# DeepNLME rivals expert-specified pharmacokinetic functions for mosunetuzumab

*Can learned functions inside an NLME framework replace predefined empirical PK functions, while maintaining predictive performance and enabling broader hypothesis exploration?*

Sreenath M. Krishnan[1,*] · Denison Kuruvilla[2,*] · Kenta Yoshida[2] · Niklas Korsbo[1] (presenting) · Chi-Chung Li[2]

[1] Pumas-AI, Inc. [2] Genentech, Inc. [*] Co-first authors.

IN COLLABORATION WITH PumasAI

ACoP 2026
Poster W-050

## Background and objective

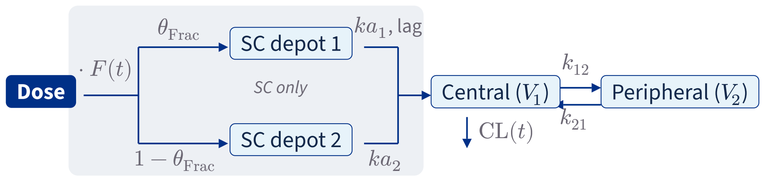
Mosunetuzumab is an off-the-shelf CD20 × CD3 T-cell–engaging bispecific antibody approved for relapsed/refractory **follicular lymphoma** after ≥ 2 prior systemic therapies [7, 8]. By binding CD20 on B cells and CD3 on T cells, it promotes immune-synapse formation and T-cell–mediated B-cell killing.

Its established popPK model [1–3] includes predefined **empirical** functions for time-varying clearance and SC bioavailability — functions that required iterative structural specification during development, thereby providing a relevant case study for potential DeepNLME applications.

Prior modelling also suggested that mosunetuzumab may operate within a relatively low, sub-saturating target-engagement range, where concentration-dependent TMDD was not identifiable from the clinical PK data.

**Objective.** We evaluated whether DeepNLME could (1) **reduce** manual functional-form specification, (2) **maintain** held-out predictive performance and (3) **enable** exploration of alternative model inputs.

## The published model



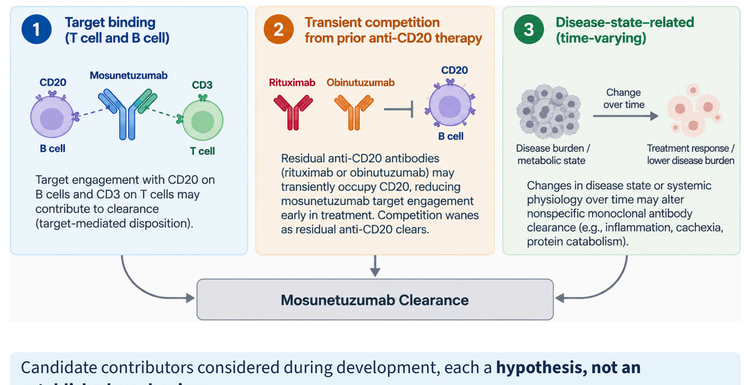

**F(t)** **CL(t)** are the two empirical functions DeepNLME replaces

$$\mathrm{CL}(t) = \mathrm{CL}_{\mathrm{base}} + (\mathrm{CL}_{\mathrm{ss}} - \mathrm{CL}_{\mathrm{base}})\left(1 - e^{-(\ln 2/\mathrm{HL}_{\mathrm{trans}})t}\right)$$

$$\mathrm{logit}\ F(t) = F_{\mathrm{base}} + (F_{\mathrm{final}} - F_{\mathrm{base}})\left(1 - e^{-(\ln 2/\mathrm{HL}_{,\mathrm{trans}\ F})t}\right)$$

*Fig 1. Both were specified as these empirical forms.*

## Mechanistic hypotheses for mosunetuzumab clearance




Candidate contributors considered during development, each a **hypothesis, not an established mechanism**.

