Inference and modeling of biological networks : a statistical-physics approach to neural attractors and protein fitness landscapes

Lorenzo Posani

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Inference and modeling of biological networks

A statistical-physics approach to neural attractors and protein fitness landscapes

Soutenue par
Lorenzo Posani
Le 07/12/2018

Ecole doctorale n° 564
Physique en Île-de-France

Spécialité
Physique

Composition du jury :

Kate Jeffery University College London	*Présidente du jury*
Matteo Marsili ICTP	*Rapporteur*
Andrea Pagnani Politecnico di Torino	*Rapporteur*
Gianluigi Mongillo Université Paris Descartes	*Examinateur*
Simona Cocco École Normale Supérieure	*Directrice de thèse*
Rémi Monasson École Normale Supérieure	*Directeur de thèse*

1794
ENS
ÉCOLE NORMALE
SUPÉRIEURE

ACKNOWLEDGMENTS

We all are products of our environment, and this thesis is no exception. During these three years I had the privilege to work alongside exceptional mentors and peers, whose ideas and suggestions pervade this manuscript on many levels.

I first must express my gratitude to Simona Cocco and Rémi Monasson for taking the commitment of managing my "non-orthodox" working hours both as a master student and as a Ph.D. candidate. The transition from "learning" to "doing" science can be harsh for students, but you both sweetened the deal by providing a brilliant example of how professional scientists think and work, and I will always be grateful for that.

My warm and sincere thanks go to my brothers-in-thesis, Marco Molari and Jerome Tubiana, who did not miss one single occasion to prove their kindness and will to help, equaled only by the sharpness of their mind. You really made me feel like home, which probably explains the unhealthy frequency of dinner at the office together! Special thanks are also due to Diego Contreras, who taught me bike and philosophy (unrelatedly), Kevin Berlemont, Clement Roussel, and Max Puelma Touzel for helping me with French and English, respectively, and Francesca Rizzato with whom I had the pleasure to collaborate on the protein fitness landscape project and whose results are partially included here.

Big thanks are due to all my colleagues and friends for making the ENS (and, more generally, Paris) such a stimulating and enjoyable environment, each with his/her peculiar character and mind. In pseudo-order of appearance: Gaia Tavoni, Alice Coucke, Tommaso Brotto, Tommaso Comparin, Anirudh Kulkarni, Quentin Feltgen, Volker Pernice, Manon Michel, Ze Lei, Tridib Sadhu, Juliane Klamser, Alexis Dubreuil, Andreas Mayer, Louis Brezin, Elisabetta Vesconi, Ivan Amelio, Quentin Marcou, Tran Huy, Jacopo Marchi, Cosimo Lupo, Victor Dagard, Michelangelo Preti, Aldo Battista, Eduarda Susin, Sebastien Wolf, Moshir Agarwal, Ohgod Youaresomany, Daniele Conti, Giulio Isacco, Federica Ferretti, Fabio Manca, Arnaud Fanthomme, Simone Blanco Malerba , Silvia Ferri, and all those who I may have forgotten, who will not take it personally since they are familiar with my total lack of episodic memory.

Finally, I'd like to dedicate a word to my life partner and lifeblood, Federica, who constantly supported me with her piercing intelligence and unconditioned tenderness. You enlightened every single step of this three-years journey, which would have been impossible without you by my side. Words can not express how thankful I am.

ABSTRACT

The recent advent of high-throughput experimental procedures has opened a new era for the quantitative study of biological systems. Today, electrophysiology recordings and calcium imaging allow for the in vivo simultaneous recording of hundreds to thousands of neurons. In parallel, thanks to automated sequencing procedures, the libraries of known functional proteins expanded from thousands to millions in just a few years. This current abundance of biological data opens a new series of challenges for theoreticians. Accurate and transparent analysis methods are needed to process this massive amount of raw data into meaningful observables. Concurrently, the simultaneous observation of a large number of interacting units enables the development and validation of theoretical models aimed at the mechanistic understanding of the collective behavior of biological systems. In this manuscript, we propose an approach to both these challenges based on methods and models from statistical physics.

The first part of this manuscript is dedicated to an introduction to the statistical physics approach to systems biology, with a particular focus on the interfaces between statistical physics, Bayesian inference, and systems biology. The intersections between these fields are presented by following the common thread of the tools and models that have been applied, during the development of this thesis, to the study of two biological systems: the navigation-memory task in the hippocampal complex (part II) and the fitness landscape of co-evolving residues in proteins (part III).

The second part is dedicated to the representation of navigation memory in the hippocampal network. We first introduce a Bayesian population-activity decoder, based on the adaptive cluster expansion (ACE) of the graphical-Ising inference, aimed at retrieving the represented cognitive map on fast time scales. We apply the decoder on CA1 data, showing that it outperforms the current standards in discriminating the recalled cognitive state, and on in-silico data, to investigate the functional meaning of the inferred neural couplings. We then apply this method to the investigation of the *flickering* phenomenology, i.e., the oscillatory behavior of the cognitive map that was observed in a recent experiment where contextual cues are abruptly changed to induce the network instability in the rodent hippocampal region CA3. We present an attractor model, subject to external and path-integrator inputs, which is shown to accurately reproduce the oscillating phenomenology of the cognitive map. By the application of the Ising decoder, we show that a number of novel predictions of the model, concerning the precision of the positional representation during the instability of the cognitive state, can be verified by a careful re-analysis of the original data. Finally, we show that the Ising model inferred from hippocampal recordings can be used to generate population activities that are coherent with low-dimensional attractors, which have been proposed as the neural mechanism underlying spatial navigation in the cognitive map.

In the third part, we employ a statistical-physics model of protein folding, called Lattice Proteins, to benchmark inference methods aimed at the reconstruction of the local fitness landscape of a protein from sequence data of homologous proteins. We first show that a sparse version of the ACE inference, which adapts the sparsity of the inferred interaction graph to the number of available data, yields superior performances than standard DCA methods in the common sub-sampled regime. We then frame the inference task in the context of the bias-variance trade-off, showing that we can optimize its retrieval performance by choosing the proper subset of the training alignment (MSA). We propose a procedure, called "focusing," aimed at finding this optimal subset from a given MSA, opening to applications on real protein datasets.

RESUMÉ

L'avènement récent des procédures expérimentales à haut débit a ouvert une nouvelle ère pour l'étude quantitative des systèmes biologiques. De nos jours, les enregistrements d'électrophysiologie et l'imagerie du calcium permettent l'enregistrement simultané in vivo de centaines à des milliers de neurones. Parallèlement, grâce à des procédures de séquençage automatisées, les bibliothèques de protéines fonctionnelles connues ont été étendues de milliers à des millions en quelques années seulement. L'abondance actuelle de données biologiques ouvre une nouvelle série de défis aux théoriciens. Des méthodes d'analyse précises et transparentes sont nécessaires pour traiter cette quantité massive de données brutes en observables significatifs. Parallèlement, l'observation simultanée d'un grand nombre d'unités en interaction permet de développer et de valider des modèles théoriques visant à la compréhension mécanistique du comportement collectif des systèmes biologiques. Dans ce manuscrit, nous proposons une approche de ces défis basée sur des méthodes et des modèles issus de la physique statistique.

La première partie de ce manuscrit est consacrée à une introduction de l'approche des systèmes biologiques par la physique statistique. Dans cette partie, l'accent est porté sur l'interface entre la physique statistique, l'inférence bayésienne et les systèmes biologiques. Les intersections entre ces domaines sont présentées en suivant le fil conducteur des outils et modèles qui ont été appliqués, lors de cette thèse, à l'étude de deux systèmes biologiques particuliers : la tâche navigation et mémoire spatiale dans le complexe hippocampique (partie II) et le paysage adaptatif de coévolution dans les protéines (partie III).

La deuxième partie est consacrée à la représentation de la mémoire spatiale dans le réseau hippocampique. Nous introduisons d'abord un décodeur Bayésien d'activité de population, basé sur l'expansion adaptative de clusters (ACE) de l'inférence d'un modèle d'Ising sur un graph, visant à récupérer la carte cognitive représentée sur des échelles de temps rapides. Nous appliquons le décodeur sur des données CA1, montrant qu'il surpasse les normes actuelles en matière de discrimination de l'état cognitif rappelé, et sur des données in-silico, pour étudier la signification fonctionnelle des couplages neuronaux inférés. Nous appliquons ensuite cette méthode à l'étude de la phénoménologie du « flickering », c'est-à-dire le comportement oscillatoire de la carte cognitive observé lors d'une expérience récente où les conditions contextuelles sont brusquement modifiées pour induire l'instabilité du réseau dans la région hippocampique CA3 du rongeur. Nous présentons un modèle d'attracteur, soumis à des entrées externes et à des intégrateurs de trajectoire, dont il est démontré qu'il reproduit avec précision la phénoménologie oscillante de la carte cognitive. Par l'application du décodeur précédent, nous montrons qu'un certain nombre de nouvelles prédictions du modèle, concernant la précision de la représentation positionnelle pendant l'instabilité de l'état cognitif, peuvent être vérifiées par une nouvelle analyse des données originales. Enfin, nous montrons que le modèle

d'Ising inféré des enregistrements de l'hippocampe peut être utilisé pour générer des activités de population cohérentes avec les attracteurs de faible dimension, qui ont été proposés comme mécanisme neuronal sous-jacent à la navigation spatiale dans la carte cognitive.

Dans la troisième partie, nous employons un modèle de physique statistique du repliement des protéines, appelé « Lattice Proteins », pour comparer les méthodes d'inférence visant à reconstruire le paysage adaptatif local d'une protéine à partir des données de séquence des protéines homologues. Nous montrons d'abord qu'une version éparse de l'inférence ACE, qui adapte la rareté du graphique d'interaction inféré au nombre de données disponibles, donne des performances supérieures à celles des méthodes DCA standard dans le régime commun sous-échantillonné. Ensuite, dans le contexte du dilemme biais-variance, nous montrons que nous pouvons optimiser le rendement de récupération de notre inférence en choisissant le sous-ensemble approprié de l'alignement des protéines (MSA). Nous proposons une procédure, appelée « focusing », visant à trouver ce sous-ensemble optimal à partir d'un alignement donné. Cette procédure pourrait avoir des applications sur des ensembles de protéines réelles.

PUBLICATIONS

This manuscript comprises the research work that I have conducted during the last three years under the supervision of Simona Cocco and Rémi Monasson at the Laboratoire de Physique Statistique de l'Ecole Normale Superieure, and includes published as well as original results.

For what concerns the hippocampus part, Chapters 5 and 6 have been published as research papers in [1, 2], in collaboration with Karel Ježek from Charles University (Prague). Some of the early results of Chapter 6 were presented at [3]. Part of the ideas and results showed in Chapter 4 have been included in [4, 5]. Chapter 7 is a work in progress at the draft stage [6].

Chapters 9, 10, and 11 present an analysis of inference models applied to the retrieval of the mutational landscape of a protein from sequence data. While these chapters mostly cover results obtained on theoretical models, this work was conducted in parallel with a postdoctoral researcher, Francesca Rizzato, who applied similar analyses to real protein datasets. The results of this collaboration will be jointly published in two future papers, now at the draft stage [7, 8].

Publications

[1] L. Posani, S. Cocco, K. Ježek, and R. Monasson, "Functional connectivity models for decoding of spatial representations from hippocampal ca1 recordings," *Journal of Computational Neuroscience*, pp. 1–17, 2017.

[2] L. Posani, S. Cocco, and R. Monasson, "Integration and multiplexing of positional and contextual information by the hippocampal network," *PLoS computational biology*, vol. 14, no. 8, p. e1006320, 2018.

Review papers, conference proceedings

[3] L. Posani, S. Cocco, K. Ježek, and R. Monasson, "Position is coherently represented during flickering instabilities of place-cell cognitive maps in the hippocampus," *BMC neuroscience - 26th Annual Computational Neuroscience Meeting (CNS* 2017)*, vol.18-1, pp.58, 2017.

[4] S. Cocco, R. Monasson, L. Posani, and G. Tavoni, "Functional networks from inverse modeling of neural population activity," *Current Opinion in Systems Biology*, 2017.

[5] S. Cocco, R. Monasson, L. Posani, S. Rosay, and J. Tubiana, "Statistical physics and representations in real and artificial neural networks," *Physica A: Statistical Mechanics and its Applications*, p. doi/10.1016/j.physa.2017.11.153, 2017.

In preparation

[6] L. Posani, S. Cocco, and R. Monasson, "Pairwise models inferred from hippocampal activity generate neural configurations typical of single or multiple low-dimensional attractors," 2018.

[7] L. Posani†, F. Rizzato†, R. Monasson, and S. Cocco, "Infer global, predict local: Bias-variance trade-off in protein fitness landscape reconstruction from sequence data," 2018 (†: joint first authors)

[8] L. Posani†, F. Rizzato†, R. Monasson, and S. Cocco, "Improved performance of DCA fitness predictions with sparsity prior from structural information," 2018 (†: joint first authors)

CONTENTS

Part I

STATISTICAL PHYSICS, INFERENCE, BIOLOGY: AN INTRIGUING SCIENTIFIC BRAID

1

BACKGROUND

1.1 STATISTICAL PHYSICS

Statistical physics, as the name suggests, is the branch of physics that deals with the *statistical* analysis of physical phenomena. It finds its roots in the kinetic theory of thermodynamics, developed mainly by James Clerk Maxwell and Ludwig Boltzmann in the 19th century, and later significantly contributed to by Willard Gibbs with his seminal work "Elementary principles in statistical mechanics" [9].

In some sense, statistical physics represents an answer to the analytical intractability of many-body systems. As proven by the famous work of Henry Poincaré in 1887 [10], a detailed deterministic description of the motion of three (or more) bodies, interacting through Newton's laws of gravitation, is impossible, since the system follows a chaotic (non-periodic) dynamics in the phase space.

However, some systems, such as gases, crystals, and amorphous solids, display a limited number of stable macroscopic behaviors, even though they are composed of a massive amount of interacting bodies (atoms or molecules). This apparent paradox is elegantly solved by the concept of *statistical equilibrium*: even though the single molecule of air follows a chaotic and fast dynamics, endlessly moving at an average speed of $\sim 400 m/s$, it is very unlikely (statistically) that a significant number of molecules will coherently move to the same direction, creating a spontaneous flow of air to one side of the room. As a result, the ensemble, or gas, is globally static, and the room is always uniformly filled by breathable air.

Statistical physics (of equilibrium) deals with many-body systems where the number of interacting units, and the nature of the interaction, is such that the global behavior of the system displays a limited number of stable equilibrium conditions. In these cases, one can derive an analytical picture, at the ensemble level, by giving up the microscopic determinism and introducing uncertainty - therefore statistics - in the theory.

To that end, the conceptual approach of statistical physics is to explain the phenomenology of a macroscopic object by deriving a macroscopic theory from the detailed laws that rule the behavior of its microscopic components.

The power of this approach lies in its generality: for the right set of phenomena, the global behavior of the ensemble does not depend of the detailed nature of its constitutive elements. Therefore, for example, the theory developed for the formation of droplets during liquid-vapor transition [11] can also be used to describe the self-sustained neural activity that encodes a spatial position in the collective state of a neural population [12]. As we will see in the next chapters, this generality has been leveraged by many (often successfully) to apply the tools of statistical physics to a wide variety of fields.

1.1.1 *The Boltzmann approach*

One of the pillars of statistical physics is the so-called Boltzmann distribution, which relates the probability for a system to be in a state to the energy of the state and the temperature. We will here review the derivation due to Boltzmann himself: the law was derived to describe the distribution of energy within a gas composed of a large number of molecules in a thermal bath, and was first given in his paper dated 1877 (see [13] for an English translation).

Let's consider an isolated idealized system composed of N interacting particles, each having kinetic energy e_i with $i \in [1, N]$. Globally, the total energy E is conserved, but particles can hit each other and exchange their energy by elastic collisions. Therefore, in time, the individual values of e_i will vary due to the continuous scattering happening at the microscopic level. The system is described by a chaotic trajectory of the $6N$-dimensional phase space vector

$$\mathbf{z}(t) = (\mathbf{q}_1(t), \dots, \mathbf{q}_N(t), \mathbf{p}_1(t), \dots, \mathbf{p}_N(t)) \quad ,$$

where $\mathbf{q}_i(t)$ and $\mathbf{p}_i(t)$ are the three-dimensional space position and three-dimensional momentum of the i-th particle.

Boltzmann started with a simplification: each e_i can take only a set of discrete values, and the exchanges happen accordingly via discrete amounts of an elementary unit ϵ. Therefore, at any time we have

$$e_i = \alpha_i \epsilon \quad ,$$

were each α is limited by 0 (below) and by the total energy $E = L\epsilon$ (above), and varies with time due to collisions. This is the equivalent of making a coarse-grained partition of the phase space into cells of finite size. The variable $\mathbf{z}(t)$, therefore, moves within this discrete (although huge) set of states. Each of this cells is called a *microstate*. A fundamental hypothesis is that the system occupies each microstate with *equally probability*, called the *ergodic hypothesis*. Boltzmann then focused on the question of *how the energy is distributed within the system*, i.e., how many particles have energy $e_i = 0$, ϵ, 2ϵ and so on. To do so, he performed a change of variable: instead of considering the microscopic kinetic state of the whole system $\mathbf{z}(t)$, he considered the *occupation vector*

$$\mathbf{n} = (n_0, n_1, \dots, n_L) \quad ,$$

where n_α is the number of molecules that have energy $\alpha\epsilon$. Now, it is clear that the coordinate change is not bijective, in the sense that each occupation vector corresponds to a different number of microstates in the phase space. The occupation vector is, therefore, a *macrostate*, i.e., a state of the system that corresponds to an entire region of the phase space. In the ergodic hypothesis, every microstate is equally probable. Therefore we can compute the relative probability of macrostates by just counting how many microstates map to each of them, i.e., computing the corresponding phase-space volume. We call this

number $W(\mathbf{n})$. To compute it, let's proceed iteratively: given an occupation vector $\mathbf{n}$, we have, out of N atoms,

$$\binom{N}{n_0} = \frac{N!}{(N-n_0)!n_0!}$$

ways of choosing n_0 molecules to whom assign the energies $e = 0$. We then are left with $N - n_0$ molecules, and we choose n_1 out of them to pick the molecules with energy $e = \epsilon$, so we multiply by $\binom{N-n_0}{n_1}$.

$$\frac{N!}{(N-n_0)!n_0!}\frac{(N-n_0)!}{(N-n_0-n_1)!n_1!} = \frac{N!}{n_0!n_1!}\frac{1}{(N-n_0-n_1)!}$$

Proceeding this way, we see that the $N - n_0 - n_1 \dots$ term is simplified at each step, until the final solution is found

$$W(\mathbf{n}) = \frac{N!}{n_0!n_1!\dots n_L!} \quad . \tag{1.1}$$

Due to the difficulty of treating factorials, we take the logarithm of $W(\mathbf{n})$ instead. With the Stirling approximation

$$\log N! \simeq N\log N - N \tag{1.2}$$

we can re-write Eq. 1.1 as

$$\log W(\mathbf{n}) \simeq N\log N - \sum_\alpha n_\alpha \log n_\alpha \quad . \tag{1.3}$$

The most likely macrostate $\mathbf{n}^* = \text{argmax}_\mathbf{n} \log W(\mathbf{n})$ is the one that, by virtue of typicality, dominates the probability when the number of particles N is very large, i.e., $P(\mathbf{n} = \mathbf{n}^*) \to 1$ when $N \to \infty$. We therefore maximize the expression in Eq. 1.3, under two important constraints: the total sum is equal to N and the total energy is equal to E. We therefore construct the functional with two Lagrange multipliers

$$\Phi[\mathbf{n}] = \log W(\mathbf{n}) - \gamma(\sum_\alpha n_\alpha - N) - \beta(\sum_\alpha n_\alpha \epsilon_\alpha - E) \tag{1.4}$$

Where, for generality, we used ϵ_α as the energy of the α-th occupation number. We then set the functional derivative of F to zero and get the expression of the occupation probability of the α-th energetic level:

$$\frac{\delta}{\delta n_\alpha}\Phi = 0 \iff n_\alpha = N\frac{e^{-\beta\epsilon_\alpha}}{\mathcal{Z}} \tag{1.5}$$

Where $\mathcal{Z} = e^\gamma = \sum_{\alpha'} e^{-\beta\epsilon_{\alpha'}}$ is obtained by imposing the conservation of the number of particles. By considering the probability of finding a particle with energy ϵ_α, i.e. $P(\epsilon_\alpha) = \frac{n^*_\alpha}{N}$, we find the so-called Boltzmann distribution:

$$\rho(\epsilon_\alpha) = \frac{1}{\mathcal{Z}}e^{-\beta\epsilon_\alpha} \tag{1.6}$$

1.1.2 *Boltzmann entropy and Helmholtz free energy*

The meaning of the second multiplier, β, can be derived by combining the computations above with the classical laws of thermodynamics (see for example [14] for a derivation). β is the *inverse temperature*,

$$\beta = \frac{1}{\kappa T} \quad , \tag{1.7}$$

where κ is called the Boltzmann constant. This constant is found in the famous definition of the Boltzmann entropy, which is the number of microstates that correspond to a given macrostate:

$$S = \kappa \log W \quad , \tag{1.8}$$

which, if we are to consider equations carved on gravestones as important, is a quite fundamental equation (see Fig.1.1).

Figure 1.1: Boltzmann gravestone in Vienna, with the equation of entropy as a function of the number of microstates. Historically, this precise form of the equation was given by Planck in 1902. Picture adapted from Wikipedia.

If we now consider the Helmholtz free energy $F = E - TS$ and plug our microscopic definition of entropy and energy we obtain

$$F = -\kappa T N \log \mathcal{Z} \quad . \tag{1.9}$$

Or, for the single particle

$$f = -\kappa T \log \mathcal{Z} \quad , \tag{1.10}$$

where $\mathcal{Z} = \sum_\alpha e^{-\beta \epsilon_\alpha}$ is the normalization factor that we encountered earlier in the derivation when we imposed the conservation of the number of particles. $\mathcal{Z}$ is called the *partition function*. From the Helmholtz free energy F we can retrieve the expected value of any quantity by differentiating F with respect to its *conjugate* variable. For example, it is straightforward to verify that the conjugate variable of the energy of a particle is β:

$$\langle e \rangle = -\frac{\partial \log \mathcal{Z}}{\partial \beta} = \sum_\alpha \epsilon_\alpha \frac{e^{-\beta \epsilon_\alpha}}{\mathcal{Z}} \quad . \tag{1.11}$$

1.1.3 *The Gibbs approach*

As we saw, the argument of Boltzmann relies on counting the configurations of energy units distributed among N particles (atoms or molecules). In these computations, we have assumed that these particles are statistically independent, since they only interact by elastic collisions. This assumption allows for the precise counting of microstates that leads to the (1.3) and, consequently, to the Boltzmann distribution. However, such an idealized computation has a limited range of application; in fact, it applies only to systems mappable to the idealized gas.

In his book of 1902 [9], Gibbs proposed a different approach. He started by the definition of *ensemble*, an idealized system composed of a great number of sub-systems. Each sub-system contains a large number of particles, such that it follows the laws of thermodynamics, but there is no idealized requirement on the nature of the interactions between its elementary constituents.

The sub-systems are instead considered as weakly interacting with each other and thermally coupled, such that heat exchanges can occur. This interaction allows for the internal energy of each sub-system to fluctuate, such that it can explore all the energetic levels. The subsystem still conserves the energy, but on *average*. This idealization is called the *Canonical ensemble*. Focusing on a single subsystem, we see that it can be in several different states s, each of energy E_s. Gibbs defined entropy for such system, which depends on the probability p_s of the system being in the state s:

$$H = -\kappa \sum_s p_s \log p_s \tag{1.12}$$

He then claimed that the equilibrium energy probability p_s is the one that maximizes the entropy H under the constraint of conserving the energy as an average over the probability distribution of states. Therefore, in order to find the equilibrium distribution,

we need to solve a constrained maximization similar to the one we have seen in the Boltzmann approach. We thus define the functional with two Lagrange multipliers

$$\Phi[p] = \sum_s p_s \log p_s - \gamma(\sum_s p_s - 1) - \beta(\sum_s p_s E_s - E) \tag{1.13}$$

By solving for the maxima we re-find the Boltzmann distribution of eq. 1.6.

$$\delta\Phi = 0 \iff p_s = \frac{1}{\mathcal{Z}} e^{-\beta E_s} \tag{1.14}$$

Where $\mathcal{Z}$ is the partition function of the canonical system

$$\mathcal{Z} = \sum_{\text{states } s} e^{-\beta E_s} \tag{1.15}$$

This time, however, the probability distribution is general for any system that is drawn from a canonical ensemble, without the need of the independence properties of an idealized gas. Therefore, once we know the energetic structure E_s of an equilibrium system, be it a solid, a liquid or generally strongly-interacting, we can always write its Boltzmann distribution.

The Gibbs approach is the foundation of all modern statistical physics. For an analytical as well as historical discussion on differences and similarities between the two approaches see for example [15, 16]. As we will see in the next section, the mathematical formalism of constrained maximization that we used to derive the Boltzmann distribution is the same that one finds in a particular class of inference problems, following the so-called *maximum-entropy principle*.

1.2 BAYESIAN INFERENCE

1.2.1 *Extended logic*

> By 'inference' we mean simply: deductive reasoning whenever enough information is at hand to permit it; inductive or plausible reasoning when – as is almost invariably the case in real problems – the necessary information is not available. But if a problem can be solved by deductive reasoning, probability theory is not needed for it; thus our topic is the optimal processing of incomplete information
>
> E.T. Jaynes

This extract appears as a footnote in the introductory part of Jaynes' book "Probability theory: the logic of science" [17], and gives a concise yet thorough definition of the inductive process we call "inference". As Jaynes suggested, in most real problems the amount of available relevant information is insufficient to find a solution by deductive reasoning. A scientist, consequently, needs to integrate the available information, following an inductive prescription, in order to reach a possible solution to said problems. Carried within the mathematical framework of probability theory, this integration process is called *statistical inference*.

In the first half of the 19th century, the combined work of R.T. Cox [18] and George Polya [19] showed that to conduct inference without violating basic logical and consistency assumptions [17], only one possible set of laws can be followed. These laws stipulate how to update one's degrees of uncertainty following a set of observations.

Remarkably, this was the set of *standard rules of probability theory*, originally given by Bernoulli in his work "Ars conjectandi" [20], and analytically developed by Laplace at the end of the 18th century. An interesting feature of the Polya-Cox result is that their prescription contains no reference to "chance" or "randomness", but instead descends by logical assumptions [17].

This result unified probability theory and statistical inference by defining a common set of principles, while at the same time reaching greater logical simplicity and widely expanding the range of possible applications of their mathematical framework.

In light of this logical unification, statistical inference is, in Jaynes' own words, nothing but "extended logic", by which problems can be quantitatively analyzed by following the sole, optimal, inductive prescription that shows consistency with a set of basic logical assumptions.

The Cox axioms

Notation Let the "degree of belief in proposition x" be denoted by $b(x)$. The negation of x (not x) is written $\bar{x}$ .The degree of belief in a conditional proposition "x, assuming proposition y to be true" is represented by $b(x|y)$.

Axiom 1 Degrees of belief can be ordered. If $b(x)$ is greater than $b(y)$ and this latter is greater than $b(z)$ then $b(x)$ is greater than $b(z)$.
$\implies$ degree of belief can be mapped onto real numbers

Axiom 2 The degree of belief in a proposition $b(x)$ and the degree of belief in its negation $b(\bar{x})$ are related, i.e. there exists a function f such that

$$b(x) = f[b(\bar{x})]$$

Axiom 3 The degree of belief in the joint proposition x, y (read x AND y) is related to the degree of belief in the conditional proposition $x|y$ and the degree of belief in the proposition y. In other words, there is a function g such that

$$b(x, y) = g[b(x|y), b(y)]$$

Consequence If a set of beliefs satisfy these axioms then they can be mapped onto probabilities satisfying $P(\text{TRUE}) = 1$, $P(\text{FALSE}) = 0$, $0 \leq P(x) \leq 1$, and the rules of probability

$$P(x) = 1 - P(\bar{x}) \tag{1.16}$$

$$P(x, y) = P(x|y)P(y) \tag{1.17}$$

(Box adapted from [21])

1.2.2 *The Bayes theorem*

The starting point of Bayesian probability theory is the "degree of belief" in a proposition x, which we encode into the probability $P(x)$. This degree can be conditional to the fact that another proposition y is true, in which case it is mapped onto the conditional probability $P(x|y)$. As shown in the box above, the probability function P that is derived from the Cox axioms satisfies the rules of probability theory (Eq.s 1.16 and 1.17). The fundamental relation of Eq. 1.17, that links the joint probability of two events $P(x, y)$ (x AND y) to the conditional probability $P(x|y)$, is known as *chain rule*. From the chain rule we can easily derive the following relation

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{1.18}$$

The formula in (1.18) is known as the Bayes theorem, named after Reverend Thomas Bayes, who first provided the equation as a way to update beliefs after new evidence in his "An Essay towards solving a Problem in the Doctrine of Chances (1763)". Bayes theorem is the foundation of all Bayesian statistics (also called the *subjective* view of probability), where probabilities are seen as degrees of belief (instead of occurrence frequences of random variables, which is called the *frequentist* view). The *frequentists*

vs. *subjectivists* is still an ongoing debate between experts of both fields. Quoting David MacKay [21], we will hereby take for granted that the Bayesian approach makes sense, and proceed consequently. For a resolute defense of the Bayesian approach, we refer the reader to Jaynes' book [17].

1.2.3 *Hypothesis testing*

One useful application of the Bayes theorem (Eq. 1.18) is the so-called Bayesian hypothesis testing: say we have two hypotheses for how a certain variable x behaves probabilistically. In other words, we have two putative probabilistic models $H_a = P^a(x)$ and $H_b = P^b(x)$. We also have collected a set of B realizations of said variable, which we call the *data* $D = x^1, x^2, \dots, x^B$. Our goal is to decide which of the two hypotheses is more likely to be true, given the available data.

For this task, it is convenient to name the terms in Eq. 1.18 to explicitly address observations (the data D) and the model (the hypothesis H). We so define the *likelihood* of our hypothesis as the probability of the data given the model $P(D|H)$, namely $P^a(D)$ and $P^b(D)$ for the two hypotheses; the *prior* $P(H)$ encodes the information that we have, for external reasons from the data, on the hypothesis H; the *evidence* $P(D)$ is the probability of the data independently of our hypothesis. The combination of these terms expressed by Eq. 1.18 defines the *posterior* of our problem, i.e., the degree of belief that we associate to the hypothesis H given the combination of the data D and our prior information:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)} \tag{1.19}$$

$$\Rightarrow \text{posterior probability} = \frac{\text{likelihood} \cdot \text{prior probability}}{\text{evidence}} \quad . \tag{1.20}$$

The task of choosing between the two hypothesis is therefore reduced to computing the two posteriors $P(H_a|D)$ and $P(H_b|D)$ and comparing their values. The hypothesis that maximizes the posterior probability is the one to be chosen as the most likely probabilistic model to explain the data.

1.2.4 *Maximum likelihood and maximum a posteriori*

Bayes theorem can be used to retrieve the parameters of a known statistical model given a set of observations. Say we want to model an N-dimensional variable

$$\mathbf{x} = (x_1, \dots, x_N)$$

for which we now know the statistical model, i.e. the probability function that regulates its behavior, up to a set of M unknown parameters,

$$\Theta = (\theta_1, \dots, \theta_M) \quad ,$$

which we need to fit to our set of observations. We write this probabilistic model as

$$P(\mathbf{x}|\Theta)$$

Where the conditional to Θ explicitly expresses the fact that we need to know the value of these parameters to describe the probability of $\mathbf{x}$. Now say we observe a set of B realizations of the variable $\mathbf{x}$, i.e. the *data* D,

$$D = \left\{\mathbf{x}^1, \ldots, \mathbf{x}^B\right\} \quad .$$

We want to find the most likely values of the parameters Θ given the evidence of D. By use of Bayes theorem, this is straightforward, since we can invert the statistical model to write the probability for the parameters given the data (what we seek) as a function of the probability of the data given the parameters (what we have, the statistical model)

$$P(\Theta|D) \propto P(D|\Theta)P(\Theta) \quad , \tag{1.21}$$

where we ignored the *evidence* term $P(D)$, since it does not depend on the parameters on which we are performing the maximization. If the observations of $\mathbf{x}$ are i.i.d. samples of the underlying probability distribution we can decompose the above into

$$P(\Theta|D) \propto \left[\prod_{k=1}^{B} P(\mathbf{x}^k|\Theta)\right] P(\Theta) \tag{1.22}$$

To work with sums instead of products, it is usually convenient to convert the equation above to the logarithm formulation

$$\log P(\Theta|D) \propto \sum_{k=1}^{B} \log P(\mathbf{x}^k|\Theta) + \log P(\Theta) \tag{1.23}$$

And our problem is solved by maximizing this posterior probability

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \left[\sum_{k=1}^{B} \log P(\mathbf{x}^k|\Theta) + \log P(\Theta)\right] \tag{1.24}$$

Depending on the complexity of the model $P(\mathbf{x}|\Theta)$ this maximization can be taken analytically, numerically, or via approximate formulas. If we do not specify any prior information, i.e. we take a flat $P(\Theta)$, the procedure is called *maximum likelihood* (ML), otherwise, if we have external prior information on the parameters that we want to include, it is called *maximum a-posteriori* (MAP).

