Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

TRANSFORMER-BASED PK/PD MODELING FOR MONOCLONAL ANTIBODY DOSING FROM SPARSE CONCENTRATION DATA

Thabo Nkosi[1]*, Lerato Molefe[1], Sipho Dlamini[2], Ayanda Mokoena[1], Kabelo Ndlovu[2]

1. ***Department of Computational Pharmaceutical Engineering, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa.***
2. ***Department of AI Drug Systems, Faculty of Pharmacy, University of the Witwatersrand, Johannesburg, South Africa.***

ARTICLE INFO

Received:
01 December 2025
Received in revised form:
19 February 2026
Accepted:
21 February 2026
Available online:
28 February 2026

Keywords: Monoclonal antibodies, Transformer, Pharmacokinetics, Pharmacodynamics, Model-informed precision dosing, Sparse data

ABSTRACT

Monoclonal antibodies are central to modern biologic therapy for inflammatory, oncologic, and immune-mediated diseases, yet their pharmacokinetics are often nonlinear, highly variable between patients, and challenging to individualize when only limited concentration measurements are available. Traditional population PK/PD modeling and Bayesian forecasting rely on predefined structural models and assumptions regarding clearance, distribution, and variability, which can be fragile when data consist of sparse or irregularly timed therapeutic drug monitoring samples. To address these limitations, this MDL article proposes a transformer-based sequence model for individualized monoclonal antibody dosing, designed to infer patient-specific exposure patterns from sparse concentration data, dosing history, and clinical covariates without requiring an explicit compartmental model. The model employs a transformer encoder to process variable-length sequences of time, concentration, dose, and covariate tuples, using self-attention to capture relationships among prior doses, measured concentrations, elapsed time, and patient factors, thereby generating predicted future concentration trajectories and dose recommendations aimed at a target exposure window. Conceptually, this approach enables individualized concentration forecasts and dose suggestions even with minimal concentration data, complementing Bayesian forecasting by learning flexible temporal patterns that are difficult to specify parametrically. By extending model-informed precision dosing to data-sparse biologic treatment settings, a transformer-based framework could enhance the clinical utility of therapeutic drug monitoring for monoclonal antibodies while maintaining the need for prospective validation and clinical oversight.

To Cite This Article: Nkosi T, Molefe L, Dlamini S, Mokoena A, Ndlovu K. Transformer-Based PK/PD Modeling for Monoclonal Antibody Dosing from Sparse Concentration Data. Pharmacophore. 2026;17(1):120-8. https://doi.org/10.51847/hfTXxQ9dlx

Introduction

Monoclonal antibodies have become key therapeutic agents because they can engage disease-relevant targets with high specificity, yet their clinical benefit depends on achieving adequate exposure over time. Their pharmacokinetics are shaped by mechanisms such as target-mediated drug disposition, nonlinear clearance, and subcutaneous absorption, all of which can vary substantially across patients [1]. This variability motivates model-informed precision dosing, where therapeutic drug monitoring and patient covariates are used to adjust treatment rather than relying solely on fixed or body-weight-based regimens [2]. For biologics such as infliximab, precision dosing is clinically attractive because exposure is linked to pharmacologic response, disease control, and the avoidance of unnecessary underexposure or overexposure.

The standard pharmacometric toolset for individualized dosing is population PK/PD modeling, often followed by Bayesian forecasting once a patient's concentration becomes available. These approaches are powerful when the structural model adequately represents drug disposition and when the available observations are informative enough to estimate individual random effects. However, sparse or irregularly collected samples can make individualized inference unstable, especially if the model cannot fully account for nonlinear clearance, immunogenicity, inflammatory burden, or other time-varying covariates [3]. Hybrid pharmacokinetic–machine learning approaches have therefore been explored as a way to preserve pharmacologic structure while improving prediction from limited patient-level information.

Transformers offer a different modeling paradigm because self-attention can learn temporal relationships without imposing a predefined compartmental structure. Temporal Fusion Transformers were introduced for interpretable multi-horizon time-

Corresponding Author: Thabo Nkosi; Department of Computational Pharmaceutical Engineering, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa. E-mail: thabo.nkosi@gmail.com.

series forecasting [4], while Informer extended attention-based forecasting to long temporal sequences [5]. In clinical data science, transformer models such as BEHRT and Med-BERT demonstrated that attention-based representations can learn clinically meaningful temporal patterns from longitudinal health records [6, 7]. These developments suggest that transformer-based pharmacometric models could learn from irregular PK/PD histories, particularly when the relevant signal is distributed across dosing events, concentration measurements, and covariate trajectories.

