Protein structure prediction

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

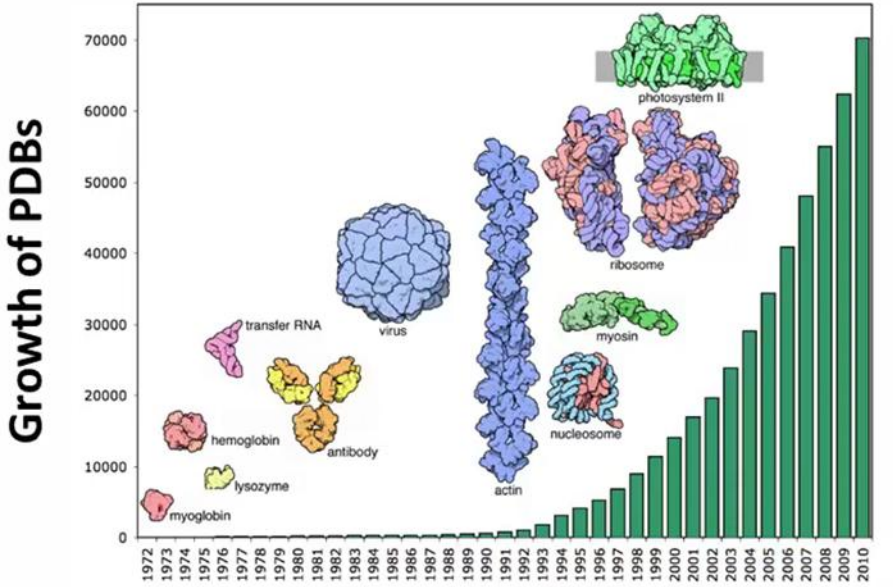
Dr. Jaspreet Kaur Dhanjal
Assistant Professor, Department of Computational Biology
Email ID: jaspreet@iiitd.ac.in

October 10, 2025

Threading or Fold Recognition

- This method heavily depends on the analysis of known protein structures.
- It is estimated there are only around 1000 to 10000 stable folds in nature.
- Irrespective of the amino acid sequence, a protein has to adopt one of these folds.
- Fold recognition is essentially finding the best fit of a sequence to a set of candidate folds.
- Select the best sequence-fold alignment using a fitness scoring function.

Unlimited structures, but limited folds



Year	structure	fold	s/f
2016	117882	1393	84.6
2015	114660	1393	82.2
2010	69605	1393	50

Threading or Fold Recognition

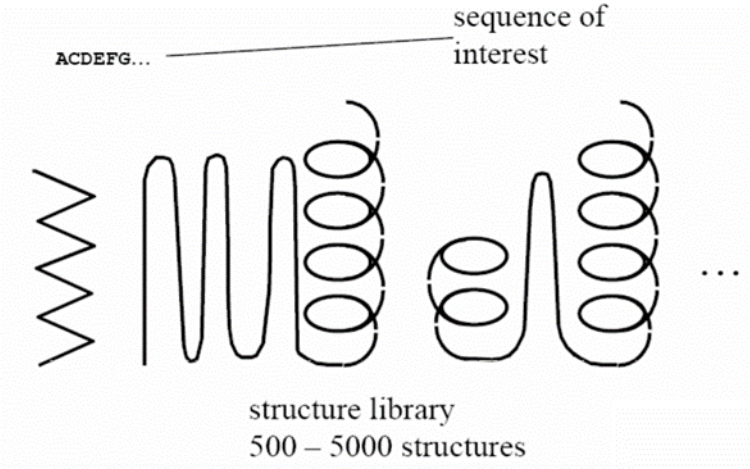
- Unlike sequence-only comparison, these methods take advantage of the extra information made available by 3D structure.
- In effect, fold prediction methods turn the protein folding problem on its head: rather than predicting how a sequence will fold, they predict how well a fold will fit a sequence.

