INSTITUT POLYTECHNIQUE DE PARIS

NNT : 20XXIPPAXXXX

Thèse de doctorat

On learning mechanistic models from time series data with applications to personalised chronotherapies

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Inria Saclay - INSERM - Institut Curie - l'École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat: Computing, Data and Artificial Intelligence

Thèse présentée et soutenue à Paris, le 18 Février 2022, par

Julien Martinelli

Composition du Jury :

Paul-Henry Cournède Professeur, CentraleSupélec, Université Paris-Saclay	Examinateur
Hervé Isambert Directeur de Recherche, Institut Curie	Rapporteur
Marc Lefranc Professeur, Université de Lille 1	Rapporteur
Céline Rouveirol Professeur, Université Paris 13 Paris-Nord	Examinatrice
Anne-Laure Huber Chargée de Recherche, Centre de Recherche en Cancérologie de Lyon	Examinatrice
Didier Gonze Professeur, Université Libre de Bruxelles	Examinateur
François Fages Directeur de Recherche, Inria Saclay	Directeur de thèse
Annabelle Ballesta Chargée de Recherche INSERM, Institut Curie	Directrice de thèse

Si maintenant, — avec la conviction que l'intuition est la source première de toute évidence, que la vérité absolue consiste uniquement dans un rapport direct ou indirect avec elle, qu'enfin le chemin le plus court est toujours le plus sûr, attendu que la médiation des concepts est exposée à bien des erreurs, — si, avec cette conviction, nous nous tournons vers les mathématiques, telles qu'elles ont été constituées par Euclide, et telles qu'elles sont restées de nos jours, nous ne pouvons nous empêcher de trouver leur méthode étrange, je dirai même absurde. Nous exigeons que toute démonstration logique se ramène à une démonstration intuitive ; les mathématiques, au contraire, se donnent une peine infinie pour détruire l'évidence intuitive, qui leur est propre, et qui d'ailleurs est plus à leur portée, pour lui substituer une évidence logique. C'est absolument, ou plutôt ce devrait être, à nos yeux, comme si quelqu'un se coupait les deux jambes pour marcher avec des béquilles.

Le Monde comme volonté et comme représentation - Arthur Schopenhauer

Abstract

Mathematical modeling of biological processes aims at providing formal representations of complex systems to enable their study, both in a qualitative and quantitative fashion. The need for explainability suggests the recourse to mechanistic models, which explicitly describe molecular interactions. Nevertheless, such models currently rely on the existence of prior knowledge on the underlying reaction network structure. Moreover, their conception remains an art which necessitates creativity combined to multiple interactions with analysis and data fitting tools. This rules out numerous applications conceivable in personalized medicine, and calls for methodological advances towards machine learning of patient-tailored models. This thesis intends to devise algorithms to learn models of dynamical interactions from temporal data, with an emphasis on explainability for the human modeler. Its applications are in the context of personalized chronotherapies, that consist in optimizing drug administration with respect to the patient's biological rhythms over the 24-hour span. Three main themes are explored: mechanistic modeling, network inference and treatment personalization. The first chapter describes the development of the first quantitative mechanistic model of the cellular circadian clock integrating transcriptomic, proteomic and sub-cellular localization data. This model has been successfully connected to a model of cellular pharmacology of an anticancerous drug, irinotecan, achieving personalization of its optimal administration timing. The second chapter introduces a novel protocol for inferring whole-body systemic controls enforced on peripheral clocks. On the long run, this approach will make it possible to integrate individual data collected from wearables for personalized chronotherapies. The third chapter presents a general algorithm to infer reactions with chemical kinetics from time series data.

Résumé

La modélisation mathématique des processus biologiques vise à fournir des représentations formelles de systèmes complexes afin d'en permettre des études qualitative et quantitative. Le besoin d'explicabilité suggère le recours à des modèles mécanistes qui décrivent explicitement les interactions moléculaires. Cependant, l'utilisation de ces derniers est conditionnée à l'existence de connaissances *a priori* sur la structure du réseau de réactions sous-jacent. En outre, leur conception demeure un art qui nécessite créativité et de multiples interactions avec les outils d'analyse et de calibration aux données expérimentales. Cela écarte de nombreuses applications imaginables en médecine de précision personnalisée, et appelle à des développements méthodologiques pour l'automatisation de l'apprentissage de modèles adaptés aux données du patient. Cette thèse participe d'un effort de conception d'algorithmes d'apprentissage de modèles d'interactions dynamiques à partir de données temporelles, avec le souci de l'explicabilité à un modélisateur humain. Elle a pour champ d'application la chronothérapie personnalisée qui vise à administrer les médicaments aux horaires optimaux en fonction des rythmes biologiques du patient sur 24h. Ainsi, trois grands thèmes sont abordés : modélisation mécaniste, inférence de réseaux et personnalisation des traitements. Le premier chapitre décrit le développement du premier modèle mécaniste de l'horloge circadienne cellulaire complètement quantitatif, intégrant des données de transcriptome, proteome et localisation sub-cellulaire. Ce modèle a été connecté avec succès à un modèle de la pharmacologie cellulaire d'un anticancéreux, l'irinotecan, afin d'en personnaliser l'horaire optimal d'administration. Le deuxième chapitre présente un protocole original d'inférence des contrôles systémiques qu'exerce le corps entier sur les horloges des tissus périphériques. Cette approche permettra, à terme, d'intégrer des données individuelles issues d'objets connectés pour la personnalisation des chronothérapies. Le troisième chapitre présente un algorithme général d'inférence de réactions avec cinétiques chimiques à partir de séries temporelles.

Contents

Chapter 1
Introduction

Historically, in most areas of Medicine, including cancer management, patients have been treated following the "one-size-fits-all" paradigm. As such, they received drug protocols that showed acceptable results for the majority of patients presenting a similar diagnosis. This convention has since been questioned by the discovery of large inter-patient variabilities in pathologies, as well as in the dynamics and drug response of healthy and diseased tissues (Weinberg et al., 2016; Stephanou et al., 2018). With this in mind, clinical research communities have shifted the focus towards a more *personalized* and *precise* medicine to improve patient outcomes.

US National Research Councils define personalised medicine as "the tailoring of medical treatment to the individual characteristics of each patient" (Disease et al., 2012). There is a growing number of examples in which this concept is influencing clinical decisions and helping shape healthcare provision (Bates, 2010; Girotti et al., 2016).

While "personalized" and "precision" are used interchangeably in the literature, there is a conceptual distinction between them. Personalized medicine refers to an approach to therapies that considers the patient genetic make-up but with attention to their preferences, beliefs, attitudes, knowledge and social context. On the other hand, precision medicine describes a model for health care delivery that relies heavily on data, analytics, and information, involving mathematical models and algorithms (Ginsburg and Phillips, 2018). These methods have been charged with the task of processing data coming from innovative technologies designed to assess biological features in cell cultures, laboratory animals or patients. Integrating such complex, heterogeneous datasets encompassing genetic, transcriptomic, proteomic and systemic information constitutes a modern challenge for mathematical modeling and statistics.

Systems medicine approaches aim to study these multi-type datasets through the design of patient digital twins (CasymConsortium, 2014; Wolkenhauer et al., 2014). According to several international consortia, this *in silico* version of the patient should be based on mathematical models that represent the detailed physiology of key intracellular pathways driving disease evolution and treatment response.

Mechanistic models are a relevant class of models fulfilling such requirements,

as their variables and parameters carry a physiological meaning, conserved across species. They permit simulations and provide us with explainable predictions, a precious perk in a biomedical setting. We will especially focus on Ordinary Differential Equation (ODE) based models. While parameter estimation is a well-understood step of model building, the identification of its structure is more tedious. Nowadays, it is achieved exclusively on the basis of an extensive review and subsequent summary of the literature by the modeler, taking the form of a Chemical Reaction Network (CRN) for instance.

Two recent advances call into question this view. The first is the automation of biological experiments combined with the increase of quality of experimental measurements. As a result, consequent streams of large-scale biological datasets, possibly involving time dependency, are available. The second is the growing need for a more precise and personalized medicine, which suggests the building of mechanistic and patient-specific models to enlighten our understanding of underlying molecular mechanisms and of inter-patient heterogeneity. One way to address these challenges is through the automation of mechanistic model structure inference from accessible datasets. A main concern of this predicament is then to account for the sparsity exhibited by biological systems (Ouma et al., 2018).

As a case study, this thesis will focus on circadian rhythms and personalized chronotherapies. Circadian rhythms represent the 24-hour patterns that are inherent to most organisms, generated by the Circadian Timing System. The latter is responsible for the adequate allocation of biological and metabolic processes over the 24-hour span (Chen et al., 2007), thus anticipating predictable changes in the environment such as light-dark cycles. This system is composed of a central pacemaker located in the brain, which coordinates most physiological signals towards peripheral clocks, found in each nucleated cell of the body. These signals, also known as systemic regulators, are in the form of biomechanical stresses, temperature cycles, hormonal variations, or nutrient exposure for instance (Ballesta et al., 2017). In turn, the peripheral circadian clocks control the timing of a broad number of cellular and physiological processes, including the cell cycle, energy metabolism, or drug pharmacology (Matsuo et al., 2003; Bass and Takahashi, 2010; Bicker et al., 2020). As a consequence, time-dependent toxicity patterns have been observed in more than 40 drugs in mouse (Lévi et al., 2010). This has led to the conception of chronotherapies, that is, the design of drug timing administration schemes in accordance with the patient's circadian rhythms to improve treatment outcomes. Chronotherapy could be further improved through personalization. One way to implement that is through the monitoring of whole-body regulators, which has been made easy by the rise of wearable sensors. Still, this task necessitates a better understanding of the influences systemic regulators have on the cellular clock.

From the methodological to the applied side, the present introduction emphasized the need for several advances. Motivated by the automation of experiments, approaches for learning mechanistic models, e.g. chemical reaction networks, have started to emerge. In principle, such algorithms could operate without any prior knowledge and feed on time series profiles. Such measurements are generally obtained in a *wild type* setting, in which knock down experiments may not be available. This makes it difficult to observe chemical reactions in an independent manner. This framework is known to raise issues when it comes to learning sparse representations, yet these are ubiquitous in biological systems. Hence, there is a necessity in the design of algorithms for inferring chemical reaction networks from time series data obtained in this context.

On a more applied position, the main challenge is directed towards the personalization of chronotherapies. One way this can be achieved is through wearable technologies, which now allow the monitoring of an increasing number of systemic biomarkers. These have proved to capture circadian variability and contain e.g. core body temperature, food intake, blood pressure, heart rate, saliva cortisol (Skarke et al., 2017). From then on, determining the precise molecular interactions between systemic biomarkers and clock genes is of extensive interest. While the latter are not fully understood, much of the cellular clock machinery itself is reported in the literature. Therefore, a method which leverages this accessible prior knowledge to inform on these undiscovered mechanisms would be relevant. In a matter of accounting for patients individual characteristics, inference should be done in a quantitative manner, so that the strength of the inferred links could be compared between subjects.

Lastly, modeling efforts are to be made at the end of the chain as well in order to connect the cellular circadian clock mechanisms to that of a drug's pharmacology. This would permit the integration of inter-patient clock variability in the design of chronotherapies.

Our contributions lie at the intersection of mechanistic modeling and machine learning, in an effort towards personalized chronotherapies. Our thesis is developed in three chapters, following a background chapter which provides the mathematical concepts that are common to systems biology and systems medicine, and an outline of the biomedical context steaming from circadian rhythms and personalized chronotherapies.

Chapter 3: A mathematical model of the circadian clock and drug pharmacology to optimize irinotecan administration timing in colorectal cancer

The quest for personalized chronotherapies can largely benefit from the development of mechanistic models. In this perspective, using newly available mouse liver circadian datasets, we have designed a novel model of the mammalian circadian

clock based on ordinary differential equations. To the best of our knowledge, it is the first model that has been calibrated on absolutely quantitative data of time-resolved clock gene mRNA and protein amounts. This permits the simulation of species concentrations in mol/L. We then proceed to connect this model with a detailed network of the cellular pharmacology of irinotecan, a widely-used anticancer drug. This combined model was successful in predicting core clock circadian variations and irinotecan cytotoxicity rhythms in two colorectal cancer cell lines, corresponding to the initial and metastatic disease of the same patient. The model therefore enables the derivation of patient-specific toxicity curves from a collection of circadian measurements of clock and irinotecan-related gene mRNAs. Upon sensitivity analysis, our model further sheds light on the potential determinants of cytotoxicity rhythms at the molecular level. On that matter, results point towards the participation of the circadian degradation of CES2 and UGT1A1, irinotecan's activating and deactivating enzymes, thus guiding future experiments and clinical trials.

Scientific production

This chapter is based on: J. Hesse, J. Martinelli, O. Aboumanify, A. Ballesta, A. Relogio, *"A mathematical model of the circadian clock and drug pharmacology to optimize irinotecan administration timing in colorectal cancer"* Computational and Structural Biotechnology Journal, 2021. (Hesse et al., 2021)

Chapter 4: Model learning to identify systemic regulators of the peripheral circadian clock

We now focus on the question of identifying the influences of whole-body biomarkers on the cellular circadian clock, thanks to time-resolved measurements of systemic regulators and core-clock gene mRNAs performed in 4 mouse classes (2 strains, 2 sexes). To do so, we developed an innovative methodology for learning unknown subparts of mechanistic models. Armed with the previously created mechanistic model of the cellular clock as prior knowledge (Chapter 3) and datasets at hand, we investigated the possible action of systemic regulators on the transcription and mRNA degradation of core clock gene expression. To that end, these subparts are isolated from the mechanistic model, leading to an approximation of the systemic control on them. This allows us to formulate the problem in a regression context, with the objective to predict this approximated systemic control from whole-body regulators. Relying on linear models, our results highlight the implication of temperature and feeding cycles on two core-clock genes transcription: *Bmal1* and *Per2*. Significant differences related to sex and genetic background are obtained for the strength of these influences. Overall, our contributions are twofold. First, we propose a novel methodology for identifying unknown subparts

of a mechanistic model while accounting for inter-subject variability. Second, by characterizing the action of whole-body regulators on the cellular clock, we enable the possibility to predict drug chronopharmacology from wearable technologies. Therefore, this paves the way for chronotherapy personalization based on systemic biomarkers monitoring.

Scientific production

This chapter is based on: J. Martinelli, S. Dulong, X. Li, M. Teboul, S. Soliman, F. Lévi, F. Fages, A. Ballesta, *Model learning to identify systemic regulators of the peripheral circadian clock"* Bioinformatics, 2021. (Martinelli et al., 2021)

Chapter 5: Reactmine: an algorithm for inferring biochemical reactions from time series data

Finally, we tackle the more general problem of inferring biochemical reactions with kinetics from time series data, lifting the requirement of complete prior knowledge on the structure of the network to recover. A crucial element of this task is to account for the sparsity of biological systems, manifested by the fact that most species share few interactions, whose mechanisms are well-described by a limited number of kinetic functions. We propose Reactmine, a learning algorithm which infers a set of preponderant reactions at stake in time-resolved data. Reactmine addresses the issue of sparsity by construction, through a sequential inference of the reactions, whose selection rests on a statistical computation over the inferred kinetics. Our algorithm is especially dedicated to the regime where reactions are observed all at a time and are therefore highly correlated, which corresponds to the common situation in Biology and is a source of difficulty for classical network inference algorithms. An evaluation on toy networks, such as a linear chain of reactions, shows that the reactions learned by Reactmine are accurate. Two real-life situations are also considered: cell cycle / circadian clock protein fluorescence videomicroscopy, and learning the controls of systemic factors on peripheral clock gene expression. In both cases, the inferred reactions are biologically meaningful and in agreement with other findings, thus demonstrating the relevance of this approach.

Scientific production

This chapter is based on:

- J. Martinelli, S. Soliman, A. Ballesta, F. Fages *"Reactmine: an algorithm for inferring biochemical reactions from time series data"* In preparation.
- J. Martinelli, J. Grignard, S. Soliman, F. Fages, *"On Inferring Reactions from Data Time Series by a Statistical Learning Greedy Heuristics"* CMSB'19. (Martinelli et al., 2019a)
- J. Martinelli, J. Grignard, S. Soliman, F. Fages, *"A statistical unsupervised learning algorithm for inferring reaction networks from time series data"* ICML'19 Workshop on Computational Biology. (Martinelli et al., 2019b)

Chapter 2

Background

This chapter aims at providing the reader with all necessary material to navigate in this dissertation. We begin with an overview of the mathematical methods spanned by systems biology and systems medicine which will be used in the thesis. These approaches will be detailed and put in perspective with others. A similar treatment will be offered for model learning techniques. Finally, an extensive description of the circadian timing system and chronotherapies will be furnished.

Contents

2.1 Mathematical concepts

The modeling of biological regulation mechanisms can be split in two main trends, quantitative and qualitative modeling. The first relies on Ordinary Differential Equations (ODEs), involving the quantitative expression of the interacting species. The second consists in assuming that the expression of each species is based on several (usually 2) discrete qualitative levels. While the latter may seem like a simplifying abstraction, it has shown to be efficient for specific questions such as the reachability of a particular state. These two types of modeling leverage many frameworks to represent biological systems dynamically, such as boolean networks, bayesian networks, agent-based models, ODEs. We refer the reader to

(Machado et al., 2011) for a detailed review. A model is always designed to answer a biological question. As such, the choice of a precise modeling framework depends on it. Likewise, the predictive power of a model is limited to a particular setup, and fades away if the experimental conditions change significantly. The remainder of this thesis will focus on quantitative modeling, using ODE-based models.

2.1.1 Ordinary differential equations based models

Consider a biological system composed of m species $\mathbf{x} = (x_1, \ldots, x_m)$. ODEs aim at representing the instantaneous rate of change of each species through time. This formulation is well-adapted to the modeling of species concentration, a continuous quantity. It writes as

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) \quad \text{with } \mathbf{x}(0) = \mathbf{x}_0 \tag{2.1}$$

where $\mathbf{f} : \mathbb{R}^m \times \Theta \to \mathbb{R}^m$ is a function called the vector field, differentiable with respect to $\mathbf{x}$. $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ is a parameter vector. ODE-based models are usually employed to describe biological networks including up to several dozens of variables before their optimization becomes computationally challenging.

Equation (2.1) is equivalent to

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}(\mathbf{x}(\tau), \boldsymbol{\theta}) \mathrm{d}\tau \tag{2.2}$$

with $\mathbf{x}_0 = \mathbf{x}(0)$ the initial state of the system. Most of the time, the integral in Equation (2.2) is not tractable, hence a numerical solution is computed, using dedicated solvers.

2.1.2 Chemical reaction networks

In an ODE system, the coordinate functions $f_i, i \in \{1, \ldots, m\}$ of the vector field $\mathbf{f}$ can be any arbitrary function as long as it is differentiable with respect to $\mathbf{x}$. Yet, systems biology has been found to be accurately pictured using only a combination of few rate functions, namely: mass action law kinetics, Michaelis-Menten kinetics and Hill kinetics (Keener and Sneyd, 2009). These rate functions can be associated with a set of reactants and products to define a chemical reaction, such as $s_{R1}R_1 + s_{R2}R_2 \implies s_{P1}P_1$, with s_i being the stoichiometry factor of species i in the reaction. In this sense, they carry a mechanistic explanation.

Definition. *A reaction is a triple (R, P, g), also written $R \overset{g}{\implies} P$. where R (resp. P) is a set of reactants (resp. products) indices and $g : \mathbb{R}^m \to \mathbb{R}_+$ is a rate function over species concentrations. A CRN is a finite set of reactions.*

Definition. *$R \overset{g}{\implies} P$ is said to:*

- *follow mass action law with parameter* k

  $\iff g(x_1 \dots, x_m) = k \prod_{r \in R} x_r^{s_r}$

  s_r *stands for the stoichiometry of element* $r \in R$

- *follow Michaelis-Menten kinetics with parameters* v_{max} *and* K_m

  $\iff g(x_1 \dots, x_m) = v_{\max} \dfrac{x_r}{K_m + x_r}$

- *follow Hill kinetics with parameters* α *and* K_m

  $\iff g(x_1 \dots, x_m) = \dfrac{x_r^\alpha}{K_m^\alpha + x_r^\alpha}$

Hill kinetics and Michaelis-Menten kinetics are defined in the case of a single reactant x_r.

Mass action law states that the rate of spontaneous chemical reactions is directly proportional to the product of the concentrations of the reactants. The factor k captures the information on reactants affinity and on conditions of the reaction such as pressure and temperature (Keener and Sneyd, 2009). Michaelis-Menten kinetics represents the rate of an enzymatic reaction, $v_{\max}$ being the maximum possible rate of production of the product and K_m the concentration of substrate such that the rate is half its maximum. Hill kinetics may reflect several mechanisms, like a protein binding to a ligand, with α quantifying the degree of affinity between both species (Goutelle et al., 2009; Huang and Ferrell, 1996; Gonze and Abou-Jaoudé, 2013) (Fig. 2.1).

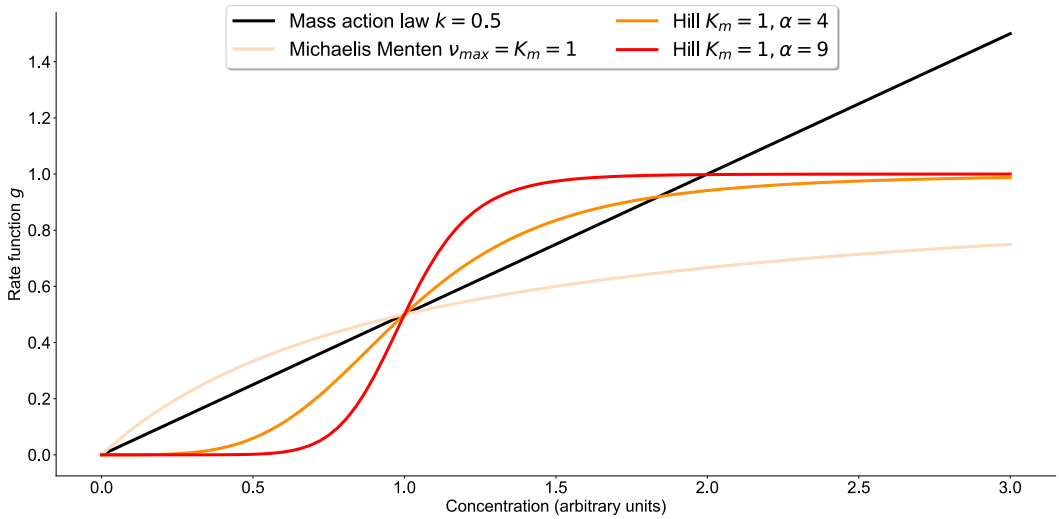


Figure 2.1: Several rate functions plotted against reactant concentration.

CRNs composed with such rate functions define a positive ODE system Fages

et al. (2015). That is, $\forall \mathbf{x}_0 \in \mathbb{R}_+^m$, $\mathbf{x}(t)$ as defined in Equation (2.2) remains non-negative for all $t \geqslant 0$.

Example. *The following ODE system*

$$\begin{cases} \frac{dA}{dt} = -k_1 A \\ \frac{dB}{dt} = k_1 A - k_2 B \\ \frac{dC}{dt} = k_2 B - k_3 C \\ \frac{dD}{dt} = k_3 C - k_4 D \\ \frac{dE}{dt} = k_4 D \end{cases} \tag{2.3}$$

leads to the CRN

$$A \overset{k_1[A]}{\Longrightarrow} B \overset{k_2[B]}{\Longrightarrow} C \overset{k_3[C]}{\Longrightarrow} D \overset{k_4[D]}{\Longrightarrow} E$$

Models encompassed by CRNs are of compelling interest as they provide a mechanistic understanding of the underlying biological processes. This enables the explanation of their predictions. The transient behavior of such CRNs largely depends on the values of their parameters so that the next section focuses on parameter acquisition for ODE-based models.

2.1.3 Parameter estimation for ODEs

Nowadays, the building of classical ODE models is composed of two interconnected components. The first is the design of its structure, which embodies the different regulations present across species, and assigns a parametric form to these regulations, using known kinetic functions. This is achieved by providing the governing equations of the system. The second consists in finding an appropriate parameterization that accurately describes the observations. This part connects the mathematical representation to the reality, as captured by time series data from biological experiments for example.

Let $\mathbf{Y} = (y_{l,i})_{\substack{1 \leqslant l \leqslant n \\ 1 \leqslant i \leqslant m}} \in \mathbb{R}^{n \times m}$ be a collection of measurements potentially corrupted by noise, performed at n different time points $\{t_l\}_{1 \leqslant l \leqslant n}$ over m species. *Model fitting* is the process of selecting a parameterization $\boldsymbol{\theta}$ of the vector field $\mathbf{f}$ and an initial condition $\mathbf{x}_0$ such that evaluating $\mathbf{x}$ (Equation (2.2)) at time points $\{t_l\}_{1 \leqslant l \leqslant n}$ generates a matrix $\hat{\mathbf{X}}(\boldsymbol{\theta}) \in \mathbb{R}^{n \times m}$, close in some sense to $\mathbf{Y}$. The closeness is assessed by a loss function $\mathcal{L}$ which depends on $\boldsymbol{\theta}$, a classical choice being the square loss:

$$\mathcal{L}(\boldsymbol{\theta}) := \|\hat{\mathbf{X}}(\boldsymbol{\theta}) - \mathbf{Y}\|_2^2 \tag{2.4}$$

with $\|\cdot\|_2$ the Frobenius norm:

$$||\mathbf{A}||_2^2 = \sum_{l=1}^{n} \sum_{i=1}^{m} a_{li}^2$$

The objective of the optimization problem is to minimize the loss $\mathcal{L}$ with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\ \mathcal{L}(\boldsymbol{\theta}) \tag{2.5}$$

This problem must often be solved numerically, since the considered ODEs rarely have closed form solution, as mentioned in Section 2.1.1. This leads to an optimization problem that is typically non convex, implying no guarantees about the existence of a unique global minimizer, while potentially presenting many local minima. In this case, $\boldsymbol{\theta}^*$ defined in Equation (2.5) may not be unique, and is therefore fixed to the value returned by the solver.

This issue can be partially alleviated with gradient matching. Instead of fitting the data $\mathbf{Y}$, its estimated derivatives $\hat{\mathbf{V}}$ are computed. From the evaluation of the vector field at data points $\mathbf{Y}$, The matrix $\mathbf{F}(\mathbf{Y}, \boldsymbol{\theta}) = (f_i(\mathbf{Y}_{l,\bullet}, \boldsymbol{\theta}))_{\substack{1\leqslant l\leqslant n\\ 1\leqslant i\leqslant m}}$ is then assembled and fitted to $\hat{\mathbf{V}}$.

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\ \mathcal{LV}(\boldsymbol{\theta}) := \|\mathbf{F}(\mathbf{Y}, \boldsymbol{\theta}) - \hat{\mathbf{V}}\|_2^2 \tag{2.6}$$

$\mathbf{Y}_{l,\bullet}$ (resp. $\mathbf{Y}_{\bullet,i}$) stands for the l^{th} row (resp. i^{th} column) of $\mathbf{Y}$. Free from the need for a numerical integration of the ODE system, the problem can be solved more efficiently from a numerical point of view. Some vector fields, in particular the linear ones, then lead to a convex problem which facilitates the subsequent minimization task.

Example. *For $\mathbf{F}(\mathbf{Y}, \boldsymbol{\theta}) = \mathbf{YA}$, $\mathcal{LV}$ is convex.*

This assumption corresponds to a mass action law system with unimolecular reactions and 0/1 stoichiometry. The mass action law coefficients that constitute the parameter vector $\boldsymbol{\theta}$ are encoded into the matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$.

Indeed, let $h : \mathbf{A} \mapsto \|\mathbf{A}\|_2^2$. h is convex and since $\mathcal{LV}$ is the composition of a convex function with an affine function $\mathbf{z} : \mathbf{A} \mapsto \mathbf{YA} - \hat{\mathbf{V}}$, $\mathcal{LV}$ is convex (Boyd and Vandenberghe, 2004).

Convex problems come as a relief, because any locally optimal point is therefore also globally optimal. This ensures the convergence of optimization algorithms. Nevertheless, gradient matching requires reliable estimates of the observed derivatives $\hat{\mathbf{V}}$, particularly under the presence of noise (Dony et al., 2019). The gain of computational time arising from the absence of the ODE system integration must be appreciated in view of the numerical error that is introduced when estimating $\hat{\mathbf{V}}$ from the real data $\mathbf{Y}$.

Finally, a regularization term is often added to Equation (2.4).

$$\mathcal{L}(\boldsymbol{\theta}) = \|\mathbf{Y} - \hat{\mathbf{X}}(\boldsymbol{\theta})\|_2^2 + \mathcal{R}(\boldsymbol{\theta}) \tag{2.7}$$

$\mathcal{R}(\boldsymbol{\theta})$ can typically be $\lambda\|\boldsymbol{\theta}\|_1$ or $\lambda\|\boldsymbol{\theta}\|_2$. $\lambda \geqslant 0$ is an hyperparameter that controls the trade-off between regularization and least square minimization. This penalization enforces additional constraints on $\boldsymbol{\theta}$ and reduces overfitting. For instance, the ℓ_1-norm is also known to enforce sparsity (Tibshirani, 1996).

Derivative-based minimization

If $\mathcal{L}$ is differentiable, the most classical approach to minimize it with respect to its parameters $\boldsymbol{\theta}$ is to perform gradient descent (GD):

$$\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{(k)}) \tag{2.8}$$

where η is the learning rate. While this is the vanilla setup, state of the art algorithms currently use more complex algorithms featuring adaptive learning rate strategies, stochasticity, coordinate-wise descent, and so on. (Ruder, 2016). For a twice differentiable function $\mathcal{L}$, upon estimation of the hessian matrix $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}$ and if the latter is invertible, Newton updates can also be used.

$$\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} - \eta [\nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{(k)})]^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{(k)}) \tag{2.9}$$

This being said, biological systems may come across as difficult to optimize for at least three reasons. To begin with, the observations are often corrupted by noise which leads to an unreliable estimation of the gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}$. Next, such systems sometimes present stiff behaviors whose numerical integration is complicated and another source of error for the computation of the gradients. Moreover, no guarantees to converge towards a global minimizer can be obtained from GD and Newton algorithms due to the non-convex nature of the loss function in classic systems biology problems. Be that as it may, these derivative-based methods have been successfully used for fitting some particular biological systems (Stapor et al., 2018).

Derivative-free minimization

Another way to minimize $\mathcal{L}$ is to resort to black-box algorithms, which assume that one can only call the function $\mathcal{L}$. No requirement over its differentiability are needed, hence why the term *derivative-free*. The heart of these methods is to implement a sampling scheme of the parameter space Θ that is sounder than the random uniform one. The latter requires a number of samples that is exponential with the dimension of the problem.

In this context, CMA-ES (Covariance Matrix Adaptation Evolutionary Strategy, Hansen and Ostermeier (2001)) has been established as a state of the art algorithm (Hansen et al., 2010). Its application to the optimization of biological systems was fruitful (Rizk et al., 2011). CMA-ES starts from a population of N parameter vectors $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}\}$ sampled according to a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It

then evaluates the function $\mathcal{L}$ for each candidate parameter vectors, sequentially or in parallel. The mean vector μ and its covariance matrix Σ are then updated in the direction of the individuals which led to the lowest value for $\mathcal{L}$, thus giving higher probability to good samples. The next iteration then follows with the sampling of a new generation according to $\mathcal{N}(\mu^{\text{new}}, \Sigma^{\text{new}})$. This goes on until a convergence criterion is met (Fig. 2.2).

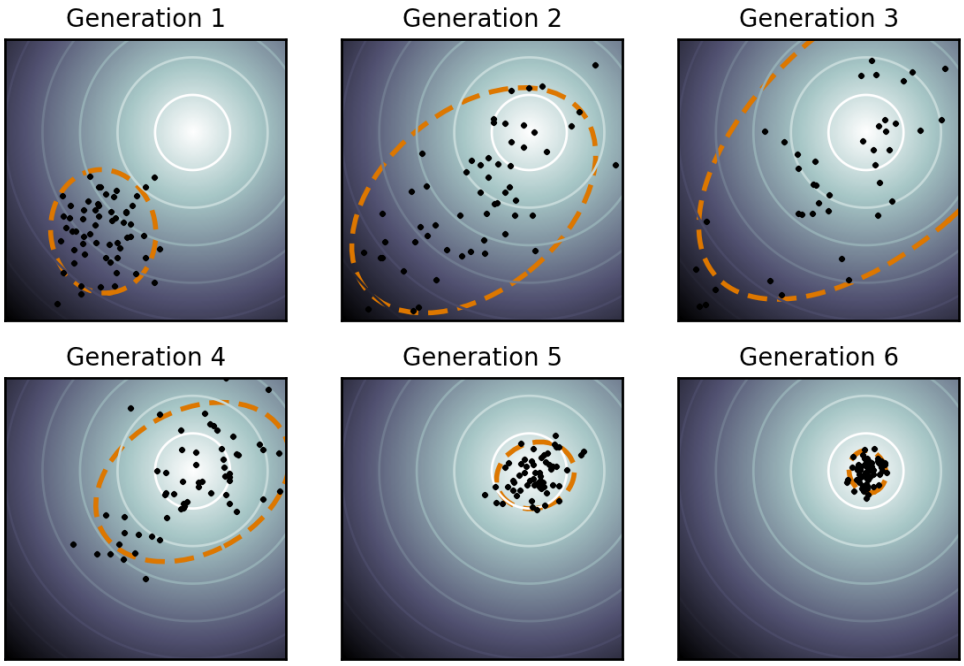


Figure 2.2: Illustration taken from Wikipedia, of an optimization run using CMA-ES on a simple two-dimensional minimization problem. The spherical optimization landscape is depicted with solid lines of equal function values. The population (dots) shows how the multi-variate normal distribution of the population (dotted line) changes during the optimization.

Black-box optimization can also be tackled through the lens of Bayesian Optimization (BO) (Frazier, 2018). Suppose that the black-box function $\mathcal{L}$ has been evaluated at few points $\{\theta^{(1)}, \ldots, \theta^{(N)}\}$. BO begins by building a surrogate function of $\mathcal{L}$, usually thanks to Gaussian Processes (GPs). GPs are ubiquitous in machine learning and can be seen as a generalization of the multivariate normal distribution to infinite dimensional spaces. They enable the possibility to set a prior on function spaces, as a way to embed some belief about the function one wishes to estimate. For example, one may want to sample functions that are only twice differentiable. The reader is referred to (Rasmussen and Williams, 2006) for a thorough analysis of the subject. In particular, based on the already observed evaluations of $\mathcal{L}$, GPs allow for computing the posterior probability distribution. This distribution de-

scribes potential values for $\mathcal{L}(\boldsymbol{\theta}^{(N+1)})$ at candidate point $\boldsymbol{\theta}^{(N+1)}$. Upon selection of this point, $\mathcal{L}$ is indeed evaluated at $\boldsymbol{\theta}^{(N+1)}$, and the posterior probability is updated. This iterative process ends after a sufficient quantity of samples has been generated. The selection of the next point is performed through a so-called acquisition function. The most common one is the expected improvement:

$$\mathrm{EI}_N(\boldsymbol{\theta}) = \mathbb{E}\left[\left[\min_{k\leqslant N} \mathcal{L}\left(\boldsymbol{\theta}^{(k)}\right) - \mathcal{L}(\boldsymbol{\theta})\right]^+ \middle| \left(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(N)}\right), \left(\mathcal{L}\left(\boldsymbol{\theta}^{(1)}\right), \dots, \mathcal{L}\left(\boldsymbol{\theta}^{(N)}\right)\right)\right] \tag{2.10}$$

with $x^+ = \max(x, 0)$ the positive part of x. Equation (2.10) measures the mean improvement between the lowest value found for $\mathcal{L}$ so far, and the potential value coming from the next parameter sampled. The expectation is taken over the posterior distribution given evaluations of $\mathcal{L}$ at $\left\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(N)}\right\}$. Under the assumption of a classical GP as surrogate, this distribution is available in closed form. Jones et al. (1998) further derived a closed-form formula for Equation (2.10), which facilitates the determination of $\boldsymbol{\theta}^{(N+1)}$ as the minimizer of EI_N.

Of note, BO and CMA-ES share similarities, in the sense that BO treats the black-box function $\mathcal{L}$ as random and puts a prior on it, whereas CMA-ES puts a prior on $\boldsymbol{\theta}$, the parameters of the function (Benhamou et al., 2020).

2.1.4 Model analysis

Once the model is fitted, a variety of sanity checks are available. We now detail two of them that will be used during this thesis: parameter identifiability and sensitivity.

Identifiability

An important aspect of a model is whether or not it is identifiable. For clarity, let us write the solution of Equation (2.1) $\mathbf{x}_{\boldsymbol{\theta}}(t)$.

Definition. *The dynamical model defined by Equation (2.1) is said to be structurally identifiable if*

$$\forall \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta, \forall t \geqslant 0, \mathbf{x}_{\boldsymbol{\theta}}(t) = \mathbf{x}_{\tilde{\boldsymbol{\theta}}}(t) \implies \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$$

In other words, the mapping $\boldsymbol{\theta} \mapsto \mathbf{x}_{\boldsymbol{\theta}}$ is one-to-one, which means that only one vector of parameters $\boldsymbol{\theta}$ should be responsible for a particular solution of the system. Should it be wrong, biological inference based on the parameters would lose some of its interest as one would be able to find another set of parameters producing the same behavior. If not all $\boldsymbol{\theta}$ satisfy this definition, the model is said to be locally structurally identifiable. Structural identifiability is independent on the available data and can be examined by dedicated tools based on rigorous mathematical

theory such as DAISY (Differential Algebra for Identifiability of SYstems, Bellu et al. (2007)).

This definition is a first step towards assessing the reliability and predictive power of a model and its parameters, independently of the problem of data availability. However, in practice, the existence of experimental datasets is an important limiting step of model design. That is why another notion of identifiability accounting for the data at hand was invented and is called *practical identifiability*. The latter can be investigated by means of profile likelihood. Assuming identically and independently, normally distributed errors in the data, minimizing the square loss defined in Equation (2.4) is equivalent to maximizing the likelihood. We then have:

Definition. *The profile likelihood of parameter θ_j, the j^{th} coordinate of $\boldsymbol{\theta}$, is*

$$\mathcal{PL}(\theta_j) = \min_{\substack{\boldsymbol{\theta}\in\Theta \\ \theta_j = a}} \mathcal{L}(\boldsymbol{\theta})$$

This constrained estimator can be obtained for a sequence of values $\{a_k\}_{1\leqslant k\leqslant N}$ in the domain of parameter θ_j. This gives rise to a curve describing how the likelihood behaves when one of its parameters is constrained to a given value. Confidence intervals for θ_j can be computed based on Wilks theorem (Wilks, 1938) and will tell us in what range is the parameter identifiable:

$$[\theta_j^-, \theta_j^+] = \{\theta_j | \, \mathcal{PL}(\theta_j) \leqslant \Delta\chi^2_{1,\alpha} + \mathcal{L}(\boldsymbol{\theta}^*)\}$$

with $\Delta\chi^2_{1,\alpha}$ the α-quantile of the chi-square distribution with one degree of freedom, and $\boldsymbol{\theta}^*$ the maximum likelihood estimator.

Definition. *(Raue et al., 2009)*

A parameter θ_j is practically identifiable if the confidence interval $[\theta_j^-, \theta_j^+]$ of its estimate is finite.

If the confidence region $[\theta_j^-, \theta_j^+]$ is infinitely extended and although the loss has a unique minimizer, the parameter is non practically identifiable.

Fig. 2.3 shows an example of both cases. At any rate, a parameter that is non structurally identifiable will never be practically identifiable.

Profile likelihood has proven useful in systems biology as it allows to investigate model identifiability for nonlinear systems from a data-driven point of view (Raue et al., 2009). One should notice that determining the identifiability of parameter θ_j requires to perform N optimizations, for the number of values that θ_j will be fixed to. This can be tedious for large systems, hence Venzon and Moolgavkar (1988) proposed a strategy to select the next value $\theta_j = a$ where the profile likelihood should be computed based on its estimated curvature.

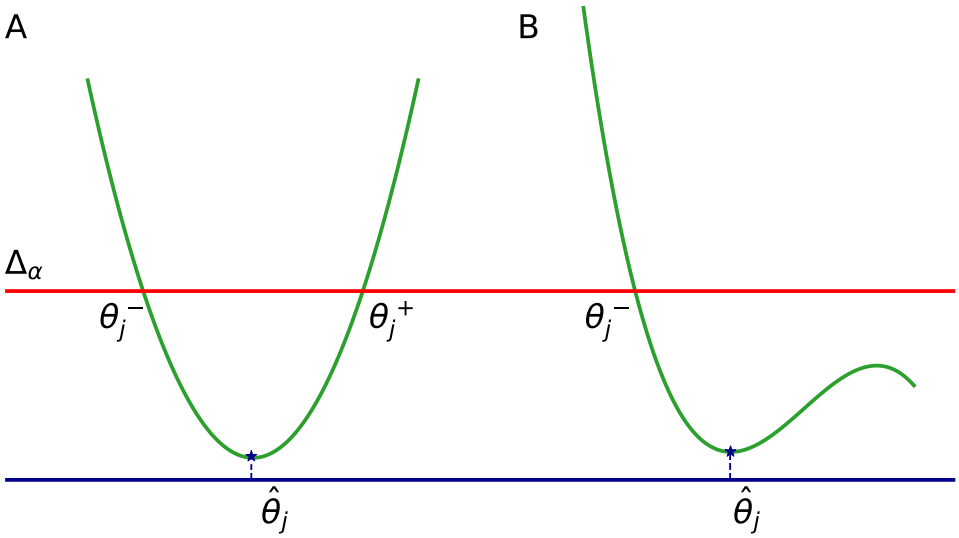


Figure 2.3: Potential likelihood profiles. Panel **(A)** shows an identifiable parameter $\hat{\theta}_j$ with confidence interval $[\theta_j^-, \theta_j^+]$. Panel **(B)** shows a non practically identifiable parameter since the profile only passes the threshold once.

The definition of practical identifiability provided here relied on profile likelihood (Raue et al., 2009; Wieland et al., 2021). Yet it is worth noticing that other definitions of this notion can be found in the literature. One of them is given by Saccomani and Thomaseth (2018), defining a system obtained through least squares minimization as practically identifiable if its sensitivity matrix $\mathbf{S}(\boldsymbol{\theta}) = \left((\partial y_i/\partial\theta_j)\theta_j\right)_{ij}$ is of full rank. The idea is that a model whose output has zero sensitivity with respect to some parameter variations is clearly indicative of non identifiability.

Sensitivity Analysis

Sensitivity analysis allows to quantify parameter influence on a specific output of the model, denoted $h(\boldsymbol{\theta}) \in \mathbb{R}$. Its application can help leverage information in several ways:

1. Understanding the relationship between the input parameters and the specified output
2. Assessing the extent of parameter uncertainty on model output variability
3. Orientating future experimental design to investigate in priority the most sensitive parameters

Sobol sensitivity analysis can be used to that end (Sobol, 2001). As it is a global method, parameters are simultaneously varied. This implies that not only are the direct contributions of each parameter quantified, but the importance of parameter interactions on model output are also computed. The theoretical ground takes

profit of Hoeffding decomposition, considering the parameter vector as a random variable. This decomposition writes an output of the model, h, as a sum of the contribution from each possible combination of parameters.

Let $L^2(\Theta)$ be the space of real-valued, square-integrable functions over Θ

$$L^2(\Theta) := \left\{ f : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}), \int_\Theta f(\boldsymbol{\theta})^2 d\boldsymbol{\theta} < +\infty \right\}$$

endowed with the usual inner product $\langle \cdot, \cdot \rangle$

$$\forall f, g \in L^2(\Theta),\ \langle f, g \rangle = \int_\Theta f(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Theorem. *(Hoeffding, 1948)*
Let $\mathbf{Z} = (Z_1, \ldots, Z_d)$ be independent variables with law $\mathbb{P}_\mathbf{Z}$ and $h : \Theta \to \mathbb{R}$ such that $h(\mathbf{Z}) \in L^2(\mathbb{P}_\mathbf{Z})$. There exists a unique expansion of h of the form

$$h(\mathbf{Z}) = h_0 + \sum_{p \in \mathcal{P}(\{1,\ldots,d\})} h_p(Z_p)$$

with:

$$\begin{aligned} h_0 &= \mathbb{E}[h(\mathbf{Z})] \\ h_p(Z_p) &= \mathbb{E}[h(\mathbf{Z})|Z_p] - \sum_{q \subseteq p} h_q(Z_q) \qquad \forall p \in \mathcal{P}(\{1, \ldots, d\}) \end{aligned}$$

All the h_ps are centered and orthogonal with respect to $L^2(\Theta, \mathbb{P}_\mathbf{Z})$.

As an example, if $p = \{1, 2\}$, $h_p(Z_p)$ corresponds to the contribution of the cooperative effect of Z_1 and Z_2 on model output $h(\mathbf{Z})$. In practice, Sobol sensitivity indices are based on the total variance formula. For random variables W and U, it writes as

$$\mathbb{V}(W) = \mathbb{E}[\mathbb{V}(W|U)] + \mathbb{V}[\mathbb{E}(W|U)] \tag{2.11}$$

This law assumes that the sample space for W is split according to the values of U. Both $\mathbb{V}(W|U)$ and $\mathbb{E}(W|U)$ are random variables. Each realization is done by first drawing U from its distribution, then sampling W from its conditional distribution given $\{U = u\}$. The first term of Equation (2.11) is the expected variance of W averaged over all values of U. From the definition of the condition variance $\mathbb{V}(\cdot|\cdot)$, it follows that $\mathbb{V}(W|U = u)$ is taken with respect to the conditional mean $\mathbb{E}(W|U = u)$. Therefore, this does not take into account the movement of the mean itself, just the variation around each, possibly varying, mean. The second term focuses on the variability of $\mathbb{E}(W|U)$, not just on $\mathbb{E}(W|U = u)$. In this sense, $\mathbb{E}[\mathbb{V}(W|U)]$ can be seen as a measure of the average within-sample variance, and $\mathbb{V}[\mathbb{E}(W|U)]$ as measuring the between-sample variance.

Substituting $h(\mathbf{Z})$ to W and Z_j to U (or $\mathbf{Z}_{\sim j} = (Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots Z_d)$ to U) the

first-order (resp. total-order) Sobol index S_j (resp. S_j^{tot}) associated with Z_j are

$$S_j = \frac{\mathbb{V}[\mathbb{E}(h(\mathbf{Z})|Z_j)]}{\mathbb{V}(h(\mathbf{Z}))} \qquad S_j^{\text{tot}} = 1 - \frac{\mathbb{V}[\mathbb{E}(h(\mathbf{Z})|\mathbf{Z}_{\sim j})]}{\mathbb{V}(h(\mathbf{Z}))}$$

S_j^{tot} measures the contribution of Z_j, accounting for all possible interactions with other parameters. Both indices are obtained by using intensive Monte Carlo simulations.

In the case of ODE-based models, each Monte-Carlo sample $h(\boldsymbol{\theta}^{(k)})$ requires a numerical integration. This leads to a computationally expensive estimation of sensitivity indices. In this respect, the use of metamodels has received large attention in the last few years. These approaches learn a statistical surrogate of the $\boldsymbol{\theta} \mapsto h(\boldsymbol{\theta})$ from a limited number of samples. Then the surrogate can be used to compute the Sobol sensitivity indices in negligible time. Metamodels can be based on Reproducing Kernel Hilbert Spaces (Huet and Taupin, 2017), Gaussian Processes or Polynomial Chaos Expansion, etc. (Le Gratiet et al., 2016).

Otherwise, local sensitivity methods can be used. These strategies compute the model output variation relative to single parameter changes. For instance, Morris screening estimates $\frac{\partial h}{\partial \theta_j}$, characterizing the effect of parameter θ_j on the output h (Morris, 1991). A fewer number of samples is required due to parameter variations being studied independently. Such approaches an be applied as a first approximation, but are not best suited for models involving nonlinear effects and interactions.

All things considered, it is worth noticing that sensitivity analysis becomes a difficult matter as soon as the number of parameters under investigation becomes too large.

2.2 Model learning

The previous section provided the reader with a summary of two important aspects of mechanistic models: parameter estimation and model analysis. These concepts rested on the implicit assumption that the structure of the model was already established. We now consider approaches focusing on learning the structure of a model.

In that matter, a substantial amount of methods have been dedicated to network inference, with the aim to output a model that lacks quantitative information, but recapitulates the different interactions between its components. Such organization is usually represented as a (possibly) directed graph. These techniques have been applied for the search of phosphoproteomic and clinical networks thanks to Answer Set Programming, Information Theory and Constraint Solvers (Ostrowski et al., 2016; Cabeli et al., 2020; Köksal et al., 2018).

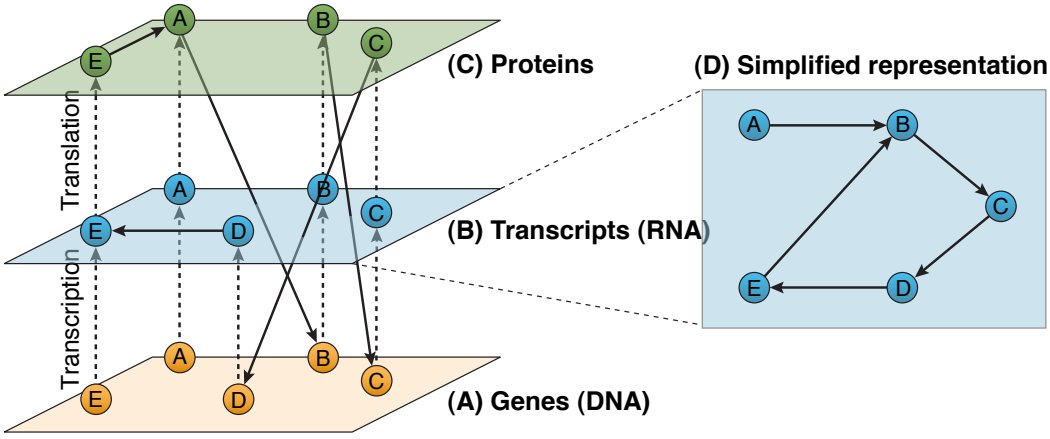


Figure 2.4: Illustrative example of a gene regulatory network. From Marbach (2009).

2.2.1 Inferring gene regulatory networks

One type of network in particular has received most of the attention, Gene Regulatory Networks (GRNs). These are composed of a collection of genes, which may interact through their RNA and protein products (Fig. 2.4). For instance, a gene may code for a protein that can bind to the DNA (a so-called transcription factor), thereby activating or inhibiting the expression of other genes (transcriptional regulation). Proteins also interact among each other, e.g., in Fig. 2.4 protein E binds to protein A, hence changing the regulatory effect of protein A on gene B. As gene expression levels are often measured in terms of mRNA concentrations, a graph is derived, with nodes being associated with mRNAs. The regulatory influences can be thought of as the projection of the different types of regulatory interactions onto the "RNA space" (Panel **(D)**, Fig 2.4) (Marbach, 2009).

The inference of GRN was further motivated by knowledge discovery problems presented in the DREAM challenge series (Stolovitzky et al., 2007). This field or research spans a broad domain of machine learning and statistics, e.g. decision trees (Huynh-Thu et al., 2010), Information Theory (Chan et al., 2017; Margolin et al., 2006) or Correlation Networks (Krumsiek et al., 2011). Furthermore, extensions to time series data have been proposed (Zoppoli et al., 2010; Huynh-Thu and Geurts, 2018; Aalto et al., 2020). By design, none of these approaches provide quantitative insights about the kinetics of the inferred interactions. Indeed, the underlying link between the target and the regulators is obtained in a non-mechanistic manner using e.g. Random Forests or Gaussian Processes. These estimators envelop a widespread class of functions, inducing high predictive power, but at the expense of explainability.

2.2.2 Inferring chemical reaction networks

In this thesis, we would like to focus on a quite new field: inferring chemical reaction networks, rather than GRNs, from time-dependent observations of species at stake. CRNs differ from GRNs as they convey a notion of dynamics of events and further enable the modeling of intricate reactions like complexations $A + B \implies C$. Inferring a CRN therefore requires to entangle both structure identification, that is the reaction network, and the search of an appropriate parameterization.

To that end, the connection between ODE-based models and CRNs allows to apply machine learning methods, originally developed for the discovery of governing equations. SINDy (Sparse Identification of Nonlinear Dynamics, Brunton et al. (2016)) proposes to learn the terms composing each ODE by selecting them among a library of d functions created from the input variables, such as polynomials, cosines, etc. They are stored in a matrix $\boldsymbol{\Theta}(\mathbf{Y}) \in \mathbb{R}^{n\times d}$, with $\mathbf{Y} \in \mathbb{R}^{n\times m}$ the data measurements. An example of such library is

$$\boldsymbol{\Theta}(\mathbf{Y}) = \begin{bmatrix} | & | & & | & | & & | & | & & | \\ 1 & \mathbf{Y}_{\bullet,1} & \dots & \mathbf{Y}_{\bullet,m} & \mathbf{Y}_{\bullet,1}\mathbf{Y}_{\bullet,2} & \dots & \mathbf{Y}_{\bullet,m-1}\mathbf{Y}_{\bullet,m} & \cos(\mathbf{Y}_{\bullet,1}) & \dots & \cos(\mathbf{Y}_{\bullet,m}) \\ | & | & & | & | & & | & | & & | \end{bmatrix}$$

where we remind that $\mathbf{Y}_{\bullet,i}$ stands for the i^{th} column of $\mathbf{Y}$. SINDy is based on the assumption that the dynamics of each variable can be expressed using only a few elements, so that techniques like sparse regression can be used to select the relevant members of the library. This translates into the following mathematical model

$$\dot{\mathbf{V}} = \boldsymbol{\Theta}(\mathbf{Y})\boldsymbol{\Xi} + \boldsymbol{\Sigma} \tag{2.12}$$

which reconstructs the estimated derivatives $\dot{\mathbf{V}} \in \mathbb{R}^{n\times m}$ as a weighted combination of library members, the weights being encompassed in the matrix $\boldsymbol{\Xi} \in \mathbb{R}^{d\times m}$, constituting the parameters of the model. $\boldsymbol{\Sigma}$ is a matrix of independent terms $\Sigma_{l,i} \sim \mathcal{N}(0, \sigma^2)$. Model fitting is performed similarly as the gradient matching approach outlined in Section 2.1.3, minimizing the following function:

$$\boldsymbol{\Xi} = \underset{\boldsymbol{\Xi}\in\mathbb{R}^{d\times m}}{\operatorname{argmin}} \|\dot{\mathbf{V}} - \boldsymbol{\Theta}(\mathbf{Y})\boldsymbol{\Xi}\|_2^2 + \lambda\|\boldsymbol{\Xi}\|_1 \tag{2.13}$$

A ℓ_1-norm penalization term is added to the optimization problem, leading to the Lasso, a popular instance of sparsity enforcing methods (Tibshirani, 1996). With this term, there is hope that spurious functions in the library $\boldsymbol{\Theta}(\mathbf{Y})$, that are irrelevant for the prediction of $\dot{\mathbf{V}}$, will be associated with a zero weight.

In systems biology, this assumption is backed up by the *scale-free* property of most regulatory networks, implying that most species share few interactions, while a small number does (Ouma et al., 2018). Combined with the fact that a limited

number of rate functions are enough to accurately describe biological reactions (Keener and Sneyd, 2009), the intuition of a sparse dynamics within the library of admissible functions is justified. However, the library $\mathbf{\Theta}(\mathbf{Y})$ stems from the input variables, biological species evolving in a dynamical system. Such a setting entails highly correlated predictors, a framework in which Lasso is known to struggle with (Zhao and Yu, 2006). As a consequence, even for a simple chain of reactions like the one used for the simulation in Fig. 2.5, this leads to learned models that fail to provide a sparse embedding of the intrinsic dynamics, despite a R^2 score near 1, as shown on panel **(B)**. The library $\mathbf{\Theta}(\mathbf{Y})$ used in this case contained first and second interactions with a bias term, but no squared term, e.g. $\mathbf{Y}^2_{\bullet,i}$.

