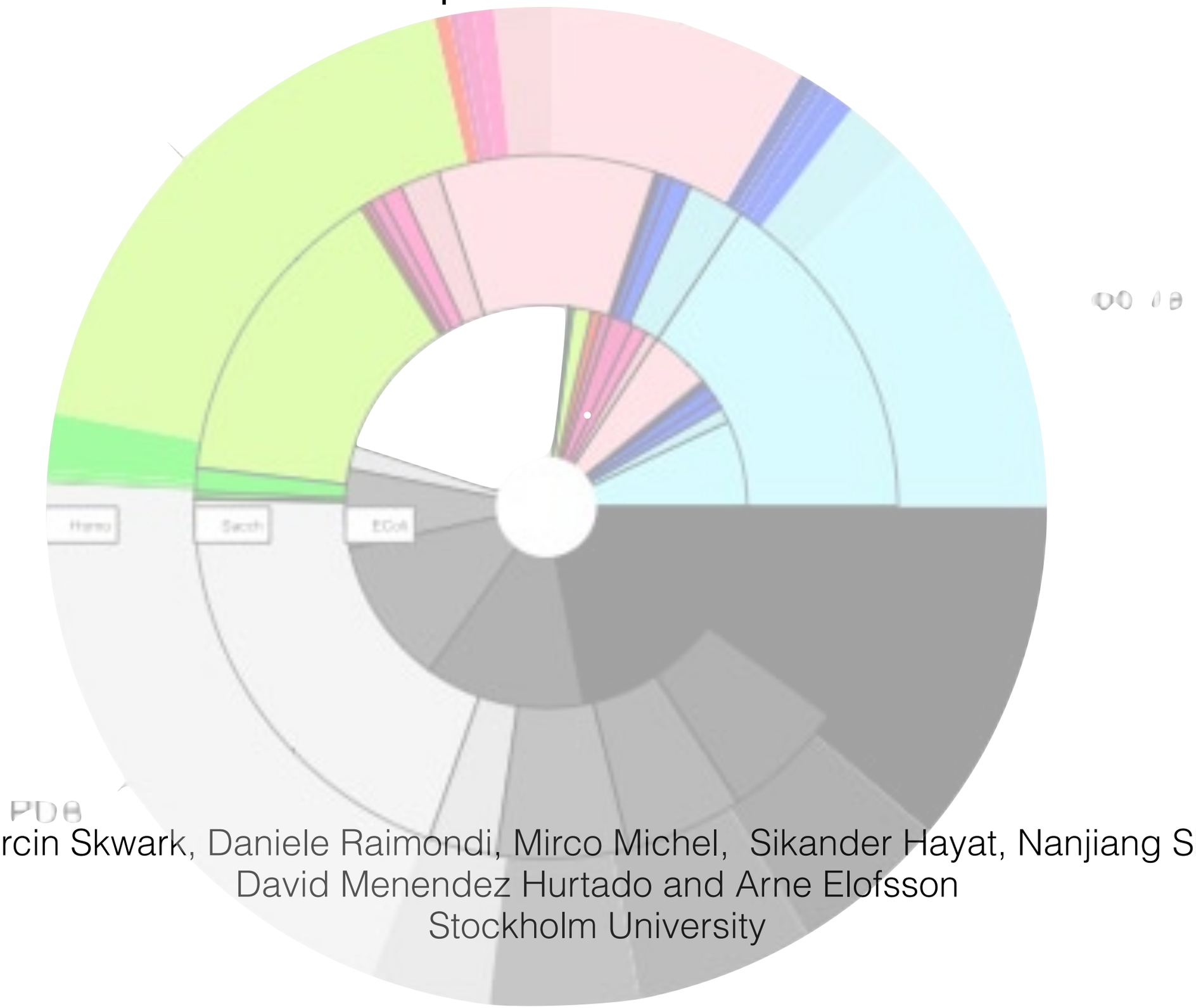
How far are we from complete structural coverage of the proteomes ?



Marcin Skwark, Daniele Raimondi, Mirco Michel, Sikander Hayat, Nanjiang Shu,
David Menendez Hurtado and Arne Elofsson
Stockholm University

Ab Initio Structure Prediction

- Introduction
- Lattice models
- Fragment based models
  - Rosetta/Robetta
- Molecular mechanics models
  - Folding@home
- Contact Predictions - the revolution

Some lides from
Howard Feldman hfeldman@blueprint.org

Ab Initio Prediction

- Predicting the 3D structure of a protein without any "prior knowledge"
- Uses when homology modeling not is possible.
- Equivalent to solving the "Protein Folding Problem"
- Similar methods useful for "Protein design"
  - Protein design is the "inverse" protein folding problem, i.e design a sequence that fold into a given fold.
  - Potentially easier and more useful

ab-initio protein structure prediction

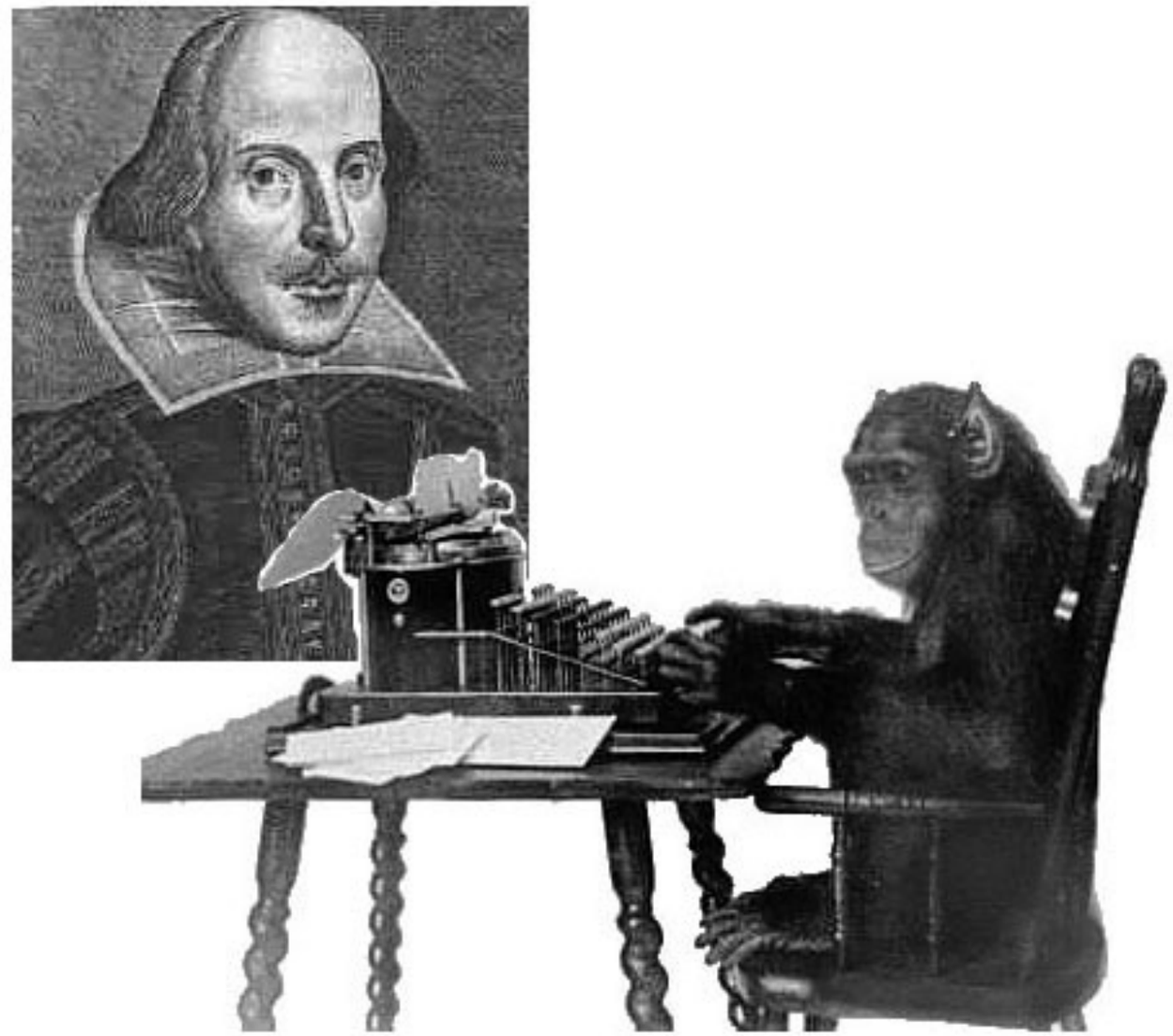
- **Optimization problem**
  - Define some initial model.
  - Define a function mapping structures to numerical values (the lower the better).
  - Solve the computational problem of finding the global minimum.
- **Simulation of the actual folding process**
  - Build an accurate initial model (including energy and forces).
  - Accurately simulate the dynamics of the system.
  - The native structure will emerge.
  - No hope due to large search space

Ab Initio Prediction

- Purists will argue must use laws of physics alone
  - But on what level ?
- However most successful methods use a blend of physics, fold recognition, and statistical probability
- Still an ongoing research problem, but becoming less essential as databases grow
  - But also useful for mini-domains and loop

Ab Initio Folding

- Two Central Problems
  - Sampling vast conformational space
  - The energy minimum problem
- The Sampling Problem (Solutions)
  - Lattice models, off-lattice models, simplified chain methods – exhaustive sampling not possible, even for small peptides
- The Energy Problem (Solutions)
  - Threading energies, packing assessment, topology assessment, physics



An infinite

number of

monkeys on

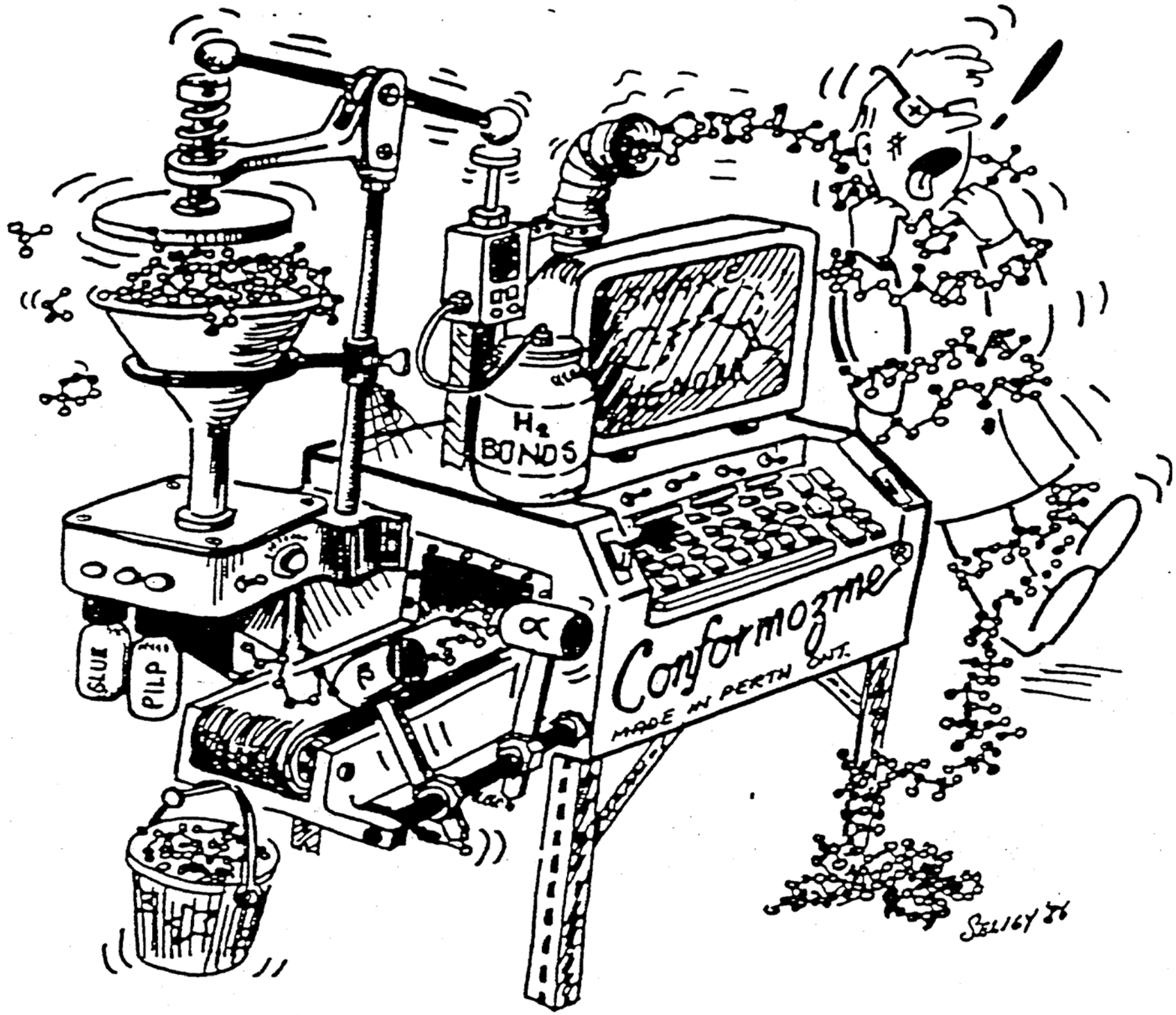
an infinite

number of

typewriters

would eventually

recreate all the works of Shakespeare, and similarly, an infinite number of CPUs could eventually fold every known protein.



CANADA'S FIRST PROTEIN FOLDING MACHINE!

Molecular mechanics based models

- Could we just use MD simulations to fold proteins.
  - Folding is in the mS to S scale
  - Current simulations is in the μS scale
  - How accurate are the energy functions
- Folding@home
  - Parallel simulations on distributed computers
    - Many mS of simulations
    - Runs on PS3 (Check our kitchen)
  - Folds small proteins
  - Can not (yet) fold big proteins.
  - Often uses implicit water models

Energy Minimization (Theory)

- Treat Protein molecule as a set of balls (with mass) connected by rigid rods and springs
- Rods and springs have empirically determined force constants
- Allows one to treat atomic-scale motions in proteins as classical physics problems

Folding@Home Distributed Computing - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://folding.stanford.edu/ Links

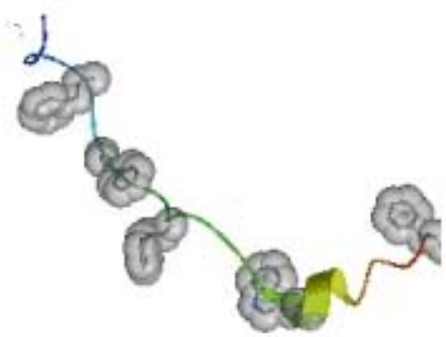
Folding@home distributed computing

Home
Download
FAQ
Forum
Help!
Education
News
Stats
Science
Results

Chinese(中文) Dutch(Nederlands) French(Français) German(Deutsch)
Italian (Italiano) Japanese(日本語) Korean(한국말) Persian(ویفارس)
Portugese(Português) Russian(Русский) Spanish(Español) Vietnamese(Tiếng Việt)

Our goal: to understand protein folding, protein aggregation, and related diseases

What are proteins and why do they "fold"? Proteins are biology's workhorses -- its "nanomachines." Before proteins can carry out their biochemical function, they remarkably assemble themselves, or "fold." The process of protein folding, while critical and fundamental to virtually all of biology, remains a mystery. Moreover, perhaps not surprisingly, when proteins do not fold correctly (i.e. "misfold"), there can be serious effects, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's, Parkinson's disease, and many cancers and cancer-related syndromes.



Results from Folding@Home

Internet

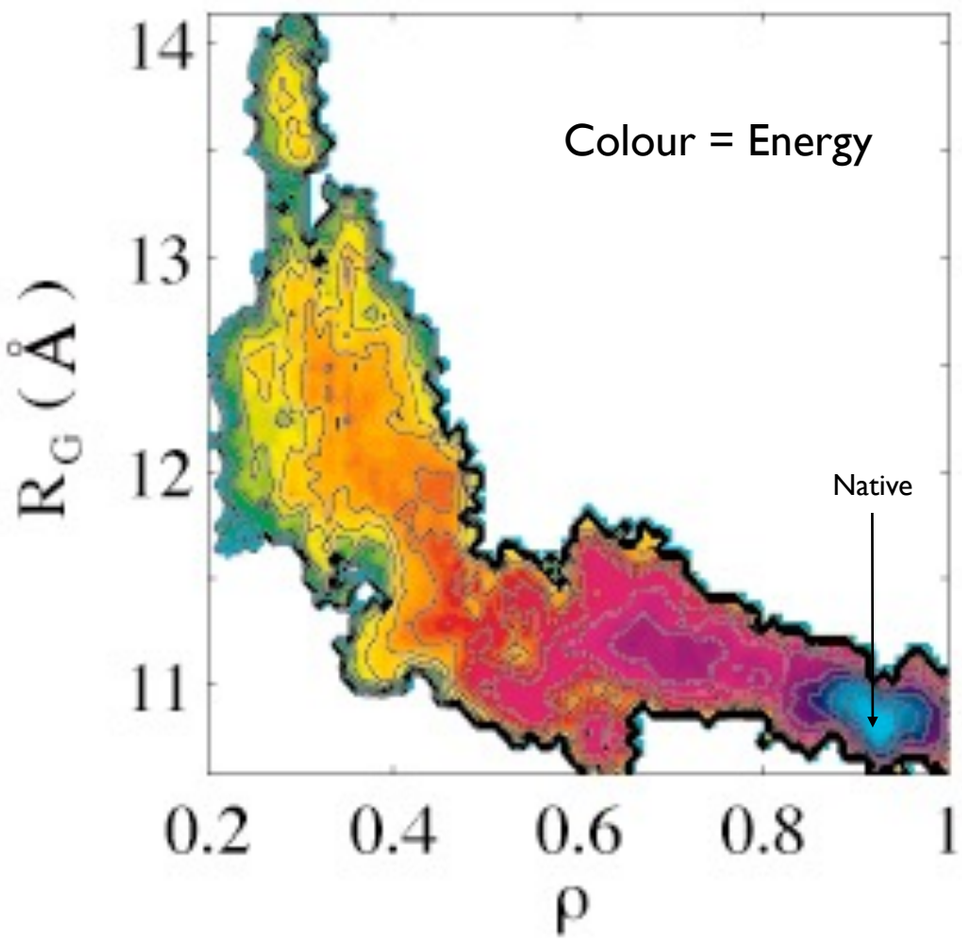
Folding@home Intro

- The work unit uses the cpu to “fold” the protein in millions of combinations and send the results back to Stanford.
- The program then downloads another work unit and repeats.
- On average 1 work unit will take anywhere from a few hours to a few days to complete on a P4 2.6Ghz CPU.

What does Folding@home do?

- Folding@home is a distributed computing project which studies protein folding, misfolding, aggregation and related diseases.
- Folding@home (F@H) uses spare cpu cycles to fold proteins in the form of Work Units (WU) and send the results to Stanford Universities servers.

The L shape of a protein folding pathway



(native contacts)

Brooks and Sheinerman's Protein G

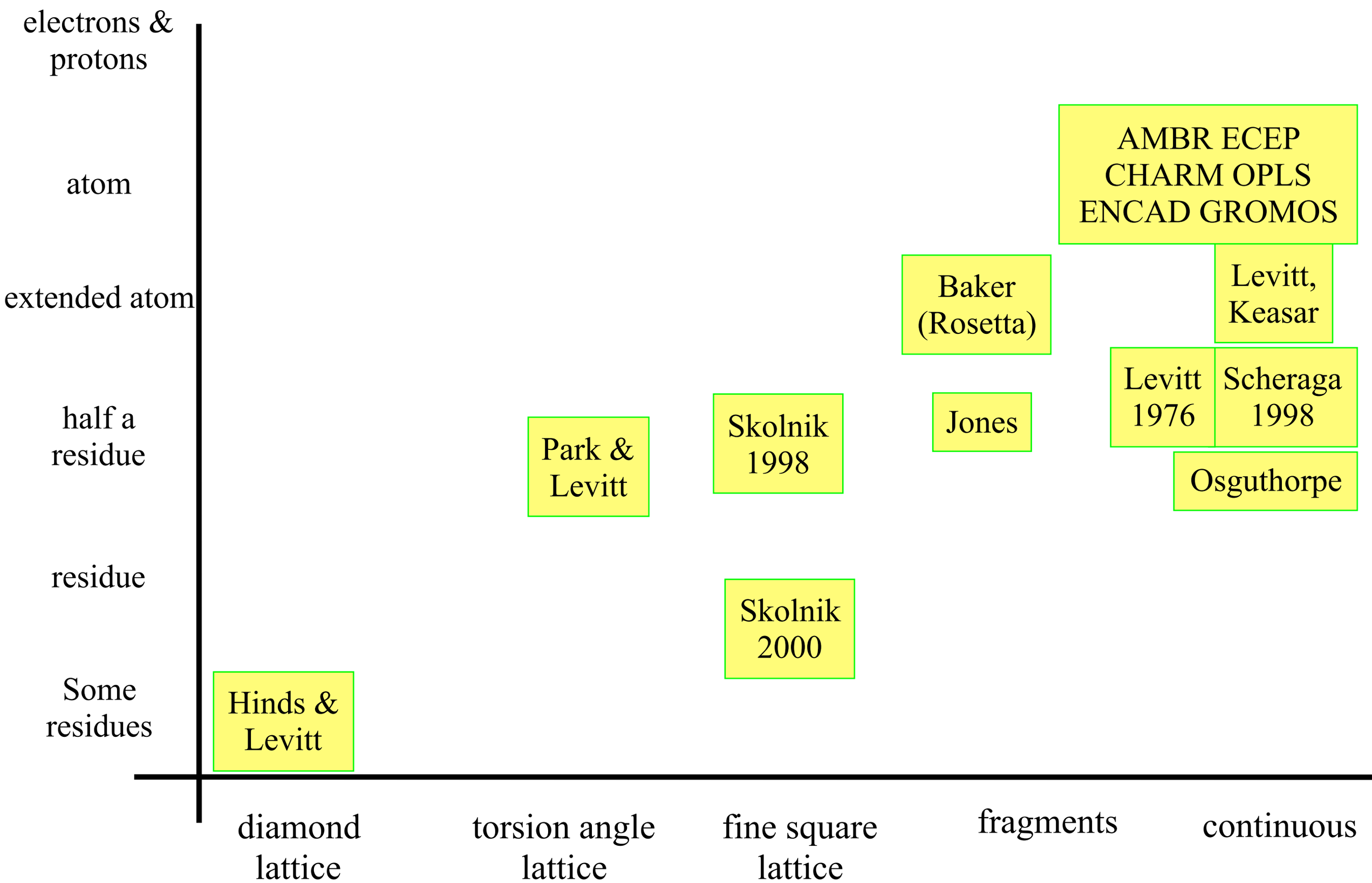
MD folding
512 Processors
Cray T3E 1 month

No "core nucleation" apparent

Folding@home video

http://www.youtube.com/watch?v=EZ1XuOgknuE

Basic element



Protein folding energy landscape

- protein energy landscape is complex, with many local minima

- believed to have a funnel-like shape, with global minimum representing native structure

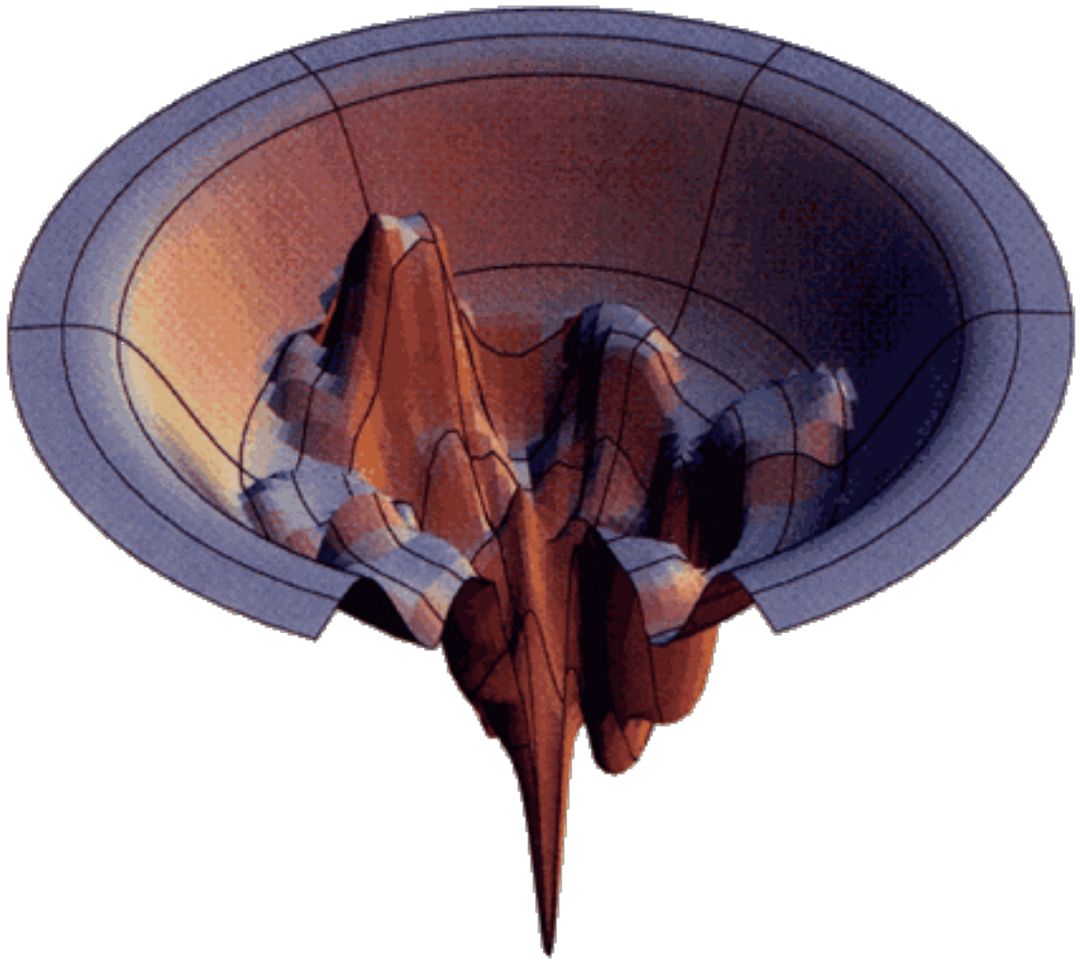


image from http://bioinfo.mshri.on.ca/

Problems with energy functions

- Not accurate enough
  - The energy difference between folded/unfolded is a often only 5-10 kcal/moles
  - 1000s of energy terms, sum of error is large
- Water
  - For accurate calculation inclusion of water is needed.
  - Implicit water models are quite slow
  - Explicit water needs time to equilibrate

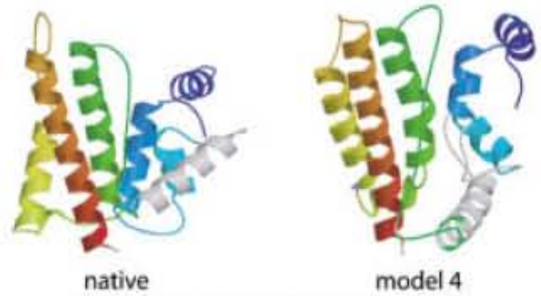
Problems (cont)

- Entropy
  - We are not searching for the energy minimum, but for the free energy minimum, i.e. MD simulations needed.
- Local minimum problem
  - The barriers are often extremely high to go from one minima to the next.
  - Sidechains cannot pass through each others

Solutions ?

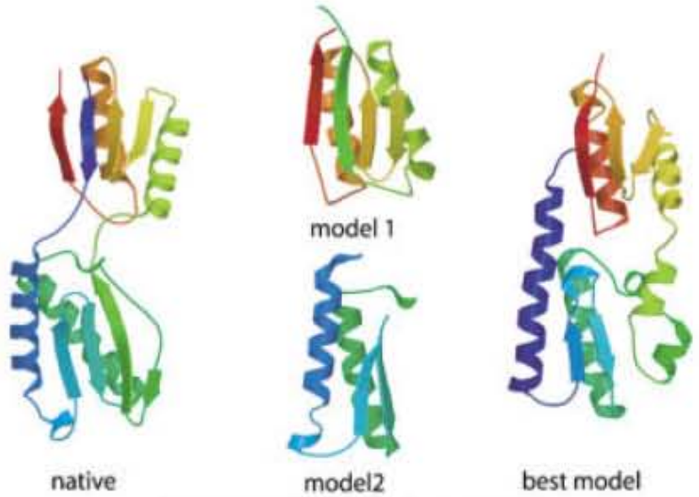
- Are there some ways around these problems
- How does proteins really fold ?
- Can we divide the problem into subproblems ?
  - Local preferences
    - Dominated by sequential information
  - Globular structures
    - Dominated by hydrophobicity
- !?!?!?! FRAGMENTS !?!?!

