

Design of Experiments for Model Discrimination Hybridising Analytical and Data-Driven Approaches

Simon Olofsson¹, Marc Peter Deisenroth^{1,2} and Ruth Misener¹

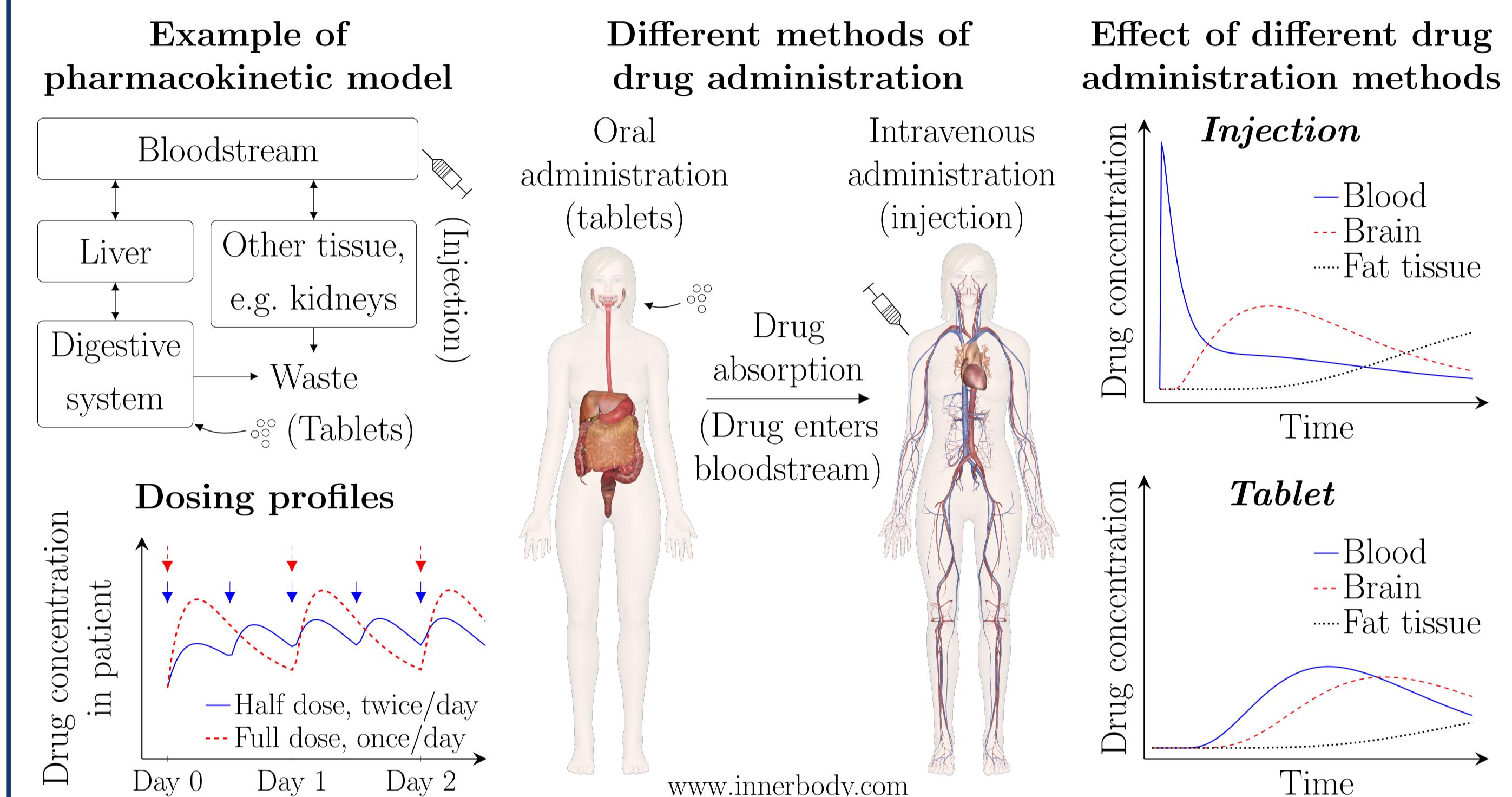
¹Dept. of Computing, Imperial College London, UK.

