Université de Montréal

Développement d'un processus d'analyse d'expériences dose-réponse par inférence Bayésienne et application à de larges jeux de données

par

Caroline Labelle

Département de biochimie et médecine moléculaire
Faculté de médecine

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Bio-informatique

31 octobre 2023

Université de Montréal

Faculté de médecine

Cette thèse intitulée

Développement d'un processus d'analyse d'expériences dose-réponse par inférence Bayésienne et application à de larges jeux de données

présentée par

Caroline Labelle

a été évaluée par un jury composé des personnes suivantes :

Morgan Craig
(président-rapporteur)

Sébastien Lemieux
(directeur de recherche)

Anne Marinier
(codirecteur)

Paul François
(membre du jury)

Olivier Lichtarge
(examinateur externe)

Rafaël Najmanovich
(représentant du doyen de la FESP)

À JF, Max et Arthur
mon phare, mon modèle de persévérance et mon réconfort

Résumé

Dans le contexte du processus de découverte de médicaments, divers composés chimiques sont développés, testés et optimisés dans l'optique d'identifier de nouvelles thérapies efficaces pour un contexte médical précis. L'efficience de ces composés se caractérise, entre autres, via des expériences de type dose-réponse. Les expérimentateurs filtrent et sélectionnent les meilleurs composés sur la base des métriques d'efficience obtenues, telles que l'IC_{50}/EC_{50} et la réponse à haute concentration (HDR).

Traditionnellement, les valeurs des métriques d'efficience sont estimées en ajustant les paramètres du modèle log-logistique à des données expérimentales. Je désigne cette approche par Levenberg-Marquardt, soit l'algorithme le plus couramment implémenté pour une régression non-linéaire par descente de gradient. Bien que Levenberg-Marquardt soit le standard dans l'analyse des expériences dose-réponse, il présente la principale limitation de ne pouvoir évaluer ou quantifier adéquatement l'incertitude des estimations des valeurs des métriques d'efficience. Cela a un impact particulièrement néfaste lorsque des réponses incomplètes ou plates sont analysées: les métriques estimées sont incorrectes et les expérimentateurs ne sont pas outillés pour en faire l'identification rapide. Ceux-ci doivent souvent se rabattre sur des évaluations visuelles des réponses, une approche peu efficace lorsque plusieurs expériences sont considérées et difficile à reproduire d'un expérimentateur à l'autre. Il existe donc un important besoin pour une méthodologie robuste et accessible qui tienne compte de l'incertitude découlant des données expérimentales et qui soit apte à quantifier l'incertitude sous-jacente des mesures d'efficacité.

La présente thèse vise à mieux outiller les expérimentateurs dans leurs processus d'analyse d'expériences dose-réponse et de prise de décisions. Pour ce faire, je propose un processus d'analyse par inférence bayésienne: les métriques d'efficience sont dès lors représentées par des distributions des valeurs les plus probables, soit des *posteriors*. Les *posteriors* représentent explicitement l'incertitude découlant des variabilités biologique, expérimentale et analytique. L'intégration de *priors* rend le processus d'inférence robuste aux expériences incomplètes ou plates, contrairement à Levenberg-Marquardt. Je démontre cette robustesse qualitativement et quantitativement via une comparaison des représentations (c.-à-d. *posterior* et estimation) pour des paires de réplicats biologiques provenant de trois larges jeux

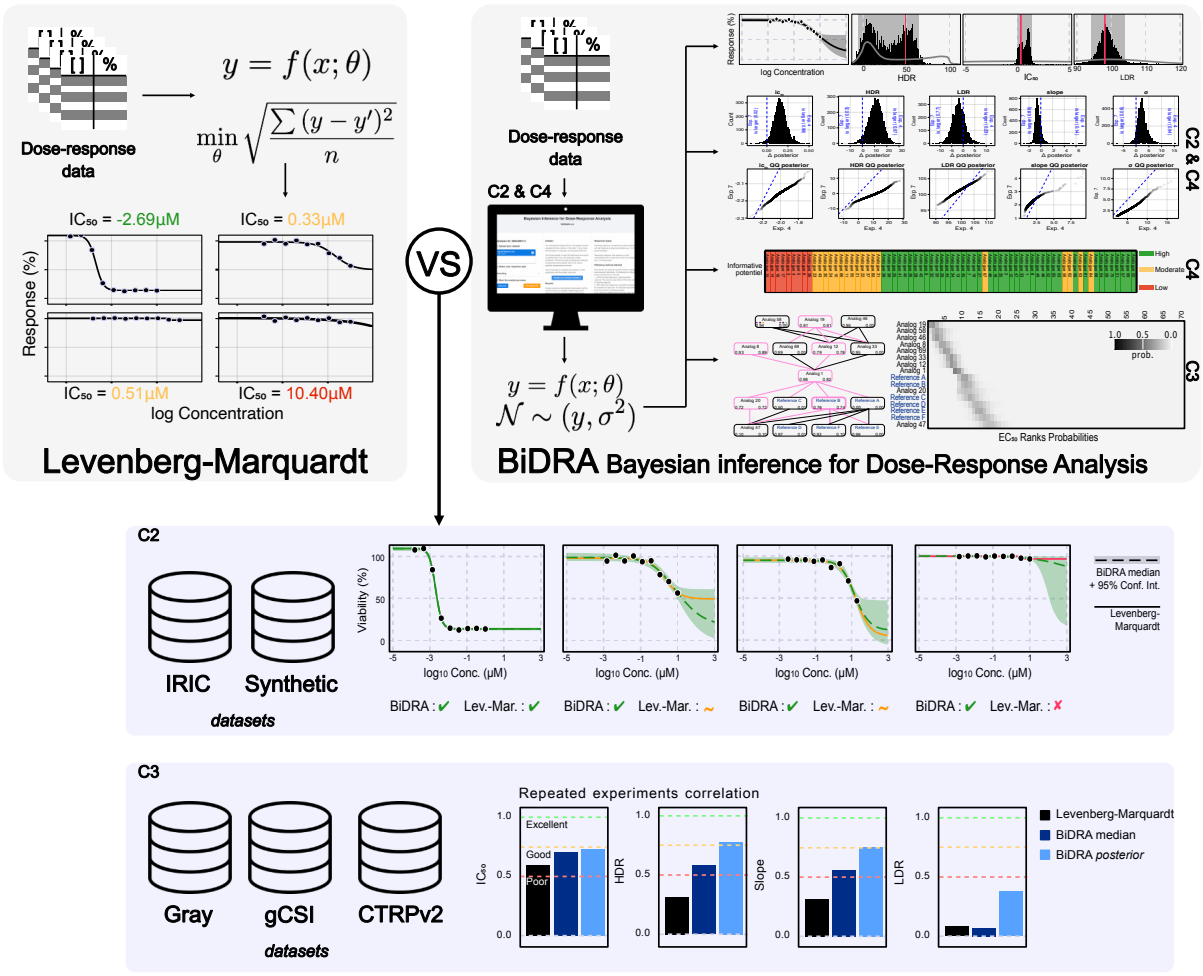


Fig. 1. Résumé illustré des travaux de thèse.

de données publics. Parallèlement à la nouvelle méthodologie proposée, je démontre pour une première fois quantitativement les lacunes de Levenberg-Marquardt. Je propose aussi diverses analyses post-inférence tirant tout le potentiel informatif des *posteriors*. Celles-ci sont plus flexibles, informatives et statistiquement valables que les analyses faites sur les estimations de Levenberg- Marquardt. Finalement, j'ai intégré le processus d'inférence et d'analyses post-inférence dans deux versions d'une interface web (BiDRA: Bayesian inference for the Analysis of Dose-Response) outillant ainsi de façon conviviale les expérimentateurs.

Mes travaux de thèse proposent une alternative robuste et accessible aux lacunes de Levenberg- Marquardt dans le contexte de la caractérisation de l'efficience de composés chimiques. De plus, les différentes démonstrations ouvrent la voie à l'intégration de l'inférence

Bayésienne pour divers types d'expériences dans le contexte du processus de découverte de médicaments.

Mots clés: Découverte de médicaments, Dose-réponse, SAR, Métriques d'efficience, Incertitude, Levenberg-Marquardt, Inférence Bayésienne, MCMC

Abstract

In the context of drug discovery, various chemical compounds are developed, tested and optimized with the aim of identifying new effective therapies for a specific medical context. The efficiency of these compounds is characterized by efficiency metrics, such as potency and efficacy (HDR), that are derived from the analysis of dose-response experiments. Experimenters filter and select compounds based on these efficiency metrics.

Traditionally, the values of efficiency metrics are estimated by adjusting the parameters of the log-logistic model to experimental data. I refer to this approach as Levenberg-Marquardt, the most commonly implemented algorithm for non-linear regression by gradient descent. Although Levenberg-Marquardt is the standard in the analysis of dose-response experiments, it presents the main limitation of not being able to adequately evaluate or quantify the uncertainty of efficiency metrics. This has a particularly harmful impact when incomplete or flat responses are analyzed: the estimated metrics are incorrect and experimenters are not equipped to quickly identify them. They often have to fall back on visual evaluations of responses, an approach that is not very effective when several experiments are considered and difficult to reproduce from one experimenter to another. There is thus a dire need for a robust and accessible methodology that account for uncertainty arising from the experimental data and is able to quantify the underlying uncertainty of efficiency metrics.

This thesis aims to provide experimenters with better tools for analyzing dose-response experiments and making decisions. To do this, I propose a Bayesian inference analysis process: the efficiency metrics are now represented by distributions of the most probable values, i.e. *posteriors*. The *posteriors* explicitly represent the uncertainty arising from biological, experimental and analytical variabilities. The integration of *priors* makes the inference process robust to incomplete or flat experiments, unlike Levenberg-Marquardt. I demonstrate this robustness qualitatively and quantitatively via a comparison of representations (i.e., *posterior* and estimation) for pairs of biological replicates from three large public datasets. Alongside the new proposed methodology, I demonstrate for the first time quantitatively the shortcomings of Levenberg-Marquardt. I also propose various post-inference analyzes drawing all the informative potential of the *posteriors*. Such analyzes are more flexible, informative and statistically valid than analyzes done on Levenberg-Marquardt estimates.

Finally, I integrated the process of inference and post-inference analyzes into two versions of a web interface (BiDRA: Bayesian inference for the Analysis of Dose-Response) thus providing tools in a user-friendly manner for experimenters.

My thesis work proposes a robust and accessible alternative to the shortcomings of Levenberg-Marquardt in the context of characterizing the efficiency of chemical compounds. Additionally, the various demonstrations pave the way for the integration of Bayesian inference to the analysis various types of experiments in the context of the drug discovery process.

Keywords: Drug Discovery, Dose-response, SAR, Efficiency metrics, Uncertainty, Levenberg-Marquardt, Bayesien inference, MCMC

Table des matières

Résumé 5

Abstract 9

Liste des tableaux 15

Liste des figures 17

Liste des sigles et des abréviations 19

Remerciements 23

Chapitre 1. Introduction 25

1.1. La découverte de médicaments 25

1.2. Les expériences de type Dose-Réponse 27

1.2.1. Les expériences et leur protocole expérimental 28

1.2.2. La caractérisation de l'efficience d'un composé 29

1.2.3. Les outils et jeux de données 32

1.2.4. Les limitations de la caractérisation standard des métriques d'efficience 34

1.3. L'inférence Bayesienne 37

1.3.1. Le théorème de Bayes 38

1.3.2. MCMC: méthodes Monte-Carlo par chaîne de Markov 39

1.3.3. Programmation probabiliste 45

1.4. L'inférence bayésienne et l'analyse du dose-réponse 47

1.4.1. But et objectifs 48

Chapitre 2. Article 1 - Enhancing the drug discovery process: Bayesian inference for the analysis and comparison of dose–response experiments 51

2.1. Introduction 53

2.1.1. Dose-response experiments 53

2.1.2. Marquardt-Levenberg ... 55
2.1.3. Bayesian Inference ... 55
2.1.4. Objectives ... 56
2.2. Methods ... 57
2.2.1. Inferring a Dose-Response Curve ... 57
2.2.2. Comparing Two Dose-Response Curves ... 58
2.2.3. Implementation ... 59
2.2.4. Dose-Response Data ... 59
2.3. Results and Discussion ... 60
2.3.1. Bayesian Inference on Dose-Response Data ... 60
2.3.1.1. Marquardt-Levenberg vs. Bayesian Inference ... 60
2.3.1.2. Defining *prior* distributions ... 63
2.3.1.3. Unresponsive Data ... 65
2.3.1.4. Inferring noise ... 66
2.3.2. Comparison of Two-Dose-Response Datasets ... 67
2.3.3. BiDRA: an Online Tool ... 69
2.4. Implications ... 70
Acknowledgements ... 72
Funding ... 72
Chapitre 3. Article 2 - Bayesian Inference as a Robust Alternative to Non-Linear Regression for Dose-Response Efficiency Metrics Assessment ... 73
3.1. Introduction ... 74
3.2. Results ... 75
3.2.1. BiDRA model and *priors* ... 75
3.2.2. Response consistency across replicates ... 75
3.2.3. Efficiency metrics conscistency across replicates ... 77
3.2.4. Control experiments ... 80
3.2.5. BiDRA's robustness ... 81
3.2.6. Post-inference analysis: compounds selection ... 82
3.2.7. Other definitions of replicates ... 84
3.3. Discussion ... 85

3.4. Methods ... 88
3.4.1. Data ... 88
3.4.2. Compounds Efficiency Metrics ... 89
3.4.2.1. Dose-response model ... 89
3.4.2.2. Levenberg-Marquardt: Estimating Efficiency Metrics ... 89
3.4.2.3. BiDRA: Inferring Bayesian *Posteriors* of Efficiency Metrics ... 90
3.4.3. Assessing concordance of biological replicates efficiency metrics ... 91
3.4.3.1. Identifying replicated experiments ... 91
3.4.3.2. Correlation coefficients ... 92
3.4.3.3. Correlation between responses of biological replicates ... 92
3.4.3.4. Correlation between Levenberg-Marquardt efficiency metrics estimates of biological replicates ... 93
3.4.3.5. Correlation between BiDRA efficiency metrics posteriors of biological replicates ... 93
3.4.3.6. Control experiment ... 94
3.4.4. Robustness evaluation ... 94
3.4.4.1. Response set completeness ... 94
3.4.4.2. IC_{50} and HDR ... 94
3.4.5. Typical Application of BiDRA: SAR Analysis and Compounds Selection ... 95
3.4.5.1. Compound ranking ... 95
3.4.5.2. DAG representation ... 95
3.4.5.3. Compounds selection ... 95

3.5. Supplementary ... 97
3.5.1. Efficiency metrics consistency across replicates ... 97
3.5.2. Area Above the Curve (AAC) consistency across replicates ... 98
3.5.3. Control experiments ... 103
3.5.4. BiDRA's robustness ... 104
3.5.5. Other definitions of replicates ... 105

Chapitre 4. Outiller davantage les expérimentateurs : évaluation du potentiel informatif d'une expérience et mise à jour de l'interface BiDRA ... 109

4.1. Prémisse de l'évaluation du potentiel informatif ... 109

4.2. Évaluation et comparaison de modèles bayésiens ... 111

4.3. Modèles comparés et données . . . 113
4.4. Assignation d'un sigle décrivant le potentiel informatif d'une expérience . . . 113
4.4.1. Validation du processus comparatif . . . 114
4.4.2. Définition de groupes de potentiel informatif . . . 115
4.4.3. Limitations . . . 122
4.4.4. Généralisation . . . 125
4.5. BiDRA V2 . . . 127
4.5.1. Implémentation . . . 128
4.6. Conclusion . . . 129
Chapitre 5. Discussion . . . 133
5.1. Le choix des jeux de données . . . 133
5.2. L'implémentation de BiDRA . . . 136
5.3. Le choix et les implications des *priors* . . . 139
5.4. La quantification de l'incertitude . . . 141
5.5. Évaluation du potentiel informatif des *posteriors* . . . 145
5.6. Accessibilité . . . 149
5.7. Conclusion: Implications et Perspectives . . . 151
Références bibliographiques . . . 155

Liste des tableaux

1 Synthetic datasets 59
2 Distributions Parameters 64

Liste des figures

1 Résumé illustré des travaux de thèse 6

2 Visualisation sommaire des étapes du processus de découverte de médicament (DDP) 27

3 Visualisation d'une courbe dose-réponse et de ses métriques d'efficience 31

4 Exemples de courbes dose-réponse et de leurs métriques d'efficience 35

5 Équation annotée du théorème de Bayes 38

6 Dose-response curve and efficacy metrics 54

7 Marquardt-Levenberg vs. Bayesian inference 61

8 *Prior* informativeness 63

9 Effects of various *prior* 64

10 Bayesian inference applied to seemingly unresponsive experimental data 66

11 A *posterior* distribution of σ 67

12 Comparison of synthetic datasets 68

13 Comparison of experimental datasets 69

14 Datasets and Bayesian model overviews 76

15 Response consistency across biological replicates 77

16 Efficiency metrics consistency across biological replicat 78

17 BiDRA's robustness across diverse types of responses 79

18 Control experiment: consistency assessment of efficiency metrics across randomly paired experi-ments 81

19 Analysis of SAR: compound selection using *posteriors* 83

20 Pearson correlation and coefficients comparisons 97

21 Concordance of the asymptotic basal responses between biological replicates 99

22 Assessment of common experimental dose for pairs of biological replicates from the gCSI dataset. 100
23 Concordance of area above the curve (AAC) between biological replicates 101
24 Exploring special cases of AAC calculation for Levenberg-Marquardt and BiDRA 102
25 Pearson correlations for control experiments. 103
26 Comparison of IC_{50} and HDR estimates and their *posterior* for Gray and CTRPv2 104
27 Comparison of HDR estimates and *posterior* uncertainty to curve completeness . 105
28 Assessment of efficiency metrics correlations for within dataset multi-replicates and across datasets singletons. 106
29 Visualisation des distributions *posterior* prédictives (ppd) des modèles BiDRA et Line appliqués aux données des Analogues 1 (réponse sigmoïde) et 4 (réponse plate) 114
30 Assignation des sigles de potentiel informatif des expériences, du jeu de données IRIC et identification d'expériences exemples. 116
31 Comparaison de métriques pour la définition des groupes et l'assignation des sigles de potentiel informatif 118
32 Analyse et comparaison des pénalités de $WAIC_K$ et WAIC pour les modèles BiDRA et Line 119
33 Analyse de la constance du processus d'assignation de sigle de potentiel informatif pour diverses métriques 123
34 Courbes dose-réponse du jeu de données IRIC. 124
35 Généralisation de l'approche d'assignation de sigles au jeu de données gCSI 126
36 Résultats types pour toutes les analyses retournées par l'interface BiDRA V2 . . . 129
37 Résultats types pour les analyses de jeux de données contenant plus d'une expérience retournés par l'interface BiDRA V2. 130

Liste des sigles et des abréviations

AAC	Air au-dessus de la courbe dose-réponse, de l'anglais *Area Above the dose-response Curve*
ADME	Absorption, Distribution, Métabolisme et Excrétion
ATP	Adénosine TrisPhosphate
AUC	Air sous la courbe dose-réponse, de l'anglais *Area Under the dose-response Curve*
BiDRA	*Bayesian nference for Dose-Response Analysis*
BRET	Transfert d'énergie par résonance bio-luminescente, de l'anglais *Bioluminscence Resonance Energy Transfer*
CCL	Lignées cellulaire cancéreuse, de l'anglais *Cancer Cell Lines*
CDF	Fonction de distribution cumulative, de l'anglais *Cumulative Distribution Function*
DAG	Graph orienté acyclique, de l'anglais *Directed Acyclic Graph*

DDP	Processus de d.couverte de médciament, de l'anglais *Drug Discovery Process*
DSS	*Drug Sensitivity Scoring*
elppd	*Expected log poitwise predictive density*
GPCR	Récepteurs couplés aux protéines G, de l'anglais *G-Protein-Coupled Receptors*
HDR	Réponse à haute concentration, de l'anglais *High-Dose Response*
HTS	Criblage à haut débit, de l'anglais *High-Thoughpout Screen*
LDR	Réponse à basse concentration, de l'anglais *Low-Dose Response*
LOESS	Régression locale, de l'anglais *Locally Estimated Scatterplot Smoothing*
LPP	Langage de Programmation Probabiliste
lppd	*Log-pointwise predictive density*
MCMC	Monte Carlo par chaînes de Markov, de l'anglais *Markov Chain Monte Carlo*

NLP	Négatif du logarithme des *posteriors*, de l'anglais *Negative Log Posterior*
PDF	Fonction de densité, de l'anglais *Probability Density Function*
PP	Programmation Probabiliste
ppd	Distribution *posterior* prédictive, de l'anglais *predictive posterior distributions*
PSRF	Facteur de réduction d'échelle potentiel *Potentiel Scale Reduction Factor*
QQ	*Quantile-to-Quantile*
RMSE	Erreur quadratique moyenne, de l'anglais *Root-Mean-Square Error*
SAR	Relation activité-structure, de l'anglais *Structure-Activity Relationship*
SMILES	*Simplified Molecular-Input Line Entry System*
TPP	Profilage terminque du protéome, de l'anglais *Thermal Proteome Profiling*
WAIC	*Widely Available Information Criterion*

Remerciements

Cette thèse de doctorat est le résultat de plusieurs années de travail, des années remplies de journées roses comme grises. Je tiens à remercier tous celles et ceux qui m'ont accompagnée, de proche comme de loin, dans cet unique périple. Vous m'avez soutenue et encouragée; grâce à vous, j'ai grandi et je termine ce chapitre de ma vie avec fierté. Vous avez fait une différence et je tiens à vous partager ma gratitude: merci.

À **Sébastien Lemieux**, mon directeur de thèse, merci pour ta patience, ton écoute et ton ouverture. Travailler sous ta supervision restera une des expériences les plus enrichissantes de ma vie, et je t'en serai toujours reconnaissante. Merci de m'avoir encouragée à explorer tous les aspects et opportunités d'une thèse, et ainsi faire un doctorat qui est réellement mien. Merci pour les mille-et-une rencontres et discussions, pour les conseils et pour le soutien. Au plaisir de continuer à travailler ensemble !

À **Anne Marinier**, ma co-directrice, et les membres de mon comité de thèse, **François Major** et **Claudia Kleinman**, merci pour votre temps. J'ai grandement apprécié nos discussions scientifiques. Celles-ci ont été sources de motivation et d'encouragements, et ont contribué à ma réussite.

À mes anciens et présents collègues du Laboratoire Lemieux, merci pour vos commentaires constructifs lors des lab meetings. Un merci tout spécial à **Assya**, **Maria Virginia**, **Safia** et **Léa**. Merci d'avoir été présentes lors de moments plus difficiles, et d'avoir partagé mon bonheur lors de réussites. Votre écoute et votre soutien ont fait une différence, et je suis fière de compter d'aussi formidables femmes et scientifiques parmi mes amies.

À mes amis d'autres laboratoires, **Savandara** et **Thomas**, merci de m'avoir fait découvrir de nouvelles facettes de la science. Votre enthousiasme à mener des projets, que ce soit pour l'AÉBINUM ou pour Sciences À La Carte, est contagieuse, enivrante et inspirante. Collaborer avec vous sur ces divers projets était complémentaire à mon doctorat et a contribué à mon amour pour la Science-avec-un-S.

Aux membres de la plateforme bio-informatique circa 2015, **Patrick**, **Éric**, **Jonathan** et **Jean-Philippe**, merci de m'avoir prise sous votre aile et de m'avoir initiée à la "vraie" bio-informatique. Cet été passé à la plateforme, à apprendre de vous, a impacté positivement

mon choix d'entreprendre des études graduées et fut le premier pas vers le doctorat que je complète aujourd'hui. Un merci tout particulier à **Geneviève** qui fut et restera un mentor pour moi. Merci pour ton temps, ton écoute et tes conseils.

Merci à **Sarah**, mon amie de cœur. Merci pour ton constant soutien et tes encouragements; merci pour ta carte; merci pour ton humour et ta simplicité.

Merci à mes parents, **Pierre** et **Antoinette**; merci de m'avoir appris à rêver et à persévérer; merci de lire chacun de mes *absracts* avec un émerveillement renouvelé; merci de célébrer mes accomplissements, petits et grands.

Merci à mon petit (grand) frère, **Maxime**; merci d'être à l'écoute, même quand je radote; merci d'être courageux et persévérant comme tu l'es, mon modèle de tous les jours.

Et à mon **Jean-François**, le plus doux des mercis. À chaque pas entrepris, tu étais à mes côtés, à me tenir la main, et pour cela, je tiens simplement à te dire: merci pour tout.

À vous tous, je vous dis le plus sincère des mercis.

Chapitre 1

Introduction

Ce premier chapitre se veut une introduction des grands thèmes abordés dans les chapitres subséquents (Chapitres 2, 3 et 4). Les efforts de recherche de la présente thèse s'installent dans le contexte général et biomédical du processus de la découverte de médicament. Celui-ci sera présenté, dans un premier temps, à la Section 1.1. Une description du contexte expérimental précis suivra à la Section 1.2. Nous y présenterons les données expérimentales considérées, soit les expériences de type dose-réponse, ainsi que la méthodologie standard utilisée pour faire leur analyse et ainsi en dériver une caractérisation de l'efficience d'un composé chimique. La Section 1.3 couvre les notions propres à l'inférence bayésienne, allant de la présentation du théorème de Bayes à l'implémentation algorithmique du processus d'inférence selon la méthode Monte-Carlo par chaînes de Markov (MCMC). Finalement, le chapitre se conclut à la Section 1.4 avec la présentation de mon projet de recherche et de ses objectifs.

1.1. La découverte de médicaments

Le processus de découverte de médicament (DDP, de l'anglais *drug discovery process*) comprend plusieurs phases et étapes, en plus d'être multidisciplinaire [1]. Le but du DDP est d'identifier des composés ou de petites molécules comme étant de potentiels candidats pour un nouveau médicament [2]. Une petite molécule (de l'anglais *small molecule*) est un composé synthétisé chimiquement dont le poids moléculaire est faible (< 500 Da) [3]. Par simplicité, nous nous limiterons à l'appellation générale "composé" pour définir une structure moléculaire d'intérêt.

Le DDP est initié par l'identification d'un besoin médical qui n'est présentement pas satisfait. L'absence d'un traitement pour une maladie donnée, la basse efficacité d'un traitement existant, ou même l'occurrence d'effets secondaires importants d'un traitement existant dont des problématiques garantes d'initier un effort de découverte de médicament [4]. Une

cible thérapeutique est par la suite identifiée et validée. Nous définissons comme cible thérapeutique une structure moléculaire naturelle dont l'interaction avec un composé génère une activité cellulaire quantifiable [5]. Un exemple de cibles connues sont les récepteurs couplés aux protéines G (GPCRs, de l'anglais *G-protein-coupled receptors*) [6]. La validation d'une cible sert à confirmer son rôle dans une condition donnée, soit celle de la maladie d'intérêt.

Plusieurs composés sont ensuite testés en laboratoire dans le but d'identifier des composés-potentiel (de l'anglais *hits*). Ceux-ci sont identifiés comme étant des composés générant une activité optimale lors d'un criblage. Généralement, les composés-potentiel sélectionnés sont puissants et leur activité suggère être liée à la cible thérapeutique d'intérêt. Des composés analogues aux composés-potentiel sont synthétisés, selon différentes séries chimiques, pour générer des composés-candidat (de l'anglais, *leads*). La relation activité-structure (SAR, de l'anglais *structure-activity relationships*) pour diverses propriétés (ex. activité, sélection de la cible, potentiel toxique) de ces composés-candidat est quantifiée expérimentalement. Un sous-ensemble de composés-candidat est identifié et leur structure chimique est optimisée. Le composé le plus optimal (selon divers critères) est sélectionné pour passer à la phase préclinique du processus de développement de médicament [2, 7]. Les différentes étapes du DDP sont présentés sommairement dans la Figure 2. De façon générale, un seul composé-candidat du DDP passe à la phase préclinique du processus de développement de médicament. En parallèle de ce deuxième processus, d'autres composés-candidat sont optimisés et conservés comme substitut dans l'éventualité que le composé-candidat principal ne passe pas les diverses phases précliniques ou cliniques [2].

Historiquement, le DDP était initié par la découverte d'un élément actif naturel provenant d'une plante ou d'un minéral, garantissant ainsi la disponibilité de la structure moléculaire d'intérêt. Les potentiels thérapeutiques n'étaient cependant étudiées qu'après l'identification de l'élément actif [1]. Aujourd'hui, ce processus est inversé, tel que décrit plus haut. Cette inversion permet de déployer des efforts de recherche de façons précises et pertinentes, en tenant compte d'un besoin médical identifié. Cela étant dit, les structures moléculaires d'intérêt ne sont pas nécessairement des produits naturels et doivent souvent être synthétisées. De plus, l'interaction entre les structures moléculaires et la cible thérapeutique n'est pas garantie et doit être testée de façon exhaustive. Cela fait du DDP un processus très coûteux en temps ainsi qu'en ressources monétaires et matérielles [4]: il est donc primordial d'optimiser le processus décisionnel du DDP et ainsi maximiser le potentiel informatif de chacune de ses étapes.

Les travaux de la présente thèse s'inscrivent dans cette optique d'optimisation: nous cherchons à mieux outiller les biologistes et chimistes médicinaux dans leur processus de sélection de composés. Nos travaux se concentrent sur l'analyse d'un type d'expériences, soit les expériences de dose-réponse (Section 1.2). La méthodologie développée et présentée

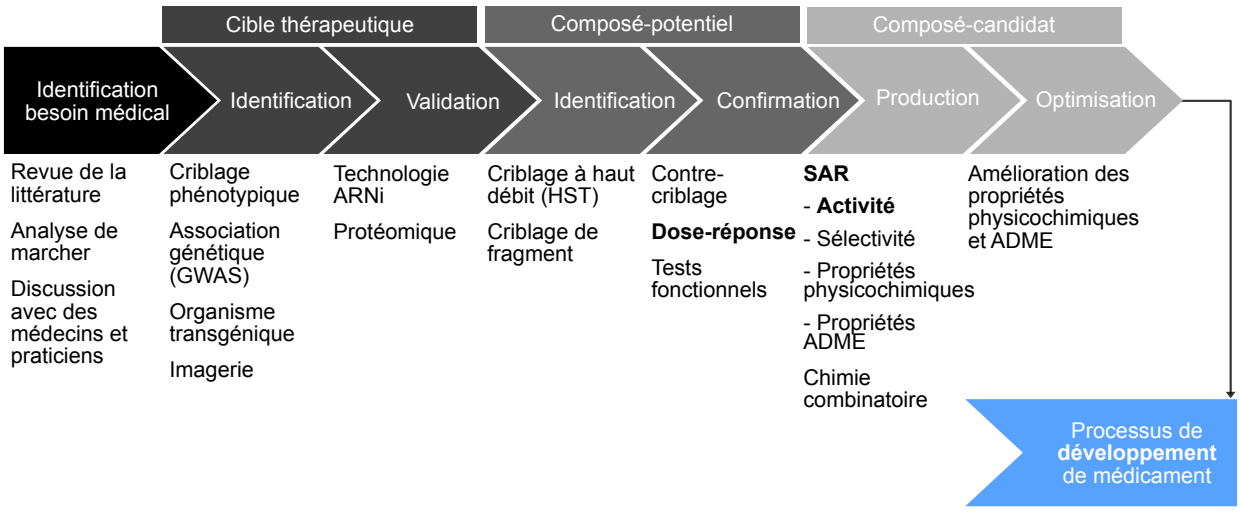


Fig. 2. Visualisation sommaire des étapes du processus de découverte de médicament (DDP). Chacune des phases générales sont identifiées par une teinte de gris et sont liées à un type de donnée: cible thérapeutique, composé-potentiel ou composé-candidat. Elles sont aussi accompagnées d'une description sommaire et non exhaustive des méthodes expérimentales typiques. Les méthodes expérimentales en gras génèrent des données pouvant être analysées par le processus présenté dans la présente thèse. ADME est un concept de pharmacocinétique décrivant le temps nécessaire pour qu'un médicament soit absorbé, distribué, métabolisé et éliminé par l'organisme étudié [8]. Figure adaptée de [2].

dans les Chapitres 1, 2 et 3 contribuent aux efforts de recherches des phases visant les composés-potentiel et -candidat (en gras, Fig. 2).

1.2. Les expériences de type Dose-Réponse

Un important nombre d'expériences entreprises dans le cadre du DDP sont dites de dose-réponse. Il y a, entre autres, celles de criblage et de criblage à haut débit (HTS, de l'anglais *high-throughpout screening*). Ces approches expérimentales permettent la caractérisation *in vitro* de composés via une quantification des réponses cellulaires générées. Le HTS a l'avantage de pouvoir caractériser un très grand nombre de composés, et ce, rapidement [7, 9]. Les expériences de transfert d'énergie par résonance bioluminescente (BRET, de l'anglais *bioluminscence resonance energy transfer*) sont un autre exemple de dose-réponse [10]. Il est à noter que les termes "dose" et "concentration" sont utilisés de façon interchangeable, bien que leur définition diffère [11]. Les appellations "dose" et "concentration" réfère à la variable de laquelle dépend la réponse cellulaire mesurée.

Des composés agonistes [8] sont testés sur des cellules dans le but d'identifier ceux générant une réponse satisfaisante. Notons que d'autres types de composés (e.g. antagonistes, agonistes inverses [8]) peuvent aussi être testés: leurs critères d'identification diffèrent cependant de ceux d'agonistes. L'efficience potentielle d'un composé est communément évaluée en quantifiant la réponse cellulaire générée pour une concentration donnée. Nous utilisons

l'appellation "criblage" (de l'anglais, *screen*) pour décrire le processus qui quantifie la réponse générée par plusieurs composés. Un criblage primaire permet d'étudier les réponses générées par une unique concentration et ainsi identifier des composés-potentiel (Fig. 2). La concentration utilisée est relativement élévée et les composés-potentiel identifiés sont ceux générant les réponses cellulaires les plus importantes [7, 12]. Un criblage dose-réponse permet quant à lui d'étudier un profil de réponses cellulaires étant donné un gradient de concentrations (Fig. 3) [13]. Ce type de criblage est notamment utilisé lors de la confirmation des composés-potentiel, de l'identification de cibles thérapeutiques [14], et d'analyse SAR pour la production et l'optimisation de composés-candidat [2]. L'effet combinatoire de composés est évalué via un criblage de synergie [15, 16]. La combinaison de composés est une avenue de recherche intéressante puisqu'elle a le potentiel d'augmenter l'efficience d'un traitement tout en réduisant les probabilités d'émergence de résistance. La combinaison de composés utilise principalement des composés connus et souvent approuvés, diminuant ainsi considérablement les coûts liés au développement de nouveaux composés [17]. Finalement, les effets d'un composé peuvent aussi être quantifiés via la relation entre les ratios BRET et les concentrations d'un composé d'intérêt [18].

La présente thèse porte sur l'analyse des résultats de criblage dose-réponse. Les prochaines sous-sections abordent leur contexte expérimental (Section 1.2.1) ainsi que le processus de quantification de l'efficience d'un composé (Section 1.2.2) et les outils présentement disponibles pour cet effet (Section 1.2.3).

1.2.1. Les expériences et leur protocole expérimental

Les criblages dose-réponse sont couramment utilisés lors des phases pré-cliniques du DDP [14]. Des cellules en culture sont exposées à une gamme de concentrations pour un nombre de composés donnés, et ce, pour une durée définie allant généralement entre 24 et 72 heures. La viabilité cellulaire est alors quantifiée via un décompte cellulaire (microscopie) ou une expérience de viabilité cellulaire ou de cytotoxité, telle que proposé par l'essai *CellTiter-Glow* de Promega [13, 19]. Cette deuxième quantification mesure, par exemple, les niveaux d'adénosine trisphophate (ATP) dans les lysats cellulaires: le niveau d'ATP (qui va produire la luminescence mesurée expérimentalement) est considéré comme proportionnel au nombre de cellules viables [20]. Les taux (%) de viabilité cellulaire (ou d'inhibition de la croissance cellulaire) sont obtenus en normalisant la luminescence obtenue pour une concentration à celles obtenues pour les contrôles négatifs (c.-à-d. absence de composé). Des contrôles positifs sont aussi utilisés pour confirmer le bon fonctionnement de l'expérience et leurs luminescences maximales est parfois utilisée dans le processus de normalisation. Ce sont ces taux de réponses cellulaires qui sont à la base de la caractérisation de l'efficience d'un composé. Nous utilisons le terme "expérience" pour définir un ensemble de réponses spécifique à un

composé et une implémentation expérimentale. Un criblage est ainsi composé de plusieurs expériences, et celles-ci peuvent être répliquées (c.-à-d. un même composé testé lors de deux implémentations expérimentales distinctes). Les expériences d'un même criblage sont généralement faites en même temps. Un criblage peut être fait manuellement ou de façon automatisée. Habituellement, les cribles manuels considèrent une plus petite librairie de composés que les cribles automatisés. Finalement, un jeu de données regroupe les expériences partageant une thématique similaire et peut ainsi contenir plusieurs criblages.

Bien que la normalisation des réponses cellulaires en taux facilite l'interprétation des données, cette manipulation est source de propagation et d'amplification du bruit présent dans les données. Ce bruit est une combinaison de variabilités biologique et expérimentale. Parmi les causes de ces variabilités, il y a la croissance cellulaire [13] et les effets de *batch* des composés testés [19], la position des puits sur une plaque multipuits, la gamme de concentrations sélectionnée et les dilutions en série [19, 21]. De récents travaux [19] ont démontré la difficulté à identifier et minimiser les causes de variabilités. L'automatisation des criblages et l'uniformité des protocoles expérimentaux se présentaient comme principales solutions [19, 22]. Ces variabilités affectent le subséquent processus de caractérisation de l'efficience d'un composé (Section 1.2.2), et il est donc primordial de les considérer lors de l'analyse, bien que cela ne soit pas trivial (Section 1.2.4).

1.2.2. La caractérisation de l'efficience d'un composé

L'efficience d'un composé est couramment caractérisée et quantifiée en termes de puissance et d'efficacité [23, 24]. La puissance (de l'anglais *potency*) est relative à la quantité de composé (c.-à.-d. la concentration) et à l'ampleur de l'effet généré [8] (Fig. 3). Une puissance est optimale lorsque qu'une importante réponse cellulaire est générée à une relativement faible concentration. L'IC_{50} et l'EC_{50} sont couramment utilisés comme références pour décrire la puissance d'un composé. Ces métriques représentent la concentration nécessaire pour générer une réponse équidistante entre les réponses minimale et maximale [23, 25]. Il existe diverses définitions pour ces métriques [26]. Par simplicité et généralité, nous utiliserons l'appellation IC_{50} lorsque la réponse mesurée est inhibée (c.-à-d.. données et courbe descendantes), et l'appellation EC_{50} lorsqu'elle ne l'est pas (c.-à-d.. données et courbe ascendantes) [25, 26]. Des variantes de l'IC_{50}/EC_{50} pour divers niveaux de réponse (e.g. IC_{90}, [27]) existent et peuvent être dérivées des métriques d'efficience de base. L'efficacité (de l'anglais *efficacy*) d'un composé est quant à elle relative à la réponse et est souvent référée par la réponse générée soit pour la concentration expérimentale la plus élevée, soit pour une concentration hypothétique infiniment grande [8, 23].

Bien que ces deux métriques soient les plus couramment utilisées, l'efficience d'un composé peut aussi être caractérisée par la pente de sa courbe dose-réponse (voir ci-bas). La pente quantifie le changement dans la réponse entre les deux plateaux expérimentaux, soit entre les réponses minimale et maximale. L'interprétation et la signification de la pente diffèrent d'un type d'expérience à l'autre. Pour une expérience de liaison au ligand (de l'anglais *ligand binding assay*), par exemple, la pente est indicative de la coopérativité entre le ligand et le récepteur (la pente est alors référée comme étant le coefficient de Hill [28]).

L'approche standard pour quantifier ces métriques d'efficience est la modélisation d'une courbe dose-réponse selon les données expérimentales (Fig. 3). Il existe différents modèles mathématiques décrivant la relation dose-réponse [29–31], le plus couramment utilisé étant le log-logistique (Équation 1) [32]. Les courbes dose-réponse obtenues à partir de ce modèle sont caractérisées par leur forme sigmoïde.

$$f(x) = a + \frac{b - a}{1 + 10^{d \cdot (x - c)}} \tag{1}$$

Le modèle log-logistique retourne une réponse $f(x)$ étant donné une concentration $\log_{10}$-transformé, x, et un ensemble $\theta = \{a, b, c, d\}$ de paramètres. Ces derniers comprennent les deux plateaux asymptotiques (a, b), le point d'inflexion (c) et la pente au point d'inflexion (d). Considérant le contexte biomédical des expériences dose-réponse, les paramètres libres θ du modèle sont interprétés tels que a et b sont respectivement les réponses à grande et faible concentrations (HDR et LDR, de l'anglais *high-* et *low-dose response*); c est l'IC_{50}/EC_{50} ; et d est la pente. L'Équation 1 devient alors l'Équation 2.

$$f(x) = HDR + \frac{LDR - HDR}{1 + 10^{pente \cdot (x - IC_{50})}} \tag{2}$$

Dès lors, nous référons à θ par "métriques d'efficience". Le LDR représente la réponse basale, soit la réponse attendue en l'absence de composé (ou pour une très faible concentration de composé). Le HDR, quant à lui, représente l'efficacité du composé telle que décrite plus haut (Fig. 3). Les valeurs des concentrations (x) et de l'IC_{50}/EC_{50} sont $\log_{10}$-transformées, bien que nous fassions, pour des questions de simplicité, abstraction de cela dans la notation.

Les valeurs des métriques d'efficience sont estimées $(\hat{\theta})$ pour une expérience dose-réponse via une régression non-linéaire. L'algorithme de Levenberg-Marquardt [33, 34] reste le plus couramment implémenté pour cette tâche dû à sa convergence rapide et stable [31]. La régression minimise la somme des résiduels entre les réponses expérimentales (y) et les réponses prédites $(\hat{y} = f(x; \hat{\theta}))$. Les valeurs de θ sont itérativement ajustées, selon une descente des gradients, pour représenter cette minimisation. L'algorithme est dit avoir convergé lorsqu'après I itérations, $\theta_I - \theta_{I-1} \leq \epsilon$. La régression non-linéaire reste l'approche la plus courante et est implémentée par divers outils (Section 1.2.3), bien qu'elle présente certaines

limitations non négligeables (Section 1.2.4). Quelques approches alternatives ont été proposées [35–40]. Celles-ci sont généralement peu utilisées, soit dû la complexité de l'approche et le manque d'outil (c.-à-d. peu accessible pour un expérimentateur), soit dû à la spécificité expérimentale de leur application (c.-à-d. n'est pas généralisable), ou soit dû à la nature des données expérimentales (c.-à-d. requiert l'application d'un nouveau protocole expérimental et ne peut être appliqué sur d'anciennes données).

Le résultat de la régression non-linéaire est visualisable via une courbe dite de dose-réponse (Fig. 3). Celle-ci représente les réponses prédites, $\hat{y}$ pour une gamme de concentrations hypothétiques et selon les valeurs estimées $\hat{\theta}$. Des exemples de telles courbes sont présentés dans la Figure 3. La caractérisation et la quantification de l'efficience d'un composé pour un contexte expérimental précis se fait par l'analyse de $\hat{\theta}$ et de la courbe dose-réponse résultante.

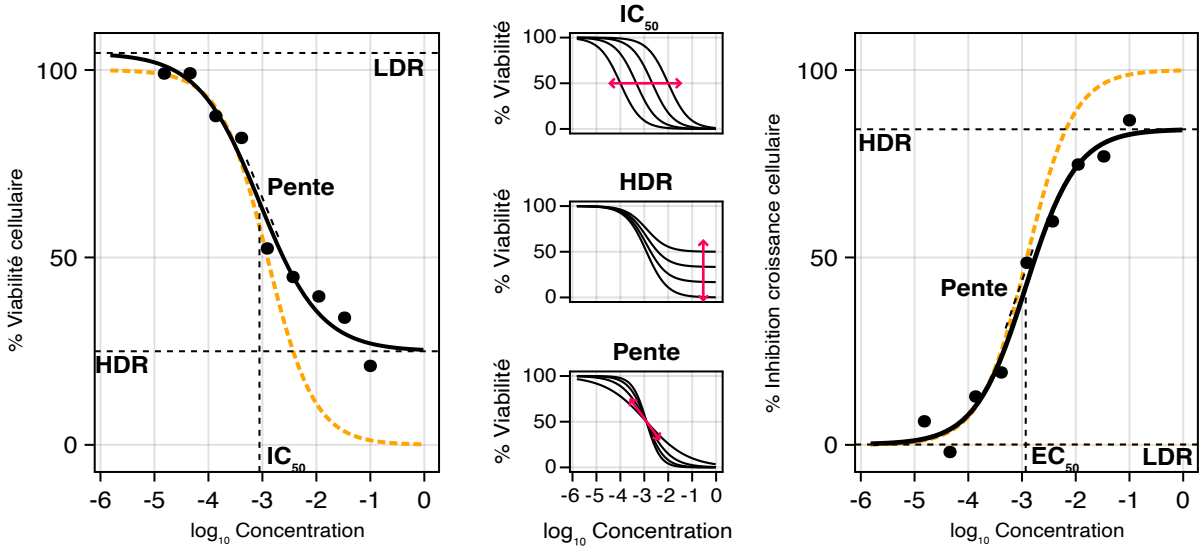


Fig. 3. Visualisation d'une courbe dose-réponse et de ses métriques d'efficience. Les panneaux de gauche et droite illustrent des exemples de réponses descendante (c.-à-d. viabilité cellulaire) et ascendante (c.-à-d. inhibition de la croissance cellulaire). Les mesures expérimentales sont représentées par les points noirs et la courbe dose-réponse estimée par régression non-linéaire par la courbe noire. Les métriques d'efficience correspondantes sont identifiées et leurs valeurs numériques sont marquées de traits hachés noirs. La courbe dose-réponse orange illustre une courbe synthétique optimale. Les effets des IC_{50} , HDR et pente sur la courbe dose-réponse sont représentés dans les panneaux du centre. Les flèches rouges illustrent la direction des effets.

La réponse considérée peut représenter un taux de viabilité comme d'inhibition. Les courbes résultantes sont alors respectivement descendante et ascendante (Fig. 3.). L'Équation 2 est représentante d'une courbe descendante. Pour une courbe ascendante, l'équation devient:

$$f(x) = LDR + \frac{HDR - LDR}{1 + 10^{pente \cdot (EC_{50} - x)}} \tag{3}$$

Les définitions mathématique et biologique de θ restent les mêmes. Notons cependant que les valeurs des LDR et HDR sont inversées. Le processus d'estimation de $\hat{\theta}$ décrit ci-haut reste le même. Par simplicité, nous référerons aux Équations 2 et 3 par l'appellation "modèle log-logistique" et assumons que l'équation appropriée est sélectionnée selon le type de réponse analysée.

L'évaluation de l'efficience d'un composé chimique se fait principalement via une analyse visuelle de la courbe dose-réponse obtenues, ainsi qu'une analyse des métriques d'efficience de base, soit la puissance (IC_{50}/EC_{50}), l'efficacité (HDR/LDR) et la pente. D'autres métriques d'efficience peuvent être calculés à partir des métriques de base. Il y a entre autre l'aire sous/au-dessus de la courbe dose-réponse (AUC/AAC, de l'anglais *area under/above the curve*) [23, 41] et le *DSS* (de l'anglais *drug sensitivity scoring*, [42]). L'AUC/AAC correspond à l'intégration du modèle log-logistique pour une gamme de concentration et pour des limites de réponses définies. La valeur obtenue combine les notions de puissance et d'efficacité. Sa comparaison d'une expérience à l'autre n'est pas triviale, considérant que la gamme de concentrations expérimentales est variantes [43]. Le DSS est une version normalisée de l'AUC/AAC qui permet de considérer les effets de toutes les métriques d'efficience. Une première normalisation est faite selon le HDR, puis une seconde selon la concentration à laquelle la réponse excède un seuil pré-défini. Le DSS est principalement utilisé pour quantifier la différence entre les réponses de cellules cancéreuses (ou autres) et de cellules contrôles [42]. Notons que ces métriques ne sont pas triviales à interpréter seules puisqu'elles résument les effets de diverses métriques, et sont surtout utilisées pour comparer plusieurs expériences selon divers critères (e.g. IC_{50}/EC_{50} et HDR).

1.2.3. Les outils et jeux de données

Tel que mentionné à la Section 1.2.2, les métriques d'efficience sont estimées via régression non-linéaire (c.-à-d. minimisation des moindres carrés). Cette approche est le standard [39, 44] et est implémentée et proposée aux expérimentateurs à travers plusieurs outils. L'accessibilité de ces outils varie en termes de coût, de convivialité et de flexibilité.

Parmi les outils les plus souvent utilisés nous retrouvons ActivityBase/SARview d'IDBS [`https://www.idbs.com/`] ainsi que GraphPad de Prism [`www.graphpad.com`]. Ces deux outils proposent une interface graphique et sont payants. ActivityBase/SARview a l'avantage d'être lié au processus expérimental de lecture des réponses et peut considérer un très grand nombre d'expériences. Les résultats sont retournés sous forme de rapport contenant, entre autres, les métriques d'efficience et les courbes dose-réponses. L'ensemble du protocole analytique doit cependant être défini avant la lecture des réponses, et ne peut être changé *a posteriori*, limitant l'expérimentateur dans ses analyses post-inférence. De plus,

l'utilisation d'ActivityBase/SARview nécessite une connaissance professionnelle du processus d'analyse et de l'interface. Alternativement, GraphPad est relativement simple à utiliser et communément employé pour l'analyse de criblage à plus petite échelle (c.-à.-d. des criblages manuels plutôt qu'automatisés). Cependant, les données de chaque expérience doivent être entrées manuellement et individuellement par l'expérimentateur. GraphPad permet d'obtenir une courbe dose-réponse ainsi que l'estimation des métriques d'efficience et quelques métriques diagnostiques et de confiance (e.g. erreur standard sur les métriques et métrique décrivant la qualité de l'ajustement). Similairement à GraphPad, IDBS propose aussi XLfit [`https://www.idbs.com/xlfit/`], une application pouvant être intégrée à Microsoft Excel. XLfit permet d'exécuter diverses régressions sur des données importées par l'expérimentateur. Tel que mentionné plus haut, ces outils requièrent l'achat de licences qui sont relativement couteuses et leur processus d'analyse complet et détaillé est souvent opaque pour l'expérimentateur. Celui-ci devient rapidement limité dans son analyse, et les données expérimentales ne sont pas toujours exploitées à leur plein potentiel [45].

À l'autre extrême des outils, nous trouvons divers langages de programmation: ceux-ci sont généralement libres d'accès (c.-à-d. aucun coût monétaire n'est lié à leur utilisation) et hautement flexibles, puisque le processus d'analyse est créé par l'expérimenter même. Les langages R [46] et Python [47] sont les plus communément utilisés par les expérimentateurs des domaines biomédicaux [48], bien que d'autres langages tels que Julia [49] et Matlab [50] soient aussi disponibles. Un expérimentateur peut définir son propre protocole d'estimations des métriques d'efficience avec, par exemple, `minpack` [51] en R et `scipy` [52] en Python, ou utiliser des librairies spécifiquement mises en place pour l'analyse d'expérience dose-réponse (e.g. `drc` [32, 53], `drda` [54] et `pGX` [55] en R, et `ECCpy` [56] en Python). Bien que flexibles, ces outils demandent que l'expérimentateur ait de bonnes habiletés en programmation et de bonnes connaissances des concepts mathématiques sous-jacents aux librairies utilisées, sans quoi les analyses peuvent involontairement mener à des résultats biaisés, voire erronés.

Plusieurs outils proposés par différentes équipes de recherche sont aussi disponibles, et ce, sous diverses formes (e.g. interface web, Jupyter Notebooks, logiciel). Ces outils couvrent plusieurs applications: la mise en place de protocoles expérimentaux [21] et analytiques complets [57], l'estimation de métriques d'efficience pour diverses conditions [44, 45, 58–60]. Ces outils proposent aussi des métriques de diagnostic de l'ajustement (e.g. R^2 et la RMSE) ainsi qu'une visualisation des résultats. Les efforts de développement de tels outils s'inscrivent généralement dans un ou plusieurs des trois objectifs suivants: (1) amélioration de la méthodologie standard (voir Section 1.2.4), (2) amélioration de la reproductibilité des résultats [44, 45, 58, 59] et de la transparence du processus analytique [21, 55, 57] et (3) exploration et visualisation de jeux de données [58, 60–62].

Le troisième objectif mentionné ci-haut concerne principalement l'exploration et la visualisation de larges jeux de données publics [41, 63–69]. Dans les dernières décennies, plusieurs équipes et instituts de recherche ont testé expérimentalement des centaines de composés sur diverses lignées cellulaires cancéreuses (CCL, de l'anglais *cancer cell lines*) [60, 61] dont ils ont aussi fait le profillage moléculaire (c.-à-d. analyses génomiques). L'objectif premier de la mise en place de ces jeux de données est de permettre l'analyse de corrélations entre la sensibilité des CCLs à des composés, et des mesures génomiques (e.g. mutations). De telles analyses permettent l'identification de nouvelles signatures prédictive de la réponse à un composé [70] et s'installent dans les efforts de recherche en médecine de précision [71]. Les jeux de données gCSI [66, 72], CTRPv2 [63, 73, 74], NCI60 [65], CCLE [41] et GDSC [70] sont des exemples de larges jeux de données pharmacogénomiques publics.

Bien que ces jeux de données soient publics, deux principaux facteurs limitent leur utilisation [60, 61]. Premièrement, l'analyse et la gestion de données brutes n'est pas un processus trivial pour tous les expérimentateurs. Cela devient encore plus complexe lorsque qu'un grand nombre d'expériences sont considérées. Deuxièmement, les annotations des CCL et des composés sont hautement inconsistantes entre les jeux de données. Il devient donc difficile pour un expérimentateur de considérer plusieurs jeux pour une même analyse.

Des outils, tels que PharmacoDB [60, 75], facilitent l'utilisation de ces larges jeux de données en proposant un processus d'harmonisation. Les données de plusieurs jeux sont ainsi accessibles et visualisables via une interface web, par exemple. Ces outils [58, 61, 75] proposent aussi un processus analytique uniforme (e.g. normalisation des réponses, calcul de métriques d'efficience) qui est appliqué à tous les jeux disponibles.

1.2.4. Les limitations de la caractérisation standard des métriques d'efficience

L'ajustement d'une courbe dose-réponse aux données expérimentales, via régression non-linéaire (algorithme Levenberg-Marquardt), est la méthodologie standard pour estimer les métriques d'efficience [39]. L'ensemble des outils présentés à la Section 1.2.3 implémente cette approche. Elle présente cependant une principale et importante limitation pouvant impacter négativement les résultats et conclusions: l'évaluation de notre confiance dans le processus analytique n'est pas implicite à celui-ci. Or, plusieurs facteurs peuvent affecter la certitude (ou incertitude) des métriques d'efficience: la complétude de la réponse (c.-à-d. la gamme de concentrations expérimentales n'est pas toujours représentative de la réponse complète et certaines métriques ne peuvent être observées expérimentalement, Fig. 4.B), la réactivité cellulaire au composé (c.-à-d. certains composés ne semblent pas générer de réponse cellulaire [76], Fig. Fig. 4.C) et la variance des réponses mesurées pour une expérience (c.-à-d. les effets de bruits biologiques et expérimentaux, Fig. 4.C).

Des méthodes pour pallier cette importante limitation sont utilisées, mais celles-ci restent découplées du processus d'estimation des métriques d'efficiences et comportent leur propre lot de limitations [77].

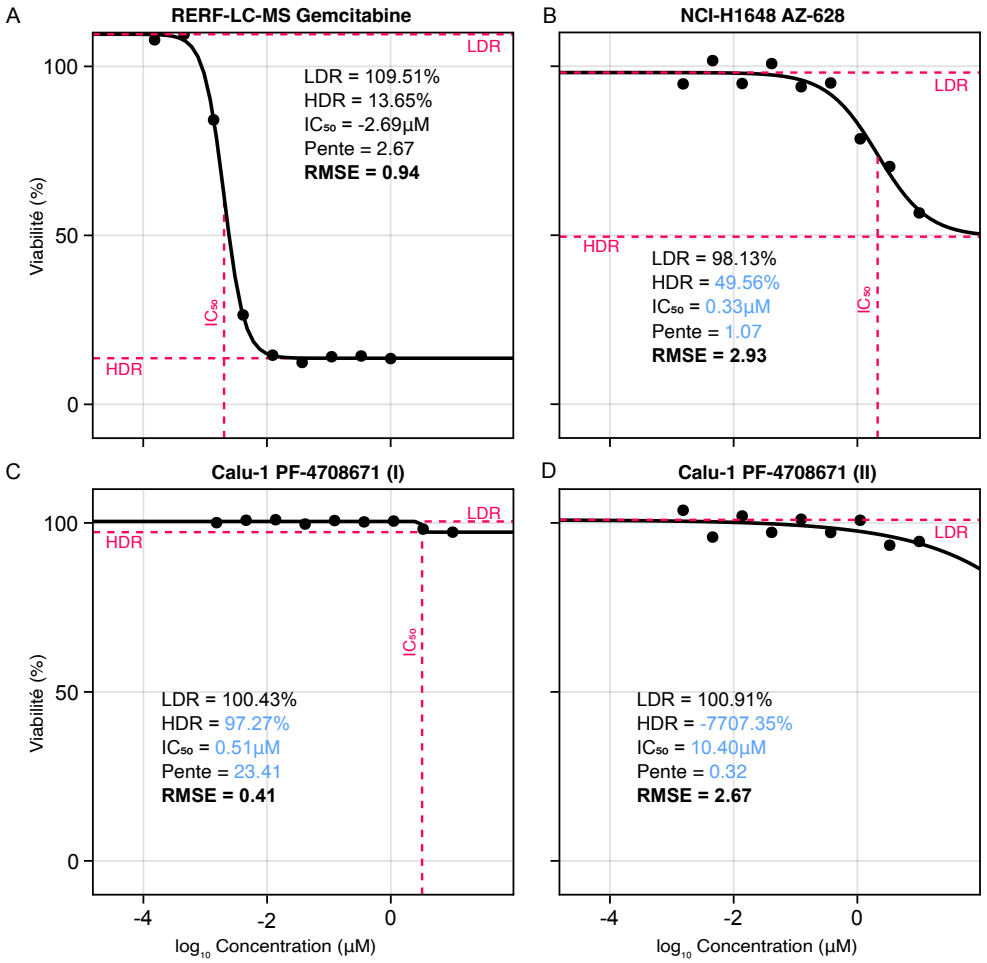


Fig. 4. Exemples de courbes dose-réponse et de leurs métriques d'efficience. Les réponses expérimentales proviennent du jeu de données gCSI. Pour chaque expérience, les courbes dose-réponse et les métriques d'efficience estimée sont représentées (le trait noir et les traits hachés rouges, respectivement). Les valeurs des métriques d'efficience et de la RMSE (qualité de l'ajustement) sont indiquées. Les valeurs incertaines, peu probables selon le contexte, ou non observées expérimentalement sont indiquées en bleu. Cette indication provient d'une évaluation qualitative et de notre intuition. **(A)** Exemple de réponse sigmoïde complète et peu bruitée. L'approche standard retourne des valeurs dont la précision et la certitude ne sont pas remises en question. **(B)** Exemple de réponse sigmoïde incomplète et bruité. L'approche standard retourne une estimation du HDR qui n'est pas soutenue par les données expérimentales: le plateau inférieur n'est tout simplement pas observé. La certitude et la précision du HDR et des IC_{50} et pente (puisque ces valeurs sont dépendantes du HDR) sont questionnables. **(C)** Exemple de réponse suggérant que le composé n'a pas d'effet sur les cellules. L'approche standard force une sigmoïde et les métriques estimées ne correspondent pas au contexte expérimental observé. **(D)** Exemple de réponse ambiguë et bruitée, oscillant entre l'absence d'effet et être incomplète. De façonsimilaire à **B** et **B**, les données et le contexte expérimental ne soutiennent pas les estimations obtenues pour les diverses métriques d'efficience.