Rosetta - David Baker – CASP 5 structure prediction competition

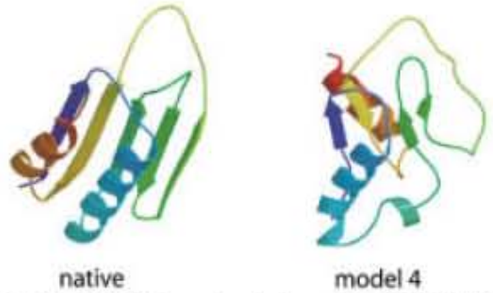


T129: HI0817 (full chain 1-182)

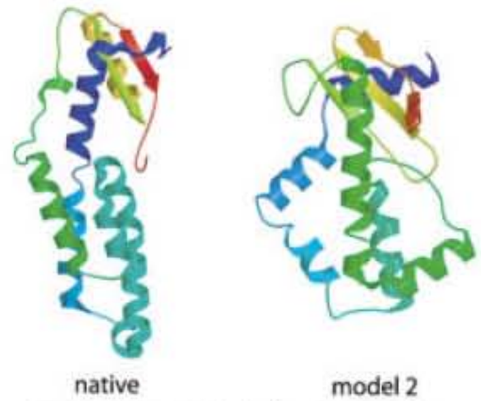
T135: Boiling stable protein (full chain 1-108)



T148: HI1034 (full chain, 1-163)



T149: yjiA (C-terminal domain, 206-318)

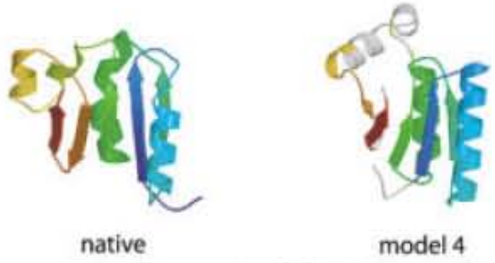


T161: HI1480 (full chain, 1-156)

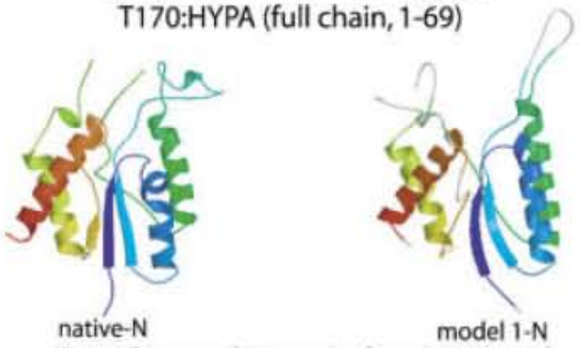
T156: MENG (full chain 1-157)

T162: (Domain 1, 1-62)

T170:HYPA (full chain, 1-69)



T191:(N-terminal domain, 1-104)



T173: Rv1170 (N-terminal region, 1-127)

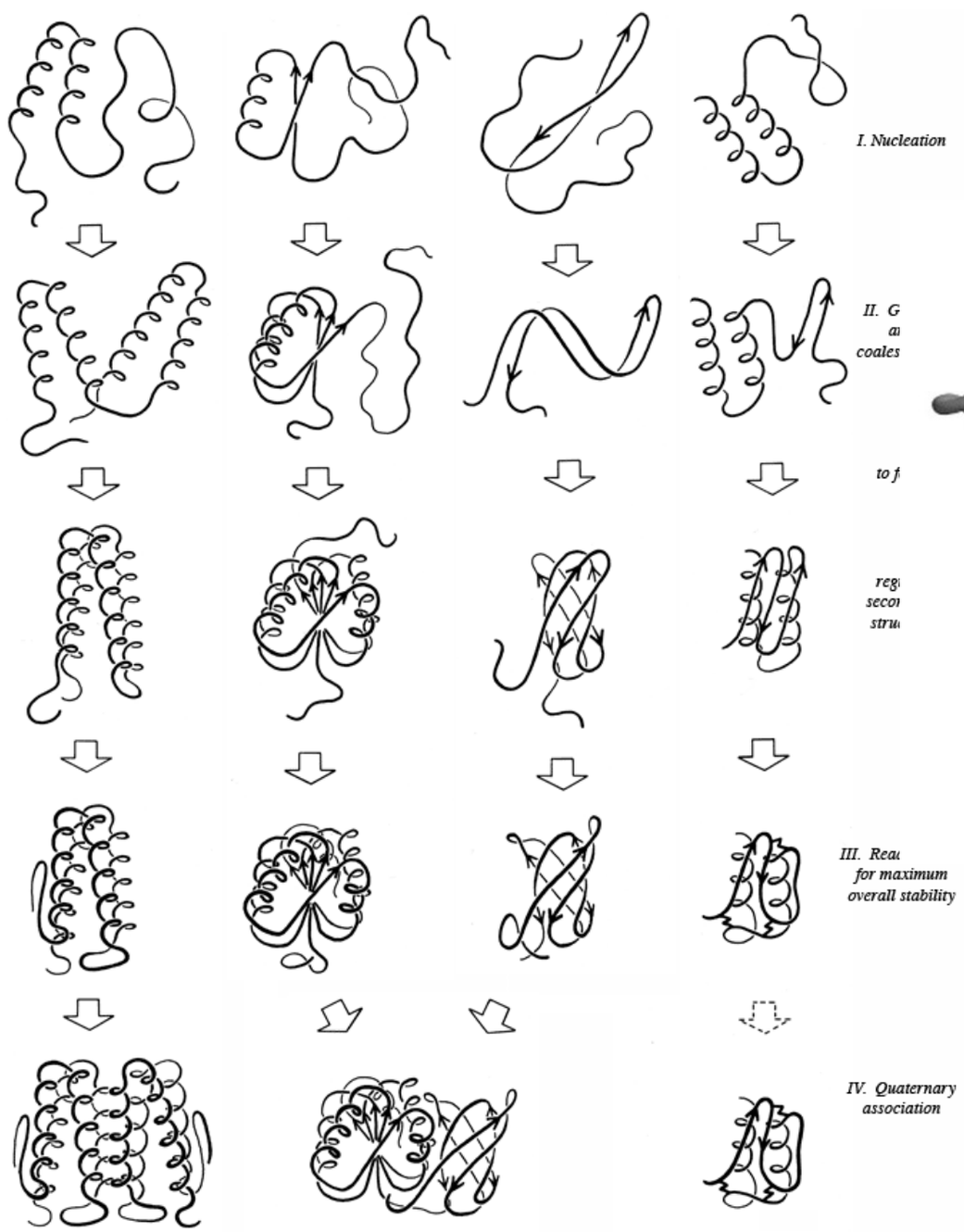
Rosetta

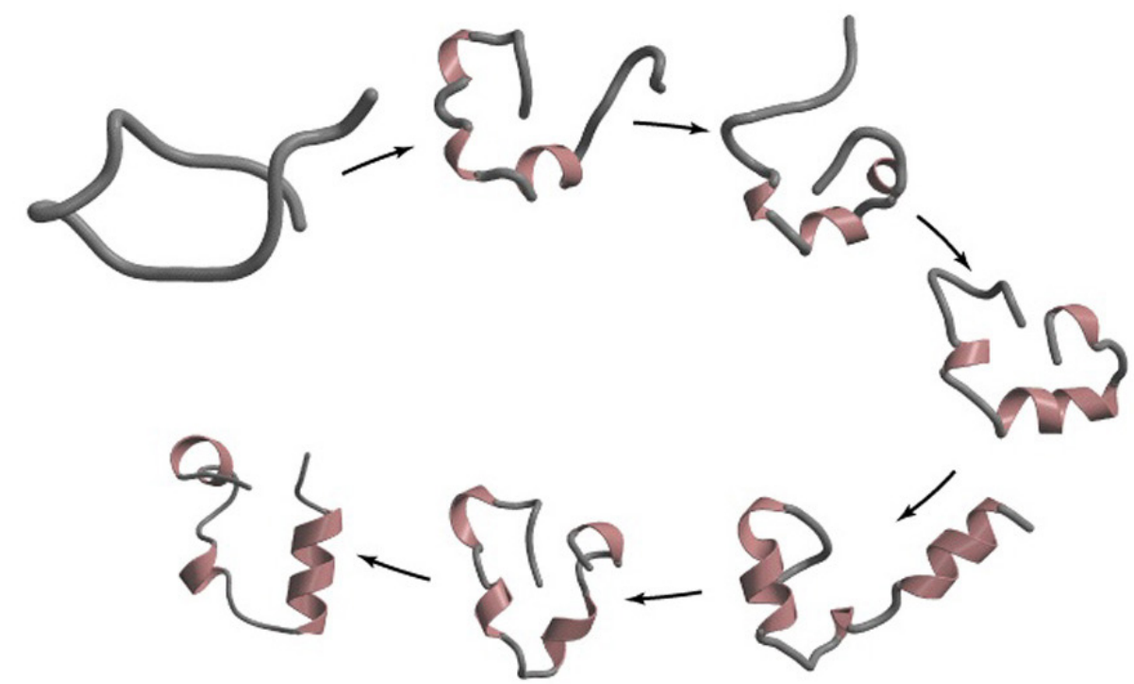
- http://www.youtube.com/watch?v=GzATbET3g54

Theory Behind Rosetta

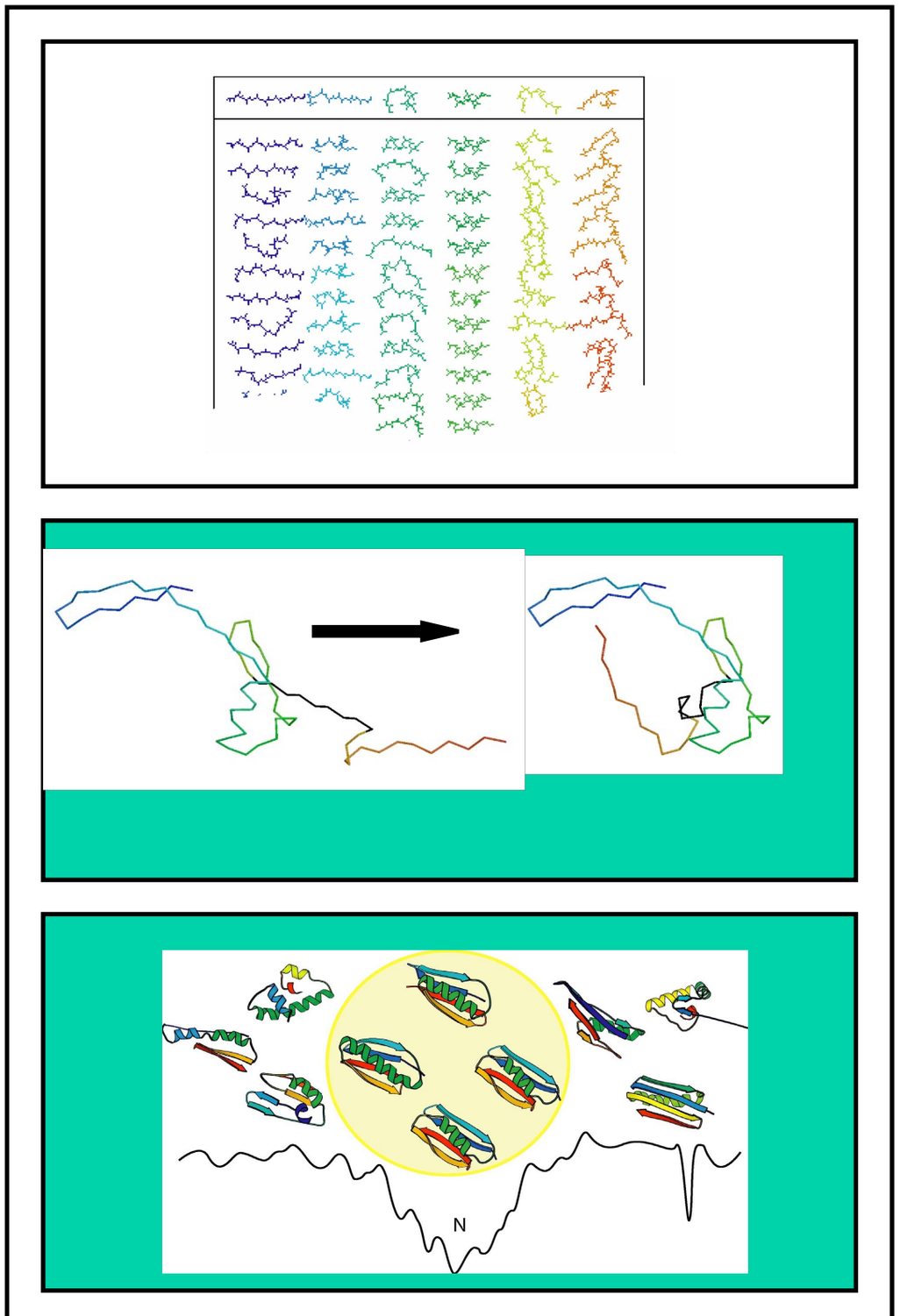
- Proteins are thought to 'collapse' from an unfolded > folded state.
- Local conformations precede and guide global conformations and tertiary structure.
- Local conformations are largely dependent on local sequence, and are finite in number.

Theory Behind Rosetta



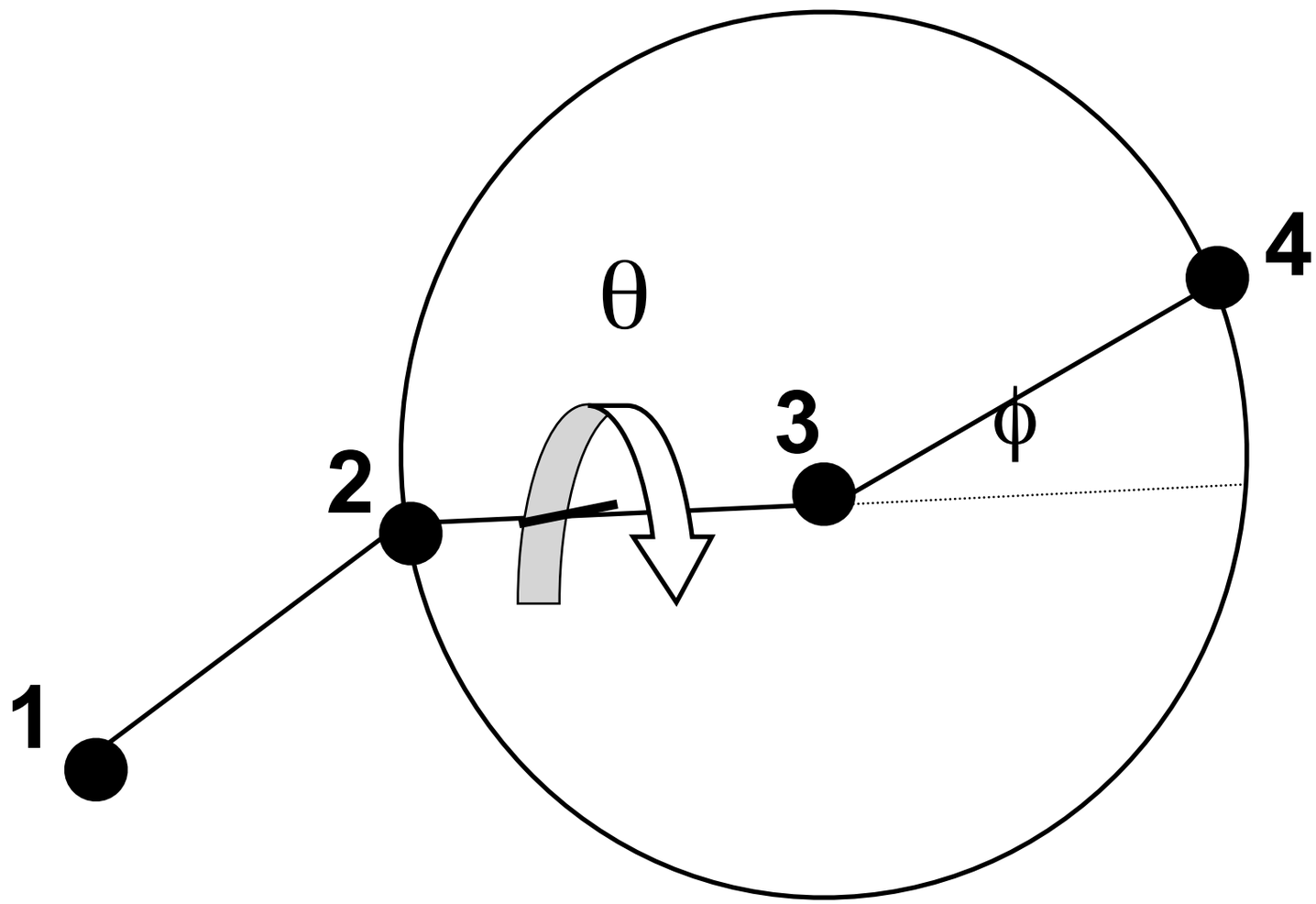


Structure Prediction with Rosetta



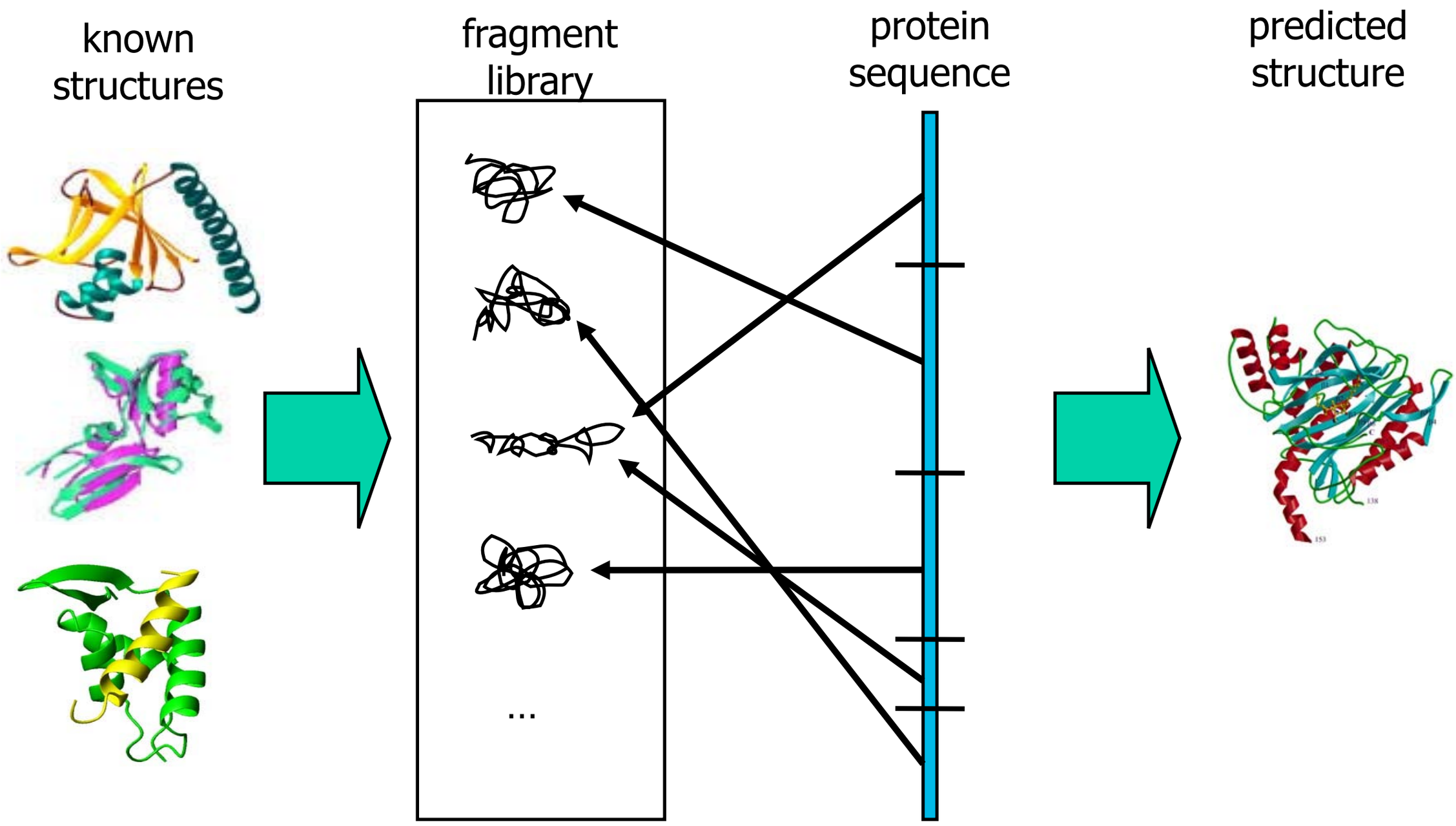
- Select fragments consistent with local sequence preferences
- Assemble fragments into models with native-like global properties
- Identify the best model from the population of decoys

Simplified Chain Representation



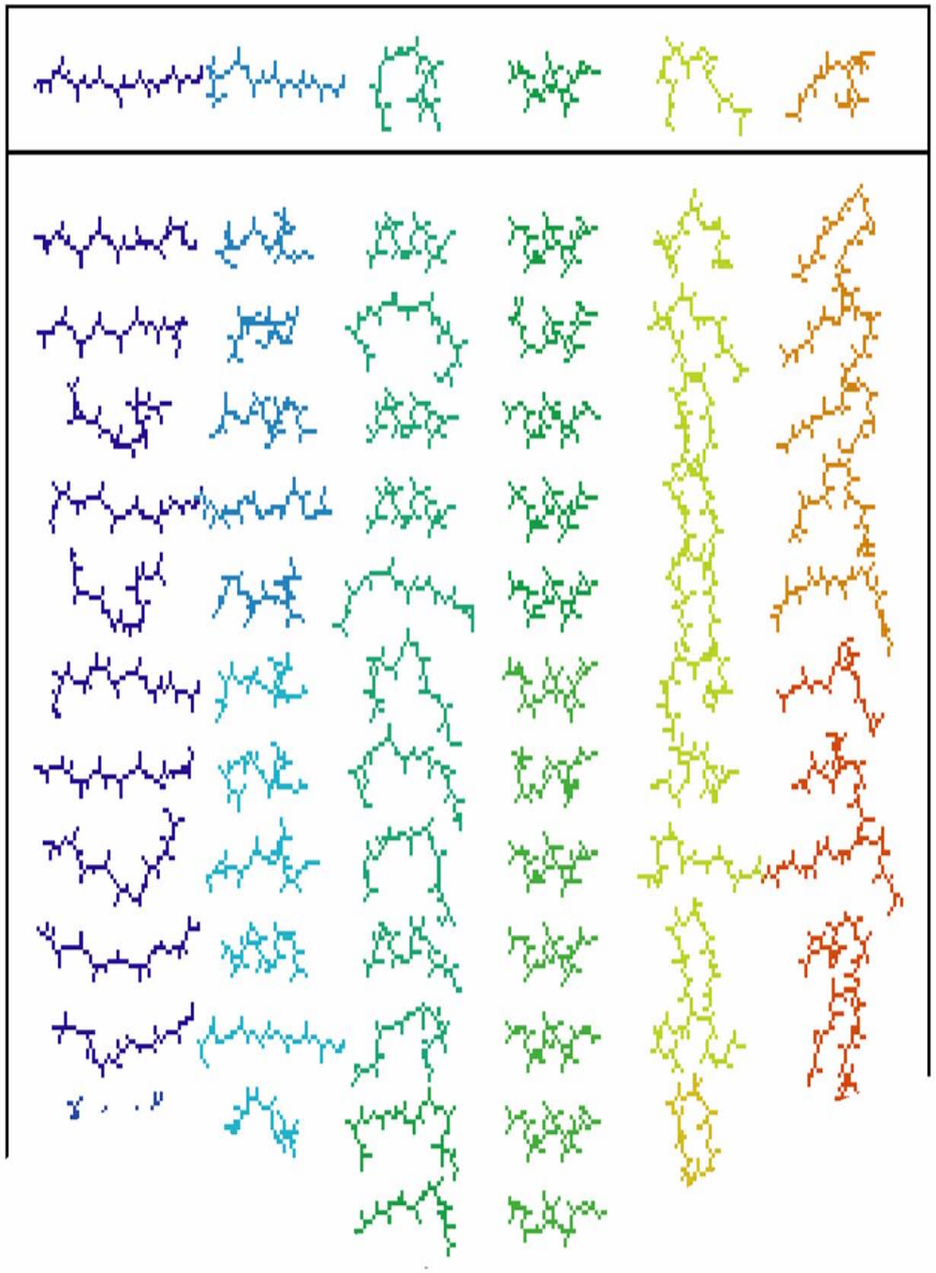
Spherical Coordinates

Assembly of sub-structural units



Build the Fragment Library-Rosetta

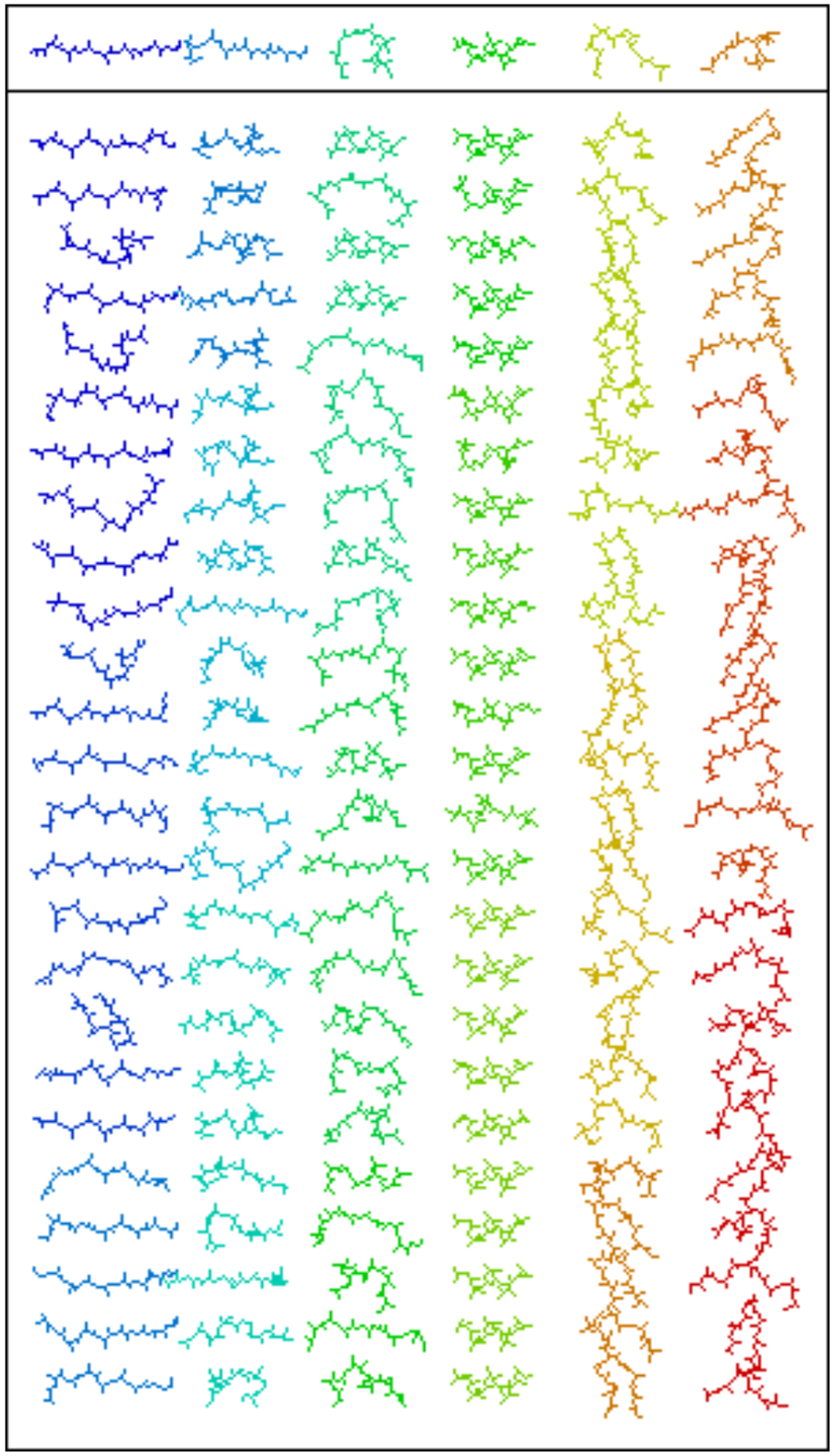
- Extract possible local structures from PDB



Generate the Fragment Library

- Select PDB template
  - Select Sequence Families
  - Each Family has a single known structure (family)
  - Has no more than 25% sequence identity between any two sequence
- Clustering the fragments
  - Generate all the fragments from the selected families

Rosetta Fragment Libraries



- 25-200 fragments for each 3 and 9 residue sequence window
- Selected from database of known structures
  - > 2.5Å resolution
  - < 50% sequence identity
- Ranked by sequence similarity and similarity of predicted and known secondary structure

Scoring Function

- Ideal energy function
  - Has a clear minimum in the native structure.
  - Has a clear path towards the minimum.
  - Global optimization algorithm should find the native structure.

Rosetta MC Energy Function

- Compactness (radius of gyration)
- Hydrophobic burial
- Polar side chain contacts (statistical pairwise potential)
- Hydrogen bonding between beta-strands
- Hard-sphere repulsion (VdW)

Fragment Insertion

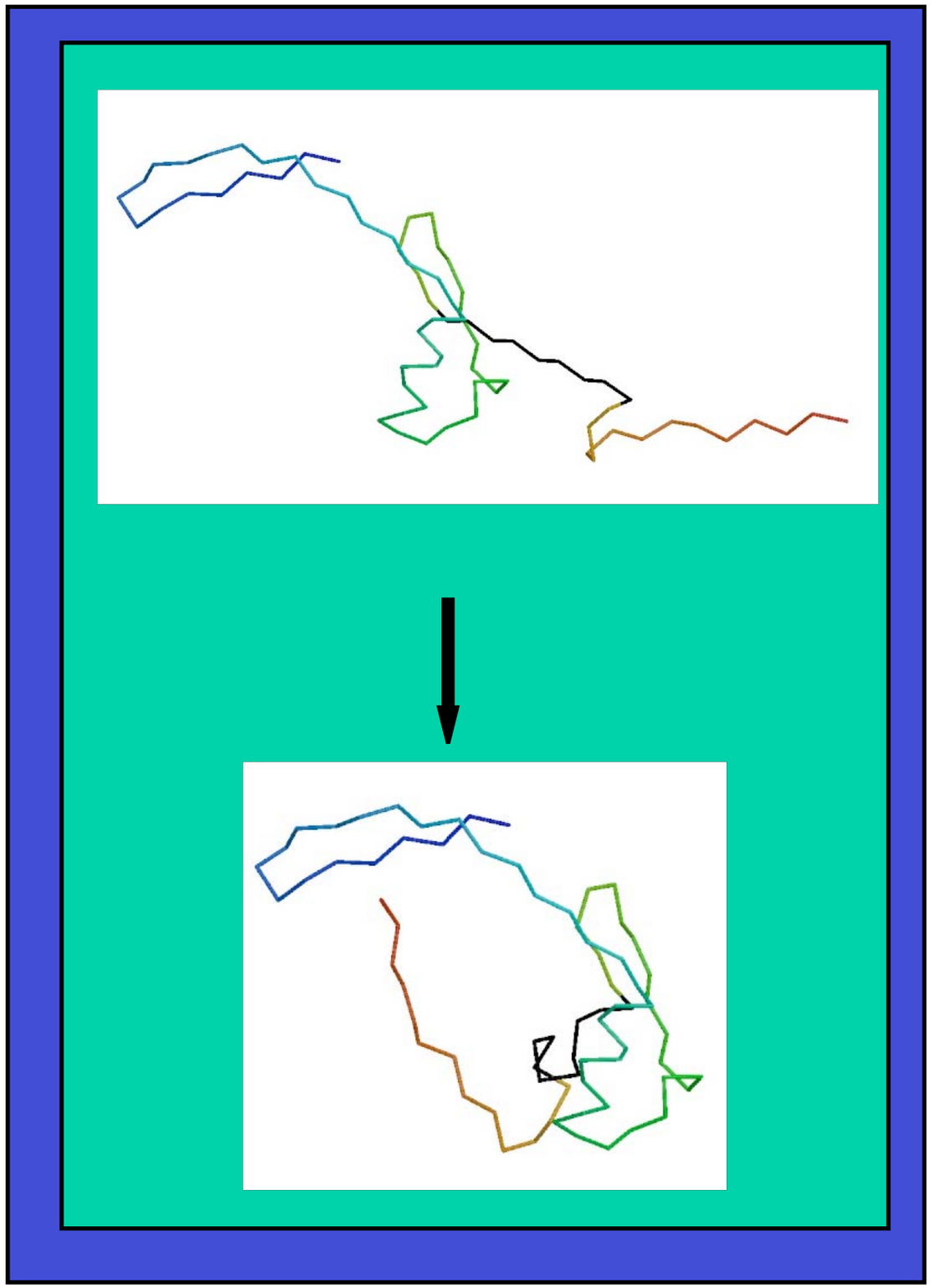
- Finds three and nine residue fragments from known library and replaces unknown torsion angles with the ‘known’ ones
- Scores all windows of three and nine residues
- Create fragment list with the 200 best three residue and 200 best nine residue fragments

Fragment Assembly

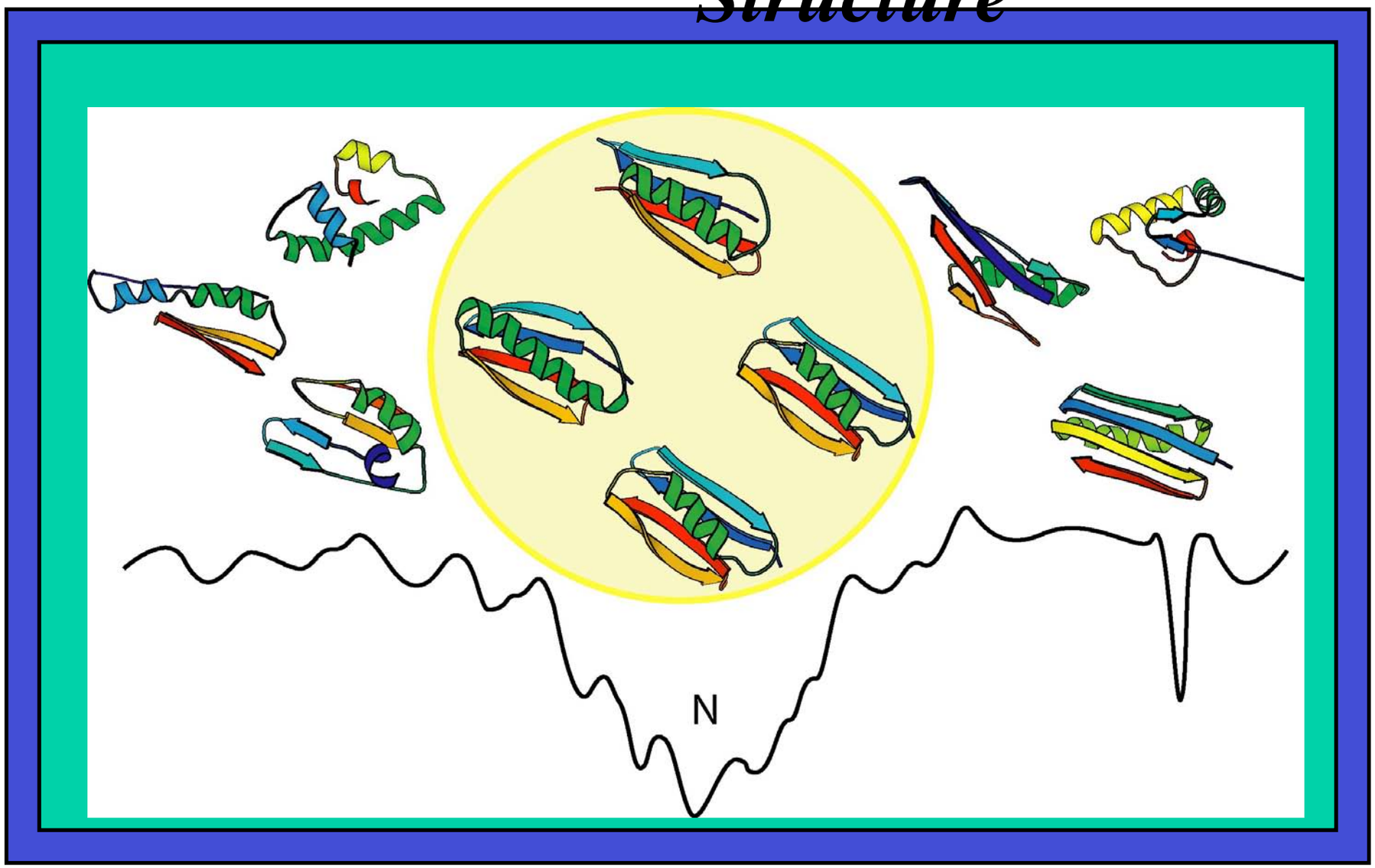
- Randomly choose a nine residue fragment from the top 25 fragments in the ranked list
  - Score this replacement, negatives are kept
- Each simulation chooses a different random start and attempts 28,000 nine residue insertions
- Next 8,000 attempted three residue insertions are scored with the overall structure

Rosetta Potential Function

- Derived from Bayesian treatment of residue distributions in known protein structures
- Reduced representation of protein used; one centroid per sidechain
- Potential Terms:
  - environment (solvation)
  - pairwise interactions (electostatics)
  - strand pairing
  - radius of gyration
  - Cβ density
  - steric overlap



Decoy Discrimination: Identifying the Best Structure



- 1000-100,000 short simulations to generate a population of 'decoys'
- Filter population to correct systematic biases
- Fullatom potential functions to select the deepest energy minimum
- Cluster analysis to select the broadest minimum
- Structure-structure matches to database of known structures

Rosetta: clustering the models

- Compare models to each other with RMSD
- Models can come from different family members
- Cutoff varied to give 80-100 members in largest cluster
- The largest clusters are assumed to contain the best structures (attractors in folding space...?)

