UNIVERSITÉ DU LUXEMBOURG

PhD-FSTC-2019-54
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 11/09/2019 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN SCIENCES DE L'INGENIEUR

by

Laurent MOMBAERTS
Born on 8 June 1989 in Liège (Belgium)

DYNAMICAL MODELING TECHNIQUES FOR BIOLOGICAL TIME SERIES DATA

Dissertation defence committee

Dr Jorge Goncalves, dissertation supervisor
Professor, Université du Luxembourg

Dr Alexander Skupin
Université du Luxembourg

Dr Alexander Webb
Professor, University of Cambridge

Dr Alexandre Mauroy
Professor, Université de Namur

Dr Rudi Balling, Chairman
Professor, Université du Luxembourg

Acknowledgements

Wow! This is it... After 4 years and thousands of hours of work, I am now writing my acknowledgements. This has been an amazing experience, and I consider myself lucky for being so well surrounded.

First and foremost, I would like to thank my supervisors, Jorge Goncalves and Alexander Skupin, for being always available, supportive, and for their valuable guidance over those years. You provided me with a welcoming and stimulating work environment, as well as countless opportunities to improve myself and learn. To my colleagues: Atte, Zuogong, Jin, Nicolo, Johan (x2), Alexandre, Stefano, Rui, Alice, Marino, Rucha, Daniele, Mehri, Andreas, René and Vladimir. I sincerelly enjoyed working with all of you and our trips to conferences, thank you. In particular, I would like to thank Alice for giving me the opportunity to work on her Zebrafish data, and for her personal investment. To Alex Webb, for his patience, determination and involvement. To Rudi Balling, for giving me the chance to live an incredible experience at the Scripps Institute, I was proud of presenting my work by your side.

To my friends, who supported me all those years. In particular, I would to thank my childhood friends Pierre and Joseph for their support and for keeping my mind busy when I needed it.

To Rachel, who has been through these 4 years almost as intensively as I did. Thank you, you have been a constant source of motivation all those years.

To my family, and those we miss.

Abstract

The present thesis is articulated over two main topics which have in common the modeling of the dynamical properties of complex biological systems from large-scale time-series data.

On one hand, this thesis analyzes the inverse problem of reconstructing Gene Regulatory Networks (GRN) from gene expression data. This first topic seeks to reverse-engineer the transcriptional regulatory mechanisms involved in few biological systems of interest, vital to understand the specificities of their different responses. In the light of recent mathematical developments, a novel, flexible and interpretable modeling strategy is proposed to reconstruct the dynamical dependencies between genes from short-time series data. In addition, experimental trade-offs and optimal modeling strategies are investigated for given data availability. Consistent literature on these topics was previously surprisingly lacking. The proposed methodology is applied to the study of circadian rhythms, which consists of complex GRN driving most of daily biological activity across many species.

On the other hand, this manuscript covers the characterization of dynamically differentiable brain states in Zebrafish within the context of epilepsy and epileptogenesis. Zebrafish larvae represent a valuable animal model for the study of epilepsy due to both their genetic and dynamical resemblance with humans. The fundamental premise of this research is the early apparition of subtle functional changes preceding the clinical symptoms of seizures. More generally, this idea, based on bifurcation theory, can be described by a progressive loss of resilience of the brain and ultimately, its transition from a healthy state to another characterizing the disease. First, the morphological signatures of seizures generated by distinct pathological mechanisms are investigated. For this purpose, a range of mathematical biomarkers that characterizes relevant dynamical aspects of the neurophysiological signals are considered. Such mathematical markers are later used to address the subtle manifestations of early epileptogenic activity. Finally, the feasibility of a probabilistic prediction model that indicates the susceptibility of seizure emergence over time is investigated. The existence of alternative stable system states and their sudden and dramatic changes have notably been observed in a wide range of complex systems such as in ecosystems, climate or financial markets.

Table of contents

Glossary

AED	Antiepileptic Drug.
ARNI	Algorithm for Revealing Network Interactions.
ATA	All-to-All.
AUPREC	Area Under the PR Curve.
AUROC	Area Under the ROC Curve.
CWT	Continuous Wavelet Transform.
db4	Daubechie 4.
DD	Dark-Dark (Constant Darkness Condition).
DE	Differential Analysis.
DTFT	Discrete Time Fourier Transform.
DWT	Discrete Wavelet Transform.
DyDE	Dynamical Differential Expression.
dynGENIE3	dynamical GEne Network Inference with Ensemble of trees.
ECG	ElectroCardioGram.

EEG	ElectroEncephaloGram.
FD	Fractal Dimension.
FPR	False Positive Rate.
GPDM	Gaussian Process Dynamical Models.
GRN	Gene Regulatory Network.
HE	Hurst Exponent.
iCheMA	Improved Chemical Model Averaging.
IED	Interictal Epileptiform Discharges.
IQR	Interquartile Ranges.
ISI	Intespike Interval.
LE	Lyapunov Exponent.
LL	Light-Light (Constant Light Condition).
LTI	Linear Time-Invariant.
ML	Machine Learning.
MODWT	Maximal Overlap Discrete Wavelet Transform.
MRA	Multi-Resolution Analysis.
NAM	Nicotinamide.
ODE	Ordinary Differential Equation.
PTX	Picrotoxin.

PTZ	Pentylenetetrazol.
RF	Random Forest.
RFE	Recursive Feature Elimination.
ROC	Receiver Operating Characteristic.
RWE	Relative Wavelet Energy.
SampEn	Sample Entropy.
SISO	Single Input-Single Output.
SNR	Signal-to-Noise Ratio.
TF	Transcription Factor.
TPR	True Positive Rate.
WT	WildType.

Chapter 1

General Introduction

1.1 Modeling of Biological Systems

Biological systems exhibit an elegant combination of complexity and efficiency under many aspects. Perhaps most remarkably, nature has developed nearly ubiquitous control mechanisms that allow a wide range of biological processes to remain stable under various internal or external perturbations. This capability of living organisms to remain in a quasi-permanent stable equilibrium despite the changing environment is a fundamental aspect of life known as homeostasis. The optimal functioning of the human body, for example, requires a tight regulation of a large array of variables such as its core temperature, blood glucose or arterial pressure. Such adaptation requires the sensing, identification and integration of external and internal stimuli at numerous scales. For instance, at the cellular level, this corresponds to a meticulous modulation of chemical balances and regulation of gene expression. Each of these processes shares a common logical structure in such that they all are under constant control of one or more complex feedback mechanisms that ensure their correct functioning. In general, the stability of this category of systems is not only illustrated by their convergence to a fixed value but it can also exhibit more complicated behaviors such as permanent oscillations around stable cycles. This is the case for heart rhythms or circadian systems. In this respect, the dynamical properties of sophisticated biological systems can only be fully comprehended by considering their underlying control mechanisms, rather than merely investigating their isolated parts. A system-level approach that accounts for dynamical interactions between biological components, thus, is of the utmost importance to understand a wide panel of biological responses, the sources and proliferation of complex, multifactorial pathogenesis or drugs effects [1, 2].

An important step in the understanding of any physical or biological phenomenon is its translation from observable behaviors into meaningful, interpretable objects known as models. Modeling is a fundamental scientific methodology that is based on the quantitative formulation of dynamical interactions between variables, perturbations and their product. When applied to biological systems, such methodology, if valid, allows to describe temporal and spatial evolution of complex biological processes, to formulate new hypotheses or to make predictions on their behavior in previous untested situations. Models may come in various shapes and complexity. In particular, their purpose can range from describing a specific process with a high degree of precision, including all pertinent details and species, to conceptually characterizing generic, global features of the phenomenon under investigation. In addition, models can be broken down into smaller pieces each concerned with different aspects of the problem. Their key commonality, however, remains their ability to provide novel insights that would not have been possible to gain otherwise. Overall, modeling is a flexible concept that involves many iterations between prediction, guided experiments and model refinement at its core.

This claim is of particular relevance for the investigation of biological systems. In contrast to human engineered systems for which the functions and properties of every individual parts are known, the analysis of biological systems typically resembles the task of learning from a machine we have never seen before. Moreover, the range of possible manipulations and observable gears of this machinery remains yet limited, while fully controllable engineered systems can be manipulated at will. Such systems involve an incredibly large number of interacting components and it is reasonable to expect that the amount of interactions among biological species is of several magnitudes larger than the quantity of its individual parts. Furthermore, dynamical interactions on one scale may yield unexpected activity at a larger scale, a phenomenon referred to as emergence. The comprehensive study of such systems constitutes an even further formidable challenge as their behavior is intrinsically stochastic, highly nonlinear, and spans across various dimensional scales, from genetic regulation to brain organization. Hence, particularly complex, or even counterintuitive behaviors might be expected. Moreover, while expensive, biomedical data are often subject to non-neglectable inter-intra variability across organisms as well as other types of uncertainties and noise sources.

A Timely Research Topic

At the crossroad between engineering, physics, mathematics and computer sciences, this particularly multidisciplinary field has been fueled by technical innovations at decreasing costs that enabled unprecedent amounts of data to be generated at an increasing resolution

across time and space. Among the most relevant examples are the relatively recent possibilities to measure gene expression at the single cell level, rather than averages of heterogeneous population of cells, or to record neurophysiological activity at the neuron level [3, 4]. Such exponential piling of multi-scale data permitted computational fields, such as machine learning, artificial intelligence and theories of systems identification, dynamical systems and systems control among others, to grow importance and greatly contribute to the biological and medical knowledge through the identification of hidden patterns in highly dimensional systems. The synergy resulting from the intercommunication between confluent scientific disciplines is at the root of almost every modern and significant advances in todays' biomedical research.

The promises carried by proper modeling of biological systems are truly exhilarating. The ability to measure the dynamical nature of complex biological systems is a crucial source of information and provides significant insights of disease progression or drug responses. To date, there is significant progress in the direction of very ambitious aims such as fully automated rule-decision models, personalized medicine, early detection of diseases or the understanding of the fundamental properties of genes [5]. A recent example is the automatic detection of heart malfunctions via an electrocardiogram (ECG) embedded in a wearable device, the Apple Watch, which received FDA approval in 2017 [6]. At the same time, novel biomarkers and therapeutic targets are being identified with the help of mathematical models for a range of acute and critical conditions (e.g. cancer, kidney injury...), models which provided better preclinical evaluation of treatment effects and real-time decision-making guidance [7, 8]. Altogether, both motives and technical possibilities make the modeling of biological systems a timely research topic.

Subject Matter

The present thesis is articulated over two main topics which have in common the modeling of the dynamical properties of complex biological systems from large-scale time series data.

On the one hand, this thesis analyzes the inverse problem of reconstructing Gene Regulatory Networks (GRN) from gene expression data. This first topic seeks to reverse-engineer the transcriptional regulatory mechanisms involved in few biological systems of interest, vital to understand the specificities of their different responses. In the light of recent mathematical developments, a novel, flexible and interpretable modeling strategy is proposed to reconstruct the dynamical dependencies between genes from short-time series data. In addition, experimental trade-offs and optimal modeling strategies are in-

vestigated for given data availability. Consistent literature on these topics was previously surprisingly lacking. The proposed methodology is applied to the study of circadian rhythms, which consists in complex GRN driving most of daily biological activity across many species.

On the other hand, this thesis covers the characterization of dynamically differentiable brain states in Zebrafish in the context of epilepsy and epileptogenesis. Zebrafish larvae represent a valuable animal model for the study of epilepsy due to both their genetic and dynamical resemblance with humans [9, 10]. The fundamental premise of this research is the early apparition of subtle functional changes preceding the clinical symptoms of seizures. More generally, this idea, based on bifurcation theory, can be described by a progressive loss of resilience of the brain and ultimately, its transition from a healthy state to another characterizing the disease [11]. First, the morphological signatures of seizures generated by distinct pathological mechanisms are investigated. For this purpose, a range of mathematical biomarkers that characterizes relevant dynamical aspects of the neurophysiological signals are considered. Such mathematical markers are later used to address the subtle manifestations of early epileptogenic activity. Finally, the feasibility of a probabilistic prediction model that indicates the susceptibility of seizure emergence over time is investigated. The existence of alternative stable system states and their sudden and dramatic changes have notably been observed in a wide range of complex systems such as in ecosystems, climate or financial markets [12, 13, 14].

Overall, the frameworks of systems identification theory, systems control theory, (non)linear time series analysis, dynamical bifurcation theory and machine learning constitute the foundations upon which both the reconstruction of gene regulatory networks and the investigation of brain vulnerability to epileptic seizure are addressed. Hereafter, the background underlying these two problematics is introduced.

1.2 Gene Regulatory Networks Inference

DNA is the main carrier of biological information. The information required by cells or group of cells for their proper functioning, however, is obviously not identical across different times, and cell types. The key to this variability is called gene expression, a process by which the genetic information is dynamically "read" by cells in order to constantly satisfy cells demands. Gene expression consists in two steps: the DNA first gets transcribed into mRNA, then is subsequently translated into its final product: proteins. Yet, only a very small proportion of our genome, approximately 2%, codes for protein;

a larger proportion is regulatory and the function of the rest is still debated [15]. While some genes are continuously expressed, a fined-tuned regulation of the expression of most genes is essential for a timely production of the proteins involved in specific molecular processes such as cell division, circadian regulation, etc. In practice, gene regulation is very often not performed by a single, isolated gene but rather by a variety of other components being themselves regulated. Such sequence of regulatory interactions forms interlocking transcriptional feedback loops that allow gene expression to be precisely controlled and ensure its robustness against perturbations of different nature. The ensemble of biochemical species and their interactions which together control genes expressions is called a Gene Regulatory Network (GRN).

While the functions of many coding protein genes have been elucidated, the vast majority of the causal map formed by the regulatory relationships between genes remains elusive. However, gene mutations or dysregulations in such regulatory mechanisms contribute to a broad range of diseases such as cancer, neurological disorder, diabetes and cardiovascular diseases [16]. The accurate identification and modeling of such transcriptional regulatory circuitry, therefore, would greatly benefit to the overall understanding of disease mechanisms and diagnosis, drug effects and contribute to advances in the field of personalized medicine [16].

The process of reverse-engineering the blue-print, or topology, of such networks and reconstructing their dynamical properties is called gene regulatory network inference, an important research topic of systems biology. The structure of such gene interaction network is assessed to be static, which means that the possibilities for physical interactions between genes are considered fixed. In the biological literature, GRN are often represented as a network of their gene-gene equivalent representation, depicting the genes (nodes of the network) and the causal interaction between them (the links of the network). This latter is an abstraction of the molecular processes involved that considers the role of Transcription Factors (TFs) and their interactions as implicit, although crucial, because genes do not interact directly but by means of their products. The concepts of gene expression, regulation and equivalent GRN representation are illustrated on Figure 1.1.

A Limited System Observability

Gene expression is measured at the transcript level through an extensive value: the level of mRNA it produces. Instead of concentrating at a specific terminal point in time, we now have the possibility to monitor the temporal progression of gene expression over time by next-generation sequencing technologies such as microarray or RNA-Seq [17].

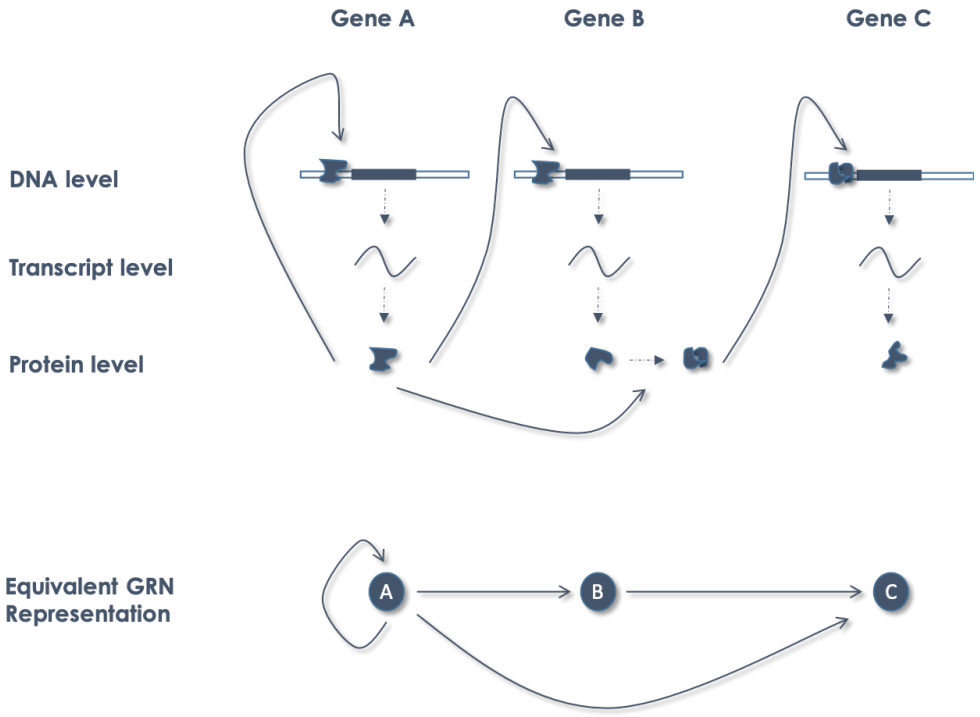

Fig. 1.1 **Genes Expression, Regulation and Equivalent Gene Regulatory Network (GRN).** Gene expression is the process by which the DNA information is dynamically read to produce proteins. The activation or repression of genes, and the amount of protein produced is controlled by proteins called Transcription Factors (TFs). In order to execute their function, they attach in the vicinity of the promoter regions of target genes.

Such data are called time series measurements, as opposed to steady state measurements, and provide the opportunity to uncover transcriptional dynamics between genes [18]. More importantly, it enables the recording and identification of transient changes in gene expression, which is particularly relevant for the analysis of cyclic processes, or in case of perturbation-response experiments [19, 20].

Nowadays, microarrays or RNA sequencing technologies simultaneously provide time series data of the expression of ten of thousands of genes at given points in time. Such recording enables the investigation of the emerging multidimensional expression patterns over time, and thereby constitutes the necessary basis to address the complexity of biological systems dynamics [19]. However, technical and other practical constraints critically restrict the number of time points at which the system can be observed, as well as the amount of replicates that can be produced. Typically, time series recordings of gene expression contain incredibly more genes (10000-20000) than time points (of the order of $\sim$15) and replicates (2-4) (Figure 1.2). Hence, characterizing microarray

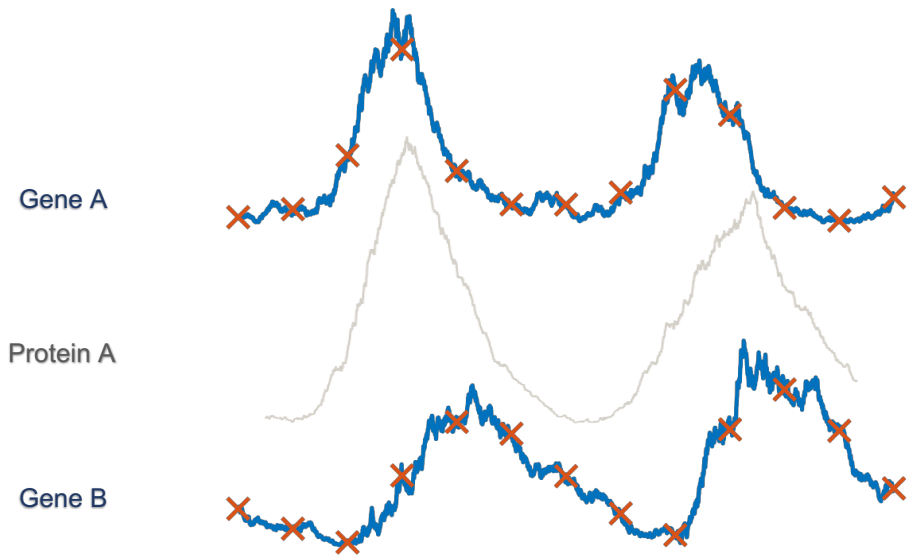

Fig. 1.2 **Illustration of the limited observability of transcriptional dynamics (GRN).** In this example, gene A activates gene B through protein A. Genes expressions are represented in blue, protein levels in grey. Protein levels are typically unavailable to us, despite its crucial role in regulation since genes do not regulate themselves directly. The red crosses represent the discrete sampled observations of the continuous expression process. In this real case example, gene expression is sampled every 4 hours.

or RNAseq data as "big data" would be misleading. Indeed, the problem is ill-posed. From a mathematical point of view, this represents an improper conditioning of the information which is described by a class of mathematical problems called undetermined. Furthermore, measuring mRNA abundance as a proxy for gene expression intrinsically overlooks the effects of other active products in the network. Mechanistic details over the system of interest are then partially visible, which constitutes a major limitation as parts of the regulatory dynamics are hidden to us [21].

Altogether, both partial measurements and subsampling lead to a general lack of observable that constitutes a fundamental challenge of gene regulatory network inference. A proper choice of sampling rate, amount of data points and replicates, are then crucial parameters to take into account for the collection of relevant data. The sub-optimality of experimental design for the investigation of GRN, however, is largely underestimated. For this purpose, general guidelines are investigated in Chapter 2 in order to select the most efficient set of experiments together with appropriate mathematical paradigm and model complexity. Indeed, while navigating experimental tradeoffs is not an entirely new

concept, the literature on the topic is surprisingly scarce or outdated.

As a summary, several challenges underly the inference of gene regulatory networks:

- Large amount of genes recorded, but few time points.
- Intrinsically stochastic processes, but few replicates.
- Partial measurements both in terms of sampling frequency and of the species involved in the transcriptional regulation.
- Noisy measurements, inherent to any observation of physical processes.
- The optimality of the experimental design is often not clear in advance, neither is the most appropriate computational approach to be undertaken, as it relies both on the biological system under investigation and on the availability of resources.

The modeling strategy, therefore, must take into account those limitations and the biological question to be answered.

Distinct Biological Questions and Modeling Approaches

Formally, gene expression over time is represented by the rate of transcription of its corresponding mRNA concentration. It can be formulated as following:

Rate of change of mRNA = Synthesis Rate - Decay Rate

Such equation describes the mRNA levels as a function of two fundamental processes: its synthesis rate from DNA and its decay. The decay, a relatively slow process compared to the synthesis, is an intrinsic function that depends on the mRNA abundance only. On the contrary, the synthesis rate depends on other genes, hereafter referred to as regulators. More generally, reverse-engineering the entire gene regulatory network corresponds to the identification of the regulators π_i and the function f for all given genes $i \in \{1,..,n\}$ (where n is the amount of genes in the system), so that

$$\frac{dy_i(t)}{dt} = f_i(\pi_i(t)) - \alpha_i y_i(t) \tag{1.1}$$

where $y_i(t)$ is the mRNA concentration of gene i at time t and the term $\alpha_i y_i(t)$ corresponds to the degradation rate of $y_i(t)$. The highly nonlinear function f_i represents

the influence of the transcription factors of the parents on the target genes, and can be represented by Michaelis-Menten or Hill type functions under mild hypothesis [22]. The time series data of the regulators π_i consist of other mRNA levels, as the protein levels are typically not available to us, and then $\pi_i(t) \subset \{y_1(t), ..., y_n(t)\}$. The concatenation of the structural regulators π_i of every gene, then, forms the gene regulatory network.

The investigation of the interactions between genes and the emergent properties of gene regulatory networks may be approached with distinct assumptions and modeling strategies. For this purpose, there exist different modeling paradigms that span across various levels of details, faithfulness to biological reality, amount of data needed for modeling or the ability to perform predictions [23, 24].

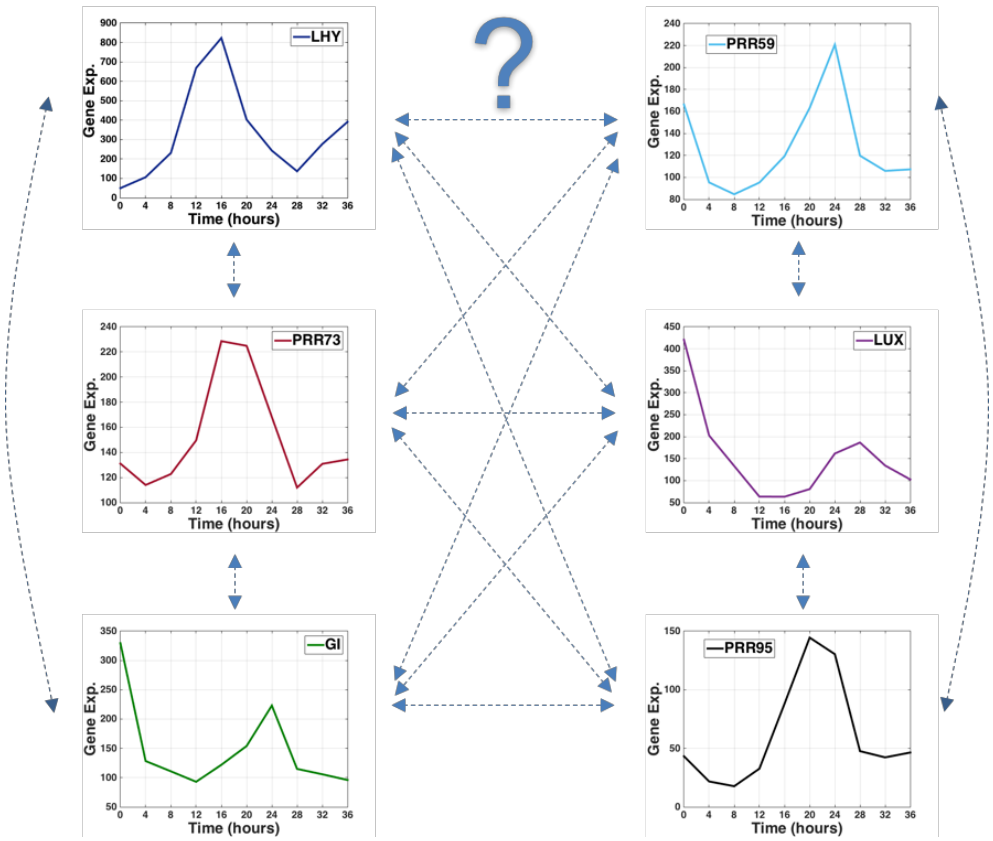

Fig. 1.3 **The network inference problem.** Gene expression is measured through the level of mRNA the genes produce. Typically, gene expression data consists in very few time points, but a large collection of genes. Inferring the topology of the network from time series data means identifying the causal interactions between genes, depicted with dotted lines.

Mainly, there exist two categories of approaches for which the biological questions fundamentally diverge. The first category assumes that at least most of the regulatory connections between genes are known (π_i is known for each gene), which allows to investigate their dynamics in details (the function f_i), i.e. through highly specific, non-linear models of gene regulation. This analysis specifically addresses the question "*How is the regulation performed ?*". However, this category of problem requires considerable prior knowledge of the network structure, which is typically achieved by extensive experimental validation. The second category favors the investigation of network topology, that is, the causal map of transcriptional processes. Such analysis focuses on answering the question "*Who is performing the regulation ?*". It corresponds to the correct identification of the regulators π_i for each gene of the network by formulating often more general hypothesis on the regulation functions f_i. This is the main focus of this thesis (Figure 1.3). Although it is clear that the topology of the network alone does not determine its dynamical behavior, it should be noted that such investigation does not constitute the end product of the study per se, but rather an intermediate and necessary step to learn the functioning of the biological system of interest [25].

Model Complexity

Learning the topology of GRN from time series data is a major challenge of systems biology for which numerous computational approaches have been introduced and many comparisons conducted to assess their respective performances [26, 27]. Their performances, however, have been shown to be crucially dependent on the studied conditions, including: data availability, sampling rate, size of the network, network topology or prior biological knowledge. Due to experimental heterogeneity, therefore, the applicability and accuracy presumptions of those algorithms remain unclear. Hence, it is important that the selected approach is relevant to the biological conditions under investigation and the question to be answered. Furthermore, different mathematical paradigms may carry fundamentally diverging assumptions, which make them more likely to correctly identify different types of regulatory interactions.

In the early stages, the investigation of gene regulatory networks circuitry was performed with association networks [18]. In such case, the association between genes is based on correlation or other informations metrics such as mutual-information between signals. Clusters of co-expressed genes are created, and their functions analyzed. Such network can be constructed either from steady-state or time series data but it does not, however, exploit the underlying dynamical information of time series. The bottom line for those attempts to parse genes into groups is based on the idea that genes that appear to

be correlated or share similar responses to perturbations often share common regulatory mechanisms. Despite not unraveling causality between genes, these networks carry valuable information as countless of biological insights and papers have successfully resulted from this approach [28].

On the opposite, our approach is focused on the reconstruction of causal relationships between genes, which carries the most potential for the understanding of regulatory mechanisms as a whole. Unmistakably, a brute force approach that would thoroughly scan through every possible causal interaction between genes and their associated dynamical behavior is typically not conceivable because of the combinatorial nature of the problem (for 10 genes and without accounting for self-regulation, 4.7^{21} possible directed network structures already exist). That being said, it should be stressed that gene regulatory networks appear to be naturally sparsely connected [29], which is a crucial property that has been widely taken advantage of for the successful development of network inference algorithms.

A central difficulty for the development of inference algorithms is to decide upon the complexity of the strategy, that is, the hypothesis on the functions f_i. For the purpose of reconstructing the structure of the network, representing regulatory functions through a simple model is an advantage, as it requires few or no detailed understanding of the system and less parameters to be estimated [30]. Such model decision takes its root in the well-known overfitting problem of statistical inference: the number of parameters to be accurately estimated may quickly become too large for the given information. As a result of overfitting the data, the model gets really good at predicting cases from which it has learned, but not in unseen data. Its general predictive potential is then very limited. A widely used strategy in the field of machine learning is to separate the data into a training group and a test group. The training group is used to estimate the parameters of the model while the test group evaluate its generalization potential. The opposite case, called underfitting, happens when the complexity of the model is not sufficient to even describe the data it is seeing in the training set. A good balance between the two is essential for a good model (Table 1.1). Such concepts are of crucial importance in the context of biomedical studies. Indeed, biological data are often scarce and noisy so that careful steps have to be taken to build knowledge sequentially. The appropriateness of different models is then a key consideration for modeling.

Train Error	Test Error	Diagnosis
Low	High	Overfitting
High	High	Underfitting
High	Low	Unusual
Low	Low	Good !

Table 1.1 **Addressing the overfitting / underfitting problem.** A widely accepted strategy is to separate the data in two groups, a training set and a test set. The former is used to estimate the parameters of the model. The latter is used to verify its generalization potential.

In general, the suitability of a network inference strategy is estimated from data simulated from toy systems that reproduce realistic experimental conditions and networks. In such case, the ability of each algorithm to accurately reconstruct the topology of the GRN is evaluated in terms of metrics borrowed from the field of machine learning: the resulting Area Under the ROC Curve (AUROC) and the Precision-Recall Curve (AUPREC). On one hand, the ROC curve represents the proportion of regulatory interactions between genes that have been discovered against the proportion of false predictions (predicting a link where this is in fact no direct interactions). On the other hand, the Precision-Recall curve displays the precision (correctly inferred links over total amount of predictions), against how much of the network has been identified. Precision-Recall curves allow for a more accurate picture of algorithms performances for sparse GRNs and are therefore commonly used for such task. The computation of those metrics is now further detailed in the next paragraph.

Both the AUROC and AUPREC require to investigate the amount of regulatory interactions, or links, that are correctly identified between genes, and those that are not. For this purpose, the large majority of network inference algorithm involves a decision threshold. Such thresholding results from the mathematical framework and often stands for the confidence that a link between two genes exists. Some algorithms formally characterize this confidence as a probability while others may be less explicit and simply act as a proxy for such probability. The procedure is relatively simple. One decreases the threshold, starting from no links being identified down to a fully connected network. For each novel link, the sensitivity/recall or True Positive Rate (TPR), precision and False Positive Rate (FPR) are computed (Figure 1.4 and Equations 1.2) and reported on the ROC and PR curves (Figure 1.5).

This thesis explores the structure of GRNs related to two specific complex biological systems in two distinct organisms. It should be noted that the mechanistic premises of

gene regulatory dynamics are inherently similar between organisms. As such, algorithms are transferrable without further adjustments. The biological systems under investigation correspond to a type of rhythmic gene regulatory network called circadian network across two plants organisms: Arabidopsis Thaliana and Barley. The properties of such networks are presented hereafter.

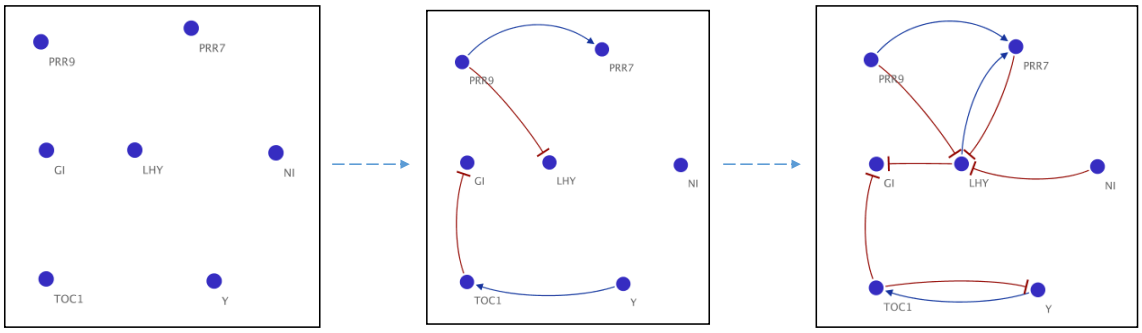

Fig. 1.4 **Gradual assessment of network inference algorithms performances.** Most of network inference algorithms involve a threshold. The thresholding often corresponds to a cut-off on the probability of appearance of links that the user is willing to set. The overall performances of algorithms are assessed for every choice of threshold, by gradually decreasing its value. This affects the amount of links that are inferred, and one can therefore assess whether the regulatory interactions are correctly identified, and evaluate the performance of the algorithm for this threshold. Realistic in silico models, for which the ground-truth is known, are used for this purpose. Blue pointed arrows and red blunt arrows represent activation and inhibition reactions respectively.

		Actual Condition	
		True	False
Predicted Condition	True	True Positive	False Positive
	False	False Negative	True Negative

Table 1.2 **Confusion Matrix.** A specific table that allows to visualize the performances of network inference algorithms. The actual condition refers to the ground truth of the network, i.e. whether a link actually exists between two genes. The predicted condition corresponds to whether a link has been predicted by the network inference algorithm. For example, if a link has been predicted where there is in fact no interactions between those genes, then it corresponds to a false positive.

$$\begin{aligned} Sensitivity &= \frac{TP}{TP+FN} \quad ; \quad Precision = \frac{TP}{TP+FP} \\ FPR &= \frac{FP}{FP+TN} \quad ; \quad Specificity = \frac{TN}{TN+FP} \end{aligned} \tag{1.2}$$

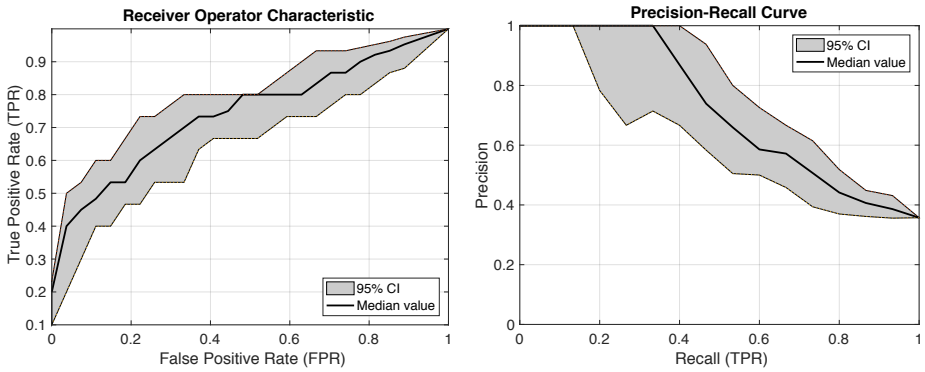

Fig. 1.5 **Resulting Receiver Operating Characteristics (ROC) and Precision-Recall (PR) curves.** The capacity of each algorithm to recover the circuitry of GRN is compared from the area under both curves. The Precision-Recall curves allow for a more accurate picture of algorithms performances for sparse networks such as GRNs. The shaded area represents the variability of this algorithm performance for data originating from the same network, but with different noise states.

1.2.1 Circadian Clocks

Many cell-signaling and transcriptional processes show pulsatile, or even oscillatory behavior. This is the case for circadian transcriptional networks, or circadian clocks. Circadian networks have recently drawn attention in 2017 as the Nobel Prize in physiology and medicine has been awarded to Jeffrey C. Hall, Michael Rosbash and Michael W. Young for their very early work on the molecular mechanisms controlling circadian rhythms.

Circadian clocks consist of complex gene regulatory networks that are responsible for maintaining synchrony of a wide range of biological processes with the daily timing of light and dark cycles resulting from Earth's rotation (Figure 1.6). Present in most organisms, such self-regulating GRN produces oscillations in gene expression with a period of about 24 hours and are continuously synchronized with the external environment by integrating environmental signals, such as light or temperature. This process of synchronization is called entrainment. Studying the mechanisms that dynamically

adjust circadian period and phase, therefore, is critical to understand the control of daily biological activities.

Conceptually, the circadian clock is composed of 3 mains components: a self-sustaining central oscillator, an input pathway that incorporates the environment conditions, and an output pathway that adjusts the metabolism (Figure 1.7). The central oscillator of the clock consists of a complex network of interlocking genes activations, inhibitions and feedback loops. The identification of the functional properties of the individual components in circadian regulatory network is challenging due to the complexity of the interlocked network. Over the past 20 years, the circadian clock of one plant, Arabidopsis Thaliana, has been intensively studied. Several mathematical models have emerged, which fit the experimental data, either in light/dark cycles or constant (light (LL) or dark (DD)) conditions, and elucidate the minimal regulatory structure [31, 32, 33]. The mechanistic basis underlying the adjustment of circadian rhythms to changing external conditions, however, has yet to be clearly elucidated.

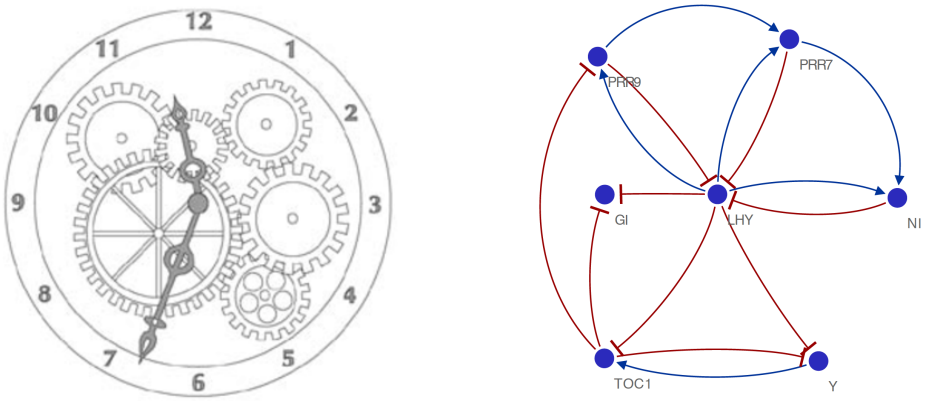

Fig. 1.6 **The Circadian Clock: A biological timekeeping mechanism.** Circadian regulatory networks are the conceptual equivalent of a watch that maintains synchrony with the external environment for most organisms to regulate a wide range of biological processes. To date, it is known to consist of a relatively small amount of genes forming an intricated network of multiple feedback loops across organisms. On the right, a GRN representation of the circadian clock of Arabidopsis Thaliana: 7 genes (nodes) and their complex regulatory interactions (blue arrows represent an activation while red arrows represent inhibition).

This thesis investigates the dynamical mechanisms that are responsible for driving circadian period in Arabidopsis Thaliana and the regulatory structure of the circadian clock of the cereal crop Barley, which represents a significant source of food and animal feed. An important step in improving the yield, a particularly relevant task, is to elucidate

the functional components of its clock. Arabidopsis clock genes reveal a high similarity in nucleotide sequences and expression patterns with the cereal crop Barley [34, 35]. However, clear genetic differences exist. It is therefore unclear how similar the circadian clocks are between barley and Arabidopsis and how the barley oscillator regulates the global transcriptome.

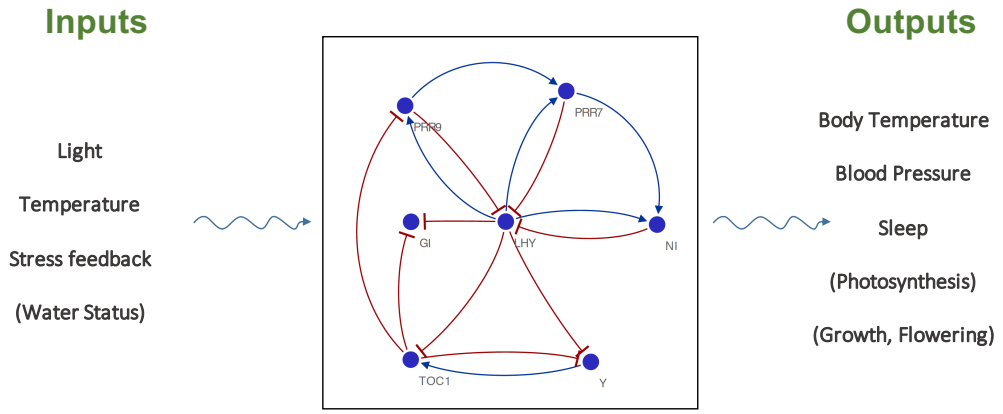

Fig. 1.7 **Inputs and Outputs of the Circadian Clock.** Conceptually, the circadian clock is composed of 3 elements. An input pathway, that integrates environmental cues, an oscillating GRN and an output pathway that regulate a wide range of biological processes. For humans, those outputs include body temperature, blood pressure or sleep, among others. For plants, further inputs may include the water status while outputs control photosynthesis, growth and flowering.

1.3 Characterizing Epileptic Seizures and Epileptogenesis

Living organisms not only create or respond to rhythms, but sometimes unintended rhythms can be generated with adverse effects. Indeed, despite their apparent robustness, complex biological systems may undergo transient or definitive fragility. Many diseases, in fact, are the result of such homeostatic failure [36].

Epilepsy, the fourth most common neurological disorder, affects approximately 1% of the world's population with approximately 30% of the patients being resistant to anti-epileptic drugs [37]. Epilepsy is defined by a state of recurrent seizures, which are characterized by excessive and a synchronous neuronal discharge produced by large regions of the brain [38]. Such neuronal synchronization takes its origin in the disruption of the mechanisms that normally create a balance between excitation and inhibition of

neurons [37]. At the fundamental level, the electrical activity of neurons is a function of its chemical milieu that creates electrical gradients. Normally, there is a high concentration of potassium inside neurons as well as a high extracellular sodium concentration, leading to a negative membrane potential (-70mV). This resting potential is sufficiently close from the activation threshold (-55mV) to simultaneously avoids neurons to constantly fire while allowing them to rapidly discharge and produce an action potential as a result of an environmental stimulus. The ionic basis of the action potential consists in a chain reaction involving the consecutive opening and closing of voltage-gated ion channels, allowing the depolarization and subsequent repolarization of the cellular membrane, and the transmission of the electrical pulse across synapses. On the opposite, the activity of the receptors of the principal inhibitory neurotransmitter γ-aminobutyric acid (GABA) at the postsynaptic sites controls chloride entry into the cells which results in an hyperpolarization of neurons, and maintain the inhibitory tone that counterbalances neuronal excitation. Hence, abnormal neuronal activity can be promoted by numerous and various mechanisms, such as the malfunctionning of sodium or potassium channels [39], or the adverse modulation of GABA receptors [40]. However, the exact origin of the seizure-generating process, or epileptogenesis, is not entirely understood. Conceptually, it is often depicted as a progressive bifurcation of the brain dynamics from a normal, less ordered state to an abnormal, synchronous state [41].

To date, more than 500 genetic mutations have been associated with epilepsy in humans [37]. Yet, genetic alterations are not the only possible causes of seizures, which can essentially be triggered by a wide range of brain functions perturbations. Such perturbations include brain insults (e.g. strokes, brain trauma, Alzheimer disease, etc.), infectious diseases or autoimmune diseases [37]. It has been proposed that both the healthy and synchronous states of the brain co-exist in its dynamical landscape [42, 43], which may explain why so many different neurological conditions are also associated with seizures. Hence, there exists a range of pathological causes that share the same conceptual outcomes: bringing the brain system in the vicinity of a seizure state [44, 45, 46]. For healthy brains, the neuronal activity is far from the seizure state and requires strong stochastic circumstances to operate a transition, such as intoxication, metabolic disturbances or brain insults [44].

The epileptic condition, however, imposes constraints on brain dynamics, reducing the threshold separating the healthy state from the seizures. More specifically, experimental evidence points towards the crucial role of a slowly changing variable describing the loss of resilience and stability of the brain systems towards seizures, which reflects the long-lasting re-organization of the brain system [10, 11]. From a certain point on,

stochastic fluctuations can precipitate the brain systems to seemingly sudden and rapid changes of condition. The idea of a slowly varying system, however, carries promises for the identification of early warning predictors of epileptic seizures. Notably, experiences of focal seizures (one hemisphere of the brain) of humans are often preceded by certain sensory or motor phenomena, known as "aura" [47].

A fundamental challenge of modern epilepsy research, therefore, is the accurate and early detection of the onset of the transition to the epileptic region from brain activity monitoring data to allow for rapid intervention and treatment. While often the most common mean employed in practice to identify a seizure is a visual inspection by a medical doctor, over the last decades, the availability of methods for the automatic detection and prediction of human epileptic seizures has dramatically increased [48, 49]. In humans, changes in the spatiotemporal patterns of brain wave activity have been observed up to 70 minutes in advance [50].

Zebrafish as an animal model

Brain activity in humans in often performed by noninvasive recordings such as ElectroEncephaloGram (EEG) from the scalp or intracranial recordings with Local Field Potentials (LFP) that monitor charges separation directly from the extracellular space. The availability of brain activity recording in humans, however, is scarse, costly and can reach ethical barriers. In addition, seizures-generating process may differ between patients, especially given the large range of causes and symptoms characterizing the epileptic condition. As a result, the heterogenic nature of seizure generation cannot be fully examined in humans, which is suboptimal for both the understanding of the underlying epileptogenic mechanisms and the development of prediction algorithms [51, 52]. As an additional challenge, seizure occurrence is occasional as humans experience clinical seizures on average for less than 0.05% of the total time of studies.

As an alternative to human experiments, a lot of interest has recently been rising in performing experiments in Zebrafish larvae (Figure 1.8) due to the easy handling, low cost and high homology with the human genome [9]. Furthermore, the central nervous system of Zebrafish, while simpler, shares functional and structural similarities with the mammalian system [53, 54]. As an animal model, specific seizures mechanisms can be triggered by means of genetic mutations or drugs [51]. Finally, Zebrafish recordings are exempt of some endogenous and exogenous factors such as the disease stage or the influence of circadian rhythms, which have been recently reported to influence seizure

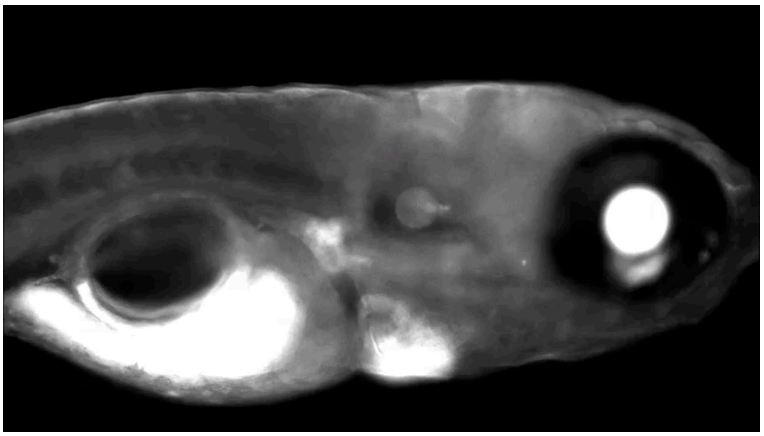

Fig. 1.8 **Zebrafish Larvae.** Zebrafish have high homology with the human genome. Furthermore, they are transparent so that neuron activity can be monitored non-invasively by neuroluminescence using transgenic fish expressing the Ca^{2+} photoprotein GFP-apoAequorin (GA). As such, they are valuable animal model for the study of epilepsy.

emergence [55].

Zebrafish larvae are of the size of millimeters, can be grown and employed in an experiment in huge numbers, have a fast reproduction rate and development, and within their fifth day of life they do not have pain receptors. Importantly, Zebrafishes are transparent, which allow imaging techniques to simultaneously monitor the spatio-temporal activity of neurons together with the LFP recordings that measure their collective activity.

Overall, being able to further characterize seizures in Zebrafish has the potential to open the door to a whelm of novel knowledge and tools to anticipate and treat the disease in humans.

Seizure Detection and Characterization

In epilepsy research, brain states are typically classified into four events (Figure 1.9). The *ictal* state refers to the seizure event per se, i.e. during the hypersynchronous activity of large assemblies of neurons. The *pre-ictal* state refers to the moments just before the seizure, where clinical symptoms of seizures are not yet apparent. The duration of the pre-ictal state is not well defined, and often arbitrarily delineated. It corresponds to a region where the brain has reached a critical state and seizures are likely to occur. The *post-ictal* state represents the recovery of the brain towards a normal state and the *interictal* phase corresponds to the moments in between seizures, during the normal functioning of the brain. The latter stage accounts for most of the patient life, which

makes epileptic seizure a rare and particularly impairing condition.

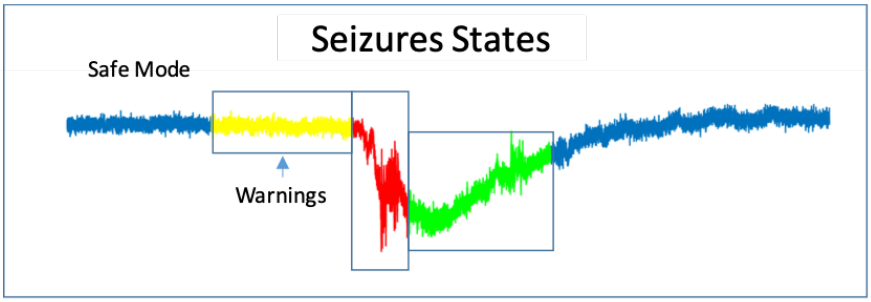

Fig. 1.9 **Clinical Seizure Terminology.** The red region of the signal corresponds to the seizure event (ictal state). The recovery of the brain towards a normal state is depicted in green (post-ictal). In yellow, the moments before the seizure, where an incoming seizure is likely to happen (pre-ictal). The blue region corresponds to the interictal state, supposedly far from ictal states.

The first step to be undertaken in the understanding of the brain conformation and mechanisms that lead to seizures events is to isolate seizures events from the entire signal. It is a non-trivial task, however, that requires expert knowledge, considerable time and experience. Yet, such task represents a very promising area for the development of automated decision systems based on automatic prediction models. For this purpose, many algorithms have been developed with varying degree of success for different organisms such as mice, humans and more recently, Zebrafish [48, 56]. Yet, EEG/LFP data recordings of brain activity typically suffers from a few limitations that hinder their effective analysis or processing [57]. In more details:

- Brain activity patterns are subject to patient intra-inter variability (Figure 1.10). This phenomenon arises from physiological differences between individuals, which vary in magnitude but severely affect the performance of models that are meant to generalize across subjects [58]. Since the ability to generalize from a first set of individuals to a second, unseen set is key to many practical applications of EEG, it is crucial to develop methods that hold high generalization potential while remaining specific enough to capture seizures events only.

- Brain recordings are highly nonlinear and non-stationary in nature, that is, their statistics vary over time. As a result, the investigation of brain patterns on a temporally-limited amount of subject-specific data might generalize poorly to data recorded at a different time on the same individual (1.11). This is crucial challenge for real-life clinical applications of EEG/LFP, which often need to work with limited amounts of data.

- The signal holds a low-signal-to-noise ratio (SNR) (Figure 1.11). EEG/LFP signals are typically contaminated with noise originating from various sources. The most prominent external source of noise is the ambient electrical 50 Hz frequency. Internal sources of noise of eventually very large amplitude include body motion, muscle activity or eye blinking [59]. Those large unrelated signals are considered artifacts that need to be removed from the signal. Finally, the brain is engaged in many different activities in time and space, which all get mixed into the overall signal captured by the electrodes.
- Finally, brain dynamics span orders of magnitude in space and time, making the dynamical activity of populations of neurons both rich and difficult to understand.

The key challenge in correctly discriminating seizures events from baseline brain activity, therefore, is constructing a predictive model that is robust to translation and deformation of signal in space, frequency, and time, due to inter- and intra-subject differences, as well as signal acquisition protocols (electrode positioning etc.).

In its core, the identification of seizure events from EEG or LFP recordings can be tackled from two different perspectives. On one hand, sub-sequences of signals from interictal and ictal phases can be retrospectively extracted by clinicians, and their differences subsequently characterized through the development of a predictive model that captures the main discriminative features of both signals. Such approach allows to evaluate the applicability range of automated approaches as well as discover novel knowledge on important discriminating features. However, it does not permit the development of algorithms for implantable devices for continuous diagnostic or therapeutic purposes. Yet, a vast majority of the current literature on automatic seizure detection focus on the development of novel and not necessarily interpretable algorithms through publicly available datasets [58], therefore overlooking their direct clinical portability. It is often referred to *offline* detection of seizures. On the other hand, seizure identification can be performed in real-time, which often represents a more subtle mathematical challenge but yet closer to a clinically relevant framework. In such approach, data are provided to the algorithms in the form of a sliding window and the analysis is focused on emitting an alarm as soon as the seizure is detected. The latter approach is called *online* automatic detection. This thesis considers both *offline* and *online* aspects for the identification and prediction of seizures events.

For this purpose, machine learning tools play a significant role in the field as it offers tools to address the high complexity of EEG/LFP signals through the integration of multivariate factors to distinguish between brain states. Traditionally, the standard

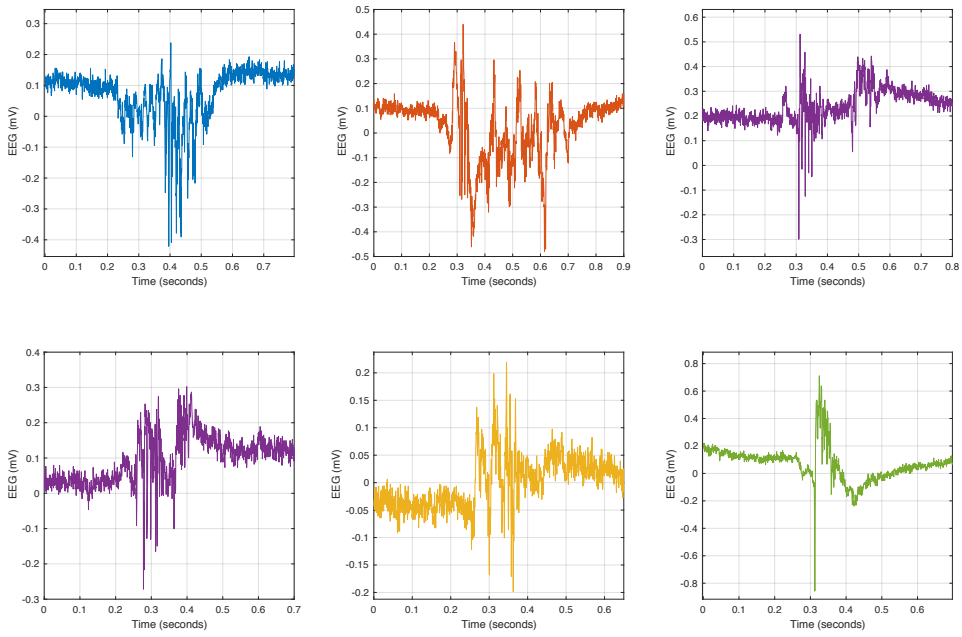

Fig. 1.10 **Subject variability of seizures-like events recorded from LFP recordings of 5 Zebrafish.** Seizures-like events are displayed here in 5 colors, representing different subjects. The characteristics of seizures vary drastically across fishes, but within fishes as well. Among those, the duration and amplitude of seizures, but also their dynamical signature. Characterizing the signature of seizures and their dynamical (in)variants is a crucial step in the understanding of seizures mechanisms.

approach undertaken by most automatic seizure detection algorithms can be summarized as: (1) extraction of a set of features from a short time window of a few seconds to few minutes and (2) classification into different epileptic stages. Features extraction typically consists in reducing the dimensionality of the original signal to a lower-dimensional space that represents its most salient characteristics. Features can be computed from raw EEG/LFP signals or from a decomposition of the signal. Furthermore, they can be extracted either from the time domain, from the frequency domain or both. It should be noted that feature engineering represents one of the most demanding steps of the EEG/LFP analysis pipeline [60, 61].

1.3.1 Predictive Modeling

The language used to describe the behavior of complex nonlinear systems characterized by transitions between dynamical regimes is of great relevance to seek for global invariants in the collective neuronal dynamics leading to seizures emergence. It is hereafter introduced.

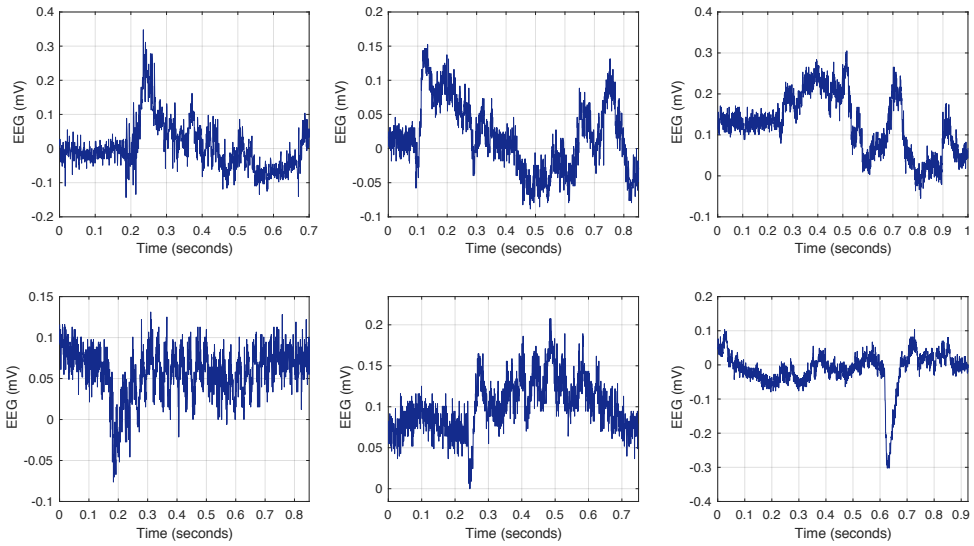

Fig. 1.11 **Abnormal LFP activity.** EEG/LFP signals hold low signal to noise ratio. This figure displays artifacts and signal variations that significantly differ from baseline activity, but do not represent seizures.

The multistability of the healthy brain and the simultaneous existence of an epileptic state can be represented by two basins of attractions, or *attractors*, separated by a seizure threshold, or *separatrix* [10]. A dynamical bifurcation is said to occur when the dynamical regimes of the system critically change (e.g. transition from a laminar fluid to a turbulent one and vice versa). Such transition can be achieved by crossing the separatrix or through geometrical modification of the dynamical landscape, which ultimately leads the nonlinear system towards another attractor. Concerning seizure emergence, such phenomenon can be achieved through multiple routes, possibly reflecting the involvement of specific cellular mechanisms [62, 63]. Dependent on the response of the system after moving beyond a tipping point, different classes of bifurcations have been defined that can lead to multiple equilibria and attractor states, e.g. homoclinic or Hopf-bifurcations [64, 65].

Of particular interest here is the subcritical Hopf bifurcation to describe the loss of system stability towards seizures and the homoclinic bifurcation to characterize system's recovery to the normal neuronal activity [10, 66]. The Hopf bifurcation describes the transition from out of the stable region to a periodic behavior while the homoclonic bifurcation describes the transition outside of the synchronous neuronal activity to leave the system with one stable point. Nevertheless, while a common mathematical framework may be formulated, dynamical specificities that account for various pathological conditions surely remain. To understand the brain predisposition to seizures, it is crucial

to elucidate the dynamical pathways through which the bifurcation point is reached.

More specifically, the transition between the healthy state and the epileptic one appears to be generally governed by the principles of critical slowing down, or slow-fast dynamical processes, where a slight perturbation or disturbance takes a significant time to be recovered by the system [45, 67, 68]. This phenomenon can be described by a pathological decrease of the sepatrix threshold over time. The transient synchronous activity of pathologically interconnected neurons during interictal and pre-ictal periods, known as interictal epileptiform discharges (IEDs), is thought to be visible manifestation of such threshold decrease (Figure 1.12) [69]. IEDs can take many shapes and their role in epileptogenesis is still a matter of debate, ranging from being an adaptive phenomenon (manifestation of feedback processes) to disruptive (feedforward mechanisms) or merely co-occurring mechanisms [70]. Mathematically, such changes of dynamics preceding the ictal region are characteristic features of subthreshold oscillations occurring in the vicinity of a dynamical bifurcation, that is, close from the separatrix [71]. Hence, the occurrence of IEDs at least corresponds to reflections of the dynamical state of the brain system, which is analogous to the definition of a pre-ictal state. Furthermore, the amplitude and frequency of IEDs allow to formulate the type of bifurcation occurring. The quantification of the role IEDs in seizure emergence, however, is challenging. Indeed, these fluctuations are notably variable [72], do not inevitably progress to a seizure [73], and are dominated by the normal baseline activity.

