In Silico Modeling of Drug Safety and Drug Release: Machine Learning Prediction of Kinase Inhibitor Adverse Events and Deconvolution Analysis of In Vivo Release from Modified Release Formulations

by

Natalie Jusko

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Pharmaceutical Sciences)
in the University of Michigan
2026

Doctoral Committee:

Professor Duxin Sun, Chair
Assistant Professor Brandon Bordeau
Professor Mingyan Liu
Professor Kerby Shedden
Professor Peter Tessier

Natalie Jusko

nmjusko@umich.edu

ORCID iD: 0009-0004-6233-816X

© Natalie Jusko 2026

Dedication

Я на гору круту крем'яную
Буду камінь важкий підіймать
І, несучи вагу ту страшную,
Буду пісню веселу співать.

Up the steep and flinty mountain
I will heave the heavy stone
and bearing that fearsome weight,
I will sing a merry song.

— Lesya Ukrainka, *Contra spem spero!*

For my parents, on whose shoulders I climb.

Acknowledgements

I have been very fortunate over the course of my education that a great number of people have seen fit to invest their time, effort, and knowledge in me.

First and foremost, I would like to extend my deepest gratitude to my advisor, Dr. Duxin Sun, for his advice, scientific input, financial support, and continuous encouragement. His willingness to go above and beyond for his students, depth of knowledge across disciplines, boundless work ethic, and tactful leadership skills carried me through my graduate studies and set the example for my career.

I would also like to thank my dissertation committee, Dr. Brandon Bordeau, Dr. Mingyan Liu, Dr. Kerby Shedden, and Dr. Peter Tessier, for their time and effort providing valuable feedback on my research, and for pushing me to see the forest for the trees.

A significant portion of graduate work is administrative, and I would have been lost long ago if Erika Zucal was not very good at what she does. Thank you to Erika for having all the answers, and to Hanning Wen and the rest of the Sun lab. Additionally, research reported in this dissertation was partially supported by the Center for Data-Driven Drug Development and Treatment Assessment (DATA), an industry-university cooperative research center partially supported by the U.S. National Science Foundation under the award number 2209546, and by DATA industry partners.

Next, I would like to acknowledge my co-workers and mentors at the Ann Arbor Pharmacometrics Group, who spurred my growth as a scientist and imparted their passion for deriving answers from complex modeling. Thank you to Dr. Emily (Briggs) Schapiro, Dr. Ryan Crass, Dr. Nancy Dolphin, Dr. Howard Bockbrader, Dr. Won Byon, Dr. Nidal al-Huniti, and particularly to Dr. Sunny Chapel, who first encouraged me to go to Michigan, often ensured I was eating properly, and for believing in me as someone worth investing in right out of my undergraduate studies.

I was also fortunate enough to cross paths with Dr. David Smith and Dr. Mohamed Kamal during my time at Michigan and the Ann Arbor Pharmacometrics Group. Their scientific guidance, life advice, generosity, and kindness will not be forgotten.

I would also like extend my gratitude to my mentors at the FDA, whose research acumen, depth of regulatory knowledge, and warmth made my time as an ORISE fellow especially meaningful. Thank you to Dr. Yuqing Gong, Dr. Andrew Babiskin, Dr. Meng Hu, Dr. Lucy Fang, and Dr. Liang Zhou.

Additionally, I would like to thank the clinical team on the glipizide project – to Merna Sitto, Patricia Londy, and Cathrin Ring for months of sleepless nights in the Michigan Clinical Research Unit, to Dr. Borko Nojkov, Dr. Graham Brady, and Dr. Jason Baker for their medical and scientific expertise, and particularly to Dr. Amit Pai, for his wealth of knowledge and encouragement.

Next, I would like to thank the friends I have made over the past decade who have filled my days with sunshine, laughter, and much needed time away from studying. Thank you to John McClay and Alex Meyer, whose Midwestern charm and willingness to go out on a Friday night carried me through the more stressful moments of graduate school. To Dr. Rose Dimitroyannis, Baunnee Martinez, and Anoushka Patel – who I met during my very first week at the University of Chicago and then spent nearly every single day with thereafter – I feel fortunate to have forged relationships as meaningful and enduring as ours, and it's been a privilege to witness your lives unfold and accomplishments accumulate since we parted ways at 5213 Dorchester Ave.

Thank you to my sister, Nicole Jusko, who knows me, can make me laugh, and can read my mind like no one else. To my mom, Małgorzata Jusko – for my entire life, you have removed all obstacles in my path and put me in positions to succeed. Your unwavering belief in the value of a prestigious education, your formidable attention to detail, and your persistent "are you alive ??" messages make me who I am today. To my family back in Poland, in particular my babci, Maria Sidor, and my wujka, Jarosław Andruch – I carry your love and pride with me in this milestone and wherever I go.

To my great love and husband of 6 years, Assistant Professor Josh Foster, I am endlessly thankful. You have supported all my goals without hesitation, endured many long flights and drives so that we might be in the same place, and built a life in Madison I cannot wait to share

with you. Without your many reminders that a PhD actually isn't very hard at all, I might not have made it.

And finally, for my dad, Dr. William Jusko, I struggle to find the words. Above all else, you told me that if I found something I loved, I would never work a day in my life. My search for this has been nonlinear, but by your infinite love, support, and example, I believe I have found it.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables xii

List of Figures xiv

Abstract xiv

Chapter 1: AI/ML for Predicting Preclinical and Clinical Drug Toxicities 1

1.1 Abstract 1

1.2 Introduction 1

1.3 A Brief Overview of AI/ML Methods for Predicting Drug Toxicity 3

1.3.1 Methodological considerations for AI/ML-based toxicity prediction 4

1.3.2 Transforming drug structures into AI/ML model inputs and features 9

1.3.3 Performance metrics in AI/ML for drug toxicity prediction 11

1.4 AI/ML Models for Predicting *In Vitro* and Preclinical Drug Toxicities 15

1.4.1 AI/ML models to predict hERG channel blockage for cardiotoxicity assessments 15

1.4.2 AI/ML models to predict drug-induced liver injury on the basis of clinical reporting 20

1.4.3 AI/ML algorithms to predict acute toxicity in animal models 23

1.4.4 AI/ML models to predict genotoxicity in vitro 25

1.4.5 AI/ML models to predict multi-endpoint toxicity in vitro, in animal models, and in humans 28

1.4.6 AI/ML-based prediction of drug toxicity via chemical-target interactions 32

1.4.7 AI/ML-based prediction of protein and peptide toxicity using in vitro and animal systems 34

1.5 AI/ML Models for Predicting Clinical Drug Toxicity 37

1.5.1 Drug-drug interactions 37

1.5.2 Patient-level adverse events for individual therapeutic agents 41

1.5.3 Pharmacogenomic predictors of drug-induced toxicity 45

1.5.4 Adverse event prediction across pharmacologically related drugs 47

1.5.5 Adverse event prediction from chemical structure and systems pharmacology 49

1.6 Challenges and Emerging Directions in AI/ML-based Drug Toxicity Prediction 53

1.6.1 Interpretability 53

1.6.2 Generalizability and benchmarking 55

1.6.3 Emerging architectures: large language models and multi-modal foundation models 56

1.7 Conclusions and Future Perspectives 58

1.8 Research Aims 62

1.9 References 62

Chapter 2: Machine Learning Predicts Clinical Adverse Events of Small Molecule Kinase Inhibitors in Cancer Patients Using On-/Off-Target Engagement and Tissue Selectivity 83

2.1 Abstract 83

2.2 Introduction 84

2.3 Methods 85

2.3.1 Data collection 85

2.3.2 Data preprocessing 89

2.3.3 Statistical framework and model development 91

2.3.4 Statistical validation, inference, and variable importance 94

2.3.5 Statistical assumptions and limitations 97

2.4 Results 98

2.4.1 Linking on-/off-target engagement and tissue selectivity to clinical adverse events for small molecule kinase inhibitors 98

2.4.2 Machine learning models incorporating on-/off-target engagement and tissue selectivity improve tissue-specific adverse event predictions 101

2.4.3 Variable importance analysis reveals target-adverse event associations across organs 102

2.4.4 External validation using neratinib confirms predictive performance of final ML model 106

2.5 Discussion 108

2.6 References 111

Chapter 3: Multi-Task Neural Network Predicts Clinical Adverse Event Incidence Across a Multi-Mechanism Cohort Using a Unified STAR Feature Space 117

3.1 Abstract 117

3.2 Introduction 118

3.3 Methods 120

3.3.1 Cohort definition and inclusion criteria 120

3.3.2 STAR feature assembly 120

3.3.3 Adverse event outcomes 121

3.3.4 Multi-task neural network architecture 122

3.3.5 Training and validation protocol 123

3.3.6 Statistical metrics and baselines 124

3.3.7 Software and implementation 125

3.4 Results 126

3.4.1 Cohort and STAR feature composition 126

3.4.2 The multi-task neural network recovers drug-specific predictive signal beyond AE base rates 127

3.4.3 Per-drug predictions trace to known target-AE mechanisms 131

3.4.4 Block-level feature ablations: no single block is independently load-bearing 132
3.4.5 Generalization across mechanism classes 133
3.4.6 Negative controls and hyperparameter sensitivity 134
3.4.7 Magnitude calibration of predicted AE incidence 136
3.4.8 Sample-size scaling and the case for cohort expansion 137
3.5 Discussion 139
3.6 References 141
Chapter 4: Perspectives on STAR-Guided AI/ML Prediction of Clinical Drug Toxicity and Future Directions 145
4.1 Summary 145
4.1.1 The translational gap and the limits of existing computational approaches 145
4.1.2 Integrating pharmacological context into AI/ML toxicity prediction 146
4.2 Implications 148
4.2.1 STAR as a triage tool for pre-first-in-human safety prediction 148
4.2.2 Mechanistic interpretability as a candidate-design feedback tool 149
4.2.3 Regulatory alignment: STAR within the FDA's New Approach Methodologies framework 150
4.3 Limitations 152
4.3.1 Feature coverage as the framework's empirical bottleneck 152
4.3.2 Magnitude calibration as a constraint of label quality rather than model capacity 153
4.3.3 Prospective generalization as the open empirical question 154
4.4 Future Directions 155
4.4.1 Empirical extensions: expanding the cohort beyond oncology 155
4.4.2 Architectural extensions: STAR-specific inductive biases 156
4.5 Conclusions 158
4.6 References 159

Chapter 5: Direct Measurement of *In Vivo* Dissolution and Release of Modified-Release Glipizide Formulations 163

5.1 Abstract 163

5.2 Introduction 164

5.3 Methods 168

5.3.1 Study design overview 168

5.3.2 Subjects, enrollment, and primary analysis population 169

5.3.3 Study products and dosing 170

5.3.4 Gastrointestinal intubation and sample collection 171

5.3.5 Bioanalytical methods 173

5.3.6 Continuous glucose monitoring 174

5.3.7 Non-compartmental analysis 175

5.3.8 Population pharmacokinetic modeling 176

5.3.9 Deconvolution of modified-release absorption 177

5.3.10 Software and reproducibility 178

5.4 Results 179

5.4.1 Subject disposition and demographics 179

5.4.2 Plasma pharmacokinetics 180

5.4.3 Regional gastrointestinal disposition 182

5.4.4 Luminal pH and bile acid environment 185

5.4.5 Urinary and fecal recovery 191

5.4.6 Continuous glucose monitoring 192

5.4.7 Paired intubation (Phase 1) versus non-intubation (Phase 2) comparison 201

5.4.8 Population pharmacokinetic modeling and deconvolution of modified-release absorption 203

5.5 Discussion 210

5.5.1 Principal findings 210
5.5.2 Release mechanism and regional disposition 211
5.5.3 Luminal environment: pH and bile acid composition 213
5.5.4 Systemic pharmacokinetics, population modeling, and deconvolution 214
5.5.5 Implications and future directions 216
5.6 Conclusions 217
5.7 References 218

List of Tables

Table 1-1. Overview of AI/ML methods used in drug toxicity prediction, with advantages and limitations for safety-critical applications. 5

Table 1-2. Performance metrics for AI/ML-based toxicity prediction, with guidance on suitability for imbalanced datasets and safety-critical decision contexts. 12

Table 2-1. Description of SMKIs used in AE model training and validation. 86

Table 2-2. Most common treatment-emergent AEs experienced by patients on small molecule kinase inhibitors. 88

Table 2-3. Hyperparameter final settings for RSF AE models. 92

Table 2-4. AE models constructed in this analysis and the organ system or tissue type associated with their occurrence. 94

Table 2-5. Predictive accuracy of Model 1 (systemic kinase inhibition) versus Model 2 (systemic plus tissue-weighted kinase inhibition) for the most common tissue-localized AEs . 102

Table 2-6. Predictive accuracy of Model 1 (systemic kinase inhibition) for severe (grade ≥ 3) hematologic AEs 102

Table 2-7. Subset of literature survey results with reported kinase-associated toxicities that match model identified KI-AE pairs. 103

Table 2-8. Predictive accuracy for five of the most frequently reported AEs in the neratinib external validation dataset 108

Table 3-1. Cohort-level performance of the multi-task neural network versus a constant-scalar predictor, the per-AE cohort-mean baseline, and single-task ridge regression. 130

Table 3-2. Cohort-level performance under block-level feature ablations. 133

Table 3-3. Per-class mean corrected per-drug Spearman ρ under leave-one-mechanism-class-out cross-validation across 12 eligible mechanism classes (≥ 2 drugs). 134

Table 3-4. Cohort metrics under two negative controls compared with the unperturbed multi-task neural network. 135

Table 3-5. Cohort metrics at subsampled cohort sizes $N \in \{30, 45, 60, 75, 86\}$ drugs, mean ± standard deviation across 20 LOOCV bootstrap seeds per cohort size. 139

Table 5-1. Demographic and baseline characteristics of the primary analysis population, by modified-release formulation arm. 180

Table 5-2. Non-compartmental pharmacokinetic parameters for plasma modified-release glipizide and for the co-administered $^{13}C_6$-glipizide oral solution reference, summarized by formulation arm. 181

Table 5-3. Population pharmacokinetic parameter estimates for the two-compartment model with first-order absorption fitted to $^{13}C_6$-glipizide oral solution plasma concentration data from the primary analysis population using all oral solution data. 204

List of Figures

Figure 1-1. Overview of AI/ML methods commonly employed in drug toxicity prediction. 9

Figure 1-2. Schematic overview of the DeepHIT framework for prediction of hERG-induced cardiotoxicity. 18

Figure 1-3. The DeepTox pipeline for computational toxicity prediction. 30

Figure 1-4. Architecture of ToxDL2.0 for multi-modal protein toxicity prediction. 36

Figure 1-5. Architecture of MSKG-DDI for drug-drug interaction prediction. 40

Figure 1-6. Multi-modal deep learning pipeline for predicting symptomatic pneumonitis in lung cancer patients undergoing combined thoracic and radiotherapy and immunotherapy. 44

Figure 1-7. Knowledge graph-based prediction of adverse drug reactions. 51

Figure 1-8. The Structure–Tissue/cell selectivity–Activity Relationship (STAR) framework for AI/ML prediction of clinical adverse events. 61

Figure 2-1. Graphic representation of random survival forest (RSF) modeling to predict adverse event (AE) status and time to AE onset of 16 small molecule kinase inhibitors (SMKIs) in 3,433 cancer patients. 93

Figure 2-2. Flow chart depicting the process of clinical report selection for systematic review. 96

Figure 2-3. Adverse events (AEs), on-/off-target binding constants (K_i), and tissue selectivity of 16 SMKIs. 99

Figure 2-4. Principal Component Analysis (PCA) of n = 442 kinase target K_i values, stratified by occurrence or non-occurrence of AEs. 100

Figure 2-5. Principal Component Analysis (PCA) of n = 36 tissue K_p values per drug compound for 16 SMKIs, stratified by occurrence or non-occurrence of AEs. 101

Figure 2-6. Final ML model with on-/off-targets and tissue selectivity of SMKIs identified novel kinase targets associated with fatigue and rash. 105

Figure 2-7. Final ML model with on-/off-targets and tissue selectivity of SMKIs identified novel kinase targets associated with diarrhea and vomiting. 106

Figure 2-8. External validation on independent SMKI dataset demonstrates good model sensitivity and specificity to predict AEs of neratinib. 107

Figure 2-9. Predicted cumulative incidence curves for common adverse events in the neratinib external validation dataset. 108

Figure 3-1. Mechanism-class composition of the modeling cohort, partitioned into 20 oncology and 14 non-oncology classes. 126

Figure 3-2. AE-label coverage of the assembled prediction matrix 127

Figure 3-3. Per-drug corrected (rank-in-column) Spearman ρ across the 80 evaluable cohort drugs (≥ 3 labeled AEs each), sorted ascending by per-drug mean. 129

Figure 3-4. Mechanism case studies for three representative cohort drugs: sunitinib (VEGFR multi-TKI), imatinib (BCR-ABL inhibitor), and erlotinib (EGFR inhibitor). 131

Figure 3-5. Calibration of the multi-task neural network's predicted AE incidence. 137

Figure 3-6. Sample-size scaling of cohort-level model performance. 138

Figure 5-1. Plasma glipizide concentration-time profiles for the modified-release (MR) glipizide tablet (5 mg) and the co-administered $^{13}C_6$-glipizide oral solution reference (1 mg). . 182

Figure 5-2. Luminal modified-release glipizide concentrations in the supernatant fraction of aspirates from the primary analysis population. 183

Figure 5-3. Luminal $^{13}C_6$-glipizide concentrations in the supernatant fraction of aspirates from the 1 mg oral solution reference, pooled across the primary analysis population (both formulation arms combined). 184

Figure 5-4. Regional gastrointestinal luminal pH by region (panels: stomach, duodenum, jejunum), measured ex vivo from each aspirate. 185

Figure 5-5. Intraluminal bile acid concentrations by class and gastrointestinal sampling site. . 187

Figure 5-6. Intraluminal concentrations of the four primary conjugated bile acid species. 188

Figure 5-7. Intraluminal concentrations of the six secondary conjugated bile acid species. 189

Figure 5-8. Intraluminal concentrations of the two primary unconjugated bile acid species. ... 190

Figure 5-9. Intraluminal concentrations of the three secondary unconjugated bile acid species. 190

Figure 5-10. Mean cumulative urinary recovery of glipizide over time, by formulation arm, for both the modified-release product (5 mg dose) and the co-administered $^{13}C_6$-glipizide oral solution reference (1 mg dose). 191

Figure 5-11. Mean cumulative fecal recovery of glipizide over time, by formulation arm, for both the modified-release product (5 mg dose) and the co-administered $^{13}C_6$-glipizide oral solution reference (1 mg dose). 192

Figure 5-12. Continuous glucose monitoring profile for subject G12 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 194

Figure 5-13. Continuous glucose monitoring profile for subject G13 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 194

Figure 5-14. Continuous glucose monitoring profile for subject G14 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 195

Figure 5-15. Continuous glucose monitoring profile for subject G15 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 196

Figure 5-16. Continuous glucose monitoring profile for subject G16 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 196

Figure 5-17. Continuous glucose monitoring profile for subject G17 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 197

Figure 5-18. Continuous glucose monitoring profile for subject G18 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 198

Figure 5-19. Continuous glucose monitoring profile for subject G19 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 199

Figure 5-20. Continuous glucose monitoring profile for subject G20 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). 199

Figure 5-21. Continuous glucose monitoring profile for subject G21 following a single oral dose of modified-release glipizide (Product 1, osmotic pump; 5 mg). 200

Figure 5-22. Continuous glucose monitoring profile for subject G22 following a single oral dose of modified-release glipizide (Product 1, osmotic pump; 5 mg). 201

Figure 5-23. Plasma glipizide concentration-time profiles by analyte and intubation versus no-intubation cohorts, shown on linear (left) and semi-logarithmic (right) scales. 203

Figure 5-24. Structural diagram of the two-compartment population pharmacokinetic model with first-order absorption fitted to the $^{13}C_6$-glipizide oral solution plasma concentration data. 204

Figure 5-25. Goodness-of-fit diagnostics for the final two-compartment population pharmacokinetic model of the $^{13}C_6$-glipizide oral solution using all oral solution data. 206

Figure 5-26. Visual predictive check for the final population pharmacokinetic model, based on 500 simulated replicates of the observed design. 207

Figure 5-27. Pooled geometric mean plasma concentration-time overlay of observed, individual-predicted (IPRED), and population-predicted (PRED) concentrations for the $^{13}C_6$-glipizide oral solution, across the primary analysis population. 207

Figure 5-28. Cumulative fraction of the modified-release glipizide dose absorbed over time, by formulation arm, reconstructed on an individual-subject basis using the Loo-Riegelman method..... 209

Figure 5-29. Observed plasma modified-release glipizide concentrations (filled circles) and fitted Weibull input function predictions across all converged subjects in each formulation arm. 210

Abstract

Drug development frequently relies on indirect or incomplete predictors of human outcomes. This dissertation develops computational and experimental methods that supply missing biological context: target and tissue context for clinical toxicity, and regional gastrointestinal context for modified-release performance. Together they improve the interpretation of clinical safety and exposure behavior. Dose-limiting toxicity and lack of efficacy drive approximately 70% of clinical-stage failures and modified-release (MR) oral products that meet plasma-based bioequivalence criteria can exhibit substantially different regional *in vivo* release that conventional pharmacokinetic assessment cannot resolve. Existing *in silico* approaches to adverse event (AE) prediction integrate neither target engagement nor tissue distribution with patient-level exposure, and biopharmaceutic characterization of MR products rests on indirect inference from dissolution and plasma pharmacokinetics rather than direct regional measurement.

A Structure-Tissue/cell selectivity-Activity-Relationship (STAR) feature framework was developed for clinical AE prediction and applied first to small molecule kinase inhibitors (SMKIs). Patient-level pharmacokinetics, kinome-wide on- and off-target binding affinities across 442 kinases, and tissue-specific drug distribution were integrated within a random survival forest trained on 3,433 patients from registrational trials of 16 FDA-approved SMKIs. Tissue-specific distribution improved prediction of tissue-localized toxicities over plasma exposure-only models, while systemic exposure remained the appropriate input for severe hematologic toxicities. Variable importance analysis recovered kinase-AE associations across organ systems, and external validation on an unseen drug, neratinib, reproduced trial-level safety patterns with high sensitivity and specificity.

To extend the framework beyond kinase inhibitors, a shared-trunk multi-task neural network was trained on a unified 3,170-dimensional STAR feature vector integrating structural descriptors, kinome-wide binding affinities, target-tissue expression, and approved clinical dose across FDA-approved small molecules spanning 20 oncology and 14 non-oncology mechanism

classes. Per-AE incidence rates were extracted from FDA prescribing labels across 79 endpoints. The model recovered a reproducible drug-specific predictive signal beyond per-AE base rates, separable from constant, ridge, and permutation-based baselines, and produced per-drug AE rankings traceable to established target-AE pharmacology for sunitinib, imatinib, and erlotinib. Generalization under leave-one-mechanism-class-out cross-validation was strongest for classes pharmacologically analogous to the training cohort. A rank-within-column corrected per-drug Spearman ρ statistic, robust to asymmetric AE label presence, was introduced as a base-rate-robust discrimination metric.

An *in vivo* biopharmaceutic methodology was developed for oral MR products, combining region-resolved oral multi-lumen gastrointestinal intubation, within-subject $^{13}C_6$-glipizide stable-isotope oral solution co-administration as a same-occasion systemic disposition reference, and population pharmacokinetics-supported Loo-Riegelman deconvolution. The methodology was implemented in healthy volunteers using two commercially available MR glipizide products of contrasting release mechanisms: an osmotic pump and a hydrophilic polymer matrix. Plasma profiles of the two products were closely similar, consistent with their bioequivalence status, while their regional luminal release patterns differed. The osmotic pump showed drug concentrations confined predominantly to the stomach, whereas the matrix product showed drug appearance distributed across stomach, duodenum, and jejunum, with the largest individual concentrations in the small bowel. The directly characterized luminal environment showed low gastric pH, near-neutral small bowel pH, and small bowel bile acid concentrations higher than gastric values but below the FaSSIF-V2 specification. Deconvolution indicated that approximately one-quarter of the MR dose had been absorbed by the end of the seven-hour window.

Together, these studies render mechanism, tissue, and regional context observable inputs to drug safety prediction and biopharmaceutic evaluation of oral modified-release products, supporting regulatory adoption of New Approach Methodologies-aligned alternatives to GLP animal toxicology and compendial dissolution testing.

Chapter 1: AI/ML for Predicting Preclinical and Clinical Drug Toxicities

1.1 Abstract

Clinical toxicity accounts for approximately 70% of drug development failures, as 40% of failures attributed to poor efficacy may often be due to dose-limiting toxicities that prevent adequate dose escalation. Conventional drug toxicity studies from *in vitro* assays and preclinical animal models often fail to predict adverse events (AEs) in humans at clinically relevant doses. To address this gap, the U.S. Food and Drug Administration has incentivized New Approach Methodologies to supplement or replace traditional animal-based toxicity testing. Recent advances in artificial intelligence (AI) and machine learning (ML) have enabled the development of predictive models for drug-induced toxicity *in vitro*, in animal models, and in clinical settings. These models leverage molecular structure to forecast outcomes such as hERG channel inhibition, Ames mutagenicity, hepatotoxicity, and LD_{50} values in animals. AI/ML approaches have also been employed to predict clinical AEs based on drug-specific pharmacogenomic targets or across drug classes in patient populations; other models aim to create AE predictions using the chemical structure and target profiles of hundreds or thousands of drugs. Despite rapid and ongoing progress, current AI/ML frameworks remain inadequate for accurately predicting clinical AEs at therapeutic doses. Future models must incorporate drug structure, on- and off-target interactions, tissue- and cell-specific selectivity, target expressions, and clinical doses to enhance predictive performance.

1.2 Introduction

Navigating the process of drug development and regulatory approval is a lengthy, expensive, and high-risk investment. On average, it takes 10 to 15 years and approximately US $1-2 billion[1] to bring a single drug successfully through phase I to III clinical testing[2,3]. Despite rigorous efforts, including target validation, high-throughput screening of vast libraries of candidate molecules, optimization for biological activity and drug-like properties, and extensive preclinical testing both *in vitro* and *in vivo*, approximately 90% of drug candidates fail after

entering phase I trials[2,3]. While 40-50% of clinical drug development failures are attributed to lack of efficacy and around 30% to toxicity, these issues are often interconnected. In many cases, a drug's lack of efficacy is tied to dose-limiting toxicities, which prevent dose escalations to therapeutic levels. Therefore, clinical dosing, efficacy, and safety must be evaluated as interdependent factors. Overall, clinical safety concerns, including dose-limiting toxicities and other adverse effects, are more likely responsible for up to 70% of drug development failures[2-4].

Traditional preclinical toxicity assessments entail comprehensive and recurrent examination of drug-like properties, reactive metabolites, nonspecific off-targets, and drug-drug interactions (DDIs), as well as standard endpoints including cardiotoxicity, mutagenicity, hepatotoxicity, and acute toxicity. However, not only are the most common *in vitro* assays and animal models expensive and time-consuming, they lack relevance to human physiology, leading to unanticipated safety concerns once drug candidates reach late-stage clinical testing. Conventional *in silico* models, *in vitro* toxicity, and GLP toxicity testing in different animal species rarely predict clinical drug toxicities presenting as adverse events (AEs) in patients at therapeutic doses. Indeed, only about half of toxicity predictions generated from rodent and nonrodent models are concordant when extrapolated to humans[5], and this drops to less than 30% when narrowed to clinical AEs in target organs[6].

Marking a paradigm shift for drug safety assessments in the 21st century, several U.S. regulatory initiatives have been announced which seek to use New Approach Methodologies (NAMs) that do not involve animal testing[7-11]. In November 2020, the U.S. Food and Drug Administration (FDA) announced its pilot program called Innovative Science and Technology Approaches for Next Drugs (ISTAND)[7], aiming to provide a pathway for novel approaches to be integrated into drug development and regulatory decision-making specifically focused on non-animal-based methods and technologies. In 2022, the National Center for Toxicological Research (NCTR) launched the AI4TOX program[8], which consists of five initiatives: AnimalGAN[12] for predicting animal toxicology data for untested chemicals through machine learning (ML) models, SafetAI for developing novel deep learning methods for toxicological endpoints, BERTox for developing advanced AI-powered natural language processing to facilitate analysis of FDA documents and public literature, PathologAI for developing an accurate and effective framework for the analysis of histopathological data, and TranslAI for developing generative artificial intelligence (AI) models that facilitate the translation of

experimental findings across domains, such as organ systems, IVIVE, and genomic technologies. In April 2025, the agency officially released its Roadmap to Reducing Animal Testing in Preclinical Safety Studies[9], an explicit, stepwise approach to reduce animal testing in preclinical safety assessments and to replace with scientifically validated NAMs, which encompass *in vitro* human-based systems, *in silico* modeling, and other innovative platforms that can collectively evaluate toxicity, immunogenicity, and pharmacodynamics (PD) in humans and provide an opportunity to improve the predictive relevance of preclinical drug testing.

Recent advancements in AI are being increasingly integrated into scientific discovery to enhance and accelerate drug development research, helping scientists generate hypotheses, design experiments, collect and interpret large amounts of data, and glean insights beyond what's possible using experimental and conventional *in silico* approaches[13,14]. Rapidly evolving ML algorithms, combined with the increasing availability of large datasets, parallel computing capabilities, and storage hardware, promise to perform toxicity assessments faster and more cost-effectively than conventional methods, all while mimicking the complexity and diversity of the human body. In this chapter, we explore how AI/ML, from simple parametric models to complex deep learning architectures, has been applied to the prediction of drug toxicity *in vitro*, in preclinical animal models, and in human patients. Although drug absorption, distribution, metabolism, and excretion (ADME) and pharmacokinetic (PK) properties are important determinants of drug toxicity, AI/ML approaches for predicting these properties have been comprehensively reviewed elsewhere[15-38]. The present review focuses specifically on AI/ML methods for drug toxicity prediction. We examine the methodological evolution of these approaches across major toxicity endpoints, assess their predictive performance and persistent limitations, and consider whether current models adequately capture the pharmacological and physiological determinants of clinical toxicity. We conclude by examining how these limitations might inform the design of the next generation of predictive models.

1.3 A Brief Overview of AI/ML Methods for Predicting Drug Toxicity

AI/ML methods have been increasingly applied to predict drug toxicity *in vitro*, in animal models and in clinical patients. Although ML is a subfield of AI, the paired abbreviation "AI/ML" has been widely adopted in pharmaceutical and regulatory sciences, including by the FDA and EMA, and is used throughout this review to reflect this convention. The following subsections outline the key methodological considerations for applying these methods to toxicity

prediction, the molecular representations used as model inputs, and the performance metrics by which models are evaluated.

1.3.1 Methodological considerations for AI/ML-based toxicity prediction

Toxicity prediction tasks impose specific computational demands that shape the choice of AI/ML method. The most fundamental distinction is between classification, in which a model predicts whether a compound falls into a discrete category such as toxic or non-toxic, and regression, in which it estimates a continuous value such as an LD50, IC50, or percent inhibition. This choice determines which algorithms, loss functions, and evaluation metrics are appropriate, and many toxicity endpoints can be framed as either task depending on whether the available data provide continuous measurements or categorical labels. A second critical choice is between models that operate on expert-defined molecular features, such as physicochemical descriptors and structural fingerprints, and those that learn representations directly from raw molecular structure through graph neural networks (NNs) or transformers. Expert-defined features benefit from decades of medicinal chemistry knowledge and are interpretable by design, but they impose a fixed representational ceiling. Learned representations can in principle capture patterns not anticipated by human experts, but they require larger training sets, and their internal features are often difficult to interpret in chemical terms.

A third consideration, particularly acute in toxicology, is the tradeoff between predictive performance and interpretability. Regulatory decision-making increasingly requires not just a prediction but a mechanistic rationale, favoring methods whose outputs can be traced to specific structural or biological features. Simple models such as logistic regression and decision trees produce inherently interpretable outputs but lack the capacity to capture complex, nonlinear structure-toxicity relationships. Deep learning architectures can model these relationships but function largely as black boxes, necessitating *post-hoc* interpretability methods whose reliability for molecular data is an active area of research. Between these extremes, tree-based ensemble methods such as random forests and gradient boosting machines often achieve competitive predictive performance while retaining some degree of feature importance attribution, which partly explains their continued dominance in toxicity prediction benchmarks despite the availability of more complex architectures.

Dataset characteristics further constrain the choice of method. Unlike domains where deep learning has been transformative, such as image recognition and natural language

processing, toxicity data cannot be generated cheaply or at scale. Every label requires an expensive biological experiment or a clinical observation, and the pace of data generation is constrained by the throughput of *in vitro* assays, animal studies, and clinical trials. As a result, toxicity datasets are typically small, ranging from a few hundred compounds for specific organ toxicities to approximately 10,000 for the largest public benchmarks such as the Toxicology in the 21st Century (Tox21) database[39], and class imbalance is pervasive, with toxic compounds often representing fewer than 10% of a dataset. These limitations necessitate resampling strategies, modified loss functions, or evaluation metrics robust to skewed distributions. The quality of toxicity labels poses an additional challenge. Solubility, permeability, and CYP inhibition are measured in standardized assays with continuous, reproducible outputs, whereas many toxicity endpoints rely on binary classifications that collapse distinct mechanisms, dose dependencies, and patient populations into a single categorical label. This imposes a ceiling on achievable model performance independent of algorithmic sophistication. Missing labels are common in multi-endpoint datasets where not all compounds have been tested across all assays, motivating multi-task learning approaches that can train across endpoints with incomplete label matrices. These practical constraints mean that the most architecturally sophisticated model is not necessarily the most appropriate for a given toxicity prediction task, and several studies reviewed here report that classical ML methods match or outperform deep learning when training data are limited.

Figure 1-1 and **Table 1-1** summarize the methods and learning strategies most commonly employed for toxicity prediction.

Table 1-1. Overview of AI/ML methods used in drug toxicity prediction, with advantages and limitations for safety-critical applications.

Method	Description	Advantages and limitations for toxicity prediction
Classical Machine Learning		
Linear (or least squares) regression	Models linear relationships between molecular features and continuous outcomes such as LD50 or IC50 values.	Interpretable and fast; limited to linear structure-toxicity relationships, which are rare for complex endpoints.
Logistic regression	Models the probability of binary outcomes such as toxic/non-toxic classification using the sigmoid function.	Interpretable coefficients indicate feature importance; struggles with high-dimensional molecular descriptors without regularization.

Method	Description	Advantages and limitations for toxicity prediction
Support vector machines	Separate classes via an optimal hyperplane; kernel functions enable nonlinear decision boundaries. Can also be applied to regression.	Effective with high-dimensional fingerprints via kernel functions; performance sensitive to kernel choice, and models are difficult to interpret.
Ridge or lasso regression	Linear models with L2 (ridge) or L1 (lasso) regularization to reduce overfitting. Applicable to both classification and regression.	Built-in feature selection (lasso) or multicollinearity handling (ridge); assumes linear relationships, limiting flexibility.
Partial least squares regression	Projects predictors and responses into a shared latent space by maximizing covariance; handles collinear descriptors.	Handles collinear descriptors well; rarely competitive with nonlinear methods on complex toxicity endpoints.
Tree-based methods (decision trees, random forest, gradient boosting)	Recursively partition the feature space into decision rules. Random forests combine many trees to reduce variance; gradient boosting builds trees sequentially to correct prior errors.	Random forests are robust and handle mixed feature types well; gradient boosting often achieves top performance on tabular toxicity data. Decision trees alone are interpretable but prone to overfitting.
Naive Bayes	Probabilistic classifier assuming conditional independence of features.	Fast and effective with sparse binary fingerprints; independence assumption is often violated for correlated molecular features.
k-nearest neighbor	Classifies or predicts based on the majority or average of the *k* most similar training compounds.	Simple and effective for local structure-activity relationships; performance degrades in high-dimensional feature spaces.
Gaussian processes	Bayesian method that defines a distribution over functions, providing predictions with uncertainty estimates.	Built-in uncertainty estimates valuable for safety applications; computationally expensive for large compound libraries.
Deep Learning Architectures		
Neural networks	Layers of interconnected neurons with nonlinear activation functions learn complex mappings from molecular features to toxicity outcomes.	Flexible and capable of learning nonlinear relationships; require large datasets and lack inherent interpretability.
Convolutional neural networks	Apply convolutional filters to extract spatial feature hierarchies	Effective for image-based molecular representations; require structured input formats and large training sets.

Method	Description	Advantages and limitations for toxicity prediction
	from structured inputs such as molecular images or 3D structures.	
Recurrent neural networks	Maintain hidden states across sequential inputs to model temporal or sequential dependencies.	Capture sequential dependencies in SMILES or longitudinal data; prone to vanishing gradients for long sequences.
Long short-term memory networks	Gated variant of recurrent neural networks designed to capture long-range dependencies in sequential data.	Address vanishing gradient problem; more complex to train and slower than transformers.
Graph neural networks	Operate on molecular graphs using message passing to update atom-level representations.	Learn directly from molecular structure without manual featurization; performance advantage over fingerprint-based methods is inconsistent across toxicity endpoints.
Generative adversarial networks	Generator and discriminator networks trained adversarially to produce realistic synthetic data.	Useful to augmenting imbalanced toxicity datasets; training is unstable and synthetic data quality is difficult to validate.
Transformers	Use self-attention mechanisms to capture global dependencies across input elements in parallel.	Scale well and capture long-range dependencies; require large training datasets and are computationally expensive.
Dimensionality Reduction and Representation Learning		
Principal component analysis (PCA)	Linear transformation that projects data onto directions of maximum variance.	Effective for reducing descriptor dimensionality; may miss nonlinear structure-toxicity relationships.
Linear discriminant analysis (LDA)	Projects data onto directions maximizing separation between known classes.	Class-aware dimensionality reduction; assumes shared covariance across toxicity classes.
Quadratic discriminant analysis (QDA)	Similar to LDA but allows each class its own covariance structure.	More flexible than LDA; requires more training data per class, often limiting for rare toxicity endpoints.
Autoencoders	Neural networks trained to reconstruct input data through a compressed bottleneck representation.	Learn nonlinear compressed representations; require careful tuning and lack interpretability of linear methods.
Learning Strategies		

Method	Description	Advantages and limitations for toxicity prediction
Multi-task learning	Trains a single model to predict multiple related toxicity endpoints simultaneously, sharing representations across tasks.	Improves prediction on data-sparse endpoints by borrowing information from related tasks; benefits are inconsistent for mechanistically dissimilar endpoints.
Transfer learning	Pretrained model on a large or related dataset is adapted to a new toxicity endpoint with limited data.	Reduces data requirements for novel endpoints; effectiveness depends on similarity between source and target tasks.
Self-supervised and contrastive learning	Model learns molecular representations from unlabeled data through pretext tasks such as masked atom prediction or contrastive pairing before fine-tuning on labeled toxicity data.	Leverages large unlabeled chemical libraries; pretraining gains do not consistently translate to toxicity endpoints.
Ensemble methods	Combine predictions from multiple models to improve robustness. Includes bagging (e.g., random forest), boosting (e.g., XGBoost), and stacking.	Generally improve performance and reduce variance; increase computational cost and reduce interpretability.
Knowledge graph methods	Represent drugs, targets, diseases, and adverse reactions as nodes in a graph with typed edges encoding biological relationships; use graph embeddings or link prediction to infer new associations.	Integrate heterogenous biological data sources for systems-level safety prediction; dependent on completeness and accuracy or underlying databases.

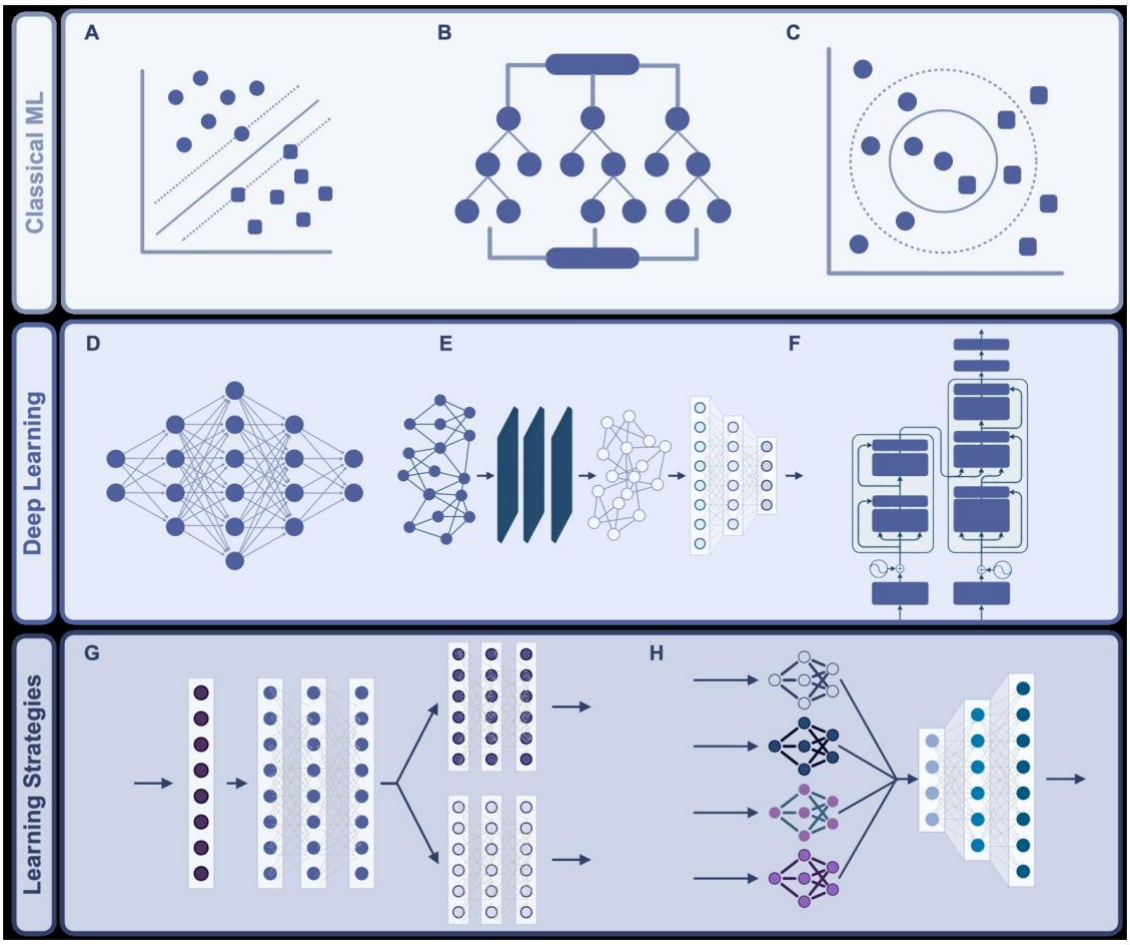


Figure 1-1. Overview of AI/ML methods commonly employed in drug toxicity prediction. Classical ML approaches (top) include (**A**) support vector machines, (**B**) random forests, and (**C**) k-nearest neighbor. Deep learning architectures (middle) include (**D**) multilayer perceptrons, (**E**) graph neural networks, and (**F**) transformers. Learning strategies (bottom) include (**G**) multi-task learning, in which a shared representation is trained across multiple toxicity endpoints, and (**H**) ensemble methods, in which predictions from diverse models are combined. Created in BioRender. Sun, D. [2026] https://BioRender.com/y48kifh.

1.3.2 Transforming drug structures into AI/ML model inputs and features

Applications of AI/ML to drug discovery and drug toxicity prediction are highly dependent on the representation of drug molecules by descriptors capable of capturing their structural characteristics in a format easily processed by machines. One of the most simplistic representations is the Simplified Molecular Input Line Entry System (SMILES) string[40], a line notation which represents primarily the heavy atoms, optionally the hydrogens and bonds as well, of a molecule with a compact and unambiguous sequence of characters. Where SMILES strings provide a human-readable, text-based representation of molecular structure, molecular structure fingerprints encode additional information about a molecule's structural characteristics, especially those substructures predictive of biological activity, in the form of a more machine-

readable continuous, discrete, or categorical vector[41]. Popular types of molecular fingerprints, each encoding slightly different aspects of chemical structure, include substructure key-based fingerprints such as PubChem fingerprints and Molecular ACCess System (MACCS) keys[42], topological or path-based fingerprints like FP2[43], Daylight, and electro-topological state (EState)[44] fingerprints, circular-style fingerprints such as the Morgan[45] or extended-connectivity fingerprints (ECFP)[46], and pharmacophore fingerprints[47-49], which include the relevant features and interactions needed for a drug molecule to be active against a given target. Recent improvements in deep NNs have now enabled the generation of neural fingerprints[50,51], where learned representations of molecular structure are created as fixed-size vectors without the need for pre-defined rule-based generation.

An alternative to the use of molecular structure fingerprints as inputs to AI/ML are molecular descriptors, which are a set of measured or computationally calculated properties of molecules. They can be binary, discrete, or continuous and are typically defined by the dimensionality of the information they capture, as 1-, 2-, 3-, or 4-D descriptors. Accordingly, 1-D descriptors generally cover physicochemical properties and molecular formulas. 2-D descriptors are defined by their derivation from the 2-D, or graph, representation of the molecule and 3-D descriptors are derived from 3-D representations of molecules. While 2-D descriptors tend to be simpler, faster to calculate, and perform well in the prediction of biological targets and physicochemical properties, 3-D descriptors offer more precise geometric and spatial information, with more use cases in the analysis of ligand-protein binding and for molecules with low structural similarity to others within the same dataset. 4-D molecular descriptors, introduced by Hopfinger et al.[52] in 1997, are derived from ensemble sampling or conformational sampling calculated through molecular dynamics simulations; they are useful in alignment problems and in identifying the active conformations of flexible molecules[53]. There are a variety of user-friendly tools available for calculation of thousands of molecular descriptors, including but not limited to RDKit, Mordred[54], CDK[55], PaDEL[56], and DRAGON. The number and dimensionality of molecular descriptors for AI/ML should be selected based on their relevance to the prediction task as well as the complexity of the problem; high-dimensional descriptors are not recommended for small datasets because they increase the dimensionality of the overall modeling task and thus make data sparser. Generally, the number of molecules in the dataset should be significantly larger than the number of dimensions of the descriptors.

The choice of molecular representation can have as large an impact on predictive performance as the choice of algorithm. Multiple studies reviewed here report that the same model architecture yields substantially different results depending on whether it receives MACCS keys, extended-connectivity fingerprints, graph-based inputs, or physicochemical descriptors, and no single representation consistently outperforms across all toxicity endpoints. This suggests that the information encoded by each representation captures different and partially complementary aspects of structure-toxicity relationships, and that the field's emphasis on algorithmic novelty may underweight the importance of input representation.

1.3.3 Performance metrics in AI/ML for drug toxicity prediction

Performance metrics for toxicity prediction models must be aligned with both the prediction task and the decision context in which the model will be used. For regression tasks, such as estimating LD50 values or IC50 concentrations, adjusted R^2 penalizes overfitting to correlated molecular descriptors, while weighted Mean Absolute Error (MAE) or Mean Bias Error (MBE) may be preferred when the directionality of prediction error has practical consequences, such as when overestimating toxicity severity leads to unnecessary compound attrition. For classification tasks, standard accuracy is misleading when classes are imbalanced, as is typical in toxicity prediction where toxic compounds represent a small fraction of any chemical library. Precision, recall, AU-ROC, and MCC are more informative in these settings, but their relative importance depends on the stage of drug development and the severity of the endpoint. In early-stage screening, recall (sensitivity) should be prioritized over precision, as failing to flag a toxic compound carries greater consequence than incorrectly flagging a safe compound, which results in unnecessary attrition but not patient harm. That said, if downstream experimental validation is resource-constrained, an unchecked false positive rate can overwhelm screening capacity, and practitioners may need to balance recall against the practical throughput of the pipeline. At later stages, when the cost of falsely terminating a promising candidate increases, precision and specificity become more relevant. For severe or irreversible endpoints such as carcinogenicity or QT prolongation, the cost asymmetry favoring recall is particularly pronounced. Area under the receiver operating characteristic curve (AU-ROC) is widely reported and measures overall discrimination ability, but for highly imbalanced datasets, the area under the precision-recall curve (AU-PRC) can provide a more informative assessment of performance

on the minority class. Matthew's Correlation Coefficient (MCC), which accounts for all four cells of the confusion matrix, is similarly recommended when class distributions are skewed.

No single metric captures all dimensions of model quality for safety applications, and reporting a single headline metric, as many studies do, obscures critical tradeoffs. A model with high AU-ROC but low precision at the operating threshold may generate too many false alarms to be practically useful in a screening cascade, while a model optimized for precision may miss rare but severe toxicities. Beyond discrimination, calibration, meaning whether a predicted probability of 0.7 corresponds to a true positive rate of approximately 70%, is essential for risk-based decision-making but is rarely assessed in the toxicity prediction literature. The choice of classification threshold is itself a decision with safety implications. A default threshold of 0.5 optimizes balanced accuracy but may be inappropriate when the costs of false negatives and false positives are asymmetric, and few studies report threshold sensitivity analyses or justify their threshold selection in the context of the intended application. Cross-study comparisons are further complicated by differences in dataset composition, label definitions, and splitting strategies. A model reporting an AU-ROC of 0.95 on a random split of a curated dataset cannot be meaningfully compared to one reporting an AU-ROC of 0.85 on a scaffold split of a noisier dataset, yet such comparisons are routinely made in the literature. Temporal or prospective validation, where models are tested on compounds discovered after the training set was assembled, provides the most realistic assessment of generalization but is rarely performed. **Table 1-2** provides an overview of the most common performance metrics employed in AI/ML analyses of drug toxicity covered by this chapter.

Table 1-2. Performance metrics for AI/ML-based toxicity prediction, with guidance on suitability for imbalanced datasets and safety-critical decision contexts. Formulas are presented in plain-text approximations.

Performance metric	Formula	Range	Description	Suitability for safety-critical contexts
Regression metrics				
R^2	$R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$	[0, 1]; R^2 = 1 is a perfect fit	Proportion of variance in outcomes explained by the model.	General purpose; does not penalize model complexity.

Performance metric	Formula	Range	Description	Suitability for safety-critical contexts
Adjusted R^2	$R^2_adj = 1 - [(n - 1)/(n - k - 1)] \times (1 - R^2)$; n = number of observations, k = number of inputs	[0, 1]; theoretical perfect fit at 1	Adjusts R^2 based on the number of features used, penalizing unnecessary variables.	Preferred over R^2 when models use large molecular descriptor sets.
Mean absolute error (MAE)	$MAE = (1/N) \times \Sigma \lvert y_i - \hat{y}_i \rvert$	[0, ∞); MAE = 0 is a perfect fit	Average magnitude of errors between predicted and actual values, in the same units as the outcome.	Treats all errors equally; appropriate when over- and under-prediction carry similar consequences.
Mean squared error (MSE)	$MSE = (1/N) \times \Sigma (y_i - \hat{y}_i)^2$	[0, ∞); MSE = 0 is a perfect fit	Average squared error between actual and predicted values.	More sensitive to large errors; useful when large prediction errors are disproportionately costly.
Root mean squared error (RMSE)	$RMSE = \sqrt{[(1/N) \times \Sigma (y_i - \hat{y}_i)^2]}$	[0, ∞); RMSE = 0 is a perfect fit	Square root of MSE, retaining the same units as the predicted variable.	Same sensitivity to large errors as MSE but more interpretable.
Classification metrics				
Accuracy	(TP + TN) / (TP + FP + FN + TN) × 100%	[0, 100]; 100% = perfect performance	Proportion of correctly classified compounds.	Misleading for imbalanced datasets; not recommended as a primary metric for toxicity prediction.
Precision	TP / (TP + FP)	[0, 1]; 1 = perfect precision	Proportion of predicted toxic compounds	Prioritize at later development stages when false positives

Performance metric	Formula	Range	Description	Suitability for safety-critical contexts
			that are truly toxic.	carry high cost (e.g., unnecessary termination of promising candidates).
Recall (sensitivity)	TP / (TP + FN)	[0, 1]; 1 = perfect recall	Proportion of truly toxic compounds correctly identified.	Prioritize in early screening and for severe endpoints where missing a toxic compound carries the greatest risk.
Specificity	TN / (TN + FP)	[0, 1]; 1 = perfect specificity	Proportion of non-toxic compounds correctly classified.	Important when minimizing unnecessary attrition of safe compounds.
F1 score	2 × (Precision × Recall) / (Precision + Recall)	[0, 1]; 1 = perfect precision and recall	Harmonic mean of precision and recall.	Useful when both false positives and false negatives matter; does not account for true negatives.
Matthews correlation coefficient (MCC)	(TP×TN − FP×FN) / √[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]	[−1, 1]; 1 = perfect prediction	Correlation coefficient between observed and predicted classifications.	Unlike accuracy, scores high only when a model performs well on both classes relative to their sizes; recommended for imbalanced toxicity datasets.
AU-ROC	Area under TP-rate vs. FP-rate curve (no closed-form expression)	[0.5, 1]; 1 = perfect performance	Summarizes discrimination performance across all classification thresholds.	Widely used but can overstate performance on highly imbalanced datasets; consider

Performance metric	Formula	Range	Description	Suitability for safety-critical contexts
				supplementing with AU-PRC.
AU-PRC	Area under precision vs. recall curve (no closed-form expression)	[0, 1]; 1 = perfect performanc e	Summarizes the trade-off between precision and recall across all classification thresholds.	More informative than AU-ROC for imbalanced datasets where the positive (toxic) class is rare.

The following sections review the application of these methods to specific toxicity endpoints, organized by the development stage at which toxicity is assessed. Preclinical applications span hERG channel blockage, drug-induced liver injury (DILI), acute toxicity, genotoxicity, multi-endpoint prediction, drug-target interaction (DTI) profiling, and protein and peptide toxicity. Clinical applications encompass DDIs, patient-level AE prediction, pharmacogenomic risk assessment, and AE prediction from chemical structure and systems pharmacology data. This progression from preclinical to clinical prediction reflects both the historical development of the field and an increasing level of data complexity, as clinical toxicity prediction must contend with patient-level variability, heterogenous reporting systems, and pharmacological context that are largely absent from preclinical datasets.

1.4 AI/ML Models for Predicting *In Vitro* and Preclinical Drug Toxicities

For conventional preclinical toxicity prediction, various *in silico* methods such as read-across, structural alerts, and quantitative structure-activity relationship (QSAR) modeling have been developed to predict the toxicity profile of chemical compounds. QSAR modeling, the most widely used of those listed, refers to an approach which correlates a compound's quantitative descriptors of chemical structure with its toxicological effect. Since their introduction to the field of toxicology in the early 1960's[57], the concepts of QSAR modeling have spread to drug development and now include more sophisticated ML algorithms for prediction.

1.4.1 AI/ML models to predict hERG channel blockage for cardiotoxicity assessments

Cardiotoxicity is among the most assessed endpoints when screening drug candidates as drug-induced clinical cardiovascular events are a major cause of drug development failures and can have life-threatening consequences for patients. Encoded by what's commonly known as the human ether-a-go-go gene (hERG), hERG proteins are potassium ion channel proteins expressed in human muscle, various brain regions, endocrine cells, and heart, and their blockage is associated with potentially fatal prolongation of the QT interval. The first ML-QSAR models to classify drug compounds as hERG blockers or non-blockers were published as early as 2002[58], and often focused on fitting just one classification algorithm on relatively small experimental datasets of IC50. Drug compounds are stratified based on a hERG activity threshold that ranges from 1 to 40 μM; while 1 and 10 μM have been used as common cutoffs for potential cardiotoxicity[59], there was and is still no widely accepted threshold for toxicity prediction. Among the earliest ML algorithms evaluated were partial least squares regression[60,61], naive Bayes[62-64], Gaussian Process[65], decision trees[66], support vector machines[67-73], and NNs[58]. The choice of activity threshold directly affects both class balance and model performance: a stringent threshold of 1 μM yields fewer positives and higher class imbalance but captures only the most potent blockers, while a permissive threshold of 10 μM produces more balanced datasets but includes compounds with marginal hERG activity that may not translate to clinical cardiotoxicity. Few studies systemically evaluate performance across multiple thresholds, though Chen et al.[74] demonstrated that binary threshold classification outperformed single threshold approaches.

In recent years, much larger hERG IC50 datasets constructed on hundreds of drug compounds became available for extraction from Tox21, BindingDB[75], PubChem[76], and ChEMBL[77] public databases. These larger datasets have not only allowed toxicologists to compare multiple ML methods in order to identify the most predictive approach, but also to construct more sophisticated deep learning model structures. In 2019, several research groups demonstrated the utility of deep NNs constructed on concatenated vectors of molecular structure features for hERG inhibition predictive modeling[78-80], using activity thresholds of either 5 or 10 μM. Zhang et al.[78] developed deep NN models for predicting hERG K+ channel blockage using various sampling methods and 2-D molecular descriptors. The optimal model (M15) with three hidden layers and undersampling achieved 78% accuracy and a 0.75 F1-score on the test set, outperforming 35 other models including random forest approaches. Cai et al.[79] developed

deephERG, a multi-task deep NN for predicting hERG channel blockade using 7,889 compounds with structurally diverse features and defined hERG experimental data. The best model achieved an AU-ROC of 0.967 on validation data, outperforming single-task deep NNs, random forests, support vector machines, and other ML approaches. The model identified 29.6% of FDA-approved drugs as potential hERG inhibitors, demonstrating its value for drug safety assessment in both discovery and post-market surveillance. Lee et al.[80] developed an NN trained on 2,130 compounds under uniform experimental conditions, achieving an AU-ROC of 0.764 and 90.1% accuracy. Validated on ten compounds in guinea pigs with 80% accuracy, the model demonstrated that deep NNs can predict both *in vitro* and *in vivo* cardiotoxicity.

In 2020, Ryu et al.[81] constructed DeepHIT (**Figure 1-2**), a multi-representation ensemble framework that processes each compound through three parallel pipelines: a deep NN trained on physicochemical and topological molecular descriptors, a deep NN trained on combined ECFP and PubChem fingerprints, and a graph convolutional network trained on molecular graphs. Rather than selecting the best-performing model, DeepHIT applies a union decision rule that classifies a compound as a hERG blocker if any of the three models predicts blocking activity, explicitly prioritizing sensitivity over specificity to reflect the asymmetric cost of missing a cardiotoxic compound in early drug discovery. Trained on 14,440 compounds with a hERG activity threshold of 10 M, the ensemble achieved a sensitivity of 0.833 and accuracy of 0.773 on a structurally independent test set, outperforming all comparator tools.

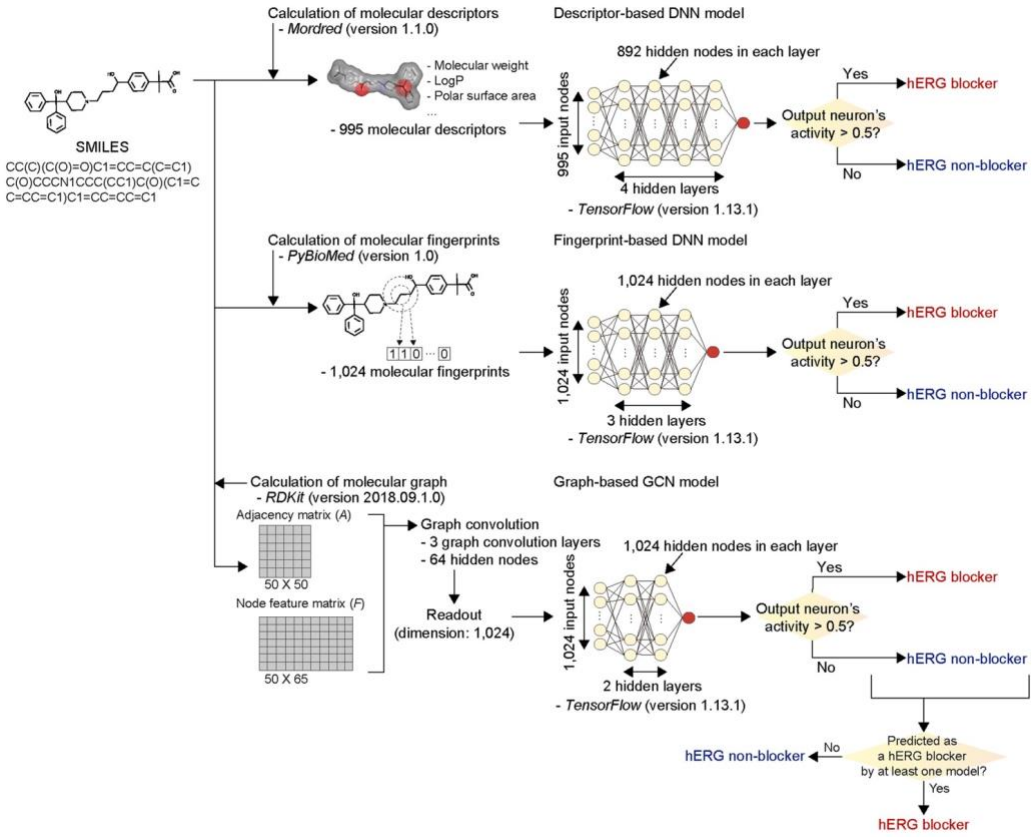


Figure 1-2. Schematic overview of the DeepHIT framework for prediction of hERG-induced cardiotoxicity. A compound provided in SMILES format is processed in parallel through three independent pipelines: a descriptor-based deep neural network operating on physicochemical and topological molecular descriptors, a fingerprint-based deep neural network operating on a combined ECFP and PubChem fingerprint vector, and a graph convolutional network operating on the molecular graph. Each model independently generates a binary prediction, and DeepHIT applies a union decision rule, classifying a compound as an hERG blocker if at least one model predicts blocking activity, to maximize sensitivity for reliable identification of cardiotoxicity risk. Ryu JY, Lee MY. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. *Bioinformatics.* 2020;36(10). By permission of Oxford University Press.

Since then, state-of-the-art ML cardiotoxicity models are consistently trained on combinations of raw or engineered molecular features and employ ensemble or consensus methods to improve prediction performance[74,82-87]. Recent work by Chen et al.[74] in 2023 employed graph convolutional networks alongside traditional ML approaches on a dataset of 4,556 compounds with patch clamp-determined IC50 values for hERG channel blockage. Their graph convolutional network models achieved superior performance with AU-ROCs above 0.952 by directly extracting molecular features from graph representations, while support vector machines and random forest models trained on Morgan fingerprints and other molecular descriptors achieved AU-ROCs of 0.884-0.950. Chen et al. also identified seven structural alerts associated with hERG blockage and demonstrated that binary threshold classification

outperformed single threshold approaches for more accurate hERG blocker prediction. Yang et al.[85] developed AttenhERG, a graph NN framework based on the Attentive-FP algorithm that achieved an AU-ROC of 0.835 on 14,322 compounds. The framework used 39 atomic descriptors (atom symbol, covalent bonds, charge, hybridization, aromaticity) and 4 bond features (type, conjugation, ring status, stereochemistry) as molecular graph inputs. The model incorporated uncertainty estimation techniques (Entropy and MC-Dropout) and dual-level attention mechanisms to capture both local atomic features and global molecular characteristics, demonstrating superior interpretability through attention weight visualization for structural optimization guidance. Similarly, Karim et al.[86] developed ensemble QSAR models using six individual approaches that achieved MCCs of 0.682-0.730 in cross-validation, surpassing previously reported models on the same dataset. Their ensemble approach demonstrated the continued value of combining multiple modeling strategies for robust hERG prediction, with external validation MCCs ranging from 0.520-0.715, highlighting the importance of rigorous external testing for model generalizability. In March 2025, Jing et al.[88] released GraphDeep-hERG, an integrated atom embedding for graph-based molecular feature generation and graph NN trained on 7,909 ChEMBL compounds. GraphDeep-hERG achieved an accuracy of 0.93 in representing molecular structures and a predictive accuracy of 0.84 on hERG inhibitor classification.

The evolution of hERG prediction models over the past two decades reflects broader trends in AI/ML, marked by a progressive shift from classical algorithms trained on small, curated datasets towards deep learning architectures capable of extracting complex structure-activity relationships from increasingly large and chemically diverse compound libraries. The availability of larger curated datasets from Tox21, ChEMBL, and BindingDB catalyzed this transition, with graph-based representations showing particular promise by encoding molecular topology directly rather than relying on pre-computed descriptors, as evidenced by the strong performance of graph convolutional network-based models such as those by Chen et al.[74] and Jing et al.[88]. Ensemble and consensus strategies have also emerged as a consistent theme, leveraging complementary feature representations to improve robustness[82,86]. However, several methodological challenges persist. hERG activity thresholds remain unstandardized across studies, ranging from 1-40 M, making direct cross-study performance comparisons difficult. External validation performance consistently lags behind cross-validation metrics, as illustrated

by the MCC drop observed by Karim et al., raising concerns about overfitting and limited chemical diversity in training sets. Furthermore, few studies have systemically addressed activity cliffs, where structurally similar compounds exhibit dramatically different hERG binding affinities, which remains a fundamental challenge for structure-based prediction.

Beyond modeling limitations, caution is warranted when extrapolating these findings to clinical cardiotoxicity. *In vitro* models, including those AI/ML algorithms based on experimental hERG IC50 thresholds, do not fully capture the complexity of cardiovascular physiology, drug metabolism, off-target effects, or patient-specific risk factors like genetic polymorphisms and comorbidities. Many drugs known to block hERG channels do not cause QT prolongation or torsades de pointes *in vivo* and conversely, some cardiotoxic agents act through mechanisms independent of hERG inhibition[89-91]. Whether hERG blockade translates to clinical cardiotoxicity depends on cardiac tissue drug concentrations at therapeutic doses and concurrent interactions with other cardiac ion channels, information that is absent from all models reviewed here. Therefore, AI/ML hERG inhibition models should be interpreted as early safety screening rather than definitive predictors of clinical cardiotoxicity, and their outputs should be interpreted within the broader context of integrated preclinical and clinical toxicity assessment.

1.4.2 AI/ML models to predict drug-induced liver injury on the basis of clinical reporting

DILI is also of significant concern during the drug development process because it is a clinically significant side effect of many drugs; however, the precise mechanisms underlying hepatotoxicity are often unclear. Large doses of hepatically cleared compounds have been shown to increase the risk of liver toxicity through the saturation of metabolic processes or the buildup of metabolites[92], but immune-mediated mechanisms of inflammation and genetic factors[93,94] may also play a role in patient risk. Drug compounds that are known to cause severe DILI in humans do not demonstrate clear hepatotoxicity in animal models[95] – preclinical evaluation of potential drug candidates is therefore highly limited, and chemical compounds used for predictive modeling are largely stratified on the basis of clinical DILI reporting.

Early research in this area evaluated mainly traditional ML algorithms such as LDA[96,97], *k*-NN[98,99], decision trees[100], or naive Bayes[101-103] on smaller, manually constructed datasets with varying criteria for DILI classification. In 2016, Chen et al.[104,105] released DILIrank, a dataset composed of 775 drugs annotated for DILI risk divided into three ranks, ‘Most-DILI-Concern’, ‘Less-DILI-concern’, and ‘No-DILI-concern’, omitting drugs without verified causality. This

DILI classification began to be incorporated into more recent studies[106-110], bolstered by its use in the 2019 Critical Assessment of Massive Data Analysis (CAMDA) cMAP Drug Safety Challenge. This standardized approach to classifying compounds as hepatotoxic or non-hepatotoxic has also proven effective with Minerali et al.[107], who achieved robust predictive performance using DILIrank categories. Their best Bayesian model based on DILI-concern classifications achieved an AU-ROC of 0.814, sensitivity of 0.741, and specificity of 0.755. Minerali et al. leveraged DILI *in vitro* data collated by Pfizer and AstraZeneca alongside physicochemical properties to enhance model performance. Nguyen-Vo et al.[106] constructed a deep learning framework using a convolutional NN combined with Morgan fingerprint-embedded features to predict DILIrank categories. The model, trained on a combined dataset of 1,597 compounds and validated on an independent set of 322 compounds, achieved an AU-ROC of 0.96, accuracy of 0.89, and MCC of 0.80. The adoption of DILIrank's framework extends beyond individual studies, as evidenced by its integration into comprehensive toxicity prediction platforms called MegaTox[111], which combines multiple ML algorithms (Bayesian, k-nearest neighbors, support vector machines, AdaBoost, and deep learning) to predict hepatotoxicity across early-stage clinical compounds and FDA-approved drugs.

DILIst[112], a version of DILIrank expanded to include 1,279 drugs, 768 DILI-positive and 511 DILI-negative from a broader range of therapeutic categories, was released in 2020. Thakker et al.[112] augmented DILIrank with four literature sources using statistical concordance criteria (>75% agreement). The expanded dataset enabled comprehensive analysis across 381 chemical subgroups using WHO's Anatomical Therapeutic Chemical (ATC) classification system, revealing DILI hotspots including antidepressants (86% DILI-positive), anti-inflammatory drugs (>90% DILI-positive), and ACE inhibitors (81% DILI-positive). This expanded dataset provides a more comprehensive reference for DILI research and regulatory applications, offering improved coverage across therapeutic categories and enhanced identification of drug class-specific hepatotoxicity patterns. In 2024, Seal et al.[113] achieved an AU-ROC of 0.79 using a series of random forest models trained on chemical structure, physicochemical properties, PK parameters, and both *in vitro* hepatotoxicity data and DILIst. The final model, DILIPredictor, reports the ability to identify chemical substructures associated with DILI and can differentiate between compounds which cause hepatotoxicity in animal models but are considered DILI-negative in humans. In January 2025, Li et al.[114] constructed a computational framework trained

on DILIrank along with manifold embedding of molecular structure (MEMS) data, encoding quantum mechanical features such as electrostatic potential and Fukui functions from molecular surfaces into 2-D embeddings processed by DeepSets architecture. Their DeepSets NN architecture ensures permutation invariance and uses self-attention mechanisms, trained on 787 molecules from DILIrank. The final model achieved an F1-score of 0.93, a recall of 0.97, precision of 0.90, and an accuracy of 0.90, outperforming conventional molecular descriptors and molecular fingerprint-based training approaches. The MEMS representation captures quantum information governing intermolecular interactions of compound structures, enabling more direct structure-activity relationships for hepatotoxicity prediction as compared to traditional molecular representations. In contrast to the structure-based approaches described above, Bergen et al.[115] demonstrated that transcriptomic profiling of primary human hepatocytes exposed to compounds at multiple concentrations, combined with pathway-level ML classification, can capture DILI mechanisms inaccessible to molecular structure alone. Their framework, ToxPredictor, achieved 88% sensitivity and 100% specificity for DILI prediction in blind validation, outperforming both structure-based *in silico* models and conventional cytotoxicity assays. Notably, the framework provides dose-resolved margin-of-safety estimates rather than binary classifications and flagged several clinical-stage compounds that failed due to hepatotoxicity, including candidates missed by preclinical animal studies.

A consistent theme across the DILI prediction literature is the outsized influence of dataset construction on model performance. The release of DILIrank and its successor DILIst represented inflection points for this area of study, providing standardized clinical annotations that enabled more meaningful cross-study comparisons and catalyzed the adoption of deep learning approaches. However, these datasets remain relatively small (fewer than 1,300 compounds in DILIst), and the binary or ternary DILI classifications they employ necessarily oversimplify a heterogeneous clinical phenomenon that encompasses dose-dependent hepatocellular injury, cholestatic patterns, and idiosyncratic reactions with distinct underlying mechanisms[93,94]. This heterogeneity poses a fundamental challenge for AI/ML models, as compounds classified under the same DILI label may cause liver injury through entirely different pathways. It is also worth noting that, unlike other preclinical toxicity endpoints discussed in this review, DILI prediction models are largely trained on clinical outcome data rather than preclinical assay measurements, reflecting the poor concordance between animal models and

human hepatotoxicity[95]. This means these models are fundamentally retrospective, identifying structural patterns associated with known clinical DILI rather than prospectively predicting hepatotoxic liability from preclinical data alone. Notably, the progression from traditional molecular descriptors towards more expressive representations, including Morgan fingerprints, graph-based encodings, and quantum mechanical surface embeddings[114], has yielded incremental performance gains, suggesting that richer structural information improves predictive capacity. Yet even the best-performing models do not account for patient-level variability driven by genetic polymorphisms, comedications, and pre-existing hepatic conditions, which are among the primary determinants of clinical DILI risk but are absent from structure-based training data. Similarly, none of these models incorporate hepatic drug concentrations or tissue-specific target engagement, which determine whether a compound reaches sufficient concentrations in the liver to cause injury at therapeutic doses. Future progress will likely depend on integrating chemical structure features with biological readouts and patient-level clinical and genomic data, as well as on expanding DILI datasets with more granular mechanistic annotations. The ToxPredictor framework illustrates this direction, achieving substantially higher sensitivity than structure-based models by incorporating transcriptomic responses and dose-resolved exposure data.

1.4.3 AI/ML algorithms to predict acute toxicity in animal models

Acute toxicity refers to rapid or severe adverse effects that result from a single dose or short-term exposure to a chemical substance in animal models. LD50 values, which represent the dose of a chemical compound that would be expected to kill 50% of the animals tested, are used to assign compounds to one of four categories for oral, dermal, and inhalation toxicity as established by the U.S. Environmental Protection Agency (EPA)[116,117]. Early ML approaches to acute toxicity prediction focused on relatively small, species-specific datasets and simple molecular descriptors, establishing baseline performance but limited by dataset size and chemical diversity. Kaiser et al.[118-120] were among the first to apply ML algorithms to the prediction of LD50 values, applying linear regression and NNs for fathead minnow, *Tetrahymena pyriformis*, and *Daphnia magna* toxicity data as early as 1998. These models use molecular descriptors, derived from the chemical structure of the compound, as inputs to the NN, and are then trained on data from 865 chemicals. The NN and linear regression approaches achieved an R^2 value of 0.9316. Their work inspired a wave of consensus ML modeling, using either classification to categorize chemicals as ‘toxic’ or ‘non-toxic’[121], regression to predict LD50 values directly[122-

[126], or a combined multi-tasking approach[127-131], utilizing molecular descriptors and fingerprints of compounds as inputs to models such as gradient boosting, support vector machines, and NNs in order to predict LD50 values. In 2006, Xue et al.[121] compared logistic regression, decision trees, k-nearest neighbors, NNs, and support vector machines for classifying compounds as toxic or non-toxic to *Tetrahymena pyriformis*, achieving accuracies of 0.71 to 0.94 using feature-selected molecular descriptors. In 2023, Ryu et al.[131] developed PredAOT by comparing various ML models such as message passing NNs, graph NNs, random forests, and support vector machines using molecular descriptors and molecular fingerprints as inputs. Their model classified compounds as 'toxic' or 'non-toxic' in rats and mice using LD50 ≤ 300 mg/kg as a cutoff. They found that random forests using SMOTE, a popular oversampling technique, had the highest accuracy of 95.9% and 93.4% in mice and rats respectively, with an AU-ROC of 0.778 and MCC of 0.551. As with hERG thresholds, the choice of LD50 cutoff for binary classification affects the sensitivity-specificity tradeoff: a threshold of 300 mg/kg captures severely toxic compounds but may miss moderately toxic agents, while higher thresholds such as 2000 mg/kg align with regulatory categories but dilute the positive class with compounds of lower toxicological concern. The absence of standardized thresholds across studies complicates cross-model comparisons, as apparent differences in performance may partly reflect differences in label definition rather than algorithmic superiority.

In 2021, a collaborative effort between the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) and the EPA National Center for Computational Toxicology (NCCT) resulted in a comprehensive repository of acute oral LD50 data on 11,992 chemicals[132], spurring the development of more complex deep learning algorithms[126,131,133,134]. In 2023, Jain et al.[135] generated the largest publicly available dataset to date, comprising over 80,000 compounds measured against 59 acute systemic toxicity endpoints. Daghighi et al.[136] used this dataset to evaluate a consensus of ML algorithms for regression. As inputs, they used images composed of 1-D molecular descriptors and subsequently trained a convolutional NN on the images. This approach yielded the highest prediction accuracy against other models such as random forests or support vector machines, with an $R^2 = 0.93$ on internal validation and $R^2 = 0.70$ on external validation, demonstrating that predicted Z-scores of LD50 value was highly correlated with experimental Z-score values. In May 2025, Mustafa et al.[137] applied deep learning architectures TabNet and TabTransformer, trained on 12,228 training and

3,057 test compounds each characterized by 801 molecular descriptors for the classification of 12 acute toxicological endpoints, including both nuclear receptor response signals and stress response indicators. TabNet and TabTransformer both utilize feature engineering, considering the molecular, topological, geometric, and 3-D features of chemical structures. Both deep learning architectures consistently outperformed traditional classifiers, achieving an AU-ROC of up to 0.96 for endpoints such as oxidative stress and p53 pathway activation predicted with chemical structure data.