Rosetta: Filtering the models

- Between 6,000 and 150,000 models generated
- Contact Order
- Generated models are biased towards simple structures
- Filter models to give correct contact order distribution for domains of that size/composition
- Sheet filter
- Add side chains, calculate atomic physical potential (to eliminate poorly packed structures)

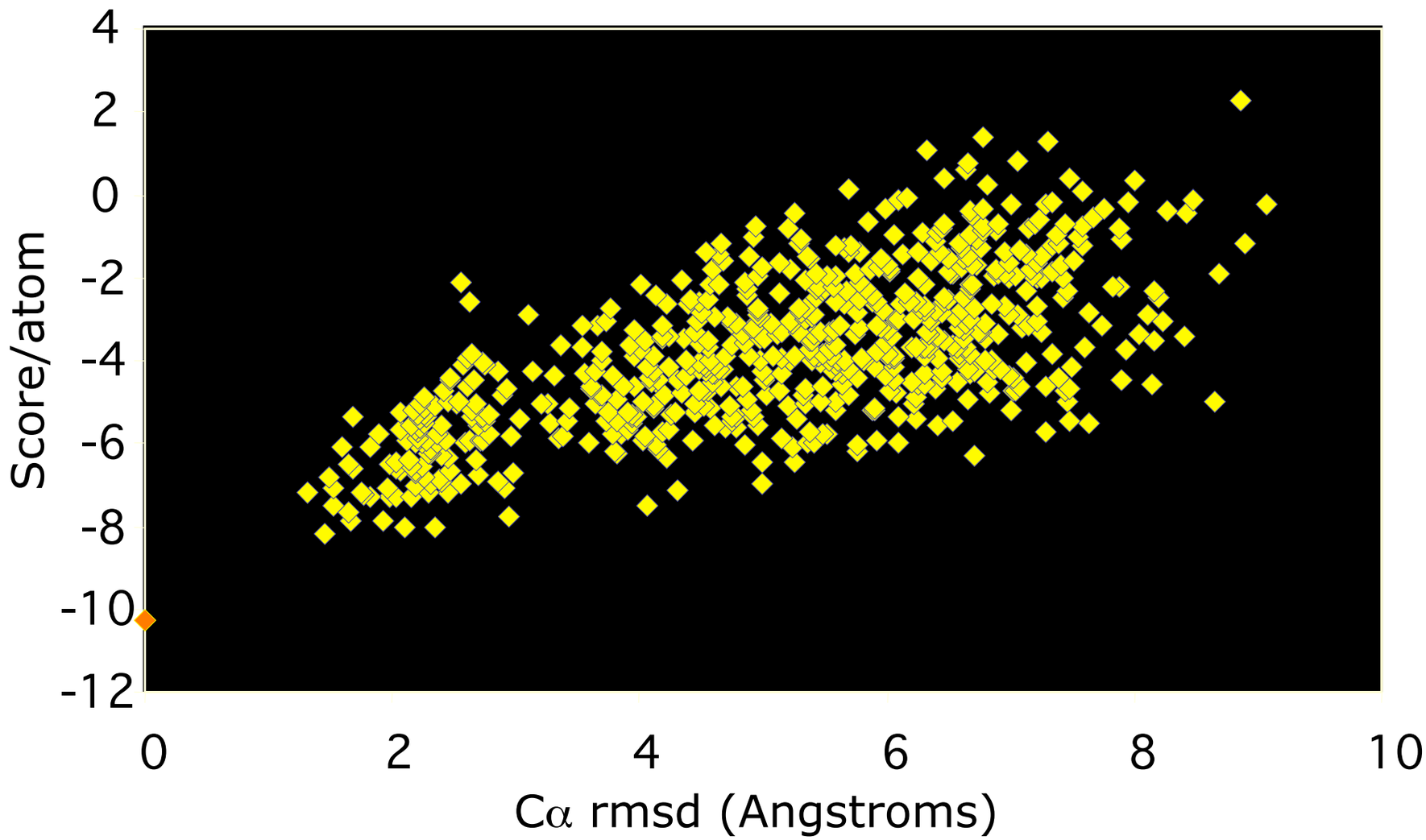
Monte Carlo optimisation

1. Initial configuration (random or extended)
2. Make a randomised MOVE on configuration
3. Measure change in quality of structure (DE)
4. IF better () ACCEPT MOVE
5. ELSIF rand ACCEPT MOVE
6. ELSE REJECT MOVE

- GO TO 2. (reduce T if you like)

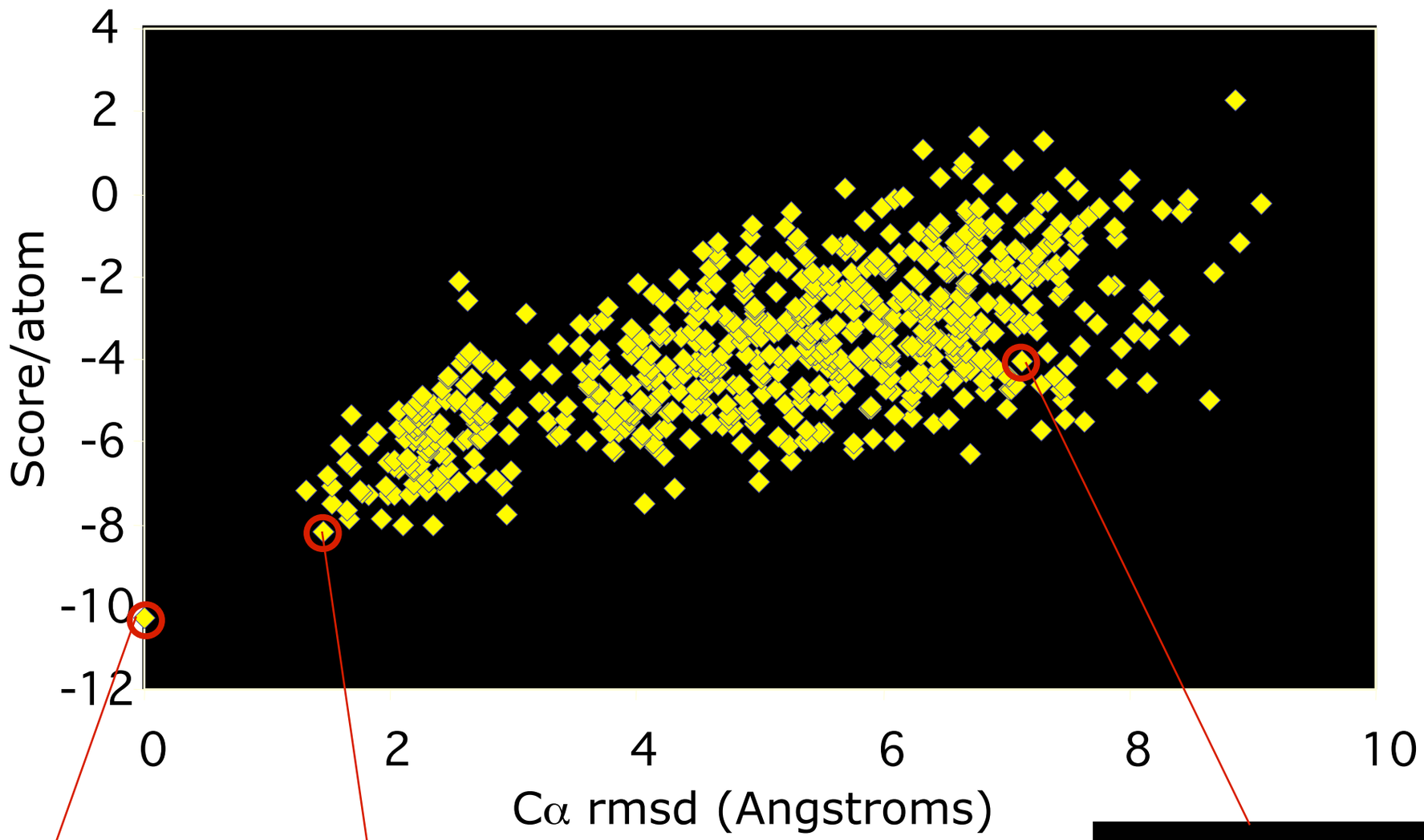
Testing of scoring functions

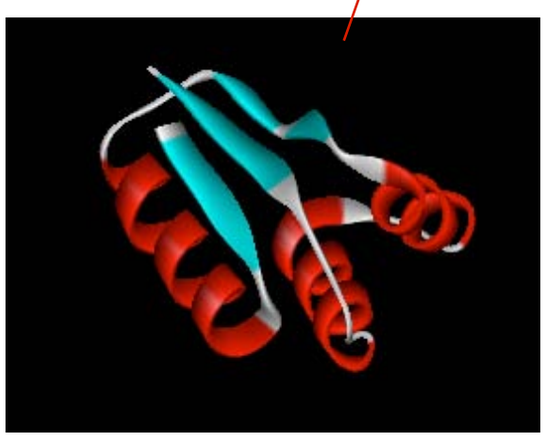
Contact scores for 1ctf decoy set (4state decoys)

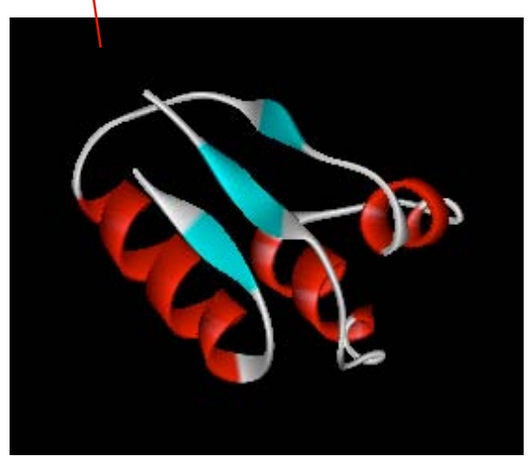


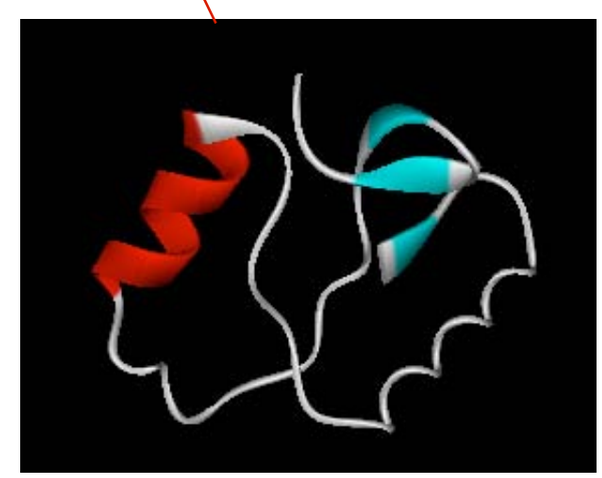
Testing of scoring functions

Contact scores for 1ctf decoy set (4state decoys)

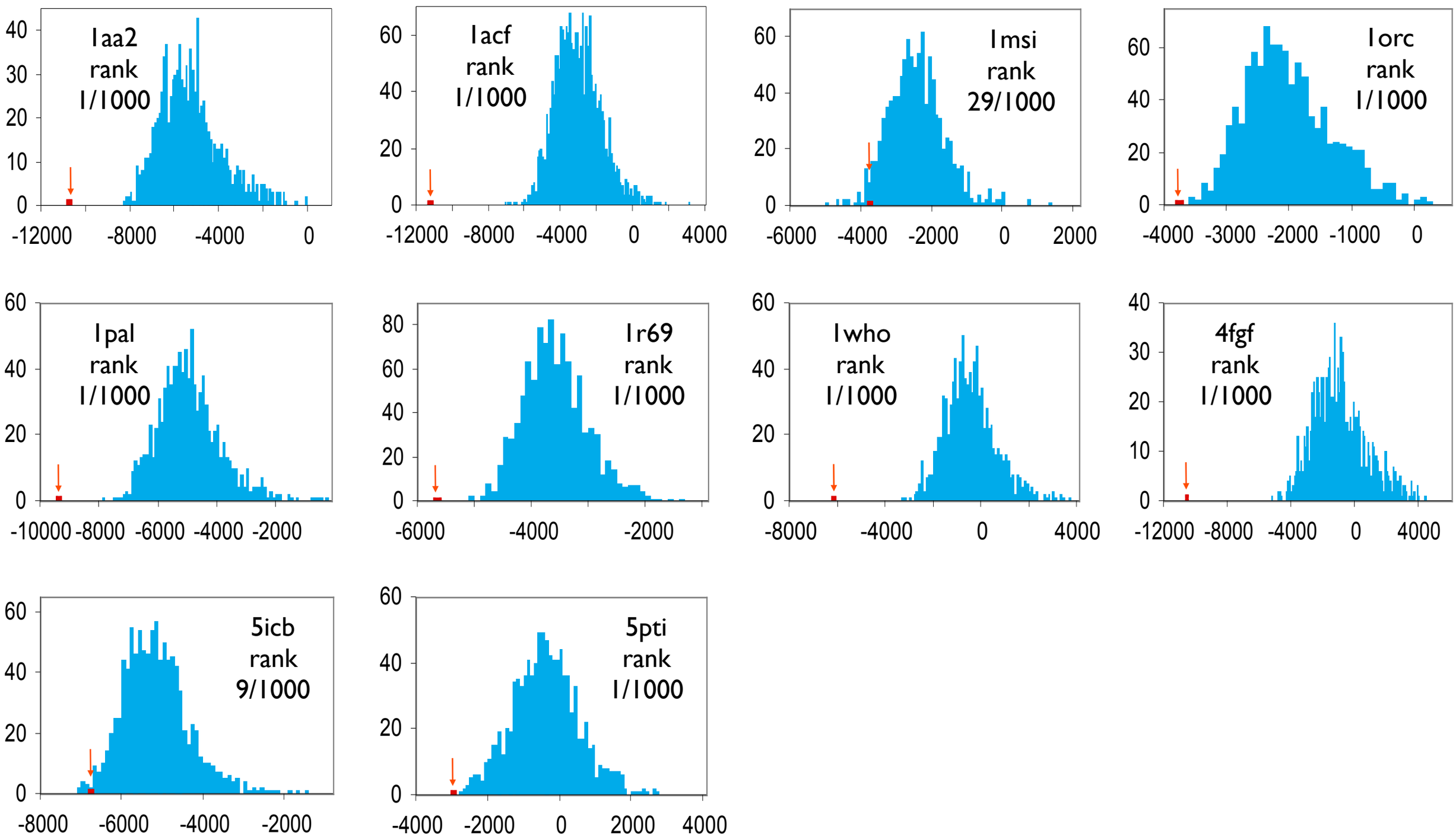




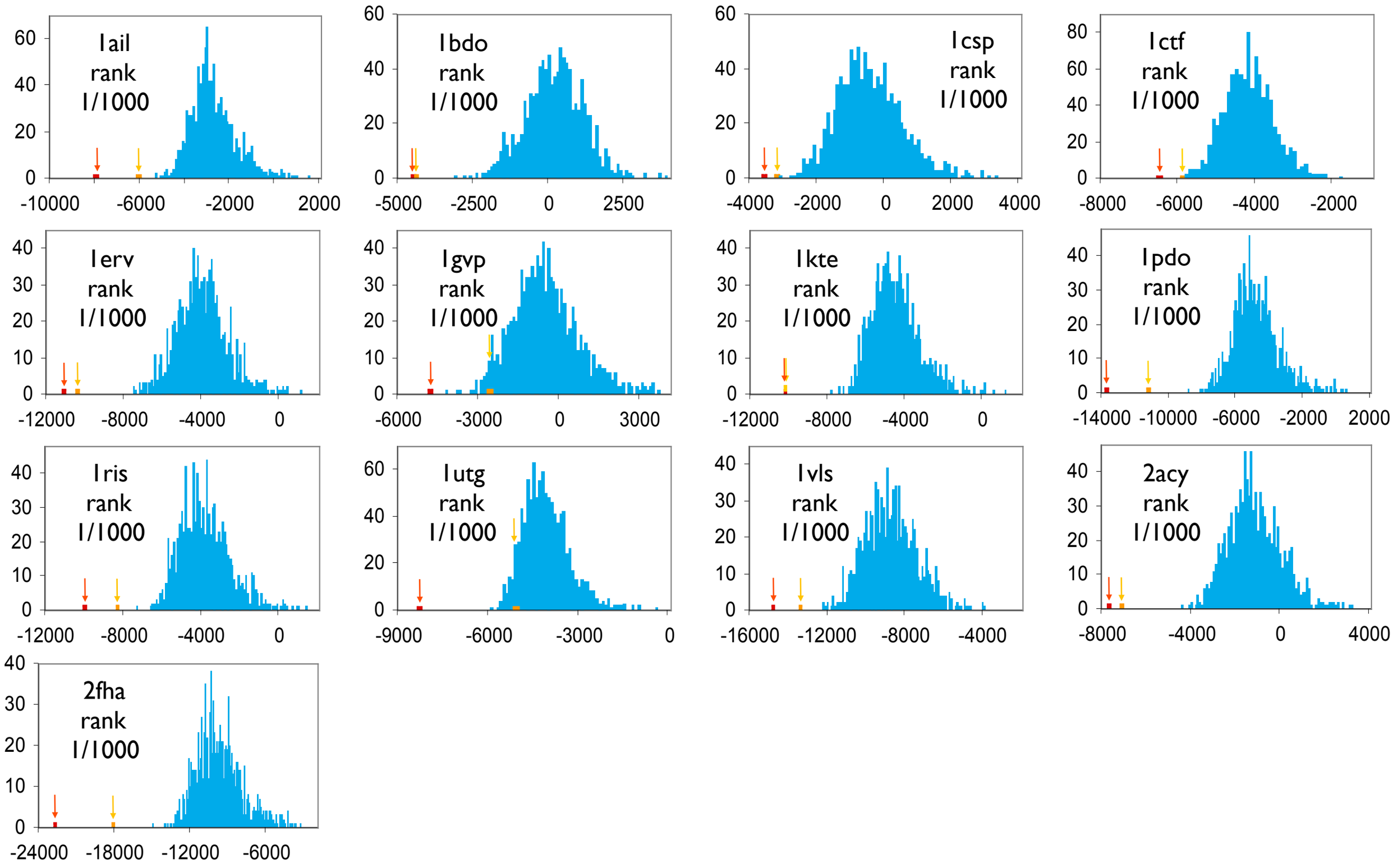




Histograms of native (red) and decoy (blue) scores for the Rosetta decoy monomers

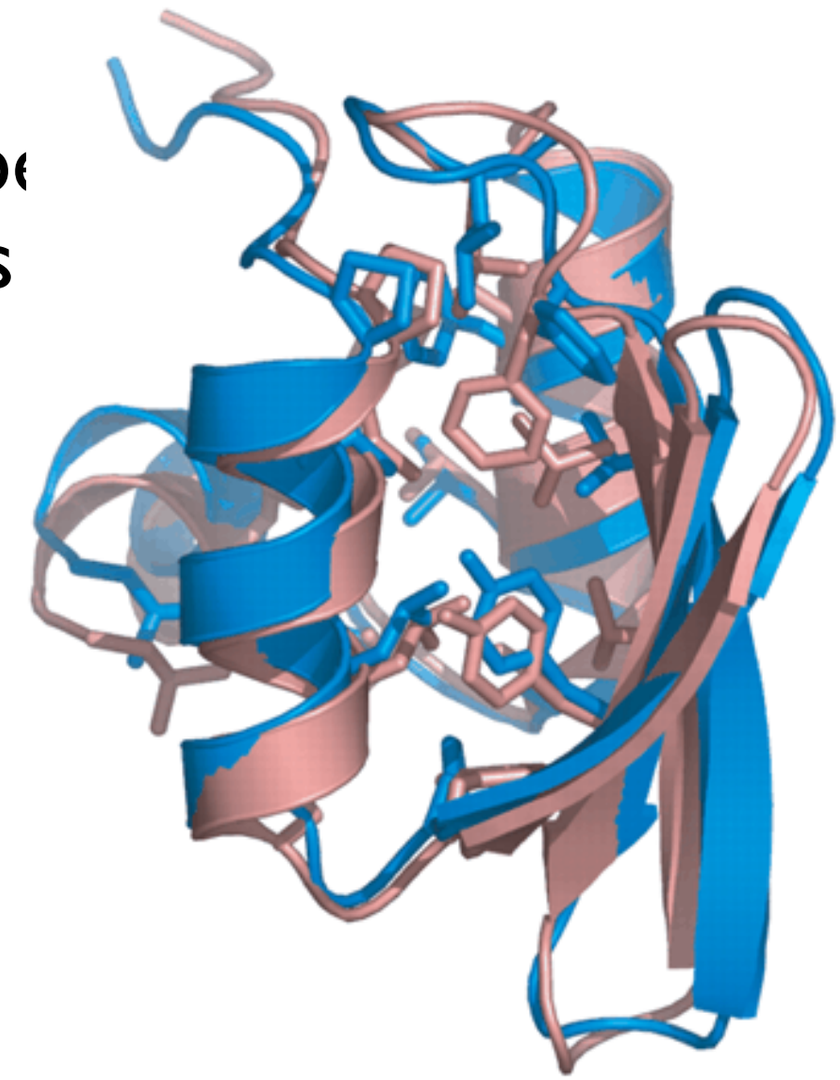


Histograms of native (red) and decoy (blue) scores for the Rosetta decoy oligomers

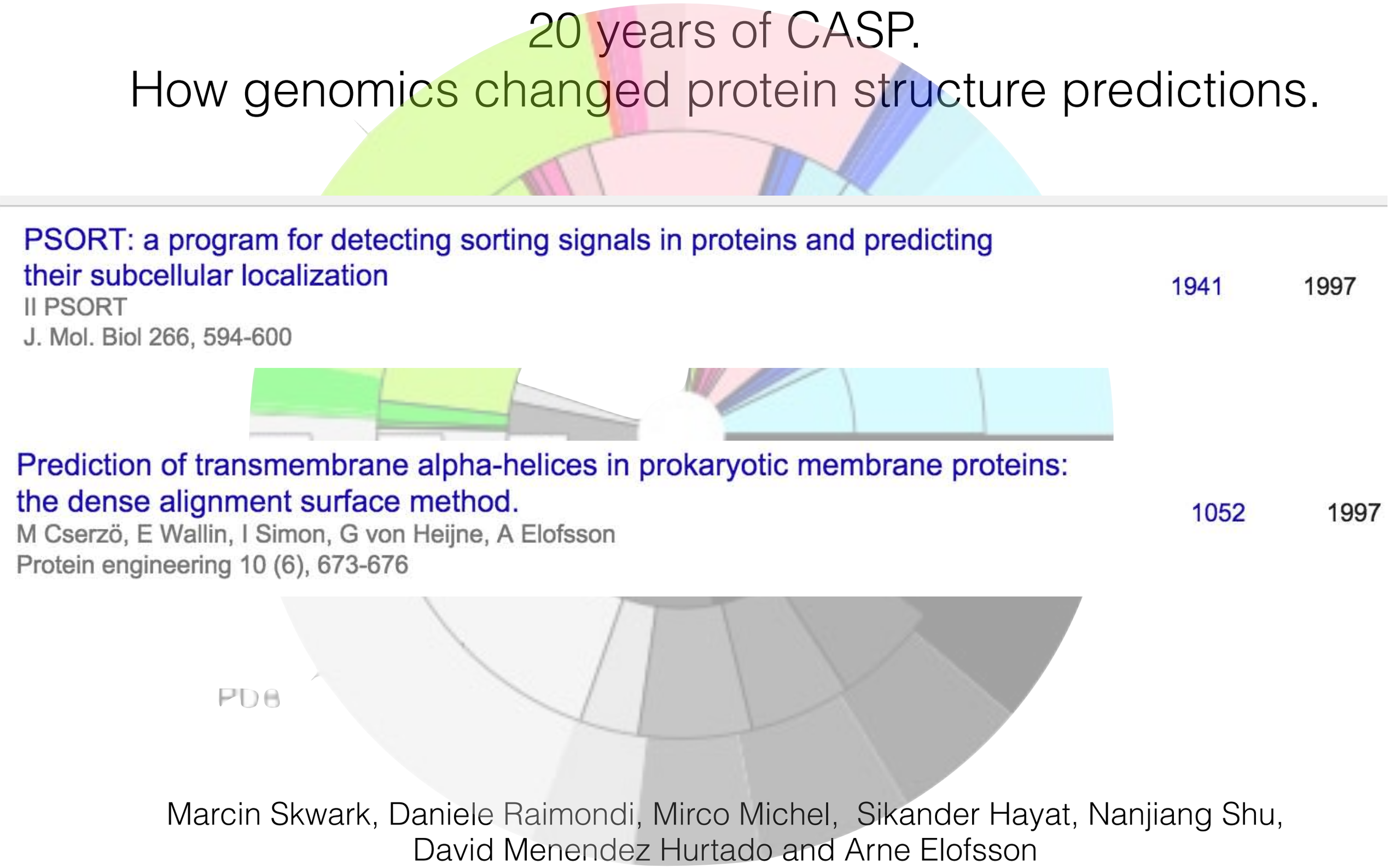


HR protocol to Rosetta

- Additional refinement step from be clusters using all atom refinements
  1. Make small dihedral changes
  2. Rebuild sidechains
  3. Minimize (in dihedral space)
  4. Evaluate energy
  5. Go To 1
- 5 out 16 small proteins < 1.5 Å

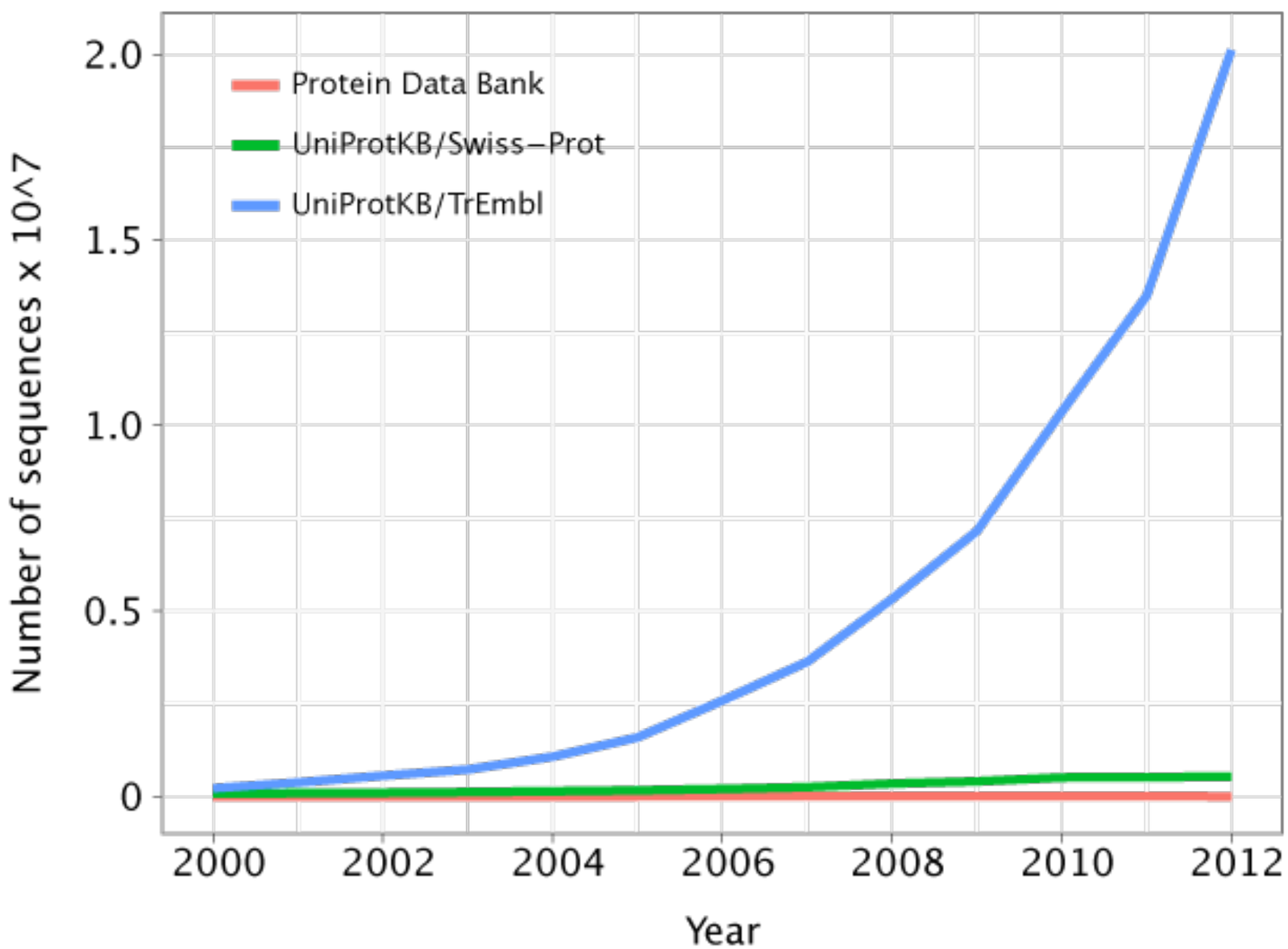


20 years of CASP.
How genomics changed protein structure predictions.

