OMNIEDGE SCIENTIFIC

AI & Machine Learning for Biologists

Classical ML • Deep Learning • Protein AI • Drug Discovery • XAI

Program Overview

This 21-day curriculum equips biologists with the practical AI and machine learning skills reshaping modern life sciences. Starting from zero programming assumption and building progressively through classical machine learning, deep learning, large language models for proteins, generative AI, and explainability, every concept is anchored to real biological questions: Can we predict cancer from expression data? Can AI fold a protein we have never seen before? Can we design new drug molecules computationally?

Foundations of AI & ML for Biology

What is AI/ML? A Biologist's Map	• AI vs Machine Learning vs Deep Learning — definitions with biological analogies • Supervised, unsupervised, and reinforcement learning with biology examples • Where ML is transforming biology: drug discovery, protein folding, genomics • Python ecosystem: NumPy, Pandas, scikit-learn, TensorFlow, PyTorch overview	✓ Distinguish AI, ML, and deep learning ✓ Map ML paradigms to biological problems ✓ Set up Python ML environment (Colab)	Google Colab, NumPy, Pandas, scikit-learn	Survey: map 5 real biological problems to the correct ML paradigm
Data Preparation & Biological Feature Engineering	• Loading biological tabular data with Pandas (CSV, Excel, TSV) • Handling missing values in clinical/genomic datasets • Feature scaling: StandardScaler, MinMaxScaler for expression data • Encoding categorical data: one-hot encoding of amino acids, nucleotides	✓ Clean and preprocess biological datasets ✓ Scale and normalise expression features ✓ Encode categorical biological variables numerically	Pandas, scikit-learn preprocessing	Preprocess a gene expression CSV: impute, scale, encode, export ready-to-use matrix

Exploratory Data Analysis & Visualisation	• Descriptive statistics on biological datasets • Correlation heatmaps of gene expression matrices (seaborn) • PCA scatter plots for sample clustering • Violin plots, box plots for comparing conditions with matplotlib/seaborn	✓ Perform EDA on biological datasets ✓ Interpret correlation patterns in expression data ✓ Visualise sample groupings with PCA	Pandas, seaborn, matplotlib, scikit-learn PCA	EDA report on a cancer gene expression dataset: correlation heatmap + PCA plot
Supervised Learning I — Classification	• Train/test split, cross-validation, overfitting explained with biology • Logistic Regression for binary disease classification • K-Nearest Neighbours for cell type prediction • Evaluation: accuracy, precision, recall, F1, confusion matrix	✓ Split data and apply logistic regression ✓ Evaluate classifiers with biological context ✓ Avoid overfitting with cross-validation	scikit-learn: LogisticRegression, KNeighborsClassifier, metrics	Classify tumour vs normal samples using gene expression features
Supervised Learning II — Regression & Trees	• Linear and Ridge Regression for predicting protein expression levels • Decision Trees for interpretable biological rules • Random Forest: ensemble method for biomarker discovery • Feature importance scores from Random Forest	✓ Apply regression to predict biological quantities ✓ Build and interpret decision trees ✓ Extract feature importances as biomarker candidates	scikit-learn: LinearRegression, DecisionTreeClassifier, RandomForestClassifier	Predict drug IC50 from molecular features using Random Forest; rank top features
Unsupervised Learning — Clustering & Dimensionality Reduction	• K-Means clustering for gene expression pattern discovery • Hierarchical clustering with dendrograms (scipy) • UMAP and t-SNE for high-dimensional biological data • Silhouette score for optimal cluster selection	✓ Apply K-Means and hierarchical clustering to omics data ✓ Visualise high-dimensional data with UMAP/t-SNE ✓ Select optimal number of clusters biologically	scikit-learn, scipy, umap-learn, seaborn	Cluster RNA-seq samples with K-Means; visualise with UMAP and label clusters
Model Evaluation, Tuning & Pipelines	• GridSearchCV and RandomizedSearchCV for hyperparameter tuning • ROC-AUC, precision-recall curves for imbalanced biological datasets	✓ Tune hyperparameters systematically ✓ Evaluate models on imbalanced biological data ✓ Build end-to-end reproducible ML pipelines	scikit-learn: Pipeline, GridSearchCV, roc_curve, joblib	Build a tuned pipeline to classify cancer subtypes; plot ROC-AUC curve

