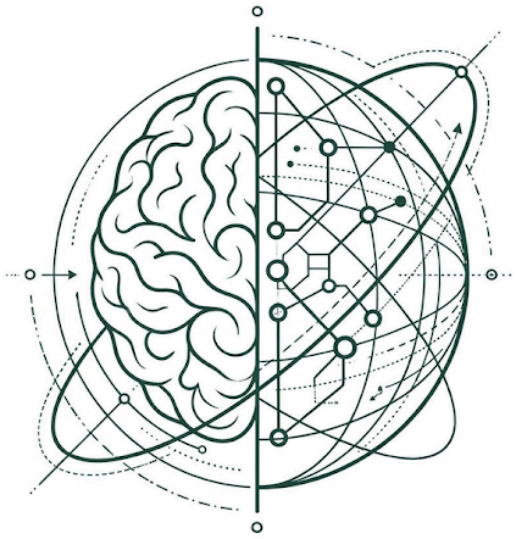
Fundusze Europejskie dla Rozwoju Społecznego
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską
NAWA NARODOWA AGENCJA WYMIANY AKADEMICKIEJ

AI in Molecular Biology, Biochemistry, and Biophysics

Institutional Handbook for IBB PAS



Introduction

0.1. Purpose and Scope of the Document

This handbook provides a comprehensive operational, regulatory, and educational framework for the scientific community of the Institute of Biochemistry and Biophysics, Polish Academy of Sciences (IBB PAN). The primary objective is to standardize and institutionalize the responsible, secure, and scientifically rigorous deployment of Artificial Intelligence (AI), Machine Learning (ML), Large Language Models (LLMs), and Generative Diffusion Models across research domains in molecular biology, biochemistry, biophysics, structural biology, and omics sciences. The document encompasses a full operational spectrum: a directory of specialized bio-AI architectures, empirical case studies and training evaluations from international mobility fellowships, key global researchers to follow at international conferences ("People to Follow"), comparative benchmarking of global external workshops, grant application methodology text blocks (NCN, ERC, Horizon Europe), FAIR data management protocols, patentability risk assessments under European Patent Office (EPO) and Polish Patent Office (UPRP) statutes, and a legally binding 12-Article Institutional AI Directive with 6 execution annex templates.

0.2. Target Audience (Who This Document Is For)

This handbook is tailored to address the distinct operational roles and responsibilities across IBB PAN:

- **Principal Investigators (PIs) and Research Group Leaders: Provides governance frameworks for supervising AI integration, managing intellectual property (IP) risks, conducting pre-synthesis biosafety screens, and drafting grant methodology proposals.**
- **PhD Candidates in the IBB PAN Doctoral School: Establishes explicit boundaries for AI usage in doctoral dissertations, disclosure requirements, and a structured 4-module bio-AI educational syllabus.**
- **Research Scientists & Bioinformaticians: Serves as a reference directory for validated bio-AI frameworks, experimental bench validation protocols, key conference speakers to prioritize, and FAIR-compliant data repository workflows. Restricts and defines AI applications in academic writing.**
- **Wet-Lab Personnel & Microscopy Specialists: Offers clear guidelines for digital image restoration (CARE, CSBDeep) without compromising image integrity or violating journal publication standards.**
- **Administrative Staff, RODO Data Protection Officers, & Technology Transfer Office (TTO): Provides legal mechanisms for GDPR/RODO compliance, spin-off provenance auditing, and protecting patentability under EPO Article 54 prior-art rules.**

0.3. How to Use This Document

The handbook is organized modularly to enable immediate navigation based on specific research, pedagogical, or administrative tasks:

1. **Evaluating International Workshops & Mobility Fellowships (Chapter 2): Consult Chapter 2 for 20 detailed case studies with workshop evaluations, training value scores (1–10), and transferred skill sets.**
2. **Identifying Key Conference Speakers & Workshops (Chapter 3): Refer to Chapter 3 for a directory of global Bio-AI thought leaders to follow at scientific conferences ("People to Follow"), a comparative benchmark table of leading international AI workshops (EMBL-EBI, EMBO, SIB, CSHL), and the 4-module doctoral curriculum.**
3. **Manuscript Preparation & Publication Ethics (Chapter 4): Consult Chapter 4 for publisher AI directives (COPE, Nature Portfolio, Science/AAAS) and digital image integrity**

guidelines.

4. **Grant Proposal Writing & Data Management (Chapter 5): Utilize Chapter 5 for ready-to-use English methodology text blocks for NCN/ERC grants and experimental bench validation matrices.**
5. **Data Security & IP Risk Assessment (Chapter 6): Refer to Chapter 6 for LLM security matrices, local open-weights deployment guidelines, and EPO patentability firewalls.**
6. **Institutional Compliance & Declarations (Chapter 7 & Annexes): Review Chapter 7 for the 12-Article Institutional Directive and download execution forms (Annexes 1–6).**

0.4. Financing Information

This handbook, along with the underlying empirical evaluation audit, was developed and funded under the following framework:

- **Program: PROM Programme – International Scholarship Exchange of PhD Candidates and Academic Staff**
- **Funding Agency: Polish National Agency for Academic Exchange (NAWA / Narodowa Agencja Wymiany Akademickiej)**
- **Co-funding Source: European Social Fund under the Knowledge Education Development Operational Programme (POWER) / European Funds for Social Development (FERS)**
- **Beneficiary Institution: Institute of Biochemistry and Biophysics, Polish Academy of Sciences (IBB PAN)**

The NAWA PROM grant supported short-term scientific mobility fellowships (ranging from 5 to 30 days) to premier international research centers across Europe and the USA, facilitating advanced training in structural biology, mass spectrometry, confocal microscopy, and bio-AI integration.

0.5. Data Sources

The synthesis and empirical recommendations within this document are derived from four primary data streams:

- **Empirical Audit of 20 NAWA PROM Mobility Reports: Individual fellowship reports submitted by IBB PAN researchers following visits to host laboratories (e.g., IGBMC Strasbourg, Univ. of Oslo, CEITEC Brno, Univ. of Edinburgh, ZMBH Heidelberg, CERM Florence, Biozentrum Basel).**
- **Institutional AI Governance Frameworks: Governance directives published by leading European and global research organizations, including EMBL, Max Planck Society, ETH Zurich, Karolinska Institutet, Harvard University, Francis Crick Institute, Wellcome Sanger Institute, CNRS, and NIH.**
- **Publisher Ethical Directives & Editorial Policies: AI policies established by the Committee on Publication Ethics (COPE), Nature Portfolio, Science/AAAS, Cell Press, EMBO Press, eLife, PNAS, and Nucleic Acids Research.**
- **Primary Bio-AI Literature & System Repositories: Technical specifications, peer-reviewed literature, and open repositories for structure prediction (AlphaFold 3, RFdiffusion, ESMFold), genomic foundation models (Evo 2, Caduceus), and smart microscopy (CARE, Cellpose).**

CHAPTER 1: Directory of AI Frameworks in Molecular Life Sciences

Artificial Intelligence in molecular life sciences includes generative models, equivariant graph neural networks (GNNs), large language models (LLMs), and Mamba state-space architectures operating on biological sequences, 3D atomic coordinates, mass spectra, and single-cell manifolds. Below is a directory detailing architectures, biological scopes, operational limits, hardware requirements, and source repositories.