Note from Eq. 1.24 that, in the limit of a very large number of observations, the prior is irrelevant compared to the data, as common sense suggests. Another important property of this formalism is that the inclusion of new data is trivial since one only needs to add one or more terms to the sum.

1.2.5 *The "max-entropy" principle*

As we saw in the previous section, the MAP and ML methods can be used to estimate the most likely parameters Θ of a statistical model, $P(\mathbf{x}|\Theta)$, given multiple independent observations of the variable $\mathbf{x}$.

It is clear that we can use these methods only if we know *a priori* what family of distributions $P(\mathbf{x})$ we ought to use as a statistical model for the analyzed problem. In practical applications, the model is usually given by external information, such as the underlying physical laws that regulate the analyzed problem or one's assumptions regarding the statistics of a process. In this case, we only have few degrees of freedom, i.e., the parameters $\Theta = (\theta_1, \ldots, \theta_M)$, whose values we can derive by using ML or MAP on the set of observations.

However, there are cases where we do not know the underlying model, but we still would like to integrate a set of observations and derive a statistical predictive model. For example, say that of a sequence of observations $\{x_k\}$ of a variable x we only know the average value, i.e. $\bar{x} = \frac{1}{B}\sum_k x_k$, and that we would like to make statistical predictions on the next outcome x_{B+1}.

The problem is: what family of parametric functions should we use if we know nothing but aggregated information such as means, correlations, and other statistical averages of our data? In this case, the degrees of freedom are *infinite*, since there are infinitely many distributions that display the given average values. The *maximum-entropy principle* addresses precisely this question.

The intuitive reasoning that lies behind this principle is that, in performing our inference procedure, we want to choose the distribution family in the *fairest possible way*, i.e., without adding any constraints to the problem that are not directly deducible from the available data. Following the principle, there is only one family that satisfies this requirement and is the one that maximizes the *Shannon entropy*, defined as

$$H[P] = -\int dx P(x) \log P(x) \tag{1.25}$$

or, in the discrete case

$$H[P] = -\sum_i P_i \log P_i \quad , \tag{1.26}$$

under the constraint of displaying the empirical average values. As an example, let's consider a discrete case where our variable x can take only a finite set of values $\{x_i\}$, each with probability P_i. The scenario is the one presented above, i.e., the only information we have is the mean value of a set of observations of the variable, $\bar{x}$. To find the maximum entropy distribution we need to solve the constrained maximization by using the Lagrange multipliers formalism, where we include a multiplier λ_0 for the normalization constraint ($\sum_i P_i = 1$) and another λ_1 for the mean value ($\sum_i x_i P_i = \bar{x}$). The constrained maximum P_i^* is the one that solves

$$0 = \frac{\delta}{\delta P_i}\left[H[P] - \lambda_0(\sum_i P_i - 1) - \lambda_1(\sum_i x_i P_i - \bar{x})\right] \quad . \tag{1.27}$$

With basic algebra we find the solution to be

$$P_i^* = \frac{e^{-\lambda_1 x_i}}{\mathcal{Z}} \quad , \tag{1.28}$$

where $\mathcal{Z} = e^{\lambda_0}$ is a normalization constant, found by applying the normalization constraint, i.e.,

$$\sum_i P_i^* = 1 \implies \mathcal{Z} = \sum_i e^{-\lambda_1 x_i} \quad . \tag{1.29}$$

The specific value of λ_1 is fixed by the constraint on the mean value

$$\sum_i x_i P_i^* = \bar{x} \tag{1.30}$$

The original formulation of this principle is due to Jaynes [22] and is based on the interpretation of the Shannon entropy as the "randomness" of the probability distribution, or "ignorance" about the realization of the random variable drawn from it. In this view, to take nothing but the data into account means to maximize our ignorance about the problem, therefore taking the most "random" possible family of distributions consistent with observables. The principle has been later derived axiomatically, claiming that no other distribution family than the one that maximizes the Shannon entropy can be used to perform inference, based on average values, without contradicting a set of consistency axioms [23].

As mentioned in the first chapter, the Gibbs-Boltzmann distribution for the canonical ensemble is of the same exponential family of the max-entropy distribution constrained to reproduce the average value $\bar{x}$. As argued by E.T. Jaynes [22], the connection between statistical mechanics and information theory is more than a formal coincidence; it instead establishes a viewpoint on statistical physics as a theory based on the state of knowledge of the experimentalist instead of the physical details of the system under consideration. The interested reader can find a detailed discussion on this connection in the works of Jaynes, see for example [17, 22, 24].

2

INTERSECTIONS

During the last decades, the range of application of statistical physics has widened enormously, extending its influence outside of the boundaries of physics per se. Its conceptual and mathematical framework has been successfully applied to problems from chemistry, biology, computer science, ecology, and even social sciences, such as sociology and economics.

This "foreign" success of statistical physics is at least in part due to its tight relationship with *statistics*. Through the development of mathematical tools derived from the statistical analysis of physical phenomena, statistical physics provided researchers with a framework that is adaptable to a broad category of problems, namely those that involve a large number of interacting units that give rise to macroscopic collective behaviors.

In this chapter, we will discuss two cases of intersection between statistical physics and foreign fields, namely **Bayesian inference** and **systems biology**. We will try to explore these vast regions by following a common thread that links to the research work presented in chapters 5, 6, 7, 10, and 11. The following sections are therefore thought to be a technical and philosophical introduction to the work presented here, more than an exhaustive historical overview of the relationship between these diverse scientific areas. Indeed, this latter would surely deserve a dedicated thesis in the sociology of science to fairly cover it in its many facets.

As we will see, the application of tools from statistical physics to Bayesian inference has brought theoretical insights that allowed for the development of performant algorithms for statistical inference and data analysis. The overlap between statistical physics and systems biology, instead, is the quantitative *top-down* modelling approach (i.e., from mathematical abstraction to observables) that allowed to understand and make novel quantitative predictions in a wide range of biological systems, from cellular motility to flocks of birds, from populations of neurons to protein folding.

Finally, the advent of powerful computers and large biological datasets has made necessary the development of a whole new category of "big data" *bottom-up* analysis methods, which are at the intersection between systems biology and Bayesian inference. Despite the great importance of recent developments in bioinformatics by inference methods applied to biological data, this intersection will not have a dedicated section. We will cover an introduction to the inference and data-analysis methods used in the present work in chapters 4 and 8.1. The interested reader is also referred to the relevant reviews in the literature [25–27].

2.1 BAYESIAN INFERENCE ∩ STATISTICAL PHYSICS

An interesting consequence of the close interrelation between statistical physics and Bayesian statistics is that some of the mathematical effort carried by physicists, in the everlasting attempt to formalize and explain physical phenomena, can be borrowed to develop new algorithms for statistical inference and, ultimately, data analysis.

A good example located at this intersection is the problem of retrieving a graph of interaction, named **network inference**. The problem, in its generality, could be phrased as

> There is a group of N agents, influencing each other via an interaction matrix $\mathbf{J}$, and whose activity $\mathbf{s} = (s_1, \dots, s_N)$ we have repeatedly collected as empirical observations; can we retrieve the interaction structure $\mathbf{J}$ from these observations?

This problem has attracted great interest, in the last decades, in the communities of statistical physics and computer science, since the nature of the agents and the interaction matrix can vary depending on the specific application. It could represent the phenomenology of opinion dynamics during a political riot, or the collective behavior of neurons in a specific brain area, or even the interaction structure of magnetic spins of atoms in a magnet. In this latter case, a simplified model of magnetic spins that can take either direction up ($s = +1$) or direction down ($s = -1$), is the so-called **Ising model** in statistical physics.

2.1.1 *The inverse Ising model*

The model is named after the physicist Ernst Ising, who invented the formalism in his doctoral thesis [28], however giving a solution only in the one-dimensional chain case. Onsager has later given the more involved solution of spins placed on a two-dimensional lattice in 1944 [29]. These solutions are an example of the so-called *direct problem*, i.e., deriving the value of observables starting from the Hamiltonian formulation of the problem.

The Hamiltonian formulation is based on the definition of the *energy* E of the system, whose state space is the hypercube of all binary vectors $\mathbf{s} = (s_1, \dots, s_N)$ of dimension N (number of spins). The energy depends on the specific state $\mathbf{s}$ through the matrix of couplings J_{ij} and a set of magnetic fields h_i:

$$E(\mathbf{s}) = -\sum_{i=1}^{N} h_i s_i - \sum_{i<j} J_{ij} s_i s_j \quad . \tag{2.1}$$

As we saw in the first chapter, in equilibrium conditions the probability of a configuration $\mathbf{s}$ is given by the Boltzmann distribution over its energy

$$P(\mathbf{s}) = \frac{1}{\mathcal{Z}(\mathbf{h}, \mathbf{J})} e^{\left(\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j\right)} \quad , \tag{2.2}$$

where we conveniently choose the temperature scale such that $\beta = 1$. Now, say that instead of starting from the energy of the system and deriving the behavior of the model, we observe a set of B configurations, i.e. the data $D = \{\mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^B\}$, and we want to retrieve the interaction matrix $\mathbf{J}$ and the field vector $\mathbf{h}$. This is called the *inverse* problem.

Since we know what probability distribution the degrees of freedom of the model are following, i.e., we know the Boltzmann distribution $P(\mathbf{s})$ in Eq. 2.2, we can apply the Bayesian framework and retrieve the most likely value of the parameters $\Theta = \{\mathbf{J}, \mathbf{h}\}$ given the evidence D. We will use the maximum-likelihood method that we described in the previous section (see Eq. 1.24) since at this point we have no particular reason to include a prior $P(\Theta)$ on the values of the parameters.

We proceed by writing the log-likelihood function, that is maximized by the solution:

$$\begin{aligned}\mathcal{L}(\Theta) := \log P(\Theta|D) &\propto \sum_{k=1}^{B} \log P(\mathbf{s}^k|\Theta) \\ &= \sum_{k=1}^{B} \left(\sum_i h_i s_i^k + \sum_{i<j} J_{ij} s_i^k s_j^k - \log \mathcal{Z}(\mathbf{h}, \mathbf{J}) \right) \\ &= B \left(\sum_i h_i \langle s_i \rangle_D + \sum_{i<j} J_{ij} \langle s_i s_j \rangle_D - \log \mathcal{Z}(\mathbf{h}, \mathbf{J}) \right) \end{aligned} \tag{2.3}$$

where the notation $\langle \rangle_D$ indicates the average over the observed data D. Note that we can interpret the equation (2.3) in terms of physical quantities: the first term is minus the mean energy estimated from the empirical observations, and the second one is minus the free energy of the system. Therefore, the log-likelihood has the same form of a *entropy*, with a minus sign (see Section 1.1.2). For this reason, in the statistical physics community, the log-likelihood maximization is also referred to as *cross-entropy minimization*.

We now apply the condition of maximum likelihood to retrieve the most likely values of the parameters $\widehat{\Theta} = \widehat{\mathbf{J}}, \widehat{\mathbf{h}} = \text{argmax}_\Theta \mathcal{L}(\Theta)$

$$\widehat{h}_i : \qquad 0 = \frac{1}{B} \frac{\partial \mathcal{L}}{\partial h_i} = \langle s_i \rangle_D - \frac{1}{\mathcal{Z}} \frac{\partial \mathcal{Z}}{\partial h_i} \iff \langle s_i \rangle_D = \langle s_i \rangle_{P(\mathbf{s})} \tag{2.4}$$

$$\widehat{J}_{ij} : \quad 0 = \frac{1}{B} \frac{\partial \mathcal{L}}{\partial J_{ij}} = \left\langle s_i s_j \right\rangle_D - \frac{1}{\mathcal{Z}} \frac{\partial \mathcal{Z}}{\partial J_{ij}} \iff \left\langle s_i s_j \right\rangle_D = \left\langle s_i s_j \right\rangle_{P(\mathbf{s})} \tag{2.5}$$

The two last terms of (2.4) and (2.5) are called *moment-matching conditions*, since they express the requirement that the correlations $\langle s_i s_j \rangle$ and the magnetizations $\langle s_i \rangle$ computed over the probability distribution $P(\mathbf{s}|\widehat{\Theta})$ have to be the same of the ones computed on the empirical data D.

An important remark is that this formalism can be worked out in the exact opposite direction. Let's say that we observe the magnetizations and correlations of a set of

interacting units and that we know that the state-space is the hypercube of binary vectors of dimension N. If we follow the principle of *maximum entropy* and we look for the least biased distribution (maximal in the Shannon entropy) that reproduces said magnetizations and correlations, we find that the solution is precisely the Ising model.

$$\delta\left[\sum_{\mathbf{s}} P(\mathbf{s})\log P(\mathbf{s}) + \sum_i \lambda_i\left(\sum_{\mathbf{s}} P(\mathbf{s})\cdot s_i - \langle s_i\rangle_D\right) + \sum_{i<j}\lambda_{ij}\left(\sum_{\mathbf{s}} P(\mathbf{s})\cdot s_i s_j - \langle s_i s_j\rangle_D\right)\right] = 0$$

$$\iff P(\mathbf{s}) = \frac{1}{\mathcal{Z}} e^{\left(\sum_i \lambda_i s_i + \sum_{i<j}\lambda_{ij} s_i s_j\right)} \tag{2.6}$$

Where the moment matching conditions of (2.4) and (2.5) are then imposed to retrieve the values of the Lagrange multipliers λ_i and λ_{ij}. The exponential model of (2.6) has a name and a history in the field of statistics, where is called *undirected pairwise graphical model* [30].

2.1.2 *Sparsity and regularization*

Having defined our task within the Bayesian framework, we are allowed to make and control assumptions on the value of the inferred parameters, $\widehat{\Theta}$, by encoding them as prior probabilities $P(\Theta)$. Several examples in the literature showed how such priors are useful (and sometimes necessary) to avoid degenerate solutions, reduce overfitting, and to speed up the convergence of the algorithms [31].

One widely used prior is the so-called ℓ_1 *regularization*, related to the LASSO regression in statistics, and introduced by [32] in the context of the inverse Ising model. The ℓ_1 regularization is an exponential prior probability in minus the ℓ_1-norm of the parameter vector, which penalizes solutions whose sum of absolute values of the inferred parameters is large:

$$P_{\ell_1}(\Theta) \propto \exp\left(-\lambda \sum_{\theta\in\Theta} |\theta|\right) \quad . \tag{2.7}$$

By defining two parameters λ_h and λ_J, that control the strength of the prior over fields and couplings, respectively, the the log-likelihood (2.3), now log-posterior, of the inverse Ising problem can be written as

$$\mathcal{L}_{\ell_1}(\Theta) = B\left(\sum_i h_i\langle s_i\rangle_D + \sum_{i<j} J_{ij}\langle s_i s_j\rangle_D - \log\mathcal{Z}(\mathbf{h},\mathbf{J})\right) - \lambda_h\sum_i |h_i| - \lambda_J\sum_{i<j}|J_{ij}| \quad , \tag{2.8}$$

where $\lambda_{h,J} = O(1)$ to ensure consistency with the requirement that the posterior is dominated by the likelihood in the presence of a large number of data. The role of this prior is to enforce a subset of the parameters to be exactly 0, effectively reducing the number of inferred parameters. For this reason, it is also referred to as a *sparsity prior*.

Another possibility to select solutions with a small absolute value of the inferred parameters is to assume a Gaussian probability distribution over the ℓ_2-norm of the parameter vector. This results in an additional quadratic penalty term in the log-posterior and is called ℓ_2 regularization. The log-posterior therefore reads:

$$\mathcal{L}_{\ell_2}(\Theta) = B\left(\sum_i h_i \langle s_i \rangle_D + \sum_{i<j} J_{ij} \langle s_i s_j \rangle_D - \log \mathcal{Z}(\mathbf{h}, \mathbf{J})\right) - \lambda_h \sum_i (h_i)^2 - \lambda_J \sum_{i<j} (J_{ij})^2 \quad (2.9)$$

This is a necessary hypothesis if one has to deal with under-sampled data where the natural solution of the inverse problem would retrieve infinitely-negative parameters (for example, missing data on one single site would lead to $\langle s_i \rangle_D = 0$ and consequently to a field $h_i = -\infty$). In this the ℓ_2 norm is equivalent but less invasive than the ℓ_1 norm, since it does not enforce a sparse solution. For a detailed discussion on the role of regularizations in the inverse Ising problem, see [31].

2.1.3 *Computational approaches*

Having derived the moment-matching conditions (2.4) and (2.5) from the maximum-likelihood approach, one could be tempted to think of the inverse Ising model as a solved task. The reality, however, is computationally much more complex.

In fact, the equations (2.4) and (2.5) cannot be solved for the parameters $\widehat{J}_{ij}$ and $\widehat{h}_i$, due to the many-body nature of the Ising model. A change in a single parameter J_{ij}, in fact, will affect multiple correlations $\langle s_{i'} s_{j'} \rangle$ and, vice-versa, a change in a single empirical correlation $\langle s_i s_j \rangle_D$ would lead to several changes to parameters $\widehat{J}_{ij}$ in the solution of the inverse problem. Moreover, the inverse Ising inference has been proven to be an NP-hard problem, i.e., there does not exist (up to now) an algorithm able to solve it in polynomial time in the number of units N.

The only way to satisfy the matching conditions is, therefore, to proceed via numerical methods. Luckily, one can prove that the log-likelihood of the Ising model is a concave function of the parameters, [33], allowing for the application of convex optimization methods to reach the solution by following the gradient of the log-likelihood in the space of the parameters. This gradient is defined as

$$\nabla \mathcal{L} = \left(\frac{\partial \mathcal{L}}{\partial h_1}, \ldots, \frac{\partial \mathcal{L}}{\partial h_N}, \frac{\partial \mathcal{L}}{\partial J_{1,2}}, \ldots, \frac{\partial \mathcal{L}}{\partial J_{1,N}}, \ldots, \frac{\partial \mathcal{L}}{\partial J_{N-1,N}}\right) \quad (2.10)$$

Where the single terms can be computed from the definition of the log-likelihood (2.3)

$$\frac{1}{B}\frac{\partial \mathcal{L}}{\partial h_i} = \langle s_i \rangle_D - \langle s_i \rangle_{P(\mathbf{s})} \quad (2.11)$$

$$\frac{1}{B}\frac{\partial \mathcal{L}}{\partial J_{ij}} = \langle s_i s_j \rangle_D - \langle s_i s_j \rangle_{P(\mathbf{s})} \quad (2.12)$$

To reach the solution, one can proceed by changing the value of the parameters iteratively, following the gradient, and checking at each iteration if the moment-matching condition is satisfied up to a given convergence criterion. For example, in the pseudocode below we stop only if each moment is matched up to a precision threshold ϵ

log-likelihood maximization by gradient ascent

loop
  for $i = 1, \ldots, N$ **do**
    $\Delta_i \leftarrow \langle s_i \rangle_D - \langle s_i \rangle_{P(\mathbf{s})}$
    for $j = 1, \ldots, N$ **do**
      $\Delta_{ij} \leftarrow \langle s_i s_j \rangle_D - \langle s_i s_j \rangle >_{P(\mathbf{s})}$
    end for
  end for
  if $(\exists i : \ |\Delta_i| > \epsilon)$ **or** $(\exists (ij) : \ |\Delta_{ij}| > \epsilon)$ **then**
    $h_i \leftarrow h_i + \eta_h \Delta_i$
    $J_{ij} \leftarrow J_{ij} + \eta_J \Delta_{ij}$
  else
    convergence has been reached, break loop
  end if
end loop

Where η_h and η_J rule the speed of movement in the parameter space, and are called *learning rates*. This pseudocode introduces the next problem, that is that for each iteration of the process we need to compute the averages $\langle s_i \rangle_{P(\mathbf{s})}$ and $\langle s_i s_j \rangle_{P(\mathbf{s})}$. However, there is no analytical solution for the *direct Ising problem* that provides us with a generic closed form for these average values given a particular choice of couplings $\mathbf{J}$ and fields $\mathbf{h}$.

Again, we need a computational approach: one way is to compute the partition function $\mathcal{Z}$ and then numerically estimate its derivatives with respect to J_{ij} and h_i, which give, respectively, the magnetization $\langle s_i \rangle_{P(\mathbf{s})}$ and the correlation $\langle s_i s_j \rangle_{P(\mathbf{s})}$ of the model. However there is one major obstacle to this approach, i.e.

The partition function $\mathcal{Z}(\mathbf{J}, \mathbf{h})$ is the sum of 2^N terms, N being the dimensionality of the problem.

It is clear that, even for a modest analysis of $N = 50$ interacting units, we can not afford to enumerate, at each iteration, all the $\sim 10^{15}$ terms that compose the partition function. For this reason, people from the fields of statistical inference, machine learning, and physics have worked to develop computational methods that solve the inverse Ising problem without the need for exactly computing the partition function. We will here enlist some results of these efforts. For a recent review on the matter see for example [34].

(a) **Boltzmann learning:** a popular technique in machine learning [35], it consists in simulating the system with Monte Carlo methods to estimate, instead of compute, the averages $\langle\rangle_{P(\mathbf{s})}$ in (2.11) and (2.12), and proceed by gradient ascent as shown in the pseudo code. The stationary point of this procedure is proven to be the correct solution for the inverse Ising problem. This technique avoids the explicit computation of the partition function, but is still computationally very demanding, since it requires to simulate the system, at each update of the parameters, for a time that is long enough to avoid dependence on the initial condition (a requirement called *thermalization*). For even a reasonably small number of units ($N \sim 100$) it is usually impossible to reach convergence in a reasonable time. However, it is safely employable for smaller systems, and its simplicity has made it one of the most popular algorithms in the field, widely used in the literature of the last decades for a great variety of problems.

(b) **Mean field:** directly inspired by theoretical approximation techniques, this method is based on the hypothesis of statistical independence of single spins s_i. This allows for the factorization of the Boltzmann distribution into $P(\mathbf{s}) = \prod_i P_i(s_i) = \prod_i \frac{1+\mu_i s_i}{2}$, where $\mu_i = \langle s_i \rangle_P$ is the magnetization of the spin i. If we assume this factorization, the max-likelihood couplings and fields can be analitically retrieved from the empirical moments, thanks to equations developed from the Gibbs free energy of the Ising model (see for example [34] for a derivation). This solution is based on the definition of the matrix of connected correlations

$$C_{ij} := \langle s_i s_j \rangle_D - \langle s_i \rangle_D \langle s_j \rangle_D \tag{2.13}$$

and reads

$$\widehat{J}_{ij} = -(\mathbf{C}^{-1})_{ij} \quad i \neq j \tag{2.14}$$

$$\widehat{h}_i = \operatorname{atanh}\langle s_i \rangle_D - \sum_{j \neq i} J^*_{ij} \langle s_j \rangle_D \tag{2.15}$$

The mean-field approach has the obvious advantage of being immediately computable (it just requires the inversion of a matrix, that can be done in $O(N^3)$ polynomial time), but its range of validity is limited to those cases in which the above factorization of $P(\mathbf{s})$ is an appropriate approximation. Its applicability has, therefore, to be checked case by case. For further reading see for example [33, 36, 37]

(c) **Tree-like graphs:** if the connectivity structure defined by the matrix $\mathbf{J}$ is tree-like, i.e. contains no or few interaction loops, the partition function can be computed in $O(N)$ time [38] by employing the so-called *Message-passing* or *Belief-propagation* methods. These methods have been shown to be exact on tree-like structures, or when loops are confined to a local scale. It has been showed that the message-passing methods are equivalent to assuming the *Bethe-Peierls approximation* [39],

an analytical tool derived in statistical physics to compute the partition function and expectation values by solving a set of non-linear equations. For a detailed exposition of these methods and their applications see [40].

(d) **Pseudo likelihood:** the algorithm derived by Ravikumar et al. [32] solves the inverse Ising problem by requiring the knowledge of the full ensemble of observed patterns $\{\mathbf{s}^k\}$, instead of the empirical averages $\langle s_i s_j \rangle_D$ and $\langle s_i \rangle_D$. It is based on an approximation that considers N independent single-spin problems, each conditioned to the value of the remaining spins

$$P(s_i|\{s_j\}_{j\neq i}) = \frac{1}{2}\left[1 + s_i \tanh\left(h_i + \sum_{j\neq i} J_{ij} s_j\right)\right] \tag{2.16}$$

By using this expression for the probability of the data given a set of parameters we can write the log-likelihood function for the i-th row of the matrix $\mathbf{J}$ and for the field h_i given a set of B empirical observations $\{\mathbf{s}^k\}$

$$\mathcal{L}_i^{PL} = \sum_k \log \frac{1}{2}\left[1 + s_i^k \tanh\left(h_i + \sum_{j\neq i} J_{ij} s_j^k\right)\right] \tag{2.17}$$

Whose maximization gives the following equalities for the solution $\widehat{\mathbf{h}}, \widehat{\mathbf{J}}$

$$\langle s_i \rangle_D = \left\langle \tanh\left(\widehat{h}_i + \sum_{j\neq i} \widehat{J}_{ij} s_j\right)\right\rangle_D \tag{2.18}$$

$$\langle s_i s_j \rangle_D = \left\langle s_i \cdot \tanh\left(\widehat{h}_i + \sum_{j\neq i} \widehat{J}_{ij} s_j\right)\right\rangle_D \tag{2.19}$$

We are therefore left with N minimization problems that can be carried out by using standard routines such as Newton method or gradient descent.

Importantly, the pseudo-likelihood approximation is proven to be asymptotically consistent [41], i.e., it retrieves the max-likelihood solution when the number of data goes to infinite. Note that the complexity class of this algorithm is polynomial in the number of parameters and in the number of data, i.e., $O(BN)$, therefore is usually much faster than Boltzmann learning. Differently from other inverse-Ising algorithms, the PL couplings are generally asymmetric, i.e. $\widehat{J}_{ij} \neq \widehat{J}_{ji}$.

This algorithm has been widely studied in the field of statistical inference [32, 42], and has lately gained popularity in the field of bioinformatics, since it has been shown to provide good results in the problem of reconstructing the 3D structure and fitness landscape of proteins starting from sequence covariation within the relevant protein family [43, 44].

(e) **Adaptive cluster expansion:** derived by Cocco and Monasson [33, 45–47], the adaptive cluster expansion (ACE) method is based on the expansion of the cross-entropy of the inverse problem

$$S(\mathbf{h}, \mathbf{J}|D) = -\log \mathcal{Z} + \sum_i h_i \langle s_i \rangle_D + \sum_{i<j} J_{ij} \langle s_i s_j \rangle_D \tag{2.20}$$

which, up to a minus sign, equals the log likelihood of the parameters given the data. The cross-entropy is expanded into several terms, each corresponding to a cluster of spins of varying size

$$S(\mathbf{h}, \mathbf{J}|D) = \sum_{\Gamma \in \mathcal{P}(N)} S_\Gamma \tag{2.21}$$

where $\mathcal{P}(N)$ is the power set, i.e., the set containing all possible subsets (unordered), of the N spins. The algorithm builds an iterative approximation of the cross-entropy that decomposes the inverse problem into a set of smaller inverse problems of increasing size. The procedure starts from the single-site and two-sites clusters, which have an analytical solution and are therefore immediate to solve:

$$S(\mathbf{h}, \mathbf{J}|D) = \sum_{\Gamma : |\Gamma| < 3} S_\Gamma + \Delta S \tag{2.22}$$

The method then proceeds iteratively to decompose the remaining ΔS. It first solves all the small inverse problems of all included clusters up to size $k = 2$, then for each cluster Γ it computes S_Γ, i.e., its contribution to the cross-entropy, and checks if it exceeds a chosen threshold θ. If it does, the cluster is marked as *significant*, and included in the expansion. In the next iteration ($k = 3$) only clusters of size three that are composed by significant clusters of size two are considered, therefore significantly reducing the total number of terms in the expansion. The procedure continues by lowering down the threshold θ and repeating the expansion until a criterion of convergence (derived from the moment matching conditions) is reached.

The intuition behind this procedure is that the number of terms in the expansion is exponentially reduced by excluding, upstream, all the irrelevant clusters. If a small cluster is irrelevant to the cross-entropy, it is unlikely that a larger one, derived from it, will be relevant. Since the complexity class of the inverse problem is exponential, solving a large number of smaller tasks is usually more convenient than solving a single large inverse problem. Therefore, the convergence of this algorithm is usually much faster than classical Boltzmann learning [33].

In chapters 3 and 8.1 we will show how the inverse problem can be applied to systems in neuroscience and bioinformatics, following a recent tradition of successful applications of this paradigm to biological systems [48–58]. We will predominantly use the ACE method.

2.2 STATISTICAL PHYSICS ∩ SYSTEMS BIOLOGY

The statistical-physics modelling of biological systems is today a widely-used and recognized approach in quantitative biology. But how can a theoretical framework that has been invented to describe the thermodynamics of inert gases be successfully applied to biological systems? We will here try to develop our humble point of view on the matter, which, far from being solved, is still the object of an ongoing debate in science and philosophy.

2.2.1 *Biology is complicated*

Biology by definition involves life. The understanding of the underlying principles of living systems is one of the main and most fascinating challenges of modern science. However, for a physicist, used to controlled mathematical models of reality, biology is utterly *complicated*. Even the most simple biological process that we can imagine is composed of a large number of diverse parts that interact by following physical, chemical, and biochemical laws on different scales. Designing a precise and all-encompassing physical model of such complexity is often a hopeless and possibly pointless task, which would require an immense number of equations and parameters that can hardly be interpreted to inspire new insights on the analyzed system.

The only reasonable approach, in this case, seems to be the reduction of the system to its elementary parts. These can be studied in controlled settings, in order to establish a detailed understanding of their individual behavior. A complete description of the whole, hopefully, will naturally emerge from the sum of its parts. This approach is known as *methodological reductionism*. The reductionist approach has proven very successful in understanding the principles that rule the behavior of elementary biological components. Its *palmarés* is adorned with the most important scientific discoveries of the last centuries, ranging from the helix structure of DNA in genetics to the Krebs cycle in biochemistry, and forms the basis for many of the well-developed topics of modern science.

Starting from the '70s, however, it has been increasingly acknowledged that this approach is limited when it comes to describing the *collective behavior* of biological systems. For example, a more detailed biophysical description of the membrane potential of neurons does not help to understand how cognitive functions are performed by the central nervous system, as much as the characterization of human biology will hardly be sufficient to understand the emergence of societies. Acknowledging the limits of reductionism has pushed researchers and philosophers to seek a complementary paradigm, focused more on the *emergent behavior* than on an accurate description of the elementary components. The diverse set of efforts that have been carried in the last decades in this direction goes under the name of "complex systems".

2.2.2 *Biology is complex*

The label *complexity* encloses a set of philosophical and scientific approaches that aim to understand how the interaction between the parts of a system gives rise to its collective behavior. These systems, composed of many interacting parts, are referred to as *complex systems*. It is usually opposed to reductionism in the sense that complex systems can not, in this view, be understood by individually studying their fundamental components. The system must instead be observed and studied as a whole, since the origin of its behavior lies in the *interaction structure*, and not in the physical nature, of its components.

Although there is no consensus on a precise definition of "complexity", there are some properties that are widely associated with "complex systems" in the literature [59,60]. Among these, the one that we think is most relevant for our modelling approach is *emergent behavior*.

The classic example of emergent behavior is the collective motion of groups of individuals, such as flocks of birds [61,62], fish schools [63], or pedestrians that move during an emergency situation [64]. In all these cases, in line with complexity theory, a detailed characterization of the motion of an isolated individual is of small use in understanding the collective one, which instead emerges from the interaction between individuals and with the external environment.

An important feature of complex systems is that their description can be developed on several different layers of increasing *coarse-grained* scale. Starting from the immense phase-space of the system, we can proceed by considering larger "pixels" on spatial, temporal, or generally abstract dimensions. The blurring effect of coarse-graining can, in some cases, average out an underlying chaotic dynamics and allowing for stable and reproducible patterns on a larger scale. A typical example is the weather conditions: even if we know that the underlying dynamics is non-linear and chaotic [65], we still can find patterns and regularities in some macroscopic variables, such as the alternation of wind velocity, temperature, pressure, and humidity, in time [60]. As we will see below, this hierarchy of description levels is one of the enabling factors for our modelling approach.

2.2.3 *Simple models of complex systems*

In the last decades, *complexity* has become pervasive in every scientific field that deals with systems composed of many interacting units: examples are economics, systems biology, sociology, climate science, ecology, and neuroscience. As the attentive reader might have noticed, this list very much resembles the set of scientific fields that we mentioned at the beginning of this chapter, where we enumerated some examples of disciplines that have been quantitatively approached through tools of statistical physics.

Indeed, from the above description of *complex system* we can see the first touching point: as it happened with the kinetic theory of gases, we know that the underlying dynamics is impossible to characterize in detail, but we still observe some regularities in the global (macroscopic) behavior of a large collections of individuals. The observation

of these patterns motivates our curiosity, and we would like to understand how these regularities emerge from the interaction between the parts.

This analogy seems to highlight statistical physics as a possible quantitative approach to the study of complex systems. However, statistical physics has been invented to model ideal gases at thermal equilibrium, which is as far as one can get from the concept of complexity. We will try to clarify this apparent contradiction by some further discussion on the rationale behind our modelling approach: *what* we, as physicists, want to model about these systems, *how* do we extract relevant observables from our abstraction of reality and, ultimately, *why* do we choose this approach, i.e. how this process can push further the understanding of the underlying mechanisms.

what As mentioned before, our approach to biological systems *can not* aim at the mathematical characterization of all the relations between the involved variables. Instead, we *isolate* what we think are the most intriguing features of a system, based on our curiosity, scientific sensibility, and intuition on how we could approach these phenomenologies through our quantitative toolset.