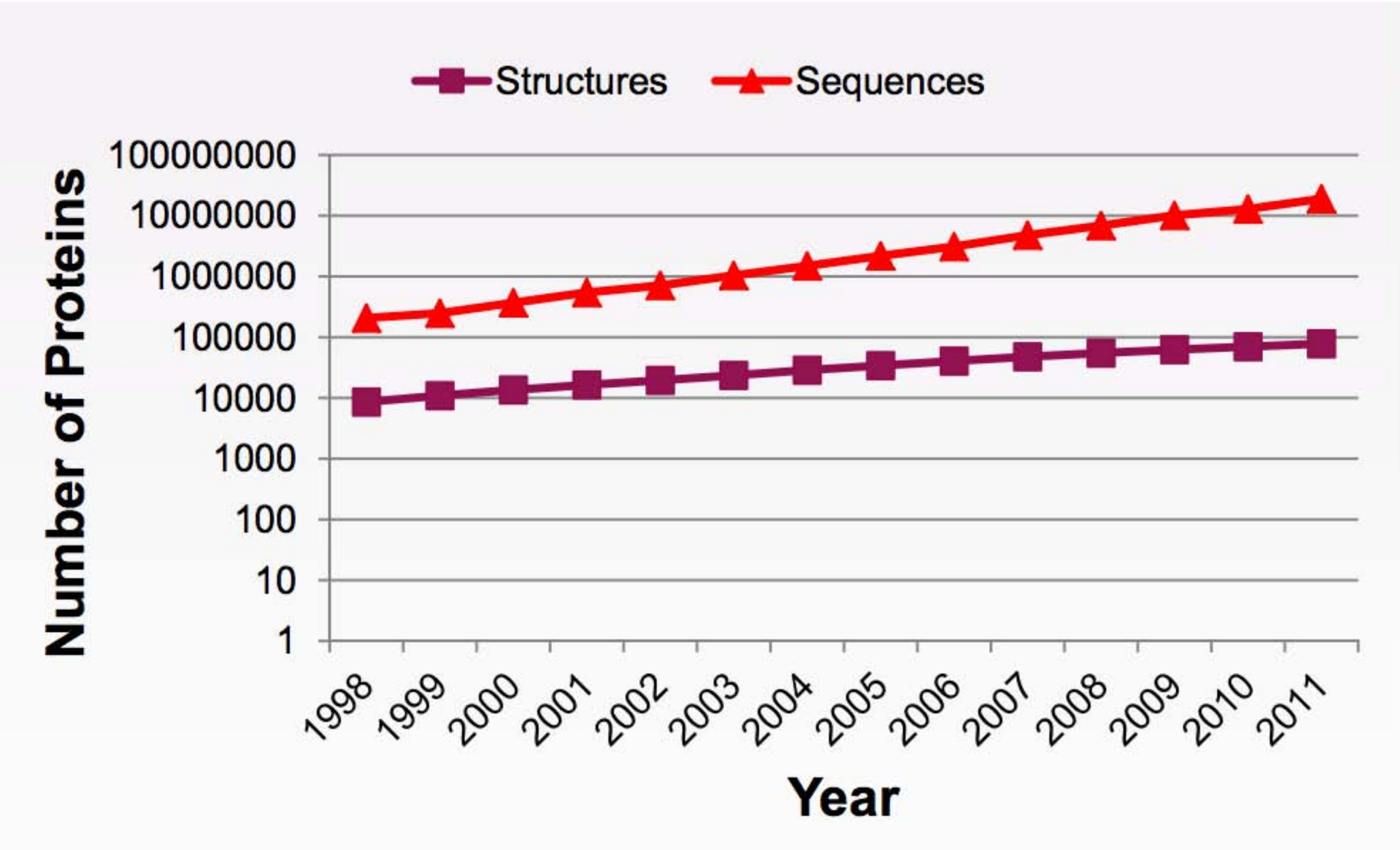


Marcin Skwark, Daniele Raimondi, Mirco Michel, Sikander Hayat, Nanjiang Shu,
David Menendez Hurtado and Arne Elofsson
Stockholm University

Number of protein sequences and structures is increasing.

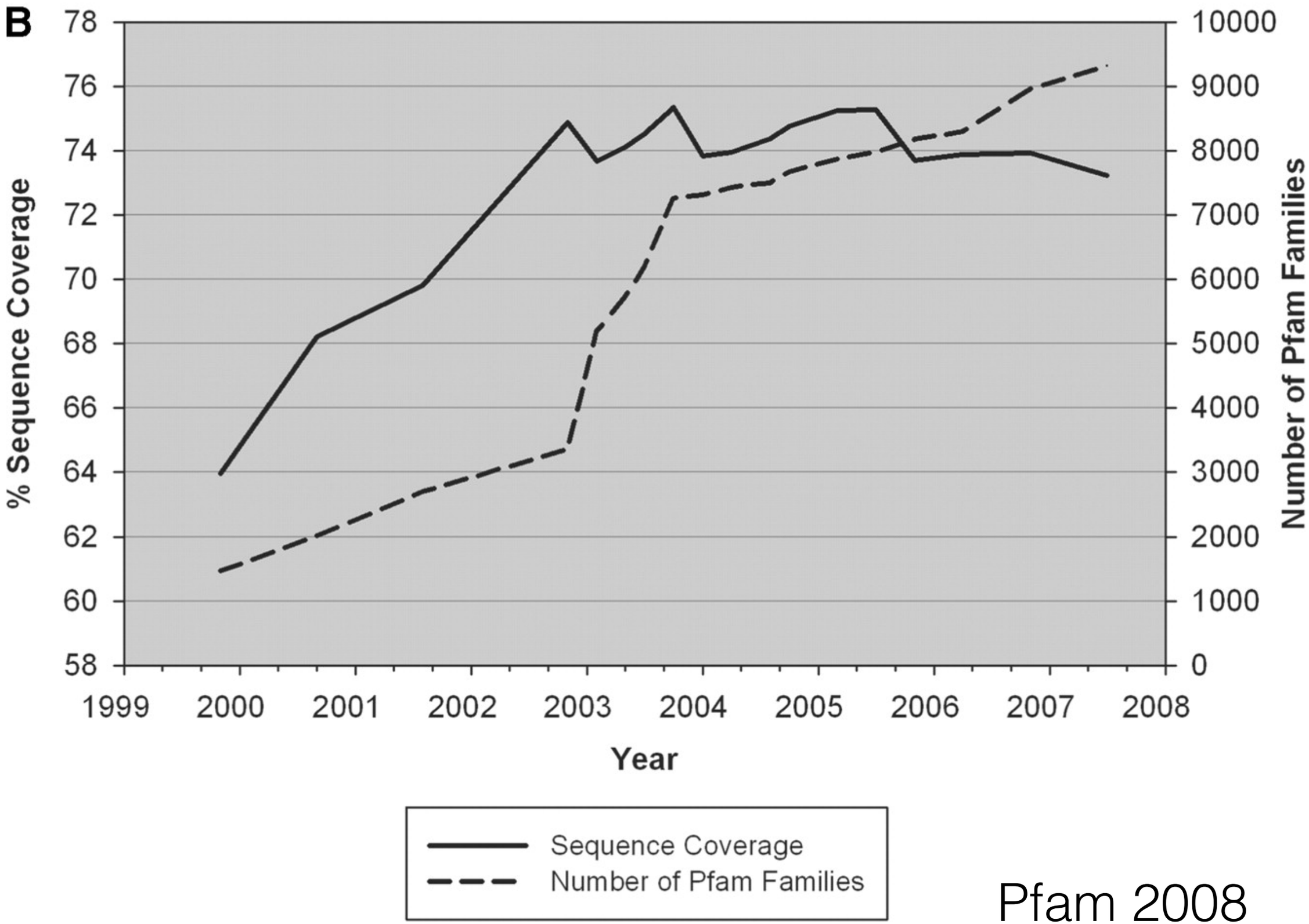


Exponential increase



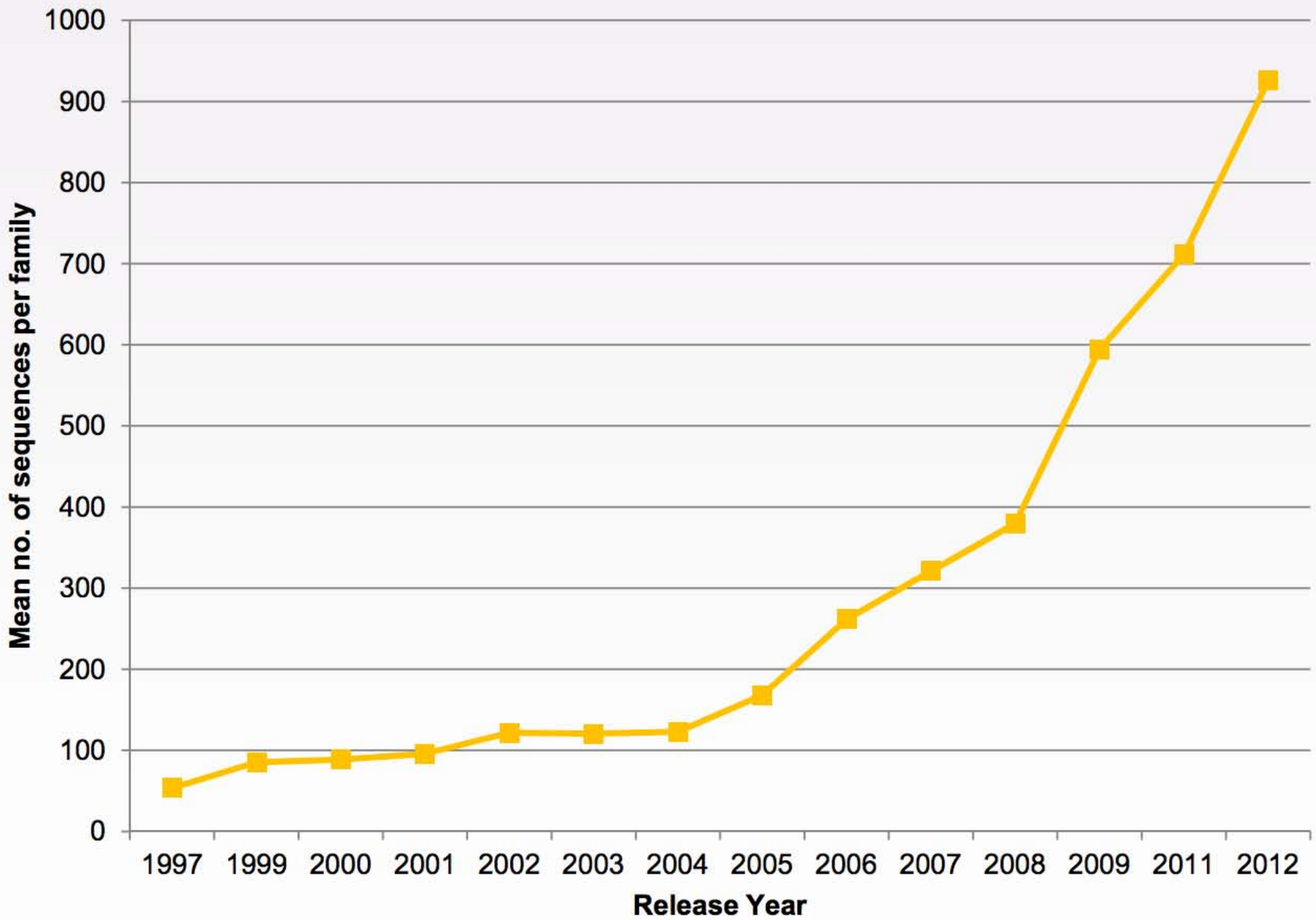
David Jones 2012

Protein Families are increasing only slowly and covers ~75% of all sequences



Pfam 2008

Sequence families are getting bigger and bigger



Pfam 2012

The late revolution - Has contact Predictions solved the protein folding problem?

Protein 3D structure computed from evolutionary sequence variation.
Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, **Sander** C.
PLoS One. 2011;6(12):e28766. doi: 10.1371/journal.pone.0028766. Epub 2011 Dec 7.
PMID: 22163331 **Free PMC Article**

Identification of direct residue contacts in protein-protein interaction by message passing.
Weigt M, White RA, Szurmant H, Hoch JA, **Hwa** T.
Proc Natl Acad Sci U S A. 2009 Jan 6;106(1):67-72. doi: 10.1073/pnas.0805923106. Epub 2008 Dec 30.
PMID: 19116270 **Free PMC Article**
Similar articles

Disentangling direct from indirect co-evolution of residues in protein alignments.
36. Burger L, van **Nimwegen E**.
PLoS Comput Biol. 2010 Jan;6(1):e1000633. doi: 10.1371/journal.pcbi.1000633. Epub 2010 Jan 1.
PMID: 20052271 **Free PMC Article**
Similar articles

Superadditive correlation.
Giraud BG, Heumann JM, Lapedes AS.
Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. **1999** May;59(5 Pt A):4983-91.
PMID: 11969452
Similar articles
Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary

Direct-coupling analysis of residue coevolution captures native contacts across many protein families.
Morcos F, Pagnani A, Lunt B, Bertolino A, **Marks** DS, Sander C, Zecchina R, Onuchic JN, Hwa T, **Weigt** M.
Proc Natl Acad Sci U S A. 2011 Dec 6;108(49):E1293-301. doi: 10.1073/pnas.1111471108. Epub 2011 Nov 21.
PMID: 22106262 **Free PMC Article**

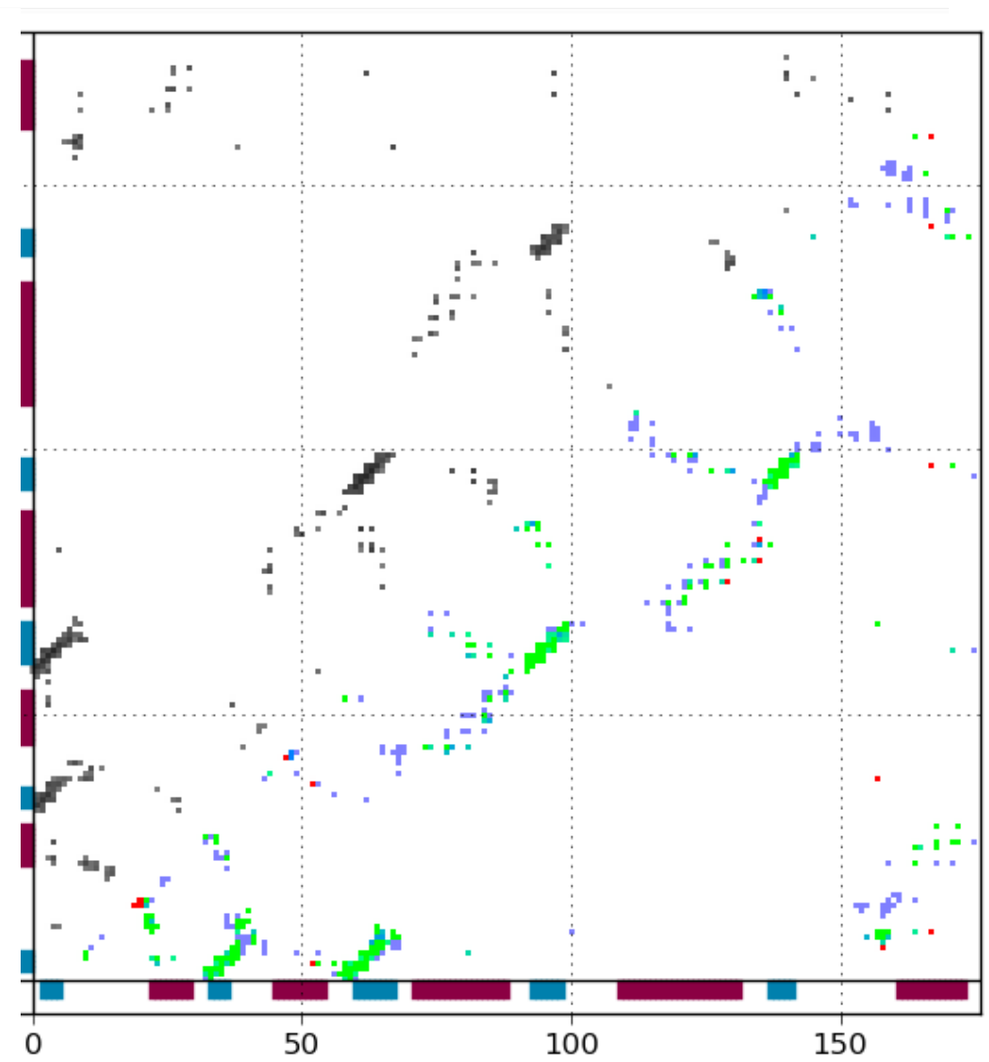
Three-dimensional structures of membrane proteins from genomic sequencing.
Hopf TA, Colwell LJ, Sheridan R, Rost B, **Sander** C, **Marks** DS.
Cell. 2012 Jun 22;149(7):1607-21. doi: 10.1016/j.cell.2012.04.012. Epub 2012 May 10.
PMID: 22579045 **Free PMC Article**

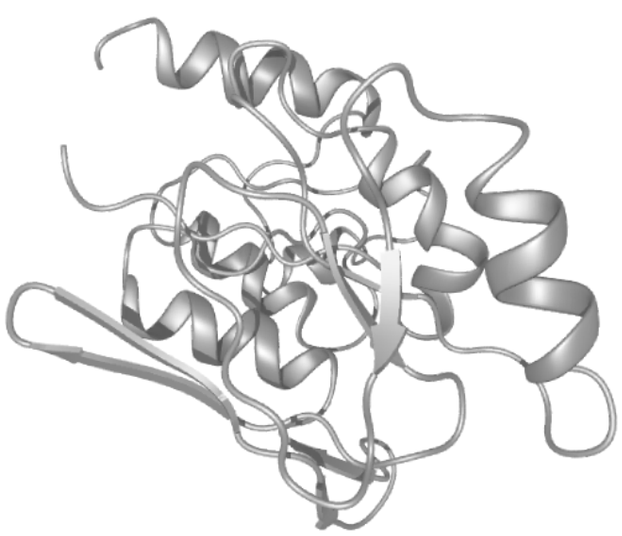
Contact based structure prediction - the Revolution occuring finaly.

Contact map

```
LLSSEKLIAIGASTGGTEAIRH
VLQPLPLSSPAVIITQHMPPGF
TRSFAERLNKLCQISVKEAEDG
ERVLPGHAYIAPGDKHMELARS
GANYQIKIHDGPPVNRHRPSVD
VLFHSVAKHAGRNAVGVILTGM
GNDGAAGMLAMYQAGAWTIAQN
EASCVVFGMPREAINMGGVSEV
VDLSQVSQQMLAKISAGQAIRI
```

Contact Prediction

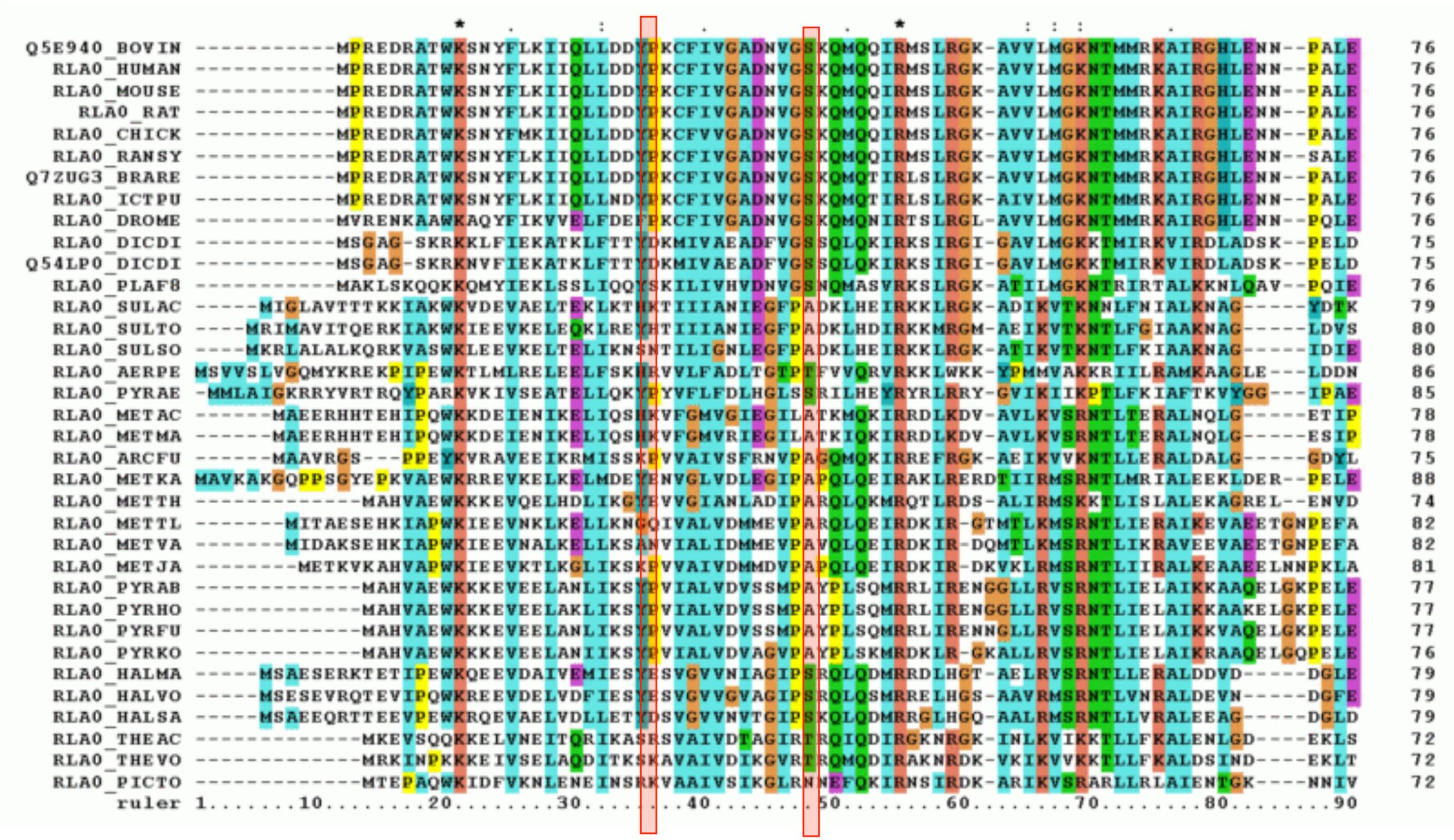




Folding

Mirco Michel

Basis for Contact prediction

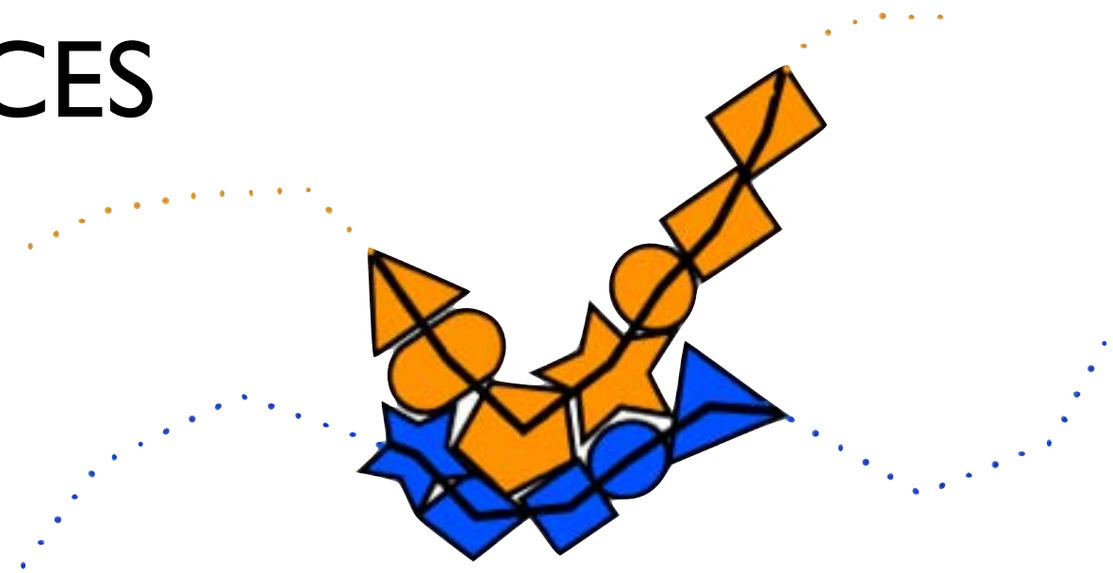


Residues in contact tend to co-evolve

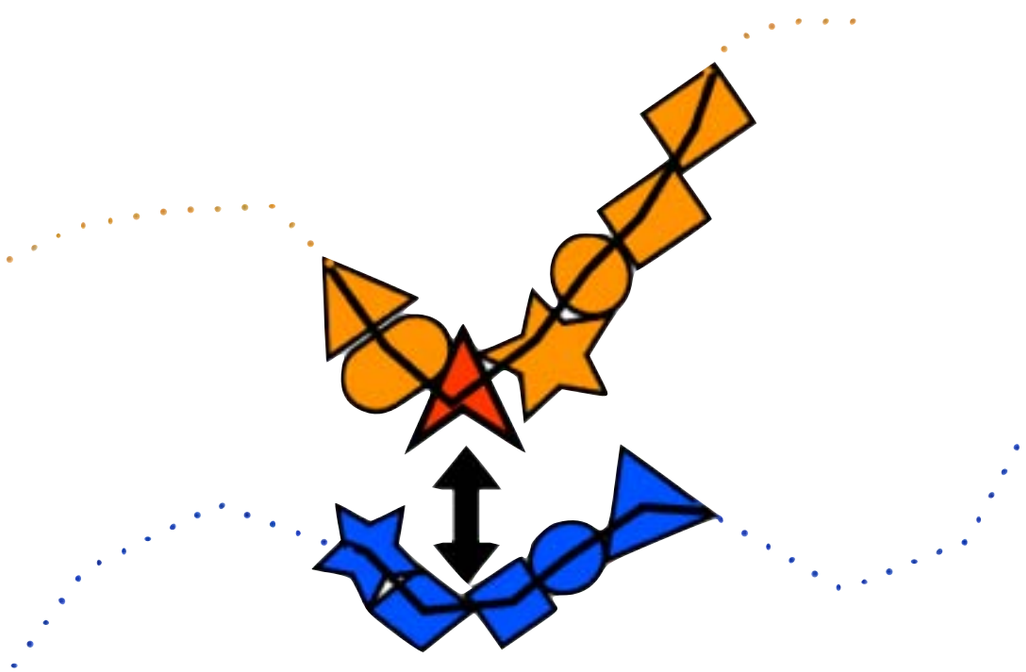
Decoupling direct interactions
from indirect ones
Giraud, Phys Rev E Stat 1999