1.1. Macromolecular Structure Prediction, Multimer Complex Assembly, & Dynamic Ensembles

1. AlphaFold 3 (Google DeepMind / Isomorphic Labs):
Architecture & Mechanism: Utilizes a Pairformer architecture coupled with a 3D generative Diffusion Module operating on raw atomic coordinates without requiring stereochemical template structures.
Biological Scope: Predicts joint 3D structures of macromolecular assemblies comprising proteins, nucleic acids (DNA, RNA), small-molecule ligands, post-translational modifications (PTMs), and inorganic ions.
Operational Limits & Failure Modes: Intrinsically disordered protein regions (IDPs) yield low pLDDT scores (<50); dynamic conformational flexibility is represented as static poses; coordination geometry for rare transition metals may exhibit stereo-chemical distortion.
Validation Requirement: High PAE (>15 Å) between domains requires experimental SEC-MALS, Cryo-EM, or site-directed mutagenesis verification.
Source: alphafoldserver.com

2. RoseTTAFold All-Atom (Baker Lab, Institute for Protein Design):
Architecture & Mechanism: A three-track neural network integrating 1D sequence representations, 2D distance maps, and 3D atomic coordinates with a diffusion-based module.
Biological Scope: Modeling of protein complexes bound to non-protein moieties, including synthetic molecules, nucleic acid scaffolds, and cofactors.
Operational Limits: Computationally intensive MSA generation step; lower accuracy on large multi-RNA tertiary assemblies compared to protein complexes.
Source: github.com/baker-laboratory/RoseTTAFold-All-Atom

3. ESMFold (Meta AI / EvolutionaryScale):
Architecture & Mechanism: Atomic structure predictor built on the ESM-2 15-billion-parameter protein language model, bypassing Multiple Sequence Alignment (MSA) construction.
Biological Scope: High-throughput monomeric structure prediction for screening metagenomic databases.
Operational Limits: Lower accuracy for orphan proteins without evolutionary homologs or complex multimeric interfaces.
Source: esmatlas.com

4. BioEmu (Microsoft Research):
Architecture & Mechanism: Neural emulator trained on molecular dynamics (MD) trajectories, combining generative AI with statistical thermodynamics to sample equilibrium conformational ensembles.
Biological Scope: Models conformational flexibility, functional state transitions, and transient pocket openings.
Operational Limits: Limited to timescales represented in training MD trajectories; sensitive to initial starting states.

Source: github.com/microsoft/bioemu

5. RFdiffusion (Baker Lab):
Architecture & Mechanism: SE(3)-equivariant denoising diffusion model trained on protein structures, generating 3D backbone coordinates from noise guided by spatial constraints.
Biological Scope: Design of protein scaffolds, target binders, symmetric oligomers, catalytic centers, and nanomaterial assemblies.
Operational Limits: Generates backbone geometries only; requires downstream sequence fitting (ProteinMPNN) and structural filtering (AlphaFold validation).
Source: github.com/rossettacommons/RFdiffusion

6. ProteinMPNN (Baker Lab):
Architecture & Mechanism: Message-passing neural network for inverse protein folding, predicting optimal amino acid sequences given a fixed 3D backbone.
Biological Scope: Sequence optimization for designed scaffolds, protein solubility enhancement, and thermal stability engineering.
Source: github.com/dauparas/ProteinMPNN

1.2. Genomic Foundation Models, Variant Prediction, & Long-Context Nucleotide Design

1. Evo 2 (Arc Institute / Stanford University):
Architecture & Mechanism: Biological foundation model utilizing StripedHyena architecture, combining convolutions with state-space dynamics to process context windows up to 1,000,000 base pairs at single-nucleotide resolution.
Biological Scope: Predicts functional impacts of coding and non-coding genetic variants; generates synthetic promoters, enhancers, CRISPR systems, and genomic constructs.
Source: github.com/arcinstitute/evo2

2. Caduceus (Stanford University):
Architecture & Mechanism: Bi-directional Mamba state-space model engineered for genomic sequences, incorporating reverse-complement equivariance.
Biological Scope: Genomic variant effect prediction, gene expression modeling, and chromatin accessibility scoring.
Source: github.com/kexinhuang/Caduceus

3. Nucleotide Transformer (InstaDeep / InstaDeepAI):
Architecture & Mechanism: Multi-species transformer foundation models trained on genetically diverse genomes.
Biological Scope: Promoter region identification, splice site prediction, histone modification annotation, and non-coding RNA classification.
Source: github.com/instadeepai/nucleotide-transformer

1.3. Optical Microscopy, Cryo-EM/ET Map Reconstruction, & Virtual Staining

1. ModelAngelo (MRC Laboratory of Molecular Biology, Cambridge):
Architecture & Mechanism: Graph neural network coupled with an amino acid language model for atomic model building into Cryo-EM density maps.
Biological Scope: Tracing of protein chains, secondary structures, and side-chain rotamers directly from 3D density maps.
Source: github.com/3dem/ModelAngelo

2. CSBDeep / CARE (Max Planck Institute of Molecular Cell Biology and Genetics):

Architecture & Mechanism: Content-Aware Image Restoration (CARE) using convolutional neural networks trained on low- and high-illumination image pairs.
Biological Scope: Restores microscopy images from low-exposure live-cell recordings, reducing phototoxicity during time-lapse imaging.
Source: csbdeep.bio

3. Cellpose & StarDist (HHMI Janelia / MPI-CBG):
Architecture & Mechanism: Segmentation algorithms utilizing vector flow fields (Cellpose) and star-convex object representations (StarDist) for cell and nuclear boundary detection.
Biological Scope: Quantification of cell morphology, nuclear counting, and spatial localization across tissues.
Source: cellpose.org | github.com/stardist/stardist

1.4. Single-Cell Omics, Spatial Transcriptomics, & Cellular Foundation Models

1. scVI / scvi-tools (Yosef Lab, UC Berkeley / Weizmann):
Architecture & Mechanism: Generative variational autoencoders (VAEs) modeling count distributions of scRNA-seq expression data.
Biological Scope: Batch-effect removal, biological integration across single-cell platforms, differential expression testing, and trajectory imputation.
Source: scvi-tools.org

2. Geneformer (Broad Institute / Harvard):
Architecture & Mechanism: Transformer foundation model pre-trained on 30 million single-cell transcriptomes encoding gene network dynamics.
Biological Scope: Models cell-state transitions, gene network rewiring under perturbations, and in silico gene knockouts.
Source: huggingface.co/ctheodoris/Geneformer

1.5. Metabolomics, Mass Spectrometry, & Natural Product Discovery Pipelines

1. SIRIUS 5 & CANOPUS (Leibniz Institute for Natural Product Research / Jena):
Architecture & Mechanism: Integrates tandem mass spectrometry (MS/MS) fragmentation trees with deep neural networks predicting molecular fingerprints (CSI:FingerID) and compound classes (CANOPUS).
Biological Scope: Automated structural annotation of unknown metabolites, lipid profiling, and natural product identification in complex biological extracts.
Operational Limits: Requires clean MS/MS precursor isolation; novel chemical scaffolds absent from training databases yield broad class annotations rather than exact structures.
Source: sirius-bioinformatics.de

