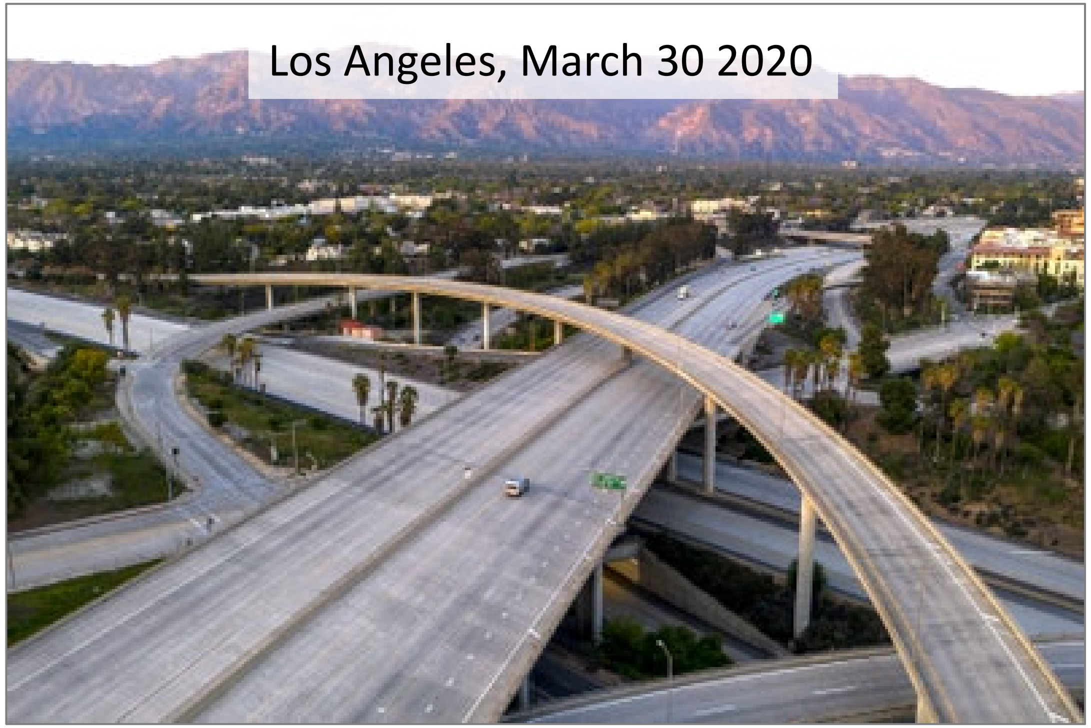
Target Classification in the 14th Round of the Critical Assessment of Protein Structure Prediction (CASP14)

Lisa Kinch, Andriy Kryshtafovych and Nick Grishin

CASP14 Target Classification during COVID-19

What used to be

Start of Registration

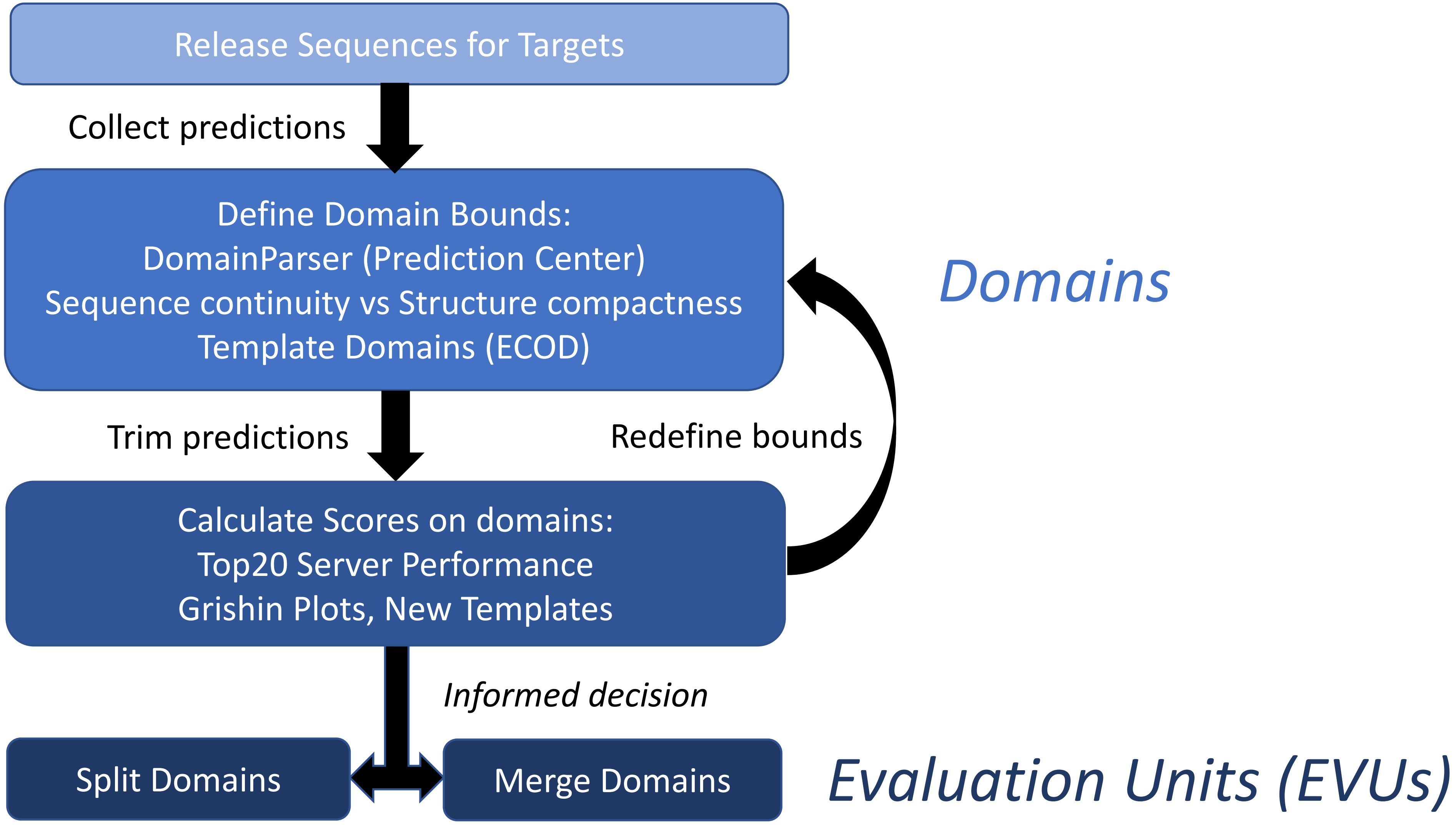


Registration for CASP14 opened,
March 9, 2020

Can we get enough targets?

CASP14 Domain Definition and Evaluation Units

What used to be



CASP14 *Pre-Evaluation* Domain Definition

What used to be

Release Sequences for Targets

Collect predictions

Define Domain Bounds:
DomainParser (Prediction Center)
Sequence continuity vs Structure compactness
Template Domains (ECOD)

Trim predictions

Redefine bounds

Calculate Scores on domains:
Top20 Server Performance
Grishin Plots, New Templates

Trimmed predictions

Original predictions

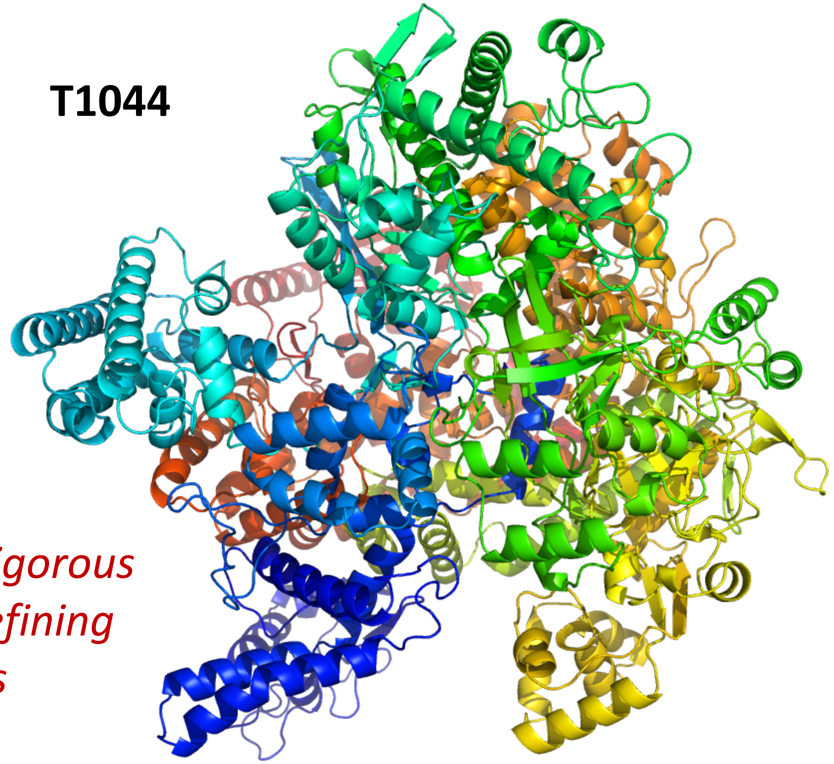
Split into EVUs

Merge Domains

CASP14 adaptation

T1044 Define Domain Bounds:
Suggestion from Experimentalist
Sequence continuity vs Structure compactness
Template Information

T1044

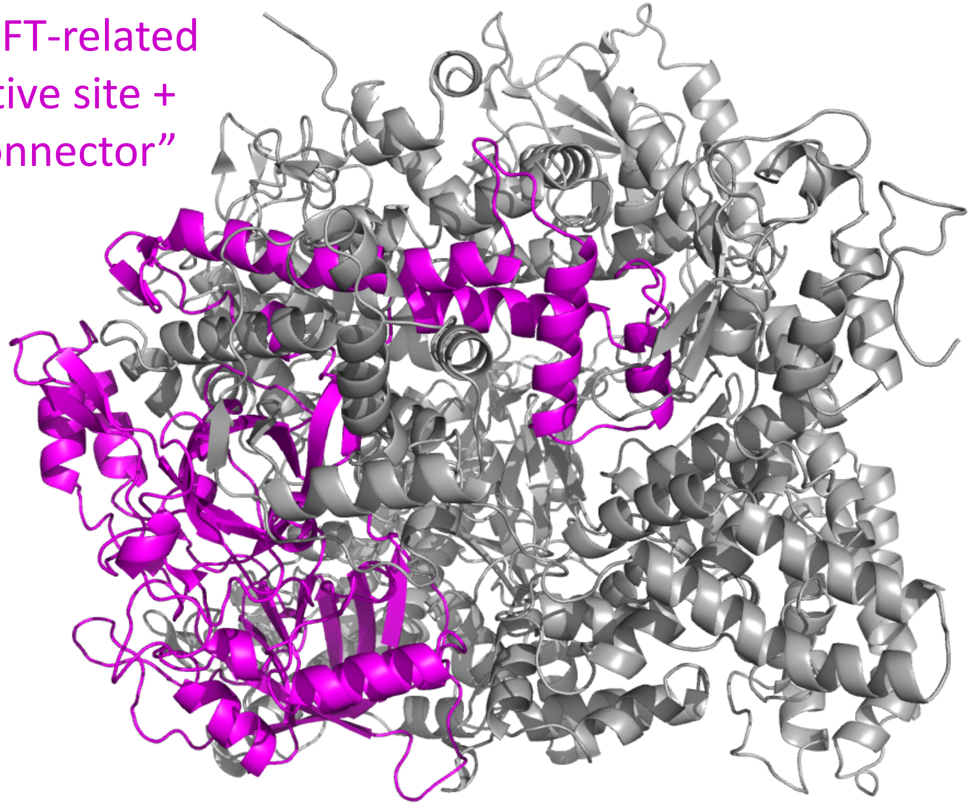


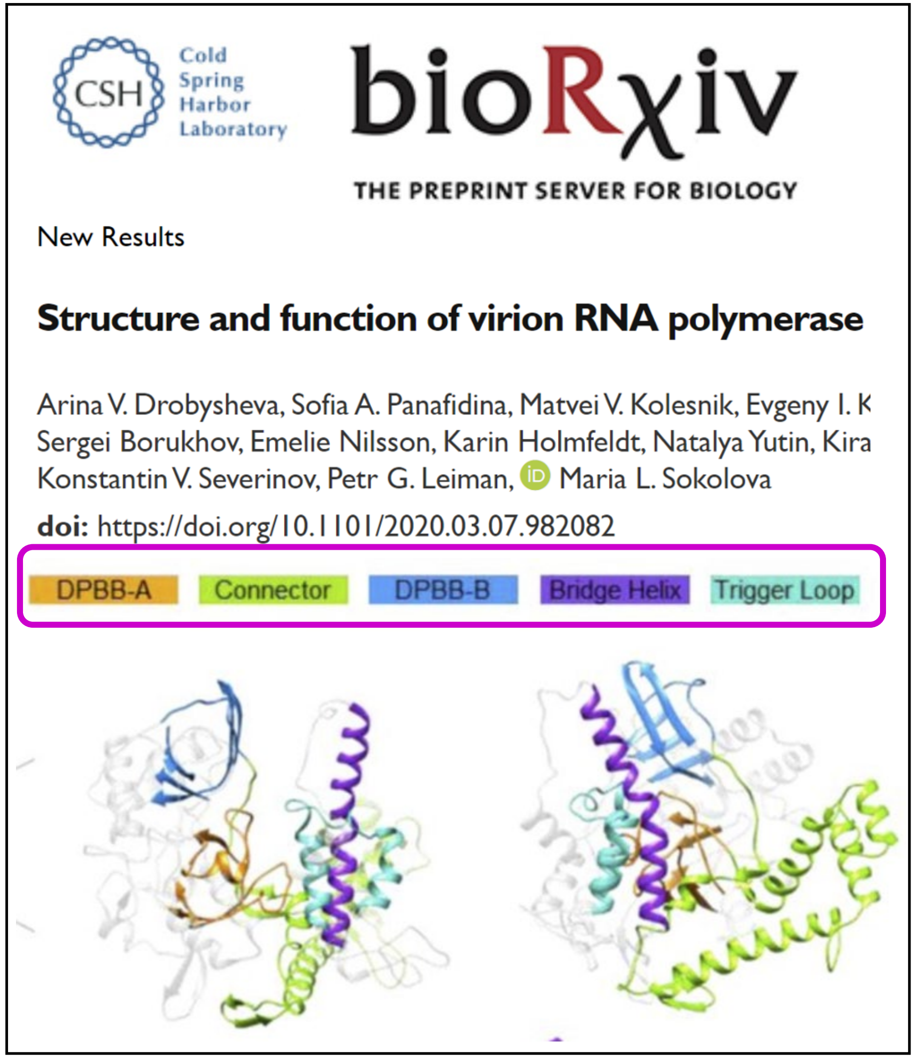
Bypasses the rigorous process for defining domains

CASP14 *Pre-Evaluation* Leak of Information

T1044: Phage DNA-dependent RNA polymerase

2xRIFT-related
Active site +
“connector”





New Results

Structure and function of virion RNA polymerase

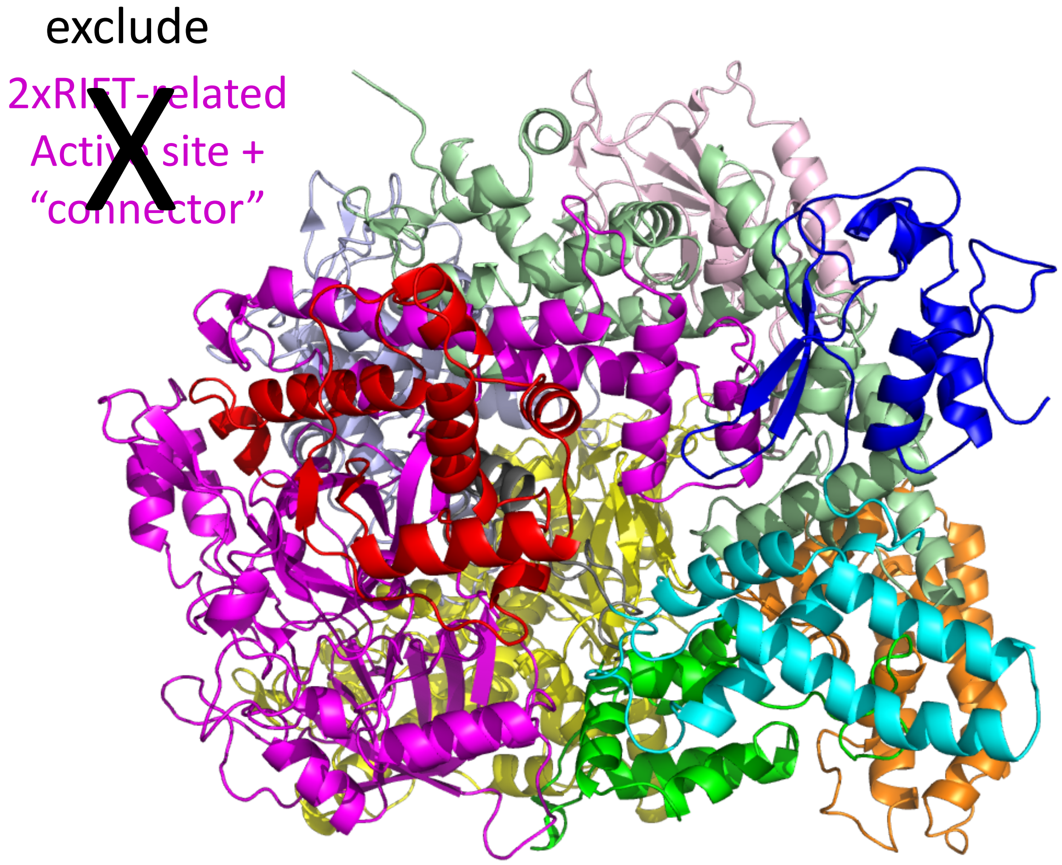
Arina V. Drobysheva, Sofia A. Panafidina, Matvei V. Kolesnik, Evgeny I. K
Sergei Borukhov, Emelie Nilsson, Karin Holmfeldt, Natalya Yutin, Kira
Konstantin V. Severinov, Petr G. Leiman, Maria L. Sokolova

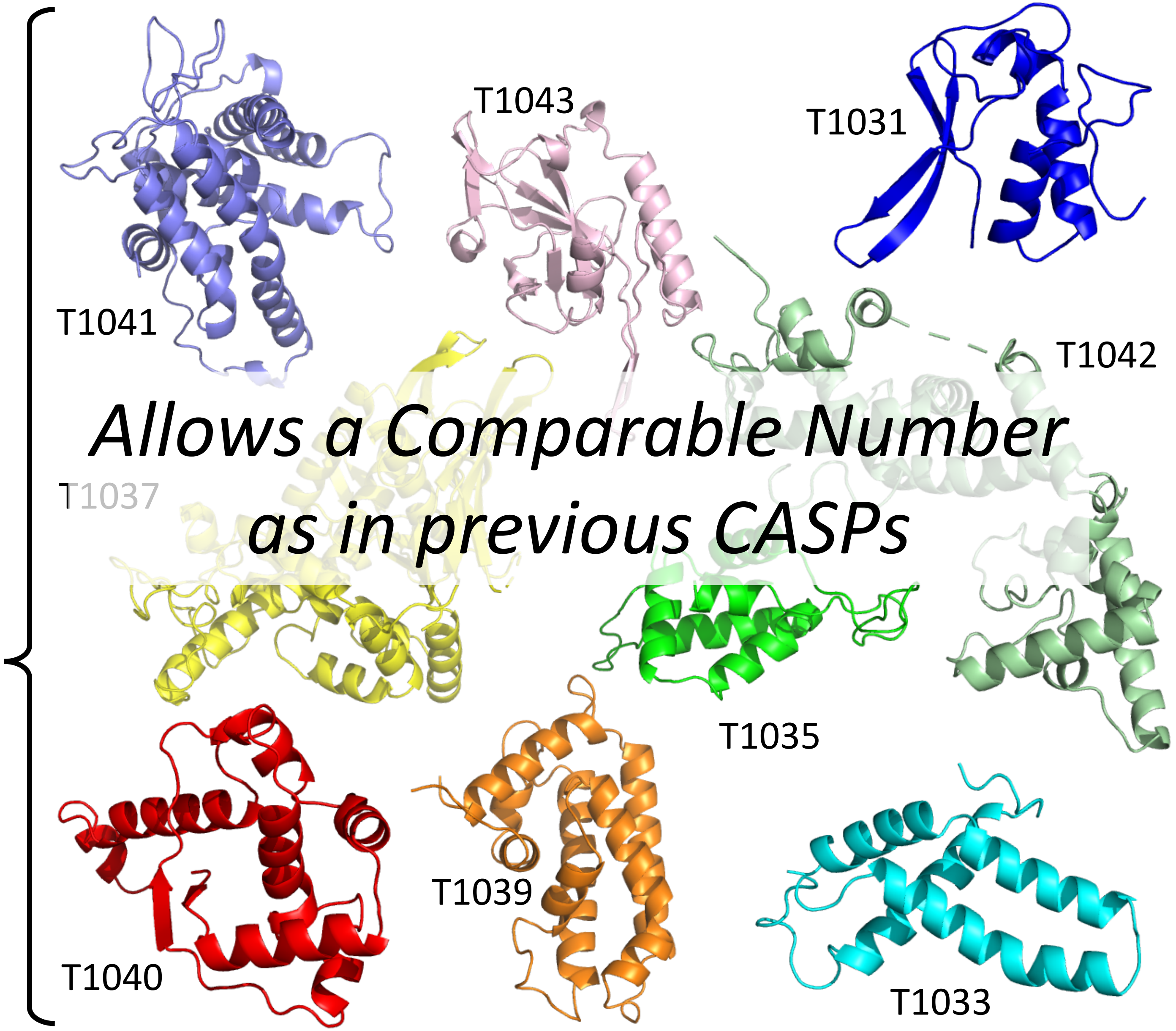
doi: https://doi.org/10.1101/2020.03.07.982082