SPATIAL PROXIMITY INDUCES SEQUENCE COUPLING

Native interactions



Unfavourable mutation



Compensating mutation

Indirect couplings

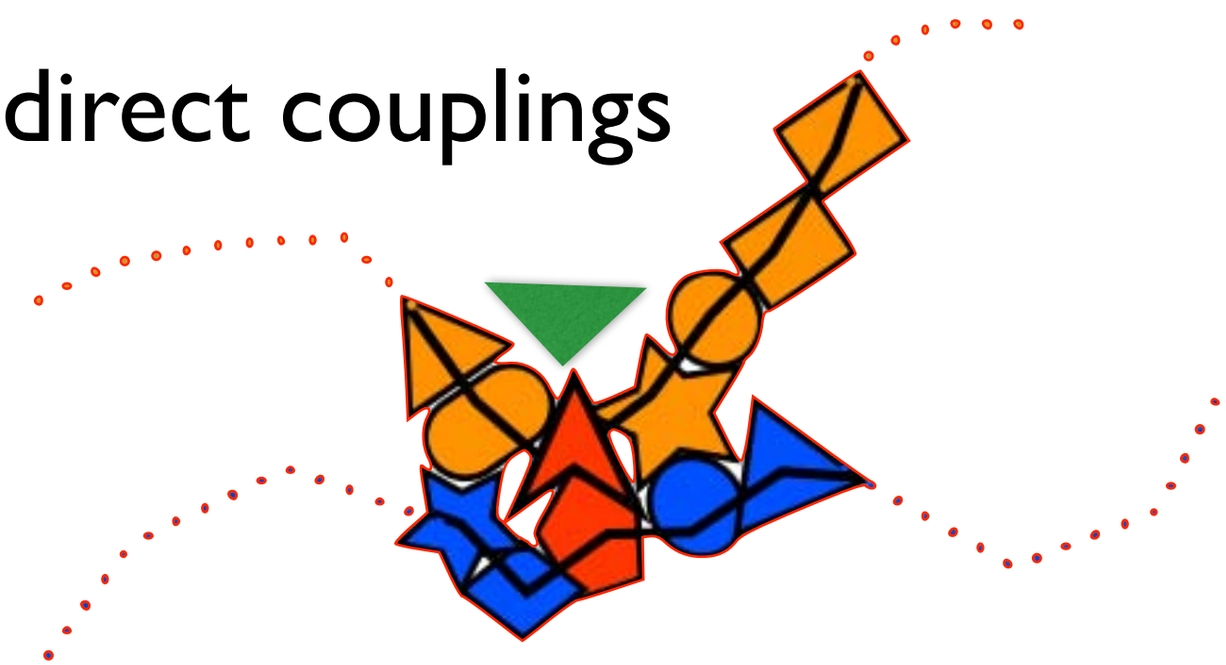
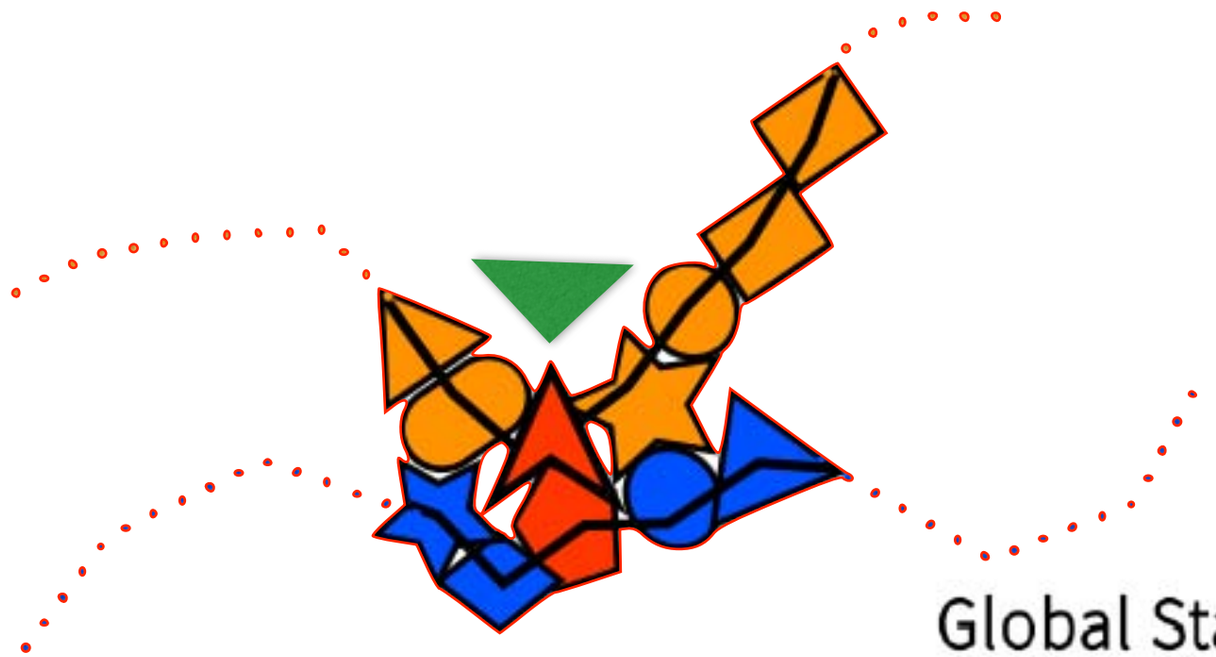
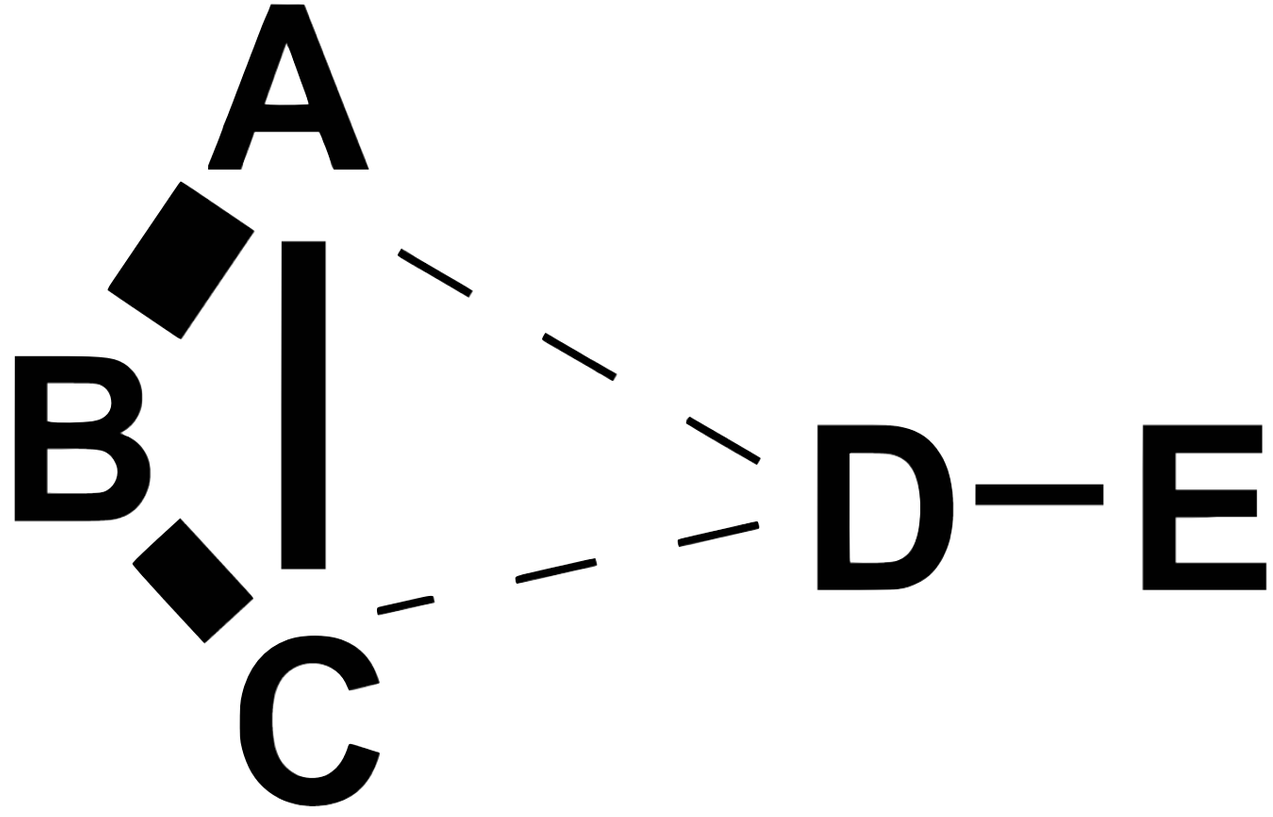


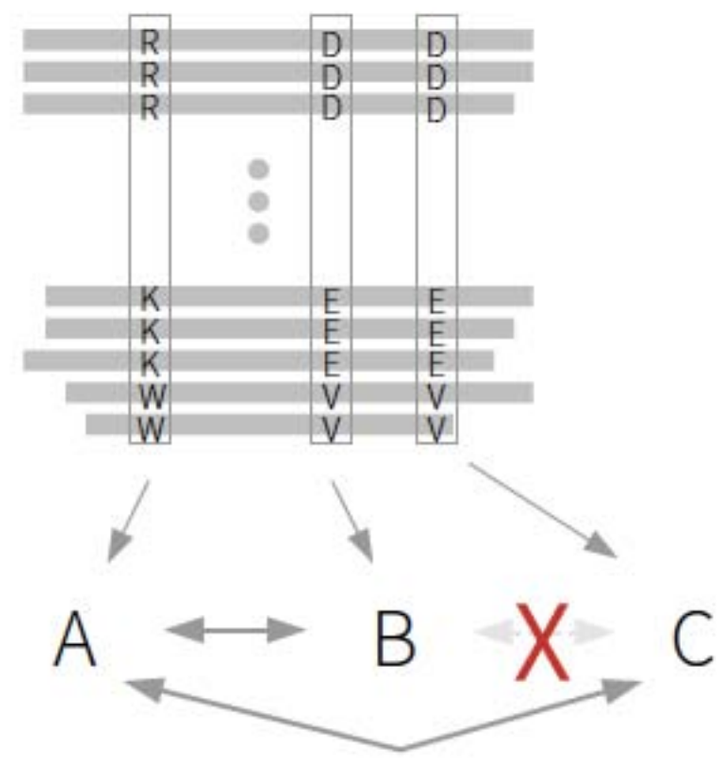
Image source: Andrea Pagnani, International School on Physics of Complex Systems 2012

Key idea in “new” contact predictors





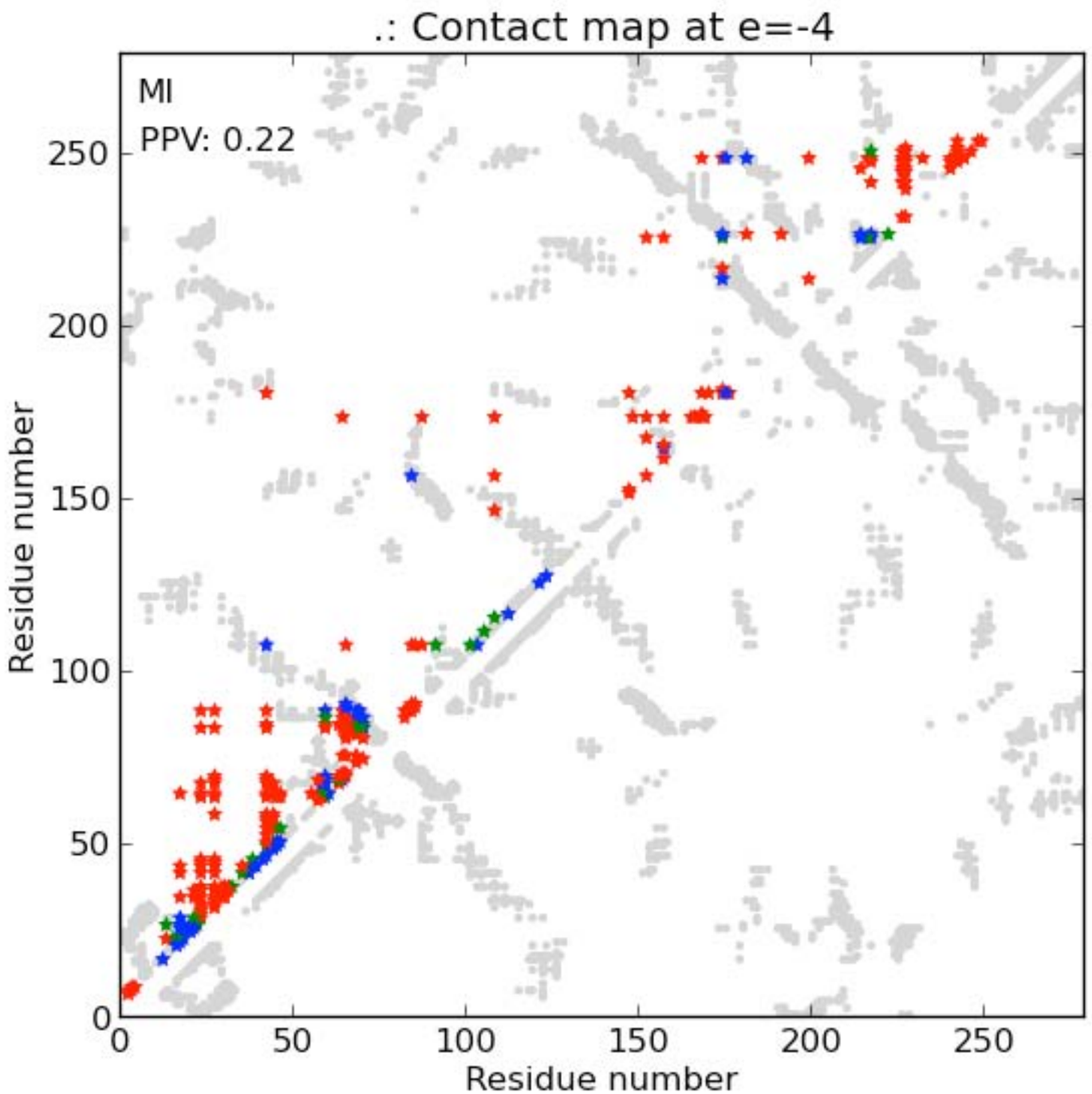
Global Statistical Model

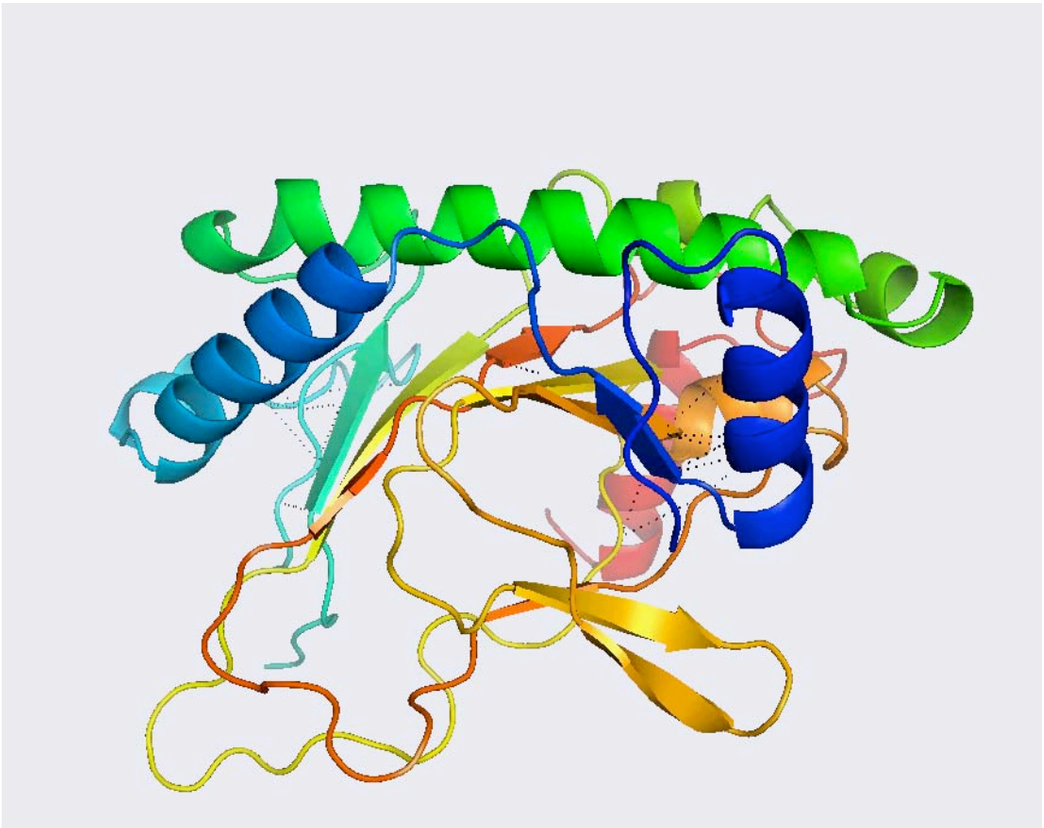


Decoupling direct interactions
from indirect ones
Giraud, Phys Rev E Stat 1999

Which is the best contact prediction method ?

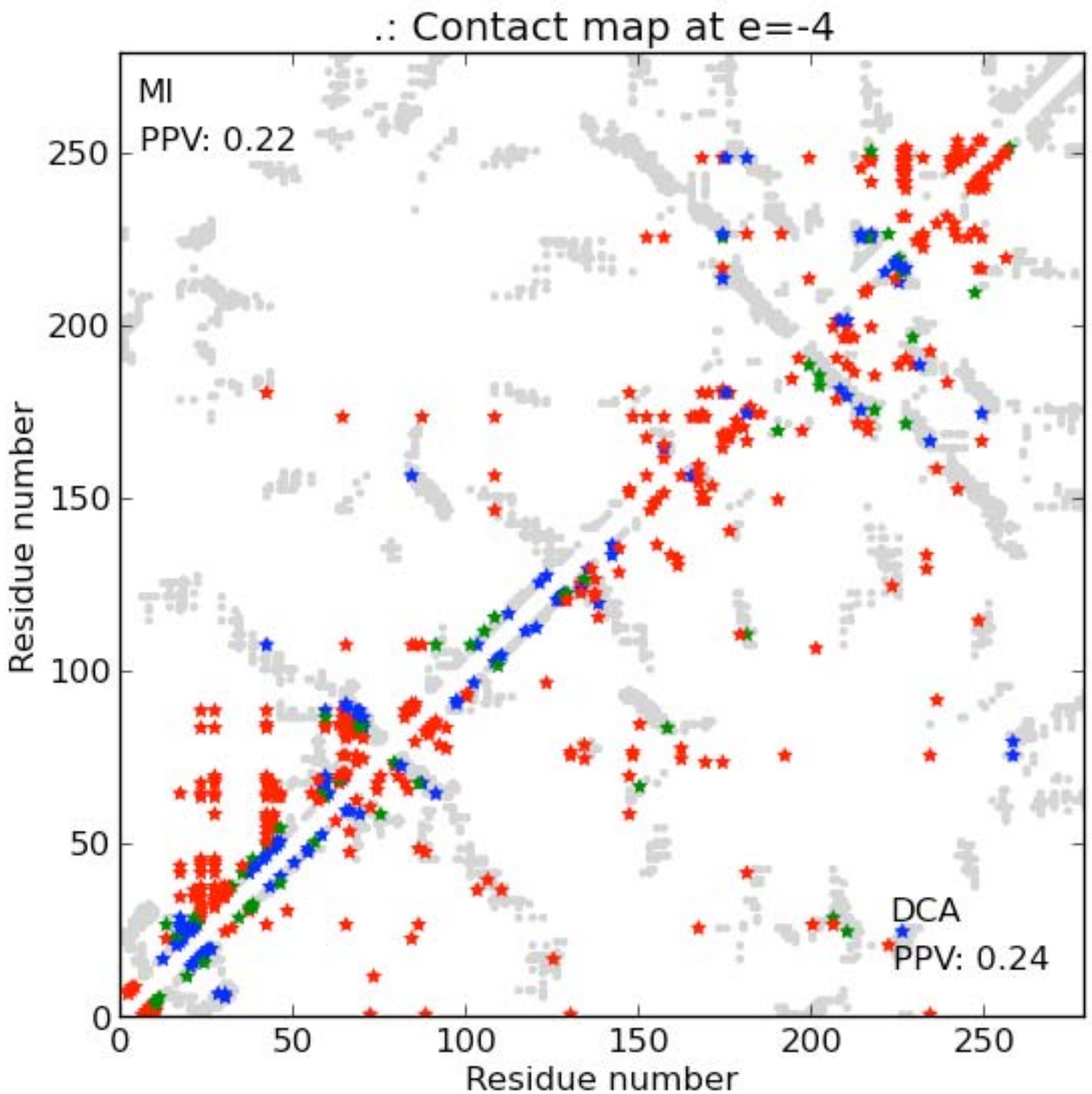
Mutual Information

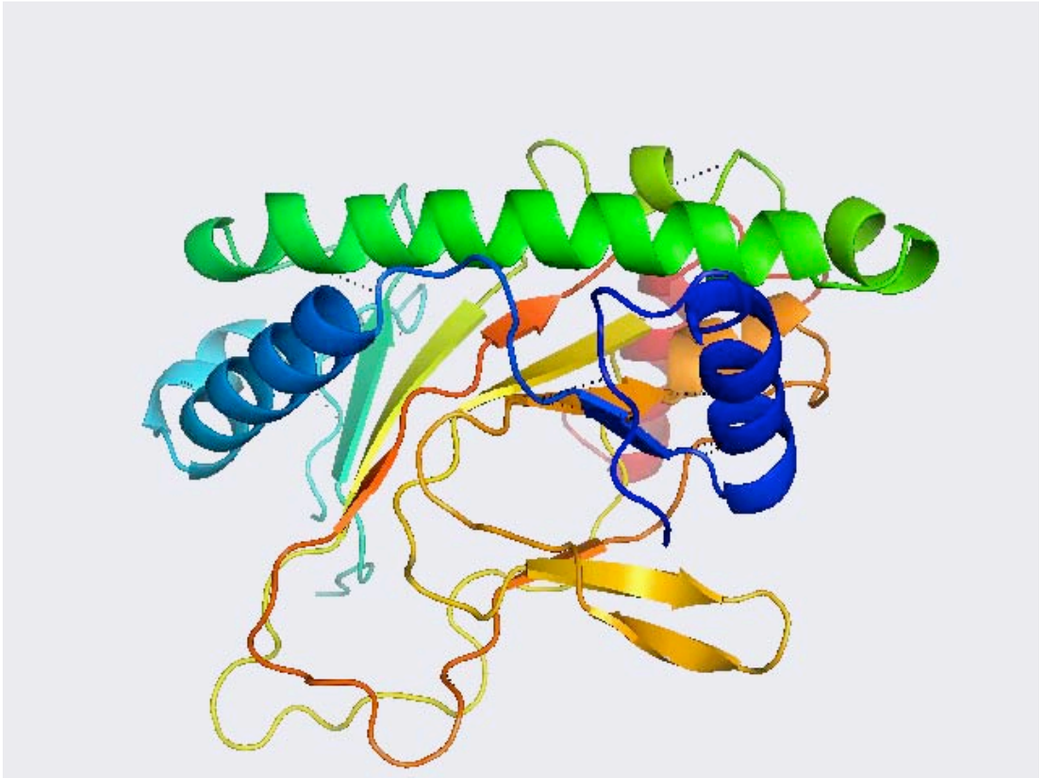




Fodor, A.A. and Aldrich, R.W.
Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. (2004).

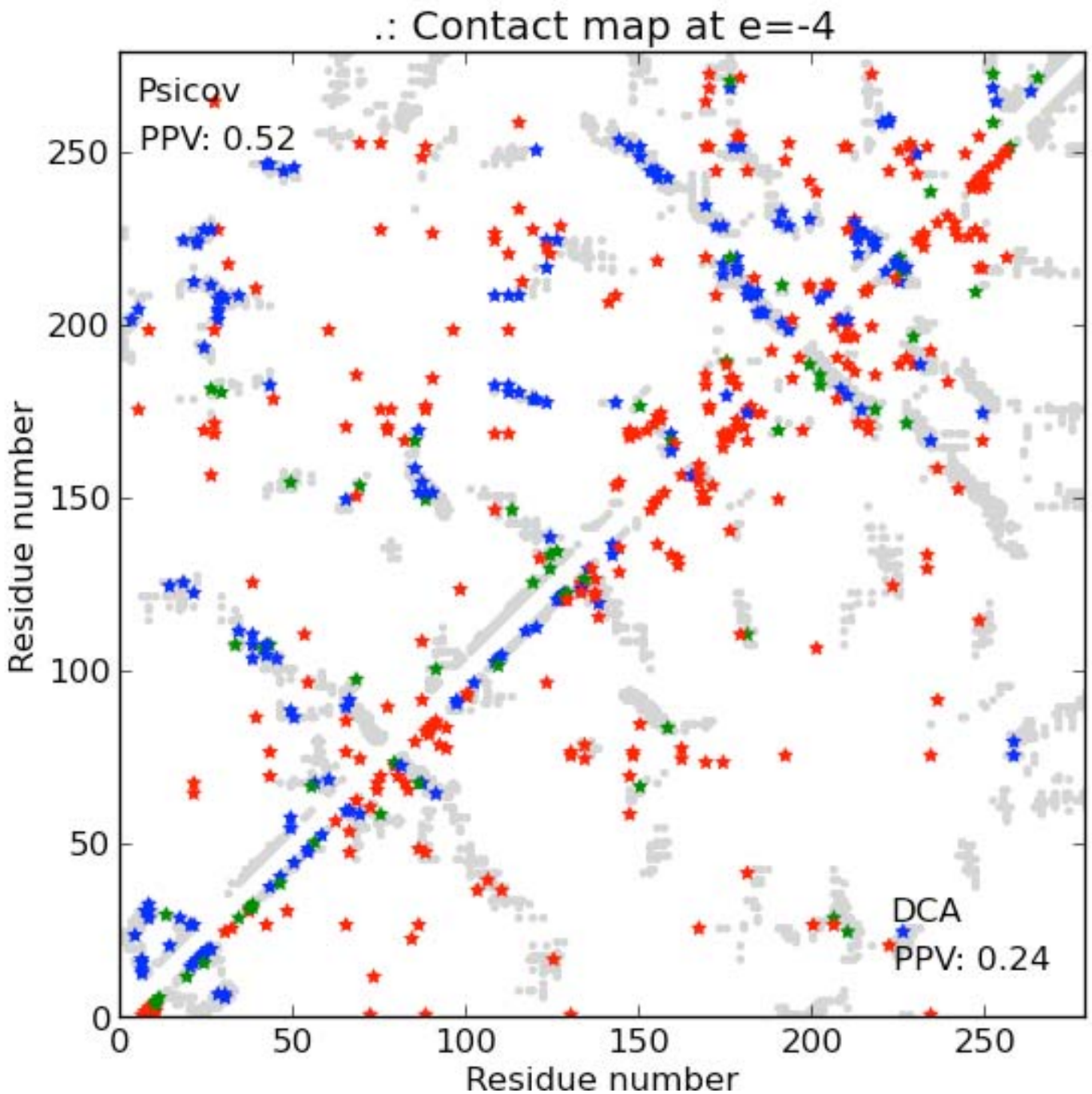
mfDCA

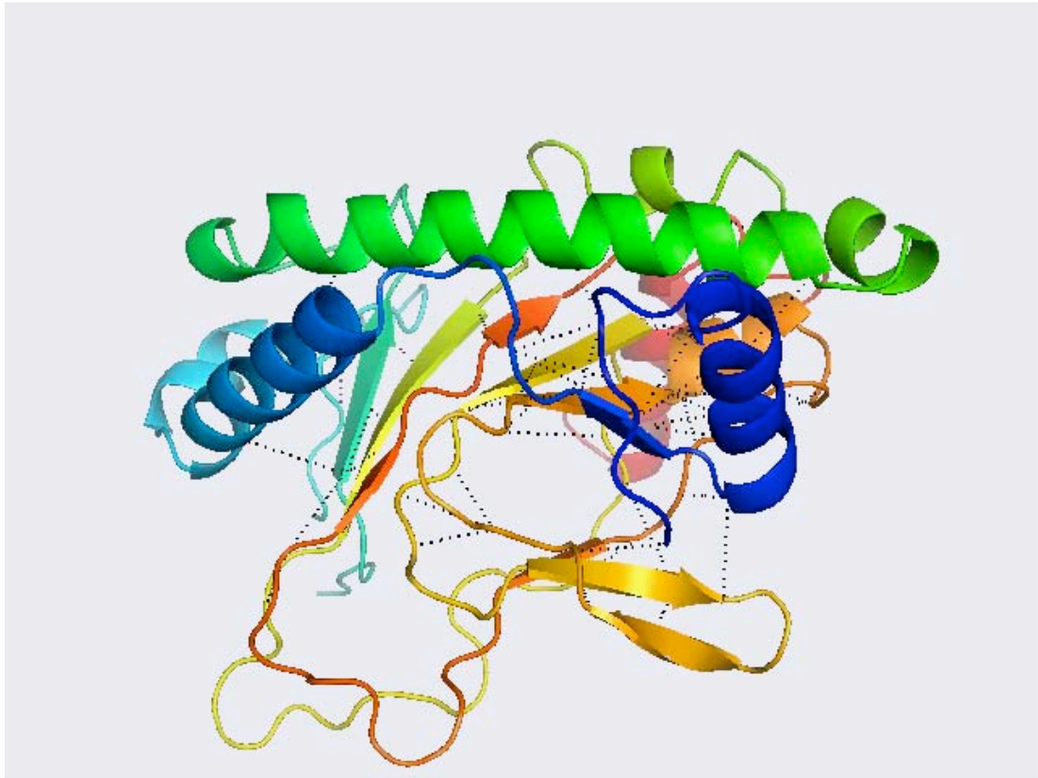




Morcos F. et al.
Direct-coupling analysis of residue coevolution captures native contacts across many protein families (2012)

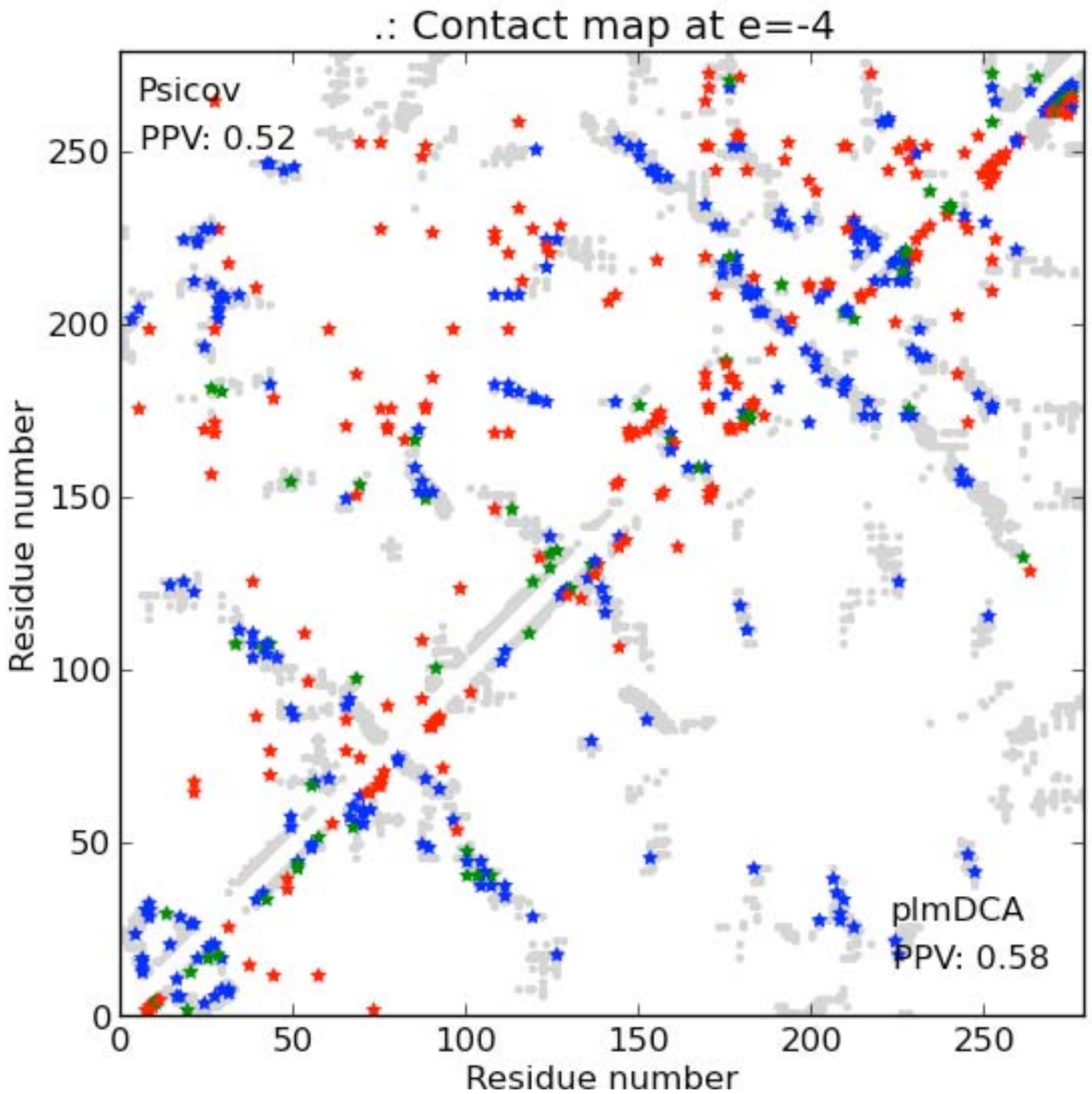
PSICOV

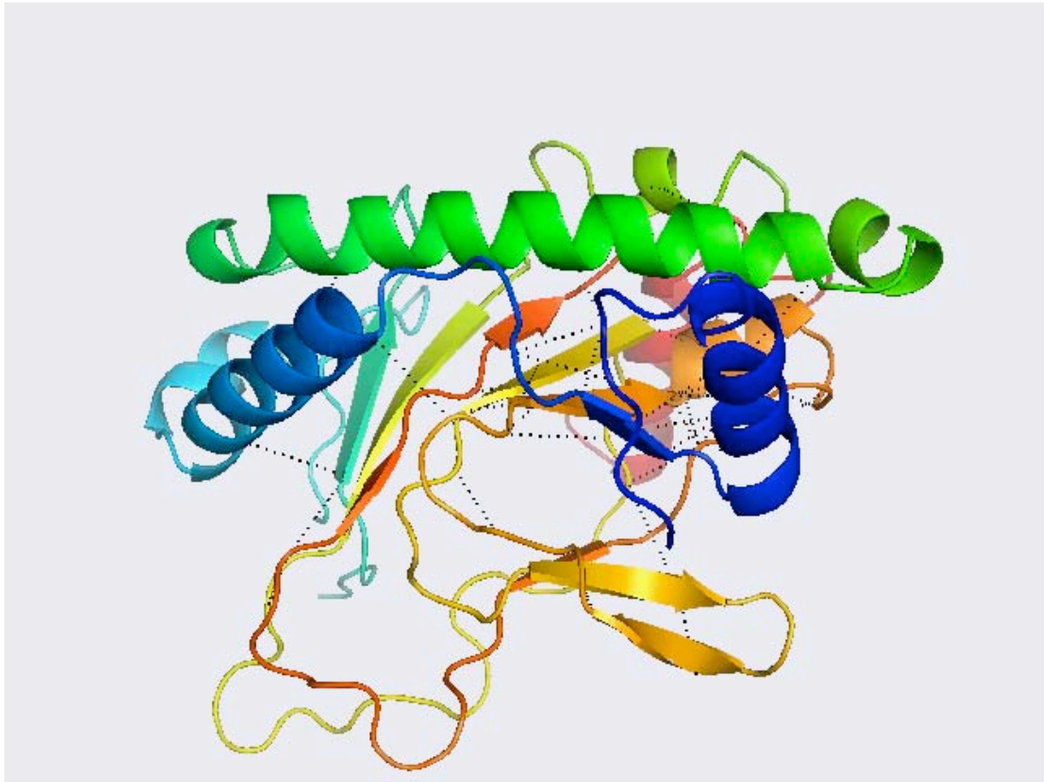




Jones DT, Buchan D.W.A, Cozetto D., Ponti M.
PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments (2012)