This article proposes a transformer-based PK/PD model for monoclonal antibody dosing from sparse concentration data. The conceptual model learns directly from time-stamped dose, concentration, and covariate sequences to predict future exposure and support individualized dose adjustment, building on early work applying transformer and natural language processing models to longitudinal PK/PD analysis [8]. Unlike a conventional compartmental model, the proposed architecture would not require explicit specification of clearance, volume, or absorption compartments before training, although such pharmacologic knowledge could still guide feature engineering and evaluation [9]. The goal is not to replace classical pharmacometrics, but to create a flexible MDL framework that can be evaluated alongside Bayesian forecasting in the data-sparse biologics settings where current tools are most strained [10].

Background

PK/PD Characteristics of Monoclonal Antibodies

Monoclonal antibodies often exhibit long half-lives, restricted distribution, saturable target binding, and nonlinear elimination, making their PK/PD behavior more complex than that of many small molecules. Target-mediated drug disposition provides one mechanistic explanation for nonlinear exposure because binding to a pharmacologic target can influence both apparent clearance and response [11]. Subcutaneous administration adds further complexity, as absorption may depend on lymphatic transport, formulation, molecular properties, and patient-level physiologic factors [12]. Covariates such as body weight, albumin, inflammatory status, immunogenicity, and disease activity can therefore alter exposure in clinically meaningful ways, motivating models that can integrate both baseline and time-varying information [13].

Population PK/PD Modeling and Bayesian Forecasting

Population PK/PD modeling describes typical pharmacokinetic behavior and between-patient variability using structural, statistical, and covariate components, then uses observed concentrations to refine individual predictions. Bayesian forecasting within model-informed precision dosing combines prior population information with patient-specific observations, but its reliability depends on the appropriateness of the structural model and the informativeness of the available samples. In sparse therapeutic drug monitoring, a single or small number of concentrations may be insufficient to distinguish high clearance from delayed absorption, nonadherence, immunogenicity, or timing errors. This is why contemporary precision dosing frameworks emphasize fit-for-purpose evaluation, clinical workflow compatibility, and the careful definition of when model recommendations should be trusted.

Transformers and Self-Attention for Time-Series

Transformers use self-attention to compare each element in a sequence with every other element, allowing the model to learn dependencies that are not restricted to adjacent time points. In time-series forecasting, architectures such as Temporal Fusion Transformer combine attention with gating and variable-selection mechanisms to represent static covariates and temporal inputs in an interpretable way [4]. Informer was designed to improve transformer scalability for long-sequence forecasting, which is relevant to monoclonal antibody dosing because exposure may evolve over repeated administration cycles [5]. Because transformer inputs can include explicit time encodings rather than relying on a fixed sampling grid, the architecture is conceptually well suited to irregular PK data where concentrations are measured at clinically convenient rather than protocolized times [14].

Deep Learning for Pharmacometrics

Deep learning has begun to enter pharmacometrics through recurrent neural networks, neural PK/PD models, neural ordinary differential equations, and hybrid machine learning–pharmacometric approaches. Long short-term memory networks have been investigated for PK/PD modeling [15], and neural-pharmacokinetic/pharmacodynamic modeling has been proposed to predict response time courses from early data [16]. Deep compartment models and neural ordinary differential equation approaches show how mechanistic concepts can be blended with flexible learning systems [17, 18]. However, the specific use of transformers for sparse monoclonal antibody concentration histories remains an emerging gap, despite early evidence that transformer-style models can support longitudinal PK/PD analysis [8].

Data Paucity and Transfer Learning in PK

A central obstacle for deep learning in pharmacometrics is that individual-level PK datasets are often small, proprietary, heterogeneous, and sparsely sampled. Reviews of machine learning in pharmacometrics emphasize that model development must address limited data, external validation, interpretability, and regulatory credibility rather than simply applying generic prediction algorithms [19]. One conceptual solution is to pretrain on simulated or pooled population PK profiles, including data generated from physiologically based or population models, and then adapt the model to a specific biologic or clinical setting [20]. Such transfer learning could be especially useful for monoclonal antibodies because shared properties of biologics

may support cross-molecule representation learning, while molecule-specific fine-tuning could capture differences in target binding, clearance, and absorption [21].

Model Development Overview

High-Level Modelling Pipeline

The proposed pipeline begins with a patient's sparse concentration history, prior dosing records, and clinical covariates, which are encoded into a variable-length temporal tensor. A transformer encoder processes this sequence and produces a latent patient representation that reflects observed exposure, dose timing, and covariate context, extending the general idea of learning PK/PD time courses from early individual data [16]. A forecasting component would then predict the future concentration–time profile, while a dosing component would recommend a dose and interval expected to achieve a predefined therapeutic exposure range. This model-oriented structure follows the broader movement toward integrating machine learning with pharmacometrics, while retaining the clinical objective of individualized dose optimization [9].

