AlphaFold2 (Protein Structure prediction)

When to use it: Choose AlphaFold2 when the highest possible accuracy in protein structure prediction is important, and there are sufficient computational resources and time. AlphaFold2 excels with proteins that have numerous homologous sequences, allowing it to build comprehensive multiple sequence alignments (MSAs) and utilize template structures when available. However, it requires significant computational power and longer processing times due to the complexity of generating MSAs and running its deep learning models. It's less suitable for orphan genes or rapidly evolving proteins with few homologs, where MSAs are sparse or non-existent.

Availability on SCINet: The underlying databases and singularity files are available on both Ceres and Atlas.

Key Features:

- **Deep Learning Architecture:** AlphaFold2 utilizes advanced deep learning techniques, specifically transformer neural networks, to predict protein structures from amino acid sequences with high accuracy.
- **High Prediction Accuracy**: Achieves near-experimental accuracy in many cases, significantly outperforming previous computational methods in the Critical Assessment of protein Structure Prediction (CASP) competitions.
- **Use of Multiple Sequence Alignments (MSAs)**: Relies heavily on evolutionary information derived from MSAs to inform its predictions, leveraging patterns conserved across homologous sequences to improve accuracy.
- **Incorporation of Template Information**: Utilizes template-based modeling by incorporating structural information from experimentally solved protein structures when available, enhancing prediction quality.
- **End-to-End Learning Model**: Integrates multiple stages of protein folding prediction into a single, end-to-end model, streamlining the process and reducing the need for separate modules or expert intervention.
- **Attention Mechanisms**: Employs attention mechanisms within its neural network architecture to capture long-range interactions between amino acids, which are crucial for accurate protein folding predictions.
- **Computational Resource Intensive**: Due to its complex architecture and reliance on MSAs and templates, AlphaFold2 requires substantial computational resources and longer processing times compared to models like OmegaFold and ESMFold.
- **Broad Applicability:** Effective in predicting structures for a wide range of proteins, including those with complex folds and multiple domains, making it a valuable tool for various biological research areas.

- **Code Availability:** The code for AlphaFold2 has been made openly available by DeepMind, allowing researchers worldwide to utilize and build upon its groundbreaking methods.

Code: https://github.com/google-deepmind/alphafold
Paper: https://doi.org/10.1038/s41586-021-03819-2
Website: https://alphafold.ebi.ac.uk/
Online servers: https://build.nvidia.com/deepmind/alphafold2 or
https://build.nvidia.com/deepmind/alphafold2-multimer or
https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
Download AlphaFold protein structures:
EBI: https://alphafold.ebi.ac.uk/download
Google Cloud: https://console.cloud.google.com/storage/browser/public-datasets-deepmind-alphafold-v4/proteomes

AlphaFold3 (Protein Structure prediction)

When to use it: AlphaFold3 may be the preferred option when you require highly accurate predictions of biomolecular structures and interactions, especially for complex systems involving proteins, DNA, RNA, and ligands. It is particularly beneficial for studying multi-domain proteins and complexes where enhanced performance is critical. If your research needs a comprehensive view of molecular interactions within a cellular context, AlphaFold3's ability to generate joint 3D structures makes it valuable. However, due to its proprietary nature, daily job submission limits (restricted to 20 jobs per day), and license restrictions imposed by DeepMind, AlphaFold3 is best suited for smaller-scale projects that can operate within these constraints. It is less suitable for large-scale analyses or scenarios where integration into local high-performance computing (HPC) systems is required.

Availability on SCINet: The code for AlphaFold3 is currently not available on SCINet. Groups are working to reproduce the workflows and make them available.

Key Features:

- **Improved Accuracy**: Delivers enhanced prediction accuracy over previous versions, excelling in modeling complex biomolecular structures and interactions.
- **Multi-Biomolecule Prediction:** Capable of predicting the structures and interactions of a variety of biomolecules, including proteins, DNA, RNA, and ligands.
- **Advanced Complex Modeling:** Provides improved performance on multi-domain proteins and complexes, facilitating a more detailed understanding of intricate biological systems.