² PROWLER.io, UK.

1. Motivation

We are interested in complex, expensive-to-evaluate physical systems. Multiple parametric models \mathcal{M}_i may try to explain and predict the behaviour of the system [1].

Example: different models for human pharmacokinetics (how drugs are affected by processes in the human body). Interpretable parametric models are required by regulatory agencies (e.g. FDA and EMA).

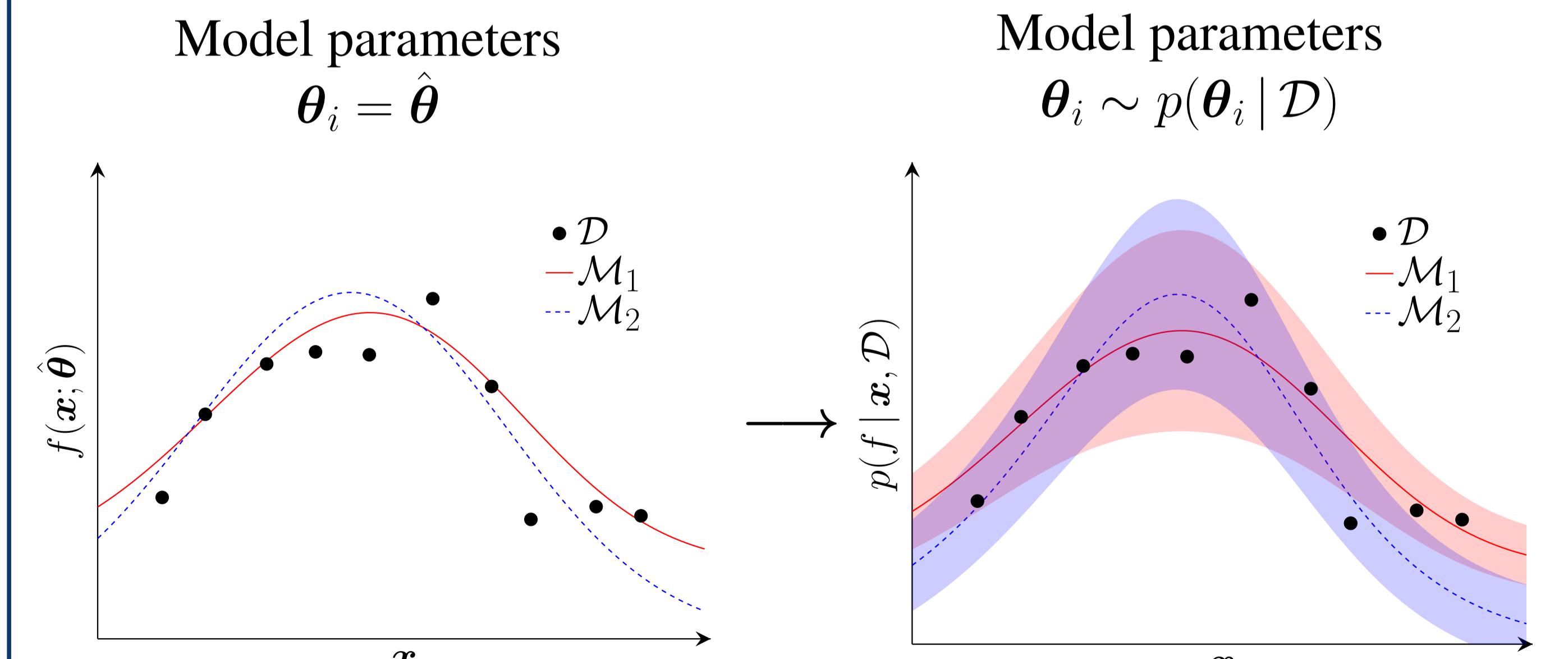


Problem: There may not be enough experimental data to discriminate between the models (i.e. find the most accurate model).

