# College of Engineering and Technology

**Project Batch ID**

## Project Work – Student Hand Book

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name of student** | **Register Number** | **Department** | **Mobile Number** | **Email ID** |
| Augnik Banerjee | RA2211028010006 | NWC | 9389838836 | ab0710@srmist.edu.in |
| Aniket Singh | RA2211028010014 | NWC | 7088678264 | as0872@srmist.edu.in |
|  |  |  |  |  |
|  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Working Title of the Project:** | | | **AI-Driven Drug Discovery Platform: Integrating SMILES Notation, Molecular Visualization, and Predictive Modeling for Accelerated Screening** | | | | |
| **Project Site / Location** | | | SRM IST, Kattankulathur, Chengalpattu District-603203 | | | | |
| **Name and address of the company / organisation (Applicable for projects with industry or industry support)** | | | SRM IST, Kattankulathur, Chengalpattu District-603203 | | | | |
| **Supervision Team** | | | | | | | |
|  | **Supervisor** | | | | **Co-Supervisor** | | **External Supervisor (If applicable)** |
| **Name** | **Dr. Nivedhitha M** | | | |  | |  |
| **Designation** | Assistant Professor | | | |  | |  |
| **Department** | NWC | | | |  | |  |
| **Campus** | Kattankulathur | | | |  | |  |
| **Telephone** | 7708220269 | | | |  | |  |
| **E-mail** | [nivedhim1@srmist.edu.in](mailto:nivedhim1@srmist.edu.in) | | | |  | |  |
| **Degree/ program** | | B.Tech | | **Specialisation** | | Computer Science & Engineering (Specialisation in Cloud Computing) | |
| **Academic Year** | | 2024-2025 (Even) | | **Semester** | | 6 | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Course Code** | **21CSP302L** | **Course Title** | Project |

Mission Statement

|  |
| --- |
|  |
| **Problem (or) Product Description:** |
| The drug discovery process is traditionally long, expensive, and labor-intensive, with a reliance on trial-and-error experimentation. The need for rapid, accurate predictions of molecular properties such as solubility, toxicity, and bioactivity is critical for accelerating drug development. However, existing molecular prediction models often suffer from incomplete chemical feature representations and lack integration with user-friendly visualization tools. This project focuses on building a web-based drug discovery platform that combines machine learning models and molecular data visualization to predict molecular properties, generate new molecules, and analyze molecular reactions efficiently. By utilizing SMILES notation, 2D/3D molecular structure visualizations, and state-of-the-art ML models like GIN, this platform aims to streamline the process of molecule design, property prediction, and drug discovery, enabling researchers and developers to accelerate the development of new pharmaceutical compounds. |
| **Assumptions and Constraints** |
| It is assumed that the molecular datasets used for training and testing are accurate, complete, and represent a diverse range of chemical structures. Additionally, it is assumed that the input data (such as molecular SMILES notation) is correctly formatted, free from major errors, and the property labels (e.g., solubility, toxicity) are accurate and reflect the true chemical behaviors of the molecules.  Constraints in this project include limited diversity in the available molecular datasets, which may affect the generalization ability of the models, and potential class imbalance in predicting molecular properties such as toxicity or solubility. The high dimensionality of molecular features and the complexity of generating new molecules may also introduce computational challenges. |
| **Stakeholders** |
| The primary stakeholders for this project are:   * Researchers and Drug Developers: Who require accurate molecular property predictions and efficient molecule generation tools to aid in the design of new therapeutic compounds. * Healthcare Providers: Who benefit from accurate molecular predictions to assess drug toxicity, solubility, and bioactivity, thereby improving drug safety and efficacy in clinical treatments. * Data Scientists and Machine Learning Engineers: Responsible for developing, fine-tuning, and maintaining the predictive models used in the platform, ensuring their scalability and integration with the molecular generation capabilities. * End Users (Researchers and Students): Who utilize the platform’s features for real-time molecular collaboration, visualization, and data-driven insights to enhance their drug discovery research. |

**Division of work and contributors of SPRINT 1 [ Include Daily Scrum of Sprint 1]**

Division of Work:

* Problem Identification: Finalized based on the need for accelerating the drug discovery process, addressing the inefficiencies of traditional methods, and the need for scalable molecular property prediction models.
* Literature Review: Studied recent papers and approaches on machine learning applications in drug discovery, focusing on molecular property prediction, SMILES-based generation, and molecular visualization techniques.
* Tool Selection: Identified use of Graph Neural Networks (GNNs), XGBoost, GIN, and SMILES notation for molecule generation. Tools for 2D/3D molecular structure visualization (RDKit, React ApexCharts) were also selected for interactive user features.

