Student Project Proposal
ActiveLearn-ED: Active Learning for Experiment Selection and Model Improvement in Drug Discovery

Master area	AI for Drug Discovery; active learning; chemoinformatics; experimental design; predictive modelling
Project type	Computational project with Python implementation, data analysis and decision-support workflow
Duration	2-3 months
Supervision	One tutor or co-tutor; weekly or bi-weekly progress meetings
Recommended level	Basic Python, pandas/scikit-learn and RDKit; basic knowledge of machine learning and drug discovery is useful
Main output	Documented notebook/workflow, ranked molecule and assay recommendations, short technical report and final presentation
Possible datasets	Tutor-provided molecule-assay tables; small public QSAR/ADME datasets; ChEMBL-derived subsets; synthetic example matrices

Short description. The project asks the student to prototype an active learning workflow that recommends the next experimental measurements in a drug discovery campaign. The first development phase selects new molecules expected to improve a predictive model. The second phase extends the same logic to the selection of the most useful molecule-assay pair, with the aim of reducing redundant or low-value experiments while increasing the information gained from each experimental cycle.

1. Background and rationale

Experimental datasets in drug discovery are typically small, noisy, heterogeneous and incomplete. A conventional machine learning workflow uses the available data to train a model and then ranks new compounds by predicted activity or property. This is useful, but it does not answer the more strategic question: which new data should be generated next to improve the model and accelerate decision-making?

Active learning addresses this problem by closing the loop between prediction and experimentation. Instead of selecting only the molecules predicted to be best, the system selects the measurements that are expected to reduce uncertainty, expand the useful applicability domain and improve future predictions. In this project, the student will implement a compact proof of concept focused on active learning for experimental design in drug discovery.

A critical aspect is that chemical novelty must be controlled. A molecule that is too different from the training set may not improve the current model; it may act through a different mechanism of action or belong to a separate SAR regime. Therefore, active learning must be combined with applicability-domain and mechanism-of-action consistency checks.

2. Project objective

The objective is to implement and test a Python workflow that recommends new experiments through two progressive active learning phases. Phase 1 selects molecules that are expected to improve a single predictive model, for example a pIC50 or ADME model. Phase 2 ranks missing molecule-assay measurements and recommends which assay should be performed next for each molecule or project stage.

The project should remain realistic for a 2-3 month student activity. The minimal successful version may use a small tabular dataset with SMILES and one endpoint. A recommended version will add uncertainty, chemical diversity, applicability-domain control and a basic assay-selection layer.

3. Specific aims

- Design a clean input format for molecules, SMILES, experimental endpoints, assay identifiers and missing values.
- Validate and standardize the input, including SMILES parsing, duplicate detection, unit conversion and basic quality flags.
- Build a baseline predictive model with uncertainty estimation using an ensemble or related strategy.
- Implement active learning scores based on uncertainty, controlled diversity, predicted value and applicability-domain risk.
- Introduce a mechanism-of-action consistency filter to separate model-building molecules from exploratory molecules.
- Extend the workflow from molecule selection to molecule-assay selection in sparse experimental matrices.
- Generate ranked output tables explaining which experiments should be performed, delayed or avoided.

4. Proposed methodology

4.1 Input data and validation

The first input will be a molecule table containing molecule_id, SMILES and one experimental endpoint y. Rows with known y values form the training set, while rows with missing y values are candidate molecules. The workflow will check that required columns are present, that SMILES can be parsed, that values are numeric and that enough measured molecules are available to train a meaningful model.

For Phase 2, the input becomes a molecule-assay matrix. Each row is a molecule and each column is an endpoint such as pIC50, solubility, PAMPA, microsomal stability or hERG. Missing cells represent possible future experiments. A second table will describe the assays, including approximate cost, duration, priority, prerequisites and experimental level.

4.2 Phase 1: active learning for molecule selection

Each molecule will be represented using Morgan fingerprints and simple physicochemical descriptors such as molecular weight, logP, TPSA, hydrogen-bond donors and acceptors, rotatable bonds and ring count. The baseline model may be a Random Forest, Gradient Boosting model or another scikit-learn estimator. Uncertainty can be estimated with an ensemble of models trained on different bootstrap samples or data splits.

For each candidate molecule, the workflow will calculate predicted value, predictive uncertainty, distance from known molecules and applicability-domain status. The active learning score will combine these terms. In the model-building stage, uncertainty and controlled diversity should be weighted more strongly than predicted value. In the optimization stage, the predicted objective can receive more weight.

