Machine Learning for Biologists

Page 1 – Core machine-learning and protein-data vocabulary

Machine learning is the practice of learning useful patterns from data. For biologists, the key is simple: know *what the model sees*, *what it predicts*, and *what its confidence means.* In a hackathon, you usually use **pretrained** models rather than training giant systems from scratch.

Typical framework of a machine learning project. In machine learning, a biological problem is translated into data, a model, and an evaluation strategy. A dataset of examples is assembled, each example containing input features and, in supervised settings, target labels. The data are separated into training, validation, and test sets. The model is trained on the training set by updating its parameters to minimize a loss function. The validation set is used during development to tune settings and compare model variants, while the test set is kept untouched until the end to measure true performance on unseen data. After training and assessment, the final model is applied to new data during inference, where it can classify, predict, score, or generate biologically relevant outputs.

A compact ML dictionary

Term	Meaning
Artificial intelligence (AI)	The broad idea of machines performing tasks that appear intelligent.
Machine learning (ML)	A subfield of AI in which a system learns useful patterns from data.
Deep learning (DL)	ML based on multi-layer neural networks that learn rich representations.
Sample / example	One observation in a dataset, such as one sequence or one sequence-structure pair.
Feature / input	The information fed into the model: sequence, image, coordinates, metadata, etc.
Label / target	The quantity the model is asked to predict: class, score, activity, structure, etc.
Dataset	The collection of examples used for development and evaluation.
Training set	The subset used to fit model parameters.
Validation set	The held-out subset used to tune settings and compare model versions.
Test set	The final held-out subset used for an honest estimate of performance.
Model	A parameterized function that maps inputs to outputs, often written as $f_\theta(x)$.
Parameters / weights	The learned numerical values inside the model.
Architecture	The overall model design (transformer, graph neural network, diffusion model, etc.).
Embedding / representation	A learned numerical encoding that captures useful biological relationships.
Loss function	A number measuring how wrong the model is; training minimizes it.
Gradient	The direction telling us how to change parameters to reduce the loss.
Optimizer	The update rule that uses gradients (for example SGD or Adam).
Learning rate	The step size used by the optimizer.
Batch / mini-batch	A small subset of examples used for one update step.
Epoch	One full pass through the training set.
Inference	Using a trained model on new data.
Supervised learning	Learning from inputs paired with known targets.
Unsupervised learning	Finding structure in data without explicit labels.
Self-supervised learning	Creating learning tasks from the data itself (for example masked-token prediction).
Generative model	A model that creates new plausible examples, such as sequences or 3D backbones.
Classification	Predicting a category.
Regression	Predicting a continuous value.
Overfitting	Memorizing the training set and failing on new data.
Underfitting	Failing to capture the important pattern.
Generalization	Performance on unseen examples.
Regularization	Any strategy used to reduce overfitting.
Calibration	Whether confidence estimates match reality.

Protein and structural-biology terms you must recognize

Term	Meaning
FASTA	Plain-text sequence format. A header starts with > and is followed by amino-acid letters.
MSA	Multiple sequence alignment: homologous sequences aligned residue-by-residue, exposing conservation and co-evolution.
PDB file	Traditional text format for 3D structures: atomic coordinates, chain IDs, residue numbers, and metadata.
mmCIF	The modern structured replacement for PDB, especially for larger or more complex entries.
Template	A known structure used as a guide during prediction.
Chain	One polymer entity in a structure (for example chain A).
Residue	One amino acid in a protein chain.
Atom coordinates	The 3D positions of atoms, usually in angstroms (A).
Backbone	The repeating N-C_α-C framework that defines the fold.
Side chain	The residue-specific chemical group attached to the backbone; often central for binding and catalysis.
Monomer / complex	A single chain vs a multi-chain assembly or protein-ligand structure.
Motif	A local structural or functional pattern that should be preserved.
Residue indexing	The exact chain IDs and residue numbers used by the tools; mistakes here cause many workflow errors.

Practical mindset: understand the inputs, understand the outputs, and never confuse model confidence with experimental truth.

The story of ML-guided protein analysis and design

Page 2 – From folding prediction to modern protein design workflows

The two central problems

Protein folding prediction asks the forward question: given a sequence, what structure is most likely?
Inverse folding asks the reverse question: given a backbone or structural constraint, which sequences are compatible with it? Modern protein design uses both directions together.

How the field evolved

Early computational protein design relied on physics-inspired scoring: steric packing, hydrogen bonding, solvation, and Monte Carlo search. This era produced major proofs of principle such as **Top7** and the first **computationally designed enzymes**. The deep-learning era accelerated when three things improved together: larger structural and sequence databases, better neural-network architectures, and enough GPU compute to train and run these models at scale.

A short timeline

Year	Milestone
2003	Top7 showed that a new protein fold could be designed with atomic-level accuracy.
2008–2011	Early computational enzyme design plus directed evolution showed the value of design-then-iterate.
2021	RoseTTAFold and AlphaFold2 transformed sequence-to-structure prediction.
2022	ProteinMPNN made backbone-conditioned sequence design much faster.
2023	RFdiffusion introduced diffusion-based generation of new protein backbones.
2024	AlphaFold3 extended prediction to biomolecular complexes.

What the main tools do

- AlphaFold / RoseTTAFold: sequence → structure. Mainly used for **analysis** and **fold-back validation**.
- ProteinMPNN: backbone → compatible sequences. A modern **inverse-folding** tool.
- RFdiffusion: noise + constraints → new backbone. A **generative design** tool.

The practical pipeline

1. **Define the target and constraints:** motif, interface, symmetry, chain count, residue locks.
2. **Generate backbones:** use RFdiffusion to create candidate structures.
3. **Design sequences:** use ProteinMPNN to propose amino-acid sequences for each backbone.
4. **Fold-back validation:** run AlphaFold or RoseTTAFold and check whether the designed sequence recovers the intended fold.
5. **Screen and test:** rank by confidence, geometry, and diversity, then move the best candidates to experiments.

Scoring terms you will see

Term	Meaning
pLDDT	AlphaFold local confidence; higher usually means the local structure is more reliable.
PAE	AlphaFold uncertainty between regions; especially useful for domains and interfaces.
RMSD	Coordinate difference between aligned structures; lower means more similar.
TM-score	Global fold similarity; higher means more similar folds.
Sequence recovery	How often an inverse-folding method recovers native-like residues on known backbones.
Contig	In RFdiffusion, a compact specification of fixed and generated segments and their chain arrangement.

The most important caveats

- **Structure is not function.** A design can have the right fold and still fail biologically.
- **Structure is not dynamics.** A single static prediction may miss the functional conformational state.
- **Confidence is not proof.** Model confidence supports decisions but does not replace experiments.

Selected references

Kuhlman *et al.* *Science* (2003); Röthlisberger *et al.* *Nature* (2008); Baek *et al.* *Science* (2021); Jumper *et al.* *Nature* (2021); Dauparas *et al.* *Science* (2022); Watson *et al.* *Nature* (2023); Abramson *et al.* *Nature* (2024).