Une première catégorie de méthodologies consiste à minimiser le besoin d'évaluer l'incertitude des métriques d'efficience. Cela est fait soit en identifiant et minimisant les effets de réponses aberrantes (de l'anglais *outliers*) [39], soit en plafonnant les réponses à des valeurs minimale et maximale [55, 78], soit en personnalisant le processus de normalisation [79]. Bien que ces manipulations de données puissent être artificielles, elles diminuent la variabilité à même une réponse. Il est aussi pratique courante d'utiliser une constante pour le LDR (e.g. 0% ou 100%) et seulement inférer les trois autres métriques d'efficience [21, 55, 62]. Ces approches contraignent l'estimation à un contexte expérimental optimal qui n'est pas nécessairement représentée par les données: les manipulations de données pré-analyse (e.g. normalisation) résultent parfois à des réponses sortant des limites optimales [0% - 100%] (Fig. 4). Les métriques d'efficience obtenues ont le risque d'être soient erronées, soit peu informatives.

Une deuxième catégorie de méthodologie consiste à définir des métriques pour évaluer l'incertitude. Des intervalles de confiance autours des métriques d'efficience sont obtenues depuis les erreurs standards (implémenté par GraphPad), par ré-échantillonnage Bootstrap [31, 80, 81] (implémenté par plusieurs par diverses librairies de programmation et par [82]), ou par simulation Monte-Carlo [31] (aussi implémenté par GraphPad). La validité et le caractère informatif des intervalles de confiance obtenus sont hautement dépendantes du nombre de concentrations (ainsi que du nombre de mesures par concentration) et de la qualité de la réponse [77]. Alternativement, la qualité de l'ajustement de la courbe dose-réponse est aussi utilisée [21, 58, 59, 62]. Celle-ci est représentée soit par la racine de l'erreur quadratique moyenne (RMSE, de l'anglais *root-mean-square error*)(Fig. 4), soit par R^2 (c.-à-d. le coefficient de détermination). Bien que ces métriques soient indicatives de la capacité du modèle à prédire la réponse et la variance observée, elles ne représentent pas réellement l'incertitude des métriques d'efficience [77]. L'expérimentateur doit extrapoler de ces métriques l'incertitude et se fier à son intuition. Ces métriques, RMSE et R^2, font aussi abstraction du contexte expérimental: l'ajustement d'une sigmoïde forcée sur des données non réactives peut être associée à une faible RMSE, par exemple, bien que les métriques d'efficience soit hautement incertaines et erronées (Fig. 4.C). Encore une fois, l'identification d'un tel cas est communément faite par l'expérimentateur via une évaluation visuelle [70, 83]. Quelques outils implémentent des comparaisons (teste-F) entre la qualité des ajustements pour le modèle sigmoïde et un modèle linéaire constant [21, 62]. Lorsque ce deuxième modèle est favorisé, les métriques d'efficience sont simplement ignorées. Bien qu'utile, cette approche ne permet d'évaluer l'incertitude des métriques, notamment lorsque la réponse est incomplète.

Il est présentement difficile, voire impossible, pour un expérimentateur d'évaluer adéquatement l'incertitude des métriques d'efficience lorsque celles-ci sont estimées via une régression non-linéaire (Levenberg-Marquardt). Il est cependant primordial de considérer l'incertitude, notamment lorsque les métriques de diverses expériences sont comparées. Cette comparaison forme la base d'un processus décisionnel qui permet de sélectionner des composés et de les progresser dans le DDP (Section 1.1). Les métriques sont couramment comparées selon la magnitude de leur différence (e.g. 10 *fold change* dans les IC_{50} [84]). Cela se complexifie lorsque plusieurs métriques doivent être considérées pour une même sélection. Des métriques telles que l'AUC/AAC [23, 41] et le DSS [42] sont alors utilisées, bien que celles-ci se basent sur les métriques estimées par régression non-linéaire. Leurs auteurs reconnaissent le besoin d'intégrer un processus quantifiant leur incertitude [42]. Un poids considérable est alors donné à l'intuition de l'expérimentateur lors de l'interprétation des résultats.

Les travaux de Haibe-Kains *et. al* [22] sont un bon exemple des importants effets engendrés par les limitations de la régression non-linéaire (Levenberg-Marquardt). Ils démontrent que les métriques d'efficience de trois larges jeux de données (CGP [70], CCLE [41] et GSK [85]) ne corrèlent globalement pas, et ce, même lorsque le processus d'estimation est uniformisé. S'est ensuivi une remise en question de la validité de ces jeux de données (lequel détient les "vraies" données?) et des conclusions tirées à partir de ceux-ci (e.g. identification de biomarqueurs). D'autres travaux ont par la suite exploré les causes de cette discordance [19, 76, 86, 87] et démontrent qu'il est possible de minimiser les effets des bruits expérimentaux via l'uniformisation des protocoles [19], bien que les bruits biologiques soient inévitables. Tous identifient un important besoin de réviser la façon dont les réponses cellulaires sont résumées en métrique d'efficience, et ainsi minimiser l'ajout d'un bruit analytique.

Notons qu'outre les travaux étudiants la correlations entre les métriques d'efficience de divers jeux de données, aucune démonstration concrète des limitations de la méthode standard n'a été fait.

1.3. L'inférence Bayesienne

Le terme inférence réfère au processus permettant de déterminer la "cause" depuis les "effets". L'inférence nous permet de tirer des conclusions se basant sur le concept de probabilité, soit la probabilité de la cause étant donné les effets. Toute hypothèse faite par rapport à la cause peut donc être acceptée, ou refusée, sur la base de la probabilité obtenue. L'inférence fréquentiste définie cette probabilité suite à une série de tests binaires. L'inférence bayésienne, elle, se base sur le théorème de Bayes pour obtenir la probabilité. Cette dernière représente alors une accumulation d'évidence pour un ensemble de causes possibles. La présente section aborde spécifiquement ce type d'inférence.

Nous présentons le théorème de Bayes dans la Section 1.3.1, puis enchaînons dans la Section 1.3.2 avec une présentation de l'approche Monte Carlo par chaînes de Markov (MCMC, de l'anglais *Markov Chain Monte Carlo*), une classe d'algorithme permettant d'implémenter efficacement l'inférence bayésienne. Finalement, nous abordons le sujet de la programmation probabiliste et l'utilisation simplifiée des MCMCs dans la Section 1.3.3.

Pour faciliter l'intégration des notions théoriques des prochaines sections à notre contexte expérimental, nous utilisons les termes θ pour dénoter la "cause" et "*observations*" pour dénoter les "effets". Plus concrètement, θ représente alors un ensemble de paramètres pouvant décrire le comportement ou la tendance de nos *observations*.

1.3.1. Le théorème de Bayes

Le théorème de Bayes (Éq. 4) [88] est à la base de l'inférence bayésienne. L'Équation 4 permet d'inférer une distribution des valeurs les plus probables de θ, étant donné un ensemble d'*observations* [89]. Le résultat de l'Équation 4 est communément appelé *posterior* (Fig. 5).

$$P(\theta|observations) = \frac{P(observations|\theta) \times P(\theta)}{P(observations)} \tag{4}$$

Plus concrètement, le théorème de Bayes ajuste la probabilité de θ pour chaque nouvelle *observation* faite. La probabilité initiale de θ, c.-à-d. sans considération pour les *observations*, est décrite par l'appellation *prior*. Les *observations* sont incorporées via le terme de vraisemblance qui décrit la probabilité d'obtenir les *observations* pour des valeurs données de θ. Finalement, le *posterior* est normalisé par la probabilité d'observer les *observations* considérées (Fig. 5). Il est à noter que le choix du *prior* peut avoir un important impact sur le *posterior* résultant, et qu'il est donc primordial de définir tout *prior* de façon éclairée. Le rôle et l'impact de cette composante du théorème sont abordés et discutés plus en détail dans les Chapitres 2, 3 et 5.

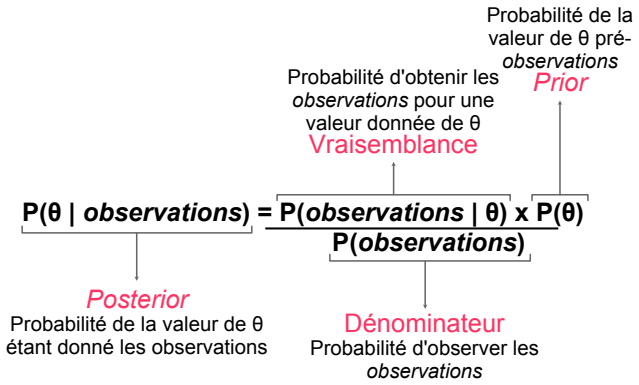


Fig. 5. Équation annotée du théorème de Bayes. Annotation de l'Équation 4 par terme.

Comme mentionné ci-haut, le dénominateur normalise le *posterior* en une densité de probabilités valide. Numériquement, le dénominateur est l'intégrale (lorsque θ est continue) du numérateur pour toutes les valeurs possibles de θ. Le résultat, $P(observations)$, est une densité de probabilité marginale représentant une distribution de probabilité pour nos *observations*. Le calcul de l'intégrale devient rapidement complexe lorsque θ représente plusieurs paramètres (intégrale multidimensionnelle). Pour des problèmes relativement complexes, il devient impossible de normaliser le *posterior* . Cependant, il est possible de remplacer le calcul exact du *posterior* (Éq. 4) par une approximation via échantillonnage [90, 91]. Le théorème de Bayes (Éq. 4) devient alors:

$$P(\theta|observations) \propto P(observations|\theta) \times P(\theta) \tag{5}$$

Le dénominateur de l'Équation 4 est ignoré lors de l'approximation par échantillonnage puisqu'il informe de la hauteur du *posterior* : or, pour échantillonner d'un *posterior* il ne nous faut que connaître sa forme et celle-ci est donnée par le numérateur de l'Équation 4 [91]. L'approche d'échantillonnage des *posteriors* forme la base des méthodes modernes computationelles [90] (Section 1.3.2).

1.3.2. MCMC: méthodes Monte-Carlo par chaîne de Markov

Tel que mentionné dans la précédente Section 1.3.1, le dénominateur de l'Équation 4 devient rapidement complexe lorsque plus d'un paramètre continu sont considérés (c.-à.-d. intégrable multidimensionnelle). Les *posteriors* peuvent être obtenus numériquement via discrétisation des paramètres continus [92] ou par quadrature [93, 94], ou être inférés via un échantillonnage Monte-Carlo. Cette dernière approche est favorisée aux deux premières puisque sa complexité n'est pas proportionnellement exponentielle à la complexité du problème et au nombre de paramètres considérés [90].

Considérons la valeur θ_A, positionnée dans l'espace θ. Nous pouvons déterminer la fréquence de θ_A depuis la fonction de densité (PDF, de l'anglais *probability density function*) décrivant l'espace θ. Dans le cas où la PDF ne nous est pas connue, la fréquence de θ_A peut être calculée de façon relative aux autres valeurs de l'espace θ. Considérant la valeur θ_B, aussi positionnée dans l'espace θ, nous pouvons calculer le ratio de l'Équation 6:

$$\begin{aligned} \frac{p(\theta_A|observations)}{p(\theta_B|observations)} &= \frac{\frac{p(observations|\theta_A)\cdot p(\theta_A)}{p(observations)}}{\frac{p(observations|\theta_B)\cdot p(\theta_B)}{p(observations)}} \\ &= \frac{p(observations|\theta_A)\cdot p(\theta_A)}{p(observations|\theta_B)\cdot p(\theta_B)} \end{aligned} \tag{6}$$

Si nous venions à calculer l'ensemble des ratios $\frac{p(\theta_A|observations)}{p(\theta_i|observations)}$ pour toutes valeurs possibles θ_X, nous obtiendrons la fréquence relative de θ_A. Tel que démontré par l'Équation 6, le dénominateur du théorème de Bayes (Éq. 4) n'est pas nécessaire pour l'évaluation de la

fréquence relative: l'information véhiculée par le *posterior* non normalisé (Éq. 5) nous permet d'obtenir les fréquences relatives de chaque valeur θ_X. L'abstraction du dénominateur n'affectera pas la forme du *posterior*, puisque le dénominateur n'est pas dépendant de θ. De plus, la hauteur relative (plutôt qu'exacte) d'un *posterior* est suffisamment informative pour en dériver des métriques informatives telles que la moyenne et la médiane [90].

Il nous est donc possible d'échantillonner les *posteriors* non normalisés pour obtenir une représentation des *posteriors* θ [91]. Par simplicité, l'utilisation du terme *posterior* réfère dès lors et pour le reste de la présente thèse aux *posteriors* non normalisés (Éq. 5).

Optimalement, l'échantillonnage d'un *posterior* serait fait de façon indépendante. Dans les faits, l'échantillonnage peut au mieux être fait indépendamment de manière pseudo-aléatoire [90, 91] avec les algorithmes d'échantillonnage par rejet (de l'anglais *rejection sampling*) [95] et d'échantillonnage par transformation inverse (de l'anglais *inverse transform sampling*) [96]. Le premier algorithme est cependant hautement inefficace lorsque plusieurs paramètres sont à échantillonner (c.-à.-d. seule une petite proportion des échantillons est acceptée et conservée); le deuxième algorithme n'est pas applicable lorsque la fonction de distribution cumulative (CDF, de l'anglais *cumulative distribution function*) n'est pas connue, comme c'est le cas pour les *posteriors*.

Alternativement, l'échantillonnage dépendant (c.-à-d. θ_{i+1} dépend de θ_i) ne nécessite pas le calcul de plusieurs ratios (Éq. 6) et a l'avantage d'être léger en termes de calculs. Plutôt que de considérer l'ensemble d'un *posterior*, l'approche MCMC (Monte Carlo par chaînes de Markov) se concentre sur des pas (de l'anglais *steps*) locaux aux localisations aléatoires (d'où l'appellation Monte-Carlo) dans l'espace θ. Les valeurs θ_i échantillonnées successivement constituent une chaîne. L'appellation « chaîne de Markov » (de l'anglais *Markov chain*) vient du fait que seul θ_i est considéré pour déterminer θ_{i+1}. Les algorithmes MCMC sont dits « sans mémoire ». L'échantillonnage dépendant a cependant le désavantage de mener à une chaîne auto-corrélée. Pour minimiser cet effet indésirable, un plus grand nombre d'échantillons est nécessaire, comparé à un échantillonnage indépendant, pour inférer adéquatement un *posterior*. La taille effective d'échantillonnage (de l'anglais *effective sample size*) réfère au nombre d'échantillons indépendants nécessaires pour obtenir une même précision d'inférence avec un échantillonnage dépendant [91].

Un *posterior* est normalement constitué de plusieurs chaînes indépendantes, initialisées à diverses valeurs [91, 97]. La convergence d'un algorithme d'inférence est définie en comparant ces dites chaînes: optimalement, il nous serait impossible de différencier les valeurs d'une chaîne de celles des autres chaînes. La métrique de base pour évaluer la convergence d'un *posterior* est le facteur de réduction d'échelle potentiel (PSRF, de l'anglais *potentiel scale reduction factor* ou $\hat{R}$) [98, 99]. Le PSRF (Éq. 7) indique s'il serait possible d'augmenter la précision du *posterior* (diminuer sa largeur) en considérant plus d'itérations. Pour un

nombre infini d'itérations, tout *posterior* converge et le PSRF = 1. En règle générale, un *posterior* est dit avoir convergé lorsque son PSRF < 1.1. Un large PSRF indique que la variance inter-chaînes (B, Éq. 8) est supérieure à la variance intra-chaîne (W, Éq. 9) [98, 99]. Un tel PSRF est notamment observé lorsque le *posterior* est multimodal et que les chaînes ont individuellement convergées vers différents modes. Notons que la PSRF assume que le *posterior* est normalement distribué. D'autres métriques de convergence ont été proposées, mais la PSRF reste la plus commune pour sa simplicité de calcul [100].

$$PSRF = \hat{R} = \sqrt{\frac{W + \frac{1}{N}(B - W)}{W}} \tag{7}$$

$$B = \frac{N}{M-1} \sum_{j=1}^{M} (\bar{\theta_j} \bar{\theta})^2 \tag{8}$$

$$\begin{aligned} W &= \frac{1}{M} \sum_{j=1}^{M} s_j^2 \\ s_j^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (\theta_{j,i} - \bar{\theta_j})^2 \end{aligned} \tag{9}$$

Dans les Équations 7 à 9, N réfère aux nombres d'itérations, M aux nombres de chaînes, et $\bar{\theta}$ à la moyenne de θ.

Considérant que chaque chaîne est initialisée indépendamment et à diverses localisations de l'espace θ, leurs premières itérations sont peu représentatives du *posterior*. Il est pratique courante d'ignorer les w premières itérations de chaque chaîne que l'on décrit comme étant des itérations d'échauffement (de l'anglais, *warmup*) [91].

De façon générale, pour chaque itération des algorithmes MCMC, une valeur θ_{prop} est échantillonnée selon une fonction spécifique considérant θ_{curr}. La valeur θ_{accept} ajoutée à la chaîne c est soit θ_{prop} (selon une probabilité α qui est relative au ratio $\frac{p(\theta_{prop}|observations)}{p(\theta_{curr}|observations)}$), soit θ_{curr} (Algo. 1).

Algorithm 1 Pseucode d'un algorithme MCMC

```
for C chaînes do
    θ_curr ← θ_0
    for N itérations do
        θ_prop ← fonction(θ_curr)
        α ← min(1, ratio)
        θ_accept ← θ_prop selon la probabilité α
        c ← c + [θ_accept]
        θ_curr ← θ_accept
    end forc[w:]
end for
```

Les prochains paragraphes présentent les principaux algorithmes MCMC dans leur forme générale.

L'**algorithme *Random Walk***, dans sa version *drunkard*, échantillonne θ_{prop} depuis une distribution uniforme et ajoute toujours cette valeur à la chaîne. Cette dernière est peu représentative de la forme du *posterior*, puisque l'exploration ne se concentre pas sur les régions de haute densité. Dans sa version *Hill Climb*, l'algorithme accepte θ_{prop} si et seulement si $p(\theta_{prop}|observations) > p(\theta_{curr}|observations)$. Le *posterior* résultant est biaisé vers le mode et est souvent non représentatif de sa réelle forme. L'algorithme *Random Walk* Metropolis [101] minimise les effets du *Hill Climb* en ne rejetant pas systématiquement θ_{prop} lorsque $p(\theta_{prop}|observations) < p(\theta_{curr}|observations)$ (Éq. 10).

$$\alpha = \begin{cases} 1 & \text{si } p(\theta_{prop}|observations) \geq p(\theta_{curr}|observations) \\ \frac{p(\theta_{prop}|observations)}{p(\theta_{curr}|observations)} & \text{sinon} \end{cases} \tag{10}$$

L'échantillonnage de θ_{prop} se fait depuis une normale centrée à θ_{curr}. La valeur de σ doit être spécifiée, et elle dictera la largeur des pas (de l'anglais *step size*): une petite valeur ralentira le processus d'exploration des régions à haute densité, et une grande valeur mènera principalement à l'exploration de régions à basse densité (c.-à-d. rejet de θ_{prop}). Dans le cas où θ comprend plusieurs paramètres, une distribution multivariée normale et une matrice de covariance sont considérées. L'algorithme *Random Walk Metropolis-Hasting* [102] permet de considérer des distributions de proposition J (c.-à-d. distributions desquelles θ_{prop} est échantillonné) asymétriques. La probabilité d'acceptation devient (Éq. 11):

$$\alpha = \frac{p(\theta_{prop}|observations)}{p(\theta_{curr}|observations)} \times \frac{J(\theta_{curr}|\theta_{prop})}{J(\theta_{prop}|\theta_{curr})} \tag{11}$$

L'algorithme *Random Walk* Metropolis a l'avantage d'être simple en termes de calculs. La convergence des *posteriors* est cependant lente et son optimisation est dépendante d'un ajustement de la largeur des pas (c.-à-d. σ de la distribution d'échantillon $\mathcal{N}(\theta_{curr}, \sigma)$). De plus, l'exploration de l'espace θ complet est limité par le rejet probabiliste des valeurs de θ se trouvant dans des régions de basse densité et par le fait que l'échantillonnage de θ_{prop} ne considère pas la forme du *posterior* (c.-à-d. ne favorise pas l'échantillonnage de valeurs se trouvant des régions à haute densité) [91].

L'**algorithme de Gibbs** [103] diffère grandement des autres algorithmes MCMC et ne peut être décrit par la forme générale présentée par l'Algorithme 1. Aucune valeur θ_{prop} est rejetée puisque l'échantillonnage se fait à partir des distributions conditionnelles de θ. Notons que cet algorithme est applicable lorsque θ est constitué d'au moins deux paramètres ($i \geq 2$). Les valeurs θ^i_{propr} sont échantillonnées successivement à même une itération. Leur ordre d'échantillonnage est aléatoire et change d'une itération à l'autre. L'échantillonnage d'une valeur θ^i_{prop} se fait depuis la distribution $p(\theta^i|\theta^{k\neq i}, y)$, et considère les plus récentes

valeurs de $\theta^{k\neq i}$. Ce type d'échantillonnage est dit « *block sampling* » [90, 91]. L'algorithme de Gibbs est généralement plus rapide et efficace que le *Random Walk* Metropolis lorsqu'il est mathématiquement possible de calculer les probabilités conditionnelles de θ. Or, cela n'est pas toujours possible. De plus, l'exploration de l'espace θ est ralenti considérablement lorsque les paramètres sont corrélés: les paramètres étant ajustés un à la fois, les déplacements dans l'espace des *posteriors* sont, eux aussi, limités à une dimension à la fois.

L'**algorithme *Hamiltonian Monte Carlo* (HMC)** [104] se base sur les dynamiques physiques hamiltoniennes (d'où l'appellation) pour sa phase d'exploration: cet algorithme est plus efficace que les *Random Walk* Metropolis et Gibbs, puisque l'obtention de θ_{prop} considère la forme des *posteriors* et n'est pas aléatoire. Pour ce faire, une variable *momentum* (m) (c.-à-d. une quantité de mouvement) est introduite. À chaque itération de l'algorithme, le *momentum* est défini aléatoirement, tel que $m \sim \mathcal{N}(0, \sum)$. Les valeurs de θ et m sont par la suite mises à jour pour L sous-itérations, résultant en une trajectoire d'exploration. La valeur finale θ^{L}_{prop} est ajoutée à la chaîne selon la probabilité α (Éq. 12):

$$\alpha = \frac{p(\theta_{prop}|observations)}{p(\theta_{curr}|observations)} \times \frac{q(m^L)}{q(m)} \tag{12}$$

Dans l'Équation 12, $q(m)$ réfère à la PDF de la distribution d'échantillonnage de m, soit $\mathcal{N}(0, \sum)$. Les *momentums* de la dernière itération, m, et de la présente exploration (suite aux ajustements des L sous-itérations), m^L, sont considérés.

Pour bien comprendre l'algorithme présenté ci-haut, abordons brièvement le concept de dynamiques hamiltoniennes et leurs implications dans l'échantillonnage de *posteriors*. Pour un système donné, l'énergie totale $H(\theta,m)$ est définie par l'Équation 13, où θ est une position (c.-à-d. des valeurs dans l'espace θ), m le *momentum*, et $U(\theta)$ et $K(m)$ les énergies potentielle et cinétique.

$$H(\theta,m) = U(\theta) + K(m) \tag{13}$$

Dans notre contexte d'exploration des *posteriors*, l'énergie potentielle est décrite par l'espace négatif du logarithme des *posteriors* (NLP, de l'anglais *negative log posterior*, Éq. 14). Lorsque l'on se déplace dans le système $H(\theta,m)$, l'énergie totale reste constante puisqu'il y a conversion des énergies potentielle et kinétique. Par exemple, lorsque nous nous déplaçons d'une région à faible densité vers une région à haute densité du *posterior* (c.-à.-d. d'un sommet à un creux du NLP), l'énergie potentielle diminue et, inversement, l'énergie cinétique augmente. Plus formellement, le système évolue dans le temps (t) selon les gradients $-\nabla U(\theta)$ (Éq. 15) et $\nabla K(m)$ (Éq. 16).

$$U(\theta) = -\log[p(observations|\theta) \times p(\theta)] \tag{14}$$

$$-\nabla U(\theta) = \frac{dm}{dt} = -\frac{\partial H}{\partial \theta} \tag{15}$$

$$\nabla K(m) = \frac{d\theta}{dt} = \frac{\partial H}{\partial m} \tag{16}$$

Ce sont ces équations hamiltoniennes qui dictent la trajectoire d'exploration pour les L sous-itérations. Les Équations 15 et 16 sont numériquement approximées pour de petit pas ϵ grâce à l'algorithme « saute-mouton » (de l'anglais *leapfrog*, Algo. 2). Pour chaque $l \in 1,...,L$ saut, la position (θ^l_{prop}) et le *momentum* (m^l) sont mis à jour de telle sorte à conserver l'énergie totale du système $H(\theta,m)$ [104].

Algorithm 2 Pseucode de l'algorithme saute-mouton

for L sauts **do**
  $m_{l+\epsilon/2} \leftarrow m_l - \frac{\epsilon}{2}\nabla U(\theta_l)$ ▷ Demi-saut d'ajustement de m
  $\theta_{l+\epsilon} \leftarrow \theta_l + \epsilon \nabla K(m_{l+\epsilon/2})$ ▷ Saut d'ajustement de θ
  $m_{l+\epsilon} \leftarrow m_{l+\epsilon/2} - \frac{\epsilon}{2}\nabla U(\theta_{l+\epsilon})$ ▷ Demi-saut d'ajustement de m
end for

L'introduction du *momentum* permet de considérer la forme des *posteriors* lors de la phase d'exploration, ce qui a pour effet de re-balancer l'échantillonnage en dirigeant l'exploration vers des régions qui seraient peu explorées par l'algorithme *Random Walk* Metropolis (c.-à-d. des régions à faible densité). L'efficacité du HMC, bien que globalement plus élevée que les *Random Walk* Metropolis et Gibbs, est largement dépendante du choix des paramètres ϵ (largeur des pas, de l'anglais *step size*), L (nombre de sauts pour l'exploration à même une itération) et $\sum$ (matrice des masses) [90]. De façon générale, $\sum$ est une matrice de masse diagonale [91, 104], et ϵ doit être défini de façon similaire à σ pour l'algorithme *Random Walk* Metropolis. Un trop large ϵ peut mener à des itérations dites « divergentes ». Cela signifie que, pour itération donnée, le chemin exploré via l'algorithme saute-mouton (Algo. 2) diverge du « vrai » chemin proposé par l'espace des *posteriors* [91]. L'exploration est arrêtée et la position θ^l_{prop} est considérée. Or, celle-ci est relativement arbitraire et non représentative des *posteriors*. Un haut taux de divergence indique que certaines régions de l'espace des *posteriors* n'ont pu être explorées adéquatement: les *posteriors* inférés sont dès lors peu représentatifs des « réels » *posteriors*. Le risque de divergence augmente lorsque l'algorithme explore des régions des *posteriors* où la courbure est prononcée [90, 91]. La valeur de L dicte la longueur de l'exploration à même une itération. Un trop petit L demanderait plus d'itérations pour assurer une bonne représentation des *posteriors*. Cependant, lorsque L est grand, l'exploration a tendance à revenir sur ses pas (de l'anglais *u-turn*) et devenir inefficace et redondante. Ce comportement est dicté par la forme du *posterior* et est prépondérant dans les régions de l'espace des *posteriors* aux courbures prononcées (rappelons-nous que l'exploration est dictée par les lois de la dynamique hamiltonienne). Le choix d'un L adéquat n'est pas trivial: il devrait être suffisamment grand pour permettre l'exploration de régions plates, sans pour autant mener à des *u-turns* inefficaces dans les régions courbées.

L'**algorithme du *No-U-Turn Sampler* (NUT-S)** [105] est une version adaptative du HMC, tel que le paramètre L est défini et ajusté à chaque itération. Simplement, NUT-S évalue la distance entre θ_{curr} et θ^l_{prop} de telle sorte que l'algorithme saute-mouton est stoppé lorsque cette distance n'augmente pas pour $l+1$. La valeur θ_{prop} est échantillonnée depuis l'ensemble des θ^l_{prop}. Hoffman et Gelman [105] adaptent aussi l'approche *dual averaging* [106] pour définir les paramètres ϵ et $\sum$ de façon adaptative, lors des itérations d'échauffement. La valeur de ϵ est définie, entre autres, selon une estimation du taux d'acceptation cible (δ, de l'anglais *target Metropolis acceptance rate*), tel que $100 \cdot (\epsilon L) = \delta\%$ itérations seraient acceptées avec l'algorithme HMC [104, 107]. Ces valeurs sont par la suite constantes pour le reste des itérations.

Les algorithmes présentés si haut sont les principaux représentants de la classe MCMC. Ceux-ci nous permettent d'inférer les *posteriors* non normalisés (Éq. 5) via des échantillonnages de valeurs θ_{prop} selon θ_{curr}. L'efficacité de cet échantillonnage diffère d'un algorithme à l'autre, le NUT-S [105] étant une version du HMC [104] hautement efficace et utilisée: l'échantillonnage θ_{prop} considère la forme des *posteriors* lors de l'exploration, et les paramètres de l'algorithme sont définis de façon adaptive.

1.3.3. Programmation probabiliste

La programmation probabiliste (PP) englobe l'implémentation et l'application des algorithmes présentés à la Section 1.3.2 pour un contexte d'inférence bien précis. Les algorithmes *Random Walk* Metropolis [101] et Gibbs [108] sont relativement simples à implémenter, bien que le deuxième puisse présenter des difficultés dans le calcul des distributions conditionnelles. L'implémentation de l'algorithme *Hamiltonian Monte Carlo* (HMC) [104, 109] et sa version adaptative, NUT-S [105], est plus complexe notamment dû aux calculs de gradients [90]. Divers outils (langages et *frameworks*) de PP [110–115] proposent des implémentations dites « boite noire » (de l'anglais *black box*) des algorithmes MCMC et simplifient leur application. Notons qu'il est tout de même important pour un utilisateur de tels outils de bien comprendre et connaître les concepts (Section 1.3.1) et mathématiques sous-jacents (Section 1.3.2): une utilisation naïve de MCMC *black box* peut facilement biaiser l'inférence et mener à de faux *posteriors* [90].

Les outils PP proposent un environnement flexible et adaptable au modèle bayésien probabiliste défini par un utilisateur, ainsi qu'un processus d'inférence automatique et optimisé. Cela libère l'utilisateur de la tâche d'implémenter puis optimiser lui-même le processus d'inférence [115]. Une telle implémentation requiert un important effort de validation du code et de son optimisation (c.-à-d. stabilité numérique et temps de calculs) pour être utilisé avec confiance lors d'analyses et par autrui (e.g. intégré à un outil d'analyse) [113–115].

De plus, la plupart des outils PP proposent des statistiques sommaires (e.g. moyenne, médiane) ainsi que des métriques de diagnostic (e.g. PSRF), aidant ainsi l'utilisateur dans le développement et l'ajustement de son modèle bayésien probabiliste. Les outils PP sont une alternative efficace et rapide à l'implémentation maison, lorsque les algorithmes génériques sont applicables au problème à résoudre [90].

Les langages de programmation probabiliste (LPP) BUGS [110, 111, 116], JAGS [112] et Stan [114] sont considérés comme les outils les plus établis [90, 91]. Les modèles probabilistes sont déclarés dans un langage qui est propre à chaque outil. Ces trois LLP utilisent des « blocs » pour définir un modèle (e.g. variables, paramètres, modèle). Stan est connu pour être plus simple que BUGS et JAGS, notamment par le fait que son langage soit impératif. Stan est aussi souvent favorisé à BUGS et JAGS pour sa rapidité et flexibilité: les *posteriors* sont échantillonnés selon l'algorithme NUT-S (version adaptative du HMC) plutôt qu'une version de l'algorithme Gibbs (BUGS et JAGS). De plus, Stan est accessible depuis diverses plateformes telles qu'en ligne de commande (`cmdStan`), R (`RStan`), et Python (`PyStan`) [117].

Plus récemment, des librairies propres à un langage de programmation sont proposées telles que `PyMC3` [113] et `Turing.jl` [115]. Le modèle bayésien n'a plus besoin d'être déclaré dans un langage différent de celui de l'échantillonneur (c.-à-d. l'algorithme MCMC). `Turing.jl` est entièrement développé en Julia et est particulièrement flexible puisque toute librairie Julie peut être intégrée et utilisée à même le modèle probabiliste [115]. `PyMC3` et `Turing.jl` implémentent l'échantillonneur HMC en plus de sa version adaptative, NUT-S. `Turing.jl` propose une approche moins contraignante à la PP en permettant l'inférence de paramètres θ par bloc, via divers échantillonneurs (c.-à-d. inférence composée [118]). Brièvement, cette approche ressemble à l'algorithme de Gibbs: pour une itération, un sous-ensemble de paramètres θ^A_{accept} est obtenu étant donné un échantillonneur A; le sous-ensemble de paramètres θ^B_{accept} est par la suite obtenu selon θ^A_{accept} et un deuxième échantillonneur B. Cette approche permet notamment d'intégrer des variables dont il est impossible de calculer le gradient (e.g. des variables discrètes) et dont les *posteriors* ne peuvent être inférés via l'algorithme HMC. L'inférence de telles variables peut être faite via un échantillonnage par simulation (e.g. *rejection sampling* [95], *sequential Monte Carlo* [119]). `Turing.jl` implémente un *Particle Gibbs* [120] tel que proposé par Andrieu *et al.* [121]. Bien que la flexibilité `Turing.jl` permette la création de plusieurs modèles probabilistes sans réelles contraintes, leur validité et inférence ne sont pas garanties. Il n'existe présentement aucune métrique de diagnostic pour valider de tels modèles, et leurs champs d'applications sont sujets de plusieurs discussions et remises en question.

Les outils PP facilitent grandement l'application d'échantillonneurs MCMC à des contextes et problèmes précis. Les échantillonneurs proposés sont optimisés et validés, permettant ainsi à un utilisateur de maximiser ses efforts sur le développement du modèle

probabiliste d'intérêt. Une discussion sur les outils PP utilisés dans la présente thèse est présentée à la Section 5.2.

1.4. L'inférence bayésienne et l'analyse du dose-réponse

Considérant les principales limitations observées et connues de l'estimation des métriques d'efficience via régression non-linéaire (Levenberg-Marquardt, Section 1.2.4), l'inférence bayésienne, telle que décrite à la Section 1.3, se présente comme une intéressante alternative. Comparée à d'autres méthodes alternatives proposées [36, 37, 122], l'inférence bayésienne ne nécessite pas de modifier le protocole expérimental des expériences dose-réponse et peut être appliqué rétroactivement à des expériences. De plus, cette approche permet d'obtenir les mêmes métriques d'efficience d'intérêt, assumant un même modèle mathématique (e.g. log-logistique, Éq. 2).

La représentation de ces métriques d'efficience sera cependant différente: plutôt que d'être représentées par des valeurs en un seul point (de l'anglais *single-point values*), elles seront représentées par des *posteriors*, soit des distributions des valeurs les plus probables. Ce type de représentation serait un important gain informatif, car l'incertitude de chaque métrique y serait intégrée et représentée explicitement. De plus, il nous est possible d'appliquer certaines contraintes sur nos métriques via l'intégration de *priors*. Ce contexte contraignant est alors beaucoup plus souple que d'utiliser une constante pour un paramètre, par exemple. De plus, le choix des *priors* serait représentatif de l'intuition globale des expérimentateurs, incorporant celle-ci dans le processus d'inférence plutôt que dans le processus d'interprétation. L'information contenue dans les *posteriors* permettraient aux expérimentateurs de mener plusieurs analyses post-inférence sur les métriques d'efficience. Ces analyses seraient statistiquement valables et tiendraient adéquatement compte de l'incertitude. De façon générale, l'inférence bayésienne a le potentiel de minimiser la propagation du bruit analytique dérivé du calcul des métriques d'efficience, et de modéliser les effets de tous bruits (c.-à-d. biologique, expérimental et analytique) dans le processus analytique.

Quelques travaux ont exploré l'application de l'inférence bayésienne aux analyses d'expériences dose-réponse (avant nos premiers travaux [123–130]; après notre première publication [77, 131–134]). L'ensemble de ces travaux ne correspond pas ou très peu au contexte expérimental décrit dans la Section 1.2: [124] infère les réponses pour l'identification de composés-potentiel (expériences évaluant la réponse à une concentration unique); [128, 131, 133] proposent une approche de *Bayesian model averaging* pour inférer la dose repère (de l'anglais *benchmark dose*); [77, 123, 125–127, 130, 134] considèrent des contextes expérimentaux précis (e.g. mesure de la croissance sphéroïde tumorale) et les modèles mathématiques utilisés sont soit des versions hautement contraintes du log-logistique (c.-à-d. 2 ou 3 paramètres), soit

d'autres modèles moins communs (e.g. exponentiel pour la probabilité d'infection, *spline* cubique); [129, 132] infèrent les métriques de synergie (dose-réponse de combinaison de composés).

La validité de l'approche bayésienne et son gain informatif sont peu mis de l'avant dans les travaux s'apparentant au contexte expérimental de la Section 1.2. En effet, l'argumentaire favorisant l'approche bayésienne à l'approche standard est souvent décrit théoriquement [77, 123] ou démontré qualitativement par comparaisons graphiques de courbes dose-réponse [77, 129, 132]. Les quelques démonstrations quantitatives sont faites à petite échelle (moins de 10 échantillons) principalement sur des données synthétiques, et se basent sur la comparaison de métriques telle que la RMSE. De plus, seuls [123, 131] proposent des outils implémentant l'approche bayésienne, sous forme de librarie R et de logiciel Python, respectivement. Outre des représentations graphiques des courbes dose-réponse et des *posteriors*, peu de méthodologies d'analyse post-inférence sont proposées aux experimentateurs. Le changement de représentation des métriques d'efficience en *posterior* peut présenter une limitation pour les expérimentateurs, notamment dans leur manipulation (e.g. comparaison de deux composés) et leur interprétation.

Divers facteurs limitent présentement l'application concrète par des expérimentateurs de l'inférence bayésienne à la caractérisation de l'efficience de composés via des expériences dose-réponse. La présente thèse se veut une opportunité à retirer ces limitations et à ainsi mieux outiller les expérimentateurs.

1.4.1. But et objectifs

Considérant le contexte expérimental décrit aux Sections 1.1 et 1.2, le but de la présente thèse est de mieux outiller les expérimentateurs dans leurs processus analytique et de prise de décisions en appliquant l'inférence bayésienne à la caractérisation de l'efficience de composés chimiques. Ce but général est décorticable en quatre objectifs concrets:

(1) Mettre en place un modèle bayésien généralisable à un large nombre d'expériences dose-réponse, telles que celles décrites dans la Section 1.2 et disponibles publiquement (Section 1.2.3) (O1);
(2) Démontrer quantitativement la robustesse et la validité de notre modèle bayésien en comparaison avec l'approche standard (Levenberg-Marquardt) (O2);
(3) Démontrer le potentiel informatif des *posteriors* dans le contexte du DDP (O3);
(4) Assurer l'accessibilité du processus d'inférence, des méthodologies d'analyse post-inférence et des résultats (c.-à-d. leur interprétabilité) (O4).

L'atteinte des quatre objectifs présentés plus haut peut être décrite en termes de résultats attendus. Le modèle bayésien mis en place (O1) serait applicable à un large éventail d'expériences de diverses provenances et ne partageant pas nécessairement le même protocole expérimental (p. ex. différentes concentrations, différents nombres de mesures). Bien que le modèle soit généralisable et commun, les conclusions seraient représentatives des données expérimentales de chaque expérience individuelle. Aucune manipulation des données ou des paramètres du modèle ne serait nécessaire, ce qui aura pour effet d'uniformiser le processus d'analyse sans compromettre la qualité des métriques d'efficience inférées. Nous nous attendons à ce que notre modèle bayésien soit plus robuste que l'approche standard (O2), et comble les lacunes de cette approche. Premièrement, les résultats du modèle bayésien seraient tous explicitement informatifs, peu importe la complétude de la réponse considérée : le résultat d'une réponse incomplète, par exemple, ne serait pas aléatoire et aberrant, comme pour l'approche standard. Deuxièmement, nous nous attendons à ce que les résultats d'expériences reproduites indépendamment soient plus constants avec notre modèle bayésien qu'avec l'approche standard. Troisièmement, les résultats du modèle bayésien seraient suffisamment précis et informatifs pour caractériser efficacement l'efficience d'un composé : contrairement à l'approche standard, il ne serait pas essentiel pour un expérimentateur de faire une évaluation visuelle des données expérimentales pour interpréter adéquatement les résultats des métriques d'efficience. Cette robustesse de notre modèle bayésien augmenterait la confiance d'un expérimentateur envers la validité des métriques d'efficience obtenue, et subséquemment, envers ses conclusions et décisions expérimentales. Nous nous attendons aussi à ce que la représentation des métriques d'efficience par des *posteriors* soit un outil plus informatif (O3) que les estimations retournées par l'approche standard. L'utilisation de *posteriors* permettrait d'établir des méthodologies d'analyse statistique et quantitative (plutôt que qualitative) pour évaluer et comparer les métriques d'efficience de divers composés. Plus d'analyses seraient applicables aux métriques d'efficience, augmentant ainsi la capacité d'un expérimentateur à répondre adéquatement à ses questions expérimentales.

Les prochains Chapitres aborderont nos objectifs de façon entremêlée. Le Chapitre 2 est sous forme d'article publié et présente un premier modèle bayésien (O1), une méthodologie pour comparer statistiquement deux expériences (O3-O4) ainsi qu'une interface web facilitant l'inférence de métrique d'efficience (O4). Une brève comparaison de l'approche bayésienne à l'approche standard est aussi faite (O2). Le Chapitre 3 est sous forme d'article soumis et présente une démonstration quantitative de la robustesse de l'approche bayésienne faite sur trois larges jeux de données publiques (O2). Les effets des limitations et lacunes de l'approche standard sont aussi démontrés quantitativement pour la première fois. Le modèle bayésien du Chapitre 1 est amélioré et plus généraliste (O1), et de nouvelles méthodologies d'analyse post-inférence sont proposées via un exemple de sélection de composés (O3-O4). Le Chapitre 4 présente une méthodologie pour quantifier et décrire le potentiel informatif

d'une expérience. Cette méthodologie guide l'expérimentateur dans son interprétation des métriques d'efficience (O4). Finalement, le Chapitre 5 présente une discussion sur les divers sujets abordés dans la thèse, soit les jeux de données, l'implémentation de l'inférence bayésienne, les *priors* , la quantification de l'incertitude, la description du potentiel informatif, l'accessibilité et les implications présentes et futures des travaux faits dans le cadre de la présente thèse.

Chapitre 2

Article 1 - Enhancing the drug discovery process: Bayesian inference for the analysis and comparison of dose–response experiments

Caroline Labelle[1], Anne Marinier[1,2], Sébastien Lemieux[1,3,4]

[1]Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, Qc, Canada
[2]Department of Chemistry, Faculty of Arts and Science, Université de Montréal, Montréal, Qc, Canada
[3]Department of Biochemistry, Faculty of Medicine, Université de Montréal, Montréal, Qc, Canada
[4]Department of Computer Science and Operations Research, Faculty of Arts and Sciences, Université de Montréal, Montreéal, QC, Canada

Cet article fut pulié dans *Bioinformatics* (DOI : 10.1093/bioinformatics/btz335).

Les principales contributions de Caroline Labelle pour cet article sont :

- Conceptualisation du projet;
- Développement des méthodologies proposées et présentées;
- Implémentation algorithmique;
- Curation des données expérimentales;
- Développement et mise en place de l'interface web;
- Rédaction du manuscrit.

La nomenclature utilisée dans cet article diffère légèrement du reste de la thèse. L'appellation "efficacité" (de l'anglais *efficacy*) réfère dans ce chapitre au concept d'efficience d'un composé et non à la valeur de son HDR. L'appellation IC_{50} réfère au point d'inflexion de la courbe dose-réponse, et ce, indépendamment du type de réponse considérée. Selon la

nomenclature décrite dans la Section 1.2.2, la métrique devrait plutôt être référée comme étant l'EC_{50} puisque les réponses considérées sont ascendantes (Fig. 6). L'algorithme Marquardt-Levenberg référé dans l'article est le même que l'algorithme Levenberg-Marquardt (Section 1.2.2). Finalement, la notion $\Delta\theta$ réfère dans ce chapitre à la différence entre deux *posteriors* pour une même métrique d'efficience. À ne pas confondre avec le ΔHDR du Chapitre 3 qui réfère à la différence entre les limites d'un intervalle de confiance pour un *posterior* HDR donnée.

Deux coquilles de la version publiée ont aussi été corrigées: (1) à la Section 2.2.1, au troisième point décrivant notre intuition, la notation LDR a été remplacé par HDR; (2) la notation Θ a été remplacée par θ dans les Sections 2.2.1 et 2.2.2.

Suite aux commentaires des membres du jury d'évaluation, une précision quant à l'implication des résultats a été ajoutée au résumé.

Résumé

Motivation: L'efficacité d'un composé chimique est typiquement testée via des expériences dose-réponse desquelles des métriques d'efficacité, telles que l'IC_{50} , peuvent être dérivées. L'algorithme Marquardt-Levenberg (régression non-linéaire) est couramment utilisé pour estimer les valeurs de ces métriques. Les analyses subséquentes sont cependant limitées et peuvent mener à des conclusions biaisées. Cette approche n'évalue pas la certitude (ou incertitude) des estimations et ne permet pas de comparer statistiquement deux expériences. Pour compenser ces lacunes, l'intuition de l'expérimentateur joue un important rôle dans l'interprétation des résultats et la formulation des conclusions. Nous proposons une méthodologie par inférence Bayésienne pour l'analyse et la comparaison d'expériences dose-réponse.

Résultats: Nos résultats démontrent le gain en information obtenu via notre approche Bayésienne, lorsque comparé à l'approche standard via l'algorithme Marquardt-Levenberg. Notre approche est apte à caractériser le bruit d'une expérience tout en inférant pour chaque métrique d'efficacité des distributions des valeurs les plus probables. Notre approche permet aussi d'évaluer la différence entre les métriques d'efficacité de deux expériences et de calculer la probabilité qu'une métrique soit supérieure à une autre. Les conclusions tirées de ces analyses sont dès lors plus précises, aidant ainsi grandement les expérimentateurs à sélectionner les composés pertinents et à optimiser les efforts de recherche.

Disponibilité et implémentation: Nous avons implémenté une interface web permettant aux expérimentateurs d'analyser une expérience de dose-réponse et de comparer statistiquement les métriques d'efficacité de deux expériences.

Mots clés : Inférence Bayésienne, Découverte de médicament, Dose-réponse

Abstract

Motivation: The efficacy of a chemical compound is often tested through dose–response experiments from which efficacy metrics, such as the IC_{50} , can be derived. The Marquardt–Levenberg algorithm (non-linear regression) is commonly used to compute estimations for these metrics. The analysis are however limited and can lead to biased conclusions. The approach does not evaluate the certainty (or uncertainty) of the estimates nor does it allow for the statistical comparison of two datasets. To compensate for these shortcomings, intuition plays an important role in the interpretation of results and the formulations of conclusions. We here propose a Bayesian inference methodology for the analysis and comparison of dose–response experiments.

Results: Our results well demonstrate the informativeness gain of our Bayesian approach in comparison to the commonly used Marquardt–Levenberg algorithm. It is capable to characterize the noise of dataset while inferring probable values distributions for the efficacy metrics. It can also evaluate the difference between the metrics of two datasets and compute the probability that one value is greater than the other. The conclusions that can be drawn from such analyzes are more precise, thus greatly helping experimenters when selecting relevant compounds and optimizing research efforts.

Availability and implementation: We implemented a simple web interface that allows the users to analyze a single dose–response dataset, as well as to statistically compare the metrics of two datasets.

Keywords: Bayesian inference, Drug discovery, Dose-response

2.1. Introduction

Drug discovery is a highly multidisciplinary process that encompasses the domains of biology, chemistry, computer science and mathematics [135]. A relevant therapeutic target is first identified, then different experiments are set up to analyze its activity under various conditions [9]. Such an approach makes it possible to deploy research efforts in a relevant and precise way, as well as in a context where there is a demand and a need for novel therapies.

The drug discovery process generates a very large amount of data, which often makes it difficult to manage and analyze experimental results. Analyses are thus often limited and omit a large amount of information. Intuition hence plays an important role when interpreting the results which can easily lead to biased conclusions. This work aims at developing a methodology that addresses these important issues in the specific context of dose–response experiments.

2.1.1. Dose-response experiments

The technological and biomedical advancements made in recent years have helped to accelerate the drug discovery process. For a specific assay, various chemical compounds are

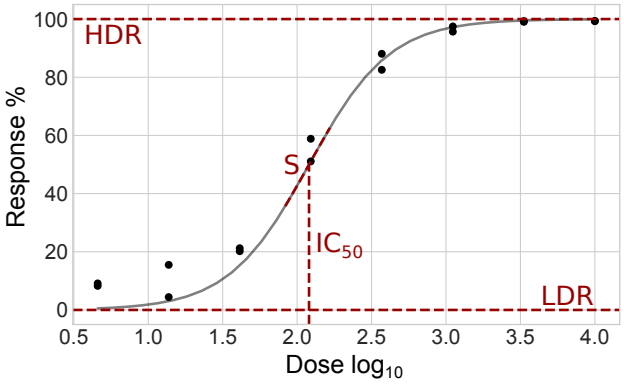


Fig. 6. Dose-response curve and efficacy metrics. Example of dose-response curve modeled by non-linear regression. The four commonly reported efficacy metrics are identified in red.

tested in order to identify those capable of generating a satisfactory response. The studied response is specific to the assay setup and can represent in- hibition of cell growth, proliferation of cells etc. High-throughput screening (HTS) allows to quantitatively characterize a very large number of compounds (several thousands per day) in an in vitro or in vivo setting. HTS also allows the rapid elimination of unfit com- pounds in the context of a specific study [9].

Screen assays are often used to assess the effectiveness of a chemical compound: it evaluates the biological response for a given dose of the compound of interest. It is possible to study single-dose responses as well as a set of responses for a dose gradient (dose–response screen). Assays can also be designed to study the effect of a combination of chemical compounds (synergistic screen). The proposed methodology described in this article is primarily applicable to dose–response screens, but its application could be widen to the other types of assay mentioned.

Dose–response screens are what we could refer to an idealized HTS experiment. It is quite typical that the effectiveness of a set of hits identified through a single-dose assay is validated by a dose–re- sponse screen [136]. For a gradient of concentrations, a compound of interest is added to well containing cells (cell lines, patient-derived cells etc.). The set of responses obtained (one for each concentration times the number of replicates) is then used to model a dose–response curve from which efficacy metrics are derived [137] (Fig. 6).

From a dose-response curve, four efficacy metrics can be derived:

IC_{50}: The dose needed to generate a mean response equidistant from minimal and maximal responses (LDR and HDR);

LDR and HDR: The asymptotic responses generated for very low and very high doses of the compound, also referred to as the plateaus of the curve (HDR: high-dose-response, LDR: low-dose response);

slope: The steepness of response transition between the two plateaus.

These metrics can be embedded into a mathematical model, the log-logistic (Eq. 17), in which a response $f(x)$ is modeled in terms of a dose x. Although there are different models [29, 30] that can be used for dose-response analysis, the log-logistic is by far the most commonly used [32].

$$f(x) = LDR + \frac{HDR - LDR}{1 + 10^{S \cdot (\log_{10} IC_{50} - \log_{10} x)}} \tag{17}$$

We often seek to identify the compound with the lowest IC_{50} [137], that is the compound capable of generating a maximal response for the lowest dose.

2.1.2. Marquardt-Levenberg

The process by which the metrics (or the model's parameters) are normally estimated is called non-linear regression. The experimental data is used to adjust the parameters of the model such that the difference between the experimental data and the dose-response curve is minimized. The regression can be identified with algorithms such as gradient descent, Gauss-Newton and Marquardt-Levenberg [33], the latter being the most widely implemented.

Various software tools are available to estimate a dose-response curve and its associated metrics [45, 59, 138]. Other tools include GraphPad, ActivityBase, the R environment and multiple Python libraries. The vast majority of these tools are not accessible to everyone, either because they are costly or because of they are complex to use. None of them allows for the comparison of two curves which limits the comparative analyzes to a qualitative numerical comparison of parameter estimates.

The non-linear regression approach, as implemented by the Marquardt-Levenberg algorithm, greatly limits the conclusions that can be made: it does not take into account the *uncertainty* of the estimated efficacy metrics. The *certainty* of the adjusted parameters and of the dose-response curve in regards to the experimental data is generally evaluated on the basis of *intuition*, based on visual inspection of the model fit. Complementary methodologies to the non-linear regression are sometimes used to compute confidence intervals. Bootstrap re-sampling [80] and Monte-Carlo simulation are among the most popular.

There is a significant need for a methodology that explicitly quantifies the reliability of the efficacy metrics taking into account the noise over the data, while adjusting the log-logistic model.

2.1.3. Bayesian Inference

Bayesian inference refers to the process of fitting a probabilistic model to a specific dataset and to represent the fitted parameters by probability distributions. The results obtained are both representative of observed and unobserved data [91].

Bayesian inference aims to infer the *posterior* probability of a hypothesis H given a dataset of evidence E and previous knowledge about H. As more elements of E are presented

to the model, the *posterior* of H is updated. The final results is a *posterior* distribution of the probability of H as described by Bayes Theorem (Eq. 18).

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)} \tag{18}$$

The probability of H given E is directly proportional to the likelihood $P(E \mid H)$ and to the *prior* distribution $P(E)$. The latter represents our intuition regarding the value of H. The *prior* is often defined by anterior evidence and observations, as well as theoretical knowledge. The likelihood evaluates the probability of obtaining E given H [139].

Given a parametric model of data $y \sim f(x \mid \theta)$, it is assumed that θ is a random variable which uncertainty can be described by a distribution, hence the *prior*. Defining the *prior* is not a trivial task and using a suboptimal *prior* can be detrimental to the analysis. In the context of dose-response, y represents an experimental response to dose x, and $\theta = \{$ IC_{50}, HDR, slope, LDR $\}$ defines the log-logistic model.

Various works have already been published on the application of Bayesian inference to the analysis of dose-response experiments [81, 123, 126, 130, 140]. Although they span a wide range of experimental contexts and their applications are well demonstrated, most methodology lacks flexibility in the type of data it can analyze. To our knowledge, no work has been done on Bayesian comparative methodology which could be beneficial to dose-response analysis.

From a software development perspective, there currently exists various platforms to facilitate the implementation and execution of probabilistic analyses. Among the most frequently cited are Stan [114] and PyMC3 [141].

2.1.4. Objectives

The methodologies currently in place limit the analysis of dose-response screens. To overcome these limitations, a significant weight is given to the *intuition* of the experimenter which can easily results in incomplete, biased and difficult to reproduce conclusions. These methodologies do not exploit the experimental data to their full informational potential and thereby impede the drug discovery process.

We aim at developing and implementing a Bayesian model for the analysis and comparison of dose-response datasets. The model incorporates the notion of *intuition* through *prior* distributions and computes the *most probable value distribution* for each of the efficacy metrics that define the log-logistic model. The comparison approach computes the *most probable value distribution* for differences between the metrics of two experiments. Finally, we want to redefine the way experimenters, such as medicinal chemists, analyze and interpret dose-response experiments by including uncertainty in their reasoning and providing them a simple and visual approach to do so.

2.2. Methods

We separated our work in three main axes: (1) the probabilistic analysis of a single dose-response dataset, (2) the comparative analysis of two dose-response datasets, and (3) the development of a web interface. The latter encapsulated the methodologies developed in the two first axes.

2.2.1. Inferring a Dose-Response Curve

We used a hierarchical Bayesian model (Eq. 19) to infer the parameters of the log-logistic model (Eq. 17) given a dataset y of dose-response data.

$$P(\theta \mid y) = \frac{P(y \mid \theta) \cdot P(\theta)}{P(y)} \tag{19}$$

For each component of θ we define a *prior* distribution $P(\theta)$. We assume that the dose-response data are normally distributed around $f(x;\theta)$ and for some shared value of σ (Eq. 20). The value of σ is also inferred but without *prior*. Its *posterior* is representative of the noise in the dataset.

$$P(y \mid \theta) \sim \mathcal{N}(f(x;\theta), \sigma^2) \tag{20}$$

To obtained the *posterior* distribution $P(\theta \mid y)$ we use the Markov chain Monte Carlo (MCMC) approach with the No-U-Turn sampler (NUTS) [114]. Summarily, we define a chain as an ensemble of values that approximate $P(\theta \mid y)$. For every i iterations of I, a set of values θ is proposed. The *posterior* $P(\theta \mid y)$ is calculated and the values of θ are appended to the chain with a probability given by the ratio of likelihood of θ and θ_i. If θ is accepted, i becomes $i+1$ and $\theta_{i+1} = \theta$; if θ is not accepted we say that the iteration as resulted in a divergence and $\theta_{i+1} = \theta_i$. Once i as reach I, a number w of the first iterations is discarded as they are *warm-up* iterations. Multiple chains C can be run in parallel and their results concatenated to generate the final *posterior* distribution, which is thus composed of $C \times (I - w)$ values.

Once $P(\theta \mid y)$ is obtained, we compute $\mathcal{N}\Big(f(x;\theta), \sigma^2\Big)$ for a wide continuous range of hypothetical x. This allow use to derive an inferred dose-response curve, which is really the sequence of median responses for hypothetical and very close to each other doses x. In the same fashion, we are also able to derive a confidence interval around the curve by aligning the $\frac{100-\alpha}{2}^{\text{th}}$ percentiles of every $\mathcal{N}\Big(f(x;\theta), \sigma^2\Big)$ for the lower bound, and the $100 - \frac{100-\alpha}{2}^{\text{th}}$ percentiles for the upper bound. By doing so, we are capable to analyze what the responses might be for untested experimental doses while characterizing their uncertainty. The data can also be analyzed by plotting the histograms of the *posterior* distributions of θ. Confidence interval and median values can easily be derived from these distributions.

We tested our Bayesian model for various setups and multiple contexts (see Section 2.3). As demonstrated in the following section, an important aspect of Bayesian inference is the definition of the *prior* distributions. The current paper only presents analyzes done on inhibition rate (%) responses (see Section 2.2.4), that is responses that range from more or less 0 to 100, and increase as the doses increase. Our general *intuition* regarding the values of the efficacy metrics is as follow :

- We would expect the IC_{50} to be around the median experimental dose (assuming an appropriate range of doses has been tested); We are assuming its value could span a very large range of hypothetical doses while above the *absence of compound* dose;
- The HDR should have a positive value and should more or less have a maximal value of 100%; We do not assume that its value is caped at nor will reach 100%;
- We would expect the slope (slope) to be positive (inhibition rate response); We do not restrict it to have a positive value;
- The LDR should be somewhere around the 0% mark.

Following these elements of *intuition*, we tested different *prior* distribution in order to assess their effects on the inferred *posterior* distributions. Our model could easily be applicable to other type of responses (eg. survival rate) by adjusting the *prior*.

2.2.2. Comparing Two Dose-Response Curves

To further our analysis approach and to propose a novel methodology, we adapted our Bayesian model so that we can infer the probability that two curves have significantly different components of θ.

Given two dose-response datasets D_1 and D_2, we are asking *What is the probability that θ_k of D_1 will be greater than that of D_2?* In order to answer this question, we evaluate the *posterior* of differences between θ_{1k} and θ_{2k} (Eq. 21)

$$P(\Delta\theta \mid \theta_1, \theta_2) \tag{21}$$

Posterior distributions are inferred for D_1 and D_2 in parallel. For every accepted θ appended to the chain, $\Delta\theta = \theta_2 - \theta_1$ is computed and stored. In the end, the w first elements are discarded, just as for the other *posterior*. We can evaluate the probability that each data has the largest value for θ_k by calculating the ratios of positive (D_1) and negative (D_2) *posterior* values. To facilitate the interpretation, we plot the histogram of the differences *posterior* with a contrasted vertical segment marking the median difference. It is also easy to calculate confidence interval and evaluate the reliability of the comparison.