Preprint Statistics (Andriy): CASP boosts Interest 5-fold in 1 week

T1044 *Pre-Evaluation* split into 9 Targets

T1044: Phage DNA-dependent RNA polymerase



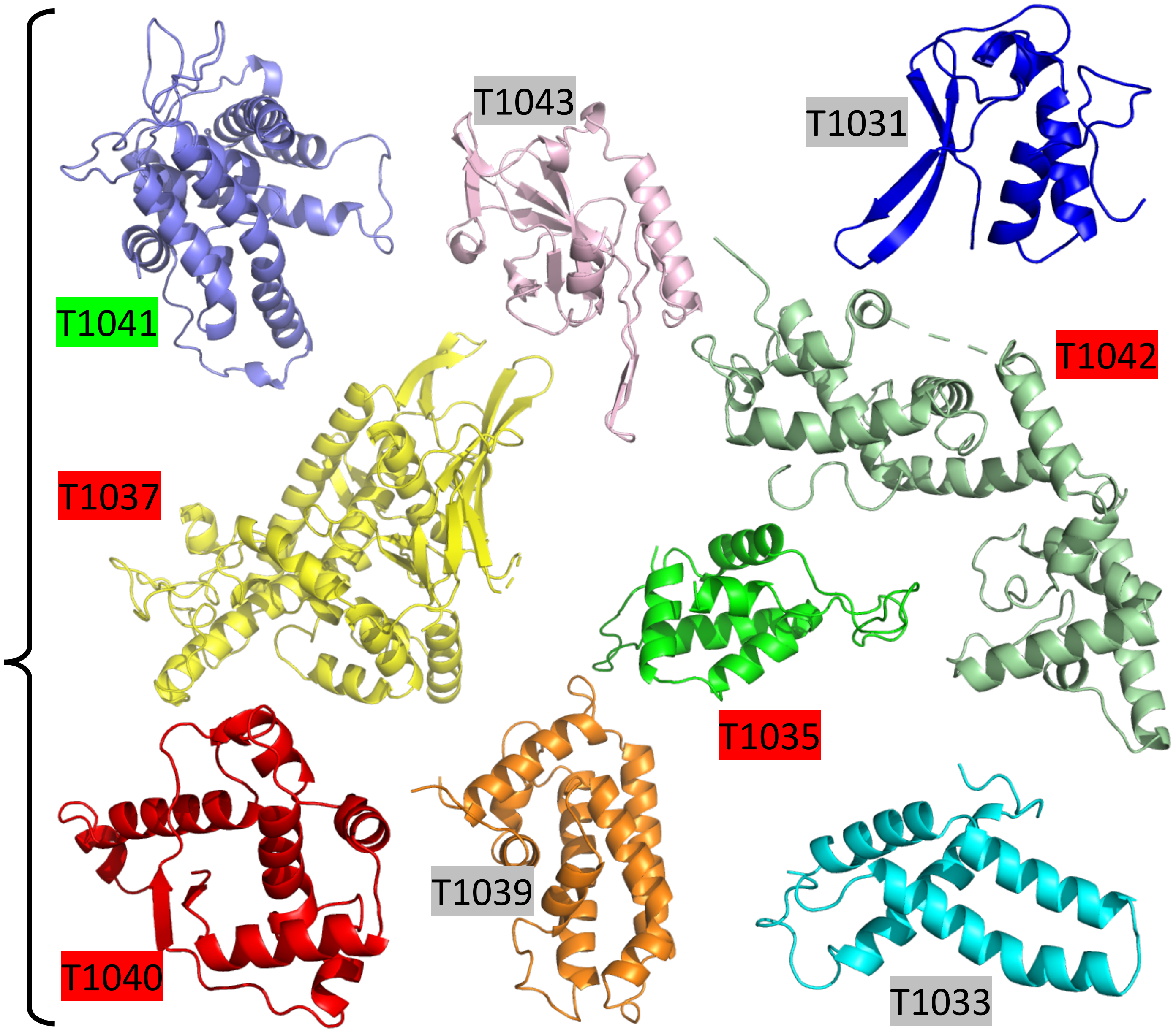


Template Information was Lacking

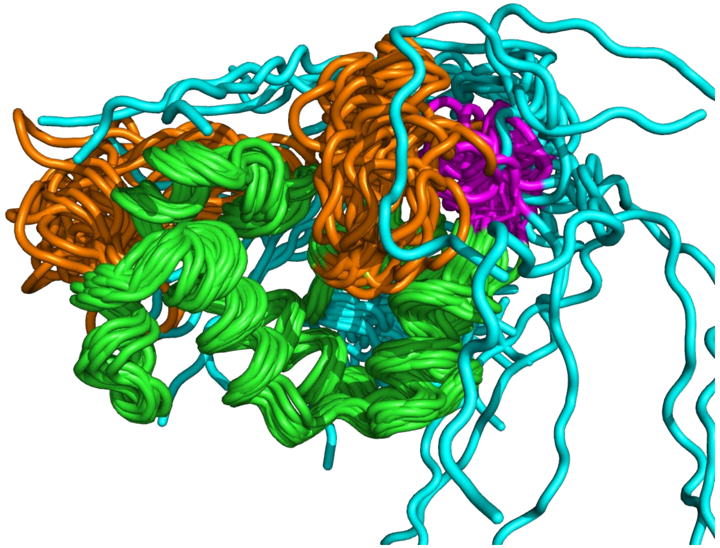
T1044: Phage DNA-dependent RNA polymerase

Evolutionary Relationships
New Fold (4)
Topology-level (4)
Distant Homolog (1)

Lack of Templates and extensive domain interactions mean Domains might not be independent folding units

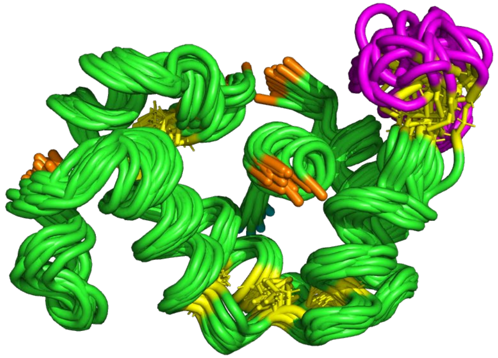


Technical Considerations for Evaluation Units

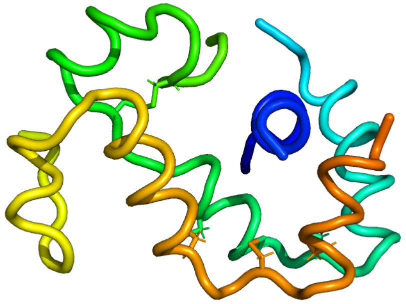


T1027: Gaussia luciferase

- NMR structure with high flexibility
- Loose ensemble
- 5 disulfide bond pairs



- Keep overlapping parts of the structure
- Trim last disulfide pair

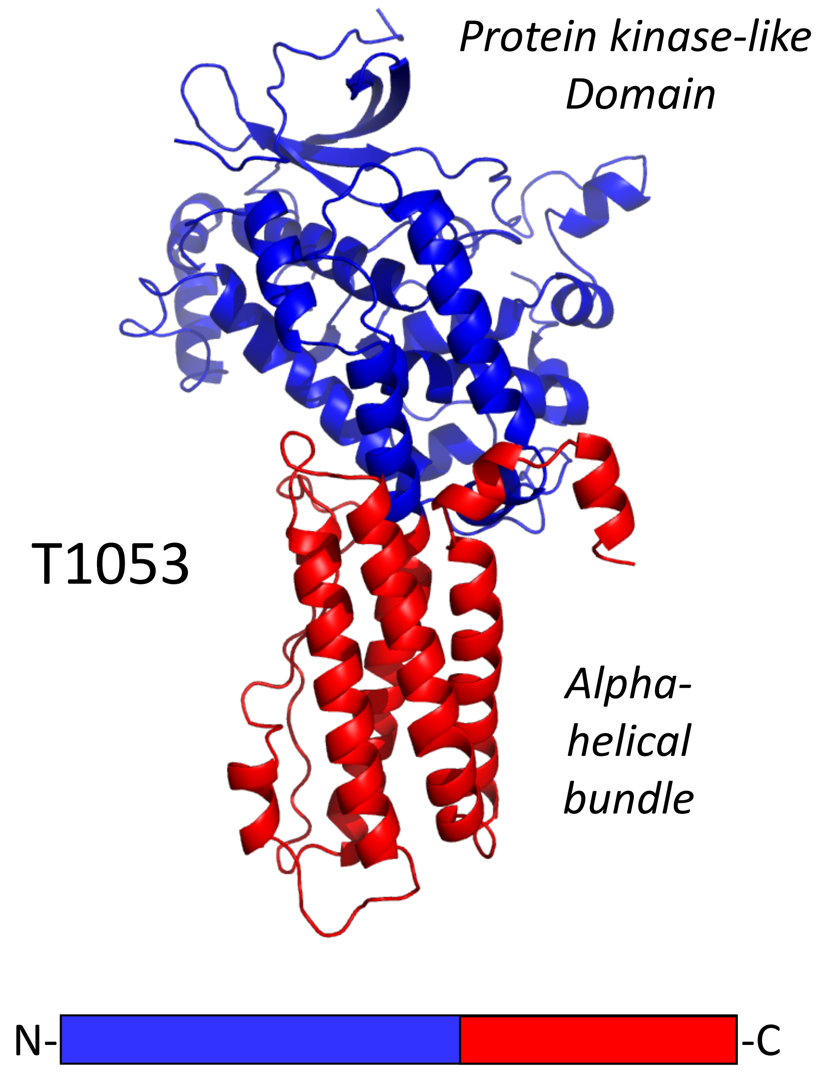


Keep trimmed model1 as the T1027 EVU

Domains Have Many Different Definitions

What is a Domain?

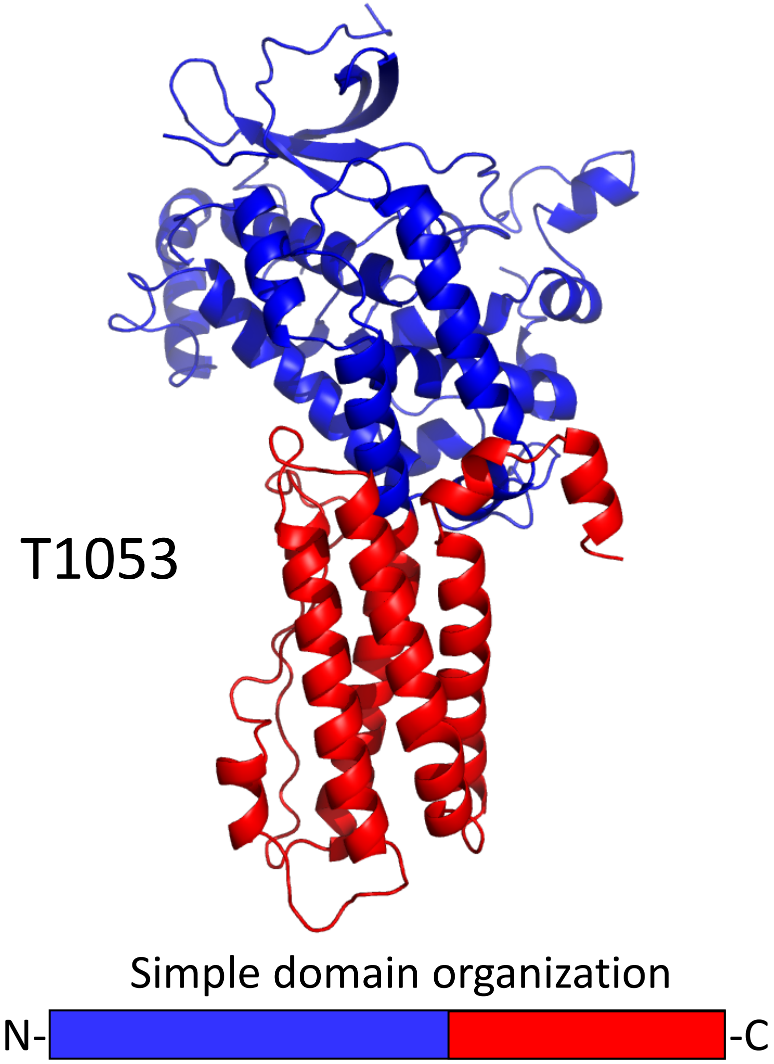
- Compact, globular substructures that have more interactions within them than with the rest of the structure



Domains are More than Compact Substructures

What is a Domain?

- Compact, globular substructures that have more interactions within them than with the rest of the structure
- Conserved, Independent folding unit that can exist in multiple contexts, i.e. serve as building blocks of evolution
- Evolution tends to preserve sequence continuity in domains



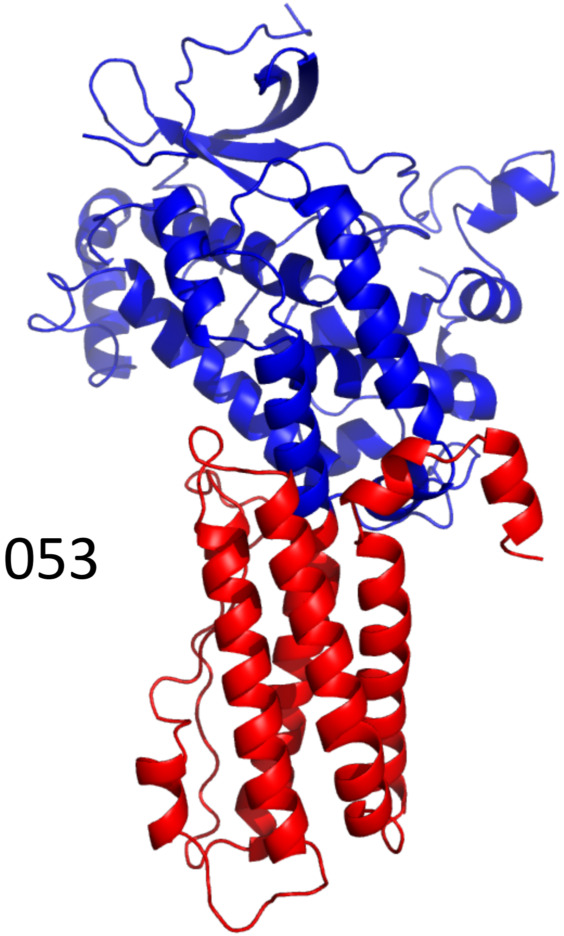
ECOD Database as a Resource for Definition

What is a Domain?

- Compact, globular substructures that have more interactions within them than with the rest of the structure
- Conserved, Independent folding unit that can exist in multiple contexts i.e. serve as building blocks of evolution
- Evolution tends to preserve sequence continuity in domains
- Evolutionary Classification of Protein Domains (ECOD) database was *an essential resource* for defining domains: prodata.swmed.edu/ecod/ (thanks Dustin!)

ECOD PMID: 25474468

T1053



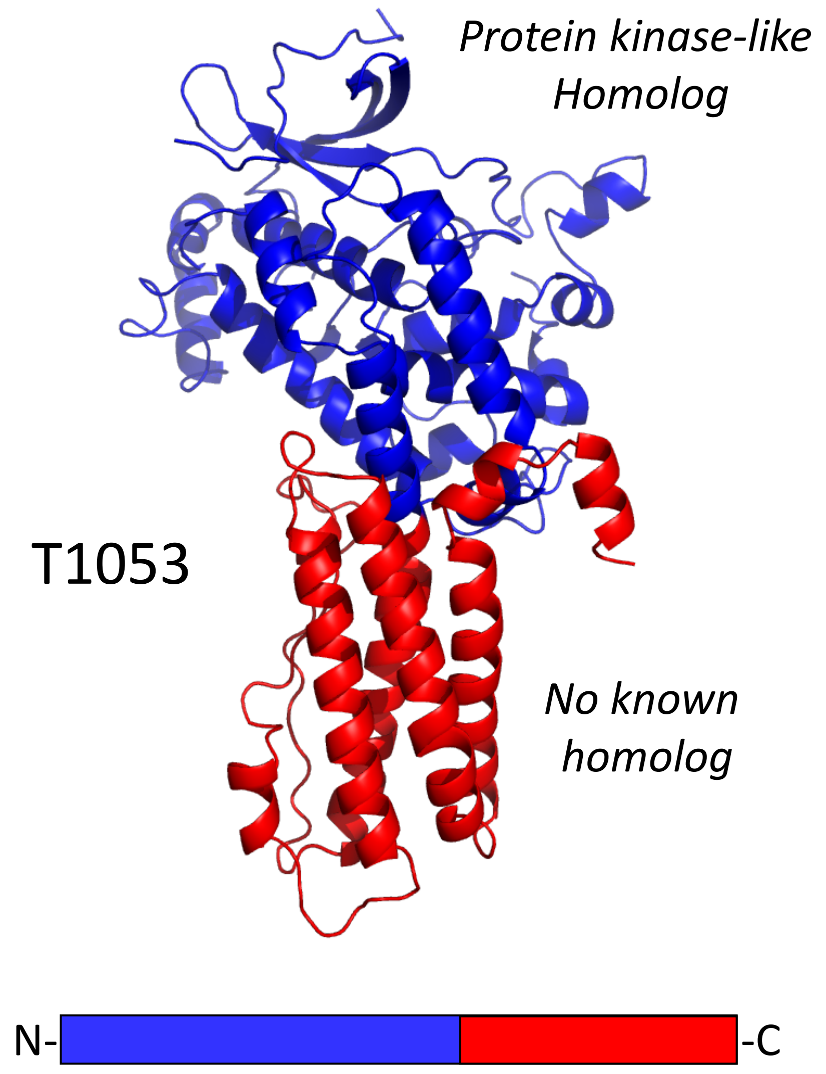
Simple domain organization

Turning Domains into Evaluation Units

Domains = Evaluation Units

- Using split domains as EVUs are required when templates have known conformation changes (example to follow)
- Using split domains as EVUs are required when they have different difficulty levels (perhaps not in the future)