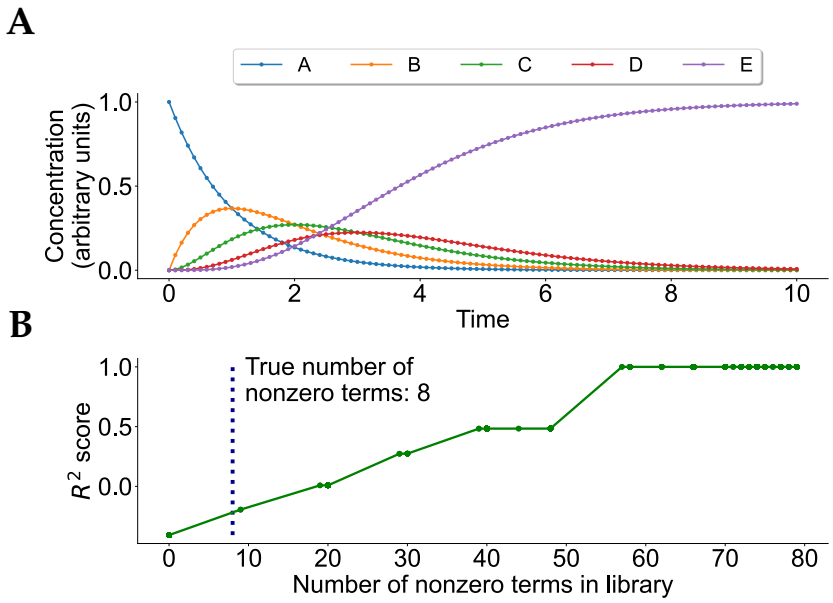


Figure 2.5: **(A)** Numerical simulation of a linear chain of reactions (Equation (2.3)) $A \Longrightarrow B \Longrightarrow C \Longrightarrow D \Longrightarrow E$, each reaction with mass action law coefficient set to 1. **(B)** R^2 score provided by SINDy as a function of the number of terms included in ODEs system, which is directly related to the choice of λ in Equation (2.13).

The reason for the failure of Lasso or other sparse optimizers (Zheng et al., 2019a) to enforce parsimony in this simple example is the high correlation of the predictors due to the parallel action of the reactions. This is investigated in Fig. 2.6. These weaknesses in such cases may be palliated by considering multiple traces obtained from different conditions where species are removed or silenced one by one at the initial time of the experiment. This makes the reactions observable in a more independent manner, produces more informative traces (Carcano et al., 2017), and in effect operates a decorrelation of the predictors.

SINDy's accuracy is deteriorated in a correlated setting

For readibility, let $\mathbf{v} = \hat{\mathbf{V}}_{\bullet,i} \in \mathbb{R}^n$, $\boldsymbol{\xi} = \boldsymbol{\Xi}_{\bullet,i} \in \mathbb{R}^d$ and $\mathbf{Z} = \boldsymbol{\Theta}(\mathbf{Y}) \in \mathbb{R}^{n \times d}$. The SINDy model defined in Equation (2.12) can be written for variable i

$$\mathbf{v} = \mathbf{Z}\boldsymbol{\xi} + \boldsymbol{\varepsilon} \quad \text{with } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$$

$\boldsymbol{\xi}$ is the true parameter. The model is assumed to be "sparse": some of the regression coefficients are exactly zero, corresponding to spurious variables for the prediction of $\mathbf{v}$. Let $J = \{j, \xi_j \neq 0\}$. The Lasso estimator of $\boldsymbol{\xi}$ is

$$\hat{\boldsymbol{\xi}}(\lambda) = \underset{\xi \in \Theta}{\operatorname{argmin}} \|\mathbf{v} - \mathbf{Z}\boldsymbol{\xi}\|_2^2 + \lambda \|\boldsymbol{\xi}\|_1$$

Definition. *An estimator $\hat{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$ is said to be sign consistent if*

$$\mathbb{P}\left(\forall j \in [\![1, d]\!],\ sgn(\hat{\xi}_j) = sgn(\xi_j)\right) \to 1 \text{ as } n \to \infty$$

Theorem. *Assume there exist positive constants M_1, M_2 and c_1, c_2 such that $0 < c_1 < c_2 < 1$ satisfy:*
(i) $\forall j \in [\![1, d]\!]$, $\frac{1}{n}(\mathbf{Z}_{\bullet,j})^T \mathbf{Z}_{\bullet,j} \leqslant M_1$
(ii) The matrix $\frac{1}{n}(\mathbf{Z}^T\mathbf{Z})_{J,J}$ is invertible.
(iii) $card(J) = O(n^{c_1})$
(iv) $n^{\frac{1-c_2}{2}} \min_{j \in J} |\xi_j| \geqslant M_2$
Let us further assume that it exists c_3 such that $0 \leqslant c_3 < c_2 - c_1$ and $m = O(e^{n^{c_3}})$ and that the irrepresentability condition holds: it exists a positive vector η such that

$$|(\mathbf{Z}^T\mathbf{Z})_{J^C,J}((\mathbf{Z}^T\mathbf{Z})_{J,J})^{-1} sgn(\boldsymbol{\xi}_J)| \leqslant \mathbf{1} - \eta \tag{2.14}$$

where the inequality holds component-wise and $\mathbf{1}$ is a $(d - card(J))$-sized vector of 1s.
Then, $\forall \lambda = O(n^{\frac{1+c_4}{2}})$ where $c_3 < c_4 < c_2 - c_1$:

$$\mathbb{P}\left(\forall j \in [\![1, d]\!],\ sgn(\hat{\xi}_j(\lambda)) = sgn(\xi_j)\right) = 1 - o(e^{-n^{c_3}}) \to 1 \text{ as } n \to \infty$$

The first term in Equation (2.14) is the correlation matrix between relevant and spurious functions of the library. The second term recapitulates the correlations between relevant functions, controlling for confounding effects. The bottleneck lies in the fact that these correlations must be low for the inequality to hold so that the theorem can be applied. This is unlikely to happen in a dynamical system, thus SINDy may not find an accurate sparse representation.

Figure 2.6

Having defined the mathematical concepts inherent to mechanistic modeling in systems biology, we now propose to apply them to the personalization of cancer chronotherapy. The next section introduces the biomedical context relative to this research area.

2.3 Circadian translational medicine

2.3.1 The circadian timing system

The Earth's rotation around its axis causes predictable changes in the environment such as light-dark cycles or temperature cycles, thereby providing organisms with options to anticipate them. Most mammalian species have developed a precise time-keeping mechanism that can measure passage of time on an approximately 24 h scale, so-called *Circadian Timing System* (CTS) (Paranjpe and Sharma, 2005).

The CTS is responsible for the adequate distribution of biological and metabolic processes over the 24-hour span (Chen et al., 2007). In mammals, the suprachiasmatic nuclei (SCN), a central pacemaker located in the brain, account for organismal entrainment to the geophysical time, primarily via light cues (top left of Fig. 2.7). The SCN coordinates most physiological signals towards peripheral clocks, placed in each nucleated cell of the body (middle of Fig. 2.7). These signals are in the form of biomechanical stresses, temperature cycles, hormonal variations (e.g., cortisol, melatonin), or nutrient exposure (Ballesta et al., 2017) for instance. Rhythmic behaviors such as feeding patterns (Greenwell et al., 2019), or physical exercise (Wolff and Esser, 2012) also impact the peripheral clocks in an SCN-independent fashion (dashed black arrow in Fig. 2.7).

The cellular circadian clock is a molecular machinery of interconnected transcriptional / translational feedback loops that produces sustained 24 h-oscillations (Ko and Takahashi, 2006). For the discovery of these precise molecular mechanisms, in *Drosophila*, Jeffrey C. Hall, Michael Rosbash and Michael W.Young were awarded the Nobel prize in Physiology or Medicine in 2017 (Lorenz et al., 1989). *Via* the regulation of clock-controlled genes (CCGs), the circadian clock controls the timing of multiple cellular and organismal processes, including the cell division cycle, DNA repair or energy metabolism and the immune system (Matsuo et al., 2003; Bass and Takahashi, 2010; Scheiermann et al., 2013) (right of Fig. 2.7). The disruption of circadian rhythms leads to deregulation in the timing of cellular processes and organ functions, and accumulating evidence points to a negative impact on human health. Several pathologies have been associated with the deregulation of the circadian system including cardiovascular diseases, metabolism disorders and cancer (El-Athman and Relógio, 2018; Ballesta et al., 2017).

From its systemic and genetics components, the CTS displays a great inter-individual heterogeneity. As a matter of fact, substantial interpersonal differences

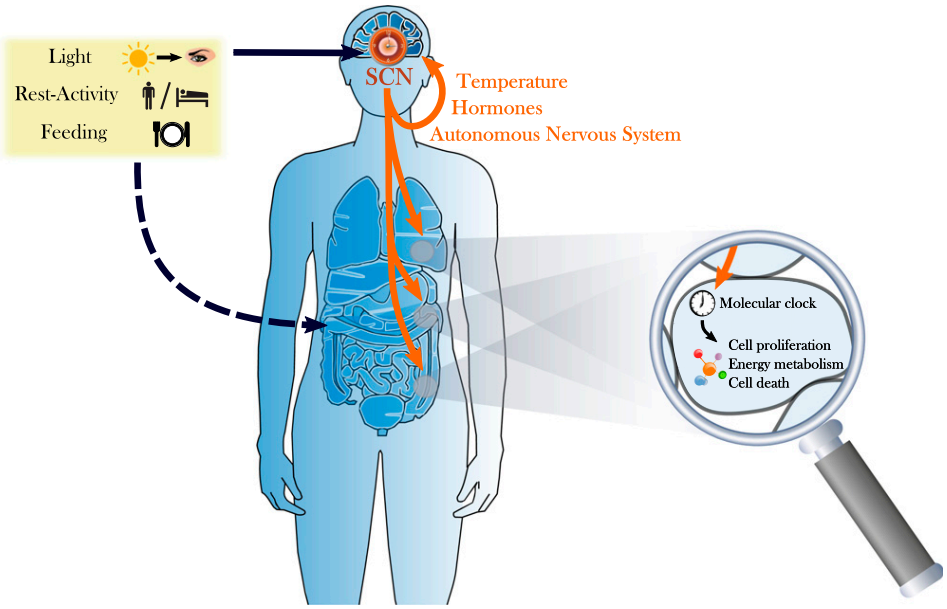


Figure 2.7: The Circadian Timing System. Figure taken from (Ballesta et al., 2017)

have been observed in several endpoints aiming to measure circadian rhythms, from chronotype questionnaires to melatonin onset timing or circadian biomarkers measured by wearables (Kim et al., 2020; Phillips et al., 2019). In a recent study, Komarzynski et al. (2019) examined the variability of core body temperature nadir (lowest level in the 24h cycle), a robust estimate of endogenous circadian phase, among a cohort of 33 healthy participants. The results revealed differences of as much as 7 hours between subjects. Sex also appears as a major determinant of circadian rhythms, as women, in general, have higher-amplitude behavioural rhythms than men and tend to be more resilient to circadian disruption (Anderson and Fitzgerald, 2020). Moreover, the PiCADo mobile eHealth platform identified significant sex- and age-related differences in circadian coordination during daily routine by combining skin temperature with rest-activity measurements, which were not detected from rest-activity data alone (Komarzynski et al., 2018). The literature further presents a variety of clock disparities at the genetic level. For instance, a variant of the human core-clock gene *PER2* has been associated with longer intrinsic circadian period (Chang et al., 2019). Likewise, a gain-of-function variant of *CRY1*, an element of the core-clock repressive feedback loop, has been related to delayed sleep phase disorder (DSPD), a common type of insomnia. This variant reduces the expression of key transcriptional targets and lengthens the period of circadian molecular rhythms, providing a mechanistic link to DSPD symptoms (Patke et al., 2017).

All these manifestations of inter-patient variability in circadian rhythms suggests the use of personalized approaches to treatment.

2.3.2 Chronotherapy

Indeed, by virtue of the implication of the circadian machinery at various levels, most physiological processes involved in the transport and metabolism of xenobiotics are regulated in a time-dependent manner (Fig. 2.8). This impacts the pharmacokinetics (PK) of numerous drugs, that may broadly vary according to the administration timing (Bollinger and Schibler, 2014; Bicker et al., 2020). On the other hand, several drugs impact on clock-regulated genes. Recent findings showed that >50% of the top 100 best-selling drugs in the United States target products of circadian genes (Zhang et al., 2014). Thus, timing drug administration may also impact drug pharmaco-dynamics (PD) and eventually treatment outcome.

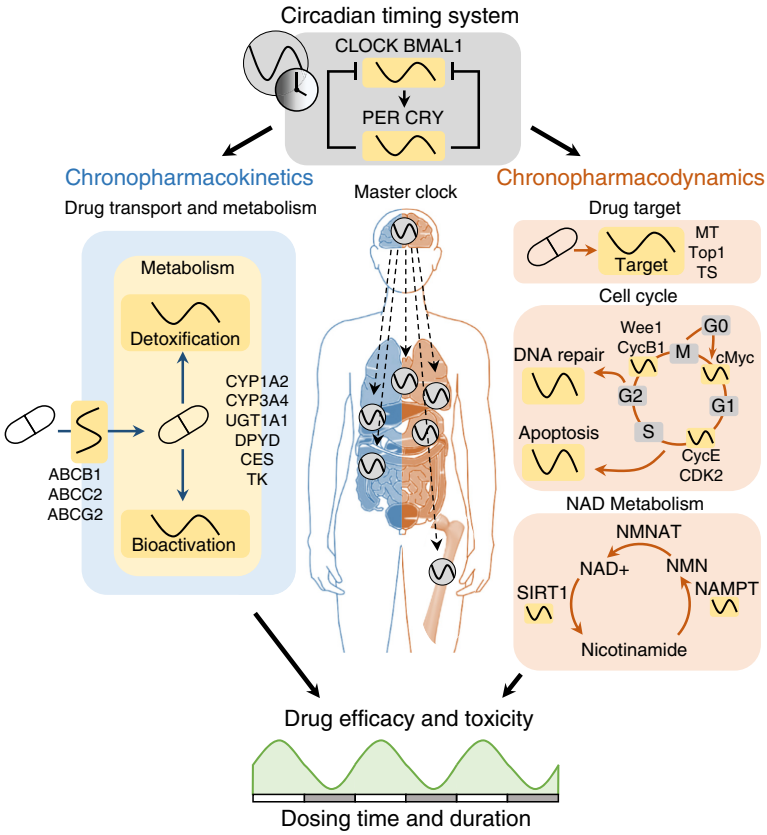


Figure 2.8: The CTS regulates key pharmacokinetic factors (e.g. drug transport and metabolism) and pharmacodynamic factors (e.g. drug targets, cell cycle, and energy metabolism). This leads to a strong dependency of drug efficacy and toxicity on dosing time. Figure taken from (Kim et al., 2020).

Such administration scheduling, designed in accordance with the patient's circadian rhythms to improve treatment outcomes, is also known as *Chronotherapy*. In the 1980's, mouse experiments showed circadian differences in drug toxicity and tissue uptake, inducing vastly heterogeneous survival rates with respect to the time of injection of oxaliplatin, an anticancer compound (Fig. 2.9) (Boughattas et al.,

1989). As an illustration, this study reported an administration time of least toxicity at ZT15 (*Zeitgeiber Time*, 15 hours after light is on), associated with a survival of 75% 40 days after treatment. The worst time of toxicity was found to be at ZT7, with a survival of 25%. This shows that not only the dose makes the poison, but dosing-time as well.

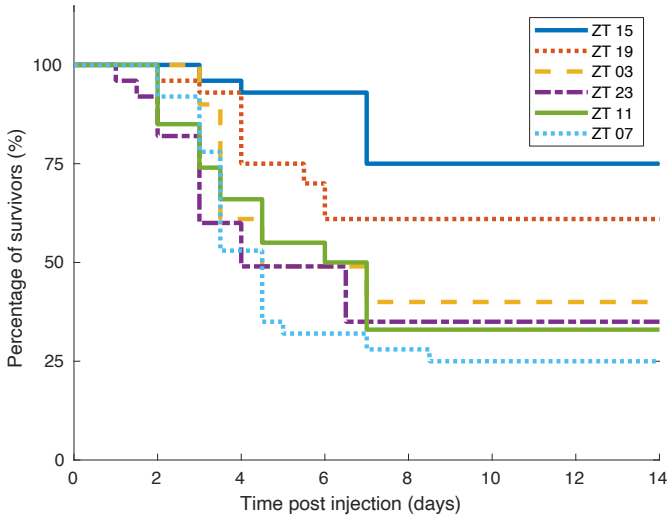


Figure 2.9: Survival curves of male B6D2F1 mice following intravenous administration of 17mg/kg of oxaliplatin at 6 different times of injection. Survival rate differed as a function of injection time. Original data were collected over 40 days (Boughattas et al., 1989). Data shown here were truncated as survival percentage stays constant from day 9 until the end of the experiment.

In the case of irinotecan, another anticancer drug, molecular interactions were investigated to understand circadian influence at the cellular level. Inter-subject heterogeneity was assessed on three different strains of C57BL/6 mouse, showing that for both sexes, there were three chronotoxicity classes with distinct patterns (Li et al., 2013). At the present day, the number of anticancer drugs demonstrating time-dependent toxicity patterns in mouse is beyond 40 (Lévi et al., 2010), more than 10 in clinical studies (Dallmann et al., 2016).

Given the reported alterations in circadian gene expression profiles of cancer cells (Relógio et al., 2014), administering anticancer treatment at a time of least toxicity to healthy tissues is likely to provide a benefit to healthy cells while still targeting the cancer cells. In addition, by timing treatment, it would be possible to increase the tolerated dose, or prevent treatment discontinuation, to achieve a more effective efficacy to the tumor cells (Ballesta et al., 2017).

Nowadays, in the field of cancer management, several clinical studies have addressed the effects of chronotherapy, with promising results (Ballesta et al., 2017; Gaspar et al., 2019; Innominato et al., 2020b; Lévi et al., 2011). Giacchetti and colleagues reported data from three international Phase III clinical trials involving

842 patients (345 females and 497 males) treated with 5-fluorouracil, leucovorin and oxaliplatin administered following either conventional infusions or a particular chronomodulated schedule (Giacchetti et al., 2012). The results showed that male patients lived significantly longer upon this time-sensitive exposure scheme compared to conventional chemotherapy. Yet, while this specific time-dependent administration strategy showed a beneficial trend in males, leading to an increase in overall survival, a decrease was reported in females undergoing this chronomodulated regimen, in comparison to a control group receiving the conventional therapy (Giacchetti et al., 2012). Moreover, a recent international clinical trial concluded that irinotecan hematological and clinical toxicities were lower for early morning administration in male and for early afternoon infusion in female colorectal cancer patients receiving the drug in combination with 5-fluorouracil and oxaliplatin (Fig. 2.10) (Innominato et al., 2020a). Such results highlight the necessity for more

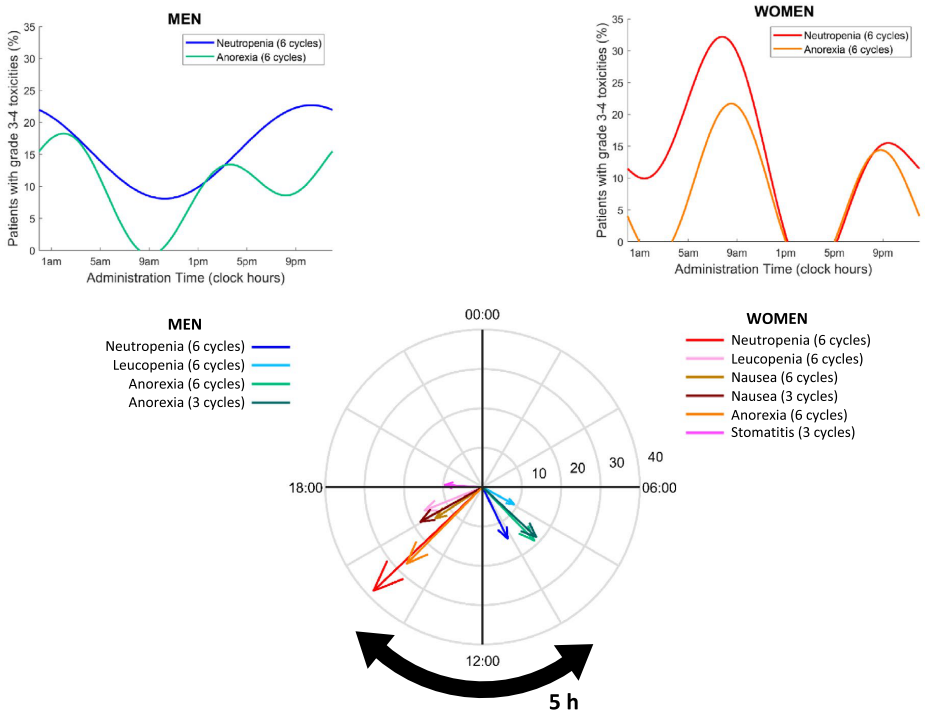


Figure 2.10: Main chronotoxicities according to irinotecan timing and sex. Top: best fit cosine curves for percentages of men and women with grade 3-4 neutropenia or anorexia over six cycles with respect to irinotecan timing. Bottom: polar plots highlighting the sex-specific clock hours associated with minimum incidence of toxicities. For each toxicity, the arrow's length and angle represent amplitude and clock time of minimum value of the best-fit cosine, respectively. Only toxicities with significant cosinor tests are displayed ($P < 0.05$). From (Innominato et al., 2020a).

work in this research area to understand inter-patient circadian discrepancies and enable safe and efficient clinical application of chronotherapy.

Chapter 3

A MATHEMATICAL MODEL OF THE CIRCADIAN CLOCK AND DRUG PHARMACOLOGY TO OPTIMIZE IRINOTECAN ADMINISTRATION TIMING IN COLORECTAL CANCER

Scientific production

The content of this chapter is contained in the article: J. Hesse, J. Martinelli, O. Aboumanify, A. Ballesta, A. Relogio, *"A mathematical model of the circadian clock and drug pharmacology to optimize irinotecan administration timing in colorectal cancer"* Computational and Structural Biotechnology Journal, 166:78 – 97, 2021.

Abstract Scheduling anticancer drug administration over 24 h may critically impact treatment success in a patient-specific manner. Here, we address personalization of treatment timing using a novel mathematical model of irinotecan cellular pharmacokinetics and -dynamics linked to a representation of the core clock and predict treatment toxicity in a colorectal cancer (CRC) cellular model. The mathematical model is fitted to three different scenarios: mouse liver, where the drug metabolism mainly occurs, and two human CRC cell lines representing an *in vitro* experimental system for human CRC progression. The model successfully recapitulates quantitative circadian datasets of mRNA and protein expression together with timing-dependent irinotecan cytotoxicity data. The model also discriminates time-sensitive toxicity between the different cells, suggesting that treatment can be optimized according to their cellular clock. The results show that the chronomodulated degradation of the protein mediating irinotecan activation, as well as an oscillation in the death rate may play an important role in the circadian variations of drug toxicity. In the future this model can be used to support personalized treatment scheduling by predicting optimal drug timing based on the patient's gene expression profile.

Contents

3.1 Introduction

In this chapter, we lay the foundations for basing chronotherapy individualization on the patient's circadian profiles of selected genes including core-clock and drug pharmacology related genes. Several patient-friendly methods for measuring clock gene expression using saliva or blood sampling have been recently validated in the clinics (Hesse et al., 2020). Such patient datasets, combined with mathematical modeling and machine learning, may allow to predict the times of least toxicity to healthy tissues, and optimal antitumor efficacy for an individual patient (Hesse et al., 2020). In particular, computational models representing the chronopharmacology of a specific drug can help to predict therapy time windows of decreased toxicity and optimal efficacy (Ballesta et al., 2017; Hesse et al., 2020). These models can also be optimized for a given patient and used to generate personalized treatment

timing indications. In the past years, several ODE (Ordinary Differential Equation) models have been developed, which either aim to model the circadian clock network (Leloup and Goldbeter, 2003; Forger and Peskin, 2003; Becker-Weimann et al., 2004; Mirsky et al., 2009; Relógio et al., 2011; Kim and Forger, 2012; Furlan et al., 2019; Woller et al., 2016; Almeida et al., 2020) or the biochemical and biophysical interplay between the circadian timing system and a given drug (Ballesta et al., 2011; Dulong et al., 2015). These chronoPK-PD models consider both the impact that the organism has on the drug, i.e. its PK, as well as the impact of the drug on the organism, i.e. its PD, and further include the control of the circadian timing system on these processes.

Currently, there is a gap between existing mathematical models for the core-clock network and mathematical models of drug PK-PD. Here we aimed at merging the cellular clock machinery with a model of the chronoPK-PD of irinotecan, an anticancer drug widely used against digestive malignancies. We generated a new mathematical model, which enables predictions of the cytotoxicity timing for irinotecan having as an input the circadian gene expression of a set of core clock and irinotecan metabolism-related mRNAs. For that, we refined and combined two previously published ODE mathematical models, a core clock from Relógio et al. (2011) and a model of irinotecan chronoPK-PD from Ballesta et al. (2011) and Dulong et al. (2015) which have been successfully used for simulating the mammalian core clock and the time-dependent cytotoxicity of irinotecan, respectively. The core-clock model was refined using newly available quantitative circadian datasets of gene and protein expression in the liver of C57Bl6 male mice. Representing the clock of the liver is important in view of predicting the drug metabolism that mainly occurs in this organ in the whole-body scenario (Mathijssen et al., 2001; de Man et al., 2018). To connect it with the PK-PD model, we extended the transcription-translation network of the core clock with a set of irinotecan-related genes. We fitted this new clock-irinotecan model with transcriptomic data from an *in vitro* colorectal cancer (CRC) experimental progression model and carried out time-dependent irinotecan treatment in both cell lines across 24 h.The CRC *in vitro* progression model includes two cell lines derived from the primary tumor (SW480) and from a metastasis site (SW620) of the same patient, which are known to display different circadian profiles (Fuhr et al., 2018).

Our mathematical model for timing of irinotecan cytotoxicity nicely reproduced mRNA circadian expression, as well as experimental data obtained via longitudinal monitoring of cytotoxicity for both cell lines. In addition, we found that particular parameters associated with *BMAL1* and *CLOCK* showed high impact on drug toxicity emphasizing the relevance of the core clock for irinotecan PK-PD. Finally, we proposed possible candidates for molecular biomarkers of irinotecan chronotherapy, which were the prodrug activation enzyme and the enzyme responsible for deactivation of SN-38, the irinotecan main metabolite.

3.2 Material and methods

3.2.1 Cell culture

SW480 (ATCC© CCL-228™), SW620 (ATCC© CCL-227™) cell lines were maintained in Dulbecco's Modified Eagle Medium (DMEM) low glucose (Lonza, Basel, CH) culture medium supplemented with 10% fetal bovine serum (FBS) (Life technologies, Carlsbad, CA, USA), 1% penicillin–streptomycin (Life technologies), 2 mM Ultragluta-mine (Lonza) and 1% HEPES (Life technologies). Cells were incu-bated at 37°C in a humidified atmosphere with 5% CO2. TheSW480 cell line originated from a surgical specimen of a primary tumour of a moderately differentiated colon adenocarcinoma (Dukes' type B) of a 51-year-old Caucasian male (blood group A,Rh +). The SW620 cell line was derived from a lymph node metastasis (Dukes' type C) taken from the same patient one year later.

3.2.2 shRNA-mediated knockdown

For the knockdown of BMAL1, a TRC lentiviral shRNA glycerol set (Dharmacon, Lafayette, CO, USA) specific for *BMAL1* was used consisting of five individual shRNAs. The construct that gave best knockdown efficiency was determined by gene expression analysis and used for further experiments.

3.2.3 RNA extraction

Total RNA isolation was performed using the RNeasy Mini kit(Qiagen, Venlo, NL) according to the supplier's manual. Medium was discarded and cells were washed twice with PBS and lysed in RLT buffer (Qiagen) prior to the purification procedure. RNA was eluted in 30 μL RNase-free water. Final RNA concentration measurement was performed using a Nanodrop 1000 (ThermoFisher Scientific).

3.2.4 c-DNA and synthesis RT-qPCR

For Real Time quantitative PCR (RT-qPCR) analysis, the extracted RNA was reverse transcribed into cDNA (4 ng/ml) using random hexamers (Eurofins MWG Operon, Huntsville, AL, USA) and Reverse Transcriptase (Life technologies). RT-qPCR was per-formed using SsoAdvanced Universal SYBR Green Supermix (Bio-Rad Laboratories, Hercules, CA, USA) in 96-well plates (see Table1for list of primers used). Human *GAPDH* (QuantiTect Primer, Qia-gen) was used as reference housekeeping gene due to its high abundance and to the lack of circadian oscillations, as confirmed by a cosinor analysis carried out in microarray and RNA-seq data for SW480 cells (Fig. 2.S2d). The qPCR reaction was performed using a CFX Connect Real-Time PCR Detection System (Biorad). Relative gene expression was calculated

using the $2^{-\Delta\Delta C_T}$ method (Livak and Schmittgen, 2001). Biological and technical replicates were included into the analysis.

Primer	Sequence (5′ → 3′)
PER2 forward	AGCCAAGTGAACGAACTGCC
PER2 reverse	GTTTGACCCGCTTGGACTTC
NR1D1 forward	CTCCATCGTCCGCATCAATC
NR1D1 reverse	AACGCACAGCGTCTCG
ARTNL forward	AACCTTCCCACAGCTCACAG
ARNTL reverse	CTCTTTGGGCCACCTTCTCC
TOP1 forward	CCAAGCATAGCAACAGTGAAC
TOP1 reverse	GAGGCTCGAACCTTTTCCTC

Table 3.1: Primers used for the RT-qPCR analysis of SW480 and SW620 cell lines. The primers for mouse can be found in the original publication (Narumi et al., 2016), in Supplementary File S2.

3.2.5 Time-dependent treatment with irinotecan

SW480 and SW620 cells were seeded in 96-well plates at 5000cells and total volume 150mL per well. The cells were synchronized by medium change at 4 different time points (6 h, 12 h, 18 h and 24 h) before treatment with 2mM of irinotecan. Cells (at 60% confluence at the start of measurements) were incubated at 37°C in a humidified atmosphere with 5% CO2. The corresponding untreated control condition was measured in parallel with the treated cells. Cytotoxicity was evaluated in real time with the IncuCyte© S3 Live-Cell system. Abundances of dead cells are measured experimentally as red florescent objects. Cytotox dyes are inert, non-fluorescent and do not enter viable cells, when added to the cell culture. In dying cells, the membrane integrity is lost, the cytotox dye enters the cells and fluorescently labels the nuclei. To prevent dependence on initial conditions, the cytotoxicity curves are shifted along the cytotoxicity axis such that the first value of all curves overlaps with the control curve. The cells are then identified and quantified by the appearance of red labelled nuclei. Because confluency saturated after 84.5 h for the control conditions, the analysis was restricted to 84.5 h, compare with Fig. 2.S7.

3.2.6 Omics data

The models were fitted to microarray time series data of 24 h sampled with an interval of 3 h for the SW480 and SW620 cells and of 48 h sampled with an interval of 2 h for the liver, which was scaled to concentrations based on RNA-seq data. The microarray data and RNA-seq data for liver tissue was published by Zhang *et al.* 2014, accession numbers GSE54650 and GSE54652 (Zhang et al., 2014). For

the SW480 and SW620 cells, the microarray time series data was published by El-Athman *et al.* 2018, accession number E-MTAB-5876 (El-Athman et al., 2018), and the RNA-seq time series data was published by El-Athman *et al.* 2019, accession number E-MTAB-7779 (El-Athman et al., 2019). To relate the microarray data with concentrations, the following steps were done for each gene separately. RNA-seq transcript data was used to calculate the temporal mean of the expression in TPM, which was then converted into mean concentrations in mol/L by a simple rescaling, see Supplementary Information for details. The microarray data was first unlogged (2^{values}) as the data was given in fold change. Then the data was rescaled such that its mean expression matched that of the RNA-seq derived mean concentration, i.e. for a time series of the original microarray data, we used $C \times 2^x / < 2^x >$, where $< . >$ denotes a temporal mean, and C is the concentration calculated for this gene based on the RNA-seq data. For gene families, genes with good oscillations were selected as representative gene for the gene family, as denoted in the figures.

3.2.7 Mathematical models

Model equations and parameters of the core clock are listed in the Supplementary Information, Equations (S1-1–18) and Table 2.S2, model equations and parameters of the clock-irinotecan model are listed in Equations (S1-29–59) and Table 2.S6. Parameter optimization was done using the evolutionary algorithm CMA-ES (Hansen and Ostermeier, 2001). Computations for the core clock were carried out on a laptop with i5 2.9 gHz dual core processor using Python's `pycma` for the optimization and Python's `scipy.integrate.odeint` for the numerical integration (method: lsoda, relative tolerance = absolute tolerance = 10^{-12}). Computations for the clock-irinotecan network were carried out on a compute cluster with the same Python packages. Model fits are restricted to oscillating mRNAs, with a minimum relative amplitude of 5%, i.e. (max–min)/max > 0.05 for each gene expression time series. The fit of the clock-irinotecan network uses the same algorithm and constraints as the core-clock model, see Supplementary Information. The cost function is extended to account for the additional genes in the network. The model variables representing proteins relevant for irinotecan PK-PD, i.e. UGT, CES, ABCB and ABCC, do not regulate other genes' expression within our transcription-translation network and are thus not constrained by the available experimental data used for the model fit. Maximal protein concentration for UGT, CES, ABCB and ABCC are scaled to the maximal concentrations used in the original model by Dulong et al. (2015). Hence, the transcription-translation network predicts toxicity based on the estimated relative amplitude and phase of the protein oscillations, and formerly determined mean absolute levels. The model of Dulong et al. (2015) explicitly involves ABCG2, which is in our case replaced by the dynamical variable ABCC with an appropriate rescaling. The existing model of irinotecan chronoPK-PD uses

protein dynamics as inputs to ultimately predict irinotecan toxicity. We replaced the cosine fit with the dynamics that result from the clock-irinotecan network. For the cell line Caco-2 (cell line derived from a human colorectal adenocarcinoma), the PK-PD model was fitted to cell death fol-lowing irinotecan treatment (Ballesta et al., 2011; Dulong et al., 2015). To fit the circadian variation in toxicity, we change the PK-PD model output by replacing the equation modelling apoptosis with two equations for the time series of alive and dead cells, see Equation (2.S58) and Equation (2.S59). As it turns out, this current model is not sufficient to reproduce the large timing-related differences in cytotoxicity observed experimentally, likely due to the small relative amplitudes of the protein oscillations, which in our model, as defined by Equations (S1-29–47), cannot be larger than the fitted mRNA oscillations. To relax this constraint, we replaced constant protein degradation for UGT, CES, ABCB and ABCC with oscillatory degradation in the final model (Lück et al., 2014). For convenience, acrophases are rescaled to the range from 0 to 1 instead of 0 to 2π.

3.2.8 Statistical analysis

The experimental toxicity profile is fitted by a harmonic regression using Matlab, significance is set to $p < 0.05$ (Hesse et al., 2020). For the cyto-toxicity data, the Area Under the Curve (AUC) is calculated using the linear trapezoidal method, using as weights w_k a vector with n elements (where n is the number of time points considered), with 1 h for the first and last element, and 2 h for the other elements, with the error associated calculated as $var(AUC) = \sum_{k=0}^{n-1} w_k^2 SEM_k^2$ where SEM_k is the standard error at the time point related to time point k, and var is the variance of the AUC calculated as $AUC = \sum_{k=0}^{n-1} w_k x_k$.

3.3 Results

The effect of a drug results from an intricate interplay between its metabolites and the organism, which is under circadian control. Regarding the anticancer agent irinotecan, multiple genes and proteins involved in its PK-PD are directly or indirectly regulated by the cellular core clock. The aim of the study was to design a mathematical model combining the core clock and irinotecan PK-PD-related elements to investigate possible cellular biomarkers predicting irinotecan chronotoxicity rhythms (Fig. 3.1**a**, top). This model was developed and calibrated for three biological systems: the healthy mouse liver, and two cell lines derived from human colorectal cancer (CRC) (Fig. 3.1**a**, bottom). The combined model was trained for the CRC cell lines using circadian datasets of mRNA levels and with experimental results on time-related irinotecan cytotoxicity. We first present a quantitative version of the core-clock model (Fig. 3.1**b**), followed by its extension to account for the clock-controlled regulation of genes involved in irinotecan PK-PD.

Dataset	Organ / Cell line	Used for model	Acquisition technique	Accession number
Narumi et al. (2016)	C57Bl6 male mouse liver (WT mouse and *Bmal1*$^{-/-}$)	Liver core-clock	mRNA: RT-qPCR proteins: mass spectrometry	NA
Zhang et al. (2014)	C57Bl6 male mouse liver	Liver transcription -translation network	mRNA: microarray and RNA-seq	GSE54650 and GSE54652
Wang et al. (2017)	C57Bl6 male mouse liver	Liver core-clock	Nuclear proteins: mass spectrometry	E-MTAB-5876
El-Athman et al. (2018)	SW480 / SW620 human CRC cell lines	CRC core-clock CRC full model	mRNA: microarray	E-MTAB-7779
El-Athman et al. (2019)	SW480 / SW620 human CRC cell lines	CRC core-clock CRC full model	mRNA: RNA-seq	E-MTAB-7779
In this publication	SW480 / SW620 control and siRNA *Bmal1*	CRC core-clock CRC full model	mRNA: RT-qPCR	NA
Dulong et al. (2015) Ballesta et al. (2011)	Caco-2 human colorectal cancer cell line	CRC full model	CPT11 and SN38 cellular PK:HPLC TOP1 activity: DotBlot	NA
In this publication	SW480 / SW620	CRC full model	CPT11 cytotoxicity	NA
Zheng et al. (2019b)	C57Bl6 male mouse liver	Liver core-clock	Cytoplasmic and nuclear proteins CLOCK and BMAL abundance: immunoprecipitation	NA
Aryal et al. (2017)	C57Bl6 male mouse liver	Liver core-clock	Cytoplasmic proteins PER and CRY: immunodepletion	NA
Schwanhäusser et al. (2011)	Mouse fibroblasts NIH3T3	Liver core-clock	mRNA transcription rates: RNA-seq	SRA030871

Table 3.2: Resources for the data used in the current study.

Finally, this model was connected to a representation of irinotecan chronoPK-PD.

3.3.1 A quantitative model of the core clock in mouse liver

To investigate the interactions between the circadian clock and irinotecan cellular PK-PD, we started by designing a quantitative model of the cellular core clock (Fig. 3.1**b**). We refined the previously published ODE model by Relógio et al. (2011), which represents the molecular mechanisms of the core clock at the cellular level based on experimental data for the mammalian SCN. Clock gene paralogs and isoforms were merged into the following model variables for mRNA elements: *Per* (*Per1, Per2, Per3*), *Cry* (*Cry1, Cry2*), *Ror* (*Rora, Rorb, Rorc*), *Rev-Erb* (*Rev-Erba, Rev-Erbb*) and *Bmal1*. We applied the same principle for model variables representing proteins and protein complexes. The dynamical variable CLOCK/BMAL representing the CLOCK/BMAL1 dimer is assumed to activate the transcription of the core-clock genes *Rev-Erb, Ror, Per* and *Cry* and the PER/CRY complex to inhibit this

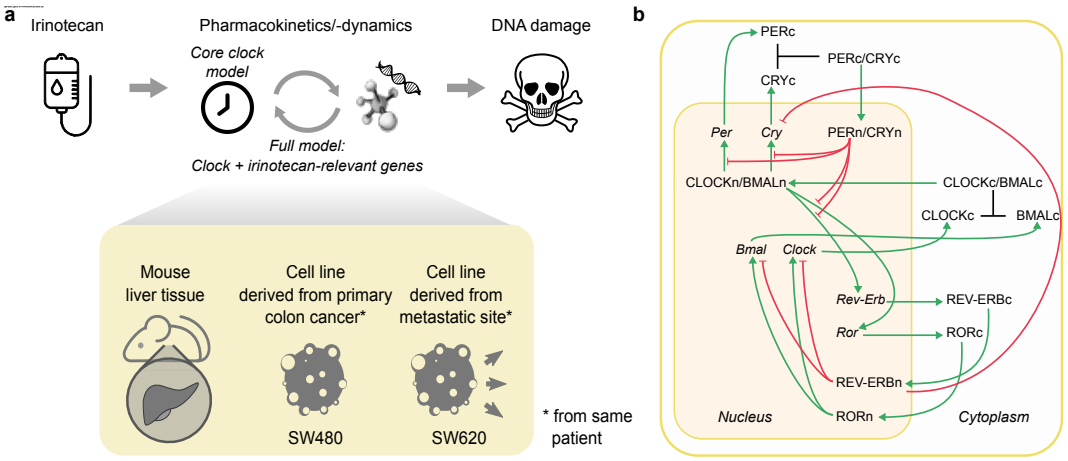


Figure 3.1: The action of the drug irinotecan involves the core clock and a set of clock-regulated genes, experimentally assessable in different cell types. a Workflow of the clock-irinotecan model construction. Irinotecan induces DNA damage and potentially cell death via its interaction with clock-controlled proteins. Mathematical models were fitted to different datasets in healthy mouse liver, and in human cancer cell lines. **b** Network representation of the core-clock model. Inhibitory interactions are presented in red with flat arrowheads, activating interactions in green with pointed arrowheads, and complex formation in black.

transcriptional activity. The model includes two main negative feedback loops. The first one involves the self-inhibition of the dynamical variables *Per* and *Cry* through the inhibition of CLOCK/BMAL by the PER/CRY complex. In addition, REV-ERB inhibits the transcription of *Cry*, thus inhibiting its own inhibition through the regulation of PER/CRY. The second feedback loop is induced by the self-repression of the dynamical variable *Bmal* through the activation of its repressor REV-ERB by CLOCK/BMAL. On the contrary, ROR, which is transcriptionally activated via CLOCK/BMAL, acts positively on *Bmal* regulation.

The Relógio et al. (2011) model differentiated between phosphorylated and unphosphorylated PER proteins. However, in the absence of time-dependent quantitative data on PER phosphorylation, we opted to simplify the PER/CRY (PC) loop and to merge the phosphorylated/unphosphorylated variables (Fig. 3.1**b**). Similarly, the equations for the dynamical variables CLOCK/BMAL and PER/CRY cytoplasmic complexes originally included both a term for complex dissociation into free proteins and for complex degradation, which were not identifiable from the available data so that the degradation terms were removed (Supplementary Information, Section S2-1.1).

We further refined the core-clock model to represent the core clock in organs relevant for irinotecan pharmacology, in particular the liver, where the drug is processed. The Relógio *et al.* model did not explicitly consider *Clock* given its lack of rhythmicity in the SCN (Reppert and Weaver, 2001). However, this is not the case in

the liver (Narumi et al., 2016). Moreover, CLOCK/BMAL1 is a key transcriptional regulator of genes involved in the irinotecan network (Yang et al., 2009; Oishi et al., 2005). Thus, we expanded the initial model by explicitly including *Clock* as follows. Similarly to the dynamical variable *Bmal*, *Clock* transcription is assumed to be positively regulated by ROR and negatively impacted by REV-ERB (Preitner et al., 2002). The cytoplasmic protein $CLOCK_C$ dimerizes with the dynamical variable $BMAL_C$ and translocates to the nucleus to form the heterodimer complex $CLOCK/BMAL_N$. The dynamical variable $CLOCK_C$ representing the cytosolic CLOCK protein is assumed not to be able to enter the nucleus, as it was not detected in the nucleus of cells not expressing *Bmal1* (Zheng et al., 2019b). Of note, BMAL1 and CLOCK nuclear protein expressions shared the same circadian phase and amplitude experimentally, suggesting that both species exist mostly as dimers in the nucleus (Wang et al., 2017) (Fig 3.1**b**). One last modification was made to the model structure to increase the accuracy of cytoplasm/nucleus transport terms. The equations now account for the ratio between the compartment volumes to ensure that the quantity of matter is conserved during transport (Supplementary Information, Section S2-1.1). The cytoplasm/nucleus volume ratio was set for mouse hepatocytes to 14 (Peters, 1984).

Our new core-clock model allows to quantitatively simulate gene and protein levels, expressed in mol/L, thus allowing for the fitting of quantitative datasets informing on absolute concentrations (Table 3.2. Parameter estimation was done using the time series data reported by Narumi *et al.* (Narumi et al., 2016). Starting from the Relogio *et al.* model, we performed a linear change of variables, mapping the original model variables to their scaled versions with respect to the maximum of the observed data (Supplementary Information, Section S2-1.3). The obtained scaled parameter values were then used as an initial guess for the subsequent parameter estimation procedures.

The liver dataset also included protein expression for $Bmal1^{-/-}$ mice (Narumi et al., 2016). Assuming that *Bmal1* knockout (KO) led to a loss of oscillations in the clock (Shimba et al., 2011), this data could be seen as a glance at the system at steady state. This enabled us to derive functional relationships to compute three transcription rate parameters as a function of the KO mice data and other parameters (Supplementary Information, Section S2-1.3), thus decreasing the number of parameters to estimate. This led to a simplification in the parameter estimation. We further reduced the number of parameters by assuming that Hill power coefficients were equal for all activators (parameter b) and all inhibitors (parameter c) of the transcription across genes. This led to a decrease of 8 parameters to be estimated while producing next to no change in the goodness of fit as expected from the argument of unidentifiability. Only *Cry* kept separated Hill coefficients due to its transcription being regulated by 3 species (the dynamical variables CLOCK/BMAL, PER/CRY and REV-ERB). The final core-clock model has 18 state variables and 58

parameters to be estimated.

The parameter estimation procedures consisted in a numerical minimization of a cost function, which was the sum of two terms (Equation (2.S28)). The first term is the least square error between the data and the model's simulation, while the second term accounts for biological constraints. These constraints were derived from co-immunoprecipitation experiments and provided bounds for complex concentrations with respect to free protein concentrations (Zheng et al., 2019b; Aryal et al., 2017). Additional constraints were specified on the bounds of parameter search intervals including those of degradation or transcription rates based on mRNA and protein half-lives and levels (Schwanhäusser et al., 2011). Fig 3.2 shows the model best-fit, which convincingly reproduced the data ($R^2 = 0.86$). *Bmal1* and *Clock* mRNA model-predicted profiles presented a similar phase but different mean levels (5.3 and 19.5 pmol/L, respectively) and relative amplitude (84% and 62% of the mean, respectively). Differences were also observed at the protein level as free BMAL1 and CLOCK protein mean levels were equal to 13.9 and 8.85 nmol/L respectively, with relative amplitudes of 35% and 25%. These differences came as a justification to the addition of the *Clock* gene into the core-clock model.

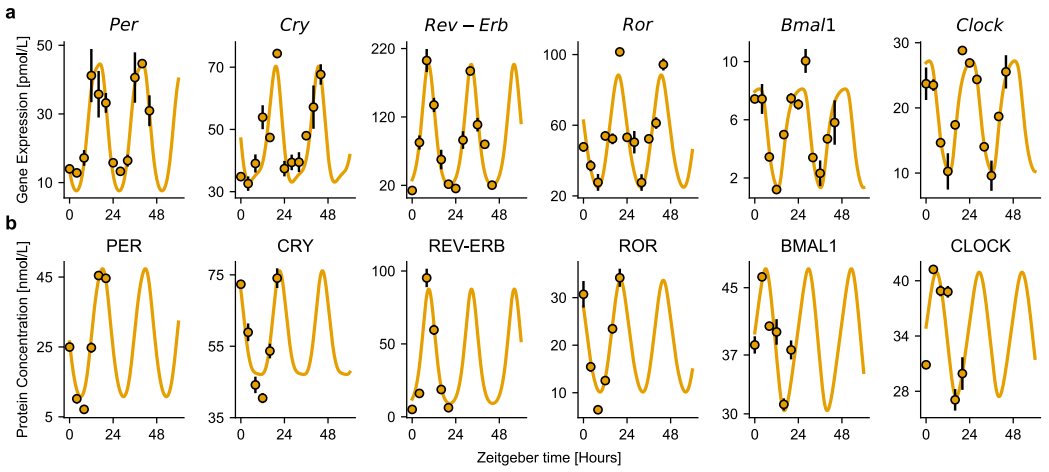


Figure 3.2: Best fit of the quantitative core-clock model to mRNA and protein circadian datasets in the mouse liver. a mRNA expression for core-clock elements in pmol/L. **b** Protein levels for core-clock elements in pmol/L. Model simulation (orange lines), experimental data used for calibration (black circles). Depicted are mean values ($n = 2$ biological replicates) $\pm$ SEM.

Validation of the model was done using an external time course dataset from mouse hepatocytes, which was not used for the model design and calibration Wang et al. (2017). This study reports a phase between 8.5 h and 10.8 h for the circadian rhythm of REV-ERB nuclear expression and a relative amplitude of 98%, while the model simulation for phase and relative amplitude were 9.7 h and 90%, respectively. Similarly, for ROR nuclear expression, the reported phase was 20.8 h, as compared

to 21.1 h for model predictions, and its relative amplitude was 80% as compared to 69% for the model. Both predictions are in close agreement with the study and serve as a validation of the model. For the other clock proteins, as this model only tracked them as complexes in the nucleus, the comparison to single-protein data was not possible. A subsequent robustness analysis was performed by analyzing whether the model could maintain sustained oscillations upon parameter perturbation. Gaussian noise was added to the best-fit parameter vector with a standard deviation of 10%, leading to oscillating simulations in 73% of the cases, thus demonstrating the model robustness (Fig. 2.S1).

3.3.2 The clock model reproduces the expression profiles of core-clock genes in CRC cell lines

To test our mathematical model in a colon cancer context, we chose a well-known *in vitro* cellular model of CRC progression, which includes two cell lines from the same patient (SW480, SW620), derived from the primary tumour and from the metastasis, respectively. We carried out a time course of 45 h (9 h - 54 h, after synchronization) with a 3 h sampling interval, for the gene expression analysis of *PER2*, *REV-ERBα* and *BMAL1 via* RT-qPCR, in either control or *shBMAL1* conditions. This dataset was combined with microarray data for the expression of *Cry*, *Ror* and *Clock* in order to calibrate the core-clock model for each of the CRC cell lines (see Section 2.6). Quantitative mean concentration levels expressed in mol/L for the clock genes of CRC cell lines were derived from an already published RNA-seq transcriptomic dataset (Zhang et al., 2014). Thus, in total, three datasets were combined for the calibration of the core-clock model for the CRC cell lines. The cytoplasm/nucleus volume ratio of the CRC cell lines was set to 5 (manual curation, using snapshots of SW480 and S620 cells from Abdulrehman *et al.* (Abdulrehman et al., 2018) Fig. 2, the cytoplasm/nucleus ratio was computed for each cell of the figure and an average value close to 5 was found for both cell lines). The transcription-translation network was assumed to be similar in either control or *shBMAL1* conditions, yet with a single different parameter to account for *shBMAL1* activity. Accordingly, the RT-qPCR datasets obtained from the control and *shBMAL1* conditions were fitted simultaneously to their respective models using the same set of parameter values with the exception of *BMAL1* basal transcription, which was allowed to differ between both conditions (Fig. 2.S3) The *shBMAL1* condition provided a view of a dampened circadian clock, due to the knockdown of *Bmal1*, which induced a lower activation power of the transcription factor CLOCK/BMAL1. Upon model calibration, a 375-fold reduction in the *BMAL1* estimated basal transcription rate was necessary to allow for a good fit of both conditions. This demonstrates the ability of the model to reproduce two different physiological scenarios by tuning a single parameter.

Concerning the SW480 cell line, the model achieved an excellent fit of the data (R^2 = 0.75) (Fig. 3.3**a**, Fig. 2.S2**a**). The fit for the SW620 was reasonable as well (R^2 = 0.67), but lacked a proper fit of *CRY, ROR* and *CLOCK* expression reported in the microarray dataset (Fig. 3.3**b**, Fig. 2.S2**b**). Oscillations of the core-clock genes, normalized to the mesor, showed large rrelative amplitudes in the healthy mouse liver than in CRC-derived cell lines, with the circadian rhythms in SW620 cells being largely dampened as compared to both other systems (Fig. 3.3**c**). The peaks of *BMAL1* mRNA levels of the best-fit model for mouse liver and SW480 cells were aligned to allow for an *in vitro*/*in vivo* systems comparison. This highlighted a moderate phase shift of 5 h (respectively 1 h) for *PER2* (respectively *REV-ERBα*) between the SW480 cell line and liver tissue. On the opposite, larger phase delays were observed in the case of the SW620 cell line. Although the three models represent different organs in different conditions, their comparison exhibited a moderate agreement between the clocks of the healthy liver and of the SW480 colorectal cancer cell line, and large differences in terms of oscillations dampening and phase differences in comparison to the clock of the SW620 metastatic colon cancer cell line.

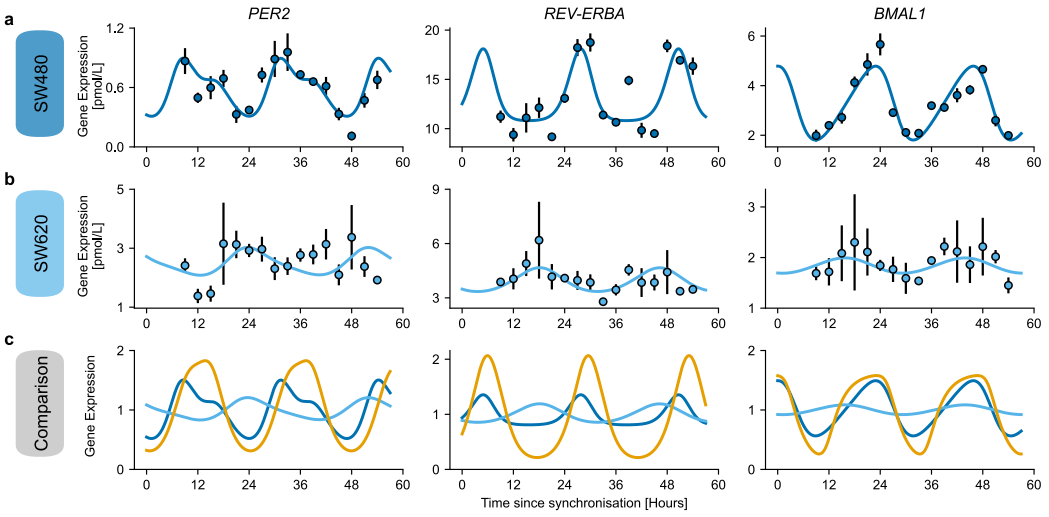


Figure 3.3: Comparison of the core-clock models fitted to healthy mouse liver or human cancer cell lines. Best fit of the quantitative core-clock model to **(a)** the SW480 cell line and **(b)** the SW620 cell line. Model simulation (line) against the RT-qPCR data used for calibration (dots), depicted as mean values (n = 3) $\pm$ SEM. **c** Comparison of the model fit for liver (orange), SW480 (dark blue) and SW620 (sky blue). *Bmal1* circadian phases were aligned for mouse liver and SW480 cell line and all gene expression profiles were normalized to the mesor to allow for comparison. See Fig. 2.S2 for the other genes of the core clock model.

For most core-clock genes, the oscillations displayed a non-cosine shape, with different intervals of high versus low gene expression, see Fig. 3.3**c** and Fig.2.S2**c**.

Overall, the here presented core-clock model, based on cellular mRNA and protein concentrations, reproduced the circadian gene expression profiles for different sets of experimental data with good precision. Thus,this model provided a reasonable starting point for the following extension with irinotecan PK-PD-related genes.