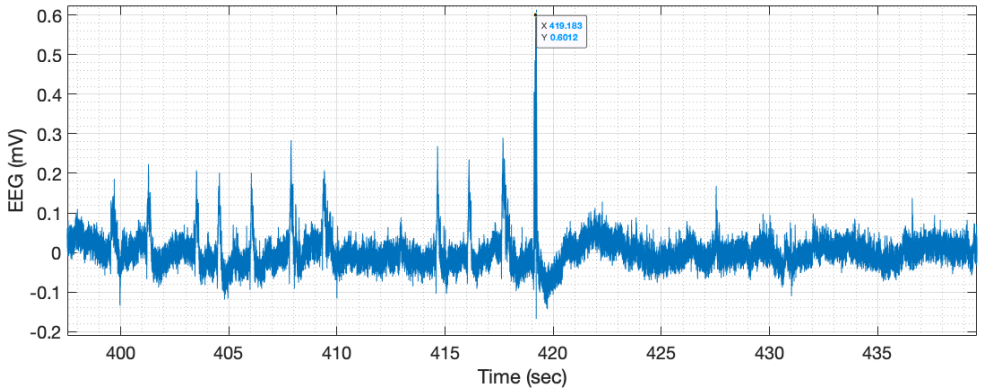

Fig. 1.12 **Interictal Epileptiform Discharges (IEDs).** The seizure has been centered in the middle of the Figure. 10 IEDs can be observed before seizure occurrence. After the seizure, the brain activity returns to a normal state, without IEDs.

In the recent years, the frontiers of computational epilepsy research have moved from seizure detection to a more challenging problem: seizure prediction [74, 75, 76].

The mathematical framework underlying the development of prediction algorithms is very similar to detection, besides that statistical classifiers do not aim at discriminating between fully developed seizure states and normal brain activity, but rather at identifying the pre-ictal region characterizing the vicinity of the separatrix. Whereas it is typically not difficult to discriminate between a healthy and a fully developed diseased state, the differentiation between healthy and pre-disease state is much harder. To date, no predictive characteristic or pre-seizure biomarker that is universal and forecast the exact time of the next seizure has been identified [76]. The lack of such biomarker may originate from a range of factors, including the general lack of long-term continuous recordings of human EEGs, or from the pooling of patients with different pathological conditions as a result of sparse data availability.

In this thesis, the aim is to keep with the general view of dynamical bifurcations to address the possible (in)variants properties of seizures types by means of distinct animal models. The objectives are twofold: contributing to the current knowledge of seizures and epileptogenesis mechanisms and pave the way for targeted and more accurate prediction models. For this purpose, a probabilistic prediction model is developed to conceptually formulate seizure occurrence as a function of the distance between the attractor states.

1.4 Thesis Objectives & Overview

Objectives

This thesis is divided into two main parts: (1) the efficient modeling of gene regulatory networks from sparse microarray data and (2) epileptic seizures and epileptogenesis from local field potentials. The predictive models and mathematical frameworks developed here aim at providing a solution to current, real-world biomedical problems. As such, "efficient" modeling refers to the optimization of model complexity, parameters estimation, model validation and the modeling conditions to provide the best solution to given problems. Furthermore, the thesis principally aims at generating meaningful biomedical knowledge, so that interpretable models are favorized throughout the whole manuscript.

The objectives respectively consider:

1. The development of efficient modeling strategies for the identification of GRN in the context of circadian networks. Such approach is data-driven, in such that two applications are targeted. On one hand, the investigation of the mechanisms responsible for driving circadian period in Arabidopsis Thaliana. On the other hand, the identification of the yet vastly unknown circadian regulatory network of Barley.

2. The development of efficient modeling strategies for the automatic extraction, characterization and prediction of pathologically distinct epileptic seizures mechanisms of zebrafish from LFP data.

Overall Contributions

In this thesis, a modelling strategy based on the identification and comparison of gene regulatory dynamics before and after a perturbation occurred in the network is first introduced. The rationale behind this approach is that not only genes, but also their interactions, are affected by a drug. This reasoning is further supported by [2, 77, 78], which highlight the fact that drugs and diseases mechanisms should be regarded as network instead of gene-centric perturbations. We designed our modelling strategy so that it could be applied to large datasets with scarce sampling (described in Chapter 2).

The Dynamical Differential Expression (DyDE) methodology uses a reverse engineering approach that favors both the accurate identification of unknown Gene Regulatory Network (GRN) topology and the interpretation of the possible dynamical changes, without the need to cover extensive experiments or to make prior assumptions of network dynamics. The modelling strategy and its application to the Arabidopsis circadian network resulted in a publication in PLoS Computational Biology, 2019 (described in Chapter 3):

- ***Mombaerts, L.*** *et al. Dynamical differential expression (DyDE) reveals the period control mechanisms of the Arabidopsis circadian oscillator. PLoS Comput. Biol. (2019).*

The flexibility and accuracy of the introduced approach has been further exploited to uncover the circadian network of barley, a yet unknown complex regulatory network. This investigation resulted in a paper, currently under submission (described in Chapter 4):

- *Lukas M. Müller*,* ***Laurent Mombaerts*****, Davis SJ, Alex A. R. Webb, Jorge Goncalves and Maria von Korff. Dynamic modelling of the barley circadian clock and transcriptome rhythmicity analysis reveal differential effects of the day-night cues and circadian clock on gene transcription. (Submitted to Plant Cell)*

In addition, the strategy developed here has also been successfully applied to another kind of GRN related to the human immune system. This study is not detailed in this manuscript, since it merely consisted in applying the developed algorithms (DyDE) to another biological system without further theoretical development. Nevertheless, such

application across organisms illustrates the flexibility of the proposed approach. This research topic resulted in the following contribution:

- *Sawlekar R., Magni S., Capelle C., Baron A.,* ***Mombaerts L.****, Zeng N., Yue Z., He F., Goncalves J. Dynamical modelling predicts novel regulatory genes of FOXP3 in humans. (To be submitted)*

This thesis also assesses the performance of recent and successful network inference strategies under a novel, multifactorial evaluation framework in order to highlight pragmatic tradeoffs in experimental design. The effects of data quantity and systems perturbations are addressed, thereby formulating guidelines for efficient resources management (Chapter 2). It is shown that data originating from transients systems dynamics are more informative for the identification of regulatory interactions between genes, which is novel and of particular relevance for oscillating networks. Furthermore, it is shown that network inference strategies do not benefit equally from data increments and across experimental conditions. This constitutes a novelty as this analysis allows to further unveil experimental trade-offs and computational performance of algorithms in such conditions. As such, it is shown that, in order to provide a comprehensive comparison between algorithms, it is necessary to perform a multifactorial analysis of the algorithm performances. Such analysis resulted in two conferences papers at the Foundations for Systems Biology in Engineering (FOSBE), in 2016 and 2019:

- ***Mombaerts L*** *et al. Optimising time series experimental design for modelling of circadian rhythms: the value of transient data. FOSBE 2016.*
- ***Mombaerts L*** *et al. A multifactorial evaluation framework for gene regulatory network reconstruction. FOSBE 2019.*

The thesis also contributes to the global knowledge of epileptic seizures and epileptogenesis. In particular, epileptic seizures and their dynamical signatures are investigated for different Zebrafish models, i.e. for both chemically induced seizures and genetic variants. So far, Zebrafish have been used either because of convenience or genetic resemblance with humans but it is not known to what extent the mechanistic processes generating seizures are similar. Investigating those topics resulted in the following contribution:

- *Oldano A,* ***Mombaerts L****, Magni S., Goncalves J., Skupin A. Machine learning classification of epileptic seizures reveals distinct dynamic mechanisms in zebrafish models. (To be submitted)*

Then, based on the previously gained knowledge, a probabilistic seizure prediction model is proposed for Zebrafish. This investigation will be considered for further publications.

The investigation of dynamical perturbations and transitions of complex adaptive systems in the context of epileptic seizures and heart arrhythmias in humans also resulted in the following book chapter:

- *Balling R, Goncalves J, Magni S,* ***Mombaerts L****, Oldano A, Skupin A. From diagnosing diseases to predicting diseases.*

Collaborations with external scientific partners also led to the following publications:

- *Tzortis I, Hadjicostis CN,* ***Mombaerts L.*** *Reconstruction of gene regulatory networks using an error filtering learning scheme. Communication, Control and Computing. 2017*
- *Aalto A, Viitasaari L, Ilnonen P,* ***Mombaerts L****, Goncalves J. Gene regulatory network inference from sparsely sampled time series data. (To be submitted)*

The overall thesis structure is illustrated on Figure 1.13.

PART I

CHAPTER 2 describes the development of the Dynamical Differential Expression (DyDE) framework, which allows both identification of the underlying gene regulatory circuitry without prior knowledge but also favors the interpretation of its dynamical properties before and after a perturbation occurred in the biological system of interest. It is shown that the modeling approach developed here is valuable tool both in terms of accuracy of identification and mechanistic interpretation. Furthermore, the optimization of experimental design is also addressed using various state-of-the-art mathematical approaches to GRN inference.

CHAPTER 3 applies the DyDE modeling framework to the identification of the mechanisms responsible for driving circadian period in Arabidopsis Thaliana. This led to two outcomes. On one hand, *PRR7* has been identified as a regulator of the pace of the circadian clock in Arabidopsis by learning the effect of nicotinamide from differentiated systems. On the other hand, the role of blue light is uncovered in the response of the circadian oscillator to nicotinamide.

CHAPTER 4 investigates the yet unknown circadian regulatory network of barley through the application of the modeling framework previously introduced. The flexibility and interpretability of the model class used is exploited to gain knowledge on the effects of light on the circadian networks and its internal coincidence.

Part II

CHAPTER 5 covers mathematical preliminaries on dynamical bifurcations, state-space embedding, a time-frequency analysis of signals called wavelet decomposition and introduces the machine learning tools used in the following chapters.

CHAPTER 6 describes the development of the automatic seizure extraction algorithm for several Zebrafish models of epilepsy. Then, a random forest approach is used to investigate the dynamical signature of seizures.

CHAPTER 7 describes the development of the probabilistic model for the prediction of seizures in Zebrafish, as well as its interpretation.

CHAPTER 8 is a discussion of the main findings and future perspectives of each topic covered in this thesis.

Talks & Posters

- Thesis presentation "Inference of gene regulatory networks" at the ERASyS International Group , Munich, Germany.
- Thesis presentation "Inference of gene regulatory networks" at the ERASyS International Group, Brussels, Belgium.
- Thesis presentation "Dynamical modelling of RNAseq time series data" at KTH, Stockholm, Sweden.
- Thesis presentation "Towards early warning signals in epilepsy and cardiac diseases" Scripps Institute, San Diego, US.
- Poster presentation "Optimising time series experimental design for modelling of circadian rhythms: the value of transient data" at FOSBE conference, Magdeburg, Germany.
- Poster presentation "A linear modelling framework for the reconstruction of gene regulatory networks" at ERNSI workshop, Lyon, France.

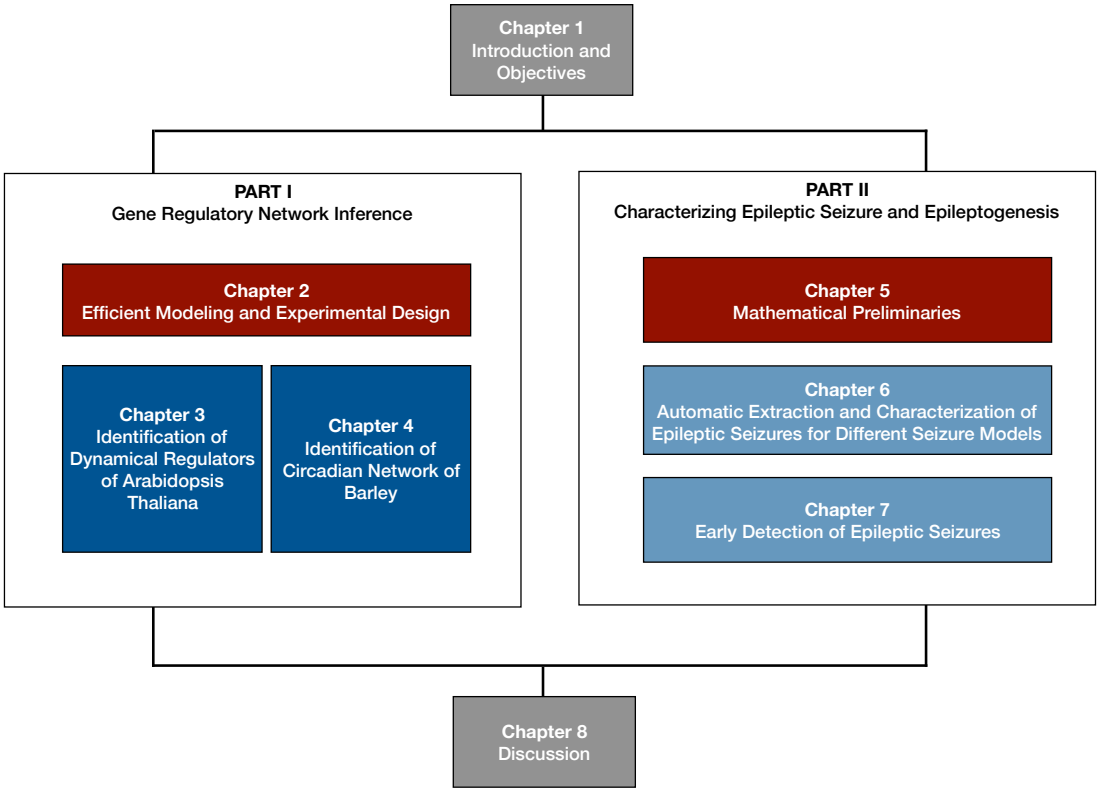

Fig. 1.13 **Thesis Overview.** Chapters colored in red depicts theoretical chapters or developments based on *in silico* generated data. Blue chapters consist in applications using real data. Dark blue chapters comprise those applications related to circadian networks in plant and crops. Light blue chapters represent the analysis of epileptic seizures in Zebrafish.

- Poster presentation "Effects of explicit derivatives computation for the identification of network inference from time series data" at ERNSI workshop, Cambridge, UK.
- Educational talk "Introduction to Machine Learning" at LCSB, Belval, Luxembourg.
- Conference talk "A multifactorial evaluation framework for gene regulatory network reconstruction" at FOSBE conference, Valencia, Spain.

Part I

Gene Regulatory Network Inference

Chapter 2

Efficient Modeling and Experimental Design for GRN Reconstruction

2.1 Contribution

The contribution of this thesis to the identification of GRN is twofold.

First, an algorithm is developed for the automatic identification of transcriptional dependences between genes given relatively small amount of time series data and no a priori assumptions on the topology of the system. This mathematical framework allows both the reliable identification of unknown GRN's topology and the interpretation of possible dynamical changes following a network perturbation. This work has been particularly motivated by the investigation of drug mechanisms of action within complex biological systems, and more specifically in the context of circadian networks. Nevertheless, it should be noted that its applicability range is not restricted as such. By means of *in silico* and real data, it is shown that the performances of the proposed mathematical framework are comparable for steady-state biological networks, and outperform most of the current state-of-the-art network inference algorithms as well.

Then, the effects of experimental design on the performances of relevant state-of-the-art mathematical models are investigated. Indeed, in the past years, many computational methods have been developed to infer the structure of gene regulatory networks from time series data. However, the applicability and accuracy presumptions of such algorithms remain unclear due to experimental heterogeneity. Hence, in practice, experimentalists are still faced with difficult questions: which method to use with an available dataset? Alternatively, with fixed amount of resources, what kind of experiments to carry out to ensure an optimal use of resources? How much can be gained by investing into few more

datapoints? The results highlight the importance of transients data and that algorithms do not benefit equally from data increments. The purpose of this study is to provide guidelines for conscientious management of biological resources by unveiling the performances of state-of-the-art network inference strategies under various experimental designs. To this end, the effects of data quantity and multi-experiment availability are assessed simultaneously on the accuracy of the topological reconstruction, thereby formulating experimental trade-offs and practical guidelines, which are yet vastly unexplored.

2.2 Network Inference and Analysis by Dynamical Differential Expression (DyDE)

Adapted from: Mombaerts, L. et al. Dynamical differential expression (DyDE) reveals the period control mechanisms of the Arabidopsis circadian oscillator. PLoS Comput. Biol. https://doi.org/10.17863/CAM.35626 (2019) [79].

2.2.1 Introduction

The identification of gene regulatory networks is a major challenge of systems biology. This is due on one hand to the complexity of the interlocked network and on the other hand to the partial observations of the species and mechanisms involved [19, 80]. To this end, numerous computational approaches have been introduced and many comparisons conducted to assess their respective performances [26, 27]. However, performances of such network reconstruction algorithms have been shown to be crucially dependent on the studied conditions, including: data availability, size of the network, network topology or prior biological knowledge [26, 81].

The network inference algorithm introduced in this chapter has been developed in response to the particularly small amount of data (10-12 datapoints) that is typically representative of microarray recordings resulting from the study of a novel GRN. Conveniently, the strategy chosen in this thesis allows the comparison of the dynamical properties of GRN, which is introduced as well. Additionally, the strategy developed is flexible and highly scalable, so that it can be applied to very large networks (10 000+ genes) to uncover fundamental regulatory interactions. In the case of the problems sought to be solved in this thesis, those properties are of particular importance, since the most relevant information are scanned across large datasets, and through slight modifications of the model across applications.

2.2.2 Model Class

This thesis introduces the development of a systematic and scalable dynamical modeling framework named Dynamical Differential Expression (DyDE). DyDE uses a simple yet consistent modeling approach to reverse-engineer comparable gene regulatory dynamics from time series data. In addition, it does not use any prior information and, hence, it is unbiased towards prior knowledge of network topology and dynamics. The general equation (1.1) of gene regulation is repeated hereafter to introduce the methodology:

$$\frac{dy_i(t)}{dt} = f_i(\pi_i(t)) - \alpha_i y_i(t) \tag{2.1}$$

where $y_i(t)$ is the mRNA concentration of gene i at time t and the term $\alpha_i y_i(t)$ corresponds to the degradation rate of $y_i(t)$. The nonlinear function f_i represents the influence of the transcription factors of the parents on the target genes. The goal of the reconstruction of the gene regulatory network, therefore, is to find the regulators π_i for each gene of the network.

Here, Linear Time-Invariant models (LTI) are used to capture the dynamics describing the rate of change of the selected mRNA. Both the synthesis rate of the output mRNA and the degradation term, therefore, appear linearly in the equation above. Linear models are a simple yet flexible class of models that represent a local approximation in time and frequency (first order truncation at a point of interest) of the underlying nonlinear, complex physical processes at play, where their properties are preserved.

The motivations for such choice are multiple:

1. While complex nonlinear models have the potential to capture the dynamical relationships between genes with great precision, it should be noted that large amount of experiments are often required to accurately reconstruct the topology of the network in such case [27]. In particular, such complexity is typically subject to overfitting (fitting the noise instead of dynamics) without sufficient data or detailed knowledge such as network topology, types of nonlinear interactions, or potentially some of the model parameters (e.g. Hill coefficients).

2. Linear Time Invariant (LTI) systems are the most widely studied class of dynamical systems. Such model benefits from a rich theory and a well-established collection of tools that makes the analysis of its dynamical properties straightforward, as contrast to detailed mechanistic models.

3. The estimation of the parameters of such models is reliable and computationally efficient.

4. LTIs have a frequency description with an easy visual interpretation, which can be used to infer their stability and performance.

5. LTIs models are highly flexible: the number of hidden variables involved (which can represent non–observable biological species) can be tuned and a best approximation of the system can be estimated with information criterion such as the AIC (Aikake Information Criterion). Furthermore, user-defined time-delays can be introduced to formulate further hypothesis on the nature of the biological process (for example, the time it takes for the regulation to happen) [82].

6. Although they inherently represent idealizations of the real, underlying complex physical processes at play, careful design and considerations based on linear theory led to good results in many different fields.

7. It is not yet clear how non-parametric methods, such as [83, 84] could be used to compare subtle changes in dynamics caused by perturbations, and pinpoint the source of those perturbations.

8. The description of biological mechanisms from time series data by LTI models has been previously studied in [85, 86] and showed that such simple linear black box model representation of circadian networks offers advantages when data are scarce. In addition, [87, 88] correctly predicted previously unknown interactions and design principles within the Arabidopsis oscillator using LTI models. It is important to notice, therefore, that although such an approach does not provide detailed functioning of the network, it is capable of describing gene regulatory interactions with a reliable degree of precision.

Mathematical Formulation

Formally, a LTI model is generally represented by the following set of equations:

$$\begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + Bu(t) + Ke(t) \\ y(t) &= Cx(t) + Du(t) + e(t) \end{aligned} \tag{2.2}$$

where $x(t)$, $u(t)$, $y(t)$ and $e(t)$ respectively represent the internal dynamics of the system, its input and its output, and the inherent white noise of the system and measurements

($e(t) \sim \mathcal{N}(0, \sigma^2)$) [89]. The matrix $A \in \mathbb{R}^{n \times n}$, vectors $B, K \in \mathbb{R}^{n \times 1}$, $C \in \mathbb{R}^{1 \times n}$, and scalar D are parameters of the system.

Here, the strategy consists in representing the transcriptional relationships between each pair of genes, one pair at a time, and identify the most relevant parents genes given both genes expressions (Figure 2.1). As such, the model investigates whether the rate of change of a particular gene $y(t)$ depends on another gene $u(t)$. For this purpose, $u(t)$ and $y(t)$ represent the time series of the gene expression level of parents genes and of the regulated gene, respectively. The LTI representation above represents a direct extension of equation (2.1) in such that only an intermediate state $x(t)$ is added to give the dynamics more flexibility towards the biological processes involved (translation, transcription, etc.). The dimension of the vector $x(t)$ defines the model order: it can be a 1-dimensional vector (direct regulation or relatively slow dynamics compared to internal dynamics), or a multi-dimensional vector (the regulation happens through intermediate steps that introduce delays and cannot be ignored).

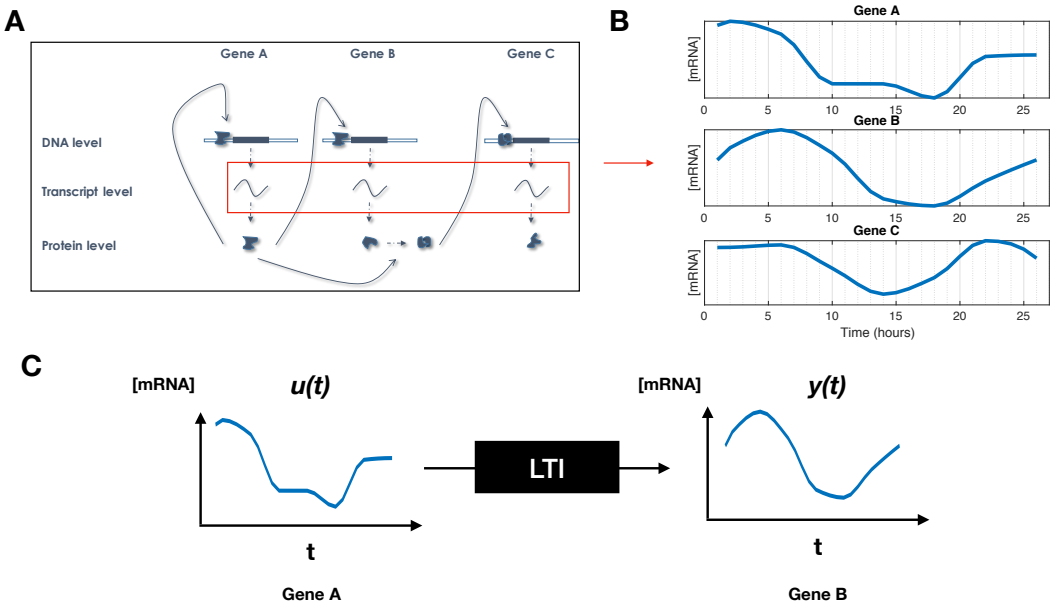

Fig. 2.1 **Mathematical Modeling Strategy. A & B** Genes expression are measured through their mRNA levels. **C** Gene regulation dynamics between genes is represented through Linear Time Invariant (LTI) systems, one pair of genes at a time.This is repeated for all input-output pairs of genes to estimate a network.

Estimating a model means finding A, B, C, D and K that reproduce the dynamics involved with a sufficiently high degree of precision. More specifically, it produces a vector $\hat{y}(t)$ using $u(t)$ and through the estimation of A, B, C, D and K that needs to be as

close as possible to the real data $y(t)$. As for the modeling of gene regulatory networks, the model has been simplified to merely estimating A, B and C. Indeed, explicitly estimating K did not improved the accuracy of the reconstruction and it is further assumed that $y(t)$ is a direct observation of the output gene, so that $D = 0$. Furthermore, C is set so that the first state is the measured gene. Finally, since gene expression measurements always hold a constant offset, an additional state of order 0 is added to account for it. Hence, the model can be rewritten as:

$$\begin{aligned} \frac{dx_{full}(t)}{dt} &= \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} x_{full}(t) + \begin{pmatrix} B \\ 0 \end{pmatrix} u(t) \\ y(t) &= \begin{pmatrix} C & 1 \end{pmatrix} x_{full}(t) + e(t) \end{aligned} \tag{2.3}$$

with

$$x_{full} = \begin{pmatrix} x \\ x_{offset} \end{pmatrix}$$

The identification of the parameters is performed using the function 'pem' implemented in $MATLAB^{TM}$ to minimize the prediction error [89]. A, B and C describe entirely the evolution of the system, which can therefore be predicted. The estimation of parameters requires low computational time: a single system between a pair of genes is typically identified within few seconds (Intel Core i5).

Each potential link between two genes is validated if the corresponding model reproduces the dynamics involved with a sufficient degree of precision, which is characterized by a high goodness of fit, defined as:

$$fitness = 100 * (1 - \frac{\sum_{k=1}^{N} \sqrt{(y - \hat{y}_k)^2}}{\sum_{k=1}^{N} \sqrt{(y - \bar{y})^2}})$$

where y is the validation data, $\bar{y}$ is the average value of the validation data, and $\hat{y}_k$ is the estimated output. MATLAB function *compare* can be used to compute the fitness of the model. A model fitness equal to 100% corresponds to a perfect identification. The choice of such metric is motivated by the dependency of noise towards the abundance of gene expression. When the distance of the true data points towards the mean is large

(represented by the denominator in the above equation), the fitness conveniently penalizes less the error term, which lies in regions where the intrinsic noise involved in the gene expression is potentially the largest.

2.2.3 DyDE Framework

In this section, one of the main contributions of the thesis is introduced. The Dynamical Differential Expression (DyDE) methodology uses the aforementioned LTI modelling framework to reverse engineer the topology of unknown GRNs while at the same time providing a dynamically reliable support to investigate the source of a perturbation (e.g. treatments or drugs) in the network. In particular, this approach does not rely on prior assumptions of the network dynamics and can be applied to short time series data with scarce sampling. In chapter 3, it is shown that the introduced methodology can reliably identify the source of a perturbation in complex regulatory systems such as the circadian network. The proposed mathematical framework is scalable and flexible, so that it can be applied to large datasets.

Identification of Transcriptional Dependencies

The first step of DyDE consists of uncovering dependencies and quantifying dynamics between every genes of the whole network with LTI models. The mathematical framework estimates a collection of Single Input-Single Output (SISO) models between pairs of genes to characterize the system dynamics. The limited number of available time points restricts the modelling of SISO systems to first and second order models. It should be noted that the use of DyDE involves a cubic spline interpolation between data points before estimating the parameters. Gaussian processes have been empirically considered as an alternative way of interpolating the data for further constraining the dynamics of the linear modeling, without noticeable improvement. Furthermore, second order systems did not improve significantly the fitness of models and resulted in a considerable increase of false positives (overfitting). Hence, the model order is restricted to one for the applications considered in the thesis. Second order models will be computed to investigate the complexity of the dynamics given that the link has been validated. For a first order model, the LTI model described the dynamics between two genes can be reduced to:

$$\frac{dy(t)}{dt} = au(t) - by(t) + c$$

where $u(t)$ and $y(t)$ represent the time series of the regulatory gene and the regulated gene, respectively. In addition, $by(t)$ corresponds to the degradation rate of gene y, $au(t)$

corresponds to the influence of $u(t)$ on the rate of $y(t)$ and c is a constant offset. The model has a total of three parameters (a, b, and c), leading to efficient solutions. A subspace initialization algorithm is chosen since it performed similarly as randomizing initial conditions–for the vast majority of models (99%), the final solution was identical with either method. This suggests that the chances of being trapped into a local minimum are negligible.

To reverse engineer the whole gene regulatory network, therefore, this modeling is independently repeated for all available pairwise genes, where each gene takes its turn as being an input and then an output to another gene. This modeling approach, therefore, generates a large amount of SISO LTI models ($n^2 - n$ models, where n corresponds to the amount of genes, and self regulation is not considered) to describe the system. Finally, the fitness of each of the pairwise regulation model subsequently serves as a proxy for the confidence level of the interactions between genes. A thresholding process, therefore, allows to recover the topology of the network (illustrated on Figure 2.2).

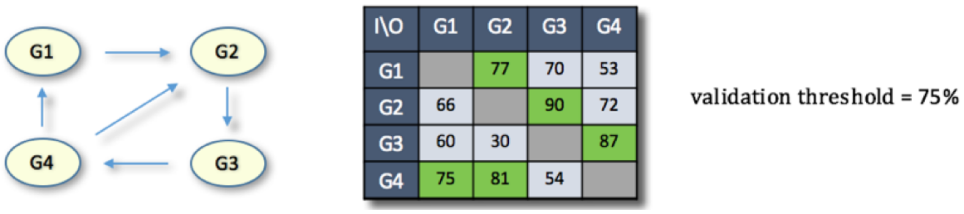

Fig. 2.2 **Network Reconstruction through Fitness Thresholding.** LTI systems are computed between each pair of genes, one pair at a time. The agreement of the model with the observable data is therefore represented by the fitness metric. The fitness values for each pair of genes of the network can be represented as a fitness matrix. A threshold that describes the level of required accuracy of the model to the data is then applied. The topology of the network is defined as the models for which the agreement with the data were above the fitness threshold. In this example, a fitness threshold of 75% has been chosen, so that only the following links are accepted: G1 to G2, G2 to G3, G3 to G4, G4 to G1, G4 to G2.

Assessing Biological Alterations using the ν-gap

The second step consists in identifying the effect of a treatment on the biological network. In general, the discovery of drug modes of action is still a costly and inefficient process, which often requires considerable prior knowledge of a biological system and/or a vast amount of data in several experimental condition (e.g. mutations). In particular, while a

treatment might affect the abundance of many transcripts, only a few links are affected, as depicted in red in Figure 2.3. Indeed, most biological systems have a large number of feedback loops. Hence, a perturbation anywhere in the network typically affects all nodes (in this case, their molecular concentration and time profiles), which makes the problem of inferring the entry point of a perturbation hard using the standard Differential Analysis (DE) of transcripts levels. The main reason is that DE only performs statistical analysis of changes in gene expression levels [23, 90]. Indeed, a complex cascading effect causes large section of the transcriptome to be differentially expressed, despite not being directly affected by the drug.

To capture the cause of the perturbation, it is proposed, instead, that key mechanisms involved in the perturbation can be captured by identification and comparison of regulatory dynamics before and after the perturbation occurred. The rationale behind this approach is that not only genes, but also their interactions, are affected by a drug. The modelling strategy was designed so that it could be applied to scarce data without the need to cover extensive experiments or to make prior assumptions of network dynamics.

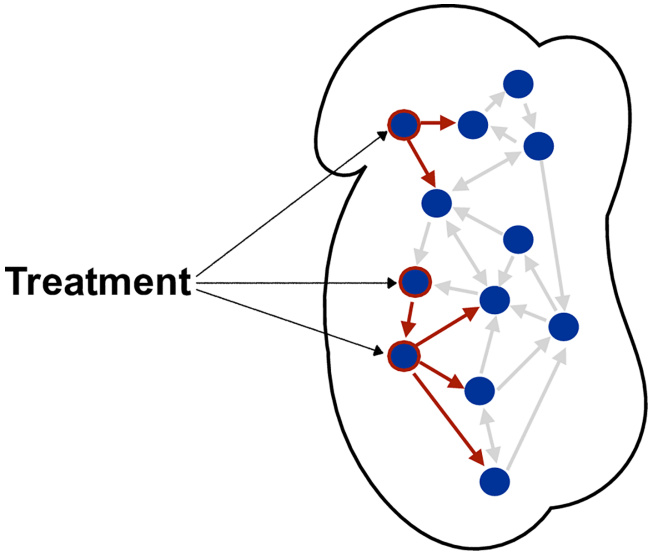

Fig. 2.3 **Treatments effects in transcriptional networks.** Treatment effects can be perceived as perturbations in molecular networks [2, 91]. In transcriptional networks, such perturbations usually only affect a very small number of regulatory links directly. For example, only the red links have been directly affected by the treatment. All other links are unchanged, although all nodes (concentrations) in the Figure have been (indirectly) affected due to cascading and feedback effects. Hence, Differential Expression (DE) might not distinguish between direct and indirect effects of a treatment. Dynamical Differential Expression (DyDE), therefore, investigates how and why changes occur, instead of simply measuring what and how much is produced by those changes.

For this purpose, two cases are of particular interest. First, a link between two genes is validated in the untreated system alone (i.e. it is not possible to find a combination of a, b and c so that the model in the treated system provides a good match with the data anymore). Second, a link is validated in both systems, but the way one gene regulates the other may change; this is a much subtler change in the dynamics of the link. The latter case requires us to compare the dynamics between both links.

Here, a rigorous and well-established tool from engineering known as the ν-gap [92] was used. To understand the control theory concepts underlying the ν-gap, some mathematical notions need to be introduced. When estimating the regulation function between two genes, the resulting model is known as "open-loop", such that the external feedbacks loops controlling both genes expression are not integrated directly in the model computation. Let P represent the equivalent transfer function of the model, the genes are known to be involved in a feedback loop, represented by the controller C, the system can be illustrated by the block diagram on Figure 2.4. Altogether, P and C represent the closed-loop behavior of the system. In fact, the closed-loop behavior of two systems can appear very close although the difference between the open-loop systems can be arbitrarily large.

This can be illustrated with the following example [92], given P_1, P_2, P_3 so that:

$$P_1 = \frac{100}{2s+1} \;\; ; \;\; P_2 = \frac{100}{2s-1} \;\; ; \;\; P_3 = \frac{100}{(s+1)^2} \tag{2.4}$$

P_1 and P_2 are very different. One is stable, the other unstable. Nevertheless, their closed loop behaviors are quite similar (for the example, C = -1):

$$\begin{aligned} \frac{P_1}{1-P_1C} &= \frac{100}{2s+101} \\ \\ \frac{P_2}{1-P_2C} &= \frac{100}{2s+99} \end{aligned} \tag{2.5}$$

Which is confirmed by a small ν-gap; $\delta(P_1,P_2) = 0.02$, while $\delta(P_1,P_3) = 0.89$. Originally developed to compare two linear models from a perturbation perspective, the ν-gap estimates the smallest amount of perturbation that is needed to transform one model into

another. In the context of biological networks, this is of particular relevance since genes regulate each others through multiple interlocked feedback loops. This then facilitates us to determine the significance of the dynamical change of a link between experimental conditions. The ν-gap returns a value between 0 to 1, quantifying whether the models are similar or very different, respectively.

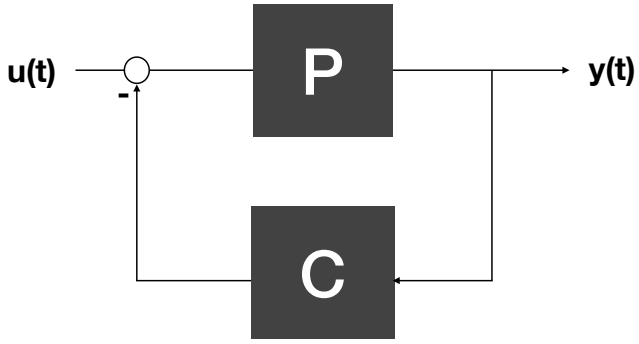

Fig. 2.4 **Block diagram representation of a feedback system.** When a perturbation occurs in the system, the dynamics between genes might be affected. P represent the model as estimated by DyDE while C represent the control effect of the rest of the biological network. Originally developed to address the stability properties of closed loops systems defined in the same feedback loop, the ν-gap essentially measures the distance, from a perturbation point of view, between linear models P.

This approach has been tested with *in silico* data in [93] which subsequently suggested that values above 0.2 could be used to infer the main target of a perturbation. The ν-gap is computed using the *gapmetric* function in MATLAB. It should be applied to all models that have been estimated in both networks.

The ν-gap can in general be computed pointwise in the frequency domain, under mild conditions [94]. Let P_1 and P_2 represent the transfer function of the open loop LTI system before and after a perturbation occurred in the system:

$$\delta_\nu(P_1,P_2) = \sup_w \psi(P_1(j\omega),P_2(j\omega)) \tag{2.6}$$

In particular, for a SISO system:

$$\psi(P_1(j\omega),P_2(j\omega)) = \frac{|P_1(j\omega)-P_2(j\omega)|}{\sqrt{1+|P_1(j\omega)|^2}\sqrt{1+|P_2(j\omega)|^2}} \tag{2.7}$$

Therefore, if the signals are concentrated around a particular range of frequencies (such as oscillating signals, such as circadian of period approx. 24 hours), the gap should be measured 'locally' around that range of frequencies only, since they dominated the model estimation in Step 1.

2.2.4 Example of DyDE application

Next, we explain the key ideas behind DyDE through a small number of genes in the Arabidopsis circadian oscillator. For example, the following model considers *TOC1* as an input and *PRR9* as an output.

$$\frac{d[PRR9](t)}{dt} = a[TOC_1](t) - b[PRR9](t) + c$$

where b represents the strength of activation or repression induced by *TOC1* on the expression rate of *PRR9*, and a corresponds to the degradation rate of *PRR9*. These parameters are estimated by minimizing the prediction error from the untreated time series for both *TOC1* and *PRR9*.

In this case, we found a model in good agreement with the data (57% fitness), suggesting that indeed *TOC1* regulates *PRR9* (Figure 2.5). Moreover, the model demonstrates that the rate of change of the concentration of *PRR9* is proportional to the concentration of *TOC1*. Note that the other way around (i.e., *PRR9* regulating *TOC1* could not be established since the respective model has a low goodness of fit (16%, Figure 2.5B). These results are consistent with the literature [33]. Hence, we would then establish a link from *TOC1* to *PRR9*, but not the other way around (Figure 2.5C).

Then, a model is estimated between *TOC1* and *PRR9* from the NAM-treated time series. From the untreated and treated time series alone, it is unclear whether the link dynamics have changed (Figure 2.5D). The optimal model parameters, however, have significantly changed. A ν-ugap of 0.5 confirms that indeed the link has been affected. This result indicates that there is large perturbation in the regulatory dynamics that links *TOC1* to *PRR9*, which, therefore, should be considered as a strong candidate for being an entry point for NAM in the system. If true, knocking down either *TOC1* or *PRR9* would therefore lead to NAM no longer affecting the clock. This analysis is then repeated for all common links between untreated and treated plants.

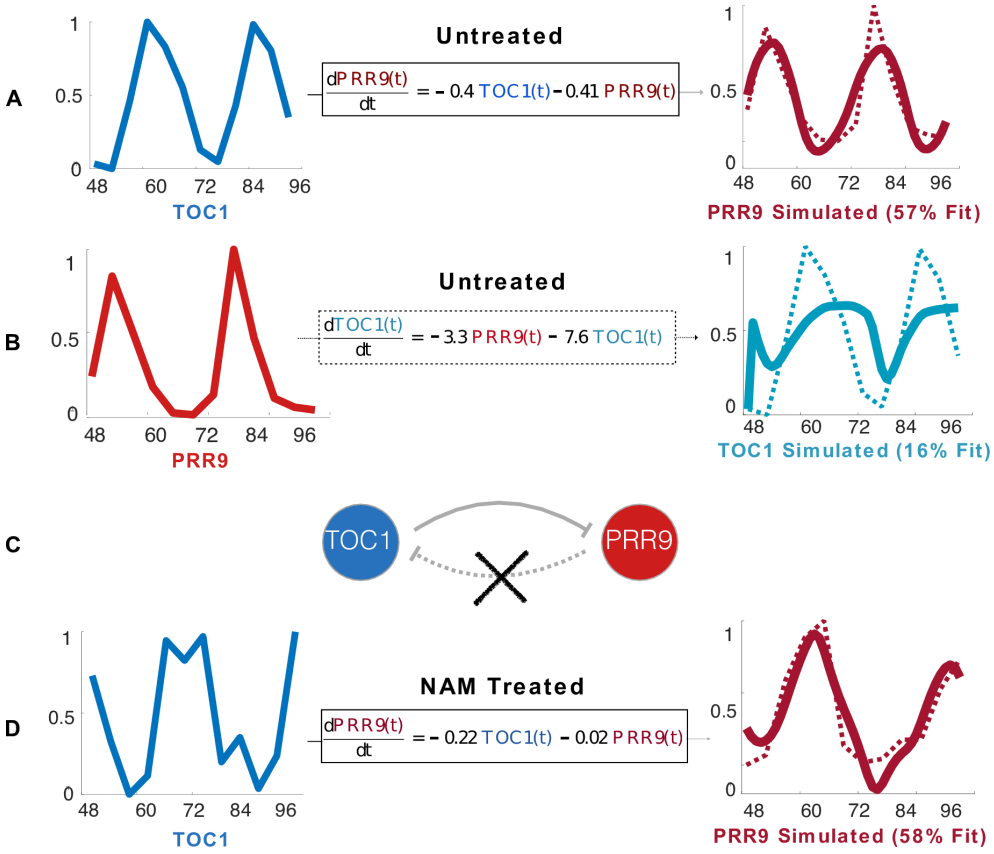

Fig. 2.5 **Network inference and analysis by dynamical differential expression (DyDE). (A)** LTI system captures the dependence of the rate of the concentration of a transcript on the concentration of another transcript. First order linear models are used to represent the dynamics between two genes. Here, a good agreement (plain line) with the data (dotted line) was found (57% goodness of fit). **(B)** The inverse regulation is considered. In this case, it is not possible to find a combination of parameters so that a first order linear model captures the dynamics involved. For this inverse regulation the model that best described the data obtained a goodness of fit of only 16%. **(C)** A threshold by which each model is (in)validated is applied on the goodness of fit of the models. As an example, a threshold of 46% would consider a link from *TOC1* to *PRR9* but not the other way around. The same threshold is applied to all models. **(D)** A first order linear model is evaluated in the presence of nicotinamide between the same species. The ν-gap is then applied to compare models **(A)** and **(D)** to quantify whether the models are similar, or significantly affected by NAM

2.3 Optimal Experimental Design and Multifactorial Benchmarking for GRN Inference

Adapted from:

- Mombaerts L et al. Optimising time series experimental design for modelling of circadian rhythms: the value of transient data. FOSBE 2016. [95]
- Mombaerts L. et al. A multifactorial evaluation framework for gene regulatory network reconstruction. FOSBE 2019. [96]

2.3.1 Introduction

Hereafter, the performance of the DyDE framework and recent and successful network inference strategies are assessed under a novel, multifactorial evaluation framework in order to highlight pragmatic trade-offs in experimental design and network reconstruction. Realistic time series datasets are generated from one rhythmic and five non-rhythmic models of gene regulatory networks that have been widely used as benchmarks in the literature [27, 26]. External interventions, i.e., gene deletion (knock-out) and chemical treatments, are explicitly simulated to provide a comprehensive picture of the performance of each algorithm under a range of experimental conditions. Then, increasingly rich multi-experiment time series datasets are provided to five state-of-the-art network inference algorithms representing distinctive mathematical paradigms. Performance is assessed using standard techniques for classification algorithms, studying the area under the receiver operating characteristics (ROC) and the precision-recall (PR) curves.

Navigating experimental trade-offs for GRN inference is not an entirely new concern, although literature on the topic is surprisingly scarce [97]. Such a multifactorial approach has never been undertaken for systematic evaluation of network reconstruction algorithms. [98] studied the trade-offs between dense and replicate sampling strategies. Their results showed that, under reasonable noise assumptions, gene expression profiles reconstructed from dense sampling are more accurate than those resulting from replicate sampling. [99] showed that at equivalent data size, short-time, gene knock-out experiments contain more information about the GRN structure than single experiment, longer recordings of non-rhythmic systems. The GRN inference algorithms used in their study, however, are no longer state-of-the-art. [100] studied the cell cycle of Saccharomyces cerevisiae as a case-study to analyze the effect of temporal resolution on the quality of the inferred network. The performance as a function of time series length resulting from a LASSO methodology [101] resembled a sigmoid shape with a plateauing effect at the

end. [102] identified previously unrecognized factors that affect inference outcomes, such as stimulus-specific experimental design and network motifs in the vicinity of a stimulus. Following the DREAM3 competition, [103] investigated strengths and weaknesses of algorithms in recognizing types of motifs that appear in gene networks. Finally, it has also been shown that, for mutual-information based techniques, the accuracy reaches a saturation point after a specific data size [104]. Algorithms based on correlation or mutual information are, however, excluded from this research as they cannot detect causality between genes [26].

In this analysis, it is reported that the algorithms do not benefit equally from data increments. Furthermore, for rhythmic systems, it is more profitable for network inference strategies to be run on long time series rather than short time series with multiple perturbations. In the case of circadian networks, it is further noted that the transitory regime that follows the switch to a new constant condition (such as constant light) has the potential to shortly reveal supplementary dynamics between genes that constitute the network.

By contrast, for the non-rhythmic systems, increasing the number of perturbation experiments yielded better results than increasing the sampling frequency. It is expected that future benchmark and algorithm design would integrate such multifactorial considerations to promote their widespread and conscientious usage.

2.3.2 Generation of Realistic Data

The use of *in silico* networks is preferable over random graphs, as they account for realistic structural properties of biological networks [105]. For example, although randomly generated networks display approximately the same power-law degree distribution of regulatory interactions, they often fail to represent important properties such as the modularity [106] or the occurrence of network motifs, which are statistically overrepresented in real complex biological networks [107]. For this analysis, one rhythmic and five non-rhythmic models of gene regulatory networks that have been widely used as benchmarks in the literature [27, 26] are used to produce realistic time series data of gene expression. The dynamical models considered for benchmarking all rely on common modeling specificities and on highly nonlinear equations explicitly integrating protein dynamics (although not observable in practice), such that gene expression is computed as:

$$\begin{aligned} \frac{dy_i}{dt} &= m_i f_i(\pi_i(t)) - \alpha_i y_i \\ \frac{dx_i}{dt} &= r_i y_i - \beta_i x_i \end{aligned} \tag{2.8}$$

where m_i corresponds to the maximum transcription rate and r_i the translation rate. α_i and β_i represent both mRNA and protein degradation rates, respectively. f_i is of the Hill or Michaelis-Menten type formula for mass-action kinetics such that:

$$f_i = \frac{I_{u,i} x_u^a(t) + (1 - I_{u,i}) k_{u,i}^a}{x_u^a(t) + k_{u,i}^a} \tag{2.9}$$

where $k_{u,i}$ corresponds to the Michaelis-Menten parameters and $I_{u,i}$ indicates whether the regulation is an activation ($I_{u,i} = 1$) or an inhibition ($I_{u,i} = 0$). x_u represent protein levels and π_i the ensemble of parents genes for gene i. Indeed, mRNA levels are not influenced directly by other mRNA levels but through their respective protein products. Finally, a represent the Hill exponent (typically equal to 2 in most models).

Furthermore, all simulations were performed based on Langevin equations (stochastic differential equations) to account for the Brownian motions of chemical species and represent the intrinsic stochasticity in the dynamics of gene regulatory networks (accounting for molecular noise in both transcription and translation processes) [108]. Given an infinitesimal updating, the standard form of the multivariate Langevin equation describes the random fluctuations of the molecular species and their evolution over time in a "well-stired" system of interest. Formally, it can be represented by:

$$\frac{dX_i(t)}{dt} = \sum_{j=1}^{M} \nu_{ji} a_j(X(t)) + \sum_{j=1}^{M} \nu_{ji} a_j^{1/2}(X(t)) \Gamma_j(t) \tag{2.10}$$

with $i = 1, ..., N$ the amount of species in the system and M the amount of reaction channels. The uppercase X_i represents the number of molecules i in the system (in this case, both mRNAs and proteins levels). M accounts for the molecular dependencies (parents genes or proteins). a_j is referred to as the propensity function. It measures the propensity for reactions to occur in the next moment and is therefore inversely proportional to the total system volume Ω, which in turn can be tuned so that it matches the noise magnitude observed in real biological systems. ν_{ji} represent the "state-change",

or stoichiometry vector. It can admit 0 entries if a molecular species does not participate in the reaction. The term $\Gamma_j(t)$ represents temporally uncorrelated, statistically independent Gaussian white noises formally defined as [108]:

$$\Gamma_j(t) \equiv \lim_{dt \to 0} \mathcal{N}(0, 1/dt) \tag{2.11}$$

The intrinsic noise is expected to have a significant impact on the behavior of the system [109]. Then, data are downsampled to resemble realistic experimental design (every 4 hours in the case of circadian experiments). Finally, the protein concentrations are not made available in the provided datasets (only mRNA concentrations levels), as current high-throughput technologies typically do not allow the monitoring of protein expression [110]. Auto-regulatory interactions are not included. The models are hereafter introduced.

The Millar Group Model of Circadian Regulation

The rhythmic gene regulatory system used as a benchmark here is a model of circadian regulation of Arabidopsis thaliana, hereafter referred to as Millar 10 (Figure 2.6A) [31]. This model describes the central circadian oscillator through the modelling of 8 genes and the intervention of several intermediate transcription factors. Moreover, regulatory interactions are either additive or multiplicative, while proteins might undergo post-translational modifications. As such, many levels of complexity encountered in real biological systems are represented. The parameters of this model were typically investigated experimentally or via simulated annealing over years of experiments.

Conceptually, it is composed of 2 interconnected feedback loops and an input pathway that incorporates external light cues in order to synchronize the plant to the surrounding light conditions. When subjected to another external regime, e.g., constant light or constant darkness, the system displays transient dynamics and reaches a new limit cycle characterized by the free-running conditions of the circadian network. The transitory regime that follows the switch to a new condition (either constant light or constant darkness) can be short before establishing the new regime. However, this time window may have the potential to shortly reveal supplementary dynamics between the genes that constitute the network. To support this hypothesis, the performances of the network inference strategies are first analyzed under such transient dynamics. For this purpose, the model has been initially simulated for 240 hours in light/dark cycles to remove initial system transients and then switched to constant light regime. Time windows of 48 hours

are then extracted from the light/dark limit cycle, at the transition to constant light and 48 hours after transition to constant light for further comparison. In the subsequent analysis, only time windows starting from the transition to constant light and up to 3 days of observations (24-36-48-60-72 hours) are considered.

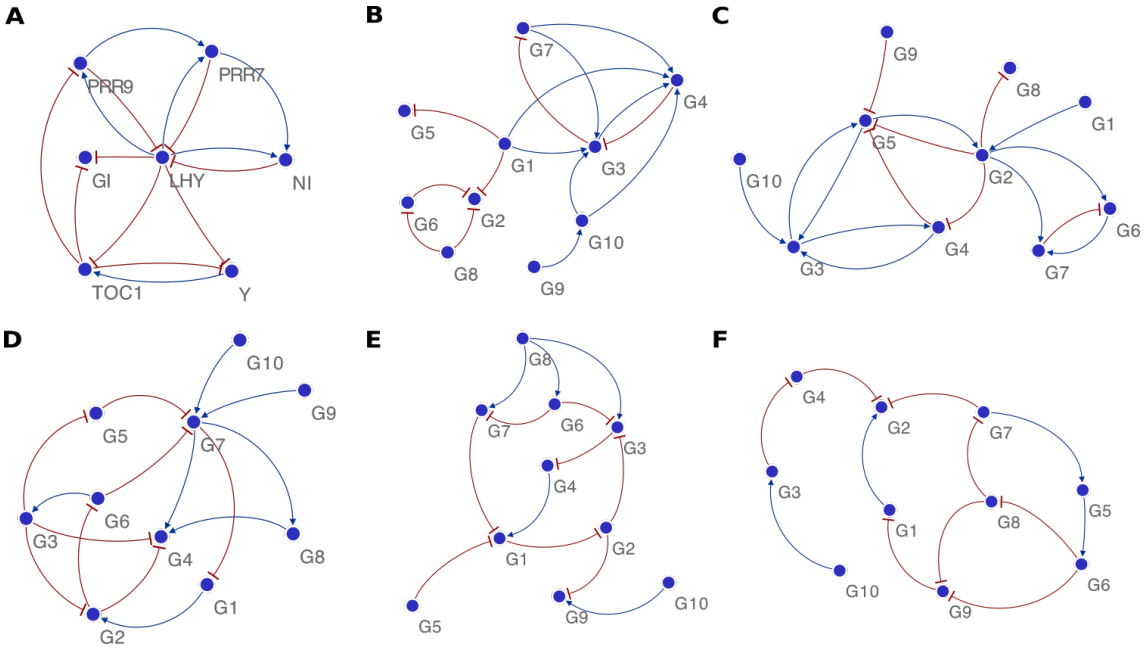

Fig. 2.6 **Gene regulatory networks used as benchmarks.** Blue pointed arrows and red blunt arrows represent activation and inhibition reactions respectively. **A.** Millar 10 Model (Rhythmic) **B.–F.** DREAM4 models 1–5

The Millar 10 model has been simulated to reproduce gene knock-out experiments. Knock-out experiments are very informative, more than knock-down experiments, as they provide network response to individual and large perturbations (genes are deleted) [26]. Knock-out experiments were simulated as in [111], by replacing the transcription rates of the targeted genes by random noise drawn from a truncated normal distribution to ensure non-negativity of the concentrations. Genes that have been knocked out are, therefore, not influenced by their structural regulators anymore. The datasets in the numerical experiments consist of a wildtype (WT) time series, and up to three randomly chosen knock-out time series at a time. This selection has been randomized 6 times to account for the uneven informative potential of different genes in the network. Furthermore, such simulations being stochastic, each experiment was replicated 10 times to provide a representative view of the performance of each computational method. In total, $950 = 10 \cdot 5 + 10 \cdot 6 \cdot 5 \cdot 3$ simulations were performed.

The DREAM Competition Models

The studied non-rhythmic models (Figure 1 B–F) originate from the DREAM4 *in silico* network inference challenge [103, 26, 112]. The structural properties of those networks are extracted from the global interactions network of E. Coli, and therefore represent biologically realistic networks (groups of genes are more connected than expected in random graphs and cycles were preferentially extracted) [113]. In such case, a mix of normal and lognormal noise is added to the overall gene expression time series data to represent measurement noise. New time series data were generated by the GeneNetWeaver software [105] which simulates the system's response to a perturbation of about a third of its nodes, followed by a relaxation back to steady state after half of the recording, when the perturbation was removed. The data characteristics are as in the challenge, with the exception that the perturbation targets were randomized, whereas in the challenge data they were preferentially carried out to cover the whole network [26].

Such data simulation offers a realistic representation of the undetermined effects of a chemical treatment on a system at rest (steady-state). The resulting data are then resampled using 3 different sampling rates (11-21-41 datapoints) to further assess the effect of changes in experimental design. Here, 10 chemical perturbations were simulated for each of the 5 available networks and replicated 3 times. Then, increasing amount of perturbation time series (up to 4 at a time) are randomly selected from those 10 generated perturbations, and provided to the inference algorithms. This random selection is performed 5 times. The perturbation targets are not known to the methods. In total, $900 = 3 \cdot 5 \cdot 3 \cdot 4 \cdot 5$ simulations were performed.

2.3.3 Network Inference Techniques

The network inference methods included in the comparison represent the function f of Equation (2.1) at various levels of complexity and through entirely different mathematical paradigms. They are Gaussian Process Dynamical Models (GPDM) [114], dynamical GEne Network Inference with Ensemble of trees (dynGENIE3) [83], Algorithm for Revealing Network Interactions (ARNI) [115] and Improved Chemical Model Averaging (iCheMA) [27]. The modeling framework involved in DyDE has been extended to handle multi-experiments data and is also included, hereafter referred to as All-to-All (ATA).

All-to-All (ATA)

The network inference methodology developed in this thesis to reverse-engineer the structure of GRNs with comparable dynamics has been extended to deal with multi-

experiments datasets. The datasets are merged together so that the dynamics to be identified are identical through all experimental conditions, assuming that the signal-to-noise ratios are similar in different experiments. The fitness score over multiple experiments is defined as:

$$fitness = 100 * (1 - \frac{\sqrt{\sum_{t \in T^1} (y_1(t) - \hat{y}_1(t|\theta))^2} + ... + \sqrt{\sum_{t \in T^n} (y_n(t) - \hat{y}_n(t|\theta))^2}}{\sqrt{\sum_{t \in T^1} (y_1(t) - \bar{y}_1(t))^2} + ... + \sqrt{\sum_{t \in T^n} (y_n(t) - \bar{y}_n(t))^2}}) \tag{2.12}$$

where n represents the amount of experimental conditions, T the sampling times of experiments, y the gene expression level, $\hat{y}$ the modeled gene expression, and $\bar{y}$ is the average value of the gene expression level.

Gaussian Process Dynamical Models (GPDM)

GPDM, a non-parametric method, models gene expression as a nonlinear stochastic differential equation:

$$\begin{cases} y_j = x(t_j) + v_j \\ dx = g(x, \theta)dt + du \end{cases} \tag{2.13}$$

where the dynamics function g is modeled as a Gaussian process with some hyper-parameters θ, v_j's correspond to measurement errors, and u is a Brownian motion. This defines gene expression as a stochastic process whose realizations can be sampled using a Markov Chain Monte-Carlo (MCMC) strategy. Network inference is based on estimating the hyperparameters of the covariance function of the GP. Multi-experiments are taken into account by assuming that all time series are produced by the same dynamics function g. Independent samplers are then constructed for trajectories x corresponding to different experiments. Performance of GPDM was recently compared to the best performers of the DREAM4 challenge and consistently shown superior in dealing with short time series data [114].

dynamical GEne Network Inference with Ensemble of trees (dynGENIE3)

The semi-parametric method dynGENIE3 is an adaptation of the GENIE3 method for time series data. GENIE3 was the best performer in the DREAM4 Multifactorial Network Inference challenge and the DREAM5 Network Inference Challenge [116]. The

transcription function f_i in Equation (2.1) is represented by an ensemble of regression trees which is estimated from the gene expression data and their derivatives, estimated using a difference approximation [83], so that Eq. 2.1 becomes:

$$\frac{y_i(t_{k+1}) - y_i(t_k)}{t_{k+1} - t_k} = f_i(\pi_i(t)) - \alpha_i y_i(t) \tag{2.14}$$

The gene regulatory network inference problem is here casted as a feature selection problem, estimated from the following learning sample:

$$LS^i_{TS} = \{(y_i(t_k), \frac{y_i(t_{k+1}) - y_i(t_k)}{t_{k+1} - t_k} + \alpha_i y_i(t_k)), k = 1, .., N-1\} \tag{2.15}$$

This semi-parametric approach provides a greater flexibility to the inference framework but complicates the comparison of dynamical properties between experimental conditions.

Algorithm for Revealing Network Interactions (ARNI)

ARNI is a recently developed method used for the estimation of network topologies that performed well in network inference from a large collection of short time series [115]. The derivatives are estimated explicitly through a difference approximation and the relationships between nodes in the network are estimated by solving a nonlinear regression problem, with a user-selected library of nonlinear basis functions.

$$\begin{aligned}\frac{dy_i(t)}{dt} = &\sum_{j=1}^{N} \Lambda_{ij} g_{ij}(y_j) + \sum_{j=1}^{N}\sum_{s=1}^{N} \Lambda_{ij}\Lambda_{is} g_{ijs}(y_j, y_s) \\ &+ \sum_{j=1}^{N}\sum_{s=1}^{N}\sum_{w=1}^{N} \Lambda_{ij}\Lambda_{is}\Lambda_{iw} g_{ijsw}(y_j, y_s, y_w) + ...\end{aligned} \tag{2.16}$$

Where g_i are basis functions. $\Lambda_{ij} \in \{0,1\}^{N\times N}$ represents the dependency matrix of the GRN so that:

$$\Lambda_{ij} = \begin{cases} 0 & \text{if } \dfrac{\delta f_i}{\delta y_j} \equiv 0 \\ 1 & \text{if } \dfrac{\delta f_i}{\delta y_j} \neq 0 \end{cases} \tag{2.17}$$

Where f_i is as in Equation (2.1). In the experiments, polynomial basis functions of degree at most 3 were used. In essence, this approach shares an important commonality with the network inference strategy developed in this thesis in such that it decomposes the dynamics between biological species of the network into pairwise dynamical units. The difference lies in the fact that more complex basis functions are typically considered here with higher order hypernetwork interactions, together with a parameter selection based on the enforcement of group sparsity via a greedy approach known as the Block Orthogonal Least Squares (BOLS) algorithm [117].

