Post-hoc model joining with Normalizing Flows for efficient and scalable multi-endpoint PKPD

Re-couple separately developed endpoint models post-hoc by replacing their priors with a learned joint density. No joint refit, no structural change.

Niklas Korsbo[1] [1]Pumas-AI · PAGE 2026 · Abstract I-094

Joining (conditionally) independent NLME models via a **causal** bridge is already standard in sequential PK→PD modelling, with clear workflow benefits. **For coupling that is shared latent biology rather than causal**, the same separable strategy extends via a **correlational** bridge: learn a joint prior over the random effects post-hoc, leaving the structural models untouched. Independently fitted endpoint models then share information at prediction time, letting observations of one endpoint sharpen predictions of the others.

Objective

The **goal** is a joint NLME model across endpoints. In reality, joint fits don't scale, tangle workflows, and are brittle once neural components enter the structural model. Here, we explore a workaround where endpoints are fitted independently to sidestep those problems, at the cost of discarding predictive dependencies. We then restore those dependencies post-hoc by re-fitting a **joint prior** on the existing per-patient posteriors, without touching the structural models.

Theory

Two endpoints, jointly modelled

For two endpoints X and Y measured on each subject, the population marginal over the random effects η is

$$p(Y, X \mid \theta) = \int p(Y, X \mid \eta, \theta)\, p(\eta \mid \theta)\, \mathrm{d}\eta.$$

Fitting this jointly is the ideal but sometimes infeasible. We can break this joint problem into simpler, independent problems through a set of assumptions:

A1 · Cond. independence given η: $p(Y, X \mid \eta, \theta) = p(Y \mid \eta, \theta)\, p(X \mid \eta, \theta)$

$$p(Y, X \mid \theta) = \int p(Y \mid \eta, \theta)\, p(X \mid \eta, \theta)\, p(\eta \mid \theta)\, \mathrm{d}\eta$$

A2, A3 · Parameter & latent-space separation: $\theta = (\theta_Y, \theta_X)$, $\eta = (\eta_Y, \eta_X)$; each endpoint owns its own, prior still joint

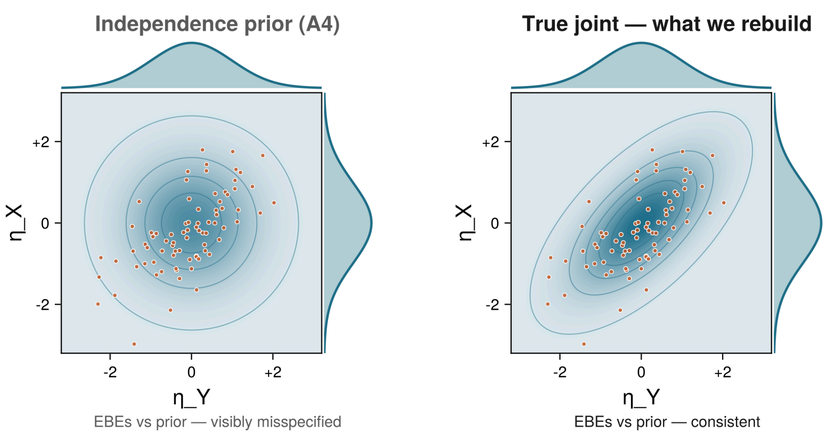
$$p(Y, X \mid \theta) = \iint p(Y \mid \eta_Y, \theta_Y)\, p(X \mid \eta_X, \theta_X)\, p(\eta_Y, \eta_X \mid \theta)\, \mathrm{d}\eta_Y\, \mathrm{d}\eta_X$$

A4 · Prior independence (the one we relax): $p(\eta_Y, \eta_X \mid \theta) = p(\eta_Y \mid \theta_Y)\, p(\eta_X \mid \theta_X)$

$$p(Y, X \mid \theta) = p(Y \mid \theta_Y) \cdot p(X \mid \theta_X)$$

Result: two endpoint marginals, fittable in isolation.

Independent prior (A4) vs. true joint



Re-coupling: backtrack A4

A1–A4 together make the structural models fittable in isolation. **A4 stands out** because it is both **impactful** (assuming zero cross-endpoint correlation is rarely benign) and **cheap to reverse**: relaxing it touches only the prior, not the structural models themselves. That makes it the natural target for a post-hoc fix.