Figure 1 illustrates the proposed transformer-based PK/PD architecture for converting sparse monoclonal antibody concentration histories, dosing events, elapsed-time information, and clinical covariates into individualized exposure forecasts, uncertainty-aware dose recommendations, and clinician-reviewed precision dosing support.

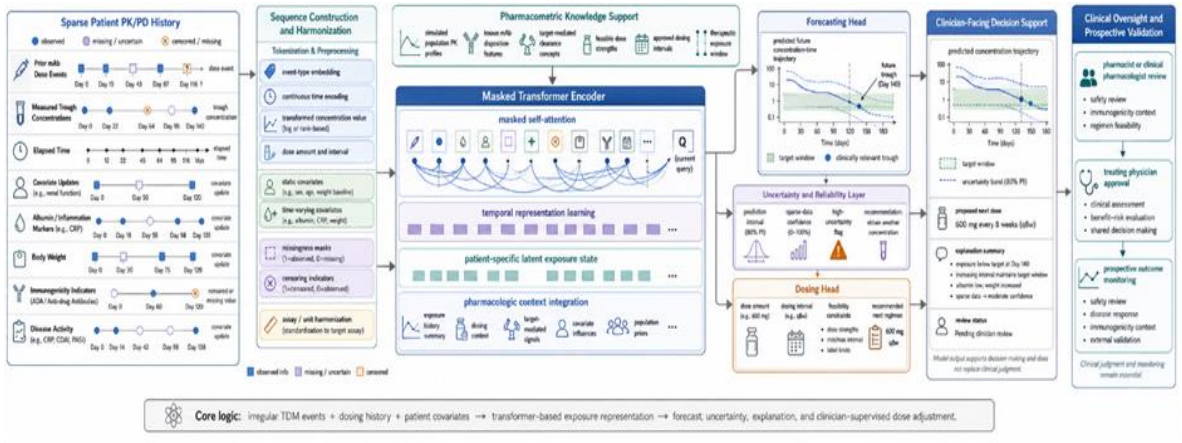

Figure 1. Transformer-Based PK/PD Architecture for Individualized Monoclonal Antibody Dosing from Sparse Concentration Histories

Core Input Features

The core input sequence would consist of time-stamped tuples representing dose events, concentration measurements, and clinically relevant covariates. Dose amount, route, nominal interval, measured concentration, and elapsed time since the prior dose are natural dynamic features, while baseline weight, age, sex, disease state, albumin, immunogenicity indicators, and endogenous IgG levels may function as static or slowly changing features [13]. In monoclonal antibody applications, the model should also allow disease activity markers to influence exposure predictions because inflammation and target burden may affect apparent clearance [11]. Similar to clinical transformer models that encode longitudinal records as temporal tokens, each PK event would be represented as part of a sequence rather than as an isolated observation [6].

Design Principles

The model is designed to be data-driven, minimally parametric, and capable of online updating as new concentration measurements become available. This design is motivated by prior work showing that deep learning models can represent pharmacokinetic patterns without requiring every component of the structural model to be manually specified [22]. Missing covariate values should be represented through masking and missingness indicators rather than silent deletion, because missingness itself may reflect clinical workflow or disease severity. At the same time, the model should remain pharmacometrically interpretable enough to be compared with classical population PK and hybrid PK–machine learning approaches during validation [10].

Data Sources and Preprocessing

Compilation of Sparse mAb PK Datasets

A practical implementation would require curated concentration–time data from therapeutic drug monitoring programs, clinical trials, and hospital databases across monoclonal antibodies such as infliximab, adalimumab, and rituximab. Infliximab is a useful initial case because model-informed dosing tools and exposure-guided decision support have already been studied in inflammatory bowel disease . Pediatric and young adult infliximab modeling has also provided an example of combining population pharmacokinetics with machine learning for individualized prediction. Multi-mAb curation would allow the transformer to learn both shared biologic disposition patterns and molecule-specific differences, but preprocessing would need to harmonize assay units, sampling times, dosing intervals, and clinical covariate definitions.

Encoding of Dosing History and Time-Aware Features

Dosing history should be encoded as a sequence of events rather than as a single cumulative exposure variable. Each token can represent either a dose administration, a concentration measurement, or a combined clinical observation, with time since the previous event included as an explicit feature. Time-aware sequence construction is consistent with transformer forecasting methods, where temporal position and covariate context help the model distinguish short-term from long-term dependencies [4]. For monoclonal antibodies with long dosing intervals, this representation is especially important because a concentration measured weeks after administration has a different interpretation from one measured near peak or trough exposure [12].

Handling Censored Concentrations and Missing Doses

Concentrations below the limit of quantification should not be discarded, because low or censored observations may carry clinically important information about high clearance, poor adherence, or inadequate exposure. Instead, the model can include a censoring indicator and a transformed concentration value that preserves the distinction between observed, censored, and missing data. Missing or uncertain dose events could be represented with a separate token type and confidence indicator, allowing the model to distinguish a documented missed dose from an unknown dosing history. This approach aligns with broader machine learning recommendations in pharmacokinetics, where uncertainty in clinical data provenance should be represented explicitly rather than hidden during preprocessing [23].