Improved Chemical Model Averaging (iCheMA)

iCheMA is a semi-mechanistic model that estimates the derivatives from the data by fitting a smooth Gaussian process to the time series. Then, genes expressions profiles are modeled using explicitly the Michaelis-Menten formula for mass-action kinetics:

$$\frac{dy_i(t)}{dt} = \sum_{u \in \pi_i} v_{u,i} \frac{I_{u,i} x_u(t) + (1 - I_{u,i}) k_{u,i}}{x_u^a(t) + k_{u,i}} \tag{2.18}$$

Network inference is then based on estimating the parameters using an MCMC approach. iCheMA goes exhaustively through all possible combinations of regulators (typically, up to 3 at a time), which makes it a computationally heavy algorithm that does not scale easily to large systems. Nevertheless, when provided with a large amount of experimental conditions (11 experiments), it was revealed as the best performer of a set of established network reconstruction algorithms applied to the inference of circadian-type regulatory networks.

Summary of GRN reconstruction strategies

Table 2.1 summarizes the properties of the methods included in the benchmark. A method is deemed a continuous-time method if it is based on continuous trajectory-fitting, or modeling from a continuous-time system. Methods estimating derivatives from the data and then solving input-output regression are deemed discrete-time methods. Combinatorial effects mean dynamics of the form $\dot{y}_i = f(y_j, y_k)$ where $f(y_j, y_k)$ cannot be represented as a sum $f(y_j, y_k) = g(y_j) + h(y_k)$. The table shows whether the methods explicitly take into account combinatorial effects. The computational time is based on 48 hours recordings (13 datapoints) of the Millar 10 model. Performance ranking is based on average AUPREC values.

Table 2.1 Summary of the properties of the different network inference methodologies introduced.

Method	Nonlinear dynamics	Continuous-time	Combinatorial effects	Hidden nodes	Computation time (s)	Performance Millar/DREAM
All-to-all		✓		(✓)[1]	49.4	2/5
GPDM	✓	✓	✓		333.4	1/1
dynGENIE3	✓		✓		0.7	4/2
ARNI	✓		(✓)[2]		1.0	3/3
iCheMA	✓		(✓)[2]		1999	5/4

[1]If higher-order dynamics are used.
[2]Discussed in the article, but not in the implementation.

2.3.4 The Value of Transients Data for Modelling Circadian Rhythms

The performances of each algorithm are here assessed in terms of the resulting Area Under the Receiver Operating Characteristic (AUROC) and the Area Under the Precision-Recall (AUPREC). Precision-Recall curves allow for a more accurate picture of algorithms performances for sparse GRNs and is commonly used for the comparison of inference algorithms. Auto-regulatory interactions are not considered.

Decomposing the time series resulting from the rhythmic model into synchronized-desynchronized states showed that, on average, the accuracy of the network reconstruction is improved by considering transient dynamics (Figure 2.8). While GPDM, ATA, and dynGENIE3 benefit —to a varying degree— from the transition to the desynchronized state, change in performance of iCheMA was not statistically significant and ARNI's performance was slightly impaired. It should be noted that a significant increase in accuracy is observed for the strategies that do not explicitly estimate derivatives. An example of the resulting ROC and PR curves is shown for the ATA strategy on Figure 2.7.

2.3.5 Experimental Tradeoffs and Optimal Strategy

Figures 2.10 displays the performance of each algorithm resulting respectively from the simulations with data from the Millar 10 model and the steady-state systems under several combinations of data types.

On one hand, these graphs show that GPDM outperforms the other approaches for almost every system and experimental configuration considered. It is outperformed by ATA in only one case with 24h wildtype only data from the Millar 10 model. This illustrates the importance of various experimental scenarios in benchmarking network inference

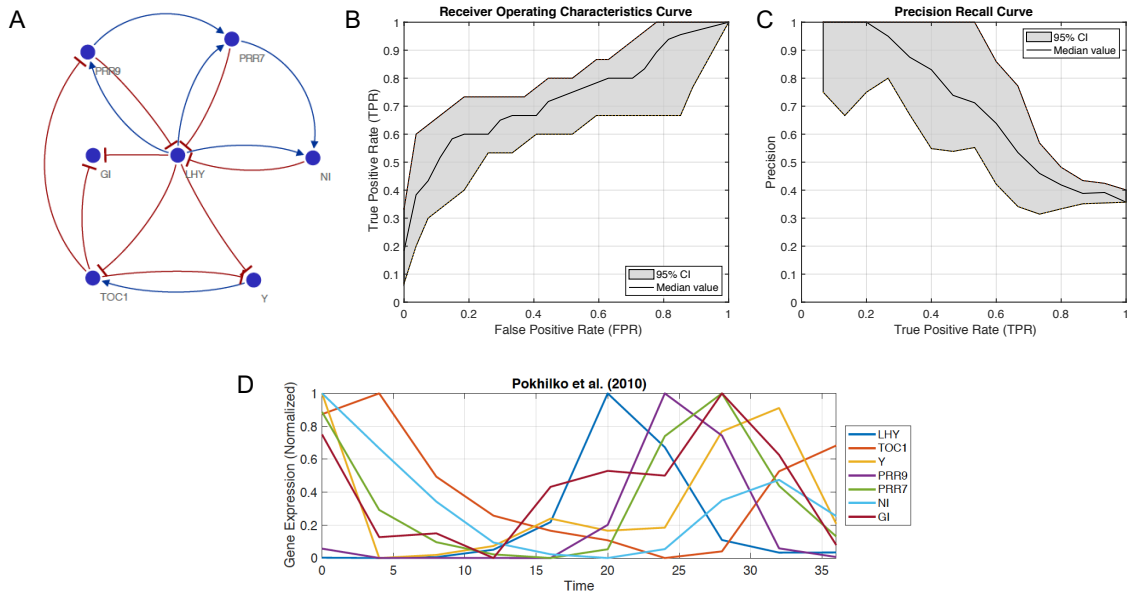

Fig. 2.7 **Example of results for the All-to-All on the Millar 10 model.** (A) The circadian network that serves as a benchmark. (B & C) The ROC and the Precision-Recall curve resulting from the application of the All-to-All to the time series gene expression data, respectively. The grey area represent the 95% confidence interval and is a result of the stochastic simulations. (D) An example of transients data for one run of stochastic simulation of the Millar 10 model (downsampled to one data point every 4 hours).

strategies and motivates the work undertaken in this paper. Interestingly, the simple pairwise low order linear modeling (ATA) seems to outperform dynGENIE3, iCheMA and ARNI in terms of AUPREC for every observation length and system perturbation considered in the rhythmic model. Only the AUROC values of the non-parametric approach ARNI exceed those of ATA for the 3 mutations case, starting from 36 hours of observation.

It is further interesting to notice that not all algorithms benefit equally from data increments. The gain of accuracy resulting from increasing the amount of data in the rhythmic model is only mild for the linear modeling strategy and iCheMA while it is significant for GPDM, dynGENIE3, and ARNI. In this regard, in average, GPDM benefits from the largest increase in accuracy whereas dynGENIE3 and ARNI compete at a slightly lower level for the experimental conditions presented. A saturation effect, however, can be observed at AUPREC values of around 0.8 for GPDM, 0.63 for the ATA, and 0.58 for ARNI.

The analysis of the DREAM competition models delivers a different view on network reconstruction as not all nodes are stimulated in a given system perturbation. For those

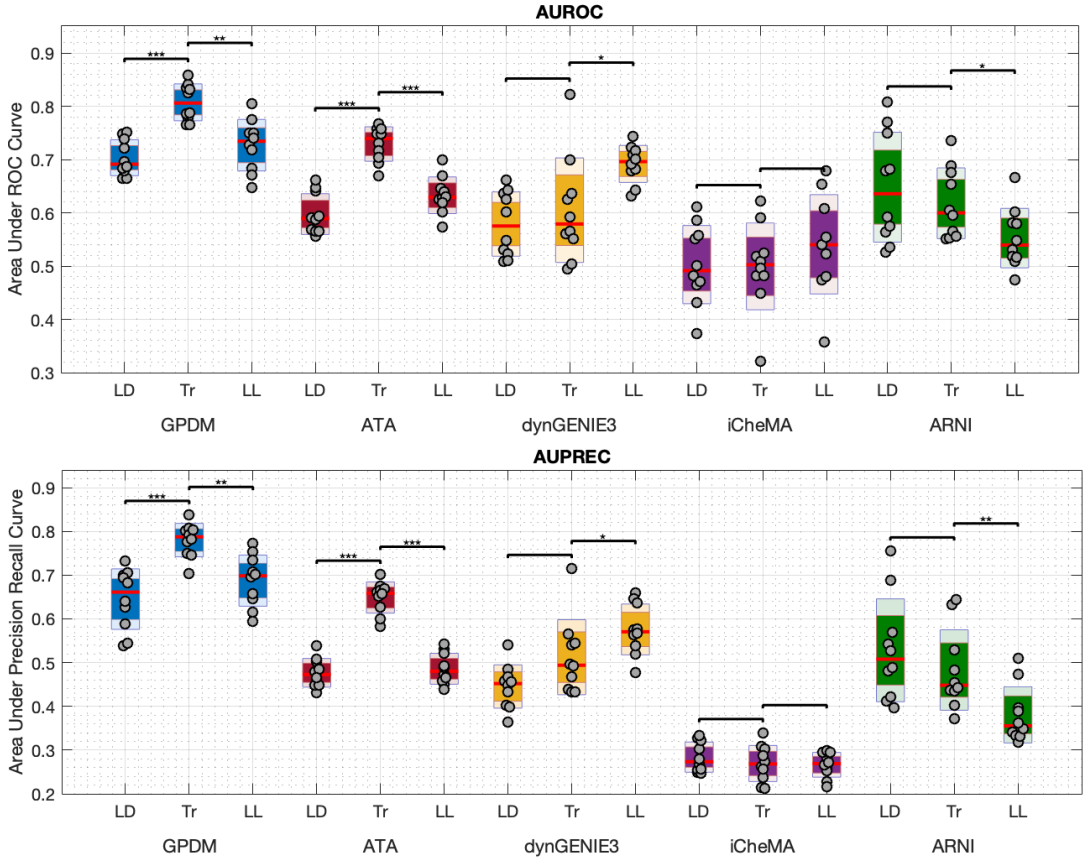

Fig. 2.8 **Evaluation of the effects of dynamical transients on the rhythmic model (Millar 10).** Statistical significance between predictions from transient dynamics and other conditions are indicated by star symbols. They are computed with the Mann-Whitney U-test (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$). The different light conditions are LD: 48h with two light-dark cycles, Tr: 48h constant light, starting right after the last dark period, LL: 48h constant light starting 48h after the transition from regular cycles to constant light

networks, the benefit of additional system perturbations is considerable as they allow investigation of novel, previously unstimulated segments of the network. In the experimental design cases presented, none of the algorithms seemed to approach a saturation point for the data combinations considered. While the GPDM succeeds in providing the best accuracy for the DREAM networks as well, dynGENIE3 ranks second, ARNI third, iCheMA fourth and ATA last. The reason why the linear modeling strategy is surpassing dynGENIE3 and ARNI for the 1 perturbation only case and does not improve for additional datasets is likely related to the partially stimulated nature of the whole dynamical system and has yet to be investigated.

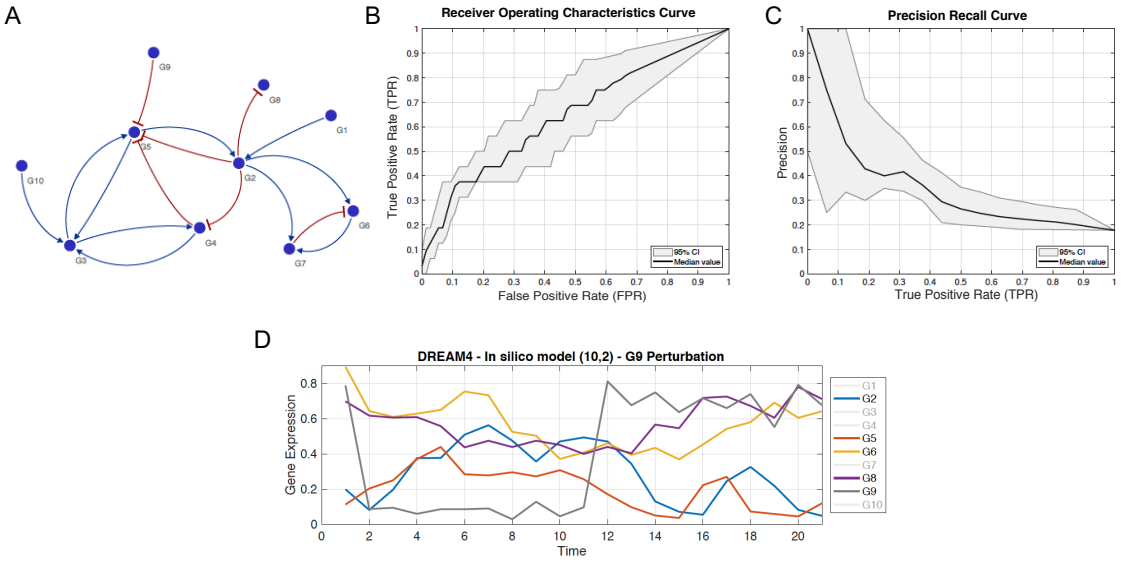

Fig. 2.9 **Example of results for the All-to-All on a DREAM Model.**(A) The DREAM model that serves as a benchmark. (B & C) The ROC and the Precision-Recall curve resulting from the application of the All-to-All to the time series gene expression data, respectively. The grey area represent the 95% confidence interval and is a result of the stochastic simulations. (D) An example of transients data for one run of stochastic simulation of the DREAM model 2 with a perturbation on node 9.

On the other hand, Figure 2.10 allows for proper visualization of experimental trade-offs. Doubling the amount of datapoints by performing another experiment does not provide the same level of information than doubling the amount of datapoints in a given experimental setup. Table 2.2 summarizes the amount of datapoints in each of the presented experimental setups. While the cost of each datapoint might not be equivalent whether it originates from a novel system perturbation or from longer recording, these tables provide insight on how to choose an appropriate experimental scenario regarding the performances of each of the algorithms presented in Figure 2.10. For instance, sequencing a gene in WT every 4 hours during 72 hours requires a similar amount of datapoints as the WT with 2 mutations for 24 hours or the WT with 1 mutation for 36 hours. In this case, the experimental design that provides the best results would be a single recording of 72 hours resulting in an AUPREC of 0.84 using GPDM, compared to 0.75 or 0.7.

Regardless of the algorithm and assuming an equivalent cost per datapoint, it can be observed that, as a general rule of thumb and for top performing strategies, it is often preferable to observe the rhythmic system for a longer amount of time. By contrast, increasing the sampling frequency of the steady-state systems only resulted in a marginal

improvement in the accuracy of network reconstruction. Surely, a lower bound on the sampling rate is required for a reliable construction of those systems but it was not reached in this analysis.

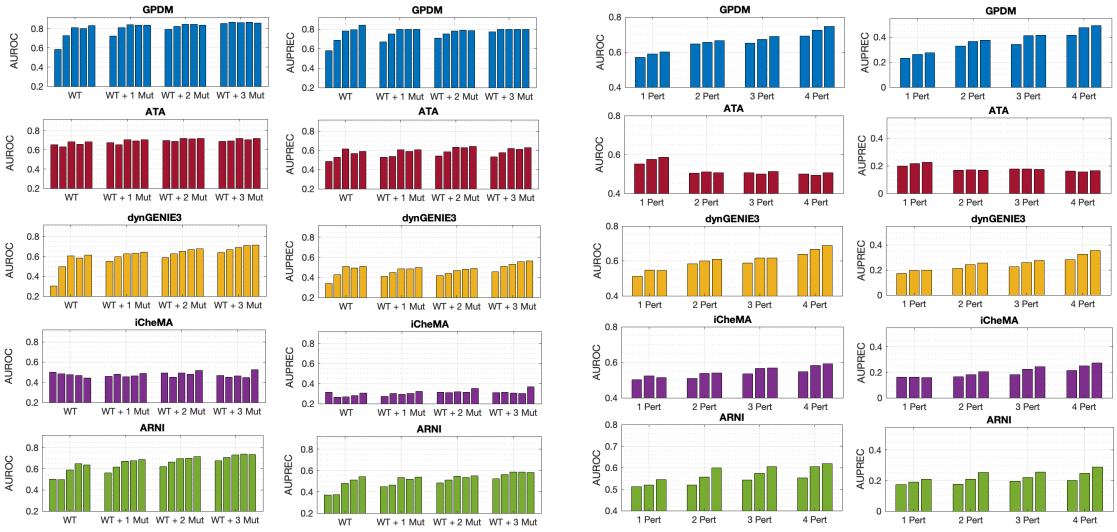

Fig. 2.10 **Area Under the ROC curve and the Precision-Recall curve resulting from the inference of the (a) Millar 10 Model (b) DREAM models, for multiple combinations of observation lengths and system perturbations.** (a) The bars are grouped by the number of additional recordings resulting from perturbations applied to the system (up to 3) and decomposed into data observation lengths of [24-36-48-60-72] hours (from left to right). (b) The bars are grouped by the amount of systems perturbations (up to 4) and decomposed into data quantity of [11-21-41] datapoints (from left to right).

Table 2.2 **Experimental value.** Left: Numbers of measurements in the Millar 10 experiments. Sampling rate is always 4h. Right: Numbers of measurements in the DREAM experiments. Window length is always 20.

Window (h)	24	36	48	60	72
WT	7	10	13	16	19
WT + 1 Mut	14	20	26	32	38
WT + 2 Mut	21	30	39	48	57
WT + 3 Mut	28	40	52	64	76

Δt	2	1	0.5
1 Pert	11	21	41
2 Pert	22	42	82
3 Pert	33	63	123
4 Pert	44	84	164

2.4 Discussion

DyDE/ATA as a network inference strategy

Undoubtedly, the concept of using a linear modeling strategy for the identification of gene regulatory dependencies is not new. However, the implementation design and the general GRN modeling framework presented in this chapter are new and appear to be particularly suited for the problems considered. A comprehensive analysis of its performance has been performed under various experimental conditions and compared to state-of-the-art network inference algorithms. In the current network inference literature, such steps are often undertaken by roughly comparing few strategies, or algorithms such as those based on correlation or mutual information that do not belong to state-of-the-art anymore. A notable exception is the attempt undertaken by the DREAM competition, 7 years ago. Even so, such comparisons are often biased in the sense that if more than one, only few experimental designs are considered. Therefore, the presented performance might be misleading and as a result, the applicability and suitability of the proposed algorithm remain unclear for most practical cases. Hence, it is not rare to witness applications papers that investigate the performance of multiple algorithms from scratch in the specific context of their experimental design and biological process. In fact, a thorough comparison of state-of-the-art network inference algorithms was not found in the current literature. Eventually, a benchmark framework that provides both computational and experimental scientists with comprehensive trade-offs was lacking. This is the main reason for which a step in this direction has been taken to provide a suitable evaluation framework together with an updated, state-of-the-art comparison of network inference algorithms under various realistic experimental designs. The field has now reached a certain maturity, so that the relative performance of the mathematical paradigms presented in this contribution are expected to remain conclusive.

In the light of the results presented in this chapter, the linear modeling approach developed has been shown as an extremely valuable tool. Being very flexible with a large supporting literature, linear models have shown that not only their performance in recovering the structure of GRN from short time series data can impressively compete, but even outperform, most of the latest algorithms for most of today's real case experiments. Moreover, they can also be interpreted and manipulated in such ways that are currently not possible with nonlinear or nonparametric models. In this respect, the comprehensiveness of linear models remains a major advantage for many practical applications.

Finally, the modeling strategy presented here has the considerable advantage of being applicable to very large datasets, or "big data", monitoring the gene expression of

thousands of genes at a time. Indeed, DyDE scales very easily as a result of its pairwise dynamical investigation.

It should be highlighted that the addition of static nonlinearities to the linear framework developed here has been considered but the preliminary results were not satisfactory.

Optimal resources management for network reconstruction

Choosing between different experimental designs and network inference strategies depends on the research question and the resource constraints. For this purpose, performing a complete cycle study involving multiple inference strategies and specific benchmarks, such as in [118], is not uncommon. However, while such analysis provides a comprehensive idea of the relevance of the inferred network topology, it represents a significant investment of both time and money, and is sometimes not even possible.

In this chapter, the general effects of data quantity and system perturbations on the accuracy of the GRN reconstruction were evaluated for one rhythmic model of gene regulation and five steady-state models. Our contribution is threefold. We showed the relevance of multifactorial benchmarks to assess the performances of network inference strategies, the importance of an appropriate choice of model complexity given data availability, and revealed pragmatic considerations for experimental designs. Depending on the cost of performing more experiments or increasing the amount of datapoints, one of those choices should be preferred.

The algorithms considered here showed consistent performances across the 6 investigated networks. All network inference strategies did not, however, benefit equally from the increasing amount of data. Nevertheless, the fact that the parameter free, Gaussian process strategy GPDM has been consistently outperforming all strategies presented is noticeable and of further interest. In addition, by looking at the data expense and the resulting reconstruction accuracy, it should be further noted that GRN inference algorithms should improve the way various time series experiments originating from the same biological system are taken into account.

[26] showed that, on average, a combination of network inference strategies leads to the best network reconstruction. As such, it is noticed that the order in which the links were inferred by each algorithm, and experiment, was different. Further research, therefore, should learn the ranks, or confidence levels, of each link in the network reconstruction process and design a proper combination of the algorithms that optimizes their

synergy, depending on the experimental conditions.

Multifactorial studies such as the one presented here require a considerable amount of simulations. Some algorithms, such as those involving MCMC sampling, took several days to run on a 24 cores workstation. As such, a complete analysis of the experimental design space is not possible but other decisional factors exist and require further inspection. Among those, [98] showed that denser sampling is preferable to additional replicates. Such strategy could be particularly profitable for transient data and to algorithms that explicitly estimate derivatives. [102] used time series data originating from the DREAM 4 challenge to show that using only half of the perturbation data (without the recovery to steady-state) might be beneficial to some algorithms. Furthermore, some methods are able to incorporate information on external inputs, such as perturbations (with the targets still unknown), which increases the average performance. In addition, in practice, gene regulatory networks are often of bigger dimension which is not always accessible to the most computationally expensive algorithms.

Finally, this study did not take into account prior knowledge of the system, which could potentially be iteratively integrated into each step of the network reconstruction. For example, a strategy such as the one presented by [119] actively optimizes the precision of the predictions by proposing the next most informative knock out. In such case, the aforementioned results would likely underestimate the resulting accuracy of the reconstruction. In fact, doubling the amount of data points by doubling the observation time or by performing an additional experiment not only provides different levels of information, but can reveal different parts of the network as well. Such strategy might be necessary to cope with the most isolated genes.

Further Notes

It is important to notice that a key commonality between the most performant network inference strategies is that they do not simultaneously reconstruct the network as a whole, but rather through "separate", most likely, set of interactions. As a conclusion, even for the most performant network inference strategies, the resulting network is not ready to be simulated to reproduce their emergent dynamical behavior already, but instead intermediate validations are yet still required.

The field has now reached a certain maturity and strengths and weaknesses of many network inference algorithms have become more evident. However, it should be noted that the amount of information that can be withdrawn from gene expression data about the

circuitry of the GRN would eventually saturate (if not already), given the partial measurements of the species involved. Furthermore, increasing amount of data are generated over time and it becomes less common to study an entirely novel complex biological system on its own. The next challenge would now eventually be heading towards focusing on the best integration of multiple experiments data, the integration of prior knowledge, the robust combination of network inference strategies or the optimization of the information that can be gained from a following-up experiment.

2.5 Strengths and Limitations of the Study

As a summary, the DyDE/ATA modeling framework developed in this thesis holds the following strengths and weaknesses:

- Strengths:
  - Simplicity, efficiency, flexibility, interpretability, comparable, scalable.
  - To date, it is the only network inference strategy that allows to infer dynamical perturbations, since it relies on control theory tools developed for linear systems.
- Limitations
  - Given additional, more informative data, it might not remain the most efficient technique to infer the gene regulatory network. However, the development of gene regulatory network inference algorithm reached a sufficient maturity so that one might now be interested in combining the results of different paradigms, or how to rely on prior knowledge to build hypothesis about the circuitry.
  - The reconstruction of the circuitry of GRN through the linear modeling strategy considered here is based on a individual, independent thresholding process that does not inherently take into account the gene expression of the others genes to compute the fitness score between two genes.

    It is showed empirically that such approach works well: the highest fitness scores are very likely to represent actual gene-gene interactions, which make its performance comparable with more complex, state-of-the-art network reconstruction algorithms. As such, it is a valuable tool that should be further considered at least as a complement to other techniques in the case the data

would be more informative (10+ datapoints).

Nevertheless, the modeling strategy typically represent a heuristic approach. Indeed, it is often considered in the network inference literature that genes can be regulated by up to 3 parents genes at a time. There are, therefore, rooms for improvements. A more biologically accurate version of the approach considered here should therefore estimate the most relevant contribution of multiple genes at a time, given the other genes expressions available.

Chapter 3

Identification of Dynamical Regulators of the Arabidopsis Thaliana Circadian Clock

Adapted from: Mombaerts, L. et al. Dynamical differential expression (DyDE) reveals the period control mechanisms of the Arabidopsis circadian oscillator. PLoS Comput. Biol. https://doi.org/10.17863/CAM.35626 (2019).

Author Contributions

- **Conceptualization:** Laurent Mombaerts, Alberto Carignano, Fiona C. Robertson, Jorge Goncalves, Alex A. R. Webb.
- **Investigation:** Fiona C. Robertson.
- **Methodology:** Laurent Mombaerts, Alberto Carignano, Timothy J. Hearn, Jin Junyang, David Hayden, Ye Yuan.
- **Resources:** Alberto Carignano, Fiona C. Robertson, Timothy J. Hearn, Zoe Rutterford, Carlos T. Hotta, Katherine E. Hubbard, Marti Ruiz C. Maria, Matthew A. Hannah.
- **Supervision:** Jorge Goncalves, Alex A. R. Webb.
- **Writing – original draft:** Laurent Mombaerts, Alberto Carignano.
- **Writing – review & editing:** Laurent Mombaerts, Timothy J. Hearn, Jorge Goncalves, Alex A. R. Webb.

3.1 Contribution

The circadian oscillator, an internal time-keeping device found in most organisms, enables timely regulation of daily biological activities by maintaining synchrony with the external environment. The mechanistic basis underlying the adjustment of circadian rhythms to changing external conditions, however, has yet to be clearly elucidated. We explored the mechanism of action of nicotinamide in Arabidopsis thaliana, a metabolite that lengthens the period of circadian rhythms, to understand the regulation of circadian period.

For this purpose, several computational methods were developped. First, a prediction model was build to distinguish between rhythmic-non rhythmic gene expression data, based on the optimization of a hand-designed skewed sinusoidal function and logistic regression. Second, to identify the key mechanisms involved in the circadian response to nicotinamide, we developed a systematic and practical modeling framework (DyDE) based on the identification and comparison of gene regulatory dynamics. While theoretically tested with in silico models [93], this is the first successful application of the ν-gap methodology.

The results showed that the developped methodology was able to recover most of the known structure of the Arabidopsis Circadian Network from a single experiment of 48 hours with a sampling rate of one data point per 4 hours. Subsequently, several novel regulatory interactions were proposed. From a biological perspective, the methodology developed for this paper allowed to identify genes that are particularly responsible for the dynamical entrainment of the circadian clock, and validated those results experimentally. On one hand, this provides additional knowledge on the dynamical effects of nicotinamide and the role of blue light in the response of the circadian oscillator. On the other hand, it provides novel knowledge on the mechanisms of synchronization of the physiological rhythms of most organisms with the environment. Finally, being flexible, highly parallelizable, the methodology was extended to the whole genome (1 000 000+ interactions) to search for previously unknown genes that have the potential to be involved in the dynamical regulation of the circadian oscillator and nicotinamide. While those results remain to be validated, promising genes were identified. Such approach was not possible for most currently highly performant state-of-the-art network inference algorithms.

Altogether, those results suggest that our methodology could be adapted to predict mechanisms of drug action in complex biological systems.

3.2 Introduction

The synchronization of physiological rhythms with the external environment is important for nearly all organisms. Circadian oscillators are internal timing devices that produce rhythms with a period of about 24 hours to regulate a wide range of biological processes. Circadian rhythms maintain synchrony with the daily timing of light and dark cycles resulting from Earth's rotation by constantly integrating environmental signals. This process of synchronization is called entrainment. Studying the mechanisms that dynamically adjust circadian period and phase, therefore, is critical to understand the control of daily biological activities.

In Arabidopsis thaliana, the circadian oscillator consists of a complex circuit of highly connected transcriptional regulators. Together, they coordinate global transcript accumulation and diverse biological processes, such as photosynthesis, hormone signaling, hypocotyl elongation and plant-pathogen interactions [120, 121, 122, 123, 124]. The light perception of the circadian oscillator is conferred by a suite of photoreceptors. The photoreceptors are split into two classes: phytochromes (principally *PHYA* and *PHYB*), that primarily sense the red portion of the spectrum [125] and cryptochromes (*CRY1* and *CRY2*) that are sensitive to blue light [126, 127, 128].

Recent studies have demonstrated a role for metabolism in regulating and entraining the circadian oscillator of Arabidopsis thaliana. The primary metabolite sucrose accelerates the circadian oscillator (i.e., reduces its period) through regulation of the morning expressed gene PSEUDO RESPONSE REGULATOR (*PRR*) 7 [129], while GIGANTEA (*GI*) has been identified as a necessary sucrose-signaling mediator in the dark [87]. Another metabolite, nicotinamide (NAM), a breakdown product of nicotinamide adenine dinucleotide (NAD), causes long period of the circadian oscillator in all organisms tested [130, 131]. The mode of action of NAM is uncertain: various mechanisms having been proposed, including inhibition of the production of the Ca^{2+}-agonist cyclic adenosine diphosphate ribose (cADPR), inhibition of polyADP ribose polymerases and histone modifications [130, 131, 132]. The goal of this study was to use NAM as a tool to identify the processes responsible for a change in circadian period, which might be required for circadian entrainment and homeostatic adjustment [133, 134, 135].

Typically, large sections of the transcriptome can be differentially expressed, despite not being directly affected by the treatment (off-targets) (Figure 2.3). Due to the large number of feedback loops involved in a complex and relatively small Gene Regulatory Network (GRN) such as the circadian clock, this effect is particularly significant as a per-

turbation anywhere in the network typically strongly affects all molecular concentrations. Furthermore, as the perturbations induced by NAM in the circadian clock are intrinsically related to changes in circadian period, a large part of the transcripts are differentially expressed. Thus, Differential Expression (DE) analysis, the traditional approach used to identify the mechanisms that alter biological behavior in response to drugs, environmental signals or genetic lesions [136], will usually fail to identify the small number of genes central to the biological perturbation. The main reason is that DE only performs statistical analysis of changes in gene expression levels [23, 90]. As an alternative to the DE analysis, the DyDE modeling framework has been used to identify and characterize differentiated regulatory dynamics between genes to capture key mechanisms involved in NAM-induced perturbations in the circadian system of Arabidopsis. By comparing changes in both topology and subtle dynamic modifications of regulatory mechanisms, we were able to considerably narrow down potential targets of NAM in the circadian clock.

The findings predicted by DyDE are experimentally tested and demonstrate the role of the circadian gene PRR7 as a key regulator of dynamics adjustment of the circadian clock. In addition, TIMING OF CAB EXPRESSION 1 (*TOC1*) and the interplay between *PRR7* and PSEUDO RESPONSE REGULATOR 9 (*PRR9*) are identified as the main mediators of the circadian system response to NAM. The modeling insights also identified alterations in *CRY2* dynamics resulting from the NAM treatment. Therefore, we also investigated the role of blue light in the circadian period change of NAM-treated plants. In particular, we found that blue light regulates circadian oscillations of $[Ca^{2+}]_{cyt}$ through a NAM-sensitive pathway. These new perspectives contribute to the understanding of the mechanistic details underlying the regulation of period of circadian oscillators. Overall, the results suggest that DyDE is a useful tool to generate reliable hypothesis from time series data for the identification of drug targets in complex biological systems.

3.3 Methods

To investigate how NAM might regulate the period of the circadian oscillator we first used statistical tools to identify those transcripts that have circadian rhythms in abundance in both untreated and NAM-treated plants. Then, we introduce the DyDE approach to characterize altered dynamics within the circadian regulatory network of NAM-treated plants. The hypothesis generated by DyDE were experimentally tested using genetic mutant and physiological experiments in different light conditions. Finally, we extended DyDE to the whole rhythmic transcriptome to further investigate clock period regulation.

3.3.1 Statistical Characterization of Circadian Transcripts

To assess whether genes are regulated by the circadian oscillator, most methods take advantage that circadian regulation of transcript abundance resemble a sinusoid. To estimate circadian period of the regulation of a particular transcript, the main idea is to find the sinusoid that most closely matches its abundance over time [137, 138]. However, in NAM-treated plants the changes in abundance of circadian-regulated transcripts have a considerable number of nonsinusoidal profiles (Figure 3.1). Furthermore, available theoretical framework usually don't hold for poorly sampled signals, especially when corrupted with significant noise.

Pseudo-sinusoidal Functions

To overcome this problem, we devised a learning approach based on "pseudo-sinusoidal" functions to properly assess the rhythmicity and the corresponding circadian period of signals from gcRMA normalized microarray data of NAM treated plants. To infer period, phase and amplitude, linear trends are eliminated by removing the best straightline fit and pseudo-sinusoidal functions are fitted to each signal to minimize the 2-norm error. Pseudo-sinusoidal functions account for many signals that are periodic but not sinusoidal. Pseudo-sinusoidal functions are constructed by joining together two sinusoids with different periods. Hence, a complete oscillation of a pseudo-sinusoidal function consists of the first sinusoid (of period p_1) in the first half-oscillation, and the second sinusoid (of period p_2) in the second half-oscillation (Figure 3.2A). The resulting period of the pseudo-sinusoidal function is defined by $p = \frac{p_1+p_2}{2}$. This can be expressed by:

$$S = \begin{cases} A * sin(\frac{2\pi}{p_1} * t + \phi_1), t \in [0, \frac{p_1}{2}] \\ \\ A * sin(\frac{2\pi}{p_2} * (t - \frac{p_1}{2} + \frac{p_2}{2}) + \phi_1), t \in [\frac{p_1}{2}, \frac{p_1}{2} + \frac{p_2}{2}] \end{cases}$$

where A is a scaling factor that accounts for the amplitude of the signal and ϕ_1 is the phase of the signal. The algorithm searches possible combinations of p_1 and p_2 to minimize the least square distance between pseudo-sinusoidal functions and the data. We allowed periods p_1 and p_2 to vary between 12 and 36 hours. A perfect sinusoid gave a high fit the for wild-type background dataset. We found that three periodic signals were highly represented in the dataset. In particular, those with p_1, p_2 equal to $p/2$, $p/2$ (pure sinusoid); $p/2+3.8$, $p/2-3.8$ (p_1 is greater than p_2); and $p/2-7.3$, $p/2+7.3$ (p_1 is smaller than p_2) (Figure 3.2).

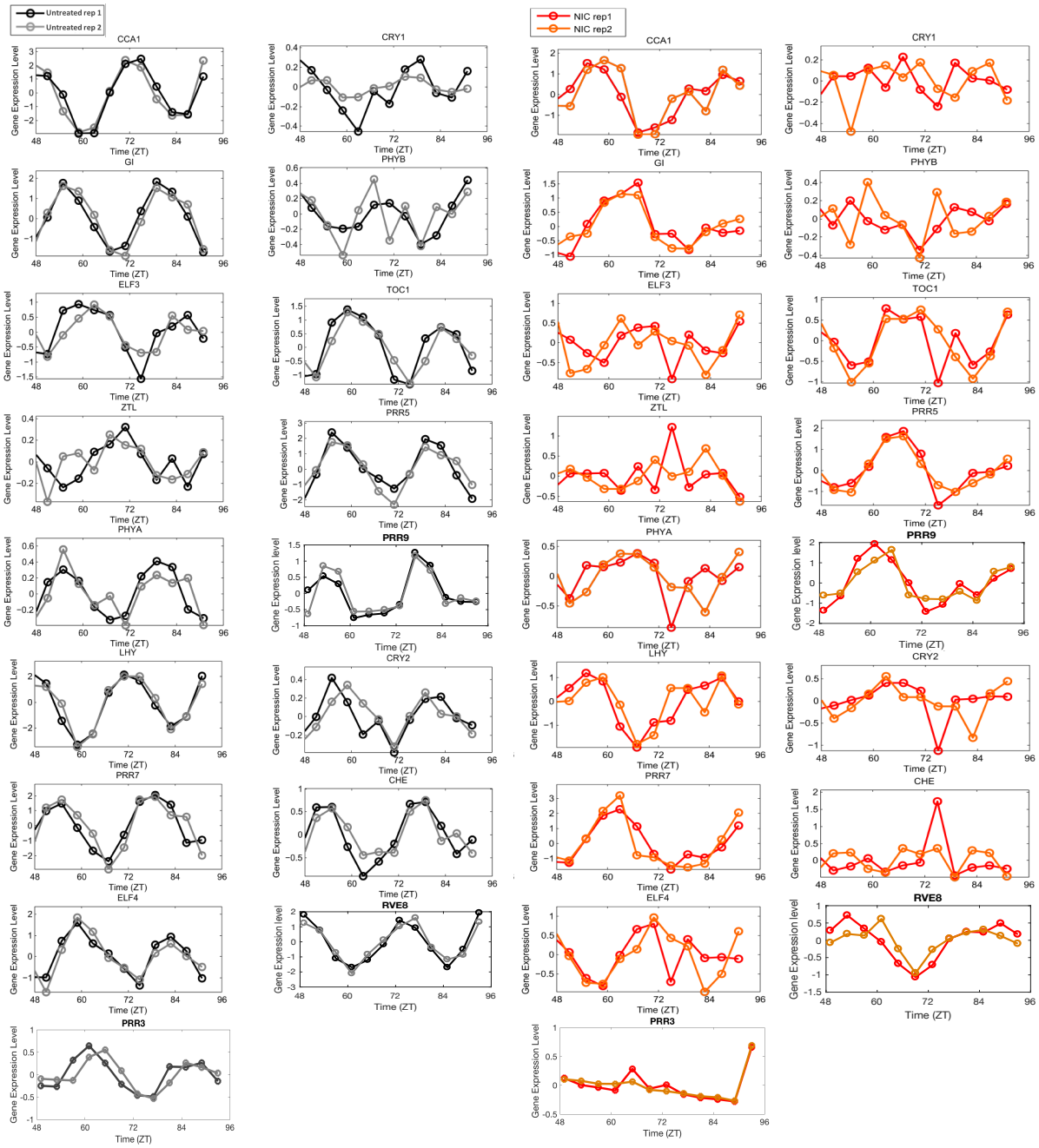

Fig. 3.1 **Detrended time series of circadian transcript in both untreated and NAM condition.** Data were gathered for 44 hours every 4 hours, 2 replicates, starting from 49 hours after the switch to constant light (i.e., third day of constant light). Data showed are detrended, so that the rhythmic pattern is clear. LUX does not appear on this list, as the probe also measured the expression of AT5G59570. CRY1, PHYB, ELF3, ZTL and CHE were not considered for the network inference step.

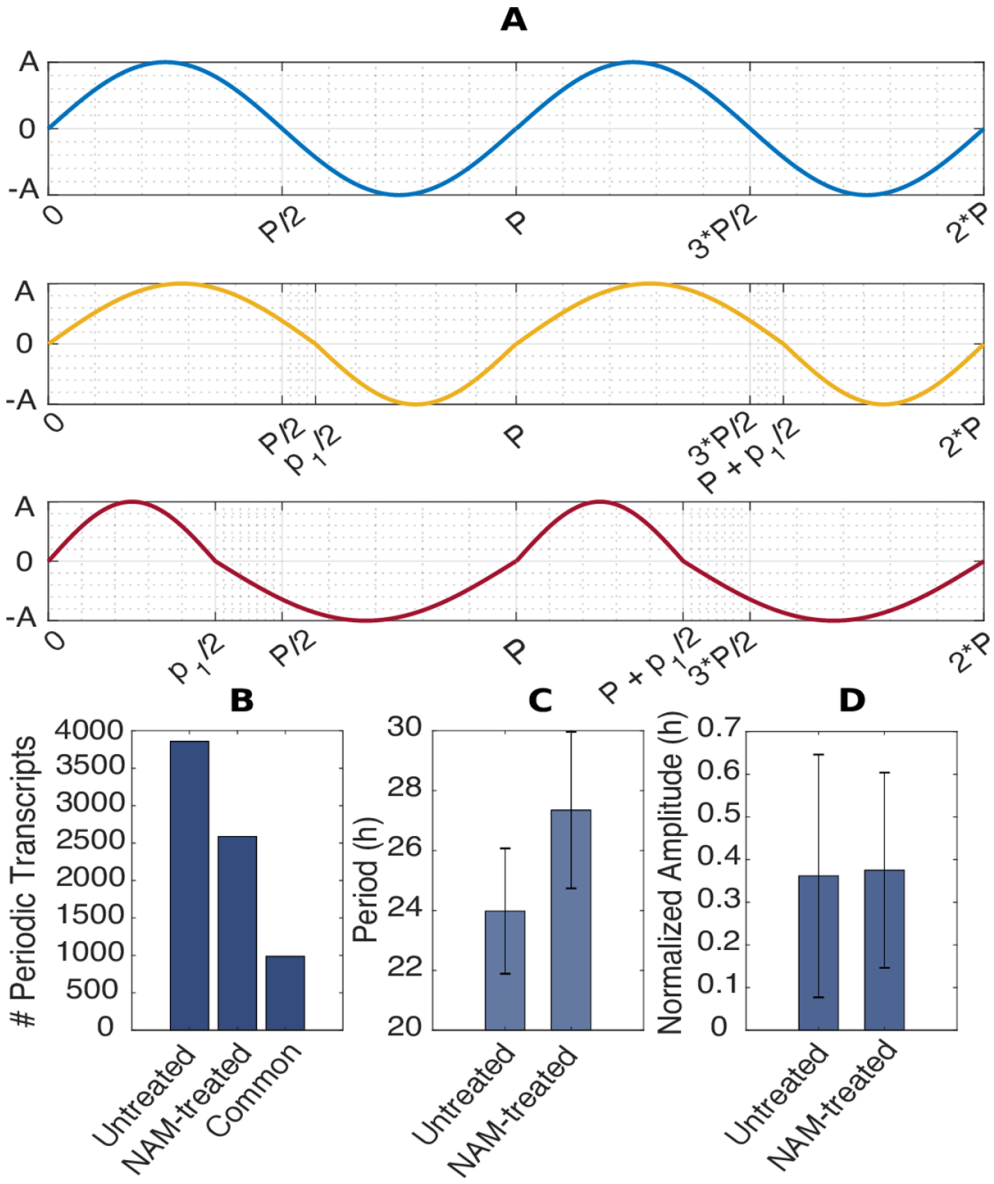

Fig. 3.2 **The effects of NAM on the circadian regulation of the transcriptome. (A)** Illustration of the shape of S. The first panel shows two period of a perfect sinusoidal shape, whereas the second panel displays the segmentation of the period P into p_1 and p_2, where p_1 is greater than p_2. p_1 and p_2 follows the formula: $P = (p_1 + p_2)/2$. The last panel displays the case were p_1 is smaller than p_2. **(B)** Number of periodic transcripts that have been identified in untreated and NAM-treated plants, as well as the intersection. **(C)** Circadian period of untreated and NAM-treated transcripts plus minus standard deviation. The mean increase of period following the NAM treatment is of 3.3h. **(D)** Amplitude analysis (normalized) for the same transcripts

Probabilistic Discriminative Model

We used a logistic regression framework to generate a probabilistic discriminative model that estimates the probability of a gene to be rhythmic given its time course data. In this case, the classification problem only contains two classes: rhythmic (C1) and arrhythmic (C2). For each transcript, a set of 8 features $x = X_1, X_2, ..., X_8$ is computed and empirically believed to be crucial to distinguish between rhythmic and arrhythmic transcripts. The features were computed from 2 signals: the first signal (A) corresponds to the average of replicates and (B) being a single replicate for which the L2-norm error with the best fitted pseudo-sinusoidal function is lower than for the other replicate. The following features were computed:

- Ratio of power in the 18–32 hours frequency range (of (A) and (B))
- L2-norm of the error to the best fit of pseudo sinusoidal function (of (A) and (B))
- The variance of the power spectrum (of (A) and (B))
- The amplitude of the best fitted pseudo-sinusoidal function (of(A) and (B))

The log of the ratio of probabilities between the two classes, also known as the log odds, is given by [139]:

$$ln(\frac{p(rhythmic|x)}{p(arrhythmic|x)}) = ln(\frac{p(C_1|x)}{p(C_2|x)}) = ln(\frac{\sigma}{1-\sigma}) = logit(\sigma)$$

The goal of the logistic regression is to estimate σ for a linear combination of the X_n features such that:

$$logit(\sigma) = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n$$

The weights b_i of the independent variables X_i were estimated using the mnrfit function in MATLAB. The algorithm is initially trained with a mix of 100 rhythmic and 100 arrhythmic transcripts randomly chosen from the dataset and visually inspected to show clear (ar)rhythmicity. Finally, the decision boundary was set so that if $p(C_1|x) > 0.5$, the gene was classified as rhythmic, and vice versa. Our approach, therefore, is inspired by the patterns observed in the dataset but not strictly constrained to pure cosine

shapes. With the inclusion of the S function, we allow the search for asymmetric signals, which represent a large part of the transcriptome. A main distinction with the previously introduced algorithms, therefore, is the data-specific, learning approach devised to allow for a wider range of periodic signals. However, this offers additional advantages such as a dedicated way to handle noise between replicates, or the information in the frequency domain of the signal, which are both learned from the data. Comparison of performances with standard periodicity assessment tools is shown on Figure 3.3.

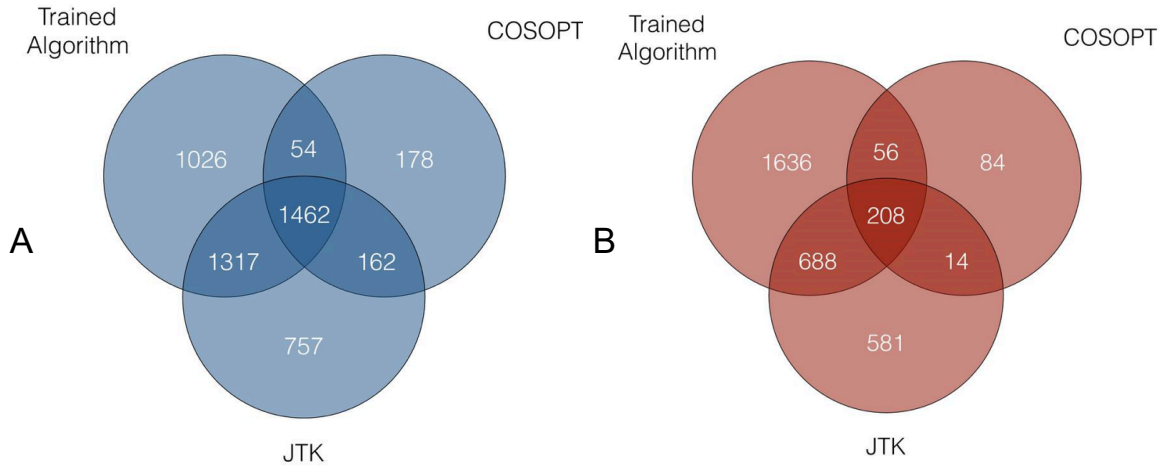

Fig. 3.3 **Assessment of circadian regulated transcripts from both the learning methodology and standard tools.** **(A)** Results correspond to untreated plants. The trained algorithm, COSOPT [138] and JTK [137] respectively identified 3859, 1856 and 3698 circadian regulated transcripts. JTK and the trained algorithm identified most of the genes labelled as periodic by COSOPT (resp. 87% and 81% of them). The rhythmicity of 75% of the genes labelled rhythmic by JTK was confirmed by the learning strategy. **(B)** Results corresponding to NAM-treated plants. The rhythmicity of 60% of the genes labelled rhythmic by JTK was confirmed by the learning strategy while 1636 novel genes were identified as rhythmic with a typical non-sinusoidal profile.

3.4 Results

We identified 3859 (18.4%) circadian-regulated transcripts for the untreated plants (Figure 3.2A). These were enriched for Gorilla terms ‘Circadian Rhythm’ and ‘Rhythmic Process’ (p = 4.07E18; GEO No. GSE19271). A total of 2588 (12.3%) transcripts were identified as rhythmic in NAM-treated plants (Figure 3.2B), with a mean increase in period from 24.0 ± 2.1 h (-NAM) to 27.4 ± 2.6 h (+NAM) (Figure 3.2C) and without a noticeable change in amplitude (Figure 3.2D).

3.4.1 DyDE applied to the Arabidopsis circadian clock genes

We considered a total of 17 known clock genes: *CCA1, LHY, PRR9, PRR7, PRR5, RVE8, GI, TOC1, ZTL, ELF4, ELF3, PHYA, PHYB, CRY1, CRY2, CHE* and *PRR3*. However, the core oscillator genes *ZTL, ELF3, PHYB, CRY1, PRR3* and *CHE* were identified as non-rhythmic in the presence of NAM, which was confirmed by visual inspection (Figure 3.1). Hence, these genes are excluded from the modeling of NAM targets as they cannot be contributing to the rhythmic dynamics of the remaining oscillator components that are measured in the presence of NAM.

As a first step, we computed models for all available pairs of the clock genes for both conditions, totaling 220 SISO models (110 in untreated and 110 in NAM). We kept only those models with good agreement with the data, i.e. above a fitness threshold. On one hand, the userdefined threshold has to be set large enough to reliably capture the dynamics involved between genes, and provide the ν-gap analysis with comparable models. On the other hand, the threshold has to be set sufficiently low to consider enough gene-to-gene relationships to detect a dynamical perturbation in the network. Here, the fitness threshold was set to 46% as we noted that below this threshold, the amount of unknown regulations dramatically raised (Figure 3.4, Supplemental Table 3 (Online Material)).

In total, 70 regulatory links were retained for untreated plants and 55 links for NAMtreated plants between the 11 clock genes. The untreated models describe 70% of the known regulatory pathways among these 11 genes (Supplemental Table 3 (Online Material); Figure 3.4). 64% of which, had the expected activation or inhibition effect. These numbers are remarkable, taking into account the model simplicity, and confirms that the majority of clock links can be represented by simple linear dynamics [85, 86, 88].

In particular, 28 links were present in the untreated samples but not in the NAM-treated samples. These 28 links form a network from now on referred to as “regulation loss” network, which captures the links abolished by NAM. In addition, 42 links are present in both conditions which form a network, so called “common” network that is common to both treated and untreated plants (Supplemental Table 3 (Online Material)).

We used the ν-gap to identify those links among the common network whose dynamics were significantly affected by NAM. Figure 3.5A and Table 3.1 depict the comparison of the dynamics of each link with the ν-gap. All regulatory interactions are somehow affected by the treatment, which is expected from the interconnected circadian network. Let us then consider the highest ν-gap values, which are associated with the following

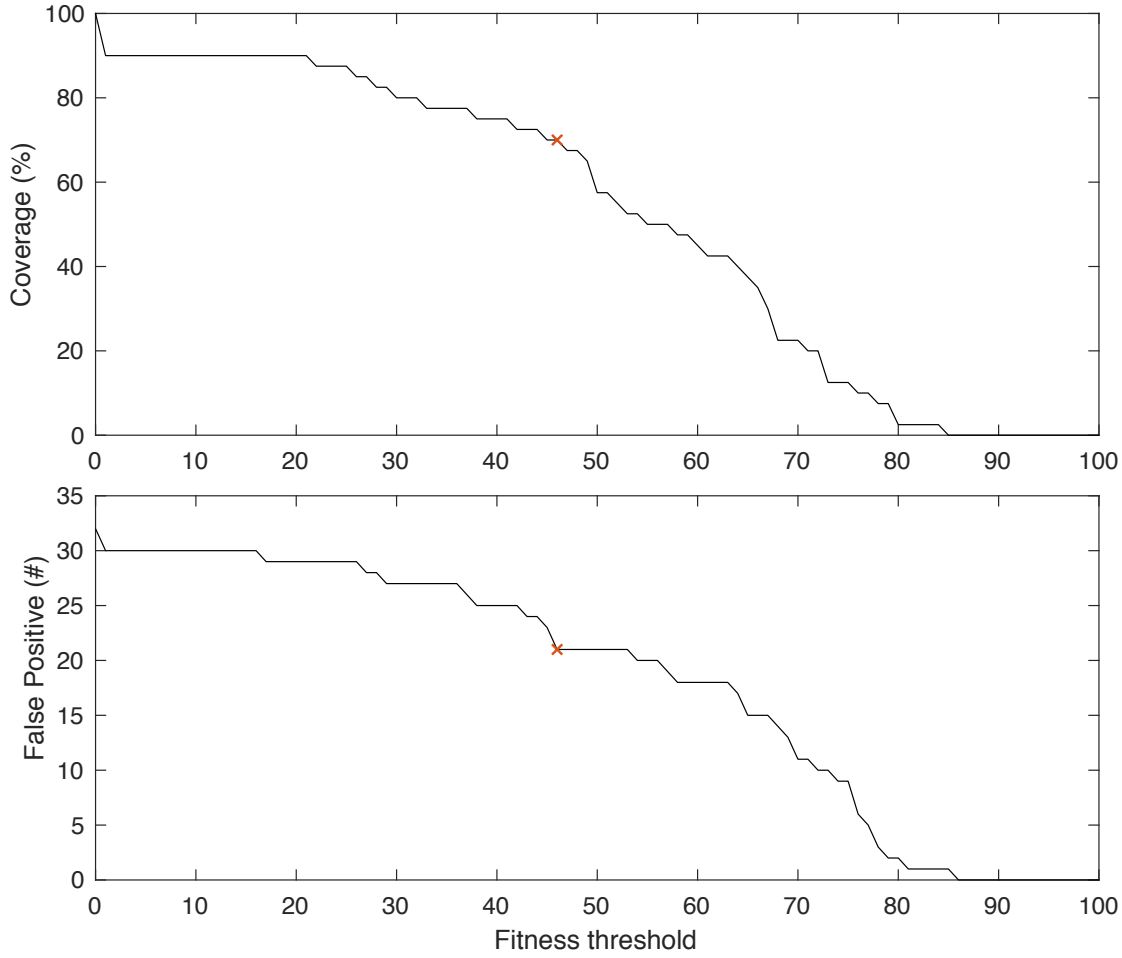

Fig. 3.4 **Coverage and false positive curves of the known regulatory links involved in the circadian oscillator of Arabidopsis Thaliana, as inferred by DyDE.** In DyDE, linear Ordinary Differential equations (ODEs) of order one are computed between each pair of genes to describe the dynamics of the whole system. To be further considered as a good approximation of the dynamics involved, each dynamical model needs to pass a validation criterion based on its agreement to the data (i.e. a user-defined threshold on the goodness of fit). Decreasing the fitness threshold leads to a better coverage (upper panel) of the system dynamics but increase the amount of false positives (lower panel). The coverage describes the amount of links inferred over the amount of total true links in the system (as defined by [33]). The number of false positive corresponds to links that are not represented in Fogelmark et al. The maximum amount of possible false positives is 32, while the total amount of links in the true system is of 40. The threshold of 46% (represented by a red cross) is chosen for this analysis with a coverage of 70% (which corresponds to 28 true positives and 21 false positives)

Table 3.1 **Sorted ν-gap values corresponding to common links between untreated and NAM (top 5).** ν-gap values computed for each link inferred in both untreated and NAM-treated networks. This table ranks the ν-gap values from the largest to the smallest.

Rank	ν-gap value	Input	Output
1	0.500	*TOC1*	*PRR9*
2	0.471	*CRY2*	*ELF4*
3	0.424	*CRY2*	*LHY*
4	0.379	*CRY2*	*RVE8*
5	0.356	*PRR9*	*CRY2*

links: *TOC1* to *PRR9* (0.5), those originating from *CRY2* to *ELF4* (0.47), LHY (0.42) and *RVE8* (0.37) and *PRR9* to *CRY2* (0.35). Interestingly, the only inferred interaction originating from *CRY2* that does not seem affected connects to *TOC1* (ν-gap of 0.06). These results suggest that a major dynamical change is induced to *CRY2* in the dynamical response of the circadian clock to NAM. In addition, the largest ν-gap value suggests that the causality within the time course data of *TOC1* and *PRR9* has changed significantly differently towards the treatment, as compared to the other parts of the circadian network.

We then used a standard network topology metric to identify the genes that are central to the drastic changes in dynamics captured by the regulation loss network. This topology metric accounts for the connectivity of a gene, i.e. the number of its incoming and outgoing links. This measure is estimated for each gene of the regulation loss network. As an example, *PRR7* has six incoming links and nine outgoing links for untreated plants. The connectivity of *PRR7* in untreated plants is then equal to 15. Among those, only six of were present in NAM-treated plants. *PRR7*, therefore, has a connectivity of nine in the regulation loss network, which correspond to a loss of 60% of its connectivity from untreated to NAM treated plants. As a result, *CCA1* (61%), *PRR7* (60%), *TOC1* (57%) exhibit the highest connectivity drop (Figure 3.5B; Table 3.2). This result identifies the biological functions of *CCA1, PRR7* and *TOC1* as being highly affected by NAM in the regulation of the circadian clock.

DyDE, therefore, identifies the regulatory dynamics of *TOC1-CRY2-CCA1-PRR7* as being predominately affected by NAM as a result of both ν-gap and connectivity analysis. Accordingly, the strong emergence of the blue light receptor *CRY2* in the ν-gap analysis suggests that nicotinamide alters the regulation of the interactions between light signaling and the circadian oscillator. These findings are further examined through mutant analysis and single wavelength light experiments.

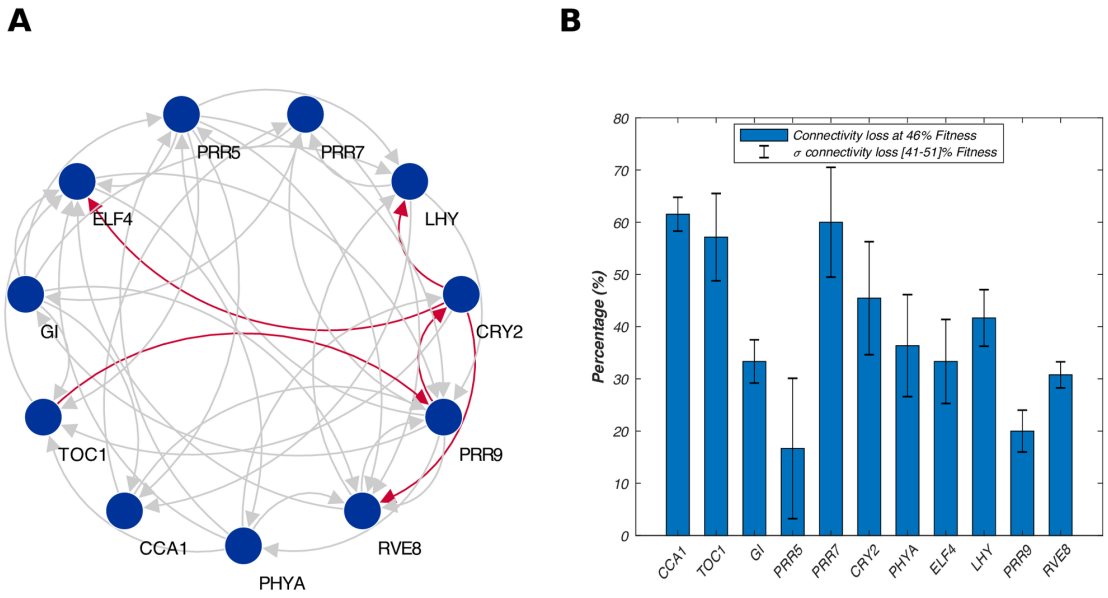

Fig. 3.5 **DyDE applied to the Arabidopsis circadian oscillator genes. (A)** Common network and ν-gap analysis. The common network displays the models that have been validated in both untreated and NAM-treated plants. A directed arrow from gene a to gene b (blue circles), therefore, represents a dynamical model that captures the dependency of b on a. Red arrows represent the models associated with the top five highest ν-gap values. **(B)** Bar plot comparing the connectivity loss (%) associated to each gene. For a particular gene, the connectivity loss corresponds to the total amount of incoming and outgoing links that were validated in untreated plants but not in NAM-treated plants. Error bars represent the standard deviation of connectivity loss for ± 5% change in fitness threshold selection.

3.4.2 PRR7/PRR9 Inter-regulation together with TOC1 are Targets of Nicotinamide

To test the predictions that *TOC1, CRY2, CCA1* and *PRR7* are associated with the effect of NAM on the circadian oscillator, the sensitivity of circadian mutants to NAM was experimentally investigated. All mutants responded to NAM with increased circadian periods, with the exception of two independent lines of the same T-DNA insertion allele of PRR7, which were insensitive (prr7-3 $p > 0.95$; prr7-11 $p > 0.95$). The insensitivity of prr7-11 to NAM was confirmed by measuring circadian rhythms of leaf movement. prr7-11 was not affected by NAM at any tested concentration, contrasting with a dosedependent effect of NAM on circadian period in other prr mutants and associated backgrounds ($R2 > 0.9$).