To backtrack A4, we need a joint prior. We fit to the **joint the posteriors imply** [3]: the per-patient posterior product, averaged over the population.

$$\underbrace{\tilde{p}(\eta_Y, \eta_X \mid \theta)}_{\text{target}} = \mathbb{E}_{(Y,X)}\left[\underbrace{p(\eta_Y \mid Y, \theta_Y)}_{\text{indep. fit}}\,\underbrace{p(\eta_X \mid X, \theta_X)}_{\text{indep. fit}}\right].$$

While the components here are independent by design, the expected joint posterior is still correlated since each data pair (X_i, Y_i) comes from the same patient and reflects a shared underlying disease state. We fit a flexible density p_φ (Gaussian or normalizing flow [1, 2]) to approximate $\tilde{p}$, leaving the structural models, observation densities, and θ unchanged. ($\tilde{p}$ is the **expected-posterior** joint; it matches the generative prior only in the idealised limit.)

Supervision and fitter choice

p_φ is fitted against the per-patient pairs of Laplace posteriors that the independent fits already produce:

$$\left\{\left(p(\eta_{Y,i} \mid Y_i, \theta_Y), p(\eta_{X,i} \mid X_i, \theta_X)\right)\right\}_{i=1}^{N}.$$

The density family for p_φ is a separate choice; each fitter delivers a different **summary** of the same supervision.

Prior	Captures
Gaussian	linear cross-correlation only; assumes Gaussian marginals
Normalizing Flow	non-Gaussian, non-linear dependence; no family to specify
Other (mean-shift, …)	different fidelity–cost trade-offs

Method

Generic recipe, applied below across three experiments:

1. **Fit each endpoint model independently** under a marginal prior. When endpoints share structure (e.g. PK in a PKPD model), fit the shared component first and carry its typical values into downstream models as fixed parameters.
2. **Compute per-patient Laplace posteriors** of the random effects on the training subjects.
3. **Fit a joint prior** $p_\varphi(\eta)$ on those per-patient posteriors (here, a Gaussian or a Normalizing Flow).
4. **Plug the learned prior back in** and evaluate on a held-out test set.

Results

Experiment 1: PK + three PD endpoints (Gaussian DGM)

Synthetic four-endpoint PKPD: oral 1-cmt PK driving three turnover-type PD endpoints (Primary, AST, ALT), conditionally independent given η_{PK}. Six PD random effects from $\mathcal{N}(0, \Omega_{PD})$ with hand-specified cross-endpoint correlations (AST↔ALT: 0.7–0.8; Primary↔AST/ALT: 0.30–0.35). 200 train / 50 tune / 500 test subjects, randomised doses. Realistic per-observation noise (4–14% residual) but enough samples per subject (7–8 per endpoint) to identify the random effects well (mean η-shrinkage ≈ 6.5%). **The random-effect distribution is Gaussian by construction: the easy case for the parametric fitter.**

Structural model. PK (oral 1-cmt):

$$\mathrm{Depot}' = -K_a\ \mathrm{Depot}, \quad \mathrm{Central}' = K_a\ \mathrm{Depot} - (CL/V_c)\mathrm{Central}, \quad C_p = \mathrm{Central}\ /V_c$$

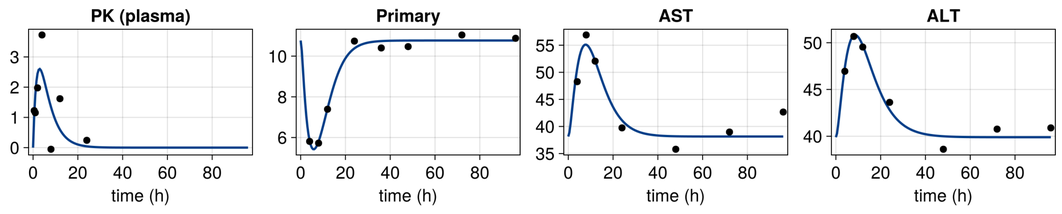
PD endpoint $e \in \{\mathrm{prim}, \mathrm{ast}, \mathrm{alt}\}$ (stimulatory turnover):

$$B_{e'} = K_{\mathrm{in},e}\left(1 + S_{\max,e}\ C_p/\left(C_p + \mathrm{SC50}_e\right)\right) - K_{\mathrm{out},e} B_e$$