2. COSMIC & DeepMASS (University of Jena / Bright-Insight):
Architecture & Mechanism: Confidence-score calibration algorithms for metabolomic identification and deep generative spectral matching.
Biological Scope: Filtering false positive assignments in untargeted metabolomics datasets.
Source: github.com/boecker-lab/sirius

1.6. Virtual Screening, Ligand Docking, & De Novo Small-Molecule Design

1. DiffDock (MIT CSAIL):
Architecture & Mechanism: Diffusion generative model over the manifold of ligand poses, modeling translation, rotation, and torsional degrees of freedom relative to target protein structures.
Biological Scope: Blind docking of small molecules to receptor proteins, catalytic site pose generation, and virtual screening.
Operational Limits: Assumes rigid protein backbone during generation unless combined with side-chain optimization tools; requires physical binding affinity verification (SPR/ITC).
Source: github.com/gcorso/DiffDock

2. GNINA & Boltz-1 (Carnegie Mellon University / MIT):
Architecture & Mechanism: Molecular docking software incorporating 3D convolutional neural networks (CNNs) for empirical scoring and pose refinement.
Biological Scope: Ligand binding affinity scoring and virtual screening filtering.
Source: github.com/gnina/gnina

1.7. Plant Biochemistry, Polyploidy Subgenome Resolution, & Metagenomic Mining

1. Subgenome Unmixing & antiSMASH 7 AI Modules (Wageningen University / Tübingen):
Architecture & Mechanism: Machine learning pipelines utilizing k-mer frequency distributions, epigenetic methylation signatures, and GNNs to resolve polyploid subgenomes and biosynthetic gene clusters (BGCs).
Biological Scope: Identification of secondary metabolite pathways in complex plant (e.g., Nicotiana, Triticum spp.), fungal and bacterial metagenomes.
Source: antismash.secondarymetabolites.org

1.8. Summary Matrix of Primary Bio-AI Frameworks

Domain	Framework / Tool	Core Biological Scope & Capabilities	Direct Repository / Link
Structure Prediction	AlphaFold 3 / RoseTTAFold AA	Multi-entity biomolecular complexes (protein, DNA, RNA, ligands, PTMs)	alphafoldserver.com
Generative Design	RFdiffusion / ProteinMPNN	SE(3) generative diffusion backbone design & inverse sequence fitting	github/rosettacommons
Genomic Foundation	Evo 2 / Caduceus	1M bp context window DNA/RNA variant prediction & genomic design	github/arcinstitute/evo2
Cryo-EM Automated	ModelAngelo / Topaz	GNN-assisted atomic model building & particle picking from 3D maps	github/3dem/ModelAngelo

Domain	Framework / Tool	Core Biological Scope & Capabilities	Direct Repository / Link
Smart Microscopy	CSBDeep CARE / Cellpose	Image denoising, cell & nuclear segmentation	csbdeep.bio
Single-Cell Omics	scVI / Geneformer / scGPT	Generative latent integration, single-cell foundation embeddings	scvi-tools.org
Metabolomics MS	SIRIUS 5 / CANOPUS	Tandem MS/MS formula prediction & compound class annotation	sirius-bioinformatics.de
Ligand Docking	DiffDock / GNINA	Diffusion-based flexible docking pose generation and CNN scoring	github/gcorso/DiffDock

CHAPTER 2: Empirical Case Studies and Workshop Training Evaluations from Foreign Mobility Fellowships (PROM NAWA IBB PAN)

2.1. Funding Framework: NAWA PROM Programme & Mobility Parameters

The empirical case studies presented in this chapter originate from an audit of individual mobility reports submitted by researchers and PhD candidates of the Institute of Biochemistry and Biophysics, Polish Academy of Sciences (IBB PAN). These research visits were funded under the PROM Programme – International Scholarship Exchange of PhD Candidates and Academic Staff, implemented by the Polish National Agency for Academic Exchange (NAWA) and co-financed by the European Social Fund under the Knowledge Education Development Operational Programme (POWER) / European Funds for Social Development (FERS).

The NAWA PROM grants provided targeted funding for short-term research mobility (ranging from 5 to 30 days) to leading European and US research institutions. The primary objectives of these visits included conducting specialized physical measurements, acquiring mastery in advanced computational or wet-lab techniques, executing joint experimental protocols, participating in host-laboratory hands-on workshops, and evaluating foreign laboratory governance practices. The audited cohort represents 20 distinct fellowship visits spanning structural biology, mass spectrometry, single-cell transcriptomics, confocal imaging, bacterial pathogenesis, and plant biochemistry.

2.2. Tiers of AI Adoption Across Partner Host Labs

- **Tier A: Domain-Specific Integration (30% of host labs): Deploying structural AI (AlphaFold 3 at IGBMC Strasbourg), deep learning microscopy processing (CEITEC Brno, Univ. Bern), or machine learning splicing site predictors (Univ. Edinburgh).**
- **Tier B: Institutional Infrastructure (15% of host labs): Establishing internal AI instances (e.g., "GPT UiO" at Univ. Oslo) or federated analytics compliant with EU AI Act standards (IIS La Fe Valencia).**
- **Tier C: Auxiliary Usage (35% of host labs): Utilizing AI for literature synthesis, proofreading, manuscript formatting, or R/Python plotting scripts (UIB Palma, IPREM Pau, IPB Halle, IMBB-FORTH).**
- **Tier D: Physical Wet-Lab Primacy (20% of host labs): Visits where AI was not applied (DSMZ Braunschweig, Weill Cornell Medicine, Curie Paris), confirming that several assays, experimental procedures, analytical frameworks (eg. liposome spectroscopy), and biopreservation techniques (eg. phage banking,) retain independent validity.**

2.3. Detailed Audited Case Studies with Workshop Training Evaluations

Case Study 1 **IGBMC Strasbourg (France) – research internship of PhD candidate**

Host Institution & Laboratory Research Scope: Institute of Genetics and Molecular and Cellular Biology (IGBMC), Department of Functional Genomics & Cancer, Strasbourg. The host laboratory specializes in mRNA decay mechanisms using X-ray crystallography, Cryo-EM, and biochemical assays.
Host Supervisor & Context: Dr. Bertrand Séraphin & Dr. Fabienne Mauxion. PROM NAWA mobility focused on viral protein-host factor interactions.
Research Focus: Structural interface between viral nucleoprotein (NP) and human poly(A)-binding protein PABPC1.

AI Methodology & Tool Stack: AlphaFold 3 and AlphaFold-Multimer modeling to generate structural hypotheses of candidate interaction surfaces
Empirical Bench Validation: Predictions served strictly as hypotheses. The models were used only to narrow down candidate interaction surfaces and to design constructs and truncated variants for the pull-down and EMSA (Electrophoretic Mobility Shift Assay) experiment.
Takeaway: Illustrates Tier A integration where AI accelerates structural hypothesis generation, while validity rests on wet-lab validation assays.

