Reinforcement Learning for Adaptive Tacrolimus Dosing with Multi-Drug Interaction Management

Anonymous Authors[1]

Abstract

Tacrolimus is a critical immunosuppressant following solid organ transplantation, but its narrow therapeutic window (4–15 ng/mL) and high inter-patient pharmacokinetic variability make dosing difficult, particularly when CYP3A4-inhibiting co-medications cause trough concentrations to rise. We formulate tacrolimus dosing with multi-drug drug-drug interaction (DDI) management as a Partially Observable Markov Decision Process (POMDP) and train an LSTM-augmented Proximal Policy Optimization (LSTM-PPO) model in a two-compartment physiologically-based pharmacokinetic (PBPK) simulator spanning four clinically relevant CYP3A4 inhibitors. The agent achieves 93.1% time in therapeutic window (TITW), a 13.9 percentage-point improvement over a proportional TDM controller. Against a memoryless ablation, the agent reduces toxic concentration events 8-fold and rejection risk six-fold, demonstrating that memory specifically prevents tail failures in DDI-exposed patients. Retrospective EHR-based validation on 6,394 real patients from the Cedars-Sinai OMOP cohort demonstrates a 5.8 percentage-point higher simulated TITW than observed EHR trough trajectories, consistent across all demographic and clinical subgroups. These results support memory-augmented RL as a promising framework for precision dosing from longitudinal, partially observed structured EHR data.

1. Introduction

Tacrolimus is a calcineurin inhibitor widely used in solid organ transplantation. By inhibiting calcineurin-mediated T-cell activation, tacrolimus prevents immune-mediated rejection of transplanted organs (Liu et al., 1991; noa, 1994). Its clinical utility is constrained by an extremely narrow therapeutic window. Typical trough targets of 4–15 ng/mL combined with high inter- and intra-patient pharmacokinetic (PK) variability (Henkel et al., 2023). Thus, small changes in systemic exposure can be clinically significant. Concentrations below the therapeutic window risks both acute and chronic graft rejection; concentrations above it cause nephrotoxicity, neurotoxicity, and increased susceptibility to opportunistic infections (Kaye et al., 2024; Larsson et al., 2026). Balancing these competing risks requires individualized, adaptive dosing.

The current standard of care relies on therapeutic drug monitoring (TDM), measuring pre-dose trough concentrations and adjusting doses empirically (Tolou-Ghamari & Palizban, 2025). Monitoring frequency ranges from daily during early post-transplant phase to quarterly during maintenance phase (Kasiske et al., 2000), making adjustments inherently reactive and lagged.

A particularly under-addressed challenge is the management of drug–drug interactions (DDIs). Tacrolimus is almost exclusively metabolized by CYP3A enzymes (principally CYP3A4 and CYP3A5) in the liver and intestinal wall (Van Gelder, 2002); co-administered drugs that inhibit these enzymes cause tacrolimus to accumulate, with trough increases ranging from 1.5 to 10-fold or greater (Jiang et al., 2025). These co-medications are usually common. Transplant patients are on tacrolimus for life which means DDI exposure is a lifelong risk. Common inhibitors include antifungals (fluconazole, voriconazole), calcium channel blockers (verapamil, diltiazem), and macrolide antibiotics (erythromycin) (Kriegl et al., 2022; Chow & Jusko, 2004a; Paterson & Singh, 1997). Current clinical guidance for DDI management consists largely of qualitative recommendations without systematic computational support or validation (Lempers et al., 2015).

Reinforcement learning (RL) offers a principled framework for sequential dosing under uncertainty, optimizing clinical objectives without requiring explicit specification of all patient parameters. With Long Short-Term Memory (LSTM), RL agents maintain memory of trough trajectories, enabling implicit inference of hidden pharmacokinetic state including

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

CYP3A4 inhibition level and patient-specific characteristics.

Previous work on RL for drug dosing has shown promise for anticoagulants (Anzabi Zadeh et al., 2023), vasopressors (Komorowski et al., 2018), and insulin dosing (Zhu et al., 2021). However, no prior work has addressed tacrolimus dosing with multi-drug DDI management, which presents unique challenges: pharmacokinetic disruption is dynamic rather than static, multiple inhibitor mechanisms, and the recovery/manifestation unfolds over days to weeks.

In this work, our main contributions are:

1. We formulate tacrolimus dose adjustment under CYP3A-mediated drug-drug interactions as a partially observable sequential decision problem, where patient-specific pharmacokinetics, enzyme activity, adherence, and DDI dynamics are latent and must be inferred from observed trough trajectories.
2. We develop a literature-calibrated PBPK simulator for tacrolimus dosing under multiple clinically relevant CYP3A4 inhibitors, and use it to train and evaluate recurrent reinforcement learning policies against rule-based, model-based, and memoryless baselines.
3. We conduct an EHR-informed retrospective evaluation on 6,394 real patients from the Cedars-Sinai OMOP cohort, across the major transplant types, demonstrating consistent improvement across demographic and clinical subgroups.

2. Methods

2.1. PBPK Simulator Environment

We developed a two-compartment physiologically-based pharmacokinetic (PBPK) simulator to serve as the training environment for the RL agent. The model tracks drug through gut, liver, and systemic compartments governed by well-established ordinary differential equations (Appendix Section A) (Pang & Rowland, 1977; Rowland et al., 1973). Drug-drug interactions are incorporated through time-varying suppression of intrinsic hepatic clearance (CL_{int}), capturing both reversible competitive inhibition and mechanism-based inactivation (MBI) (Silverman, 1995; Mayhew et al., 2000; Venkatakrishnan et al., 2007). Four clinically relevant CYP3A4 inhibitors were modeled spanning a range of inhibition mechanisms and magnitudes: fluconazole (He et al., 2020), verapamil (De Carlo et al., 2025), clarithromycin (Cheung & Senior, 2016) and erythromycin (Chiang et al., 2019). Drug-specific inhibition parameters were extracted and the models validated from published PBPK models and in vitro studies (Appendix Table 4).