Taking a little creative liberty, and in all probability being incautious from a philosophical point of view, we might define this approach as *phenomenological reductionism*: instead of investigating the system by separating it into its fundamental physical constituents, we separate it into the *phenomenological* ones, choosing to individually investigate the behaviors that pose the most interesting (and hopefully treatable) scientific questions.

This coarse-graining approach is not new to physicists. The same idea of neglecting some degrees of freedom, in order to isolate the relevant variables, is at the basis of many mathematical models in physics (e.g., the Ising model to describe phase transitions in magnets). The difference is that, in biological systems, the coarse-graining happens on a much larger scale. Moreover, the relevant phase space is harder to precisely define, leading to approximations that are much more difficult to control.

There is a famous quote, from the statistician George Box, that well summarizes this idea: "Essentially, all models are wrong, but some are useful". In the case of mathematical modelling of complex biological systems, we know for sure that our models are *very* wrong. Hopefully, they can also be of some use in understanding the mechanisms by which biological systems achieve their rich and fascinating phenomenology.

how Once we conceptually isolated the behavior that we want to understand, the second step of our approach is to (1) use our trained mathematical intuition to choose the relevant state space, i.e. what degrees of freedom $\mathbf{s} = (s_1, s_2, \dots, s_N)$ we are going to model, (2) choose the relevant parameters $\Theta = (\theta_1, \theta_2, \dots, \theta_M)$ that we think are responsible for the phenomenology we want to investigate (for example the coupling strength between two terms in the equation, the network matrix of interactions, ...), and (3) define a function that mathematically relates them. This function is called the *Hamiltonian* of a system.

$$\mathcal{H}(\mathbf{s}, \Theta) \tag{2.23}$$

In analogy with its meaning in physics, where it encodes the energy of the model, the Hamiltonian here represents the global "stress" or "cost function" of the system. The state-space vector $\mathbf{s}$ will, therefore, evolve in time to minimize this quantity. If we have reasons to believe that the system explores the energy landscape at equilibrium, subject to a certain level of noise (modelled by the inverse temperature β), the Hamiltonian description becomes equivalent to a *probabilistic* one, thanks to the Boltzmann distribution (1.6):

$$P(\mathbf{s}) = \frac{1}{\mathcal{Z}} e^{-\beta \mathcal{H}(\mathbf{s},\Theta)} \tag{2.24}$$

This procedure of choosing a subset of relevant degrees of freedom, which also goes by the name of *dimensionality reduction*, poses a series of challenging practical tasks [66]. First, among all the possible variables that we might include in the model, we do not know which are the relevant ones, i.e., the ones that drive the behavior that we want to reproduce. Second, in practice, we usually can observe only a subset of the relevant variables. Third, all observations are typically noisy, which becomes critical if we have a limited amount of data. Some approaches, for example, based on information theory criteria [66, 67], have been proposed to isolate the most informative degrees of freedom in an under-sampled system.

why Once we have chosen the degrees of freedom and the parameters that control their interactions, we need to analyze the Hamiltonian $\mathcal{H}(\mathbf{s}, \Theta)$ to get a description of how the model behaves with respect to the value of its parameters.

As we saw in the first chapter, the kinetic theory of gases is analytically solvable with paper and pencil, i.e., the tool set accessible to a physicist from the 19th-century. As for today, the theoretical methods of statistical physics have developed to treat out-of-equilibrium and even strongly-interacting systems, by exact, approximated or numerical methods.

By employing these tools, we can get insights on how the values of the parameters, $\Theta = (\theta_1, \theta_2, \ldots, \theta_M)$, drive the behavior of the model. Sometimes we can use an analytical approach to derive the *phase diagram* of the model, i.e., a low-dimensional description that shows how the system separates into a small number of qualitative behaviors depending on the values of specific parameters [68, 69]. Other times we might need numerical simulations, or a combination of analytics and simulations, to get insights on the effect of parameters on the behavior.

If we are lucky, we will find a region of the parameter space in which our simplified model reproduces the investigated phenomenology. In this case, we generalize the observed behavior to an abstract set of mathematical rules, which could in principle be implemented by entirely different systems. In other words, we achieved a step towards generalization, satisfying our physicist pursuit for simplification and synthesis.

If we are *very* lucky, our simplified model will also display some unexpected behavior that is successively verified by further analysis of the real system. In this fortunate case our model is not only *descriptive*, but also *predictive*. A predictive model provides strong

support to the hypothesis that the modelled mechanism is factually implemented by the real system, therefore shedding light on the underlying principles that rule the analyzed phenomenology.

Overall, the rationale behind our statistical-physics approach to biological systems can be summarized as: assume a mechanism as the explanation for an observed complex behavior; design a model that relates some coarse-grained degrees of freedom of the system with a set of parameters; work out the model to see if our mechanism is descriptive and, hopefully, predictive for the complex phenomenology of the real system.

Part II

NAVIGATION AND MEMORY IN THE HIPPOCAMPAL COMPLEX: MODELLING AND INFERENCE OF ATTRACTORS FROM NEURAL RECORDINGS

3 BACKGROUND

3.1 NAVIGATION AND MEMORY IN THE HIPPOCAMPUS

3.1.1 *The hippocampus*

The hippocampus is a region of the mammalian brain located in the medial temporal lobe, part of the so-called limbic system of the brain (Fig. 3.1). All vertebrate species, including reptiles and birds, have a homologous region. The hippocampus is one of the brain regions that attracted the most interest from psychologists and neuroscientists, mostly due to its crucial role in spatial navigation and episodic memory.

The first evidence of an involvement of the hippocampus in memory processes was observed in the well-known case of the patient H.M., who suffered from severe anterograde amnesia after he received a bilateral hippocampal ablation as a treatment for epilepsy [71]. This first case has since been followed by a substantial number of observations that confirmed a strong correlation between hippocampal lesions and impairment of the formation and consolidation of *declarative* (i.e., involving conscious recall) memories in human patients [72, 73]. Experiments on monkeys showed that the hippocampus is crucial for the formation and recall, but not for the storage, of memories [74].

In rodents, the hippocampus has been extensively studied for its role in spatial navigation and spatial memory, i.e., the process of storing and recalling a *cognitive map* of an environment. The term "cognitive map", coined by the American psychologist Edward Tolman [75], refers to an allocentric representation of the surroundings embedded in a Euclidean metric, which enables navigation through the cognition of the spatial distances between locations and objects. Tolman himself gave the first striking evidence for map-based navigation in rodents in a famous experiment that showed the ability, in rats, of elaborating shortcuts to a known reward position [76]. The role of the hippocampus in the formation and retrieval of cognitive maps in rodents has been demonstrated by numerous experiments typically involving the memorization of spatial locations. One notable example is the Morris water maze: the rat, who is an able swimmer but dislikes being in the water, is trained to swim to a hidden platform in a specific location within a pool of milky water. A healthy animal quickly learns the position of the platform, showing an average time to goal that sharply decreases with time [77]. Morris and colleagues compared this performance profile to the one resulting from animals whose hippocampus had previously been lesioned, showing a significant drop in performance for the damaged group compared to the control one [78].

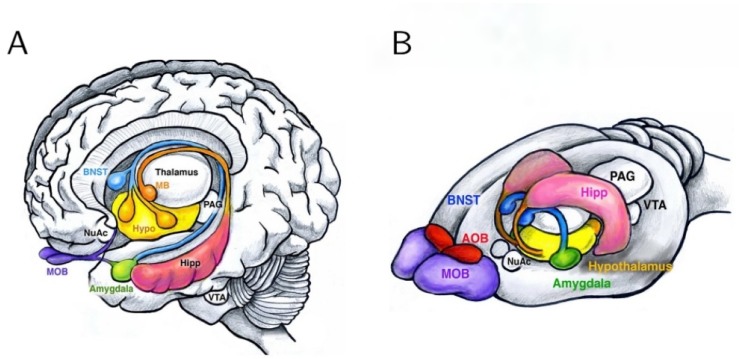


Figure 3.1: **The limbic system in human and rodent.** Main structures of the human and rodent limbic system. (A) Human brain showing the amygdala (green), bed nucleus of stria terminalis (BNST, blue), hypothalamus (yellow), and hippocampus (pink). The hippocampus (pink) attaches to the mamillary bodies (orange) through the fimbria-fornix. Olfactory inputs are received by the olfactory bulbs (MOB, purple). Other structures include the nucleus accumbens (NuAc), ventral tegmental area (VTA), and the periaqueductal gray (PAG). (B) Similar structures are found in rodents. Figure and caption adapted from [70]

3.1.2 *Place cells*

The connection between the hippocampal region and spatial navigation received remarkable support in 1971 when O'Keefe and Dostrovsky discovered a population of hippocampal pyramidal cells which fired only when the animal crossed specific regions within an environment [79]. The sharp firing specificity of these neurons granted them the name of "place cells", and their spatial receptive fields were named "place fields".

Decades of subsequent research aimed at characterizing place cells properties in diverse contexts. In a given environment, place fields corresponding to different place cells are centered around different locations, and the whole population can globally cover the full surroundings [80, 81]. An essential property of place cells is that their firing specificity is stable over time, such that after a period of exploration of an environment the same place fields are observed if the animal is placed in the same settings even after weeks [82]. Place fields are also resistant to small perturbations and continuous transformations of external landmarks [83–85]. The specific positioning of place fields are however flexible and might shift or entirely re-arrange upon drastic changes in external landmarks and boundaries [84, 86], odors [87], or even abstract variables such as contextual conditions or the task to be performed [87, 88]. For different environments, place fields can either re-position in a supposedly random way, a property called "global remapping" [83, 84, 89, 90], or keep the same spatial positioning and change their mean firing rate, called "rate remapping" [91]. An example of remapping in CA3 and CA1

sub-regions of the hippocampus is shown in Fig. 5.1. Thanks to said properties, namely spatial selectivity, stability over time, and discrimination of different environments, the place-cells population has been proposed as a suitable candidate for the neurological basis of the cognitive map [92].

3.1.3 *Head-direction cells*

Place cells are not the only neurons that display a sharp spatial selectivity in their firing properties. Soon after the discovery by O'Keefe and Dostrovsky, neurons in the septal presubiculum were shown to respond to specific orientations of the head of the animal [95, 96], hence called *head-direction cells*. Cells responsive to the direction of the motion have later been discovered in other regions, such as the entorhinal cortex [97], the anterior and lateral dorsal thalamic nuclei [98], the lateral mammillary nucleus [99], the retrosplenial cortex [100] and the striatum [101], suggesting that the directional signal could be computed in brain regions external to the hippocampal formation [102]. Since HD cells fire allocentrically and depending only on the ongoing direction of the animal (not on the specific location within the environment) they have been interpreted as the "compass" used for navigation in the cognitive map [96]. HD cells primarily rely on external landmarks to represent the motion direction [103], although they are known to respond to self-motion cues [104] as well as contextual conditions [105] when visual information is unavailable or unreliable [86].

3.1.4 *Grid cells and path integration*

Recently, E. and MB. Moser discovered neurons in the medial entorhinal cortex (MEC) that exhibit an hexagonal, periodic grid-like spatial selectivity, hence denominated *grid cells* [85, 106]. Grid cells are characterizable by the period and the orientation of the grid. The period, or grid spacing, is similar for cells that are nearby in the MEC (a property called "topography") and increases along the dorsoventral axis of the cortex [107]. A crucial property of grid cells is that they do not change their mutual relation in different environments, i.e., the superposition of their firing fields is constant independently on external conditions, as the firing grids coherently shift and rotate across different familiar environments [108].

The context-independency of relative spatial encoding of grid cells, in contrast to the more elaborate variability of place cells, led to their interpretation as a putative substrate for the representation of a universal metric for navigation [85, 109–111]. A context-independent metric is necessary to enable *path integration*, i.e., the process of updating one's cognition of self-location based on the estimation of linear and angular direction and velocity from proprioception and vestibular information, which allows for navigation in a known environment even in the absence of visual guidance. Also known as "dead reckoning", path integration is implemented by species from all the animal reign, such as ants [112], bees [113], spiders [114], birds [115], rodents [116], and humans [117].

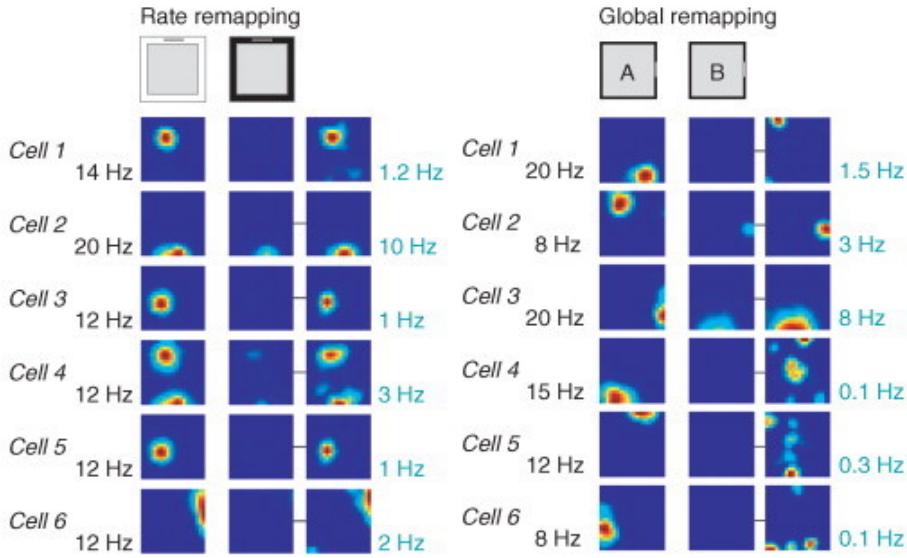


Figure 3.2: **Global remapping and rate remapping.** Colour-coded rate maps showing rate remapping and global remapping in six pairs of CA3 place cells (dark blue = 0 Hz; red = maximum firing rate, as shown on the far left and right of each row). Rats were tested in boxes with a different colour configuration in a constant location (rate remapping) or in identical boxes in different locations (A and B) (global remapping). In each panel, the left column shows rate maps for the condition where the cell had the highest peak rate (black or white in the rate remapping condition; room A or room B in the global remapping condition). Peak rates are indicated to the left. The middle column shows rate maps for the same cells in the condition with a lower peak rate. The scale is the same as for the left column. The right column contains the same data as the middle, but the colour maps are now scaled to their own maximum values (indicated to the right of each map). Note that firing locations remained constant in the rate remapping condition, whereas the intensities of firing differed strongly. In the global remapping conditions, both firing locations and firing rates were changed. Figure and caption from [93], modified from [94].

3.1.5 *What inputs drive the firing of place cells?*

The complex response to different modalities of place-cells firing, combined with the known involvement of the hippocampus in episodic memory, suggests that the sharp sensitivity of place cells to spatial position might reflect the relevancy of location as an essential feature, for rodents, of events that are represented and memorized as points belonging to a more general contextual space.

Over the last decades, a significant amount of research has been carried to probe and isolate the effect of different sensory modalities on the firing properties of place cells. As previously mentioned, place cells respond to visual stimuli such as the presence and orientation of a distal cue [84, 89, 118]. It has been shown that place fields strongly depend on the positioning of environment boundaries: they get stretched proportionally to changes in the geometry of the environment, and the positioning of a border in the middle of the room causes a double place field to emerge at the same relative location with the original border [119–121]. Neurons that primarily respond to the presence of borders, called *boundary-vector cells*, have been identified in the subiculum [122], and have been proposed as a plausible upstream source of visual information to hippocampal place cells [123]. Together, these results suggest that visual information is an essential factor in determining the firing of place cells.

Another critical determinant for the activity of place cells is path integration [124–126], possibly carried by the grid population in the MEC and projected to the hippocampus via perforant paths to the dentate gyrus [127–129]. The hippocampal cognitive map, defined as the set of place field in a known environment, has been shown to be stable in the dark, providing evidence for self-motion-based navigation in the hippocampal population. An ingenious tool to isolate path integration from visual information is to create a mismatch between the two modalities, either by moving a reward site [126], either by virtual-reality experiments, where the gain of the rotating ball where the rat is walking on can be controlled with respect to the virtual motion [124, 130]. When external and vestibular inputs are put into conflict, a significant subpopulation of place cells showed a response to movement independently on visual cues, while other remained anchored to visual information, suggesting that the mutual strength of the received path-integrator and visual inputs might vary from one place cell to the other [130].

A model of a Fourier-like integration of grid cells with different grid spacings to form a place field was hypothesized as a mechanism for the formation of place cells selectivity [110, 131–135]. However, as appealing as this model might sound to the theoreticians' ear, there is substantial evidence for place fields to be relatively independent to grid cell firing [129]. For example, place fields appear earlier than stable grid fields during development [136, 137], and stabilize faster than grid fields [138] during the exploration of a novel environment. Finally, the disruption of grid fields does not affect the formation of new place fields in a novel environment [139], or their recall and stability in a familiar environment [140]. Overall, it is clear that grid cells and place cells jointly contribute to the navigation task, but the precise nature of their relationship is still debated [108, 123, 129].

3.1.6 *The "teleportation" experiment*

What we described so far concerns the statistical properties of place cells firings, as these characterizations are usually performed by averaging observations over many sessions of the same experimental protocol. An interesting question concerns the dynamics of the cognitive map in the hippocampus, i.e., how the cognitive representation responds, on short timescales, to a manipulation of the received inputs.

As we saw, different contextual conditions within the same environment might correspond to different cognitive maps, which are recalled to navigate the surroundings under context-specific conditions [141]. The ability to perform a rapid change of the recalled cognitive map to differently coordinate the same input information upon variation of the context, a phenomenon called "dynamic grouping" [142], is an essential cognitive task for animals living in a complex environment. For example, the sudden appearance of a predator during foraging should cause a prompt recall of a cognitive representation that allows for rapid and precise navigation towards escape routes and shelters. The newly-recalled cognitive state should ignore other information, such as the position of food, that was of particular importance until the moment just before.

Jezek and colleagues investigated the question of the recall dynamics of different cognitive maps by the so-called "teleportation" experiment [143]. In the training session, the rat was let free to explore, in the dark, two boxes (A and B) that are identical in shape and differ by placement of light cues on the top. By doing that, the animal is supposed to rely on these light conditions during navigation in the cognitive map. The experiment was conducted by connecting the two boxes by a long corridor. This way, the animal was shown to form two separate cognitive maps for the two environments by global remapping (see Fig. 3.3).

In the test session, the rat was placed to explore one box, say A, and after some minutes the light conditions were abruptly switched to box B, effectively "teleporting" the rodent from one box to the other. What Jezek and colleagues found was that the cognitive map did not immediately switch to the representation corresponding to the new external conditions. Instead, in the immediate proximity of the light switch, the cognitive state oscillated between the two maps before reaching a stable representation after a few seconds (see Fig. 3.4). This oscillatory behavior was interpreted as a sign of an attractor dynamics taking place in the network of place cells of the hippocampal subregion CA3, possibly caused by the mismatch of suddenly-changed visual inputs and a path-integrator input that delays to update its internal representation (e.g., the shift and rotation observed in [108]).

As we will show in Chapter 6, a theoretical attractor model that incorporates these ingredients accurately reproduces the flickering behavior observed in [143]. By application of a position-independent method for decoding of the cognitive map, which will be the object of Chapters 4 and 5, we will show that a set of novel predictions of the model, concerning the precision of self-location during flickering and stable conditions, can be verified by re-analysis of the original data, adding evidence in support of the attractor picture.

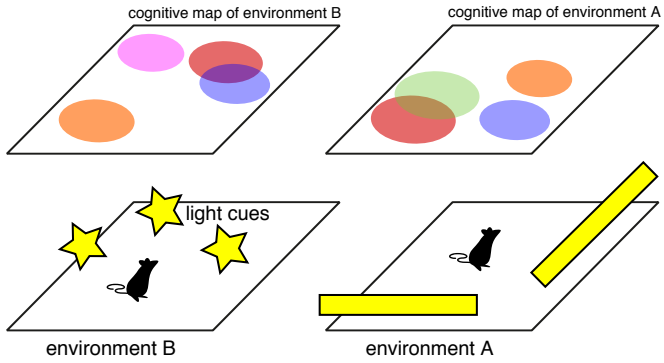


Figure 3.3: **The teleportation experiment: reference sessions.** The rodent is trained, in the dark, to recognize two boxes that are identical in shape and differ by placement of visual light cues on the top of the box. The training is conducted such that global remapping occurs and two separate cognitive maps are formed for the two environments.

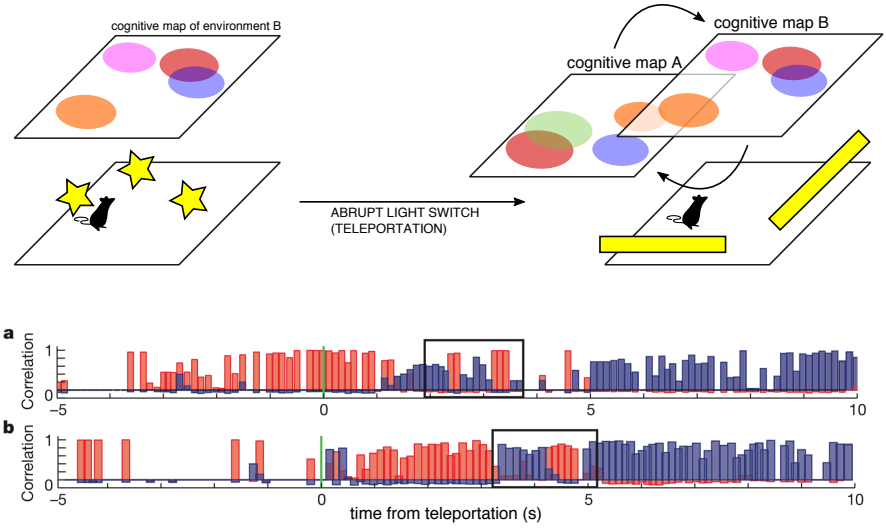


Figure 3.4: **The teleportation experiment: test session.** After the abrupt switch of visual context the internal representation of the environment oscillates between the two training boxes A and B, identified by analysis of the correlation (represented in red and blue for the two maps) between the firing activity in a theta bin and the rate vector registered at the rodent's position in the two reference sessions. Bottom panel from Jezek et al. [143].

3.2 MEMORY AND ATTRACTOR NEURAL NETWORKS

3.2.1 *The Hebbian theory of memory*

The mechanism by which the brain forms new memories was first theorized by the Canadian psychologist Donald Hebb and goes by the name of "Hebbian theory of learning". The theory is often summarized as "neurons that fire together wire together": the synaptic strength between two neurons is reinforced if the pre-synaptic and the post-synaptic neurons are activated simultaneously, e.g., from a common input. In Hebb's own words:

> *Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability . . . When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.* [144]

Since Hebb's original formulation, many studies have investigated the physiological basis of synaptic potentiation. The long-lasting strengthening of the synaptic connection between two neurons is called long-term potentiation, or LTP [145, 146]. In the hippocampus, the best-known mechanism that enables LTP is the transduction of electrical signals into chemical ones that activate the potentiation mechanisms in both the pre-synaptic and post-synaptic neurons, mediated by the N-methyl-D-aspartate (NDMA) receptor complex [147]. Selective inactivation of NDMA receptors was shown to significantly affect the performance of tasks involving spatial memory in rodents, for example in the Morris water-maze experiment [148], confirming the role of LTP in spatial-memory consolidation in the hippocampal complex [149–151].

3.2.2 *The Hopfield model*

In the early 80's, J. Hopfield formalized Hebb's visionary intuition in the celebrated Hopfield model for storage and retrieval of zero-dimensional (i.e., points in the state space) memories [152]. The idea behind the Hopfield model is that a network of N binary neurons $s_i \in \{0, 1\}$, interacting by a matrix J_{ij} via the coupled Hamiltonian

$$\mathcal{H}(\mathbf{s}) = -\sum_{i<j} J_{ij}\, s_i s_j \quad , \tag{3.1}$$

can store a set of P random N-binary patterns $\boldsymbol{\xi}^k = (\xi_1^k, \xi_2^k, \ldots, \xi_N^k)$ in their connection matrix J_{ij} by following Hebb's rule of learning, i.e.,

$$J_{ij} = \frac{1}{N} \sum_{k=1}^{P} \xi_i^k \xi_j^k \quad . \tag{3.2}$$

The network was studied in its dynamical properties, either noiseless (in the original paper) either by including stochasticity through a non-zero temperature in the Gibbs measure of the Hamiltonian in Eq. 3.1. The patterns $\boldsymbol{\xi}^k$ were shown to be *point attractors*, i.e., stable states, of the dynamics of the network, hence achieving their memorization. Each stable point was shown to correspond to a basin of attraction that leads the activity to the memorized state, allowing the network to perform pattern completion, a well-known feature of memory. Despite its simplicity, the Hopfield model attracted considerable interest from the statistical physics community, since it reproduced non-trivial features of memory systems and captured essential properties of neurons (e.g., the thresholded linear sum of the inputs) while still being analytically treatable by tools from the mathematics of disordered systems. A few years later, Amit and Sompolinsky applied these tools to characterize the phase diagram of the model, showing that the network can store up to αN patterns, N being the number of interacting neurons, before affecting the stability of the system [153]. The seminal book from Amit [153] still represents, at the present day, an essential milestone in the theory of attractor-based memory models, which goes under the name of Attractor Neural Network (ANN).

3.2.3 *CANN: Continuous-Attractor Neural Network*

As we just saw, the memorized states in the Hopfield model are single patterns of activity separated by energy walls which define the corresponding basins of attraction. One could be tempted to extend this model to the hippocampal spatial memory of an environment, by representing each position $\mathbf{r} = (x, y)$ as a pattern of activity where the active neurons are the place cells whose fields are centered in the proximity of $\mathbf{r}$. However, the self-location within the cognitive map is a continuous variable: the population activity should be able to seamlessly shift from one represented position $\mathbf{r}$ to any nearby location $\mathbf{r} + \delta\mathbf{r}$. Therefore, such an extension of the Hopfield model should involve correlated patterns that collectively form a continuous valley, instead of a set of isolated minima, in the energy space. By doing so, the network activity would be free to move on said iso-energetic manifold, or *chart*, that represents a "continuous" attractor of the network dynamics.

Following the success of the Hopfield model, a significant amount of theoretical work has been carried to develop and study such continuous extensions of the model. The ensemble of theoretical efforts carried in this direction by physicists, mathematicians, and neuroscientists goes by the name of Continuous-Attractor Neural Networks, or CANN. The underlying idea of CANN applied to the cognitive map in the hippocampus is that neurons that have overlapping place fields will display a correlated spiking activity when the rat walks through the overlap region. As a consequence, these neurons will reinforce their mutual synaptic strength by Hebbian learning. The cognitive map is, therefore, stored in the connectivity matrix by a connection J_{ij} that is proportional to the place-field superposition of neuron i and neuron j. The first model based on these ideas was presented by Tsodyks & Sejnowski [154]. In their model, place fields of N place

cells s_i are lined up on a one-dimensional track, each centered at a position x_i, and the connection J_{ij} are set as

$$J_{ij} = J_0 \exp\left(-\frac{|x_i - x_j|}{\sigma}\right) - J_1 \tag{3.3}$$

Where J_1 acts as global inhibition term, necessary to keep a stable global activity. The dynamics of the resulting spiking network was shown to display a set of attractor states, each represented as a coherent *bump* of activity, in the one-dimensional place-field space, peaked around a different position x (Fig. 3.5, left panel). Such bump was shown to be free to move along the linear track without the need for high energetic jumps, effectively reproducing a continuous (1D) attractor state.

The idea of a bump of coherent activity as the neural correlate of self-location in the cognitive map was soon generalized to two-dimensional and multiple maps by Samsonovich & McNaughton [155], who proposed a mechanism for path integration in the attractor network enabled by *shifter cells* that displace the bump according to vestibular inputs [109, 155]. The learning rule proposed in [155], as an extension of the Hopfield model, sets the coupling as a linear sum of the contributions of each single map $m = 1, \dots, M$:

$$J_{ij} = \sum_{m=1}^{M} \exp\left(-\frac{||\mathbf{r}_i^m - \mathbf{r}_j^m||^2}{\sigma^2}\right) \quad , \tag{3.4}$$

where $\mathbf{r}_i^m$ is the location of the place-field center of the cell i in the map m. When one cognitive map is retrieved, the activity is organized as a coherent $2D$ bump on the corresponding chart, while looking scattered and uninformative on all the other charts (Fig. 3.5, right panel).

Successive works have further characterized the statistical and dynamical properties of these attractor models [156–160], for example by computing the number of possible charts that can be stored in the network [12]. Over the last two decades, the CANN framework has been been proposed as a model for the stability of several phenomenologies in different regions of the brain, such as population code in the visual cortex [161,162], motor control [163, 164], parametric working memory [165, 166], as well as neural integration in spatial-responsive populations such as head-direction cells [103, 167, 168] and grid cells [131, 169]. Thanks to accumulating evidence of attractor-like dynamics in various brain regions [111, 170–172], the CANN framework is today broadly recognized as a plausible mechanism for stable neural representations at the population level [109, 111].

3.3 OUTLINE OF THE FOLLOWING CHAPTERS

In the next chapters, we will apply methods inspired from the the CANN paradigm to synthetic and real neural hippocampal data.

In Chapter 4, we will test the Ising inference (Section 2.1) as a decoder aimed at retrieving which attractor state is expressed from the in-silico activity of a small subsample of place cells, generated from the multi-charts attractor model of Monasson & Rosay [12].

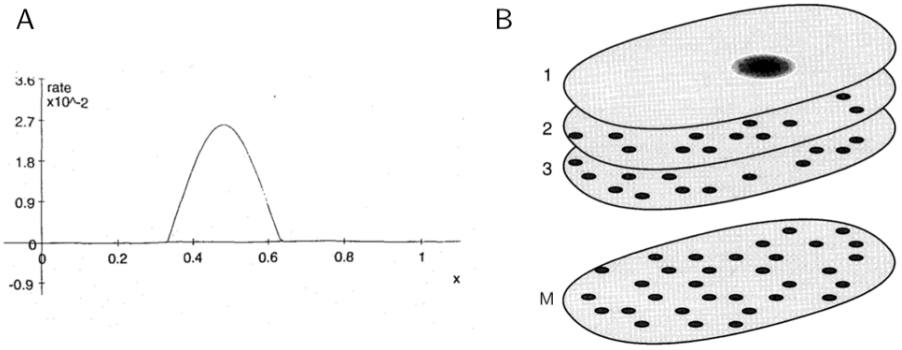


Figure 3.5: **Bump of activity in attractor models.** **A** one-dimensional continuous attractors of Tsodyks & Sejnowski. The profile represents the firing rate of all the neurons in the network in one of the stable attractor state. Each neuron i is labeled by its corresponding place-field center's position x_i. Figure and caption adapted from [154]. **B** two-dimensional multi-chart model of Samsonovich & McNaughton: the set of M charts, composed of the same neural units. Activity that is well localized on one of the charts (chart 1 in the picture) looks scattered on other charts (2, 3, M). Figure and caption adapted from [155].

We will then characterize the relationship between the inferred functional couplings and the "real" ones.

Chapter 5 is dedicated to an application of the Ising decoder to multi-array recordings in CA1, showing that it outperforms the current standards in discriminating the recalled cognitive state on short ($\sim$theta) time scales. Importantly, the Ising decoder does not need additional information (as the position of the rat) to decode the cognitive map from neural activity, opening to applications on other brain regions where there is no clear external correlate for the activity.

In Chapter 6 we will introduce an attractor model for the teleportation experiment of Jezek et al. [143]. The memory network is subject to external and path-integrator inputs and accurately reproduces the flickering phenomenology of cognitive maps observed in the original experiment. The attractor model makes several novel predictions that are verified by a careful re-analysis of the original data through the application of the Ising decoder.

Finally, in Chapter 7, we show that a single network model inferred from hippocampal recordings of two distinct cognitive maps can be used to generate population activities that are coherent with two separate low-dimensional attractors, suggesting that the inferred functional connectivity preserves some fundamental structure of the underlying attractor network.

4

INFERRING THE ATTRACTOR STATE FROM POPULATION ACTIVITY: THE "SUBSAMPLING" PROBLEM

A significant part of this chapter is adapted from [5].

4.1 INTRODUCTION

The problem of decoding the firing activity of a neural population to retrieve which cognitive state is internally represented at a given time has a natural application in many experiments that involve electrophysiology or calcium-imaging simultaneous recording of several neurons [4]. Spatial memory in the hippocampus provides an example: as we saw in the previous chapter, a widely-accepted theory for spatial representation is that during navigation in a specific cognitive map the collective state of the neural population is organized as a coherent bump of activity on the corresponding chart. The population activity, therefore, follows a constrained dynamics determined by the connectivity structure and by the expressed cognitive map, exploring the corresponding continuous attractor, i.e., a low-dimensional flat manifold of the energy landscape.