Goal: Design a new, maximally informative experiment \mathbf{x}^* .

2. Model Parameter Uncertainty

Account for the model parameter uncertainty arising from parameter estimation using noisy experimental data \mathcal{D} .



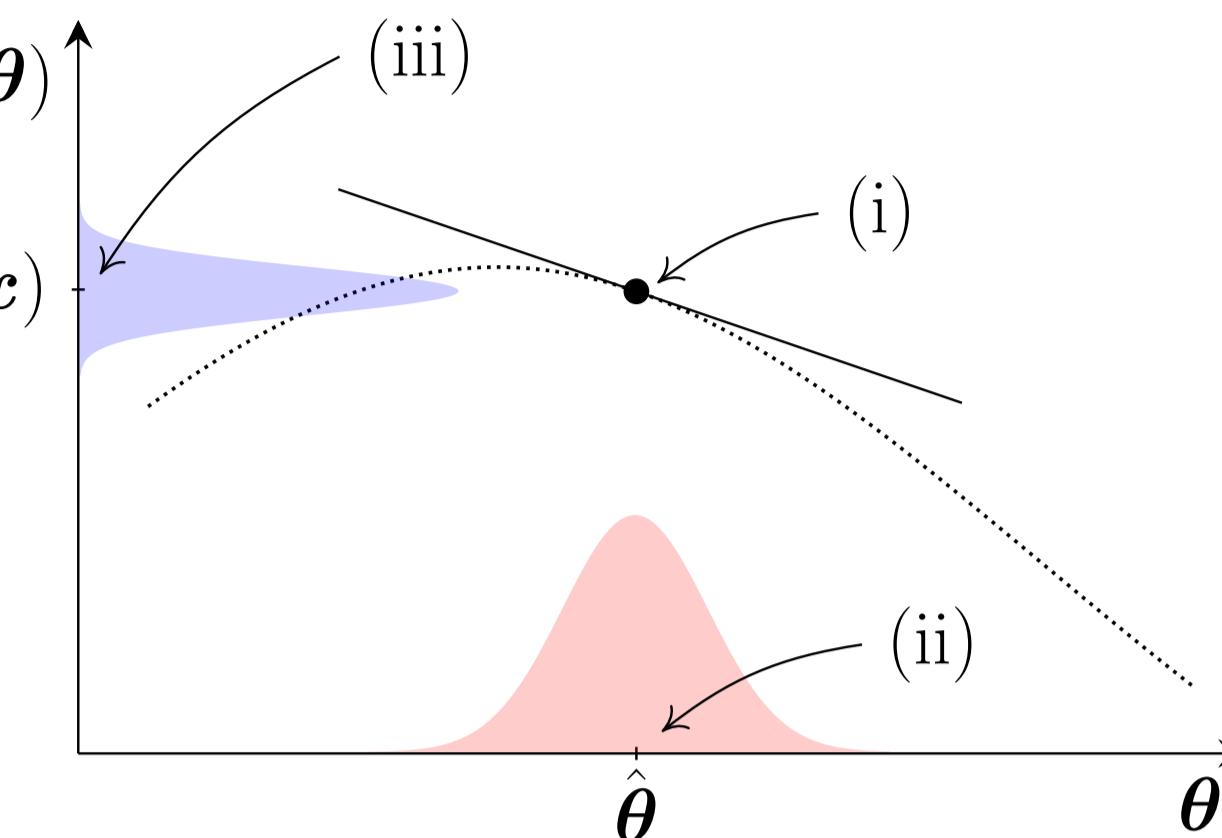
Objective: Find \mathbf{x}^* that maximises the divergence between distributions

$$p(f_i | \mathbf{x}, \mathcal{D}) = \int p(f_i | \mathbf{x}, \theta_i) p(\theta_i | \mathcal{D}) d\theta_i, \quad i = 1, \dots, M \quad (1)$$

