

Predictor@home: A Multiscale, Distributed Approach for Protein Structure Prediction

C. An, M. Taufer, C. L. Brooks III - The Scripps Research Institute (TSRI)

Financial support through the NIH (MMTSB P41 RR12255) and NSF (CTBP PHY-0216576 and PHY-0225630)

Public-Resource Computing

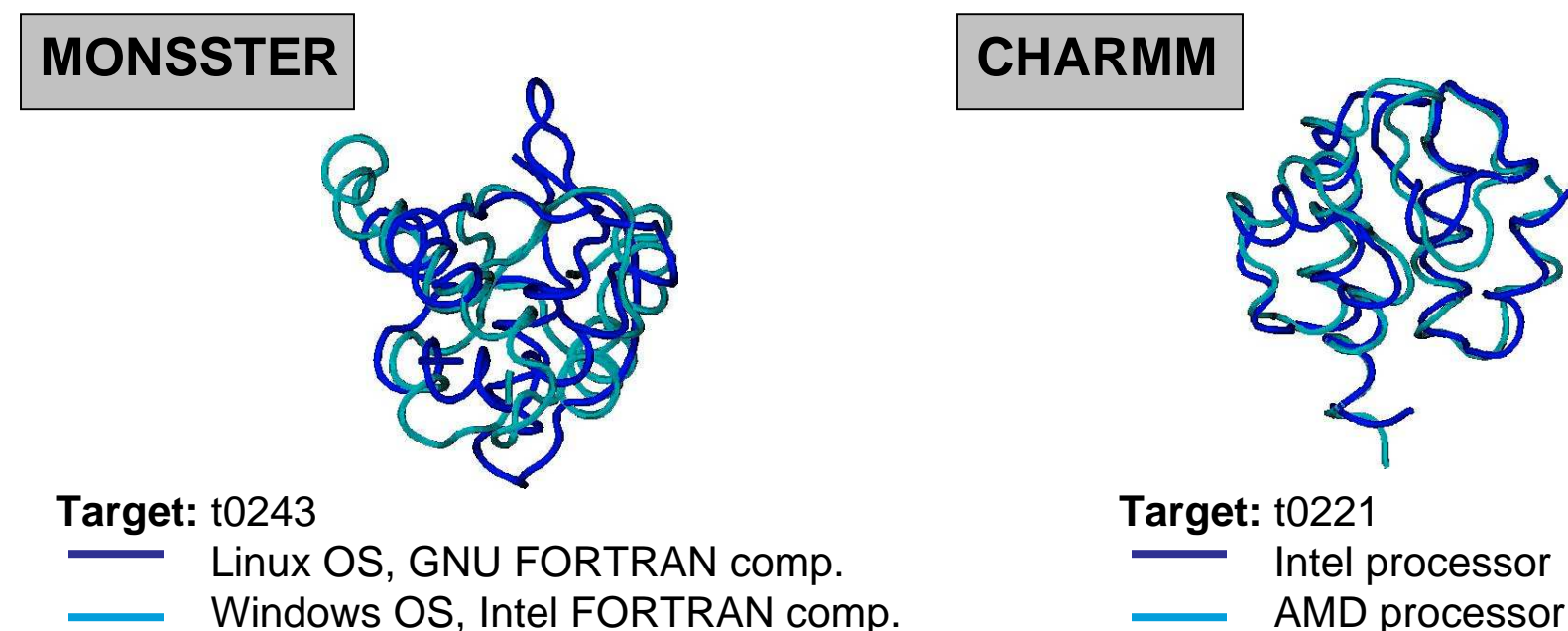
- Public-Resource Computing → PCs connected to the Internet and Intranets
- Their number is large and still growing
- 80%-90% of their CPU time is idle time
- Advantage of commodity-priced hardware and open-source software offers good price/performance ratio



Public-resource systems are highly **heterogeneous, volatile** computational environments → well-suited applications require checkpointing techniques and robust computations

Homogeneous Redundancy

- Replication for detecting errors in public-resource systems:
 - Hardware malfunctions
 - Incorrect software modifications
 - Malicious attacks
- MD and MC simulations can produce different outcomes for replicas of same computations on public-resource systems:
 - MD and MC simulations are subject to positive Lyapunov exponents → Computation results are highly sensitive to initial simulation states and might be **divergent**
 - Initial simulation states include machine architecture, operating system, specific compiler and compiler flags