	• scikit-learn Pipelines: preprocessing + model in one object • Saving and loading models with joblib for reproducibility			

Deep Learning for Biological Sequences & Images

Neural Networks & Deep Learning Foundations	• Perceptron, activation functions (ReLU, sigmoid, softmax) — biological neuron analogy • Feedforward networks: layers, weights, backpropagation • Building an MLP with Keras Sequential API • Loss functions and optimisers: binary crossentropy, Adam	✓ Build and train a multi-layer perceptron in Keras ✓ Understand forward and backward propagation ✓ Apply appropriate loss functions for classification	TensorFlow, Keras, NumPy	Train a neural network to classify cancer type from gene expression profiles
Sequence Encoding & Embeddings for Genomics	• One-hot encoding of DNA and protein sequences • k-mer frequency vectors for sequence representation • Word2Vec-style embeddings for amino acids (ProtVec concept) • Embedding layers in Keras for biological sequences	✓ Encode DNA and protein sequences for ML input ✓ Generate k-mer feature vectors ✓ Use embedding layers for sequence data	NumPy, TensorFlow/Keras, Biopython	Encode 500 DNA sequences as one-hot and k-mer matrices; compare classification accuracy
Convolutional Neural Networks for Sequence & Image Data	• 1D CNNs for motif detection in DNA sequences • Conv1D, MaxPooling1D, Flatten, Dense layers in Keras • 2D CNNs for microscopy image classification • Transfer learning intro: using pre-trained image models for cell images	✓ Build 1D CNNs for sequence motif detection ✓ Apply 2D CNNs to biological image data ✓ Understand transfer learning for small datasets	TensorFlow/Keras: Conv1D, Conv2D, MaxPooling	Train a 1D CNN to detect splice sites in DNA sequences
Recurrent Neural Networks & LSTMs for Sequences	• RNNs for sequential biological data: time series gene expression • LSTM and GRU layers to capture long-range sequence dependencies • Protein secondary structure prediction as a sequence labelling task • Bidirectional LSTMs for reading sequences in both directions	✓ Build LSTM models for biological sequence tasks ✓ Apply bidirectional LSTMs to protein sequences ✓ Model temporal gene expression with RNNs	TensorFlow/Keras: LSTM, GRU, Bidirectional	Predict protein secondary structure (helix/sheet/coil) from amino acid sequence

Transformers & Protein Language Models	• Attention mechanism — biological sequence context analogy • Transformer architecture overview (BERT, GPT concept) • ESM-2 (protein language model) for protein representation • ProtTrans, ESMFold: embedding proteins and predicting properties	✓ Understand attention and transformer architecture ✓ Use pre-trained protein language models (ESM-2) ✓ Extract protein embeddings for downstream ML tasks	HuggingFace transformers, fair-esm, ESM-2	Extract ESM-2 embeddings for 50 proteins; cluster by function using embeddings
AlphaFold2 & Protein Structure Prediction	• AlphaFold2 architecture: multiple sequence alignment + structure module • Running AlphaFold2 via ColabFold (Google Colab interface) • Interpreting pLDDT confidence scores and PAE plots • Visualising predicted structures with py3Dmol and nglview	✓ Run AlphaFold2 predictions via ColabFold ✓ Interpret confidence metrics (pLDDT, PAE) ✓ Visualise and compare predicted structures	ColabFold, py3Dmol, nglview, Bio.PDB	Predict and visualise the structure of a novel protein; annotate confidence regions
Generative Models — VAEs & GANs in Biology	• Variational Autoencoders (VAE) for drug molecule generation • Generative Adversarial Networks (GAN) concept for synthetic biology data • Latent space exploration: interpolating between molecule properties • SMILES representation of molecules; RDKit for chemistry	✓ Understand VAE architecture and latent spaces ✓ Generate new molecule representations with VAEs ✓ Interpret latent space biology	TensorFlow/Keras, RDKit, DeepChem	Train a VAE on SMILES drug molecules; sample new molecule candidates from latent space

