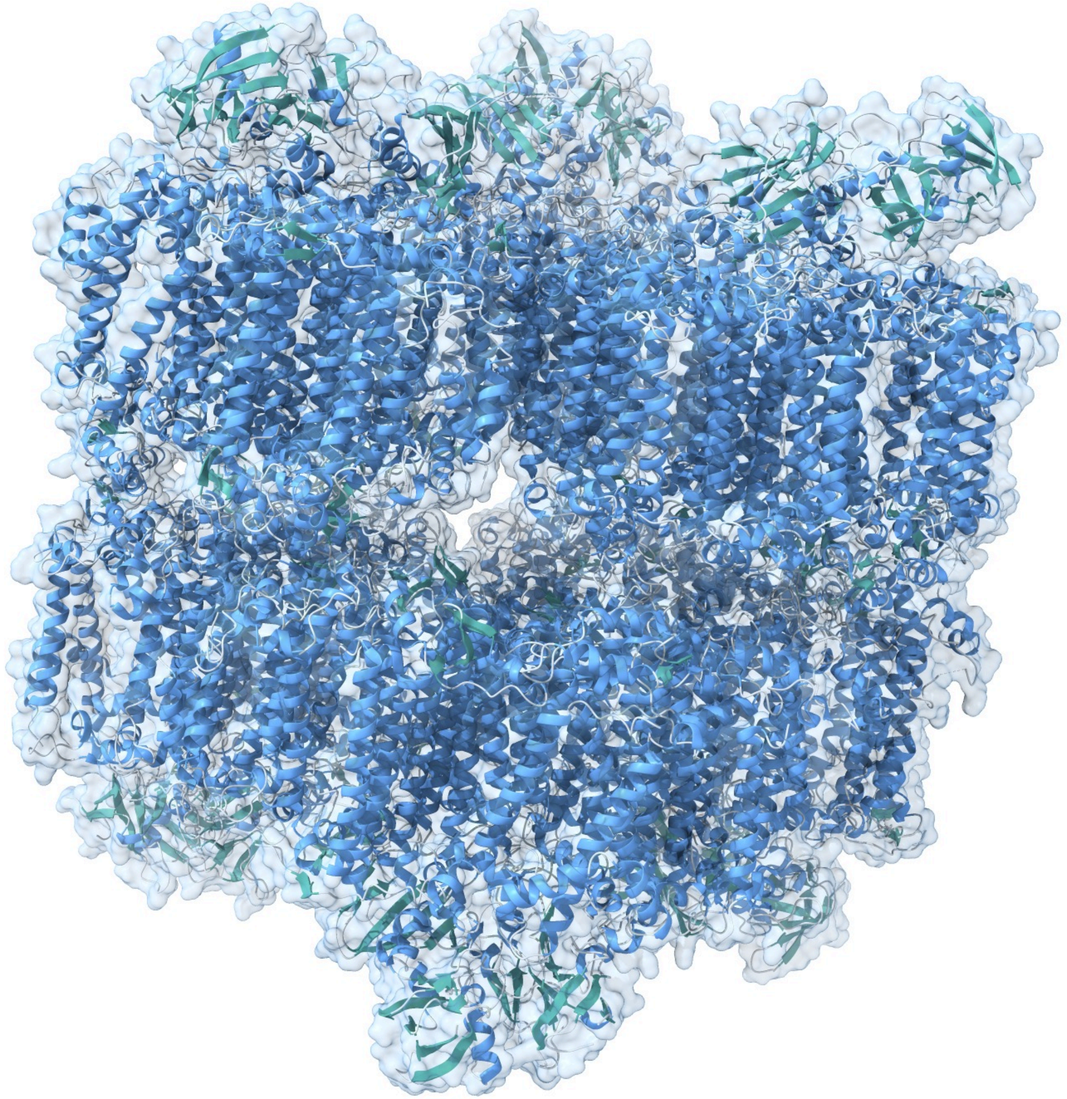
2026 UPDATE:

A Comprehensive Introduction to AI for Proteins

(Almost) Everything You Need to Know To Get Started with Computational Protein Design



Read the full article: www.tamarind.bio/blog/a-comprehensive-introduction-to-ai-for-proteins

Structure Prediction & Docking

- AlphaFold2: The progenitor of the explosion in BioML advances in recent years. AlphaFold2 blew away every other methodology on the CASP benchmark, both for single-chains and complexes, enabling new applications for structure-based work.
- Since AF2, DeepMind has released AlphaFold3, but with no commercial use allowed. As an alternative, many groups have released reproductions from the authors' paper. These are mostly similar in performance, here are some of the nuances:
  - Chai-1: Restraints allow for describing binding sites between proteins, supporting substantially increased docking accuracy, may be useful for Igs
  - Boltz-2/1x: Specific improvements for physical accuracy of protein-small molecule complexes, may be useful for enzymes or small molecules targeting proteins, can also predict binding affinity to small molecules
- AlphaRED: Predicted structures combined with Rosetta-based physics approaches for improved docking pose quality.
- Structure predictors specialized for specific tasks or modalities are often quicker or more accurate for their respective niches, tools like PLACER, AbodyBuilder, TCRModel2, and others will be discussed in their own sections
- Along with the actual structures produced for a given sequence, models like AlphaFold also provide various different confidence metrics, which can serve to predict the fitness and stability of the sequence's fold.