2.2. Patient Variability

Each simulated episode samples a virtual patient from validated population distributions. CYP3A5 genotype (**1/*1*, **1/*3*, **3/*3*) is drawn from population frequencies and applies discrete clearance multipliers, accounting for 40–50% of total PK variability (Hesselink et al., 2003; Thölking et al., 2014). Residual inter-individual variability is captured by a log-normal clearance multiplier (CV 50%), with weight-scaled volume of distribution (Staatz & Tett, 2004). Stochastic episode-level variability includes bioavailability noise (V. et al., 2024), assay imprecision (Christians et al., 2015; Koster et al., 2013), food effects (Bekersky et al., 2001), and non-adherence (Prendergast & Gaston, 2010; Dew et al., 2007), collectively producing trough variability consistent with published transplant cohorts (Appendix Table 3 and Appendix A.4).

2.3. POMDP Formulation

We formulate the tacrolimus dosing problem as a Partially Observable Markov Decision Process (POMDP). The observation space consists of an 5-dimensional vector visible to the agent at each timestep: normalized trough concentration, DDI active flag, normalized episode day, and dose statistics (mean/SD over last 5 steps).

The hidden state includes DDI drug identity, drug-specific potency parameters (K_i, k_{inact}, K_I), patient pharmacokinetic parameters (CL, V_d), CYP3A5 genotype, missed doses, food effects, and true CYP3A4 enzyme activity. This is never observed by the agent and it has to be inferred. This partial observability faithfully mirrors clinical reality, where clinicians observe only trough concentrations and medication lists without access to underlying enzymatic state.

The action space consists of 7 discrete dose levels expressed as fractions of the 2.0 mg BID baseline dose: $\{0.0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50\}$, corresponding to actual doses of $\{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ mg twice daily. Episodes run for 45 days with one dosing decision per simulated day.

2.4. Reward Function

The reward function combines an immediate trough signal with a cumulative episode signal: $R_{\text{total}} = R_{\text{obs}} + 0.5 \times R_{\text{range}}$. R_{obs} rewards troughs within the therapeutic window [4–15 ng/mL] with a score scaled toward the clinical optimum (9.5 ng/mL), applies graded penalties in the subtoxic out-of-range zones [3–4 and 15–20 ng/mL], and a fixed penalty of -3.0 at or beyond the critical thresholds ($C \leq 3$ or $C \geq 20$ ng/mL), treating severe safety violations as categorical rather than graded errors. $R_{\text{range}} = 2 \times (\text{steps in range}/\text{total steps}) - 1$ incentivizes sustained

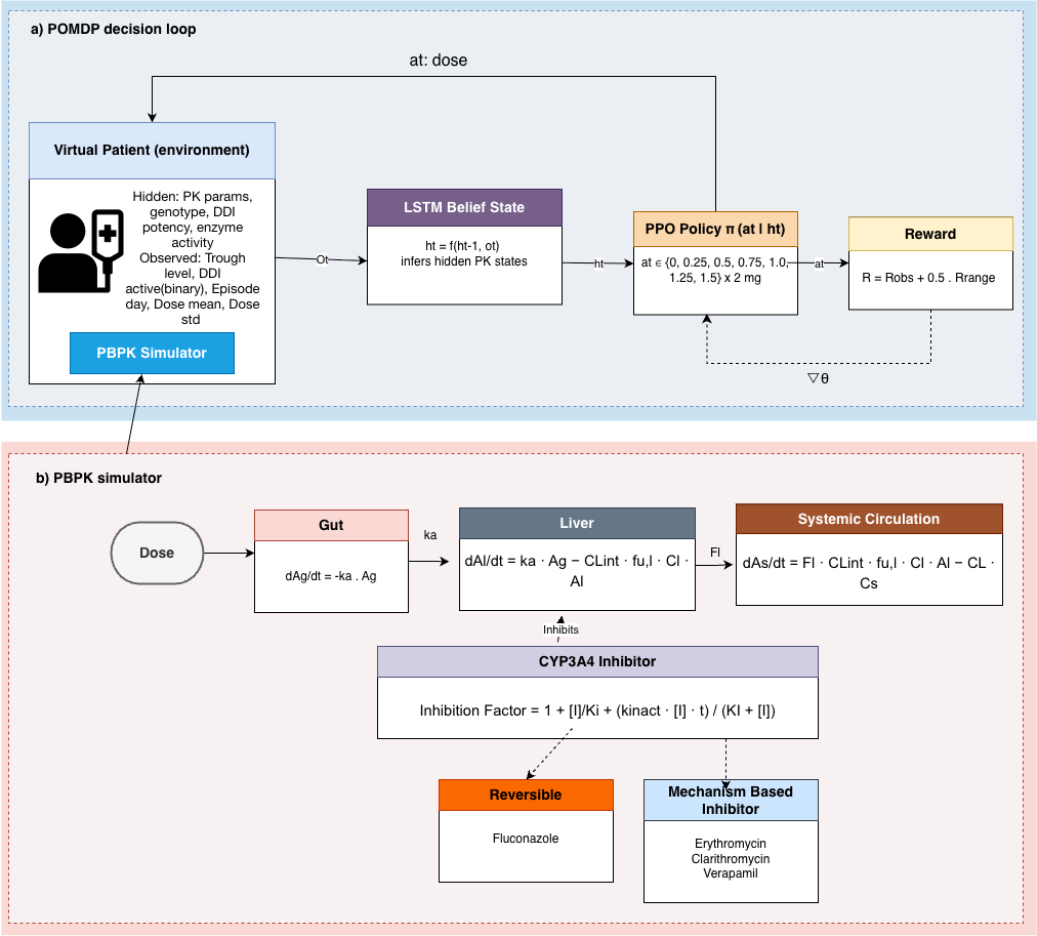

Figure 1. POMDP decision loop and PBPK simulator architecture.

therapeutic control across the episode rather than reactive recovery after excursions.

