

Harshini Suresha

📍 London, UK ☎ +44 7474549265 ✉ harshinisuresha7@gmail.com 🎓 [Google Scholar](#)
💼 harshini.suresha@kcl.ac.uk 🌐 harshinisuresha.me 🗣 [Harshini-Suresha](#) in [harshini-suresha](#)

Research & Projects

Immune Cell Transcriptomics & Metabolic Modeling

IISc Research Internship Dissertation — Integrated Bulk and Single-Cell Transcriptomic and Metabolic Differences in Human and Mouse Immune Cells

Jun 2025 – Jun 2026

IISc | Bangalore, India

Supervisor: Dr. Payel Roy

- Analysed bulk and single-cell RNA-seq datasets to investigate transcriptional and metabolic differences across T-cell populations in human and mouse models of chronic inflammatory disease.
- Processed bulk RNA-seq data using FastQC, MultiQC, Trimmomatic, FastP, STAR, Quant mode, edgeR and DESeq2 for quality control, trimming, alignment, and differential gene expression analysis.
- Performed WGCNA to identify gene modules associated with immune-cell states and disease-related phenotypes, followed by module-trait correlation analysis and identification of hub genes.
- Applied constraint-based metabolic modelling using COMPASS to map transcriptomic profiles with a genome-scale metabolic network, infer reaction-level metabolic activity, quantify flux differences, and identify differentially active metabolic pathways and state-specific metabolic programs.
- Conducted single-cell RNA-seq analysis using Seurat, performing quality control, normalization, dimensionality reduction, clustering, cell-type annotation and Trajectory Inference & Pseudotime Analysis to characterize cellular heterogeneity and identify distinct immune-cell populations.

In-Silico & In-Vitro Antimicrobial Resistance

Bachelor's Capstone Dissertation — Exploring the Quorum Quenching Properties of *Premna serratifolia* on *Pseudomonas aeruginosa*

Aug 2023 – Dec 2025

PES University | Bangalore, India

Supervisor: Dr. Reshma S. V.

- Conducted a detailed wet-lab and computational study on quorum quenching in *Pseudomonas aeruginosa*, integrating plant-based phytochemical screening with microbiological approaches.
- Performed in-silico molecular docking, virtual screening, and molecular dynamics simulations using Discovery Studio BIOVIA, OpenBabel, DiffDock-L, AlphaFold 3, PyRx, PyMOL, LigPlot+, GROMACS, Chimera, PyQt5, AMBER, MMPBSA, SASA, pkCSM, ImageJ, SwissADME, SwissDock.
- Analysed 38 phytochemical compounds from *Premna serratifolia* against 38 quorum-sensing-related target proteins, generating heatmaps, ridge, violin, dot, network, pair, scatter, bar, box, PCA, volcano, correlation, clustering and histogram plots, along with molecular interaction maps.
- Executed in-vitro assays including agar well diffusion, disk diffusion, broth dilution MIC, pyocyanin quantification, swarming motility, and biofilm inhibition using the crystal violet assay to evaluate antibacterial, quorum-sensing-associated, and anti-virulence effects of *Premna serratifolia* extracts.
- Performed SEM imaging studies to visualize changes in biofilm architecture following plant-extract treatment, identifying structural disruption and reduced extracellular matrix formation.

AI-Driven RNA Therapeutics & ASO Design

RNA Therapeutic-Design Platform: Mechanism Arbitration and Candidate Generation Across Nine Therapeutic Goals

Jun 2026 – Sep 2026

NCBS | Bangalore, India

Supervisor: Dr. Tanay Bhatt

- Developed a single-pass mechanism-arbitration pipeline spanning 41 RNA therapeutic mechanisms across 9 therapeutic goals, integrating target validation, transcript-derived sequence design, RNA accessibility, and mechanism-specific scoring to generate ranked candidate oligos.
- Implemented and integrated 19 mechanism-specific features spanning SpliceAI-based splice prediction, ViennaRNA RNAplfold accessibility modelling, Ensembl/NCBI gene annotation & transcript validation, miRNA targeting, promoter accessibility, RNA-binding protein sites, alternative polyadenylation, promoter switching, and intron retention, with organism-aware resolvers, genome-scale reference datasets, provenance tracking, and citation-gated evidence handling.
- Analysed 165,237 benchmark rows across RNase H (159,109), splice-switching (2,181), and siRNA (3,947) modalities to benchmark mechanism-specific performance, characterize modality-level patterns, and evaluate therapeutic design strategies.
- Evaluated seven ML models across 12 controlled experiments (E1–E12), investigating class imbalance, gene-grouped splitting, ranking performance, cross-chemistry transfer, feature attribution, RNA architecture, thermodynamics, RNA-FM embeddings, and ML-versus- heuristic performance.
- Implemented genome-scale, multi-organism feature resolution with source-backed reference data, organism isolation, provenance tracking, confidence caps, and explicit Unresolved/Not-applicable states to prevent unsupported therapeutic inferences.
- Developed and validated the mechanism-arbitration API and workflow, with 84 mechanism-arbitration tests and 11 feature-level tests, including regression coverage for identified defects.

Scleroderma & Fibroblast Single-Cell Transcriptomics

Single-Cell Transcriptomic Profiling of Mouse Fibroblast Heterogeneity in Scleroderma

Jun 2026 – Sept 2026

NCBS | Bangalore, India

Supervisor: Dr. Tanay Bhatt

- Conducting 10x Genomics-based single-cell RNA-seq analysis to investigate scleroderma-associated heterogeneity, comparing wild-type and phenotype fibroblast populations across multiple age points.
- Performed single-cell dataset integration and batch-effect correction across multiple experimental timepoints to resolve fibroblast subpopulations and distinguish biological from technical variation.
- Characterizing disease-associated fibroblast states through canonical marker-gene analysis, pathway enrichment, GO/KEGG/Reactome and GSEA-based functional analysis, and scoring ECM, collagen, TGF- β , inflammatory, interferon, senescence, and myofibroblast activation signatures.
- Performing cell-cycle scoring, CNV inference, and gene-set pathway activity analysis to characterize proliferative states, genomic variation, and disease-associated transcriptional programs.
- Trajectory inference and pseudotime analysis using Monocle3 to reconstruct fibroblast state transitions, infer developmental trajectories, and characterize differential gene-expression dynamics across pseudotemporal states.
- Analysing cell–cell communication and regulatory networks using CellChat, CellPhoneDB, and Cell2cell to identify ligand–receptor interactions, signalling pathways, cellular niches, transcription-factor regulons, and gene regulatory programs across age points.