## 1 Workflow: can DeepNLME reduce manual functional-form iteration?

### Two ways to search the same space

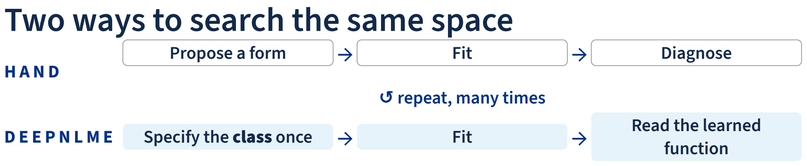



*Fig 2. Conventional: iterative form selection. DeepNLME: specify the function class once, then rank fits automatically.*

**Learned instead, inside the otherwise unchanged NLME model:**

$$\mathrm{CL}_i(t) = \mathrm{NN}_w\left(t, \mathrm{BWT}_i, \eta_{V_1,i}, \eta_{\mathrm{NN},i}\right) \quad F_i(t) = \mathrm{expit}\left(\mathrm{NN}_w\left(t, \eta_{F,i}\right)\right)$$

Both spend **5 random effects on IV: 3 inside CL(t)** — on $\mathrm{CL}_0$, $\mathrm{CL}_{ss}$, $\mathrm{HL}_{trans}$ in the closed form, on the network here — and 2 on the volumes. $\eta_{V_1}$ is an input because the network's own are independent by construction: the only route left for the CL–$V_1$ correlation.

**Same patients** **Same compartments** **Same number of random effects**

**ANSWER:** DeepNLME reduced the need to manually specify and evaluate candidate functional forms. **No mechanistic structure is lost in this case**, because the replaced components were already empirical in the closed-form model [1–3].

WHY IT MATTERS The potential gain is reduced manual functional-form iteration, **not** shorter runtime per fit. No controlled development-time comparison was performed.

## Methods

**Design.** IV first: CL(t) replaced — closed form d100 against neural dp100c. SC added next: F(t) replaced — closed form d200a against dp202c (IV fit held fixed) and dp203c (re-estimated jointly).

**Data.** 439 IV and 228 SC patients (GO29781; NCT02500407). Subject-level 70 / 15 / 15 split (IV 302 / 66 / 71; SC 160 / 30 / 38). The tuning set was used for model selection; **the test set was withheld from all modelling decisions.**

**Software.** DeepPumas v0.9 [6], Julia 1.11.7. Softplus constrains CL > 0.

**Validation.** Three random restarts per neural model, on both routes, with **early stopping** on the tuning set and restarts ranked the same way. Per-fit runtime is longer for the neural model: on IV, about 70 min against 6 min in Pumas, or 39 min in NONMEM.

**Note.** Poster values supersede the submitted abstract following QC correction of a data-entry issue (a CMAX / CTROUGH swap affecting 13 subjects) and duplicate observation records.

### What the substitution left alone

Median and full range across the three restarts (d100 is a single run). Not one lucky fit.

| Parameter | d100 closed form | dp100c neural | Δ | range over restarts |
|---|---|---|---|---|
| $V_1$ (L) | 5.298 | 5.329 | +0.6 % | 5.32 – 5.34 |
| Q (L/d) | 1.389 | 1.422 | +2.3 % | 1.40 – 1.43 |
| σ residual | 0.240 | 0.236 | −1.6 % | 0.234 – 0.236 |
| Ω on $V_1$ | 0.115 | 0.120 | +4.8 % | 0.120 – 0.123 |
| $\theta_{BWT}$ on $V_1$ | 0.537 | 0.623 | +16 % | 0.59 – 0.66 |
| $V_2$ (L) | 6.057 | 5.502 | −9.2 % | 5.36 – 5.56 |
| Ω on $V_2$ | 0.045 | 0.036 | −19 % | 0.031 – 0.042 |
| $\theta_{BWT}$ on $V_2$ | 0.693 | 0.322 | −54 % | 0.31 – 0.34 |
| $\theta_{BWT}$ on CL | 0.750 | — | — | — |

***Unchanged by the substitution**: $V_1$, Q, the residual error and the variance on $V_1$ — each within a few per cent and pinned to under 3 % over restarts. What moves is the **body-weight coefficients** and the peripheral compartment. BWT is now also a network input, so its influence is shared with the network rather than carried by the structural terms alone; $\vartheta_{BWT}$ on CL is **not separately identified here** and is left blank rather than compared. The two Ω rows sit in different covariance blocks, so are not strictly comparable.*

## 2 Performance: does the learned function maintain predictive performance?

### A. IV clearance: relatively robust

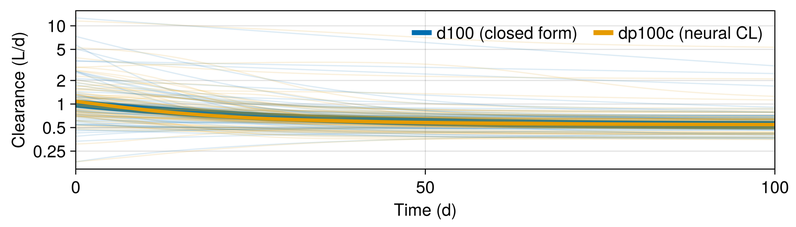



*Fig 3. Population and individual CL(t). Blue: closed form. Gold: neural, given only time, body weight and η. Bold: population prediction at 78 kg; faint: all 66 tune subjects. Log axis, nothing clipped.*