As a consequence, the set of neural patterns expressed during the representation of a specific cognitive map will show statistical properties that are typical of the attractor state and, consequently, of the recalled cognitive map. In this framework, the *decoding problem* can be tackled by learning the statistical properties of each different cognitive state (in this case, the different cognitive maps) and classifying new observations accordingly, a problem which is deeply connected to high-dimensional classification in machine learning. However, in real electrophysiology applications, one typically records $O(10^{1-2})$ neurons out of the $O(10^{4-5})$ of the neural population that participates to the collective attractor dynamics. An important question, therefore, concerns the retrievability of these state-dependent statistics from neural recordings in the case of a strongly-subsampled population, an issue that we call the *subsampling problem*.

Hereafter, we describe an attempt to draw a parallel between experimental conditions of multi-array recordings and the theoretical model for spatial memory in the continuous-attractor neural network (CANN) framework. We first design a Monte Carlo simulation that mimics an experiment with two memorized environments, referred to as A and B. We simulate single-environment reference sessions by forcing the activity to explore local minima corresponding to the memorized environments in a system with a relatively large number $(N = 1,000)$ of neurons. We then address the question if a small, randomly selected, set of neurons (here, $N_{sam} = O(10)$ out of $1,000$) could provide enough information to perform the decoding procedure and infer the time course of the spatial representations from neural activity. As place cells are non-topographical,

i.e., cells that are physically nearby in the hippocampus can have distant place fields, recording a spatially located population of cells can be thought of to be equivalent to a random subsample in the place-field abstract space.

We approach the decoding problem on the test session using Ising and independent models learned from the two reference sessions and test their decoding capability in a binary classification problem on a test session composed by samples from both states. Finally, we investigate the relationship between the real ("microscopic") couplings and the inferred functional connectivity, showing that this last preserves the spatial correlate of the microscopic connectivity.

4.2 METHODS

Monasson-Rosay CANN multi-chart model

The model that we use to generate in-silico neural activity is the one of Monasson & Rosay [12, 159, 160]. We will here briefly review its principal ingredients. For a full characterization of the statistical and dynamical properties, as well as the replica computation of its phase diagram, please refer to the cited literature.

As an extension of the Hopfield model, the model is based on binary neurons. The N place cells are modeled by binary units s_i equal to 0 (silent state) or 1 (active state). These neurons interact together through excitatory couplings J_{ij}. Moreover, they interact with inhibitory interneurons, whose effect is to maintain the total activity of the place cells to a fraction f of active cells (global inhibition). We also assume that there is some stochasticity in the response of the neurons, controlled by a noise parameter T. All these assumptions come down to considering that the network states are distributed according to the Gibbs distribution associated to the Hamiltonian of Eq. 3.1, restricted to configurations of spins $\mathbf{s}$ such that

$$\sum_i s_i = fN \,. \tag{4.1}$$

We want to store $L+1$ environments (or charts) in the coupling matrix, indexed by ℓ, each defined as a random permutation π^ℓ of the N neurons' place fields. This models the experimentally observed remapping of place fields from one map to the other. With this definition, an environment is said to be stored when activity patterns localized in this environment are stable states of the dynamics. In other words, the configurations where active neurons have neighbouring place fields in this environment are equilibrium states. To make this possible, we assume a Hebbian prescription for the couplings J_{ij} that is a straightforward extension of the Hopfield synaptic matrix to the case of quasi-continuous attractors. This rule is illustrated in Figure 4.2, and is mathematically described as follows:

- additivity: $J_{ij} = \sum_{\ell=0}^{L} J_{ij}^{\ell}$ where the sum runs over all the environments.

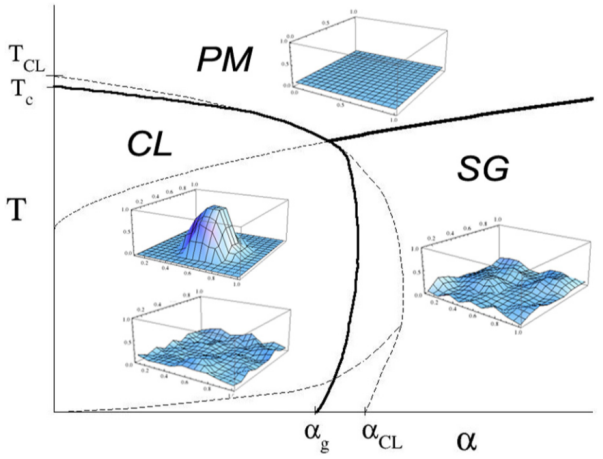


Figure 4.1: Sketch of the phase diagram in the plane of neural noise, T, and number of environments per neuron, α. Thick solid lines: transitions between phases. Thin dashed lines: stability region of each phase against fluctuations. Insets show the corresponding activity profiles in the $2D$ model (averaged over one round of Monte Carlo simulations after thermalization). In the clump phase we represent the same activity profile in the retrieved environment (top), and in another stored environment (bottom). Figure and caption adapted from [158].

- potentiation of excitatory couplings between units that may become active together when the animal explores the environment:

$$J_{ij}^{\ell} = \begin{cases} \frac{1}{N} & \text{if } d_{ij}^{\ell} \leq d_c \\ 0 & \text{if } d_{ij}^{\ell} > d_c \end{cases}, \tag{4.2}$$

  where d_{ij}^{ℓ} is the distance between the place-field centers of i and j in the environment ℓ; for instance, in dimension $D = 1$, $d_{ij}^{\ell} = \frac{1}{N}|\pi^{\ell}(i) - \pi^{\ell}(j)|$.

d_c represents the distance over which place fields overlap. In practice, it is chosen so that, in each environment, each neural cell is coupled to a fraction w of the other cells (its neighbours); in dimension $D = 1$ again, we may choose $d_c = \frac{w}{2}$. The $\frac{1}{N}$ factor in Eq. 4.2 ensures that the total input received by a cell remains finite as N goes to infinity, a limit case in which exact calculations become possible [11].

Monasson & Rosay formally characterized the behavior of the model, in this limit case, as a function of the number of stored charts (αN) and the noise factor T. Importantly, they showed the existence of a region of the parameters space corresponding to a *clump* phase of the system ("CL" in Fig. 4.1) where the activity condensates into a coherent bump on one of the stored charts. At high noise level (high T), the system is said to be in

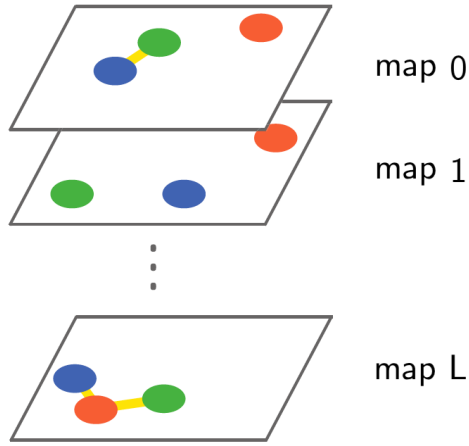


Figure 4.2: Remapping and connectivity rule in the model, illustrated with three units and $L+1$ two-dimensional environments. The place field centers of the units are displayed respectively in red, blue and green. Thick yellow lines indicate the excitatory couplings between cells with nearby place fields in each environment. These place fields overlap; here, for the sake of clarity, only the centers of the place fields are represented.

the paramagnetic phase ("PM" in Fig. 4.1), where no coherent representation is formed and the population activity is comparable to random. Finally, when the number of stored maps exceeds a critical capacity ($\alpha(T)$) the system falls into the spin glass (SG) phase, a behavior characterized by the presence of many local minima where the effective noise induced by the competition between maps freezes the activity and no spatial or map selectivity is achieved.

In the case of a single continuous attractor, the bump behaves as a quasi-particle with little deformation [159]. This quasi-particle undergoes a pure diffusion with a diffusion coefficient that can be computed exactly from first principles, i.e., from the knowledge of microscopic flipping rates of spins in Monte Carlo simulations. The bump can be driven by imposing an external force on the spins, i.e., by acting on the fields h_i of the model (see Eq. 4.5 below).

In the presence of multiple maps, the disorder in the couplings due to the additive storage creates an effective free-energy landscape for the bump of activity in the reference environment. The free-energy barriers scale typically as $\sqrt{N}$, and are correlated over space lengths of the order of the bump size, see [158]. In one dimension, the bump therefore effectively undergoes Brownian motion in the Sinai potential, with strongly activated diffusion. In higher dimension, diffusion is facilitated with respect to the 1D case. In addition to moving in the reference environment, the bump can also spontaneously jump between maps. A full characterization of the phenomenology of these spontaneous transitions can be found in [160].

Sampling and sub-sampling of in-silico activity

We used the model just described to sample generated activity from two cognitive maps, referred to as A and B. Parameters are carefully chosen such that the *clump* phase is maintained: the bump thoroughly explores the environment, and no spontaneous transitions occur. In other words, the system stays in one of the two maps during the whole simulation; this mimics a single-environment exploration of a rodent during training sessions, see for example [143]. Two *reference sessions* are defined using the first half (5,000 steps) of each simulation, and a *test session* is constructed by concatenating the second halves, for a total of 10,000 total time steps, see Fig. 4.3. Parameters used in the following analysis are: $T = 0.006$, $N = 1000$, $w = 0.05$, $f = 0.1$. Specifically, we conduct simulations as follows:

- First we define two 1D environments, hereby referred to as A and B, through their two random place-field permutations, denoted by π^A, π^B.
- From these two environments, two coupling matrices J^M, $M \in \{A, B\}$, are created using learning prescription described in Eq. 4.2:

$$J_{ij}^M := \begin{cases} \frac{1}{N} & \text{if} \quad \frac{1}{N}\left|\pi^M(i) - \pi^M(j)\right| \leq \frac{w}{2} \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{4.3}$$

- A unique coupling matrix J is then constructed as point-sum of the two single-environment matrices: $J_{ij} = J_{ij}^A + J_{ij}^B$.
- simulations are performed, with $n = 10^4$ Monte Carlo steps, each one starting from an initial neuronal condition localized in one of the two reference environments M. To maintain the total activity constant, we select, at each algorithm step, one active spin $s_i = 1$ and one silent spin $s_j = 0$. The flip trial is then defined as the joint flip of these spins.
- an additional small force is added to make the bump exhaustively explore the one-dimensional map, by an asymmetric term in the energy. This results in a left-right asymmetry in the Monte Carlo acceptance rule:

$$\Delta E = \sum_{k \neq i,j} (J_{ik} - J_{jk}) s_k + A^M(i,j) \tag{4.4}$$

  with $A^M(i,j)$ being a right-pulling force in the environment M, namely

$$A^M(i,j) := \frac{A}{fN^2} \times \left(\pi^M(i) - \pi^M(j) + N\epsilon_M(i,j)\right) \tag{4.5}$$

  where A controls the magnitude of the pulling force, $\pi^M(i)$ is the position occupied by the place field of neuron i in environment M, and $\epsilon_M \in \{-1, 0, 1\}$ ensures periodic boundary conditions.

We then select a random subset of 33 neurons out of the 1000 used for the simulation. The activity of these neurons in time are then used to construct two reference session and one test session, that will be used in the performance assessment of inference models described in the next section.

Inference of the cognitive state from subsampled activity

The patterns collected in the reference sessions are samples of neural activity constrained to explore the continuous attractor that corresponds to one memorized cognitive map. We can, therefore, use these samples to learn two statistical models, one for each map, that can, in turn, be embedded in a Bayesian framework to classify new patterns from a test session. Formally, we need to infer a probability density function over the neural patterns $\mathbf{s}$ for each brain state M, $P(\mathbf{s}|M)$. These probability distributions can be used to decode the internal state M given an observation (a neural pattern) $\mathbf{s}$ in the test session by maximizing the log-likelihood

$$\mathcal{L}(M|\mathbf{s}) = \log P(\mathbf{s}|M) \ . \tag{4.6}$$

This inference framework relies on the definition of a parametric probability function, whose parameters are inferred by solving the corresponding *inverse problem* from reference data. According to the max-entropy principle, our choice is to use the family of graphical models [21, 22, 30] as parametric probabilistic functions. Depending on the reference sample size and the complexity of representations we can invert the Independent model, which accounts for the different average activations of neurons in different brain states, or make a step further and include correlations between neuron activities, defining an Ising model for each state M.

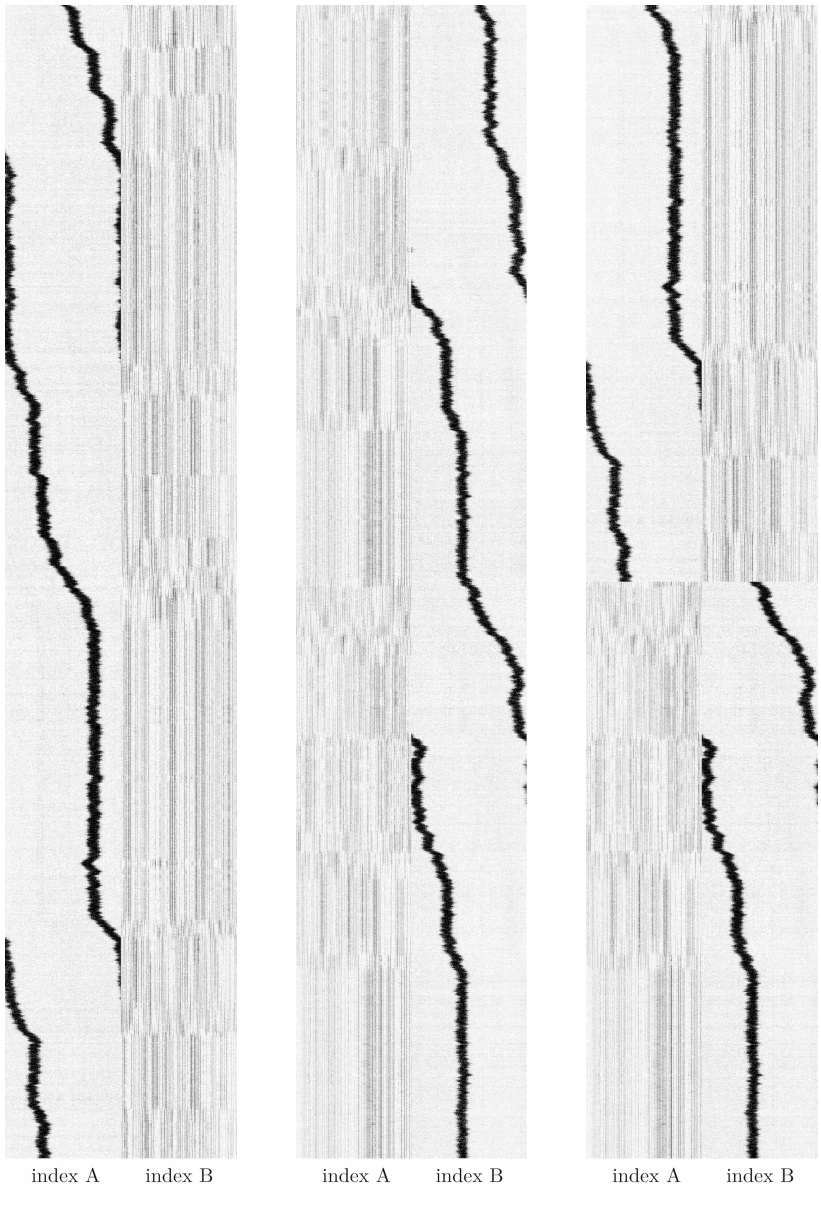
$$P(\mathbf{s}|M) = \frac{\exp\left(\sum_i h_i^M s_i + \sum_{i<j} J_{ij}^M s_i s_j\right)}{\mathcal{Z}^M(h,J)} \tag{4.7}$$

where $\mathcal{Z}^M$ is the normalization constant ("partition function"). The core steps of the Ising decoding procedure are:

Train For each brain state M, (a) collect samples of neural pattern in a known brain state M (reference session), and compute the frequencies p_i^M and pairwise joint frequencies p_{ij}^M of the recorded neurons; (b) find the Ising model that reproduces the same quantities on average, i.e. such that $\langle s_i \rangle = p_i^M$ and $\langle s_i s_j \rangle = p_{ij}^M$, where $\langle \cdot \rangle$ denotes the average over the probability distribution $P(\mathbf{s}|M)$. This is a highly non-trivial computational problem, reviewed in Section 2.1.

Test Given a neural pattern from the test session $\mathbf{s}^t$, compute the log-likelihood of each brain state, and decode the internal state as the most likely one

$$M^t = argmax_M \ \mathcal{L}(M|\mathbf{s}^t) \ . \tag{4.8}$$



(a) Session A (b) Session B (c) Test Session

Figure 4.3: Monte Carlo simulation sessions of our memory model in the case of two 1D environments (random permutations), denoted by A and B. X-axis: states of the system $\mathbf{s}(t)$ (black dots correspond to active neurons $s_i = 1$ and white dots to silent cells, $s_i = 0$), with neurons ordered in increasing order of their place field centers in the A (left part of columns) or B (right part of columns) permutations. Y-axis: time in MC rounds, increasing from top to down. The bump is forced to move rightwards with an external force, see [158]. In columns **(a)** and **(b)**, the system is initialized with a localized bump of activity in environments, respectively, A and B. Column **(c):** Test simulations composed of the second halves of simulations reported in **(a)** and **(b)** used for decoding purposes, see text. Parameter values: $T = 0.006$, $N = 1000$, $w = 0.05$, $f = 0.1$.

This framework, therefore, allows for the decoding of the neural representation from the observed neural pattern only. The same procedure has been applied to experimental data from the hippocampus, showing good performance in retrieving the explored environment from neural activity [1, 2], and to other brain regions, see for instance [48, 51, 173, 174].

4.3 RESULTS

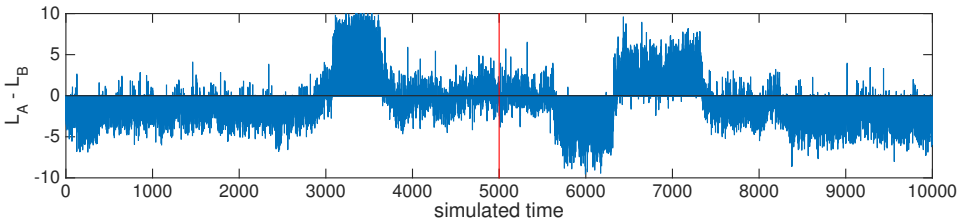
Decoding the cognitive map

As a measure of decoding precision we use the true positive rate (TPR), i.e. the overall fraction of correctly-classified neural pattern. By applying the Ising and the independent models to the $10,000$ patterns in the test session we obtain:

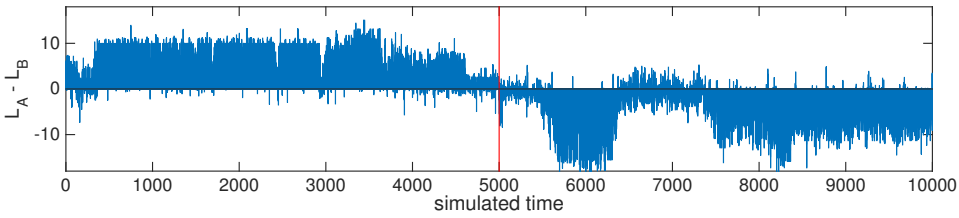
$$\begin{aligned} \text{Ising model}: \ \text{TPR} &= 0.928 \\ \text{Independent model}: \ \text{TPR} &= 0.491 \end{aligned} \tag{4.9}$$

The difference between the two models, shown in Fig. 4.4, is remarkable. The independent model, in which all couplings are set to zero, accounts only for the average firing rates of the cells. It shows no decoding capability at all, with a TPR compatible with random guessing. This difference could be expected from the fact that the localized bump of activity, which represents the position of the rat within the retrieved a map, moves along the entire environment during reference sessions. Hence the average activity of all cells is close to f in both maps. The independent model, which only uses information on averages to decode the activity, is, therefore, unable to achieve useful discrimination. Conversely, the Ising model exhibits an impressive performance in the decoding task. As shown in Fig. 4.4b, the time course of the likelihood difference $\Delta\mathcal{L}$ allows us to decode the spatial representation as a function of time unambiguously. This difference is also clear from the scatter plot of the likelihoods in the test session, which shows a well-separated pattern in the plane, contrary to the Independent model (Fig. 4.5).

One natural question is how the performance of the Ising model scales with the number of subsampled neurons N_{sam}. If we assume translational invariance of the problem, we can approximate the TPR with the probability that a pattern generated when the bump is at a given position, say $x = 0$, in one of the two maps, say A, is correctly classified by the Ising model. As a first approximation, let's assume that every neuron inside the bump, i.e., positioned in the interval $[-N\frac{f}{2}, N\frac{f}{2}]$, is active. We treat the subsampling problem as a draw of N_{sam} i.i.d. random positions in the map. To further simplify the classification problem, we assume that every time two or more neurons out of the sampled N_{sam} fall inside the active bump (i.e., in the interval $[-N\frac{f}{2}, N\frac{f}{2}]$ in map A), the map-specific coupling terms in the likelihood will give us enough information to decode the map correctly. Conversely, in the case of only 1 or 0 neurons sampled inside the bump, we will assume that the guess of the environment is equivalent to random chance.

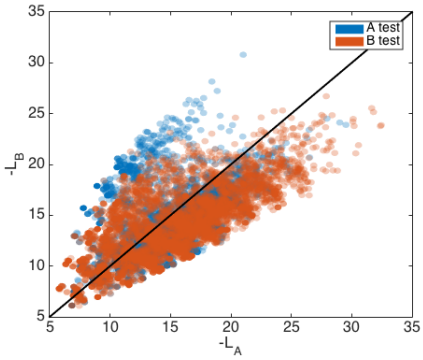


(a) Independent model

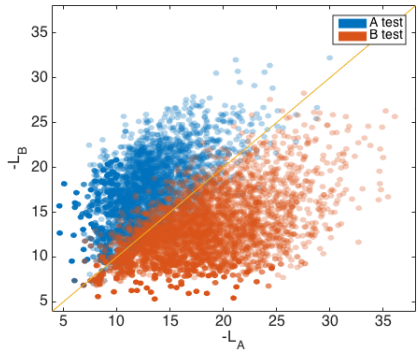


(b) Ising model

Figure 4.4: Log-likelihood difference $\mathcal{L}_A(t) - \mathcal{L}_B(t)$ along the test session using independent model and Ising model on the montecarlo test session. The first half of the test session is sampled from environment **A**, the second half from environment **B**.



(a) Independent model



(b) Ising model

Figure 4.5: Log-likelihood scatters computed from the Independent (a) and Ising (b) models. Each dot represents the value of $-\mathcal{L}_A$ and $-\mathcal{L}_B$ for each neural configuration $\mathbf{s}^t$ during the Monte Carlo test session.

Therefore, defining N^A_{bump} the number of neurons in the bump of environment A, we will assume

$$TPR(N_{sam}) = \begin{cases} \frac{1}{2} & \text{if } N^A_{bump} = 1 \text{ or } 0 \\ 1.0 & \text{if } N^A_{bump} \geq 2 \end{cases} = 1 - \frac{1}{2} \cdot P(N^A_{bump} < 2) \tag{4.10}$$

The problem is therefore reduced to the easy computation of the probability of having N^A_{bump} smaller than 2

$$P(N^A_{bump} < 2) = (1-f)^{N_{sam}} + fN_{sam}(1-f)^{N_{sam}-1} \quad , \tag{4.11}$$

that easily gives the predicted form for the TPR of the Ising model as a function of the bump width f and the number of sampled neurons:

$$TPR(N_{sam}, f) \simeq 1 - \frac{1}{2}(1 - f - f \cdot N_{sam})(1-f)^{N_{sam}-1} \quad . \tag{4.12}$$

This form has an inflection point at

$$N^*_{sam} = 1 - \frac{1}{f} - \frac{2}{\log(1-f)} \simeq \frac{1}{f} \quad , \tag{4.13}$$

that gives us a gross estimate of the number of neurons that we should sample, given the mean activity f, in the absence of noise. As for our case, we know that the noise is proportional to f^2. We therefore use $\tilde{f} = f(1-f)$ that gives us the estimation of $N^*_{sam} \simeq 11$. As shown in Fig. 4.6, this simple theory is well verified by numerical results.

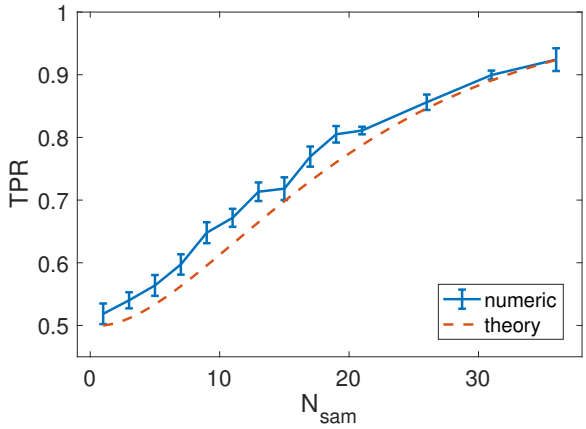


Figure 4.6: **Dependence of Ising TPR from the number of sampled neurons.** Numerical estimations for a given N_{sam} are obtained by performing n different subsamplings and performing the TPR test on the same reference/test sessions described in text. The errorbars represent the standard error on the mean obtained on a variable number (from $n = 3$ to $n = 25$) of repetitions, depending on the value of N_{sam}.

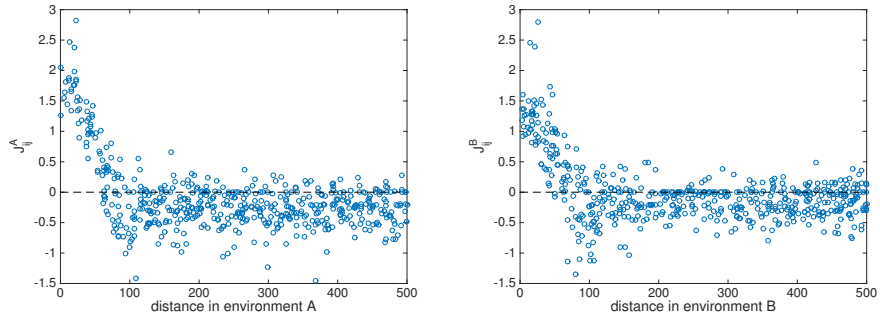
Inferred vs. true couplings

The application of inference routines to a simulated neural network allows us to investigate the relationship between functional couplings, i.e., the inferred J_{ij} in the inverse

Ising model, and the real coupling strength, defined in Eq. 4.3. We show in Fig. 4.7 the couplings inferred between the neurons as functions of the distances between their place-field centers in each map. We observe that:

- Couplings decay very rapidly with the distance, on a typical scale compatible with both $w\,N$ and $f\,N$, and the width of the bump; Note that w, f have similar values in the simulations. At long distances, couplings are independent of distance and equal to a negative value. The presence of many long-range inhibitory couplings, clearly visible in the histograms of Fig. 4.8, is a natural consequence of the constraint in Eq. 4.1 on the level of activity.

- The magnitude of coupling at small distances, $\sim 2-3$ in Fig. 4.7, is much larger than the one of the 'true' couplings in the model, equal to $J^0 = \frac{1}{TN} = 0.167$. This discrepancy highlights the *effective* nature of the inferred couplings, which would coincide with the true couplings only in the limit of perfect spatial sampling ($N_{sam} = N$).



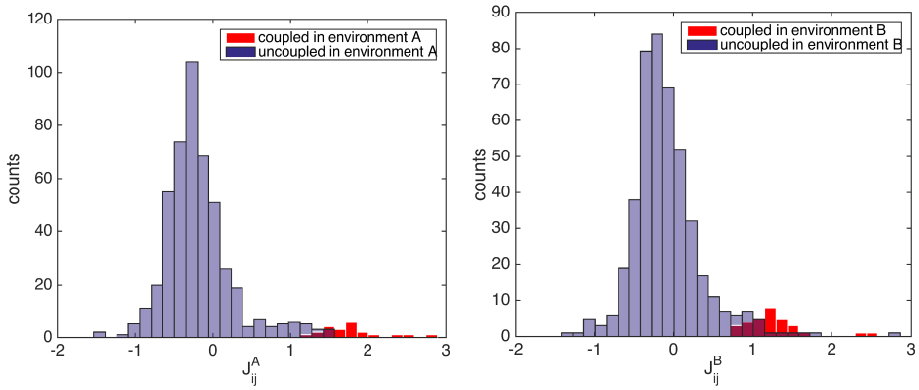
(a) Environment A (b) Environment B

Figure 4.7: Inferred coupling J_{ij} vs. distance $\left|\pi^M(i) - \pi^M(j)\right|$ between the place-field centers of the corresponding neurons in environment $M = A$ (left) and $M = B$ (right).

To better understand the relationship between the inferred and the real couplings we can compute the statistical moments of the neurons in the CANN model. As explained above, by forcing the bump to homogeneously explore the whole space at an almost-constant speed we cause all neurons to have the same average activity in both environments:

$$p_i^A = p_i^B = f\,, \quad \forall i\,. \tag{4.14}$$

In the clump phase of the model, we can also estimate correlations from the joint probability that two neurons are active. In the large-N limit, the true couplings vanish as they scale as $1/N$. Hence, spins become two-by-two independent in a ground state of the



(a) Environment A (b) Environment B

Figure 4.8: Relationship between true couplings and inferred couplings. In purple, historgram of inferred couplings. In red, inferred couplings corresponding to truly connected neurons in the environment.

Hamiltonian, that is when the bump is centered around a given position x. Conditioned to x, and defining the position of the place-field center of cell i in map M through

$$x_i = \frac{\pi^M(i)}{N} \,, \tag{4.15}$$

we have

$$\langle s_i \rangle_x = \rho(x_i - x), \ \langle s_j \rangle_x = \rho(x_j - x), \ \langle s_i s_j \rangle_x - \langle s_i \rangle_x \langle s_j \rangle_x \sim \frac{1}{N} \,. \tag{4.16}$$

In the above expression, we implicitly centered ρ at $x = 0$. However, we have to average over the position x of the bump that moves across the environment (Fig. 4.3). Doing so, we obtain the pairwise activity, see Eq. 36 and Fig. 12 in [160]:

$$p^M_{ij} = \int dx \, \rho(x_i + x) \, \rho(x_j + x) \quad \forall i, j \,. \tag{4.17}$$

This effective matrix of pairwise activities, therefore, depends on the map, which explains why the Ising model, contrary to the Independent model, is map-specific and can efficiently decode the representation. However, the effective correlation between neurons, $p^M_{ij} - f^2$, does not scale as $\frac{1}{N}$: the Ising couplings are thus effective interactions, not simply related to the true couplings in the model. Since this discrepancy follows from the fact that only a small fraction of neurons is observed and included in the inference, a scenario that is very frequent in real applications, we expect this statement to hold also for the functional couplings inferred from real recordings and their physiological, synaptic counterparts.

5

FUNCTIONAL CONNECTIVITY MODELS FOR DECODING OF SPATIAL REPRESENTATIONS FROM HIPPOCAMPAL CA1 RECORDINGS

This Chapter was published by myself, S. Cocco, and R. Monasson, in collaboration with Karel Jezek from Charles University, in [1]. It focuses on an application of the Ising inference to the task of decoding the represented cognitive map on short time scales (ranging from ~ 10 to ~ 1000 ms) from population activity recorded in the hippocampal region CA1. The Ising decoder is compared to the current standards, which rely on the similarity between the activity vector and two reference vectors computed at the specific location of the animal during reference sessions (e.g., used by Jezek et al. [143]). The Ising model is shown to outperform the other models on all time scales. Crucially, the Ising decoder relies only on the correlation and average activations observed during the reference sessions. Therefore, it can decode the cognitive map without the need of knowing the precise value of an external correlate (in this case, the position of the animal), opening to applications on other brain regions where there is no obvious physical correlate of the neural activity. Finally, the proposed decoder is applied to CA1 data from the teleportation experiment [143], where contextual cues are switched abruptly (hence the name "teleportation") to trigger an instability of the recalled cognitive map. We report a long-term instability of the post-teleportation map that persists over all the session. This long-term effect could not be retrieved with less-performant decoding methods, due to the low orthogonality of cognitive maps in the CA1 region.

ABSTRACT Hippocampus stores spatial representations, or maps, which are recalled each time a subject is placed in the corresponding environment. Across different environments of similar geometry, these representations show strong orthogonality in CA3 of hippocampus, whereas in the CA1 subfield a considerable overlap between the maps can be seen. The lower orthogonality decreases reliability of various decoders developed in an attempt to identify which of the stored maps is active at the moment. Especially, the problem with decoding emerges with a need to analyze data at high temporal resolution. Here, we introduce a functional-connectivity-based decoder, which accounts for the pairwise correlations between the spiking activities of neurons in each map and does not require any positional information, *i.e.* any knowledge about place fields. We first show, on recordings of hippocampal activity in constant environmental conditions, that our decoder outperforms existing decoding methods in CA1. Our decoder is then applied to data from teleportation experiments, in which an instantaneous switch between the environment identity triggers a recall of the corresponding spatial representation. We test the sensitivity of our approach on the transition dynamics between the respective memory states (maps). We find that the rate of spontaneous state shifts (flickering) after a

teleportation event is increased not only within the first few seconds as already reported, but this instability is sustained across much longer (> 1 min.) periods.