Explainable AI & Plant Disease Detection

TriggerNet: A Novel Explainable AI Framework for Red Palm Mite Detection and Multi-Model Comparison and Heuristic-Guided Annotation

Nov 2024 – Feb 2025 Supervisor: Sunil KC ICCET Conference | Anna University, Chennai

- Developed TriggerNet, an explainable AI framework integrating Grad-CAM, RISE, FullGrad, and TCAV to generate visual explanations for deep-learning models in plant classification and disease detection; the first authored work was presented at the 13th ICCET conference.
- Constructed a hierarchical interpretability pipeline that dynamically activates explanation modules under high prediction uncertainty or ambiguous class boundaries, while fusion of attribution methods improved explanation coverage and p-score consistency over individual interpretability methods.
- Built and evaluated a multi-model classification framework using RGB images across 11 plant species, comparing CNN, EfficientNet, MobileNet, Vision Transformer, ResNet50, and InceptionV3 architectures alongside Random Forest, SVM, and KNN classifiers.
- Implemented a four-class Red Palm Mite disease classification framework covering Healthy, Yellow Spots, Reddish Bronzing, and Silk Webbing, using Snorkel-based heuristic weak supervision to reduce manual annotation requirements and improve dataset labelling.
- Achieved 95.1% classification accuracy with EfficientNet + Random Forest, 94.2% with ResNet50 + SVM, and 94.4% detection accuracy with YOLOv8, with CNN-based detection reaching 95%; disease-level F1-scores reached 0.87 for Silk Webbing and 0.86 for Reddish Bronzing, with weighted precision and recall of approximately 0.82.
- Developed the computational workflow using Python, TensorFlow, Keras, OpenCV, NumPy, Pandas, Matplotlib, and Scikit-learn, integrating image preprocessing, model inference, and explainability outputs with an interactive Streamlit application for real-time plant infestation detection.

Neuro-Symbolic AI & Causal Pathway Discovery

PathSymetic: Neuro-Symbolic Causal Hypothesis Generation for Mechanistic Pathway Discovery in Genomic Systems

Mar 2025 – Aug 2025 | AAAI 2026 (Bridge-LM Reasoning, AI4Science for Research & SPARTA)

- First-authored PathSymetic, selected within the top 15% of accepted submissions at AAAI 2026 workshops, recognizing its integration of neuro-symbolic reasoning and causal inference.
- Developed PathSymetic, a neuro-symbolic framework integrating biological knowledge graphs, pathway ontologies, causal structure learning, and deep-learning embeddings to generate interpretable mechanistic pathway hypotheses from genomic data.
- Engineered a hybrid inference pipeline combining graph neural networks (GNNs), rule-based symbolic reasoning, and LLM-derived embeddings with biological knowledge bases including KEGG, Reactome, and BioGRID for pathway-level causal discovery.
- Constructed gene–pathway causal graphs from multi-omics evidence and biological knowledge graphs, integrating gene–gene, gene–pathway, and pathway-level relationships with ontology-based constraints for mechanistic hypothesis generation and evidence-weighted prioritization.
- Evaluated PathSymetic against GNN-based and pathway-enrichment baselines, achieving an AUPR of 0.81, F1-score of 0.77, Precision@10 of 0.94, and Hit@10/Hit@20 of 98.4%/99.6%; applied evidence-based ranking combining symbolic constraints and learned representations to prioritize experimentally supported pathways and identify novel mechanistic hypotheses.

Multi-Method Model Interpretability

A Multi-Method Interpretability Framework for Probing Cognitive Processing in Deep Neural Networks Across Vision and Biomedical Domains

Jan 2025 – Jun 2025 NeurIPS 2025 (SPIGM, Reliable ML from Unreliable Data workshops)

- First-authored research recognized among the top papers at NeurIPS 2025 workshops for contributions to interpretable and trustworthy deep learning across diverse applications in plants, vision, and biomedical domains.
- Engineered TriggerNet, a multi-method interpretability framework integrating Grad-CAM, RISE, FullGrad, and TCAV to generate class-discriminative visual explanations and quantify model reasoning across vision, agricultural, and biomedical imaging domains.
- Evaluated CNN, EfficientNet, MobileNet, Vision Transformer, and ResNet50 architectures across Red Palm Mite-affected plant images, MedMNIST PathMNIST and OrganMNIST, and CIFAR-10, spanning 11 plant species and four disease stages.
- Integrated Snorkel-based weak supervision for automated annotation and evaluated explanation quality using p-score and Brier score alongside classification performance, enabling quantitative assessment of interpretability and prediction reliability.
- Achieved classification accuracies of 97.3% on plant images, 94.2% on PathMNIST, and 92.8% on CIFAR-10, with a 21.4% improvement in p-score and 16.7% reduction in Brier score over baselines.
- Extended TriggerNet to plant-disease detection using YOLOv8, achieving 94.4% detection accuracy and class-wise F1-scores of 0.87 for Silk Webbing and 0.86 for Reddish Bronzing across disease-stage categories.
- Quantified interpretability fidelity using Pointing Game accuracy, Deletion/Insertion AUC, saliency-concept alignment, and IoU against expert annotations, demonstrating improved localization and semantic alignment over individual attribution methods.