De novo design

- Generation of a protein from close to nothing as input, not quite function, but typically a target molecule e.g. another protein, or a small molecule. The majority of these protocols do also require a structure for the target of interest

- Miniprotein binders: An up-and-coming modality of proteins ~50-200 residues long, that are uniquely successful in the de novo design paradigm. If a design passes in silico thresholds of tools like BindCraft, 10-20 can have 10-100% hit rates with median KD 1-30 nM after single-round expression. RFdiffusion+MPNN+AlphaFold is another protocol to watch.

- Antibodies: De novo design for Ig or VHH binders are less established, especially relative to minibinders. However, we've seen substantive improvements, with reliably up to 10-20% hit rates against diverse targets from tools like Boltzgen, Germinal, RFantibody, mBer and more.

- Peptides: Structure-based methods for linear and cylic peptides including Boltzgen, BindCraft, PXDesign, RFpeptides and more have shown very valuable immediate hit rates.

Optimization with Inverse Folding

- “Inverse” of a structure prediction task, going from structure to novel sequence that is predicted to fold into that structure. Note that the input can be a model structure made by the aforementioned structure prediction tools.
  - ProteinMPNN: Given a list of indices from the input structure, replace those positions with residues predicted to be fitter, while maintaining the same structure. These tend to be more soluble, stable, and express well due to the training data having these attributes.
  - SolubleMPNN: A specialized version of ProteinMPNN on soluble proteins, interesting results in GPCR solubilization and general fitness optimization.

- Antibodies
  - AntiFold & IgDesign: Specialized offshoots from ProteinMPNN for antibody-antigen complexes to replace CDRs for improved binding affinity.

- Stability
  - ThermoMPNN
    - Scores every possible point mutation for its effect on stability (ddG). Good quantitative results in various benchmarks.
      At Tamarind.Bio we’ve anecdotally found the predictions to correlate well with our users’ wet lab results as well.
  - HyperMPNN
    - ProteinMPNN trained on hyper-thermophiles, biased towards more stable variants.

Active Learning

- Lab-in-the-loop: Genentech/Prescient Design's approach to incorporating wet lab data into custom, specialized antibody sequence generators.
  - 1,800 variants over four rounds against EGFR, IL-6, HER2, OSM, starting from animal-immunization and repertoire-mined leads
  - Performance: Every target yielded 3-100× affinity gains; ten leads hit ~100 pM KD—well within therapeutic range—after only four design/test cycles.
  - NOS/LaMBO-2 and DyAb are the generative steps, along with many developability predictors at each round, trained on data.
- Active Learning-assisted Directed Evolution (ALDE) combines uncertainty-aware machine-learning models with iterative wet-lab screening to navigate epistatic sequence space far more efficiently than conventional DE.
  - In three rounds, ALDE optimized five active-site residues of a cyclopropanase to raise product yield from 12 % to 93 %, and simulations across public datasets indicate it consistently outperforms standard DE strategies.
- EVOLVEpro
  - Requiring only ~10 assays per cycle to map sequence to function and drive multi-objective optimization. In benchmarks and six diverse experimental campaigns (antibodies, CRISPR nuclease, prime editor, serine integrase, T7 RNA polymerase), it delivered up to 40 to 100-fold performance gains—decisively outclassing zero-shot PLM guesses and conventional directed evolution, and showing that PLM-guided, data-sparse iteration is now the method to beat.