5.1 INTRODUCTION

Over the recent decades, multi-cell recording techniques have provided insights into the nature of brain representations and their internal dynamics. While many works have focused on the input-output transfer functions in primary sensory systems (visual, olfactory, etc.), understanding functions corresponding to complex representations in higher cortical circuits is very hard as they are often based on mixed selectivities [175]. In relatively rare cases, such as in the entorhino-hippocampal system, a highly processed neural activity can be reliably correlated with behavior. The so-called 'place cells' in the CA1 and CA3 of hippocampus exhibit sharp spatially tuned and environment specific activity [79], see Fig. 5.1. Collective activity of the place–cell population coding for the environment defines its neural representation, or *map*. Simultaneous recording of multiple place-cell activity thus allows one to identify a general memory state of the network (specific map), as well as to decode the accurate position of the rodent in the corresponding environment [176].

Recently, [143] have studied the dynamics of transient change between the spatial maps encoding two different environments in CA3 at high temporal resolution (ca 120 ms time windows). The two environments differed by light cues that could be switched instantaneously ('teleportation procedure'), while the animal hippocampal neural activity was recorded to monitor the course of activation of the proper spatial map. An unstable state generally emerged for some seconds after the light switch, as both maps started to flicker back and forth. This phenomenon, called flickering, was identified through measure of the similarity between the place-cell population activity and its averaged patterns across both environments, recorded earlier in respective reference sessions. Typically, a given 120 ms time window activity of the test data strongly correlated with the average reference activity in one map, and had essentially no correlation with the reference activity in the other map.

Success of such comparison-based decoding methods reflects the strong orthogonality of spatial maps in CA3: across two environments, activity of place cells broadly differ in their mean frequencies and receptive field locations, see Fig. 5.1(b). Hence, simple map decoders, essentially assuming that cells fire independently of each other, are sufficient to reliably identify the representation expressed by the animal. In contrast, remapping between environments (especially of similar geometry) is less orthogonal in CA1 as it shows higher number of cells firing at corresponding places across rooms, see Fig. 5.1(a). The population activity vector often correlates well with both concurrent reference templates, which hinders the use of comparison-based methods for map decoding.

Here we address the challenging goal of map decoding in CA1 by introducing a probabilistic graphical model for the neural activity configurations in each map. Graphical models, in which a functional connectivity network accounts for the pairwise correlation structure between neuronal firing events in the recorded population [21], have been

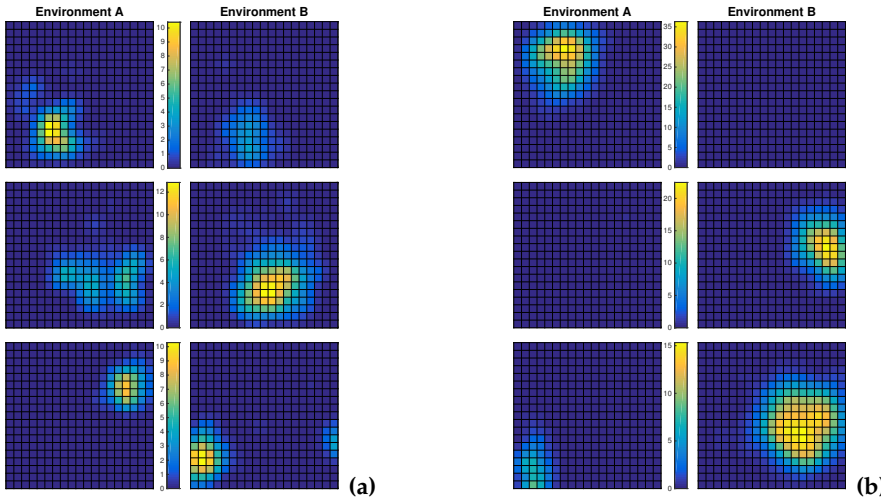


Figure 5.1: **Hippocampal representations are less orthogonal in CA1 (a) than in CA3 (b).** Each panel shows six firing fields from CA1 (a) and CA3 (b) corresponding to three place cells (rows) in the recorded neuronal population, computed from 10 min recordings of the activity during free exploration of environments A and B (same 60×60 cm square shapes; spatial bins : 3×3 cm). Whereas CA3 coding is highly sparse and representations are largely orthogonal, CA1 population shows higher amount of cells active in corresponding locations across the two rooms, with peak rates (color scale) changing from one environment to the other. The non-orthogonality of environment representations in CA1 makes identification of the represented map from neural activity difficult compared to the situation in CA3. CA3 data were taken from [143]. Colorbars show average firing rate in Hz.

applied to various areas so far [48,49], e.g. to estimate the information conveyed by [53] or the activity of [51,52] retinal ganglion cells in the presence of visual stimuli, to detect learning-related changes in functional connectivity in the prefrontal cortex [54,55].

We apply our graphical-model decoder to already published [143] and some new recordings of the hippocampal activity in CA1, performed within the teleportation setup of [143]. Our decoder shows very good performances in terms of precision and statistical properties in CA1. It allows us, in particular, to identify transitions between spatial representations in CA1 in a statistically robust way. Remarkably, we find that the frequency of these flickering events is increased even minutes after a teleportation switch.

It is important to stress that, in contradistinction with previously used map decoders, ours does not use any position information. It can therefore be applied to decode and study the dynamics of general brain states with unknown input correlates, the only working hypothesis being that we dispose of reference sessions to build statistical models of the corresponding internal states.

5.2 RESULTS

Decoding methods and number of parameters used

We start by presenting map-decoding methods and their performances. For each environment, A and B, we have two recorded sessions with constant light cues: the first one, called reference session, is used to infer the decoder parameters. The second one, called test session, is used for cross-validation, *i.e.* to assess the performances of the decoder. We compare the performances of five different decoders, described in Methods, Section 5.4. Our decoders mainly differ by the fact that they may use or not knowledge of the rat positions and of the spatial rate maps (place fields). They are also based on simple comparison methods or on more sophisticated probabilistic frameworks.

Rate-map based decoders require the computation of the rate maps during the reference session. Knowledge of the position $\vec{x}(t)$ of the rats and of the neural firing rates $r_i(t)$ as a function of time t allows one to build the rate maps, that is, the average firing rate of each cell i as a function of the rat position $\vec{x}$, $r_i^{(m)}(\vec{x})$ for environment $m = A, B$. The similarities between those reference population activities and the activity measured during the test sessions may then be used as a simple estimator of the map retrieved by the rodent. We consider two such comparison-based approaches, called *Dot Product* and *Pearson* [143]. A more sophisticated decoder, called *Poisson*, consists in assuming that each place cell i fires with a Poisson process, with average rate $r_i^{(m)}(\vec{x})$ when the rodent is at position $\vec{x}$, and in estimating the likelihood of the test spiking activity with this multiple Poisson process and for maps $m = A$ and B. The posterior distribution for the (binary) map variable m can then be computed, and we decode the map as the one with larger posterior probability. Poisson is based on a more solid probabilistic framework than Dot Product and Pearson, while making use of the same rate maps estimated from the reference sessions.

Activity-only decoders do not need any information about rat position and place fields. Those models provide approximate expressions for the probability distribution of population neural activity over short time bins, *i.e.* of binary (silent or active neuron in the time bin) strings of length N (the number of recorded neurons). The *independent-cell* model is the simplest maximum-entropy model [22]; it reproduces the N average activities of the neuron only. The second model, called *Ising* in statistical physics, is a graphical model that, in addition, reproduces the pairwise correlations between the neural activities in a time bin [22,45,51]. The Ising model requires the inference of pairwise effective couplings between every two cells, which we have performed with the Adaptive Cluster Expansion method [45, 177]. Similarly to Poisson, the independent-cell and Ising models provide estimates of the likelihood of the population activity in a time bin, and can be used to compute the posterior distribution for the map variable, m, and to decode the retrieved map through maximization over m.

As a consequence the numbers of parameters to be learned from the reference sessions vary a lot with the decoders. For N recorded neurons (38 in one of the data sets studied here, see Materials) and a discretization of the environment into S (=20×20 in the present analysis) spatial bins, the numbers of parameters to be extracted from the reference sessions are, respectively $N = 38$ for the independent-cell decoder, $\frac{1}{2}N(N+1) = 741$ for the Ising decoder, and $N \times S = 15,200$ for the Poisson, Pearson, and Dot Product decoders.

Cross validation of map-decoding methods

Inferred Ising couplings are fingerprints of environment representation in CA1

As a result of rate remapping taking place in CA1 (Fig. 5.1) the populations of active cells in the two environments are similar. This property can be seen from the comparison of the inputs $\{h_i\}$ in the Ising models inferred in the reference sessions of the two environments, see Fig. 5.2. The input h_i to place cell i takes similar values across the environments; its value is indicative of the average firing rate of the cell (Methods, Section 5.4).

Distinction between the neural representations of the environments in CA1 can, however, be drawn from the correlational structure of firing events in the place-cell population. Place cells with overlapping firing fields in one environment are indeed more likely to be simultaneously active during the animal's exploration, and their activities are thus correlated. Due to remapping the amplitudes of these correlations are specific to each environment. The inferred Ising couplings $\{J_{ij}\}$, which capture the direct correlation between cells i, j not mediated by other recorded cells (Methods, Section 5.4), are different from one environment to the other, as shown in Fig. 5.2.

The set of effective couplings $\{J_{ij}\}$ is therefore a *fingerprint* of the environment [178], which we can exploit to distinguish between maps, *i.e.* to decode the neural representation. Note that these effective, functional couplings are not directly related to the physiological synaptic interactions, which are not accessible from the data.

Pairwise correlations are also environment specific, see Appendix B, Fig. 5.9. However, inferring effective couplings allows us to *score* any configuration of the population activity, that is, to quantitatively assess its similarity with typical activities in each environment, as shown below. This score is, in practice, given by the Ising probability, see Methods, Eq. (6.2), and heavily relies on the inferred couplings and inputs. Scoring is not possible from the knowledge of the mean activity and pairwise correlations.

Comparison of performances of map-decoding methods

We present a systematic study of the performances of map-decoding methods in CA1 within the framework of binary-decoder theory, see Methods, Section 5.4 for a detailed description. Results are reported in Fig. 5.3.

We plot in Fig. 5.3 (a) the Receiver Operating Characteristic (ROC) curve for the Ising and Pearson decoders. Briefly speaking, ROC curve shows the value of the True Positive

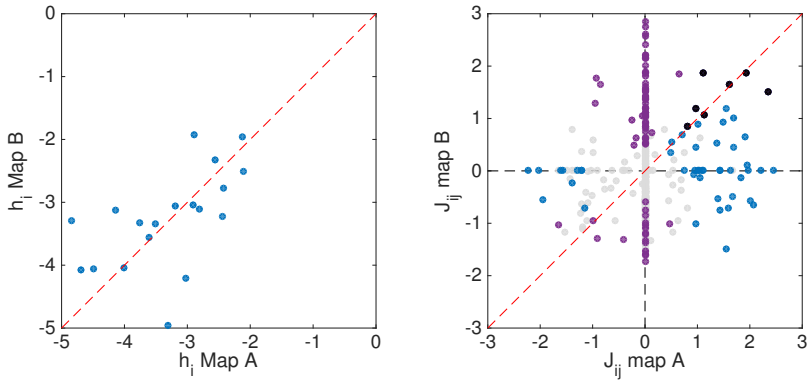


Figure 5.2: **Comparison of inferred Ising parameters across the two maps**. Top: Inputs h_i of the Ising models inferred from reference sessions. Only values greater than -5, corresponding to a firing rate of c.a. 0.05 Hz in the independent-cell model, are shown. Bottom: Couplings J_{ij} of the Ising models inferred from reference sessions. Dots are colored with reference to their relative statistical error (due to finite sampling) $\frac{|J|}{\Delta J}$: Unreliable couplings, *i.e.* such that $\frac{|J|}{\Delta J} < 3$ in both maps, are shown in grey (note the presence of many zero couplings produced by ACE). Couplings that are reliable only in one map are shown with purple (A) and blue (B) dots. Couplings reliable in both maps are shown in black. Analysis performed with discretization time bin $\Delta t = 120$ ms.

Rate (fractions of time bins in reference session for environment A for which the decoder rightly decodes map A) as a function of the False Positive Rate (fractions of time bins in reference session for environment B for which the decoder erroneously recognizes map A). A random decoder would have equal values for TPR and FPR, and lies on the diagonal line of the unit square in Fig. 5.3 (a). A perfect decoder would always recognize map A in environment A and never in environment B, and would thus correspond to TPR $= 1$, FPR $= 0$. Varying the threshold for significance of the decoder changes both the values of TPR and FPR, with the resulting ROC shown in Fig. 5.3 (a). We observe that the Ising decoder shows much better performances than the Pearson decoder. An alternative representation of the decoder performances is given by the Precision-Recall curve, shown in Fig. 5.3 (b), see Methods, Section 5.4 for definition.

A measure of the accuracy of the decoder is given by the integral of the ROC curve, called Area Under the Curve (AUC), which ranges from 0.5 for a random decoder to 1 for a perfect decoder. To compare the five decoders we plot in Fig. 5.3 (c) their AUC values as a function of the elementary time bin Δt, ranging from 10 ms to 1 s. The Ising model, which takes into account the correlational structure of the population activity, has higher decoding precision and retrieval capacity than other decoders in CA1 recordings (Fig. 5.3 (a,b)). As a consequence, in terms of AUC (Fig. 5.3 (c)), *Ising* is generally the most performant model, followed by *Poisson* and lastly by equally-performant *Pearson*

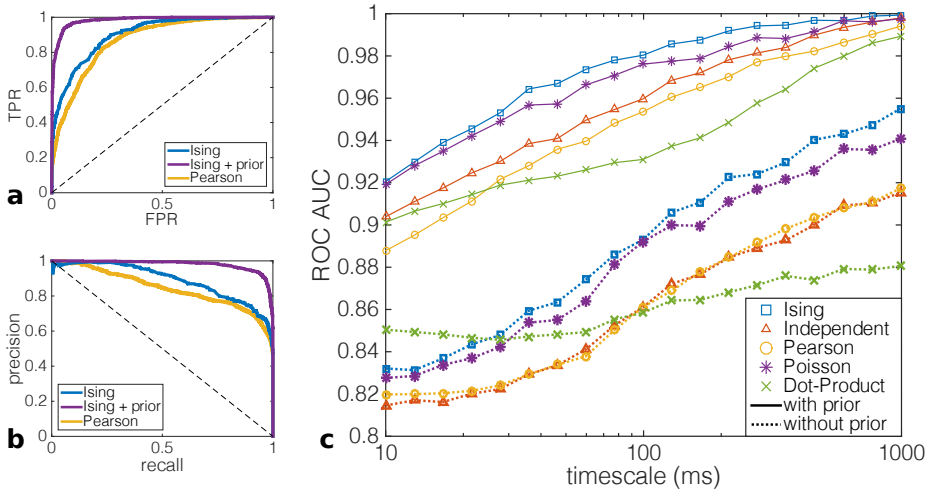


Figure 5.3: **Performances of spatial representation decoders.** ROC **(a)** and Precision-Recall **(b)** curves computed at fixed time scale $\Delta t = 120$ ms for a combination of two test sessions in environments A and B, recorded in CA1. Maps A and B correspond, respectively, to positive and negative predictions, see Table A.1. The True Positive Rate, also called Recall, is the number of true positive predictions divided by the total number of positive events. The False Positive rate is the number of false positive predictions, divided by the total number of negative events. Precision is defined as the fraction of identified positive events that are true positives. **(c)** performances of *Ising*, *Independent-cell*, *Poisson*, *Pearson*, and *Dot Product* decoders (with and without the addition of a continuity prior) as functions of the discretization time scale Δt, applied to CA1 neural recordings. Full and dashed curves correspond to predictions, respectively, without and with continuity prior; in the latter case the correlation C in Eq. (5.13) decays over $t_0 = 2$ time bins (Methods, Section 5.4 and Fig. 5.5 (a)).

and *independent-cell* decoders. *Dot Product* method is the best performant on very short time scales (< 20 ms), but its performance increases very slowly with the time bin width, and as a consequence it has the worst performance for $\Delta t > 100$ ms.

This behavior has an explanation in terms of sensitivity of the different models to the average number of active neurons per time bin. Bayesian models, whose predictions do not depend on the specific position of the rat at each time, rely on information conveyed by activity alone. As a consequence, when the number of simultaneously active neurons for each time bin is very small, Bayesian models may be less accurate than decoders that take into account spatial information, like Dot Product.

As a general feature we observe that the performances of all decoders improve for larger discretization time scales (Fig. 5.3 (c)). This result does not come from better

inference of the Ising parameters, as couplings remain remarkably unchanged as Δt varies, see Appendix C. The increase in performance may be simply understood as follows. Decoding performances were evaluated from the fraction of time bins in which the decoded map matched the one of the external environment evoked by the light conditions. In test sessions with stable external environment for several minutes, it is natural that merging larger portion of data results in more stable decoded maps, and, hence, in a larger fraction of correctly decoded maps. Similarly, improvement in decoding stability is obtained through the introduction of a continuity prior, which prevents switching back and forth between spatial maps in nearby time bins, see below for further discussion.

Performance of Ising decoder with number of recorded cells and duration of recording

We further analysed the behavior of the Ising decoder (as the most performant amongst presented methods) upon varying the number of recorded cells and the duration of the recording through subsampling the reference session data. As expected, the performance of the Ising decoder improves with the number of neurons and the duration of the reference data sets, see Fig. 5.4. We observe that fluctuations from subsample to subsample shrinks as the number of retained neurons increases, an effect that mirrors the heterogeneity of spatial and environment-related firing properties of single neurons. A relatively small subsample of the reference session, *e.g.* of duration ~ 1 min, suffices to compute a good estimate of the average firing rates, yielding performances similar to the independent model (Fig. 5.3, red curve, $\Delta t = 120$ ms).

Map decoding with continuity prior

Map decoding can be combined with a continuity prior that enhances persistence in the decoded maps over consecutive time bins, see Methods, Section 5.4. The motivation for the continuity prior is two-fold. First, in situations where the latency between a delivery of external stimulus and the network state change is the main parameter to be measured (e.g. after pharmacology treatment, etc.), one needs to search for a single time point of the state transition. This can be achieved by imposing a strong continuity prior, allowing for the presence of a single transition between maps along the whole recording session.

Secondly, with moderate continuity prior, dynamical events (such as state transitions) can be detected with more precision, at the price of discarding events that happen on time scales shorter than the temporal resolution set by the prior strength. To estimate this temporal resolution, we compute, for a fixed prior strength K, the correlation $C(\tau)$ between decoded maps in two time bins that are τ bins apart, see Methods, Eq. (5.13). This correlation decays exponentially with τ, see Fig. 5.5 (a). A persistence 'time' τ_0 can be computed through an exponential fit of the correlation: τ_0 is the characteristic number of bins over which decoded maps are persistent. Its value can be chosen at our convenience by tuning the prior strength parameter K, see Fig. 5.5 (b) and Methods, Section 5.4. Hence, we can choose a temporal resolution τ_0 and exploit the noise-cancelling property of the continuity prior over larger time scales.

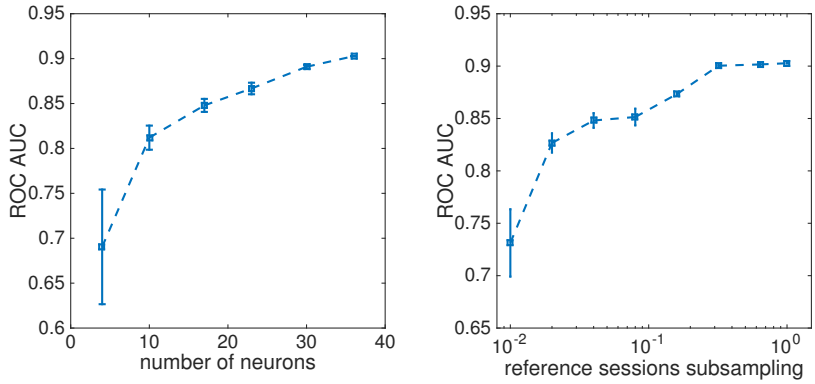


Figure 5.4: **Performance of Ising decoder in subsampled conditions**. Top: Performance of Ising decoder for $\Delta t = 120$ ms time bins vs. number n of cells employed in the inference and decoding routines. For each value of n results are averaged over 10 randomly-chosen subsamples of cells (among the $N = 36$ recorded neurons). Bottom: Performance of the decoder as a function of the fraction of the reference session recording (subsampled from the total recordings of duration $T = 509$ and $T = 551$ seconds). For each duration considered results were averaged over 3 random subsamples of reference data.

Unless otherwise specified we set in the following the characteristic persistence time to the small value $\tau_0 = 2$ time bins. As shown in Fig. 5.3, use of this weak continuity prior enhances decoding performances with the Ising method. The AUC increases by about 10%, see Fig. 5.3 (c). For direct comparison, if one instead increases the time-bin resolution Δt by a factor 2, the increase in AUC is much lower (Fig. 5.3 (c)): for instance, Ising AUC is equal to 0.90 for $\Delta t = 120$ ms and to 0.92 for $\Delta t = 240$ ms, while it reaches 0.98 for $\Delta t = 120$ ms with a continuity prior such that $\tau_0 = 2$ time bins. This result shows that imposing a continuity prior is a more efficient way to reduce statistical errors in the decoding than considering larger time bins.

The observed increase in performance due to the application of the continuity prior can be explained from different perspectives. First, as explained in the Methods section and observed in Fig. 5.5 (a), the overall procedure introduces short-range correlations (decaying over a tunable time scale) between time bins. The resulting effect is a smoothing filter, similar to a convolution with a sliding averaging window, which acts as a noise-cancelling filter, and improves decoding precision. Secondly, the application of the prior enhances the stability of the decoded maps. This improves the decoding performance since, as pointed before, the test session is such that light conditions remain stable for long times (minutes) before the switch.

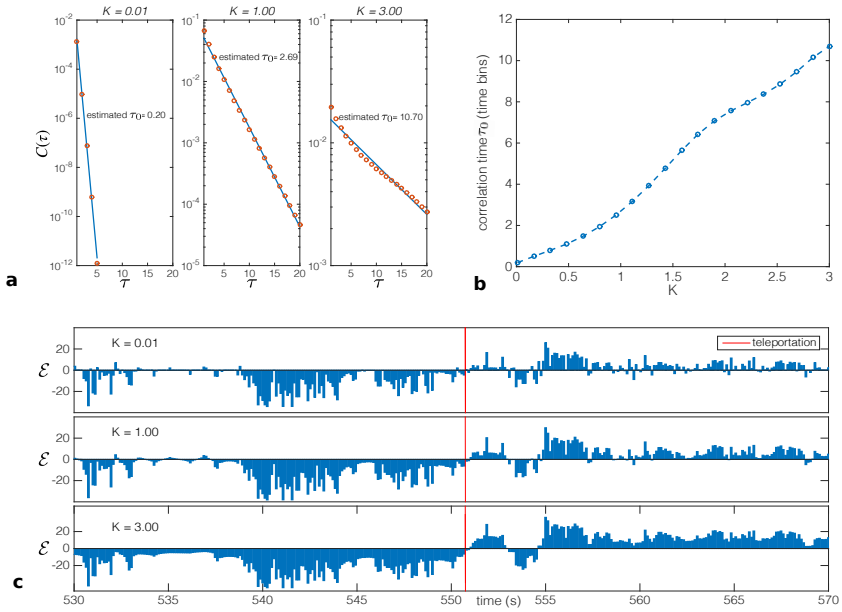


Figure 5.5: **Continuity prior for map decoding. (a)** Correlation (Methods, Eq. (5.13)) between maps decoded in two time bins as a function of their separation τ (measured in units of time bins), for three values of the prior strength K. Correlations are well fitted by exponential decaying functions, over a characteristic number of bins τ_0. **(b)** Value of τ_0 as a function of the prior strength K. **(c)** Application of the prior on CA1 teleportation session for different values of prior strength parameter K. Difference in log-probabilities of the neural activity configurations over time bons t. Ising decoder, with a discretization time bin $\Delta t = 120$ ms.

Transitions between maps in "teleportation" experiment

Brain hippocampal memory circuitry is a dynamic system expressing distinct states of activity - neural representations of surrounding space - with attractor properties [143, 172, 179]. We applied our Ising decoder to dynamically identify those states to CA1 recordings in the 'teleportation' setup introduced in [143], in which the appearance of recording box is abruptly changed by switching between two familiar light cue settings (A and B, respectively) while the laboratory rat continuously explores it (Methods, Section 5.4). This procedure was shown to induce a rapid exchange of corresponding hippocampal representations in CA3, including periods of instability with spontaneous fast flickering between them. The CA1 recordings considered here include both data published in [143], and new recordings, see Methods section. Transitions between the maps were identified based on activity models of representations A and B, respectively, inferred from reference

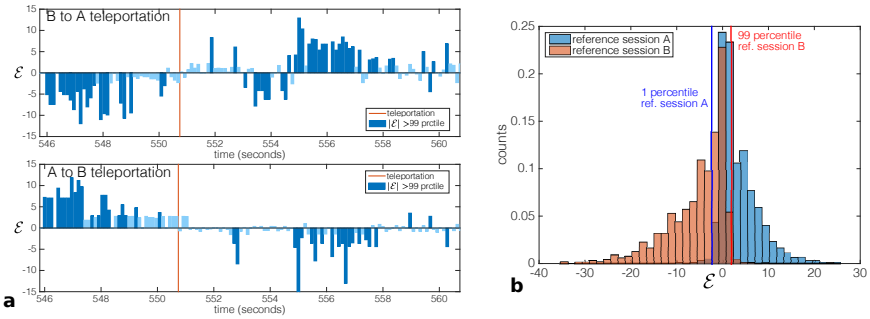


Figure 5.6: **Log-probabilities of neural activities around teleportation events.** Difference $\mathcal{E}$ of the log-probabilities, Eq. (5.1), computed with the Ising decoder applied to the neural activity recorded in a teleportation session (**a**), with light-cue switch from environment B to A (top) and from A to B (bottom). The light switch is marked with a red line, predictions higher than 99 percentile value of reference sessions are colored in dark blue, weaker prediction are colored in light blue. Panel **(b)** shows the distributions of differences of log-probabilities in reference sessions. A percentile value θ in $[0, 100]$ (normally in the interval $[90, 100]$) is defined. We consider a test time bin as significantly decoded as A only if the log-probability difference $\mathcal{E}$ of the activity configuration in the time bin is higher than the θ percentile value of reference session B, and as B only if its value is lower than the $100 - \theta$ percentile value of reference session A. The underlying reasoning is to decode a test time bin as A only if it is very unlikely that it comes from reference population B, and vice-versa.

recordings in both environments under stable conditions preceding the 'teleportation test session'.

To illustrate performance of Ising method in the post-teleportation kinetics of network state expression, we used four teleportation events recorded in hippocampal CA1 in three rats. Representative evolution of the difference in log-probabilities $\mathcal{E}$, see Eq. (5.1), of the neural activities, computed with the models inferred for the two maps from the reference sessions, is shown before and after two instances of teleportation events in Fig. 5.6(a). The criterion for accepting given bin as corresponding to representation of environment A or B, respectively, was set to match 1% error derived from stable reference sessions, see Fig. 5.6 (b). This ensures that a time bin is identified as A only if there is 99% (or higher) confidence that this difference in log-probabilities cannot be found in environment B (and vice-versa) under reference conditions.

Teleportation procedure induces long-term network instability

To characterize the kinetics of network state development, we identified the amount of time bins expressing a neural representation that was incongruent (non-corresponding) with the present environment, *i.e.* coding the environment presented before the teleportation. We estimated the short-term effect within interval of the first 10 seconds, and a possible long-term effect in the period that begun after 30 seconds after the telepor-

tation has elapsed. The rates of incongruent bins are shown in Fig. 5.7. The amount of non-corresponding events per time bin raised from the baseline levels before the teleportation 0.013 ± 0.002 SEM to 0.046 ± 0.021 measured within the first 10 seconds after the teleportation (short-post effect). However, this increase was not significant, probably due to combination of large variability within the short evaluated interval (10 seconds in contrast to order of minutes of baseline state before the teleportation) and frequent empty bins (no cell active in $40.5\% \pm 5.8$ SEM of all bins).

Interestingly, the rate of flickering remained significantly increased beyond 30 seconds after the teleportation (0.034 ± 0.021 SEM, $F = 19.38$, $p < 0.01$). [143] used temporal binning that reflected local theta oscillation ($6 - 11$ Hz) in the hippocampal circuitry. While all the results reported so far were obtained with a fixed, regular binning with a similar rate ($\Delta t = 120$ ms, *i.e.* about 8 Hz), we decided to re-analyze the teleportation data in a natural theta binning as done by [143]. We detected the phase of local theta oscillation based on minimum place-cell activity criterion, and the corresponding timestamps were used to define the temporal bins. We got the same pattern of results as with fixed binning (pre = 0.014 ± 0.002 SEM, short post = 0.041 ± 0.030 SEM, $p > 0.05$; long post = 0.030 ± 0.005 SEM, $F = 6.15$, $p < 0.05$), yielding non-significant increase within the first 10 seconds and a significant increase after 30 seconds following the teleportation, respectively.

Last of all, we analyzed once more this teleportation data, this time with the Pearson correlation-based decoder. Neither in fixed nor theta-based binning this decoder returned significant differences between the pre teleportation and any of the post (short and long) teleportation intervals ($p > 0.05$ in all cases). This finding provides further evidence for the results depicted in Fig. 5.3, that is, for the better performance of Ising method over Pearson decoder for hippocampal CA1 data.

Identification of transitions with strong continuity prior

Our network state-decoding procedure with continuity prior can be used to detect internal state-shifts under predefined criteria. For instance, when the prior strength is brought to extreme values the decoding procedure discards the fast instability-driven dynamics and, instead, returns a single state transition time point that reflects the evolution of log-likelihood values across the continuum of temporal bins. Taking as an illustration the CA1 teleportation session in Fig. 5.8, we see that the response of network activity state to the teleportation event is identified with high accuracy. This is a valuable tool to measure the most probable moment of network remapping even under widely fluctuating dynamics.

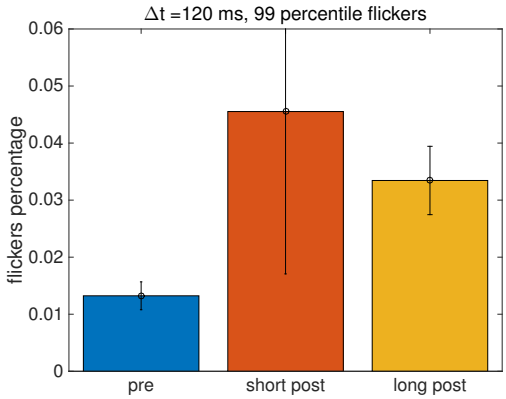


Figure 5.7: **Teleportation enhances network instability over both short- and long-term periods.** Percentage of temporal bins expressing the environment-incongruent coding computed in rest conditions (pre), during the first 10 seconds after a light switch (short post), and in long-term period after the teleportation event (long post, more than 30 seconds after light switch). Only bins expressing $\mathcal{E}$-values higher than 99 percentile of reference sessions have been taken into account; Similar results are obtained with 90 and 95 percentiles. Results were averaged over a total of four sessions recorded from three different animals. Recording durations before (pre)/after (post) teleportation equal to, respectively, 10/8, 9/9 minutes (33 cells), 12/11 minutes (17 cells), and 2.5/3 minutes (20 cells). Analysis performed with Ising environment decoder with discretization time bin $\Delta t = 120$ ms.

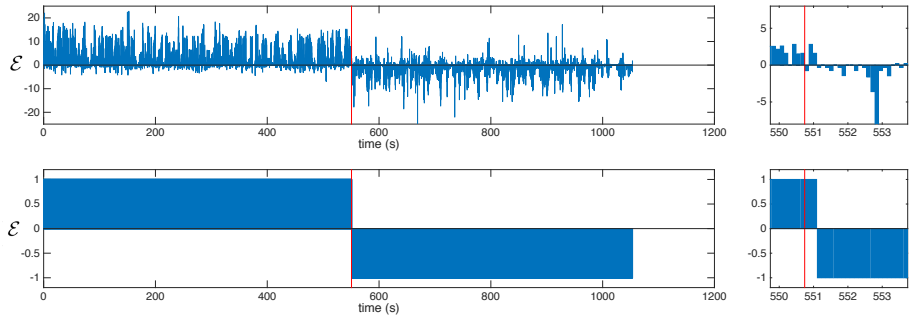


Figure 5.8: **Network state transitions identified by implementation of continuity prior.** Ising decoder and Viterbi algorithm with strong continuity constraint applied to neural activity in a CA1 teleportation session with enlarged examples. Light switches are marked with red lines. Analysis performed with time bin $\Delta t = 120$ ms.

5.3 DISCUSSION

Graphical models for brain state identification in the absence of input correlates. Methods for decoding spatial representations considered in this work can be divided in two classes, depending on whether they make use of positional information or not. Remarkably, the latter methods do not show worse performances than the former approaches. In CA3, efficient decoding does not require the use of sophisticated probabilistic models: due to the quasi-orthogonality of maps, the simple independent-cell decoder, which compares the activity at any time to the average activities in environments A and B irrespectively of the rat position, shows very good performances [180]. In CA1, the similarity in the firing fields across environments constrained us to consider a graphical model, the pairwise Ising model, which not only captures the average activity of the place cells but also their pairwise correlations. The higher performance of the Ising model, combined with the lower number of parameters involved in the inference process compared to firing field-based decoding methods, suggests that the correlational structure of neural firing activities conveys essential information about the internal representation of memorized environments.

