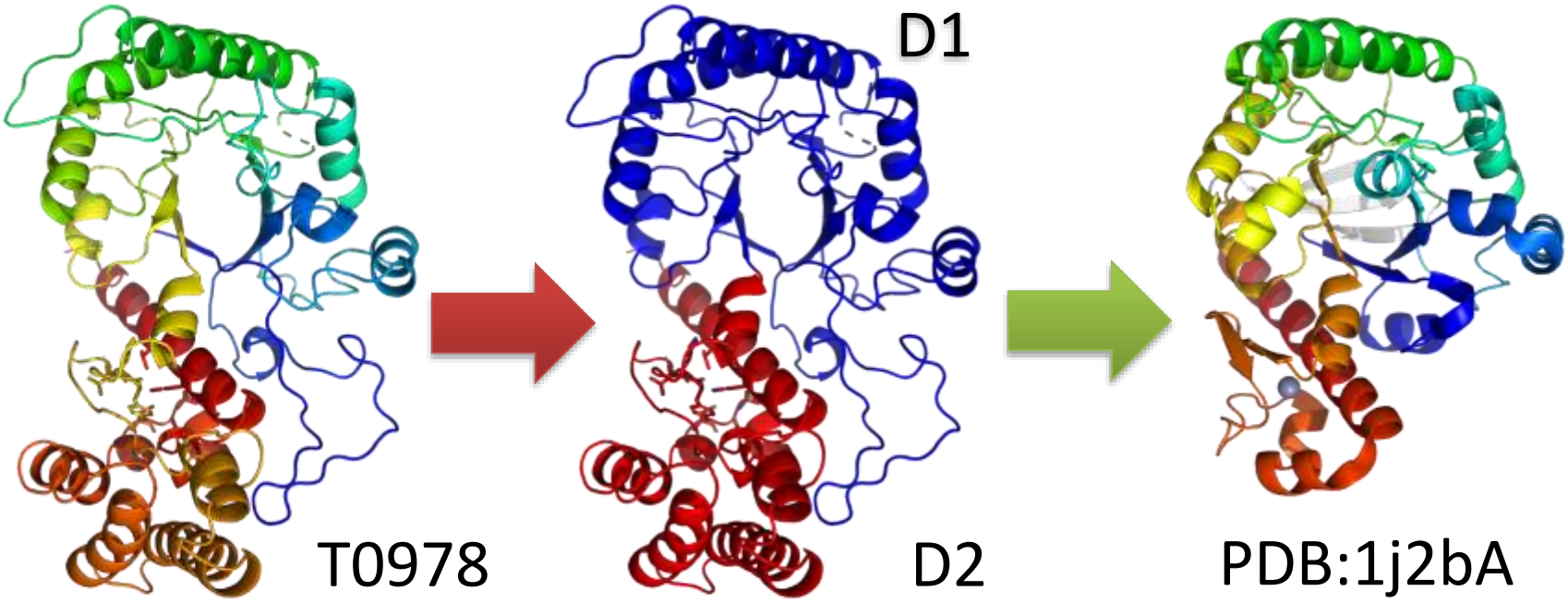
2 Evolutionary-Based Assignment into Categories

- ECOD classification using sequence/structure templates
- Assign “High Accuracy Modeling” (formerly Template Based Modeling) or “Topology” (formerly Free Modeling)
  - CASP12-like difficulty/performance plot
  - Boundary cases examined with PCA and score clustering

CASP13 Targets: Example EVU Definition

Split targets into domains

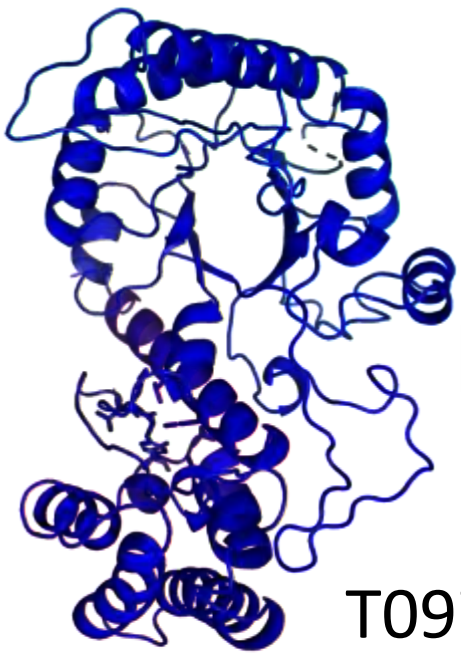
Check templates for consistency



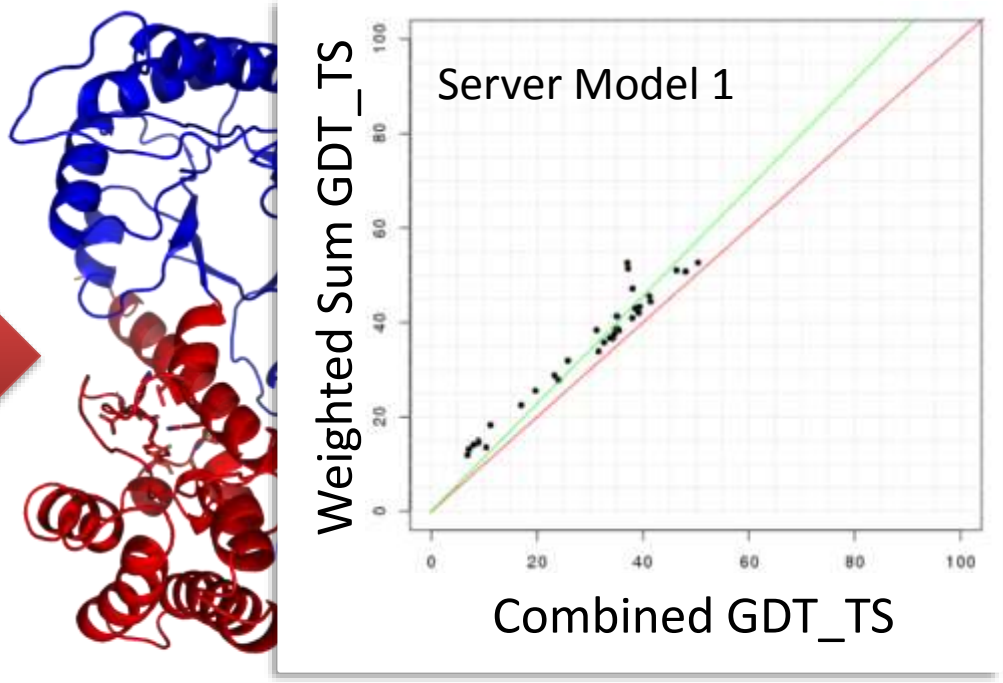
CASP13 Targets: Example EVU Definition

Split targets into domains

Compare performance (Grishin Plot)

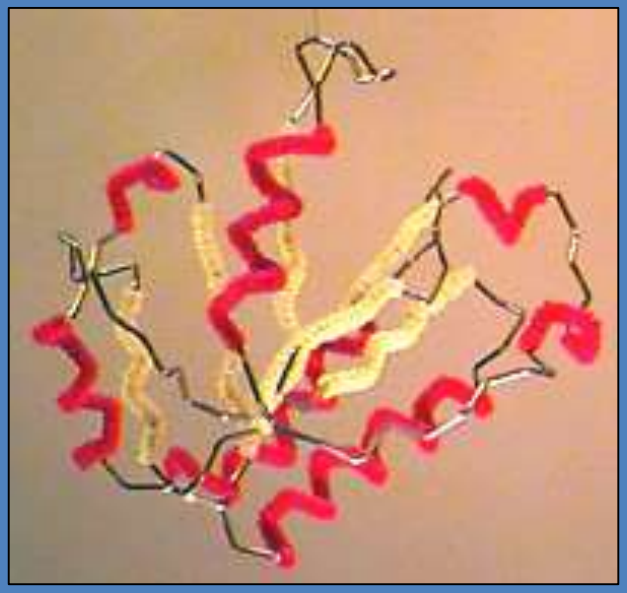


T0978

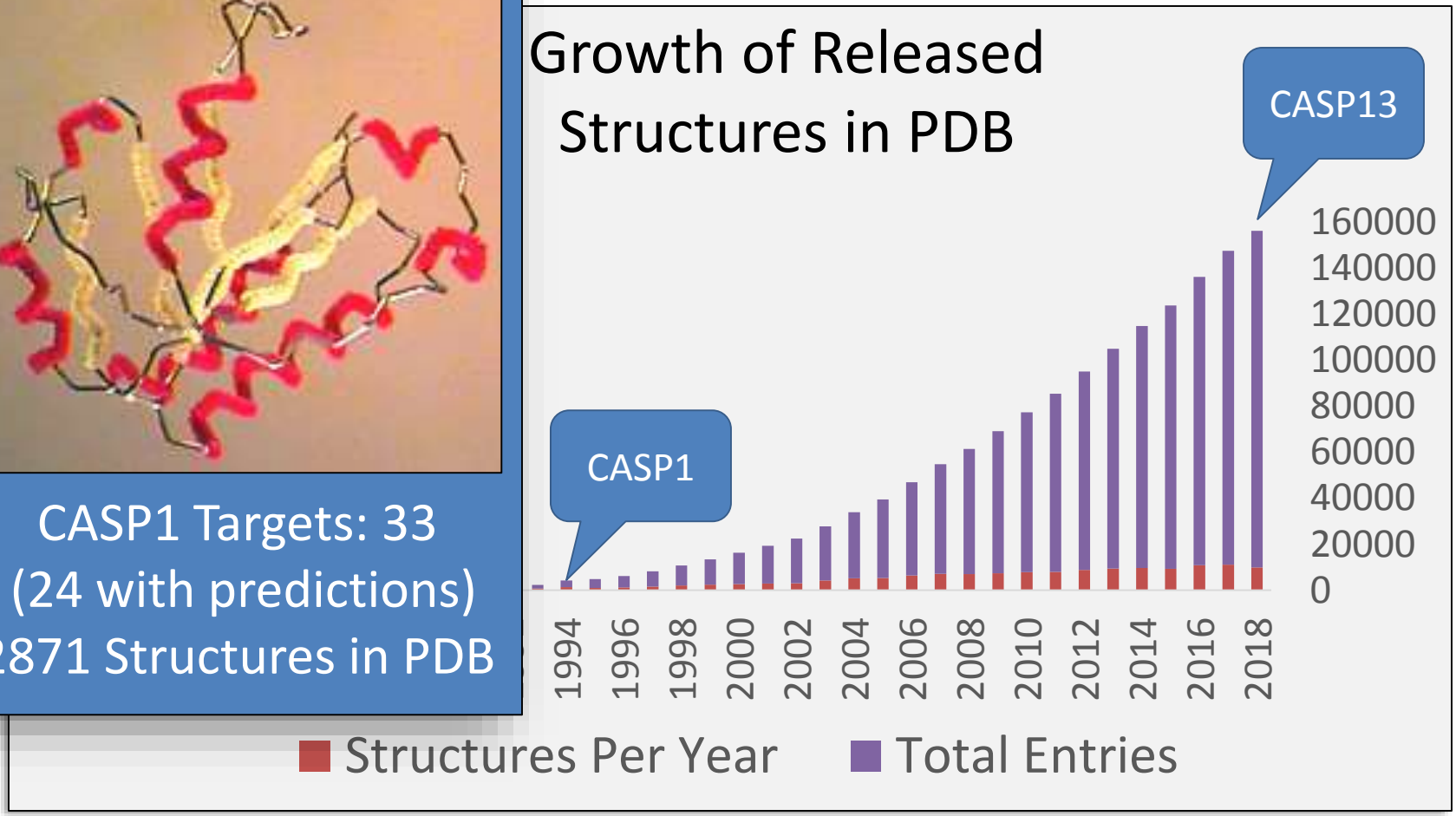


Treat as single EVU

CASP Targets and the Transformation of Structure Biology



CASP1 Targets: 33
(24 with predictions)
2871 Structures in PDB



CASP Targets and the Transformation of Structure Biology

1

Increasing number, size *and* complexity of PDB

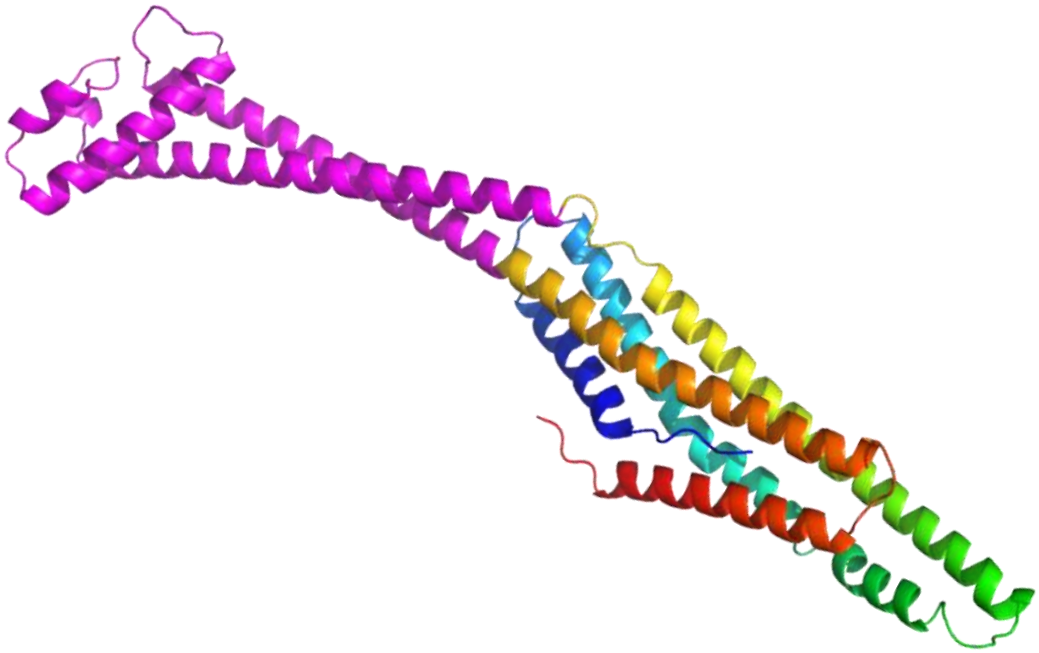
- Expanding available templates: 14,622 PDB structures in 2018
- Transformaing methods: 2541 EM structures in 2018
- More examples of protein dynamics, domain interactions, molecular complexes and membrane proteins

CASP13 Target Type	Number
Tertiary Structure Prediction	90
(all-group)	82
Hetero-Multimer	13
Refinement	31
Assisted Prediction	60
Transmembrane	5*
EM Targets	7

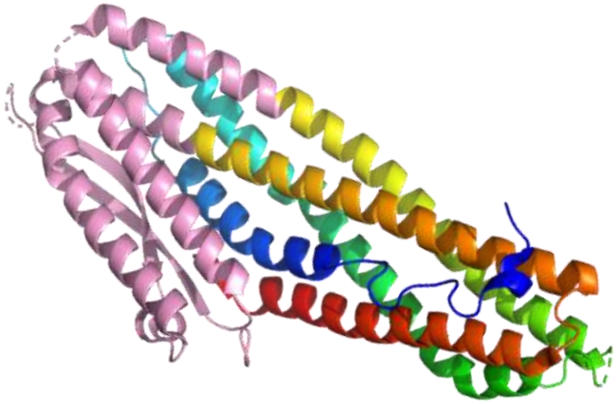
* T0950 exists in soluble AND transmembrane states