- Neural and closed-form CL(t) profiles **nearly overlapped** despite the neural form being learned rather than specified.
- The VPC is comparable too (Fig 4).

| +25.75 / +12.00 | 1.08 vs 1.01 | 0.55 vs 0.54 |
|---|---|---|
| ΔLL tune / held-out test neural vs closed form | L/d clearance at first dose neural vs closed form | L/d clearance at Day 100 neural vs closed form |

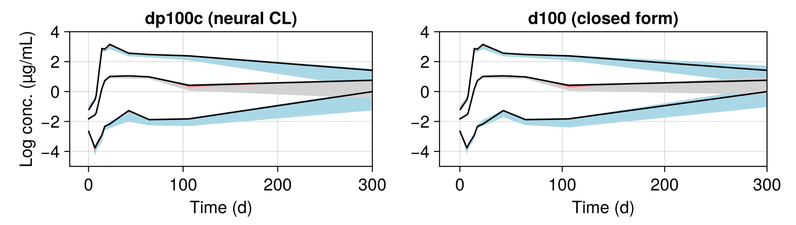



*Fig 4. Visual predictive check, IV route, 500 replicates. Observed 5th, 50th and 95th percentiles against their simulated 95 % intervals; red marks where an observed percentile falls outside one. The two models are near-indistinguishable.*

### B. SC bioavailability: structurally sensitive

| Model | SC bioavailability | ΔLL tune | ΔLL test |
|---|---|---|---|
| d200a | closed-form logit F(t) | 0.00 | 0.00 |
| dp202c | neural F(t), IV fixed | +5.53 | +34.32 |
| dp203c | neural F(t), disposition re-opened | +6.28 | +26.99 |

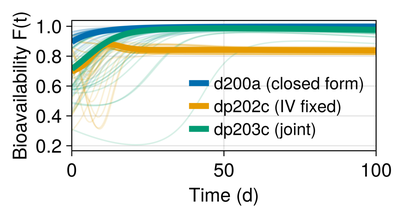



*Fig 5. F(t) by specification; faint: the 30 tune subjects. **All three rise from their starting value to a higher plateau** — the networks were free not to. Plateau (0.84 to 1.00), rate and shape differ: only the closed form is monotone, by construction; both networks reverse late, appreciably only in dp202c.*

- Both neural F(t) specifications **improved held-out likelihood** against the closed form.
- **Sequential or joint changes the answer**: holding the IV estimates fixed leaves F(t) to absorb what that locks in.
- The two shapes differ while scoring within a log-likelihood unit on tune — indicating low practical identifiability at this point.

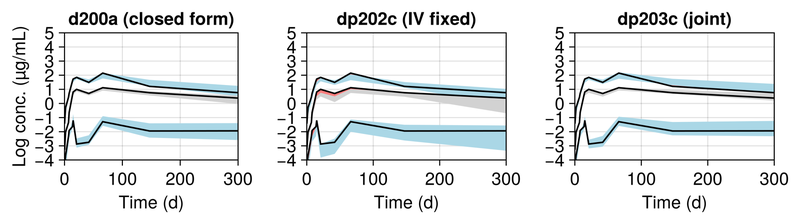



*Fig 6. SC visual predictive checks, 500 replicates, same settings as Fig 4. The closed form and the joint fit are near-indistinguishable; the **IV-fixed** fit is less well calibrated.*

## Held-out predictive performance

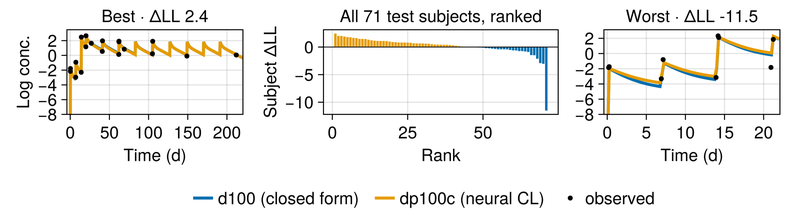



*Fig 7. Held-out test subjects (n = 71) ranked by ΔLL (neural − closed form); gold favours the neural model. The flanking panels are the two extremes of that ranking, each drawn over its own observed window.*

- **Improvement is distributed across the majority of subjects**, not driven by a small subset.
- **A likelihood difference is not a visual claim**: the worst-ranked subject differs by **11.5 LL** with the two fits barely distinguishable by eye.

