# Evaluating model evaluation: Shall we trust a model whose parameter estimation yields low RSEs or successful output messages?

Authors: Stanislav Tsitkov[1], Konstantinos Biliouris[2], Etienne Pigeolet[3], Ramprasad Ramakrishna[2], Yu-Yun Ho[4], Andrew Stein[2]

[1]Department of Biomedical Engineering, Columbia University, New York, USA; [2]Novartis Institutes for Biomedical Research, Cambridge, MA, USA; [3]Novartis Pharma, Basel, Switzerland; [4]Novartis Pharma, New Jersey, USA
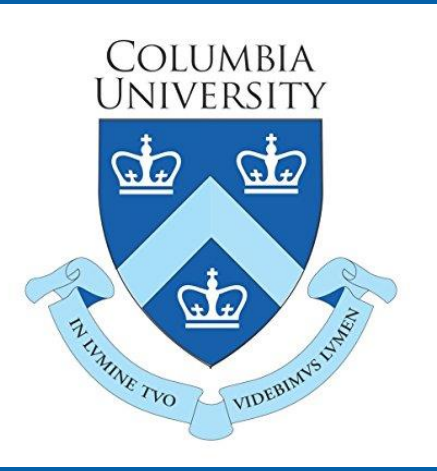
Background

- Traditionally, modelers comfortably accept a model when the NONMEM/Monolix parameter estimation comes with low relative standard errors (RSEs) and no error messages
- Achieving these outputs does not guarantee model convergence/identifiability and additional model evaluation may be needed
- Log Likelihood Profiles (LLPs) and Log Likelihood Waterfalls (LLWs) [1] can be employed for additional model evaluation
- We implement real-world examples to demonstrate the value of model evaluation implementing LLPs and LLWs

Methods

Models used

- Ten in-house or literature models [2] describing preclinical or clinical data from small molecules or antibodies were implemented in the analysis (Table 1)
- The models were built in NONMEM or Monolix. Their output included no error messages (suggesting a successful estimation) and the RSEs of all estimates were below 40%

Model evaluation

The models listed in Table 1 were evaluated using LLPs and LLWs:

- LLPs are generated by fixing one parameter at a time at different values, refitting all the other parameters, and then plotting the OFV change vs the fixed parameter. Fully identified models have a "U-shaped" LLP for each parameter, with the minimum corresponding to the parameter value that best fits the data. LLPs were automatically generated in PsN
- LLWs are created by fitting the same model 100 times with different initial estimates and then plotting the sorted OFV vs the run-number. When there is a single global optimum, the LLW is a flat line. When there are a multiple local optima, the LLW has multiple steps. When the model does not converge at all, the LLW is an increasing function wherein the OFV for each run differs