Example: T0950 Multiple Conformations 1

T0950: membrane embedded conformation



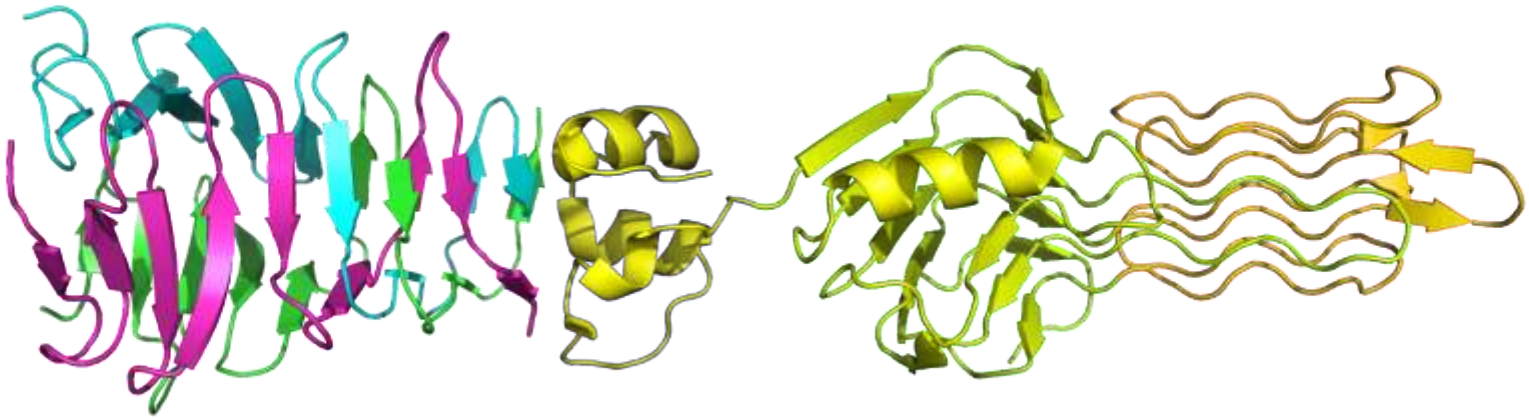
Top Template (4k1p): soluble conformation



Attempted split based on **conformation change**, but kept as single EVU

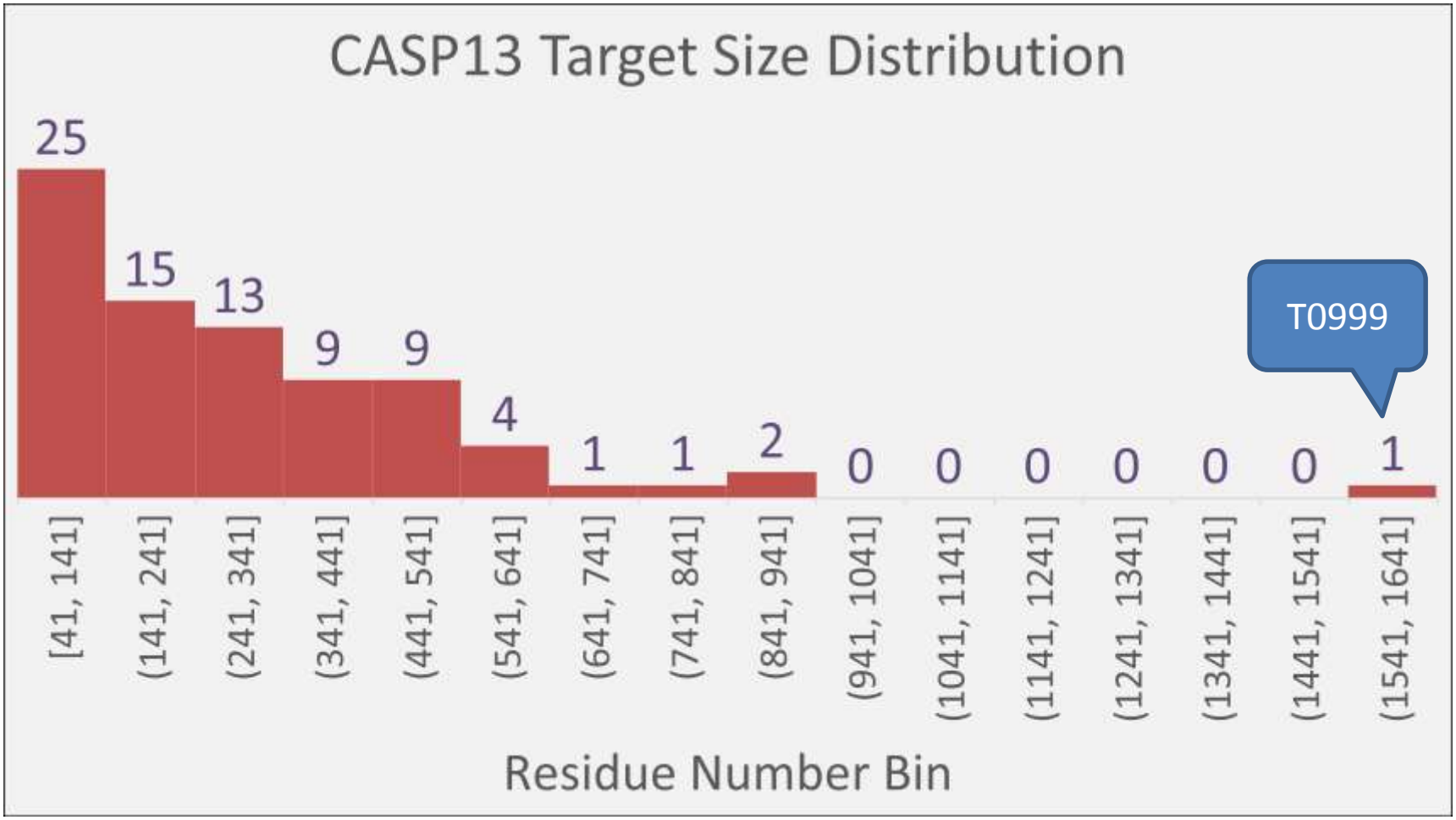
Example: T0953 Complex Interaction Topology 1

- S1 adopts an obligate trimer interaction (cyan, pink, green)
- S2 adopts an intertwined, non-sequential, multidomain topology with a low sequence complexity domain (blue, green, and red)



S1 S2

CASP13 Targets: Large Structures



14 Targets > 500 residues

CASP13 Targets: Large Structures

T0999: AroM polypeptide, contains 5 central enzymes of the shikimate pathway fused in one chain =1589 residues