**ANSWER:** Learned functions reproduced major time-varying PK behaviour with comparable-to-improved held-out performance, **with the form learned rather than specified**.

WHY IT MATTERS The flexible specification also made **structural uncertainty in F(t) visible**, a form of model diagnostics less readily available from predefined functions.

## 3 Scientific learning: does the learned function enable broader hypothesis exploration?

### Model comparison

Each hypothesis in the **mechanistic-hypotheses panel** becomes a choice of **network inputs, not a new equation**.

| Model | What the network was given | ΔLL tune | ΔLL test |
|---|---|---|---|
| d100 | closed-form reference | 0.00 | 0.00 |
| **dp100c** | t, BWT, η | **+25.75** | **+12.00** |
| dp105 | t, competing mAbs, η | +25.51 | +22.76 |
| dp101 | t, CPP, BWT, η | +20.66 | +21.66 |
| dp102 | t, CPP, mAbs, BWT, η | +21.43 | +18.08 |
| dp103 | CPP, mAbs, η (no t) | −80.85 | −55.20 |
| dp104 | mAbs, η (no t) | −80.99 | −51.69 |

*ΔLL versus the closed-form reference d100. Tune n = 66, test n = 71. t = time; BWT = body weight; CPP = predicted central concentration; mAbs = prior rituximab / obinutuzumab exposure.*

### What the comparisons show

- **Removing explicit time substantially worsened model performance.** dp104 and dp105 share every input but time, and dp104 was given the wider hidden layer (15 vs 10 units), yet without time the fit collapses by ≈ 106 LL on tune.
- **Neither predicted concentration nor prior anti-CD20 exposure displaced time** as the dominant input. dp101 and dp102 carry strictly more information than dp100c and could match it by zeroing the added inputs, yet land 4–5 LL below on tune, the scale of the fitting noise here.

This analysis **did not establish the mechanism** underlying time-varying clearance. These data show that something changes over time; they do not localise where in the model it sits.

**ANSWER:** Flexible model inputs allowed additional candidate relationships to be explored without redesigning a parametric function for each hypothesis.

WHY IT MATTERS The framework can redirect model-development effort from choosing equations toward **evaluating alternative scientific hypotheses**.

**Acknowledgements.** We thank the gRED DevSci Innovation Fund and Committee for supporting internal innovation; Tu Mai and Joy Hsu for early use-case ideation; Sushuma Gunupudi for project management support; and gRED CPLT and DSLT members for their support of the project.

## Overall

DeepNLME may **reduce effort spent searching across functional forms**, while maintaining model performance and allowing more effort to be directed toward **evaluating alternative scientific hypotheses**.

**Differences across refits make unidentifiability visible.** A fixed structural component is assumed and carries no uncertainty of its own — even where the data could not have distinguished between candidate forms with different biological implications. A learned function can show that, as heterogeneity in form at similar likelihood.

**References.**
1. Bender BC, et al. Clin Transl Sci. 2024;17(6):e13825.
2. Kuruvilla D, et al. ACoP 2025, Poster T-046.
3. Bender BC, et al. ACoP 2025, Poster S-117.
4. Budde LE, et al. J Clin Oncol. 2022;40(5):481–491.
5. Matasar MJ, et al. Am J Hematol. 2025. doi:10.1002/ajh.70225.
6. DeepPumas v0.9, Pumas-AI, Inc.; Julia 1.11.7.
7. US Food and Drug Administration. Lunsumio (mosunetuzumab-axgb) approval. 2022.
8. European Medicines Agency. Lunsumio EPAR; conditional MA 3 Jun 2022.

**Disclosures.** D. Kuruvilla, K. Yoshida and C.-C. Li are employees of Genentech, Inc. S. M. Krishnan and N. Korsbo are employees of Pumas-AI, Inc.
**Presenting author.** Niklas Korsbo · niklas@pumas.ai

**AI disclosure.** The authors used Anthropic Claude (Claude Code) as a tool throughout this work, directing it to write and run analysis and plotting code, to produce figures, and to draft and typeset this poster. The questions asked, the models specified, the judgements made and the conclusions drawn are the authors' own, and the authors take full responsibility for the content.

Scan for this poster