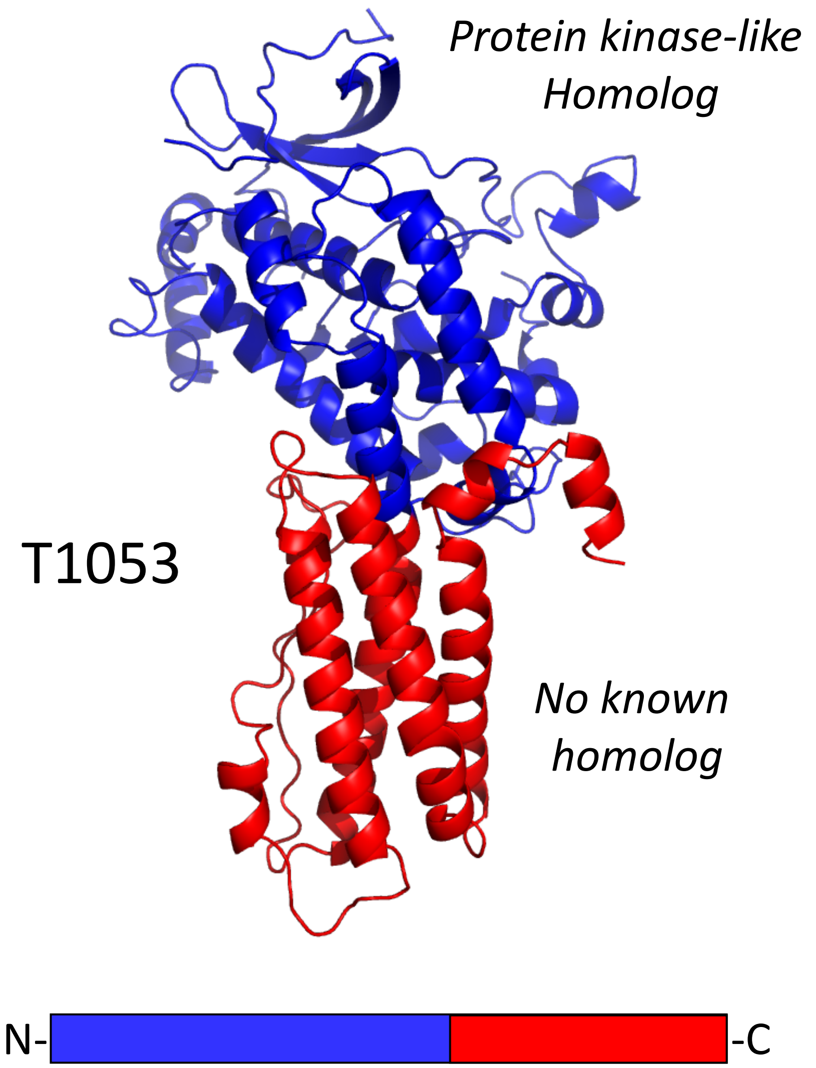
For CASP14 we tried to ***keep domains together****;*
If not, we evaluated domain interactions in a separate assessment



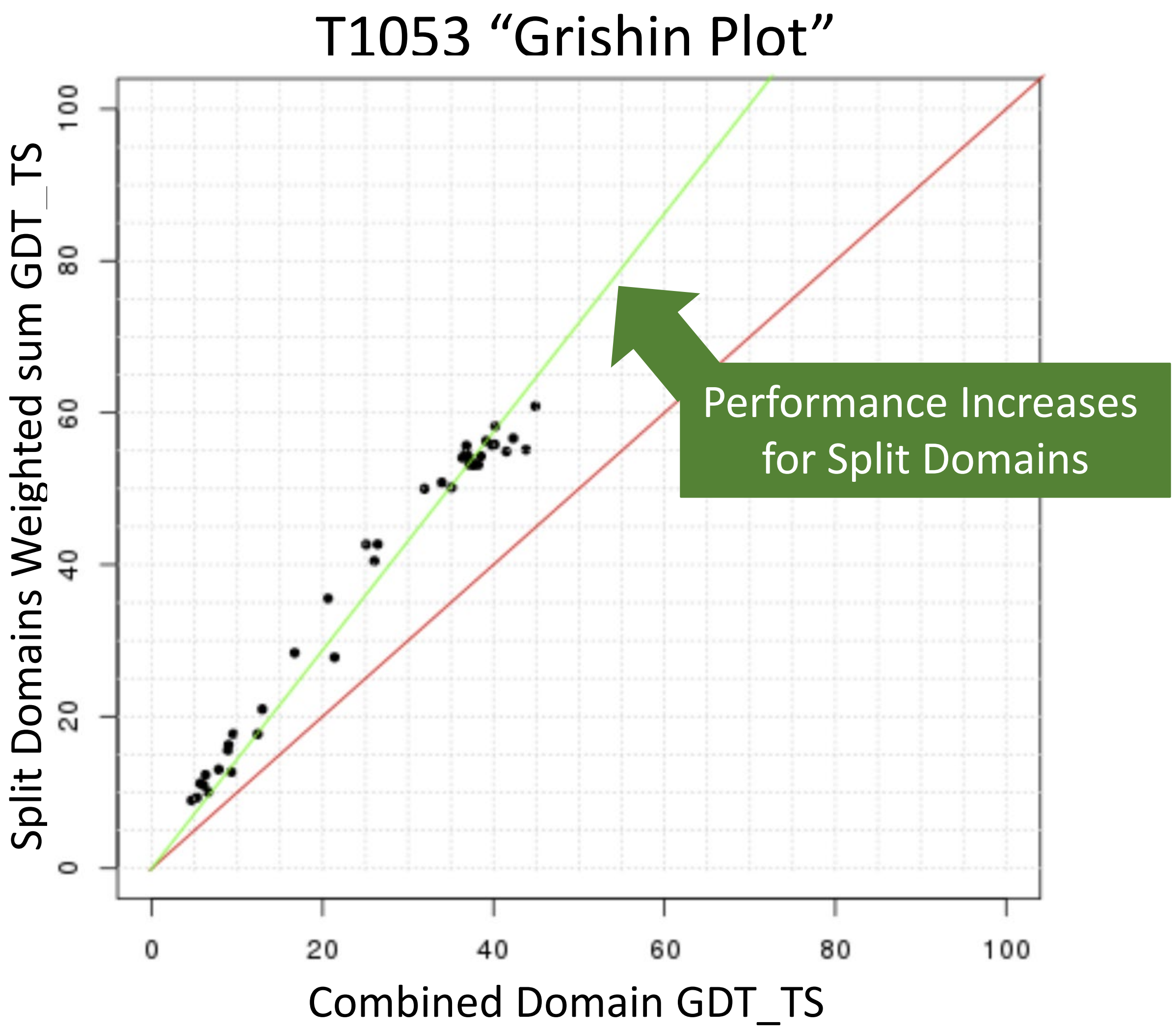
Turning Domains into Evaluation Units

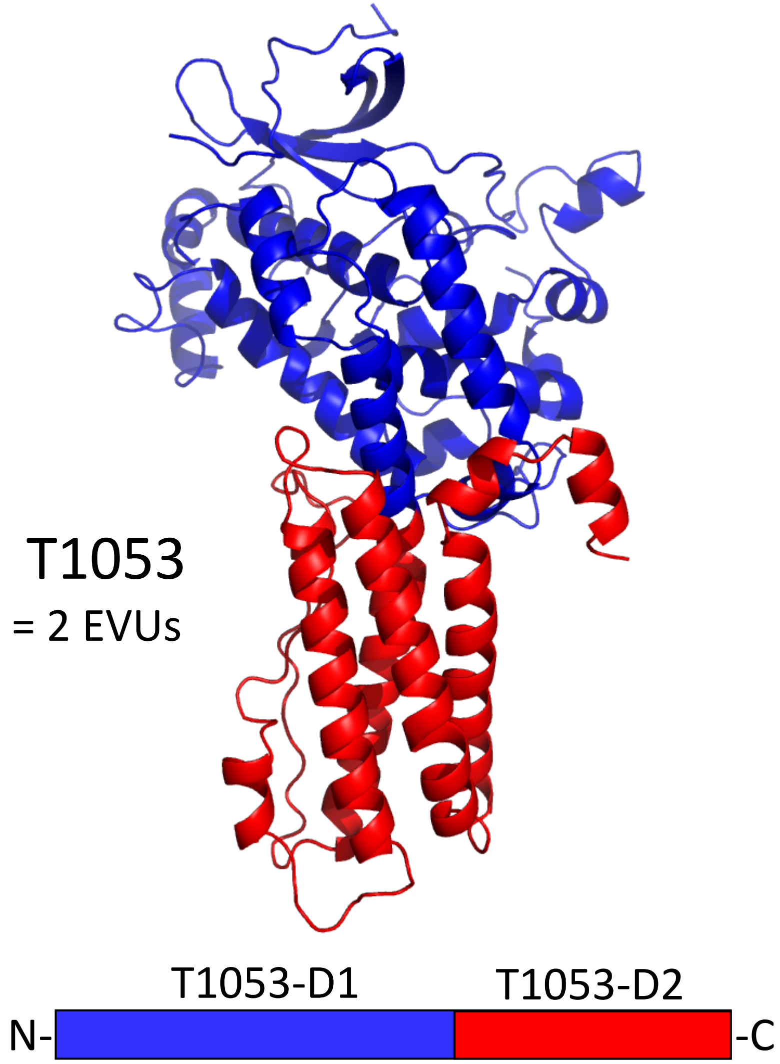
Domains = Evaluation Units

- Using split domains as EVUs are required when templates have known conformation changes (example to follow)
- Using split domains as EVUs are required when they have different difficulty levels
- Decisions to split or merge are based on group performance: traditionally evaluated using “Grishin Plots”



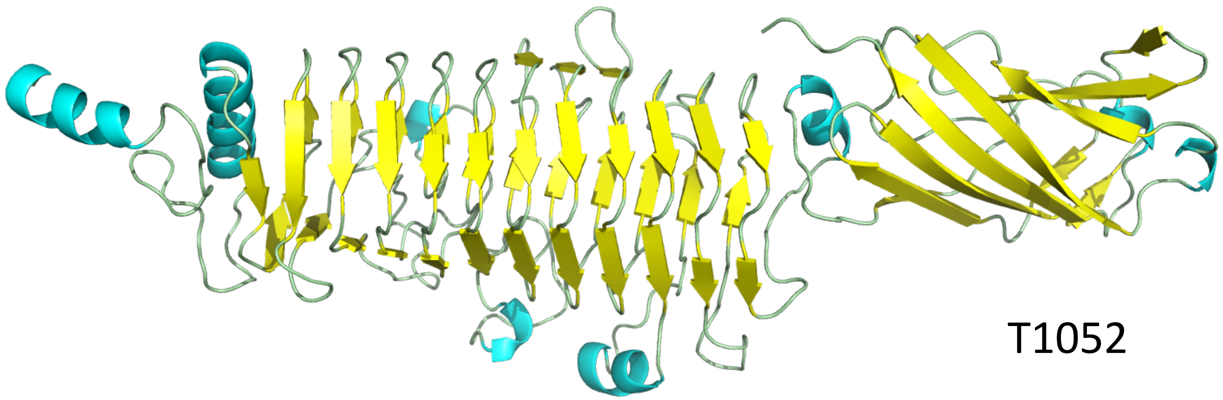
Grishin Plots Inform Decisions to Split Targets



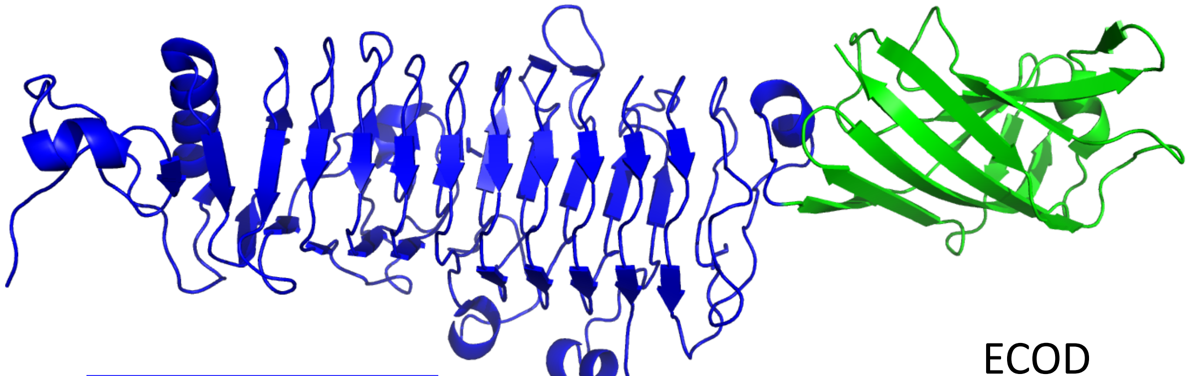


Merging Target Domains as Evaluation Units

No Need to Split when Good Templates Exist



T1052



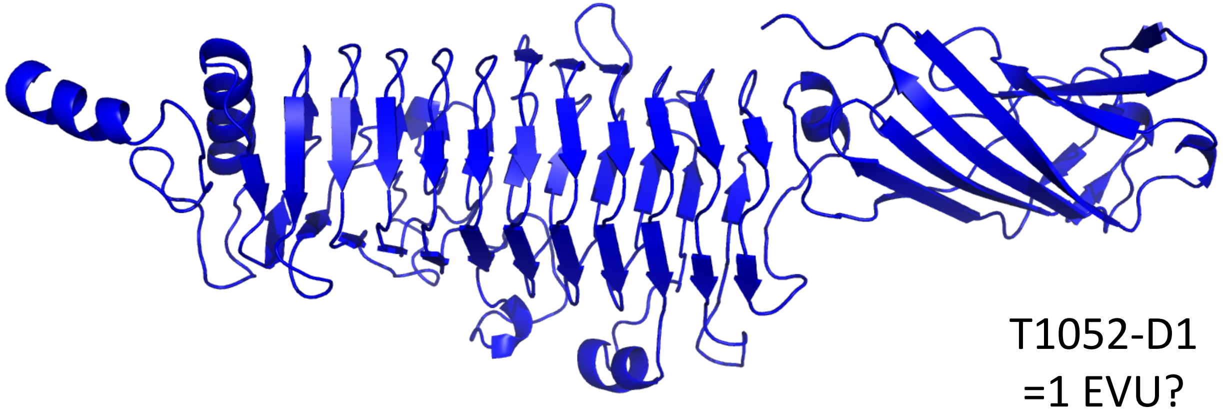
Domain in virus attachment proteins

Pectin lyase-like

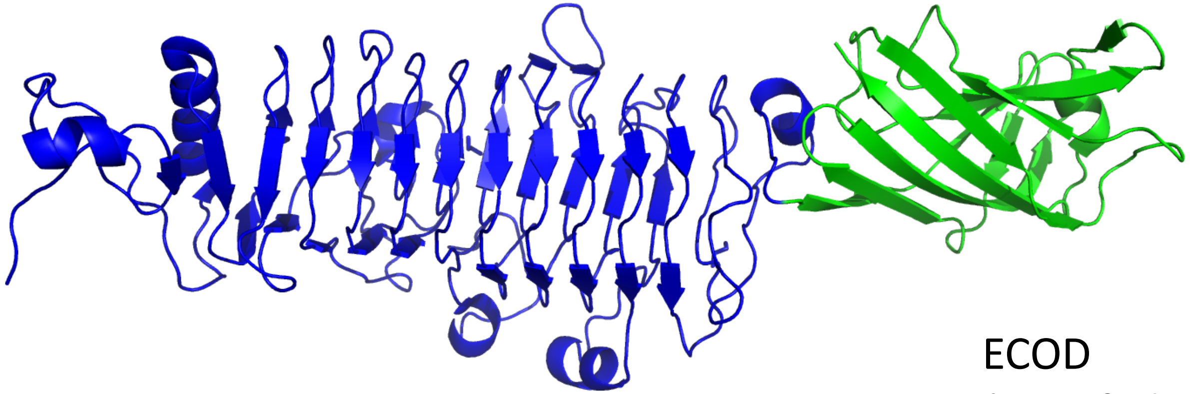
ECOD template 6f7dA = 2 domains

Merging Target Domains as Evaluation Units

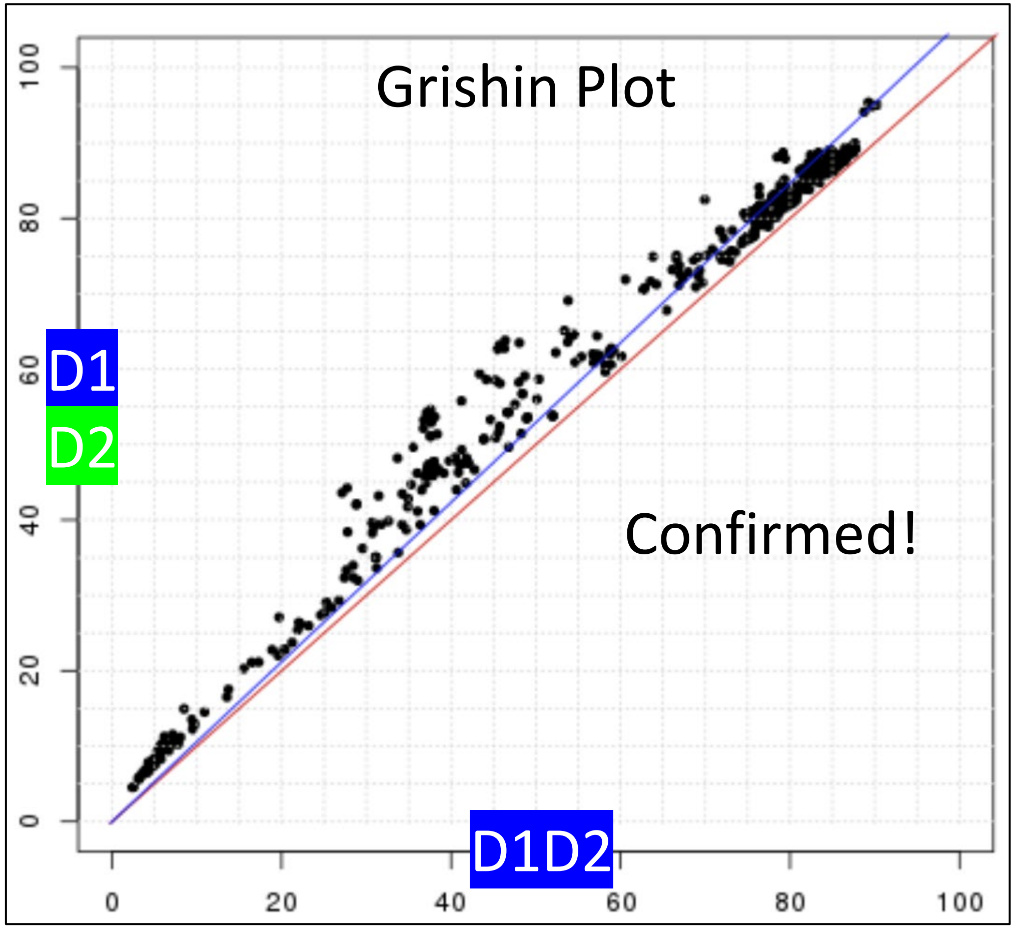
No Need to Split when Good Templates Exist



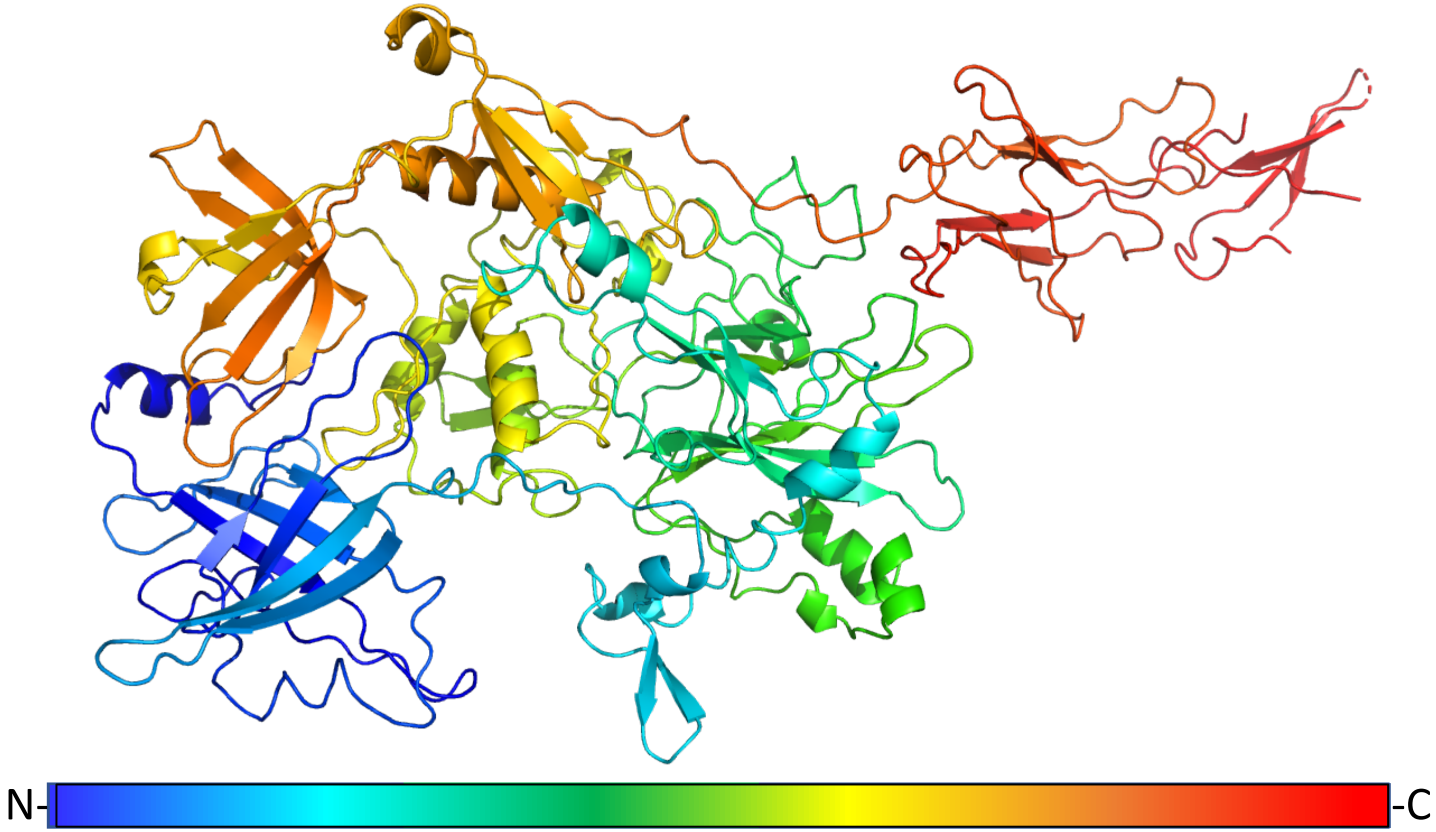
T1052-D1
=1 EVU?