DHAP

↓

3-Dehydroquinate

↓

3-Dehydro Shikimate

↓

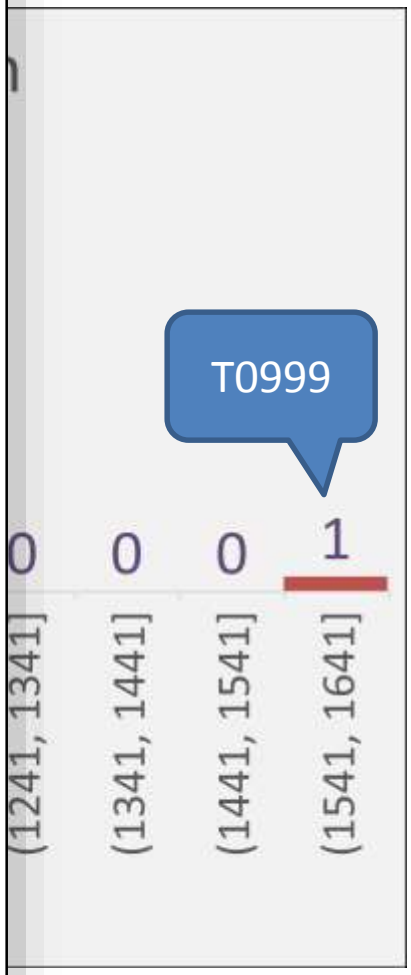
Shikimate

↓

Shikimate 3 Phosphate

↓

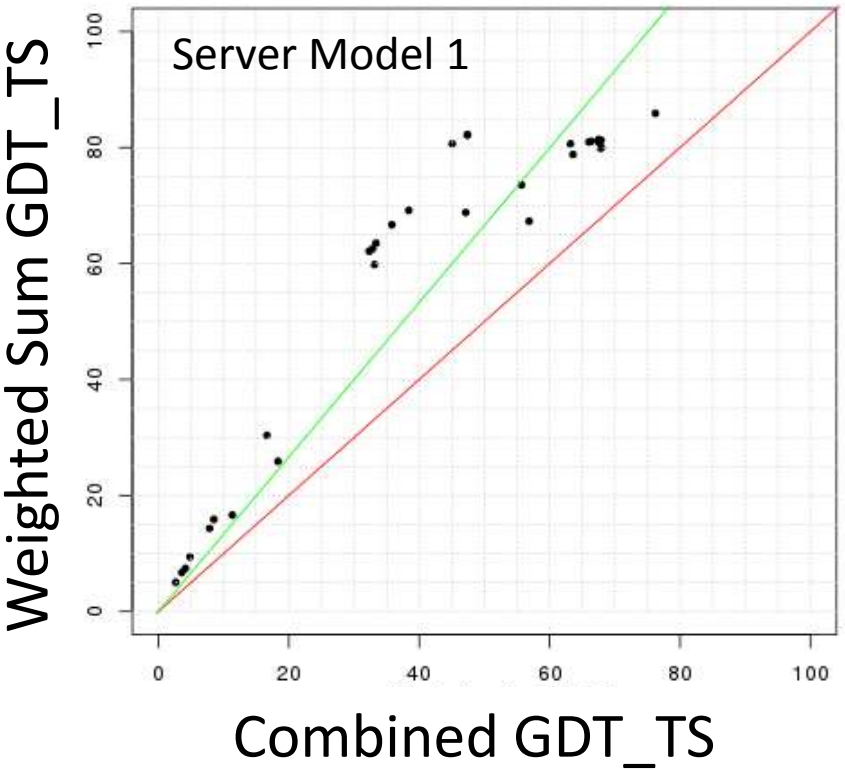
5-enolpyruvylshikimate-3-phosphate



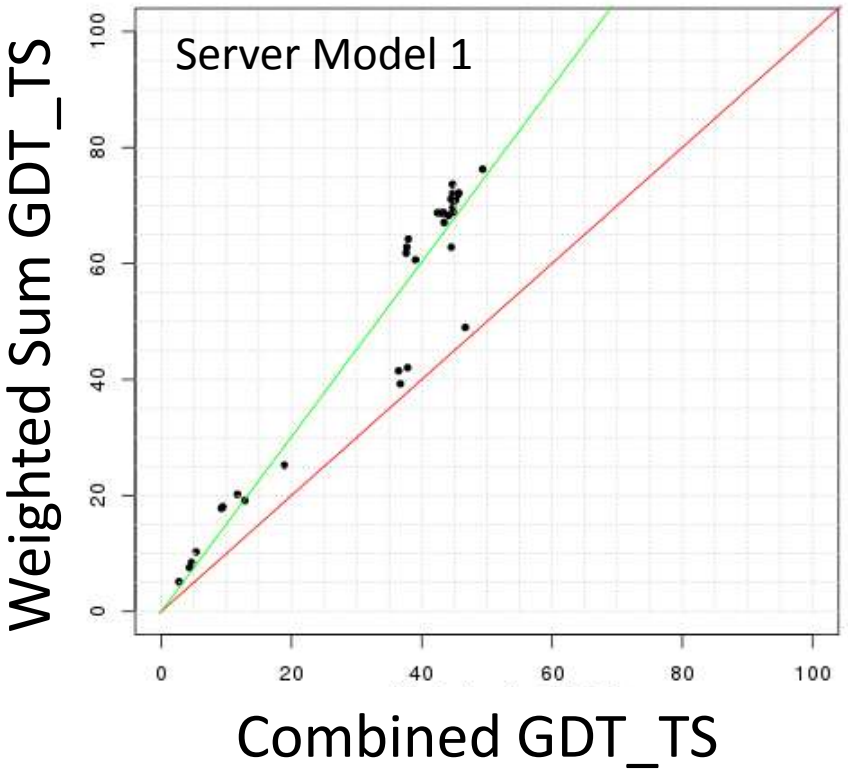
T0999: Difficulty in Domain Splitting

1

Grishin Plot: Doman4&5



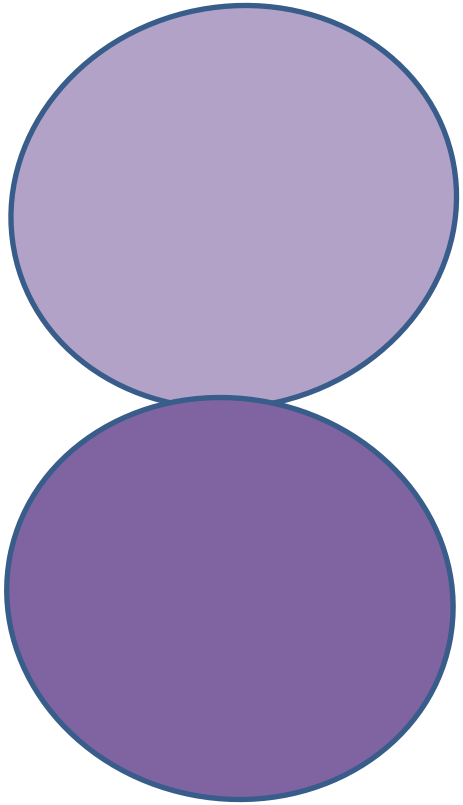
Grishin Plot: Domain3&5



T0999: EPSP synthase Conformation Change

1

T0999
Domain 2



Open
template

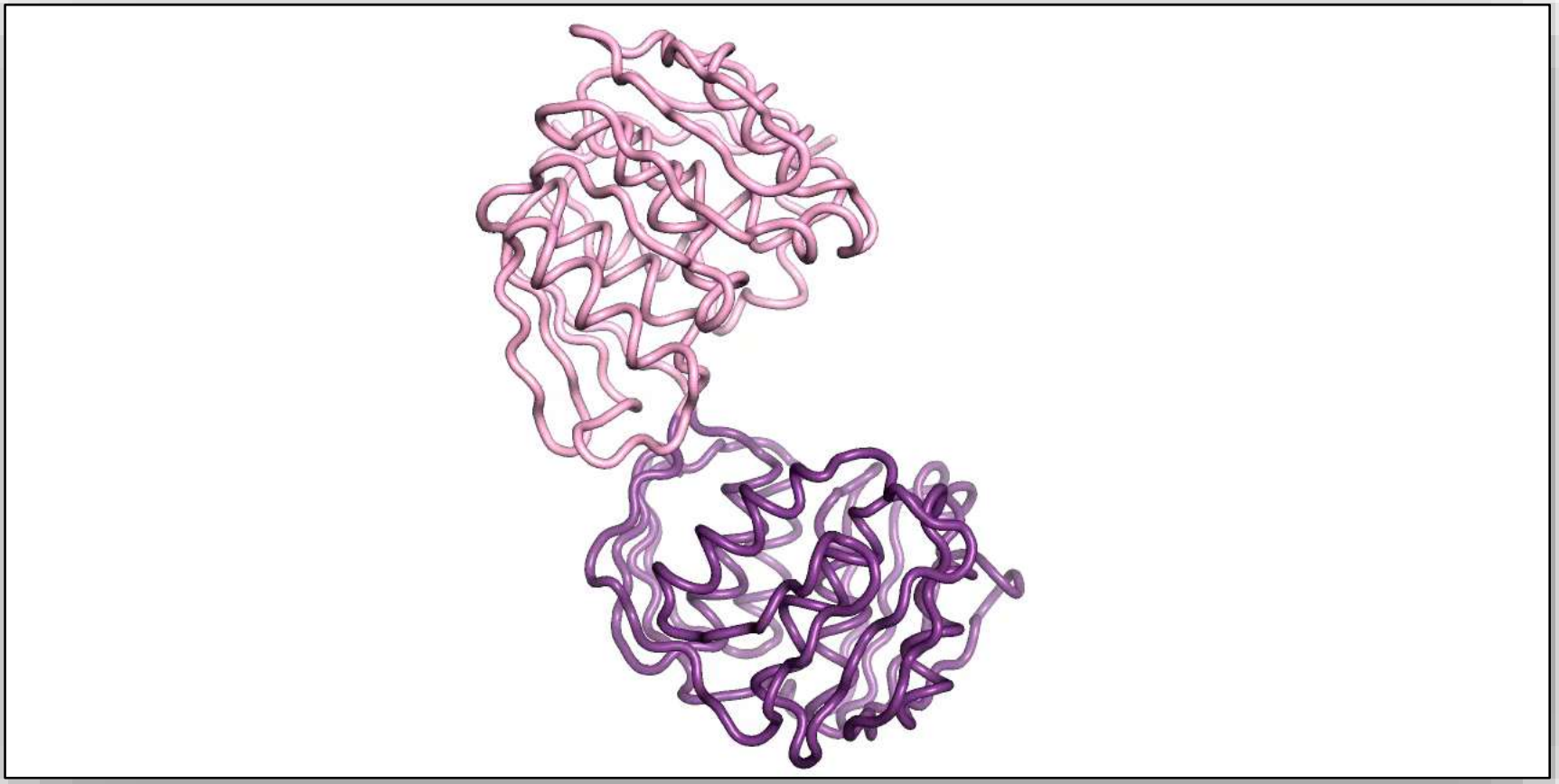
Closed
template

T0999: EPSP synthase Conformation Change

T0999 Domain 2

Open template

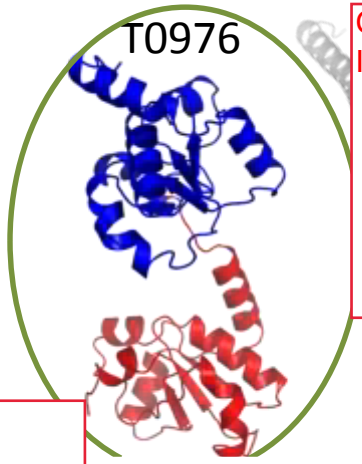
Closed template



Splits (Duplications) 1

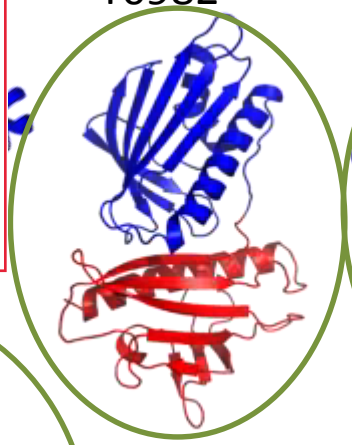
T0957s1

T0976

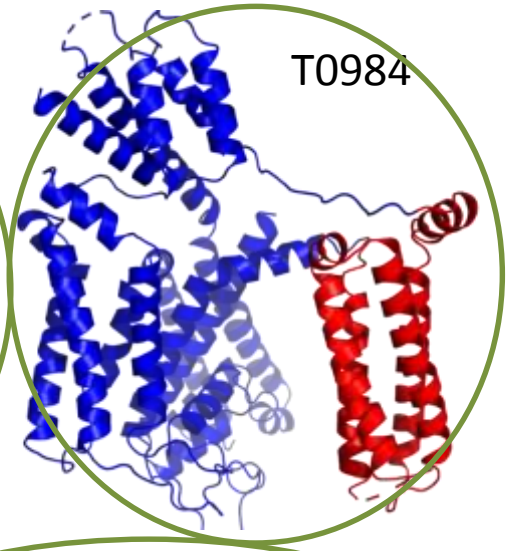


CASP:
Image redacted

T0982



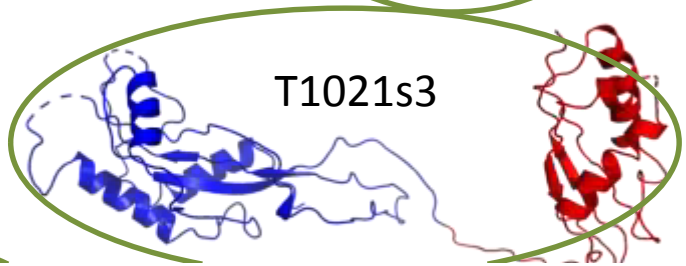
T0984



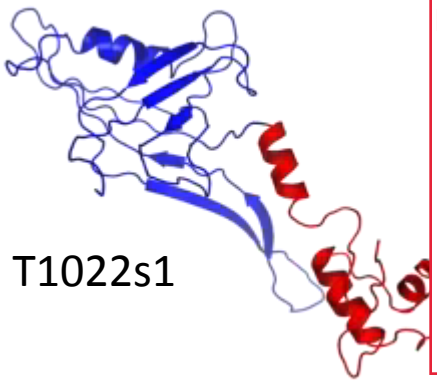
CASP:
Image redacted

CASP:
Image redacted

T1021s3



T1022s1



CASP:
Images redacted

CASP13 Targets: Domains in Numbers

Split Targets into Multiple EVUs

- Split 19 Targets
- 52 EVUs (many duplications)

Multidomain Targets not Split

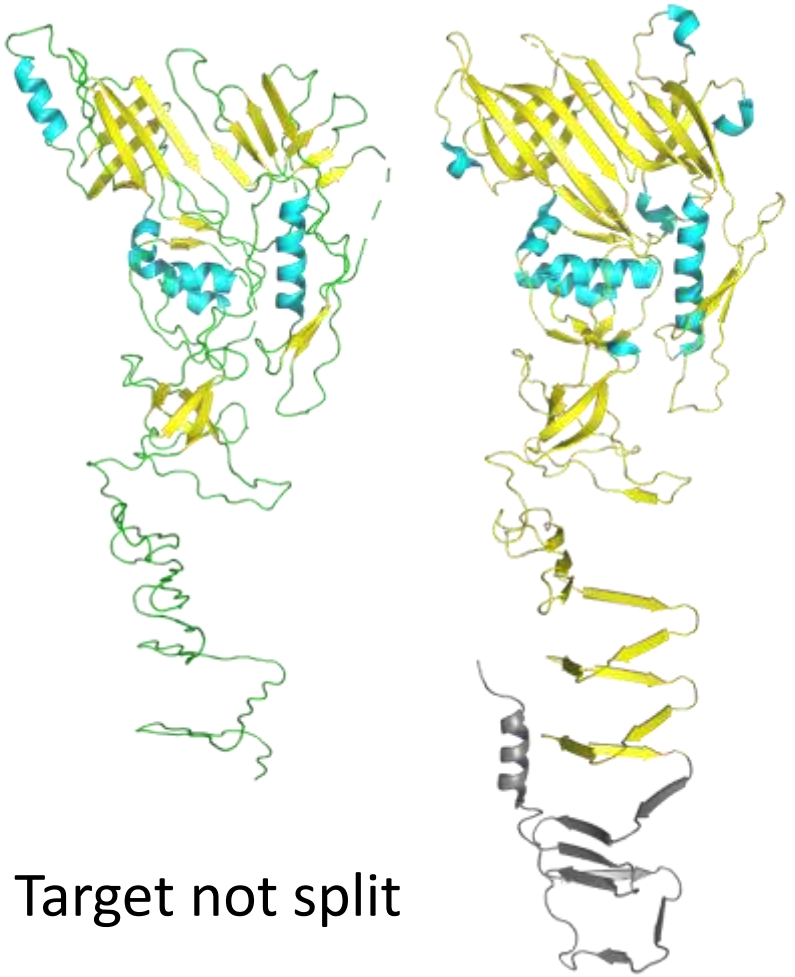
- 13 Targets (3 EVUs)
- 37 Unsplit Domains (6 unsplit EVUs)