The consolidation of large public repositories, from the NICEATM-NCCT collection of nearly 12,000 chemicals to the Jain et al. dataset of over 80,000 compounds, has been the primary catalyst for methodological advancement in this area, enabling a shift from classical algorithms trained on small descriptor sets to deep architectures including convolutional NNs and tabular transformers. However, gains in model complexity have not translated to proportional improvements in external validity: Daghighi et al.[136] report $R^2 = 0.93$ on internal validation but $R^2 = 0.70$ on external data, a gap that likely reflects mechanistic heterogeneity inherent to the LD50 endpoint itself. Acute lethality is a whole-organism outcome that can arise through CNS depression, cardiac arrest, respiratory failure, hepatic necrosis, or renal failure, among other pathways. A model that captures structure-lethality correlations for one mechanism of death will not necessarily generalize to another, and no current architecture explicitly disentangles these mechanistic contributions.

Compounding this, the LD50 data aggregated in these repositories derive from different species, routes of administration, and laboratories spanning decades, introducing variability that is biological and procedural rather than chemical. More broadly, while useful for early hazard detection, LD50-based models should not be assumed to predict severe human toxicities without complementary data from chronic toxicity assays, mechanistic analyses, or clinical observations at therapeutically relevant dose levels.

1.4.4 AI/ML models to predict genotoxicity in vitro

The genotoxicity of a chemical compound refers to its ability to damage DNA sequence or structure, which can lead to cancer or heritable damage. Preclinical assessment of genotoxicity during drug development according to FDA guidance[138] includes assessment of mutagenicity in a bacterial reverse gene mutation test, also known as the Ames test[139], which has been shown to detect relevant genetic changes in a majority of rodent and human carcinogens. *In silico*

assessments of genotoxicity have evolved to incorporate AI/ML algorithms commonly constructed on Ames mutagenicity benchmark datasets developed by Hansen et al.[140] in 2009, comprising approximately 6,500 chemical compounds along with their biological activity, or the 1,547 chemicals in the Carcinogenic Potency Database (CPDB)[141], published in 2008. The early ML approaches applied to the classification of drug compounds as 'genotoxic' or 'non-genotoxic' use 1-D molecular descriptors as model inputs and employ support vector machines[142-144], naive Bayes[145], decision tree[146,147], LightGBM[148], and NN[149] algorithms. Ramesh and Veerappapillai[147] developed ML models using 272 compounds with micronucleus assay data trained on six types of molecular fingerprints, including PubChem and MACCS keys, and 325 chemical descriptors. Feature selection using assessment of variance, correlation filtering, and recursive feature elimination reduced their inputs to 42 descriptors and six fingerprints of compound structures to train six classifiers. Among all models, the random forest algorithm trained on molecular fingerprints achieved accuracies up to 0.97 and F1-scores of 0.95 during external validation and identified chlorocyclohexane and trimethylamine as structural alerts predictive of genotoxicity. LightGBM[148] demonstrated superior performance across Tox21 and mutagenicity datasets while achieving significantly faster computational times than XGBoost, deep NNs, random forest, and support vector machines through rigorous Bayesian optimization-integrated nested 10-fold cross-validation. The algorithm's leaf-wise tree growth strategy maintained high predictivity while resolving traditional gradient boosting's scalability limitations, enabling rapid prediction of large compound libraries. Chu et al.[150] applied 8 classification algorithms to the 7,617 training and 731 test compounds in the Xu et al.[149] mutagenicity dataset. The top-performing support vector machines and XGBoost models achieved an AU-ROC, sensitivity, and specificity greater than 0.65 on the test data. Chu et al. constructed 112 predictive models using chemical fingerprints combined with molecular properties, with y-randomization validation confirming model robustness. Their best models outperformed commercial tools including StarDrop and TEST mutagenicity prediction systems when validated against Class A mutagenic compounds from the Ames/QSAR International Challenge Project[151]. Titania[152], a modeling framework composed of *k*-nearest neighbors and random forest algorithms trained on 777 different 1-D and 2-D structural descriptors generated via Mold2 software for the prediction of nine endpoints, including mutagenicity, was released in April 2025 on a web-based application designed to democratize access to advanced

computational tools. The Titania platform incorporates comprehensive QSPR/QSTR models using 5,738 compounds from the Hansen et al.[140] dataset, achieving robust mutagenicity prediction performance (accuracy = 0.767, MCC = 0.529) through random forest algorithms. Mittal et al.[153] took a mechanistically grounded approach to carcinogenicity prediction with Metabokiller, an ensemble framework that constructs six independent classifiers targeting distinct carcinogen-associated biochemical properties, including electrophilicity, epigenetic modification potential, genomic instability, oxidative stress, proliferative capacity, and anti-apoptotic response, then combines their outputs via gradient boosting. Trained on experimentally validated carcinogens, Metabokiller achieved 98% recall on an independent held-out set and was subsequently applied to screen the entire human metabolome, identifying previously uncharacterized carcinogenic metabolites that were validated experimentally through genomic instability assays and malignant transformation studies.

Recent development of predictive genotoxicity models leverages convolutional NNs and graph convolutional networks to learn directly from low-level encoded data without the need for chemical feature extraction. In 2021, Hung and Gini[154] constructed a graph convolutional network which uses SMILES strings to produce 2-D molecular graphs as model inputs which are then processed through an NN-based structure to predict the outcome of the Ames test on the Hansen et al.[140] dataset. They achieved superior performance as compared to traditional descriptor-based approaches while automatically identifying toxicologically relevant substructures, including known structural alerts such as α, β-unsaturated carbonyls and aromatic amines, through attention mechanisms. The model achieved full dataset coverage and outperformed existing commercial prediction tools. Similarly, Tran et al.[155] developed AMPred-CNN, which converts molecular SMILES to 224×224 pixel RGB images as convolutional NN inputs, achieving exceptional performance (accuracy = 0.899, AU-ROC = 0.954, MCC = 0.803). Tran et al. compared five distinct molecular featurization methods, ECFP6, MACCS, RDKit2D, Mordred descriptors, and molecular images, across traditional ML algorithms and convolutional NNs, demonstrating that image-based molecular representations combined with convolutional NNs significantly outperform classical ML algorithms trained on conventional molecular descriptors.

Most models reviewed here are trained on Ames mutagenicity datasets, principally Hansen et al.[140] and Xu et al.[149] which capture only one mechanism of genotoxicity, frameshift

and base-pair substitution mutagenicity in bacterial strains. This is notable given that FDA guidance[138] requires a battery of genotoxicity tests including Ames, *in vitro* chromosomal aberration or micronucleus, and *in vivo* micronucleus assays; models that predict only Ames outcomes address one component of a multi-assay regulatory requirement, and a compound that passes an Ames prediction may still fail on chromosomal aberration or aneuploidy endpoints. Of the studies reviewed here, only Ramesh and Veerappapillai[147] train on micronucleus assay data. Metabokiller offers a complementary strategy, demonstrating that integrating multiple mechanistic properties beyond mutagenicity alone can substantially improve carcinogenicity recall, though this approach addresses a broader endpoint than the Ames-specific models that dominate the field. It is also worth considering whether these models are learning toxicological patterns beyond existing expert knowledge. The genotoxicity field has a long history of rule-based structural alert systems[156,157], and Hung and Gini[154] note that their graph convolutional network's attention mechanisms identify known alerts such as unsaturated carbonyls and aromatic amines, which may be interpreted as validation of the model or as evidence that it is recapitulating established structure-activity rules with greater statistical power rather than discovering novel predictive features. While Chu et al.[150] and Hung and Gini[154] report improvements over commercial tools such as StarDrop and TEST, the margins are often modest, and the practical case for adopting newer ML architectures over established, regulatory-accepted systems remains to be clearly demonstrated.

1.4.5 AI/ML models to predict multi-endpoint toxicity in vitro, in animal models, and in humans

In parallel to models developed for individual preclinical toxicity endpoints, ML-QSAR and deep-QSAR approaches have been applied to construct comprehensive, multi-endpoint prediction models. The Tox21 program, a federal collaboration involving the NIH, the EPA, and the FDA, was established more than a decade ago to support the development of more advanced toxicity assessment methods. As part of the 2014 Tox21 Challenge[158], the consortium published assay activity data, including nuclear receptor responses, cytotoxicity and gene toxicity, and chemical structures on approximately 10,000 compounds and challenged researchers to develop computational models that can predict compounds' interference in biochemical pathways using only chemical structure data. Early approaches applied classical ML algorithms with engineered molecular features to these datasets. Banerjee et al.[159] found that a random forest approach based

on MACCS fingerprints and a subset of 13 molecular descriptors performed best for classification of Tox21 nuclear receptor and stress response pathway assay outcomes, and subsequently released ProTox-II[160], a publicly available web server incorporating molecular similarity, pharmacophore-based features, and fragment propensities into 33 ML models spanning acute toxicity, hepatotoxicity, mutagenicity, cytotoxicity, and immunotoxicity endpoints. Alves et al.[161] took a related approach with Multi-Descriptor Read Across (MuDRA), which uses varying types of chemical descriptors simultaneously for similarity assessment across six endpoints including hERG inhibition, hepatotoxicity, and Ames mutagenicity. These efforts demonstrated that multi-endpoint toxicity prediction was feasible with classical ML, but the endpoint-dependent variability in optimal features and algorithms suggested that a unified approach to diverse toxicity mechanisms would prove elusive.

The introduction of deep learning architectures raised the question of whether increased model complexity could overcome these limitations. A wide range of deep NN structures have been applied to multi-endpoint toxicity prediction, including random NNs[162,163], convolutional NNs[164,165], graph convolutional NNs[166,167], visible NNs[168], spiking NNs[169], and message-passing NNs[170]. The winner of the Tox21 challenge, DeepTox[171], employed a five-stage pipeline comprising data cleaning, heterogeneous descriptor computation, multi-task deep NN training across all 12 endpoints jointly, cluster-based cross-validation to prevent inflated performance from structurally similar compounds, and calibrated ensemble prediction (**Figure 1-3**). However, even within the DeepTox pipeline, deep NNs did not uniformly dominate: support vector machines achieved comparable mean AU-ROC (0.832 vs. 0.837 for deep NNs on the final test set), and the margin of improvement over random forests was modest, suggesting that the advantage of deep learning over classical ML for these endpoints is narrower than the architectural complexity might imply. Subsequent work introduced graph-based and consensus strategies to improve robustness. Tan et al.[172] combined graph transformer architecture with a graph NN to capture both local and global structural information, achieving competitive results across all Tox21 endpoints. Galati et al.[173] developed VenomPred, a consensus framework that averages prediction probability across the top-performing models from 20 candidates built on five molecular fingerprints and four ML algorithms, achieving specificity, recall, and precision ranging from 0.63 to 0.91 across mutagenicity, carcinogenicity, hepatotoxicity, and estrogenic effect endpoints. De Sá et al.[174] developed toxCSM, which evaluated ten ML algorithms across

36 endpoints using toxicophores, graph-based signatures, and similarity-based scores, with feature selection guided by transfer learning, achieving AU-ROC values from 0.68 to 0.99 and generally outperforming DeepTox. A consistent pattern across these studies is that well-characterized, single-mechanism assays such as nuclear receptor binding yielded higher performance than complex, multi-mechanism endpoints, suggesting that the difficulty lies less in the algorithmic approach than in the biological heterogeneity of the endpoints themselves.

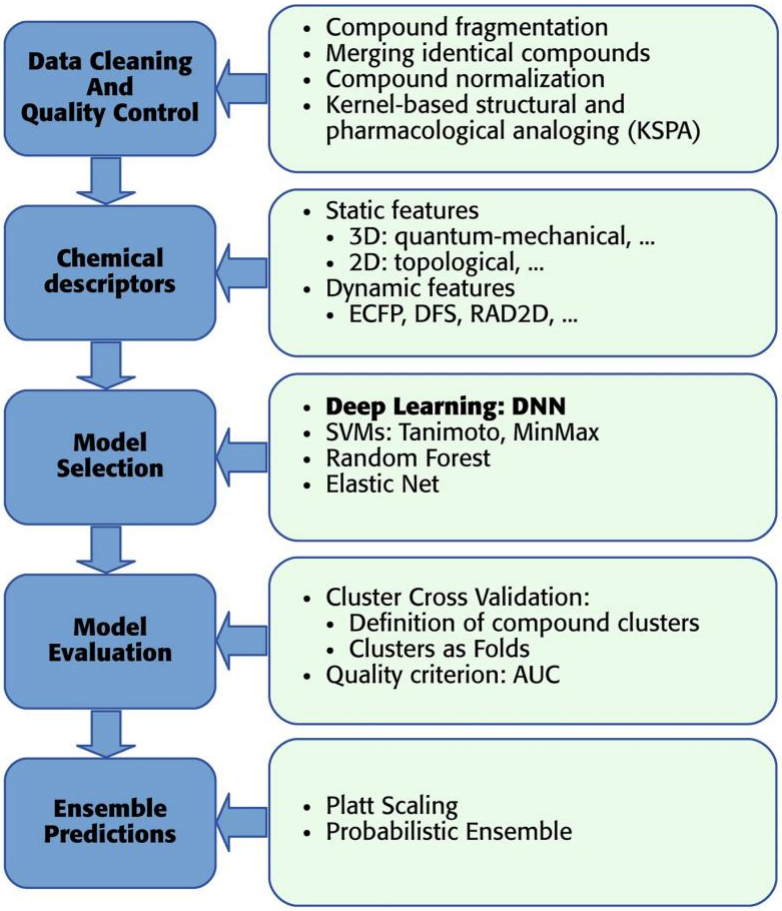


Figure 1-3. The DeepTox pipeline for computational toxicity prediction. The pipeline proceeds through five sequential stages. In data cleaning, raw compound inputs are fragmented, deduplicated, and normalized, with optional augmentation structurally or assay-correlated compounds in public databases. In the chemical descriptor stage, both expert-defined molecular features, including physicochemical properties and 2,500 toxicophore substructure indicators, and algorithmically generated fingerprints are computed. In model selection, multi-task deep neural networks trained jointly across all 12 toxicity endpoints are supplemented by support vector machines, random forests, and elastic nets. In model evaluation, cluster-based cross-validation on ECFP4 Tanimoto similarity ensures structural dissimilarity between training and test folds. In the ensemble stage, per-model outputs are calibrated via Platt scaling and combined through a probabilistic ensemble to produce final toxicity predictions. Reproduced from Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. *Front Environ Sci.* 2016;3:80. https://doi.org/10.3389/fenvs.2015.00080. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

More recent approaches have explored whether multi-task learning and self-supervised pretraining can leverage cross-endpoint information to improve predictions on individual tasks. Yuan Li et al.[175] developed a multi-task model using molecular fingerprints, 2-D and 3-D descriptors, and molecular graphs as inputs to deep NNs and graph convolutional networks, achieving AU-ROC scores of 0.897 and 0.912 on Tox 21 nuclear receptor and stress response outcomes by averaging predicted probabilities across model types. Boonsom et al.[176] constructed a stacked ensemble of 36 models trained with twelve molecular fingerprints to simultaneously predict hERG toxicity, mTOR toxicity, cytotoxicity, and mutagenicity, achieving AU-ROC scores from 0.807 to 0.876. Wu et al.[177] developed MGA, a multi-task graph attention framework that jointly learns regression and classification across 33 toxicity endpoints, generating transferable toxicity features whose utility was confirmed by successful transfer to hepatotoxicity and urinary tract toxicity endpoints not seen during training. Yu et al.[178] took a different approach with GCLmf, a self-supervised contrastive learning graph NN that uses molecular fragments as hard negatives, achieving state-of-the-art performance on 16 of 27 classification benchmarks and four of six regression benchmarks across cardiotoxicity, hepatotoxicity, acute oral toxicity, and Ames mutagenicity. These multi-task and self-supervised frameworks implicitly assume that molecular representations predictive of one toxicity endpoint transfer meaningfully to others. Yet the consistently uneven performance across endpoints, with ion channel binding and receptor activation assays outperforming mechanism-dependent outcomes such as immune-mediated hepatotoxicity or genotoxicity arising from DNA-reactive metabolites, suggests that shared representations may offer limited benefit for mechanistically dissimilar tasks.

The appearance of rapid progress in multi-endpoint toxicity prediction may also be partly an artifact of benchmark fragmentation. When the original Tox21 Challenge was incorporated into widely used benchmark suites such as MoleculeNet[179], alterations to data splits and label imputation rendered cross-study comparisons unreliable. A recent analysis by Ebner et al.[180] reinstated the original 2015 splits and demonstrated that the competition winner DeepTox and self-normalizing NNs from 2017 remained among the top-performing methods, raising the question of whether increasingly complex architectures has yielded meaningful gains for these endpoints.

1.4.6 AI/ML-based prediction of drug toxicity via chemical-target interactions

The prediction of drug-target interactions provides a valuable tool to predict drug toxicity through the inhibition or activation of critical biological targets. Traditional *in vitro* or *in vivo* target specificity and selectivity evaluations are time-consuming and labor-intensive and can suffer from reproducibility issues and limited physiological relevance. Recent AI/ML approaches leverage deep learning and graph-based molecular representations to integrate chemical structure, bioactivity data, and on-/off-target interaction profiles into mechanistically informative toxicity predictions. These models can also capture latent structure-activity relationships and predict interactions for understudied targets where empirical data are sparse.

Chen et al.[181], Yamanishi et al.[182], and Amano et al.[183] developed toxicity prediction models trained on known chemical-protein interactions and side effects, outperforming classical chemical structure-based modeling. Troger et al.[184] combined molecular docking into the pesticide rotenone binding site at respiratory complex I (CI) with pharmacophore-based screening and ML into a workflow which could identify compounds prone to cause mitochondrial toxicity via CI inhibition. Li et al.[185] created a kinome-scale ML prediction framework based on a multi-task deep NN trained on bioactivity data across 391 human kinases and approximately 32,000 chemical compounds. Selectivity scores were determined by setting Kd = 1 M and probability of 50% as the activity threshold for reported and predicted inhibitory activities, respectively. Facilitating network information sharing across different kinases thereby compensating for limited bioactivity data associated with any one kinase, the final model yielded an AU-ROC over 0.90, an F1-score of 0.74, and a balanced accuracy of 0.83. KinomeX, a kinome-wide activity predictor for small molecules based on a multi-task deep NN classification algorithm, was released by Li et al.[186] in 2019. Using ECFP4 fingerprints as 2-D representations of molecular structure and a probability threshold of 50% for active binding, they trained the model on 391 kinases and roughly 32,000 compounds and applied to predict overall selectivity as well as selectivity against a preselected sub-family of kinases for a given small molecule drug. KinomeX achieved an AU-ROC of 0.90 on an internal test dataset, outperforming comparable single task NN models, while also achieving an AU-ROC of 0.86 in a case-study on experimental data, and novel predicted ‘off-target’ activities have been confirmed by experimental data.

In 2024, Ren et al.[187] presented KinomeMETA, an ML framework which integrates a model-agnostic meta-learning algorithm with a graph attention NN-based molecular representation model, trained on more than 160,000 distinct chemical compounds and 661 kinases. To construct the final classifier, bioactivity data were converted to two classes: positive (pKi/pKd/pIC50/pEC50 ≥ 6 or % inhibition ≥ 80) and negative (pKi/pKd/pIC50/pEC50 < 6 or % inhibition < 80). By employing a cluster-wise training strategy to avoid bias from imbalanced kinase groups, KinomeMETA demonstrates MCCs above 0.6 on kinome-wide validation tasks. Specifically, for 412 predictable wild-type human kinases, MCCs of 189 kinases were higher than 0.8, while only 12 reported MCCs less than 0.4. KinomeMETA also demonstrates good performance on identification of novel hit compounds through large-scale virtual screening; for example, they used the framework to select scaffold inhibitors for PKMYT1 and drug compounds highly selective against understudied FGFR mutants and noted good agreement between inhibitory probabilities and experimental inhibitory profiles. Furthermore, even on limited data, KinomeMETA performs well; on 20-shot learning, KinomeMETA achieved an MCC of 0.68 and an AU-ROC of 0.91. Liu et al.[188] applied multi-task graph NN models using the Attentive-FP algorithm for molecular representation and learning to predict interactions between over 600 marketed and withdrawn drugs and a panel of multi-species secondary or off-target proteins corresponding to seven target families, including GPCR, ion channel, enzyme, kinase, transporter, and nuclear receptor groups. Using the off-target prediction model to infer interaction probabilities against 242 targets for any given molecule, they then utilized the resulting multi-dimensional off-target probability vectors to assign drugs according to the ATC system as well as predict a ‘toxic’ or ‘non-toxic’ designation, as defined by labels collected from DrugBank, ChEMBL, and the Clintox[179] dataset. Achieving average AU-ROC scores above 0.70 for toxicity prediction, the predicted off-target profiles serve as structural-biological fingerprints capable not only of clustering drugs by pharmacological class but also mechanistically inferring potential adverse reactions.

The introduction of a mechanistic intermediate, the predicted off-target interaction profile, between chemical structure and toxicity outcome distinguishes this approach from the structure-activity models reviewed in Sections 1.3.1-1.3.5 and provides a layer of biological interpretability that purely structural models lack. This advantage is most apparent for kinase-focused models such as KinomeX[186] and KinomeMETA[187], where the target space is well-

characterized and binding assay data are abundant. However, the approach is inherently constrained by the completeness of the known target space; toxicities arising from mechanisms not represented in the target panel, such as reactive metabolite formation, mitochondrial dysfunction through non-target-mediated pathways, or immune-mediated idiosyncratic reactions, will be invisible to these models. Notably, no study reviewed here directly benchmarks DTI-based toxicity predictions against structure-only models for the same endpoints, leaving the added predictive value of the target interaction layer unquantified.

1.4.7 AI/ML-based prediction of protein and peptide toxicity using in vitro and animal systems

Peptide-based drug development has been active for over half a century, and *in silico* techniques have become increasingly common for the identification and characterization of the therapeutic and toxicological properties based on amino acid sequences of proteins and peptides. Compared to small molecules, their high specificity makes peptides attractive candidates for treatment from cancer and cardiovascular diseases to irritable bowel syndrome and respiratory distress[189,190]. However, therapeutic peptides also demonstrate weak membrane permeability and poor *in vivo* stability, as well as clinical toxicities such as intestinal wall disruption, erythrocyte and lymphocyte toxicity, free radical production, enzymopathic and immunopathic tissue damage, and cytotoxicity[191]. The wealth of structure- and sequenced-based information now available in the postgenomic era necessitates the development of highly flexible and robust AI/ML protein and peptide toxicity prediction models. Mall et al.[192] developed VISH-Pred, an ensemble framework that combines multiple fine-tuned ESM2 language models with traditional classifiers such as LightGBM and XGBoost to predict the toxicity of protein and peptides, addressing limitations of previous methods like CSM-Toxin which required an impractical optimal threshold of 0.968 due to severe class imbalance issues. To handle the massive class imbalance in their training data (91:1 ratio), they applied an under-sampling technique by dividing the dataset into smaller, balanced groups, enabling the use of the standard 0.5 classification threshold. Their approach achieved excellent results with an F1-score of 0.759 and an MCC of 0.737, outperforming existing methods by over 10% including CSM-Toxin, while only requiring the protein's primary sequences as input.

Over the last decade, a multitude of web-based applications for the prediction of protein and peptide toxicity have been made accessible. In 2013, Gupta et al.[193] proposed ToxinPred, a support vector machine-based method which utilizes features such as amino acid composition,

dipeptide composition, and the identification of toxic motifs or regions derived from sequence information to predict the toxicity of peptides and proteins. Sharma et al.[194] developed ToxinPred2, designed for large protein sequences. The tool employs a combined methodology that merges BLAST similarity searches, MERCI motif identification, and ML algorithms utilizing compositional characteristics. This extends the functionality beyond ToxinPred, which was optimized for shorter peptide sequences. The approach discovered toxicity-associated motifs (GCYCG, MKTLL, TLLLTL, LLLTLV). It also determined that cysteine, glycine, lysine, and tryptophan residues are predominantly found in toxic proteins. It also demonstrated excellent predictive capability, with AU-ROC of 0.99 and MCC of 0.91 on the most comprehensive protein toxicity dataset available. ToxinPred3[195] was released in 2023, now updated to leverage artificial NNs and long short-term memory networks with fixed sequence length instead of traditional ML algorithms for prediction, and achieves an AU-ROC of 0.93 on independent data. Brandes et al.[196] created ProteinBERT, a natural language processing model which treats amino acids as words and protein sequences as sentences and using an attention mechanism, can capture complex correlations between distant residues. Morozov et al.[197] developed CSM-Toxin using a transformer deep learning model that treats protein sequences like sentences, achieving an MCC of 0.67 and outperforming existing methods like ToxinPred2 (MCC: 0.46). The approach requires only raw amino acid sequences and automatically learns toxicity patterns without manual feature engineering, representing a shift from traditional ML to modern deep learning for protein toxicity prediction. In April 2025, Zhu et al.[198] constructed ToxDL2.0 (**Figure 1-4**), a multi-modal deep learning model that integrates structural, evolutionary, and functional domain information for protein toxicity prediction. The model constructs residue-level protein graphs from AlphaFold2-predicted 3-D structures, with node features initialized from the ESM-2 protein language model[199] and processes them through a graph convolutional network to produce structural embeddings. A separate domain embedding module encodes InterPro-annotated functional domains, and the two representations are concatenated and classified by a dense network. Trained on a dataset of 4,900 toxic and 9,800 non-toxic proteins derived from UniProt, ToxDL2.0 achieved an AU-ROC of 0.945 and outperformed all existing methods including ToxinPred2, CSM-Toxin, and its predecessor ToxDL on an independent holdout set of proteins deposited after the training data cutoff. Ablation analysis confirmed that the graph

convolutional module was the dominant contributor to performance, with its removal reducing AU-PRC by over 25% on independent data.

Deep learning-based methods including TOXIFY[200], ToxDL[201], ATSE[202], ToxIBTL[203], and VISH-Pred offer end-to-end solutions that eliminate the requirement for manual feature extraction by learning directly from protein sequences. ToxGin[204] integrates 3-D peptide structures with sequence data through graph isomorphism networks, achieving an AU-ROC of 0.91 on external data. Moukheiber et al.[205] took a different approach, predicting compound toxicity through protein-compound interactions rather than chemical descriptors. Using random forest with SMOTE+ENN resampling on the Tox21 dataset, they achieved AU-ROC values up to 0.92 and identified interconnected biological pathways contributing to toxicity, including DNA damage response, NF-B signaling, and MAPK activation. Wei et al.[203] developed ToxIBTL, which combines convolutional and recurrent architectures with information bottleneck optimization and transfer learning, achieving 96.0% accuracy and outperforming existing methods by 3-64.5% in F1-score across peptides of varying lengths.

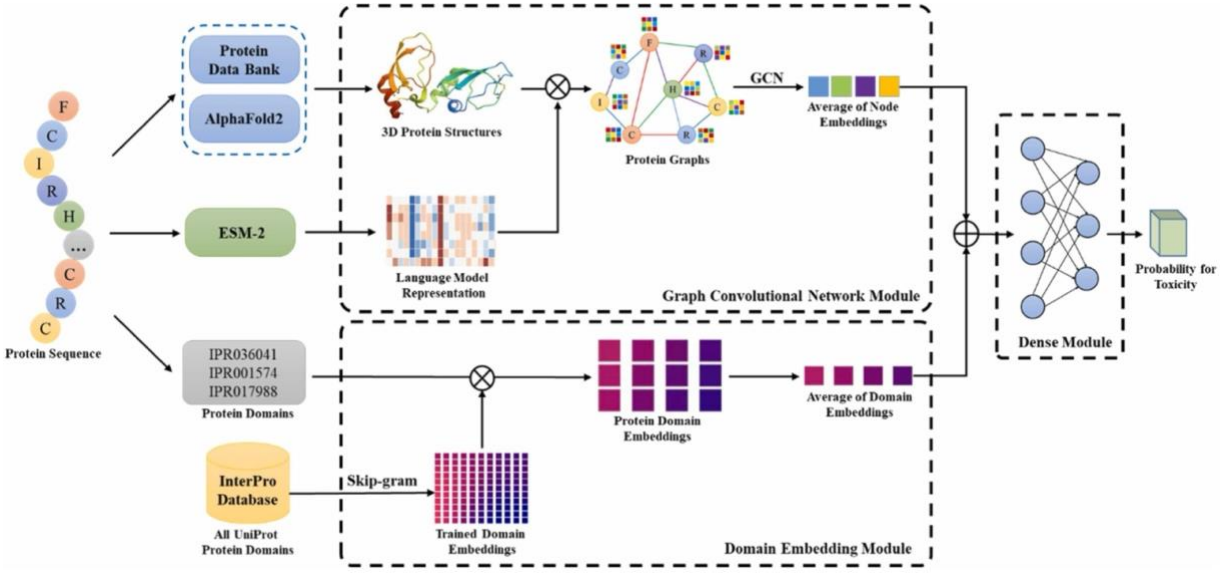


Figure 1-4. Architecture of ToxDL2.0 for multi-modal protein toxicity prediction. The model comprises three modules: a graph convolutional network module that constructs residue-level protein graphs from AlphaFold2-predicted 3-D structures with node features initialized from ESM-2 protein language model embeddings; a domain embedding module that encodes functional protein domains identified through InterPro annotation; and a dense module that concatenates both embeddings and applies a multilayer perceptron to output a toxicity probability. Reprinted from *Computational and Structural Biotechnology Journal*, Vol. 27, Zhu L, Fang Y, Liu S, Shen H-B, De Neve W, Pan X, ToxDL 2.0: Protein toxicity prediction using a pretrained language model and graph neural networks, pp. 1538–1549, Copyright (2025), with permission from Elsevier.

The rapid proliferation of protein and peptide toxicity prediction tools from support vector machine-based methods like ToxinPred to transformer architectures like CSM-Toxin and graph-based frameworks like ToxGin, reflects both the growing importance of biologic therapeutics and the availability of curated sequence-toxicity datasets. A recurring theme is the severity of class imbalance: toxic proteins represent a small fraction of known sequences, requiring techniques such as under-sampling and SMOTE resampling[205] that may introduce their own biases. More fundamentally, most models in this area predict a binary toxic/non-toxic classification from primary sequence alone, without distinguishing among the diverse mechanisms by which peptides and proteins cause toxicity, such as cytolytic membrane disruption, enzymatic activity, receptor-mediating signaling, or immunogenicity. As the therapeutic peptide pipeline expands to include increasingly complex modalities such as cyclic peptides, stapled peptides, and peptide-drug conjugates, the relevance of models trained predominantly on naturally occurring toxins and linear peptide sequences to these engineered molecules remains to be established.

1.5 AI/ML Models for Predicting Clinical Drug Toxicity

Predicting drug toxicity from molecular structure and preclinical data addresses whether a compound has the potential to cause harm. Translating these predictions to clinical safety outcomes introduces fundamentally different challenges: the data are more heterogeneous, the endpoints are more complex, and the pharmacological context of the patient, including dose, concomitant medications, organ function, and genetic background, becomes a critical determinant of whether a molecular liability manifests as a clinical AE. AI/ML approaches to clinical toxicity prediction span a range of tasks, from predicting interactions between co-administered drugs to forecasting patient-level AEs for individual therapies, across drug classes, and from chemical structure and systems pharmacology data. The choice and availability of high-quality patient data is an ongoing barrier, as is the explainability and trustworthiness of model development and validation processes.

1.5.1 Drug-drug interactions

DDIs occur when the pharmacological effects of one drug compound are altered by another drug compound in a combination therapy regimen. The mechanism by which this interaction occurs can be highly complex and either PK or PD in nature. However, while best

practices for PK examination of DDIs have been well-documented[206,207], PD DDI studies in preclinical drug development are largely limited to high-throughput *in vitro* screening studies. Due to the absence of well-defined examination into PD DDIs at the *in vivo* systems-level, the translational impact of PD DDIs identified early in the drug development process remains ambiguous[208]. Despite the importance of better understanding of DDIs in aging societies where older patients are more likely to be prescribed multiple drugs for multiple indications, clinical examination of DDIs is considered highly challenging due to the high cost and time associated with evaluating a substantial number of drug combinations. To address these issues, *in silico* modeling approaches including AI/ML have been implemented to predict DDI effects in patients using data from diverse sources.

The first ML framework, DeepDDI, applied to the prediction of DDIs was constructed by Ryu et al.[209] in 2018. The deep NN structure accepts SMILES strings as structural information along with a pair of drug names, and from these inputs can predict multiple DDI types for each drug-pair and provide suggestions for alternative drug compounds which produce similar pharmacological effects without toxicity. DeepDDI, along with a large majority of AI/ML models for DDI prediction, were developed using the DrugBank[210] dataset, which was released in 2006 and as of its last update in 2018 now contains information on 1,413,413 DDIs from 6,231 drug compounds based on *in vitro* and *in vivo* data. Lee et al.[211] employed a similar approach by constructing an autoencoder to input compound structural similarity profiles, gene ontology term similarity profiles, and target gene similarity profiles of known drug pairs into a multilayer perceptron to predict the multilabel classification for 106 DDI types in human patients derived from *in vivo* and *in vitro* studies. Their model demonstrated improved performance when all three similarity profiles were combined and performed better than existing methods such as support vector machines, random forests, and the principal component analysis-based DeepDDI. Deng et al.[212] developed DDIMDL, assembling features from the drug using chemical substructures, targets associated, pathways, and enzymes. They reduce feature dimensionality using similarity matrices and train a deep NN which combines the sub-models to learn from cross-modality representations of drug-drug pairs and predict DDI events in patients with an AU-ROC of 0.921. Luo et al.[213] created a deep learning model based on drug-induced transcriptome data. Their model uses a graph convolutional autoencoder network to create an embedding from the transcriptome data, then a long short-term memory network to predict DDIs, achieving F1,

precision, and recall scores over 0.94. Kim and Nam[214] proposed DeSIDE-DDI, a deep learning-based framework for predicting DDIs based on drug-induced gene expression data *in vitro*. A feature generation model is used to predict gene expression signatures from molecular fingerprints and compound properties, and the model takes the features and uses gated linear units to then predict the probability of an interaction. DeSIDE-DDI outperformed existing methods and achieved AU-ROC scores of 0.889, 0.640 and 0.553 on unseen, one-unseen, and both-unseen data. Consensus ML models without utilizing deep learning have also shown potential. In 2022, Wang et al.[215] developed a consensus ML model to predict DDIs related to five important CYP450 enzymes for drugs that are inhibitors and non-inhibitors. They compared random forest and XGBoost models, testing four different molecular fingerprints separately; the models were then used in a consensus ML approach, achieving an accuracy of 0.85 on all five enzymes.

Given the complexity of large DDI datasets, DDI prediction tasks are particularly well-suited for more sophisticated input feature representations. Briand et al.[216] combined molecular docking with linear discriminant analysis, quadratic discriminant analysis, and multilayer perceptrons to identify compounds susceptible to DDIs through inhibitory effects on human metabolite CES1. Graph NN-based approaches have become the dominant paradigm for DDI prediction, with frameworks such as DPDDI[217], GCN-BMP[218], MUFFIN[219], GMPNN-CS[220], and MFFGNN[221] all learning from graph-structured molecular data and employing multilayer perceptrons for binary or multi-label DDI classification. These models achieve AU-ROC values generally exceeding 0.9, with incremental gains coming from increasingly sophisticated feature fusion strategies: MFFGNN[221] combines topological information from molecular graphs with chemical context from SMILES sequences, while GMPNN-CS[220] uses gated message-passing to extract features from graph representations. Wang et al.[222] published KnowDDI in 2024, which enhances drug representation by leveraging neighborhood information from biomedical knowledge graphs and learning individual knowledge subgraphs for each drug pair to provide interpretable DDI predictions. Chen et al.[223] developed MSKG-DDI (**Figure 1-5**), which encodes each drug through two parallel modules: a knowledge-embedded message-passing NN that learns structural representations from molecular graphs, and a graph NN that aggregates neighborhood information from a drug knowledge graph encoding protein targets, chemical substructures, and enzyme relationships. The two representations are fused through a

transformer-based multi-head attention layer to produce drug-pair DDI predictions. On DrugBank, the full model achieved an AU-ROC of 0.997 and outperformed all baselines including MUFFIN and DeepDDI across binary, multi-class, and multi-label prediction tasks. Ablation analysis confirmed that both components contribute complementary information, and the molecular graph module enables predictions for novel drugs absent from the training knowledge graph, partially addressing the cold-start problem that limits purely knowledge graph-based approaches.

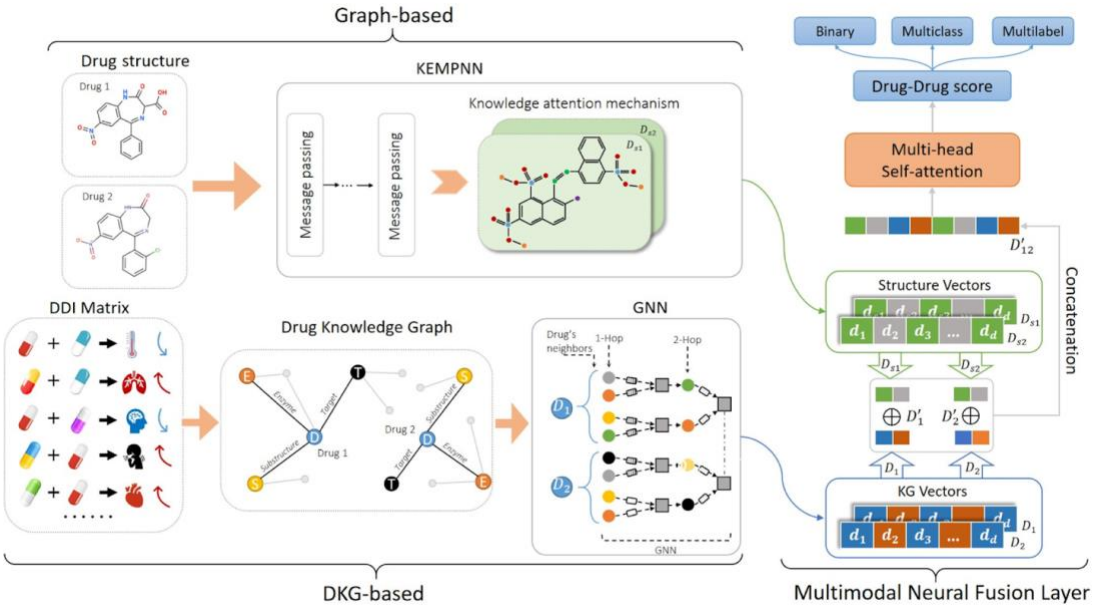


Figure 1-5. Architecture of MSKG-DDI for drug-drug interaction prediction. The framework consists of two parallel representation modules and a multi-modal fusion layer. In the graph-based module (upper pathway), SMILES strings are converted to molecular graphs and processed through knowledge-embedded message-passing neural networks to produce structural drug representations. In the drug knowledge graph-based module (lower pathway), a knowledge graph encoding relationships between drugs, protein targets, chemical substructures, and enzymes is processed by a graph neural network to produce knowledge-based drug representations. In the multi-modal fusion layer, the structural and knowledge graph representations for each drug are concatenated and passed through a multi-head self-attention transformer black to produce drug-pair interaction predictions. Reprinted from *Computers in Biology and Medicine*, Vol. 169, Chen S, Semenov I, Zhang F, Yang Y, Geng J, Feng X, Meng Q, Lei K, An effective framework for predicting drug–drug interactions based on molecular substructures and knowledge graph neural network, 107900, Copyright (2024), with permission from Elsevier.

Across the DDI prediction literature, a notable trend is the rapid escalation of model complexity, from the initial deep NN framework of DeepDDI[209] to increasingly elaborate architectures incorporating knowledge graphs, transcriptomic data, molecular docking, and multi-modal fusion layers. DrugBank has served as the dominant data source, and most models are evaluated on DDI-type classification tasks derived from this database. While reported AU-ROC values are often high, several concerns temper these results. The DDI prediction task as

typically formulated is fundamentally different from the toxicity endpoints discussed in Sections 1.3 and the remainder of 1.4, as it involves predicting whether a known pharmacological interaction exists between two compounds rather than predicting a *de novo* toxicological outcome. Performance drops substantially when models are challenged with drug pairs where one or both compounds are absent from the training set, as demonstrated by the declining AU-ROC of DeSIDE-DDI from 0.889 on seen pairs to 0.553 on both-unseen pairs[214]. This suggests that current models may be learning drug-specific memorization patterns rather than generalizable interaction rules. Moreover, the clinical significance of a DDI is strongly dependent on the doses and systemic exposures at which both drugs are co-administered, yet no model reviewed here incorporates dosing information as a predictive feature. Furthermore, the overwhelming reliance on DrugBank, which catalogs known interactions largely from literature and labeling rather than from systematic clinical pharmacology studies, raises questions about the completeness and reliability of negative labels, as the absence of a documented interaction does not confirm the absence of a clinically meaningful one. The field would benefit from greater integration of mechanistic PK and PD data, validation against prospectively identified clinical DDIs, and clearer stratification of interaction types by clinical significance rather than broad categorical labels.

1.5.2 Patient-level adverse events for individual therapeutic agents

Predicting whether an individual patient will experience a toxic outcome from a specific therapeutic agent requires a fundamentally different modeling approach from predicting whether a compound has the potential to cause harm. Rather than learning from chemical structure alone, these models integrate patient-specific clinical covariates, including demographics, laboratory values, imaging features, and PK exposure, with treatment information to predict individual risk. ML models trained on 45 distinct pre-implant features, including demographics, physical characteristics, heart failure data, LV assessment data, ECG data, comorbid conditions, surgical interventions, and medication classes, aimed to predict all-cause mortality or heart failure hospitalization at twelve months post cardiac resynchronization therapy for 595 patients in the COMPANION trial[224]. A random forest classification algorithm with 550 trees produced the best performance with an AU-ROC of 0.74 (95% CI: 0.72, 0.76) and identified prior history of renal disease, time from heart failure diagnosis, New York Heart Association class, QRS duration, and history of intermittent atrial arrhythmia as the most influential dataset features. The model

demonstrated superior risk stratification compared to current clinical guidelines, creating patient quartiles with an 8-fold difference in survival between the highest and lowest risk groups, substantially outperforming the 1.5-fold difference achieved by traditional bundle branch block morphology and QRS duration classifications. The authors emphasized that such models could be integrated into electronic health records to automatically calculate predicted event probabilities during pre-implant consultations, facilitating improved shared decision-making between clinicians and patients. Modongo et al.[225] employed a similar approach, first fitting a two-compartment PK model to plasma concentration levels of amikacin and then a Classification and Regression Tree analysis to identify clinical factors predictive of ototoxicity in 28 multidrug-resistant pulmonary tuberculosis patients on amikacin. Their final model identified cumulative days of therapy and cumulative AUC as primary predictors of drug-induced ototoxicity and achieved an AU-ROC of 0.97 on unseen data. Considerable effort has been applied to the prediction of individual chemotherapy outcomes for therapies such as FOLFIRINOX[226], the IVA protocol (Ifosfamide, Vincristine, Actinomycin D) for rhabdomyosarcoma[227], and oxaliplatin and/or fluorouracil-based regimens[228,229], also using PK, dose, and laboratory data such as baseline neutrophil and platelet counts as model inputs. Cuplov and André[227] used a gradient boosting regression approach to predict the likelihood of neutropenia and thrombocytopenia in 24 pediatric rhabdomyosarcoma patients undergoing IVA chemotherapy. Input features included baseline neutrophil and platelet counts, weight-normalized dosing of ifosfamide, and time since treatment cycle start, as well as dynamic blood cell profiles. Gradient boosting outperformed lasso and ridge regression models, achieving MAEs between predicted and observed neutrophil and platelet counts of 1.0 and 72.8 G/L, respectively. They developed a web-based application to provide cycle-specific predictions, enabling dose adjustment recommendations to minimize severe hematological toxicities.

More recently, deep learning algorithms have been applied to clinical trial data to analyze patient safety outcomes such as adverse event rate and perform prediction tasks more complex than classification; in 2024, Wysokiński and Dreczka[230] designed a deep NN to predict clozapine and norclozapine concentrations, toxicity, and therapeutic dose range using demographic and clinical data from 69 schizophrenia patients. Their final model facilitates therapeutic drug monitoring using only readily available clinical information and achieves an RMSE of 0.6276 for norclozapine concentrations and an RMSE of 0.8523 for clozapine concentrations. The model

integrated readily available clinical parameters including demographic data (sex, age, BMI), clozapine dosing, inflammatory markers (C-reactive protein), and pharmacogenetic factors (numbers of CYP1A2 and CYP3A4 substrates, inhibitors, and inducers), addressing the clinical need for therapeutic drug monitoring in patients receiving high-risk antipsychotic medications. The deep learning architecture employed separate models for each metabolite using an 84-layer network for clozapine concentration prediction and a 64-layer network for norclozapine concentration prediction. Patients were then classified into two groups based on their predicted clozapine or norclozapine concentrations, normal or toxic, where toxic was defined as clozapine levels > 550 ng/mL or norclozapine levels > 270 ng/mL. The ML model achieved good performance in toxicity classification, with 99% accuracy for identifying patients at risk for toxic concentration levels of clozapine and 93% accuracy for toxic concentration levels of norclozapine; it can also provide personalized therapeutic dose range recommendations through a freely available web application, demonstrating successful clinical translation of deep learning models for real-world therapeutic drug monitoring applications.

In January 2025, Yang et al.[231] created a deep learning model to predict symptomatic pneumonitis in 261 lung cancer patients undergoing thoracic radiotherapy in combination with immunotherapy trained on depth image features extracted from deep learning combined with radiomics and clinical characteristics (**Figure 1-6**). The methodology integrated 107 traditional radiomic features extracted from bilateral lung parenchyma using PyRadiomics[232], deep image features using ResNet34 architecture with transfer learning, and 12 clinical parameters including ECOG performance status, T-stage, and inflammatory markers (NLR, PLR, LMR, SII). They demonstrated that the final model which utilizes an integrated and dimensionally reduced combination of clinical data, traditional radiomic features, and deep image features and connects them to a deep NN-based prediction layer yields the highest prediction accuracy (AU-ROC = 0.922, sensitivity = 0.879, specificity = 0.789), significantly outperforming both traditional ML algorithms and deep learning model runs which relied solely on clinical data. Clinically, 36.02% of patients developed Grade 2+ symptomatic pneumonitis, with the model identifying ECOG performance status (HR 3.322, $p = 0.027$) and T-stage (HR 1.501, $p = 0.003$) as independent predictors of toxicity occurrence.

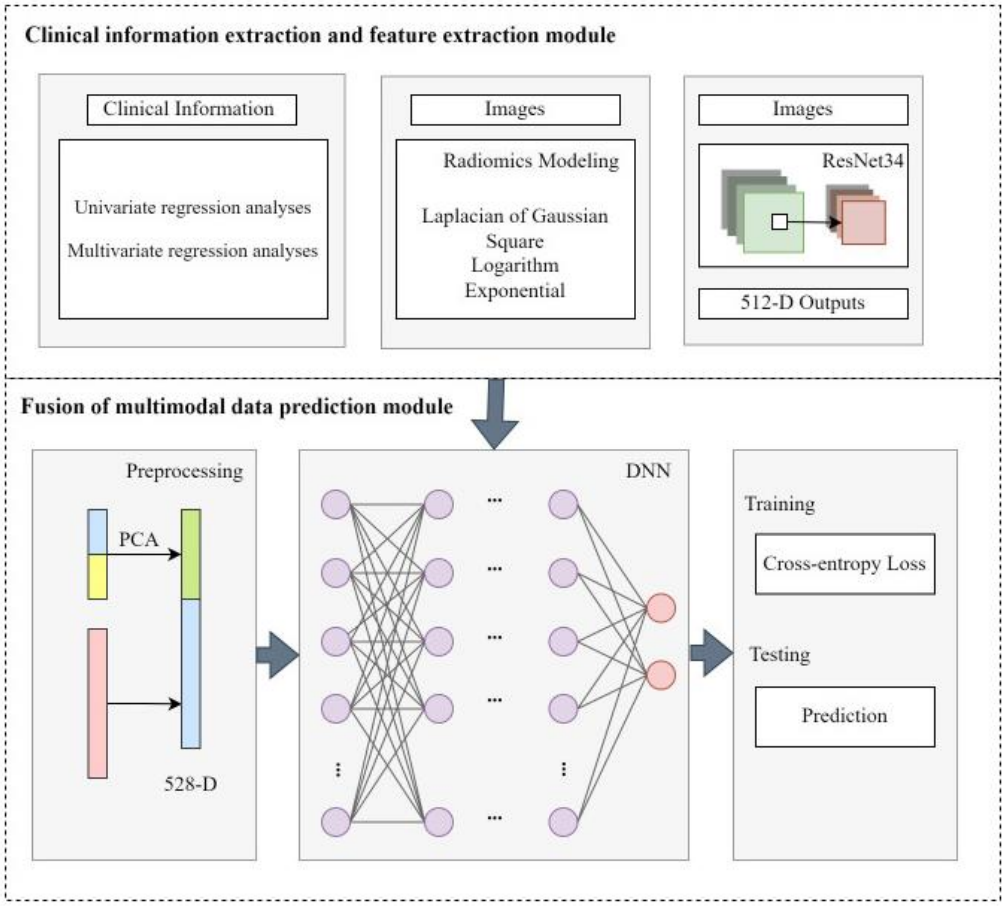


Figure 1-6. Multi-modal deep learning pipeline for predicting symptomatic pneumonitis in lung cancer patients undergoing combined thoracic and radiotherapy and immunotherapy. The framework integrates three parallel feature streams: clinical baseline characteristics including performance status, tumor stage, dosimetric parameters, and inflammatory indices; handcrafted radiomic features extracted from pre-radiotherapy CT images of the bilateral lung parenchyma; and deep image features produced by a pre-trained ResNet34 encoder applied to the same CT images. All three feature sets are independently normalized and dimensionality-reduced, then concatenated and passed through a fully connected deep neural network for binary classification. Reproduced from Yang M, Ma J, Zhang C, Zhang L, Xu J, Liu S, et al. Multimodal data deep learning method for predicting symptomatic pneumonitis caused by lung cancer radiotherapy combined with immunotherapy. *Front Immunol.* 2024;15:1492399. https://doi.org/10.3389/fimmu.2024.1492399. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A distinguishing feature of these models is their integration of patient-specific covariates, including demographics, comorbidities, laboratory values, and PK exposure. However, the sample sizes in this area are small by ML standards, with most studies comprising fewer than 300 patients, raising concerns about statistical power and the risk of overfitting to idiosyncratic features of individual datasets. The high reported accuracies in some studies, such as 99% toxicity classification accuracy for clozapine in Wysokiński and Dreczka[230], should be interpreted cautiously given these small cohorts and the absence of multi-site external validation.

1.5.3 Pharmacogenomic predictors of drug-induced toxicity

With the dissemination of new genome sequencing technologies, particularly the advent of next-generation sequencing (NGS)[233], large-scale pharmacogenomic studies are now feasible in laboratory and clinical settings. The most commonly inherited genetic variants include single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs), copy number variations (CNVs), and a varying number of tandem repeats[234,235]; those involved in either processes or target mechanisms of action and post-target signaling can lead to drug toxicity. The identification of genetic variants associated with differential patient response to drugs has greatly expanded, yet conventional genomewide association study methods remain limited in their ability to capture complex, nonlinear, or indirect linkages in multi-dimensional association networks of SNPs and toxicity data. ML algorithms are particularly well-suited to the high-dimensional analysis required for pharmacogenomic studies and have been implemented successfully to unravel polygenic mechanisms of drug-induced toxicity.

In 2021, Bose et al.[236] applied a maximum relevance minimum redundancy algorithm to an analysis of the effect of antiretroviral therapy on apoptosis pathway genes for 16 patients and 16 healthy volunteers. Their final model identified two upregulated proapoptotic genes, DFFA and TNFRSF1A, as the most important predictors of toxicity and achieved a prediction accuracy of 0.86. Nielsen et al.[237] integrated pediatric patient information on age, sex, and SNPs into ML models including regression, random forest, AdaBoost, and artificial NNs to predict risk of asparaginase-associated pancreatitis in acute lymphoblastic leukemia. Their clinical baseline model (only age and sex included) achieved an AU-ROC of 0.62, improving to 0.67 with genetic variants and further to 0.80 with 30 AAP-associated SNPs. The most predictive features included rs10273639 (PRSS1-PRSS2), rs10436957 (CTRC), and age (1-7 years). An ensemble model achieved AU-ROC of 0.83 with a sensitivity of 0.62 and 37% positive predictive value at 0.8 prediction threshold. Models were validated on independent European (AU-ROC: 0.84) and non-European (AU-ROC: 0.72) cohorts. Notably, integrating clinical features like asparaginase dosing parameters boosted performance to AU-ROC of 0.86, demonstrating the clinical utility of combined genomic-clinical models for personalized toxicity prediction.

Guo et al.[238] constructed a dataset of 24 SNPs in the PI3K/AKT pathway from 259 locally advanced cervical cancer patients treated with neoadjuvant chemotherapy and employed a random forest model combined with a Mean Decrease in Impurity approach to identify the

most predictive genotypic model features of toxicity. Guo et al. used one-hot encoding to transform the 24 SNPs into 70 genotype features, comparing chemotherapy toxicity grades 1-2 versus grade 3 across neurological, gastrointestinal, and hematological toxicity types. Specifically, gastrointestinal toxicity was primarily linked to the GA genotype of Akt2 rs4558508, TC genotype of PTEN rs17431184, and GA genotype of Akt1 rs1130233, while hematological toxicity showed strongest associations with heterozygous AG in Akt2 rs7259541, CT genotype of Akt1 rs2494739, and CC genotype of PTEN rs926091. Notably, Akt2 rs7259541 ranked as the most important feature for both hematological and neurological toxicity prediction in the Mean Decrease in Impurity analysis, demonstrating the clinical utility of pathway-specific genetic profiling for personalized platinum-based chemotherapy toxicity risk assessment. Their analysis associated a higher risk of neurological toxicity in patients with homozygous AA in Akt2 rs7259541 than in those with AG or GG genotypes, along with additional sets of loci attributed to elevated risk of gastrointestinal and hematological toxicities.

In 2024, Arab et al.[239] used a combination of NNs and GANs, called BRI-Net, to identify genetic markers associated with cisplatin-induced ototoxicity in a cohort of 362 pediatric cancer patients. The methodology integrated biological domain knowledge through hierarchical network architecture (SNPs → genes → pathways → outcomes) and employed adversarial training to remove confounding bias from dose intensity, learning unbiased data representations to enhance causal inference. Following comprehensive feature engineering that reduced 6.7 million SNPs to 8,641 with established biological connections, the model underwent rigorous statistical validation using Randomized Block Design with McNemar's test. Two biomarkers within the CERS6 and TLR4 genes were associated with cisplatin-induced AEs: CERS6 rs13022792 as a risk factor (p = 3 x 10-4, OR = 2.37, 37% vs 26% allele frequency in cases vs controls) and TLR4 rs10759932 as a protective factor (p = 7 x 10-4, OR = 0.36, 9% vs 19% allele frequency in cases vs controls). Clinically, 63.8% of patients developed ototoxicity overall, with risk stratification showing 72% incidence in CERS6 variant carriers versus 55% in non-carriers, and 46% incidence in TLR4 variant carriers versus 70% in non-carriers. Further *in silico* analysis suggests that ML-identified SNPs may also affect molecular mechanisms implicated in cisplatin-induced hearing loss.

Incorporating SNP data into ML frameworks can meaningfully improve predictive performance beyond clinical baseline models, as illustrated by Nielson et al.[237], where AU-ROC

improved from 0.62 with clinical features alone to 0.83 with SNP integration. However, most studies in this area use candidate-gene or pathway-focused SNP selection strategies, as in Guo et al.[238] with 24 SNPs in the PI3K/AKT pathway, which may miss polygenic contributions outside the selected pathways. Cohort sizes are also small, and the populations studied are predominantly of limited ethnic diversity, which may reduce the generalizability of identified genetic associations to broader patient populations.

1.5.4 Adverse event prediction across pharmacologically related drugs

AI/ML models trained on large patient-level datasets which capture information on multiple drugs within the same activity or structure class have great potential to uncover new mechanism- and exposure-based associations to AEs while generating powerful personalized safety outcomes for patients. Because high quality clinical trial data is time-consuming and costly to generate, most such analyses use retrospective cohort studies for patients exposed to any medication within a certain class or on any medication at all to identify which reported AEs can be attributed to their drug regimen and further analyze any associated risk factors[240-249]. Jamshidi et al.[250] developed the first multi-vaccine comparative framework for personalized AE prediction analyzing 19,943 participants across six COVID-19 vaccines using a novel two-stage modeling approach, first integrating baseline demographic and medical history features for first dose predictions, then incorporating first dose AEs as additional predictors for enhanced second dose modeling, for a total of 54 input parameters. Their dataset included patients who received either the AZD1222, Sputnik V, BBiBP-CorV, COVAXIN, BNT162b2, or mRNA-1273 vaccines. Jamshidi et al. achieved predictive performance with AU-ROCs ranging from 0.620-0.727 for first dose and 0.777-0.906 for second dose predictions. The best performing model uses a logistic regression algorithm to predict the possibility of fever, fatigue, headache, nausea, chills, joint pain, muscle pain, and local side effects after first and second vaccine doses, and identifies age, history of COVID-19 infection, and history of cancer as input factors with a substantial impact on prediction performance. Notably, the inclusion of first dose side effects significantly improved second dose predictions.

In 2022, Gong et al.[251] at the FDA constructed a random survival forest model trained on exposure and safety data from 4,638 patients in the registrational clinical trials of 16 small molecule kinase inhibitors as well as the dissociation constants of those 16 SMKIs against 442 wild-type and mutated human protein kinases. The methodology integrated individualized kinase

inhibition profiles that normalized steady-state drug exposure with molecular binding affinity for each patient-kinase pair. This is the first ML time-to-event analysis to leverage patient-level PK exposure to prospectively associate on- and off-target binding with individual adverse event prediction for 16 compounds within the same drug class. Their final models predict time to AE occurrence at the patient-level with an average predictive accuracy of 0.701, achieving exceptional performance for severe grade 4 life-threatening events (C-index: 0.864-0.920) and demonstrate strong predictive capability across diverse AE types including dermatitis acneiform (0.852) and hypertension (0.782). The models identify correlations between on- and off-target kinase binding and AEs through variable importance analysis, validating well-established clinical associations such as VEGFR-hypertension and EGFR-diarrhea while discovering novel mechanistic insights, with external validation on independent datasets confirming generalizability (R^2 = 0.49-0.92, $p < 0.05$). In 2026, Jusko et al.[252] advanced this approach by incorporating on-/off-target driven tissue selectivity, demonstrating improvements in the prediction of AEs associated with specific organ systems as compared to prediction with systemic PK exposure alone. The final RSF models improve the prediction accuracy for rash, nausea, fatigue, and headache events, achieving C-indices of 0.814, 0.772, 0.799, and 0.669 respectively, and identified novel targets in organs associated with rash and fatigue. The improved ML model was validated using external data from a neratinib monotherapy trial and achieved sensitivities and specificities greater than 0.95 for the prediction of clinical AEs.

Gong et al.[251] and Jusko et al.[252] demonstrate that training across multiple drugs within a pharmacological class can disentangle target-mediated toxicity from drug-specific confounders, and their integration of patient-level PK exposure with molecular binding affinity produces individualized kinase inhibition profiles that connect mechanism to clinical outcome. This approach is enabled by the extensive kinome-wide profiling infrastructure available for kinase inhibitors; whether analogous class-level models can be constructed for drug classes with less systematically characterized target landscapes, such as immunomodulators or CNS-active agents, remains an open question. Additionally, the reliance on registrational clinical trial data, while ensuring data quality, means that these models are trained on populations subject to restrictive inclusion and exclusion criteria that typically exclude patients with significant comorbidities, organ impairment, extremes of age, or complex polypharmacy regimens. Toxicities that emerge preferentially in these subpopulations may therefore be underrepresented in the training data,

potentially limiting model generalizability to the broader patient populations encountered in post-market clinical practice.

1.5.5 Adverse event prediction from chemical structure and systems pharmacology

A complementary approach to patient-level clinical prediction is to forecast AEs directly from a drug's chemical structure and its interactions with the biological systems it engages. By incorporating off-target binding profiles, protein interaction networks, and structured relationships between drugs, targets, indications, and clinical outcomes, these models attempt to bridge the gap between molecular properties and clinical safety without requiring patient-level data. Bender et al.[253] constructed a series of naive Bayes models to classify chemical structures by their binding activity to preclinical safety targets as well as their clinical adverse event rate as early as 2007. By combining both models, they link on- and off-target binding predictions to adverse event predictions and report good agreement between predicted target-AE pairs and experimentally validated associations in the literature. Over the next two decades, a wave of logistic regression, Bayesian, support vector machines, decision trees, and NN classification algorithms were developed on datasets expanded to include protein-protein interaction networks and gene ontology annotations, as well as other biologic and phenotypic features[254-264]. Much of the research in this area relies on public databases such as the Side Effect Resource (SIDER) database[265], which was released in 2010 and maintained until 2015; it currently contains labeling data on 1,430 drugs, 5,868 clinical AEs, and 139,756 drug-AE pairs. Galeano et al.[266] report a matrix decomposition algorithm which learns a small set of latent biological activity features to predict the categorical frequencies of unknown side effects for drugs with a small number of determined side effect frequencies trained on 759 drugs and 994 side effects, sourced in part from SIDER v4.1. In 2023, Sharma et al.[267] compared multi-task deep NNs and single task deep NNs, trained on Morgan fingerprints and SMILES of chemical structures. The multi-task deep NNs, which aims to predict *in vivo*, *in vitro*, and clinical toxicity, successfully leverages transfer learning and the SMILES embedding to predict clinical outcomes, with results comparable to state-of-the-art models in AU-ROC and balanced accuracy. Their results demonstrate a minimal need for *in vivo* data for clinical toxicity predictions and use contrastive explanations to explain the NN's predictions and important features.

Cao et al.[264] also leveraged off-target interactions to predict drug toxicity in humans, using the MACCS fingerprints of chemical structures and the target representation of drugs

against 477 different human protein targets. Using the SIDER database, they trained an NN to predict the five most common AEs, such as nausea, headache, vomiting, diarrhea, dizziness, and the resulting model achieved accuracy and F1-scores greater than 0.9 for four of the five AEs, including headache, nausea, diarrhea, and dizziness. The results further demonstrate that off-target interactions are crucial in predicting toxicity, and the potential of predicting clinical toxicity without *in vivo* or *in vitro* toxicity information. Yu et al.[268] developed a multi-structural deep learning framework which uses a convolutional NN with an Inception structure and a multi-head Self-Attention module to jointly learn from SMILES sequence-based word embeddings, substructure-based molecular fingerprints, and chemical structure-based graph embedding for the pairwise probability prediction of drug side effects. In 2024, Farnoush et al.[269] constructed AI/ML models which include chemical and molecule features with and without demographic data, using data from the FDA Adverse Event Reporting System (FAERS) database. They compare random forest and deep learning architectures, focused on predicting 30 different adverse reactions, which include common effects such as nausea and diarrhea as well as rare reactions such as psychotic disorders and liver disorders. The models, assessed by AU-ROC, achieve scores of 0.611 and 0.674, and an analysis of feature importance revealed that five were demographic factors, specifically age and sex.

The principles of systems pharmacology are also highly germane to knowledge graph mining[270-272], which refers to the extraction of useful patterns from structured networks of interconnected data, often visualized as a graph. Bean et al.[273] constructed a knowledge graph integrating drug-target, target-indication, and drug-adverse reaction relationships for 524 marketed drugs and developed a prediction algorithm based on enrichment analysis to identifying properties overrepresented among drugs known to cause each adverse reaction (**Figure 1-7**). The method achieved an AU-ROC of 0.92 for classifying known drug-adverse reaction associations and correctly predicts 67.3% of deliberately deleted associations in cross-validation, compared to 20.0% for decision trees, 14.5% for support vector machines, and 14.3% for logistic regression. Predictions were validated against adverse reaction mentions extracted by natural language processing from anonymized electronic health records, where the trained model outperformed random predictions for 92.3% of tested adverse reactions, providing real-world clinical evidence for drug-adverse reaction associations not found in public databases. Zhao et al.[274] employed a knowledge graph mining approach on the Galeano et al.[266] dataset by first

aggregating chemical structure and side effect semantic similarities, known drug-side effect frequencies, and pre-trained word vector representations into a drug-side effect bipartite knowledge graph, then extracting unified embeddings and feeding them into a multilayer perceptron to predict the frequency scores of drug-side effect pairs. In 2025, Chen et al.[275] also employed a knowledge graph of drugs and proteins and used multiple generalized matrix factorization (GMF) models to predict drug-adverse drug reaction pairs, drawing data from SIDER and DrugBank and relabeling negative samples via an ensemble method with label weighting. The model outperforms other baseline models on benchmarks such as AU-ROC and F1-scores, highlighting the potential in methods used to balance negative samples and the ensemble method with label weighting.

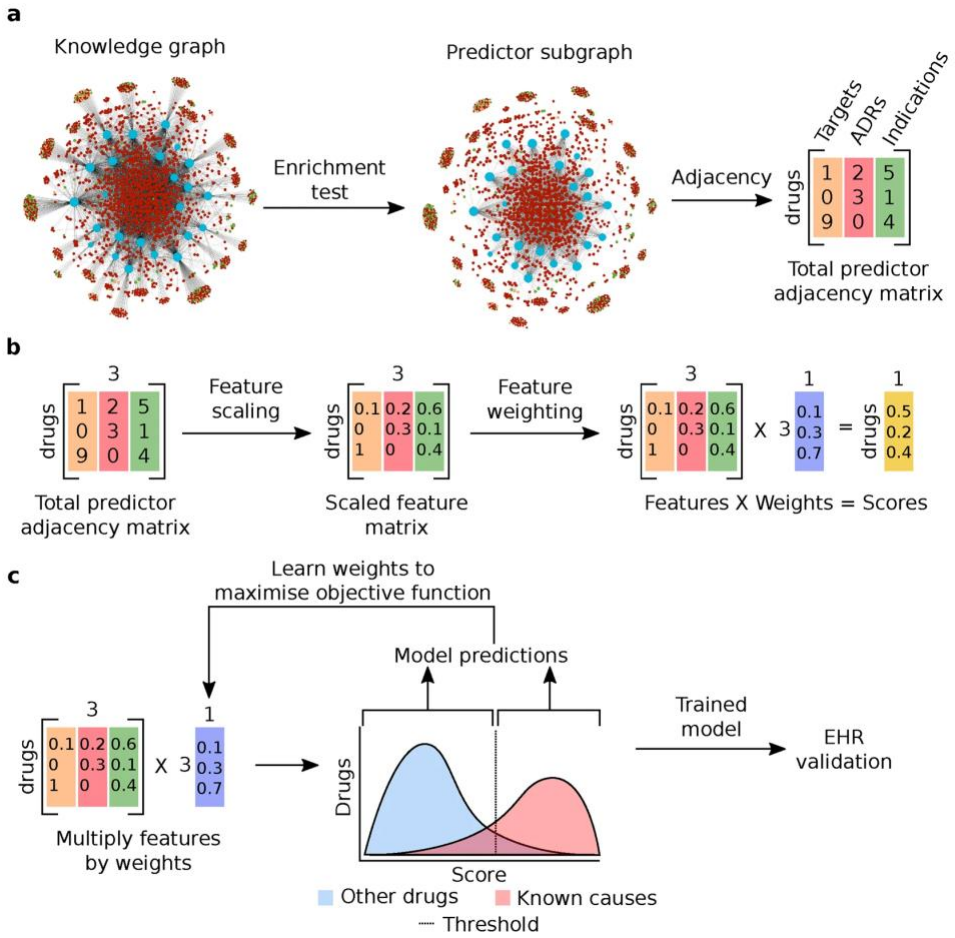


Figure 1-7. Knowledge graph-based prediction of adverse drug reactions. The pipeline proceeds through three stages. In the feature extraction stage (a), a knowledge graph integrating drug-target relationships, clinical indications, and known adverse drug reactions for 524 marketed drugs is queried to identify properties significantly overrepresented among drugs known to cause each reaction. In the scoring stage (b), each drug's associations with

the enriched predictors are normalized and combined into a predictive score. In the validation stage (c), drugs not currently linked to the reaction but scoring above a learned threshold constitute novel predictions, where are prospectively validated against adverse event mentioned extracted from electronic health records. Reproduced from Bean DM, Wu H, Iqbal E, Dzahini O, Ibrahim ZM, Broadbent M, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci Rep.* 2017;7(1):16416. https://doi.org/10.1038/s41598-017-16674-x. © The Author(s) 2017. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Collectively, these analyses demonstrate the growing sophistication of AI/ML approaches for predicting drug toxicity by integrating chemical structure, off-target interactions, clinical reporting data, and systems-level knowledge. Advances in graph NNs, multi-task deep learning, and transfer learning have enabled models to approximate clinical toxicity profiles, including drug-AE associations and off-target effects, without relying solely on *in vivo* or *in vitro* data. However, the generalizability of such analyses remains limited by several factors, including sparse and imbalanced clinical datasets, inconsistent annotation and reporting of AEs, lack of temporal exposure data, and limited incorporation of patient-level variables such as age, sex, comorbidities, and concomitant medications. A further structural limitation concerns the construction of negative training labels. Because databases such as SIDER and FAERS primarily catalog reported drug-AE associations, negative labels must be inferred, often by assuming that the absence of a reported association indicates the absence of a true clinical risk. The reliability of this assumption is questionable, as unreported associations may reflect incomplete surveillance, low event frequency, or insufficient time on market rather than genuine safety. While some approaches have attempted to address label noise through ensemble relabeling methods[275], the fundamental asymmetry between well-documented positive associations and uncertain negative labels remains a pervasive challenge for the field. Additionally, high prediction accuracy for common AEs such as nausea, headache, and diarrhea should be interpreted cautiously, as these events occur at high base rates across many drug classes, and strong performance may reflect prevalence patterns rather than meaningful structure-activity relationships. Many AI/ML models are trained on curated endpoints and surrogate labels, such as binary toxic or non-toxic flags or AE frequency categories, which are unlikely to capture the underlying mechanisms of severe or delayed onset toxicities. Consequently, while these modeling frameworks offer substantial predictive value for early-stage drug screening and hypothesis generation, their clinical translation requires careful validation against temporal patient-level toxicity data and integration of PK and systems biology models.

1.6 Challenges and Emerging Directions in AI/ML-based Drug Toxicity Prediction

1.6.1 Interpretability

In drug safety assessments, where model fit or final parameters often elucidate a molecule's mechanism of action or drug-like properties, AI/ML's opaque decision-making can pose a major hurdle. Interpretability, referring to why and how certain predictions are made, enhances a model's credibility and utility, particularly when handling clinical patient data. Currently, there is little consensus on what interpretability is in AI/ML and how it can be evaluated for benchmarking[276]. For "black box" deep learning architectures, methods such as SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME) offer insights into input molecular feature contributions by approximating local model behavior or distributing importance scores across input dimensions. However, general purpose explainable AI tools were not originally designed to accommodate molecular graphs, hierarchical chemical embeddings, or multi-modal toxicity endpoints, which limits their direct applicability in drug development pipelines[277]. More domain-specific approaches such as Positional SHAP (PoSHAP)[278], which incorporates positional encoding in sequential inputs, or MetA-Train to Explain (MATE)[279], which aligns meta-learning objectives with interpretability constraints, have shown early promise in bridging this gap.

Graph-based explainability methods represent a more structurally informed alternative for molecular property prediction. GNNExplainer[280] identifies the minimal subgraph and node features that most influence a graph NN's prediction by maximizing mutual information between the model output and candidate subgraph structures, and has been applied to mutagenicity prediction where it correctly identified carbon rings, nitro groups, and mutagenic motifs. However, the subgraphs identified by perturbation-based methods are not guaranteed to correspond to chemically meaningful fragments; they may split functional groups across separate atoms or highlight disconnected regions, producing explanations that are difficult for medical chemists to act upon. Wu et al.[281] addressed this with substructure mask explanation, which constrains attributions to chemically valid substructures defined by established fragmentation schemes such as BRICS, Murcko scaffolds, and common functional groups. Applied to mutagenicity and hERG cardiotoxicity prediction, the method correctly identified known toxicophores and detoxifying groups, and its function group attributions for hERG were highly correlated with lipophilicity, consistent with established medicinal chemistry understanding.

Notably, the method also flagged cases where the model had learned incorrect structure-activity relationships due to training data imbalance, demonstrating the potential for explainability methods to serve as quality control tools for model reliability. Attention-based interpretability within graph NNs offers a complementary intrinsic approach, as models such as AttenhERG[85] generate atom- or bond-level importance weights as a byproduct of prediction. These graph-native methods are better suited to molecular data than general-purpose tools, but importantly they identify statistical associations between substructures and outcomes rather than mechanistic explanations of why those substructures cause toxicity.

The interpretability challenge is particularly acute for toxicity prediction relative to other ADME properties. For a model predicting solubility or lipophilicity, a black-box prediction may be acceptable as a screening filter because the downstream decision is typically whether to prioritize a compound for further optimization. For a model informing a safety decision, such as whether a compound advances to first-in-human studies or whether a clinical hold should be imposed, regulators and safety pharmacologists need to understand the mechanistic basis of the prediction, because different mechanistic drivers (structural alerts, off-target interactions, reactive metabolite formation) require different risk mitigation strategies. A toxicity prediction that cannot be traced to a plausible mechanism is difficult to act upon even if it is accurate. This suggests that explainability methods for toxicity models should be evaluated not only by their fidelity to model behavior but also by their ability to generate actionable mechanistic hypotheses. The regulatory stakes of this challenge are made explicit in the FDA's 2025 draft guidance on the use of AI to support regulatory decision-making for drug and biological products[282], which proposes a risk-based credibility assessment framework scaled to the influence of the model on the decision and consequence of an incorrect prediction. Under this framework, an AI model used as the primary basis for a safety decision on a serious toxicity endpoint would be classified as high risk, requiring rigorous documentation of model architecture, training data provenance, independent test set evaluation, uncertainty quantification, and early engagement with the agency. For toxicity prediction models, this means that interpretability is not merely a scientific preference but an emerging regulatory expectation, and that the field's current reliance on *post-hoc* explainability methods applied to opaque architectures may prove insufficient to meet the evidentiary standards that the framework envisions.

1.6.2 Generalizability and benchmarking

The generalizability of AI/ML-based toxicity prediction models remains a critical concern as well. Generalizability refers to a model's ability to maintain predictive performance on novel chemical scaffolds or patient populations not represented in the training data. Overfitting to narrow chemical domains, batch-specific assay artifacts, or specific patient subgroups can severely limit a model's translational relevance. A pervasive methodological issue is the choice of data splitting strategy: many studies report performance using random splits of compound datasets, which allow structurally similar molecules to appear in both training and test sets, inflating apparent accuracy. Scaffold-based splitting, in which compounds sharing the same Murcko scaffold are assigned entirely to one partition[51,179], provides a more realistic assessment of generalization to novel chemotypes but typically yields more conservative performance estimates. The gap between random-split and scaffold-split performance is itself informative, as it indicates the degree to which a model relies on local structural memorization rather than learning transferable structure-toxicity relationships.

A related challenge is the activity cliff problem. Structurally similar compounds can exhibit markedly different toxicity profiles, and conversely, structurally dissimilar compounds can share toxicity mechanisms through convergent pharmacology. This discontinuity in structure-activity relationships is particularly pronounced for toxicity endpoints, where small structural modifications can alter metabolic activation, off-target binding, or reactive metabolite formation in ways that are not captured by molecular fingerprints or learned graph representations. Models that rely heavily on structural similarity for prediction are inherently vulnerable to such cliffs, regardless of their architectural complexity.

The availability of standardized benchmarks such as the Therapeutic Data Commons (TDC)[283], Tox21, and MoleculeNet has improved the consistency of model comparisons, but benchmark performance does not necessarily translate to predictive utility in drug development. Toxicity endpoints consistently rank among the most difficult tasks on these leaderboards, in part because the underlying labels are inherently noisier than those for physicochemical or PK properties. A systematic benchmarking study by Wu et al.[284] using the TOXRIC database reinforces this concern: across 1,381 ToxCast and Tox21 endpoints evaluated with standard ML algorithms, over 80% yielded average F1-scores below 0.5, attributable to small sample sizes and severe class imbalance, with over a third of endpoints exhibiting imbalance ratios exceeding 9:1.

For solubility or CYP inhibition, the ground truth is a continuous measurement from a standardized *in vitro* assay. For DILI or clinical AEs, the label is a categorical judgment aggregated across heterogenous reporting systems, patient populations, doses, and durations of exposure, introducing irreducible noise that no amount of algorithmic refinement can overcome. The incorporation of uncertainty quantification and out-of-distribution detection methods have been increasingly recognized as important in this context[285,286], not only for assessing model reliability but because knowing that a model is uncertain about a compound may be more actionable than the point prediction itself, given the asymmetric cost of false negatives in safety assessment.