In contrast, toc1-2 and TOC1-ox had significantly greater responses to NAM than wild type. These results support our predictions that NAM induces dynamical changes specifically to PRR7 and TOC1. No dramatic changes of period, however, were observed

Table 3.2 **Connectivity loss corresponding to each gene, for untreated and NAM-treated networks.** Values are displayed for a fitness threshold of 46%.

Gene	Connectivity (Untreated)	Loss of Connectivity (count)	Loss of Connectivity (%)
CCA1	13	8	61.5
TOC1	14	8	57.1
GI	12	4	33.3
PRR5	12	2	16.7
PRR7	15	9	60
CRY2	11	5	45.5
PHYA	11	4	36.4
ELF4	12	4	33.4
LHY	12	5	41.7
PRR9	15	3	20
RVE8	13	4	30.8

for cry2-1 and cca1-11, suggesting that these might not contribute directly to the response to NAM.

Finally, derived from the ν-gap analysis, the possible change in the dynamical behavior of *PRR9* in mediating the effect of NAM on the clock was evaluated with a prr7-3 and prr9-10 double mutant. prr7-3 and prr9-10 had an epistatic interaction, with the double mutant responding to NAM by a 5.3 ± 1.6 h increase of period, more than either the insensitive prr7-3 or the oversensitive prr9-10 alone. The epistasis of prr9-10 to prr7-3 was confirmed at all concentrations of NAM tested.

3.4.3 Nicotinamide-induced Changes in Period are Associated with a Blue Light Signaling Pathway

The mutant analysis did not confirm the modeling dynamical perturbation of *CRY2* in the response to NAM. However, *CRY2* is one of a pair of cryptochrome blue light photoreceptors and so mutant analysis might not be the most appropriate tool to investigate the role of the blue light photoreceptor. To investigate further we also investigated the role of blue light in the response to NAM using monochromatic light conditions. High frequency measurements of the circadian promoter fusions PRR9:LUC, PRR7:LUC, TOC1:LUC, CCA1:LUC, LHY:LUC and GI:LUC were collected in the presence or absence of 20 mM nicotinamide under constant blue or red light.

In the absence of blue light, NAM was without effect on the circadian period or amplitude of CCA1:LUC and other promoter:luciferase fusions. This demonstrates that input pathways associated with blue light are sensitive to NAM. Under blue light exposure, all promoter:luciferase fusions considered had an increase in period in the presence of NAM. Under red light exposure, the period response was either negligible (PRR9:LUC, CCA1:LUC, LHY:LUC, GI:LUC) or negative (PRR7:LUC, TOC1:LUC). These results suggest that blue light increase the response of circadian period to NAM, while red light decrease its responsiveness.

Having previously proposed that the effects of NAM on the circadian system are due to the inhibition of the production of the Ca^{2+}-agonist cADPR [130], we tested if the response to NAM of prr7-11 is due to altered Ca^{2+} signaling. We investigated, therefore, the inhibitory effects of NAM on circadian $[Ca^{2+}]_{cyt}$ oscillations in prr7-11 and in light signaling mutants in red and blue light. 20 mM NAM was equally effective in abolishing circadian rhythms of $[Ca^{2+}]_{cyt}$ in both Col-0, prr7-11 and prr7-3 prr9-10. This suggests either that there are multiple sites of action of NAM or that PRR7 is downstream of the effects of NAM on $[Ca^{2+}]_{cyt}$.

In constant blue light, there were robust oscillations of $[Ca^{2+}]_{cyt}$ in plants with functional *CRY1* photoreceptors, being abolished in cry1 and, cry1cry2 but unaffected by cry2, phototropins and Phy loss-of-function mutants. Under blue light, NAM abolished $[Ca^{2+}]_{cyt}$ oscillations but did not reduce oscillations further in cry1 or cry1cry2. High amplitude oscillations of $[Ca^{2+}]_{cyt}$ were dependent on blue light because in constant red light, $[Ca^{2+}]_{cyt}$increased early in each cycle but without a subsequent decrease. This red light-induced increase in $[Ca^{2+}]_{cyt}$ was dependent on *PHYB*.

To examine the role of *PHYB* further we measured $[Ca^{2+}]_{cyt}$ in PhyB-ox and determined that in these plants $[Ca^{2+}]_{cyt}$ was rhythmic with a sinusoidal period of 25.0 ± 0.5 h in constant red light. NAM was without effect on $[Ca^{2+}]_{cyt}$ in constant red light, even in the PHYB-ox background demonstrating that blue light regulates circadian oscillations of $[Ca^{2+}]_{cyt}$ through a NAM-sensitive pathway. This pathway appears to be required for the major oscillatory dynamics of $[Ca^{2+}]_{cyt}$.

3.4.4 Extension of DyDE to the Rhythmic Transcritome

DyDE was further adapted to explore the rhythmic genome for additional targets for NAM and novel clock genes. For this purpose, models were computed between each pair of the 988 genes that were scored rhythmic in both untreated and NAM treated conditions,

resulting in 2 million models corresponding to potential interactions.

We selected the models that exhibit the highest goodness of fit (over 80%) in both untreated and NAM-treated plants to minimize the identification of erroneous interactions and computed their ν-gap value to investigate dynamics affected by NAM. As a result, out of ten, two models only were retained with a ν-gap > 0.2. These models identified the regulation of *AT5G35970* (P-loop containing nucleoside triphosphate hydrolases superfamily protein) by *AT2G21860* (violaxanthin de-epoxidase-like protein) and the regulation of *ATG21660 (GRP7/CCR2)* by *AT1G78600 (LZF1/BBX22)* as being altered by NAM. The regulation of *AT5G35970* by *AT2G21860* may be important as *AT5G35970* is identified by DyDE as being a hub regulated by four circadian oscillator genes. The second link is easier to explain because *GRP7* along with *GRP8* forms a slave oscillator driven by the circadian clock that regulates ABA responses [140]. *GRP7* is an RNA binding protein regulated by ADP ribosylation [141]. As the enzymes that perform ADP ribosylation are inhibited by NAD, this could suggest a role for nicotinamide inhibiting ADP ribosylation of an oscillator or slave oscillator component.

Then, the fitness threshold was released to 60% to further investigate novel clock components. For this purpose, we searched for those genes for which models can be computed from/ to clock components. Models with a ν-gap value above 0.2 were discarded as a consistency criterion. Finally, candidates were ranked according to their connectivity with the clock. As a result, 20 high potential genes were isolated. The whole genome analysis of clock input and output hubs and the ν-gap analysis suggest interesting roles for previously characterized genes, including *AT3G47500* (CYCLING DOF FACTOR3) [142], *AT4G38960* (*BBX19*) [143], *AT1G78600 (BBX22)* [144, 145], *AT3G22840 (CRY3)* [146], *AT1G28330 (DRM1)*, *AT2G33830 (DRM2)* [147, 148] and uncharacterized genes including *AT5G35970*.

3.5 Discussion

Here, we considered the problem of inferring the entry point of a treatment in an organism from limited time series data (in this case, the circadian clock in Arabidopsis). For this purpose, we used simple dynamical models to capture gene regulatory dynamics and compare those under different scenarios without making a priori assumptions on the structure of the network. Subsequently, we showed that simple dynamical models have the potential to identify crucial dynamical perturbations for complex systems such as the circadian clock. However, it should be stressed that, as for the sole purpose of identifying

the topology of the underlying network, our method competes well with the current state-of-the-art of network inference strategies.

We further devised a learning algorithm to capture the specific pattern of oscillating wave forms of genes affected by NAM. Since the period of oscillations of central clock genes increases from 24 (wildtype) to roughly 28 (NAM) hours, we focused on those genes. For a relatively small number of genes, DyDE efficiently narrowed down possible targets of NAM that could then be verified experimentally. Since it is likely that other genes may be targets of NAM, we further applied DyDE to all 988 circadian genes that were scored rhythmic in both untreated and NAM treated conditions.

DyDE identified important changes in the regulatory dynamics of *PRR7, TOC1, CCA1* and the blue light photoreceptor, *CRY2*, resulting from the treatment of plants to NAM as well as suggesting a mediating role of *PRR9*. Mutants analysis confirmed DyDE predictions of altered activity of *PRR7, TOC1* and *PRR9* and blue/red light experiments demonstrated that the effect of NAM is blue light dependent. The latter also demonstrated that blue light regulates circadian oscillations of $[Ca^{2+}]_{cyt}$ through a NAM-sensitive pathway.

The involvement of *PRR7* with the dynamic adjustment of circadian period in response to nicotinamide, revealed by the insensitivity of prr7-11 and prr7-3 to NAM and confirmed by leaf movements analysis, is interesting because *PRR7* is also required for the response of the circadian oscillator to sugars [129, 149]. *PRR7*, however, is not a direct target for NAM in the circadian oscillator because *PRR7* is not required for the response to NAM, as demonstrated by the hyper-sensitivity to NAM of the prr7-3 prr9-10 double mutant. Together, the insensitivity of prr7-3 and prr7-11 to NAM and hypersensitivity in the prr7-3 prr9-10 double mutant indicates that *PRR7* and *PRR9* regulate a component or pathway influenced by NAM and that *PRR7* might act upstream of *PRR9* in this regulation. The levels of expression of *PRR7* and 9 appear to regulate the pace of the circadian oscillator through feedback with *CCA1/LHY* and by acting as toggle switching the oscillator from a morning state when *CCA1/LHY* are high to an evening state when *TOC1* is high [150, 151].

Additionally, the blue-light dependency of both circadian oscillations of $[Ca^{2+}]_{cyt}$ and NAM regulation of circadian period might suggest that Ca^{2+} is associated with the response of the oscillator to NAM. Furthermore, we recently reported that *CALMODULIN-LIKE 24 (CML24)*, is a Ca^{2+}-dependent regulator of circadian period and that its effects are NAM sensitive [152]. A caveat to this argument is that our methodology identified

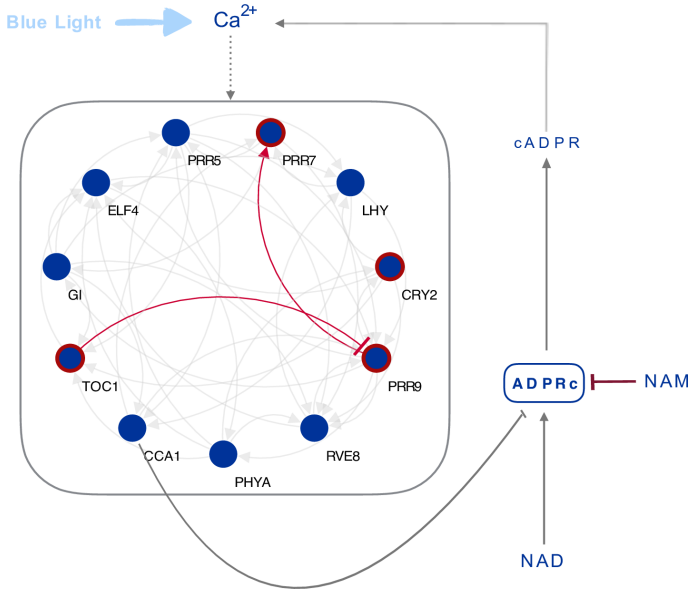

Fig. 3.6 **A blue light dependent module regulates the response of the circadian oscillator to NAM.** NAM might regulates the circadian oscillator through regulation of cADPR dependent circadian oscillations of $[Ca^{2+}]$. *CCA1* is a repressor of ADPRc. ADPRc generation of cADPR and $[Ca^{2+}]$ oscillations is inhibited by NAM. Both the effects of NAM on the circadian oscillator and circadian oscillations of $[Ca^{2+}]$ are blue light dependent. The regulation of $[Ca^{2+}]$ on the circadian oscillator is indicated by a dotted line. NAM could also regulate the circadian oscillator by $[Ca^{2+}]$ -independent events. We determined that the NAM-induced changes in circadian period are mediated principally by the interaction between *PRR7* and *PRR9*, as well as *TOC1*. These interactions are shown in red in the model.

CRY2 regulation of the transcriptional network being altered by NAM but the circadian oscillations of $[Ca^{2+}]_{cyt}$ were dependent on *CRY1*. NAM can also affect the oscillator through Ca^{2+}-independent mechanisms [132]. We propose that a module of circadian oscillator components *PRR7* and *9*, *TOC1* and a Ca^{2+} signaling network contribute to the blue light-dependent response of the circadian oscillator to NAM that regulates circadian period 3.6.

Then, extension of DyDE to the whole circadian genome has also identified components outside the core oscillator that might also be involved in response to NAM, including the regulation of *GRP7/CCR2* by *LZF1/BBX22* and these will be candidates for future investigation. Remarkably, five genes out of 22 that were isolated in our genome analysis are known to interact with circadian regulators (*BX19, CYCLING DOF FACTOR3*) [142, 143], have been previously implicated in circadian regulation (*BBX22*) [144, 145], in blue light signaling (*CRY3*) [146] or are being downregulated by ABA/cADPR (*GRP7, BBX19, BBX22*) [140, 141, 153]. This result is encouraging and opens the door to the

identification of novel drivers of circadian rhythms in Arabidopsis.

Overall, we suggest that the description of gene regulatory dependencies and the quantification of changes in dynamics computed by DyDE provide reliable hypotheses for the investigation of drug targets in complex gene regulatory networks, which has a broad range of applications in systems biology.

3.6 Strengths and Limitations of the Study

The DyDE modeling methodology has been applied to the circadian regulatory network of Arabidopsis Thaliana. This strategy holds the following strengths and weaknesses:

- Strengths
  - Identifying the source of a perturbation in a GRN is a challenging task, especially when large uncertainties remain about the circuitry of the underlying network. Indeed, the identification of Arabidopsis clock genes and their interaction is yet a dynamic field, with various novel interactions or novel cores genes being added over the years. The approach proposed here allows to provide predictions that are independent of the current consensus of the core network.
  - In the previous chapter, it has been shown that DyDE is capable of providing reliable predictions of the network while relying on small amounts of datapoints only. Given the length of NAM-treated recordings, this approach appeared to be particularly suited. In particular, it is showed that the DyDE not only provides good prediction of the underlying circuitry of the GRN, but also gives reliable estimation of the dynamical properties of genes to genes relationships.
  - The simultaneous inference and comparison of gene to gene dynamics is particularly new to the field, and showed promising performance. Such comparison of dynamic is not a straightforward task for nonlinear or nonparametric models. In this sense, the use of a linear model is a considerable advantage.
- Limitations
  - While the linear modeling strategy provided seemingly accurate predictions, there remains room for improving the specificity of the perturbation targets. For example, the modeling strategy only considers pairwise interactions, which is a heuristic approach to identify links that are very likely to exist,

but biological perturbations can have more complex behaviors that involve concurrent genes, especially in such partially observable systems consisting of many feedback loops.

Chapter 4

Predicting the Transcriptional Network of the Barley Circadian Oscillator

Adapted from: Lukas M. Müller*, Laurent Mombaerts*, Artem Pankin, Davis Seth, Alex A. R. Webb, Jorge Goncalves and Maria von Korff. Dynamic modelling of the barley circadian clock and transcriptome rhythmicity analysis reveal differential effects of the day-night cues and circadian clock on gene transcription. *(Submitted to Plant Cell)*

4.1 Contribution

As a first step in the characterization of a novel complex system, this analysis aims at inferring the main gene regulatory interactions that shape the circadian network of barley.

The All-to-All methodology was used together with the ν-gap to infer for the first time a transcriptional network between circadian genes in barley. For this purpose, the capability of the developed methodology to recover accurately few links with high confidence from limited data, together with its scalability, flexibility and interpretability, were of particular importance for this study. Indeed, multi-input LTI systems were subsequently built to explicitly integrate light patterns and further characterize genes dynamics in the whole transcriptome. The respective contribution of each input was visualized through Bode plots and their response magnitude quantified. This analysis enabled to mathematically support that an unneglectable part of the genes for which the phase is not correlated between diel and free-running conditions were mostly driven by external light.

4.2 Introduction

Barley (*Hordeum vulgare L.*) crop is the fourth most important cereal (preceded only by wheat, rice and maize) and one of the most versatile cereal in such that it has adapted to different global climates outside regions where other cereals live, from artic to tropical regions [154]. It is a major source of animal feed and underlies the brewing industry, for which Europe is the leading exporter. It can also be consumed as human food directly but represents merely 6% of its production [154]. Identifying adaptation strategies is critical to mitigate the negative effects of climatic variability on agriculture [155].

In plants, the circadian system controls many agronomically important processes, such as metabolism, growth, photosynthesis, and flowering time [156]. It has been suggested that the circadian clock is key to improving adaptation and performance of crop plants [157, 35]. Putative circadian oscillator genes have been identified in the monocot crop barley based on their homology with the Arabidopsis clock genes [34, 158]. Although the circadian oscillator genes diversified via duplication independently between the monocot and eudicot clades, their structure and expression patterns remained highly similar [34, 157, 35]. For example, in monocots, the morning expressed MYB-like transcription factor CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) is the only ortholog of the Arabidopsis paralogs AtCCA1 and LATE ELONGATED HYPOCOTYL (AtLHY) [159, 34]. HvCCA1 overexpression in Arabidopsis causes arrhythmia, suggesting circadian functionality [160]. AtCCA1 and LHY suppress the PSEUDO RESPONSE REGULATORs (PRRs), which duplicated independently from three ancient PRR genes after the divergence of monocots and eudicots such that the orthologous relationship within the PRR3/7 and PRR5/9 clades of Arabidopsis and monocot plants cannot be immediately resolved [161]. Partial complementation of Arabidopsis prr7-11 by HvPRR37 suggests that the barley gene might retain some functionality of the Arabidopsis orthologue [160]. However, PRR37 orthologs in monocots, PPD1 in barley and wheat [162, 163] and SbPRR37 in sorghum (Sorghum bicolor) [164], are major determinants of photoperiod sensitivity and flowering time, whereas natural variation in PRR genes in Arabidopsis did not have any notable effect on flowering time [165]. For EARLY FLOWERING 4 (ELF4), which in Arabidopsis forms an evening complex with ELF3 and LUX ARRHYTHMO (LUX), several ELF4-like homologs in monocots exist, including HvELF4-like 4 that can complement an Arabidopsis Atelf4 null mutant [166, 167]. Circadian gene expression is altered in the two barley mutants, early maturity 8 and 10 (eam8, eam10) , which carry functional mutations in homologs of the Arabidopsis circadian clock regulators EARLY FLOWERING 3 (ELF3) and LUX ARRHYTHMO (LUX1), respectively [168, 169, 170].

We generated diel and circadian RNAseq datasets of four barley genotypes, the spring barley Bowman (BW WT) and three derived introgression lines with mutations in HvELF3 (BW290), HvLUX1 (BW284), and EARLY MATURITY 7 (EAM7) (BW287) [168, 170]. The candidate gene for EAM7 has not yet been identified, but loss of EAM7 function accelerates flowering by abolishing sensitivity to the photoperiod [171]. We used the RNAseq time-course data to analyse the effects of barley clock genes on diel and circadian transcriptome oscillations including changes in phase and period under constant conditions and light and dark cycles. Dynamical modelling allowed us to predict a molecular structure of the barley circadian oscillator and to uncover how circadian oscillator components interact with day/night cues to regulate the global transcriptome in barley.

4.3 Circadian and Environmental Regulation of the Barley Transcriptome

4.3.1 Rhythmic Analysis

To characterize oscillating barley transcriptomes, we generated the RNAseq datasets from the barley cultivar Bowman and the derived introgression lines carrying mutations in HvELF3 (BW290), HvLUX1 (BW284) and HvEAM7 (BW287) grown under two different conditions - diel night/day cycles (ND; 12h/12h) and under constant light and temperature (LL) (Figure 4.1). Among 18,500 transcripts expressed in all the investigated lines, 84% were scored rhythmic under ND in Bowman. By contrast, under LL, about 23% of the transcripts were rhythmic, which is a distinctive feature of clock-regulated genes. The gene ontology (GO) analyses revealed that, in Bowman under LL, the circadian controlled transcripts were primarily related to the processes of regulation of DNA-dependent transcription, translation, electron transport, signal transduction, responses to salt stress and cold, and metabolic processes, including amino-acid, sucrose and starch metabolism. The molecular functions of the circadian controlled transcripts in Bowman in LL were primarily represented by protein, zinc ion and ATP binding, DNA and nucleotide binding, and sequence-specific DNA-binding transcription factor activity GO terms (Figure 4.2).

We found that the majority of the transcripts expressed rhythmically under LL were also rhythmic in ND (20% of all the transcripts). This demonstrated that about one-quarter of the Bowman transcriptome is modulated by the circadian clock. However, the largest proportion of the rhythmic transcripts in ND required daily external environmental cues

for the rhythmic expression.

The large impact of external transitions on transcriptome oscillations independent of the clock was further supported by the analysis of the Hvelf3 plants deficient in the circadian clock regulation. In Hvelf3, no transcript rhythms were detected under LL demonstrating that a functional HvELF3 is required for self-sustained transcriptome oscillations in barley. Environmental cues under ND restored oscillatory dynamics in the Hvelf3 loss-of-function line with 83% of the global transcriptome being rhythmic in the BW290 plants. The number and the identity of oscillating transcripts were comparable between BW290 and Bowman plants under diel cycles. In Hvlux1 plants, only 2% of the expressed transcripts were rhythmic under LL suggesting that, like HvELF3, HvLUX1 is required for free-running oscillations under LL. Once again, ND cycles were sufficient to restore transcriptional rhythms in the Hvlux1 mutant, i.e. 75% of the transcriptome oscillated in BW284 plants under ND. Mutation of the EAM7 locus in BW287 reduced the pervasiveness of circadian transcriptional oscillations but did not completely abolish them because 8% of the expressed transcripts cycled in LL in BW287, about a third of the number of the oscillating transcripts in Bowman. Under ND, 80% of the global transcriptome was rhythmic in BW287 and 72% of the rhythmic transcripts were common between BW287 and the background Bowman plants.

Our data demonstrate that cycles of light and temperature and the circadian oscillator drive rhythmic expression in barley. HvELF3, HvLUX1 and EAM7 contribute to free-running oscillations under constant conditions while environmental rhythms are sufficient to drive rhythmic expression in the absence of a free-running oscillator.

4.3.2 Bimodal phase distribution

To investigate temporal expression patterns of the circadian-regulated transcripts under free running conditions, we estimated the phase and the period of every circadian-regulated transcript in the two genotypes that sustained free-running circadian rhythms, Bowman and BW287. In Bowman, the distribution of the circadian transcriptome expression phase followed a symmetrical bimodal pattern with the highest number of transcripts peaking shortly before the transitions to subjective days and nights (Figure 4.2a). By contrast, in BW287 this phase pattern of the cumulative circadian transcriptome was not evident (Figure 4.2a). These findings indicated that EAM7 is required to modulate the characteristic bimodal pattern of the circadian transcriptome expression in barley. The period estimates of the oscillating transcripts under LL ranged between 22 h and 34 h in Bowman and BW287 and followed a bell shaped distribution with mean periods of 27.5

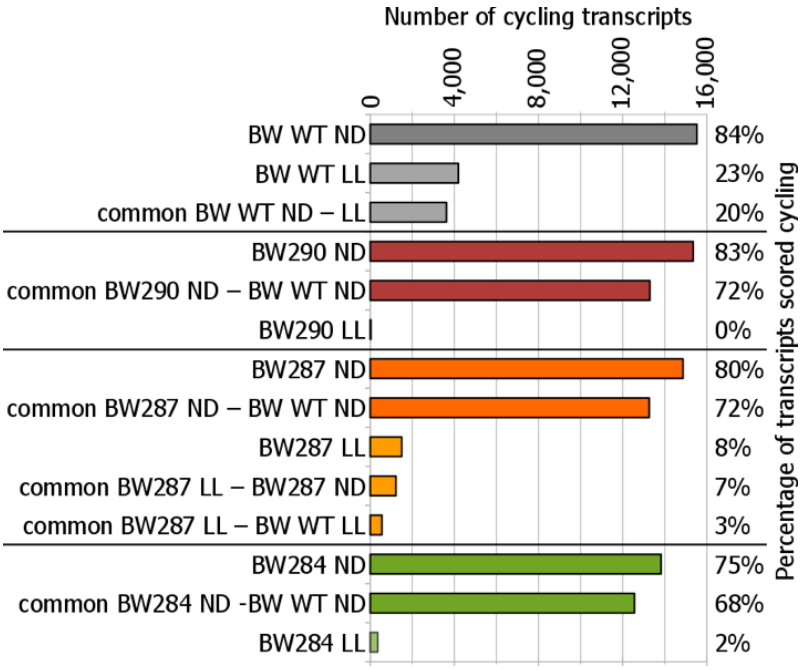

Fig. 4.1 **Fraction of transcripts detected as oscillating through computational analysis.** BW WT: Background Bowman; BW290, BW287 and BW284: Clock mutant genotypes in the Bowman background; ND: night/day cycles; LL: free-running conditions of constant light and temperature. Fractions refer to a total of 18,500 transcripts expressed in all genotypes.

h and 27.9 h in Bowman and BW287, respectively (Figure 4.2b, c). In both Bowman and BW287, the standard deviation of the period distribution was higher under LL (6 h) than under ND (2.5 h) (Figure 4.2b, c). This could arise from either the uncoupled nature of cellular oscillations in free-running conditions or is a consequence from the period estimation as the signal amplitude was lower in LL than in ND. A longer mean period of oscillating expression patterns in BW287 suggested that the free-running period under LL was extended in BW287 compared with Bowman.

4.3.3 Phase regulation

Next, we investigated the transcriptome oscillations under the diel ND conditions. In all genotypes, including those that were arrhythmic in LL, the mean of the period distribution was consistent with the enforced 24-h diel cycle and ranged between 23.5 and 23.6 h. The phase was bimodally distributed over the day/night cycle in Bowman so that for the highest number of transcripts the peak of expression occurred before dawn and dusk and, the number of transcripts with the peak expression during the night and day was the lowest (Figure 4.3a). This pattern was comparable with the phase distribution under LL

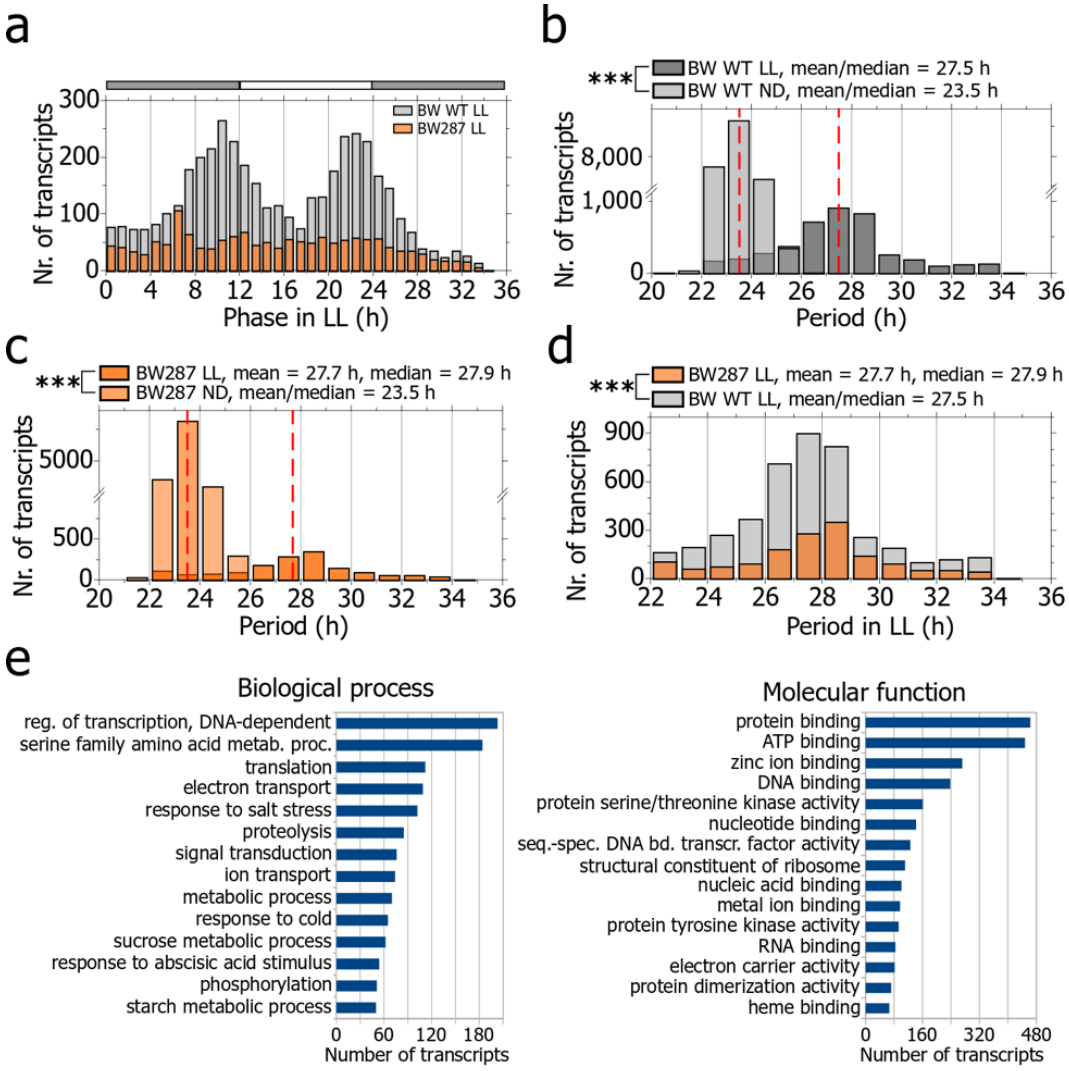

Fig. 4.2 **Distribution of the period and the phase of the oscillating transcriptome in constant light and their involvement in biological processes and molecular functions. a)** Phase distribution in constant light (LL) of Bowman (BW WT) and BW287. Grey-white bars indicate the subjective night (grey) and subjective day (white) in constant light conditions. **b, c)** Period distribution of the oscillating transcriptome in constant light (LL) in comparison with night/day cycles (ND) in b) Bowman wild-type (BW WT) and c) BW287. **d)** Comparison of the period distribution in constant light (LL) between Bowman (BW WT) and BW287. **e)** Top-15 categories of the GeneOntology terms for biological processes and molecular function of the transcripts oscillating in Bowman (BW WT) in constant light (LL).

(Figure 4.2a) and the transcripts that oscillated in both LL and ND were also bimodally distributed under the diel cycles (Figure 4.3a). This suggested that the bimodal distribution of transcriptome-wide gene expression is, at least partly, under control of the circadian clock.

The analysis of the clock mutants, however, suggested that the bimodal phase distribution under ND is controlled by both the circadian clock and day/night cues. In Hvelf3 the phase was bimodally distributed under diel cycles similar to Bowman, however the quantitative characteristics of the phase distribution differed. Namely, in Hvelf3, the phase distribution showed higher peaks at dawn and dusk and deeper troughs during the night or the day than in Bowman (Figure 4.3b). A large number of the transcripts that peaked around the night-to-day and day-to-night transition in Hvelf3 (Figure 4.3b) peaked during the day or the night in Bowman (Figure 4.3c,d). his demonstrated that HvELF3 modulates timing of peak expression of multiple transcripts in day/night cycles. This effect was apparently completely or partially independent of the oscillator defect that causes arrhythmia in the Hvelf3 plants under LL since the phase distribution in Hvlux1 mutants under ND was comparable to Bowman (Figure 4.3f), even though self-sustained circadian oscillations were also absent in this genotype under LL conditions (Figure 4.1). This was also evident from the transcriptome-wide comparison of the phase between the barley clock mutants with Bowman under ND. Here, the phase distributions strongly correlated between Hvlux1 and Bowman (Pearson correlation ρ=0.97,R^2=0.94) while the phase distributions in BW290 and Bowman were correlated to a lower degree (Pearson correlation ρ=0.93, R^2=0.86), even though both mutant genotypes harbor an arrested oscillator under LL conditions (Figure 4.1).

Day/night cycles had strong effects on the phase distribution of the transcriptome as demonstrated by the analysis of the BW287 (eam7) transcriptome. Whereas the phase distribution was not bimodal in BW287 under LL (Figure 4.2d), under ND, the phase distribution was bimodal similar to the one in Bowman (Figure 4.3e). Consistently, the phase distributions under ND were highly correlated between BW287 and Bowman (Pearson correlation ρ=0.96,R^2=0.92). Consequently, external cues under ND controlled the phase of the global transcriptome in BW287 to peak at the night/day transitions despite the circadian defects observed in BW287 under LL. Together, these results demonstrated that the bimodal distribution of the phase in diel cycles is controlled by both day/night cues and the clock component HvELF3. The genetic defects and their underlying circadian phenotypes in BW284 and BW287 have limited effects on the phase of the global oscillating transcriptome in diel cycles despite their strong transcriptional phenotypes under LL.

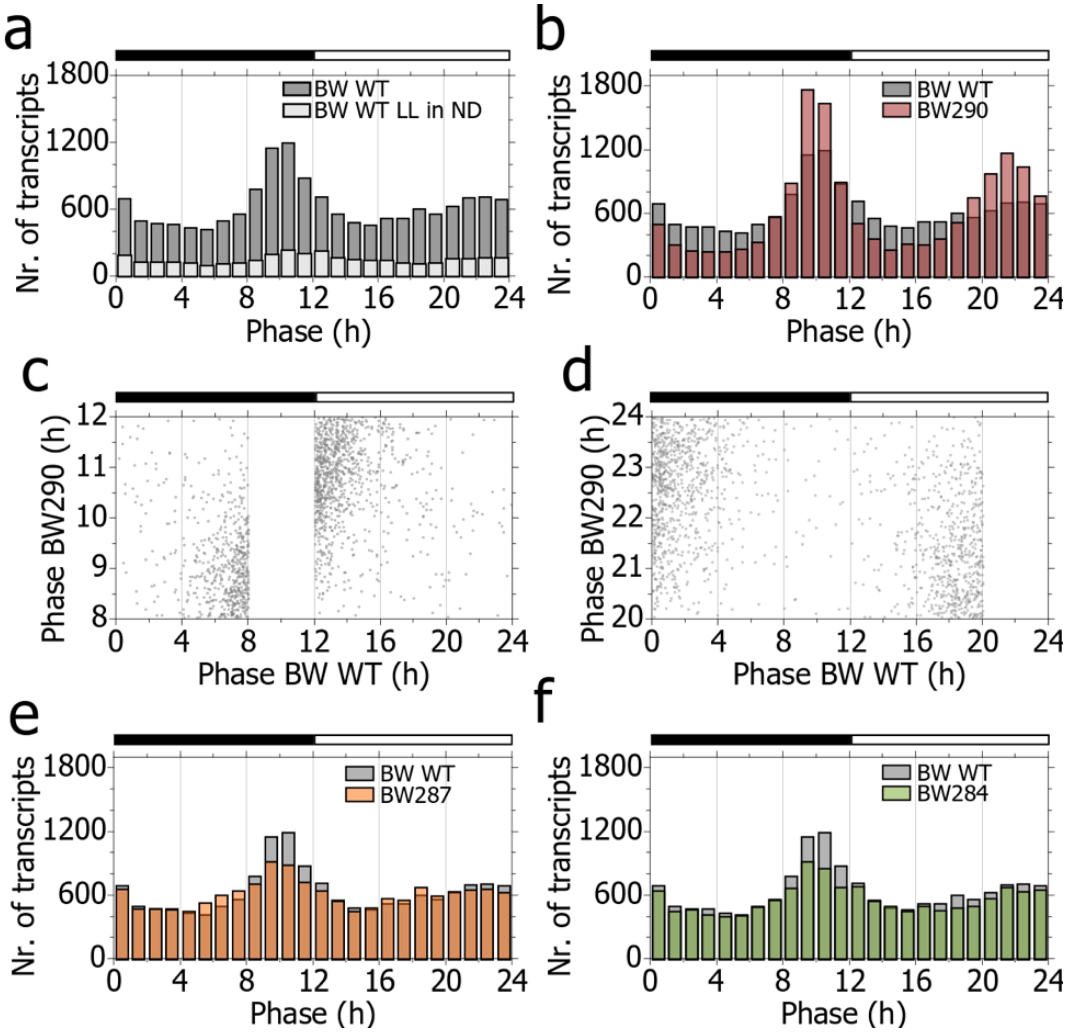

Fig. 4.3 **Distribution of the phase of the oscillating transcriptome in night/day cycles. a)** Phase distribution in Bowman in night/day (ND) cycles for the global oscillating transcriptome (BW WT) and those transcripts detected oscillating in both night/day cycles and constant light (BW WT LL in ND). **b), e), f)** Phase distribution in diel cycles in a) BW290, e) BW287 and f) BW284 in comparison with the Bowman wild-type background (BW WT). **c), d)** Detailed views of b). Time point of peak expression (phase) of transcripts that peaked in Bowman (BW WT) but not in BW290 before the c) night-to-day transition (time points between 8h-12h in b) and d) the day-to-night transition (time points between 20h-24h in b)).

4.4 Prediction of the Central Clock Mechanisms of Barley

4.4.1 Inferring Barley Clock Components

To reduce the identification of erroneous interactions we filtered all circadian transcripts for those that were homologous to Arabidopsis genes representing transcription factors that were labeled "circadian" (www.geneontology.org), resulting in 138 transcripts. Importantly, only genes that exhibit unambiguous dynamics are further considered, in the sense of signal to noise ratio. This filtering step is necessary to ensure that we are not identifying dynamics out of noise. Hence, genes for which their amplitude of oscillation is lower than an arbitrary set value of 20 CPM in the last 24 hours were removed. The choice of such filter is motivated by both the transitional nature of constant light data, which typically shows a large decrease of amplitude after few hours in barley, and the dependency of noise on gene expression level. Furthermore, genes that are constantly up/down regulated without exhibiting further significant dynamics were similarly discarded. This was performed by detrending the 24 last hours of constant light data before applying the same filtering criterion. Consequently, out of 138, 49 and 47 genes passed the filtering criterions respectively in WT and M287 datasets. M284 (LUX mutation) and M290 (ELF3 mutation) datasets were not considered in the following network inference analysis as the clock has been perceptibly broken by such mutation. Finally, 7 genes (Hv.21080, Hv.22191, Hv.23289, Hv.32914, Hv.33010, Hv.6793, MLOC7084.3) were manually discarded from both subsets list of candidates as they were not DNA binding transcription factors but rather enzymes in a metabolical process, leaving 42 and 40 transcripts for modeling respectively in the WT and M287 dataset.

Our data suggested that HvELF3 and HvLUX1 are integral components of the barley oscillator as they were necessary to sustain transcriptome oscillations under LL (Figure 1). Therefore, we hypothesized that modeling a transcriptional network around HvELF3 and HvLUX1 could identify the regulatory relationships that shape the circadian clock in barley. We followed an approach that searches the dynamic dependencies of HvELF3 and HvLUX1 expression on other transcripts. Unfortunately, ELF3 transcript was discarded through the filtering step and could not be used to infer dynamical interactions. Then, to investigate the potential regulators of LUX, a collection of independent 1st order LTI models was estimated from each of the transcript to LUX in the Bowman background. In each case, the parameters are estimated so that they together provide the best possible fit to the LUX time course data. This step takes the following form:

$$\begin{aligned} \frac{[LUX]_t}{dt} &= a_1 u_1(t) - b_1[LUX]_t + c_1 \\ &\dots \\ \frac{[LUX]_t}{dt} &= a_n u_n(t) - b_n[LUX]_t + c_n \end{aligned} \tag{4.1}$$

where n corresponds to the number of candidates, so that 42 models are finally computed. Each model is characterized by a fitness metric that ranges from 0 to 100%, representing its capability of describing the original regulatory system between genes. A gene, therefore, would be further considered as a regulator for LUX if the model is capable of reproducing the shape of LUX with a sufficient degree of precision, which is characterized by a high goodness of fit. In this case, the fitness threshold was set to 60% to restrain false positives predictions of regulatory interactions while accounting for sufficient gene regulatory models to describe the system of interest. The threshold was chosen so that the inference strategy correctly identified 62% of the links it predicted, which corresponds to 38% of the entire circuitry of the network being accurately recovered (including the distinction between inhibition and activation) based on the in-silico benchmarks of Chapter 2. This strategy aims at keeping the links with the highest confidence only. These numbers are remarkable, considering the complexity of the network and the amount of different experimental conditions investigated. As a result, 20 models were validated.

4.4.2 Predicting the Circadian Transcriptional Network

To obtain a whole system representation of the regulatory interactions involved in the barley circadian clock, the interactions between the potential regulators for which the corresponding model was validated were estimated, as in (Equation 4.1). This step produced a total amount of (21)*20 = 420 models (as we did not consider self-regulation, and LUX is included to evaluate a potential feedback to its regulators), among which 79 were validated using the same fitness threshold than previously. We further narrowed down the resulting regulatory interactions to the most relevant ones by estimating the consistency of these models using the filtered M287 dataset. For this purpose, we identically evaluated 1st order LTI models for each of the previously identified regulations, when possible, and evaluated their goodness of fit in the M287 experimental condition. A fitness threshold of 60% was then applied on all the models computed. The resulting network is represented in Figure 4.4.

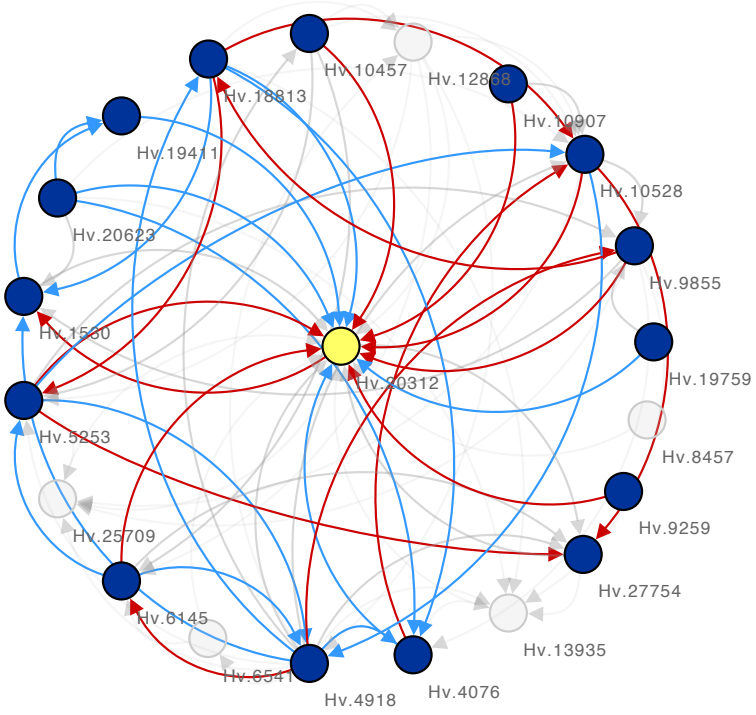

Fig. 4.4 **Predicted regulatory interactions between the regulators of LUX1 that have been identified in Bowman and further validated using the BW287 dataset.** The grey nodes represent those transcripts that did not pass the noise filter in constant light, in BW287. The grey links represent those models for which their counterpart was not validated using the BW287 dataset. Colored links represent the models that have been validated in both Bowman and BW287. A blue arrow represents an activation while a red arrow represents repression.

The networks computed from Bowman and BW287 were highly comparable so that we only considered those regulatory links for our further analysis that were identified in both datasets. Finally, we compared the dynamical features of pairs of analogous models (WT and M287) using the well-established metric called the ν-gap. As to keep the links of highest relevance, we considered models that hold a small ν-gap value only, as in [79]. Indeed, high gaps values can either suggests a link being affected by the mutation or a sign for a false positive. Hence, it is reasonable to discard the high gap links from this network as a first attempt of unveiling the core mechanisms of the clock. [93] suggested that values above around 0.2 could be used to infer the main target of perturbation. Therefore, models holding a ν-gap above 0.2 between WT and M287 conditions were discarded from the core network. This quality check is motivated by the assumption that two LTI models that share identical dynamical properties while describing the same genetic regulation, with a relatively high goodness of fit, in two different experimental conditions are more likely to correctly identify the regulatory dynamics between the genes, thereby reducing the chance

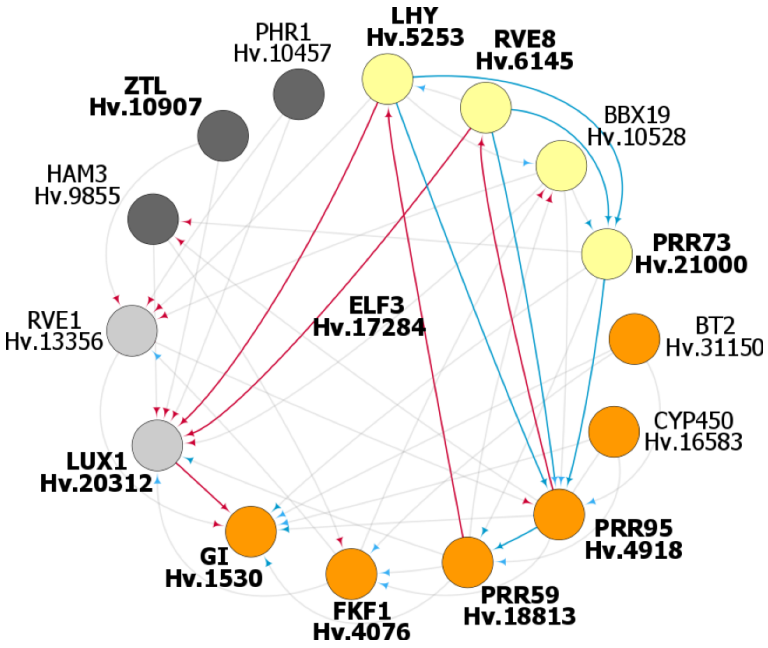

Fig. 4.5 **The putative circadian network of the barley oscillator as predicted from time series expression data. Genetic evidence but no model prediction allowed placing HvELF3 as a core clock component.** The figure displays the inferred components and interactions that constitute the barley circadian transcriptional network. Circadian clock components are represented by circles and sorted in clockwise direction for the time point of peak expression starting with HvLHY at dawn (yellow: morning, orange: evening, grey: night). The regulatory interactions are represented by directed arrows, where activation is marked in blue and inhibition in red. The components printed in bold and the links highlighted in color are consistent with key components and key regulatory principles present in circadian clock models from Arabidopsis. The clock components in barley are named after their closest Arabidopsis homolog and identified by barley UniGenes.

of identifying inexistent regulatory interactions. As a result, 6 regulatory interactions were filtered out (Hv.10528 to Hv.27754, Hv.1530 (GI) to Hv.19411, Hv.19411 to Hv.20312 (LUX), Hv.19759 (TOC1) to Hv.20312 (LUX), Hv.5253 (LHY) to Hv.27754, Hv.9855 to Hv.18813 (PRR59)). It is interesting to note that this is where the only link of TOC1 with the central clock is lost. On the resulting conjunction network, we noted that PRR95 (Hv.4918) appeared as a hub with 8 connections, while the mean node connectivity being 3. Providing that LUX has 11 connections while being the origin of the graph, this suggested a significant role for PRR95 in the regulation of the core circadian genes. We repeated, therefore, the search for regulators of PRR95 (as in Equation 4.1), computed their interactions in both datasets and checked their consistency. Consequently, 4 genes were added to the final network (RVE1 (Hv.13356), Hv.16583, PRR73 (Hv.21000) and

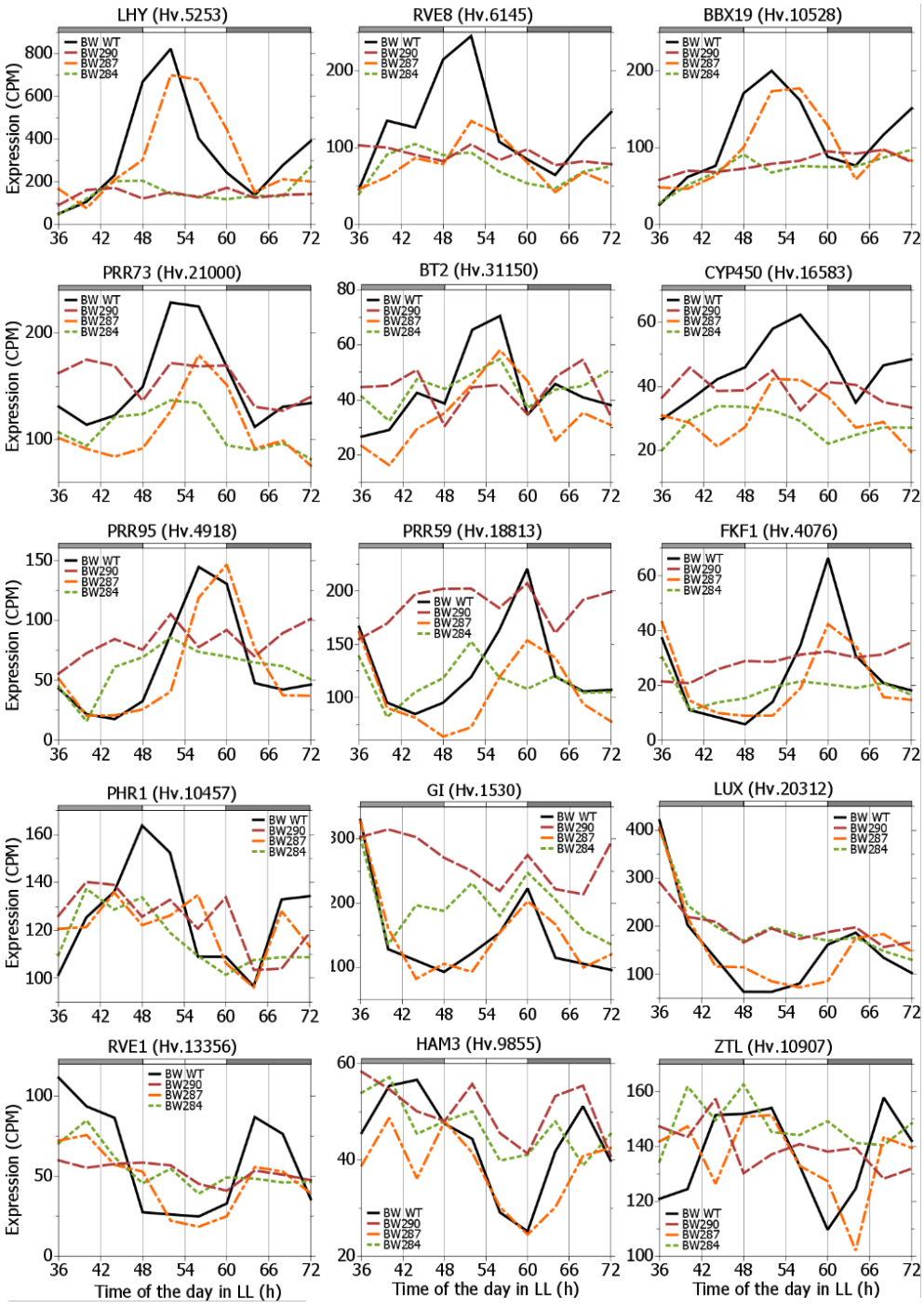

Fig. 4.6 **Expression profiles of predicted core clock components of the barley oscillator in free-running conditions of constant light.**

BT2 (Hv.31150)) (Figures 4.5, 4.6). Further extension of the network around other genes did not include additional genes to the core clock system.

In addition to LUX1 as a core of the model (Hv.20312), the predicted components of barley circadian clock were barley homologs of LHY (Hv.5253) [172], REVEILLE 8 (RVE8) (Hv.6145) [157], PRR73 (Hv.21000) [173], PRR95 (Hv.4918) [173], PRR59 (Hv.18813) [174], FLAVIN-BINDING, KELCH REPEAT, F-BOX 1 (FKF1) (Hv.4076) [175], GIGANTEA (GI) (Hv.1530) [87], and ZEITLUPE (ZTL) (Hv.10907) [176]. Homology of the predicted barley clock components with well characterized Arabisopsis core clock genes validated our approach to predict core components of barley clock. Therefore, based on the timing of the peak expression starting with HvLHY expression at subjective dawn, we arranged the predicted components into a model of the barley circadian clockwork (Figure 4.5).

In addition to the barley homologs of known Arabidopsis oscillator genes, our analysis suggests several previously uncharacterized components of barley circadian clock – HvBBX19 (Hv.10528) and REVEILLE1 (HvRVE1) (Hv.13356) (Figure 4.5). In the model, both HvBBX19 and HvRVE1 regulate HvPRR95 and are regulated by HvLHY (Figure 4.5). Such connectivity of BBX and RVE1 with the known clock components suggests that they might be a part of the oscillator network in barley. AtBBX19 and AtRVE1 have been proposed to have connections to the Arabidopsis oscillator [177, 178], suggesting that our network modeling has identified two candidate oscillator components in barley. The modeling predicted that HvRVE1 represses HvPRR95 and HvBBX19 activates HvPRR73 and HvPRR95. Another predicted component of barley circadian clock was a homolog of HAIRY MERISTEM3 (HAM3) (Hv.9855). Based on the model, HvHAM3 is regulated by HvPRR73 and HvPRR95 and regulates HvLUX1 and HvFKF1 (Figure 4.5), whereas, in Arabidopsis, HAM3 plays a role in cell differentiation and cell polarity. The model predicted that barley homologs of BTB AND TAZ DOMAIN PROTEIN 2 (BT2) (Hv.31150), CYTOCHROME 450 (CYP450) (Hv.16583), and PHOSPHATE STARVATION RESPONSE 1 (PHR1) (Hv.10457) are part of the core circadian oscillator in barley. However, all of these genes were predicted to regulate clock components but were not regulated themselves by the clock genes (Figure 4.5).

4.5 Modeling the Effect of the Light Signaling Pathway

Hereafter, the flexibility of the modeling strategy considered is exploited to model the relative contribution of light signaling to circadian regulated genes. To this end, we used

3642 transcripts that were identified as oscillating in both diel and free-running conditions in the wild-type Bowman background. As a reference, we selected a formerly identified clock gene peaking in the morning, HvLHY (Hv.5253), with a range of delays integrated into the model to implicitly represent the clock input, as devised in [82]. The structure of such models is schematically represented in Figure 4.7A. This way, the light input is incorporated on two levels: explicitly, through the light input and implicitly through the clock pathway. Mathematically:

$$\frac{dy(t)}{dt} = a_1 u_{light}(t - \mu_{light}) + a_2 u_{LHY}(t - \mu_{light}) - by(t)$$

Where μ_{light} was assumed to be binary (1 = light; 0 = dark). We fixed the light delay μ_{light} to 0h to represent the effect of rapid light signaling on the transcripts, and computed delays ranging from 0 to 8h, every 0.2h, for LHY. The delay value that provided the best fit to the data was selected independently for each transcript. Ultimately, models were validated if they succeeded in capturing the regulatory dynamics involved with a goodness of fit > 60%.

In systems theory, Bode plots are used to visualize the frequency response of linear models. The frequency response of a model represents the response of an input signal through the model. The magnitude and phase of the resulting output signals are therefore visible for each possible input. In our case, we use the magnitude response of the signal to assess the relative contribution of the inputs $u_{light}(t)$ and $u_{LHY}(t)$ in each of the validated model, at a frequency of 24h (or .262 rad h^{-1}). The contribution of each input is computed in dB (decibels). In [82] it was observed that, for Arabidopsis, a threshold of 7 dB could be used to differentiate the contribution of each of the input signals. Therefore, if the magnitude of the response of the light input was 7dB higher than the contribution of the clock (represented by LHY potentially delayed), the circadian regulated gene (the output of the model) was considered mostly driven by light. Conversely, it was considered as being driven by the clock. If the magnitude difference was lesser than 7dB, then the circadian regulated gene was considered regulated by both inputs equally. The methodology is summarized in Figure 4.7B as well as a presentation of the results.

The analysis estimated that 43% of the transcripts that oscillate in both day/night cycles and constant light were predominantly controlled by the circadian clock in light/dark cycles and that 47% were co-regulated by the circadian clock and light/dark cues (Figure 4.8a). Only 10% of the transcripts were primarily controlled by light/dark cues (Figure

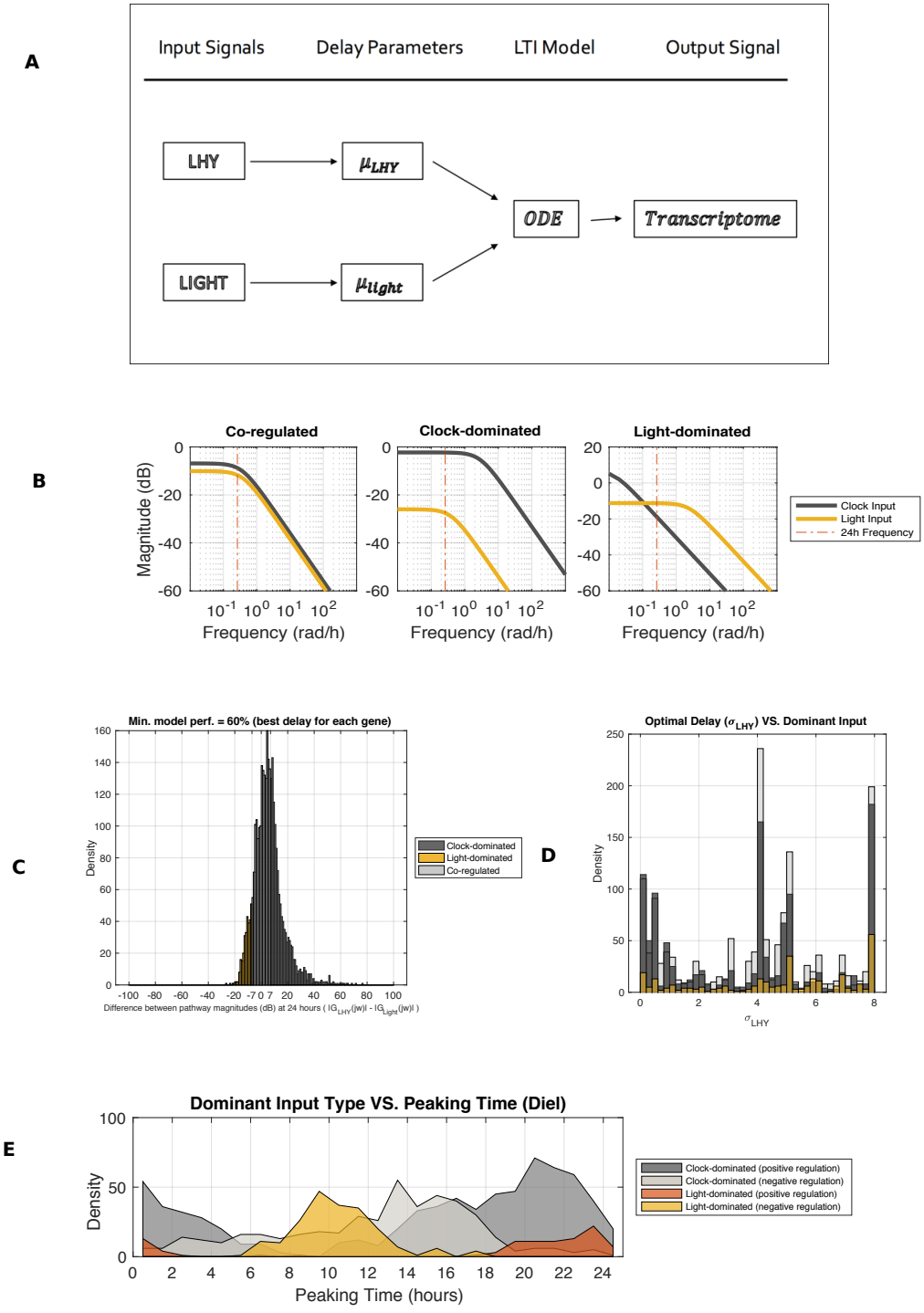

Fig. 4.7 **Modeling the light regulation.** **(A)** Schematics of the model class **(B)** Examples of Bode magnitude graphs describing each category of dominant input (light-dominated, clock-dominated, co-regulated) **(C)** Distribution of differences in Bode amplitude for each transcript. **(D)** Optimal delay for each model, displayed for each dominant input. **(E)** Correlation between peaking time and dominant input type.

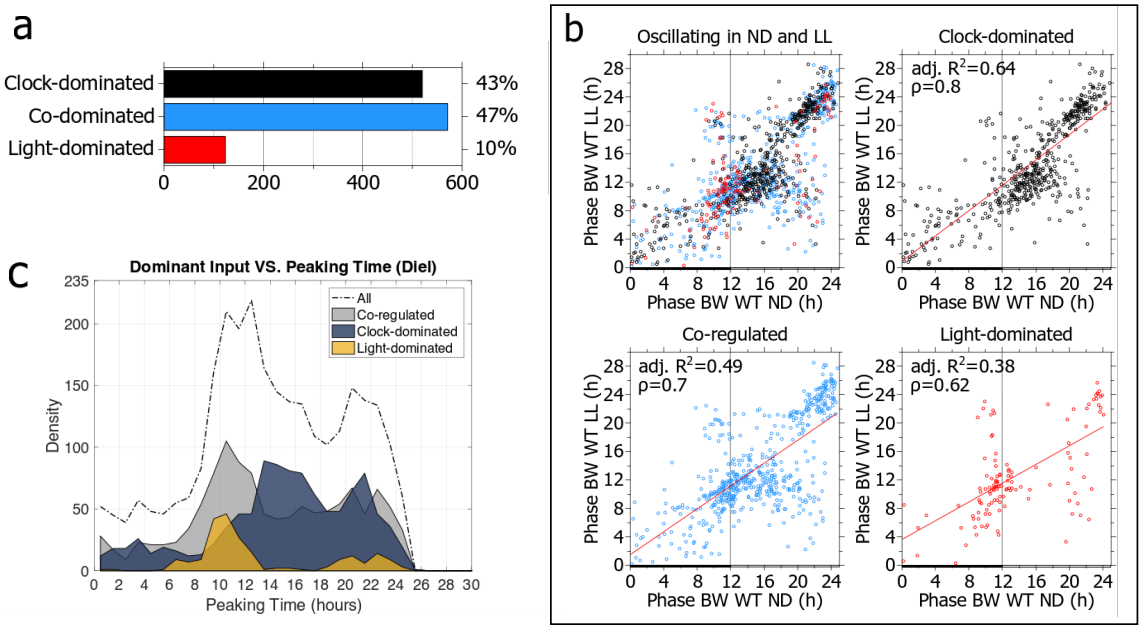

Fig. 4.8 **Relationship between external and internal cues to regulate the phase of the barley transcriptome. a)** Fractions of transcripts identified as clock-dominated, co-dominated by the clock and light and light-dominated by the Bode-analysis. **b)** Phase relationship between diel cycles (ND) and constant light (LL) for all transcripts oscillating in ND and LL and those dominated by the circadian clock, co-regulated by the circadian clock and light and light-dominated. **c)** Phase distribution of clock-dominated, co-regulated and light-dominated transcripts in diel cycles (ND).