Results

Models

Our analysis demonstrated that 6/10 models were either unidentifiable or poorly converged, although all NONMEM models had covariance steps that converged and no error messages (Table 1). A model was deemed unidentifiable if at least one of its LLPs was not "U-shaped", whereas a model was considered poorly converged if a) most of its LLP minimization runs failed, or b) its LLPs had an erratic shape
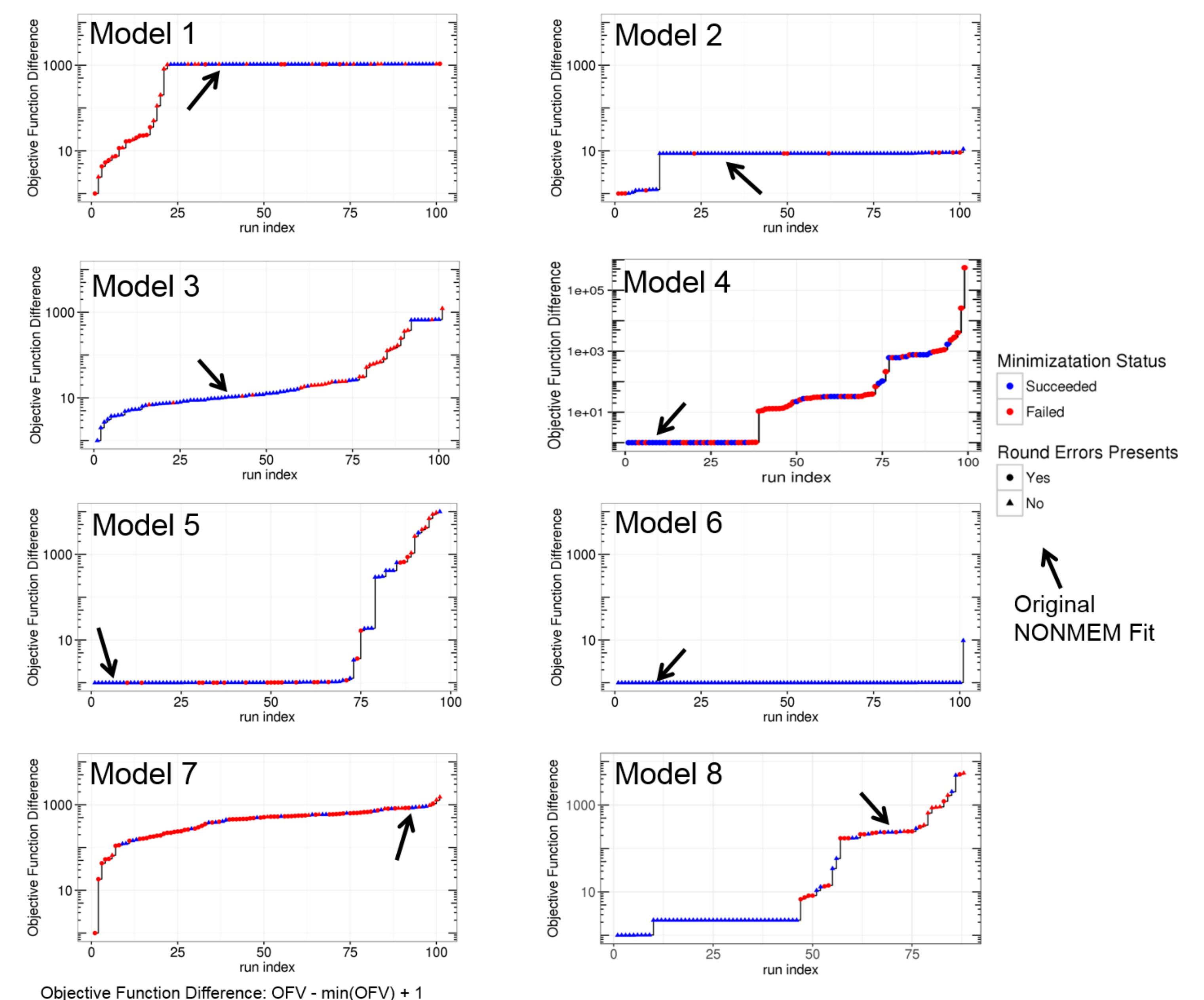
Representative LLPs
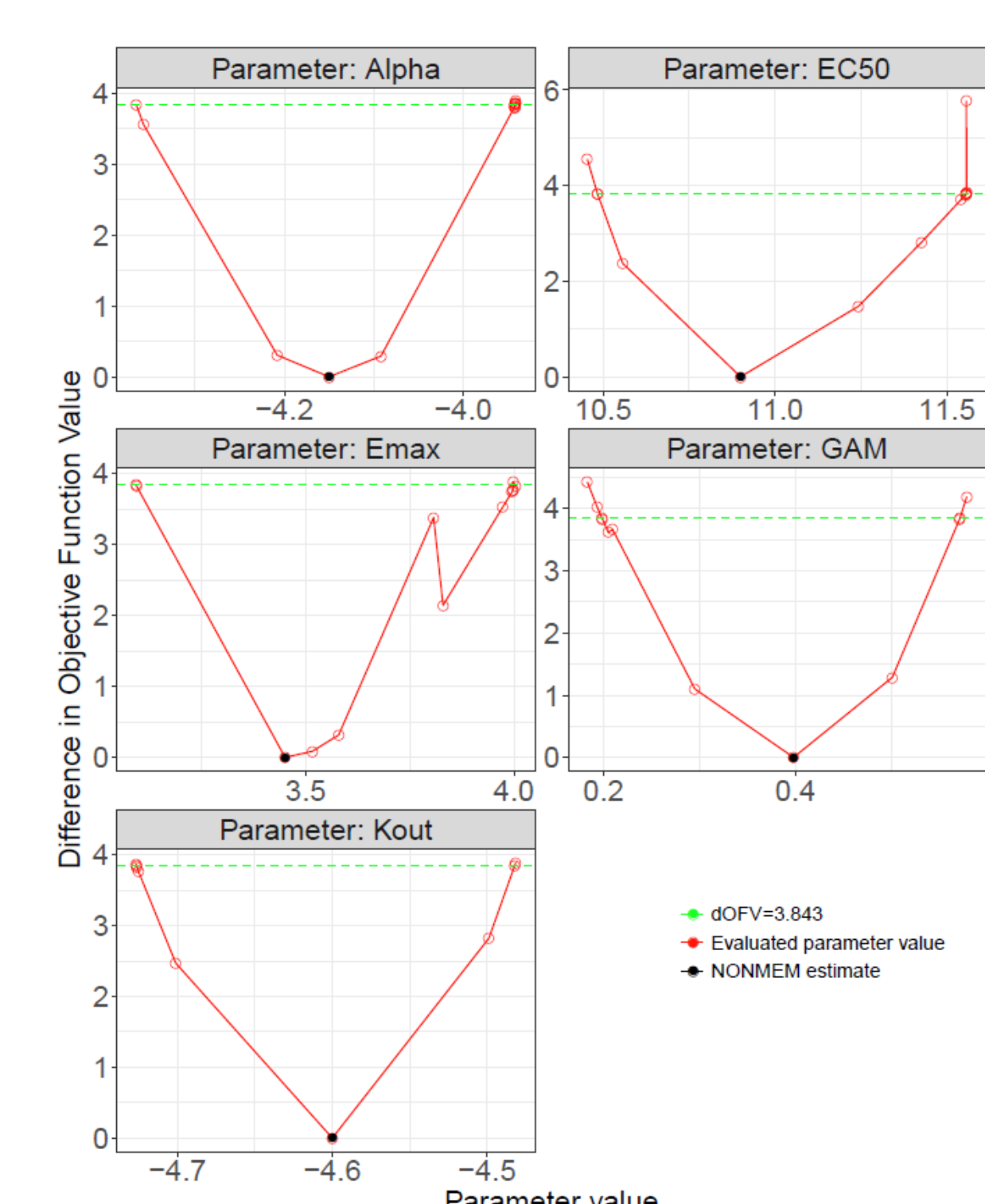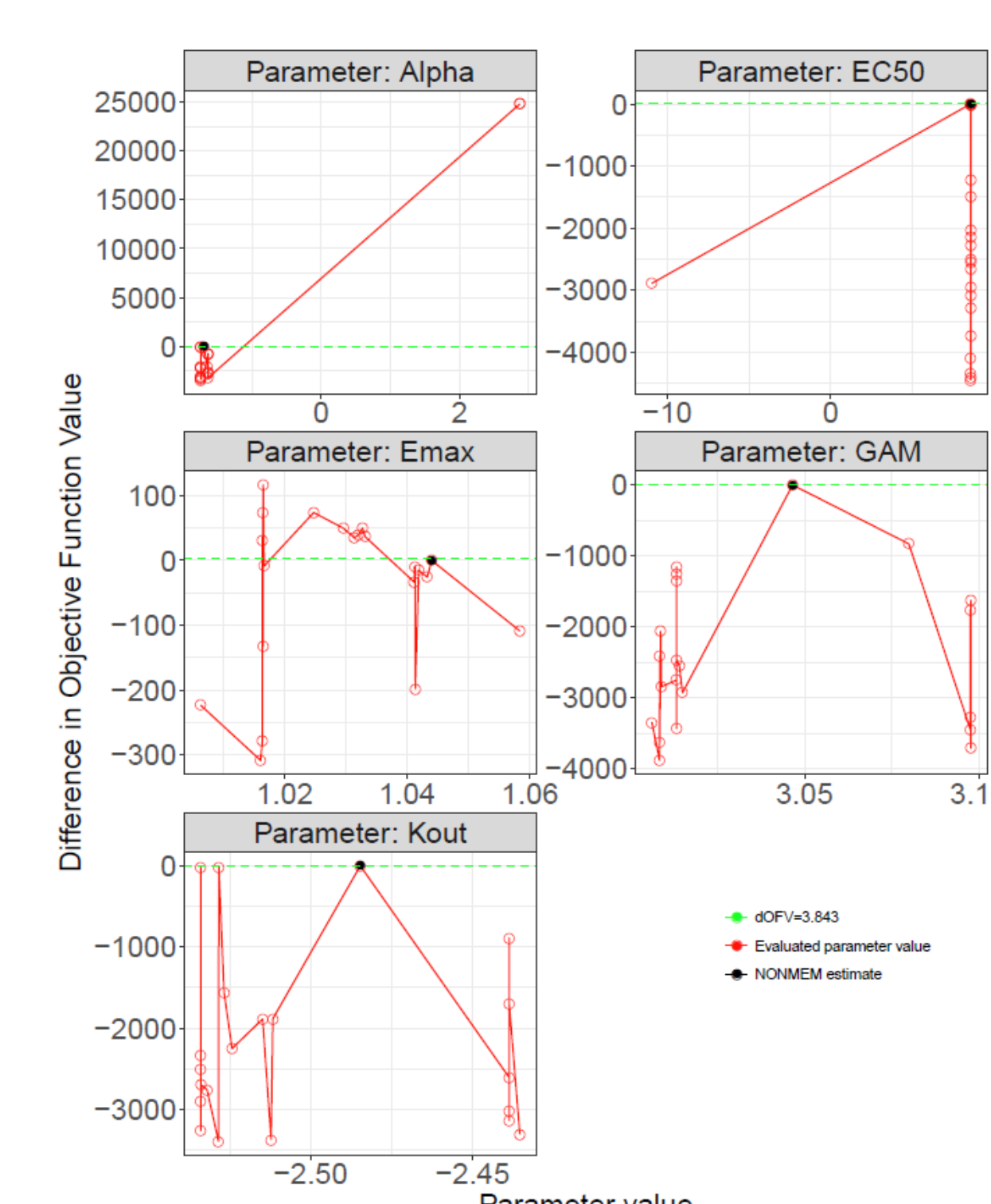
Figure 1: LLPs for model 5 (left) and 7 (right). The only difference between model 5 and model 7 is simply that the former has one additional random effect. The LLPs of model 5 are "U-shaped", while the LLPs of model 7 are not. This indicates that model 5 is identifiable but model 7 is not