1.6.3 Emerging architectures: large language models and multi-modal foundation models

The recent proliferation of large language models (LLMs) and multi-modal foundation models in biomedical research raises the question of whether these architectures can address the limitations of current toxicity prediction approaches. Early investigations suggest a mixed picture. Yang et al.[287] compared GPT-4 and GPT-4o against conventional deep learning and ML models for predicting bone, neurological, and reproductive toxicity, finding that LLM performance was comparable to but did not consistently exceed that of purpose-built models. A notable practical distinction of LLMs is that they require no specialized modeling expertise to deploy: a user can query a general-purpose LLM with a drug name and receive a toxicity assessment without constructing a training dataset, selecting molecular descriptors, or validating a model. This accessibility lowers the barrier to entry for toxicity screening but also increases the risk that predictions are generated and acted upon without the domain expertise needed to evaluate their reliability or provenance. LLMs are also sensitive to prompting strategy and molecular representation format; Park et al.[288] showed that toxicity prediction performance varied substantially between zero-shot, few-shot, and chain-of-thought prompting, and that IUPAC nomenclature yielded better results than SMILES strings for the same compounds. Unlike conventional ML models, where predictions are deterministic for a given input, LLM outputs can vary with prompt phrasing and stochastic decoding, introducing reproducibility challenges for regulatory or screening applications.

LLMs have also been evaluated for DDI detection, where ChatGPT, Gemini, and Claude demonstrated high sensitivity for identifying known DDIs from pharmacovigilance case reports but low specificity, frequently reporting nonexistent interactions in cases without true

DDIs[289,290]. These results suggest that general-purpose LLMs, while capable of encoding broad chemical and biomedical knowledge, do not yet offer clear advantages over domain-specific architectures for quantitative toxicity prediction tasks. A potentially more impactful application of LLMs lies in data curation rather than direct prediction. Silberg et al.[291] developed UniTox, a framework that leverages LLMs to extract and standardize drug-induced toxicity annotations from FDA drug labels, addressing a persistent bottleneck in the construction of high-quality training datasets.

Foundation models pre-trained on large molecular corpora represent a more architecturally substantive development. Ross et al.[292] demonstrated that MoLFormer, a SMILES-based transformer pre-trained on 1.1 billion molecules using self-supervised masked language modeling, outperformed supervised graph NNs on multiple MoleculeNet property prediction benchmarks after fine-tuning. Méndez-Lucio et al.[293] took a complementary approach with MolE, a graph-based foundation model pre-trained on over 842 million molecular graphs using self-supervised learning, which achieved state-of-the-art performance on 10 of 22 ADMET prediction tasks in the TDC benchmark, though notably not on the toxicity-specific endpoints (hERG, Ames, DILI, LD_{50}), where existing models retained superior performance. Three-dimensional foundation models such as Uni-Mol+[294] incorporate spatial conformational information beyond the topological graph, though their advantages over graph-based models on toxicity-specific tasks remain to be systematically established. Several other foundation models, including MoleculeGPT[295] and BioT5[296], have been evaluated on toxicity benchmarks such as Tox21, ClinTox, and SIDER, and the tool-augmented agent ChemCrow[297] has been proposed for chemical property prediction, though none have demonstrated clear advantages over existing approaches on toxicity-specific endpoints. The pre-train-then-fine-tune paradigm these models employ may help address the data scarcity problem that limits many toxicity prediction tasks, by transferring chemical knowledge learned from large unlabeled corpora to smaller, labeled toxicity datasets. Analogous developments in protein language models, including ESM-2 and ProteinBERT, underpin the pretraining-driven frameworks for protein and peptide toxicity discussed in Section 3.7 (ToxDL2.0, VISH-Pred), suggesting that the foundation model paradigm extends across both small-molecule and biologic modalities. Multi-modal foundation models extend this concept further by jointly processing chemical structures, genomic perturbation data, and clinical outcomes within a unified framework, enabling the capture of

cross-modal relationships that single-input architectures cannot represent. Knowledge graph approaches pursue a related strategy by encoding relationships across drugs, targets, and AEs within structured networks. Huang et al.[298] demonstrated this with MADRIGAL, a multi-modal model that integrates structural pathway, cell viability, and transcriptomic data to predict clinical safety outcomes of drug combinations across 953 endpoints, with extensions incorporating patient genomic profiles for individualized risk prediction, illustrating both the promise and persistent gaps of multi-modal approaches.

Addressing the current limitations of these architectures will require methodological developments aligned to each. The reproducibility and false positive issues affecting general purpose LLMs in toxicity and DDI prediction may be mitigated by constrained or structured decoding, retrieval-augmented generation grounded in curated toxicity databases, and tool-augmented agents that delegate quantitative prediction to validated specialist models while reserving the LLM for synthesis and reasoning. The inconsistent transfer of pretraining gains to toxicity-specific endpoints, evident in the MolE results on the TDC benchmark, suggests that toxicity-aware pretraining objectives, multi-task fine-tuning that explicitly couples structural features with mechanistic readouts, and curriculum learning from data-rich to data-sparse endpoints may better align self-supervised representations with downstream safety prediction. UniTox illustrates a third direction in which LLMs may be most impactful: not as direct predictors but as scalable curators that address the data bottleneck on which all of these architectures ultimately depend. Even with these advances, however, neither foundation models nor multi-modal architectures resolve the broader input feature limitation identified throughout this review. If training labels remain binary toxicity classifications derived from heterogenous clinical reporting, and if pharmacological variables such as tissue-specific drug concentrations, target engagement, and clinical dosing are absent from the input feature space, more expressive architectures will learn more complex representations of the same incomplete data.

1.7 Conclusions and Future Perspectives

Across every major toxicity endpoint reviewed here, from hERG channel blockage to DILI to clinical AEs, AI/ML models have achieved high reported accuracy on benchmark datasets, yet their predictions remain difficult to translate into clinical safety decisions. Conventional preclinical assessments, including *in vitro* assays and GLP-compliant animal studies, similarly fail to predict AEs at clinically relevant human doses, a translational gap that

has motivated regulatory initiatives such as the FDA Modernization Act of 2021[299] to validate NAMs. The reasons are not primarily algorithmic; the pharmacological variables that determine whether a molecular liability becomes a clinical adverse event are largely absent from the input feature space of current models. Cardiac tissue drug concentrations are absent from hERG models. Hepatic exposure and zone-specific metabolism are absent from DILI models. Clinical dosing information is absent from DDI models. Patient-specific factors such as pharmacogenomic variation, organ function, and concomitant medications are incorporated in only a small subset of clinical AE models.

Recognizing that clinical dose, efficacy, and safety are immutably interconnected, we propose a Structure–Tissue/cell selectivity–Activity Relationship (STAR) framework that can predict clinical AEs from chemical structure by incorporating mechanistically meaningful, clinically relevant features. Building on our recent work[300,301], STAR integrates 5 core variables:

1. **Chemical structure** (**S**): Encoded as SMILES strings, molecular fingerprints, or graph embeddings for AI/ML input.
2. **Potency/specificity** (**PS**): Determines on- and off-target liabilities that influence both efficacy and toxicity.
3. **Target expression in normal *versus* diseased tissue** (**EX_n *vs.* EX_T**): Defines exposure-response relationships and potential off-target effects.
4. **Tissue/cell selectivity** (**TS**): On- and off-target binding selectivity in normal tissues at clinically relevant doses, which shape drug safety beyond drug-like properties alone.
5. **Clinical dose to reach therapeutic concentrations** (**Cp**): Determined by the PS and TS balance, this dose reflects what is realistically achievable and tolerable in humans.

The potential predictive power of STAR derives from the interplay among these five variables. Tissue and cell selectivity defines which normal tissues are exposed to pharmacologically active drug concentrations, while potency and specificity determine whether that exposure results in on-target toxicity, off-target liability, or both. The clinical dose reached to achieve therapeutic concentrations is itself a function of the potency and selectivity balance: a highly potent and selective compound may achieve efficacy at doses that produce minimal off-target exposure, whereas a less selective compound may require doses at which off-target effects

become dose-limiting. Target expression in normal versus diseased tissue further modulates this relationship, as a target that is highly expressed in the tumor but minimally expressed in cardiac tissue presents a fundamentally different safety profile from one that is ubiquitously expressed. Chemical structure, while already the dominant input to existing AI/ML models, encodes these pharmacological properties only indirectly, and incorporating the downstream consequences of structure would enable models to capture the mechanistic determinants of toxicity that current approaches miss.

Assembling the training data for such a model is feasible with existing resources, though curation remains a substantial undertaking. Chemical structure data for FDA-approved drugs are readily available from PubChem, UniChem[302], and DrugBank. On- and off-target binding affinities can be sourced from ChEMBL, BindingDB, and target class-specific resources such as Davis et al.[303] and KiBA[304] for kinase inhibitors, with computational target prediction methods filling gaps for understudied compounds. Target expression across normal and diseased tissues can be derived from the Human Protein Atlas[305] and integrative portals such as UCSC Xena[306], which aggregates GTEx[307] normal tissue expression with TCGA[308] and other disease-specific transcriptomic resources. Tissue-plasma partitioning coefficients required to estimate tissue-specific drug exposure can be predicted from physicochemical properties using established mechanistic equations or supplemented by experimental measurements where available. Clinical dosing and systemic exposure data are documented in FDA drug labels[309] and can be supplemented by population PK models and PBPK predictions where direct measurements are unavailable. AE data can be extracted directly from the adverse reaction tables in FDA drug labeling, which provide incidence rates for specific AEs at defined doses and can be standardized using terminologies such as SNOMED[310], as well as from FAERS and SIDER, though as this review has emphasized, the latter sources lack the dose, duration, and patient-level context that would maximize their value as training labels. The principal bottleneck is not the existence of these data sources but their integration into a unified, ML-ready format with consistent identifiers, standardized units, and explicit handling of missing values across the five variable domains.

As shown in **Figure 1-8**, the proposed framework organizes these five variables into a multi-task NN that learns to predict clinical AE incidence directly from compound-level inputs assembled from the data sources described above. The framework is designed to be evaluated on

predictive accuracy, calibration, generalization to unseen drugs, and per-AE feature attribution. Once curated for a reference set of 2000 to 5000 FDA-approved drugs with corresponding AE data, the STAR framework could be trained to predict clinical AEs prior to first-in-human trials. By directly modeling the molecular, mechanistic, and pharmacological drivers of clinical toxicity, this approach has the potential to substantially reduce clinical trial failure rates and increase the success of drug candidates advancing to market.

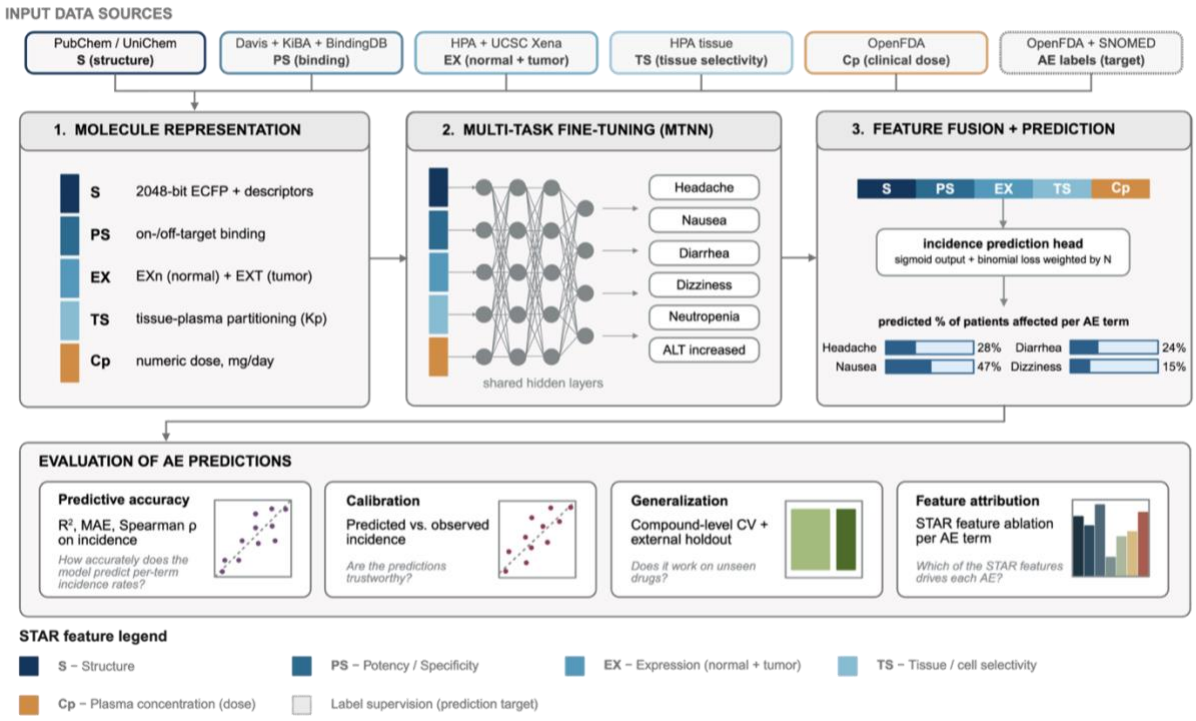


Figure 1-8. The Structure–Tissue/cell selectivity–Activity Relationship (STAR) framework for AI/ML prediction of clinical adverse events. Input data are drawn from public databases for each of the five core variables: chemical structure (S) from PubChem and UniChem; on- and off-target binding affinities (PS) from Davis et al.[303], KiBA, and BindingDB; target expression in normal and tumor tissue (EX) from the Human Protein Atlas and UCSC Xena; tissue–plasma partitioning (TS) from the Human Protein Atlas; and clinical dose (Cp) and AE labels from OpenFDA normalized by SNOMED. The pipeline proceeds through three stages: multi-component molecule representation combining structural fingerprints, binding affinities, expression profiles, partitioning coefficients, and clinical dose; multi-task neural network fine-tuning across shared hidden layers to jointly learn incidence rates for clinically important AE terms; and feature fusion with a sigmoid incidence prediction head trained with binomial loss weighted by patient sample size to output predicted percentages of patients affected per AE term. AE predictions are evaluated through four complementary analyses: predictive accuracy on per-term incidence rates, calibration of predicted versus observed incidence, generalization to unseen drugs through compound-level cross-validation and external holdout, and feature ablation to determine which STAR variables drive each AE prediction. Created in BioRender. Sun, D. (2026) https://BioRender.com/q6a4n4x.

The next chapter presents a first concrete instantiation of this framework, applied to small molecule kinase inhibitors as a clinically important and mechanistically tractable test case. Chapter 2 develops a machine learning model that integrates patient-level pharmacokinetics,

kinome-wide on- and off-target binding profiles, and tissue-specific drug distribution, four of the five variables defined above, to predict the occurrence and time to onset of clinical adverse events in 3,433 cancer patients treated with 16 FDA-approved small molecule kinase inhibitors.

1.8 Research Aims

The preceding review establishes that existing *in silico* approaches to clinical toxicity prediction integrate chemical structure with either target binding or reported adverse event associations, but none represents target engagement, tissue distribution, target expression, and clinical exposure jointly, and none directly measures the *in vivo* behavior of the dosage form that determines systemic exposure in the first place. This dissertation addresses both gaps through three specific aims:

1. Construct a STAR-guided ML framework which predicts the clinical AEs of small molecule kinase inhibitors in cancer patients using kinome-wide binding and tissue selectivity.
2. Assemble data on the five key STAR features for a broader range of small molecules, >1000-3000 FDA-approved drugs.
3. Construct, train, and validate STAR-guided ML models to predict clinical safety profiles for small molecules spanning multiple mechanism classes.

1.9 References

1. Hinkson IV, Madej B, Stahlberg EA. Accelerating Therapeutics for Opportunities in Medicine: A Paradigm Shift in Drug Discovery. *Front Pharmacol*. 2020;11:770. doi:10.3389/fphar.2020.00770
2. Dowden H, Munro J. Trends in clinical success rates and therapeutic focus. *Nat Rev Drug Discov*. Jul 2019;18(7):495–496. doi:10.1038/d41573-019-00074-z
3. Takebe T, Imai R, Ono S. The Current Status of Drug Discovery and Development as Originated in United States Academia: The Influence of Industrial and Academic Collaboration on Drug Discovery and Development. *Clin Transl Sci*. Nov 2018;11(6):597–606. doi:10.1111/cts.12577
4. Harrison RK. Phase II and phase III failures: 2013-2015. *Nat Rev Drug Discov*. Dec 2016;15(12):817–818. doi:10.1038/nrd.2016.184
5. Olson H, Betton G, Robinson D, et al. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol*. Aug 2000;32(1):56–67. doi:10.1006/rtph.2000.1399
6. Tamaki C, Nagayama T, Hashiba M, et al. Potentials and limitations of nonclinical safety assessment for predicting clinical adverse drug reactions: correlation analysis of 142 approved drugs in Japan. *J Toxicol Sci*. 2013;38(4):581–98. doi:10.2131/jts.38.581

7. U.S. Food and Drug Administration. Innovative Science and Technology Approaches for New Drugs (ISTAND) Pilot Program. Updated 05/16/2025. Accessed 07/25/2025, 2025. https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/innovative-science-and-technology-approaches-new-drugs-istand-pilot-program
8. U.S. Food and Drug Administration. Artificial Intelligence (AI) Program for Toxicology at NCTR. Updated 11/26/2024. Accessed 07/25/2025, 2025. https://www.fda.gov/about-fda/nctr-research-focus-areas/artificial-intelligence
9. U.S. Food and Drug Administration. FDA Announces Plan to Phase Out Animal Testing Requirement for Monoclonal Antibodies and Other Drugs. Updated 04/10/2025. Accessed 07/25/2025, 2025. https://www.fda.gov/news-events/press-announcements/fda-announces-plan-phase-out-animal-testing-requirement-monoclonal-antibodies-and-other-drugs
10. Roadmap to Reducing Animal Testing in Preclinical Safety Studies. 2025. https://www.fda.gov/files/newsroom/published/roadmap_to_reducing_animal_testing_in_preclinical_safety_studies.pdf
11. FDA-NIH Workshop: Reducing Animal Testing. 2025. https://www.fda.gov/news-events/fda-meetings-conferences-and-workshops/fda-nih-workshop-reducing-animal-testing-07072025
12. Chen X, Roberts R, Liu Z, Tong W. A generative adversarial network model alternative to animal studies for clinical pathology assessment. *Nat Commun*. Nov 6 2023;14(1):7141. doi:10.1038/s41467-023-42933-9
13. Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence. *Nature*. Aug 2023;620(7972):47–60. doi:10.1038/s41586-023-06221-2
14. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem Rev*. Sep 25 2019;119(18):10520–10594. doi:10.1021/acs.chemrev.8b00728
15. Alsultan A, Aljutayli A, Aljouie A, Albassam A, Woillard JB. Leveraging machine learning in limited sampling strategies for efficient estimation of the area under the curve in pharmacokinetic analysis: a review. *Eur J Clin Pharmacol*. Feb 2025;81(2):183–201. doi:10.1007/s00228-024-03780-9
16. Arav Y. Advances in Modeling Approaches for Oral Drug Delivery: Artificial Intelligence, Physiologically-Based Pharmacokinetics, and First-Principles Models. *Pharmaceutics*. Jul 24 2024;16(8)doi:10.3390/pharmaceutics16080978
17. Bassani D, Parrott NJ, Manevski N, Zhang JD. Another string to your bow: machine learning prediction of the pharmacokinetic properties of small molecules. *Expert Opin Drug Discov*. Jun 2024;19(6):683–698. doi:10.1080/17460441.2024.2348157
18. Danishuddin, Kumar V, Faheem M, Woo Lee K. A decade of machine learning-based predictive models for human pharmacokinetics: Advances and challenges. *Drug Discov Today*. Feb 2022;27(2):529–537. doi:10.1016/j.drudis.2021.09.013
19. Fox T, Kriegl JM. Machine learning techniques for in silico modeling of drug metabolism. *Curr Top Med Chem*. 2006;6(15):1579–91. doi:10.2174/156802606778108915
20. Gawehn E, Greene N, Miljkovic F, et al. Perspectives on the use of machine learning for ADME prediction at AstraZeneca. *Xenobiotica*. Jul 2024;54(7):368–378. doi:10.1080/00498254.2024.2352598
21. Gill J, Moullet M, Martinsson A, et al. Comparing the applications of machine learning, PBPK, and population pharmacokinetic models in pharmacokinetic drug-drug interaction

prediction. *CPT Pharmacometrics Syst Pharmacol*. Dec 2022;11(12):1560–1568. doi:10.1002/psp4.12870
22. Guo W, Dong Y, Hao GF. Transfer learning empowers accurate pharmacokinetics prediction of small samples. *Drug Discov Today*. Apr 2024;29(4):103946. doi:10.1016/j.drudis.2024.103946
23. Huang S, Xu Q, Yang G, Ding J, Pei Q. Machine Learning for Prediction of Drug Concentrations: Application and Challenges. *Clin Pharmacol Ther*. May 2025;117(5):1236–1247. doi:10.1002/cpt.3577
24. Li Y, Wang Z, Li Y, et al. A Combination of Machine Learning and PBPK Modeling Approach for Pharmacokinetics Prediction of Small Molecules in Humans. *Pharm Res*. Jul 2024;41(7):1369–1379. doi:10.1007/s11095-024-03725-y
25. Lim MA, Yang S, Mai H, Cheng AC. Exploring Deep Learning of Quantum Chemical Properties for Absorption, Distribution, Metabolism, and Excretion Predictions. *J Chem Inf Model*. Dec 26 2022;62(24):6336–6341. doi:10.1021/acs.jcim.2c00245
26. M M, S M, Kate SS, et al. A review on integrated machine learning and deep learning driven artificial intelligence models for pharmacokinetics and toxicokinetics predictions, and their application. *Drug Metab Dispos*. Jan 22 2026;54(3):100240. doi:10.1016/j.dmd.2026.100240
27. Maltarollo VG, Gertrudes JC, Oliveira PR, Honorio KM. Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin Drug Metab Toxicol*. Feb 2015;11(2):259–71. doi:10.1517/17425255.2015.980814
28. McComb M, Bies R, Ramanathan M. Machine learning in pharmacometrics: Opportunities and challenges. *Br J Clin Pharmacol*. Feb 2022;88(4):1482–1499. doi:10.1111/bcp.14801
29. Obrezanova O. Artificial intelligence for compound pharmacokinetics prediction. *Curr Opin Struct Biol*. Apr 2023;79:102546. doi:10.1016/j.sbi.2023.102546
30. Ota R, Yamashita F. Application of machine learning techniques to the analysis and prediction of drug pharmacokinetics. *J Control Release*. Dec 2022;352:961–969. doi:10.1016/j.jconrel.2022.11.014
31. Shah A, Bi F, Yang J. Machine learning to predict food effects during drug development: a comprehensive review. *J Cheminform*. Dec 10 2025;18(1):8. doi:10.1186/s13321-025-01131-z
32. Talkington AM, Cao Y, Kearsley AJ, Lai SK. Opportunities for machine learning and artificial intelligence in physiologically-based pharmacokinetic (PBPK) modeling. *Adv Drug Deliv Rev*. Dec 2025;227:115716. doi:10.1016/j.addr.2025.115716
33. Tanihata S, Iwata H. Next-Generation Artificial Intelligence for ADME Prediction in Drug Discovery: From Small Molecules to Biologics. *Yonago Acta Med*. Feb 2026;69(1):1–13. doi:10.33160/yam.2026.02.001
34. Tao L, Zhang P, Qin C, et al. Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools. *Adv Drug Deliv Rev*. Jun 23 2015;86:83–100. doi:10.1016/j.addr.2015.03.014
35. Tran TTV, Tayara H, Chong KT. Recent Studies of Artificial Intelligence on In Silico Drug Distribution Prediction. *Int J Mol Sci*. Jan 17 2023;24(3)doi:10.3390/ijms24031815
36. Uno M, Nakamaru Y, Yamashita F. Application of machine learning techniques in population pharmacokinetics/pharmacodynamics modeling. *Drug Metab Pharmacokinet*. Jun 2024;56:101004. doi:10.1016/j.dmpk.2024.101004

37. Varela-Rey I, Bandin-Vilar E, Toja-Camba FJ, et al. Artificial Intelligence and Machine Learning Applications to Pharmacokinetic Modeling and Dose Prediction of Antibiotics: A Scoping Review. *Antibiotics (Basel)*. Dec 10 2024;13(12)doi:10.3390/antibiotics13121203
38. Venkataraman M, Rao GC, Madavareddi JK, Maddi SR. Leveraging machine learning models in evaluating ADMET properties for drug discovery and development. *ADMET DMPK*. 2025;13(3):2772. doi:10.5599/admet.2772
39. Richard AM, Huang R, Waidyanatha S, et al. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem Res Toxicol*. Feb 15 2021;34(2):189–216. doi:10.1021/acs.chemrestox.0c00264
40. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*. 1988/02/01 1988;28(1):31–36. doi:10.1021/ci00057a005
41. Capecchi A, Probst D, Reymond JL. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J Cheminform*. Jun 12 2020;12(1):43. doi:10.1186/s13321-020-00445-4
42. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*. Nov–Dec 2002;42(6):1273–80. doi:10.1021/ci010132r
43. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform*. Oct 7 2011;3:33. doi:10.1186/1758-2946-3-33
44. Roy K, Mitra I. Electrotopological state atom (E-state) index in drug design, QSAR, property prediction and toxicity assessment. *Curr Comput Aided Drug Des*. Jun 2012;8(2):135–58. doi:10.2174/157340912800492366
45. Morgan HL. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*. 1965/05/01 1965;5(2):107–113. doi:10.1021/c160017a018
46. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. May 24 2010;50(5):742–54. doi:10.1021/ci100050t
47. McGregor MJ, Muskal SM. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J Chem Inf Comput Sci*. May–Jun 1999;39(3):569–74. doi:10.1021/ci980159j
48. Mason JS, Cheney DL. Library design and virtual screening using multiple 4-point pharmacophore fingerprints. *Pac Symp Biocomput*. 2000:576–87. doi:10.1142/9789814447331_0055
49. Stiefl N, Watson IA, Baumann K, Zaliani A. ErG: 2D pharmacophore descriptions for scaffold hopping. *J Chem Inf Model*. Jan–Feb 2006;46(1):208–20. doi:10.1021/ci050457y
50. Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*. 2015;28
51. Yang K, Swanson K, Jin W, et al. Correction to Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model*. Dec 23 2019;59(12):5304–5305. doi:10.1021/acs.jcim.9b01076
52. Hopfinger AJ, Wang S, Tokarski JS, et al. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *Journal of the American Chemical Society*. 1997/10/01 1997;119(43):10509–10524. doi:10.1021/ja9718937
53. Bajusz D, Rácz A, Héberger K. Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In: Chackalamannil S, Rotella DP, Ward SE, eds. *Comprehensive Medicinal Chemistry III*. 3rd ed. Elsevier; 2017:329–378.

54. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. *J Cheminform*. Feb 6 2018;10(1):4. doi:10.1186/s13321-018-0258-y
55. Willighagen EL, Mayfield JW, Alvarsson J, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform*. Jun 6 2017;9(1):33. doi:10.1186/s13321-017-0220-4
56. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. May 2011;32(7):1466–74. doi:10.1002/jcc.21707
57. Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*. 1962/04/01 1962;194(4824):178–180. doi:10.1038/194178b0
58. Roche O, Trube G, Zuegge J, Pflimlin P, Alanine A, Schneider G. A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *Chembiochem*. May 3 2002;3(5):455–9. doi:10.1002/1439-7633(20020503)3:5<455::AID-CBIC455>3.0.CO;2-L
59. Siramshetty VB, Chen Q, Devarakonda P, Preissner R. The Catch-22 of Predicting hERG Blockade Using Publicly Accessible Bioactivity Data. *J Chem Inf Model*. Jun 25 2018;58(6):1224–1233. doi:10.1021/acs.jcim.8b00150
60. Cianchetta G, Li Y, Kang J, et al. Predictive models for hERG potassium channel blockers. *Bioorg Med Chem Lett*. Aug 1 2005;15(15):3637–42. doi:10.1016/j.bmcl.2005.03.062
61. Keseru GM. Prediction of hERG potassium channel affinity by traditional and hologram qSAR methods. *Bioorg Med Chem Lett*. Aug 18 2003;13(16):2773–5. doi:10.1016/s0960-894x(03)00492-x
62. Sun H. An accurate and interpretable bayesian classification model for prediction of HERG liability. *ChemMedChem*. Mar 2006;1(3):315–22. doi:10.1002/cmdc.200500047
63. Wang S, Li Y, Wang J, et al. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol Pharm*. Apr 2 2012;9(4):996–1010. doi:10.1021/mp300023x
64. Liu LL, Lu J, Lu Y, et al. Novel Bayesian classification models for predicting compounds blocking hERG potassium channels. *Acta Pharmacol Sin*. Aug 2014;35(8):1093–102. doi:10.1038/aps.2014.35
65. Obrezanova O, Csanyi G, Gola JM, Segall MD. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J Chem Inf Model*. Sep–Oct 2007;47(5):1847–57. doi:10.1021/ci7000633
66. Gepp MM, Hutter MC. Determination of hERG channel blockers using a decision tree. *Bioorg Med Chem*. Aug 1 2006;14(15):5325–32. doi:10.1016/j.bmc.2006.03.043
67. Tobita M, Nishikawa T, Nagashima R. A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors. *Bioorg Med Chem Lett*. Jun 2 2005;15(11):2886–90. doi:10.1016/j.bmcl.2005.03.080
68. Song M, Clark M. Development and evaluation of an in silico model for hERG binding. *J Chem Inf Model*. Jan–Feb 2006;46(1):392–400. doi:10.1021/ci050308f
69. Leong MK. A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. *Chem Res Toxicol*. Feb 2007;20(2):217–26. doi:10.1021/tx060230c
70. Jia L, Sun H. Support vector machines classification of hERG liabilities based on atom types. *Bioorg Med Chem*. Jun 1 2008;16(11):6252–60. doi:10.1016/j.bmc.2008.04.028

71. Shen MY, Su BH, Esposito EX, Hopfinger AJ, Tseng YJ. A comprehensive support vector machine binary hERG classification model based on extensive but biased end point hERG data sets. *Chem Res Toxicol*. Jun 20 2011;24(6):934–49. doi:10.1021/tx200099j
72. Yap CW, Cai CZ, Xue Y, Chen YZ. Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol Sci*. May 2004;79(1):170–7. doi:10.1093/toxsci/kfh082
73. Doddareddy MR, Klaasse EC, Shagufta, Ijzerman AP, Bender A. Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases. *ChemMedChem*. May 3 2010;5(5):716–29. doi:10.1002/cmdc.201000024
74. Chen Y, Yu X, Li W, Tang Y, Liu G. In silico prediction of hERG blockers using machine learning and deep learning approaches. *J Appl Toxicol*. Oct 2023;43(10):1462–1475. doi:10.1002/jat.4477
75. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. Jan 4 2016;44(D1):D1045–53. doi:10.1093/nar/gkv1072
76. Kim S, Chen J, Cheng T, et al. PubChem 2025 update. *Nucleic Acids Res*. Jan 6 2025;53(D1):D1516–D1525. doi:10.1093/nar/gkae1059
77. Zdrazil B, Felix E, Hunter F, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res*. Jan 5 2024;52(D1):D1180–D1192. doi:10.1093/nar/gkad1004
78. Zhang Y, Zhao J, Wang Y, et al. Prediction of hERG K+ channel blockage using deep neural networks. *Chem Biol Drug Des*. Sep 2019;94(5):1973–1985. doi:10.1111/cbdd.13600
79. Cai C, Guo P, Zhou Y, et al. Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. *J Chem Inf Model*. Mar 25 2019;59(3):1073–1084. doi:10.1021/acs.jcim.8b00769
80. Lee HM, Yu MS, Kazmi SR, et al. Computational determination of hERG-related cardiotoxicity of drug candidates. *BMC Bioinformatics*. May 29 2019;20(Suppl 10):250. doi:10.1186/s12859-019-2814-5
81. Ryu JY, Lee MY, Lee JH, Lee BH, Oh KS. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. *Bioinformatics*. May 1 2020;36(10):3049–3055. doi:10.1093/bioinformatics/btaa075
82. Kim H, Park M, Lee I, Nam H. BayeshERG: a robust, reliable and interpretable deep learning model for predicting hERG channel blockers. *Brief Bioinform*. Jul 18 2022;23(4)doi:10.1093/bib/bbac211
83. Liu M, Zhang L, Li S, et al. Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints. *Toxicol Lett*. Oct 10 2020;332:88–96. doi:10.1016/j.toxlet.2020.07.003
84. Liu J, Khan MKH, Guo W, et al. Machine learning and deep learning approaches for enhanced prediction of hERG blockade: a comprehensive QSAR modeling study. *Expert Opin Drug Metab Toxicol*. Jul 2024;20(7):665–684. doi:10.1080/17425255.2024.2377593
85. Yang T, Ding X, McMichael E, et al. AttenhERG: a reliable and interpretable graph neural network framework for predicting hERG channel blockers. *J Cheminform*. Dec 23 2024;16(1):143. doi:10.1186/s13321-024-00940-y
86. Karim A, Lee M, Balle T, Sattar A. CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. *J Cheminform*. Aug 16 2021;13(1):60. doi:10.1186/s13321-021-00541-z

87. Siramshetty VB, Nguyen DT, Martinez NJ, Southall NT, Simeonov A, Zakharov AV. Critical Assessment of Artificial Intelligence Methods for Prediction of hERG Channel Inhibition in the "Big Data" Era. *J Chem Inf Model*. Dec 28 2020;60(12):6007–6019. doi:10.1021/acs.jcim.0c00884
88. Jing Y, Zhang Y, Zhao G, et al. GraphDeep-hERG: Graph Neural Network PharmacoAnalytics for Assessing hERG-Related Cardiotoxicity. *Pharm Res*. Apr 2025;42(4):579–591. doi:10.1007/s11095-025-03848-w
89. Redfern WS, Carlsson L, Davis AS, et al. Relationships between preclinical cardiac electrophysiology, clinical QT interval prolongation and torsade de pointes for a broad range of drugs: evidence for a provisional safety margin in drug development. *Cardiovasc Res*. Apr 1 2003;58(1):32–45. doi:10.1016/s0008-6363(02)00846-5
90. Sager PT, Gintant G, Turner JR, Pettit S, Stockbridge N. Rechanneling the cardiac proarrhythmia safety paradigm: a meeting report from the Cardiac Safety Research Consortium. *Am Heart J*. Mar 2014;167(3):292–300. doi:10.1016/j.ahj.2013.11.004
91. Gintant GA. Preclinical Torsades-de-Pointes screens: advantages and limitations of surrogate and direct approaches in evaluating proarrhythmic risk. *Pharmacol Ther*. Aug 2008;119(2):199–209. doi:10.1016/j.pharmthera.2008.04.010
92. Wang K, Zhang S, Marzolf B, et al. Circulating microRNAs, potential biomarkers for drug-induced liver injury. *Proc Natl Acad Sci U S A*. Mar 17 2009;106(11):4402–7. doi:10.1073/pnas.0813371106
93. David S, Hamilton JP. Drug-induced Liver Injury. *US Gastroenterol Hepatol Rev*. Jan 1 2010;6:73–80.
94. Mosedale M, Watkins PB. Drug-induced liver injury: Advances in mechanistic understanding that will inform risk management. *Clin Pharmacol Ther*. Apr 2017;101(4):469–480. doi:10.1002/cpt.564
95. Guidance for Industry: Drug-Induced Liver Injury — Premarketing Clinical Evaluation (U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER)) 38035−38036 (2009).
96. Hultin-Rosenberg L, Jagannathan S, Nilsson KC, et al. Predictive models of hepatotoxicity using gene expression data from primary rat hepatocytes. *Xenobiotica*. Oct–Nov 2006;36(10-11):1122–39. doi:10.1080/00498250600861801
97. Cruz-Monteagudo M, Cordeiro MN, Borges F. Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. *J Comput Chem*. Mar 2008;29(4):533–49. doi:10.1002/jcc.20812
98. Rodgers AD, Zhu H, Fourches D, Rusyn I, Tropsha A. Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure-activity relationship method. *Chem Res Toxicol*. Apr 19 2010;23(4):724–32. doi:10.1021/tx900451r
99. Liew CY, Lim YC, Yap CW. Mixed learning algorithms and features ensemble in hepatotoxicity prediction. *J Comput Aided Mol Des*. Sep 2011;25(9):855–71. doi:10.1007/s10822-011-9468-3
100. Chen M, Hong H, Fang H, et al. Quantitative structure-activity relationship models for predicting drug-induced liver injury based on FDA-approved drug labeling annotation and using a large collection of drugs. *Toxicol Sci*. Nov 2013;136(1):242–9. doi:10.1093/toxsci/kft189
101. Ekins S, Williams AJ, Xu JJ. A predictive ligand-based Bayesian model for human drug-induced liver injury. *Drug Metab Dispos*. Dec 2010;38(12):2302–8. doi:10.1124/dmd.110.035113

102. Zhang H, Ding L, Zou Y, et al. Predicting drug-induced liver injury in human with Naive Bayes classifier approach. *J Comput Aided Mol Des*. Oct 2016;30(10):889–898. doi:10.1007/s10822-016-9972-6
103. Liu Z, Shi Q, Ding D, Kelly R, Fang H, Tong W. Translating clinical findings into knowledge in drug safety evaluation--drug induced liver injury prediction system (DILIps). *PLoS Comput Biol*. Dec 2011;7(12):e1002310. doi:10.1371/journal.pcbi.1002310
104. Chen M, Vijay V, Shi Q, Liu Z, Fang H, Tong W. FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov Today*. Aug 2011;16(15-16):697–703. doi:10.1016/j.drudis.2011.05.007
105. Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov Today*. Apr 2016;21(4):648–53. doi:10.1016/j.drudis.2016.02.015
106. Nguyen-Vo TH, Nguyen L, Do N, et al. Predicting Drug-Induced Liver Injury Using Convolutional Neural Network and Molecular Fingerprint-Embedded Features. *ACS Omega*. Oct 6 2020;5(39):25432–25439. doi:10.1021/acsomega.0c03866
107. Minerali E, Foil DH, Zorn KM, Lane TR, Ekins S. Comparing Machine Learning Algorithms for Predicting Drug-Induced Liver Injury (DILI). *Mol Pharm*. Jul 6 2020;17(7):2628–2637. doi:10.1021/acs.molpharmaceut.0c00326
108. Liu A, Walter M, Wright P, et al. Prediction and mechanistic analysis of drug-induced liver injury (DILI) based on chemical structure. *Biol Direct*. Jan 18 2021;16(1):6. doi:10.1186/s13062-020-00285-0
109. Aguirre-Plans J, Pinero J, Souza T, et al. An ensemble learning approach for modeling the systems biology of drug-induced injury. *Biol Direct*. Jan 12 2021;16(1):5. doi:10.1186/s13062-020-00288-x
110. Chen Z, Zhao M, You L, et al. Developing an artificial intelligence method for screening hepatotoxic compounds in traditional Chinese medicine and Western medicine combination. *Chin Med*. May 17 2022;17(1):58. doi:10.1186/s13020-022-00617-4
111. Rudik AV, Bezhentsev VM, Dmitriev AV, et al. MetaTox: Web Application for Predicting Structure and Toxicity of Xenobiotics' Metabolites. *J Chem Inf Model*. Apr 24 2017;57(4):638–642. doi:10.1021/acs.jcim.6b00662
112. Thakkar S, Li T, Liu Z, Wu L, Roberts R, Tong W. Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity. *Drug Discov Today*. Jan 2020;25(1):201–208. doi:10.1016/j.drudis.2019.09.022
113. Seal S, Williams D, Hosseini-Gerami L, et al. Improved Detection of Drug-Induced Liver Injury by Integrating Predicted In Vivo and In Vitro Data. *Chem Res Toxicol*. Aug 19 2024;37(8):1290–1305. doi:10.1021/acs.chemrestox.4c00015
114. Li T, Li J, Jiang H, Skiles DB. Deep Learning Prediction of Drug-Induced Liver Toxicity by Manifold Embedding of Quantum Information of Drug Molecules. *Pharm Res*. Jan 2025;42(1):109–122. doi:10.1007/s11095-024-03800-4
115. Bergen V, Kodella K, Srikrishnan S, et al. A large-scale human toxicogenomics resource for drug-induced liver injury prediction. *Nat Commun*. Nov 13 2025;16(1):9860. doi:10.1038/s41467-025-65690-3
116. Strickland J, Clippinger AJ, Brown J, et al. Status of acute systemic toxicity testing requirements and data uses by U.S. regulatory agencies. *Regul Toxicol Pharmacol*. Apr 2018;94:183–196. doi:10.1016/j.yrtph.2018.01.022

117. Label review manual chapter 7: precautionary statements (U.S. Environmental Protection Agency) (2012).
118. Kaiser KL, Niculescu SP. Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (Pimephales promelas): a study based on 865 compounds. *Chemosphere*. Jun 1999;38(14):3237–45. doi:10.1016/s0045-6535(99)00553-6
119. Niculescu SP, Kaiser KL, Schultz TW. Modeling the toxicity of chemicals to Tetrahymena pyriformis using molecular fragment descriptors and probabilistic neural networks. *Arch Environ Contam Toxicol*. Oct 2000;39(3):289–98. doi:10.1007/s002440010107
120. Kaiser KL, Niculescu SP. Modeling acute toxicity of chemicals to Daphnia magna: a probabilistic neural network approach. *Environ Toxicol Chem*. Feb 2001;20(2):420–31.
121. Xue Y, Li H, Ung CY, Yap CW, Chen YZ. Classification of a diverse set of Tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. *Chem Res Toxicol*. Aug 2006;19(8):1030–9. doi:10.1021/tx0600550
122. Serra JR, Jurs PC, Kaiser KL. Linear regression and computational neural network prediction of tetrahymena acute toxicity for aromatic compounds from molecular structure. *Chem Res Toxicol*. Nov 2001;14(11):1535–45. doi:10.1021/tx010101q
123. Lei T, Li Y, Song Y, Li D, Sun H, Hou T. ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. *J Cheminform*. 2016;8:6. doi:10.1186/s13321-016-0117-7
124. Ai H, Wu X, Zhang L, et al. QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. *Ecotoxicol Environ Saf*. Sep 15 2019;179:71–78. doi:10.1016/j.ecoenv.2019.04.035
125. Bouhedjar K, Boukelia A, Khorief Nacereddine A, Boucheham A, Belaidi A, Djerourou A. A natural language processing approach based on embedding deep learning from heterogeneous compounds for quantitative structure-activity relationship modeling. *Chem Biol Drug Des*. Sep 2020;96(3):961–972. doi:10.1111/cbdd.13742
126. Daghighi A, Casanola-Martin GM, Timmerman T, Milenkovic D, Lucic B, Rasulev B. In Silico Prediction of the Toxicity of Nitroaromatic Compounds: Application of Ensemble Learning QSAR Approach. *Toxics*. Dec 1 2022;10(12)doi:10.3390/toxics10120746
127. Singh KP, Gupta S, Rai P. Predicting acute aquatic toxicity of structurally diverse chemicals in fish using artificial intelligence approaches. *Ecotoxicol Environ Saf*. Sep 2013;95:221–33. doi:10.1016/j.ecoenv.2013.05.017
128. Fan T, Sun G, Zhao L, Cui X, Zhong R. QSAR and Classification Study on Prediction of Acute Oral Toxicity of N-Nitroso Compounds. *Int J Mol Sci*. Oct 3 2018;19(10)doi:10.3390/ijms19103015
129. Lunghini F, Marcou G, Azam P, et al. Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context. *SAR QSAR Environ Res*. Dec 2019;30(12):879–897. doi:10.1080/1062936X.2019.1672089
130. Vukovic K, Gadaleta D, Benfenati E. Methodology of aiQSAR: a group-specific approach to QSAR modelling. *J Cheminform*. Apr 3 2019;11(1):27. doi:10.1186/s13321-019-0350-y
131. Ryu JY, Jang WD, Jang J, Oh KS. PredAOT: a computational framework for prediction of acute oral toxicity based on multiple random forest models. *BMC Bioinformatics*. Feb 24 2023;24(1):66. doi:10.1186/s12859-023-05176-5
132. Mansouri K, Karmaus AL, Fitzpatrick J, et al. CATMoS: Collaborative Acute Toxicity Modeling Suite. *Environ Health Perspect*. Apr 2021;129(4):47013. doi:10.1289/EHP8495

133. Bo T, Lin Y, Han J, Hao Z, Liu J. Machine learning-assisted data filtering and QSAR models for prediction of chemical acute toxicity on rat and mouse. *J Hazard Mater*. Jun 15 2023;452:131344. doi:10.1016/j.jhazmat.2023.131344
134. Lane TR, Harris J, Urbina F, Ekins S. Comparing LD(50)/LC(50) Machine Learning Models for Multiple Species. *J Chem Health Saf*. Mar 27 2023;30(2):83–97. doi:10.1021/acs.chas.2c00088
135. Jain S, Siramshetty VB, Alves VM, et al. Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods. *J Chem Inf Model*. Feb 22 2021;61(2):653–663. doi:10.1021/acs.jcim.0c01164
136. Daghighi A, Casanola-Martin GM, Iduoku K, Kusic H, Gonzalez-Diaz H, Rasulev B. Multi-Endpoint Acute Toxicity Assessment of Organic Compounds Using Large-Scale Machine Learning Modeling. *Environ Sci Technol*. Jun 11 2024;58(23):10116–10127. doi:10.1021/acs.est.4c01017
137. Mustafa FM, Al-Hussainy AF, Doshi H, et al. TabNet and TabTransformer: Novel Deep Learning Models for Chemical Toxicity Prediction in Comparison With Machine Learning. *J Appl Toxicol*. May 1 2025;doi:10.1002/jat.4803
138. S2(R1) Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use (U.S. Food and Drug Administration) (2012).
139. Ames BN, McCann J, Yamasaki E. Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test. *Mutat Res*. Dec 1975;31(6):347–64. doi:10.1016/0165-1161(75)90046-1
140. Hansen K, Mika S, Schroeter T, et al. Benchmark data set for in silico prediction of Ames mutagenicity. *J Chem Inf Model*. Sep 2009;49(9):2077–81. doi:10.1021/ci900161g
141. Fitzpatrick RB. CPDB: Carcinogenic Potency Database. *Med Ref Serv Q*. Fall 2008;27(3):303–11. doi:10.1080/02763860802198895
142. Kumar P, Ma X, Liu X, et al. Effect of training data size and noise level on support vector machines virtual screening of genotoxic compounds from large compound libraries. *J Comput Aided Mol Des*. May 2011;25(5):455–67. doi:10.1007/s10822-011-9431-3
143. Sharma A, Kumar R, Varadwaj PK, Ahmad A, Ashraf GM. A comparative study of support vector machine, artificial neural network and bayesian classifier for mutagenicity prediction. *Interdiscip Sci*. Sep 2011;3(3):232–9. doi:10.1007/s12539-011-0102-9
144. Fan D, Yang H, Li F, et al. In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. *Toxicol Res (Camb)*. Mar 1 2018;7(2):211–220. doi:10.1039/c7tx00259a
145. Zhang H, Kang YL, Zhu YY, et al. Novel naive Bayes classification models for predicting the chemical Ames mutagenicity. *Toxicol In Vitro*. Jun 2017;41:56–63. doi:10.1016/j.tiv.2017.02.016
146. Fujita Y, Honda H, Yamane M, Morita T, Matsuda T, Morita O. A decision tree-based integrated testing strategy for tailor-made carcinogenicity evaluation of test substances using genotoxicity test results and chemical spaces. *Mutagenesis*. Mar 6 2019;34(1):101–109. doi:10.1093/mutage/gey039
147. Ramesh P, Veerappapillai S. Prediction of Micronucleus Assay Outcome Using In Vivo Activity Data and Molecular Structure Features. *Appl Biochem Biotechnol*. Dec 2021;193(12):4018–4034. doi:10.1007/s12010-021-03720-8

148. Zhang J, Mucs D, Norinder U, Svensson F. LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets. *J Chem Inf Model*. Oct 28 2019;59(10):4150–4158. doi:10.1021/acs.jcim.9b00633
149. Xu C, Cheng F, Chen L, et al. In silico prediction of chemical Ames mutagenicity. *J Chem Inf Model*. Nov 26 2012;52(11):2840–7. doi:10.1021/ci300400a
150. Chu CSM, Simpson JD, O'Neill PM, Berry NG. Machine learning - Predicting Ames mutagenicity of small molecules. *J Mol Graph Model*. Dec 2021;109:108011. doi:10.1016/j.jmgm.2021.108011
151. Honma M, Kitazawa A, Cayley A, et al. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. *Mutagenesis*. Mar 6 2019;34(1):3–16. doi:10.1093/mutage/gey031
152. Koutroumpa NM, Antoniou M, Varsou DD, et al. Titania: an integrated tool for in silico molecular property prediction and NAM-based modeling. *Mol Divers*. Aug 2025;29(4):3555–3573. doi:10.1007/s11030-025-11196-5
153. Mittal A, Mohanty SK, Gautam V, et al. Artificial intelligence uncovers carcinogenic human metabolites. *Nat Chem Biol*. Nov 2022;18(11):1204–1213. doi:10.1038/s41589-022-01110-7
154. Hung C, Gini G. QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction. *Mol Divers*. Aug 2021;25(3):1283–1299. doi:10.1007/s11030-021-10250-2
155. Tran TTV, Tayara H, Chong KT. AMPred-CNN: Ames mutagenicity prediction model based on convolutional neural networks. *Comput Biol Med*. Jun 2024;176:108560. doi:10.1016/j.compbiomed.2024.108560
156. Ashby J, Tennant RW. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat Res*. May 1991;257(3):229–306. doi:10.1016/0165-1110(91)90003-e
157. ICH Harmonised Guideline: Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk M7 (U.S. Food and Drug Administration) (2014).
158. National Center for Advancing Translational Sciences. Tox21 Data Challenge 2014. Online. National Center for Advancing Translational Sciences, National Institutes of Health. Accessed 07/25/2025, https://tripod.nih.gov/tox21/challenge/
159. Banerjee P, Siramshetty VB, Drwal MN, Preissner R. Computational methods for prediction of in vitro effects of new chemical structures. *J Cheminform*. 2016;8:51. doi:10.1186/s13321-016-0162-2
160. Banerjee P, Eckert AO, Schrey AK, Preissner R. ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res*. Jul 2 2018;46(W1):W257–W263. doi:10.1093/nar/gky318
161. Alves VM, Golbraikh A, Capuzzi SJ, et al. Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure-Activity Relationship Models. *J Chem Inf Model*. Jun 25 2018;58(6):1214–1223. doi:10.1021/acs.jcim.8b00124
162. Grenet I, Yin Y, Comet JP. G-Networks to Predict the Outcome of Sensing of Toxicity. *Sensors (Basel)*. Oct 16 2018;18(10)doi:10.3390/s18103483

163. Peng Y, Zhang Z, Jiang Q, Guan J, Zhou S. TOP: A deep mixture representation learning method for boosting molecular toxicity prediction. *Methods*. Jul 1 2020;179:55–64. doi:10.1016/j.ymeth.2020.05.013
164. Stahl N, Falkman G, Karlsson A, Mathiason G, Bostrom J. Deep Convolutional Neural Networks for the Prediction of Molecular Properties: Challenges and Opportunities Connected to the Data. *J Integr Bioinform*. Dec 5 2018;16(1)doi:10.1515/jib-2018-0065
165. Yuan Q, Wei Z, Guan X, et al. Toxicity Prediction Method Based on Multi-Channel Convolutional Neural Network. *Molecules*. Sep 17 2019;24(18)doi:10.3390/molecules24183383
166. Chen J, Si YW, Un CW, Siu SWI. Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network. *J Cheminform*. Nov 27 2021;13(1):93. doi:10.1186/s13321-021-00570-8
167. Saravanan KM, Wan JF, Dai L, Zhang J, Zhang JZH, Zhang H. A deep learning based multi-model approach for predicting drug-like chemical compound's toxicity. *Methods*. Jun 2024;226:164–175. doi:10.1016/j.ymeth.2024.04.020
168. Hao Y, Romano JD, Moore JH. Knowledge-guided deep learning models of drug toxicity improve interpretation. *Patterns (N Y)*. Sep 9 2022;3(9):100565. doi:10.1016/j.patter.2022.100565
169. Nascimben M, Rimondini L. Molecular Toxicity Virtual Screening Applying a Quantized Computational SNN-Based Framework. *Molecules*. Jan 31 2023;28(3)doi:10.3390/molecules28031342
170. Zhou Y, Ning C, Tan Y, et al. ToxMPNN: A deep learning model for small molecule toxicity prediction. *J Appl Toxicol*. Jul 2024;44(7):953–964. doi:10.1002/jat.4591
171. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Original Research. *Frontiers in Environmental Science*. 2016;3:80. doi:10.3389/fenvs.2015.00080
172. Tan Z, Zhao Y, Zhou T, Lin K. Hi-MGT: A hybrid molecule graph transformer for toxicity identification. *J Hazard Mater*. Sep 5 2023;457:131808. doi:10.1016/j.jhazmat.2023.131808
173. Galati S, Di Stefano M, Martinelli E, et al. VenomPred: A Machine Learning Based Platform for Molecular Toxicity Predictions. *Int J Mol Sci*. Feb 14 2022;23(4)doi:10.3390/ijms23042105
174. de Sa AGC, Long Y, Portelli S, Pires DEV, Ascher DB. toxCSM: comprehensive prediction of small molecule toxicity profiles. *Brief Bioinform*. Sep 20 2022;23(5)doi:10.1093/bib/bbac337
175. Yuan Li Y, Chen L, Pu C, et al. Co-model for chemical toxicity prediction based on multi-task deep learning. *Mol Inform*. May 2023;42(5):e2200257. doi:10.1002/minf.202200257
176. Boonsom S, Chamnansil P, Boonseng S, Srisongkram T. ToxSTK: A multi-target toxicity assessment utilizing molecular structure and stacking ensemble learning. *Comput Biol Med*. Feb 2025;185:109480. doi:10.1016/j.compbiomed.2024.109480
177. Wu Z, Jiang D, Wang J, Hsieh CY, Cao D, Hou T. Mining Toxicity Information from Large Amounts of Toxicity Data. *J Med Chem*. May 27 2021;64(10):6924–6936. doi:10.1021/acs.jmedchem.1c00421
178. Yu X, Chen Y, Chen L, et al. GCLmf: A Novel Molecular Graph Contrastive Learning Framework Based on Hard Negatives and Application in Toxicity Prediction. *Mol Inform*. Jan 2025;44(1):e202400169. doi:10.1002/minf.202400169

179. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*. Jan 14 2018;9(2):513–530. doi:10.1039/c7sc02664a
180. Ebner AB, Christoph; Topf, Sonja; Luukkonen, Sohvi; Schimunek, Johannes; Klambauer, Günter. *Measuring AI Progress in Drug Discovery: A Reproducible Leaderboard for the Tox21 Challenge*. 2025. arXiv:2511.14744. https://arxiv.org/abs/2511.14744
181. Chen L, Huang T, Zhang J, et al. Predicting drugs side effects based on chemical-chemical interactions and protein-chemical interactions. *Biomed Res Int*. 2013;2013:485034. doi:10.1155/2013/485034
182. Yamanishi Y, Pauwels E, Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces. *J Chem Inf Model*. Dec 21 2012;52(12):3284–92. doi:10.1021/ci2005548
183. Amano Y, Honda H, Sawada R, et al. In silico systems for predicting chemical-induced side effects using known and potential chemical protein interactions, enabling mechanism estimation. *J Toxicol Sci*. 2020;45(3):137–149. doi:10.2131/jts.45.137
184. Troger F, Delp J, Funke M, et al. Identification of mitochondrial toxicants by combined in silico and in vitro studies – A structure-based view on the adverse outcome pathway. *Computational Toxicology*. 2020/05/01/ 2020;14:100123. doi:https://doi.org/10.1016/j.comtox.2020.100123
185. Li X, Li Z, Wu X, et al. Deep Learning Enhancing Kinome-Wide Polypharmacology Profiling: Model Construction and Experiment Validation. *J Med Chem*. Aug 27 2020;63(16):8723–8737. doi:10.1021/acs.jmedchem.9b00855
186. Li Z, Li X, Liu X, et al. KinomeX: a web application for predicting kinome-wide polypharmacology effect of small molecules. *Bioinformatics*. Dec 15 2019;35(24):5354–5356. doi:10.1093/bioinformatics/btz519
187. Ren Q, Qu N, Sun J, et al. KinomeMETA: meta-learning enhanced kinome-wide polypharmacology profiling. *Brief Bioinform*. Nov 22 2023;25(1)doi:10.1093/bib/bbad461
188. Liu J, Gui Y, Rao J, et al. In silico off-target profiling for enhanced drug safety assessment. *Acta Pharm Sin B*. Jul 2024;14(7):2927–2941. doi:10.1016/j.apsb.2024.03.002
189. Vishnoi S, Matre H, Garg P, Pandey SK. Artificial intelligence and machine learning for protein toxicity prediction using proteomics data. *Chem Biol Drug Des*. Sep 2020;96(3):902–920. doi:10.1111/cbdd.13701
190. Wang L, Wang N, Zhang W, et al. Therapeutic peptides: current applications and future directions. *Signal Transduct Target Ther*. Feb 14 2022;7(1):48. doi:10.1038/s41392-022-00904-4
191. Khan F, Niaz K, Abdollahi M. Toxicity of Biologically Active Peptides and Future Safety Aspects: An Update. *Curr Drug Discov Technol*. 2018;15(3):236–242. doi:10.2174/1570163815666180219112806
192. Mall R, Singh A, Patel CN, Guirimand G, Castiglione F. VISH-Pred: an ensemble of fine-tuned ESM models for protein toxicity prediction. *Brief Bioinform*. May 23 2024;25(4)doi:10.1093/bib/bbae270
193. Gupta S, Kapoor P, Chaudhary K, et al. In silico approach for predicting toxicity of peptides and proteins. *PLoS One*. 2013;8(9):e73957. doi:10.1371/journal.pone.0073957
194. Sharma N, Naorem LD, Jain S, Raghava GPS. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief Bioinform*. Sep 20 2022;23(5)doi:10.1093/bib/bbac174

195. Rathore AS, Choudhury S, Arora A, Tijare P, Raghava GPS. ToxinPred 3.0: An improved method for predicting the toxicity of peptides. *Comput Biol Med*. Sep 2024;179:108926. doi:10.1016/j.compbiomed.2024.108926
196. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*. Apr 12 2022;38(8):2102–2110. doi:10.1093/bioinformatics/btac020
197. Morozov V, Rodrigues CHM, Ascher DB. CSM-Toxin: A Web-Server for Predicting Protein Toxicity. *Pharmaceutics*. Jan 28 2023;15(2)doi:10.3390/pharmaceutics15020431
198. Zhu L, Fang Y, Liu S, Shen HB, De Neve W, Pan X. ToxDL 2.0: Protein toxicity prediction using a pretrained language model and graph neural networks. *Comput Struct Biotechnol J*. 2025;27:1538–1549. doi:10.1016/j.csbj.2025.04.002
199. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. Mar 17 2023;379(6637):1123–1130. doi:10.1126/science.ade2574
200. Cole TJ, Brewer MS. TOXIFY: a deep learning approach to classify animal venom proteins. *PeerJ*. 2019;7:e7200. doi:10.7717/peerj.7200
201. Pan X, Zuallaert J, Wang X, et al. ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics*. Jan 29 2021;36(21):5159–5168. doi:10.1093/bioinformatics/btaa656
202. Wei L, Ye X, Xue Y, Sakurai T, Wei L. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform*. Sep 2 2021;22(5)doi:10.1093/bib/bbab041
203. Wei L, Ye X, Sakurai T, Mu Z, Wei L. ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics*. Mar 4 2022;38(6):1514–1524. doi:10.1093/bioinformatics/btac006
204. Yu Q, Zhang Z, Liu G, Li W, Tang Y. ToxGIN: an In silico prediction model for peptide toxicity via graph isomorphism networks integrating peptide sequence and structure information. *Brief Bioinform*. Sep 23 2024;25(6)doi:10.1093/bib/bbae583
205. Moukheiber L, Mangione W, Moukheiber M, et al. Identifying Protein Features and Pathways Responsible for Toxicity Using Machine Learning and Tox21: Implications for Predictive Toxicology. *Molecules*. May 8 2022;27(9)doi:10.3390/molecules27093021
206. Guidance for industry : drug metabolism/drug interaction studies in the drug development process : studies in vitro. (United States) (1997).
207. Guidance for industry : in vivo drug metabolism/drug interaction studies--study design, data analysis, and recommendations for dosing and labeling. (United States) (1999).
208. Niu J, Straubinger RM, Mager DE. Pharmacodynamic Drug-Drug Interactions. *Clin Pharmacol Ther*. Jun 2019;105(6):1395–1406. doi:10.1002/cpt.1434
209. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug-drug and drug-food interactions. *Proc Natl Acad Sci U S A*. May 1 2018;115(18):E4304–E4311. doi:10.1073/pnas.1803294115
210. Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res*. Jan 5 2024;52(D1):D1265–D1275. doi:10.1093/nar/gkad976
211. Lee G, Park C, Ahn J. Novel deep learning model for more accurate prediction of drug-drug interaction effects. *BMC Bioinformatics*. Aug 6 2019;20(1):415. doi:10.1186/s12859-019-3013-0

212. Deng Y, Xu X, Qiu Y, Xia J, Zhang W, Liu S. A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics*. Aug 1 2020;36(15):4316–4322. doi:10.1093/bioinformatics/btaa501
213. Luo Q, Mo S, Xue Y, et al. Novel deep learning-based transcriptome data analysis for drug-drug interaction prediction with an application in diabetes. *BMC Bioinformatics*. Jun 11 2021;22(1):318. doi:10.1186/s12859-021-04241-1
214. Kim E, Nam H. DeSIDE-DDI: interpretable prediction of drug-drug interactions using drug-induced gene expressions. *J Cheminform*. Mar 4 2022;14(1):9. doi:10.1186/s13321-022-00589-5
215. Wang NN, Wang XG, Xiong GL, et al. Machine learning to predict metabolic drug interactions related to cytochrome P450 isozymes. *J Cheminform*. Apr 15 2022;14(1):23. doi:10.1186/s13321-022-00602-x
216. Briand E, Thomsen R, Linnet K, Rasmussen HB, Brunak S, Taboureau O. Combined Ensemble Docking and Machine Learning in Identification of Therapeutic Agents with Potential Inhibitory Effect on Human CES1. *Molecules*. Jul 29 2019;24(15)doi:10.3390/molecules24152747
217. Feng YH, Zhang SW, Shi JY. DPDDI: a deep predictor for drug-drug interactions. *BMC Bioinformatics*. Sep 24 2020;21(1):419. doi:10.1186/s12859-020-03724-x
218. Chen X, Liu X, Wu J. GCN-BMP: Investigating graph representation learning for DDI prediction task. *Methods*. Jul 1 2020;179:47–54. doi:10.1016/j.ymeth.2020.05.014
219. Chen Y, Ma T, Yang X, Wang J, Song B, Zeng X. MUFFIN: multi-scale feature fusion for drug-drug interaction prediction. *Bioinformatics*. Sep 9 2021;37(17):2651–2658. doi:10.1093/bioinformatics/btab169
220. Nyamabo AK, Yu H, Liu Z, Shi JY. Drug-drug interaction prediction with learnable size-adaptive molecular substructures. *Brief Bioinform*. Jan 17 2022;23(1)doi:10.1093/bib/bbab441
221. He C, Liu Y, Li H, et al. Multi-type feature fusion based on graph neural network for drug-drug interaction prediction. *BMC Bioinformatics*. Jun 10 2022;23(1):224. doi:10.1186/s12859-022-04763-2
222. Wang Y, Yang Z, Yao Q. Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. *Commun Med (Lond)*. Mar 28 2024;4(1):59. doi:10.1038/s43856-024-00486-y
223. Chen S, Semenov I, Zhang F, et al. An effective framework for predicting drug-drug interactions based on molecular substructures and knowledge graph neural network. *Comput Biol Med*. Feb 2024;169:107900. doi:10.1016/j.compbiomed.2023.107900
224. Kalscheur MM, Kipp RT, Tattersall MC, et al. Machine Learning Algorithm Predicts Cardiac Resynchronization Therapy Outcomes: Lessons From the COMPANION Trial. *Circ Arrhythm Electrophysiol*. Jan 2018;11(1):e005499. doi:10.1161/CIRCEP.117.005499
225. Modongo C, Pasipanodya JG, Zetola NM, Williams SM, Sirugo G, Gumbo T. Amikacin Concentrations Predictive of Ototoxicity in Multidrug-Resistant Tuberculosis Patients. *Antimicrob Agents Chemother*. Oct 2015;59(10):6337–43. doi:10.1128/AAC.01050-15
226. Oyaga-Iriarte E, Insausti A, Sayar O, Aldaz A. Prediction of irinotecan toxicity in metastatic colorectal cancer patients based on machine learning models with pharmacokinetic parameters. *J Pharmacol Sci*. May 2019;140(1):20–25. doi:10.1016/j.jphs.2019.03.004
227. Cuplov V, Andre N. Machine Learning Approach to Forecast Chemotherapy-Induced Haematological Toxicities in Patients with Rhabdomyosarcoma. *Cancers (Basel)*. Jul 17 2020;12(7)doi:10.3390/cancers12071944

228. Cao K, Yeung J, Wei MYK, et al. Improving the prediction of chemotherapy dose-limiting toxicity in colon cancer patients using an AI-CT-based 3D body composition of the entire L1-L5 lumbar spine. *Support Care Cancer*. Dec 21 2024;33(1):45. doi:10.1007/s00520-024-09108-8
229. Froicu EM, Oniciuc OM, Afrasanie VA, et al. The Use of Artificial Intelligence in Predicting Chemotherapy-Induced Toxicities in Metastatic Colorectal Cancer: A Data-Driven Approach for Personalized Oncology. *Diagnostics (Basel)*. Sep 19 2024;14(18)doi:10.3390/diagnostics14182074
230. Wysokinski A, Dreczka J. Clozapine Toxicity Predictor: Deep neural network model predicting clozapine toxicity and its therapeutic dose range. *Psychiatry Res*. Dec 2024;342:116256. doi:10.1016/j.psychres.2024.116256
231. Yang M, Ma J, Zhang C, et al. Multimodal data deep learning method for predicting symptomatic pneumonitis caused by lung cancer radiotherapy combined with immunotherapy. *Front Immunol*. 2024;15:1492399. doi:10.3389/fimmu.2024.1492399
232. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. Nov 1 2017;77(21):e104–e107. doi:10.1158/0008-5472.CAN-17-0339
233. Satam H, Joshi K, Mangrolia U, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology (Basel)*. Jul 13 2023;12(7)doi:10.3390/biology12070997
234. Ismail S, Essawi M. Genetic polymorphism studies in humans. *Middle East Journal of Medical Genetics*. 2012;1(2)
235. Yu Y, Werdyani S, Carey M, Parfrey P, Yilmaz YE, Savas S. A comprehensive analysis of SNPs and CNVs identifies novel markers associated with disease outcomes in colorectal cancer. *Mol Oncol*. Dec 2021;15(12):3329–3347. doi:10.1002/1878-0261.13067
236. Bose E, Paintsil E, Ghebremichael M. Minimum redundancy maximal relevance gene selection of apoptosis pathway genes in peripheral blood mononuclear cells of HIV-infected patients with antiretroviral therapy-associated mitochondrial toxicity. *BMC Med Genomics*. Dec 1 2021;14(1):285. doi:10.1186/s12920-021-01136-1
237. Nielsen RL, Wolthers BO, Helenius M, et al. Can Machine Learning Models Predict Asparaginase-associated Pancreatitis in Childhood Acute Lymphoblastic Leukemia. *J Pediatr Hematol Oncol*. Apr 1 2022;44(3):e628–e636. doi:10.1097/MPH.0000000000002292
238. Guo L, Wang W, Xie X, Wang S, Zhang Y. Machine learning for genetic prediction of chemotherapy toxicity in cervical cancer. *Biomed Pharmacother*. May 2023;161:114518. doi:10.1016/j.biopha.2023.114518
239. Arab A, Kashani B, Cordova-Delgado M, et al. Machine learning model identifies genetic predictors of cisplatin-induced ototoxicity in CERS6 and TLR4. *Comput Biol Med*. Dec 2024;183:109324. doi:10.1016/j.compbiomed.2024.109324
240. Lee JS, Paintsil E, Gopalakrishnan V, Ghebremichael M. A comparison of machine learning techniques for classification of HIV patients with antiretroviral therapy-induced mitochondrial toxicity from those without mitochondrial toxicity. *BMC Med Res Methodol*. Nov 27 2019;19(1):216. doi:10.1186/s12874-019-0848-z
241. Yu Z, Ji H, Xiao J, et al. Predicting Adverse Drug Events in Chinese Pediatric Inpatients With the Associated Risk Factors: A Machine Learning Study. *Front Pharmacol*. 2021;12:659099. doi:10.3389/fphar.2021.659099
242. Hu Q, Wu B, Wu J, Xu T. Predicting adverse drug events in older inpatients: a machine learning study. *Int J Clin Pharm*. Dec 2022;44(6):1304–1311. doi:10.1007/s11096-022-01468-7

243. Jiang D, Song Z, Liu P, Wang Z, Zhao R. A prediction model for severe hematological toxicity of BTK inhibitors. *Ann Hematol*. Oct 2023;102(10):2765–2777. doi:10.1007/s00277-023-05371-7
244. Zhu SY, Yang TT, Zhao YZ, Sun Y, Zheng XM, Xu HB. Interpretable machine learning model predicting immune checkpoint inhibitor-induced hypothyroidism: A retrospective cohort study. *Cancer Sci*. Nov 2024;115(11):3767–3775. doi:10.1111/cas.16352
245. Xiao Y, Xiao L, Zhang Y, et al. Prediction of tumor lysis syndrome in childhood acute lymphoblastic leukemia based on machine learning models: a retrospective study. *Front Oncol*. 2024;14:1337295. doi:10.3389/fonc.2024.1337295
246. Xiong Y, Liu X, Wang Q, et al. Machine learning-based prediction model for the efficacy and safety of statins. *Front Pharmacol*. 2024;15:1334929. doi:10.3389/fphar.2024.1334929
247. Zhang Z, Zhu M, Jiang W. Risk Factors Analysis of Cutaneous Adverse Drug Reactions Caused by Targeted Therapy and Immunotherapy Drugs for Oncology and Establishment of a Prediction Model. *Clin Transl Sci*. Jan 2025;18(1):e70118. doi:10.1111/cts.70118
248. Shi Z, Huang L, Wang H. Predictive modeling of pediatric drug-induced liver injury: Dynamic classifier selection with clustering analysis. *Digit Health*. Jan–Dec 2025;11:20552076251330078. doi:10.1177/20552076251330078
249. Wang Y, Lin X, Sun Y, et al. Development and Validation of a Novel Model to Discriminate Idiosyncratic Drug-Induced Liver Injury and Autoimmune Hepatitis. *Liver Int*. Feb 2025;45(2):e16239. doi:10.1111/liv.16239
250. Jamshidi E, Asgary A, Kharrazi AY, et al. Personalized predictions of adverse side effects of the COVID-19 vaccines. *Heliyon*. Jan 2023;9(1):e12753. doi:10.1016/j.heliyon.2022.e12753
251. Gong X, Hu M, Liu J, et al. Decoding kinase-adverse event associations for small molecule kinase inhibitors. *Nat Commun*. Jul 27 2022;13(1):4349. doi:10.1038/s41467-022-32033-5
252. Jusko NM, Cao A, Sun D. Machine learning model predicts clinical adverse events of small molecule kinase inhibitors in cancer patients using on-/off-target engagement and tissue selectivity. *Clin Pharmacol Ther*. 2026:In press.
253. Bender A, Scheiber J, Glick M, et al. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*. Jun 2007;2(6):861–73. doi:10.1002/cmdc.200700026
254. Hammann F, Gutmann H, Vogt N, Helma C, Drewe J. Prediction of adverse drug reactions using decision tree modeling. *Clin Pharmacol Ther*. Jul 2010;88(1):52–9. doi:10.1038/clpt.2009.248
255. Scheiber J, Jenkins JL, Sukuru SC, et al. Mapping adverse drug reactions in chemical space. *J Med Chem*. May 14 2009;52(9):3103–7. doi:10.1021/jm801546k
256. Huang LC, Wu X, Chen JY. Predicting adverse side effects of drugs. *BMC Genomics*. Dec 23 2011;12 Suppl 5(Suppl 5):S11. doi:10.1186/1471-2164-12-S5-S11
257. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*. May 18 2011;12:169. doi:10.1186/1471-2105-12-169
258. Liu M, Wu Y, Chen Y, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc*. Jun 2012;19(e1):e28–35. doi:10.1136/amiajnl-2011-000699

259. Lopes TJ, Shoemaker JE, Matsuoka Y, Kawaoka Y, Kitano H. Identifying problematic drugs based on the characteristics of their targets. *Front Pharmacol*. 2015;6:186. doi:10.3389/fphar.2015.00186
260. Spiro A, Fernandez Garcia J, Yanover C. Inferring new relations between medical entities using literature curated term co-occurrences. *JAMIA Open*. Oct 2019;2(3):378–385. doi:10.1093/jamiaopen/ooz022
261. Zhou H, Cao H, Matyunina L, et al. MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug Side Effects, Indications, Efficacy, and Modes of Action. *Mol Pharm*. May 4 2020;17(5):1558–1574. doi:10.1021/acs.molpharmaceut.9b01248
262. Wang CS, Lin PJ, Cheng CL, Tai SH, Kao Yang YH, Chiang JH. Detecting Potential Adverse Drug Reactions Using a Deep Neural Network Model. *J Med Internet Res*. Feb 6 2019;21(2):e11016. doi:10.2196/11016
263. Kang L, Duan Y, Chen C, et al. Structure-Activity Relationship (SAR) Model for Predicting Teratogenic Risk of Antiseizure Medications in Pregnancy by Using Support Vector Machine. *Front Pharmacol*. 2022;13:747935. doi:10.3389/fphar.2022.747935
264. Cao A, Zhang L, Bu Y, Sun D. Machine Learning Prediction of On/Off Target-driven Clinical Adverse Events. *Pharm Res*. Aug 2024;41(8):1649–1658. doi:10.1007/s11095-024-03742-x
265. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. Jan 4 2016;44(D1):D1075–9. doi:10.1093/nar/gkv1075
266. Galeano D, Li S, Gerstein M, Paccanaro A. Predicting the frequencies of drug side effects. *Nat Commun*. Sep 11 2020;11(1):4575. doi:10.1038/s41467-020-18305-y
267. Sharma B, Chenthamarakshan V, Dhurandhar A, et al. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. *Sci Rep*. Mar 25 2023;13(1):4908. doi:10.1038/s41598-023-31169-8
268. Yu L, Xu Z, Qiu W, Xiao X. MSDSE: Predicting drug-side effects based on multi-scale features and deep multi-structure neural network. *Comput Biol Med*. Feb 2024;169:107812. doi:10.1016/j.compbiomed.2023.107812
269. Farnoush A, Sedighi-Maman Z, Rasoolian B, Heath JJ, Fallah B. Prediction of adverse drug reactions using demographic and non-clinical drug characteristics in FAERS data. *Sci Rep*. Oct 9 2024;14(1):23636. doi:10.1038/s41598-024-74505-2
270. Zhang W, Yue X, Liu F, Chen Y, Tu S, Zhang X. A unified frame of predicting side effects of drugs by using linear neighborhood similarity. *BMC Syst Biol*. Dec 14 2017;11(Suppl 6):101. doi:10.1186/s12918-017-0477-2
271. Yu L, Cheng M, Qiu W, Xiao X, Lin W. idse-HE: Hybrid embedding graph neural network for drug side effects prediction. *J Biomed Inform*. Jul 2022;131:104098. doi:10.1016/j.jbi.2022.104098
272. Zhou F, Khushi M, Brett J, Uddin S. Graph neural network-based subgraph analysis for predicting adverse drug events. *Comput Biol Med*. Dec 2024;183:109282. doi:10.1016/j.compbiomed.2024.109282
273. Bean DM, Wu H, Iqbal E, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci Rep*. Nov 27 2017;7(1):16416. doi:10.1038/s41598-017-16674-x
274. Zhao H, Zheng K, Li Y, Wang J. A novel graph attention model for predicting frequencies of drug-side effects from multi-view data. *Brief Bioinform*. Nov 5 2021;22(6)doi:10.1093/bib/bbab239

275. Chen J, Han F, He M, Shi Y, Cai Y. A novel weighted pseudo-labeling framework based on matrix factorization for adverse drug reaction prediction. *BMC Bioinformatics*. Feb 17 2025;26(1):54. doi:10.1186/s12859-025-06053-z
276. Holzinger A, Carrington A, Muller H. Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *Kunstliche Intell (Oldenbourg)*. 2020;34(2):193–198. doi:10.1007/s13218-020-00636-z
277. Karim MR, Islam T, Shajalal M, et al. Explainable AI for Bioinformatics: Methods, Tools and Applications. *Brief Bioinform*. Sep 20 2023;24(5)doi:10.1093/bib/bbad236
278. Dickinson Q, Meyer JG. Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences. *PLoS Comput Biol*. Jan 2022;18(1):e1009736. doi:10.1371/journal.pcbi.1009736
279. Spinelli I, Scardapane S, Uncini A. A Meta-Learning Approach for Training Explainable Graph Neural Networks. *IEEE Trans Neural Netw Learn Syst*. Apr 2024;35(4):4647–4655. doi:10.1109/TNNLS.2022.3171398
280. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: Generating Explanations for Graph Neural Networks. Curran Associates, Inc.; 2019:9240–9251.
281. Wu Z, Wang J, Du H, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat Commun*. May 4 2023;14(1):2585. doi:10.1038/s41467-023-38192-3
282. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products: Draft Guidance for Industry (U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER)) (2025).
283. Huang K, Fu T, Gao W, et al. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. Curran Associates, Inc.; 2021:
284. Wu L, Yan B, Han J, et al. TOXRIC: a comprehensive database of toxicological data and benchmarks. *Nucleic Acids Res*. Jan 6 2023;51(D1):D1432–D1445. doi:10.1093/nar/gkac1074
285. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J Chem Inf Model*. Aug 24 2020;60(8):3770–3780. doi:10.1021/acs.jcim.0c00502
286. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J Chem Inf Model*. Jun 22 2020;60(6):2697–2717. doi:10.1021/acs.jcim.9b00975
287. Yang H, Xiu J, Yan W, et al. Large Language Models as Tools for Molecular Toxicity Prediction: AI Insights into Cardiotoxicity. *J Chem Inf Model*. Mar 10 2025;65(5):2268–2282. doi:10.1021/acs.jcim.4c01371
288. Park J PY, Song M, Park S, Lee D, Baek S, Kang J. CoTox: Chain-of-Thought-Based Molecular Toxicity Reasoning and Prediction. *arXiv preprint*. 2025;
289. Sicard J, Montastruc F, Achalme C, et al. Can large language models detect drug-drug interactions leading to adverse drug reactions? *Ther Adv Drug Saf*. 2025;16:20420986251339358. doi:10.1177/20420986251339358
290. Radha Krishnan RP, Hung EH, Ashford M, et al. Evaluating the capability of ChatGPT in predicting drug-drug interactions: Real-world evidence using hospitalized patient data. *Br J Clin Pharmacol*. Dec 2024;90(12):3361–3366. doi:10.1111/bcp.16275

291. Silberg J, Swanson K, Simon E, et al. UniTox: Leveraging LLMs to Curate a Unified Dataset of Drug-Induced Toxicity from FDA Labels. *medRxiv*. 2024:2024.06.21.24309315. doi:10.1101/2024.06.21.24309315
292. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*. 2022/12/01 2022;4(12):1256–1264. doi:10.1038/s42256-022-00580-7
293. Mendez-Lucio O, Nicolaou CA, Earnshaw B. MolE: a foundation model for molecular graphs using disentangled attention. *Nat Commun*. Nov 12 2024;15(1):9431. doi:10.1038/s41467-024-53751-y
294. Lu S, Gao Z, He D, Zhang L, Ke G. Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-Mol. *Nat Commun*. Aug 19 2024;15(1):7104. doi:10.1038/s41467-024-51321-w
295. Liu YD, Sirui; Zhou, Sheng; Fan, Wenqi; Tan, Qiaoyu. MolecularGPT: Open Large Language Model (LLM) for Few-Shot Molecular Property Prediction. *arXiv preprint arXiv:240612950*. 2024;
296. Pei Q, Zhang W, Zhu J, et al. BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations. Association for Computational Linguistics; 2023:1102–1123.
297. A MB, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. *Nat Mach Intell*. 2024;6(5):525–535. doi:10.1038/s42256-024-00832-8
298. Huang Y, Su X, Ullanat V, et al. Multimodal AI predicts clinical outcomes of drug combinations from preclinical data. *ArXiv*. Sep 24 2025;
299. 117th Congress. FDA Modernization Act of 2021. S 2952. Washington, DC: U.S. Government Publishing Office; 2021.
300. Sun D, Macedonia C, Chen Z, et al. Can Machine Learning Overcome the 95% Failure Rate and Reality that Only 30% of Approved Cancer Drugs Meaningfully Extend Patient Survival? *J Med Chem*. Sep 26 2024;67(18):16035–16055. doi:10.1021/acs.jmedchem.4c01684
301. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B*. Jul 2022;12(7):3049–3062. doi:10.1016/j.apsb.2022.02.002
302. Chambers J, Davies M, Gaulton A, et al. UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J Cheminform*. Jan 14 2013;5(1):3. doi:10.1186/1758-2946-5-3
303. Davis MI, Hunt JP, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol*. Oct 30 2011;29(11):1046–51. doi:10.1038/nbt.1990
304. Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J Chem Inf Model*. Mar 24 2014;54(3):735–43. doi:10.1021/ci400709d
305. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. Jan 23 2015;347(6220):1260419. doi:10.1126/science.1260419
306. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol*. Jun 2020;38(6):675–678. doi:10.1038/s41587-020-0546-8
307. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. Sep 11 2020;369(6509):1318–1330. doi:10.1126/science.aaz1776

308. Zhang K, Wang H. [Cancer Genome Atlas Pan-cancer Analysis Project]. *Zhongguo Fei Ai Za Zhi*. Apr 2015;18(4):219–23. doi:10.3779/j.issn.1009-3419.2015.04.02
309. Kass-Hout TA, Xu Z, Mohebbi M, et al. OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. *J Am Med Inform Assoc*. May 2016;23(3):596–600. doi:10.1093/jamia/ocv153
310. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279–90.