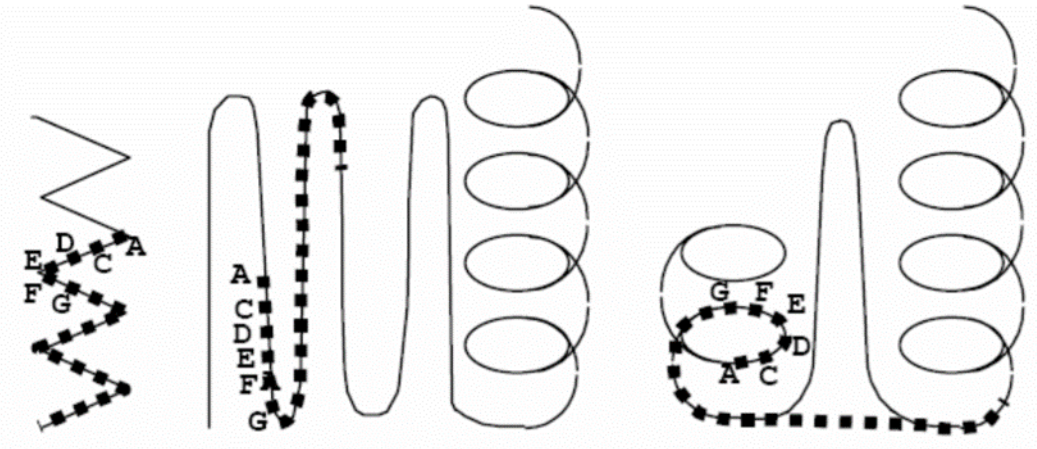
Analysis of protein structures: Microenvironment characterization

Describe structures at multiple levels of detail using a comprehensive set of properties:

Atom based properties	→	Type, Hydrophobicity, Charge
Residue based properties	→	Type, Hydrophobicity
Chemical group	→	Hydroxyl, Amide, Carbonyl, etc.
Secondary structure	→	a-Helix, b-Strand, Turn, Loop
Other properties	→	VDW volume, B-factor, Mobility, Solvent accessibility

Threading/Fold recognition – Basic strategy





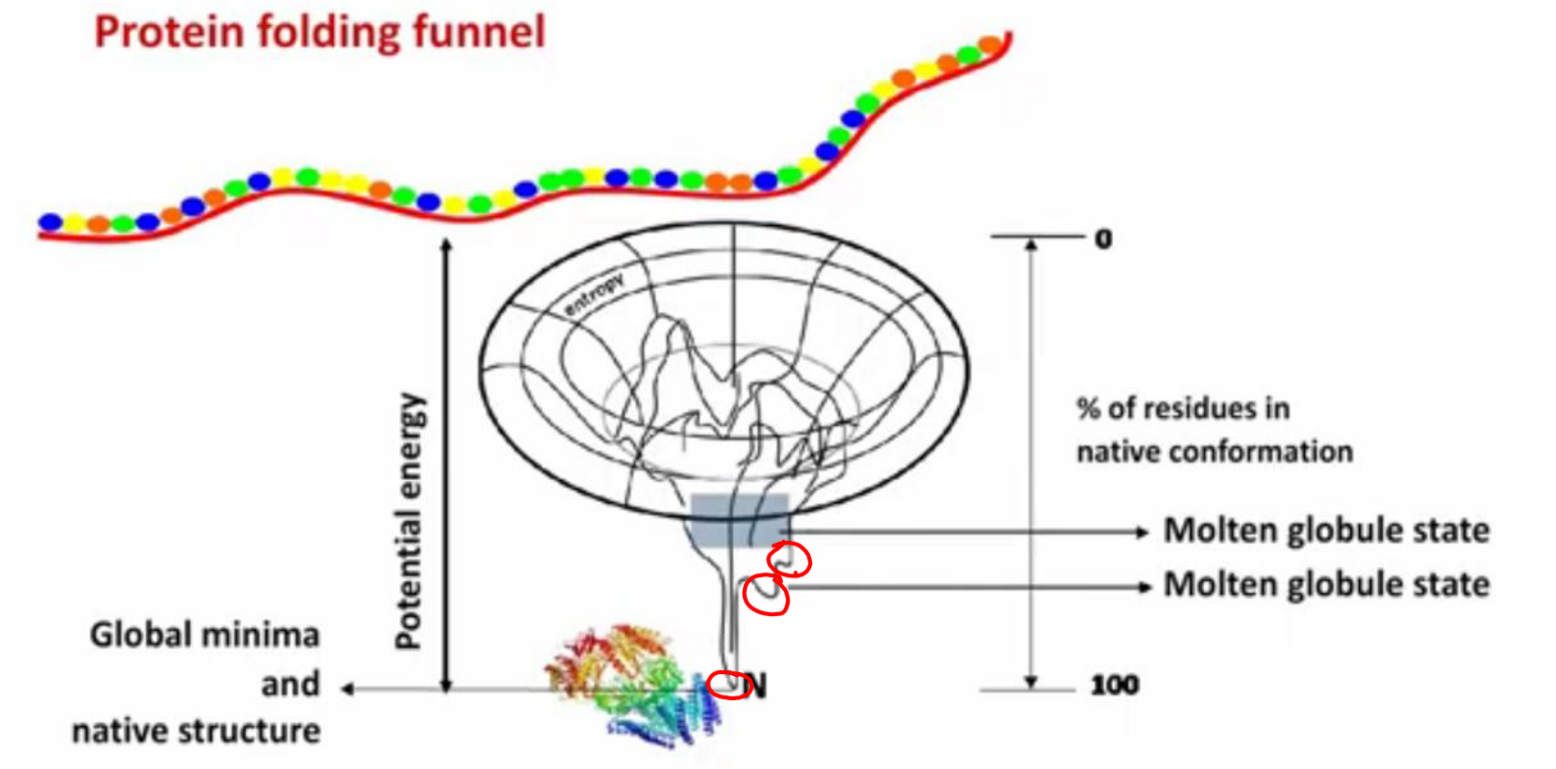
A practical approach for fold recognition