Random effects $\eta_{PD} \sim \mathcal{N}(0, \Omega_{PD})$ multiply endpoint-specific parameters. Observations:

$$\mathrm{DV}_{pk} \sim \mathcal{N}(C_p, \sigma_{pk}), \quad \mathrm{DV}_e \sim \mathcal{N}(B_e, \sigma_e).$$

One representative subject: individual fit (line) over observations (dots)



Quantitative comparison (test, $n = 500$)

Model	ΔLL (nats)	% of joint
Independent prior	0.0	0%
Normalizing Flow	**337.6 ± 22.8**	**89.1%**
Gaussian (post-hoc)	**371.6 ± 22.4**	**98.0%**
Gaussian (jointly fit)	**379.0 ± 28.3**	**100%**

ΔLL: total test-LL gain over the independent prior across $n = 500$ subjects (± SE). **Normalizing Flow**: coupling flow, 2 coupling × 1 hidden, median over 5 seeds. **Gaussian (post-hoc)**: second moment of the pooled per-patient posterior samples. **Gaussian (jointly fit)**: free 6×6 Ω by MAP/FOCE, the joint fit we'd run **instead of** the modular approach. Since the DGM is Gaussian this benchmark is **correctly specified**, so its gap to the post-hoc Gaussian is the cost of sequential fitting, not structural misspecification.

Modular post-hoc joining recovers most of the jointly-fit gain on a Gaussian DGM. The Normalizing Flow pays ≈ 11 pp for flexibility it doesn't need here. The 2 pp gap between the two Gaussians is mostly the bias inherited from Laplace approximations computed under the misspecified independent prior; the joint fit iterates that bias away.

Experiment 2: 2-D toy with non-Gaussian joint prior

Two endpoints, one scalar random effect each. Per-subject true random effects (η_A, η_B) drawn from the **yang half** of the unit disk (a non-convex teardrop, zero linear correlation by symmetry). Eight noisy observations per subject per endpoint. We apply the same Gaussian and Normalizing Flow joining on the per-subject Laplace posteriors of (η_A, η_B).

Structural model (per endpoint $e \in \{A, B\}$):

$$\eta_e \sim \mathcal{N}(0, 1), \quad y_{e,t} \sim \mathcal{N}(\eta_e, \sigma^2), \quad t = 1..8$$

DGM truth: in simulation $(\eta_A, \eta_B) \sim \mathrm{Uniform}(\text{yang half})$, not the assumed Gaussian. Marginally standard normal, jointly non-convex, beyond any Gaussian prior.

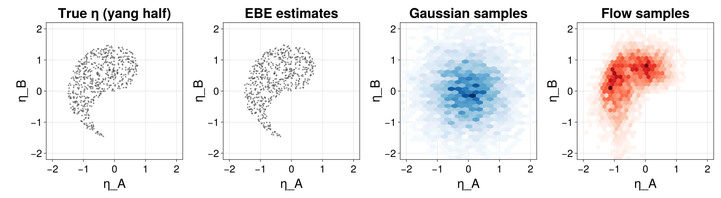


Figure 1: True η (yang half), per-subject EBEs, and samples from each joined prior. The Gaussian recovers zero linear correlation (the DGM has none) and bleeds into the empty quadrant; the Flow recovers the comma shape. Train ΔLL: Gaussian ≈ 0, Flow +512 nats.

When the latent coupling is non-Gaussian, the density family decides the outcome: the Flow recovers the teardrop and +512 nats, while the Gaussian gains nothing. The flexibility that cost the Flow ≈ 11 pp on Experiment 1's Gaussian truth is exactly what pays here, recovered from independently-fit models with the structure untouched.