Antibodies

Affinity Maturation

- Inverse Folding
  - As discussed, Inverse Folding involves starting from a protein (complex) structure, assigning residues to replace, and fills in those positions while maintaining of the starting structure/function. This often results in more “fit” proteins in terms of stability and expression due to the protocols being trained on solved structures.
  - IgDesign & Antifold replace CDRs given an antibody-antigen complex
    - This often yields better binders than the original, though there’s still a need for experimental validation of ~100 binders
    - A common strategy in deploying these protocols is to generate a very large number of sequences in silico and picking the top ~100. I.e. you might do a million sequences and select the top scoring ones for wet lab validation.
- Language Models
- Similarly, sequence-based models such as ESM, AntiBERTy, AbLang can replace(known as masking) arbitrary residues
- Since language models are trained on sequences found in nature, they can tend to suggest germline mutations.
- Efficient Evolution: Language model which suggests point mutations to improve binding affinity. Tested on antibodies but may work for other proteins. The authors tested 20 or less designs over 2 rounds for 7 antibodies, finding up to sevenfold improvement for 4 mature antibodies and up to 160-fold improvement for 3 unmature antibodies.
- Combination of Inverse Folding+Language Models
  - An interesting result released recently is a combination of AbLang and ProteinMPNN for higher fitness sequence design processes.
  - Language models often revert sequences to the germline, whereas inverse folding tools tend to stick to a relatively conservatives set of residues to maintain the original structure.
  - The authors test 96 trastuzumab variants with CDRH3 loops redesigned with the method and found that it generated thirty-six HER2 binders, compared to three out of 96 designs generated by ProteinMPNN alone
- Active learning: see previous

Antibodies

Structure Prediction

- AlphaFold/Chai/Boltz/OpenFold are still the standard for AbAg complex structure prediction. As of the writing of this post, AlphaFold3+Rosetta is the highest quality antibody-antigen docking protocol available.
- ImmuneBuilder: A collection of the Deane Lab (Oxford)'s Immune Protein Structure Prediction Tools
  - AbodyBuilder2: Predict the structure of the VH-VL chains not bound to an antigen
  - NanobodyBuilder: VHH, single-chain structure prediction
  - TCRModel2: T Cell Receptor structure prediction

Antibodies

De novo design

- There has been a recent explosion of de novo design tools for antibodies along with many other protein modalities. We saw these protocols go from one in ten thousand tested being a hit to double-digit success rates.

- BoltzGen: designs proteins, peptides, nanobodies, and other molecules that bind to target structures or small molecules. Achieves 60-70% success rates for nanobody and protein binders against novel targets with only 15 designs tested.

- mBER is an open-source protein design framework specifically engineered for antibody-format binder design. By leveraging structural templates and sequence conditioning within the ColabDesign framework, mBER enables backpropagation-based design through AlphaFold-Multimer to produce high-affinity VHH (nanobody) binders.

- Germinal: A generative framework for designing high-affinity antibodies against specific protein epitopes. Unlike general protein design tools, Germinal is optimized specifically for antibody-format binders, creating functional complementarity-determining regions (CDRs) onto user-specified frameworks while preserving favorable therapeutic developability profiles.

- RFantibody: The authors show a validated, influenza-targeting VHH, along with scFvs (for both light and heavy chains) to TcdB and a Phox2b peptide-MHC complex. With this, de novo design of antibodies targeting specific epitopes becomes possible, albeit with some needed affinity optimization after the initial round of generated binders.

- JAM from Nabla Bio, along with Chai-2 from Chai discovery have published results in de novo design of antibodies using their proprietary platforms

Antibodies

Developability & Scoring

- Machine learning: The next generation of developability predictors tend to combine physical properties for an input mAb to feed into machine learning methods to evaluate developability quantitatively.
- TAP: The default approach to evaluate an antibody for developability is to extract physics-based features from a model structure (such as hydrophobic patches), and comparing those against the same features in clinical stage antibodies. If one of the five properties of the Therapeutic Antibody Profiler are not matching those of clinical stage antibodies, it is a sign developability risks.
- Immunogenicity: DeepImmuno and TLimmuno can predict immunogenicity of any peptide/HLA combination.
- Physical properties: Analyzing hydrophobic and charged surface patches can help identify residues to mutate for reducing aggregation and improving viscosity, such as using Masif for surface embeddings.
  - Viscosity: Deep Viscosity
  - Aggregation: Aggrescan3D
  - Solubility: Netsolp predicts solublity with moderate correlation

Antibodies

Humanization

- BioPhi: Statistical and computational methods to analyze sequence alignments and identify human-like regions in antibodies. By comparing the non-human antibody sequences against large datasets of human antibody sequences, BioPhi predicts which regions should be modified for effective humanization while maintaining the original antibody's binding affinity.
- Sapiens: Using deep learning models trained on human antibody data to predict optimal modifications for humanization. It focuses on preserving the structural and functional integrity of the antibody while maximizing its compatibility with the human immune system.