3.3.3 Filling the gap: Connecting the core clock with irinotecan PK-PD related genes

We extended the core-clock network with eight clock-controlled genes relevant for irinotecan pharmacology as depicted in Fig. 3.4, named in the following clock-irinotecan network. The elements added to the core-clock model are involved in irinotecan metabolism, transport, and pharmacodynamics. Irinotecan is a prodrug, which needs to be converted into its active metabolite, SN-38, through the enzymatic activity of CES2 (Carboxylesterase2) (Xu et al., 2002). Subsequently, UGT1A1 (uridine diphosphate glucuronosyl-transferase 1A1) regulates the conversion of SN-38 into its inactivated form, SN-38G (Mathijssen et al., 2001). The ATP-Binding Cassette (ABC) transporters ABCB1, ABCC1, ABCC2 and ABCG2 control the efflux of these molecules out of the cell (Mathijssen et al., 2003). Central to irinotecan action, SN-38 binds to the protein TOP1 (DNA topoisomerase 1), which under normal conditions releases the supercoiling and torsional tension of DNA by transiently cleaving and rejoining one strand of the DNA, thereby controlling DNA topology during replication and transcription. SN-38 binds to DNA-TOP1 complexes to stabilize them. This leads to double-stranded breaks, erroneous transcription and likely cell death (Pommier, 2006; Smith et al., 2006). Besides these proteins directly relevant for irinotecan PK-PD, the clock-irinotecan network contains three elements that act as transcription factors for the above-mentioned genes, DBP (D site of albumin promoter (albumin D-box) binding protein), which is also considered as a core-clock element, NFIL3 (Nuclear factor, interleukin 3 regulated), and PPARα (Peroxisome proliferator-activated receptor alpha), a regulator of liver lipid metabolism that also acts as transcription factor for UGT1A1, which deactivates SN38 (Ciotti et al., 1999; Ueda et al., 2005; Gascoyne et al., 2009). For the clock-irinotecan network, we only consider the ABC transporter ABCB1 for irinotecan efflux and ABCC1 for SN38 and SN38G efflux as *ABCC2* showed less clear circadian oscillations and *ABCG2* did not appear in our RNA-seq data for the studied cell lines. The mathematical description of the clock-irinotecan network contains 39 equations and 115 parameters (Table 2.S6, Equations S1-29–47). In the clock-irinotecan network, oscillations are inherited from the core clock to genes outside of the core-clock network via the CLOCK/BMAL1 complex, ROR and REV-ERB. Accordingly, most connections go from the core clock to the remaining elements. Only the inhibition of *REV-ERB* by NFIL3 and the inhibition of *BMAL1* by TOP1 provides feedback from the irinotecan-related genes to the core clock (Onishi and Kawano, 2012; Zhao et al.,

2018). From the clock-irinotecan network fitted to experimental mRNA expression data, we use the mRNA dynamics of *UGT1A1*, *CES2*, *ABCB1* and *ABCC1* as an input to the protein dynamics and the PK-PD model to predict irinotecan toxicity, see below.

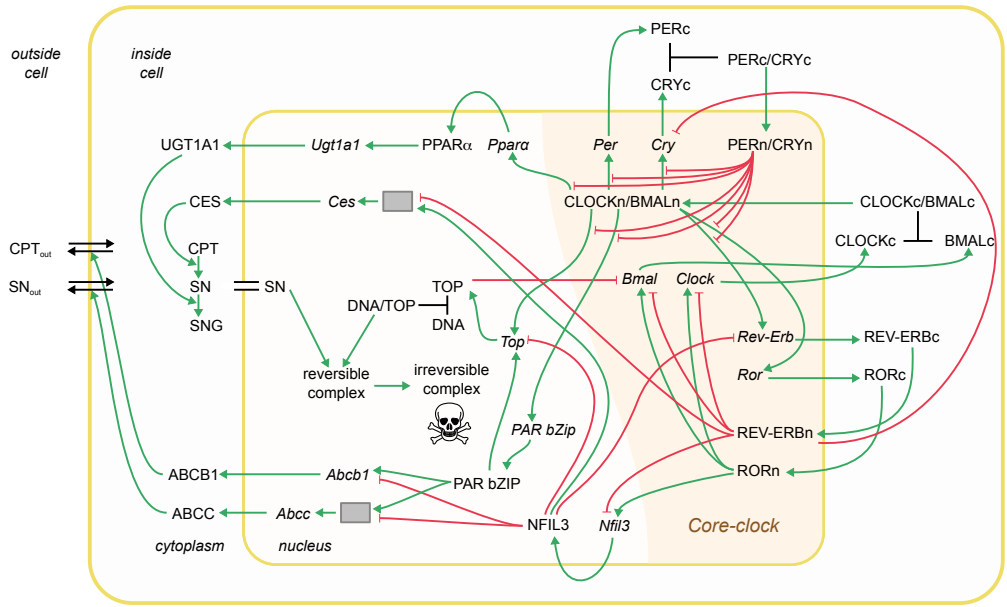


Figure 3.4: Model of the interplay between irinotecan PK-PD and the core clock. Irinotecan treatment is simulated by a transcriptional-translational network that comprises the core clock, irinotecan-relevant genes, and the PK-PD of irinotecan. Different types of interactions are represented among the elements of the network: inhibition (red arrows with flat arrowhead) and activation interactions (green arrows); complex formation (black lines). Grey boxes represent post-transcriptional sub-networks necessary for a model fit to the data. The double black line indicates equal concentrations between nucleus and cytoplasm, the double black line with arrowheads indicates CPT-11 and SN-38 cellular transport inside and outside of the cell. All indicated molecular interactions are based on experimental evidence from a number of different sources and corresponding references are provided in Table 2.S3.

The clock-irinotecan network was fitted to experimental time-series datasets of mRNA concentrations (El-Athman et al., 2019), available for mouse liver (Zhang et al., 2014) and extracted from microarray and RNA-seq data for CRC cell lines, see Section 2 for details. We fitted the clock-irinotecan model to the temporal dynamics of mouse liver, as well as to untreated SW480 and SW620 cells, which resulted in R^2 scores of 0.72, 0.61 and 0.40, respectively (Fig 3.5**a-c**) see Fig. 2.S4 for an example with all genes. For acrophases and relative amplitudes of the model fits see Fig. 2.S5. Periods predicted by the model are 23.5 h for the liver (in accordance with literature values for the circadian period of mice (Ripperger et al., 2011)), 21.6 h for the SW480 cells and 28.8 h for SW620 cells (within the range of previously reported values (Relógio et al., 2014; Fuhr et al., 2018)).

A first version of the model assumed direct (i.e. one step) regulation of irinotecan-related gene transcription by elements of the core clock (Equations S1-29–40, i.e.

Fig. 3.4 without the grey boxes). While this restriction did not hamper the fitting of most genes, the resulting best-fit curves for *CES2* and *ABCC1* mRNA levels were phase shifted compared to corresponding microarray data in SW cell lines. This originated from large phase delays between the clock-controlled regulators and the expression of the regulated genes. For example, *CES2*, which showed clear circadian oscillations in the SW480 cell line (harmonic regression for a 24 h period results in p-value = 0.014, acrophase = 0.09 rad$/2\pi$, relative amplitude = 30%), peaked seemingly before its two regulators, *REV-ERBA* and *NFIL3*, see Fig. 2.S5. Thus, more intermediate elements might play a role in the network. As a simple solution, we extended the model for *CES2* clock-controlled transcription by a simple chain of post-transcriptional modifications (compare Equation (2.S41) to Equation (2.S47) and see Fig. 3.4, grey boxes). To cover the phase delay between *Ces2* and *NFIL3* of 0.60 rad$/2\pi$ (12.9 h, phase delay between *Ces2* and *REV-ERB*α is 0.85 rad$/2\pi$, i.e. 18.2 h), three uni-directional activation steps, with one parameter for both activation and degradation rates, were required. As *ABCC1* showed a similar problem, we added an analogue set of intermediate reactions for *ABCC1* transcription, for which two steps were sufficient, as the phase delay with its regulators was smaller (phase delay between *ABCC1* and *NFIL3* of 0.39 rad$/2\pi$, i.e. 8.4 h, phase delay between *ABCC1* and *REV-ERB*α of 0.64 rad$/2\pi$, i.e. 13.8 h), see Fig. 2.S5. The fit of the additional genes does not reduce the quality of the fit of the core-clock model, with R^2 scores for the core clock of the full fit of 0.93, 0.78 and 0.57 compared to 0.84, 0.67 and 0.52 for a fit of only the core clock using the rescaled microarray data (see Section 2), for liver tissue, SW480 and SW620 cell lines, respectively. Lower R^2 scores for the full fit likely result from the longer optimization required for a good fit of all genes, as compared to the optimization required for fitting only the core clock genes. From liver to SW480 to SW620, the relative amplitude of the oscillation was reduced for genes of the core clock and for genes directly regulated by the core clock, whereas this amplitude reduction was relaxed for genes only indirectly controlled by the core clock (Fig. 3.5**d** and Fig.2.S5**b**).

From the fit of the gene expression data, we obtained a calibrated model computing clock and irinotecan-related mRNA circadian rhythms. However, mRNAs need to be translated into proteins that eventually interact with the drug. Hence, the link between the clock-irinotecan network and the PK-PD model for treatment toxicity was assumed via the protein dynamics and allows us to investigate the interplay between the circadian clock and irinotecan action. We designed a new model of protein dynamics of UGT1A1, CES2, ABCB1 and ABCC1, which are the inputs for the irinotecan chronoPK-PD model (Equation (2.S31) with Equation (2.S57)), replacing the forced cosine function utilized in the original model by Dulong et al. (2015). The protein dynamics contained a term for protein translation including mRNA levels computed by the clock-irinotecan model, together with a circadian process of degradation as suggested for many proteins (Lück et al., 2014). Magnitude, am-

plitude and phase of the circadian degradation are fitted to cytotoxicity data; the translation rate is set to 1 as protein abundances are re-scaled in the PK-PD model, see Section 2 (Lück et al., 2014).

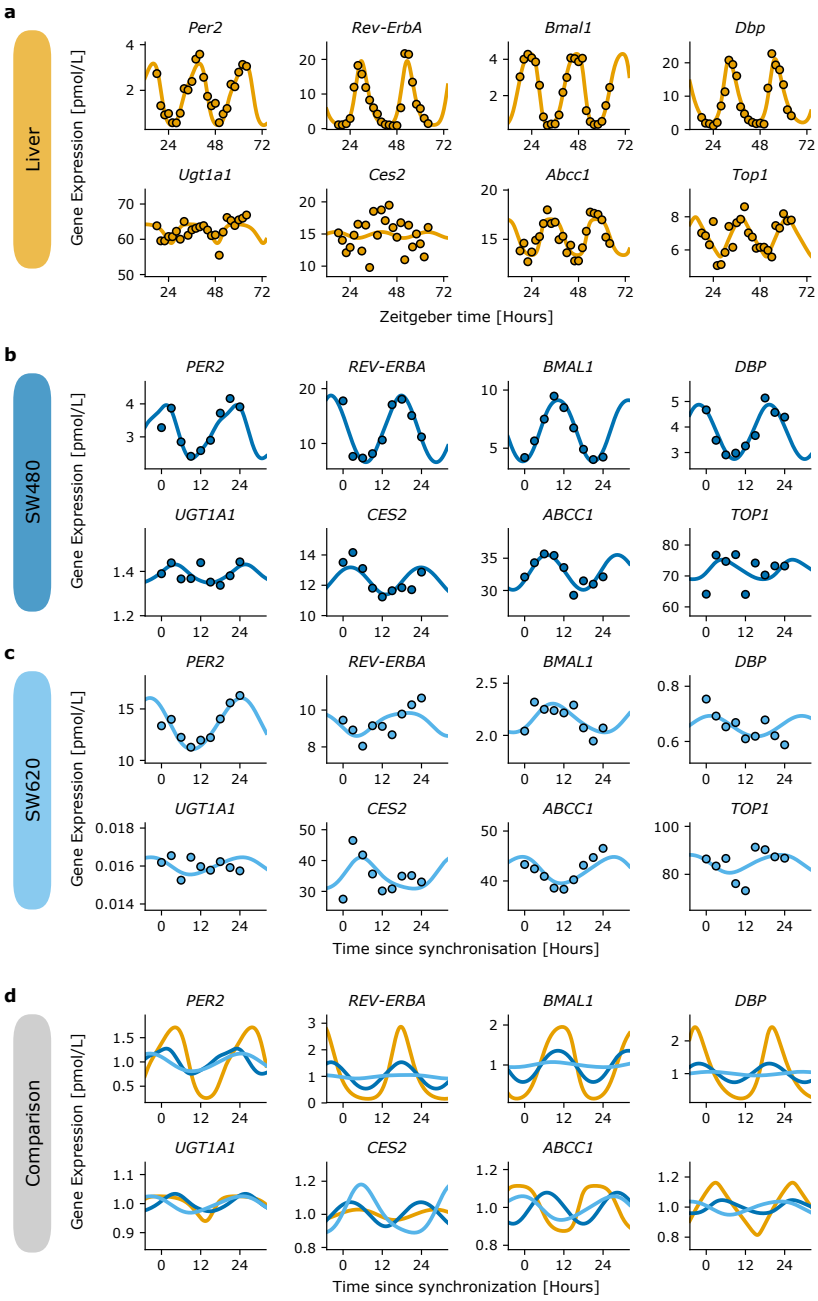


Figure 3.5: Comparison of the fitted clock-irinotecan network for healthy mouse liver and human cancer-derived cells. Selected gene expression and fit of the clock-irinotecan network for **(a)** mouse liver data, **(b)** SW480 and **(c)** SW620, all data without treatment. **d** Comparison of the model output when fitted to liver (orange), SW480(dark blue) and SW620 (sky blue). Profiles were normalized to the mesor, and *BMAL1* phases were aligned between mouse liver and the SW480 cell line to allow for comparison.

3.3.4 The full clock-irinotecan model recapitulates different chronotoxicity rhythms for CRC cells

To investigate the putative effects of time-dependent treatment in CRC, SW480 and SW620 cells were synchronized by media change and treated with 2μM of irinotecan at four different circadian times (CT after synchronization: 6 h, 12 h, 18 h and 24 h). The SW480 cell line exhibited a circadian cytotoxicity response to treatment (harmonic regression with the period of the model fit, p= 0.043 for SW480, see Fig. 2.S7; not significant for SW620). SW480 cells showed the highest toxicity when irinotecan was administered 24 h after synchronization, while the lowest toxicity was observed when irinotecan was administered 12 h post synchronization (acrophase of 0.006 $\pm$ 0.03 rad/2π,Fig. 3.6). The differences in cytotoxicity values between different treatment time points where higher in SW480 as compared to SW620 cells, which resulted in larger circadian amplitudes for the SW480 toxicity rhythm (Fig. 3.6**b**). Here it is relevant to notice that, in the absence of treatment, the number of dead cells in SW480 cultures is higher than for SW620 cell cultures (ratio of the area under the curve of SW480 and SW620 is 2.46 $\pm$ 0.07,Fig. 3.6**a** and **d**) pointing to cell death and cell cycle differences between the tumor and the metastasis-derived cells.

To allow for the comparison of the model with this experimental data, we supplemented the model by Dulong et al. (2015) with two cell population equations that explicitly track the number of alive and dead cells. An exponential growth and a first-order natural cell death were assumed in both control and treated conditions. Irinotecan was assumed to act negatively on cell proliferation and survival through DNA damage formation, and a circadian oscillation in the cell death rate was added, see Equation (2.S58) and Equation (2.S59). Parameters of the original model were kept unchanged apart from the formation rate of the irreversible complex which had to be adapted for a successful fit, and which ended up being reduced as compared to its former estimation.

Using the mRNA dynamics computed by the clock-irinotecan model, the irinotecan PK-PD model allows to fit the circadian dynamics of cell death (Fig. 3.6). The best-fit full model generated a circadian cell death profile that agreed with the toxicity phase of the experimental data for the SW480 cell line. The model also recapitulated a different toxicity profile for the SW620 cell line, supporting the hypothesis that the same drug at the same concentration could lead to different responses based on the time of treatment administration and on the cancer clocks. Interestingly, while the highest and lowest cytotoxicity trends were the same in both cell lines, the overall response to the cytotoxic effect of the drug was higher in SW480 (derived from the primary tumour) in comparison to SW620 (derived from a metastasis, but from the same patient). This also alludes to a role of the cellular clock profile in treatment outcome, as the two cell lines have different oscillatory

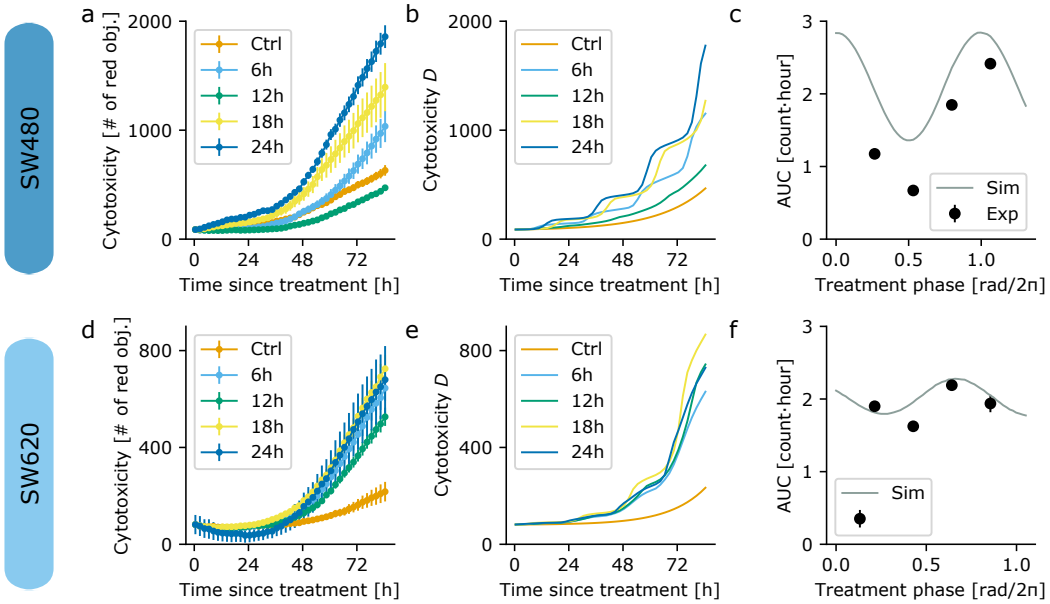


Figure 3.6: Fitting of the time-dependent treatment from human cancer cell lines. **a** Experimentally measured cytotoxicity curves, estimated by measuring red fluorescent objects (see Methods), for SW480 cells that are untreated (Ctrl), or treated at indicated time points with irinotecan (6 h, 12 h, 18 h or 24 h after synchronization). Time is aligned to treatment onset. **b** Best-fit of the extended PK-PD model (shown is the number of dead cells, the dynamical variable D of Equation (2.S59)) to the experimental cytotoxicity data of the SW480 cell line. **c** Area Under the Curve (AUC) of treated SW480 cells normalized by the untreated control (dots), compared with the area under the curve of the best-fit model (grey line). **d** Experimentally measured cytotoxicity curves for SW620 cells untreated (Ctrl), or treated at indicated time points with irinotecan (6 h, 12 h, 18 h or 24 h after synchronization). **e** Best-fit of the extended PK-PD model to the experimental cytotoxicity data of the SW620 cell line. **f** Area under the curve of treated SW620 cells normalized by the untreated control (dots), compared with the area under the curve of the best-fit model (grey line).

patterns. We further tested a simplified version of the equation for the protein dynamics assuming constant, i.e. non-circadian, protein degradation (Equation (2.S31) with constant degradation rate). As anticipated by the mathematical analysis, toxicity oscillation amplitudes were drastically reduced to approximately 1% of the mesor and were then much smaller than those observed experimentally. Yet, the proposed simplified clockPK-PD model with minor adaptations to the experimental settings gave reasonable toxicity phases using non-circadian protein dynamics with an appropriately chosen degradation rate (Fig. 2.S6), without fitting the model to the circadian toxicity values obtained experimentally, see Fig. 2.S7.

To test for the sensitivity of this final model to parameter variations, we evaluated parameter sensibility of a set of 123 parameters with respect to the phase

and amplitude of irinotecan circadian toxicity profiles (i.e. the curves depicted in Fig. 3.6**c** and **f**) by calculating Sobol sensitivity total order indices, see Fig. 2.S9. A close agreement was found between the parameter sensitivity on the phase and on the amplitude of the drug chronotoxicity rhythms. This analysis highlighted the impact of the protein dynamics on the toxicity profile, most importantly the relevance of the phases of the circadian degradation of CES2 (parameter phiCes from Fig. 2.S9) and UGT1A1 (parameter phiUgt) and the amplitude of CES2 (parameter ampCes). Besides those parameters, several core-clock elements - in particular parameters associated with the loop formed by *ROR, BMAL1* and *CLOCK* (maximal transcription rates, degradation rates, production rates) - showed high sensitivity, probably because existence of most oscillations depends on the core clock. The feedback from irinotecan-relevant genes to the core clock, through the inhibition of *REV-ERB* by NFIL3 (parameter `i_RevNfil`) and the inhibition of *BMAL1* by TOP1 (parameter `i_BmalTop`), only showed a weak impact on the toxicity curve. All parameters associated with ABC transporters showed a low impact on the toxicity profile.

3.4 Discussion

The circadian clock regulates the timing of various crucial molecular pathways including drug metabolism, apoptosis, DNA damage repair and cell cycle (Sancar and Brunner, 2014; Sancar et al., 2010; Gaucher et al., 2018). The malfunctioning of these pathways is involved in cancer onset and progression. On the other hand, several drugs used in cancer treatment target genes, which are expressed in a circadian manner and also the metabolism of these drugs is carried out by circadian-regulated genes and proteins. Hence, timing treatment in accordance with the patient's circadian timing system is likely to contribute to improved treatment outcome, and several studies have shown promising results using chronotherapy in cancer treatment (Ballesta et al., 2017). We developed a novel mathematical model of irinotecan cellular PK-PD, which links the core clock with predicted treatment toxicity for CRC cells. The model simulations highlighted the existence of time-dependent toxicity for the different cells, which was different for the tumour-derived cell line as compared to the metastasis-derived cell line. Results suggest that, in addition to gene expression, the dynamics of proteins, with circadian variation in their degradation, plays an important role in the timing of drug toxicity. In particular the phase (the time of maximum expression) and amplitude (difference between minimum and maximum) of the circadian oscillation in protein degradation of CES2, which controls the activation of irinotecan, seems to be relevant in shaping the toxicity profile. Moreover, elements associated with core-clock genes, such as *BMAL1* or *CLOCK* showed high sensitivity which proved the importance of the core-clock parameters on irinotecan toxicity.

3.4.1 A comprehensive mathematical model for circadian regulation of irinotecan PK-PD

Our clock-irinotecan model can be fitted to different scenarios providing different circadian toxicity profiles for CRC cells. The core-clock model was initially developed from multiple datasets of mammalian SCN cells (Relógio et al., 2011), and was successfully refined here using quantitative measurements of the clock of the mouse liver, and of SW480 and SW620 cell lines. Regarding the model of irinotecan PK-PD, it was designed based on extensive datasets in Caco-2 cells (Dulong et al., 2015), and was further validated in both SW480 and SW620 cell lines. Using SW480 and SW620 cell lines here provided a proof of principle that a personalization of the model to other cell lines was possible. Hence, models of both transcription-translation clock network and irinotecan PK-PD were validated in several *in vitro* and *in vivo* experimental settings, which argues in favour of their reliability.

The reduction in fit quality from liver to SW480 to SW620 cells likely results from the decreasing amplitudes of circadian oscillations, see Fig. 2.S5**b** (Relógio et al., 2014; Fuhr et al., 2018). Indeed, assuming that the experimental data of the same gene, yet from different biological sources (liver or CRC cell lines), shows the same level of noise, arising from biological stochasticity or from experimental constraints, larger oscillation amplitudes lead to higher signal to noise ratios, which facilitates fitting. In addition, the fact that the core-clock model did not fully fit the SW620 data was an indication that clock gene and protein interactions may be impaired in this cell line or at least different from the ones implemented in the model. It is also important to notice that each patient (or healthy individuals) and each cancer are unique. Accordingly, different cell lines,patients or healthy individuals (Basti et al., 2021) have specific clock phenotypes, and this requires personalization of treatment. Thus, any clinical application requires that our model is fitted to the individual patient, or to groups of patients with similar clock molecular profiles.

Another ODE-based model of the mouse liver clock was designed by Woller et al. (2016) to investigate the effect of feeding cycles on liver circadian rhythms. That model was developed to address a different question as compared to this study and could not be readily used here, as for instance the energy metabolism part was out of the scope here. In addition, that model did not include different compartments for the nucleus and the cytoplasm, which we were able to do thanks to the recent publication of data on clock-gene subcellular trafficking (Wang et al., 2017). Further, we included the gene *Clock* to the model, to incorporate available data on this gene and investigate its importance in the clock machinery. Finally, the model integrates both mRNA and protein circadian datasets in a quantitative manner, meaning that it does not only predict the phase and relative amplitude of the gene expression data as existing models do (Relógio et al., 2011; Woller et al., 2016), but also computes the protein expression levels, such information being critical

for the connection to PK-PD models. A very interesting perspective for future studies would be to consider coupling the model by Woller et al. (2016) with ours to investigate the impact of feeding/fasting cycles on irinotecan chronopharmacology, as liver enzymes involved in the drug PK showed variations according to food intake (Okyar et al., 2019)

Our core-clock model represents intracellular regulatory feed-back loops that implicitly include extrinsic circadian regulators such as temperature or light/dark cycles. Such external synchronizers were not present in our cell culture setting, so that the SW480 and SW620 models are likely to represent the actual events at stake. On the opposite, external or systemic regulators have a great influence on the mouse liver clock. This precise question was the topic of another of our recent studies and the main concern of Chapter 4, in which we explicitly model the influence of temperature cycles and food intake on the core clock in four classes of mice (2 strains, 2 sexes) (Martinelli et al., 2021).

Regarding CES2 modelling, we chose not to connect its protein degradation rate directly to the core clock, since there is no published data regarding the existence or absence of such molecular links. Thus, instead of including unreliable reactions to the model, we preferred to estimate the circadian rhythm of CES2 protein degradation directly from the data. The parameter sensitivity analysis evidenced the importance of this part of the model and strongly advocates the generation of additional biological results about the circadian control of CES2 protein degradation.

Our experimental results using a CRC *in vitro* cellular system highlight the existence of cytotoxicity differences resulting from chronomodulated treatment, which were further emphasized by our simulations. Compared to the predicted toxicity acrophase of the SW480 cell line, the toxicity acrophase of the SW620 cell line is delayed by about four hours. In particular, the metastasis-derived cell line showed the lowest variations in cytotoxicity among different treatment times. Our data also shows a difference in terms of drug resistance between the two CRC cell lines, which, we hypothesize, could be overcome if using treatment times of high cytotoxicity with the same amount of drug, or potentially by increasing dosages in times of least toxicity. Yet, these are still speculative ideas, which need more studies and validation in a clinical setting. The small number of cell lines included in our experimental setup present some limitations to the generalization of our findings and thus further investigation of time-dependent treatment with a higher variety of cell lines and anticancer treatment agents needs to be carried out in future research.

3.4.2 Personalized models to optimize timing in cancer treatment

Chronotherapeutic studies aim at increasing treatment efficacy and minimizing toxicity for healthy cells leading to a reduction of the side effects for patients (Lévi et al., 2010). Previous clinical results have shown that personalization is a key

element of successful chronotherapy outcome, for example males and females have shown different toxicities depending on treatment timing (Anderson and Fitzgerald, 2020; Innominato et al., 2020a; Li et al., 2013). Sex should be considered as a relevant determinant of circadian rhythms and optimal drug timing in the light of recent preclinical and clinical findings (Anderson and Fitzgerald, 2020; Innominato et al., 2020a; Li et al., 2013). Here, the mouse liver data was obtained from male mice, and the cell lines were derived from a human male, so that the sex specificity seems out of the scope of this study. However, we have started to investigate the impact of sex on the circadian timing system as mentioned above and did find significant sex-related differences in the shape and intensity of systemic controls on the core clock in a mouse study (Martinelli et al., 2021). These differences are analyzed in the next Chapter, Section 4.4.3. Several sex-specific datasets related to irinotecan chronotoxicity are available in mice and in patients, and further modelling work would allow to investigate molecular determinants of male/female differences in irinotecan response with respect to timing (Ballesta et al., 2017).

One aspect of personalized chronotherapy is an adaptation of medication timing to the patient's internal time, which can be best assessed by a combination of mathematical modelling and machine learning. Existing models of irinotecan PK-PD offer to predict the optimal drug timing based on circadian rhythms of proteins involved in irinotecan pharmacology, in the organ of interest (e.g. liver, or intestine) or in the tumour. However, such datasets are unlikely to be obtained in the clinics on an individual patient basis as it would involve multiple around-the-clock biopsies. This obviously raises questions of feasibility and ethics regarding benefit/risk ratios. Furthermore, circadian datasets on irinotecan-related proteins would not be informative for personalizing the timing of other drugs, in particular the ones usually combined with irinotecan (e.g. 5-fluorouracil, oxaliplatin). Instead, our new combined model provides the option of computing irinotecan best timing from circadian rhythms of core-clock mRNA levels. The major advantage of measuring core-clock genes - and not directly drug-related genes - is that it can be done in any organ, since the peripheral core clock is synchronized across healthy tissues as suggested by mouse and baboon studies (Zhang et al., 2014; El-Athman et al., 2019; Mure et al., 2018). Several patient-friendly methods for measuring clock-gene expression in saliva or blood samples have been recently validated in the clinics (see Gaspar et al. (2019) for a review). Furthermore, strong oscillations, clearly above noise level, are expected in core-clock gene expression, which facilitates the characterisation of their circadian profiles and reduces the number of needed time points to do so (Basti et al., 2021). In addition, a newly available statistical algorithm offers to derive clock gene mRNA circadian rhythms from a single-time-point measurement of 10 clock genes (Vlachou et al., 2020). Such methodology could potentially be used to predict clock gene variations when only one time point is available, which is often the case for the tumour. Our combined model could

then be used to infer irinotecan personalized best timing from clock-gene expression. As such chronoPK-PD models could be developed for any other drug, optimal timing could be derived for multiple compounds from a single dataset of clock gene mRNA circadian variations. In addition, instead of simplifying a patient's complex circadian profile by an estimate of a value associated with their circadian time, our model has the potential to fit the circadian rhythms of the patient based on their personal gene expression data from peripheral tissues (e.g. saliva (Basti et al., 2021)). Thus, in a clinical application, the model can be fitted both to the tumour clock and to the healthy peripheral clock of the patient. Several therapeutic strategies may then be considered from maximizing efficacy, or minimizing side effects, of a given drug dose, to more advanced approaches aiming to optimize antitumor efficacy under strict tolerability constraints (Ballesta et al., 2011). To exemplify the power of our model for personalization, it was fitted to two different cell lines derived from human CRC with cell line-specific toxicity profiles, which are different in the metastasis-derived cells as compared to the primary tumour cells likely due to a disruption of the circadian profile and an alteration of metabolism in the former cells (Fuhr et al., 2018). Overall, the proposed approach provides a promising direction for mechanism-based chronotherapy personalization in the clinical setting. The sensitivity analysis of the circadian toxicity profile is in accordance with the previously published sensitivity analysis by Dulong et al. (2015), highlighting especially the importance of CES2, which is responsible for activation of irinotecan. Using our fitted mathematical model, changes in toxicity in response to relevant alterations in core-clock or protein dynamics can in principle be predicted based on circadian data for core-clock and drug-pharmacology genes. Further, the model can also be adapted for patients with alterations in irinotecan PK-PD proteins, such as patients with increased sensitivity against irinotecan due to a reduced UGT1A1 activity (reduced deactivation of SN-38) (Fujii et al., 2019), or patients with a decreased sensitivity to irinotecan due to an over-expression of ABC transporters, which leads to a faster drug removal from the cell (Fletcher et al., 2016).

The influence of the core clock on the toxicity profile supports a dependence of optimal treatment times on the personal circadian rhythm of patients, in accordance with previous reports (Ballesta et al., 2017). In particular, several core-clock parameters associated with *BMAL1* and *CLOCK* (*BMAL1* degradation rate, *CLOCK* activation rate, cytosolic *BMAL1* degradation rate) show high sensitivity in the model, highlighting the relevance of the core clock for irinotecan PK-PD. This is particularly relevant for cancer patients, who often show alteration in their circadian rhythms that might be further changed during hospitalization, as bedridden patients seem to have disrupted circadian rhythms (Rida et al., 2019; Neikrug et al., 2012; Ikegami et al., 2019; Roenneberg and Merrow, 2016; Innominato et al., 2014). This suggests that even during a treatment frame of a few weeks, optimal treatment times might be shifted by a flattening of the circadian rhythms. Light therapy

might help to stabilize toxicity profiles, as it has been shown to improve circadian oscillations in breast cancer patients (Neikrug et al., 2012). Also melatonin administration or pharmacological modulation of core-clock genes may have a positive impact on cancer therapy (Gil-Martín et al., 2019; Sulli et al., 2019). We here report differently timed toxicity peaks for CRC cell lines. Naively, one would assume that cancer cells show less robust oscillations compared to healthy cells, but this remains to be shown in future research. While the toxicity of Caco-2 cells in the model of Dulong et al. (2015) was predicted following a repeated 2-hour treatment, we here predict toxicity phase following a 84.5-h long treatment. The situation in the patient is most likely somewhere between these values as irinotecan terminal half-life after a 30-min infusion to colorectal cancer patients in the morning was approximately equal to 12 h (Lévi et al., 2010). As the treatment administration scheme may present complex and chronomodulated shapes, irinotecan whole-body pharmacokinetics must be precisely modelled in order to faithfully predict plasma and tissue exposure concentrations. A mathematical model relating infusion pump and administration schedules to predict actual drug concentrations in the body has been developed (Hill et al., 2020), and a corresponding extension could be used to further improve predictions of the here presented cellular model in a whole-body context. Thus, for personalized medical treatments the personalization of mathematical models is key, using easily accessible patient data to predict unassessable information relevant for medication.

3.5 Conclusion

Our clock-irinotecan model can be further optimized in a personalized manner and may be used to predict the toxicity profile of a particular patient upon fitting his or her molecular circadian profile. The model can be additionally used to investigate whether the differential regulation of PK-PD elements, for example via additional medication with melatonin, can result in circadian toxicity profiles that would support chronotherapy in irinotecan-treated cancers (Gil-Martín et al., 2019). Altogether, these findings highlight the relevance of investigating the effect of chronomodulated therapy in a clinical setting as it may contribute to providing better personalized medical treatment with higher efficacy and lower cytotoxicity, leading to a decrease of side effects and an increase of life quality for the patient.

This chapter has focused on optimal drug timing in the context of irinotecan administration at the CRC cell level. Using an explicit modeling of the circadian clock control on irinotecan PK-PD, this approach provided the option of computing irinotecan best timing from circadian rhythms of core-clock and pharmacological gene mRNA levels. This being said, two key aspects of the circadian timing system (CTS) were out of the scope of this cellular study. First, the influence of external or

systemic regulators on the peripheral clocks was not accounted for in this *in vitro* setting where the main synchronizers such as temperature or nutrient exposure were constant over time. This research topic is of great importance since the precise molecular interactions between clock genes and synchronizers remain unclear. Second, although the approach addresses treatment personalization through fitting of the cellular circadian clock and irinotecan PK-PD of the individual patient, whole-body sex-specific and genetic background CTS differences were not investigated from a mechanistic point of view. The combination of these two aspects calls for an inference method to identify the influences of extracellular synchronizers on peripheral clocks, in a personalized fashion. These prospects lie at the core of the next Chapter.

Chapter 4

Model learning to identify systemic regulators of the peripheral circadian clock

Scientific production

The content of this chapter is based on the article: J. Martinelli, S. Dulong, X. Li, M. Teboul, S. Soliman, F. Lévi, F. Fages, A. Ballesta, *Model learning to identify systemic regulators of the peripheral circadian clock"* Bioinformatics, i401-9, 2021.
The code for the figures as well as an introductive jupyter notebook is available on GitLab https://gitlab.inria.fr/julmarti/model-learning-mb21eccb

Abstract Personalized medicine aims at providing patient-tailored therapeutics based on multi-type data towards improved treatment outcomes. Chronotherapy that consists in adapting drug administration to the patient's circadian rhythms may be improved by such approach. Recent clinical studies demonstrated large variability in patients' circadian coordination and optimal drug timing. Consequently, new eHealth platforms allow the monitoring of circadian biomarkers in individual patients through wearable technologies (rest-activity, body temperature), blood or salivary samples (melatonin, cortisol), and daily questionnaires (food intake, symptoms). A current clinical challenge involves designing a methodology predicting, from circadian biomarkers, the patient peripheral circadian clocks and associated optimal drug timing.

The mammalian circadian timing system being largely conserved between mouse and humans yet with phase opposition, the study was developed using available mouse datasets. We investigated at the molecular scale the influence of systemic regulators (e.g. temperature, hormones) on peripheral clocks, through a model learning approach involving systems biology models based on ordinary differential equations. Using as prior knowledge our existing circadian clock model, we derived an approximation for the action of systemic regulators on the expression of three core-clock genes: *Bmal1*, *Per2* and *Rev-Erbα*. These time profiles were then fitted with a population of models, based on linear regression. Best models involved a modulation of either *Bmal1* or *Per2* transcription most likely by temperature or nutrient exposure cycles. This agreed with biological knowledge on temperature-dependent control of *Per2* transcription. The strengths of systemic regulations were found to be significantly different according to mouse sex and genetic background.

Contents

4.1 Introduction

Recent clinical findings concluded to a large impact of patients' sex, genetic background, and lifestyle on drug optimal timing, thus highlighting the need for indi-

vidualized chrono-infusion schemes to further improve treatment outcome. This demand has initiated the development of eHealth platforms dedicated to the follow up of key circadian biomarkers in individuals (Kim et al., 2020). For instance, the PiCaDo platform, that integrates data from wearable sensors recording rest-activity, position and skin-surface temperature, was validated for safe home-based assessment of patient's rhythms (Innominato et al., 2018; Komarzynski et al., 2018). Such dynamical patient characteristics may be combined to measurements of key markers in blood or salivary samples, like melatonin and cortisol, and to food diary keeping track of nutrient intake over the day. However, nowadays, there does not exist a methodology integrating these patients' circadian datasets for the prediction of the temporality of the circadian timing system (CTS) and of personalized drug timing. This study aims at proposing the first steps of such an approach.

From a mechanistic point of view, the information needed for the personalization of chronotherapy consists in the circadian variations of proteins involved in drug PK-PD, which can then be integrated in ODE-based models as demonstrated in Chapter 3 (Section. 3.3.4). However, such internal measurements are unlikely to be assessed around the clock in individual patients due to the invasive nature and high frequency of the clinical acts that would be required. As a consequence, clinical datasets comprising both circadian biomarkers and circadian rhythms of clock or pharmacological genes in peripheral organs in the same individuals are very sparse. This means that purely statistical approaches cannot be applied here. To take on the long-term challenge of predicting the personalized drug timing from non-invasive patients' circadian datasets, we present a three-step approach, summarized in Fig. 4.1. In brief, it aims to model the influence of whole-body systemic regulators on the peripheral clocks that, in turn, control the circadian rhythms of key pharmacological enzymes from which drug PK-PD and optimal timing can be inferred.

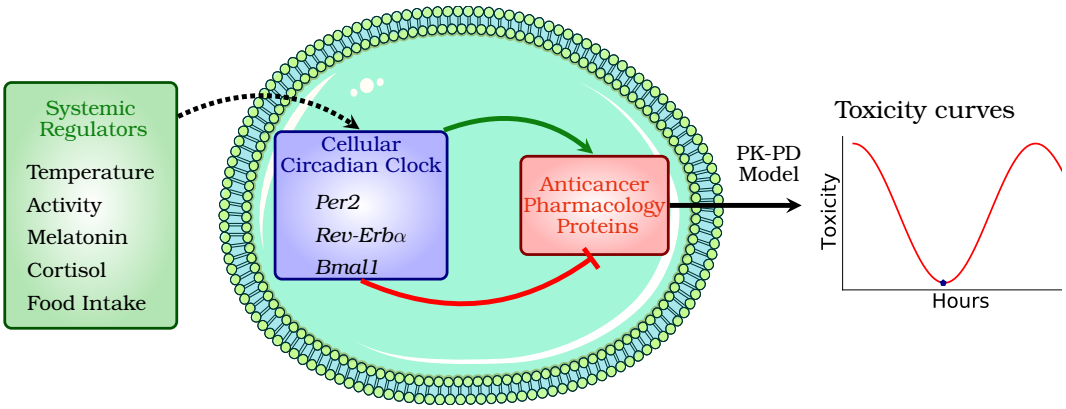


Figure 4.1: Long-term aim of cancer chronotherapies personalization from non-invasive monitoring of systemic regulators.

Chapter 3 provided a representation based on mechanistic models of the connections between the cellular circadian clock and key pharmacological enzymes (middle part of Fig. 4.1). It also presented the computation of administration time-dependent toxicity curves through their linking with drug PK-PD modeling (right part of Fig. 4.1). The case of irinotecan, a widely-used anticancer compound was investigated here. This chapter focuses on the precise circadian influence of whole-body regulators on the expression of core clock genes in peripheral organs (left part of Fig. 4.1). To that end, we designed a model learning methodology to identify systemic regulators of the peripheral circadian clock from temporal data. The approach built on the circadian clock model presented in Chapter 3 as the embodiment of the prior knowledge available on the molecular interactions underlying the clock. The objective was to infer precise mechanisms of whole-body control from available circadian time series of systemic regulators and clock gene expression.

Some existing machine learning techniques for network inference combine the use of time series data with a facilitated integration of prior knowledge. In the context of ODE-based models, this is accounted for as known regulatory mechanisms or kinetic rates in the ODEs (Huynh-Thu and Geurts, 2018; Aalto et al., 2020). This being said, none of these approaches provide quantitative insights about the inferred interactions. Indeed, the underlying kinetics between the target and the regulators are obtained in a non-mechanistic manner using e.g. boosted decision trees or gaussian processes. Consequently, dealing with datasets involving multiple related groups or individuals as can be the case in clinical trials including patients of different sex or genetic background, would not be possible. Hence, there is a need for the design of a network inference method able to handle prior knowledge in a context where quantitative patient-specific information needs to be accounted for in the inferred dynamical model.

We here relied on systems biology and systems pharmacology approaches that offer to dynamically model, through ODEs, key intracellular pathways. Model variables and parameters carry physical meaning that is conserved across species, so that sub-model structures and parameter values can be validated in pre-clinical settings and further integrated in patient models, as in a multi-scale pipeline. This is particularly handy in the case of the mammalian CTS which is largely conserved between mouse and humans yet with phase opposition. Thus, we developed our model learning approach using extensive circadian datasets available in four classes of mice (2 strains, 2 sexes) as a first step towards clinical application. After describing the available mouse datasets, we will expose our approach of model learning and then present the results obtained in terms of biological predictions.

4.2 Available data: circadian biomarkers and liver clock gene expression in four mouse classes

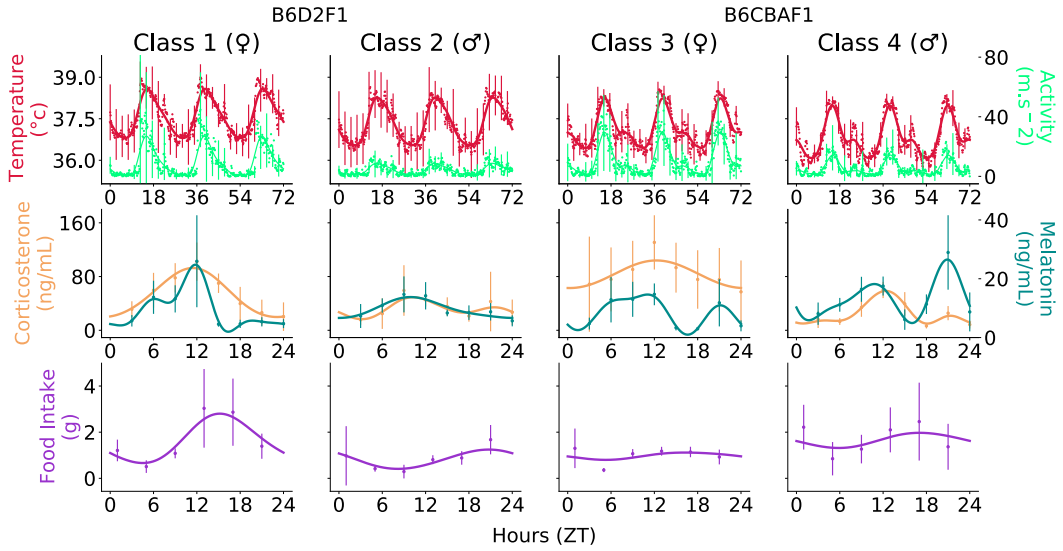


Figure 4.2: Circadian biomarkers in four mouse classes. Raw data are represented with dots (average) and error bars (standard deviations). For the sake of readability, error bars were only displayed every 2.5 hours for the first line. Solid lines stand for the mean function obtained by fitting a Gaussian Process.

This study aiming to identify the control of systemic regulators on the cellular circadian clock was based on extensive circadian datasets available in both male and female mice of B6D2F1 and B6CBAF1 strains (Ahowesso et al., 2011; Li et al., 2013). Class 1 and 2 were defined as female and male B6D2F1 mice, Class 3 and 4 as female and male B6CBAF1 mice, respectively. For each mouse class, five systemic biomarkers were measured around the clock, which were body temperature, rest-activity, food intake, plasma corticosterone and melatonin (Fig. 4.2). The first two biomarkers were captured by an implanted sensor providing data every 10 minutes for 72 hours, with up to 8 biological replicates per point (Ahowesso et al., 2011). For the plasma corticosterone and melatonin, the time resolution was 3 hours with 3 biological replicates (Ahowesso et al., 2011; Li et al., 2000). Finally, the amount of food in a cage housing 3 mice was weighted every 4 hours using a precision scale. The value measured for food intake consists of the amount of food at time T_1 minus the amount of food at the next circadian time T_2 (Ali and Kravitz, 2018). 3 biological replicates were used per time point. Circadian rhythms were validated using Cosinor for all classes for temperature, rest-activity and melatonin ($P < 0.05$). Concerning corticosterone, all classes but class 2 displayed circadian rhythms ($P = 0.08$). Food intake was predicted to display circadian variations for Class 1 and 2, only ($P = 0.16$ and $P = 0.31$ for class 3 and 4, respectively).

Significant sex differences could be observed for instance in rest-activity profiles in terms of mesor as well as relative circadian amplitudes, although the phases were similar. Conversely, temperature profiles were virtually identical across classes. Overall, phases are well-preserved from one class to another for all biomarkers. Furthermore, mRNA circadian concentrations of the core-clock genes *Bmal1*, *Per2* and *Rev-Erbα* were measured in the mouse liver for the four classes (Figure 3 from Li et al. (2013)). Circadian rhythms were validated for all genes and classes (Cosinor $P < 0.05$). Gene expressions were quite alike classwise in terms of phases. All datasets were preprocessed using Gaussian processes with a 24h-periodic kernel (Fig. 4.2, Rasmussen and Williams (2006)).

4.3 Model Learning Approach

4.3.1 Accounting for direct and indirect action of systemic regulators on the clock

The five measured circadian biomarkers (rest-activity, temperature, food intake, corticosterone and melatonin) are considered as possible systemic regulators of the clock. We here focus on liver cells which do not express receptors to melatonin so that we do not anticipate any direct control of this feature on the clock. It is thus integrated in the study as a negative control. Regulators may have either immediate or time-shifted interactions with clock genes. Indeed, intermediate species are likely to be involved in the influence of these regulators on the clock. This would induce time delays as compared to the biomarkers data. For instance, temperature increase may lead to an enhanced expression of Heat-shock proteins (HSP) which then interact with clock genes (Kornmann et al., 2007). Such cascade of events would induce a phase shift between the action of the direct regulator (e.g., HSP) and the data of the corresponding biomarker (e.g., temperature). Let us assume that the regulator z_1 produces the species Z_1 through a linear kinetics with rate constant k_1, an explicit formula is obtained for Z_1 as:

$$\dot{Z}_1(t) = k_1 z_1(t) \Rightarrow Z_1(t) = k_1 \int_0^t z_1(s)ds \tag{4.1}$$

Hence, direct action of the five regulators are represented by the corresponding circadian biomarker data $\bar{z}_j$ and indirect actions are included through *integral regulators* $\bar{Z}_j$:

$$\bar{\mathbf{z}} = (\bar{z}_j, \bar{Z}_j)_{1 \leqslant j \leqslant 5} \text{ where } Z_j(t) = \int_0^t z_j(s)ds \tag{4.2}$$

k_1 being incorporated into parameters of the statistical models, see below.

4.3.2 Setting a regression problem, using an ODE-based model of the liver circadian clock

In order to identify the action of systemic regulators on the cellular circadian clock, we settled for a model-based approach, which enables us to derive a mathematical expression for the approximation of this action. This approximation relies on several hypothesis described in this section. We use the ODE-based model of the mouse liver circadian clock presented in the previous Chapter, Section 3.3.1, which recapitulates the molecular interactions between clock genes and their transcription, nuclear transport and degradation (Fig. 3.1**b**). Briefly, CLOCK/BMAL dimer is assumed to enhance the transcription of clock genes *Rev-Erbα*, *Rorγ*, *Per2*, and *Cry1* and PER/CRY complex to inhibit this transcriptional activation. The model includes two main negative feedback loops. The first one involves the self-inhibition of *Bmal1* through the activation of its repressor REV-ERB by the dimmer CLOCK/BMAL. On the opposite, ROR whose expression is also increased by CLOCK/BMAL presence, acts positively on *Bmal1* modulation. The second feedback loop is induced by the self-repression of *Per2* and *Cry1* gene expression through the inhibition of CLOCK/BMAL transcriptional activity by the PER/CRY protein complex. In addition, REV-ERB inhibits *Cry1* gene expression, thus inhibiting its own inhibition through the modulation of PER/CRY level. In this mathematical model, the expression of gene x is typically described by the following differential equation:

$$\frac{dx}{dt} = V_{\max}\text{Transc}(\mathbf{M}, \boldsymbol{\gamma}) - \alpha x \tag{4.3}$$

The right term of Equation (4.3) accounts for gene mRNA degradation occurring at constant rate α. $V_{\max}$ stands for the gene transcription level in the absence of modulators. The function Transc embodies the action of modulatory species $\mathbf{M}$ on x transcription through Hill-like kinetics terms parametrized by $\boldsymbol{\gamma}$. For instance, the positive action of the ROR protein on *Bmal1* transcription and the counter inhibitory part from REV-ERB action (Guillaumond et al., 2005) are modelled as:

$$\text{Transc}_{Bmal1} = \frac{1 + \gamma_1 \left(\frac{\text{ROR}}{\gamma_2}\right)^{\gamma_3}}{1 + \left(\frac{\text{REV-ERB}}{\gamma_4}\right)^{\gamma_5} + \left(\frac{\text{ROR}}{\gamma_2}\right)^{\gamma_3}} \tag{4.4}$$

where γ_1 is a fold transcription ratio parameter, γ_2, γ_4 are modulation ratio parameters and γ_3, γ_5 are Hill coefficients.

The available model represents the liver circadian organization as a dynamic purely driven by intracellular feedback loops and does not explicitly include the influence of systemic cues such as temperature or hormonal exposure which yet contribute to the liver circadian clock robustness (Ballesta et al., 2017). A key question lays in the molecular links between the cellular clock and those systemic

circadian regulators. Hence, they will be included in a new form of the mouse liver clock model as follows. We consider that the action of systemic regulators $\mathbf{z}$ on the circadian cellular clock is done by a forcing function f, and any feedback from the clock to the systemic regulators is neglected. Two regulations are considered as multiplicative action of the regulators on either gene transcription or gene mRNA degradation so that the dynamics of a gene x in an *in vivo* scenario can be written as one of the following equations:

Hypothesis H1:

$$\begin{aligned} \frac{dx^{\text{vivo}}}{dt} &= f(\mathbf{z}) V_{\max} \text{Transc}(\mathbf{M}, \boldsymbol{\gamma}) - \alpha x^{\text{vivo}} \\ \Leftrightarrow f(\mathbf{z}) &= \frac{\frac{dx^{\text{vivo}}}{dt} + \alpha x^{\text{vivo}}}{\text{Transc}(\mathbf{M}, \boldsymbol{\gamma})} \end{aligned} \tag{4.5}$$

Hypothesis H2:

$$\begin{aligned} \frac{dx^{\text{vivo}}}{dt} &= V_{\max} \text{Transc}(\mathbf{M}, \boldsymbol{\gamma}) - f(\mathbf{z}) \alpha x^{\text{vivo}} \\ \Leftrightarrow f(\mathbf{z}) &= \frac{V_{\max} \text{Transc}(\mathbf{M}, \boldsymbol{\gamma}) - \frac{dx^{\text{vivo}}}{dt}}{x^{\text{vivo}}} \end{aligned} \tag{4.6}$$

where $f(.) \leftarrow V_{\max} f(.)$ for Equation (4.5) and $f(.) \leftarrow \alpha f(.)$ for Equation (4.6). Incorporating $V_{\max}$ and α into the residual trajectories bypasses the need for any assumption on their values as they will be merged with parameters of the considered statistical models, see below.

For *Bmal1*, *Per2* and *Rev-Erbα*, x^{vivo} can be estimated from the gene expression data $\bar{x}^{\text{vivo}}$ available in the four mouse classes. Similarly, the five potential systemic regulators $\mathbf{z}$ which are rest-activity, temperature, food intake, corticosterone and melatonin can be set equal to their measurements in the mouse classes $\bar{\mathbf{z}}$. Upon discretization over the time grid $\{t_i\}_{1 \leqslant i \leqslant N}$, at which the Gaussian processes used for data preprocessing are evaluated, Equation (4.5) and Equation (4.6) can be transformed as:

$$f(\bar{\mathbf{z}}(t_i)) \approx \frac{\frac{\Delta \bar{x}^{\text{vivo}}(t_i)}{\Delta t_i} + \alpha \bar{x}^{\text{vivo}}(t_i)}{\text{Transc}(\mathbf{M}, \boldsymbol{\gamma})} := y(t_i) \qquad \textbf{(H1)} \tag{4.7}$$

$$f(\bar{\mathbf{z}}(t_i)) \approx \frac{V_{\max} \text{Transc}(\mathbf{M}, \boldsymbol{\gamma}) - \frac{\Delta \bar{x}^{\text{vivo}}(t_i)}{\Delta t_i}}{\bar{x}^{\text{vivo}}(t_i)} := y(t_i) \quad \textbf{(H2)} \tag{4.8}$$

Each function y is called a **residual trajectory**. The goal of the study is to identify all possible functions f that would properly fit all residual trajectories y, given the systemic biomarkers measurements in the four mouse classes.

We now define a model learning problem. For the sake of simplicity, we will study the case where systemic regulators only act on either the transcription or

the degradation of a single gene. This gene is either *Bmal1*, *Per2* or *Rev-Erbα* for which we have mRNA level data. For each of the six scenarios (3 genes, action on transcription or degradation), let us consider the following learning samples, for a given class of mice

$$\left\{\left(\bar{\mathbf{z}}(t_i), y(t_i)\right),\ i \in [\![1, N-1]\!]\right\}$$

The problem of finding the optimal functions f can be addressed in a regression setting, by solving

$$\underset{\hat{f}\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{N-1}\sum_{i=1}^{N-1}\left(y(t_i) - \hat{f}(\bar{\mathbf{z}}(t_i))\right)^2 \tag{4.9}$$

for a given family of estimator functions $\mathcal{F}$, such as linear functions or tree-based functions.

$\bar{\mathbf{z}}(t_i)$ are given by the datasets on the circadian rhythms of the five regulators in the four mouse classes so that the principal issue is now to compute the residual trajectories y. They are computed by Equation (4.7) and Equation (4.8) which include: i) $\bar{x}^{\text{vivo}}$ gene expression which are set equal to mouse liver mRNA levels of either *Bmal1*, *Per2* and *Rev-Erbα*; ii) parameters α, $V_{\max}$ and γ which are unknown at this stage, iii) time-resolved concentrations of the modulatory species $\mathbf{M}$ for which no data is available. To estimate the needed parameters and circadian profiles of modulators, we will investigate the circadian clock of liver cells cultured *in vitro*, that is under constant influence or complete absence of the five whole-body regulators.