Chapter 2: Machine Learning Predicts Clinical Adverse Events of Small Molecule Kinase Inhibitors in Cancer Patients Using On-/Off-Target Engagement and Tissue Selectivity

2.1 Abstract

Adverse events (AEs) of small molecule kinase inhibitors (SMKIs) at therapeutic doses in cancer patients are largely unpredictable in phase I-III studies and clinical use, despite extensive preclinical toxicity testing under good laboratory practice conditions. To address this gap, we developed a machine learning (ML) framework to predict the occurrence and time to onset of clinical AEs caused by SMKIs using on-/off-target engagement and tissue/cell selectivity. The analysis included 1939 unique AEs from 3433 patients treated with 16 SMKIs. On-/off-target engagement was evaluated by linking the inhibition (Ki) constants and expression of 442 kinase targets to SMKI exposure, expressed as dose-normalized AUC across plasma and 36 tissues. For each AE, we constructed random survival forest models composed of ensembles of binary decision trees, evaluated predictive accuracy using the concordance index, and applied variable importance (VIMP) measures to identify kinase targets potentially responsible for tissue-specific AEs. The final models successfully predict the most common AEs (rash, nausea, fatigue, headache) along with the most severe hematological AEs (neutropenia, leukopenia, lymphopenia, thrombocytopenia, anemia). VIMP analyses highlighted previously unrecognized kinases potentially involved in tissue-specific AE profiles. External validation using data from a Phase II neratinib monotherapy trial demonstrated strong model performance, yielding Pearson correlation coefficients (PCCs) ≥ 0.87 between predicted and observed AE incidences. These findings show that integrating exposure, on-/off-target engagement, and tissue-specific selectivity enables robust prediction of the likelihood of SMKI-associated AEs for both blood- and organ-related toxicities, offering a scalable approach at both the patient- and population-level.

2.2 Introduction

Clinical toxicity, as measured by the occurrence of adverse events (AEs), both limits the practical use of approved oncology drugs in cancer patients and determines the success or failure of cancer drug development in phase I-III clinical trials. AEs at therapeutic doses occur in only a subset of treated patients and are consequently challenging to predict, yet for those who do experience them they lead to dose reductions, treatment discontinuation, and diminished quality of life. Toxicity drives approximately 30% of drug development failures and lack of efficacy, often traceable to dose-limiting toxicities that prevent optimal dosing, drives another 40%[1-3]. Preclinical GLP-compliant animal studies are rarely predictive of clinical AEs at therapeutic doses, which, along with long-standing ethical concerns, has prompted U.S. Food and Drug Administration (FDA) initiatives to identify and validate New Approach Methodologies (NAMs) as alternatives[4,5]. Artificial intelligence (AI) and machine learning (ML) approaches have increasingly been explored as NAMs for drug safety prediction, though most remain limited in their ability to predict patient-level clinical outcomes at therapeutic doses.

Prior ML approaches to drug safety prediction have followed two broad strategies that proved difficult to translate into clinical safety decisions. Compound-level screening uses chemical structure, molecular fingerprints, or predicted target interactions to flag potentially toxic molecules before clinical testing[6-15], but cannot capture the complex interplay of pharmacokinetics (PK), target binding, and physiological context that determines whether a compound causes toxicity in an individual patient. Patient-level approaches apply ML to electronic health records, laboratory values, and pharmacogenomics to predict individual AE risk[16-24], but typically treat AEs as binary classification endpoints rather than time-to-event outcomes, limiting their ability to inform when toxicities are likely to emerge. More mechanistic approaches have leveraged drug-target interaction networks and systems pharmacology to link off-target binding profiles to clinical AEs, though these typically lack patient-level exposure data and temporal resolution[14,15,25,26]. Random survival forests (RSF)[27] and related survival methods[28-30] address the temporal limitation by modeling censored time-to-event data, but have rarely been applied to drug toxicity[24,31]. One such application, Gong et al.[32], used RSF to predict AE onset for 16 kinase inhibitors using patient-level exposure and kinase binding affinities; their models performed well for hematologic toxicities but showed limited accuracy for organ-localized AEs

such as rash and diarrhea, likely reflecting the absence of tissue-specific drug distribution as a determinant of local target engagement and, consequently, organ-specific toxicity.

SMKIs have become central to targeted cancer therapy, with 94 compounds now FDA-approved across a wide range of malignancies[33]. The approximately 518 kinases encoded in the human genome share a highly conserved ATP binding site, and SMKIs that target this site inevitably inhibit dozens of off-target kinases at therapeutic concentrations[34]. This broad off-target activity produces toxicity patterns that vary widely, ranging from severe hematologic and cardiovascular events to the more common AEs affecting the skin, gastrointestinal tract, lungs, kidney, and thyroid[35]. Critically, plasma drug exposure does not necessarily reflect exposure in the tissues where these toxicities occur; tissue-level drug accumulation modulates local target engagement, which in turn drives organ-specific toxicity. We posit that predicting such AEs requires integrating tissue-specific drug distribution with kinome-wide binding profiles, and that current *in silico* frameworks relying solely on plasma exposure are unlikely to capture these relationships.

We developed an ML framework to predict both individual risk trajectories and population-level AE incidence over time for SMKIs. The model integrates patient-level PK, kinome-wide on-/off-target binding profiles, and tissue-specific drug distribution to capture both systemic and organ-localized toxicities. We trained and evaluated the model using individual AE records and exposure data from 3,433 patients enrolled in registrational trials of 16 FDA-approved SMKIs. Such a framework could support individualized safety monitoring, informed drug selection for patients at elevated risk of specific toxicities and generate trial-level safety forecasts for SMKIs in development, reducing the dose modifications and treatment discontinuations that compromise therapeutic outcomes.

2.3 Methods

2.3.1 Data collection

Preprocessed datasets containing exposure data and AE records for the 16 selected kinase inhibitors were reported in a previous study[32]. Further described in **Table 2-1**, the 16 SMKIs included are as follows: axitinib (Inlyta), bosutinib (Bosulif), crizotinib (Xalkori), dasatinib (Sprycel), erlotinib (Tarceva), gefitinib (Iressa), midostaurin (Rydapt), nintedanib (Ofev),

nilotinib (Tasigna), pazopanib (Votrient), ruxolitinib (Jakafi), sorafenib (Nexavar), sunitinib (Sutent), tofacitinib (Xeljanz), trametinib (Mekinist), and vandetanib (Caprelsa).

Table 2-1. Description of SMKIs used in AE model training and validation.

Drug Compound	Primary Target	Therapeutic Indication	Year of First Approval	Approval Procedure	Accessed Trial(s)*
Axitinib (Inlyta)	VEGFR et al.	RCC	2012	Standard review	NCT00678392
Bosutinib (Bosulif)	BCR-Abl	CML	2012	Standard review Orphan drug	NCT00261846
Crizotinib (Xalkori)	ALK	NSCLC (ALK+ or ROS1+)	2011	Priority review Orphan drug	NCT00585195
Dasatinib (Sprycel)	BCR-Abl et al.	CML	2006	Priority review Orphan drug	NCT00481247
Erlotinib (Tarceva)	EGFR	NSCLC	2004	Rolling review	NCT00556712
Gefitinib (Iressa)	EGFR	NSCLC (EGFR L858R mutation+)	2003	Standard review Orphan drug	NCT00206219
Midostaurin (Rydapt)	FLT3	AML, aggressive systemic mastocytosis	2017	Priority review Fast track & breakthrough therapy Orphan drug	NCT00782067 NCT03512197
Nintedanib (Ofev)	EGFR et al.	idiopathic pulmonary fibrosis	2014	Priority review Fast track & breakthrough therapy Orphan drug	NCT01335464 NCT01335477
Nilotinib (Tasigna)	BCR-Abl	CML	2007	Standard review Orphan drug	NCT00471497
Pazopanib (Votrient)	VEGFR	RCC, soft tissue sarcoma	2009	Standard review	NCT00334282
Ruxolitinib (Jakafi)	JAK	myelofibrosis	2011	Priority review Orphan drug	NCT01243944
Sorafenib (Nexavar)	VEGFR et al.	RCC, hepatocellular carcinoma, thyroid cancer	2005	Priority review Orphan drug	NCT00984282
Sunitinib (Sutent)	VEGFR et al.	RCC, GIST	2006	Priority review	NCT00077974

Drug Compound	Primary Target	Therapeutic Indication	Year of First Approval	Approval Procedure	Accessed Trial(s)*
Tofacitinib (Xeljanz)	JAK3	rheumatoid arthritis	2012	Standard review	NCT00814307 NCT00856544 NCT00853385 NCT00847613 NCT00960440
Trametinib (Mekinst)	MEK	melanoma	2013	Standard review Orphan drug	NCT01245062
Vandetanib (Zelboraf)	EGFR, VEGFR	medullary thyroid cancer	2011	Priority review Orphan drug	NCT00410761
External validation					
Neratinib (Nerlynx)	EGFR, HER2, HER4	HER2-overexpressed or amplified breast cancer	2017	Standard review Fast track	NCT00300781 NCT00777101

*Studies are referred to by their ClinicalTrials.gov Identifier. **RCC**: renal cell carcinoma; **CML**: chronic myelogenous leukemia; **NSCLC**: non-small cell lung cancer; **AML**: acute myeloid leukemia; **GIST**: gastrointestinal stromal tumor.

Adverse event records. AE terms were normalized and aggregated using the Medical Dictionary for Regulatory Activities (MedDRA, Version 21.0). AE severity was graded according to the Common Terminology Criteria for Adverse Events (CTCAE, Version 3.0). The full dataset includes 59,604 total AE records. The most frequent AEs modeled, their CTCAE grade definitions, and their overall incidence across the 16 inhibitors are shown in **Table 2-2**.

Table 2-2. Most common treatment-emergent AEs experienced by patients on small molecule kinase inhibitors. National Cancer Institute Common Terminology Criteria for Adverse Events version 5.0 descriptions for grades 1-4 are shown. Grade 3 is the most severe grade for nausea, fatigue, headache, weight loss, and PPES.

	Grade 1	Grade 2	Grade 3	Grade 4
diarrhea	• Increase of < 4 stools over baseline	• Increase of 4-6 stools per day over baseline	• Increase of ≥ 7 stools/day over baseline • Incontinence • Hospitalization indicated • Limiting self-care ADL	• Life-threatening consequences • urgent intervention indicated Gr 1 – 4: **43.1%**
rash	• Papules and/or pustules covering <10% BSA, which may or may not be associated with symptoms of pruritus or tenderness	• Papules and/or pustules covering 10-30% BSA, which may or may not be associated with symptoms of pruritus or tenderness • Associated with psychosocial impact • Limiting instrumental ADL	• Papules and/or pustules covering >30% BSA with moderate or severe symptoms • Limiting self-care ADL • Associated with local superinfection with oral antibiotics indicated	• Life-threatening consequences • urgent intervention indicated • Papules and/or pustules covering any % BSA, which may or may not be associated with symptoms of pruritus or tenderness and are associated with extensive superinfection with IV antibiotics indicated Gr 1 – 4: **26.3%**
nausea	• Loss of appetite without alteration in eating habits	• Oral intake decreased without significant weight loss, dehydration or malnutrition	• Inadequate oral caloric or fluid intake • Tube feeding, TPN, or hospitalization indicated	- Gr 1 – 3: **23.2%**
fatigue	• Fatigue relieved by rest	• Fatigue not relieved by rest • Limiting instrumental ADL	• Fatigue not relieved by rest, limiting self-care ADL	- Gr 1 – 3: **20.4%**
decreased appetite	• Loss of appetite without alteration in eating habits	• Oral intake altered without significant weight loss or malnutrition • Oral nutritional supplements indicated	• Associated with significant weight loss or malnutrition • Tube feeding or TPN indicated	• Life-threatening consequences • Urgent intervention indicated Gr 1 – 3: **18.7%**
headache	• mild pain	• moderate pain • limiting instrumental ADL	• severe pain • limiting self-care ADL	- Gr 1–3: **15.3%**
vomiting	• 1-2 episodes (separated by 5 minutes) in 24 hours	• 3-5 episodes (separated by 5 minutes) in 24 hours	• ≥ 6 episodes (separated by 5 minutes) in 24 hours; tube feeding, TPN, or hospitalization indicated	• life-threatening consequences Gr 1–4: **14.4%**
weight loss	• 5 to < 10% decrease in body weight from baseline; intervention not indicated	• 10 to < 20% decrease from baseline; nutritional support indicated	• ≥ 20% decrease from baseline; tube feeding or TPN indicated	- Gr 1–3: **13.8%**
hypertension	• pre-hypertension (systolic BP 120-139 mmHg or diastolic BP 80-89 mmHg)	• stage 1 hypertension (systolic 140-159 mmHg or diastolic BP 90-99 mmHg) recurrent or persistent (≥ 24 hours) • symptomatic increase by > 20 mmHg (diastolic) or to > 140/99 mmHg if previously WNL • medical intervention needed	• stage 2 hypertension (systolic BP ≥ 160 mmHg diastolic BP ≥ 100 mmHg) • > 1 drug or more intensive therapy than previously used indicated • medical intervention indicated	• life-threatening consequences (e.g., malignant hypertension, transient or permanent neurologic deficit, hypertensive crisis) • urgent intervention indicated Gr 1–4: **13.5%**
PPES	• minimum skin changes or dermatitis (e.g., erythema, edema, or hyperkeratosis) without pain	• skin changes (e.g., peeling, blisters, bleeding, edema, hyperkeratosis) with pain • limiting instrumental ADL	• severe skin changes (e.g., peeling, blisters, bleeding, edema, or hyperkeratosis) with pain • limiting self-care ADL	- Gr 1–4: **12.8%**

Notes: Instrumental ADL: preparing meals, shopping for groceries or clothes, using the telephone, managing money, etc. Self-care ADL: bathing, dressing and undressing, feeding self, using the toilet, taking medications, and not bedridden.
Abbreviations: ADL, activities of daily living; AE, adverse event; BSA, body surface area; Gr, grade; TPN, total parenteral nutrition; BP, blood pressure; WNL, within normal limits; PPES, palmar-plantar erythrodysesthesia syndrome.

Individual drug exposures. SMKI exposure was included at the patient level through computed measurements of total dose-normalized drug exposure at steady state (AUC_{SS}). All patients who received at least one dose and had at least one evaluable PK measurement were included in the final dataset. If a population PK study report was not included in an SMKI's New Drug Application (NDA) package, a population PK model was constructed to estimate patient-level drug exposure in NONMEM (Version 7.4, Icon Development Solutions, Ellicott City, MD).

Kinase target inhibition profiles. The dissociation constants (K_i) of the 16 SMKIs against 442 kinases, including disease-linked wild-type and mutant variants, were measured using a competitive binding assay and retrieved from a previous study[36]. Any missing K_i entries in the kinome panel (unmeasured drug–kinase pairs) were imputed at 10,000 nM (10 μM), corresponding to the no-inhibition default at the screening threshold; this is a known simplification rather than a formal missing-data model.

Tissue distribution data. The tissue–plasma partition coefficients (K_p) included in this analysis were derived from quantitative whole-body autoradiography studies of ^{14}C-labeled drug in rodents, which were reported in their NDA packages. The 36 organs or tissue types with non-missing ^{14}C-labeled data for the 16 SMKIs were bone marrow, lymph nodes, spleen, thymus, brown fat, white fat, brain, spinal cord, non-pigmented skin, pigmented skin, adrenal gland, pancreas, pituitary gland, thyroid, bile, kidney, liver, urinary bladder, eye, epididymis, preputial gland, prostate gland, seminal vesicles, testis, lung, exorbital lachrymal gland, harderian gland, intraorbital lachrymal gland, salivary gland, bone, heart, and skeletal muscle.

To account for the fact that kinase targets are differentially expressed across tissues, we also collected target expression data for kinases across organs and tissue types. The target expression levels were obtained from the Human Protein Atlas (version 24.0) as shown in our previous study[14,37]. The tissues and cell types included were narrowed to the 36 types also available in the K_p dataset. The target expression levels in each tissue or cell type were classified as 'Not detected' = 0, 'Low' = 1, 'Medium' = 2, 'High' = 3 for inclusion in the model.

2.3.2 Data preprocessing

All available clinical and non-clinical data were included in the final analysis dataset. Patient demographic information (age [years], sex [M, F]), individual drug exposure, kinase dissociation constants, and tissue distribution information all served as model inputs. The

outcome for each AE model was time from treatment initiation to first occurrence of that AE. Patients who did not experience the AE were right-censored at treatment end.

Tissue/cell selectivity. To quantify drug accumulation in tissue for all 16 SMKIs, the R-based package 'PKNCA' was applied to the ^{14}C-labeled tissue distribution data. For each tissue, AUC_{TLAST} was computed, and Equation 2.1 was then used to calculate the tissue–plasma partition coefficient for each tissue and SMKI:

$$K_p = AUC_{tissue}/AUC_{plasma} \tag{2.1}$$

On-/off-target engagement. The kinase inhibition effect by each of the 16 SMKIs was estimated from drug exposure and K_i values of each of the 16 SMKIs against 442 kinases in blood or various tissues, defined according to Equation 2.2:

$$kinase\ inhibition\ effect\ \propto\ AUC_{i,j}/K_{i,j,k} \tag{2.2}$$

where $AUC_{i,j}$ is the total drug exposure in the plasma over one dosing interval at steady state of kinase inhibitor *j* for patient *i* and $K_{i,j,k} = (K_{i,j,1}, K_{i,j,2}, \ldots, K_{i,j,k})$ is the vector of dissociation constants for kinase inhibitor *j* against *k* kinase targets. A similar expression was derived for the impact of kinase inhibition in the 36 various tissues:

$$kinase\ inhibition\ effect\ in\ tissue\ \propto\ AUC_{sys} \cdot K_{p,tissue}/K_i \tag{2.3}$$

where $K_{p,tissue}$ is the tissue-plasma partition coefficient for kinase inhibitor *j* in a given tissue.

Reporting odds ratio (ROR). ROR analysis was performed to assess whether specific AEs were disproportionately reported for each SMKI versus all other SMKIs. For each drug–AE pair, ROR was calculated as $ROR = ad / bc$, where *a* and *b* are the numbers of patients on the drug who did and did not experience the AE, respectively, and *c* and *d* are the corresponding numbers for patients not on the drug[38].

Principal component analysis (PCA). PCA is a linear dimensionality reduction technique that was applied separately to the kinase inhibition (K_i) dataset and the tissue distribution (K_p) dataset to visualize clustering patterns across compounds and to assess whether underlying structure in the predictors aligned with observed toxicity patterns. PCA was implemented using the 'prcomp' function in R (version 4.4.2); all features were standardized prior to decomposition, and the first two principal components were used for visualization.

2.3.3 Statistical framework and model development

The occurrence and time to onset of AEs were modeled with random survival forests (RSF)[27], a fully nonparametric ensemble method for right-censored survival data. RSF was selected over alternatives for three reasons specific to the structure of these data. First, the kinase predictor matrix is severely collinear: the first ten principal components of the K_i matrix capture approximately 90% of the variance across 442 kinases, and several AE endpoints have an events-to-predictor ratio below the conventional 10:1 threshold for stable Cox regression[39], precluding standard penalized Cox approaches without further dimension reduction that would discard biologically relevant variables. Second, several hazards are non-proportional in stratified Kaplan-Meier inspection, violating the central assumption of Cox regression. Third, RSF natively delivers the three quantities required by the analysis—patient-level conditional survival functions, trial-level cumulative incidence at a clinically relevant horizon, and predictor importance rankings—within a single ensemble.

For observed data (T_i, δ_i, x_i), the target functional is the conditional cumulative hazard $H(t|x) = -\log S(t|x)$. RSF estimates H by an ensemble of B survival trees, each grown on an independent bootstrap resample; the approximately e^{-1} (~37%) of observations excluded from each bootstrap form the out-of-bag (OOB) set used for internal error estimation. At each candidate node, m_{try} predictors are sampled at random and the split maximizing the two-sample log-rank statistic between daughter nodes is selected, following the survival-tree splitting criterion of Segal[40] and of LeBlanc and Crowley[41]. Within each terminal node ℓ, the cumulative hazard is estimated by the Nelson–Aalen estimator $\widehat{H_\ell}(t) = \sum d_j/Y_j$, the counting-process estimator of H, consistent and asymptotically normal under independent right-censoring[42,43]. Per-tree estimates are averaged to give $\widehat{H}(t|x) = B^{-1}\sum_b \widehat{H}^{(b)}(t|x)$, with predicted survival $\hat{S}(t|x) = e^{-\widehat{H}(t|x)}$; consistency of the RSF ensemble is established in Ishwaran and Kogalur (2010)[44].

A separate RSF was fit for each AE. The outcome was defined as time from treatment initiation to first occurrence of that AE at any CTCAE grade, following Gong et al.[32]; patients who did not experience the AE were right-censored at treatment end, and follow-up was administratively censored at 700 days. Hyperparameters were tuned by leave-one-out cross-validation on a representative subset of common AEs, optimizing median concordance index (C-index); the selected configuration was then frozen and applied uniformly across all AE-specific

models (**Table 2-3**). The split rule, number of trees, and number of random split points were tuned; *m*try and minimum terminal node size were left at their R package defaults.

Table 2-3. Hyperparameter final settings for RSF AE models.

Hyperparameter	Final setting
'splitrule'	'logrank'
'ntree'	200
'nsplit'	10
'mtry'	√p
'nodesize'	15

To evaluate whether tissue-specific pharmacology improves AE prediction beyond systemic exposure, we constructed three AE-specific RSF model variants using progressively enriched kinase feature representations. All three variants share the same baseline patient-level inputs (age, sex, and individual SMKI exposure). Model 1 is the simplest variant, using systemic kinase inhibition features only (AUC/K_i for each of the 442 profiled kinases) and is the direct analog of the specification in Gong et al.[32]. Model 2 extends Model 1 by adding tissue-weighted kinase inhibition features ($AUC \cdot K_p/K_i$), testing whether drug distribution to the tissue where each AE manifests improves prediction over plasma exposure alone. Model 3 extends Model 1 by adding tissue- and expression-weighted inhibition features ($AUC \cdot K_p \cdot expression/K_i$) to test whether further weighting tissue exposure by target abundance adds information. Models 2 and 3 are thus parallel extensions of a common baseline, each testing a distinct pharmacological hypothesis. An overview of data integration and the RSF modeling workflow is shown in **Figure 2-1**. Models were implemented using the "randomForestSRC" package in R (Version 3.6.0).

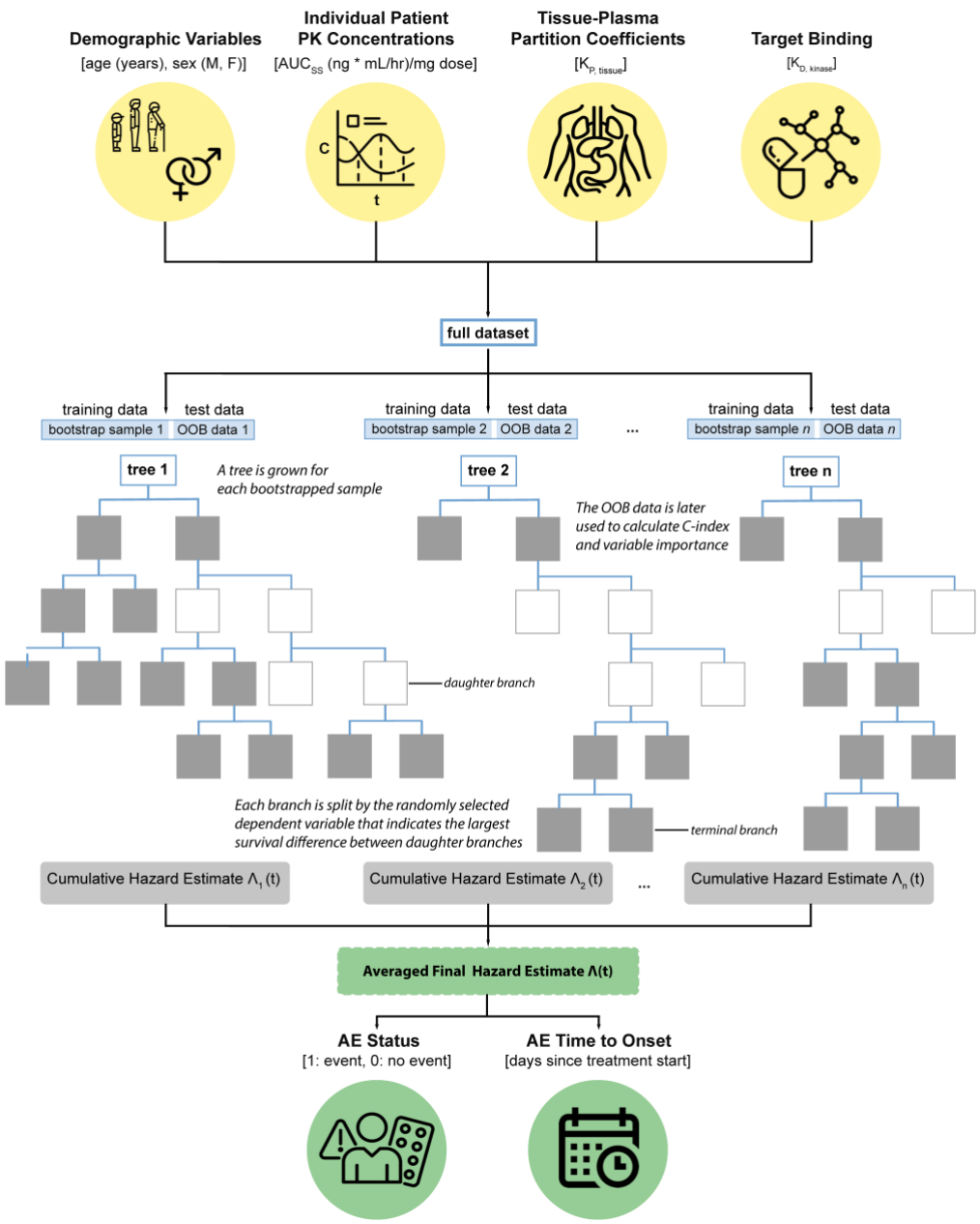


Figure 2-1. Graphic representation of random survival forest (RSF) modeling to predict adverse event (AE) status and time to AE onset of 16 small molecule kinase inhibitors (SMKIs) in 3,433 cancer patients.

Because Model 2 weights kinase inhibition by tissue-specific exposure, each AE was first assigned the tissue in which its toxicity manifests. AE terms standardized to MedDRA carry a System Organ Class (SOC), which was used to map each AE to the corresponding tissue in the Kp dataset (**Table 2-4**). For gastrointestinal AEs, stomach, large intestine, and small intestine were each evaluated as candidate tissues; small intestine Kp values gave the highest predictive

accuracy across all GI-related AEs and were carried forward. Hematologic and other systemic AEs were modeled without a tissue weighting.

Table 2-4. AE models constructed in this analysis and the organ system or tissue type associated with their occurrence. Stomach, large intestine, and small intestine were all tissue types considered as potentially associated with gastrointestinal (GI) disorders. Separate AE models were constructed using each GI tissue type, and the use of small intestine Kp values were found to produce the highest prediction accuracy across all GI-related AEs considered.

System Organ Class	Adverse Event	K_P Tissue Type
Blood and lymphatic system disorders; general disorders and administration site conditions; investigations	Anemia Hypertension Leukopenia Lymphopenia Neutropenia Thrombocytopenia	n/a (systemic)
Gastrointestinal disorders	Decreased appetite Diarrhea Nausea Vomiting Weight loss/ decreased weight	small intestine
Nervous system disorders	Fatigue Headache	brain (entire; avg)
Skin and subcutaneous tissue disorders	PPES Rash	skin (pigmented)

2.3.4 Statistical validation, inference, and variable importance

Model performance was evaluated at two levels that target different clinical use cases: patient-level discrimination of who will experience a given AE, and AE term-level classification of which toxicities will occur frequently in an unseen cohort.

Patient-level discrimination. Discrimination was quantified by the concordance index (C-index)[45,46], applied to RSF ensemble-mortality predictions on held-out test partitions. The C-index measures agreement between predicted risk scores and observed event-time ordering (0.5 = no discrimination, 1.0 = perfect ordering) and is the survival-analysis analog of the area under the ROC curve. C-indices were computed using the Noether-variance implementation in the 'survcomp' package (version 1.56.0); Harrell's C[45] and Uno's C[47] are defensible alternatives and

noted as sensitivity checks. Internal validation followed Gong et al.[32]: 500 iterations of random 80:20 train/test splitting of the full patient cohort without replacement; at each iteration a new forest was fit on the 80% training partition, and the C-index was computed on the held-out 20%. The median test-set C-index across the 500 iterations is reported as the point estimate and the empirical 5th and 95th percentiles as the 90% interval. Model 1 versus Model 2 was compared via the overlap of these marginal 90% CIs, matching the reporting convention of Gong et al.; a paired per-iteration ΔC on matched splits is the more powerful alternative and is a natural extension given that split seeds could be shared across the three model variants.

Variable importance. Per-kinase variable importance (VIMP)[48,49] was computed using the random-daughter-assignment procedure of Ishwaran et al., the default in randomForestSRC: for predictor X_k, OOB cases are routed to a randomly chosen daughter at any node that splits on X_k, and $VIMP_k$ is the resulting increase in OOB prediction error $(1 - C)$ relative to the unperturbed ensemble. Larger VIMP values indicate stronger associations between a predictor and the AE outcome. This statistic splits importance between collinear predictors and is downward-biased for each of them; conditional permutation importance[50] is the standard remedy and is noted as future work. To stabilize rankings, each AE model was refit ten times with independent random seeds and VIMP scores averaged across runs.

To provide biological face validity for the VIMP-ranked associations, a systematic PubMed search (1 January 2010 to 24 March 2025) was conducted for clinical reports linking kinase targets to the AEs modeled here, screened according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (**Figure 2-2**). Model-identified pairs supported by experimental or clinical evidence, with proposed mechanisms, are reported in **2.3.3** (**Table 2-5**).

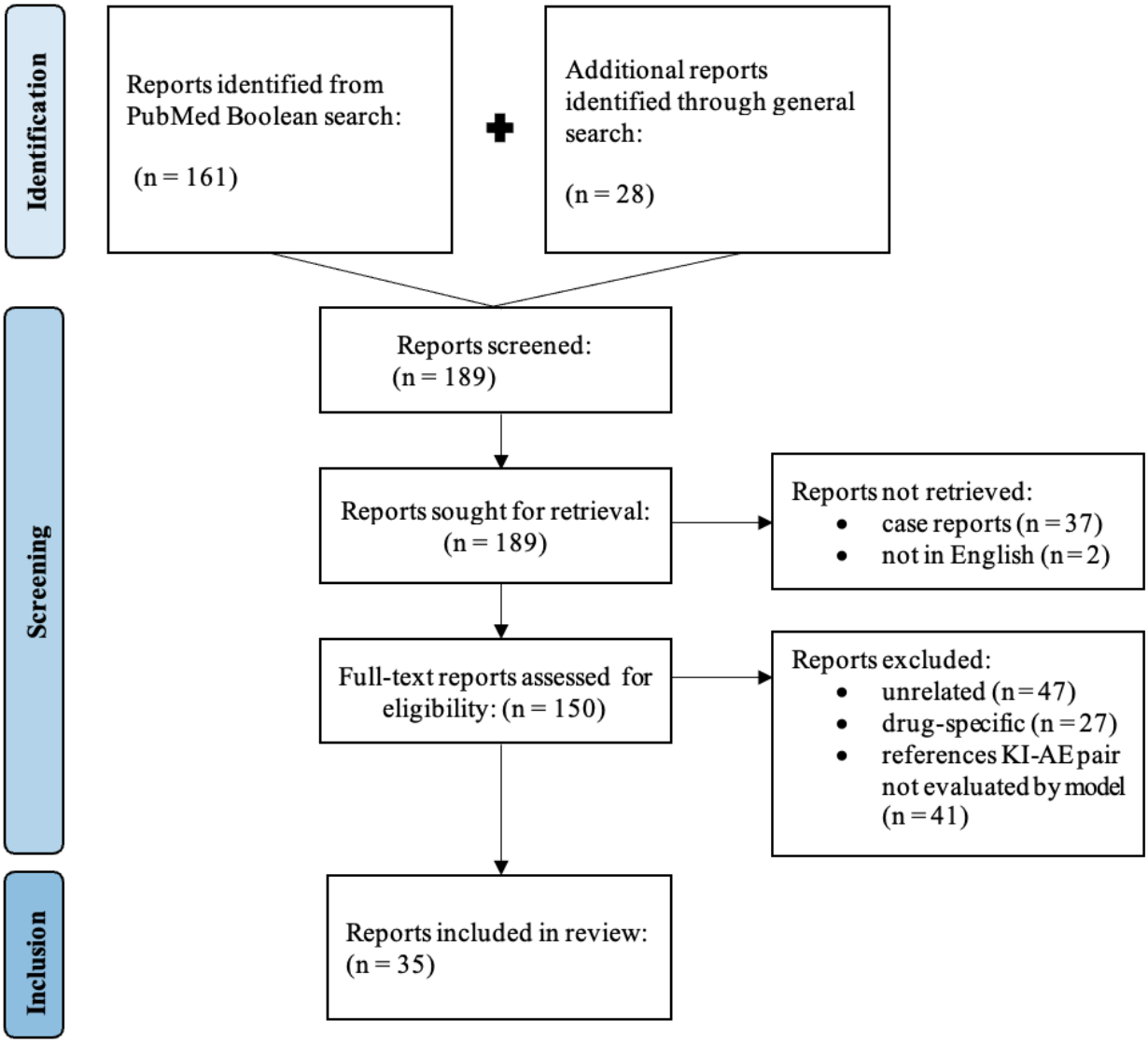


Figure 2-2. Flow chart depicting the process of clinical report selection for systematic review. The structure of the chart follows the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Data (PRISMA) 2020 flow diagram template.

External validation. External validation was performed on neratinib, an SMKI excluded from training and hyperparameter tuning, using exposure and safety data from a Phase II randomized trial of neratinib in 116 patients with HER2+ advanced breast cancer at therapeutic doses of 240 mg/day[51]. Patient-level discrimination was assessed by C-index under leave-one-out cross-validation. Trial-level AE classification, distinct from patient-level discrimination, was assessed by averaging patient-level cumulative incidence predictions at 229 days (the mean treatment duration) across the cohort to yield a predicted population-level incidence for each AE, which was then compared against the observed trial incidence as a measure of calibration.

Pearson correlation between predicted and observed incidences was computed on AEs observed in ≥ 10% of patients and on AEs the model predicted to occur in ≥ 10%; sensitivity and specificity of AE classification against the 10% clinical reporting threshold were computed across the full AE panel.

2.3.5 Statistical assumptions and limitations

Four assumptions and limitations of the statistical framework warrant explicit discussion.

Drug-level clustering and manifold coverage. Patients are nested within 16 drugs, and the 500-iteration 80:20 procedure splits at the patient level rather than by any clustering structure in the predictor space. Because the 16 drugs are a non-random, mechanistically clustered sample of pharmacological feature space (the first ten PCs of the K_i matrix capture ~90% of the variance), random patient-level resampling cannot probe extrapolation to underrepresented regions of that space. Leave-one-drug-out cross-validation across the 16 training drugs or leave-one-cluster-out in pharmacological feature space, would provide a more stringent test of extrapolation than the single-drug neratinib validation alone, and is a natural extension.

No Cox benchmark. RSF was chosen over penalized Cox regression because of expected nonlinear and interactive kinase-AE relationships, but proportional-hazards violations were assessed informally via stratified incidence curves rather than a Schoenfeld residual test[39], and no penalized Cox baseline was fit[52]. A future sensitivity analysis fitting a penalized Cox model with elastic-net regularization on the same feature spaces, even if expected to underperform, would more rigorously substantiate the choice of nonparametric ensemble.

VIMP is associative, not causal. Two distinct confounding concerns apply to the kinase–AE associations identified by VIMP.

Kinase-level confounding: a kinase correlated with the true causal driver can carry the predictive signal by proxy, and the forest has no basis for distinguishing the two because it selects splits by maximizing the log-rank statistic rather than by identifying mechanism, a concern heightened by the severe collinearity of the K_i matrix.

Confounding by indication: the 16 SMKIs share few common indications, and no common control arm exists, so background rates of events such as fatigue, nausea, diarrhea, anemia, and headache in advanced-cancer populations are attributed to the SMKI the patient was taking. A kinase preferentially inhibited by drugs used in an indication with a high background AE rate can therefore receive a high VIMP for that AE without any mechanistic role. Literature

corroboration via the systematic review provides biological face validity rather than a formal statistical test against either failure mode, and model-identified pairs should be treated as hypothesis-generating.

Multiplicity. Fitting a separate RSF per AE produces many implicit comparisons on VIMP rankings with no formal family-wise error rate or false discovery rate correction. Because VIMP is used here for ranking rather than binary significance decisions, literature concordance serves as the operative check; a permutation null under outcome shuffling[53] is the standard formal alternative and a natural extension. Multi-task learning across AE endpoints, in which a single model jointly predicts incidence across multiple AEs, addresses this concern at the level of model architecture and is the focus of subsequent work.

2.4 Results

2.4.1 Linking on-/off-target engagement and tissue selectivity to clinical adverse events for small molecule kinase inhibitors

To examine the clinical AE profiles of the 16 SMKIs in relation to their primary targets, we first performed an ROR analysis linking the ten most common AEs, their relative reporting frequencies, and the primary target of each compound (**Figure 2-3A**). SMKIs sharing a primary target did show some overlap in AE profile: the EGFR inhibitors erlotinib and gefitinib were both associated with elevated risk of rash, and the JAK inhibitors ruxolitinib and tofacitinib share an elevated risk of headache. Most SMKIs, however, produced broad and heterogeneous AE profiles that could not be attributed to their primary target alone, implicating off-target kinase inhibition and tissue-specific exposure as additional determinants of clinical toxicity.

The binding landscapes of the 16 SMKIs against 442 kinases (**Figure 2-3B**) confirmed extensive off-target activity at therapeutic concentrations. Plasma concentrations of these drugs in cancer patients typically range from 1 to 5 μM, and each SMKI was found to inhibit dozens of kinase targets with K_i values below 0.5 μM, consistent with simultaneous *in vivo* inhibition of many off targets. PCA of the K_i matrix across all 442 kinases showed that these engagement profiles are high-dimensional: the first two principal components explained only ~40% of the variance (PC1: 27.0%, PC2: 13.4%), and patients stratified by AE status showed substantial overlap between class centroids with no reduction in within-group variance (**Figure 2-4**). On-

/off-target binding alone therefore has limited power to discriminate patients who will and will not experience specific clinical AEs.

Tissue–plasma partition coefficients across the 36 profiled tissues (**Figure 2-3C**) showed that tissue exposure is equally heterogeneous, varying widely across both tissues and compounds. Many SMKIs showed elevated exposure in tissues clinically associated with common toxicities, including skin, gastrointestinal organs, liver, kidney, and bone marrow. PCA on the K_p matrix captured more of the variance in the first two components (PC1: 52.6%, PC2: 12.2%) but again showed tight clustering with minimal separation between AE groups (**Figure 2-5)**. The dispersed, high-dimensional structure of both feature spaces and the absence of class separation in either indicates that SMKI-driven clinical AEs reflect interactions between off-target binding and tissue-specific accumulation that no single feature class can capture in isolation.

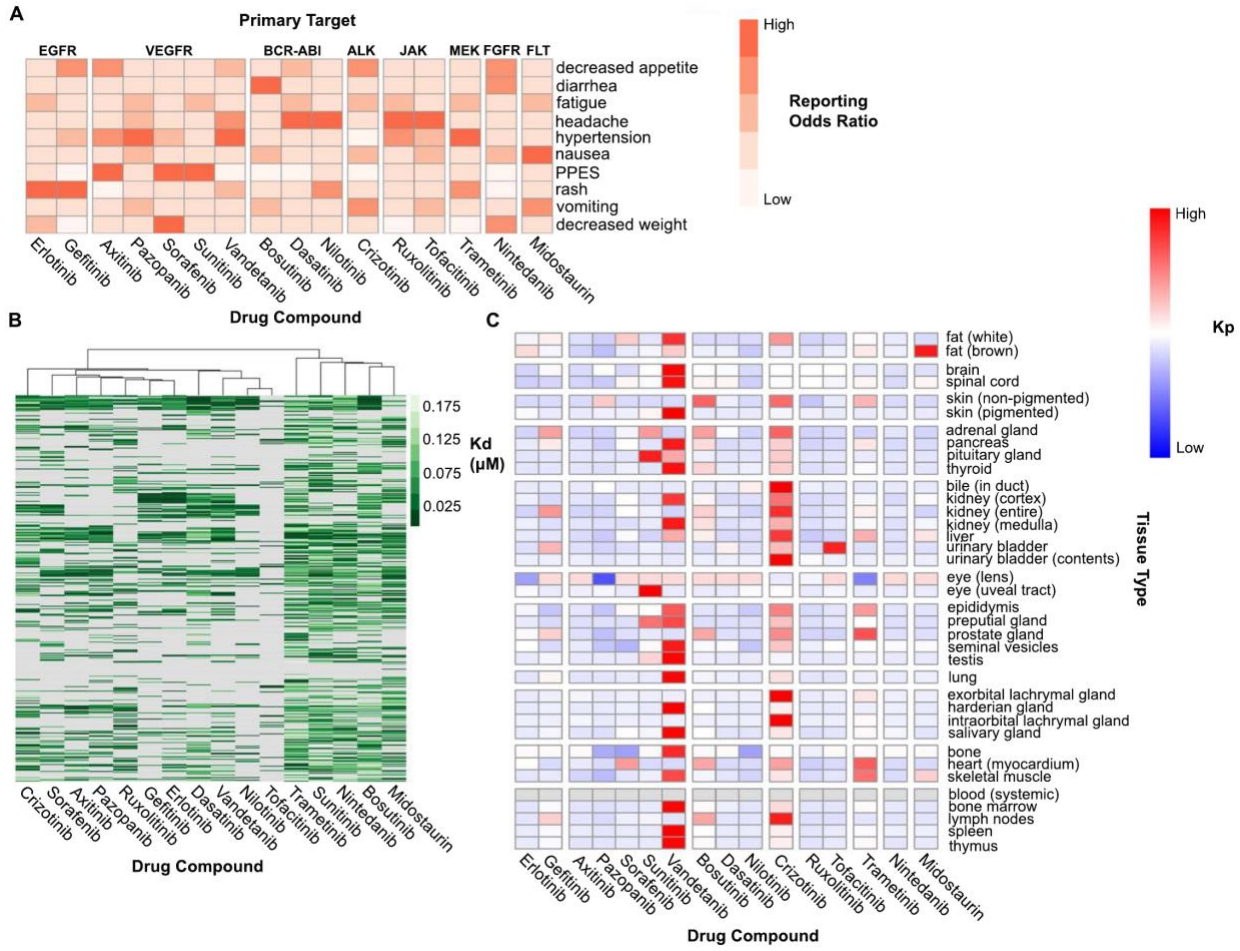


Figure 2-3. Adverse events (AEs), on-/off-target binding constants (K_i), and tissue selectivity of 16 SMKIs. Normalized reporting odds ratio (ROR) profiles for the top ten AEs common to all 16 SMKIs in the dataset. SMKIs

are grouped and labeled by primary target. (**B**) Heatmap of K_i values for 16 SMKIs across 442 kinase targets. Darker green cells indicate a lower K_i or higher binding affinity, and lighter green cells indicate a higher K_i or weaker binding affinity. Gray cells indicate combinations that were evaluated but for which binding was weak (K_i > 10 μM) or not detected in a 10 μM primary screen. (**C**) Heatmap of tissue–plasma partition coefficients (K_p) for 16 SMKIs in 36 tissues. SMKIs are grouped and labeled by primary target. Blue cells indicate a low K_p value ($0 < K_p < 1$). White or gray cells indicate a K_p value close to 1. Red cells indicate $K_p > 1$.

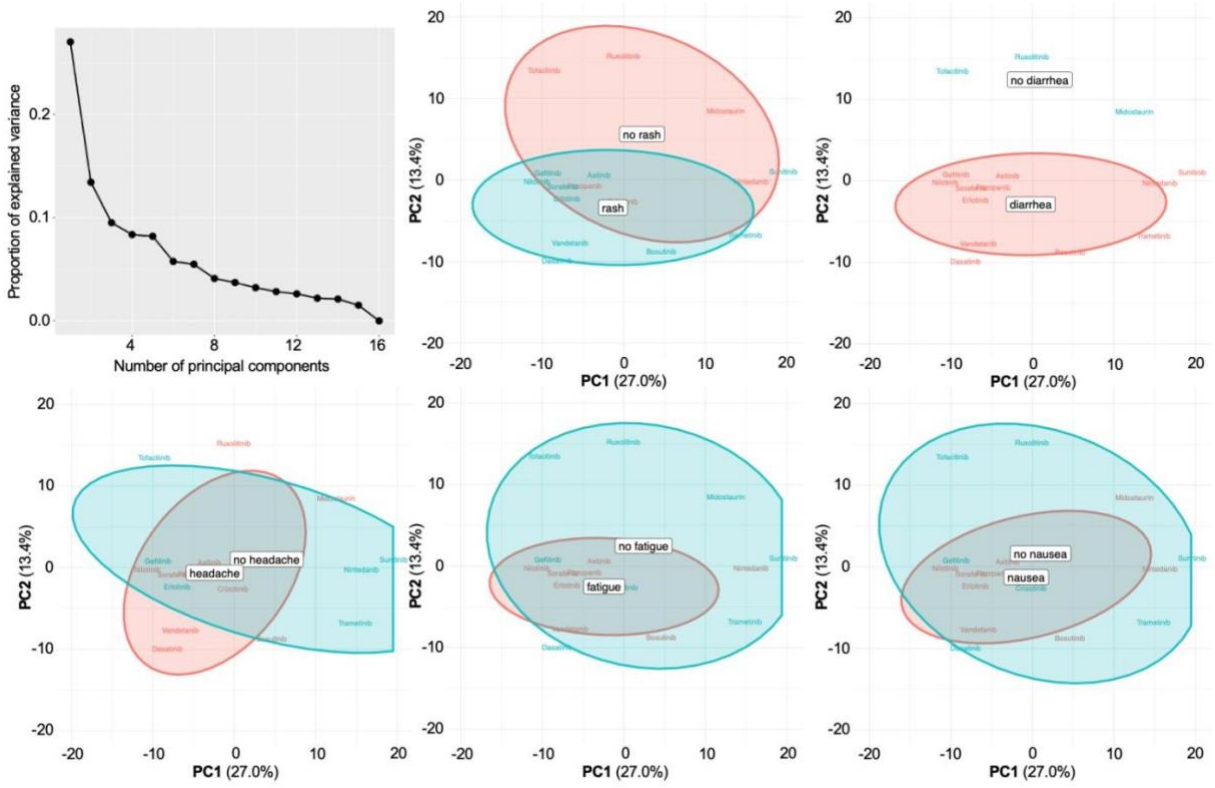


Figure 2-4. Principal Component Analysis (PCA) of n = 442 kinase target K_i values, stratified by occurrence or non-occurrence of AEs. The top left panel is a scree plot visualizing the importance of each principal component (PC). The first 10 PCs contain almost 90% of the total information in the Ki dataset. The remainder of the panels are biplots which visualize the similarities and dissimilarities between drug compounds, stratified by occurrence or non-occurrence of a specific AE, as captured by the first two PCs. Drug compounds that are grouped together are associated with Ki values that are positively correlated. Moving clockwise, the AEs depicted are rash, diarrhea, nausea, fatigue, and headache.

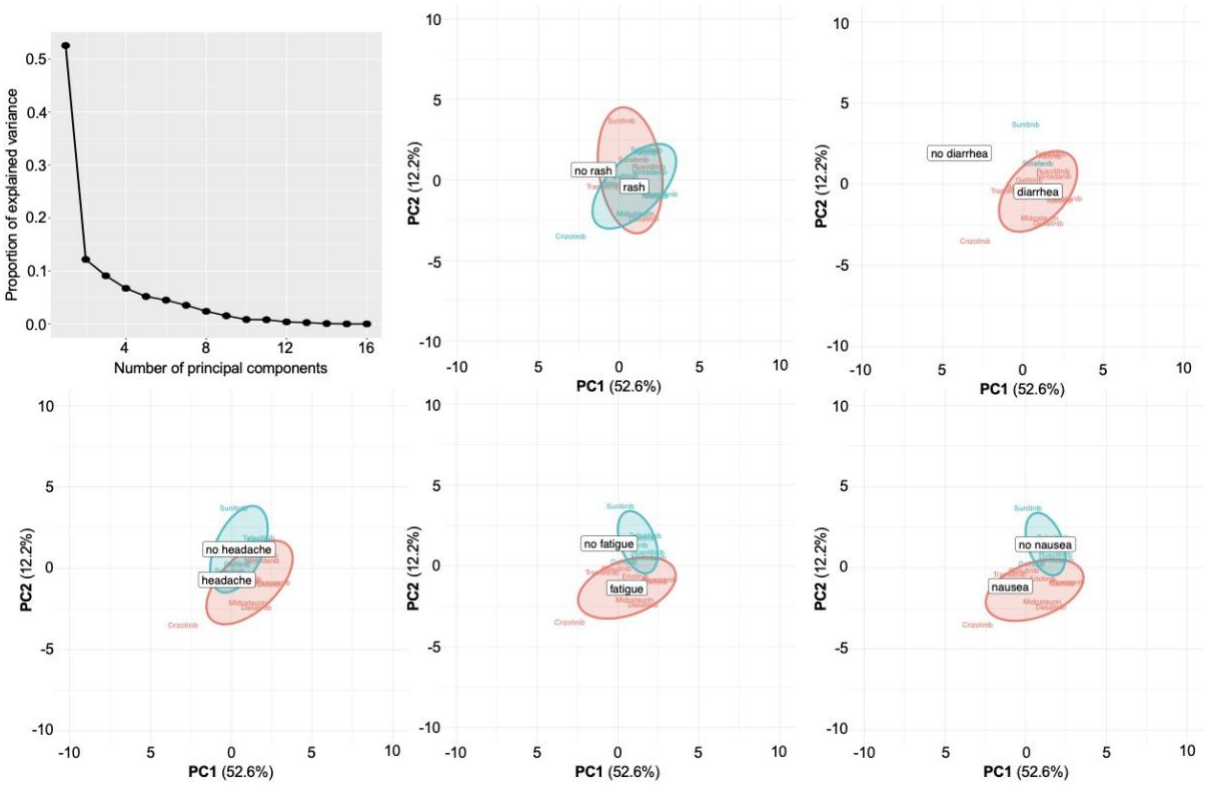


Figure 2-5. Principal Component Analysis (PCA) of n = 36 tissue K_p values, stratified by occurrence or non-occurrence of AEs. The top left panel is a scree plot visualizing the importance of each principal component (PC). The first 5 PCs contain almost 90% of the total information in the Kp dataset. The remainder of the panels are biplots which visualize the similarities and dissimilarities between drug compounds, stratified by occurrence or non-occurrence of a specific AE, as captured by the first two PCs. Drug compounds that are grouped together are associated with Kp values that are positively correlated. Moving clockwise, the AEs depicted are rash, diarrhea, nausea, fatigue, and headache.

2.4.2 Machine learning models incorporating on-/off-target engagement and tissue selectivity improve tissue-specific adverse event predictions

We compared the predictive performance of Models 1 and 2 across AE endpoints associated with different organ systems. Model 2, which incorporates tissue-specific drug distribution, achieved the highest predictive accuracy as measured by C-index for four of the most common clinical toxicities, including rash, nausea, fatigue, and headache (**Table 2-5**). Model 2 improved median C-index by 0.03-0.07 over Model 1, with the largest gain observed for headache (from 0.601 to 0.669). Because severe hematologic toxicities arise from SMKI effects on circulating hematopoietic cells in the blood, plasma exposure rather than tissue-weighted exposure is the pharmacologically relevant driver, and we therefore evaluated these endpoints using Model 1 alone. Model 1 achieved strong predictive accuracy for severe hematologic AEs (median C-index 0.694-0.880; **Table 2-6**), with the highest performance for lymphopenia (0.880)

and thrombocytopenia (0.830). Together, these results identify the AE classes for which tissue-aware modeling adds predictive value beyond plasma exposure alone.

Table 2-5. Predictive accuracy of Model 1 (systemic kinase inhibition) versus Model 2 (systemic plus tissue-weighted kinase inhibition) for the most common tissue-localized AEs, reported as median concordance index (C-index) with 90% intervals.

Adverse event (% patients affected)	Model 1 – Median C-index (90% CI)	Model 2 – Median C-index (90% CI)
Rash (26.3)	0.781 (0.765, 0.797)	0.814 (0.799, 0.824)
Nausea (23.2)	0.708 (0.685, 0.731)	0.772 (0.745, 0.799)
Fatigue (20.4)	0.744 (0.734, 0.761)	0.799 (0.774, 0.836)
Headache (15.3)	0.601 (0.578, 0.619)	0.669 (0.618, 0.697)

Table 2-6. Predictive accuracy of Model 1 (systemic kinase inhibition) for severe (grade ≥ 3) hematologic AEs, reported as median concordance index (C-index) with 90% intervals. Model 2 was not applied to hematologic endpoints because plasma exposure, rather than tissue-weighted exposure, is the pharmacologically relevant driver of toxicity to circulating hematopoietic cells.

Adverse event (% patients affected)	Median C-index (90% CI)
Anemia (7.8)	0.694 (0.667, 0.737)
Thrombocytopenia (6.5)	0.830 (0.774, 0.854)
Neutropenia (4.2)	0.772 (0.749, 0.807)
Leukopenia (2.7)	0.794 (0.729, 0.831)
Lymphopenia (0.9)	0.880 (0.776, 0.909)

2.4.3 Variable importance analysis reveals target-adverse event associations across organs

To identify kinases likely mediating tissue-specific AEs, we performed VIMP analysis across AE-specific RSF models (**Figures 2-6** and **2-7**). For each AE, cumulative incidence curves were stratified by the top-ranked kinase predictor. Validated kinase-AE associations from a systematic literature survey (**Figure 2-2**) are highlighted in bold, with supporting references and proposed mechanisms summarized in **Table 2-7**.

Table 2-7. Subset of literature survey results with reported kinase-associated toxicities that match model identified KI-AE pairs.

System Organ Class	Adverse Event	Kinase	Reference	Proposed Mechanism of Action
Skin and subcutaneous tissue disorders	Rash	EGFR (ErbB-1)	[54-71]	Disruption of EGFR signaling leads to an inflammatory response characterized by the release of chemokines and the infiltration of inflammatory cells in the skin.
		ErbB family	[72,73]	The ErbB family is highly structurally related to EGFR, its first discovered member.
Gastrointestinal disorders	Diarrhea	EGFR (ErbB-1)	[54,59-61,63,64,66,68,72-75]	EGFR inhibition affects the physiologic processes of normal proliferation and repair of cells with rapid turnover that line the alimentary tract; also thought to result from excess chloride secretion.
		ErbB family	[73]	The ErbB family is highly structurally related to EGFR, its first discovered member.
		VEGFR	[54,63,74-81]	VEGFR inhibition causes perturbations of blood flow in the intestinal mucosa that resemble ischemia and hypoxia, leading to diarrhea.
		FLT	[63,82,83]	While its primary role is in blood cell development, inhibition of FLT might also affect intestinal epithelial cells and mucosa, impairing the absorption of water and other substances.
		AXL	[80,82,84]	AXL signaling aids in the regulation of intestinal immune response and its disruption can lead to

				reduced antimicrobial peptides, an overexpression of inflammatory cytokines, and dysbiosis.
		BCR-ABL1	[54,61]	-
		MEK	[54,85]	MEK and its downstream effectors are suggested to help regulate the claudin dependent assembly of functional tight junctions in intestinal epithelial cells; their dysregulation leads to intestinal toxicity.
		AURK	[54]	-
		PI3K	[54,86-88]	Deregulation of T-cells by inhibition of PI3K produces downstream effects on colonic epithelial cells.

For rash, both Models 1 and 2 identified EGFR inhibition as a top predictor; Model 2 further distinguished EGFR variants associated with drug accumulation in skin, tending to rank these more highly than variants reflecting systemic exposure alone. Model 1 prioritized EGFR G719C and Model 2 prioritized skin-localized EGFR S752_I759del; both showed clear separation between high and low inhibition groups in the corresponding cumulative incidence curves, with higher inhibition associated with greater and earlier risk of rash (**Figure 2-6B**). ERBB4 and ABL1 also ranked among the top predictors.

For fatigue, both models ranked FLT1 and AURKB among top predictors. Model 2 prioritized brain-localized variants, with brain-localized PLK4 emerging as the top predictor. Cumulative incidence curves stratified by the top-ranked predictor showed clear separation for both models, but stratification was greater and onset earlier for Model 2 when patients were grouped by brain-localized PLK4 inhibition (**Figure 2-6A**).

For diarrhea, the models identified multiple kinases with literature-supported gastrointestinal toxicity associations, including EGFR, VEGFR, and FLT[75,89,90]. Model 1 prioritized FLT3 and EGFR, while Model 2 ranked GI-localized predictors more highly overall. Cumulative incidence curves stratified by the top-ranked predictor showed clearer separation for

Model 2, with higher inhibition of AXL associated with greater and earlier risk of diarrhea (**Figure 2-7A**). Additional VIMP analyses and cumulative incidence plotting for vomiting are provided in **Figure 2-7B**.

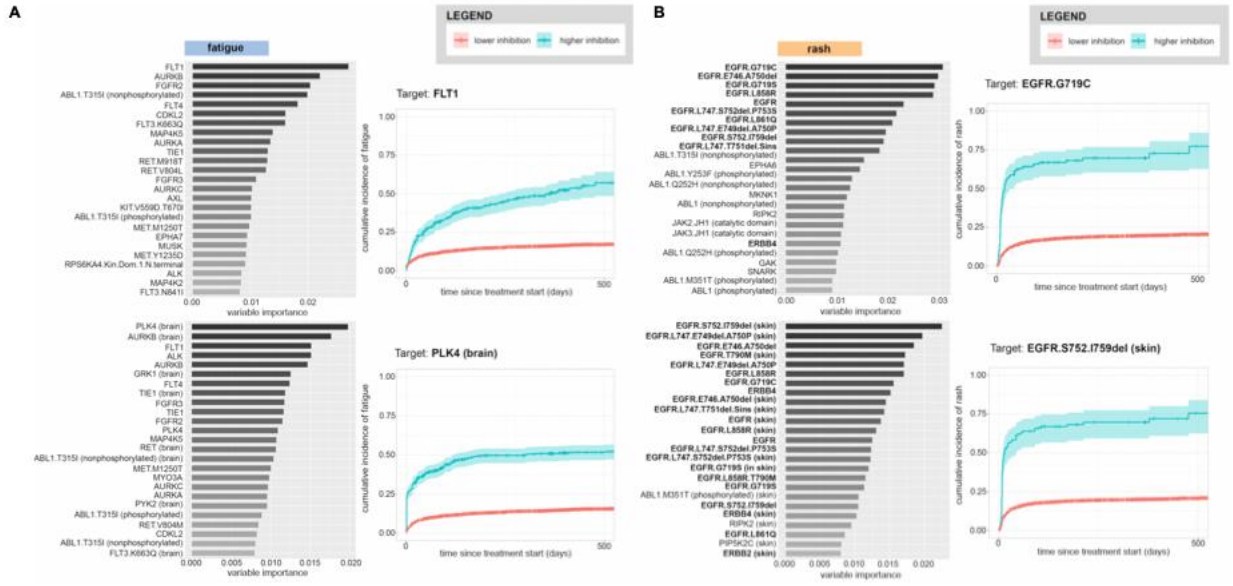


Figure 2-6. Final ML model with on-/off-targets and tissue selectivity of SMKIs identified novel kinase targets associated with fatigue and rash. VIMP assessments (left columns) of n = 444 predictive variables and cumulative incidence curves (right columns) stratified by top kinase target for two representative AE model runs. Each point on the cumulative incidence curve is the average of the ensemble survival function over all patients for a given time, with error bands showing 95% pointwise confidence intervals for all patients together. (**A**) Kinases associated with fatigue. (**B**) Kinases associated with rash. Top panel of each subfigure displays VIMP assessments and cumulative incidence curves of ML Model 1 using on-/off-target and plasma drug exposure to identify possible kinases associated with fatigue or rash. Bottom panel displays the corresponding output of ML Model 2 using on-/off-target and tissue drug exposure. VIMP scores for each kinase were computed in each AE-specific model and averaged across 10 iterations to produce a stable estimate of each variable's contribution to prediction. Kinases shown in bold within the VIMP plots were validated by experimental or clinical findings identified in the literature survey.

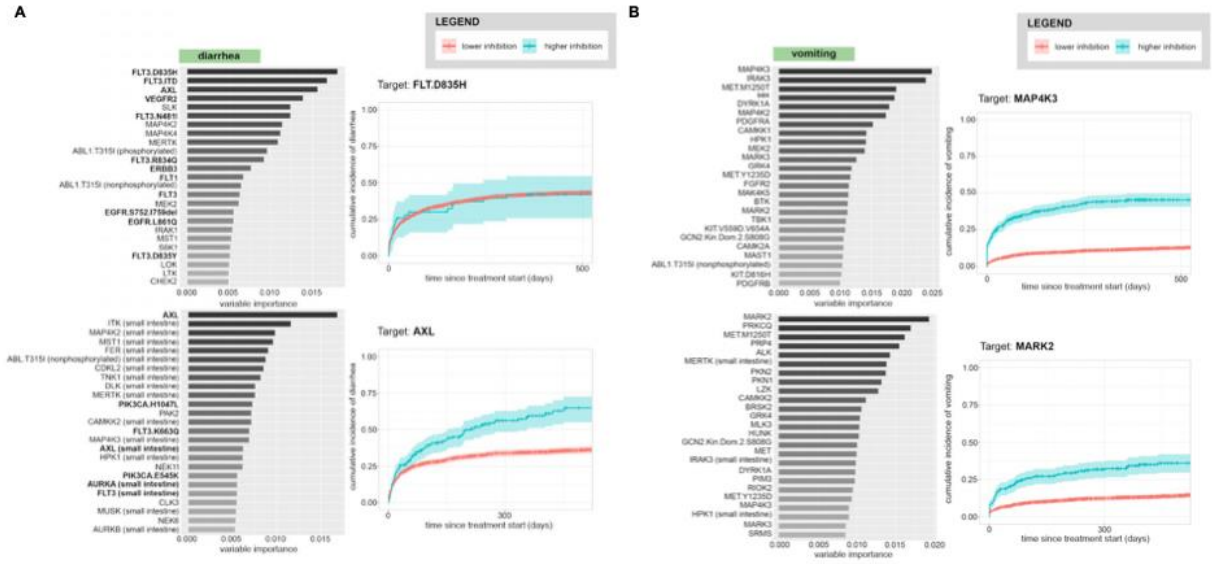


Figure 2-7. Final ML model with on-/off-targets and tissue selectivity of SMKIs identified novel kinase targets associated with diarrhea and vomiting. VIMP assessments (left columns) of n = 444 predictive variables and cumulative incidence curves (right columns) stratified by top kinase target for two representative AE model runs. Each point on the cumulative incidence curve is the average of the ensemble survival function over all patients for a given time, with error bands showing 95% pointwise confidence intervals for all patients together. (**A**) Kinases associated with diarrhea. (**B**) Kinases associated with vomiting. Top panel of each subfigure displays VIMP assessments and cumulative incidence curves of ML Model 1 using on-/off-target and plasma drug exposure to identify possible kinases associated with fatigue or rash. Bottom panel displays the corresponding output of ML Model 2 using on-/off-target and tissue drug exposure. VIMP scores for each kinase were computed in each AE-specific model and averaged across 10 iterations to produce a stable estimate of each variable's contribution to prediction. Kinases shown in bold within the VIMP plots were validated by experimental or clinical findings identified in the literature survey.

2.4.4 External validation using neratinib confirms predictive performance of final ML model

To assess the generalizability of Model 2, we performed external validation on neratinib, a clinically approved EGFR, HER2, and HER4 inhibitor not included in model training. We used clinical exposure and safety data from a Phase II randomized trial of neratinib in 116 patients with HER2+ advanced breast cancer at therapeutic doses of 240 mg/day.

Figure 2-8 compares predicted AE probabilities against observed frequencies in the neratinib trial, mirroring how safety data are typically reported in regulatory submissions where population-level AE incidence informs labeling and clinical decision-making. The left panel plots, for each AE frequently observed in the trial (observed incidence ≥ 10%), the model's predicted probability; the right panel plots, for each AE the model predicted to occur frequently, the observed incidence. Strong correlations in both panels ($R = 0.87$, $p = 3.0 \times 10^{-5}$ on the left; $R = 0.89$, $p = 1.8 \times 10^{-14}$ on the right) indicate that the predicted AE incidences closely track trial-level incidences across the full set of AE terms. This Pearson correlation is a calibration

assessment—whether the predicted population-level incidence at 229 days agrees with the observed trial incidence—distinct from the patient-level discrimination quantified by the C-index. Using the 10% threshold for classification, the model achieved 100% sensitivity and > 95% specificity for frequently occurring AEs; some low-incidence events were conservatively overpredicted, but specificity was preserved given the large number of true negatives correctly identified as unlikely.

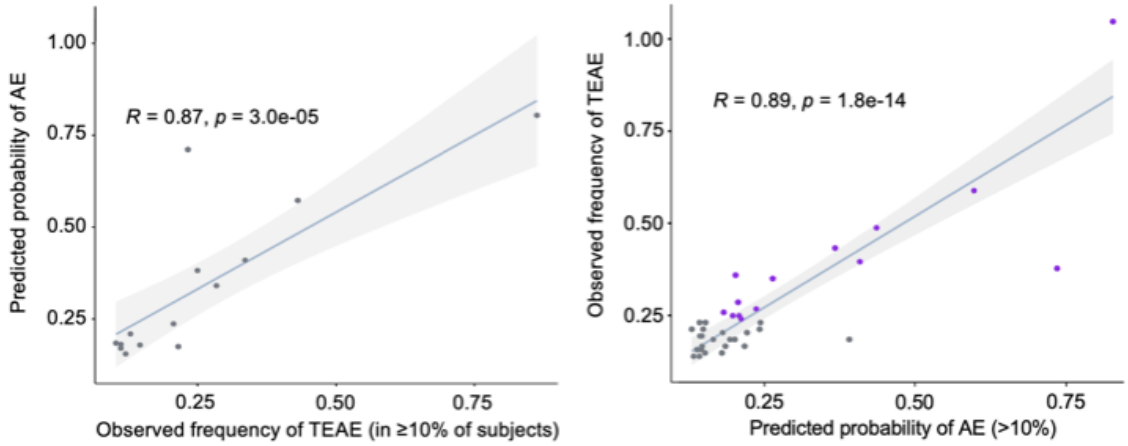


Figure 2-8. External validation on independent SMKI dataset demonstrates good model sensitivity and specificity to predict AEs of neratinib. Plotting results based on external validation performed on the neratinib clinical trial dataset. The left panel depicts a scatterplot of predicted probabilities versus all frequently reported AEs (in ≥ 10% of study subjects). The right panel depicts a scatterplot of the corresponding observed AE frequencies versus all AEs predicted to occur in ≥ 10% of study subjects. Each point on either plot represents a unique clinical AE-specific model applied to the external validation cohort. On each plot, the blue line represents the fitted linear regression and the gray band represents the 95% CI. Purple points represent AEs with both predicted probabilities and observed incidences ≥ 10%, and Pearson's r (two-tailed) and p-values are displayed directly on each plot.

Patient-level cumulative incidence was predicted for each AE across the cohort (**Figure 2-9**). Because leave-one-out cross-validation is computationally intensive, per-AE C-indices were computed for the five most frequently reported AEs (**Table 2-6**); values ranged from 0.732 to 0.808, with narrow confidence intervals indicating stable patient-level risk rankings despite the modest cohort size.

Table 2-8. Predictive accuracy for five of the most frequently reported AEs in the neratinib external validation dataset, reported as median concordance index (C-index) with 90% intervals, by Model 2.

Adverse Event (% patients affected)	**Median C-index (90% CI)**
Diarrhea (85.3%)	0.732 (0.687, 0.774)
Nausea (43.1%)	0.751 (0.718, 0.790)
Vomiting (33.6%)	0.746 (0.718, 0.798)
Fatigue (25.0%)	0.797 (0.770, 0.820)
Rash (23.3%)	0.808 (0.789, 0.837)

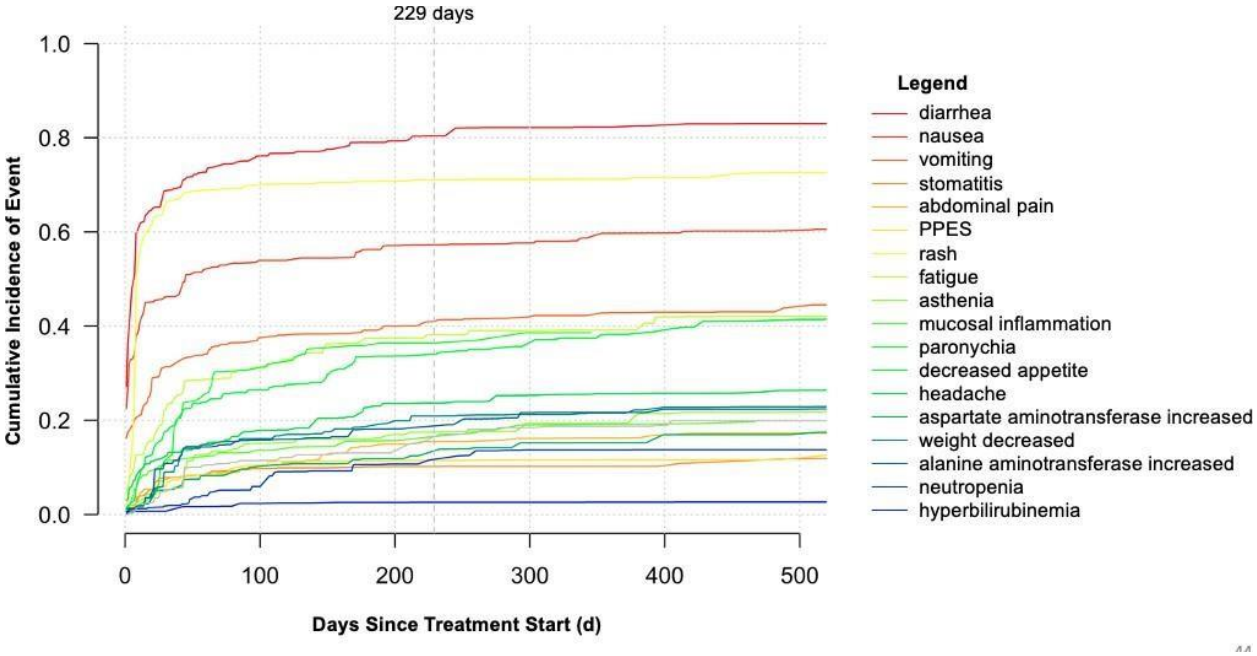


Figure 2-9. Predicted cumulative incidence curves for common adverse events in the neratinib external validation dataset. Each curve represents the RSF-predicted cumulative probability of a given AE onset over time. The vertical dashed line marks 229 days, corresponding to the mean treatment duration in the neratinib trial and the prediction horizon used for summary risk estimates.