LLW plots

Figure 2: LLWs of the 8 in-house models. For models 2, 5 and 6, the majority of runs converge to the same minima, indicating good convergence. For models 1, 3 and 7 most of the runs converge to a different OFV (or fail) or their LLPs had an erratic shape (data not shown), thus indicating poor convergence

Further exploration of identifiability and convergence

The condition number and the correlation between the parameters were also investigated (Table 1):

Table 1: Models used in this analysis and their corresponding information

Model type (model number)	Number of fixed effects fit (Thetas)	Number of random effects fit (Etas)	Source	Evaluation outcome	Max correlation	Condition number
Mechanistic model (1)	13	6	In-house	Poor convergence	0.998	1,000,000
Mechanistic model (2)	16	10	In-house	Unidentifiable	0.987	20,000
Custom PKPD model (3)	5	1	In-house	Unidentifiable	0.962	872
PK model (4)	5	3	In-house	Identifiable	0.95	844
Indirect response (5*)	5	1	In-house	Identifiable	0.946	191
Ligand binding (6)	14	5	In-house	Identifiable	0.946	172
Indirect response (7*)	5	2	In-house	Poor convergence	0.930	712
PK model (8)	7	3	In-house	Requires further exploration	0.794	111
Ligand binding (9)	7	0	Literature [2]	Unidentifiable	Not calculated	Not calculated
Ligand binding (10)	7	0	Literature [2]	Unidentifiable	Not calculated	Not calculated

*Models 5 and 7 are structurally identical but differ in the number of Etas

Conclusions

- Using 10 real-world examples we demonstrate how LLPs and LLWs can be used to evaluate model identifiability and convergence. The code for the LLP and LLW analysis in Monolix and Nonmem/PsN is publicly available [4]
- Implementing LLWs and LLPs in model evaluation can aid in simplifying the model and improving convergence properties (Figure 1). This not only accelerates the estimation process, but also increases the likelihood that the model converges upon potential dataset updates
- Unidentifiable models had a >0.98 correlation between parameters, whereas the condition number did not appear to be informative. Additional exploration of how identifiability is related to the parameter correlation and condition number is warranted

References

1. Raue Andreas, et al., PloS one 8.9 (2013): e74335
2. Stein Andrew, bioRxiv (2017): 123240
3. Lavielle Marc and Aarons Leon JPKPD, 43.1 (2016): 111-122
4. https://github.com/Novartis/MonolixIdent