- Although fold prediction methods are not 100% accurate, the methods are still very useful.
- Run many different methods on many sequences from your homologous protein family. After all these runs, one can build up a consensus picture of the likely fold.
- Remember that a correct fold may not be at the top of the list, but it is likely to be in the top 10 scoring folds.
- Think about the function of your protein, and look into the function of the predicted folds.
- Don’t trust the alignments, rather use them as starting points.

Programs available for Threading/Fold recognition

- **Phyre:** http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
- **RAPTORX:** http://raptorx.uchicago.edu/StructurePrediction/predict/
- **PROSPECT:** http://compbio.ornl.gov/structure/prospect/
- **NovaFold:** http://www.dnastar.com/t-products-NovaFold.aspx
- **Hhpred:** http://toolkit.tuebingen.meg.de/hhpred
- **I-TASSER:** http://zhanglab.ccmb.med.umich.edu/I-TASSER

Ab initio protein modelling



Ab initio modelling protocol

- Protein representation: assign rules and restrictions to be applied
- Energy functions: defines total potential energy of atoms
- Conformational search: identify global minimum energy state
- Modal selection: choose the best native-like structure from a pool of decoy structures

Protein representation

- Protein is represented as three dimensional coordinates in space or set of dihedral angle pairs.
- Torsion angles are restricted to a finite set of values
- Hydrogen atoms are given importance in polar residues
- Restrictions are laid for non-polar hydrogen atoms; only a finite set of dihedral angles are considered, and bulky side chains are replaced by single pseudo-atoms.

Function for potential energy calculations

Bond stretching

Angle bending

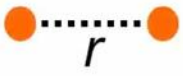
Dihedral torsion

$$V_{\text{total}} = \sum_{\text{bonds}} K_b (r - r_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\phi [1 + \cos(n\phi - \gamma)]$$

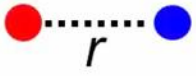
$$+ \sum_{\text{Hbonds}} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + \sum_{\substack{\text{van der Waals} \\ i,j \text{ pairs}}} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum_{\substack{\text{electrostatictic} \\ i,j \text{ pairs}}} \frac{q_i q_j}{\varepsilon r_{ij}}$$

H-bonding

van der Waals



Electrostatic



A_{ij}, B_{ij} are Lennard Jones parameters, ij: couple of atoms; q_i, q_j: Coulomb Charge parameters; K_b, K_q, K_ϕ: constants, r: distancs; t: time