Transformer Architecture for PK/PD Time-Series

Input Sequence Representation

Each event in the patient history would be converted into a token containing a type embedding, a value embedding, and a continuous time encoding. Dose and concentration values could be transformed to stabilize scale, while categorical indicators would distinguish dose administrations, PK samples, laboratory values, and covariate updates. This token-based formulation parallels transformer models for structured health records, where clinical events are embedded into a common representation that can be processed longitudinally [7]. For PK/PD use, the key adaptation is that token meaning must preserve pharmacologic timing, because the same concentration can imply different clearance behavior depending on dose amount and elapsed time [8].

Table 1 defines the event-token structure through which sparse monoclonal antibody dosing histories, concentration measurements, time intervals, covariates, and uncertainty signals can be represented for transformer-based PK/PD forecasting.

Table 1. Event-Token Architecture for Transformer-Based PK/PD Modeling of Monoclonal Antibody Dosing

Event-token domain	Representative variables encoded	Temporal role in sparse PK/PD sequence	Transformer modeling function	Pharmacometric interpretation added by the token
Dose administration token	Dose amount, route, infusion or injection date, nominal dosing interval, dose confidence indicator	Defines the exposure-generating event from which future concentrations are interpreted	Provides an anchor for attention across subsequent concentration and covariate events	Helps the model distinguish low exposure caused by insufficient dose, long interval, missed dose, or altered clearance
Concentration measurement token	Measured drug level, assay unit, sample date, trough or non-trough status, censoring flag	Supplies sparse observed evidence of realized exposure	Allows the model to compare observed concentrations with prior doses and elapsed time	Supports individualized inference of apparent clearance, absorption delay, or underexposure
Continuous time token	Time since previous event, time since last dose, absolute treatment day, dosing-cycle position	Preserves irregular sampling intervals without forcing a fixed time grid	Enables time-aware attention across unevenly spaced clinical events	Prevents a concentration from being interpreted without its pharmacokinetic timing context
Static patient covariate token	Body weight, age, sex, diagnosis, baseline disease category, baseline albumin	Provides patient-level context that may influence typical exposure	Conditions the latent patient representation across the entire sequence	Supports between-patient variability modeling without requiring predefined random effects
Time-varying clinical covariate token	Albumin, inflammatory markers, disease activity, endogenous IgG, immunogenicity indicators	Captures changing physiologic or immunologic context over treatment	Allows attention to identify covariate shifts that modify future exposure	Helps interpret nonlinear clearance, target burden, inflammation-related exposure loss, or immunogenicity-driven changes
Missingness and uncertainty token	Missing covariate flag, uncertain dosing history, delayed sample, undocumented interruption	Represents incomplete clinical information explicitly	Prevents silent deletion or misleading imputation of uncertain events	Signals when the observed history may be insufficient for confident dose adjustment
Censored concentration token	Below-limit-of-quantification indicator, transformed low value, assay detection limit	Preserves clinically meaningful low-exposure evidence	Allows the model to learn from censored observations without treating them as absent	Identifies possible high clearance, nonadherence, delayed dosing, or inadequate therapeutic exposure
Molecule-context token	Monoclonal antibody name, route, target class,	Supports learning across multiple antibodies while	Enables transfer learning and molecule-specific fine-tuning	Separates shared biologic disposition patterns from antibody-

	formulation, expected dosing interval	retaining molecule-specific context		specific clearance or absorption behavior
Decision-time token	Current visit date, available evidence window, candidate next dose date, target exposure window	Marks the point at which the model must forecast and recommend	Restricts prediction to information available before the clinical decision	Aligns model output with real therapeutic drug monitoring workflow rather than retrospective fitting

Transformer Encoder with Masked Self-Attention

The transformer encoder would use masked self-attention so that predictions for future exposure are based only on information available up to the clinical decision time. Self-attention allows a trough concentration to attend to a distant prior dose, an earlier high concentration, or a relevant covariate shift, which is valuable when many intervening weeks contain no measurements [5]. In contrast to recurrent architectures, the attention mechanism does not require information to pass sequentially through every intermediate step, which may help when PK sampling is sparse and unevenly spaced. This design builds on the broader use of transformer and natural language processing models in longitudinal pharmacology and clinical time-series prediction [8, 14].

