Student Project Proposal

OTO-QSTR: OECD-compliant quantitative structure-toxicity relationship (QSTR) modeling for predicting drug-induced ototoxicity (OTO)

Master area	AI for Drug Discovery; QSAR; Computational Toxicology; Cheminformatics; AI/ML; Drug Safety Assessment
Project type	Computational cheminformatics and machine learning project using curated ototoxicity datasets and OECD-compliant QSTR methodologies
Duration	2-3 months
Supervision	One tutor or co-tutor; weekly or bi-weekly progress meetings
Recommended level	Python programming, basic machine learning, statistics, and introductory cheminformatics. Familiarity with molecular descriptors and molecular fingerprints is advantageous.
Main output	OECD-compliant QSAR/QSTR model, reproducible computational workflow, technical report, and final presentation.
Possible data	Curated ototoxic and non-ototoxic compounds obtained from published literature, regulatory reports, DrugBank, ChEMBL, PubChem, FDA adverse event resources, or tutor-provided datasets.

Short description. Drug-induced ototoxicity is a serious adverse effect that can lead to irreversible hearing loss and vestibular dysfunction. Early identification of compounds with ototoxic potential is essential for reducing drug-development attrition and improving patient safety.

This project aims to develop an Organisation for Economic Co-operation and Development (OECD)-compliant Quantitative Structure–Toxicity Relationship (QSTR) model capable of predicting ototoxicity directly from molecular structure. The student will curate a high-quality dataset, calculate molecular descriptors and fingerprints, build predictive machine learning models, validate them according to OECD principles, and identify structural features associated with ototoxicity risk.

1. Background and rationale

Ototoxicity remains one of the most challenging adverse drug reactions encountered during clinical therapy. Several classes of drugs, including aminoglycosides, platinum-based chemotherapeutics, loop diuretics, and antimalarial agents, have been associated with auditory damage.

Experimental toxicological evaluation is expensive, labor-intensive, and often unavailable during the early stages of drug discovery. Computational approaches such as QSTR modeling provide a rapid and cost-effective strategy for identifying toxicity liabilities before synthesis and biological testing.

To ensure scientific reliability and potential regulatory relevance, modern QSTR models should follow the OECD principles for QSAR/QSTR validation. These principles promote transparency, reproducibility, interpretability, and predictive reliability.

This project directly supports AI-driven drug discovery by integrating cheminformatics, machine learning, and predictive toxicology.

2. Project objective

The objective of this project is to develop a reproducible and interpretable OECD-compliant QSAR/QSTR model for predicting drug-induced ototoxicity from chemical structure. The project aims to identify molecular features associated with ototoxicity and evaluate the predictive performance, applicability domain, and mechanistic relevance of the resulting models.

3. Specific aims

- Curate a reliable ototoxicity dataset.
- Standardize chemical structures.
- Calculate molecular descriptors and fingerprints.
- Explore chemical diversity and descriptor distributions.
- Perform feature selection.

- Build OECD-compliant QSTR models.
- Define the applicability domain.
- Perform internal and external validation.
- Conduct Y-randomization testing.
- Identify molecular determinants of ototoxicity.
- Generate interpretable toxicity rules and structural alerts.

4. Proposed methodology

5.1 Data curation

- Collection of compounds with experimentally supported ototoxicity annotations.
- Structure standardization.
- Salt removal and duplicate elimination.
- Endpoint harmonization.

5.2 Molecular representation

Descriptors:

- Constitutional descriptors
- Physicochemical descriptors
- Topological descriptors
- Electronic descriptors

Fingerprints:

- MACCS
- Morgan fingerprints (ECFP4 and ECFP6)

5.3 Chemical space analysis

Methods:

- PCA
- UMAP
- Similarity mapping
- Descriptor distribution analysis

5.4 Feature selection

Methods:

- Variance filtering
- Correlation filtering
- Mutual information
- Recursive Feature Elimination (RFE)
- Random Forest importance

5.5 Model development

Algorithms:

- Random Forest
- Support Vector Machine
- Gradient Boosting
- XGBoost (optional)
- Logistic Regression

Hyperparameter optimization:

- Grid Search
- Optuna

5.6 Validation and robustness

Validation procedures:

- Stratified k-fold cross-validation
- External test set validation
- Y-randomization
- Applicability domain assessment

5.7 Mechanistic interpretation

Interpretability methods:

- SHAP analysis
- Feature importance ranking
- Structural alert identification
- Descriptor contribution analysis

5. Indicative work plan

Period	Main task	Expected output
Weeks 1-2	Data collection and curation	Curated ototoxicity dataset
Weeks 3-4	Descriptor calculation and exploration	Descriptor matrix and chemical-space analysis
Weeks 5-6	Feature selection	Reduced descriptor set
Weeks 7-8	Model development	Initial QSAR/QSTR models
Weeks 9-10	Validation and applicability domain analysis	Robustness assessment
Weeks 11-12	Interpretation and reporting	Final model, technical report and presentation.

6. Expected deliverables

- Curated ototoxicity dataset.
- Descriptor and fingerprint database.
- OECD-compliant QSTR workflow.
- Predictive machine-learning models.
- Applicability domain assessment.
- Y-randomization analysis.
- Mechanistic interpretation of toxicity.
- A final technical report and a 10–15-minute presentation.

7. Skills acquired by the student

- Chemical data curation.
- Molecular descriptor generation.
- QSTR model development.
- OECD model validation principles.
- Applicability domain analysis.
- Machine learning in toxicology.
- Model interpretation and explainable AI.
- Drug-safety assessment.

8. Suggested software and resources

The project can be implemented using:

- RDKit
- Mordred
- Pandas
- NumPy
- Scikit-learn
- Optuna
- SHAP
- Matplotlib
- UMAP
- XGBoost (optional)

9. Optional extensions

Possible extensions include:

- Multi-class toxicity classification (low, moderate, high ototoxicity).
- Deep learning approaches using graph neural networks.
- Consensus modeling and model ensembling.
- Comparison between descriptor-based and fingerprint-based models.
- Integration with other toxicity endpoints.

10. Evaluation criteria

Criterion	What should be assessed
Scientific clarity	The toxicity problem and modeling strategy are clearly defined.
OECD compliance	Adherence to all five OECD principles
Technical implementation	Reproducibility and documentation
AI relevance	Models effectively use machine learning for toxicity prediction.
Interpretability	Mechanistic understanding of toxicity
Critical discussion	The final report discusses limits, uncertainty and possible future extensions.

Note for tutors. The minimal successful project should include a curated binary ototoxicity dataset, descriptor generation, one fingerprint representation, at least two machine-learning algorithms, applicability-domain analysis, and Y-randomization validation. External validation and mechanistic interpretation are strongly encouraged.