Machine Learning & Computational Drug Discovery

Predicting Bioactivity of SARS-CoV-2 3C-like Proteinase Inhibitors Using Machine Learning

Dec 2024 – Feb 2025 PES University | Bangalore, India Supervisor: Sunil KC

- Curated 3,425 ChEMBL 3CLpro compounds with measured IC_{50} values, kept exact nM measurements, took the median of duplicates, and labelled molecules active ($\leq 1 \mu M$) or inactive ($\geq 10 \mu M$), dropping the ambiguous middle band to reduce label noise.
- Featurized molecules as 2048-bit RDKit Morgan fingerprints and benchmarked RF, XGBoost, LightGBM on a stratified 80/20 split, comparing them on AUC-ROC, F1, precision, recall.
- Stress-tested the models with scaffold-based splits, calibration analysis and review of the most confident misclassifications to check performance on chemically new molecules, and used SHAP bit attribution and Tanimoto similarity to link predicted actives to recurring chemical scaffolds.
- Built an end-to-end cheminformatics workflow integrating ChEMBL Web Services, RDKit, Scikit-learn, XGBoost, and LightGBM for bioactivity retrieval, molecular featurization, model training, and comparative evaluation and performed molecular similarity, structure-activity analysis to characterize chemical patterns among predicted 3CLpro inhibitors.
- Developed visualization and interactive analysis workflows using Seaborn and Streamlit to explore cutoff tuning, molecular features, model predictions, and comparative classifier performance.

Computational Bioprocess Engineering & Biofuels

Advanced Biorefinery Strategies for Sustainable Biofuel Production - *In-Silico* Modeling

Oct 2023 – Dec 2024

PES University | Bangalore, India

Supervisor: Dr. Kavitha S H

- Designed and implemented MATLAB-based kinetic simulation frameworks for bioethanol production using *Saccharomyces cerevisiae*, modelling microbial growth, substrate consumption, ethanol formation, and product inhibition across diverse biomass feedstocks.
- Integrated Monod and Michaelis–Menten kinetics with mass-balance equations, yield coefficients, maintenance terms, and temperature-dependent microbial growth and decay parameters to simulate fermentation dynamics and predict bioethanol production.
- Parameterized fermentation models using substrate concentration, yeast cell density, ethanol concentration, temperature, and pH, including a fermentation temperature of 30°C, initial glucose concentration of 220 kg/m³, and *S. cerevisiae* concentration of 1.5 kg/m³ for model simulations.
- Developed feedstock-specific simulation models for tropical fruit residues, including mango, banana, and pineapple, and extended kinetic modelling to potato peels, sugarcane bagasse, and *Saccharina latissima* marine algae to assess substrate-specific bioethanol production potential.
- Conducted a comparative literature analysis of 250+ studies covering tropical & temperate fruit residues, vegetable waste, lignocellulosic biomass, and marine algae, evaluating feedstock composition, pre-treatment, hydrolysis, fermentation, *S. cerevisiae* strains, reported ethanol yields.
- Analysed simulated time-course profiles of glucose depletion, yeast biomass growth, and ethanol accumulation to characterize fermentation kinetics, substrate limitation, product inhibition, and steady-state behaviour for sustainable biofuel process optimization.

NGS & Transcriptomic Data Analysis

RNA-Seq Data Analysis and Transcriptomics Workflow - Differential Expression Analysis

Oct 2024 – Jan 2025

PES University | Bangalore, India

- Processed paired-end RNA-seq datasets through an end-to-end transcriptomics workflow, performing adapter trimming with Cutadapt, splice-aware alignment with STAR, and reference-based mapping against the DM6 genome using GTF annotations on the Galaxy platform.
- Implemented multi-level sequencing and alignment quality control using MultiQC, evaluating read quality, mapping efficiency, uniquely mapped reads, splice-junction annotation, alignment statistics, and genome-wide coverage distributions across samples.
- Performed gene-level expression quantification using FeatureCounts, including strandedness assessment and optimization of counting parameters to generate sample-level expression count matrices for downstream transcriptomic analysis.
- Processed count matrices for differential gene-expression analysis using DESeq2 and edgeR, enabling identification of transcriptional changes and disease/condition expression signatures.
- Conducted alignment-level quality assessment through BAM-file inspection and IGV visualization, examining read coverage, exon-level alignment, splice junctions, and transcript structure to validate mapping and expression-quantification outputs.
- Integrated sequence preprocessing, reference-guided alignment, quality assessment, gene quantification, and differential expression into a reproducible RNA-seq analysis workflow, establishing a computational pipeline for downstream transcriptomic interpretation.

Machine Learning & Cardiovascular Risk Prediction

Machine Learning Approaches for Cardiovascular Risk Prediction in Cohort Data

Jul 2025 - Sep 2025

IISc | Bangalore, India

Supervisor: Payel Roy

- Designed an interpretable machine-learning framework for cardiovascular risk prediction from longitudinal cohort data, integrating clinical, lifestyle, laboratory, and immunological features to model heterogeneous disease-risk profiles.
- Integrated classical machine-learning models including Logistic Regression, Support Vector Machines, and Random Forests with gradient boosting, neural networks, Random Survival Forests, and DeepSurv for nonlinear and time-to-event risk modelling.
- Incorporated single-cell transcriptomic profiles, TCR repertoires, APOB-specific immune responses, and clinical biomarkers using dimensionality reduction and representation learning to capture high-dimensional immune states.
- Developed a multi-model survival framework combining Random Survival Forests, DeepSurv, SVM, and Random Forests to integrate clinical, molecular, and immunological features into unified hazard and risk estimates.
- Established evaluation using concordance index (C-index), integrated Brier score, time-dependent AUC, calibration curves, decision-curve analysis, cross-cohort transfer, and feature-set ablation to assess predictive performance, calibration, clinical utility, and generalizability.
- Aligned longitudinal clinical and immune trajectories to model temporal shifts, irregular sampling, evolving baseline characteristics, and dynamic changes in cardiovascular risk, enabling integration of time-dependent clinical measurements with evolving immune states for longitudinal risk prediction.