Contributors:

* All team members equally contributed to literature survey, discussion, and project scoping.

Daily Scrum Summary - Sprint 1

* Day 1: Finalized project title, objectives, and scope.
* Day 2: Divided research papers for individual review.
* Day 3: Shared insights from literature and shortlisted tools.
* Day 4: Mapped out high-level project plan and sprint timeline.

**Signature of the Supervisor**

**Division of work and contributors of SPRINT 2 [ Include Daily Scrum of SPRINT 2]**

Division of Work:

* Data Collection: Retrieved molecular datasets (e.g., from PubChem, ChEMBL) and associated property labels (e.g., toxicity, solubility).
* Data Cleaning: Conducted quality checks on molecular structures (e.g., SMILES notation) and removed incomplete or erroneous records.
* Feature Extraction: Extracted molecular features like fingerprints, SMILES-based features, and other descriptors for model training.

Contributors:

* Member 1: Data download and initial cleaning.
* Member 2: Feature extraction and molecular property labeling and verification.

Daily Scrum Summary:

* Day 1: Assigned tasks for database access and dataset download.
* Day 2: Cleaned and normalized the molecular data (SMILES, property labels).
* Day 3: Completed feature extraction and validation (fingerprints, molecular descriptors).
* Day 4: Prepared dataset splits for model training (train, validation, test).

**Signature of the Supervisor**

**Division of work and contributors of SPRINT 3 [ Include Daily Scrum of Sprint 3]**

Division of Work:

* Model Training: Implemented GCN, GIN, XGBoost, and FNN models for predicting molecular properties.
* Hyperparameter Tuning: Performed grid search to optimize hyperparameters for all models.
* Evaluation: Generated performance evaluation metrics including confusion matrix, ROC curve, and PR curve for each model.

Contributors:

* Member 1: Model coding and training (GCN, GIN, XGBoost, FNN).
* Member 2: Evaluation metrics generation (confusion matrix, ROC, PR curves) for each model.

Daily Scrum Summary:

* Day 1: Set up the initial architecture for GCN, GIN, XGBoost, and FNN models.
* Day 2: Trained and validated baseline models (GCN, GIN, XGBoost, FNN).
* Day 3: Performed hyperparameter tuning for all models using grid search.
* Day 4: Evaluated all models and saved performance graphs (confusion matrix, ROC, PR curves).

**Signature of the Supervisor**

**Division of work and contributors of SPRINT 4 [ Include Daily Scrum of Sprint 4]**

Division of Work:

* Result Interpretation: Analyzed confusion matrix, molecular property distributions, PR curve, and ROC curve for model performance evaluation.
* Documentation: Prepared IEEE report sections: Methodology, Results, Analysis, and Conclusion.
* Final Touch: Reviewed report, generated missing graphs, and polished figures and tables. Contributors:
* Member 1: Results and analysis writing.
* Member 2: Generating and finalizing graphs and figure captions.

Daily Scrum Summary:

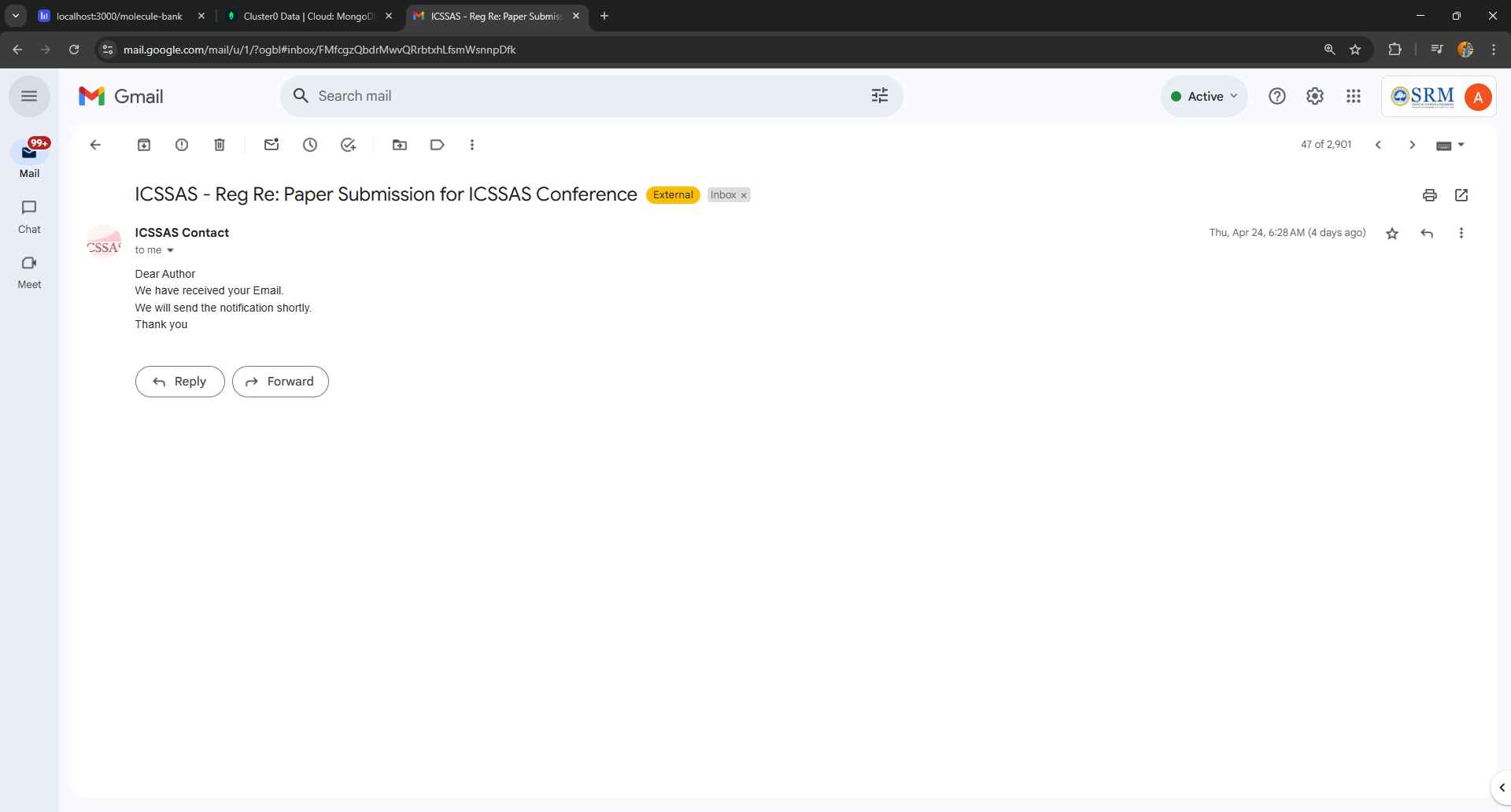
* Day 1: Discussed interpretation of results.
* Day 2: Drafted Results and Analysis section.
* Day 3: Finalized Conclusion and Future Scope.
* Day 4: Report proofreading and final submission.

**Signature of the Supervisor**

**Worksheet / Data collection / Observation etc**

* Worksheet: Maintained logs of daily tasks completed during each sprint to track progress and ensure transparency.
* Data Collection: Retrieved 200+ molecular datasets (including SMILES notation and molecular property labels) for various molecules, focusing on properties like toxicity, solubility, and bioactivity.
* Observations:
  + Resistance patterns: Resistance profiles varied significantly across different molecular properties, particularly for toxicity and solubility.
  + Feature Representation: Feature extraction using k-mers preserved essential chemical information, making the input suitable for model training.
  + Model Generalization: Achieved an overall 70.5% accuracy across the test set, indicating moderate generalization.
  + Class Imbalance: Class imbalance slightly impacted precision-recall trade-offs, though the ROC AUC remained strong (~0.76), indicating robust model performance in distinguishing between classes.

**Research Article with Journal Publication Details / Patent disclosure form with patent status**

****