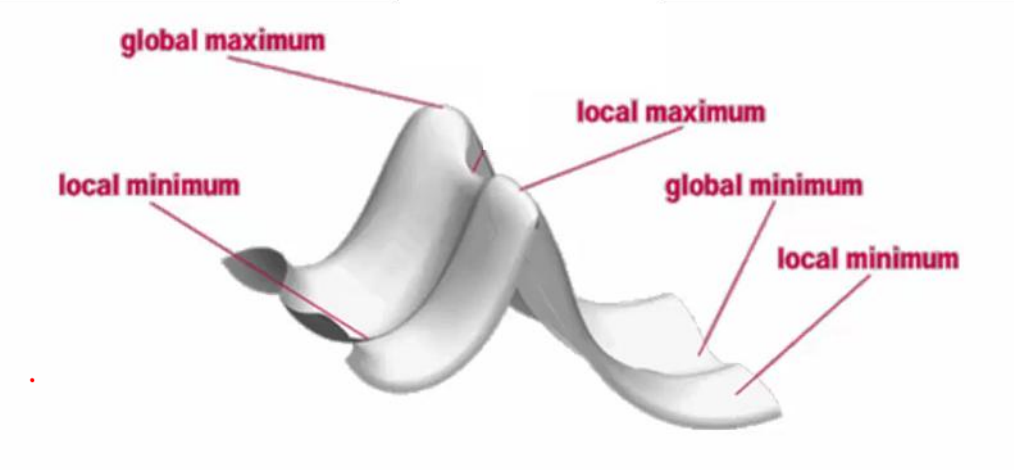
Conformational search

- To produce possible conformations
- Search potential energy surface and locate the global minimum (native)

Methods:

- Energy Minimization
- Monte Carlo
- Genetic Algorithm
- Molecular dynamics
- Simulated annealing

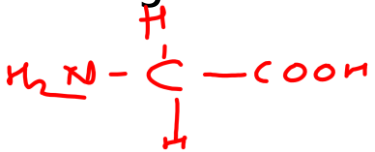
Protein Energy Landscape

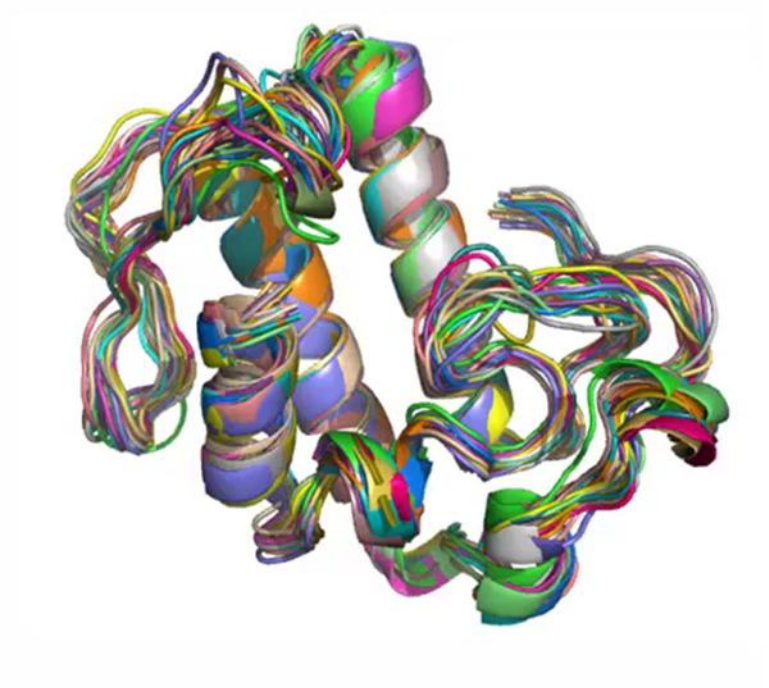


Challenges

Loop structure prediction:

- Loops are vey integral for protein folding and protein function
- Loops exhibit greater structural variability than helices and strands
- Loop prediction is often a limiting factor on fold recognition methods





Programs available for *Ab initio* protein modelling

- **ROBETTA:** http://robetta.bakerlab.org

De novo Automated structure prediction analysis tool used to infer protein structural from protein sequence data

- **PROTINFO:** http://protinfo.compbio.washington.edu

De novo protein structure prediction web server utilizing simulated annealing for generation and different scoring functions for selection of final five conformers

- **SCRATCH:** http://www.igb.uci.edu/servers/psss.html

Protein structure and structural features prediction server which utilizes recursive neural networks, evolutionary information, fragment libraries and energy

- **ASTRO-FOLD:** http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1303441/

Astro-fold: first principles tertiary structure prediction based on overall deterministic framework coupled with mixed integer optimization

- **ROKKY:** http://www.proteinsilico.org/rokky/rokky-p/

De novo structure prediction by the simfold energy function with the multi-canonical ensemble fragment assembly

- **BHAGEERATH:** http://www.scfbio-iitd.res.in/bhageerath

Energy based methodology for narrowing down the search space of small globular proteins