GPCRs & Target Engineering

- GPCR Solubilization
  - Computational design of soluble and functional membrane protein analogues
    - The authors achieve designs for "complex protein topologies and [enrich] them with functionalities from membrane proteins, with high experimental success rates, leading to a de facto expansion of the functional soluble fold space"
    - After using AlphaFold2 to continuously test different replacement sequences, the authors feed these predicted structures to Soluble ProteinMPNN and find that the sequences produced via MPNN show significant experimental success.
- WRAPS / AI-designed nano disc alternative proteins
  - Another approach to this problem is to design proteins that replicate the effect of detergent, i.e. to keep the protein stable and water soluble in the absence of the cells's lipid bilayer. The authors introduce the de novo protein category "Water-soluble RFdiffused Amphipathic Proteins".
- Stability
  - ThermoMPNN Scores every possible point mutation for its effect on stability (ddG). Good quantitative results in various benchmarks. At Tamarind.Bio we've anecdotally found the predictions to correlate well with our users' wet lab results as well.
- Miniprotein binder design
  - An aspect of de novo design we discussed previously can be used to add mass to targets to allow Cryo-EM work, or to fix a protein in an active/inactive state. See below for the best practices for de novo miniprotein binder design, same goes for VHH design.

Antibody Informatics

- Numbering
  - IMGT, Kabat, Chothia, among others let us define CDRs and framework regions in standardized ways.
  - With these standard numbering approaches, we can then compare frequencies at given positions and germlines to identify liabilities, infrequent amino acid occurrences etc. Notably, this is how tools like BioPhi evaluate humanness, by comparing to human subsets of databases like the OAS.
- Post-translational modifications and Liabilities
  - Some PTMs are relatively straightforward to identify, e.g. just a substring of resides might be a risky part of an antibody.
  - Some are not as clear, and machine learning methods exist to predict, e.g. N-Linked Glycosylation to varying degrees of success
- Scoring & Developability
  - Tools like PROPERMAB from Regeneron are following in that realm, evaluating properties like:
    - Basic composition & charge
    - Structure-derived:
      - Surface patches (global & CDR)
      - Solvent-accessible areas
      - Charge distribution
      - Electric & hydrophobic moments
      - Hydrophobic-potential score
      - Spatial statistics (clustering)
      - Aromatic counts

Bioinformatics Utilities

- Multiple Sequence Alignment
  - MMseqs2: fast sensitive clustering & alignment for large datasets (uniprot and other databases)
  - HMMER: search/alignment; excels at detecting remote homologs
  - BLAST: gold-standard heuristic local alignment
  - IgBlast: BLAST variant specialized for antibodies and Igs
  - MAFFT: high-accuracy iterative refinement for large MSAs
  - Clustal Omega: scalable progressive alignment using guide-trees
  - MUSCLE: fast, accurate aligner; solid default for most datasets
- Databases
  - PlabDab: curated antibody sequence/structure database
  - SabDab: structural antibody database
    - TheraSabDab: therapeutic antibody subset of SabDab with developability metadata
  - PDB: 3-D structures of macromolecules
  - AlphaFold DB: predicted structures for >200 M proteins
  - UniProtKB: comprehensive protein sequence & functional annotation
  - Pfam: HMM profiles of protein families/domains
  - InterPro: integrated signatures from Pfam, SMART, TIGRFAMs, etc.
  - Observed Antibody Space (OAS): >2 B raw NGS antibody sequences
  - CATH / SCOPe: hierarchical structural classification of proteins
  - GPCRdb: receptor structures, ligands & mutational data
  - RCSB Ligand Expo: chemical components present in PDB entries

Simulation, Molecular Dynamics & Mechanics

- Molecular dynamics (MD) integrates Newton's equations of motion to predict atomistic trajectories, capturing conformational changes, stability, binding/unbinding, and free-energy landscapes from femtoseconds to milliseconds.
  - OpenMM: GPU-accelerated, Python-native MD engine; easy custom workflows
  - GROMACS: highly optimized MD for biomolecules; outstanding parallel performance
  - AMBER (pmemd/sander): force-field development and advanced free-energy workflows
  - NAMD: scalable MD for very large systems; CHARMM force-field support
  - CHARMM: versatile MD & energy minimization toolkit
  - Desmond: high-throughput MD with replica-exchange, REST, and FEP capabilities
  - LAMMPS flexible engine for coarse-grained and materials simulations
  - Anton / Anton 2: purpose-built supercomputer enabling micro- to millisecond simulations
  - Rosetta Relax / FastRelax: all-atom energy minimization & refinement
  - Enhanced Sampling Plugins
    - PLUMED: metadynamics, umbrella sampling, adaptive biasing force, etc.
    - Colvars (NAMD/CHARMM): collective variable framework, replica-exchange, string method
  - Coarse-Grained & Implicit Solvent Frameworks
    - MARTINI: CG force-field for proteins, lipids & membranes
    - AWSEM / SMOG: structure-based potentials for folding and pathway studies