4.3.3 A model of the *in vitro* liver cellular circadian clock

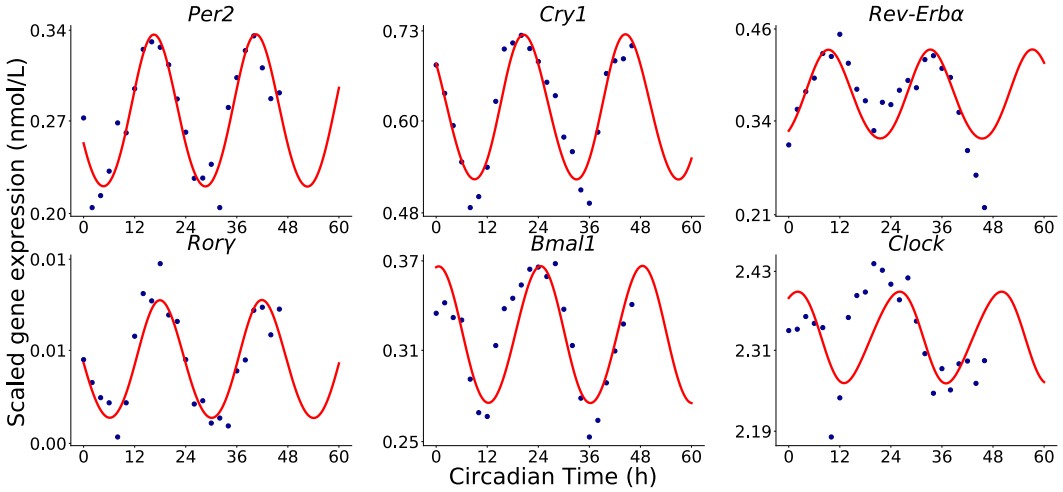


Figure 4.3: Best-fit of the *in vitro* cellular clock model (red curves) to mRNA levels of six clock genes measured in MMH-D3 cell culture (blue dots)

We leveraged time-resolved mRNA expression of six clock genes measured in immortalized MMH-D3 mouse hepatocytes using microarray technology (Atwood et al., 2011). This cell line is often used as a surrogate for healthy hepatocytes. Cells were cultured in standard conditions in which they are exposed to constant temperature and access to nutrients, in the absence of mechanical stress, melatonin or corticosterone addition. Under these *in vitro* conditions, the influence of all regulators are constant over time so that gene expression can be expressed by Equation (4.3). The existing model of the *in vivo* mouse liver clock can thus be used to represent the *in vitro* circadian clock, yet after parameter adaptation based on cell culture data in which the clock is not under rhythmic controls. The MMH-D3 gene expression datasets were used to adjust the parameters of the *in vitro* clock model starting from estimates of the in vivo model (Hesse et al., 2021). Parameter estimation is performed similarly as described in Section 3.2.7. RT-qPCR data from primary mouse hepatocytes culture were used to scale microarray intensities to obtain absolute values of mRNA intracellular concentrations, as required for modeling purpose (Feillet et al., 2016). The fitted *in vitro* model succeeded in capturing the oscillatory behavior of the six clock genes (Fig. 4.3). A total of 10 optimal parameter sets were obtained from different runs of the optimization algorithm, both leading to the same reasonable fit of the data. The result can then be thought of as the cellular clock contribution isolated from the rhythmic influence of the systemic regulators. It will be used to identify specific regulators of the cellular clock in the *in vivo* setting.

4.3.4 Computing residual trajectories for the *in vivo* scenario using the *in vitro* clock model

In Section 4.3.2, we derived an expression for the approximation of the action of systemic regulators on the cellular circadian clock under **(H1)** or **(H2)** . Equation (4.9) formulates the problem in a regression setting, which requires the computation of the residuals trajectories y. The latter necessitates parameter values for α, $V_{\max}$ and γ, as well as the concentrations of the modulatory species $\mathbf{M}$ (Equation (4.7), Equation (4.8)): REV-ERB and ROR for *Bmal1*, CLOCK/BMAL and PER/CRY for both *Per2* and *Rev-Erbα*.

While the adjustment of our circadian clock model to *in vitro* data provided estimates for these quantities, one can question their reliability in the *in vivo* setting. Given that, we have decided to identify leading systemic regulators based on the prediction of **multiple** residual trajectories obtained by varying parameter values of the *in vitro* model. This reduces the dependence of future inference on these estimates, thus ensuring the robustness of the method as functions of systemic regulators f would have to be optimal for numerous different liver clocks. Let $\boldsymbol{\theta}$ be the model parameter vector. For selected coordinates j, we apply an additive

Gaussian noise to *in vitro* values as follows:

Hypothesis H3:

$$\theta_j^{\text{vivo}} = \theta_j^{\text{vitro}} + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}\left(0, \frac{\theta_j}{\sigma}\right) \tag{4.10}$$

with σ a scaling factor, in practice set to 10. The relevant coordinates j are composed of two sets. The first set corresponds to the parameters $\alpha, V_{\max}$ and γ involved in Equation (4.7) and Equation (4.8). These parameters are different for each gene *Per2*, *Bmal1* and *Rev-Erbα*. The second is the set of model parameters that have the greatest impact on the time-concentration profile of modulator species **M**. These are best suited to make the modulators deviate from their *in vitro* concentrations. They were determined for each species through global sensitivity analysis in which outputs are defined as the circadian mean, amplitude or phase of the temporal profile of **M** (Fig. 4.S1, (Sobol, 2001)). For each of these characteristics l and each modulator m, we selected the parameter set $\mathcal{P}_{m,l}$ comprised of the p most sensible parameters according to Sobol sensitivity indices. Then, the intersection of $\mathcal{P} = \bigcap_{l,m} \mathcal{P}_{l,m}$ was computed. p was chosen such that $\#\mathcal{P} = 5$ where # is the cardinal of a set. Among these parameters were found 3 degradation parameters for *Bmal1*, *Clock* and CLOCK/BMAL$_N$ as well as 2 cytoplasmic protein production parameters for CLOCK$_C$ and BMAL$_C$. All selected sensible parameters were related to the CLOCK/BMAL loop.

Under **(H3)** , additive gaussian noise is applied to each of the 10 *in vitro* parameter sets and fed to the model to compute corresponding clock variables time profiles. Considering multiple optimal parameter sets allows us to reduce the parameter uncertainty related to the lack of constraints. Selection criteria are applied in order to only select realistic clocks: i) variable concentrations outputted should be periodic with period between 20 and 28h, and display relative amplitude above 5%, ii) the phase difference between the nuclear variables REV-ERB and ROR, and between PER/CRY and CLOCK/BMAL complexes should be larger than 6 h, as this two couples are made of variables that play antagonist roles (Ko and Takahashi, 2006). If all these criteria are met, the model simulation and its associated parameter set are kept and the corresponding trajectory from Equation (4.7) or Equation (4.8) is computed, using the perturbed parameter vector. This procedure is repeated until n trajectories are obtained, for each mouse class and gene. In practice n was set to 2000. Inter-class differences of circadian amplitudes and phases could be observed between the trajectories generated for each of the four mouse classes as a result of variations present in the clock gene expression data (Fig. 4.4).

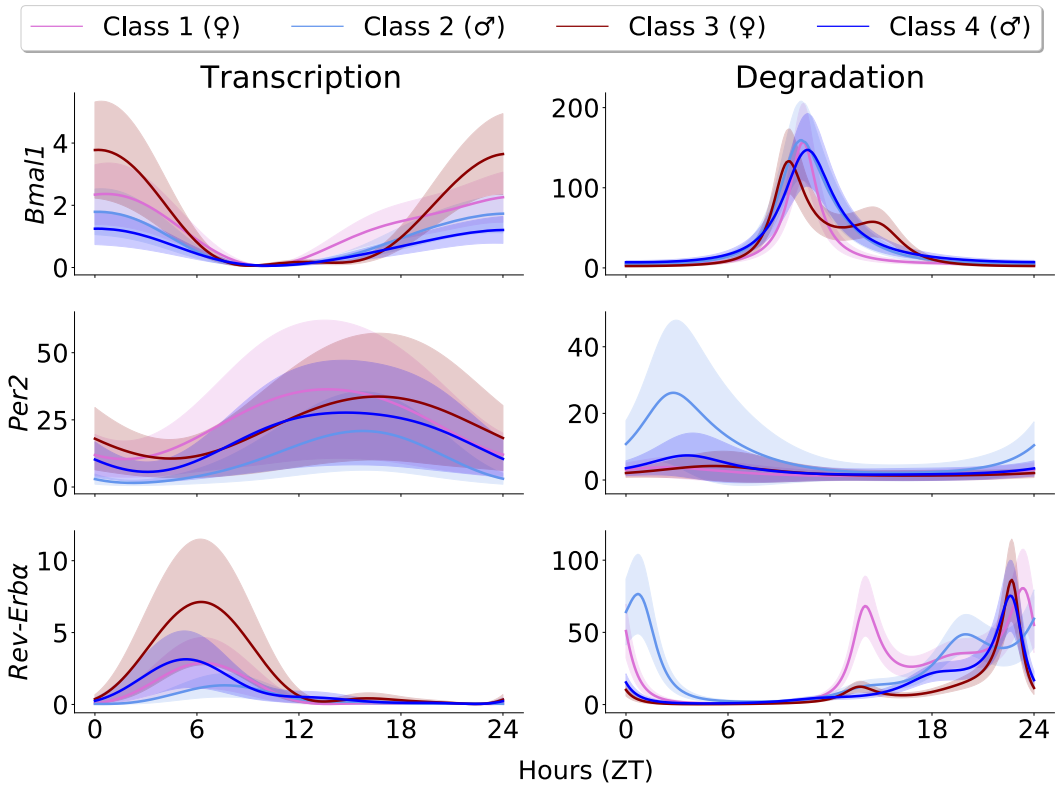


Figure 4.4: Mean and standard deviation of the selected residual trajectories obtained for each clock gene. Left (resp. right) panels under **(H1-H3)** (resp. **(H2-H3)**).

4.3.5 Identifying the action of systemic regulators as a linear regression problem

The most straightforward way to solve the problem evoked in Equation (4.9) is to compute an estimator of f thanks to linear regression. This is biologically meaningful as chemical reactions can often be written using the law of mass action that assumes linear kinetics. Besides, estimators provided in this case are easy to interpret as the contribution of a regulator z_j is represented with a weight β_j:

$$\hat{f}(\bar{\mathbf{z}}(t_i)) = \sum_j \beta_j \bar{z}_j(t_i) \tag{4.11}$$

We assume the same model structure, i.e. active regulators for all mouse classes. Only weights $\boldsymbol{\beta}$ can vary across mouse strains and sexes. From a biological point of view, this is equivalent as saying that the involved regulators are the same whatever the mouse category, although the strength of their influence may vary classwise. For the sake of simplicity, any model containing both a regulator and its corresponding integral regulator is ruled out. This constraint ensures that in a model, a systemic regulator has only one way to act on the gene: either directly or

indirectly. Thus, a regression model can include at most 5 terms. To select the first term, 10 choices are possible, then 8, then 6, etc since once a regulator is chosen, its associated integral regulator cannot be selected for the current model. We end up with $(10 \times 8 \times 6 \times 4 \times 2)/5! = 32$ possible models involving exactly five regulators. The general formula below shows that there is a total of 242 models when including one to five regulators.

$$\sum_{r=1}^{5} \frac{1}{r!} \prod_{j=5-r+1}^{5} 2j \tag{4.12}$$

Considering that there are n residual trajectories $y_k^{(c)}$ for each of the four classes, the learning samples become:

$$\left\{ \left(\bar{\mathbf{z}}^{(c)}(t_i), y_k^{(c)}(t_i) \right),\ i \in [\![1, N-1]\!],\ c \in [\![1,4]\!], k \in [\![1,n]\!] \right\}$$

For each class c, let us define the loss of a given model $\hat{f}$ parametrized by $\boldsymbol{\beta}_k^{(c)} = \left(\beta_{k,j}^{(c)}\right)_{j \in [\![1,10]\!]}$ applied to the class regulator data $\bar{\mathbf{z}}^{(c)}$ against trajectory $y_k^{(c)}$ as:

$$\ell(y_k^{(c)}, \bar{\mathbf{z}}^{(c)}, \boldsymbol{\beta}_k^{(c)}) := \frac{1}{N-1} \sum_{i=1}^{N-1} \left(y_k^{(c)}(t_i) - \sum_j \beta_{k,j}^{(c)} \bar{z}_j^{(c)}(t_i) \right)^2 \tag{4.13}$$

The total error associated to this model is defined as the average of the errors of each residual trajectories across the four classes. It is computed as,

$$\mathcal{E}\left(y, \boldsymbol{\beta}, \bar{\mathbf{z}}\right) := \frac{1}{4n} \sum_{c=1}^{4} \sum_{k=1}^{n} \min_{\boldsymbol{\beta}_k^{(c)}} \ell\left(y_k^{(c)}, \bar{\mathbf{z}}^{(c)}, \boldsymbol{\beta}_k^{(c)}\right) \tag{4.14}$$

Finally, to allow comparison of errors and coefficients for different trajectories, both the inputs and outputs of the regression problem are standardized with zero mean and standard deviation one. Therefore the loss between a trajectory y_k and an empty model is 1. One can see this value as an upper bound for the performance of a model, providing an assessment of the goodness of fit.

4.3.6 Regulator importance through Shapley values

An important question in an inference setting is to determine the precise set of relevant features in terms of prediction. Here, the fact that we deal with only ten features coupled with the low complexity cost of linear regression tolerates an exhaustive search over the whole regulators model space. Consequently, our inference considers large linear regression models as a first step and focus on smaller models thereafter. The first step of our method to identify relevant regulators for each clock gene uses Shapley values. Shortly, Shapley values stem from Game

Theory and allocate to each feature z_j a value ϕ_j that represents the effect of including that feature on model predictions. It is computed as the following weighted average:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{\#S!\,(\#F - \#S - 1)!}{\#F!} \left(\hat{f}_{S \cup \{j\}}(\bar{\mathbf{z}}_{S \cup \{j\}}) - \hat{f}_S(\bar{\mathbf{z}}_S) \right) \tag{4.15}$$

where F is the set of all feature indices, S a subset of F and $\mathbf{z}_S$ the vector of features with indices in S. Since the effect of z_j depends on other features, the model differences are computed for all possible subsets of features (Peters, 2015). This approach was recently extended to handle any machine learning model such as tree-based models or neural networks (Lundberg and Lee, 2017). For linear models, one can derive a simpler formula: $\phi_j(t_i) = \beta_j \bar{z}_j(t_i)$.

We computed Shapley values for all possible regulator models involving 5 features, which is the maximum size of the model if excluding concomitant direct and indirect action of the same regulator. From Equation (4.12), there are 32 such admissible subsets of regulators $\mathbf{z}_S$. We call I the set containing all possible indice subsets S of cardinal 5, excluding those containing indices of direct and indirect actions of the same regulator. Pipeline 1 shows the procedure to compute the mean absolute Shapley values.

4.4 Results

4.4.1 Action of systemic regulators on clock gene transcription

Our first aim is to investigate possible actions of systemic regulators on the transcription of the three clock gene for which we have mRNA data: *Bmal1*, *Per2* and *Rev-Erbα*. Thus, in this section, we consider action of regulators in the form of **(H1)** and residual trajectories are computed under **(H3)** (Fig. 4.4, left column).

The importance of each regulator was assessed through the computation of Shapley values for each possible linear estimator $\hat{f}$ (Fig. 4.5). One should notice from Pipeline 1 that these values are averaged across mouse classes, residual trajectories, and time points. Remarkably, the lowest score is achieved by Melatonin and its indirect version $\int$Melatonin, for all three clock genes. This means that according to the Shapley values metric and based on linear regression models, the melatonin is the least relevant contributor to the prediction of the trajectories y. This is in agreement with biological knowledge and thus provides a partial validation of the approach. Conversely, Shapley values yielded as leading regulator $\int$Temperature for all three genes, advocating for a strong effect of temperature cycles on the cellular clock, yet through indirect actions involving an intermediate species.

While Shapley values give a coarse-grained ranking of the regulators, another level of granularity can be achieved. As mentioned earlier, the small dimension of

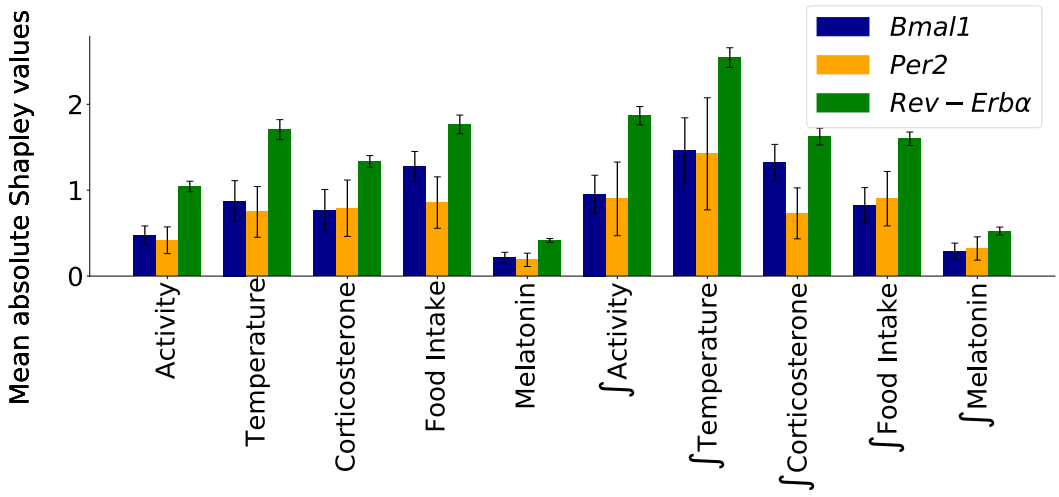


Figure 4.5: Mean absolute Shapley values for all features and each gene under **(H1-H3)** . Standard deviations computed across residual trajectories.

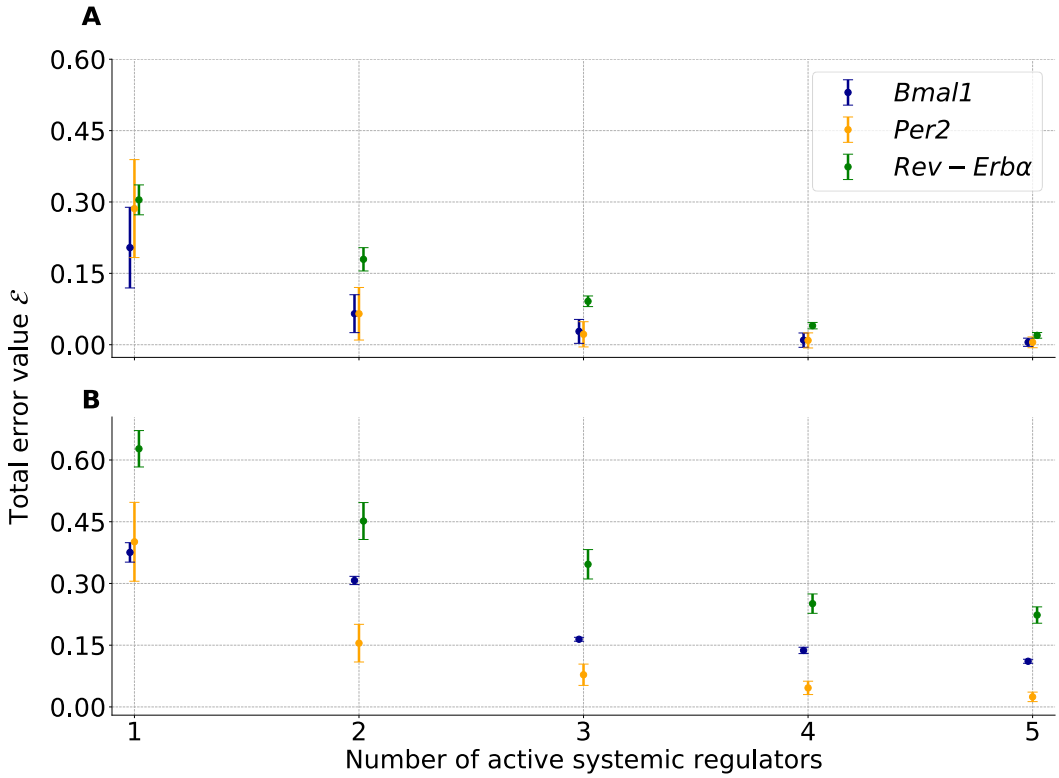


Figure 4.6: For each gene, the total error of the best-fitting model depending on its number of nonzero terms is reported for Transcription (**A**) and Degradation (**B**). Standard deviations are taken across residual trajectories.

the problem allows for an exhaustive search of all possible linear models. Under the constraint that no regulator is found twice in the same model with both a direct

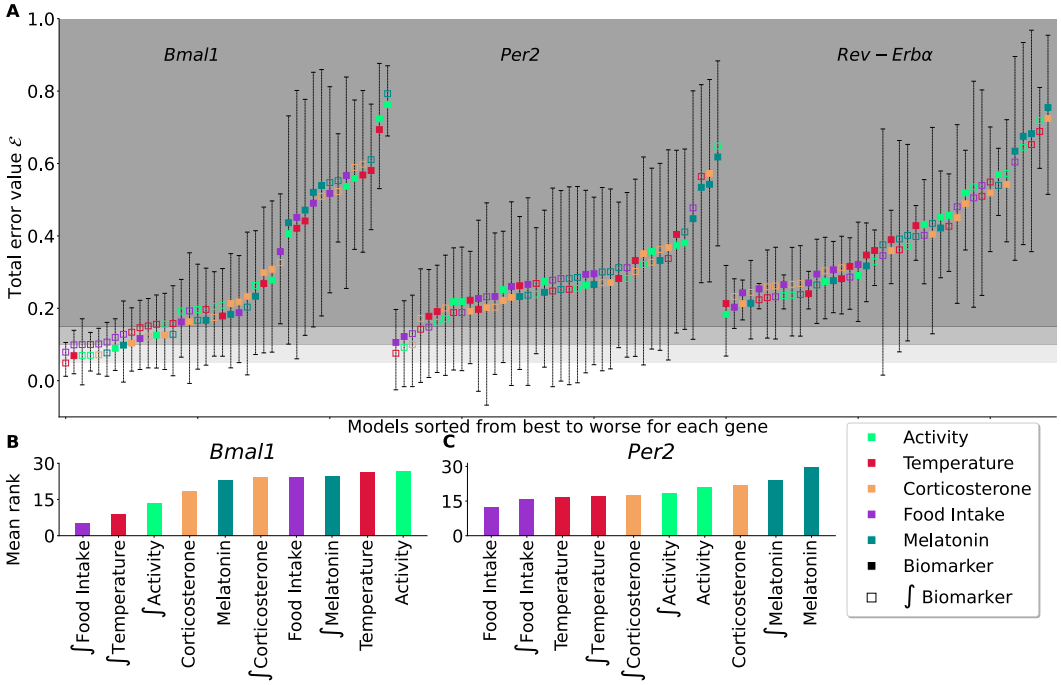


Figure 4.7: **(A)** Total error for each of the 40 2-term models representing a systemic control on either *Bmal1*, *Per2* or *Rev-erbα* transcription. Means and standard deviations were computed across residual trajectories under **(H1-H3)** . Colors indicate regulators involved in each model, with the top square referring to the dominant regulator. Areas defined by different shades of gray refer to thresholds of total error equal to 0.05, 0.1 and 0.15. **(B-C)** Mean rank of regulators among the 40 2-term models, from lowest to highest, for models impacting *Bmal1* or *Per2* transcription.

and indirect action, there are 242 models (Equation (4.12)). For each gene, Fig. 4.6 displays the total error $\mathcal{E}$ of the best model across residual trajectories, involving from 1 to 5 regulators. Model overfitting was investigated as follows. For each residual trajectory, time points were shuffled and divided in 4 folds on which cross validation was performed. A close agreement between training and testing total errors was found indicating that overfitting was not an issue for any considered number of systemic regulators (Fig. 4.S2). As expected, the total error decreased as terms were added to the models. The slope was found to be the steepest when moving from 1-term models to 2-term models for all genes, demonstrating the superiority of the latter in terms of balance between degrees of freedom and goodness of fit. Furthermore, for *Bmal1*, best 1-term, 2-term and 3-term models were nested, thus Fisher test could be applied. These models were $\int$Temperature, $\int$Temperature + $\int$Food Intake and finally $\int$Temperature + $\int$Food Intake + $\int$Activity. The 2-term model, was found to be significantly better than the 1-term and the 3-term model ($P < 0.05$). Hence, we now focus on 2-term models which were all fitted to residual trajectories (Fig. 4.7). For each gene, there exists exactly $\frac{10\times8}{2!} = 40$ such models. For

each model, the dominant term is defined as the regulator with indice j maximizing the following quantity: $\frac{1}{4n}\sum_{k=1}^{n}\sum_{c=1}^{4}|\beta_{k,j}^{(c)}|$.

The best model including a control of *Rev-Erbα* transcription achieved a poor fit to data with a total error of 0.2 which led us to discard all models for *Rev-Erbα* (Fig. 4.7, Fig. 4.S3) For both *Bmal1* and *Per2*, one can observe a large preponderance of the regulators food intake and temperature among the top ranked models (Fig. 4.7). These regulators end up with the lowest mean ranks across all systemic regulators, though only through an indirect action for *Bmal1*. For *Per2*, temperature ends up being the most present systemic regulator, involved, through direct or indirect action, in 6 out of the 10 best-performing models. This is consistent with $\int$Temperature having the highest Shapley value for this gene (Fig. 4.5). Moreover, this finding is in agreement with the observation of an effect of the temperature on *Per2* transcription through Heat Shock Proteins reported in (Kornmann et al., 2007) and provides a form of validation of our approach.

Over all 2-term models fitted for each gene, melatonin first rankings as a leading biomarker were found to be quite high: 28^{th}, 22^{th} and 20^{th} for *Bmal1*, *Per2* and *Rev-Erbα*, respectively. This comes as further validation of this approach as melatonin is included here as a negative control since liver cells do not express its receptors.

4.4.2 Action of systemic regulators on clock gene mRNA degradation

In this section, we search for possible actions of the systemic regulators on clock gene mRNA degradation, assuming **(H2-H3)** hold. Fig. 4.4 (right column) shows the residual trajectories computed with the method described in Section 4.3.4. These time profiles appear strongly nonlinear, with sharp peaks for *Bmal1* and *Rev-Erbα* as a result of the division by the clock gene concentrations which are close to zero for certain circadian time window. Such shapes suggest that a systemic regulation of clock gene mRNA degradation would lead to an unstable control that would explode during some interval of the 24h span. This type of behavior is unlikely to derive from the realisation of natural biological processes which mostly produce robust patterns over time. Consequently, the same analysis as in the transcription case yielded to the exclusion of all models. First, for *Bmal1*, a minimum of three to four terms is necessary to achieve a proper fit of the residual trajectories, with respective total errors of 0.16 and 0.14 (Fig. 4.6**B**). In the case of *Rev-Erbα*, even 5-term models are far from producing reasonable fits, with a total error of 0.22 for the best 5-term model. For *Per2*, as in the transcription case, the slope of the total error was found to be the steepest when moving from 1-term models to 2-term model, so that all models with number of terms greater than 2 are rejected. However, with a total error of 0.31, 0.19 and 0.46 for *Bmal1*, *Per2* and *Rev-Erbα* respectively, there is no 2-term model providing a good fit of the trajectories. Total errors of all 2-term models are presented in Fig. 4.S4. Overall, we conclude under **(H2-H3)** , that there

is no admissible models involving a linear action of the regulators on clock gene mRNA degradation.

4.4.3 Mouse class differences

As data for four mouse classes (2 strains, 2 sexes) are available, we can investigate the effect of sex and genetic background on the regulators action. Indeed, one perk of linear models is their simplicity when it comes to providing explanations: the impact of a feature on the prediction is determined by the weight associated to this feature. This enables the study of mouse class differences in terms of regulator weights. Weight distributions were estimated for each mouse class from best-fit parameters obtained across all trajectories through kernel density estimation (Fig. 4.8). Using all 2-term models for both *Bmal1* and *Per2* under **(H1-H3)**, we performed two-way ANOVA, asking whether or not genetic background or sex is statistically significantly impacting regulator weights. In that event, the values of regulator weights in a given model, obtained by fitting each residual trajectory, are considered as realisations of a random variable. For *Bmal1* (resp. *Per2*), 38 (resp. 37) out of 40 models agreed on the statistically significant influence of sex and genetic background on the extent of regulators influence ($P < 0.05$). Interactions between both factors were also found to account for differences in regulators weights in 38 (resp. 37) models for *Bmal1* (resp. *Per2*). Models failing to uncover statistically significant differences were all associated with a total error above 0.2.

This finding matches the fact that circadian rhythms display sex differences in mice and in humans (Ballesta et al., 2017). Moreover, previous findings demonstrated different optimal timing of the anticancer drug irinotecan in these four mouse classes (Li et al., 2013).

A closer look to the weight distributions is given for *Bmal1* and *Per2*'s best fitting model (Fig. 4.8). Interestingly, large inter-class differences can be found for the regulator weights. The best 2-term model integrating a control of *Bmal1* transcription, involves the joint action of Food intake and Temperature, probably through intermediate species. Food intake appears to act mostly negatively on Bmal1 transcription in female and male B6D2F1 mice (Classes 1 and 2) and positively in B6CBAF1 mice (Classes 3 and 4). For Temperature, the exact reverse situation is observed. For each mouse strain, sex-specific differences are also present in the distribution modes and standards deviations, Class 1 displaying the largest variability across trajectory best-fit parameters. Next, the best model targeting *Per2* expression involves a direct positive regulation of the gene mRNA transcription by Food Intake and an indirect mostly negative influence of temperature for all mouse classes. The distributions of Food Intake weight present different shapes for Class 1 and 2, with a higher mode for Class 2, while being analogous for Class 3 and 4. Regarding the indirect action of Temperature, the distribution for

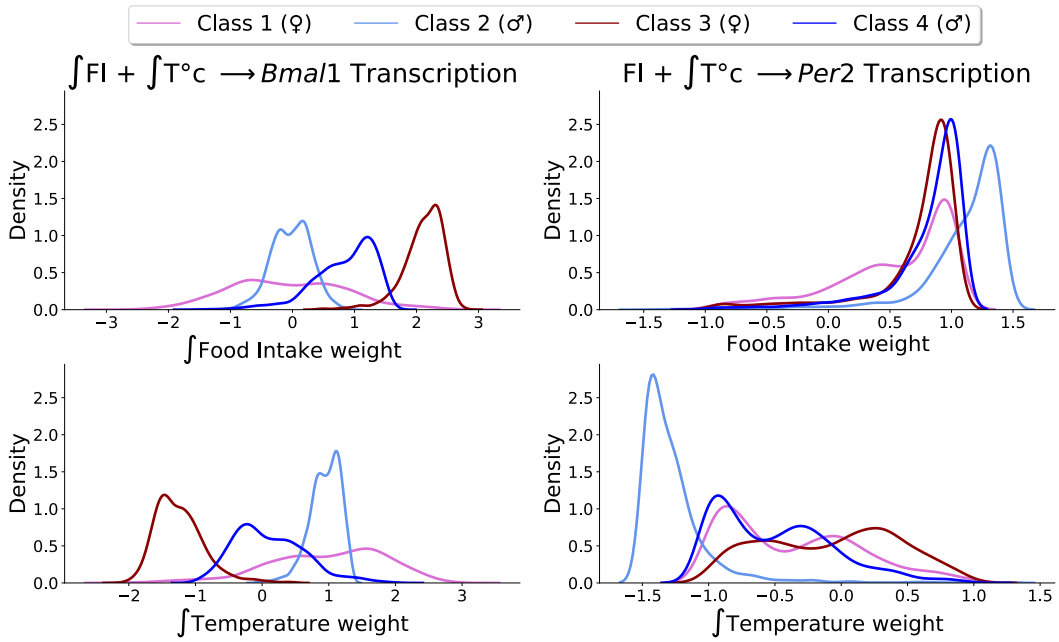


Figure 4.8: Density plot of the coefficients of the best 2-term model for *Bmal1* and *Per2*, computed across residual trajectories for each mouse class. FI: Food Intake, T°c: Temperature.

Class 2 is almost set apart from the others with a lower mode, hence a stronger negative impact. The fact that parameter distributions can have classwise opposite modes raises a few questions as this would imply that systemic regulators could act either positively or negatively on gene transcription across mouse classes. As a start, model identifiability was assessed by means of profile likelihood, a method determining practical and structural identifiability (Raue et al., 2009). Fig. 4.S5 and Fig 4.S6 show that the parameters of the best 2-term models for *Bmal1* and *Per2* are indeed identifiable for each class. Subsequently, similar systemic regulator weight signs across classes was enforced and models which did not initially meet this constraint, i.e. 23 for *Bmal1* and 32 for *Per2*, were re-optimized. A 1.5-fold average increase in total errors from unconstrained to constrained optimization was found (Fig. 4.S7). Altogether, these findings demonstrate that models with opposite signs were reasonable and better fit data than constrained models.

4.5 Discussion

We have presented a model learning methodology to identify systemic regulators of the peripheral circadian clocks. The theoretical approach comprises two key steps. The first was based on the integration of extensive prior knowledge on the mammalian circadian timing system into an ODE-based circadian clock model. The

comparison of this calibrated model with available circadian datasets allowed the derivation of an approximation for the action of the regulators on the clock in the form of residual trajectories. In a second step, using a linear regression framework, the task of inferring systemic regulators of the clock was interpreted as a model selection problem. The latter involving a small number of features, an exhaustive exploration of the regulator model space could be performed. Thus, we used Shapley values to draw inference on the importance of each regulator from large regression models and acquired a more fine-grained understanding with smaller models afterwards.

Our approach produces explainable linear models that mechanistically represent the action of the measured regulators on clock genes in two mouse strains. The focus was given to five regulators for which measurements were accessible: biomechanical stresses (derived from rest-activity), body temperature, nutrient exposure (derived from food intake), plasma melatonin and corticosterone. Given the available mRNA data, we were able to investigate systemic regulation of *Bmal1*, *Per2* and *Rev-Erbα* mRNA transcription or degradation. Models involving a modulation of mRNA degradation were all rejected, as well as those impacting *Rev-Erbα* transcription. Hence, all admissible models included regulation of either *Bmal1* or *Per2* transcription. Temperature was found to affect *Per2* transcription in an indirect manner, which was in line with temperature dependency of the expression of HSPs that interact with clock genes (Kornmann et al., 2007). Similarly, melatonin which was included as a negative control was not involved in the best models. Lastly, the large predominance of food intake in the best fitting models agreed with recent experimental findings (Greenwell et al., 2019). Indeed, modulating meal timing and composition impacts liver clock genes time profiles as, for instance, dampened oscillations were shown in mice subjected to high-fat-diet, whereas time restricted high-fat-diet restored regular circadian rhythms (Hatori et al., 2012; Li et al., 2010). Arrhythmic feeding does not cause liver clock genes to lose oscillations in mice, a behavior which is well reproduced by our models (Greenwell et al., 2019). A subsequent step would be to study the precise molecular mechanisms linking energy metabolism and the clock which requires the design of dedicated systems biology frameworks (Woller et al., 2016). Next, our approach assumes independence of the systemic regulators while this may not be the case for all of them. However, independence of temperature and food intake seems to have been validated in experiments, where different feeding patterns led to similar temperature profiles in mice (Greenwell et al., 2019). Lastly, classwise opposite action of the systemic regulators was found to be necessary to ensure reasonable fit of the trajectories. Biologically speaking, differences in influences of regulators are plausible. It may imply that a systemic regulator activates different regulatory pathways for each mouse class as a consequence of different gene expression levels. For instance, it was recently found that ubiquitin associated pathways regulated the cellular clock

only in female and not in male mice (Mekbib et al., 2020).

The theoretical approach developed here could be extended to handle more complex model learning scenario. We have focused on the identification of systemic regulators on gene mRNA degradation and transcription, the latter also being the starting point of gene regulatory network learning algorithm such as Dyngenie3 (Huynh-Thu and Geurts, 2018). In our case, conditionally to the availability of additional data on other species present in the clock model (proteins and protein complexes), this method could be applied to search for systemic regulations on any process included in the ODEs (e.g., nuclear translocation or protein production). For larger problems, exhaustive model search could be replaced by machine learning methods like sparse multi task regression, to find a parsimonious set of optimal predictors while enforcing the same structure across mouse classes (Lozano and Swirszcz, 2012). Otherwise, nonlinear models can be searched for with sparse regression tools (Brunton et al., 2016) or virtually any machine learning algorithm, whose output can be explained, e.g by SHAP (Lundberg and Lee, 2017). Although in practice, ensuring similar structure across classes might be difficult in some cases.

As a perspective, the best models inferred from this study will be integrated back in our ODE-based clock model and parameters will be updated based on available data. The validated models will then be tested in dedicated preclinical experiments. Such an approach has been successfully employed using a small number of ODE-based models and allowed discovering new molecular interactions between clock genes and the protein p53 (Gotoh et al., 2016). The next step will be the scaling of the model for humans in order to predict molecular clocks from the measurements of circadian biomarkers using wearable technologies. This will shortly be possible thanks to the availability of clinical datasets including both clock gene expression in the oral mucosa and longitudinal measurements of circadian biomarkers in the same individuals. Such human model of the CTS could then be connected to drug chronoPK-PD models to derive patient-specific optimal timing, as the one built in Chapter 3.

This chapter has presented an innovative model learning method to infer systemic regulators of the peripheral circadian clock. It is particularly amenable when prior knowledge exists and can be encompassed in an ODE-based model to be further completed. Missing parts are then assumed to be located at specific places like gene transcription or gene mRNA degradation, a commonly-used hypothesis in such methods (Gotoh et al., 2016; Huynh-Thu and Geurts, 2018). Next, model selection can be applied. This being said, extensive prior knowledge may not always be available when trying to infer regulatory networks. That is why the next chapter is dedicated to the design of a new model learning algorithm focusing on inferring the most important reactions when nothing is known on the underlying molecular interactions at stake.

Appendix

Parameter estimation

Calibration of the cellular clock model against MMH-D3 hepatocytes mRNA expression was performed using the evolutionary algorithm CMA-ES (Hansen and Ostermeier, 2001), considered state of the art for parameter optimization in settings where gradient information cannot be easily accessed, typically in systems biology. The cost function was defined as the sum of squares between the time-resolved mRNA expression and the model simulated timecourses.

Exhaustive model search with linear regression was performed using the Scikit-learn package in Python (Pedregosa et al., 2011). Computations were done on a laptop with a 2.9GHz Intel core I5 dual core.

Supplementary Pipeline

Pipeline 1: Shapley values computation.

Result: Shapley values summed over classes and trajectories

initialization $\phi = 0_{\mathbb{R}^{10\times(N-1)}}$

for *all sets of regulator indices* $S \in I$ **do**

  for $c = 1, \ldots, 4$ **do**

    for $k = 1, \ldots, n$ **do**

      $\beta_k^{(c)} = 0_{\mathbb{R}^{10}}$

      $\beta_{k,S}^{(c)} \leftarrow \underset{\beta_{k,S}^{(c)}}{\operatorname{argmin}}\, \ell(y_k^{(c)}, \bar{z}_S^{(c)}, \beta_{k,S}^{(c)})$

      for $i = 1, \ldots, N-1$ **do**

        $\phi_S(t_i) \leftarrow \phi_S(t_i) + |\beta_{k,S}^{(c)}\, \bar{z}_S^{(c)}(t_i)|$

$\phi \leftarrow \dfrac{\phi}{4n\#I}$

Supplementary figures

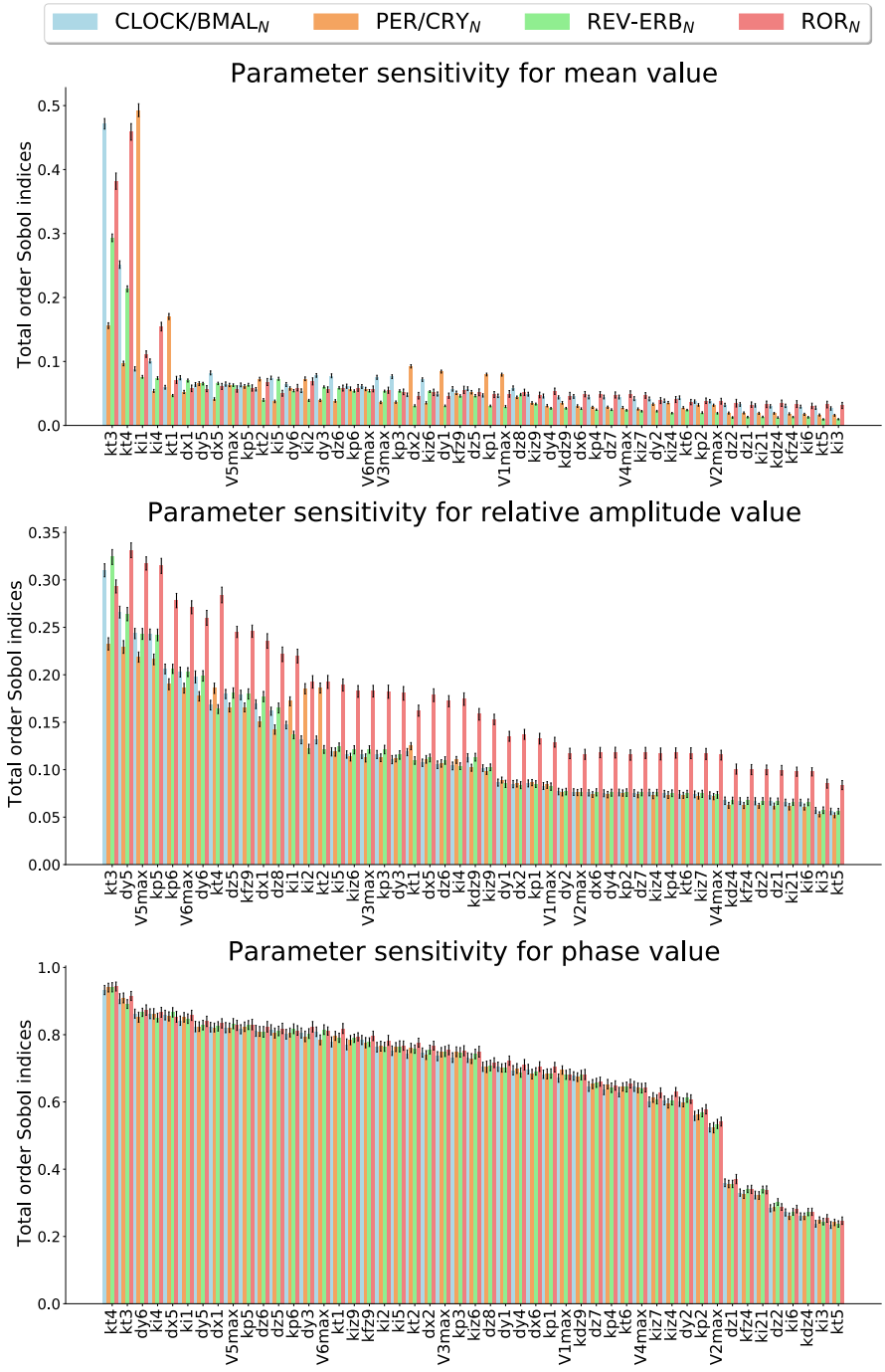


Figure 4.S1: Sensibility analysis performed on the mean, relative amplitude and phase of the modulatory variables of the *in vitro* circadian clock model. Total order sobol indices plotted, with estimated standard deviations. Parameter definitions provided in Supplementary Section S2-1.2.

Figure 4.S2: For each gene, the total error of the best-fitting model depending on its number of nonzero terms is reported under **(H1-H3)**. Timepoints were shuffled and divided in 4 folds on which 4-fold cross validation was performed.

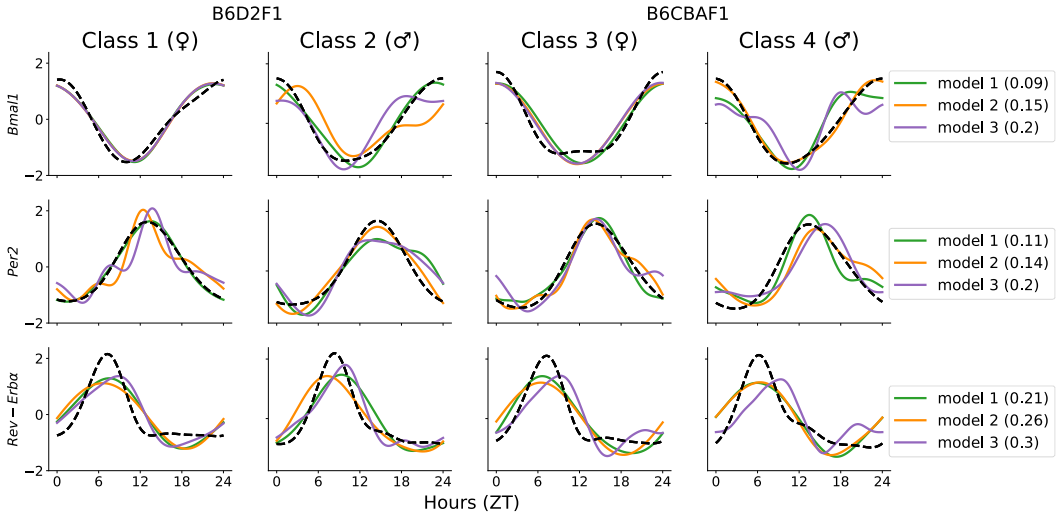


Figure 4.S3: Fit of one residual trajectory (dashed black curves) for each gene and each class. 3 different models (solid colored curves) are used for the sake of error visualization.

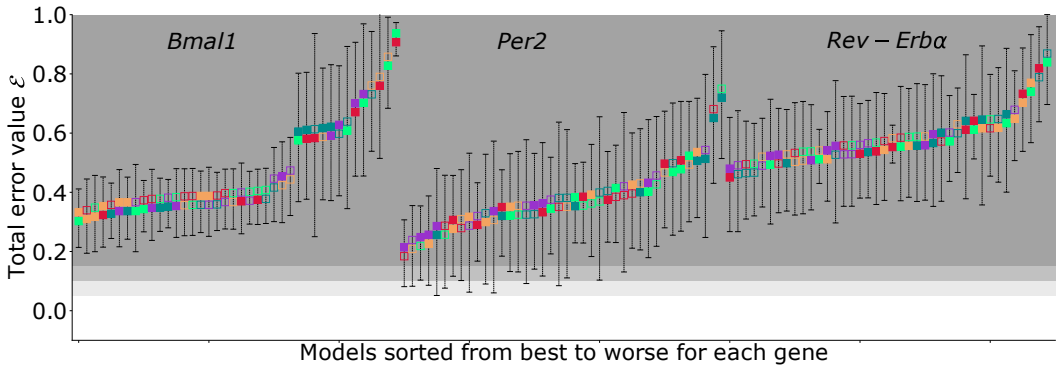


Figure 4.S4: Total error for each of the 40 2-terms models, for all genes. Mean and standard deviation were computed across residual trajectories under **(H2-H3)** . Colored squares at the mean describe the regulators involved for each model, with the top square referring to the dominant regulator. Areas defined by different shades of gray refer to thresholds of total error equal to 0.05, 0.1 and 0.15.

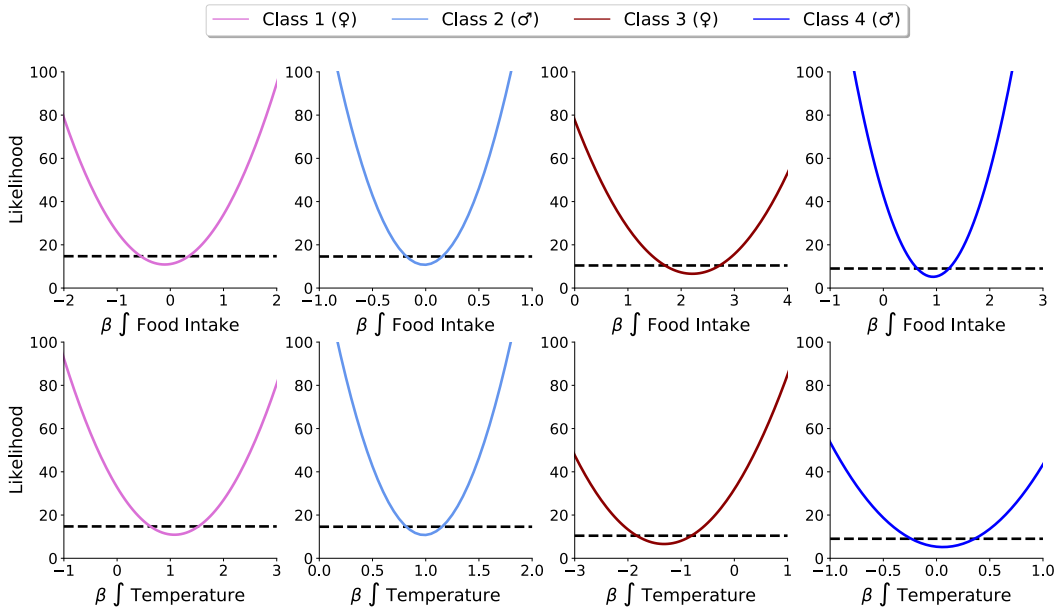


Figure 4.S5: Profile likelihood obtained for *Bmal1* best 2-term model under **(H1-H3)**. Black dashed lines delimit the 95% confidence interval in which the parameters are identifiable.

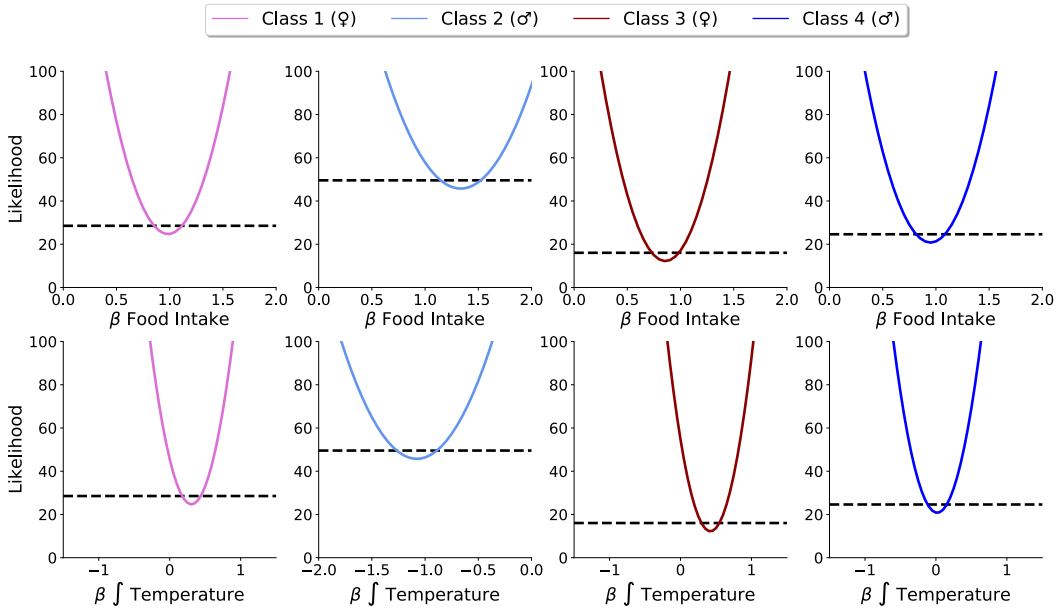


Figure 4.S6: Profile likelihood obtained for *Per2* best 2-term model under **(H1-H3)**. Black dashed lines delimit the 95% confidence interval in which the parameters are identifiable.

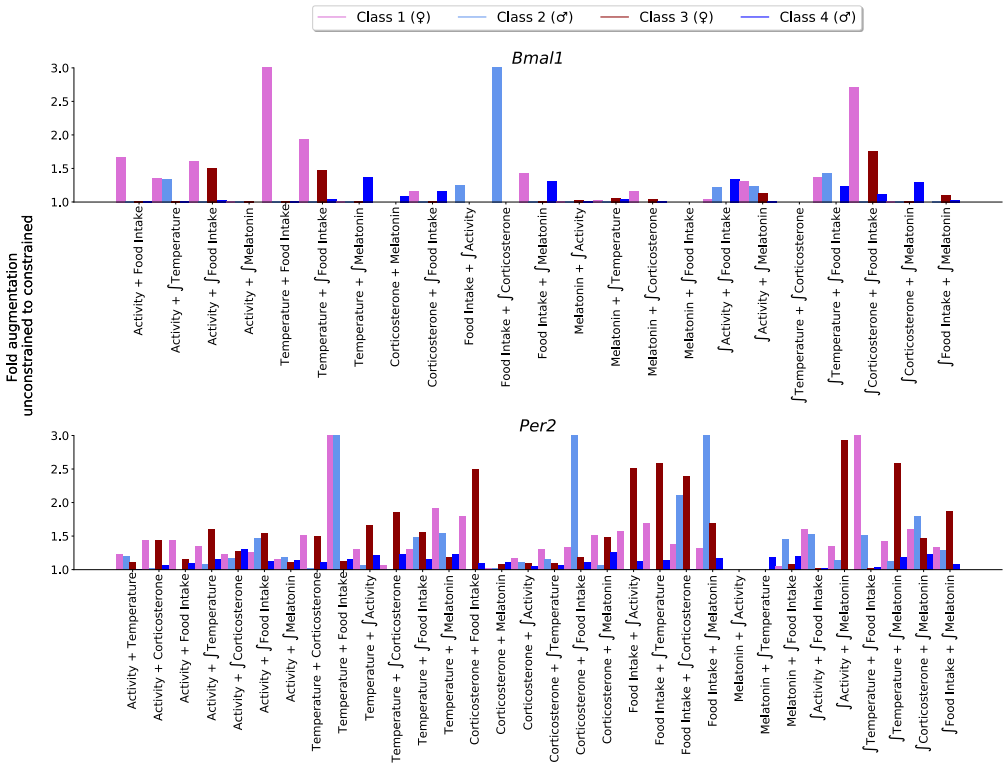


Figure 4.S7: Under **(H1-H3)**, for each 2-term model whose unconstrained optimal parameters did not meet the classwise sign equality constrain, the fold augmentation between the total error in the unconstrained case versus the constrained case is plotted. Only classes where the sign of at least one regulator was modified from unconstrained to constrained optimization are plotted.

Chapter 5

Reactmine: an algorithm for inferring biochemical reactions from time series data

Scientific production

The content of this chapter is based on the following articles:

- J. Martinelli, S. Soliman, A. Ballesta, F. Fages *"Reactmine: an algorithm for inferring biochemical reactions from time series data"* In preparation, 2022.
- J. Martinelli, J. Grignard, S. Soliman, F. Fages, *"On Inferring Reactions from Data Time Series by a Statistical Learning Greedy Heuristics"* International Conference on Computational Methods in Systems Biology, 352-355, 2019.
- J. Martinelli, J. Grignard, S. Soliman, F. Fages, *"A statistical unsupervised learning algorithm for inferring reaction networks from time series data"* International Conference on Machine Learning-Workshop on Computational Biology, 2019.

The method which is presented is implemented in the package Reactmine available on GitLab https://gitlab.inria.fr/julmarti/crninf

Abstract Inferring biochemical reactions from the observation of time series data is a challenge which arises from the need for mechanistic understanding of biological phenomena and from the rapid growth of experimental data availability. This motivates the design of algorithms to suggest preponderant molecular interactions and help the human modelers in their effort of reconstructing intracellular pathways. ODE-based inference methods nowadays resort to least squares regression combined with sparsity-enforcing penalization, e.g. LASSO, leveraging the idea that in biological networks, not all variables are connected. These approaches can benefit from *knockout* experiments, where some genes are silenced, thus enabling the observation of different influences in an independent

manner. However, these *knockout* experiments may not be available when dealing with real data and the input time series may only originate from *wild type* measurements in which all reactions are thus present at all time. In this setting, we observe that the current off-the-shelf learning methods fail to learn sparse influence or reaction networks.

We present Reactmine, a reaction inference algorithm which enforces sparsity by proceeding in a sequential fashion, with the ambition of discovering the preponderant reaction candidates responsible for the observed time series data, rather than aiming at a complete reconstruction of the underlying network. We first conduct performance evaluation on synthetic data as well as parameter sensitivity analysis, and then apply Reactmine to two sets of real data: one from cell cycle / circadian clock protein fluorescence videomicroscopy, and one from biomedical experiments for the task of learning the controls of systemic factors on peripheral clock gene expression. We show that in both cases, Reactmine succeeds in inferring meaningful reactions.

Contents

5.1 Introduction

With the automation of biological experiments and the increase of quality of cell measurements, automating the building of mechanistic models from data becomes conceivable and a necessity for many new applications. Traditionally, the structure of these models, e.g. chemical reaction networks, is inferred from an extensive review and subsequent compilation of the literature by the modeler. More recently, efforts have been made to develop so-called model learning algorithms to assist humans in that burden in order to automate model structure design, for instance in situations where time series measurements are the only observations available. The exercise of inferring the mechanism of action of new compounds, which may be drug candidates, is such an example (Vertes et al., 2018). Therefore, there is a need to develop algorithms to learn dynamical models from time series data without relying on complete knowledge of the structure of the molecular mechanisms.

In this setting, extensive literature concerning gene regulatory network (GRN) inference is available, partly motivated by experimental design (King et al., 2004) or knowledge discovery problems presented in the DREAM challenge (Stolovitzky et al., 2007). A GRN consists in a graph $G = (W, E)$: if a transcription factor W_i binds to the promoter region of a target gene W_j, then the graph involves an edge E_{ij}. GRN inference algorithms feature a wide range of machine learning methods, e.g. Decision Trees (Huynh-Thu and Geurts, 2018), Information Theory (Zoppoli et al., 2010) or Gaussian Processes (Aalto et al., 2020). Let us also mention the work from Aubin-Frankowski and Vert (2020) which leverages the ODE framework but for GRN inference only, with an output limited to a ranking of reactions without the dynamics.