- Fuzzy comparisons → not applicable to MD and MC replicas

Homogeneous Redundancy
distribution of replicas of the same computation only among "numerically equivalent" PCs, i.e. that compute identical floating-point results

Structure Prediction Through the Public-Resource Computing Paradigm

Our objective:

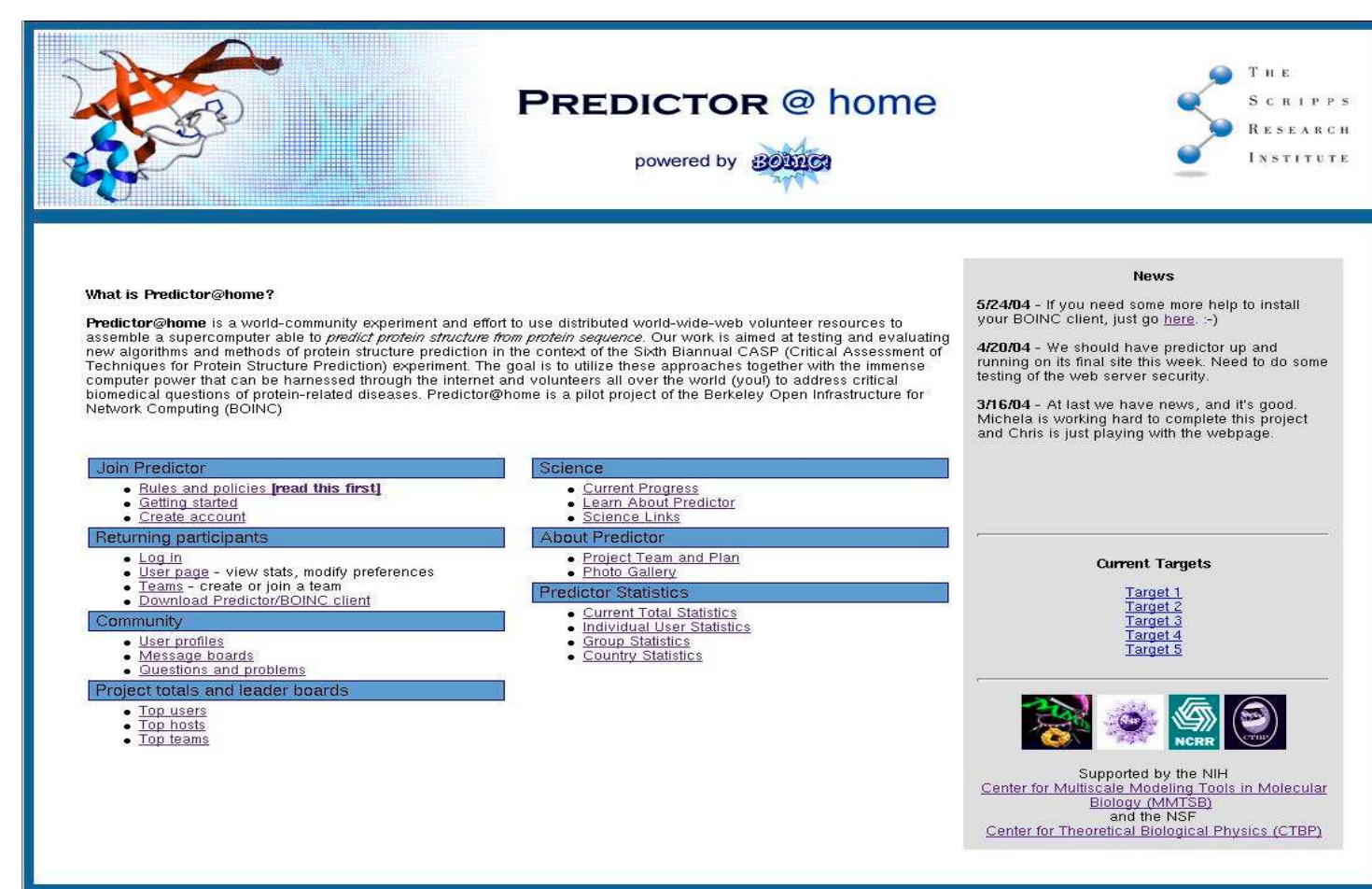
Improve upon previous protocols for protein structure prediction by augmenting conformational sampling

Our approach:

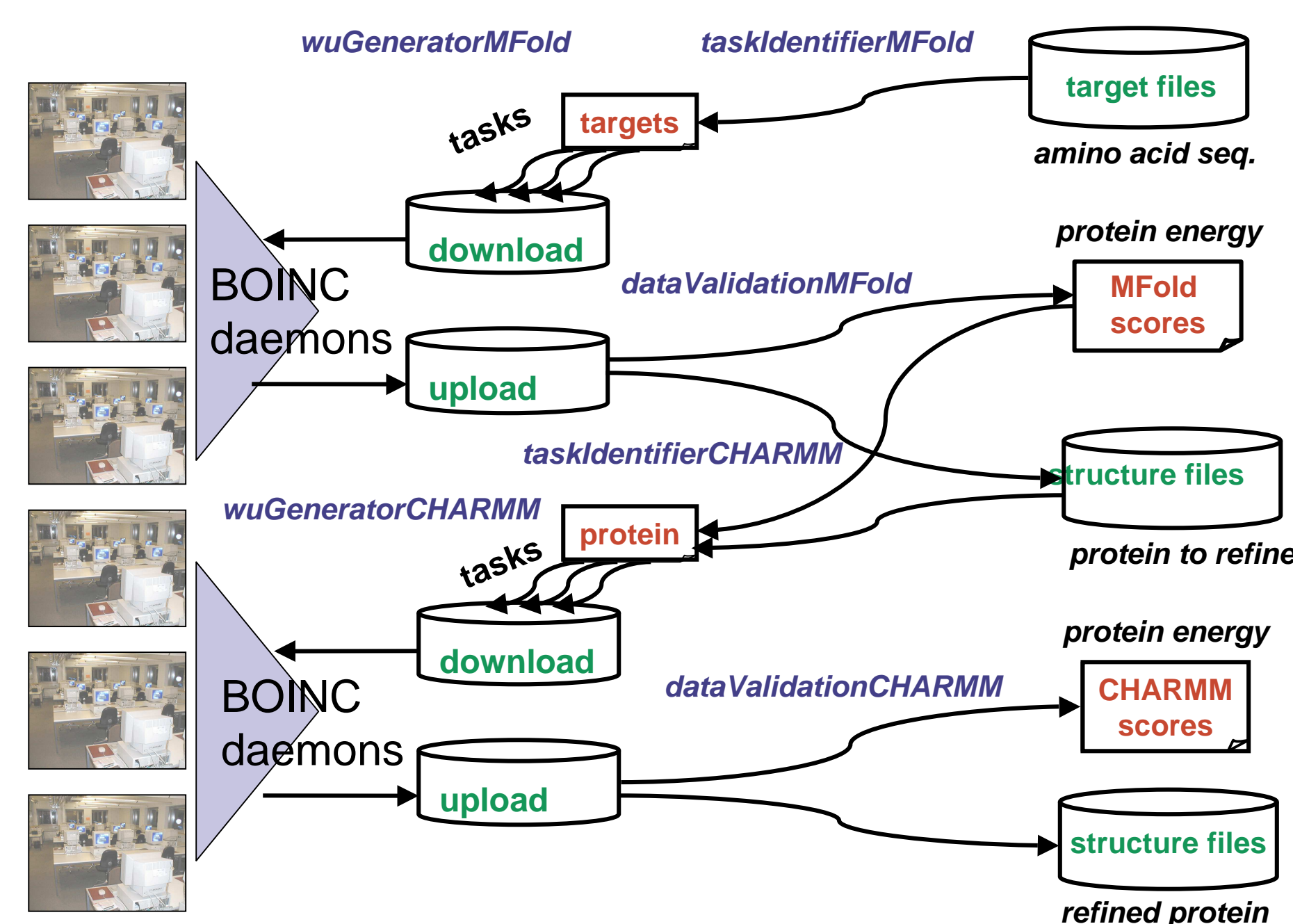
Deployment of a "structure prediction supercomputer" based on the public-resource computing paradigm

Our final goal:

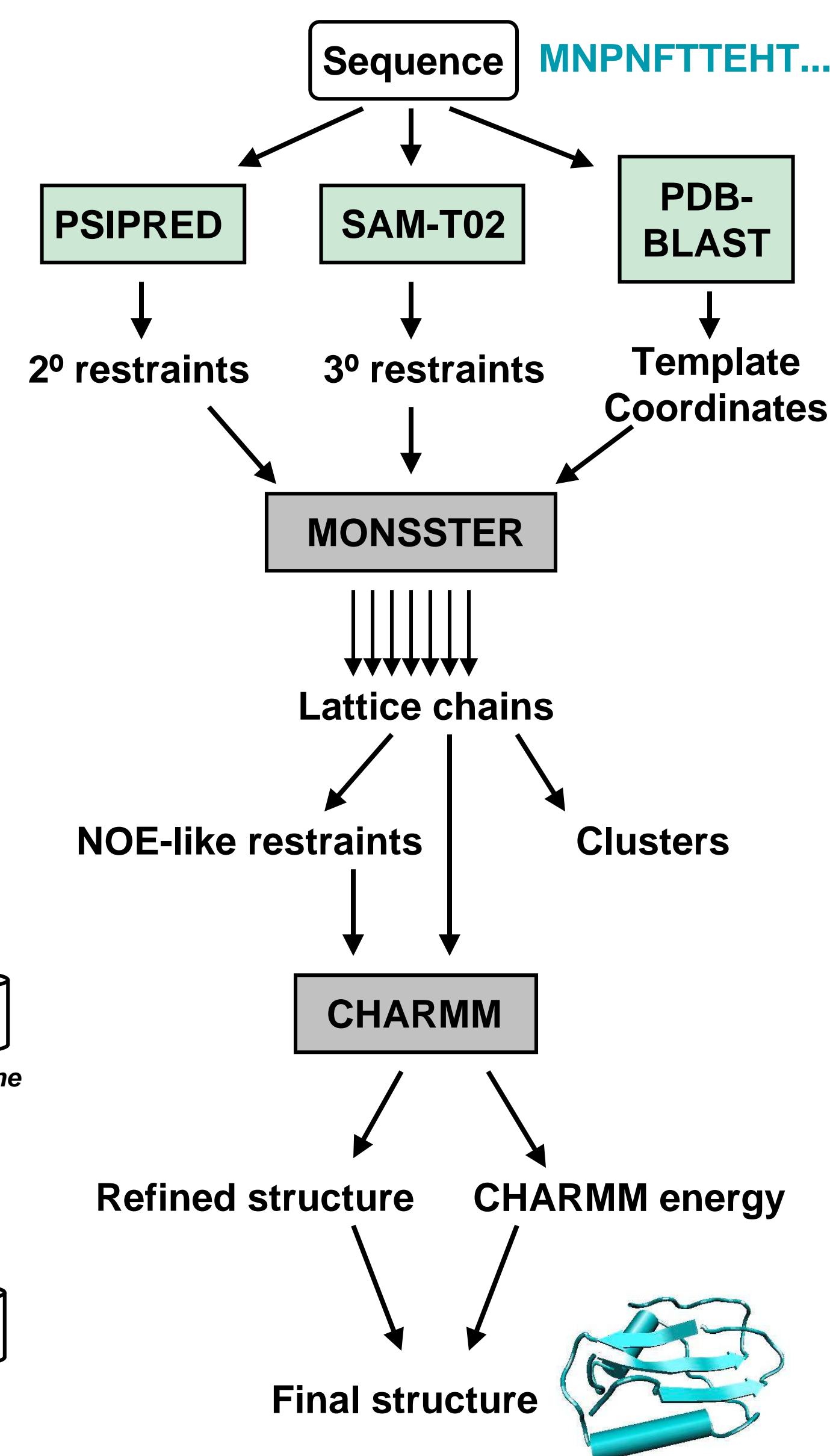
Testing the hypothesis that the significantly increased sampling afforded by distributed computing can improve structure prediction



Predictor@home Framework



Predictor@Home Pipeline



BOINC - Berkeley Open Infrastructure for Network Computing

- BOINC is an open-source platform for public-resource computing with built-in support for distributed computing on heterogeneous PCs connected to the Internet.
- Volunteers donate unused cycles of their personal computers by installing the BOINC client and attaching a project
- This platform is used by other public projects such as SETI@home, Climateprediction.net, and Einstein@home