3. Existing Methods

Analytical Approach (e.g. [2])

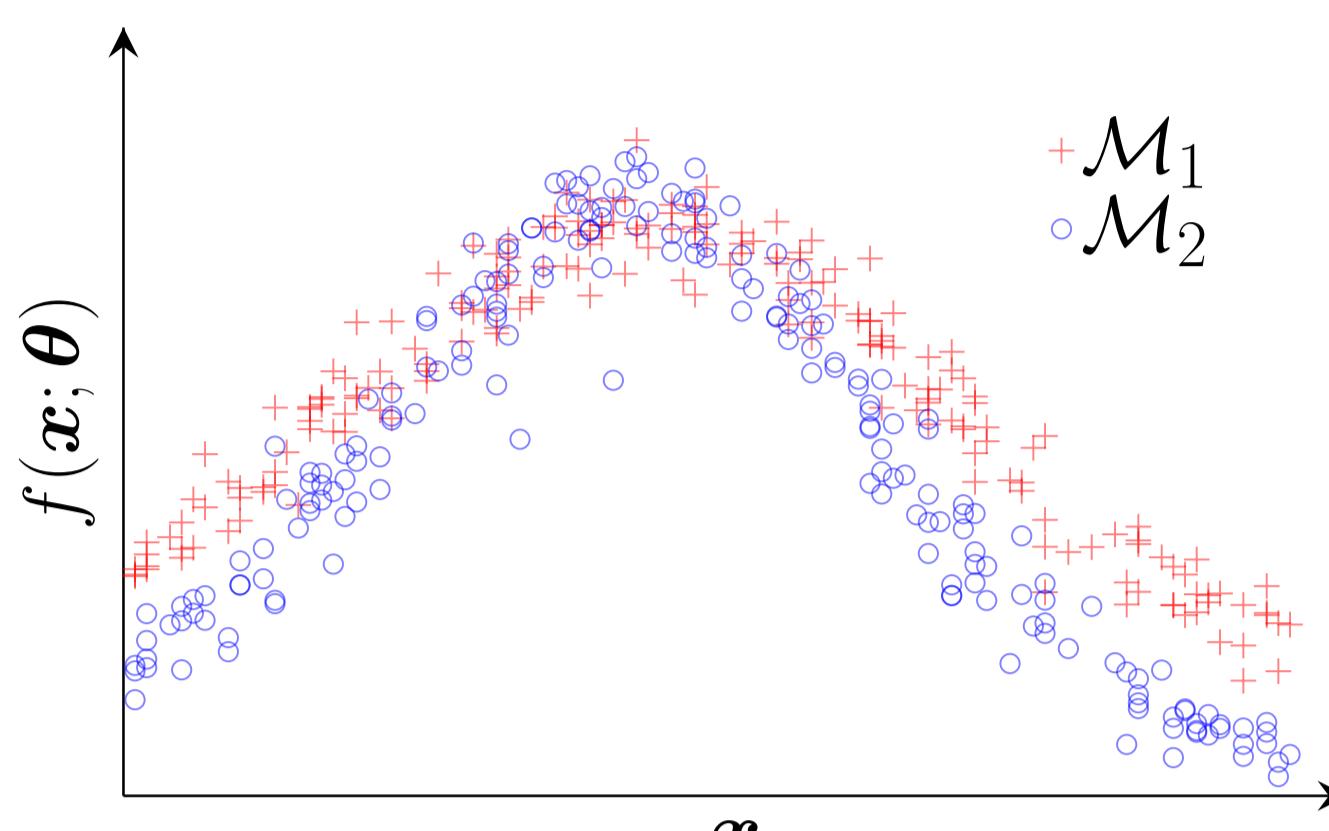
- (i) Linearise model around $\theta = \hat{\theta}$.
 - (ii) Assume $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma_\theta)$.
 - (iii) Eq. (1): $p(f | \mathbf{x}, \mathcal{D}) \approx \mathcal{N}(\check{\mu}, \check{\Sigma})$
 - (iv) Use analytical divergence criteria.
- + Closed-form, computationally cheap.
- Linearised models; requires $\partial f_i / \partial \theta_i$.