Multi-Omics & Computational Immunology

Stimulation-Resolved Multi-Omics Analysis of Cardiovascular & Immune Cohorts

Jun 2025 – Jun 2026 IISc | Bangalore, India Supervisors: Prof. Dipankar Nandi & Dr. Payel Roy

- Co-developed multi-omics workflows for stimulation-resolved cardiovascular and immune cohorts using CD3, CD28, PMA, and P3971 perturbation panels to characterize T-cell activation, signalling rewiring, and biomarker-resolved immune responses.
- Engineered end-to-end bulk and single-cell RNA-seq pipelines encompassing read-level processing with Fastp and Trimmomatic, splice-aware alignment and UMI-aware quantification using STAR, STARsolo, and CellRanger, and gene-abundance estimation with FeatureCounts, HTSeq, and Salmon-Alevin.
- Implemented stringent sequencing and expression QC using FastQC and MultiQC, followed by batch-aware correction with Harmony and scVI and doublet detection using Scrublet and DoubletFinder to improve cohort-scale signal quality.
- Applied DESeq2, edgeR, limma-voom, MAST, and generalized linear mixed models (GLMMs) for gene- and cluster-level differential analysis, accounting for inter-sample variation and extracting stimulation-associated transcriptional signatures.
- Built reproducible and portable computational workflows using Linux & Bash, Conda-based environment management, SLURM-based HPC infrastructure, and containerized execution with Docker and Singularity, enabling scalable and consistent deployment of computational analyses across heterogeneous computing environments.

AI & Microbiome-Targeted Therapeutics

Harnessing Bioinformatics and AI for Microbiome-Targeted Discovery of Plant-Derived Therapeutics: Integrating Phytochemistry, Multi-Omics & Nutrigenomics

Jun 2024 – Apr 2025

PES University | Bangalore, India

Supervisor: Dr. Reshma S V

- Synthesized bioinformatics, cheminformatics, and AI-driven approaches for the discovery and prioritization of plant-derived compounds with microbiome-modulating therapeutic potential.
- Applied BLAST, QSAR Toolbox, RDKit, OpenBabel, Pipeline Pilot, KNIME, MetaboAnalyst, and Cytoscape across sequence analysis, molecular descriptor generation, QSAR/ADMET profiling, workflow automation, multi-omics analysis, and compound–gene–disease network construction.
- Evaluated AI methodologies including classification, clustering, and deep learning for modelling compound–microbiome interactions, microbial biotransformation, metabolite production, and host-response pathways.
- Developed a systems-level framework integrating metagenomics, metabolomics, transcriptomics, and phytochemical data to model compound–microbiome–host interactions and microbiome-mediated therapeutic mechanisms.
- Extended the framework to personalized nutrigenomics by integrating microbial metabolism, host genetic variation, nutrient-responsive pathways, and AI-based prediction of phytochemical responses.

Structural Bioinformatics & Molecular Modelling

Structural Analysis and Preparation of the Angiotensin II Type 1 Receptor (AT1R)

Jan 2025 – Feb 2025

PES University | Bangalore, India

- Analysed the 2.10 Å X-ray structure of anti-AT1R nanobody AT118i4h32 (PDB 7T83; 8 chains × 128 residues) using BIOVIA, PyMOL, and annotating its CDR and framework regions, disulfide bond, His-tag, and checking crystallographic parameters (Matthews coefficient, solvent content).
- Annotated the 128-residue sequence into framework and CDR regions and profiled charge and Kyte–Doolittle hydrophobicity per region. This identified the arginine- and tyrosine-rich CDR1 (+3) and the 12-residue CDR3 as the loops most likely to be flexible, and confirmed the conserved Cys24–Cys97 disulfide.
- Identified preparation issues before topology building: 76 of 1,024 deposited residues unmodelled, four citrate ligands, an author-assigned monomeric assembly (justifying single-chain simulation), and no histidine protonation choices outside the His-tag.
- Prepared the system in BIOVIA (removed crystallographic waters and citrate, selected one chain, added polar waters) and planned the GROMACS protocol: minimization, NVT/NPT equilibration, production, with a box-size and run-length estimator.
- Ran a 50 ns GROMACS simulation (ff14SB, TIP3P, 300 K, 0.15 M NaCl). Backbone RMSD plateaued near 0.18 nm after 20 ns, with mean Rg 1.45 nm. RMSF peaked at CDR3, then CDR2 and the C-terminal tag, and simulated flexibility followed the crystal B-factor profile.
- Validated simulated flexibility against the crystal by converting per-residue RMSF to B-factor equivalents ($B = \frac{8\pi^2}{3} \times \text{RMSF}^2$) and comparing with the deposited PDB B-factors.
- Built an interactive project website and Streamlit app with a sequence lab, .xvg trajectory viewer, RMSF-to-B-factor converter and run planner, plus a Colab notebook.

DOE & Statistical Bioprocess Optimization

Design of Experiments and Multivariate Statistical Optimization of *Aspergillus awamori* MK788209 Cellulase Production from Agricultural Waste

Aug 2024 – Dec 2024

PES University | Bangalore, India

Supervisor: Dr. S H Kavitha

- Optimized solid-state fermentation of *Aspergillus awamori* MK788209 for cellulase production from agricultural waste using statistical design-of-experiments and response surface methodology.
- Applied Plackett–Burman Design to screen 11 process variables and identify the dominant factors influencing cellulase production, extended the analysis with leave-one-out and bootstrap stability checks, power analysis of the 12-run screen, and Monte Carlo comparison of one-factor-at-a-time, grid and CCD strategies.
- Implemented response surface methodology with random-block effects to account for block-level variation and improve estimation of factor effects during experimental optimization.
- Applied logistic regression, Weibull regression, Gaussian process models to characterize process-response relationships, factor effects, nonlinear response behaviour across experimental conditions.
- Generated and interpreted three-dimensional response surfaces, contour maps, interaction plots, and model diagnostics to investigate factor interactions and identify optimal operating regions for cellulase production.
- Performed ANOVA and statistical model assessment, obtaining an R^2 of 0.9878, adjusted R^2 of 0.9792, and predicted R^2 of 0.9136 for the optimized cellulase production model.
- Showed that a substrate $\times$ time interaction reverses the effect of incubation time, and that the stationary point is a saddle with the best condition on the design boundary. Identified a lower-cost alternative operating point giving 87% of peak titre in 58% of the time with half the substrate.