This comparative methodology takes into account the *uncertainty* of θ which is currently ignored when comparing two dose-response curve. We tested our approach on both synthetic and experimental results, and the results proved to be more informative than the simple qualitative comparison.

2.2.3. Implementation

Our Bayesian model is implemented in the modeling language Stan [114]. We use 4 chains of 2000 iterations and 1000 warm-ups to compute the *posterior*. We use the PyStan interface (v2.18.0.0) to work with Stan in the Python (v3.0.0) environment. Our plots are generated with Matplotlib (v3.0.2). When comparing our model to the Marquardt-Levenberg algorithm, we used the *optimize* package of Scipy (v1.2.0) with default settings to implement the non-linear regression.

For our web interface, we use Flask (v1.0.2) and Python on the server side. On the client side, standard HTML5 and JavaScript is used as well as Jinja and Bootstrap (v3.3.7). Interactivity is mainly provided by the use of jQuery (v2.1.1).

2.2.4. Dose-Response Data

We use various datasets to test and demonstrate the efficacy of our proposed approach. We use both synthetic and experimental datasets.

Using synthetic data allows us to evaluate the efficacy of the various approaches tested in a controlled environment. These data are generated from the log-logistic model (Eq. 17). For a given set of 10 hypothetical doses x and defined $\theta = \{\text{IC}_{50}, \text{HDR}, \text{slope}, \text{LDR}\}$, we compute the associated $f(x;\theta)$ responses. Noise is added to dataset by sampling from $\mathcal{N}\left(y_j, \sigma^2\right)$ for each response y_j. We used multiple σ to test how well our methodology dealt with noise. The various synthetic datasets used in Section 2.3 are described in Table 1. When referencing a synthetic dataset, we use the label of Table 1 to which we add the σ value in subscript. For instance, A_0 would describe a dataset with an IC_{50} of 2.15, a HDR of 60 and a Gaussian noise of $\sigma = 0.1$.

Table 1. Synthetic datasets

Label	HDR	IC_{50}	σ
A	60	2.15	{0.1, 10 }
B	60	2.0	{0.1, 5, 10 }
C	90	2.15	{0.1, 5, 10 }

slope = 0.8 and LDR = 0.0

We also used real experimental data to demonstrate the application of our proposed methodology. The datasets E_1, E_2 and E_3 are from a single assay and represents different compounds. The compounds were tested at 8 concentrations against patient-derived leukemic cell. The response measured is representative of cell growth inhibition rate (%). The experimental data was obtained through the Leucegene project.

Our proposed Bayesian model is unaffected by the number of replicates R (number of measured responses for each concentration). R varies from one experimental setting to another: to demonstrate the flexibility of our approach, we generated synthetic datasets with $R = \{1, 3\}$ and used experimental datasets with $R = 2$.

2.3. Results and Discussion

Results presented in this section are obtained by analyzing both synthetic and experimental datasets (Section 2.2.4). Most of the figures are adaptation from our web interface (Section 2.3.3).

2.3.1. Bayesian Inference on Dose-Response Data

We evaluated the efficacy and limits of our Bayesian model in various experimental contexts. We first compared its results to those obtained by non-linear regression (Marquardt-Levenberg algorithm). We then assessed the effects of various *prior* in order to define the most appropriate. Lastly, we discussed the inferred σ *posterior* distributions for multiple datasets.

2.3.1.1. Marquardt-Levenberg vs. Bayesian Inference

We did not optimized the Bayesian *prior* but they were chosen wisely. As for the Marquardt-Levenberg algorithm, we tested it for both the four-parameters (4P) log-logistic model (Eq. 17) and a two-parameters model (2P). In the 2P model, only the IC_{50} and slope parameters are estimated: the HDR and LDR are fixed to constant values, 100 and 0 respectively. These three approaches are applied to two synthetic datasets with varying noise (A_0 and A_{10}) and on an experimental dataset (E_1). The results are reported in Figure 7.

Both Bayesian inference and Marquardt-Levenberg 4P generate the expected values when the data has very little noise (A_0, black dataset). The median HDR and adjusted LDR are the same (60.1) and contrary to the Marquardt-Levenberg 2P, they stay around the 60% mark. As expected, Marquardt-Levenberg 2P generates a dose-response curve that is not representative of the dataset as its HDR is fixed at 100%. This forces the IC_{50} to shift to the right (3.62) and the slope to flatten (0.286).

When the data is noisier (A_{10}, orange dataset), Marquardt-Levenberg 4P generates curves that resemble the expected model the most. Its curve is steeper (1.36), which can be explained by its high LDR (11.1). The estimated HDR is as expected (58.9) and there is a small shift in the IC_{50} (2.00). The Marquardt-Levenberg 2P curve is mostly the same as for the A_0 dataset. Interestingly, the Bayesian inference results differ from the previous ones. First, the confidence interval (95%) surrounding the curve is significantly larger. Second, the median HDR now reaches well above 60% (86.5), creating a shift in the IC_{50} to the right

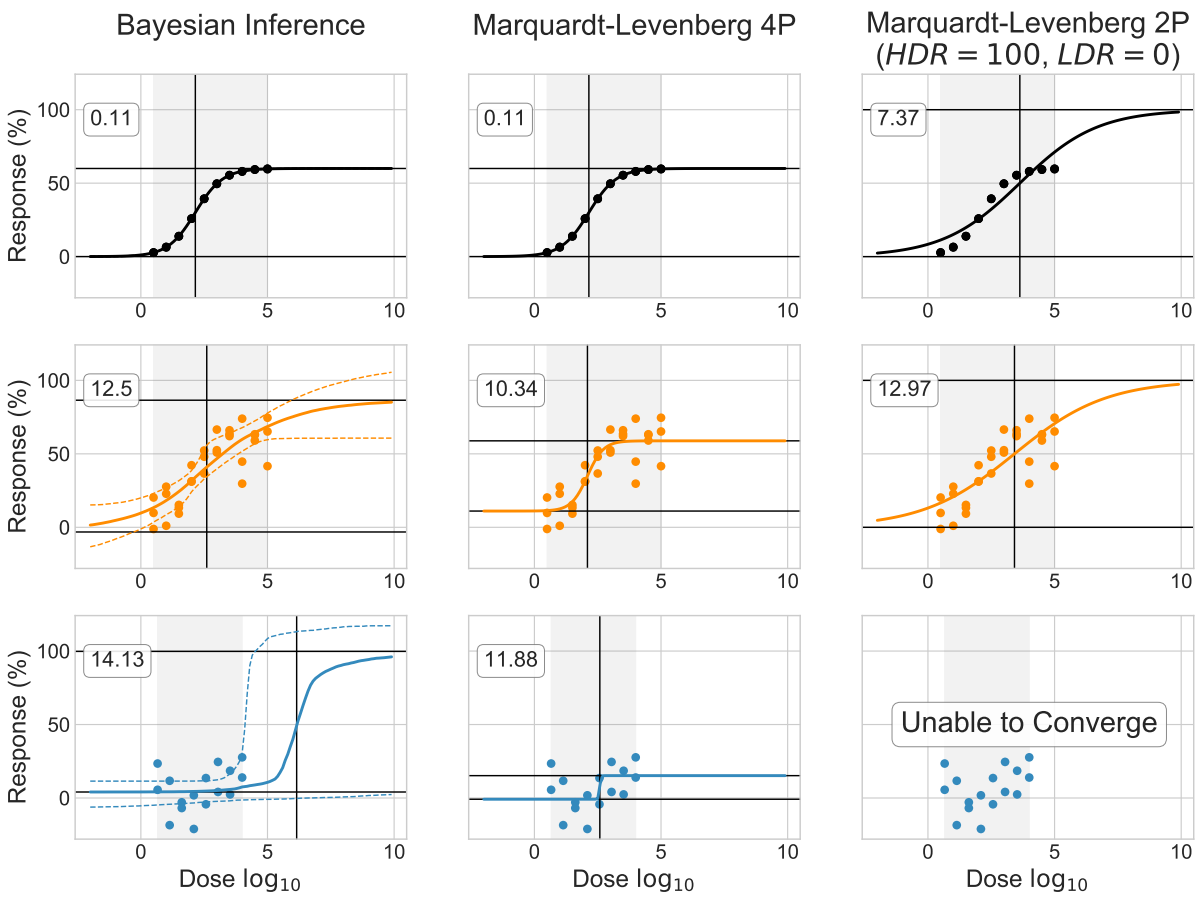


Fig. 7. Marquardt-Levenberg vs. Bayesian inference. Both synthetic datasets A_0 (black) and A_{10} (orange) have triplicate ($R = 3$). The experimental dataset E_1 (blue) displays no response in the range of doses tested. Our Bayesian model is used to estimate dose-response curves with a 95% confidence interval. Marquardt-Levenberg estimates the parameters of the four-parameters (4P) log-logistic model (Eq. 17) and of the two-parameters (2P) model (only the IC_{50} and slope parameters are estimated: the HDR and LDR are fixed to constant values, 100 and 0 respectively). HDR and LDR (median values for Bayesian Inference) are represented by horizontal black segments; IC_{50} (median value for Bayesian Inference). Root mean square error (RMSE) value is identified for each curve. For Bayesian Inference, we used the median curve to compute the residuals

(2.60) and a flatter response (slope of 0.290). Even though the curve *seems* to represent well the data, the median parameters do not approximate those expected, with the exception of the LDR (-3.17).

Compared to the two datasets presented above, E_1 (blue dataset) completely breaks both Marquardt-Levenberg 4P and 2P. The latter is simply unable to converge (when using Scipy's implementation, see Section 2.2.3). As for the former, it returns a very low HDR (15.2) and an unrealistically steep slope (18.9). Confusingly, the IC_{50} estimated could lead to erroneously conclude that the compound is active ($IC_{50} = 2.58$). The Bayesian inference curve better models the absence of response over the range of doses tested. The curve's inflexion, or IC_{50}, is largely out of the experimental range and reaches a median value of 6.23.

The confidence interval surrounding the right side of the curve (outside of the experimental doses range) is extremely wide: its bound span from approximately 120% to 0%. The dataset E_1 is not sufficient to infer precisely efficacy metrics, but sufficient to clearly indicate the lack of response for this compound over the range of doses tested. Finally, since the Marquardt-Levenberg 4P directly minimizes the RMSE, it is not surprising that it achieves overall lower values.

It is interesting to interpret the results of Figure 7 by comparing how each methodology handles the concept of *intuition*. Marquardt-Levenberg 4P has no implemented consideration for it: only the data is considered when computing the estimates for the parameters of the model. This greatly limits the analysis to the range of experimental doses, as the approach assumes that the lower and upper response plateaus have been experimentally observed. This explained the unrealistic dose-response curve obtained for E_1 (absence of the high dose plateau). If we were to analyze this dataset only by looking at its IC_{50}, which is common practice, we would conclude that the tested compound is somewhat active. If we were to further our analysis to the other parameters, we would be puzzled by the very low HDR. The dataset would most likely be discarded because of the small distance between the LDR and HDR estimates and/or because of the unusual shape of the curve. The decision to discard E_1 is entirely based on *intuition* from the experimenter. The Marquardt-Levenberg 2P does take into account the notion of *intuition* in its implementation, but in an extreme way. By fixing the HDR and LDR to constant values, we imply that our *intuition* is rather a *certitude*. Again, this methodology is highly limiting, since our *intuition* prevails over the data. This is exactly what happened during the analysis of A_0. In the case where the data does not fit our *intuition* (E_1), the algorithm simply does not converge and the dataset is discarded. Neither of the Marquardt-Levenberg methodologies are capable of considering both the data and our *intuition* in a complementary fashion: it is one or the other. As demonstrated in Figure 7 this can highly bias our conclusions.

Our proposed Bayesian inference methodology is a good alternative to the problematic Marquardt-Levenberg. The use of *prior* allows us to incorporate the notion *intuition* into the computation in a less drastic way than Marquardt-Levenberg 2P. Thus, the resulting dose-response curve can be expanded to doses that were not tested experimentally. This approach also allows for the quantification of *uncertainty*, which neither of the Marquardt-Levenberg approaches do. For instance, we can conclude with certainty that E_1 does not support an IC_{50} within the range of doses tested and the compound can be eliminated from further studies.

Even though Bayesian inference is better suited for the analysis of dose-response data than the Marquardt-Levenberg algorithm, it still presents some limitations as demonstrated by the analysis of A_{10}: inappropriate *prior* combined with high noise can skew the results (Figure 7). The following section discusses this topic in more details.

2.3.1.2. Defining *prior* distributions

We must think of *prior* as safety nets: when the data is insufficient, the inference gradually falls back on the *prior* distributions. It is thus important to use appropriate *prior* that best represent the experimental context. The process of defining the most suitable *prior* for θ is referenced as *prior elicitation*. It can either be based on consensus notions regarding θ [142], or on beliefs [143]. The latter corresponds to our aim of mathematically implementing the notion of *intuition*.

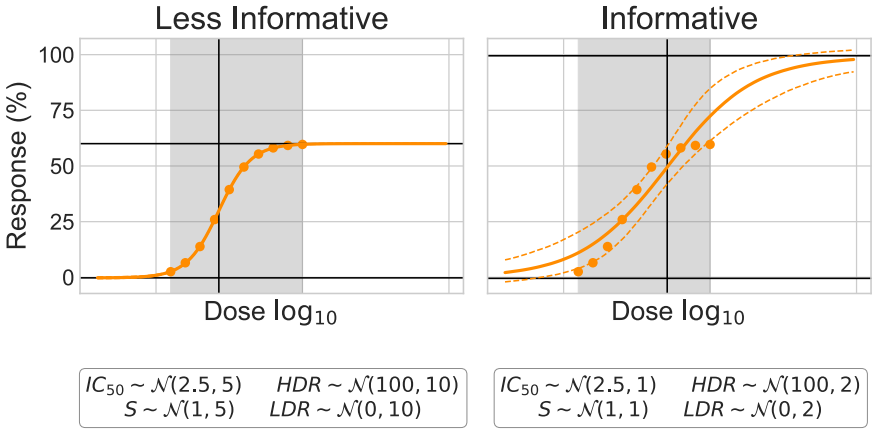


Fig. 8. *Prior* informativeness. Informativeness can be described by the wideness of the distribution. "Informative" (narrow) *prior* prevail on the data and the inference is biased. "Less Informative" (wide) *prior* do not overshadow the data and the curve is inferred with high certainty. We used the A_0 dataset with $R = 1$. The median HDR and LDR are represented by black horizontal segments, and the median IC_{50} vertical segments. RMSE value is identified for each curve. We used the median curve to compute the residuals

When comparing both Bayesian inferences, it is clear that the "Informative" *prior* are not suited to the data (Figure 8). Even though both HDR *prior* are centered at 100%, the "Less Informative" *prior* does not prevail over the data and parameters can be inferred as expected (2.14, 60.1, 0.801, 0.032 for the IC_{50}, HDR, slope and LDR respectively). The second curve is reminiscent of the one obtained when using Marquardt-Levenberg 2P (Figure 7). In such case, the *prior* are highly restrictive and do not complement the data, causing the inferred curve to mainly be representative of the *prior* themselves.

Figure 8 illustrates the effect of *prior* informativeness on 10 data points ($R = 1$). The undesirable effects of "Informative" *prior* can be counterbalanced by giving more data points to the Bayesian model. For example, A_0 dataset with $R = 5$ prevails on the "Informative" setup. In the context of dose-response analysis, it is not always possible to generate large dataset due to cost and material limitations. *Prior* should thus be defined by less informative distributions.

We tested various setups of *prior* distributions (Table 2) in order to establish the ones that can generalized to multiple experiments with similar contexts. Again, we used the

synthetic dataset A_{10} with $R = 1$. The dose-response curves and *posterior* distributions are presented in Figure 9.

Table 2. Distributions Parameters

Description	IC_{50}	HDR	slope	LDR
More Informative Normal Dist.	$\mathcal{N}(2.5, 5)$	$\mathcal{N}(100, 10)$	$\mathcal{N}(1, 5)$	$\mathcal{N}(0, 10)$
Less Informative Uniform Dist.	$\mathcal{U}(-15, 45)$	$\mathcal{U}(0, 150)$	$\mathcal{U}(-10, 10)$	$\mathcal{U}(-50, 50)$
Less Informative Normal Dist.	$\mathcal{N}(\hat{x}, 10)$	$\mathcal{N}(100, 20)$	$\mathcal{N}(0.5, 10)$	$\mathcal{N}(0, 20)$

Note: $\hat{x}$: median of experimental doses

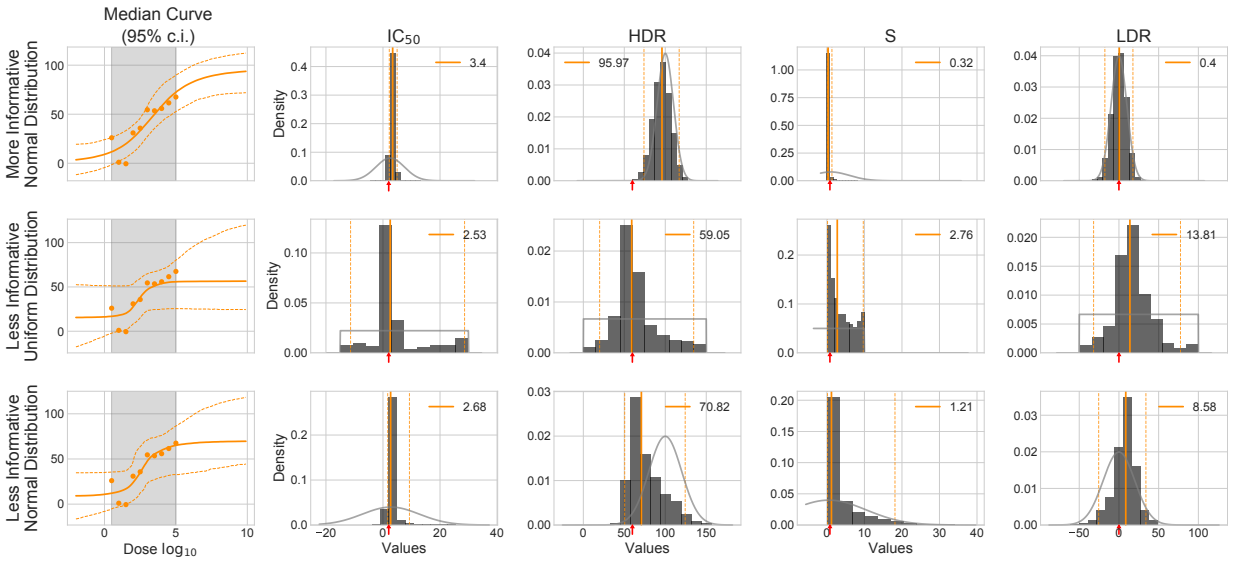


Fig. 9. Effects of various *prior*. Three different *prior* settings (Table 2) are tested on the A_{10} synthetic dataset with $R = 1$. Dose-response curves are plotted against the data and with a 95% confidence interval. The *posterior* of each parameter is represented by an histogram. The colored vertical segments represent the median value (continuous) and its 95% confidence interval (hashed). The numerical median value is indicated in the legend. The expected values (used to generate the synthetic data) is identified by a red arrow on the x-axis. The light gray segment superimposed on the histogram illustrates the contour of the *prior* distribution.

The "More Informative Normal Dist." *prior* resulted in a higher than expected HDR (96.0) which generates a shift to the right in the IC_{50} (3.40). The slope is also flattened by this high HDR and its value diverges greatly from the expected one. Interestingly, the HDR *posterior* is highly similar to the *prior*. Similarly, the LDR *posterior* is also matching its *prior*. When looking at the data, we notice that there are no clearly define upper and lower plateaus: the inference must thus rely mainly on the *prior* to define these regions of the dose-response curve. Even though the *prior* distributions are not highly informative, they are still too informative and force the HDR to reach the theoretical optimal 100.0% even though it is not directly supported by the data.

The "Less Informative Uniform Dist." *prior* are the less informative out of the three settings. Only the median HDR is approaching the expected values (59.1) but its confidence interval (95%) is quite large. The other inferred parameters do not resemble those expected, which is not surprising considering the noise present in the data. When comparing the *posterior* distributions to the *prior*, we noticed that they were bound by very similar limits with the exception of the slope, which has a lower bound of 0.

The "Less Informative Normal Dist." *prior* seems to be a good compromise between the two previously described settings. The median values are not as expected but this can be explained by the noise in the data, mainly in the low dose region. The median HDR is however not too far from the 60% mark. It is interesting to notice the shift between the *posterior* and the *prior* of that parameter, which is not observed in the other two settings.

Overall, normally distributed *prior* ($\mathcal{N}(\mu, \sigma^2)$) appear more appropriate. The uniform distributions *prior* ($\mathcal{U}(\alpha, \beta)$) are too uninformative: when data is insufficient, the distribution values suggested by the *prior* are all equally probable which has the same effect as adding a large amount of new noisy data. This could explain the very large confidence intervals when using uniform *prior*, with the exceptions of the slope. In addition to the lack of informativeness in regards to the most probable value, uniform distributions are constrained by their α and β parameters. For instance, the slope *posterior* abruptly stops at 10 which is incidentally the defined β we selected for the slope uniform *prior*. Comparatively, the normal distribution is not bound and each distribution values as its own probability. We also adjusted our intuition of μ for both the IC_{50} and slope *prior* (Table 2). Assuming the experimental doses are sufficient and range on a two-folds scale, we could expected the IC_{50} to be near the median experimental dose.

We will be using the "Less Informative Normal Dist." *prior* as default settings for now on.

2.3.1.3. Unresponsive Data

So far, we mainly used synthetic datasets to explore the application of our Bayesian model to the analysis of dose-response data. To asses the extend of the applicability of our approach, we applied it to the analysis of a seemingly unresponsive experimental dataset, E_1 (Figure 10). This type of response is frequent during the drug discovery process and it is of the utmost importance that the analysis approach applied can confidently assess that this compound has an IC_{50} value above the range of doses tested and must be discarded.

On this specific dataset, the responses never reach more than 30% and there is no clear tendency. The inferred dose-response curve is mainly flat for the entire experimental doses range. The median IC_{50} is high (9.37). As we expected it, the HDR *posterior* is highly reminiscent of the *prior*: the data did not give any indication regarding the response at very high doses. All the parameters' confidence interval are quite large. We are not able

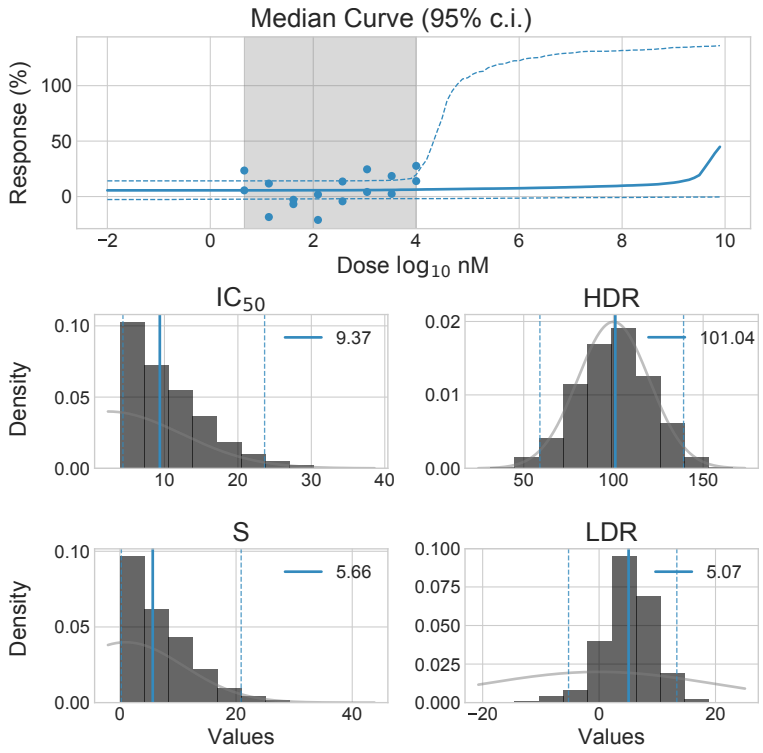


Fig. 10. Bayesian inference applied to seemingly unresponsive experimental data. Results obtained for the experimental data E_1 using our default *prior* settings. Parameters *posterior* are represented by histograms. Their median values is identified by the colored full vertical segments and the values are reported in the legend. The colored dashed vertical segments mark the 95% confidence interval bounds. The light gray segment superimposed on the histogram illustrates the contour of the *prior* distributions.

to determine with certainty the efficacy metrics of the tested compound. We can however conclude with certainty its IC_{50} is bigger than 5 $\log_{10} nM$, which is enough to discard this compound as ineffective. Such a high certainty conclusion can not be made on seemingly unresponsive dataset with commonly used Marquardt-Levenberg algorithm methodology, without resorting to *ad hoc* rules.

2.3.1.4. Inferring noise

Our Bayesian model also infers a *posterior* distribution for σ (Eq. 20), which describe the amount of noise in the dataset. For synthetic datasets (A_0 and A_{10}), the median σ is close to the actual σ used to generate the data (Figure 11). It is true that we used a Gaussian noise when generating the data, and that our Bayesian model assumes that the responses are from independent identical normal distributions. That being said, the median σ for the experimental dataset E_1 is very close to the standard deviation of responses for this dataset (Figure 11), which corresponds to interpreting the dose-response as flat and corresponding to the LDR plateau.

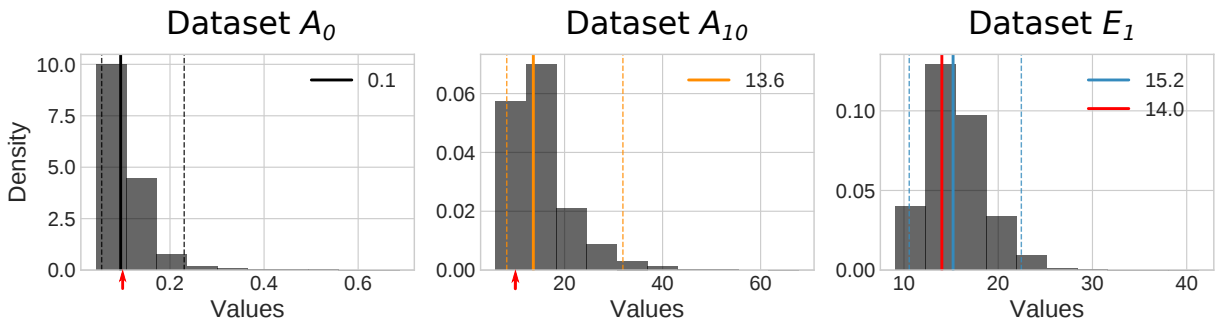


Fig. 11. A *posterior* distribution of σ. *Posterior* distributions obtained by applying our Bayesian model to the synthetic datasets A_0 and A_{10} (black and orange respectively), and to experimental dataset E_1 (blue). The median σ are represented by the full segments and their values are reported in the legend. The dashed segments mark the bounds of a 95% confidence interval. For the two synthetic datasets, the real value of σ is identified by a red arrow on the x-axis. For the experimental dataset, the standard deviation of the responses is represented by the red full segment and its value is reported in the legend.

2.3.2. Comparison of Two-Dose-Response Datasets

To further the analysis of dose-response data, we proposed a novel comparison methodology.

As mentioned in Section 2.2.2, the comparison is done by inferring the *posterior* of the difference between two values of an efficiency metric. From these *posterior*, we can derive the probability that a dataset has the largest value for a given efficiency metric. The uncertainty identified through the individual dose-response inference is carried to our comparison analysis, which allows to characterize the uncertainty of the difference.

When comparing the synthetic datasets B_0 and C_0, we can conclude with great certainty that the C_0 IC_{50} is larger than that of B_0, even though the difference between the value is quite small (~ 0.15). We can also conclude with great certainty that C_0 has a higher HDR than B_0. The precision of both datasets ($\sigma = 0.1$) allows us to draw these conclusions without doubt.

In contrast, when comparing B_5 to C_5 we can not make such conclusion. These two datasets share the same parameter values as B_0 and C_0, respectively, but they were generated with increased noise ($\sigma = 5$). The inferred IC_{50} are more uncertain and their *posterior* distributions overlap. When comparing their respective median IC_{50}, one could easily concludes that the B_5 dataset has a larger IC_{50} than the C_5 dataset ($2.29 > 2.15$) and that the B_5 compound is thus less effective than that of C_5. This conclusion is highly biased: the uncertainty of the inferred IC_{50} does not allow for the identification of significantly greater value as demonstrated by the ΔIC_{50} *posterior*. If we take a look at the comparison of HDR values, we notice that the uncertainty does not affect the comparison: the values are different enough that the two *posterior* do not overlaps. We thus can conclude with certainty the C_5 HDR is greater than the B_5 HDR even when considering their respective uncertainty.

Similar results have been observed when comparing the two highly noisy ($\sigma = 10$) that are B_{10} and C_{10}. The difference in median IC_{50} is even greater but the Δ IC_{50} *posterior* tends more toward the expected conclusion. The HDR comparison is still highly convincing despite an higher level of uncertainty in the individual HDR *posterior*.

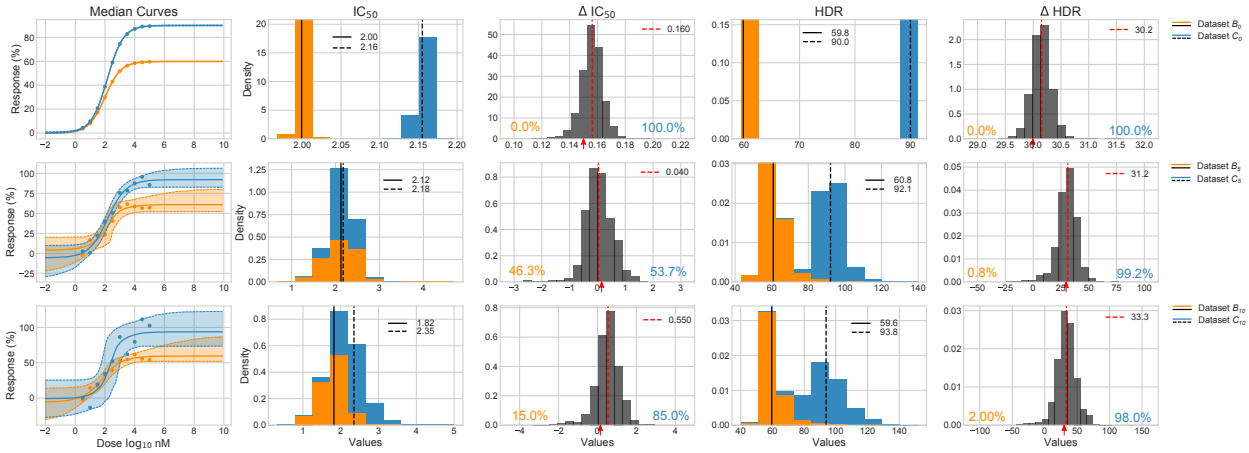


Fig. 12. Comparison of synthetic datasets. Three pairs of synthetic datasets with $R = 1$ are compared: B_0 to C_0, B_5 to C_5 and B_{10} to C_{10}. Each pair of datasets differs in their IC_{50} and HDR values. The 95% confidence interval of each median curve is represented by the colored shaded regions. For both IC_{50} and HDR, the stacked individual *posterior* are represented by the colored histograms. Their median are marked by black segments. The numerical values are reported in the legend. The Δ IC_{50} and Δ HDR *posterior* are represented by gray histograms. The true difference is identified by a red arrow on the x-axis (0.15 for the IC_{50} and 30 for the HDR). The median values of the differences *posterior* are identified by the red hashed segments, and the numerical values are reported in the legend. The probability (in %) that a dataset has the largest value for a given parameter is identified on the graph in the color corresponding to the dataset.

To highlight the informative gain of our comparative approach, we compared and analyzed two experimental datasets (E_2 and E_3) using two methodologies: (1) the commonly used numerical comparison of IC_{50} and (2) our differences *posterior* approach.

Numerical comparison When comparing the IC_{50} median values, we notice they differ by 0.11 $\log_{10}$ nM (Figure 13.A) which is equivalent to approximately 32 nM. We would conclude that the E_3 dataset has a larger IC_{50} than that of the E_2 dataset. The E_2 compound is thus seemingly more effective than the E_3 compound.

Differences *posterior* When we first look at the Δ IC_{50} *posterior* we can not conclude that one of the IC_{50} is greater than the other. The IC_{50} were not inferred with enough certainty, because of the noise present in the data, for us to conclude that their values are significantly different. The HDR are however significantly different, despite the great uncertainty of E_2 HDR (Figure 13.A). The Δ HDR *posterior* identify the E_3 dataset as the one with the overall largest HDR. It is also interesting to note that the difference between the two HDR is quite large, with median difference of almost 23% (Figure 13.A). The E_3 also have the overall largest slope (Figure 13.A). When combining all of these information,

we can conclude that the E_3 compound is more effective at generating a maximal response than the compound of E_2.

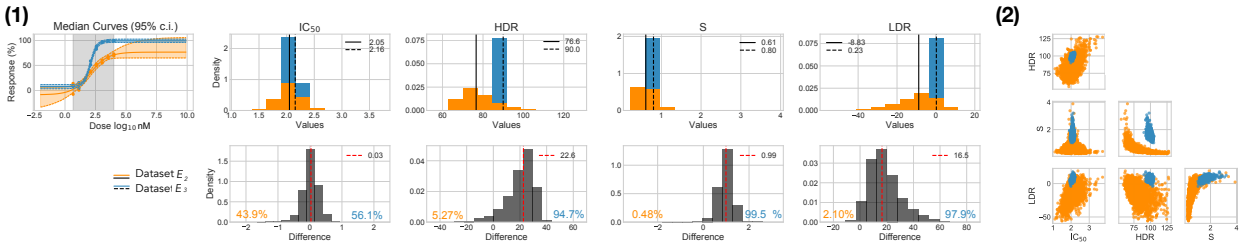


Fig. 13. Comparison of experimental datasets. Datasets E_2 and E_3 are compared. (1) The 95% confidence interval of each median curve is represented by the colored shaded regions. The individual *posterior* are represented by the stacked colored histograms. The median values are indicated by black segments and the numerical values are reported in the legend. The Δ *posterior* are shown as gray histograms with their median values represented by red hashed segments. The probability (in %) that a dataset has the largest value for a given efficiency metric is identified on the graph in the color corresponding to the dataset. (2) Pairwise comparison of the individual *posterior*. Each dot a single value of the *posterior* for a given efficiency metric. Datasets are identified by their colors.

The two conclusions greatly differ and the one drawn from the numerical comparison is biased. The numerical comparison methodology is highly limited as it only considers one efficacy metric and does not consider the uncertainty associated to its values. It is preferable to consider all four metrics to get a more complete characterization of the efficacy of a compound. We must also evaluate the probabilities of certainty on the metrics as well as on the comparison itself to ensure our conclusions are as precise as is appropriate.

Lastly, interpreting pairwise plots of the *posterior* distributions can also help to draw informed conclusions. This sort of representation can identify inter-parameter dependencies which should be considered when analyzing *posterior* distributions. We can observe in Figure 13.B that both datasets are distinguishable by pairing their HDR and slope, which was not observable from the analysis of the histograms of Figure 13.A.

2.3.3. BiDRA: an Online Tool

The two previous sections demonstrated how well and how much more information can be gathered when using our proposed Bayesian methodology for the analysis and comparison of dose-response data. The conclusions drawn from such analyses are less prone to bias compared to other commonly used methodologies. We are aware that the implementation and subsequent application of our Bayesian approach is not within everyone's reach. We thus decided to develop an easy-to-use web interface, *BiDRA* (Bayesian inference for Dose-Response Analysis).

The interface proposes both the analysis of a single dataset (Sections 2.2.1 and 2.3.1) and the comparison of two datasets (Sections 2.2.2 and 2.3.2). For both analyses, the user simply uploads the dataset(s) in a CSV format with the first column corresponding to the

doses and the second representing the associated responses. It is important that the doses be log-transformed since we are using the log-logistical model (Eq. 17). The data type must then be specified: *Inhibition* if the response increases with the dosage; *Activity* is the response decreases as the dosage increases. The HDR and LDR *prior* are adjusted according to the response type. We suggest default *prior* distributions (Section 2.3.1.2), assuming the data represent some sort of rate (%). The user can however easily specify his own μ and σ for each parameter.

The results are returned in both a figure and in a table. For the single dataset inference, the median dose-response curve as well as the *posterior* of all efficacy metrics and the σ are plotted. The returned results are similar to Figure 10. For the two datasets analysis, the individual inference plots are returned as well a figure describing the comparison. The latter includes the stacked individual *posterior* as well as the differences *posterior*. As an example, figure 13.A was obtained from BiDRA. For every computed *posterior*, we return its median and the bounds for 10%, 5% and 1% confidence intervals in a table.

The interface is accessible (https://bidra.bioinfo.iric.ca/) and does not require any authentication. The interface is not connected to any database and the analysis are not saved. We plan on adjusting the interface as our work progresses (see Section 2.4).

2.4. Implications

We propose in this paper a Bayesian inference methodology for the analysis of dose-response data. This approach is then extended to directly infer differences in efficacy metrics between two dose-response experiments.

Our approach addresses two limitations of the commonly used Marquardt-Levenberg algorithm: first, it yields a single point estimate for each efficacy metrics, with no assessment of the uncertainty for these values. The experimenter is then left to decide on whether to accept or reject a given fit based on its *intuition*. This process is typically manual leading to possible biases and difficulty to reproduce analysis results. The second limitation is that the Marquardt-Levenberg algorithm relies entirely on the experimental data to estimate the efficacy metrics. In cases where the data is insufficient to determine one of the efficacy metrics, this algorithm will settle for the mostly likely value without consideration for experimentally sound boundaries. These limitations are compounded by the fact that there exists no methodology to support direct comparison of dose-response curves besides numerically comparing the efficacy metrics. The Bayesian inference approach we describe here allows us to incorporate in the analysis of dose-response the notion of experimental *intuition* to guide the identification of plausible ranges for each of the efficacy metrics. This reduces the necessity for careful inspection of curve fitting and provide a sound statistical framework to communicate the reliability of estimates to the experimenter. Our approach

shares similarities to the ones presented in [81, 130, 140, 144, 145] as it implements a simple hierarchical Bayesian model. We consider as part of our analysis all efficacy metrics of the log-logistic model (Eq. 17). We also propose a novel and informative approach to compare two dose-response curves, again unambiguously conveying estimates uncertainty as *posterior* distributions of the differences for efficacy metrics of interest. In practice, these distribution are either communicated as a probability that one value is larger for one experiment than in the other, or as a confidence interval on the difference.

As mentioned by [145], the Bayesian inference still have some limitations even though it provides numerous advantages when compared to the usual non-linear regression approach. As for Marquardt-Levenberg, computation time increases with the number of data points under consideration: the analysis of A_{10} ($R = 3$) (Figure 9) took ~ 0.8 seconds, while the analysis of A_{10} ($R = 1$) (Figure 9) took ~ 0.5 seconds (Intel, i9-7920X). Comparatively, the comparison of B_{10} and C_{10} (both $R = 1$) (Figure 12) took ~ 3 seconds. In most practical settings, the computational time necessary for these analysis is insignificant to the time required for actually performing the experiments being analyzed. A more important limitation to consider is the difficulty to clearly express the relative weight of the *prior* in the analysis. As shown in Section 2.3.1.2, an inappropriate *prior* can greatly alter the *posterior* distributions. This effect is mostly seen for the HDR and LDR as they often depend on extrapolation of the experimental data. As a general rule of thumb, the *prior* informativeness should not outweigh the data information and least informative *prior* should be favored in most situations.

That being said, we do think our approach to directly compare two dose-response will provide a useful tool to support the drug discovery process, either at the stage of secondary validation following a primary screen or during compound optimization. Considering distributions of *probable values* instead of single point estimates brings more depth to interpretation efficacy metrics and supports better informed decision from the experimenters. These benefits are also attained through a method that better support automated analysis as we greatly reduced the necessity for manual inspection of each fit.

Finally, we would also like to emphasis the flexibility of the proposed framework. We are currently exploring the use of this approach in the context of primary screens based on high-throughput, single-dose assays or to the more complex context of two-compounds synergistic dose-response assays. There is currently no established methodologies for the analysis of these types of assay. We think that Bayesian inference would be highly beneficial and could help to more reliably identify compound *hits* as well as better quantification of compounds interactions.

Acknowledgements

The authors would like to thank Geneviève Boucher, the Sauvageau's Lab (IRIC) and the Leucegene team for their help and contributions.

Funding

This work has been supported by Genome Canada and Genome Quebec.

Chapitre 3

Article 2 - Bayesian Inference as a Robust Alternative to Non-Linear Regression for Dose-Response Efficiency Metrics Assessment

Caroline Labelle[1], Petr Smirnov[4,5], Maud David[1], Mario Callejo[1], Benjamin Haibe-Kains[4,5,6,7,8,9], Anne Marinier[1,3], Sébastien Lemieux[1,2]

[1]Institut de Recherche en Immunologie et Cancérologie (IRIC), Québec, Canada
[2]Department of Biochemistry and Molecular Medicine, Université de Montréal, Québec, Canada
[3]Department of Chemistry, Université de Montréal, Québec, Canada
[4]Princess Margaret Cancer Center, University Health Network, Toronto, Canada
[5]Medical Biophysics, University of Toronto, Toronto, Canada
[6]Vector Institute for Artificial Intelligence, Toronto, Canada
[7]Ontario Institute for Cancer Research, Toronto, Canada
[8]Department of Computer Science, University of Toronto, Toronto, Canada
[9]Department of Biostatistics, Dalla Lana School of Public Health, Toronto, Canada

Cet article fut soumis à *Scientific Report*.

Les principales contributions de Caroline Labelle pour cet article sont:

- Conceptualisation du projet;
- Développement des méthodologies proposées et présentées;
- Implémentation algorithmique;
- Curation des données expérimentales;
- Rédaction du manuscrit.

La nomenclature utilisée dans cet article est constante avec celle de Chapitres 1, 4 et 5. Les Figures 20 à 28 font partie du matériel supplémentaire de l'article original. Ces figures sont référées

dans le texte avec la notation "Supp. Fig.". À noter que la numération des figures est continue et ne recommence pas pour les figures en supplémentaire. L'ensemble du matériel supplémentaire est présenté à la Section 3.5.

Résumé

L'efficience d'un composé est communément caractérisée par des métriques dérivées de réponses cellulaires. Des décisions expérimentales sont faites sur la base de conclusions découlant de l'analyse de ces-dites métriques. Dans le présent article, nous démontrons les lacunes de la méthodologie standard résumant les réponses expérimentales en métriques d'efficience, tout en proposant une méthodologie Bayésienne robuste (BiDRA). Nous démontrons la robustesse de BiDRA quantitativement et illustrons son applicabilité avec une analyse exemple de sélection de composé.

Mots clés : Inférence Bayésienne, Dose-réponse, Métriques d'efficiences divergente, Incertitude, SAR, Sélection de composé

Abstract

Compounds' efficiency is commonly characterized by metrics derived from cellular responses. Experimental decisions are often based on conclusions arising from the analysis of said metrics. In this paper, we highlight the shortcomings of the standard methodology used to summarize responses into efficiency metrics, while proposing a more robust method based on Bayesian inference (BiDRA). We demonstrate BiDRA's robustness qualitatively and quantitatively and demonstrate its applicability with a compound's selection example.

Keywords: Bayesian inference, Dose-response, Effeciency metrics discrepancies, Uncertainty, SAR, Compound selection

3.1. Introduction

Large-scale dose-response screens are used to test the efficiency of therapeutic agents (compounds) across various conditions [22]. Compound efficiency is assessed through the analysis of various metrics, such as the IC_{50}/EC_{50} , the high- and low-dose responses (HDR and LDR) and the slope [26]. These metrics are normally estimated through non-linear regression based on experimental responses, referenced hereafter by the most used algorithm, Levenberg-Marquardt [33, 34]. This specific approach presents important limiting factors to the underlying analysis. For instance, the estimation solely considers observable data which is considered as the whole truth whereas incomplete responses can represent an important portion of the dataset [76]. The obtained efficiency metrics are thus unreliable and incorrect. Experimenters use various pre- and post-analysis data manipulation methods to overcome these shortcomings, but most are based on intuition as efficiency metrics estimates lack uncertainty measurements [77]. While these limitations are known and have been observed, there is no quantifiable large-scale demonstration of their impact on compounds efficiency assessment and analysis. The more general drug discovery process would be positively impacted by tools and processes that allow experimenters to draw informed conclusions through

uncertainty. Bayesian inference [90, 91] proposes itself as well-suited alternative to Levenberg-Marquardt. We [146] and other teams [77, 125, 129, 145, 147, 148] have proposed Bayesian models to infer efficiency metrics *posteriors*. These previous works, however, lack a global demonstration of the robustness of Bayesian inference in contrast to the standard Levenberg-Marquardt. Such demonstration would validate the proposed models and could facilitate and justify the transition to analysis processes and tools that utilize Bayesian inference.

Using three large publicly available datasets (Gray [69, 149], gCSI [72, 150] and CTRPv2 [63, 73, 74], Fig. 14.A), we quantitatively compared the abilities of Levenberg-Marquardt and Bayesian inference to accurately assess efficiency metrics for various experimental contexts. We propose a revised version of our Bayesian model BiDRA [146] that is generalizable to various experiments. Our analyses clearly demonstrate Levenberg-Marquardt limitations and their impact on the analysis process, as well as BiDRA's robustness in similar contexts. Our results also highlight the impact of the method used to summarize responses into efficiency metrics. For instance, we observed that *posteriors* enhanced metrics reproducibility when considering replicates. Such results are of great interest as there has been evidence of significant discrepancies between metrics of experimental replicates [22, 86], rendering the use of these large datasets questionable and almost obsolete. We also demonstrate the informative power of *posteriors* through a compound selection analysis from a structure-activity relationship (SAR) study.

3.2. Results

3.2.1. BiDRA model and *priors*

The use of *priors* when inferring efficiency metrics plays an important role as it allows the experimenter to mathematically incorporate his/her intuition regarding the experimental context [142, 143]. *Priors* are soft constraints on the possible values for each metric, and they thus act as a safety net when the experimental data is insufficient for the inference to accurately sample metrics values to form the *posteriors*. When we first introduced BiDRA [146], we demonstrated the importance of using appropriate, weakly informative *priors*. We revisited and redefined our *priors* to better represent the possible and plausible range of responses and concentrations (Fig. 14.B). The HDR and LDR *priors*' validity and generalization were confirmed via a visual inspection of the distributions of viability responses for the highest (HDR) and lowest (LDR) doses over the datasets (Fig. 14.C). Similarly, the wide range of experimental concentrations (Fig. 14.C) and the difficulty of constraining the range of possible values for the IC_{50} justified the use of a wide and weakly informative *prior* for this metric (Fig. 14.B). *Priors* were not mathematically fitted to the experimental data distributions.

3.2.2. Response consistency across replicates

Response consistency across biological replicates was assessed as a baseline indicator of reproducibility. When numerically comparing shared-concentration response from biological replicates

A

Dataset	Experiments	Cell-line	Drug	Replicate (R=2)	LM Converge
Gray	9,525	72	109	1,604	6,506
gCSI	16,680	569	44	2,044	12,316
CTRPv2	395,200	887	544	7,647	260,838
IRIC	70	1	63	-	-

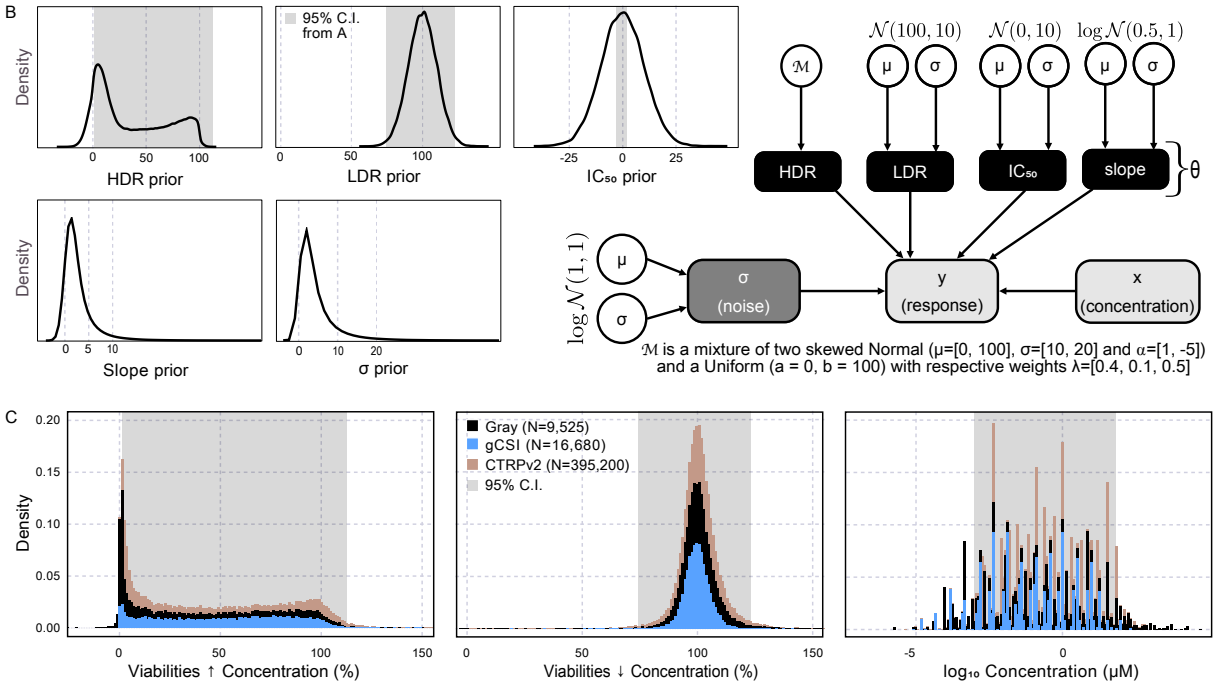
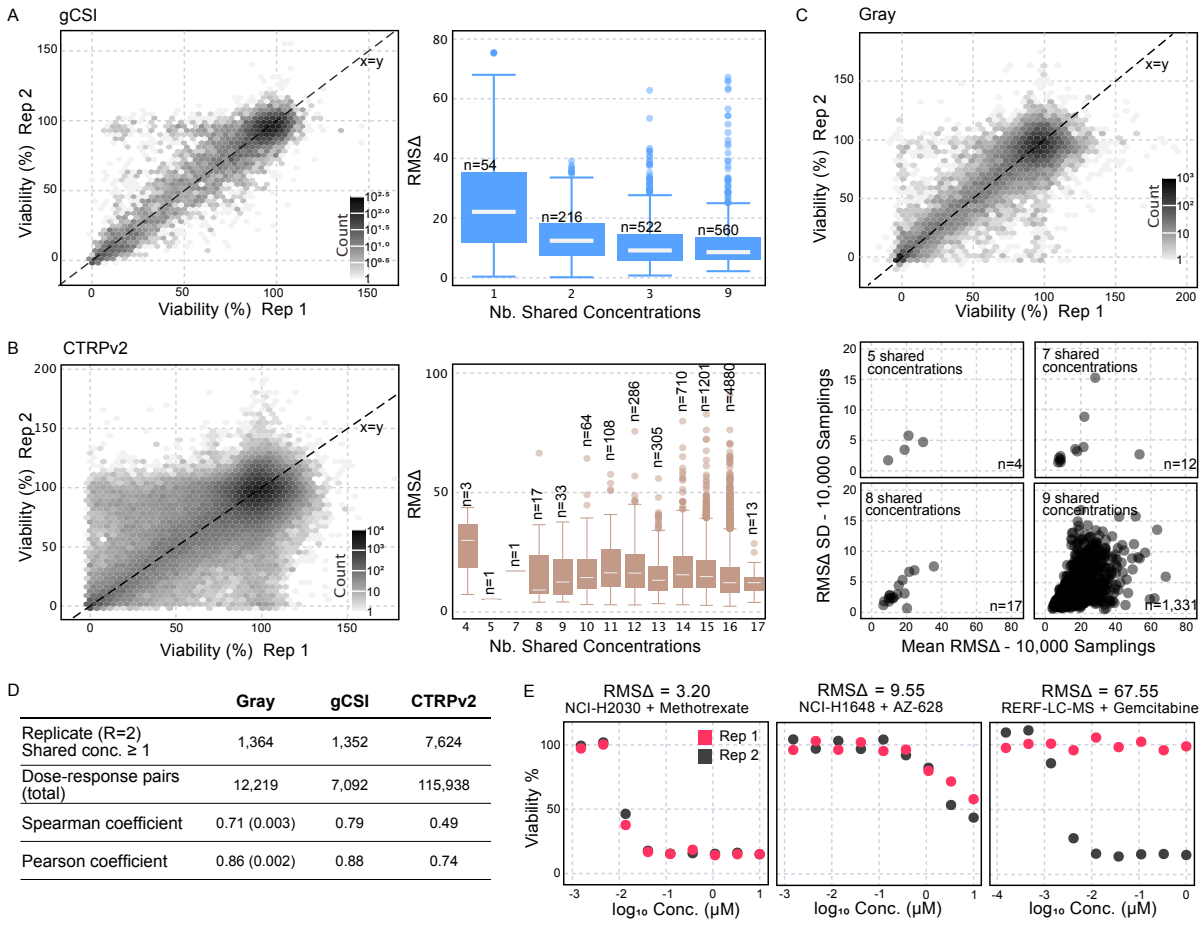


Fig. 14. Datasets and Bayesian model overviews. **(A)** Datasets description with counts of unique occurrences. **(B)** *Priors* and BiDRA Bayesian model. Shaded regions over the HDR, LDR and IC_{50} densities are defined as in **C** (in the same order). BiDRA's schematic representation: *priors* distribution parameters are represented by circles; efficiency met-rics by black boxes; the likelihood variance by a gray box; and experimental data by light gray boxes. **(C)** Distributions of experimental responses for the highest (↑) and lowest (↓) concentrations, and experimental concentrations across the Gray, gCSI and CTRPv2 datasets (421,405 experiments considered). Shaded regions represent 95% confidence intervals (also represented in **B**).

(Fig. 15.A-C), we found that all three datasets had consistent response, with their respective Pearson correlation coefficients being above or close to 0.75 (Gray: 0.86, gCSI: 0.88 and CTRPv2: 0.74) (Fig. 15.D). Furthermore, most individual pairs of replicates have a RMSΔ that suggest good replications of response values (RMSΔ < 20, Fig. 15.E). Only 0.5% of all pairs considered share a single concentration (54 pairs in gCSI, Fig. 15.A) making it difficult to evaluate consistency. Pairs with a large number of shared concentrations have globally a lower RMSΔ, suggesting that their responses are consistent (Fig. 15.A-C). These results make it reasonable to expect efficiency metrics of biological replicates to be consistent, considering they are derived from consistent viability responses.



	Gray	gCSI	CTRPv2
Replicate (R=2) Shared conc. ≥ 1	1,364	1,352	7,624
Dose-response pairs (total)	12,219	7,092	115,938
Spearman coefficient	0.71 (0.003)	0.79	0.49
Pearson coefficient	0.86 (0.002)	0.88	0.74

Fig. 15. Response consistency across biological replicates. Hexbins plots in **A**, **B** and **C** illustrate the comparison of shared-concentration responses across biological replicates. Boxplots in **A** and **B** illustrate the range of RMSΔ for diverse number of shared concentrations. Numbers of pairs of biological replicates considered for each number of shared concentrations are denoted by n. **(A)** Response consistency assessment for gCSI. **(B)** Response consistency assessment for CTRPv2. **(C)** Response consistency assessment forGray. Hexbins plot is for a single random pairing of within-concentration replicates. RMSΔ results are presented as mean and standard deviation (SD) values for 10,000 random samplings and pairings of within-concentration replicates. **(D)** Analysis metrics summary by dataset. For Gray, correlation coefficient values are reposted as mean and (standard deviation) from 10,000 random samplings and pairings of within-concentration replicates. **(E)** Examples of biological replicates from gCSI with highly consistent (RMSΔ = 3.20), consistent (RMSΔ = 9.55) and discrepant (RMSΔ = 67.55) responses.

3.2.3. Efficiency metrics conscistency across replicates

Consistency of efficiency metrics across biological replicates was then assessed for three metrics representations (Levenberg-Marquardt estimates, BiDRA *posteriors*' median values and BiDRA quantile-to-quantile (QQ) *posteriors*) and three groupings of pairs based on responses completeness (All, Incomplete and Complete Pairs) (Fig. 16 and Supp. Fig. 20.A). We used standard deviation (SD) of response to group experiments: small SD (< 20) are characteristic of response leading to seemingly unresponsive (e.g., PF-4708671 in Fig. 17.A) or incomplete curves (e.g., AZ-628 and

Lapatinib in Fig. 17.A), whereas larger SD ($\geq$ 20) are generally indicative of complete curves (Gemcitabine in Fig. 17.A).

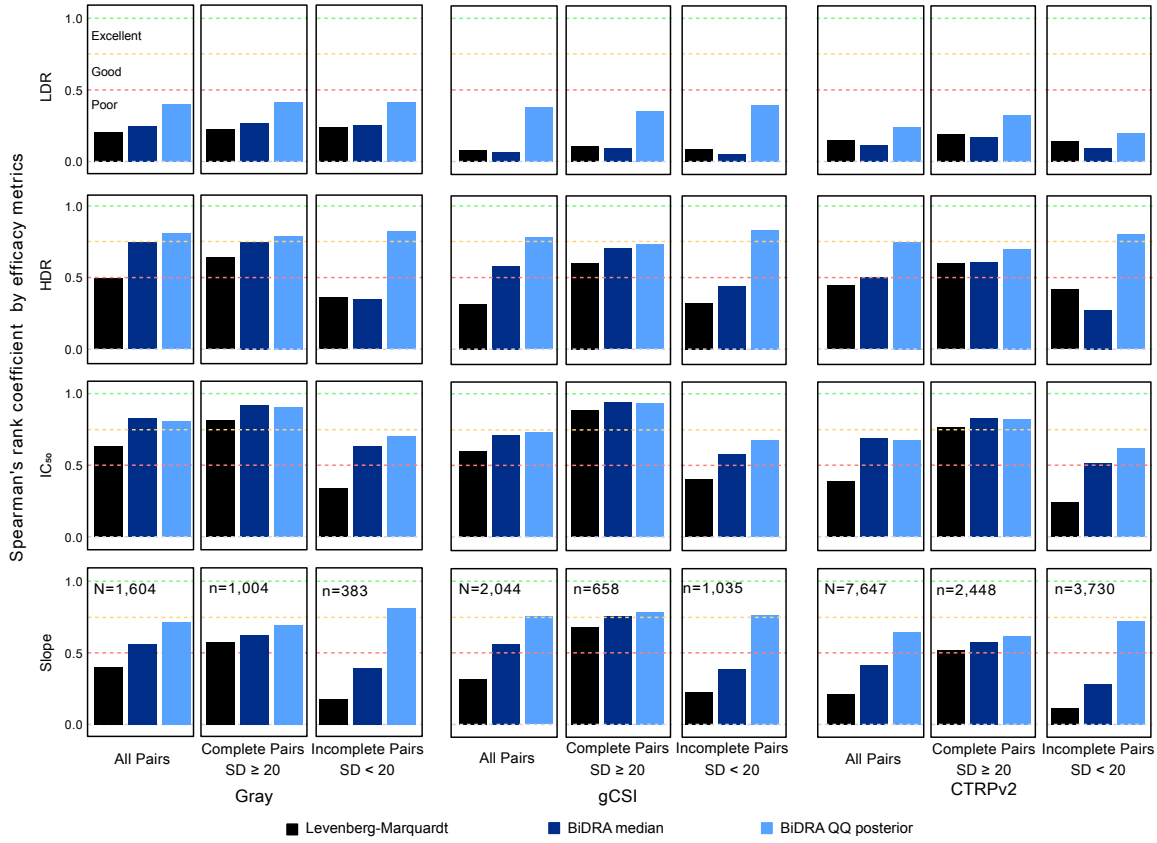


Fig. 16. Efficiency metrics consistency across biological replicat. Spearman's rank correlation coefficient across efficiency metrics (rows), efficiency metrics represen-tations (colors), datasets (column triplets), and pairs groupings (columns). The latter are defined by response completeness (i.e., response standard deviation, SD). Complete and Incomplete Pairs con-sist of two replicates with the same completeness status. Numbers of pairs of biological replicates considered for each grouping and dataset are denoted by N (All) and n (subsets).

