**Research Paper 1**

Predicting drug-target interactions  
using machine learning with  
improved data balancing and  
feature engineering

**Problem Solved:**  
The study addressed the challenges in predicting Drug-Target Interactions (DTIs), including data imbalance, the complexity of biochemical representations, and the limitations of traditional drug discovery methods. The proposed hybrid framework aimed to improve prediction accuracy and scalability in computational drug discovery.

**Dataset Used:**  
The researchers utilized datasets from BindingDB, specifically:

* **BindingDB-Kd**: 52,284 DTI pairs, 10,665 unique drugs, 1,413 unique proteins.
* **BindingDB-Ki**: 375,032 DTI pairs, 174,662 unique drugs, 3,070 unique proteins.
* **BindingDB-IC50**: 991,486 DTI pairs, 549,205 unique drugs, 5,078 unique proteins.

**Methods/Models Applied:**

1. **Feature Engineering**:
   * **Drug Features**: Extracted using MACCS keys (structural fingerprints).
   * **Target Features**: Represented using Amino Acid Composition (AAC) and Dipeptide Composition (DC).
2. **Data Balancing**:
   * **Generative Adversarial Networks (GANs)**: Generated synthetic data for the minority class to address imbalance.
3. **Machine Learning Models**:
   * **Random Forest Classifier (RFC)**: Primary model for DTI prediction.
   * **Comparison Models**: Decision Tree (DTC), Multilayer Perceptron (MLP), Fully Connected Neural Network (FCNN), Multi-Head Attention FCNN (MHA-FCNN), and state-of-the-art models like DeepLPI, BarlowDTI, and Komet.

**Results Achieved:**  
The GAN+RFC hybrid model achieved outstanding performance across all datasets:

* **BindingDB-Kd**: Accuracy = 97.46%, ROC-AUC = 99.42%.
* **BindingDB-Ki**: Accuracy = 91.69%, ROC-AUC = 97.32%.
* **BindingDB-IC50**: Accuracy = 95.40%, ROC-AUC = 98.97%.  
  The model outperformed existing methods in sensitivity, specificity, and error metrics (e.g., MAE, RMSE).

**Gaps/Limitations:**

1. **Model Scope**:
   * Did not incorporate transformer-based deep learning models, which could capture long-range dependencies more effectively.
2. **Feature Fusion**:
   * Limited use of advanced techniques to combine diverse data representations (e.g., graph-based features).
3. **Data Scarcity**:
   * Few-shot learning methods were not explored to handle scenarios with limited labeled data.
4. **Statistical Significance**:
   * The Friedman test indicated no statistically significant differences between models (p-value = 0.4289), suggesting further refinement may be needed.

**Future Work:**  
The authors proposed integrating transformer-based models, advanced feature fusion techniques, and few-shot learning to enhance robustness and generalizability in future studies.

**Research Paper 2**

Predicting Drug-Target Interaction with Machine Learning

**1. Problem Solved**

The thesis addresses the challenge of **predicting drug-target interactions (DTIs)**—a critical but time-consuming and expensive step in drug discovery. Traditional experimental methods (e.g., lab testing) are costly and slow, while computational approaches often suffer from low accuracy or high false positives. Machine learning (ML) offers a faster, more scalable solution to identify potential drug-protein interactions, accelerating drug development.

**2. Datasets Used**

Several datasets were referenced across studies:

* **Gold Standard Dataset (Yamanishi et al., 2008)**:
  + Used in random forest studies (Shi et al., 2019; Chen et al., 2025).
  + Contains drug-protein pairs categorized into enzyme, ion channel, GPCR, and nuclear receptor classes.
* **DrugBank & KEGG**:
  + Used in kernel-based studies (Kuang et al., 2017) for similarity matrices.
* **FDA-approved drug datasets**:
  + Applied in deep learning studies (Wen et al., 2017; Lee et al., 2019) to predict interactions without protein-class restrictions.