To address the risk of introducing molecules with a different mechanism of action, candidates will be classified as inside-domain, frontier or outside-domain. When structural or target information is available, additional checks may include similarity to known ligands, pharmacophore consistency, docking in the same binding site, preservation of key interactions and absence of obvious assay-interference alerts. The output should distinguish molecules recommended for model improvement from exploratory molecules that should not be automatically merged into the main training set.

4.3 Phase 2: active learning for assay selection

The second phase does not require a new multi-objective model at the beginning. The simplest implementation uses separate single-task models for each endpoint and a common decision layer that scores the missing cells of the molecule-assay matrix. The question becomes: should the next experiment measure pIC50 for molecule M3, solubility for molecule M1, PAMPA for molecule M2, or hERG for molecule M5?

For each missing molecule-assay pair, the workflow will estimate uncertainty, expected decision impact, project priority, cost, time and experimental risk. The score should prioritize measurements that can change a go/no-go decision, remove a bottleneck, avoid later expensive assays, or improve a model that is currently limiting project decisions. This phase is therefore mainly a strategy to reduce useless experiments and increase information per euro or per week spent.

An optional advanced extension is a multi-task or matrix-completion model. This may be useful when the matrix contains many correlated endpoints and many missing values. However, multi-task modelling should be treated as an extension, not as a requirement for the first working prototype.

5. Indicative work plan

Period	Main task	Expected output
Weeks 1-2	Problem definition and data schema	Short technical note describing input tables, endpoint definitions, assay metadata and decision criteria.
Weeks 3-4	Data validation and feature generation	Notebook functions for SMILES validation, descriptor calculation and train/candidate split.
Weeks 5-6	Predictive model and uncertainty	Baseline model, cross-validation metrics and ensemble-based uncertainty estimates.
Weeks 7-8	Phase 1 active learning ranking	Ranked molecule table with predicted value, uncertainty, diversity, domain status and recommendation type.
Weeks 9-10	Phase 2 assay-aware decision layer	Sparse molecule-assay matrix, missing-cell scoring, cost/time weighting and assay recommendations.
Weeks 11-12	Final integration and reporting	Final notebook, example outputs, technical report and 10-15 minute presentation.

6. Expected deliverables

- A documented Python notebook or small code repository implementing the active learning workflow.
- A validated input example with old measured data, candidate molecules and, for Phase 2, a sparse molecule-assay matrix.
- A Phase 1 ranking table identifying molecules recommended for model building, exploratory testing or exclusion.

- A Phase 2 ranking table identifying the most useful molecule-assay measurements, including cost/time-aware prioritization.
- Basic visualizations of uncertainty, chemical diversity, applicability domain and selected experimental batches.
- A final technical report of approximately 3-5 pages and a final presentation of approximately 10-15 minutes.

7. Skills acquired by the student

- Practical implementation of active learning in a drug discovery context.
- Handling of molecular structures, descriptors, sparse experimental matrices and model-ready data tables.
- Use of uncertainty, diversity and applicability-domain concepts to guide experiment selection.
- Design of transparent decision scores that combine scientific value, cost, time and experimental priority.
- Critical understanding of the difference between predictive modelling, mechanism-of-action consistency and decision optimization.

8. Suggested software and resources

The project can be implemented in Python using pandas, NumPy, RDKit, scikit-learn, scipy, matplotlib and optionally NetworkX for graph-style representations. The tutor should provide a small molecule dataset or help the student prepare a manageable public dataset. ChEMBL-derived subsets, small QSAR benchmarks or synthetic sparse matrices can be used for testing the logic before moving to internal experimental data.

9. Evaluation criteria

Criterion	What should be assessed
Scientific clarity	The student clearly defines the active learning question, the endpoint, the experimental decision and the limits of the dataset.
Technical implementation	The code is readable, documented, reproducible and able to generate ranked experimental recommendations.
Active learning relevance	The workflow uses uncertainty, diversity and domain control rather than simple predicted activity alone.
Decision value	The Phase 2 layer clearly prioritizes experiments by information, cost, time and expected impact on decisions.
Critical discussion	The report discusses limitations, mechanism-of-action risks, data quality and possible multi-task extensions.

10. Optional extensions

Possible extensions include conformal prediction, comparison of acquisition functions, batch diversity optimization, integration of docking or pharmacophore filters, use of graph neural networks, multi-task learning, matrix completion, or a dashboard for interactive experiment planning. For stronger students, the workflow may be connected to a real internal project, provided that data curation and confidentiality constraints are clearly defined.

Note for tutors. The minimal successful outcome is a reproducible Phase 1 active learning prototype. Phase 2 should be implemented as a decision layer over missing molecule-assay measurements. Multi-task learning is optional and should not obscure the central educational objective: using active learning to choose experiments that improve models and reduce low-value experimental work.