2.5. LSTM-PPO Architecture

We train a Proximal Policy Optimization (PPO) agent augmented with Long Short-Term Memory (LSTM), implemented via Stable-Baselines3-Contrib RecurrentPPO. The LSTM hidden state (h_t, c_t) functions as a learned belief state over unobserved pharmacokinetic parameters, implicitly inferring patient-specific characteristics and enzyme activity from trough trajectory. The actor outputs a probability distribution over 7 discrete dose levels; the critic estimates expected cumulative reward $V(s_t)$. Full details in Appendix Table 5.

2.6. Baseline Methods

We compare against five baseline methods. **Constant Dose** (2.0 mg BID uniformly), **Random** (uniform sampling), **Proportional Controller** ($K_p = 0.5$), **Model Predictive Control** (forward PBPK simulation with population parameters), and **PPO Vanilla (MLP)** (feedforward policy without LSTM). Full details in Appendix Section B.

2.7. Evaluation and EHR-Based Retrospective Analysis

Primary evaluation used 500,000 simulated episodes per method, reporting mean TITW, toxic events ($C \geq 20$ ng/mL), and rejection risk ($C \leq 3$ ng/mL) with 95% bootstrap CIs.

The EHR retrospective analysis used 6,394 tacrolimus patients from the Cedars-Sinai OMOP Clinical Data Warehouse (CDM v5.4; snapshot 2024-12-10; coverage 2015–2024), comprising 312,798 trough measurements over a median follow-up of 2.3 years. Inclusion criteria required $\geq$3 tacrolimus measurements, adult age ($\geq$18), and valid trough concentrations (0–100 ng/mL). Patients on concurrent cyclosporine were excluded. DDI exposure was defined as a concurrent prescription of a CYP3A4 inhibitor overlapping the measurement date by $\geq$3 days.

Table 1. Performance metrics over simulated evaluation episodes. TITW is the percentage of evaluated timepoints with trough concentration within the therapeutic window (4–15 ng/mL). Toxic Episodes and Rejection-risk Episodes report the percentage of episodes with at least one trough concentration $\geq$ 20 ng/mL or $\leq$ 3 ng/mL, respectively.

Method	TITW (%)	Toxic Episodes (%)	Rejection-risk Episodes (%)
LSTM-PPO	**93.1 (91.5–94.7)**	**0.06 (0.04–0.07)**	**2.51 (2.41–2.61)**
PPO Vanilla	92.6 (91.0–94.2)	0.47 (0.43–0.51)	15.32 (15.08–15.55)
Proportional	79.2 (76.7–81.7)	0.83 (0.77–0.89)	6.08 (5.92–6.23)
MPC	74.4 (71.7–77.1)	0.63 (0.58–0.68)	9.23 (9.04–9.42)
Const. Dose	58.1 (55.0–61.2)	9.93 (9.74–10.13)	14.60 (14.36–14.83)
Random	56.8 (53.7–59.9)	9.60 (9.41–9.79)	16.89 (16.65–17.13)

3. Results

3.1. Main Performance Comparison

Table 1 presents performance metrics for all methods. The LSTM-PPO agent achieves 93.1% TITW (95% CI: 91.5–94.7%), significantly outperforming all baseline methods.

MPC, despite having direct access to the PBPK model equations, achieves only 74.4% TITW, showing that accurate inference of patient-specific parameters is required, which the LSTM learns implicitly through trajectory observation.

Comparing LSTM-PPO against PPO Vanilla isolates the recurrent memory mechanism's contribution. The two methods achieve statistically indistinguishable TITW (93.1% vs. 92.6%), suggesting a memoryless policy can match *average* performance through an implicitly conservative dosing strategy. However, there is a six-fold difference in rejection risk (2.51% vs. 15.32%) and an eight-fold difference (0.06% vs. 0.47%) in toxic event rate. This implies that memory specifically prevents the tail failures in DDI or unusual PK phenotypes rather than improving average performance.

3.2. EHR-Informed Retrospective Evaluation

Retrospective validation on the Cedars-Sinai OMOP cohort included 6,394 tacrolimus patients with a median of 49 measurements per patient. The cohort comprised kidney (68%), liver (19%), heart (8%), and lung (5%) transplant recipients. Clinical TITW across the full cohort was 84.5% (CI: 83.9–85.2%). The RL agent achieved 90.3% , an improvement of 5.8 percentage points.

Table 2 presents retrospective validation results stratified by patient characteristics and DDI exposure. Improvement is consistent across all demographic strata. The most striking results occur where current practice struggles: patients with poor baseline control improve by 49.0pp, and high trough variability patients improve by 9.1pp (vs 2.3pp for low-variability). Across DDI conditions, the agent maintains >91% TITW. Strong mechanism-based inhibitors such as Erythromycin and Clarithromycin show largest percentage point improvements.

Table 2. EHR-based retrospective evaluation. *Top:* Subgroup analysis (OMOP, $n = 6{,}394$). *Bottom:* DDI stratification (sim). pp: percentage points.

Subgroup Analysis				
Subgroup	n	Clin.	RL	Δpp
Age <50y	2,024	86.1	90.1	+4.0
Age 50–65y	2,345	85.3	90.1	+4.7
Age ≥65y	2,025	82.0	90.6	+8.5
Male	3,919	84.5	90.0	+5.5
Female	2,475	84.5	90.6	+6.1
Poor (<50%)	173	38.3	87.4	+49.0
Mod. (50–75%)	898	67.7	90.9	+23.2
Good (>75%)	5,323	88.9	90.2	+1.4
Low var.	3,198	87.2	89.5	+2.3
High var.	3,196	81.8	91.0	+9.1
DDI Stratification				
DDI Drug	n	Clin.	RL	Δpp
Verapamil	1,912	85.0	91.8	+6.8
Fluconazole	282	86.7	91.1	+4.4
Erythromycin	164	81.6	91.3	+9.7
Clarithromycin	61	81.8	93.8	+12.0

4. Discussion and Conclusion

We developed a memory-augmented RL framework for tacrolimus dose adjustment under CYP3A-mediated drug–drug interactions and evaluated it in simulation and an EHR-based retrospective analysis of 6,394 tacrolimus-treated patients. The policy achieved 5.8 percentage-point higher simulated TITW than observed EHR trough trajectories, with the largest gains in patients with poor baseline control and high trough variability.