A substantial advantage of this approach is that it can be effectively applied to other brain regions with much weaker correlation between the local activity and its inputs, *e.g.* the prefrontal cortex, or without any known input-output relation. The use of graphical models does not require any knowledge about the network inputs, as activity states are identified based on a (high-dimensional) fit of the correlation structure of the spiking data [49]. The core idea, first put forward in the context of retinal data modeling [51], is that the model obtained after inference of the functional network is an approximate (albeit quantitatively accurate, compared to principal-component based approaches [181]) description of the distribution of activities characterizing specifically one brain state. Provided that we have at our disposal different data sets for well-identified states (here, the reference sessions) we may later use the inferred model to decode the activity at any time. This approach has recently been applied to identify transient activation of memory-related cell assemblies in the rat prefrontal cortex [54, 55]. We expect, owe to its generality and its applicability to very fast time scales (down to ~ 10 msec), further applications in future. Note that our decoding approach, based on the inference of effective pairwise couplings, could be extended to higher-order interactions. In the hypothetical situation of distinguishing between brain states that differ in high-order statistics, *e.g.* in the frequencies of 3-cell firing events only, the inference of these high-order effective interactions would be necessary to obtain an efficient decoder.

Let us also remark that, once the Ising parameters have been inferred from reference sessions corresponding to the possible states (here, maps), the computation of the log probability difference $\mathcal{E}(\{s_{i,t}\})$ is very fast, as it requires $O(N^2)$ operations only. Our decoder could therefore be applied online, provided the neural activity configuration are available, *e.g.* through automatic spike sorting, at any time. One potential issue here is that fast spike sorting may introduce error in the activity variables, leading to wider and more overlapping distributions of $\mathcal{E}$, see Fig. 5.6 (b). Maintaining high precision in the

decoding would still be possible if the confidence threshold θ is increased, but at a price of smaller recall, see Fig. 5.3 (b).

Functional connectivity-based models for map decoding: Ising and other models. The Ising decoder introduced in this article yields the highest performance on all time scales Δt in CA1 (Fig. 5.3). While we have here mostly considered the activity vectors as discretized in regular-spaced time windows of duration Δt, our approach was also easily extended to process activity in elementary windows in correspondence to Theta cycles. It would be interesting to pursue the latter analysis to deepen our understanding of the role of Theta oscillations for the dynamics of transitions [143] in CA1, and to assess the plausibility of the different transition scenarios (temporary disappearance or coexistence of both map representations) put forward by theoretical studies [182].

In this regard, repeating the present study with probabilistic models capable of capturing some aspects of the activation dynamics in recorded spiking sequences, such as Generalized-Linear Models [183], could be potentially interesting. Contrary to their Ising model counterparts effective couplings in the GLM approach are not necessarily symmetric, and may reflect specific ordering in neuron activations. However, some basic assumptions underlying GLM, such as the Poissonian nature of firing events are questionable for hippocampal place cell activity [184]. Another potentially interesting alternative is provided by reverse engineering of networks of Integrate-and-Fire neurons [185–187], which were already applied to recordings, *e.g.* of retinal data with tens of neurons.

Instabilities in hippocampal space representations. In the CA3 area of hippocampus, patterns of place-cell activity across different environments behave as uncorrelated network states with attractor properties [172]. Transitions between those hippocampal activity states were recently studied based on recordings taken during a free exploration in two environments in an experimental paradigm shown to induce rapid switches [143]. In the present paper we used multiunit recordings from hippocampal area CA1. Both CA3 and CA1 are parts of the entorhino-hippocampal loop, an essential circuit for spatial memory and navigation in mammalian brain. Despite being directly connected in series (CA3 signalling into CA1), they very much differ in their architecture - while CA3 is organized as a recurrent network with attractor properties, CA1 has a feed forward structure - and in their connections with other brain areas involved into space representation [188].

Use of the Ising model allowed us to robustly decode the memory state expressed in the CA1 network, with temporal resolution high enough to reflect natural time patterning of activity provided by local theta oscillation (ca. 6-11 Hz). We could track the network state kinetics following the sensory input switch. In agreement with previous report in CA3 [143], we detected a high degree of flickering in CA1 following the switch.

Moreover, when analyzing the development of post-teleportation population vector activity on a long-term scale (20-60 sec), we found sustained network instability in CA1 spanning far beyond the 10 seconds interval reported in [143], see Fig. 5.7. This effect is statistically significant with the Ising decoder but could not have been discovered with

simpler, correlation-based methods. The presence of long-term instabtility in CA1 is rather surprising as the network usually reaches a relative stability within a couple of seconds after the cue switch [143]. An occasional delayed spontaneous flickering was described in CA3 as a result of repetitive teleportation within a short time period (every 40-60 seconds) [143]. This suggests prolonged (though rare) flickering effect might be present in both CA3 and CA1. In our data the persistent instability in CA1 came after one or two teleportation events on a given day, respectively.

What mechanism can account for this observation? The current view considers the short term (up-to 10 sec.) instability as a product of teleportation-induced conflict between a sudden change in the allothetic visual input (another environment presence) and a non-corresponding idiothetic signaling (no self-motion tracked traversal). Within couple of seconds the idiothetic input seems to reset as the rate of flickering dramatically decreases to levels close to the baseline steady state. The fact an occasional flickering is present longer both in CA3 and CA1 can have more reasons. The autoassociative character of CA3 is capable to store and express stable patterns of activity, but also to associate between different simultaneously active ensembles in the network. After teleportation, despite an attractor separation on a theta frame binning has been proved, an occasional overlap between both representations is present as well. Such brief coactivation of concurrent maps can eventually lead to their binding by collateral synapses or by detecting and learning their conjunction by CA1 [189]. Such a linkage could, under appropriately ambiguous or noisy input (e.g. encountering an odor mark dropped in the concurrent lighting conditions), eventually lead to a rare completion of the concurrent activity state. Other, rather speculative, possibility is that the observed activation of the other representation in CA3 and CA1 could be related to a reflection of past configuration of the external world, eventually to an expectation of another coming change of environment identity, so far of unknown mechanisms. Whatever input triggers the long-term flickering, these transient episodes do not occur during sharp wave/ripple complexes as they were present during strong theta network oscillation without any apparent increase of population activity. A further insight that is beyond the scope of this report is necessary to provide a better understanding of the origin and characteristics of long-term dynamics of transition between distinct hippocampal network states.

5.4 METHODS

Experimental methods

Electrode preparation and surgery. Single unit neuronal activity was recorded in three adult Long Evans male rats in hippocampal subfields CA1. Rats were implanted with a "hyperdrive" allowing for an independent positioning of 16 tetrodes organized into an ellipsoid bundle. Tetrodes were twisted from 17 um insulated platinum-iridium wire (90% and 10%, respectively, California Fine Wire Company). Impedance of electrode tips was adjusted by platinum plating to 120 – 250 kOhm (at 1 kHz). Anesthesia was introduced by placing the rat into a plexiglas chamber with seal top filled with isoflurane vapour. Then the animal was shaved and placed into the stereotaxic frame and continued the isoflurane delivery with a face mask. Breathing, heart action and reflexes were monitored continuously. Hyperdrive was then implanted above the right dorsal hippocampus at coordinates AP 3.8 mm and ML 3.2 mm relative to bregma. Stainless steel screws and dental acrylic were used to stabilize the implant on the skull. Two of the screws served as the hyperdrive ground.

Tetrode position. The tetrodes were slowly approached towards CA1 or to CA3 within 2-3 weeks after the surgery while the rat was resting in a comfortable pot on a pedestal. To maintain stable recordings, electrodes were not moved at all before and during the experiment on a given day. The recording reference electrode was positioned in corpus callosum. Additional reference for EEG was placed in stratum lacunosum moleculare.

Recording procedures. Neural activity was recorded while the rat was behaving in an apparatus described by [143]. Signal was recorded differentially against the reference tetrode. Hyperdrive was connected to a multichannel, impedance matching, unity gain headstage and its output conducted through a 82-channel commutator to a Neuralynx digital 64 channel data acquisition system. Signal was band-pass filtered at 600 Hz–6 kHz. Unit waveforms above individually set thresholds (45-70 μV) were time-stamped and digitized at 32 kHz. Position of the light emitting diodes on the headstage was tracked at 50 Hz to assess the animal's position. For the purpose of this study only data from intervals when the rat's movement speed exceeded 5 cm/sec were used. Broadband EEG from each tetrode was recorded continuously at 2000 Hz.

Spike sorting and cell classification. Spikes were sorted manually using 3D graphical cluster-cutting software (SpikeSort, Neuralynx) The feature space consisted of three-dimensional projections of multidimensional waveform amplitudes and energies. Autocorrelation and crosscorrelation functions were used as additional separation tools. Putative pyramidal cells were distinguished from putative interneurons by average rate, spike width and occasional complex spikes.

Histology. After the experiment was finished, the rat was overdosed with a barbiturate and was perfused intracardially with saline followed by 4 % formaldehyde. Brain coronal

sections (30 μm) were stained with cresyl violet. Traces of all 14 tetrode locations were identified. Each tip location was considered as the place in the section before the tissue damage became negligible. Only recordings from tetrodes with their tips in CA1 were used in this study.

Behavioral procedure. Animals were first pre-trained according to the procedure described in [143]. Briefly speaking, the apparatus consisted of two identical black plastic boxes (60 $\times$ 60 cm, 50 cm in height). The two environments differed only by sets of light cues, one placed on the upper rim of the box, the second was positioned under the semi-transparent floor with an additional cue on one wall, respectively. There were no other visual cues present as the experiment was otherwise carried in darkness provided by surrounding light-proof curtains. The training consisted of four phases. Initially, the two boxes were connected with an alley so the rat could freely explore both of them within three 20 min. sessions for 3 days. In the second phase, after the first 20 min. session, the alley was removed and the animal was placed into box A or B, respectively, in a quasi-random manner so that it received two 10 min. sessions in each of them, respectively. The next day the rat received two 10 min. sessions in each environment as the day before. Then we removed the double maze and replaced it with a single box equipped with both sets of lights that was presented at the original locations with just one cue set switched on at the given session. The rat was given another two 10 min. sessions in each environment that day. Finally, the next day, after two sessions in the original locations, the box was presented in a central location. Again, the animal was presented another two 10 min sessions in each environment, respectively, in a quasi-random order. In all stages, the running sessions were separated by a 20 min. break in the resting pot. On the test day, both environments were presented in two "reference" recording sessions (10 min each). After a 20 minutes break, the test session begun. The animal was inserted to the box with one set of lights on, and the lights were switched between the both sets after couple of minutes of recording.

Data structure

Cross validation of environment decoding methods. For the validation of environment decoding methods (Section 5.2) a total amount of four recording sessions were used. Two of them, one in the environment A and one in the environment B, called *reference* sessions, were used to infer activity models and reference statistics. The other two (again one in environment A and one in environment B) have then been used as *test* sessions, *i.e.* to assess the performance of our method for decoding which environment is internally-represented by the rodent.

Teleportation sessions. In the post-teleportation analysis shown in section 5.2 we used recordings from three experiments performed in three different animals (one of them was already used in the original [143] study). Each data set included two reference sessions for both environments and one or two teleportation sessions, each containing one single

light switch. The switch between light cues was in total performed four times (direction balanced, A to B or vice versa), and the activity was recorded for some minutes before and after the teleportation.

Map decoding methods

We consider two classes of decoders: *Rate-map based decoders*, which expressly use the knowledge of place fields and the rat trajectory as an input, and *Activity-only decoders* that do not rely on any information about the correspondence between position and neural firing. Throughout this section neural activities are binned with time resolution Δt; we define the number of spikes of neuron i in time bin t, $n_{i,t}$, and the binary activity, $s_{i,t} = \min(n_{i,t}, 1)$. Little information is lost when considering s instead of n as long as Δt is smaller than the typical inter-spike interval of the cells.

Activity-only decoders

Bayesian approach to map decoding. We introduce probabilistic models for the distribution of activities $\{s_i\}_{i=1\ldots N}$ in a time bin, $P(\{s_i\}, \Theta)$. Those models are parametrized by a set of variables, Θ, which are fitted to maximize the likelihood of the data in reference sessions. Two sets of parameters $\Theta^{(m)}$ are fitted, one for each reference session $m = A, B$. We then define the difference in log–probabilities

$$\mathcal{E}(\{s_i\}) = \log\left[\frac{P(\{s_i\}|\Theta^{(A)})}{P(\{s_i\}|\Theta^{(B)})}\right] . \tag{5.1}$$

The sign of the quantity $\mathcal{E}(\{s_{i,t}\})$ may be used to decode the map in time bin t. Significance levels, based on the percentiles of the distribution of $\mathcal{E}$ can be imposed, see Results, Section 5.2.

Independent-cell model. The simplest way to model the firing properties of the neural population is to assume that the neural activities s_i are independent from cell to cell. For each map m, the probability distribution P is parametrized by a set $\Theta^{(m)} = \{h_i^{(m)}\}$ of N 'inputs' $h_i^{(m)}$:

$$P^{(m)}\left(\{s_i\}|\Theta^{(m)}\right) = \prod_i \frac{e^{h_i^{(m)} s_i}}{1 + e^{h_i^{(m)}}} . \tag{5.2}$$

Each input parameter is fitted in order to match the average value of s_i with $P^{(m)}$ and the mean value $\mu_i^{(m)}$ of $s_{i,t}$ across the time bins t in reference session relative to map m. This procedure yields $h_i^{(m)} = \log[\mu_i^{(m)}/(1 - \mu_i^{(m)})]$.

Graphical Ising model. A more accurate probabilistic model for the activity of the cell population is obtained when pairwise correlations between neural activities s_i in a time bin are taken into account. For each map m, we introduce couplings $J_{ij}^{(m)}$ to express the

conditional probability that cell i is active given the activity of cell j. The probability distribution $P^{(m)}$ is now parametrized by the set $\Theta^{(m)} = \{h_i^{(m)}, J_{ij}^{(m)}\}$ of N inputs $h_i^{(m)}$ and $\frac{1}{2}N(N-1)$ couplings $J_{ij}^{(m)}$:

$$P^{(m)}\left(\{s_i\}|\Theta^{(m)}\right) = \frac{\exp\left(\sum_i h_i^{(m)} s_i + \sum_{i<j} J_{ij}^{(m)} s_i s_j\right)}{\mathcal{Z}^{(m)}[\{h_i^{(m)}, J_{ij}^{(m)}\}]} \tag{5.3}$$

where $\mathcal{Z}^{(m)}$ is a normalization constant. Parameters $h_i^{(m)}$ and $J_{ij}^{(m)}$ are computed to match the average values of s_i and $s_i s_j$ with P and, respectively, the mean values of $s_{i,t}$ and $s_{i,t}s_{j,t}$ across the time bins t in reference session relative to map m. This hard computational problem can be approximately solved with the Adaptive Cluster Expansion (ACE) algorithm [33, 46, 49, 177], which provides estimates of the parameters $\{h_i^{(m)}, J_{ij}^{(m)}\}$ and $\mathcal{Z}^{(m)}$ in Eq. (6.2).

Adaptive Cluster Expansion (ACE). The log-likelihood of the model parameters given the neural activities, $\log P$, is regularized, *i.e.* added a term penalizing large couplings. It is expanded as a sum of contributions corresponding to clusters (subsets) of variables [33, 177]. Clusters of increasing sizes are recursively built from smaller clusters and added to the expansion if their contributions to the log-likelihood exceed some threshold value. The value of the threshold is iteratively decreased, until the 1- and 2-point statistics of the data are reproduced (within the expected sampling accuracy). This iterative procedure builds the simplest network (smallest number and sizes of selected clusters) able to reproduce the low order statistics of the data and avoids overfitting. Statistical error bars on the inferred inputs and coupling parameters are estimated [177]. The threshold value, the number, and maximal size of selected clusters at convergence are given in Appendix A.

Rate-map based decoders

Computation of rate maps. The squared box is partitioned into a 20×20 grid of 3×3 cm^2 bins, and the rat position during the two reference sessions is discretized with respect to this grid. The coordinates (x_t, y_t) associated to time bin t correspond to the first spatial bin visited by the rat in the time interval $[t - \Delta t; t]$. We define the average firing rate $r_i^{(m)}(x, y)$ as the total number of spikes emitted by neuron i in the reference session m when the rat is at position (x, y), divided by the total time $T^{(m)}(x, y)$ spent by the animal in this spatial bin. These rate maps are then smoothed to fill missing bins through discrete cosine transform [190].

Pearson decoder. The observed firing pattern at time t ,$\{n_{i,t}\}_{i=1\ldots N}$, is compared to the average firing rates in map m, $\left\{r_i^{(m)}(x_t, y_t)\right\}_{i=1\ldots N}$, in the position (x_t, y_t) occupied by the animal at the same time [143]. This comparison is made through the Pearson correlation

$$\mathcal{C}^{(m)}(\{n_{i,t}\}) = \frac{\langle n\, r^{(m)}(x_t, y_t)\rangle_t - \langle n\rangle_t \langle r^{(m)}(x_t, y_t)\rangle_t}{\sqrt{\left(\langle n^2\rangle_t - \langle n\rangle_t^2\right)\left(\langle r^{(m)}(x_t, y_t)^2\rangle_t - \langle r^{(m)}(x_t, y_t)\rangle_t^2\right)}} \tag{5.4}$$

where the notation $\langle f\rangle_t := \frac{1}{N}\sum_{i=1}^{N} f_{i,t}$ denotes the average of the quantity $f_{i,t}$ over the N neurons i in time bin t. The decoding of the map in time bin t is done according to the sign of

$$\mathcal{E}(\{n_{i,t}\}) = \mathcal{C}^{(A)}(\{n_{i,t}\}) - \mathcal{C}^{(B)}(\{n_{i,t}\})\ . \tag{5.5}$$

Dot-product decoder. The second method used in [143] compares directly the activity to the firing rates at the rat position. The decoding of the map m is done according to the sign of

$$\mathcal{E}(\{n_{i,t}\}) = \langle n\, r^{(A)}(x_t, y_t)\rangle_t - \langle n\, r^{(B)}(x_t, y_t)\rangle_t\ . \tag{5.6}$$

Bayesian Poisson rate model. This model assumes that each neuron fires independently according to a Poisson statistics, with a position-dependent firing rate $r_i^{(m)}(x, y)$ in map m. The probability of the number of spikes $\{n_i\}$ emitted by the neural cells in a time bin when the rat is at position (x, y) reads

$$\begin{aligned} P^{(m)}\left(\{n_i\}|\ (x, y)\right) &= \\ &= \prod_i \frac{\left(r_i^{(m)}(x, y)\,\Delta t\right)^{n_i}}{n_i!}\ e^{-r_i^{(m)}(x,y)\,\Delta t} \end{aligned} \tag{5.7}$$

The prior probability over positions is

$$P^{(m)}(x, y) = T^{(m)}(x, y)/T^{(m)}\ , \tag{5.8}$$

where $T^{(m)}$ is the total recording time in reference session m. Assuming that both maps m are *a priori* equally likely, we obtain the probability of the activity conditioned to map m by marginalizing over positions

$$P(\{n_i\}|m) = \sum_{x,y} P\left(\{n_i\}|(x, y)\right) \times P^{(m)}(x, y)\ . \tag{5.9}$$

We then define the log-ratio

$$\mathcal{E}(\{n_{i,t}\}) = \log\left[\frac{P(\{n_{i,t}\}|A)}{P(\{n_{i,t}\}|B)}\right]\ , \tag{5.10}$$

whose sign will be used to decode the map in time bin t.

decoder output	A	B	A	B
cue	A	A	B	B
denomination	True Positive	False Negative	False Positive	True Negative

Table 5.1: Denominations used for the four possible events, depending on the output of the decoder and on the environment-defining cue. The cue is not changed throughout the reference session.

Performance measure of a binary decoder

To quantitatively assess decoding performance of map-decoding methods we refer to binary classifier theory [191–194].

Receiver Operating Characteristic (ROC) diagrams. A standard framework to assess the performance of binary decoders is the so-called ROC diagram [191]. For each time bin t the decoder outputs either map A or map B. To match the vocables used in the ROC framework we will arbitrarily say that the output is *positive* if the map is decoded to be A, and *negative* if the map is predicted to be B. If the output of the decoder matches the environment defined by the light cues at the same time t, the prediction is said to be *True*, otherwise it is said to be *False*. For instance, a time bin such that the decoder predicts A, in agreement with the cues, corresponds to a True Positive event. The 2×2 possible events are shown in Table A.1. Two important quantities are: the True Positive Rate (TPR, also called Recall), that is, the number of true positive predictions divided by the total number of positive events, and the False Positive Rate (FPR), that is, the number of false positive predictions, divided by the total number of negative events. In other words, the TPR measures the fraction of time bins with A–cues that are correctly decoded as A, while the FPR is the fraction of time bins with B–cues that are incorrectly predicted to be A.

Our binary decoders are all based on thresholding the *estimator* variable $\mathcal{E}$. Within the Bayesian framework, for instance, we compute $\mathcal{E}$ as the difference between the logarithms of the posterior probabilities of A and B, and output Positive if the difference is larger than $\theta = 0$, Negative otherwise. The value of the significance threshold θ can be arbitrarily changed, with the consequence of modifying the TPR and FPR values. A ROC curve shows the parametric plot of TPR vs. FPR as the threshold varies, and describes a curve in the unit square, see Results, Section 5.2. The two extreme points of the ROC curves have coordinates (0,0), and (1,1); (0,0) is obtained for a very large significance threshold θ, the decoder never outputs Positive and both TPR and FPR vanish; (1,1) is obtained when the significance threshold is very low, the decoder always outputs Positive and both TPR and FPR are equal to unity. Very good decoders are such that the TPR is close to unity, while maintaining a very low value for the FPR. A random-guessing decoder would give equal values for the TPR and FPR, and the ROC curve would coincide with the diagonal of the unit square.

A complementary measure of decoding performances is the Precision versus Recall (or TPR) curve, obtained by scanning the values of the significance threshold θ, see Results, Section 5.2. The Precision is defined as the number of true positive events, divided by the total number of positive predictions. When lowering the significance threshold the Precision decreases from 1 to 0, while the Recall increases from 0 to 1.

Area Under the Curve (AUC). A quantitative measure of the decoding performances is the Area Under the (ROC) Curve [191]. According to this measure, the ideal decoder has AUC $= 1$, while random guessing would give AUC $= 0.5$. Note that this measure is invariant with respect to the arbitrary choice of assigning *positive* value to environment A: if we assign positive to B and negative to A instead of the previous choice, ROC curves will undergo a symmetry transformation with respect to the top-left/bottom-right diagonal, resulting in an identical area under the curve. This is granted by the fact that positive and negative values are mutually exclusive and complementarily cover the whole data set: for each θ value the fraction of False Positive (B decoded as A) equals one minus the fraction of True Negative (B decoded as B) events.

Continuity prior for map decoding

A continuity prior can be included in map inference in order to reduce noise in the decoding and highlight clusters of contiguous transited time bins. To do so, we consider the output $\{\mathcal{E}_t\}$ of the map decoder (see Section 5.4); for Bayesian decoders $\mathcal{E}_t$ is the difference between the log-likelihoods of the two maps $m_t = +1$ and -1 in time bin t. We then introduce a prior, controlled by a strength parameter K, which favors persistence between decoded maps in nearby time bins. Informally speaking, K is the cost (in log-likelihood) we are willing to pay for flipping the map index in time bin t predicted by the sign of $\mathcal{E}_t$ to its opposite value, if it then matches the map indices of the neighboring time bins, $t-1$ or $t+1$. The prior may thus be effective in changing the map prediction m_t if the differences between $\mathcal{E}_{t-1}$, $\mathcal{E}_t$, $\mathcal{E}_{t+1}$, ... are of the order of K (in absolute value). Two situations are encountered: (1) for some decoders, *e.g.* Pearson, $\mathcal{E}_t$ takes value in $[-2;2]$, and the variations of $\mathcal{E}$ over successive time bins is bounded; (2) for other decoders, *e.g.* Independent-Cell, Poisson and Ising, the difference between $\mathcal{E}_t$ and $\mathcal{E}_{t+1}$ can take arbitrarily large values and show wide fluctuations as t varies across the recording. In the latter case, a uniform prior K is unadequate in large portions of the recording. To circumvent this difficulty we introduce a scale factor $\beta < 1$, and multiply all outcomes $\mathcal{E}_t$ by this factor. As a result we get a smoother time course of $\mathcal{E}_t$ over the time index t, on which a uniform prior can now be applied.

The joint probability of the time sequence of map predictions $\{m_t\}$ reads

$$P(m_1, m_2, \ldots, m_T) = \tag{5.11}$$
$$= \frac{1}{\mathcal{Z}} \exp\left(\frac{\beta}{2}\sum_{t=1}^{T} \mathcal{E}_t\, m_t + K \sum_{t=1}^{T-1} m_t\, m_{t+1}\right)$$

where $\mathcal{Z}$ is a normalization coefficient. To decode the map in time bin t we compute the marginal probability P_t over m_t from the joint distribution P. Exploiting the analogy with the one-dimensional Ising model of statistical physics, this computation can be done with the transfer matrix method, also called dynamic programming, in a time scaling linearly with the total number of time bins. Then the outcome of our combined decoder+prior is

$$\mathcal{E}_t^{\text{decoder+prior}} = \frac{1}{\beta} \log \left[\frac{P_t(m_t = +1)}{P_t(m_t = -1)} \right] . \tag{5.12}$$

The presence of the $\frac{1}{2}$ and $\frac{1}{\beta}$ factors in, respectively, Eq. (5.11) and Eq. (5.12) ensure that, for $K = 0$, $\mathcal{E}_t^{\text{decoder+prior}}$ and $\mathcal{E}_t$ coincide. In practice we choose $\beta = \frac{1}{|\mathcal{E}_0|}$, where $\mathcal{E}_0 := \max_t \{|\mathcal{E}_t|\}$.

Induced correlation as a function of K. The transfer matrix technique allows us to compute also the correlation between the maps decoded τ bins apart, defined as

$$C(\tau) = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \left(\langle m_t m_{t+\tau} \rangle - \langle m_t \rangle \langle m_{t+\tau} \rangle \right) \tag{5.13}$$

where the angular-bracket notation denotes the average over the probability distribution in Eq. (5.11). $C(\tau)$ decays exponentially with τ, over a characteristic 'time' monotically growing with K in Eq. (5.11), see Results, Section 5.2.

acknowledgements We are indebted to S. Rosay, who contributed to the early stage of the data analysis. We are grateful to J. Tubiana and A. Treves for useful discussions and suggestions. This study benefited from partial fundings from the CNRS-InphyNiTi INFERNEUR project and from GACR 15-20008S, Q-39 and NPU I LO1503 of the Czech Republic.

5.5 SUPPLEMENTARY INFORMATION

ACE inference convergence details

The ACE inference procedure of Ising model parameters was applied with L_2-norm regularization of strength $\gamma = 5/B$, where B is the total number of time bins [177]. Details on the convergence are given in Table 5.2. The full code for Adaptive Cluster Expansion can be downloaded from the GitHub repo `https://github.com/johnbarton/ACE/`.

Comparison of neuron activities across spatial maps

Similarly to Fig. 5.2 where we compare the Ising parameters inferred from the population activity in the two environments A, B, we show in Fig. 5.9 the probabilities of firing of all

session	N	$\theta(\times 10^{-3})$	S	K_C	N_C
cv. A	36	0.89	5.31	7	390
cv. B	36	0.08	6.21	8	4789
t. I-II A	33	0.21	4.82	4	509
t. I-II B	33	1.4	4.33	5	198
t. III A	17	1.1	1.60	3	27
t. III B	17	0.81	1.76	3	34
t. IV A	20	0.89	5.45	4	117
t. IV B	20	0.99	4.51	6	1347

Table 5.2: Studied sessions (cv. = cross-validation, t. = teleportation, followed by number of teleportation and environment) number of recorded cells (N) and ACE parameters at convergence: threshold θ for cluster selection, cross-entropy (in natural log.), maximal size K_C and number N_C of selected clusters. The algorithms stops when the relative errors on single-neuron frequencies and pairwise connected correlations become smaller than unity [177].

cells i (in fixed time bins with $\Delta t = 120$ ms) and the pariwise correlations (defined as the probability that cells i, j fire together in a bin minus the product of their individual firing probabiltiies). We see that no substantial correlation is found in the pairwise statistics of cells across the two environments.

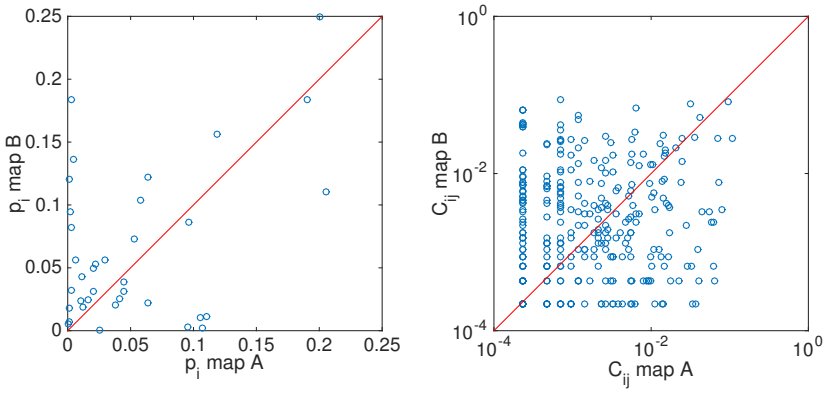


Figure 5.9: Comparison between correlations and averages of the two maps of the cross-validation reference sessions. Fixed binning with $\Delta t = 120$ ms.

Dependence of J_{ij} on temporal binning

Couplings inferred for time-bin duration $\Delta t = 120$ ms are compared to the ones inferred for $\Delta t = 10$ ms in Fig. 5.10. Many couplings are very similar across the two binning choices. Differences, in particular null couplings in just one of the two cases, mostly arise from sampling differences. For 10 ms time windows, it is rare to find two neurons active within the same time bin, while, for larger time bins, there is a smaller number B of time bins, which forces us to consider larger ACE threshold θ. Couplings inferred using the theta-binning discretization procedure for data are very similar to the ones inferred using a fixed time binning of 120 ms (average duration of theta cycles), see Fig. 5.10. A discussion of the independence of Ising couplings from the bin duration Δt was done by [50].

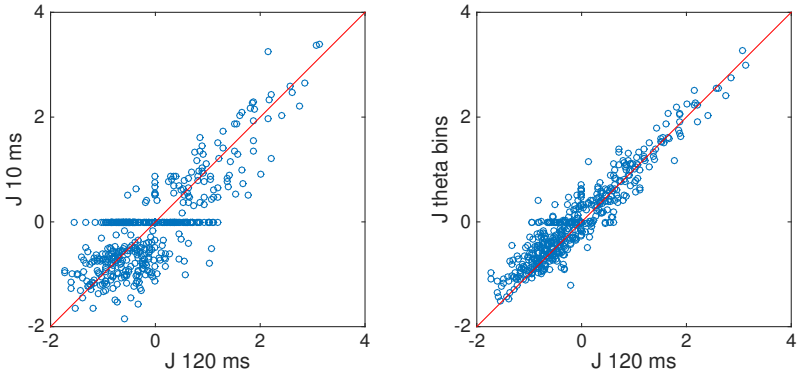


Figure 5.10: Left: Scatter plot of couplings inferred with time bin $\Delta t = 120$ ms vs. $\Delta t = 10$ ms (fixed time bin discretization procedure, from cross-reference data set). Right: Scatter plot of couplings inferred with fixed time bin vs. theta-binning procedure ($\Delta t = 120$ ms, from cross-reference data set).

6

INTEGRATION AND MULTIPLEXING OF POSITIONAL AND CONTEXTUAL INFORMATION BY THE HIPPOCAMPAL NETWORK

This Chapter was published by myself, S. Cocco, and R. Monasson in [2]. It features an analysis of the CA3 data of [143], courtesy of Karel Jezek from Charles University.

ABSTRACT The hippocampus is known to store cognitive representations, or maps, that encode both positional and contextual information, critical for episodic memories and functional behavior. How path integration and contextual cues are dynamically combined and processed by the hippocampus to maintain these representations accurate over time remains unclear. To answer this question, we propose a two-way data analysis and modeling approach to CA3 multi-electrode recordings of a moving rat submitted to rapid changes of contextual (light) cues, triggering back-and-forth instabitilies between two cognitive representations ("teleportation" experiment of Jezek et al). We develop a dual neural activity decoder, capable of independently identifying the recalled cognitive map at high temporal resolution (comparable to theta cycle) and the position of the rodent given a map. Remarkably, position can be reconstructed at any time with an accuracy comparable to fixed-context periods, even during highly unstable periods. These findings provide evidence for the capability of the hippocampal neural activity to maintain an accurate encoding of spatial and contextual variables, while one of these variables undergoes rapid changes independently of the other. To explain this result we introduce an attractor neural network model for the hippocampal activity that process inputs from external cues and the path integrator. Our model allows us to make predictions on the frequency of the cognitive map instability, its duration, and the detailed nature of the place-cell population activity, which are validated by a further analysis of the data. Our work therefore sheds light on the mechanisms by which the hippocampal network achieves and updates multi-dimensional neural representations from various input streams.