**3. Methods/Models Applied**

Three ML approaches were compared:

**A. Random Forest (RF)**

* **Key Studies**: Shi et al. (2019), Chen et al. (2025).
* **Preprocessing**:
  + **Drug representation**: Molecular fingerprints (FP2, PaDEL).
  + **Protein representation**: Pseudo-PSSM (PsePSSM) or sequence descriptors.
  + **Dimensionality reduction**: Lasso (Shi et al.) vs. Random Projection (Chen et al.).
  + **Class imbalance**: SMOTE (oversampling) vs. NearMiss (undersampling).
* **Model**: RF classifier with 5-/10-fold cross-validation.

**B. Kernel Methods**

* **Key Studies**: Cichonska et al. (2017), Kuang et al. (2017).
* **Approach**:
  + **KronRLS**: Kernel-based regression to predict binding affinities (Cichonska et al.).
  + **KMDR**: Kernel matrix dimensionality reduction for classification (Kuang et al.).
* **Input**: Drug-drug and protein-protein similarity matrices (Kronecker product/sum).

**C. Deep Learning (DL)**

* **Key Studies**: Lee et al. (2019), Wen et al. (2017).
* **Models**:
  + **DeepConv-DTI (CNN)**: Processes raw protein sequences to extract residue patterns (Lee et al.).
  + **DeepDTI (Deep Belief Network)**: Uses RBMs for unsupervised feature learning (Wen et al.).
* **Input**: Drug fingerprints + protein descriptors (Wen et al.) or raw sequences (Lee et al.).

**4. Results Achieved**

* **Random Forest**:
  + High accuracy (>90%) and AUPR scores across protein classes (Shi et al., Chen et al.).
  + Outperformed decision trees, Naive Bayes, and logistic regression.
* **Kernel Methods**:
  + KronRLS achieved **0.77 correlation** with experimental binding data (Cichonska et al.).
  + KMDR surpassed RLS and SLP in AUC/AUPR metrics (Kuang et al.).
* **Deep Learning**:
  + **DeepConv-DTI outperformed DeepDTI** and other DL models (Lee et al.).
  + DeepDTI surpassed RF in true positive rate, accuracy, and AUC (Wen et al.).

**5. Gaps and Limitations**

* **Data Limitations**:
  + Performance drops for **novel drugs with unknown targets** (e.g., Cichonska et al.’s correlation fell to 0.66 for poorly characterized drugs).
  + Datasets are often **imbalanced** (far more non-interacting pairs), requiring synthetic sampling (SMOTE/NearMiss).
* **Model Limitations**:
  + **RF**: Less effective for complex feature extraction compared to DL.
  + **Kernel Methods**: Depend heavily on similarity matrices, which may not generalize.
  + **DL**: Computationally expensive; requires GPUs and large datasets.
* **Practical Challenges**:
  + **Ethical/regulatory barriers**: Liability concerns, lack of industry awareness (Wubineh et al., 2023).
  + **Not a standalone solution**: ML cannot fully replace wet-lab validation (especially for novel targets).

**Conclusion**

The thesis demonstrates ML’s potential to **streamline DTI prediction**, with deep learning showing the most promise for future applications. However, gaps in **generalizability, interpretability, and integration** into real-world drug pipelines remain. Future work should focus on:

1. **Hybrid approaches** (ML + experimental validation).
2. **Explainable AI** to build trust in pharmaceutical applications.
3. **Larger, diverse datasets** to improve predictions for novel drugs.

**2. Your Understanding + Project Direction**

**Summary of Key Takeaways from the Papers**

1. **Problem Addressed**: Both papers focus on predicting **Drug-Target Interactions (DTIs)** using machine learning to overcome the limitations of traditional experimental methods (high cost, time consumption) and computational approaches (low accuracy, data imbalance).
2. **Datasets**: Different datasets were used (BindingDB, Gold Standard, DrugBank, KEGG), highlighting the need for standardized benchmarks.
3. **Methods**:
   * **Paper 1** used **GANs for data balancing** and **Random Forest (RFC)** with feature engineering (MACCS keys, AAC, DC).
   * **Paper 2** compared **Random Forest, Kernel Methods, and Deep Learning**, with DL models (e.g., DeepConv-DTI) showing strong performance.
4. **Gaps Identified**:
   * Lack of **transformer-based models** (e.g., BERT for proteins).
   * Limited **feature fusion** (e.g., graph-based drug representations).
   * Challenges with **novel drugs/targets** (poor generalization).
   * Need for **explainability** and **hybrid ML + experimental validation**.