In simulation, the agent achieves 93.1% TITW with eight-fold toxic event reduction and six-fold rejection risk reduction. The LSTM maintains memory of trough trajectories, enabling inference of hidden pharmacokinetic state and anticipation of delayed DDI dynamics. Memoryless policies achieve similar average TITW through conservative dosing but fail catastrophically for DDI/unusual PK patients.

Limitations include reliance on a simulated training environment that necessarily simplifies real patient complexity, an incomplete DDI drug set (ritonavir and antiretroviral combinations are notable omissions), and the absence of observed administered doses in the OMOP data, precluding direct comparison against clinical decisions.

Critical next steps include prospective clinical validation and eventually integrating into clinical workflows.

References

A Comparison of Tacrolimus (FK 506) and Cyclosporine for Immunosuppression in Liver Transplanta-

tion. *New England Journal of Medicine*, 331(17): 1110–1115, October 1994. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJM199410273311702. URL http://www.nejm.org/doi/abs/10.1056/NEJM199410273311702.

Anzabi Zadeh, S., Street, W. N., and Thomas, B. W. Optimizing warfarin dosing using deep reinforcement learning. *Journal of Biomedical Informatics*, 137:104267, January 2023. ISSN 1532-0480. doi: 10.1016/j.jbi.2022.104267.

Arroyo-Currás, N., Ortega, G., Copp, D. A., et al. High-precision control of plasma drug levels using feedback-controlled dosing. *ACS Pharmacology & Translational Science*, 1(2):110–118, 2018. doi: 10.1021/acsptsci.8b00033.

Bekersky, I., Dressler, D., and Mekki, Q. A. Effect of low- and high-fat meals on tacrolimus absorption following 5 mg single oral doses to healthy human subjects. *Journal of Clinical Pharmacology*, 41(2):176–182, February 2001. ISSN 0091-2700. doi: 10.1177/00912700122009999.

Cheung, K. K.-T. and Senior, P. A. Tacrolimus toxicity in islet transplantation due to interaction with macrolides. *Clinical Diabetes and Endocrinology*, 2:2, 2016. doi: 10.1186/s40842-016-0019-7.

Chiang, L.-H., Wu, T.-H., Tsai, T.-C., and Lee, W.-C. Coadministration of erythromycin to increase tacrolimus concentrations in liver transplant recipients. *Transplantation Proceedings*, 51(5):1439–1441, 2019. doi: 10.1016/j.transproceed.2019.01.145.

Chow, F.-S. and Jusko, W. J. Immunosuppressive interactions among calcium channel antagonists and selected corticosteroids and macrolides using human whole blood lymphocytes. *Drug Metabolism and Pharmacokinetics*, 19(6):413–421, 2004a. doi: 10.2133/dmpk.19.413.

Chow, F.-S. and Jusko, W. J. Immunosuppressive interactions among calcium channel antagonists and selected corticosteroids and macrolides using human whole blood lymphocytes. *Drug Metabolism and Pharmacokinetics*, 19(6):413–421, December 2004b. ISSN 1347-4367. doi: 10.2133/dmpk.19.413.

Christians, U., Vinks, A. A., Langman, L. J., Clarke, W., Wallemacq, P., van Gelder, T., Renjen, V., Marquet, P., and Meyer, E. J. Impact of Laboratory Practices on Interlaboratory Variability in Therapeutic Drug Monitoring of Immunosuppressive Drugs. *Therapeutic Drug Monitoring*, 37(6):718–724, December 2015. ISSN 1536-3694. doi: 10.1097/FTD.0000000000000205.

De Carlo, A., Tosca, E. M., and Magni, P. Precision Dosing in Presence of Multiobjective Therapies by Integrating Reinforcement Learning and PK-PD Models: Application to Givinostat Treatment of Polycythemia Vera. *CPT: pharmacometrics & systems pharmacology*, 14(6):1018–1031, June 2025. ISSN 2163-8306. doi: 10.1002/psp4.70012.

Dew, M. A., DiMartini, A. F., De Vito Dabbs, A., Myaskovsky, L., Steel, J., Unruh, M., Switzer, G. E., Zomak, R., Kormos, R. L., and Greenhouse, J. B. Rates and Risk Factors for Nonadherence to the Medical Regimen After Adult Solid Organ Transplantation. *Transplantation*, 83(7):858–873, April 2007. ISSN 0041-1337. doi: 10.1097/01.tp.0000258599.65257.a6. URL https://journals.lww.com/00007890-200704150-00005.

He, J., Yu, Y., Yin, C., Liu, H., Zou, H., Ma, J., Yang, W., Liu, Y., Zhong, L., and Chen, X. Clinically significant drug-drug interaction between tacrolimus and fluconazole in stable renal transplant recipient and literature review. *Journal of Clinical Pharmacy and Therapeutics*, 45(2):264–269, April 2020. ISSN 1365-2710. doi: 10.1111/jcpt.13075.

Henkel, L., Jehn, U., Thölking, G., and Reuter, S. Tacrolimus-why pharmacokinetics matter in the clinic. *Frontiers in Transplantation*, 2:1160752, 2023. ISSN 2813-2440. doi: 10.3389/frtra.2023.1160752.

Hesselink, D. A., van Schaik, R. H. N., van der Heiden, I. P., van der Werf, M., Gregoor, P. J. H. S., Lindemans, J., Weimar, W., and van Gelder, T. Genetic polymorphisms of the CYP3A4, CYP3A5, and MDR-1 genes and pharmacokinetics of the calcineurin inhibitors cyclosporine and tacrolimus. *Clinical Pharmacology and Therapeutics*, 74(3):245–254, September 2003. ISSN 0009-9236. doi: 10.1016/S0009-9236(03)00168-1.

Hovorka, R., Canonico, V., Chassin, L. J., et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological Measurement*, 25(4):905–920, 2004. doi: 10.1088/0967-3334/25/4/010.