2.5 Discussion

Our ML framework accurately predicted both the most frequently occurring AEs and the most common severe AEs in 3,433 cancer patients treated with 16 FDA-approved SMKIs, and reproduced trial-level safety patterns for a completely unseen drug (neratinib) in external validation. By integrating patient-level pharmacokinetics, kinome-wide on-/off-target binding,

and tissue-specific drug distribution, the framework establishes a mechanism-informed basis for predicting clinical AEs prospectively rather than retrospectively. Two specific applications follow directly from this capability: anticipating which patients are at elevated risk of which toxicities given their individual exposure and target engagement profile, informing dose selection and proactive monitoring; and forecasting the safety profile of an SMKI in clinical development before sufficient trial data are available, informing trial design, endpoint selection, and labeling.

The results identify a meaningful distinction between AE classes for which tissue-aware modeling adds predictive value and those for which it does not. For tissue-localized AEs (rash, nausea, fatigue, headache), incorporating K_p values to estimate local drug concentrations at the site of toxicity reveals exposure–response relationships obscured by systemic PK parameters alone. For severe hematologic toxicities, where circulating hematopoietic cells are exposed directly to plasma drug concentrations, systemic exposure remains the appropriate metric, and Model 1 achieves comparable predictive accuracy without the added complexity of tissue features.

The identified kinase-AE associations include both established and novel targets. For rash, EGFR signaling is critical for keratinocyte proliferation and differentiation, and its inhibition is known to trigger inflammatory chemokine release and immune cell infiltration in the skin[57]. The identification of ERBB4 is consistent with its structural homology to EGFR and shared downstream signaling[91,92], while ABL1 represents a potential association warranting further investigation. For diarrhea, the identified targets reflect distinct mechanisms: EGFR inhibition disrupts chloride secretion regulation in intestinal epithelium, while VEGFR inhibition causes intestinal mucosal ischemia through vascular regression[75,89]. The association with FLT likely reflects the multikinase inhibitor profiles of agents such as sunitinib and sorafenib, which potently inhibit both FLT3 and VEGFR; selective FLT3 inhibitors show notably lower rates of gastrointestinal toxicity, suggesting the diarrhea signal is VEGFR-mediated rather than FLT-specific[90]. The convergence of multiple validated targets on diarrhea is consistent with the clinical observation that diarrhea is among the most frequent toxicities reported across the SMKI class. For fatigue, while the mechanistic basis remains poorly understood, the association of FLT1 (VEGFR1) with fatigue across models is consistent with fatigue as a recognized class effect of anti-angiogenic therapies[93]. The identification of PLK4 and AURKB as brain-localized predictors of fatigue may represent novel targets for further study.

Several methodological considerations warrant discussion. On-/off-target binding profiles were limited to 442 human kinases, potentially missing additional uncharacterized off-targets. Further, while the use of AUC/K_i as a proxy for kinase inhibition is a common heuristic in translational modeling, our estimates of kinase activity should be interpreted as approximations of relative inhibition potential rather than mechanistically precise occupancy metrics. Tissue exposures were estimated from rodent-derived K_p values combined with human plasma AUCs, which necessarily assumes interspecies similarity. However, K_p values can vary due to species-specific physiological differences, including tissue composition, transporter expression, and binding proteins, potentially affecting model precision. Since K_p values cannot be measured in humans, similarity of K_p values between animals and humans was assumed, consistent with traditional physiologically-based PK modeling. Censoring in clinical trial data is potentially dependent rather than non-informative: a patient who discontinues treatment because of severe toxicity is censored at the moment a related AE is most likely, creating dependent censoring akin to a competing-risks structure. RSF assumes independent censoring; because we could not distinguish censoring causes at the patient level, some bias in hazard estimation may result, though this would likely be conservative for the AEs of primary interest. A formal competing-risks extension (e.g., Aalen–Johansen subdistribution hazards within an RSF framework) is a natural extension. Finally, incorporating tissue-specific target expression (Model 3) did not further improve prediction accuracy as was expected. This suggests that tissue drug exposure, as modeled by $AUC{\cdot}K_p$, already captures the effective distribution of on-/off-target interactions *in vivo*, reflecting underlying patterns of kinase abundance across tissues[2,14].

Future work will extend this framework beyond kinase inhibitors to broader drug classes. By compiling on-/off-target binding profiles, tissue distribution data, and clinical AE data for a much larger set of FDA-approved drugs, we aim to develop generalized models capable of predicting clinical safety profiles prior to first-in-human trials. Integration of additional features, including target expression in disease versus normal tissues and drug-like physicochemical properties, may further improve predictive accuracy and mechanistic interpretability. An expanded framework of this kind would support prediction of clinical safety profiles for new drug candidates before first-in-human trials, advancing NAM-aligned alternatives to GLP animal toxicology.

2.6 References

1. Harrison RK. Phase II and phase III failures: 2013-2015. *Nat Rev Drug Discov*. Dec 2016;15(12):817–818. doi:10.1038/nrd.2016.184
2. Sun D, Macedonia C, Chen Z, et al. Can Machine Learning Overcome the 95% Failure Rate and Reality that Only 30% of Approved Cancer Drugs Meaningfully Extend Patient Survival? *J Med Chem*. Sep 26 2024;67(18):16035–16055. doi:10.1021/acs.jmedchem.4c01684
3. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B*. Jul 2022;12(7):3049–3062. doi:10.1016/j.apsb.2022.02.002
4. Adashi EY, O'Mahony DP, Cohen IG. The FDA Modernization Act 2.0: Drug Testing in Animals is Rendered Optional. *Am J Med*. Sep 2023;136(9):853–854. doi:10.1016/j.amjmed.2023.03.033
5. Roadmap to Reducing Animal Testing in Preclinical Safety Studies (2025).
6. Bai C, Wu L, Li R, Cao Y, He S, Bo X. Machine Learning-Enabled Drug-Induced Toxicity Prediction. *Adv Sci (Weinh)*. Apr 2025;12(16):e2413405. doi:10.1002/advs.202413405
7. Toni E, Ayatollahi H, Abbaszadeh R, Fotuhi Siahpirani A. Machine Learning Techniques for Predicting Drug-Related Side Effects: A Scoping Review. *Pharmaceuticals (Basel)*. Jun 17 2024;17(6)doi:10.3390/ph17060795
8. Nguyen DA, Nguyen CH, Mamitsuka H. A survey on adverse drug reaction studies: data, tasks and machine learning methods. *Brief Bioinform*. Jan 18 2021;22(1):164–177. doi:10.1093/bib/bbz140
9. J SG, P D, P E. Enhancing drug discovery in schizophrenia: a deep learning approach for accurate drug-target interaction prediction - DrugSchizoNet. *Comput Method Biomec*. Jan 25 2025;28(2):170–187. doi:10.1080/10255842.2023.2282951
10. Galeano D, Paccanaro A. Machine learning prediction of side effects for drugs in clinical trials. *Cell Rep Methods*. Dec 19 2022;2(12):100358. doi:10.1016/j.crmeth.2022.100358
11. Yamanishi Y, Pauwels E, Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces. *J Chem Inf Model*. Dec 21 2012;52(12):3284–92. doi:10.1021/ci2005548
12. Chen L, Huang T, Zhang J, et al. Predicting drugs side effects based on chemical-chemical interactions and protein-chemical interactions. *Biomed Res Int*. 2013;2013:485034. doi:10.1155/2013/485034
13. Amano Y, Honda H, Sawada R, et al. In silico systems for predicting chemical-induced side effects using known and potential chemical protein interactions, enabling mechanism estimation. *J Toxicol Sci*. 2020;45(3):137–149. doi:10.2131/jts.45.137
14. Cao A, Zhang L, Bu Y, Sun D. Machine Learning Prediction of On/Off Target-driven Clinical Adverse Events. *Pharm Res*. Aug 2024;41(8):1649–1658. doi:10.1007/s11095-024-03742-x
15. Chen J, Han F, He M, Shi Y, Cai Y. A novel weighted pseudo-labeling framework based on matrix factorization for adverse drug reaction prediction. *BMC Bioinformatics*. Feb 17 2025;26(1):54. doi:10.1186/s12859-025-06053-z
16. Zhang J, Cui X, Yang C, et al. A deep learning-based interpretable decision tool for predicting high risk of chemotherapy-induced nausea and vomiting in cancer patients prescribed highly emetogenic chemotherapy. *Cancer Med*. Sep 2023;12(17):18306–18316. doi:10.1002/cam4.6428
17. Jian C, Chen S, Wang Z, et al. Predicting delayed methotrexate elimination in pediatric acute lymphoblastic leukemia patients: an innovative web-based machine learning tool

developed through a multicenter, retrospective analysis. *BMC Med Inform Decis Mak*. Aug 3 2023;23(1):148. doi:10.1186/s12911-023-02248-7
18. Hussain Z, De Brouwer E, Boiarsky R, et al. Joint AI-driven event prediction and longitudinal modeling in newly diagnosed and relapsed multiple myeloma. *NPJ Digit Med*. Jul 29 2024;7(1):200. doi:10.1038/s41746-024-01189-3
19. Seyednasrollah F, Koestler DC, Wang T, et al. A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer. *JCO Clin Cancer Inform*. Nov 2017;1:1–15. doi:10.1200/CCI.17.00018
20. Yang M, Ma J, Zhang C, et al. Multimodal data deep learning method for predicting symptomatic pneumonitis caused by lung cancer radiotherapy combined with immunotherapy. *Front Immunol*. 2024;15:1492399. doi:10.3389/fimmu.2024.1492399
21. Oyaga-Iriarte E, Insausti A, Sayar O, Aldaz A. Prediction of irinotecan toxicity in metastatic colorectal cancer patients based on machine learning models with pharmacokinetic parameters. *J Pharmacol Sci*. May 2019;140(1):20–25. doi:10.1016/j.jphs.2019.03.004
22. Guo L, Wang W, Xie X, Wang S, Zhang Y. Machine learning for genetic prediction of chemotherapy toxicity in cervical cancer. *Biomed Pharmacother*. May 2023;161:114518. doi:10.1016/j.biopha.2023.114518
23. Zhang Z, Zhu M, Jiang W. Risk Factors Analysis of Cutaneous Adverse Drug Reactions Caused by Targeted Therapy and Immunotherapy Drugs for Oncology and Establishment of a Prediction Model. *Clin Transl Sci*. Jan 2025;18(1):e70118. doi:10.1111/cts.70118
24. Timilsina M, Tandan M, Novacek V. Machine learning approaches for predicting the onset time of the adverse drug events in oncology. *Mach Learn Appl*. Sep 15 2022;9doi:ARTN 100367
10.1016/j.mlwa.2022.100367
25. Bean DM, Wu H, Iqbal E, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci Rep*. Nov 27 2017;7(1):16416. doi:10.1038/s41598-017-16674-x
26. Zhao H, Zheng K, Li Y, Wang J. A novel graph attention model for predicting frequencies of drug-side effects from multi-view data. *Brief Bioinform*. Nov 5 2021;22(6)doi:10.1093/bib/bbab239
27. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *The Annals of Applied Statistics*. 2008;2(3):841–860, 20.
28. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. Feb 26 2018;18(1):24. doi:10.1186/s12874-018-0482-1
29. Lee C, Yoon J, Schaar MV. Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data. *IEEE Trans Biomed Eng*. Jan 2020;67(1):122–133. doi:10.1109/TBME.2019.2909027
30. Zhao L, Feng D. Deep Neural Networks for Survival Analysis Using Pseudo Values. *IEEE J Biomed Health Inform*. Nov 2020;24(11):3308–3314. doi:10.1109/JBHI.2020.2980204
31. Huang Y, Li J, Li M, Aparasu RR. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC Med Res Methodol*. Nov 13 2023;23(1):268. doi:10.1186/s12874-023-02078-1
32. Gong X, Hu M, Liu J, et al. Decoding kinase-adverse event associations for small molecule kinase inhibitors. *Nat Commun*. Jul 27 2022;13(1):4349. doi:10.1038/s41467-022-32033-5

33. Roskoski R, Jr. Properties of FDA-approved small molecule protein kinase inhibitors: A 2026 update. *Pharmacol Res*. Feb 2026;224:108107. doi:10.1016/j.phrs.2026.108107
34. Zhang J, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer*. Jan 2009;9(1):28–39. doi:10.1038/nrc2559
35. Shyam Sunder S, Sharma UC, Pokharel S. Adverse effects of tyrosine kinase inhibitors in cancer therapy: pathophysiology, mechanisms and clinical management. *Signal Transduct Target Ther*. Jul 7 2023;8(1):262. doi:10.1038/s41392-023-01469-6
36. Davis MI, Hunt JP, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol*. Oct 30 2011;29(11):1046–51. doi:10.1038/nbt.1990
37. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. Jan 23 2015;347(6220):1260419. doi:10.1126/science.1260419
38. Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol Drug Saf*. Aug 2004;13(8):519–23. doi:10.1002/pds.1001
39. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika*. 1994;81(3):515–526. doi:10.2307/2337123
40. Segal MR. Regression Trees for Censored Data. *Biometrics*. 1988;44(1):35–47. doi:10.2307/2531894
41. LeBlanc M, Crowley J. Survival Trees by Goodness of Split. *Journal of the American Statistical Association*. 1993;88(422):457–467. doi:10.2307/2290325
42. Aalen O. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*. 1978;6(4):701–726.
43. Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer-Verlag; 1993.
44. Ishwaran H, Kogalur UB. Consistency of Random Survival Forests. *Stat Probab Lett*. Jul 1 2010;80(13-14):1056–1064. doi:10.1016/j.spl.2010.02.020
45. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. May 14 1982;247(18):2543–6.
46. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. Feb 28 1996;15(4):361–87. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4
47. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. May 10 2011;30(10):1105–17. doi:10.1002/sim.4154
48. Breiman L. Random Forests. *Machine Learning*. 2001/10/01 2001;45(1):5–32. doi:10.1023/A:1010933404324
49. Manual on setting up, using, and understanding random forests v3. 1. https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.
50. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics*. Jul 11 2008;9:307. doi:10.1186/1471-2105-9-307
51. Martin M, Bonneterre J, Geyer CE, Jr., et al. A phase two randomised trial of neratinib monotherapy versus lapatinib plus capecitabine combination therapy in patients with HER2+ advanced breast cancer. *Eur J Cancer*. Dec 2013;49(18):3763–72. doi:10.1016/j.ejca.2013.07.142

52. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw*. Mar 2011;39(5):1–13. doi:10.18637/jss.v039.i05
53. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. 2009:908–913.
54. Dy GK, Adjei AA. Understanding, recognizing, and managing toxicities of targeted anticancer therapies. *CA Cancer J Clin*. Jul–Aug 2013;63(4):249–79. doi:10.3322/caac.21184
55. Agero AL, Dusza SW, Benvenuto-Andrade C, Busam KJ, Myskowski P, Halpern AC. Dermatologic side effects associated with the epidermal growth factor receptor inhibitors. *J Am Acad Dermatol*. Oct 2006;55(4):657–70. doi:10.1016/j.jaad.2005.10.010
56. Joshi SS, Ortiz S, Witherspoon JN, et al. Effects of epidermal growth factor receptor inhibitor-induced dermatologic toxicities on quality of life. *Cancer*. Aug 15 2010;116(16):3916–23. doi:10.1002/cncr.25090
57. Lacouture ME, Anadkat MJ, Bensadoun RJ, et al. Clinical practice guidelines for the prevention and treatment of EGFR inhibitor-associated dermatologic toxicities. *Support Care Cancer*. Aug 2011;19(8):1079–95. doi:10.1007/s00520-011-1197-6
58. Potthoff K, Hofheinz R, Hassel JC, et al. Interdisciplinary management of EGFR-inhibitor-induced skin reactions: a German expert opinion. *Ann Oncol*. Mar 2011;22(3):524–535. doi:10.1093/annonc/mdq387
59. Melosky B, Hirsh V. Management of Common Toxicities in Metastatic NSCLC Related to Anti-Lung Cancer Therapies with EGFR-TKIs. *Front Oncol*. 2014;4:238. doi:10.3389/fonc.2014.00238
60. Takeda M, Okamoto I, Nakagawa K. Pooled safety analysis of EGFR-TKI treatment for EGFR mutation-positive non-small cell lung cancer. *Lung Cancer*. Apr 2015;88(1):74–9. doi:10.1016/j.lungcan.2015.01.026
61. Gharwan H, Groninger H. Kinase inhibitors and monoclonal antibodies in oncology: clinical implications. *Nat Rev Clin Oncol*. Apr 2016;13(4):209–27. doi:10.1038/nrclinonc.2015.213
62. Chu CY, Chen KY, Wen-Cheng Chang J, Wei YF, Lee CH, Wang WM. Taiwanese Dermatological Association consensus for the prevention and management of epidermal growth factor receptor tyrosine kinase inhibitor-related skin toxicities. *J Formos Med Assoc*. Jun 2017;116(6):413–423. doi:10.1016/j.jfma.2017.03.001
63. Cabanillas ME, Ryder M, Jimenez C. Targeted Therapy for Advanced Thyroid Cancer: Kinase Inhibitors and Beyond. *Endocr Rev*. Dec 1 2019;40(6):1573–1604. doi:10.1210/er.2019-00007
64. Zhou JY, Liu SY, Wu YL. Safety of EGFR-TKIs for EGFR mutation-positive non-small cell lung cancer. *Expert Opin Drug Saf*. May 2020;19(5):589–599. doi:10.1080/14740338.2020.1753697
65. Coleman EL, Olamiju B, Leventhal JS. Potentially life-threatening severe cutaneous adverse reactions associated with tyrosine kinase inhibitors (Review). *Oncol Rep*. Mar 2021;45(3):891–898. doi:10.3892/or.2020.7911
66. Nixon S, Duquette D, Doucette S, Larouche JF. How We Manage Patients with Indolent B-Cell Malignancies on Bruton's Tyrosine Kinase Inhibitors: Practical Considerations for Nurses and Pharmacists. *Curr Oncol*. Apr 18 2023;30(4):4222–4245. doi:10.3390/curroncol30040322

67. Singer S, Tan SY, Dewan AK, et al. Cutaneous eruptions from ibrutinib resembling epidermal growth factor receptor inhibitor-induced dermatologic adverse events. *J Am Acad Dermatol*. Jun 2023;88(6):1271–1281. doi:10.1016/j.jaad.2019.12.031
68. Galitzia A, Maccaferri M, Mauro FR, Murru R, Marasca R. Chronic Lymphocytic Leukemia: Management of Adverse Events in the Era of Targeted Agents. *Cancers (Basel)*. May 24 2024;16(11)doi:10.3390/cancers16111996
69. Lu CW, Chen CB, Chiu TM, et al. Consensus of the Taiwanese dermatological association and Taiwan Lung Cancer Society on the prevention and management of tyrosine kinase inhibitor-related skin toxicities in patients with non-small cell lung cancer: An updated version incorporating Taiwanese treatment experience. *J Formos Med Assoc*. May 2025;124(5):402–412. doi:10.1016/j.jfma.2024.07.029
70. Macdonald JB, Macdonald B, Golitz LE, LoRusso P, Sekulic A. Cutaneous adverse effects of targeted therapies: Part I: Inhibitors of the cellular membrane. *J Am Acad Dermatol*. Feb 2015;72(2):203–18; quiz 219–20. doi:10.1016/j.jaad.2014.07.032
71. Reguiai Z, Bachet JB, Bachmeyer C, et al. Management of cutaneous adverse events induced by anti-EGFR (epidermal growth factor receptor): a French interdisciplinary therapeutic algorithm. *Support Care Cancer*. Jul 2012;20(7):1395–404. doi:10.1007/s00520-012-1451-6
72. Arrieta O, Vega-Gonzalez MT, Lopez-Macias D, et al. Randomized, open-label trial evaluating the preventive effect of tetracycline on afatinib induced-skin toxicities in non-small cell lung cancer patients. *Lung Cancer*. Jun 2015;88(3):282–8. doi:10.1016/j.lungcan.2015.03.019
73. Ran F, Liu Y, Wang C, et al. Review of the development of BTK inhibitors in overcoming the clinical limitations of ibrutinib. *Eur J Med Chem*. Feb 5 2022;229:114009. doi:10.1016/j.ejmech.2021.114009
74. Enokida T, Tahara M. Management of VEGFR-Targeted TKI for Thyroid Cancer. *Cancers (Basel)*. Nov 4 2021;13(21)doi:10.3390/cancers13215536
75. Liu J, Yan S, Du J, et al. Mechanism and treatment of diarrhea associated with tyrosine kinase inhibitors. *Heliyon*. Mar 30 2024;10(6):e27531. doi:10.1016/j.heliyon.2024.e27531
76. Yuan A, Kurtz SL, Barysauskas CM, Pilotte AP, Wagner AJ, Treister NS. Oral adverse events in cancer patients treated with VEGFR-directed multitargeted tyrosine kinase inhibitors. *Oral Oncol*. Nov 2015;51(11):1026–1033. doi:10.1016/j.oraloncology.2015.09.003
77. Cabanillas ME, Takahashi S. Managing the adverse events associated with lenvatinib therapy in radioiodine-refractory differentiated thyroid cancer. *Semin Oncol*. Feb 2019;46(1):57–64. doi:10.1053/j.seminoncol.2018.11.004
78. Si X, Zhang L, Wang H, et al. Management of anlotinib-related adverse events in patients with advanced non-small cell lung cancer: Experiences in ALTER-0303. *Thorac Cancer*. Mar 2019;10(3):551–556. doi:10.1111/1759-7714.12977
79. Nervo A, Ragni A, Gallo M, et al. Symptomatic Biliary Disorders During Lenvatinib Treatment for Thyroid Cancer: An Underestimated Problem. *Thyroid*. Feb 2020;30(2):229–236. doi:10.1089/thy.2019.0355
80. Schwartz G, Darling JO, Mindo M, Damicis L. Management of Adverse Events Associated with Cabozantinib Treatment in Patients with Advanced Hepatocellular Carcinoma. *Target Oncol*. Aug 2020;15(4):549–565. doi:10.1007/s11523-020-00736-8
81. Liu J, Nicum S, Reichardt P, et al. Assessment and management of diarrhea following VEGF receptor TKI treatment in patients with ovarian cancer. *Gynecol Oncol*. Jul 2018;150(1):173–179. doi:10.1016/j.ygyno.2018.03.058

82. Weitzman SP, Cabanillas ME. The treatment landscape in thyroid cancer: a focus on cabozantinib. *Cancer Manag Res*. 2015;7:265–78. doi:10.2147/CMAR.S68373
83. Vergoulidou M. More than a Decade of Tyrosine Kinase Inhibitors in the Treatment of Solid Tumors: What We Have Learned and What the Future Holds. *Biomark Insights*. 2015;10(Suppl 3):33–40. doi:10.4137/BMI.S22436
84. Saeedi BJ, Carr HE, Higgins PDR, Steiner CA. AXL: A novel therapeutic target in IBD. *Adv Pharmacol*. 2024;101:141–157. doi:10.1016/bs.apha.2024.10.009
85. Kuang AG, Mohajir W, Panneerselvam K, et al. Diarrhea and colitis related to immune checkpoint inhibitor and BRAF/MEK inhibitor therapy. *Ann Gastroenterol*. Jan–Feb 2023;36(1):45–53. doi:10.20524/aog.2023.0762
86. Sanaei MJ, Razi S, Pourbagheri-Sigaroodi A, Bashash D. The PI3K/Akt/mTOR pathway in lung cancer; oncogenic alterations, therapeutic opportunities, challenges, and a glance at the application of nanoparticles. *Transl Oncol*. Apr 2022;18:101364. doi:10.1016/j.tranon.2022.101364
87. Shan W, Wu G, Huang Y, et al. Adverse events in lymphoma patients treated with phosphoinositide 3 kinase Inhibitor in clinical trials: a meta-analysis. *Ann Hematol*. Aug 2022;101(8):1741–1753. doi:10.1007/s00277-022-04876-x
88. Breal C, Beuvon F, de Witasse-Thezy T, et al. Management of Gastro-Intestinal Toxicity of the Pi3 Kinase Inhibitor: Optimizing Future Dosing Strategies. *Cancers (Basel)*. Apr 13 2023;15(8)doi:10.3390/cancers15082279
89. Schmidinger M. Understanding and managing toxicities of vascular endothelial growth factor (VEGF) inhibitors. *EJC Suppl*. Sep 2013;11(2):172–91. doi:10.1016/j.ejcsup.2013.07.016
90. Larrosa-Garcia M, Baer MR. FLT3 Inhibitors in Acute Myeloid Leukemia: Current Status and Future Directions. *Mol Cancer Ther*. Jun 2017;16(6):991–1001. doi:10.1158/1535-7163.MCT-16-0876
91. Yarden Y, Sliwkowski MX. Untangling the ErbB signalling network. *Nat Rev Mol Cell Biol*. Feb 2001;2(2):127–37. doi:10.1038/35052073
92. Topalan E, Buyukgungor A, Cigdem M, et al. A Structural Insight Into Two Important ErbB Receptors (EGFR and HER2) and Their Relevance to Non-Small Cell Lung Cancer. *Arch Pharm (Weinheim)*. Apr 2025;358(4):e2400992. doi:10.1002/ardp.202400992
93. Takahashi S. Fatigue and its management in cancer patients undergoing VEGFR-TKI therapy. *Expert Opin Drug Saf*. Mar 2022;21(3):397–406. doi:10.1080/14740338.2021.1969360

Chapter 3: Multi-Task Neural Network Predicts Clinical Adverse Event Incidence Across a Multi-Mechanism Cohort Using a Unified STAR Feature Space

3.1 Abstract

Clinical adverse event (AE) prediction methods that operate on chemical structure or known AE associations alone do not sufficiently represent the pharmacological mechanisms by which a molecular liability becomes a clinical toxicity, and most clinical safety resources express AEs as reporting signals rather than incidence rates with known sample sizes. We assembled a cohort of 86 FDA-approved small molecules spanning 20 oncology and 14 non-oncology mechanism classes, each represented by a unified 3,170-dimensional Structure-Tissue/cell selectivity-Activity-Relationship (STAR) feature vector integrating structural fingerprints and descriptors, kinome-wide on-/off-target binding affinities harmonized across four public databases, normal and tumor target-tissue expression, tissue/cell selectivity, and approved clinical dose. Per-AE incidence rates were extracted from the adverse-reactions sections of FDA prescribing labels across 79 endpoints normalized to SNOMED CT, and a shared-trunk multi-task neural network was trained under a binomial-*N*-weighted cross-entropy loss. Under a base-rate-robust per-drug rank metric, the model recovered a reproducible drug-specific predictive signal beyond per-AE base rates (corrected per-drug Spearman $\rho = +0.079$, 95% stability interval [+0.033, +0.127]), cleanly separated from constant, cohort-mean, ridge, Y-permutation, and random-feature baselines. Per-drug predicted AE rankings traced to established target-AE pharmacology for sunitinib, imatinib, and erlotinib. Under leave-one-mechanism-class-out cross-validation, generalization was strongest for classes pharmacologically analogous to the training cohort. A rank-within-column corrected per-drug Spearman statistic, robust to asymmetric AE label presence in FDA labeling data, is introduced as a base-rate-robust discrimination metric. By placing structurally and mechanistically diverse small molecules on a common pharmacological footing from public data alone, the framework offers a route to triaging likely adverse event profiles across mechanism classes before first-in-human testing, complementing structure- and target-panel-based screening in early safety assessment.

3.2 Introduction

We previously established that integrating patient-level pharmacokinetics (PK), kinome-wide on- and off-target binding, and tissue-specific drug distribution improves the prediction of clinical adverse events (AEs) in cancer patients treated with small molecule kinase inhibitors, with predictive gains particularly for tissue-localized toxicities. The framework as constructed depends on three resources unavailable at scale: individual patient-level exposure data from registrational trials, dense binding panels across a structurally constrained target family, and a single mechanistic class within which off-target liability can be cleanly attributed to inhibition of paralogous kinases. Patient-level PK is absent from public sources for most FDA-approved drugs, and mechanism-aware AE prediction has so far been demonstrated almost exclusively within the kinase inhibitor class, where the available binding infrastructure is densest[1,2]. Extending the framework beyond kinase inhibitors therefore requires a shift in both the modeling architecture and the input data: from patient-level time-to-event prediction within one drug class to compound-level incidence prediction across many, using inputs that can be assembled from public databases without recourse to individual patient trial data.

Compound-level AE prediction has been advanced along several methodological lines. Matrix decomposition over SIDER frequency data recovers AE rates for held-out drugs from latent biological activity features but does not encode on- or off-target engagement explicitly[3], and subsequent extensions through similarity-based deep learning and graph attention networks have improved frequency prediction without incorporating pharmacological features beyond drug structure and known AE associations[4,5]. Multi-task deep neural networks trained on Morgan fingerprints and SMILES achieve performance comparable to state-of-the-art benchmarks and show that transfer learning across *in vitro*, *in vivo*, and clinical labels can compensate for the scarcity of clinical training data, but such frameworks treat AEs are binary classification targets and typically rely on chemical structure as the sole input modality[6]. A complementary line of work links AE prediction directly to drug-target engagement; large-scale integration of off-target binding panels with marketed-drug AE associations has demonstrated that pharmacological target profiles capture clinical safety signals not visible from chemical structure alone[7], and pharmaceutical-industry safety panels built on this principle have been adopted as a standard component of preclinical attrition reduction[8]. Graph convolutional and knowledge graph

approaches in corporate richer relational information among drugs, targets, and AE terms[9-11], and a multi-task framework has recently been developed to predict the seriousness class of drug-adverse drug reaction interactions[12]; these models predominantly draw labels from FAERS, whose absence of sample size information conflates incidence with reporting intensity. Prior work from our group combined ConPLex-predicted IC50 values across 477 protein targets with tissue-specific protein expression to predict clinical AEs for 759 FDA-approved drugs, providing the first integration of on-/off-target binding with tissue expression at the compound level but by predicting each AE endpoint with an independent classifier[13]. Across the field, three properties are missing in combination: a single feature representation that unifies the variables of the Structure-Tissue/cell selectivity-Activity (STAR) framework[14,15]; AE labels expressed as incidence rates with known trial-arm sample sizes, recoverable from the adverse-reactions sections of FDA prescribing labels; and a multi-task architecture in which supervision from one drug's AE profile informs predictions for pharmacologically related compounds across mechanism classes.

We hypothesize that a multi-task neural network trained on a unified STAR feature space across a cohort of FDA-approved drugs spanning multiple mechanism classes will (i) recover compound-level discrimination of per-AE incidence beyond what per-AE base rates alone explain, under a base-rate-robust per-drug rank metric; (ii) produce per-drug predicted AE rankings that trace to established target-AE pharmacology in well-represented mechanism classes; and (iii) generalize to held-out mechanism classes to the extent that those classes share pharmacological context with the rest of the cohort. To test these hypotheses, we assembled a cohort of FDA-approved small molecules spanning 20 oncology and 14 non-oncology mechanism classes, each represented by a 3,170-dimensional STAR feature vector comprising structural fingerprints and descriptors (S), kinome-wide on-/off-target binding affinities harmonized across four public databases (PS), normal and tumor target-tissue expression from the Human Protein Atlas and UCSC Xena (EX), and approved clinical daily dose from FDA prescribing labels (Cp). We extracted per-AE incidence rates from the adverse-reactions sections of FDA prescribing labels normalized to SNOMED CT preferred terms and trained a shared-trunk multi-task neural network to predict incidence simultaneously across 79 AE endpoints under a binomial-N-weighted cross-entropy loss. We evaluate the framework under compound-level leave-one-out and leave-one-mechanism-class-out cross-validation, against constant

predictor, per-AE cohort mean, and single-task ridge regression baselines, and under Y-permutation and random feature negative controls. The framework is the first, to our knowledge, to place small molecules from kinase, GPCR, nuclear receptor, proteasome, and other non-kinase mechanism classes on a common pharmacological footing for cross-class AE incidence prediction and provides both an evaluation protocol and a methodological correction that subsequent work can adopt.

3.3 Methods

3.3.1 Cohort definition and inclusion criteria

Starting from an initial pool of 1,147 FDA-approved small molecules with structured AE data extracted from prescribing labels, we applied three coverage filters to define the modeling cohort: (i) at least one measured on- or off-target binding affinity in a unified kinase-binding daaset; (ii) a parseable SMILES string for structural fingerprint generation; and (iii) at least one clinical AE recorded in the FDA label's adverse-reactions section after terminology normalization. The resulting cohort contains 86 small molecule drug compounds spanning oncology and non-oncology mechanism classes.

3.3.2 STAR feature assembly

Each drug compound in the cohort is represented by a five-block STAR feature vector, with each block capturing a distinct dimension of the STAR framework. The total per-compound dimensionality is 3,170.

Structure (S). The S block contains 2,056 dimensions per compound. 2,048 of those dimensions are a Morgan circular fingerprint at radius 2, generated from RDKit[16] from each compound's SMILES and retained as binary indicators without further normalization. The remaining eight dimensions are physicochemical descriptors – cLogP, molecular weight, topological polar surface area, number of rotatable bonds, hydrogen bond donor and acceptor counts, aromatic ring count, and the fraction of sp^3 carbons – each z-scored using per-fold training-set statistics during model training.

Potency/specificity (PS). The PS block contains 938 dimensions per compound, capturing on-/off-target binding affinities across 469 genes. The total set of genes captured represent the union of compound-binding genes across four public databases: Davis et al.[17], KIBA[18], BindingDB[19], and Drug Target Commons (DTC)[20], restricted to those with non-missing measurements for ≥ 3

cohort drug compounds. Affinities are expressed at pK_d values (negative $\log_{10}$ of the dissociation constant in M). Where multiple sources report a (compound, gene) measurement, the priority order of Davis et al., BindingDB, DTC, then KIBA is adopted to reflect per-source data quality and homogeneity, with KIBA scores from the original publication empirically calibrated against Davis et al. pKd values prior to the merge to bring them onto a common log-kinetic-constant axis. Each (compound, gene) cell is paired with a 0/1 measured-presence mask.

Expression (EX). The EX block contains 87 dimensions per compound, summarizing the target-gene expression in normal and tumor tissues across a representative tissue panel relevant to the PS gene set. Normal tissue expression is expressed as log2(nTPM, normalized transcripts per million) and drawn from the Human Protein Atlas (HPA; v23.0)[21]; tumor tissue expression for the same subset of genes is drawn from UCSC Xena TCGA pan-cancer dataset[22]. The block summarizes target-gene expression patterning across tissues rather than encoding per-gene expression directly, providing the model with information about which compartments the PS feature set targets are most likely to engage.

Tissue/cell selectivity (TS). The TS block contains 87 dimensions per compound. Each dimension is a binding strength-weighted target expression score across normal and tumor tissues, computed on a linear scale expression following the formulation of Cao et al.[13].

Clinical dose (Cp). The Cp block contains 2 dimensions per compound: the approved daily dose extracted from the dosage and administration section of each compound's FDA prescribing label, expressed a $\log_{10}$(milligrams per day), and a 0/1 measured mask column. Body surface area regimens (mg/m^2) and weight-based regimens (mg/kg) are standardized to per-dosing-day equivalents using reference values of 1.7m^2 body surface area and 70 kg body weight.

3.3.3 Adverse event outcomes

Source data and extraction. FDA labels were obtained from the openFDA drug label corpus (https://open.fda.gov/drug/label/)[23], supplemented by targeted single-record API queries for cohort drugs whose most current label was absent from the bulk download at corpus assembly. From each label's adverse-reactions section we extracted (drug, AE term, incidence rate, study-arm sample size) tuples by parsing both HTML tabular content and free-text content using a custom HTML row-parser for the former and regular-expression matching of AE-term-followed-by-incidence-rate patterns for the latter. Reported incidence rates were parsed to a numeric value with a censored-flag column for label conventions ("<X%", "≤ X%", "less than X",

"approximately X", "about X", "up to X"); the numeric value was retained as the bound itself. Raw extraction produced 106,171 (drug, AE-string, incidence) rows across the cohort and the broader openFDA corpus.

AE terminology mapping and matrix construction. Extracted AE strings were case-folded, whitespace-canonicalized, and stripped of trailing severity-grade qualifiers (e.g. "nausea grade 3-4" to "nausea") before mapping to SNOMED CT concepts via a three-pass pipeline: a pre-curated CTCAE/MedDRA-to-SNOMED crosswalk (the primary pass; 5,602 AE-string entries anchored via UMLS CUIs), followed by exact case-folded matches against the SNOMED CT Fully Specified Name index (AE-scope only) and against SNOMED CT synonym strings (SNOMED CT US Edition release 20260301; UMLS Metathesaurus 2025AB)[24,25]. Where a single AE string matched multiple SNOMED concepts, the mapping was retained only if all candidates resolved to the same concept identifier. Rows that remained unmapped, matched out-of-AE-scope concepts, contained grade-level qualifiers without a base AE term, or had missing or non-finite incidence values were excluded. Where multiple rows mapped to the same (drug, SNOMED-CID) cell, incidence was aggregated by sample-size-weighted mean using the trial-arm *N* reported alongside each row. Endpoints with fewer than 5 labeled drugs were retained in the source data for audit but excluded from the modeling vocabulary.

3.3.4 Multi-task neural network architecture

Drugs sharing pharmacological mechanism typically share AE patterns: heme-toxic agents present with overlapping hematologic AEs across cytotoxic and targeted classes; VEGFR-pathway inhibitors share hypertension and proteinuria. A multi-task architecture is the natural choice for this structure, learning a single shared representation of drug toxicity from each compound's STAR feature vector and producing predictions for all 79 AE endpoints simultaneously, so that supervision from one pharmacologically related compound's AE profile informs predictions for related compounds rather than each AE prediction being trained in isolation. The architecture itself is intentionally simple, a feedforward multilayer perceptron without STAR-specific inductive biases, chosen to isolate the contribution of feature engineering from architectural choices.

Shared trunk. The shared trunk is composed of two linear layers (3,170 $\rightarrow$ 128 $\rightarrow$ 64) with ReLU activation and dropout (p = 0.3) between layers, holding 414,144 trainable parameters.

Per-endpoint heads. The per-endpoint heads are composed of a single linear layer ($64 \rightarrow 1$) followed by a sigmoid activation, with one independent head per AE endpoint (79 total). Per-endpoint heads contribute 5,135 trainable parameters in aggregate ($\approx$ 65 parameters per head). The approximately 80:1 ratio of shared trunk to head-specific weights concentrates the model's representational capacity in the shared layer, allowing each AE prediction to draw on supervision from the other 78 AE endpoints.

Loss function. Predictions are trained against a masked, binomial-*N*-weighted binary cross-entropy:

$$\mathcal{L} = -\frac{\sum_{ij} m_{ij}\, N_{ij}[y_{ij}\, log\hat{p}_{ij} + (1 - y_{ij}) \log(1 - \hat{p}_{ij})]}{\sum_{ij} m_{ij}\, N_{ij}} \tag{3.1}$$

where y_{ij} is the observed incidence (fraction) for drug i and AE j; $\hat{p}_{ij}$ is the predicted incidence; $m_{ij} \in \{0, 1\}$ is the label-presence mask; and N_{ij} is the underlying trial-arm sample size. The binomial weighting causes endpoints backed by larger clinical datasets to contribute proportionally more to training, and unlabeled (drug, AE) cells contribute zero gradient.

Optimizer and training schedule. An AdamW optimizer with learning rate 1×10^{-3} and weight decay 1×10^{-3} was implemented. Training proceeds for up to 2,000 epochs with early stopping at patience 100. These hyperparameters were selected via a $3 \times 3 \times 3$ grid sweep over learning rate $\in \{3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}\}$, dropout $\in \{0.0, 0.3, 0.5\}$and hidden dimensions $\in \{(64, 32), (128, 64), (256, 128)\}$; sensitivity to this choice is reported in **3.3.7**.

3.3.5 Training and validation protocol

Compound-level leave-one-out cross-validation (LOOCV). Each of the drugs in the final training cohort is held out in turn; the model is trained from scratch on the remaining drug compounds and used to predict the held-out drug's full 79-endpoint AE profile. The procedure is repeated such that every drug receives an honest out-of-sample prediction for every AE for which it is labeled. Feature normalization is computed using only the training fold's statistics within each LOOCV iteration.

Multi-seed bootstrap. To quantify variability arising from random weight initialization, the full LOOCV procedure is repeated 500 times with distinct initialization seeds. The 500-seed bootstrap allows reporting of bootstrap confidence intervals for all aggregate metrics as well as per-cell prediction intervals on the predicted incidence matrix.

Leave-one-mechanism-class-out (LOOMCO) cross-validation. Compound-level LOOCV permits chemical similarity leakage; when one EGFR inhibitor is held out, the seven other EGFR inhibitors remain in the training set, allowing the model to memorize EGFR-class AE patterns. To address this, we additionally evaluate the model under LOOMCO: all drugs belonging to a given mechanism class (≥ 3 drugs; 12 eligible classes) are held out simultaneously, the model is trained on the remaining classes, and predictions for the held-out class are recorded. This provides a stricter test of generalization to unseen pharmacological contexts.

3.3.6 Statistical metrics and baselines

Cohort-level metrics. Three cohort-level metrics are reported. Macro Mean Absolute Error (MAE) is the mean of per-AE mean absolute errors and is reported primarily as a calibration reference; for sparse heavy-tailed outcomes such as AE incidence, MAE is dominated by per-AE base rates. Mean per-head Pearson *r* is the Pearson correlation between predicted and observed incidence vectors for each AE, averaged across the 79 endpoints; this metric is base-rate-invariant and serves as the measure of cross-drug ranking performance per AE. Mean per-drug Spearman ρ is the rank correlation between each drug's predicted and observed AE profile vectors, averaged across drugs with ≥ 3 labeled AEs.

The primary per-drug metric is a corrected (rank-in-column) per-drug Spearman $\bar{\rho}$: for each AE column the observed and predicted incidence is replaced with their percentile rank across the cohort of drugs labeled for that AE (using average rank handling for ties), then compute per-drug Spearman ρ across the percentile ranks of that drug's labeled AEs. The rank-in-column normalization removes the per-AE base rate axis without imposing a parametric model of the base rate, and the resulting metric is undefined for a constant predictor, which is the desired null behavior. We report corrected $\bar{\rho}$ as the primary discrimination metric, alongside raw and legacy-residual $\bar{\rho}$ for completeness.

Baseline comparators. The constant-predictor null assigns a single scalar to every (drug, AE) cell and establishes the corrected metric's expected null. The per-AE cohort mean baseline predicts each AE's leave-one-out cohort mean incidence for every drug and establishes the calibration reference for macro MAE. The single-task ridge regression baseline fits independent per-endpoint RidgeCV models (GCV-tuned over $\alpha \in \{0.1, 1, 10, 100, 1000\}$) under the same

LOOCV protocol as the multi-task NN, isolating the contribution of the multi-task shared trunk from the STAR feature representation itself.

Negative controls. To verify that the model recovers genuine pharmacological signal rather than label or feature noise, we additionally run two negative controls. Under Y-permutation, AE labels are shuffled within each AE column, preserving per-AE marginal incidence statistics but removing the drug-AE coupling. Under random-feature, the input feature matrix is replaced with random noise matching each block's marginal statistics (Bernoulli bit-fraction for Morgan fingerprints; Gaussian with matched mean and variance for continuous blocks). Both controls are evaluated under the corrected per-drug $\bar{\rho}$ and should return $\bar{\rho} \approx 0$.

Calibration analysis. We quantified the absolute calibration of the multi-task neural network's predicted incidence against observed incidence using four measurements applied to the 500-seed bootstrap mean predictions on which the corrected per-drug $\bar{\rho}$ is computed. Expected calibration error (ECE) discretizes predictions into 10 equal-width bins on [0, 1] and reports the count-weighted mean absolute difference between each bin's mean predicted incidence and mean observed incidence; ECE is reported both cohort-pooled and per-AE head. Brier score is the mean squared error between predicted and observed incidence, reported pooled across cells (macro Brier) and per-AE head. Compression ratio is the ratio of the standard deviation of predicted incidence to the standard deviation of the observed incidence across all labeled data; a value of 1.0 indicates predictions span the observed range, and values below 1.0 quantify shrinkage of predictions towards per-AE means. AE class-stratified ECE values are computed by grouping the 79 SNOMED-defined AE endpoints into pharmacological categories (hematologic, gastrointestinal, dermatologic, cardiovascular, neurologic, hepatic/renal, electrolyte), then pooling ECE within each category. We additionally compute per-head isotonic post-calibration, which uses isotonic regression to rescale each AE head's predicted incidence values to better match observed values while preserving their rank order; we report the median post-calibration Brier score as an in-sample upper bound on the calibration improvement achievable by *post-hoc* rescaling alone, not as a cross-validated estimate.

3.3.7 Software and implementation

All analyses were performed in Python 3.11. The multi-task neural network was implemented in PyTorch 2.0[26], with AdamW optimization[27]. Molecular fingerprints and physicochemical descriptors were computed using RDKit 2024.03[28]. Ridge regression baselines

and isotonic post-calibration were implemented using the RidgeCV and IsotonicRegression classes from scikit-learn 1.4[29]. Data manipulation, statistical tests, and figure generation used NumPy 1.26, pandas 2.2, SciPy 1.13, and matplotlib 3.8.

3.4 Results

3.4.1 Cohort and STAR feature composition

The final modeling cohort comprises 86 FDA-approved small molecules spanning 20 oncology and 14 non-oncology mechanism classes (**Figure 3-1**). Oncology drugs (n = 63) are concentrated in the kinase inhibitor families EGFR/HER (n = 8), VEGFR/multi-TKI (n = 6), PI3K/AKT/mTOR (n = 6), ALK/ROS1 (n = 5), BCR-ABL (n = 4), and cytotoxic agents (n = 6), in addition to smaller classes spanning CDK4/6, BRAF, BTK, JAK, IDH, proteasome inhibitors, and immunomodulary drug compounds (IMiDs). Non-oncology drugs are predominantly antipsychotics (n = 5), antimuscarinics, SGLT2 inhibitors, HIV antiretrovirals, and other mechanism classes added to test generalization beyond kinase pharmacology.

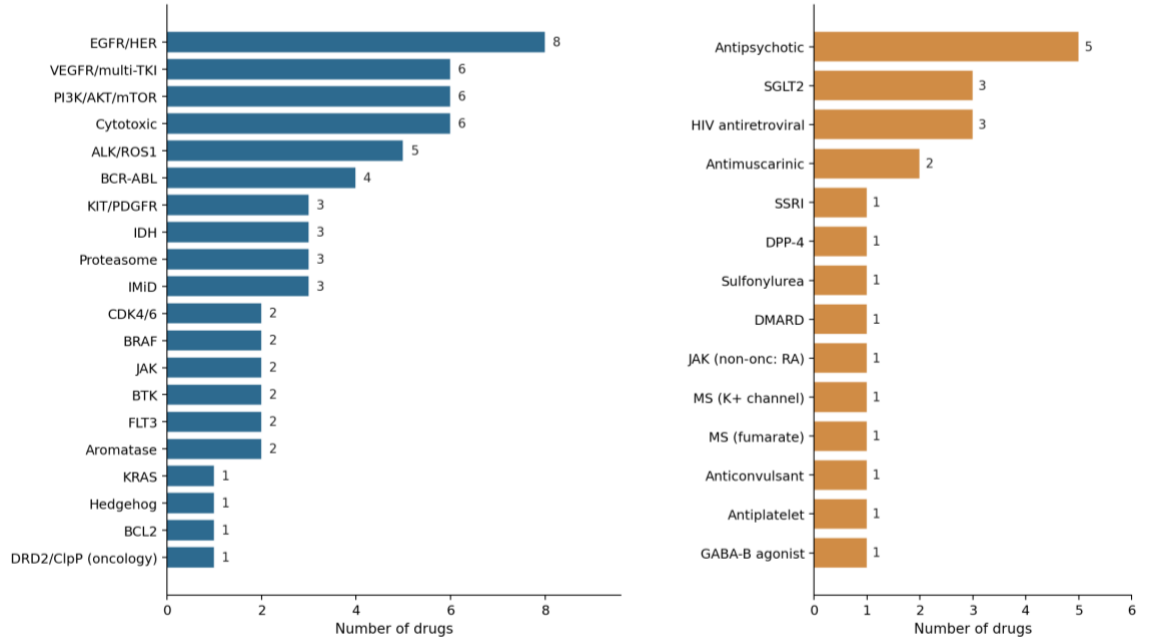


Figure 3-1. Mechanism-class composition of the modeling cohort, partitioned into 20 oncology and 14 non-oncology classes. Counts on each bar indicate the number of drugs assigned to that class.

The five STAR feature blocks total 3,170 dimensions per compound. Per-drug median Morgan-bit count is 59 (range: 11-106), reflecting heterogeneity in molecular complexity across the cohort. Per-drug median PS target count is 10 (range: 1-402 over the 469 gene set), indicating

wide variation in available *in vitro* kinome profiling. EX and Cp block coverage are complete by construction.

The assembled AE prediction matrix contains 1,491 labeled cells (21.9% density; **Figure 3-2**). Per drug, the median number of labeled AEs is 15 (range: 0-56); per AE, the median number of labeled drugs is 17 (range: 5-64). The labeled-cell distribution skews toward high prevalence AEs such as nausea, diarrhea, anemia, and fatigue, which are reported across most drugs in the cohort. Rarer AEs are labeled for fewer drugs but contribute mechanism-specific signal to the multi-task supervision.

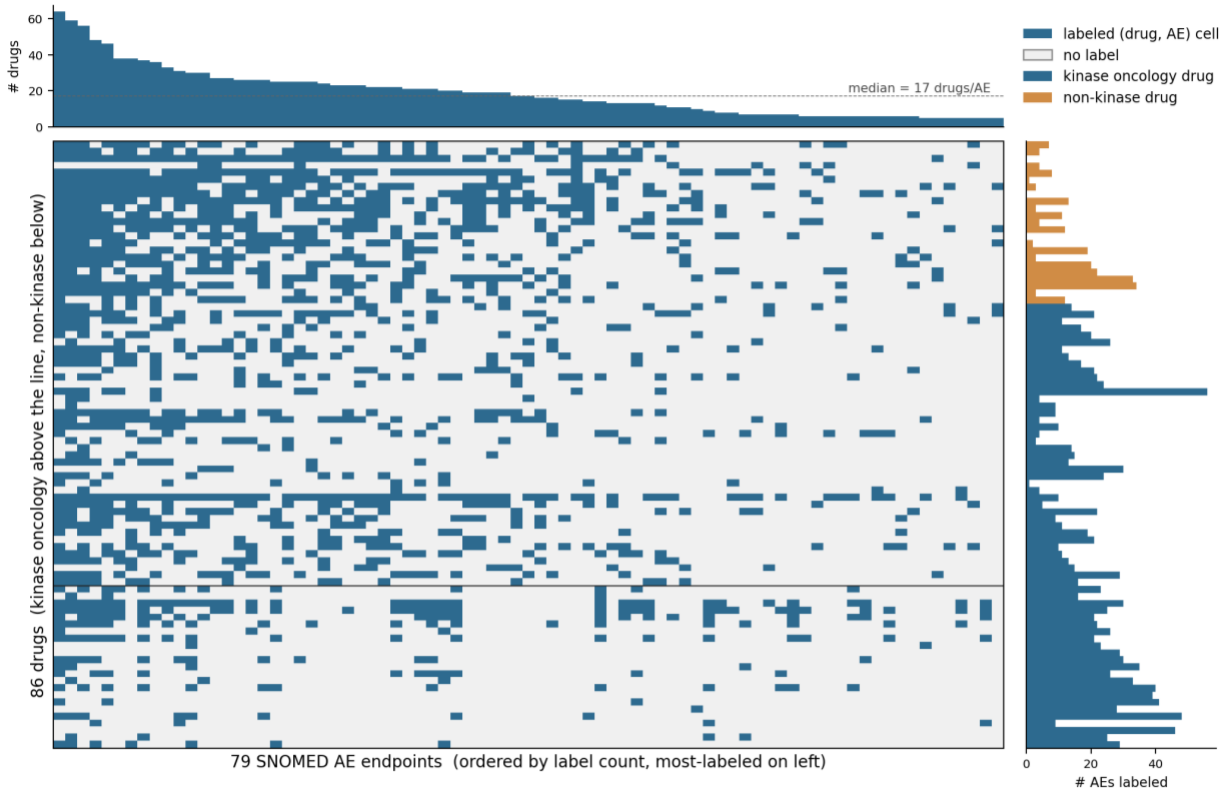


Figure 3-2. AE-label coverage of the assembled prediction matrix (center: full coverage heatmap; top/right: marginal histograms of per-drug and per-AE label counts).

3.4.2 The multi-task neural network recovers drug-specific predictive signal beyond AE base rates

The multi-task neural network recovers a modest but reproducible drug-specific predictive signal beyond what AE base rates alone explain. Across the 500-seed bootstrap on compound-level LOOCV, the cohort-wide mean corrected (rank-in-column) per-drug Spearman $\bar{\rho}$ is +0.079 (across-seed 95% stability interval [+0.033, +0.127]) (**Table 3-1**, **Figure 3-3**). A constant predictor returns a null result on this metric by construction, and the per-AE cohort

mean predictor returns $\bar{\rho} \approx 0$; Y-permutation and random-feature negative controls (**3.3.7**) return +0.028 ± 0.037 and -0.081 ± 0.046 respectively. Single-task ridge regression on the same input features and LOOCV protocol returns $\bar{\rho} = -0.080$, placing the multi-task neural network's bootstrap interval cleanly above ridge and both negative controls. On a per-(seed × AE) basis, the multi-task neural network outperforms ridge on 56.6% of head Pearson values.

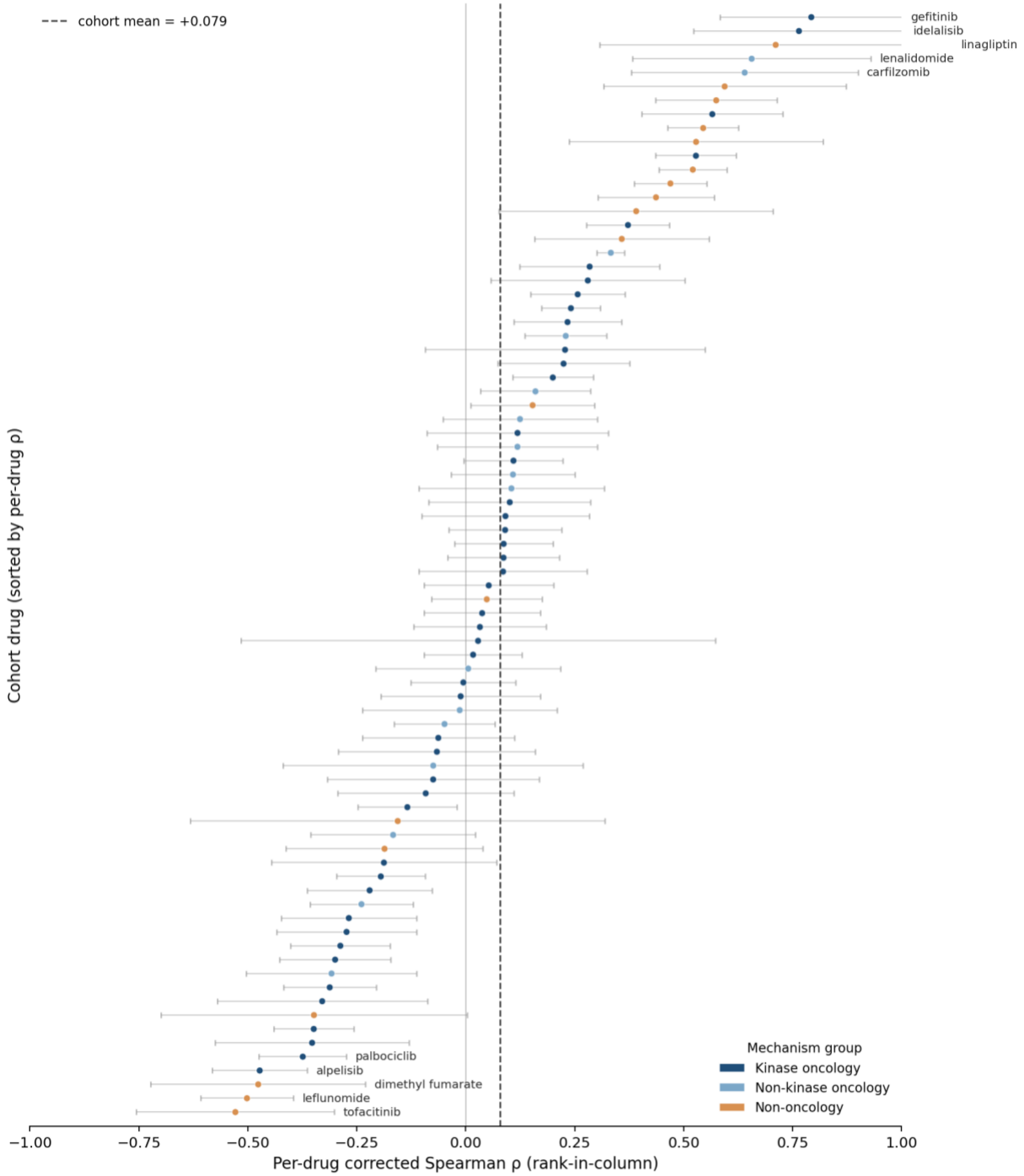


Figure 3-3. Per-drug corrected (rank-in-column) Spearman $\overline{\rho}$ across the 80 evaluable cohort drugs (≥ 3 labeled AEs each), sorted ascending by per-drug mean. Points are 500-seed bootstrap means; error bars are 500-seed bootstrap standard deviations. Color encodes mechanism group: deep blue, kinase oncology (EGFR/HER, VEGFR/multi-TKI, ALK/ROS1, BCR-ABL, PI3K/AKT/mTOR, CDK4/6, BRAF, KIT/PDGFR, JAK, BTK, IDH, FLT3, KRAS); mid blue, non-kinase oncology (cytotoxic, proteasome, IMiD, BCL2, aromatase, hedgehog, other); amber, non-oncology (antipsychotic, antimuscarinic, SSRI, SGLT2, DPP-4, sulfonylurea, antiretroviral, DMARD, MS, anticonvulsant, antiplatelet, GABA-B). Dashed vertical line marks the cohort mean of per-drug means (+0.079). Of 80 evaluable drugs, 48 (60%) have positive per-drug bootstrap mean. The five most-positive and five most-negative drugs are labeled by name.

The magnitude of the corrected per-drug signal ($\bar{\rho} \approx +0.08$) is notably smaller than what the legacy residual per-drug $\bar{\rho}$ suggests (+0.310 [+0.273, +0.350]). AE labels in FDA prescribing information are reported only when an event meets a manufacturer-defined frequency threshold, so a (drug, AE) pair is labeled selectively when the incidence is elevated above an implicit per-drug floor. Under this asymmetric labeling pattern, several intuitively reasonable per-drug metrics inflate spuriously. In particular, the legacy residual $\bar{\rho}$ variant, computed by subtracting each AE's leave-one-out cohort-mean incidence from both predicted and observed values then taking per-drug Spearman ρ of the residuals, yields $\bar{\rho} \approx +0.36$ even for a constant predictor. As a result, its absolute magnitude is dominated by an artifact of the asymmetric label presence selection and should not be assessed as a measure of drug-specific signal. It is retained in **Table 3-1** for transparency and as a sensitivity reference.

As an additional sensitivity analysis, we also report raw per-drug $\bar{\rho}$ (**Table 3-1**), which conflates drug-specific signal with the base-rate and overall toxicity correlation. On this metric the multi-task neural network achieves $\bar{\rho} = +0.383$ (95% CI [+0.216, +0.408]). The convergence of the three estimators on raw $\bar{\rho}$ but their separation on corrected $\bar{\rho}$ demonstrates that the corrected metric isolates the component of cross-drug ranking accuracy that is not explained by per-AE base rates.

Table 3-1. Cohort-level performance of the multi-task neural network versus a constant-scalar predictor, the per-AE cohort-mean baseline, and single-task ridge regression. Multi-task values are reported as mean across the 500-seed bootstrap with the across-seed 95% stability interval in brackets; ridge values are reported as the point estimate with 95% bootstrap CIs from a 10,000-resample cohort-level bootstrap.

Metric	Constant predictor	Per-AE cohort mean	Single-task ridge regression	Multi-task neural network
Macro MAE	-	-	0.084 [0.070, 0.103]	0.087 [0.085, 0.090]
Mean per-head Pearson r	-	-	-0.083 [−0.168, +0.004]	+0.022 [−0.016, +0.064]
Per-drug $\bar{\rho}$ (raw)	-	+0.399	+0.316 [+0.216, +0.408]	+0.383 [+0.334, +0.425]
Per-drug $\bar{\rho}$ (legacy residual)	+0.363	-	+0.128 [+0.049, +0.208]	+0.310 [+0.273, +0.350]
Per-drug $\bar{\rho}$ (corrected)	-	-	−0.080	+0.079 [+0.033, +0.127]

3.4.3 Per-drug predictions trace to known target-AE mechanisms

To test whether the model's predictions are mechanistically interpretable rather than reflecting statistical pattern-matching on AE base rates, we examined the relationship between each drug's PS binding profile and its predicted AE profile across three representative case studies: sunitinib (VEGFR/PDGFR/KIT/FLT3 inhibitor), imatinib (BCR-ABL inhibitor), and erlotinib (EGFR inhibitor) (**Figure 3-4**).

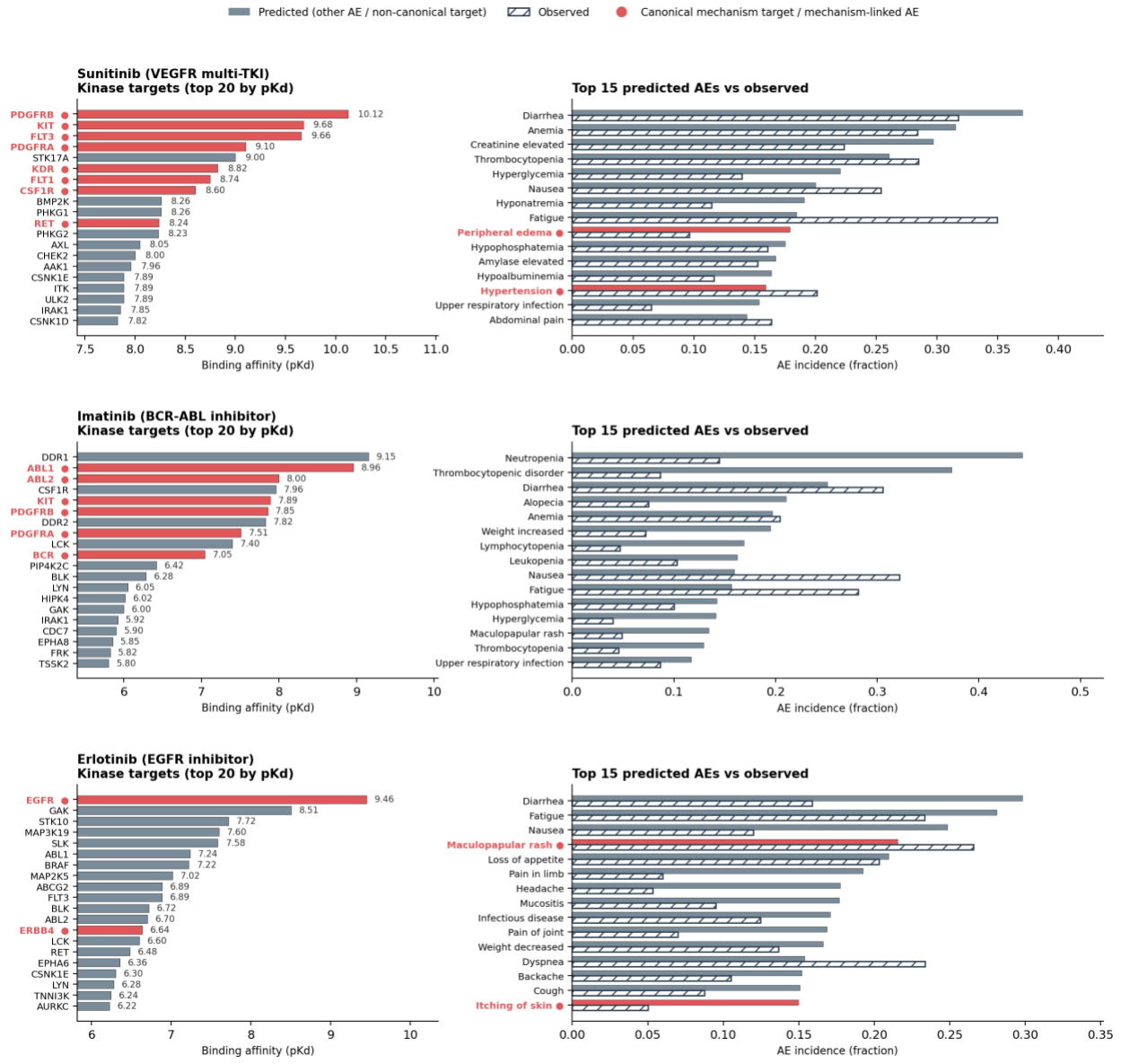


Figure 3-4. Mechanism case studies for three representative cohort drugs: sunitinib (VEGFR multi-TKI), imatinib (BCR-ABL inhibitor), and erlotinib (EGFR inhibitor). For each drug, the left panel shows the top 20 kinase targets ranked by binding affinity (pK_d), with canonical mechanism-class targets highlighted in red; the right panel shows

the top 15 model-predicted AEs ranked by predicted incidence (filled bars) alongside observed clinical-trial incidence (hatched bars), with canonical mechanism-linked AEs highlighted in red.

For sunitinib, seven of the top eight PS targets are members of the anti-angiogenic axis (KDR, FLT1, PDGFRA, PDGFRB, KIT, FLT3, CSF1R, all at $pK_d > 8.2$). The model's top predicted AEs mix base-rate-common AEs (diarrhea, anemia, creatinine elevation, thrombocytopenia) with mechanism-specific safety signals: hypertension (predicted 20% versus observed 16%) and peripheral edema (predicted 10% versus observed 18%), both canonical VEGFR2-pathway toxicities[30]. These predictions are elevated above the cohort base rate, demonstrating that the model leverages sunitinib's specific binding profile rather than predicting average toxicity. Sunitinib's corrected per-drug ρ is +0.30, well above the cohort mean of +0.08.

For erlotinib (EGFR-selective inhibitor), the highest affinity target is EGFR itself (pK_d = 9.46), followed by HER kinase family members. The top predicted AEs include EGFR-pathway dermatologic toxicities (maculopapular rash, xerosis) elevated above their cohort base rates, in addition to the universal gastrointestinal and constitutional AEs[31]. Erlotinib's corrected per-drug ρ is +0.26.

For imatinib (BCR-ABL/KIT/PDGFR), six of the top kinase targets (ABL1, ABL2, KIT, PDGFRA, PDGFRB, BCR) are established mechanism targets; the top predicted AEs are dominated by hematologic toxicities (neutropenia, thrombocytopenia, anemia, lymphocytopenia), consistent with imatinib's BCR-ABL-driven myelosuppression profile[32]. Imatinib's corrected per drug ρ is +0.12.

3.4.4 Block-level feature ablations: no single block is independently load-bearing

To quantify the contribution of each STAR feature block to the multi-task neural network's predictive performance, we ablated each of the four primary feature blocks (S, PS, EX, Cp) by zeroing its columns and retraining the model under the same 20-seed LOOCV bootstrap (**Table 3-2**).

Under the corrected (rank-in-column) per-drug $\bar{\rho}$ metric, no single block ablation degrades cohort discrimination beyond one standard deviation of the full-feature seed distribution. Removing Structure (S) drops the cohort mean from +0.079 (full bootstrap) to +0.066 ± 0.026; removing PS, EX, or Cp produces drops to +0.077 ± 0.023, +0.070 ± 0.017, and +0.077 ± 0.026 respectively. All four ablation means lie within one standard deviation of the full feature mean, indicating that the cohort-level discrimination signal is distributed across blocks

rather than concentrated in any single feature block. Marco MAE is largely unchanged across all ablations (within ± 0.001 of the full feature value of 0.087), confirming that block ablation moves the model only modestly along both the calibration and discrimination axes.

Table 3-2. Cohort-level performance under block-level feature ablations. Each row reports the cohort metric mean ± standard deviation across the 20-seed LOOCV bootstrap after zeroing the indicated STAR block. The full-feature row at the top is the 500-seed bootstrap (no ablation).

Configuration	Macro MAE	Per-head Pearson	$\bar{\rho}$ (raw)	$\bar{\rho}$ (legacy residual)	$\bar{\rho}$ (corrected)
Full features (500 seeds)	0.087 ± 0.001	+0.022 ± 0.020	+0.383 ± 0.023	+0.310 ± 0.020	+0.079 ± 0.024
No S (structure)	0.085 ± 0.001	+0.012 ± 0.014	+0.361 ± 0.019	+0.237 ± 0.021	+0.066 ± 0.026
No PS (binding)	0.084 ± 0.001	+0.042 ± 0.016	+0.366 ± 0.018	+0.319 ± 0.018	+0.077 ± 0.023
No EX (expression)	0.088 ± 0.001	+0.017 ± 0.025	+0.382 ± 0.021	+0.319 ± 0.019	+0.070 ± 0.017
No Cp (dose)	0.088 ± 0.001	+0.024 ± 0.023	+0.384 ± 0.023	+0.313 ± 0.019	+0.077 ± 0.026

This contrasts with the legacy residual $\bar{\rho}$ metric, on which the same set of ablations shows a seemingly large S-block effect. This effect is consistent with the diagnostic finding that the legacy residual $\bar{\rho}$ is partly an artifact of asymmetric label-presence selection; under the corrected metric the S-block effect attenuates to within seed noise.

3.4.5 Generalization across mechanism classes

To address the chemical-similarity leakage permitted by compound-level LOOCV, we additionally evaluated the model under LOOMCO cross-validation across 12 eligible mechanism classes (≥ 2 drugs each; **Table 3-3**). Pooled across 52 held-out drugs, the mean corrected per-drug $\bar{\rho}$ under LOOMCO is +0.167 (standard deviation: 0.391), with substantial cross-class variability indicating that generalization is not uniform across pharmacological contexts. Per-class performance is strongest for small classes within mechanism analogues in the training cohort (SGLT2 $\bar{\rho}$ = +0.750, n = 2; antipsychotic +0.620, n = 5; IMiD +0.408, n = 3), and for canonical kinase-inhibitor classes (EGFR/HER +0.250, n = 8; PI3K/AKT/mTOR +0.224, n = 6; VEGFR/multi-TKI +0.142, n = 6). Per-class performance is weakest for mechanism classes

whose AE patterns do not align with the rest of the cohort: proteasome inhibitors ($\bar{\rho}$ = −0.292, n = 3), BCR-ABL (−0.178, n = 3), and IDH (−0.118, n = 3). The large standard deviation of the pooled estimate reflects this heterogeneity rather than seed-level noise.

Table 3-3. Per-class mean corrected per-drug Spearman $\bar{\rho}$ under leave-one-mechanism-class-out cross-validation across 12 eligible mechanism classes (≥ 2 drugs). Classes are ordered by descending mean $\bar{\rho}$. The pooled-across-classes mean weights each held-out drug equally.

Held-out mechanism class	n drugs held out	n training drugs	Mean corrected $\bar{\rho}$
SGLT2	2	84	+0.750
Antipsychotic	5	81	+0.620
IMiD	3	83	+0.408
EGFR/HER	8	78	+0.250
PI3K/AKT/mTOR	6	80	+0.224
VEGFR/multi-TKI	6	80	+0.142
Cytotoxic	5	81	+0.076
ALK/ROS1	5	81	+0.012
KIT/PDGFR	3	83	-0.009
IDH	3	83	-0.118
BCR-ABL	3	83	-0.178
Proteasome	3	83	-0.292
Pooled across 52 held-out drugs	52	-	+0.167 (SD 0.391)

3.4.6 Negative controls and hyperparameter sensitivity

To verify that the multi-task neural network's corrected per-drug $\bar{\rho}$ reflects genuine drug-feature-AE coupling rather than fitting noise, we evaluated the model under two negative controls (**Table 3-4**). The first control, Y-permutation, shuffles which AE labels belong to which drug while preserving how frequently each AE is reported across the cohort; this destroys the drug-specific link but leaves the cohort-wise common-AE structure intact, so a model that learned only the per-AE base rates would score the same as on the real data. The second control, random-feature, replaces the drug's STAR inputs with marginal-statistic-matched random noise carrying no information about the compound. The primary prediction result, corrected per-drug

$\bar{\rho}$, collapses under both controls: from +0.079 (full model) to +0.028 under Y-permutation and to -0.081 under random-feature, confirming that the model depends on both a real drug-AE coupling and on the STAR features carrying actual pharmacological information. Calibration also degrades, with macro MAE rising from 0.087 (full model) to 0.097 (Y-permutation) to 0.126 (random-feature). The full model versus random-feature separation is unambiguous, whereas the full model versus Y-permutation separation is supportive but not strict, reflecting the small absolute magnitude of the cohort-level drug-specific signal at the present cohort size.

The per-head Pearson correlation, which measures how well the model predicts a single AE's incidence rankings across drugs, is higher under Y-permutation (+0.065) than under the full model (+0.022). Only 21 of 79 AE endpoints have ≥ 30 labeled drugs in the cohort, so most single-AE correlations are noisy regardless of model quality, and a model that simply learned each AE's cohort-mean incidence (which Y-permutation preserves by construction) can match or exceed the full model on this single-AE metric. The chapter's cohort-level discrimination claims are therefore indexed to the corrected per-drug $\bar{\rho}$ metric, which integrates information across all 79 AEs per drug rather than testing each in isolation; per-head Pearson is retained in **Table 3-4** for transparency but is not the operative null discriminator.

Table 3-4. Cohort metrics under two negative controls compared with the unperturbed multi-task neural network. Both controls were run with 20 LOOCV bootstrap seeds; the full model is shown with 500-seed across-seed intervals.

Configuration	Macro MAE	Per-head Pearson	$\bar{\rho}$ (raw)	$\bar{\rho}$ (legacy residual)	$\bar{\rho}$ (corrected)
Full model	0.087 [0.085, 0.090]	+0.022 [−0.016, +0.064]	+0.383 [+0.334, +0.425]	+0.310 [+0.273, +0.350]	+0.079 [+0.033, +0.127]
Y-permutation	0.097 [0.094, 0.101]	+0.065 [+0.005, +0.114]	+0.286 [+0.219, +0.335]	+0.124 [+0.062, +0.174]	+0.028 [−0.031, +0.100]
Random feature	0.126 [0.121, 0.132]	−0.118 [−0.196, −0.034]	+0.300 [+0.256, +0.348]	+0.223 [+0.163, +0.264]	−0.081 [−0.158, +0.001]

A 27-cell grid sweep over learning rate, dropout, and hidden dimensions at the training cohort scale (81 fits total at 3 seeds per cell) characterizes the multi-task neural network's sensitivity to hyperparameter choice. Ranked by corrected per-drug $\bar{\rho}$, the best configuration is

learning rate = 1×10^{-3}, a dropout of 0.3, hidden dimensions = (128, 64), which achieves corrected $\bar{\rho} = +0.091 \pm 0.025$ across three seeds. The worst configuration (learning rate 3×10^{-3}, dropout 0.5, hidden = (128, 64)) achieves $+0.015 \pm 0.007$. The dominant determinant of corrected $\bar{\rho}$ across the sweep is the joint pairing of dropout with learning rate: configurations with dropout 0.3 and intermediate learning rate (1×10^{-3}) cluster at the top, while configurations with dropout 0 or with the most aggressive learning rate cluster lower. The bootstrap configuration is at the top of the corrected-metric sweep, confirming that the headline result is not sensitive to a different (possibly cohort-specific) hyperparameter choice.