Prediction and Dosing Heads

The architecture would contain a forecasting head for future concentration prediction and a dosing head for individualized dose recommendation. The forecasting head could estimate the concentration expected at clinically relevant future times, while the dosing head could search the predicted exposure space for a dose and interval expected to maintain the target range. Such an approach is conceptually related to machine learning exposure-response modeling, where prediction is linked to an actionable dosing or therapeutic objective rather than treated as a purely statistical endpoint [24]. The dosing head should also be constrained by feasible dose strengths, approved intervals, and clinical rules, because model-informed precision dosing must remain compatible with safe prescribing workflows [2].

Handling Sparse and Irregularly Sampled Data

Continuous Time Encoding

Sparse monoclonal antibody PK data rarely follow a regular time grid, so the model should encode elapsed time directly rather than treating each observation as equally spaced. A learnable continuous time function could be added to the token embedding, allowing the transformer to distinguish a concentration measured shortly after dosing from one collected near the end of a dosing interval [4]. This design is consistent with attention-based time-series forecasting, where temporal context and covariate structure are modeled jointly rather than forced into a rigid compartmental schedule [5]. In pharmacometric terms, continuous time encoding would help the model respect clinically meaningful differences in absorption, distribution, and elimination phases without requiring these phases to be explicitly parameterized [12].

Attention-Based Imputation of Missing Context

When many weeks pass without observed concentrations, the transformer can use attention to identify which earlier doses, measurements, and covariate values remain informative for the current prediction. This is not imputation in the conventional sense of filling a complete concentration–time table; rather, the model learns a latent representation of missing context during the forward pass. Deep compartment models have shown that flexible neural structures can learn pharmacokinetic trajectories from incomplete time-series data [17], and explainable neural-network PK models suggest that learned representations can still be interrogated after training [25]. For monoclonal antibody dosing, attention-based context reconstruction would be especially useful when trough samples are missed, recorded late, or collected at clinically irregular intervals.

Evaluating Prediction Uncertainty on Scarce Data

Prediction uncertainty is essential when the model is asked to recommend a dose from very few concentration measurements. Quantile regression, ensemble methods, or Monte Carlo dropout could be used conceptually to produce prediction intervals, allowing the model to signal whether the available evidence is sufficient for dose adjustment [23]. This uncertainty layer would be aligned with the clinical logic of model-informed precision dosing, where a recommendation should be accepted only when the model is adequately informed by patient-specific data and clinically plausible assumptions. If uncertainty remains high, the system should recommend obtaining an additional concentration before changing therapy, rather than presenting an overconfident dosing decision.

Model Interpretability and Dosing Decision Support

Attention-Based Explanation of Dose Recommendations

Attention weights can be used to highlight which prior concentrations, dose events, and covariates most influenced the forecast, creating a transparent bridge between the patient history and the dose recommendation. This interpretability goal is consistent with Temporal Fusion Transformer design, where attention and variable-selection components support examination of temporal drivers of prediction [4]. In PK/PD modeling, Shapley additive explanations have also been used to interpret neural-network pharmacokinetic predictions, showing that post hoc explanation can help make data-driven models more acceptable

to pharmacometric users [25]. For monoclonal antibody dosing, explanation should focus on clinically recognizable factors such as recent trough level, dose interval, body weight, inflammatory burden, and evidence of altered clearance.

Integration into Clinical Dashboards

The model output should be displayed in a familiar pharmacometric dashboard format, with predicted concentration–time curves, the target exposure window, uncertainty intervals, and the proposed next dose shown together. Existing model-informed precision dosing software has been evaluated partly on whether it satisfies clinical workflow needs, not only on whether it implements sound pharmacokinetic calculations. In infliximab therapy, model-informed tools that forecast trough exposure and relate it to disease status illustrate how pharmacometric predictions can be translated into clinician-facing interfaces. A transformer-based dashboard should therefore present recommendations as decision support, allowing pharmacists, clinical pharmacologists, and treating physicians to review the predicted exposure profile before accepting or modifying the proposed regimen.

Integration Into Clinical Pharmacometrics Workflows

Real-Time Dose Optimization in TDM Practice

In therapeutic drug monitoring practice, the transformer model could be invoked when a new concentration result becomes available and used to update the patient's predicted exposure profile. This workflow parallels current model-informed precision dosing, where observed concentrations are combined with prior information to guide dose selection, but the transformer would learn the temporal mapping directly from patient histories and covariates [2]. For infliximab induction therapy, precision dosing concepts have already been linked to efforts to reduce exposure variability and support therapeutic targets. A transformer-based system would be expected to fit best as a rapid advisory layer that converts sparse monitoring data into a proposed dosing action while preserving clinician review.