Applied AI in Bioinformatics & Drug

Graph Neural Networks for Molecular Biology	• Graphs as molecules: atoms as nodes, bonds as edges • Message passing neural networks (MPNN) concept • PyTorch Geometric and DGL-LifeSci for molecular GNNs • Predicting molecular toxicity and bioactivity with GNNs	✓ Represent molecules as graphs for GNN input ✓ Build and train a simple molecular GNN ✓ Predict molecular properties from graph structure	PyTorch Geometric, DGL-LifeSci, RDKit, DeepChem	Predict ESOL aqueous solubility for 100 drug molecules using a GNN
AI for Drug Discovery & Virtual Screening	• QSAR modelling: molecular descriptors as ML features	✓ Build QSAR models for bioactivity prediction	DeepChem, RDKit, scikit-learn, mordred	QSAR model to predict hERG cardiotoxicity from molecular fingerprints

	• DeepChem for molecular property prediction (ADMET) • Docking score prediction vs physics-based docking • Active learning for hit identification in virtual screening	✓ Use DeepChem for ADMET property prediction ✓ Apply active learning to prioritise compounds		
Deep Learning for Genomics & Variant Effect Prediction	• Predicting gene expression from DNA sequence (Enformer concept) • Variant effect scoring with CNN-based models • DeepSEA / Basset architecture for regulatory genomics • Interpreting models: saliency maps and in-silico mutagenesis	✓ Apply CNNs to regulatory sequence prediction ✓ Score variant effects on gene expression ✓ Use saliency maps to interpret model decisions	TensorFlow/Keras, Biopython, Captum (PyTorch)	Train a CNN to predict enhancer vs non-enhancer sequences; compute saliency maps
AI for Medical Imaging & Digital Pathology	• Histopathology image classification with CNNs • Transfer learning with ResNet50 / EfficientNet for cell images • Data augmentation strategies for limited biomedical image datasets • Grad-CAM visualisation for model interpretability in pathology	✓ Fine-tune a pre-trained CNN on pathology images ✓ Apply augmentation to overcome limited data ✓ Visualise CNN decisions with Grad-CAM	TensorFlow/Keras, torchvision, OpenCV, tf-explain	Fine-tune EfficientNet on H&E histology images to classify tumour grade
Single-Cell AI — scRNA-seq & Cell Type Annotation	• scVI: variational autoencoder for scRNA-seq integration • Scanpy + scANVI for automated cell type annotation • SCGAN and CellGAN for synthetic cell generation • Trajectory inference with PAGA and Monocle3 concepts	✓ Use scVI for batch-corrected scRNA-seq integration ✓ Automate cell type annotation with scANVI ✓ Infer differentiation trajectories from scRNA-seq	scvi-tools, Scanpy, anndata, PyTorch	Integrate two scRNA-seq batches with scVI; annotate cell clusters automatically
Explainable AI (XAI) & Model Interpretability in Biology	• SHAP values for feature importance in biological ML models • LIME for local explanations of individual predictions • Attention weight visualisation in transformer-based protein models • Biological validation of ML-identified features and biomarkers	✓ Apply SHAP to explain gene importance in classifiers ✓ Use LIME for individual patient prediction explanations ✓ Validate ML biomarkers against biological literature	SHAP, LIME, Captum, matplotlib	Apply SHAP to Random Forest cancer classifier; identify and validate top biomarker genes

Capstone — End-to-End AI Biology Pipeline	• Problem framing: biological question to ML pipeline design • Full pipeline: data ingestion → preprocessing → model training → evaluation → interpretation • Model card and reproducibility: documenting assumptions, data, metrics • Presenting AI biology results to mixed scientific audiences	✓ Design and execute a complete ML biology pipeline ✓ Document model decisions and limitations transparently ✓ Present findings to both computational and wet-lab audiences	All prior tools integrated; Colab, MLflow (optional)	Full AI pipeline on a chosen dataset: sequence, omics, or imaging with written report

Ready to Build Real-World AI Solutions in Biology?

Join the next generation of researchers combining Biology + AI + Innovation.

APPLY NOW

Visit Website : www.omniedgesci.com