The problem tackled here concerns a quite new field: inferring chemical reaction networks (CRNs) instead of GRNs. A CRN can be represented as a bipartite graph $G = (U, W, E)$ where U and W are two disjoint sets representing respectively the reactions and the species. Every edge connects a vertex in W to one in U. Of note, the indegree of the reaction nodes can be above one, which allows for bimolecular reactions like complexations ($A + B \implies C$). Finally, a notion of dynamics is conveyed by a CRN, using classical rate functions such as the mass action law, or Michaelis Menten kinetics. These aspects make the above described methods hardly suitable for inferring CRNs. Instead, current tools in this context, such as SINDy (Sparse Identification of Nonlinear Dynamics, Brunton et al. (2016)), rely on ODE modelling of the system, and propose to learn the terms composing each ODE by selecting them among a library of potentially nonlinear functions. The main assumption is that the dynamics of each variable can be expressed using only a few elements, so that techniques like sparse regression can be used to select the relevant members of the library. This being said, selecting the ground truth sparse set of predictors is a task best achieved provided two hypotheses are satisfied: low correlations between the true predictors and the spurious ones, and low partial correlations among the set of true predictors (Zhao and Yu, 2006). We observe that these conditions can be met in datasets composed of multiple initial states with various combinations of absent and present species, allowing the reactions to be witnessed in an independent manner (e.g. *knockout* experiments). However, in the context of inferring reactions from experimental time series data, such type of variations of the initial conditions is rarely available. Therefore, we have decided to restrict ourselves to the most general case dealing with traces only obtained in a *wild type* setting. This entails a correlated framework.

We present Reactmine, a sequential algorithm which infers biochemical reactions one at a time with their kinetics, and enforces sparsity by construction. The chapter is organized as follows. In the Methods section, we introduce our inference workflow, its core algorithm with the statistical formulae used, and a comparison to related work from the literature. In the Results section, an evaluation on synthetic

data obtained from several toy CRNs is first carried out, as well as a parameter sensitivity analysis. Then, Reactmine is applied on two real-world examples: first on single cell data of the cell cycle and the circadian clock obtained through videomicroscopy, to infer regulation reactions, and, secondly, on the circadian data presented in Chapter 4, for the task of inferring the influence of systemic factors on peripheral clock gene expression.

5.2 Materials and methods

5.2.1 Settings and notations

Bold lower (resp. upper) case letters denote vectors (resp. matrices). Unless stated otherwise, sets are represented with capital letters. For a matrix $\mathbf{M}$, $\mathbf{M}_{l,\bullet}$ (resp. $\mathbf{M}_{\bullet,i}$) stands for its l^{th} row (resp. i^{th} column). We observe a system $\mathbf{y}$ describing the evolution of m biological species at n discrete time points $\{t_l\}_{1\leqslant l\leqslant n}$. Focusing on the one-trace case for clarity, a data matrix can be defined:

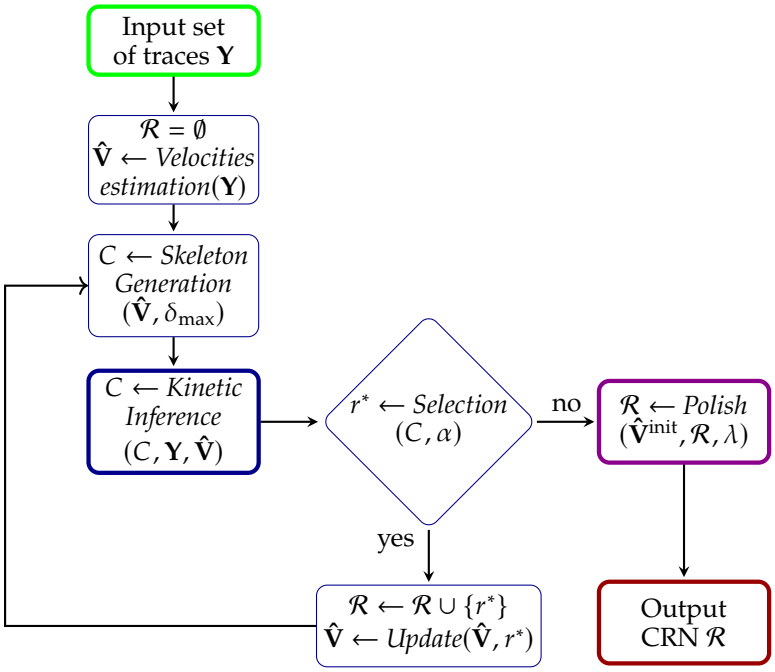
$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,m} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,m} \end{bmatrix}$$

as well as a matrix of observed derivatives, called velocities: $\mathbf{V} = (v_{l,i})_{\substack{1\leqslant l\leqslant n \\ 1\leqslant i\leqslant m}} \in \mathbb{R}^{n\times m}$. The extension to multiple traces is straightforward through matrix concatenations.

We restrict ourselves to biochemical networks displaying a $0/1$ stoichiometry. A reaction is formally defined here as a triple (R, P, f), where R (resp. P) is the set of reactant (resp. product) indices. $f : \mathbb{R}_+^n \to \mathbb{R}_+$ is a rate function over molecular concentrations. We begin by assuming mass action law kinetics for all reactions. Reaction rates are thus the product of the reactant species concentrations by some rate constant k. We also consider the possibility for the reaction to be catalyzed by a species c that is neither consumed nor produced in the reaction, but needs to be present for it to occur. In this case, $R \leftarrow R \cup \{c\}$ and $P \leftarrow P \cup \{c\}$. (R, P, f) can be written $R \overset{f}{\Longrightarrow} P$ or $R \overset{k}{\Longrightarrow} P$ for a mass action law reaction with parameter k. A chemical reaction network (CRN) is defined as a finite set of reactions. Finally, a reaction can also be associated with a stoichiometry vector $\mathbf{s} \in \{-1, 0, 1\}^m$ where $\forall i \in [\![1, m]\!]$, $s_i = 1$ if $i \in P \setminus R \cap P$, -1 if $i \in R \setminus R \cap P$, 0 otherwise.

5.2.2 Algorithm workflow

In this section, the main steps of Reactmine are described (Fig. 5.1).



Reactmine workflow in a nutshell

- Data $\mathbf{Y}$ reporting the evolution of biological species is used as input.
- The inferred CRN is initialized to $\mathcal{R} = \emptyset$. Velocities $\hat{\mathbf{V}}$ are estimated from $\mathbf{Y}$ with finite differences.
- Species whose ratio of velocities is in the order of magnitude given by parameter $\delta_{\max}$ form a reaction skeleton (R, P). Skeletons are collected for each time point t_l and stored in a set C.
- Each reaction skeleton $(R, P) \in C$ is equipped with a rate function $\hat{f}$ from which a loss criterion is derived.
- The reaction $r^* = (R^*, P^*, f^*)$ minimizing the loss criterion is selected.
- If the loss associated to r^* is below a threshold parameter α it is appended to $\mathcal{R}$. Its effect on $\hat{\mathbf{V}}$ is removed, followed by a new round.
- Otherwise, the inferred CRN goes through a global optimization with sparsity-inducing penalization managed by parameter λ.
- The final CRN $\mathcal{R}$ is returned.

Figure 5.1

Generation of reaction skeletons

Most of the time, direct access to measurements for $\mathbf{V}$ is impossible, hence an estimator $\hat{\mathbf{V}}$ is obtained using finite differences. In the event of real data, one could use a smoothing procedure like the one described in (Aubin-Frankowski and Vert, 2020).

Once the velocities $\hat{\mathbf{V}}$ have been estimated, reaction skeleton candidates (R, P), that is to say reactions devoid of their kinetics, are generated. Let us consider $\hat{\mathbf{V}}_{l,\bullet}$, the velocity of the system for an arbitrary time point t_l, and let

$$i^{\max} = \underset{i}{\operatorname{argmax}} \, |\hat{v}_{l,i}| \tag{5.1}$$

Species of index $i^{\max}$ has the highest velocity at time t_l, and then participates in a preponderant reaction that is further explored. To determine the other components of the reaction, the following set is computed:

$$I_\delta(t_l) = \left\{ i \in \{1, \dots, m\} \setminus i^{\max}, \; |\hat{v}_{l,i^{\max}}| \leqslant \delta |\hat{v}_{l,i}| \right\}, \text{ with } \delta \geqslant 1 \tag{5.2}$$

Species with indices belonging to $I_\delta(t_l)$ have similar absolute velocities as $y_{i^{\max}}$, the extent of similarity being defined by δ. Those species will be part of the reaction as well. Depending on the sign of the variation of the corresponding species, elements of $I_\delta(t_l)$ belongs either to the reactant set R of the reaction, or to the set of products P. Likewise, if $I_\delta(t_l)$ is an empty set, according to the sign of $\hat{v}_{l,i^{\max}}$, a synthesis reaction $\emptyset \implies x_{i^{\max}}$ or degradation reaction $y_{i^{\max}} \implies \emptyset$ will be obtained. This computation is done for all time points $\{t_l\}_{1 \leqslant l \leqslant n}$ and for a sequence of increasing δ values belonging to $[1, \delta_{\max}]$, such as $\{1, 1.1, \dots, \delta_{\max} - 0.1, \delta_{\max}\}$. $\delta_{\max}$ is a parameter of the algorithm, representing the maximum absolute fold change allowed between the variations of species involved in a reaction, typically equal to 3. This value significantly above 1 accounts for the fact that $\hat{v}_{l,i^{\max}}$ might not be completely explained by only one reaction. The set of all candidate reaction skeletons $C = \{(R_q, P_q)\}_q$ (or equivalently the stoichiometry matrix of the candidates $\mathbf{S}$ as there is no catalyst yet at this step) is therefore obtained from the union of all skeletons generated, for all $\delta \in \{1, \dots, \delta_{\max}\}$ and for all t_l, $l \in \{1, \dots, n\}$, removing duplicates.

For a stoichiometry vector $\mathbf{s}$ generated, a support set $\text{supp}(\mathbf{s}) = \{(t_l, \delta_l), \; l \in L\}$ is computed. This set contains the L time points where $\mathbf{s}$ originated, and the value of the threshold δ relative to each time point. If for a fixed time point t_l, multiple δ values led to the same stoichiometry vector, i.e. for $\delta \neq \delta', I_\delta(t_l) = I_{\delta'}(t_l)$, then $\delta_l = \min(\delta, \delta')$.

Inference of reaction kinetics constants

The next step consists in assigning a mass action law rate function to the candidate reaction skeletons, to completely define the reactions. (R, P, f) follows the law of mass action with parameter k if $\forall j \in R \cup P$, $\forall l \in \{1, \dots, n\}$

$$v_{l,j} = s_j f(\mathbf{Y}_{l,\bullet}) = s_j k \prod_{u \in R} y_{l,u} \tag{5.3}$$

where we recall that s_j is the stoichiometry of species y_j in the reaction. Let

$$\mathcal{E}(R) = \left\{ l, \prod_{u \in R} y_{l,u} > 0 \right\}$$

Using the finite differences estimate $\hat{\mathbf{V}}_{\bullet,j}$, one can provide an estimator of k $\forall j \in R \cup P$:

$$\hat{k}_j = \frac{s_j}{\#\mathcal{E}(R)} \sum_{l \in \mathcal{E}(R)} \frac{\hat{v}_{l,j}}{\prod_{u \in R} y_{l,u}} \tag{5.4}$$

This estimator is designed to realize an equality in mean between observed kinetics and inferred kinetics. The former is represented by the numerator, the latter by the denominator times $\hat{k}_j$.

Selection of the best reaction candidate

In order to compare reaction candidates between them, an interesting criterion to look at is certainly the statistical quality of the inferred kinetics. The variance of the mass action law coefficient estimate can be estimated itself for each species involved in the reaction:

$$\sigma_j = \frac{1}{\#\mathcal{E}(R)} \left(\sum_{l \in \mathcal{E}(R)} \frac{\hat{v}_{l,j}}{\prod_{u \in R} y_{l,u}} - \hat{k}_j \right)^2 \tag{5.5}$$

It is worth noticing however that there is a relationship between mean and variance when estimating the kinetics of different reactions: a slow reaction will tend to produce a low variance, compared to a faster reaction. We thus consider the coefficient of variation (CV),

$$\rho_j = \frac{\sigma_j}{|\hat{k}_j|} \tag{5.6}$$

measured for each reactant or product of the reaction, and introduce more precisely the species index that minimizes it:

$$j^* = \underset{j \in R \cup P}{\operatorname{argmin}}\, \rho_j \tag{5.7}$$

on which we rely to estimate k. The complete reaction is therefore $r = (R, P, \hat{f})$ with $\hat{f} : \mathbf{y} \mapsto \hat{k} \prod_{u \in R} y_u$ and $\hat{k} = \hat{k}_{j^*}$. This process is done over all reaction skeleton candidates. Then, identifying the best reaction boils down to the minimization of a composite loss term including the CV along with the penalty term defined in Equation (5.9). For a stoichiometry vector $\mathbf{s} \in \mathbf{S}$:

$$\mathcal{L}(\mathbf{s}) = \rho(\mathbf{s}) - \gamma(\mathbf{s}) \tag{5.8}$$

where the species selected to compute $\rho(\mathbf{s})$ is the one that minimizes it (Equation (5.7)). The penalty coefficient $\gamma(\mathbf{s})$ can be computed for each reaction skeleton as follows:

$$\gamma(\mathbf{s}) := \frac{\#\text{supp}(\mathbf{s})}{n} \left(\frac{1}{\#\text{supp}(\mathbf{s})} \sum_{l \in L} \delta_l \right)^{-1} \tag{5.9}$$

The first term encourages skeletons inferred over a large number of time points, as a mark of a frequently occurring relationship between involved species. For the second term: the smaller the δ thresholds required for species to assemble in a skeleton for particular time points, the larger its value. This means that reaction skeletons created from similarly varying species are favored.

The acquisition of the best reaction $r^* = (R^*, P^*, f^*)$, from the pool of candidate stoichiometry vectors $\mathbf{S}$ is outlined in Algorithm 1.

Algorithm 1: Reaction candidate selection

Result: Selected stoichiometric vector $\mathbf{s}^*$, kinetics k^*, loss $\mathcal{L}(\mathbf{s}^*)$

Initialize tabulars for $\hat{k}, \sigma, \rho$ and $\mathcal{L}$

for $\mathbf{s} \in \mathbf{S}$ **do**

  $R = \{i, s_i = -1\}, P = \{i, s_i = 1\}$

  for $j \in R \cup P$ **do**

    Compute $\hat{k}_j(\mathbf{s})$ and $\sigma_j(\mathbf{s})$ from Equation (5.4) and (5.5)

  end

  $\hat{k}(\mathbf{s}), \sigma(\mathbf{s}) = \underset{j \in R \cup P}{\text{argmin}} \dfrac{\sigma_j(\mathbf{s})}{|\hat{k}_j(\mathbf{s})|}$

  $\rho(\mathbf{s}) = \dfrac{\sigma(\mathbf{s})}{|\hat{k}(\mathbf{s})|}$

  $\mathcal{L}(\mathbf{s}) = \rho(\mathbf{s}) - \gamma(\mathbf{s})$

end

$\mathbf{s}^* = \underset{\mathbf{s} \in \mathbf{S}}{\text{argmin}}\, \mathcal{L}(\mathbf{s})$

$k^* = \hat{k}(\mathbf{s}^*)$

return $\mathbf{s}^*, k^*, \mathcal{L}(\mathbf{s}^*)$

The best reaction being identified, it is now accepted and added to the CRN $\mathcal{R}$ if it satisfies $\mathcal{L}(\mathbf{s}^*) < \alpha$. The threshold α is one parameter of the algorithm. A

typically acceptable value for the first term of Equation (5.8) is below 1, indicating that the variance of the estimator does not overcome the mean. The second term, defined in Equation (5.9), is related to the number of time points where a reaction skeleton was witnessed, and with which threshold values δ. For a range of values $[\delta_{\min}, \delta_{\max}]$, this term lives in $[\frac{1}{\delta_{\max}}, \frac{1}{\delta_{\min}}]$. Using the range $[\delta_{\min} = 1, \delta_{\max} = 3]$, $\gamma(\mathbf{s})$ being subtracted, a reasonable value of the threshold α would lie close to 0 and up to 1.

In the event where the best reaction r^* fails to satisfy this condition, Algorithm 1 is performed again, this time with catalyzed reactions. To that end, Equation (5.3) is modified:

$$v_{l,j} = s_j k \prod_{u \in R \cup \{c\}} y_{l,u} \tag{5.10}$$

$\forall j \in R \cup P$, and c can be any species. The optimal catalyst c^* for a particular reaction is the species providing the lowest CV, in which case $R \leftarrow c^*$ and $P \leftarrow c^*$. Lastly, the best catalyzed reaction is selected if its associated loss value is below α.

Velocities update

Once one reaction (R^*, P^*, f^*) is accepted, it is added to the CRN $\mathcal{R}$, and its effect on the velocity data is removed as follows:

$$\hat{\mathbf{V}} \leftarrow \hat{\mathbf{V}} - \begin{pmatrix} f^*(\mathbf{Y}_{1,\bullet}) \\ \vdots \\ f^*(\mathbf{Y}_{n,\bullet}) \end{pmatrix} \mathbf{s}^{*T} \tag{5.11}$$

Starting from these new velocities, another set of candidate reaction skeletons can be computed, initiating the inference of the next reaction. Reactions are selected and added to the CRN $\mathcal{R}$ until the loss value associated to the best inferred reaction no longer satisfies the threshold α. Upon successful inference of the unknown reactions, the entries of $\hat{\mathbf{V}}$ will eventually be close to zero. Together with the fact that skeletons are generated to explain the highest variations in the velocities, this iterative process prioritizes eliminating large variations in the system.

Postprocessing polishing step

Let $\hat{\mathbf{V}}^{\text{init}}$ be the initial matrix of the estimated velocities, i.e. before the removal of the effect of the inferred reactions. Once a CRN has been built from the iterative inference of p reactions, an additional *polish* step can be applied. The objective is to remove fallacious reactions that have been inferred. This is done by means of sparse regression. From a CRN $\mathcal{R} = \{(R_q, P_q, f_q)\}_{1 \leqslant q \leqslant p}$ and data matrix $\mathbf{Y}$, one can

construct a matrix $\mathbf{F}(\mathbf{Y}, \mathbf{k}) \in \mathbb{R}^{n \times p}$:

$$\mathbf{F}(\mathbf{Y}, \mathbf{k}) := \begin{bmatrix} | & & | & & | \\ f_1(\mathbf{Y}_{l,\bullet}) & \dots & f_q(\mathbf{Y}_{l,\bullet}) & \dots & f_p(\mathbf{Y}_{l,\bullet}) \\ | & & | & & | \end{bmatrix} \tag{5.12}$$

with n the number of time points. The q^{th} column of $\mathbf{F}(\mathbf{Y}, \mathbf{k})$ is a vector describing the rate of the reaction (R_q, P_q, f_q) at each time point, with $\mathbf{k}$ being the vector of reaction kinetic parameters. Combined with the stoichiometry matrix of the CRN $\mathbf{S}^* \in \mathbb{R}^{p \times m}$, we can formulate an optimization problem:

$$\mathbf{k} = \underset{\mathbf{k} \in \mathbb{R}_+^p}{\operatorname{argmin}} \|\hat{\mathbf{V}}^{\text{init}} - \mathbf{F}(\mathbf{Y}, \mathbf{k})\mathbf{S}^*\|_2^2 + \lambda \|\mathbf{k}\|_1 \tag{5.13}$$

In particular, if c_q is the catalyst species of reaction (R_q, P_q, f_q)

$$\mathbf{F}(\mathbf{Y}, \mathbf{k}) = \begin{bmatrix} | & & | & & | \\ \prod\limits_{i \in R_1 \cup \{c_1\}} y_{l,i} & \cdots & \prod\limits_{i \in R_q \cup \{c_q\}} y_{l,i} & \cdots & \prod\limits_{i \in R_p \cup \{c_p\}} y_{l,i} \\ | & & | & & | \end{bmatrix} \operatorname{diag}(\mathbf{k}) \tag{5.14}$$

The optimization starts with an initial guess set as $\mathbf{k} = (k_1, \dots, k_p)^T$. It is worth noticing that the least squares term compares the inferred and observed velocities, rather than the data measurements $\mathbf{Y}$ and a numerical integration of the inferred CRN, which allows avoiding the resolution of a non-convex optimization problem. Indeed, in our case, Equation (5.14) shows that the inferred velocities are written as a weighted linear combination of reaction effects, which makes the minimization problem convex. Furthermore, if $p \leqslant n$ with p the number of reactions and n the number of observations, there is a unique solution (Tibshirani, 2012). Next, the ℓ_1 penalty term favors parsimony in the vector of kinetics: coordinates of $\mathbf{k}$ that are shrink to 0 in the optimization represent reactions that are removed from the final CRN. The regularization parameter λ controls the sparsity level. While the iterative selection of reactions treats the problem from a local point of view, inferring reactions one by one independently of one another, this final polishing step allows us to jointly adjust the parameters of all reactions, removing those for which a value of 0 is found.

A plot showing how the velocities are updated after each reaction inference is presented in Fig. 5.2, taking as an example a linear chain of reactions (Fig. 5.3). Additionally, at each iteration, the 5 best ranked candidates reactions are displayed, and the time points leading to their inference are shown with specific colors, thus providing explanations for the selection of the reaction.

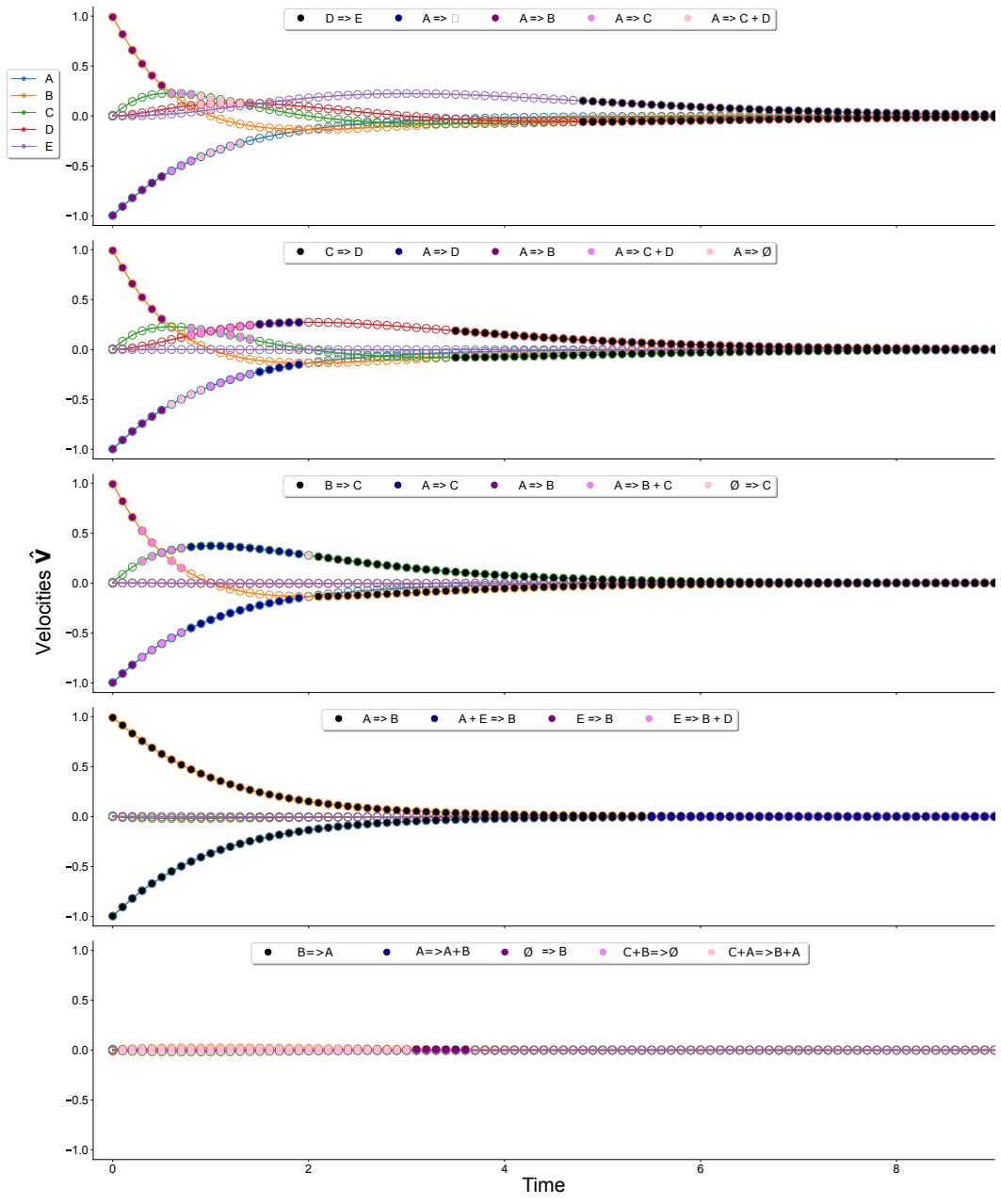


Figure 5.2: Plot of the support of inferred reactions for the chain CRN $A \overset{1}{\Longrightarrow} B \overset{1}{\Longrightarrow} C \overset{1}{\Longrightarrow} D \overset{1}{\Longrightarrow} E$. Each row represents the velocities $\hat{\mathbf{V}}$ after the effect of a reaction is removed, except for the first row. At each step, the 5 top ranked reactions (from left to right in the legend) in terms of loss value are printed on top. For each time point, the color filling the dot associated to a species velocity reveals what reaction was considered. For this example, Reactmine parameters were set to $\alpha = 0.25$, $\lambda = 0.7$, $\delta_{\max} = 3$.

Extension for Michaelis-Menten kinetics inference

Although the workflow of the algorithm was detailed in a mass action law framework, Reactmine is also compatible with other forms of kinetics, provided that the measure of quality of reaction parameters defined in Equation (5.6) can be computed. In this section, we demonstrate how Michaelis-Menten kinetics can also be considered for the inferred reactions. A single-reactant reaction (R, P, f) follows Michaelis Menten kinetics if $\forall j \in R \cup P,\ \forall l \in \{1, \dots, n\}$

$$v_{l,j} = s_j f(\mathbf{Y}_{l,\bullet}) = s_j v_{\max} \frac{y_{l,u}}{y_{l,u} + K_m} \tag{5.15}$$

where $v_{\max}$ and K_m are parameters, and $R = \{u\}$. As $y_u \to +\infty$, $|v_j| \to v_{\max}$. Besides, $|v_j|$ being an increasing function of y_u, an estimator of $v_{\max}$ can be obtained $\forall j \in R \cup P$

$$\tilde{v}_{\max,j} = \max_{l \in \{1,\dots,n\}} |\hat{v}_{l,j}| \tag{5.16}$$

Then, K_m is defined as the value of reactant concentration for which the associated velocity is equal to $\frac{v_{\max}}{2}$. Since measurements are only available at discrete time points, one has

$$\hat{K}_{m,j} = y_{l^*,u} \quad \text{with } l^* = \underset{l \in \{1,\dots,n\}}{\operatorname{argmin}} \left| |\hat{v}_{l,j}| - \frac{\tilde{v}_{\max,j}}{2} \right| \tag{5.17}$$

Once an estimator for K_m has been provided, we apply the same principle as in Equation (5.4) to obtain a new estimator of $v_{\max}$, using the whole dataset, $\forall j \in R \cup P$

$$\hat{v}_{\max,j} = \frac{s_j}{\#\mathcal{E}(R)} \sum_{l \in \mathcal{E}(R)} \hat{v}_{l,j} \frac{\hat{K}_{m,j} + y_{l,u}}{y_{l,u}} \tag{5.18}$$

The coefficient of variation ρ_j is then computed based on $\hat{v}_{\max,j}$, and the velocities update step in Equation (5.11) remains unchanged. Finally, the polishing step only re-optimizes the estimated $\hat{\boldsymbol{v}}_{\max}$, as optimizing for $\hat{\boldsymbol{K}_m}$, in Equation (5.13) is a non convex problem.

Reactmine computational performances

We now detail the time complexity of Reactmine.

Proposition. The computational time complexity to infer one reaction (R, P, f) is $O(nmI)$ where n is the number of time points, m the number of species, and $I = |R \cup P|$.

Proof. The inference of reaction kinetics constants involves the computation of a mean for each species present in the reaction (Equation (5.4)), which is $O(nI)$. In the worst-case, a lookup for a catalyst species is necessary, at a cost of $O(nIm)$.

The velocities update step is $O(nI)$ (Equation (5.11)). Generating reaction skeletons requires the computation of the species displaying highest velocities for each time point, which is $O(nm)$ (Equation (5.1)). After that, the sets $I_\delta(t_l)$ are obtained with δ ranging in a grid of several values, but in practice, this step is vectorized. The time complexity for the inference of one reaction is therefore $O(nmI)$. □

This leads to very fast inference results, with less than a few minutes for datasets with up to 20000 time points. Computations were executed on a laptop with i5 2.9 gHz dual core processor. Lastly, it is worth noting that the generation of skeletons and the association of a kinetic parameter are two easily parallelizable steps. The fact that Reactmine displays a linear time complexity with respect to the number of samples advocates for a more time-consuming reaction inference protocol, for instance with a selection of the K best reactions at each step. Such an advance is compatible with the current structure of Reactmine.

5.2.3 Comparison with SINDy

The SINDy method

Some methods originally designed to discover the dynamics of physical systems can be applied to our problem of inferring a CRN from time series data. In particular, the SINDy system (Brunton et al., 2016), starting from temporal measurements, aims at providing a reconstruction of the velocities in the following way:

$$\hat{\mathbf{V}} = \boldsymbol{\Theta}(\mathbf{Y})\boldsymbol{\Xi} \tag{5.19}$$

Like Equation (5.12), $\boldsymbol{\Theta}(\mathbf{Y}) \in \mathbb{R}^{n\times p}$ is a library of functions constructed from the input variables $\mathbf{Y}$ including, for instance, first to m-order polynomial interactions, the sin and cos functions, or even more sophisticated user-defined functions. The dynamics of each variable is then captured by a weighted combination of library members, the weights being encompassed in $\boldsymbol{\Xi}$. Because it is thought that the expression of the dynamics should be sparse within the library $\boldsymbol{\Theta}(\mathbf{Y})$, SINDy proposes to obtain $\boldsymbol{\Xi}$ using sparse regression.

$$\boldsymbol{\Xi} = \underset{\boldsymbol{\Xi}\in\mathbb{R}^{p\times m}}{\operatorname{argmin}} ||\hat{\mathbf{V}} - \boldsymbol{\Theta}(\mathbf{Y})\boldsymbol{\Xi}||_2^2 + \lambda||\boldsymbol{\Xi}||_1 \tag{5.20}$$

which is similar to Reactmine's polishing step defined in Equation (5.13), although one should notice that for Reactmine, the parsimony is achieved at the reaction level. On the opposite, SINDy enforces sparsity over the components of the right-hand side of the ODEs of each species, which may only be parts of a reaction, leading to a larger number of parameters to estimate.

A recent update of this approach now makes it possible to enforce sparsity using non convex regularizers such as the ℓ_0 penalty. We refer to the article from Cham-

pion et al. (2020) for further details. In our experiments, we use the `pySINDy` package with SR3 optimizer and ℓ_0 regularizer. For fair comparison in an elementary reactions setup with mass action law and stoichiometry at most 1, $\mathbf{\Theta}(\mathbf{Y})$ is here restricted to interactions up to the second order, and a bias term. Hence $p = \left(1 + \frac{m(m+1)}{2}\right)$.

$$\mathbf{\Theta}(\mathbf{Y}) := \begin{bmatrix} | & | & & | & | & & | \\ 1 & \mathbf{Y}_{\bullet,1} & \dots & \mathbf{Y}_{\bullet,m} & \mathbf{Y}_{\bullet,1}\mathbf{Y}_{\bullet,2} & \dots & \mathbf{Y}_{\bullet,m-1}\mathbf{Y}_{\bullet,m} \\ | & | & & | & | & & | \end{bmatrix} \tag{5.21}$$

Associated to a positive weight, the bias term corresponds to a synthesis reaction. First order interactions translate to reactions such as $A \overset{k}{\Longrightarrow} B + C$ for $A \neq \emptyset$, with the special case $A \in \{B, C\}$ corresponding to a catalyzed synthesis. Second order interactions encompass reactions of the form $A + B \overset{k}{\Longrightarrow} C + D$ for $\{A, B\} \neq \emptyset$. Again, the case $A \veebar B \in \{C, D\}$ corresponds to a catalyzed reaction, with $\veebar$ referring to the exclusive OR.

The GRISLI method

Somewhat related to SINDy is the algorithm GRISLI originally designed for inferring Gene Regulatory Networks from single-cell data (Aubin-Frankowski and Vert, 2020). GRISLI solves a similar problem as SINDy, but estimates velocities $\hat{\mathbf{V}}$ thanks to a weighted average of finite differences, with weights defined by a spatio-temporal kernel $K\left(\mathbf{Y}_{l,\bullet}, \mathbf{Y}_{l',\bullet}\right)$, which quantifies how $\mathbf{Y}_{l',\bullet}$ is believed to be useful to estimate the velocity at $\mathbf{Y}_{l,\bullet}$. Subsequently, Equation (5.20) is solved multiple times, each time with a bootstrapped sample of $\mathbf{Y}$ and $\hat{\mathbf{V}}$. This leads to frequencies of apparition of each term. GRISLI considers a library $\mathbf{\Theta}(\mathbf{Y})$ made with first-order interactions only, which makes it a particular case of SINDy for the optimization part. This last point means that GRISLI cannot infer complexation reactions. Furthermore, GRISLI outputs frequencies that are difficult to compare with the output of Reactmine and SINDy. For these reasons, we decided to compare with SINDy only.

Comparing SINDy and Reactmine

Rather than fitting the velocities - except during the polishing step - Reactmine's intention is to iteratively infer chemical reactions that summarize the highest variations of the system. Consequently, a comparison with SINDy based on the goodness of fit of each inferred model seems inadequate here. In addition, as SINDy is based on a global optimization, for a large enough library there exist a parametrization such that the data are perfectly reconstructed, from the principle of overfitting (Fig. 2.5). Instead, to allow for comparison between inference results of both Reactmine and SINDy, one should notice that there is a mapping between the CRN

inferred by Reactmine and the weight matrix Ξ associated to the library $\Theta(\mathbf{Y})$ defined in Equation (5.19), as detailed in Section 5.2.3. Then, for any synthetic CRN, the ground truth matrix Ξ^* is known and one can compute classification metrics between Ξ and Ξ^*. In this case, there is 3 possibilities for the weights of the matrix Ξ: negative, zero or positive values. The precise value of the weight is not assessed here, only its sign, which carries the mechanistic information. A common accuracy metric in classification is the F_1-score, defined as the harmonic mean between precision and recall. For a category h, these quantities are defined as:

$$\begin{aligned}
\text{Precision}_h &= \frac{\#\text{True positives}}{\#\text{True positives} + \#\text{False Positives}} \\
\text{Recall}_h &= \frac{\#\text{True positives}}{\#\text{True positives} + \#\text{False Negatives}} \\
F_1\text{-score}_h &= 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

We define the macro F_1-score as the mean of the F_1-scores computed for each category (Santos et al., 2011).

Finally, as SINDy cannot account easily for Michaelis Menten kinetics due to the non convex nature of the cost function with respect to parameters such as K_m, Reactmine is restricted to mass action law kinetics and the evaluation to mass action law CRNs. Numerical integration of the CRNs is performed using the Python package `scipy.integrate.odeint` with the default integrator `lsoda`.

5.3 Results

5.3.1 Evaluation on synthetic toy CRNs

We begin with an evaluation of Reactmine on synthetic examples for which a ground truth CRN is accessible. Panel **(A)** of Fig. 5.3 shows the different examples studied for evaluation purposes. Each CRN is designed to test a specific pattern. For instance, the reactant-parallel CRN is a network where two species, C and E, produce the same output species D, an instance of interest with respect to the sequential structure of Reactmine. In what follows, Reactmine $\delta_{\max}$ parameter is set to 3 and $\alpha = 0.25$. These values are discussed in the below sensitivity analysis (Fig. 5.5). For comparison, SINDy inference results are also reported. A classifier predicting -1, 0 and 1s uniformly at random is also included for comparison.

Inference from a single trace

A first scenario is analyzed: evaluation of Reactmine using a single so-called canonical trace $\mathbf{Y} \in \mathbb{R}^{n \times m}$ which is an observation of all species at all time points obtained

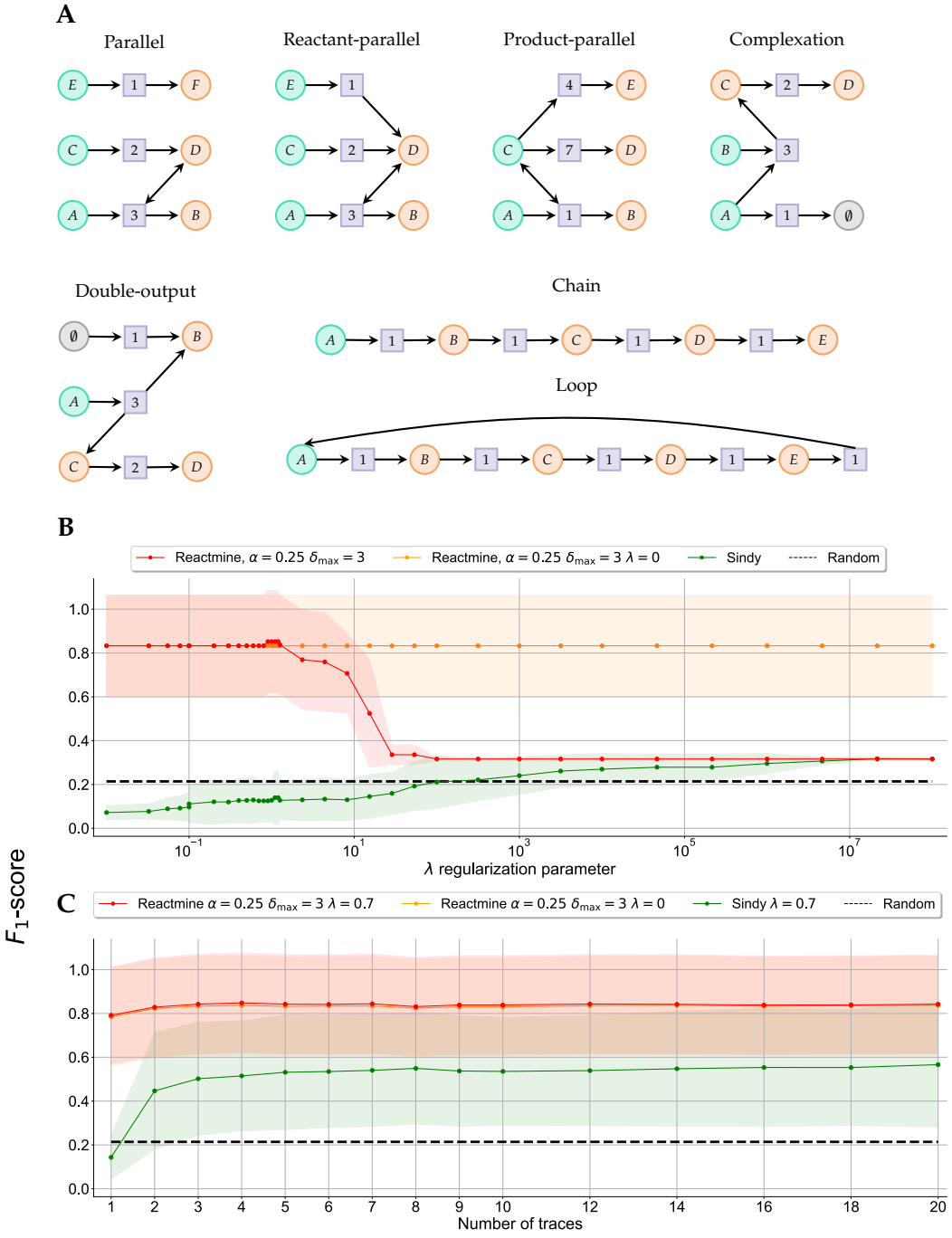


Figure 5.3: **(A)** Toy CRNs used for the evaluation on synthetic data. Squared blue-filled nodes are reaction nodes, with mass action law parameter value written inside. Green nodes stand for nonzero species at initial state. Double-headed arrows represent catalysts. **(B)** Inference metrics of Reactmine and SINDy averaged over all toy CRNs. A single canonical trace is considered and performances are given as a function of λ the regularization parameter. **(C)** Performances as a function of the number of traces included. For both panels, standard deviations are computed across CRNs.

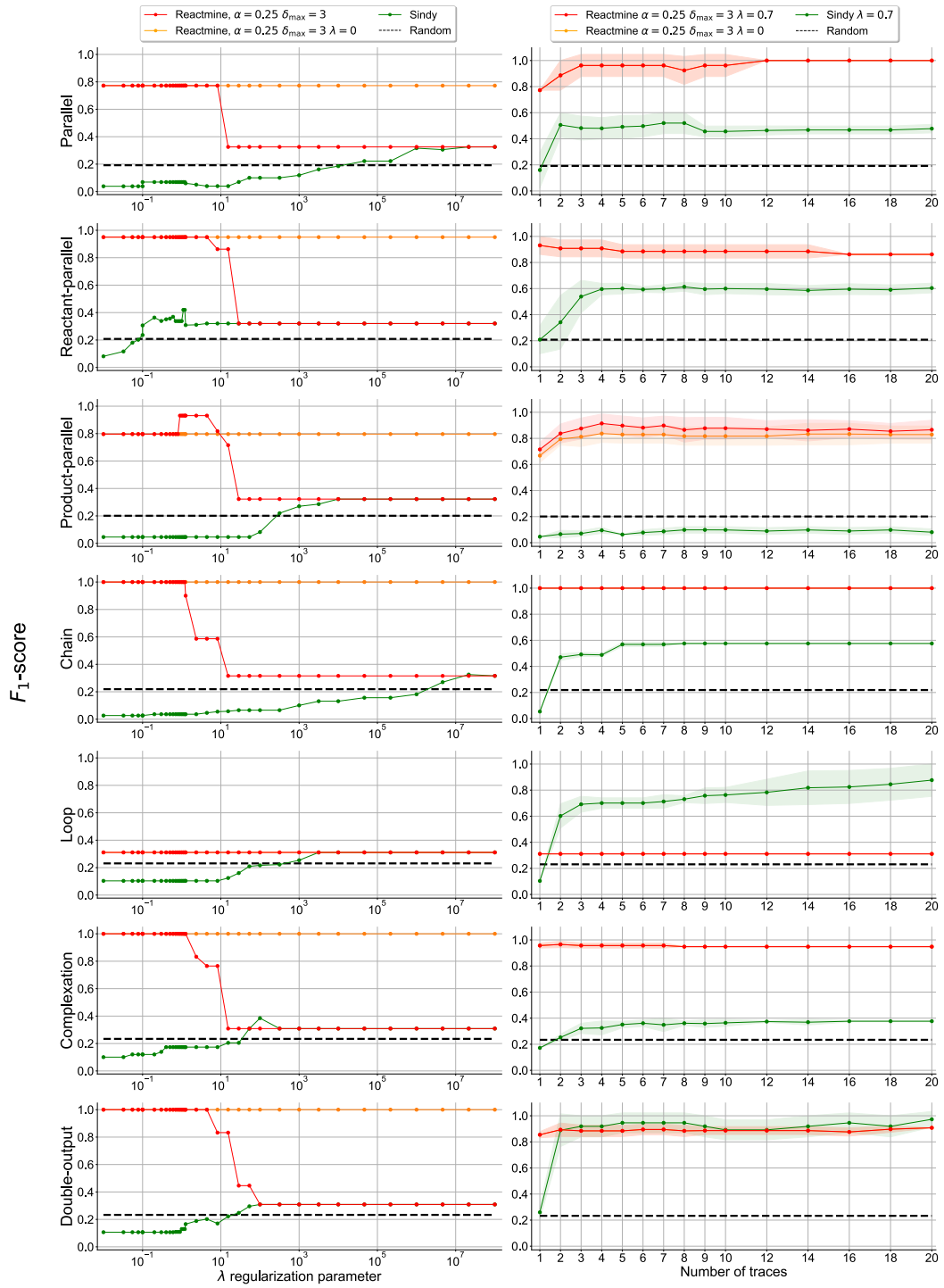


Figure 5.4: Inference results of Reactmine and Sindy for each toy CRN. The first column considers a single canonical trace and gives the performances as a function of λ the regularization parameter of both methods. The second considers multiple traces. Standard deviations originate from the fact that the evaluation was performed on different traces 6 different times.

by numerical integration of the CRNs (Fig. 5.S1). Simulations run for a time horizon $T = 10$ in arbitrary time unit, with time step $\Delta t = 0.1$. We consider an initial state $\mathbf{y}(0) \in \mathbb{R}^m$ ensuring that all reactions of the network will be witnessed during the simulation. For instance, in the chain CRN containing five species, $\mathbf{y}(0) = (1, 0, 0, 0, 0)$ is such an initial state. $(0, 0, 1, 1, 0)$ is not: reactions $A \overset{1}{\Longrightarrow} B$ and $B \overset{1}{\Longrightarrow} C$ will never occur. For each CRN depicted in Panel **(A)** of Fig. 5.3, non-zero species at the initial state are marked in green. Performances of Reactmine, Sindy and the random classifier for this scenario are reported in the Panel **(B)** of Fig. 5.3, with the accuracy metrics being averaged across all CRNs shown as a function of λ the regularization parameter. In this regime, Reactmine succeeds in reconstructing the unknown CRNs for a large, reasonable range of λ values, whereas SINDy performs barely better than a random classifier. It is worth noticing that in the limit $\lambda \to +\infty$, the F_1-score converges to approximately $\frac{1}{3}$. This is due to the fact that with sufficiently enough regularization, the inferred models are empty, which means that all the 0's of the matrix $\mathbf{\Xi}^*$ are correctly predicted, thus one of the three labels is well-predicted.

The first column of Fig. 5.4 goes into the details of the performances for each network. Reactmine succeeds in inferring the chain, complexation and double-output CRNs. The reactant-parallel CRNs is almost correctly recovered, with the exception of reaction $A + D \overset{3.2}{\Longrightarrow} B$ inferred in place of $A + D \overset{3}{\Longrightarrow} B + D$. For the parallel CRN, the F_1-score drops below 0.8, as the reaction $A \Longrightarrow B$ is inferred in place of $A + D \Longrightarrow B + D$. On both examples, the incapacity of Reactmine to correctly infer the catalyst reaction $A + D \Longrightarrow B + D$ stems from the fact that the value for the threshold α is too high, thus allowing the acceptance of non catalyzed reactions. A lower value for α leads to a perfect reconstruction. Interestingly, for the product-parallel CRN, reactions $C \overset{13.2}{\Longrightarrow} D$ and $C \overset{4.8}{\Longrightarrow} E$ were inferred. While the structure is accurate, the kinetics are poorly estimated compared to the ground truth reactions $C \overset{7}{\Longrightarrow} D$ and $C \overset{4}{\Longrightarrow} E$. This large bias amounts to updates of the velocity matrix $\hat{\mathbf{V}}$ that are too powerful, creating variations instead of removing only what corresponds to the current reaction. Reactmine then tries to explain these new velocities with reactions that do not belong to the ground truth CRN. However, as shown by the red curve, there exists a range of λ values for which the polish step shrinks to 0 the rate of these unwanted reactions. Finally, the results are mediocre for the loop CRN, with an empty CRN being output. This can be explained by the structure of our method: its sequential nature makes it sensitive to the order of inference of reactions. For the loop CRN, there does not exist any species that is only an input or output of the network, making the choice of a first reaction difficult since the mass action law estimators cannot be computed reliably reaction per reaction.

Inference from multiple traces with same initially present species

The second scenario (Panel **(C)** of Fig. 5.3, right column of Fig. 5.4) is based on multiple traces. Non-zero species at initial state are the same as for the canonical trace. For the present species, the initial concentrations are sampled uniformly in $[1, 5]$. For example, $\mathbf{y}(0) = (2, 0, 0, 0, 0)$ and $\mathbf{y}(0) = (4.2, 0, 0, 0, 0)$ are two possible initial states for the chain. Reactmine recovers, in this case, similar performances as in the previous setup involving only a single trace, on average. On the opposite, using more data points seems to improve the results of SINDy. In particular, SINDy almost matches the results of Reactmine concerning the double-output CRN. Furthermore, it also succeeds in perfectly reconstructing the loop CRN, while the performances of Reactmine on this CRN are unchanged with an empty CRN being inferred. On the other hand, in this multiple traces regime, the product parallel and parallel CRNs can be perfectly reconstructed by Reactmine for a number of possible data configurations, as shown by the standard deviations.

5.3.2 Reactmine parameter sensitivity

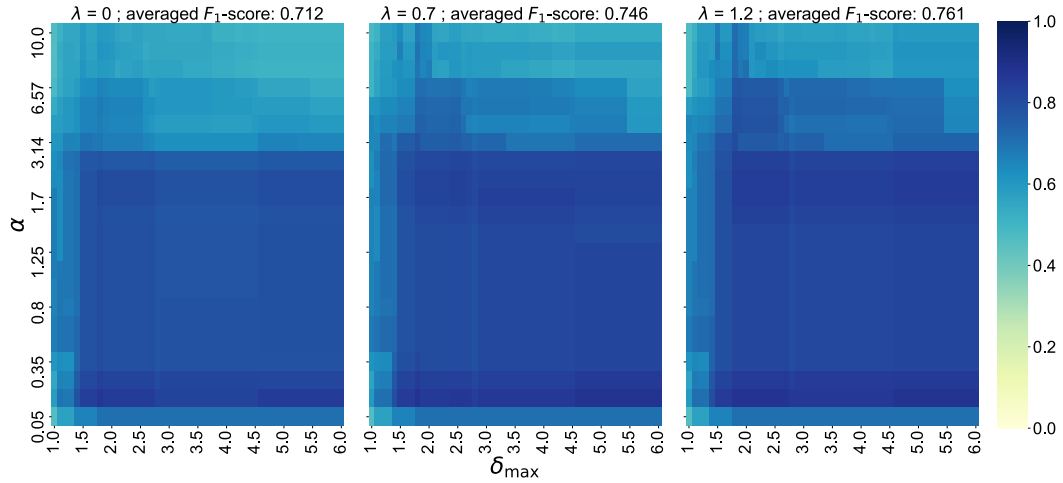


Figure 5.5: Sensitivity of Reactmine to α, $\delta_{\max}$ and λ parameter change. The measured output is the F_1-score, averaged across all toy CRNs. On top of each subplot, the F_1-score averaged across the grid of $(\alpha, \delta_{\max})$ values is reported.

In this section, we study the impact of parameter variations on the inference results of Reactmine. The previous evaluation considered fixed values of the reaction acceptance threshold $\alpha = 0.25$ and species variation pairing threshold $\delta_{\max} = 3$, as well as a fixed regularization parameter $\lambda = 0.7$ in the second scenario where the number of traces varies (Panel **(C)** of Fig. 5.3). Using as input a single canonical trace, the F_1-score is computed and averaged across CRNs for varying values of α, $\delta_{\max}$ and λ (Fig. 5.5). α appears to be the most sensitive parameter, with a range

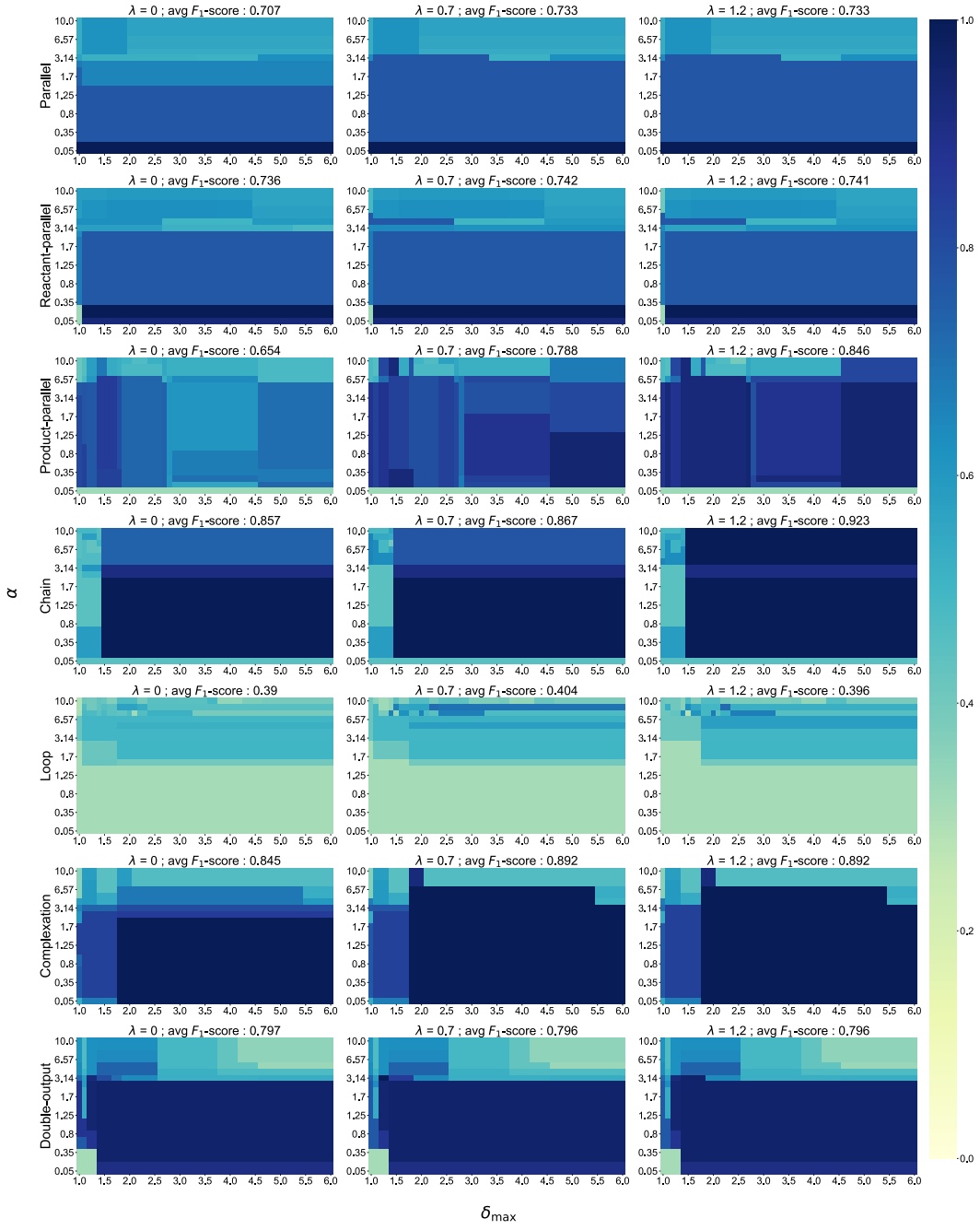
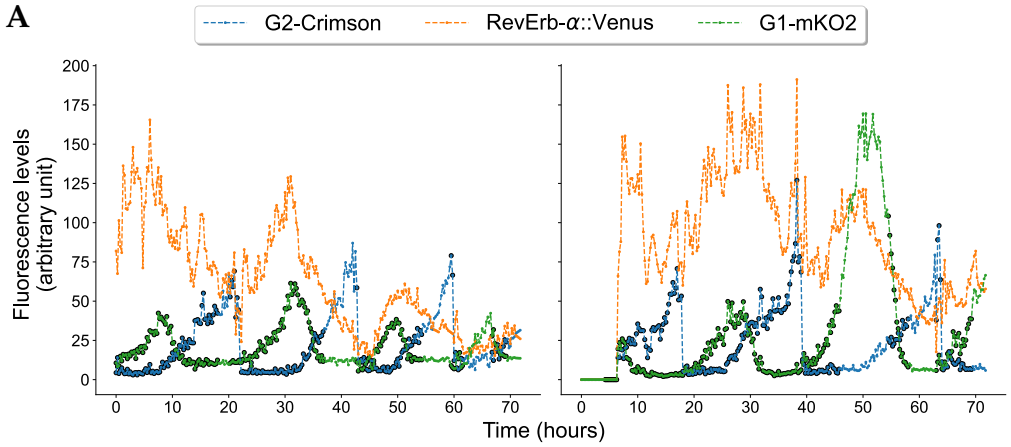


Figure 5.6: Sensitivity of Reactmine to α, $\delta_{\max}$ and λ parameter change, for each CRN. The measured output is the F_1-score. On top of each subplot, the F_1-score averaged across the grid of $(\alpha, \delta_{\max})$ values is reported.

of acceptable values roughly between 0.1 and 2.5. The dependence of the F_1-score with respect to $\delta_{\max}$ seems smaller. As expected, a sufficiently large value for $\delta_{\max}$ is needed in order to generate reaction skeletons other than synthesis and degradation. Fig. 5.6 displays the results for each CRN and shows a clear difference in

parameter sensitivity between the parallel CRN and non parallel CRNs, such as the chain for instance. This finding suggests a fine-tuning of the value of $\delta_{\max}$ *a priori* according to some expected degree of connectivity of the underlying network. Some sensitivity of $\delta_{\max}$ can also be observed when α is fixed to high values. This can be understood: large α values ends up in too many reactions being accepted. To moderate this issue, $\delta_{\max}$ should be set to a lower value. Concerning λ, computing an average of the F_1-score across all values taken by α and $\delta_{\max}$ on the grid gives a value of 0.712 for $\lambda = 0$, of 0.746 for $\lambda = 0.7$ and of 0.761 for $\lambda = 1.2$. This increasing trend is confirmed for all CRNs but the reactant-parallel and loop ones. Sensibility to λ is particularly well demonstrated in the case of the product-parallel CRN, as already observed in Section 5.3.1. This indicates that the post processing polishing step is able to shrink the kinetics of fallacious reactions to zero.