Jiang, Y., Wang, L., Liu, J., Jiang, L., Li, K., Lv, Z., Zhou, S., and Shao, F. Physiologically Based Pharmacokinetic Modeling to Evaluate Drug-Drug Interactions of Tacrolimus With Ritonavir, a CYP3A Irreversible Inhibitor: Applications for Dosing Optimization in Transplant Patients. *Clinical Pharmacology in Drug Development*, 14(10):797–808, October 2025. ISSN 2160-7648. doi: 10.1002/cpdd.1572.

Kasiske, B. L., Vazquez, M. A., Harmon, W. E., et al. Recommendations for the outpatient surveillance of renal

transplant recipients. *Journal of the American Society of Nephrology*, 11(Suppl 15):S1–S86, 2000.

Kaye, A. D., Shah, S. S., Johnson, C. D., et al. Tacrolimus- and mycophenolate-mediated toxicity: Clinical considerations and options in management of post-transplant patients. *Current Issues in Molecular Biology*, 47(1):2, 2024. doi: 10.3390/cimb47010002.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, November 2018. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-018-0213-5. URL https://www.nature.com/articles/s41591-018-0213-5.

Koster, R. A., Alffenaar, J.-W. C., Greijdanus, B., and Uges, D. R. A. Fast LC-MS/MS analysis of tacrolimus, sirolimus, everolimus and cyclosporin A in dried blood spots and the influence of the hematocrit and immunosuppressant concentration on recovery. *Talanta*, 115:47–54, October 2013. ISSN 1873-3573. doi: 10.1016/j.talanta.2013.04.027.

Kriegl, L., Boyer, J., Egger, M., and Hoenigl, M. Antifungal stewardship in solid organ transplantation. *Transplant Infectious Disease*, 24:e13855, 2022. doi: 10.1111/tid.13855.

Larsson, A., Saldeen, J., Cedernaes, J., et al. Decreasing tacrolimus concentrations in routine therapeutic drug monitoring data indicate adherence to updated therapeutic goals. *Biomedicines*, 14(1):94, 2026. doi: 10.3390/biomedicines14010094.

Lempers, V. J. C., Martial, L. C., Schreuder, M. F., et al. Drug-interactions of azole antifungals with selected immunosuppressants in transplant patients: strategies for optimal management in clinical practice. *Current Opinion in Pharmacology*, 24:38–44, 2015. doi: 10.1016/j.coph.2015.07.002.

Liu, J., Farmer, J. D., Lane, W. S., Friedman, J., Weissman, I., and Schreiber, S. L. Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes. *Cell*, 66(4):807–815, August 1991. ISSN 0092-8674. doi: 10.1016/0092-8674(91)90124-h.

Mayhew, B. S., Jones, D. R., and Hall, S. D. An in vitro model for predicting in vivo inhibition of cytochrome P450 3A4 by metabolic intermediate complex formation. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, 28(9):1031–1037, September 2000. ISSN 0090-9556.

Pang, K. S. and Rowland, M. Hepatic clearance of drugs. I. Theoretical considerations of a "well-stirred" model and a "parallel tube" model. Influence of hepatic blood flow, plasma and blood cell binding, and the hepatocellular enzymatic activity on hepatic drug clearance. *Journal of Pharmacokinetics and Biopharmaceutics*, 5(6):625–653, December 1977. ISSN 0090-466X. doi: 10.1007/BF01059688. URL http://link.springer.com/10.1007/BF01059688.

Paterson, D. L. and Singh, N. Interactions between tacrolimus and antimicrobial agents. *Clinical Infectious Diseases*, 25(6):1430–1440, 1997. doi: 10.1086/516138.

Prendergast, M. B. and Gaston, R. S. Optimizing medication adherence: an ongoing opportunity to improve outcomes after kidney transplantation. *Clinical journal of the American Society of Nephrology: CJASN*, 5(7):1305–1311, July 2010. ISSN 1555-905X. doi: 10.2215/CJN.07241009.

Rowland, M., Benet, L. Z., and Graham, G. G. Clearance concepts in pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics*, 1(2):123–136, April 1973. ISSN 0090-466X. doi: 10.1007/BF01059626.

Sheppard, L. C. Computer control of the infusion of vasoactive drugs. *Annals of Biomedical Engineering*, 8(5–6):431–434, 1980. doi: 10.1007/BF02363444.

Silverman, R. B. [10] Mechanism-based enzyme inactivators. In *Methods in Enzymology*, volume 249, pp. 240–283. Elsevier, 1995. ISBN 978-0-12-182150-0. doi: 10.1016/0076-6879(95)49038-8. URL https://linkinghub.elsevier.com/retrieve/pii/0076687995490388.

Staatz, C. E. and Tett, S. E. Clinical Pharmacokinetics and Pharmacodynamics of Tacrolimus in Solid Organ Transplantation:. *Clinical Pharmacokinetics*, 43(10):623–653, 2004. ISSN 0312-5963. doi: 10.2165/00003088-200443100-00001. URL http://link.springer.com/10.2165/00003088-200443100-00001.

Thölking, G., Fortmann, C., Koch, R., Gerth, H. U., Pabst, D., Pavenstädt, H., Kabar, I., Hüsing, A., Wolters, H., Reuter, S., and Suwelack, B. The tacrolimus metabolism rate influences renal function after kidney transplantation. *PloS One*, 9(10):e111128, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0111128.

Tolou-Ghamari, Z. and Palizban, A.-A. Tacrolimus pharmacotherapy: Infectious complications and toxicity in organ transplant recipients; an updated review. *Current Drug Research Reviews*, 17:301–310, 2025. doi: 10.2174/0125899775259326231212073240.

V., S., M., R., M. S., S., and S., P. Understanding the factors influencing pharmacokinetics of tacrolimus. *International Journal of Research in Medical Sciences*, 12(5):1769–1775, April 2024. ISSN 2320-6012, 2320-6071. doi: 10.18203/2320-6012.ijrms20241273. URL https://www.msjonline.org/index.php/ijrms/article/view/13384.