- **Comprehensive Interaction Modeling:** Employs a methodology that allows for the generation of joint 3D structures of interacting biomolecules, enhancing the understanding of their functions within the cellular environment.
- **Proprietary Software:** Not open-source, which limits customization, transparency, and the ability to integrate the software into existing computational pipelines.
- **Online Access Only:** Available exclusively through an online platform; cannot be installed on local machines or HPC systems, which may affect workflow integration and data privacy considerations.
- **Daily Job Limit:** Users are currently restricted to submitting 20 job submissions per day, potentially limiting throughput for projects requiring extensive computational predictions.
- **License Restrictions:** Use is subject to DeepMind's terms and conditions, with restrictions on commercial use and redistribution, which may impact collaborative and commercial projects.

Paper: https://doi.org/10.1038/s41586-024-07487-w
Website: https://alphafoldserver.com/

OmegaFold (Protein Structure prediction)

When to use it: Use OmegaFold for protein structure prediction when you require a balance between prediction speed and accuracy without heavy reliance on computational resources. Like ESMFold, OmegaFold does not depend on MSAs or template structures, making it suitable for proteins with limited or no homologous sequences. OmegaFold often provides higher accuracy than ESMFold, especially for proteins that are difficult to align. It's a good choice for modeling orphan genes, novel proteins, or fast-evolving genes where MSAs are unavailable or unreliable. The computational efficiency of OmegaFold makes it practical for large-scale analyses or situations where resources are limited. It is also possible to run OmegaFold on CPU nodes although it is quite a bit slower.

Availability on SCINet: Module on Ceres. Need to install code and packages on Atlas.

Key Features:

- **Single Sequence Input:** OmegaFold predicts protein structures using only individual amino acid sequences, without the need for multiple sequence alignments (MSAs) or evolutionary information.
- **Deep Learning Architecture:** Employs advanced deep learning models, specifically transformer neural networks, to capture intricate patterns within protein sequences and predict their three-dimensional structures.

- **Fast Prediction Times:** By bypassing the computationally intensive steps of generating MSAs and searching for templates, OmegaFold offers quicker prediction times compared to methods like AlphaFold2.
- **Suitable for Orphan and Fast-Evolving Genes:** Ideal for proteins with few or no homologous sequences, such as orphan genes or rapidly evolving proteins, where MSAs are sparse or unavailable.
- **Moderate to High Accuracy:** Provides higher accuracy than other single-sequence prediction methods, often approaching the performance of MSA-based models for certain proteins.
- **Computational Efficiency:** Requires fewer computational resources, making it accessible for researchers with limited access to high-performance computing facilities.
- **No Dependence on Templates or MSAs:** Operates independently of template structures and evolutionary profiles, relying solely on the information encoded in the single protein sequence.
- **Scalability for Large-Scale Analyses:** Its efficiency and reduced resource requirements make OmegaFold suitable for large-scale protein structure prediction projects.
- **Applicable to Diverse Protein Families:** Capable of modeling a wide range of proteins, including those that are difficult to analyze with MSA-dependent methods due to low sequence similarity.
- **Facilitates Novel Protein Research:** A valuable tool for studying newly discovered proteins, synthetic sequences, or engineered proteins where evolutionary data is lacking.

Code: https://github.com/HeliXonProtein/OmegaFold
Paper: https://doi.org/10.1101/2022.07.21.500999

ESMFold (Protein Structure prediction)

When to use it: It is recommended to use ESMFold when you need rapid protein structure predictions with reduced computational demands. ESMFold is advantageous for proteins lacking extensive evolutionary information, such as orphan genes or fast-evolving sequences, because it does not rely on MSAs or templates. It operates using protein language models trained solely on sequence data, allowing for faster computation times. While ESMFold offers speed and efficiency, its accuracy is generally lower than that of AlphaFold2, making it more suitable for preliminary analyses or cases where approximate models are acceptable. It is also more suitable for predictions on a whole proteome scale or for all possible isoforms.