Structure-Based Ligand Profiling & Metalloprotease Inhibition

Structural and Interaction-Based Analysis of MMP-7 Inhibition Using Molecular Docking

Nov 2025 – Feb 2026

PES University | Bangalore, India

- Docked the 50 most potent known MMP-7 inhibitors from ChEMBL (ChEMBL4073) into the catalytic site of PDB 1MMQ (1.90 Å) using AutoDock Vina, after energy minimization and multi-conformer ligand preparation.
- Characterized the 1MMQ active site from crystal coordinates: a five-coordinate catalytic Zn^{2+} (His218, His222, His228 and both hydroxamate oxygens at 2.16 and 2.20 Å; $\tau_5 = 0.16$, distorted square pyramid), with 13 polar contacts and 16 residues within 4.5 Å of the inhibitor.
- Assessed each pose on affinity and Zn^{2+} chelation separately, since Vina has no explicit metal term, using closest O– Zn^{2+} distance and a bidentate-like contact criterion benchmarked against the crystal geometry. The top hit, ChEMBL41497 (–9.25 kcal/mol, O– Zn^{2+} 2.3 Å), chelated, while the second-ranked ligand (–9.05) did not (2.65 Å) derived using Pareto-front analysis.
- Defined a dual-criterion hit-selection rule combining Vina affinity with chelation fidelity (closest O– Zn^{2+} within 0.30 Å of the 2.18 Å crystal mean, bidentate-like contact), and explored trade-offs with weighted ranking and Pareto-front analysis across affinity, ligand efficiency and Zn^{2+} geometry.
- Built an interactive report website, a browser-based Docking Lab, a Streamlit app (also runnable in Docker) and a Colab notebook so the screen can be reproduced.

Comparative Genomics & Behavioural Genomics

Genomics of Behaviour in *Apis mellifera*: Comparative Genome Annotation, Sequence Analysis & Evolutionary Profiling

Jan 2026 – Apr 2026

PES University | Bangalore, India

- Investigated the genomic basis of honeybee social behaviour, focusing on the genetic and molecular architecture underlying foraging, division of labour, caste differentiation, pheromone-mediated communication, and behavioural plasticity.
- Explored the *Apis mellifera* genome using NCBI Genome Browser, BeeBase, and Ensembl Metazoa to examine chromosome organization, gene annotations, genomic coordinates, regulatory features, and alternative transcript structures.
- Performed sequence homology and comparative genomic analysis using BLAST to identify homologous sequences across honeybee and related insect genomes and investigate evolutionary conservation of behaviour-associated genes.
- Applied InterProScan for protein-domain and family-level annotation, connecting sequence features with functional classifications relevant to gene function and behavioural phenotypes.
- Conducted multiple-sequence alignment using MAFFT and the NCBI MSA Viewer, followed by phylogenetic analysis using MEGA and RAxML to compare homologous protein sequences across species, assessing conserved and variable regions for evolutionary and functional interpretation.
- Integrated quantitative-trait locus (QTL) evidence and behavioural-genomics datasets with annotated honeybee genomic regions to examine candidate loci associated with foraging, division of labour, nest defence, and other social-behavioural phenotypes, linking genomic variation to complex behavioural traits.
- Combined the comparative sequence analysis using RefSeq, BeeBase, HymenopteraMine with behavioural-genomics literature to examine genes including *for* and *per*, epigenetic regulation, pheromone signalling, and molecular pathways associated with age-dependent division of labour.
- Synthesized genomic and behavioural evidence using NCBI Genome Browser, InterProScan, and IGV to investigate haplodiploidy, caste differentiation, social immunity, olfactory specialization, and evolutionary diversification across social insects.

Protein Structural & Functional Bioinformatics

Multi-Level Structural and Functional Annotation of Neuropilin-2 (NRP2)

Aug 2025 – Oct 2025

PES University | Bangalore, India

- Ran an eight-layer in-silico annotation of NRP2 (UniProt O60462) using about 20 tools and databases, integrating sequence, topology, domain, structure and interaction-network evidence to examine its roles in neuronal signalling, angiogenesis and disease.
- Performed comparative sequence analysis (BLAST, COBALT, Clustal Omega) to build multiple-sequence alignments and phylogenetic relationships across homologues and species, and characterized composition and physicochemical properties with ProtParam (molecular weight, pI, instability and aliphatic indices, hydrophathy).
- Characterized NRP2 sequence composition and physicochemical properties using UniProt and ProtParam, evaluating molecular weight, theoretical pI, amino-acid composition, instability index, aliphatic index, and hydrophathy.

- Predicted membrane topology using TMHMM and DeepTMHMM, identifying transmembrane architecture from sequence-level Hidden Markov Model and deep-learning-based predictions.
- Mapped proteolytic cleavage susceptibility across the NRP2 sequence using PeptideCutter, profiling enzyme-specific cleavage sites and sequence regions potentially accessible to proteolytic processing.
- Resolved the domain architecture (a1/a2 CUB, b1/b2 FV/VIII, MAM, transmembrane helix, cytoplasmic tail) with InterProScan, PROSITE, CDD and PDBsum, and compared experimentally determined NRP2 structures in Mol* across crystallographic and cryo-EM entries.
- Built NRP2 protein–protein interaction networks (IntAct, STRING) and analysed connectivity (density, clustering coefficient, path length) and functional partner classes, including semaphorin–plexin and VEGF–receptor signalling.
- Built an interactive website and Colab notebook with in-browser tools: a sequence lab (ProtParam-style properties, hydropathy and TM-candidate plots, protease sites), an MSA conservation and UPGMA tree viewer, a force-directed interactome and a motif and PTM scanner.

ML for Healthcare & Predictive Analytics

Comparative Machine Learning Frameworks for Cardiovascular, Metabolic, Oncological & Neurodevelopmental Disease Prediction