Van Gelder, T. Drug Interactions with Tacrolimus:. *Drug Safety*, 25(10):707–712, 2002. ISSN 0114-5916. doi: 10.2165/00002018-200225100-00003. URL http://link.springer.com/10.2165/00002018-200225100-00003.

Venkatakrishnan, K., Obach, R. S., and Rostami-Hodjegan, A. Mechanism-based inactivation of human cytochrome P450 enzymes: strategies for diagnosis and drug–drug interaction risk assessment. *Xenobiotica*, 37(10-11): 1225–1256, November 2007. ISSN 0049-8254, 1366-5928. doi: 10.1080/00498250701670945. URL http://www.tandfonline.com/doi/full/10.1080/00498250701670945.

Zhu, T., Li, K., Herrero, P., and Georgiou, P. Basal Glucose Control in Type 1 Diabetes Using Deep Reinforcement Learning: An *In Silico* Validation. *IEEE Journal of Biomedical and Health Informatics*, 25(4): 1223–1232, April 2021. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2020.3014556. URL https://ieeexplore.ieee.org/document/9159862/.

A. PBPK Simulator Equations

A.1. Oral Absorption and Systemic Tacrolimus Model

The simulator implements a simplified PBPK-inspired oral pharmacokinetic model with first-order absorption and time-varying CYP3A-mediated clearance. It is designed to capture clinically relevant tacrolimus exposure and CYP3A-mediated DDI dynamics rather than to represent a full whole-body PBPK model.

A.1.1. Gut Absorption

The gut compartment models oral absorption with first-order kinetics:

$$\frac{dA_g}{dt} = -k_a A_g, \tag{1}$$

where A_g is the amount of tacrolimus in the gut (mg) and k_a is the absorption rate constant (h^{-1}). We set $k_a = 0.8\ \text{h}^{-1}$ based on published tacrolimus pharmacokinetic estimates (Staatz & Tett, 2004).

A.1.2. Systemic Compartment

The systemic compartment receives the absorbed fraction of the oral dose and eliminates tacrolimus through time-varying apparent clearance:

$$\frac{dA_s}{dt} = F \cdot k_a A_g - \text{CL}(t) \cdot C_s, \tag{2}$$

where A_s is the amount of tacrolimus in systemic circulation (mg), F is oral bioavailability, $\text{CL}(t)$ is time-varying apparent clearance (L/h), and

$$C_s = \frac{A_s}{V_d} \tag{3}$$

is the systemic concentration in mg/L. Reported tacrolimus troughs are converted to ng/mL as:

$$C_{\text{trough}} = 1000 \cdot \frac{A_s}{V_d}. \tag{4}$$

A.2. Drug–Drug Interaction Modeling

CYP3A4 inhibitors reduce tacrolimus clearance by decreasing effective CYP3A enzyme activity. We model this through time-varying apparent clearance:

$$\text{CL}(t) = \text{CL}_0 \cdot M_{\text{CYP3A5}} \cdot M_{\text{IIV}} \cdot \frac{E(t)}{1 + [I](t)/K_i}, \tag{5}$$

where CL_0 is baseline clearance, M_{CYP3A5} is the CYP3A5 genotype-specific clearance multiplier, M_{IIV} is a log-normal inter-individual variability multiplier, $[I](t)$ is the inhibitor concentration at the enzyme site, K_i is the reversible inhibition constant, and $E(t)$ is the fraction of active CYP3A enzyme remaining.

For reversible inhibitors such as fluconazole, enzyme activity is assumed to recover immediately after inhibitor discontinuation:

$$E(t) = 1. \tag{6}$$

For mechanism-based inhibitors, enzyme activity follows:

$$\frac{dE}{dt} = k_{\text{deg}} \cdot (1 - E(t)) - \frac{k_{\text{inact}}[I](t)}{K_I + [I](t)} \cdot E(t), \tag{7}$$

where k_{inact} is the maximal enzyme inactivation rate, K_I is the inhibitor concentration producing half-maximal inactivation, and $k_{\text{deg}} = \ln(2)/t_{1/2,\text{enzyme}}$ controls enzyme turnover. We use $t_{1/2,\text{enzyme}} \approx 36$ hours for CYP3A4. This formulation captures delayed recovery after mechanism-based inhibition because enzyme activity returns only through de novo enzyme synthesis.

Table 3. PBPK Model Parameters and Variability.

Parameter	Symbol	Distribution	Reference
Absorption rate	k_a	0.8 h^{-1} (fixed)	(Staatz & Tett, 2004)
Volume of distribution	V_d	1.0 × Weight L/kg	(Staatz & Tett, 2004)
Baseline clearance	$CL_{int,0}$	LogNormal(22, CV=50%) L/h	(Staatz & Tett, 2004)
CYP3A5 *1/*1	CL multiplier	1.40×	(Hesselink et al., 2003)
CYP3A5 *1/*3	CL multiplier	1.21×	(Hesselink et al., 2003)
CYP3A5 *3/*3	CL multiplier	1.00× (reference)	(Hesselink et al., 2003)
Bioavailability	F	0.25 ± 0.10	(Bekersky et al., 2001)
Unbound fraction	f_u	0.01 (fixed)	(Staatz & Tett, 2004)

Table 4. CYP3A4 Inhibitor Pharmacodynamic Parameters.

Inhibitor	Mechanism	K_i (μM)	k_{inact} (h^{-1})	K_I (μM)
Fluconazole	Reversible	12.5	—	—
Verapamil	Weak MBI	8.2	0.05	15.0
Erythromycin	Strong MBI	3.5	0.12	8.5
Clarithromycin	Strong MBI	2.8	0.15	6.2

MBI: Mechanism-based inactivation. Parameters extracted from (Mayhew et al., 2000; Venkatakrishnan et al., 2007; He et al., 2020).

A.3. Patient Variability Parameters

Inter-individual pharmacokinetic variability is captured through the following distributions (Table 3 and Table 4).