Case Study 2: University of Oslo (Norway) – research internship of PhD candidate

Host Institution & Laboratory Research Scope: Department of Biosciences & Nordic EMBL Partnership for Molecular Medicine (NCMM), University of Oslo. The host lab focuses on computational biology, translation dynamics, and direct RNA sequencing methodologies (Oxford Nanopore Technologies) to resolve post-transcriptional modifications.
Host Supervisor & Context: Prof. Eivind Valen. PROM NAWA visit to analyze mitochondrial non-coding RNA processing.
Research Focus: Profiling human mitochondrial transcriptome surveillance by MTPAP utilizing Nano3P-seq Nanopore direct cDNA sequencing.
AI & Infrastructure: No AI tools were used. The University of Oslo offers its students and employees a dedicated ChatGPT platform (GPT UiO), where the data is securely stored on the UiO's servers, is not shared outside and is not used for further training of AI models.
Data Privacy Alignment: Genomic sequencing reads and transcript maps were processed within isolated local servers without external cloud transmission.
Takeaway: Example of Tier B infrastructure. UiO's self-hosted wrapper allows researchers to leverage LLM code generation while enforcing data confidentiality.

Case Study 3: Health Research Institute La Fe (Valencia, Spain) – research internship of IBB employee

Host Institution & Laboratory Research Scope: Medical Research Institute Hospital La Fe, Neonatal Research Group. The lab is equipped with high-resolution chromatography-mass spectrometry (GC-MS, LC-MS/MS), Tunable Resistive Pulse Sensing (TRPS), and ATR-FTIR spectroscopy for clinical biomarker discovery in pediatric and maternal biofluids.
Host Supervisor & Context: Dr. Julia Kuligowski. PROM NAWA mobility centered on extracellular vesicles (EV) isolation and characterization.
Research Focus: Human milk extracellular vesicle (EV) lipidomics and biophysical characterization.
AI & Governance: Deployment of Google AI workflow frameworks ("Umiejętności Jutra 3.0") for spectral feature extraction and metabolomic clustering. Aligned with Spanish *Esquema Nacional de Seguridad (ENS)* and EU AI Act standards.
Empirical Bench Validation: Synthetic training datasets generated via SYNTHIA were validated against reference lipid standards.
Takeaway: Demonstrates compliance with European data privacy regulations when processing clinical biological samples.

Case Study 4: IBMCP Valencia (Spain) – Postdoctoral fellow

Host Institution & Laboratory Research Scope: Institute for Plant Molecular and Cellular Biology (IBMCP), CSIC-Polytechnic University of Valencia. Research scope encompasses plant-pathogen coevolution, artificial small RNAs, plant immunity, and agroinfiltration gene-silencing technologies.
Host Supervisor & Context: Dr. Alberto-Tomás Carbonell-Olivares. PROM NAWA fellowship on artificial microRNA design against phytopathogens.
Research Focus: Artificial microRNA (amiRNA) biogenesis in study of plant-microbe interactions
AI Policy: AI treated as a co-pilot. Used for target prediction script optimization (TargetFinder wrappers) followed by manual code audits.
Empirical Bench Validation: Transient agroinfiltration assays and Northern blot quantification of

amiRNA processing efficiency.
Takeaway: Highlights the necessity of manual verification for AI-generated bioinformatic scripts before execution.

Case Study 5: CEITEC Masaryk University (Brno, Czechia) – PhD candidate

Host Institution & Laboratory Research Scope: Central European Institute of Technology (CEITEC MU), CELLIM Core Facility. Core facility specializing in high-end confocal, STED, and light-sheet microscopy, single-molecule fluorescence spectroscopy, and automated image quantitative processing.
Host Supervisor & Context: Dr. Milan Esner. PROM NAWA visit for advanced fluorescence imaging and proximity assay quantification.
Research Focus: Proximity Ligation Assay (PLA) for human DNA polymerase iota utilizing Zeiss LSM800/880 confocal microscopy.
AI Scope: Deep learning image segmentation (StarDist) and low-light image restoration (CSBDeep CARE) to restore low-illumination confocal z-stacks.
Biological Benefit: Reduced laser exposure power by 75%, preserving sample fluorescence from photobleaching while recovering PLA focus spots. AI can reconstruct high-fidelity structures from low-light, noisy images. This opens the door for scientists to explore delicate living specimens without phototoxic damage. In addition, generative AI models allow "in-silico" or virtual staining. What the software is doinghere is looking at images taken in a bright field, without labels, and predicting where fluorescent markers would bind. This saves immense time and chemical resources. Machine learning does more than just analyze images, but also reveals hidden spatial relationships and complex
phenotypic morphologies through explainable AI frameworks to clarify its predictive pathways, leading to deeper and more robust insights into human health and disease.
Takeaway: Example of uniform deep-learning image denoising compliant with image integrity rules.

Case Study 6: University of Bern (Switzerland) – PhD candidate

Host Institution & Laboratory Research Scope: Institute of Cell Biology, University of Bern. The host laboratory focuses on organelle dynamics, proteostasis, age-related neurodegeneration, and cellular extrusion mechanisms using *Caenorhabditis elegans* transgenic models and automated fluorescence tracking.
Host Supervisor & Context: Department of Cell Biology. PROM NAWA mobility to measure neuronal exopher formation under stress.
Research Focus:To study impact of maternal stress–induced alterations in exopher production on in post-embryonic growth in *C. elegans*.
AI Application: Convolutional neural network (CNN) segmentation pipelines for worm body axis tracing and exopher vesicle counting over 48-hour time-lapse movies. Established automated longitudinal growth analysis workflow. Applied deep learning–based image analysis pipeline. Quantified developmental timing and growth dynamics.
Empirical Validation: Manual annotation of 500 reference frames confirmed segmentation precision > 94%. Validation performed via worm sorter (COPAS) assays.
Takeaway: Demonstrates computer vision replacing manual image analysis in model organism phenotyping.

Case Study 7 University of Edinburgh (UK) – Postdoctoral Fellow

Host Institution & Laboratory Research Scope: Institute of Cell Biology, School of Biological Sciences, University of Edinburgh. Specializes in genome architecture, DNA supercoiling topological domains, minor spliceosome assembly (U12-type introns), and RNA processing kinetics using RNA-seq and ChIP-seq.
Host Supervisor & Context: RNA Biology Division. PROM NAWA visit to map splice site motifs in non-canonical intron processing.
Research Focus: Genome-wide mapping of negative DNA supercoiling domains and alternative splicing regulation by the human minor spliceosome.

AI Application: Sequence pattern recognition algorithms for predicting minor intron U12-type splice site selection probabilities.
Empirical Validation: RT-PCR gel electrophoresis quantification and Sanger sequencing confirmation of predicted spliced isoforms.
Takeaway: Integration of machine learning motif discovery with transcriptomics validation.