Data-Driven Approach (e.g. [3])

Monte Carlo-based methods to solve $\mathbf{x}^* = \arg \max_{\mathbf{x}} \mathbb{E}_{\mathbf{y}, \theta} [U(\mathbf{y}, \mathbf{x}, \theta) | \mathcal{D}]$ for design utility function $U(\cdot, \cdot, \cdot)$.

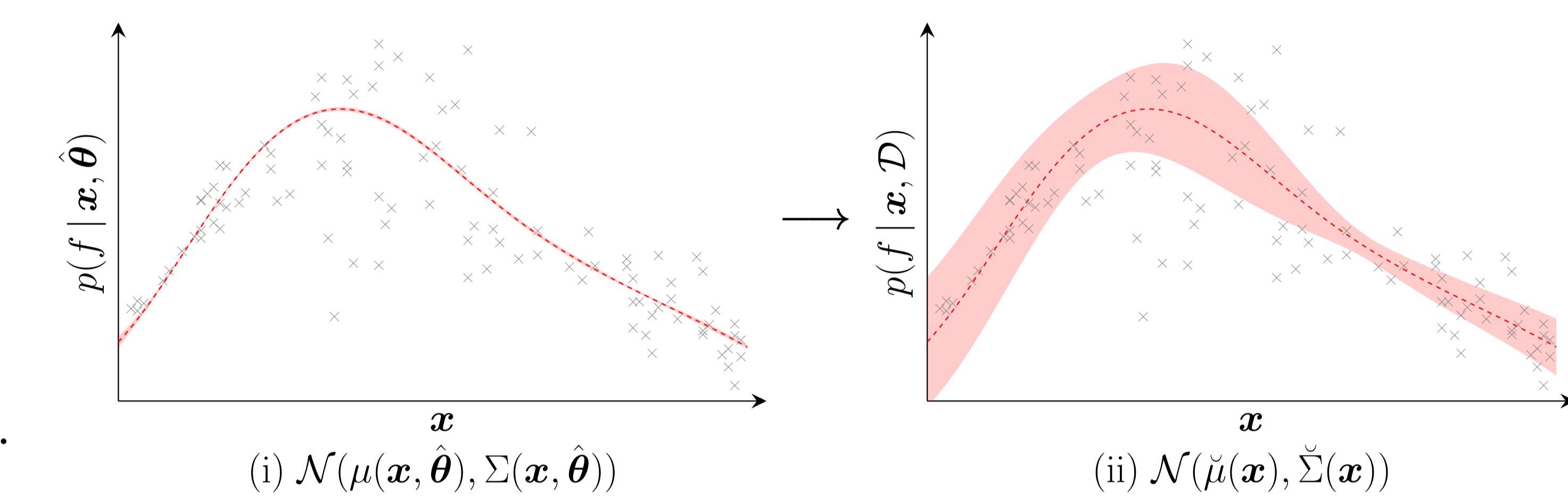
- + Accommodates black-box models.
- Computationally expensive.
- No closed-form $p(f | \mathbf{x}, \mathcal{D})$.



4. Our Approach

Hybridising the Analytical and Data-Driven Approaches

- (i) Train GP surrogate models on data from model evaluations, with sampled \mathbf{x} and θ .
- (ii) Apply approximations from analytical approach to GP surrogate models.



5. Results

Comparing to the 'exact' analytical approach:

- $j = 1, \dots, 12$ (analytical) case studies.
- $\ell = 1, \dots, N$ trials for each case study j .