Interestingly, Pearson and Spearman's rank coefficients are consistent for both of BiDRA representations, while Spearman rank coefficient seems to be a better indicator of consistency for Levenberg-Marquardt estimates as it is insensitive to outliers (Supp. Fig. 20.B). Pearson is more robust to the high frequency of unresponsive cell-lines or incomplete dose-response [151], explaining the lower correlations for Levenberg-Marquardt estimates. We thus focused our analysis on Spearman's rank coefficients as it underestimates the correlation and still demonstrates BiDRA's superiority. Results for Pearson correlations are shown in the Supplementary. The conclusions regarding BiDRA's robustness and superiority are held regardless of the correlation coefficient selected.

As expected, we found that BiDRA's representations better correlate than Levenberg-Marquardt estimates in all configurations tested. BiDRA *posteriors* have good ($\rho \geq 0.5$) and excellent ($\rho \geq 0.75$) correlations for HDR, IC_{50} and slope across all three datasets (except for

CTRPv2 slope), while Levenberg-Marquardt estimates have mainly poor correlations, except for Gray's and gCSI's IC_{50} which have good correlations (All pairs in Fig. 16). The LDR has the poorest correlation, regardless of the metric representation. It represents the basal asymptotic response which is experimentally measured through negative controls. It has, consequently, little variability, explaining the observed poor correlation coefficients (Supp. Fig. 21).

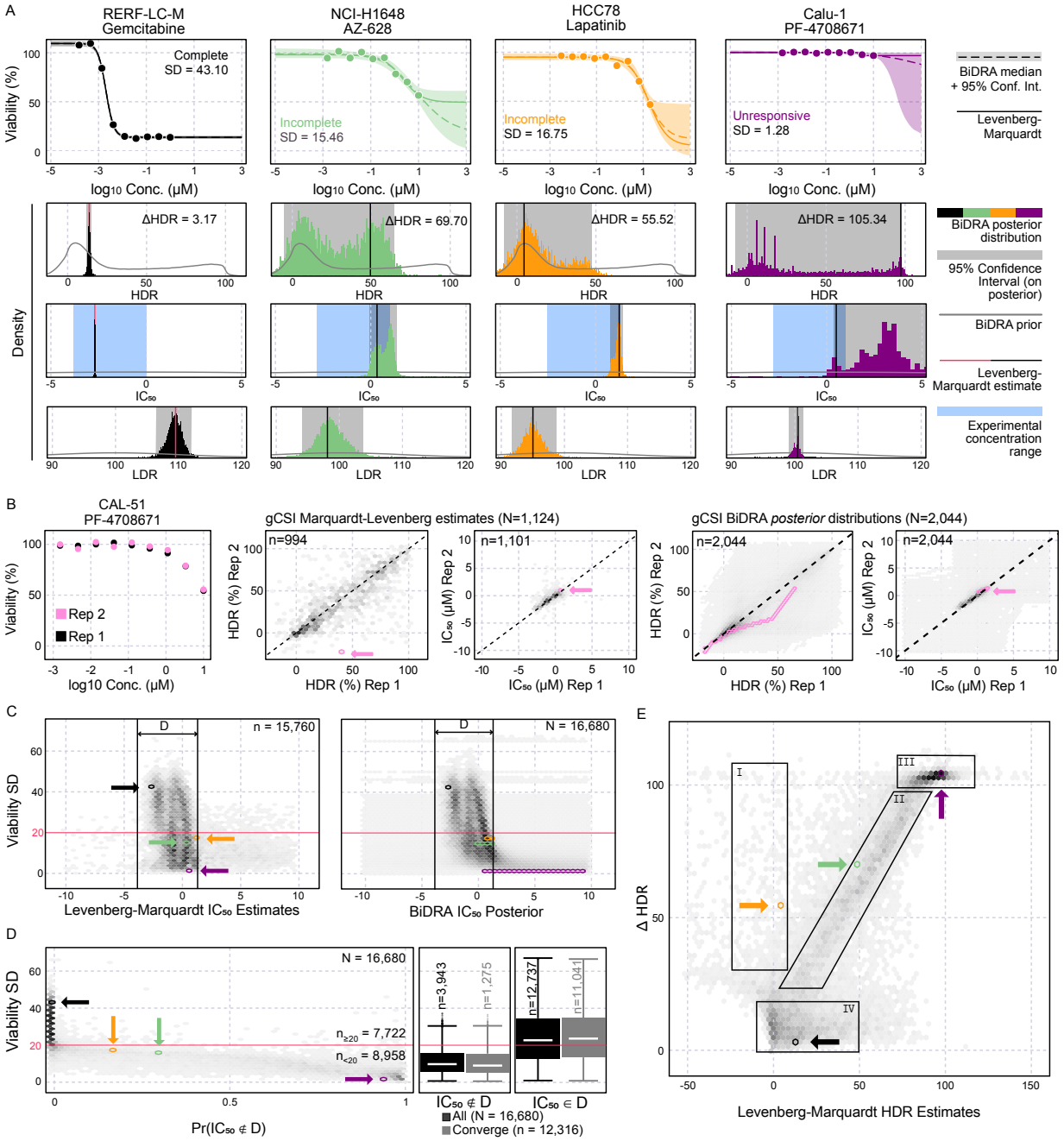


Fig. 17. BiDRA's robustness across diverse types of responses. *(see next page)*

Fig. 17 ***(previous page).*** **(A)** Experiments with different response completeness status (Complete, Incomplete and Unrespon-sive), as defined by their response standard deviation (SD) values. Estimated (Levenberg-Marquardt) and inferred (BiDRA) dose-response curves are also represented. HDR, IC_{50} and LDR estimated values (Levenberg-Marquardt) and *posteriors* (BiDRA) are shown in individual plots. These experiments are referred to in **C**, **D** and **E** by their respective color. **(B)** PF-4708671 biological replicate with incomplete and concordant responses. HDR and IC_{50} estimates and *posteriors* from both experiments (pink) are compared against all of gCSI pairs of biological replicates. Dashed lines represent the identity diagonal. **(C)** IC_{50} estimates (Levenberg-Marquardt) and *posteriors* (BiDRA) compared to response completeness (SD). The bounds of the 95% interval for experimental concentrations are represented by the black vertical lines and denoted by D. Numbers of experiments represented by the plots are denoted by N (All) and n (subset). **(D)** (Left) Comparison between response completeness (SD) and the probability of an IC_{50} of being outside the experimental concentration range. (Right) Comparison between response completeness (SD) and IC_{50} estimates observability, based on Levenberg-Marquardt convergence status. Numbers of experiments considered for the comparisons are denoted by N (All) and n (subsets) **(E)** Comparison of HDR estimates (Levenberg-Marquardt) to *posterior* wideness i.e., uncertainty (ΔHDR, BiDRA). Results presented in **C**, **D** and **E** are representative of the gCSI dataset. Results for the Gray and CTRPv2 datasets are reported in Supplementary Figure 26

Whole *posteriors* are more informative than single-point values, as demonstrated by the correlations from various groupings of replicates (Fig. 16 and Supp. Fig. 20.A). Complete Pairs of biological replicates (both experiments have a SD $\geq$ 20) are overall consistent ($\rho \geq 0.5$) for HDR, IC_{50} and Slope. All three representations have similar correlation coefficients, but BiDRA *posteriors* are still the most consistent across all three datasets. Levenberg-Marquardt estimates lead to extremely poor correlations when considering Incomplete Pairs of biological replicates (both experiments have a SD $<$ 20), regardless of dataset and efficiency metrics. BiDRA *posteriors* are, however, still consistent when considering such pairs. Correlations for All Pairs are representative of the Complete to Incomplete pairs ratio for each dataset. We also assessed the consistency of areas above the curve (AAC) values, which are computed from the basic efficiency metrics (LDR, HDR, IC_{50} and slope) and for a shared range of experimental concentrations between two replicates (Supp. Fig. 22). The observed results are consistent with the aforementioned efficiency metrics results. BiDRA's AAC *posteriors* had either good ($\rho \geq 0.5$) or excellent ($rho \geq 0.75$) correlations across all three datasets and pairs groupings (Supp. Fig. 23). We also highlighted the flexibility of *posteriors* compared to single-point estimate values when calculating derived efficiency metrics, such as the AAC (Supp. Fig. 24).

3.2.4. Control experiments

The validity of our comparative analysis was confirmed through a negative control experiment of randomly generated pairs of experiments (10 repetitions, Fig. 18 and Supp. Fig. 25). As expected, Levenberg-Marquardt estimates of random experiment pairings do not correlate, regardless of pairs grouping.

To our initial surprise, we observed grouping-specific correlations between *posteriors* of randomly paired experiments. For instance, random pairs of complete experiments (both experiments have a SD $\geq$ 20) show poor ($\rho < 0.5$) or no correlation ($\rho \approx 0$) between their HDR, IC_{50} and slope. When considering incomplete random pairs, the correlations between HDR and between

slope become good ($\rho \geq 0.5$). The correlation between IC_{50} is still poor ($\rho < 0.5$) but has increased compared to that of complete random pairs. These results are explained by the soft constraints introduced on the effciency metrics through the *priors*. Experiments with complete response and well-defined dose-response curve are less likely to fall back on *priors*, as their experimental data is sufficient for the inference to sample the *posterior* distributions. Thus, the resulting *posteriors* are more precise and experiment-specific, and differentiable from *posteriors* of other experimental contexts (i.e., different cell-line and/or compound). Experiments with incomplete response, on the other hand, tend to rely more on the *priors* as their experimental data is insufficient to accurately infer the metrics from the desired log-logistic model. Since *priors* are shared across all experiments and datasets (Fig. 14), it is expected for *posteriors* to heavily incorporate them to correlate, regardless of experimental contexts. Correlations for all randomly generated pairs are thus representative of the ratio of incomplete to complete pairs of experiments for each dataset.

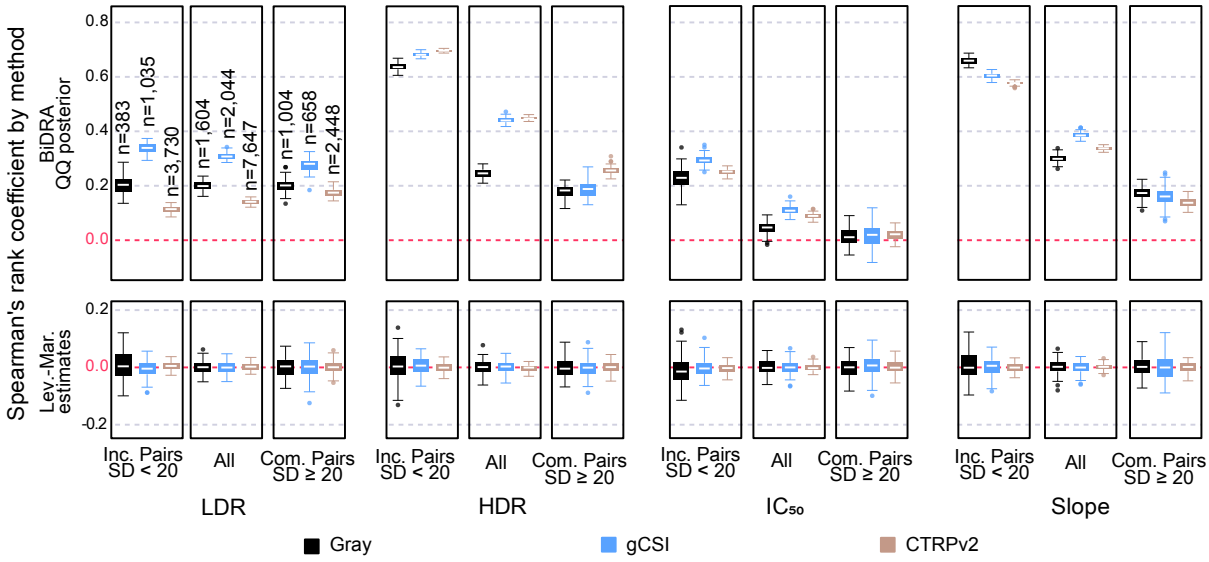


Fig. 18. Control experiment: consistency assessment of efficiency metrics across randomly paired experi-ments. Spearman's rank correlation coefficient across random pairings of experiments for 10 repetitions for each efficiency metrics (column triplets), efficiency metrics representations (rows), datasets (colors), and pairs groupings (columns). The latter are defined by response completeness (i.e., response standard deviation, SD). Complete and Incomplete Pairs consist of two experiments with the same completeness status. Numbers of random pairs considered for each grouping and dataset are denoted by n.

3.2.5. BiDRA's robustness

The results presented in Figure 16 suggest that the main differences between the two methodologies arise from their handling of experiments with incomplete or unresponsive dose-response curves (Fig. 17.A). The PF-4708671 pair (Fig. 17.B) demonstrates BiDRA's robustness when assessing incomplete dose-response curves. Levenberg-Marquardt IC_{50} estimates are somewhat discordant (0.69μM for Rep1 and. 1.21μM for Rep2) while BiDRA's QQ correlation suggest concordant IC_{50} . Levenberg-Marquardt HDR estimates are arbitrary (the experimental responses of either replicate

do not support the estimated values) and discordant (41.65% for Rep1 and –18.39% for Rep2), suggesting the two experiments were not well replicated even though the viabilities are nearly identical. Alternatively, BiDRA's wide HDR *posteriors* mostly align on the diagonal, confirming that both experiments are concordant, while having highly uncertain HDRs.

Algorithmically, Levenberg-Marquardt forces a fit onto the observed responses which results, in some cases, in unreliable estimated efficiency metrics. For instance, we observed that Levenberg-Marquardt mostly estimates observable IC_{50} , meaning that the estimated concentration is within the experimental concentration range (denoted by D in Fig. 17.C and Fig. 1713). We think this conclusion is an overestimation of the true amount of observable IC_{50} as demonstrated by the estimated IC_{50} for Calu-1 + PF-4708671 (purple in Fig. 17.C). Indeed, in gCSI, for instance, 12,737 IC_{50} estimates (76.36% of experiments) were within the experimental concentration range. For the same dataset, we only identified 6,501 complete responses (SD $\geq$ 20, 38.97%). Even though an incomplete response can be informative regarding the IC_{50} location, the estimated value would most likely be inaccurate. We also observed similar distributions when comparing viability SD to the number of observed and unobserved IC_{50} , regardless of the algorithm convergence status (Fig. 17.D and Supp. Fig. 26). Alternatively, we found that incomplete responses had a higher probability of having their IC_{50} *posterior* outside of the experimental concentration range ($Pr(IC_{50} \notin D)$ in Fig. 17.D and Supp. Fig. 26). BiDRA handles the uncertainty of an efficiency metric by returning a wider distribution, as demonstrated by the comparisons of ΔHDR to Levenberg-Marquardt estimated HDR (Fig. 17.E and Supp. Fig. 26) and to response completeness (SD values, Supp. Fig. 27). We identified many experiments with highly uncertain *posterior* and thus most likely unsupported Levenberg-Marquardt estimates close to the optimal 0% (orange and I in Fig. 17.E). Interestingly, we notice a trend for which experiments with higher HDR estimates correlate with uncertain *posterior* (green and II in Fig. 17.E). This group most likely corresponds to experiments with incomplete curves. We are also able to clearly identify the problematic seemingly unresponsive curves for which Levenberg-Marquardt forces an HDR where there should not be one (purple and III in Fig. 17.E). It is worth noting that for complete curves, both methodologies lead to similar and certain metric representations (black and IV in Fig. 17.E).

3.2.6. Post-inference analysis: compounds selection

Given the IRIC dataset (Fig. 14.A), we are interested in finding out if there is any analog compound that is better than our reference compound. We define better as having a higher efficacy (large HDR) and being more potent than the reference compound (smaller EC_{50}). We applied BiDRA to the 70 dose-response experiments and collected *posteriors* for each of the standard efficiency metrics: LDR, HDR, EC_{50} and Slope. We will be using these metrics representations to complete our compounds selection.

From the EC_{50} rank probabilities we identified 16 experiments with a probability of at least 0.90 of being amongst the 20 first ranks (purple in Fig. 19.A). From these, 10 are Analogs and 6 are replicates of the Reference. Overall, experiments seem to be differentiable in terms of EC_{50}. Interestingly, we noticed subgroups of compounds (e.g., Analogs 58 and 46) that have similar

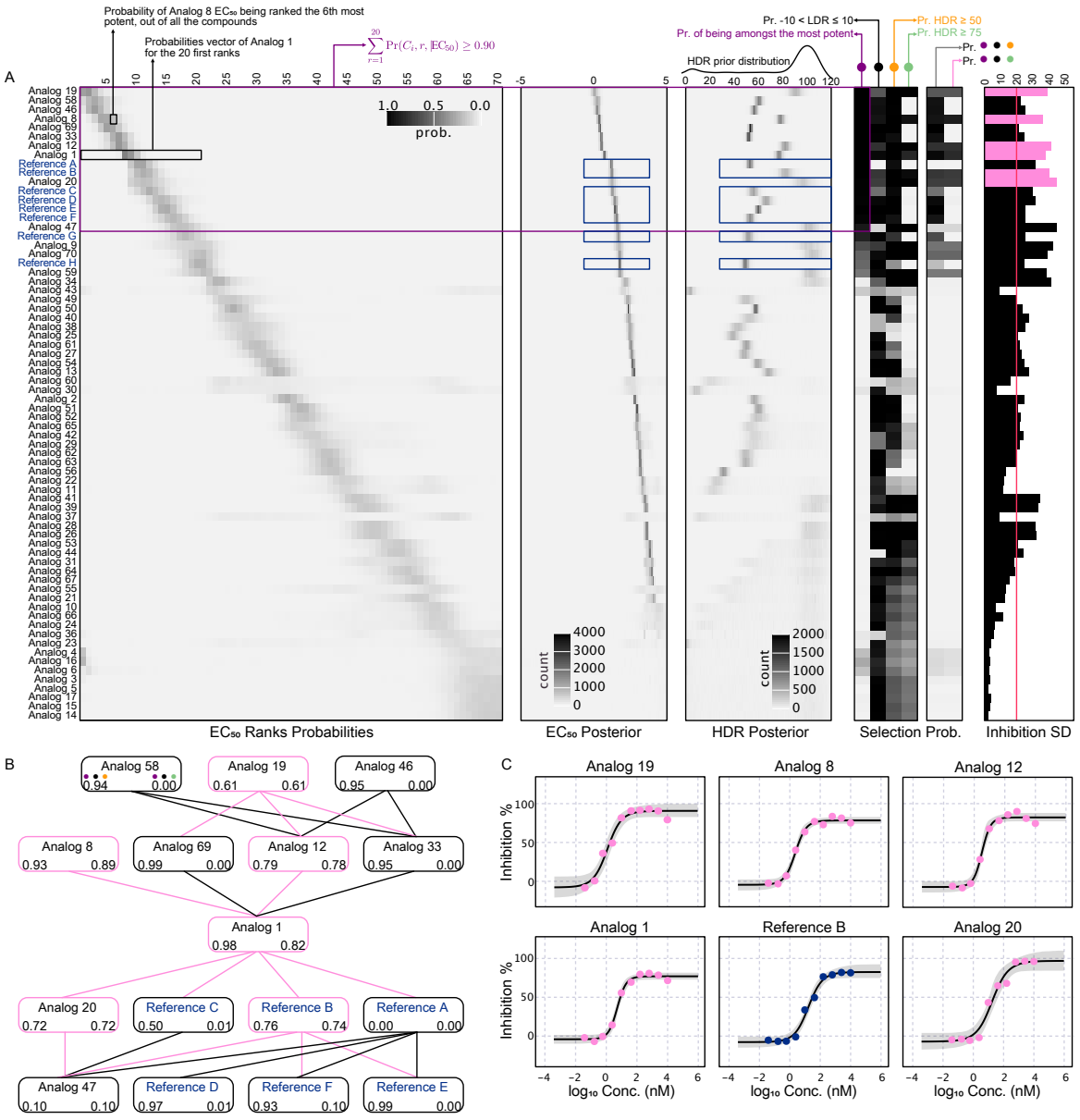


Fig. 19. Analysis of SAR: compound selection using *posteriors*. **(A)** EC_{50} ranking and selection criteria probabilities. The heatmaps from left to right illustrate: (1) EC_{50} ranks in descending order of potency with experiments sorted based on the ascending order of their EC_{50} *posterior* median values; Experiments that are most likely to be amongst the 20 most potent are highlighted by the purple rectangle. (2) EC_{50} *posteriors* sorted as in (1). (3) HDR posteri-or sorted as in (1). (4) Selection criteria probabilities based on *posteriors*; High probabilities are illustrated in black, while smaller probabilities are closer to white. (5) Global selection probabilities for two combinations of (4). Response completeness (SD) of each experiment is illustrated by the bar plot. Experiments selected in **C** are highlighted in pink. **(B)** EC_{50} DAG. Only the experiments from the most potent subset are considered (purple in **A**). Global selection probabilities are identified for each experiment. **(C)** Median dose-response curves of experiments selected from **A**. 95% confidence intervals on the curves are represented by shaded regions.

ranking probabilities, suggesting they have similar *posteriors* and that it would be incorrect to rank these experiments in any specific order amongst themselves. We were also pleased to observe

that replicates for the Reference had EC_{50} *posterior* that ranked similarly (blue in Fig. 19). The slight shifts in rank probabilities are most likely due to the "precision" of the EC_{50} *posterior* and is a direct demonstration of how well *posterior* accounts for biological and analytical variabilities. EC_{50} and HDR of Reference replicates mostly overlap (except Reference B which has a higher HDR) (Fig. 19.A). It is worth noting that the globally low and uncertain HDR observed are characteristic of the cell-line used and was expected. Further experiments were done to enhance HDR while keeping the EC_{50} potent but are not shown in this paper.

We used a directed acyclic graph (DAG) to summarized statistically significant EC_{50} comparisons ($\alpha = 0.05$) form our 16-experiment subset (Fig. 19.B). If there exist a path between two nodes of DAG, then the EC_{50} *posteriors* of the corresponding two experiments are statistically different ($\alpha = 0.05$). For instance, Analog 58 is statistically more potent than Analog 12.

We obtained probability values by experiment for a set of selection criteria (Fig. 19.A): (•) Experiment must be amongst the most potent i.e. its EC_{50} must have a probabilities of at least 0.90 of being amongst the 20 first ranks; (•) Experiment must have a basal response (LDR) between -10 and 10, to ensure quality control over the normalization of response; (•) Experiment must have a high efficacy i.e. its HDR must be at least 50%; (•) Experiment must have a high efficacy i.e. its HDR must be at least 75%. We obtained global selection probabilities by combining the probabilities of meet-ing each aforementioned criteria for individual experiment (Fig. 19.A). The two global probabilities differ in terms of the high efficacy criteria (HDR). As we are interested in potent compounds with high efficacy, we used the second global selection probability (•••). We identified 6 potent compounds with a high efficacy (Fig. 19.C) with complete dose-response curves (SD $\geq$ 20) (pink in Fig. 19).

From these results, we are thus capable of identifying Analogs 1, 8, 12 and 19 as better than the Reference.

3.2.7. Other definitions of replicates

We defined replicates as having at most two experiments within a single dataset. We thus exclude cell-line + compound pairs that were replicated more than twice, and those that were replicated across datasets. BiDRA's superiority compared to Levenberg-Marquardt was also confirmed for two other definitions of replicates.

We first considered multi-replicate or cell-line + compound pairs that were tested at least three times within a dataset (Supp. Fig. 28.A). We calculated the proportion of R = 2 replicates and R $\geq$ 3 replicates. We found that gCSI (15.52% vs. 1.97%) and CTRPv2 (1.98% vs. 0.11%) both have more R = 2 replicates (Supp. Fig. 28.B). The results obtained with the initial definition (R = 2) are representa-tive of the general replicates tendency of those datasets. Gray on the other, has minor difference between the number of cell-line + compound pairs for both types of replicates (31.94% vs. 22.60%) (Supp. Fig. 28.B). BiDRA's *posteriors* have overall higher correlation coefficient than Marquardt-Levenberg estimates when considering multi-replicates (R $\geq$ 3) (Supp. Fig. 28.C).

We then considered experimental replicates or across dataset replicates. Most cell-line + compound pairs of Gray, gCSI and CTRPv2 were singletons, meaning that they were tested once

within a dataset. Singletons represent 45.46%, 80.46% and 97.91% of each dataset, respectively (Supp. Fig. 28.B). We considered singletons that were replicated across pairs of datasets. It is worth noting that Gray, gCSI and CTRPv2 are not the most appropriate datasets for such analysis as most dataset's pairs share a small number of singletons. Only gCSI vs. CTRPv2 should be truly considered in this context (Supp. Fig. 28.B). While observed correlations are poorer than that of biological replicates, our preliminary assessment of efficiency metrics replicability between experimental replicates further supports BiDRA's superiority (Supp. Fig. 28.D).

3.3. Discussion

The drug discovery process is a costly multi-phase process that spans multiple years and requires many resources [12, 152]. Every decision made, such as identifying and selecting compounds with desired efficiency characteristics, bare a relevant and significant impact on the complete process. Experimenters thus need to be equipped with the most appropriate tools28 to make such decisions in an informed and unbiased fashion. While the frequentist non-linear regression (Levenberg-Marquardt algorithm) is widely used and implemented by various tools [45, 58, 59, 138], we argue that Bayesian inference, namely our BiDRA model, is much more robust and suited to the inference of efficiency metrics.

Levenberg-Marquardt main disadvantages, namely its dependance on the number of experimental concentrations and datapoints [76, 153], and the lack of direct measure to describe estimates uncertainty [77], are commonly known. We now have quantitively demonstrated the effect of Levenberg-Marquardt's shortcomings on the estimated efficiency metrics for three large datasets. The algorithm is limited to observable responses, making the estimated values reliability and uncertainty depend on the accuracy of the experimental responses and dose range. For complete well-defined sigmoidal responses (SD $\geq$ 20), the algorithm can accurately estimate efficiency metrics. When responses are either incomplete or flat (unresponsive cells, SD $<$ 20), some or all efficiency metrics estimates are extreme, unreliable, and unsupported by the experimental context. This behavior is well demonstrated by our assessment of IC_{50} observability: most IC_{50} values are estimated to be within the experimental dose range (Fig. 17.C and Supp. Fig. 26). Due to the lack of statistical measurement of estimates uncertainty [77, 154], experimenters must rely on goodness of fit metric (such as the squared-root of the sum of residual, RMSE [146]), convergence status or qualitative inspection of the fitted dose-response curve. Unfortunately, goodness of fit is not indicative of context-specific estimates reliability [77, 146] as is convergence status (Fig. 17.D and Supp. Fig. 26). Visual inspection [154] is also a poor alternative as it is difficult to reproduce (experimenter-specific) and can be tedious to carry out due to the number of experiments to consider. Efficiency metrics estimates plausibility can also be enhanced through data manipulation such as dataset filtering [155] (e.g., filtering out flat response experiment), experimental responses caping (e.g., responses cannot exceed 100%) and estimated efficiency metrics values caping [60, 75] (e.g., caping IC_{50} values to the largest experimental dose [76]). This sort of manipulation increases analytical variability which, added to the inevitable biological variability, can biased conclusions and altered the decision-making process. It also eliminates potential useful information regarding

compound and cell-line sensitivity [76]. Levenberg-Marquardt shortcomings can be particularly damaging when doing exploratory screens as incomplete and flat responses can represent an important portion of the dataset [76]. The task of selecting compounds through uncertainty becomes increasingly difficult.

Bayesian inference presents itself as a flexible and robust alternative to Levenberg-Marquardt [77, 129, 145, 146]. Flexible because it can introduce multiple soft constraints on efficiency metrics through the *priors*, and it allows for various post-inference analysis that leads to statistically sound conclusions; robust because it can handle incomplete and flat responses, and it can infer efficiency metrics through uncertainty arising from biological and analytical variabilities. We proposed here an improved version of our BiDRA model [146] (Fig. 14.B). We defined weakly informative *priors* for all parameters (Fig. 14.B) that incorporate both experimenter intuition and soft constraints on the experimental context. The validity of our HDR, LDR and IC_{50} *priors* were confirmed through qualitative evaluation of minimal and maximal responses, and experimental dose values across three large datasets (421,405 experiments, Fig. 14.C). We decided on using a normal sampling distribution (likelihood) as it best represent our knowledge of the data [132] and the data itself (the amount of response is often insufficient to accurately characterize the distribution of the data variance). The proposed model is generalized and well suited for a wide range of dose-response experiments (e.g., Gray, gCSI, CTRPv2 and IRIC datasets, Fig. 14.A). Previously proposed models often present *priors* that are specific to a given dataset and/or experimental contexts [77, 145]. Furthermore, we also considered the four-parameter log-logistic model compared to lesser parameterized versions of the model [129, 145]. While the latter can be appropriate in some experimental context [145], we found that it can be too constraining and have similar effect to an overly precise *prior* [146]. We thus recommend inferring all efficiency metrics as they can all be informative in to regards to compound sensitivity [23, 76]. For instance, LDR *posterior* can be indicative of an improper response normalization or of hormesis [30], and be used for quality control assessment (Fig. 19.A). We also demonstrated how the *posterior* of basic efficiency metrics can be used to derive other efficiency metrics such as the area above the curve [22, 60] (AAC, Supp. Fig. 22-24) or DSS [42].

An important and novel contribution of our present work is the quantitative demonstration of Bayesian inference robustness compared to Levenberg-Marquardt. While previous work, including our own, were mainly based on qualitative analysis and comparison (e.g., simulation studies, applicability demonstration on real life experimental data)[77, 129, 145, 146], we aimed at establishing a large-scale quantitative benchmark demonstrating Bayesian inference robustness. Such demonstration validates and justifies the need for past and future proposed Bayesian models for the analysis ofdose-response data. When comparing efficiency metrics (HDR, IC_{50} , slope and AAC) across biological replicates, we found that BiDRA's *posteriors* had either Good (≥ 0.5) or Excellent (≥ 0.75) correlations, compared to Levenberg-Marquardt estimates which had overall Poor (< 0.5) correlations. We decided to use Spearman's rank coefficient as it demonstrated BiDRA superiority while favoring Levenberg-Marquardt estimates correlation. Pearson correlation coefficient makes a better account of noise resulting from high frequency of unresponsive cell-lines or incomplete response [151], which are poorly handled by Levenberg-Marquardt. The fact that BiDRA's superiority is independent of correlation coefficient (Pearson and Spearman's rank coefficients were consistent

across *posterior* correlations, Supp. Fig. 20.B) is further demonstration that Bayesian inference is more robust than Levenberg-Marquardt, especially when considering incomplete responses. We still observed Good and Excellent correlations between *posteriors* when considering pairs of biological replicates with incomplete response (SD < 20), while Levenberg-Marquardt estimates plummeted to even poorer correlations (Fig. 16 and Supp. Fig. 23.B). While these high correlations are in part explained by the incorporation of soft constraints through *priors*, we emphasize that *posteriors* better account for experimental data and are informative representation of efficiency metrics. We observed that correlations between incomplete (SD < 20) true biological replicates (Fig. 16) were always higher than those of incomplete (SD < 20) random pairs of experiments (Fig. 18). The differences in HDR correlation are small since the experimental HDR plateau is often missing from incomplete responses. The differences in IC_{50} correlation are, however, bigger: ≥ 0.5 for true replicates and< 0.4 for random pairs. This result highlights the experiment-specific information that can be retrieved, even from incomplete response, through Bayesian inference. Experiments with incomplete response are still relevant a are informative [76]. Furthermore, the uncertainty of each efficiency metric can be easily and accurately assessed through *posteriors* (e.g., confidence interval and median values). Our comparative assessment of discrepancies between efficiency metrics of biological replicates quantitatively demonstrates Bayesian inference, and specifically our BiDRA model, superiority and robustness.

Another novel element of our current work is our demonstration of a complete post-inference analysis for compounds selection from a SAR screen (Fig. 19). While Bayesian inference proposes a robust alternative to Levenberg-Marquardt, the interpretation and usage of *posteriors* can be daunting and non-trivial. When developing a novel methodology, it is important to emphasize its accessibility and utility [131]. Using the IRIC dataset, we highlighted several post-inference methodologies such as efficiency metric ranking, experiment comparisons and evaluation of selection criteria. The obtained results accounted for both experimental and analytical variabilities, considered the entire dataset (regardless of response completeness), and are statistically sound. Such results are unreachable while using Levenberg-Marquardt's estimates.

Finally, our results suggest that *posterior* could be beneficial in other analysis context. For instance, our preliminary results assessing discrepancies between efficiency metrics of experimental replicates suggest that *posterior* better correlates than single-point estimates. Even though the correlation is smaller [22, 86] than that observed for biological replicates, this leads us to believe that the usage of *posterior* distributions would be beneficial when considering and analyzing dose-response experiments from multiple sources. The subject of discrepancies (across and within large scale pharmacogenomics) could be thoroughly revised while considering *posteriors* as an efficiency metric representation. Many papers that address efficiency metrics discrepancies concentrate on the difficulty of accurately and experimentally replicating experiments (across laboratories and research centers) and the underlying challenge of combining dose-response datasets [19]. While biological variability undoubtedly affects consistency of efficiency metrics, our present results suggest that efficacy metrics representation is also an important driver of discrepancies. *Posteriors* could also be used in machine learning. There is a great interest in developing models that can predict drug sensitivity based on compound representations [156–159]. Such approaches have the potential,

among other things, to facilitate compound repositioning [157] and thus alleviate the financial cost of drug discovery [160]. Scientists face many challenges when developing a model to predict compound sensitivity, the main ones being: (1) datasets often contain too little samples and (2) sensitivity metrics used to train models are discrepant [159]. While the former can lead to overfitting [161], the latter makes it hard to accurately train a model (its accuracy is bound by the noise from the data); to produce a model that is generalizable to multiples cell-lines and/or datasets; and to cross-validate the model performance on another dataset than the one used for training [162]. Sampling from *posteriors* during training could likely increase the limits on prediction accuracies reported by Xia *et al.* [159]. The lower rate of discrepancies observed while using *posteriors* could also facilitate the pooling of multiple datasets and the cross-validation of models on different datasets. Finally, using *posteriors* while training a model could also minimize the risk of overfitting by acting as a source of data augmentation, especially when working with a small and limiting number of samples.

The use of Bayesian inference, through BiDRA, highlights the increased robustness when compared to the conventional Levenberg-Marquardt, particularly in the context of incomplete or unresponsive dose-responses. Our proposed methodology does not require the experimenter to manually curate datasets based on dose-response curve completeness (which is tedious and subjective), nor does it require to manipulate, and curate estimated metrics to exclude inconclusive fittings. While there are considerable efforts made to standardize response measurement protocols [13, 19, 163], true biological variability will always be present43 and incomplete or unresponsive curves are to be expected frequently. BiDRA makes it possible to robustly account for such situations when analyzing and interpreting efficiency metrics.

3.4. Methods

3.4.1. Data

For our robustness demonstration, we considered three large publicly available datasets (Gray [69, 149], gCSI [72, 150] and CTRPv2 [63, 73, 74]), and for our *posteriors* application demonstration, we used a smaller in-house dose-response screen (IRIC) (Fig. 14.A). The latter can be made available upon request with compound identification anonymized.

Gray, gCSI and CTRPv2 have different coverage of cancer cell-lines and compounds and contain replicated experiments for various pairings of cell-line and compound (Fig. 14.A). We mainly considered biological replicates of two experiments (R = 2). In the supplementary materials, we discuss results obtained for the analysis of biological replicates with more than two experiments (R $\geq$ 3, Supp. Fig. 28). We are also limiting our main analysis to biological replicates, i.e., within-dataset replicates. A preliminary analysis of experimental or across-dataset pairs of replicates is presented in the sup-plementary materials (Supp. Fig. 28).

The Gray, gCSI and CTRPv2 datasets were downloaded from PharmacoDB [75]/ORCESTRA [57]53 and were handled using PharmacoGx [55]. The retrieved normalized responses represent cell

viability (%, descending dose-response curve) and the retrieved concentration values were $\log_{10}$-transformed (μM). We excluded experiments with at least one response greater than 200% and/or less than –50%. In total, we considered 9,525, 16,680 and 395,200 individual dose-response experiments (i.e., unique dose-response curves) from each dataset, respectively (Fig. 14.A).

The IRIC screen is a subset of a larger structure-activity relationship (SAR) study. The aim of this screen was to identify analogous compounds that have a high efficacy and are more potent than a reference compound. The subset used in this paper contains 70 dose-response experiments on a single leukemic cell-line. In total, 62 analog and 1 reference compounds were tested (Fig. 14.A). A luminescence assay was used to quantify cell-response. We normalized luminescence values by the mean luminescence values of negative control (DMSO + cell-line) based on plate-location (shared-row). Values were normalized to represent responses of cell-growth inhibition rate (%, ascending dose-response curve).

3.4.2. Compounds Efficiency Metrics

3.4.2.1. Dose-response model

We use the four-parameter log-logistic model [44, 153] to describe the experimental dose-response relationship:

$$f(x, \theta) = HDR + \frac{LDR - HDR}{1 + 10^{slope \cdot (x - IC_{50})}} \tag{22}$$

In Equation 22, $\log_{10}$-transformed experimental concentration is referred to as x and its generated response given a set of parameters θ is $f(x,\theta)$.The free parameters (θ) of Equation 22 are hereafter referred to as efficiency metrics.

- LDR/HDR: the two asymptotic plateaus of the dose-response curve expressed as the low- and high-dose responses. Their interpretation is dependent on the type of response analyzed. For instance, when considering a decreasing dose-response relationship, the LDR represents the basal maximal responses, while the HDR represents the minimal response.
- IC_{50} : the inflexion points of the dose-response curve. It represents the concentration needed to generate a response that is half of the maximal response. Since we are considering $\log_{10}$ concentrations, values for IC_{50} are also $\log_{10}$. We use IC_{50} , for simplicity. When considering ascending dose-response curves (IRIC's dataset), we reference this metric as EC_{50} .
- Slope: the slope around the inflexion point (IC_{50}/EC_{50}). This metric is often referred to as the Hill coefficient. It is indicative of the stability of the compound.

For our main analyses, we only considered the four main efficiency metrics (θ) as described above. There exists alternative metrics, such as area under/above the curve [43] (AUC/AAC, Section 3.5.2) and drug sensitivity score (DSS, [42]), that could be derived from θ.

3.4.2.2. Levenberg-Marquardt: Estimating Efficiency Metrics

We estimated efficiency metrics through non-linear regression. To do so, we used the Levenberg-Marquardt algorithm [33, 34] as implemented by `LsqFit.jl` [`https:`

//github.com/JuliaNLSolvers/LsqFit.jl]. We used p_0 = [100, 0, 0, 1] to initiate θ = LDR, HDR, IC_{50} , slope, respectively, and used the default maximum number of iterations (1,000). The algorithm can converge before reaching 1,000 iterations. We retrieved the convergence status (true or false) of the algorithm for the maximum number of iterations for each experiment (Fig. 14A.).

We estimated efficiency metrics for each experiment of Gray, gCSI and CTRPv2. From an estimated θ_{LM}, we can produce a dose-response curve by calculating $f(x,\theta_{LM})$ for a given range of hypothetical concentrations (x). Examples of such curve are represented by solid lines in Figure 17.A. The corresponding θ_{LM} used to generate these curves are identified by the black/red vertical lines in the metrics subplots. We refer to the efficiency metrics obtained with this approach as Levenberg-Marquardt estimates.

3.4.2.3. BiDRA: Inferring Bayesian *Posteriors* of Efficiency Metrics

BiDRA (Bayesian inference for Dose-Response Analysis)9 is our Bayesian model (Fig. 14.B) that infers *posteriors* of efficiency metrics (θ). Similarly to Levenberg-Marquardt, BiDRA assumes that experimental responses are normally distributed around $f(x,\theta)$ with a shared standard deviation (σ) (Eq. 24).

$$P(\theta|y) \propto P(y|\theta) \cdot P(\theta) \tag{23}$$

$$P(y|\theta) \sim \mathcal{N}(f(x,\theta), \sigma^2) \tag{24}$$

More formerly, the probability of θ given some experimental response y is proportional to the product of the likelihood of observing said response given θ, and θ *priors*. We iterate this process over several iterations and chains, resulting in distributions of the most probable values for each efficiency metric. We refer to these distributions as *posteriors*.

The *priors* are *priori* information given to the model before considering any observation (i.e., experimental responses). Namely, we use *priors* as soft constraints on the range of possible and plausible values for each efficiency metric. They are distribution of values used during sampling, when the observations are insufficient to accurately orient the inference and the sampling. They allow us to computationally integrate both the experimenter's intuition and the experimental context. Our defined *priors* are weakly informative so that they do not outweigh experimental responses. We explored and tested various *priors* before defining the followings:

- HDR $\sim M$, where M is a mixture of two skewed Normal (μ = [0, 10], σ = [10, 20] and α= [1, -5]) and a $\mathcal{U}$(0,100) with respective weights λ = [0.4, 0.1, 0.5]
- LDR $\sim \mathcal{N}$(100,10)
- $IC_{50} \sim \mathcal{N}$(0,10)
- slope $\sim \log\mathcal{N}$(0.5, 1)
- $\sigma \sim \log\mathcal{N}$(1,1)

It is to be noted that the above listed *priors* are for decreasing dose-response curves (Gray, gCSI and CTRPv2). For increasing curves (IRIC), the LDR and HDR *priors* are adjusted to be representative of the response range.

Our choice of *priors* well fit the experimental context at hand, as demonstrated in Figure 14.C. When considering distributions of responses generated by the lowest experimental concentrations (basal response), we observe slight variation justifying our more precise/informative LDR *prior*. The distributions of responses generated by the largest experimental concentrations are, on the other hand, highly variable. Responses mainly range between 100% and 0%. Our wide HDR *prior* is thus appropriate. We did not use experimental responses to model our *priors*, we rather use them to express our intuition regarding the efficiency metrics. *Priors* are shared across experiments and datasets. We used log-normal *priors* for the slope and σ for the skewness and to ensure that the sampled values are positive. Because we define plateaus as LDR and HDR instead of "min" and "max", the slope is positive.

BiDRA is based on the No-U-Turn sampler (NUT-s, [105]), a Markov chain Monte Carlo (MCMC) algorithm. The obtained *posteriors* are a result of 4 combined MCMC chains of 2,000 iterations each (1,000 discarded warm-ups and 1,000 samplings). For this paper, BiDRA was implemented in the Julia language [`https://julialang.org/`] [49] using `Turing.jl` [`https://turing.ml/stable/`] [115] and `MCMCChains.jl` [`https://github.com/TuringLang/MCMCChains.jl`] with source code available here: `https://github.com/lemieux-lab/bidra_robustness`.

We inferred efficiency metrics *posteriors* independently for each experiment of Gray, gCSI, CTRPv2 and IRIC.

Given θ *posteriors*, we can calculate the median dose-response curve as well as confidence interval bounds (95%, represented by shaded regions in Figure 17.A). The calculation to obtain such curve is as follow:

(1) For each i sampling of the inference across all chains, calculate $f(x;\theta_i)$ for a range of hypothetical concentrations. The result is a response *posterior* for each of the hypothetical concentrations.

(2) Calculate the median and the [$\alpha/2$, 1 - $\alpha/2$] percentiles from each response *posterior*. The precision of the confidence interval is defined by α (e.g. 0.05).

Alternatively, one can plot individual curves for each sampling of the inference and get a global picture of the inference itself.

Examples of median dose-response curves and their corresponding *posteriors* are shown in Figure 17.A. *Posteriors* (colored histograms) are also contrasted to *priors* (light gray density lines) and Levenberg-Marquardt estimates (black/red vertical lines).

3.4.3. Assessing concordance of biological replicates efficiency metrics

3.4.3.1. Identifying replicated experiments

We define replicates as individual experiments for which the same compound was tested on the same cell-line. We considered pairs of biological replicates (R = 2) and ignored replicates of more than two experiments (R $\geq$ 3). As the aim of the present paper is to demonstrate BiDRA's robustness, we also limited our analysis to within-dataset replicates (i.e., biological).

3.4.3.2. Correlation coefficients

As there is no consensus on the metric to use [76, 86], we used Pearson (r) and Spearman's rank (ρ) correlation coefficients to quantify concordance between efficiency metrics of pairs of biological replicates. The Pearson coefficient measures the linear correlation between two variables and is sensitive to the presence of outliers. It is commonly used when comparing Levenberg-Marquardt estimates [151]. The Spearman's rank coefficient measures how well the relationship between two variables can be described by a monotonic function. More formally, it is the Pearson coefficient between two ranked variables. Contrary to the Pearson coefficient, the Spearman coefficient is insensitive to outliers. Comparison of both correlation coefficients across all dataset and efficiency metrics is presented in Supplementary Figure 20.B.

We describe correlation and consistency between two groups of measurements as followed:

- If correlation metric < 0.5 then the correlation is said to be **poor**, and the measurements are said to be discrepant.
- If correlation metric ≥ 0.5 then the correlation is said to be **good**, and the measurements are said to be somewhat concordant.
- If correlation metric ≥ 0.75 then the correlation is said to be **excellent**, and the measurements are said to be concordant.

3.4.3.3. Correlation between responses of biological replicates

We first considered correlation between responses of pairs of biological replicates. The aim of this analysis was to establish if we could expect some correlation between their efficiency metrics, as the latter derives from individually measured responses.

Replicated experiments do not always share the same concentrations. We limited our analysis to the comparison of shared-concentration responses (Fig. 15). Most pairs of replicates shared at least one concentration and were thus comparable: 84% of Gray, 66% of gCSI and 99% of CTRPv2. Gray's and CTRPv2's replicates share at least 4 and 5 concentrations, respectively. gCSI have a fraction (3%) of pairs of replicates with a single shared concentration. All three datasets have most of their replicates sharing many concentrations.

We assessed concordance by calculating correlation coefficients (Pearson and Spearman's rank) and by calculating the root mean square of response differences (RMSΔ) as in Equation 25. A small RMSΔ is indicative of similar response sets, while a larger RMSΔ correspond to dissimilar response sets (Fig. 15.E).

$$RMS\Delta = \sqrt{\sum_{i=1}^{N} \frac{(y_{1i} - y_{2i})^2}{N}} \tag{25}$$

In Equation 25, N is the number of shared concentrations between two replicated experiments, y_{1i} and y_{2i} are respectively the i^{th} responses of the first (1) and second (2) replicates of the pair. The assignment of "first" and "second" replicate is arbitrary and has no impact on the analysis. When interpreting RMSΔ results (Fig. 15), we need to consider the value of N. As previously mentioned, most replicates share multiple concentrations, and we can thus consider the RMSΔ as representative of the general relationship between the two replicated experiments.

gCSI and CTRPv2 response comparisons are straightforward. For each shared concentration between two replicates there are two responses to compare (Fig. 15.A and .B). We compared, respectively, 1,352 and 7,624 pairs of replicates by considering a total of 7,092 and 115,938 pairs of shared-concentration responses. Gray's response comparison was done differently since the experiments have multiple response measurements for each concentration. Given a shared concentration for a pair of replicates, we randomly paired a single response from each replicate. We calculated RMSΔ for a set of random pairings. We repeated this process 10,000 times and reported the mean RMSΔ and its standard deviation (SD) (Fig. 15.C). A total of 1,364 pairs of replicates were considered, across 12,219 shared-concentration response pairings.

3.4.3.4. Correlation between Levenberg-Marquardt efficiency metrics estimates of biological replicates

The assessment of correlation between efficiency metric estimates is straightforward. We considered 1,604, 2,044 and 7,647 pairs of biological replicates for Gray, gCSI and CTRPv2 respectively. We calculated Pearson and Spearman's rank correlation coefficients across all four efficiency metrics.

We considered various groups of biological replicates: (1) all pairs, (2) pairs with both experiments having complete dose-response curves (SD $\geq$ 20, see Section 3.4.4.1), and (3) pairs with both experiments having incomplete or unresponsive dose-response curves (SD $<$ 20).

3.4.3.5. Correlation between BiDRA efficiency metrics posteriors of biological replicates

When assessing correlation between pairs of *posteriors*, we considered two metric representations (Fig. 16): (1) *posterior* median values, and (2) complete quantile-to-quantile *posteriors* (BiDRA QQ).

For the first representation, we calculated correlation coefficients (Pearson and Spearman) on pos-terior median values across 1,604, 2,044 and 7,647 pairs of biological replicates for Gray, gCSI and CTRPv2 respectively. The main drawback of this representation is the abstraction of uncertainty. For instance, two experiments with correlating median values could in fact be discrepant: one *posterior* could be large and uncertain (inference heavily relied on the *prior*), while the second *posterior* could be more precise (inference heavily relied on the experimental data). We thus consider the obtained correlation coefficients for this representation to be slightly optimistic.

For BiDRA QQ *posterior*, we extrapolated the concept of comparing the median (50^{th} quantiles) of each replicate, to comparing every quantile (over all inference samplings, i.e., 4,000 samplings for the present analysis). *Posteriors* are first sorted, and the individual quantiles of replicates are paired. Similar distributions, when plotted, follow the diagonal. We calculated correlation coefficients on QQ *posteriors* across 1,604, 2,044 and 7,647 pairs of biological replicates for Gray, gCSI and CTRPv2 respectively. This representation allows to consider uncertainty when assessing if two experiments are concordant. Although less conventional, the strength of this representation is well demonstrated with the PF-4708671 replicates (Fig. 17.B).

3.4.3.6. Control experiment

To confirm the validity of our correlation analysis and results, we conducted a negative control experiment. We randomly paired two experiments (regardless of the drug and cell-line used) from the replicates pool. We considered the same number of replicates pairs as in our main correlation analysis. We calculated correlation coefficients for these sets of random pairs of experiments, across all efficacy metrics, and for Levenberg-Marquardt estimates and BiDRA QQ *posterior*. We repeated this experiment 10 times and represented the results as box plots of correlation coefficients (Fig. 18).

We considered various groups of random pairs of experiments: (1) randomly pair all experiments, (2) randomly pairs experiments with complete dose-response curve (SD $\geq$ 20, see Response Set Completeness section below), and (3) randomly pairs experiments with incomplete dose-response curve (SD $<$ 20).

3.4.4. Robustness evaluation

We found that BiDRA's robustness stems from its handling of experiments with incomplete or unresponsive curves, as demonstrated in Figure 17. To demonstrate and explain BiDRA's robustness, we compared curves completeness to inferred and estimated IC_{50} and HDR.

3.4.4.1. Response set completeness

We expected our assessment of concordance between efficiency metrics of biological replicates to be in part indicative of curve completeness and we thus needed a metric to quantify completeness. We opted to use the standard deviation (SD) of the response set of an experiment. A small SD is most likely indicative of an incomplete or an unresponsive curve, whereas a large SD is most likely indicative of a complete curve. Visually, one would describe a curve completeness based on the presence (or absence) of both plateaus, and the number responses that forms the high-concentration plateau.

After visual inspection of multiple responses sets, we approximated the completeness threshold to be around 20. We thus characterize experiments with SD $<$ 20 as being most likely incomplete or unresponsive, and experiments with SD $\geq$ 20 as being complete (Fig. 17.A).

We used this completeness metric to categorize curves and assess BiDRA's robustness in contrast to Levenberg-Marquardt (Fig. 17.C to .E).

3.4.4.2. IC_{50} and HDR

The completeness of a curve will affect metrics estimation and inference. We focus our analysis on the IC_{50} and HDR, two metrics that are of great interest to experimenters. It is worth noting that the LDR is less affected by curve completeness as it represents the basal response.

To quantify the observability of an estimated IC_{50} , we compared the estimates to the range of experimental concentrations (denoted by D) specific to each experiment (Fig. 17.C). We can then classify estimates as observable ($IC_{50} \in D$) or unobservable ($IC_{50} \notin D$). We also apply the concept

to IC_{50} *posteriors* and calculated the probability of an IC_{50} of being unobservable ($P(IC_{50} \notin D)$) (Fig. 17.D).

To quantify the uncertainty of *posteriors*, we calculate the difference between the two bounds of its 95% confidence interval (denoted by Δ). We found that this metric was better suited for exper-imenter's interpretation than standard deviation (SD) or median absolute deviation (MAD).

3.4.5. Typical Application of BiDRA: SAR Analysis and Compounds Selection

Analysis and interpretation of a set of *posterior* distribution might not be trivial to experimenters. We demonstrate the utility of using such representation when ranking compounds and evaluating selection criteria.

3.4.5.1. Compound ranking

Given a set of N compounds and their corresponding *posteriors*, we assess their rank probabilities for a given metric θ_i. Each θ_i *posterior* are sorted (ascending or descending, depending on the metric). We then assign a rank to the N values of each sorted sampling position. As a result, we obtain a rank *posterior* for each compound (C), from which we compute the probability of being at each rank $r \in 1,...,N$ (Eq. 26).

$$\Pr(C, r|\theta_i) \tag{26}$$

We thus obtain rank probabilities distribution for each of our N compounds and present these results in the form of a heatmap. To facilitate the visualization, we sort compounds based on *posterior* median values. Given rank probabilities, we can identify compounds that have a probability of at least p of being in the m first ranks (R) (Eq. 27).

$$\sum_{r=1}^{R} \Pr(C, r|\theta_i \geq p) \tag{27}$$

3.4.5.2. DAG representation

We found that directed acyclic graphs (DAGs) are an effective way to illustrate an overview of the relationships amongst a set of N compounds.

To build a DAG, each possible pair of compounds is compared as described in our initial paper [146]. If two compounds (C_i, C_j) are significantly different for a metric of interest at a significance threshold α, then an edge is added between nodes i and j, with redundant edges removed by transitive reduction58.

3.4.5.3. Compounds selection

The two approaches described above facilitate compound relationship visualization and compound selection. They help experimenters identify the compounds that are most likely the best amongst a set, and which are significantly better than a given compound. Finally, both approaches can be paired with various selection criteria, as exemplified with our structure activity relationship (SAR) analysis example (Fig. 19). Given a criterion and a *posterior*, we can empirically compute

the probability of meeting said criterion. Multiple criteria can be combined, resulting in a global selection probability. While using BiDRA and the resulting *posteriors*, many experimental questions can thus be adequately and statistically answered, while minimizing analytical bias.

3.5. Supplementary

3.5.1. Efficiency metrics consistency across replicates

Figure 20 presents Pearson correlation coefficients and coefficients comparisons.Figure 21 illustrates the concordance of the asymptotic basal responses between biological replicates.

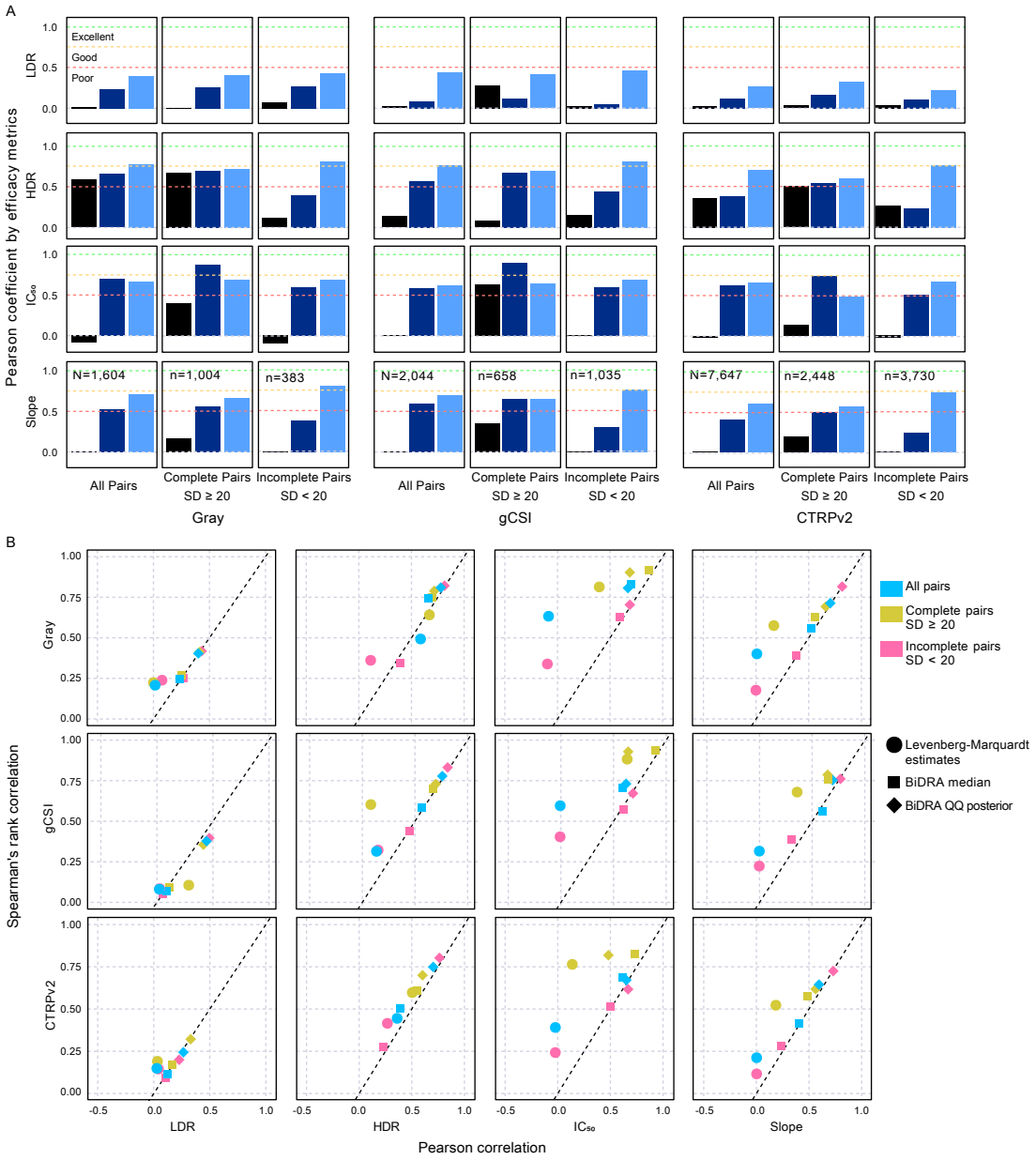


Fig. 20. Pearson correlation and coefficients comparisons. Figure 20: **(A)** Pearson correlation coefficient across metrics, datasets, and biological replicates groups. [*Caption continues on next page*]

Fig. 20 *(previous page).* Rows are metric-specific, column groups are dataset-specific, and columns are group-specific. Bars color is representative of the type of representation used: Levenberg-Marquardt estimate (black), BiDRA *posterior* median values (dark blue), and BiDRA *posterior* (light blue). We considered three grouping of pairs of replicates: all pairs, pairs with both experiments having complete dose-response curve (Complete pairs, SD $\geq$ 20), and pairs with both experiments having incomplete curve (Incomplete pairs, SD $<$ 20). The number of biological replicates considered for each group and each dataset are denoted by N (total) and n (subset). **(B)** We compared two widely used correlation coefficients: Pearson's (x-axis) and Spearman's rank (y-axis). Coefficients represent concordance level between efficiency metrics of biological replicates. We assessed concordance for all replicates (blue), pairs with both experiments having complete dose-response curve (Complete pairs, SD $\geq$ 20), and pairs with both experiments having incomplete curve (Incomplete pairs, SD $<$ 20). Efficiency metrics are represented by Levenberg-Marquardt estimates (circles), BiDRA *posterior* medians (squares) and BiDRA *posteriors* (diamonds). Rows are dataset-specific, and columns are efficiency metric-specific. The identity diagonal ($x = y$) is represented by the dashed line.

3.5.2. Area Above the Curve (AAC) consistency across replicates

We computed AACs values starting from the LDR, from Levenberg-Marquardt's estimates values, as well as from BiDRA's *posteriors*. For the latter, we computed an AAC value for each sampling of the inference, and thus obtained a *posterior* of the AAC itself.

