Comprehensive Lecture Summary: Computational Modeling of Protein Structure and Interactions

Tahamid Siam (us79)

May 1, 2026

Color Key:
Red = Source 1 (Summary 1: Docking & Scoring)
Yellow = Source 2 (Summary 2: AlphaFold)
Green = Source 3 (Summary 3: PPI & Deep Learning)
Blue = My Additions (Bridging Logic)

The grand challenge of modern computational biology lies not only in determining the shapes of isolated proteins but in understanding how they physically assemble to execute cellular functions.

The protein folding problem involves elucidating a protein's 3D structure directly from its amino acid sequence. This is immensely complicated because a given sequence could have multiple low-energy folded representations, and structures depend heavily on environmental conditions. Historically, progress was driven by three main wet-lab techniques: X-ray crystallography (analyzing diffraction patterns of crystals), Nuclear Magnetic Resonance (NMR) spectroscopy (measuring electromagnetic shifts to determine constraints), and Cryo-electron microscopy (Cryo-EM). While these methods have successfully populated the Protein Data Bank (PDB) with over 200,000 structures, they remain highly costly and difficult. Furthermore, proteins rarely act alone; computational modeling of protein-protein interactions (PPIs) matters because these interactions drive core cellular processes, their disruption is linked to disease, and their interfaces are increasingly viewed as tractable targets for therapeutic modulation.

The scale of this biological problem is staggering, with a lower bound estimate of roughly 50,000,000 PPIs in the human proteome alone. This mismatch between experimental capacity and biological reality creates a major need for computation to answer four core questions: Do proteins A and B bind? How strongly do they bind? Where do they bind? And what bound conformations do they adopt?

To bridge this gap, the field has increasingly relied on evolutionary data and the rapid advancement of deep learning architectures. Before end-to-end deep learning became dominant, genomic logic—such as gene fusion signals and interologs (interactions known in one species projected onto homologs in another)—was used to externalize the idea that evolution preserves functional partnerships. Coevolutionary models operate at a finer resolution using Direct-Coupling Analysis (DCA). By applying maximum-entropy principles to paired multiple sequence alignments (MSAs), DCA identifies compensatory

mutations, successfully separating direct residue couplings from indirect correlations to infer physical contacts.

This heavy reliance on MSAs set the stage for the AlphaFold revolution, catalyzed by the Critical Assessment of Structure Prediction (CASP) competition. AlphaFold1 initially cast structural prediction as an "image recognition problem," applying a convolutional neural network (CNN) to an input matrix to predict pairwise contact probabilities, which were then used to minimize an energy function. AlphaFold2 optimized this by framing it as a "language processing problem." It utilizes a Transformer-based Evoformer to process MSA and pairwise information directly into a structure block for inference. More recently, AlphaFold3 integrated a diffusion module, allowing for improved resolution of complex molecular assemblies, including protein-ligand structures. Today, AlphaFold-style co-folding can generate acceptable-quality structures for 50-70% of stable heterodimers, representing a massive paradigm shift where networks learn strong priors over protein geometry, allowing sampling to begin much closer to plausible complexes.

However, proposing a plausible geometry via deep learning is only one part of the pipeline; rigorous physical and thermodynamic validation remains essential. A practical mental model splits PPI modeling into three stages: partner prediction, structural modeling, and energetic interpretation. In classical physics-based docking, bound complexes must exhibit shape complementarity and favorable energetics. FFT-based rigid-body docking makes this practical by turning an expensive 6D search into fast correlation calculations. Rosetta Dock builds on this with multiscale Monte Carlo searches—using standard Metropolis acceptance steps ($p_{accept} = \min(1, e^{-\Delta E/kT})$)—to refine side-chains. When experimental data is sparse, tools like HADDOCK convert interface information from NMR or mutagenesis into energetic restraints, severely shrinking the search space.

This computational docking workflow fundamentally separates into two distinct challenges: sampling (generating candidate decoy complexes) and scoring (ranking those decoys so near-native structures appear better than incorrect ones). A major hurdle in this process is that proteins undergo bound-unbound structural changes during interaction. This conformational change is quantified using Root-Mean-Square Deviation (RMSD), which compares matched atoms only after an optimal rigid-body alignment (like Kabsch-alignment) removes irrelevant global rotations and translations.

Ultimately, assessing the quality of these predicted structural interactions requires rooting the models in fundamental thermodynamics. Binding affinity measures how strongly proteins associate, governed by the thermodynamic relationship:

$$\Delta G^{\circ}_{bind} = -RT \ln K_a = RT \ln K_d$$

Stronger binding corresponds to a smaller dissociation constant (K_d) and a more negative binding free energy (ΔG). To evaluate if scoring functions align with these realities, researchers rely on DockQ, a continuous quality score ranging from 0 to 1. DockQ mathematically combines the fraction of native contacts recovered (F_{nat}), interface RMSD, and ligand RMSD, applying a mapping function so that it smoothly reproduces the categorical CAPRI classes (Incorrect, Acceptable, Medium, and High). A successful scoring function should exhibit strong negative Pearson or Spearman correlation with DockQ, meaning lower energy scores reliably track with higher structural accuracy across an entire set of decoys.

Post-processing pipelines often deploy Molecular Dynamics (MD) or end-point approximations like MM/PBSA ($\Delta G_{bind} \approx \Delta E_{MM} + \Delta G_{solv} - T\Delta S$) to evaluate whether

an interface remains intact and to identify "hot spots"—interface residues that contribute disproportionately to binding free energy.

Despite these immense theoretical and computational strides, severe limitations continue to bottleneck the field. Scoring difficulty is highly dependent on the physical geometry of the target interface. Interfaces that are heavily intertwined or contact-rich provide strong geometric constraints, making them easier to accurately score. Conversely, flat interfaces are highly ambiguous because countless incorrect translations or rotations still look superficially plausible. Furthermore, deep learning models like AlphaFold are highly constrained by MSA depth; shallow alignments yield inaccurate predictions. They also struggle with the "protein design problem" (generating sequences for a desired structure) and cannot easily model the impact of single point mutations. Critically, these models are trained on in vitro ground truths from crystallography or Cryo-EM, which often fail to reflect the true physiological in vivo structures dictated by cellular crowding and non-steric effects.

In short, the field is that there is no universally best PPI method. Performance remains poor for transient interactions and highly disordered systems. The frontier of computational structural biology relies not on a single algorithm, but on integrative workflows that combine learned deep-learning geometric priors, explicit molecular physics, curated evolutionary evidence, and continuous orthogonal experimental validation.