3.4.7 Magnitude calibration of predicted AE incidence

The multi-task neural network recovers per-AE rankings with limited absolute incidence calibration (**Figure 3-5**). Pooled across the 1,491 labeled (drug, AE) data, the expected ECE is 0.037, the macro Brier score is 0.017, and the standard deviation of predicted incidence is 0.70 times that of observed incidence, indicating that predictions are pulled approximately 30% toward the per-AE cohort mean relative to the observed spread. Per-head median ECE is 0.055.

In addition, calibration varies by AE class. Lab-finding AEs (hepatic, renal, electrolyte) calibrate best (per-class ECE 0.023-0.027), while symptom-based AEs (cardiovascular, neurologic, dermatologic) calibrate worst (per-class ECE 0.053-0.060). This pattern is consistent with laboratory AEs being reported in FDA labels with more consistent quantitative methodology than symptom-based AEs. Per-head isotonic post-calibration on the cohort data recovers an in-sample upper bound of approximately halved Brier score, but cross-validated calibration on a larger cohort is necessary for dose-finding quantitative use.

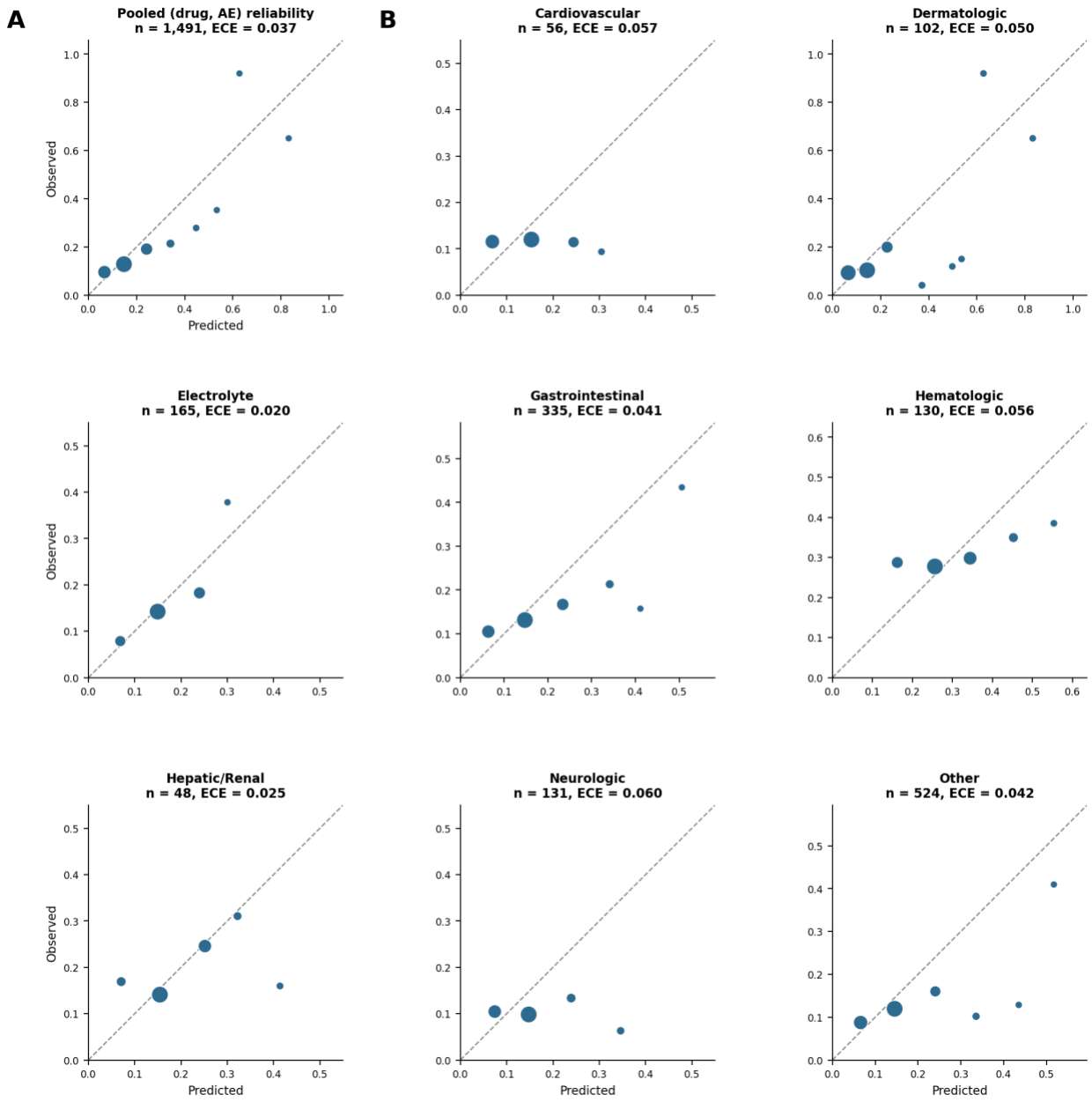


Figure 3-5. Calibration of the multi-task neural network's predicted AE incidence. **(A)** Pooled reliability diagram across all labeled (drug, AE) cells. Marker size encodes per-bin cell count; the dashed line marks perfect calibration. Pooled Brier score = 0.017 and predicted-to-observed standard-deviation ratio = 0.70 indicate that predicted incidence is compressed approximately 30% toward per-AE cohort means. **(B)** Per-AE-class reliability diagrams across eight pharmacological categories. Per-class cell count and ECE annotated above each panel.

3.4.8 Sample-size scaling and the case for cohort expansion

To characterize whether the framework would benefit from further cohort expansion, we evaluated cohort metrics at subsampled cohort sizes $N \in \{30, 45, 60, 75, 86\}$ drugs (**Figure 3-6**, **Table 3-5**). Across the range, the cohort macro MAE improves monotonically from 0.150 ± 0.031 at $N = 30$ to 0.087 ± 0.001 at $N = 86$, and raw per-drug $\bar{\rho}$ rises from +0.244 to +0.390.

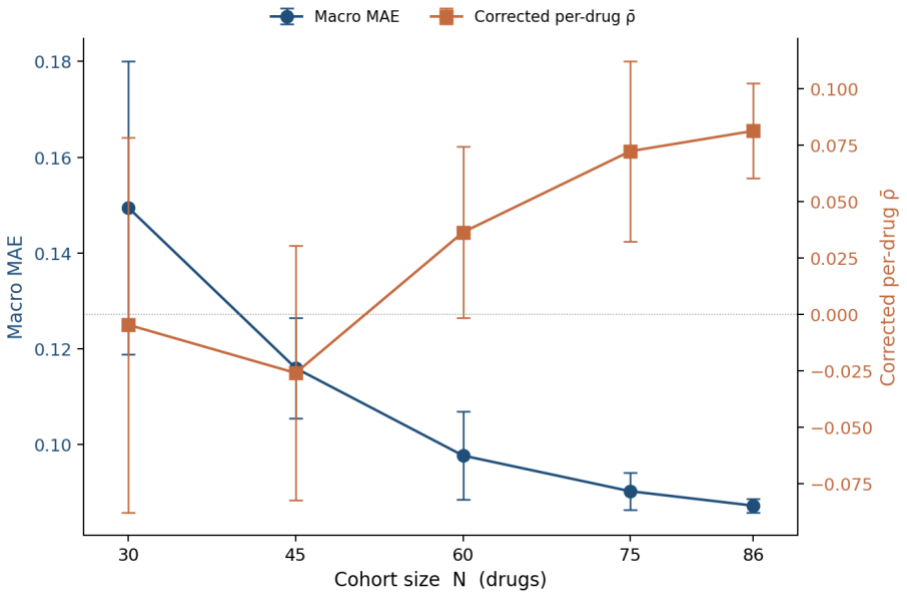


Figure 3-6. Sample-size scaling of cohort-level model performance. Macro mean absolute error (blue, left axis) and corrected per-drug Spearman $\bar{\rho}$ (copper, right axis) are reported as mean ± standard deviation across 20 LOOCV bootstrap seeds at each cohort size $N \in \{30, 45, 60, 75, 86\}$.

Under the corrected per-drug $\bar{\rho}$ metric, the cohort-level discrimination signal emerges with cohort size rather than appearing fully formed at small N. Corrected $\bar{\rho}$ is statistically indistinguishable from zero at $N = 30$ (-0.005 ± 0.083) and $N = 45$ (-0.026 ± 0.056), crosses into the positive regime by $N = 60$ ($+0.036 \pm 0.038$), and continues to grow to $+0.072 \pm 0.040$ at $N = 75$ and $+0.081 \pm 0.021$ at the full $N = 86$. The monotonic emergence is consistent with the multi-task shared trunk requiring a critical mass of compounds across mechanism classes to recover a useful joint representation; below $N \approx 60$ drugs the per-AE supervision is too sparse to identify drug-specific structure beyond per-AE base rates. The curve has not visibly plateaued at $N = 86$: the across-seed standard deviation at $N = 86$ (0.021) is smaller than the across-seed mean (+0.081), suggesting that cohort expansion is the most promising direction for improving cohort-level discrimination. We note that the legacy residual $\bar{\rho}$ in this table is essentially constant across N (from +0.32 to +0.30), masking the genuine sample-size dependence and providing further illustration that the legacy metric is dominated by the base-rate-asymmetry artifact rather than by drug-specific signal.

Table 3-5. Cohort metrics at subsampled cohort sizes $N \in \{30, 45, 60, 75, 86\}$ drugs, mean ± standard deviation across 20 LOOCV bootstrap seeds per cohort size.

***N* drugs**	***n* seeds**	**Macro MAE**	**$\bar{\rho}$ (raw)**	**$\bar{\rho}$ (legacy residual)**	**$\bar{\rho}$ (corrected)**
30	20	0.150 ± 0.031	+0.244 ± 0.095	+0.317 ± 0.067	-0.005 ± 0.083
45	20	0.116 ± 0.010	+0.290 ± 0.059	+0.287 ± 0.053	-0.026 ± 0.056
60	20	0.098 ± 0.009	+0.345 ± 0.039	+0.281 ± 0.038	+0.036 ± 0.038
75	20	0.090 ± 0.004	+0.369 ± 0.033	+0.295 ± 0.023	+0.072 ± 0.040
86	20	0.087 ± 0.001	+0.390 ± 0.015	+0.303 ± 0.017	+0.081 ± 0.021

3.5 Discussion

The STAR feature framework places small molecules spanning kinase, GPCR, nuclear receptor, proteasome, and other mechanism classes onto a common pharmacological footing for compound-level AE prediction, extending mechanism-aware approaches beyond the kinase inhibitor class within which they had previously been demonstrated. A shared-trunk multi-task neural network trained on this representation recovers a reproducible drug-specific predictive signal beyond what per-AE base rates explain (corrected per-drug Spearman ρ +0.079, 95% stability interval [+0.033, +0.127]) that is separable from constant, per-AE cohort mean, ridge, Y-permutation, and random feature null baselines. Per-drug case studies show that the framework operates at mechanistic resolution: sunitinib's predicted AE profile elevates hypertension and peripheral edema above the cohort base rates, consistent with its multi-target inhibition of the VEGFR axis[33] (KDR, FLT1, PDGFR, KIT, FLT3, CSF1R, all bound at pK_d > 8.2); erlotinib's elevates EGFR-pathway dermatologic toxicities including maculopapular rash and xerosis[31]; and imatinib's is dominated by hematologic toxicities consistent with BCR-ABL-driven myelosuppression[32]. In each case the predicted ranking reflects the drug's specific kinome binding profile rather than the cohort-average toxicity expectation, demonstrating that the multi-task trunk learns drug-specific structure rather than reproducing the per-AE base rate.

Cross-class generalization, evaluated under LOOMCO cross-validation, is markedly heterogenous and provides the clearest view of which compound classes the present framework can and cannot extend to. Strong per-class generalization concentrates in mechanism classes that share pharmacological context with the rest of the cohort, either through close mechanism analogues (SGLT2 inhibitors ρ +0.750, antipsychotics +0.620, IMiDs +0.408) or through well-

represented kinase inhibitor families with shared target-class AE patterns (EGFR/HER +0.250, PI3K/AKT/mTOR +0.224, VEGFR/multi-TKI +0.142). Weak or negative per-class generalization concentrates in mechanism classes whose AE patterns do not align with the rest of the cohort (proteasome inhibitors ρ -0.292, BCR-ABL -0.178, IDH -0.118). This pattern is the expected behavior of a multi-task architecture: the shared trunk transfers supervision where pharmacological similarity supports it and fails where it does not, providing a direct diagnostic of where cohort expansion is most needed.

A methodological contribution of this chapter is the identification and correction of an artifact in the per-drug Spearman ρ statistic when applied to AE labels with asymmetric label presence[34-36]. The rank-within-column percentile-normalized variant developed here removes the per-AE base-rate axis without imposing a parametric base rate model and returns the expected null behavior for a constant predictor, providing a base-rate-robust discrimination measure that future cross-class AE prediction work on FDA labeling data can adopt. Three additional statistical considerations frame the present results: the corrected ρ is computed against the 80 cohort drugs with ≥ 3 labeled AEs, a threshold chosen to ensure stable rank estimation; per-AE-head Pearson correlations are individually noisy across the 79 endpoints, so the cohort-level discrimination signal aggregates information that is below statistical significance at the per-head level; and under the corrected metric, no single STAR block ablation degrades discrimination beyond the seed-level noise floor, indicating that the cohort-level signal is broadly distributed across feature families rather than localized in any single block and that joint ablations or a larger cohort would be required to identify per-block contributions.

Several pharmacological features of the framework also bound how its predictions should be interpreted. The PS block remains kinome-centric, reflecting the relative density of publicly available kinome binding panels relative to other target families; compounds whose principal targets fall outside the kinase superfamily are represented largely through their incidental off-target kinase activity, which is sparse and incomplete. Target tissue engagement is proxied through population-mean expression in healthy donor tissue from the Human Protein Atlas and tumor tissue from UCSC Xena, without accounting for cell-type heterogeneity within tissues or for disease-state expression changes between healthy donor and patient tissue. Clinical dose is represented as a single population-level value standardized to 1.7 m^2 body surface area and 70 kg body weight, without representation of inter-patient pharmacokinetic variability or dose-response

curves, a notable simplification from the patient-level exposure modeling that anchored our earlier kinase inhibitor framework. The time-to-event characteristics of AEs are flattened to incidence, so rapid-onset and chronic-onset AEs at equal incidence are treated equivalently. AE labels themselves are noisy and partially confounded: incidence aggregates trial populations that differ in indication, comorbidities, concomitant medications, and demographics; reporting thresholds and AE-term granularity vary across manufacturers and regulatory submissions; AE incidence in oncology labels is partially confounded by underlying disease progression and by AEs of concomitant cytotoxic chemotherapy; and rare or idiosyncratic AEs (immune-mediated reactions, pharmacogenomically-driven hepatotoxicity, severe cutaneous reactions) are systematically underrepresented at the population-incidence level reported in prescribing labels and therefore outside the framework's resolution regardless of feature quality[37,38].

The sample-size scaling result is the strongest argument that the present cohort is below the scale at which the STAR framework reaches its full predictive potential. Cohort-level corrected ρ is statistically indistinguishable from zero at $N = 30$ and $N = 45$, crosses into the positive regime by $N = 60$, and continues to grow monotonically through $N = 86$ without visible plateau. The highest-leverage direction for cohort expansion is therefore the addition of compounds from mechanism classes currently sparse in the cohort (proteasome, BCR-ABL, IDH, KIT/PDGFR), with secondary value in densifying the kinase inhibitor backbone for which the present cohort already provides reasonable coverage. Broader extensions, including expansion of the PS target data beyond the kinome, incorporation of patient-level pharmacokinetic descriptors into the Cp block, refinement of the AE outcome standardization, prospective external validation, and integration with chemical-language or graph neural network molecular embeddings, are promising directions for future work.

3.6 References

1. Gong X, Hu M, Liu J, et al. Decoding kinase-adverse event associations for small molecule kinase inhibitors. *Nat Commun*. Jul 27 2022;13(1):4349. doi:10.1038/s41467-022-32033-5
2. Jusko NM, Cao A, Sun D. Machine Learning Model Predicts Clinical Adverse Events of Small Molecule Kinase Inhibitors in Cancer Patients Using On-/Off-Target Engagement and Tissue Selectivity. *Clin Pharmacol Ther*. May 26 2026;doi:10.1002/cpt.70343
3. Galeano D, Paccanaro A. Machine learning prediction of side effects for drugs in clinical trials. *Cell Rep Methods*. Dec 19 2022;2(12):100358. doi:10.1016/j.crmeth.2022.100358

4. Zhao H, Wang S, Zheng K, Zhao Q, Zhu F, Wang J. A similarity-based deep learning approach for determining the frequencies of drug side effects. *Brief Bioinform*. Jan 17 2022;23(1)doi:10.1093/bib/bbab449
5. Xu X, Yue L, Li B, et al. DSGAT: predicting frequencies of drug side effects by graph attention networks. *Brief Bioinform*. Mar 10 2022;23(2)doi:10.1093/bib/bbab586
6. Sharma B, Chenthamarakshan V, Dhurandhar A, et al. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. *Sci Rep*. Mar 25 2023;13(1):4908. doi:10.1038/s41598-023-31169-8
7. Lounkine E, Keiser MJ, Whitebread S, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*. Jun 10 2012;486(7403):361–7. doi:10.1038/nature11159
8. Bowes J, Brown AJ, Hamon J, et al. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat Rev Drug Discov*. Dec 2012;11(12):909–22. doi:10.1038/nrd3845
9. Anastopoulos IN, Herczeg CK, Davis KN, Dixit AC. Multi-Drug Featurization and Deep Learning Improve Patient-Specific Predictions of Adverse Events. *Int J Environ Res Public Health*. Mar 5 2021;18(5)doi:10.3390/ijerph18052600
10. Krix S, DeLong LN, Madan S, et al. MultiGML: Multimodal graph machine learning for prediction of adverse drug events. *Heliyon*. Sep 2023;9(9):e19441. doi:10.1016/j.heliyon.2023.e19441
11. Bean DM, Wu H, Iqbal E, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci Rep*. Nov 27 2017;7(1):16416. doi:10.1038/s41598-017-16674-x
12. Zhao H, Ni P, Zhao Q, et al. Identifying the serious clinical outcomes of adverse reactions to drugs by a multi-task deep learning framework. *Commun Biol*. Aug 24 2023;6(1):870. doi:10.1038/s42003-023-05243-w
13. Cao A, Zhang L, Bu Y, Sun D. Machine Learning Prediction of On/Off Target-driven Clinical Adverse Events. *Pharm Res*. Aug 2024;41(8):1649–1658. doi:10.1007/s11095-024-03742-x
14. Sun D, Macedonia C, Chen Z, et al. Can Machine Learning Overcome the 95% Failure Rate and Reality that Only 30% of Approved Cancer Drugs Meaningfully Extend Patient Survival? *J Med Chem*. Sep 26 2024;67(18):16035–16055. doi:10.1021/acs.jmedchem.4c01684
15. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B*. Jul 2022;12(7):3049–3062. doi:10.1016/j.apsb.2022.02.002
16. Morgan HL. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*. 1965/05/01 1965;5(2):107–113. doi:10.1021/c160017a018
17. Davis MI, Hunt JP, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol*. Oct 30 2011;29(11):1046–51. doi:10.1038/nbt.1990
18. Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J Chem Inf Model*. Mar 24 2014;54(3):735–43. doi:10.1021/ci400709d
19. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. Jan 4 2016;44(D1):D1045–53. doi:10.1093/nar/gkv1072

20. Tang J, Tanoli ZU, Ravikumar B, et al. Drug Target Commons: A Community Effort to Build a Consensus Knowledge Base for Drug-Target Interactions. *Cell Chem Biol*. Feb 15 2018;25(2):224–229 e2. doi:10.1016/j.chembiol.2017.11.009
21. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. Jan 23 2015;347(6220):1260419. doi:10.1126/science.1260419
22. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol*. Jun 2020;38(6):675–678. doi:10.1038/s41587-020-0546-8
23. Kass-Hout TA, Xu Z, Mohebbi M, et al. OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. *J Am Med Inform Assoc*. May 2016;23(3):596–600. doi:10.1093/jamia/ocv153
24. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. Jan 1 2004;32(Database issue):D267–70. doi:10.1093/nar/gkh061
25. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279–90.
26. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc.; 2019:Article 721.
27. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. presented at: International Conference on Learning Representations (ICLR); 2019; New Orleans, LA, USA.
28. *RDKit: Open-source cheminformatics*. Version 2024.03. 2024. https://www.rdkit.org
29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12(85):2825–2830.
30. Robinson ES, Khankin EV, Choueiri TK, et al. Suppression of the nitric oxide pathway in metastatic renal cell carcinoma patients receiving vascular endothelial growth factor-signaling inhibitors. *Hypertension*. Dec 2010;56(6):1131–6. doi:10.1161/HYPERTENSIONAHA.110.160481
31. Joshi SS, Ortiz S, Witherspoon JN, et al. Effects of epidermal growth factor receptor inhibitor-induced dermatologic toxicities on quality of life. *Cancer*. Aug 15 2010;116(16):3916–23. doi:10.1002/cncr.25090
32. Druker BJ, Talpaz M, Resta DJ, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N Engl J Med*. Apr 5 2001;344(14):1031–7. doi:10.1056/NEJM200104053441401
33. Schmidinger M. Understanding and managing toxicities of vascular endothelial growth factor (VEGF) inhibitors. *EJC Suppl*. Sep 2013;11(2):172–91. doi:10.1016/j.ejcsup.2013.07.016
34. Spearman C. The proof and measurement of association between two things. By C. Spearman, 1904. *Am J Psychol*. Fall–Winter 1987;100(3-4):441–71.
35. Heckman JJ. Sample Selection Bias as a Specification Error. *Econometrica*. 1979;47(1):153–161. doi:10.2307/1912352
36. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. Jan 22 2003;19(2):185–93. doi:10.1093/bioinformatics/19.2.185
37. Seruga B, Sterling L, Wang L, Tannock IF. Reporting of serious adverse drug reactions of targeted anticancer agents in pivotal phase III clinical trials. *J Clin Oncol*. Jan 10 2011;29(2):174–85. doi:10.1200/JCO.2010.31.9624

38. Uetrecht J. Idiosyncratic drug reactions: past, present, and future. *Chem Res Toxicol*. Jan 2008;21(1):84–92. doi:10.1021/tx700186p

Chapter 4: Perspectives on STAR-Guided AI/ML Prediction of Clinical Drug Toxicity and Future Directions

4.1 Summary

4.1.1 The translational gap and the limits of existing computational approaches

Clinical drug toxicity accounts for approximately 30% of attrition in late-stage clinical trials and remains the leading cause of post-approval withdrawal of marketed drugs[1,2]. Despite decades of investment in preclinical safety assessment, fewer than one in seven drug candidates entering Phase 1 trials reaches FDA approval, with concordance between preclinical animal models and human target-organ toxicity falling below 30% for the toxicity categories that most often drive clinical attrition[3,4]. The persistence of this gap has motivated a turn toward computational prediction of clinical toxicity from preclinical data, with artificial intelligence and machine learning (AI/ML) approaches positioned as the methodology most likely to close the translational gap at scale. The predictive capacity of these approaches, however, is bounded not by algorithmic sophistication but by the scope of the pharmacological information available to them.

Existing computational frameworks integrate chemical structure with target binding data or reported adverse event (AE) associations, but they do not adequately represent the complex mechanisms through which a molecular liability becomes a clinical adverse event. Target binding affinity alone is insufficient: a compound that potently inhibits a target may produce no clinical toxicity if the target is not expressed in the relevant tissue, if the drug does not distribute to that tissue at therapeutic concentrations, or if the clinical dose does not achieve exposures capable of engaging the target at all. Among small molecule kinase inhibitors with VEGFR2 inhibition, for example, only those with an exposure margin below approximately 13-fold show elevated hypertension reports in FAERS, despite all of them engaging the target *in vitro*[5]. A computational framework that does not represent target engagement, tissue distribution and

expression, and clinical exposure jointly will produce false positives at the target-engagement level and false negatives at the clinical outcome-level.

4.1.2 Integrating pharmacological context into AI/ML toxicity prediction

The Structure-Tissue/cell selectivity-Activity Relationship (STAR) framework[1,6] decomposes the causal chain from administered drug molecule to clinical toxicity into five interdependent variables: chemical structure, which determines the molecular interactions a compound can form; potency and specificity, which determine the strength and selectivity of those interactions across on- and off-targets; target expression in normal versus diseased tissue, which determines where in the body engagement is biologically meaningful; tissue selectivity, which determines whether the drug distributes to those tissues at therapeutically relevant concentrations; and the clinical dose required to reach therapeutic exposure, which determines whether the administered regimen achieves the concentrations necessary for engagement *in vivo*. Clinical toxicity emerges from the joint state of all five variables, with each variable's contribution conditioned on the others. Target engagement is contingent on local unbound drug concentrations in the relevant tissue, which is in turn contingent on systemic exposure, tissue-specific distribution, and the underlying expression of the target in that tissue.

For the small molecule kinase inhibitor class, as established in Chapter 2, a random survival forest framework integrating patient-level pharmacokinetics, kinome-wide on- and off-target binding affinities, and tissue-specific drug distribution recovers patient-level temporal risk estimates for clinical AEs, with particular gains for tissue-localized toxicities. The four STAR variables included there (potency and specificity, target expression, tissue selectivity, and dose) were sufficient to demonstrate that integrating tissue-weighted exposure with on-/off-target engagement improves discrimination over plasma exposure alone for rash, nausea, fatigue, and headache (median C-index 0.669 to 0.814), while plasma exposure remains the appropriate predictor for severe hematologic toxicities arising in circulating cells (median C-index 0.694 to 0.880). Variance importance analysis recovered established target-AE associations (EGFR with rash[7], VEGFR with diarrhea and hypertension[8,9], FLT1 with fatigue[10]) and identified candidate associations that warrant mechanistic follow-up. External validation on neratinib, a drug withheld from training, reproduced trial-level safety patterns with 100% sensitivity and >95% specificity at a 10% incidence threshold and Pearson correlations >0.87 between predicted and observed AE frequencies.

Chapter 3 extended the framework from a single mechanism class to a cohort of FDA-approved small molecules spanning 20 oncology and 14 non-oncology mechanism classes, replacing patient-level time-to-event prediction with compound-level prediction in a multi-task neural network. Each compound was represented by a unified 3,170-dimensional STAR feature vector spanning all five framework variables: a Structure block (S, Morgan fingerprints and physicochemical descriptors from SMILES); a Potency/Specificity block (PS, pKd values and measured-presence masks across genes pooled from Davis et al.[11], KIBA[12], BindingDB[13], and Drug Target Commons[14]); an Expression block summarizing target-gene expression in normal and tumor tissues from the Human Protein Atlas[15] and UCSC Xena TCGA[16] (EX); a Tissue/cell selectivity block computed as binding-strength-weighted expression scores following Cao et al.[17] (TS); and a Clinical dose block derived from FDA prescribing labels (Cp). AE labels were extracted from the adverse-reactions section of openFDA[18] prescribing information and mapped to SNOMED CT[19] concepts. Leveraging the full STAR framework, the multi-task neural network recovered a reproducible drug-specific predictive signal beyond per-AE base rates (cohort-mean corrected per-drug Spearman $\rho = +0.079$, 95% stability interval [+0.033, +0.127]), cleanly separated from constant, per-AE cohort mean, ridge, Y-permutation, and random-feature baselines. Per-drug case studies confirmed that the predictions trace to mechanism; sunitinib's predicted AE profile elevates hypertension and peripheral edema above the cohort base rate, consistent with its multi-target inhibition of the VEGFR axis; erlotinib's elevates EGFR-pathway dermatologic toxicities; and imatinib's is dominated by hematologic toxicities consistent with BCR-ABL-driven myelosuppression. Leave-one-mechanism-class-out cross-validation across 12 eligible classes returned a pooled mean corrected ρ of +0.167 with strong generalization to mechanism analogues of the training cohort (SGLT2 +0.750, antipsychotics +0.620, immunomodulatory drugs +0.408; EGFR/HER +0.250, PI3K/AKT/mTOR +0.224, VEGFR/multi-TKI +0.142).

Taken together, both implementations of the STAR framework demonstrate that a unified representation of the interdependent determinants of toxicity recovers patient-level temporal risk within the kinase inhibitor class and compound-level incidence across mechanism classes, resolving safety signals that target binding or chemical structure alone cannot capture. Three contexts now bear on the framework's onward development: the drug development decisions it

can usefully inform, the boundaries of its current applicability, and the extensions that would consolidate it into a translational tool.

4.2 Implications

4.2.1 STAR as a triage tool for pre-first-in-human safety prediction

The decisions made between lead optimization and first-in-human (FIH) dosing set the trajectory of a drug development program. Once a candidate advances, the sponsor must assemble the investigator's brochure, file an Investigational New Drug (IND) application, design the Phase 1 protocol, justify the starting dose, and establish a monitoring strategy. Each of these in turn depends on a predicted safety profile that is currently assembled from a heterogenous mix of inputs, including Good Laboratory Practice (GLP)-compliant animal toxicology studies, *in vitro* secondary pharmacology panels of variable target coverage, chemical structure-based liability assessments, and, where available, exposure margins from prior-in-class agents. With clinical development now accounting for the majority of the approximately $2.6 billion cost of bringing a new drug to market[20], the value of a triage step that re-ranks pre-FIH candidates on the basis of likely clinical AE profile is large in expectation even when the per-candidate prediction is imperfect.

The triage role STAR is positioned for is defined by what current pre-FIH safety screenings leave unanswered. ICH S7A[21] mandates a core battery of *in vivo* animal studies covering cardiovascular, central nervous system, and respiratory function in animals, with hERG and *in vivo* QT assessment governed by the related ICH S7B[22]. These regulatory guidelines address acute physiological function but not the broader range of organ-specific and tissue-localized AEs that drive most clinical attrition. The guideline itself is now over two decades old and is the subject of an active revision effort that explicitly calls for integration of human-relevant *in vitro* and *in silico* methods[23]. Outside regulatory standards, industry has converged on secondary pharmacology screening against curated target panels: the 44-target Bowes consensus introduced in 2012[24] and the recently expanded 77-target consortium panel from the IQ DruSafe group[25] are widely used and increasingly requested in submissions[26], though neither is required by ICH S7A. These panels return a binding liability vector that identifies which targets the candidate engages but does not commit to which engagements will produce clinical AEs, at what incidence, or in which patient populations. The interpretive step from panel output to predicted

AE profile is currently performed by safety pharmacologists drawing on prior-in-class experience and on independent computational tools, none of which integrates the panel output with tissue exposure and dose into a single prediction. By contrast, the STAR framework collapses this interpretive step into the model: rather than returning a target-engagement vector for downstream translation, it returns predicted incidence directly on the AE endpoints that prescribing-label safety sections, Phase 1 stopping rules, and clinical monitoring plans are written against, with the prediction conditioned on the therapeutic exposure context in which the candidate will actually be dosed.

The use case the framework's present predictive performance most directly supports, consistent with current model-informed drug development (MIDD) practice[27], is a model-based tool for candidate selection committees and Phase 1 protocol design teams that ranks compounds on relative AE risk and flags tissue-localized toxicities for prospective monitoring, alongside rather than in place of the *in vitro* and *in vivo* animal data such teams already consider. Panel-based screening and the STAR framework are thus complementary rather than competing, with panel screening detecting target-level liability without representing exposure context and STAR predicting clinical toxicity under exposure context across the AE endpoints that frame downstream decisions.

4.2.2 Mechanistic interpretability as a candidate-design feedback tool

Current explainability methods for toxicity prediction operate on chemical structure inputs and produce attributions over fingerprint bits, graph nodes, or atom-level features[28-30]. As Chapter 1 describes, these attributions identify statistical associations between structural motifs and toxicity outcomes rather than the mechanisms by which those motifs manifest as clinical AEs. STAR's interpretability, in contrast, is anchored at the level of pharmacological mechanism: feature blocks correspond to distinct biological constructs (target engagement, tissue-plasma partitioning, structural liability, clinical exposure), and per-block ablations and per-target attributions return interpretations in terms a medicinal chemist or safety pharmacologist can act upon. The STAR-guided random survival forest framework in Chapter 2 demonstrated this concretely through its variable importance analysis, which recovered EGFR as the predominant predictor of dermatologic AEs, VEGFR family members as predictors of hypertension and gastrointestinal events, and FLT1 as a predictor of fatigue, providing not only attributions but candidate target-AE associations warranting mechanistic follow-up.

With this mechanistic interpretability leveraged as a candidate design feedback tool, a program optimizing a new compound can evaluate not only whether the optimized scaffold removes a target-level liability flagged by a secondary pharmacology panel, but also whether the STAR model's predicted AE incidence on the relevant endpoint shifts in response. Where the prediction shift is conditional on tissue and exposure context, the optimization becomes mechanistically tractable rather than reduced to a binary structural alert outcome. For tissue-localized AEs, the model's attribution surface decomposes by design into target engagement and tissue distribution components, enabling design-stage triage of which liability axis (off-target affinity, tissue penetration, dose margin) is the most cost-effective to address for any given candidate.

The mechanistic interpretability of the STAR framework also positions it as a hypothesis-generation tool for the downstream target-validation work that establishes whether candidate target-AE associations are causal. Chapter 2's variable importance analysis identified specific candidate associations of this kind, including ABL1 as a contributor to dermatologic AEs and PLK4 and AURKB as brain-localized predictors of fatigue, none of which is established to the same degree as the EGFR-rash and VEGFR-hypertension associations the model also recovered. The same approach extends naturally to the cross-class cohort of Chapter 3, where per-AE attribution over the potency and specificity block can surface candidate target-AE pairs across non-kinase target families that can be confirmed *in vitro* or in animal models. Within this framing, the framework's predictions are most usefully treated as ranked hypotheses about target-mediated AE etiology rather than as standalone predictions, with the interpretability serving as the bridge between cohort-scale ML signal and the compound- or target-scale experimental validation that establishes mechanism.

4.2.3 Regulatory alignment: STAR within the FDA's New Approach Methodologies framework

The regulatory environment for *in silico* toxicology has evolved considerably in the past five years. The FDA Modernization Act 2.0 (December 2022) removed the statutory requirement that all drug applications include animal testing data and explicitly recognized *in silico* methods as one class of acceptable alternative[31]. The FDA's 2025 Roadmap to Reducing Animal Testing in Preclinical Safety Studies operationalizes this by defining New Approach Methodologies (NAMs) as *in vitro* human-based systems, organ-on-chip platforms, and *in silico* predictive

models, with the FDA's stated goal of integrating qualified NAMs into IND and New Drug Application (NDA) review[32]. Within this framework, computational toxicology submissions are governed by an emerging set of fit-for-purpose qualification expectations: a NAM is qualified for a specific context of use (a defined AE endpoint, a defined drug class, a defined decision the prediction supports) rather than as a general purpose tool, and qualification rests on demonstrated predictive performance under defined uncertainty bounds, transparent methods, and reproducible implementation[33].

The STAR framework aligns with the FDA's NAM guidelines along three specific axes. First, the framework is mechanistically interpretable at the regulatory level expected of qualified NAMs. Predicted toxicity outcomes are explicitly conditioned on structure, target engagement, tissue distribution, and dose, and per-block ablations identify which features drive predictions. This contrasts with structure-only deep learning approaches, which can deliver comparable or higher discrimination metrics but provide attributions only at the structural level[34,35], an interpretive resolution that has historically been a barrier to regulatory adoption. Second, the framework's uncertainty quantification is explicitly defined. Bootstrap confidence intervals, per-class generalization metrics (LOOMCO), and per-prediction calibration metrics (**3.3.8**) are reported alongside point estimates, providing the uncertainty-bounded outputs that NAM qualification expects. Third, the framework's training data are drawn from public sources using documented and reproducible pipelines, supporting the reproducibility expectations of the Drug Development Tool qualification process[33].

Two specific qualification pathways align with the framework's present state. The Drug Development Tool (DDT) qualification program through CDER's MIDD pilot accepts model-based decision support tools for defined contexts of use, with biomarker, clinical outcome assessment, and animal model qualification programs as precedents. The STAR framework's per-AE class calibration profile directly suggests an initial context of use focused on laboratory-finding AEs (hepatic, renal, electrolyte) where calibration is strongest and clinical thresholds are quantitatively defined. The second pathway is the ongoing ICH S7A revision, which the FDA has identified as the most likely near-term regulatory mechanism for incorporating qualified *in silico* safety methods into preclinical safety assessment alongside ICH-required animal studies[23]. STAR's positioning as a complement to secondary pharmacology panel screening rather than a

replacement for animal toxicology aligns with the integration model the ICH S7A revision envisions.

4.3 Limitations

4.3.1 Feature coverage as the framework's empirical bottleneck

The cohort-level corrected per-drug $\bar{\rho}$ of +0.079 reported in **3.3.2** is modest in absolute magnitude. Three candidate sources of this ceiling warrant examination: a limitation of the multi-task neural-network architecture, a limitation of the cohort scale, or a limitation of the STAR feature representation itself. The analysis points to feature coverage rather than architecture as the principal bottleneck. The block-level ablations demonstrated that no single STAR feature block independently dominates discrimination, indicating that the cohort-level signal is broadly distributed across the available features. The sample size scaling showed monotonic emergence of corrected $\bar{\rho}$ with cohort size without plateau at $N = 86$, indicating that inclusion of additional drug compounds would yield further discrimination signal. Both findings position the framework's headline magnitude as scaling with the addition of further compounds to the training cohort, rather than as a ceiling imposed by the model architecture.

The potency and specificity block is the most consequential feature coverage bottleneck. The 469-gene unified kinase dataset is dense for kinase inhibitors (per-drug median is 10 measured targets, with deeply profiled compounds such as sunitinib reaching 376 targets) but sparse for compounds whose principal targets fall outside the kinase superfamily. The LOOMCO validation demonstrated that for proteasome inhibitors, BCR-ABL inhibitors, and IDH inhibitors, all classes whose targets are absent or represented only by a few kinases, were the mechanism classes where the framework most under-performed (corrected $\bar{\rho}$ between -0.118 and -0.292). Expanding the potency and specificity feature block is therefore the most direct route to improving the framework's predictive ceiling. Several target families are under-covered relative to their pharmacological importance: G-protein-coupled receptors (GPCRs) are the largest drug-target family in approved therapeutics and are essentially absent from the present PS vocabulary; nuclear hormone receptors are sparsely represented; the proteasome and other non-kinase enzyme families are nearly absent; ion channel families (Nav, Cav, KCNQ, hERG) are present only insofar as some compounds in the cohort have been profiled against them in BindingDB; and drug-metabolizing enzymes (CYP450, UGT) are absent. Resources for filling these gaps

exist[36,37], with the barrier being the heterogenous reporting of binding affinity units, assay formats, and target ontologies across data sources. A second-generation potency and specificity block that combines the kinome datasets with a curated GPCR panel, nuclear receptor panel, ion channel panel, and a CYP-inhibition module would substantially close this coverage gap, and would correspondingly extend the cohort's pharmacological diversity, increasing the framework's expected generalization across non-kinase mechanism classes.

4.3.2 Magnitude calibration as a constraint of label quality rather than model capacity

The calibration analysis in Chapter 3 established that the multi-task neural network's predicted incidence is compressed approximately 30% toward per-AE cohort means relative to observed incidence (predicted-to-observed standard-deviation ratio of 0.70), with pooled expected calibration error (ECE) of 0.037 and per-class ECE ranging from 0.020 (electrolyte) to 0.060 (neurologic). This compression towards the mean is a direct consequence of the masked binomial-N-weighted binary cross-entropy training objective, which regularizes predictions toward per-AE base rates when per-AE base rates when per-AE supervision is sparse, and is the architectural mechanism that supports the framework's role as a discrimination tool rather than as a quantitative dose prediction tool. At issue is whether the calibration limit is a property of the model that improved architecture can address, or a property of the labels that improved training cannot.

Several lines of evidence point to label properties rather than model capacity as the principal calibration constraint. First, the per-AE class breakdown in **3.3.8** shows that calibration quality varies substantially across AE classes, with laboratory-finding AEs calibrating better than symptom-based AEs. This variation is consistent with how AE incidence is measured and reported across compounds: laboratory AEs are defined by quantitative thresholds (lab value above or below reference range, specified in Common Terminology Criteria for Adverse Events, CTCAE or in label-specific terms) and reported with consistent denominators (n patients with at least one out-of-range measurement out of n at-risk patients), whereas symptom-based AEs are reported qualitatively by patients, aggregated heterogeneously across trials, and subject to substantial inter-trial variability in how the same clinical phenomenon is captured in case report forms (CRFs) and ultimately in the prescribing label. The model's calibration cannot exceed the consistency of its training data, and the per-class calibration variation reflects this. Second, AE incidence derived from FDA labeling is itself a noisy and partially confounded outcome variable.

Patient populations aggregated across indications, comorbidities, concomitant medications, demographic variables, variable reporting thresholds across manufacturers and AE classes, disease progression, and the systematic underrepresentation of rare and idiosyncratic AEs are all sources of noise that are fixed at the data source and are not removable by retraining or by architectural changes[38].

Third, in-sample per-head isotonic post-calibration reduces median Brier score by approximately half, indicating that the residual calibration error after the multi-task neural network's training is partly correctable by post-hoc rescaling. This isotonic recovery is, however, an in-sample upper bound, where under cross-validated calibration recovery would be smaller. The implication for the STAR framework's practical use is that the predicted incidence values should be interpreted as ranks within the per-AE cohort rather than as absolute incidence estimates, and where absolute incidence is the required output, the framework should be paired with explicit per-AE post-calibration on independent data. Where the label source can be improved – for example, by training on patient-level AE outcomes from registrational trial datasets rather than on aggregated prescribing label incidence, or by training on AE endpoints defined with consistent quantitative thresholds – the calibration ceiling can be expected to lift correspondingly. The calibration limit in the framework's present form is therefore label-bounded rather than architecture-bounded, and the appropriate architectural response is to preserve the discrimination quality the framework already achieves while improving the label inputs that constrain absolute calibration.

4.3.3 Prospective generalization as the open empirical question

Two cross-validation protocols anchor the STAR framework's internal evidence for generalization. Compound-level leave-one-out cross-validation (LOOCV) holds each drug out in turn and evaluates the model on that drug's AE labels using a model trained on the remaining drugs, providing per-drug out-of-sample predictions across the cohort. Leave-one-mechanism-class-out cross-validation (LOOMCO) holds out entire mechanism classes simultaneously, providing a stricter test of generalization to compound classes whose chemical and pharmacological neighbors were absent from training. Both protocols are necessary and informative, but neither constitutes prospective generalization to compounds drawn from outside the training cohort.

The mechanism-class heterogeneity observed under LOOMCO suggests that generalization is strong (corrected $\bar{\rho} > 0.4$) where the held-out class has close pharmacological analogues in the training cohort (SGLT2, antipsychotics, immunomodulatory drugs) and where the held-out class is itself well represented in the cohort. Generalization is negative where the held-out class has no close analogue in the training cohort. The framework's prospective generalization on a new mechanism class is therefore expected to depend on whether the prospective class has in-cohort analogues, and the LOOMCO matrix provides a heuristic for which classes are likely to generalize well, and which would require cohort expansion before prospective use.

A separate constraint on prospective external evaluation is the prediction-output calibration limit. The compressed predicted incidence spread means that, for a compound whose true AE pattern differs substantially from cohort averages, the model's predictions cluster too narrowly to support the within-column rank normalization that the corrected per-drug $\bar{\rho}$ requires. Strict time-stratified prospective validation on post-2024 new molecular entities therefore require two prerequisites: cohort expansion in the mechanism classes previously identified, so that an external set has in-cohort pharmacological analogues; and treatment of the prediction-output calibration limit, either through architectural changes or through post-hoc rescaling against an independent validation set.

What remains to be established is the cohort scale and feature breadth at which the framework achieves prospective generalization on a defined population of new molecular entities. The sample size scaling analysis projects that a cohort of approximately N = 200 compounds with broader non-kinase coverage would push corrected $\bar{\rho}$ into the +0.15 to +0.20 range, at which magnitude prospective evaluation against a well-matched external set becomes meaningfully informative. Whether the framework achieves qualification as a clinical safety prediction tool therefore depends jointly on cohort scale expansion, potency and specificity coverage expansion, and the prospective validation that these expansions enable.

4.4 Future Directions

4.4.1 Empirical extensions: expanding the cohort beyond oncology

The current STAR framework's training data comprises 86 FDA-approved small molecules, with oncology overrepresented relative to its share of the FDA-approved small

molecule population as a consequence of the potency and specificity inclusion criterion that defines the final cohort. Expanding this cohort along three specific axes is expected to be the most informative direction for the STAR framework's discrimination magnitude and pharmacological breadth.

The first axis is densifying mechanism classes that LOOMCO identified as under-represented, including proteasome inhibitors, BCR-ABL inhibitors, IDH inhibitors, KIT/PDGFR inhibitors, antimetabolites, and S1P modulators. Adding two to three new compounds in each of these classes would lift them about the $n \geq 5$ threshold at which mechanism-class generalization estimates become stable.

The second axis is expanding non-oncology mechanism breadth. Many therapeutically important small molecule classes are absent or thinly represented in the current cohort, including anticoagulants (Factor Xa inhibitors, direct thrombin inhibitors), antiarrhythmics (Na+ and K+ channel modulators), antimicrobials (β-lactams, fluoroquinolones, antifungal azoles, antiretroviral integrase inhibitors), psychotropics beyond the antipsychotic class currently included, analgesics, and metabolic agents beyond the SGLT2 and DPP-2 classes.

The third axis is temporal expansion to post-2024 new molecular entities. The openFDA prescribing label corpus updates continuously, and approximately 20-30 new small molecule drug approvals enter the corpus annually in recent years. An expanded drug cohort assembled with the same inclusion criteria would double the cohort to approximately $N = 170$ compounds. With cohort-expansion-driven discrimination scaling, this N would be expected to deliver corrected $\bar{\rho}$ in the +0.12 to +0.15 range. A cohort assembled to deliberately span the under-covered mechanism classes named above, rather than relying on natural FDA approval flow, could plausibly approach $N = 200$ within a defined assembly horizon and the +0.15-0.20 projection that **3.3.7** supports.

4.4.2 Architectural extensions: STAR-specific inductive biases

The multi-task neural network of Chapter 3 was deliberately architecturally minimal; chosen to isolate the contribution of the STAR feature representation from any architectural inductive bias the result might inherit. Once the contribution of the feature framework is established, architectural extensions that encode STAR-specific inductive biases constitute the appropriate next direction. Three classes of extension are technically tractable on the present cohort and address specific limitations of the architectural baseline.

The first class is per-AE block-attention layers that allow the model to learn distinct STAR-block weightings for different AE endpoints. Under the current shared-trunk multi-layer perceptron, every AE head shares the same trunk representation, and the per-head linear weights determine how that representation maps to per-AE predicted incidence. A more pharmacologically faithful architecture would produce per-block intermediate representations (one each for S, PS, EX, TS, Cp) and learn per-AE attention weights over these representations. This would allow, for example, hepatic AE heads to weigh PS and EX heavily, reflecting the role of hepatic-expressed off-target engagement in hepatic toxicities, and dermatologic AE heads to weight S and PS more heavily, reflecting the role of direct structural pharmacology in skin-localized toxicity for many drug classes. Block-attention layers would additionally yield per-prediction interpretability of the form, "for sunitinib's predicted hypertension, 60% of the prediction was driven by the PS block, with VEGFR2 the dominant contributing target." This is the form of attribution that renders the framework useful for candidate design feedback.

The second class is explicit cross-block interaction modules that operationalize the PS × EX coupling that is currently latent in the TS block. The current framework uses a binding-strength-weighted expression score following Cao et al.[17], and the block-level ablation analysis demonstrated it to be approximately redundant with PS × EX combinations that the shared trunk learns implicitly. A learned cross-block interaction module, for instance a bilinear PS-by-EX interaction layer or a self-attention layer over PS-EX pairs, would replace the current TS construction with a learned interaction representation and would naturally extend to non-kinase target panels where the Cao et al. TS formula does not directly apply. The same module class supports cross-block interactions with Cp (a dose-by-target-engagement interaction that captures whether predicted clinical exposure is sufficient to engage each target), addressing the observation that the current framework treats Cp as a flat scalar feature rather than as an exposure-context modulator on PS.

The third class is uncertainty quantification at the per-prediction level rather than at the cohort-level bootstrap. The current framework reports cohort-mean corrected $\bar{\rho}$ with 500-seed bootstrap confidence intervals, and per-drug predicted incidence with per-drug seed standard deviations from the bootstrap. For clinical decision support, however, per-drug-AE confidence intervals, encoding the model’s confidence in each individual predicted incidence, are the operative uncertainty quantification. Deep ensembles, Bayesian last-layer methods, and Monte

Carlo dropout each provide per-prediction uncertainty estimates at modest training time cost[39-41]. Per-prediction uncertainty is of particular value in the STAR framework's intended pre-FIH triage use case, where a decision committee can respond differently to a low-incidence prediction made with high confidence than to one accompanied by wide uncertainty bounds.

A fourth direction, deferred to longer-horizon work, is integration with multi-modal molecular embeddings. The current framework's structure block uses Morgan circular fingerprints, a fixed representation that captures local chemical structure but does not encode learned chemical language or geometric representations. Recent multi-modal molecular foundation models such as ChemBERTa, MolFormer, MolE, and UniMol[42-45], produce learned chemical embeddings that have demonstrated improvements over Morgan fingerprints across diverse property prediction benchmarks. Concatenating learned chemical embeddings with the STAR mechanism-aware features, or training a STAR-aware adapter on a pretrained molecular foundation model, would provide an orthogonal axis along which to improve the framework's discrimination beyond the per-block ablation gains that single block additions can deliver.

4.5 Conclusions

As shown through two implementations of the STAR framework, integrating mechanism-aware pharmacological context into AI/ML toxicity prediction recovers signal that target binding or chemical structure alone cannot. The random survival forest established in Chapter 2 demonstrated patient-level temporal risk prediction within a single mechanism class. The multi-task neural network constructed in Chapter 3 demonstrated compound-level AE incidence prediction across an expanded cohort spanning oncology and non-oncology mechanism classes. The two implementations differ in their cohort scope, prediction target, and modeling architecture, but converge on the methodological claim that a feature representation jointly encoding structure, target engagement, tissue distribution, and exposure recovers drug-specific safety signals that prior representations have not adequately captured.

The framework's principal contribution is methodological in nature rather than an advance in predictive performance. The cohort-level magnitudes are bounded both by the pharmacological information captured in the present STAR feature set and by the inherent noise of prescribing-label AE incidence as a quantitative training target. Within those limits, the framework offers a reusable feature taxonomy, a base-rate-robust evaluation protocol, and a

transparently reported empirical evidence base on which future cross-class AE prediction work can build.

The translational direction toward which STAR points is an *in silico* pre-FIH safety prediction tool integrating mechanism-aware features, discrimination conditioned on pharmacological context, and uncertainty quantification aligned with regulatory expectations into the candidate selection and Phase 1 protocol design workflows of MIDD. At a moment when drug development attrition has made the translational gap acute, the STAR framework provides both an empirical demonstration that pharmacological context unlocks signal beyond what structure or target binding alone can recover, and a methodological foundation on which the next generation of NAMs can build.

4.6 References

1. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B*. Jul 2022;12(7):3049–3062. doi:10.1016/j.apsb.2022.02.002
2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol*. Jan 2014;32(1):40–51. doi:10.1038/nbt.2786
3. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. Apr 1 2019;20(2):273–286. doi:10.1093/biostatistics/kxx069
4. Tamaki C, Nagayama T, Hashiba M, et al. Potentials and limitations of nonclinical safety assessment for predicting clinical adverse drug reactions: correlation analysis of 142 approved drugs in Japan. *J Toxicol Sci*. 2013;38(4):581–98. doi:10.2131/jts.38.581
5. Maciejewski M, Lounkine E, Whitebread S, et al. Reverse translation of adverse event reports paves the way for de-risking preclinical off-targets. *Elife*. Aug 8 2017;6doi:10.7554/eLife.25818
6. Sun D, Macedonia C, Chen Z, et al. Can Machine Learning Overcome the 95% Failure Rate and Reality that Only 30% of Approved Cancer Drugs Meaningfully Extend Patient Survival? *J Med Chem*. Sep 26 2024;67(18):16035–16055. doi:10.1021/acs.jmedchem.4c01684
7. Joshi SS, Ortiz S, Witherspoon JN, et al. Effects of epidermal growth factor receptor inhibitor-induced dermatologic toxicities on quality of life. *Cancer*. Aug 15 2010;116(16):3916–23. doi:10.1002/cncr.25090
8. Liu J, Yan S, Du J, et al. Mechanism and treatment of diarrhea associated with tyrosine kinase inhibitors. *Heliyon*. Mar 30 2024;10(6):e27531. doi:10.1016/j.heliyon.2024.e27531
9. Robinson ES, Khankin EV, Choueiri TK, et al. Suppression of the nitric oxide pathway in metastatic renal cell carcinoma patients receiving vascular endothelial growth factor-signaling inhibitors. *Hypertension*. Dec 2010;56(6):1131–6. doi:10.1161/HYPERTENSIONAHA.110.160481
10. Takahashi S. Fatigue and its management in cancer patients undergoing VEGFR-TKI therapy. *Expert Opin Drug Saf*. Mar 2022;21(3):397–406. doi:10.1080/14740338.2021.1969360
11. Davis MI, Hunt JP, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol*. Oct 30 2011;29(11):1046–51. doi:10.1038/nbt.1990

12. Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J Chem Inf Model*. Mar 24 2014;54(3):735–43. doi:10.1021/ci400709d
13. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. Jan 4 2016;44(D1):D1045–53. doi:10.1093/nar/gkv1072
14. Tang J, Tanoli ZU, Ravikumar B, et al. Drug Target Commons: A Community Effort to Build a Consensus Knowledge Base for Drug-Target Interactions. *Cell Chem Biol*. Feb 15 2018;25(2):224–229 e2. doi:10.1016/j.chembiol.2017.11.009
15. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. Jan 23 2015;347(6220):1260419. doi:10.1126/science.1260419
16. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol*. Jun 2020;38(6):675–678. doi:10.1038/s41587-020-0546-8
17. Cao A, Zhang L, Bu Y, Sun D. Machine Learning Prediction of On/Off Target-driven Clinical Adverse Events. *Pharm Res*. Aug 2024;41(8):1649–1658. doi:10.1007/s11095-024-03742-x
18. Kass-Hout TA, Xu Z, Mohebbi M, et al. OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. *J Am Med Inform Assoc*. May 2016;23(3):596–600. doi:10.1093/jamia/ocv153
19. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279–90.
20. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ*. May 2016;47:20–33. doi:10.1016/j.jhealeco.2016.01.012
21. S7A Safety Pharmacology Studies for Human Pharmaceuticals (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use) (2001).
22. The Non-Clinical Evaluation of the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation) by Human Pharmaceuticals S7B (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use) (2005).
23. Valentin JP, Leishman D. 2000-2023 over two decades of ICH S7A: has the time come for a revamp? *Regul Toxicol Pharmacol*. Mar 2023;139:105368. doi:10.1016/j.yrtph.2023.105368
24. Bowes J, Brown AJ, Hamon J, et al. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat Rev Drug Discov*. Dec 2012;11(12):909–22. doi:10.1038/nrd3845
25. Brennan RJ, Jenkinson S, Brown A, et al. The state of the art in secondary pharmacology and its impact on the safety of new medicines. *Nat Rev Drug Discov*. Jul 2024;23(7):525–545. doi:10.1038/s41573-024-00942-3
26. Papoian T, Chiu HJ, Elayan I, et al. Secondary pharmacology data to assess potential off-target activity of new drugs: a regulatory perspective. *Nat Rev Drug Discov*. Apr 2015;14(4):294. doi:10.1038/nrd3845-c1
27. Wang Y, Zhu H, Madabushi R, Liu Q, Huang SM, Zineh I. Model-Informed Drug Development: Current US Regulatory Practice and Future Considerations. *Clin Pharmacol Ther*. Apr 2019;105(4):899–911. doi:10.1002/cpt.1363

28. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017; Long Beach, California, USA.
29. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: Generating Explanations for Graph Neural Networks. Curran Associates, Inc.; 2019:9240–9251.
30. Dickinson Q, Meyer JG. Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences. *PLoS Comput Biol*. Jan 2022;18(1):e1009736. doi:10.1371/journal.pcbi.1009736
31. Han JJ. FDA Modernization Act 2.0 allows for alternatives to animal testing. *Artif Organs*. Mar 2023;47(3):449–450. doi:10.1111/aor.14503
32. Roadmap to Reducing Animal Testing in Preclinical Safety Studies. 2025. https://www.fda.gov/files/newsroom/published/roadmap_to_reducing_animal_testing_in_preclinical_safety_studies.pdf
33. Qualification Process for Drug Development Tools Guidance for Industry and FDA Staff (U.S. Food and Drug Administration) (2020).
34. Galeano D, Paccanaro A. Machine learning prediction of side effects for drugs in clinical trials. *Cell Rep Methods*. Dec 19 2022;2(12):100358. doi:10.1016/j.crmeth.2022.100358
35. Sharma B, Chenthamarakshan V, Dhurandhar A, et al. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. *Sci Rep*. Mar 25 2023;13(1):4908. doi:10.1038/s41598-023-31169-8
36. Halip L, Avram S, Curpan R, et al. Exploring DrugCentral: from molecular structures to clinical effects. *J Comput Aided Mol Des*. Dec 2023;37(12):681–694. doi:10.1007/s10822-023-00529-x
37. Heinzke AL, Zdrazil B, Leeson PD, et al. A compound-target pairs dataset: differences between drugs, clinical candidates and other bioactive compounds. *Sci Data*. Oct 21 2024;11(1):1160. doi:10.1038/s41597-024-03582-9
38. Guidance for Industry: Adverse Reactions Section of Labeling for Human Prescription Drug and Biological Products — Content and Format (U.S. Food and Drug Administration) (2006).
39. Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. 2017; https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf
40. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. presented at: Proceedings of The 33rd International Conference on Machine Learning; 2016; Proceedings of Machine Learning Research. https://proceedings.mlr.press/v48/gal16.html
41. Wilson AG, Izmailov P. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. 2020; https://proceedings.neurips.cc/paper_files/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf
42. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv*. 2020;arXiv:2010.09885
43. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*. 2022/12/01 2022;4(12):1256–1264. doi:10.1038/s42256-022-00580-7

44. Mendez-Lucio O, Nicolaou CA, Earnshaw B. MolE: a foundation model for molecular graphs using disentangled attention. *Nat Commun*. Nov 12 2024;15(1):9431. doi:10.1038/s41467-024-53751-y
45. Zhou G, Gao Z, Ding Q, et al. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. 2023:

Chapter 5: Direct Measurement of *In Vivo* Dissolution and Release of Modified-Release Glipizide Formulations

5.1 Abstract

Oral modified-release (MR) products that meet plasma-based bioequivalence criteria can release drug differently along the gastrointestinal tract, yet conventional pharmacokinetic and compendial dissolution testing characterize this *in vivo* release only indirectly. We developed an integrated *in vivo* biopharmaceutic methodology for oral MR products that combines region-resolved multi-lumen gastrointestinal intubation, within-subject co-administration of a $^{13}C_6$-glipizide stable-isotope oral solution as a same-occasion systemic disposition reference, and population pharmacokinetics-supported Loo-Riegelman deconvolution. The methodology was implemented in healthy adult volunteers using two commercially available MR glipizide products of contrasting release mechanism: an osmotic pump and a hydrophilic polymer matrix. The primary analysis population comprised 16 volunteers (7 osmotic pump, 9 matrix) who received a single 5 mg MR dose with a 1 mg $^{13}C_6$-glipizide oral solution reference, with luminal fluid sampled from the stomach, duodenum, and jejunum over a seven-hour window. Plasma concentration profiles of the two products were closely similar, consistent with their bioequivalence status, whereas their regional luminal release patterns differed substantially. Drug from the osmotic pump was confined predominantly to the stomach, whereas drug from the matrix product appeared across the stomach, duodenum, and jejunum, with the largest individual concentrations in the small bowel. The directly characterized luminal environment showed low gastric pH, near-neutral small bowel pH, and small bowel bile acid concentrations above gastric values but below the FaSSIF-V2 specification. Loo-Riegelman deconvolution indicated that approximately one-quarter of the administered MR dose had been absorbed by the end of the seven-hour intubation window. By rendering regional *in vivo* release and the local luminal environment directly observable for products that plasma-based bioequivalence testing cannot distinguish, the methodology supports mechanism-aware biopharmaceutic evaluation of oral

modified-release formulations and can inform patient-centric quality standards and the assessment of modified-release generics.

5.2 Introduction

Type 2 diabetes mellitus is a chronic metabolic disorder characterized by insulin resistance and progressive pancreatic beta-cell dysfunction and remains among the most prevalent non-communicable diseases worldwide[1]. Pharmacological management typically begins with lifestyle modification and metformin, with additional agents introduced as glycemic targets demand. Sulfonylureas, of which glipizide is a second-generation representative, remain widely prescribed because they achieve clinically meaningful reductions in hemoglobin A1c at low cost and are available in both immediate-release and modified-release oral dosage forms[2]. Glipizide acts by binding to the sulfonylurea receptor SUR1 on the pancreatic beta-cell membrane, closing ATP-sensitive potassium channels, depolarizing the cell, and triggering calcium-dependent insulin secretion[3]. The clinical profile of glipizide is dominated by two pharmacokinetically driven concerns. The first is iatrogenic hypoglycemia, which follows directly from the insulinotropic mechanism and is exacerbated when plasma concentrations rise steeply or remain elevated in the fasting state. The second is substantial between-subject variability in systemic exposure following oral administration, which complicates dose individualization and contributes to inconsistent glycemic response among patients receiving the same nominal regimen[4-6].

Modified-release (MR) oral dosage forms have been developed for glipizide to address these concerns. By extending the interval over which drug is released into the gastrointestinal tract, an MR formulation reduces the peak-to-trough oscillation in plasma concentration, attenuates the sharpness of the postprandial insulinotropic stimulus, and permits once-daily administration in place of the two-to-three daily doses typically required for the immediate-release product[7-9]. Several MR glipizide products are currently approved and commercially available in the United States. The two products examined in this study represent distinct modified-release mechanisms: Product 1, an osmotic pump product, and Product 2, a generic hydrophilic polymer matrix product. Both products are FDA-approved for once-daily administration, with the generic matrix product having demonstrated bioequivalence to the reference listed drug under conventional plasma-based criteria[10]; yet the two products are constructed around mechanistically distinct release principles, and whether this mechanistic

difference translates into meaningful differences in *in vivo* regional release and gastrointestinal disposition has not been directly addressed in published human studies of these specific products.

Osmotic pump tablets are bilayer devices comprising an active drug layer and an osmotically active push layer, the two layers being enclosed in a semipermeable polymeric membrane into which one or more delivery orifices have been drilled by laser[11]. After oral administration, water is drawn into the tablet core across the semipermeable membrane under an osmotic gradient, the push layer expands, and a hydrated drug suspension is extruded through the delivery orifice at a rate that is governed primarily by the membrane's water permeability and the osmotic driving force rather than by the hydrodynamics of the surrounding medium. Consequently, osmotic pump formulations are widely regarded as exhibiting release profiles that are relatively insensitive to changes in gastrointestinal pH, hydrodynamic agitation, and regional transit, and are commonly cited as approaching the ideal of a zero-order oral delivery device.[12] A direct corollary, and one that motivates the inclusion of an osmotic pump arm in the present study, is that *in vivo* release from such a product would be expected to localize predominantly to the region of the gastrointestinal tract through which the intact device is transiting, with comparatively little dependence on regional pH or bile acid environment.

Hydrophilic polymer matrix tablets operate on a fundamentally different principle. The tablet core contains the drug dispersed in a polymeric matrix, most commonly a high-molecular-weight cellulose derivative such as hydroxypropyl methylcellulose, which upon contact with aqueous gastrointestinal fluid hydrates to form a gel layer through which drug is released by a combination of diffusion through the hydrated matrix and erosion of the matrix surface[13]. The rate and cumulative extent of release from such systems depend on the hydration kinetics of the polymer, the solubility and diffusivity of the active ingredient in the hydrated gel layer, and the mechanical integrity of the matrix under the shear and deformation conditions encountered along the gastrointestinal tract. Matrix systems are therefore more susceptible in principle to variations in regional fluid volume, luminal pH, bile acid concentration, and mechanical transit than are osmotic pumps, and their *in vivo* release profile may differ appreciably from that predicted by compendial dissolution testing conducted under idealized conditions. Product 2 is a hydrophilic matrix tablet and represents, in the present study, the matrix arm of a deliberate mechanism-level comparison with the osmotic pump arm. A direct corollary, and one that motivates the inclusion of a matrix arm as the mechanistic counterpart to the osmotic pump arm, is that *in vivo* release

from such a product would be expected to vary with the regional gastrointestinal environment encountered during transit, and may therefore yield a regional release pattern that is spatially and temporally distinct from that of an osmotic pump device administered under the same conditions.

The biopharmaceutic characterization of an oral MR product traditionally rests on indirect inference. *In vitro* dissolution testing provides a measurement of cumulative drug release into a single well-stirred medium under fixed conditions, and plasma pharmacokinetics provides a measurement of drug appearance in the systemic circulation; the *in vivo* release process itself, occurring in the gastrointestinal lumen, is inferred from these surrogate observations rather than measured directly. The quantitative relationship between *in vitro* dissolution performance and *in vivo* pharmacokinetic behavior is formalized by the regulatory construct of the *in vitro-in vivo* correlation (IVIVC). A Level A IVIVC, in which the *in vitro* cumulative release profile is predictive of *in vivo* cumulative absorption on a point-to-point basis, is the most stringent correlation class recognized by the Food and Drug Administration and, when established, can support the use of dissolution testing as a surrogate for bioequivalence in the context of post-approval formulation or manufacturing changes[14-18]. The successful establishment of a Level A IVIVC, however, rests on the assumption that *in vitro* dissolution captures the rate-limiting step of *in vivo* drug appearance in systemic circulation. For osmotic pump products whose release is mechanism-limited rather than environment-limited, this assumption is usually defensible; for matrix products whose release is sensitive to the gastrointestinal environment, the assumption may be challenged by regional variations in fluid volume, pH, and luminal content that are poorly reproduced by any single *in vitro* apparatus[19,20].

Complementary to the IVIVC paradigm, the Patient-Centric Quality Standards (PCQS) framework links critical product quality attributes to clinical performance through mechanistic understanding of the *in vivo* release, dissolution, and absorption process, with physiologically based biopharmaceutics modeling (PBBM) increasingly used as the integrative platform[21,22]. Inputs that support such a framework include regional luminal drug concentrations, regional fluid volume, luminal pH, and bile acid composition, each of which can modulate the release and dissolution of an oral MR product in ways that single vessel compendial testing cannot reproduce. These inputs are not recoverable from plasma pharmacokinetic data alone; their measurement requires direct sampling from the relevant gastrointestinal regions.

Direct *in vivo* sampling of gastrointestinal luminal fluid in healthy human volunteers, by means of indwelling multi-lumen intubation catheters, has been developed over the past several decades as a means of quantifying the regional environment encountered by an oral dosage form during transit[23-25]. Intubation studies conducted in our laboratory and by other groups have quantified region-specific fluid volume, pH, bile acid composition, and drug concentration across the stomach, duodenum, and jejunum following administration of a range of oral test products, and have demonstrated that these parameters can differ substantially between regions and between fed and fasted states[26-28]. Applied to an MR product, such sampling yields a direct measurement of the drug concentration delivered to each region during the absorption window, a directly observed, rather than inferred, characterization of *in vivo* release. Co-administration of a stable-isotope-labeled oral solution on the same occasion as the MR dose provides each subject with their own systemic disposition reference, enabling deconvolution of the *in vivo* absorption profile against an individually characterized disposition function rather than against a population average.

Despite the regulatory centrality of MR glipizide products and their widespread clinical use, methodological approaches that integrate region-resolved luminal sampling with within-subject systemic disposition characterization have not been routinely deployed alongside conventional pharmacokinetic assessment of oral MR products in healthy human volunteers. The present study develops and demonstrates such an integrated methodology, combining oral multi-lumen gastrointestinal intubation for region-resolved sampling of luminal fluid from the stomach, duodenum, and jejunum with within-subject co-administration of a stable-isotope ($^{13}C_6$-glipizide) oral solution to provide an individual-level systemic disposition reference. Two commercially available MR glipizide products of contrasting release mechanism, an osmotic pump product and a hydrophilic polymer matrix product, were used as test products to exercise the methodology across mechanistically distinct release types. The study has three aims. The first is to demonstrate that the integrated methodology resolves region-specific *in vivo* release behavior together with the local luminal environment (pH, bile acid composition) into which release occurs. The second is to cross-validate the regional luminal observations against an independent, plasma-based reconstruction of the *in vivo* absorption process, obtained by population-pharmacokinetic-supported deconvolution of each subject's MR profile against their own same-occasion isotope disposition reference. The third is to consider the implications of the

methodology for mechanism-aware biopharmaceutic evaluation of oral MR products, and its potential role within regulatory frameworks, such as PCQS, that link product quality attributes to *in vivo* performance through directly observed regional behavior.

5.3 Methods

5.3.1 Study design overview

The study employed an open-label, two-phase, parallel-arm design in healthy adult volunteers to characterize the *in vivo* dissolution, regional gastrointestinal disposition, and systemic pharmacokinetics of modified-release (MR) glipizide formulations. Volunteers received a single 5 mg oral dose of one of two commercially available MR glipizide products: Product 1 (osmotic pump) or Product 2 (hydrophilic polymer matrix) co-administered with a 1 mg oral solution of $^{13}C_6$-glipizide as a within-subject pharmacokinetic reference. Formulation assignment was made at the investigator's discretion to achieve approximately balanced enrollment between arms.

In Phase 1, volunteers underwent gastrointestinal (GI) intubation to permit direct sampling of luminal fluid from the stomach, duodenum, and jejunum over a seven-hour period following dose administration. Serial plasma, urine, and fecal samples were collected over 78 hours. Continuous glucose monitoring (CGM) sensors were applied prior to dosing and remained in place for at least 24 hours to detect and manage hypoglycemia, a known adverse effect of glipizide in fasting subjects.[29] Rifaximin 200 mg was administered concurrently with the MR tablet as a non-absorbed luminal marker, enabling estimation of total GI fluid volume from rifaximin concentrations measured in aspirated samples[25,30].

Phase 2 was a subsequent, matched non-intubation period in which the same 5 mg MR dose and 1 mg $^{13}C_6$-glipizide reference were administered under identical fasting, fluid-load, and sampling conditions. In place of GI intubation, volunteers ingested a SmartPill Motility Testing System capsule (Medtronic Limited, Watford, United Kingdom), an ingestible wireless device that records regional pH, pressure, and temperature during transit. The inclusion of the SmartPill permitted assessment of GI motility in the absence of the intubation catheter. Phase 2 enrollment was planned to follow Phase 1 after a five to ten day washout period; however, the SmartPill device was discontinued by the manufacturer during the study, terminating further Phase 2 enrollment. Three volunteers (G1, G4, and G5) completed both phases. The paired Phase 1

versus Phase 2 data from these three subjects provide the only within-subject assessment of the pharmacokinetic impact of the intubation procedure and are analyzed separately (**5.3.7**; **Figure 5-23**).

All study drugs were administered with 240 mL of glucose solution delivered in two steps: an initial 40 mL of 14% glucose solution containing the $^{13}C_6$-glipizide reference, followed by a 200 mL 20% glucose solution administered together with the MR glipizide tablet and rifaximin. Glucose co-administration served two purposes: to provide a dissolution medium for the isotope reference, and to supply a fixed caloric load intended to partially attenuate the hypoglycemic effect of glipizide in fasting subjects by controlling gastric emptying rate.

The study was approved by the University of Michigan Medical School Institutional Review Board, conducted in accordance with the Declaration of Helsinki[31], ICH Good Clinical Practice guidelines[32], and registered on ClinicalTrials.gov (NCT05159427). Written informed consent was obtained from all volunteers prior to any study procedures. The primary analysis population consisted of the healthy volunteers who completed Phase 1 drug administration and provided evaluable plasma pharmacokinetic data; further details of subject disposition are presented in **5.3.1**.

5.3.2 Subjects, enrollment, and primary analysis population

Healthy adult volunteers were recruited through public advertising on umhealthresearch.org and by flyer distribution. Eligibility was based on inclusion and exclusion criteria developed in consultation with the study physicians and approved by the Institutional Review Board. Inclusion criteria required age between 18 and 55 years, body mass index (BMI) between 18.5 and 35.0 kg/m^2, the ability to independently provide informed consent, and a negative serum pregnancy test for women of child-bearing potential. Exclusion criteria encompassed 36 specified conditions that fell into four principal categories: clinical conditions that would confound the pharmacokinetic or pharmacodynamic assessment (including type 1 diabetes mellitus, diet-controlled type 2 diabetes, glucose-6-phosphate dehydrogenase deficiency, and gastrointestinal disorders affecting transit, motility, or absorption); hypersensitivity to any of the study drugs or related compounds (including sulfonamides, lidocaine, rifaximin, and adhesive components used in the intubation procedure); concomitant medications or substances expected to interact with glipizide or affect gastric physiology; and

factors compromising study feasibility (including inability to fast for 19 hours, pregnancy, and recent clinical trial participation).

Between September 30, 2020 and March 29, 2026, a total of 37 individuals completed informed consent and the screening visit. Of these, 3 were screen failures and 34 met the inclusion and exclusion criteria, of whom 30 were approved for further enrollment by the study physicians. A total of 16 volunteers successfully completed Phase 1 drug administration and provided evaluable plasma pharmacokinetic data, constituting the primary analysis population. Three additional volunteers (G1, G4, and G5) subsequently completed Phase 2 procedures, permitting within-subject comparison of the intubation phase with a non-intubation phase in these paired subjects.

Written informed consent was obtained electronically through a HIPAA- and 21 CFR Part 11-compliant signature platform prior to any study procedures, in accordance with the Declaration of Helsinki and ICH Good Clinical Practice guidelines. Each volunteer's understanding of the consent material was confirmed through a nine-item questionnaire administered by the study coordinator or investigator prior to signing. Demographic and baseline characteristics of the primary analysis population are summarized in **Table 5-1**.

5.3.3 Study products and dosing

Volunteers were administered a single 5 mg oral dose of one of two commercially available modified-release (MR) glipizide products: Product 1 (osmotic pump) or Product 2 (hydrophilic polymer matric). Product 1 is an osmotic pump formulation, which delivers drug through a laser-drilled orifice at a rate driven by the osmotic uptake of water into an expanding push-layer within the tablet core. Product 2 is a hydrophilic polymer matrix formulation, in which drug release is governed by hydration, swelling, and erosion of the polymeric tablet matrix in contact with gastrointestinal fluid. These two release mechanisms represent the principal technologies used in commercially available MR glipizide products and were selected to permit cross-mechanism comparison between the osmotic pump product and the matrix product. The administered 5 mg dose corresponds to the recommended therapeutic dose in patients with type 2 diabetes mellitus and is lower than the 10 mg dose specified for healthy volunteers in the FDA product-specific guidance for extended-release glipizide tablets[10].

Concurrently with each MR dose, volunteers received a 1 mg oral solution of $^{13}C_6$-glipizide as a within-subject pharmacokinetic reference. The isotope solution was prepared by

dissolving 1 mg of $^{13}C_6$-glipizide in 40 mL of 14% glucose solution (United States Pharmacopeia grade sterile water and dextrose), with pH adjustment to facilitate dissolution of glipizide as a weak acid[33]. Because the isotope is delivered as a solution, it undergoes immediate systemic absorption without requiring tablet disintegration or dissolution, permitting direct estimation of individual disposition parameters for each volunteer. These individually estimated disposition parameters serve as the reference against which the *in vivo* release profile of the MR product is subsequently characterized by deconvolution (**5.2.9**). Differentiation of $^{13}C_6$-glipizide from unlabeled glipizide during bioanalysis was achieved by liquid chromatography with tandem mass spectrometry, which resolves the six-dalton mass difference between the two forms.

All study drugs were administered with 240 mL of glucose solution delivered in two sequential steps: the 40 mL of 14% glucose solution containing the isotope reference was administered first, followed by 200 mL of 20% glucose solution administered together with the MR glipizide tablet and a 200 mg rifaximin tablet.

5.3.4 Gastrointestinal intubation and sample collection

Regional luminal sampling in Phase 1 was performed using a multi-lumen gastrointestinal catheter with four aspiration ports (Mui Scientific, Mississauga, Ontario, Canada), with target port locations in the stomach, duodenum, and jejunum. The catheter was inserted orally by a trained gastrointestinal physiology laboratory technician under the supervision of the study physician. A topical 4% lidocaine solution was applied to the oropharynx, and the catheter was lubricated with lidocaine jelly, to suppress the gag reflex and facilitate passage. Correct placement of the catheter, including confirmation that the distal end had advanced past the pylorus, was verified by abdominal fluoroscopy performed by a trained technician during the insertion procedure. Once correct position had been confirmed, the catheter was secured externally with medical tape or a Velcro tube holder. The full placement procedure generally required approximately four hours to complete, allowing for adjustment of catheter position and for the volunteer to become tolerant of the catheter. Prior to study drug administration, the study team confirmed that the topical anesthetic had sufficiently dissipated to permit safe swallowing of the dose, typically at approximately one hour after lidocaine application.