Jan 2026 - Mar 2026

Independent Projects | GitHub

- Developed supervised ML workflows for cardiovascular, metabolic, oncological, and neurodevelopmental disease prediction, integrating exploratory data analysis, feature engineering, missing-value imputation, feature scaling, model selection, hyperparameter optimization, quantitative evaluation.
- Implemented artificial neural networks for heart-disease prediction, developing nonlinear classification workflows to model relationships between clinical attributes and cardiovascular diseases.
- Conducted comparative diabetes-prediction experiments using K-Nearest Neighbors, systematically evaluating feature preprocessing, StandardScaler/MinMaxScaler normalization, outlier handling, hyperparameter selection, and validation strategies; improved test accuracy from 70.1% to 75.2%.
- Developed breast-cancer classification workflows using K-Nearest Neighbors and Logistic Regression, incorporating mean-value imputation, StandardScaler normalization, 10-fold cross-validation for neighbour selection, GridSearchCV with RepeatedStratifiedKFold for hyperparameter optimization.
- Engineered behavioural and demographic predictors for Autism Spectrum Disorder classification and compared Logistic Regression, Decision Trees, Random Forest, SVM, Gradient Boosting, and XGBoost models across alternative feature representations and preprocessing strategies.
- Applied systematic model-selection and hyperparameter-tuning strategies using GridSearchCV, RandomizedSearchCV, cross-validation, and repeated stratified validation to evaluate generalization and identify robust model configurations across heterogeneous healthcare datasets.
- Evaluated predictive performance using accuracy, precision, recall, F1-score, confusion matrices, classification reports, and ROC–AUC, comparing model discrimination, class-specific performance, and generalization across disease-prediction tasks.
- Built reproducible Python-based analytical workflows using Pandas, NumPy, Scikit-learn, XGBoost, LightGBM, Matplotlib, Seaborn, and Joblib for data preprocessing, statistical exploration, feature transformation, model training, hyperparameter optimization, evaluation, and visualization.

Data Structures & Algorithms (DSA)

RouteOpt - Design and Implementation of Graph-Based Route Optimization, Shortest-Path Algorithms & Efficient Data Structures

Jun 2024 - Aug 2024/ Dept. of Computer Science, PES University/ Supervisor: Badri Prasad V R

- Designed a graph-based navigation system representing locations as vertices and routes as weighted edges (generated grid maps of 36 to 100 locations), supporting shortest-path computation, route discovery, and efficient location search over large-scale map data.
- Implemented and compared Dijkstra (binary-heap priority queue, $O((V+E) \log V)$), A*, Bellman-Ford ($O(V \cdot E)$) and BFS on identical maps and start-destination pairs, measuring nodes visited, edges checked, heap activity, route cost and runtime.
- Built a benchmarking framework that runs all four algorithms across every pair on the 36-location map (630 pairs) and across sampled pairs on 64- and 100-location maps. It reports route quality against the cheapest route and the effect of wider road-weight ranges, and measures corner-to-corner scaling on grids from 25 to 900 nodes in Python (up to 6,400 live in the browser), with all datasets exported as CSV.
- Implemented BFS and DFS for connectivity, cycle detection and reachability, and Kahn's topological sort for dependency-ordered directed graphs, including cycle detection.
- Built a Trie for prefix autocomplete of location names ($O(k)$ per prefix) and a hash table for $O(1)$ lookup of locations and metadata.
- Developed an interactive site and Streamlit app that animate each search, with a "compare all four" mode, plus a Colab notebook so every chart can be reproduced.

Operating Systems & Systems Programming

KernelSim - Comprehensive Operating System Simulation, Process Scheduling, Memory Management & Concurrent Resource Management System

Jun 2024 - Aug 2024/ Dept. of Computer Science, PES University/ Supervisor: Chandravva Hebbi

- Designed and implemented a modular OS simulator in C/C++ modelling process creation, PCBs and state transitions, CPU scheduling, virtual memory, synchronization, IPC and file management, inter-process communication, and file-system operations.
- Implemented FCFS, SJF, Priority, Round Robin, and Multilevel Feedback Queue scheduling algorithms, generating Gantt-chart execution traces and comparing waiting time, turnaround time, response time, throughput, and CPU utilization.
- Built a virtual-memory subsystem with paging, page-table address translation, page-fault handling and FIFO, LRU and Optimal replacement, analysing fault rates across frame counts and reproducing Belady's anomaly. Modelled TLB effective access time, page-size trade-offs and flat vs two-level page tables.
- Implemented concurrency with POSIX threads, mutexes and semaphores, demonstrating race conditions (lost updates with and without a mutex) and enumerating all 20 interleavings of a three-step counter increment. Implemented producer-consumer, dining philosophers, and deadlock detection and avoidance (Banker's algorithm).
- Developed IPC with pipes, shared memory and message queues, and a simulated file system with file creation, read/write, directory structures, allocation, fragmentation analysis on a 32-block disk.

- Built a Linux workflow with C/C++, system calls, shell scripting and process-level resource monitoring, plus an interactive site, Streamlit app and Colab notebook with a live scheduler runner, time-travel debugger and “be the scheduler/pager” challenges.

Design & Analysis of Algorithms (DAA)

AlgoBench - Advanced Algorithm Design, Computational Complexity Analysis, Optimization Paradigms, Algorithmic Performance Evaluation & Large-Scale Benchmarking System

Sep 2024 - Dec 2024/ Dept. of Computer Science, PES University/ Supervisor: Jayapriya D

- Implemented and benchmarked algorithms across eight paradigms (divide and conquer, dynamic programming, greedy, backtracking, branch and bound, graph algorithms, randomized and approximation) in 14 benchmark suites, including sorting, Fibonacci, 0/1 knapsack, shortest paths, MST (Kruskal vs Prim), max-flow (Edmonds–Karp), N-Queens, TSP, LCS and vertex cover.
- Derived theoretical bounds using recurrences, recursion trees, the Master theorem and amortized analysis, applying Big- O , Big- Ω and Big- Θ notation; built a Master-theorem solver and an amortized-cost demonstration for dynamic arrays.
- Designed a benchmark protocol: every optimized algorithm is checked against a reference (sorted(), DP optimum, Bellman–Ford) before timing, using best-of-repeated-runs timing, deterministic node and operation counts where time is noisy, and tracemalloc memory profiling.
- Fitted measured performance against theoretical complexity using $c \cdot f(n)$, R^2 , empirical exponents on a 162-point Python reference dataset, quantified greedy knapsack solution quality against the dynamic programming optimum, vertex-cover approximation ratios against exact solutions.
- Studied measurement noise (warm-up and percentile analysis over 60 runs) and the effect of randomization (random-pivot quicksort over 1,000 trials), and documented what the benchmarks can and cannot show (relative scaling and crossover points, but not absolute speeds or adversarial worst cases)
- Implemented backtracking and branch-and-bound strategies for combinatorial search problems, incorporating state-space trees, pruning heuristics, bounding functions, and search-space reduction to improve computational tractability.
- Mapped NP-completeness polynomial reductions ($\text{SAT} \rightarrow 3\text{-SAT} \rightarrow \text{Clique/Vertex Cover} \rightarrow \text{TSP}$, and $\text{Subset Sum} \rightarrow \text{Knapsack}$), and built an interactive site, Streamlit app and Colab notebook with live benchmarks, algorithm animations and tools to estimate the complexity of your own code.