To compare AACs of biological replicates (Supp. Fig. 23), we first had to establish the common range of experimental concentrations for each pair of experiment (Supp. Fig. 22). We confirmed that these common range were of reasonable size to assess AACs values (Supp. Fig. 22.C). We assessed AAC concordance using the same approach as for the basic efficiency metrics. Results are presented in Supplementary Figure 23.

As expected, BiDRA's representations (*posterior* medians and whole *posteriors*) show higher concordance than Levenberg-Marquardt's estimates, regardless of curve completeness (Supp. Fig. 23.B). Furthermore, all of BiDRA's correlation are either good ($\geq$ 0.50) or excellent ($\geq$ 0.75), keeping in line with our *prior* observations (Supp. Fig 23.B). AAC comparison results are similar to the ones made above for the basic efficiency metrics.

In addition to observing higher concordance with BiDRA's representations, we also notice experiments for which we are unable to calculate an AAC based on Levenberg-Marquardt's estimates. For instance, we compared 2,034 pairs of AACs for the gCSI datasets out of 2,044 possible pairs. The ten missing pairs had one experiment for which we were unable to calculate an AACs. When looking at such experiments (Supp. Fig. 24.A), we notice that they mostly represent unresponsive cell lines (9 experiments out of 10) and that at least one efficiency metrics was extreme. There is also one instance for which LDR $<$ HDR. These results emphasize pitfalls of using Levenberg-Marquardt, especially when the standard sigmoidal dose-response curve is ill-defined or absent.

When comparing *posteriors*, we were able to compare all possible pairs of experiments. We do, however, observe the unexpected presence of negative AAC values (Supp. Fig. 23.A). We hypothesize that these values are caused when, for a single inference sample, the LDR $<$ HDR. In such instances, since we are calculating the AAC from the LDR, we are actually getting the area under the curve and thus a negative AAC. We confirmed our hypothesis by looking at the LDR and HDR values of inference samplings resulting in a negative AAC (Supp. Fig. 24.B). The vast

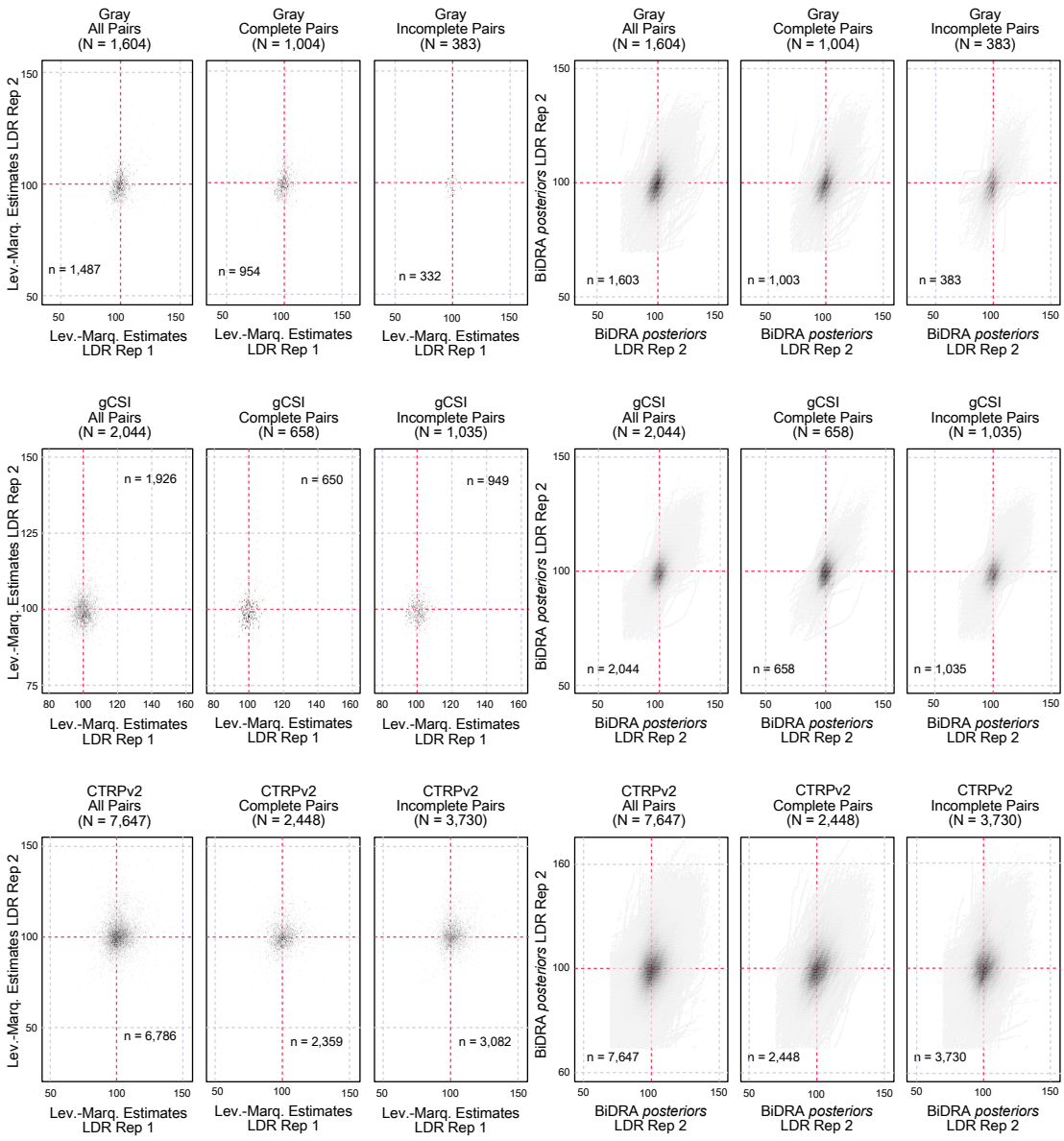


Fig. 21. Concordance of the asymptotic basal responses between biological replicates. Illustration of the lack of variability in LDR estimates and *posteriors* across biological replicates of all three datasets. Rows are dataset-specific, and column-triplet are method-specific. The number of pairs of replicates considered in each plot is identified by n and N (total number). Various grouping of replicates is considered: (1) all pairs, (2) pairs with both experiments having complete dose-response curve (Complete pairs, SD ≥ 20), and pairs with both experiments having incomplete curve (Incomplete pairs, SD <20). As expected, LDR comparison, regardless of representing or grouping, is centered at 100% (red axis).

majority of these samplings are associated with incomplete or unresponsive dose-response curves. It is important to note that these inference samplings only represent a fraction of many experiments (less than 35% of complete *posteriors*). These results further demonstrate BiDRA's robustness and flexibility:

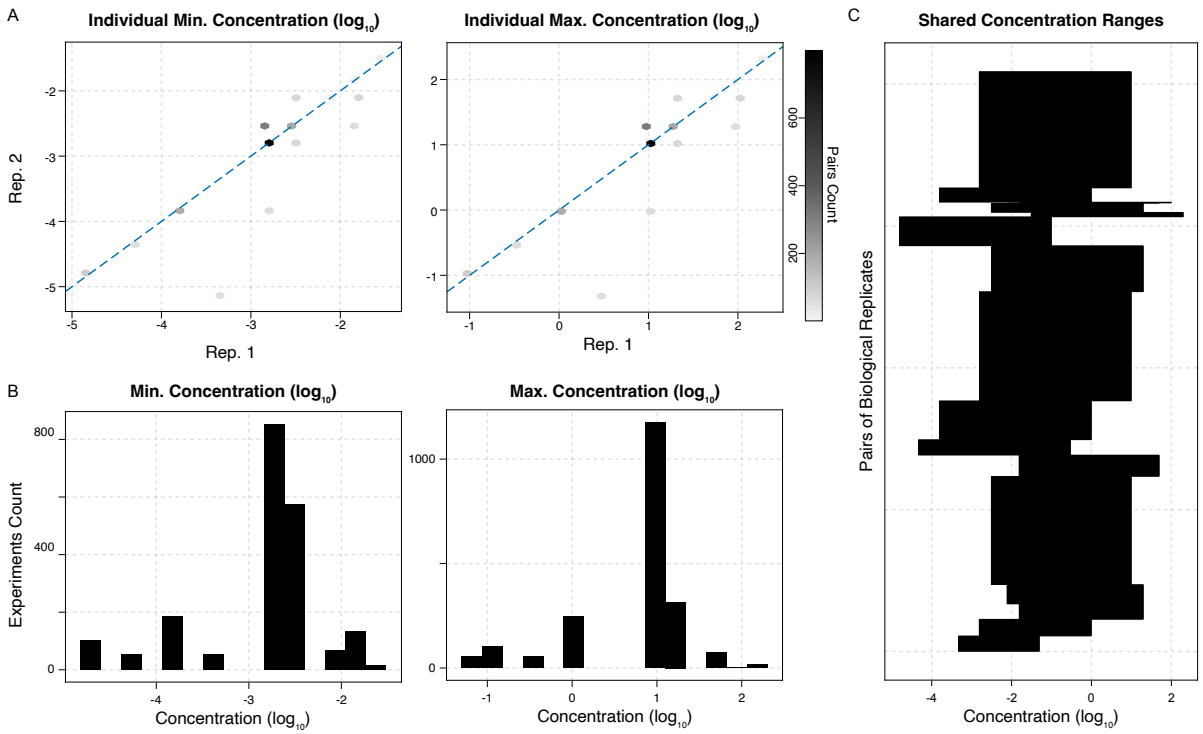


Fig. 22. Assessment of common experimental dose for pairs of biological replicates from the gCSI dataset. **(A)** Comparison of the minimal and maximal experimental dose between pairs of replicates. Replicates that share the same range fall on the diagonal (x=y). Color of each dot is representative of the number of pairs, as defined by the color scale on the far right. **(B)** Distribution of the minimal and maximal experimental dose of each experiment that is part of a pair of replicates (n = 4,088). **(C)** Visualisation of the length of common dose ranges across all pairs of replicates.

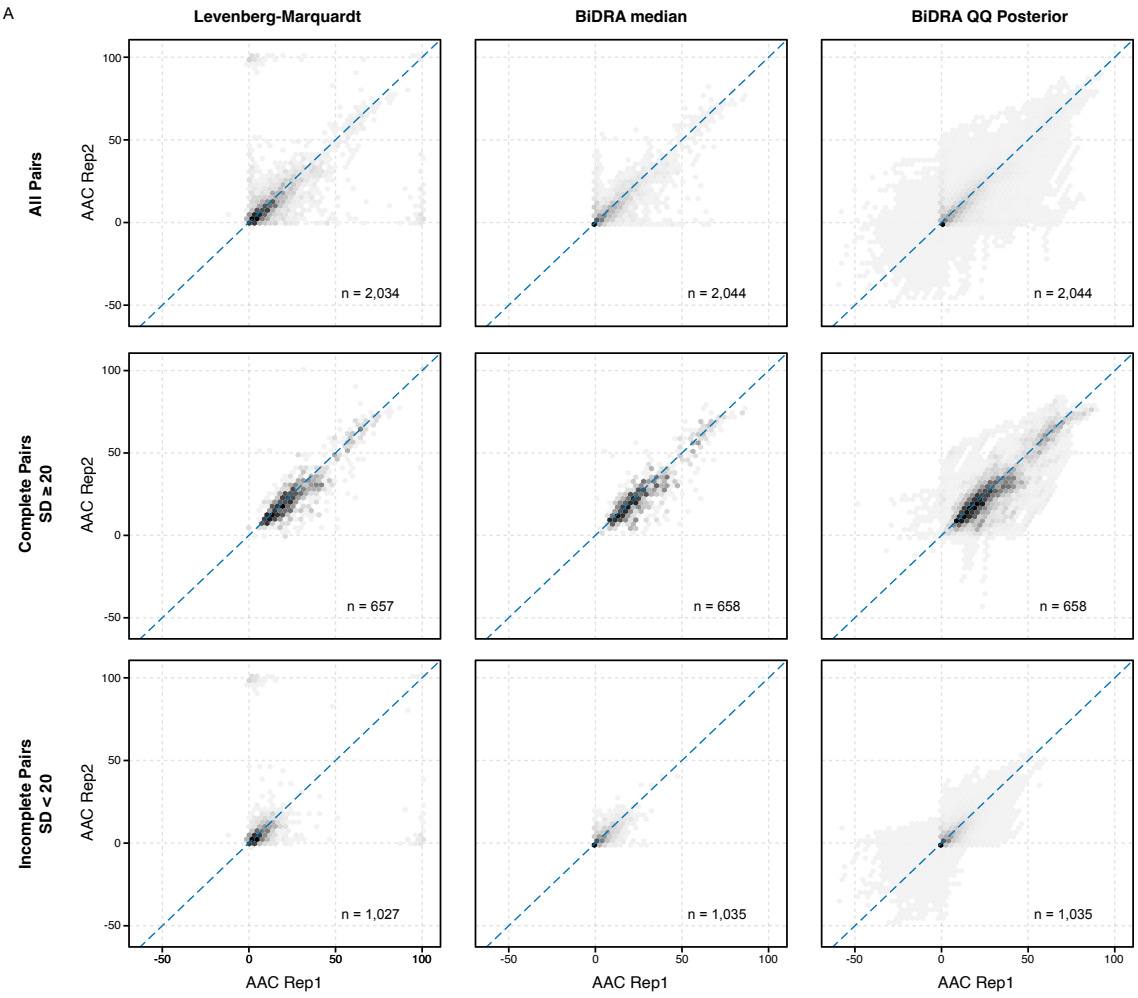
(1) BiDRA is less likely to return extreme values for efficiency metrics as it considers the experimental context through the *priors*.

(2) BiDRA allows the log-logistic model to "flip" its plateaux (descending to ascending curve) when needed. In contrast, Levenberg-Marquardt has a harder time to do so based on the initiating parameters used.

For these reasons, we choose to keep the negative AAC obtained with BiDRA as we consider them to be informative of the underlying experimental data (Supp. Fig. 23.A).

The AACs results discussed above were consistent across all three datasets.

A



B

		Spearman's Rank Correlation			Pearson Correlation		
		Gray	gCSI	CTRPv2	Gray	gCSI	CTRPv2
All pairs	Lev.-Marq.	0.75 (n=1,586)	0.56 (n=2,034)	0.51 (n=7,378)	0.57	0.42	0.24
	BiDRA median	**0.89 (n=1,604)**	0.69 (2,044)	**0.74 (n=7,647)**	**0.90**	**0.81**	**0.76**
	BiDRA posterior	**0.89 (n=1,604)**	**0.70 (n=2,044)**	**0.74 (n=7,647)**	**0.90**	**0.81**	**0.76**
Complete Pairs (SD ≥ 20)	Lev.-Marq.	0.83 (n=1,002)	0.81 (n=657)	0.73 (n=2,435)	0.74	0.86	0.68
	BiDRA median	**0.87 (n=1,004)**	**0.81 (n=658)**	**0.73 (n=2,448)**	**0.87**	0.87	0.75
	BiDRA posterior	**0.87 (n=1,004)**	**0.81 (n=658)**	**0.73 (n=2,448)**	**0.87**	0.87	0.75
Incomplete Pairs (SD < 20)	Lev.-Marq.	0.43 (n=371)	0.45 (n=1,027)	0.25 (n=3,515)	0.18	0.11	0.10
	BiDRA median	0.67 (n=383)	0.63 (n=1,035)	0.57 (n=3,730)	**0.80**	0.65	**0.67**
	BiDRA posterior	**0.72 (n=383)**	**0.67 (n=1,035)**	**0.63 (n= 3,730)**	0.77	**0.69**	0.65

Fig. 23. Concordance of area above the curve (AAC) between biological replicates. **(A)** Illustration of concordance between AAC values for the gCSI dataset. Each column is specific to a metric representation. arious grouping of replicates is considered: (1) all pairs, (2) pairs with both experiments having complete dose-response curve (Complete pairs, SD $\geq$ 20), and pairs with both experiments having incomplete curve (Incomplete pairs, SD $<$ 20). [*Caption continues on next page*]

Fig. 23 ***(previous page).*** Various grouping of replicates is considered: (1) all pairs, (2) pairs with both experiments having complete dose-response curve (Complete pairs, SD $\geq$ 20), and pairs with both experiments having incomplete curve (Incomplete pairs, SD $<$ 20).The number of pairs considered for each comparison is indicated by n. Correlation coefficient for these comparisons are reported in **B**. (B) Correlation coefficients (Spearman's rank and Pearson) for AACs comparison across all three datasets. Columns are dataset-specific while rows are method specific. Various grouping of replicates is considered: (1) all pairs, (2) pairs with both experiments having complete dose-response curve (Complete pairs, SD $\geq$ 20), and pairs with both experiments having incomplete curve (Incomplete pairs, SD $<$ 20). The corresponding number of pairs considered for each comparison is denoted in parenthesis by n. The same numbers were considered for both correlation coefficients. Correlation coefficients are classified as "poor" in red ($<$ 0.50), "good" in yellow ($\geq$ 0.50) and "excellent" in green. Bold coefficients are highlighted as the highest (i.e. best) for a given dataset and pair grouping. In case of equality, the "best" is selected as the one with the most pairs of experiments considered.

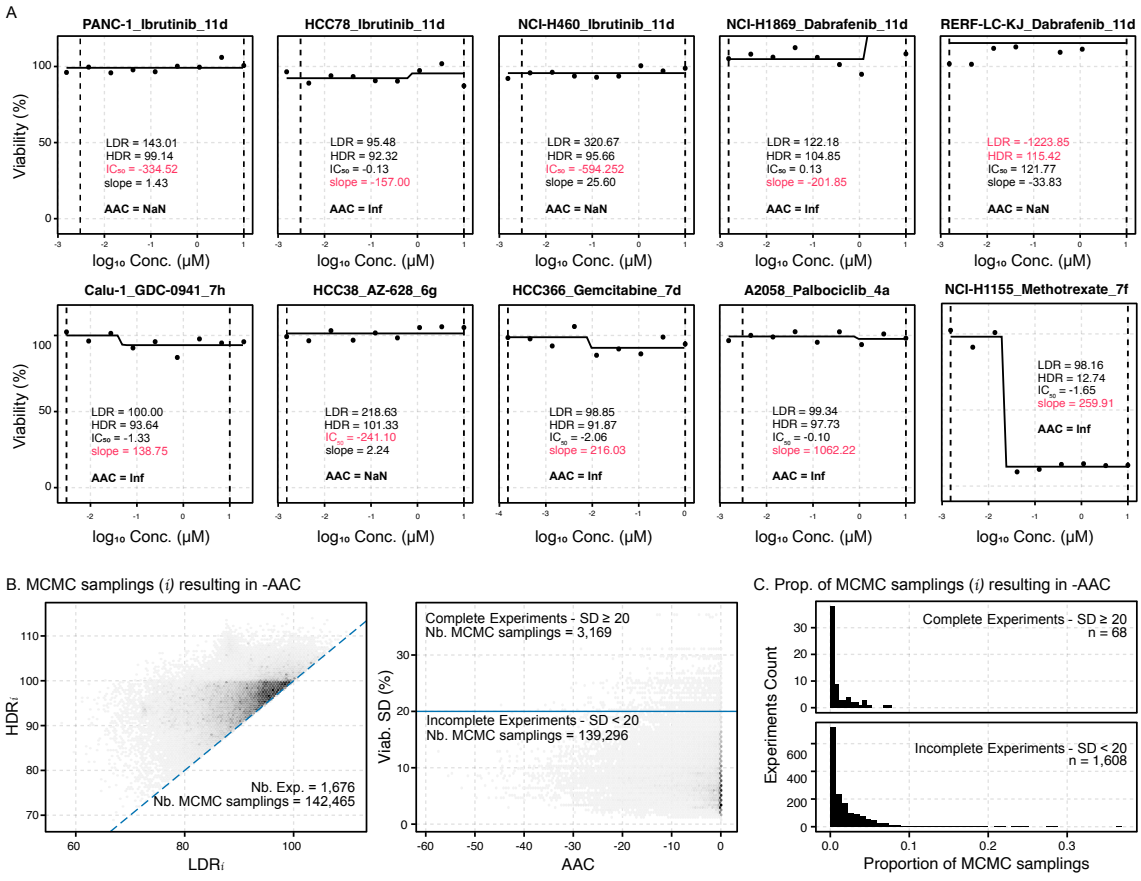


Fig. 24. Exploring special cases of AAC calculation for Levenberg-Marquardt and BiDRA. **(A)** Experiments for which the AAC calculation resulted in a Nan or Inf when using Levenberg-Marquardt estimates. This is in part due to Levenberg-Marquardt difficulty to estimate metrics when responses are characteristic of unresponsive curve, as it can return extreme values that are not supported by the experimental context. Such values are highlighted in red in each plot. The resulting dose-response curve is also plotted. [*Caption continues on next page*]

Fig. 24 ***(previous page).*** The experimental dose range considered for the AAC calculation is also denoted by the two dashed lines. The range is common to the experiments' replicates (not depicted). **(B)** (left) Comparison of LDR and HDR values for every inference sampling that resulted in a negative AAC. Across all experiments part of a pairs of replicates (n = 4,088), 1,676 had at least one inference sampling that resulted in a negative AAC. A total of 142,465 inference samplings are considered. (right) Comparison of negative AAC by inference sampling, to curve completeness. Experiments are considered to be "complete" when the standard deviation of their responses is ≥ 20, and "incomplete" when it is < 20. **(C)** Distribution of proportions of inference sampling that resulted in a negative AAC for each experiment considered in B. The upper panel is specific to complete experiments (n=68) and the lower panel is specific to incomplete experiments (the majority, n=1,608).

3.5.3. Control experiments

Figure 25 shows Pearson correlations for control experiments.

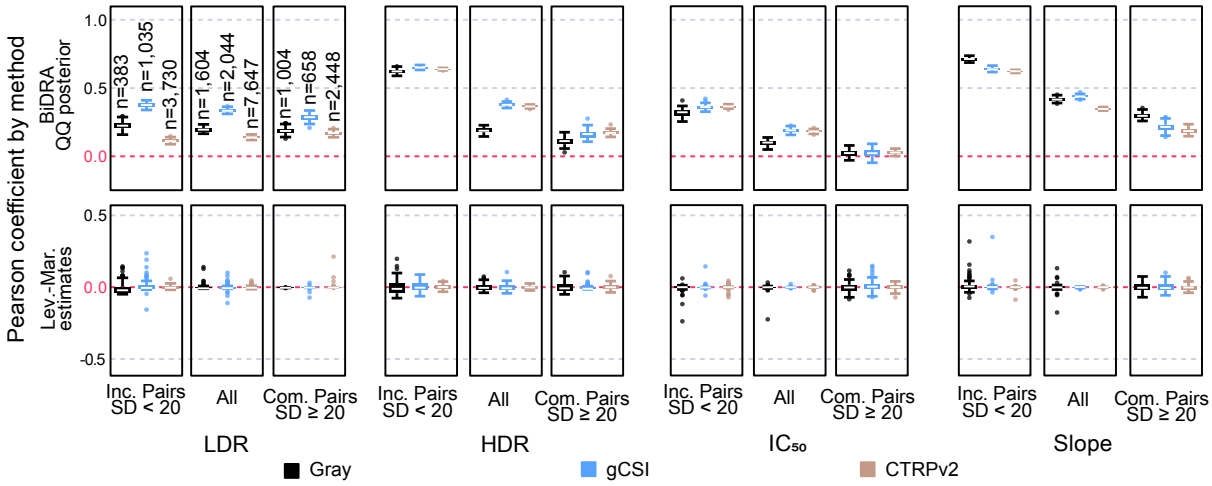


Fig. 25. Pearson correlations for control experiments.Experiments from different subset were randomly paired, regardless of the compound or cell-line used. Pearson correlation coefficients were calculated on these random pairings for 10 repetitions (represented as boxplots). Correlations are represented across metrics, datasets, and random pairings groups. Boxplots color is dataset specific. Three experiments subsets were considered: (1) all experiments with biological replicates (R =2) (All), (2) experiments with complete curves (SD $\geq$ 20), and (3) experiments with incomplete or unresponsive curves (SD $<$ 20). Correlation coefficients were calculated for both Levenberg-Marquardt estimates and BiDRA's *posteriors*.

3.5.4. BiDRA's robustness

Figure 26 illustrates the compararison of estimates and *posteriors* of IC_{50} and HDR while Figure 27 presents the comparison of HDR estimates and *posterior* uncertainty to curve completeness. Both Figures show results for the Gray and CTRPv2 datasets.

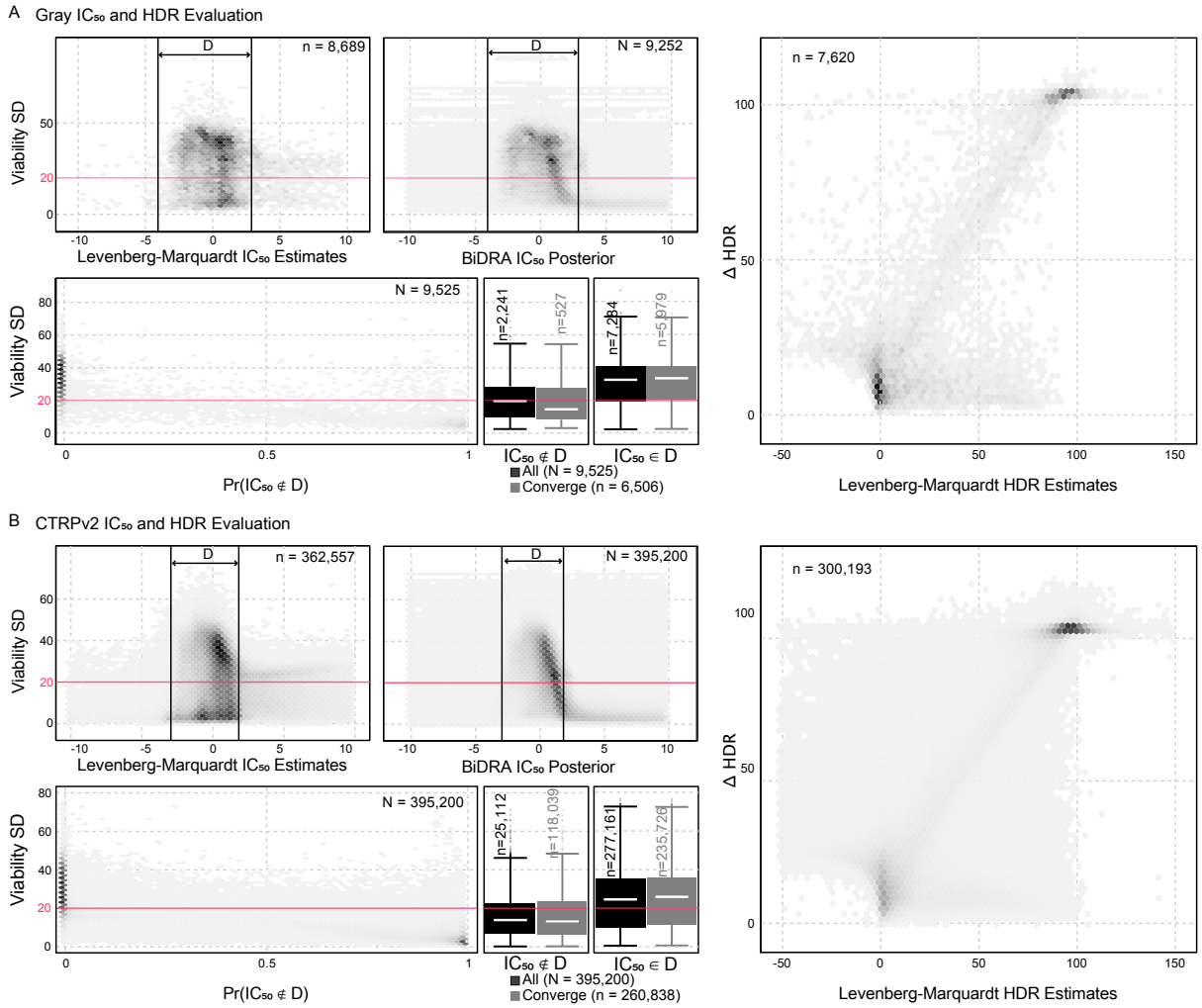


Fig. 26. Comparison of IC_{50} and HDR estimates and their *posterior* for Gray and CTRPv2. **(A)** Analysis of estimates and *posterior* for the Gray dataset. Complete description is identical to **B**. **(B)** The two upper left panels compare IC_{50} estimates (left) and IC_{50} *posteriors* (right) to response viability (SD). The red line divides experiment between complete response sets (SD $\geq$ 20) and incomplete/unresponsive response sets (SD $<$ 20). To establish the experimental dose range, denoted by D, we considered all experiments range and calculated 95% confidence interval bounds. The bounds are identified by black vertical segments. All experiments (N) were considered for the analysis, but only a subset is represented for the estimates plot (n). The two lower left panels compare IC_{50} classification (within/outside the experimental dose range) to viability SD. The large right panel compare Lev.-Marq. HDR estimates to BiDRA's uncertainty measurements for that same metric, ΔHDR.

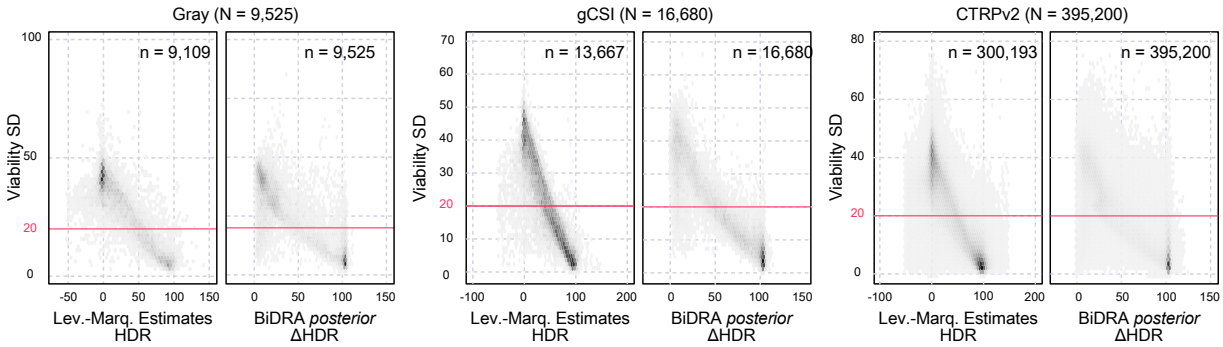


Fig. 27. Comparison of HDR estimates and *posterior* uncertainty to curve completeness. We define curve completeness by the standard deviation (SD) of a set of responses. Experiments with SD ≥ 20 are categorize as having a complete curve, while experiments with SD < 20 categorize as having an incomplete or unresponsive curve. HDR *posterior* uncertainty is denoted by ΔHDR and represent the difference between the bounds of the 95% confidence interval. While many experiments' HDR estimates follow the expected diagonal, there is still several experiments with small SD and unsupported HDR estimates. There are also the experiments for which Levenberg-Marquardt returned very high HDR estimates (most likely unresponsive curve) and incidentally unsupported and inaccurately observable IC_{50} estimates. The number of experiments represent in each plot is denoted by n while the total number of experiments for each dataset is denoted by N.

3.5.5. Other definitions of replicates

Figure 28 pressent correlations and other metrics when considering other definitions of replicates.

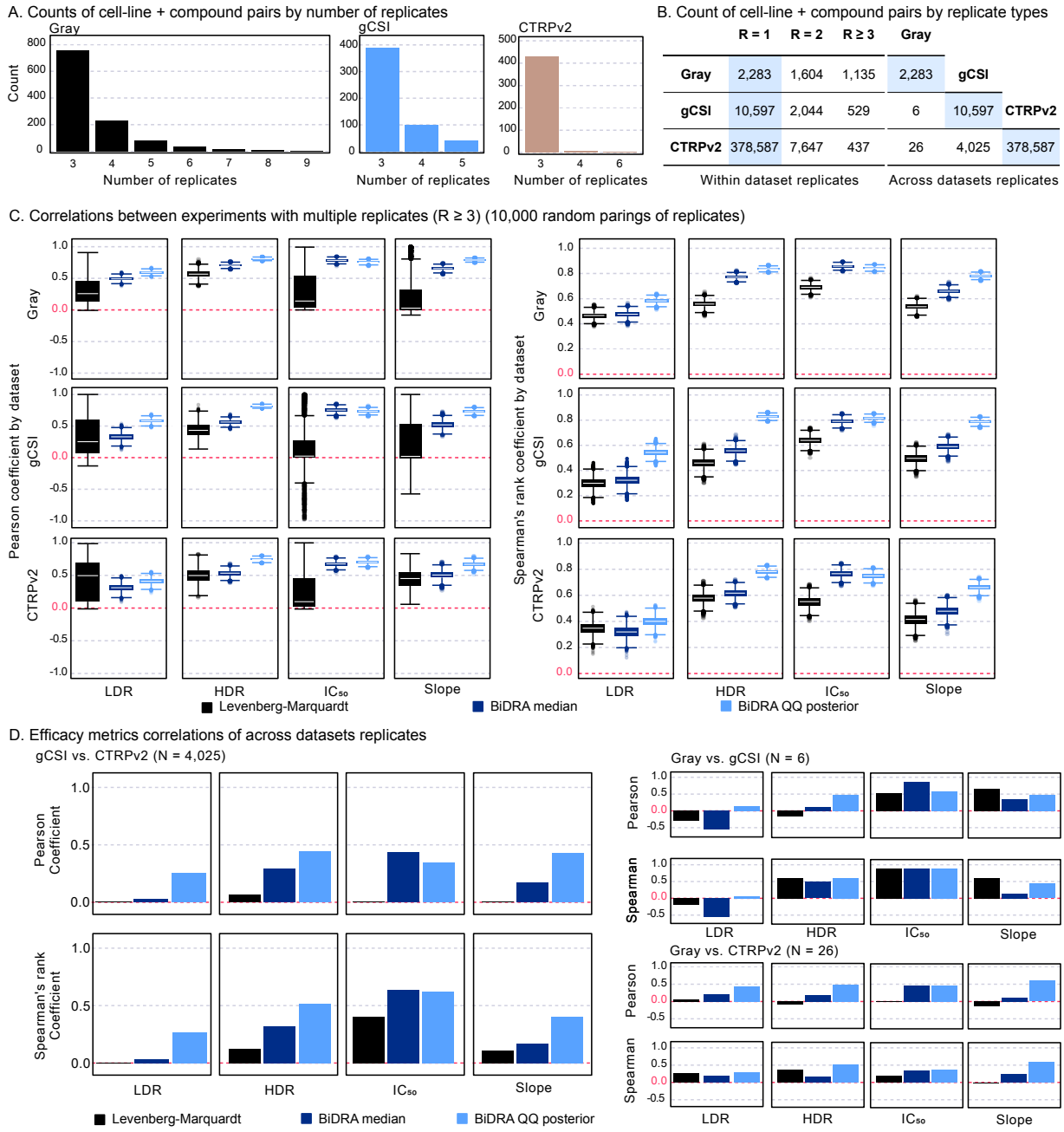


Fig. 28. Assessment of efficiency metrics correlations for within dataset multi-replicates and across datasets singletons. **(A)** To generate the coefficients referenced in **C**, we randomly paired two experiments for each cell-line + compound pair. The total number replicates for each of these pairs differ and is illustrated by the bar charts. Most cell-line + compound pairs have 3 replicated experiments. Colors are dataset specific. **(B)** The leftmost table contains the number of cell-line + compound pairs by datasets and replicate types. R = 1 are singletons (referenced in **D**), R = 2 are our initial definition of replicates (referenced in the main text), and R ≥ 3 are multi-replicates (referenced in **A** and **C**). The rightmost table contains the number of shared singleton (R = 1) for different dataset pairings. Singletons numbers are highlighted by blue shadowed cells. **(C)** Multi-replicates (R ≥ 3) correlation coefficients across efficiency metrics and datasets. Box plots illustrate correlation coefficients for 10,000 repetitions of replicates random pairings. [*Caption continues on next page*]

Fig. 28 ***(previous page).*** Colors are specific to metric representation (Levenberg-Marquardt estimates, BiDRA *posterior* median and BiDRA QQ *posterior*). The 0 baseline is highlighted in red. The number of cell-line + compound pairs considered in each dataset is indicated in the R $\geq$ 3 column of **B**.**(D)**Correlation coefficients across efficiency metrics of singletons shared by two datasets. gCSI vs. CTRPv2 is the most reliable results as the two other comparisons have very little singletons in common (the results are still represented in the right panels). Colors are specific to metric representation (Levenberg-Marquardt estimates, BiDRA *posterior* median and BiDRA QQ *posterior*). The 0 baseline is highlighted in red. The number of cell-line + compound pairs considered in each dataset comparison is indicated by N (and referenced in the rightmost table of **B**).

Chapitre 4

Outiller davantage les expérimentateurs : évaluation du potentiel informatif d'une expérience et mise à jour de l'interface BiDRA

Dans le Chapitre 2, une première interface web BiDRA est présentée. Celle-ci facilite, pour un expérimentateur, l'accès et l'utilisation de la méthodologie bayésienne présentée. Dans le Chapitre 3, la robustesse du processus bayésien est démontrée de façons qualitative comme quantitative, et en comparaison à l'approche traditionnelle par régression non-linéaire (Marquardt-Levenberg). Le processus est aussi appliqué à de larges jeux de données publics, soit les jeux de données Gray [69, 149], gCSI [66, 72] et CTRPv2 [63, 73, 74]. L'utilisation des *posteriors* de diverses expériences (jeux de données complets ou sous-ensembles) est aussi mise de l'avant: les gains analytiques et informationnels de l'utilisation des *posteriors* dans un contexte de sélection de composés pour une variété de critères d'efficience est clairement démontré et mis en valeur (Fig. 19).

Suite aux travaux du Chapitre 3, deux besoins supplémentaires liés au processus d'inférence ont été constatés : (1) une évaluation du potentiel informatif d'une expérience et de ces *posteriors*, puis (2) le renouvellement de l'interface BiDRA permettant notamment l'analyse de plus de deux expériences.

Le présent chapitre présentera une méthode analytique permettant de catégoriser les expériences selon leur potentiel informatif pour ainsi guider et outiller les expérimentateurs dans l'interprétation de leurs résultats. De plus, une nouvelle version de l'interface BiDRA sera présentée.

4.1. Prémisse de l'évaluation du potentiel informatif

Divers exemples d'analyse post-inférence exploitant l'information contenue dans les *posteriors* sont présentés dans les Chapitres 2 et 3. Parmi ceux-ci, il y a la comparaison de deux expériences (Section 2.2.2), la comparaison ordonnée de plus de deux expériences (Section 3.4.5.1), le calcul

de probabilités de sélection pour différents critères d'intérêt (Section 3.4.5.3), ainsi que la visualisation des relations statistiques entre les métriques d'un groupe d'expériences (Section 3.4.5.2). Ces exemples d'approches sont complémentaires au processus d'inférence et hautement pertinentes au processus décisionnel d'un expérimentateur. Bien que l'incertitude d'une métrique d'efficience puisse être représentée par le *posterior* même, il n'est pas toujours trivial pour un expérimentateur d'évaluer la quantité d'information contenue dans ce *posterior*. Cela peut être fait en inspectant visuellement les données expérimentales et/ou en comparant le *posterior* au *prior*. Or, cette approche qualitative est peu efficace lorsque plusieurs expériences doivent être considérées, et les conclusions peuvent différer d'un expérimentateur à l'autre.

Je propose dans les prochaines sections une méthodologie computationnelle permettant d'attribuer à toute expérience un sigle de couleur (●, ●, ●) décrivant le potentiel informatif d'une expérience. Ce processus se veut comme répliquant la démarche qualitative faite par les expérimentateurs lors d'évaluation visuelle.

La notion de "potentiel informatif" réfère à la quantité d'information contenue dans les données d'une expérience. Cette information limite les conclusions pouvant être tirées depuis les *posteriors* inférés. Par exemple, le potentiel informatif d'une réponse dite "plate" (Analogues 3 et 15, Fig. 30.A) est faible (●): les données expérimentales ne peuvent être utilisées pour inférer une IC_{50}/EC_{50} , un HDR ou une pente. Peu d'information peut être puisée des *posteriors* de ces métriques, outre qu'ils sont hautement incertains. Inversement, une réponse sigmoïde définie (Analogues 18 et 54, Fig. 30.A) a un potentiel informatif élevé (●): les données expérimentales sont suffisamment informatives pour inférer des *posteriors* précis des métriques d'efficiences.

L'évaluation du potentiel informatif d'une expérience est dépendante du modèle choisi pour représenter la relation dose-réponse. Les exemples mentionnés dans le précédent paragraphe considèrent l'utilisation du modèle BiDRA et de la fonction log-logistique (Éq. 2). Considérons plutôt le modèle Line, soit un modèle linéaire constant où la pente est toujours égale à 0. Le potentiel informatif d'une réponse plate devient alors plus élevé que celui d'une réponse sigmoïde définie. Pour déterminer le potentiel informatif d'une expérience, il est donc primordial d'évaluer la capacité du modèle choisi (c.-à-d. modèle BiDRA) à représenter la réponse expérimentale.

Un second élément à considérer est la complétude de la réponse. Par exemple, une réponse sigmoïde incomplète (Analogue 31, Fig. 30.A) a un potentiel informatif plus élevé qu'une réponse plate, mais moins élevé qu'une réponse sigmoïde définie. Le potentiel informatif d'une telle expérience est décrit comme étant "modéré" (●).

La capacité du modèle à représenter les données et la complétude des réponses forment la base de l'inspection visuelle faite par un expérimentateur. Ceux-ci forment aussi la base de la méthode d'évaluation du potentiel informatif décrit dans les prochaines sections. La Section 4.2 présente les divers concepts utilisés dans le processus d'évaluation du potentiel informatif, tel que décrit, démontré et appliqué à la Section 4.4.

4.2. Évaluation et comparaison de modèles bayésiens

Il est pratique courante d'évaluer un modèle par sa capacité à représenter de nouvelles données n'ayant pas encore été observées. Cette pratique se base sur le concept qu'un modèle est une interprétation humaine de la réalité et devrait donc être représentatif de celle-ci [97]. Lors de la comparaison de modèles, la capacité et la précision prédictives de ceux-ci sont évaluées et comparées.

La capacité prédictive d'un modèle s'évalue via ses **distributions *posterior* prédictives** (ppd, de l'anglais *predictive posterior distributions*) [97]. Une ppd représente la probabilité d'une nouvelle observation $\tilde{y}$ étant donné les observations y originales (Éq. 28) [164, 165].

$$\begin{aligned} ppd &= p(\tilde{y} \mid y) \\ &= \int_{all\theta} p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta \\ &= E_{post}(p(\tilde{y} \mid \theta)) \end{aligned} \tag{28}$$

Dans l'Équation 28, E_{post} dénote l'espérance des *posteriors* $p(\tilde{y} \mid \theta)$. θ dénote l'ensemble des paramètres de la fonction modélisant $\tilde{y}$ et y.

Comme pour le théorème de Bayes (Éq. 4), il n'est pas trivial d'obtenir les ppd de façon analytique (Éq. 28), puisque les vraies valeurs de θ sont inconnues [164]. Une approche numérique par échantillonnage est plutôt utilisée pour estimer les ppd. Pour chaque itération $i \in [1, I]$, θ_i est échantillonné depuis les *posteriors* tel que $\theta_i \sim p(\theta \mid y)$; une donnée fictive y_i' est aussi échantillonnée de la distribution de vraisemblance telle que $y_i' \sim p(y \mid \theta_i)$ [97]. Les ppd obtenus sont comparées aux observations y pour déterminer qualitativement si les ppd capturent bien les attributs des données (Fig. 29). Considérant le contexte expérimental de la présente thèse (Section 1.2), l'attribut principal à répliquer est la forme de la courbe dose-réponse.

La visualisation des ppd est un outil utile pour l'évaluation qualitative de la capacité prédictive d'un modèle. Pour une évaluation quantitative, une estimation de la précision de la capacité prédictive peut être calculée. Une description de ce calcul est faite dans les prochains paragraphes.

Optimalement, la précision de prédiction d'un modèle se calculerait en comparant les ppd de nouvelles observations $\tilde{y}$, soit $p(\tilde{y} \mid y)$ (Éq. 28), à la véritable distribution $f(y)$ du processus de génération de données. En d'autres termes, nous mesurerions la capacité de prédiction « hors-échantillon » (de l'anglais *out-of-sample*) du modèle [97]. Pour un ensemble de n données, cette mesure de précision est la *expected log pointwise predictive density*, ou la elppd (Éq. 29) [164]. Or, nous n'avons pas accès à $\hat{y}$ et la véritable fonction $f(y)$ nous est inconnue [164].

$$elppd = \sum_{i=1}^{n} E_f\big(\log[p(\tilde{y}_i) \mid y]\big) \tag{29}$$

Dans l'Équation 29, E_f dénote l'espérance de la fonction $f(y)$, et y dénote toujours l'ensemble des observations utilisées par le processus d'inférence.

La ***log-pointwise predictive density*** (lppd) est une approximation de la elppd et se base sur les données y observées (Éq. 30), plutôt que sur de nouvelles données $\tilde{y}$ (Éq. 29).

$$\begin{aligned} lppd &= \sum_{i=1}^{n} \log \int_{tout\theta} p(y_i \mid \theta) p(\theta \mid y) d\theta \\ &= \sum_{i=1}^{n} \log \left[E_{post}(p(y_i \mid \theta)) \right] \end{aligned} \tag{30}$$

Tel que mentionné plus haut, le vrai θ nous est inconnu. Il est cependant possible d'approximer avec une certaine précision la valeur de la lppd en utilisant les *posteriors* de θ, assumant un nombre d'échantillons d'inférence S suffisamment large (Éq. 31) [164]. Désormais, l'appellation lppd sera utilisée pour désigner le résultat de l'Équation 31.

$$lppd \approx \sum_{i=1}^{n} \log \Big(\frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta_s) \Big) \tag{31}$$

La lppd, telle que décrite par l'Équation 31, est une surestimation de la elppd puisqu'il y a réutilisation des données utilisées pour inférer les *posteriors*. Pour obtenir une estimation plus représentative ($\widehat{elppd}$), un coefficient de pénalité, ε, est appliqué à la lppd (Éq. 32)

$$\widehat{elppd} = lppd - \varepsilon \tag{32}$$

Le résultat de l'Équation 32 est référé comme étant un « **critère d'information** » (de l'anglais *information criteria*), pour des raisons historiques. De plus, pour les mêmes raisons historiques, le critère d'information est communément représenté par la déviance [97], transformant l'Équation 32 par l'Équation 33.

$$\text{critère d'information} = -2 \cdot \widehat{elppd} \tag{33}$$

Il existe diverses méthodes pour estimer la précision de prédiction hors échantillon d'un modèle, le critère Watanabe-Akaike (ou le *widely available information criterion*, WAIC) [166] étant le plus communément utilisé [97, 164]. J'ai comparé les critères WAIC et WAIC_k à la lppd pour le présent contexte d'évaluation du potentiel informatif. Tel que démontré à la Section 4.4, le critère WAIC_k semble être le plus pertinent et approprié pour ce contexte précis.

Pour le WAIC standard (Éq. 36), la lppd est pénalisée (Éq. 35) par une approximation du nombre effectif de paramètres (Éq. 34). Ce nombre est calculé via la variance des paramètres individuels pour l'ensemble des n données. Nous utilisons la variance des *posteriors* inférés pour calculer cette approximation [167, 168].

$$\varepsilon_{\text{waic}} = \sum_{i=1}^{n} V_{s=1}^{S} (\log[p(y_i \mid \theta_s)]) \tag{34}$$

$$\widehat{elppd}_{\text{waic}} = lppd - \varepsilon_{\text{waic}} \tag{35}$$

$$\text{WAIC} = -2 \cdot (lppd - \varepsilon_{\text{waic}}) \tag{36}$$

Le WAIC_k (Éq. 37) est une version simplifiée du WAIC: plutôt que d'utiliser le nombre effectif de paramètres, nous utilisons le nombre réel, k, de paramètres, soit 5 et 2 pour les modèles BiDRA et

Line respectivement.

$$\text{WAIC}_k = -2 \cdot (lppd - k) \tag{37}$$

Étant donné deux modèles, les métriques décrites ci-haut sont comparées dans le but d'identifier le modèle qui maximise la lppd ou minimise WAIC et WAIC_k. Pour simplifier la comparaison et la visualisation des résultats, j'utilise la différence (Δ) entre les valeurs d'une même métrique pour les deux modèles.

La Section 4.4 démontre l'application des concepts présentés ci-haut dans le contexte de l'évaluation du potentiel informatif d'une expérience.

4.3. Modèles comparés et données

Le modèle bayésien BiDRA est le même que celui décrit dans le Chapitre 3 (Fig. 14). Le modèle bayésien Line assume, tout comme BiDRA, que les données sont normalement distribuées autours de $\mu = f(\theta_{line})$ et pour un écart-type σ.

$$y \sim \mathcal{N}(\theta_{line}, \sigma) \tag{38}$$

La variable θ_{line} est constante pour l'ensemble des concentrations et représente l'instance où le LDR et le HDR sont égaux. Les *priors* utilisés sont similaires à ceux de BiDRA, soit $\theta_{line} \sim \mathcal{N}(0, 10)$ et $\sigma \sim log\mathcal{N}(1, 1)$. Le *prior* de θ_{line} est le même que celui du LDR (c.-à-d. la réponse basale). Nous considérons, dans les prochains exemples, une réponse d'inhibition de la croissance cellulaire, justifiant une réponse basale centrée à 0% (ou 100%, selon le type de réponse).

L'inférence de θ et θ_{line} est faite de façon similaire: l'algorithm NUT-S (tel qu'implémenté par la librarie `Turing.jl` [115]), est utilisé sur 4 chaines et pour 2,000 itérations (les 1,000 premières itérations sont utilisées en guise de *warmup* et sont exclues des *posteriors*).

Les analyses présentées dans les prochaines sections ont principalement été menées sur le jeu de données IRIC (70 expériences) présenté dans le Chapitre 3. Le jeu de données gCSI [72, 150], aussi présenté dans le Chapitre 3, est utilisé pour démontrer l'applicabilité et la validité de l'approche proposée sur un large jeu de données (Fig. 35).

4.4. Assignation d'un sigle décrivant le potentiel informatif d'une expérience

Les prochaines sections abordent la comparaison concrète des modèles BiDRA et Line (Section 4.3) ainsi que l'utilisation de cette comparaison pour catégoriser les expériences en termes de leur potentiel informatif. Les limitations (Section 4.4.3) et la généralisation (Section 4.4.4) du processus présenté sont aussi discutées.

4.4.1. Validation du processus comparatif

J'ai dans un premier temps confirmé la validité de la comparaison des modèles BiDRA et Line comme outil pour déterminer le potentiel informatif d'une expérience. Pour ce faire, la capacité prédictive de chaque modèle pour différents types d'expériences a été évaluée via une visualisation des distributions *posteriors* prédictives (ppd). La Figure 29 illustre cette évaluation pour deux exemples d'expériences provenant du jeu de données IRIC (Chapitre 3): l'Analogue 1 représente une réponse sigmoïde définie et l'Analogue 4 une réponse plate. Pour chaque modèle (Section 4.3), les ppd par concentration expérimentale (Fig. 29.A) sont estimées depuis les 4,000 échantillons des *posteriors* (θ et θ_{line}).

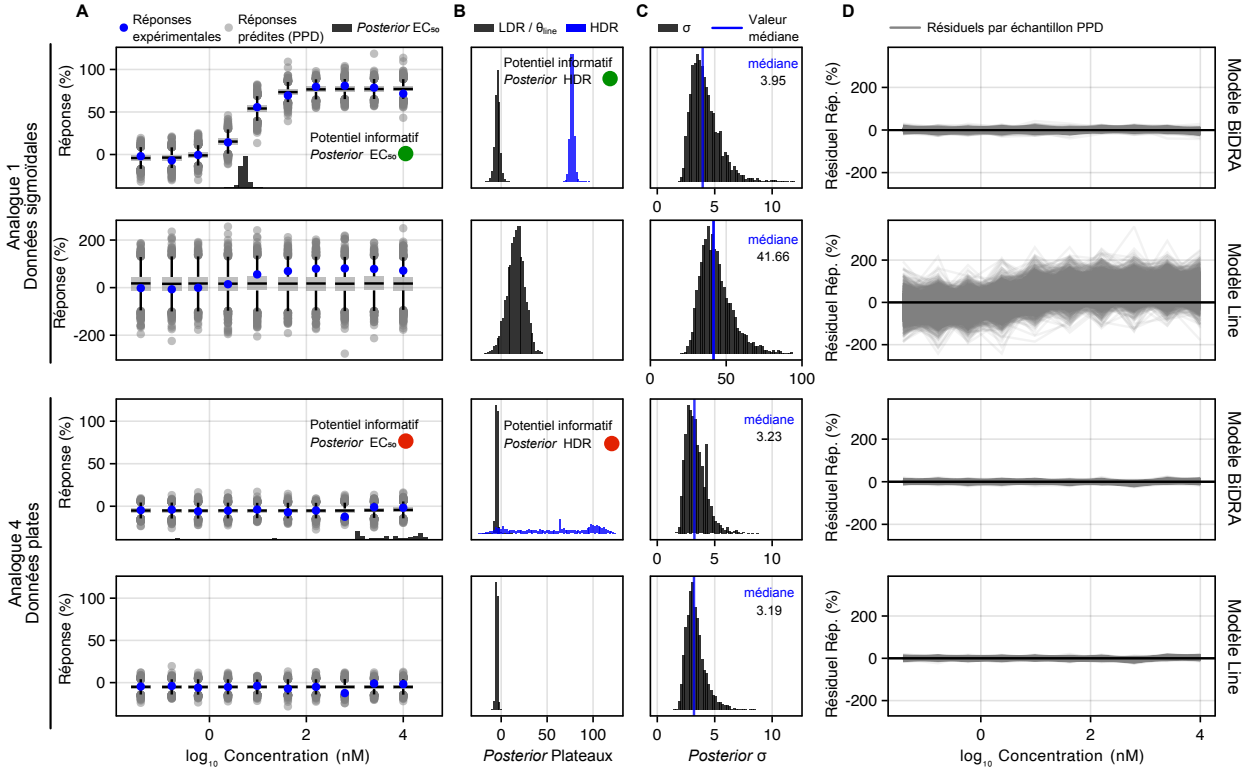


Fig. 29. Visualisation des distributions *posterior* prédictives (ppd) des modèles BiDRA et Line appliqués aux données des Analogues 1 (réponse sigmoïde) et 4 (réponse plate). **(A)** Données dose-réponse expérimentales (bleues) superposées aux données prédites y'_i obtenues pour 4,000 échantillonnages θ_i des *posteriors* (grises). Les y'_i constituent les ppd par concentration expérimentale. Pour chaque Analogue, les y'_i des modèles BiDRA (graphique du haut) et Line (graphique du bas) sont présentées. Le *posterior* de l'EC_{50} est représenté par un histogramme noir dans les graphiques pour le Modèle BiDRA. Les sigles hypothétiques du potentiel informatif de ces expériences sont identifiés. **(B)** Représentation des *posteriors* des plateaux pour les deux modèles, soit θ_{line} pour le modèle Line (noir), et LDR et HDR pour le modèle BiDRA (noir et bleu respectivement). Les sigles du potentiel informatif sont identifiés. **(C)** Représentation des *posteriors* du σ de la fonction de vraisemblance. Les valeurs médianes des *posteriors* sont identifiées. **(D)** Résiduels (réponses expérimentales vs. données prédites c.-à-d. $y_i - y'_i$) pour l'ensemble des 4,000 échantillonnages θ_i des *posteriors*. Ce sont les résiduels des données représentées en **A**.

Tel qu'attendu, seul le modèle BiDRA réplique la réponse sigmoïde (Analogue 1, Figure 29.A et .D), le modèle Line est invalide et non représentatif de la relation dose-réponse. L'ensemble des

posteriors associés à ce type de réponse est précis et suffisamment informatif quant aux valeurs des métriques d'efficience (Fig. 29.A et .B, ●)

De façon intéressante, les deux modèles ont une capacité similaire à répliquer une réponse plate (Analogue 4, Figure 29.A et .D). Tel que démontré dans le Chapitre 3, le modèle BiDRA est robuste aux expériences incomplètes ou plates grâce à l'incorporation des *priors* peu informatifs. Les *posteriors* LDR/θ_{line} (Fig. 29.B) et σ des deux modèles sont semblables (Fig. 29.B et .C) puisque les données expérimentales supportent l'inférence précise de ces métriques pour toute expérience (comparé au HDR, par exemple) et puisque les *priors* sont les mêmes pour les deux modèles. Nous remarquons que le *posterior* EC_{50} excède majoritairement la gamme de concentration expérimentale (Fig. 29.A, hors graphique), et que le *posterior* HDR est large et incertain (Fig. 29.B). Ces *posteriors* sont peu informatifs quant aux valeurs précises des métriques, et sont caractéristiques d'une expérience au faible potentiel informatif (●). Ces résultats démontrent que, de façon générale, lorsque la capacité prédictive des deux modèles est équivalente, le potentiel informatif d'une expérience est faible (●). Notons aussi que l'évaluation du potentiel informatif d'une expérience est particulièrement pertinente pour l'analyse et l'interprétation des *posteriors* HDR, IC_{50}/EC_{50} et pente.

Bien que les résultats de l'analyse qualitative des ppd n'aient rien de surprenant, ceux-ci confirment la validité de la méthode comparative. Ils sont aussi une démonstration supplémentaire de la robustesse de notre modèle BiDRA à inférer les métriques d'efficience peu importe le type de réponse considérée.

4.4.2. Définition de groupes de potentiel informatif

L'analyse des ppd abordée ci-haut est essentiellement qualitative et donc difficile à appliquer dans un processus d'analyse automatisé. De plus, l'assignation d'un sigle est tranchée (● ou ●), et il est difficile de définir ce qui caractériserait une expérience au potentiel informatif modéré (●).

Nous suggérons plutôt de définir des groupes de potentiel informatif (assignées à chacun des sigles ●, ● et ●) qui sont numériquement caractérisables par deux métriques décrivant la complétude de la réponse, et le modèle mathématique décrivant le mieux la relation dose-réponse. Pour illustrer et justifier la définition de ces groupes, j'ai identifié cinq expériences provenant du jeu de données IRIC (Fig. 30.A). Chaque expérience et son potentiel informatif attendu sont décrits ci-bas.

- Les Analogues 18 et 54 représentent des expériences ayant une réponse sigmoïde complète, pour différentes valeurs d'efficacité (HDR). Le potentiel informatif attendu de ces expériences est haut (●), puisque leurs *posteriors* respectifs sont précis et, notamment, indicatifs des valeurs HDR et EC_{50} .
- L'Analogue 31 représente une expérience ayant une réponse sigmoïde incomplète. Le potentiel informatif attendu de cette expérience est modéré (●). Les *posteriors* HDR et EC_{50} sont incertains, bien qu'informatifs. Par exemple, nous pouvons établir que le HDR est fort probablement supérieur à 50%.
- Les Analogues 15 et 3 représentent des expériences ayant une réponse plate. Le potentiel informatif attendu de ces expériences est faible (●) puisque les données expérimentales ne

suggèrent pas de valeur pour le HDR ni pour l'EC_{50} (ces paramètres ne sont pas observés expérimentalement). Leurs *posteriors* sont dès lors incertains et peu informatifs.