Gastrointestinal fluid samples were aspirated from each available catheter port at nominal times of 15, 30, and 45 minutes and at 1, 1.5, 2, 2.5, 3, 4, 5, 6, and 7 hours following

administration of the modified-release dose, with a target aspirate volume of approximately 1 mL per port per timepoint. A pre-dose aspirate was collected where feasible to support analytical baseline characterization. Because luminal fluid availability varied with regional content and catheter position, samples were collected as available, and the actual volume recovered at each timepoint was recorded; some port-timepoint combinations yielded less than the target volume or no recoverable fluid. The pH of each aspirate was measured *ex vivo* using a calibrated benchtop pH meter immediately after collection, and samples were subsequently aliquoted and stored at -80 °C pending bioanalysis. The catheter was removed by the gastrointestinal physiology technician at the seven-hour timepoint by gentle, gradual withdrawal, after which the volunteer was provided a standardized meal. Actual collection times were recorded for all samples and were used in place of nominal times in all subsequent pharmacokinetic analyses.

Plasma samples for pharmacokinetic analysis were collected over a total period of 78 hours following the modified-release dose. An indwelling intravenous catheter was placed prior to dosing and maintained with a saline lock, and a pre-dose blood sample (T0) was drawn immediately before study drug administration. Following dosing, serial blood samples were collected through the indwelling catheter at nominal times of 10, 20, 30, and 45 minutes and at 1, 1.5, 2, 2.5, 3, 4, 5, 6, 7, 8, and 10 hours. The indwelling catheter was removed after the 10-hour collection, and the volunteer was discharged with instructions for outpatient return visits at nominal times of 18, 30, 54, and 78 hours post-dose, at which additional samples were obtained by peripheral venipuncture. All blood samples were collected into 6 mL sodium heparin tubes, from which approximately 1 to 3 mL of plasma was separated by centrifugation and stored in aliquots at −80 °C pending bioanalysis. Sample collection was targeted to within ±10 minutes of the nominal time during the dosing visit and to within ±2 hours of the nominal time at the outpatient return visits; actual collection times were recorded and were used in place of nominal times in all subsequent pharmacokinetic analyses.

Urine and feces samples were collected in sterile containers throughout the 78-hour post-dose period to permit characterization of mass balance and to support quantification of excreted parent drug and metabolites. Pre-dose samples were collected where feasible to provide analytical baseline matrices. During the dosing visit, all urine and feces voided by the volunteer were collected at the study site; for the period between discharge and each subsequent outpatient return visit, volunteers were provided with sterile collection containers and written instructions

for home collection, with returned samples logged on arrival at each visit. All urine and fecal samples were stored at -80 °C pending bioanalysis. The plasma, urine, and fecal sample collection schedules described above were applied identically in the non-intubation Phase 2 visits completed by the three paired volunteers (G1, G4, and G5).

5.3.5 Bioanalytical methods

Quantification of glipizide and $^{13}C_6$-glipizide was performed by ultra-performance liquid chromatography with tandem mass spectrometry (UPLC-MS/MS) on a Waters ACQUITY UPLC system coupled to a Waters Xevo TQD triple-quadrupole mass spectrometer operated in positive electrospray ionization mode with multiple reaction monitoring (Waters, Milford, Massachusetts, United States). Chromatographic separation was achieved on a Waters C18 column (2.1 × 50 mm, 1.7 μm) using a gradient of 0.1% formic acid in water and 0.1% formic acid in acetonitrile at 0.6 mL/min. The MRM transitions monitored were *m/z* 446.21 $[M+H]^+ \rightarrow$ 103.03 for glipizide and *m/z* 452.06 $[M+H]^+ \rightarrow$ 92.97 for $^{13}C_6$-glipizide, with glipizide-d_{11} as the deuterated internal standard (Cayman Chemical, Ann Arbor, Michigan, United States). Calibration standards were prepared in the matrix of interest across 11 concentration levels spanning 0.5 to 1,000 ng/mL, and standard curves were fitted by weighted ($1/x^2$) least-squares regression of the analyte-to-internal-standard peak area ratio against nominal concentration. The method was validated in each matrix for linearity, selectivity, accuracy, precision, recovery, matrix effects, carryover, dilution integrity, and analyte stability, in accordance with United States Food and Drug Administration guidance on bioanalytical method validation[34].

The core UPLC-MS/MS conditions described above were applied to all four biological matrices with matrix-specific sample preparation steps. Plasma samples (100 μL) were subjected to direct methanol protein precipitation with internal standard, followed by centrifugation and injection of the supernatant. Gastrointestinal fluid samples (10 μL) were first diluted 10-fold into blank human plasma to reproduce the matrix of the calibration standards and were then processed following the plasma protein precipitation procedure. Urine samples (100 μL) were processed directly by methanol protein precipitation against urine-matched calibration standards. Fecal samples were weighed and homogenized in three volumes of methanol, and an aliquot of the resulting supernatant was further processed by methanol protein precipitation against fecal-matrix-matched calibration standards. Across all four matrices, the lower limit of quantification for glipizide and $^{13}C_6$-glipizide was 0.5 ng/mL; observations below this concentration were

classified as below the limit of quantification (BLQ) and are identified as such in subsequent tabulations and analyses. The treatment of BLQ observations in pharmacokinetic analyses is described in **5.2.8**.

A panel of 15 bile acids was quantified in gastrointestinal fluid samples using a separate UPLC-MS/MS method on the same Waters ACQUITY UPLC platform coupled to a Waters TQD tandem-quadrupole mass spectrometer operated in negative electrospray ionization mode with multiple reaction monitoring. The panel comprised the five principal endogenous bile acids (cholic acid, chenodeoxycholic acid, deoxycholic acid, lithocholic acid, and ursodeoxycholic acid) together with their glycine- and taurine-conjugated derivatives, yielding 15 analytes spanning the primary and secondary, conjugated and unconjugated classes. Chromatographic separation was performed on a CORTECS T3 column (2.1 × 30 mm, 2.7 μm) at 40 °C using a gradient of 0.01% formic acid with 0.2 mM ammonium formate in water and in a 50:50 (v/v) isopropanol-acetonitrile co-solvent, at 1.0 mL/min. Quantification employed nine deuterated bile acid internal standards and a 10-point calibration curve ranging from 5 to 250,000 ng/mL fitted by weighted ($1/x^2$) least-squares regression. The lower limit of quantification varied by analyte across the range 5 to 100 ng/mL, and the method was validated for linearity, accuracy, precision, recovery, matrix effects, and stability in accordance with FDA bioanalytical method validation guidance[34].

5.3.6 Continuous glucose monitoring

Because glipizide is a sulfonylurea with a well-characterized capacity to lower blood glucose in fasting subjects, interstitial glucose was monitored continuously during and after the dosing visit beginning with subject G12, following an interim protocol amendment approved by the Institutional Review Board. Continuous monitoring was performed using the FreeStyle Libre 3 system (Abbott Diabetes Care, Alameda, California, United States), an FDA-approved sensor-based device that reports interstitial glucose at 1-minute intervals and transmits data to the clinical study team in real time through the LibreLinkUp application. The sensors were applied to the posterior aspect of the volunteer's upper left and right arms at least one hour prior to study drug administration, in accordance with the manufacturer's warm-up requirement, and remained in place for at least 24 hours following the dose. After discharge from the clinical research unit, interstitial glucose continued to be monitored remotely through LibreLinkUp for approximately a

further 12 hours, with glucose alarms set to alert both the volunteer and the responsible study team member if readings fell below 90 mg/dL[35].

Hypoglycemia was managed in accordance with the University of Michigan Hospital Hypoglycemia Treatment Algorithm for Adult Patients. The approach to glucose management evolved over the course of the study as clinical experience accumulated. Early subjects enrolled under the initial amendment were managed by intermittent intravenous rescue, in which a 25 mL intravenous push of 50% dextrose (D50) was administered whenever continuous glucose measurements declined to or below 90 mg/dL. In subsequent subjects, a prophylactic continuous intravenous infusion of 10% dextrose (D10) at 50-100 mL/h was initiated immediately prior to dose administration and maintained through the end of the dosing visit, with the infusion rate adjusted at the discretion of the study physician; the prophylactic D10 infusion substantially reduced the frequency of D50 rescue boluses required. Following discharge, volunteers were provided with oral glucose tablets (four tablets totaling 15 g of carbohydrate) and were instructed to consume them and to contact the study team whenever the continuous glucose monitor alarm indicated a reading below 70 mg/dL. For the purposes of subsequent summary reporting, hypoglycemia was classified using American Diabetes Association definitions, with Level 1 hypoglycemia defined as glucose below 70 mg/dL and Level 2 as glucose below 54 mg/dL[2].

Continuous glucose monitoring data were therefore available for the 11 subjects dosed after the protocol amendment (G12 through G22). Of these, 9 received the hydrophilic matrix product and 2 received the osmotic pump product. This distribution precludes a meaningful between-formulation comparison of postprandial glucose trajectories, and continuous glucose monitoring is accordingly presented in this study as a safety and feasibility measure and as an exploratory pharmacodynamic readout of the pooled cohort, rather than as a formulation-discriminating endpoint.

5.3.7 Non-compartmental analysis

Non-compartmental analysis (NCA) of plasma concentration-time data was performed separately for the modified-release glipizide and the $^{13}C_6$-glipizide oral solution for each subject using the PKNCA package (version 0.12.1) in R (version 4.5.2). The following parameters were estimated for each analyte: maximum observed plasma concentration (C_{max}), time to maximum concentration (T_{max}), area under the plasma concentration-time curve extrapolated to infinity (AUC_{inf}), the apparent terminal elimination rate constant (λ_z), and the terminal elimination half-

life ($t_{1/2}$). The apparent terminal elimination rate constant λ_z was estimated by log-linear regression of the last three or more measurable concentrations in the apparent terminal phase, and $t_{1/2}$ was calculated as $\ln(2)/\lambda_z$.

5.3.8 Population pharmacokinetic modeling

A population pharmacokinetic (PopPK) model was developed for the co-administered $^{13}C_6$-glipizide oral solution to derive individual post-hoc estimates of systemic disposition parameters for subsequent use in the deconvolution of modified-release absorption profiles (**5.2.9**). The isotope-only analysis leverages the within-subject design: because the isotope solution is absorbed directly from an aqueous dosing vehicle without requiring tablet disintegration or dissolution, its plasma concentration-time profile reflects systemic disposition on the same occasion as the co-administered modified-release tablet, and so provides a subject- and occasion-matched reference against which the modified-release absorption profile can be deconvolved.

The analysis dataset comprised all plasma concentrations of $^{13}C_6$-glipizide from all volunteers in the primary analysis population, yielding 296 observation records at actual recorded times post-dose. Below-quantification-limit observations were retained in the analysis and handled using the Beal M5 method, in which BLQ records are substituted with one-half the lower limit of quantification and treated as measured observations; this approach was selected to retain information from the low-concentration terminal phase[36]. Population modeling was performed in NONMEM (version 7.5, ICON plc, Dublin, Ireland) using the first-order conditional estimation method with interaction (FOCE-I). Dataset assembly, exploratory graphics, and post-processing of NONMEM output were performed in R (version 4.5.2) using base R and the tidyverse family of packages; control streams, runs, and output files were managed directly without an intervening run-management layer.

Structural model development followed a standard sequential approach. One- and two-compartment disposition models were evaluated, with absorption parameterized as first-order, zero-order, first-order with absorption lag time, sequential transit-compartment, and parallel first-order schemes. A two-compartment model with first-order absorption was selected as the structural base on the basis of objective function value, condition number, relative standard error on individual parameter estimates, visual inspection of goodness-of-fit plots, and physiological plausibility of parameter estimates. Interindividual variability (IIV) was modeled using

exponential random effects on the structural parameters and was retained where inclusion produced a significant reduction in objective function value (ΔOFV >3.84 for one degree of freedom, $p < 0.05$) and where η-shrinkage remained acceptable. Proportional and combined (proportional plus additive) residual error models were evaluated; a proportional residual error model was selected for the final model.

Adequacy of the final model was evaluated graphically and statistically. Goodness-of-fit was assessed through the standard four-panel diagnostic display of observed versus population-predicted concentrations, observed versus individual-predicted concentrations, conditional weighted residuals versus time, and conditional weighted residuals versus population-predicted concentrations, with a locally estimated scatterplot smoother (LOESS) overlaid to aid visual inspection of systematic trends. The distribution and shrinkage of the empirical Bayes estimates (η) were inspected for each random effect, and the condition number of the covariance matrix was examined as a check on overall identifiability. Simulation-based adequacy was evaluated using a visual predictive check, with 1,000 simulated replicates of the observed design binned by nominal time after dose, and the observed median and 5th and 95th percentiles overlaid on the corresponding simulated medians and 90% prediction intervals. Pooled goodness-of-fit across subjects was additionally summarized by plotting the geometric mean of observed, individual-predicted, and population-predicted concentrations against time on both linear and semi-logarithmic axes. Final population parameter estimates, together with their relative standard errors and the estimated IIV variances and η-shrinkage, are reported in Table 5.3. Individual post-hoc estimates of clearance (CL), central and peripheral volumes of distribution (V_2 and V_3), and intercompartmental clearance (Q) were extracted from the NONMEM phi-file and used to derive subject-specific two-compartment micro-rate constants (k_{10}, k_{12}, k_{21}) for use as the disposition input to the deconvolution analyses described in **5.2.9**.

5.3.9 Deconvolution of modified-release absorption

The *in vivo* absorption profile of each modified-release formulation was characterized on an individual-subject basis using three complementary approaches, each leveraging the subject-specific disposition parameters derived from the population pharmacokinetic model of the co-administered isotope. Because the oral isotope solution and the modified-release tablet were administered to the same subject on the same occasion, systemic disposition of the modified-release drug can be treated as known on a per-subject, per-occasion basis, and the absorption

profile can be separated from disposition without assuming a common population disposition function.

The primary approach was the Loo-Riegelman method for drugs exhibiting two-compartment disposition, which yields a non-parametric estimate of the cumulative fraction of the modified-release dose absorbed at each observed sampling time without requiring an assumed functional form for the absorption process[37,38]. The individual two-compartment micro-rate constants k_{10}, k_{12}, and k_{21} derived from the population pharmacokinetic model were used as the disposition input for each subject. The cumulative fraction absorbed was expressed as a percentage of the total amount absorbed, with the denominator calculated as k_{10} multiplied by the modified-release AUC extrapolated to infinity, and the extrapolation of the modified-release AUC used log-linear regression of the last three measurable modified-release plasma concentrations. The Loo-Riegelman method was selected as the primary deconvolution approach, in preference to numerical deconvolution, because it directly uses the individual micro-rate constants obtained from the same dosing occasion as the modified-release product, thereby removing between-occasion and between-subject disposition variability from the reconstructed absorption profile.

As a complementary parametric approach, Weibull absorption models were fitted to each subject's modified-release plasma concentration-time profile by numerical convolution with the individual two-compartment disposition function obtained from the population pharmacokinetic model[39,40]. A single-Weibull input function was fitted for subjects with 5 to 9 measurable modified-release concentrations, and a double-Weibull input function, permitting a biphasic release profile, was fitted for subjects with 10 or more measurable modified-release concentrations. Subjects with fewer than 5 measurable modified-release concentrations were not included in the Weibull deconvolution analysis.

5.3.10 Software and reproducibility

All numerical analyses in this study were performed in the open-source statistical computing environment R (version 4.5.2) and in NONMEM (version 7.5, ICON plc, Dublin, Ireland), with dataset assembly, non-compartmental analysis, deconvolution, descriptive statistics, and post-processing of population pharmacokinetic output performed in R, and population pharmacokinetic estimation performed in NONMEM. Non-compartmental analysis used the PKNCA package (version 0.12.1). Data manipulation relied on the tidyverse family of

packages, and all figures in this study were generated using the ggplot2 graphics package using a consistent chapter-wide set of typographic and color conventions.

5.4 Results

5.4.1 Subject disposition and demographics

The primary analysis population for this study comprises 16 healthy adult volunteers who received one of the two modified-release glipizide formulations under study and provided evaluable plasma pharmacokinetic data: 7 who received Product 1 (osmotic pump) and 9 who received Product 2 (hydrophilic matrix). Demographic and baseline characteristics of the primary analysis population are summarized in **Table 5-1** and were broadly comparable between the two formulation arms.

Between September 30, 2020 and March 29, 2026, 37 individuals completed informed consent and the screening visit. Three were screen failures and four were not approved for further enrollment by the study physicians, yielding 30 volunteers approved for enrollment; of these, ten withdrew prior to dosing and 20 completed Phase 1 drug administration in the parent study. The 16 volunteers in the primary analysis population for this study are the subset of Phase 1 completers who received the two formulations analyzed here. Three of the 16 primary-analysis volunteers (G1, G4, and G5), all in the osmotic pump arm, subsequently completed matched non-intubation Phase 2 procedures, providing the within-subject paired data analyzed in **5.3.7**.

Table 5-1. Demographic and baseline characteristics of the primary analysis population, by modified-release formulation arm.

Characteristic	Osmotic pump (n = 7)	Hydrophilic matrix (n = 9)
Sex, n (%)		
Male	3 (42.9)	4 (44.4)
Female	4 (57.1)	5 (55.6)
Race, n (%)		
White	5 (71.4)	7 (77.8)
Black or African American	1 (14.3)	1 (11.1)
Asian	0 (0.0)	1 (11.1)
Other	1 (14.3)	0 (0.0)
Ethnicity, n (%)		
Hispanic or Latino	2 (28.6)	1 (11.1)
Not Hispanic or Latino	4 (57.1)	8 (88.9)
Age (years)		
Mean ± SD	26.1 ± 4.0	26.8 ± 8.9
Median (range)	26.0 (21.0 - 31.0)	22.0 (18.0 - 44.0)
Weight (kg)		
Mean ± SD	77.2 ± 17.3	72.4 ± 16.0
Median (range)	86.4 (55.0 - 98.0)	71.7 (53.5 - 108.0)
Height (cm)		
Mean ± SD	172.4 ± 12.3	175.4 ± 8.5
Median (range)	175.0 (154.9 - 187.0)	179.6 (160.0 - 186.7)
BMI (kg/m²)		
Mean ± SD	25.7 ± 3.9	23.4 ± 4.1
Median (range)	24.7 (21.0 - 30.9)	22.4 (18.5 - 33.2)

5.4.2 Plasma pharmacokinetics

Plasma concentration-time profiles of modified-release glipizide and of the co-administered $^{13}C_6$-glipizide oral solution reference are shown in **Figure 5-1**, with formulation arm on rows (osmotic pump, hydrophilic matrix) and scale on columns (linear, semi-logarithmic). For both formulations, geometric mean modified-release profiles rose from near zero to a peak between approximately 8 and 12 hours post-dose, reaching geometric mean peaks of approximately 150 ng/mL for the osmotic pump arm and approximately 100 ng/mL for the hydrophilic matrix arm, and declined monotonically thereafter through the remainder of the 78-hour sampling window. The two arms were closely similar in the shape and magnitude of their

geometric mean modified-release profiles across both linear and semi-logarithmic panels, and the plasma profile alone did not visibly distinguish between release mechanisms. This observation motivates the regional luminal observations reported in **5.3.3** and the deconvolution analysis reported in **5.3.8** as complementary, methodologically independent characterizations of *in vivo* release. Individual modified-release observations exhibited substantial scatter around the geometric mean trajectory in both arms, with single-subject concentrations reaching approximately 270 ng/mL during the early-to-mid post-dose interval and remaining measurable above the lower limit of quantification through approximately 70 hours post-dose in some matrix-arm subjects. Non-compartmental pharmacokinetic parameters (C_{max}, T_{max}, AUC_{inf}, and $t_{1/2}$) by formulation are reported in **Table 5-2**.

Table 5-2. Non-compartmental pharmacokinetic parameters for plasma modified-release glipizide and for the co-administered $^{13}C_6$-glipizide oral solution reference, summarized by formulation arm. Values are presented as geometric mean (geometric coefficient of variation, %) for C_{max}, and AUC_{inf}; as median (range) for T_{max}; and as arithmetic mean (coefficient of variation, %) for terminal half-life.

Parameter	Osmotic pump (n = 7)	Hydrophilic matrix (n = 9)
Modified-release glipizide (5 mg)		
Cmax (ng/mL)	149.3 (41%)	99.3 (88%)
Tmax (h)	8.51 (7.00–18.37)	9.75 (4.00–35.17)
AUCinf (ng·h/mL)	2618.1 (135%)	2351.2 (95%)
t½ (h)	6.1 (17%)	8.0 (29%)
$^{13}C_6$-glipizide oral solution (1 mg)		
Cmax (ng/mL)	74.1 (59%)	75.9 (35%)
Tmax (h)	0.42 (0.17–0.52)	0.67 (0.37–2.00)
AUCinf (ng·h/mL)	430.5 (59%)	383.7 (34%)
t½ (h)	4.4 (53%)	4.6 (21%)

Plasma profiles for the co-administered 1 mg $^{13}C_6$-glipizide oral solution are overlaid in **Figure 5-1** (dashed lines for the geometric mean; open squares for individual observations). The isotope geometric mean profile exhibited a single, sharply defined absorption phase reaching a geometric mean peak of approximately 75 ng/mL within the first hour post-dose, followed by an apparent monoexponential decline visible on the semi-logarithmic panels that approached the lower limit of quantification by approximately 30 to 45 hours post-dose. The isotope profile was visually closely similar in shape and magnitude between the two formulation arms (**Table 5-2**). This shape is characteristic of an oral solution whose appearance in plasma is rate-limited by

absorption rather than by tablet disintegration or dissolution and is consistent with its intended role as a subject- and occasion-matched disposition reference for the modified-release analyte.

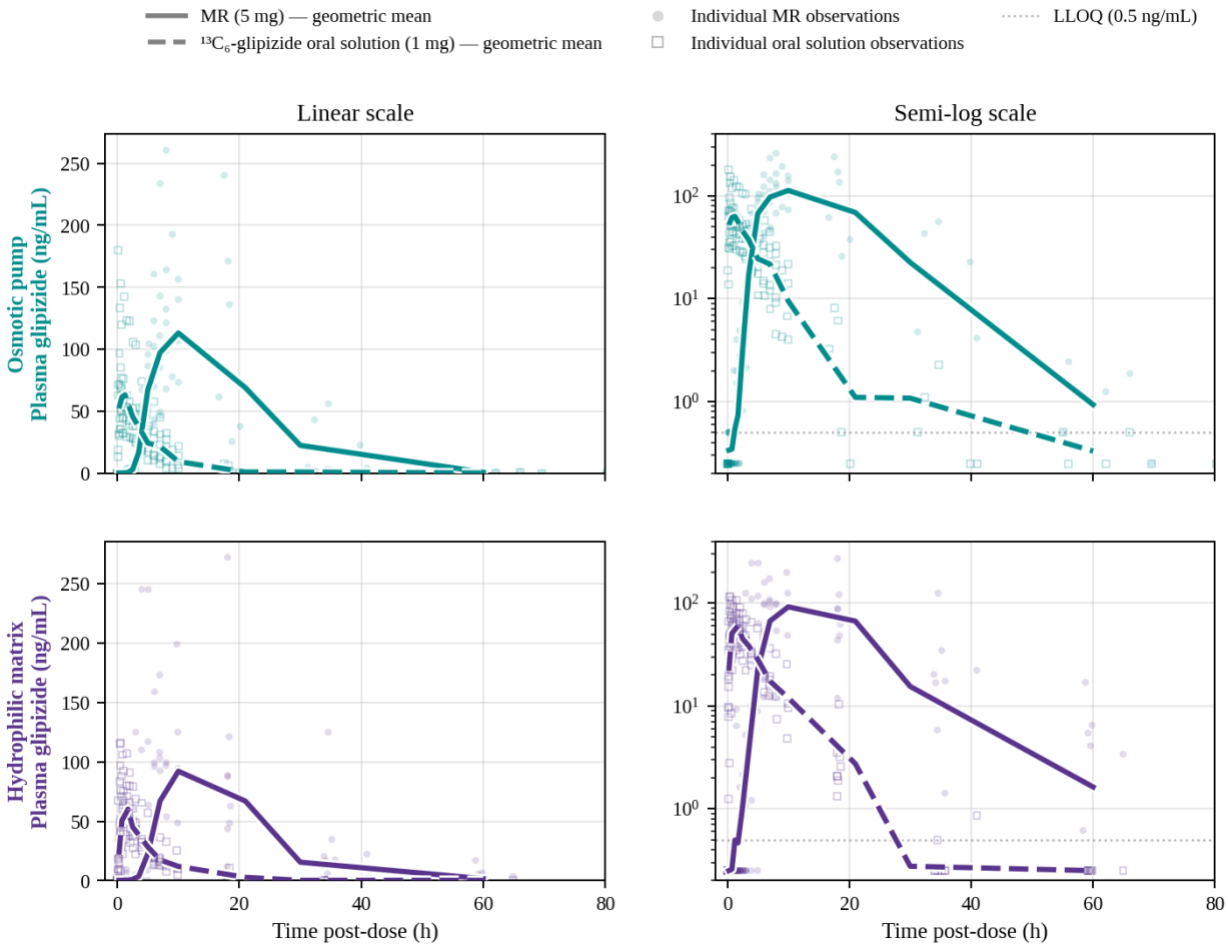


Figure 5-1. Plasma glipizide concentration-time profiles for the modified-release (MR) glipizide tablet (5 mg) and the co-administered $^{13}C_6$-glipizide oral solution reference (1 mg). Panels are arranged with formulation arm on rows (osmotic pump, top; hydrophilic matrix, bottom) and concentration scale on columns (linear, left; semi-logarithmic, right). Solid colored lines are geometric mean MR profiles and filled circles are individual MR observations; dashed colored lines are geometric mean oral solution profiles and open squares are individual oral solution observations. The horizontal dotted line on the semi-logarithmic panels indicates the lower limit of quantification (0.5 ng/mL).

5.4.3 Regional gastrointestinal disposition

Luminal modified-release glipizide concentrations measured by direct aspiration from the stomach, duodenum, and jejunum are shown by formulation arm and region in **Figure 5-2**. The two formulation arms produced different regional patterns of luminal drug appearance. For the osmotic pump arm, measurable luminal glipizide was largely confined to the stomach: stomach observations rose progressively over the seven-hour sampling window from near-zero pre-dose values to a scattered distribution of measured concentrations, with most observations falling

below 200 ng/mL and individual observations reaching approximately 400, 500, and 830 ng/mL at 4 to 7 hours post-dose. Duodenum and jejunum observations from the osmotic pump arm were almost entirely at or below the lower limit of quantification throughout the sampling window, with a single measurable jejunal observation of approximately 18 ng/mL at three hours.

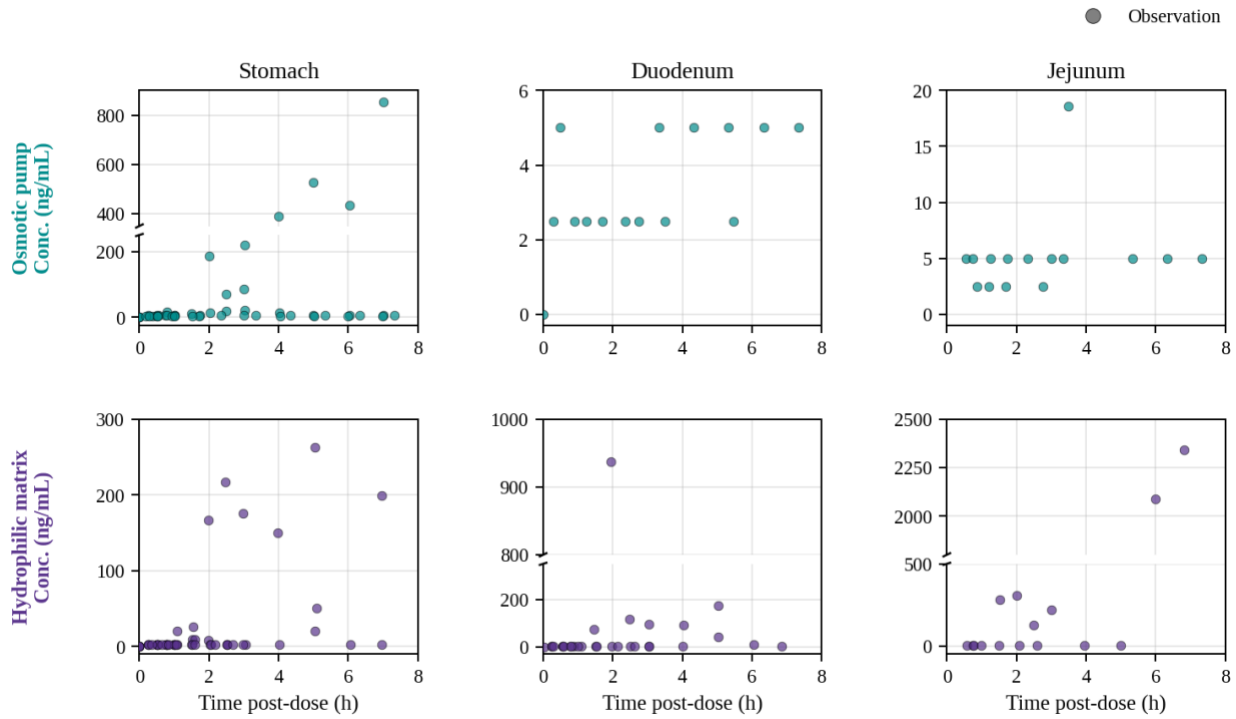


Figure 5-2. Luminal modified-release glipizide concentrations in the supernatant fraction of aspirates from the primary analysis population. Panels are arranged with formulation arm on rows (osmotic pump, top; hydrophilic matrix, bottom) and intraluminal sampling site on columns (stomach, duodenum, jejunum); the x-axis is time post-dose (h) and the y-axis is luminal glipizide concentration (ng/mL). Individual aspirate observations are shown as filled circles. Each panel uses an independent linear y-axis scaled to the data within that panel; axis breaks are used where individual high-leverage observations exceed the principal data range. Below-quantification-limit observations are plotted at imputed values (lower limit of quantification, 0.5 ng/mL or 0.25 ng/mL).

For the hydrophilic matrix arm, measurable luminal glipizide appeared in all three regions, with the largest individual observations occurring in the small bowel rather than in the stomach. Matrix-arm stomach observations distributed across a broader range than for the osmotic pump arm, with several observations between 150 and 265 ng/mL across 2 to 7 hours post-dose. Duodenum observations included a single high-leverage observation of approximately 940 ng/mL at approximately 2 hours post-dose, with a scattered background of lower observations through the remainder of the sampling window. Jejunum observations included two large late-window observations of approximately 2150 and 2330 ng/mL at 6 to 7 hours post-dose in two matrix-arm subjects, with a long lower-concentration tail across the rest of the population.

The matrix-arm pattern is therefore characterized by spatially and temporally distributed luminal appearance with a small number of high-magnitude individual observations, in marked contrast to the stomach-confined pattern of the osmotic pump arm.

Luminal concentrations of the co-administered $^{13}C_6$-glipizide oral solution, pooled across the primary analysis population, are shown by region in **Figure 5-3**. The regional pattern was that of rapid proximal appearance with progressively attenuated and earlier-resolving observations from proximal to distal. Stomach observations included two high early observations of approximately 5700 and 3550 ng/mL within the first 30 minutes post-dose; the remaining stomach observations were predominantly within 200 to 1500 ng/mL during the first hour and declined to near-zero values by approximately three hours post-dose. Duodenum observations included one large early observation of approximately 1130 ng/mL near 30 minutes, with the remaining observations clustered below approximately 600 ng/mL early and declining through the sampling window. Jejunum observations were of substantially lower magnitude than either proximal region throughout, with two early observations of approximately 180 and 230 ng/mL near 30 minutes followed by a flat tail of observations predominantly below 80 ng/mL through the remainder of the sampling window. The regional pattern of isotope luminal appearance is consistent with rapid gastric emptying of an oral aqueous solution, progressive dilution into the small bowel, and the absence of any sustained distal release.

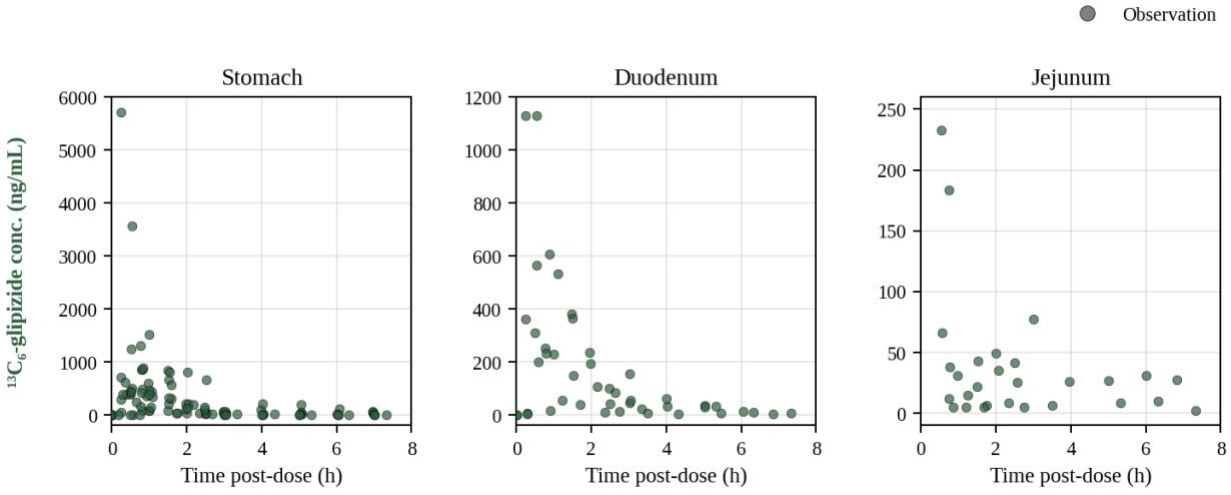


Figure 5-3. Luminal $^{13}C_6$-glipizide concentrations in the supernatant fraction of aspirates from the 1 mg oral solution reference, pooled across the primary analysis population (both formulation arms combined). Panels are arranged with intraluminal sampling site on columns (stomach, duodenum, jejunum); the x-axis is time post-dose (h) and the y-axis is luminal $^{13}C_6$-glipizide concentration (ng/mL). Individual aspirate observations are shown as filled circles. Each panel uses an independent linear y-axis scaled to the data within that panel.

5.4.4 Luminal pH and bile acid environment

Regional luminal pH measurements obtained *ex vivo* from each aspirate are shown by region in **Figure 5-4**. The three regions occupied distinctly different pH ranges. Duodenum and jejunum pH readings clustered between approximately 6 and 8 across the seven-hour sampling window, with no apparent time trend in either region and a small number of individual observations falling to approximately pH 3.5 to 4.7 in single aspirates. Stomach pH readings, in contrast, distributed bimodally: the majority of observations clustered in a typical low gastric pH range between approximately 1 and 3, while a smaller second cluster fell between approximately pH 6.5 and 8 and overlapped the pH range observed in the duodenum and jejunum. The bimodality was sustained throughout the seven-hour sampling window, with both modes populated at all sampled timepoints. Inspection at the subject level indicated that the bimodality was subject-driven, with a minority of volunteers contributing the high-pH stomach observations and the remainder distributed within the typical low gastric pH range. Glipizide is a weak acid with a *p*Ka of approximately 5.9, and its aqueous solubility is strongly pH-dependent, increasing by roughly two orders of magnitude between typical gastric pH and the near-neutral pH of the small bowel; the regional pH distribution observed here therefore defines a luminal environment in which glipizide solubility is low in the stomach of most subjects, comparable to small-bowel solubility in the minority subset of subjects with a higher-pH stomach environment, and consistently higher upon entry into the duodenum and jejunum.

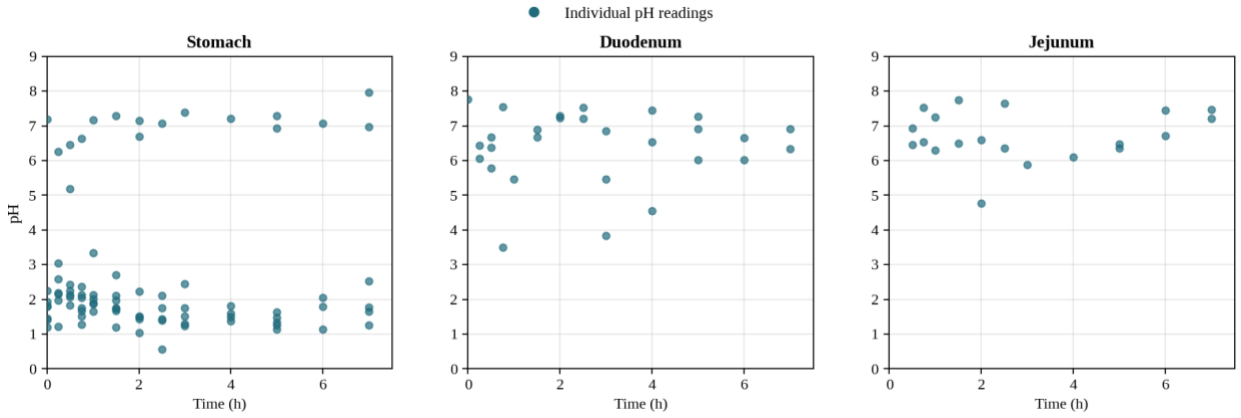


Figure 5-4. Regional gastrointestinal luminal pH by region (panels: stomach, duodenum, jejunum), measured ex vivo from each aspirate. Individual pH readings shown as filled circles.

Beyond the pH-driven contribution to apparent aqueous solubility, bile acids present in the duodenum and jejunum further modulate the solubilization of poorly soluble weak acids through mixed-micelle formation. Regional bile acid concentrations therefore provide the *in vivo* solubilization context against which the regional release patterns of the two MR test articles can be interpreted. Concentrations are summarized at the class level (primary unconjugated, primary conjugated, secondary unconjugated, secondary conjugated) in **Figure 5-5** and resolved into the 15 individual bile acid species in **Figures 5-6** to **5-9**, by gastrointestinal sampling site (stomach, duodenum, jejunum). Time-averaged geometric mean concentrations across the seven-hour post-dose sampling window differed substantially between regions and between bile acid classes. Within the small bowel, primary conjugated species were the dominant class. Time-averaged geometric mean concentrations of primary conjugated species were approximately 0.7 mM in the duodenum and 2.4 mM in the jejunum; secondary conjugated species were several-fold lower (approximately 0.18 mM and 0.8 mM, respectively); and unconjugated species were several orders of magnitude lower again, in the sub-micromolar range (geometric means below 0.001 mM at both small-bowel ports). Duodenum-to-stomach geomean ratios were approximately 30- to 40-fold for both conjugated classes, and jejunum-to-stomach ratios were approximately 120- to 180-fold. The biopharmaceutic consequence for an MR glipizide product is that drug released into the duodenum or jejunum enters a luminal environment with bile acid concentrations approximately 30- to 40-fold higher than those measured in the stomach, supporting bile-acid-mediated solubilization of glipizide in the small bowel that is not available in the gastric compartment. Considered together with the regional pH observations, the regional luminal environment characterized here is one in which proximal release places drug into a low pH, low bile acid environment with limited solubilization capacity, while distal release places drug into a near-neutral, bile-acid-supplemented environment that supports both ionization and bile acid-mediated solubilization of glipizide.

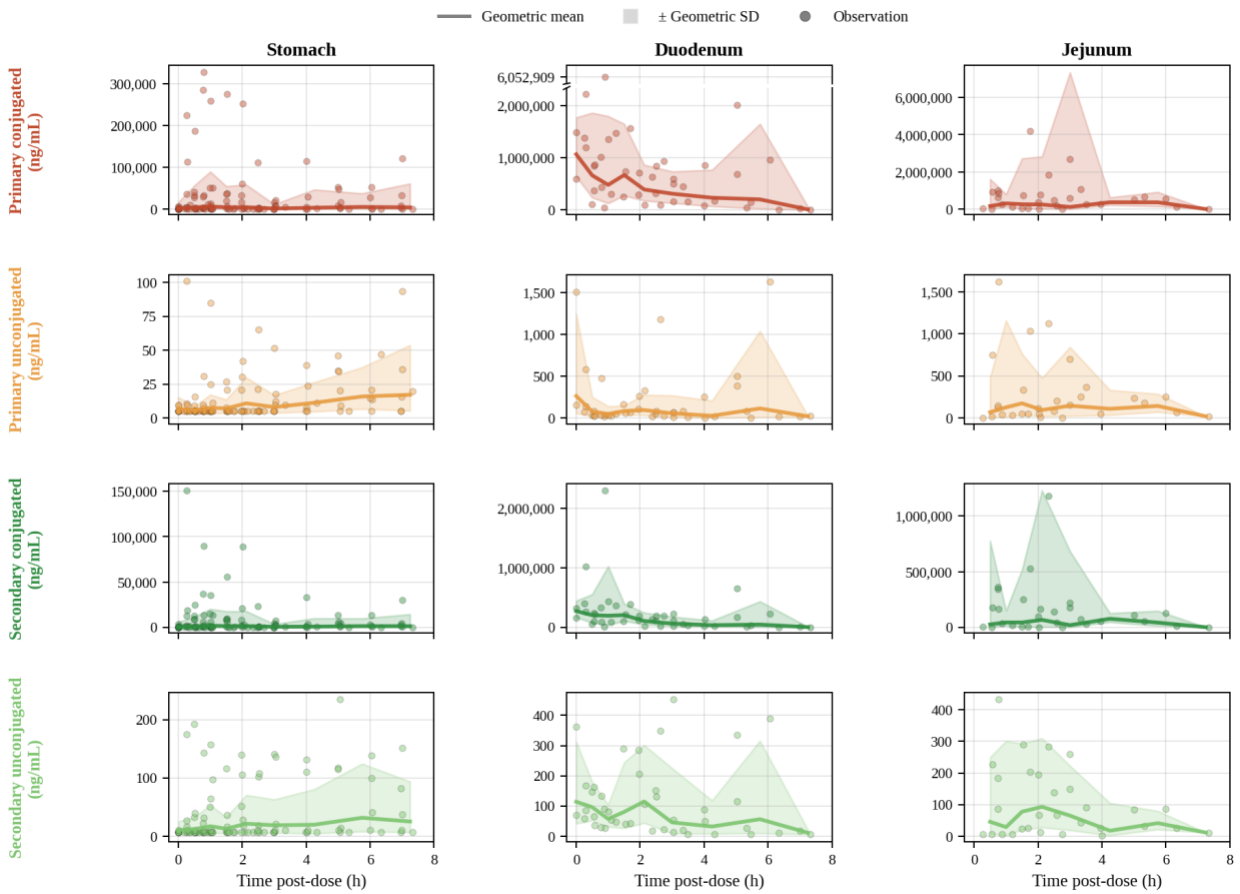


Figure 5-5. Intraluminal bile acid concentrations by class and gastrointestinal sampling site. Panels are arranged with bile acid class as rows (primary conjugated, primary unconjugated, secondary conjugated, secondary unconjugated; row labels color-coded as in the panels) and sampling site as columns (stomach, duodenum, jejunum); the x-axis is time post-dose (h) and the y-axis is class-summed bile acid concentration (ng/mL). Solid lines are geometric means and shaded bands are ± 1 geometric standard deviation across the primary analysis population; individual observations are shown as filled circles. Each panel uses an independent linear y-axis scaled to the data within that panel; axis breaks are used where individual high-leverage observations exceed the principal data range.

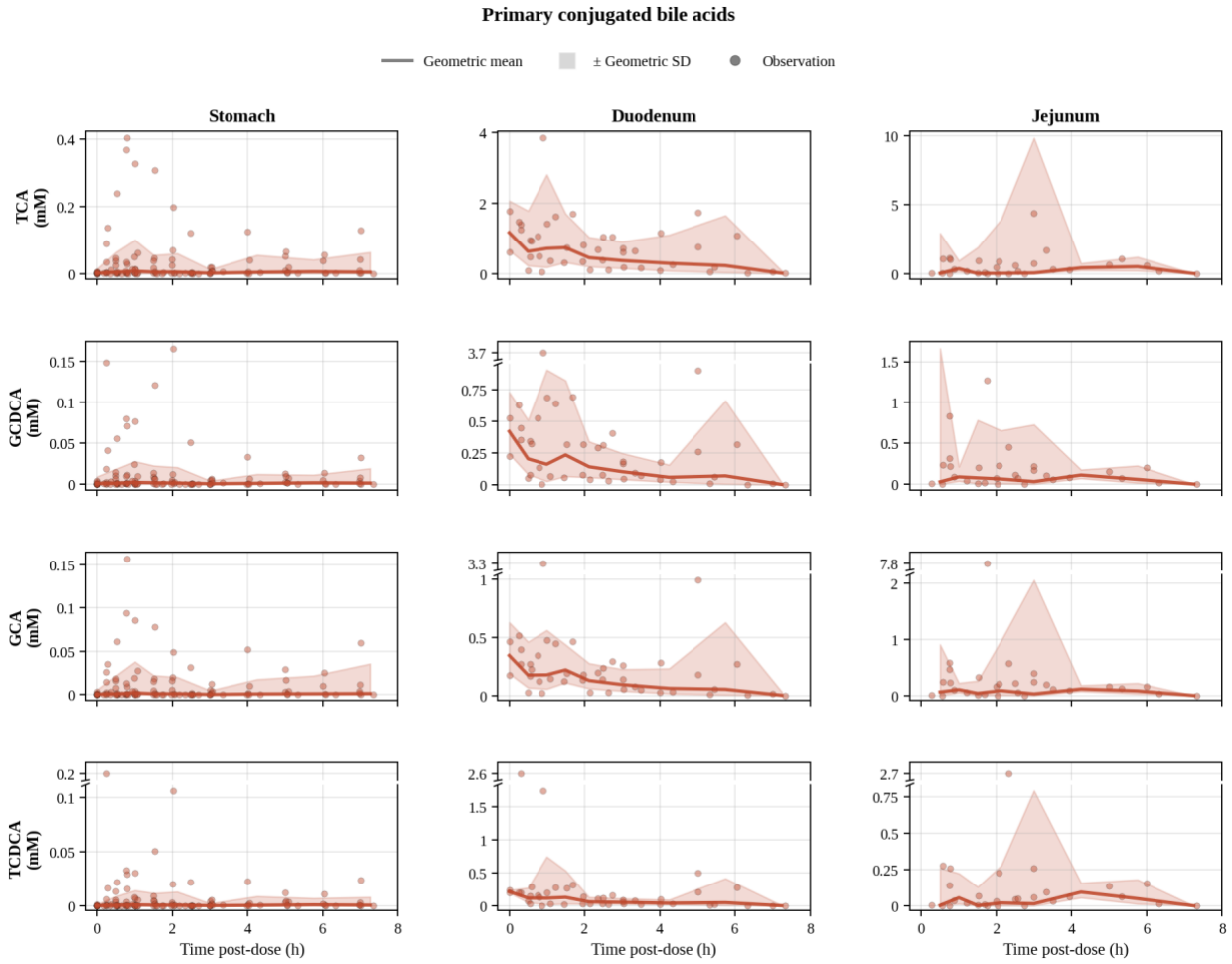


Figure 5-6. Intraluminal concentrations of the four primary conjugated bile acid species. Panels are arranged with species as rows (TCA, taurocholic acid; GCDCA, glycochenodeoxycholic acid; GCA, glycocholic acid; TCDCA, taurochenodeoxycholic acid) and sampling site as columns (stomach, duodenum, jejunum); the x-axis is time post-dose (h) and the y-axis is concentration of the indicated species (mM). Solid lines are geometric means and shaded bands are ± 1 geometric standard deviation across the primary analysis population; individual observations are shown as filled circles. Each panel uses an independent linear y-axis scaled to the data within that panel; axis breaks are used where individual high-leverage observations exceed the principal data range.

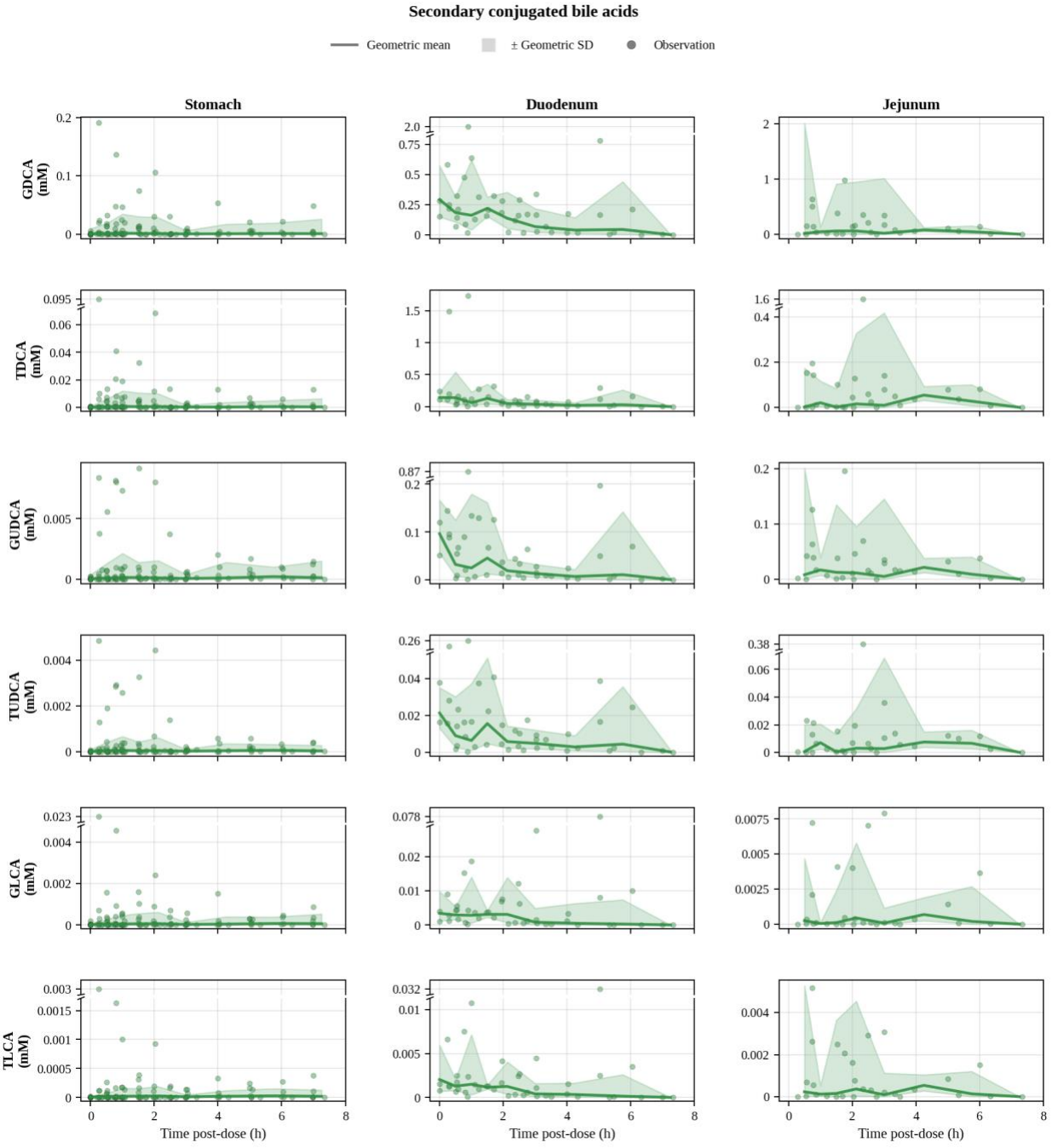


Figure 5-7. Intraluminal concentrations of the six secondary conjugated bile acid species. Panels are arranged with species as rows (GDCA, glycodeoxycholic acid; TDCA, taurodeoxycholic acid; GUDCA, glycoursodeoxycholic acid; TUDCA, tauroursodeoxycholic acid; GLCA, glycolithocholic acid; TLCA, taurolithocholic acid) and sampling site as columns (stomach, duodenum, jejunum); the x-axis is time post-dose (h) and the y-axis is concentration of the indicated species (mM). Solid lines are geometric means and shaded bands are ± 1 geometric standard deviation across the primary analysis population; individual observations are shown as filled circles. Each panel uses an independent linear y-axis scaled to the data within that panel; axis breaks are used where individual high-leverage observations exceed the principal data range.

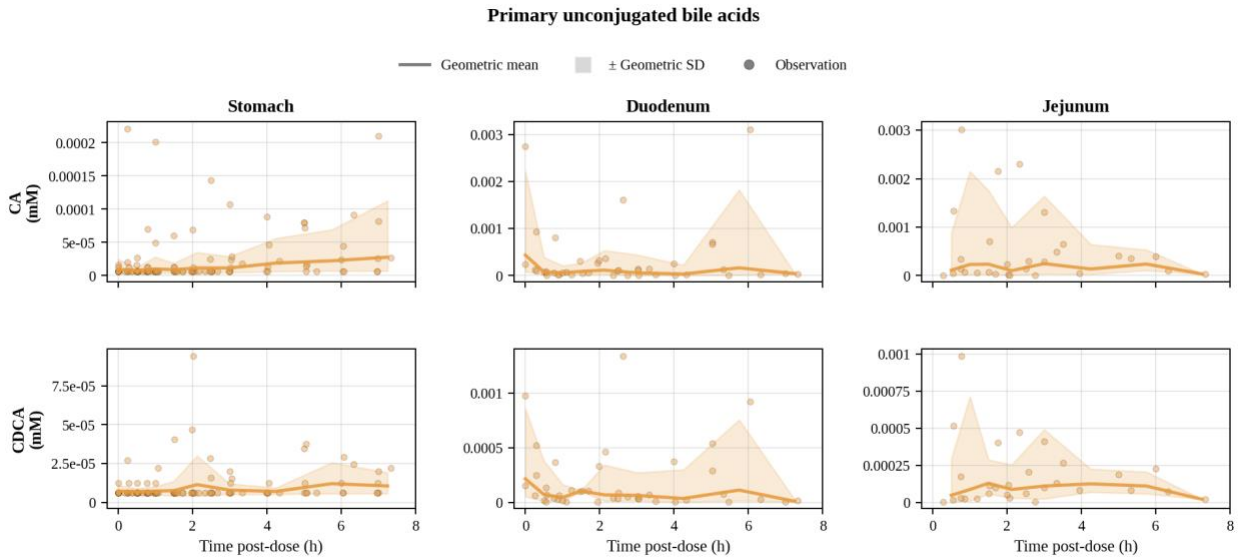


Figure 5-8. Intraluminal concentrations of the two primary unconjugated bile acid species. Panels are arranged with species as rows (CA, cholic acid; CDCA, chenodeoxycholic acid) and sampling site as columns (stomach, duodenum, jejunum); the x-axis is time post-dose (h) and the y-axis is concentration of the indicated species (mM). Solid lines are geometric means and shaded bands are ± 1 geometric standard deviation across the primary analysis population; individual observations are shown as filled circles. Each panel uses an independent linear y-axis scaled to the data within that panel.

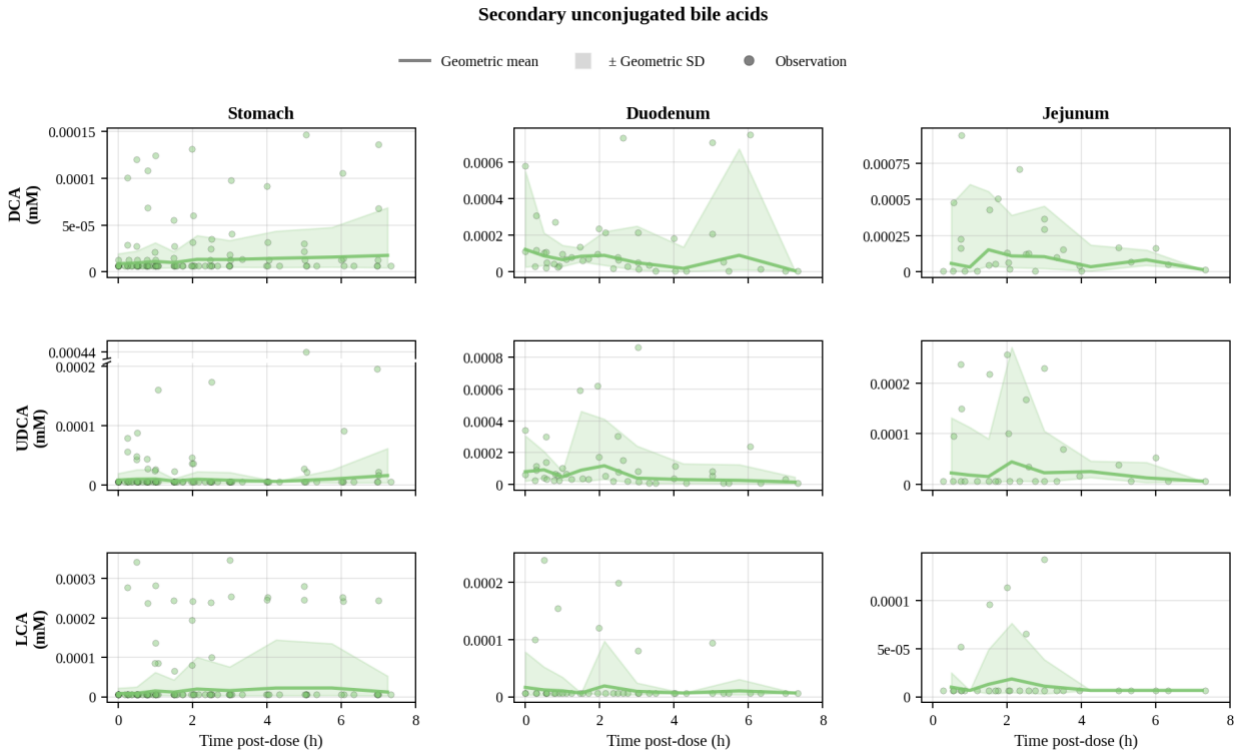


Figure 5-9. Intraluminal concentrations of the three secondary unconjugated bile acid species. Panels are arranged with species as rows (DCA, deoxycholic acid; UDCA, ursodeoxycholic acid; LCA, lithocholic acid) and sampling site as columns (stomach, duodenum, jejunum); the x-axis is time post-dose (h) and the y-axis is concentration of the indicated species (mM). Solid lines are geometric means and shaded bands are ± 1 geometric standard deviation

across the primary analysis population; individual observations are shown as filled circles. Each panel uses an independent linear y-axis scaled to the data within that panel; axis breaks are used where individual high-leverage observations exceed the principal data range.

5.4.5 Urinary and fecal recovery

Cumulative urinary and fecal recovery of parent drug over the 78-hour collection window, expressed as arithmetic mean percent of administered dose, are shown by formulation arm in **Figures 5-10** (urinary) and **5-11** (fecal) for both the modified-release product (5 mg dose) and the co-administered $^{13}C_6$-glipizide oral solution (1 mg dose). Across both arms and both analytes, cumulative urinary and fecal recovery of unchanged parent drug accounted for less than 10% of the administered dose, consistent with the predominant role of hepatic metabolism in glipizide elimination[4].

Cumulative urinary recovery curves rose over the first approximately 30 hours post-dose and reached a plateau thereafter in both arms and for both analytes, indicating effectively complete capture of the urinary elimination phase within the 78-hour collection window. Mean cumulative urinary recovery of the modified-release analyte plateaued at approximately 6% of dose in the osmotic pump arm and approximately 5% of dose in the hydrophilic matrix arm. Mean cumulative urinary recovery of the $^{13}C_6$-glipizide oral solution plateaued at approximately 3% of dose in the osmotic pump arm and approximately 5% of dose in the hydrophilic matrix arm.

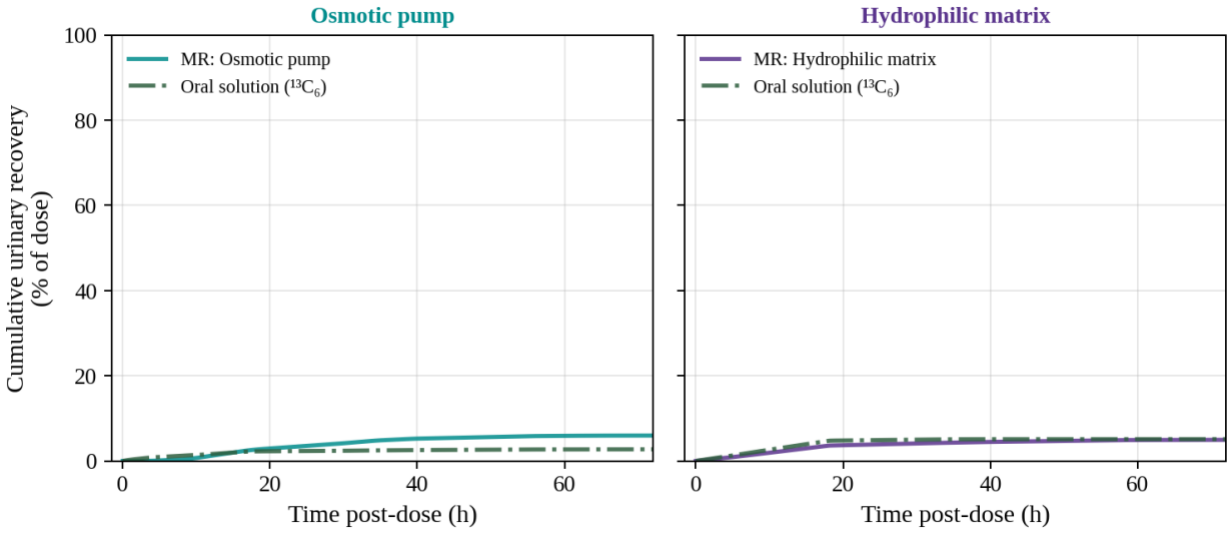


Figure 5-10. Mean cumulative urinary recovery of glipizide over time, by formulation arm, for both the modified-release product (5 mg dose) and the co-administered $^{13}C_6$-glipizide oral solution reference (1 mg dose). Each panel shows the arm-specific oral solution profile alongside the corresponding modified-release profile. Recovery is expressed as a percentage of the administered dose. Data are shown through approximately 72 hours post-dose, where all subjects contributed data.

Cumulative fecal recovery curves rose more gradually and plateaued at lower values than the corresponding urinary curves. Mean cumulative fecal recovery of the modified-release analyte reached approximately 1.5% of dose in the osmotic pump arm and approximately 3% of dose in the hydrophilic matrix arm by the end of the collection window. Mean cumulative fecal recovery of the $^{13}C_6$-glipizide oral solution remained near 1% of dose in both arms.

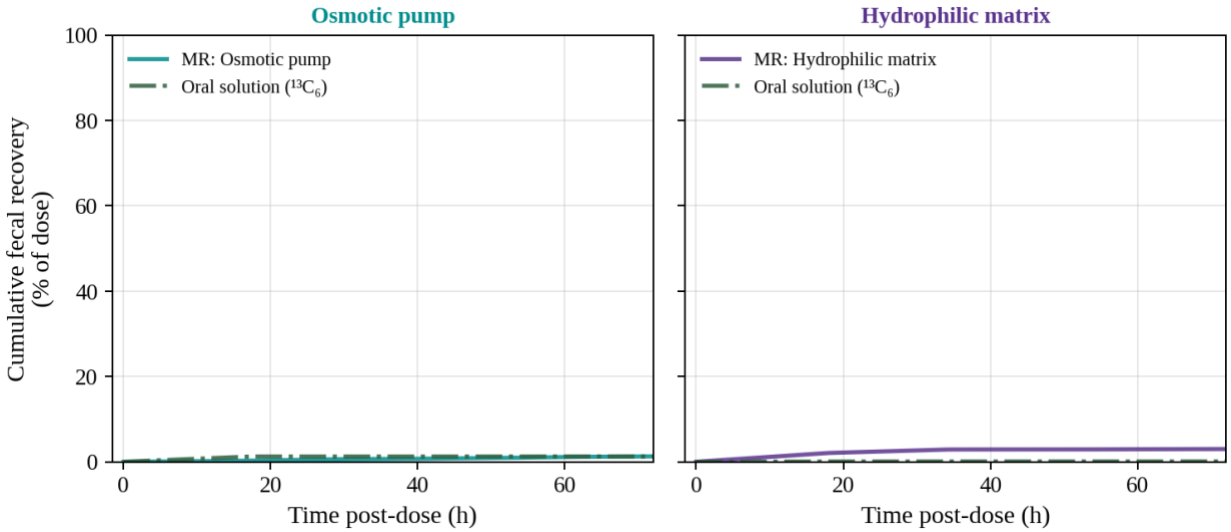


Figure 5-11. Mean cumulative fecal recovery of glipizide over time, by formulation arm, for both the modified-release product (5 mg dose) and the co-administered $^{13}C_6$-glipizide oral solution reference (1 mg dose). Each panel shows the arm-specific oral solution profile alongside the corresponding modified-release profile. Recovery is expressed as a percentage of the administered dose. Data are shown through approximately 72 hours post-dose, where all subjects contributed data.

5.4.6 Continuous glucose monitoring

Continuous interstitial glucose monitoring (CGM) was performed in subjects G12 through G22 (n = 11), with individual subject profiles shown in **Figures 5-12** through **5-22**. Each per-subject plot displays simultaneous glucose traces from a left-arm and a right-arm sensor against time relative to dose, with two horizontal reference lines indicating the 90 mg/dL hypoglycemia prevention threshold and the 70 mg/dL hypoglycemia treatment threshold; the corresponding euglycemic and hypoglycemic ranges are shaded in green and red, respectively. Rescue and prophylactic interventions are annotated on each plot as filled inverted triangles (D50 IV push), open inverted triangles (oral glucose tablets), and a lavender horizontal bar (D10 IV infusion interval).

An early post-dose rise in interstitial glucose, peaking at approximately 130 to 175 mg/dL within the first hour after dosing, was visible in most subjects and is attributable to the 20% glucose dosing vehicle co-administered with the modified-release tablet (and, in subjects who

received prophylactic D10 infusion, additionally to the infusion itself) rather than to a drug-related pharmacodynamic event. Beyond this early vehicle- and infusion-associated peak, interstitial glucose ran near pre-dose baseline through the in-clinic observation period in most subjects, within or close to the euglycemic range.

Excursions below the 90 mg/dL hypoglycemia prevention threshold occurred in 9 of the 11 post-amendment subjects, and excursions below the 70 mg/dL hypoglycemia treatment threshold occurred in 6 subjects, prompting protocol-defined rescue interventions that are annotated on each per-subject plot. The hypoglycemia management protocol evolved during the post-amendment phase of the study in two stages. Under the initial post-amendment protocol, intravenous administration of 50% dextrose (D50) by push was used as the rescue intervention for confirmed hypoglycemic events occurring during the intubation period, with subjects receiving repeated D50 boluses as needed when interstitial glucose fell below the treatment threshold; oral glucose tablets were used as the corresponding rescue intervention for events occurring after catheter removal. Following observations that several subjects experienced recurrent hypoglycemic excursions despite repeated rescue under this protocol, the protocol was amended to incorporate a prophylactic continuous intravenous infusion of 10% dextrose (D10) for subjects whose interstitial glucose approached or fell below the 90 mg/dL prevention threshold during the intubation period.

The evolving hypoglycemia management protocol had a visible impact on the per-subject glucose trajectories. Subjects managed under the initial D50-only protocol exhibited multiple sequential below-threshold excursions, with each event requiring rescue and post-rescue glucose returning to or near the baseline range only briefly before declining again. Subjects managed with the subsequent addition of D10 infusion exhibited more sustained correction of below-threshold trajectories and fewer subsequent rescue events during the infusion interval, although below-threshold excursions could recur after infusion discontinuation in some subjects. The CGM observations are accordingly reported here as a per-subject safety and feasibility characterization of the integrated methodology rather than as a formulation-discriminating pharmacodynamic endpoint.

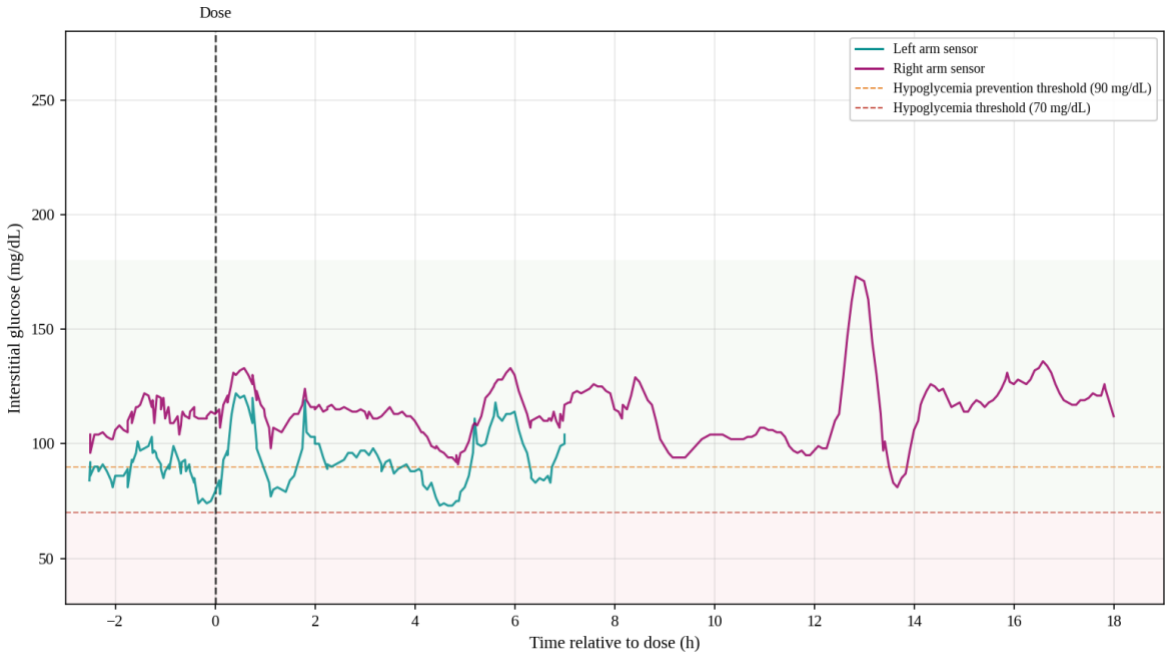


Figure 5-12. Continuous glucose monitoring profile for subject G12 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red).

