Title: Advancing AI-Driven Drug Discovery: Solving Data and Biology Challenges for Pharmedics

Prepared in Reflection to: Prof. Andreas Bender's Talk, Oncode Accelerator Summit '24

Submitted by: Alagappan Alagappan

Date: 1-07-2025

Introduction

At the Oncode Accelerator Summit '24, Prof. Andreas Bender posed a critical question: "How do we bring the right drug to the right patient in the right way?" This challenge drives Pharmedics' mission to reduce error-costs and save lives through AI-driven drug discovery. However, barriers like scarce high-quality labeled data, complex biological (omics) data, and patient variability hinder progress. This report proposes sophisticated yet practical solutions to these challenges, blending biological insight with advanced AI to advance precision medicine. Written for Pharmedic's Product Technology Department, it balances technical depth with intuitive explanations to ensure clarity for all stakeholders.

Mitigating Key Challenges

1. Overcoming Scarce Labeled Data

High-quality labeled data, such as known drug side effects, is hard to obtain due to costly experiments and ethical constraints. Imagine trying to learn a new language with only a few example sentences — AI faces a similar problem in drug discovery. To address this:

- **Transfer Learning:** Think of this as using a pre-learned dictionary. We start with AI models trained on large public datasets like ChEMBL (chemical structures) or LINCS (gene responses), then fine-tune them with smaller, specific datasets for side effects. For example, a model pre-trained to predict drug mechanisms of action on LINCS data can be fine-tuned for ADR prediction, reducing the need for labeled examples.
- **Active Learning:** Instead of labeling every sample, the AI picks the most confusing ones for experts to label, like a student asking questions only on tough topics. This maximizes the value of limited expert time by prioritizing samples where model uncertainty is highest.
- **Synthetic Data with VAEs:** Variational autoencoders (VAEs) act like artists creating realistic synthetic patient gene profiles. These generative models learn

the distribution of real omics data, then simulate new data points that are biologically plausible. Synthetic samples are validated through similarity in the latent space (e.g., t-SNE or cosine similarity) to ensure accuracy.

Why This Works: Transfer learning borrows knowledge from existing data, active learning focuses labeling efforts, and VAEs expand datasets ethically. To ensure trust, synthetic data is validated by biologists to match real-world patterns, aligning with Pharmedics' goal of safe, reliable solutions.

2. Simplifying Complex Omics Data

Omics data (e.g., gene expression, protein interactions) is like a massive, interconnected web of biological signals, complicated by noise and scale. To make it manageable:

- **Graph Neural Networks (GNNs):** Picture a city map where genes and proteins are buildings, and their interactions are roads. GNNs learn how signals travel through this map, capturing relationships like how one gene affects another. These models integrate knowledge from biological graphs such as Reactome or STRING and predict how drugs might disrupt these systems.
- **Autoencoders for Noise Reduction:** These act like noise-canceling headphones, compressing omics data to keep only the important biological signals, like key gene patterns, while filtering out batch effects and experimental noise.
- **Transformers for Time-Series:** Drugs cause changes in cells over time, like a story unfolding. Transformers use self-attention to track how gene expression evolves over time, modeling long-range dependencies more effectively than traditional RNNs.

Why This Works: GNNs respect the interconnected nature of biology, autoencoders clean up messy data, and transformers capture time-based changes, making models more accurate and aligned with real biological processes.

3. Handling Patient Variability

Every patient is unique due to genetics, lifestyle, or other medications, like how no two people have the same fingerprint. To account for this:

- **Clustering for Similarity:** Group patients with similar biological traits (e.g., similar gene mutations) using clustering algorithms like k-means. This is like sorting people by clothing size before tailoring, ensuring models fit specific groups.

- **Meta-Learning (MAML):** This teaches AI to adapt quickly to new patients, like a chef learning to cook in any kitchen. Model-Agnostic Meta-Learning (MAML) trains models that can quickly personalize individual patient data with minimal retraining.
- **Ensemble Models:** Combine a GNN for omics data with a tree-based model (e.g., XGBoost) for clinical data (e.g., age, dosage). This is like consulting two experts — one for molecular details, one for patient history — for a complete picture.

Why This Works: Clustering simplifies variability by grouping similar patients, MAML enables personalized predictions, and ensembles integrate diverse data, ensuring tailored treatments that support Pharmedics' mission of precision medicine.

Building Robust AI Models

1. Model Choices

To handle omics data and predict side effects (adverse drug reactions, ADRs), we choose models that capture complex patterns:

- **Graph Neural Networks (GNNs):** Ideal for modeling drug-protein or gene-gene interactions, like mapping a social network. GNNs learn structural relationships, making them effective for modeling polypharmacy side effects and drug impact pathways [1].
- **Transformers:** These are great for sequence data, like tracking gene changes over time, ensuring we capture dynamic drug effects.
- **XGBoost:** A tree-based model that handles clinical data well, robust to missing information, and works for diverse patient groups.

2. Implementation Steps

To build these models:

1. **Data Preparation:** Clean omics data (e.g., log-transformation with Scanpy) and encode drugs as molecular graphs (using RDKit). Normalize clinical data to ensure consistency.
2. **Training:** Use PyTorch Geometric for GNNs and Hugging Face for transformers. Apply transfer learning from datasets like LINCS. Use federated learning to train across hospitals without sharing private data, ensuring ethical compliance.
3. **Evaluation:** Measure accuracy with AUROC for ADR predictions and Precision@K for top drug recommendations. Use SHAP to explain why the AI made a prediction, building clinician trust.

4. **Clinical Integration:** Deliver predictions via a REST API to electronic health records (EHRs), providing real-time ADR risk scores. Estimate prediction uncertainty to guide doctors on reliability.

3. Real-World Example

In oncology, GNNs predicted ADRs for cancer drugs by modeling drug-protein interactions, achieving an AUROC of 0.85 [1]. By clustering patients based on tumor gene profiles, the model improved predictions for immunotherapy side effects, showing how these methods deliver precise, life-saving insights.

Ethical and Clinical Impact

To align with Pharmedics' mission, we prioritize ethical AI. Federated learning protects patient privacy by training models without sharing sensitive data. SHAP explanations ensure doctors understand AI decisions, fostering trust. By reducing ADR risks through precise predictions, these solutions minimize errors and save lives, directly supporting Pharmedics' vision.

Conclusion

Scarce labeled data, complex omics, and patient variability are tough challenges, but solvable with smart AI. Transfer learning, GNNs, and patient clustering make models accurate and personalized. By integrating these into clinical workflows with ethical safeguards, we ensure AI doesn't just compute — it saves lives. This approach positions Pharmedics at the forefront of precision medicine.

References

[1] Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13), i457–i466. https://doi.org/10.1093/bioinformatics/bty294