4.8a). This is consistent with the expected underrepresentation of light/dark controlled transcripts in a set of genes that oscillate in the absence of environmental cues.

We also investigated the phase relationship between driven and free-running conditions for transcripts predicted to be under the clock control, light control and co-regulation by light and the clock by the Bode analysis (Figure 4.8b, c). The clock-dominated transcripts revealed the highest correlation of the phase between day/night cycles and constant light (R^2=0.54, Figure 4.8b) and the light dominated transcripts the lowest (R^2=0.38, Figure 4.8b). The correlation of the phase of co-regulated transcripts lay in between (PR^2=0.49, Figure 4.8b). This is consistent with the anticipated regulation because transcripts dominated by the circadian clock are expected to preserve the phase against changing light conditions, whereas transcripts dominated by light cues are expected to reflect the changes in light. These findings suggested that the Bode analysis predicted the main regulatory principles that determine the phase of oscillating transcription in day/night cycles. Namely, it suggested that about 40% of the transcripts with clock-

maintained oscillations reveal a phase dominated by the circadian clock in diel cycles. For the remaining 60% of the transcripts with clock-maintained oscillations, the peak of their expression is under the control of light signaling pathways or co-regulated by light signaling and clock. This finding highlights the importance of light signaling pathways to regulate the phase of oscillating transcription even for the transcripts, which rhythmicity is maintained by the circadian clock.

4.6 Discussion

The circadian clock was estimated to control ∼25% of the expressed transcripts under constant conditions and regulated transcripts to peak in a bimodal pattern before subjective dusk and dawn. The circadian-controlled transcripts were primarily related to the processes of DNA dependent transcription, translation, amino-acid and carbon metabolism, stress responses, electron transport, and signal transduction. Similarly, Arabidopsis, rice (Oryza sativa) and poplar (Populus trichocarpa) exhibited circadian regulation of between 8 and 30% of the global transcriptomes [122, 179, 180, 181]. Furthermore, the cycling transcriptomes in Arabidopsis and rice were also enriched for transcripts involved in transcription, translation, and amino acid and carbon metabolism suggesting the conservation of the circadian control of metabolic pathways and transcriptional networks among mono- and dicotyledonous plants [179, 182, 183].

Our data demonstrated that the expression phase under LL conditions was generally not a strong predictor of the transcript phase under ND conditions. However, the phasing of clock-controlled transcripts showed higher correlations between constant and day-night conditions than that of light-controlled transcripts indicating that the clock influences expression phases of circadian transcripts also under diel conditions. Overall, day/night cues imposed the strongest control on transcriptome oscillations. Foremost, the hvelf and hvlux1 mutants with no cycling transcriptome under LL conditions, were characterized by transcriptome oscillations under ND comparable to wild type Bowman. Second, the majority of circadian transcripts was regulated by light/temperature or a combination of the clock and light/temperature cues under ND conditions. In this context it is interesting to note, that hvelf3 and hvlux1 mutants with a disrupted circadian clock, have been used to breed for barley cultivars adapted to Northern European environments with strong diurnal and seasonal changes in light and temperatures [168, 170, 184]. Neither of the two arrhythmic mutants (hvelf3, hvlux1) have been reported to display any obvious impairment in photosynthesis and growth under conditions of pronounced photo- and thermocycles in contrast to the corresponding Arabidopsis mutants [168, 170, 185]. Similarly, [186] have

reported that an osgi mutant in the field was not affected in photosynthesis and yield. Only under atypical growing conditions with late transplanting dates in the field, fertility was significantly reduced in osgi plants, indicating a loss of seasonal adaptability. Our data suggested that diel cycles could compensate circadian defects in the barley clock mutants, increase the number of oscillating transcripts compared to free-running conditions and strongly influence the time point of transcript peak expression.

While the number of cycling transcripts was not different between the hvelf3 and hvlux1 mutants and Bowman, we observed quantitative variation in the phase distribution under diel conditions between the three genotypes. HvELF3 altered the timing of transcript oscillations in day/night cycles by suppressing expression at the light and dark interfaces. This effect was apparently completely or partially independent of the oscillator defect that causes arrhythmia in the hvelf3 plants under LL. Loss of HvELF3, but not of HvLUX1, altered clock gene expression and transcriptome regulation in diel cycles although both mutants had a disrupted circadian clock. These results suggest that HvELF3 is a strong regulator of the diel transcriptome in barley and that this regulation is uncoupled from the role of HvELF3 in the endogenous oscillator. Our data supported the notion that HvELF3 in barley, similar to Arabidopsis [187, 188], mediates diel inputs of light and temperature into the oscillator during the night and that this regulation mitigates or even complements defects in the circadian network. This would explain why clock gene expression and transcriptome regulation was restored to wild-type levels in day/night cycles in the hvlux1-mutant BW284 and the clock mutant BW287 but not the Hvelf3-mutant BW290.

We also observed that HvELF3 has a specific function in the distribution of peak expression of the global transcriptome because only the Hvelf3-mutant but not the hvlux1- or eam7 mutants revealed transcriptional phenotypes in diel cycles. This is interesting because out of the Arabidopsis core components of the Evening Complex (EC) ELF3-ELF4-LUX only LUX has been identified as a transcription factor with direct DNA binding activity [189]. It has been shown that LUX provides DNA binding specificity for the EC at a large number of loci and recruits ELF3 to target loci [190]. On the other hand, chromatin immune precipitation experiments demonstrated that ELF3 had many more significant binding sites than LUX suggesting that ELF3 also binds independently of LUX [190]. Our results suggest that ELF3 has a strong effect on the transcriptional regulation of many different target genes. The different effects of hvelf3 and hvlux1 on the diel transcriptome may also be caused by the different nature of the underlying mutations, while the hvelf3 mutant line carries a premature stop codon leading to a truncated HvELF3 protein, the hvlux1 mutant is characterized by a single amino-acid exchange

in the Myb-domain which is important for the binding to cognate DNA sequences and regulation of their target genes [168, 170]. Generally, it was reported before that different clock mutants (lhycca1, prr7prr9, gi and toc1) affected specific sets of genes and proteins [183]. Consequently, different clock genes may control very different functions and output targets apart from their common role in maintaining circadian clock oscillations.

Based on RNA time series data we modeled a possible barley clock as a basis for understanding its effects on physiology, metabolism, and agronomic performance. It is important to emphasize that the resulting interactions between the individual components of the clock represent one of the possible solutions of the barley circadian clock circuit, which may serve as a null model in future studies aimed to experimentally resolve composition and regulation of this clock. The identification of gene regulatory networks is a major challenge of systems biology. The methodology followed here (referred to as "All-to-All" in Chapter 2) aims at providing reliable predictions of interactions between genes given the specific informative potential of the generated dataset. Furthermore, it does not rely on prior knowledge of the network, and is therefore unbiased.

Our modeling strategy used HvLUX1 to reveal the circadian circuitry, which therefore appeared as a major hub in the barley clock. Nevertheless, this predicted central role of HvLUX1 is consistent with the loss of self-sustained rhythms in the hvlux1 mutant. Unlike HvELF3 and HvELF4, HvLUX1 comprises known DNA binding domains suggesting that the transcriptional regulation of the EC converges on HvLUX1 [191]. Our model predicted that HvLUX1 represses HvGI and is itself repressed by HvLHY, consistent with the suggested repression of HvGI by the EC and CCA1/LHY repressing the Evening Complex in Arabidopsis [157, 33].

Further, the regulatory predictions suggested that HvLHY and HvRVE8 are activators of HvPRR73 and HvPRR95 in the morning and, at the same time, repress HvLUX1. The morning activation of the HvPRRs through HvLHY and HvRVE8, together with the repression of HvLHY and HvREV8 through the HvPRRs later in the day, are also a key regulatory principle of the Arabidopsis clock [33, 157]. This suggests that the regulatory links between HvLHY, HvRVE8, and the HvPRRs are conserved between barley and Arabidopsis, despite the independent evolutionary history of LHY-like and PPR-like genes in the barley and Arabidopsis clades [34, 159, 161].

Our model suggested that HvPRR73, the first PRR expressed in barley in the morning, activates HvPRR95, which, in turn, activates HvPRR59 such that HvPRR73, HvPRR95 and HvPRR59 are expressed in a sequential cascade. This resembles predictions by [31]

who described the PRRs as a series of activators in the Arabidopsis clock, while other models have predicted that direct interactions between the PRRs are negative and directed from the later PRRs in the sequence to the earlier ones [33, 192, 193]. However, the sequential regulation of the PRRs during the day appears to be a common feature of the circadian clock in both barley and Arabidopsis, while the sequence of expression of PRR genes is altered between Arabidopsis and barley. In Arabidopsis, the sequence of PRR expression starts with PRR9 and ends with PRR5, while it the sequential PRR expression wave started with PRR73 and ended with PRR59 in our data [157]. Despite up-regulation of PRR73, PRR59, and PRR95 in the hvelf3 and hvlux1 mutant plants, our model did not predict repression of the PRR genes by HvLUX1, which is a key feature of the Arabidopsis clock [33, 157]. However, while our LTI modeling strategy can reliably identify few links with high confidence with respect to the informative potential of the dataset investigated, it is likely that very complex regulatory interactions might not be identified.

BBX19, RVE1, and HAM3 had several connections to putative barley core clock genes suggesting that our network modeling identified three new candidate oscillator components in barley. While the three genes have already been proposed to have connections to the Arabidopsis oscillator, they have not been modeled as an integral part of the circadian clock but rather as clock outputs in Arabidopsis [177, 178]. HAM3 has been described as a gene controlling cell differentiation and polarity in Arabidopsis, but it shows diurnal and circadian expression oscillations that are consistent with a gene involved in circadian clock regulation [194]. RVE1 is a transcription factor homologous to the central clock genes CCA1/LHY and RVE8 and might have evolved functions in the circadian clock in barley [177]. BBX19 with two conserved zinc finger B-boxes acts as a gatekeeper of EC formation by mediating degradation of ELF3 [143] and is together with BBX32 co-expressed and forms a protein complex with LHY in poplar. Its close homolog, BBX32, is part of a regulatory loop with CCA1and/or LHY, because overexpression of BBX32 increases both their expression and circadian period length [195]. Further work will be required to test the hypothesis on additional clock genes generated by the network modeling. The model also predicted that barley homologs of BT2, CYP450, PHR1 are part of the core circadian oscillator in barley. However, these genes were only predicted to regulate other clock components and were not regulated themselves by clock genes. They therefore displayed a low connectivity within the circadian network, consistent with their known functions outside the central clock [196, 197, 198]. Therefore, these components might provide input into the circadian network but are probably not components of the barley oscillator.

4.7 Strengths and Limitations of the Study

The present study holds the following strengths and limitations:

- Strengths
  - The circadian system is key to improving adaptation and performance of crop plants. However, the core components of the clock, their interactions and the integration mechanisms of the light input are largely unknow. The findings and datasets presented herre are a valuable resource for exploring the circadian regulatory systems in crop plants.
  - The performances of our methodology were evaluated on in silico time series data generated in condition that replicated those of the current dataset (no prior knowledge of the system, 48 hours of transient data, 4 hours sampling rate). As a result, the modeling strategy optimizes the informative potential of the dataset.
  - Another advantage of the methodology here lies in the fact that the interactions suggested are independent of each other. Hence, invalidating one causal interaction between genes does not affect the others. This approach was particularly relevant to construct a network of candidates around LUX and for the general analysis of a novel complex system.
  - The findings of the key components of the clock are consistent with the current understanding of the genes involved in clock systems in Aradidopsis.
  - The flexibility of the methodology further allowed to study the contribution of the light input on the regulation of the transcriptome.
- Limitations
  - From a general point of view, while many interactions have been proposed, extensive validations will be required to sequentially build knowledge of this novel system.

Part II

Epileptic Seizure and Epileptogenesis Characterization

Chapter 5

Mathematical Preliminaries

5.1 Introduction

This chapter describes the mathematical preliminaries underlying the following investigation of both seizure characterization and early detection. Characterizing seizures signature and the subtle signal abnormalities emerging during epileptogenesis requires tools that allow to address the highly complex and multivariate nature of electrophysiological data. Machine learning provides such tools that can identify complex patterns in very large amounts of data. Machine learning plays a central role in many modern biomedical applications (computer aided diagnosis, medical image categorization, among others) and in the following analysis. Generally, it is a modeling process that encompasses a large range of algorithms that allow to formulate a relationship - with various degrees of complexity - between data and a specific outcome. Each algorithm has its own specificities, such as the assumptions it makes, the objective, the type of results or interpretability. When the outcomes for some data are known, and the task is to observe both the inputs x_i and outputs y_i of the system under study to formulate data-driven hypothesis or further categorize unseen data, the modeling process is called *supervised learning*. In epilepsy research as for most other applications in different fields, supervised learning represents the vast majority of machine learning applications.

The typical supervised machine learning framework consists in assembling a series of observations (x_i, y_i) and feeding them into a learning algorithm to identify the mapping $y_i = f(x_i)$ (Figure 5.1). Given a sufficiently large amount of observations or realizations, such approach allows to build new knowledge on a given topic of interest. However, the more complex the underlying phenomenon f is, the larger the amount of observations needs to be. Due to the multiscale interactions of complex molecular and cellular-level processes, understanding the brain system is one of the greatest challenges in contempo-

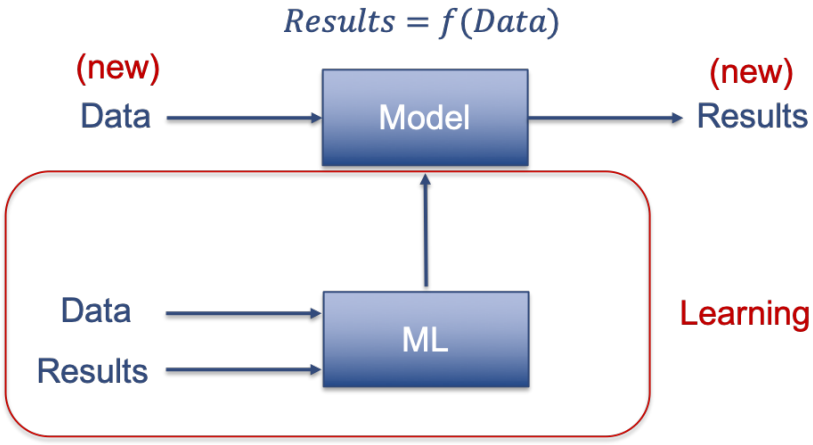

Fig. 5.1 **Typical Modeling Process of Supervised Learning Approaches.** The objective is to find a function f so that it maps the data x_i to the results y_i. The model parameters are trained using past experience, that is, using data for which the results are known. This is the Machine Learning step (ML). The hope is that the model will be close enough from the real system to be useful for all sets of inputs to be encountered in practice. In such case, the model accurately captured the mapping $y_i = f(x_i)$. Such strategy implies that such mapping, or pattern, exists and that sufficient data are available for its identification.

rary science [199]. Furthermore, electrophysical signals such as those generated by LFP typically monitor the activity of hundreds or thousands of neurons at a time, therefore blending relevant information together. A particularly successful strategy to analyze those signals consists in reducing the high dimensionality of the original signal into a lower dimensional space that represents its most salient characteristics or an abstraction of its behavior. Those characteristics, or *features*, are signal transformations that aim at guiding the algorithm towards the most important information underlying the data. Hence, less relevant information are filtered out and the accuracy of the model is increased. In practice, however, the relevant information are often hidden or untrivial so that they are mostly unknown. Domain-specific knowledge (e.g. clinical expertise), therefore, is a considerable asset to boost the performances of machine learning algorithms. Alternatively, designing features such that they improve the performances ot the algorithm means that novel and important aspects of the data were previously overlooked and are now captured. Features engineering, therefore, is a task that serves both the purposes of enhancing the performance of the model by approaching the phenomenon f more closely and of generating novel knowledge on the properties of the system. As such, it is a challenging task that requires meticulous analysis of the data and literature review.

The neurophysiological data in this thesis are time-series data monitoring of brain activity in Zebrafish larvae. The machine learning algorithms that will be used are here

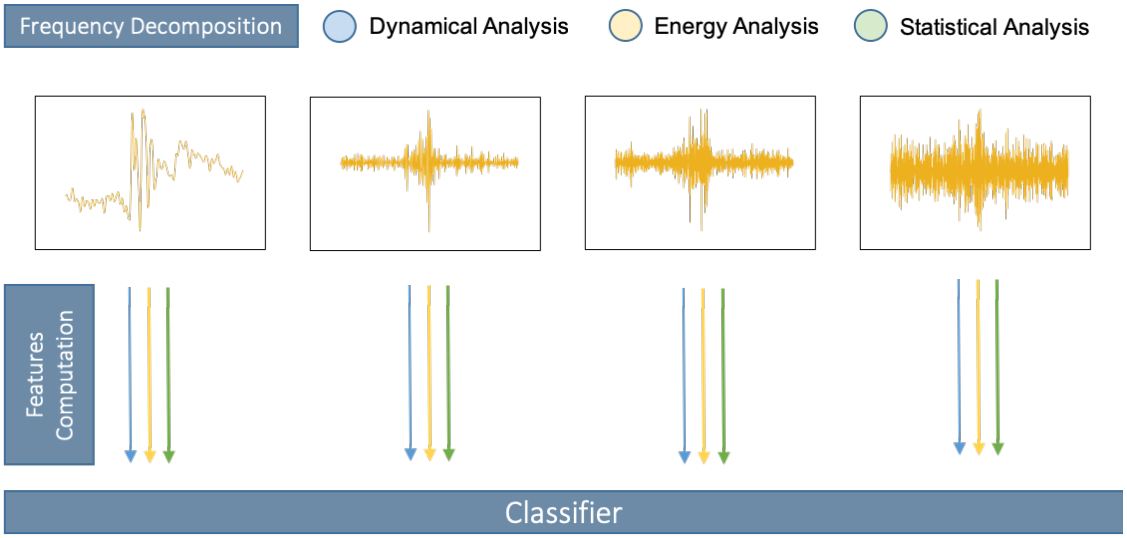

Fig. 5.2 **Signal Decomposition, Features Computation and Signal Classification.** In the case of LFP signal analysis, the signal is first decomposed into different relevant frequency bands (represented by the 4 frequency bands in yellow). Then, features that are believed to describe the most salient characteristics of the signal are then computed on each of the resulting decomposed signals. In particular, features that describe the dynamical properties of the signal, its (relative) energy and statistical properties are investigated. Then, the overall feature vector is provided to a prediction model (classifier) that learns to discriminate between the most important aspects of the signal.

introduced in more details, together with the mathematical technicalities required to transform those signals into features x_i. In order to create meaningful knowledge on the underlying pathological biological process of epileptogenesis, interpretable models and features are preferentially used, as opposed to black-box strategies which don't offer straightforward or human-interpretable decision rules. Furthermore, an important premise of our approach is that the signal consists in a superposition of different functional mechanisms, e.g. a number of oscillating frequency components, operating at different time or spatial scales. Notably, such composition of the signal has been taken benefit of to improve the performances of numerous automatic seizures detection algorithms in humans [48]. Therefore, further decomposition of the signal into several frequency sub-bands (either arbitrarily large or into physiologically relevant ones) was considered to capture and enhance the neuronal activity that is not directly obvious from the full-spectrum recording [200, 201, 202, 203]. As such, features characterizing the neurophysiological signals are here computed from both the original signal and the frequency sub-bands. Figure 5.2 illustrates the approach undertook.

For this purpose, the concepts underlying a time-frequency transformation of the signal called wavelet transform are first introduced. The wavelet transform consists in the analysis of the frequency components of the signal, but account for transients signals so that it can be applied to non-stationary data. Specifically, it has the potential to locate and quantify brief and specific neurophysiological abnormalities from LFP signals. Second, the features derived from those signals are detailed. The features used in this thesis aim at capturing the main dynamical properties of the signal, that is, changes in brain mechanisms over time that would indicate the propensity and progression of the condition and therefore potentially serve as biomarkers of epileptogenesis. Many features are considered and believed to capture important aspects of the neurophysiological signals. Indeed, it is unlikely that a single feature, or biomarker, would suffice to reliably indicate the presence of a particular epileptogenic process [44]. In particular, features originating from dynamical systems theory, statistics and nonlinear time series analysis are emphasized. Finally, the statistical model used to discriminate brain signals using such features is introduced. The random forest algorithm was further considered for its flexibility, efficiency, interpretability and robustness to noise.

5.2 Wavelet Transform

The Fourier transform is a well known signal transformation that decomposes the signal into its constituent frequencies. However, one assumption of the Fourier transform is that the data are stationary, that is, the statistics of the data are constant over time. This is clearly not the case for signals describing neurophysiological activity, because endogenous and spontaneous processes occur. In this sense, the analysis of such signal requires tools that account for the intrinsically transient and spontaneous dynamical nature of the signal (Figure 5.3).

The wavelet transform can be thought of as an extension of the classic Fourier transform, except that, instead of working in the frequency domain only, it provides a time-frequency localization describing the dynamical patterns of the frequency structure in a time-series. In the case of Fourier transform, we might be able to determine all the frequencies present in the signal, but not when they are present. Alternatively, the wavelet transform uses a convolution kernel that has a compact support, that is, it vanishes outside a certain window interval. The latter is called the wavelet, which essentially consists in small oscillations that are highly localized in time [204]. Because a convolution in the time domain is equivalent to a multiplication in the frequency domain, the frequency-band intersection between the data and the wavelet is therefore localized in both time and

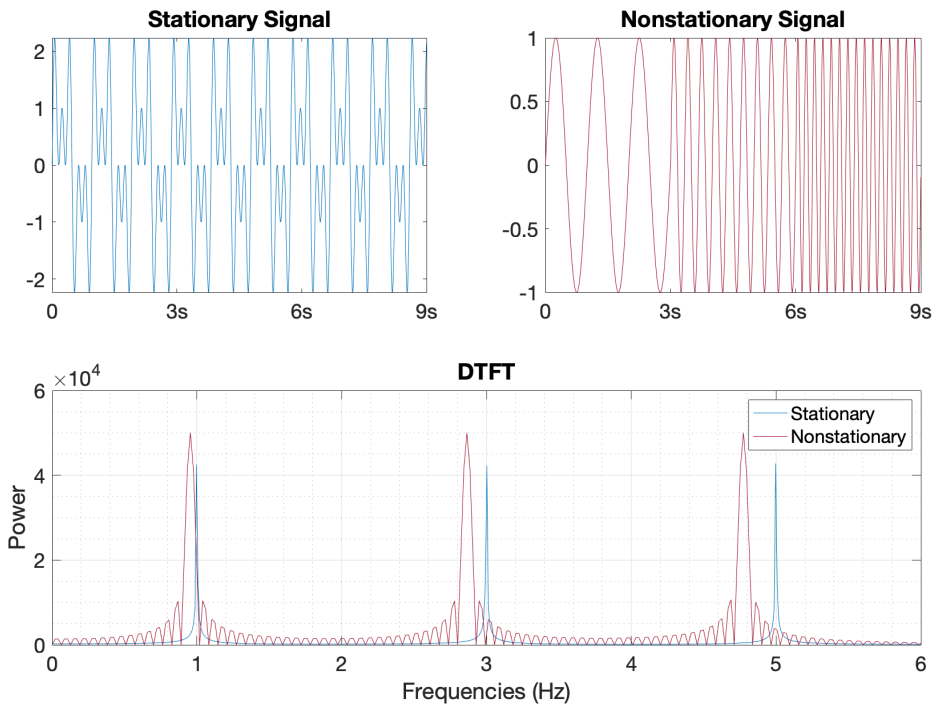

Fig. 5.3 **Example of Stationary VS Nonstationary signals.** Upper left: Stationary. Upper right: Nonstationary. The main frequencies of both signals are the same but they appear sequentially on the nonstationary signal. The variance, therefore, is not constant. Below is the power spectra of the signal. Not only the Discrete Time Fourier Transform (DTFT) does not provide information on the sequential changes in frequencies occuring in the data, but it also biases the resulting frequency decomposition due to sudden dynamical changes.

frequency. This property makes wavelets well-suited for the analysis of data with sharp or transient discontinuities such as those occuring prior or during epileptic events.

Wavelets are flexible tools. Indeed, the wider the bandwidth of the kernel is, the less temporally localized the information is but the more precise the information on the frequency is. This crucial parameter, referred to as the number of wavelet cycles, defines the trade-off between temporal and frequency precision. Hence, unlike sine waves, wavelets do not contain energy in a single frequency but rather, in a range of frequencies characterized by the wavelet pattern. As a consequence, the resulting frequency information at each time point is a weighted sum of the frequency information of surrounding time points, with the weight decreasing with increasing distance away from the center of the wavelet. The mathematical formulations of both the Continuous

Wavelet Transform (CWT) and the Discrete Wavelet Transform (DWT) are hereafter detailed.

5.2.1 Continuous Wavelet Decomposition

Formally, the continuous wavelet decomposition (CWT) is written as [204]:

$$\gamma(s,\tau) = \int f(t)\Psi^*_{s,\tau}(t)dt \tag{5.1}$$

where $*$ denotes complex conjugation. This equation shows how a function $f(t)$ is decomposed into a set of basis functions $\Psi_{s,\tau}$ called the wavelets. The variables s and τ are the new dimensions, scale and translation, after the wavelet transform. The wavelets are generated from a single basic wavelet $\Psi(t)$, the so-called mother wavelet, by scaling and translation:

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Psi(\frac{t-\tau}{s}) \tag{5.2}$$

The wavelet transform, therefore, requires to specify the mother wavelet from which the basis functions will be constructed. There exist many wavelet shapes for which the suitability varies given certain kind of signals. However, not any function can be used as a mother wavelet. The mother wavelet should be smooth, oscillatory and carry finite energy. The choice of a particular optimal shape of the mother wavelet for signal decomposition requires extensive investigation which was beyond the scope of this study. We used Daubechies kernel functions that have good localizing properties both in temporal and frequency domains. More specifically, they have shown promising results regarding the analysis of LFP signals [200, 201, 202, 203].

For completeness, the inverse wavelet transform is written as:

$$f(t) = \int\int \gamma(s,\tau)\Psi_{s,\tau}(t)d\tau ds \tag{5.3}$$

An example of such time-frequency analysis is illustrated on Figure 5.4, where the non-stationary signal of Figure 5.3 is analyzed with a continuous wavelet transform. The effect of the number of cycles of the wavelet, that determines its bandwidth, is

also illustrated. Importantly, as the wavelet transform is calculated by convolving the signal with wavelets for continuous values of both scaling and translation, the obtained decomposition is highly redundant and computationally heavy. To overcome those problems and make the wavelet decomposition usable in practice, the discrete wavelet transform (DWT) was used.

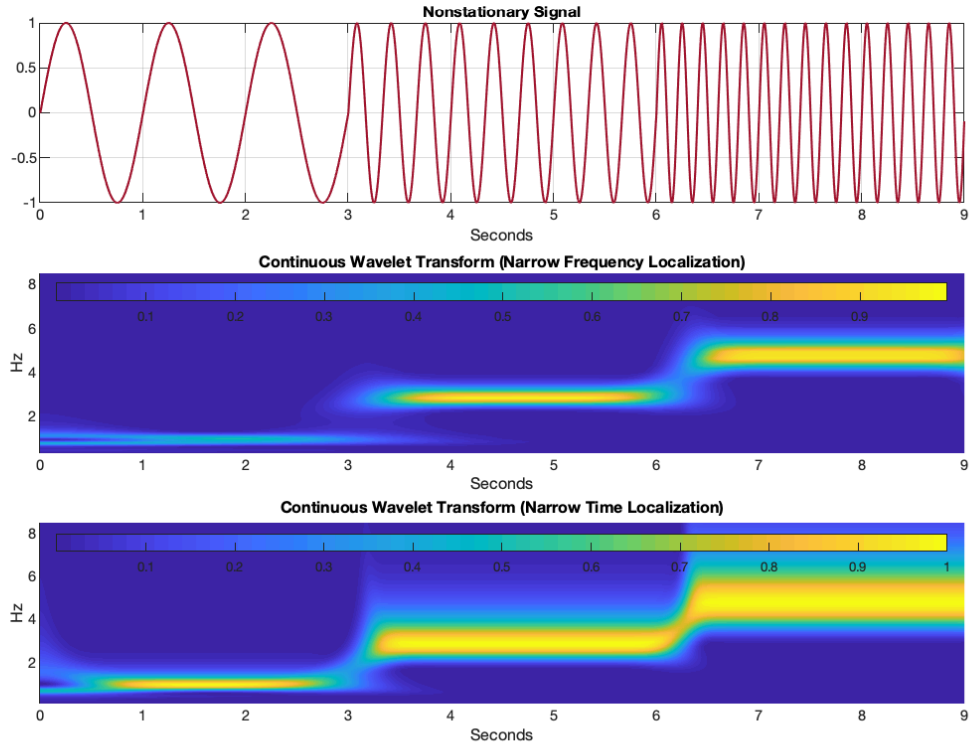

Fig. 5.4 **The CWT maps a one-dimensional signal into a two-dimensional time-scale joint representation that displays the power spectra of the signal over time.** The first panel shows the original signal while the two following ones display the CWT decomposition of the signal. The middle panel displays wavelets for which the bandwidth is larger, so that the transformation displays a narrow frequency localization but less precise in time. Conversely, a smaller bandwidth shows a more accurate localization in time than in frequencies. Changes points in the underlying dynamics of the system are visible from both panels, but the narrow bandwidth is more suited for detecting transient activations whereas the larger bandwidth is more sensitive to long activations at specific frequencies. As a trade-off to identify transients neurophysiological events, the Daubechies 4 (db4) wavelet has been selected in this thesis.

5.2.2 Discrete Wavelet Decomposition

Discrete wavelet transform discretizes the signal decomposition into an orthonormal basis, which has many benefits in signal analysis. As an orthonormal signal decomposition, the

DWT has been widely used in engineering or mathematics with applications ranging from signal processing or denoising to data compression, since noise is therefore uncorrelated at the input and output. This is achieved by modifying the wavelet representation so that [205]:

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}}\Psi(\frac{t - k\tau_0 s_0^j}{s_0^j}) \tag{5.4}$$

with j and k integers and $s_0 > 1$ is a dilation step. The translation factor τ_0 depends on the dilation step. The time-scale space is now sampled at discrete intervals. Usually, $s_0 = 2$, so that the decomposition is called dyadic and $\tau_0 = 1$ so that the sampling is also dyadic on the time axis. The schematic dyadic DWT decomposition is illustrated on Figure 5.5. The DWT decomposes a given signal by passing it through a series of related low-pass $h[n]$ and high-pass filters $g[n]$, for which the coefficients correspond exactly to those of the wavelet coefficients for a discrete set of child wavelets originating from the mother wavelet $\Psi(t)$. The range of the bandwidth, therefore, decreases exponentially.

The signal $X[n]$ is therefore approximated by increasingly finer details such that:

$$X[n] = \sum_{k=-\infty}^{\infty} 2^{l/2} a_{l_0}[k] h[2^{l_0} n - k] + \sum_{l=l_0}^{\infty} \sum_{k=-\infty}^{\infty} 2^{l/2} d_l[k] g[2^{l_0} n - k] \tag{5.5}$$

where l represents the scale index and the coefficients (respectively called approximation a_j and details d_j) in the above expansion are calculated by:

$$\begin{aligned} a_j[k] &= \sum_{k=-\infty}^{\infty} 2^{l/2} x[n] h[2^l n - k] \\ d_j[k] &= \sum_{k=-\infty}^{\infty} 2^{l/2} x[n] g[2^l n - k] \end{aligned} \tag{5.6}$$

The decrease by a factor of two of the wavelet coefficients at each step may, however, represent a limiting factor to carry out statistical analysis. Furthermore, the coefficients are not aligned with the events in the time-series. In the following paragraph, the nondecimated wavelet transform and the multi-resolution analysis are introduced.

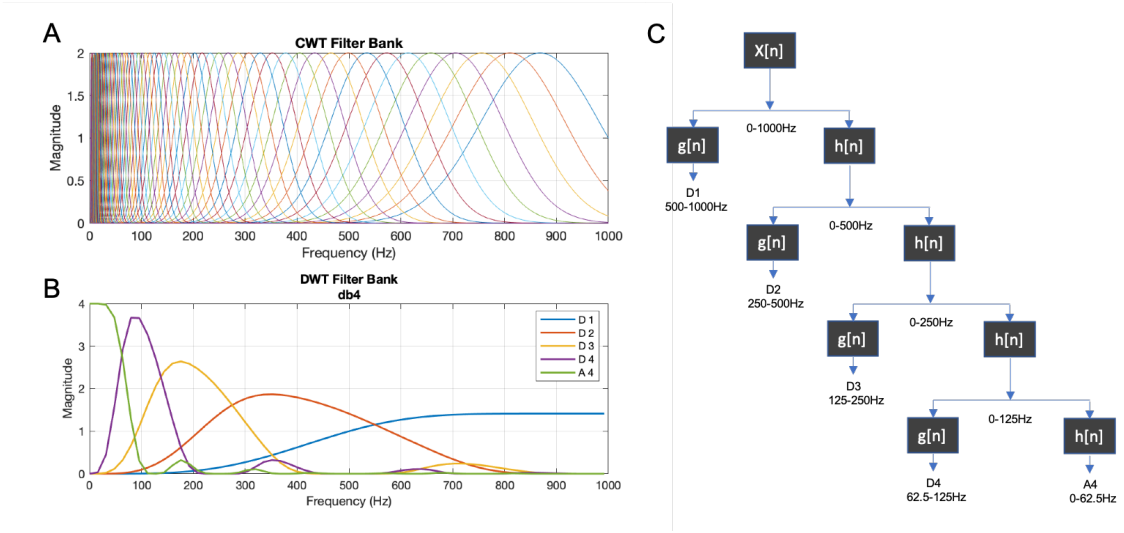

Fig. 5.5 **Dyadic Wavelet Decomposition.** **A** Filter bank of the Continuous Wavelet Transform (CWT). **B** Filter bank of the Discrete Wavelet Transform (DWT) using the Daubechie 4 wavelet (db4). The CWT, as opposed to the DWT which is an orthonormal transform, is a highly redundant signal decomposition. **C** To decompose the signal following the DWT, the signal is sequentially passed through a series of high pass g and low pass h filters. At each step, the signal is downsampled by two, since the output signals only hold half of the original frequency bandwidth. The outputs of the high pass and low pass filters are called *details* and *approximations*, respectively.

5.2.3 Nondecimated Wavelet Transform and Multi-resolution Analysis

The nondecimated wavelet transform, or maximal overlap WT (MODWT), is called overdetermined, that is, it is not an orthonormal transform anymore, and therefore introduces some redundancy in the decomposition. Similarly to the DWT, the MODWT is defined in terms of a pyramidal algorithm. While computationally heavier, the MODWT has several advantages over the DWT [206]:

- The details and approximations coefficients are not downsampled by power of two anymore, therefore enabling a more statistically relevant analysis of the decomposed signal.
- The MODWT is a zero-phase filter, so that the events in the original signal are exactly lined up with those of the decomposed signal.
- The coefficients of the MODWT are norm-preserving, that is, the sum of the energy of the wavelet coefficients is equal to the energy (L2 norm squared) of the original signal.

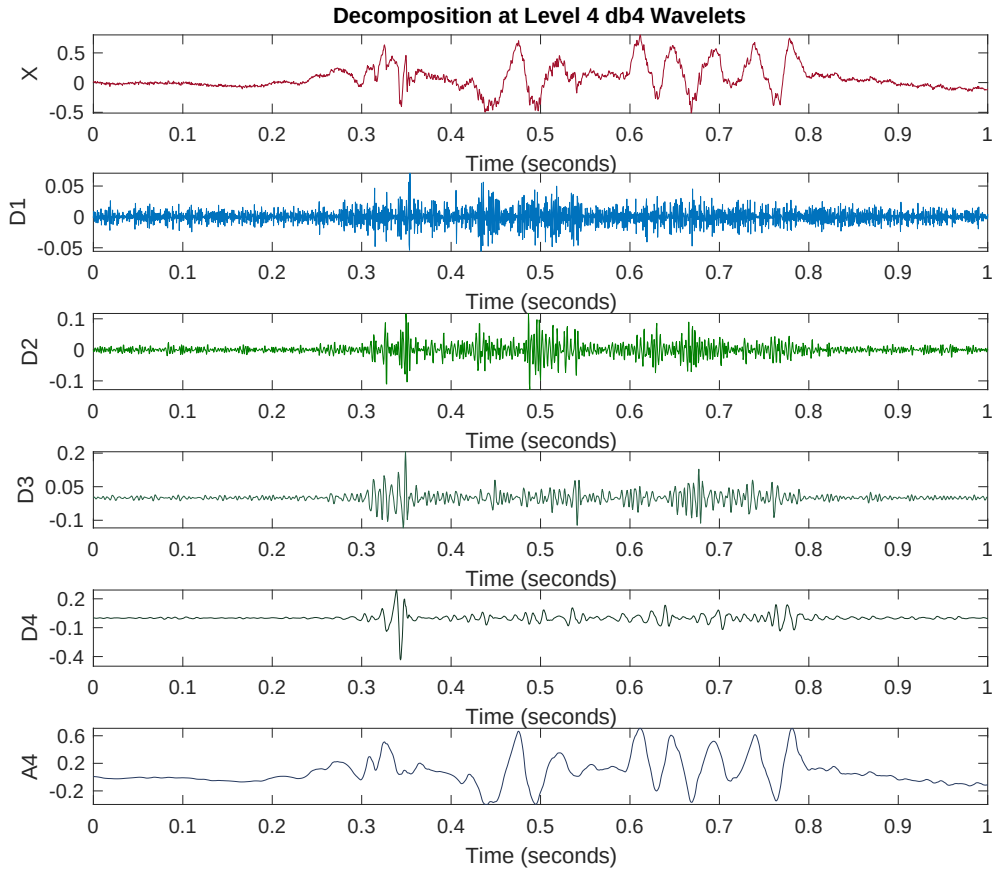

Fig. 5.6 **Maximal Overlap Discrete Wavelet Transform (MODWT) of a LFP Signal.** The signal is transformed into its details and approximation coefficients via a dyadic decomposition. As compared to the DWT, the amount of coefficients is not downsampled by a factor 2 at each step. Importantly, the events are lined up across frequency bands and the original LFP signal.

- It is shift invariant, so that the MODWT does not depend on the starting point of the analysis. Therefore, it can be used to perform a multi-resolution analysis (MRA). The MRA is not norm-preserving, but the sum of the components of the MRA, element by element, form the original time-series.

To achieve this, the filters h and g are renormalized so that:

$$\bar{h}_l \equiv \frac{h_l}{\sqrt{2}} \quad ; \quad \bar{g}_l \equiv \frac{g_l}{\sqrt{2}} \tag{5.7}$$

That is, the filters have the same widths but do not result in a downsampling of the signal by 2^l at each step anymore. Figure 5.6 displays an example of the decomposition

of a LFP signal with the MODWT into 4 levels, using the db4 wavelets.

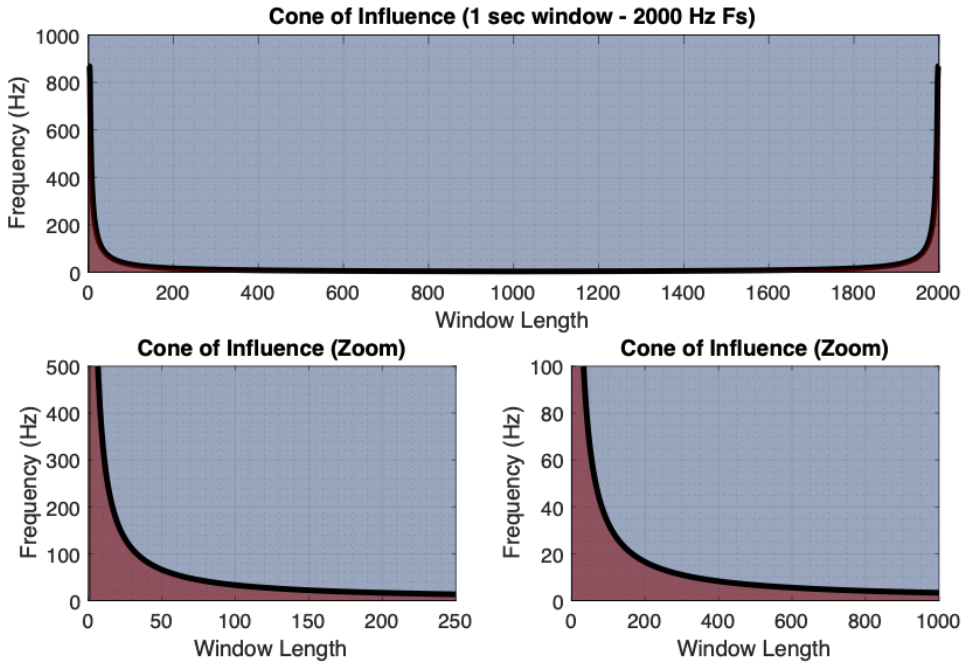

Fig. 5.7 **Cone of Influence (Boundary Effects).** Although there are several ways to reduce signal distortions at the edges, the signal cannot be decomposed into arbitrarily smaller scales (low frequencies) without loss of accuracy. The cone of influence shows areas potentially affected by edge-effects artefacts. Those effects arise from regions where the signal convolution with the wavelets extend beyond the edge of the observable signal. In red, the regions where boundary effects occur. In particular for those analysis that divide the signals into rolling windows, boundary effects will apply. There is no strict mathematical rule to determine the cone of influence. Because the bandwidth of wavelets decay exponentially in time, the borders of the cone are approximated with a time constant $1/e$, so that one time-domain standard deviation of the wavelet is removed at both ends.

As a final note, it is important to emphasize that the signal cannot be decomposed into infinitely smaller scales, as boundary effects occur. Boundary effects are typical signal distortions that appear at the edges of the resulting decomposed signal as a consequence of the convolution with the (wavelet) kernel on a finite-length window. Several strategies can be applied to reduce such effects, such as zero-padding or symmetrization of the data. In this thesis, the symmetrization strategy is used so that the external values of the convolution are recovered by symmetrically replicating those. The range of the boundary effect phenomenon is depicted on Figure 5.7 for a signal window of 1 second, sampled at 2000 Hz.

5.3 Features Engineering

Feature engineering plays a key role in machine learning, statistical pattern recognition and data mining. The capability of the algorithms to learn the hidden structure within the data heavily relies on the balance between data availability, data and system complexity, data representation and the outputs of the model to be constructed, among others. As a very complex, highly dimensional nonlinear system, the investigation of brain dynamics typically requires a considerable amount of data to thoroughly learn its key underlying principles. In addition, biomedical systems are inherently stochastic and their monitoring may suffer from low signal to noise ratio, intra-inter variance between patients and a generally scarce availability. Navigating through the space of possible brain mechanisms to formulate a model, therefore, is a challenging process even for the most sophisticated machine learning tools.

Data representation concerns the meaningful reduction of the high dimensionality, or data transformation, of the original data into features. Features engineering consists in representing the most salient characteristics of the signal or an abstraction of their properties, such as quantifying trends or whether a signal contains a recurrent pattern, respectively. As a consequence, the machine learning algorithm does not see the raw data anymore, but the features only, which are in turn mapped to a certain desired output. Hence, designing appropriate features is a crucial step which requires a comprehensive analysis of the signal and of the mathematical, physics, and biological literature. Doing so confers the machine learning algorithm more robustness to noisy or ambiguous information, but also further guides it to likely better results.

The choice of particular features is here inspired by the concepts of dynamical bifurcations introduced in Chapter 1. The key idea is to investigate distinct dynamical aspects of the signal, with the aim of correlating those with specific neuronal and epileptogenesis mechanisms. In particular, feature computation is performed on both the original and decomposed signals. Indeed, it has been shown that signs of signal complexity and organization were depending on the clinical frequency bands investigated, so that a wide range of scales has to be considered for a comprehensive analysis [200]. In other words, scale mixing blurs the effectiveness of features for the design of prediction algorithms [61]. Hence, the algorithm will be provided with more information and left to choose the most relevant scales upon which to rely for the identification of informative predictors.

To date, empirical studies have revealed few general biomarkers of electrical activity, or precipitating factors, that appear near a dynamical bifurcation towards seizure: interictal

epileptiform discharges (IEDs), high frequency oscillations and abnormal changes in the electrical background activity [11]. The shape, occurrence and duration of those events, however, appear to be highly variable across patients, and do not inevitably progress to a seizure, so that no universal biomarker has yet been identified [44, 73]. The features used in this thesis cover the aforementioned patterns of brain activity, but are not restricted to them.

5.3.1 Features Description

The list of transformations of the original neurophysiological signal into features is detailed hereafter. Overall, signal transformations can be separated into 4 classes that contract the original signal into distinct dynamical aspects:

- System memory: Hurst Exponent (HE), Fractal Dimension (FD).
- Signal complexity: Sample Entropy (SampEn), Lyapunov Exponent (LE).
- Data distribution: statistical moments (variance, skewness, kurtosis).
- Energy distribution: Relative Wavelet Energy (RWE).

Beside detecting seizures events, the goal is to reflect the underlying changes in the stability of the brain system and its susceptibility to seizure occurrence. Generally, LFP signals, which reflects the collective activity of a very number of dynamically coupled neurons, appears to gain in predictability, that is, a general loss in complexity towards seizure occurrence [11]. Such complexity loss has been reported to vary according to the scale at which they were investigated (i.e. on different frequency bands) [61]. At this stage, it is important to highlight that the metrics characterizing system memory (HE, FD) and the LE are not computed from the decomposed signal data (from DWT), as the resulting information does not have a clear mathematical nor biological support.

Various complexity measures have been developed over the years to distinguish random signals (e.g. the brain at rest) to more organized ones (e.g. the brain during seizures events). On one hand, the entropy exactly measures the randomness of the information in a signal (and can be applied to nonlinear stochastic systems as well) [207]. On the other hand, the Hurst Exponent, the Fractal Dimension and the Lyapunov Exponent can also be thought as signal complexity measures [208, 209], albeit measuring the dynamical predictability of the system by conceptually different approaches that are detailed hereafter. For example, the FD is a local property of roughness while the HE reflects the long-memory dependence of the system [210]. Overall, empirical

studies showed that complexity measures such as those proposed here eventually behave qualitatively consistently in the context of large deviations of dynamics in physiological signals (e.g. seizures) [61].

Lyapunov Exponent

The early days of the application of complex computational tools inspired by dynamical systems theory (and more precisely chaos theory) to the analysis of neurophysiological data have led to the investigation of fluctuations in the Lyapunov Exponent in the vicinity of a critical bifurcation. While more recent research suggested that such metric empirically estimated from noisy time series suffers from substantial limitations to identify an approaching transition, its value for discriminating significantly different systems states (i.e. between interictal and ictal activity) has been widely supported [211, 212, 213].

The Lyapunov Exponent (LE) is a dynamical invariant of nonlinear systems that measures the exponential growth of an infinitesimal line segment in the phase-space. In other words, the LE describes the evolution of a dynamical system from its trajectory in a embedded space. The exponential growth of its trajectory echoes the sensibility of the system to small perturbations, and hence, its predictability. A chaotic system holds at least one positive LE. Typically, this measure can be calculated exactly given the multi-dimensional nonlinear equations that describe the system. However, estimating it from the observable one-dimensional behavior of a system is a non-trivial task. [214] demonstrated that the trajectories of this kind of system can be recovered from a state-space embedding given a sufficiently high amount of data and relatively low signal to noise ratio. Given two neighboring points in the state-space at time 0 and a time t, the distance between them is a function of time $\Delta x(x_0,t)$, the largest LE is formally defined as:

$$\lambda = \lim_{t\to\infty}(\frac{1}{t}\times \ln(\frac{|\Delta x(x_0,t)|}{|\Delta x|})) \tag{5.8}$$

with

$$\Delta x(x_0,t) \sim \Delta x e^{\lambda t}, \Delta x \to 0 \tag{5.9}$$

In this thesis, the algorithm proposed by [215] to estimate the largest LE from time-series data has been used. The algorithm consists in two steps. First, the user needs to

define an embedding space for the time series data, and then the LE is estimated by the aforementioned equation. The embedding space is defined such that:

$$X(i) = (x(t), x(t+\tau), x(t+2*\tau), ..., x(t+(m-1)*\tau)) \tag{5.10}$$

which requires prior estimation of two parameters: the embedding dimension m and the delay τ. Such embedding of time series creates a trajectory in the m dimensional space. As an example, periodic time series would become closed phase space orbits. In practice, the delay τ is chosen so that it corresponds to the smallest value that maximizes the independence of the coordinates of the embedded vector [45].

For this purpose, a metric (here, the mutual information) is computed between the signal and a delayed τ version of itself. The delay τ that corresponds to this criterion is then the first minima of the average mutual information. The mutual information $I(X;Y)$ quantifies the amount of information random variables have in common. It follows:

$$I(X;Y) = \sum_i \sum_j P(X_i, Y_j) \log \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)} \tag{5.11}$$

which can be computed empirically from time-series using histograms. The next step is to estimate the embedding dimension, which is obtained using Cao's method [216]. As for τ, bigger is not necessarily better, as a single noisy point in the time series will affect m points. Therefore, the aim is to obtain the smallest m that conserves a topologically correct result. The strategy followed here is a type of false near neighbor (FNN) algorithm. For each dimension m, the distance between each close neighbor of each point is computed. The dynamics is considered as being correctly unfolded in the selected embedding dimension when those distances do not change significantly anymore. Note that estimating both parameters m and τ simultaneously can also be a good strategy but was not chosen here for simplicity. Mathematically:

$$E(d) = \frac{1}{N-\tau d} \sum_{i=1}^{N-\tau d} |x_{i+\tau d} - x_{n(i,d)+\tau d}| \tag{5.12}$$

where $i = 1, 2, ..N - \tau d$ and $x_{i+\tau d}$ is the nearest neighbor of $x_{n(i,d)+\tau d}$ in the d-dimensional space. Then, the embedding dimension is modified to obtain a measure of the distance $E(d)$ as a function of d so that the ratio $E1(d)$ is defined:

$$E1(d) = \frac{E(d+1)}{E(d)} \tag{5.13}$$

In practice, seizure detection is relatively tolerant towards the exact choice of those parameters [215], but it might impair the detection of more subtle dynamical changes such as those related to epileptogenesis.

The state-space embedding and the estimation of the largest LE from time series data originating from real (physiological) system suffer from several issues and limitations that are briefly discussed below.

- **Data length.** Although looser bounds can be defined, the original embedding theorem requires an infinite amount of data to unfold the dynamics of the true underlying dynamical system [213]. Hence, obtaining robust estimations of the LE might reveal itself challenging. As a rule of thumb, it is important to limit the use of high embedding dimension m with relatively small amount of data (e.g. estimating the manifold of the attractor in a six-dimensional embedding space using 100 datapoints should not be regarded as a robust representation of the system).
- **Very high-dimensional systems.** In practice, the nature of biological systems is often not only high-dimensional but also spans across (unobserved) multiple scales. In such case, it is not clear whether the actual attractor of the system can be reconstructed from spatio-temporal dynamics contracted into univariate times series data.
- **Noise.** Typically, noise in the time series will affect m of the points in the m-dimensional embedding space, which can substantially hinder the estimation of the dynamical trajectories.

Signal Entropy

The entropy is a quantity that measures the signal disorder, that is, signal complexity. A more complex signal is typically less predictable. The entropy, as opposed to the LE, can be applied to both deterministic chaotic and stochastic systems, and are suited for noisy and finite-length experimental data [217]. In particular, entropy measures receive a considerable interest for quantifying the complexity of biomedical time-series data, such as for the diagnosis of diverse brain functional or pathological states (e.g. Alzheimer) [218, 219], or heart rate variability [220, 221].

A robust measure of the entropy from physiological signals called the Sample Entropy (SampEn) has been used in this thesis [60]. The SampEn is capable of detecting changes in the signal, which are not reflected in peak occurrences or amplitudes [222]. Large values of the SampEn mean that the underlying system generating the time series data is more complex, and its future behavior less predictable. SampEn is a function of 3 parameters m, n, R, given by the following formula:

$$SampEn(m,r,N) = -ln(\frac{C_{m+1}(r)}{C_m(r)}) \tag{5.14}$$

where m corresponds to the embedding dimension, r the tolerance parameter used in the Heaviside function and N the amount of data points. C_m is the correlation integral defined as:

$$C_m(r) = \frac{\{\text{number of all pairs (i,j) with } |x_i^m - x_j^m| < r, i \neq j\}}{\{\text{number of all pairs, i.e. } (N-m+1)(N-m)\}} \tag{5.15}$$

where $|x_i^m - x_j^m|$ represents the distance between points x_i^m and x_j^m. x_i^m and x_j^m correspond to all possible pairs of points in the embedded vector. Hence, the SampEn is a measure of how close two consecutive data points remain similar in the next point $(m+1)$. It has been shown that SampEn has a better statistical validity for m = 1 or 2 and the range of the tolerance parameter r around $0.1 \times \sigma$ to $0.25 \times \sigma$ [60]. Parameters values were here chosen so that m = 2 and $r = 0.2 \times \sigma$.

Fractal Dimension

The fractal dimension characterizes the local self-similarity in the signal. It can be also thought as a quantification of the roughness or correlation structure of a time-series as the scale becomes infinitesimally fine. Interestingly, fractal behavior has been observed in a large number of physical, biological and financial systems [223, 224, 225, 226]. Conveniently, the FD is particularly appropriate to capture transient events in the data and does not require the reconstruction of the attractor in a multidimensional state-space (such as the SampEn or LE). As such, it is a very computationally efficient measure.

There exist many methods to estimate the fractal dimension from time series data, but they all share the following scheme [227]:

- A certain numerical property Q of the time-series is estimated as a function of scale ε.
- The scale ε is made infinitesimally small ($\varepsilon \to 0$) and a power law is derived ($Q(\varepsilon) \propto \varepsilon^{\beta}$) such that:
  - The scaling exponent β is a linear function of the fractal dimension D.
  - D is estimated through linear regression of $\log Q(\varepsilon)$ on $\log \varepsilon$.

Here, the fractal dimension has been estimated through the computation of an extensive measure: the empirical variogram. Variograms can be interpreted as a statistically more efficient and robust estimator of fractal dimension than the classical box-counting algorithms [227]. The robustness of the estimator is of particular importance for the analysis of neurophysiological signals, since the aim is to differentiate between subtle brain functions from short recordings hindered by a relatively high signal to noise ratio.

Originally developed to describe the degree of spatio-temporal dependence of spatial random field, the variogram can be naturally generalized to time series data. The variogram γ of a stochastic process X_t is defined as :

$$\gamma(t) = \frac{1}{2}\mathbb{E}(X_u - X_{u+t})^2 \tag{5.16}$$

which corresponds to one-half times the expectation of the square of an increment at lag t. Hence, for a stationary Gaussian process, the variogram satisfies $\gamma(t) = |ct|^{\alpha}$, with α (0,2[being the fractal index, which therefore relates the fractal dimension to the variogram in such that:

$$D = d + 1 - \frac{\alpha}{2} \tag{5.17}$$

where d corresponds to the topological dimension (1 for time series, ≥ 2 for surfaces). Interestingly, this formulation allows to relate the properties of a Gaussian stochastic process to the measure of the fractal dimension, which then directly relates to the covariance function $\sigma(t) = cov(X_u, X_{t+u})$, so that:

$$\gamma(t) = \sigma(0) - \sigma(t) \tag{5.18}$$

Estimating the moments of γ from time series data can be typically formulated as:

$$\hat{V}(l/n) = \frac{1}{2(n-l)} \sum_{i=l}^{n} |X_{i/n} - X_{(i-l)/n}|^p \tag{5.19}$$

where p corresponds to a power index here chosen to be equal to 1 (so-called madogram) and $|.|$ denotes the norm. The fractal dimension D is obtained via regression fit of $\log \hat{V}(t)$ on $\log t$, which therefore yields the following robust variogram estimator, via equation (6.17) [227]:

$$\hat{D} = 2 - \frac{1}{p} \frac{log\hat{V}_p(2/n) - log\hat{V}_p(1/n)}{log2} \tag{5.20}$$

For time series data, the fractal dimension is bounded between 1 and 2. Hence, 1.5 corresponds to serially uncorrelated processes. The rougher the time series are (i.e. it fills more space), the larger the fractal dimension is. On the opposite, a smooth curve is associated with a low value of FD.

Hurst Exponent

The Hurst Exponent (HE) is an estimate of the long-term memory dependence of time series data. Conceptually, it can be used to estimate the presence of long-term feedback processes in physiological signals. The HE and FD are two notions that are closely linked to each other, but are only equal for strictly self-similar processes, where the local properties of the system are reflected in the global ones [226]. This is obviously not the case for nonstationary processes such as those observables from neurophysiological signals, that is, LFP or EEG.

The HE is bounded between 0 and 1. Low numbers of HE indicate a mean-reversive process, while values close to 1 represent a trend persistence in the signal. A random process is characterized by a HE of 0.5. The HE is here computed via the Detrended Fluctuation Analysis (DFA) algorithm [228], which works in the following way, given a time series X_t:

- X_t is detrended so that (with $\langle x \rangle$ representing the mean of the signal):

$$X_t = \sum_{i=1}^{t} (x_i - \langle x \rangle) \tag{5.21}$$

- A straight-line fit Y_t is estimated by minimizing the squared error and the Root-Mean-Squared (RMS) deviation from the trend. The fluctuation $F(n)$ is obtained:

$$F(n) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(X_t - Y_t)^2} \tag{5.22}$$

- Steps 1 and 2 are repeated for a range of different window sizes, obtained by dividing the original signal X_t into arbitrarily smaller windows. The values of $F(n)$ and reported against n on a log-log graph.

- The scaling coefficient corresponding to the Hurst Exponent is calculated by estimating the slope of the resulting curve on the log-log graph.

Statistical Moments

During recordings of resting brain activity, the signal is well described by simple linear statistics. Indeed, despite originating from a large collection of coupled neurons, their average activity converges to a Gaussian probability distribution, even if the individual processes are non-Gaussian, because neurons spiking are largely uncorrelated. However, a variety of empirical studies showed that often those distribution, while conforming with the assumption of simple probability distribution, are often heavy-tailed as a result of the erratic modulation or synchronization of the firing rates of ensemble of neurons [229, 230, 231, 232]. Hence, higher statistical moments are here considered. The skewness is a measure of the asymmetry of a distribution:

$$\gamma_1 \equiv \frac{\langle (x - \langle x \rangle)^3 \rangle_{p(x)}}{\sigma^3} \tag{5.23}$$

where $\sigma^2 \equiv \langle (x - \langle x \rangle)^2 \rangle_{p(x)}$ is the variance of x with respect to $p(x)$. A positive skewness means the distribution has a heavy tail to the right, and inversely.

The kurtosis is a measure of how peaked around the mean a distribution is:

$$\gamma_2 \equiv \frac{\langle (x - \langle x \rangle)^4 \rangle_{p(x)}}{\sigma^4} \tag{5.24}$$

A distribution with a positive kurtosis has more mass around its mean than would a Gaussian with the same mean and variance, and inversely. The kurtosis is defined so that

a Gaussian has a kurtosis equals to 3. Typically, multiscale recruitment of neurons into large synchronous ensembles in the vicinity of a bifurcation characterized by a critical slowing down causes the variance to shrink and the kurtosis to increase [233, 234].

Computing those values from the wavelets coefficients of the decomposed signal accounts for the estimation of the conditional spectral moments around those specific frequencies originating from the dyadic decomposition. Hence, it is capable of capturing transient high frequency oscillatory phenomena.

Relative Wavelet Energy

The MODWT partitions the energy across the various scales and scaling coefficients:

$$||X||^2 = \sum_{l=1}^{l_0} ||d_l||^2 + ||a_{l_0}||^2 \tag{5.25}$$

Where X is the input data, d_j are the detail coefficients at scale l and the a_{l_0} are the final-level scaling coefficients. Therefore, we define the relative wavelet energy:

$$RWE_\delta = \frac{E_\delta}{E_{tot}} \tag{5.26}$$

Conceptually, computing the relative wavelet energy captures the spectral moments of the signal, which can be used to evaluate the frequency slowing down and hence, an indicator of the approaching transition to a different dynamic regime. Indeed, it has been demonstrated that energy switches occur between frequency bands in the event of an approaching seizure, or in the general case of critical dynamical transition [234].