Case Study 8: UIB Palma (Spain) – Postdoctoral Fellow

Host Institution & Laboratory Research Scope: Multidisciplinary Translational Oncology Group, University of the Balearic Islands (UIB), Palma de Mallorca. Research focuses on energy metabolism, mitochondrial function, oxidative stress, aging, and the hormonal environment in tumor progression
Host Supervisor & Context: The group is co-led by Dr. Jordi Oliver Oliver and Dr. María del Pilar Roca Salom
Research Focus: proteomics-based secretome study of breast cancer cells.
AI Scope: At Universitat de les Illes Balears (UIB), AI tools are used for text editing and literature search assistance. In accordance with UIB ethical standards and regulations, no confidential experimental data are uploaded to AI models. All AI generated outputs are verified by the researchers.
Takeaway: Tier C auxiliary AI usage with defined intellectual property boundaries.

Case Study 9: PREM Pau (France) – PhD candidate

Host Institution & Laboratory Research Scope: Institute of Analytical Sciences and Physico-Chemistry for Environment and Materials, CNRS / Université de Pau et des Pays de l'Adour. The laboratory is an international benchmark center for trace element speciation, high-resolution ICP-MS, metallomics, and isotopic analysis.
Host Supervisor & Context: prof. dr hab. Joanna Szpunar, Metallomics and metalloproteomics
Research Focus: characterization of copper-glutathione complexes.
AI Scope: AI-based tools were used exclusively to support literature research
Takeaway: AI tools are mainly used to support scientific work, particularly for language editing and reference formatting. Good practices emphasize the importance of transparency, verification of AI-generated content, and protection of confidential research data. Ultimately, researchers are responsible for verifying the accuracy and reliability of any AI-generated content used in their work..

Case Study 10: PB Halle (Germany) – PhD candidate

Host Institution & Laboratory Research Scope: Leibniz Institute of Plant Biochemistry (IPB), Halle (Saale). The host institute specializes in natural product chemistry, metabolism and protein biochemistry, cell and plant biology, as well as in synthetic biology and biotechnology
Host Supervisor & Context: Department of Biochemistry of Plant Interactions
PROM NAWA mobility for learning proteomics workflow for enrichment and detection of ADP-ribosylated proteins and peptides.
Research Focus: Identification of host proteins modified by a type three *effector* of *Pseudomonas syringae.*
AI Scope: AI-based tools used for searching and summarising literature, to improve the clarity and linguistic quality of scientific texts, while ensuring that the scientific content remained the responsibility of the author.
Takeaway: A transparent approach to the use of AI-based tools in scientific communication. .

Case Study 11: IMBB-FORTH Heraklion (Greece) – PhD candidate

Host Institution & Laboratory Research Scope: Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH). The laboratory specializes in homeostasis of proteins and RNA molecules, role of biocondensates, and *Arabidopsis thaliana* and *Nicotiana benthamiana* as models.
Host Supervisor & Context: stress adaptation in plants. PROM NAWA fellowship to analyze biocondensates‘ composition

Research Focus: The Proxitome-RNA-Capture Approach and FRAP analysis.
AI Methodology & Scope: Generative AI was used only as an auxiliary tool: for language editing of protocol and report texts and for drafting and debugging data-analysis scripts. All outputs were verified before use. AI was not used to generate, alter, or interpret primary experimental data, and responsibility for the scientific content remains entirely my own. The host institution's approach to AI, treating it as a verification and drafting tool rather than as a source of scientific authority, exemplifies the best-practice approach in this field.Every AI-generated text or code fragment was treated as a draft to be verified rather than as a finished product: wording was checked against the primary literature and against my own experimental records, and scripts were tested on data with a known outcome before being applied. The division of labour was kept explicit: AI was used for language editing, structuring and debugging only, never to generate, interpret or extrapolate experimental results, No confidential material or third-party personal data were entered into AI tools.
Takeaway: Demonstrates Tier C auxiliary role.

Case Study 12 ZMBH Heidelberg (Germany) – PhD candidate

Host Institution & Laboratory Research Scope: Center for Molecular Biology of Heidelberg University (ZMBH),. Research focuses on Synthetic biology and synthetic genomics approaches in microbial model systems
Host Supervisor & Context: Daniel Schindler lab. Passive regulatory mechanisms in genetic systems
Research Focus: Effects of genomic architecture on gene expression and cellular fitness in yeasts
AI Methodology & Scope: AI- assisted tools were used to support the preparation of scripts for data visualization and plotting. AI was used solely as a programming aid to improve coding efficiency and did not perform the scientific analysis, interpret experimental results, or generate research conclusions. All codes produced with AI assistance were reviewed, tested, and validated. The underlying experimental data, analyses, and scientific interpretations remained under our full responsibility. No confidential or unpublished research data were shared beyond the scope permitted by institutional policies, and AI was used in accordance with the principles of research integrity, transparency, and responsible scientific practice.
Takeaway: Auxiliary use of AI-tools with strict privacy.

Case Study 13: CERM Florence (Italy) – PhD Candidate

Host Institution & Laboratory Research Scope: Magnetic Resonance Center (CERM), University of Florence. World-recognized research facility for high-field Solution and Solid-State NMR spectroscopy (600–1200 MHz), structural biology of metalloproteins, and intracellular NMR dynamics.
Host Supervisor & Context: Bioinorganic Structural Biology Division. PROM NAWA mobility to analyze the structural and redox events surrounding the metal cofactor of a metalloenzyme.
Research Focus: NMR analysis of superoxide dismutase
AI Methodology & Scope: AI used only as a tool to write script for data visualization
Workshop & Training Value Assessment: Attended CERM's 4-day advanced international workshop Takeaway: High-field structural NMR requires strict local data privacy controls to protect unreleased structural assignments.

Case Study 14: ECSEQ Bioinformatics Berlin (Germany) – PhD student

Host Institution & Laboratory Research Scope: ECSEQ Bioinformatics Training & R&D Center, Berlin. Specializes in Next-Generation Sequencing (NGS) data architecture, clinical epigenomics, Reduced Representation Bisulfite Sequencing (RRBS), and pipeline containerization.
Host Supervisor & Context: Epigenomics Computational Group. PROM NAWA visit to optimize quality control metrics in DNA methylation pipelines.
Research Focus: Automated quality control, adaptor trimming, and methylation bias correction for single-cell bisulfite sequencing datasets.
AI Methodology & Scope: usage in genomics, particularly for large-scale sequencing data

analysis, variant interpretation, and bioinformatics workflows. Deep learning scripts for detecting abnormal sequencing tile quality patterns and automated HTML summary report generation using custom R markdown templates.
Takeaway: Demonstrates automated quality monitoring in high-throughput sequencing pipelines.

Case Study 15: University of Iowa (USA) – Postdoctoral fellow

Host Institution & Laboratory Research Scope: Carver College of Medicine, Department of Microbiology and Immunology, University of Iowa. Research focuses on bacterial pathogenesis, *Staphylococcus aureus* biofilm architecture, quorum sensing (agr system), and host immune evasion.
Host Supervisor & Context: Bacterial Pathogenesis Laboratory. PROM NAWA mobility to construct transposon mutant libraries.
Research Focus: Construction and phenotypic screening of *S. aureus* knockouts involved in extracellular matrix dispersion.
AI Methodology & Scope: Zero AI reliance (Tier D). The research relied entirely on traditional molecular biology protocols, allelic exchange plasmid construction, transposon mutagenesis, and physical phenotypic assays.
Empirical Bench Validation: Phenotypes were confirmed via microtiter plate biofilm crystal violet staining, confocal laser scanning microscopy of glass-bottom chamber slides, and Western blotting of secreted proteases.
Takeaway: Highlights core areas of bacterial genetics where physical wet-lab assays remain the sole valid methodology without AI intervention.