AUTHOR SUMMARY As an animal moves in space and receives external sensory inputs, it must dynamically maintain the representations of its position and environment at all times. How the hippocampus, the brain area crucial for spatial representations, achieves this task, and manages possible conflicts between different inputs remains unclear. We propose here a comprehensive attractor neural network-based model of the hippocampus and of its multiple input streams (including self-motion). We show that this model is capable of maintaining faithful representations of positional and contextual information, and resolves conflicts by adapting internal representations to match external cues. Model predictions are confirmed by the detailed analysis of hippocampal recordings of a rat submitted to quickly varying and conflicting contextual inputs.

6.1 INTRODUCTION

Following the discovery of place cells, which specifically fire at determined positions in space [80], the hippocampus was recognized as an essential brain area for spatial representations and memories. These cognitive representations, or maps, actually code for more than position in physical space, and are also strongly informative about context [195], including physical features of the background, such as visual landmarks, light, odors, auditory stimuli, as well as more abstract conditions, such as the emotional state or the task to be performed [84, 87–89, 118, 196].

A fundamental property of the hippocampus is its capacity to memorize multiple cognitive maps [80, 94, 108, 197]. This property may result from specific recurrent synaptic connectivity in the hippocampal CA3 region [198, 199], and can be theoretically understood in the framework of continuous attractor neural networks (CANN) [154, 200]. Thanks to the remapping properties of place cells, multiple maps can be memorized in the same connectivity matrix with almost no interference between them [155, 156, 160, 201].

Cognitive maps may be retrieved when the animal explores again the corresponding environments, or be quickly and intermittently recalled depending on the most relevant behavorial information at that moment [141]. Different sources of inputs to the hippocampus concur to form, recall, and dynamically maintain cognitive maps [202]. Changes in visual cues and landmarks may substantially affect place field shape and positioning [84]. The Path Integrator (PI), capable of integrating proprioceptive, vestibular and visual flow inputs and possibly supported by the grid-cell network in the medial-enthorinal cortex (mEC) [109], allows the animal to update the neural representation during navigation [116]. The path integrator is itself sensitive to other sources of inputs, and undergoes reset in case of large disagreement with external landmarks or sensory information [126].

Insights about how these different inputs contribute to hippocampal representations were recently obtained by studying the effects of mismatches between path-integration and visual sensory information, in particular in virtual reality settings [124, 203]. In another study Jezek et al showed how abrupt changes in visual context (light conditions) during active exploration by a rodent resulted in fast flickering between context-associated maps in CA3 on the theta time scale [143] (Fig. 6.1A). Though they are largely artificial, these conditions offer a rare window on the fast dynamics of the place-cell population, and on how this dynamics is shaped by the various inputs.

Despite these studies, how contextual and PI inputs are combined by the hippocampal network to produce cognitive maps and accurate positional encoding is not fully understood yet. In this work, we carefully reanalyze and model the experiment of Jezek et al to address this issue. We first introduce of a dual inference method capable of extracting reliably and independently the encoded map [1] and the encoded position [176, 204] from the recorded spiking activity alone (Fig. 6.1B). Our dual decoder allows us to robustly show that the hippocampal activity always encodes the correct location in the retrieved map, even during the fast, unstable dynamics of the cognitive maps, as put forward in [143]. To explain this robust encoding, we propose a CANN model of the hippocampal circuitry, capable of storing multiple cognitive maps; the model is fed by visual-cue

and path-integration inputs projecting on the place-cell populations supporting those maps [157]. The path integrator is, in turn, influenced by the hippocampal activity, closing an interaction loop between the hippocampus and the mEC [123]. Our model not only reproduces the flickering phenomenology and the stable encoding of position, but also makes several precise predictions on the dynamics of cognitive maps, the relative strength of inputs, and the intricate activation of place-cell populations supporting the two maps. These predictions are corroborated by a further detailed analysis of Jezek et al's data. Our work therefore proposes explicit mechanisms by which the hippocampus could be capable of encoding various contextual and self-locomotion information in multi-dimensional representations, and of updating them accurately on fast time scales.

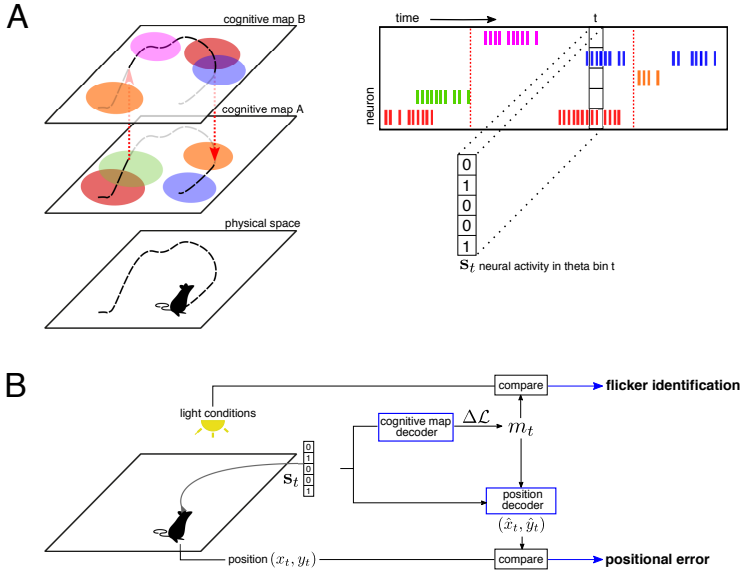


Figure 6.1: **Neural encoding and decoding of cognitive maps and position. A. Schematic description of Jezek et al's experiment.** As a rodent is moving in an environment, its position $\mathbf{r} = (x, y)$ is tracked over time, and a population of place cells is recorded (see raster plot). The activity of each cell is then binarized (0: silent cell, 1: active cell) to define the activity pattern $\mathbf{s}_t$ in theta cycle t. One out of two cognitive maps, established during the training sessions, is recalled at any time; change of maps are located by vertical dashed arrows. Place cells may have place fields in both cognitive maps (red, orange, and blue cells) or in one map only (green and purple cells). In addition, pair of neurons may be active simultaneously or not depending on the map; for instance the red and blue neurons have overlapping place fields in map B, but not in A. Hence, pairwise correlations are a fingerprint of the map. **B. Sketch of the dual decoder.** The neural activity alone is used to decode the retrieved map m_t as a function of time, and then to infer the position of the animal based on the place fields in the decoded cognitive map. Mismatches between the decoded maps and the external light cues define flickers over time. The distance between the predicted and real positions defines the positional error ϵ_t.

6.2 RESULTS

Jezek et al trained a rodent in two environments (square boxes), equal in size and shape, but differing by their light conditions [143]. A population of 34 CA3 place cells was recorded during reference sessions with fixed light conditions, and shown to define environment-specific maps, denoted by A and B. In a subsequent test session, taking place in a single box, instantaneous switches between environmental light conditions triggered the instability of the recalled cognitive map, which flickered back and forth between the two corresponding environments.

The neural activity $\mathbf{s}_t$ of the population in any theta cycle t encodes information on the context (the set of rules that connects position to activity, i.e. the place fields defining the cognitive map) as well as on the specific position within the environment (Fig. 6.1A). We first introduce a dual decoder, able to independently infer the cognitive map and the position, at high temporal resolution. By comparing the inferred position to the true animal location, we then assess how precisely the position is represented in the population activity, irrespectively of the cognitive map in which it is neurally encoded (Fig. 6.1B).

Functional network-based decoding of the cognitive map dynamics

Due to the global remapping properties of CA3, the intensities and mutual superpositions of place fields are specific to each environment (Fig. 6.1A). Consequently, the average firing rates and pairwise correlations of the place-cell population define a fingerprint of the corresponding cognitive map [4, 5]. We use the reference session recordings in each environment m (A or B) to compute this fingerprint statistics. We then build a model $P_m(\mathbf{s})$ that approximates the probability of observing the neural activity $\mathbf{s}$ when the cognitive map m is recalled. This model relies on the inference of a functional network of couplings between the place cells, reproducing the fingerprint statistics of map m [4, 205] (Methods).

Given the activity $\mathbf{s}_t$ recorded in theta bin t during the test session, we then compare the two probabilistic models P_A and P_B to estimate which map m is more likely to have generated $\mathbf{s}_t$. The log-ratio

$$\Delta\mathcal{L}(\mathbf{s}_t) = \log\left[\frac{P_A(\mathbf{s}_t)}{P_B(\mathbf{s}_t)}\right] \tag{6.1}$$

indicates whether the neural activity $\mathbf{s}_t$ is more similar to the neural patterns encountered in map A than to the ones of map B (large and positive $\Delta\mathcal{L}$), or typical of B and not of A (large in absolute value and negative $\Delta\mathcal{L}$). Comparing $\Delta\mathcal{L}$ to a statistical significance threshold allows us to infer the map m_t (Methods). If the decoded map m_t is discordant with the imposed light conditions the theta bin is identified as a *flicker*.

As a control, we check that $\Delta\mathcal{L}$ is mostly positive in reference sessions for environment A and negative for B, see Fig. 6.2A. Applying the decoder to the test session, we observe the presence of flickers, see Fig. 6.2B (yellow bins); flickers were first found in [143] with correlation-based methods requiring knowledge of the true position of the animal. An

analysis of the temporal correlation of these flickers reveals that they typically persist over ~ 6 theta bins (Methods and Fig. 6.2C); hence, cognitive maps show some inertia extending beyond the theta scale.

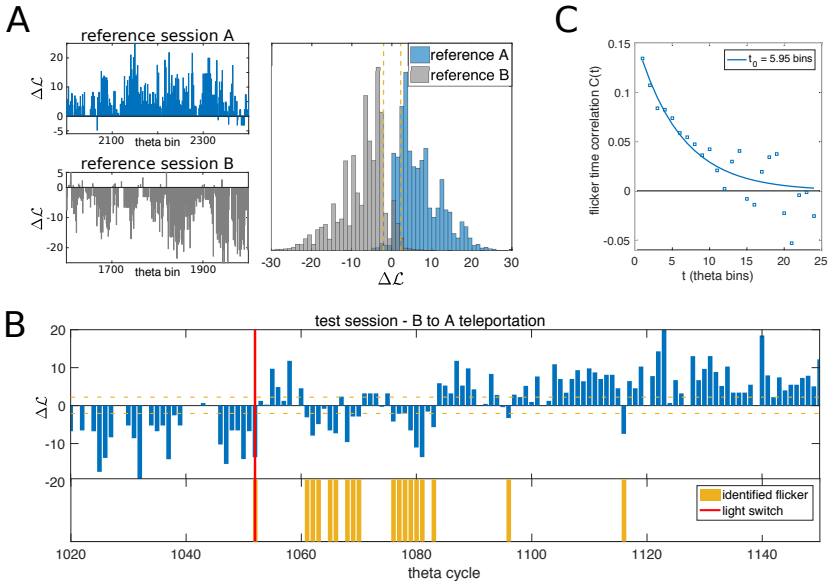


Figure 6.2: **Decoding the cognitive map from neural activity. A.** Map-decoding procedure applied to constant-environment reference sessions. The log-ratio $\Delta\mathcal{L}_t$, Eqn. (1), which is the difference of likelihoods of map A and B given the recorded neural pattern in theta bin t (Methods), is mostly positive in the reference session with constant light-cue evoking A (blue) and mostly negative in the reference session B (grey). **B.** Time course of $\Delta\mathcal{L}$ in a portion of the test session around one light switch (red vertical line). The emergence of flickers (disagreement between the recalled map and the light cue) is clearly visible after the light switch. Yellow horizontal dashed lines show the statistical threshold applied in map decoding and flicker identification ($|\Delta\mathcal{L}| > L_0 = \log 10$, Methods). Yellow bars represent the identified flicker instabilities, i.e. significative discordances ($\Delta\mathcal{L} < -L_0$) between the decoded cognitive map (here, B) and the post-switch light conditions (here, A). **C.** time correlation of flickers, computed with significance threshold $L_0 = \log 10$. The exponential fit shows that correlations extend over ~ 6 theta bins, highlighting the tendency of the cognitive map to persist beyond the theta cycle.

Position is accurately encoded even during flickering instabilities of the cognitive map

To assess if the fast dynamics of cognitive maps affects the quality of positional encoding we next re-use the neural activity pattern $\mathbf{s}_t$ in theta bin t, this time to infer the position of the animal. A naive Bayesian decoder [176, 204] takes as an input the above-decoded map m_t (Fig. 6.1B) and uses its place fields to estimate the position. The distance between the inferred position, $\hat{\mathbf{r}}_t$, and the true position, $\mathbf{r}_t$, defines the positional error ϵ_t. As shown

in Fig. 6.3A, the positional error ϵ_t (blue line) is independent of the time elapsed after the light switch, and has a value comparable to the one obtained in fixed-environment conditions (blue dashed line). This result crucially depends on the fact that position is estimated according to the decoded map m_t, which varies with time t. For comparison, in Fig. 6.3B we show the error if we decode the position according to the new, post-switch map (green line) or to the old, pre-switch one (red line) at all times. Both procedures result in similar, higher errors right after the light-switch, where flickers are frequent. The error with the post-switch map eventually decrease to fixed-environment value after few seconds, due to the rarity of flickers long after the light switch.

In summary, the output of our map decoder, m_t, can be interpreted as the correct cognitive state to read the positional code, see Fig. 6.2A. Even in the presence of fast dynamical flickers of the cognitive map, the location of the animal is robustly and coherently represented at all times. Our findings show that the hippocampus representation encodes both positional and contextual information in an independent and accurate way. Interestingly, the positional error computed with the map opposed to the decoded one (orange line in Fig. 6.3A) shows a significant reduction in the first seconds after the light switch; this non-trivial effect will be explained in detail in the next sections.

Continuous attractor neural network model for the interplay between path integrator, visual cues, and memory

The findings above suggest that the stream of positional information to the hippocampus is maintained despite the presence of rapid changes of cognitive representations following the abrupt modification of visual cue (V) after the light switch. A natural hypothesis is that the path-integrator (PI) sends to the hippocampus information relative to the position in the 'old' map [108, 157], competing with the visual cue input associated to the 'new' map. To formalize this assumption, we introduce a continuous attractor neural network (CANN) model that contains the minimal ingredients to understand the effect of conflicting PI and V stimuli onto the hippocampal activity. In the CANN paradigm for memory storage and retrieval of cognitive maps [12, 154–156], the animal location at a certain time is represented as a self-sustained bump of neural activity. The bump is localized in the current position within a two-dimensional manifold, where place cells are embedded according to the positions of their place field centers in the real environment. We generalize this classical model by including two informational inputs on the memory network, from allothetic (visual cues) and idiothetic (path integrator) stimuli. The proposed interaction model is composed of the following four ingredients, see Fig. 6.4A and Methods:

(a) A CANN *memory*, including excitatory recurrent connections of strength γ_J and global inhibition, designed to store and support two cognitive maps m, denominated A and B, which mimics the status of the CA3 place-cell network after learning of the two 'environments'. This model of stochastic neurons was described and studied in [12, 158, 160].

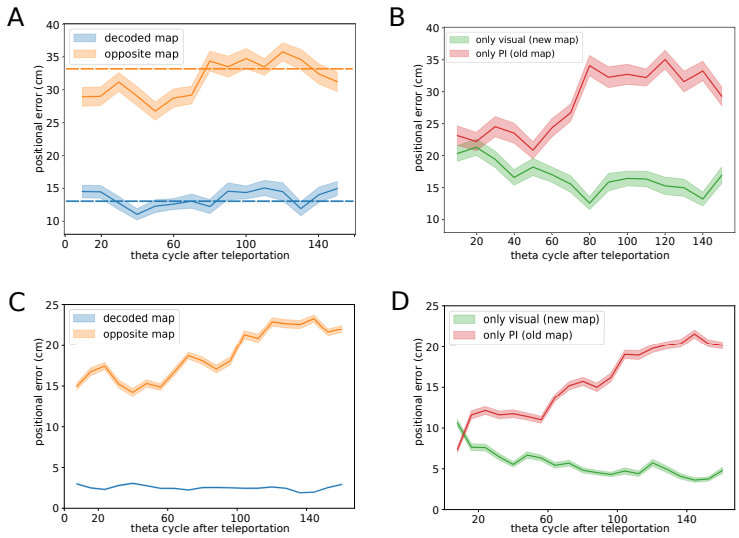


Figure 6.3: **Positional error as a function of time from the teleportation light switch. A. Recorded data.** Positional error computed as a function of time from the teleportation (in units of theta cycle). In each theta bin t, the map decoder is used to select which map m_t (set of place fields) to use to infer the position from the neural activity (Fig. 1A). When using the decoded map m_t, the positional error (blue line) is comparable to the one found in constant-environment conditions (blue dashed line). The orange line shows the positional error if the opposite map (alternative to m_t) is used for inferring position. The positional error is significantly reduced with respect to constant-environment conditions (orange dashed line) in the vicinity of the teleportation, and reaches comparable values after $\sim$ 80 theta bins. Shaded areas represent the standard error computed over 15 teleportation events. **B. Recorded data.** Same as in panel A but using a fixed map of reference for position inference, irrespectively from the decoded map. Red and green lines show results with, respectively, the 'old' (pre-teleportation) and the 'new' (post-teleportation) maps. **C & D. Simulations of CANN model**: same analysis as in panels A & B, computed over 15 simulated teleportation events, with the same trajectory of the rodent as in the experimental data. Model parameters: $N = 400$ neurons, $\gamma_I = 0.0025$, $\gamma_V = \gamma_{PI} = 0.4$, $\gamma_W = 6.25$, $\beta = 15$, see Methods and Figs. B&C in S1 Text for detailed discussion of the choice of parameters.

(b) A *visual-cue input* of amplitude γ_V onto place cells whose place fields match the current position of the rat, $\mathbf{r}_t$, in the cognitive map corresponding to the external light cue. The latter is denoted by $V = A$ or B.

(c) A *path-integrator input* of amplitude γ_{PI} that projects onto place cells whose place fields match the current position in the cognitive map corresponding to its own internal cognitive state [108], denoted by the variable $PI = A$ or B.

(d) An *effective feedback* from the hippocampal network to the path integrator, which stochastically maintains coherence between the recalled cognitive map and the path-integrator state.

From a functional point of view, the CANN model mostly behaves, for a fixed position **r** of the rodent, as an effective two-state model for the hippocampal activity, as sketched in Fig. 6.5A. These two states correspond to the activity localized in map A or B; their probabilities are controlled by the intensity of, respectively, the path-integrator and visual-cue inputs. Note that the emergence of two well separated collective states from the microscopic CANN model is intrinsically due to the presence of recurrent connections, see Fig. 6.5A; A characterization of the effective barrier between the states is reported in S1 Text (see Fig. A in S1 Text). The height of the barrier, controlled by the parameter γ_J, and the amount of stochasticity in the individual neural dynamics are crucial ingredients to determine the dynamics of the model. In particular, these variables control the time-correlation of flickers (Methods); For the chosen simulation parameters, the time correlation decays over ~ 7 theta bins (Fig. 6.5B) in accordance with data (Fig. 6.2C).

The typical outcome of a simulated experiment is shown in Fig. 6.4B. During the exploration phase preceding the light switch, the visual (b) and path-integrator (c) inputs jointly contribute to the stability of the internal representation of the position. A localized bump of activity, sustained by the recurrent connections (a), can be observed in the pre-switch map (Fig. 6.4A, left), say, $m = PI = V = A$. Right after the switch, the hippocampal network receives conflicting streams of information: $PI = A$ differs from $V = B$. The path integrator is still activating place cells coding for the current position of the animal in the 'old' map, while the visual stream points to neurons coding for the same position in the 'new' map (Fig. 6.4A, center). This results in a *conflict* between the two bump representations, which are mutually incompatible due to the orthogonality of global remapping. Flickering is produced as an alternance between these two possible states, $m = PI = A$ and $m = V = B$. During this conflicting phase (Fig. 6.4, shaded region), the feedback (d) from the memory network to the path integrator tries to achieve coherence between the hippocampal and path-integrator states. When the bump is in the visually-driven, post-switch map ($m = B$), incoherence is strong, and the path integrator is more likely to be reset. Realigning the path integrator state with the external cue, $PI = V = m = B$, brings the conflict phase to an end, and the hippocampal state reaches stability (Fig. 6.4A, right).

Despite its conceptual simplicity, the model shows a rich phenomenology and reproduces in a strikingly-accurate manner the results of the analysis of the CA3 teleportation recordings. In Fig. 6.4D we show a representative time trace of the log-ratio $\Delta\mathcal{L}$ (Eqn. (1)) in a simulated teleportation session. Alternate intervals of positive and negative $\Delta\mathcal{L}$ signal the presence of map instability, as in [143], following the light switch (red vertical lines) and the path-integrator realignment (green vertical lines). Applying to the simulated data the same positional-error analysis as for the recorded data (Fig. 6.3A&B), we observe the same qualitative picture, see Fig. 6.3C&D. In particular, position is coherently encoded in the recalled cognitive map at all times.

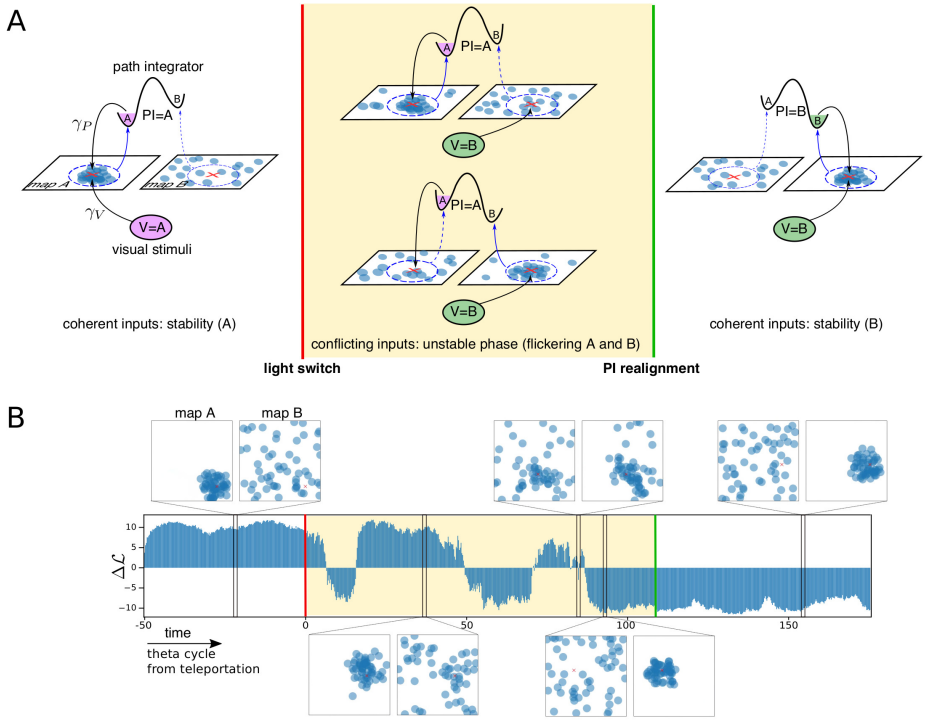


Figure 6.4: **CANN model for interplay between path integrator, external stimuli, and memory. A. Phenomenology of the CANN model.** The model is composed of a recurrent hippocampal network that has memorized two cognitive maps (place fields dispositions) denominated A and B, a path integrator input (PI), and a visual input (V). In the left panel both PI and V are activating place cells whose place-field centers (shown by blue dots) correspond to the position of the rodent (red cross X) in the cognitive map A. The activity is said to be localized in a bump around X in t map A, while it appears as sparse and uninformative if interpreted with respect to the place-field locations in map B. A feedback projection (blue arrows) from the hippocampal state (bump) to the path-integrator state (purple) maintains the stability of the system by enforcing that the retrieved hippocampal map and the PI state agree. After the light conditions have been switched (teleportation, red line), V projects on place cells encoding position X in map B. The two hippocampal cognitive maps are therefore in conflict, and the bump of activity is alternatively localized in A (center-top) or in B (center- bottom). When the hippocampal activity is localized in the cognitive map B, the feedback projection tries to realign the internal state of PI along the corresponding map. Once the realignment has succeeded (green line), both inputs are back to a coherent state, and stability is reached in the cognitive map relative to the post-teleportation external light conditions. **B.** Time trace of the log-likelihood difference $\Delta\mathcal{L}$ (Eqn. (6.13) in Methods) in a simulated teleportation session; flickers can be observed during the conflicting phase following teleportation (shaded region). Prior to teleportation, and after the PI is realigned, inputs are coherent, and the system is stable: the sign of $\Delta\mathcal{L}$ is constant, mirroring the localization of the bump in one map. Screenshots of the activity projected on the two cognitive maps are shown for five different times. From left to right: bump localized in map A, bump localized in map A during the conflicting phase, mixed state during the conflicting phase, bump localized in map B during the conflicting phase, bump localized in map B. The video of the simulation can be found in SI.

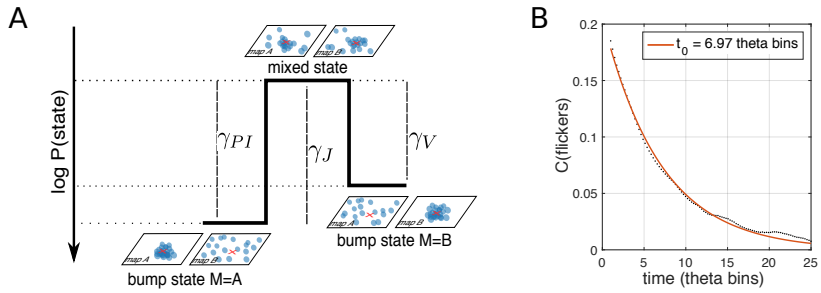


Figure 6.5: **CANN model for interplay between path integrator, external stimuli, and memory. A.** Representation of the effective model for the activity bump and effects of parameters. The input strengths, γ_{PI} and γ_V, contribute to push the hippocampal activity towards the corresponding cognitive states. Increasing the strength of recurrent connections, γ_J, results in an effective barrier separating the two collective hippocampal states, giving rise to well formed bumps in either map A (left) or in map B (right). Due to this effective trap the bump state remains localized in either map for more than a single theta bin. **B.** Temporal correlation of flickering events (theta bins with cognitive state opposite to external light conditions) decay over c.a. 7 theta bins in simulated data. Same model parameters as in Fig. 6.3.

Flickering frequency is constant throughout the conflicting phase, whose duration is exponentially distributed

Our model predicts that (1) the duration of the conflict phase, i.e. the time elapsed from a light switch to the subsequent PI realignment, is exponentially distributed (Fig. 6.6A, right panel); (2) during the conflict phase, the flickering frequency i.e. the percentage of theta bins identified as flickers, is constant and independent of time (Fig. 6.6B, right panel).

In order to test these two predictions on CA3 recordings, we introduce a method to disentangle the flickering dynamics of the cognitive map and the realignment of the PI, the latter bringing an end to the former. We first infer the most likely PI-realignment time for each light-switch event, given the sequence of identified flickers (Methods). The outcomes are shown as green lines in Fig. 6.6D, and correctly separate conflicting phases (rich in flickering events) from coherent periods (during which the hippocampal representation is much more stable). The distribution of conflicting phase durations is approximately exponential in agreement with model prediction (1), with decay time $\tau = 53$ theta bins (Fig. 6.6A, left panel). Dividing the test session into conflicting and coherent phases, we compute the frequency of flickers in the conflicting phase only. Consistently with the model prediction (2), the frequency of flickers is independent of the delay after the switch, with about 60% of theta bins in the conflicting phase carrying flickers (Fig. 6.6B, left panel).

Similar frequencies of flickers, close to one half, are obtained in the model when the two inputs have comparable strengths ($\gamma_{PI} \simeq \gamma_V$ in Fig. 6.4B, see also Fig. C in S1 Text). A testable consequence of this balance is that the distributions of the sojourn times (durations of the periods in which the neural activity persists in a cognitive map, see Methods) in map A and in map B are similar. This prediction is confirmed by a further analysis of the CA3 recordings: the two distributions of the sojourn times are both exponential, with roughly the same decay times (see Fig. E in S1 Text). This common time scale is related to the correlation time of the flickers (Fig. 6.2C & 6.4C), see Methods.

The combination of properties (1) and (2) explains the exponential decay in the frequency of flickers with the delay after the switch reported in [143] and [206]. While the frequency of flickers is constant and large in the conflicting phase, and constant and very low in the coherent phase, the duration of the conflicting phase is exponentially distributed. Hence the frequency of flickering theta bins, irrespectively of the phase, shows the same exponential decay, see Fig. 6.6C (right panel: simulated experiment, left panel: analysis of CA3 recordings). A detailed analysis of the data provides overwhelming statistical support to our two-fold explanation compared to a simple exponential decay of the flickering frequency (logarithmic likelihood-ratio test ~ 150, see Methods).

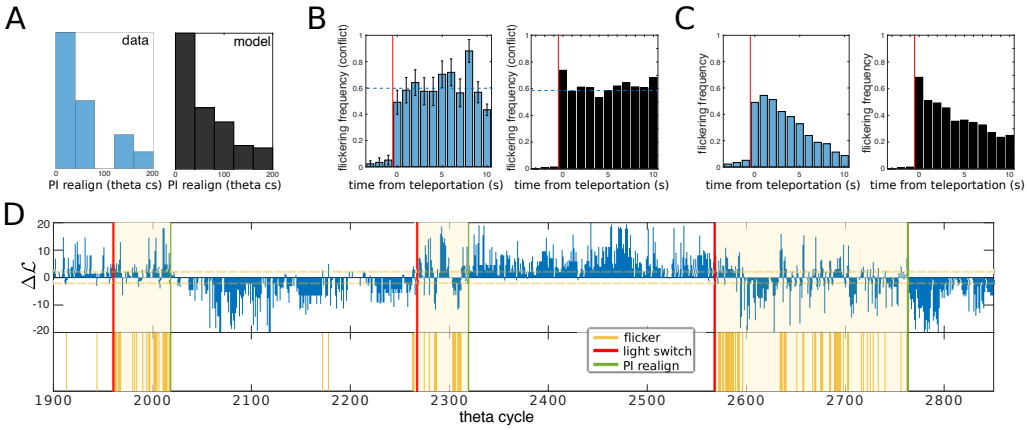


Figure 6.6: **Flickering rate is homogenous within each conflicting period. A.** Distribution of the path-integrator realignment times in a simulated session (right) and inferred from the recorded data (left). **B.** Mean Frequency of Flickers (MFF) during the conflicting phase, binned over 8 theta bins ($\sim$ 1 second) intervals. MFF is constant during the conflicting phase, both in the model (right) and in recordings (left). Realignment times inferred from data were obtained with a Bayesian procedure (Methods); histograms show flickering frequency in each time bin t normalized with respect to the fraction of conflicting phases, out of 15 teleportation events, that survived at least up to time t. **C.** Convolution of the two distributions shown in panels A and B, i.e. the MFF computed on the full test session, shows an apparent exponential decay both in the model (right) and in data (left, similar to the analysis shown in [143]). **D.** Map decoder output $\Delta\mathcal{L}$ and inferred PI-realignment times for the experimental test session; 3 out of 15 teleportations are shown. Light switches are marked with red lines, inferred PI-realignment times are marked with green lines. Identified conflicting periods are shaded. Simulated data were obtained with the same model parameters as in Fig. 6.3.

Neural encoding of position reflects the presence of input mismatches

Our model allows us to better understand the subtle differences between the neural encodings of position in the conflicting and the coherent phases. In the latter phase, both path-integrator and visual inputs point to the neurons with place fields overlapping the rodent position $\mathbf{r}$ in map. During the conflicting phase, the two inputs excite the two place-cell populations centered in $\mathbf{r}$ in their respective maps, respectively, $m = PI$ and $m = V$. Hence, while the bump of activity is mostly localized in one of the two maps (varying over time), some dispersion may be expected due to these incoherent inputs.

Mixed activity states, in which two (distinct) populations of neurons encoding the same position in the two maps are active, can be occasionally observed in the snapshots of the simulated activity in Fig. 6.4D, e.g. around theta bin $t = 80$. The over-dispersion present during the conflicting phase has two consequences. First, the accuracy in position encoding is expected to be lower in the conflicting phase that in the coherent phase, see Fig. 6.7A, left panel. Secondly, the loss in accuracy is not due to some random noise in the neural activity, but to a transient bump-like activity in the 'wrong' map, opposite to the decoded one. This effect is clearly seen when we choose the opposite map to infer the rodent position. While this choice leads to very poor prediction during the coherent phase, the positional error is significantly reduced during the conflicting phase (Fig. 6.7B, left panel).