Experiment 3: Sparse primary endpoint, dense biomarkers

Primary endpoint (PE): observed only at baseline and end-of-study ($t = 0$, $t = 96$ wks); linear in time with subject-specific baseline + slope. **Biomarkers** (AST, ALT, U/L): weekly through the study; inhibitory indirect-response (drug brings elevated baselines toward normal). 96-week study, Q12W × 8 doses. 200 train / 500 test subjects. Cross-endpoint coupling in the DGM ($\mathrm{cor} \approx 0.4$–0.65). Normalizing Flow architecture transferred from Exp 1.

Task: at interim ($t = 48$), with PK + AST + ALT pre-interim plus the single baseline PE observation, predict PE at $t = 96$.

Structural model. PK identical to Exp 1. PE (linear in time, no PK driving):

$$\mathrm{PE}(t) = \mathrm{PE}_{\mathrm{base}} + \mathrm{PE}_{\mathrm{slope}} \cdot t, \quad \mathrm{DV}_{\mathrm{PE}} \sim \mathcal{N}\left(\mathrm{PE}(t), \sigma_{\mathrm{prim}}\right)$$

Biomarker $e \in \{\mathrm{ast}, \mathrm{alt}\}$ (inhibitory turnover; elevated baseline):

$$B_{e'} = K_{\mathrm{in},e}\left(1 - I_{\max,e}\ C_p/\left(C_p + \mathrm{IC50}_e\right)\right) - K_{\mathrm{out},e} B_e, \quad B_{e(0)} = K_{\mathrm{in},e}/K_{\mathrm{out},e}$$

Subject-specific: $\mathrm{PE}_{\mathrm{base}} = \mathrm{tvPE}_{\mathrm{base}} \cdot e^{\eta_{\mathrm{prim},2}}$, $\mathrm{PE}_{\mathrm{slope}} = \mathrm{tvPE}_{\mathrm{slope}} + \omega_{\mathrm{PEsl}} \eta_{\mathrm{prim},1}$, $K_{\mathrm{in},e} = \mathrm{tv}K_{\mathrm{in},e} \cdot e^{\eta_{e,2}}$, $\mathrm{IC50}_e = \mathrm{tvIC50}_e \cdot e^{\eta_{e,1}}$. Joint prior: $\eta = \left(\eta_{\mathrm{prim}}, \eta_{\mathrm{ast}}, \eta_{\mathrm{alt}}\right) \sim \mathcal{N}(0, \Omega_{PD})$.

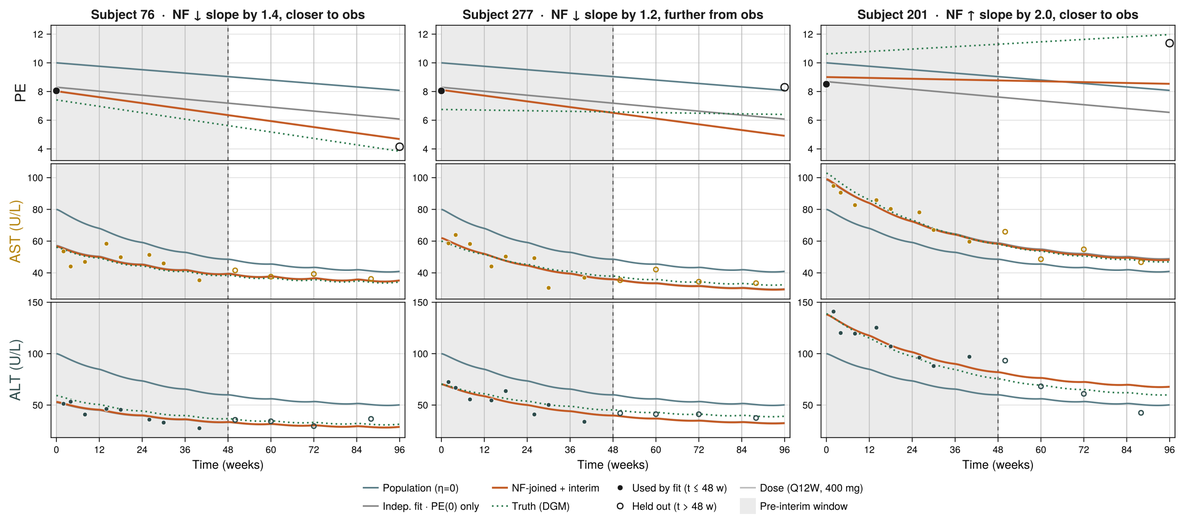


Figure 2: Three subjects spanning the Flow's per-subject effect (helper / miss / helper). Hollow markers: held out. Truth = noise-free DGM trajectory; obs = Truth + $\mathcal{N}(0, \sigma^2_{\mathrm{prim}})$.