5.3.3 Evaluation on real videomicroscopy data



$$\begin{cases} \frac{dG2}{dt} = -0.645 + 0.02G1 \\ \frac{dRev}{dt} = 314.77\frac{G2}{12.72 + G2} - 288.81\frac{Rev}{8.59 + Rev} \\ \frac{dG1}{dt} = -440.84\frac{G1}{8.4 + G1} \end{cases} \qquad \begin{cases} \frac{dG2}{dt} = -0.645 + 0.02G1 \\ \frac{dRev}{dt} = -0.881 \\ \frac{dG1}{dt} = 0.722 \end{cases}$$

Figure 5.7: Inference results on the videomicroscopy data. **(A)** Plots of fluorescence levels reported in two cells among the 91 of the dataset. Black-circled data points corresponds to points where the reaction $G1 \implies G2$ was inferred by Reactmine. **(B)** ODEs representation of the CRN inferred by Reactmine. Parameters of the algorithm were set to $\delta_{\max} = 3$, $\alpha = 6.5$, $\lambda = 0.7$. **(C)** System of ODEs learned by SINDy with $\lambda = 0.04$.

In this section we investigate time series data obtained by biological experiments

on proliferating NIH3T3 embryonic fibroblast cells. These data have been used to develop a coupled model of the cell cycle and the circadian clock in this cell line in (Traynard et al., 2016). The reported experiments have been done using time lapse videomicroscopy and cell tracking using different fluorescent reporters for the cell cycle and the circadian clock observed during 72 hours (Feillet et al., 2015). This cell line was modified to include three fluorescent markers of the circadian clock and the cell cycle: the RevErb-α::Venus clock gene reporter (Nagoshi et al., 2004) for measuring the expression of the circadian protein RevErbα, and the Fluorescence Ubiquitination Cell Cycle Indicators (FUCCI), Cdt1 and Geminin, two cell cycle proteins which accumulate during the G1 and S/G2/M phases respectively, for measuring the cell cycle phases (Sakaue-Sawano et al., 2008). The cells were left to proliferate in regular medium supplemented with different concentrations of FBS (20% in this data set). Long-term recording was performed in constant conditions with one image taken every 15 minutes during 72 hours. A dataset constituted of 91 cells could be assembled from these data. Panel **(A)** of Fig. 5.7 shows plots of the fluorescence levels obtained in two cells as an illustration of the high cell variability and noise level displayed by the data. For this reason, a smoothing of the curves with a sliding window of 2.5 hours was applied.

Reactmine was run with a threshold of reaction acceptance α equal to 6.5 to account for high noise levels. Three reactions could be discovered by Reactmine, all using Michaelis Menten kinetics. The first reaction inferred was $G1 \implies G2$, a production of the variable accounting for phases S/G2/M through G1 disapperance. Secondly, the reaction $G2 \implies Rev$, an action of the phases S/G2/M of the cell cycle on the circadian clock, encompassed by the variable Rev, was predicted. This result is in agreement with the findings of the modeling study Traynard et al. (2016). Finally, the reverse reaction $Rev \implies G2$ was inferred. Besides these three reactions, two reactions were also inferred but had their kinetics set to zero by the polishing step: $G2 \implies Rev + G1$ and $Rev \implies \emptyset$. In Fig. 5.7, $G1$ and $G2$, fluorescence levels relative to 2 out of the 91 cells have been colored with black dots for the time points that led to the inference of the reaction $G1 \implies G2$. This enables us to observe that the inferred reaction explains low fluorescence levels, and does not completely account for the stiff transitions generated by large variations. Fig. 5.7 Panel **(B)** displays the ODE representation of the inferred dynamical models by Reactmine and SINDy. The latter is run using only a mass action law based library.

The ODE system found by SINDy is less biologically meaningful, as it involves degradation terms of constant rates for species $G2$ and RevErb-α. An activation of $G2$ by $G1$ is uncovered, though with a quite low rate compared to the other terms. Finally, a synthesis of $G1$ is inferred. Increasing λ above the value of 0.04 gives ODEs made of constant terms only.

5.3.4 Detection of circadian systemic controls on liver clock gene expression

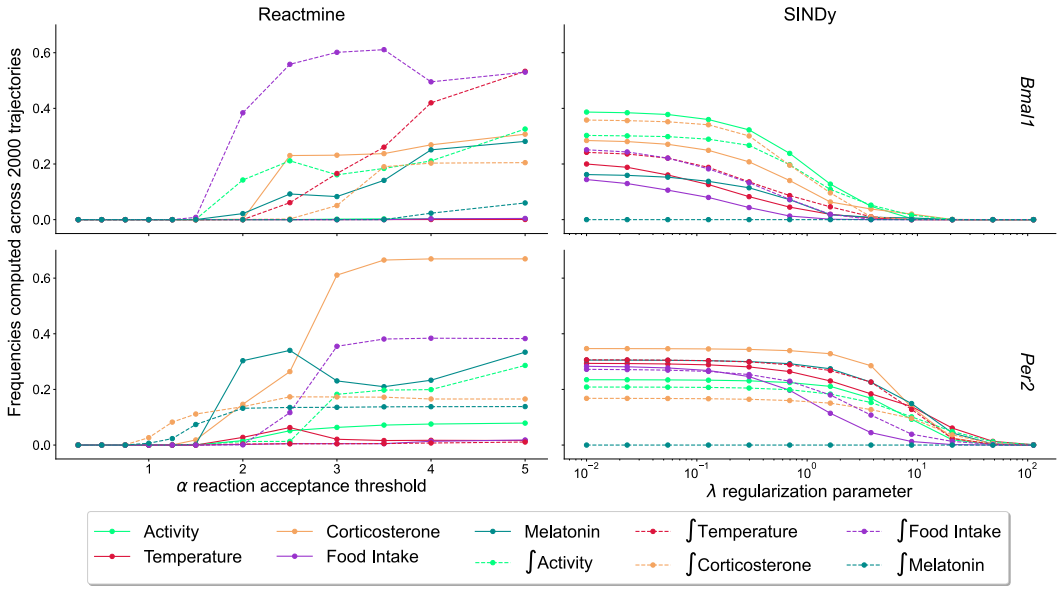


Figure 5.8: Inference of systemic drivers of clock gene transcription by Reactmine and SINDy. For Reactmine, λ was set at 0.7 and $\delta_{\max}$ to 3.

In this section, we retrieve parts of the data used in (Martinelli et al., 2021) and studied in Chapter 4. These data collected in mice are comprised of measurements of circadian rhythms of five systemic regulators: body temperature, rest-activity rhythms, food intake, plasma corticosterone and melatonin, as well as circadian expression data of 2 core-clock genes in the liver: *Bmal1* and *Per2*. Here, the objective is to determine the circadian control of systemic regulators on liver clock gene transcription. Indirect actions of these regulators through intermediate species were also computed under the form of *integral regulators*, such as $\int$Temperature, as previously detailed in Equation (4.1).

Using a model of the liver circadian clock developed in (Hesse et al., 2021) and presented in Chapter 3, an approximation y of the action of these systemic regulators on clock gene expression can be derived, thanks to Equation (4.7), as detailed in Section 4.3. To explain this term, catalyzed reactions $z \implies z + y$ are searched for, where z can be any regulator. This amounts to run Reactmine, bypassing the skeleton generation step and seeking multiple catalyzed synthesis of y including the five regulators and five integral regulators as possible candidates for these reactions. As in Chapter 4, a process of data augmentation was carried out to obtain $N = 2000$ versions of the residuals $\{y^{(j)}\}_{1 \leqslant j \leqslant N}$, under different circadian clock conditions. This permits a CRN to be inferred for each $y^{(j)}$, and frequencies of apparition of regulators to be computed for both Reactmine and SINDy.

Fig. 5.8 recapitulates the results, as a function of λ for SINDy, and as a function of α for Reactmine, with $\delta_{\max} = 3$ and $\lambda = 0.7$. Interestingly, Reactmine is able to discriminate between regulators for sufficiently large α. For regulation of *Bmal1*, indirect action of Temperature and Food Intake on gene transcription are the two most inferred drivers across the 2000 trajectories. Besides, for $\alpha = 5$, the ranking of frequencies for the 6 first regulators closely matches the order found in (Martinelli et al., 2021), where similar linear models were used. For the regulators of *Per2* transcription, Reactmine favors an action of Corticosterone. Except for the high frequency given to $\int$Food Intake, these rankings do not agree with the findings from (Martinelli et al., 2021). This can be justified by the sequential nature of the algorithm, which picks the best reaction and then update its effect before inferring another reaction, whereas regression-based methods jointly estimate the effect of regulators. Therefore, differences between the reactions inferred by the two approaches are likely to appear depending on whether the best models for an increasing number of variables are nested or not. In this particular case, it turns out that for the regulation of *Bmal1* transcription, the best 2-term model found in Chapter 4 was $\int$Temperature + $\int$Food Intake (Fig. 4.7), while the best 1-term model featured $\int$Temperature. On the other hand, regarding *Per2*, Corticosterone was found to be the best 1-term model, which is not nested with the best 2-term model, Food Intake + $\int$Temperature. Hence why the differences with Reactmine.

About SINDy, except for the indirect action of melatonin, defining a cutoff between the regulators is difficult, whatever value λ takes, meaning that no regulator really stands out. Though, it is worth noting that, for *Per2* and for $\lambda < 0.5$, regulators food intake and temperature display the highest frequencies, both through direct and indirect action. Aside from the high frequency of corticosterone, those results are similar to the one obtained in Chapter 4.

5.4 Discussion

We have presented an algorithm for the inference of biochemical reactions from time series data. Reactmine is based on a sequential process, which we view as a different way to handle parsimony compared to current approaches based on sparse regression. By design, this technique does not aim at a perfect reconstruction of the data. The latter requires a global point of view during the optimization process which is difficult to conjugate with sparsity if interventional data is not usable to reduce correlations among the system. Instead, the purpose of Reactmine is rather to learn a set of reactions that summarizes a biological system, even in a correlated framework. On synthetic data from toy networks, this led to promising results in a low trace regime, in contrast to SINDy, suffering from the fact that the variables are highly correlated. The loop CRN is a good instance of a hard network to tackle for our method. The lack of input or output species makes the association of kinetics

unreliable, inducing reactions with great variance. Although, upon testing of high values for α, some of reactions in the network could be recovered.

Reactmine depends on three parameters, $\delta_{\max}$ for the maximum absolute fold change allowed between the variations of species involved in a reaction, and α for the reaction acceptance threshold. λ controls the level of ℓ_1-regularization in a final optional polishing step. A precise tuning of these parameters can give improved results on specific examples, e.g. the product-parallel and loop CRNs. Nevertheless, a sensitivity analysis highlighted the presence of a fairly large region in which results are consistent to a good level. Reasonable bounds were provided for α. In the case of λ, the computation of a Pareto front between model complexity and least squares could be leveraged.

The application to videomicroscopy data required the adjustment of the reaction acceptance threshold α at a high value to account for considerable noise levels as well as the recourse to Michaelis Menten Kinetics. It resulted in the inference of biologically meaningful reactions. At each iteration, an accepted reaction represents the best candidate across all time points, with respect to the acceptation criterion. It is worth noticing that despite a large cell heterogeneity, three meaningful reactions could be selected. Of note, with a reaction comes the possibility to display the time points over which it was inferred, conferring an explainable nature of the algorithm. In an ultimate application, inference of systemic drivers of clock gene transcription was carried out using Reactmine, recovering the results obtained in (Martinelli et al., 2021) for *Bmal1*, but hardly for *Per2*. This fact could be explained by the sequential nature of Reactmine.

The approach can be straightforwardly enriched with the integration of prior knowledge, under the form of already known reactions, in a preprocessing step of the data. For instance, the knowledge of the half life of a protein can be formulated into a degradation reaction. Using the velocities update step, one just has to remove the effect of this reaction and start from there.

Further improvements will take advantage of the linear time complexity in the number of observations for the inference of a reaction. One way to proceed is to infer K reactions at each step instead of 1. This would lead to a research tree, where each path defines a CRN. A CRN could be given a score for instance defined as the worse coefficient of variation among its reactions. Metaheuristics such as Limited Discrepancy Search could then be applied to mitigate the computational burden (Korf, 1996). Another direction consists in finding strategies for updating previously inferred reactions when a new reaction is inferred, either in terms of kinetics or structure.

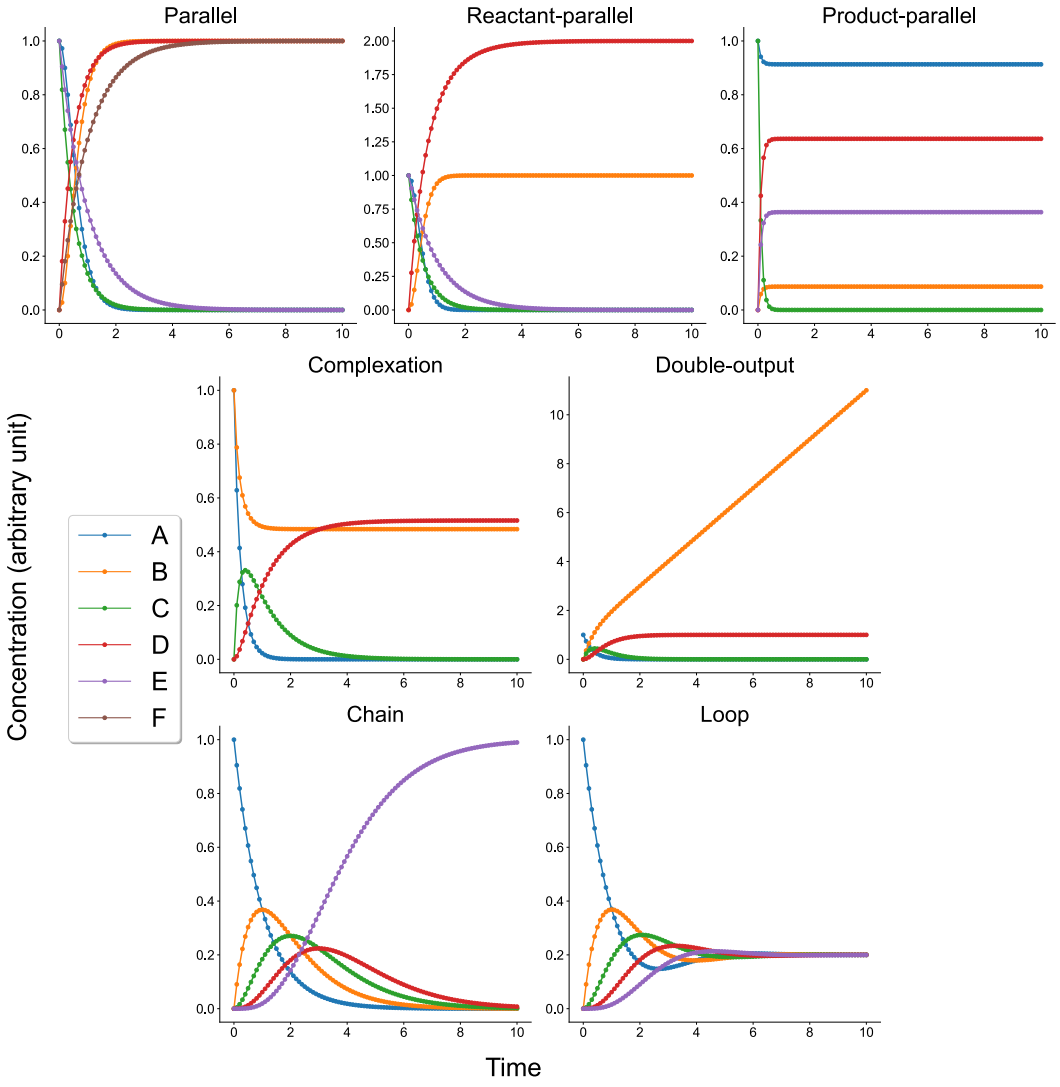


Figure 5.S1: Numerical simulation of the toy CRNs.

Chapter 6

Conclusion

This PhD was aimed towards the inference of mechanistic models from time series data, with a significant focus on the personalization of chronotherapies. Therefore, the work done in this project lies at the intersection between methodological advances and biological findings for individualizing medicine. The results include the conception of two new methods for model learning and the development of key theoretical tools leading to new insights into physiological mechanisms governing circadian medicine. We now provide a summary of the contributions and discuss their future directions.

6.1 Learning mechanistic models from temporal data: two innovative methodologies

Automating the design of mechanistic models is a challenge motivated by the considerable number of available data thanks to recent technological development and the need for more personalized modeling to assist medicine individualization.

In Chapter 4, we presented a new methodology for model learning applied to the investigation of possible whole-body dynamical regulators controlling peripheral circadian clocks of the liver. Data at hand was comprised of time-dependent measurements of systemic regulators and core-clock gene expression in four mouse classes. The objective was to uncover the action of the former on the latter, in a classwise manner to further inform chronotherapy personalization. Therefore, it required to learn both the structure and the parameters of a model.

We created and calibrated a model of the cellular circadian clock in Chapter 3 which served as a starting point to discover new biological findings. Often, when looking to represent a physiological phenomenon, one has already access to some form of prior knowledge such as gene regulation maps, degradation rates or subcellular distribution of species. Being able to integrate such information in order to help identifying unknown interactions from temporal data constitutes a key asset that cannot be readily leveraged in purely statistical approaches. Instead, we started by assuming that known processes could be recapitulated by a mechanistic model, composed of ODEs, whose parameters had been estimated from experimental

data (Chapter 3). Undiscovered regulatory links are then considered as missing subparts of the model, which were hypothesized to be located at specific areas. More precisely, as accessible data consisted in circadian rhythms of clock gene mRNA amounts, we assumed that the systemic influences occurred at either the gene mRNA transcription or degradation level. This enabled the computation of an approximation of this action termed *residual trajectories* y for each mouse class. Next, our workflow featured a process of data augmentation, spanning multiple modified versions of the residuals $(y_1, \ldots, y_N)$ based on biologically relevant perturbations of the postulated mechanistic model.

The generated trajectories were fitted by means of linear models, that is, weighted combinations of systemic regulators and of their indirect counterpart, *integral regulators*. Such variables account for the fact that a regulator may not act directly on gene expression, but perhaps through intermediate species. This provides a way to include hidden variables in the models. Linear combinations including both direct and indirect concomitant action of the same factor were discarded. Model fitting of all trajectories led to parameter distributions specific to each model and mouse class. Such distributions enabled a fine-grained comparison of the impact whole-body regulators have on different mouse classes. Furthermore, an investigation of classwise differences could be performed by means of ANOVA, treating the obtained parameter vectors as different realisations of a random variable. It is worth mentioning that without resorting to data augmentation, pointwise estimates would have led to a normally distributed vector of parameters. Yet, the distributions obtained by introducing a biological noise at the clock level often exhibited heavy tails or bimodality, far from a gaussian behavior.

One should notice that, after isolating model subparts to learn, the next steps of the methodology, model calibration and model selection, are actually quite flexible. We settled for the exhaustive fitting of linear models, as a straightforward choice to evaluate regulator relative importance among mouse classes. Yet, any other machine learning algorithm whose output can be explained, e.g. with SHAP (Lundberg and Lee, 2017), could have been employed. These include random forests or feed-forward neural networks. As for model selection, a motivational example of how it could be adapted is to consider a larger number of systemic regulators, adding for instance blood pressure, heart rate, light intensity, etc. Increasing the number of regulators from 5 to 10, thus from 10 to 20 variables when including their integral counterpart, the number of admissible models grows from 242 to 59048. Performing an exhaustive search then becomes prohibitive. In this case, sparsity-enforcing approaches like Lasso are better suited. However, Lasso does not impose structures to be equal for each mouse class, i.e. for the same regularization level λ, fitting the trajectories could lead to different whole-body regulators being selected across classes. To ensure this does not happen, multi-task learning could be leveraged (Lozano and Swirszcz, 2012).

A more general context of growing interest would be to infer chemical reaction networks from automated experiments, e.g. assessing the mechanism of action of a new compound (Vertes et al., 2018). In this situation, the requirement of integrating prior knowledge is lifted and learning algorithms may only rely on temporal data at hand. Chapter 5 introduces Reactmine, a novel model learning methodology that is applicable in such scenario.

In biological systems, species typically interact with only a few others (Ouma et al., 2018), and the dynamics of these interactions may be well-described by a few rate functions such as mass action law (Keener and Sneyd, 2009). These observations translate into a sparse representation of each species dynamics. This embedding is difficult to recover in a *wild type* setting, where all reactions are witnessed at the same time and hence highly correlated.

In this precise *wild type* framework, Chapter 5 presented Reactmine, a novel inference algorithm. Whereas other approaches aimed at inferring a sparse system of ODEs, Reactmine views the problem in terms of CRNs, and enforces sparsity through a sequential inference of the reactions. This choice also ensures positivity of the inferred system. Our algorithm depends on 3 parameters: $\delta_{\max}, \alpha$ and λ. $\delta_{\max}$ corresponds to the maximum absolute fold change allowed between the variations of species involved in a reaction. α, to the reaction acceptance threshold conditionally to its inferred kinetics, and λ to the amount of ℓ_1 regularization in the final CRN polishing step. Applying Reactmine to synthetic data generated through toy networks gave promising results. The same exercise on real data required to increase the α threshold to account for noise. Reactions inferred were biologically meaningful and in agreement with other findings, thus demonstrating the relevance of Reactmine as a tool to aid the human modeler.

Our algorithm can be applied to gain intuition on the preponderant reactions at stake in a collection of time series measurements. The evaluation (Section 5.3.1) revealed that the sequential nature of Reactmine makes it more suitable for biological structures that are expected to display input/output behavior. Cyclic topologies like the Loop CRN constitute one limitation at the moment as their kinetics cannot be estimated reliably.

An immediate perspective would be to take advantage of the low time complexity of this algorithm to extend the sequential inference from 1 to K reactions at each step. This would give raise to a research tree, in which each path leads to a different CRN being inferred. Metaheuristics like Limited Discrepancy Search could then be applied for efficient exploration of this tree (Korf, 1996). The selection of the optimal CRN could for instance be done with respect to the minimization of the worse coefficient of variation for the reactions contained in it.

Finally, an unrealistic assumption implicitly made by Reactmine is that all the variables of the reactions to learn are observed. In the long range, a desirable

objective would be to account for hidden entities, as a mechanism that is likely to be at stake in problems which are not properly fitted with documented species only. Introducing hidden variables in an inference pipeline is a strategy that is not dissimilar to how genetic algorithms work. For instance, the algorithm proposed by Degrand et al. (2019) starts from a population of mass action law ODE systems with fixed initial conditions. Subsequently, it makes them evolve with standard operations, including species removal, removal of a monomial in a species's ODE, addition of a new variable, etc. Each of these candidate systems has its parameters and initial conditions fitted to some evaluations of a function one wishes to approximate. Then, a selection phase keeps a fraction of the best individuals, and the process is iterated. Borrowing ideas to these algorithms might be fruitful to improve Reactmine.

On a more high-level perspective, let us jointly examine the proposed methods. The original protocol detailed in Chapter 4 isolates the action of unidentified species on a particular target. It then admits any inference algorithm to fit the detached model subpart, e.g. with Random Forests or Reactmine (Section 5.3.4). The choice depends on the level of interpretability on the mechanisms one wishes to achieve. This points out the overlap of the method with Reactmine. However, a conceptual difference lies in the fact that the available information is used to derive an approximation, which is then fitted through a joint optimization, for preciseness. In the other hand, Reactmine performs a sequential inference of the reactions, in a matter of addressing sparsity, but may result in biased reaction parameters since they are obtained in a separate manner. This makes the selection order an important determinant of how things will play out, which results in Reactmine not being able to enforce similar structure for different mouse classes. Reactmine is therefore better suited in a setting where nothing is known and will output a set of reactions that explains the most preponderant variations present in the data in a mechanistic manner.

6.2 Personalizing chronotherapies

With wearable technologies, the eased acquisition of whole-body biomarkers data in individual patients enabled us to develop a three-step approach. First, the influence of systemic regulators on the cellular circadian clock was investigated. Next, a mechanistic description of the control exercised by the peripheral clock on key pharmacological enzymes was provided. This allowed the prediction of drug chronopharmacology, thus contributing to the personalization of chronotherapies from wearable sensors.

In Chapter 3, we began by providing a novel model of the liver circadian clock. This model was fitted using fully quantitative and time-resolved datasets both for

clock gene mRNAs and proteins. As a result, the simulated species concentrations are expressed in mol/L and can account for the mesor of gene expression, not only for the phase and relative amplitude as existing mathematical models of the clock already do. An explicit modeling of cytoplasmic and nucleic versions of proteins was done, so that the action of transcription factors like CLOCK/BMAL or PER/CRY over clock-controlled genes is faithfully represented. Both these specificities were thought of with an ulterior motive, that our model will be connected with models representing drug chronopharmacology. In particular with the PK-PD network of irinotecan. A simplified version of such a connection was already done, e.g. in (Dulong et al., 2015), but using autonomous cosine-based rhythms as a proxy for the circadian control of protein activity.

Sensitivity analysis of the global model was also carried out, exploring parameter effects on toxicity profiles, namely their phase and relative amplitude. Besides the involvement of clock gene transcription and degradation, it pointed out the non negligible impact of circadian degradation of irinotecan-related enzymes, CES2 and UGT1A1. The latter activates irinotecan into its metabolite, SN38, while the former detoxifies it. This advocates in favor of an experimental validation of these actions.

Among other things, our mechanistic description permits to personalize the optimal time of administration, with respect to the acquisition of circadian measurements of clock and irinotecan-related genes mRNAs. While collecting such circadian expressions might still be costly, and invasive for the patient, a new approach allows to determine the time profile of clock genes based on a single time point (Vlachou et al., 2020). This bypasses the need of multiple around-the-clock biopsies on an individual patient basis thus greatly simplifying the process.

Chapter 3 laid the foundations for the personalization of drug administration timing from gene expression measurements. On the other side, Chapter 4 was devoted to the integration of inter-patient variability at the whole-body level.

To that end, the mechanistic model developed in Chapter 3 was employed to investigate the influences of systemic regulators on the peripheral circadian clock. At the cellular level, the available data concerned gene expression of 3 clock genes, *Bmal1*, *Per2* and *Rev-Erbα*. Therefore, influences were assumed to occur on gene mRNA transcription and degradation, such hypothesis being biologically relevant. For instance *Per2*'s transcription was found to be enhanced by Heat Shock Proteins (Kornmann et al., 2007).

By considering a linear action of whole-body regulators on clock gene transcription, two main drivers were identified for *Bmal1* and *Per2*: temperature and feeding patterns, potentially through intermediate species. The action of Food Intake has also been massively reported in the literature since modulating meal timing and composition impacts liver clock genes time profiles (Greenwell et al., 2019). As an example, dampened oscillations were shown in mice subjected to high-fat-diet,

whereas time restricted high-fat-diet restored regular circadian rhythms (Hang et al., 2012; Li et al., 2010). Importantly, the strength of these inferred regulations were found to be statistically different according to mouse sex and genetic background. This may imply that a systemic regulator activates different regulatory pathways depending on the mouse class as a consequence of different gene expression levels.

On the opposite, there was no evidence pointing towards a systemic control of *Rev-Erbα* transcription, under the restriction to linear models. Likewise, there was no indication of a systemic control of gene mRNA degradation for all 3 genes. Indeed, the obtained approximation profiles in this case were found to be strikingly nonlinear, displaying sharp peaks. It seems unlikely for this behavior to derive from the realisation of natural biological processes. Of note, the absence of data on protein levels in the four mouse classes prevented us from studying possible systemic controls on clock gene translation or protein degradation, which thus cannot be ruled out. Similarly, systemic controls relying on regulators other than those measured in the utilized dataset cannot be excluded.

Upon selection of Temperature and Food Intake as main drivers of *Bmal1* and *Per2*'s transcription, the subsequent stage consists in their integration back into the cellular clock model, in the ODEs. This will involve a step of model refitting to data. If a satisfying data fit is obtained, experimental validation may be performed by modulating the inferred regulators and observing the circadian clock for instance.

Next, the obtained model could be scaled for humans. For that, we take advantage of the well-preserved character of the cellular circadian clock across mouse and human, up to a phase opposition. Multi-scale approaches are then permitted, in which the same structure would be kept, and selected parameters will be changed through species scaling. This allows us to avoid designing and fitting such a mechanistic model solely based on human data which would have been cumbersome due to the scarcity of human circadian datasets. This is one perk of multi-scale modeling, facilitated by mechanistic models.

Another kind of validation of the method can be looked for on the clinical domain. Yet, an issue lies in finding circadian datasets featuring variables both at the cellular and systemic level, for the same individuals. Such material can be found in (Akashi et al., 2010). The study showed that clock gene oscillations were well characterized in hair follicle cells. This permits to account for the peripheral clock at different times in a much easier way than if relying on biopsies. This article further reports rest-activity rhythms, meal hours, saliva melatonin and cortisol measurements for 6 individuals. Altogether, these datasets could be used to validate the inferred regulations, but in a human context, with gene expression coming from hair follicles instead of liver. This disparity seems to be manageable as phase differences of clock genes are well conserved across organs in baboons (Mure et al., 2018). Let us also mention the forthcoming Multidom clinical trial. The goal of that study is to assess the efficacy and toxicities of mFOLFIRINOX, the combination of

5-fluorouracil, irinotecan and oxaliplatin, in 42 patients with pancreatic cancer in real time. This drug cocktail represents the standard treatment for this pathology, but induces severe toxicities, limiting its use to patients in good general condition. In this perspective, several biomarkers like rest-activity and temperature variations will be monitored by a chest sensor every minutes. Body weight will be collected by a connected scale and daily questionnaires filled by the patients through telecommunicating tablets. The democratization of systemic biomarkers data collection opens up new perspectives for personalized chronotherapies.

Supplementaries of Chapter 2

S2-1 Quantitative core-clock model

S2-1.1 Derivation of the quantitative core-clock model from Relógio *et al.*

The model of Relógio *et al.*

We started from the model of Relógio et al. (2011), for which a graphical description is provided Fig. 3.1**b** of the main text. In this model, paralogs and isoforms are merged into global species, e.g. *Cry* = *Cry1* + *Cry2*. In what follows, we will use the notations adopted in Relógio et al. (2011) for the parameters and variables. In Section S2-1.1, we describe the changes in model structure that we have done.

Simplification of the PER/CRY loop

The model of Relógio et al. (2011) contains two parallel PER/CRY loops which account for the fact that PER proteins may exist in phosphorylated or unphosphorylated forms. Considering the lack of quantitative data on PER phosphorylation, we simplified the PER/CRY loop by merging phosphorylated and unphosphorylated species. The new variables refer to the total amount of PER proteins and are defined as:

- $\mathrm{PER}_C^{tot} \leftarrow \mathrm{PER}_C^* + \mathrm{PER}_C$ ($z_2 \leftarrow z_2 + z_3$)
- $\mathrm{PER/CRY}_C^{top} \leftarrow \mathrm{PER^*/CRY}_C + \mathrm{PER/CRY}_C$ ($z_4 \leftarrow z_4 + z_5$)
- $\mathrm{PER/CRY}_N^{top} \leftarrow \mathrm{PER^*/CRY}_N + \mathrm{PER/CRY}_N$ ($x_2 \leftarrow x_2 + x_3$)

Writing the ODEs for these three new variables results in:

$$\begin{aligned}\frac{d(PER_C^* + PER_C)}{dt} &= \frac{d(z_2 + z_3)}{dt} \\ &= k_{p_1} y_1 + k_{d_{z_5}} z_5 + k_{d_{ph_{z_3}}} z_3 - k_{f_{z_5}} z_1 z_2 - k_{ph_{z_2}} z_2 - d_{z_2} z_2 \\ &+ k_{ph_{z_2}} z_2 + k_{d_{z_4}} z_4 - k_{d_{ph_{z_3}}} z_3 - k_{f_{z_4}} z_1 z_3 - d_{z_3} z_3\end{aligned}$$

$$\begin{aligned}\frac{d(PER^*/CRY_C + PER/CRY_C)}{dt} &= \frac{d(z_4 + z_5)}{dt} \\ &= k_{f_{z_5}} z_1 z_2 + k_{f_{z_4}} z_1 z_3 + k_{e_{x_2}} z_2 + k_{e_{x_3}} x_3 \\ &- k_{i_{z_4}} z_4 - k_{i_{z_5}} z_5 - k_{d_{z_4}} z_4 \\ &- k_{d_{z_5}} z_5 - d_{z_4} z_4 - d_{z_5} z_5\end{aligned}$$

$$\begin{aligned}\frac{d(PER^*/CRY_N + PER/CRY_N)}{dt} &= \frac{d(x_2 + x_3)}{dt} \\ &= k_{i_{z_4}} z_4 + k_{i_{z_5}} z_5 - k_{e_{x_2}} x_2 - k_{e_{x_3}} x_3 \\ &- d_{x_2} x_2 - d_{x_3} x_3\end{aligned}$$

The phosphorylation parameters $k_{d_{ph_{z_3}}}$ and $k_{ph_{z_2}}$ naturally disappear. The right hand sides of these three ODEs can be re-written in terms of the new variables $z_2 + z_3$, $z_4 + z_5$ and $x_2 + x_3$ by assuming equal parameters for PER phosphorylated and unphosphorylated forms for a given reaction (e.g. nuclear transport, protein degradation, . . .). According to the parameter table provided in Relógio et al. (2011), this is a reasonable assumption. For each reaction, the new parameter was set equal to the mean of parameters obtained in Relógio et al. (2011) for phosphorylated and unphosphorylated PER proteins, as an initial guess for parameter estimation.

Removal of cytoplasmic complexes degradation parameters

Equations for CLOCK/BMAL and PER/CRY cytoplasmic complexes originally included both a term for complex dissociation into CLOCK, BMAL and PER, CRY free proteins and for complex degradation (i.e./ immediate disappearance). These two redundant terms induced problems of model identifiability as there is no available data on any of those two molecular events taken separately that would allow for reliable parameter estimation. Hence, degradation parameters for the complexes in the cytoplasm were removed. This is equivalent to assuming that the complex needs first to dissociate before the proteins can be degraded.

Refining the CLOCK/BMAL subnetwork

The proven action of CLOCK/BMAL on the expression of genes involved in the irinotecan network implies that the variable CLOCK/BMAL$_N$ is an important exit point from the circadian clock model to the irinotecan network (Yang et al., 2009; Oishi et al., 2005). Therefore, particular attention must be paid to faithfully modelling the dynamics of CLOCK/BMAL nuclear level. Next, while the inital clock model did not explicitly represent *Clock* for the reason that it was found arrhyth-

mic in the SCN (Reppert and Weaver, 2001), the argument does not hold in other tissues such as the liver (Narumi et al., 2016). For these reasons, we decided to extend the original clock model and included the *Clock* gene as a state variable. *Clock* transcription is regulated through RORE elements family (Preitner et al., 2002) so that ROR and REV-ERB were assumed to act on its transcription, respectively in a positive and negative manner. The cytoplasmic protein CLOCK_C dimerizes in the cytoplasm with BMAL_C (Zheng et al., 2019b; Lowrey and Takahashi, 2004). CLOCK/BMAL_C then translocates to the nucleus and becomes the variable CLOCK/BMAL_N. We assumed that BMAL nuclear protein level was negligible as it has been observed that BMAL and CLOCK nuclear protein expressions share the same circadian phase and amplitude, suggesting that both species exist majoritarily in complexed forms (Wang et al., 2017).

Accounting for the cytoplasm/nucleus volume ratio in shuttling dynamics

The mathematical model in Relógio et al. (2011) did not focus on cellular compartmentalization, and thus potentially different volumes for nucleus vs cytoplasm were not considered. This difference can be very large as the fraction of the total cell volume that occupies the nucleus is approximately equal to 10% in mammalian cells (see Bionumbers). Thus, the equation terms representing the transport between the cytoplasm and the nucleus need to be scaled to ensure the conservation of the total species quantity (e.g./ for nuclear import, what leaves the cytoplasm should be equal to what enters the nucleus). We chose to only modify the equations of the cytoplasm compartment in which all nuclear transport terms are multiplied by the cytoplasm/nucleus volume ratio:

$$\frac{dz_6}{dt} = k_{p_3} y_3 - \frac{v_c}{v_n} k_i z_6 - d_{z_6} z_6$$

S2-1.2 Mathematical description of the quantitative core-clock model

List of variables

Table 2.S1 lists the state variables of the core-clock model.

Model equations

CLOCK/BMAL_C

$$\frac{dz_9}{dt} = k_{f_{z_9}} z_8 z_5 + \frac{v_c}{v_n} k_{e_{x_1}} x_1 - \frac{v_c}{v_n} k_{i_{z_9}} z_9 - k_{d_{z_9}} z_9 \tag{2.S1}$$

CLOCK/BMAL_N

Variable name	Species name
x_1	CLOCK/BMAL_N
x_2	PER/CRY_N^{tot}
x_5	REV-ERB_N
x_6	ROR_N
y_1	*Per*
y_2	*Cry*
y_3	*Rev-Erb*
y_4	*Ror*
y_5	*Bmal*
y_6	*Clock*
z_1	CRY_C
z_2	PER_C^{tot}
z_4	PER/CRY_C^{tot}
z_5	CLOCK_C
z_6	REV-ERB_C
z_7	ROR_C
z_8	BMAL_C
z_9	CLOCK/BMAL_C

Table 2.S1: List of state variables of the core-clock model.

$$\frac{dx_1}{dt} = k_{i_{z_9}} z_9 - k_{e_{x_1}} x_1 - d_{x_1} x_1 \tag{2.S2}$$

CLOCK_C

$$\frac{dz_5}{dt} = k_{p_6} y_6 + k_{d_{z_9}} z_9 - k_{f_{z_9}} z_8 z_5 - d_{z_5} z_5 \tag{2.S3}$$

Rev-Erb

$$\frac{dy_3}{dt} = V_{3_{\max}} \frac{1 + g\left(\frac{x_1}{k_{t_3}}\right)^b}{1 + \left(\frac{x_2}{k_{i_3}}\right)^c \left(\frac{x_1}{k_{t_3}}\right)^b + \left(\frac{x_1}{k_{t_3}}\right)^b} - d_{y_3} y_3 \tag{2.S4}$$

Ror

$$\frac{dy_4}{dt} = V_{4_{\max}} \frac{1 + h\left(\frac{x_1}{k_{t_4}}\right)^b}{1 + \left(\frac{x_2}{k_{i_4}}\right)^c \left(\frac{x_1}{k_{t_4}}\right)^b + \left(\frac{x_1}{k_{t_4}}\right)^b} - d_{y_4} y_4 \tag{2.S5}$$

REV-ERB_C

$$\frac{dz_6}{dt} = k_{p_3} y_3 - \frac{v_c}{v_n} k_{i_{z_6}} z_6 - d_{z_6} z_6 \tag{2.S6}$$

ROR_C

$$\frac{dz_7}{dt} = k_{p_4} y_4 - \frac{v_c}{v_n} k_{i_{z_7}} z_7 - d_{z_7} z_7 \tag{2.S7}$$

REV-ERB_N

$$\frac{dx_5}{dt} = k_{i_{z_6}} z_6 - d_{x_5} x_5 \tag{2.S8}$$

ROR_N

$$\frac{dx_6}{dt} = k_{i_{z_7}} z_7 - d_{x_6} x_6 \tag{2.S9}$$

Clock

$$\frac{dy_6}{dt} = V_{6_{\max}} \frac{1 + j \left(\frac{x_6}{k_{t_6}}\right)^b}{1 + \left(\frac{x_5}{k_{i_6}}\right)^c + \left(\frac{x_6}{k_{t_6}}\right)^b} - d_{y_6} y_6 \tag{2.S10}$$

Bmal

$$\frac{dy_5}{dt} = V_{5_{\max}} \frac{1 + i \left(\frac{x_6}{k_{t_5}}\right)^b}{1 + \left(\frac{x_5}{k_{i_5}}\right)^c + \left(\frac{x_6}{k_{t_5}}\right)^b} - d_{y_5} y_5 \tag{2.S11}$$

BMAL_C

$$\frac{dz_8}{dt} = k_{p_5} y_5 + k_{d_{z_9}} z_9 - k_{f_{z_9}} z_8 z_5 - d_{z_8} z_8 \tag{2.S12}$$

Per

$$\frac{dy_1}{dt} = V_{1_{\max}} \frac{1 + a \left(\frac{x_1}{k_{t_1}}\right)^b}{1 + \left(\frac{x_2}{k_{i_1}}\right)^c \left(\frac{x_1}{k_{t_1}}\right)^b + \left(\frac{x_1}{k_{t_1}}\right)^b} - d_{y_1} y_1 \tag{2.S13}$$

Cry

$$\frac{dy_2}{dt} = V_{2_{\max}} \frac{1 + d\left(\frac{x_1}{k_{t_2}}\right)^e}{1 + \left(\frac{x_2}{k_{i_2}}\right)^f \left(\frac{x_1}{k_{t_2}}\right)^e + \left(\frac{x_1}{k_{t_2}}\right)^e} \frac{1}{1 + \left(\frac{x_5}{k_{i_{21}}}\right)^{f_1}} - d_{y_2} y_2 \tag{2.S14}$$

CRY_C

$$\frac{dz_1}{dt} = k_{p_2} y_2 + k_{d_{z_4}} z_4 - k_{f_{z_4}} z_1 z_2 - d_{z_1} z_1 \tag{2.S15}$$

PER_C^{tot}

$$\frac{dz_2}{dt} = k_{p_1} y_1 + k_{d_{z_4}} z_4 - k_{f_{z_4}} z_1 z_2 - d_{z_2} z_2 \tag{2.S16}$$

PER/CRY_C^{tot}

$$\frac{dz_4}{dt} = k_{f_{z_4}} z_1 z_2 + \frac{v_c}{v_n} k_{e_{x_2}} x_2 - \frac{v_c}{v_n} k_{i_{z_4}} z_4 - k_{d_{z_4}} z_4 \tag{2.S17}$$

PER/CRY_N^{tot}

$$\frac{dx_2}{dt} = k_{i_{z_4}} z_4 - k_{e_{x_2}} x_2 - d_{x_2} x_2 \tag{2.S18}$$

List of model parameters of the core-clock model

Parameter	Name	Liver	SW480	SW620
Degradation rates for nuclear proteins or nuclear protein complexes [hour^{-1}]				
d_{x_1}	CLOCK/BMAL	0.2565	0.1017	0.1806
d_{x_2}	PER/CRY_N^{tot}	0.1267	0.7392	0.3232
d_{x_5}	REV-ERB_N	1.7383	2.4484	0.1183
d_{x_6}	ROR_N	1.3175	0.6792	1.6102
Degradation rates for mRNAs [hour^{-1}]				
d_{y_1}	*Per*	0.4977	0.2913	1.3323
d_{y_2}	*Cry*	0.4194	0.134	0.0347
d_{y_3}	*Rev-Erb*	0.6743	2.4956	2.2398
d_{y_4}	*Ror*	0.2352	0.035	0.0312
d_{y_5}	*Bmal*	0.3198	2.5	1.9949
d_{y_6}	*Clock*	0.3235	0.3381	0.2074
Degradation rates for cytoplasmic proteins [hour^{-1}]				
d_{z_1}	CRY_C	0.3125	2.4994	2.9975

d_{z_2}	PER_C	0.0662	0.0409	0.0364
d_{z_5}	$CLOCK_C$	0.6229	1.6832	2.3992
d_{z_6}	$REV\text{-}ERB_C$	0.3378	0.0414	0.032
d_{z_7}	ROR_C	0.7913	0.1356	0.1363
d_{z_8}	$BMAL_C$	0.3014	1.1466	2.6224
	Reaction rates for complex formation [nmol$\times L^{-1}\times$hours^{-1}]			
$k_{f_{z_9}}$	$CLOCK/BMAL_C$	0.0657	0.0169	0.0563
$k_{f_{z_4}}$	PER/CRY_C^{tot}	0.0318	0.0257	0.0341
	Reaction rates for complex dissociation [hours^{-1}]			
$k_{d_{z_9}}$	$CLOCK/BMAL_C$	0.279	0.6536	2.9995
$k_{d_{z_4}}$	PER/CRY_C^{tot}	0.021	0.3111	0.3573
	Transcription rates [nmol$\times L^{-1}\times$ hours^{-1}]			
$V_{1_{max}}$	*Per*	0.1431	2.5165	5.5144
$V_{2_{max}}$	*Cry*	0.0496	0.0523	0.1499
$V_{3_{max}}$	*Rev-Erb*	0.0079	0.0267	0.0057
$V_{4_{max}}$	*Ror*	0.0393	0.0004	0.0018
$V_{5_{max}}$	*Bmal*	0.0027	0.1713	0.0203
$V_{6_{max}}$	*Clock*	0.0027	0.0073	0.0022
	Activation/inhibition rates [nmol$\times L^{-1}$]			
k_{t_1}	*Per*-activation rate	0.066	0.0065	0.0189
k_{i_1}	*Per*-inhibition rate	1.1912	0.5495	0.5672
k_{t_2}	*Cry*-activation rate	241.2015	18.7184	22.6995
k_{i_2}	*Cry*-inhibition rate	0.0223	0.0021	0.0016
$k_{i_{21}}$	*Cry*-inhibition rate	11.5034	15.0881	19.7684
k_{t_3}	*Rev-Erb*-activation rate	0.0716	0.0186	0.0054
k_{i_3}	*Rev-Erb*-inhibition rate	14.9091	15.6085	39.5099
k_{t_4}	*Ror*-activation rate	4.6929	10.2838	45.7059
k_{i_4}	*Ror*-inhibition rate	0.0012	0.0015	0.0023
k_{t_5}	*Bmal*-activation rate	18.4408	10.1313	3.5061
k_{i_5}	*Bmal*-inhibition rate	10.5553	15.1897	98.7832
k_{t_6}	*Clock*-activation rate	0.01593	0.4121	1.1779
k_{i_6}	*Clock*-inhibition rate	46.9265	31.3405	165.0875
	Transcription fold activation (dimensionless)			
a	*Per*	30.0384	1.6157	4.0879
d	*Cry*	6.0566	107.1115	94.8115
g	*Rev-Erb*	198.347	22.2163	32.5303
h	*Ror*	9.7537	195.1387	120.3863
i	*Bmal*	12	12	12
j	*Clock*	3.3507	17.8863	11.5117
	Production rates [molecules $\times$ mRNA$^{-1}\times$ hour^{-1}]			
k_{p_1}	PER_C^{tot}	384.9242	17175.9753	4058.6511

k_{p_2}	CRY_C	469.6348	13313.783	9622.3143
k_{p_3}	REV-ERB_C	2695.93	33397.5015	78769.4987
k_{p_4}	ROR_C	338.6881	18.1801	161.066
k_{p_5}	$BMAL_C$	803.5155	1706.2512	16006.2689
k_{p_6}	$CLOCK_C$	288.4416	2047.4364	2858.2369
	Import/Export rates [hour^{-1}]			
$k_{i_{z_4}}$	PER/CRY_C^{tot}	0.03407	0.0549	0.0506
$k_{i_{z_6}}$	REV-ERB_C	0.4057	0.0135	0.0018
$k_{i_{z_7}}$	ROR_C	0.0011	0.0044	0.0113
$k_{i_{z_9}}$	CLOCK/$BMAL_C$	0.001	0.0011	0.0012
$k_{e_{x_1}}$	CLOCK/$BMAL_N$	0.4025	0.2495	0.9313
$k_{e_{x_2}}$	PER/CRY_N^{tot}	0.00005	0.003	0.0014
	Hill coefficients of transcription (dimensionless)			
b	activation	8	6.1184	7.5482
c	inhibition	4.5568	6.425	5.8353
e	*Cry*-activation	5.1910	7.9585	5.2423
f	*Cry*-inhibition	7.9525	7.9968	6.5434
f_1	*Cry*-inhibition	1	7.9587	3.5099
	Volume proportion (dimensionless)			
v_c	cytoplasm	0.93	0.8	0.8
v_n	nucleus	0.07	0.2	0.2

Table 2.S2: List of parameters and optimal values for each dataset.

Hill coefficients can be related to the number of boxes in the promoter of the target gene and should therefore remain to low values (e.g. 3 for Rev-Erb Korenčič et al. (2012)). However, the transcription term in the equations is only semi-mechanistic and encompasses all molecular events from the activation of transcription regulators (e.g. CLOCK/BMAL1) in the nucleus to the production of mRNA molecules ultimately reaching the cytoplasm. Thus, such term can implicitly model more complex networks such as activation cascades (Gonze and Abou-Jaoudé, 2013), which may explain why Hill coefficients may present higher values than expected for few genes. When fitting the mouse liver clock data, constraining all Hill coefficients to lower values (i.e. below 5) did not yield satisfactory results in terms of model fit. Therefore, the coefficients were constrained between 1 and 8 to allow larger values. A similar finding is reported in Woller et al. (2016).

S2-1.3 Model calibration

Gene and protein circadian expression datasets and their preprocessing

Datasets reported in Narumi et al. (2016) were used to estimate parameters of the quantitative core-clock model in mouse liver. These are time series of gene and protein circadian expression for the core-clock species, measured in the liver of C57BL6 male mice synchronized with 12 hours of light and 12 hours of darkness. Gene expression data was obtained through RT-PCR while the proteomics data were obtained through a novel mass spectrometry workflow allowing for the quantification of protein concentrations. Both datasets had a time resolution of 4 hours up to 48 h (resp. 24 h) for the genomics (resp. proteomics) study. Two mice were analyzed per time point. Our data processing workflows for genes and proteins is summarized below.

Genes:

- Sum every paralog into one species (e.g. *Per* = *Per1* + *Per2* + *Per3*) for each model variable, if need be
- Multiply every expression by the intracellular concentration of the reference gene *Tbp*. *Tbp* concentration was estimated using the value reported in Schmidt and Schibler (1995) expressed in molecules/cell. This quantity was converted from molecules/cell into mol/L by dividing by the Avogadro number and dividing by the cell volume assumed to be equal to 1pL.

Proteins:

- Average the concentration of all identified peptides for each protein present in the model.
- Sum every paralog-induced proteins into one species (e.g. PER = PER1 + PER2 + PER3).
- Convert values from molecules/cell to mol/L by dividing by the Avogadro number and dividing by the cell volume assumed to be equal to 1pL.

In the case of proteins, the data acquisition technique provided us with total intracellular protein amounts so that the mapping to the outputs of the clock model was not straightforward. Indeed, the model contains variables in different compartments (nucleus, cytoplasm) and under different forms (free or in complex). Hence, protein data was mapped to aggregated variables for BMAL, CLOCK, PER, CRY, REV-ERB and ROR which were defined as follows:

$$BMAL_{tot} = \frac{v_c}{v_t} BMAL_C + \frac{v_c}{v_t} CLOCK/BMAL_C + \frac{v_n}{v_t} CLOCK/BMAL_N \quad (2.S19)$$

$$CLOCK_{tot} = \frac{v_c}{v_t} CLOCK_C + \frac{v_c}{v_t} CLOCK/BMAL_C + \frac{v_n}{v_t} CLOCK/BMAL_N \quad (2.S20)$$

$$PER_{tot} = \frac{v_c}{v_t} PER_C + \frac{v_c}{v_t} PER/CRY_C + \frac{v_n}{v_t} PER/CRY_N \quad (2.S21)$$

$$CRY_{tot} = \frac{v_c}{v_t} CRY_C + \frac{v_c}{v_t} PER/CRY_C + \frac{v_n}{v_t} PER/CRY_N \quad (2.S22)$$

$$REV\text{-}ERB_{tot} = \frac{v_c}{v_t} REV\text{-}ERB_C + \frac{v_n}{v_t} REV\text{-}ERB_N \quad (2.S23)$$

$$ROR_{tot} = \frac{v_c}{v_t} ROR_C + \frac{v_n}{v_t} ROR_N \quad (2.S24)$$

where v_c (resp. v_n) is the cytoplasm (resp. nucleus) volume and v_t the total cell volume, so that $v_t = v_c + v_n$.

The nuclear amount of free proteins were assumed to be negligible since BMAL and CLOCK nuclear protein expressions shared the same circadian phase and amplitude, suggesting that both species exist majoritarily in complexed forms (Wang et al., 2017).

Derivation of a fully quantitative model

The previous clock model by Relógio *et al.* was expressed in arbitrary units (a.u.), due to the lack of absolutely quantitative information. The objective of this study was to design a fully quantitative model, able to predict not only the circadian phase and amplitude of genes and proteins, but also their mean levels, which could be of crucial importance in the context of pharmacology, where an optimal dose is sought for. Starting from the initial model, we have derived a first quantitative model through the following pipeline.

Let x be a model variable in a.u. and let x' be the scaled variable in **mol/L** s.t.

$$x' = x \times \frac{x^D_{\max}}{x^M_{\max}} = x \times x_{\max}$$

where $x^D_{\max}$ is the maximum value from the genes and protein data of Narumi et al. (2016), expressed in mol/L, and $x^M_{\max}$ the maximum value in the simulation of the original model, thus expressed in a.u.. Then $x \in [0, x^M_{\max}]$ and $x' \in [0, x^D_{\max}]$. This change of variable induces a scaling on the parameters which provided us with new parameter estimates that were used as initial guess in the parameter estimation

procedure. For instance, scaling the equation for z_6 results in:

$$\frac{dz_6}{dt} = k_{p_3} y_3 - \frac{v_c}{v_n} k_{iz6} z_6 - d_{z_6} z_6$$

$$\frac{dz'_6}{dt} = k_{p_3} \frac{z_{6\max}}{y_{3\max}} y'_3 - \frac{v_c}{v_n} k_{iz6} z'_6 - d_{z_6} z'_6$$

This showed that only the parameter k_{p_3} needed to be scaled whereas the parameters $k_{i_{z_6}}$ and $k_{d_{z_6}}$ were unchanged by the change of model unit. Scaled parameters belong to the following families: complex formation rates, transcription rates, activation/inhibition rates and production rates.

Reducing the number of free parameters using clock protein expression in *Bmal1*$^{-/-}$ mice

The article of Narumi et al. (2016) from which circadian genomics and proteomics data were used here for parameter estimation also featured data in *Bmal1* KO conditions. Clock protein concentrations were quantified in the liver of *Bmal1*$^{-/-}$ mice sacrificed at CT4 and CT16. These datasets were used to further constrain the model parameter estimation as follows. In a nutshell, three transcription parameters $V_{\max}$ could be expressed with respect to these datasets and to the other parameters of the model, thus leading to a reduction of the number of parameters to estimate.