ECOD
template 6f7dA
= 2 domains



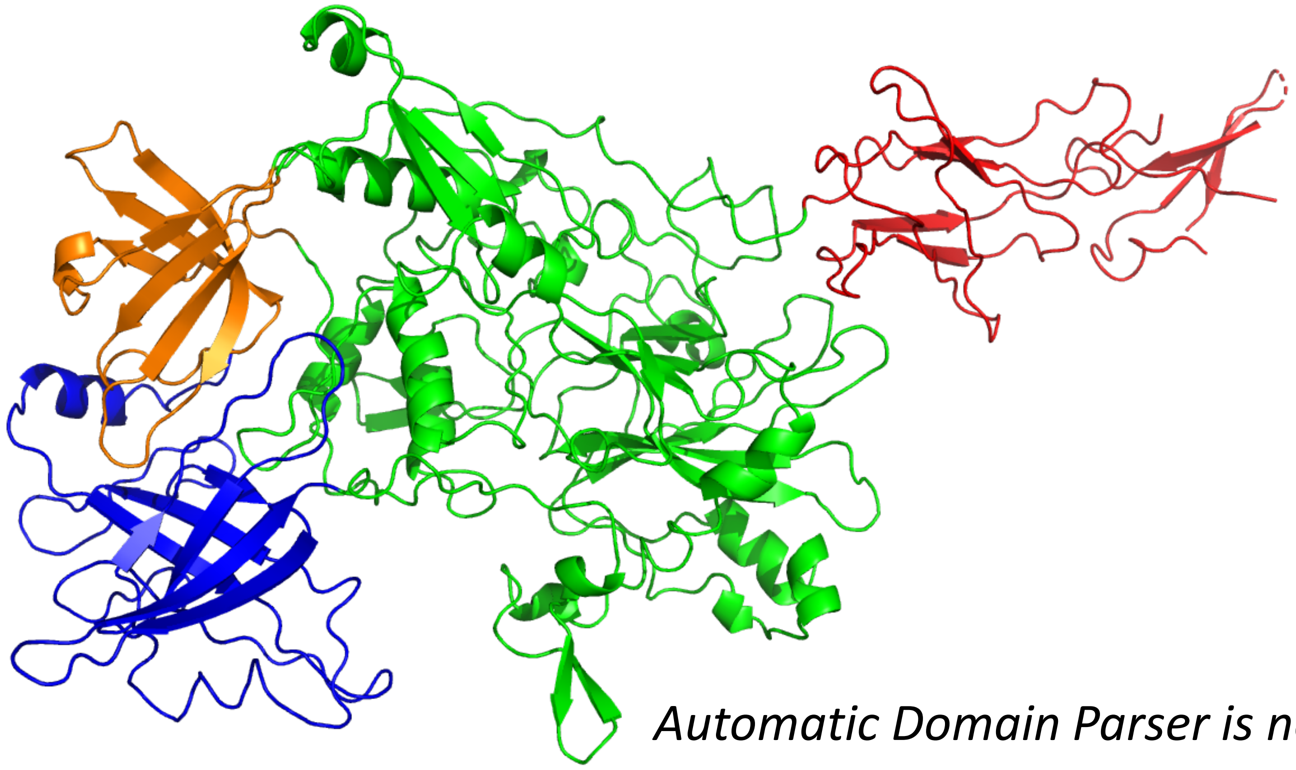
Some Domain Definitions are Difficult



T1061: *E.coli* phage tail

- Complex domain organization

Some Domain Definitions are Difficult

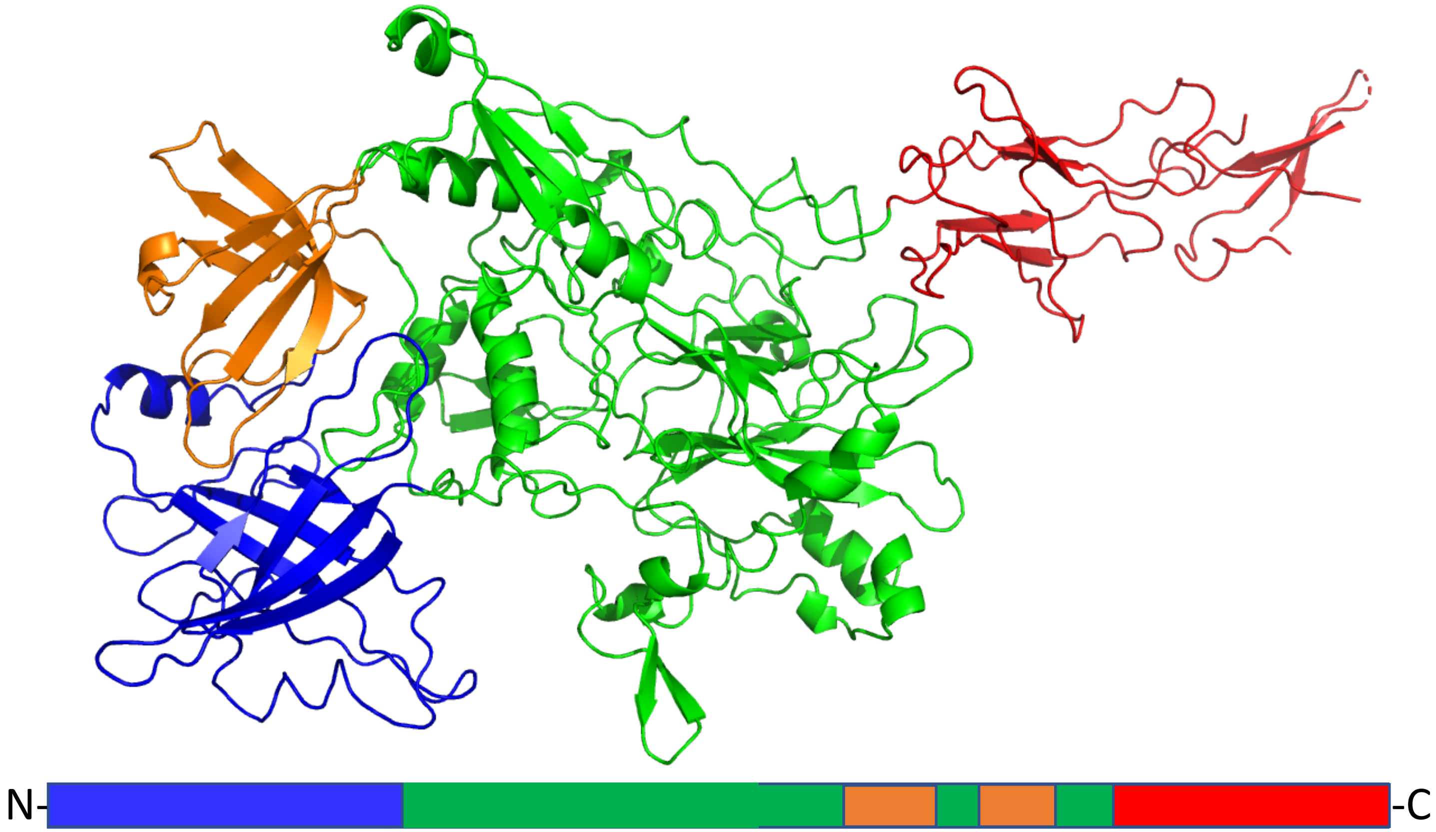


Automatic Domain Parser is non continuous

T1061: *E.coli* phage tail

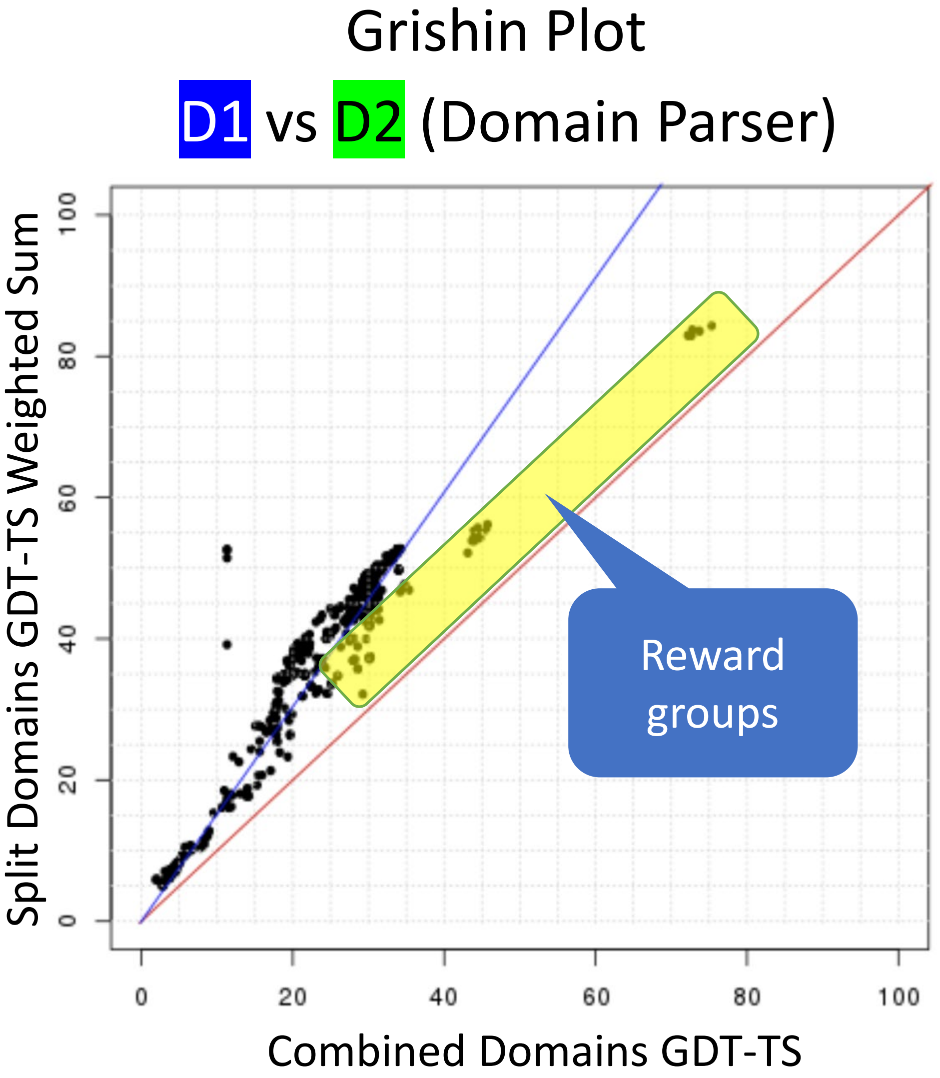
- Complex domain organization
- Domain parser and Ddomain split differently (4 vs 5)

Some Domain Definitions are Difficult

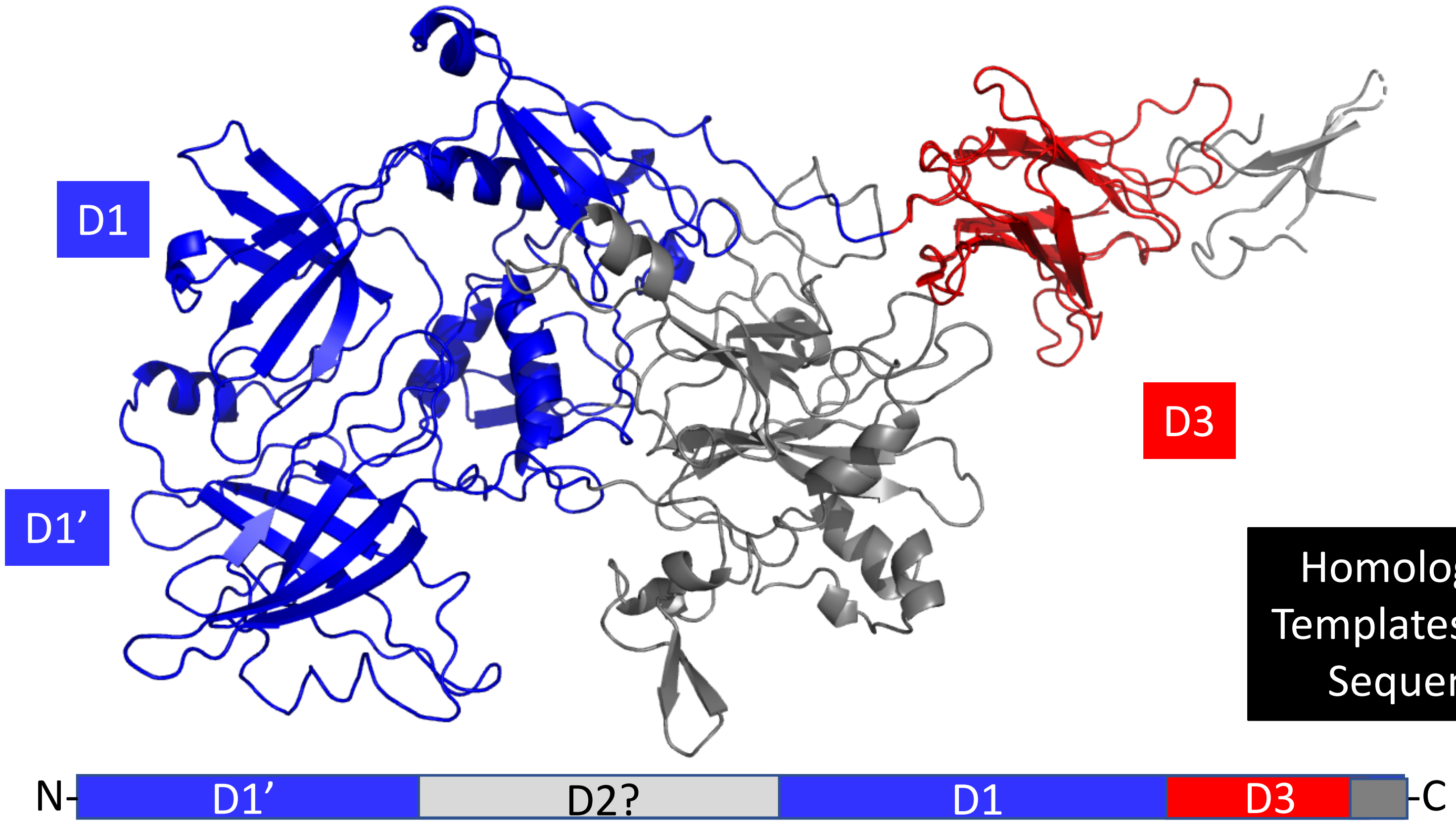


T1061: *E.coli* phage tail

- Complex domain organization
- Domain parser and Ddomain split differently (4 vs 5)
- Grishin Plot has multiple clouds



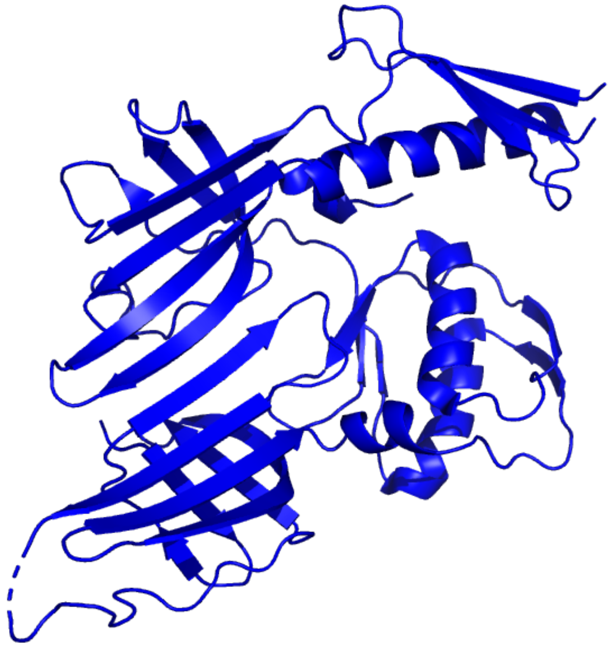
Homologous Templates Suggest Domain Bounds



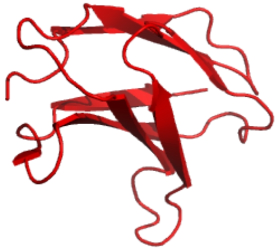
T1061: *E.coli* phage tail

- Complex domain organization
- Domain parser and Ddomain split differently (4 vs 5)
- Grishin Plots have multiple clouds
- Templates for blue and red domains

Homologous Templates from Sequence



D1’ D1 3cddF Template
4 domains: RIFT-related, N0 domain, insert, and RIFT-related *but 1EVU*