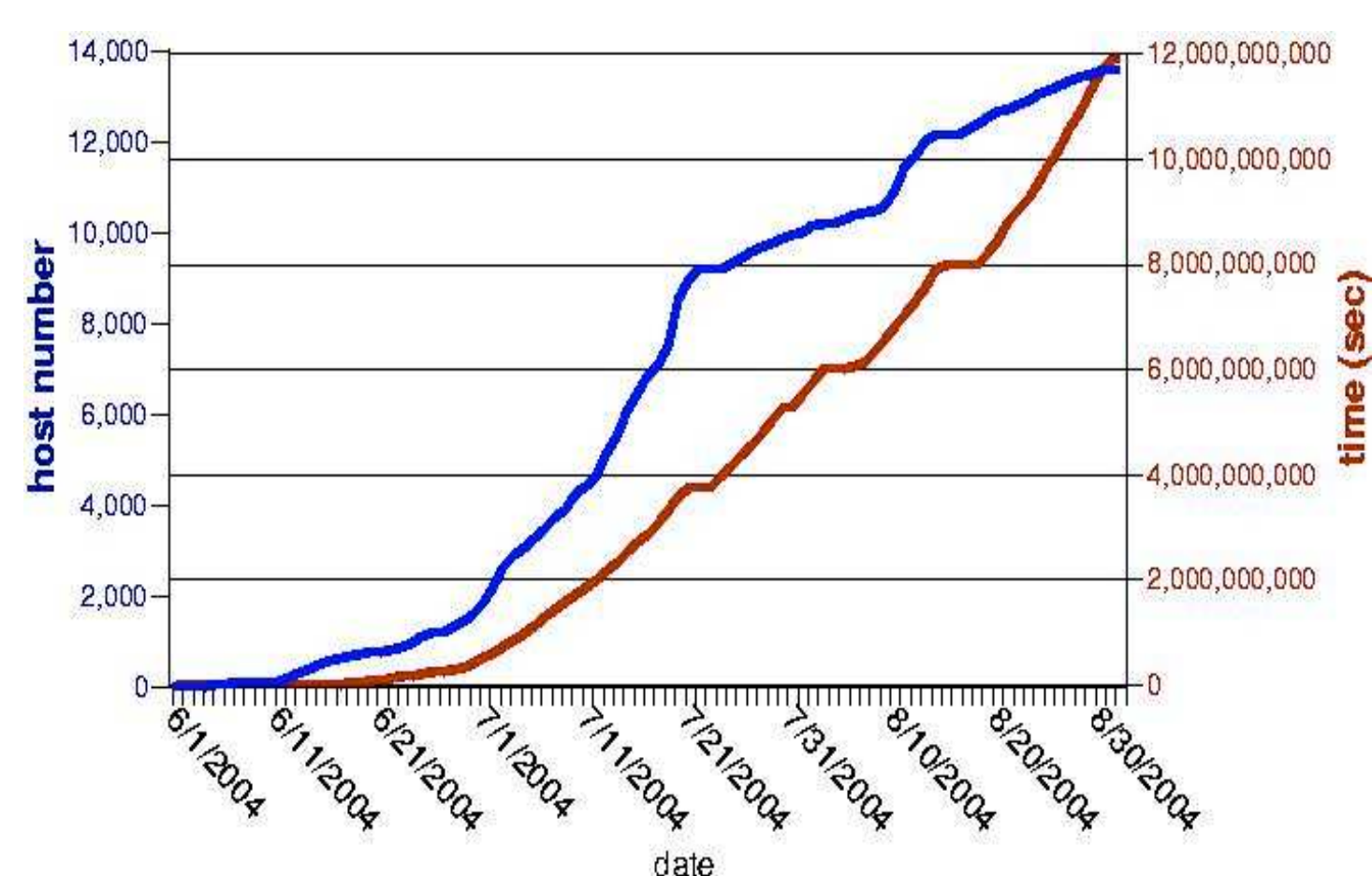
MONSSTER - Modeling of New Structures from Secondary and Tertiary Restraints

- Uses a Monte Carlo Algorithm, a simple knowledge-based force field, and the SICHO lattice model, developed by Skolnick Group
- SICHO model: "**Side CHain Only**" -- one lattice point per residue, at side chain center of mass. Each residue type has own volume and interaction properties
- The computational efficiency of sampling and the smooth protein folding landscape resulting from this model makes it ideal for distribution of conformational sampling tasks