Case Study 16: Curie Institute Paris (France) – PhD candidate

Host Institution & Laboratory Research Scope: Department of Biology, Ecole Normale Superior-PSL University. The mechanisms that underlie the stability, organization, and plasticity of genetic and epigenetic information.
Training: "Quantitative Approaches to DNA Damage and Repair"
AI Methodology & Scope: AI- and machine learning-related topics in the context of quantitative analysis of DNA damage and repair. Machine learning can be used to predict DNA break sites from genomic data, classify genomic features, and support data integration in bioinformatic workflows. Related quantitative tools also demonstrated how AI-based approaches can help with microscopy image analysis, mutation cluster inference, and modelling chromatin or polymer dynamics. The training also emphasized good practices such as using high-quality datasets, documenting methods clearly, and ensuring reproducibility.
Takeaway AI tools in DNA damage and repair used as supportive analytical methods rather than standalone conclusions, requiring careful validation, proper controls, and critical interpretation. A key point is that AI outputs should be checked against biological knowledge and, whenever possible, confirmed experimentally.

Case Study 17 DSMZ Braunschweig (Germany) – Postdoctoral fellow

Host Institution & Laboratory Research Scope: Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig. Global biobank authority for bacterial strain isolation, phage banking, taxonomic archiving, and phage-host specificity profiling.
Host Supervisor & Context: Bacteriophage Research & Biobanking Unit. PROM NAWA fellowship for lytic phage isolation against multidrug-resistant pathogens.
Research Focus: Isolation, host-range profiling, and plaque morphology categorization of novel lytic bacteriophages targeting Pseudomonas aeruginosa.
AI Methodology & Scope: Zero AI predictive tools deployed during strain isolation. Phage collection protocols followed classical double-layer agar plaque assays.
Empirical Bench Validation: Biological purity and plaque morphology, host-range killing efficiency quantified via spot testing and liquid culture lysis kinetics.
Takeaway: Biological biobanking relies on physical strain isolation and empirical plaque verification.

Case Study 18 **Weill Cornell Medicine NYC (USA) – Postdoctoral fellow**

Host Institution & Laboratory Research Scope: Department of Biochemistry, Weill Cornell Medical College, New York. Specializes in membrane biophysics, lipid-protein interactions, endocytic trafficking, and quantitative fluorescence spectroscopy.
Host Supervisor & Context: Membrane Biophysics Laboratory. PROM NAWA mobility to measure membrane order parameter changes.
Research Focus: Determination of membrane phase order and fluidity in liposomal models using C-Laurdan steady-state fluorescence spectroscopy.
AI Methodology & Scope: Zero AI intervention (Tier D). Generalized Polarization (GP) spectra were calculated directly from emission intensities at 440 nm and 490 nm using physical formulas.
Empirical Bench Validation: Temperature-dependent phase transition curves benchmarked against synthetic dipalmitoylphosphatidylcholine (DPPC) lipid standards across phase transitions (25°C to 55°C).
Takeaway: Direct physical spectroscopy yields reliable biophysical metrics without complex AI transformations.

CHAPTER 3: Key Researchers to Follow, International AI Workshops, and Doctoral Training Syllabus

3.1. Key Bio-AI Thought Leaders & Conference Speakers to Follow ("People to Follow")

For researchers and PhD candidates attending international conferences (e.g., ISMB/ECCB, EMBO Symposia, Gordon Research Conferences, CASP, Protein Society Meeting, IUCr, Protein Design Conference), keynotes and sessions led by the following principal investigators and computational bio-leaders represent priority destinations:

Researcher / Thought Leader	Affiliation & Lab	Key Bio-AI Contribution & Focus Area	Priority Conferences
David Baker	Institute for Protein Design, Univ. of Washington (USA)	Pioneer of *de novo* protein design, RFdiffusion, ProteinMPNN, RoseTTAFold All-Atom. 2024 Nobel Laureate.	EMBO Symposia, Protein Society, GRC Protein Folding
Demis Hassabis & John Jumper	Google DeepMind / Isomorphic Labs (UK)	Creators of AlphaFold 2 and AlphaFold 3 multi-entity biomolecular structure prediction diffusion engines.	ISMB/ECCB, CASP, NeurIPS (AI in Science)
Sergey Ovchinnikov	MIT Center for Data-Driven Life	ColabFold co-creator, evolutionary sequence	3DSIG, RECOMB,

Researcher / Thought Leader	Affiliation & Lab	Key Bio-AI Contribution & Focus Area	Priority Conferences
	Sciences (USA)	alignment optimization, hallucination-based protein design.	Biophysical Society
Mohammed AlQuraishi	Columbia University, Dept. of Systems Biology (USA)	OpenFold developer, Recurrent Geometric Networks (RGN), open-source biomolecular AI models.	ISMB, Critical Assessment of Structure Prediction (CASP)
Brian Hie	Stanford University / Arc Institute (USA)	Lead developer of Evo 2 (1M bp context genomic foundation model), viral escape language models.	CSHL Computational Biology, RECOMB
Christina Theodoris	Broad Institute of MIT and Harvard (USA)	Creator of Geneformer single-cell transcriptomic foundation model for gene network modeling.	ASHG, Single Cell Genomics, Nature Conferences
Florian Jug	Fondazione Human Technopole (Italy) / MPI-CBG	Pioneer of CARE/CSBDeep, Noise2Void, deep learning for live-cell optical microscopy restoration.	ELMI, Focus on Microscopy (FOM), IEEE ISBI
Carsen Stringer & Marius Pachitariu	HHMI Janelia Research Campus (USA)	Creators of Cellpose and Suite2p vector-flow segmentation and neural optical imaging pipelines.	Janelia Conferences, COSYNE, Society for Neuroscience
Sebastian Böcker	Leibniz Inst. / Friedrich Schiller Univ. Jena (Germany)	Lead developer of SIRIUS 5, CANOPUS, CSI:FingerID for tandem mass-spectrometry metabolite elucidation.	ASMS, Metabolomics Society, ECCB

Researcher / Thought Leader	Affiliation & Lab	Key Bio-AI Contribution & Focus Area	Priority Conferences
Sjors Scheres & Kresimir Bepler	MRC Laboratory of Molecular Biology, Cambridge (UK)	RELION developer (Scheres) and ModelAngelo GNN map tracer (Bepler) for Cryo-EM atomic modeling.	3DEM Gordon Conference, IUCr Congress

3.2. Evaluation Framework for Bio-AI Scientific Workshops & Hands-on Courses

To guide researchers and PhD candidates at IBB PAN in selecting external training opportunities, all international workshops, summer schools, and courses are evaluated against four standardized operational metrics:

1. Hands-on Computational / Practical Ratio: Proportion of course time dedicated to live code execution, data processing, or wet-lab validation versus theoretical lectures.
2. Target Proficiency Level: Classification from Introductory (basic Linux/Python) to Advanced (custom model architecture, GPU scaling, structural refinement).
3. Access & Prerequisite Threshold: Prerequisites required for entry (e.g., prior programming experience, raw dataset submission, competitive application).
4. Practical Training Value Score (1–10): Quantitative score assessing direct applicability of acquired skills to IBB PAN research projects and grant execution.