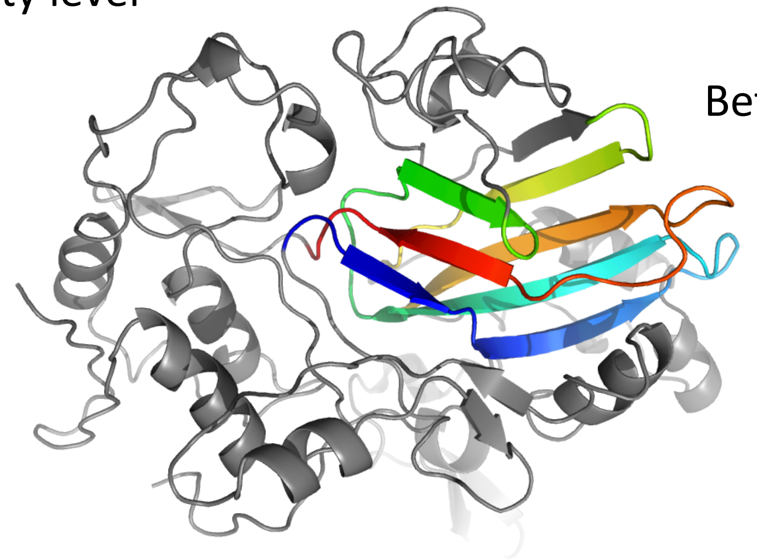


D3 1ten Top LGA_S Template
Immunoglobulin-related

Topology-level Insert is More Difficult: Suggests a Split

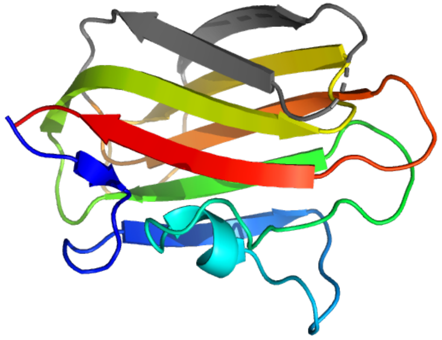
= different difficulty level
So 3 EVUs

D2

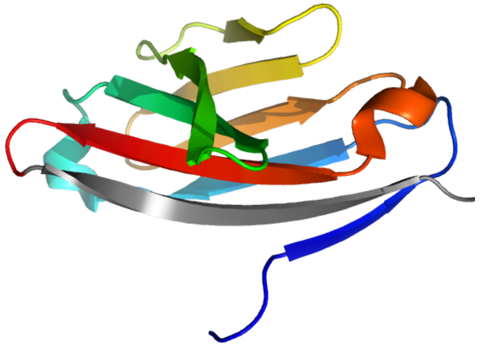


Beta-sandwich

Topology Templates From Structure



2yc2 Top LGA_S Template
Intraflagella Transport Protein 25
jelly-roll

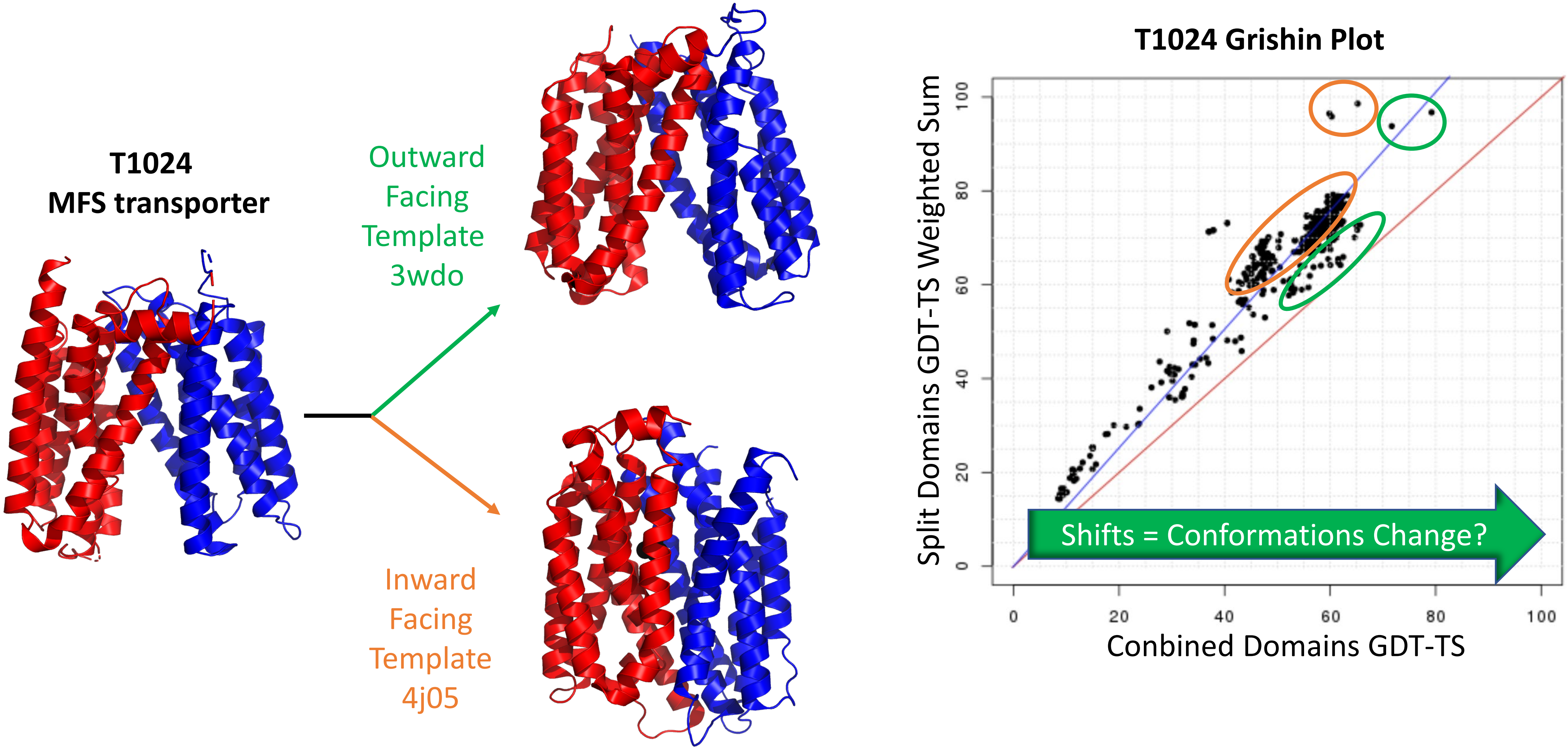


2frg Top Dali Template
human TLT1
Immunoglobulin-like β-sandwich

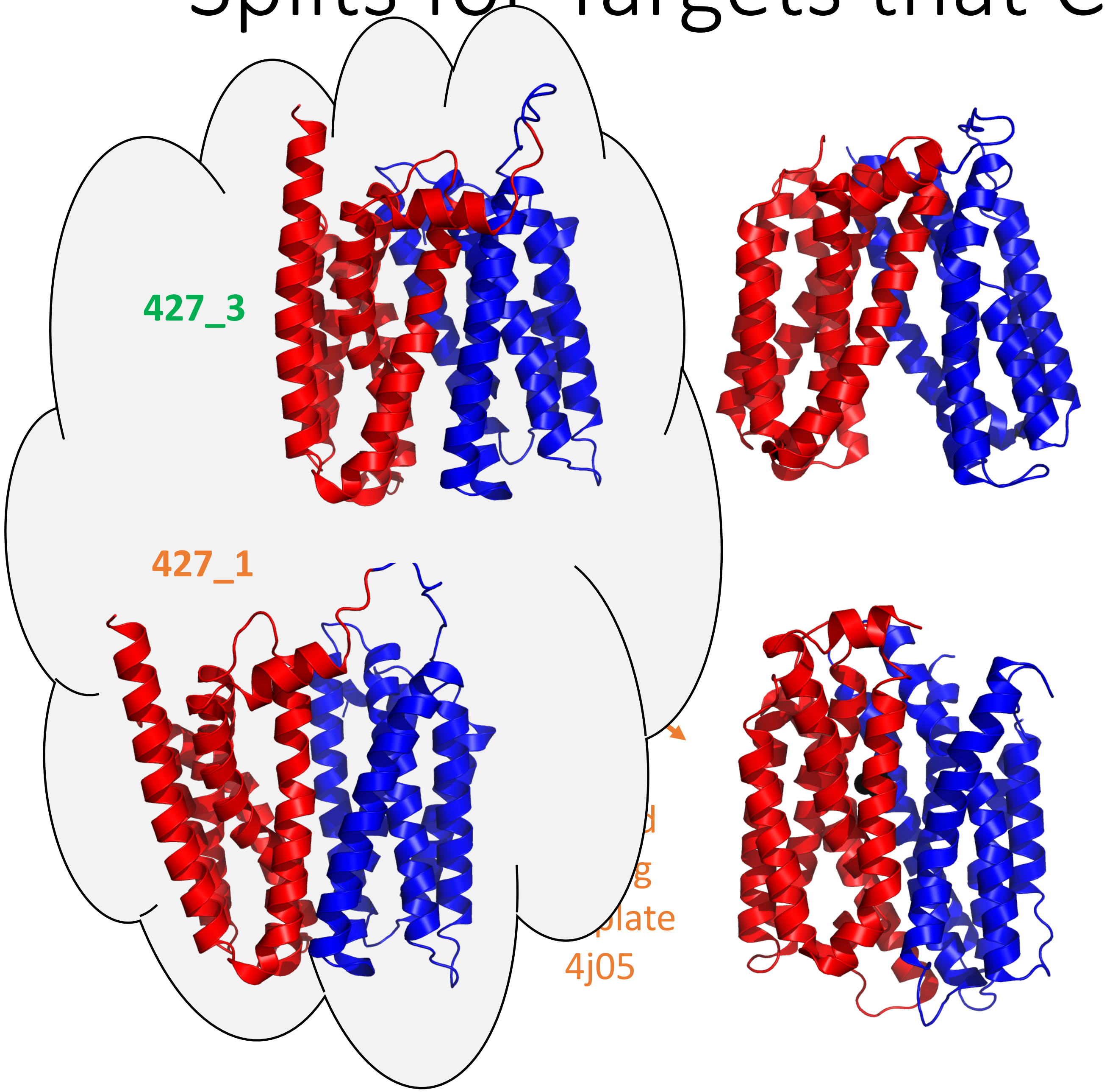
T1061: *E.coli* phage tail

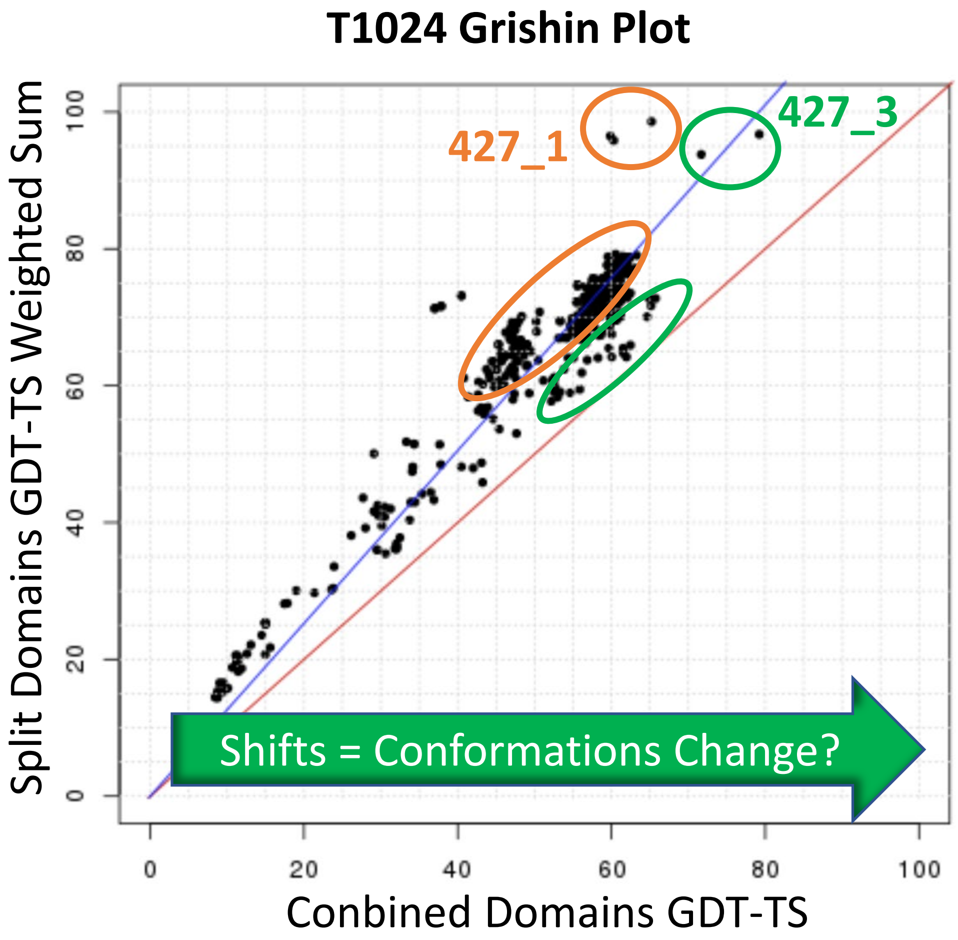
- Complex domain organization
- Domain parser and Ddomain split differently (4 vs 5)
- Grishin Plots have multiple clouds
- Templates for blue and red domains

Splits for Targets that Change Conformation



Splits for Targets that Change Conformation





4 Similar Targets: T1024, T1050, T1100, T1101

CASP14 Domains and EVUs in Numbers

67 Targets: 9 pre-split from 1 structure, 11 from 6 interacting structures

48 Single EVU Targets
6 have 18 merged ECOD domains

19 Split Domain Targets (48 EVUs)
4 have 10 merged ECOD domains

=96 Targets for classification into Topology-level (FM) and High Accuracy-level (TBM)

Evolutionary Relationships to known Templates help Classification

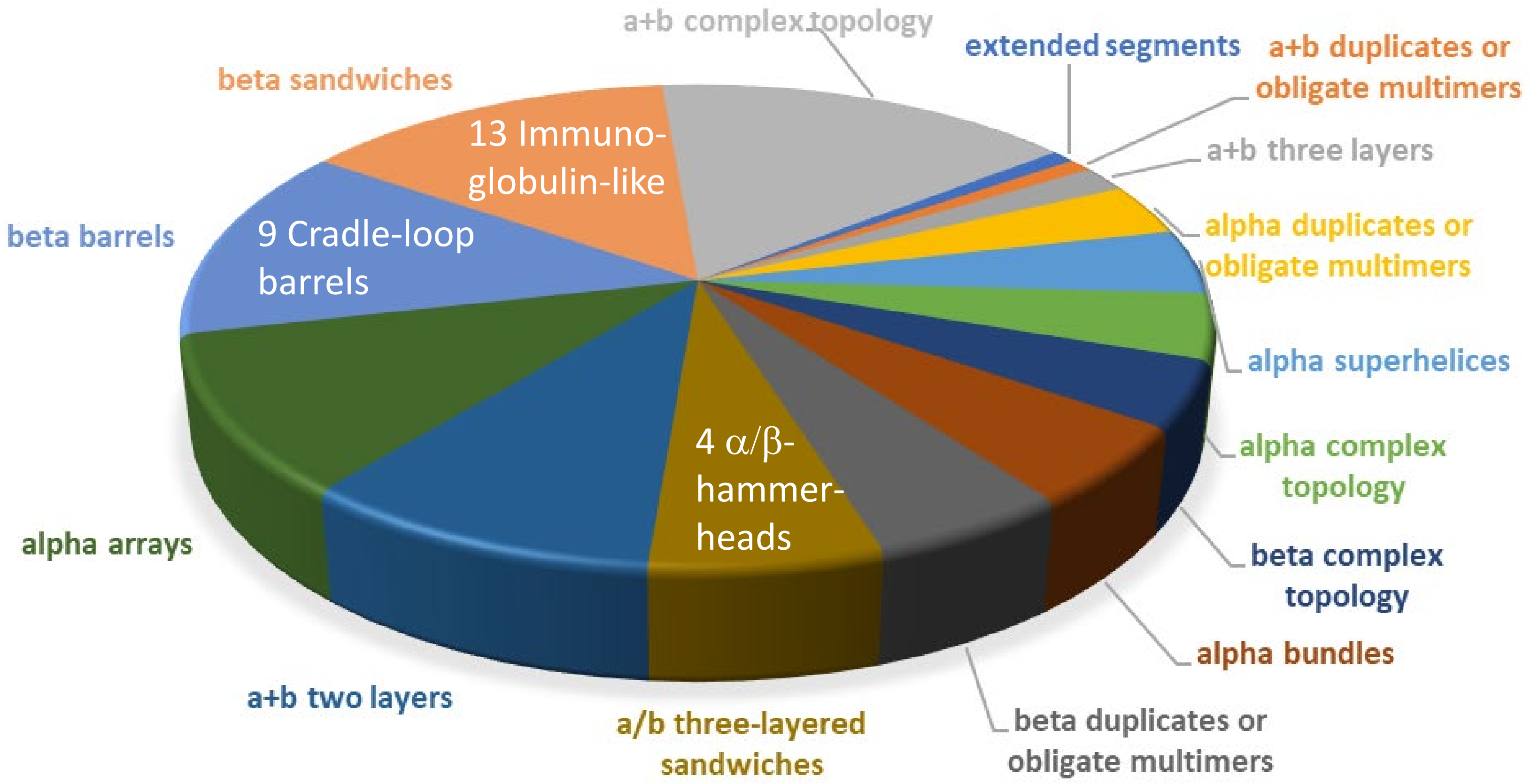
Evolution-Based Classification of CASP14 EVUs

67 Targets: 9 pre-split from 1 structure, 11 from 6 interacting structures

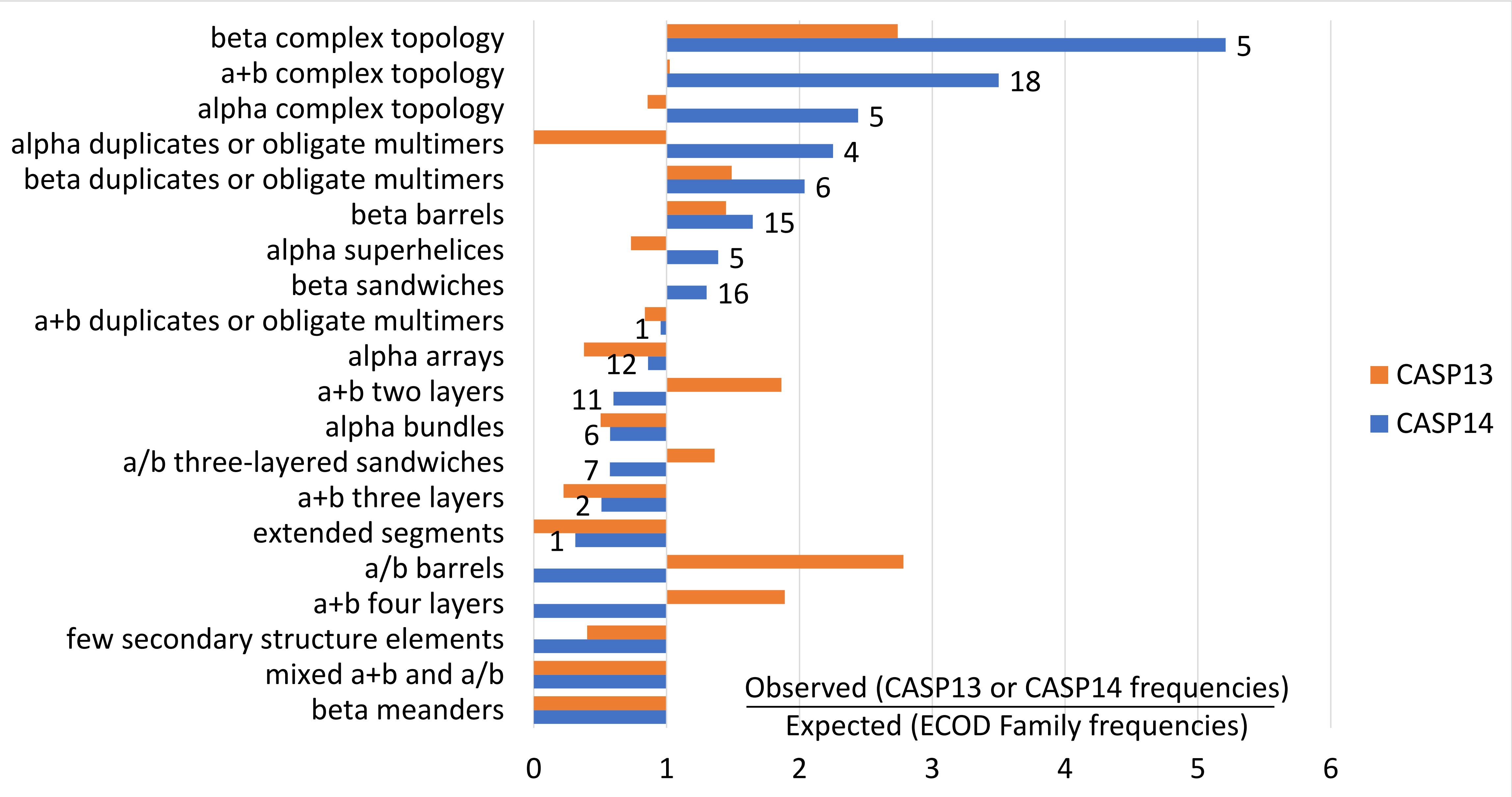
48 Single EVU Targets
6 have 18 merged ECOD domains