A.4. Stochastic Variability Sources

Beyond inter-individual PK variability, the simulator incorporates episode-level stochastic noise to capture real-world measurement and adherence variability:

- **Assay imprecision**: Trough measurements are subject to CV=15% analytical error (Christians et al., 2015; Koster et al., 2013)
- **Food effects**: Random meal timing introduces bioavailability variation of ±20% (Bekersky et al., 2001)
- **Non-adherence**: 5% probability of missed doses per day (Prendergast & Gaston, 2010; Dew et al., 2007)
- **Hematocrit variation**: Whole blood trough concentrations vary with hematocrit (CV=8%) (Koster et al., 2013)

These combined sources produce trough coefficient of variation consistent with clinical transplant cohorts (CV ≈ 30–45%).

B. Drug-Drug Interaction Model Validation

Table 5. DDI Fold-Change Validation: PBPK Model vs. Literature

Inhibitor	Mechanism	Literature Range	PBPK Predicted	Reference
Fluconazole	Reversible	1.5–3.0×	2.93×	(He et al., 2020)
Verapamil	Weak MBI	2.0–3.0×	1.56×	(Chow & Jusko, 2004b)
Erythromycin	Strong MBI	1.4–4.6×	4.36×	(Chiang et al., 2019)
Clarithromycin	Strong MBI	2.0–6.0×	5.51×	(Cheung & Senior, 2016)

MBI: Mechanism-based inactivation. Fold-change represents ratio of tacrolimus trough concentration with DDI vs. without DDI. PBPK predictions based on population simulation (n=100 virtual patients per drug). Literature ranges compiled from clinical studies and case reports.

C. RL Agent Architecture and Training

Table 6 presents the complete LSTM-PPO network architecture and training hyperparameters used in this work.

Table 6. LSTM-PPO Architecture and Training Hyperparameters.

Component	Details
Network architecture	
Input dimension	5
MLP layer 1	Linear(5→128) + ReLU
MLP layer 2	Linear(128→128) + ReLU
LSTM hidden units	128
LSTM layers	1
Actor head	Linear(128→7) + Softmax
Critic head	Linear(128→1)
Shared LSTM	`shared_lstm=True`
Critic LSTM	`enable_critic_lstm=False`
PPO hyperparameters	
Clip ratio ε	0.2
GAE λ	0.95
Entropy coefficient	0.01
Value function coefficient	0.5
Learning rate	3×10^{-4} (Adam)
Batch size	64
Epochs per rollout	10
n_{steps}	128
Training timesteps	500,000
Framework	
Implementation	Stable-Baselines3-Contrib
Action space	7 discrete dose levels
Dose levels (mg BID)	{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0}
Episode length	45 days

D. Baseline Methods

We compare LSTM-PPO against five baseline methods spanning rule-based, control-theoretic, model-based, and learning-based approaches.

D.1. Constant Dose

The Constant Dose baseline administers a fixed dose of 2.0 mg twice daily (BID) to all patients regardless of observed trough concentrations, weight, or other patient characteristics. This represents a naive approach with no personalization or feedback control. The 2.0 mg BID dose (4.0 mg total daily dose) was selected as a common starting dose in clinical practice for average adult patients.

D.2. Random

The Random baseline selects doses uniformly at random from the seven discrete dose levels available to the RL agent: 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0 mg BID. This baseline establishes a lower bound on performance and serves as a sanity check that all methods are learning meaningful dosing strategies rather than random behavior.

D.3. Proportional Controller

The Proportional (P) Controller implements a classic feedback control strategy that adjusts dosing based on the error between current trough concentration and target (Sheppard, 1980):

$$d_{t+1} = d_t + K_p \cdot (C_{\text{target}} - C_{\text{trough},t}) \tag{8}$$

where d_t is the current dose, $C_{\text{trough},t}$ is the observed trough concentration, $C_{\text{target}} = 9.5$ ng/mL is the center of the therapeutic window, and $K_p = 0.5$ is the proportional gain. The proportional gain was tuned empirically to balance responsiveness with stability. Doses are clipped to the range [0.5, 3.0] mg BID and rounded to the nearest available discrete dose level. The controller initializes with a weight-normalized dose: $d_0 = 0.1 \times$ weight (kg) mg/day, converted to BID dosing.

This baseline represents a simple, interpretable feedback strategy commonly employed in clinical settings for drugs with narrow therapeutic windows (Arroyo-Currás et al., 2018).

D.4. Model Predictive Control (MPC)

The MPC baseline leverages the PBPK simulator directly to optimize dosing via forward simulation (Hovorka et al., 2004). At each decision point, MPC solves the following optimization problem:

$$d_t^* = \arg \max_{d_t, \ldots, d_{t+H-1}} \sum_{k=t}^{t+H-1} R(C_k) \tag{9}$$

where $H = 7$ days is the planning horizon, C_k is the predicted trough concentration at time step k, and $R(\cdot)$ is the same reward function used for RL training (combining immediate concentration reward and time-in-window reward).

PBPK Simulation: Forward predictions use the PBPK model equations (Appendix A) with *population-average* pharmacokinetic parameters rather than patient-specific values. Specifically, MPC assumes:

- Clearance: Population mean CL = 15.0 L/h
- Volume of distribution: Population mean V = 85.0 L
- CYP3A5 genotype: Expresser (default)
- DDI inhibition strength: Not known (assumes no DDI)

Optimization: We use the COBYLA (Constrained Optimization BY Linear Approximation) algorithm from SciPy to solve the optimization problem, with dose bounds [0.5, 3.0] mg BID. The optimization is warm-started from the previous dose to improve convergence.

Key Limitation: MPC has privileged access to the true PBPK model structure and equations, which RL does not. However, MPC does not know patient-specific PK parameters, CYP3A5 genotype, or current DDI status. While MPC observes the DDI active flag (indicating *that* a DDI is present), it does not know which specific inhibitor is being administered or its potency, and therefore cannot adjust its population-average parameters accordingly. This baseline tests whether direct model-based optimization with population parameters can outperform learned policies that implicitly infer patient-specific parameters from observed trajectories.