PLMDCA

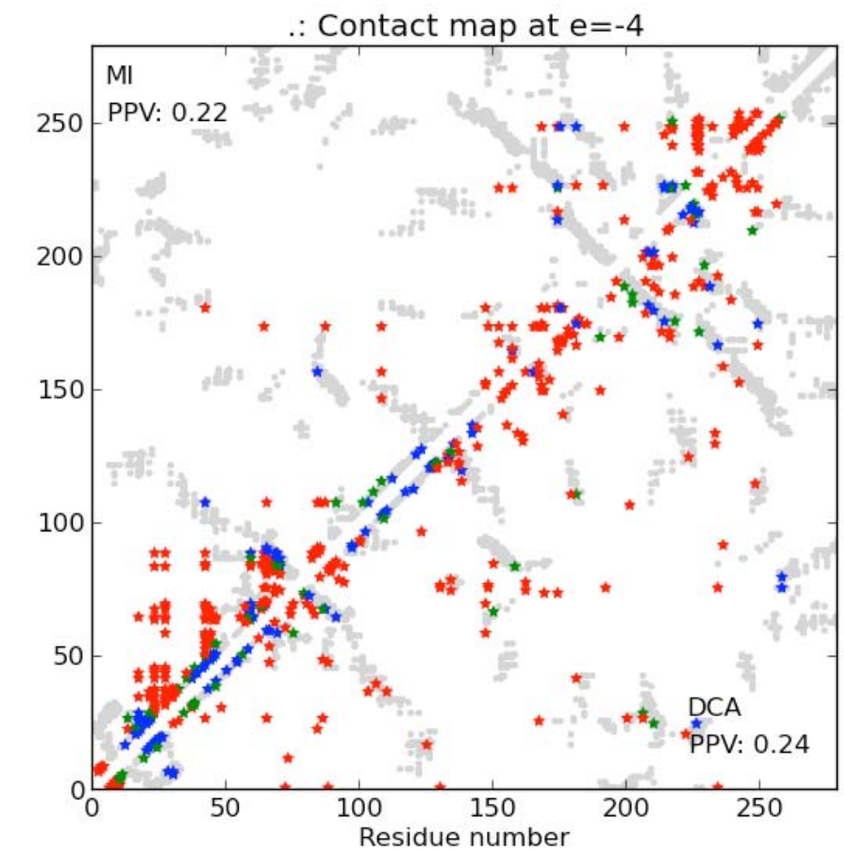


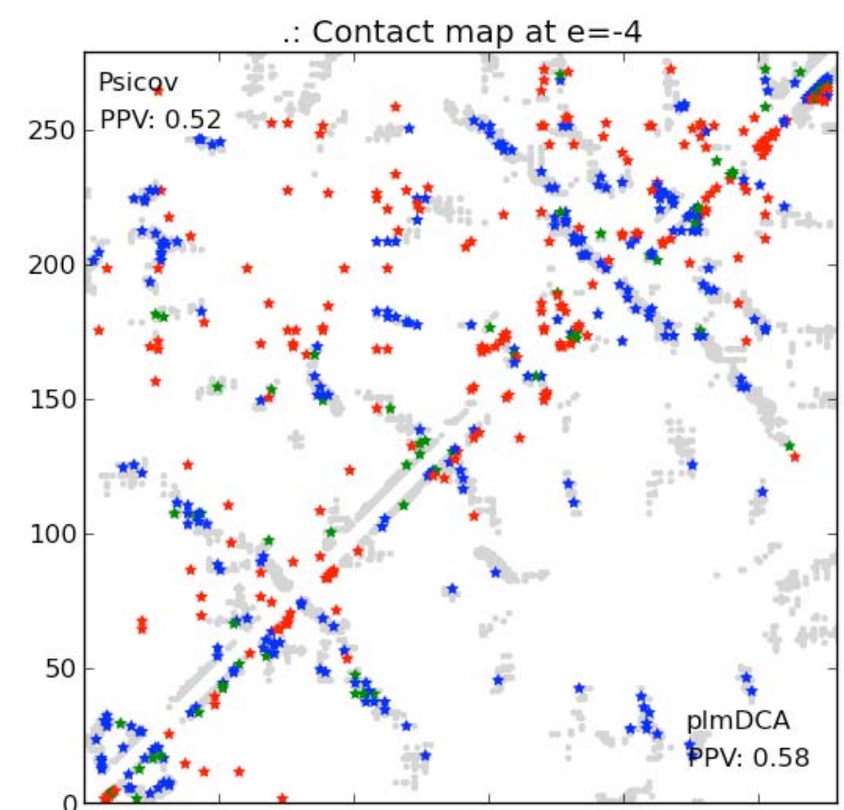


M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell,
Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models (2012)

Factors affecting contact prediction

- Different alignment construction methods
- Homology cutoffs
- Underlying representations and regularization methods





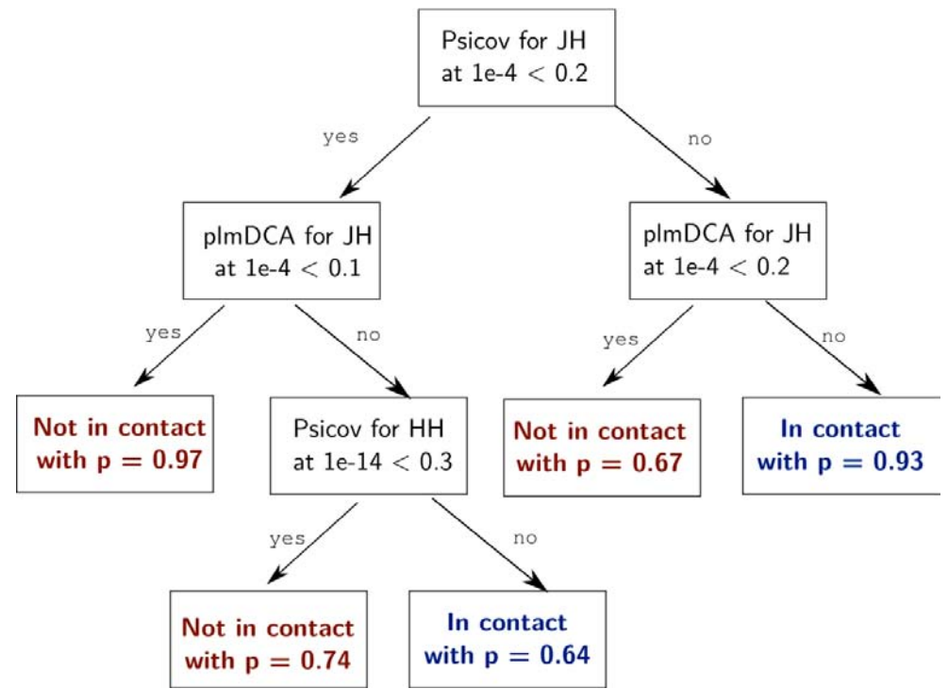
PconsC: Ensemble method for contact prediction

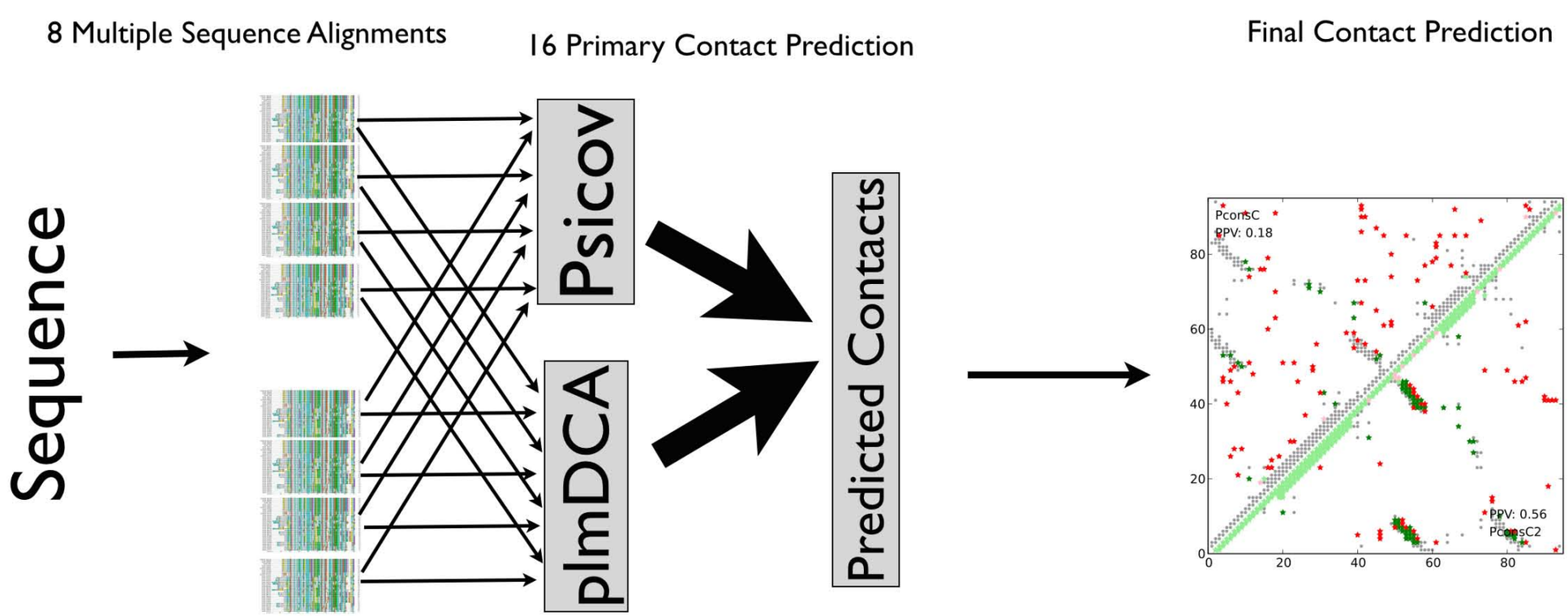
Random forest method reconciling diverse DI-based predictions

- 4 e-value cut offs
- 2 homology search methods
- 2 contact prediction methods
- 100 decision tress

Optimised on set of 48 non-homologous proteins with known structure.

Averaging probabilities of individual trees, instead of voting



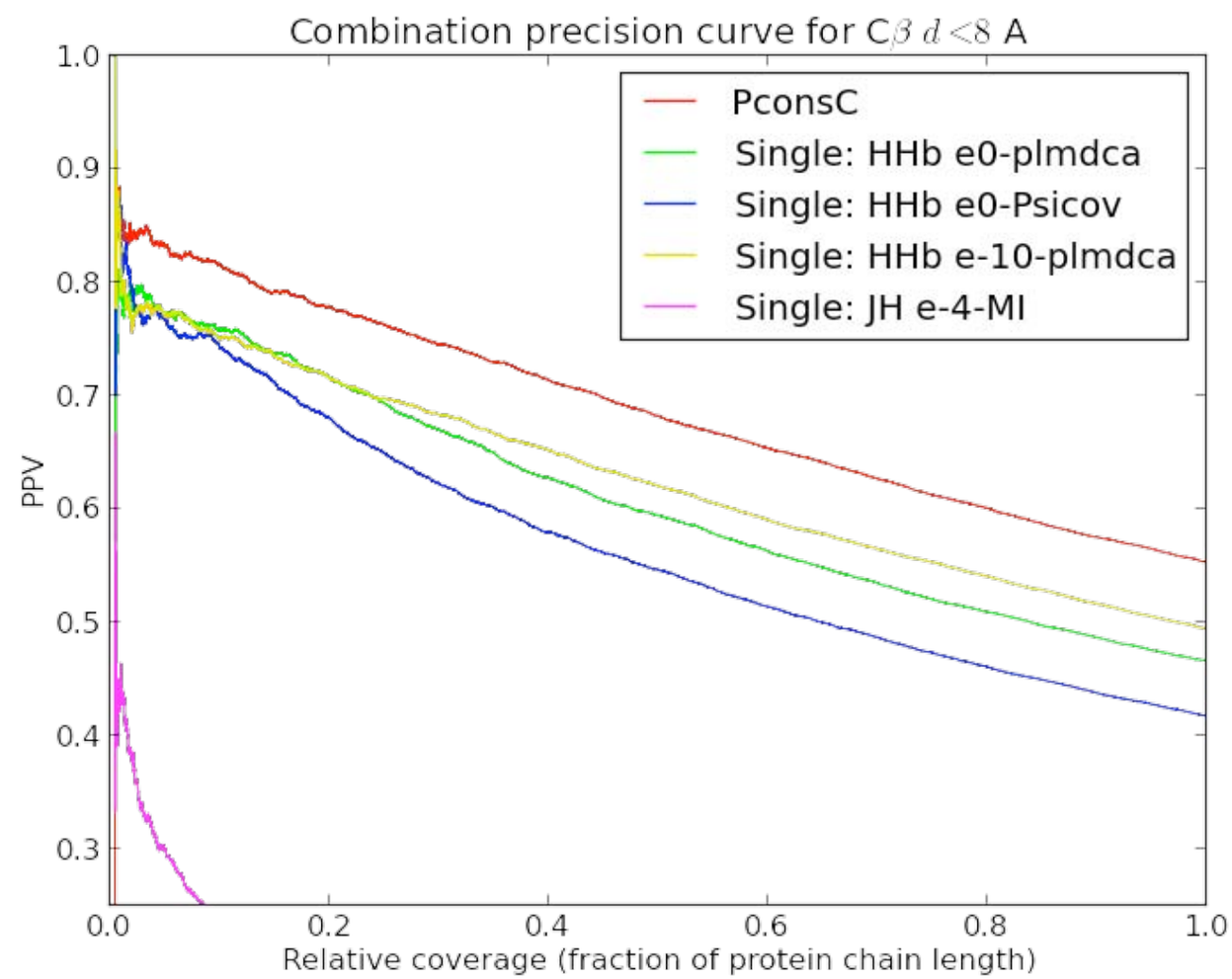


PconsC: Performance

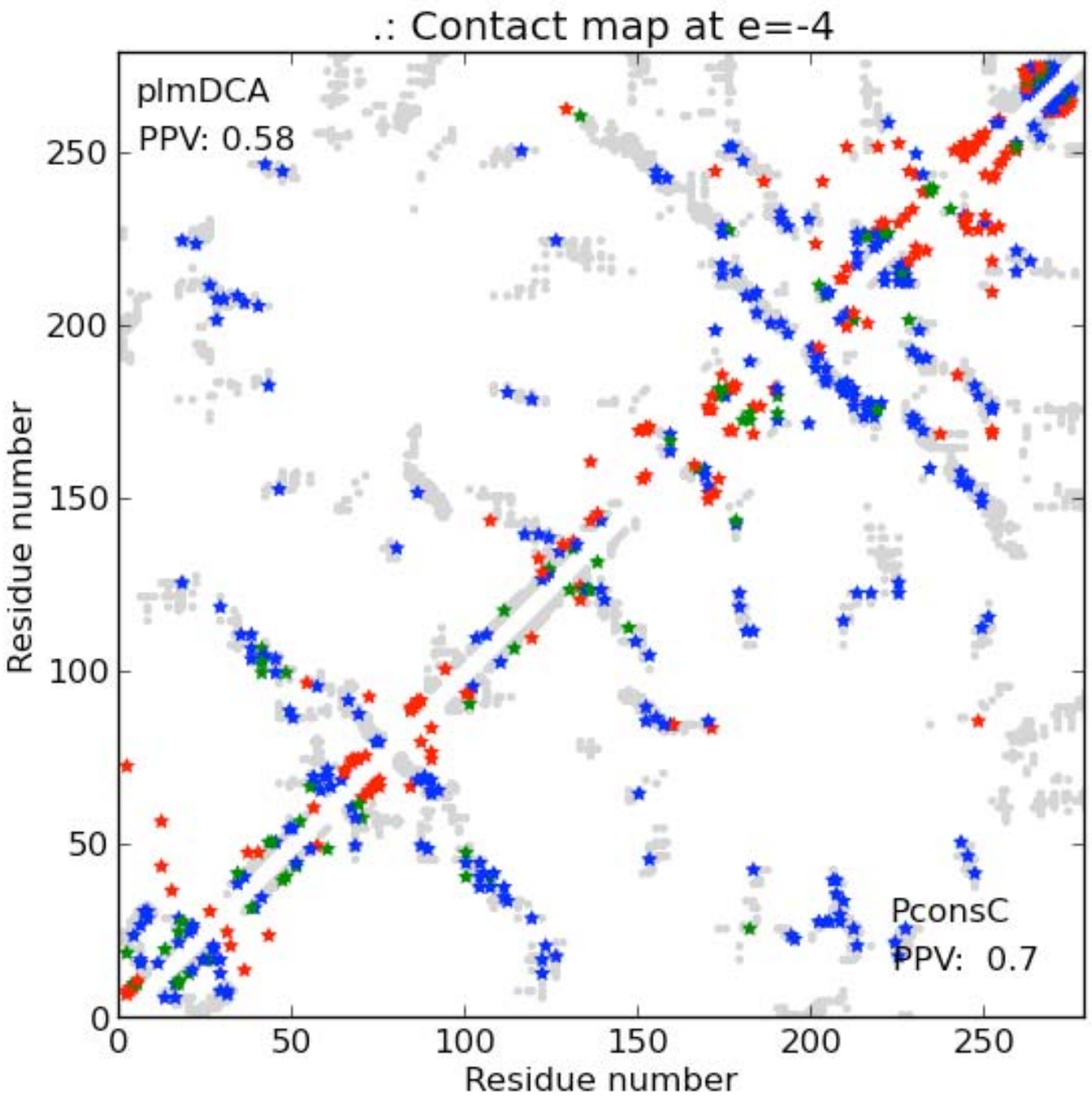
Performance on set of 150 small proteins used in PSICOV development is ~20% better than the individual methods.

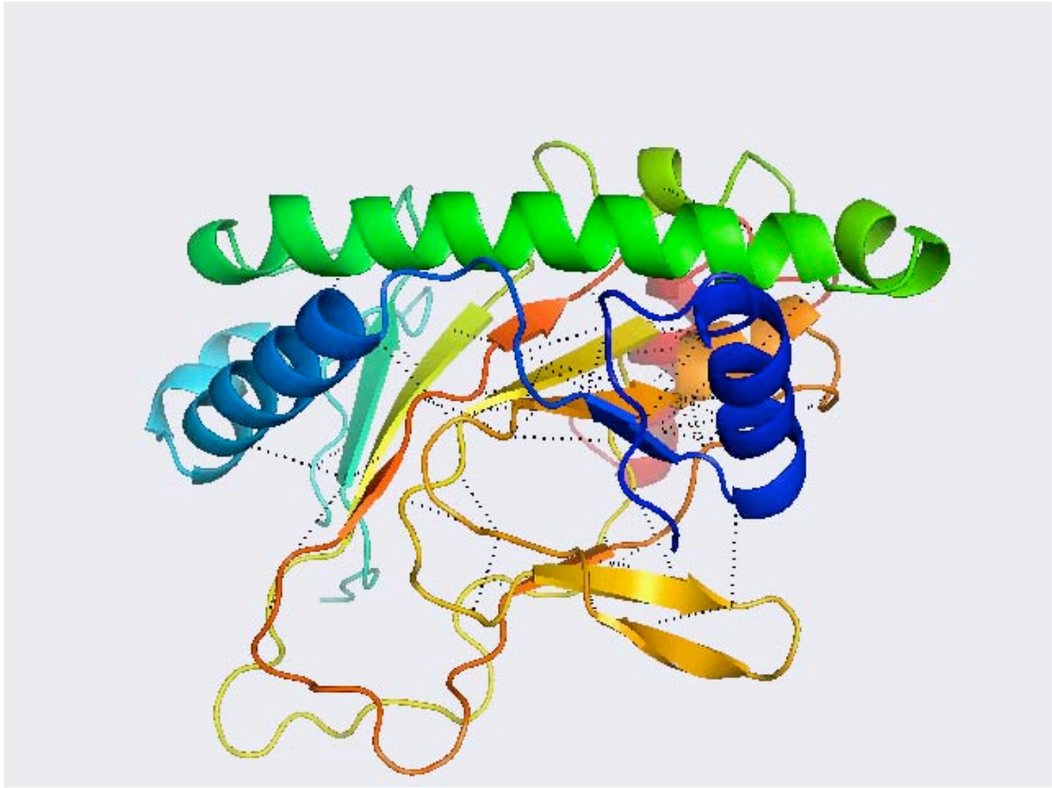
Reasons for performance:

- artifact cancelation
- contact reinforcing
- increased coverage



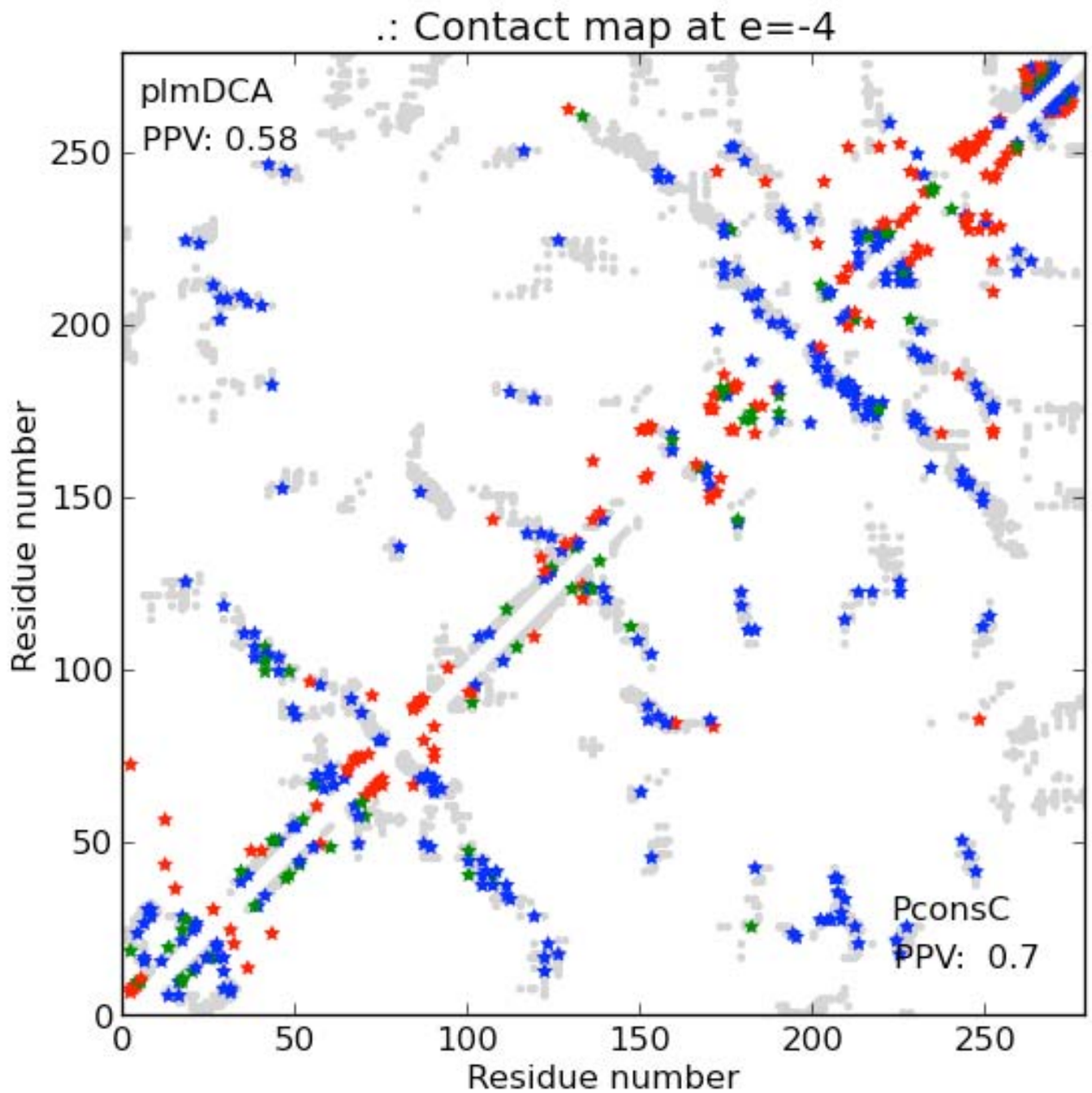
PconsC: Example





Combining alignment methods, cutoffs and prediction methods

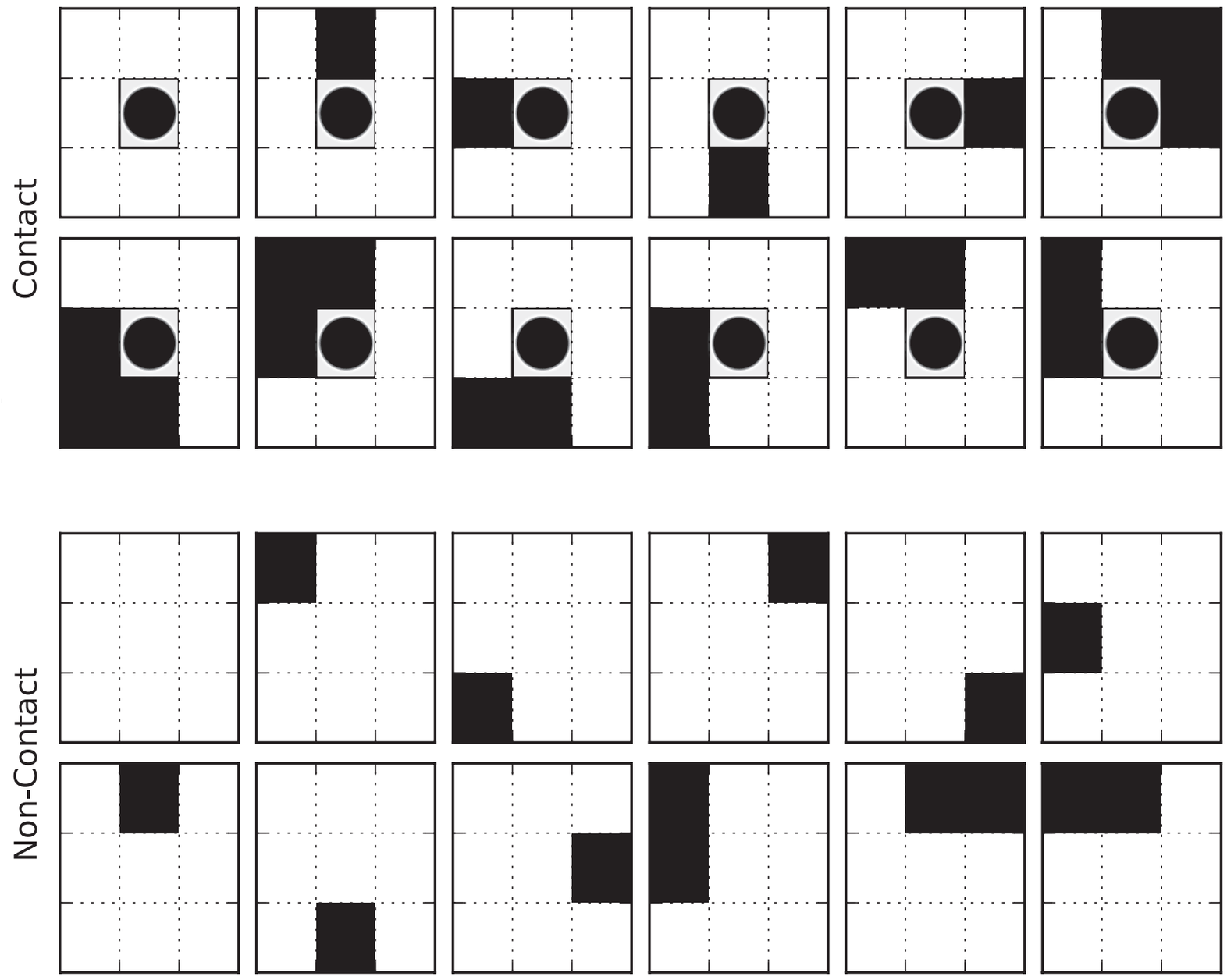
Even better performance ?