19 Split Domain Targets (48 EVUs)
4 have 10 merged ECOD domains

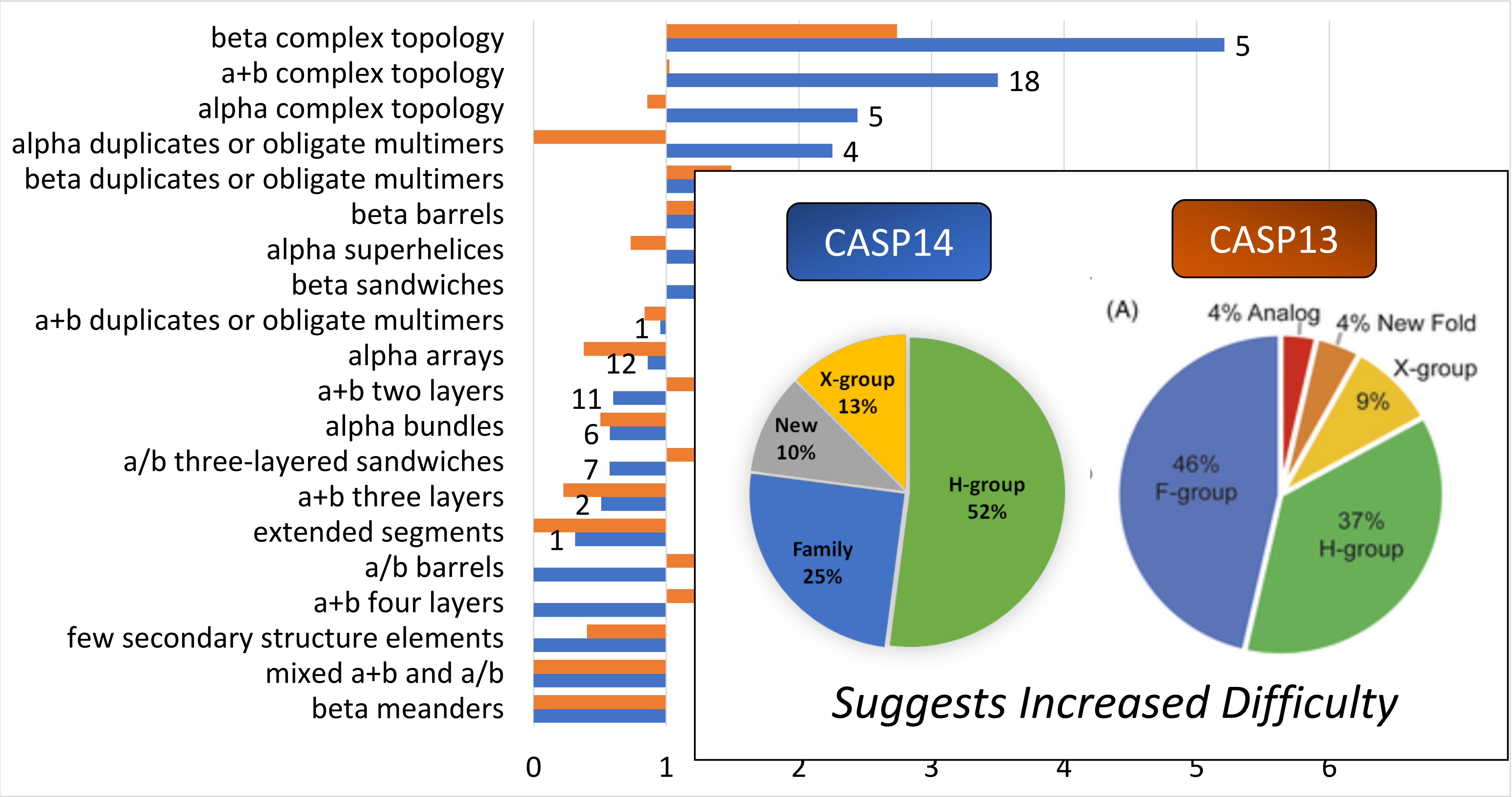
ECOD Classification based on distance to template:	
Class	**Definition**
Family (24EVU)	Template is in the same cdd
H-group (50 EVU)	Template is homologous
X-group (12 EVU)	Topological similarities
New (10 EVU)	Unique combination of SSEs



ECOD Architectures: CASP14 compared to CASP13



ECOD Relationships: CASP14 compared to CASP13

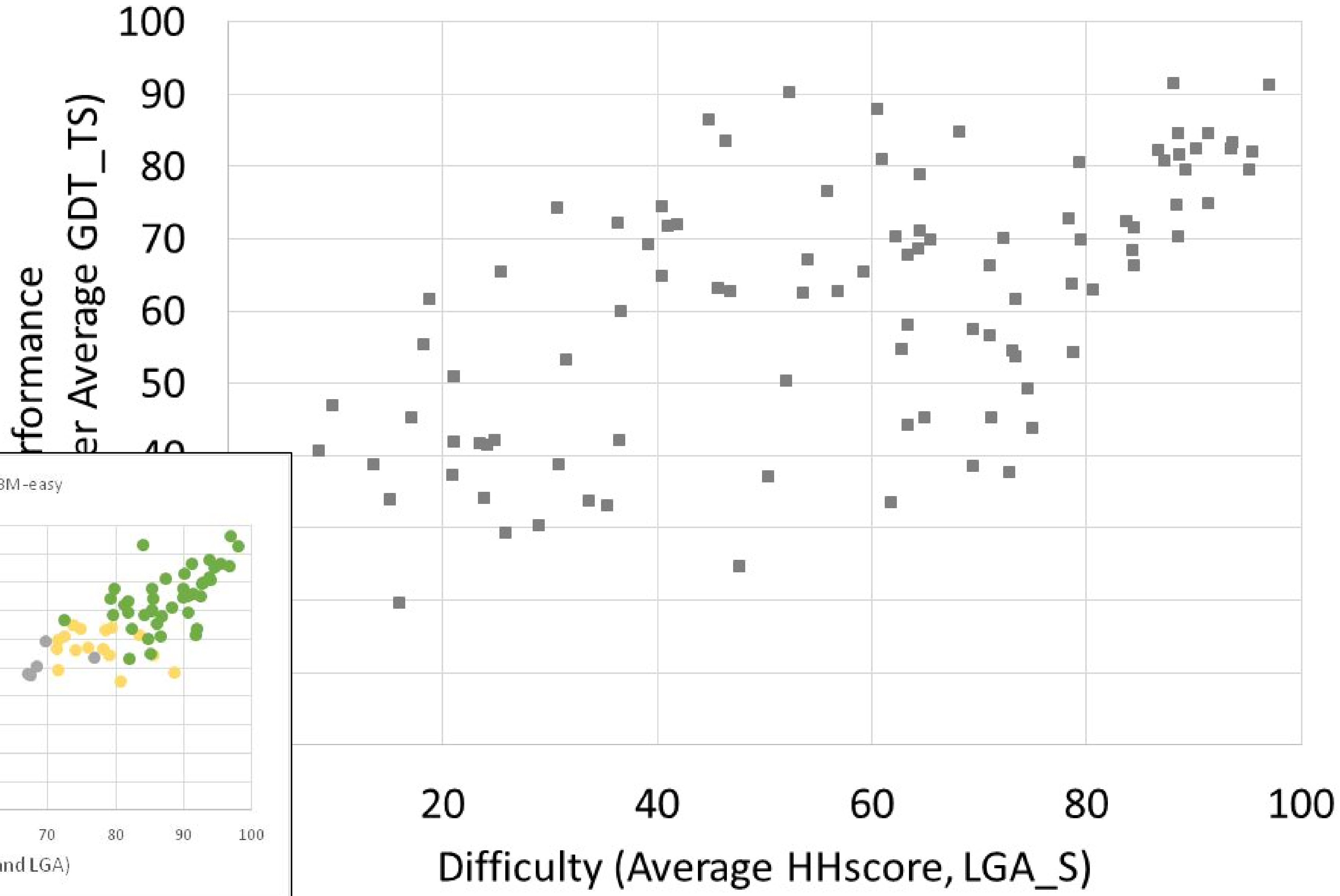


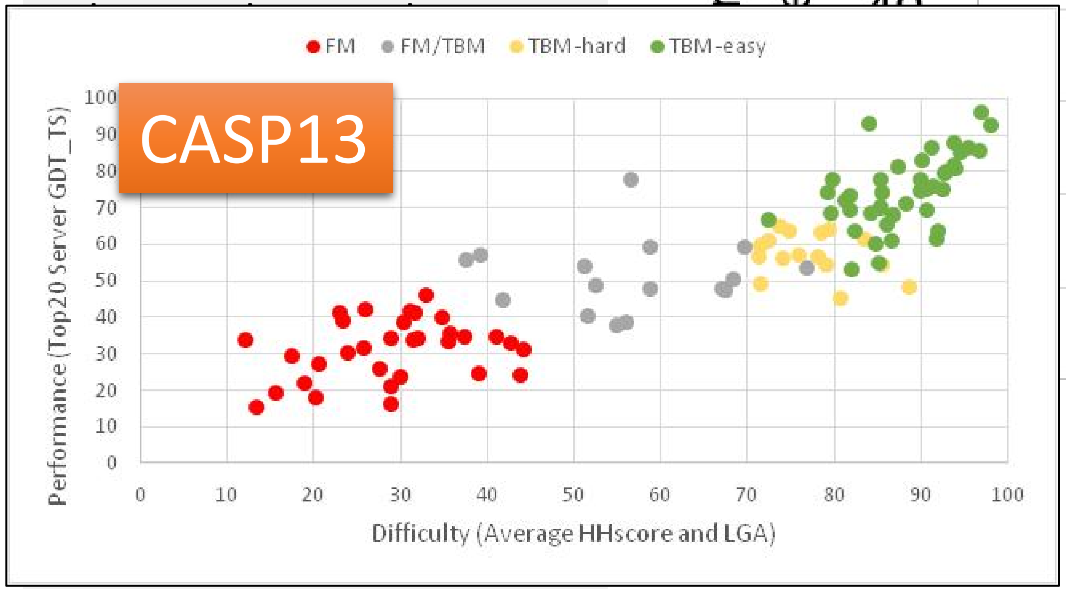
Traditional CASP Classification Plot: *Scatter is Broad*

EVU Classification Scores

Hhscore = HHprobability x Coverage for Chosen Template

Use higher HHscore from 2 methods: Uniprot100 or PDB70 for query profile





Traditional CASP Classification Plot: *Scatter is Broad*

EVU Classification Scores

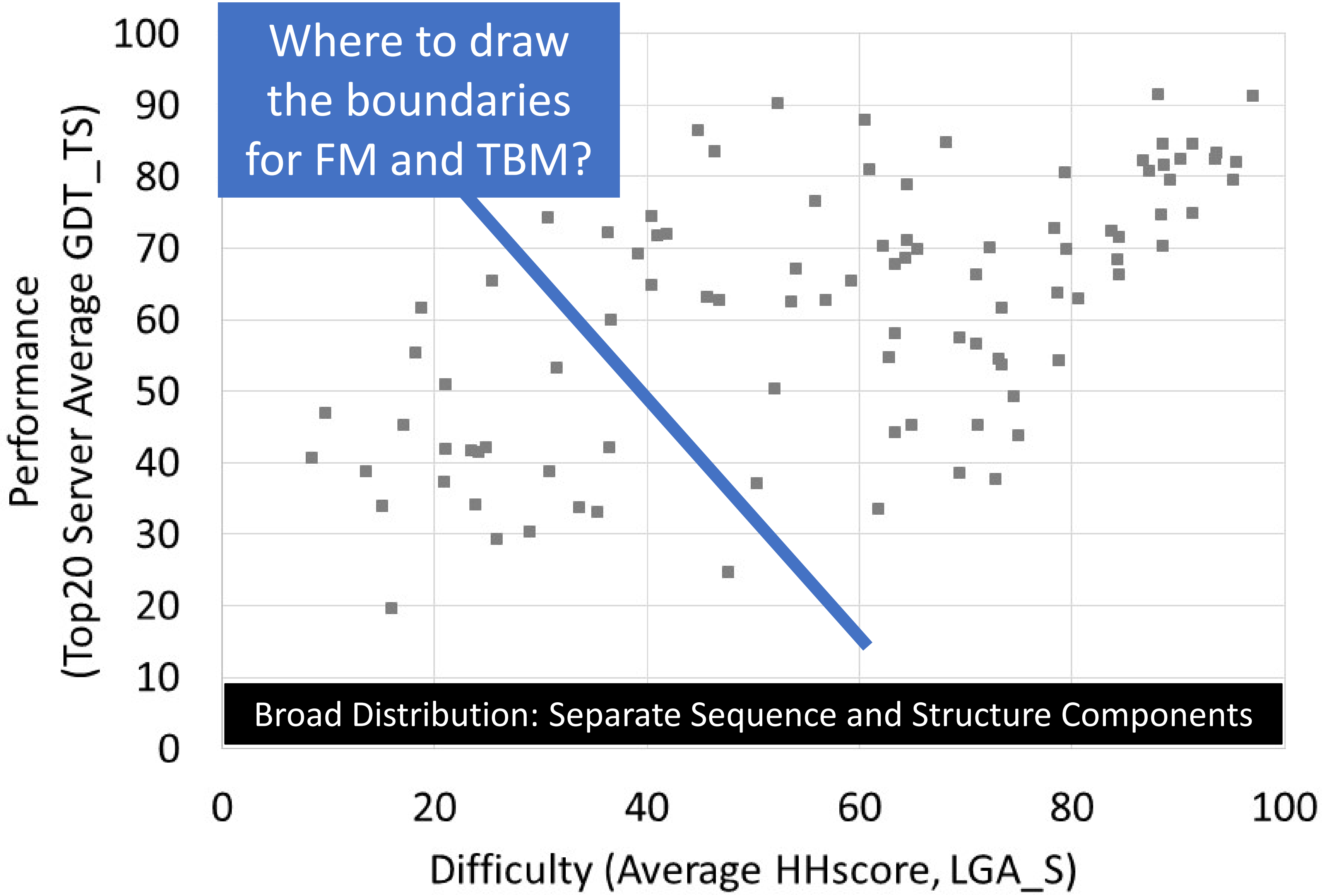
Hhscore = HHprobability x Coverage for Chosen Template

Use higher HHscore from 2 methods: Uniprot100 or PDB70 for query profile

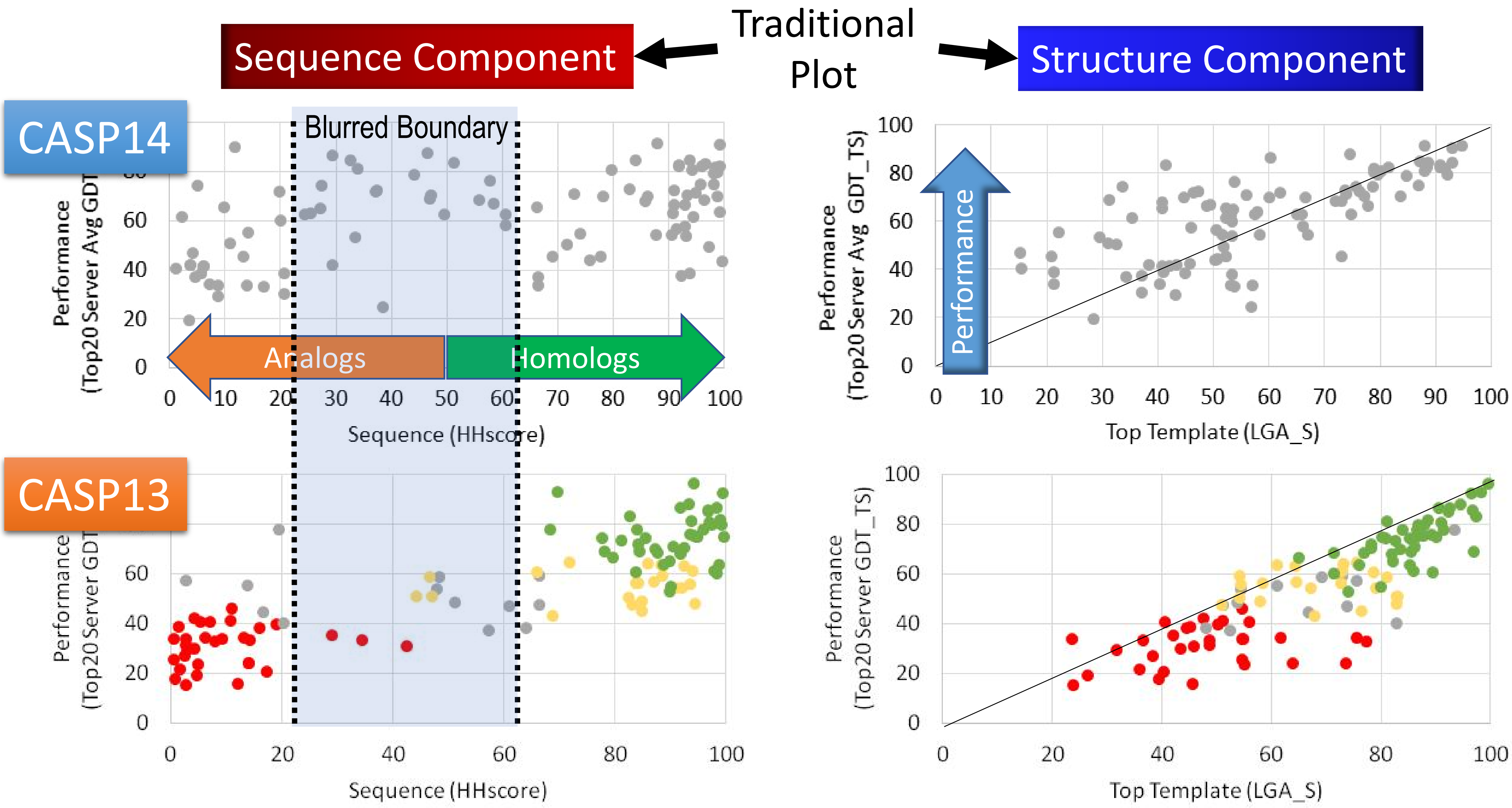
Select Rank1 template unless

- Max HHscore > for alternate homolog
- Lower rank homolog replaces analog

Top LGA_S from homolog or analogous fragment



What Contributes to Broadened Scatter?