D.5. PPO Vanilla (MLP)

PPO Vanilla uses the same Proximal Policy Optimization algorithm and training configuration as LSTM-PPO, but replaces the recurrent LSTM architecture with a feedforward multi-layer perceptron (MLP). This isolates the contribution of recurrent memory to performance.

Architecture:

- **Input:** Same 5-dimensional observation vector as LSTM-PPO
- **Policy Network (Actor):** MLP with two hidden layers of [128, 128] units, ReLU activations, output layer with 7 units (one per dose level) and softmax activation
- **Value Network (Critic):** MLP with two hidden layers of [128, 128] units, ReLU activations, single output unit for state value $V(s)$

Training: Identical hyperparameters to LSTM-PPO (learning rate, batch size, GAE-λ, clipping, entropy coefficient) to ensure fair comparison. Training runs for the same number of environment steps (500,000) across multiple patients.

Key Difference: PPO Vanilla is *memoryless*—it cannot integrate information across multiple time steps. Each dosing decision is made based solely on the current observation, without access to the historical trough trajectory that would enable inference of patient-specific PK parameters or DDI dynamics. This makes PPO Vanilla a reactive rather than anticipatory policy.

E. EHR-Based Retrospective Evaluation: Detailed Results

Figure 2 presents a comprehensive view of the retrospective EHR-based evaluation on 6,394 real patients from the Cedars-Sinai OMOP cohort.

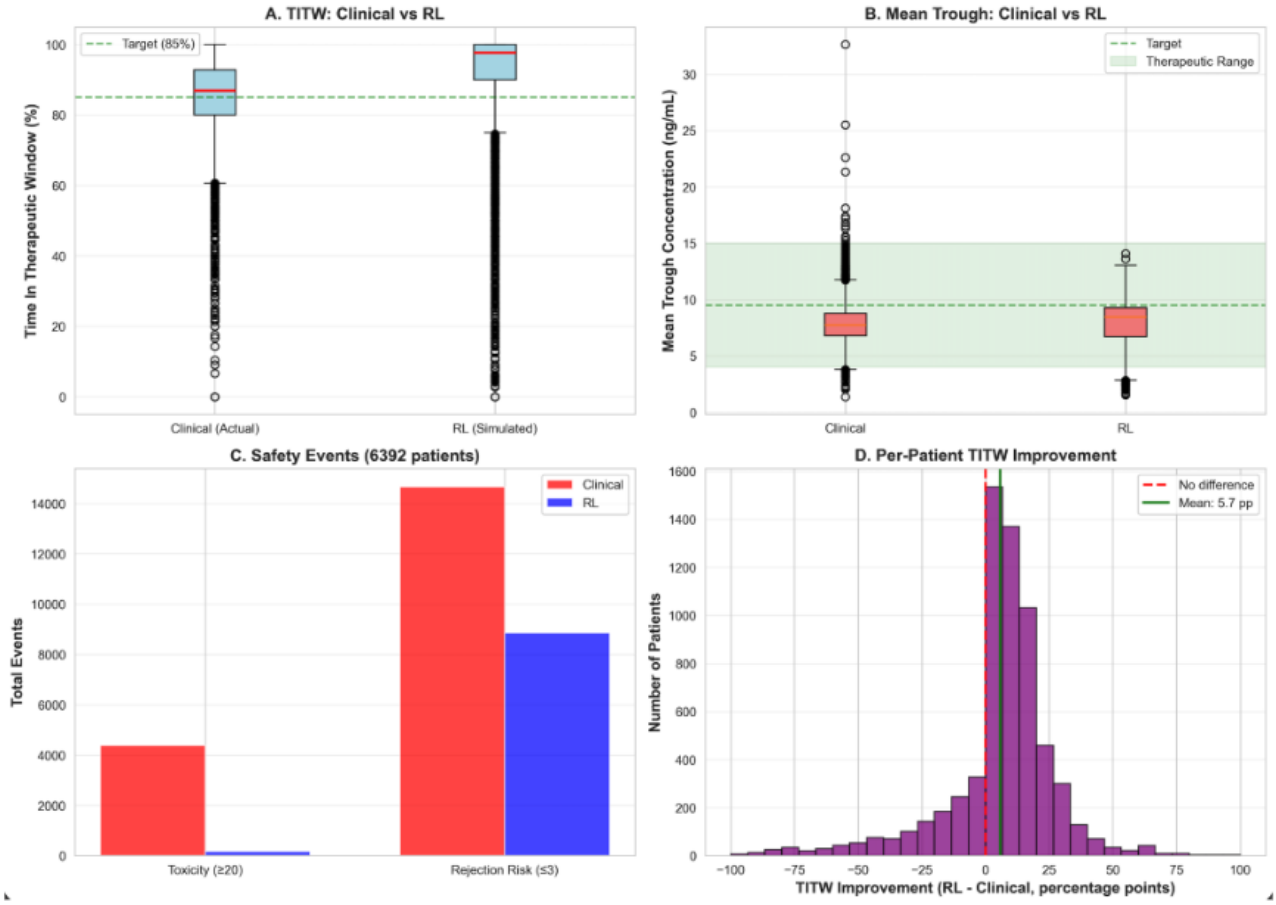

Figure 2. EHR-informed retrospective evaluation comparing observed EHR trough trajectories with simulated trough trajectories under the learned RL policy across 6,394 patients. **(A)** Time in therapeutic window (TITW) comparison showing median improvement with reduced variability. The target TITW of 85% (dashed line) is exceeded by the RL agent. **(B)** Mean trough concentration distributions for both approaches, demonstrating tighter control around the therapeutic range (4–15 ng/mL, shaded region) with the RL agent. **(C)** Safety events across the full cohort: the RL agent dramatically reduces both toxicity events ($C \geq 20$ ng/mL; red bars) and rejection risk events ($C \leq 3$ ng/mL; purple bars) . **(D)** Per-patient TITW improvement distribution showing that 92.6% of individual patients (5,923/6,394) experienced improvement with the RL agent, with mean improvement of +5.8 percentage points. The distribution is right-skewed, indicating substantial gains for patients with poor baseline control while avoiding harm to well-controlled patients.