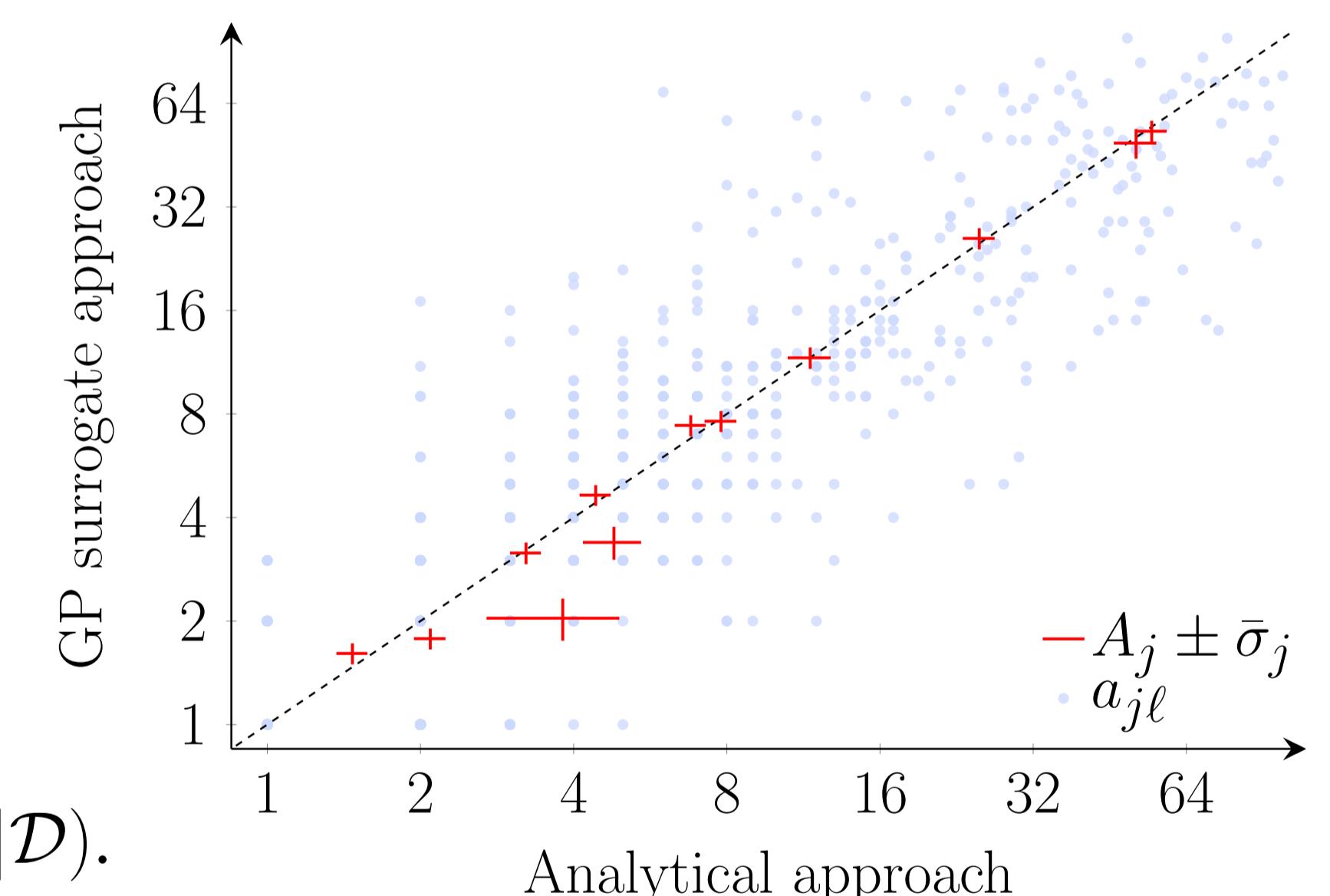
$a_{j\ell}$: extra experiments needed to identify the data-generating model in trial ℓ of case study j .