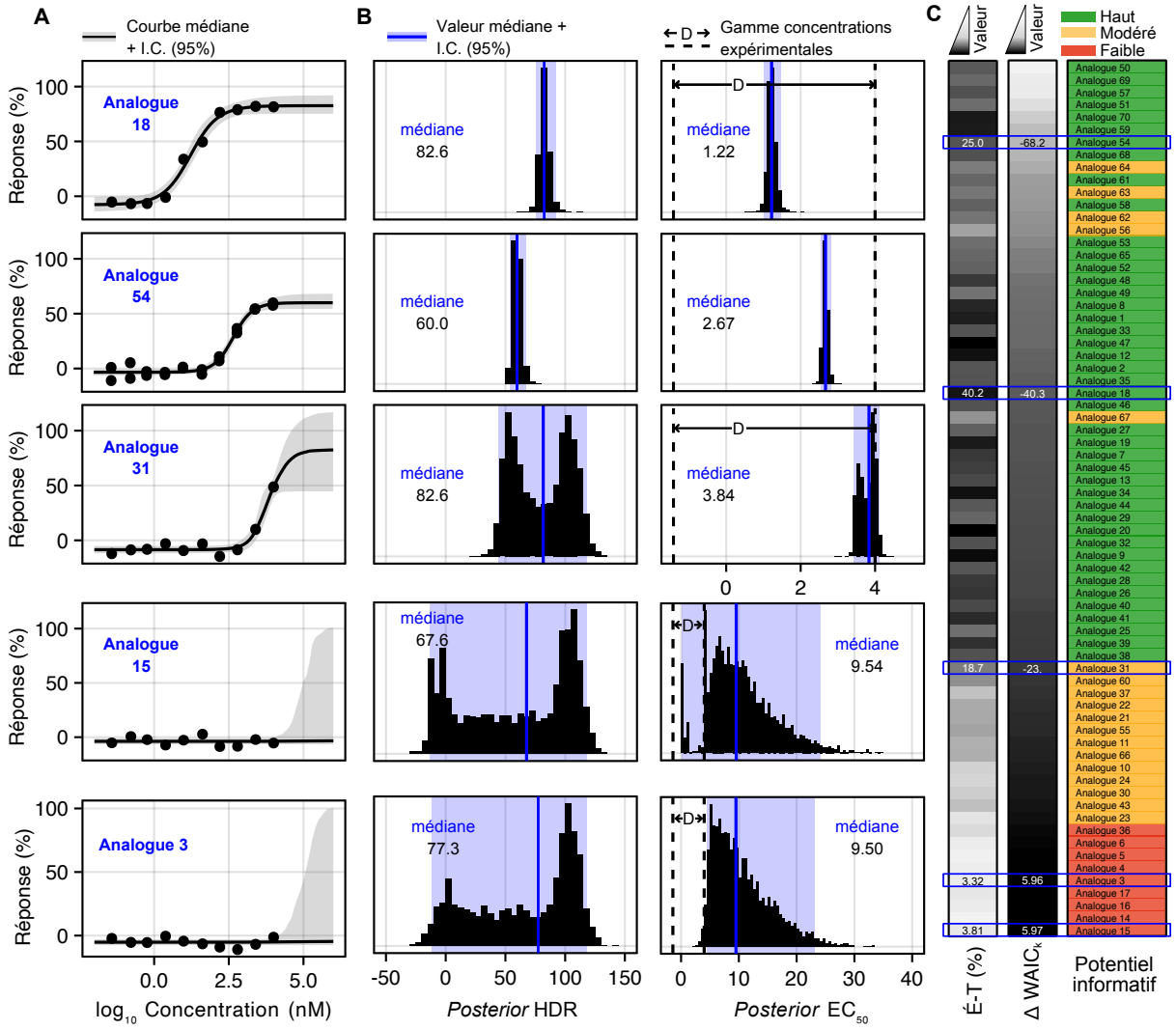


Fig. 30. Assignation des sigles de potentiel informatif des expériences du jeu de données IRIC et identification d'expériences exemples. Cinq (5) expériences sont identifiées comme des exemples types des différents sigles : Analogues 18 et 54 pour le sigle de haut potentiel informatif (●), Analogue 31 pour celui du potentiel modéré (●), et Analogues 15 et 3 pour le sigle de faible potentiel (●). **(A)** Données expérimentales et courbe dose-réponse. Cette dernière est représentée par une courbe médiane (trait gras) et un intervalle de confiance de 95% (région ombrée), tous deux obtenus depuis les *posteriors* du modèle BiDRA. **(B)** Représentations des *posteriors* HDR (gauche) et EC_{50} . Les valeurs médianes (traits bleus) et des intervalles de 95% (régions ombrées bleues) sont aussi identifiés. Dans le cas des EC_{50} , la gamme des concentrations expérimentales est identifiée par D, et ses bornes par les traits hachés. **(C)** Mise en contexte de l'assignation d'un sigle de potentiel informatif pour les 70 expériences du jeu de données IRIC. (Gauche) Écart-type des réponses expérimentales. (Centre) Δ $WAIC_k$. Les expériences ont été ordonnancées selon cette valeur de façon ascendante (du haut au bas). (Droite) Assignation des sigles de couleurs et identification des expériences. Les expériences exemples de **A** et **B** sont identifiées d'un encadré bleu et leurs valeurs sont identifiées numériquement.

Tel que mentionné ci-haut, les groupes sont en partie définis par le modèle (BiDRA ou Line) à favoriser pour représenter la relation dose-réponse. Nos résultats qualitatifs (Section 4.4.1, Fig. 29) et notre intuition suggèrent que les expériences privilégiant le modèle BiDRA (c.-à-d. log-logistique) auraient un potentiel informatif soit haut (●), soit modéré (●). Inversement, les expériences privilégiant le modèle Line auraient un faible potentiel informatif (●). Il est donc primordial de définir une métrique apte à identifier le modèle à privilégier pour représenter la réponse d'une expérience.

Trois métriques de quantification de la capacité prédictive d'un modèle sont considérées: (1) lppd, (2) WAIC_k et (3) WAIC. La lppd, car elle est l'approximation la plus directe de la elppd (elle forme d'ailleurs la base du WAIC_k et WAIC); le WAIC_k, parce qu'il est une version simplifiée et plus pénalisante du WAIC; et le WAIC, parce qu'il est présentement reconnu comme étant la meilleure alternative au processus de validation croisée [97]. La validation croisée (de l'anglais *cross-validation*) serait l'approche optimale. Elle est cependant coûteuse en termes de temps de calcul et donc peu appropriée dans le présent contexte pour outiller les expérimentateurs [164, 167, 168]. Les trois métriques ciblées se basent sur les *posteriors* et leur information, et sont alors décrites comme étant entièrement bayésienne [97, 164], comparées à d'autres critères d'information tels que les AIC et DIC [169]. Nos analyses du Chapitre 3 démontrent que de résumer les *posteriors* en une seule valeur (e.g. médiane) abstrait de l'information pertinente à l'analyse. Il est donc préférable d'utiliser une approche considérant l'ensemble des *posteriors*. La Figure 31.A compare les lppd, WAIC_k et WAIC des deux modèles pour l'ensemble du jeu de données IRIC. Les séparations et regroupement généraux d'expériences sont très similaires pour les trois métriques (fig. 31.A). Elles se différencient cependant par le nombre d'expériences favorisant le modèle Line: la lppd et le WAIC sont semblables avec 1 et 3 expériences, respectivement, et se distinguent des 9 expériences identifiées par le WAIC_k.

La lppd et le WAIC semblent sous-estimer le nombre d'expériences ayant une réponse plate (Fig. 34). Tel que démontré par plusieurs résultats de cette thèse (Fig. 7, 10, 16, 17 et 29), le modèle BiDRA est robuste aux expériences ayant une réponse plate. Souvenons-nous de la ressemblance des résultats de l'Analogue 4 lors de l'analyse des ppd (Fig. 29): pour la gamme de concentrations expérimentales, les deux modèles répliquaient bien la réponse observée. La lppd évalue les capacités des modèles à prédire ce qui est observable, et dans le cas d'une réponse plate, ces capacités sont quasiement identiques pour les deux modèles. Nous observons dans la Figure 31.A (lppd) plusieurs expériences, dont les Analogues 15 et 3, sur la diagonale d'identité. Ces expériences sont associées à un faible écart-type des réponses (< 5, Fig. 31.B) Par les ressemblances des lppds, le choix du modèle à privilégier (c.-à-d. la plus grande lppd) est sensible à de minime variance dans l'échantillonnage des *posteriors* (Fig. 33). Nous observons aussi pour ces expériences des pénalités WAIC similaires pour les deux modèles (Fig. 32.B, diagonale d'identité). Cette pénalité représente le nombre effectif de paramètres, soit le nombre de paramètres non contraints pour un modèle donné. Dans le cas d'une réponse plate, ce nombre est autours de 2 pour les deux modèles et représente le LDR/θ_{line} et le *sigma* de la fonction de vraisemblance. Les paramètres HDR, pente et EC_{50} sont, pour ce type de réponse, contraints aux *priors* puisque les données expérimentales ne les supportent pas. Comme pour la lppd, le choix du modèle à favoriser depuis la comparaison des valeurs WAIC est sensible aux variances de l'échantillonnage des *posteriors*. Notons, que le WAIC

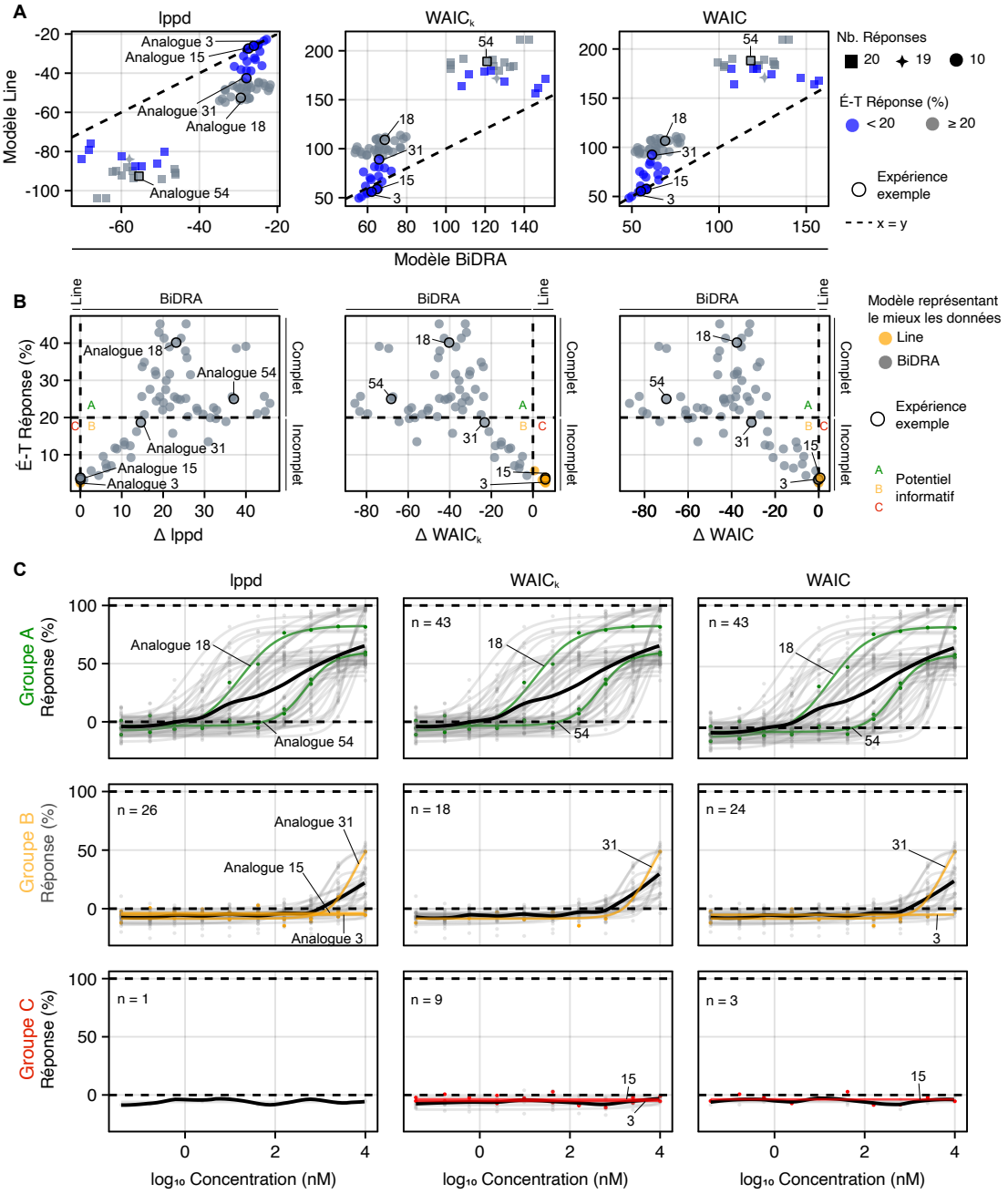


Fig. 31. Comparaison de métriques pour la définition des groupes et l'assignation des sigles de potentiel informatif. Les différentes expériences exemples de la Figure 30 sont identifiées dans **A**, **B** et **C**. **(A)** Comparaison des valeurs de métriques calculées pour les modèles BiDRA et Line. La ligne d'identité (x=y) est identifiée par le trait haché. La couleur des expériences est relative à la complétude de leur réponse (écart-type); la forme est elle définie par le nombre de réponses pour l'expérience. **(B)** Comparaison de la différence des métriques pour les deux modèles (Δ) et l'écart-type de réponse. Le trait haché $\Delta = 0$ distingue la sélection de Modèle : la couleur des expériences représente le Modèle favorisé. Les groupes de potentiel informatif sont identifiés : Groupe A/•, Groupe B/• et Groupe C/groupC. **(C)** Courbes médianes et données expérimentales de chaque groupe selon l'assignation pour une métrique. La tendance globale des réponses d'un groupe est représentée par une régression locale (LOESS, de l'anglais *locally estimated scatterplot smoothing*) (trait gras noir). Le nombre total d'expériences par groupe est dénoté par n.

est plus sensible que la lppd par l'ajout du calcul de pénalité, dépend lui aussi des *posteriors* (Fig. 33.B et .C). La lppd et le WAIC sont difficilement différentiables entre les deux modèles lorsque la réponse est plate, et cela mène à une sous-estimation du nombre d'expériences ayant ce type de réponse.

Le WAIC_k semble quant à lui bien identifier les expériences aux réponses plates ((Fig. 34)). Sa pénalité est beaucoup plus stricte que celle de WAIC: plutôt que d'utiliser le nombre effectif de paramètres, nous utilisons le nombre réel de paramètres, soit 2 et 5 pour les modèles Line et BiDRA respectivement. Pour une même lppd, le modèle Line est moins pénalisé et donc favorisé. Cela résulte en un groupe d'expérience ($n = 9$) sous la diagonale d'identité du graphique WAICk de la Figure 31.A. Une inspection visuelle des réponses expérimentales de ces neuf expériences (Analogues 3, 4, 5, 6, 14, 15, 16, 17 et 36) confirme qu'elles ont une réponse plate et que leur potentiel informatif est faible (●, Fig. 34).

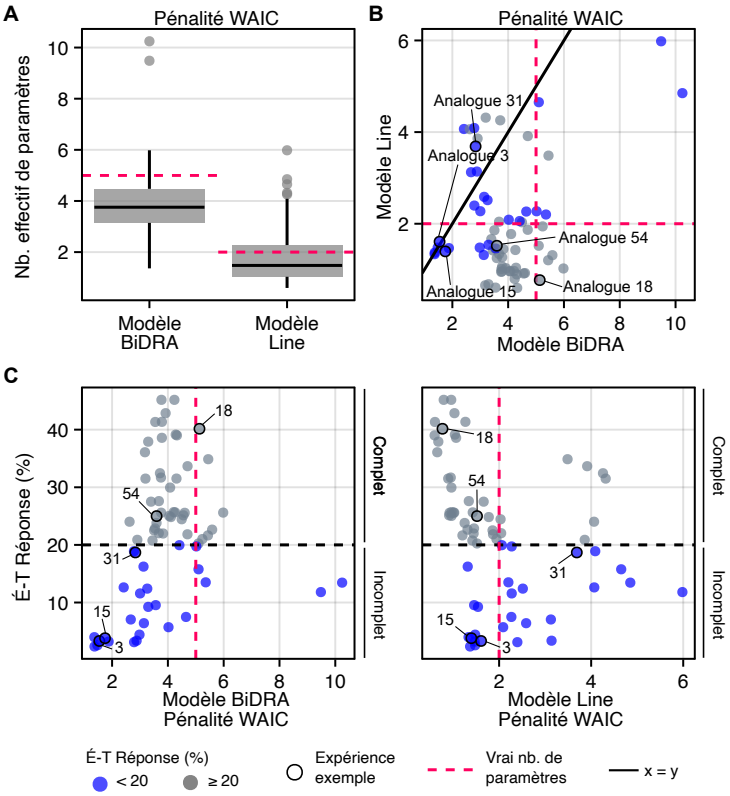


Fig. 32. Analyse et comparaison des pénalités de WAIC_K et WAIC pour les modèles BiDRA et Line. Les différentes expériences exemples de la Figure 30 sont identifiées dans **A**, **B** et **C**. Les traits hachés rouges indiquent les pénalités de WAIC_k, les vrais nombres de paramètres. **(A)** Comparaison des pénalités de WAIC (c.-à.-d. le nombre effectif de paramètres) pour les Modèles BiDRA et Line, par expérience. **(B)** Comparaison numérique et individuelle des pénalités de WAIC pour les modèles BiDRA et Line. La ligne d'identité (x=y) est identifiée par le trait noir. La couleur des expériences est relative à la complétude de leur réponse (écart-type). **(C)** Comparaison des pénalités WAIC aux écart-types des réponses, soit la complétude des expériences, pour le modèle BiDRA (gauche) et le modèle Line (droit). La couleur des expériences est idem qu'en **B**.

Le modèle à privilégier pour une expérience peut être déterminé en comparant la différence (Δ) d'une métrique donnée entre les deux modèles, à la ligne de base 0. La visualisation de la Figure 31.B illustre bien l'association des expériences au modèle à favoriser. Tel que mentionné plus haut, le nombre d'expériences privilégiant le modèle Line est inférieur à celui des expériences privilégiant le modèle BiDRA. La séparation faite par ΔWAIC_k est supportée par les données expérimentales (Fig. 34), contrairement à celles de Δlppd et ΔWAIC. Les expériences privilégiant le modèle BiDRA couvrent quant à elles un éventail de réponses, allant de la sigmoïde définie (Analogue 18, par exemple) à la réponse quasi-plate (Analogue 24, par exemple) (Fig. 31.C et 34). Pour différencier les expériences ayant un haut potentiel informatif de celles ayant un potentiel informatif modéré, nous utilisons l'écart-type (É-T) des réponses tel que présenté dans le Chapitre 3: une expérience est considérée comme ayant une réponse complète lorsque É-T ≥ 20, et inversement, une expérience est dite incomplète lorsque É-T ≤ 20. Les expériences à la réponse plate sont comprises dans la catégorie des réponses incomplètes.

La combinaison ΔWAIC_k et É-T permet de définir trois groupes d'expériences, chacun associé à un sigle de potentiel informatif (Fig. 31.B). Cette combinaison est la plus représentative de nos attentes (Fig. 34) et la plus appropriée à la tâche d'évaluation du potentiel informatif (Fig. 31.C). Les groupes se définissent comme suit:

Groupe A / • Expériences ayant une réponse sigmoïdale ($\Delta\text{WAIC}_k \leq 0$) complète et définie (É-T ≥ 20);

Groupe B / • Expériences ayant une réponse sigmoïdale ($\Delta\text{WAIC}_k \leq 0$) incomplète (É-T ≤ 20);

Groupe C / • Expériences ayant une réponse plate ($\Delta\text{WAIC}_k \geq 0$ et É-T ≥ 20).

La combinaison de ces deux métriques pour définir les groupes d'assignation semble plus appropriée que l'utilisation unique d'une ou l'autre des métriques, tel que démontrée par les résultats de la Figure 31.B. En effet, la séparation des expériences en trois groupes est plus nuancée et mieux justifiée, bien que cette approche comporte certaines limitations (Section 4.4.3). Les résultats de la Figure 31.A illustrent bien les difficultés à différencier les potentiels informatifs en ne considérant que la quantification de la capacité prédictive de chaque modèle. Premièrement, l'échelle de grandeur des valeurs est dépendante du nombre de réponses expérimentales considérées (Fig. 31.A). Il est difficile d'identifier une valeur seuil pour définir les différents groupes, applicable et généralisable à plusieurs expériences. L'utilisation de la métrique Δ ramène la comparaison sur une échelle commune à toute expérience et le seuil 0 est naturel (il n'est pas défini de façon subjective). Nous observons que les expériences se regroupent selon la complétude de leur réponse (bleu vs. gris, Fig. 31.A). La combinaison du Δ et de l'É.-T. est descriptive du potentiel informatif et appropriée au regroupement d'expériences en différentes catégories de potentiel informatif (haut •, modéré • et faible •).

Le processus d'évlauation du potentiel informatif d'une expérience, tel que décrit ci-haut, se base sur le ΔWAIC_k, bien que le WAIC_k ne soit pas communément utilisé pour quantifier la capacité

prédictive d'un modèle. Or, les résultats de la Figure 31 démontrent la supériorité de cette métrique en comparaison aux métriques plus standard telles la lppd et le WAIC. Je justifie l'utilisation du WAIC_k (et ΔWAIC_k) en quatre points.

Premièrement, nous connaissons et avons démontré à plusieurs reprises la robustesse du modèle BiDRA à bien représenter tous les types de réponses (sigmoïdes complète et incomplète, et plate). Nous ne comparons pas les modèles BiDRA et Line dans l'optique de déterminer lequel devrait être utilisé pour inférer les métriques d'efficience pour une expérience donnée: dans les faits, nous utiliserons toujours le modèle BiDRA, et considérerons les *posteriors* pour les HDR, $\text{IC}_{50}/\text{EC}_{50}$, LDR et pente. Nous n'utilisons pas la comparaison de modèle dans son contexte courant. Considérant que le modèle Line est imbriqué dans le modèle BiDRA (θ_{line} est similaire au LDR dans le cas d'une réponse plate), nous cherchons en fait à établir la complexité nécessaire pour bien représenter les données expérimentales d'une expérience. Si nous posons la question « Quel modèle représente adéquatement les données », nous savons que répondre « le modèle BiDRA » ne sera jamais faux. Or, si nous posons plutôt la question « Pour quel modèle arrivons-nous à prédire tous ou la majorité des paramètres? » la réponse variera d'une expérience à l'autre. En pénalisant la lppd par le nombre de paramètres k du modèle, nous répondons à cette deuxième question. Considérant le contexte d'évaluation du potentiel informatif d'une expérience, cette deuxième question devient plus pertinente que la première. Prenons l'exemple des Analogues 3 et 15. Leur potentiel informatif est faible (●). ΔWAIC_k nous indique qu'il est difficile d'inférer les cinq paramètres du modèle BiDRA (WAIC_k Line < WAIC_k BiDRA) pour ces expériences. Elles sont donc assignées au Groupe C (●), puisque leur É.-T. est inférieur à 20 (Fig. 31). Inversement, ΔWAIC nous informe que les deux modèles sont similaires en termes de capacité prédictive pour ces expériences, car leurs lppd et pénalité WAIC sont similaires (Fig. 31 et 32). La faible différence entre les pénalités WAIC des deux modèles résulte en une classification de l'Analogue 3 dans le Groupe B (●) et de l'Analogue 15 dans le Groupe C (●). Cet exemple démontre bien le gain et la validité de l'utilisation du ΔWAIC_k plutôt que ΔWAIC.

Deuxièmement, ΔWAIC_k est très similaire à ΔWAIC pour l'assignation d'expériences n'ayant pas une réponse plate. Dans de tels cas, les lppd des deux modèles se différencient suffisamment que la pénalité ajoutée n'a que peu d'effet. Prenons l'exemple des Analogues 18, 54, et 31. Ceux-ci sont assignés aux Groupes A (●), A (●) et B (●), respectivement, selon ΔWAIC et ΔWAIC_k. Dans le cas des Analogues 18 et 54, les deux métriques favorisent le modèle BiDRA (Fig. 31) bien que leur pénalité soit la plus élevée pour ce modèle (Fig. 32). Dans le cas de l'Analogue 31, le WAIC pénalise plus le modèle Line (Fig. 32) suggérant que celui-ci n'est pas à favoriser. WAIC_k pénalise, lui, d'avantage le modèle BiDRA. Or, les deux pénalités restent très semblables, et lorsque soustraites aux lppds, les conclusions restent les mêmes (Fig. 31).

Troisièmement, le ΔWAIC_k considère, tout comme le ΔWAIC, l'incertitude entourant l'inférence des divers paramètres puisqu'il se base sur la lppd.

Quatrièmement, ΔWAIC_k résulte à des assignations plus stables d'une instance d'inférence à une autre. Bien que la forme d'un *posterior* reste la même d'une inférence à l'autre, les valeurs échantillonnées varieront. Cela a comme impact de faire varier légèrement la lppd ainsi que la pénalité WAIC, puisque toutes deux sont calculées à partir des S échantillonnages constituant les

posteriors. La variation sera moindre pour le WAIC_k car ses pénalités sont constantes. La Figure 33 démontre les effets de ces variations sur l'assignation du potentiel informatif pour 100 répétitions d'inférence. Notons que les expériences assignées au Groupe A (●) ne fluctuent pas puisque cette assignation dépend principalement des valeurs des É-T de réponse qui sont indépendantes du processus d'inférence. Le nombre d'expériences assignées aux Groupes B (●) et C (●) varient plus lorsque l'assignation se base sur les lppds ou les WAIC (Fig. 33.A et .B). Le nombre est plus constant lorsque WAIC_k est utilisé: le nombre médian est obtenu pour 92% des répétitions (Fig. 33A et B). Les expériences souffrant le plus de la variation d'assignation lppd et WAIC sont celles ayant une réponse plate (Fig. 33.C et Fig. 34 pour référence). Cette variation s'explique par la similitude des valeurs llpd et des pénalités WAIC: les valeurs Δ varient légèrement autours du 0, changeant l'assignation d'une répétition à l'autre. Pour certaines expériences, ces variations résultent à une assignation quasi aléatoire (50/50). C'est notamment le cas des Analogues 14, 15, 16 et 17 (Fig. 33C). Considérant que le processus d'assignation du potentiel informatif vise à outiller les expérimentateurs, il est préférable que celui-ci soit le plus constant possible (sans ambiguïté et reproductible). WAIC_k stabilise l'assignation, et ce, pour des analyses répétées (Fig. 33C).

4.4.3. Limitations

L'approche d'évaluation du potentiel informatif présenté à la Section 4.4 combine la complétude de la réponse (É-T) et l'évaluation des capacités prédictives des modèles BiDRA et Line. L'assignation d'une expérience à un groupe de potentiel informatif est donc représentative des variations expérimentales puisque le processus exploite les *posteriors*. Bien que ce processus réponde à nos besoins et attentes (Fig. 34), il présente tout de même certaines limitations.

Premièrement, nous utilisons un seuil strict sur l'écart-type (É-T) des réponses expérimentales. Ce seuil a été sélectionné de façon arbitraire: après inspection sur trois larges jeux de données, un écart-type des réponses de 20% semblait bien séparer les expériences dites « complètes » (É-T $\geq$ 20%) de celles dites « incomplètes » (É-T $<$ 20%) (Chapitre 3). La forme de la réponse n'est cependant pas explicitement considérée pour définir sa complétude. L'Analogue 56, par exemple, semble avoir un faible HDR (sous 50%) bien que la courbe suggère une forme sigmoïde définie. La valeur de son ΔWAIC_k (-55.82) favorise le modèle BiDRA au modèle Line. L'assignation d'un potentiel informatif modéré (●) est dû à la valeur de l'écart-type (13.50). Alternativement, l'Analogue 25 présente aussi une réponse sigmoïde avec un faible HDR (sous 50%), mais son potentiel informatif est identifié comme haut (●). Son ΔWAIC_k (-24.92) privilégie aussi le modèle BiDRA, son écart-type est supérieur à 20 (20.76). Cette différence entre les Analogues 25 et 56 est causée par le seuil appliqué à l'écart-type des réponses. Pour des expériences ayant un faible HDR, l'écart-type est naturellement plus bas. Pour passer le seuil de l'écart-type, le plateau du HDR doit être bien défini et couvrir plusieurs concentrations. Dans le cas précis de l'Analogue 56, le HDR semble être défini par 20% des réponses et sur 2 concentrations; le HDR de l'Analogue 25 est, lui, défini par 30% des réponses et couvre 3 concentrations. Le calcul de l'écart-type des réponses étant affecté par le nombre de réponses et l'ajout d'un seuil strict sur celui-ci, limite l'assignation des expériences dans le groupe A (●). Parallèlement, ce seuil limite aussi l'assignation au groupe

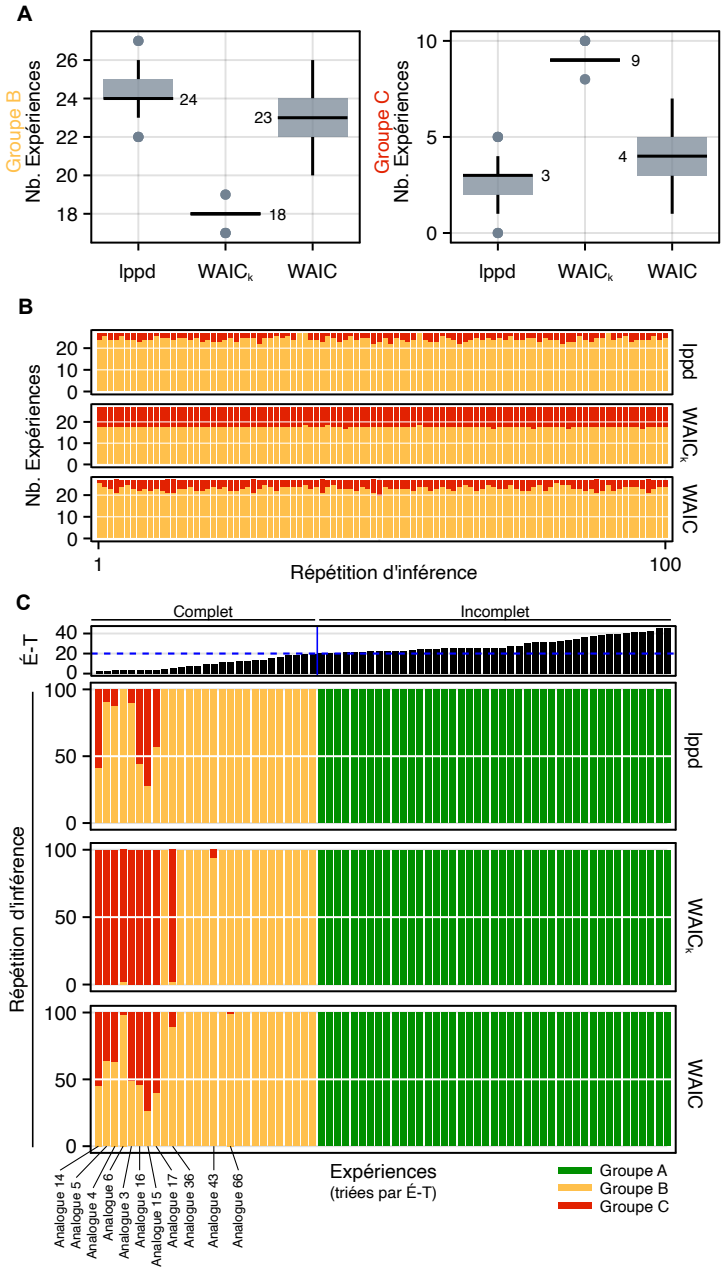


Fig. 33. Analyse de la constance du processus d'assignation de sigle de potentiel informatif pour diverses métriques. **(A)** Distributions du nombre d'expériences associées aux Groupe B/●(gauche) et Groupe C/●(droit) pour 100 répétions d'inférence. Les valeurs médianes de chaque distribution sont indiquées. **(B)** Visualisation de la variabilité des valeurs en **A**. Chaque bar représente la somme du nombre d'expériences du Groupe B/●et du Groupe C/●. **(C)** Assignation du groupe/potentiel informatif de chaque expérience pour 100 répétitions d'inférence. Les expériences sont ordonnées de façon ascendante selon l'écart-type de leur réponse, tel que présenté par les barres noires du haut. Les expériences ayant plusieurs assignations pour une ou l'autre des métriques sont identifiées dans le bas.

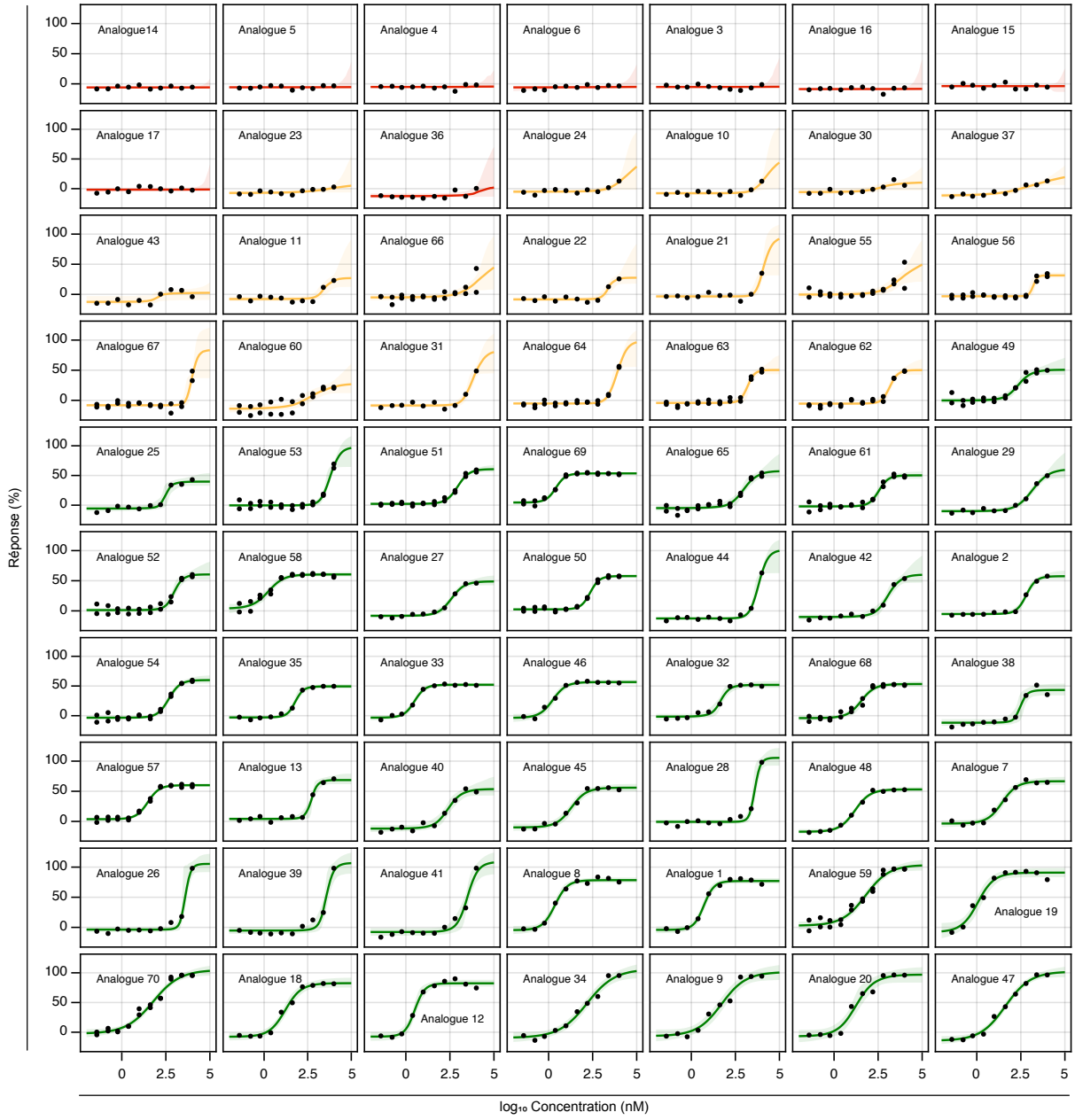


Fig. 34. Courbes dose-réponse du jeu de données IRIC. Représentations des données expérimentales et des courbes inférées par le modèle BiDRA pour chaque expérience du jeu de données IRIC. Les courbes sont représentées par la courbe médiane et un intervalle de confiance de 95%. La couleur de la courbe est représentative de l'assignation du sigle de potentiel informatif, tel qu'identifié dans la Figure 30. Les expériences sont ordonnées selon les écart-types de réponse (ascendant), soit le même ordre que dans la Figure 33.C .

B (●). Les Analogues 44 et 53 ont un écart-type légèrement supérieur à 20 (24.02 et 20.88, respectivement) et sont systématiquement assignés au Groupe A (●). Or, une inspection des réponses de ces expériences met en évidence qu'une seule mesure de réponse diffère du plateau du LDR. En d'autres mots, ces expériences ne sont pas réellement complètes. Les *posteriors* de leurs HDR,

EC_{50} et slope ne sont que modérément informatif (●): il nous est impossible de définir précisément les valeurs de ces paramètres, bien que nous puissions conclure, par exemple, que les HDR seront supérieurs à 60%. Pour ces expériences, la réponse de la plus haute concentration expérimentale est si élevée que l'écart-type des réponses dépasse le seuil de 20. Notons que la limitation engendrée par l'utilisation de l'écart-type affecte 3 expériences du jeu de données IRIC qui en compte 70 au total (Fig. 34).

Deuxièmement, il est difficile de différentier analytiquement une tendance biologique (e.g. augmentation de la réponse à une concentration donnée) d'une variance causée par des bruits biologiques, expérimentaux et/ou analytique. Cette difficulté et limitation n'est pas spécifique à la présente méthode bien qu'elle soit observée. Prenons l'exemple de l'Analogue 23. Une inspection visuelle identifierait une réponse plate et donc un faible potentiel informatif (●). Nous observons tout de même une légère augmentation graduelle et soutenue des réponses associées aux cinq dernières concentrations. Cette tendance se traduit par un $\Delta WAIC_k$ qui privilégie le modèle BiDRA. Il est aussi intéressant de noter que les Δlppd et ΔWAIC privilégient, eux aussi, le modèle BiDRA (Fig. 33.C, entre les Analogues 17 et 36). Alternativement, l'Analogue 36 est lui assigné au Groupe C (●). L'écart-type des réponses des Analogues 23 et 36 sont très similaires (4.42 et 5.72, respectivement). Ces expériences diffèrent par leur $\Delta WAIC_k$ (-2.87 et 0.79, respectivement), et par leurs *posteriors*. Dans le cas de l'Analogue 23, la variabilité des dernières réponses est analytiquement interprétée comme étant une progression vers le point d'inflexion (EC_{50}); chez l'Analogue 36, cette variabilité est plutôt interprétée comme étant du bruit. La différenciation du type de réponse pour ces deux expériences relève essentiellement de l'intuition d'un expérimentateur et est difficile à reproduire analytiquement. L'assignation des Analogues 23 et 36 aux Groupes B (●) ou C (●) est donc limitée par l'incapacité de la méthode d'inférence à distinguer une tendance d'une variance. L'intégration de l'intuition humaine dans le processus analytique n'est pas triviale, bien que nous le fassions via l'utilisation de *prior* dans notre processus d'inférence (Chapitre 2).

Malgré ces deux principales limitations, le processus présenté à la Section 4.4.2 reste le plus approprié à la tâche d'assignation du potentiel informatif des *posteriors* décrite plus haut. En effet, l'utilisation d'un critère d'information tel que le $WAIC_k$ pour décrire et comparer la capacité prédictive d'un modèle par rapport à un autre est pratique courante et recommandée [90, 91]. De plus, l'utilisation du $\Delta WAIC_k$ limite l'utilisation de seuils aléatoires. La combinaison informative de l'écart-type des réponses et du $\Delta WAIC_k$ nous permet de retourner une description simple à interpréter (c.-à-d. un sigle de couleur par expérience, soit ●, ●ou ●) pour un expérimentateur, et constante. Finalement, l'évaluation du potentiel informatif d'une expérience se veut un outil pour guider l'interprétation des *posteriors*: cette information ne devrait pas être utilisée pour exclure des expériences.

4.4.4. Généralisation

Le jeu de données IRIC (Fig. 34, 70 expériences) a été utilisé pour présenter et valider la méthodologie d'évalution du potentiel informatif présentée à la Section 4.4.2. La taille relativement

petite du jeu de données facilite la visualiation des résultats et leur interprétation qualitative (Fig. 30 et 34) et quantitative (Fig. 31).

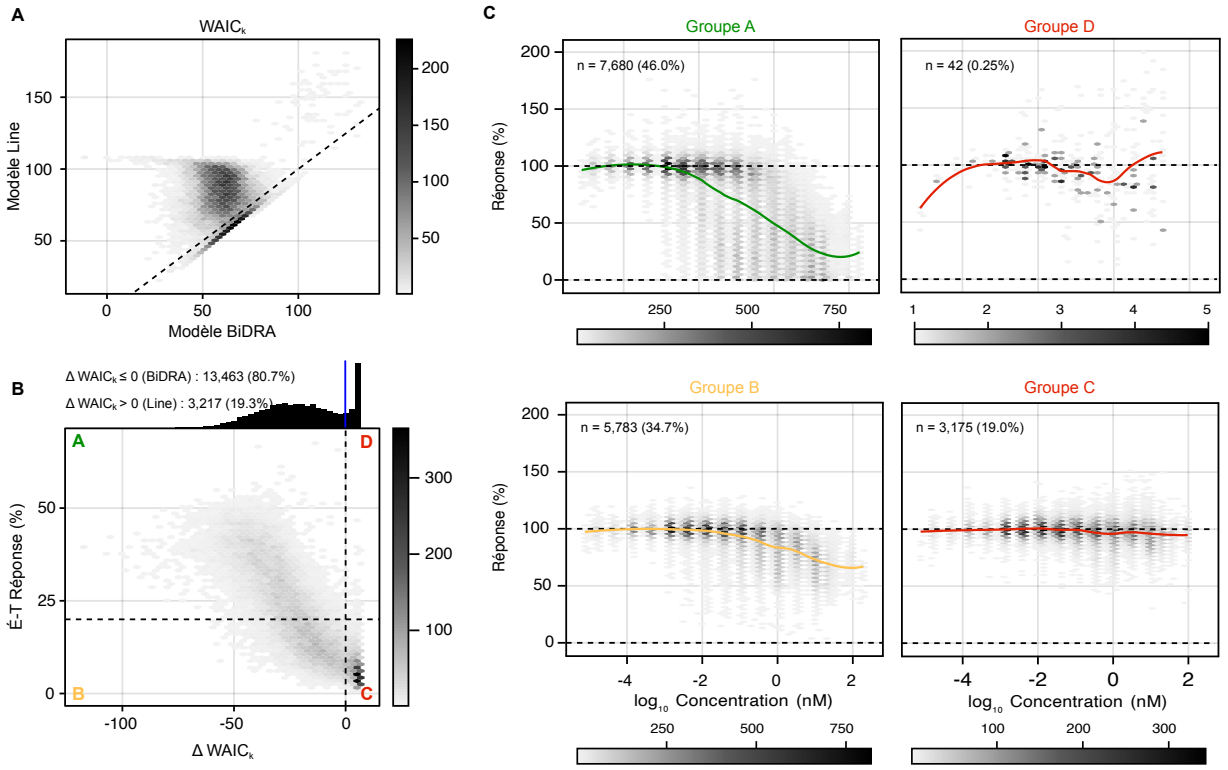


Fig. 35. Généralisation de l'approche d'assignation de sigles au jeu de données gCSI. **(A)** Comparaison des valeurs WAIC$_k$ pour les modèles BiDRA et Line pour chaque expérience du jeu de données gCSI (N = 16,680). La ligne d'identité (x=y) est identifiée par le trait haché noir. **(B)** Comparaison des valeurs de Δ WAIC$_k$ et d'écart-type de réponse. Les groupes d'assignation de sigles sont démarqués par les traits hachés noirs (Δ WAIC$_k$ = 0 et É-T = 20), et identifié par les annotations A, B, C et D. La distribution des valeurs Δ WAIC$_k$ est représentée par l'histogramme noir. **(C)** Représentation des tendances globales des réponses pour chaque groupe de **B**. Les traits de couleurs représentent la régression locale (LOESS) et les nombres totaux d'expériences par groupe sont dénotés par n.

La méthode assignant une expérience à un groupe selon son potentiel informatif se généralise bien à un jeu de données plus large, tel que le gCSI (Fig. 35, 16,688 expériences). Les résultats sont d'autant plus intéressants lorsque l'on considère qu'il serait difficile, voire impossible, d'évaluer le potentiel informatif de façon qualitative.

De façon générale, les résultats obtenus pour gCSI (Fig. 35) sont semblables à ceux obtenus pour le jeu de données IRIC (Fig. 31). Un peu moins de 20% des expériences de gCSI semblent avoir une réponse plate. Nous remarquons qu'un quatrième groupe d'expériences a été créé, soit le D (•). Ce dernier regroupe des expériences ayant un haut écart-type des réponses ($\geq 20\%$) et un ΔWAIC$_k$ supérieur à 0. Selon notre interprétation de ces deux métriques, ces expériences seraient probablement complètes (É-T $\geq 20\%$) et ce serait le modèle Line qui permettrait d'inférer le plus de paramètres précisément (ΔWAIC$_k > 0$). Or, cela semble être impossible considérant le contexte

expérimental des expériences dose-réponse. La tendance globale (LOESS) des réponses d'expériences associées au Groupe D (•) indique la présence d'une importante variance (Fig. 35C). Ces expériences semblent avoir des réponses variables et extrêmes. Leur potentiel informatif s'apparente à celui des expériences du Groupe C (•). Outre ce nouveau groupe, les résultats d'assignation du jeu de données gCSI correspondent aux définitions décrites précédemment (Section 4.4.2). Notre processus d'évaluation du potentiel informatif peut ainsi être appliqué à de large jeux de données, et outiller les expérimentateurs dans leur processus d'interprétation et d'analyse des *posteriors*.

4.5. BiDRA V2

Un des principaux objectifs du présent travail est de mieux outiller les expérimentateurs dans leur processus d'analyse et de prise de décisions. Pour ce faire, il est primordial de rendre les processus d'inférence et d'analyse décrit dans les Chapitres 2 et 3, ainsi que dans la précédente Section 4.4 accessibles. L'accessibilité à de tels processus se définit de différentes façons (Section 5.6). Dans le Chapitre 2, l'interface web BiDRA est présenté [`https://bidra.bioinfo.iric.ca/`]. Suite aux développements et améliorations apportés aux processus et modèle BiDRA (Chapitre 3 et Section 4.4), une deuxième version de l'interface web BiDRA est mise à disposition. Celle-ci peut être trouvée au `https://bidrav2.bioinfo.iric.ca/` et sera référée comme étant l'interface web BiDRA V2.

La première version de l'interface web BiDRA (Chapitre 2) permet l'analyse d'une expérience ainsi que l'analyse comparative de deux expériences. Les résultats sont retournés sous forme de figures et d'une table des intervalles de confiance. L'interface web BiDRA a été révisée pour mieux répondre aux besoins des expérimentateurs, tel que recommandé par des collaborateurs. L'interface web BiDRA V2 se différencie de la version originale en quatre principaux aspects: (1) le modèle BiDRA utilisé, (2) les entrées (données et paramètres) fournies par l'utilisateur, (3) le nombre d'expériences pouvant être considérés, et (4) les résultats retournés par l'interface.

Cette nouvelle version utilise le modèle BiDRA révisé et présenté au Chapitre 3, et utilisé dans la Section 4.4 du présent chapitre. L'algorithme NUT-s est utilisé sur quatre chaînes de 1,000 itérations (en plus de 1,000 itérations d'échauffement).

L'utilisateur doit fournir un fichier (`.csv`) contenant ses données et préciser le type de réponse considérée (c.-à-d. ascendante ou descendante). L'utilisateur n'a plus besoin de spécifier les *priors* à utiliser. Les résultats du Chapitre 3 démontrent bien que les *priors* du modèle BiDRA sont assez peu informatifs pour que le modèle soit généraliste. Tel que mentionné précédemment (Chapitres 2 et 3) et discuté dans la Section 5.3, de mauvais *priors* peuvent biaiser et altérer les résultats et conclusions. Le concept de *prior* peut être abstrait pour certains utilisateurs et il n'est pas toujours trivial de les définir de telle sorte à ne pas biaiser les résultats de l'inférence. De plus, l'utilisation de *priors* définis et constants assure la reproductibilité des résultats pour les utilisateurs de l'interface web BiDRA V2.

Le fichier de données fourni par l'utilisateur peut contenir plusieurs expériences. Contrairement à la première implémentation de l'interface web BiDRA, BiDRA V2 permet l'analyse de N expériences. Celles-ci doivent cependant partager le même type de réponse (c.-à-d. ascendantes ou descendantes). Les expériences sont différentiables par des identifiants (ID) uniques spécifiés par l'utilisateur. À titre indicatif, le processus d'analyse complet via l'interface web BiDRA V2, prend un peu moins de 3 minutes (Intel Xeon 6230, 4 *threads*) pour cinquante expériences ayant chacune 10 concentrations et une réponse par concentration. Notons que le temps de calcul est dépendant de la taille des expériences: sauf indication contraire, les temps mentionnés réfèrent à des expériences ayant 10 mesures (c.-à-d. 10 concentrations et une réponse par concentration).

Similairement à la première implémentation, une figure de la courbe dose-réponse (médiane et intervalle de confiance de 95%) et des *posteriors* est retournée pour chaque expérience. Les valeurs médianes ainsi que les valeurs des bornes des intervalles de confiance (95%) sont indiquées pour chaque *posterior* (Fig. 36.A). En plus de la figure, un fichier (`.csv`) contenant les valeurs des *posteriors* est aussi retourné pour chaque expérience. L'utilisateur a ainsi la possibilité et la flexibilité de mener ces propres analyses post-inférence. Pour tout jeu de données soumis, une analyse du potentiel informatif de ses expériences est faite. Chaque expérience se voit attribuer un sigle de couleur, tel qu'établi et décrit à la Section 4.4. Les résultats de cette analyse sont retournés sous forme de figure où l'ID unique d'une expérience est suivi d'un sigle de couleur (Fig. 36.B).

Pour un jeu de données considérant deux expériences, une analyse comparative similaire à celle de l'interface BiDRA originale est faite. Une figure des *posteriors* de différences est retournée, en plus d'une comparaison quantile-à-quantile (QQ) des *posteriors* (Fig. 37.A). Cette deuxième représentation aide les utilisateurs à évaluer la ressemblance entre les différents *posteriors*. La comparaison des *posteriors* par paires de métriques proposée dans la première version n'est pour le moment pas disponible. Ce résultat n'est pas nécessairement intuitif à interpréter par les utilisateurs et, bien qu'il permette d'identifier les dépendances inter-paramètres, il est très peu considéré.

Finalement, pour un jeu de données contenant au moins trois expériences, une analyse d'ordonnancement des IC_{50}/EC_{50} (Fig. 37.B) et des HDR est menée. Celle-ci est similaire à la démonstration faite sur le jeu de données de l'IRIC dans le Chapitre 3. Les IC_{50}/EC_{50} sont ordonnées de façon ascendante. Pour les HDR, l'ordonnancement est dépendant du type de réponse considérée.

L'ensemble des résultats (Figures et fichiers CSV) est contenu dans un répertoire compressé (`.zip`) qui est automatiquement téléchargé via le fureteur de l'utilisateur, une fois le processus d'analyse complété. Cette nouvelle version de l'interface BiDRA propose de nouvelles analyses post-inférence pour des jeux de données complets en plus de simplifier les interactions de l'utilisateur.

4.5.1. Implémentation

L'ensemble de l'interface BiDRA V2 a été implémenté en Julia (V1.8.3). Le modèle bayésien et l'inférence sont faits à l'aide des librairies `Turing` [115] (V0.26.0) et `MCMCChains` (V6.0.3). Les figures retournées à l'utilisateur sont générées grâce à la librairie `CairoMakie` (V0.10.6).

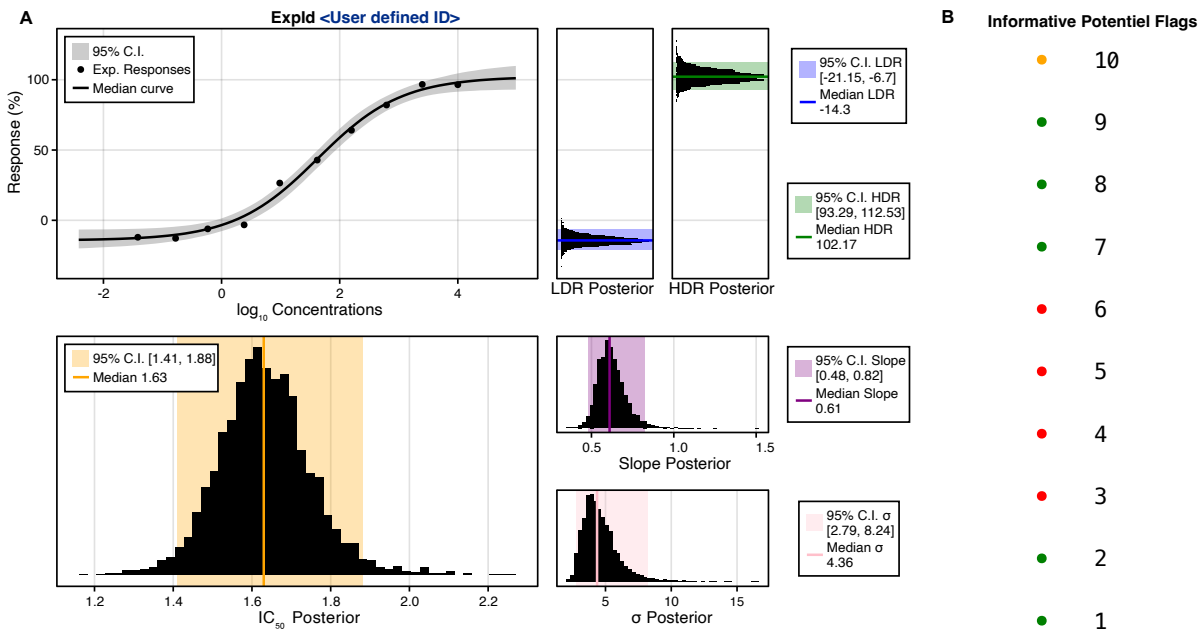


Fig. 36. Résultats types pour toutes les analyses retournées par l'interface BiDRA V2. Figures retournées à l'utilisateur pour tout jeu de données, peu importe le nombre d'expériences considérées. **(A)** Courbe médiane dose-réponse inférée et son intervalle de confiance (95%). Les *posteriors* des métriques d'efficience de base sont représentés par des histogrammes noirs. Leurs valeurs médianes et leurs intervalles de confiances (95%) sont représentées en couleur et indiquées dans la légende. Le *posterior* de la valeur σ de la fonction de vraisemblance est aussi représenté. **(B)** Association du sigle de potentiel informatif pour chaque expérience d'un jeu de données. L'exemple représenté considère 10 expériences dont les IDs sont un numéro allant de 1 à 10.

L'interface web, quant à elle, utilise les librairies `Genie` (V5.18.1) et `HTTP` (V1.9.6) pour le côté serveur, et les librairies `Stipple` et `StippleUI` pour l'interactivité. Le choix de cette infrastructure ainsi que la décision de migrer de Python à Julia entre les deux versions de l'interface sont discutés dans la Section 5.2.

4.6. Conclusion

Le présent chapitre fait suite aux travaux présentés dans les Chapitres 2 et 3 en répondant à deux besoins explicites formulés par les expérimentateurs, soit l'identification rapide du type de réponse (plate, incomplète ou complète) et l'analyse simultanée et la comparaison de plusieurs composés. Dans un premier temps, une méthode analytique permettant de catégoriser les expériences selon leur potentiel informatif est formulée (Section 4.1) puis présentée (Section 4.4). Cette catégorisation combine l'évaluation de la complétude de la réponse considérée (SD) et la capacité du modèle sigmoïde à représenter cette même réponse (ΔWAIC_k). L'approche présentée se veut une alternative automatique (via l'interface web BiDRA) et quantitative à l'évaluation visuelle des réponses par un expérimentateur. Cela assure l'uniformité et la réplicabilité (Section 4.4.2)

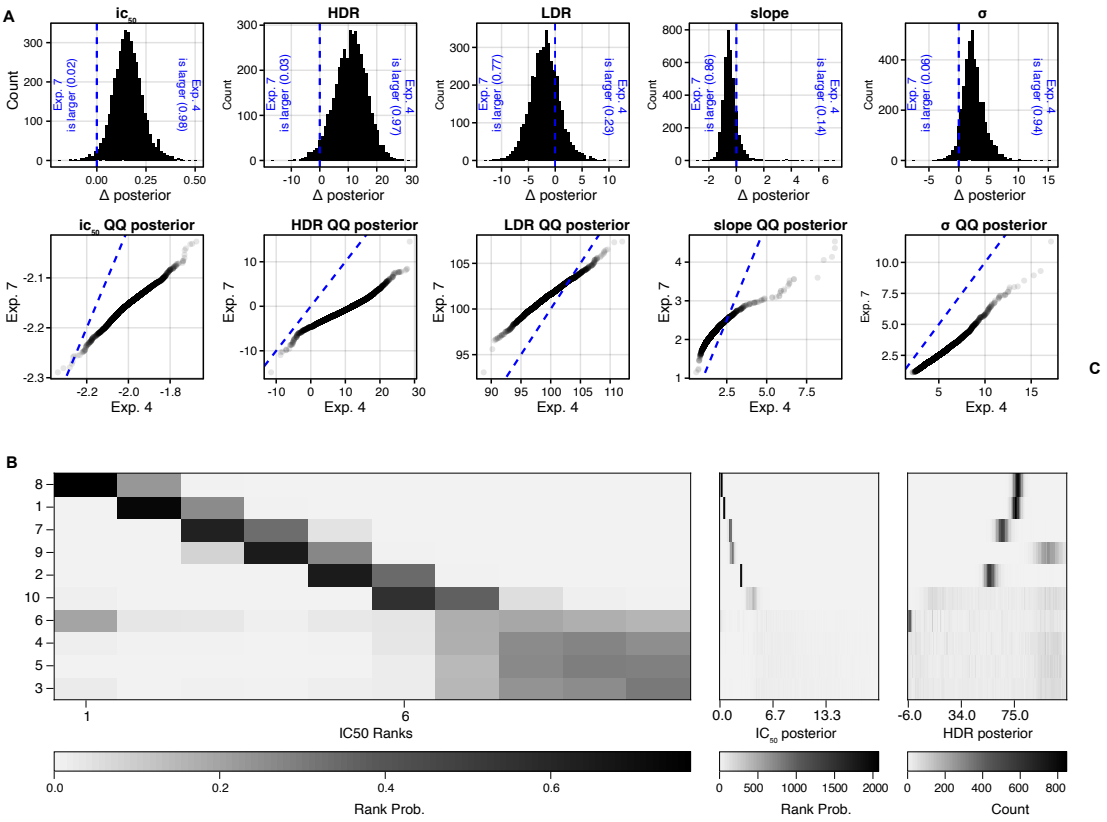


Fig. 37. Résultats types pour les analyses de jeux de données contenant plus d'une expérience retournés par l'interface BiDRA V2. Figures retournées à l'utilisateur pour tout jeux de données comportant deux ou au moins trois expériences. **(A)** Résultats de l'analyse comparative de deux expériences. La premère rangée de graphiques représentent les *posteriors* des différences pour les métriques d'efficience et le paramètre σ. La ligne de base 0.0 est indiquée d'un très bleu haché. Les probabilités (en pourcentage) qu'une expérience ait une métrique supérieure à celle de l'autre expérience sont indiquée en bleu, et ce pour chaque métrique. La deuxième rangée de graphiques représente les comparisons des *posteriors* QQ pour les métriques d'efficience et le paramètre σ. La diagonale d'identité (x = y) est représentée d'un trait bleu haché. **(B)** Résultats de l'analyse d'ordonnancement pour au moins deux expériences. L'exemple représenté considère 10 expériences dont les IDs sont un numéro allant de 1 à 10. De gauche à droite: *heatmap* représentant les probabilités d'ordonnancement des IC_{50}/EC_{50} de chaque expérience; *heatmap* des *posteriors* des IC_{50}/EC_{50} ; *heatmap* des *posteriors* des HDRs.

du processus de catégorisation d'expériences, en plus d'être efficace lorsque plusieurs expériences sont considérées (Section 4.4.4). Bien que les *posteriors* soient des représentations informatives des métriques d'efficience, leur interprétation n'est pas toujours triviale : la catégorisation des expériences selon leur réponse (c.-à-d. plate, incomplète ou incomplète) informe et guide les expérimentateurs dans leur interprétation. Dans un deuxième temps, une version revisitée de l'interface web BiDRA est présentée (Section 4.5). Cette nouvelle version propose deux principaux avantages sur la version originale : (1) l'analyse simultannée de plusieurs expériences et (2) de nouvelles analyses post-inférence, dont l'ordonnancement de *posteriors* (Section 3.2.6) et la catégorisation du potentiel informatif tel que présenté dans le présent chapitre. Cette nouvelle version facilite

l'intégration des *posteriors*dans un processus d'analyse plus générale, et ce de façon automatique sans que l'expérimentateur ait à manipuler lui-même les *posteriors*.