3.3. Directory & Comparative Benchmarking of International External Bio-AI Workshops

Workshop / Course Title	Organizing Institution	Core Training Focus & AI Technologies	Hands-on Ratio	Score
AlphaFold & Protein Structure Prediction Workshop	EMBL-EBI Train Online / Hinxton (UK)	=AlphaFold 3 parameter tuning, multimer complexes, pLDDT/PAE evaluation, PDB API programmatic parsing.	60% Hands-on	9.5 / 10
Machine Learning for Life Sciences Summer School	SIB Swiss Institute of Bioinformatics / Lausanne	Supervised/unsupervised ML, scVI single-cell VAE models, Random Forest classification for omics data.	70% Hands-on	9.0 / 10
Cryo-EM Image Processing &	Wellcome Connecting	ModelAngelo GNN atomic tracing, Topaz	75% Hands-o	9.5 / 10

Workshop / Course Title	Organizing Institution	Core Training Focus & AI Technologies	Hands-on Ratio	Score
Model Building Course	Science / Cambridge (UK)	particle picking, RELION / cryoSPARC deep learning workflows.	n	
EMBO Practical Course: Computational Protein Design	EMBO / IPD Seattle / Heidelberg	RFdiffusion backbone generative design, ProteinMPNN sequence optimization, PyMOL filtering.	80% Hands-on	10.0 / 10
CSHL Computational Genomics Course	Cold Spring Harbor Laboratory (USA)	Genomic foundation models (Evo 2, Caduceus), long-read variant effect prediction, transcriptomics.	65% Hands-on	9.0 / 10
PLGrid / EuroHPC Bio-Computing & GPU Scaling Workshop	CYFRONET AGH / PLGrid Infrastructure (Poland)	Singularity containerization, SLURM job scheduling for PyTorch, distributed multi-GPU execution.	85% Hands-on	9.5 / 10
ELIXIR Training: Bio-Image Data Analysis with Deep Learning	ELIXIR Europe Core Facilities	CARE restoration, Cellpose 3 segmentation, StarDist nuclear tracking, ImageJ Macro / Python integration.	80% Hands-on	9.0 / 10

3.4. Detailed 4-Module Syllabus for IBB PhD Candidates & Workshop Format Execution

MODULE 1: Computational Foundations, Environment Isolation, and Scripting Integrity
Target Audience: 1st Year PhD Candidates in Biological Sciences.
Topics Covered:

- **Linux Command-Line Interface (CLI) navigation, SSH remote server execution, and file management.**
- **Python environments for biological data using Conda / Mamba isolated virtual environments.**
- **BioPython and PyMOL visualization scripting for structure automated rendering.**
- **Git version control, commit discipline, and GitHub/Zenodo code archiving standards.**
- **Manual validation of LLM-generated code: detecting hallucinated function calls and computational syntax errors.**

Practical Deliverable: Students construct a version-controlled Git repository containing a validated BioPython script that parses PDB atomic structures and outputs domain contact maps.
MODULE 2: Macromolecular Structure Prediction and Conformational Analysis
Target Audience: PhD Candidates focusing on Structural Biology, Biochemistry, and Biophysics.
Topics Covered:

- **AlphaFold 3 and ESMFold parameter setup, input formatting, and Multiple Sequence Alignment (MSA) controls.**
- **Running ColabFold on local GPU clusters (PLGrid supercomputing nodes).**
- **Interpretation of structural confidence metrics: pLDDT scores, Predicted Aligned Error (PAE) matrices, and interface quality scores.**
- **Conformational ensemble sampling using BioEmu and molecular dynamics emulators.**
- **Identification of failure modes: intrinsically disordered regions, metal binding site distortions, and multimer interface false positives.**

Practical Deliverable: Generation and critical evaluation of a multi-protein interaction complex prediction, complete with PAE map analysis and designed mutant verification proposals.

MODULE 3: Generative Biomolecule Design, Inverse Folding, and Docking

Target Audience: Advanced PhD Candidates and Postdoctoral Researchers.

Topics Covered:

- **Generative scaffold design using SE(3)-equivariant diffusion models (RFdiffusion).**
- **Inverse protein folding and sequence optimization using ProteinMPNN.**
- **Small-molecule flexible docking and pose prediction using DiffDock and GNINA.**
- **PyMOL structural alignment, hydrogen bonding network analysis, and steric clash evaluation.**
- **Integrating computational design outputs with physical wet-lab synthesis plans (gene synthesis, expression vectors, SPR assays).**

Practical Deliverable: Complete in silico design pipeline producing 3 candidate protein binder sequences against a specified target, including physical assay validation protocols.

MODULE 4: HPC Execution, Data Security, Ethics, and FAIR Governance

Target Audience: All PhD Candidates prior to Mid-Term Evaluation.

Topics Covered:

- **Containerization technologies (Singularity, Docker) for executing computational pipelines on HPC clusters (PLGrid).**
- **European General Data Protection Regulation (GDPR / RODO) compliance when processing human sequencing data.**
- **Intellectual Property (IP) protection, European Patent Office (EPO Article 54) prior-art rules, and avoiding cloud data leaks.**
- **Journal image integrity standards: uniform image restoration vs. prohibited local manipulations.**
- **Drafting AI disclosure statements for scientific publications (Annex 1) and doctoral dissertations (Annex 2).**

Practical Deliverable: Submission of a completed Data Management Plan (DMP) and AI Safety Self-Audit for the candidate's PhD research projects

CHAPTER 4: Benchmarking of AI Practices in Scientific Research & Publishing

4.1. Directives of Research Institutions

- **European Molecular Biology Laboratory (EMBL): Mandates Open Science principles. AI models used in data analysis must disclose model architecture, weights, and training dataset provenance.**
- **Max Planck Society (Germany): Prioritizes Intellectual Property (IP) protection and PI accountability. Forbids transmitting unreleased sequences, crystallographic map files, or mass-spec data to third-party public models.**
- **ETH Zurich & Karolinska Institutet: Enforces physical wet-lab validation for biomedical AI predictions. *In silico* predictions of drug binding or target affinity require empirical bench testing.**
- **Harvard University & Cambridge University: Directives dictate that generative AI cannot substitute original scholarship in doctoral dissertations. AI usage for literature review or coding must be disclosed.**
- **Francis Crick Institute & Wellcome Sanger Institute (UK): Enforce bio-data governance. Human genomic sequences must remain within secure enclave environments. Cloud LLM tools are segmented from sequence storage.**
- **CNRS (France) & National Institutes of Health (NIH, USA): NIH rules prohibit peer reviewers from uploading grant proposals or manuscripts to external generative AI systems.**

4.2. Publisher Policies & Journal Directives

Publisher / Journal	AI Authorship & Text Policy	Peer-Review Confidentiality & Method Disclosure
COPE	AI tools cannot be listed as authors. LLMs lack legal agency and capacity to sign copyright agreements.	Authors must declare tool name, version, and purpose in Methods or Acknowledgments.
Nature Portfolio	Prohibition on AI authorship. Generative AI images/diagrams forbidden unless explicitly part of methodology.	Reviewers are prohibited from uploading submitted manuscripts to AI tools.
Science / AAAS	AI-generated text or figures cannot be used without prior editorial authorization. Uninformed AI text generation is treated as misconduct.	Methodological transparency required for machine learning models used in data processing.