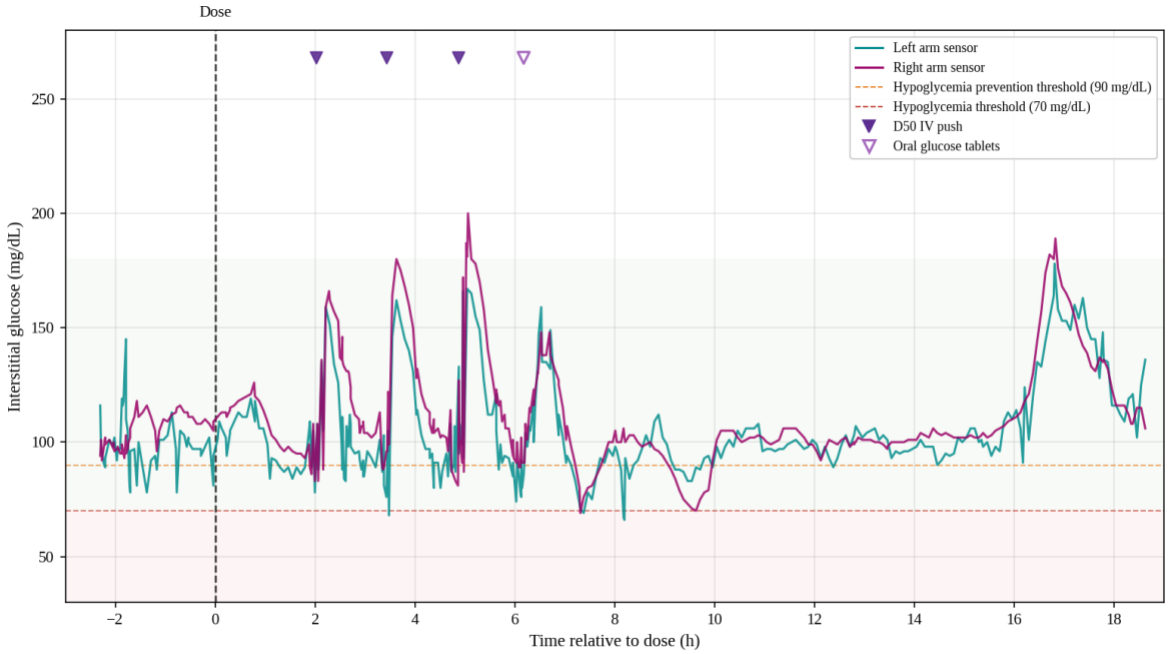


Figure 5-13. Continuous glucose monitoring profile for subject G13 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia

treatment threshold (red). Filled inverted triangles denote intravenous administration of 50% dextrose (D50) and the open inverted triangle denotes oral glucose tablet administration, each given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

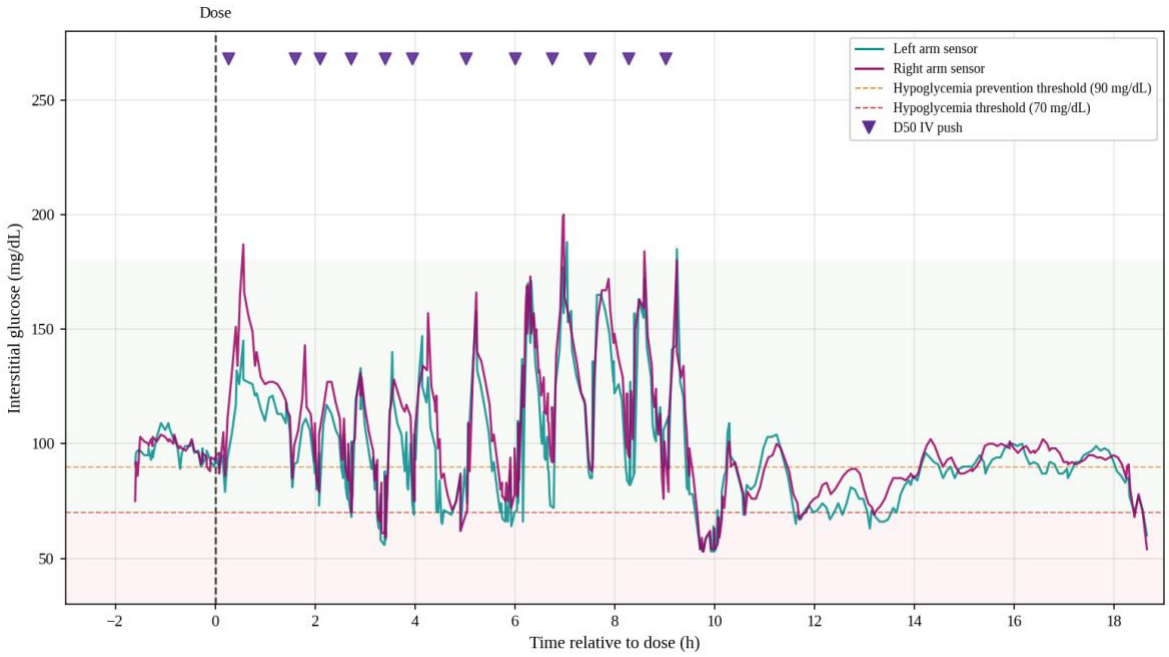


Figure 5-14. Continuous glucose monitoring profile for subject G14 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red). Filled inverted triangles denote intravenous administration of 50% dextrose (D50), given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

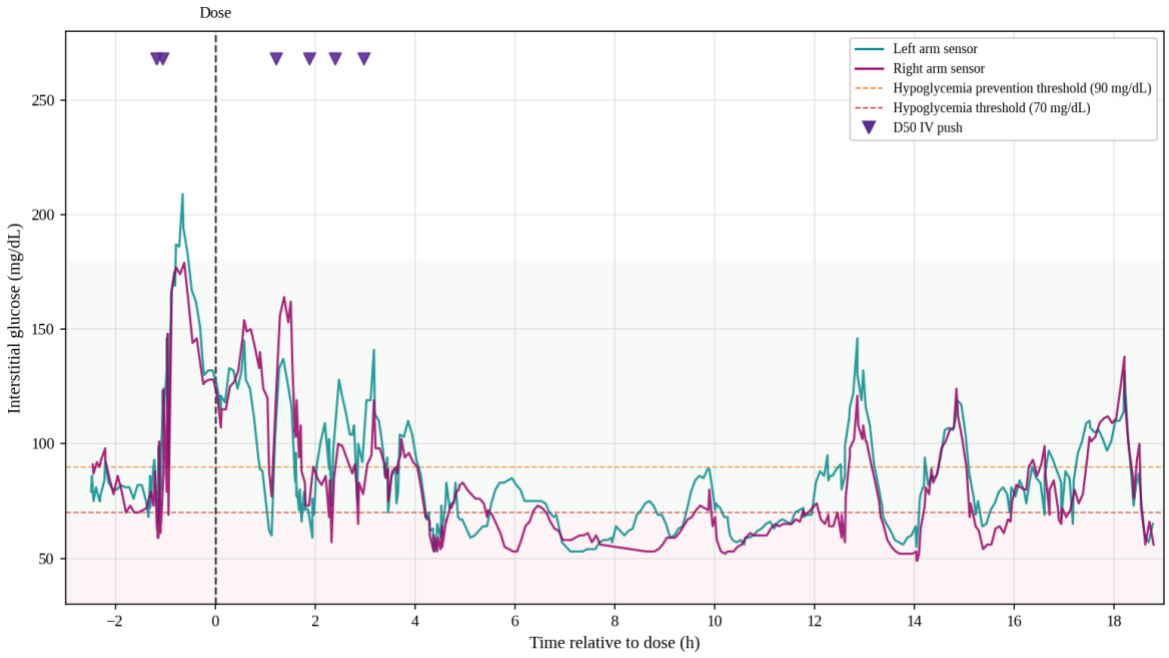


Figure 5-15. Continuous glucose monitoring profile for subject G15 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red). Filled inverted triangles denote intravenous administration of 50% dextrose (D50), given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

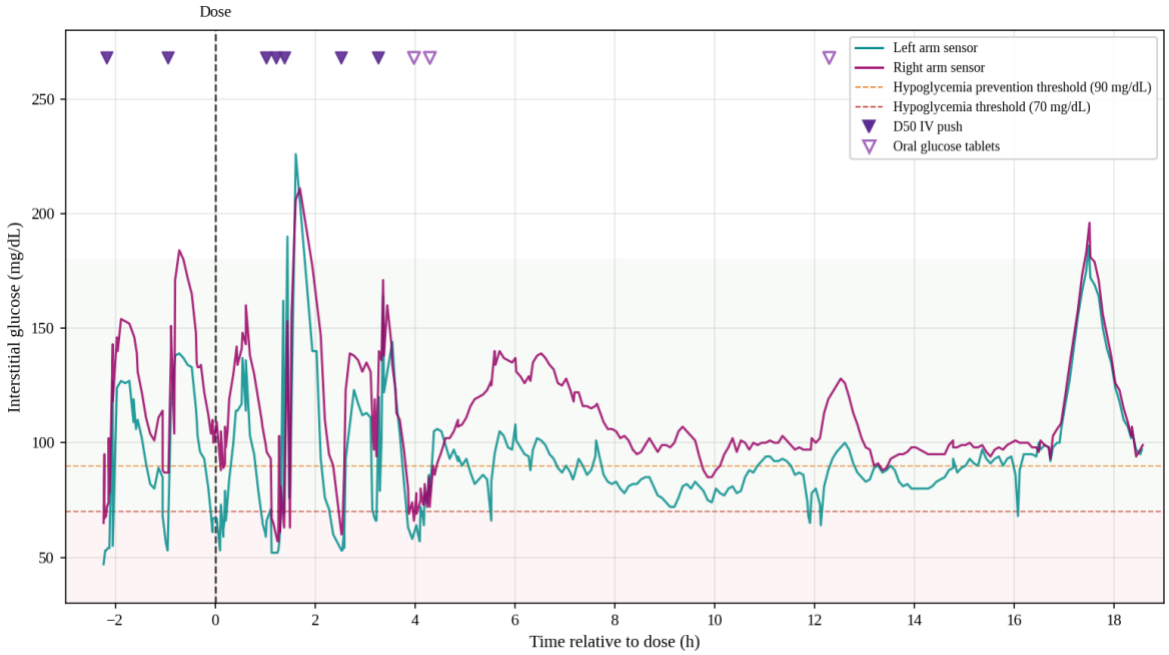


Figure 5-16. Continuous glucose monitoring profile for subject G16 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal

dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red). Filled inverted triangles denote intravenous administration of 50% dextrose (D50) and the open inverted triangle denotes oral glucose tablet administration, each given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

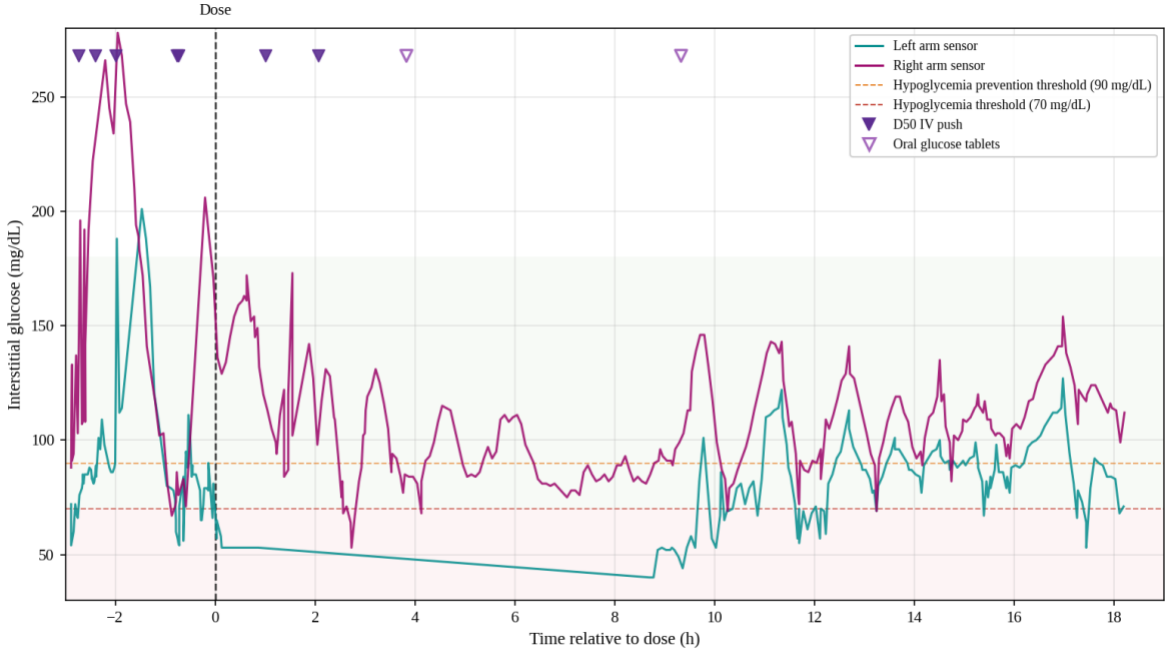


Figure 5-17. Continuous glucose monitoring profile for subject G17 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red). Filled inverted triangles denote intravenous administration of 50% dextrose (D50) and the open inverted triangle denotes oral glucose tablet administration, each given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

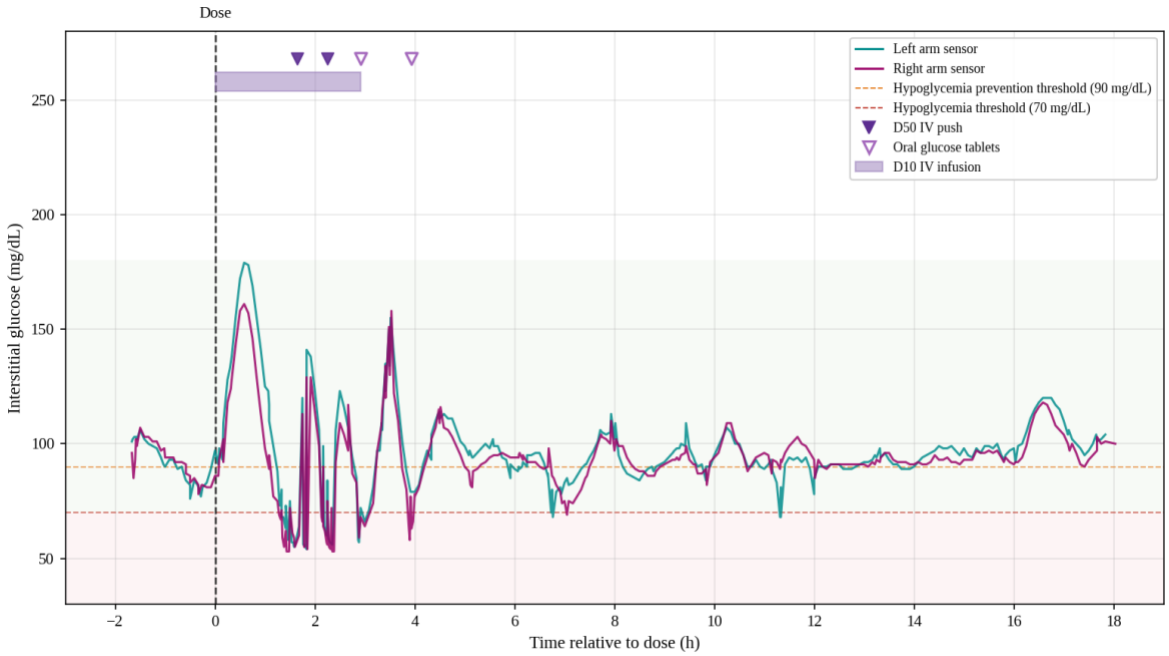


Figure 5-18. Continuous glucose monitoring profile for subject G18 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red). Filled inverted triangles denote intravenous administration of 50% dextrose (D50) by push, the open inverted triangle denotes oral glucose tablet administration, and the purple shaded bar denotes the duration of a 10% dextrose (D10) intravenous infusion, each given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

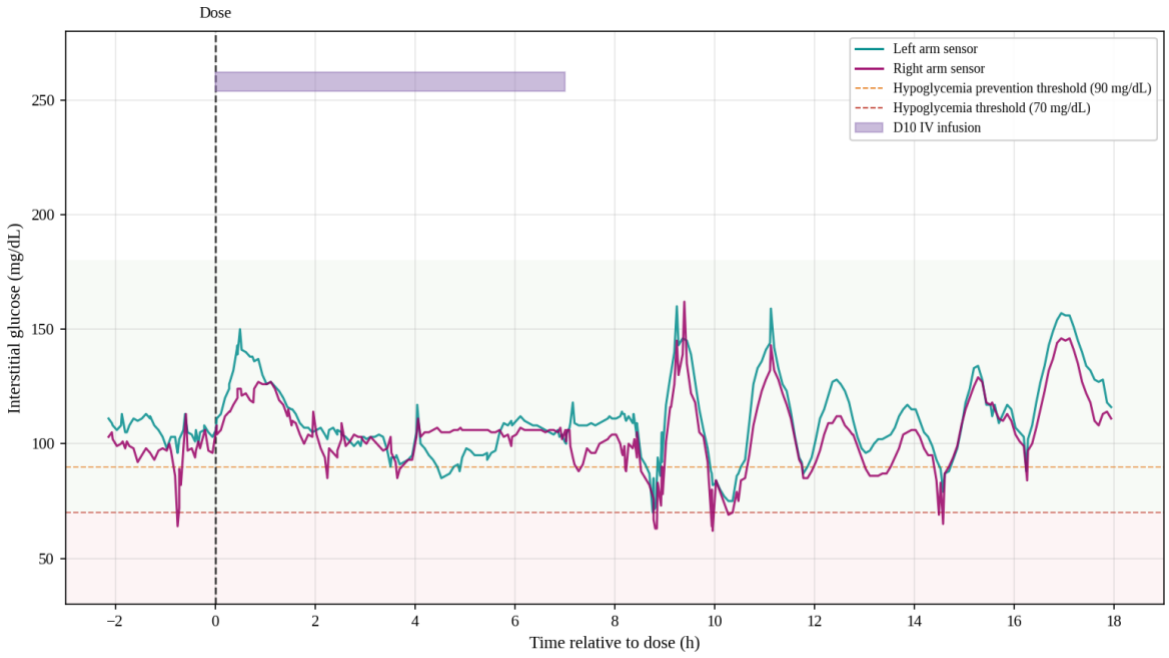


Figure 5-19. Continuous glucose monitoring profile for subject G19 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red). The purple shaded bar denotes the duration of a 10% dextrose (D10) intravenous infusion, given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

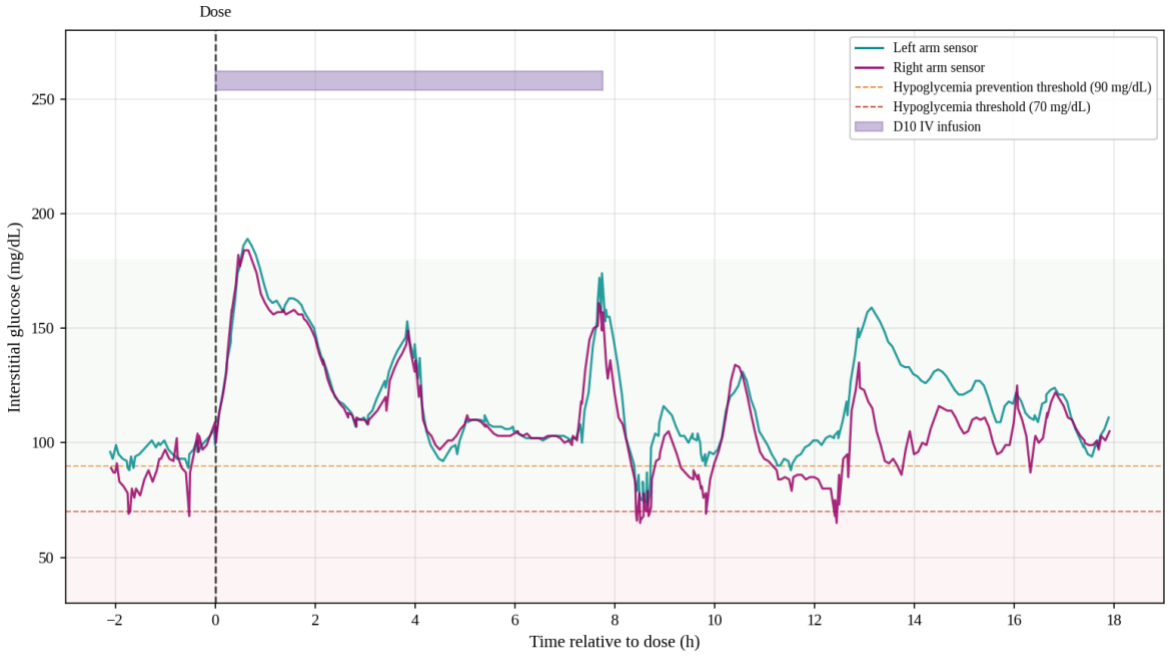


Figure 5-20. Continuous glucose monitoring profile for subject G20 following a single oral dose of modified-release glipizide (Product 2, hydrophilic polymer matrix; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal

dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red). The purple shaded bar denotes the duration of a 10% dextrose (D10) intravenous infusion, given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

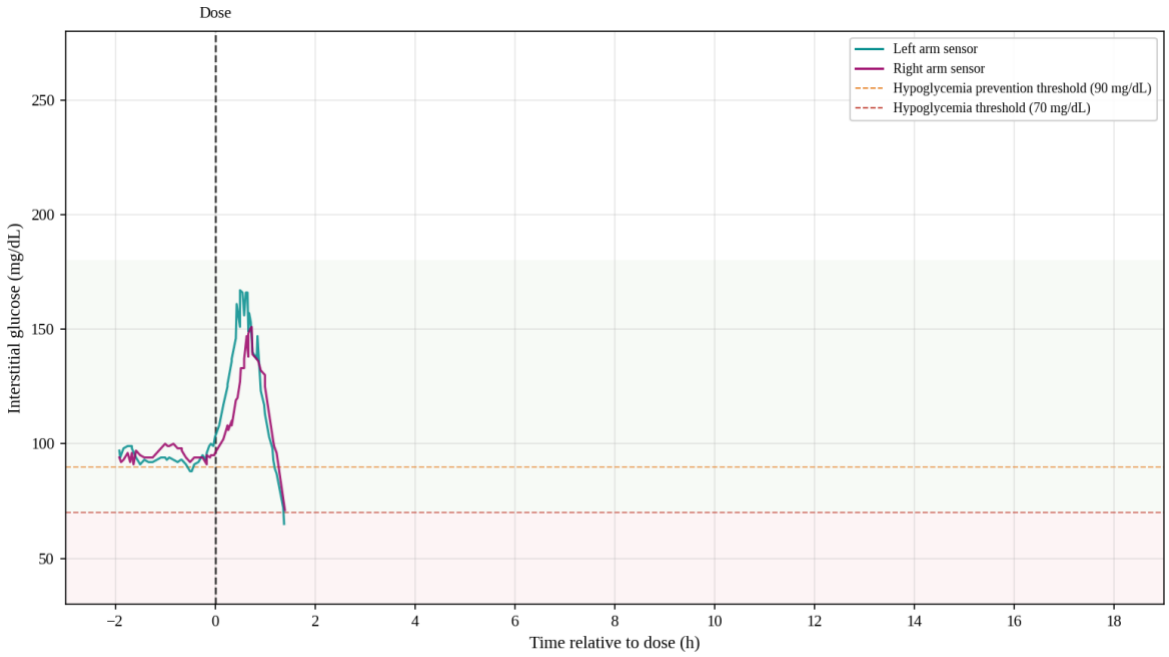


Figure 5-21. Continuous glucose monitoring profile for subject G21 following a single oral dose of modified-release glipizide (Product 1, osmotic pump; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red).

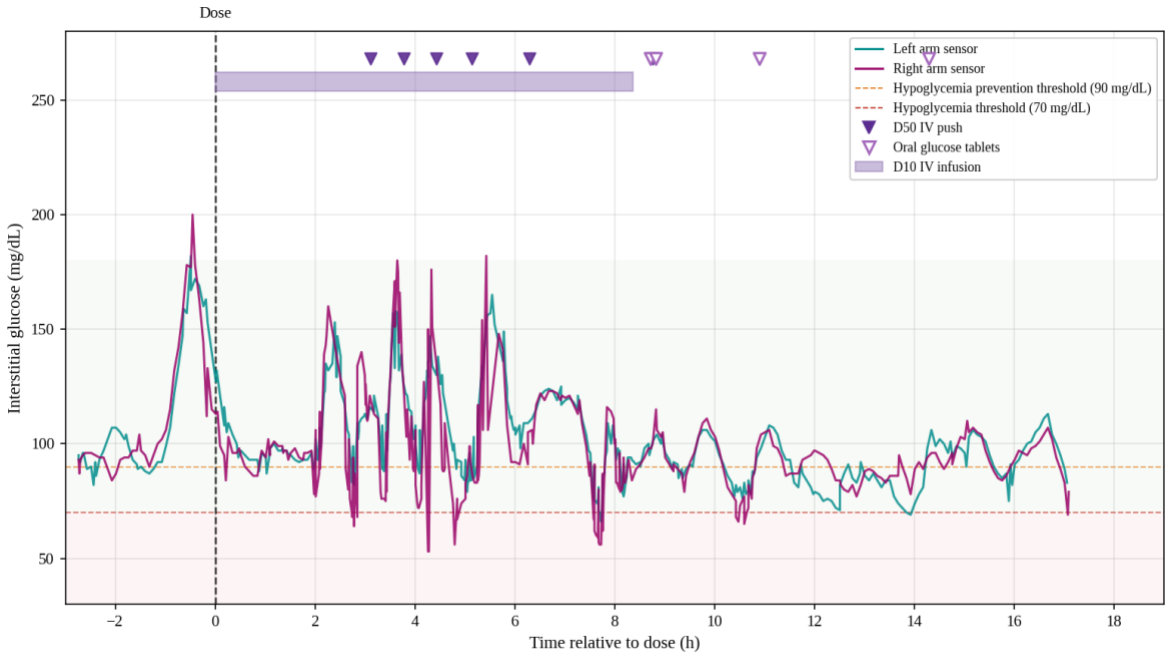


Figure 5-22. Continuous glucose monitoring profile for subject G22 following a single oral dose of modified-release glipizide (Product 1, osmotic pump; 5 mg). The vertical dashed line marks the time of study drug administration. Left and right upper-arm sensor traces are shown in teal and magenta, respectively. Horizontal dotted lines indicate the 90 mg/dL hypoglycemia prevention threshold (orange) and the 70 mg/dL hypoglycemia treatment threshold (red). Filled inverted triangles denote intravenous administration of 50% dextrose (D50) by push, the open inverted triangle denotes oral glucose tablet administration, and the purple shaded bar denotes the duration of a 10% dextrose (D10) intravenous infusion, each given as rescue therapy in response to symptomatic or protocol-defined hypoglycemia.

5.4.7 Paired intubation (Phase 1) versus non-intubation (Phase 2) comparison

Three volunteers (G1, G4, and G5), all in the osmotic pump arm, completed both the Phase 1 intubation period and the Phase 2 non-intubation period prior to the termination of Phase 2 enrollment, providing a within-subject methodological check on whether the presence of the intubation catheter perturbs systemic disposition of the analytes. Grouped concentration-time profiles for the modified-release analyte and the co-administered $^{13}C_6$-glipizide oral solution reference are shown by analyte and cohort in **Figure 5-23**, on linear and semi-logarithmic scales. The intubation cohort is represented by solid lines and filled circles and the no-intubation cohort by dashed lines and open squares; lines depict geometric means within each cohort and points depict individual observations.

The $^{13}C_6$-glipizide profiles for the intubation and no-intubation cohorts overlapped closely across the sampling window, with peak concentrations, time-to-peak, and terminal-phase slopes all visually concordant between cohorts on both linear and semi-logarithmic scales. This

concordance provides a positive methodological check that the systemic disposition of the oral reference analyte was not materially altered by the presence of the intubation catheter, supporting the use of the within-subject isotope reference as an unbiased disposition reference for deconvolution of the modified-release profile within the integrated methodology. The modified-release profiles, by contrast, did not overlap between cohorts. The geometric mean peak modified-release concentration was approximately 135 ng/mL in the intubation cohort and approximately 165 ng/mL in the no-intubation cohort (both peaks occurring at approximately 8 to 10 hours post-dose), and the no-intubation cohort geometric mean ran visibly above the intubation cohort geometric mean from the peak onward through the remainder of the sampling window on both linear and semi-logarithmic scales. Modified-release concentrations remained quantifiable in the no-intubation cohort at approximately 70 hours post-dose while intubation cohort concentrations approached the lower limit of quantification at the equivalent time. Because the $^{13}C_6$-glipizide reference rules out a catheter-induced change in systemic disposition, the modified-release divergence is most parsimoniously attributable either to a catheter-related perturbation of modified-release product behavior in the lumen (for example, by displacement of the dosage form, alteration of local hydrodynamics, or perturbation of gastrointestinal motility during transit) or to within-subject between-occasion variability in modified-release absorption. The n = 3 structure of the paired comparison does not permit these candidate explanations to be distinguished, and any quantitative interpretation of the modified-release divergence beyond its qualitative direction is not supported. The methodological implication for the integrated platform as deployed here is that the isotope-based disposition reference remains internally valid, while modified-release release and absorption may not be entirely catheter-invariant in the osmotic pump arm and warrants confirmation in a larger paired sample in any future deployment of the methodology.

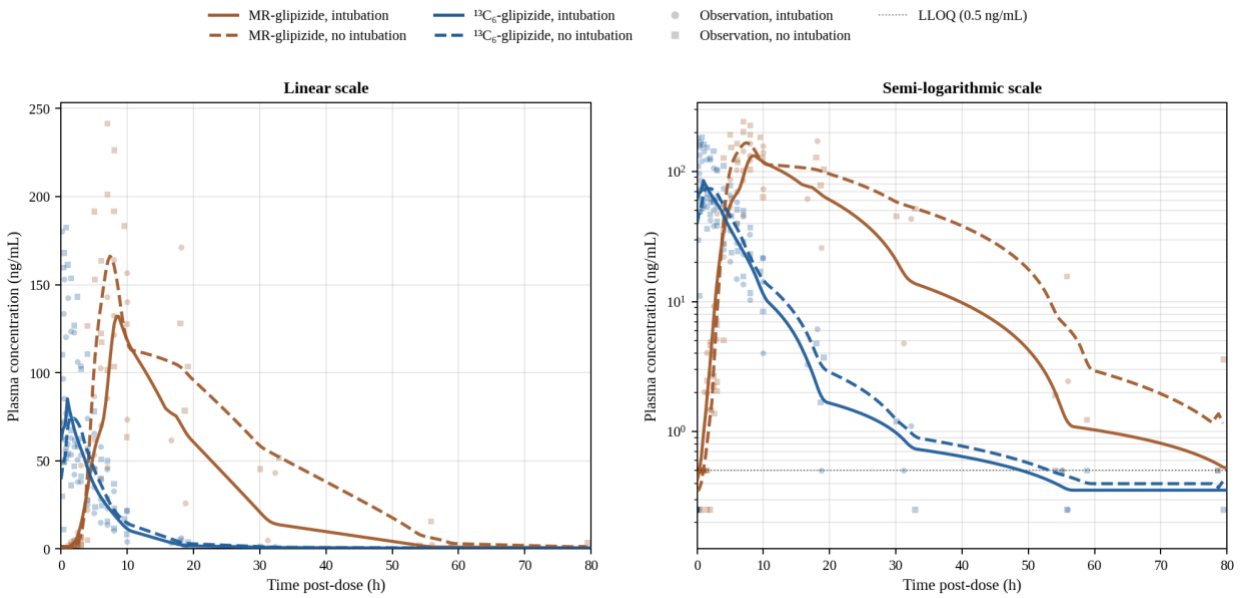


Figure 5-23. Plasma glipizide concentration-time profiles by analyte and intubation versus no-intubation cohorts, shown on linear (left) and semi-logarithmic (right) scales. MR-glipizide profiles are shown in orange and the co-administered $^{13}C_6$-glipizide oral solution reference profiles in blue; solid lines and circles correspond to the intubation cohort and dashed lines and squares correspond to the no-intubation cohort. Lines are geometric means and points are individual observations. The horizontal dotted line on the semi-logarithmic panel indicates the lower limit of quantification (0.5 ng/mL).

5.4.8 Population pharmacokinetic modeling and deconvolution of modified-release absorption

A two-compartment model with first-order absorption and proportional residual error (**Figure 5-24**) was developed against the isotope-only dataset of 296 plasma concentrations from all volunteers in the primary analysis population and converged successfully under the first-order conditional estimation method with interaction (FOCE-I). Final population parameter estimates, relative standard errors, interindividual variability variances, and η-shrinkage are reported in **Table 5-3**. Apparent clearance (CL/F = 2.50 L/h) and apparent central volume of distribution (V_2/F = 12.3 L) were estimated with relative standard errors near 17%; the absorption rate constant (ka = 3.54 h^{-1}) was estimated less precisely (relative standard error 33%) and exhibited substantial interindividual variability (IIV 139% CV), reflecting between-subject differences in early absorption behavior of the oral solution. The peripheral compartment was characterized by a small apparent peripheral volume (V_3/F = 1.93 L) and a low apparent intercompartmental clearance (Q/F = 0.26 L/h), implying slow redistribution between central and peripheral compartments and a peripheral compartment of modest mass relative to the central compartment.

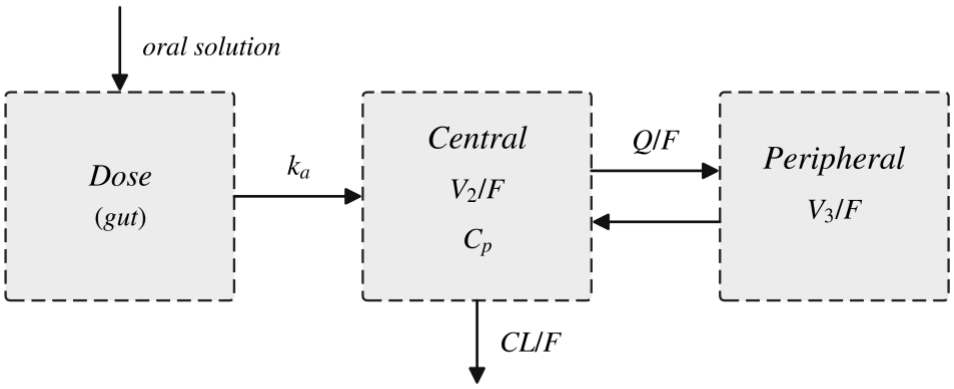


Figure 5-24. Structural diagram of the two-compartment population pharmacokinetic model with first-order absorption fitted to the $^{13}C_6$-glipizide oral solution plasma concentration data. k_a, first-order absorption rate constant; V_2/F, apparent central volume of distribution; V_3/F, apparent peripheral volume of distribution; Q/F, apparent intercompartmental clearance; CL/F, apparent total systemic clearance.

Table 5-3. Population pharmacokinetic parameter estimates for the two-compartment model with first-order absorption fitted to $^{13}C_6$-glipizide oral solution plasma concentration data from the primary analysis population using all oral solution data. Estimates are reported with their relative standard errors from the covariance step. Interindividual variability (IIV) is reported as percent coefficient of variation derived from the variance of the corresponding exponential random-effect term. Shrinkage (η-shrinkage) is reported as the percent reduction in the standard deviation of the empirical Bayes estimates relative to the population estimate.

Parameter	Estimate	Relative standard error (%)	IIV, % CV (shrinkage, %)
Structural model parameters			
Apparent clearance, *CL/F* (L/h)	2.50	17.2	45.0 (5.1)
Apparent central volume, *V_2/F* (L)	12.3	17.0	37.7 (7.3)
Absorption rate constant, *ka* (h^{-1})	3.54	33.4	139 (11.8)
Apparent intercompartmental clearance, *Q/F* (L/h)	0.26	29.9	—
Apparent peripheral volume, *V_3/F* (L)	1.93	21.1	—
Residual error			
Proportional error, σ^2 (variance)	0.21	70.8	—

Adequacy of the final model was assessed graphically and by visual predictive check. Goodness-of-fit diagnostics (**Figure 5-25**) showed conditional weighted residuals scattered symmetrically around zero across population-predicted concentrations (panel **B**), observations tracking individual predictions tightly along the line of identity throughout the concentration range (panel **C**), and population predictions showing the broader scatter expected from a model with substantial interindividual variability (panel **D**). The conditional-weighted-residuals-versus-

time display (panel **A**) showed a localized upward drift in the LOESS smoother beyond approximately 15 hours post-dose, indicating a modest underprediction of plasma concentrations in the late post-dose window. The visual predictive check (**Figure 5-26**) showed observed median and 5th/95th percentiles falling within the corresponding simulated 90% prediction intervals through approximately the first 15 hours; beyond approximately 15 to 20 hours, the observed median ran above the simulated median, visible on the semi-logarithmic panel. The pooled geometric mean overlay (**Figure 5-27**) indicated that the individual-predicted (IPRED) geometric mean tracked the observed geometric mean closely throughout the sampling window on both linear and semi-logarithmic scales, while the population-predicted (PRED) geometric mean diverged downward from the observed and IPRED curves beyond approximately 15 hours post-dose. The terminal-phase feature is therefore population-level rather than individual-level: subject-specific *post-hoc* predictions captured each subject's terminal phase, while the population-mean disposition function did not. Because the deconvolution analyses described below use individual *post-hoc* disposition parameters rather than the population mean, this terminal-phase population-level feature does not propagate into the per-subject absorption reconstruction; the underlying mechanism of the feature is interpreted in the Discussion (**5.4.4**).

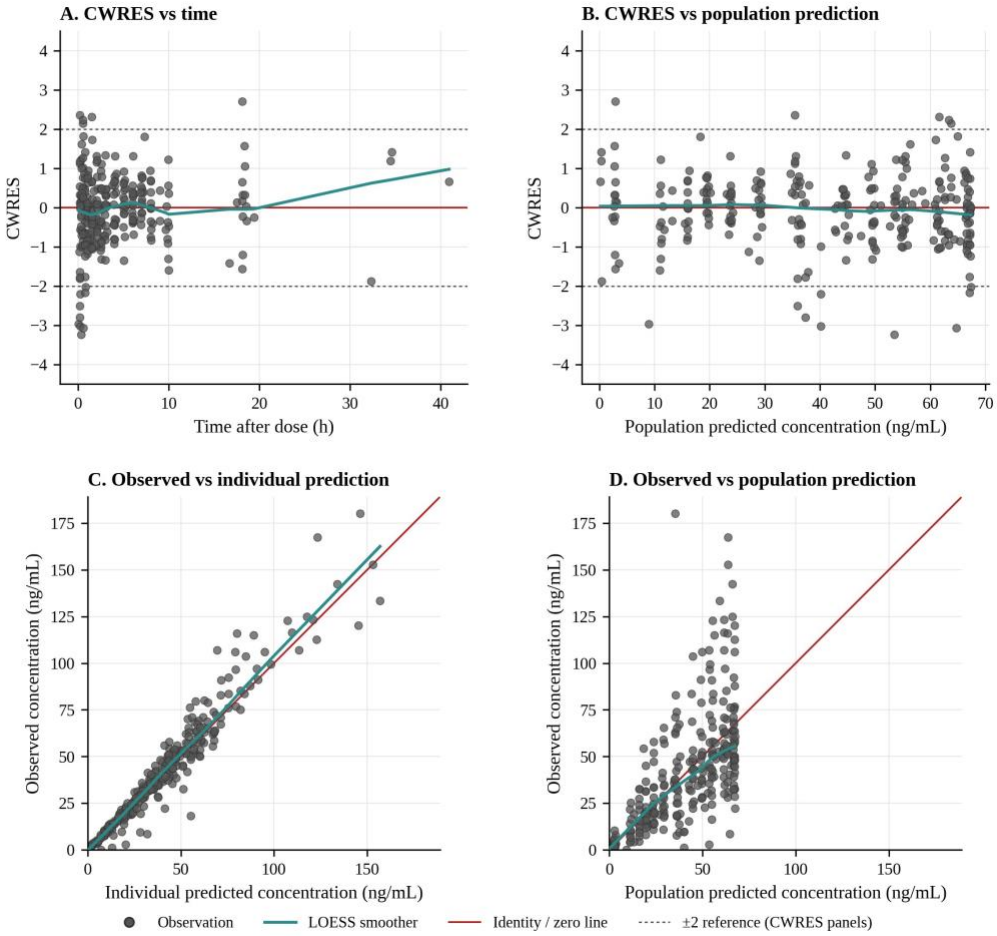


Figure 5-25. Goodness-of-fit diagnostics for the final two-compartment population pharmacokinetic model of the $^{13}C_6$-glipizide oral solution using all oral solution data. Panel **A**: conditional weighted residuals (CWRES) versus time after dose. Panel **B**: CWRES versus population-predicted concentration. Panel **C**: observed versus individual-predicted concentration. Panel **D**: observed versus population-predicted concentration. In panels A and B, the solid red horizontal line is at zero and dashed lines at ±2. In panels C and D, the solid red diagonal is the line of identity. A locally estimated scatterplot smoother (LOESS) is overlaid in teal in each panel.

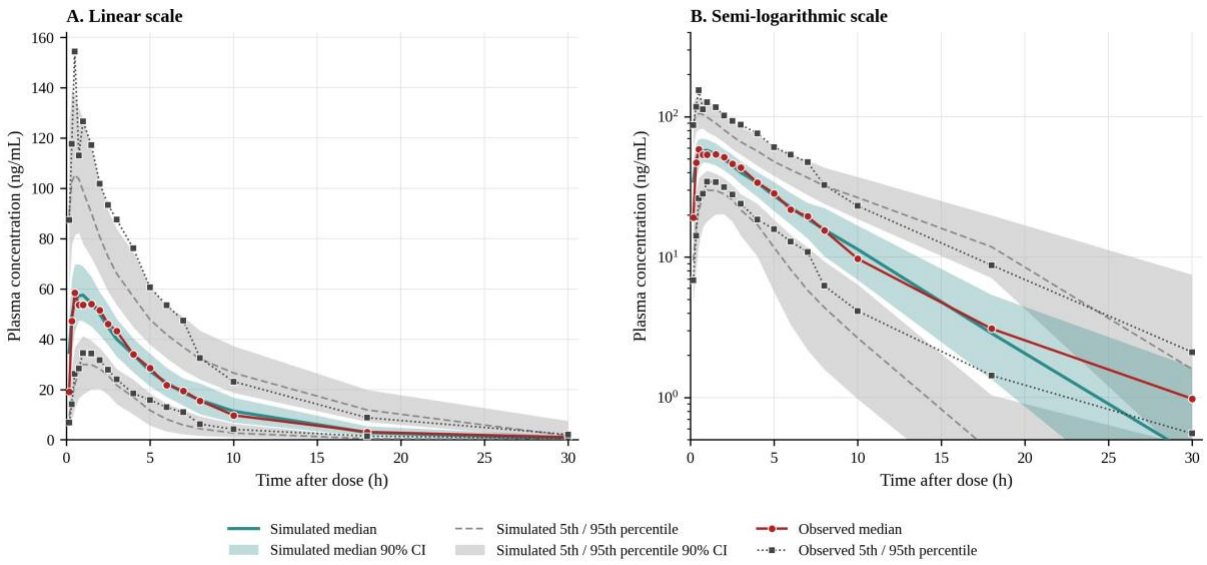


Figure 5-26. Visual predictive check for the final population pharmacokinetic model, based on 500 simulated replicates of the observed design. Observed plasma $^{13}C_6$-glipizide concentrations are summarized by the observed median (red circles) and 5th and 95th percentiles (black squares). Simulated percentiles are shown as the simulated median (teal solid line) and 5th/95th percentiles (gray dashed lines), with their associated 90% confidence intervals as shaded bands. Panels show linear (left) and semi-logarithmic (right) scales.

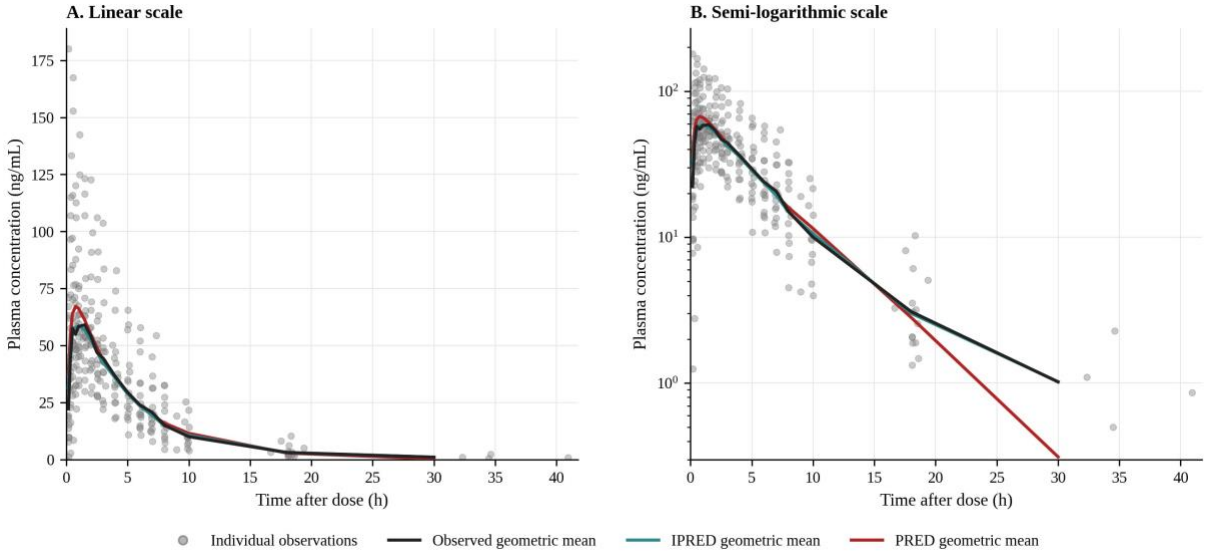


Figure 5-27. Pooled geometric mean plasma concentration-time overlay of observed, individual-predicted (IPRED), and population-predicted (PRED) concentrations for the $^{13}C_6$-glipizide oral solution, across the primary analysis population. Individual observations shown as translucent gray points. Panels show linear (left) and semi-logarithmic (right) scales.

Cumulative fraction absorbed over time for the modified-release dose was reconstructed on an individual-subject basis using the Loo-Riegelman method for drugs exhibiting two-compartment disposition[37,38], employing the subject-specific micro-rate constants k_{10}, k_{12}, and k_{21}

derived from each subject's *post-hoc* parameters of the population pharmacokinetic model of the co-administered isotope (**5.2.8**). Deconvolution was performed for all subjects in the primary analysis population with sufficient measurable modified-release concentrations to support reconstruction; one volunteer in the osmotic pump arm (G21) was excluded on that basis, leaving 14 subjects (6 osmotic pump, 8 hydrophilic matrix) in the cumulative-fraction-absorbed analysis. Aggregated cumulative-fraction-absorbed curves by formulation arm are shown in **Figure 5-28**; arm-mean curves rose from near zero to approximately 80% of the absorbed dose by 25 to 30 hours post-dose and approached an asymptote near 100% by approximately 60 to 70 hours, confirming recovery of the full absorbed dose over the sampling window and supporting the Loo-Riegelman reconstruction as a quantitatively complete characterization of the *in vivo* absorption process. The arm-level shapes were broadly similar across the two formulations, with the osmotic pump arm rising slightly faster through the 5- to 20-hour window than the hydrophilic matrix arm. Subject-level summary statistics from the cumulative-fraction-absorbed reconstructions were as follows: at the end of the seven-hour gastrointestinal sampling window, only a modest fraction of dose had been absorbed (median values approximately 25% for the osmotic pump arm and approximately 23% for the hydrophilic matrix arm); the median time to 50% cumulative absorption was approximately 18 hours in both arms; and the median time to 80% cumulative absorption was approximately 26 hours for the osmotic pump arm and approximately 34 hours for the hydrophilic matrix arm.

Parametric absorption profiles were additionally fitted to each subject's modified-release plasma concentration-time data by numerical convolution of a Weibull input function[39,40] with the individual two-compartment disposition function from the population pharmacokinetic model. Subjects with 10 or more measurable modified-release concentrations (n = 10) were fitted using a double-Weibull input function permitting biphasic release; subjects with 5 to 9 measurable concentrations (n = 2) were fitted using a single-Weibull input function; and three subjects with fewer than five measurable concentrations (G15, G16, and G17, all in the hydrophilic matrix arm) were not subjected to Weibull fitting and were retained in the Loo-Riegelman analysis only. Of the 12 Weibull fits attempted, 10 converged successfully; the two non-converged fits (G8, osmotic pump arm; G12, hydrophilic matrix arm) are excluded from the Weibull summary. Fit quality across converged subjects was variable, with coefficient-of-determination values ranging from approximately 0.13 to 0.98; aggregated Weibull fit by

formulation arm overlaid on observed concentration-time data are shown in **Figure 5-29**. Fitted Weibull absorption lag times had median values of approximately 1.1 hours (range 0 to 1.7 h) for the osmotic pump arm and approximately 2.7 hours (range 1.0 to 5.0 h) for the hydrophilic matrix arm, providing per-subject parametric summaries of the absorption-onset behavior reconstructed by the methodology.

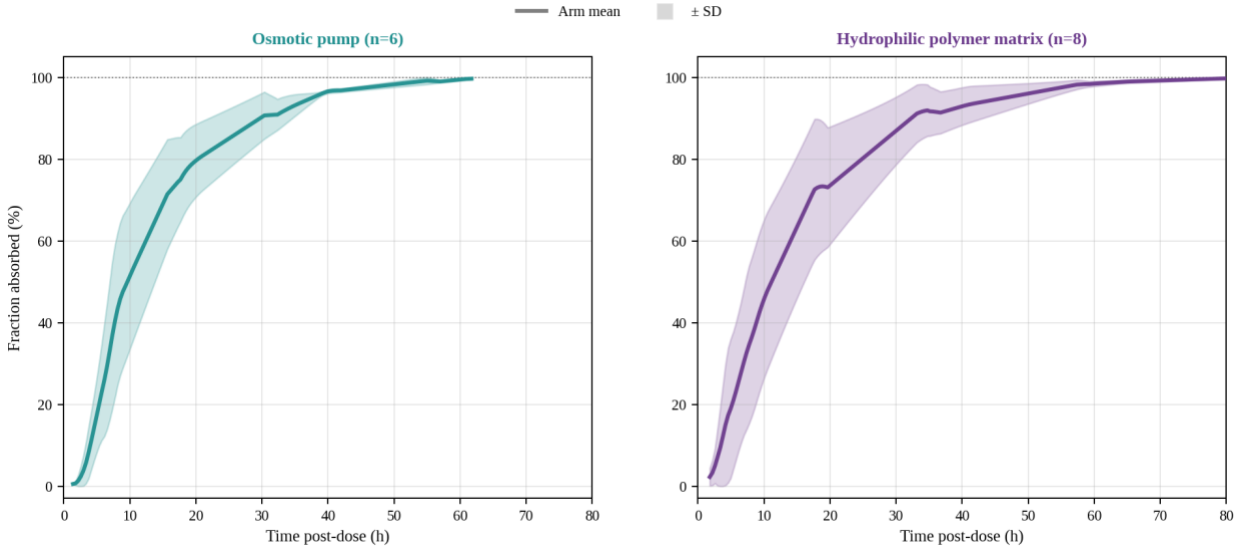


Figure 5-28. Cumulative fraction of the modified-release glipizide dose absorbed over time, by formulation arm, reconstructed on an individual-subject basis using the Loo-Riegelman method with subject-specific two-compartment disposition micro-rate constants from the population pharmacokinetic model of the isotope oral solution. Lines are arm means; shaded bands are ± 1 SD.

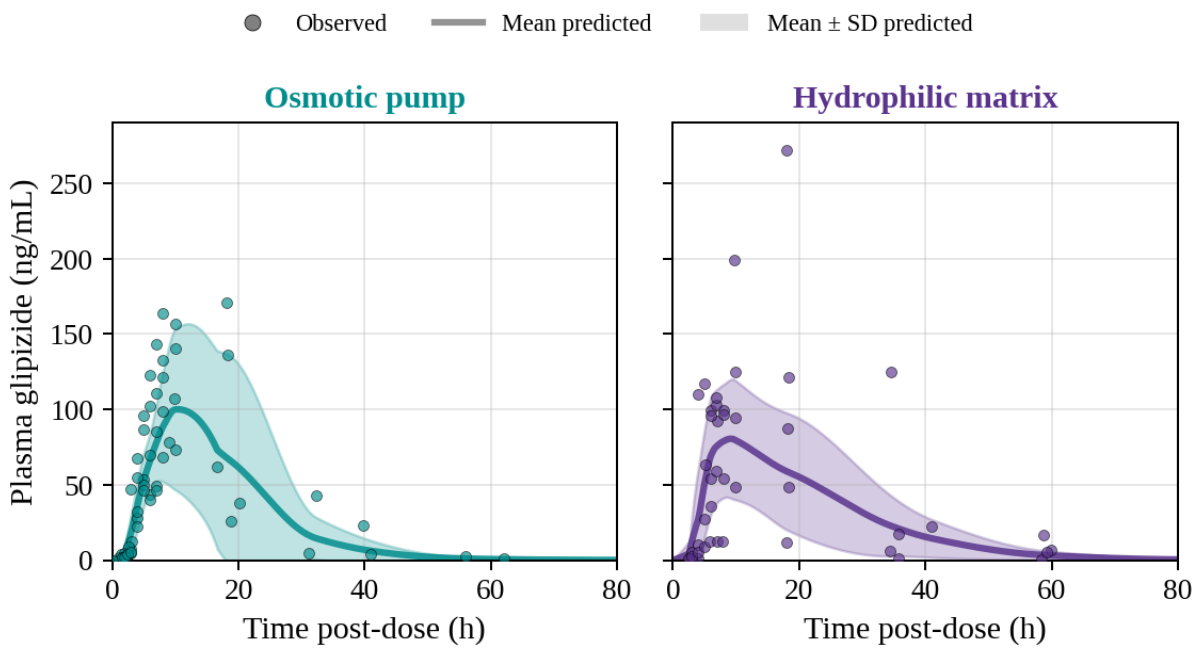


Figure 5-29. Observed plasma modified-release glipizide concentrations (filled circles) and fitted Weibull input function predictions across all converged subjects in each formulation arm. Panels show, for each arm, the individual observed measurements and the mean ± 1 standard deviation of the fitted predicted concentrations, reconstructed on a uniform time grid from each subject's post-hoc Weibull parameters and two-compartment disposition function.

5.5 Discussion

5.5.1 Principal findings

This study developed and demonstrated an integrated methodology for the *in vivo* biopharmaceutic characterization of oral modified-release products, combining regional-resolved oral multi-lumen gastrointestinal intubation with within-subject $^{13}C_6$-glipizide oral solution co-administration as a same-occasion systemic disposition reference. The methodology was exercised using two commercially available MR glipizide products of contrasting release mechanism (Product 1, osmotic pump; Product 2, hydrophilic polymer matrix), and produced four classes of *in vivo* observation per subject on a single dosing occasion: regional luminal drug concentrations, regional luminal environment (pH and bile acid composition), plasma pharmacokinetics of the MR product, and plasma pharmacokinetics of the same-occasion isotope disposition reference.

The two MR products produced different regional luminal release patterns: the osmotic pump product showed luminal drug concentrations confined predominantly to the stomach with only sparse appearance in the duodenum or jejunum, while the hydrophilic matrix product

showed luminal drug appearance distributed across all three regions, with the largest individual concentrations observed in the small bowel rather than in the stomach. The plasma pharmacokinetic profiles of the two products, by contrast, were closely similar and did not visibly distinguish between release mechanisms. The directly characterized luminal environment was that of a typical low gastric pH in most subjects, a near-neutral small-bowel pH, and small-bowel bile acid concentrations approximately 30- to 40-fold higher than gastric concentrations, supporting bile-acid-mediated solubilization of glipizide in the small bowel. Loo-Riegelman deconvolution of each subject's MR profile against their own same-occasion isotope reference indicated that only approximately one-quarter of the administered MR dose had been absorbed by the end of the seven-hour intubation sampling window in both arms, with the remainder absorbed over the following day; the temporal correspondence between regional luminal observations and the plasma-deconvolved absorption reconstruction supports the methodology as a coherent characterization of the *in vivo* release and absorption process.

A within-subject paired comparison of intubation and non-intubation phases in three osmotic-pump-arm subjects indicated that the systemic disposition of the oral isotope reference was not materially altered by the presence of the intubation catheter, supporting the use of the within-subject isotope reference as an unbiased disposition reference.

5.5.2 Release mechanism and regional disposition

For the osmotic pump product, luminal modified-release drug concentrations were dominated by the stomach over the seven-hour sampling window and were either at or near the lower limit of quantification, or only sparsely elevated above it, in the duodenum and jejunum. This stomach-dominated, small-bowel-sparing regional pattern is consistent with the expected mechanics of an osmotic pump device, in which release occurs by extrusion of a hydrated drug suspension through the delivery orifice at a rate set principally by membrane water permeability and the osmotic driving force rather than by local hydrodynamics or pH[11]. Considered against the regional luminal environment characterized in 5.3.4, the biopharmaceutic implication of stomach-dominated release for a glipizide product is that drug released into the lumen enters predominantly a low pH, low bile acid environment with limited solubilization capacity for the great majority of the dose, and only enters a near neutral, bile acid-supplemented environment to the extent that gastric emptying transports already-released drug into the small bowel. For the subset of subjects whose stomach pH fell into the higher pH cluster identified in **5.3.4**, this

proximal solubilization environment was substantially altered, although the regional drug appearance pattern itself remained stomach dominated.

For the hydrophilic polymer matrix product, the largest luminal concentrations were observed in the small bowel rather than in the stomach and appeared during the later part of the intubation sampling window. The duodenum showed a single high-leverage individual observation of approximately 940 ng/mL at approximately 2 hours post-dose with a scattered background of lower observations, and the jejunum showed two large late-window individual observations of approximately 2150 and 2330 ng/mL at 6 to 7 hours post-dose in two matrix-arm subjects, with a long lower-concentration tail across the rest of the population. The matrix regional pattern is consistent with the expected mechanics of a hydrophilic polymer matrix tablet, for which release proceeds by hydration, swelling, and erosion of the polymer matrix in contact with gastrointestinal fluid and is therefore expected to interact with the regional gastrointestinal environment encountered during transit[13]. Considered against the regional luminal environment characterized in **5.3.4**, the biopharmaceutic implication of distally distributed release is that drug released into the lumen enters predominantly the near-neutral, bile-acid-supplemented environment of the duodenum and jejunum, where both pH-driven ionization and bile acid-mediated solubilization support glipizide solubilization.

Considered alongside one another, the two regional patterns observed here illustrate a methodological point that bears directly on the assumptions underlying classical *in vitro-in vivo* correlation. The compendial dissolution apparatus conventionally used for modified-release characterization tests the formulation under a single, well-stirred, pH-controlled condition that does not reproduce the mechanical and chemical environment of either the proximal small bowel or the distal jejunum. For formulations whose release is well approximated by such conditions, classical correlation approaches based on summary exposure metrics may suffice. For formulations whose release exhibits the spatially and temporally heterogeneous behavior, the integrated luminal methodology supplies regional release information that conventional summary endpoints cannot capture, regardless of whether such products satisfy plasma-based bioequivalence criteria.

Despite these markedly different regional release patterns, the plasma concentration-time profiles of the two products were closely similar and did not visibly distinguish between mechanisms (**Figure 5-1**). This pairing, distinct regional luminal patterns alongside similar

plasma profiles, is the central biopharmaceutic observation of the study, and it has two related implications. First, the regional luminal information directly observed by the methodology is not recoverable from plasma pharmacokinetic data alone, even when both products are characterized on the same occasion within the same individuals. Second, two MR glipizide products with mechanistically distinct release behavior can produce closely similar plasma exposure profiles in healthy volunteers, an observation that is consistent with their regulatory bioequivalence status under conventional plasma-based criteria but that does not by itself imply equivalent regional *in vivo* release behavior. The methodology developed in this study makes the regional release behavior directly observable as a separate product attribute, available alongside plasma exposure for the biopharmaceutic evaluation of MR products.

5.5.3 Luminal environment: pH and bile acid composition

The regional luminal environment characterized by direct *ex vivo* and *in vitro* analysis of aspirated samples in this study, entailing typical low gastric pH in most subjects, near-neutral small bowel pH, and small bowel bile acid concentrations dominated by primary conjugated species at time-averaged geometric means of approximately 0.7 mM in the duodenum and 2.4 mM in the jejunum, shows the proximal-to-distal pattern reported by other investigators using oral multi-lumen intubation in fasted healthy adult volunteers[23,24,28], and provides directly observed *in vivo* values rather than assumed or simulated conditions for the interpretation of glipizide release behavior in the same subjects.

The directly observed small-bowel bile acid concentrations also bear comparison with the bile salt content of biorelevant dissolution media used to characterize oral solid dosage forms *in vitro*. Fasted-state simulated intestinal fluid (FaSSIF) and its updated version FaSSIF-V2 specify a sodium taurocholate concentration of 3 mM as the fasted-duodenum benchmark, a value broadly consistent with the weighted-mean total bile salt concentration of approximately 3.9 mM reported in a recent meta-analysis of human duodenal aspirate studies[41]. The total conjugated bile acid concentration observed in this study (approximately 0.9 mM in the duodenum and 3.3 mM in the jejunum, summing the primary and secondary conjugated classes) sits within the lower-to-mid portion of the published literature range. The duodenum value is roughly 3- to 4-fold below the FaSSIF and FaSSIF-V2 benchmark of 3 mM and roughly 4-fold below the fasted-duodenum weighted mean of 3.87 mM compiled in the recent meta-analysis[41]; the jejunum value is closely comparable to the FaSSIF benchmark and to the meta-analysis weighted mean for the proximal

jejunum (2.30 mM). Across the fasted-duodenum studies compiled by the meta-analysis, individual study values span approximately 0.04 to 36 mM with most study means clustering between 1 and 10 mM[42,43], and the values observed here fall in the lower portion of this reported range. The volunteers in the present study were fasted prior to dosing, with study drugs administered with a 240 mL glucose solution delivering approximately 46 g of dextrose (**5.2.1**, **5.2.3**). Although the dextrose load might have been expected to provide some stimulation of bile release, the duodenal bile acid concentrations observed here remained well below the fasted-state literature weighted mean of 3.87 mM and far below the fed-state weighted mean of 10.34 mM. Whether this reflects between-cohort biological variability, the timing of aspirate collection relative to interdigestive migrating motor complex cycling, dilution by the glucose vehicle, or methodological differences in aspirate handling cannot be discerned from the present data; the observed values are nonetheless within the published range for fasted human duodenal aspirate studies. Because glipizide is a BCS class II weakly acidic drug whose solubility is enhanced by bile-acid-mediated micellar solubilization, the use of FaSSIF or FaSSIF-V2 to characterize *in vitro* release of an MR glipizide product would test the formulation under bile salt conditions higher than those observed in the present cohort. Whether this difference is biopharmaceutically meaningful depends on whether the bile salt concentration is rate-limiting for *in vitro* dissolution at the formulation's release rate; for products whose release rate is mechanism-limited (such as the osmotic pump arm here), it likely is not, while for products whose release is environment-modulated (such as the matrix arm here), the difference between FaSSIF-level and observed-level bile salt concentration is a candidate source of *in vitro-in vivo* discrepancy. The methodology developed in this study contributes to this debate by providing per-subject *in vivo* bile acid measurements that can be used either to verify or to refine the bile salt content specified in biorelevant media for fasted-state characterization of oral MR products.

5.5.4 Systemic pharmacokinetics, population modeling, and deconvolution

Within the integrated methodology, the population pharmacokinetic step serves a defined role: to provide each subject's *post-hoc* systemic disposition parameters for use as the per-subject disposition function in the Loo-Riegelman deconvolution of that same subject's modified-release plasma profile, removing between-occasion and between-subject disposition variability from the reconstructed absorption profile. A two-compartment model with first-order absorption adequately described the isotope plasma profile across the primary analysis population, with

individual-predicted geometric means tracking observed geometric means closely throughout the sampling window on both linear and semi-logarithmic scales (**Figure 5-27**). The terminal-phase divergence between population-predicted and observed geometric means beyond approximately fifteen hours post-dose, visible in the goodness-of-fit and visual predictive check displays, is a population-level rather than individual-level feature: subject-specific *post-hoc* parameters captured each subject's terminal phase, while the population-mean disposition function did not. Because the deconvolution analyses use individual *post-hoc* disposition parameters rather than the population predictions, this terminal-phase population-level feature does not propagate into the per-subject absorption reconstruction. The high interindividual variability estimated on the absorption rate constant (139% CV) is consistent with substantial between-subject differences in early absorption behavior of the oral solution and reflects the contribution that within-subject co-administration of the isotope reference is intended to capture: by deconvolving each subject's modified-release profile against their own absorption-and-disposition reference, the methodology accommodates this variability subject-by-subject rather than averaging across it.

The Loo-Riegelman cumulative-fraction-absorbed reconstructions provide the plasma-based cross-validation that the integrated methodology requires. The reconstructions indicated that only approximately one-quarter of the administered modified-release dose had been absorbed by the end of the seven-hour intubation sampling window in both arms, with the remainder absorbed over the following day. The regional luminal data accordingly capture the early portion of the overall release and absorption process rather than its entirety, a quantitative methodological observation that bears on the design of future studies using the methodology, where extending the intubation sampling window or supplementing it with non-intubated sampling at later timepoints would capture release behavior corresponding to a larger fraction of the absorbed dose.

The Weibull-derived absorption lag times provide a further point of cross-section synthesis between the plasma-based and luminal observations. The longer median lag time observed for the matrix arm (approximately 2.7 hours, versus approximately 1.1 hours for the osmotic pump arm) is consistent with the mechanistic expectation that a hydrophilic matrix tablet requires an interval of polymer hydration before drug release begins, and is temporally consistent with the delayed appearance of luminal drug in the duodenum and jejunum ports of the matrix arm observed directly in the regional intubation data (**5.3.3**). The longer tail of the

matrix arm's cumulative-absorption curve (median time to 80% cumulative absorption of approximately 34 hours, versus approximately 26 hours for the osmotic pump arm) is temporally consistent with the late-window distal jejunum drug appearance observed in the regional luminal data of two matrix-arm subjects. The temporal correspondence between the regional luminal and plasma-deconvolved characterizations supports the methodology as an internally consistent characterization of the *in vivo* release and absorption process, while the primary analysis population, of which only a subset contributed to each region of the luminal sampling, limits the strength of any quantitative cross-section claims that can be drawn.

5.5.5 Implications and future directions

The methodology developed and demonstrated in this study has implications for *in vitro-in vivo* correlation, biopharmaceutic evaluation of MR products, and the regulatory frameworks that link product quality attributes to clinical performance. Classical IVIVC rests on the assumption that compendial dissolution testing under a single set of well-stirred, pH-controlled conditions captures the rate-limiting step of *in vivo* drug appearance[14-18]. For formulations whose release is well approximated by such conditions, including mechanism-limited release devices whose rate is set by formulation properties rather than by the surrounding environment, classical correlation approaches based on summary exposure metrics may suffice. For formulations whose release exhibits spatially and temporally heterogeneous behavior, however, the integrated luminal methodology supplies regional release information that conventional summary endpoints cannot capture, regardless of whether such products satisfy plasma-based bioequivalence criteria[19,20]. The two MR glipizide products studied here illustrate this distinction directly: their plasma profiles were closely similar (consistent with their regulatory bioequivalence status), while their regional luminal release patterns differed, information available only through direct regional sampling.

Within the broader PCQS framework, which links critical product quality attributes to clinical performance through mechanistic understanding of the *in vivo* release, dissolution, and absorption process[22], the methodology contributes a directly observed regional release attribute available alongside plasma exposure for the biopharmaceutic evaluation of MR products. The per-subject regional luminal environment characterization produced by the methodology, including the bile acid concentrations whose comparison with FaSSIF benchmarks was discussed in **5.4.3**, provides explicit *in vivo* covariate information that can be used either to anchor PBBM

models against directly observed regional values or to refine the composition of biorelevant media for the formulation class under study. The methodology is in principle extensible to other oral MR dosage forms of differing release mechanism, and to other active pharmaceutical ingredients for which regional gastrointestinal behavior is suspected of contributing to observed pharmacokinetic variability.

Several methodological extensions follow from the present work. Pairing the regional *in vivo* intubation approach with biorelevant multi-compartment dissolution apparatus[44,45] would permit direct comparison of *in vivo* and *in vitro* regional release profiles under matched conditions; combining intubation sampling with imaging-based assessment of tablet transit and integrity would clarify the relationship between regional luminal drug appearance and dosage-form location during transit; and extending the post-dose sampling window beyond the current seven hours, either by prolonging intubation in tolerant subjects or by supplementing intubation with non-intubated sampling at later timepoints, would capture release behavior corresponding to a larger fraction of the absorbed dose than the approximately one-quarter captured here. The within-subject paired comparison performed in three osmotic-pump-arm subjects (**5.3.7**) supports the catheter-invariance of the systemic disposition of the oral isotope reference but leaves open whether the catheter is similarly invariant for the modified-release tablet itself; a larger paired comparison, ideally including matrix-arm subjects, would address this open methodological question. Application of the methodology to a broader class of oral MR products, including those for which conventional IVIVC has historically been challenging, would enable regional *in vivo* release behavior to be treated as an explicit, directly observed product attribute within mechanism-aware regulatory and biopharmaceutic evaluation.

5.6 Conclusions

This study developed and demonstrated an integrated *in vivo* biopharmaceutic characterization methodology that combines region-resolved oral multi-lumen gastrointestinal intubation, within-subject stable-isotope ($^{13}C_6$-glipizide) oral solution co-administration, and population-pharmacokinetic-supported Loo-Riegelman deconvolution to characterize the regional release, regional luminal environment, and systemic pharmacokinetic behavior of oral modified-release products on a single dosing occasion per subject. Two commercially available MR glipizide products of contrasting release mechanism, an osmotic pump and a hydrophilic polymer matrix, were used as test products to evaluate the methodology. The two products

produced different regional luminal release patterns despite closely similar plasma pharmacokinetic profiles, illustrating the complementary information that direct regional measurement contributes beyond conventional plasma pharmacokinetic assessment of oral MR products. The methodology provides per-subject *in vivo* observations of release behavior, luminal environment, and systemic exposure that can serve as a directly observed input to mechanism-aware biopharmaceutic evaluation of oral MR products and is extensible in principle to other oral MR dosage forms and other active pharmaceutical ingredients for which regional gastrointestinal behavior contributes to pharmacokinetic variability.

5.7 References

1. Genitsaridi I, Salpea P, Salim A, et al. 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. *Lancet Diabetes Endocrinol*. Feb 2026;14(2):149–156. doi:10.1016/S2213-8587(25)00299-2
2. American Diabetes Association Professional Practice Committee for D. Introduction and Methodology: Standards of Care in Diabetes-2026. *Diabetes Care*. Jan 1 2026;49(Supplement_1):S1–S5. doi:10.2337/dc26-SINT
3. Lebovitz HE, Feinglos MN. Mechanism of action of the second-generation sulfonylurea glipizide. *Am J Med*. Nov 30 1983;75(5B):46–54. doi:10.1016/0002-9343(83)90253-x
4. Khan H, Zamir A, Imran I, et al. Clinical pharmacokinetics of glipizide: a systematic review. *Expert Opin Drug Metab Toxicol*. Jan 2025;21(1):69–79. doi:10.1080/17425255.2024.2402478
5. Talattof A, Price JC, Amidon GL. Gastrointestinal Motility Variation and Implications for Plasma Level Variation: Oral Drug Products. *Mol Pharm*. Feb 1 2016;13(2):557–67. doi:10.1021/acs.molpharmaceut.5b00774
6. Yang F, Liu L, Chen L, et al. OATP1B3 (699G>A) and CYP2C9*2, *3 significantly influenced the transport and metabolism of glibenclamide and glipizide. *Sci Rep*. Dec 24 2018;8(1):18063. doi:10.1038/s41598-018-36212-7
7. Jamzad S, Fassihi R. Development of a controlled release low dose class II drug-Glipizide. *Int J Pharm*. Apr 7 2006;312(1-2):24–32. doi:10.1016/j.ijpharm.2005.12.037
8. Berelowitz M, Fischette C, Cefalu W, Schade DS, Sutfin T, Kourides IA. Comparative efficacy of a once-daily controlled-release formulation of glipizide and immediate-release glipizide in patients with NIDDM. *Diabetes Care*. Dec 1994;17(12):1460–4. doi:10.2337/diacare.17.12.1460
9. Leaf E, King JO. Patient outcomes after formulary conversion from immediate-release to extended-release glipizide tablets. *Am J Health Syst Pharm*. Mar 1 1999;56(5):454–6. doi:10.1093/ajhp/56.5.454
10. Food and Drug Administration. *Draft Guidance on Glipizide Extended-Release Tablets*. 2008. https://www.accessdata.fda.gov/drugsatfda_docs/psg/Glipizide_ERtab_20329_RC7-08.pdf
11. Almoshari Y. Osmotic Pump Drug Delivery Systems-A Comprehensive Review. *Pharmaceuticals (Basel)*. Nov 18 2022;15(11)doi:10.3390/ph15111430

12. Banerjee A, Verma PR, Gore S. Controlled porosity solubility modulated osmotic pump tablets of gliclazide. *AAPS PharmSciTech*. Jun 2015;16(3):554–68. doi:10.1208/s12249-014-0246-0
13. Nokhodchi A, Raja S, Patel P, Asare-Addo K. The role of oral controlled release matrix tablets in drug delivery systems. *Bioimpacts*. 2012;2(4):175–87. doi:10.5681/bi.2012.027
14. Food and Drug Administration. *Extended Release Oral Dosage Forms: Development, Evaluation, and Application of In Vitro/In Vivo Correlations*. 1997. https://www.fda.gov/media/70939/download
15. Dickinson PA, Lee WW, Stott PW, et al. Clinical relevance of dissolution testing in quality by design. *AAPS J*. Jun 2008;10(2):380–90. doi:10.1208/s12248-008-9034-7
16. Heimbach T, Suarez-Sharp S, Kakhi M, et al. Dissolution and Translational Modeling Strategies Toward Establishing an In Vitro-In Vivo Link-a Workshop Summary Report. *AAPS J*. Feb 11 2019;21(2):29. doi:10.1208/s12248-019-0298-x
17. Nguyen MA, Flanagan T, Brewster M, et al. A survey on IVIVC/IVIVR development in the pharmaceutical industry - Past experience and current perspectives. *Eur J Pharm Sci*. May 1 2017;102:1–13. doi:10.1016/j.ejps.2017.02.029
18. Suarez-Sharp S, Li M, Duan J, Shah H, Seo P. Regulatory Experience with In Vivo In Vitro Correlations (IVIVC) in New Drug Applications. *AAPS J*. Nov 2016;18(6):1379–1390. doi:10.1208/s12248-016-9966-2
19. Gao Z, Ngo C, Ye W, et al. Effects of Dissolution Medium pH and Simulated Gastrointestinal Contraction on Drug Release From Nifedipine Extended-Release Tablets. *J Pharm Sci*. Mar 2019;108(3):1189–1194. doi:10.1016/j.xphs.2018.10.014
20. Wang K, Li Y, Chen B, et al. In Vitro Predictive Dissolution Test Should Be Developed and Recommended as a Bioequivalence Standard for the Immediate-Release Solid Oral Dosage Forms of the Highly Variable Mycophenolate Mofetil. *Mol Pharm*. Jul 4 2022;19(7):2048–2060. doi:10.1021/acs.molpharmaceut.1c00792
21. Abend A, Heimbach T, Cohen M, Kesisoglou F, Pepin X, Suarez-Sharp S. Dissolution and Translational Modeling Strategies Enabling Patient-Centric Drug Product Development: the M-CERSI Workshop Summary Report. *AAPS J*. Apr 9 2018;20(3):60. doi:10.1208/s12248-018-0213-x
22. Mackie C, Arora S, Seo P, et al. Physiologically Based Biopharmaceutics Modeling (PBBM): Best Practices for Drug Product Quality, Regulatory and Industry Perspectives: 2023 Workshop Summary Report. *Mol Pharm*. May 6 2024;21(5):2065–2080. doi:10.1021/acs.molpharmaceut.4c00202
23. Yu A, Baker JR, Fioritto AF, et al. Measurement of in vivo Gastrointestinal Release and Dissolution of Three Locally Acting Mesalamine Formulations in Regions of the Human Gastrointestinal Tract. *Mol Pharm*. Feb 6 2017;14(2):345–358. doi:10.1021/acs.molpharmaceut.6b00641
24. Koenigsknecht MJ, Baker JR, Wen B, et al. In Vivo Dissolution and Systemic Absorption of Immediate Release Ibuprofen in Human Gastrointestinal Tract under Fed and Fasted Conditions. *Mol Pharm*. Dec 4 2017;14(12):4295–4304. doi:10.1021/acs.molpharmaceut.7b00425
25. Paixao P, Bermejo M, Hens B, et al. Gastric emptying and intestinal appearance of nonabsorbable drugs phenol red and paromomycin in human subjects: A multi-compartment stomach approach. *Eur J Pharm Biopharm*. Aug 2018;129:162–174. doi:10.1016/j.ejpb.2018.05.033

26. Yu A, Jackson T, Tsume Y, et al. Mechanistic Fluid Transport Model to Estimate Gastrointestinal Fluid Volume and Its Dynamic Change Over Time. *AAPS J*. Nov 2017;19(6):1682–1690. doi:10.1208/s12248-017-0145-x
27. Mudie DM, Murray K, Hoad CL, et al. Quantification of gastrointestinal liquid volumes and distribution following a 240 mL dose of water in the fasted state. *Mol Pharm*. Sep 2 2014;11(9):3039–47. doi:10.1021/mp500210c
28. Hens B, Tsume Y, Bermejo M, et al. Low Buffer Capacity and Alternating Motility along the Human Gastrointestinal Tract: Implications for in Vivo Dissolution and Absorption of Ionizable Drugs. *Mol Pharm*. Dec 4 2017;14(12):4281–4294. doi:10.1021/acs.molpharmaceut.7b00426
29. Lei Y, Yan Y, Huang L, et al. Pharmacokinetics and Bioequivalence of Two Combination Metformin/Glipizide Tablets under Fasting and Fed Conditions in Chinese Healthy Subjects: A Randomized, Open-Label, Crossover Study. *Journal of Clinical Pharmacy and Therapeutics*. 2024;2024(1):9200636. doi:https://doi.org/10.1155/2024/9200636
30. Palenca I, Franzin SB, Pesce M, Sarnelli G, Esposito G. Rifaximin's therapeutic spectrum: approved indications and experimental insights into emerging uses. *Pharmacol Rep*. Feb 23 2026;doi:10.1007/s43440-026-00830-0
31. World Medical Association. *Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants*. 2024. https://www.wma.net/policies-post/wma-declaration-of-helsinki/
32. International Council for Harmonisation. *ICH Harmonised Guideline: Good Clinical Practice (GCP) E6(R3)*. 2025. https://database.ich.org/sites/default/files/ICH_E6(R3)_Step4_FinalGuideline_2025_0106.pdf
33. Bruni G, Ghione I, Berbenni V, et al. The Physico-Chemical Properties of Glipizide: New Findings. *Molecules*. May 24 2021;26(11)doi:10.3390/molecules26113142
34. Food and Drug Administration. *Bioanalytical Method Validation: Guidance for Industry*. 2018. https://www.fda.gov/files/drugs/published/Bioanalytical-Method-Validation-Guidance-for-Industry.pdf
35. Nguyen A, White JR. FreeStyle Libre 3. *Clin Diabetes*. Winter 2022;41(1):127–128. doi:10.2337/cd22-0102
36. Beal SL. Ways to fit a PK model with some data below the quantification limit. *J Pharmacokinet Pharmacodyn*. Oct 2001;28(5):481–504. doi:10.1023/a:1012299115260
37. Loo JC, Riegelman S. New method for calculating the intrinsic absorption rate of drugs. *J Pharm Sci*. Jun 1968;57(6):918–28. doi:10.1002/jps.2600570602
38. Yu A, Koenigsknecht MJ, Hens B, et al. Mechanistic Deconvolution of Oral Absorption Model with Dynamic Gastrointestinal Fluid to Predict Regional Rate and Extent of GI Drug Dissolution. *AAPS J*. Nov 11 2019;22(1):3. doi:10.1208/s12248-019-0385-z
39. Piotrovskii VK. The use of Weibull distribution to describe the in vivo absorption kinetics. *J Pharmacokinet Biopharm*. Dec 1987;15(6):681–6. doi:10.1007/BF01068420
40. Bressolle F, Gomeni R, Alric R, Royer-Morrot MJ, Necciari J. A double Weibull input function describes the complex absorption of sustained-release oral sodium valproate. *J Pharm Sci*. Oct 1994;83(10):1461–4. doi:10.1002/jps.2600831019
41. Santos LGA, Musther H, Bala N, et al. Gastrointestinal Bile Salt Concentrations in Healthy Adults Under Fasted and Fed Conditions: A Systematic Review and Meta-Analysis for Mechanistic Physiologically-Based Pharmacokinetic (PBPK) Modelling. *AAPS J*. Jan 22 2025;27(1):31. doi:10.1208/s12248-025-01016-x

42. Persson EM, Gustafsson AS, Carlsson AS, et al. The effects of food on the dissolution of poorly soluble drugs in human and in model small intestinal fluids. *Pharm Res*. Dec 2005;22(12):2141–51. doi:10.1007/s11095-005-8192-x
43. Bergstrom CA, Holm R, Jorgensen SA, et al. Early pharmaceutical profiling to predict oral drug absorption: current status and unmet needs. *Eur J Pharm Sci*. Jun 16 2014;57:173–99. doi:10.1016/j.ejps.2013.10.015
44. Stupak I, Pavlokova S, Vyslouzil J, Dohnal J, Culen M. Optimization of Dissolution Compartments in a Biorelevant Dissolution Apparatus Golem v2, Supported by Multivariate Analysis. *Molecules*. Nov 23 2017;22(12)doi:10.3390/molecules22122042
45. Katona MT, Nagy-Katona L, Szabo R, Borbas E, Tonka-Nagy P, Takacs-Novak K. Multi-Compartmental Dissolution Method, an Efficient Tool for the Development of Enhanced Bioavailability Formulations Containing Poorly Soluble Acidic Drugs. *Pharmaceutics*. Feb 24 2023;15(3)doi:10.3390/pharmaceutics15030753