Availability on SCINet: Module on Ceres. Need to install Python code and packages on Atlas**.**

Key Features:

- **Protein Language Model Based**: ESMFold leverages the Evolutionary Scale Modeling (ESM) protein language model developed by Meta AI, which is trained on sequence data only and does not rely on evolutionary information from multiple sequence alignments (MSAs).
- **Fast Prediction**: Due to its reliance on protein language models, ESMFold can generate structural predictions more quickly than AlphaFold2, as it bypasses the need for building MSAs and other computationally expensive steps.
- **Single Sequence Input**: Unlike AlphaFold2, which requires a protein sequence and evolutionary context (i.e., MSAs), ESMFold can predict protein structures from just a single sequence, making it applicable even for proteins with few homologs or low-quality MSAs.
- **Reduced Model Size**: ESMFold is built on a smaller and more efficient model compared to AlphaFold2, trading off a slight reduction in prediction accuracy for much faster computation times.
- **Lower Computational Resource Requirement**: ESMFold typically uses fewer computational resources because of the simpler model design and reduced reliance on large input data sets (such as MSA profiles or templates).
- **Moderate Accuracy**: While ESMFold is less accurate than AlphaFold2 in predicting protein structures, especially for complex and longer proteins, it still provides reasonable accuracy for many protein families, particularly when speed is a priority.
- **No Dependence on Templates**: ESMFold does not use template-based modeling, unlike AlphaFold2, which incorporates template information from experimentally solved structures when available.

Code: https://github.com/facebookresearch/esm
Paper: https://doi.org/10.1126/science.ade2574
Website: https://esmatlas.com/
Server: https://build.nvidia.com/meta/esmfold

FoldSeek (Protein Structure search)

When to use it: Foldseek is a valuable tool in protein biology research for efficiently comparing protein structures on a large scale. It is particularly advantageous for high-throughput structural searches, enabling rapid screening of query structures against large structure databases to identify potential homologs. Foldseek can be used in functional annotation, where it reveals structurally similar proteins with known functions, which is especially useful when sequence similarity is low and traditional alignment methods do not work. In evolutionary studies, Foldseek aids in uncovering distant protein relationships undetectable through sequence alignment alone. Its optimized performance makes it ideal for resource-limited environments, offering a balanced approach to speed and sensitivity. Furthermore, Foldseek serves as a

valuable supplement to sequence alignment tools, enhancing insights into protein function and evolutionary relationships through structural comparison.

Availability on SCINet: Need to install the code on Ceres or Atlas.

Key Features:

- **Ultra-Fast Structural Alignment**: Designed for speed, FoldSeek can compare millions of protein structures rapidly, significantly outperforming traditional structural alignment tools in terms of computational efficiency.
- **Scalability**: Capable of handling very large datasets (including PDB, SwissProt, and CATH), making it suitable for proteome-wide analyses or screening large structural databases.
- **Sensitive Detection of Structural Similarities**: Employs advanced algorithms to detect subtle structural similarities between proteins, even when they share low sequence identity.
- **Rigid-Body Alignment**: Focuses on rigid structural alignments, providing accurate superpositions of protein backbones without accounting for conformational flexibility.
- **User-Friendly Interface**: Offers command-line tools and potentially web-based interfaces that are accessible and easy to integrate into existing workflows.
- **Integration with Structural Databases**: Compatible with major protein structure databases, facilitating seamless searches and analyses.
- **Resource Efficiency**: Requires less computational power and time compared to more complex alignment tools, making it accessible for researchers with limited resources.
- **Flexible Output Formats**: Supports various output formats such as TSV (Tab-Separated Values) and HTML, facilitating easy integration with data analysis pipelines and providing user-friendly visualization of results.

Comparison with Tools Like FATCAT:

- **Flexibility vs. Speed**: FATCAT (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists - https://fatcat.godziklab.org/) allows for flexible alignments that accommodate conformational changes such as hinge movements and twists. This makes FATCAT more sensitive in detecting structural similarities involving proteins with flexible regions or significant conformational variability.
- **Alignment Approach**: FoldSeek focuses on rigid-body alignments, optimizing for speed and computational efficiency. It is ideal when you need to perform quick searches and the proteins of interest do not require flexible alignment to detect similarities.
- Sensitivity to Conformational Changes: FATCAT may be more appropriate when studying proteins that undergo conformational changes or when the flexibility of the protein structure is crucial for the analysis. FoldSeek may not detect similarities that

depend on flexible alignments, potentially missing relationships that involve structural rearrangements.