A_j : average of $\{a_{j1}, \dots, a_{jN}\}$.

$\bar{\sigma}_j$: standard error $\text{std}(a_{j1}, \dots, a_{jN})/\sqrt{N}$

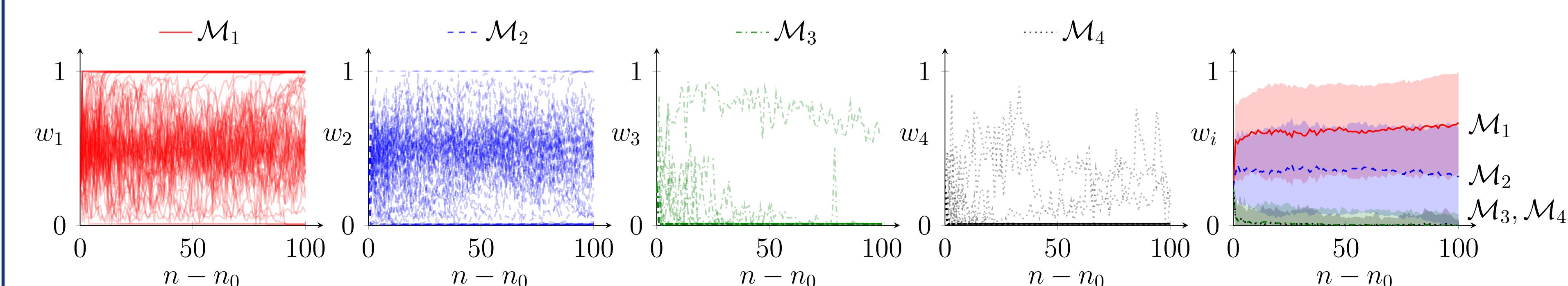
n_0 : Number of initial data points.

w_i : Normalised posterior ratio $p(\mathcal{M}_i | \mathcal{D}) / \sum_j p(\mathcal{M}_j | \mathcal{D})$.



Result	Criterion	With \mathcal{M}_2			Without \mathcal{M}_2		
		[2]	[4]	[5]	[2]	[4]	[5]
A_j		20.1	39.8	29.6	15.8	21.9	9.7
Correct [%]		15.9	9.5	33.3	89.5	77.2	95.6
Incorrect [%]		7.9	0.0	7.9	6.1	0.9	1.8
Draw [%]		76.2	90.5	58.7	4.4	21.9	2.6

A more difficult, non-analytical case study. \mathcal{M}_1 and \mathcal{M}_2 are almost indistinguishable. Our approach successfully avoids introducing overconfidence. If \mathcal{M}_2 is removed from the model set, \mathcal{M}_1 is correctly identified as the data-generating model.



6. Conclusions

- We use GP surrogates to hybridise analytical and data-driven approaches to design of experiments for model discrimination.
- Trade-off between accuracy and computational complexity.
- Results: GP surrogate approximation performs well; hybridised approach practically effective.

References

- [1] S. P. Asprey and S. Macchietto, *Comput Chem Eng*, vol. 24, 2000.
- [2] G. E. P. Box and W. J. Hill, *Technometrics*, vol. 9, no. 1, 1967.
- [3] J. Vanlier *et al.*, *Bioinformatics*, vol. 28, no. 8, 2012.
- [4] G. Buzzi-Ferraris *et al.*, *Chem Eng Sci*, vol. 45, no. 2, 1990.
- [5] C. Michalik *et al.*, *Ind Eng Chem Res*, vol. 49, 2010.

Contact

E-mail: simon.olofsson15@imperial.ac.uk
Website: <https://www.doc.ic.ac.uk/~so2015/>

Acknowledgement. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no.675251, and an EPSRC Research Fellowship to R.M. (EP/P016871/1).