Single Domain Targets

- 46 Targets
- Include extended segments, many with elaborations

T1022s2 multidomain

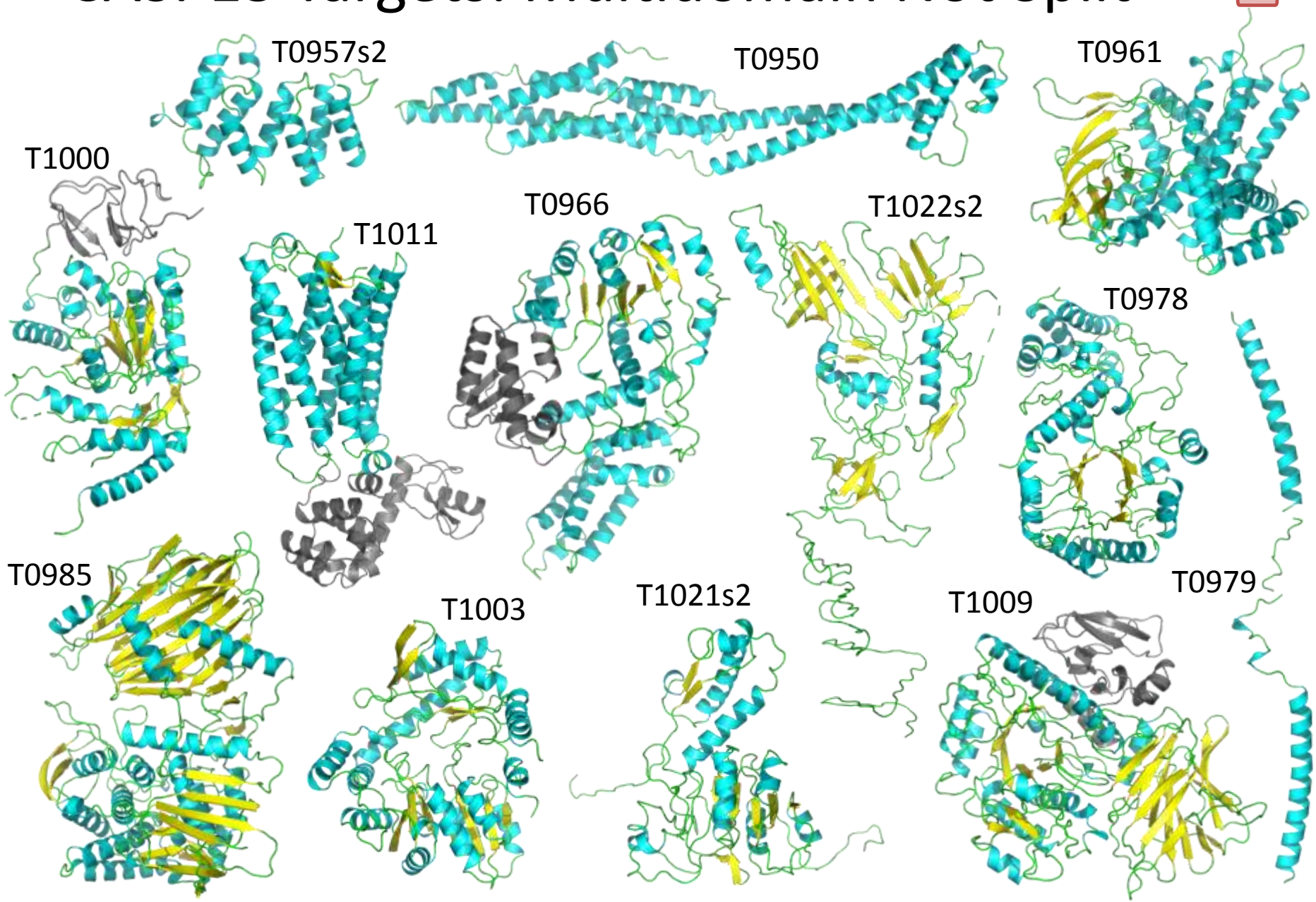
4mtk template



Target not split

CASP13 Targets: Multidomain Not Split

1



CASP13 Targets: Single Domain EVUs

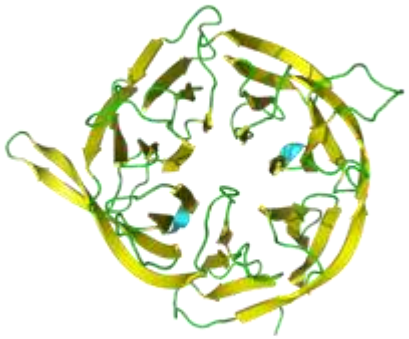
1

T0949

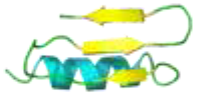
T0951

T0953s1

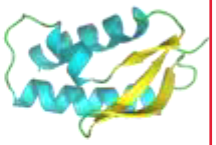
T0954



T0955



T0958



T0959

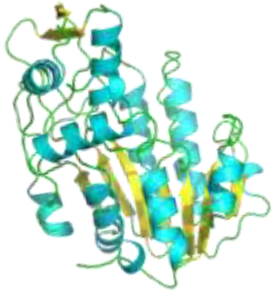
CASP:
Image redacted

T0962

CASP:
Image redacted

T0964

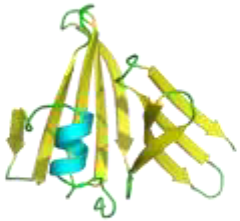
T0965



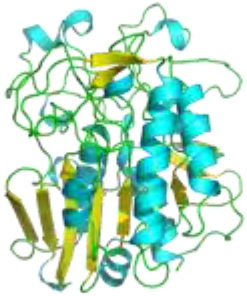
T0967

T0968s1

T0968s2



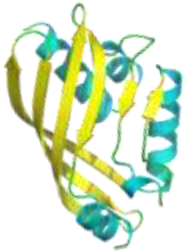
T0969



T0970

CASP:
Image
redacted

T0971



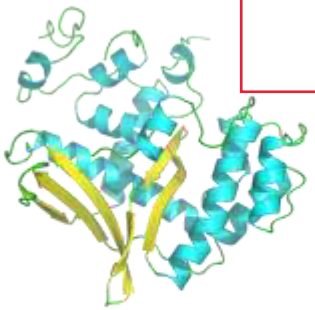
T0973

CASP:
Image redacted

T0974s1

CASP:
Image redacted

T0975



T0980s1

T0983

T0986s1

T0986s2

CASP13 Targets: Single Domain EVUs

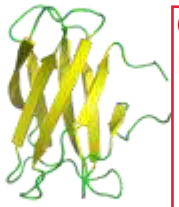
1

90

T0991

CASP:
Image
redacted

T0992

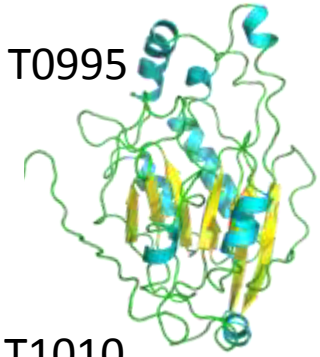


T0993s1

T0993s2

CASP:
Images redacted

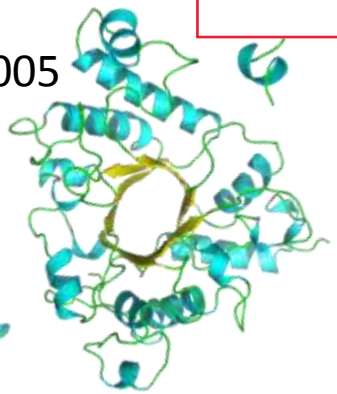
T0995



T0997

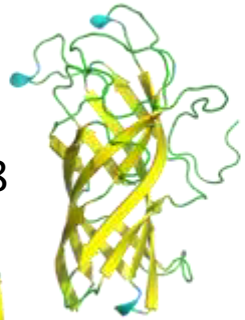
CASP:
Image redacted

T1005

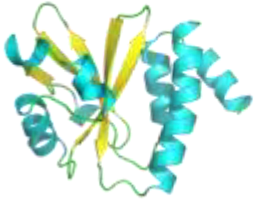


T1006

T1010



T1001



T1013

CASP:
Image redacted

T1015s1

T1015s2

CASP:
Images redacted

T1008

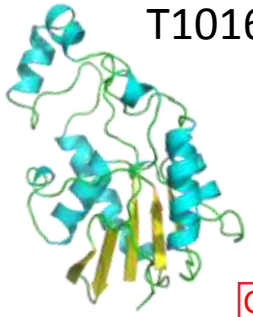
T1017s2

CASP:
Image redacted

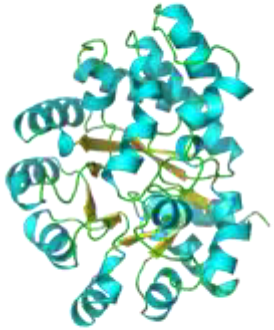
T1017s1

CASP:
Image redacted

T1016



T1018



T1019s2

CASP:
Image redacted

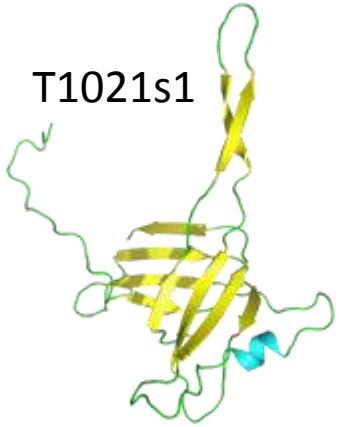
T1019s1

CASP:
Image redacted

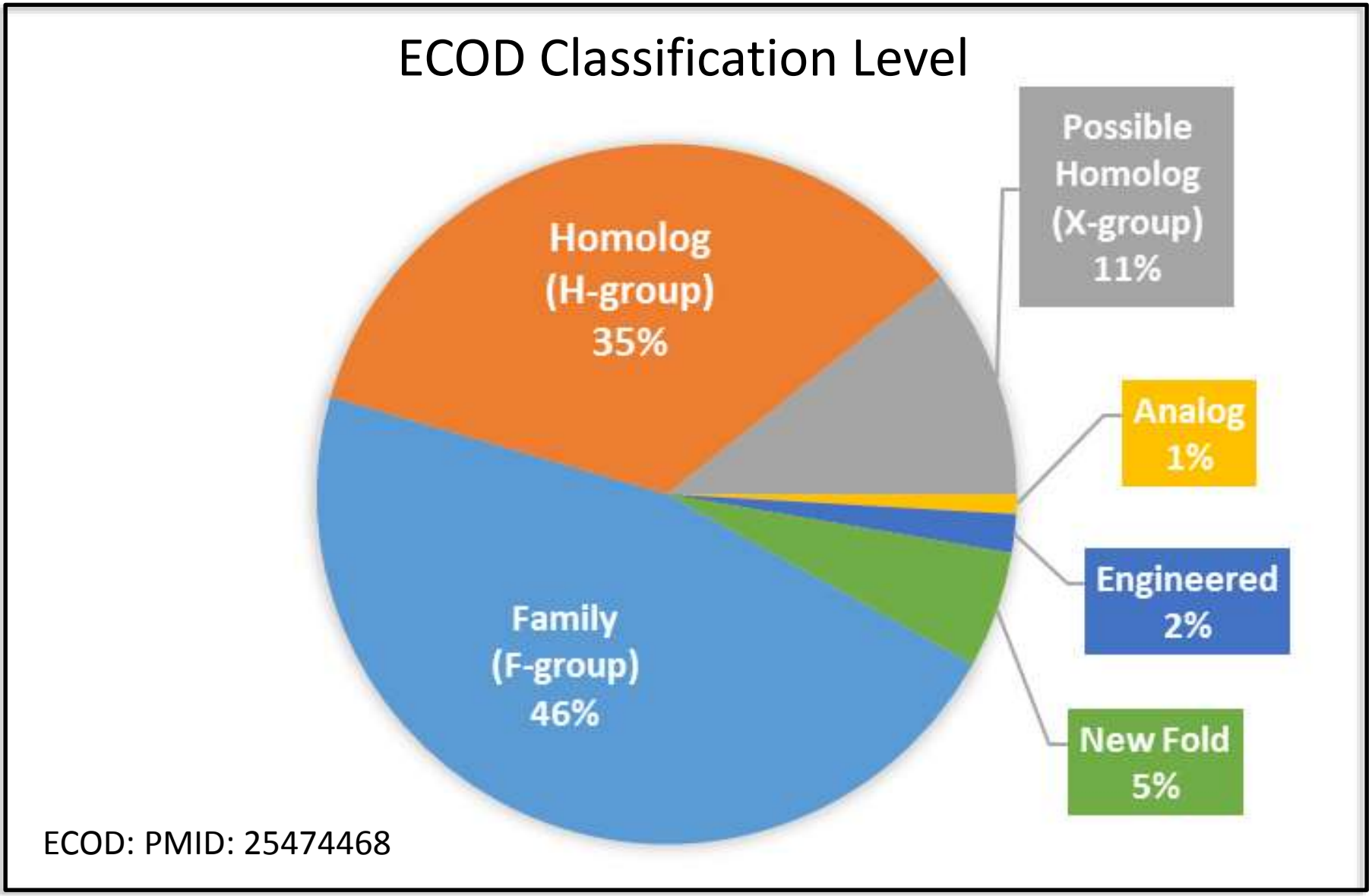
T1020

CASP:
Image redacted

T1021s1



CASP13 Targets: Position in Evolution

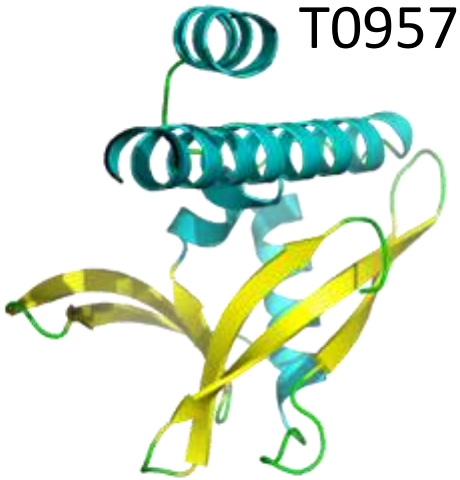


ECOD: PMID: 25474468

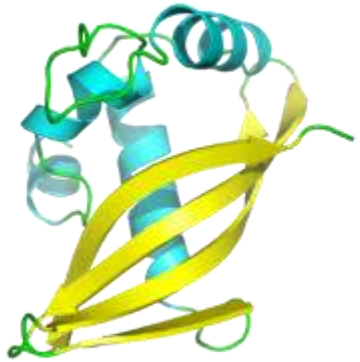
CASP13 Targets: Homology Level Classification 2

Fast evolving RelE Nuclease Toxins

T0957s1



T0968s1



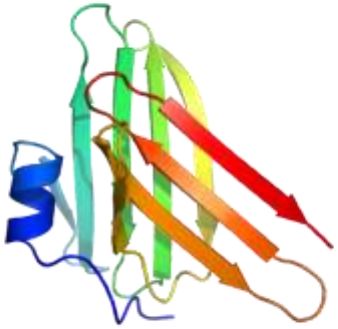
T0980s1

3kxe RelE

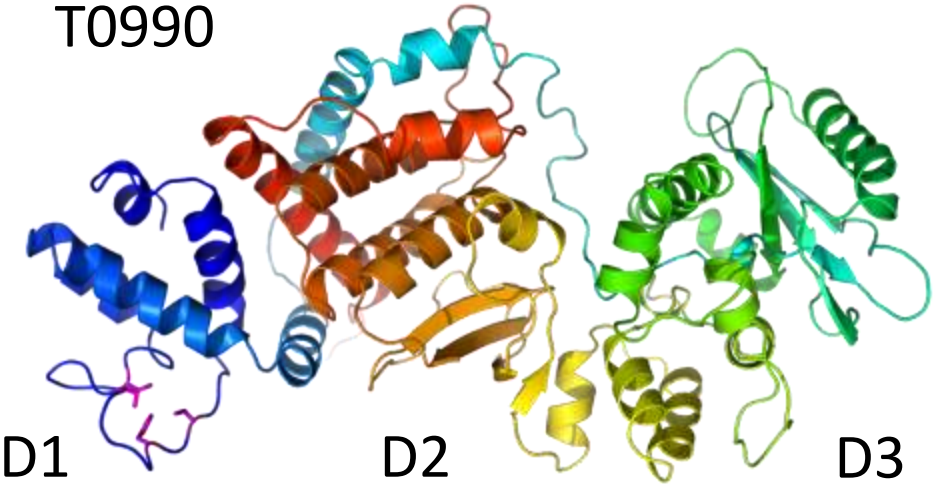
5mje RelE

CASP13 Targets: 6 Potential New Folds

T0968s2



T0990

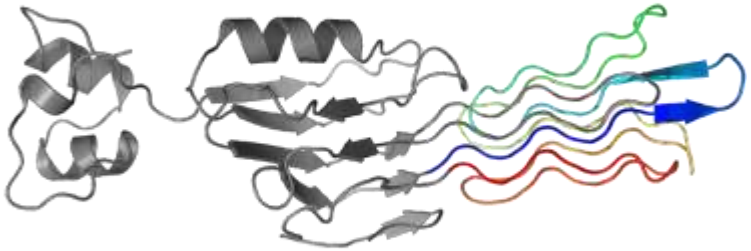


D1 D2 D3

T1017s2

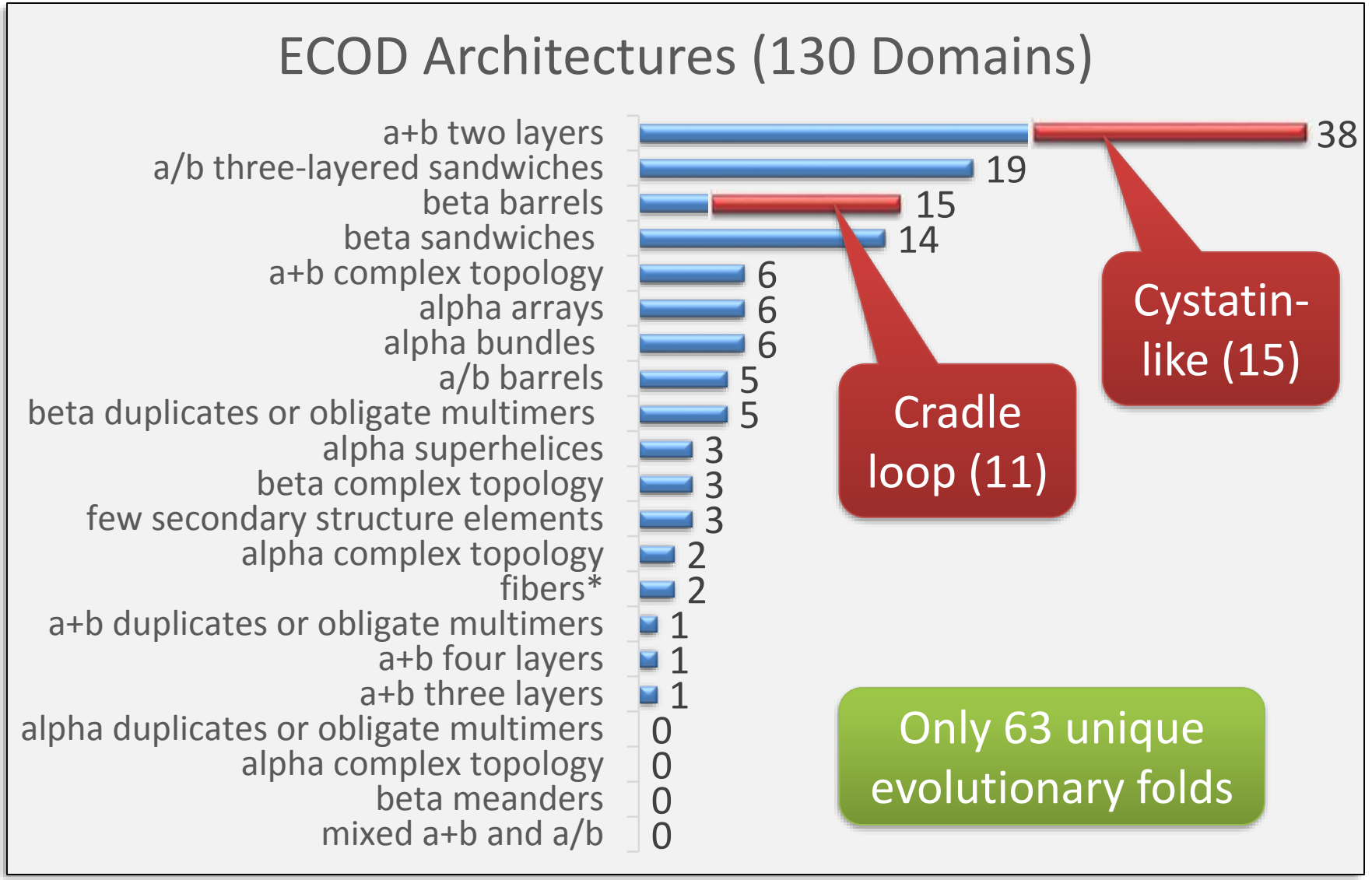
CASP:
Image redacted

T0953



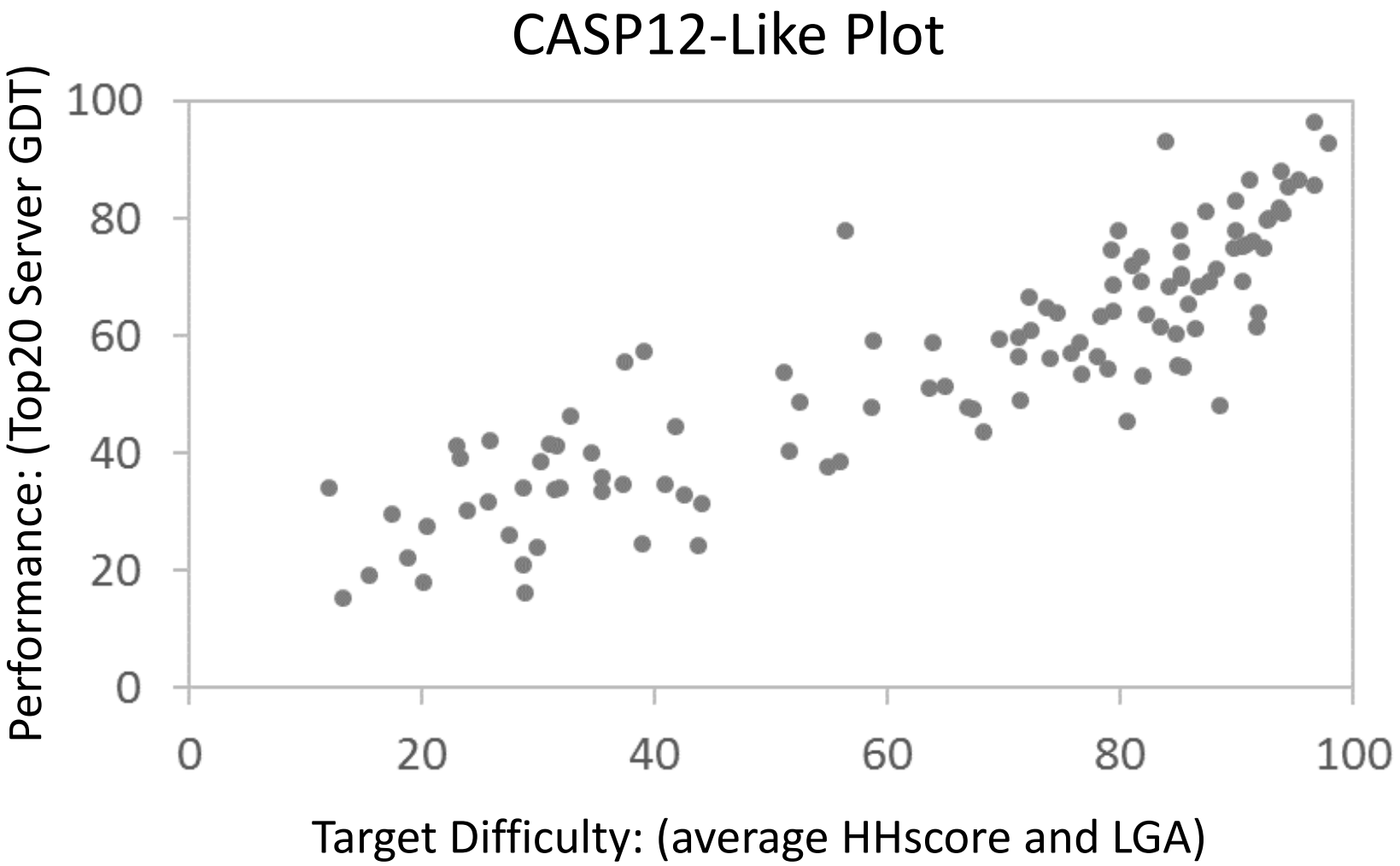
D1 D2 D3

CASP13 Targets: Position in Evolution



CASP13 Targets: Assign EVUs into Categories

2



Where to make the boundary?