5.4 Random Forest as a Statistical Model for Classification

Among machine learning methods, Random Forest (RF) notably benefits from several advantages such as their interpretability, as opposed to black-box classification algorithms such as Support Vector Machine (SVM) or Artificial Neural Networks (ANN), and their great generalization and noise robustness properties which take their origin in their inherent bagging scheme. Moreover, RF offers a flexible framework with only few hyperparameters, the possibility to tailor the objective function and to formulate its results

in a probabilistic fashion [235]. As such, RF have been applied to a broad spectrum of biomedical research, mostly formulated as classification tasks [236]. Nevertheless, they can also be applied to solve regression and multiple class clustering tasks, among others.

Random forests belong to a category of machine learning algorithms called *ensemble* techniques which combine the results of multiple models, thereby reducing the overfitting potential of the general model [237]. In this case, the RF algorithm constructs an ensemble of decorrelated decision trees and assembles their individual predictions, such as through the averaging of the predicted probabilities.

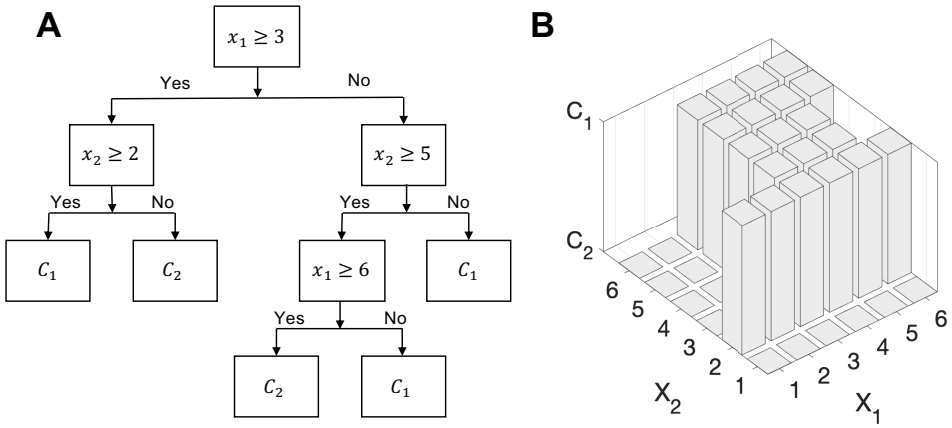

Fig. 5.8 **Decision Tree. A** A Hypothetical decision tree that classifies the data into two groups C_1 and C_2 given 2 discrete features x_1 and x_2. **B** Corresponding classification function / surface.

A decision tree is a classifier that consists in a sequence of binary decisions organized in a hierarchical fashion (Figure 5.8) [238]. Individual decision trees are intuitively appealing, since they are based on a recursive dichotomic partitioning of the data following an optimal separating decision rule at each node. Each decision tree of the RF only sees a subset of the data (with replacement, so that the size of the input data set remains identical), a process referred to as bagging. In the training phase, decision trees can be built using different decision rules, or split criterion, with a random subset of features for each tree. At each split, the data are divided into two groups so that the data are more similar within each group than across groups, picking the feature that produces the most separation. One of the most widely used split criterion is the Gini index [235], which measures the similarity of data in each group, or their "purity". For example, for regression tasks, the measure of impurity is the variance of the data in each group.

Hence, the data are iteratively split into two groups until reaching a stopping criterion (a leaf node) where the data are attributed to a certain class. When no stopping criterion is specified, the construction of the tree stops when each leaf node consists either in a single data point or a group of data belonging to the same class. Deep trees (i.e. involving a large amount of data split) are capable of approximating any arbitrarily complex functions for classification, regression or clustering task. However, they tend to overfit the data. Thus, three stopping criteria can be defined to prune trees. Those are (1) the maximum depth of the tree (2) the minimum number of samples per leaf and (3) the minimum amount of samples required to split a node. Given the relatively small amount of hyperparameters, the optimal set of values can be evaluated through grid search.

Estimating a tree is computationally fast so that it can be applied to large datasets. Furthermore, they are naturally suited for multiclass classification problems. The drawback is that individual decision trees tend to be sensitive to the set of data considered. Hence, they are particularly suitable for ensemble algorithms such as the RF, which computes the average performance of individual models each computed from a bootstrap replica of the data (e.g. random selection of samples with replacement).

Discrimination between classes can be visualized with the ROC and PR curves. Unless stated otherwise, performances of the algorithms are evaluated at classification thresholds that optimize the *sensitivity* and *specificity* of the detection. This optimization corresponds to the choice of the point in the upper left corner on the ROC curve. Altogether, the performances of the classification tasks are reported with their confusion matrix, the *precision*, *sensitivity*, *specificity* and their respective area under the ROC (AUROC) and PR curves (AUPREC).

It is worth noticing that, while the simple rules of single decision trees make their interpretation straightforward, it is not necessarily the case for a large collection of decisions trees generated from random data samples. Nevertheless, the value of each individual feature can still be estimated by considering the accumulated decrease in the split-criterion due to the use of this variable in each tree. As a result, a metric that characterizes the relative importance of each feature can be computed. Feature importance is particularly valuable to estimate the features that are the most valuable to discriminate between several classes. When the features are related to biologically meaningful intensities, such strategy permits the evaluation of the most important aspects of the processes involved.

5.4.1 Feature Selection

Generally, the feature engineering step provides a set of features for which it is not known in advance which of those are the most important. In order to reduce the risk of overfitting the data (when the number of training patterns is comparatively small to the amount of features) and improving the performances of the algorithm, one strategy is to further reduce the dimensionality of the feature space. The problem of feature selection is well-known in the machine learning field.

Depending on the number of features considered, some feature selection strategy can be preferred. In this thesis, a relatively small amount of features is considered, so that a greedy optimization technique based on *recursive feature elimination* (RFE) has been chosen to find the best performing subset of features [7]. First, the model is trained using the initial set of features and individual features importance is estimated. Then, the least important features are recursively removed from the features set until a stopping criterion is reached. Here, the stopping criterion has been chosen so that there is no further statistically significant improvements of model performances. By doing so, recursive feature elimination has the advantage of implicitly taking into account multivariable associations. Furthermore, the RFE has been demonstrated to be more robust to data overfitting than other methods [7].

Chapter 6

Detection and Characterization of Epileptic Seizure Events in Zebrafish

6.1 Contribution

Seizure occurrence in Zebrafish larvae is of several orders of magnitude higher than for humans. Their manual identification from LFP signals, therefore, is a time-consuming task. Hence, a novel and highly performant automatic seizure extraction algorithm is developed and applied to recordings of scn1lab, PTX-treated and PTZ-treated Zebrafish LFP recordings. This approach constitutes a novel framework for the *offline* (i.e. retrospective) automated extraction of seizures events from neurophysiological signals, which is typically performed manually.

In addition, the morphological signature of seizures of scn1lab, PTZ-treated and PTX-treated Zebrafish is investigated. For the first time, it is shown that a combination of dynamical biomarkers computed from the brain dynamical activity during seizures events correlate with specific biological mechanisms of the disease. For this purpose, this study leverages the large amount of Zebrafish seizures (923 in total) generated by distinct pathological mechanisms to formulate a discriminative model by means of interpretable machine learning techniques. As a result, the model developed is capable of differentiating between scn1lab, PTZ-treated and PTX-treated Zebrafish larvae seizures with a high degree of precision.

6.2 Introduction

To date, more than 500 genetic mutations have been associated with the epileptic condition in humans [37]. Yet, only few of these can precisely pinpoint the circuitry involved in seizures emergence [239]. Furthermore, genetic factors account for approximately 30% of the causes of recurrent seizures in patients [239]. In fact, seizures can be triggered by a wide range of brain insults (e.g. strokes, brain trauma, Alzheimer disease, etc.), infectious diseases or autoimmune diseases [37]. The current identification of the aetiology of epilepsy, however, remains principally based on its resulting phenotype (e.g. spasms, impaired awareness), or on the spatial aspect of seizures (localized or generalized) [37], which does not allow specific diagnosis of the underlying epileptogenetic mechanisms.

Recently, an universal framework based on dynamical bifurcation theory has been proposed to describe the initiation and termination of seizures in mice, humans and Zebrafish [10]. However, while general characteristics of the brain functional reorganization towards seizures can be formulated, distinct pathological aspects of the disease have been shown to influence seizure-onset patterns [62]. In particular, [62] has investigated the presence of static patterns such as bursts of polyspike or low-voltage fast activity at seizure onsets, but without being able to distinguish the underlying epileptogenetic mechanisms, as each pattern was shared by at least two or more pathologies.

To the best of our knowledge, the classification of the effects of the underlying pathology on the resulting morphology of the neurophysiological signals generated during seizures events has never been attempted with general dynamical measures. Such approach, however, has the potential to contribute to the understanding of the relationship between the collective neuronal dynamics as recorded by LFP and the specific biological processes involved in the disruption of the balance of neuronal activity. Furthermore, this classification is of broad significance for the precise diagnosis and personalized treatment of the epileptic condition. Ultimately, automatic detection and prediction algorithms would further benefit from a better understanding of morphological patterns that are subject-specific.

In this regard, the use of Zebrafish represents a considerable advantage to elucidate the specificity of dynamical patterns of seizures from LFP. Indeed, seizures in Zebrafish are several orders of magnitude more frequent than for humans, which allows a more robust evaluation of the intra-inter variability of their patterns across recordings. Furthermore, as an animal model, seizures mechanisms can be selectively triggered by several means (induced by drugs, or inherited by mutations) [51]. Chemoconvulsants drugs inserted

in the bathing medium, for instance, can reproduce acute seizures events, which might be subsequently used for rapid screening of antiepileptic drugs (AED). On the opposite, Zebrafish mutations often seek to reproduce the spontaneous and recurrent aspect of chronic seizures in humans. Finally, seizures induced in Zebrafish models have shown to closely resemble those in mammals from both physiological and behavioral aspects [240].

In this chapter, the correlation between the morphology of seizures events in neurophysiological signals and the underlying epileptogenic mechanism is investigated from LFP recordings of Zebrafish larvae. For this purpose, three seizures models were used: a genetic mutant (scn1lab) and two seizures inducing drugs, picrotoxin (PTX) and pentylenetetrazol (PTZ). On one hand, the sodium voltage-gated channel alpha subunit 1 mutation (scn1a) is a loss of function mutation which induces a dysfunction in the sodium voltage-gated channels. Voltage-gated sodium channels have a critical role in the generation and propagation of action potentials in the central and peripherical nervous systems. In the initial phase of the action potential, voltage-gated sodium channels are activated as a response of a membrane depolarization, causing the voltage across the neuronal membrane to increase. The membrane is then repolarized in response to a fast spontaneous inactivation of voltage-gated sodium channels, promoted by the activation of voltage-gated potassium channels. This is the falling phase of the action potential, characterized by a decrease in voltage across the membrane. Voltage-gated sodium channels then undergo a recovery phase, or deinactivation, where the inactivation gates reopen and the activation gates close, until they are ready to participate to another action potential. A common form of dysfunction in the sodium voltage-gated channel induced by mutations associated with epilepsy is a defect in their inactivation, caused by incomplete closure of the inactivation gate [241]. As a result, non-inactivating Na^+ current may facilitate neuronal hypercitability by reducing the threshold for an action potential to be triggered. However, the exact biophysical mechanisms of epileptogenesis caused by mutations of voltage-gated sodium channels remain uncertain. In particular, mutations in this channel have been found to cause genetic epilepsy and are specifically involved in Dravet syndrome (also known as severe myoclonic epilepsy of infancy) which is a rare form of early childhood epilepsy [242]. For this genetic model of Zebrafish, seizure emergence occurs naturally. On the other hand, Picrotoxin (PTX) and pentylenetetrazol (PTZ) are two chemoconvulsant drugs which bind to the GABA receptors inhibiting the flux of chloride ions in the post-synaptic neuron generating seizures by blocking the inhibitory synapse. For these two models, seizures were triggered by adding the drugs in the medium of wild type animals. In total, the data consisted in LFP recordings of 33 scn1lab, 31 PTX and 10 PTZ Zebrafish, each of 30 minutes. This rather large

Zebrafish individual cohort allowed to formulate statistically significant relationships between seizures morphology and the source of the epileptic condition.

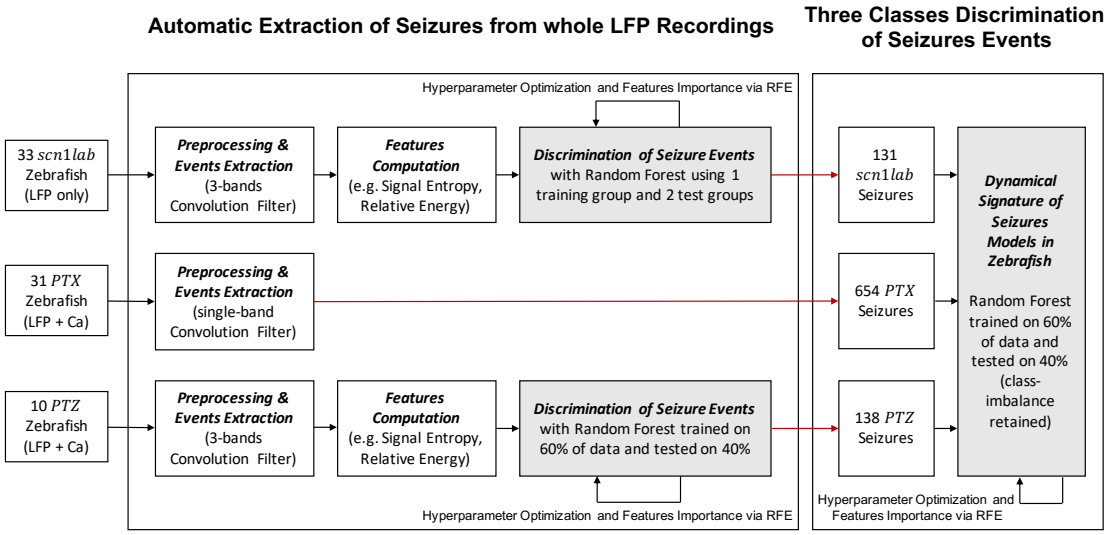

Fig. 6.1 **Summary of the Modeling Strategy to Automatically Detect and Characterize Seizures Originating from Different Zebrafish Models.** LFP recordings from 3 Zebrafish models (one genetic mutant and 2 seizures inducing drugs) are analyzed and used to extract a total of 923 seizures. Calcium imaging was used to validate seizures occurrence in LFP. The first step consists in extracting those seizures from the entire LFP signals. For this purpose, every abnormal brain activity has been extracted with a multi-resolution convolution filter and seizures subsequently discriminated with a machine learning approach. Once extracted for each Zebrafish model, the distinct characteristics of seizure types are investigated, which is formulated as a multi-class classification problem. Grey boxes represent those steps that involve the design of machine learning algorithms. Finally, the red lines illustrate seizure availability.

The investigation of the distinctive features between Zebrafish models is presented in two main steps (Figure 6.1). First, seizures are automatically extracted from retrospective LFP recordings using a newly introduced multi-resolution, convolution-based framework. Then, the investigation of the (in)variants features between seizures events is formulated as a multi-class classification task. To those ends, several interpretable machine learning prediction models were developed and their performance assessed. As a result, it is shown that the underlying epileptogenesis mechanisms can be distinguished with a high degree of accuracy by the use of dynamical measures applied at different scales of the signal. Notably, those measures are not subject- but pathology-specific.

6.3 Automatic Extraction of Seizures

6.3.1 Signal Processing and Events Extraction

The automatic extraction of seizures is a multi-resolution, convolution-based algorithm. The key ideas consist in extracting every abnormal brain dynamical behavior from the recordings and further classifying them into (non-)seizure events.

The original LFP signal is downsampled from 100kHz to 2000Hz. Indeed, it is very unlikely that any relevant biologically related mechanism would occur at frequencies higher than 2000Hz. Then, the 50Hz signal artefact was removed with a notch filter for which the Q factor was adapted to remove $\sim$ 2.5-3Hz around the notch. Finally, for the extraction step, the resulting signal is decomposed into a 10th level multi-resolution framework using the Maximal Overlap Discrete Wavelet Transform (MODWT) with a Daubechie 4 (db4) wavelet filter (see Chapter 5 for mathematical details). The wavelet transform is a time-frequency transformation of the signal that is particularly suited for the analysis of brief, transients events in non-stationary data. Such decomposition of the signal into its frequency bands (either arbitrarily large or into physiologically relevant ones) was considered in order to enhance the difference between neuronal activities that operate at different scales, which is not obvious from the full-spectrum recording.

Hence, the original signal is decomposed into its frequency sub-bands through a recursive filtering that follows a dyadic decomposition (following a power of 2) into increasingly smaller frequency bands. Conveniently, such dyadic decomposition of the original signal sampled at 2000Hz isolates the neuronal activity of Zebrafish into frequency bands that correspond to physiologically relevant brain rhythms in humans, thereby allowing straight comparison. The following frequency bands are then obtained: [1000-2000]Hz, [500-1000]Hz, [250-500]Hz, [125-250]Hz, [62-125]Hz, [31-62]Hz (gamma waves), [16-31]Hz (beta waves), [8-16]Hz (alpha waves), [4-8]Hz (theta waves), [2-4]Hz (high delta waves) and [0-2]Hz (low delta waves).

A brain activity signal was considered as abnormal if the LFP signal reaches the tail of its distribution in some frequency sub-bands of interest for seizures (Figure 6.2). To extract such activity from the entire recording, the signals of 3 frequency sub-bands ([62-125Hz],[31-62Hz],[15-31Hz]) are squared and passed through a rectangular convolution filter. The length of the convolution filter was chosen to be 1 second for scn1lab and PTX recordings and 2 seconds for PTZ recordings, because of the longer duration of seizures events in the latter. Those frequency bands were chosen for both their

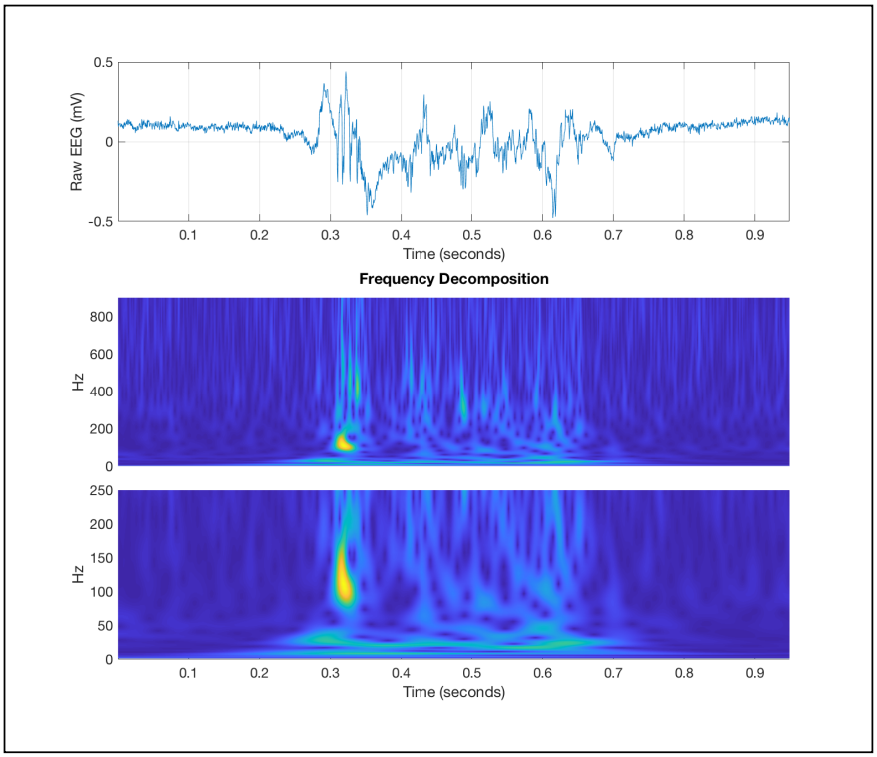

Fig. 6.2 **Continuous Wavelet Transform (CWT) of a typical (scn1lab) Zebrafish seizure.** The top panel represents the seizure as recorded from Local Fields Potential (LFP). The middle and top panel depicts the time-frequency decomposition of the seizure. The bottom one corresponds to a zoomed-in version (in the frequencies) of the middle one. On one hand, it can be seen that the very beginning of the seizure is marked by a peak in the high frequencies, with most of its energy centered around 100Hz. On the other hand, there is a pronounced change in the energy levels around the low frequencies during the whole seizure event.

sensitivity and specificity to seizures events. Then, a general threshold is applied on each frequency band to extract every seizure candidate. The threshold is chosen sufficiently low to extract every seizure from the signal but sufficiently high to frame the seizure event only. This threshold was chosen to be the 90th percentile of the distribution of the resulting convolved signals for scn1lab mutants and PTZ treated zebrafish. For PTX zebrafish, it was chosen to be the 70th percentile of the [125-250] Hz frequency band.

Finally, a machine learning algorithm is trained to differentiate between artefacts, interictal events and seizures. Here, Random Forests (RF) will be used for their interpretability, efficiency and robustness to noise. More specifically, the aim is to identify the features that contribute the most to the discrimination of seizure events from background activity, which carry potentials for further biomarkers development.

6.3.2 Discrimination of Seizure Events

A supervised RF classification task was implemented to automatically discriminate between seizures and non-seizures events. The following paragraphs detail the development of the predictive models for the extraction of both PTZ and scn1lab mutants seizures. As regards to PTX seizures, the thresholding procedure introduced previously was sufficient to extract seizures events only with a high degree of accuracy, so that training a statistical model was not required for those recordings specifically.

scn1lab Zebrafish

The candidates, i.e. extracted signals comprising seizures and non-seizures events, were first separated into three groups. One group to train the machine learning algorithm, another to validate it and the last one to ensure its generalization potential on totally unseen recordings. The first and second groups account for seizures originating from the same sub-set of individuals. The last group accounts for individuals that the algorithm has never seen at all. The purpose of this separation is to further avoid overfitting of seizure patterns due to a larger inter-variability than intra-variability of patterns originating from different LFP recordings (seizures events within a single recording tend to be similar).

The groups were organized so that the first and second groups together consisted in 14 recordings, while the third group consisted in 17 recordings from another set of individuals. Each recording has been performed on a different Zebrafish so that 31 individuals were used in total.

Candidates extracted from the first and second groups (608) were manually separated by experts into seizures and non-seizures events, therefore unveiling 84 seizures events. 60% of the seizures events was then randomly selected to train the prediction model (first group) and the remaining 40% for its validation (second group). The proportion of seizures and non-seizures candidates in each group was kept identical to retain class imbalance.

Features for the RF algorithm were selected based on their potential to discriminate brain dynamical behavior of different natures (during seizures and during background activity). For mathematical details, intuitive explanations and further motivations for the choice of each of the features mentioned hereafter, the reader is referred to the mathematical preliminaries of Chapter 5. On one hand, the Lyapunov Exponent (LE) computed from a reconstruction of the brain dynamics via delay-coordinate embedding in the phase-space will be used to characterize the loss of brain resilience and its temporal organization

during seizure and non-seizure events. On the other hand, the entropy (SampEn) of the signal will be estimated to investigate signal's predictability. Overall, both measures characterize the complexity of the brain temporal dynamics observed from LFP signals. The Hurst Exponent (HE) and the Fractal Dimension (FD), which correspond to measures of the system's memory, were not considered here due to the brief duration of candidate signals. It has been further reported that changes in the spatiotemporal patterns of neural activity during and towards seizures events often exhibit observable changes in the energy repartition over frequency sub-bands [200]. As such, the Relative Wavelet Energy (RWE) of each frequency band resulting from the MODWT decomposition of each candidate was considered. Finally, statistical measures (variance, skewness and kurtosis) were additionally computed over each frequency bands to characterize the distribution of high and low frequency components and potential differences, as dynamical changes might not be spread out equally across the entire spectrum [229]. To avoid boundary effects to dominate the signal information at the smallest scales (see Chapter 5 for further details on multi-resolution analysis with the wavelet transform), each candidate signal was decomposed following a 8-th level wavelet decomposition with db4 wavelets. The following frequency bands are then obtained for each seizure events: [1000-2000]Hz, [500-1000]Hz, [250-500]Hz, [125-250]Hz, [62-125]Hz, [31-62]Hz (gamma waves), [16-31]Hz (beta waves), [8-16]Hz (alpha waves) and [0-8]Hz (theta waves and delta waves).

To summarize, the following features were computed (50 in total):

- LE, SampEn, Variance, Skewness and Kurtosis from the original signal, leading to 5 features.
- SampEn, Variance, Skewness and Kurtosis for each of the signal sub-bands, leading to $4 \times 9 = 36$ features.
- The RWE for each frequency sub-band, leading to 9 features.

Due to the imbalance in the amount of seizures against non-seizures events proposed as candidates, a class-sensitive cost function was chosen to optimize the RF algorithm in order to raise the penalty resulting from missing a seizure event. Furthermore, the following hyperparameters were optimized using a grid search and 3 fold cross-validation to obtain the best possible generalization trade-off: number of trees in the forest (or estimators), maximum depth of the decision trees, minimum number of samples to make a split and minimum number of samples to be defined as a leaf. The best classifications results were obtained for 30 estimators, a maximum depth of 10 splits, 2 samples minimum to make a split and one single sample minimum to be defined as a leaf.

		Predicted Condition	
	Group 2	Seizure	Interictal Activity
Actual Condition	Seizure	32	1
	Interictal Activity	1	210

		Predicted Condition	
	Group 3	Seizure	Interictal Activity
Actual Condition	Seizure	46	1
	Interictal Activity	3	582

Table 6.1 **Confusion Matrices for the Automatic Extraction of Seizures from scn1lab Zebrafish.** On the left-hand side, the results for group 2, which consists in a set of seizures originating from the same individuals as from the training set. Those results correspond to a *precision* 97%, a *specificity* of 99.5% and a *sensitivity* of 97%. On the right-hand side, the results for group 3, which consists in seizures extracted from previously completely unseen individuals. Those results correspond to a *precision* 93.9%, a *specificity* of 99.5% and a *sensitivity* of 97.9%.

The confusion matrices describing the resulting classification performances are displayed on Table 6.1. A *precision* of 97%, 99.5% *specificity* and 97% *sensitivity* were obtained for the test set that corresponds to partially seen recordings (group 2, 14 individuals). For this test set, the discrimination algorithm showed a *AUROC* value of 0.999 and *AUPREC* of 0.997 (Figure 6.3). Performances of the RF on totally unseen recordings (group 3, 17 individuals), were as following: 93.9% *precision*, 99.5% *specificity* and 97.9% *sensitivity*. For this generalization test set, the discrimination algorithm showed a *AUROC* value of 0.999 and *AUPREC* of 0.994 (Figure 6.3). In biomedical applications such as the one presented here, a particularly important aim is to avoid missing information related to the clinical condition under investigation, while providing as few false negatives as possible. Here, it is worth noticing that only two seizure events have been misclassified as non-seizure events over the entire course of the classification task over both groups, which totals 876 candidates. Furthermore, the performances between the two test groups are very much similar, hence suggesting that the predictive model captures the key characteristics of seizures events across Zebrafish larvae. The performances of the classifier were further evaluated on their capability of discriminating seizures from a "normal" brain behavior (baseline) and subsequently achieved a perfect classification. This result suggests that the automatic detection algorithm could be further extended for the purpose of a real-time monitoring framework.

The importance of each feature for distinguishing between seizure and non-seizure events was investigated via Recursive Feature Elimination (RFE). Following this approach, multiple RF classifiers were trained using a subset of features of decreasing size and their performances estimated, hence ranking predictors according to their contribution to the model outcome. It should be noted, however, that the features that are

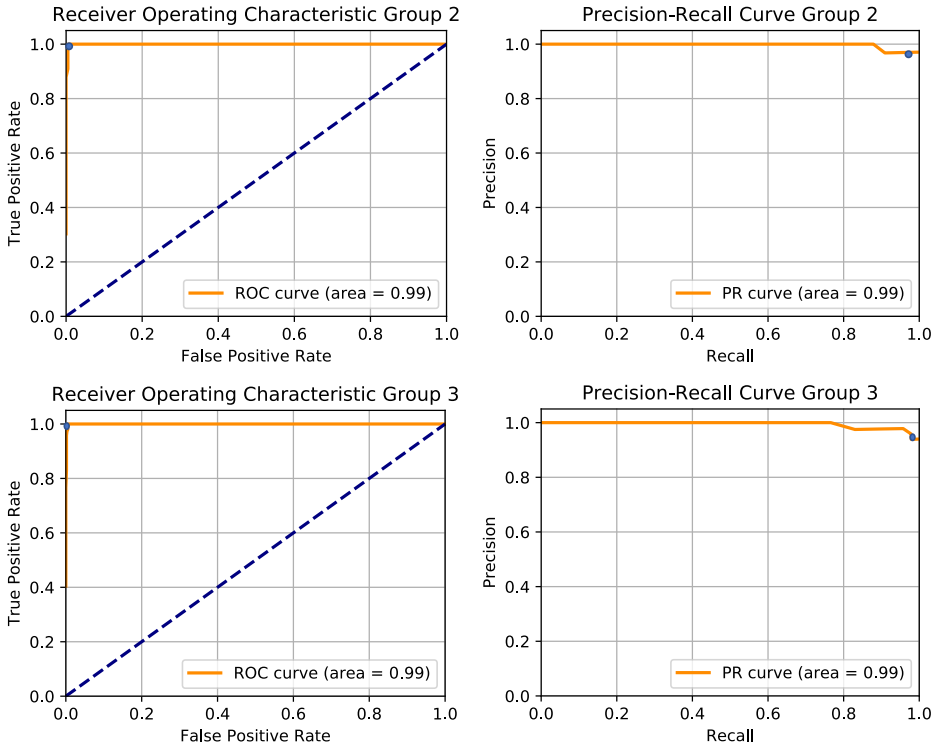

Fig. 6.3 **Receiver Operating Characteristic (left) and Precision-Recall Curve (right) of the Automatic Detection of scn1lab Seizures.** The optimal decision thresholds are represented by the blue dot on all graphs.

removed at last do not necessarily account for those that are individually the most relevant. Only the subset of the top ranked features taken together is optimal in this sense. Notably, such strategy conveniently accounts for multivariable associations between features.

As a result, the information across the highest frequency bands was selected as highly discriminative of seizures events within the signal (among other brain dynamical states or artefacts). More precisely, the entropy and variance of frequency sub-bands [250-500Hz] and [500-1000Hz] appeared in the top most discriminative features and accounted for most of the total feature importance ($\sim$70%) (Figures 6.4 and 6.5 and Table 6.2). This result is consistent with the literature that describes the appearance of very high frequency oscillations (HFO) during epileptic seizure events [69, 243, 244, 245, 246, 247, 248, 249, 250].

Further analysis of seizures candidates revealed that a majority of non-seizures events were capturing so called interictal events (IEDs) from the signal. It is interesting to notice that, as a classification between seizures and non-seizures candidates, our analysis further supports that IEDs and seizures hold distinct temporal and spectral features [251].

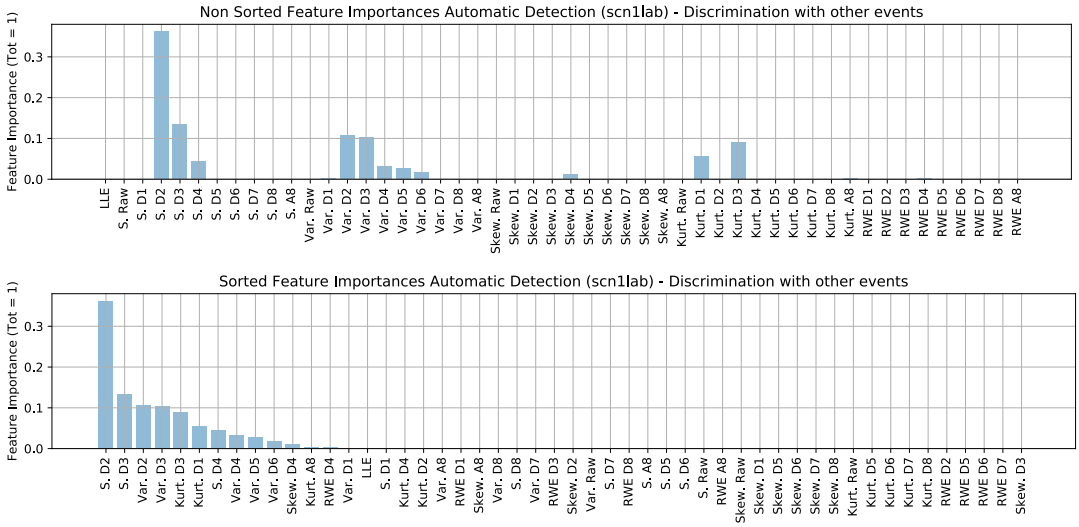

Fig. 6.4 **Feature Importance for the Automatic Detection of Seizures Events in scn1lab Zebrafish.** The entropy (S in the graph), energy (RWE) and statistical moments (Var.; Kurt.; Skew.) were computed from both the original signal and for each frequency sub-bands. Each sub-band has been assigned a code which corresponds to the decomposition level. Hence, D* and A* respectively corresponds to the detail and approximation coefficients of the discrete wavelet transform. Then, D1 = [1000-2000]Hz, D2 = [500-1000]Hz, D3 = [250-500]Hz, D4 = [125-250]Hz, D5 = [62-125]Hz, D6 = [31-62]Hz, D7 = [16-31]Hz, D8 = [8-16]Hz and A8 = [0-8]Hz.

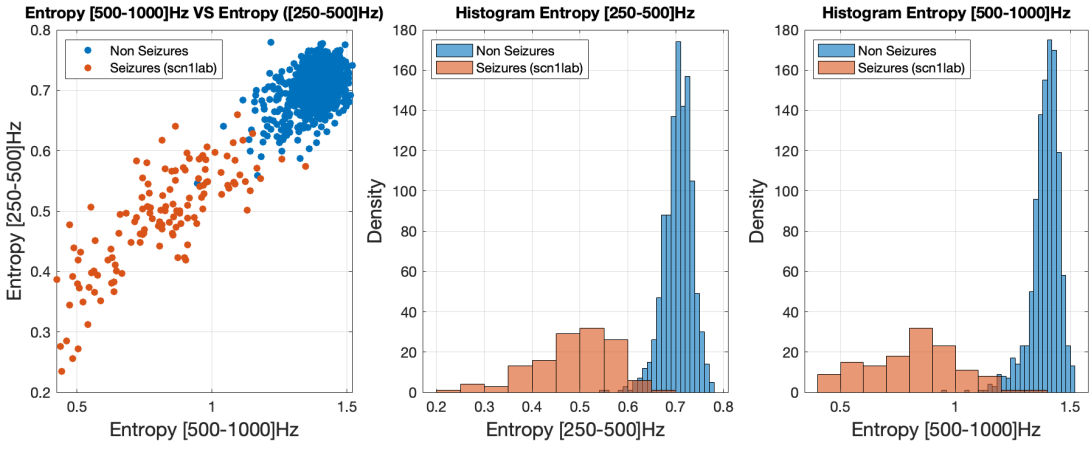

Fig. 6.5 **Top 2 Most Important Features for the Automatic Detection of Seizures Events in scn1lab Zebrafish.** The top 2 tanked features are displayed (Entropy in the [250-500]Hz and [500-1000]Hz frequency bands). Those 2 features together account for 50% of the overall feature importance. As such, clusters are clearly visible from the first panel.

	Seizures (IQR)	Interictal Activity (IQR)
Entropy [500-1000]Hz	$0.82[0.65-0.95]$	$1.40[1.36-1.43]$
Entropy [250-500]Hz	$0.49[0.43-0.55]$	$0.70[0.68-0.72]$
Variance [500-1000]Hz	$2.92[1.74-4.54](e^{-4})$	$0.40[0.30-0.53](e^{-4})$
Variance [250-500]Hz	$3.16[1.84-4.99](e^{-4})$	$0.26[0.20-0.37](e^{-4})$

Table 6.2 **Feature Importance represented in interquartile ranges (IQR) for the Discrimination of Seizure vs Non Seizures Events in scn1lab Zebrafish.**

PTZ Zebrafish

The same modeling procedure has been applied to classify candidates extracted from entire recordings of brain activity in PTZ treated Zebrafish. Hence, features selection and model optimization were performed identically. However, due to the relatively smaller set of available recordings (10 individuals), the data have only been separated into a training set (60% of the candidates) and a test set (40% of the candidates), that is, without a second test set where the candidates originates from entirely unseen individuals. Seizures events were validated by the co-occurrence of a peak in the recordings of calcium activity.

As a result, the training set consisted in 87 seizures and 376 non-seizures events, while the test set comprised 51 seizures and 259 non-seizures events. Hyperparameters were tuned so that the model structure that provided the best discriminative algorithm was obtained with the following values: 120 estimators, a maximum depth of 20 splits, 2 samples minimum to make a split and one single sample minimum to be considered as a leaf. A *precision* of 79.3%, 95.4% *specificity* and 90.2% *sensitivity* were obtained for the test set, suggesting that either differentiating the seizures from background activity in PTZ treated Zebrafish is a more challenging task than for scn1lab mutants for an automatic extraction algorithm or that PTZ induced seizures have more diversified patterns. The algorithm showed a *AUROC* value of *0.982* and AUPREC value of 0.913 (Figure 6.6). The confusion matrix is illustrated on Table 6.3.

Feature importance was computed as before with RFE. As a result, the brain activity captured by the kurtosis and the entropy in the relatively high frequency bands ([125-250]Hz and [250-500]Hz) were identified as the most discriminative between the seizures and non seizures candidates extracted by the multi-resolution convolution filter (Figures

		Predicted Condition	
		Seizure	Interictal Activity
Actual Condition	Seizure	46	5
	Interictal Activity	12	247

Table 6.3 **Confusion Matrix for the Automatic Extraction of Seizures from PTZ induced seizures.** A class-sensitive cost function has been designed to further penalize the amount of false negative, so that the algorithm favors the detection of seizures over false positive predictions.

6.7 and 6.8 and Table 6.4). This result further supports that IEDs and seizures hold distinct temporal and spectral features in PTZ-induced seizures as well, since the non-seizures candidates mostly consist of IEDs. Furthermore, the contribution of each individual feature appears more diluted than for the model that discriminates between seizures and non seizures in scn1lab Zebrafish, which suggests that the bifurcation towards the seizure state in PTZ-treated Zebrafish is the result of more complex dynamical mechanisms that span across additional scales (i.e. frequency bands, or different dynamical relationships between neurons). More specifically, starting from phenomena occurring at 125Hz and above, up to 1000Hz, as compared to [250-1000]Hz in mutant (scn1lab) seizures.

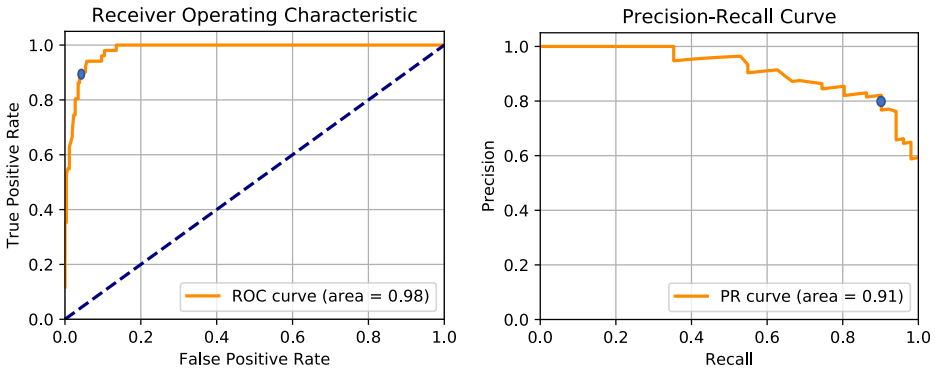

Fig. 6.6 **Receiver Operating Characteristic (left) and Precision-Recall Curve (right) of the Automatic Detection of PTZ Seizures.** The optimal decision threshold is represented by the blue dot on both graphs.

6.4 A Dynamical Signature of Seizure Models in Zebrafish

One of the major issues in identifying consistent biomarkers for epilepsy is its wide and heterogenous range of pathological mechanisms. Hence, the investigation of epileptogenic mechanisms is intrinsically hindered by the inherent difficulty of constituting a

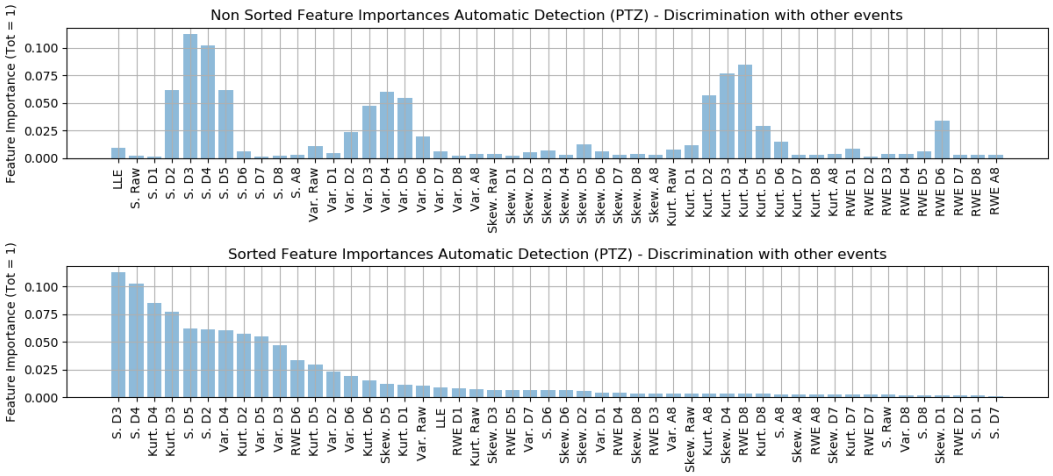

Fig. 6.7 **Feature Importance for the Automatic Detection of Seizures Events in PTZ Zebrafish.** The entropy (S in the graph), energy (RWE) and statistical moments (Var.; Kurt.; Skew.) were computed from both the original signal and for each frequency sub-bands. Each sub-band has been assigned a code which corresponds to the decomposition level. Hence, D* and A* respectively corresponds to the detail and approximation coefficients of the discrete wavelet transform. Then, D1 = [1000-2000]Hz, D2 = [500-1000]Hz, D3 = [250-500]Hz, D4 = [125-250]Hz, D5 = [62-125]Hz, D6 = [31-62]Hz, D7 = [16-31]Hz, D8 = [8-16]Hz and A8 = [0-8]Hz.

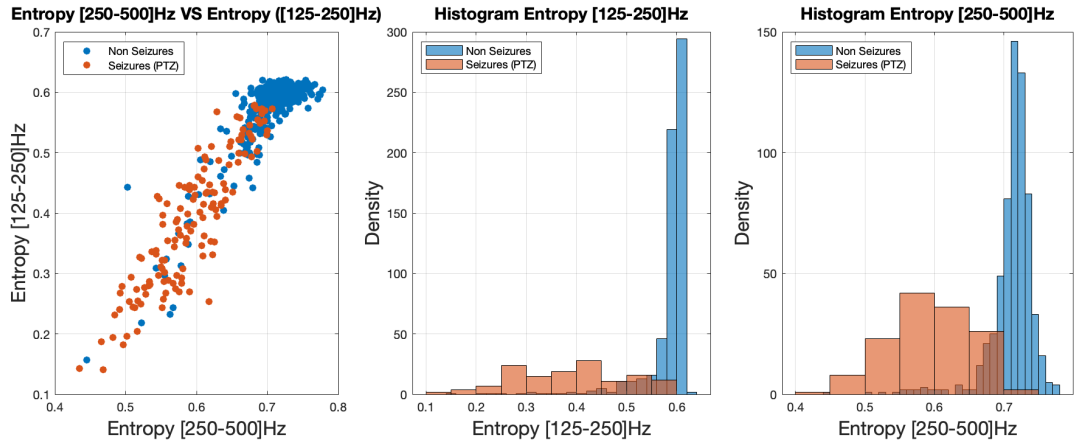

Fig. 6.8 **Top 2 Most Important Features for the Automatic Detection of Seizures Events in PTZ Zebrafish.** The top 2 tanked features are displayed (Entropy in the [250-500]Hz and [125-250]Hz frequency bands). Those 2 features together only account for 22% of the overall feature importance. As such, clusters are not clearly differentiable by only illustrating the relationship between the top 2 ranked features.

	Seizures (IQR)	Interictal Activity (IQR)
Entropy [250-500]Hz	0.59 [0.55 - 0.63]	0.72 [0.70 - 0.73]
Entropy [125-250]Hz	0.39 [0.29 - 0.48]	0.60 [0.59 - 0.61]
Kurtosis [250-500]Hz	11.95 [8.88 - 17.20]	3.19 [3.02 - 3.06]
Kurtosis [125-250]Hz	17.33 [13.86 - 28.61]	3.31 [2.98 - 4.51]

Table 6.4 **Feature Importance represented in interquartile ranges (IQR) for the Discrimination of Seizure vs Non Seizures Events in PTZ-treated Zebrafish.**

database of patients sharing an identical condition. In this sense, the use of an animal model is particularly suited to study distinct seizures mechanisms. This study takes advantage of a large cohort of Zebrafish individuals (74 in total) with 3 distinct epileptogenic mechanisms to explore the association between LFP signals patterns and the corresponding cellular mechanisms responsible for seizures emergence. An attempt is made to discriminate seizures patterns according to their associated epileptogenic mechanisms. This problem is formulated as a three-class classification problem where a RF algorithm is trained to differentiate between PTX, PTZ and scn1lab Zebrafish individuals based on the morphology of their respective seizures. Features weights are retained from the classification task and further analyzed to gain novel biological insights on the dependence of the intrinsic characteristics of seizures towards their generating mechanism.

The Random Forest algorithm was used to discriminate the seizures types based on the Zebrafish model it originates from. The features considered were the SampEn, the RWE and the statistical moments. Indeed, due to the brief duration of mutant seizures, features characterizing system memory, i.e. the Hurst Exponent and the Fractal Dimension (see Chapter 5), were not computed. Furthermore, the Lyapunov Exponent was not used as a potentially discriminative dynamical feature due to the intrinsic lack of robustness of its estimation from short-time series data generated by nonlinear deterministic processes (see Chapter 5 for details). The MODWT decomposition was performed on 8 levels with db4 wavelets filters. As a summary, the following features were used (49 in total):

- SampEn, Variance, Skewness and Kurtosis from the original signal, leading to 4 features.

- SampEn, Variance, Skewness and Kurtosis for each of the signal sub-bands, leading to $4 \times 9 = 36$ features.
- The RWE for each frequency sub-band, leading to 9 features.

6.4.1 Two Classes Classification (Drug - Mutant)

First, the problem of differentiating between drug-induced seizures and spontaneous seizures of the scn1lab Zebrafish mutant is addressed. The data were separated into two groups, the training set and the test set, which respectively comprised of 60% and 40% of the total amount of available seizures. Class imbalance has been maintained in both training set and test set, which therefore consisted in 479/74 Drug/Mutants seizures and 313/57 Drug/Mutants seizures respectively. The objective function of the RF has been modified to account for the high class imbalance, so that the performance of the algorithm are not artificially inflated because of incorrectly predicting outcomes to be part of the largest class only (in this case, drug-induced seizures). The hyperparameters of the RF were optimized using a grid search and 3 fold cross-validation. The best classification performances were obtained for the following hyperparameters: 160 estimators, a maximum depth of 20 splits, 2 samples minimum to make a split and one single sample minimum to be considered as a leaf. The algorithm obtained a perfect classification with an AUROC value of 1, so that both the *sensitivity* and *specificity* were of 100%. As a result, we suggest that there exist significant dynamical differences between the seizures patterns that have been produced by two distinct pathways, that is, drug induced (PTX or PTZ) or by a genetic mutation (scn1lab).

Feature importance was obtained as before and the results displayed on Figures 6.9 and 6.10. Interestingly, the entropy in both the lowest frequency band (theta and delta waves) and very high frequencies ([500-1000Hz]) account for a significantly large importance in discriminating between the seizures being induced by the drugs and those that are naturally occurring with scn1lab Zebrafish. As a conclusion, the main differences between drug-induced seizures and mutant ones occur simultaneously on the very high and very low frequency bands. The distributions of both the entropy in the [0-8]Hz and [500-1000]Hz frequency sub-bands are detailed on Table 6.6 (IQR).

6.4.2 Two Classes Classification (PTX - PTZ)

The problem of differentiating the morphology of seizures induced by the two types of drugs, i.e. PTX and PTZ, is here addressed. The data were separated into two groups, the training set and the test set, which respectively comprised of 60% and 40% of the total

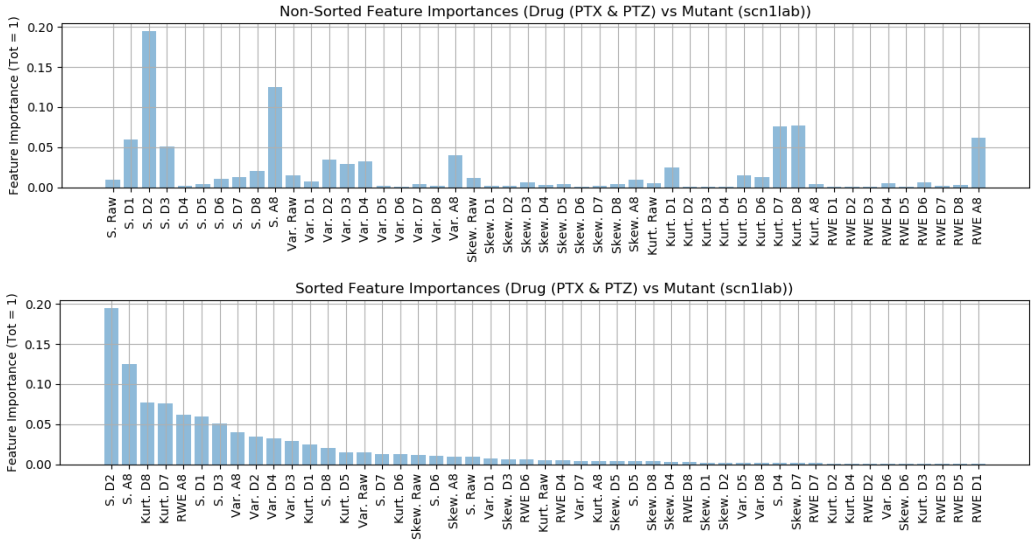

Fig. 6.9 **Feature Importance for Two Classes Classification Task (both drug induced seizures (PTX - PTZ) against the genetic variant scn1lab Zebrafish).** The entropy (S in the graph), energy (RWE) and statistical moments (Var.; Kurt.; Skew.) were computed from both the original signal and for each frequency sub-bands. Each sub-band has been assigned a code which corresponds to the decomposition level. Hence, D* and A* respectively corresponds to the detail and approximation coefficients of the discrete wavelet transform. Then, D1 = [1000-2000]Hz, D2 = [500-1000]Hz, D3 = [250-500]Hz, D4 = [125-250]Hz, D5 = [62-125]Hz, D6 = [31-62]Hz, D7 = [16-31]Hz, D8 = [8-16]Hz and A8 = [0-8]Hz.

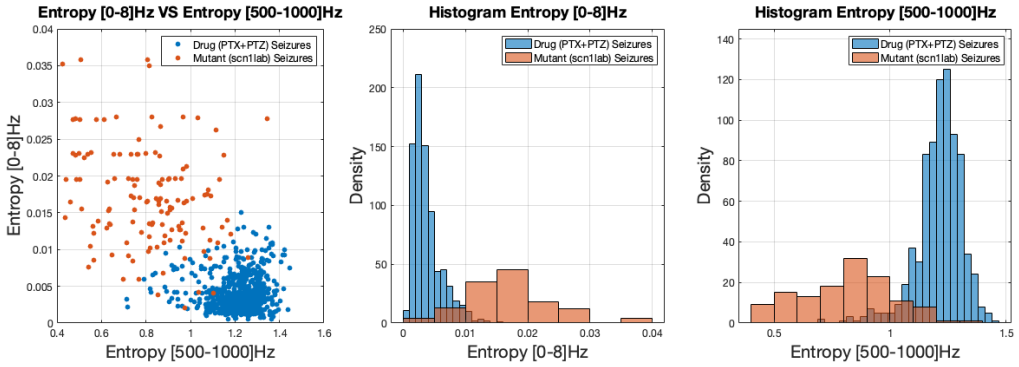

Fig. 6.10 **Feature Importance for Two Classes Classification Task (both drug induced seizures (PTX - PTZ) against the genetic variant scn1lab Zebrafish).** The top 2 tanked features are displayed (Entropy in the [0-8]Hz and [500-1000]Hz frequency bands). Those 2 features respectively represent the theta and delta waves and the very high frequency bands, which together account for 32% of the overall feature importance. As such, clusters are clearly visible from the first panel.

		Predicted Condition	
		Mutant Seizure (scn1lab)	Drug Seizure (PTX + PTZ)
Actual Condition	Mutant Seizure (scn1lab)	57	0
	Drug Seizure (PTX + PTZ)	0	313

Table 6.5 **Confusion Matrix of the Two-Class Classification Task (Drug (PTX-PTZ) and Mutant (scn1lab) seizures).**

	Drug Seizures (PTX + PTZ) (IQR)	Mutant Seizures (scn1lab) (IQR)
Entropy [500-1000]Hz	1.22 [1.17 - 1.28]	0.82 [0.65 - 0.95]
Entropy [0-8]Hz	$0.31[0.22-0.46](e^{-02})$	$1.65[1.27-2.12](e^{-02})$

Table 6.6 **Top 2 Ranked Feature Importance represented in interquartile ranges (IQR) for the Discrimination of Drug-induced seizures (PTX + PTZ) and Mutant seizures (scn1lab) Zebrafish.**

amount of available seizures. Class imbalance has been maintained in both training set and test set, which therefore consisted in 393/83 PTX/PTZ seizures and 262/55 PTX/PTZ seizures respectively. The objective function of the RF has been modified to account for the high class imbalance. The hyperparameters of the RF were optimized using a grid search and 3 fold cross-validation. The best classification performances were obtained for the following hyperparameters: 180 estimators, a maximum depth of 10 splits, 2 samples minimum to make a split and one single sample minimum to be considered as a leaf. The algorithm obtained an AUROC value of 0.986, a *sensitivity* of 96.4% and a specificity of 94.7% (Figure 6.11). The confusion matrix at the optimal decision threshold is displayed on Table 6.7. Hence, distinguishing PTX seizures out of both PTX and PTZ drug-induced seizures can be performed with a relatively high degree of accuracy.

Feature importance was obtained as before and the results displayed on Figures 6.12 and 6.13. In a similar observation than for the distinction between drugs-induced and mutant seizures, the entropy seems to account for a large part of overall feature importance, meaning that seizures hold an intrinsically different dynamics in each of those Zebrafish models as well. However, we note significant differences with the previously developed

prediction model. In particular, the discrimination between the PTX and PTZ-induced seizures occurs mostly in the low frequencies. Hence, the Relative Wavelet Energy (RWE) in the delta-theta waves ([0-8]Hz) and the entropy in the gamma waves ([31-62]Hz) together account for the two most important features. The distributions of both the RWE in the [0-8]Hz and the entropy in the [31-62]Hz frequency sub-bands are detailed on Table 6.8 (IQR).

		Predicted Condition	
		PTZ Seizure	PTX Seizure
Actual Condition	PTZ Seizure	53	2
	PTX Seizure	14	248

Table 6.7 **Confusion Matrix of the Two-Class Classification Task (PTX-PTZ induced seizures).**

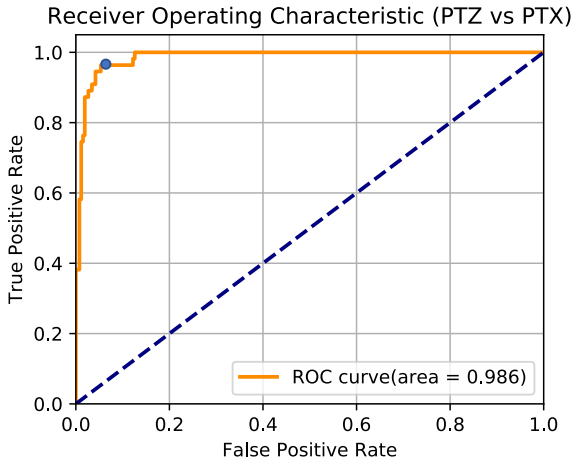

Fig. 6.11 **Two Classes Receiver Operating Characteristic of the discrimination of PTX and PTZ-induced seizures.** The optimal decision thresholds are represented by the blue dot on all graphs.

6.4.3 Three Classes Classification (PTX - PTZ - scn1lab)

In this paragraph, a predictive model is trained to distinguish between all kinds of seizures types together, that is, between PTX-treated, PTZ-treated and scn1lab mutant Zebrafish seizures. From a clinical point of view, this experimental setup is more realistic since

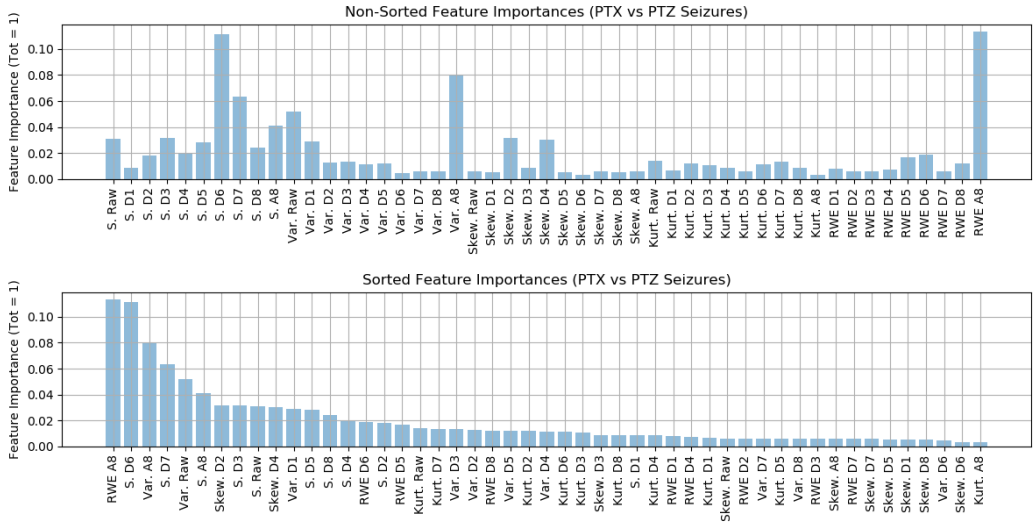

Fig. 6.12 **Feature Importance for Two Classes Classification Task (drug induced seizures PTX and PTZ).** The entropy (S in the graph), energy (RWE) and statistical moments (Var.; Kurt.; Skew.) were computed from both the original signal and for each frequency sub-bands. Each sub-band has been assigned a code which corresponds to the decomposition level. Hence, D* and A* respectively corresponds to the detail and approximation coefficients of the discrete wavelet transform. Then, D1 = [1000-2000]Hz, D2 = [500-1000]Hz, D3 = [250-500]Hz, D4 = [125-250]Hz, D5 = [62-125]Hz, D6 = [31-62]Hz, D7 = [16-31]Hz, D8 = [8-16]Hz and A8 = [0-8]Hz.

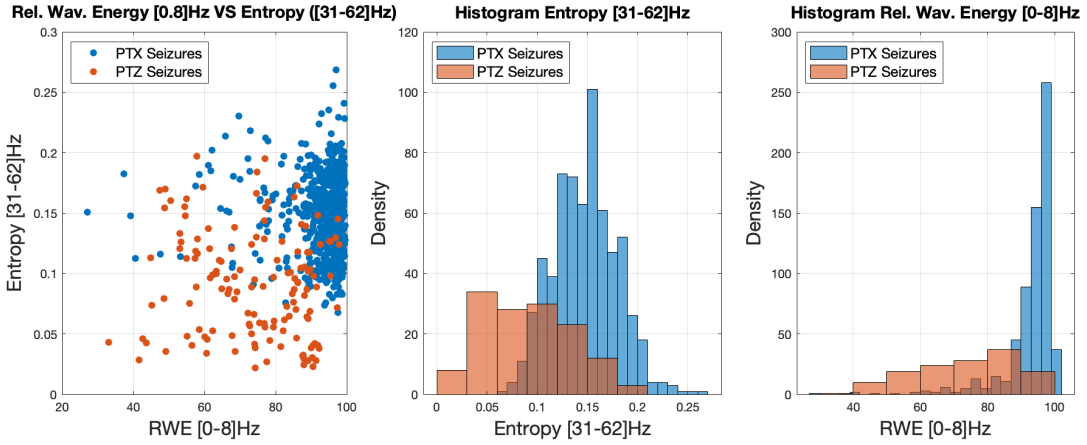

Fig. 6.13 **Feature Importance for Two Classes Classification Task (drug induced seizures (PTX - PTZ) and genetic variant scn1lab).** The top 2 tanked features are displayed (Relative Wavelet Energy in the [0-8]Hz and entropy in the [31-62]Hz frequency bands). Those 2 features together account for 22% of the overall feature importance. As such, clusters are only slightly visible on the first panel.