Complementing, Not Replacing, Classical Pharmacometrics

The transformer model should be positioned as complementary to population PK/PD modeling rather than as a replacement for mechanistic pharmacometrics. Classical models remain essential for study design, mechanistic interpretation, extrapolation, regulatory communication, and simulation of unobserved scenarios, especially when their assumptions are well supported by data [9]. Comparative work on scientific machine learning and population pharmacokinetic/pharmacodynamic models emphasizes that each approach has distinct strengths and should be evaluated according to its intended use [10]. In data-sparse monoclonal antibody TDM, the transformer could serve as an alternative forecasting engine when Bayesian updating is unstable, while population PK models could still provide priors, simulated training profiles, and mechanistic plausibility checks [20].

Evaluation Strategy

Predictive Accuracy

Predictive evaluation should compare the transformer's concentration forecasts with those from a Bayesian population PK model on the same sparse test histories. Metrics such as root mean squared error, median prediction error, individual prediction error, and calibration of prediction intervals could be defined prospectively, but this conceptual article does not report numerical outcomes. Deep learning methods for drug concentration prediction have emphasized the need for external validation and clinically meaningful forecast horizons rather than isolated in-sample performance [26]. The evaluation should also examine whether errors differ by sampling density, dosing interval, route of administration, disease activity, and covariates known to influence monoclonal antibody exposure [13].

Table 2 provides a validation and governance framework for determining whether transformer-based monoclonal antibody dosing recommendations are accurate, calibrated, interpretable, clinically feasible, and safe for prospective use.

Table 2. Validation and Clinical-Governance Framework for Transformer-Based Monoclonal Antibody Precision Dosing

Evaluation or governance layer	Core question addressed	Recommended assessment strategy	Failure mode detected	Required clinical safeguard
Sparse-history predictive accuracy	Can the model forecast future concentrations when only one or a few prior measurements are available?	Compare predicted concentration–time profiles against observed follow-up concentrations using sparse test histories	Good performance only in densely sampled patients but unstable predictions in real TDM settings	Report performance stratified by number of prior concentrations and sampling irregularity
Bayesian comparator evaluation	Does the transformer add value beyond classical population PK/Bayesian forecasting?	Benchmark against an established population PK model with Bayesian updating on identical patient histories	Apparent model benefit caused by an unfair or weak comparator	Use matched input data, matched forecast horizons, and clinically meaningful exposure endpoints
Calibration and uncertainty reliability	Are prediction intervals trustworthy enough to guide dose adjustment?	Evaluate calibration of prediction intervals, uncertainty width, and high-uncertainty alerts	Overconfident recommendations from insufficient data	Require "obtain additional concentration" output when uncertainty exceeds a predefined threshold

Dose decision quality	Would the recommended dose plausibly achieve the therapeutic exposure window?	Simulate next-dose scenarios and estimate probability of target attainment under feasible dose strengths and intervals	Accurate concentration fitting but clinically unusable dose suggestions	Constrain recommendations to approved regimens, local protocols, and prescriber-defined limits
Covariate and subgroup robustness	Does model performance remain stable across clinically important patient groups?	Stratify errors by body weight, albumin, inflammatory burden, disease activity, age group, route, and immunogenicity status	Systematic underprediction or overprediction in high-risk subgroups	Require subgroup performance reporting before deployment
Cross-mAb generalization	Can the model transfer learned biologic PK representations across monoclonal antibodies?	Train on selected antibodies and test or fine-tune on held-out antibodies	Model memorizes molecule-specific patterns and fails on new biologics	Use molecule-context encoding, external validation, and molecule-specific calibration before clinical use
Interpretability and explanation review	Can clinicians understand why a dose recommendation was generated?	Inspect attention summaries, variable-attribution outputs, and clinically recognizable explanation narratives	Black-box recommendation that cannot be reconciled with patient history	Present recent trough, dose interval, body weight, inflammation, and immunogenicity signals alongside the recommendation
Data provenance and preprocessing audit	Are concentration, dosing, and covariate inputs reliable enough for decision support?	Audit assay units, sample timing, dose documentation, missingness, censoring, and EHR extraction logic	Incorrect recommendation caused by timing error, unit mismatch, or undocumented missed dose	Flag uncertain inputs and require manual review before dose acceptance
Prospective workflow validation	Does the model improve real therapeutic drug monitoring practice safely?	Evaluate in prospective clinical workflow with pharmacist, pharmacologist, and physician review	Retrospective accuracy does not translate into usable clinical decision support	Deploy first as advisory support with human approval and outcome monitoring
Post-deployment monitoring	Does performance remain stable over time and across sites?	Track prediction errors, alert frequency, override rates, target attainment, and safety outcomes	Model drift due to changing assays, protocols, populations, or biologic use patterns	Maintain periodic recalibration, governance review, and suspension criteria

Dose Decision Quality

Dose decision quality should be assessed by whether model-recommended regimens would be expected to place future concentrations within a predefined therapeutic exposure window while respecting feasible prescribing constraints. The comparison set could include historical physician-prescribed doses, Bayesian-guided doses, and hybrid pharmacokinetic–machine learning recommendations, with outcomes interpreted conceptually rather than as claims of superiority. Exposure-response machine learning methods provide a useful framework because they connect prediction to therapeutic action, not merely to concentration fitting [24]. For biologics, dose decision evaluation should also consider immunogenicity, target burden, and nonlinear clearance because these mechanisms can alter whether a recommended dose is clinically plausible [11].