Contacts in contact maps are not randomly distributed

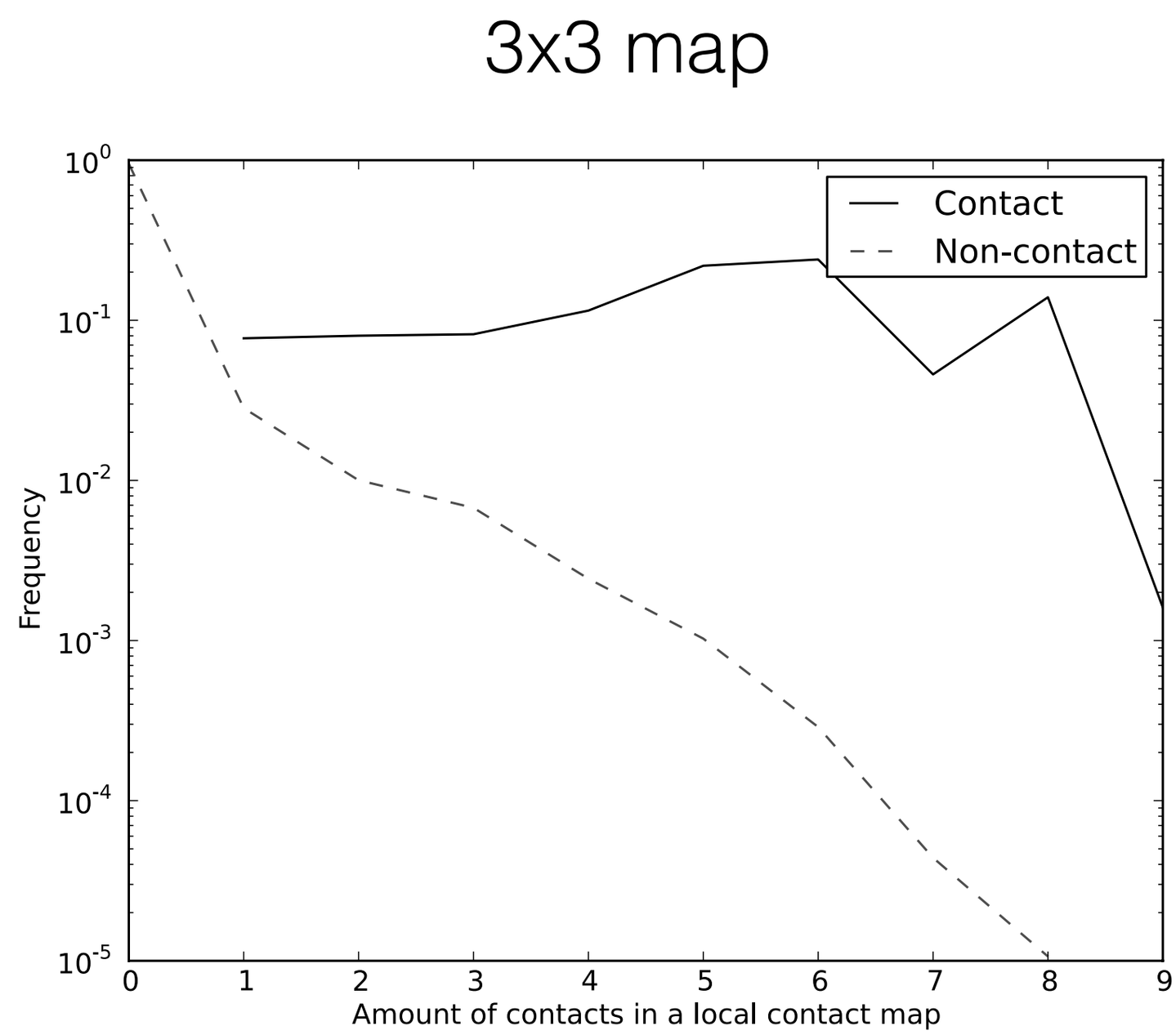
Most frequent 3x3 contact patterns

- Some patters in a map are more frequent than others
- Some patterns are forbidden.
- Contact predictors should use this, by:
  - Specific rules
  - Machine Learning

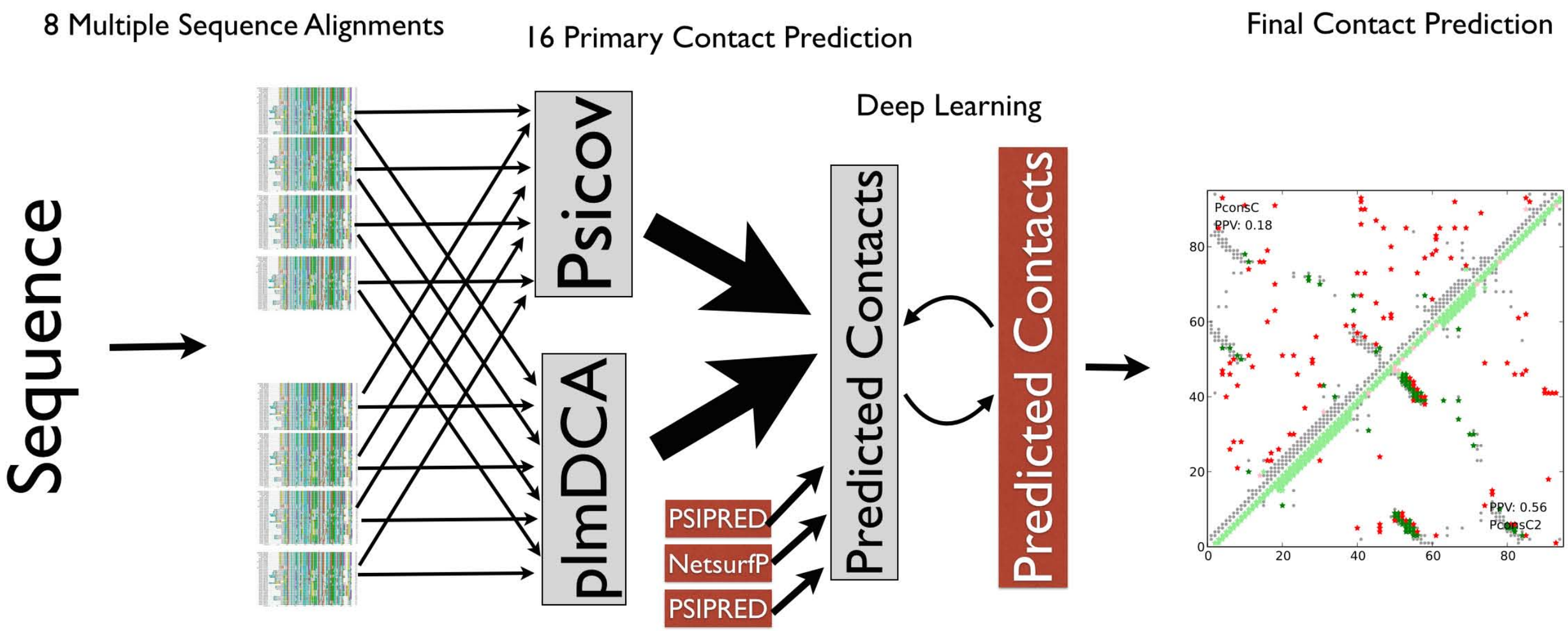


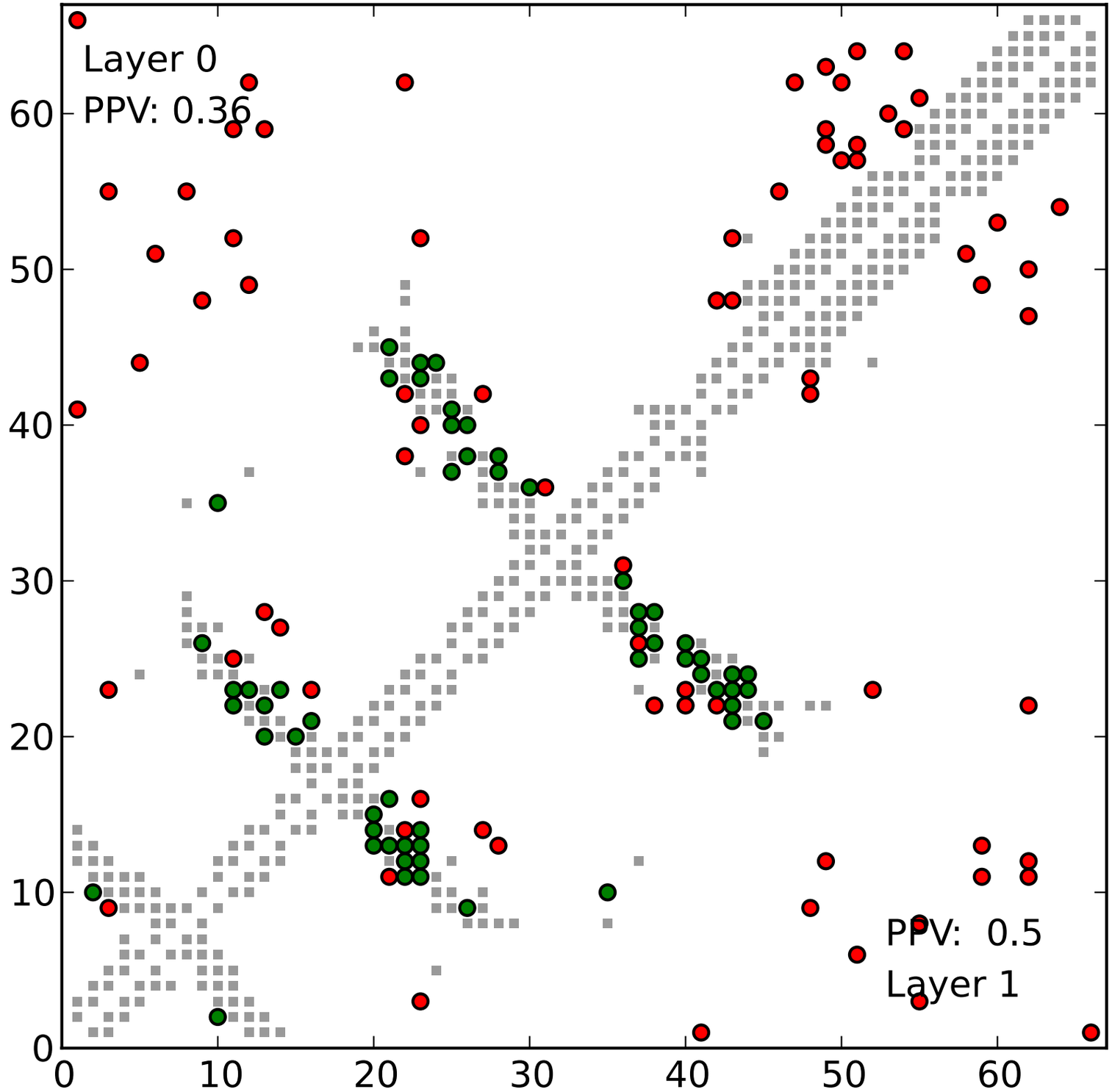
Contacts and non-contacts

- Features of contact maps:
  - Contacts close to other contacts
  - Mainly empty
  - Visible diagonal patterns

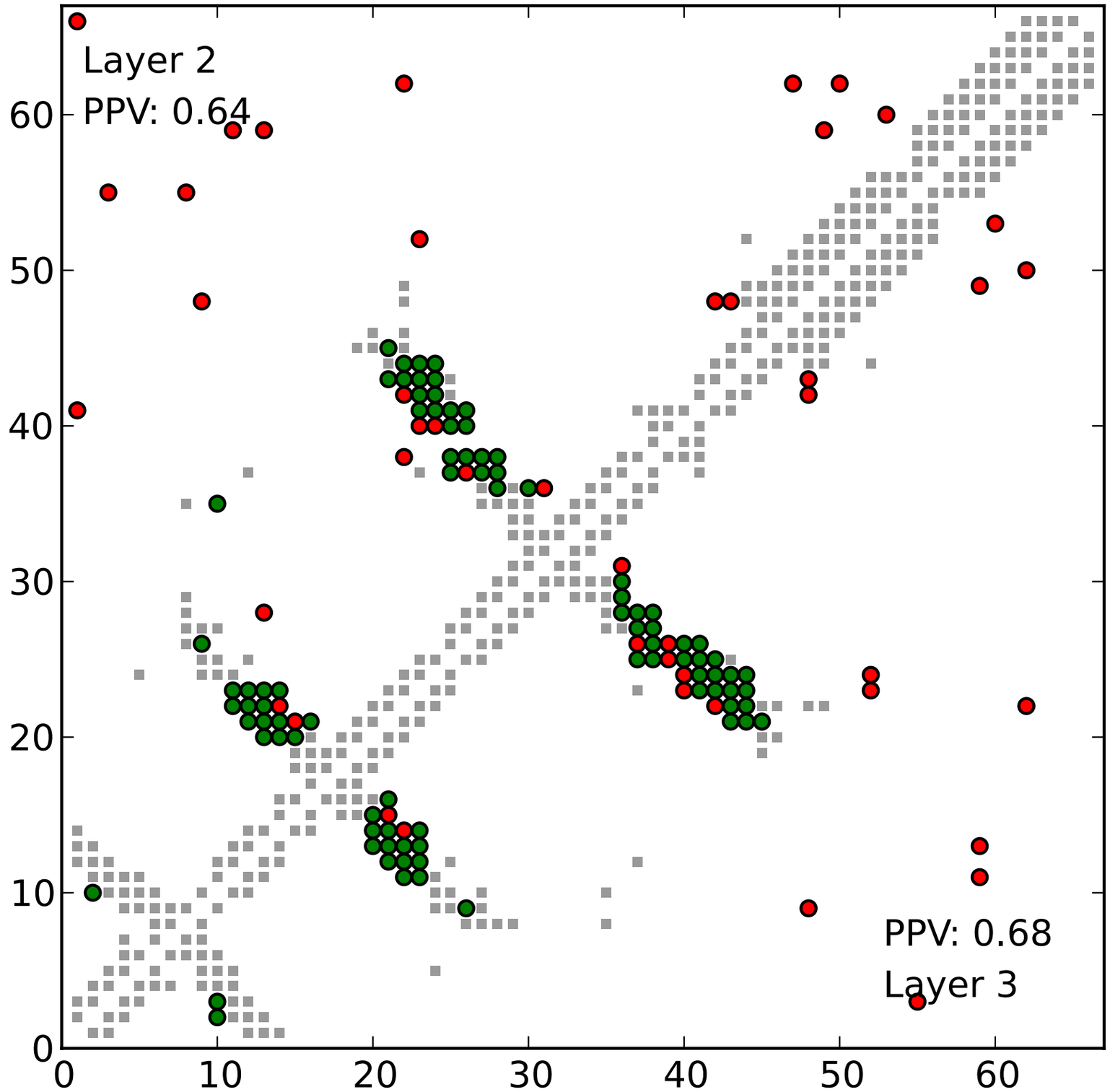


PconsC2 pipeline

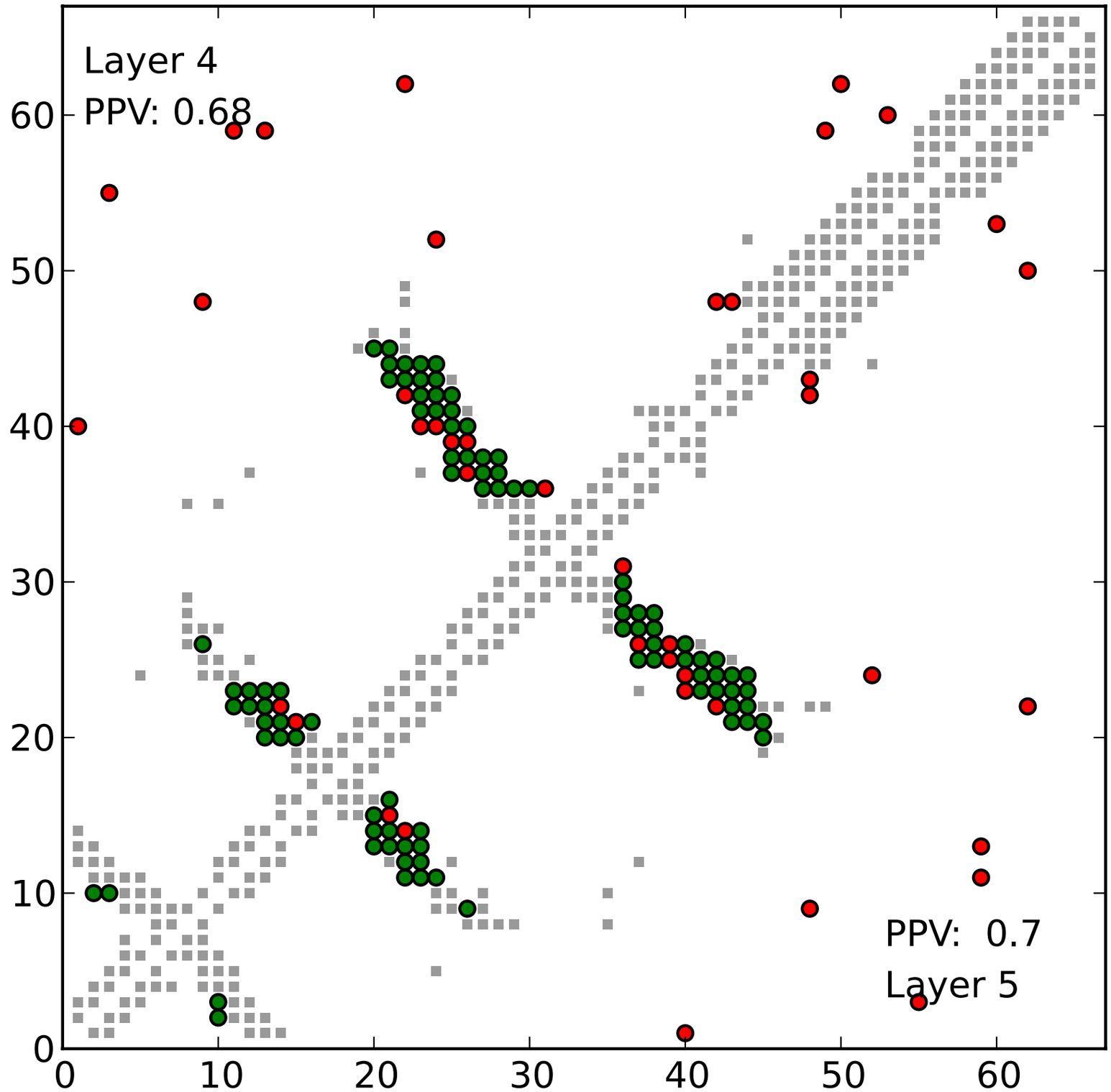




(b) 1pcf:A Layer 0-1

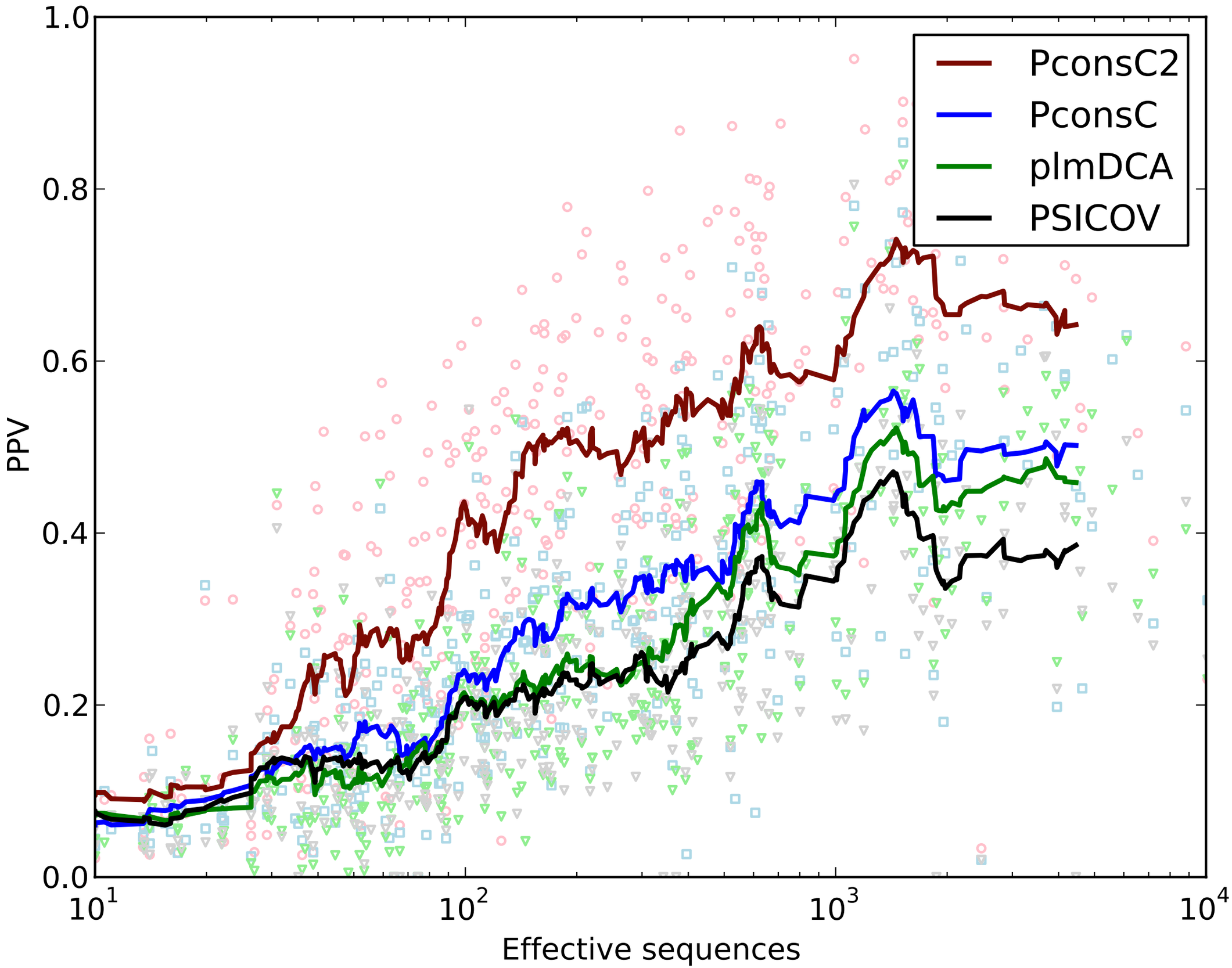


(c) 1pcf:A Layer 2-3

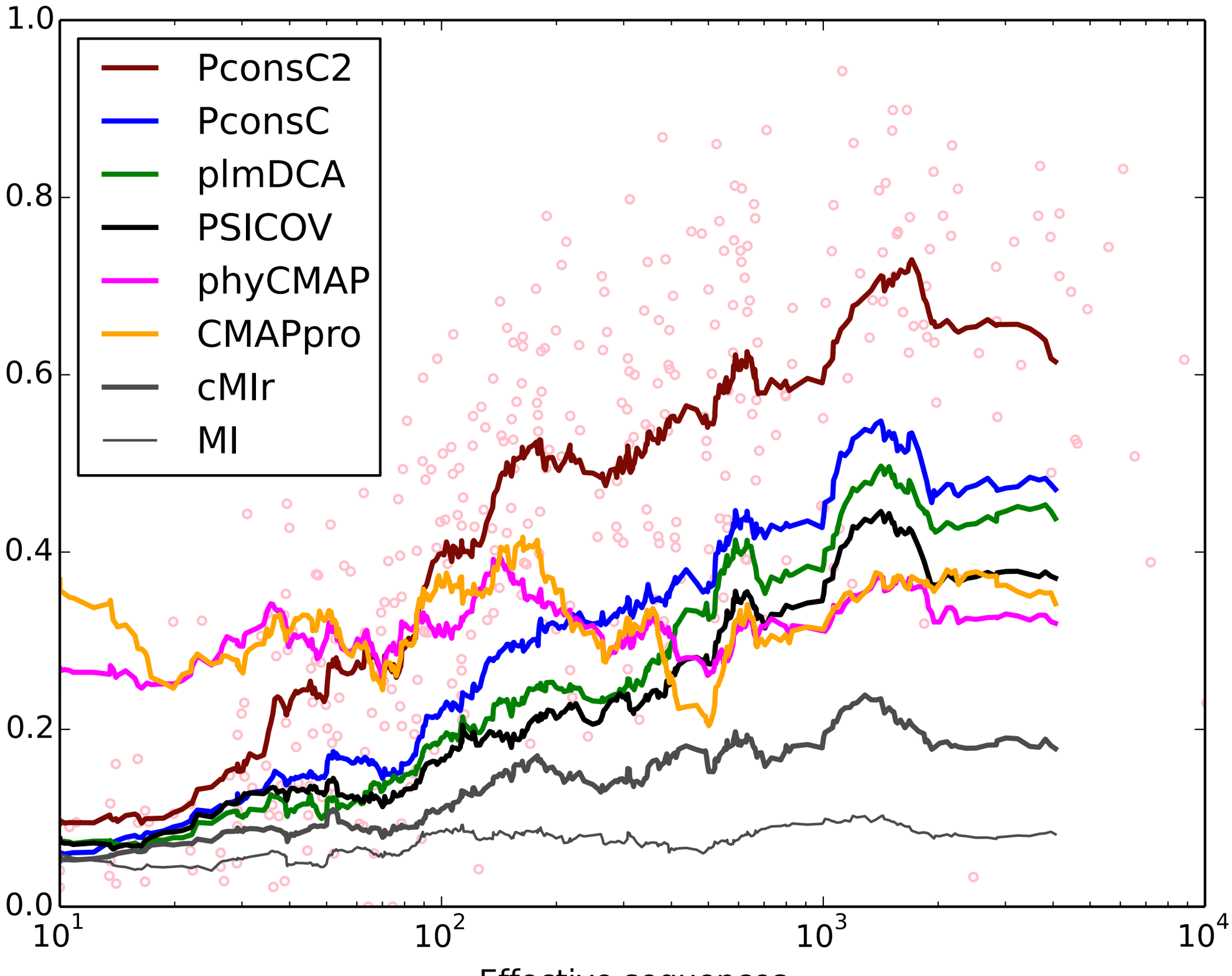


(d) 1pcf:A Layer 4-5

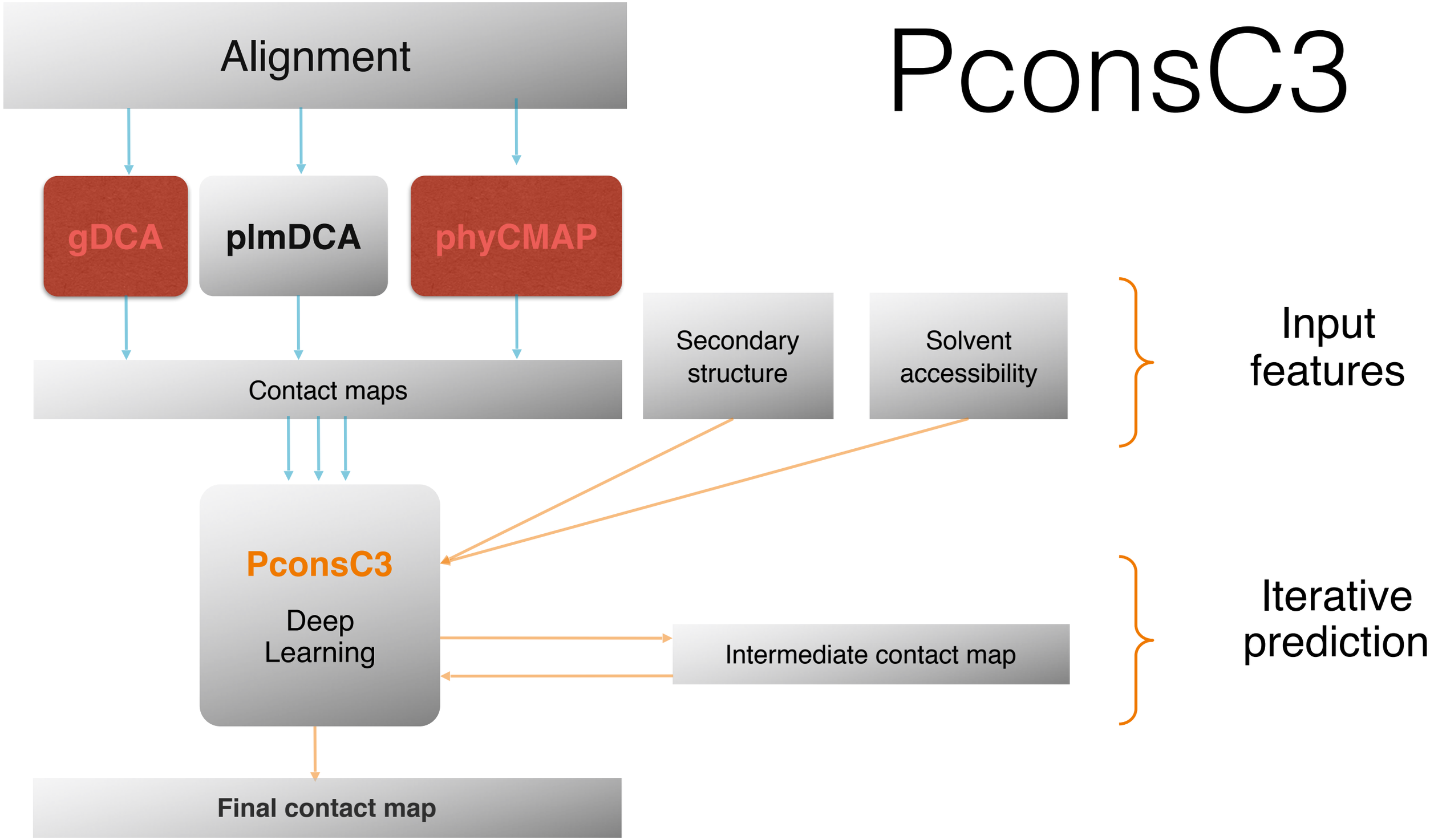
PconsC2 improves predictions for all family sizes.