CASP13 Targets: Establish Classification Bounds 2

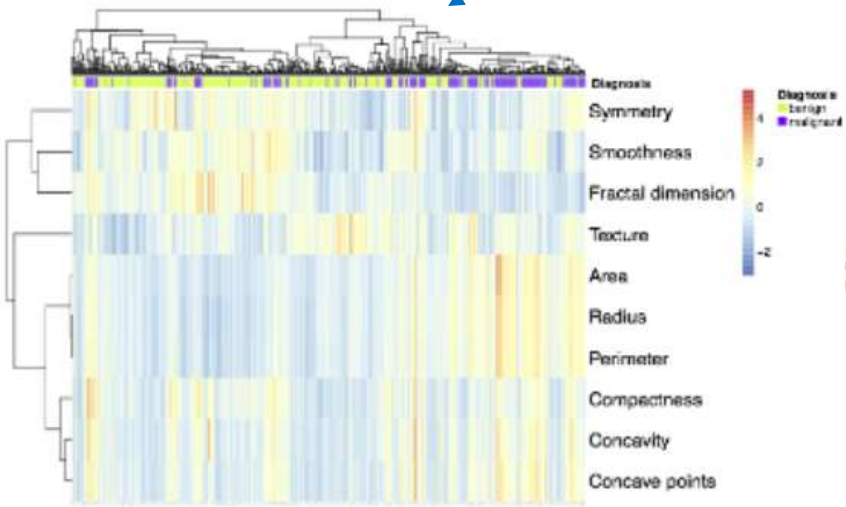
W566–W570 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* *Published online 12 May 2015*
doi: 10.1093/nar/gkv468

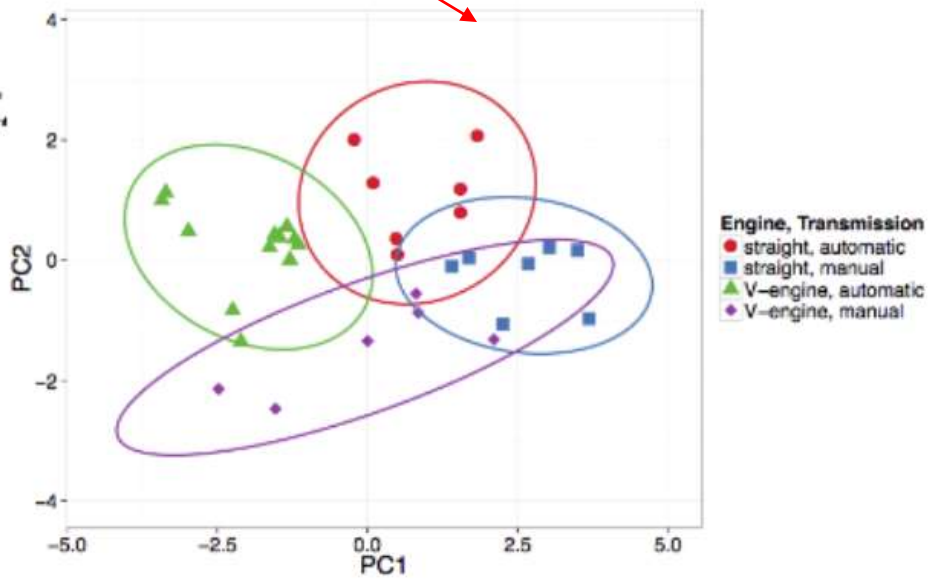
ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap

Tauno Metsalu and Jaak Vilo*

Heatmap: Data Clustering

PCA: Distribution of Measures





CASP13 Targets: ClustVis Clustering Methods

Choose *Measures* from Prediction Center:

HHPRED Probability

HHPRED Coverage

HHscore (Probability*Coverage)

LGA template

Performance (Top20 Server First Models: average GDT_TS)

Performance (All server models: average GDT_TS)

Performance (Top server model GDT_TS)

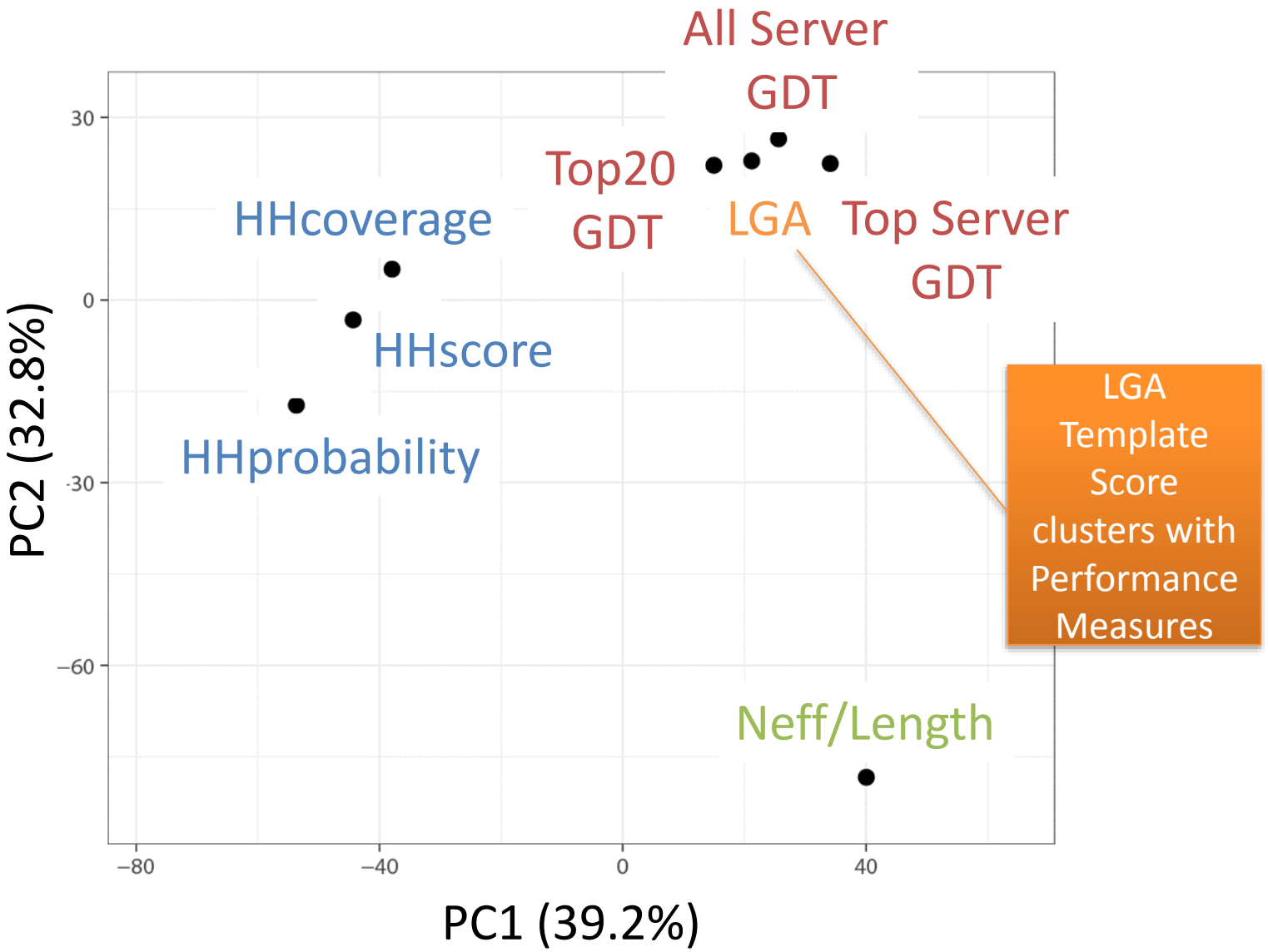
Neff from Max (psi, Hhblitz) Neff/len per domain

Transform into T-scores (for every Target = 111 EVUs)

$$T_{Target} = (Target_{measure} - Average_{measure}) / (St.Dev._{measure}/sqrt(111))$$