CHARMM - Chemistry at HARvard Macromolecular Mechanics

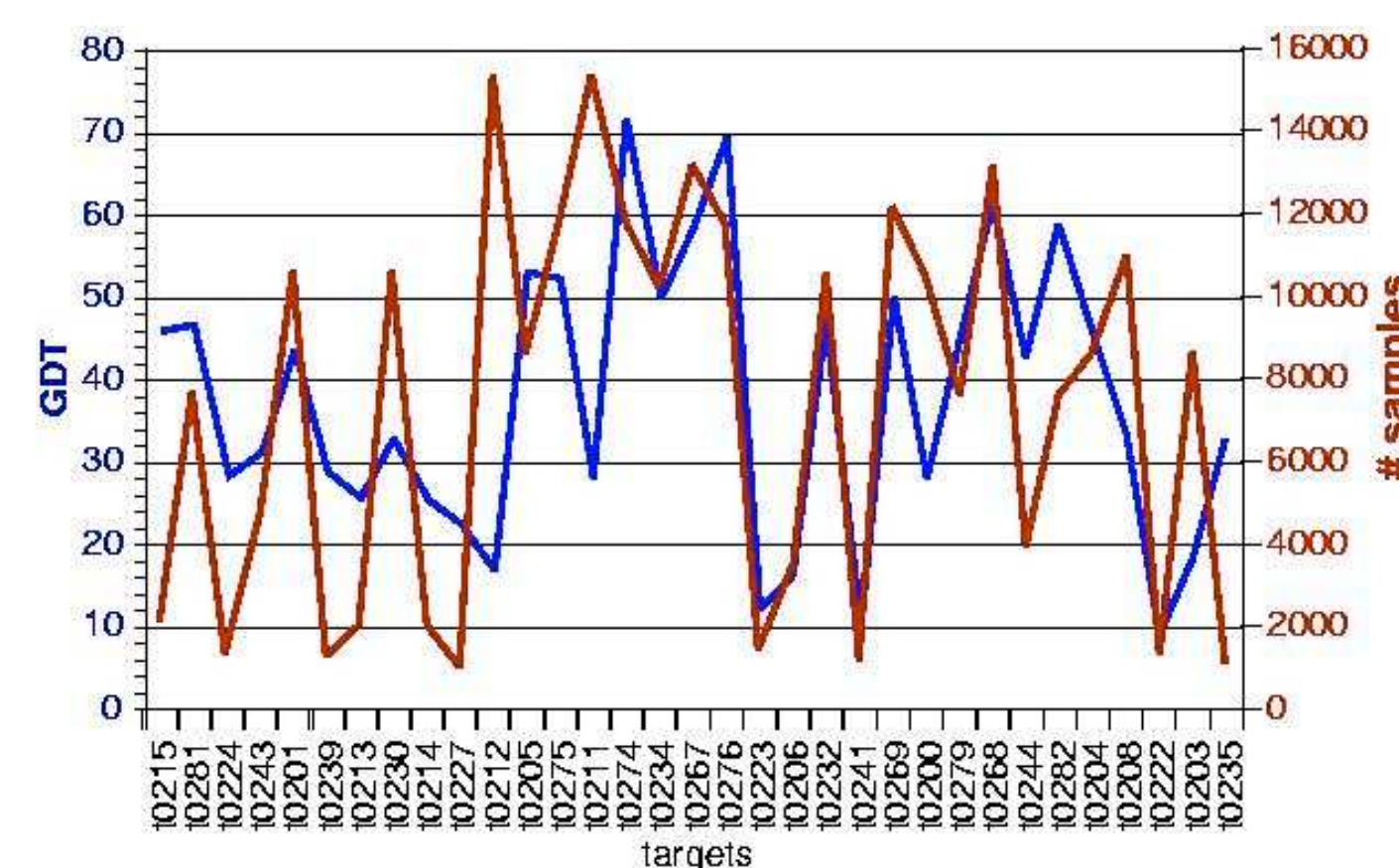
- CHARMM is used for all-atom structural refinement and scoring of the generated candidate structures.
- CHARMM uses classical mechanical methods for simulated annealing **molecular dynamics** simulations (MD) and energy minimization
- An accurate all-atom model with a physics-based force field and a Generalized Born representation of the solvent (GBMV)