Generalization Across mAbs

Generalization should be evaluated by training on some monoclonal antibodies and testing on others, thereby assessing whether the model learns transferable biologic PK/PD representations. This is important because antibody molecules can share broad disposition features while differing in Fc-fusion structure, subcutaneous absorption, target biology, biophysical properties, and clinical clearance behavior [27]. Machine learning models that use molecular dynamics simulations, surface descriptors, or biophysical assays to predict antibody properties suggest that structural and developability information could eventually be linked to PK representation learning [28, 29]. A successful cross-mAb evaluation would not require the transformer to ignore molecule-specific mechanisms, but it should show that pretraining on related biologics could support adaptation when direct patient-level data for a new antibody are limited [21].

Limitations

Need for Adequate Pre-Training Data

A transformer-based PK/PD model would require adequate pretraining data before it could provide reliable individualized dosing support. For newly approved monoclonal antibodies, patient-level concentration histories may initially be too limited to learn molecule-specific clearance, absorption, target-mediated effects, and immunogenicity patterns. Simulated data from population PK, physiologically based PK, or mechanistic antibody models could help initialize the system, but simulated profiles may not fully represent real-world adherence, assay variability, covariate missingness, or disease heterogeneity. Therefore, early deployment should be conservative, with uncertainty estimates and prospective validation used to determine when the model is ready for clinical decision support.

Not a Replacement for Clinical Judgment

The proposed model would not incorporate every clinical factor that may influence a dosing decision, including acute illness, contraindications, infection risk, drug interruptions, drug–drug interactions, pregnancy, organ dysfunction, or patient preference. It would also not determine whether treatment should continue, switch, or stop when the therapeutic question is primarily clinical rather than pharmacokinetic. The appropriate role of the model is to support dose selection under professional oversight, not to automate biologic prescribing. Clinicians should interpret every recommendation alongside diagnosis, response, safety, laboratory results, and the broader therapeutic plan.

Conclusion

A transformer-based PK/PD model for monoclonal antibody dosing would provide a flexible framework for learning from sparse concentration histories. By encoding dose events, measured concentrations, elapsed time, and patient covariates as a temporal sequence, the model could infer individualized exposure patterns without requiring an explicit compartmental structure.

The main strength of this approach is its ability to operate on few, irregularly spaced observations while still producing clinically actionable forecasts. When paired with uncertainty estimation and interpretable attention-based explanations, the model could provide dose recommendations that are transparent enough for clinical pharmacology review.

Important challenges remain before such a model could be used in practice. These include the need for adequate pretraining data, careful handling of inter-antibody differences, robust external validation, and prospective testing in real therapeutic drug monitoring workflows.

The next step is collaborative development of multi-center, multi-monoclonal antibody datasets that preserve dosing history, concentration timing, covariates, and clinical outcomes. Such datasets would allow transformer-based PK/PD models to be evaluated rigorously as decision-support tools for individualized biologic dosing.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Jones HM, Zhang Z, Jasper P, Luo H, Avery LB, King LE, et al. A physiologically-based pharmacokinetic model for the prediction of monoclonal antibody pharmacokinetics from in vitro data. CPT Pharmacometrics Syst Pharmacol. 2019;8(10):738-47.
2. Tyson RJ, Park CC, Powell JR, Patterson JH, Weiner D, Watkins PB, et al. Precision dosing priority criteria: drug, disease, and patient population variables. Front Pharmacol. 2020;11:420.
3. Peletier LA, Jansson-Löfmark R, Gabrielsson J. Comparisons of basic target-mediated drug disposition (TMDD) and ligand facilitated target removal (LFTR). Eur J Pharm Sci. 2021;162:105835.
4. Lim B, Arık SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast. 2021;37(4):1748-64.
5. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35(12):11106-15.
6. Li Y, Rao S, Solares JR, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. Sci Rep. 2020;10(1):7155.
7. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4(1):86.
8. Cheng Y, Hu H, Dong X, Hao X, Li Y. Exploring transformer model in longitudinal pharmacokinetic/pharmacodynamic analyses and comparing with alternative natural language processing models. J Pharm Sci. 2024;113(5):1368-75.
9. Koch G, Pfister M, Daunhawer I, Wilbaux M, Wellmann S, Vogt JE. Pharmacometrics and machine learning partner to advance clinical data analysis. Clin Pharmacol Ther. 2020;107(4):926-33.
10. Valderrama D, Teplytska O, Koltermann LM, Trunz E, Schmulenson E, Fritsch A, et al. Comparing scientific machine learning with population pharmacokinetic and classical machine learning approaches for prediction of drug concentrations. CPT Pharmacometrics Syst Pharmacol. 2025;14(4):759-69.
11. An G. Concept of pharmacologic target-mediated drug disposition in large-molecule and small-molecule compounds. J Clin Pharmacol. 2020;60(2):149-63.
12. Hu S, Datta-Mannan A, D'Argenio DZ. Physiologically based modeling to predict monoclonal antibody pharmacokinetics in humans from in vitro physiochemical properties. MAbs. 2022;14(1):2056944.