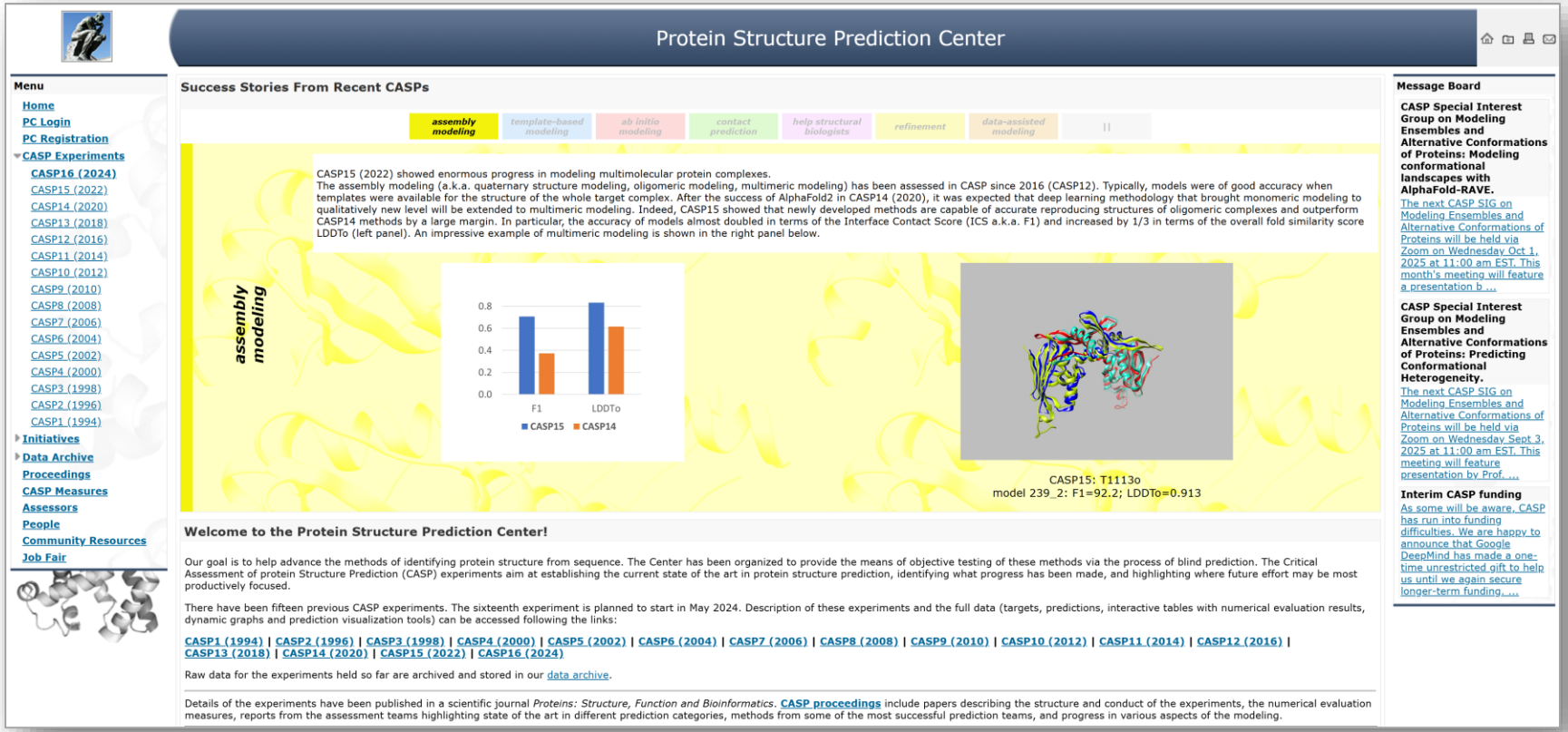
- **CHARMM:** https://www.charmm.org/charmm/

(Chemistry at HARvard Molecular Mechanics) is a program for macromolecular simulations, including energy minimization, molecular dynamics and Monte Carlo simulations

- **NAMD:** http://www.ks.uiuc.edu/Research/namd/

is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems

CASP (Critical Assessment of Structure Prediction)



CASP (Critical Assessment of Structure Prediction)

Detailed description of the experiment

CASP (Critical Assessment of Structure Prediction) is a community wide experiment to determine and advance the state of the art in computational structural biology. Every two years, participants are invited to submit models for a set of macromolecules and macromolecular complexes (proteins, RNA, ligands) for which the experimental structures are not yet public. In the latest CASP round, CASP16 in 2024, nearly 100 research groups from around the world submitted more than 80,000 models on 100+ modeling entities yielding 300 targets in five prediction categories. Independent assessors then compare the models with experiment. Assessments and results are published in a special issue of the journal PROTEINS (check the latest CASP15 issue here).