**My Project Plan**

**Objective**: Develop an **improved DTI prediction model** by integrating **transformer-based architectures** (for protein sequences) and **graph neural networks (GNNs)** (for drug molecules) while addressing data imbalance and interpretability.

**What I Will Do Differently**

1. **Model Architecture**:
   * Use **ProtBERT** (protein sequence embeddings) + **GNNs** (for drug molecular graphs) instead of traditional fingerprints/PsePSSM.
   * Implement **attention mechanisms** to improve feature fusion.
2. **Data Handling**:
   * Apply **contrastive learning** for few-shot scenarios (better generalization for novel drugs).
   * Use **SMOTE + GANs** for robust data balancing.
3. **Explainability**:
   * Integrate **SHAP/LIME** to interpret predictions (critical for pharmaceutical trust).
4. **Hybrid Validation**:
   * Combine predictions with **docking simulations** (e.g., AutoDock Vina) for higher confidence.

**Improvements Over Existing Solutions**

| **Existing Approach** | **Limitation** | **My Improvement** |
| --- | --- | --- |
| RFC + GANs (Paper 1) | No transformers/GNNs | **ProtBERT + GNNs** for better feature learning |
| Kernel Methods (Paper 2) | Relies on similarity matrices | **End-to-end deep learning** (no manual feature engineering) |
| DeepConv-DTI (Paper 2) | Limited interpretability | **Explainable AI (SHAP)** for model transparency |
| Standard SMOTE | Synthetic samples may not be realistic | **GANs + contrastive learning** for better minority-class generation |

**Why My Version Should Be Selected**

**Higher Accuracy**: Transformer + GNN fusion captures complex interactions better than RFC or CNNs.  
 **Better Generalization**: Contrastive learning helps with **novel drug-target predictions**.  
**Interpretability**: SHAP/LIME explanations make the model **trustworthy for real-world use**.  
**Practical Integration**: Hybrid ML + docking validation bridges the gap between computation and wet-lab testing.

### References

**1.**

**Title:** *Predicting drug-target interactions using machine learning with improved data balancing and feature engineering*  
**Authors:**

* Md. Alamin Talukder (International University of Business Agriculture and Technology, Dhaka, Bangladesh)
* Mohsin Kazi (King Saud University, Riyadh, Saudi Arabia)
* Ammar Alazab (Victoria University & Torrens University Australia)

**Source/Publisher:** *Scientific Reports* (Nature Portfolio)  
**DOI:** <https://doi.org/10.1038/s41598-025-03932-6>  
**Download/View Link:**

* [Nature Scientific Reports](https://www.nature.com/articles/s41598-025-03932-6)

**Dataset Source**

**Dataset:** BindingDB (for DTI prediction)  
**Source:** Therapeutics Data Commons (TDC)  
**Link:** <https://tdcommons.ai/multi_pred_tasks/dti/#bindingdb>

**2.**

1. **Title**: Predicting Drug-Target Interactions Using Machine Learning  
   **Authors**:  
   • Lilisa Särkiö  
   **Source/Publisher**: Bachelor's Thesis, Bioinformation Technology Programme  
   **Download/View Link**: [Predicting Drug-Target Interaction with Machine Learning](https://aaltodoc.aalto.fi/bitstreams/5464ed75-250e-48be-b110-c1a3fe854f89/download)

**Dataset Source**  
**1.Dataset**: Gold Standard Dataset for Drug-Target Interaction Prediction  
**Authors**: Yamanishi et al.  
**Source**: Bioinformatics  
**DOI**: <https://doi.org/10.1093/bioinformatics/btn162>  
**Download/View Link**:  
• Oxford Academic Bioinformatics

**2.Dataset**: DrugBank  
**Source**: University of Alberta  
**Link**: [https://go.drugbank.com](https://go.drugbank.com/)

**3.Dataset Source**  
**Dataset**: KEGG DRUG  
**Source**: Kanehisa Laboratories  
**Link**: <https://www.genome.jp/kegg/drug/>