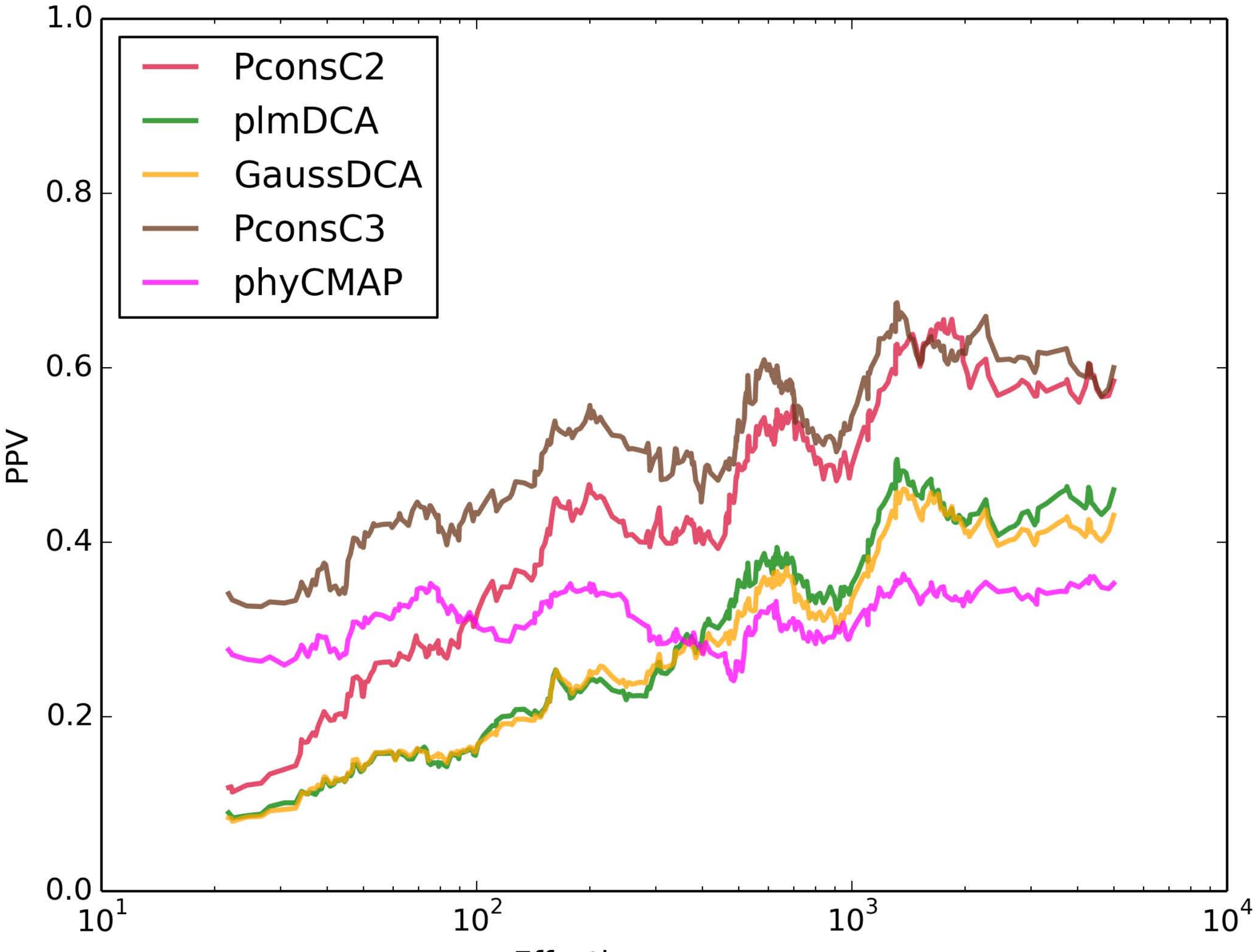
Older methods do better for smaller families



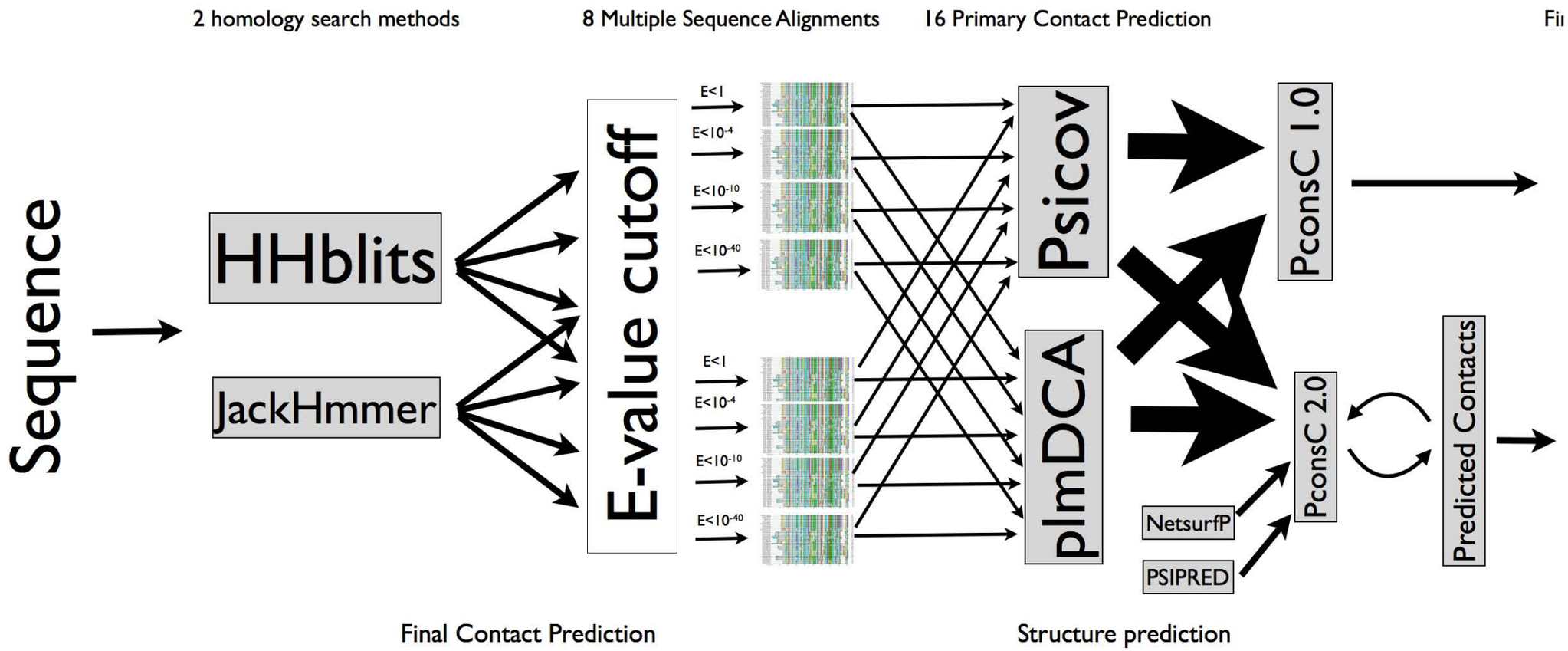
Can we handle small protein families ?

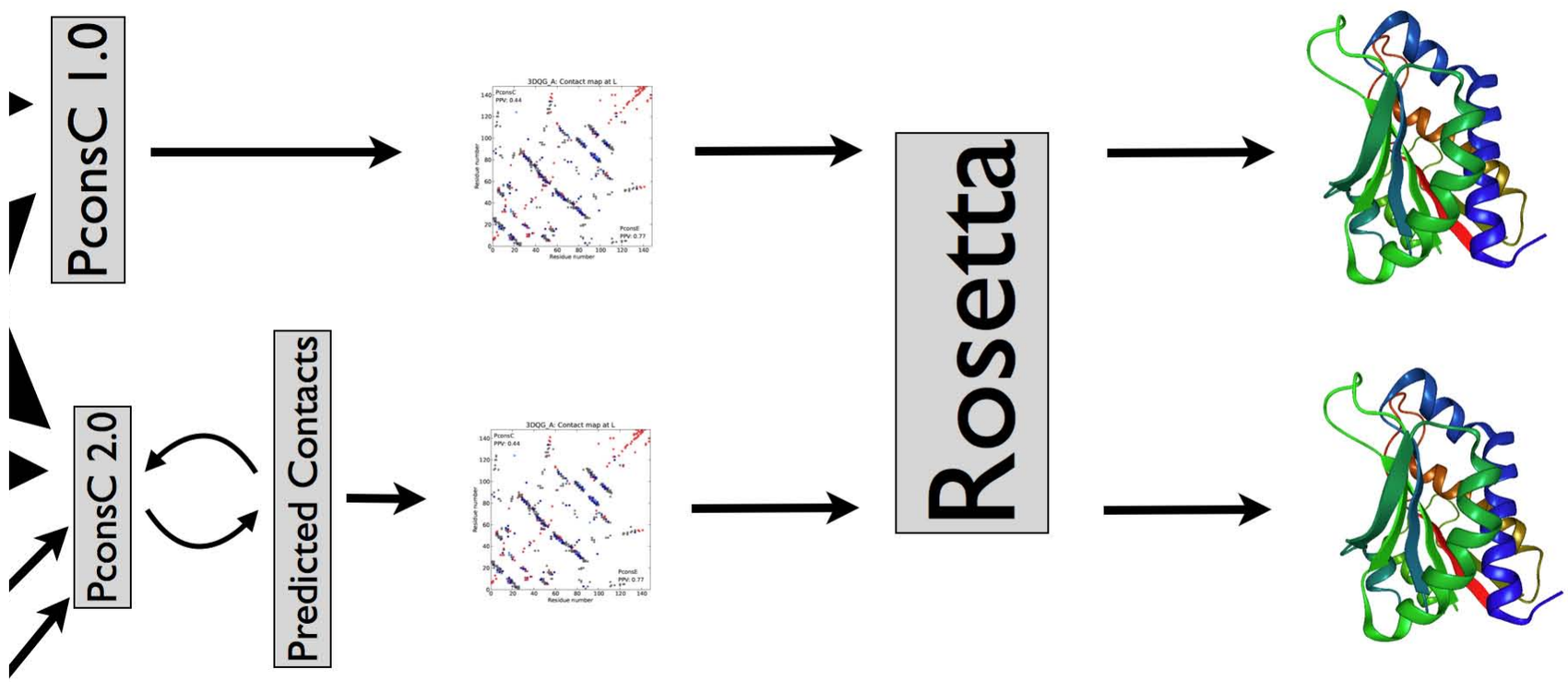


PconsC3 performance

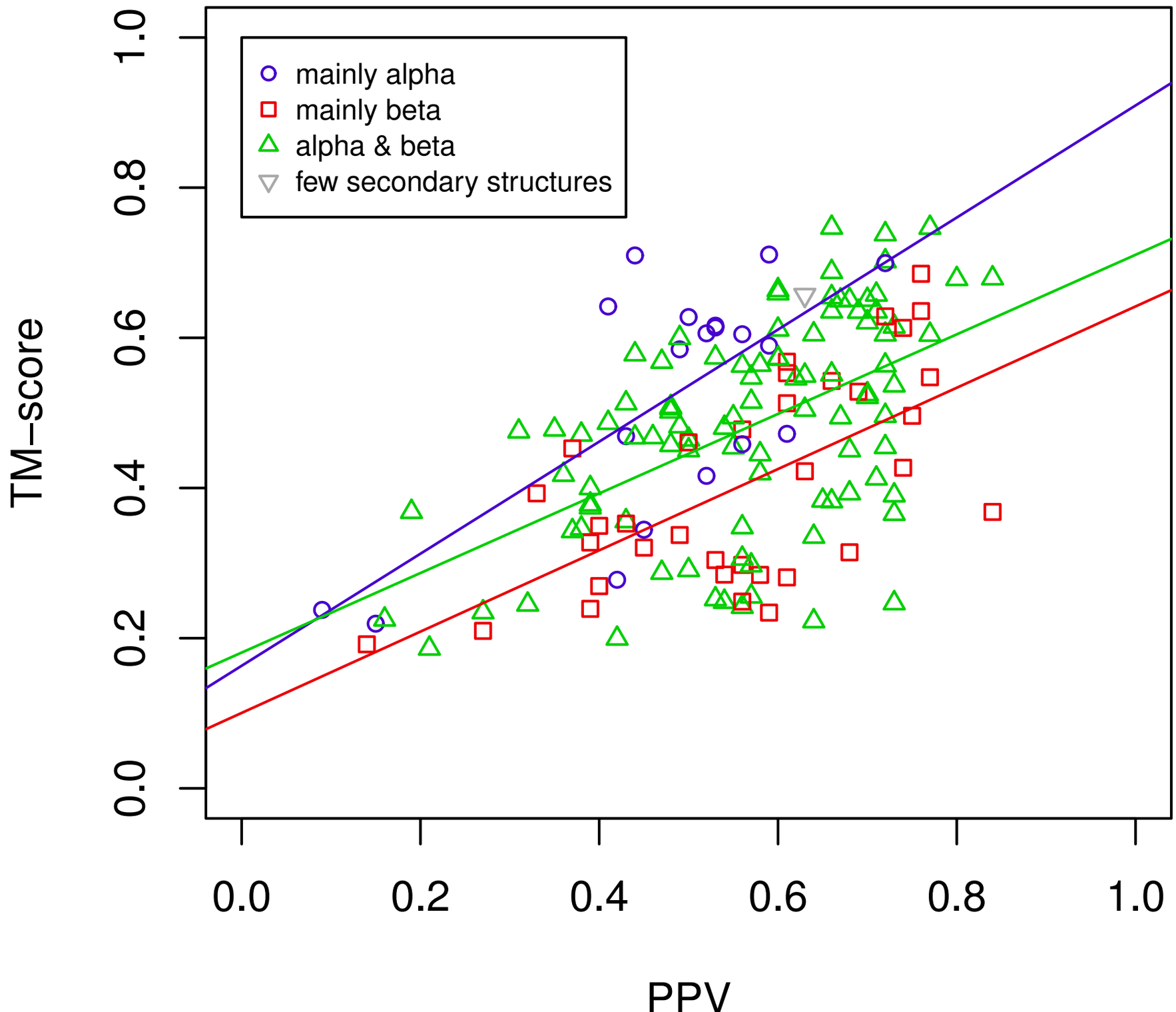


PconsFold


2 homology search methods
8 Multiple Sequence Alignments
16 Primary Contact Prediction
Sequence
HHblits
JackHmmer
E-value cutoff
E<1
E<10-4
E<10-10
E<10-40
Psicov
plmDCA
PconsC 1.0
PconsC 2.0
NetsurfP
PSIPRED
Predicted Contacts

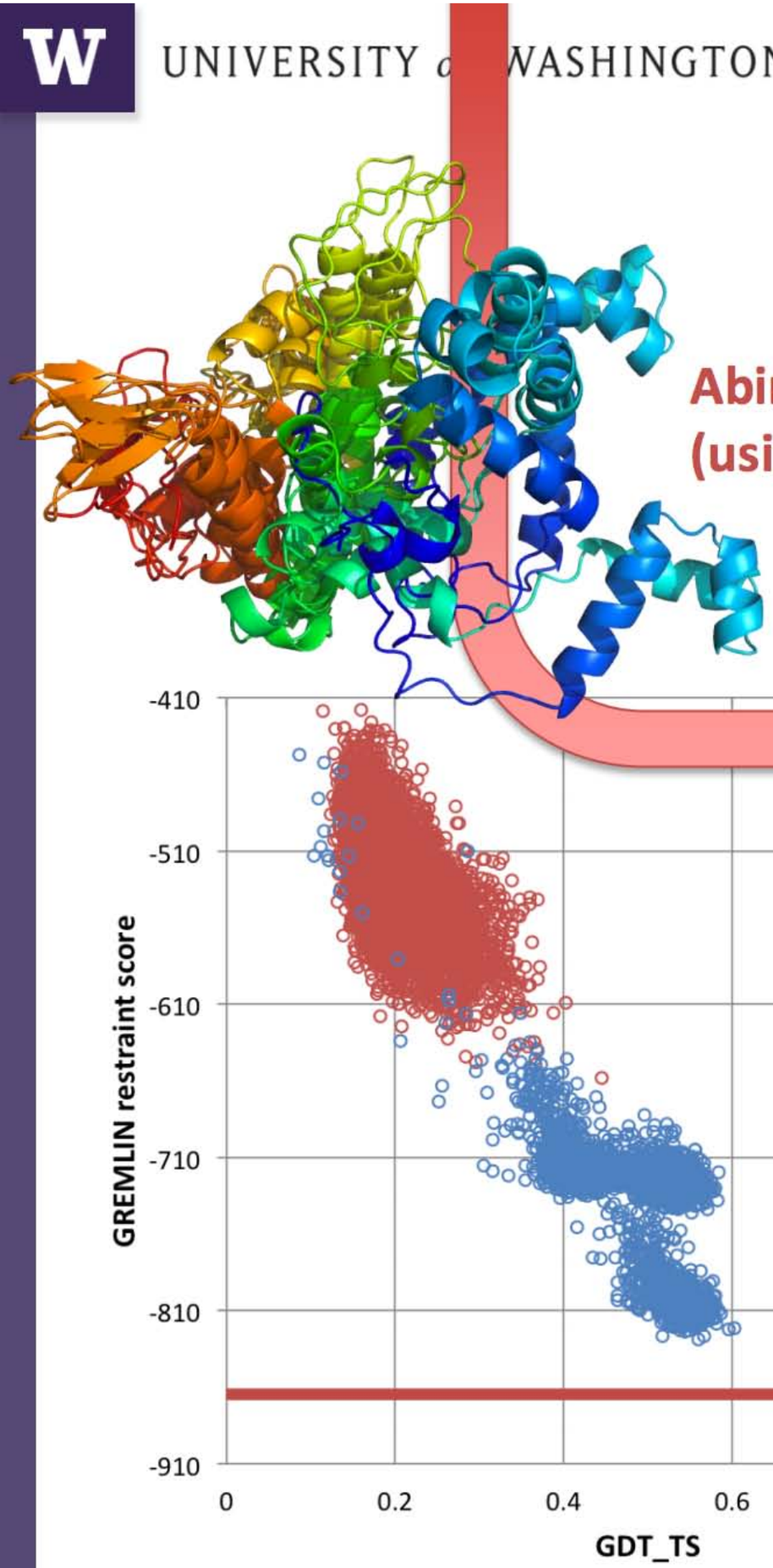

Final Contact Prediction
Structure prediction
PconsC 1.0
PconsC 2.0
Predicted Contacts
Rosetta

Contacts are crucial

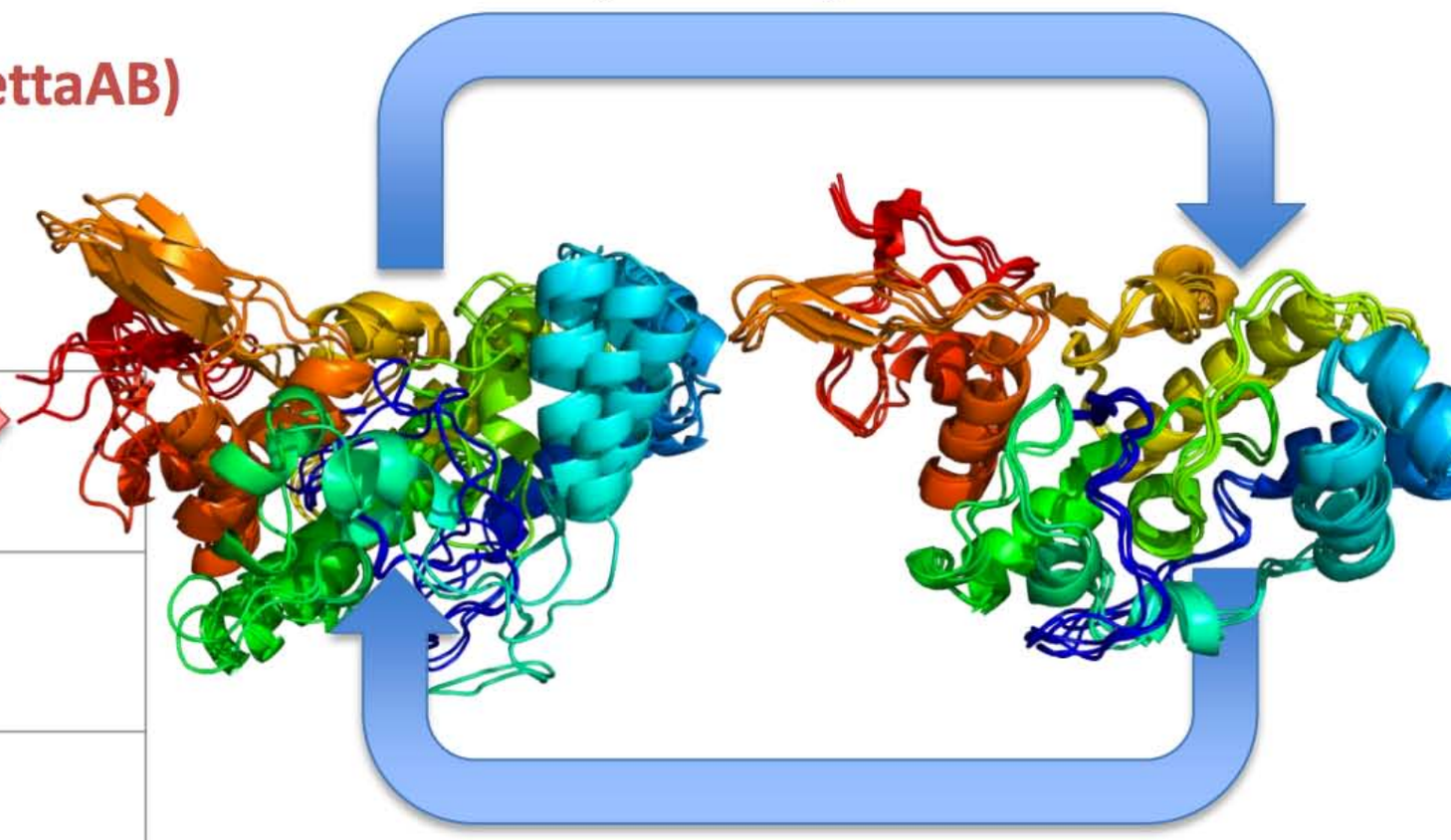


UNIVERSITY of WASHINGTON

Pipeline



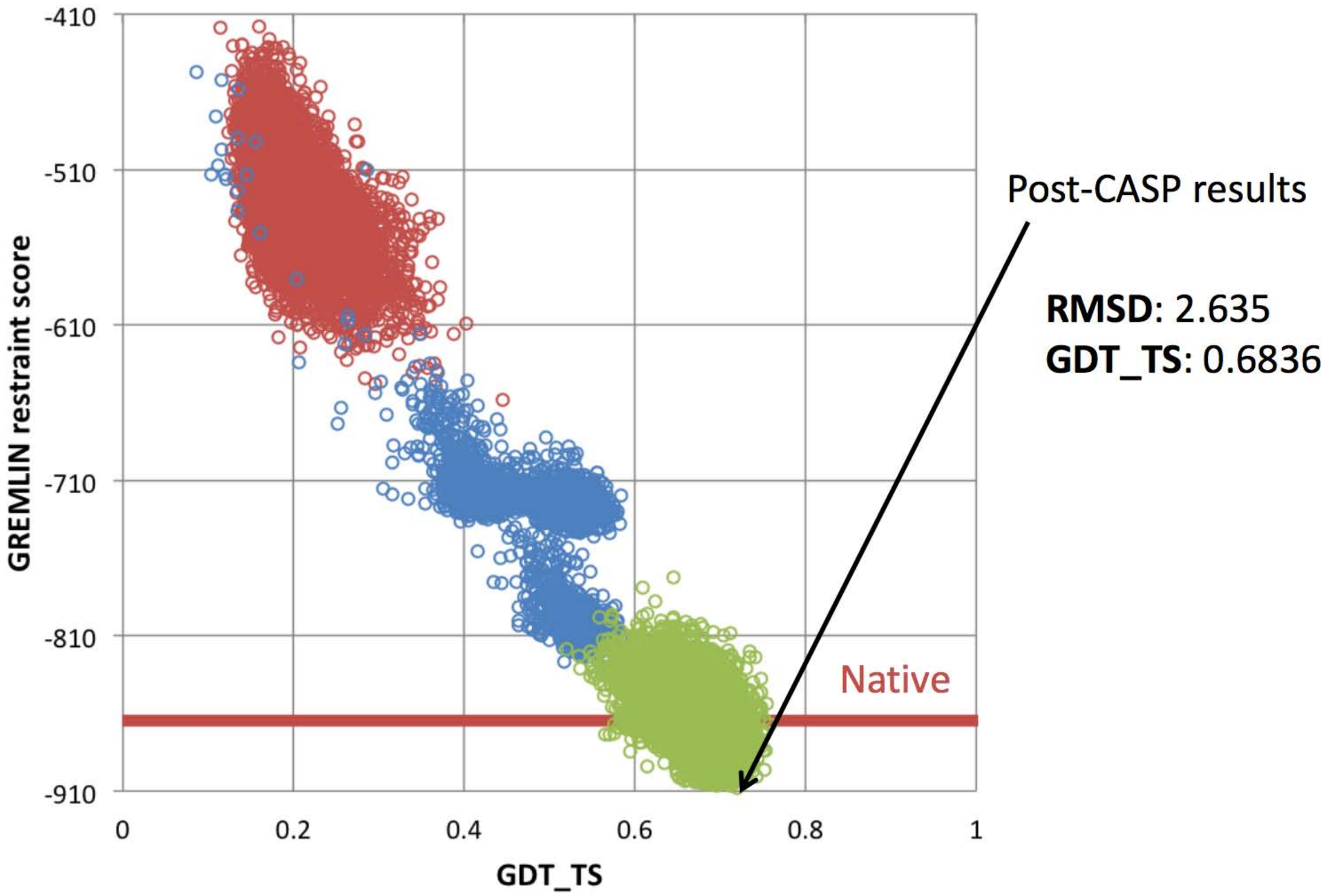
Abinitio
(using RosettaAB)



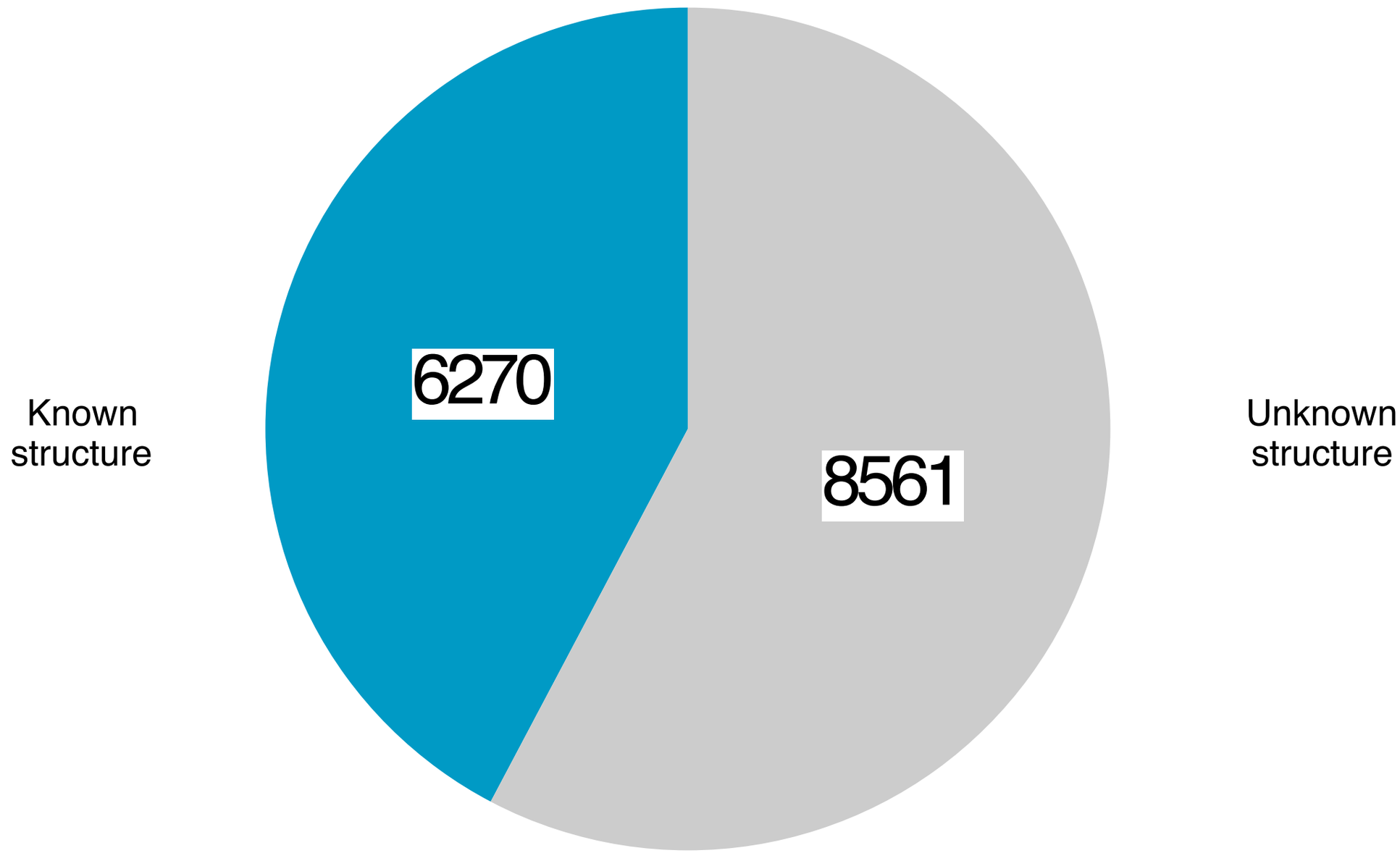
Hybridize (using RosettaCM)
Fragment insertion
(20 trials)

Repeat until **deadline**
or **convergence**.

Post-CASP repeat until **convergence**

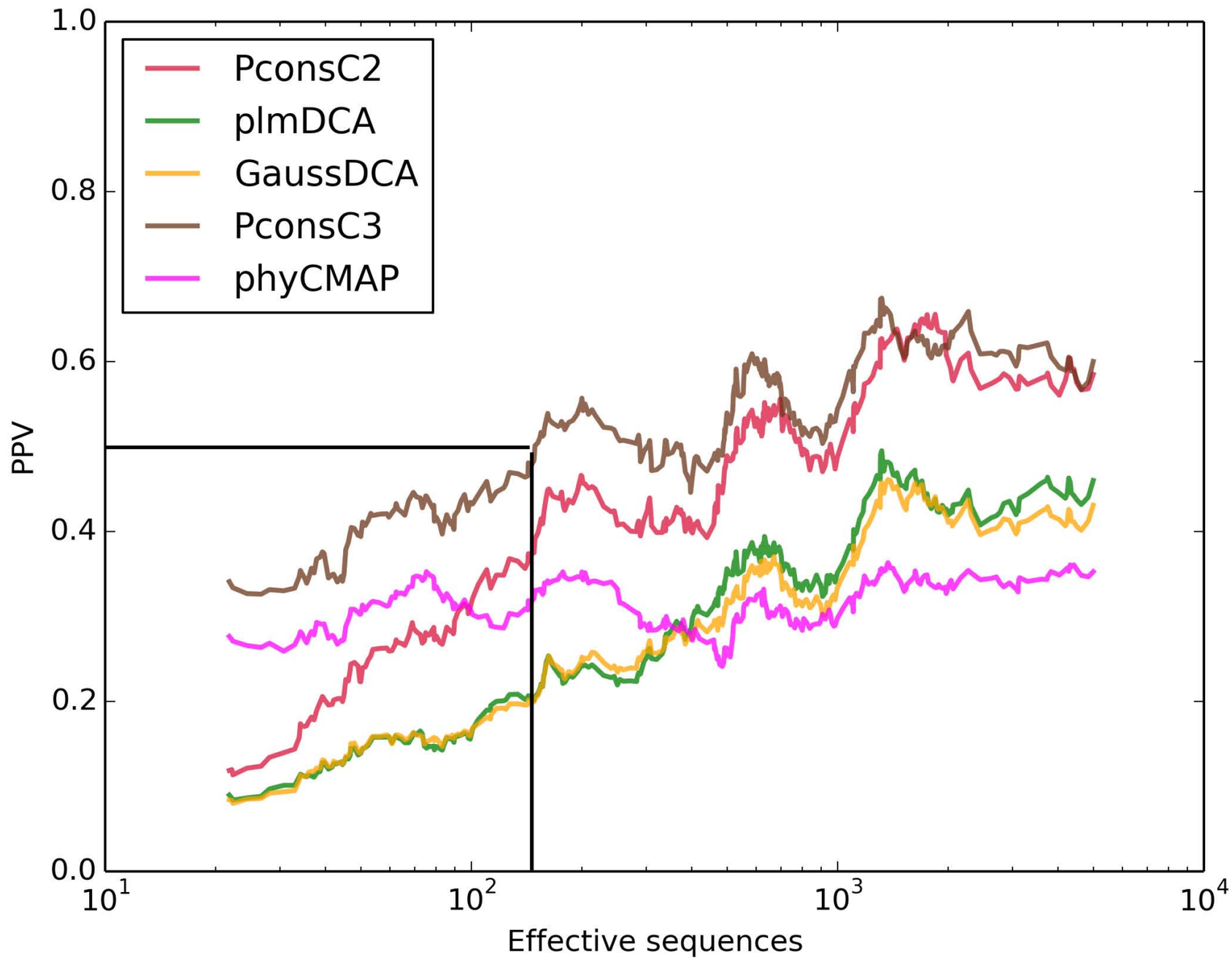


How useful for large scale predictions are contact predictions today?



Pfam familes

We need 100-1000 effective sequences for good predictions.



Predictions could be done for >5000 families !