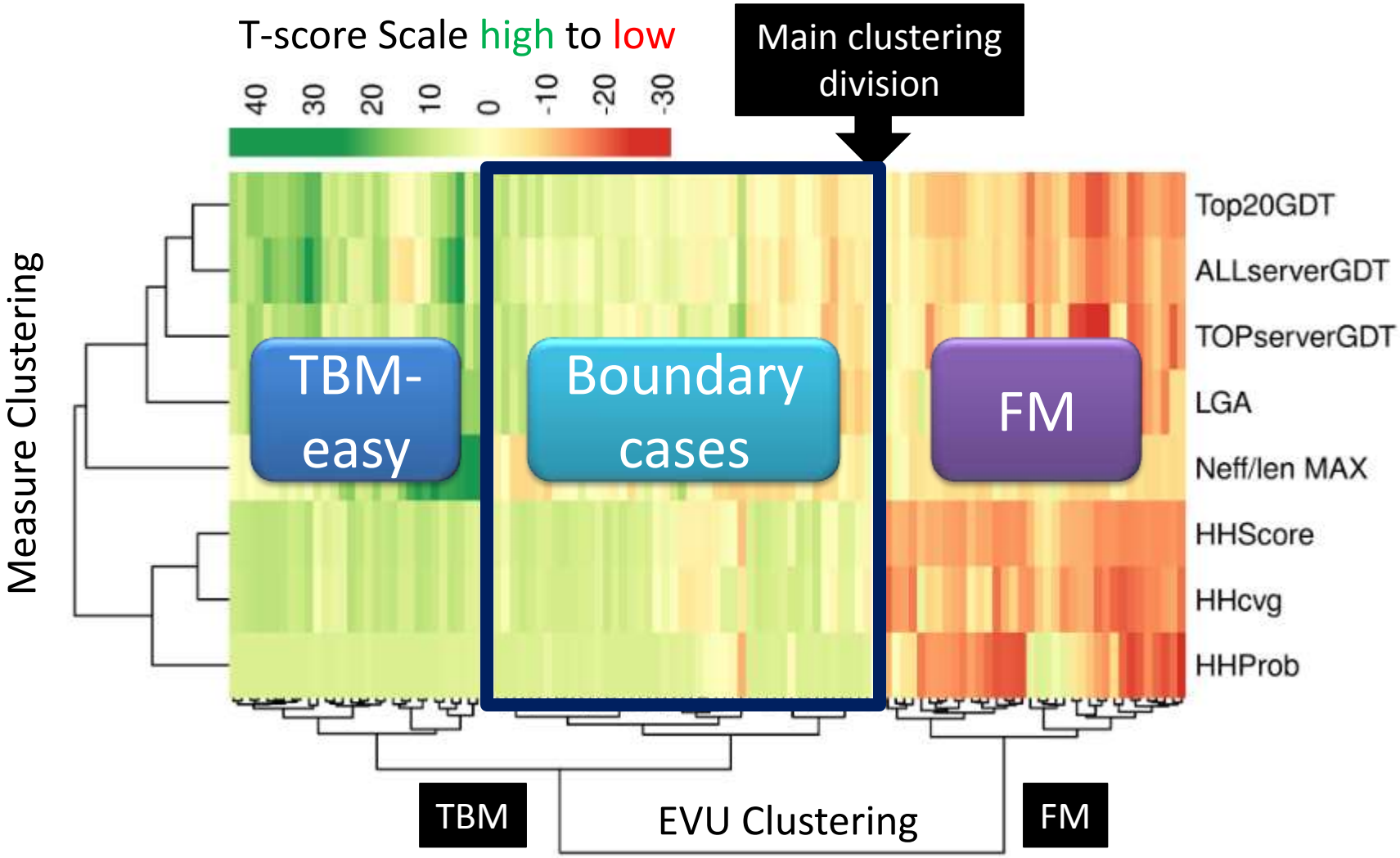
Same as Z-score except

CASP13 Targets: PCA on unscaled T-scores 2

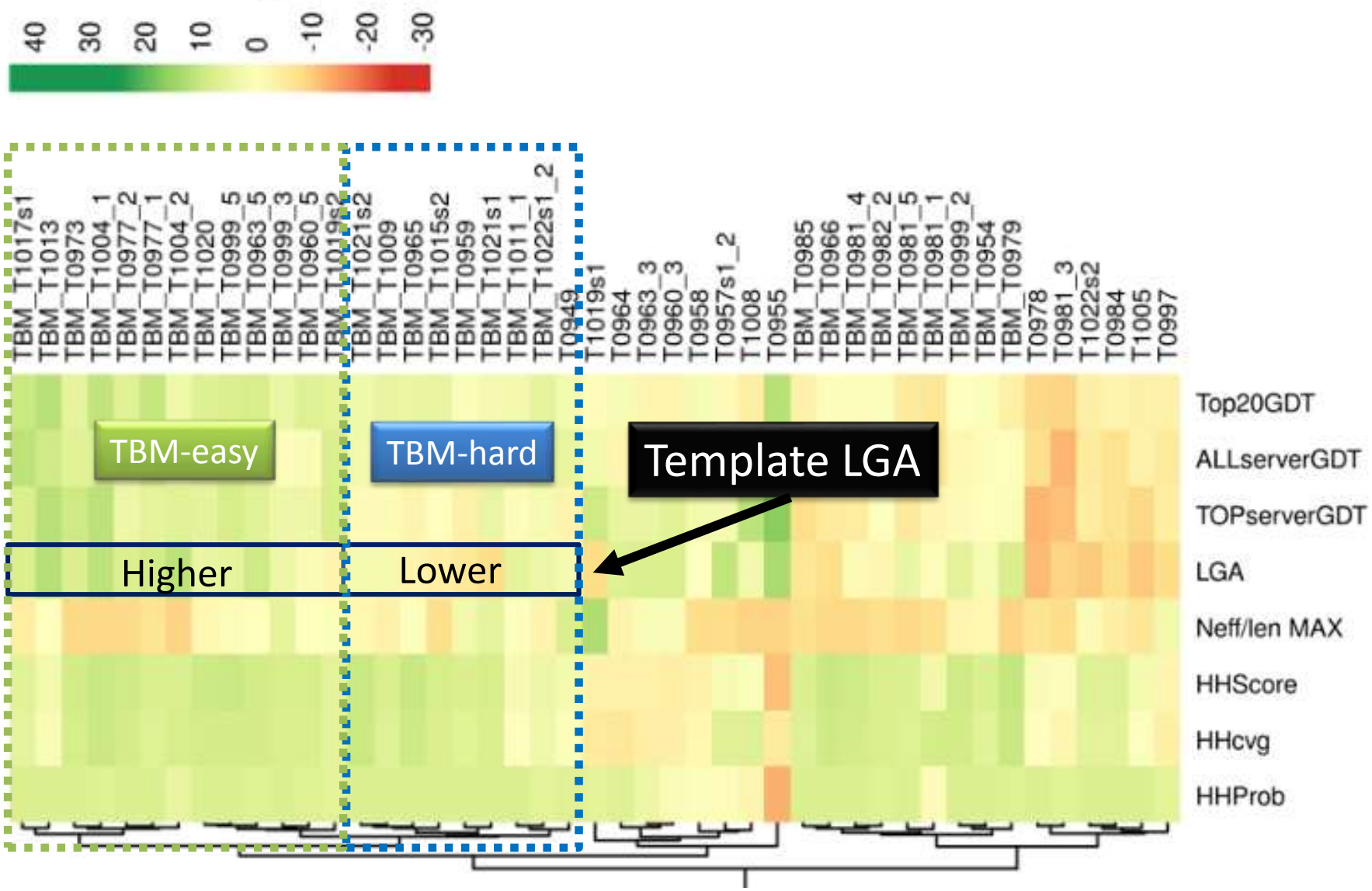


CASP13 Targets: Clustered Heatmap

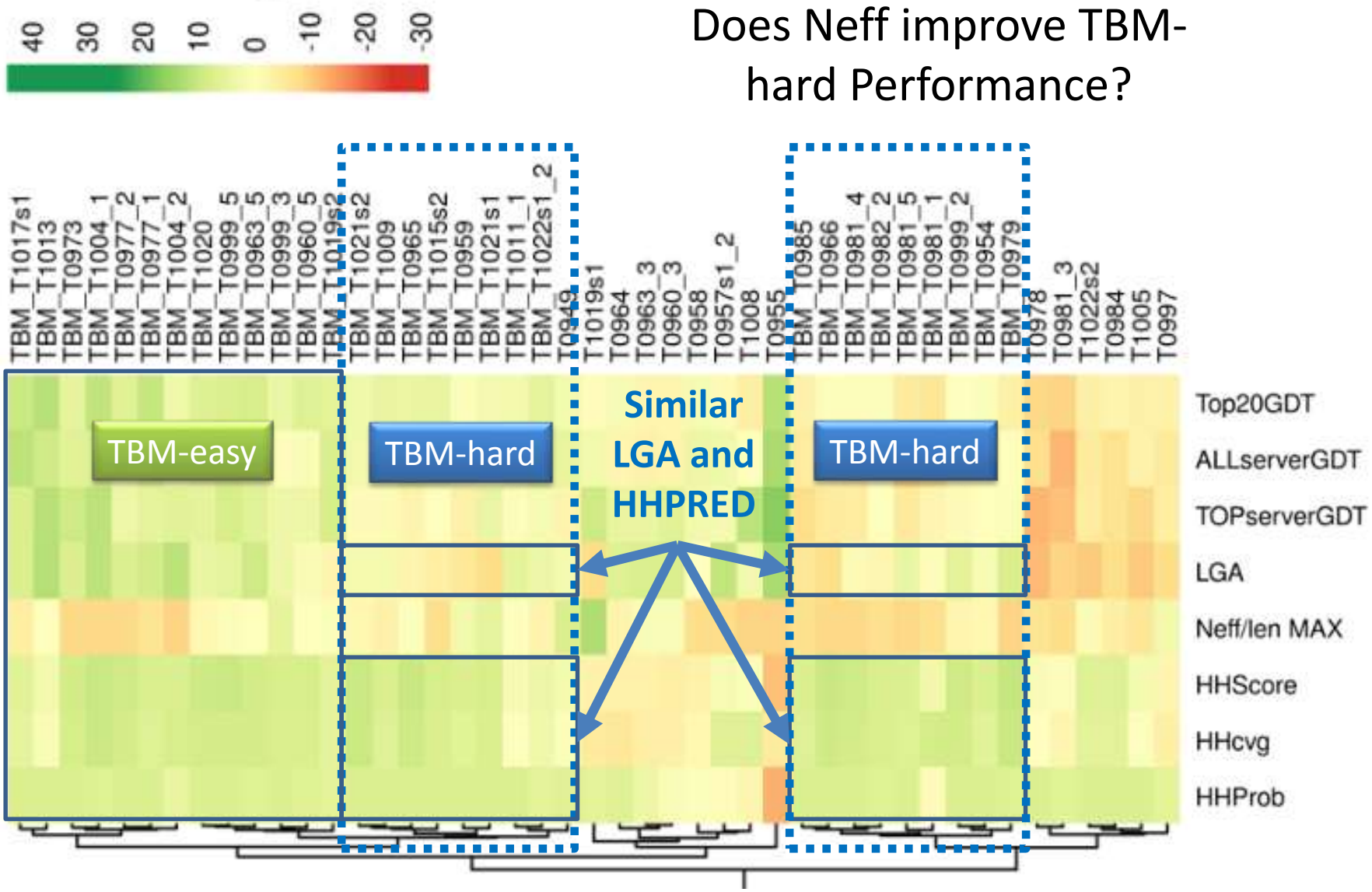
8 Rows and 111 Columns: clustered by Euclidean Distance and Ward Linkage



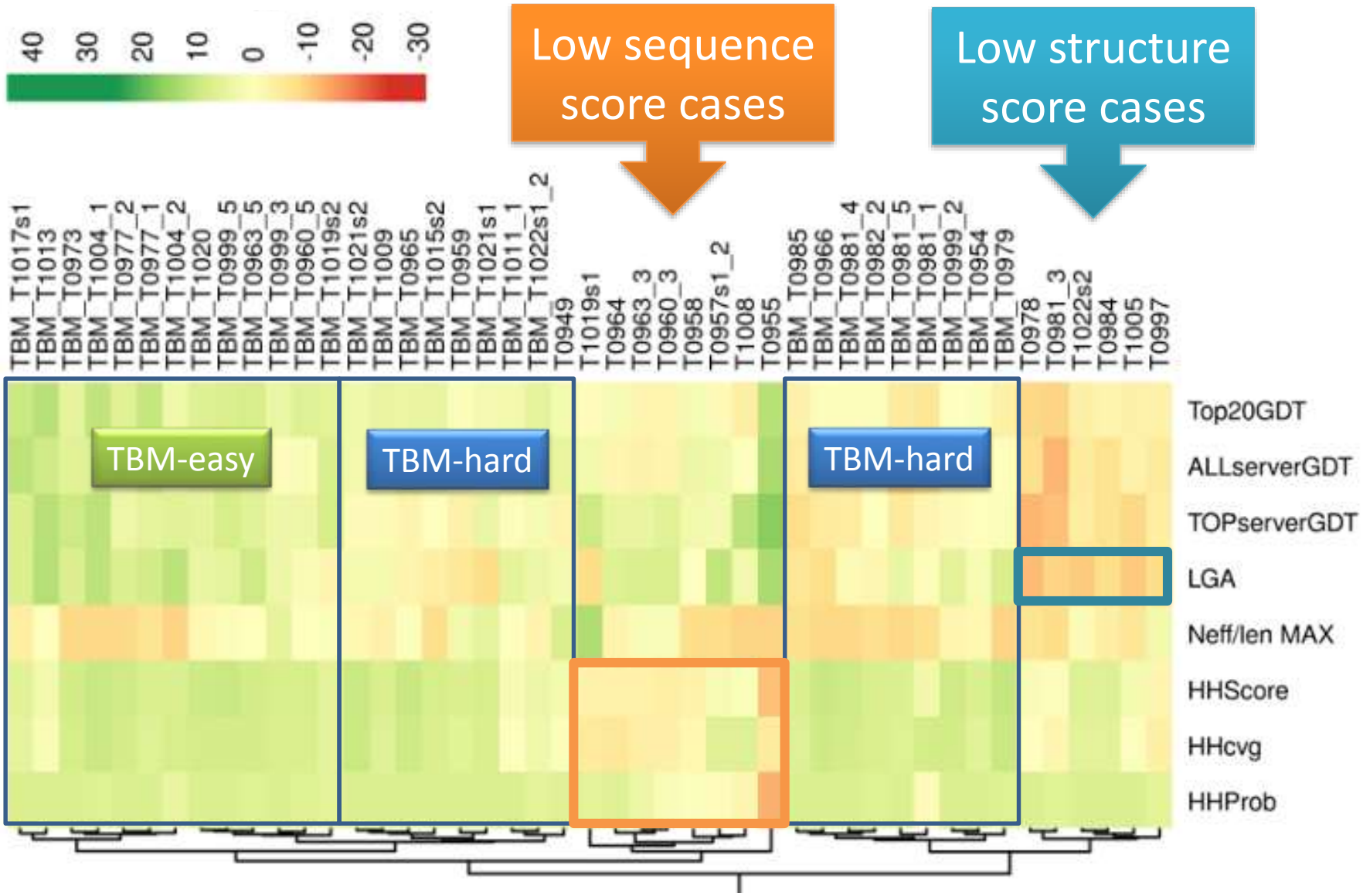
Boundary EVU Clusters Define Categories 2



Boundary EVU Clusters Define Categories 2

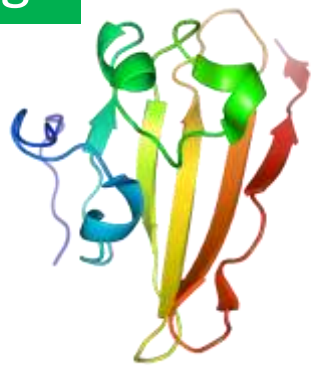


Boundary EVU Clusters Require Inspection 2



Boundary Cases: Low Sequence Score Examples 2

Homologs



T0960_3/T0963_3 Top LGA: 5nxfB

HPRED 82% (0.54) LGA 83.2

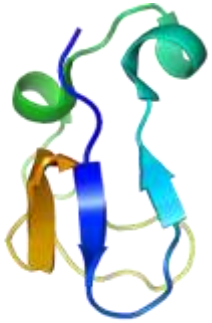
Homologs

T0964 Top LGA: 5t7aA

HHPRED 98% (0.47) LGA 81.2

Homologs

CASP:
Image redacted



T01019s1 Top Hhpred: 6rxnA

(permuted rubredoxin) rubredoxin

HHPRED 95.6% (0.5) LGA 54.3

Homologs

T0958 Top HHpred: 2kim

(SSE shifts in HTH) HTH

HHPRED 81.4% (0.59) LGA 69.2

Boundary Cases: Low Sequence Score Examples 2

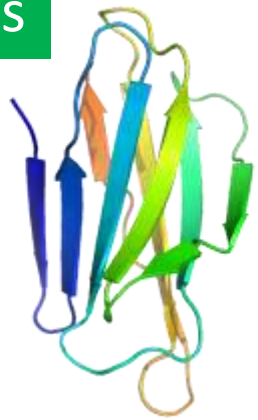
TBM-hard

Homologs

T0960_3/T0963_3 Top LGA: 5nx

HPRED 82% (0.54) LGA 83.2

Homologs



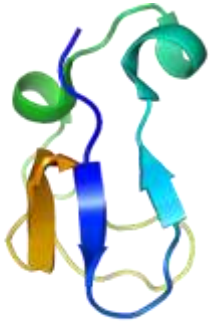
T0964 Top LGA: 5t7aA

HHPRED 98% (0.47) LGA 81.2

TBM/FM

Homologs

CASP:
Image redacted



T01019s1 Top Hhpred: 6rx

(permuted rubredoxin) rubredoxin

HHPRED 95.6% (0.5) LGA 54.3

Homologs

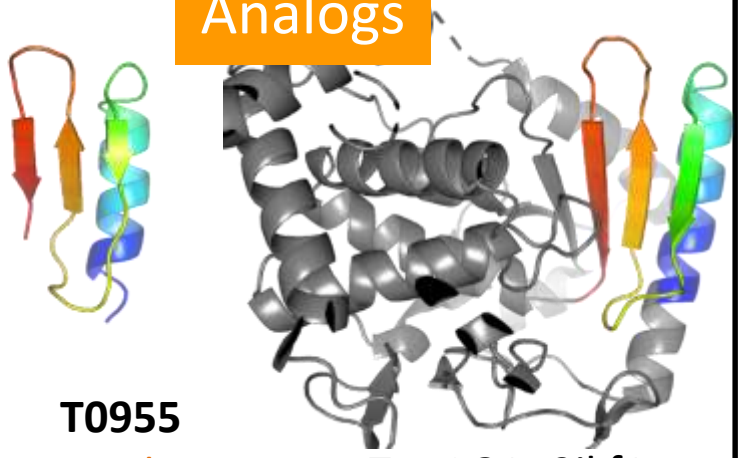
T0958 Top HHpred: 2kim

(SSE shifts in HTH) HTH

HHPRED 81.4% (0.59) LGA 69.2

Boundary Cases: Low Sequence Score Examples 2

Analogs

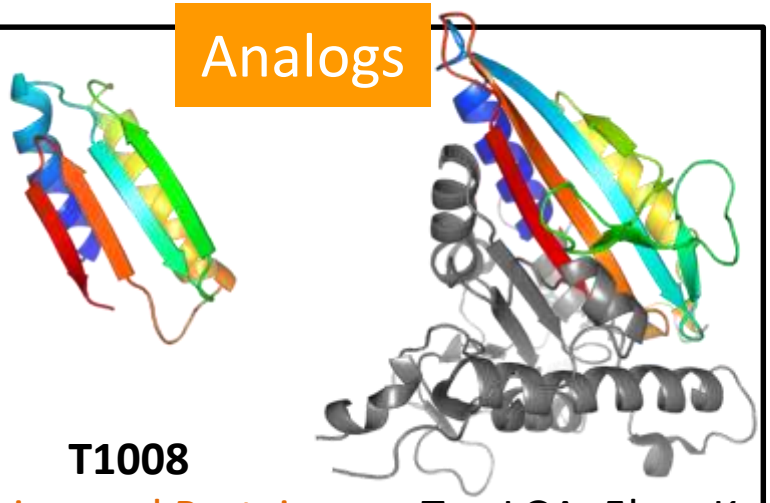


T0955

Engineered Protein Top LGA: 2ikfA

HHPRED 39.9% (0.49) LGA (93.4)