Goals | Categories | Timetable | Registration | Targets | Format | Assessment | Results | Conference | Organizers

Background and goals

The goals of CASP are to provide rigorous assessment of computational methods for modeling macromolecular structures and complexes so as to advance the state of the art. Recent CASPs saw enormous jumps in the accuracy of computed structures, first in CASP14 (2020) for single proteins and domains, with many models competitive in accuracy with experiment, and second, in CASP15 (2022), with a large increase in the accuracy of protein complexes. These advances are primarily the result of the successful application of deep learning methods, particularly AlphaFold2 and other methods built around it. Major interest in the field is now centered around the further potential impact of deep learning methods. In response to that, in the 2022 CASP15, modeling categories were realigned. We are maintaining those categories in CASP16. We will also continue our close collaborations with CAPRI (for protein complexes) and RNA puzzles (for RNA structure).

Modeling categories

CASP16 categories are as follows:

- **Single Proteins and Domains**
  As in previous CASPs, the accuracy of single proteins and where appropriate single protein domains will be assessed, using the established metrics. The major emphasis is now on the fine-grained accuracy of models, whether limitations related to sequence alignment depth and target size are surmounted, and whether interdomain relationships are accurately captured. There is also interest how well the many new deep learning methods perform, including those using large language models.
- **Protein Complexes**
  As in recent CASPs, the ability of current methods to correctly model subunit-subunit and protein-protein interactions will be assessed. We will again work in close collaboration with our CAPRI partners. There was enormous progress in this category in the last CASP, but accuracy was not yet as high as for single proteins, so there is substantial room for a further advance. New in this CASP is the option of predicting stoichiometry. Where possible, targets will initially be released without that information, models collected, followed by re-release with that data provided.
- **Accuracy Estimation**
  Members of the community will again be invited to submit accuracy estimates for multimeric complexes and inter-subunit interfaces provided by others. There is no longer a category for general methods of estimating single protein structure accuracy, since in recent CASPs estimates provided by model builders have been consistently more reliable. However, there will be an emphasis on the reliability of accuracy estimates provided with submitted structures, both overall and at the individual amino acid level. Note that all accuracy estimates are in plddt units, not Angstroms.
- **Nucleic acid (NA) structures and complexes**
  An RNA structure category was introduced in the previous CASP and the results were interesting and provocative. In particular, it appeared that deep learning methods were not yet as effective as more traditional ones for this type of macromolecule. Has that now changed? This CASP we expect to include RNA and DNA single structures and complexes, and complexes of these with proteins.
- **Protein - organic ligand complexes**
  The last round of CASP included this category for the first time. Results indicated that, as with RNA structure, deep learning methods were not yet competitive with more traditional approaches. So there is considerable interest in whether that has now changed. In addition to ligands integral to protein targets, we expect to have several target sets related to drug design.
- **Macromolecular conformational ensembles**
  Following the success of deep-learning methods for single structures, it is increasingly important to assess methods for predicting structure ensembles, and CASP included this category for the first time in 2022. While it was clear deep learning methods have considerable potential for generating ensembles, the best procedures are still hotly debated with many new papers appearing. In CASP16, we expect to have a variety of targets for both protein and RNA ensembles.
- **Integrative modeling**
  Deep learning methods combined with sparse experimental data such as SAXS and chemical crosslinking are now being used extensively to obtain the structure of large marcomolecular complexes. To assess effectiveness of these approaches, CASP is reintroducing this category of modeling, provided appropriate targets will be available.

Timetable

- April 2, 2024 - Start of the registration for CASP16 prediction experiment.
- April 16, 2024 - Start of the testing of server connectivity ("dry run" for server predictors).
- May 1, 2024 - Release of the first CASP16 modeling targets.
- June/July 2024 - Early bird registration for the December CASP16 conference.
- July 31, 2024 - Last date for releasing targets.
- August 31, 2024 - End of the modeling season.
- Early September 2024 - Collection of abstracts describing the methods used in CASP16.
- August-October 2024 - Evaluation of predictions.
- November 2024 - Invitations to groups with the most accurate models and the most interesting methods to give talks at the CASP16 conference.
- November 2024 - Program of the conference finalized.
- December 2024 - CASP16 Conference (tentatively, December 1-4).

Registration