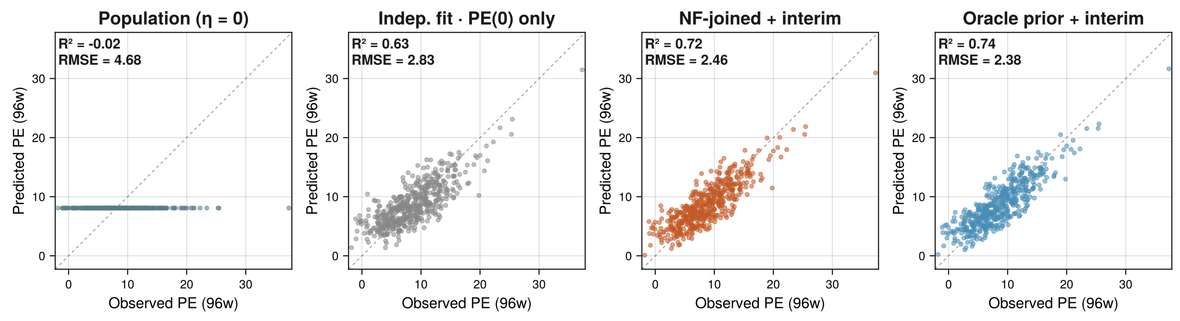


Figure 3: PE at end-of-study, 500 test subjects, panels ordered by increasing prior information. **Oracle**: EBE under the true population Ω_{PD} on the same interim data, the achievable optimum under this data scenario.

The Normalizing Flow ($R^2 = 0.72$) is within **1.8 pp of the Oracle** ($R^2 = 0.74$, the achievable optimum under interim data), even in this scenario where the biomarkers are only weakly informative about the endpoint.

Discussion

Two bridges, one workflow

Sequential modelling already trades a little correctness for workflow: fix the upstream model and propagate it downstream (PK→PD) rather than fitting simultaneously. That **causal bridge** is standard practice. The **correlational bridge** we apply here pursues the same goal for shared-latent coupling: re-fit a joint prior instead of fitting jointly. Both relax the full joint fit for the same reasons (separable, plannable, tractable).

The two are **complementary, not interchangeable**: which one suits a problem depends on what we know about the data and what we want to model, and one is rarely a drop-in replacement for the other. Each carries a real loss that is benign or damaging by situation. Sequential modelling discards upstream uncertainty and can inherit bias from the fixed upstream fit. The correlational bridge, having replaced mechanism with association, can fit the training distribution yet fail to extrapolate under intervention or a shift in what drives the data. Use the causal bridge where a mechanism links the endpoints, the correlational bridge where they share latent structure with no mechanism to exploit. We claim no general optimality. A full joint fit is the better answer when the workflow allows.

Why join post-hoc?

- **Plannability.** Each endpoint owns its protocol. Joining is a separate, restartable step.
- **Scalability.** Joint fits with neural components are slow and initialisation-sensitive. Post-hoc joining avoids them. This was the original motivation (DeepNLME [4]).
- **Reuse.** Re-couple a past-study model without touching its structure.
- **Mixed families.** Couple an NLME with any latent-variable model that exposes a tractable posterior.

The payoff: endpoint models built in isolation that still share information at prediction time, for one density fit and no structural change. In DeepPumas [5] the join itself is a single function call on the assembled models.

References. [1] Dinh et al. (ICLR 2017) · [2] Papamakarios et al. (JMLR 2021) · [3] Contento and Tarek (PAGE 2024, Abstr 11038) · [4] Korsbo et al., DeepNLME (under review) · [5] DeepPumas (Pumas-AI).

Contact. niklas@pumas.ai

Funding. PumasAI. **COI.** PumasAI sells software and services related to the poster content.

More from Pumas-AI