La catégorisation et l'interface web présentées dans ce chapitre sont deux outils qui facilitent l'accessibilité des méthodes présentées aux Chapitres 2 et 3 tant en termes d'utilisation (appliquer le modèle bayésien à des données spécifiques) qu'en termes d'interprétation des résultats et de formulation de conclusions valides.

Chapitre 5

Discussion

Le but de la présente thèse est de mieux outiller les expérimentateurs dans leur processus analytique et de prise de décisions en appliquant l'inférence bayésienne à la caractérisation de l'efficience de composés chimiques. Pour ce faire, un modèle bayésien a été dans un premier temps mis en place (Chapitre 1) puis révisé (Chapitre 3). Le modèle proposé est hautement généralisable et peut être appliqué à diverses expériences. Nous avons par la suite démontré de façon quantitative la robustesse du-dit modèle en comparaison à l'approche standard (Levenberg-Marquardt) (Chapitre 3) : la représentation des métriques d'efficience par *posteriors* est plus informative et représentative de la réponse expérimentale sous-jacente à l'analyse. De plus, elle est une représentation explicite de l'incertitude entourant les métriques. L'application de notre modèle bayésien permet donc de faire une meilleure caractérisation de l'efficience d'un composé. De par leur caractère informatif, les *posteriors* peuvent être utilisés dans divers processus post-inférence pour répondre à des questions expérimentales précises (Chapitres 3 et 4). Cela est particulièrement utile dans le contexte de sélection de composés. Les conclusions sont statistiquement valides et tiennent compte de l'incertitude biologique, comme analytique. Finalement, l'ensemble des méthodologies présentées sont intégrées dans une interface web, BiDRA (Chapitres 1 et 4), simplifiant ainsi leur accessibilité. Un expérimentateur peut donc facilement inférer et analyser des métriques d'efficience pour un nombre de composés d'intérêt, et tirer profit de la robustesse de notre modèle bayésien et de l'informativité de ses *posteriors*.

Le présent chapitre propose des discussions sur divers thématiques abordées dans les Chapitres 2, 3 et 4. Ces thématiques couvrent le choix des jeux de données (Section 5.1), l'implémentation du processus d'inférence (Section 5.2), le choix (et ses implications) des *priors* (Section 5.3), la quantification de l'incertitude (Section 5.4), l'évaluation du potentiel informatif (Section 5.5) et l'accessibilité (Section 5.6). Les implications des travaux de la présente thèse ainsi que les perspectives futures sont abordées dans la conclusion (Section 5.7).

5.1. Le choix des jeux de données

Nous faisons usage de données synthétiques (c.-à-d. des données non-expérimentales générées de façon artificielle, Chapitre 2) et expérimentales (Chapitres 2, 3 et 4). Ces premières, provenant

d'un environnement contrôlé, sont utilisées comme contrôles positifs dans le but de confirmer et de démontrer la validité de la méthode proposée [123, 129, 134]: la vérité nous étant connue (e.g. Table 1), nous pouvons la comparer avec les résultats obtenus. Les données expérimentales sont quant à elles utilisées dans le but de démontrer l'applicabilité de la méthode dans un réel contexte.

Les données synthétiques du Chapitre 2 considèrent deux caractéristiques changeantes, soit la variabilité de la réponse et les valeurs de métriques d'efficience (e.g. IC_{50}/EC_{50} et HDR). La première caractéristique nous a permis de mettre en contexte le *posterior* du σ de la fonction de vraisemblance (Fig. 11). La deuxième caractéristique nous a permis de démontrer l'importance de bien définir les *priors* (Fig. 8 et 9). La combinaison des deux caractéristiques nous a permis de mettre de l'avant les gains analytiques de l'utilisation d'une approche bayésienne, en comparaison de la méthode par régression non-linéaire (Levenberg-Marquardt). En effet, cette dernière est limitée par la variance de la réponse et la forme de la courbe (spécialement lorsque certains paramètres du modèle log-logistique sont constants, Fig. 7). Finalement, l'utilisation de données synthétique représente un important outil pour faciliter la compréhension et l'interprétation des résultats obtenus, puisqu'il est facile de comparer ceux-ci à la "vérité" (e.g. la comparaison de deux expériences dans la Figure 12). Cela permet, entre autres, aux expérimentateurs d'avoir confiance en la nouvelle méthodologie proposée, dans ce cas-ci, notre modèle bayésien BiDRA. Les données expérimentales utilisées dans le Chapitre 2 solidifient cette confiance en démontrant que BiDRA est tout aussi applicable à des données expérimentales (Fig. 10 et 13).

Dans les Chapitres 3 et 4, les données expérimentales furent utilisées pour quatre principaux objectifs: (1) démontrer l'applicabilité des méthodes proposées à un grand nombre d'expériences de diverses sources, (2) démontrer la flexibilité et la robustesse du processus analytique pour différents types de réponse (c.-à-d. avec complétude de niveaux variés), (3) démontrer l'utilité des *posteriors* pour répondre à des questions expérimentales lors d'analyses post-inférence, et (4) démontrer quantitativement les lacunes et limitations de l'approche par régression non-linéaire (Levenberg-Marquardt).

Le jeu de données interne de l'IRIC a été obtenu au travers de collaborations avec la Plateforme de chimie médicinale et fut principalement utilisé pour répondre au troisième objectif. Cette collaboration nous a permis de définir des questions expérimentales et d'ainsi établir des méthodologies d'analyse post-inférence permettant d'y répondre. De plus, la taille du jeu de données facilite les phases d'exploration lors de la mise en place d'un processus analytique. Le processus est ensuite validé sur un plus large jeu de données (Chapitre 4).

Les trois jeux de données publics utilisés dans le Chapitre 3 furent obtenus grâce à l'outil PharmacoGx [55] et la banque de données PharmacoDB [60, 75]. Seules les réponses normalisées (% de viabilité cellulaire) furent récupérées. Nous avons fait abstraction des processus de manipulation de données (e.g. plafonnement des réponses à des valeurs maximale et minimale) et d'estimation des métriques d'inférence (c.-à-d. selon le modèle log-logistique à 3 paramètres) disponibles. Premièrement, la manipulation de données (c.-à-d. modifier ou retirer des valeurs) modifient les observations et génèrent des métriques d'efficience artificielle pouvant être erronées. Et deuxièmement, nous souhaitions démontrer la robustesse de BiDRA dans un contexte expérimental optimal

sans manipulation des données. Nous avons donc calculé les métriques d'efficience par régression non-linéaire (Levenberg-Marquardt) sur les données brutes. Les jeux ont été sélectionnés de telle sorte à atteindre les objectifs un (1), deux (2) et quatre (4), mentionnés ci-haut. J'ai utilisé trois jeux de données différents pour corroborer les résultats obtenus et faire des démonstrations globales.

CTRPv2 [63, 73, 74] est le plus large jeux de données d'expériences dose-réponse faites sur des lignées cellulaires cancéreuses (CCl) [73]. gCSI fut quant à lui conçu spécialement pour investiguer de façon indépendante [66, 72] les divergences entre les jeux de données CCLE et GDSC [22]. Ce jeu possède la plus haute corrélation entre les réponses de réplicats biologiques (Fig. 15), et malgré cela, BiDRA est nettement supérieur à Levenberg-Marqaurdt lorsque les métriques d'efficience de ces mêmes réplicats sont comparées (Fig. 16). Gray [69, 149] est d'intérêt car il contient plusieurs expériences ayant des réponses extrêmes (allant jusqu'à 200% de viabilité cellulaire) démontrant ainsi la flexibilité de BiDRA. De plus, son protocole expérimental diffère légèrement de celui de CTRPv2 et gCSI, et considère plusieurs réponses par concentration. Cela est intéressant pour notre analyse comparative des méthodologies, car les résultats obtenus sont très similaires à ceux obtenus pour CTRPv2 et gCSI (Fig. 16).

Les trois jeux de données sélectionnés présentent aussi une bonne variété d'expériences en termes de la variabilité et de la complétude des réponses expérimentales. De plus, leurs couvertures variables de CCL et de composés (Fig. 14.A) démontrent bien la généralisation du modèle BiDRA proposé.

Bien que ces trois jeux de données aient permis d'atteindre les objectifs établis, il serait potentiellement intéressant de reproduire les analyses comparatives de réplicats biologiques sur d'autres larges jeux données tels que NCI60 [170–172], CCLE [41] et GDSC [70, 173]. Cela solidifierait davantage la démonstration du gain analytique lorsque l'on utilise une approche bayésienne pour l'inférence de métriques d'efficiences.

Finalement, dans l'optique détendre la démonstration de robustesse et d'application de BiDRA et de ses *posteriors* (Section 5.7), d'autres jeux de données pourraient être utilisés. Par exemple, les jeux CCLE et GDSC seraient pertinents pour l'évaluation des divergences entre jeux de données puisqu'ils ont un grand nombre d'expériences en commun et ce sont les jeux de données utilisés lors de l'analyse de divergence originale [22]. Il serait aussi intéressant de mener des analyses de divergences sur les jeux de données de LINCS [19] et FIMM [67], puisque bien que ces jeux soient relativement petits, ils ont été conçus spécialement pour l'évaluation et la quantification des variances expérimentales (e.g. entre différents centres de recherche, entre différents expérimentateurs) et biologiques. De façon similaire à gCSI, ces jeux de données constituent de bonnes bases pour démontrer la robustesse de BiDRA. Un autre exemple est le développement de modèles d'apprentissage automatique et l'intégration des *posteriors*. Dans ce contexte, de larges jeux de données tel que le NCI60 ou le CTRPv2 sont nécessaires. De plus, une représentation des composées doit être accessible. Le type de représentation le plus communément utilisé est le SMILES [174] (de l'anglais *simplified molecular-input line entry system*), soit un descriptif de la structure moléculaire présenté sous la forme de chaîne de caractères. Les SMILES ne sont pas directement accessibles sur toutes les plateformes et il n'est pas toujours trivial de faire une association entre un identifiant (e.g. nom

d'un composé) et un SMILES: il arrive souvent que cette étape de curation (automatique comme manuelle) diminue grandement le nombre d'expériences disponibles. Le jeu de données NCI60 est associé à une liste d'identifiant (c.-à-d. SID de la banque de données PubCHEM) pour les structures de chaque composé, facilitant ainsi l'accès à ces données. Le jeu de NCI60 serait le plus approprié dans un contexte de développement d'approches par apprentissage machine.

5.2. L'implémentation de BiDRA

Tel que présenté dans le Chapitre 1, il existe divers algorithmes pour simuler l'échantillonnage de *posterior* par *Markov Chain Monte Carlo* (MCMC). Des principaux algorithmes (Section 1.3.2), le NUT-S a été sélectionné pour sa rapidité, généralité et efficacité. Nous avons très peu considéré l'échantillonneur de Gibbs, car son efficacité est grandement affectée lorsque les paramètres du modèle sont corrélés, comme cela peut être le cas dans notre contexte d'analyse. Les échantillonneurs Metropolis et HMC demandent une certaine précision dans la définition de leurs paramètres de réglage (c.-à-d. matrice de covariance pour Metropolis et le *learning rate* et le nombre de pas pour le HMC) sans quoi la convergence est ralentie. Pour optimiser le processus d'inférence, ces paramètres doivent être expérience-spécifique. Or, cela est peu souhaitable dans le contexte dans lequel nous souhaitons appliquer une même approche à diverses expériences. Nous avons observé que les *posterior* de Metropolis, HMC et NUT-S étaient comparables et que le choix de l'algorithme affectait principalement l'efficacité de l'inférence en termes du nombre d'itérations et de temps de calculs. Ce sont pour ces raisons que l'algorithme NUT-S a été sélectionné pour échantillonner les *posteriors* de BiDRA.

Nous avons optimisé le modèle BiDRA et ses *priors* en considérant les métriques de diagnostic standard, soit le PSRF et le taux de divergence. Les *priors* de notre premier modèle BiDRA (Chapitre 2) étaient plus contraignants que ceux de notre modèle revisité (Chapitre 3). Cela avait pour effet d'augmenter le nombre d'itérations divergentes dans le cas de réponses incomplètes ou plates. Dans notre contexte d'analyse, un haut taux de divergence peut être informatif et n'est pas nécessairement associé à un mauvais modèle. Premièrement, ce ne sont pas toutes les expériences qui mènent à un haut taux de divergence. Deuxièmement, nous sommes tout de même limités dans la définition de notre modèle, car nous souhaitons modéliser des réponses cellulaires selon le modèle log-logistique. Et troisièmement, le taux de divergence est informatif quant à la confiance que l'on devrait avoir envers nos *posteriors*. La métrique du taux de divergence est cependant peu intuitive à interpréter et nous avons tout de même révisé nos *priors*. Les *priors* du Chapitre 3 sont moins contraignants et diminuent considérablement le taux d'itérations divergentes.

Le nombre de chaînes minimales recommandé, soit quatre, est utilisé pour faciliter la manipulation post-inférence. Pour cette même raison, le nombre d'itérations total par chaîne est 2000, dont 1000 utilisées pour l'échauffement. Les chaînes obtenues se mixtent globalement bien (PSFR) et sont suffisamment larges pour former une bonne représentation des *posteriors*, tout en facilitant leur manipulation pour les étapes d'analyse post-inférence.

L'algorithme NUT-S utilisé pour les analyses et applications web présentées aux Chapitres 2, 3 et 4 est implémenté au travers d'outils de programmation probabiliste (PP) existant. Dans le cas du Chapitre 2, Stan [114] fut utilisé au travers de la librairie Python `pyStan` [117], et pour les Chapitres 3 et 3, `Turing.jl` [115] fut utilisé.

Le choix d'utiliser un outil PP plutôt qu'une implémentation dite maison fut confirmé par des travaux entrepris en collaboration avec Cameo Sameshima (étudiante stagiaire, 2019). Les trois algorithmes mentionnés plus haut furent implémentés (C/C++) et leurs efficacités comparées à celles des implémentations de Stan/`pyStan`. Bien qu'une implémentation maison permette plus de flexibilité (par exemple, dans les données retournées), son optimisation est la principale limitation. Le temps de calcul des implémentations maison était supérieur à celui de Stan/`pyStan`, bien que les résultats étaient comparables. Un important travail de validation et d'optimisation (implémentation, stabilité numérique et rapidité) serait nécessaire pour pouvoir utiliser ces implémentations maison dans le processus d'inférence et d'analyse proposé dans le présent travail. N'ayant apporté aucune modification aux algorithmes, l'implémentation maison deviendrait alors équivalente à celle disponible depuis un outil PP, justifiant ainsi la simplicité à utiliser des outils tel que Stan/`pyStan`. Dans les prochains paragraphes, nous discuterons des choix d'outils PP utilisés, Stan/`pyStan` ainsi que `Turing.jl`.

Stan [114] est un langage de programmation probabiliste (LPP) établi et développé par Andrew Gelman, une référence dans le domaine de l'inférence bayésienne. Plusieurs ressources [90] se basent sur ce LPP pour présenter des concepts et des exemples. La décision d'utiliser ce LPP était logique, et simplifiait grandement le processus d'apprentissage par la quantité de ressources et de documentations disponibles [90, 91, 117]. BUGS [116] et JAGS [112] sont des LPPs alternatifs à Stan (Section 1.3.3). Stan se différencie de ces deux derniers en étant un langage plus flexible en permettant de déclarer le type des variables, en intégrant l'utilisation de variables locales et de syntaxe conditionnelle [90, 114]. Les échantillonneurs MCMC de Stan sont construits sur la base du *Hamiltonian Monte Carlo* (HMC et NUT-S) qui est plus robuste que les échantillonneurs de Gibbs et Metropolis utilisés par BUGS and JAGS [91, 114]. Finalement, le LPP Stan peut facilement être intégré aux codes d'autres langages de programmation plus communs via des libraires telles que `pyStan` et `RStan` [117]. Ce dernier avantage facilite l'intégration de l'inférence au processus d'analyse d'expérience dose-réponse complet qui se faisait initialement en Python. Une alternative à Stan/`pyStan` est PyMC3 [113], qui est entièrement en Python. PyMC3 utilise aussi les algorithmes HMC et NUT-S et semble être aussi efficace que Stan/`pyStan`. Le choix d'utiliser Stan/`pyStan` s'est fait sur la base des ressources et documentations disponibles, tant pour la théorie que pour l'application.

L'inférence est relativement rapide via Stan/`pyStan`. Par exemple, le temps d'inférence pour les Analogues 1 et 4 du Chapitre 4 est de 0.5 et 0.8 seconde, respectivement (Intel i9-7920X, 4 *threads*). L'utilisation de Stan/`pyStan` comporte aussi quelques limitations. `pyStan` est peu documenté en comparaison à Stan, et il est parfois difficile d'accéder à toutes les fonctionnalités de Stan. De plus, définir des *priors* personnalisés (tel que la mixture décrite dans le Chapitre 3) est complexe et il n'est pas trivial de valider la distribution créée. Finalement, il est impossible d'inférer directement

le *posterior* d'un paramètre discret et d'inférer via un échantillonneur composé, deux avenues explorées lors de la mise en place du processus d'évaluation du potentiel informationnel décrit dans le Chapitre 4.

Les limitations décrites ci-haut ont motivé la transition vers la librairie `Turing.jl`, et plus globalement, vers le langage de programmation Julia [49]. Julia est un jeune (2012) langage de programmation *open source* dynamique facilitant les analyses numériques. `Turing.jl` est une librairie permettant la PP. La syntaxe utilisée est intuitive et permet l'intégration de tout code Julia (natif comme provenant d'autres librairies) dans la déclaration du modèle. Dans ce contexte, la personnalisation de *priors* devient triviale (e.g. mixture pour le *prior* du HDR). Le processus d'inférence est efficace, car il se base lui aussi sur les échantillonneurs HMC (HMC et NUT-S). `Turing.jl` propose aussi l'inférence de variables discrètes via un échantillonneur de type *particle MCMC* [120, 121], et la combinaison d'échantillonneurs. Cette dernière fonctionnalité permet d'inférer simultanément les variables discrètes (*particle MCMC*) et continues (HMC) d'un même modèle. Cette approche fut testée de façon infructueuse, dans le contexte expérimental du présent travail. Cette expérience est brièvement discutée dans la Section 5.5.

Les limitations principales rencontrées lors de l'utilisation de `Turing.jl` sont principalement liées à la nouveauté du langage et de la librairie: la documentation peut être incomplète et dépassée. Cependant, l'accès au code source est simple, et celui-ci est en pur Julia.

L'inférence avec `Turing.jl` est plus lente qu'avec Stan/`pyStan`. Par exemple, le temps d'inférence pour les Analogues 1 et 4 (Chapitre 4) est de 1.67 et 2.20 secondes, respectivement (Intel i9-7920X, 4 *threads*). L'ajout de l'inférence pour le modèle constant (Section 4.3) est minime, soit 0.3 et 0.06 seconde pour les Analogues 1 et 4 respectivement (Intel i9-7920X, 4 *threads*). Les temps de calculs rapportés ne comprennent pas les premières compilations du modèle pour Stan/`pyStan`, et du code pour `Turing.jl`. Tels qu'attendus, les résultats sont les mêmes entre les deux outils, pour un même modèle. Les *posteriors* se différencient cependant entre les versions de BiDRA (Chapitres 2 et 3) puisque les *priors* ne sont pas les mêmes, notamment ceux du HDR. Cela affecte principalement les expériences aux réponses incomplètes ou plates, et ce sont principalement les *posteriors* des HDR et IC_{50}/EC_{50} qui diffèrent entre les deux implémentations.

Le choix d'implémenter la deuxième version de BiDRA (Chapitres 3 et 4) et du processus d'analyse en Julia s'est fait pour deux raisons. Premièrement, la flexibilité de `Turing.jl` lors de la déclaration du modèle facilite l'utilisation de *priors* personnalisés telle la mixture utilisée pour le HDR. Ce *prior* semble en effet mieux représenter les valeurs HDR. Deuxièmement, les analyses post-inférence ainsi que la gestion des données et *posteriors* sont optimisés en Julia. L'inférence même est plus lente, mais un gain en temps de calcul global est observé pour le processus complet (c.-à-d. génération de figures, comparaison de deux expériences, calculs des *posteriors* de rangs et du potentiel informatif). Par exemple, l'analyse d'une expérience depuis l'interface web BiDRA (Chapitre 2, [146]) prend environ 6 secondes, comparé à 4 secondes pour l'analyse de la même expérience depuis l'interface web BiDRA V2 (Chapitre 4) (Intel Xeon 6230, 4 *threads*). La différence est encore plus importante pour l'analyse de deux expériences, soit approximativement 20 et 7

secondes pour les interfaces web BiDRA et BiDRA V2 respectivement. Le gain en rapidité est tout aussi intéressant lorsque nous considérons que l'interface web BiDRA V2 retourne une analyse du potentiel informatif des *posteriors*. À titre informatif, l'analyse de 10 et 50 expériences depuis BiDRA V2 prend approximativement 50 secondes et 3 minutes (Intel Xeon 6230, 4 *threads*), et inclut les analyses d'ordonnancement des HDR et IC_{50}/EC_{50} . Notons que le temps de calcul est dépendant de la taille des expériences: les temps mentionnés si haut référent à des expériences ayant 10 mesures (c.-à-d. 10 concentrations et une réponse par concentration). Les analyses présentées au Chapitre 3 ont été faites en considérant un important nombre d'expériences. Par exemple, 1 580.8E6 valeurs sont considérées pour l'analyse des *posteriors* d'une métrique pour le jeu de données complet CTRPv2 (Fig. 26.B). Bien que ces analyses n'aient jamais été faites à une si grande échelle en Python, j'anticipe que cela pourrait être problématique, notamment pour la génération des figures et la manipulation des données en tant que telles (e.g. création et manipulation des *dataframes*).

Notons que l'interface web BiDRA V2 ne devrait pas être utilisée pour l'analyse de larges jeux de données (> 50). L'utilisation du code publique est à favoriser. Pour un jeux de données, un *script* `Bash` lance le processus d'inférence pour un nombre b de *batches* (c.-à-d. divers sous-groupes d'expériences) parallèles. La valeur de b est dépendante de l'infrastructure computationnelle disponible. À titre indicatif, l'inférence des *posteriors* de Gray, gCSI et CTRPv2 tels qu'utilisés dans les analyses des Chapitres 3 et 4 ont respectivement pris 2.10, 2.54 et 38.75 heures ($b = 20$ et Intel Xeon 6230).

La plus récente implémentation de BiDRA et de son interface web (Chapitre 4) en Julia permet de maximiser les analyses post-inférence tout en minimisant le temps d'attente d'un utilisateur.

5.3. Le choix et les implications des *priors*

Les *priors* sont à la base l'inférence bayésienne et représentent l'incertitude « pré-observations » des paramètres d'un modèle. Le théorème de Bayes (Éq. 4) infère le *posterior* d'un paramètre au fur et à mesure que de nouvelles données sont observées: les *priors* sont en quelque sorte notre point de départ et leur choix peut grandement impacter l'inférence [175].

Nous démontrons dans le Chapitre 2 de façon drastique les biais que peuvent introduire de « mauvais » *priors* (Fig. 8). Un mauvais *prior* est ici défini comme étant trop informatif: l'information qu'il véhicule en lien avec un paramètre donné est précise et contraignante à une solution. Il est pratique courante de décrire un *prior* en termes de son caractère informatif. Un *prior* hautement informatif est généralement associé à une quantification subjective de l'incertitude tandis qu'un *prior* non informatif est généralement plus objectif [91]. Le caractère informatif d'un *prior* peut aussi se situer entre ces deux extrêmes, comme nous verrons plus loin.

Bien que l'intégration de *priors* au processus analytique ajoute implicitement une dimension subjective à celui-ci, nous souhaitons favoriser des *priors* qui tendent vers l'objectivité. Nous ne souhaitons pas « donner la réponse » au processus d'inférence, car nous ne connaissons pas la « vraie » réponse. Nous souhaitons inférer les paramètres d'un modèle (dans le contexte précis de

cette thèse, les métriques d'efficience) qui sont représentatifs des données observées. L'exemple de la Figure 8 démontre bien la problématique à définir des *priors* hautement informatifs, et dans ce cas-ci, biaiser le processus d'inférence vers une réponse erronée et peu supportée par les données. Lorsque des *priors* non informatifs sont définis, la forme des *posteriors* relève principalement de la fonction de vraisemblance [175]. Les observations pèsent alors plus que les *priors*, et sont mieux représentées par les *posteriors* (Fig. 9).

La distribution uniforme est généralement utilisée pour décrire un *prior* non informatif [91, 175]. Ce type de *priors* est comparé à des *priors* normaux dans le Chapitre 2 (Fig. 9). Bien que d'autres modèles bayésiens utilisent des *priors* uniformes [145, 147], nous avons décidé d'utiliser des *priors* normaux. Les travaux de [145, 147] sont présentés pour des contextes et des jeux de données précis. Or, nous avons comme objectif de mettre en place un modèle bayésien généraliste et applicable à plusieurs expériences provenant de divers jeux de données. Il devient alors difficile de déterminer des limites (α et β) pour des différents *priors* uniformes. Par exemple, les bruits biologique et expérimental résultent à des réponses excédant la plage des valeurs attendues (c.-à-d. [0%, 100%]) ce qui affecte aussi les valeurs des LDR et HDR. De façon similaire, il est difficile de déterminer la limite supérieure (β) pour le *prior* de l'IC_{50}/EC_{50} . Une solution serait d'utiliser des limites extrêmes ou de déterminer des limites qui sont spécifiques à l'expérience. Dans le premier cas, il semble incorrect d'assumer que toutes les valeurs d'IC_{50}/EC_{50} , même celles se rapprochant des limites extrêmes, aient la même probabilité. Dans le deuxième cas, dériver un *prior* directement des observations augmentent la subjectivité de celui-ci et peut biaiser le résultat. Or, tel que démontré dans le Chapitre 3, les données ne supportent pas nécessairement l'inférence de l'ensemble des métriques d'efficience, et certaines d'entre elles restent non observables pour une expérience. Les expériences ayant une réponse plate en sont un bon exemple: les IC_{50}/EC_{50} , HDR et pente ne sont pas observés expérimentalement. C'est d'ailleurs pour ces raisons que le *prior* de l'IC_{50}/EC_{50} a été réévalué entre les modèles des Chapitres 2 et 3.

Nous avons donc défini des *priors* continus et principalement normaux. Notons que le *prior* normal sur l'IC_{50}/EC_{50} considère les concentrations $\log_{10}$-transformées et ne donne ainsi aucun poids à des valeurs négatives (les valeurs seront au minimum infiniment petites). La distribution normale nous permet de définir des *priors* flexibles et ne nécessite pas l'utilisation de valeurs extrêmes. De plus, une étude portant sur le choix des *priors* dans un contexte de dose-réponse a démontré que les résultats obtenus étaient similaires et comparables pour plusieurs *priors* testés (gamme, uniforme et normal). Les auteurs ont identifié l'étroitesse (de l'anglais, *tighness*) des *priors* comme étant le facteur limitant principalement le processus d'inférence [125]. Cela concorde avec notre analyse des *priors* en termes de caractère informatif (Fig. 9).

Le caractère informatif d'un *prior* varie de un à l'autre. Considérons la plus récente version du modèle BiDRA, soit celle présentée au Chapitre 3. Le *prior* du LDR est relativement informatif puisque la vraie valeur de cette métrique est en quelque sorte connue: elle représente la réponse basale, soit la réponse en absence de composé. Peu importe le type de réponse, les données supportent l'inférence de ce paramètre. Il y a très peu de variance entre les valeurs de LDR d'une expérience à l'autre (Fig. 15.C). C'est d'ailleurs pour cette raison que nous observons de basses correlations entre

les LDR d'expériences répliquées, et ce, peu importe le type de représentation (Supp. Fig. 21). Inversement, le *prior* du HDR est lui très peu informatif. La mixture utilisée fut définie en combinant les intuitions et attentes de divers collaborateurs (bio-informaticiens, chimistes médicinaux, biologistes) [143], puis validée par évaluation visuelle des réponses aux concentrations maximales de 421,405 expériences (Fig. 14.C). Les *priors* de la pente et de l'écart-type (vraisemblance) sont tous deux des distributions log-normales. Cela assure que ces variables soient positives (Hennessey et al., 2010). Finalement, l'IC_{50}/EC_{50} a un *prior* très peu informatif de telle sorte que l'inférence de ce paramètre relève principalement des données. Celui-ci permet aussi de considérer plusieurs expériences peu importe leur gamme de concentrations expérimentales: le *prior* étant tellement large, la partie couvrant les concentrations expérimentales se rapproche d'une distribution uniforme (Fig. 14).

Bien qu'il serait possible de définir d'autres *priors* (e.g. Uniforme, Gamma), les *priors* décrits ci-haut et dans le Chapitre 3 représentent adéquatement notre incertitude pré-observation pour chacun des métriques d'efficience. L'intuition des expérimentateurs (*prior elicitation* [143]) est intégrée dans le processus d'inférence de telle sorte à ne pas biaiser les résultats. L'analyse contrôle des corrélations de métriques d'efficience (Fig. 18) démontre bien le rôle et l'impact des *priors*: les paires de réplicats biologiques ont toujours un coefficient de corrélation supérieur aux paires aléatoires, et ce, indépendamment de la complétude des réponses. Cela illustre bien le fait que le poids donné aux observations excède celui des *priors*, même dans les cas incertains (e.g. réponse incomplète, SD < 20) où l'inférence retombe partiellement sur ces-derniers.

Revenons brièvement sur le sujet de la subjectivité du processus analytique, la critique principale des approches bayésienne [91]. Les *priors* confèrent une dimension subjective à l'inférence des métriques d'efficience, et tel que démontré dans le Chapitre 2 cela peut facilement biaiser les résultats. Or, nous pouvons argumenter que cette subjectivité est explicite et transparente [175] contrairement aux approches fréquentistes dont la subjectivité est implicite à travers les différentes suppositions (de l'anglais *assumptions*) faites. Par exemple, lorsque le modèle log-logistique à trois ou deux paramètres est utilisé, nous assumons des valeurs constantes pour un ou les plateaux (LDR/HDR, 0% ou 100%). Ce choix est hautement subjectif et n'est pas toujours représentatif des données. Fixer les paramètres du modèle à des valeurs constantes est comparable à définir un *prior* hautement informatif (Fig. 8). Il est donc justifiable d'utiliser une approche bayésienne et ainsi minimiser les effets néfastes d'une trop grande subjectivité en définissant des *priors* non ou peu informatifs. Ces contraintes souples permettent, entrent autres, de recentrer l'inférence dans le contexte expérimental, ce qui est impossible à faire via Levenberg-Marquardt.

5.4. La quantification de l'incertitude

Le développement du modèle et la mise en place du processus BiDRA découlent de la difficulté à évaluer et quantifier l'incertitude des métriques d'efficience estimées avec l'approche standard (Levenberg-Marquardt). Or, il est important pour un expérimentateur d'évaluer adéquatement sa confiance dans les résultats obtenus pour en faire une analyse complète et ainsi tirer des conclusions

valables. Diverses sources de variabilités biologique et expérimentale peuvent affecter le niveau d'incertitude des métriques d'efficience. Il y a, entre autres, les inévitables variabilité systémique et aléatoire du processus expérimental, et les variabilités biologiques des cellules et composés utilisés [19, 147]. De plus, l'information disponible pour inférer ou estimer les métriques d'efficience est dépendante du nombre de réponses mesurées. Les métriques d'efficience et leur précision sont affectées par la qualité et la quantité d'information véhiculée par les réponses expérimentales, d'où l'importance de considérer leur incertitude lors du processus d'analyse. Tel qu'expliqué dans le Chapitre 1, il est impossible de quantifier de façon précise l'incertitude des estimations de Levenberg-Marquardt [77, 147]. L'incertitude est souvent décrite par la qualité de l'ajustement (de l'anglais, *goodness of the fit*) ou approximée par le calcul d'erreur standard ou l'application d'un ré-échantillonnage de type (Bootstrap) [31]. Des méthodes sont aussi proposées pour corriger la variabilité expérimentale via un prétraitement des données [39, 55, 78, 79]. Cette alternative amplifie l'incertitude en introduisant un bruit analytique.

L'analyse des corrélations entre les réponses de réplicats biologiques illustrent bien les variations expérimentale et biologique (Fig. 15). Malgré cela, nos résultats démontrent que les expériences d'un même jeu de données sont globalement bien répliquées (Fig. 15.D). Peu de travaux évaluent l'incertitude à même les réponses expérimentales [76], bien que cette analyse soit informative de l'impact du processus analytique [19]. Les différences dans les concentrations expérimentales limitent cependant une telle évaluation. Ces différences diminuent le nombre de réponses comparables, et l'évaluation de la corrélation peut être sur-représentative, comme sous-représentative, de la réalité du jeu de données. Par exemple, 66.14% des paires de réplicats du jeu gCSI sont considérées (33.86% des paires n'ayant aucune concentration commune). Ces différences peuvent être expliquées par le fait qu'une expérience est parfois répliquée pour ajuster la gamme des concentrations expérimentales et ainsi optimiser la représentation de la réponse globale. Les jeux Gray et CTRPv2 considèrent 85.04% et 99.70%, respectivement, des paires de réplicats totales. Leurs corrélations (Fig. 15.D) sont plus représentatives des jeux globaux. Malgré les différences de concentration entre réplicats, il est raisonnable de s'attendre à un bon niveau de corrélation entre les métriques d'efficience calculées depuis les réponses expérimentales qui sont globalement concordantes.

L'analyse des réponses expérimentales faite au Chapitre 3 a l'avantage d'être découplée du processus analytique duquel les métriques d'efficience sont dérivées. Le niveau de concordance par paire de réplicats est représenté par la RMSΔ, plutôt que par l'aire entre deux courbes doses-réponse [86]. Cette approche permet de considérer principalement l'effet des variabilités biologique et expérimentale, bien que le processus de normalisation contribue à l'ajout d'un bruit analytique. Il serait intéressant de comparer les réponses brutes (e.g. luminescence illustrant les niveaux d'ATP des cellules) de réplicats et ainsi éliminer tout bruit analytique. Or, les données téléchargées depuis PharamcoDB [60, 75] (Gray, gCSI et CTRPv2) sont déjà normalisées (% viabilité) à l'aide d'un processus standardisé intégré à PharmacoGx [55]. Cela étant dit, les résultats de la Figure 15 démontre bien qu'il y a concordance dans les réponses, malgré l'incertitude liée aux variances biologiques et expérimentales, et à la normalisation. Notamment, on observe que les paires partageant au moins 8 concentrations (le nombre de concentrations recommandées pour bien représenter la

réponse complète [19]), ont pour la majorité une RMSDΔ sous 20 et ont des réponses relativement semblables. Ces paires, partageant au moins 8 concentrations, représentent 22.50%, 99.63% et 84.04% du total des paires de réplicats pour les jeux gCSI, CTRPv2 et Gray, respectivement.

La difficulté à obtenir un bon niveau de corrélation (0.5) avec les métriques d'efficience estimées est indicateur d'un important bruit analytique causé par la régression (Levenberg-Marquardt, Fig. 16 et Supp. Fig. 20.A). Haibe-Kains *et al.* [22, 87] ont présenté pour la première fois les importants écarts entre les IC_{50}/EC_{50} et AAC estimés d'expériences répliquées. Les variabilités expérimentales et biologiques furent principalement explorées [19, 83], tandis que les limitations de la méthodologie Levenberg-Marquardt furent implicitement mentionnées comme possibles sources de ces écarts [76, 83, 86, 87, 155].

Les travaux du Chapitre 3 démontrent quantitativement, et pour la première fois, que le choix de la méthode résumant les réponses en métrique d'efficience affecte la constance des métriques d'expériences répliquées (Fig. 16 et Supp. Fig. 20 et 28). Jusqu'à présent, le manque d'alternative à Levenberg-Marquardt rendait cette démonstration difficile. Les quelques travaux portant sur l'application de l'inférence bayésienne à l'inférence des métriques d'efficience se basaient essentiellement sur une démonstration théorique [145] et/ou qualitative [77, 147] similaires à ceux présentés dans le Chapitre 2. Les quelques démonstrations qualitatives se résument à la comparaison des RMSE (Fig. 7, [147]) et l'utilisation d'expériences synthétiques (Fig. 9, [129, 145]). La RMSE est une métrique décrivant la qualité plutôt que la validité d'un ajustement, et n'est donc pas représentative de l'incertitude des métriques. Cela explique pourquoi nous observons une petite RMSE pour une sigmoïde forcée sur une réponse plate (bleue, Fig. 7), confirmant l'invalidité de cette métrique pour décrire notre confiance dans une courbe dose-réponse et ses métriques d'efficience. De plus, ces démonstrations sont faites sur un nombre restreint d'expériences, réelles comme synthétiques, et, contrairement aux travaux du Chapitre 2, ne sont pas toujours faites en comparaison avec Levenberg-Marquardt. Le Chapitre 3 présente une analyse comparative unique et à grande échelle. En plus de démontrer le gain à considérer l'incertitude des métriques (celles-ci sont dès lors plus concordantes), la robustesse de l'inférence et les s de Levenberg-Marquardt sont aussi démontrées quantitativement (Fig. 17 et Fig. Supp. 24, 26 et 27). De plus, la comparaison considère les deux types de représentations (c.-à-d. valeur estimée et *posterior* inféré) sans manipulation additionnelle (e.g. plafonnement des LDR et HDR). Il est bien démontré que les *posteriors* sont d'importantes sources d'informations (Fig. 17).

Il est aussi intéressant de constater les différences de corrélations, considérant que les deux approches considèrent une distribution normale des réponses. Dans le cas de la méthode standard (Levenberg-Marquardt), cela est implicite à la méthodologie. Pour BiDRA, nous avons défini la fonction de vraisemblance en assumant cette normalité. Il nous semblait inadéquat d'utiliser une distribution log-normale et ainsi forcer les réponses à être positives [129], puisque ce nest pas toujours le cas (Fig. 4). Les expériences considérées dans nos analyses ne contiennent pas assez de données (peu de réponse pour une même concentration) pour évaluer adéquatement leur modélisation. Considérant ce contexte, l'utilisation d'une distribution normale et le partage d'un σ nous paraissait le plus approprié [77, 147]. Il serait intéressant de considérer des expériences ayant

plusieurs réponses pour une même concentration: il nous serait alors possible de tester, par exemple, l'inférence d'un σ indépendant pour chaque concentration. Cela nous permettrait aussi d'étudier d'avantage les rôles et effets des diverses sources de variabilité (c.-à-d. biologique, expérimentale et analytique) sur l'inférence des métriques d'efficience.

L'inférence bayésienne, et plus précisément le modèle BiDRA (Fig. 14.B), est une robustesse et puissante alternative à Levenberg-Marquardt pour l'évaluation des métriques d'efficience.

Les *posteriors* obtenus représentent les distributions des valeurs les plus probables pour chaque métrique d'efficience. Depuis ces *posteriors*, l'incertitude d'une métrique peut être quantifiée et représentée de diverses façons. Visuellement, il est utile de représenter les échantillons d'inférence par un histogramme (Fig. 36). La largeur de la distribution est indicatrice de l'incertitude. Il est aussi possible de dériver les limites d'un intervalle d'une confiance (e.g. 95%), leur écart étant aussi représentatif de l'incertitude. La médiane d'un *posterior* devrait toujours être interprétée en considérant un intervalle de confiance. Lorsqu'un *posterior* est unimodal, la médiane est souvent parmi les valeurs les plus probables et peut être interprété ainsi. Ce type de *posterior* est fréquemment associé aux métriques qui peuvent être inférées principalement depuis la réponse expérimentale (e.g. LDR, réponse complète et sigmoïde Fig. 36.A). Lorsqu'un *posterior* est plutôt multimodal, souvent associé à une importante incertitude, la médiane s'éloigne des valeurs les plus probables et son interprétation dans le contexte biologique est désuète. Ce type de *posterior* est obtenu lorsque la réponse est incomplète ou plate (Vert, Orange et Mauve, Fig. 17.A). Dans de tel cas, résumer un *posterior* en une seule valeur abstrait une importante quantité d'information pertinente. Cela est bien démontré par les grandes différences entre les corrélations des médianes et des *posteriors*, sur les trois jeux de données, pour les paires de réplicats incomplets (Fig. 16). Bien que les *posteriors* complets soient à favoriser, la médiane résume mieux l'information que la moyenne, notamment dans les cas où le *posterior* est unimodal et présente une certaine asymétrie (e.g. Pente, σ, Fig. 36.A). Ces deux métriques ne garantissent cependant pas que la valeur obtenue fasse réellement partie du *posterior*. Le mode des *posteriors* pourrait aussi être calculé et utilisé comme métrique référence. Dans le cas de *posteriors* unimodaux, cette valeur serait concrètement parmi les plus probables. Le mode serait aussi plus représentatif que la médiane dans le cas de *posterior* asymétrique unimodal. L'inconvénient avec l'utilisation du mode vient des cas où le *posterior* serait uniforme ou multimodale.

Les *posteriors* complets devraient toujours être priorisés pour représenter les métriques d'efficience. Or, leur interprétation et manipulation lors d'analyses subséquentes ne sont pas triviales considérant que les expérimentateurs sont habitués à la représentation des estimations Levenberg-Marquardt par une valeur en un seul point (de l'anglais *single-point value*). Les travaux présentant un modèle bayésien pour l'évaluation de métrique d'efficience [77, 129, 145, 147] ne démontrent pas ou que très peu l'utilité concrète des *posteriors* lors d'analyse post-inférence. Dans le Chapitre 2 une méthode comparant deux expériences est présentée. Celle-ci permet notamment de déterminer si les métriques sont statistiquement différentes, pour un niveau de signification α donné. La comparaison de métriques ne se base plus qu'uniquement sur un seuil de magnitude de différence (e.g. 10 *fold change* entre les IC_{50}): il est possible pour un expérimentateur d'identifier de faibles

différences significatives, comme de rejeter de larges différences non significatives. Dans le Chapitre 3, l'analyse d'un jeu de données complet (70 expériences) est donnée en exemple. Nous y présentons une façon de sélectionner des expériences (ou composés) selon divers critères. L'utilisation des *posteriors* permet ainsi aux expérimentateurs de combiner facilement des métriques d'efficience. Contrairement aux calculs de métriques telles que l'AUC/AAC et le DSS, les probabilités obtenues selon divers critères et métriques sont simples et intuitives à interpréter. Les *posteriors* sont utilisés pour dériver des probabilités de rangs, soit la probabilité qu'une expérience est une métrique plus petite (ou plus grande) que toutes les autres expériences du jeu. Se basant sur la méthode comparative du Chapitre 2, une visualisation par DAG (de l'anglais *directed acyclic graph*) des relations statistiques est proposée. Celle-ci permet de visualiser rapidement et simplement un jeu de données complet. Finalement, dans le Chapitre 4, nous utilisons les *posteriors* pour classer les expériences en termes de leur potentiel informatif. Les résultats de ces diverses méthodologies considèrent l'incertitude découlant des bruits biologique, expérimental et analytique. De plus, la présentation de telles méthodologies participe aux efforts d'accessibilités tels que discutés dans la Section 5.6. Il est important d'outiller adéquatement de potentiels utilisateurs, sans quoi l'approche proposée est inutilisable et devient désuète.

Nous avons bien démontré la capacité de notre modèle BiDRA à considérer et représenter l'incertitude au travers des *posteriors*. En comparaison, Levenberg-Marquardt ne permet tout simplement pas de quantifier l'incertitude. Pour un expérimentateur, cela résulte en une incapacité à évaluer adéquatement la confiance qu'il peut avoir dans les métriques d'efficience obtenues. Bien que ne pouvant pas être quantifiée depuis les métriques estimées, l'incertitude reste présente et est inévitablement propagée dans les analyses subséquentes, telles que le calcul de nouvelles métriques (AAC, Figures 23 et 24) et la sélection de composés (Fig. 19). Nous avons démontré que les *posteriors* sont quant à eux robustes à plusieurs types d'incertitude (e.g. réponse incomplète, réponse plate) et que leur utilisation dans des analyses post-inférence (Fig. 13, 19 et 30.C) mène à des conclusions statistiquement valables ainsi que représentantes des données et de leur contexte expérimental.

5.5. Évaluation du potentiel informatif des *posteriors*

L'évaluation du potentiel informatif va de pair avec la quantification de l'incertitude. Une telle évaluation est difficile lorsque la méthodologie se base sur les estimations de Levenberg-Marquardt. La RMSE est utilisée, bien qu'elle ne soit pas optimale (Section 5.4). L'évaluation visuelle des données expérimentales et de la courbe dose-réponse estimée reste la méthode évaluative la plus couramment utilisée. Or, tel que mentionné précédemment, cette approche est subjective (l'évaluation peut différer d'un expérimentateur à l'autre) et fastidieuse, notamment lorsque plusieurs expériences sont considérées. Notre confiance dans les résultats inférés (e.g. IC_{50}/EC_{50} et courbe dose-réponse) ainsi que leur précision sont limitées par la qualité et le potentiel informatif des données. Il n'est cependant pas anormal d'observer des réponses incomplètes ou plates [76], notamment lors de criblages à haut débit (HTS) et des phases exploratoires du processus de découverte de médicaments (DDP). Une bonne évaluation du potentiel informatif d'une expérience devient

un important outil d'interprétation pour l'expérimentateur et le guide dans son interprétation des résultats. Le processus d'évaluation du potentiel informatif présenté au Chapitre 4 se base sur des concepts établis (Section 4.2) et s'inscrit dans les efforts de rendre l'ensemble du processus BiDRA, méthodologie comme résultats, accessible, et se veut objectif en comparaison à l'évaluation visuelle.

Tel que discuté à la Section 4.4, la métrique ΔWAIC_k différentie le mieux les expériences ayant une réponse plate de celles ayant une réponse sigmoïde. De plus, le ΔWAIC_k étant plus robustes aux variations d'échantillonnage des *posteriors*, l'assignation d'une expérience à un groupe de potentiel informatif est plus stable et réplicable (Fig. 33). L'écart-type (É.-T.) des réponses nous permet principalement de différentier les réponses sigmoïdes incomplètes de celles qui sont mieux définies. L'utilité de la combinaison des deux métriques (c.-à-d. É.-T. et ΔWAIC_k) est bien démontrée par notre analyse de généralisation du processus sur le jeu de données gCSI (Section 4.4.4). De cette analyse, nous identifions un nouveau groupe d'expériences, soit le Groupe D. Ces expériences sont caractérisées d'un large É.-T. (≥ 20) bien que leur réponse soit mieux décrite par le modèle Line que le modèle BiDRA ($\Delta\text{WAIC}_k > 0$). Ces expériences présentent des réponses hautement variables, pouvant résulter de problèmes expérimentaux (e.g. instabilité de la lignée cellulaire). L'É.-T. seul classifierait mal ces expériences, bien que le ΔWAIC_k serait adéquat.

Le processus présenté au Chapitre 4 se base sur des hypothèses posées après inspection et visualisation d'un large nombre d'expériences provenant de divers jeux de données. Bien que ce processus se veuille objectif, l'utilisation de seuil sur les métriques choisies (c.-à-d. É.-T. et ΔWAIC_k) ajoute une subjectivité au processus. Cela étant dit, le seuil sur les É.-T., bien qu'arbitraire, semble bien départir les types de réponses et est principalement limitant lorsqu'une expérience présente un faible HDR (Section 4.4.3). Dans de tels cas, le plateau formant le HDR doit être défini sur plusieurs concentrations (e.g. Analogues 56 vs. 26, Fig. 34). Le deuxième seuil utilisé est sur le ΔWAIC_k. Celui-ci n'est cependant pas arbitraire puisque ΔWAIC_k=0 signifie qu'il n'y a mathématiquement pas de différence entre les capacités prédictives des deux modèles. Nous n'avons pas appliqué de seuil sur les valeurs de WAIC_k, car (1) elles varient selon le nombre de concentrations et de réponses mesurées, et (2) le seuil serait lui aussi arbitraire.

Il serait possible de définir d'autres seuils sur le ΔWAIC_k pour différentier les réponses sigmoïdes. Par exemple, nos résultats sur le jeu de données de l'IRIC suggèrent une séparation à $\Delta\text{WAIC}_k = 20$ (Fig. 31.B). Les expériences ayant un $\Delta\text{WAIC}_k \geq 20$ seraient du Groupe (●) et celles dont $0 \leq \Delta\text{WAIC}_k < 20$ seraient du Groupe B (●). Or, l'Analogue 31 se verrait attitrer un potentiel informatif élevé (●), ce qui ne coïncide pas avec nos attentes et observations (Fig. 30.A). Il est certain qu'une analyse à plus grande échelle (similaire à celle du Chapitre 3) devrait être faite pour déterminer un seuil additionnel sur le ΔWAIC_k. Cependant, nous semblons tirer profit du fait que l'évaluation présentée au Chapitre 4 se base sur diverses sources d'informations, soit la réponse et ses variations (c.-à-d. É.-T.), ainsi que la représentation des réponses par les *posteriors* (c.-à-d. ΔWAIC_k). Il serait aussi possible de définir une autre métrique que l'É.-T. pour caractériser le type de réponse. Tel que discuté dans la Section 4.4.3, l'É.-T. ne considère pas à proprement dit la forme et la complétude de la réponse, et tel que démontré par notre analyse du jeu de données gCSI, un large É.-T. n'est pas toujours garant d'une réponse sigmoïde (Groupe D, Fig. 34). La

complétude d'une courbe est souvent définie par l'évaluation de la composition de ces plateaux [55]. Les réponses des dernières concentrations sont comparées de telle sorte à déterminer s'il y a présence d'un plateau défini. Une telle approche présente deux principales limitations : (1) un minimum de deux seuils arbitraires doit être défini (c.-à-d. la différence entre les réponses, et le nombre de concentrations constituant un plateau défini) et (2) il devient difficile de différentier un plateau défini d'une réponse plate, ou de différentier le HDR du LDR. L'utilisation de l'É.-T. ne considère qu'un seuil arbitraire, limitant la subjectivité de l'approche, et cette métrique identifie plus facilement les réponses plates. De plus, nos hypothèses quant à la forme et à la complétude de la réponse depuis l'É.-T. sont confirmées (ou infirmées) par le ΔWAIC_k. La combinaison de ces deux métriques est complémentaire et permet d'évaluer adéquatement le potentiel informatif de la majorité des expériences.

L'inférence composée [118, 176] (de l'anglais *composable inference*) telle que proposée par `Turing.jl` [115] fut explorée comme alternative au processus proposé au Chapitre 4. Cette approche permet de « composer » un algorithme d'échantillonnage en combinant des blocs d'inférence définis. Il devient donc possible de combiner divers échantillonneurs pour différents sous-ensembles de paramètres et tirer avantage de leurs bénéfices individuels. Il est dès lors possible d'intégrer des variables discrètes à un modèle, et d'inférer celles-ci à l'aide d'un échantillonneur de type Monte-Carlo séquentiel (SMC, de l'anglais *sequential Monte-Carlo*) [115], tel que le *Particle Gibbs* (PG) [119, 120].

Dans le contexte précis de l'évaluation du potentiel informatif, une variable discrète, z, représenterait le choix du modèle mathématique (c.-à-d. log-logistique ou linéaire constant) décrivant le mieux la réponse observée. L'avantage de cette approche est l'inférence d'un *posterior* sur z permettant ainsi d'évaluer la probabilité que la réponse soit sigmoïde ou plate, en plus des *posteriors* des métriques d'efficience pour les deux modèles. `Turing.jl` permet de créer facilement un tel modèle d'inférence composé. La détermination du potentiel informatif d'une expérience se ferait sur la base du *posterior* de z, et ce, sans utilisation de seuil sur diverses métriques. Cela étant dit, plusieurs difficultés se présentent lors de l'application concrète d'une telle approche.

Premièrement, l'optimisation de l'échantillonneur composé n'est pas triviale. Les métriques de validation standards décrites par Gelman [91] ne peuvent être utilisées, car elles ont été conçues pour évaluer les résultats d'inférence se basant sur des approches par gradients (e.g. HMC, NUTS). De plus, il est impossible d'identifier les itérations divergentes et ainsi explorer les cas où les échantillonneurs explorent inadéquatement l'espace θ. Il est d'autant plus difficile de valider un modèle.

Deuxièmement, à chaque itération de l'échantillonneur composé, une valeur est échantillonnée pour l'ensemble des variables du modèle. La valeur de z détermine quel modèle (log-logistic ou linéaire constant) sera considéré pour l'étape d'échantillonnage des métriques d'efficience: seuls les paramètres de ce modèle seront considérés lors de l'exploration de l'espace θ. Selon l'implémentation de `Turing.jl`, les paramètres du modèle non considéré sont tout de même ajoutés à ses *posteriors* respectifs, sans évaluation. Il est donc important de soustraire ces valeurs des *posteriors* finaux. Or, selon la distribution de z, certains *posteriors* pourraient avoir très peu de valeurs et être

donc peu représentatifs. Une solution est de combiner les *posteriors* communs aux deux modèles (e.g. LDR et θ_{line}, les σ). Le θ_{line} pourrait être combiné au LDR et/ou au HDR. L'inconvénient à combiner les *posteriors* vient de leur exploration indépendante: la forme du *posterior* final pourrait ne pas être biaisée vers différentes régions. De plus, les paramètres IC_{50}/EC_{50} et pente restent avec des *posteriors* sous-échantillonnées, car il n'y a aucun équivalent dans le modèle Line. Cela devient hautement problématique dans des cas où le choix du modèle mathématique est incertain (e.g. Analogues 23 et 24, Figure 34). Bien qu'un *posterior* sur le paramètre z soit utile et informatif, l'inférence composée a le potentiel de compliquer l'interprétation et l'utilisation des *posteriors* des métriques d'efficience. Augmenter le nombre d'itérations permettrait de mieux représenter les *posteriors* même dans des situations de sous-échantillonnage. Cela a pour désavantage d'augmenter le temps de calculs et, comme nous le discuterons plus bas, l'inférence composée est déjà relativement lente.

Troisièmement, les échantillonneurs, bien qu'optimaux individuellement, ne résultent pas nécessairement en un échantillonneur composé optimal. Par exemple, la vitesse d'inférence de NUT-s et HMC peut déstabiliser un échantillonneur tel que le *Particle Gibbs* [115]. Cela résulte en une inférence qui bloque dans une région spécifique des *posteriors*. Une solution à ce problème est d'augmenter le nombre de particules considérées par le PG. De plus, l'algorithme NUT-S est parfois moins performant que le HMC dans le contexte d'inférence composée [115]. Cela pourrait être en partie dû aux échantillonnages d'échauffement. Normalement, ceux-ci sont rejetés. Or, ces échantillonnages sont considérés par le PG pour inférer le paramètre z. Pour solution, nous avons testé d'augmenter le nombre total d'itérations de l'échantillonneur composé et de rejeter les n premiers échantillons du PG. Cette approche jumelée à une augmentation du nombre de particules ralentit considérablement le processus d'inférence pour une expérience: pour 50 répétitions d'inférence sur les Analogues 1 et 4 (Fig. 29), nous obtenions une moyenne de 35 secondes par expérience (Intel i9-7920X, 4 *threads*). L'inférence se faisait sur 4 chaînes de 1 000 itérations (1 000 itérations d'échauffement) et utilisait 15 particules pour le PG. Alternativement, la moyenne passait à 20 secondes (Intel i9-7920X, 4 *threads*) par expériences lorsque l'algorithme HMC était utilisé pour les mêmes nombres d'itérations et de particules. Pour le HMC, nous avons défini le nombre de sauts (*leapforg steps*) $L = 100$ et le *learning rate* $\epsilon = 0.01$. L'avantage à utiliser cet algorithme plutôt que le NUT-S est l'absence des échantillonnages d'échauffement. Cela a cependant pour désavantage de devoir définir et optimiser les paramètres de l'échantillonneur (c.-à-d. nombre de *leapforg steps* et le *learning rate*). Ceux-ci doivent être optimisés de telle sorte à diminuer le temps de calcul et être applicable à plusieurs expériences, ce qui n'est pas trivial.

Les temps d'inférence mentionnés ci-haut sont relativement lents. Comparativement, l'inférence des modèles BiDRA et Line du Chapitre 4 prennent respectivement 1.67 et 0.3 seconde pour l'Analogue 1, et 2.20 et 0.06 seconde pour l'Analogue 4 (Intel i9-7920X, 4 *threads*). De plus, le processus d'inférence et d'analyses post-inférence proposé par l'interface BiDRA V2 (Section 4.5) nécessite un peu moins de 6 secondes (Intel i9-7920X, 4 *threads*) pour les deux expériences. Bien que l'inférence composée nécessiterait de moins de calculs post-inférence, le processus proposé au Chapitre 4 reste plus optimal en termes de temps et d'applicabilité à de larges jeux de données et pour un éventail d'expériences. Bien que l'inférence composée semblait être une alternative

intéressante, elle est très peu applicable dans le contexte d'analyse de la présente thèse, tel que démontré par les diverses difficultés et désavantages discutés ci-haut.

Le processus d'évaluation du potentiel informatif d'une expérience présenté au Chapitre 4 comprend quelques limitations. Tel que discuté ci-haut et dans la Section 4.4.3, celles-ci affectent peu les résultats finaux et les alternatives existantes sont peu applicables ou présentent plus de limitations. L'effet des limitations est aussi minimisé lorsque nous considérons que le but de l'évaluation du potentiel informatif n'est pas d'exclure une expérience, mais bien de guider l'interprétation des résultats. Rappelons-nous que ceux-ci divergent grandement des résultats communément obtenus avec Levenberg-Marquardt et qu'ils ne sont pas toujours triviaux à interpréter (Section 5.6). De plus, une expérience à faible potentiel informatif (•, Fig. 36.B) se retrouvent au bas d'une analyse d'ordonnancement (Fig. 37.B). L'ensemble des analyses du Chapitre 4 démontre bien la validité et l'applicabilité du processus d'évaluation du potentiel informatif d'une expérience. Celui-ci outille davantage les expérimentateurs et est un processus complémentaire à ceux présentés aux Chapitres 2 et 3.

5.6. Accessibilité

Avec les avancements computationnels et technologiques des dernières décennies, nous observons un certain débalancement entre la disponibilité d'outils et les méthodes proposées dans la littérature [131]. Cette même observation fut rapidement faite dans le cadre de la présente thèse. Il existe divers outils pour estimer les métriques d'efficience d'expérience (Section 1.2.3), mais il existe très peu d'outils pour inférer et analyser les *posteriors* de ces métriques (Section 1.4).

L'utilisation proactive d'une approche bayésienne par des expérimentateurs devient difficile, voire impossible. Bien que la robustesse du processus BiDRA et les lacunes de la méthode standard aient été démontrées dans le Chapitre 3, l'intégration d'un processus d'inférence est limité par l'accessibilité de la méthode. Considérant qu'outiller les expérimentateurs pour les assister dans leur processus analytique et décisionnel est l'un des objectifs de la présente thèse, il était important de rendre accessible l'ensemble du processus BiDRA. Le concept d'accessibilité se définit ici en deux parties. Premièrement, l'expérimentateur doit avoir accès à la méthode proposée. Il doit pouvoir l'utiliser et obtenir des résultats. Deuxièmement, les résultats doivent être accessibles en termes d'interprétabilité. L'expérimentateur doit être en mesure de comprendre et d'utiliser les résultats obtenus, sans quoi la première partie de la définition devient futile.