	PTZ Seizures (IQR)	PTX Seizures (IQR)
Rel. Wav. Energy [0-8]Hz (%)	75.98 [61.39 - 87.60]	95.35 [91.94 - 97.92]
Entropy [31-62]Hz	$0.89[0.53-1.24](e^{-01})$	$1.49[1.25-1.69](e^{-01})$

Table 6.8 **Top 2 Ranked Feature Importance represented in interquartile ranges (IQR) for the Discrimination of Drug-induced seizures (PTX + PTZ).**

seizures can be triggered by a significantly wide range of pathological causes and as such, patient datasets are intrinsically vastly heterogeneous. Hence, this study investigates the possibility for the specific identification of the biological mechanisms responsible for seizure emergence from the resulting dynamical patterns of seizures. For this purpose, the problem of differentiating seizures types by their underlying pathological mechanisms is here formulated as three-class classification problem. In particular, Random Forest was used to gain insights on the most important signal characteristics that have the potential to serve as biomarkers. Such methodology evaluates the signal properties that remain similar across class and those that remain similar within a single class and different from the others. The latter can be further used to discriminate the underlying mechanisms of seizure genesis. Interestingly, it is already worth noticing that seizures lasted on average 4.91 ± 2.62, 6.91 ± 5.07 and 0.97 ± 0.36 seconds for PTX, PTZ and mutants zebrafish models respectively, which could already serve as a valuable predictor. The respective durations of seizures for the scn1lab and the PTZ-treated Zebrafish models are in line with the findings of [56]. Nevertheless, the focus is here on the morphology of seizures themselves across different resolutions and biological triggers.

The data were separated into two groups, the training set and the test set, which respectively comprised of 60% and 40% of the total amount of available seizures. Class imbalance has been maintained in both training set and test set, which therefore consisted in 397/82/74 PTX/PTZ/scn1lab seizures and 257/56/57 PTX/PTZ/scn1lab seizures respectively. The objective function of the RF has been modified to account for class imbalance, so that samples weights in the overall performances of the algorithm are modified according to their occurrence in their respective class. The following hyperparameters were optimized using a grid search and 3 fold cross-validation to obtain the best possible performances on the test set: number of trees in the forest (or estimators), maximum

		Predicted Condition	
		PTX	PTZ + scn1lab
Actual Condition	PTX	247	10
	PTZ + scn1lab	6	107

		Predicted Condition	
		PTZ	PTX + scn1lab
Actual Condition	PTZ	53	3
	PTX + scn1lab	14	300

		Predicted Condition	
		scn1lab	PTX + PTZ
Actual Condition	scn1lab	57	0
	PTX + PTZ	0	313

Table 6.9 **Confusion Matrices of the Three-Class Classification Task (PTX - PTZ - scn1lab).** The algorithm is less efficient in discriminating between PTX and PTZ than scn1lab seizures and any others.

	Specificity (%)	Sensitivity (%)	Support
PTX Seizures	97.3	96.1	257
PTZ Seizures	95.5	94.6	56
scn1lab Seizures	100	100	57
Weighted avg/total	97.4	96.5	370

Table 6.10 **Classification Report for the Three-Class Classification Task (PTX - PTZ - scn1lab).** *Specificity* and *sensitivity* are reported as a one-versus-all classification performance (in %). It is worth noticing that scn1lab seizures benefit from perfect precision, meaning that once a scn1lab seizure has been identified, it almost certainly belongs to the mutant seizure class. The support value illustrates the amount of seizures considered.

	PTX Seizures (IQR)	PTZ Seizures (IQR)	scn1lab Seizures (IQR)
RWE [0-8]Hz (%)	95.4 [91.9 - 97.9]	76.0 [61.4 - 86.0]	53.1 [36.2 - 68.7]
Kurtosis [8-16]Hz	6.36 [4.62 - 9.35]	6.72 [5.41 - 8.91]	2.24[1.97 - 2.93]
Entropy [0-8]Hz	$0.29[0.20-0.40](e^{-02})$	$0.56[0.36-0.77](e^{-02})$	$1.65[1.26-2.11](e^{-02})$

Table 6.11 **Top 3 Ranked Features Importance represented in interquartile ranges (IQR) for the discrimination of seizures originating from different animal models (PTX-PTZ-scn1lab).**

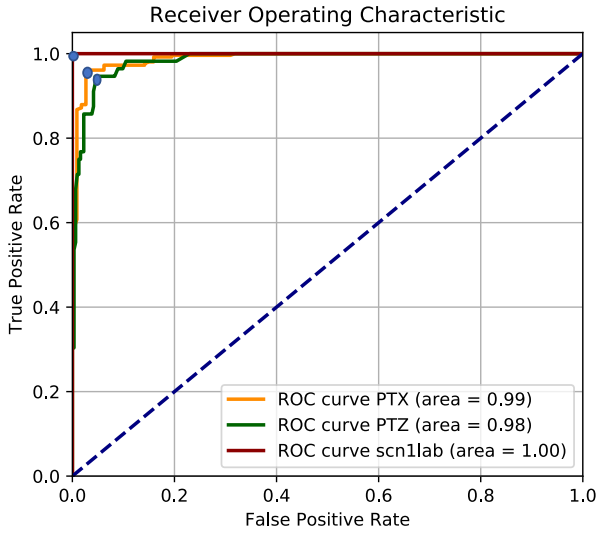

Fig. 6.14 **Multi-Class Receiver Operating Characteristic of the discrimination of seizures originating from different animal models (PTX-PTZ-scn1lab).** The optimal decision thresholds are represented by the blue dot on all curves.

depth of the decision trees, minimum number of samples to make a split and minimum number of samples to be defined as a leaf. The best classifications results were obtained for 160 estimators, a maximum depth of 20 splits, 2 samples minimum to make a split and one single sample minimum to be defined as a leaf. For a three-class classification problems, the *specificity* and *sensitivity* are computed for a one-versus-all identification (Table 6.10). In other words, their respective performance metrics correspond to the task of discriminating the current class against all the others. The confusion matrices of the resulting classifications performances on the test set at the optimal operating points are displayed on Tables 6.9 and 6.10 and the ROC curves on Figure 6.14.

The importance of each feature in the classification was computed as before. In particular, 3 features accounted for the highest discriminative potential across Zebrafish models. The features were the following: the relative energy (RWE) of the [0-8]Hz frequency band (theta and delta waves), the kurtosis of the [8-16]Hz (alpha waves) and the entropy (SampEn) of the [0-8]Hz frequency sub-band (Figures 6.15, 6.16, 6.17 and Table 6.11). Notably, among the top 4 contributing features are those that describe the brain activity at the lowest frequency bands ([0-8]Hz, i.e. theta and delta waves). The appearance of the kurtosis of the signal in the alpha frequency sub-bands is an interesting predictor. Indeed, the kurtosis is an indicator of data distribution around the mean. A lower kurtosis value in the [8-16]Hz frequency sub-band for scn1lab seizures illustrates a larger amount of wide oscillations, which might suggest a more prominent neuronal

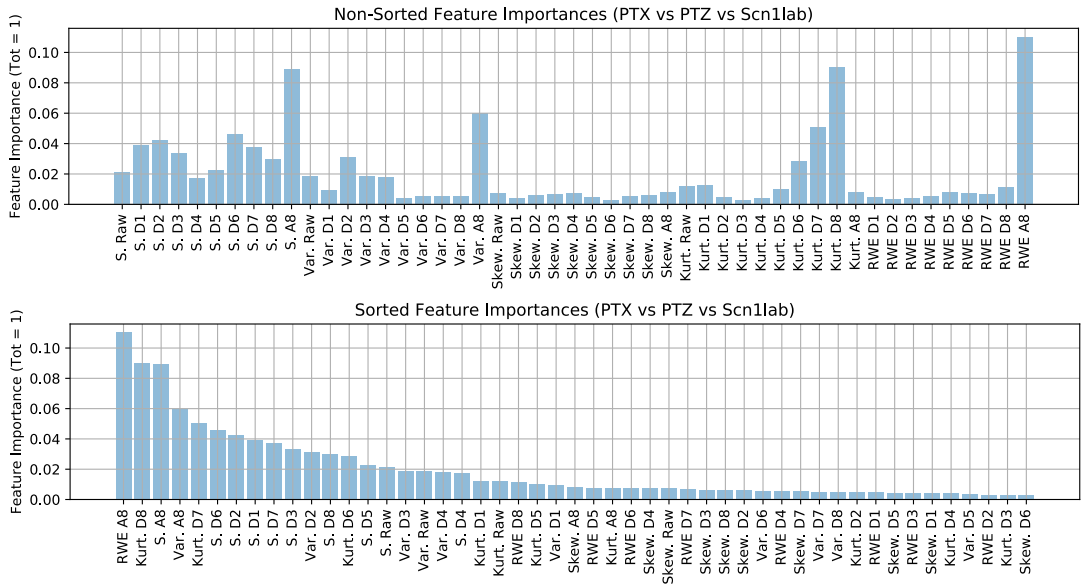

Fig. 6.15 **Feature Importance for Three Classes Classification Task (PTX - PTZ - scn1lab).** The entropy (S in the graph), energy (RWE) and statistical moments (Var.; Kurt.; Skew.) were computed from both the original signal and for each frequency sub-bands. Each sub-band has been assigned a code which corresponds to the decomposition level. Hence, D* and A* respectively corresponds to the detail and approximation coefficients of the discrete wavelet transform. Then, D1 = [1000-2000]Hz, D2 = [500-1000]Hz, D3 = [250-500]Hz, D4 = [125-250]Hz, D5 = [62-125]Hz, D6 = [31-62]Hz, D7 = [16-31]Hz, D8 = [8-16]Hz and A8 = [0-8]Hz.

activity at this scale. Moreover, quantitative measures of signal complexity and dynamics appeared consistently among the top ranked features (6 out of 10). In total, the entropy accounts for 38% of feature importance compared to 22% for the kurtosis, 18% for the variance, 16% for the relative energy and 6% for the skewness of each frequency sub-band. This result suggests that the seizure specific dynamics, as recorded by LFP, holds the potential to distinguish between the underlying epileptogenic mechanisms. In other words, intrinsic dynamical patterns can be extracted from seizures generated by distinct biological mechanisms such as those resulting from the scn1lab genetic variant, PTX-induced and PTZ-induced seizures.

Overall, the resulting high classification performance supports that such discrimination can be done in Zebrafish with a relatively high degree of accuracy. Eventually, designing a two-steps classification model could improve the classification performances by first identifying seizures that are very likely to belong to scn1lab Zebrafish, and then further distinguish between PTZ-treated and PTX-treated Zebrafish.

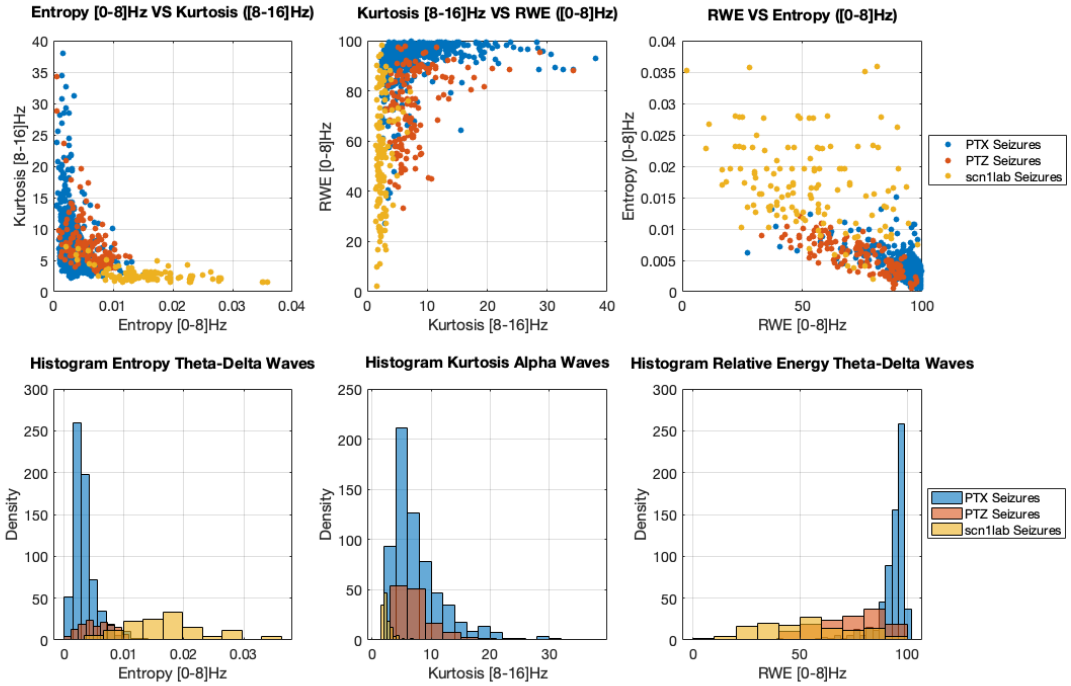

Fig. 6.16 **Feature Importance for the Three Classes Classification Task (PTX - PTZ - scn1lab).** The top 3 ranked features are displayed (Energy and Entropy in the [0-8]Hz range and Kurtosis in the [8-16]Hz range). Interestingly, the mutant seizures seem to form a separate cluster in all 3 comparisons, which is also visible from the histograms of the bottom panels. Table 6.11 displays the distribution (IQR) of the top 3 ranked predictors.

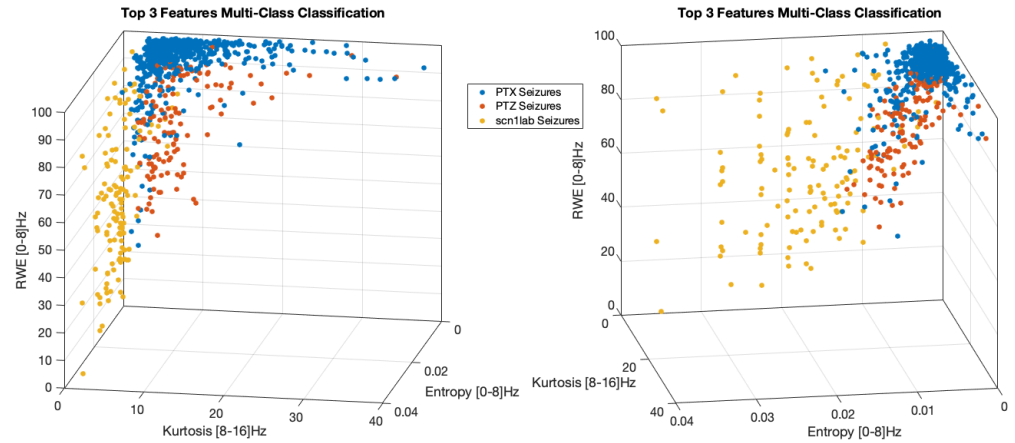

Fig. 6.17 **3D Scatter Plot of Feature Importance for the Three Classes Classification Task (PTX - PTZ - scn1lab).** The top 3 ranked features are displayed (Energy and Entropy in the [0-8]Hz range and Kurtosis in the [8-16]Hz range).

6.5 Discussion

The automatic monitoring of seizures events in animal models is essential for the identification of epileptic phenotypes and the evaluation of the effects of AED [252]. In this chapter, a highly efficient automatic detection algorithm based on a novel mathematical framework was proposed to extract seizures of scn1lab mutants, PTX and PTZ-treated Zebrafish. To the best of our knowledge, similar approaches to automatically extract seizures events from Zebrafish LFP or EEG signals were only attempted in [56, 253]. For the former, a cross-correlation index between multichannel electrodes was computed to detect events of neuronal activity that differ from baseline, which did not allow precise extraction of seizure events. For the latter, an approach based on non-interpretable machine learning tools was proposed. It uses a number of statistical and frequency measures computed from LFP signals of a genetic mutant (scn1lab) and a chemically-induced model of seizures (PTZ), which then makes their detection approach directly comparable with our study. As a result, the performances of the approach proposed in [56] did not reached high-level trade-offs standards between specificity and sensitivity towards seizure events in scn1lab mutants. Indeed, for the genetic mutant model, the sensitivity was of 70.8%, the specificity of 99.8% and the precision of 78.0%, which is significantly lower than the performance of our algorithm: 97% sensitivity, 99.5% specificity and precision 97%. On the contrary, the performances of both algorithms on the automatic detection of PTZ-induced seizures were more comparable with: 60.6% sensitivity, 99.7% specificity and 94.9% precision for their algorithm and 90.2% sensitivity, 95.4% specificity and 79.3% precision for the approach proposed in this thesis. While the direct comparison of the performances of the proposed method with its counterpart in humans or mice is not straightforward, we note that such levels of accuracy are at the level of the most recent and performant algorithms for automatic seizure extraction [48]. In addition, a notable advantage of our approach is that the length of seizures does not need to be specified a priori, thereby allowing greater flexibility. Finally, our knowledge-based approach allowed us to identify the importance of each feature in the decision system, which is typically not feasible with most of the automatic detection algorithms [48].

In particular, we note that the top 2 most important features for the automatic detection of seizures events in scn1lab Zebrafish involve the entropy of the signal at very high frequencies ([250-1000]Hz) (Figures 6.4 and 6.5). Detailed analysis of the distribution of this dynamical measure at different scales of the signal revealed that during seizures events, the entropy is significantly lower than for the interictal activity, which is consistent with the synchronized behavior of neurons during seizures (Table 6.2). Indeed, neurons synchronization is reflected by a more organized, less complex and more predictable

signal in LFP data. Furthermore, the presence of high frequency oscillations has previously been associated with seizures events in humans [69, 243, 244, 245, 246, 247, 248, 249, 250]. Similarly, we report that PTZ-induced seizures can be further distinguished from the normal brain activity with high frequency signal ([125-500]Hz) (Figures 6.7 and 6.8, Table 6.4), which has never been suggested so far. Altogether, those results support that high-to-very-high frequency signals carry most potential for accurate seizure detection, which is a specificity of seizure events across humans, mice and Zebrafish models.

The extracted seizure patterns (923) were then subsequently compared. As a result, we propose that specific neuronal dynamics *during* seizures events retain some of the information that characterizes the pathological mechanisms underlying the disease. This proposal is in line with the recent idea of an universal mathematical framework to analyze seizure patterns from the amplitude and frequency scaling of oscillations during the pre-ictal state [10], in the sense it suggests that dynamical measures hold the potential to further reveal specific brain functional machinery.

We note that the patterns of inherited and induced seizures in Zebrafish can be discriminated with a very high degree of accuracy (perfect accuracy for the data used in this thesis). For this task, the entropy computed at both very high ([500-1000]Hz) and low ([0-8]Hz) frequencies appear as important mathematical biomarkers (Figures 6.9 and 6.10). The distribution of the entropy values at the very high frequencies was significantly lower for scn1lab seizures than for drug-induced seizures, and vice versa for the low frequencies (Table 6.6). The fact that drugs and mutants seizure patterns can be discriminated from very high frequency oscillations ([500-1000]Hz) is consistent with the results of the automatic seizures extraction algorithm. Interestingly, the significant difference lies in the distribution of the entropy at the lowest frequency band ([0-8]Hz - delta and theta waves), which might represent a biomarker of drug induced-seizures. Indeed, the comparison of PTX and PTZ-induced seizure patterns revealed that most of the discriminative potential lies in the analysis of the low frequency bands, with most of the energy of PTX seizures being distributed over the [0-8]Hz frequency band (IQR: 95.35 [91.91 - 97.92] %). It is worth noticing that the morphological patterns of PTX and PTZ-induced seizures could not be perfectly separated (illustrated by a smaller precision of the classification algorithm), which might either be explained by shared biological mechanisms or because PTZ treated Zebrafish significantly displayed more irregular patterns. As a final note, the signal of PTZ-induced seizures is more predictable, that is, with more neuronal synchronization in the [31-62]Hz frequency band (gamma waves) (Figures 6.12 and 6.13, Table 6.8).

Then, a three-class classification task was proposed to distinguish between the pathological causes of seizures, which constitutes a more clinically relevant setup. The results show that seizures patterns originating from different Zebrafish models can be reliably differentiated (Figure 6.14), which hold further potential for the precise diagnosis and personalized therapy of the epileptic condition in humans. Notably, dynamical measures based on the entropy of the signal accounted for most of the feature discriminative potential, as compared to the statistical and energy measures. Finally, the fact that the kurtosis of the signal was generally chosen as a feature of high importance is of further interest.

6.6 Strengths and Limitations of the Study

The study presented in this chapter has the following strengths and limitations:

- Strengths
  - A large number of events was considered to formulate a robust statistical model of seizures dynamical patterns originating from distinct pathological causes of seizures.
  - While the inter-variability of seizures patterns is higher than their intra-variability, it did not affect the classification algorithm. We suggest that the set of mathematical biomarkers proposed in this chapter are robust metrics.
  - The use of both performant and interpretable machine learning tools allowed to classify seizures in a knowledge-oriented approach, which contrasts with most of the available literature on seizure detection.
  - In particular, the wavelet decomposition of the signal at different resolutions allowed to simultaneously gain in model accuracy and formulate hypothesis on the specific functioning of the Zebrafish brain at those scales.
- Limitations
  - Only one dynamical metric could be computed due to the brief duration of seizure events. Indeed, the Fractal Dimension and the Hurst Exponent did not provide consistent results for the short time-series data considered (during seizures events across recordings of mutants, PTX and PTZ treated Zebrafish). Alternatively, other nonlinear metrics such as the permutation entropy [254] or the multiscale entropy [255] could be further considered to improve the performances of the algorithms proposed.

– Translation of findings from animal models to human applications is not straightforward. Indeed, although the electrical patterns generated during epileptic events share fundamental characteristics with those in humans, the direct applicability of the currently available tools and methods developed for characterizing seizure dynamics in Zebrafish or rodents is not yet perfectly clear. The investigation of the possible cross-application of the tools proposed in this thesis and in the literature has the potential to reveal fundamental resemblances or dissimilarities in the epileptic mechanism between species, which is of great interest for further evaluation of a variety of anti-epileptic treatments or therapies.

Chapter 7

Prediction of Epileptic Seizure Events in Zebrafish

7.1 Contribution

For the first time, the predictability of seizures of both scn1lab and PTZ-treated Zebrafish larvae is investigated from LFP recordings. In particular, changes in the frequency of Interspike Epileptiform Discharges (IED) in the vicinity of seizures events are investigated as a distinguishable feature of the system transition from the healthy state to the epileptic seizure state. The results show that, on average, a reshaping of the distribution of IED occurs in the proximity of a seizure event.

In addition, a probabilistic prediction model is designed to evaluate the potential of seizure predictability from restrospective recordings by integrating multiple dynamical measures. The results show that the neurophysiological signals recorded by LFP in Zebrafish larvea brains (at least) partially retain some of the information about the functional reorganization of the system towards seizure. This finding is illustrated by significantly above-average performances of the prediction algorithm in the last 10 seconds before seizure emergence.

Finally, the feasibility of an automated seizure prediction algorithm from continuous recordings of Zebrafish's brain activity with LFP signals has been analyzed. In a continuous setup, however, robust classification of events between pre-ictal and interictal could not be reliably achieved from LFP signals only, which is due to an overall lack of precision of the predictions. The latter can be attributed to various causes, including the difficulty of formulating the "false positive" event, the lack of distinctive information in LFP signals, or the inherent variability in the electrode placement.

7.2 Introduction

The seemingly unpredictable nature of epileptic seizures may have a particularly debilitating effect on the patient's quality of life, especially for individuals with drug-resistant disease. However, to date, no universal biomarker has been found to consistently forecast the likelihood of seizure events across individuals [256].

In the recent years, the field of computational epilepsy research has generally acknowledged the feasibility of seizure prediction [74, 75, 76]. Notably, in 2013, a large clinical trial showed promises for the anticipation of seizures for some patients using surgically implanted devices [73]. In 2014, a seizure prediction competition was conducted on the Kaggle platform (kaggle.com) by crowdsourcing long-term intracranial recordings of canine brain activity with naturally occurring seizures [257]. As a result, the best ranked algorithms of the competition reached precision levels that were significantly higher than chance predictors. Together, those study provided convincing evidences that seizures are not random events, but rather that a reorganization of the brain activity occurs preceding a seizure [258, 259, 260, 261], which can be inferred by prospective algorithms from the neuronal activity recorded by intracranial electrodes. It is worth noticing, however, that the performances of the algorithms were not equivalent across individuals, so that the suitability of the prediction systems differed across patients. Despite those advances, important improvements are still required in both the sensitivity and specificity of the prediction algorithms for a robust implementation in a clinical setting.

Since then, progress has been made in the understanding of the mechanistic processes involved in the brain transition to seizures. During interictal and pre-ictal periods, the presence of pathologically connected groups of neurons may manifest as transient episodes of synchronous activity. Interspike Epileptiform Discharges (IED) are recurrent and morphologically similar events that are clearly distinguishable from the background LFP activity. Their duration is typically short compared to seizures events. IEDs have been observed in humans, mice and, more recently, in Zebrafish [10]. The role for IED has been a matter of debate within the concerned community. On one hand, studies examining changes in IED properties in humans have found that their frequency can both increase and decrease in advance of a seizure [72], and that this relationship might be subject-specific. In such case, IED may either hold pro-seizure (feedforward mechanisms) or anti-seizure effects (feedback mechanisms) [11, 262, 263]. Conceptually, the anti-seizure effect would be achieved by suppressing the brain neuronal activity, shifting the dynamical state back towards a more resilient state, while the pro-seizure effect would be achieved by lowering the "seizure threshold" and bringing the brain system closer to the

separatrix. On the other hand, it has been proposed that IED and seizure events are not dependent, but merely co-occurrent events [264]. However, this difference might be due to the existence of different underlying cellular mechanisms of epileptogenesis, that is, different dynamical routes to seizure initiation [265]. In addition, the robust quantification of the role of IEDs in seizure emergence is challenging. Indeed, the patterns of such fluctuations are notably variable across individuals and pathologies [72], do not inevitably progress to a seizure [73], and are dominated by the normal baseline activity. Recently, a general framework has been proposed to interpret the emergence of such patterns as the visible manifestation of the slow bifurcation from the healthy state of the brain towards the epileptic state. Based on dynamical bifurcation theory, IED are then interpreted as the subthreshold oscillations that typically occur when a nonlinear system is pushed out of its stable regime in the vicinity of a dynamical bifurcation, so-called Hopf bifurcation. Hence, the occurrence of IED at least corresponds to a reflection of the dynamical state of the brain system, which is analogous to the definition of a pre-ictal state.

The aim of this chapter is threefold. First, investigating the existence of dynamical changes in IED as a distinguishable feature of brain activity near the transition towards seizure events. Second, diverse measures of brain dynamical activity are computed in the vicinity of a seizure event and statistically compared to the baseline activity with machine learning tools. Third, based on the predictions models estimated in the previous step, the relevance of a clinically realistic setup to continuously evaluate the likelihood of a seizure emergence over time is analyzed.

7.3 Sub-threshold Oscillations towards Seizures Events

In our dataset, IEDs have been observed for both scn1lab mutants and PTZ-treated Zebrafish. However, clear and consistent morphological LFP patterns were only identified across scn1lab Zebrafish. On the opposite, they are not visible at all for PTX-treated Zebrafish, which might suggest that PTX brings the brain already in a state very close to the seizure state, where stochastic perturbations are the main trigger of rapid critical bifurcation towards seizures.

The aim of this section is to investigate the changes in subthreshold oscillations occurrence over time and evaluate their potential for characterizing the closeness of the brain system state to a sudden seizure event in scn1lab mutant Zebrafish. Indeed, fluctuations of such morphological biomarkers have the potential to indicate an approaching dynamical bifurcation.

7.3.1 Methods

Based on the detection algorithm developed in Chapter 6, pre-ictal windows of 60 seconds were extracted before seizures onsets for each seizure detected in scn1lab Zebrafish larvae. Seizures that were not at least separated by 60 seconds from the previous seizure onset were discarded. In total, 104 seizures were considered across 33 Zebrafish recordings.

Subthresholds oscillations were identified by first bandpass filtering the pre-ictal windows in the [0.5-16]Hz frequency band using the MODWT and then localizing the peaks occurrence using the build-in *findpeaks* function in Matlab. The threshold for the detection of peaks occurrence has been selected by visual inspection to maximize the detection rate (Figure 7.1). It has been suggested in [10, 11] that subthreshold oscillations are a reflection of the system state toward seizures. Hence, a change in the subthreshold dynamics would be a marker of the brain susceptibility to seizures. To analyze this hypothesis, the preictal windows were separated into sub-windows of 20, 10 and 5 seconds and their occurrence investigated (Figure 7.2). Statistical significances between their distributions over time were analyzed with the Mann-Whitney U-test. As a result, distributions of the closest and farthest distance towards seizures on each subgraph showed statistical significance: p-values of 0.0325, 0.0031 and 0.0028 were observed for the 20, 10 and 5 seconds windows respectively. No statistical significances were observed between the two closest distributions towards seizures onsets. While there is only a relatively small amount of observable IEDs per pre-ictal subwindow, the latter observation might be the result of a slow change in the brain system dynamics followed by a faster transition toward seizures.

The dynamical behavior of subthreshold oscillations were further analyzed by reporting their interspike interval (ISI) over time. Indeed, variations of intervals between subthreshold oscillations are typical hallmarks of an approaching dynamical bifurcation. To vizualize the progressive transition of the brain state dynamics, the distribution of ISI was represented using the 20 seconds subwindow size. The results are represented on Figure 7.3. It can be observed that the occurrence of subthreshold oscillations becomes faster as the seizure is approaching, which is depicted by a smaller ISI average ($T_{ave}(ISI)$) of the 20 seconds window that is the closest from the seizure (in blue).

Morphologically similar events like IEDs are the result of punctual and synchronous interactions between large population of neurons. Hence, these results support that functional changes in the brain system of Zebrafish larvae are occurring at an interpretable scale before a seizure, and therefore that the latter might be predicted from LFP signals. This topic is further investigated in the following section.

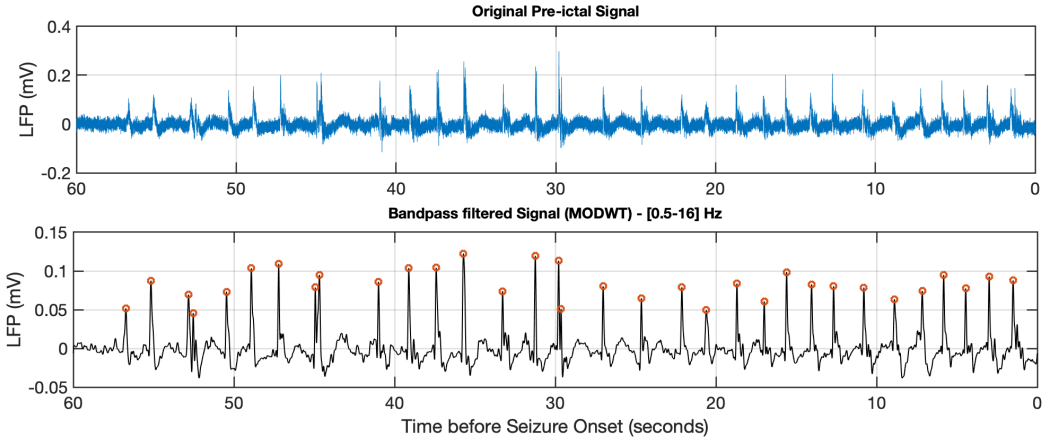

Fig. 7.1 **Illustration of the Automatic Detection Framework for Subthreshold Oscillations.** 104 Pre-ictal windows of 60 seconds were considered before seizure onset. The top panel represents the original signal while the bottom panel represents the bandpass filtered signal ([0.5-16]Hz). Red circles depict the peak automatically detected using the *findpeaks* function in Matlab.

7.4 A Probabilistic Model for Seizure Prediction from LFP

In its core, the mathematical framework behind the development of a prediction model is not different than of a detection model. Instead of differentiating between the presence and the absence of an event, the discrimination is performed between the signal preceding the event and the absence of such event. If such difference exists and a reliable classification can be performed, the model is said to predict the upcoming emergence of this event. Hence, in a machine learning framework, the key difference lies in the data that are used to train and validate the statistical classifiers. Mathematically, reliable predictions of the diseases states of highly dimensional systems such as the brain represent a much harder task than for their mere detection, as early manifestation of diseases are often barely perceptible from clinically observable signals. In addition, this task is hindered by the inherent variability of disease progression across individuals. Robust predictive modeling, therefore, requires a sufficiently large collection of examples to learn the hidden patterns of pre-disease states. This claim is of particular relevance for the identification of pre-ictal patterns from LFP recordings since it represents the average neuronal activity of a large region of the Zebrafish brain.

Typically, a seizure prediction methodology can be addressed by two distinct approaches. In the first approach, pre-ictal and interictal windows are extracted and a

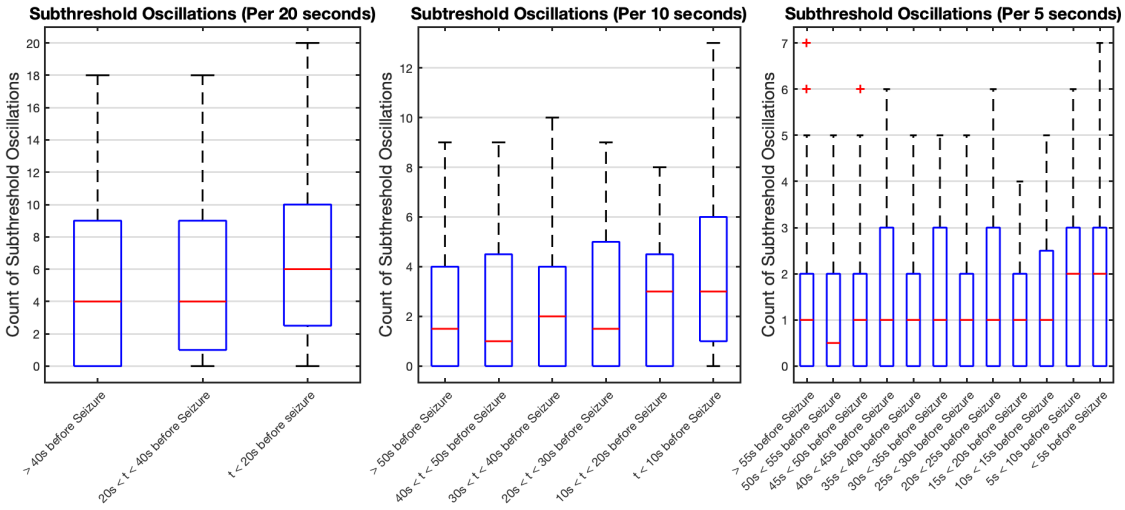

Fig. 7.2 **Counts of Subthreshold Oscillations as a Function of Distance before Seizure Onset.** 104 Pre-ictal windows of 60 seconds were considered before seizure onset. The occurrence of subthreshold oscillations was investigated by dividing the pre-ictal windows into subwindows of 20, 10 and 5 seconds.

model is trained to discriminate the signals originating from different states. In such case, the ictal and post-ictal segments do not contribute to seizure prediction, so those signals are discarded. This strategy is referred to *offline* prediction, since it relies on retrospective signal extraction and differentiation. This strategy can be used to learn the most distinctive characteristics between signals classes and as a proof-of-concept of system's predictability. The second approach consists of the use of a sliding window across the entire recording to continuously classify the signal into either interictal or pre-ictal regions. This approach is referred to as *online* prediction. A particular advantage of the latter strategy is that it represents a clinically relevant setup where the location of seizures events in the signal are not known a priori.

This section addresses both approaches. First, a retrospective analysis of pre-ictal events is performed to investigate the predictability of seizure events in scn1lab mutant and PTZ-treated Zebrafish larvae from LFP signals. Then, the suitability of the prediction models under a continuous monitoring framework is evaluated.

7.4.1 Offline Prediction of Seizures Events

Sequences of 10 seconds before seizure events were extracted from the pre-ictal windows previously considered. Such window size has been selected to maximize the discriminative potential between the interictal and pre-ictal patterns. At this stage, it is worth noticing that if the window size is chosen too short, capturing the slowly varying aspect

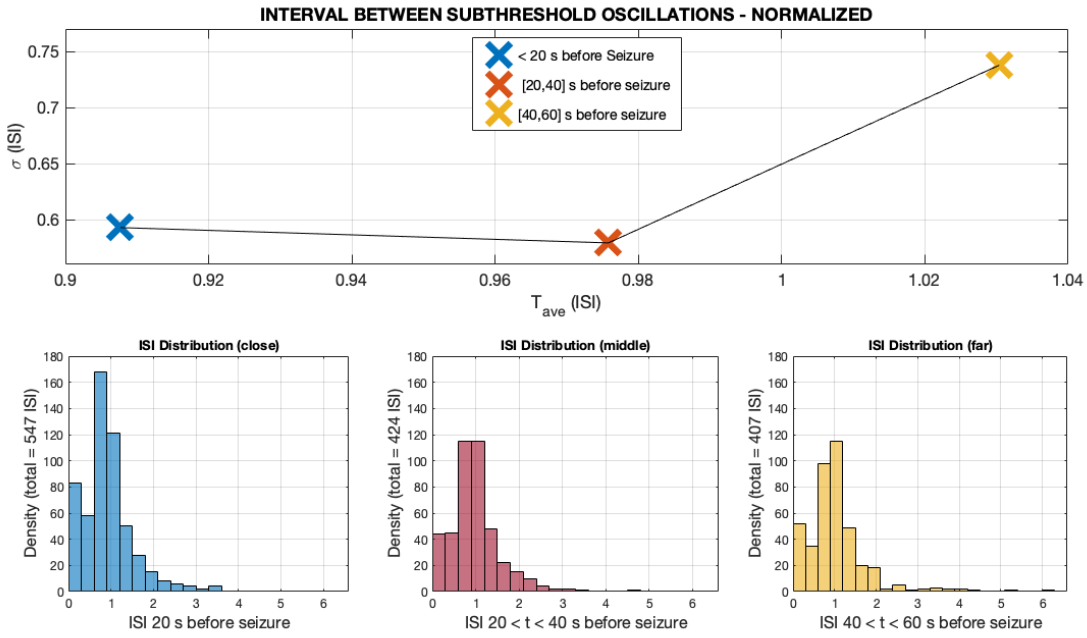

Fig. 7.3 **Evolution of ISI Distributions as a Function of the Distance towards Seizures.** ISI are normalized by the average ISI per pre-ictal window. The top panel represents the average between ISI ($T_{ave}(ISI)$) versus the standard deviation ($\sigma(ISI)$) for the distribution of specific 20 seconds subwindows.

of the system would not be possible, so that prediction of seizure emergence has to be performed using a sufficiently long-time window to reflect the reorganization of the brain system. Then, interictal windows of 10 seconds duration were randomly extracted from the recordings under two conditions:

- The interictal window must be distant by at least 60 seconds before a seizure onset as detected with the automatic detection algorithm developed in Chapter 6.
- The interictal window must be separated by at least 5 seconds of another seizure.

In particular, a larger amount of interictal events have been randomly extracted to account for the considerable variability of interictal patterns. The data are separated into three groups. Two groups are used to train and validate the prediction model. The last group consists of previously unseen recordings from another set of Zebrafish individuals, hence constituting the external validation group. Such separation intends at evaluating the effects of intra-inter variability between Zebrafish individuals on the performances of the classification algorithm.

The Random Forest algorithm was used to generate a predictive model due to both its high performance and robustness to noise inherited from its bagging scheme. For

scn1lab Zebrafish, pre-ictal windows that did not contained IED were discarded from the training set, which ultimately showed better performance on both test groups. Such result further supports that IED are representative of the system state towards seizures. For scn1lab Zebrafish, 28 pre-ictal windows and 134 interictal windows were used to train the model, 17 pre-ictal windows and 91 interictal windows were used to test the model on the same population of Zebrafish and 32 pre-ictal windows and 160 interictal windows were used to externally validate the model on unseen recordings. An external validation group was not formed for PTZ-treated Zebrafish due to the smaller number of recordings (11). Hence, the corresponding training and test group consisted in 86/407 pre-ictal/interictal windows and 51/278 pre-ictal/interictal windows, respectively. The following features (52) were computed to capture the signature of pre-ictal events:

- The Relative Wavelet Energy (RWE) of each frequency band (8th-level decomposition with db4 wavelets), totalling 10 features.
- The statistical moments (variance, skewness and kurtosis) of each frequency band, totalling 30 features.
- The Entropy of the original signal and of the wavelets coefficients of each frequency bands, totalling 9 features.
- The Hurst Exponent (HE) and Fractal Dimension (FD) computed from the raw signal, to reflect system's memory.
- The count of IED computed from the bandpass filtered window considered, due to its increase before seizure onset (Figures 7.2 and 7.3).

The objective function of the RF has been modified to account for the class imbalance, so that a class-sensitive cost function was optimized to raise the penalty resulting from missing a pre-ictal event. Feature selection was considered with RFE, but did not improve the performance of the prediction models, which is due to the fact that feature importance was uniformly spread across the 52 features. The following hyperparameters were optimized using a grid search and 3 fold cross-validation to obtain the best possible generalization trade-off: number of trees in the forest (or estimators), maximum depth of the decision trees, minimum number of samples to make a split and minimum number of samples to be defined as a leaf. The best classification results were obtained for 30/120 estimators (resp. scn1lab/PTZ), a maximum depth of 10 splits, 2 samples minimum to make a split and one single sample minimum to be defined as a leaf. Probability calibration was performed on hold-out data to reflect the observed probability of event occurrence.

The ROC and Precision-Recall curves are displayed on Figure 7.4 for scn1lab and 7.5 for PTZ-treated Zebrafish. As a result, the prediction algorithms were capable of differentiating between pre-ictal and interictal events to a moderate degree in both cases. On one hand, the scn1lab prediction model showed an AUROC value of 0.84 and an area under the Precision-Recall curve of 0.48 for the internal validation group. For the external validation group, the AUROC value was of 0.74 and the area under the Precision-Recall curve of 0.33. At the optimal decision point, the model showed the following classification performances: 42.8% *precision*, 82.4% *specificity* and 70.5% *sensitivity* for the internal validation group and 35.7% *precision*, 77.5% *specificity* and 62.5% *sensitivity* for the external validation group. The corresponding confusion matrices are displayed on Table 7.1. The PTZ prediction model, on the other hand, showed both a lower AUROC value of 0.70 and lower an area under the Precision-Recall curve of 0.28. At the optimal decision point, the model showed the following classification performances: 24.2% *precision*, 64% *specificity* and 62.7% *sensitivity*. The corresponding confusion matrix is displayed on Table 7.2.

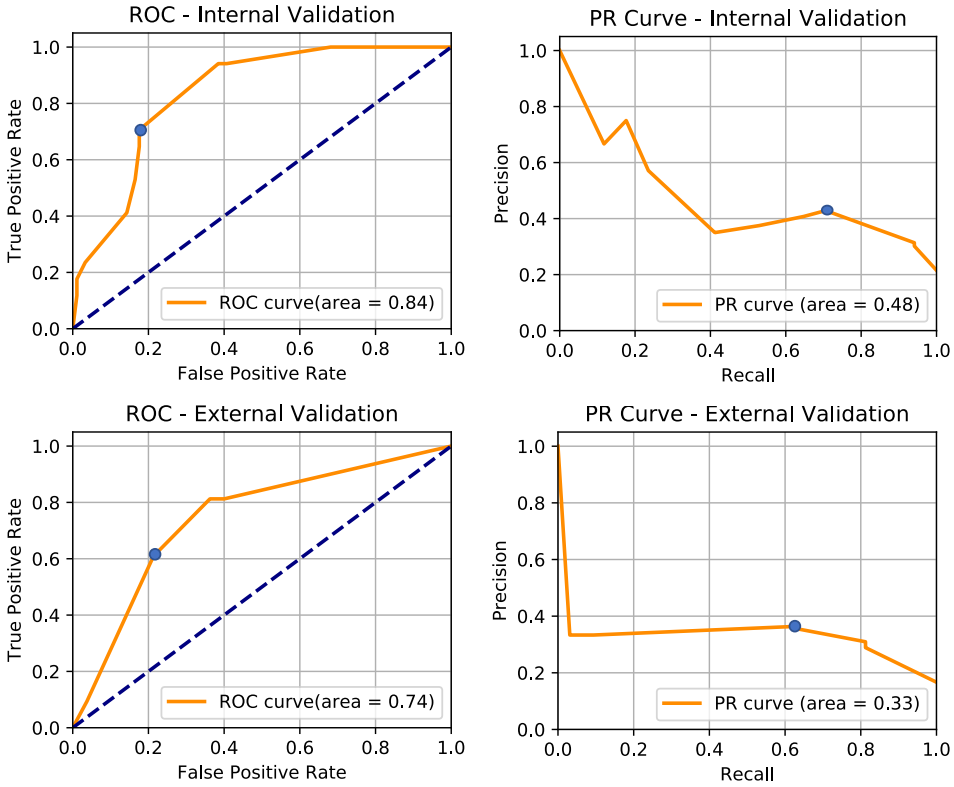

Fig. 7.4 **Receiver Operating Characteristic (left) and Precision-Recall Curve (right) for the Prediction of scn1lab Seizures from Retrospectively Extracted Pre-ictal Windows of 10 Seconds Duration.** The optimal decision thresholds are represented by a blue dot on each graph.

		Predicted Condition	
		Pre-ictal	Interictal
Actual Condition	Pre-ictal	12	5
	Interictal	16	75

	Predicted Condition	
	Pre-ictal	Interictal
Pre-ictal	20	12
Interictal	36	124

Table 7.1 **Confusion Matrices for the Prediction of Seizure Events from scn1lab Zebrafish.** On the left-hand side, the result of the internal validation, which consists of a set of pre-ictal segments extracted from the same recordings as the training set. Those results correspond to a *precision* of 42.8%, a *specificity* of 82.4% and a *sensitivity* of 70.5%. On the right-hand side, the results for the external validation group, which consists of seizures extracted from previously completely unseen recordings. Those results correspond to a *precision* of 35.7%, a *specificity* of 77.5% and a *sensitivity* of 62.5%.

While not being sufficiently high for a robust discrimination of pre-ictal segments, the aforementioned classification performances are consistently higher than random guesses (16.6% precision), suggesting that a change in brain system dynamics is indeed occurring at least in the latest 10 seconds before spontaneous seizure emergence in both scn1lab and PTZ Zebrafish larvae. The low precision and area under the Precision-Recall curve, however, might indicate that such brain reorganization is not specific to the seizure state. Nevertheless, it is worth noticing that the concept of false positive is hard to formulate in the context of seizure prediction. Indeed, while the algorithm might be sensitive enough to identify characteristic patterns of brain system dynamics before a seizure, such dynamical state might not ultimately lead to seizures events. In other swords, the state of the brain is brought close from the "seizure threshold" but the inherent stochastic noise might not be sufficient enough to make it cross the separatrix, hence reverting the system back to a more healthy state over time. In such case, the algorithm would ideally identify the brain system as being in the vicinity of a seizure event, but this outcome would be considered as a false positive since no seizure is observed in the following 60 seconds. Furthermore, it is possible that the prediction algorithm appear to be incorrectly predicting seizures while in fact correctly identifying smaller epileptic events [72].

7.4.2 Online Prediction of Seizures Events

In this section, a clinically relevant setting is tested to automatically predict the emergence of seizure event over time using the previously trained probabilistic models. For this purpose, a sliding window of 10 seconds is passed through the time series data and used to sequentially estimate the probability of the segment to belong to the pre-ictal state or to the interictal state of the brain. Several mathematical setups were considered:

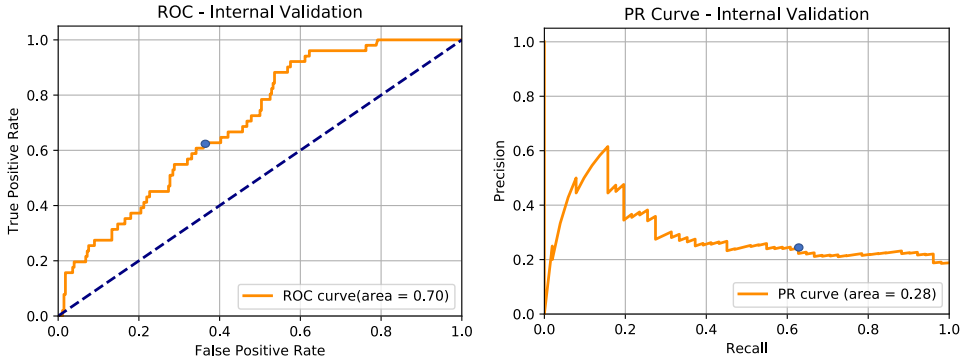

Fig. 7.5 **Receiver Operating Characteristic (left) and Precision-Recall Curve (right) for the Prediction of PTZ Seizures from Retrospectively Extracted Pre-ictal Windows of 10 Seconds Duration.** The optimal decision thresholds are represented by a blue dot on each graph.

		Predicted Condition	
		Pre-ictal	Interictal
Actual Condition	Pre-ictal	32	19
	Interictal	100	178

Table 7.2 **Confusion Matrix for the Prediction of Seizure Events from PTZ-treated Zebrafish.** Those results correspond to a *precision* of 24.2%, a *specificity* of 64% and a *sensitivity* of 62.7%.

- The sliding window has been passed through the data both with and without overlapping (80% overlap was considered, that is, increments of 2 seconds per 10 seconds window).
- Various operating points (i.e. probability thresholds) were evaluated to maximize the overall accuracy of predictions.

For each window, an alarm is raised whenever the output of the predictive model is classified as pre-ictal. An alarm was considered as a false prediction if it was separated by at least by 60 seconds from a seizure event. The functioning of the prediction algorithm is illustrated on Figure 7.6. Performances were reported in terms of false positive rate, sensitivity and specificity per Zebrafish recording on both the internal and external validation sets (Figure 7.7).

For scn1lab mutants, the best performances were obtained for the following decision points on the ROC curves: 94.1% *sensitivity* and 59.3% *specificity* for the internal validation and 81.2% *sensitivity* and 63.7% *specificity* for the external validation set. On

average, the prediction model produced 0.94 false positive per minute for the internal validation group and 0.32 false positive per minute for the external validation group, which is significantly lower than for random classifiers (2.14 FP/min and 2.25 FP/min respectively).

For PTZ-treated Zebrafish, the best performances were obtained for the following decision point: 45.1% *sensitivity* and 77.3% *specificity*. On average, the prediction model produced 1.32 false positive per minute, as compared to 1.81 FP/min for a random classifier. The results for each recording are displayed on Figure 7.8.

It is worth mentioning that, allowing the algorithm to trigger an alarm for a threshold different than the one defined by the optimal operating point (represented in blue on Figures 7.4 and 7.5) permitted an overall increase of the online performances. Finally, while not drastically affecting the outcomes of the algorithm, the performances were maximized without the use of an overlapping sliding window. Nevertheless, the trade-off between the sensitivity and specificity of the models appears to be highly subject-specific and significantly higher when the model has been partially trained on the same individual (corresponding to the internal validation group), as illustrated by Figure 7.7, so that further tuning of a predictive algorithm might be required to improve the suitability of the algorithm to each individual.

As a conclusion, when formulated as an online prediction approach, the model is lacking overall precision towards pre-ictal events. This result is consistent with the ROC curves previously evaluated on retrospectively extracted pre-ictal segments. Indeed, in a continuous approach, the data segments that are seen by the algorithm are not perfectly aligned with the future emergence of seizure, hence decreasing its overall accuracy.

7.5 Discussion

In this chapter, it has been shown that the occurrence of IED is increasing in the vicinity of a seizure event for scn1lab mutants Zebrafish (Figure 7.3). In particular, the distributions of ISI at the 3 distances considered closely resemble the results of a time-dependent Poisson process for which the mean distance between events decreases over time. The reshaping of this distribution is further interesting in such that a decrease of the standard deviation can be observed towards seizure, suggesting that the stochastic component of IED emergence is reduced. In line with the results of [10], we suggest that IED carry information on the likelihood of seizure emergence, despite the involved stochasticity. It is important to notice that the relationship between the time towards seizure and the

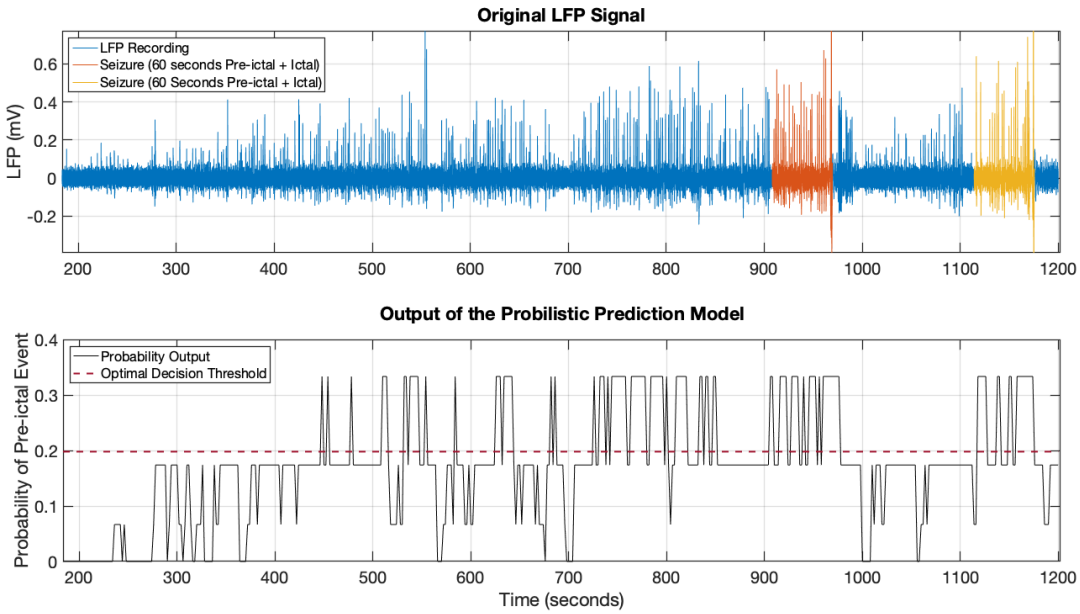

Fig. 7.6 **Example of the behaviour of the automatic prediction algorithm implemented with a sliding window across the entire Zebrafish LFP recording.** The top panel depicts the LFP recordings and the seizures events with 60 seconds of pre-ictal window. The bottom panel represents the algorithm output probabilities when provided with the last 10 seconds of LFP signal. The optimal decision threshold to maximize the accuracy of the online prediction algorithm is represented by a red dotted line. Above this threshold, LFP windows were considered as pre-ictal.

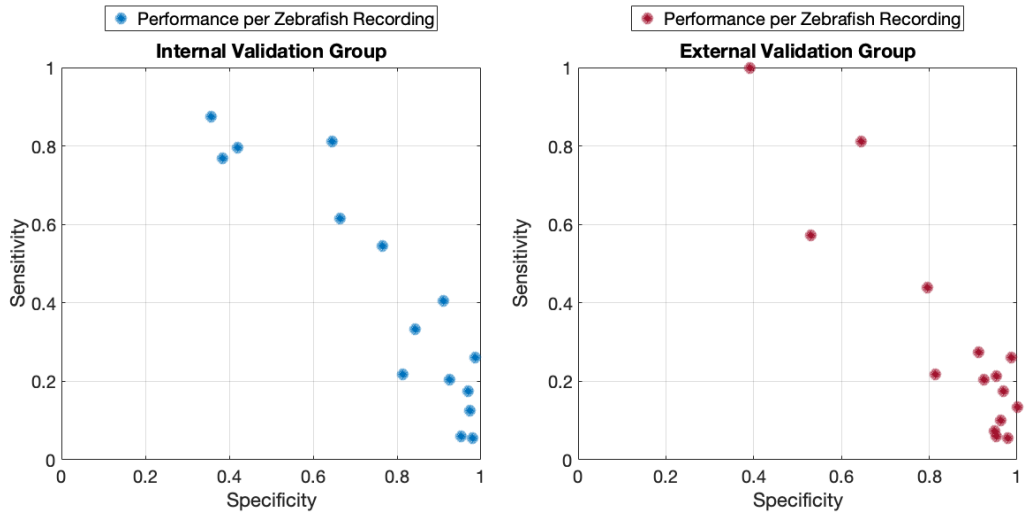

Fig. 7.7 **Performances of the prediction model applied to the continuous monitoring of scn1lab Zebrafish LFP.** Performances are displayed as a dot that represents the trade-off between specificity and sensitivity for each Zebrafish of each validation group. A random prediction algorithm typically produces results around (0.5,0.5). Ideal classification performances are located on the top right corner of the graph.

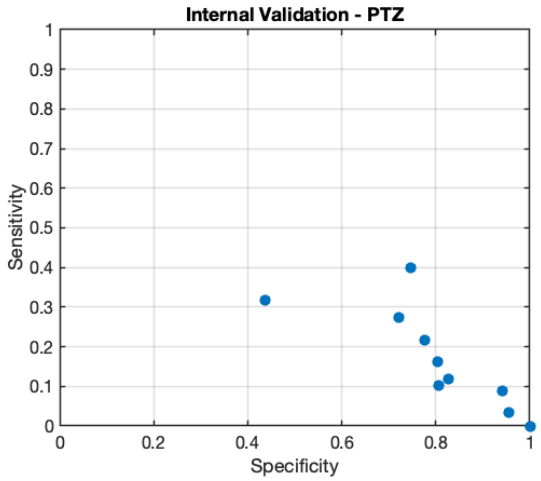

Fig. 7.8 **Performances of the prediction model applied to the continuous monitoring of PTZ-treated Zebrafish LFP.** Performances are displayed as a dot that represents the trade-off between specificity and sensitivity for each Zebrafish. A random prediction algorithm typically produces results around (0.5,0.5). Ideal classification performances are located on the top right corner of the graph.

distribution of ISI did not, however, consistently apply for all seizures and Zebrafish considered, but rather results from the average behavior of IED across 104 seizures.

Those results further permitted the formulation of a pre-ictal window that was subsequentially used to design a prediction algorithm. Indeed, a common mistake in the anticipation of a critical dynamical transition is to look at changes in dynamical markers in the close vicinity of the bifurcation point. However, in the case of scn1lab epileptic seizures, signs of an emerging transition were identified from LFP signals at a scale of (at least) 10 seconds before a seizure event.

Finally, the study demonstrated the feasability of seizure prediction in scn1lab mutant and PTZ-treated Zebrafish by designing prediction algorithms for which the performances lie above those of random predictors. However, the overall lack of precision of the algorithms, even in reproducible experimental conditions (provided by the use of an animal model such as Zebrafish larvae), highlights the challenge underlying the anticipation of seizure events. Indeed, large variabilities in the performances of the prediction algorithm were found across individuals. As such, there remain significant theoretical and technological hurdles in developing a clinically efficient continuous prediction system.

As a further note, it is worth mentioning that hand-designed features might not be the best predictors of seizure likelihood. For instance, deep learning techniques

approaches have shown a median gain in accuracy of 5.4% across all relevant studies [58]. In addition, given sufficiently long recordings of individuals, adaptive or personalized strategies might greatly benefit to the performances of the prediction algorithms. For instance, [73, 266] used the very first seizures of patients to calibrate model parameters or select more informative electrodes. From a general point of view, however, single LFP traces in Zebrafish are unlikely to provide a full understanding of seizure genesis given the spatiotemporal nature of the problem. Nevertheless, the results suggests that LFP signals contain information on the current brain system state, so that it should be at least used in combination with other recording mechanisms, such as calcium imaging. To the best of our knowledge, combining physiological information with recordings of neurophysiological activity has never been attempted for seizure prediction.

7.6 Strengths and Limitations of the Study

The study presented in this chapter has the following strenghts and limitations:

- Strengths
  - The study confirms the presence of changing IED dynamics in another animal model of epilepsy.
  - The analysis presented in this chapter is the first to study seizure predictability from LFP signals in Zebrafish.
- Limitations
  - While PTZ-treated Zebrafish exhibited IED in the LFP signal, their extraction was particularly challenging due to patterns variability. The development of further signal processing tools might be required to address IED occurrence in such model.

Chapter 8

Conclusion

In this thesis, multiple tools and models were developed to describe and predict the dynamical properties of biological systems from time series data. Biological systems are inherently complex, stochastic and their observation is limited. Therefore, generating meaningful biomedical knowledge from large amounts of data required careful calibration of model complexity and design of proper validation frameworks. For this purpose, this thesis integrated technical background originating from various fields such as system identification, systems control, nonlinear dynamical systems theory, time series analysis, statistics, signal processing and machine learning.

Two problems were addressed: the inference of dynamical relationships between genes from short time series data and the characterization of dynamically differentiable brain states in Zebrafish from rich time series data (originally sampled at 100kHz). In both cases, the aim was to provide a reliable and interpretable modeling framework that suited the informative potential of each dataset, which involved a thorough biological and computational literature review, as well as a close collaboration with biologists.

This chapter briefly summarizes the main findings of each chapter and provides potential future directions for research.

8.1 Gene Regulatory Networks

8.1.1 Main Findings

In **chapter 2**, a novel modeling framework based on LTI systems and control theory tools was proposed to simultaneously reverse-engineer the structure of GRNs and provide comparable dynamical description of their underlying mechanisms. We showed that our

approach is particularly suitable for short time series data, so that reliable predictions can be formulated in practical cases. The performances of our approach were compared to state-of-the-art network inferences methodologies under various experimental conditions. For this purpose, realistic in silico data were generated from biologically relevant GRN models by means of stochastic differential equations. This approach led to two main findings, which were subsequently presented in [95, 96]. First, we showed that, for the sole purpose of recovering the structure of GRN from time series data, the performances of the introduced linear modeling approach compete with, but even outperform, most of the latest algorithms in the field. Second, the study unveiled general effects of data quantity and system perturbations on the accuracy of the GRN reconstruction. In particular, for rhythmic systems such as the circadian clock, sampling the dynamical behavior of the system in its transition from one dynamical state to another has the potential to significantly improve the accuracy of the reconstruction. Furthermore, we reported that currently available network inference algorithms do not benefit equally from data increments, thereby revealing pragmatic considerations for experimental designs. Hence, for rhythmic systems, it is generally more profitable for network inference strategies to be run on long time series rather than short time series with multiple perturbations. By contrast, for the non-rhythmic systems, increasing the number of perturbations yielded better results than increasing the sampling frequency.