Knockout of *Bmal1* is known to result in an arrhythmic circadian clock (Shimba et al., 2011) , as the transcription factor CLOCK/BMAL, a key component of the two main negative feedback loop, is absent. The obtained protein concentrations time series are flat and we can reasonably assume that the system is at equilibrium. Accounting for the absence of BMAL, hence of CLOCK/BMAL, Equation (2.S4) for *Rev-Erb* mRNA level (y_3), gives at steady state:

$$\begin{aligned} \frac{dy_3}{dt} &= 0 \\ \Leftrightarrow V_{3_{\max}} - d_{y_3} y_3 &= 0 \\ \Leftrightarrow \frac{V_{3_{\max}}}{d_{y_3}} &= y_3 \end{aligned}$$

Plugging this expression into Equation (2.S6) for REV-ERB$_C$ (z_6):

$$k_{p_3}\frac{V_{3_{\max}}}{d_{y_3}} - \frac{v_c}{v_n}k_{i_{z_6}}z_6 - d_{z_6}z_6 = 0$$
$$\Leftrightarrow (\frac{v_c}{v_n}k_{i_{z_6}} + d_{z_6})\frac{d_{y_3}}{k_{p_3}}z_6 = V_{3_{\max}}$$

We can now use the Rev-Erb protein level experimentally observed in KO mice, we have $Rev^{DataKO} = v_c z_6 + v_n x_5$. Equation (2.S8) for REV-ERB$_N$ (x_5) at steady state gives $k_{i_{z_6}}z_6 - d_{x_5}x_5 = 0$ which is used to express x_5 in terms of z_6. Finally,

$$V_{3_{\max}} = (\frac{v_c}{v_n}k_{i_{z_6}} + d_{z_6})\frac{d_{y_3}}{k_{p_3}}\frac{Rev^{DataKO}}{v_c + v_n\frac{k_{i_{z_6}}}{d_{x_5}}}, \tag{2.S25}$$

which results in an equation for $V_{3_{\max}}$ only depending on the KO data and on the parameters of the model. The procedure is exactly the same for $V_{4_{\max}}$ which leads to the following formula:

$$V_{4_{\max}} = (\frac{v_c}{v_n}k_{i_{z_7}} + d_{z_7})\frac{d_{y_4}}{k_{p_4}}\frac{Ror^{DataKO}}{v_c + v_n\frac{k_{i_{z_7}}}{d_{x_6}}} \tag{2.S26}$$

For $V_{6_{\max}}$, at steady state, the equation for CLOCK mRNA (y_6) becomes:

$$V_{6_{\max}}\frac{1 + j\left(\frac{x_6}{k_{t_6}}\right)^b}{1 + \left(\frac{x_5}{k_{i_6}}\right)^c + \left(\frac{x_6}{k_{t_6}}\right)^b} = d_{y_6}y_6$$

From the above computations,

$$x_5 = \frac{k_{iz_6}}{d_{x_5}}\frac{Rev^{DataKO}}{v_c + v_n\frac{k_{i_{z_6}}}{d_{x_5}}}$$

Similarly, an expression for x_6 can be derived as:

$$x_6 = \frac{k_{iz_7}}{d_{x_6}}\frac{ROR^{DataKO}}{v_c + v_n\frac{k_{i_{z_7}}}{d_{x_6}}}$$

Hence, the Hill-function-like term is then entirely determined and will be denoted H. Using Equation (S2-1.2), y_6 can be expressed in terms of z_5:

$$\frac{V_{6_{\max}}}{d_{y_6}} k_{p_6} H = d_{z_5} z_5$$

We assumed that the observed CLOCK protein expression level in KO mice only corresponded to cytoplasmic CLOCK (z_5), as co-expression experiments performed in Kondratov et al. (2003) showed a reduced nuclear fraction of CLOCK to the nucleus in the absence of BMAL1. Hence we have: $Clock^{DataKO} = z_5$ and finally:

$$V_{6_{\max}} = \frac{d_{y_6}}{k_{p_6} H} d_{z_5} Clock^{DataKO} \tag{2.S27}$$

For $V_{1_{\max}}$ and $V_{2_{\max}}$, the complexation of PER^*_C and CRY_C introduces additional variables as well as non linear terms in the protein equations. In this case, the existence of an equation linking $V_{1_{\max}}$ and $V_{2_{\max}}$ to other parameters becomes conditional to the satisfaction of some stability criterion, leading to cumbersome computations in an optimization pipeline, therefore these parameters remained estimated. Similarly, $V_{5_{\max}}$ which corresponds to *Bmal* transcription, could not be derived this way.

Fitting the circadian time series through a least square approach

The first part of the parameter fitting procedure aimed to reproduce the circadian genomics and proteomics time-resolved data from Narumi et al. (2016). In order to do so, we defined a cost function $L(\theta)$ as the sum of squared residuals across species $s \in S$, divided by the maximum value of a species's concentration along time:

$$L(\theta) = \sum_{s \in S} \left(\sum_{t_i \in \mathcal{T}_g} \sum_{j \in \mathcal{J}} \left(\frac{y^{(s)}_{t_i,j} - \hat{y}^{(s)}_{t_i,j}(\theta)}{y^{(s)}_{\max}} \right)^2 + \sum_{t_i \in \mathcal{T}_p} \sum_{j \in \mathcal{J}} \left(\frac{Y^{(s)}_{t_i,j} - \hat{Y}^{(s)}_{t_i,j}(\theta)}{Y^{(s)}_{\max}} \right)^2 \right) \tag{2.S28}$$

where y (resp. Y) is the mRNA data (resp. protein data). $\hat{y}$ is the model simulation for the mRNA expression, $\hat{Y}$ refers to the convex combination of protein species defined in Equation (S2-1.3) whose model simulation will be compared to a protein Y in the data. $\mathcal{T}_g = \{0, 4, 8, \ldots, 44\}$ (resp. $\mathcal{T}_p = \{0, 4, 8, \ldots, 24\}$) are the circadian times of data measurement for the mRNAs (resp. proteins), $\mathcal{J} = \{1, 2\}$ is the number of the replica. Finally, for each species $s \in S$, $y^{(s)}_{\max} = \max_{\substack{t_i \in \mathcal{T}_g \\ j \in J}} y^{(s)}_{t_i,j}$ and $Y^{(s)}_{\max} = \max_{\substack{t_i \in \mathcal{T}_p \\ j \in J}} Y^{(s)}_{t_i,j}$

Dividing by the maximum values ensures that genes and proteins have the same importance in the cost function even though their mean levels are different by several orders of magnitude.

Constraining the parameter estimation with additional biological knowledge

It is well known in systems biology that fitting even large amounts of data with a model including numerous parameters can end up in many adequate parameter sets (Gutenkunst et al., 2007). Therefore, we used biological knowledge as a way to constraint the parameter estimation.

The first type of constraints considered acts directly on the parameter search interval to ensure biological coherence, based on experimental data. The transcription rates of ≈ 5,000 genes were measured by Schwanhäusser et al. (2011). Using the maximum of these rates μ, transcription rates of the model can be upper bounded.

Indeed, considering for instance Equation (2.S4) for *Rev-Erb*:

$$\frac{dy_3}{dt} = V_{3_{\max}} \frac{1 + g\left(\frac{x_1}{k_{t_3}}\right)^v}{1 + \left(\frac{x_2}{k_{i_3}}\right)^w \left(\frac{x_1}{k_{t_3}}\right)^v + \left(\frac{x_1}{k_{t_3}}\right)^v} - d_{y_3} y_3$$

$$\implies \lim_{\substack{x_1 \to +\infty \\ x_2 \to 0}} \frac{dy_3}{dt} = g V_{3_{\max}} - d_{y_3} y_3$$

This gives us $g V_{3_{\max}} \leqslant \mu$. Similarly we can derive a constraint for each other gene of the model. In practice we set $\mu = 5\text{nmol} \times L^{-1} \times \text{hours}^{-1}$ for each gene.

Due to their connection with half-lives, degradation parameters can also be constrained so as to represent plausible half-lives. Then again, the data reported by Schwanhäusser et al. (2011) can be useful. We then chose to bound degradation parameters between $[10^{-4}, 3]$

The second type of constraints included act on the model outputs, constraining them to a desired behavior. The protein data acquisition technique provided total protein amount but information on the relative quantities of proteins either free or in complexe was missing. Using co-immunoprecipitation data from Zheng et al. (2019b) and immunodepletion from Aryal et al. (2017), we arrived at the final expression to bind the concentration of the complexes.

$$0.15\, CLOCK_{tot} \leqslant v_c CLOCK/BMAL_C + v_n CLOCK/BMAL_N \leqslant 0.85\, CLOCK_{tot}$$
$$PER/CRY_C^{tot} \geqslant 0.5 PER_C^{tot}$$

Where $\text{CLOCK}_{tot} = v_c(\text{CLOCK}_C + \text{CLOCK/BMAL}_C) + v_n\ \text{CLOCK/BMAL}_N$

All these constraints were incorporated in the optimization in such a way that

a resample of the parameter vector was initiated whenever any of them was violated, therefore the parameter vector responsible for such constraint violation was discarded.

S2-1.4 Robustness analysis

We assessed the robustness of our model against parameter perturbations in terms of oscillatory behaviour. More precisely, in this case, let $Y = f(\theta)$ be the boolean variable indicating whether or not a model parametrized by a vector θ outputs periodic simulation or not. The parameter vector θ will be slightly perturbed around its base value for each coordinate, and is therefore seen as a random variable. A grid $[0.9 * \theta, 1.1 * \theta]$ is then produced and Y evaluated for up to 100000 samples within that grid. 73% of these evaluations led to periodic simulations.

Sobol sensitivity total order indices were then computed (Sobol, 2001). The total-order-index S_{T_i} measures the contribution of a variable θ_i in the output variance of a variable Y, including all variance caused by θ_i's interactions of any order with the other variables θ_j.

$$S_{T_i} = \frac{E_{\theta \sim i}\left(\mathrm{Var}_{\theta_i}(Y|\theta_{\sim i})\right)}{\mathrm{Var}(Y)}$$

where $\mathrm{Var}(\cdot|\cdot)$ is the conditional variance, and $\theta_{\sim i}$ is the set of all parameters except θ_i.

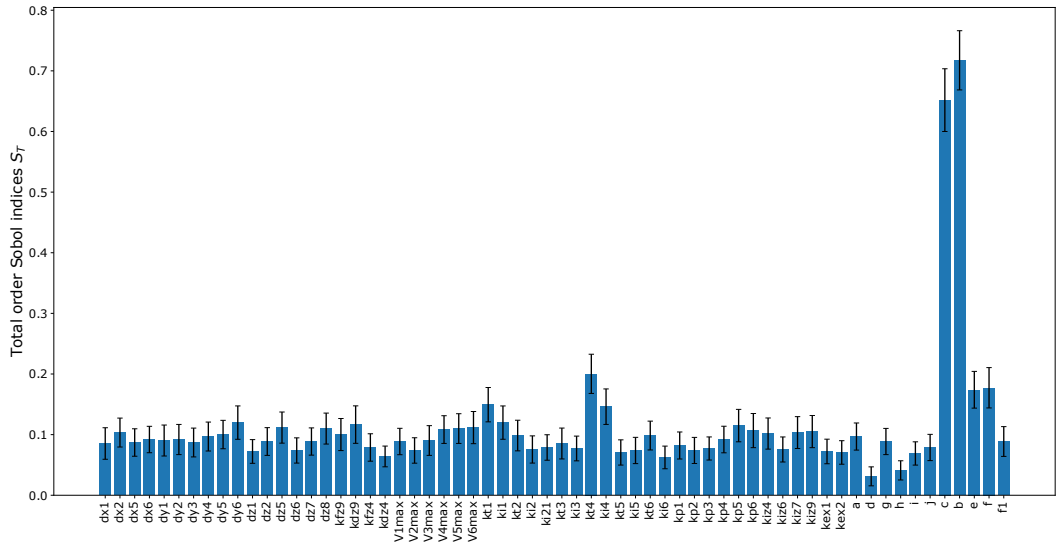


Figure 2.S1: Total order Sobol indices. 100 000 simulations of the model were used for the estimation.

The fact that mutual Hill coefficients b and c produces high values for total order indices was expected as these parameters are present in five equations.

S2-2 Additional figures to the core-clock model

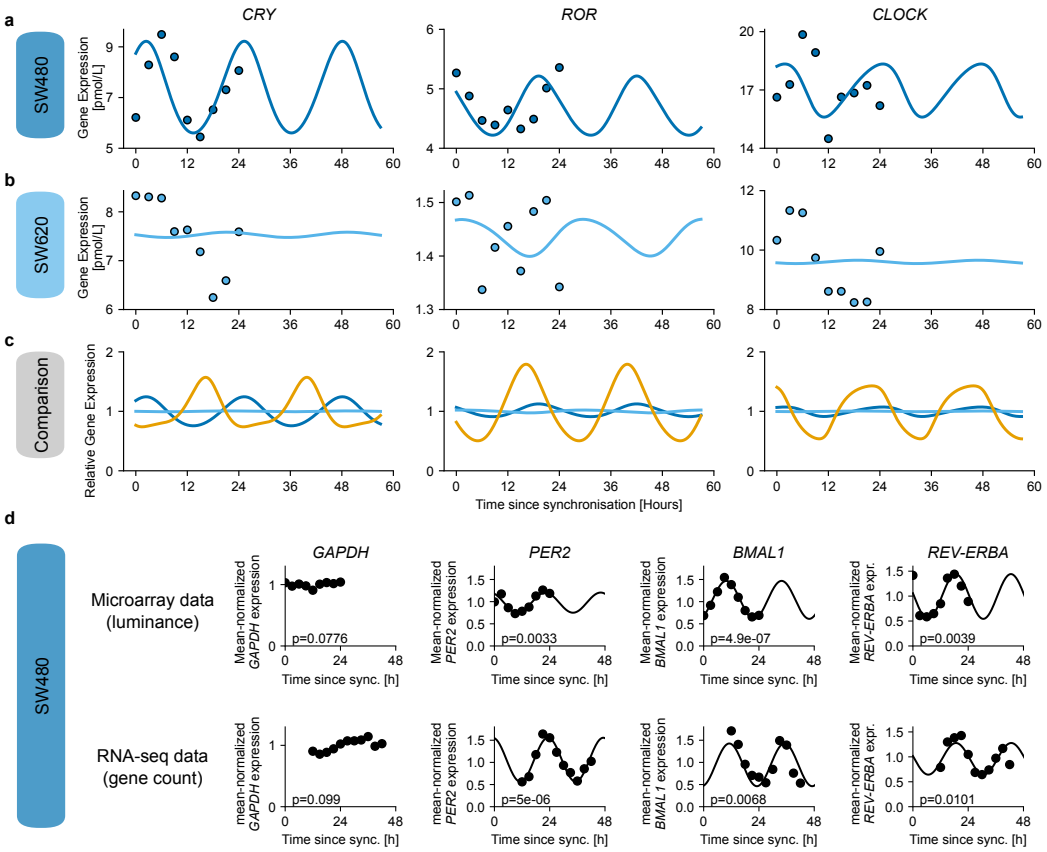


Figure 2.S2: Fit of the core-clock model to cell lines derived from human cancer. This supplements Fig. 3.3 of the main text, showing the remaining genes of the core-clock. **a** Fit (line) and experimental data (dots) for the SW480 cell line and **(b)** the SW620 cell line. **c** Comparison of the model fit for liver (orange), SW480 (dark blue) and SW620 (sky blue). *Bmal1* circadian phases were aligned for mouse liver and SW480 cell line and all gene expression were normalized to the mesor to allow for comparison. **d** Comparison of microarray and RNA-seq time series data for the reference gene GAPDH and the core-clock genes measured for the RT-qPCR data of the SW480 and SW620 cell lines. GAPDH is particularly well suited as reference gene as it is expressed at a much higher abundance compared to core-clock genes, and it does not show circadian oscillations, as confirmed by a cosinor analysis with non-significant p-values for GAPDH. Experimental mean-normalized measures (dots) with the result from a harmonic regression (lines), using a p-value threshold of p=0.05.

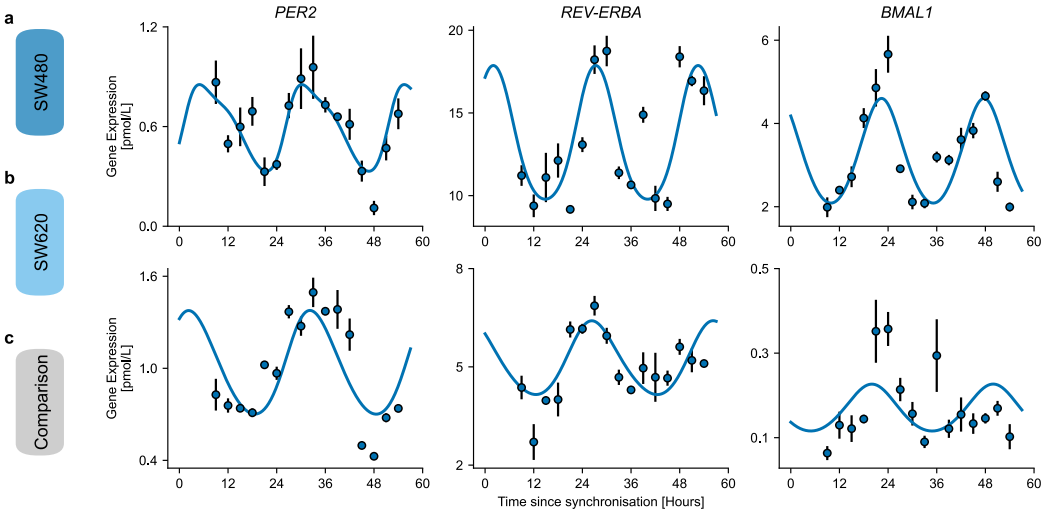


Figure 2.S3: SW480 control and Bmal knock-down can be fitted by similar core-clock models. a Fit of the core-clock model to qRT-PCR data of the SW480 cell line in Ctrl condition. **b** Fit of the core-clock model to qRT-PCR data of the SW480 cell line in shBmal condition, varying only *Bmal1* transcription rate.

S2-3 The clock-irinotecan model

Gene name (A)	Action	Gene name (B)	Reference
CLOCK	**Binds to**	BMAL1	Relógio et al. (2011)
CLOCK:BMAL1	**Activates**	*PER*	Reppert and Weaver (2001) Ueda et al. (2002)
CLOCK:BMAL1	**Activates**	*CRY*	Reppert and Weaver (2001) Ueda et al. (2002)
CLOCK:BMAL1	**Activates**	*ROR*	Reppert and Weaver (2001) Ueda et al. (2002)
CLOCK:BMAL1	**Activates**	*REV-ERB*	Reppert and Weaver (2001) Ueda et al. (2002)
PER	**Binds to**	CRY	Lee et al. (2001)
PER/CRY	**Inhibits**	CLOCK:BMAL1 transcription	Zhang and Kay (2010)
ROR	**Activates**	*BMAL1*	Guillaumond et al. (2005)
REV-ERB	**Inhibits**	*BMAL1*	Guillaumond et al. (2005)
REV-ERB	**Inhibits**	*CRY*	Liu et al. (2008)
CLOCK:BMAL1	**Activates**	*DBP* (*PAR bZip*)	Ripperger and Schibler (2006)
CLOCK:BMAL1	**Activates**	*HLF* (*PAR bZip*)	Takahashi (2017)
ROR	**Activates**	*NFIL3* (*E4BP4*)	Takahashi (2017)
REV-ERB	**Inhibits**	*NFIL3* (*E4BP4*)	Takahashi (2017)
CLOCK:BMAL1	**Activates**	*PPARa*	Oishi et al. (2005)
DBP	**Activates**	*TOP1*	Yang et al. (2009)
NFIL3	**Inhibits**	*TOP1*	Yang et al. (2009)
CLOCK:BMAL1	**Activates**	*TOP1*	Yang et al. (2009)
TOP1	**Inhibits**	*BMAL1*	Onishi and Kawano (2012)
NFIL3	**Activates**	*CES2*	Zhao et al. (2018)
NFIL3	**Inhibits**	*REV-ERB*	Zhao et al. (2018)
REV-ERB	**Inhibits**	*CES2*	Zhao et al. (2018)
DBP (PAR bZip)	**Activates**	*ABCC2*	Yu et al. (2019)
NFIL3	**Inhibits**	*ABCC2*	Yu et al. (2019)
HLF (PAR bZip)	**Activates**	*ABCB1*	Murakami et al. (2008)
NFIL3	**Inhibits**	*ABCB1*	Murakami et al. (2008)
PPARa	**Activates**	*UGT1A1*	Senekeo-Effenberger et al. (2007)

Table 2.S3: Overview over the connections in the translation-transcription network, with references for experimental reports of the connection.

The clock-irinotecan model extends the core-clock model from Section S2-1.2, with the dynamics of *Bmal1* and *Rev-Erb* as stated below, by the interactions as depicted in Fig. 3.4 of the main text.

The variables and parameters of the core-clock model are used for elements belonging to the core-clock. Additional dynamic variables of the clock-irinotecan model are stated in Supplementary Table 2.S4. For genes only the first letter is uppercase, proteins are set in uppercase, concentrations are denoted with square brackets [.]. For simplicity, the model does not explicitly differentiate between cytosolic and nuclear proteins for irinotecan PK/PD-related genes.

Species Name	Variable name
Bmal1	y_5
Rev-Erb	y_3
Ces2	Ces
Ugt1a1	Ugt
Abcb1	$Abcb$
Abcc	$Abcc$
Pparα	$Ppar$
Top1	Top
PAR bZip	Par
Nfil3	$Nfil$
CES1	CES
UGT1A1	UGT
ABCB1	$ABCB$
ABCC	$ABCC$
PPARα	$PPAR$
TOP1	TOP
PAR bZIP	PAP
NFIL3	$NFIL$

Table 2.S4: List of dynamical state variables representing mRNAs and proteins. For the irinotecan-related genes, the model uses the same variable names for mRNA and proteins, the latter in uppercase.

S2-3.1 Feedback to the core-clock: Transcription of *Bmal1* and *Rev-Erb*

The transcription of *Bmal1* and *Rev-Erb* is replaced by the following equations, which implement the feedback from *Top1* and *Nfil3*.

Bmal1

$$\frac{\mathrm{d}y_5}{\mathrm{d}t} = V_{5_{\max}} \frac{1 + i\left(\frac{x_6}{k_{t_5}}\right)^b}{1 + \left(\frac{x_5}{k_{i_5}}\right)^c + \left(\frac{x_6}{k_{t_5}}\right)^b} \frac{1}{1 + \left(\frac{[\mathrm{TOP}]}{i_{\mathrm{BmalTop}}}\right)^c} - d_{y_5} y_5 \tag{2.S29}$$

Rev-erb

$$\frac{\mathrm{d}y_3}{\mathrm{d}t} = V_{3_{\max}} \frac{1 + g\left(\frac{x_1}{k_{t_3}}\right)^b}{1 + \left(\frac{x_2}{k_{i_4}}\right)^c \left(\frac{x_1}{k_{t_3}}\right)^b + \left(\frac{x_1}{k_{t_3}}\right)^b} \frac{1}{1 + \left(\frac{[\mathrm{NFIL}]}{i_{\mathrm{RevNfil}}}\right)^c} - d_{y_3} y_3 \tag{2.S30}$$

S2-3.2 Equations of the network connecting core-clock and irinotecan dynamics

Translation

The protein is degraded and grows by translation, where d_{PROTEIN} is the degradation rate, and r_{PROTEIN} is a translation rate that either describes only the translation of the gene to the cytoplasmic protein (first four variables of Table 2.S1, or both the translation step as well as the import of this protein into the nucleus (last four variables of Table 2.S1).

For the elements of Table 2.S4 the step from genes to proteins has the same structure:

$$\frac{\mathrm{d}PROTEIN}{\mathrm{d}t} = r_{\mathrm{PROTEIN}}\, Gene - d_{\mathrm{PROTEIN}}\, PROTEIN \tag{2.S31}$$

Transcription

The transcription of all variables of Table 2.S4 and of *Bmal1* and *Rev-Erb* follow dynamics with the following structure:

$$\frac{\mathrm{d}Gene}{\mathrm{d}t} = V_{\mathrm{Gene}}\, \mathbb{T}(Gene) - d_{\mathrm{Gene}}\, Gene, \tag{2.S32}$$

where V_{Gene} is the maximal transcription rate of the gene $Gene$, d_{Gene} is the degradation rate of the gene, and $\mathbb{T}(Gene)$ is the transcription function as defined below, that includes the interactions between different elements.

Transcription functions For simplicity, the Hill coefficients of transcription for activation and inhibition, b and c, are the same for all equations.

Pparα

$$\mathbb{T}(Ppar) = \frac{1 + f_{\mathrm{Ppar}}\left(\frac{x_1}{a_{\mathrm{Ppar}}}\right)^b}{1 + \left(\frac{x_2}{i_{\mathrm{Ppar}}}\right)^c \left(\frac{x_1}{a_{\mathrm{Ppar}}}\right)^b + \left(\frac{x_1}{a_{\mathrm{Ppar}}}\right)^b} \tag{2.S33}$$

PAR bZip

Parameter name	Parameter symbol
transcription function as stated below	$\mathbb{T}(Gene)$
maximal transcription rate of the gene $Gene$	V_{Gene}
degradation rate of the gene $Gene$	d_{Gene}
transcription fold activations	f_{Gene}
activation rates	a_{Gene}
inhibition rates for inhibition by one protein	i_{Gene}
inhibition rate of TOP1 on *Bmal1*	i_{BmalTop}
inhibition rate of NFIL3 on *Rev-Erb*	i_{RevNfil}
activation rate of CLOCK/BMAL on *Top1*	a_{Top}
activation rate of PAR bZIP on *Top1*	a_{TopPar}
inhibition rate of PER/CRY on *Top1*	i_{Top}
inhibition rate of NFIL3 on *Top1*	i_{TopNfil}
Hill coefficient of activation	b
Hill coefficient of inhibition	c
rate of translation* of the protein $PROTEIN$	r_{PROTEIN}
degradation rate of the protein $PROTEIN$	d_{PROTEIN}

Table 2.S5: List of model parameters. $*$ For PPARα, TOP1, PAR bZIP and NFIL3, r_{PROTEIN} is the rate of translation and the import of the protein into the nucleus.

$$\mathbb{T}(Par) = \frac{1 + f_{\text{Par}}\left(\frac{x_1}{a_{\text{Par}}}\right)^b}{1 + \left(\frac{x_2}{i_{\text{Par}}}\right)^c \left(\frac{x_1}{a_{\text{Par}}}\right)^b + \left(\frac{x_1}{a_{\text{Par}}}\right)^b} \tag{2.S34}$$

Ugt1a1

$$\mathbb{T}(Ugt) = \frac{1 + f_{\text{Ugt}}\left(\frac{[\text{PPAR}]}{a_{\text{Ugt}}}\right)^b}{1 + \left(\frac{[\text{PPAR}]}{a_{\text{Ugt}}}\right)^b} \tag{2.S35}$$

Nfil3

$$\mathbb{T}(Nfil) = \frac{1 + f_{\text{Nfil}}\left(\frac{x_6}{a_{\text{Nfil}}}\right)^b}{1 + \left(\frac{x_5}{i_{\text{Nfil}}}\right)^c + \left(\frac{x_6}{a_{\text{Nfil}}}\right)^b} \tag{2.S36}$$

Ces

$$\mathbb{T}(Ces) = \frac{1 + f_{\text{Ces}}\left(\frac{[\text{NFIL}]}{a_{\text{Ces}}}\right)^b}{1 + \left(\frac{x_5}{i_{\text{Ces}}}\right)^c + \left(\frac{[\text{NFIL}]}{a_{\text{Ces}}}\right)^b} \tag{2.S37}$$

Abcb1

$$\mathbb{T}(Abcb) = \frac{1 + f_{\text{Abcb}} \left(\frac{[\text{PAR}]}{a_{\text{Abcb}}}\right)^b}{1 + \left(\frac{[\text{NFIL}]}{i_{\text{Abcb}}}\right)^c + \left(\frac{[\text{PAR}]}{a_{\text{Abcb}}}\right)^b} \tag{2.S38}$$

Abcc

$$\mathbb{T}(Abcc) = \frac{1 + f_{\text{Abcc}} \left(\frac{[\text{PAR}]}{a_{\text{Abcc}}}\right)^b}{1 + \left(\frac{[\text{NFIL}]}{i_{\text{Abcc}}}\right)^c + \left(\frac{[\text{PAR}}{a_{\text{Abcc}}}\right)^b} \tag{2.S39}$$

Top1

$$\mathbb{T}(Top) = \frac{1 + f_{\text{Top}} \left(\frac{x_1}{a_{\text{Top}}}\right)^b}{1 + \left(\frac{x_2}{i_{\text{Top}}}\right)^c \left(\frac{x_1}{a_{\text{Top}}}\right)^b + \left(\frac{x_1}{a_{\text{Top}}}\right)^b} \frac{1 + f_{\text{TopPar}} \left(\frac{[\text{PAR}]}{a_{\text{TopPar}}}\right)^b}{1 + \left(\frac{[\text{NFIL}]}{i_{\text{TopNfil}}}\right)^c + \left(\frac{[\text{PAR}]}{a_{\text{TopPar}}}\right)^b} \tag{2.S40}$$

Implementation of post-transcriptional modifications

For the fit of the SW480 cell line and the liver tissue, we replace Equation (2.S32) with $\mathbb{T}$ from Equation (2.S39) and Equation (2.S37) with the following set of equations, which allow us to improve the fit of *CES2* and *ABCC*:

$$\frac{\mathrm{d}Ces^*}{\mathrm{d}t} = V_{\text{Ces}} \mathbb{T}(Ces^*) - d_{\text{Ces}}\, Ces^* \tag{2.S41}$$

$$\frac{\mathrm{d}Ces^{**}}{\mathrm{d}t} = s_{\text{Ces}^*} Ces^* - d_{\text{Ces}^*}\, Ces^{**} \tag{2.S42}$$

$$\frac{\mathrm{d}Ces^{***}}{\mathrm{d}t} = s_{\text{Ces}^*} Ces^{**} - d_{\text{Ces}^*}\, Ces^{***} \tag{2.S43}$$

$$\frac{\mathrm{d}Ces}{\mathrm{d}t} = s_{\text{Ces}^*} Ces^{***} - d_{\text{Ces}^*}\, Ces \tag{2.S44}$$

$$\frac{\mathrm{d}Abcc^*}{\mathrm{d}t} = V_{\text{Abcc}} \mathbb{T}(Abcc^*) - d_{\text{Abcc}}\, Abcc^* \tag{2.S45}$$

$$\frac{\mathrm{d}Abcc^{**}}{\mathrm{d}t} = s_{\text{Abcc}^*} Abcc^* - d_{\text{Abcc}^*}\, Abcc^{**} \tag{2.S46}$$

$$\frac{\mathrm{d}Abcc}{\mathrm{d}t} = s_{\text{Abcc}^*} Abcc^{**} - d_{\text{Abcc}^*}\, Abcc \tag{2.S47}$$

with $\mathbb{T}(Ces*)$ and $\mathbb{T}(Abcc*)$ given by Equation (2.S37) and Equation (2.S39) replacing Ces by $Ces*$ and $Abcc$ by $Abcc*$.

$$\frac{d[CPT_{out}]}{dt} = \frac{V_{in}}{V_{out}}(-k_{upCPT}[CPT_{out}] + \frac{V_{effCPT}[\text{ABCB}][CPT_{in}]}{K_{effCPT} + [CPT_{in}]}) \tag{2.S48}$$

$$\frac{d[CPT_{in}]}{dt} = k_{upCPT}[CPT_{out}] - \frac{V_{effCPT}[\text{ABCB}][CPT_{in}]}{K_{effCPT} + [CPT_{in}]} - \frac{V_{ces}[\text{CES}][CPT11_{in}]}{K_{ces} + [CPT11_{in}]} \tag{2.S49}$$

$$\frac{d[SN_{out}]}{dt} = \frac{V_{in}}{V_{out}}(-k_{upSN}[SN_{out}] + \frac{V_{effSN}[\text{ABCC}][SN_{in}]}{K_{effSN} + [SN_{in}]}) \tag{2.S50}$$

$$\begin{aligned}\frac{d[SN_{in}]}{dt} = {} & k_{upSN}[SN_{out}] - \frac{V_{effSN}[\text{ABCC}][SN_{in}]}{K_{effSN} + [SN_{in}]} + \frac{V_{ces}[\text{CES}][CPT_{in}]}{K_{ces} + [CPT_{in}]} \\ & - \frac{V_{ugt}[\text{UGT}][SN_{in}]}{K_{ugt} + [SN_{in}]} - k_{f2}[DNATOP1][SN38_{in}] + k_{r2}[Compl]\end{aligned} \tag{2.S51}$$

$$\begin{aligned}\frac{d[TOP1]}{dt} = {} & k_{ftop} - k_{dtop}[TOP] - k_{f1}[TOP1][DNA_{free}] \\ & + k_{r1}[DNATOP1] + k_{r2}[Compl]\end{aligned} \tag{2.S52}$$

$$\begin{aligned}\frac{d[DNATOP1]}{dt} = {} & k_{f1}[TOP][DNA_{free}] - k_{f2}[DNATOP1][SN_{in}] \\ & - k_{r1}[DNATOP1]\end{aligned} \tag{2.S53}$$

$$\frac{d[Compl]}{dt} = k_{f2}[DNATOP1][SN38_{in}] - k_{r2}[Compl] - k_{Irr}[Compl] \tag{2.S54}$$

$$\frac{d[Icompl]}{dt} = k_{Irr}[Compl] \tag{2.S55}$$

$$\frac{d[Apop]}{dt} = k_{apop}([Compl] + [Icompl]) \tag{2.S56}$$

Pharmacodynamics/-kinetics

For the pharmacodynamics/-kinetics of irinotecan (CPT11), the model is supplemented by the following equations. These equations correspond to equations (1) to (10) from Ballesta et al. (2011); Dulong et al. (2015), with the explicit tracking of SNG (5-6) removed.

The original model was used to predict apoptosis for Caco-2 cells. The following modifications were used in a first attempt to simulate cytotoxicity for CRC cell lines. We used the model to predict cell death for SW480 and SW620 cells by replacing the cosine fit with the dynamics that result from the clock-irinotecan network, assuming protein translation according to Equation (2.S31) with constant degradation rate, chosen as described below. For the cell line Caco-2 (cell line derived from a human colorectal adenocarcinoma), the PK/PD model was fitted to cell death following irinotecan treatment (Ballesta et al., 2011; Dulong et al., 2015). The model

by Dulong et al. (2015) assumes that not only drug-induced DNA damage, but also the apoptosis mechanism itself shows a circadian oscillation. The amplitude of the latter oscillation is for simplicity set to zero for the CRC cell lines (k_{apop} =const. in Equation (2.S56)) and we compare experimental cytotoxicity to "drug-induced DNA damage" in the model, Apop=0. The formulation of Equation (2.S31) implies that acrophases and relative amplitudes of the proteins depend only on the degradation rates, while translation rates only affect absolute amplitudes (Lück et al., 2014). To exemplify this numerically, we run 1000 implementations of the model with parameters drawn randomly from a uniform distribution between 1 to 100000 for the translation rates, and 0.01 to 3 for the protein degradation rates. As expected, only the protein degradation significantly affects the relative amplitude and phase (Fig 2.S6). We choose for the proteins UGT1A1, CES2, ABCB and ABCC a degradation rate of 1.22 hour^{-1} and a translation rate of 45716.3 hour^{-1}, which entails large oscillation amplitudes and phase delays for the proteins compared to their mRNA around 4 hours. Maximal protein concentration for UGT1A1, CES2, ABCB and ABCC are scaled to the maximal concentration used in the original model. As the protein concentrations predicted by the transcription-translation model are rescaled, the prediction of toxicity is based on the relative amplitude and the phase of the protein oscillations, but not on their absolute levels. The model of Dulong et al. (2015) explicitly involves ABCG, which is in our case replaced by ABCC with an appropriate rescaling. The PK/PD irinotecan model by Dulong et al. (2015) simulates cell death relative to control for each time point since the start of treatment. Compared to the original implementation of the model, we consider the area under the curve as representative for toxicity. To be comparable with the experimental analysis, we simulate for different treatment times a long treatment duration (30 hours) and calculated the area under the curve. We considered the total amount of cell death for different treatment times as prediction curve for the experimentally measured cytotoxicity, without considering the oscillation amplitude, which is dependent on the simulation time: for short treatment durations, the acrophase of the resulting toxicity curve depends on treatment duration, but for long treatment durations above 24 hours, the predicted acrophase is stable, and the treatment duration is only affecting the amplitude of the oscillation, which we hence do not consider. Indeed, as the model by Dulong et al. (2015) does not implement proliferation, the resulting toxicity can only inform on the phase of the toxicity curve, but not on its amplitude. The resulting toxicity profile seems to predict the phase of the toxicity maximum for SW480 cells, whereas the oscillation amplitude of the toxicity profile is strongly underestimated compared to the experimental data, see Supplementary Fig. 2.S7.

In an extension of the model, we introduce exponential proliferation and cell death, and introduce circadian oscillations in the protein translation. We replace the simple protein dynamics of Equation (2.S31) with constant degradation rate

d_{PROTEIN} with protein dynamics that include a circadian oscillation in the degradation rate, i.e.

$$d_{\text{PROTEIN}}(t) = \gamma_{\text{PROTEIN}}(1 + A_{\text{PROTEIN}} \cos(\omega t + \phi_{\text{PROTEIN}})) \tag{2.S57}$$

where ω is 2π divided by the period of the numerical fit of the respective cell line, γ_{PROTEIN}, A_{PROTEIN} and ϕ_{PROTEIN} are magnitude, amplitude and phase of the circadian protein degradation for the proteins UGT1A1, CES2, ABCB and ABCC. We replace Equation (2.S56) with explicit equations for the amount of living cells, N, and dead cells, D,

$$\begin{aligned}\frac{\mathrm{d}N}{\mathrm{d}t} &= k_{\text{prol}}(1 - p_{\text{treat}}([Compl] + [Icompl]))N \\ &\quad - (k_{\text{apop}}(1 + A_{\text{apop}} \cos(\omega t + \phi_{\text{apop}}))([Compl] + [Icompl]) \\ &\quad + k_{\text{control}})N \end{aligned} \tag{2.S58}$$

$$\begin{aligned}\frac{\mathrm{d}D}{\mathrm{d}t} &= (k_{\text{apop}}(1 + A_{\text{apop}} \cos(\omega t + \phi_{\text{apop}}))([Compl] + [Icompl]) \\ &\quad + k_{\text{control}})N \end{aligned} \tag{2.S59}$$

where ω is 2π divided by the period of the numerical fit of the respective cell line. k_{prol} and k_{control} are the proliferation and cytotoxicity rates of the untreated control, p_{treat} scales the changes in proliferation due to treatment, k_{apop}, A_{apop} and ϕ_{apop} are the parameters (magnitude, amplitude and phase) of the circadian variation in cytotoxicity; these parameters are fitted, simultaneously with the parameters of the circadian protein dynamics, such that the number of dead cells, D, fits the cytotoxicity dynamics of the incucyte experiments, see Fig. 3.6 of the main text.

List of model parameters for the clock-irinotecan model

Fitting of the clock-irinotecan model to liver tissue and SW480 and SW620 cell lines results in parameters shown in Table 2.S6. Fig. 2.S4 shows the resulting fit exemplary for the SW480 cell line. Fig. 2.S5 compares acrophases (timing of the first peak relative to the period) and relative amplitudes ([max-min]/max) for the three different model fits.

Parameter	Name	Liver	SW480	SW620
Degradation rates for nuclear proteins or nuclear protein complexes [hour^{-1}]				
d_{x_1}	CLOCK/BMAL	0.079	0.788	0.1
d_{x_2}	PER/CRY$^{tot}_N$	0.474	1.26	2.76
d_{x_5}	REV-ERB$_N$	0.285	2.43	2.49
d_{x_6}	ROR$_N$	2.47	0.798	2.35
Degradation rates for mRNAs [hour^{-1}]				

d_{y_1}	*Per*	2.35	0.543	2.48
d_{y_2}	*Cry*	2.49	2.5	0.0301
d_{y_3}	*Rev-Erb*	0.46	0.432	2.43
d_{y_4}	*Ror*	0.157	2.44	0.818
d_{y_5}	*Bmal*	1.26	2.49	0.156
d_{y_6}	*Clock*	1.31	0.581	2.47
	Degradation rates for cytoplasmic proteins [hour^{-1}]			
d_{z_1}	CRY_C	0.0472	0.413	0.153
d_{z_2}	PER_C	0.0455	2.46	2.08
d_{z_5}	$CLOCK_C$	1.75	2.32	0.371
d_{z_6}	$REV\text{-}ERB_C$	0.116	0.314	2.09
d_{z_7}	ROR_C	0.0788	1.01	0.584
d_{z_8}	$BMAL_C$	2.5	0.629	2.36
	Reaction rates for complex formation [mol$\times L^{-1}\times$hours^{-1}]			
$k_{f_{z_9}}$	$CLOCK/BMAL_C$	0.394	0.0121	0.0209
$k_{f_{z_4}}$	PER/CRY_C^{tot}	0.402	0.000822	0.0442
	Reaction rates for complex dissociation [hours^{-1}]			
$k_{d_{z_9}}$	$CLOCK/BMAL_C$	2.98	2.77	2.99
$k_{d_{z_4}}$	PER/CRY_C^{tot}	1.61	1.11	2.98
	Transcription rates [mol$\times L^{-1}\times$ hours^{-1}]			
$V_{1_{max}}$	*Per*	9.79e-12	1.98e-08	1.03e-08
$V_{2_{max}}$	*Cry*	4.54e-09	2.89e-11	7.72e-11
$V_{3_{max}}$	*Rev-Erb*	2.99e-13	5.7e-09	2.33e-11
$V_{4_{max}}$	*Ror*	9.71e-12	1.49e-11	1.37e-12
$V_{5_{max}}$	*Bmal*	5.8e-12	2.57e-11	7.57e-13
$V_{6_{max}}$	*Clock*	8.16e-13	6.48e-12	9.46e-11
	Activation/inhibition rates [mol$\times L^{-1}$]			
k_{t_1}	*Per*-activation rate	1.48e-10	1.92e-10	2.04e-10
k_{i_1}	*Per*-inhibition rate	8.22e-11	1.97e-11	1.11e-10
k_{t_2}	*Cry*-activation rate	1.76e-09	1.61e-09	7.22e-10
k_{i_2}	*Cry*-inhibition rate	1.13e-13	7.75e-13	2.4e-13
$k_{i_{21}}$	*Cry*-inhibition rate	4.18e-10	3.45e-08	4.69e-08
k_{t_3}	*Rev-Erb*-activation rate	1.05e-10	1.28e-13	1.63e-11
k_{i_3}	*Rev-Erb*-inhibition rate	4.59e-09	1.25e-10	9.56e-10
k_{t_4}	*Ror*-activation rate	2.29e-10	8.41e-08	5.42e-08
k_{i_4}	*Ror*-inhibition rate	1.91e-11	3.3e-12	1.48e-10
k_{t_5}	*Bmal*-activation rate	3.83e-09	1.43e-07	1.56e-07
k_{i_5}	*Bmal*-inhibition rate	7.3e-09	2.37e-08	5.39e-09
k_{t_6}	*Clock*-activation rate	4.94e-11	1.06e-10	9.06e-09
k_{i_6}	*Clock*-inhibition rate	1.7e-08	5.49e-08	4.02e-09
	Transcription fold activation (dimensionless)			

a	*Per*	5.98	22.6	3.18
d	*Cry*	1.62	79.3	1.85e+02
g	*Rev-Erb*	1.98e+02	1.2	25.3
h	*Ror*	28.0	1.84e+02	1.42
i	*Bmal*	12	12	12
j	*Clock*	3.49	24.0	56.3
	Production rates [molecules $\times$ mRNA$^{-1}\times$ hour^{-1}]			
k_{p_1}	PER_C^{tot}	4.61e+03	2e+04	4.28e+04
k_{p_2}	CRY_C	7.57e+04	9.49e+04	6.66e+03
k_{p_3}	REV-ERB_C	5.1e+03	1.82e+05	9.5e+03
k_{p_4}	ROR_C	6.52e+02	6.49e+02	6.37
k_{p_5}	BMAL_C	3.11e+03	9.64e+04	2.97e+04
k_{p_6}	CLOCK_C	1.23e+03	1.32e+03	5.58e+02
	Import/Export rates [hour^{-1}]			
$k_{i_{z_4}}$	PER/CRY_C^{tot}	0.0276	0.0556	0.0165
$k_{i_{z_6}}$	REV-ERB_C	0.894	0.00737	0.985
$k_{i_{z_7}}$	ROR_C	0.00104	0.00797	0.00173
$k_{i_{z_9}}$	CLOCK/BMAL_C	0.0136	0.001	0.0013
$k_{e_{x_1}}$	CLOCK/BMAL_N	0.00903	0.0217	0.335
$k_{e_{x_2}}$	PER/CRY_N^{tot}	0.00495	0.0163	0.0129
	Hill coefficients of transcription (dimensionless)			
b	activation	7.4	2.83	1.2
c	inhibition	2.61	3.44	3.01
e	*Cry*-activation	2.56	7.93	7.98
f	*Cry*-inhibition	1.29	3.57	4.49
f_1	*Cry*-inhibition	1.0	2.67	6.16
	Volume proportion (dimensionless)			
v_c	cytoplasm	0.93	0.8	0.8
v_n	nucleus	0.07	0.2	0.2
	Transcription fold activation (dimensionless)			
f_{Ppar}	*PPAR*	1.04	57.5	10.1
f_{Par}	*PAR*	50.8	1.16	1.12
f_{Ugt}	*UGT*	1.89e+02	9.62	12.5
f_{Ces}	*CES*	5.49	1.92	92.3
f_{Nfil}	*NFIL*	3.98	3.91	2.31
f_{Abcb}	*ABCB*	14.8	1.25	3.54
f_{Abcc}	*ABCC*	8.44	24.5	23.8
f_{Top}	*TOP*	1.17	1.0	1.34e+02
f_{TopPar}	*TOPPAR*	2.21	3.61	2.96
	Activation/inhibition rates [nmol$\times L^{-1}$]			
a_{Ppar}	*PPAR*	2.64e-11	6.13e-09	7.55e-12

i_{Ppar}	*PPAR*	6.62e-10	3.05e-11	1.29e-10
a_{Par}	*PAR*	3.21e-11	1.58e-12	2.4e-10
i_{Par}	*PAR*	1.19e-10	7.14e-11	5.04e-11
a_{Ugt}	*UGT*	6.58e-12	3.7e-11	7.32e-12
a_{Ces}	*CES*	4.42e-12	5.67e-07	2.17e-13
i_{Ces}	*CES*	1.07e-10	2.29e-09	6.83e-10
a_{Nfil}	*NFIL*	1.2e-07	9.73e-09	2.99e-08
i_{Nfil}	*NFIL*	1.1e-08	2.23e-08	2.71e-09
a_{Abcb}	*ABCB*	9.84e-09	6.13e-12	5.22e-08
i_{Abcb}	*ABCB*	1.23e-12	1.04e-11	1.1e-11
a_{Abcc}	*ABCC*	1.69e-09	9.63e-12	8.72e-08
i_{Abcc}	*ABCC*	2.3e-11	1.07e-09	1.52e-11
i_{BmalTop}	*BMAL*	3e-07	6.4e-06	3.77e-06
a_{Top}	*TOP*	2.08e-10	4.09e-10	1.75e-11
i_{Top}	*TOP*	1.11e-10	4.14e-11	7.52e-10
a_{TopPar}	*TOP*	7.56e-12	1.1e-08	4.04e-10
i_{TopNfil}	*TOP*	6.68e-10	1.04e-10	5.06e-10
i_{RevNfil}	*REVNFIL*	7.76e-06	7.02e-07	1.41e-07
	Transcription rates [nmol$\times L^{-1}\times$ hours^{-1}]			
V_{Ppar}	*PPAR*	2.99e-12	4.09e-12	4.86e-10
V_{Par}	*PAR*	1.44e-11	4.58e-09	1.3e-10
V_{Ugt}	*UGT*	1.02e-12	1.49e-12	1.06e-14
V_{Ces}	*CES*	1.11e-11	8.38e-09	4.67e-11
V_{Nfil}	*NFIL*	4.06e-11	5.95e-13	8.33e-12
V_{Abcb}	*ABCB*	1.54e-13	3.02e-11	1.55e-12
V_{Abcc}	*ABCC*	7.55e-11	6.74e-12	1.33e-09
V_{Top}	*TOP*	3.94e-13	3.72e-10	6.02e-12
	Production rates [molecules $\times$ mRNA$^{-1}\times$ hour^{-1}]			
r_{PPAR}	PPAR	1.19e+04	5.39e+03	5.06e+03
r_{PAR}	PAR	7.62e+03	1.25e+03	5.12e+03
r_{NFIL}	NFIL	1.56e+04	2.36e+03	4.39e+02
r_{TOP}	TOP	5.93e+03	2.13e+04	2.51e+03
	Degradation rates [hour^{-1}]			
d_{Ppar}	*PPAR*	0.0587	0.306	0.152
d_{PPAR}	PPAR	0.0934	0.219	0.915
d_{Par}	*PAR*	0.331	0.161	0.0694
d_{PAR}	PAR	0.872	1.59	1.77
d_{Ugt}	*UGT*	2.98	1.35	2.83
d_{Ces}	*CES*	0.0776	0.0503	2.14
d_{Nfil}	*NFIL*	2.82	0.0353	0.032
d_{NFIL}	NFIL	0.202	2.33	2.31

d_{Abcb}	*ABCB*	0.0132	6.02	0.131
d_{Abcc}	*ABCC*	1.49	0.398	3.18
d_{Top}	*TOP*	0.083	0.0314	0.177
d_{TOP}	TOP	2.94	1.12	2.23
	Post-transcriptional modification parameters [hour^{-1}]			
d_{CesStar}	*CESSTAR*	0.0618	0.112	7.76
$d_{\text{ToCesStar}}$	*TOCESSTAR*	2.92	0.153	2.95

Table 2.S6: List of parameters for the clock-irinotecan model.

S2-4 Additional figures to the clock-irinotecan model

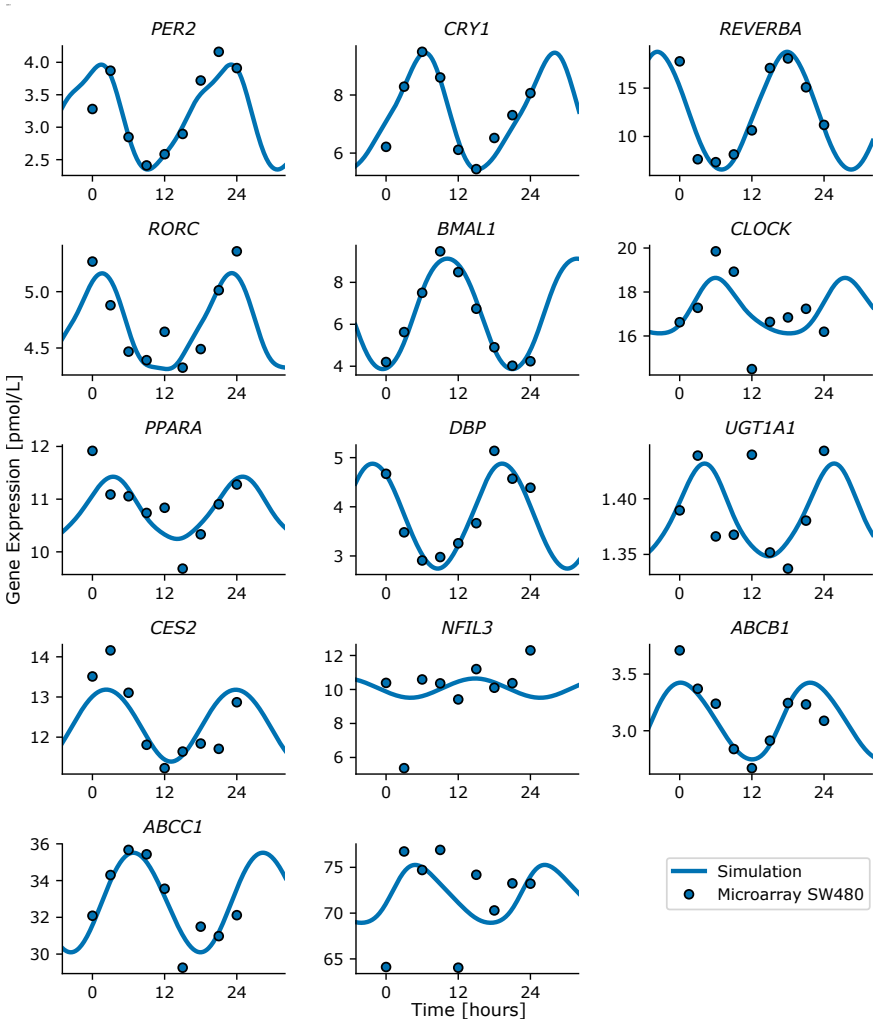


Figure 2.S4: Example of a full model fit. Model fit (line) of the mRNA expression data (dots) for the SW480 cell line.

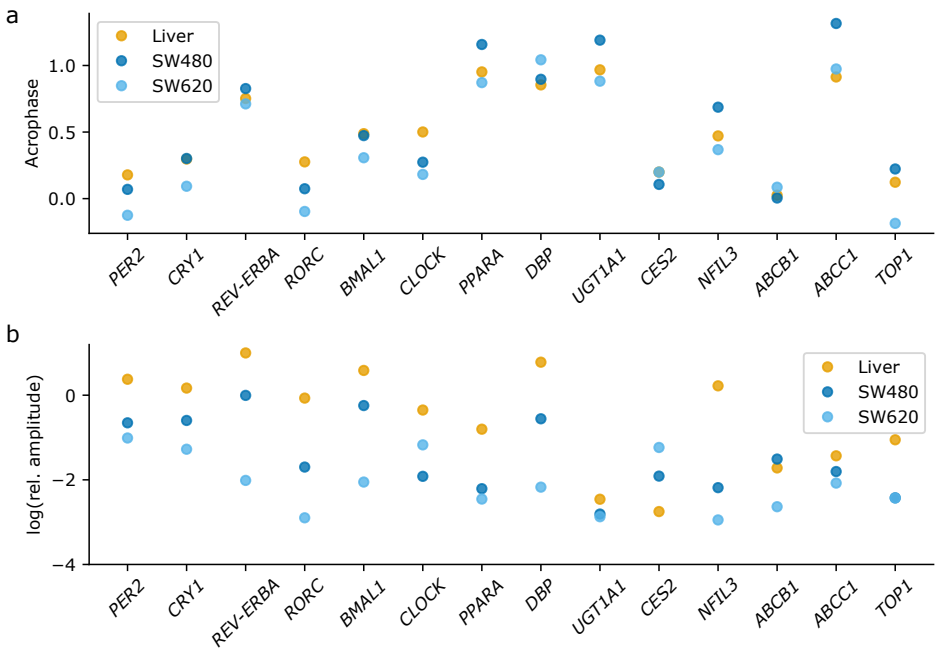


Figure 2.S5: Oscillation comparison for different model fits. Comparison between liver, SW480 and SW620 regarding **(a)** acrophase (rescaled to the interval from zero to one, values below zero and above one are for visualization and represent the phase modulo one) and **(b)** relative amplitude.

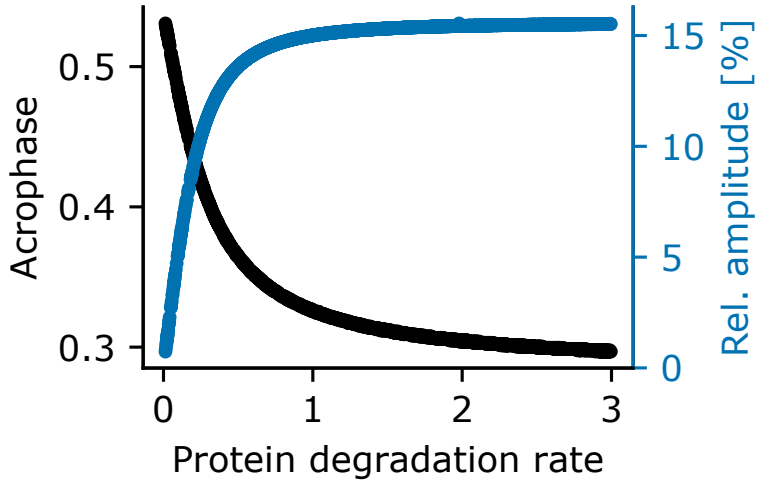


Figure 2.S6: Translation parameters of the irinotecan-relevant proteins CES2, UGTA1A, ABCB and ABCC. Scanning the range of possible translation parameters shows that only the protein degradation rate changes acrophase and relative amplitude significantly. Compromising between a large oscillation amplitude and a phase delay of at least around 4 hours between mRNA and protein peak, we choose a degradation rate of 1.2.