Analogs

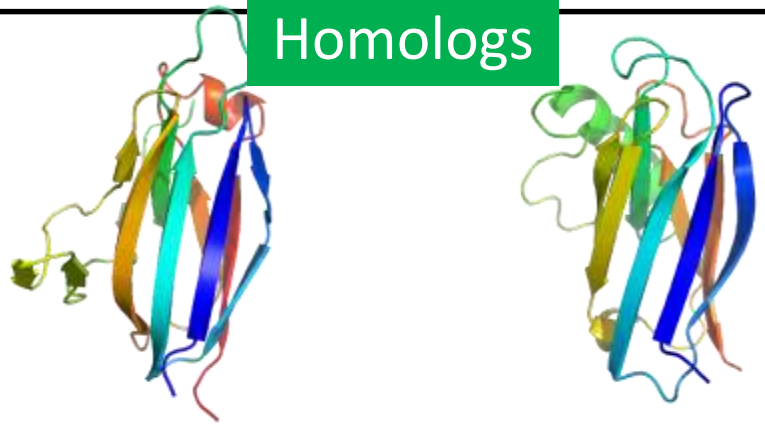


T1008

Engineered Protein Top LGA: 5hnwK

HHPRED 71.6% (0.85) LGA 73.9

Homologs



T0949 Top HHpred: 1azcB

HHPRED 99.5% (0.67) LGA 72.9

Analogs

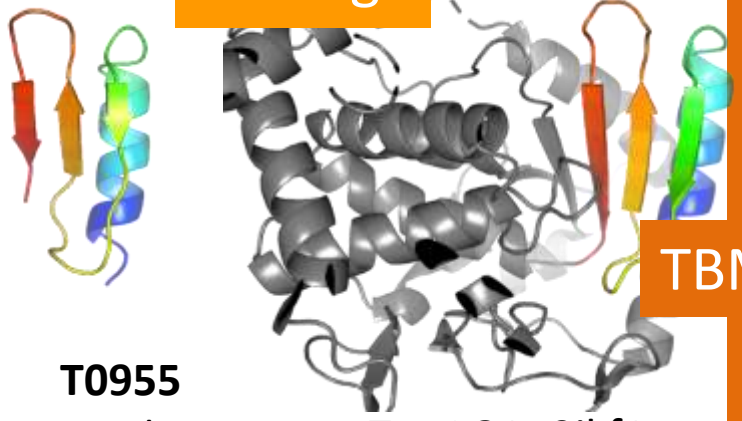
T0957s1_2 Top LGA: 6cudA

Simple Structure Transmembrane

HHPRED 75.7% (0.85) LGA 88.97

Boundary Cases: Low Sequence Score Examples 2

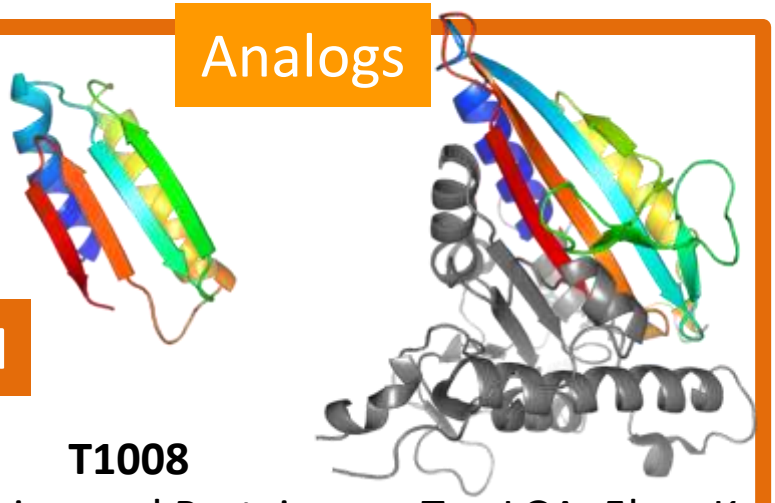
Analogs



TBM/FM

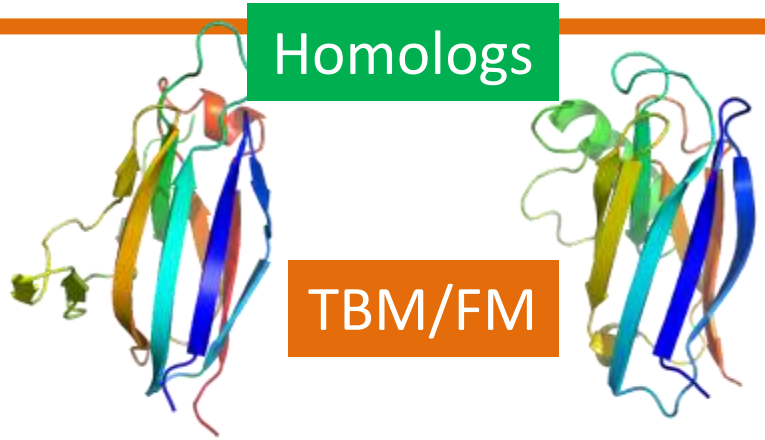
T0955
Engineered Protein Top LGA: 2ikfA
HHPRED 39.9% (0.49) LGA (93.4)

Analogs



T1008
Engineered Protein Top LGA: 5hnwK
HHPRED 71.6% (0.85) LGA 73.9

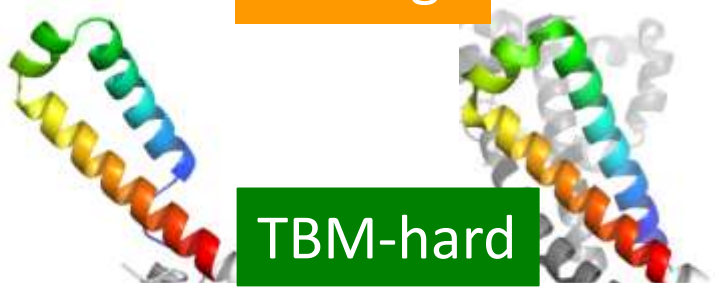
Homologs



TBM/FM

T0949 (requested) Top HHpred: 1azcB
HHPRED 99.5% (0.67) LGA 72.9

Analogs

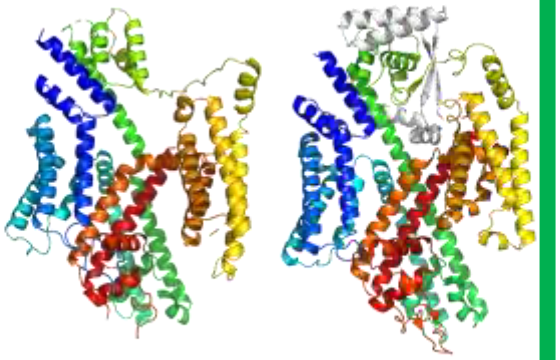


TBM-hard

T0957s1_2 Top LGA: 6cudA
Simple Structure Transmembrane
HHPRED 75.7% (0.85) LGA 88.97

Boundary Cases: Low Structure Score Examples 2

T0984 HHPRED



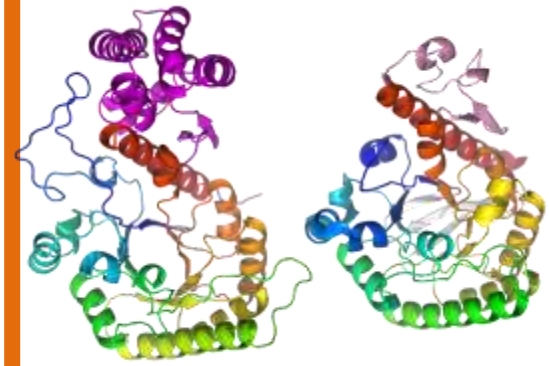
Hhscore 82.4; LGA 54.3

T0981_3 HHPRED

CASP:
Image redacted

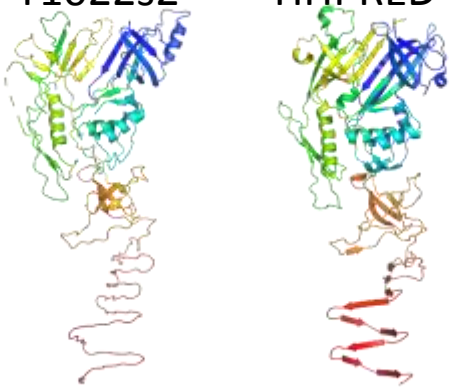
Hhscore 57.2; LGA 52.6

T0978 HHPRED



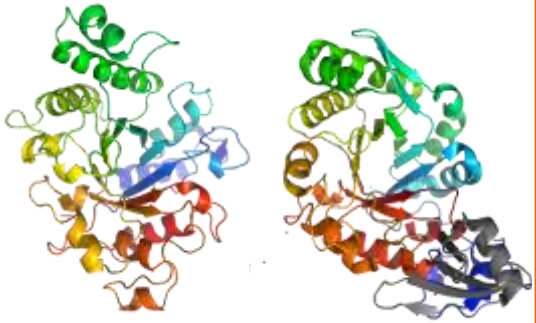
Hhscore 63.9; LGA 48.1

T1022s2 HHPRED



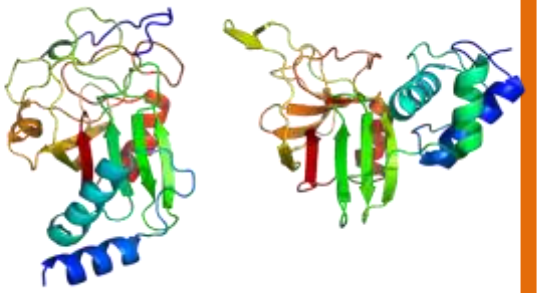
Hhscore 83; LGA 50.9

T1005 HHPRED



Hhscore 66.4; LGA 51.1

T0997 HHPRED

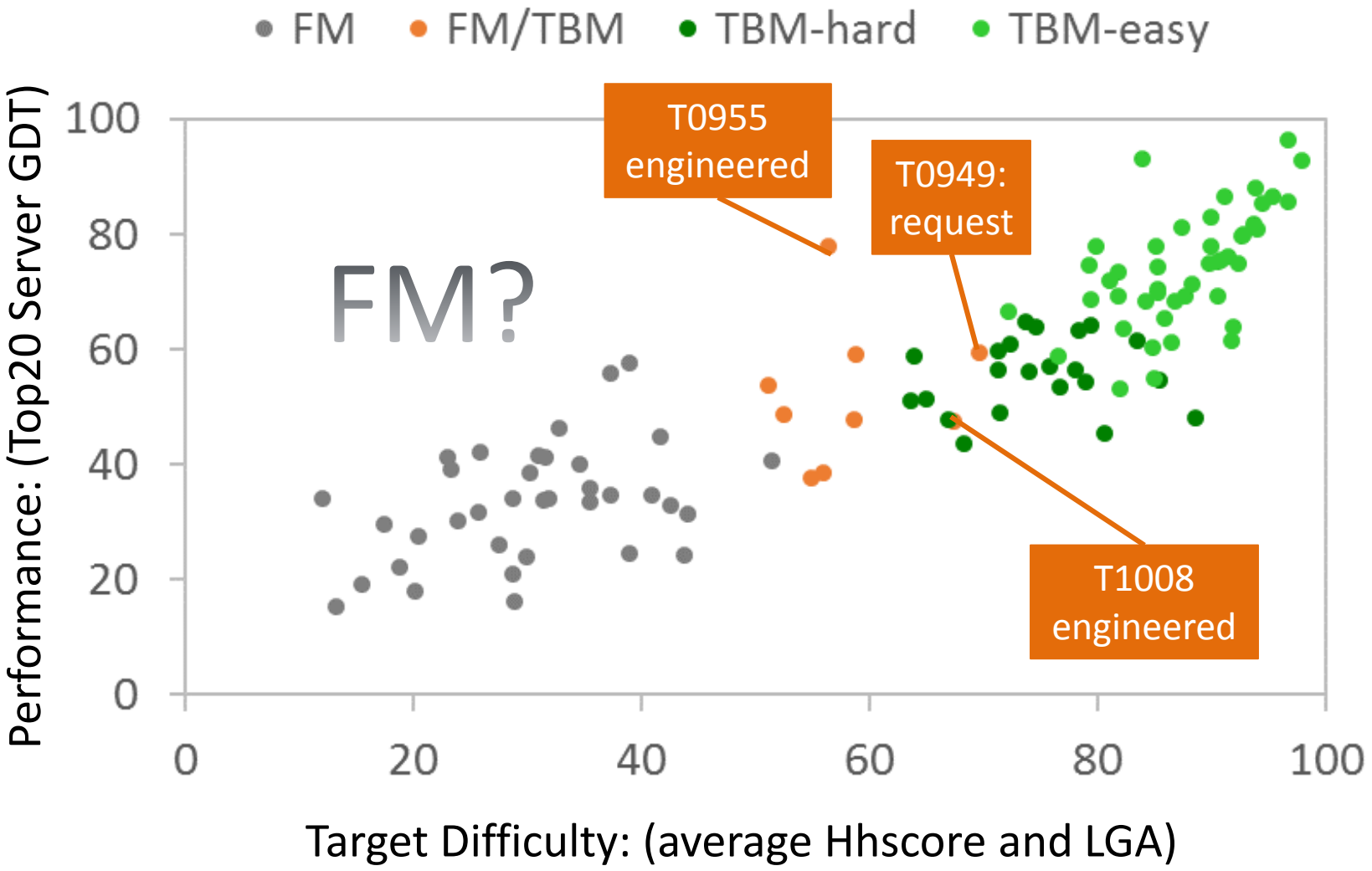


Hhscore 51.1; LGA 53.9

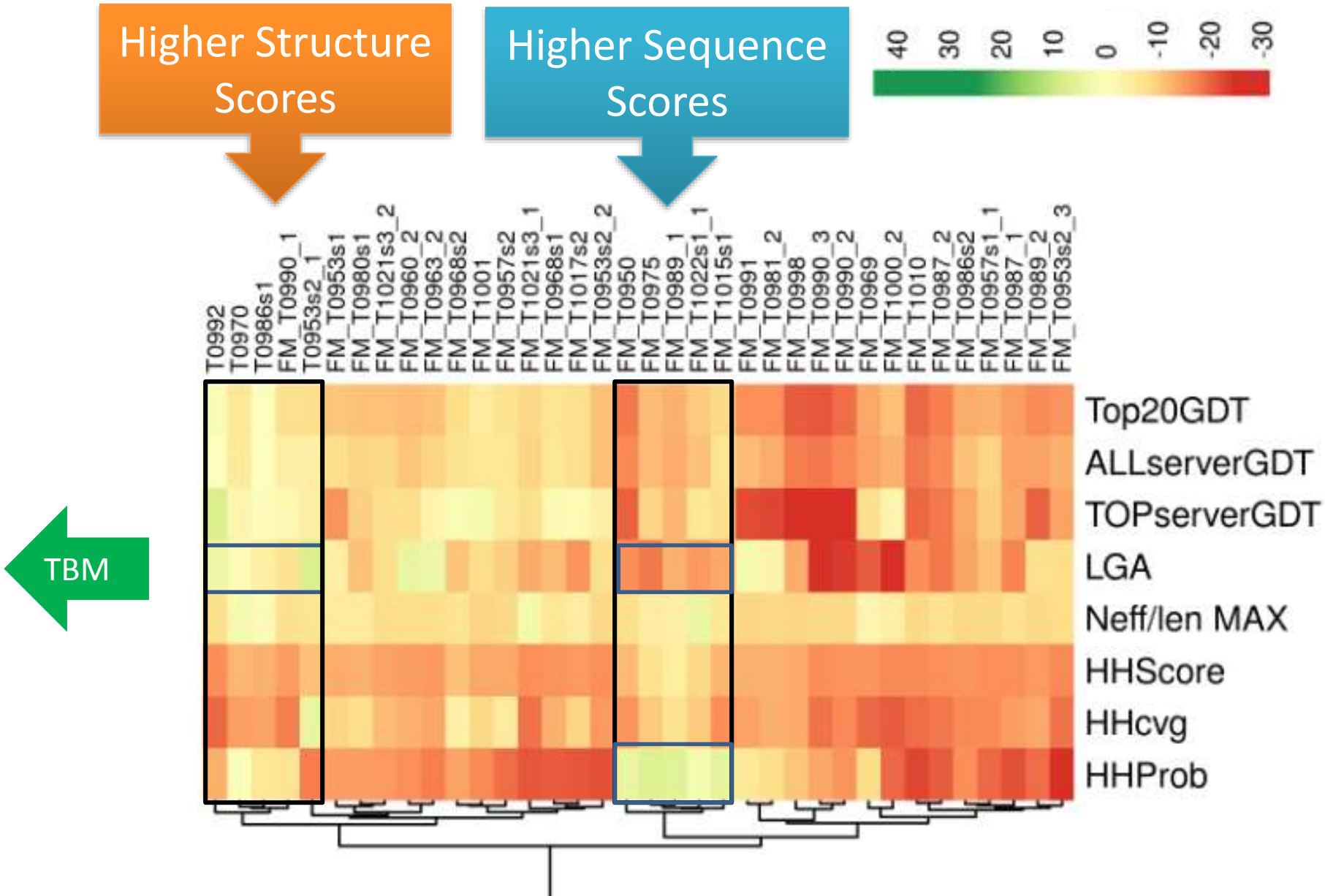
TBM-hard
Relatively small insertions

TBM/FM
More insertions/SSE shifts

Assigned Boundary Cases Mapped To Plot 2



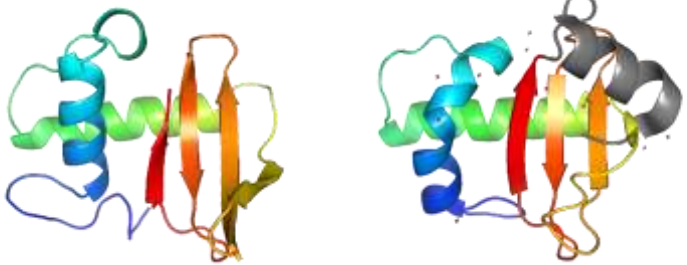
FM Boundary EVU Clusters Require Inspection 2