Publisher / Journal	AI Authorship & Text Policy	Peer-Review Confidentiality & Method Disclosure
Cell Press & EMBO Press	AI can assist in proofreading human-written text. AI cannot interpret biological data or generate conclusions.	Declarations must specify tool versions and prompt scopes within the Methods section.
Nucleic Acids Res. / PNAS	Release of custom bio-AI code, trained weights, and datasets in open repositories (GitHub/Zenodo).	Reviewers must certify no manuscript text or data was entered into external AI platforms.

4.3. Digital Image Integrity Standards

- **Permissible Adjustments: Global linear brightness/contrast adjustments across the entire frame; validated deep learning denoising (e.g., CARE) applied uniformly, provided raw unedited images are deposited in open repositories (BioImage Archive).**
- **Prohibited Manipulations: Selective local enhancement/suppression of specific blot bands or fluorescent cells, artificial background noise suppression concealing non-specific binding, or generative structure synthesis.**

CHAPTER 5: AI Integration in Research Projects, Grant Proposals, and FAIR Data Management

5.1. Proposal Methodology Text Blocks (English) for NCN and ERC Applications

TEMPLATE A: Structural Bio-AI & De Novo Binder Design Methodology

"To generate target-specific macromolecular binders against [Target Name], we will deploy an integrated computational-experimental pipeline combining generative deep learning with physical bench validation. Initial 3D structural scaffolds will be generated de novo using the SE(3)-equivariant diffusion model RFdiffusion (v1.1.0), conditioned on target binding site coordinates derived from high-resolution Cryo-EM maps / PDB entries [PDB ID]. Sequence optimization for predicted backbones will be performed using ProteinMPNN, selecting candidate sequences based on predicted folding stability and solubility metrics.

Structural fidelity and complex interaction interfaces will be independently evaluated in silico using AlphaFold 3 / RoseTTAFold All-Atom, filtering constructs displaying interface predicted Local Distance Difference Test (pLDDT) scores > 82 and Predicted Aligned Error (PAE) < 4.5 Å across interaction domains. All computational predictions will serve strictly as structural hypotheses. Top-ranked synthetic genes will be codon-optimized, synthesized, and expressed in [E. coli / HEK293] systems. Physical binding kinetics and thermodynamic stability will be experimentally quantified via Surface Plasmon Resonance (SPR / Biacore) and Isothermal Titration Calorimetry (ITC). Computational infrastructure will utilize secure local GPU nodes hosted at IBB PAN and PLGrid supercomputing allocations, ensuring zero external transmission of unreleased structural assets."

TEMPLATE B: High-Throughput Omics & Genomic Foundation Model Methodology

"For genome-wide functional variant impact prediction, we will leverage the Evo 2 long-context genomic foundation model (Arc Institute), capable of processing single-nucleotide resolution across 1M bp genomic context windows. Variant effect scoring will be benchmarked against experimental deep mutational scanning (DMS) datasets. Single-cell RNA sequencing (scRNA-seq) integration across experimental conditions will be performed using the scVI deep generative framework (scvi-tools v1.2), accounting for zero-inflated negative binomial count distributions and removing batch effects while preserving biological cell-state manifolds.

All raw sequencing datasets (FASTQ/BAM) will be maintained within RODO-compliant local storage servers at IBB PAN. AI-driven bioinformatic scripts generated during data analysis will be manually benchmarked against standard bioconductor workflows, version-controlled via Git, and deposited in public Zenodo repositories upon manuscript submission in full compliance with FAIR data principles."

5.2. FAIR Data & Bench Reproducibility Verification

Prediction Category	Required Confidence Metric	Mandatory Bench Validation Assay
Structural Complex (AF3)	pLDDT > 80, PAE < 5 Å	Yeast Two-Hybrid (Y2H),

Prediction Category	Required Confidence Metric	Mandatory Bench Validation Assay
		SEC-MALS, Co-IP, or Cryo-EM
De Novo Binder (RFdiffusion)	*In silico* RMSD < 1.5 Å, MPNN score	Surface Plasmon Resonance (SPR), MST, or ITC binding affinity
Splicing Variant (ML Predictor)	Probability score > 0.85	RT-PCR gel quantification / Sanger sequencing

CHAPTER 6: LLM & Data Security Evaluation for IBB PAN Management

6.1. Technical, Privacy, and Intellectual Property Matrix

Model Platform	Bio-Scripting (Python/R/PyMOL)	Literature Synthesis & Context Window	RODO / Data Privacy / IP Security	Strategic Recommendation
Claude 3.5 Sonnet / Opus	High performance on bio-scripting	Logical synthesis, low hallucination rates	API/Team: No training, strict privacy	Primary choice for manuscript writing & bio-scripting.
OpenAI GPT-4o / o1	Strong reasoning capabilities	Multi-step reasoning	Enterprise / Azure: GDPR compliant	Option for multi-modal blot/gel parsing & math.
Google Gemini 1.5 Pro / 2.0	Requires prompt refinement	2M token context window	Vertex AI / Enterprise isolation	Suitable for bulk PDF literature reviews.
DeepSeek R1 / V3 (Open Weights)	Math & logic verification	Structured reasoning chains	Public API: Cloud outside EU. Local: Private	Deploy open weights locally on PLGrid/IBB GPU nodes.
Local Open Weights (Llama 3.3)	Depends on size/fine-tuning	Hardware limited context	On-premise execution	Mandatory for unreleased sequence/patent data.

6.2. Patentability Risks under EPO & UPRP Rules

Under European Patent Office (EPO Article 54 EPC) and Polish Patent Office (UPRP) statutes, an invention must possess absolute novelty worldwide. Transmitting unreleased sequences, crystallographic coordinates, mass-spectrometry files, or chemical structures into public cloud LLM interfaces (where default terms permit data logging or model retraining) constitutes public disclosure / state of the art (*prior art*). This destroys patentability. Pre-patent research assets at IBB PAN must be processed on local GPU servers or enterprise-contracted instances with zero data retention clauses.

6.3. Technology Transfer & Spin-off Guardrails

When IBB PAN research yields potential spin-off commercial assets or patent filings, the IBB Technology Transfer Office (TTO) mandates that AI models used during discovery undergo an IP Audit. Authors must produce logs of model execution, verifying that no third-party proprietary data was incorporated into training or inference prompts.