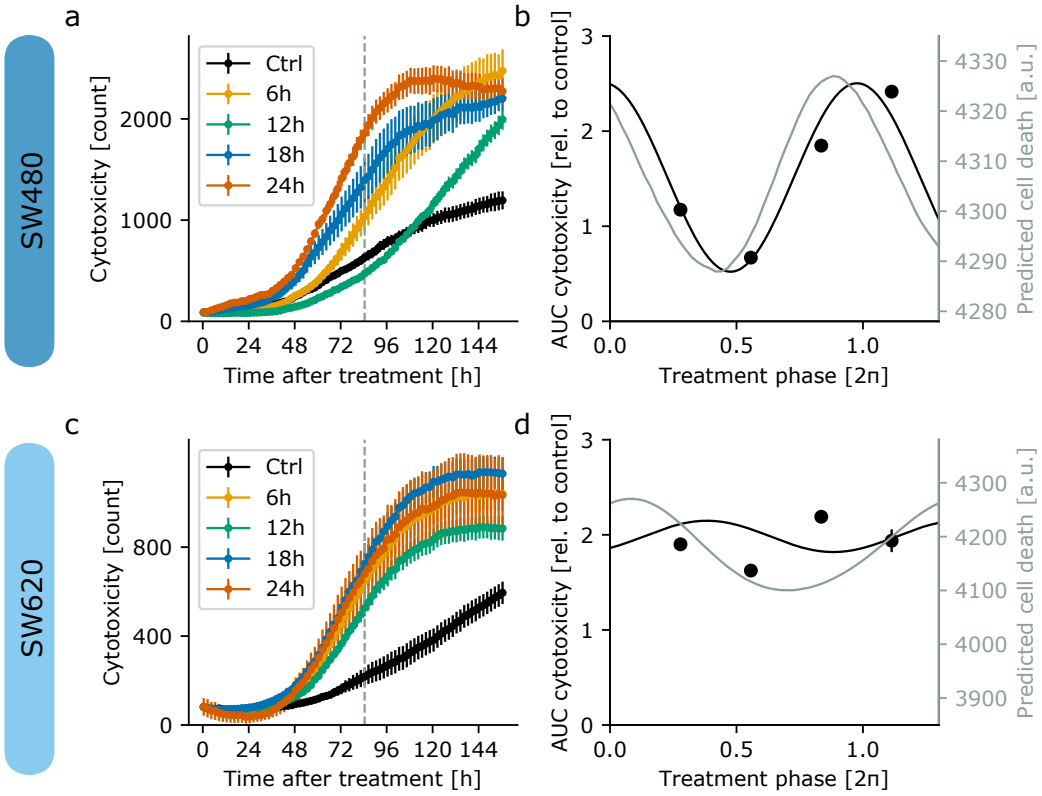


Figure 2.S7: Predictions of the original PK-PD model. **a** Cytotoxicity measured as the number of dead cells, counted as number of red objects in the experimental setup, for the full duration of the experiment. For the analysis, the time till 84.5 hours (grey dashed line) is used to prevent saturation effects. **b** Area under the curve of the experimental data (till 84.5 hours) and the harmonic regression line predicted by cosinor analysis (black line) compared with the model predictions using the model from Dulong et al. (2015) with replace protein dynamics, adapted to CRC cells as described in Section S2-3.2. The predicted phase of maximal toxicity (grey line) fits for the SW480 cell line, but the predicted amplitude is too low. **c** Cytotoxicity for the SW620 cell line. **d** Area under the curve for the experimental data (till 84.5 hours) for SW620 cell line compared to model predictions. Harmonic regression showed no significant result for the SW620 cell line.

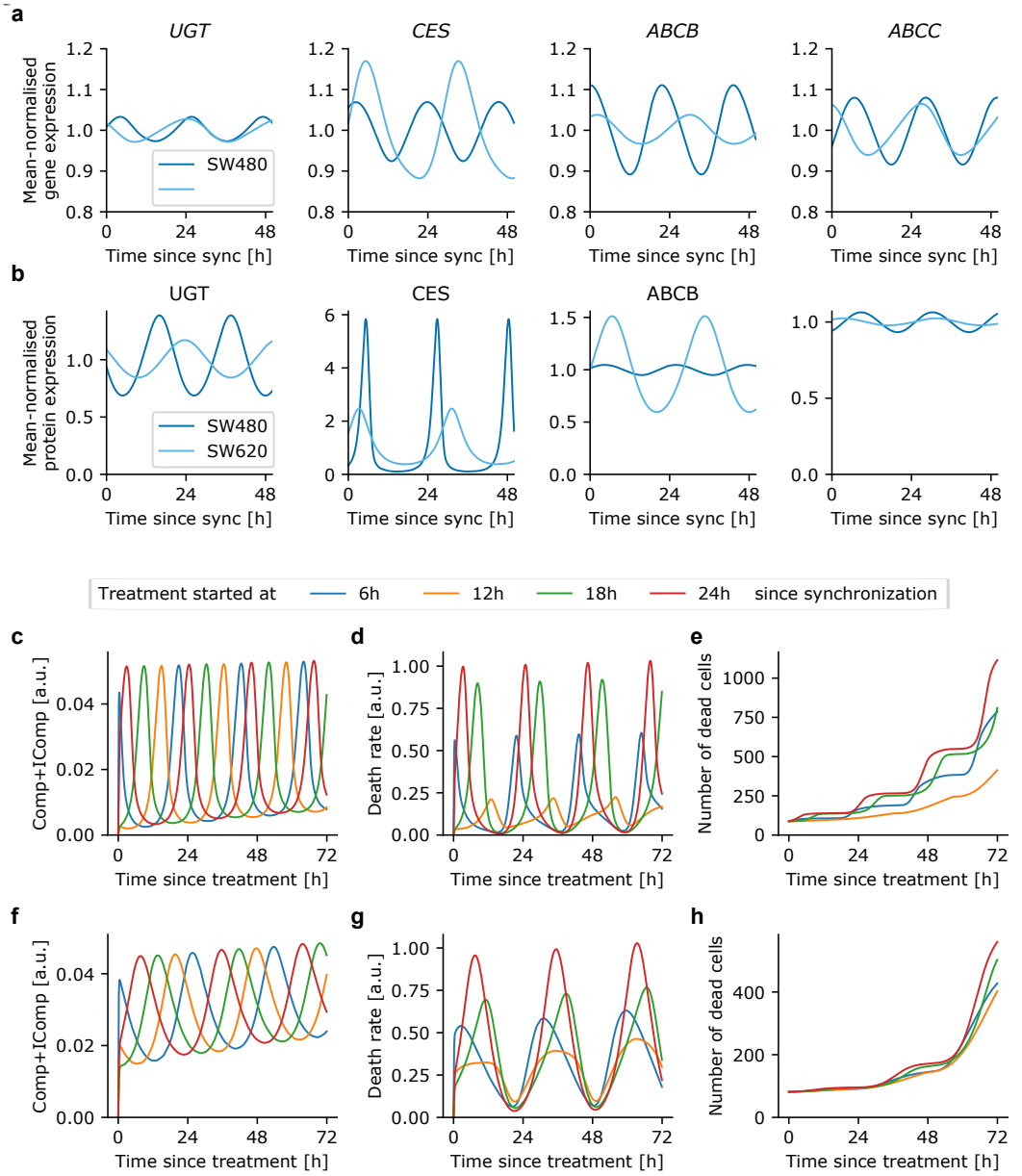


Figure 2.S8: Irinotecan-induced cytotoxicity depends on treatment time and CRC cell line. a-b Cell lines differ in the expression of irinotecan-relevant genes **(a)** and proteins **(b)**. Time is aligned to the experimental synchronization of the cells to which the model was fitted. **c-d** The action of irinotecan shows dynamics specific for the SW480 **(c)** and SW620 **(d)** cell line. Time is aligned to treatment onset. Right panel: The total irinotecan complex abundance (Compl + Icompl, see Equation (2.S54) and Equation (2.S55)) hinders successful cell division. Middle panel: The death rate resulting from irinotecan treatment (right-hand side of Equation (2.S59)) results from an interference of irinotecan complex abundance and apoptosis modulation, which results in different mean death rates depending on the treatment time. Left panel: The number of dead cells (D, see Equation (2.S59)) diverges over time depending on treatment time.

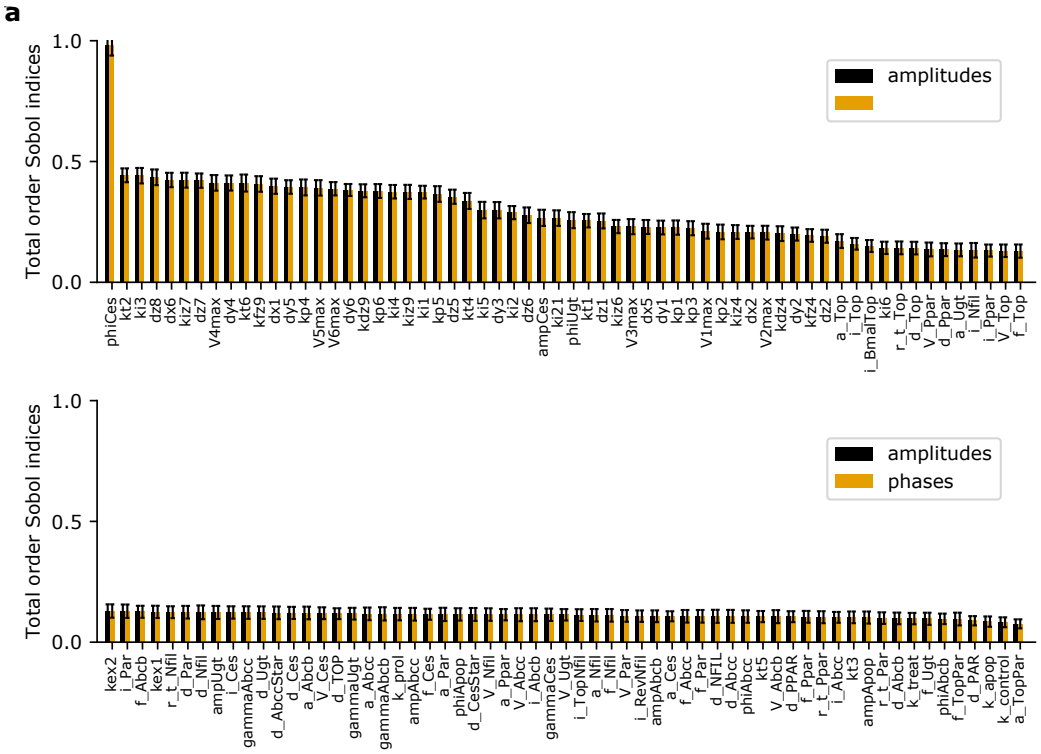


Figure 2.S9: Sensitivity analysis of the full clock-PK-PD model. Total order Sobol indices estimated for the clock-irinotecan model with respect to phase and amplitude of irinotecan circadian toxicity profile, for details see Section S2-1.4. Transcription fold changes and Hill coefficient were excluded to reduce the number of parameters. For the circadian degradation associated with the dynamical variables CES, UGT, ABCB and ABCC, phi plus variable name denotes the phase, amp plus variable name denotes the amplitude, and gamma plus variable name denotes the mesor of the degradation oscillation. 10 000 simulations of the model were used for the estimation. **a** Parameters with high sensitivity (more sensitive half). **b** Parameters with low sensitivity (less sensitive half).

Bibliography

Aalto, A., Viitasaari, L., Ilmonen, P., Mombaerts, L., and Gonçalves, J. (2020). Gene regulatory network inference from sparsely sampled noisy data. *Nature Communications*, 11(1).

Abdulrehman, G., Xv, K., Li, Y., and Kang, L. (2018). Effects of meta-tetrahydroxyphenylchlorin photodynamic therapy on isogenic colorectal cancer sw480 and sw620 cells with different metastatic potentials. *Lasers in Medical Science*, 33:1581–1590.

Ahowesso, C., Li, X.-M., Zampera, S., Peteri-Brunbäck, B., Dulong, S., Beau, J., Hossard, V., Filipski, E., Delaunay, F., Claustrat, B., and Lévi, F. (2011). Sex and dosing-time dependencies in irinotecan-induced circadian disruption. *Chronobiology International*, 28(5):458–470.

Akashi, M., Soma, H., Yamamoto, T., Tsugitomi, A., Yamashita, S., Yamamoto, T., Nishida, E., Yasuda, A., Liao, J. K., and Node, K. (2010). Noninvasive method for assessing the human circadian clock using hair follicle cells. *Proceedings of the National Academy of Sciences of the United States of America*, 107(35):15643–15648.

Ali, M. A. and Kravitz, A. V. (2018). Challenges in quantifying food intake in rodents. *Brain Research*, pages 188–191.

Almeida, S., Chaves, M., and Delaunay, F. (2020). Transcription-based circadian mechanism controls the duration of molecular clock states in response to signaling inputs. *Journal of Theoretical Biology*, 484.

Anderson, S. T. and Fitzgerald, G. A. (2020). Sexual dimorphism in body clocks. *Science*, 369(6508):1164–1165.

Aryal, R. P., Kwak, P. B., Tamayo, A. G., Gebert, M., Po-Lin-chiu, Walz, T., and Weitz, C. J. (2017). Macromolecular assemblies of the mammalian circadian clock. *Molecular Cell*, 67(5):770–782.

Atwood, A., DeConde, R., Wang, S. S., Mockler, T. C., Sabir, J. S. M., Ideker, T., and Kay, S. A. (2011). Cell-autonomous circadian clock of hepatocytes drives rhythms in transcription and polyamine synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(45):18560–18565.

Aubin-Frankowski, P.-C. and Vert, J.-P. (2020). Gene regulation inference from single-cell rna-seq data with linear differential equations and velocity inference. *Bioinformatics*, 36(18):4774–4780.

Ballesta, A., Dulong, S., Abbara, C., Cohen, B., Okyar, A., Clairambault, J., and Levi, F. (2011). A combined experimental and mathematical approach for molecular-based optimization of irinotecan circadian delivery. *Plos Computational Biology*, 7(9).

Ballesta, A., Innominato, P. F., Dallmann, R., Rand, D. A., and Lévi, F. A. (2017). Systems chronotherapeutics. *Pharmacological Reviews*, 69(2):161–199.

Bass, J. and Takahashi, J. S. (2010). Circadian integration of metabolism and energetics. *Science*, 330(6009):1349–1354.

Basti, A., Yalçin, M., Herms, D., Hesse, J., Aboumanify, O., Li, Y., Aretz, Z., Garmshausen, J., El-Athman, R., Hastermann, M., Blottner, D., and Relógio, A. (2021). Diurnal variations in the expression of core-clock genes correlate with resting muscle properties and predict fluctuations in exercise performance across the day. *BMJ Open Sport & Exercise Medicine*, 7.

Bates, S. (2010). Progress towards personalized medicine. *Drug Discovery Today*, 15(3):115–120.

Becker-Weimann, S., Wolf, J., Herzel, H., and Kramer, A. (2004). Modeling feedback loops of the mammalian circadian oscillator. *Biophysical Journal*, 87(5):3023–3034.

Bellu, G., Saccomani, M. P., Audoly, S., Leontina, and D'Angiò (2007). Daisy: A new software tool to test global identifiability of biological and physiological systems. *Computer Methods and Programs in Biomedicine*, 88(1):52–61.

Benhamou, E., Saltiel, D., Laraki, R., and Atif, J. (2020). BCMA-ES: a conjugate prior Bayesian optimization view. working paper or preprint.

Bicker, J., Alves, G., Falcão, A., and Fortuna, A. (2020). Timing in drug absorption and disposition: The past, present, and future of chronopharmacokinetics. *British Journal of Pharmacology*, 177(10):2215–2239.

Bollinger, T. and Schibler, U. (2014). Circadian rhythms - from genes to physiology and disease. *Swiss Medical Weekly*.

Boughattas, N. A., Lévi, F., Fournier, C., Lemaigre, G., Roulon, A., Hecquet, B., Mathé, G., and Reinberg, A. (1989). Circadian rhythm in toxicities and tissue uptake of 1,2-diamminocyclohexane(*trans*-1)oxalatoplatinum(ii) in mice. *Cancer Research*, 49(12):3362–3368.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, pages 3932–3937.

Cabeli, V., Verny, L., Sella, N., Uguzzoni, G., Verny, M., and Isambert, H. (2020). Learning clinical networks from medical records based on information estimates in mixed-type data. *Plos Computational Biology*.

Carcano, A., Fages, F., and Soliman, S. (2017). Probably Approximately Correct Learning of Regulatory Networks from Time-Series Data. In *CMSB'17: Proceedings of the fifteenth international conference on Computational Methods in Systems Biology*, volume 10545, pages 74–90.

CasymConsortium, C. (2014). The casym roadmap implementation of systems medicine across europe.

Champion, K., Zheng, P., Aravkin, A. Y., Brunton, S. L., and Kutz, J. N. (2020). A unified sparse optimization framework to learn parsimonious physics-informed models from data.

Chan, T. E., P.H.Stumpf, M., and C.Babtie, A. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems*, 5:251–267.

Chang, A.-M., Dufy, J. F., Buxton, O. M., Lane, J. M., Aeschbach, D., Anderson, C., Bjonnes, A. C., Cain, S. W., Cohen, D. A., Frayling, T. M., Gooley, J. J., Jones, S. E., Klerman, E. B., Lockley, S. W., Munch, M., Rajaratnam, S. M. W., Rueger, M., Rutter, M. K., Santhi, N., Scheuermaier, K., Reen, E. V., N.Weedon, M., Czeisler, C. A., Scheer, F. A. J. L., and Saxena, R. (2019). Chronotype genetic variant in per2 is associated with intrinsic circadian period in humans. *Scientific Reports*, 9(1).

Chen, Z., Odstrcil, E. A., Tu, B. P., and McKnight, S. L. (2007). Restriction of dna replication to the reductive phase of the metabolic cycle protects genome integrity. *Science*, 316(5833):1916–1919.

Ciotti, M., Basu, N., Brangi, M., and S.Owens, I. (1999). Glucuronidation of 7-ethyl-10-hydroxycamptothecin (sn-38) by the human udp-glucuronosyltransferases encoded at the *UGT1* locus. *Biochemical and Biophysical Research Communications*, 260(1):199–202.

Dallmann, R., Okyar, A., and Lévi, F. (2016). Dosing-time makes the poison: Circadian regulation and pharmacotherapy. *Trends in Molecular Medicine*, 22.

de Man, F. M., Goey, A. K. L., van Schaik, R. H. N., Mathijssen, R. H. J., and Bins, S. (2018). Individualization of irinotecan treatment: A review of pharmacokinetics, pharmacodynamics, and pharmacogenetics. *Clinical Pharmacokinetics*, 57:1229–1254.

Degrand, E., Hemery, M., and Fages, F. (2019). On chemical reaction network design by a nested evolution algorithm. In *CMSB'19: Proceedings of the seventeenth international conference on Computational Methods in Systems Biology*, Lecture Notes in BioInformatics. Springer-Verlag.

Disease, C., Sciences, B., Studies, D., and Council, N. (2012). *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*.

Dony, L., He, F., and Stumpf, M. (2019). Parametric and non-parametric gradient matching for network inference: A comparison. *BMC Bioinformatics*, 20.

Dulong, S., Ballesta, A., Okyar, A., and Lévi, F. (2015). Identification of circadian determinants of cancer chronotherapy through in vitro chronopharmacology and mathematical modeling. *Molecular Cancer Therapeutics*, 14(9):2154–2164.

El-Athman, R., Fuhr, L., and Relógio, A. (2018). A systems-level analysis reveals circadian regulation of splicing in colorectal cancer. *EBioMedicine*, 33:68–81.

El-Athman, R., Knezevic, D., Fuhr, L., and Relógio, A. (2019). A computational analysis of alternative splicing across mammalian tissues reveals circadian and ultradian rhythms in splicing events. *International Journal of Molecular Sciences*, 20(16).

El-Athman, R. and Relógio, A. (2018). Escaping circadian regulation: An emerging hallmark of cancer? *Cell Systems*, 6(3):266–267.

Fages, F., Gay, S., and Soliman, S. (2015). Inferring reaction systems from ordinary differential equations. *Theoretical Computer Science*, 599:64–78.

Feillet, C., Guérin, S., Lonchampt, M., Dacquet, C., Åke Gustafsson, J., Delaunay, F., and Teboul, M. (2016). Sexual dimorphism in circadian physiology is altered in LXRα deficient mice. *PLoS One*, 11(3).

Feillet, C., van der Horst, G. T. J., Levi, F., Rand, D. A., and Delaunay, F. (2015). Coupling between the circadian clock and cell cycle oscillators: Implication for healthy cells and malignant growth. *Frontiers in Neurology*, 6(May):1–7.

Fletcher, J. I., Williams, R. T., Henderson, M. J., Norris, M. D., and Haber, M. (2016). Abc transporters as mediators of drug resistance and contributors to cancer cell biology. *Drug Resistance Updates*, 26:1–9.

Forger, D. B. and Peskin, C. S. (2003). A detailed predictive model of the mammalian circadian clock. *Proceedings of the National Academy of Science of the United States of America*, 100(25):14806–14811.

Frazier, P. I. (2018). A tutorial on bayesian optimization.

Fuhr, L., El-Athman, R., Scrima, R., Cela, O., Carbone, A., Knoop, H., Li, Y., Hoffmann, K., Laukkanen, M. O., Corcione, F., Steuer, R., F.Meyer, T., Mazzoccoli, G., Capitanio, N., and Relógio, A. (2018). The circadian clock regulates metabolic phenotype rewiring via hkdc1 and modulates tumor progression and drug response in colorectal cancer. *EBioMedicine*, 33:105–121.

Fujii, H., Yamada, Y., Watanabe, D., Matsuhashi, N., Takahashi, T., Yoshida, K., and Suzuki, A. (2019). Dose adjustment of irinotecan based on ugt1a1 polymorphisms in patients with colorectal cancer. *Cancer Chemotherapy and Pharmacology*, 83:123–129.

Furlan, A., Jacquier, M., Woller, A., Héliot, L., Duez, H., Staels, B., and Lefranc, M. (2019). Mathematical models converge on pgc1α as the key metabolic integrator of sirt1 and ampk regulation of the circadian clock. *Proceedings of the National Academy of Science of the United States of America*, 116(27):13171–13172.

Gascoyne, D. M., Long, E., Veiga-Fernandes, H., de Boer, J., Williams, O., Seddon, B., Coles, M., Kioussis, D., and Brady, H. J. M. (2009). The basic leucine zipper transcription factor e4bp4 is essential for natural killer cell development. *Nature Immunology*, 10:1118–1124.

Gaspar, L. S., Álvaro, A. R., Carmo-Silva, S., Mendes, A. F., Relógio, A., and Cavadas, C. (2019). The importance of determining circadian parameters in pharmacological studies. *British Journal of Pharmacology*, 176(16):2827–2847.

Gaucher, J., Montellier, E., and Sassone-Corsi, P. (2018). Molecular cogs: Interplay between circadian clock and cell cycle. *Trends in Cell Biology*, 28(5):368–379.

Giacchetti, S., P.A.Dugué, Innominato, P., Bjarnason, G., Focan, C., Garufi, C., Coudert, S. T. B., Iacobelli, S., Smaaland, R., Tampellini, M., Adam, R., Moreau, T., and Lévi, F. (2012). Sex moderates circadian chemotherapy effects on survival of patients with metastatic colorectal cancer: a meta-analysis. *Annals of oncology*, 23(12):3110–3116.

Gil-Martín, E., Egea, J., Reiter, R. J., and Romero, A. (2019). The emergence of melatonin in oncology: Focus on colorectal cancer. *Medical Research Reviews*, 39(6):2239–2285.

Ginsburg, G. and Phillips, K. (2018). Precision medicine: From science to value. *Health Affairs*, 37:694–701.

Girotti, M., Gremel, G., Lee, R., Galvani, E., Rothwell, D., Viros, A., Mandal, A., Lim, K. H. J., Saturno, G., Furney, S., Baenke, F., Pedersen, M., Rogan, J., Swan, J., Smith, M., Fusi, A., Oudit, D., Dhomen, N., Brady, G., and Marais, R. (2016). Application of sequencing, liquid biopsies, and patient-derived xenografts for personalized medicine in melanoma. *Cancer discovery*, 6.

Gonze, D. and Abou-Jaoudé, W. (2013). The goodwin model: Behind the hill function. *PLoS One*, 8(8).

Gotoh, T., Kim, J. K., Liu, J., Vila-Caballer, M., Stauffer, P. E., Tyson, J. J., and Finkielstein, C. V. (2016). Model-driven experimental approach reveals the complex regulatory distribution of p53 by the circadian factor period 2. *Proceedings of the National Academy of Science of United States of America*, 113(47):13516–13521.

Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., and Maire, P. (2009). The hill equation: A review of its capabilities in pharmacological modelling. *Fundamental & clinical pharmacology*, 22:633–48.

Greenwell, B. J., Trott, A. J., Beytebiere, J. R., Pao, S., Bosley, A., Beach, E., Finegan, P., Hernandez, C., and Mene, J. S. (2019). Rhythmic food intake drives rhythmic gene expression more potently than the hepatic circadian clock in mice. *Cell Reports*, 27(3):649–657.

Guillaumond, F., Dardente, H., Giguère, V., and Cermakian, N. (2005). Differential control of Bmal1 circadian transcription by REV-ERB and ROR nuclear receptors. *Journal of Biological Rhythms*, 20(5):391–403.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *Public Library of Science Computational Biology*, 3(10):1871–1878.

Hang, N., Chelliah, Y., Shan, Y., andSeung Hee Yoo, C. A. T., Partch, C., Green, C. B., Zhang, H., and Takahashi, J. S. (2012). Crystal structure of the heterodimeric clock:bmal1 transcriptional activator complex. *Science*, 337(6091):189–194.

Hansen, N., Auger, A., Ros, R., Finck, S., and Pošík, P. (2010). Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009. In *Proceedings of the 12th annual conference comp on Genetic and evolutionary computation - GECCO '10*, page 1689, New York, New York, USA. ACM Press.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195.

Hatori, M., Vollmers, C., Zarrinpar, A., DiTacchio, L., Bushong, E. A., Gill, S., Leblanc, M., Chaix, A., Joens, M., Fitzpatrick, J. A., H.Ellisman, M., and Panda, S. (2012). Time-restricted feeding without reducing caloric intake prevents metabolic diseases in mice fed a high-fat diet. *Cell Metabolism*, 15(6):848–860.

Hesse, J., Malhan, D., Yalçin, M., Aboumanify, O., Basti, A., and Relógio, A. (2020). An optimal time for treatment—predicting circadian time by machine learning and mathematical modelling. *Cancers*, 12(11).

Hesse, J., Martinelli, J., Aboumanify, O., Ballesta, A., and Relógio, A. (2021). A mathematical model of the circadian clock and drug pharmacology to optimize irinotecan administration timing in colorectal cancer. *Computational and Structural biology*, 19:5170–5183.

Hill, R. J. W., Innominato, P. F., Lévi, F., and Ballesta, A. (2020). Optimizing circadian drug infusion schedules towards personalized cancer chronotherapy. *Plos Computational Biology*, 16(1).

Huang, C.-Y. and Ferrell, J. E. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *PNAS*, 93(19):10078–10083.

Huet, S. and Taupin, M.-L. (2017). Metamodel construction for sensitivity analysis. *ESAIM: Proceedings*, Volume 60, 2017(60):27–69.

Huynh-Thu, V. A. and Geurts, P. (2018). dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Scientific Reports*, 8.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9).

Ikegami, K., Refetoff, S., Cauter, E. V., and Yoshimura, T. (2019). Interconnection between circadian clocks and thyroid function. *Nature Reviews Endocrinology*, 15:590–600.

Innominato, P., Komarzynski, S., Karaboué, A., Ulusakarya, A., Bouchahda, M., Haydar, M., Bossevot-Desmaris, R., Mocquery, M., Plessis, V., and Lévi, F. (2018). Home-based e-health platform for multidimensional telemonitoring of symptoms, body weight, sleep, and circadian activity: Relevance for chronomodulated administration of irinotecan, fluorouracil-leucovorin, and oxaliplatin at home-results from a pilot study. *Journal of Clinical Oncology Clinical Cancer Informatics*, (2):1–15.

Innominato, P. F., Ballesta, A., Huang, Q., Focan, C., Chollet, P., Karaboué, A., Giacchetti, S., Bouchahda, M., Adam, R., Garufi, C., and Lévi, F. A. (2020a). Sex-dependent least toxic timing of irinotecan combined with chronomodulated chemotherapy for metastatic colorectal cancer: Randomized multicenter eortc 05011 trial. *Cancer Medicine*.

Innominato, P. F., Karaboué, A., Focan, C., Chollet, P., Giacchetti, S., Bouchahda, M., Ulusakarya, A., Torsello, A., Adam, R., Lévi, F. A., and Garufi, C. (2020b). Efficacy and safety of chronomodulated irinotecan, oxaliplatin, 5-fluorouracil and leucovorin combination as first- or second-line treatment against metastatic colorectal cancer: Results from the international eortc 05011 trial. *International Journal of Cancer*, 148(10):2512–2521.

Innominato, P. F., Roche, V. P., Palesh, O. G., Ulusakarya, A., Spiegel, D., and Lévi, F. A. (2014). The circadian timing system in clinical oncology. *Annals of Medicine*, 46(4).

Jones, D., Schonlau, M., and Welch, W. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492.

Keener, J. and Sneyd, J. (2009). *Mathematical Physiology: I: Cellular Physiology*. Interdisciplinary Applied Mathematics №8/1. Springer-Verlag New York, 2 edition.

Kim, D. W., Zavala, E., and Kim, J. K. (2020). Wearable technology and systems modeling for personalized chronotherapy. *Current Opinion in Systems Biology*, 21:9–15.

Kim, J. K. and Forger, D. B. (2012). A mechanism for robust circadian timekeeping via stoichiometric balance. *Molecular Systems Biology*, 8(1).

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B., and Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252.

Ko, C. H. and Takahashi, J. S. (2006). Molecular components of the mammalian circadian clock. *Human Molecular Genetics*, 15(2):271–277.

Komarzynski, S., Bolborea, M., Huang, Q., Finkenstädt, B., and Lévi, F. (2019). Predictability of individual circadian phase during daily routine for medical applications of circadian clocks. *Journal of Clinical Investigation Insight*, 4(18).

Komarzynski, S., Huang, Q., Innominato, P. F., Maurice, M., Arbaud, A., Beau, J., Bouchahda, M., Ulusakarya, A., Beaumatin, N., Breda, G., Finkenstädt, B., and Lévi, F. (2018). Relevance of a mobile internet platform for capturing inter- and intrasubject variabilities in circadian coordination during daily routine: Pilot study. *Journal of Medical Internet Research*, 20(6):924–930.

Kondratov, R. V., Chernov, M. V., Kondratova, A. A., Gorbacheva, V. Y., Gudkov, A. V., and Antoch, M. P. (2003). Bmal1-dependent circadian oscillation of nuclear clock: posttranslational events induced by dimerization of transcriptional activators of the mammalian clock system. *Genes and Development*, 17(15):1921–1932.

Korenčič, A., Bordyugov, G., Košir, R., Rozman, D., Goličnik, M., and Herzel, H. (2012). The interplay of cis-regulatory elements rules circadian rhythms in mouse liver. *PLoS One*, 7(11).

Korf, R. E. (1996). Improved limited discrepancy search. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI'96, page 286–291.

Kornmann, B., Schaad, O., Bujard, H., Takahashi, J. S., and Schibler, U. (2007). System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock. *Public Library of Science Biology*, 5(2):180–189.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*.

Köksal, A., Beck, K., Cronin, D., McKenna, A., Camp, N., Srivastava, S., MacGilvray, M., Bodik, R., Wolf-Yadlin, A., Fraenkel, E., Fisher, J., and Gitter, A. (2018). Synthesizing signaling pathways from temporal phosphoproteomic data. *Cell reports*, 24:3607–3618.

Le Gratiet, L., Marelli, S., and Sudret, B. (2016). Metamodel-based sensitivity analysis: polynomial chaos expansions and Gaussian processes. In *Handbook of Uncertainty Quantification - Part III: Sensitivity analysis*.

Lee, C., Etchegaray, J.-P., Cagampang, F. R., Loudon, A. S., and Reppert, S. M. (2001). Posttranslational mechanisms regulate the mammalian circadian clock. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):855–867.

Leloup, J.-C. and Goldbeter, A. (2003). Toward a detailed computational model for the mammalian circadian clock. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12):7501–7506.

Li, X.-M., Delaunay, F., Dulong, S., Claustrat, B., Zampera, S., Fujii, Y., Teboul, M., Beau, J., and Lévi, F. (2010). Cancer inhibition through circadian reprogramming of tumor transcriptome with meal timing. *Cancer Research*, 70(8):3351–3360.

Li, X.-M., Liu, X.-H., Filipski, E., Metzger, G., Delagrange, P., Jeanniot, J.-P., and Lévi, F. (2000). Relation of atypical melatonin rhythm with two circadian clock outputs in B6D2F1 mice. *Americal Journal of Physiology Regulatory, Integrative and Comparative Physiology*, 278(4):924–930.

Li, X.-M., Mohammad-Djafari, A., Dumitru, M., Dulong, S., Filipski, E., Siffroi-Fernandez, S., Mteyrek, A., Scaglione, F., Guettier, C., Delaunay, F., and Lévi, F. (2013). A circadian clock transcription model for the personalization of cancer chronotherapy. *Cancer Research*, 73(24):7176–7188.

Liu, A. C., Tran, H. G., Zhang, E. E., Priest, A. A., Welsh, D. K., and Kay, S. A. (2008). Redundant function of REV-ERBα and β and non-essential role for Bmal1 cycling in transcriptional regulation of intracellular circadian rhythms. *PLoS Genet*, 4(2):e1000023–e1000023.

Livak, K. J. and Schmittgen, D. (2001). Analysis of relative gene expression data using real-time quantitative pcr and the $2^{-\Delta\Delta C_T}$ method. *Methods*, 25(4):402–408.

Lorenz, L. J., Hall, J. C., and Rosbash, M. (1989). Expression of a drosophila mrna is under circadian clock control during pupation. *Development.*, 107(4):869–880.

Lowrey, P. L. and Takahashi, J. S. (2004). Mammalian circadian biology: Elucidating genome-wide levels of temporal organization. *Annual Review of Genomics and Human Genetics*, 5:407–441.

Lozano, A. C. and Swirszcz, G. (2012). Multi-level lasso for sparse multi-task regression. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML'12, page 595–602, Madison, WI, USA. Omnipress.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc.

Lévi, F., Karaboué, A., Gorden, L., Innominato, P. F., Saffroy, R., Giacchetti, S., Hauteville, D., Guettier, C., Adam, R., and Bouchahda, M. (2011). Cetuximab and circadian chronomodulated chemotherapy as salvage treatment for metastatic colorectal cancer (mcrc): safety, efficacy and improved secondary surgical resectability. *Cancer Chemotherapy and Pharmacology*, 67:339–348.

Lévi, F., Okyar, A., Dulong, S., Innominato, P. F., and Clairambault, J. (2010). Circadian timing in cancer treatments. *Annual Reviews of Pharmacology and Toxicology*, 50:377–421.

Lück, S., Thurley, K., Thaben, P. F., and Westermark, P. O. (2014). Rhythmic degradation explains and unifies circadian transcriptome and proteome data. *Cell Reports*, 9:741–751.

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., and Rocha, I. (2011). Modeling formalisms in systems biology. *AMB Express*, 1(45).

Marbach, D. (2009). *Evolutionary Reverse Engineering of Gene Networks*. PhD thesis, École Polytechnique Fédérale de Lausanne.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7.

Martinelli, J., Dulong, S., Li, X.-M., Teboul, M., Soliman, S., Lévi, F., Fages, F., and Ballesta, A. (2021). Model learning to identify systemic regulators of the peripheral circadian clock. *Bioinformatics*, 37(1):i401–i409.

Martinelli, J., Grignard, J., Soliman, S., , and Fages, F. (2019a). On inferring reactions from data time series by a statistical learning greedy heuristics. In *CMSB'19: Proceedings of the seventeenth international conference on Computational Methods in Systems Biology*, Lecture Notes in BioInformatics. Springer-Verlag.

Martinelli, J., Grignard, J., Soliman, S., and Fages, F. (2019b). A statistical unsupervised learning algorithm for inferring reaction networks from time series data. In *ICML Workshop on Computational Biology*, Long Beach, CA, USA.

Mathijssen, R. H., van Alphen, R. J., Verweij, J., Loos, W. J., Nooter, K., Stoter, G., and Sparreboom, A. (2001). Clinical pharmacokinetics and metabolism of irinotecan (cpt-11). *Clinical Cancer Research*, 7(8):2182–2194.

Mathijssen, R. H. J., Marsh, S., Karlsson, M. O., Xie, R., Baker, S. D., Verweij, J., Sparreboom, A., and McLeod, H. L. (2003). Irinotecan pathway genotype analysis to predict pharmacokinetics. *Clinical Cancer Research*, 9(9):3246–3253.

Matsuo, T., Yamaguchi, S., Mitsui, S., Emi, A., Shimoda, F., and Okamura, H. (2003). Control mechanism of the circadian clock for timing of cell division in vivo. *Science*, 302(5643):255–259.

Mekbib, T., Suen, T.-C., Rollins-Hairston, A., Smith, K., Armstrong, A., Gray, C., Owino, S., Baba, K., Baggs, J. E., Ehlen, J. C., Tosini, G., and DeBruyne, J. P. (2020). A novel female-specific circadian clock mechanism regulating metabolism. bioRxiv.

Mirsky, H. P., Liu, A. C., Welsh, D. K., Kay, S. A., and III, F. J. D. (2009). A model of the cell-autonomous mammalian circadian clock. *Proceedings of the National Academy of Sciences of the United States of America*, 106(27):11107–11112.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174.

Murakami, Y., Higashi, Y., Matsunaga, N., Koyanagi, S., and Ohdo, S. (2008). Circadian Clock-Controlled Intestinal Expression of the Multidrug-Resistance Gene mdr1a in Mice. *Gastroenterology*, 135(5):1636–1644.e3.

Mure, L., Le, H., Benegiamo, G., Chang, M., Rios, L., Jillani, N., Ngotho, M., Kariuki, T., Dkhissi-Benyahya, O., Cooper, H., and Panda, S. (2018). Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science*, 359:eaao0318.

Nagoshi, E., Saini, C., Bauer, C., Laroche, T., Naef, F., and Schibler, U. (2004). Circadian gene expression in individual fibroblasts: cell-autonomous and self-sustained oscillators pass time to daughter cells. *Cell*, 119:693–705.

Narumi, R., Shimizu, Y., Ukai-Tadenuma, M., Ode, K. L., Kanda, G. N., Shinohara, Y., Sato, A., Matsumoto, K., , and Ueda, H. R. (2016). Mass spectrometry-based absolute quantification reveals rhythmic variation of mouse circadian clock proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 113(24):3461–3467.

Neikrug, A. B., Rissling, M., Trofimenko, V., Liu, L., Natarajan, L., Lawton, S., Parker, B. A., and Ancoli-Israel, S. (2012). Bright light therapy protects women from circadian rhythm desynchronization during chemotherapy for breast cancer. *Behavioral Sleep Medicine*, 10(3).

Oishi, K., Shirai, H., and Ishida, N. (2005). Clock is involved in the circadian transactivation of peroxisome-proliferator-activated receptor alpha (pparalpha) in mice. *Biochemical Journal*, 386:575–581.

Okyar, A., Kumar, S. A., Filipski, E., Piccolo, E., Ozturk, N., Xandri-Monje, H., Pala, Z., Abraham, K., de Jesus Gomes, A. R. G., Orman, M. N., Li, X.-M., Dallmann, R., Lévi, F., and Ballesta, A. (2019). Sex-, feeding-, and circadian time-dependency of p-glycoprotein expression and activity - implications for mechanistic pharmacokinetics modeling. *Scientific Reports*, 9.

Onishi, Y. and Kawano, Y. (2012). Rhythmic binding of topoisomerase i impacts on the transcription of *Bmal1* and circadian period. *Nuclear Acids Research*, 40(19):9482–9492.

Ostrowski, M., Paulevé, L., Schaub, T., Siegel, A., and Guziolowski, C. (2016). Boolean network identification from perturbation time series data combining dynamics abstraction and logic programming. *Biosystems*, 149:139–153.

Ouma, W., Pogacar, K., and Grotewold, E. (2018). Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. *PLOS Computational Biology*, 14:e1006098.

Paranjpe, D. A. and Sharma, V. K. (2005). Evolution of temporel order in living organisms. *Journal of Biological Rhythms*, 3(7).

Patke, A., Murphy, P. J., Onat, O. E., Krieger, A. C., Özçelik, T., Campbell, S. S., and Young, M. W. (2017). Mutation of the human circadian clock gene cry1 in familial delayed sleep phase disorder. *Cell*, 169(2):203–215.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peters, H. (2015). *Game Theory*. Springer.

Peters, R. (1984). Nucleo-cytoplasmic flux and intracellular mobility in single hepatocytes measured by fluorescence microphotolysis. *The EMBO Journal*, 3(8).

Phillips, A. J. K., Vidafar, P., Burns, A. C., McGlashan, E. M., Anderson, C., Rajaratnam, S. M. W., Lockley, S. W., and Cain, S. W. (2019). High sensitivity and interindividual variability in the response of the human circadian system to evening light. *Proceedings of the National Academy of Sciences of the United States of America*, 116(24):12019–12024.

Pommier, Y. (2006). Topoisomerase i inhibitors: camptothecins and beyond. *Nature Reviews Cancer*, 6:789–802.

Preitner, N., Damiola, F., Lopez-Molina, L., Zakany, J., Duboule, D., Albrecht, U., and Schibler, U. (2002). The orphan nuclear receptor rev-erbalpha controls circadian transcription within the positive limb of the mammalian circadian oscillator. *Cell*, 110(2):251–260.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929.

Relógio, A., Thomas, P., Medina-Pérez, P., Reischl, S., Bervoets, S., Gloc, E., Riemer, P., Mang-Fatehi, S., Maier, B., Schäfer, R., Leser, U., Herzel, H., Kramer, A., and Sers, C. (2014). Ras-mediated deregulation of the circadian clock in cancer. *Plos Genetics*, 10(5).

Relógio, A., Westermark, P. O., Wallach, T., Schellenberg, K., Kramer, A., and Herzel, H. (2011). Tuning the mammalian circadian clock: Robust synergy of two loops. *Public Library of Science Computational Biology*, 7(12).

Reppert, S. M. and Weaver, D. R. (2001). Molecular analysis of mammalian circadian rhythms. *Annual Review of Physiology*, 63:647–676.

Rida, P., Syed, M. I., and Aneja, R. (2019). Time will tell: Circadian clock dysregulation in triple negative breast cancer. *Frontiers in Bioscience-Scholar*, 11(1):178–192.

Ripperger, J. A., Jud, C., and Albrecht, U. (2011). The daily rhythm of mice. *FEBS Letters*, 585(10):1384–1392.

Ripperger, J. A. and Schibler, U. (2006). Rhythmic CLOCK-BMAL1 binding to multiple E-box motifs drives circadian Dbp transcription and chromatin transitions. *Nature Genetics*, 38(3):369–374.

Rizk, A., Batt, G., Fages, F., and Soliman, S. (2011). Continuous valutations of temporal logic specifications with applications to parameter optimization and robustness measures. *Theoritical Computer Science*, 412(26):2827–2839.

Roenneberg, T. and Merrow, M. (2016). The circadian clock and human health. *Current Biology*, 26(10):R432–R443.

Ruder, S. (2016). An overview of gradient descent optimization algorithms.

Saccomani, M. and Thomaseth, K. (2018). The union between structural and practical identifiability makes strength in reducing oncological model complexity: A case study. *Complexity*, 2018:1–10.

Sakaue-Sawano, A., Kurokawa, H., Morimura, T., Hanyu, A., Hama, H., Osawa, H., Kashiwagi, S., Fukami, K., Miyata, T., Miyoshi, H., Imamura, T., Ogawa, M., Masai, H., and Miyawaki, A. (2008). Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*, 132(3):487–498.

Sancar, A., Lindsey-Boltz, L. A., Kang, T.-H., Reardon, J. T., Lee, J. H., and Ozturk, N. (2010). Circadian clock control of the cellular response to dna damage. *FEBS Letters*, 584(12):2618–2625.

Sancar, G. and Brunner, M. (2014). Circadian clocks and energy metabolism. *Cellular and Molecular Life Sciences*, 71:2667–2680.

Santos, A., Canuto, A., and Neto, A. (2011). A comparative analysis of classification methods to multi-label tasks in different application domains. *International Journal of Computer Information Systems and Industrial Management Applications*, 3.

Scheiermann, C., Kunisaki, Y., and Frenette, P. S. (2013). Circadian control of the immune system. *Nature reviews immunology*, 13:190–198.

Schmidt, E. and Schibler, U. (1995). High accumulation of components of the rna polymerase ii transcription machinery in rodent spermatids. *Development For advances in developmental biology and stem cells*, 112(8):2373–2383.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473:337–342.

Senekeo-Effenberger, K., Chen, S., Brace-Sinnokrak, E., Bonzo, J. A., Yueh, M.-F., Argikar, U., Kaeding, J., Trottier, J., Remmel, R. P., Ritter, J. K., Barbier, O., and Tukey, R. H. (2007). Expression of the Human UGT1 Locus in Transgenic Mice by 4-Chloro-6-(2,3-xylidino)-2-pyrimidinylthioacetic Acid (WY-14643) and Implications on Drug Metabolism through Peroxisome Proliferator-Activated Receptor alpha Activation. *Drug Metabolism and Disposition*, 35(3):419.

Shimba, S., Ogawa, T., Hitosugi, S., Ichihashi, Y., Nakadaira, Y., Kobayashi, M., Tezuka, M., Kosuge, Y., Ishige, K., Ito, Y., Komiyama, K., Okamatsu-Ogura, Y., and andMasayuki Saito, K. K. (2011). Deficient of a clock gene, brain and muscle arnt-like protein-1 (bmal1), induces dyslipidemia and ectopic fat formation. *Public Library of Science One*, 6(9).

Skarke, C., Lahens, N., Rhoades, S., Campbell, A., Bittinger, K., Bailey, A., Hoffmann, C., Olson, R., Chen, L., Yang, G., Price, T., Moore, J., Bushman, F., Greene, C., Grant, G., Weljie, A., and FitzGerald, G. (2017). A pilot characterization of the human chronobiome. *Scientific Reports*, 7.

Smith, N. F., D.Figg, W., and Sparreboom, A. (2006). Pharmacogenetics of irinotecan metabolism and transport: An update. *Toxicology in Vitro*, 20(2):163–175.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280.

Stapor, P., Fröhlich, F., and Hasenauer, J. (2018). Optimization and profile calculation of ODE models using second order adjoint sensitivity analysis. *Bioinformatics*, 34(13).

Stephanou, A., Fanchon, E., Innominato, P., and Ballesta, A. (2018). Systems biology, systems medicine, systems pharmacology: The what and the why. *Acta Biotheoretica*, 66.

Stolovitzky, G., Monroe, D., and Califano, A. (2007). Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences*, 1115(1):1–22.

Sulli, G., Lam, M. T. Y., and Panda, S. (2019). Interplay between circadian clock and cancer: New frontiers for cancer treatment. *Trends in Cancer*, 5(8):475–494.

Takahashi, J. S. (2017). Transcriptional architecture of the mammalian circadian clock. *Nature Review Genetics*, 18(3):164–179.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J. (2012). The lasso problem and uniqueness.

Traynard, P., Feillet, C., Soliman, S., Delaunay, F., and Fages, F. (2016). Model-based investigation of the circadian clock and cell cycle coupling in mouse embryonic fibroblasts: Prediction of reverbα up-regulation during mitosis. *Biosystems*, 149:59–69.

Ueda, H. R., Chen, W., Adachi, A., Wakamatsu, H., Hayashi, S., Takasugi, T., Nagano, M., Nakahama, K.-i., Suzuki, Y., Sugano, S., Iino, M., Shigeyoshi, Y., and Hashimoto, S. (2002). A transcription factor response element for gene expression during circadian night. *Nature*, 418(6897):534–539.

Ueda, H. R., Hayashi, S., Chen, W., Sano, M., Machida, M., Shigeyoshi, Y., Iino, M., and Hashimoto, S. (2005). System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nature Genetics*, 37:187–192.

Venzon, D. J. and Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 37(1):87–94.

Vertes, A., Avar, A.-B. A. P., Korte, A. R., Li, H., Nemes, P., Parvin, L., Stopka, S., Hwang, S., Sahab, Z. J., Zhang, L., Bunin, D. I., Knapp, M., Poggio, A., Stehr, M.-O., Talcott, C. L., Davis, B. M., Dinn, S. R., Morton, C. A., Sevinsky, C. J., and Zavodszky, M. I. (2018). Inferring mechanism of action of an unknown compound from time series omics data. In *CMSB'18: Proceedings of the seventeenth international conference on Computational Methods in Systems Biology*, Lecture Notes in BioInformatics. Springer-Verlag.

Vlachou, D., Bjarnason, G. A., Giacchetti, S., Lévi, F., and Rand, D. A. (2020). Timeteller: a new tool for precision circadian medicine and cancer prognosis. bioRxiv.

Wang, J., Mauvoisin, D., Martin, E., Atger, F., Galindo, A. N., Dayon, L., Sizzano, F., Palini, A., Kussmann, M., Waridel, P., Quadroni, M., Dulić, V., Naef, F., and Gachon, F. (2017). Nuclear proteomics uncovers diurnal regulatory landscapes in mouse liver. *Cell Metabolism*, 25(1):102–117.

Weinberg, B. A., Marshall, J. L., Hartley, M., and Salem, M. E. (2016). A paradigm shift from one-size-fits-all to tailor-made therapy for metastatic colorectal cancer. *Clinical advances in hematology & oncology*, 14(2):116–128.

Wieland, F.-G., Hauber, A. L., Rosenblatt, M., Tönsing, C., and Timmer, J. (2021). On structural and practical identifiability. *Current Opinion in Systems Biology*, 25:60–69.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypothesis. *Annals of Mathematical Statistics*, 9(1):60–62.

Wolff, G. and Esser, K. (2012). Scheduled exercise phase shifts the circadian clock in skeletal muscle. *Medicine and science in sports and exercise*, 44:1663–70.

Wolkenhauer, O., Auffray, C., Brass, O., Clairambault, J., Deutsch, A., Drasdo, D., Gervasio, F., Preziosi, L., Maini, P., Marciniak-Czochra, A., Kossow, C., Kuepfer, L., Rateitschak, K., Ramis-Conde, I., Ribba, B., Schuppert, A., Smallwood, R., Stamatakos, G., Winter, F., and Byrne, H. (2014). Enabling multiscale modeling in systems medicine. *Genome Medicine*, 6(3).

Woller, A., Duez, H., Staels, B., and Lefranc, M. (2016). A mathematical model of the liver circadian clock linking feeding and fasting cycles to clock function. *Cell Reports*, 17(4):1087–1097.

Xu, G., Zhang, W., Ma, M. K., and McLeod, H. L. (2002). Human carboxylesterase 2 is commonly expressed in tumor tissue and is correlated with activation of irinotecan. *Clinical Cancer Research*, 8(8):2605–2611.

Yang, F., Nakajima, Y., Kumagai, M., Ohmiya, Y., and Ikeda, M. (2009). The molecular mechanism regulating the autonomous circadian expression of topoisomerase i in nih3t3 cells. *Biochemical and Biophysical Research Communications*, 380(1):22–27.

Yu, F., Zhang, T., Zhou, C., Xu, H., Guo, L., Chen, M., and Wu, B. (2019). The Circadian Clock Gene Bmal1 Controls Intestinal Exporter MRP2 and Drug Disposition. *Theranostics*, 9(10):2754–2767.

Zhang, E. E. and Kay, S. A. (2010). Clocks not winding down: unravelling circadian networks. *Nature Reviews Molecular Cell Biology*, 11:764–776.

Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E., and Hogenesch, J. B. (2014). A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of The National Academy of Sciences of the United States of America*, 111:16219–16224.

Zhao, M., Zhang, T., Yu, F., Guo, L., and Wu, B. (2018). E4bp4 regulates carboxylesterase 2 enzymes through repression of the nuclear receptor rev-erbα in mice. *Biochemical Pharmacology*, 152:293–301.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zheng, P., Askham, T., Brunton, S. L., Kutz, J. N., and Aravkin, A. Y. (2019a). A unified framework for sparse relaxed regularized regression: SR3. *IEEE Access*, 7:1404–1423.

Zheng, X., Zhao, X., Zhang, Y., Tan, H., Qiu, B., Ma, T., Zeng, J., Tao, D., Liu, Y., Lu, Y., and Ma, Y. (2019b). Rae1 promotes bmal1 shuttling and regulates degradation and activity of clock: Bmal1 heterodimer. *Cell Death and Disease*, 10(2).

Zoppoli, P., Morganella, S., and Ceccarelli, M. (2010). Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*.

ECOLE DOCTORALE

Titre: Sur l'apprentissage automatique de modèles mécanistes dynamiques à partir de données temporelles avec application aux chronothérapies personnalisées

Mots clés: apprentissage statistique, médecine de précision, modélisation mécaniste, rythmes circadiens

Résumé: La modélisation mathématique des processus biologiques vise à fournir des représentations formelles de systèmes complexes afin d'en permettre des études qualitative et quantitative. Le besoin d'explicabilité suggère le recours à des modèles mécanistes qui décrivent explicitement les interactions moléculaires. Cependant, l'utilisation de ces derniers est conditionnée à l'existence de connaissances *a priori* sur la structure du réseau de réactions sous-jacent. En outre, leur conception demeure un art qui nécessite créativité et de multiples interactions avec les outils d'analyse et de calibration aux données expérimentales. Cela écarte de nombreuses applications imaginables en médecine de précision personnalisée, et appelle à des développements méthodologiques pour l'automatisation de l'apprentissage de modèles adaptés aux données du patient. Cette thèse participe d'un effort de conception d'algorithmes d'apprentissage de modèles d'interactions dynamiques à partir de données temporelles, avec le souci de l'explicabilité à un modélisateur humain. Elle a pour champ d'application la chronothérapie personnalisée qui vise à administrer les médicaments aux horaires optimaux en fonction des rythmes biologiques du patient sur 24h. Ainsi, trois grands thèmes sont abordés : modélisation mécaniste, inférence de réseaux et personnalisation des traitements. Le premier chapitre décrit le développement du premier modèle mécaniste de l'horloge circadienne cellulaire complètement quantitatif, intégrant des données de transcriptome, proteome et localisation sub-cellulaire. Ce modèle a été connecté avec succès à un modèle de la pharmacologie cellulaire d'un anticancéreux, l'irinotecan, afin d'en personnaliser l'horaire optimal d'administration. Le deuxième chapitre présente un protocole original d'inférence des contrôles systémiques qu'exerce le corps entier sur les horloges des tissus périphériques. Cette approche permettra, à terme, d'intégrer des données individuelles issues d'objets connectés pour la personnalisation des chronothérapies. Le troisième chapitre présente un algorithme général d'inférence de réactions avec cinétiques chimiques à partir de séries temporelles.

Title: On learning mechanistic models from time series data with applications to personalised chronotherapies

Keywords: machine learning, precision medicine, mechanistic modeling, circadian rhythms

Abstract: Mathematical modeling of biological processes aims at providing formal representations of complex systems to enable their study, both in a qualitative and quantitative fashion. The need for explainability suggests the recourse to mechanistic models, which explicitly describe molecular interactions. Nevertheless, such models currently rely on the existence of prior knowledge on the underlying reaction network structure. Moreover, their conception remains an art which necessitates creativity combined to multiple interactions with analysis and data fitting tools. This rules out numerous applications conceivable in personalized medicine, and calls for methodological advances towards machine learning of patient-tailored models. This thesis intends to devise algorithms to learn models of dynamical interactions from temporal data, with an emphasis on explainability for the human modeler. Its applications are in the context of personalized chronotherapies, that consist in optimizing drug administration with respect to the patient's biological rhythms over the 24-hour span. Three main themes are explored: mechanistic modeling, network inference and treatment personalization. The first chapter describes the development of the first quantitative mechanistic model of the cellular circadian clock integrating transcriptomic, proteomic and sub-cellular localization data. This model has been successfully connected to a model of cellular pharmacology of an anticancerous drug, irinotecan, achieving personalization of its optimal administration timing. The second chapter introduces a novel protocol for inferring whole-body systemic controls enforced on peripheral clocks. On the long run, this approach will make it possible to integrate individual data collected from wearables for personalized chronotherapies. The third chapter presents a general algorithm to infer reactions with chemical kinetics from time series data.

Institut Polytechnique de Paris
91120 Palaiseau, France