In **chapter 3**, the mechanisms of action of nicotinamide, a metabolite that lengthens the period of circadian rhythms, were investigated as a mean to understand the regulation of circadian period in Arabidopsis Thaliana [79]. From a mathematical point of view, this study involved the development of a prediction model to distinguish between rhythmic and non-rhythmic gene expression data based on a hand-designed skewed sinusoidal function and logistic regression. Then, the Dynamical Differential Expression (DyDE) modeling framework was used to identify the sources of subtle dynamical changes in the circadian network of Arabidopsis Thaliana, which were subsequently validated. In particular, the methodology was capable of recovering most of the dynamical structure of the circadian network from a single experiment of 48 hours with a sampling rate of one data point per 4 hours. From a biological perspective, the study further provided novel knowledge on the dynamical effects of nicotinamide and the role of blue light in the circadian oscillator.

In **chapter 4**, the network inference methodology was adapted to reconstruct the main regulatory interactions that shape the circadian network of Barley, using relatively few information. Based on the performances of the algorithm on in silico data, few links were identified with a high degree of confidence. The flexibility of the methodology was

further exploited to integrate light patterns and further characterize the contribution of light on the regulation of the whole transcriptome.

8.1.2 Future Perspectives

The network inference algorithms considered as a comparison to our approach carried fundamentally different mathematical assumptions. Further investigation of the individual model predictions revealed that the dynamical relationships inferred by each algorithm were different. As such, future research may want to consider the potential synergy between algorithms to improve the overall accuracy of the network reconstruction step. Finally, it is worth mentioning that the tools developed in this thesis aimed a reverse-engineering a GRN of interest without prior knowledge of the network. While this is a necessary condition for the study of the main gene regulatory interactions that shape a novel complex system of interest, developing a methodology that explicitly allows to integrate prior knowledge might be required in the future. Finally, single-cell sequencing technologies provide information on the per-cell variability of gene expression, which has great potential for many applications of GRN inference. For this purpose, a more probabilistic-oriented formulation of gene expression will be necessary to take advantage of those data.

8.2 Epileptic Seizure and Epileptogenesis Characterization

8.2.1 Main Findings

Chapter 6 and 7 addressed the characterization and prediction of brain states in Zebrafish from LFP signals. The study involved the development of 1) an automatic seizure detection algorithm from LFP signals, 2) two classification models to differentiate between seizures generated by distinct pathological mechanisms, 3) a multi-class prediction model to evaluate the discriminative potential of the morphological signatures of seizures using heterogenic data, 4) a pattern recognition algorithm to investigate the occurrence of interictal spikes in the signal and 5) an online seizure prediction model. In particular, we showed that the automatic extraction of seizures from LFP recordings of Zebrafish can be performed with a very high degree of accuracy. Furthermore, as a knowledge-based approach, the most important dynamical aspects of the signal and their contribution to the classification performances were subsequently analyzed for the models 1), 2) and 3). The latter revealed that seizures induced by drug or genetic variants models in Zebrafish can be reliably discriminated using the dynamical measures proposed in this thesis. Finally, for

the first time, we showed that the neurophysiological information captured by LFP signals in PTZ-treated and scn1lab Zebrafish (at least) partially retains some of the information about the functional reorganization of the brain system towards seizure. The approach, however, highlighted inherent technical and theoretical hurdles in seizure prediction.

8.2.2 Future Perspectives

In Zebrafish, calcium imaging of the neuronal activity across regions of the brain is expected to provide further information on the spatiotemporal reorganization of brain activity in the vicinity of a seizure event. Furthermore, calcium imaging can capture heart rhythms activity, which has the potential to serve as a biomarker for diagnostic and therapeutic purposes in humans. Indeed, in humans, epileptic seizures events have been associated with co-occurring cardiovascular dysfunctions [267]. Notably, patients with a long-lasting or a more severe epileptic condition seem to be more prone to chronic disruptions of cardiovascular functions. In this regard, the integration of additional physiological signals such as body temperature or respiratory rate have the potential to enhance the performances of seizure anticipation algorithms [76]. External factors, such as the time of the day or month can also be exploited to improve the accuracy of predictive algorithms, as patient's circadian profile has been shown to influence seizure emergence [268].

References

[1] Peter Csermely et al. *Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review*. 2013. DOI: 10.1016/j.pharmthera.2013.01.016.

[2] Antonio del Sol et al. *Diseases as network perturbations*. 2010. DOI: 10.1016/j.copbio.2010.07.010.

[3] Thomas Panier et al. "Fast functional imaging of multiple brain regions in intact zebrafish larvae using Selective Plane Illumination Microscopy". In: *Frontiers in Neural Circuits* (2013). ISSN: 1662-5110. DOI: 10.3389/fncir.2013.00065.

[4] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. *Single-cell sequencing-based technologies will revolutionize whole-organism science*. 2013. DOI: 10.1038/nrg3542.

[5] Kun Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. *Artificial intelligence in healthcare*. 2018. DOI: 10.1038/s41551-018-0305-z.

[6] Mintu P. Turakhia et al. "Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study". In: *American Heart Journal* (2019). ISSN: 10976744. DOI: 10.1016/j.ahj.2018.09.002.

[7] Isabelle Guyon et al. "Gene selection for cancer classification using support vector machines". In: *Machine Learning* (2002). ISSN: 08856125. DOI: 10.1023/A:1012487302797.

[8] Konstantina Kourou et al. *Machine learning applications in cancer prognosis and prediction*. 2015. DOI: 10.1016/j.csbj.2014.11.005.

[9] G. A. Hortopan, M. T. Dinday, and S. C. Baraban. "Zebrafish as a model for studying genetic aspects of epilepsy". In: *Disease Models & Mechanisms* (2010). ISSN: 1754-8403. DOI: 10.1242/dmm.002139.

[10] Viktor K. Jirsa et al. "On the nature of seizure dynamics". In: *Brain* (2014). ISSN: 14602156. DOI: 10.1093/brain/awu133.

[11] Wei Chih Chang et al. "Loss of neuronal network resilience precedes seizures and determines the ictogenic nature of interictal synaptic perturbations". In: *Nature Neuroscience* (2018). ISSN: 15461726. DOI: 10.1038/s41593-018-0278-y.

[12] Marten Scheffer et al. *Catastrophic shifts in ecosystems*. 2001. DOI: 10.1038/35098000.

[13] Annelies J. Veraart et al. "Recovery rates reflect distance to a tipping point in a living system". In: *Nature* (2012). ISSN: 00280836. DOI: 10.1038/nature10723.

[14] Andrew G. Haldane and Robert M. May. *Systemic risk in banking ecosystems*. 2011. DOI: 10.1038/nature09659.

[15] Tim R. Mercer, Marcel E. Dinger, and John S. Mattick. *Long non-coding RNAs: Insights into functions*. 2009. DOI: 10.1038/nrg2521.

[16] Tong Ihn Lee and Richard A. Young. *Transcriptional regulation and its misregulation in disease*. 2013. DOI: 10.1016/j.cell.2013.02.014.

[17] Zhong Wang, Mark Gerstein, and Michael Snyder. *RNA-Seq: A revolutionary tool for transcriptomics*. 2009. DOI: 10.1038/nrg2484.

[18] Patrik D'Haeseleer, Shoudan Liang, and Roland Somogyi. "Genetic network inference: From co-expression clustering to reverse engineering". In: *Bioinformatics* (2000). ISSN: 13674803. DOI: 10.1093/bioinformatics/16.8.707.

[19] Michael Hecker et al. "Gene regulatory network inference: Data integration in dynamic models-A review". In: *BioSystems* (2009). ISSN: 03032647. DOI: 10.1016/j.biosystems.2008.12.004.

[20] Florian Markowetz and Rainer Spang. *Inferring cellular networks - A review*. 2007. DOI: 10.1186/1471-2105-8-S6-S5.

[21] Dov Greenbaum et al. *Comparing protein abundance and mRNA expression levels on a genomic scale*. 2003. DOI: 10.1186/gb-2003-4-9-117.

[22] Hiroaki Kitano. *Foundations of systems biology*. The MIT Press Cambridge, Massachusetts London, England, 2001.

[23] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. *Studying and modelling dynamic biological processes using time-series gene expression data*. 2012. DOI: 10.1038/nrg3244.

[24] Guy Karlebach and Ron Shamir. "Modelling and analysis of gene regulatory networks. TL - 9". In: *Nature reviews. Molecular cell biology* (2008). ISSN: 1471-0072. DOI: 10.1038/nrm2503.

[25] S. Mangan and U. Alon. "Structure and function of the feed-forward loop network motif". In: *Proceedings of the National Academy of Sciences* (2003). ISSN: 0027-8424. DOI: 10.1073/pnas.2133841100.

[26] Daniel Marbach et al. "Wisdom of crowds for robust gene network inference". In: *Nature Methods* (2012). ISSN: 15487091. DOI: 10.1038/nmeth.2016.

[27] Andrej Aderhold, Dirk Husmeier, and Marco Grzegorczyk. "Approximate Bayesian inference in semi-mechanistic models". In: *Statistics and Computing* (2017). ISSN: 15731375. DOI: 10.1007/s11222-016-9668-8.

[28] Sipko van Dam et al. "Gene co-expression analysis for functional classification and gene-disease predictions". In: *Briefings in bioinformatics* (2018). ISSN: 14774054. DOI: 10.1093/bib/bbw139.

[29] Robert D. Leclerc. "Survival of the sparsest: Robust gene networks are parsimonious". In: *Molecular Systems Biology* (2008). ISSN: 17444292. DOI: 10.1038/msb.2008.52.

[30] Richard Bonneau et al. "The inferelator: An algorithn for learning parsimonious regulatory networks from systems-biology data sets de novo". In: *Genome Biology* (2006). ISSN: 1474760X. DOI: 10.1186/gb-2006-7-5-r36.

[31] Alexandra Pokhilko et al. "Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model". In: *Molecular Systems Biology* (2010). ISSN: 17444292. DOI: 10.1038/msb.2010.69.

[32] Polly Yingshan Hsu and Stacey L. Harmer. *Wheels within wheels: The plant circadian system*. 2014. DOI: 10.1016/j.tplants.2013.11.007.

[33] Karl Fogelmark and Carl Troein. "Rethinking Transcriptional Activation in the Arabidopsis Circadian Clock". In: *PLoS Computational Biology* (2014). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1003705.

[34] Chiara Campoli et al. "Functional characterisation of HvCO1, the barley (Hordeum vulgare) flowering time ortholog of CONSTANS". In: *Plant Journal* (2012). ISSN: 09607412. DOI: 10.1111/j.1365-313X.2011.04839.x.

[35] Claire Bendix, Carine M. Marshall, and Frank G. Harmon. *Circadian Clock Genes Universally Control Key Agricultural Traits*. 2015. DOI: 10.1016/j.molp.2015.03.003.

[36] Maya E. Kotas and Ruslan Medzhitov. *Homeostasis, Inflammation, and Disease Susceptibility*. 2015. DOI: 10.1016/j.cell.2015.02.010.

[37] Orrin Devinsky et al. "Epilepsy". In: *Nature Reviews Disease Primers* 4 (May 2018), 18024 EP –. URL: https://doi.org/10.1038/nrdp.2018.24.

[38] Robert S. Fisher et al. "ILAE Official Report: A practical clinical definition of epilepsy". In: *Epilepsia* (2014). ISSN: 15281167. DOI: 10.1111/epi.12550.

[39] Miriam H. Meisler et al. "Identification of Epilepsy Genes in Human and Mouse". In: *Annual Review of Genetics* (2001). ISSN: 0066-4197. DOI: 10.1146/annurev.genet.35.102401.091142.

[40] David M. Treiman. "GABAergic mechanisms in epilepsy". In: *Epilepsia*. 2001. DOI: 10.1046/j.1528-1157.2001.042Suppl.3008.x.

[41] Fernando Lopes da Silva et al. "Epilepsies as Dynamical Diseases of Brain Systems: Basic Models of the Transition Between Normal and Epileptic Activity". In: *Epilepsia* (2003). ISSN: 0013-9580. DOI: 10.1111/j.0013-9580.2003.12005.x.

[42] Michael Breakspear. *Dynamic models of large-scale brain activity*. 2017. DOI: 10.1038/nn.4497.

[43] F. Freyer et al. "Bistability and Non-Gaussian Fluctuations in Spontaneous Cortical Activity". In: *Journal of Neuroscience* (2009). ISSN: 0270-6474. DOI: 10.1523/jneurosci.0754-09.2009.

[44] Jerome Engel et al. "Epilepsy biomarkers". In: *Epilepsia* (2013). ISSN: 00139580. DOI: 10.1111/epi.12299.

[45] M. Breakspear et al. "A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis". In: *Cerebral Cortex* (2006). ISSN: 10473211. DOI: 10.1093/cercor/bhj072.

[46] Fernando H. Lopes da Silva et al. "Dynamical Diseases of Brain Systems: Different Routes to Epileptic Seizures". In: *IEEE Transactions on Biomedical Engineering* (2003). ISSN: 15582531. DOI: 10.1109/TBME.2003.810703.

[47] Soheyl Noachtar and Astrid S. Peters. *Semiology of epileptic seizures: A critical review*. 2009. DOI: 10.1016/j.yebeh.2009.02.029.

[48] U. Rajendra Acharya et al. "Automated EEG analysis of epilepsy: A review". In: *Knowledge-Based Systems* (2013). ISSN: 09507051. DOI: 10.1016/j.knosys.2013.02.014.

[49] U. Rajendra Acharya, Yuki Hagiwara, and Hojjat Adeli. *Automated seizure prediction*. 2018. DOI: 10.1016/j.yebeh.2018.09.030.

[50] S. Gigola et al. "Prediction of epileptic seizures using accumulated energy in a multiresolution framework". In: *Journal of Neuroscience Methods* (2004). ISSN: 01650270. DOI: 10.1016/j.jneumeth.2004.03.016.

[51] Ludmyla Kandratavicius et al. *Animal models of epilepsy: Use and limitations*. 2014. DOI: 10.2147/NDT.S50371.

[52] Dean R. Freestone, Philippa J. Karoly, and Mark J. Cook. *A forward-looking review of seizure prediction*. 2017. DOI: 10.1097/WCO.0000000000000429.

[53] A. M. Stewart et al. *Molecular psychiatry of zebrafish*. 2015. DOI: 10.1038/mp.2014.128.

[54] E. P. Rico et al. "Zebrafish neurotransmitter systems as potential pharmacological and toxicological targets". In: *Neurotoxicology and Teratology* (2011). ISSN: 08920362. DOI: 10.1016/j.ntt.2011.07.007.

[55] Maxime O. Baud et al. "Multi-day rhythms modulate seizure risk in epilepsy". In: *Nature Communications* (2018). ISSN: 20411723. DOI: 10.1038/s41467-017-02577-y.

[56] Borbála Hunyadi et al. "Automated analysis of brain activity for seizure detection in zebrafish models of epilepsy". In: *Journal of Neuroscience Methods* (2017). ISSN: 1872678X. DOI: 10.1016/j.jneumeth.2017.05.024.

[57] Nima Bigdely-Shamlo et al. "The PREP pipeline: standardized preprocessing for large-scale EEG analysis". In: *Frontiers in Neuroinformatics* (2015). ISSN: 1662-5196. DOI: 10.3389/fninf.2015.00016.

[58] Yannick Roy et al. "Deep learning-based electroencephalography analysis: a systematic review". In: *Journal of Neural Engineering* (2019). ISSN: 1741-2560. DOI: 10.1088/1741-2552/ab260c.

[59] Michal Teplan et al. "Fundamentals of EEG measurement". In: *Measurement science review* 2.2 (2002), pp. 1–11.

[60] OLIVER FAUST and MURALIDHAR G. BAIRY. "NONLINEAR ANALYSIS OF PHYSIOLOGICAL SIGNALS: A REVIEW". In: *Journal of Mechanics in Medicine and Biology* (2012). ISSN: 0219-5194. DOI: 10.1142/s0219519412400155.

[61] Jianbo Gao, Jing Hu, and Wen Wen Tung. "Complexity measures of brain wave dynamics". In: *Cognitive Neurodynamics* (2011). ISSN: 18714080. DOI: 10.1007/s11571-011-9151-3.

[62] Piero Perucca, François Dubeau, and Jean Gotman. "Intracranial electroencephalographic seizure-onset patterns: Effect of underlying pathology". In: *Brain* (2014). ISSN: 14602156. DOI: 10.1093/brain/awt299.

[63] Kevin Staley. "Molecular mechanisms of epilepsy". In: *Nature Neuroscience* (2015). ISSN: 15461726. DOI: 10.1038/nn.3947.

[64] Steven H. Strogatz and Ronald F. Fox. "Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering". In: *Physics Today* (1995). ISSN: 0031-9228. DOI: 10.1063/1.2807947.

[65] Peter Ashwin et al. “Tipping points in open systems: Bifurcation, noise-induced and rate-dependent examples in the climate system”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2012). ISSN: 1364503X. DOI: 10.1098/rsta.2011.0306.

[66] M. A. Kramer et al. “Human seizures self-terminate across spatial scales via a critical transition”. In: *Proceedings of the National Academy of Sciences* (2012). ISSN: 0027-8424. DOI: 10.1073/pnas.1210047110.

[67] Sergio Rinaldi and Marten Scheffer. “Geometric analysis of ecological models with slow and fast processes”. In: *Ecosystems* (2000). ISSN: 14329840. DOI: 10.1007/s100210000045.

[68] Christian Kuehn. “A mathematical framework for critical transitions: Bifurcations, fastslow systems and stochastic dynamics”. In: *Physica D: Nonlinear Phenomena* (2011). ISSN: 01672789. DOI: 10.1016/j.physd.2011.02.012.

[69] Pariya Salami et al. “Dynamics of interictal spikes and high-frequency oscillations during epileptogenesis in temporal lobe epilepsy”. In: *Neurobiology of Disease* (2014). ISSN: 1095953X. DOI: 10.1016/j.nbd.2014.03.012.

[70] Premysl Jiruska et al. *Synchronization and desynchronization in epilepsy: Controversies and hypotheses*. 2013. DOI: 10.1113/jphysiol.2012.239590.

[71] J. A. White, T. Budde, and A. R. Kay. “A bifurcation analysis of neuronal subthreshold oscillations”. In: *Biophysical Journal* (1995). ISSN: 00063495. DOI: 10.1016/S0006-3495(95)79995-7.

[72] Philippa J. Karoly et al. “Interictal spikes and epileptic seizures: Their relationship and underlying rhythmicity”. In: *Brain* (2016). ISSN: 14602156. DOI: 10.1093/brain/aww019.

[73] Mark J. Cook et al. “Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: A first-in-man study”. In: *The Lancet Neurology* (2013). ISSN: 14744422. DOI: 10.1016/S1474-4422(13)70075-9.

[74] Isabell Kiral-Kornek et al. “Epileptic Seizure Prediction Using Big Data and Deep Learning: Toward a Mobile System”. In: *EBioMedicine* (2018). ISSN: 23523964. DOI: 10.1016/j.ebiom.2017.11.032.

[75] William C. Stacey. *Seizure Prediction Is Possible?Now Let’s Make It Practical*. 2018. DOI: 10.1016/j.ebiom.2018.01.006.

[76] Levin Kuhlmann et al. *Seizure prediction ? ready for a new era*. 2018. DOI: 10.1038/s41582-018-0055-2.

[77] Dan Jones. “Pathways to cancer therapy”. In: *Nature Reviews Drug Discovery* (2008). ISSN: 14741776. DOI: 10.1038/nrd2748.

[78] Albert Pujol et al. *Unveiling the role of network and systems biology in drug discovery*. 2010. DOI: 10.1016/j.tips.2009.11.006.

[79] Laurent Mombaerts et al. “Dynamical differential expression (DyDE) reveals the period control mechanisms of the Arabidopsis circadian oscillator”. In: *PLoS Computational Biology* (2019). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006674.

[80] Gabriele Lillacci and Mustafa Khammash. "Parameter estimation and model selection in computational biology". In: *PLoS Computational Biology* (2010). ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000696.

[81] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks". In: *Frontiers in Cell and Developmental Biology* (2014). ISSN: 2296-634X. DOI: 10.3389/fcell.2014.00038.

[82] N. Dalchau et al. "Correct biological timing in Arabidopsis requires multiple light-signaling pathways". In: *Proceedings of the National Academy of Sciences* (2010). ISSN: 0027-8424. DOI: 10.1073/pnas.1001429107.

[83] Vân Anh Huynh-Thu and Pierre Geurts. "DynGENIE3: Dynamical GENIE3 for the inference of gene networks from time series expression data". In: *Scientific Reports* (2018). ISSN: 20452322. DOI: 10.1038/s41598-018-21715-0.

[84] Huanfei Ma, Kazuyuki Aihara, and Luonan Chen. "Detecting causality from nonlinear dynamics with short-term time series". In: *Scientific Reports* (2014). ISSN: 20452322. DOI: 10.1038/srep07464.

[85] Neil Dalchau. *Understanding biological timing using mechanistic and black-box models*. 2012. DOI: 10.1111/j.1469-8137.2011.04004.x.

[86] Daniel Trejo Banos, Andrew J. Millar, and Guido Sanguinetti. "A Bayesian approach for structure learning in oscillating regulatory networks". In: *Bioinformatics* (2015). ISSN: 14602059. DOI: 10.1093/bioinformatics/btv414.

[87] N. Dalchau et al. "The circadian oscillator gene GIGANTEA mediates a long-term response of the Arabidopsis thaliana circadian clock to sucrose". In: *Proceedings of the National Academy of Sciences* (2011). ISSN: 0027-8424. DOI: 10.1073/pnas.1015452108.

[88] Eva Herrero et al. " EARLY FLOWERING4 Recruitment of EARLY FLOWERING3 in the Nucleus Sustains the Arabidopsis Circadian Clock ". In: *The Plant Cell* (2012). ISSN: 1040-4651. DOI: 10.1105/tpc.111.093807.

[89] Ljung Lennart. "System identification: theory for the user". In: *PTR Prentice Hall, Upper Saddle River, NJ* (1999), pp. 1–14.

[90] Riet De Smet and Kathleen Marchal. *Advantages and limitations of current network inference methods*. 2010. DOI: 10.1038/nrmicro2419.

[91] P. Perez-Garcia et al. *Mapping the Core of the Arabidopsis Circadian Clock Defines the Network Structure of the Oscillator*. 2012. DOI: 10.1126/science.1219075.

[92] G. Vinnicombe. "A ν-gap distance for uncertain and nonlinear systems". In: 2003. DOI: 10.1109/cdc.1999.831313.

[93] Alberto Carignano et al. "Assessing the effect of unknown widespread perturbations in complex systems using the ν-gap". In: *Proceedings of the IEEE Conference on Decision and Control*. 2015. ISBN: 9781479978861. DOI: 10.1109/CDC.2015.7402698.

[94] Kemin Zhou and John Comstock Doyle. *Essentials of robust control*. Vol. 104. Prentice hall Upper Saddle River, NJ, 1998.

[95] Laurent Mombaerts, Alexandre Mauroy, and Jorge Gonçalves. "Optimising time-series experimental design for modelling of circadian rhythms: the value of transient data". In: *IFAC-PapersOnLine* (2016). ISSN: 24058963. DOI: 10.1016/j.ifacol.2016.12.111.

[96] Laurent Mombaerts et al. "A multifactorial evaluation framework for gene regulatory network reconstruction". In: *arXiv preprint arXiv:1906.12243* (2019).

[97] Samiul Haque et al. *Computational prediction of gene regulatory networks in plant growth and development*. 2019. DOI: 10.1016/j.pbi.2018.10.005.

[98] Emre Sefer, Michael Kleyman, and Ziv Bar-Joseph. "Tradeoffs between Dense and Replicate Sampling Strategies for High-Throughput Time Series Experiments". In: *Cell Systems* (2016). ISSN: 24054720. DOI: 10.1016/j.cels.2016.06.007.

[99] Florian Geier, Jens Timmer, and Christian Fleck. "Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge". In: *BMC Systems Biology* (2007). ISSN: 17520509. DOI: 10.1186/1752-0509-1-11.

[100] Johan Markdahl et al. "Experimental design trade-offs for gene regulatory network inference: An in silico study of the yeast Saccharomyces cerevisiae cell cycle". In: *2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017*. 2018. ISBN: 9781509028733. DOI: 10.1109/CDC.2017.8263701.

[101] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* (1996). DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[102] Joseph J Muldoon et al. "Network inference performance complexity: a consequence of topological, experimental and algorithmic determinants". In: *Bioinformatics* (2019). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz105.

[103] D. Marbach et al. "Revealing strengths and weaknesses of methods for gene network inference". In: *Proceedings of the National Academy of Sciences* (2010). ISSN: 0027-8424. DOI: 10.1073/pnas.0913357107.

[104] Vijender Chaitankar et al. "Time lagged information theoretic approaches to the reverse engineering of gene regulatory networks". In: *BMC Bioinformatics* (2010). ISSN: 14712105. DOI: 10.1186/1471-2105-11-19.

[105] Thomas Schaffter, Daniel Marbach, and Dario Floreano. "GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods". In: *Bioinformatics* (2011). ISSN: 13674803. DOI: 10.1093/bioinformatics/btr373.

[106] E. Ravasz et al. "Hierarchical organization of modularity in metabolic networks". In: *Science* (2002). ISSN: 00368075. DOI: 10.1126/science.1073374.

[107] Shai S. Shen-Orr et al. "Network motifs in the transcriptional regulation network of Escherichia coli". In: *Nature Genetics* (2002). ISSN: 10614036. DOI: 10.1038/ng881.

[108] Daniel T. Gillespie. "Chemical Langevin equation". In: *Journal of Chemical Physics* (2000). ISSN: 00219606. DOI: 10.1063/1.481811.

[109] Maria Luisa Guerriero et al. *Stochastic properties of the plant circadian clock*. 2012. DOI: 10.1098/rsif.2011.0378.

[110] Francesca Finotello and Barbara Di Camillo. “Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis”. In: *Briefings in Functional Genomics* (2015). ISSN: 20412657. DOI: 10.1093/bfgp/elu035.

[111] Andrej Aderhold, Dirk Husmeier, and Marco Grzegorczyk. “Statistical inference of regulatory networks for circadian regulation”. In: *Statistical Applications in Genetics and Molecular Biology* (2014). ISSN: 15446115. DOI: 10.1515/sagmb-2013-0051.

[112] Robert J. Prill et al. “Towards a rigorous assessment of systems biology models: The DREAM3 challenges”. In: *PLoS ONE* (2010). ISSN: 19326203. DOI: 10.1371/journal.pone.0009202.

[113] Daniel Marbach et al. “ Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods ”. In: *Journal of Computational Biology* (2009). ISSN: 1066-5277. DOI: 10.1089/cmb.2008.09tt.

[114] Atte Aalto et al. “Continuous time Gaussian process dynamical models in gene regulatory network inference”. In: *arXiv preprint arXiv:1808.08161* (2018).

[115] Jose Casadiego et al. “Model-free inference of direct network interactions from nonlinear collective dynamics”. In: *Nature Communications* (2017). ISSN: 20411723. DOI: 10.1038/s41467-017-02288-4.

[116] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. “Inferring regulatory networks from expression data using tree-based methods”. In: *PloS one* 5.9 (2010), e12776.

[117] Angshul Majumdar and Rabab K Ward. “Fast group sparse classification”. In: *Canadian Journal of Electrical and Computer Engineering* 34.4 (2009), pp. 136–144.

[118] Piyush B. Madhamshettiwar et al. “Gene regulatory network inference: Evaluation and application to ovarian cancer allows the prioritization of drug targets”. In: *Genome Medicine* (2012). ISSN: 1756994X. DOI: 10.1186/gm340.

[119] S. M.Minhaz Ud-Dean and Rudiyanto Gunawan. “Optimal design of gene knockout experiments for gene regulatory network inference”. In: *Bioinformatics* (2016). ISSN: 14602059. DOI: 10.1093/bioinformatics/btv672.

[120] S J Davis and A J Millar. “Watching the hands of the Arabidopsis biological clock.” In: *Genome biology* (2001). ISSN: 1474-760X.

[121] Shigeru Hanano et al. “Multiple phytohormones influence distinct parameters of the plant circadian clock”. In: *Genes to Cells* (2006). ISSN: 13569597. DOI: 10.1111/j.1365-2443.2006.01026.x.

[122] Michael F. Covington et al. “Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development”. In: *Genome Biology* (2008). ISSN: 14747596. DOI: 10.1186/gb-2008-9-8-r130.

[123] Laura C. Roden and Robert A. Ingle. “Lights, Rhythms, Infection: The Role of Light and the Circadian Clock in Determining the Outcome of Plant?Pathogen Interactions”. In: *The Plant Cell* (2009). ISSN: 1040-4651. DOI: 10.1105/tpc.109.069922.

[124] A. Graf et al. “Circadian control of carbohydrate availability for growth in Arabidopsis plants at night”. In: *Proceedings of the National Academy of Sciences* (2010). ISSN: 0027-8424. DOI: 10.1073/pnas.0914299107.

[125] Nathan C. Rockwell, Yi-Shin Su, and J. Clark Lagarias. "PHYTOCHROME STRUCTURE AND SIGNALING MECHANISMS". In: *Annual Review of Plant Biology* (2006). ISSN: 1543-5008. DOI: 10.1146/annurev.arplant.56.032604.144208.

[126] Qing-Hua Li and Hong-Quan Yang. "Cryptochrome Signaling in Plants?" In: *Photochemistry and Photobiology* (2007). DOI: 10.1562/2006-02-28-ir-826.

[127] Andrés Romanowski et al. "Phytochrome, Carbon Sensing, Metabolism, and Plant Growth Plasticity". In: *Plant Physiology* (2017). ISSN: 0032-0889. DOI: 10.1104/pp.17.01437.

[128] John M. Christie et al. *Plant flavoprotein photoreceptors*. 2015. DOI: 10.1093/pcp/pcu196.

[129] Michael J. Haydon et al. "Photosynthetic entrainment of the Arabidopsis thaliana circadian clock". In: *Nature* (2013). ISSN: 00280836. DOI: 10.1038/nature12603.

[130] Antony N. Dodd et al. "The Arabidopsis circadian clock incorporates a cADPR-based feedback loop". In: *Science* (2007). ISSN: 00368075. DOI: 10.1126/science.1146757.

[131] Gad Asher et al. "Poly(ADP-Ribose) Polymerase 1 Participates in the Phase Entrainment of Circadian Clocks to Feeding". In: *Cell* (2010). ISSN: 00928674. DOI: 10.1016/j.cell.2010.08.016.

[132] J. Malapeira, L. C. Khaitova, and P. Mas. "Ordered changes in histone modifications at the core of the Arabidopsis circadian clock". In: *Proceedings of the National Academy of Sciences* (2012). ISSN: 0027-8424. DOI: 10.1073/pnas.1217022110.

[133] Serge Daan and Jürgen Aschoff. "The entrainment of circadian systems". In: *Circadian Clocks*. Springer, 2001, pp. 7–43.

[134] Till Roenneberg, Serge Daan, and Martha Merrow. "The art of entrainment". In: *Journal of Biological Rhythms* 18.3 (2003), pp. 183–194.

[135] Motohide Seki et al. "Adjustment of the Arabidopsis circadian oscillator by sugar signalling dictates the regulation of starch metabolism". In: *Scientific Reports* (2017). ISSN: 20452322. DOI: 10.1038/s41598-017-08325-y.

[136] Charlotte Soneson and Mauro Delorenzi. "A comparison of methods for differential expression analysis of RNA-seq data". In: *BMC Bioinformatics* (2013). ISSN: 14712105. DOI: 10.1186/1471-2105-14-91.

[137] Michael E. Hughes, John B. Hogenesch, and Karl Kornacker. "JTK-CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets". In: *Journal of Biological Rhythms* (2010). ISSN: 07487304. DOI: 10.1177/0748730410379711.

[138] Martin Straume. "DNA Microarray Time Series Analysis: Automated Statistical Assessment of Circadian Rhythms in Gene Expression Patterning". In: *Methods in Enzymology* (2004). ISSN: 00766879. DOI: 10.1016/S0076-6879(04)83007-6.

[139] Nasser M Nasrabadi. "Pattern recognition and machine learning". In: *Journal of electronic imaging* 16.4 (2007), p. 049901.

[140] Shuqing Cao et al. "AtGRP7 is involved in the regulation of abscisic acid and stress responses in arabidopsis". In: *Cellular and Molecular Biology Letters* (2006). ISSN: 16891392. DOI: 10.2478/s11658-006-0042-2.

[141] Zheng Qing Fu et al. “A type III effector ADP-ribosylates RNA-binding proteins and quells plant immunity”. In: *Nature* (2007). ISSN: 14764687. DOI: 10.1038/nature05737.

[142] Takato Imaizumi et al. “Plant science: FKF1 F-box protein mediates cyclic degradation of a repressor of CONSTANS in Arabidopsis”. In: *Science* (2005). ISSN: 00368075. DOI: 10.1126/science.1110586.

[143] Chang-Quan Wang et al. “The Transcriptional Regulator BBX19 Promotes Hypocotyl Growth by Facilitating COP1-Mediated EARLY FLOWERING3 Degradation in Arabidopsis”. In: *The Plant Cell* (2015). ISSN: 1040-4651. DOI: 10.1105/tpc.15.00044.

[144] Chiung Swey Joanne Chang et al. “LZF1, a HY5-regulated transcriptional factor, functions in Arabidopsis de-etiolation”. In: *Plant Journal* (2008). ISSN: 09607412. DOI: 10.1111/j.1365-313X.2008.03401.x.

[145] Takeshi KUMAGAI et al. “ The Common Function of a Novel Subfamily of B-Box Zinc Finger Proteins with Reference to Circadian-Associated Events in Arabidopsis thaliana ”. In: *Bioscience, Biotechnology, and Biochemistry* (2008). ISSN: 0916-8451. DOI: 10.1271/bbb.80041.

[146] Tanja Göbel et al. “Flavin Adenine Dinucleotide and N5,N10-Methenyltetrahydrofolate are the in planta Cofactors of Arabidopsis thaliana Cryptochrome 3”. In: *Photochemistry and Photobiology*. 2017. DOI: 10.1111/php.12622.

[147] Silvia Gonzali et al. “Identification of sugar-modulated genes and evidence for in vivo sugar sensing in Arabidopsis”. In: *Journal of Plant Research* (2006). ISSN: 09189440. DOI: 10.1007/s10265-005-0251-1.

[148] Georgina M. Rae et al. “DRM1 and DRM2 expression regulation: Potential role of splice variants in response to stress and environmental factors in Arabidopsis”. In: *Molecular Genetics and Genomics* (2014). ISSN: 16174623. DOI: 10.1007/s00438-013-0804-2.

[149] Alexander Frank et al. “Circadian Entrainment in Arabidopsis by the Sugar-Responsive Transcription Factor bZIP63”. In: *Current Biology* (2018). ISSN: 09609822. DOI: 10.1016/j.cub.2018.05.092.

[150] Takayuki Ohara et al. “Gene regulatory network models in response to sugars in the plant circadian system”. In: *Journal of Theoretical Biology* (2018). ISSN: 10958541. DOI: 10.1016/j.jtbi.2018.08.020.

[151] Ignasius Joanito et al. “An incoherent feed-forward loop switches the Arabidopsis clock rapidly between two hysteretic states”. In: *Scientific Reports* (2018). ISSN: 20452322. DOI: 10.1038/s41598-018-32030-z.

[152] María Carmen Martí Ruiz et al. “Circadian oscillations of cytosolic free calcium regulate the Arabidopsis circadian clock”. In: *Nature Plants* (2018). ISSN: 20550278. DOI: 10.1038/s41477-018-0224-8.

[153] Juan Pablo Sánchez, Paula Duque, and Nam Hai Chua. “ABA activates ADPR cyclase and cADPR induces a subset of ABA-responsive genes in Arabidopsis”. In: *Plant Journal* (2004). ISSN: 09607412. DOI: 10.1111/j.1365-313X.2004.02055.x.

[154] Caterina Tricase et al. “Economic analysis of the barley market and related uses”. In: *Grasses as Food and Feed*. IntechOpen, 2018.

[155] Jørgen E Olesen et al. “Impacts and adaptation of European crop production systems to climate change”. In: *European Journal of Agronomy* 34.2 (2011), pp. 96–112.

[156] Kathleen Greenham and C. Robertson McClung. *Integrating circadian dynamics with physiological processes in plants*. 2015. DOI: 10.1038/nrg3976.

[157] Polly Yingshan Hsu, Upendra K. Devisetty, and Stacey L. Harmer. “Accurate timekeeping is controlled by a cycling activator in Arabidopsis”. In: *eLife* (2013). ISSN: 2050084X. DOI: 10.7554/eLife.00473.

[158] Cristiane P.G. Calixto, Robbie Waugh, and John W.S. Brown. “Evolutionary Relationships Among Barley and Arabidopsis Core Circadian Clock and Clock-Associated Genes”. In: *Journal of Molecular Evolution* (2015). ISSN: 00222844. DOI: 10.1007/s00239-015-9665-0.

[159] Naoki Takata et al. “Molecular phylogeny and expression of poplar circadian clock genes, LHY1 and LHY2”. In: *New Phytologist* (2009). ISSN: 0028646X. DOI: 10.1111/j.1469-8137.2008.02714.x.

[160] Jelena Kusakina et al. “Barley Hv CIRCADIAN CLOCK ASSOCIATED 1 and Hv PHOTOPERIOD H1 are circadian regulators that can affect circadian rhythms in arabidopsis”. In: *PLoS ONE* (2015). ISSN: 19326203. DOI: 10.1371/journal.pone.0127449.

[161] Naoki Takata et al. “Phylogenetic footprint of the plant clock system in angiosperms: Evolutionary processes of Pseudo-Response Regulators”. In: *BMC Evolutionary Biology* (2010). ISSN: 14712148. DOI: 10.1186/1471-2148-10-126.

[162] Adrian Turner et al. “Botany: The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley”. In: *Science* (2005). ISSN: 00368075. DOI: 10.1126/science.1117619.

[163] James Beales et al. “A Pseudo-Response Regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (Triticum aestivum L.)” In: *Theoretical and Applied Genetics* (2007). ISSN: 00405752. DOI: 10.1007/s00122-007-0603-4.

[164] R. L. Murphy et al. “Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum”. In: *Proceedings of the National Academy of Sciences* (2011). ISSN: 0027-8424. DOI: 10.1073/pnas.1106212108.

[165] Ian M. Ehrenreich et al. “Candidate gene association mapping of Arabidopsis flowering time”. In: *Genetics* (2009). ISSN: 00166731. DOI: 10.1534/genetics.109.105189.

[166] Karen A. Hicks, Tina M. Albertson, and D. Ry Wagner. “EARLY FLOWERING3 Encodes a Novel Protein That Regulates Circadian Clock Function and Flowering in Arabidopsis”. In: *The Plant Cell* (2007). ISSN: 10404651. DOI: 10.2307/3871295.

[167] Elsebeth Kolmos et al. “Integrating ELF4 into the circadian system through combined structural and functional studies”. In: *HFSP Journal* (2009). ISSN: 19552068. DOI: 10.2976/1.3218766.

[168] S. Faure et al. “Mutation at the circadian clock gene EARLY MATURITY 8 adapts domesticated barley (Hordeum vulgare) to short growing seasons”. In: *Proceedings of the National Academy of Sciences* (2012). ISSN: 0027-8424. DOI: 10.1073/pnas.1120496109.

[169] S. Zakhrabekova et al. “Induced mutations in circadian clock regulator Mat-a facilitated short-season adaptation and range extension in cultivated barley”. In: *Proceedings of the National Academy of Sciences* (2012). ISSN: 0027-8424. DOI: 10.1073/pnas.1113009109.

[170] Chiara Campoli et al. “HvLUX1 is a candidate gene underlying the early maturity 10 locus in barley: Phylogeny, diversity, and interactions with the circadian clock and photoperiodic pathways”. In: *New Phytologist* (2013). ISSN: 0028646X. DOI: 10.1111/nph.12346.

[171] L. W. Gallagher, K. M. Soliman, and H. Vivar. “Interactions among Loci Conferring Photoperiod Insensitivity for Heading Time in Spring Barley”. In: *Crop Science* (2010). DOI: 10.2135/cropsci1991.0011183x003100020003x.

[172] Tsuyoshi Mizoguchi et al. “LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in Arabidopsis”. In: *Developmental Cell* (2002). ISSN: 15345807. DOI: 10.1016/S1534-5807(02)00170-3.

[173] Eva M. Farré et al. “Overlapping and distinct roles of PRR7 and PRR9 in the Arabidopsis circadian clock”. In: *Current Biology* (2005). ISSN: 09609822. DOI: 10.1016/j.cub.2004.12.067.

[174] Norihito Nakamichi et al. “ PSEUDO-RESPONSE REGULATORS 9, 7, and 5 Are Transcriptional Repressors in the Arabidopsis Circadian Clock ”. In: *The Plant Cell* (2010). ISSN: 1040-4651. DOI: 10.1105/tpc.109.072892.

[175] Antoine Baudry et al. “ F-Box Proteins FKF1 and LKP2 Act in Concert with ZEITLUPE to Control Arabidopsis Clock Progression ”. In: *The Plant Cell* (2010). ISSN: 1040-4651. DOI: 10.1105/tpc.109.072843.

[176] Paloma Más et al. “Targeted degradation of TOC1 by ZTL modulates circadian function in Arabidopsis thaliana”. In: *Nature* (2003). ISSN: 00280836. DOI: 10.1038/nature02163.

[177] R. Rawat et al. “REVEILLE1, a Myb-like transcription factor, integrates the circadian clock and auxin pathways”. In: *Proceedings of the National Academy of Sciences* (2009). ISSN: 0027-8424. DOI: 10.1073/pnas.0813035106.

[178] C.-Q. Wang et al. “BBX19 Interacts with CONSTANS to Repress FLOWERING LOCUS T Transcription, Defining a Flowering Time Checkpoint in Arabidopsis”. In: *The Plant Cell* (2014). ISSN: 1040-4651. DOI: 10.1105/tpc.114.130252.

[179] Sergei A. Filichkin et al. “Global profiling of rice and poplar transcriptomes highlights key conserved Circadian-controlled pathways and cis-regulatory modules”. In: *PLoS ONE* (2011). ISSN: 19326203. DOI: 10.1371/journal.pone.0016907.

[180] Todd P. Michael et al. “A morning-specific phytohormone gene expression program underlying rhythmic plant growth”. In: *PLoS Biology* (2008). ISSN: 15449173. DOI: 10.1371/journal.pbio.0060225.

[181] Malia A. Gehan et al. *Transcriptional networks-crops, clocks, and abiotic stress*. 2015. DOI: 10.1016/j.pbi.2015.01.004.

[182] Alexander Graf and Alison M. Smith. *Starch and the clock: The dark side of plant productivity*. 2011. DOI: 10.1016/j.tplants.2010.12.003.

[183] Alexander Graf et al. “Parallel analysis of arabidopsis circadian clock mutants reveals different scales of transcriptome and proteome regulation”. In: *Open Biology* (2017). ISSN: 20462441. DOI: 10.1098/rsob.160333.

[184] Artem Pankin et al. “Mapping-by-sequencing identifies HvPHYTOCHROME C as a candidate gene for the early maturity 5 locus modulating the circadian clock and photoperiodic flowering in barley”. In: *Genetics* (2014). ISSN: 19432631. DOI: 10.1534/genetics.114.165613.

[185] Ermias Habte et al. “Osmotic stress at the barley root affects expression of circadian clock genes in the shoot”. In: *Plant, Cell and Environment* (2014). ISSN: 13653040. DOI: 10.1111/pce.12242.

[186] Takeshi Izawa et al. “ Os- GIGANTEA Confers Robust Diurnal Rhythms on the Global Transcriptome of Rice in the Field ”. In: *The Plant Cell* (2011). ISSN: 1040-4651. DOI: 10.1105/tpc.111.083238.

[187] H. G. McWatters et al. “The ELF3 zeitnhmer regulates light signalling to the circadian clock”. In: *Nature* (2000). ISSN: 00280836. DOI: 10.1038/35047079.

[188] B. Thines and F. G. Harmon. “Ambient temperature response establishes ELF3 as a required component of the core Arabidopsis circadian clock”. In: *Proceedings of the National Academy of Sciences* (2010). ISSN: 0027-8424. DOI: 10.1073/pnas.0911006107.

[189] Anne Helfer et al. “LUX ARRHYTHMO encodes a nighttime repressor of circadian gene expression in the Arabidopsis core clock”. In: *Current Biology* (2011). ISSN: 09609822. DOI: 10.1016/j.cub.2010.12.021.

[190] Daphne Ezer et al. “The evening complex coordinates environmental and endogenous signals in Arabidopsis”. In: *Nature Plants* (2017). ISSN: 20550278. DOI: 10.1038/nplants.2017.87.

[191] Dmitri A. Nusinow et al. “The ELF4-ELF3-"LUX complex links the circadian clock to diurnal control of hypocotyl growth”. In: *Nature* (2011). ISSN: 00280836. DOI: 10.1038/nature10182.

[192] Isabelle Carré and Siren R. Veflingstad. *Emerging design principles in the Arabidopsis circadian clock*. 2013. DOI: 10.1016/j.semcdb.2013.03.011.

[193] W. Huang et al. “Mapping the core of the Arabidopsis circadian clock defines the network structure of the oscillator”. In: *Science* (2012). ISSN: 10959203. DOI: 10.1126/science.1219075.

[194] Gang Li et al. “Coordinated transcriptional regulation underlying the circadian clock in Arabidopsis”. In: *Nature Cell Biology* (2011). ISSN: 14657392. DOI: 10.1038/ncb2219.

[195] Prateek Tripathi et al. “Arabidopsis B-BOX32 interacts with CONSTANS-LIKE3 to regulate flowering ”. In: *Proceedings of the National Academy of Sciences* (2017). ISSN: 0027-8424. DOI: 10.1073/pnas.1616459114.

[196] Shuxin Ren et al. “ Regulation of Telomerase in Arabidopsis by BT2 , an Apparent Target of TELOMERASE ACTIVATOR1 ”. In: *The Plant Cell* (2007). ISSN: 1040-4651. DOI: 10.1105/tpc.106.044321.

[197] Søren Bak et al. “The Arabidopsis book”. In: *The Arabidopsis book / American Society of Plant Biologists* (2011). ISSN: 1543-8120. DOI: 10.1199/tab.0144.

[198] Rajendra Bari et al. “PHO2, MicroRNA399, and PHR1 Define a Phosphate-Signaling Pathway in Plants”. In: *Plant Physiology* (2006). ISSN: 0032-0889. DOI: 10.1104/pp.106.079707.

[199] John Lisman. “The challenge of understanding the brain: where we stand in 2015”. In: *Neuron* 86.4 (2015), pp. 864–882.

[200] Dragoljub Gajic et al. “Classification of EEG signals for detection of epileptic seizures based on wavelets and statistical pattern recognition”. In: *Biomedical Engineering: Applications, Basis and Communications* 26.02 (2014), p. 1450021.

[201] Hojjat Adeli, Samanwoy Ghosh-Dastidar, and Nahid Dadmehr. “A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy”. In: *IEEE Transactions on Biomedical Engineering* (2007). ISSN: 00189294. DOI: 10.1109/TBME.2006.886855.

[202] Samanwoy Ghosh-Dastidar, Hojjat Adeli, and Nahid Dadmehr. “Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection”. In: *IEEE Transactions on Biomedical Engineering* (2007). ISSN: 00189294. DOI: 10.1109/TBME.2007.891945.

[203] Yatindra Kumar, M. L. Dewal, and R. S. Anand. “Epileptic seizures detection in EEG using DWT-based ApEn and artificial neural network”. In: *Signal, Image and Video Processing* (2014). ISSN: 18631711. DOI: 10.1007/s11760-012-0362-9.

[204] C Sidney Burrus et al. *Introduction to wavelets and wavelet transforms: a primer*. Vol. 1. Prentice hall New Jersey, 1998.

[205] Chun-Liu Liu. “A Tutorial of the Wavelet Transform”. In: *History* (2010).

[206] Michael R Chernick. “Wavelet Methods for Time Series Analysis”. In: *Technometrics* (2009). ISSN: 0040-1706. DOI: 10.1198/tech.2001.s49.

[207] Mateo Aboy et al. “Characterization of sample entropy in the context of biomedical signal analysis”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*. 2007. ISBN: 1424407885. DOI: 10.1109/IEMBS.2007.4353701.

[208] Elizabeth Bradley and Holger Kantz. “Nonlinear time-series analysis revisited”. In: *Chaos* (2015). ISSN: 10541500. DOI: 10.1063/1.4917289.

[209] Geoff Boeing. “Visual Analysis of Nonlinear Dynamical Systems: Chaos, Fractals, Self-Similarity and the Limits of Prediction”. In: *Systems* (2016). ISSN: 2079-8954. DOI: 10.3390/systems4040037.

[210] Tilmann Gneiting and Martin Schlather. “Stochastic Models That Separate Fractal Dimension and the Hurst Effect”. In: *SIAM Review* (2005). ISSN: 0036-1445. DOI: 10.1137/s0036144501394387.

[211] M. Dämmig and F. Mitschke. “Estimation of Lyapunov exponents from time series: the stochastic case”. In: *Physics Letters A* (1993). ISSN: 03759601. DOI: 10.1016/0375-9601(93)90865-W.

[212] Elif Derya Übeyli. “Lyapunov exponents/probabilistic neural networks for analysis of EEG signals”. In: *Expert Systems with Applications* (2010). ISSN: 09574174. DOI: 10.1016/j.eswa.2009.05.078.

[213] Ying Cheng Lai et al. “Controlled test for predictive power of Lyapunov exponents: Their inability to predict epileptic seizures”. In: *Chaos* (2004). ISSN: 10541500. DOI: 10.1063/1.1777831.

[214] Floris Takens. “Detecting strange attractors in turbulence”. In: 1981. DOI: 10.1007/bfb0091924.

[215] Alan Wolf et al. “Determining Lyapunov exponents from a time series”. In: *Physica D: Nonlinear Phenomena* (1985). ISSN: 01672789. DOI: 10.1016/0167-2789(85)90011-9.

[216] Liangyue Cao. “Practical method for determining the minimum embedding dimension of a scalar time series”. In: *Physica D: Nonlinear Phenomena* (1997). ISSN: 01672789. DOI: 10.1016/S0167-2789(97)00118-8.

[217] M. Palus. “Nonlinearity in normal human EEG: cycles, temporal asymmetry, nonstationarity and randomness, not chaos.” In: *Biological cybernetics* (1996). ISSN: 03401200. DOI: 10.1007/s004220050304.

[218] Jonghwa Kim and Elisabeth André. “Emotion recognition based on physiological changes in music listening”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2008). ISSN: 01628828. DOI: 10.1109/TPAMI.2008.26.

[219] Justin Dauwels, Francois Vialatte, and Andrzej Cichocki. “Diagnosis of Alzheimer’s Disease from EEG Signals: Where Are We Standing?” In: *Current Alzheimer Research* (2010). ISSN: 15672050. DOI: 10.2174/1567210204558652050.

[220] Pedro Fonseca et al. “Sleep stage classification with ECG and respiratory effort”. In: *Physiological Measurement* (2015). ISSN: 13616579. DOI: 10.1088/0967-3334/36/10/2027.

[221] By Andreas Voss et al. *Methods derived from nonlinear dynamics for analysing heart rate variability*. 2009. DOI: 10.1098/rsta.2008.0232.

[222] Joshua S. Richman and J. Randall Moorman. “Physiological time-series analysis using approximate entropy and sample entropy”. In: *American Journal of Physiology-Heart and Circulatory Physiology* (2017). ISSN: 0363-6135. DOI: 10.1152/ajpheart.2000.278.6.h2039.

[223] Luciano Pietronero and Erio Tosatti. *Fractals in physics*. Elsevier, 2012.

[224] Antonio Di Ieva et al. *Fractals in the neurosciences, part II: Clinical applications and future perspectives*. 2015. DOI: 10.1177/1073858413513928.

[225] Gerhard Werner. “Fractals in the nervous system: Conceptual implications for theoretical neuroscience”. In: *Frontiers in Physiology* (2010). ISSN: 1664042X. DOI: 10.3389/fphys.2010.00015.

[226] Tiziana Di Matteo. “Multi-scaling in finance”. In: *Quantitative finance* 7.1 (2007), pp. 21–36.

[227] Tilmann Gneiting, Hana Ševčíková, and Donald B. Percival. “Estimators of Fractal Dimension: Assessing the Roughness of Time Series and Spatial Data”. In: *Statistical Science* (2012). ISSN: 0883-4237. DOI: 10.1214/11-sts370.

[228] C. K. Peng et al. “Mosaic organization of DNA nucleotides”. In: *Physical Review E* (1994). ISSN: 1063651X. DOI: 10.1103/PhysRevE.49.1685.

[229] G. Deco et al. “Key role of coupling, delay, and noise in resting brain fluctuations”. In: *Proceedings of the National Academy of Sciences* (2009). ISSN: 0027-8424. DOI: 10.1073/pnas.0901831106.

[230] Nir Friedman et al. “Universal critical dynamics in high resolution neuronal avalanche data”. In: *Physical Review Letters* (2012). ISSN: 00319007. DOI: 10.1103/PhysRevLett.108.208102.

[231] John M. Beggs and Dietmar Plenz. "Neuronal Avalanches in Neocortical Circuits". In: *The Journal of Neuroscience* (2003). ISSN: 0270-6474. DOI: 10.1523/jneurosci.23-35-11167.2003.

[232] James A. Roberts et al. "Critical role for resource constraints in neural models". In: *Frontiers in Systems Neuroscience* (2014). DOI: 10.3389/fnsys.2014.00154.

[233] Michael Breakspear and Stuart Knock. "Kinetic models of brain activity". In: *Brain Imaging and Behavior* (2008). ISSN: 19317557. DOI: 10.1007/s11682-008-9033-4.

[234] Christophe Trefois et al. *Critical transitions in chronic disease: Transferring concepts from ecology to systems medicine*. 2015. DOI: 10.1016/j.copbio.2014.11.020.

[235] Trevor Hastie et al. "The elements of statistical learning: data mining, inference and prediction". In: *The Mathematical Intelligencer* 27.2 (2005), pp. 83–85.

[236] Yanjun Qi. "Random forest for bioinformatics". In: *Ensemble machine learning*. Springer, 2012, pp. 307–323.

[237] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[238] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[239] Jeffrey Noebels. "Pathway-driven discovery of epilepsy genes". In: *Nature Neuroscience* (2015). ISSN: 15461726. DOI: 10.1038/nn.3933.

[240] S. C. Baraban et al. "Pentylenetetrazole induced changes in zebrafish behavior, neural activity and c-fos expression". In: *Neuroscience* (2005). ISSN: 03064522. DOI: 10.1016/j.neuroscience.2004.11.031.

[241] Alfred L. George. *Inherited disorders of voltage-gated sodium channels*. 2005. DOI: 10.1172/JCI25505.

[242] Scott C. Baraban, Matthew T. Dinday, and Gabriela A. Hortopan. "Drug screening in Scn1a zebrafish mutant identifies clemizole as a potential Dravet syndrome treatment". In: *Nature Communications* (2013). ISSN: 20411723. DOI: 10.1038/ncomms3410.

[243] John G.R. Jefferys et al. *Mechanisms of physiological and epileptic HFO generation*. 2012. DOI: 10.1016/j.pneurobio.2012.02.005.

[244] John G.R. Jefferys et al. "Limbic Network Synchronization and Temporal Lobe Epilepsy". In: *Jasper's Basic Mechanisms of the Epilepsies*. 2013. DOI: 10.1093/med/9780199746545.003.0014.

[245] Maeike Zijlmans et al. "High-frequency oscillations as a new biomarker in epilepsy". In: *Annals of Neurology* (2012). ISSN: 03645134. DOI: 10.1002/ana.22548.

[246] Elena Urrestarazu et al. "Interictal high-frequency oscillations (10-500 Hz) in the intracerebral EEG of epileptic patients". In: *Brain* (2007). ISSN: 00068950. DOI: 10.1093/brain/awm149.

[247] Premysl Jiruska et al. *Electrographic high-frequency activity and epilepsy*. 2010. DOI: 10.1016/j.eplepsyres.2009.11.008.

[248] P. Jiruska et al. "High-Frequency Network Activity, Global Increase in Neuronal Activity, and Synchrony Expansion Precede Epileptic Seizures In Vitro". In: *Journal of Neuroscience* (2010). ISSN: 0270-6474. DOI: 10.1523/jneurosci.0535-10.2010.

[249] Julia Jacobs et al. “Interictal high-frequency oscillations (80-500 Hz) are an indicator of seizure onset areas independent of spikes in the human epileptic brain”. In: *Epilepsia* (2008). ISSN: 00139580. DOI: 10.1111/j.1528-1167.2008.01656.x.

[250] J. Jacobs et al. *High-frequency oscillations (HFOs) in clinical epilepsy*. 2012. DOI: 10.1016/j.pneurobio.2012.03.001.

[251] L. Federico Rossi et al. “Focal cortical seizures start as standing waves and propagate respecting homotopic connectivity”. In: *Nature Communications* (2017). ISSN: 20411723. DOI: 10.1038/s41467-017-00159-6.

[252] Aristea S. Galanopoulou et al. “Epilepsy therapy development: Technical and methodologic issues in studies with animal models”. In: *Epilepsia* (2013). ISSN: 00139580. DOI: 10.1111/epi.12295.

[253] Soon Gweon Hong et al. “A Novel Long-term, Multi-Channel and Non-invasive Electrophysiology Platform for Zebrafish”. In: *Scientific Reports* (2016). ISSN: 20452322. DOI: 10.1038/srep28248.

[254] Christoph Bandt and Bernd Pompe. “Permutation entropy: a natural complexity measure for time series”. In: *Physical review letters* 88.17 (2002), p. 174102.

[255] Madalena Costa, Ary L Goldberger, and C-K Peng. “Multiscale entropy analysis of biological signals”. In: *Physical review E* 71.2 (2005), p. 021906.

[256] Levin Kuhlmann et al. “Epilepsyecosystem.org: Crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG”. In: *Brain* (2018). ISSN: 14602156. DOI: 10.1093/brain/awy210.

[257] Benjamin H. Brinkmann et al. “Crowdsourcing reproducible seizure forecasting in human and canine epilepsy”. In: *Brain* (2016). ISSN: 14602156. DOI: 10.1093/brain/aww045.

[258] Sridhar Sunderam, Ivan Osorio, and Mark G. Frei. “Epileptic seizures are temporally interdependent under certain conditions”. In: *Epilepsy Research* (2007). ISSN: 09201211. DOI: 10.1016/j.eplepsyres.2007.06.013.

[259] Gaoxiang Ouyang et al. “Using recurrence plot for determinism analysis of EEG recordings in genetic absence epilepsy rats”. In: *Clinical Neurophysiology* (2008). ISSN: 13882457. DOI: 10.1016/j.clinph.2008.04.005.

[260] Mark J. Cook et al. “The dynamics of the epileptic brain reveal long memory processes”. In: *Frontiers in Neurology* (2014). ISSN: 16642295. DOI: 10.3389/fneur.2014.00217.

[261] Eulalie Joelle Ngamga et al. “Evaluation of selected recurrence measures in discriminating pre-ictal and inter-ictal periods from epileptic EEG data”. In: *Physics Letters, Section A: General, Atomic and Solid State Physics* (2016). ISSN: 03759601. DOI: 10.1016/j.physleta.2016.02.024.

[262] Massimo Avoli and Marco de Curtis. *GABAergic synchronization in the limbic system and its role in the generation of epileptiform activity*. 2011. DOI: 10.1016/j.pneurobio.2011.07.003.

[263] Gilles Huberfeld et al. “Glutamatergic pre-ictal discharges emerge at the transition to seizure in human epilepsy”. In: *Nature Neuroscience* (2011). ISSN: 10976256. DOI: 10.1038/nn.2790.

[264] J. Gotman and M. G. Marciani. "Electroencephalographic spiking activity, drug levels, and seizure occurence in epileptic patients". In: *Annals of Neurology* (1985). ISSN: 15318249. DOI: 10.1002/ana.410170612.

[265] Massimo Avoli et al. "Specific imbalance of excitatory/inhibitory signaling establishes seizure onset pattern in temporal lobe epilepsy". In: *Journal of neurophysiology* 115.6 (2016), pp. 3229–3237.

[266] Leon D. Iasemidis et al. "Adaptive Epileptic Seizure Prediction System". In: *IEEE Transactions on Biomedical Engineering* (2003). ISSN: 15582531. DOI: 10.1109/TBME.2003.810689.

[267] Orrin Devinsky. "Effects of Seizures on Autonomic and Cardiovascular Function". In: *Epilepsy Currents* (2004). ISSN: 1535-7597. DOI: 10.1111/j.1535-7597.2004.42001.x.

[268] Philippa J. Karoly et al. "The circadian profile of epilepsy improves seizure forecasting". In: *Brain* (2017). ISSN: 14602156. DOI: 10.1093/brain/awx173.