Database Management Systems & SQL

DataSphere - Polyglot Database Architecture, Advanced SQL Analytics, NoSQL Data Management & Graph-Based Data Processing System

Jan 2025 - Mar 2025/ Dept. of Computer Science, PES University/ Supervisor: Priya Badrinath

- Designed a polyglot architecture for an e-commerce dataset, placing orders and payments in MySQL (ACID, joins), product and review documents in MongoDB (variable attributes) and the customer–product–follow network in Neo4j (multi-hop traversal). The choice for each was based on data shape, access pattern and consistency needs, with MySQL as the system of record.
- Engineered the MySQL relational layer using ER modelling, relational algebra, functional dependencies, normalization through 3NF/BCNF, primary and foreign keys, complex joins, nested queries, CTEs, recursive CTEs, window functions, stored procedures, triggers, and transaction management.

- Implemented advanced MySQL indexing and query-optimization strategies using B⁺-tree indexes, composite indexes, covering indexes, query execution plans, EXPLAIN/EXPLAIN ANALYZE, cardinality estimation, join-order optimization, predicate pushdown, query rewriting, and materialized views.
- Modelled MongoDB documents (embed vs reference rules, JSON-schema validation, compound and text indexes following the ESR rule, aggregation pipelines) and Neo4j graphs (Cypher, variable-length paths, shortest path, PageRank and degree centrality, recommendations within 3 hops).
- Implemented Neo4j graph-based data modelling using nodes, relationships, properties, graph traversal, Cypher queries, variable-length paths, pattern matching, centrality analysis, and relationship-based querying for highly connected datasets.
- Implemented ACID transaction workflows in MySQL and investigated concurrency control through serializability, two-phase locking, isolation levels, deadlock detection, MVCC concepts, commit/rollback semantics, and concurrent transaction scheduling.
- Designed a cross-database benchmark protocol with warm-up exclusion, median/p95/p99 latency, cold vs. warm cache testing, and one-variable-at-a-time experiments; verified execution plans using EXPLAIN, .explain() and PROFILE, and built predictors for high-value customers, order amounts and review ratings across all three databases.
- Built cross-database data-integration and benchmarking workflows to evaluate query latency, throughput, indexing performance, storage overhead, transaction behaviour, graph traversal efficiency, document-query performance, and scalability across MySQL, MongoDB, and Neo4j.

Algorithms, Operating Systems & Database Systems

GenomeSearch - High-Performance Genomic Sequence Search, Algorithmic Indexing, Graph-Based Workflow Scheduling, Concurrent Processing & Data Management System

Jan 2026 – May 2026

Independent Software Project

- Built an integrated system for indexing, searching, storing and processing DNA and protein sequences, with algorithmic search, a workflow scheduler, a concurrent execution engine and a PostgreSQL store sharing one job model so each stage can be benchmarked alone or end to end.
- Implemented k-mer indexing, hash-table retrieval, trie prefix search and Smith–Waterman alignment (local and global, with a two-row memory optimization when only the score is needed), and analysed time and space complexity. Measured the index build-versus-scan break-even, k-mer occurrence density and seed sensitivity under mutation.
- Designed a dependency-aware workflow scheduler using a DAG, topological sorting and a priority queue that starts the longest ready task first. Benchmarked it on a 32-task, six-stage bioinformatics workflow with critical-path analysis, speedup against worker count and scheduler overhead.
- Developed a multi-process and multi-threaded execution engine using OS-level process management, CPU scheduling, synchronization primitives, inter-process communication, and controlled resource allocation for concurrent sequence-analysis jobs.
- Designed and normalized a relational PostgreSQL database for sequences, samples, analysis jobs, execution metadata, and search results, implementing primary/foreign keys, constraints, joins, transactions, and database indexing.
- Built a reproducible benchmarking framework covering runtime, memory, scalability and database query performance on real NCBI RefSeq data (a 154 kb Arabidopsis chloroplast genome and a 9.6 kb Yersinia pestis plasmid), with an interactive site, Streamlit workbench and Colab notebooks.

AI & ML in Precision Agriculture & Bioengineering

Synergistic AI Architectures and Bioengineered Systems for Next-Generation Agricultural Pest Detection and Control

Oct 2024 - Dec 2024

PES University| Bangalore, India

Supervisor: Dr. Dinesh M S

- Reviewed 150+ sources on AI-driven pest detection, evaluating CNN architectures (EfficientNet, ResNet, Inception), vision transformers (ViT), DETR-based object detection, and handcrafted feature descriptors (SIFT, HOG) for precision agriculture.
- Examined Random Forest, SVM, and gradient-boosting methods (XGBoost, LightGBM, CatBoost), alongside ensemble strategies, for pest classification and predictive modelling.
- Evaluated bioengineered pest-control approaches integrating logic gates, neuromorphic circuits, biosensors and 3D bioprinting for biomimetic traps, emphasizing precision targeting and reduced ecological impact.
- Assessed edge-AI architectures and sensor-fusion strategies for real-time pest monitoring, considering on-device inference, multimodal data integration, computational latency and low-power deployment for precision agriculture.