Il est possible d'accéder au processus BiDRA de deux façons. L'ensemble des programmes (Python et Julia) liés au processus sont disponibles via GitHub [`https://github.com/lemieux-lab`]. Cette forme d'accessibilité est de plus en plus répandue. Le processus d'inférence et d'analyse est aussi accessible via l'une des deux versions de l'interface web BiDRA (V1: `https://bidra.bioinfo.iric.ca/`, V2: `https://bidrav2.bioinfo.iric.ca/`). À notre connaissance, nous sommes les premiers à proposer un outil du genre. Tout récemment, Wheeler *et al.* ont

proposé ToxicR [177], une librairie R permettant d'utiliser des approches bayésienne et de *model averaging*. Or, cet outil requiert d'un expérimentateur des habilités en programmation et une excellente compréhension de l'inférence bayésienne pour pouvoir en définir les différents paramètres. De plus, seule la dose repère (de l'anglais *benchmark dose*) peut être inférée. Le manque d'outils a motivé la création des interfaces web BiDRA: inférer des *posteriors* pour les métriques d'efficience devient simple pour tout expérimentateur. Il est vrai qu'il n'est pas toujours trivial de transformer une méthodologie en un outil convivial, que ce soit à cause du temps à investir ou du manque de ressources (e.g. serveur hôte) et d'habiletés (e.g. développement web). Or, développer un tel outil est hautement bénéfique: cela permet à quiconque intéressé par la méthodologie proposée de la tester et l'utiliser rapidement. En retour, un expérimentateur peut nous communiquer ses commentaires et ainsi nous aider dans le développement de la méthodologie. Les analyses post-inférences proposées dans la deuxième version découlent de cette communication. L'outil proposé peut ultimement correspondre aux besoins des expérimentateurs. De plus, la démonstration convaincante de robustesse faite au Chapitre 3 crée une certaine confiance chez les expérimentateurs et les encourage à utiliser l'outil mis à leur disposition.

Tel que mentionné plus haut, l'accessibilité des résultats, en termes d'interprétabilité est tout aussi important que l'accès à la méthodologie. La méthodologie bayésienne proposée retourne les métriques d'inférence sous une nouvelle forme, soit des *posteriors*. Cette représentation diffère grandement de celle retournée par l'approche traditionnelle (Levenberg-Marquardt), soit des valeurs en un seul point (de l'anglais *single-point values*). Il est important d'outiller les utilisateurs de telle sorte qu'ils comprennent cette nouvelle représentation et puissent tirer tout le potentiel informatif, sans quoi ils ne seront pas portés à utiliser la méthodologie proposée. Cette pour cette raison que l'interface web BiDRA V2 retourne (1) l'ensemble des valeurs des *posteriors* ainsi que (2) des figures illustrant les résultats de certaines analyses post-inférence. Le premier résultat permet à tout utilisateur de mener ces propres analyses post-inférences. Le deuxième résultat permet à des utilisateurs d'analyser et d'interpréter facilement leurs données, sans manipulation supplémentaire de leur part. Ce deuxième point rend la représentation de métriques par des *posteriors* accessible pour un plus grand nombre d'expérimentateurs. Outre les processus d'analyse décrits dans les Chapitres 2, 3 et 4, très peu de travaux présentent ou proposent des méthodes pour utiliser des *posteriors* [132, 145].

Nous maximisons l'utilisation du processus d'inférence BiDRA en le rendant accessible via une interface web et en proposant des méthodes d'analyse post-inférence. Les expérimentateurs sont alors bien outillés pour répondre à certaines questions expérimentales tout en tenant compte de l'incertitude entourant leurs données et métriques. Les conclusions tirées sont plus précises et sont statistiquement valables.

5.7. Conclusion: Implications et Perspectives

Notre but de mieux outiller les expérimentateurs par un processus analytique et de prise de décisions appliquant l'inférence bayésienne à la caractérisation de l'efficience de composés chimiques a été atteint au travers de quatre objectifs concrets, tels que définis à la Section 1.4.1.

Considérant la méthode standard par régression non-linéaire (Levenberg-Marquardt) pour obtenir les métriques d'efficience, il est impossible de quantifier et d'évaluer explicitement l'incertitude des métriques. Des méthodologies, découplées du processus d'inférence, proposent des approximations de l'incertitude qui représentent parfois une quantification artificielle ou non représentative (Section 1.2.4). Quelques travaux proposent l'application de l'inférence bayésienne à l'analyse d'expériences dose-réponse de tout genre et pour des contextes bien précis qui ne correspondent pas à celui de la présente thèse (Section 1.4).

Les travaux présentés dans les Chapitres 2, 3 et 4 sont des alternatives aux importantes limitations de l'approche standard (Levenberg-Marquardt). L'inférence de *posteriors* pour représenter les métriques d'efficience considère explicitement l'incertitude causée par des bruits biologiques comme expérimentaux. Cette façon de représenter les métriques d'efficience minimise l'ajout et la propagation potentielle d'un nouveau bruit, soit le bruit analytique. L'évaluation de l'incertitude via l'inférence bayésienne permet d'éliminer les étapes de manipulation de données: les données aberrantes et les réponses incomplètes comme plates ne posent pas de limitations. Leurs *posteriors* sont représentatifs du potentiel informatif de ces expériences et peuvent être analysés tels quels, contrairement à la méthode standard (Levenberg-Marquardt) qui, dans de tels cas, retourne des métriques artificielles (dû à la manipulation des données) ou erronées. Les expérimentateurs n'ont plus besoin d'éliminer des expériences sur la base que la méthodologie utilisée ne peut retourner des métriques d'efficience valables et représentatives des données expérimentales. L'inférence bayésienne permet de considérer toute expérience lors d'une méta-analyse, telle que celle présentée par l'exemple de la Figure 19. Cela est particulièrement utile, car toute expérience est informative, bien qu'à différent niveau: l'inférence bayésienne permet d'extraire cette information. De plus, dû aux limitations de la méthode standard (Levenberg-Marquardt), l'intuition de l'expérimentateur joue un important rôle lors de l'interprétation des résultats. Nous intégrons cette intuition à même le calcul des métriques d'efficience via les *priors*. Cela permet entre autres d'uniformiser cette intuition en définissant des *priors* peu informatifs et généralistes, tels que ceux proposés et utilisés dans notre interface web BiDRA V2 (Section 4.5). Les *priors* permettent aussi d'appliquer de souples contraintes aux métriques d'efficience, de telle sorte à ne pas obtenir des valeurs aberrantes lorsque les données sont insuffisantes pour supporter une inférence précise.

Nos travaux diffèrent en plusieurs points des autres qui appliquent l'inférence bayésienne à l'analyse des expériences dose-réponse. Nous inférons les quatre métriques d'efficience de base et notre modèle est généraliste et applicable à plusieurs expériences, tel que démontré dans le Chapitre 3. Nous avons aussi démontré que les *posteriors* des métriques d'efficience d'expériences répliquées corrèlent considérablement mieux que les métriques estimées (Levenberg-Marquardt) pour ces mêmes expériences, et ce, même lorsque les réponses sont incomplètes ou plates. Cette

analyse démontre quantitativement et à grande échelle que l'inférence bayésienne est plus robuste et globalement supérieure à la méthode standard (Levenberg-Marquardt). Nous sommes les premiers à avoir fait une démonstration de la sorte et celle-ci justifie le développement de nouvelles méthodologies bayésiennes pour l'analyse d'expérience dose-réponse, tout en validant celles proposées par le passé. En parallèle, nous avons aussi démontré les effets des limitations de l'approche standard (Levenberg-Marquardt). Bien que ceux-ci soient connus théoriquement et observés par les expérimentateurs, ils n'ont jamais été démontrés quantitativement.

Ces travaux et démonstrations, motivent et justifient une transition vers l'utilisation d'une approche par inférence bayésienne pour obtenir les métriques d'efficience depuis des expériences dose-réponse; nos divers outils d'analyses, présentés aux Chapitres 2, 3 et 4, facilitent cette transition pour les expérimentateurs. Nos interfaces web BiDRA et BiDRA V2 (Sections 2.3.3 et 4.5) rendent accessible le processus d'inférence et facilitent son application à divers jeux de données. De plus, l'ensemble du code développé est publiquement accessible. Nous facilitons aussi l'accès et l'interprétabilité des résultats (c.-à-d. les *posteriors*) en présentant diverses méthodes d'analyse post-inférence. Leur utilité dans le processus d'analyse est démontrée via différents exemples concrets (Sections 2.3.2, 3.2.6 et 4.4). De plus, ces méthodes sont implémentées et accessibles depuis nos interfaces web. Ces méthodologies outillent les expérimentateurs dans divers processus de sélection de composés. Par exemple, elles permettent à un expérimentateur d'identifier les composés présentant une différence statistique dans leur IC_{50}/EC_{50} lorsque testés sur une lignée cellulaire d'intérêt et sur une lignée contrôle (Safa-tahar-henni *et al.*, en préparation). Ces outils ont été développés et mis en place de telle sorte que l'accessibilité ne soit pas un facteur limitant à l'application de l'inférence bayésienne à la caractérisation de l'efficience de composés chimiques.

Bien que nos travaux aient des implications concrètes et utiles dans le DDP, ils proposent aussi diverses perspectives intéressantes, allant de l'application de l'inférence Bayésienne à de nouveaux contextes, à l'exploitation des *posteriors* du modèle BiDRA dans de nouveaux travaux de recherche.

Le processus de normalisation des réponses est une source de bruit analytique. Ce bruit est considéré implicitement, au travers des réponses normalisées. Il serait intéressant d'intégrer le processus de normalisation au modèle bayésien utilisé pour inférer les métriques d'efficience. Cela aurait le potentiel d'uniformiser le processus de normalisation et aurait l'avantage d'intégrer explicitement le bruit analytique du processus dans l'incertitude des métriques d'efficience. Une modélisation des effets de plaques [19, 178] pourraient être intégrée au modèle bayésien. Il serait aussi possible d'échantillonner les valeurs du LDR directement de la distribution des contrôles négatifs, ou de définir une nouvelle fonction de vraisemblance qui intègrerait le calcul de la normalisation. Cette deuxième option permettrait aussi de revisiter la modélisation des réponses selon une distribution normale et pour un σ commun (Section 5.4). L'ajout du processus de normalisation au modèle bayésien pourrait se faire de façon découplée (deux étapes distinctes du processus général) ou intégrée via l'ajout d'un niveau au modèle hiérarchique. Le processus général d'inférence serait aussi applicable aux expériences dose-réponse considérant la croissance tumorale de base. Les travaux de [24, 122] ont démontré l'impact du taux de croissance de base des cellules sur le calcul des métriques d'inférence. Ils proposent ainsi de normaliser les réponses après un temps d'incubation

T selon des contrôles aux temps T_0 et T. Pour des expériences ayant les données nécessaires, cette normalisation pourrait être intégrée au processus d'inférence bayésienne.

Certains composés favorisent la croissance cellulaire à de faibles concentrations, un phénomène appelé "hormèse" [29]. Le modèle log-logistique ne peut pas modéliser adéquatement ce phénomène: la tendance hormétique est considérée comme une variance de la réponse à de faibles concentrations, et le LDR inféré est plus incertain. La modélisation de l'hormèse, via le modèle Brain-Cousens [29], pourraient être considérée soit de façon similaire à la méthode présentée dans le Chapitre 4 (comparaison des modèles log-logistique et Brain-Cousens), soit en considérant le modèle d'hormèse comme modèle de base. La deuxième option est particulièrement intéressante, car le modèle log-logistique deviendrait un cas spécial du modèle Brain-Cousens. Les deux modèles diffèrent d'un paramètre, soit le $h(x)$ modélisant la tendance hormétique. En absence d'hormèse, ce paramètre devient $h(x) = 0$. L'analyse des *posteriors* du modèle Brain-Cousens nous permettrait de définir la probabilité $P(h(x) \neq 0)$, soit la probabilité qu'il y ait hormèse. Similairement à l'intégration du modèle Brain-Cousens, d'autres modèles pourraient être considérés ou comparés au modèle log-logistique. Par exemple, le modèle Weibull [179, 180] est une version asymétrique du modèle log-logistique. Il est particulièrement utile lorsque les réponses sont principalement faibles (e.g. Analogues 26, 28, 39 et 41 Fig. 34) ou élevées, selon le type de réponse considérée (ascendante ou descendante). Dans de tels cas, il serait intéressant d'évaluer et comparer les deux modèles (Chapitre 4) ainsi que les *posteriors*. Il est à considérer cependant que la valeur du point d'inflexion du modèle Weibull n'est pas nécessairement l'IC_{50}/EC_{50} , dû à l'asymétrie de la courbe.

La robustesse du processus d'inférence bayésienne ayant été démontré, nous pouvons appliquer le processus d'inférence et les méthodologies d'analyse post-inférence à la modélisation de divers types de relations expérimentales. De nouveaux modèles respectant et représentant adéquatement les données devront être mis en place. Par exemple, nos méthodologies de comparaison et d'ordonnancement d'expériences seraient utiles lors de l'analyse des ratios BRET pour diverses conditions. Les expériences BRET permettent de monitorer les interactions protéine-protéine [181] et notamment de quantifier la liaison de ligands à des GPCRs [18]. Les différences entre les métriques de diverses conditions sont parfois subtiles et peuvent être influencées par la variance biologique des mesures de ratios à même une concentration. L'ajustement de courbe via régression non-linéaire est communément utilisée pour ce type de données (e.g Prism). Un autre contexte expérimental qui pourrait bénéficier du processus d'inférence est le profilage thermique du protéome (TPP, de l'anglais *thermal proteome profiling*) [182]. Brièvement, cette approche permet d'évaluer les changements dans la stabilité thermique de protéines en comparant les profils contrôle et traitement. Ces profils considèrent une quantification relative (fraction) du taux de protéines non dénaturées pour une température données. Pour identifier une différence entre deux conditions, une liste de critères numériques doit être remplie, et ce, pour au moins deux répliquats [182]. Ce type de méthodologie peut être facilement répliqué avec notre méthodologie d'évaluation de critères se basant sur les *posteriors* (Chapitre 3). Il serait ainsi possible d'évaluer la probabilité qu'il y ait un changement de stabilité thermique des protéines en présences d'un composé, par exemple. Considérant que les limitations de l'approche standard pour modéliser des relations (Levenberg-Marquardt, Section

1.2.4) restent les mêmes pour divers contextes expérimentaux, l'inférence bayésienne devient une robuste alternative pour ces problèmes de modélisation.

Il serait intéressant d'utiliser l'inférence bayésienne pour interpréter des cribles de synergie. Ces cribles sont analysés en comparant les réponses expérimentales aux réponses attendues étant un modèle de synergie (e.g. additif [183]). Les différences entre ces réponses déterminent s'il y a synergie. Notre modèle bayésien pourrait être modifié de telle sorte à intégrer un de ces modèles (e.g. Bliss [184], ZIP [185]). L'évaluation de la synergie se ferait en comparant les *posteriors* des différences entre les réponses observées et les réponses attendues selon le modèle: il nous serait donc possible d'évaluer les probabilités de synergie. Quelques travaux ont exploré cette approche, notamment pour les modèles Loewe [129] et d'agent unique le plus élevé (HSA, de l'anglais *highest single agent*) [132]. Il serait aussi intéressant d'évaluer le potentiel informatif de comparer statistiquement les différentes métriques et courbes obtenues par notre modèle BiDRA de base dans le contexte de synergie. Par exemple, comparer les résultats obtenus pour des monothérapies ou pour un composé à dose constante et un composé à dose changeante, et évaluer s'il y a une différence dans les métriques d'efficience.

Finalement, nos travaux de recherche proposent différentes possibilités d'exploitation des *posteriors* tels qu'inférés par notre modèle BiDRA. Ceux-ci seraient bénéfiques à une ré-évaluation des différences entre les métriques d'efficience d'expériences répliquées dans plusieurs jeux de données [22, 86]. Les résultats préliminaires présentés au Chapitre 3 démontrent une hausse de corrélation lorsque les *posteriors* sont utilisés. Les *posteriors* et leurs comparaisons statistiques sont aussi utiles au processus d'identification de biomarqueurs prédictifs de la réponse à un composé [186]. Les *posteriors* sont aussi sources d'information pertinente dans un contexte d'apprentissage automatique. Les modèles d'apprentissage machine [156–159] permettent de prédire l'efficience (selon divers métriques) étant donné la représentation d'un composé (e.g. SMILES [174]). Cela a le potentiel d'évaluer *in silico* plusieurs composés et d'en sélectionner un sous-ensemble à tester expérimentalement, diminuant ainsi les coûts associés à de telles expériences. L'utilisation des *posteriors* lors de l'entraînement des modèles permettrait, entre autres, de réduire les risque d'*overfitting* dû à un petit nombre d'échantillons d'entraînement [159, 161]. L'utilisation des *posteriors* deviendrait une méthode d'*augmentation de données* (de l'anglais *data augmentation*) [187]. De plus, la robustesse des *posteriors* et leur concordance pour différentes expériences répliquées faciliterait l'utilisation de divers jeux de données pour l'entraînement et l'évaluation de modèles automatiques [159].

Nos travaux contribuent concrètement aux efforts de recherche entrepris dans le cadre du processus de découverte de médicament (DDP), autant par leurs implications que par leurs persepectives. L'application de l'inférence bayésienne à la caractérisation de l'efficience de composés chimiques, via les méthodes et outils présentés dans la présente thèse, est une source d'information pertinente au processus décisionnel des expérimentateurs.

Références bibliographiques

[1] Ana Sofia Pina, Abid Hussain, and Ana Cecília A. Roque. An Historical Overview of Drug Discovery. In Ana Cecília A. Roque, editor, *Ligand-Macromolecular Interactions in Drug Discovery: Methods and Protocols*, Methods in Molecular Biology, pages 3–12. Humana Press, Totowa, NJ, 2010.

[2] Sandeep Sinha and Divya Vohora. Chapter 2 - Drug Discovery and Development: An Overview. In Divya Vohora and Gursharan Singh, editors, *Pharmaceutical Medicine and Translational Clinical Research*, pages 19–32. Academic Press, Boston, January 2018.

[3] Qingxin Li and CongBao Kang. Mechanisms of Action for Small Molecules Revealed by Structural Biology in Drug Discovery. *International Journal of Molecular Sciences*, 21(15):5262, July 2020.

[4] JP Hughes, S Rees, SB Kalindjian, and KL Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, 2011.

[5] Peter Imming, Christian Sinning, and Achim Meyer. Drugs, their targets and the nature and number of drug targets. *Nature Reviews Drug Discovery*, 5(10):821–834, October 2006. Number: 10 Publisher: Nature Publishing Group.

[6] Shimeng Guo, Tingting Zhao, Ying Yun, and Xin Xie. Recent progress in assays for GPCR drug discovery. *American Journal of Physiology-Cell Physiology*, 323(2):C583–C594, August 2022. Publisher: American Physiological Society.

[7] Konrad H. Bleicher, Hans-Joachim Böhm, Klaus Müller, and Alexander I. Alanine. Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery*, 2(5):369–378, 2003.

[8] Geoffrey M. Currie. Pharmacology, Part 1: Introduction to Pharmacology and Pharmacodynamics. *Journal of Nuclear Medicine Technology*, 46(2):81–86, June 2018.

[9] Paweł Szymański, Magdalena Markowicz, and Elżbieta Mikiciuk-Olasik. Adaptation of High-Throughput Screening in Drug Discovery—Toxicological Screening Tests. *International Journal of Molecular Sciences*, 13(1):427–452, January 2012. Number: 1 Publisher: Molecular Diversity Preservation International.

[10] Yao Xu, David W. Piston, and Carl Hirschie Johnson. A bioluminescence resonance energy transfer (BRET) system: Application to interacting circadian clock proteins. *Proceedings of the National Academy of Sciences*, 96(1):151–156, January 1999. Publisher: Proceedings of the National Academy of Sciences.

[11] Piyush Sharma and Alexis Dunham. Pharmacy Calculations. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023.

[12] Amancio Carnero. High throughput screening in drug discovery. *Clinical and Translational Oncology*, 8(7):482–490, 2006.

[13] Mario Niepel, Marc Hafner, Mirra Chung, and Peter K. Sorger. Measuring Cancer Drug Sensitivity and Resistance in Cultured Cells. *Current Protocols in Chemical Biology*, 9(2):55–74, 2017.

[14] Schenone M, Dančík V, Wagner Bk, and Clemons Pa. Target identification and mechanism of action in chemical biology and drug discovery. *Nature chemical biology*, 9(4), April 2013. Publisher: Nat Chem Biol.

[15] Ting-Chao Chou. Theoretical Basis, Experimental Design, and Computerized Simulation of Synergism and Antagonism in Drug Combination Studies. *Pharmacological Reviews*, 58(3):621–681, 2006.

[16] Ting-Chao Chou. Drug Combination Studies and Their Synergy Quantification Using the Chou-Talalay Method. *Cancer Research*, 70(2):440–446, 2010.

[17] Julie Foucquier and Mickael Guedj. Analysis of drug combinations: current methodological landscape. *Pharmacology Research & Perspectives*, 3(3):e00149, 2015.

[18] Leigh A. Stoddart, Elizabeth K. M. Johnstone, Amanda J. Wheal, Joëlle Goulding, Matthew B. Robers, Thomas Machleidt, Keith V. Wood, Stephen J. Hill, and Kevin D. G. Pfleger. Application of BRET to monitor ligand binding to GPCRs. *Nature Methods*, 12(7):661–663, July 2015. Number: 7 Publisher: Nature Publishing Group.

[19] Mario Niepel, Marc Hafner, Caitlin E. Mills, Kartik Subramanian, Elizabeth H. Williams, Mirra Chung, Benjamin Gaudio, Anne Marie Barrette, Alan D. Stern, Bin Hu, James E. Korkola, LINCS Consortium, Joe W Gray, Marc R. Birtwistle, Laura M Heiser, and Peter K Sorger. A Multi-center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. *Cell Systems*, 9(1):35–48, 2019.

[20] Nicola Tolliday. High-Throughput Assessment of Mammalian Cell Viability by Determination of Adenosine Triphosphate Levels. *Current Protocols in Chemical Biology*, 2(3):153–161, 2010.

[21] Marc Hafner, Mario Niepel, Kartik Subramanian, and Peter K. Sorger. Designing Drug-Response Experiments and Quantifying their Results. *Current Protocols in Chemical Biology*, 9(2):96–116, 2017.

[22] Benjamin Haibe-Kains, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C. Jin, Andrew H. Beck, Hugo J. W. L. Aerts, and John Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389–393, 2013.

[23] Mohammad Fallahi-Sichani, Saman Honarnejad, Laura M Heiser, Joe W Gray, and Peter K Sorger. Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nature Chemical Biology*, 9(11):708–714, 2013.

[24] Marc Hafner, Mario Niepel, and Peter K Sorger. Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics. *Nature Biotechnology*, 35(6):500–502, 2017.

[25] Aubhishek Zaman and Trever Bivona. Quantitative Framework for Bench-to-Bedside Cancer Research. *Cancers*, 14:5254, October 2022.

[26] Richard R. Neubig, Michael Spedding, Terry Kenakin, and Arthur Christopoulos. International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on Terms and Symbols in Quantitative Pharmacology. *Pharmacological Reviews*, 55(4):597–606, December 2003. Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Review.

[27] Elena Postnikova, Yu Cong, Lisa Evans DeWald, Julie Dyall, Shuiqing Yu, Brit J. Hart, Huanying Zhou, Robin Gross, James Logue, Yingyun Cai, Nicole Deiuliis, Julia Michelotti, Anna N. Honko, Richard S. Bennett, Michael R. Holbrook, Gene G. Olinger, Lisa E. Hensley, and Peter B. Jahrling. Testing therapeutics in cell-based assays: Factors that influence the apparent potency of drugs. *PLOS ONE*, 13(3):e0194880, 2018.

[28] James N. Weiss. The Hill equation revisited: uses and misuses. *The FASEB Journal*, 11(11):835–841, 1997.

[29] P. Brain and R. Cousens. An equation to describe dose responses where there is stimulation of growth at low doses. *Weed Research*, 29(2):93–96, 1989.

[30] Edward J Calabrese. Hormesis: changing view of the dose-response, a personal account of the history and current status. *Mutation Research/Reviews in Mutation Research*, 511(3):181–189, 2002.

[31] Caroline Labelle. *Méthodologie pour l'analyse de données de criblage: application à l'étude de la leucémie myéloïde aiguë*. Mémoire, Université de Montréal, Montréal, 2018.

[32] Christian Ritz. Toward a unified approach to dose–response modeling in ecotoxicology. *Environmental Toxicology and Chemistry*, 29(1):220–229, 2010.

[33] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.

[34] Donald W Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[35] J. S. Witte and S. Greenland. A nested approach to evaluating dose-response and trend. *Annals of Epidemiology*, 7(3):188–193, April 1997.

[36] Tuomas P. J. Knowles, Christopher A. Waudby, Glyn L. Devlin, Samuel I. A. Cohen, Adriano Aguzzi, Michele Vendruscolo, Eugene M. Terentjev, Mark E. Welland, and Christopher M. Dobson. An Analytical Solution to the Kinetics of Breakable Filament Assembly. *Science*, 326(5959):1533–1537, December 2009. Publisher: American Association for the Advancement of Science.

[37] Walter W. Focke, Isbe van der Westhuizen, Ndeke Musee, and Mattheüs Theodor Loots. Kinetic interpretation of log-logistic dose-time response curves. *Scientific Reports*, 7(1):2234, 2017.

[38] Yuhong Wang, Ajit Jadhav, Noel Southal, Ruili Huang, and Dac-Trung Nguyen. A Grid Algorithm for High Throughput Fitting of Dose-Response Curve Data. *Current Chemical Genomics*, 4:57–66, 2010.

[39] Thuy Tuong Nguyen, Kyungmin Song, Yury Tsoy, Jin Yeop Kim, Yong-Jun Kwon, Myungjoo Kang, and Michael Adsetts Edberg Hansen. Robust dose-response curve estimation applied to high content screening data analysis. *Source Code for Biology and Medicine*, 9(1):27,

December 2014.

[40] Rui-Ru Ji, Nathan O. Siemers, Ming Lei, Liang Schweizer, and Robert E. Bruccoleri. SDRS—an algorithm for analyzing large-scale dose–response data. *Bioinformatics*, 27(20):2921–2923, 2011.

[41] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels, Jill Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E. Myer, Barbara L. Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey, William R. Sellers, Robert Schlegel, and Levi A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[42] Bhagwan Yadav, Tea Pemovska, Agnieszka Szwajda, Evgeny Kulesskiy, Mika Kontro, Riikka Karjalainen, Muntasir Mamun Majumder, Disha Malani, Astrid Murumägi, Jonathan Knowles, Kimmo Porkka, Caroline Heckman, Olli Kallioniemi, Krister Wennerberg, and Tero Aittokallio. Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Scientific Reports*, 4(1):5193, 2014.

[43] Jeremy D. Scheff, Richard R. Almon, Debra C. DuBois, William J. Jusko, and Ioannis P. Androulakis. Assessment of Pharmacologic Area Under the Curve When Baselines are Variable. *Pharmaceutical Research*, 28(5):1081–1089, 2011.

[44] A DeLean, P J Munson, and D Rodbard. Simultaneous analysis of families of sigmoidal curves: application to bioassay, radioligand assay, and physiological dose-response curves. *American Journal of Physiology-Endocrinology and Metabolism*, 235(2):E97–102, 1978.

[45] Sudhindra R. Gadagkar and Gerald B. Call. Computational tools for fitting the Hill equation to dose–response curves. *Journal of Pharmacological and Toxicological Methods*, 71:68–76, 2015.

[46] R Core Team. R: A Language and Environment for Statistical Computing, 2021.

[47] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[48] Elisabeth Roesch, Joe G. Greener, Adam L. MacLean, Huda Nassar, Christopher Rackauckas, Timothy E. Holy, and Michael P. H. Stumpf. Julia for biologists. *Nature Methods*, 20(5):655–664, May 2023. Number: 5 Publisher: Nature Publishing Group.

[49] Jeff Bezanson, Stefan Karpinski, Viral B. Shah, and Alan Edelman. Julia: A Fast Dynamic Language for Technical Computing, September 2012.

[50] The MathWorks nc. MATLAB version: 9.13.0 (R2022b), 2023.

[51] Timur V. Elzhov, Katharine M. Mullen, Andrej-Nikolai Spiess, and Ben Bolker. minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds, September 2023.

[52] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[53] Stevan Z. Knezevic, Jens C. Streibig, and Christian Ritz. Utilizing R Software Package for Dose-Response Studies: The Concept and Data Analysis. *Weed Technology*, 21(3):840–848, 2007. Publisher: [Cambridge University Press, Weed Science Society of America].

[54] Alina Malyutina, Jing Tang, and Alberto Pessia. drda: An R package for dose-response data analysis. preprint, Bioinformatics, June 2021.

[55] Petr Smirnov, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, Adrian She, Catharina Olsen, Mark Freeman, Heather Selby, Deena M.A. Gendoo, Patrick Grossmann, Andrew H. Beck, Hugo J.W.L. Aerts, Mathieu Lupien, Anna Goldenberg, and Benjamin Haibe-Kains. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, 32(8):1244–1246, 2016.

[56] Christoph Schanzenbach, Fabian C. Schmidt, Patrick Breckner, Mark G. Teese, and Dieter Langosch. Identifying ionic interactions within a membrane using BLaTM, a genetic tool to measure homo- and heterotypic transmembrane helix-helix interactions. *Scientific Reports*, 7(1):43476, March 2017. Number: 1 Publisher: Nature Publishing Group.

[57] Anthony Mammoliti, Petr Smirnov, Minoru Nakano, Zhaleh Safikhani, Christopher Eeles, Heewon Seo, Sisira Kadambat Nair, Arvind S. Mer, Ian Smith, Chantal Ho, Gangesh Beri, Rebecca Kusko, Massive Analysis Quality Control (MAQC) Society Board of Directors, Eva Lin, Yihong Yu, Scott Martin, Marc Hafner, and Benjamin Haibe-Kains. Orchestrating and sharing large multimodal data for transparent and reproducible research. *Nature Communications*, 12(1):5797, October 2021.

[58] Nicholas A. Clark, Marc Hafner, Michal Kouril, Elizabeth H. Williams, Jeremy L. Muhlich, Marcin Pilarczyk, Mario Niepel, Peter K. Sorger, and Mario Medvedovic. GRcalculator: an online tool for calculating and mining dose–response data. *BMC Cancer*, 17(1):698, December 2017.

[59] Giovanni Y. Di Veroli, Chiara Fornari, Ian Goldlust, Graham Mills, Siang Boon Koh, Jo L Bramhall, Frances M. Richards, and Duncan I. Jodrell. An automated fitting procedure and software for dose-response curves with multiphasic features. *Scientific Reports*, 5(1):14701, 2015.

[60] Petr Smirnov, Victor Kofia, Alexander Maru, Mark Freeman, Chantal Ho, Nehme El-Hachem, George-Alexandru Adam, Wail Ba-alawi, Zhaleh Safikhani, and Benjamin Haibe-Kains. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Research*, 46(D1):gkx911, 2017.

[61] Alexander L. Ling, Weijie Zhang, Adam Lee, Yunong Xia, Mei-Chi Su, Robert F. Gruener, Sampreeti Jena, Yingbo Huang, Siddhika Pareek, Yuting Shan, and R. Stephanie Huang. Simplicity: web-based visualization and analysis of high-throughput cancer cell line screens. preprint, Pharmacology and Toxicology, September 2023.

[62] Alexander L R Lubbock, Leonard A Harris, Vito Quaranta, Darren R Tyson, and Carlos F Lopez. Thunor: visualization and analysis of high-throughput dose–response datasets. *Nucleic Acids Research*, 49(W1):W633–W640, 2021.

[63] Amrita Basu, Nicole E. Bodycombe, Jaime H. Cheah, Edmund V. Price, Ke Liu, Giannina I. Schaefer, Richard Y. Ebright, Michelle L. Stewart, Daisuke Ito, Stephanie Wang, Abigail L. Bracha, Ted Liefeld, Mathias Wawer, Joshua C. Gilbert, Andrew J. Wilson, Nicolas Stransky, Gregory V. Kryukov, Vlado Dancik, Jordi Barretina, Levi A. Garraway, C. Suk-Yee Hon, Benito Munoz, Joshua A. Bittker, Brent R. Stockwell, Dineo Khabele, Andrew M. Stern, Paul A. Clemons, Alykhan F. Shamji, and Stuart L. Schreiber. An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. *Cell*, 154(5):1151–1161, 2013.

[64] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and Mathew J. Garnett. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(Database issue):D955–D961, 2013.

[65] Douglas T. Ross, Uwe Scherf, Michael B. Eisen, Charles M. Perou, Christian Rees, Paul Spellman, Vishwanath Iyer, Stefanie S. Jeffrey, Matt Van de Rijn, Mark Waltham, Alexander Pergamenschikov, Jeffrey C. F. Lee, Deval Lashkari, Dari Shalon, Timothy G. Myers, John N. Weinstein, David Botstein, and Patrick O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24(3):227–235, March 2000. Number: 3 Publisher: Nature Publishing Group.

[66] Peter M. Haverty, Eva Lin, Jenille Tan, Yihong Yu, Billy Lam, Steve Lianoglou, Richard M. Neve, Scott Martin, Jeff Settleman, Robert L. Yauch, and Richard Bourgon. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature*, 533(7603):333–337, 2016.

[67] John Patrick Mpindi, Bhagwan Yadav, Päivi Östling, Prson Gautam, Disha Malani, Astrid Murumägi, Akira Hirasawa, Sara Kangaspeska, Krister Wennerberg, Olli Kallioniemi, and Tero Aittokallio. Consistency in drug response profiling. *Nature*, 540(7631):E5–E6, 2016.

[68] Amar Koleti, Raymond Terryn, Vasileios Stathias, Caty Chung, Daniel J. Cooper, John P. Turner, Dušica Vidovic, Michele Forlin, Tanya T. Kelley, Alessandro D'Urso, Bryce K. Allen, Denis Torre, Kathleen M. Jagodnik, Lily Wang, Sherry L. Jenkins, Christopher Mader, Wen Niu, Mehdi Fazel, Naim Mahi, Marcin Pilarczyk, Nicholas Clark, Behrouz Shamsaei, Jarek

Meller, Juozas Vasiliauskas, John Reichard, Mario Medvedovic, Avi Ma'ayan, Ajay Pillai, and Stephan C. Schürer. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Research*, 46(D1):D558–D566, January 2018.

[69] Anneleen Daemen, Obi L. Griffith, Laura M. Heiser, Nicholas J. Wang, Oana M. Enache, Zachary Sanborn, Francois Pepin, Steffen Durinck, James E. Korkola, Malachi Griffith, Joe S. Hur, Nam Huh, Jongsuk Chung, Leslie Cope, Mary Jo Fackler, Christopher Umbricht, Saraswati Sukumar, Pankaj Seth, Vikas P. Sukhatme, Lakshmi R. Jakkula, Yiling Lu, Gordon B. Mills, Raymond J. Cho, Eric A. Collisson, Laura J. van't Veer, Paul T. Spellman, and Joe W. Gray. Modeling precision treatment of breast cancer. *Genome Biology*, 14(10):R110, December 2013.

[70] Mathew J. Garnett, Elena J. Edelman, Sonja J. Heidorn, Chris D. Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I. Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J. Milano, Graham R. Bignell, Ah T. Tam, Helen Davies, Jesse A. Stevenson, Syd Barthorpe, Stephen R. Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick O'Brien, Jessica L. Boisvert, Stacey Price, Wooyoung Hur, Wanjuan Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey A. Engelman, Sreenath V. Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S. Gray, Jeffrey Settleman, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Sridhar Ramaswamy, Ultan McDermott, and Cyril H. Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012.

[71] Eoghan R. Malone, Marc Oliva, Peter J. B. Sabatini, Tracy L. Stockley, and Lillian L. Siu. Molecular profiling for precision cancer therapies. *Genome Medicine*, 12(1):8, January 2020.

[72] Christiaan Klijn, Steffen Durinck, Eric W Stawiski, Peter M Haverty, Zhaoshi Jiang, Hanbin Liu, Jeremiah Degenhardt, Oleg Mayba, Florian Gnad, Jinfeng Liu, Gregoire Pau, Jens Reeder, Yi Cao, Kiran Mukhyala, Suresh K Selvaraj, Mamie Yu, Gregory J Zynda, Matthew J Brauer, Thomas D Wu, Robert C Gentleman, Gerard Manning, Robert L Yauch, Richard Bourgon, David Stokoe, Zora Modrusan, Richard M Neve, Frederic J de Sauvage, Jeffrey Settleman, Somasekar Seshagiri, and Zemin Zhang. A comprehensive transcriptional portrait of human cancer cell lines. *Nature Biotechnology*, 33(3):306–312, 2015.

[73] Matthew G. Rees, Brinton Seashore-Ludlow, Jaime H. Cheah, Drew J. Adams, Edmund V. Price, Shubhroz Gill, Sarah Javaid, Matthew E. Coletti, Victor L. Jones, Nicole E. Bodycombe, Christian K. Soule, Benjamin Alexander, Ava Li, Philip Montgomery, Joanne D. Kotz, C. Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Daniel A. Haber, Clary B. Clish, Joshua A. Bittker, Michelle Palmer, Bridget K. Wagner, Paul A. Clemons, Alykhan F. Shamji, and Stuart L. Schreiber. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology*, 12(2):109–116, 2016.

[74] Brinton Seashore-Ludlow, Matthew G. Rees, Jaime H. Cheah, Murat Cokol, Edmund V. Price, Matthew E. Coletti, Victor Jones, Nicole E. Bodycombe, Christian K. Soule, Joshua

Gould, Benjamin Alexander, Ava Li, Philip Montgomery, Mathias J. Wawer, Nurdan Kuru, Joanne D. Kotz, C. Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Joshua A. Bittker, Michelle Palmer, James E. Bradner, Alykhan F. Shamji, Paul A. Clemons, and Stuart L. Schreiber. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery*, 5(11):1210–1223, 2015.

[75] Nikta Feizi, Sisira Kadambat Nair, Petr Smirnov, Gangesh Beri, Christopher Eeles, Parinaz Nasr Esfahani, Minoru Nakano, Denis Tkachuk, Anthony Mammoliti, Evgeniya Gorobets, Arvind Singh Mer, Eva Lin, Yihong Yu, Scott Martin, Marc Hafner, and Benjamin Haibe-Kains. PharmacoDB 2.0: improving scalability and transparency of in vitro pharmacogenomics analysis. *Nucleic Acids Research*, 50(D1):D1348–D1357, 2021.

[76] Raziur Rahman, Saugato Rahman Dhruba, Kevin Matlock, Carlos De-Niz, Souparno Ghosh, and Ranadip Pal. Evaluating the consistency of large-scale pharmacogenomic studies. *Briefings in Bioinformatics*, 20(5):1734–1753, 2019.

[77] Felice Carlo Simeone and Anna Luisa Costa. Quantifying uncertainty in dose–response screenings of nanoparticles: a Bayesian data analysis. *Nanotoxicology*, 16(2):135–151, February 2022.

[78] Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 528(7580):84–87, December 2015.

[79] Iurie Caraus, Abdulaziz A. Alsuwailem, Robert Nadon, and Vladimir Makarenkov. Detecting and overcoming systematic bias in high-throughput screening technologies: a comprehensive review of practical issues and methodological solutions. *Briefings in Bioinformatics*, 16(6):974–986, 2015.

[80] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, January 1979. Publisher: Institute of Mathematical Statistics.

[81] Michael P. Cummings, Scott A. Handley, Daniel S. Myers, David L. Reed, Antonis Rokas, Katarina Winka, and Bruce Rannala. Comparing Bootstrap and Posterior Probability Values in the Four-Taxon Case. *Systematic Biology*, 52(4):477–487, 2003.

[82] Maria Dilleen, Günter Heimann, and Ian Hirsch. Non-parametric estimators of a monotonic dose–response curve and bootstrap confidence intervals. *Statistics in Medicine*, 22(6):869–882, 2003.

[83] Mehdi Bouhaddou, Matthew S. DiStefano, Eric A. Riesel, Emilce Carrasco, Hadassa Y. Holzapfel, DeAnalisa C. Jones, Gregory R. Smith, Alan D. Stern, Sulaiman S. Somani, T. Victoria Thompson, and Marc R. Birtwistle. Drug response consistency in CCLE and CGP. *Nature*, 540(7631):E9–E10, 2016.

[84] Shuai Chang, Daomin Zhuang, Wei Guo, Lin Li, Wenfu Zhang, Siyang Liu, Hanping Li, Yongjian Liu, Zuoyi Bao, Jingwan Han, Hongbin Song, and Jingyun Li. The Antiviral Activity of Approved and Novel Drugs against HIV-1 Mutations Evaluated under the Consideration of Dose-Response Curve Slope. *PLOS ONE*, 11(3):e0149467, March 2016.

[85] Joel Greshock, Kurtis E. Bachman, Yan Y. Degenhardt, Junping Jing, Yuan H. Wen, Stephen Eastman, Elizabeth McNeil, Christopher Moy, Ronald Wegrzyn, Kurt Auger, Mary Ann

Hardwicke, and Richard Wooster. Molecular Target Class Is Predictive of In vitro Response Profile. *Cancer Research*, 70(9):3677–3686, 2010.

[86] Zhaleh Safikhani, Petr Smirnov, Mark Freeman, Nehme El-Hachem, Adrian She, Quevedo Rene, Anna Goldenberg, Nicolai J. Birkbak, Christos Hatzis, Leming Shi, Andrew H. Beck, Hugo J.W.L. Aerts, John Quackenbush, and Benjamin Haibe-Kains. Revisiting inconsistency in large pharmacogenomic studies. *F1000Research*, 5:2333, 2016.

[87] Zhaleh Safikhani, Nehme El-Hachem, Rene Quevedo, Petr Smirnov, Anna Goldenberg, Nicolai Juul Birkbak, Christopher Mason, Christos Hatzis, Leming Shi, Hugo JWL Aerts, John Quackenbush, and Benjamin Haibe-Kains. Assessment of pharmacogenomic agreement. *F1000Research*, 5:825, May 2016.

[88] Thomas Bayes and null Price. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, January 1997. Publisher: Royal Society.

[89] Jorge López Puga, Martin Krzywinski, and Naomi Altman. Bayes' theorem. *Nature Methods*, 12(4):277–278, April 2015. Number: 4 Publisher: Nature Publishing Group.

[90] B. Lambert. *A Student's Guide to Bayesian Statistics*. SAGE Publications, 2018.

[91] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, November 2013.

[92] Christopher A. Sims. Discrete Approximations to Continuous Time Distributed Lags in Econometrics. *Econometrica*, 39(3):545–563, 1971. Publisher: [Wiley, The Econometric Society].

[93] A. O'Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, November 1991.

[94] Marc Kennedy. Bayesian quadrature with non-normal approximating functions. *Statistics and Computing*, 8(4):365–375, December 1998.

[95] Bernard D. Flury. Acceptance–Rejection Sampling Made Easy. *SIAM Review*, 32(3):474–476, September 1990. Publisher: Society for Industrial and Applied Mathematics.

[96] Sheehan Olver and Alex Townsend. Fast inverse transform sampling in one and two dimensions, July 2013.

[97] Ben Lambert. Evaluation of model fit and hypothesis testing. In *A Student's Guide to Bayesian Stastistics*, page 520. SAGE Publication, illustrated edition, 2018.

[98] Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992. Publisher: Institute of Mathematical Statistics.

[99] Stephen P. Brooks and Andrew Gelman. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, December 1998.

[100] Vivekananda Roy. Convergence diagnostics for Markov chain Monte Carlo, October 2019.

[101] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[102] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970. Publisher: [Oxford University Press, Biometrika Trust].

[103] Alan E. Gelfand and Adrian F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[104] Radford M. Neal. *MCMC using Hamiltonian dynamics*. May 2011. arXiv:1206.1901 [physics, stat].

[105] Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv*, 2011. _eprint: 1111.4246.

[106] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, August 2009.

[107] Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability.

[108] George Casella and Edward I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, August 1992.

[109] Simon Duane, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, September 1987.

[110] David J. Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, October 2000.

[111] David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067, 2009.

[112] Martyn Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Working Papers*, 2003.

[113] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, April 2016. Publisher: PeerJ Inc.

[114] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017.

[115] Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. *International Conference on Artificial Intelligence and Statistics*, pages 1682–1690, 2018.

[116] David Spiegelhalter, Andrew Thomas, Nicky Best, and Wally Gilks. BUGS 0.5 Bayesian inference Using Gibbs Sampling Manual. *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*, pages 1–59, 1996.

[117] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, 2023.

[118] Robert Zinkov and Chung-chieh Shan. Composing inference algorithms as program transformations, July 2017.

[119] Frank Wood, Jan Willem van de Meent, and Vikash Mansinghka. A New Approach to Probabilistic Programming Inference, July 2015.

[120] Nicolas Chopin and Sumeetpal S Singh. On particle Gibbs sampling. *arXiv*, 2013. _eprint: 1304.1887.

[121] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342, June 2010.

[122] Marc Hafner, Mario Niepel, Mirra Chung, and Peter K Sorger. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature Methods*, 13(6):521–527, 2016.

[123] Ross H Johnstone, Rémi Bardenet, David J Gavaghan, and Gary R Mirams. Hierarchical Bayesian inference for ion channel screening dose-response data. *Wellcome Open Research*, 1:6, March 2017.

[124] Ivo D. Shterev, David B. Dunson, Cliburn Chan, and Gregory D. Sempowski. Bayesian Multi-Plate High-Throughput Screening of Compounds. *Scientific Reports*, 8(1):9551, 2018.

[125] L. W. Huson and N. Kinnersley. Bayesian fitting of a logistic dose–response curve with numerically derived priors. *Pharmaceutical Statistics*, 8(4):279–286, 2009.

[126] Joe Collis, Anthony J. Connor, Marcin Paczkowski, Pavitra Kannan, Joe Pitt-Francis, Helen M. Byrne, and Matthew E. Hubbard. Bayesian Calibration, Validation and Uncertainty Quantification for Predictive Modelling of Tumour Growth: A Tutorial. *Bulletin of Mathematical Biology*, 79(4):939–974, April 2017.

[127] Andrew P Grieve and Michael Krams. ASTIN: a Bayesian adaptive dose–response trial in acute stroke. *Clinical Trials*, 2(4):340–351, August 2005. Publisher: SAGE Publications.

[128] David Ohlssen and Amy Racine. A Flexible Bayesian Approach for Modeling Monotonic Dose–Response Relationships in Drug Development Trials. *Journal of Biopharmaceutical Statistics*, 25(1):137–156, 2015.

[129] Violeta G. Hennessey, Gary L. Rosner, Robert C. Bast Jr, and Min-Yu Chen. A Bayesian Approach to Dose–Response Assessment and Synergy and Its Application to In Vitro Dose–Response Studies. *Biometrics*, 66(4):1275–1283, 2010.

[130] Michael J. Messner, Cynthia L. Chappell, and Pablo C. Okhuysen. Risk Assessment for Cryptosporidium: A Hierarchical Bayesian Analysis of Human Dose Response Data. *Water Research*, 35(16):3934–3940, 2001.

[131] Matthew W. Wheeler, Sooyeong Lim, John S. House, Keith R. Shockley, A. John Bailer, Jennifer Fostel, Longlong Yang, Dawan Talley, Ashwin Raghuraman, Jeffery S. Gift, J. Allen Davis, Scott S. Auerbach, and Alison A. Motsinger-Reif. ToxicR: A computational platform in R for computational toxicology and dose–response analyses. *Computational Toxicology*, 25:100259, February 2023.

[132] Haoting Zhang, Carl Henrik Ek, Magnus Rattray, and Marta Milo. SynBa: improved estimation of drug combination synergies with uncertainty quantification. *Bioinformatics*, 39(Supplement_1):i121–i130, June 2023.

[133] Suji Jang, Kan Shao, and Weihsueh A. Chiu. Beyond the cancer slope factor: Broad application of Bayesian and probabilistic approaches for cancer dose-response assessment. *Environment International*, 175:107959, May 2023.

[134] Tasnim Hamza, Andrea Cipriani, Toshi A Furukawa, Matthias Egger, Nicola Orsini, and Georgia Salanti. A Bayesian dose–response meta-analysis model: A simulations study and application. *Statistical Methods in Medical Research*, 30(5):1358–1372, 2021. _eprint: 2004.12737.

[135] M. Rudin and E. E. Kim. Imaging in Drug Discovery and Early Clinical Trials. *Journal of Nuclear Medicine*, 48(6):1037–1037, June 2007.

[136] The academic pursuit of screening. *Nature Chemical Biology*, 3(8):433–433, 2007.

[137] Caroline Pabst, Jana Krosl, Iman Fares, Geneviève Boucher, Réjean Ruel, Anne Marinier, Sébastien Lemieux, Josée Hébert, and Guy Sauvageau. Identification of small molecules that support human leukemia stem cell activity ex vivo. *Nature Methods*, 11(4):436–442, 2014.

[138] I El Naqa, G Suneja, P E Lindsay, A J Hope, J R Alaly, M Vicic, J D Bradley, A Apte, and J O Deasy. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose–volume outcome relationships. *Physics in Medicine and Biology*, 51(22):5719–5735, 2006.

[139] José M. Bernardo and Adrian F. M. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[140] Michael K. Smith and Scott Marshall. A Bayesian design and analysis for dose-response using informative prior information. *Journal of Biopharmaceutical Statistics*, 16(5):695–709, 2006.

[141] John Salvatier, Thomas Wiecki, and Christopher Fonnesbeck. Probabilistic Programming in Python using PyMC, July 2015. arXiv:1507.08050 [stat].

[142] Ming-Hui Chen, Joseph G. Ibrahim, and Constantin Yiannoutsos. Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(1):223–242, 1999. Publisher: [Royal Statistical Society, Wiley].

[143] Isabelle Albert, Sophie Donnet, Chantal Guihenneuc-Jouyaux, Samantha Low-Choy, Kerrie Mengersen, and Judith Rousseau. Combining Expert Opinions in Prior Elicitation. *Bayesian Analysis*, 7(3):503–532, September 2012. Publisher: International Society for Bayesian Analysis.

[144] Joe Collis, Michael R. Hill, James R. Nicol, Philip J. Paine, and Jonathan A. Coulter. A hierarchical Bayesian approach to calibrating the linear-quadratic model from clonogenic survival assay data. *Radiotherapy and Oncology*, 124(3):541–546, 2017.

[145] Ross H Johnstone, Rémi Bardenet, David J Gavaghan, and Gary R Mirams. Hierarchical Bayesian inference for ion channel screening dose-response data. *Wellcome Open Research*, 1:6, 2016.

[146] Caroline Labelle, Anne Marinier, and Sébastien Lemieux. Enhancing the drug discovery process: Bayesian inference for the analysis and comparison of dose–response experiments. *Bioinformatics*, 35(14):i464–i473, 2019.

[147] Dapeng Zhang and Jin Xu. A Bayesian design for finding optimal biological dose with mixed types of responses of toxicity and efficacy. *Contemporary Clinical Trials*, 127:107113, April 2023.

[148] Wesley Tansey, Christopher Tosh, and David M. Blei. A Bayesian model of dose-response for cancer drug studies. *The Annals of Applied Statistics*, 16(2):680–705, June 2022. Publisher: Institute of Mathematical Statistics.

[149] Laura M. Heiser, Anguraj Sadanandam, Wen-Lin Kuo, Stephen C. Benz, Theodore C. Goldstein, Sam Ng, William J. Gibb, Nicholas J. Wang, Safiyyah Ziyad, Frances Tong, Nora Bayani, Zhi Hu, Jessica I. Billig, Andrea Dueregger, Sophia Lewis, Lakshmi Jakkula, James E. Korkola, Steffen Durinck, François Pepin, Yinghui Guan, Elizabeth Purdom, Pierre Neuvial, Henrik Bengtsson, Kenneth W. Wood, Peter G. Smith, Lyubomir T. Vassilev, Bryan T. Hennessy, Joel Greshock, Kurtis E. Bachman, Mary Ann Hardwicke, John W. Park, Laurence J. Marton, Denise M. Wolf, Eric A. Collisson, Richard M. Neve, Gordon B. Mills, Terence P. Speed, Heidi S. Feiler, Richard F. Wooster, David Haussler, Joshua M. Stuart, Joe W. Gray, and Paul T. Spellman. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8):2724–2729, February 2012.

[150] Timothy M Errington, Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Joelle Lomax, and Brian A Nosek. An open investigation of the reproducibility of cancer biology research. *eLife*, 3:e04333, 2014.

[151] Petr Smirnov, Ian Smith, Zhaleh Safikhani, Wail Ba-alawi, Farnoosh Khodakarami, Eva Lin, Yihong Yu, Scott Martin, Janosch Ortmann, Tero Aittokallio, Marc Hafner, and Benjamin Haibe-Kains. Evaluation of statistical approaches for association testing in noisy drug screening data. *BCM Bioinformatics*, 23:188, 2022. _eprint: 2104.14036.

[152] Nurken Berdigaliyev and Mohamad Aljofan. An overview of drug discovery and development. *Future Medicinal Chemistry*, 12(10):939–947, May 2020. Publisher: Future Science.

[153] Steven S. Seefeldt, Jens Erik Jensen, and E. Patrick Fuerst. Log-Logistic Analysis of Herbicide Dose-Response Relationships. *Weed Technology*, 9(2):218–227, 1995.

[154] Michael C. Newman. "What exactly are you inferring?" A closer look at hypothesis testing. *Environmental Toxicology and Chemistry*, 27(5):1013–1019, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1897/07-373.1.

[155] Zhaleh Safikhani, Nehme El-Hachem, Petr Smirnov, Mark Freeman, Anna Goldenberg, Nicolai J. Birkbak, Andrew H. Beck, Hugo J. W. L. Aerts, John Quackenbush, and Benjamin Haibe-Kains. Consistency in large pharmacogenomic studies. *Nature*, 540(7631):E1–E2, 2016.

[156] Yongkang Zhan, Jifeng Guo, C L Philip Chen, and Xian-Bing Meng. iBT-Net: an incremental broad transformer network for cancer drug response prediction. *Briefings in Bioinformatics*, page bbad256, July 2023.

[157] Min Li, Yake Wang, Ruiqing Zheng, Xinghua Shi, Yaohang Li, Fang-Xiang Wu, and Jianxin Wang. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(2):575–582, 2018.

[158] Ran Su, Xinyi Liu, Leyi Wei, and Quan Zou. Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods*, 166:91–102, 2019.

[159] Fangfang Xia, Jonathan Allen, Prasanna Balaprakash, Thomas Brettin, Cristina Garcia-Cardona, Austin Clyde, Judith Cohn, James Doroshow, Xiaotian Duan, Veronika Dubinkina, Yvonne Evrard, Ya Ju Fan, Jason Gans, Stewart He, Pinyi Lu, Sergei Maslov, Alexander Partin, Maulik Shukla, Eric Stahlberg, Justin M Wozniak, Hyunseung Yoo, George Zaki, Yitan Zhu, and Rick Stevens. A cross-study analysis of drug response prediction in cancer cell lines. *arXiv*, 2021. _eprint: 2104.08961.

[160] Guangxu Jin and Stephen T.C. Wong. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discovery Today*, 19(5):637–644, May 2014.

[161] George Adam, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precision Oncology*, 4(1):19, 2020.

[162] Austin Clyde, Tom Brettin, Alexander Partin, Maulik Shaulik, Hyunseung Yoo, Yvonne Evrard, Yitan Zhu, Fangfang Xia, and Rick Stevens. A Systematic Approach to Featurization for Cancer Drug Sensitivity Predictions with Deep Learning. *arXiv*, 2020. _eprint: 2005.00095.

[163] Abhishekh Gupta, Prson Gautam, Krister Wennerberg, and Tero Aittokallio. A normalized drug response metric improves accuracy and consistency of anticancer drug sensitivity quantification in cell-based screening. *Communications Biology*, 3(1):42, 2020.

[164] Evaluating, comparing and expanding models. In *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[165] Maria Maddalena Barbieri. Posterior Predictive Distribution. In *Wiley StatsRef: Statistics Reference Online*, pages 1–6. John Wiley & Sons, Ltd, 2015.

[166] Sumio Watanabe. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[167] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, November 2014.

[168] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27:1413–1432, 2017.

[169] Renate Meyer. Deviance Information Criterion (DIC). In *Wiley StatsRef: Statistics Reference Online*, pages 1–6. John Wiley & Sons, Ltd, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat07878.

[170] M. C. Alley, D. A. Scudiero, A. Monks, M. L. Hursey, M. J. Czerwinski, D. L. Fine, B. J. Abbott, J. G. Mayo, R. H. Shoemaker, and M. R. Boyd. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. *Cancer Research*, 48(3):589–601, February 1988.

[171] Michael R. Boyd and Kenneth D. Paull. Some practical considerations and applications of the national cancer institute in vitro anticancer drug discovery screen. *Drug Development*

Research, 34(2):91–109, 1995.

[172] William C. Reinhold, Margot Sunshine, Hongfang Liu, Sudhir Varma, Kurt W. Kohn, Joel Morris, James Doroshow, and Yves Pommier. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Research*, 72(14):3499–3511, July 2012.

[173] Francesco Iorio, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K. Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S. Gray, Daniel A. Haber, Michael R. Stratton, Cyril H. Benes, Lodewyk F. A. Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J. Garnett. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754, July 2016.

[174] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, 1988.

[175] B. Lambert. Priors. In *A Student's Guide to Bayesian Statistics*. SAGE Publications, 2018.

[176] Vikash Mansinghka, Daniel Selsam, and Yura Perov. Venture: a higher-order probabilistic programming platform with programmable inference, March 2014.

[177] Matthew W. Wheeler, Jose Cortinas, Marc Aerts, Jeffery S. Gift, and J. Allen Davis. Continuous Model Averaging for Benchmark Dose Analysis: Averaging Over Distributional Forms. *Environmetrics*, 33(5):e2728, August 2022.

[178] Keith R. Shockley, Shuva Gupta, Shawn F. Harris, Soumendra N. Lahiri, and Shyamal D. Peddada. Quality Control of Quantitative High Throughput Screening Data. *Frontiers in Genetics*, 10:387, 2019.

[179] J. O. Rawlings and W. W. Cure. The Weibull Function as a Dose-Response Model to Describe Ozone Effects on Crop Yields1. *Crop Science*, 25(5):cropsci1985.0011183X002500050020x, 1985.

[180] Zhongheng Zhang. Parametric regression model for survival data: Weibull regression model as an example. *Annals of Translational Medicine*, 4(24):484, December 2016.

[181] Johan Bacart, Caroline Corbel, Ralf Jockers, Stéphane Bach, and Cyril Couturier. The BRET technology and its application to screening assays. *Biotechnology Journal*, 3(3):311–324, 2008.

[182] Holger Franken, Toby Mathieson, Dorothee Childs, Gavain M A Sweetman, Thilo Werner, Ina Tögel, Carola Doce, Stephan Gade, Marcus Bantscheff, Gerard Drewes, Friedrich B M Reinhard, Wolfgang Huber, and Mikhail M Savitski. Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nature Protocols*, 10(10):1567–1593, October 2015.

[183] S. Loewe. The problem of synergism and antagonism of combined drugs. *Arzneimittel-Forschung*, 3(6):285–290, June 1953.

[184] Wei Zhao, Kris Sachsenmeier, Lanju Zhang, Erin Sult, Robert E. Hollingsworth, and Harry Yang. A New Bliss Independence Model to Analyze Drug Combination Data. *Journal of Biomolecular Screening*, 19(5):817–821, 2014.

[185] Bhagwan Yadav, Krister Wennerberg, Tero Aittokallio, and Jing Tang. Searching for Drug Synergy in Complex Dose–Response Landscapes Using an Interaction Potency Model. *Computational and Structural Biotechnology Journal*, 13:504–513, 2015.

[186] Petr Smirnov, Sisira Kadambat Nair, Farnoosh Abbas-Aghababazadeh, Nikta Feizi, Ian Smith, Trevor J. Pugh, and Benjamin Haibe-Kains. Meta-analysis of preclinical pharmacogenomic studies to discover robust and translatable biomarkers of drug response, October 2022. Pages: 2022.10.22.513279 Section: New Results.

[187] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, December 2022.