Computing Time



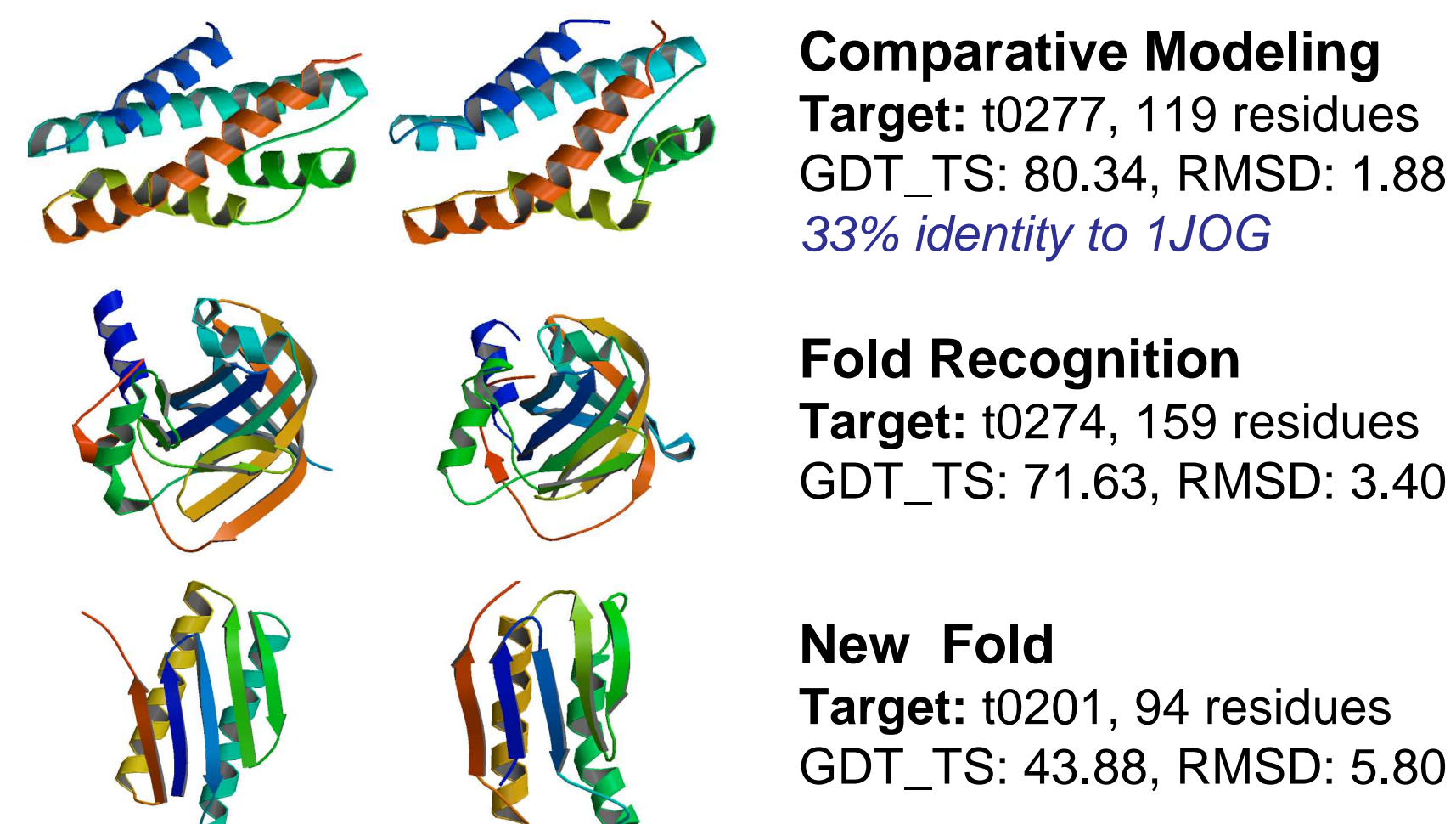
Incremental compute time for protein conformational sampling and refinement computations over CASP6

Prediction Quality vs. Sampling



Correlation between sampling quantity and prediction quality for medium difficulty targets

Structural Comparisons



Comparisons of experimental structures (left) vs. structures generated by Predictor@home (right)

Conclusions

- The public-resource computing paradigm is a powerful way for distributing structure prediction tasks:
 - Millions of workunits processed, **12 billion seconds** of computation over 3 months of CASP6
 - Integrity and security of results is guaranteed in an efficient way by deploying homogeneous redundancy and strict equality comparisons
- An increase in sampling by **1.5-2 orders of magnitude** brought about by Predictor@home has led to improved quality of structure predictions
- The structure prediction pipeline can still benefit from improvement, particularly in the template selection and refinement stages