Cluster Data to Help Confirm Classification Bounds

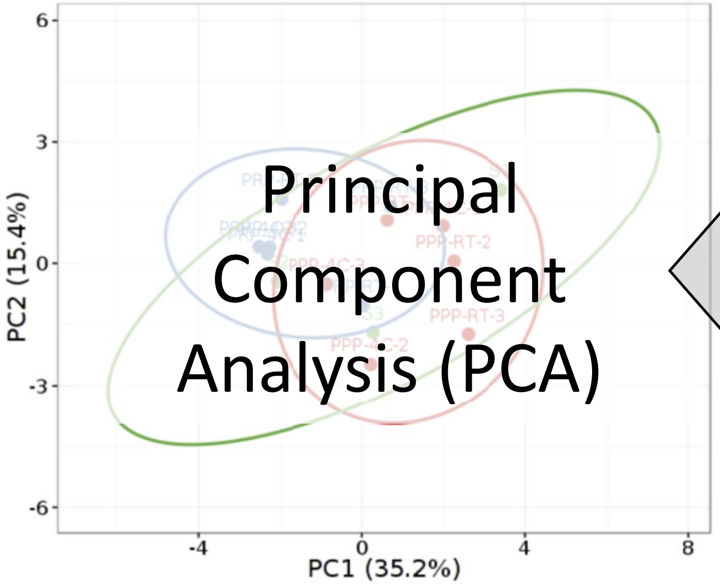
		Target1	Target2	Target3	Target4	Target5	Target6	Target7	Target8	Target96	
	feature1										*Classification*
	feature2										*ECOD Level*
Hhscore	score1	s11	s12	s13	s14	s15	s16	s17	s18	s196	
LGA_S	score2	s21	s22	s23	s24	s25	s26	s27	s28	s296	
Top20server	score3	s31	s32	s33	s34	s35	s36	s37	s38	s396	
Dali%self	score4	s41	s42	s43	s44	s45	s46	s47	s48	s496	
DaliCoverage	score5	s51	s52	s53	s54	s55	s56	s57	s58	s596	
Neff (%max)	score6	s61	s62	s63	s64	s65	s66	s67	s68	s696	
%parent Template	score7	s71	s72	s73	s74	s75	s76	s77	s78	s796	

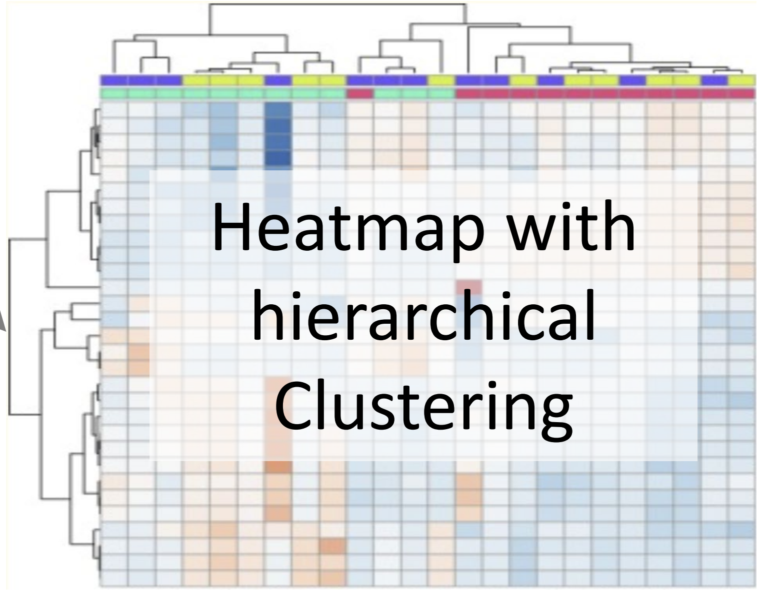
Features

Data Matrix

eigenvectors

PC1 PC2 ...

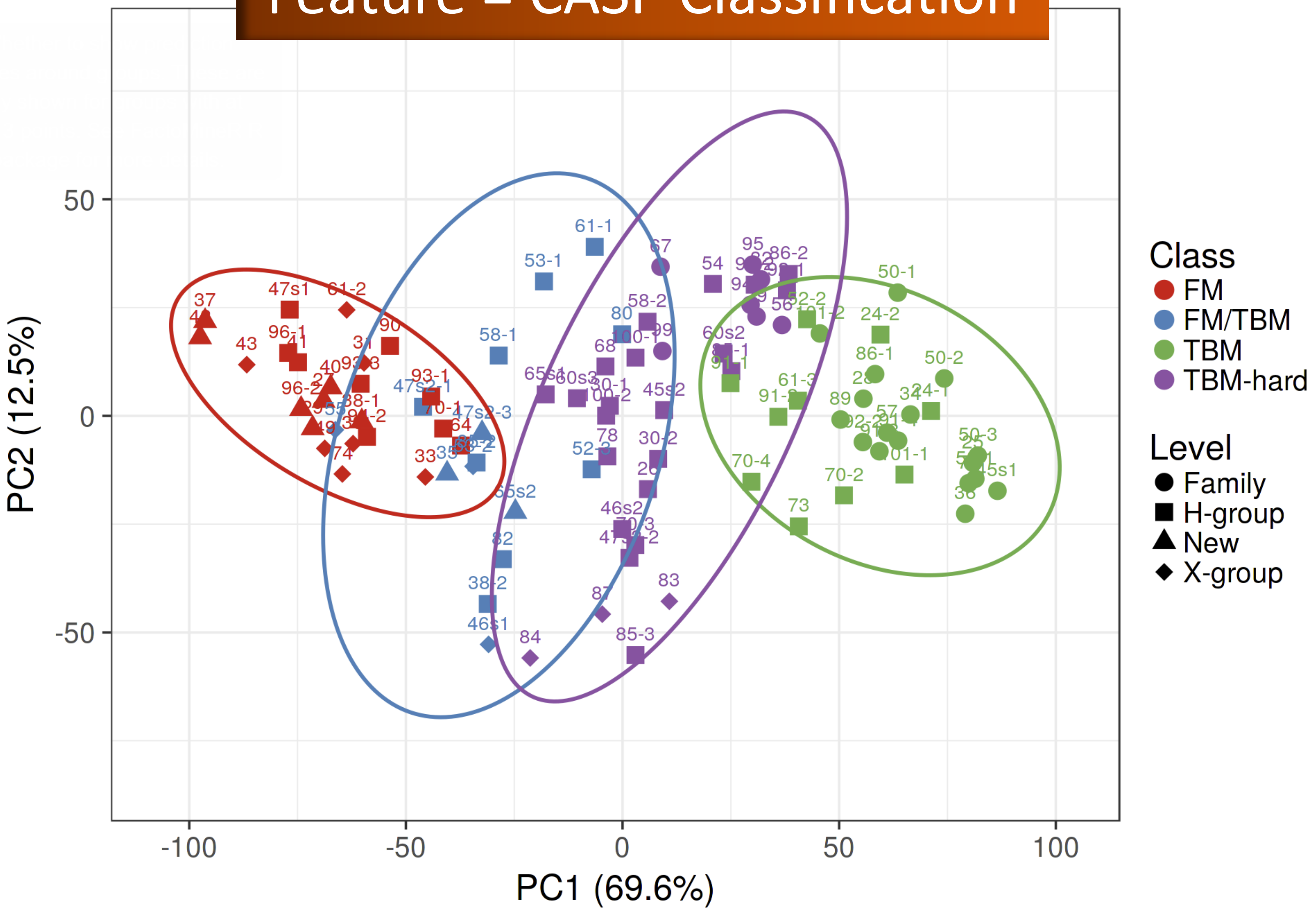




PMID: 25969447

PCA Plot of Targets Roughly Separates Classes

Feature = CASP Classification



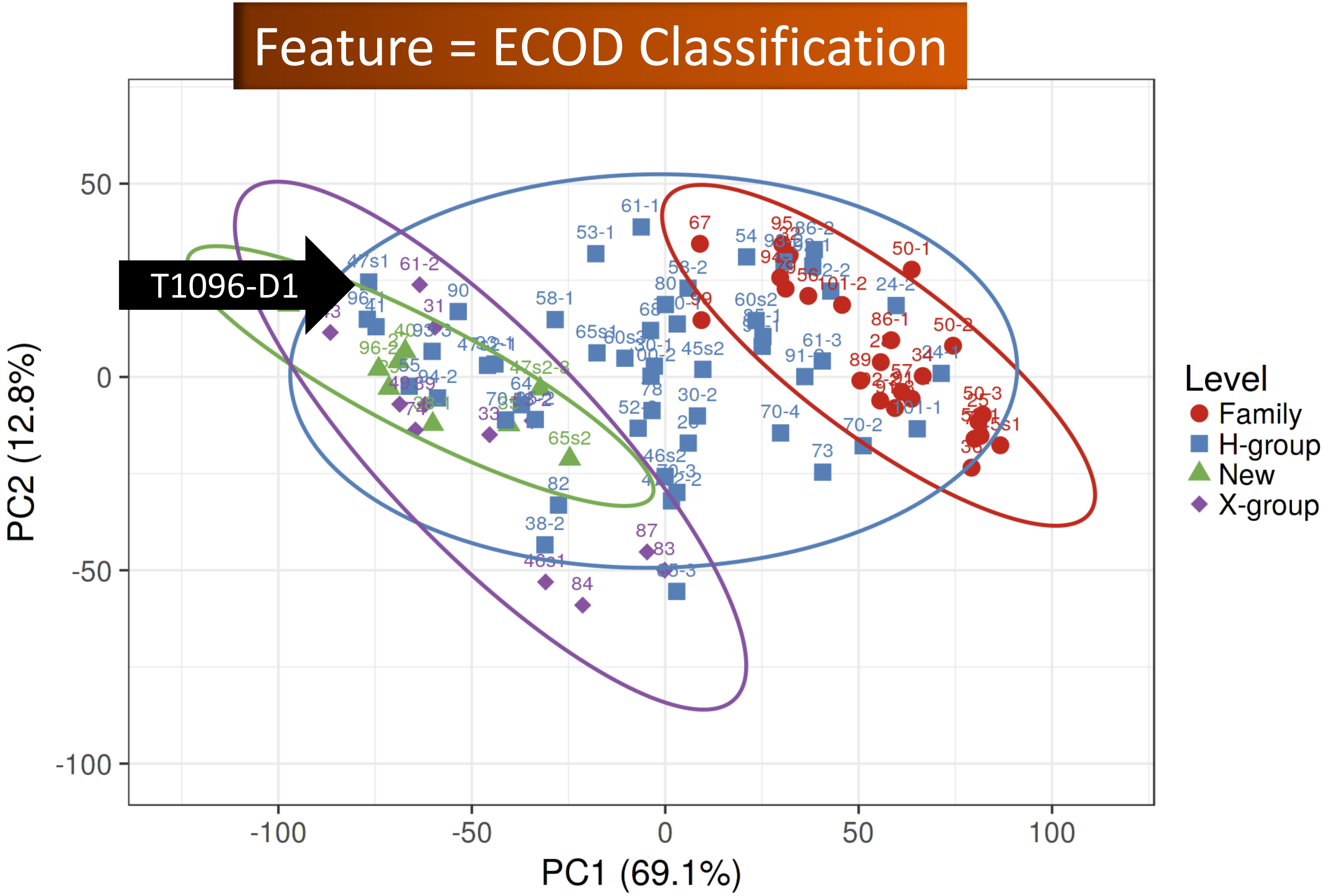
Scores Used:
HHscore
%parentTBM
Neff%max
performance
TopLGA
Dali%self
DaliCvg

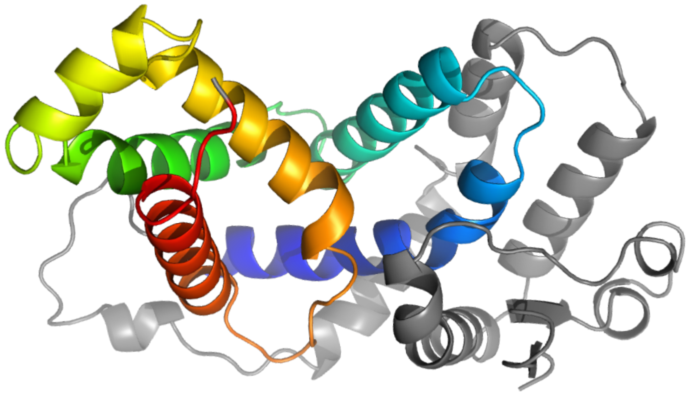
Data Preprocessing:
No scaling, rows centered

PCA Method:
SVD with Imputation

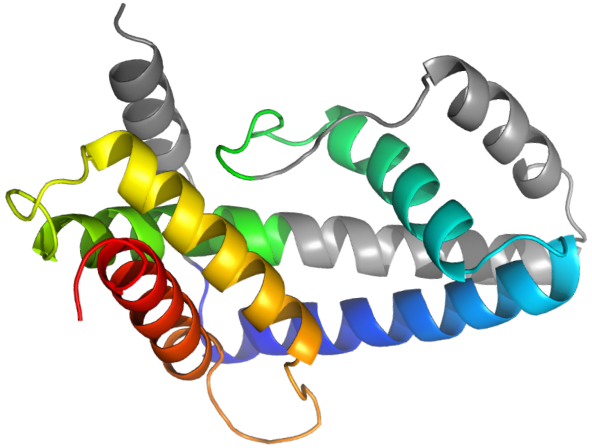
Prediction ellipses:
Probability 0.95

PCA Plot of Targets Roughly Separates ECOD Groups



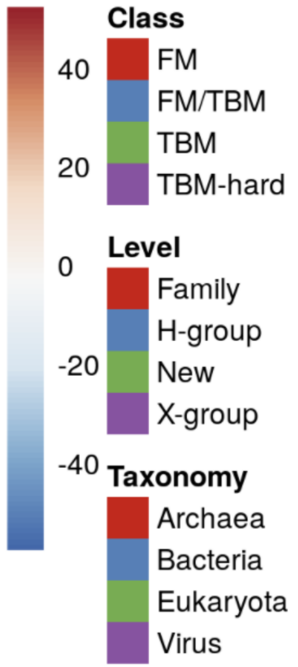


T1096-D1
Phage RNA Pol Subunit

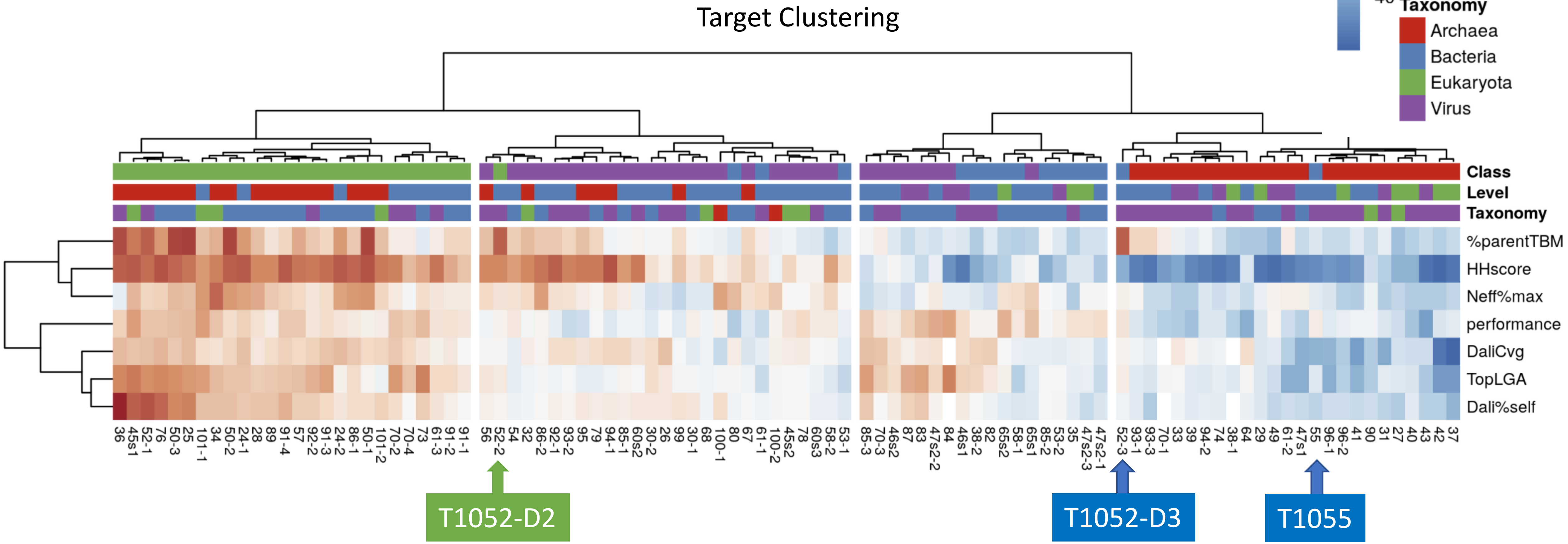


3les RNA Pol
Sigma Factor

Heatmap Clusters Targets by Classes

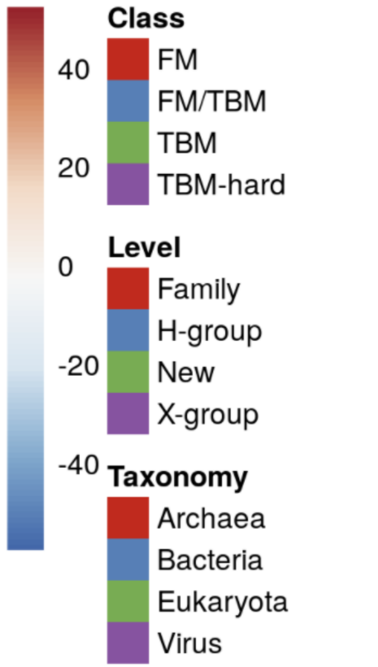


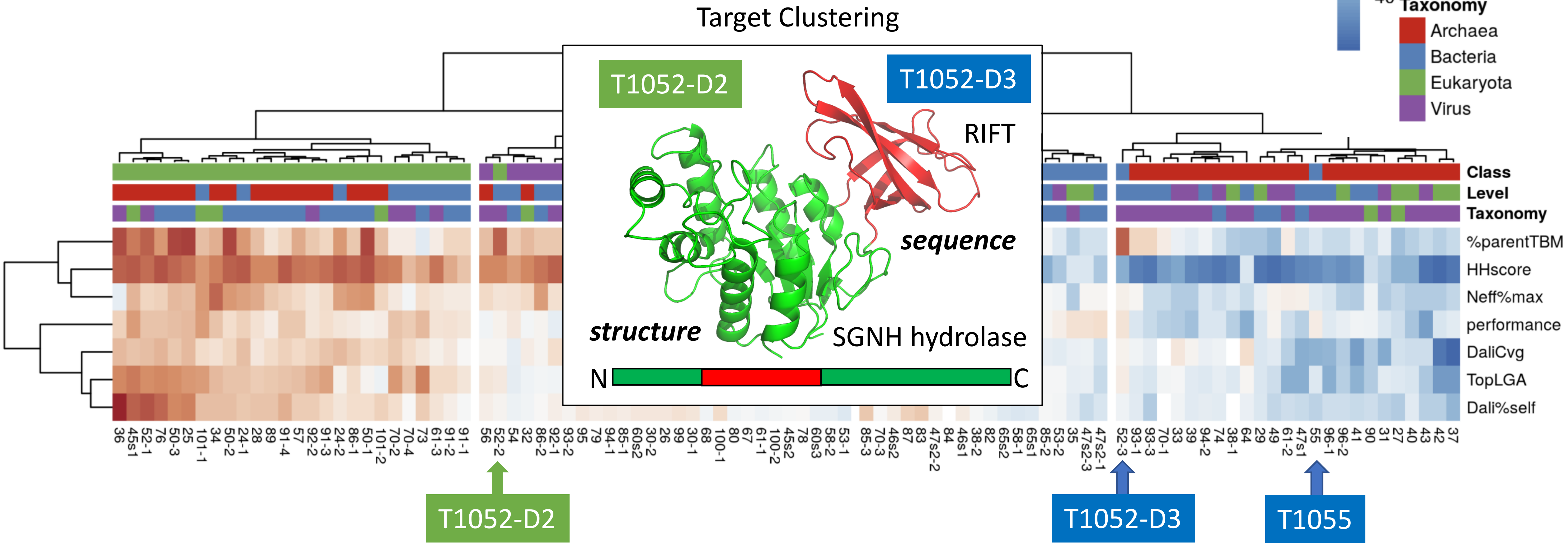
No scaling is applied to rows. Imputation is used for missing value estimation. Rows are clustered using correlation distance and Ward linkage. Columns are clustered using Euclidean distance and Ward linkage. 7 rows, 96 columns.



Heatmap Clusters Targets by Classes

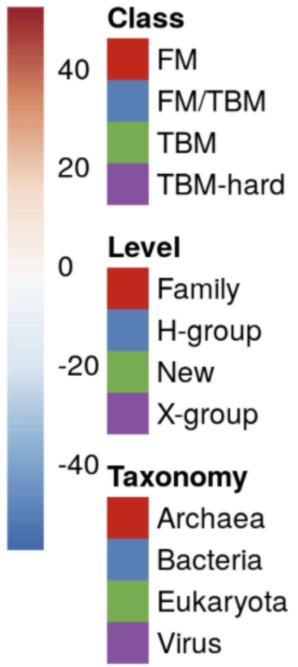
No scaling is applied to rows. Imputation is used for missing value estimation. Rows are clustered using correlation distance and Ward linkage. Columns are clustered using Euclidean distance and Ward linkage. 7 rows, 96 columns.