Code: https://github.com/steineggerlab/foldseek
Paper: https://doi.org/10.1038/s41587-023-01773-0
Website: https://search.foldseek.com/search

ESM-Variant (Protein Variant Effect Prediction)

When to use it: ESM-Variant can be used in your research to predict the functional consequences of amino acid substitutions in protein sequences. This tool is particularly useful when assessing the potential impact of missense variants on protein function, conducting high-throughput analysis of numerous protein variants, or working with proteins lacking evolutionary data. ESM-Variant is also useful for analyzing novel proteins with limited experimental data or complementing existing variant analysis pipelines to enhance accuracy. Specifically, it is beneficial in scenarios such as identifying disease-associated mutations, evaluating large-scale genomic data, or investigating engineered proteins, making it a fast and easy-to-use resource for researchers and bioinformaticians seeking to understand the effects of protein variations.

Availability on SCINet: Need to install the code and Python packages on Ceres or Atlas.

Key Features:

- **Protein Language Model-Based Predictions**: Utilizes Evolutionary Scale Modeling (ESM) protein language models trained on vast amounts of sequence data to capture deep contextual information about protein sequences.
- **Single Sequence Input**: Operates without the need for multiple sequence alignments (MSAs) or evolutionary conservation data, making it suitable for proteins lacking homologs.
- **Quantitative Variant Scoring:** Provides consequence scores for amino acid substitutions, quantifying the predicted impact on protein function and stability.
- **High-Throughput Capability**: Efficiently processes large datasets of variants, facilitating genome-wide association studies and large-scale mutational analyses.
- **Applicability to Diverse Proteins**: Effective across a wide range of proteins, including those that are poorly characterized or without known structures.
- **No Structural Data Required**: Does not rely on 3D structural information, allowing predictions for proteins without resolved structures.
- **Ease of Integration**: It can be incorporated into existing bioinformatics pipelines and workflows, enhancing the versatility of variant analysis strategies.
- **Open-Source Availability:** Available as open-source software, allowing for customization, transparency, and community-driven improvements.

Code: https://github.com/ntranoslab/esm-variants
Paper: https://github.com/ntranoslab/esm-variants
Examples of ESM-Variant data for:
Maize: https://www.maizegdb.org/effect/maize/
Fusarium: https://www.maizegdb.org/effect/fusarium/index.php

RFDiffusion (Protein Binder Design)

When to use it: RFDiffusion is a tool for researchers and bioengineers focused on designing novel protein binders for targeted interactions on protein surfaces. It is particularly effective in creating de novo protein backbone structures that can bind to specified regions, such as active sites or regulatory interfaces, making it a valuable resource where precision binding is needed. RFDiffusion is also used in structural biology studies that explore protein function by enabling the design of proteins that probe specific interaction regions. When no natural binders exist for a protein of interest, RFDiffusion provides a solution by generating entirely new proteins with desired binding properties. Its flexibility allows for customizable binding interfaces, making it ideal for engineering proteins with precise affinity and specificity at targeted hot spots.

Availability on SCINet: Need to install the code on Ceres or Atlas.

Key Features:

- **Diffusion Probabilistic Model**: Utilizes denoising diffusion probabilistic models (DDPMs) to generate diverse protein structures that can bind to target regions, capturing the complexity of protein conformational space.
- **Hot Spot Targeting**: Allows specification of binding hot spots on the target protein, enabling focused design of binders to these critical regions.
- **De Novo Protein Design**: Generates novel protein sequences and structures not limited to existing protein scaffolds, expanding the possibilities for unique binders.
- **High-Throughput Generation**: Capable of producing a large number of candidate binders efficiently, facilitating screening and selection processes.
- **Integration with Structural Data**: Leverages structural information of the target protein, enhancing the accuracy and relevance of the designed binders.
- **Customizable Constraints**: Offers the ability to impose structural and sequence constraints, such as secondary structure preferences or amino acid composition, tailoring the designs to specific requirements.
- **User-Friendly Interface**: Provides accessible tools and comprehensive documentation, enabling users with varying levels of expertise to utilize the software effectively.
- **Open-Source Availability**: Available as open-source software, allowing for transparency, customization, and community contributions.

- **Facilitates Experimental Validation**: Generates protein designs that can be synthesized and tested experimentally, bridging computational predictions with laboratory applications.

Code: https://github.com/RosettaCommons/RFdiffusion
Paper: https://doi.org/10.1038/s41586-023-06415-8
Website: https://www.bakerlab.org/2023/07/11/diffusion-model-for-protein-design/
Servers:
https://colab.research.google.com/github/sokrypton/ColabDesign/blob/v1.1.1/rf/examples/diffusion.ipynb
https://build.nvidia.com/ipd/rfdiffusion