Patient-Reported Drug Review Analytics & Ranking

Condition-Stratified Drug Ranking Using Bayesian Shrinkage and Interactive Analytics

Apr 2025 - Jul 2025

Supervisor: Sunil K C

- Engineered an end-to-end analytics pipeline over 215,063 heterogeneous patient-reported drug reviews from UCI/Drugs.com, covering 836 conditions and 3,654 drugs. Curated the dataset to 212,698 records by removing 1,171 scraping-induced label contaminations and normalising entities and text encodings.
- Developed a shrinkage-based drug-ranking method across 8,586 drug-condition pairs, improving the stability of ratings for sparsely reviewed drugs by incorporating condition-level averages.
- Characterised the rating distribution as strongly U-shaped, with modes at ratings of 1 and 10, demonstrating that mean-only summaries can misrepresent patient-reported sentiment. Incorporated distribution-aware visualisation into the interface to support more nuanced interpretation.
- Detected temporal non-stationarity in the corpus: mean ratings declined from approximately 8.9 in 2008 to 6.0 in 2017, while annual review volume increased roughly sixfold from its early-2010s level. Documented potential reviewer-population and platform drift as limitations affecting longitudinal comparisons.
- Developed an interpretable, condition-specific drug-ranking framework integrating shrinkage-based estimation, review-count transparency and rating-distribution analysis, with precomputed sufficient statistics enabling real-time sensitivity analysis and efficient ranking updates; investigated temporal distribution shifts and documented potential reviewer-population drift, sparse-data bias and confounding limitations.
- Architected a zero-backend inference layer that precomputes sufficient statistics for each condition-drug stratum, enabling interactive ranking, reverse drug lookup, and live shrinkage-parameter sensitivity analysis entirely client-side in an approximately 4 MB application.
- Explicitly bounded the system's inferential claims in an application-level validity analysis, acknowledging the absence of causal identification, self-selection bias, uncontrolled confounding (including dose, severity, and treatment duration), and the possibility of duplicate or fabricated reviews.

Lab Protocol Automation & Validation

Computational Framework for Structured Laboratory Protocol Representation, Automated Experimental Workflow Validation, and Reproducibility Assessment

Aug 2026 - Oct 2026

Independent Software project

- Developed a static checker that simulates liquid-handling protocols in software before a robot executes them, tracking liquid volume in every source tube and well and flagging steps that cannot work as written. Each failure is reported with its step number, rule code and a one-line reason.
- Implemented six interpretable validation rules (pipette range, reagent exhaustion, well overflow, non-existent wells, missing positive/negative controls, and an opt-in dead-volume check). The single-pass simulation runs in $O(S)$ time and validates typical workflows of up to 25 steps in under about 2 ms in the browser.
- Defined a JSON protocol schema inspired by the Laboratory Automation Protocol (LAP) format (ACS Synthetic Biology, 2023), parameterised by plate format, reagent stocks, pipette limits, well capacity, dead volume and control layout. Limits were grounded in Opentrons OT-2 pipette ranges and labware definitions, with plate maxima from Corning and NEST datasheets.
- Designed a two-tier evaluation. Benchmarks used randomly generated protocols (8–24 steps) with one known fault injected per protocol (300 cases per error type), with Wilson 95% confidence intervals and seed-reproducible results. Real-world validation covered four transcribed Opentrons Protocol Library protocols (serial dilution, restriction digest, PCR distribution) with 0 false alarms, and a retrospective check on shipped library faults that caught an example-input overdraw.
- Built an interactive zero-dependency, zero-backend web application (vanilla HTML/JS, no build step), with an animated 96-well plate simulator, a click-to-build protocol composer that re-checks after every action, a paste-and-validate interface that keeps data in the browser, in-browser analysis charts (reagent levels, plate state, limit sensitivity), and Method, Results and Report pages.

BodyTrack Atlas: Open Imaging Dataset Index

Anatomy-Linked Cross-Species Index Dataset Discovery with In-Browser Analysis

Aug 2026 - Oct 2026

Independent Software project

- Built an interactive discovery platform indexing 500+ open imaging datasets across 11 model organisms (human, mouse, rat, macaque, marmoset, pig, rabbit, zebrafish, Xenopus, Drosophila, C. elegans), navigated through a three-step organism, anatomical region, dataset workflow so researchers can locate relevant data by the body part they study rather than by repository.
- Aggregated and curated 400+ human entries from the OpenMedLab Awesome-Medical-Dataset catalogue and the TCIA collections archive, alongside curated model-organism resources. Every source link was HTTP-verified, and each entry carries mirrors, a suggested analysis pipeline and a citation.
- Designed species-specific region profiles, with a clickable human body figure (skin, muscle, skeleton and organ layers, front and back views) and embedded 3D microCT/CT models for other species where free scientific models exist. Region selection automatically switches the 3D model and the dataset list.
- Implemented a dataset workbench with a built-in image and CSV analyser that runs entirely client-side, so data can be inspected without upload or server-side processing. A lesion-timeline demo shows longitudinal tracking with hotspot and alert panels.

Projects Portfolio

Project	Links
3CLPro Bioactivity	Website App GitHub
DataSphere-ML	Website App GitHub Colab
Genome Search	Website App GitHub Colab
AlgoBench	Website App GitHub Colab
ML Healthcare AI	Website App GitHub Colab
RouteOpt	Website App GitHub Colab
KernelSim	Website App GitHub Colab
NeuroLipin	Website GitHub Colab
Angiotensin AT1R	Website App GitHub Colab
Cellulase-by-Design (DoE)	Website App GitHub Colab
MMP-7 Inhibition	Website Docking Lab App GitHub Colab
Patient Drug Analysis	Website App GitHub Colab
Honey Bee Genomics	Website GitHub Colab
LabCheck (Lab Protocol)	Website GitHub
BodyTrack Atlas	Website GitHub

Paper Implementations

Implementation	Source Paper	Links
scRNA-seq Benchmark Replication	Nature Biotechnology (DOI: 10.1038/s41587-020-0469-4)	Website GitHub Data
RNA-seq Replicability Replication	PLOS Computational Biology (DOI: 10.1371/journal.pcbi.1011630)	Website GitHub Data
NetPert Twist1 Atlas	PLOS Computational Biology (DOI: 10.1371/journal.pcbi.1012195)	Website GitHub Data
iDrug Reproduction	PLOS Computational Biology (DOI: 10.1371/journal.pcbi.1008040)	Website GitHub Data
InterProScan 6 Repro Lab	Bioinformatics Advances (DOI: 10.1093/bioadv/vbag141)	Website GitHub Data
ProFAB Replication	Briefings in Bioinformtics (DOI: 10.1093/bib/bbac627)	Website GitHub Data
MAGICAL Reproduction	PLOS Computational Biology (DOI: 10.1371/journal.pcbi.1012336)	Website GitHub Data