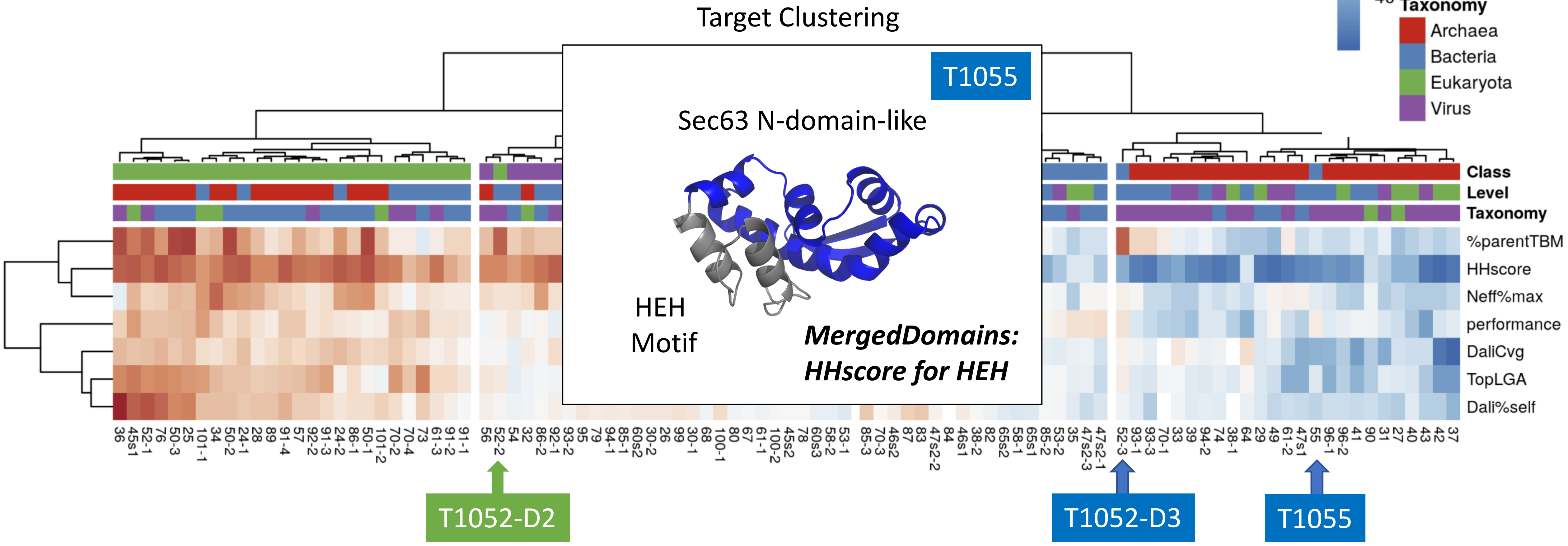




Heatmap Clusters Targets by Classes

No scaling is applied to rows. Imputation is used for missing value estimation. Rows are clustered using correlation distance and Ward linkage. Columns are clustered using Euclidean distance and Ward linkage. 7 rows, 96 columns.



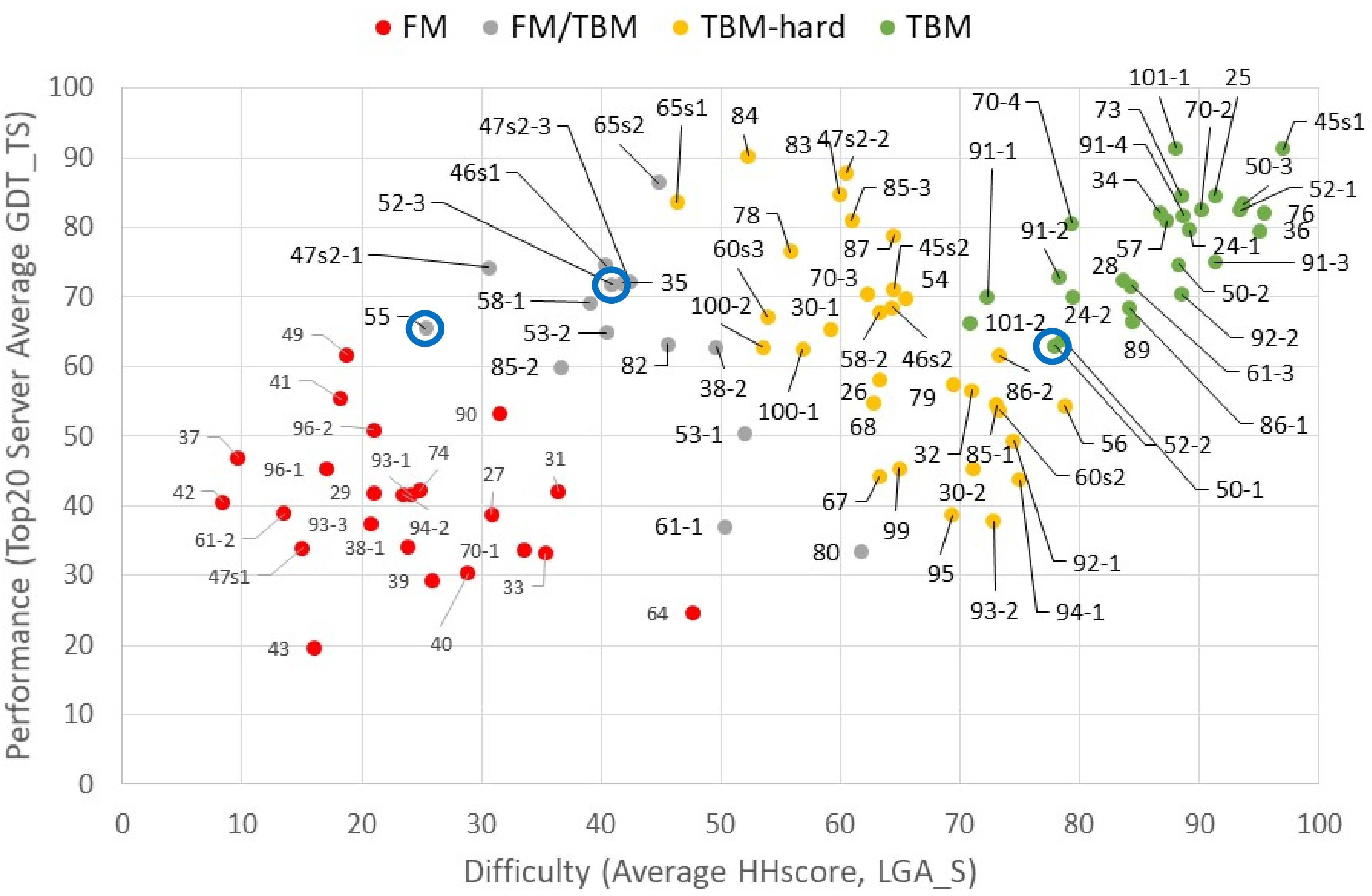


Traditional CASP Classification Plot: Outliers

Domains at the edge:
i.e. near the boundary in the traditional classification

Most domains were classified by the traditional scatter (to be consistent with CASP13)

T1055, T1052-D2 and T1052-D3 cluster differently by heatmaps, but are classified by the scatter



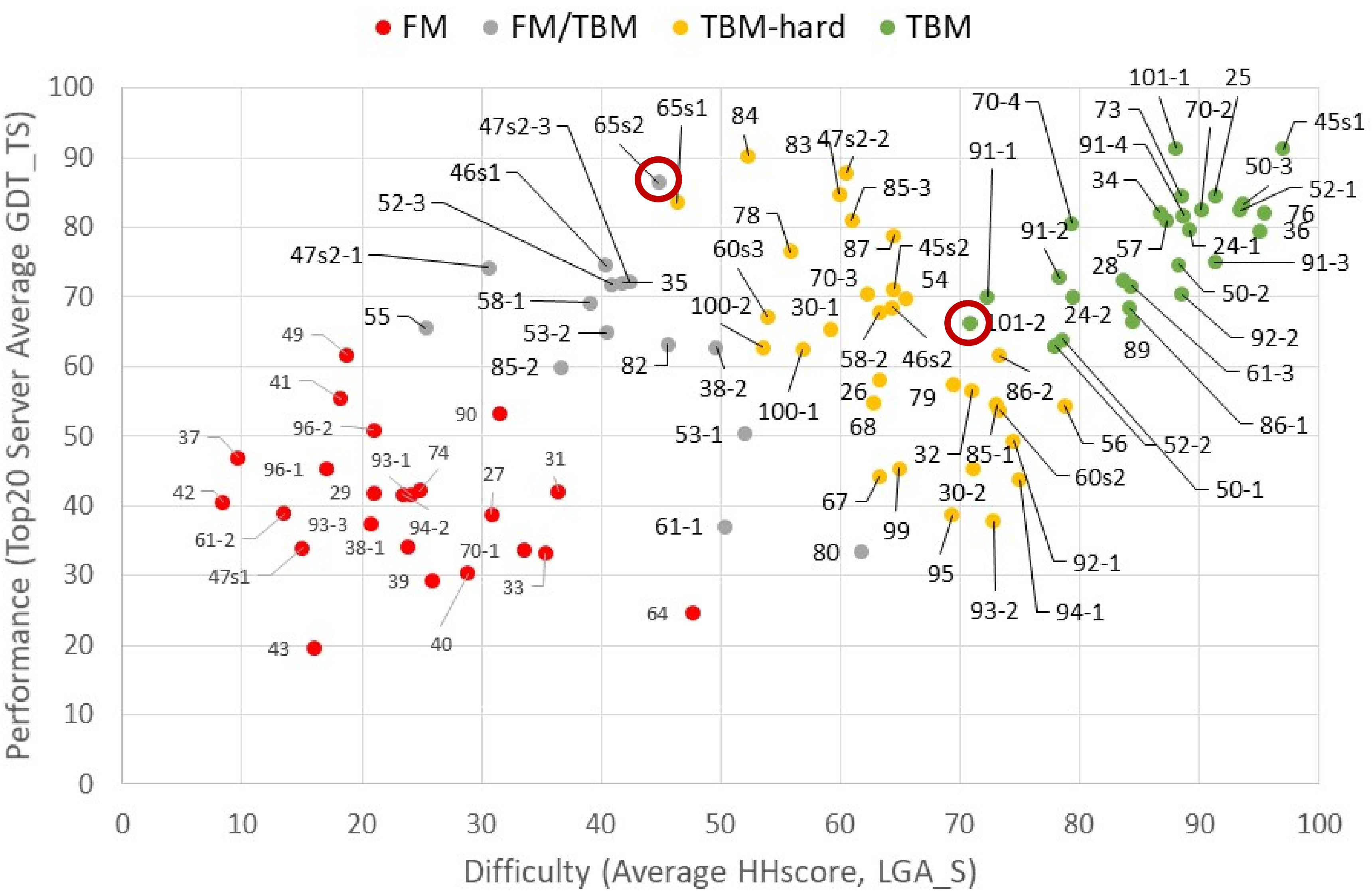
Traditional CASP Classification Plot: Outliers

Domains at the edge:
i.e. near the boundary in the traditional classification

Most domains were classified by the traditional scatter (to be consistent with CASP13)

T1055, T1052-D2 and T1052-D3 cluster differently by heatmaps, but are classified by the scatter

T101-2 and T1065s2 cluster differently by the scatter, but are classified by the heatmap groups



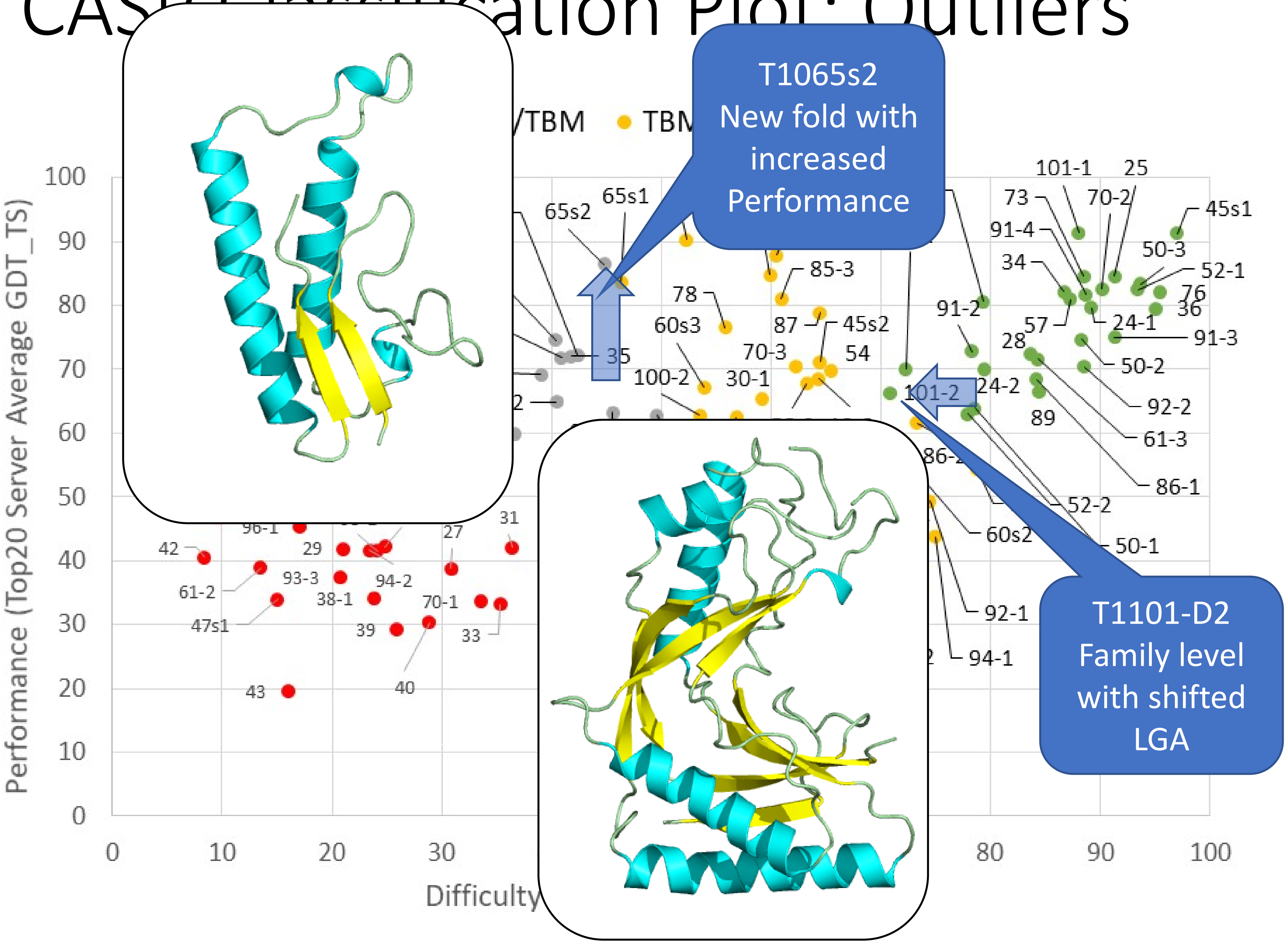
Traditional CASP Classification Plot: Outliers

Domains at the edge:
i.e. near the boundary in the traditional classification

Most domains were classified by the traditional scatter (to be consistent with CASP13)

T1055, T1052-D2 and T1052-D3 cluster differently by heatmaps, but are classified by the scatter

T101-2 and T1065s2 cluster differently by the scatter, but are classified by the heatmap groups



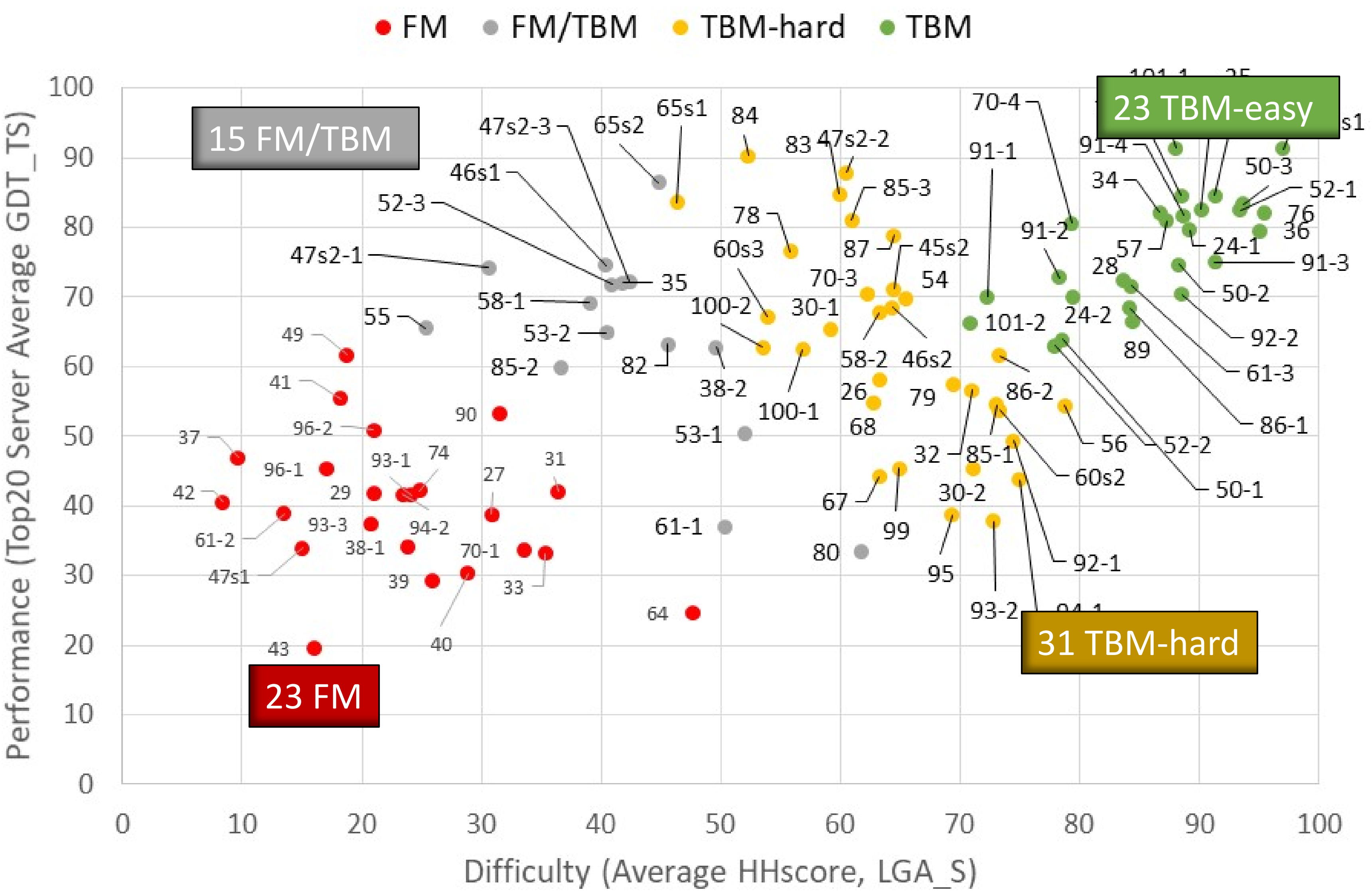
Final Traditional CASP Classification Plot

Domains at the edge:
i.e. near the boundary in the traditional classification

Most domains were classified by the traditional scatter (to be consistent with CASP13)

T1055, T1052-D2 and T1052-D3 cluster differently by heatmaps, but are classified by the scatter

T101-2 and T1065s2 cluster differently by the scatter, but are classified by the heatmap groups



Thank You!

Collaborators
Nick Grishin (UTSW)
Dustin Schaeffer (UTSW)
Jimin Pei (UTSW)
Andriy Kryshtafovych (Prediction Center)

CASP Assessors
Andrei Lupas (High Accuracy Models)
Alfonso Valencia (Contacts)
Daniel Rigden (Refinement)
Ezgi Karaca (Assembly)
Chaok Seok (Model Accuracy)
Sandor Vajda (Function)

CASP Organizing Committee
John Moult, CASP chair and founder; IBBR, University of Maryland, USA
Krzysztof Fidelis, founder, University of California, Davis, USA
Andriy Kryshtafovych, University of California, Davis, USA
Torsten Schwede, University of Basel, Switzerland
Maya Topf, Birkbeck, University of London, UK

hhmi Howard Hughes Medical Institute