13. Bei R, Thomas J, Kapur S, Woldeyes M, Rauk A, Robarge J, et al. Predicting the clinical subcutaneous absorption rate constant of monoclonal antibodies using only the primary sequence: a machine learning approach. MAbs. 2024;16(1):2352887.
14. Madan S, Lentzen M, Brandt J, Rueckert D, Hofmann-Apitius M, Fröhlich H. Transformer models in biomedicine. BMC Med Inform Decis Mak. 2024;24(1):214.
15. Liu X, Liu C, Huang R, Zhu H, Liu Q, Mitra S, et al. Long short-term memory recurrent neural network for pharmacokinetic-pharmacodynamic modeling. Int J Clin Pharmacol Ther. 2021;59(2):138.
16. Lu J, Bender B, Jin JY, Guan Y. Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling. Nat Mach Intell. 2021;3(8):696-704.
17. Janssen A, Leebeek FW, Cnossen MH, Mathot RA, Fijnvandraat K, Coppens M, et al. Deep compartment models: a deep learning approach for the reliable prediction of time-series data in pharmacokinetic modeling. CPT Pharmacometrics Syst Pharmacol. 2022;11(7):934-45.
18. Bräm DS, Steiert B, Pfister M, Steffens B, Koch G. Low-dimensional neural ordinary differential equations accounting for inter-individual variability implemented in Monolix and NONMEM. CPT Pharmacometrics Syst Pharmacol. 2025;14(1):5-16.
19. Janssen A, Bennis FC, Mathot RA. Adoption of machine learning in pharmacometrics: an overview of recent implementations and their considerations. Pharmaceutics. 2022;14(9):1814.
20. Keutzer L, You H, Farnoud A, Nyberg J, Wicha SG, Maher-Edwards G, et al. Machine learning and pharmacometrics for prediction of pharmacokinetic data: differences, similarities and challenges illustrated with rifampicin. Pharmaceutics. 2022;14(8):1530.
21. Patidar K, Pillai N, Dhakal S, Avery LB, Mavroudis PD. Development of an mPBPK machine learning framework for early target pharmacology assessment of biotherapeutics. Sci Rep. 2025;15(1):4198.
22. Liu G, Brooks L, Canty J, Lu D, Jin JY, Lu J. Deep-NCA: A deep learning methodology for performing noncompartmental analysis of pharmacokinetic data. CPT Pharmacometrics Syst Pharmacol. 2024;13(5):870-9.
23. Huang S, Xu Q, Yang G, Ding J, Pei Q. Machine learning for prediction of drug concentrations: application and challenges. Clin Pharmacol Ther. 2025;117(5):1236-47.
24. Liu C, Xu Y, Liu Q, Zhu H, Wang Y. Application of machine learning based methods in exposure–response analysis. J Pharmacokinet Pharmacodyn. 2022;49(4):401-10.
25. Ogami C, Tsuji Y, Seki H, Kawano H, To H, Matsumoto Y, et al. An artificial neural network− pharmacokinetic model and its interpretation using Shapley additive explanations. CPT Pharmacometrics Syst Pharmacol. 2021;10(7):760-8.
26. Khusial R, Bies RR, Akil A. Deep learning methods applied to drug concentration prediction of olanzapine. Pharmaceutics. 2023;15(4):1139.
27. Tomasoni D, Paris A, Visintainer R, Cook KD, Chen A, Figueroa I, et al. Predicting Aberrant Fc-fusion Protein Pharmacokinetics from In Silico Structural Properties and Physiologically Based Pharmacokinetic (PBPK) Modeling. AAPS J. 2026;28(3):87.
28. Wu IE, Kalejaye L, Lai PK. Machine learning models for predicting monoclonal antibody biophysical properties from molecular dynamics simulations and deep learning-based surface descriptors. Mol Pharm. 2024;22(1):142-53.
29. Grinshpun B, Thorsteinson N, Pereira JN, Rippmann F, Nannemann D, Sood VD, et al. Identifying biophysical assays and in silico properties that enrich for slow clearance in clinical-stage therapeutic antibodies. MAbs. 2021;13(1):1932230.