FM Boundary: High Structure Score Examples

Templates Identified by Some Servers

T0986s1 Top LGA Template

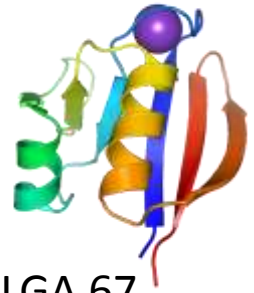


HHPRED 13.7; LGA 61

Homologs: TBM/FM

T0970 Top LGA Template

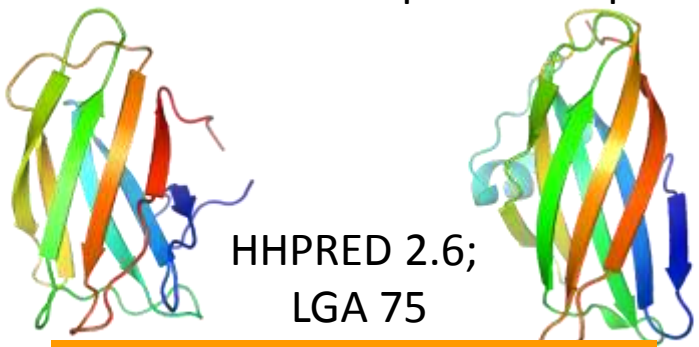
CASP: Image redacted



HHPRED 16.7; LGA 67

Homologs: TBM/FM

T0992 Top LGA Template

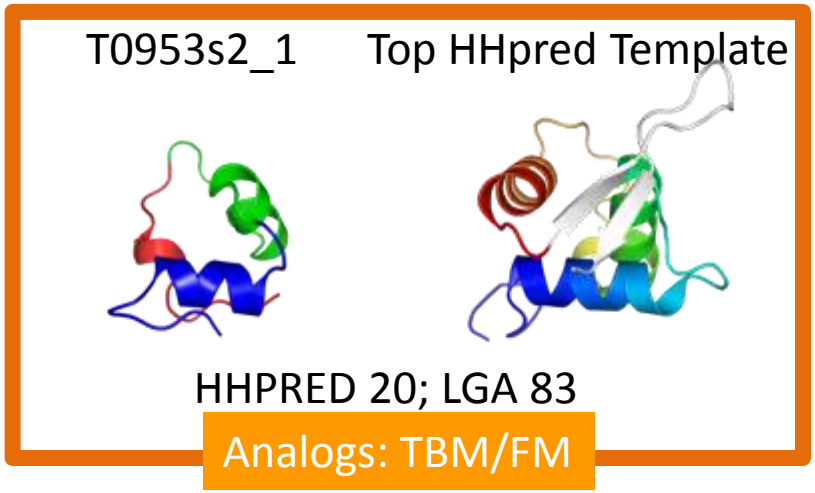


HHPRED 2.6; LGA 75

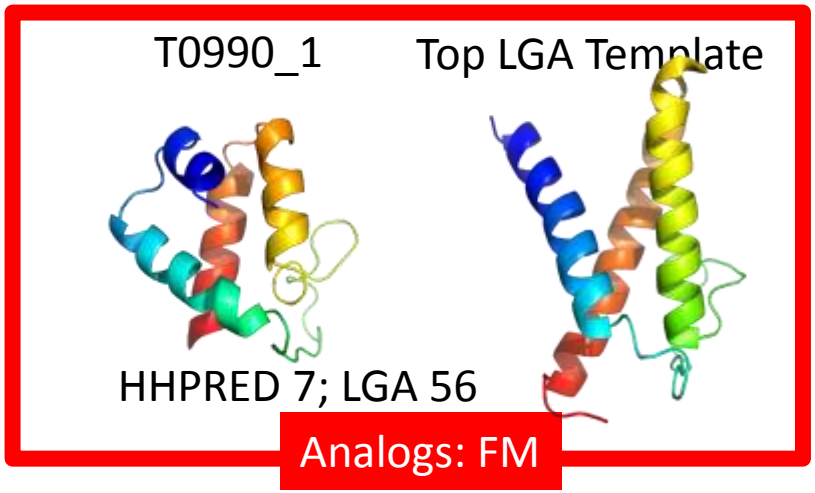
Possible Homologs: TBM/FM

FM Boundary: High Structure Score Examples

2



Simple HHPred Structure Template



Poor HHPred Structure Template; Top LGA not found

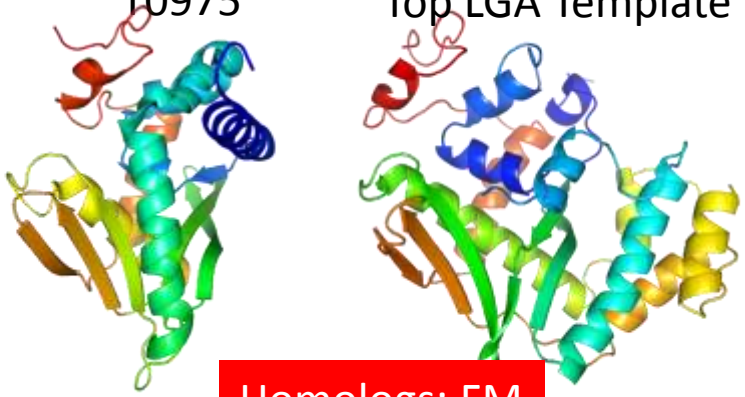
FM Boundary: High Sequence Score Examples 2

Conformation changes; movement of SSEs; insertions

T0950 Top LGA Template

Homologs: FM

T0975 Top LGA Template



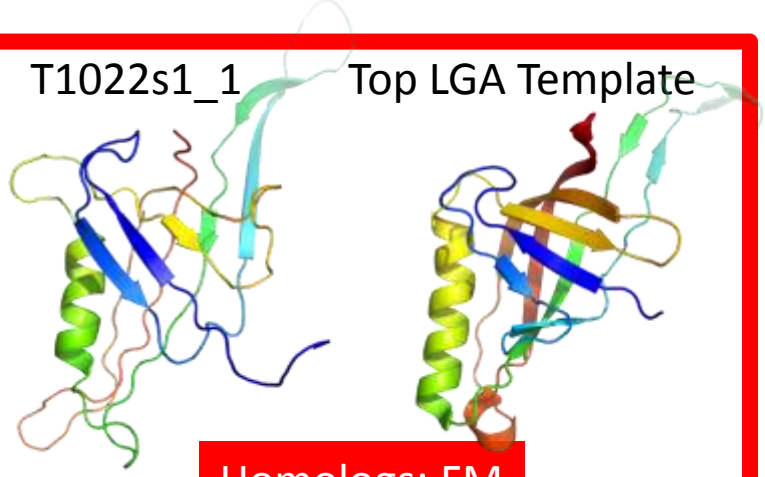
Homologs: FM

T0989_1 Top LGA Template

CASP:
Image redacted

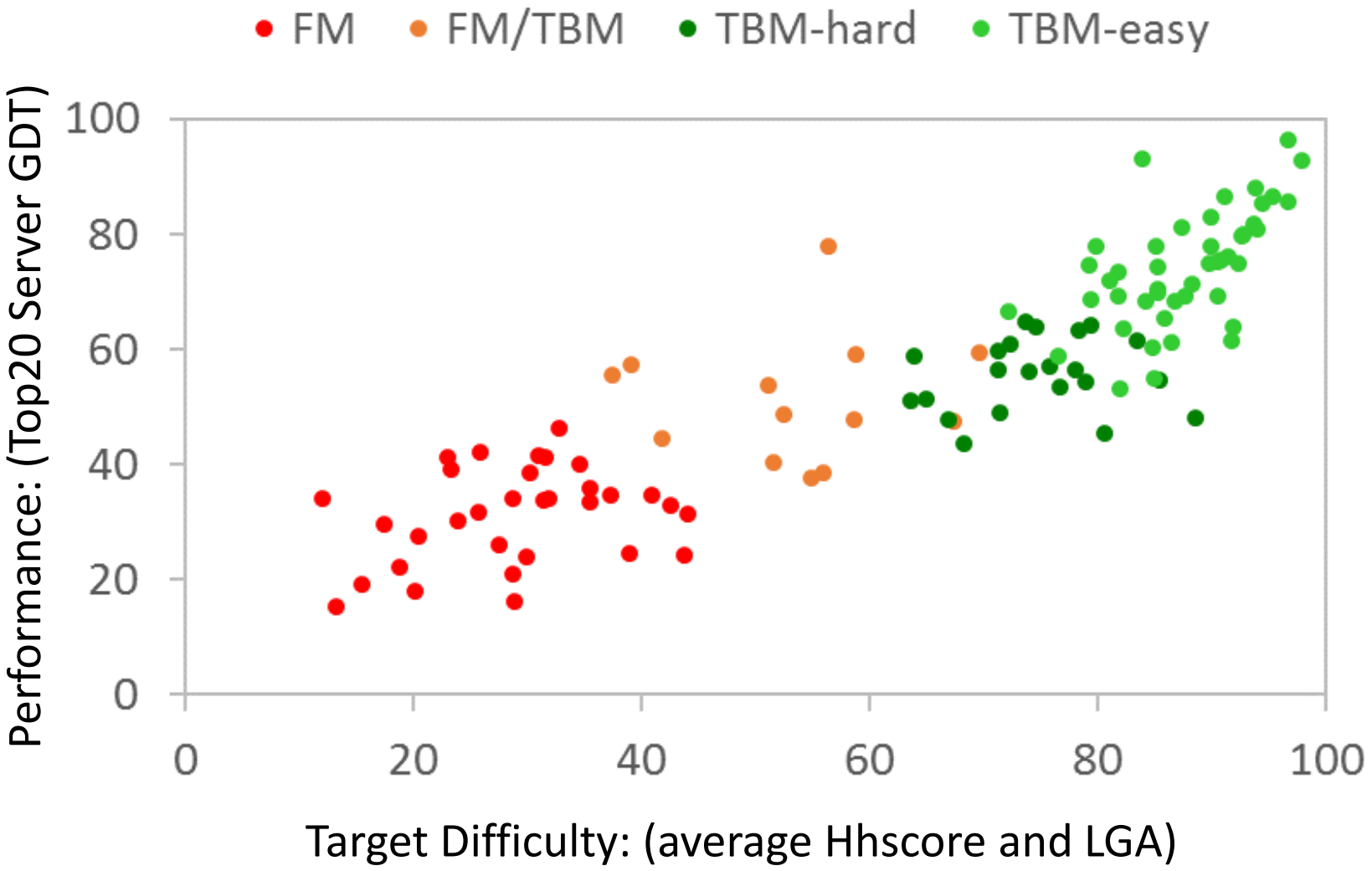
Homologs: FM

T1022s1_1 Top LGA Template

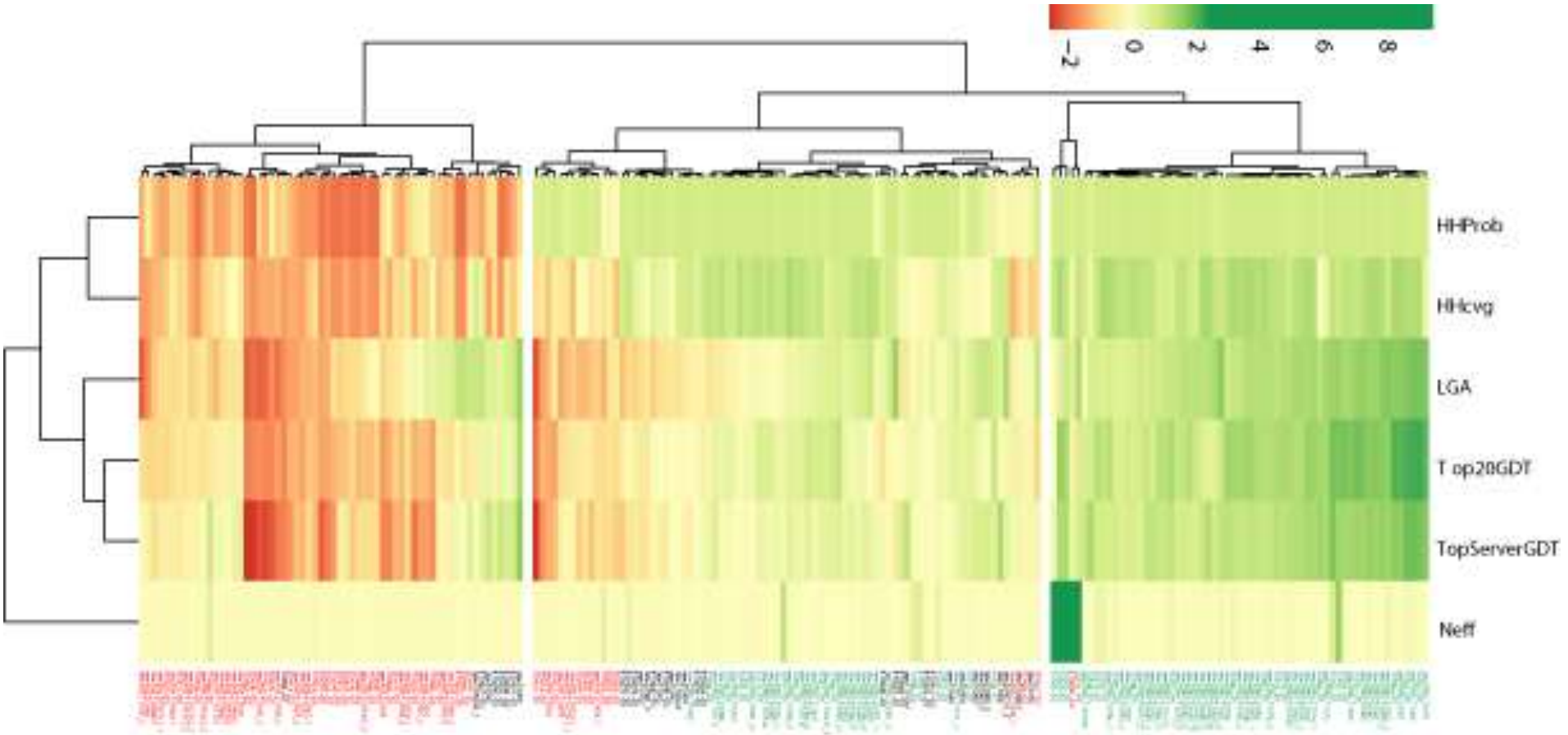


Homologs: FM

Assigned Boundary Cases Mapped To Plot 2



The END


-2
0
2
4
6
8
HHProb
HHcvg
LGA
Top20GDT
TopServerGDT
Neff