To test these two predictions in CA3 recordings, we combine our positional analysis and our PI-realignment time inference procedure. In Fig. 6.7A (right), we compare the distributions of positional errors computed with the decoded map (according to the sign of $\Delta\mathcal{L}$) during conflicting and coherent phases (blue and red, respectively). Consistently with the model predictions, the positional error is significantly increased during the conflicting phase (ANOVA $p < 8 \times 10^{-8}$; conflicting: 14.7 ± 0.5 SEM, coherent: 12.3 ± 0.1 SEM). When computed with the opposite map, the positional error is obviously much higher than its counterpart computed with the decoded map, but a substantial decrease is found in the conflicting phase compared to the coherent phase, see Fig. 6.7B, right panel (ANOVA $p < 5 \times 10^{-23}$; conflicting: 27.8 ± 0.6 SEM, coherent: 34.2 ± 0.2 SEM), in full agreement with the model prediction. This effect also explains the relatively low value of the positional error obtained with the opposite map right after the switch, i.e. deep into the conflicting phase, compared to later times, see Fig. 6.3A&C. While this phenomenology is clear, it could in principle be affected by the presence of visual inputs projecting onto place cells during flickering events, i.e. when the 'opposite' map agrees with the external cues. In order to analyze the effect of the path integrator alone, we have restricted the analysis to theta bins whose decoded maps agreed with the visual inputs, i.e. to non-flickering theta bins. Results, shown in Fig. 6.7C&D, are still statistically significant and in strong agreement with the model predictions (decoded map: ANOVA $p < 1 \times 10^{-12}$; conflicting: 17.1 ± 0.7 SEM, coherent: 12.3 ± 0.2 SEM; opposite map: ANOVA $p < 1 \times 10^{-7}$; conflicting: 28.9 ± 0.95 SEM, coherent: 34.2 ± 0.3 SEM). Our findings are robust against changes in the statistical threshold L_0 for map decoding in the identification of conflict/coherent phases (Methods), see Fig. I in S1 Text.

The over-dispersion of the neural bump during the conflicting phase can also be observed from the reduction in (the absolute value of) the log-ratio, $|\Delta\mathcal{L}|$, see Eqn. (1). This quantity can be interpreted as a proxy for the completeness of the bump in one single map (Methods), larger $|\Delta\mathcal{L}|$ corresponding to large bumps in either of the two maps and randomly scattered activity in the other map (Fig. 6.4A&D). We find that the absolute value of $\Delta\mathcal{L}$ is significantly reduced during the conflicting phase in CA3 data, see Fig. 6.7E (left panel, ANOVA $p < 10^{-15}$; conflicting: 5.51 ± 0.16 SEM, coherent: 7.19 ± 0.08 SEM). Figure 6.7F shows the bimodal nature of the distributions of $\Delta\mathcal{L}$ in the conflicting and coherent phases. While the reference and coherent-phase distributions coincide, the conflicting-phase distribution is more narrow, due to the overdispersion of the bump (reference $\sigma^2 = 80.7$, coherent $\sigma^2 = 79.2$, conflict $\sigma^2 = 47.3$). This result provides further evidence for the predictive power of the CANN model.

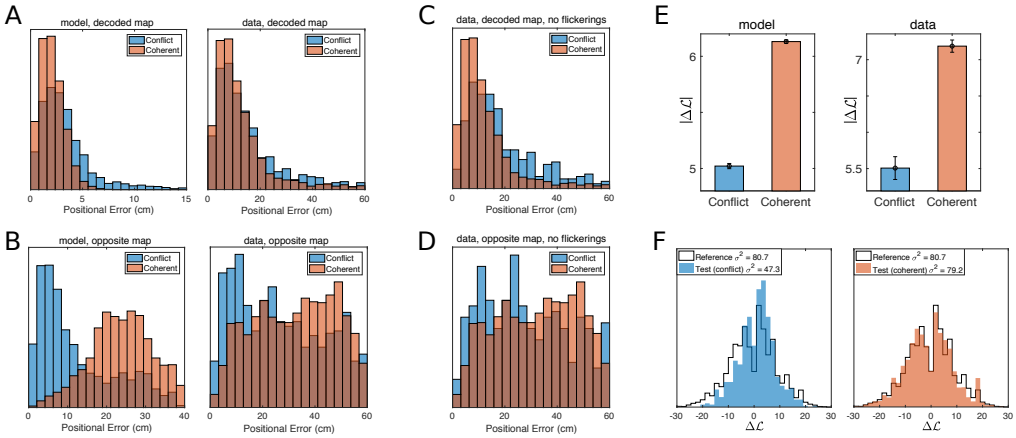


Figure 6.7: **Distributions of positional errors in conflicting and coherent phases. A.** Distributions of positional errors in the decoded cognitive map during coherent/stable (red) and conflicting/unstable (blue) phases. The model predicts (left panel) that the mean error is significantly increased during conflicting periods, mirroring the dispersion of the neural representation induced by the conflicting input streams. The same phenomenology is observed in recordings (right panel). **B.** The over-dispersion of the neural code in the 'correct' cognitive state (reported in panel A) is caused by an increased precision of the positional representation in the 'opposite' map, i.e. the one where the bump is not localized. This effect is significant in the model, left panel, and in the recorded data, right panel. **C, D.** Same analysis after exclusion of flickering theta bins. **E.** The absolute value of the log-ratio $\Delta\mathcal{L}$, Eqn. (1), is significantly reduced during the conflicting phase, both in the model (left) and in the data. $|\Delta\mathcal{L}|$ is a proxy for the completeness and stability of the bump (Methods). **F** Distributions of $\Delta\mathcal{L}$ in the test session during the conflicting (left, blue) and the coherent (right, red) phases, compared to the distribution in the two reference sessions (black contour, same as in Fig. 6.2A).

6.3 DISCUSSION

Our statistical inference-based data analysis allows us to quantify how well the CA3 neural activity encode various cognitive maps, and the position therein. Correlation-based procedures, e.g. used in [143], decode the cognitive state by comparing the instantaneous population activity to the average activity recorded in reference sessions at the same position of the rodent. Our functional-network based map decoder, instead, relies on the fact that the joint pairwise spiking activity of neurons is a fingerprint of the cognitive map [1, 4]. It does not need any knowledge of the sensory correlate (here, position), and could be used to decode generic brain states in other areas.

The fast dynamics of cognitive maps studied in [143] and here results from an unrealistic sensory situation. Imposing artificial conflicts between inputs and studying their consequences is a standard approach to unveil the circuitry underlying the processing of multimodal sensory information in the hippocampus [124, 203] as well as in other brains areas, see for instance [207] for an illustration in the primary visual cortex where mismatches involve sensory and motor inputs. However, fast retrieval of functionally relevant maps, characterized by grouping and cognitive control, has also been observed in realistic settings, in which a behaving animal is required to maintain representations of two distinct spatial frames [141].

The position of the animal was accurately inferred at all times from the spiking activity using the place fields of the retrieved cognitive map (Fig. 6.1B). As a main finding, we show that the hippocampus maintains high-quality encoding of the position even if the contextual variable undergoes fast dynamical changes. This is explained in the model by the fact that inputs point to place cells coding for the physical position in both competing maps (Fig. 6.4A), and that the bump of activity is most often localized around these place cells in one map, and scattered all over the other map. Similar findings were reported in [143] (main text, Fig. 3d and Supplementary Text Fig. 8), within a statistical framework assuming a priori the consistency of positional representation during flickering events, as the cognitive map was decoded by comparing the neural activity to the mean-activity vectors at the recorded real position of the rat. The emergence of unambiguous, non-mixed representations was also underlined in [143], and shown to take place in the second half of the theta cycle. However, the detailed analysis of the CA3 recordings and of the model data shows a loss of quality of the bump state (reduction in absolute value of log-ratio $|\Delta\mathcal{L}|$) and an increased quality of position decoding in the opposite map (Fig. 6.7), providing evidence for the presence of partially mixed states.

Our model for the retrieval of hippocampal cognitive maps in the presence of inputs from the path integrator and visual cues is based on CANN theory [123, 129]. Two-dimensional CANN attractors, were previously applied to networks of place [154, 156] and grid [169, 208] cells. Indirect experimental evidence supporting CANN is now accumulating in various animals and brain areas. Evidence of a ring-shaped attractor region associated to head direction representation was recently reported in the drosophila central brain [170]. Attractor dynamics has also been associated to behavioral observation in a study on the monkey prefrontal cortex [171]. As for space representation, experimental

support for attractor behaviour has been found in hippocampal CA1 [172] as well in grid cell [111] recordings. Further indirect evidence is provided by the pattern of connectivity in CA3, compatible with its functional role as an auto-associative attractor network [199], as suggested long ago based on anatomical and computational considerations [189, 198], and by the active nature of dendrites of mEC neurons, which enhances the robustness of attractors under environmental changes [209].

The detailed analysis of the CA3 recordings done here provides another indirect support for CANN theoretical framework, when multiple (two) cognitive maps are memorized. Memorization of the two attractors is obtained, in the model, by adding the corresponding connectivity matrices into the unique CANN connectivity matrix [155, 156, 160, 201]. A detailed theoretical study of the mechanisms for transition from map to map was obtained in the absence of inputs, i.e. for spontaneous transitions induced by neural noise only [12]. A similar picture is found here in the presence of visual-cue inputs pointing to the 'new' map, while the path-integrator inputs point to the 'old' map in the conflicting phase. As inputs are of comparable magnitude, no single map is favored. The stochastic fluctuations resulting from the noise of the individual neurons are sufficient for the system to cross the activation barrier between the two memory states (maps) of the network, see Fig. 6.4B and Fig. A in S1 Text. The hippocampal network jumps intermittently from one cognitive map to the other, reproducing the flickering events experimentally identified and described in [143]. Transition rates between the two maps increase with the neural noise, modeled here by the parameter β, see Eqn. (6.12) in Methods and Fig. D in S1 Text. Neural noise relative to the population activity could also be effectively increased through the introduction of periodic (theta and gamma) modulations of the activity into the model [157, 158, 210]. The presence of rhythms is known to facilitate memory formation and integration of information [211, 212]. While theta oscillations can help produce flickering events as previously reported [157, 206], our work shows that such periodic modulations are not necessary. Transitions could also be facilitated by particular 'confounding' landmarks or positions in space, where the maps happen to be locally similar [12, 155].

The present model reproduces accurately all the observed flickering properties, without any need for a post-learning short-term plasticity of the CA3 network hypothesized in [206]. In particular, our model predicts that the flickering frequency is independent from the time spent after the teleportation event in the conflicting phase (Fig. 6.6B). This finding is at first sight in disagreement with the exponential decay of the flickering frequency reported in [143, 206]. However, the latter was obtained as a result of an averaging over many teleportation events. For a single event, accurate data analysis shows that our constant flickering rate hypothesis, when combined with the exponentially distributed realignment time of the path integrator (Fig. 6.6A), is much more likely than an exponential decreasing scenario.

Our model is based on the existence of two streams of inputs conveying, respectively, external landmark and self-navigation information. Recent studies have pointed to the grid cells network in mEC as the possible region that supports path integration, as their firing patterns are maintained in the dark [85], and the relative phases of grid cells

seem to be largely unaffected by global remapping between environments of similar shapes [108, 213]. CANN-based approaches have been proposed to model grid-cell networks [169, 208], differing from hippocampal CANN mostly by the short-range nature of the inhibitory couplings. In much the same way the microscopic hippocampal CANN proposed here can effectively be reduced to a 2-state model (Fig. 6.4B), we expect CANN models for the grid-cell networks to be approximately described by a 2-state model, corresponding to the PI aligned with map A or B [108, 214, 215]. This motivates the simple model for the PI we have considered here.

In addition to sending projections towards the hippocampus, our model PI receives a feedback from the CANN, greatly increasing the probability of transition to the state agreeing with the instantaneous cognitive map [102, 123, 216]. Eventually, the state of the PI is realigned along the visual cue inputs, which stops the conflicting phase. Our model effectively implements a ratchet mechanism, locking the system into the coherent phase after a conflicting transient. Realignment of the path integrator based on visual landmarks is an important functional property, intended to limit the accumulation of errors in position estimation [217], and observed for large mismatch between external and internal inputs [126]. From a physiological point of view, projections exist from CA1 to mEC [123], and have been shown to be important for the formation of grid cells [216]. Hence, the feedback from the CANN, thought here to model CA3 activity, to the PI should be understood as effective.

As recently reported in [213], the impairement of the mEC grid firing resulted in a loss of path integrator in behaving rodents. As in our model, the recall of the pre-teleportation map, and, therefore, the whole flickering phenomenology are driven by the input stream from the path integrator to the CA3 network, we conjecture that flickering instabilities would disappear upon grid-cell impairement. Simultaneous recordings of mEC and CA3, as in [108], would be extremely useful to test our predictions and better describe the effect of the path integrator on the cognitive status of CA3.

6.4 METHODS

Cognitive map decoding from neural activity

Theta bins are identified with the Hilbert transform procedure of [143]. The activity of the N recorded neurons is binarized into each theta bin t: $s_{i,t} = 1$ if neuron i is active in bin t, 0 otherwise. For each cognitive map $m = A, B$ a *Ising-model* probability distribution $P_m(\mathbf{s})$ for the activity configurations $\mathbf{s} = (s_1, s_2, ..., s_N)$ is inferred,

$$P_m(\mathbf{s}) = \frac{1}{\mathcal{Z}_m} \exp\left(\sum_i h_i^{(m)} s_i + \sum_{i<j} J_{ij}^{(m)} s_i s_j\right) , \tag{6.2}$$

where $\mathcal{Z}_m$ is a normalization constant. Couplings (J^m) and fields (h^m) are determined such that the pairwise correlations and average activities in the neural population computed from P_m match their experimental counterparts in the reference session of environment m. These *inverse Ising problems* are solved using the Adaptive Cluster Expansion algorithm [33, 45, 46, 177]. The inferred models (6.2) are then used to dynamically decode the map m_t during the test session (**s**) [1], based on the log-ratio of the probabilities of the activity configuration in time bin t in the two environments (main text Eqn. [1]), with the result

$$m_t = \begin{cases} \text{A} & \text{if } \Delta\mathcal{L}(\mathbf{s}_t) > L_0 , \\ \text{B} & \text{if } \Delta\mathcal{L}(\mathbf{s}_t) < -L_0 , \end{cases} \tag{6.3}$$

where the threshold L_0 is chosen according to the required statistical confidence. We generally set $L_0 = \log 10 \simeq 2.3$.

After having decoded the map m_t in theta bin t, we define the flicker variable f_t, equal to 1 if m_t does not match the light cue in theta bin t, to 0 otherwise.

Temporal correlation of flickers and sojourn times

The time correlation of flickering events for delay τ is defined as

$$C(\tau) = \frac{1}{T_{tot}} \sum_{i=1}^{S-1} \sum_{t=T_i}^{T_{i+1}-\tau} f_t f_{t+\tau} - \left(\frac{1}{T_{tot}} \sum_{i=1}^{S-1} \sum_{t=T_i}^{T_{i+1}-\tau} f_t\right)^2 \tag{6.4}$$

where S is the total number of switch events in the recorded data ($S = 16$ in [143]), T_i is the theta bin index of switch i ($< S$), and $T_{tot} = T_S$ is the total number of theta bins in the test session. The time correlation $C(\tau)$ is typically exponentially decaying, with a decay time τ_0, see Fig. 6.2C.

The correlation time τ_0 is related to the sojourn time of the neural bump in the cognitive maps, defined as a sequence of contiguous theta bins decoded in the same map, see S1 Text. Theta bins whose $|\Delta\mathcal{L}|$ are lower than the threshold L_0 are considered as belonging to the same map as the last statistically significant time bin. The distribution of sojourn times in each map is shown in Fig. E in S1 Text.

Position decoding from neural activity

The arena is discretized into 60×60 squared bins of 1 cm^2 each, with integer coordinates (x, y) [143]. For each reference environment $m \in$ (A, B) we construct the binary rate map, $p_i^{(m)}(x, y)$, equal to the average of $s_{i,t}$ over all theta bins t in which the rat is at position (x, y). Position is then decoded according to the naive Bayesian framework [218]: the probability of the activity configuration $\mathbf{s}_t = \{s_1, s_2, ..., s_n\}$ in theta bin t and at fixed position (x, y) reads

$$P_m(\mathbf{s}_t|x, y) = \prod_{i=1}^{N} \left[p_i^{(m)}(x, y) \cdot s_i + \left(1 - p_i^{(m)}(x, y)\right) \cdot (1 - s_i) \right] . \tag{6.5}$$

Once m is known, e.g. either through the map decoder or due to constant experimental conditions, the position of the rodent can be reconstructed from the recorded neural activity through

$$(\hat{x}_t, \hat{y}_t) = \arg\max_{(x,y)} \left[P_m(\mathbf{s}_t|x, y) \times T_m(x, y) \right] , \tag{6.6}$$

where the maximum is computed over the 60×60 possible positions. $T_m(x, y)$ is the number of theta bins spent by the rodent at position (x, y) during the reference session of map m; we use it as a prior to favor positions where the rodent is more likely to be, irrespectively of the neural activity.

Continuous Attractor Neural Network model for hippocampal activity

The hippocampal population includes N place cells. For each cell $i = 1...N$ the place-field centers coordinates, $\mathbf{r}_i^A$ and $\mathbf{r}_i^B$, are drawn uniformly and independently at random in the squared environments, respectively, A and B. The linear size of each square is denoted by L.

Neural activities are represented by binary variables: $s_{i,t} = 0$ or 1 if neuron i is, respectively, silent or active in time bin $t = 1, 2, 3, ...$ The duration of a time bin is the theta cycle over K; results reported here were obtained with $K = 4$, which corresponds to approximately 30 ms.

The total input received by neuron i at time t is

$$H_{i,t} = \sum_{j \neq i} J_{ij}\, s_{j,t} + h_i^{(V)}(\mathbf{r}) + h_i^{(PI)}(\mathbf{r}) . \tag{6.7}$$

The three terms on the right hand side of Eqn. [6.7] represent, in order:

- the input due to recurrent connections in the hippocampal network. The underlying assumption is that connections have emerged from learning during the exploration of the two environments by the rodent: Place cells that turned out to be simultane-

ously active in either environment have developed positive couplings. Couplings are defined through

$$J_{ij} = J_{ij}^A + J_{ij}^B \quad \text{with} \quad \begin{cases} J_{ij}^A = \gamma_J \times \phi(\mathbf{r}_i^A - \mathbf{r}_j^A) \\ J_{ij}^B = \gamma_J \times \phi(\mathbf{r}_i^B - \mathbf{r}_j^B) \end{cases} , \tag{6.8}$$

where γ_J controls the strength of the connections, and

$$\phi(\mathbf{r}) = \frac{L^2}{N \times 2\pi\sigma^2} \, e^{-\frac{|\mathbf{r}|^2}{2\sigma^2}} . \tag{6.9}$$

Parameter σ in Eqn. [6.9] is the spatial scale corresponding to the width of the place fields. The prefactor in Eqn. [6.9] ensures that ϕ is dimensionless and that, on average, the sum of the Gaussian factors $\phi(\mathbf{r} - \mathbf{r}_i^m)$ over all neurons i is close to unity for every possible position $\mathbf{r}$ and map m.

- the visual input

$$h_i^{(V)}(\mathbf{r}) = \gamma_V \times \begin{cases} \phi(\mathbf{r}_i^A - \mathbf{r}) & \text{if } V = A \, , \\ \phi(\mathbf{r}_i^B - \mathbf{r}) & \text{if } V = B \, , \end{cases} \tag{6.10}$$

  depends on the position $\mathbf{r}$ of the rodent. We again assume that, during the exploration of the environment, visual-cue projections onto place cells have been strengthened through learning. For simplicity we use the same function ϕ as in the recurrent connections, see Eqn. (6.8), to characterize the portion of environment in which visual cues project onto a specific place cell i.

- the path-integrator input

$$h_i^{(PI)}(\mathbf{r}) = \gamma_{PI} \times \begin{cases} \phi(\mathbf{r}_i^A - \mathbf{r}) & \text{if } PI = A \, , \\ \phi(\mathbf{r}_i^B - \mathbf{r}) & \text{if } PI = B \, . \end{cases} \tag{6.11}$$

  PI inputs have the same functional dependence over space as visual-cue related inputs. The amplitudes of both input types are tuned by the parameters γ_{PI} and γ_V.

All neurons undergo stochastic updating of their activities from time bin $t \to t+1$ according to their total inputs. The activity of neuron i at time $t+1$ is chosen to be

$$s_{i,t+1} = \begin{cases} 0 \text{ with probability } \frac{1}{1+e^{\beta(H_{i,t}-\theta)}} \\ \\ 1 \text{ with probability } \frac{e^{\beta(H_{i,t}-\theta)}}{1+e^{\beta(H_{i,t}-\theta)}} \, . \end{cases} \tag{6.12}$$

To enforce global inhibition in the population activity, the value of the threshold θ is dynamically adjusted so that an average fraction f of the neurons is active at any

time. Parameter β controls the amount of noise in the neural dynamics. For $\beta \to 0$ neuron activities are random and independent of their inputs, while, for $\beta \to \infty$, they deterministically follow the signs of the inputs (after subtraction of the threshold θ). The average activity of cell i at time $t+1$ is therefore a monotonously increasing sigmoidal function of its total input $H_{i,t}$ at time t, with maximal slope equal to $\beta/4$ in $H_{i,t} = \theta$.

The properties of this CANN model in the absence of any visual and PI inputs, i.e. for $\gamma_V = \gamma_{PI} = 0$, were analytically studied in [12, 158, 160], see S1 Text for further discussion. The log-ratio $\Delta\mathcal{L}$ defined in Eqn. (1) for the decoding of cognitive maps has a direct counterpart in our CANN model as the difference between the contributions to the log-probability of an activity configuration $\mathbf{s}$ when the bump is localized in maps A and B,

$$\Delta\mathcal{L}(\mathbf{s}) = \sum_{i<j} (J^A_{ij} - J^B_{ij})\, s_i\, s_j \ . \tag{6.13}$$

Mechanism for path-integrator realignment

The path integrator is described as a two-state model, $PI = A$ or B. Its dynamics is stochastic and Markovian: in each time bin t, the state PI can jump into state PI' with transition probabilities $R(PI \to PI')$, independently of the previous states. The feedback from the hippocampal network to the path integrator is expressed in the dependence of R on the hippocampal map M_t at time t. To favor transitions to the state PI' agreeing with the current map M_t, we introduce the following witness function for the presence of the bump in map $m = A, B$:

$$W^{(m)}(\mathbf{s}, \mathbf{r}) = \sum_{i=1}^{N} s_i\, \phi\left(\mathbf{r} - \mathbf{r}_i^m\right) \ , \tag{6.14}$$

where $\mathbf{s}$ and $\mathbf{r}$ are, respectively, the activity configuration and the position of the rodent at time t. Due to the normalization of ϕ in Eqn. [6.9], we expect $W^{(m)}$ to be close to one for the retrieved map $m = M_t$ and to be much smaller for the opposite map.

We impose the preference for realigning the path-integrator state in accordance with the hippocampal map through the ratio between the two reciprocal transition probabilities,

$$\frac{R(\ PI = B \to PI' = A\ )}{R(\ PI = A \to PI' = B\ )} = e^{\gamma_W \left(W^{(A)}(\mathbf{s},\mathbf{r}) - W^{(B)}(\mathbf{s},\mathbf{r})\right)} \ . \tag{6.15}$$

Here, γ_W is a positive parameter allowing us to tune the strength of the preference. If the hippocampal bump of activity is localized in, say, map A, the right hand side of Eqn. [6.15] will be strongly positive, and the probability of realigning the path integrator to $PI' = A$ will be much larger than the probability of the reciprocal transition.

A solution to the constraint expressed by Eqn. [6.15] is given by

$$\begin{aligned} R(\ PI = B \to PI' = A\ ) &= R_0 \times e^{\gamma_W \left(W^{(A)}(\mathbf{s},\mathbf{r}) - W^{(B)}(\mathbf{s},\mathbf{r})\right)/2} \ , \\ R(\ PI = A \to PI' = B\ ) &= R_0 \times e^{-\gamma_W \left(W^{(A)}(\mathbf{s},\mathbf{r}) - W^{(B)}(\mathbf{s},\mathbf{r})\right)/2} \ , \end{aligned} \tag{6.16}$$

where R_0 is a positive number. In the absence of bias ($\gamma_W = 0$), the inverse of R_0 may be interpreted as the average time scale between two realignments of the path-integrator state.

The model for the path-integrator dynamics is entirely defined by the transition probabilities in Eqn. [6.17] and the probability conservation identities:

$$\begin{aligned} R(\ PI = A \to PI' = A\ ) + R(\ PI = A \to PI' = B\ ) = 1\ , \\ R(\ PI = B \to PI' = A\ ) + R(\ PI = B \to PI' = B\ ) = 1\ . \end{aligned} \tag{6.17}$$

Inference of path-integrator realignment times

Defining τ as the PI-realignment time, $t = 0$ as the time bin corresponding to the light switch and T as the time bin corresponding to the next switch (end of analyzed data), we assume the probability $p(t)$ for time bin t to be a flickering event to be

$$p(t) = \begin{cases} p_0, & \text{if } 1 \le t \le \tau\ , \\ p_e, & \text{if } \tau + 1 \le t \le T\ . \end{cases} \tag{6.18}$$

Here, p_0 is the constant flickering probability, and p_e is the baseline decoding error, see S1 Text for discussion of the values of parameters p_0 and p_e.

We write the log-likelihood of the parameter τ as a function of the identified flickering sequence $\mathbf{f} = \{f_t\}$ as follows:

$$\begin{aligned} \log P(\mathbf{f} \mid \tau, p_0, p_e) = \log p_0 \times \textstyle\sum_{t=1}^{\tau} f_t + \log(1-p_0) \times \sum_{t=1}^{\tau}(1-f_t) + \\ + \log p_e \times \textstyle\sum_{t=\tau+1}^{T} f_t + \log(1-p_e) \times \sum_{t=\tau+1}^{T}(1-f_t) \end{aligned} \tag{6.19}$$

We then maximize this log-likelihood over τ to infer the most likely value τ^* of the realignment time. The procedure is repeated for all light switches, see Fig. G in S1 Text.

Independence of frequency of flickers from delay after light switch

We consider two hypothesis:

(a) H_{decay} = the flickering probability depends on time as a decaying function that can be inferred from data, that can be inferred from the full test session (15 lght switches)

(b) $H_{constant}$ = the flickering probability is constant throughout the conflicting period of varying duration, which can be inferred from data (see previous section).

Following the Bayesian information criterion [219], we parametrize each model with the same number of variables. For hypothesis H_{decay}, we estimate the flickering frequency as a function of time from the average frequency computed over the full test session (Fig. 6.6C, bottom), in bins of one second-width (8 theta cycles), up to 15 seconds after the light switch. For later delays (>15 s) the flickering frequency is set to a baseline error

probability, $p_e = 0.01$. For hypothesis $H_{constant}$, we infer the most likely PI realignment times for each one of the 15 light-switch events (see Section above). The associated flickering probability p_t is then set to $p_0 = 0.55$ until the inferred PI realignment time τ^*, and equal to the baseline probability $p_e = 0.01$ afterwards. We then compute the likelihoods of both hypothesis given the observed data (identified flickering theta bins $\mathbf{f}$) through

$$\ell(\text{hypothesis}|\text{data}) = \sum_{t=1}^{T} \Big[\log p_t \times f_t + \log(1-p_t) \times (1-f_t) \Big] , \tag{6.20}$$

where T is the total length of analyzed session (number of time bins between two consecutive light switches). The above expression is then summed over all light-switch events. We define the difference of the two log-likelihoods as

$$\Delta\ell = \ell(H_{constant}|\text{data}) - \ell(H_{decay}|\text{data}) . \tag{6.21}$$

The constant flickering frequency hypothesis $H_{constant}$ is extremely more likely ($\Delta\ell \sim 150$) than the decaying model H_{decay}. The result is robust against changes in the parameters, see Fig. F in S1 Text.

ACKNOWLEDGMENTS

We are grateful to K. Jezek for providing us with the data of [143] and for fruitful discussions. We thank C. Schmidt-Hieber for insightful discussions and comments. We thank F. Rizzato, J. Tubiana, and S. Wolf for critical reading of the manuscript.

7

PAIRWISE MODELS INFERRED FROM HIPPOCAMPAL ACTIVITY GENERATE NEURAL CONFIGURATIONS TYPICAL OF SINGLE OR MULTIPLE LOW-DIMENSIONAL ATTRACTORS

7.1 INTRODUCTION

In chapters 5 and 6 we investigated the task of decoding the cognitive map from the observation of the neural activity of a hippocampal population. We framed the decoding problem in a Bayesian setup: from the statistical properties of neural patterns observed in controlled conditions (in our case, the two cognitive maps of the two environments A and B) we were able to classify patterns of activity registered in unknown external conditions [1, 2]. The statistical model used for the binary classification was the Ising model accounting for empirical pairwise correlations of neurons. Neural correlations, more than the average activities, can be considered as "fingerprint" of the expressed cognitive map [2], especially in the presence of low-orthogonality between firing rates of the two cognitive states (e.g., rate remapping in CA1). The decoding was performed by inferring two models, parametrized with map-specific fields $h_i^{A,B}$ and functional couplings $J_{ij}^{A,B}$,

$$P(\mathbf{s} \mid M) = \frac{1}{\mathcal{Z}^M} \exp\left(\sum_i h_i^M s_i + \sum_{i<j} J_{ij}^M s_i s_j\right) \quad , \tag{7.1}$$

and using them to score the likelihood of a new pattern given one or the other cognitive map as the log-probability of the pattern given the map $M \in [A, B]$. The max-entropy principle [22] ensures that the graphical Ising model is the least biased statistical model constrained to first and second-order statistics, i.e., average activations and correlations. Other approaches that rely on the definition of a functional connectivity between neurons have been proposed to fit a probabilistic model on population activity, which include RBMs, GLM, integrate-and-fire [183, 185–187].

An important feature of statistical models for population activity is that they can be used to *generate* new neural patterns, by Monte Carlo sampling of the corresponding probability distribution [220]. One important question concerns the generative power of the models, i.e., how well they can produce new patterns that are functional for the cognitive task. This question is central for future applications attempting at creating devices to emulate the activity of brain areas (Brain-Computer Interfaces). With few exceptions [221], very little has been done, so far, in this direction. In the case of the Ising model, it is known that 3-body correlations are well-reproduced [51], though there are deviations [50], despite being not directly included in the inference. Conversely, some observables are not necessary fitted well by pairwise models (e.g., the number of active

neurons) and ad-hoc modifications of models have been proposed [222]. Yet, it is unclear which observables matter functionally, so the question of whether pairwise Ising models are generative or not remains open.

Here we address this issue in the case of the Ising models inferred from hippocampal place-cell recordings analyzed in Chapter 6 [2, 143]. Place cells encode for the spatial location of the rat in the environment, as well as for contextual information, such as olfactory stimuli, emotional state, or task to be performed [83–88]. Therefore, in the case of hippocampal activity, functionality has a precise meaning regarding the spatial correlate of the activity.

We generate activity patterns by simulating the inferred model with Metropolis Monte Carlo sampling and test them on their spatial correlate. We show that the generated activity configurations are meaningful, in that they code for positions of a virtual animal in the environment. We also show that a model inferred from the joint collection of patterns from the two distinct environments generates bimodal activity configurations, coding for well-defined position in one of the two environments only, with spontaneous stochastic transitions from one map to the other. Therefore, the inferred Ising model captures and preserves the functional relation between neurons, on a coarse-grained scale, that can be used to generate new activity patterns that are functional for navigation within one or more cognitive maps.

7.2 ATTRACTOR-LIKE BEHAVIOR OF THE INFERRED MODEL: SINGLE MAP

Functional meaning of inferred couplings J_{ij}

In the continuous-attractor neural network theory (see Chapter 3.2), the neural dynamics, during the exploration of a familiar environment, is confined to the manifold of the corresponding cognitive map. As a consequence, the neural correlations fitted by the Ising inference will reflect structural information (synaptic connection) as well as functional one (being confined to the attractor). As suggested in Chapter 4, the smaller is the subpopulation that we observe, the more the correlation will be dominated by the functional constraints. In our case (~ 30 neurons over $\sim 100,000$), we thus expect the inferred couplings J_{ij} to primarily encode functional aspects of the neural population, leading to a model able to generate patterns that are functional if interpreted within the cognitive map.

Coherently with results on theoretical models [5] (see Chapter 4), we find that the inferred couplings have a direct relation to the distance between the centers of place fields, see Fig. 7.1. Couplings are positive at small distances (comparable to place-field size), and negative at larger distances. This behavior of the connectivity is observed for both maps and is compatible with attractor models of spatial navigation, where short-range excitation leads to the formation of a localized activity bump while long-range inhibition keeps a stationary average activity.

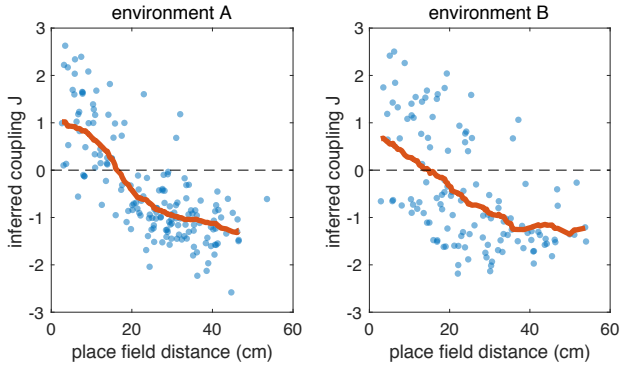


Figure 7.1: **Couplings between pairs of neurons vs. distances between their place-field centers** computed on reference sessions from CA3 recordings of Jezek et al. [143], discretized and binarized in time bins of $\Delta t = 120$ ms. Red lines are obtained by sliding average on the x-axis.

Generated activity configurations are diverse and statistically consistent

To test the generative power of the inferred model, we sampled, via Metropolis Monte Carlo, from the model inferred from discretized ($\Delta t = 120$ ms) neural activity sampled during the exploration of a single environment (A). As a first analysis, we tested the diversity of the generated activity with respect to the training session by computing the minimum Hamming distance of each generated pattern $\widehat{\mathbf{s}}$ to the training set (the experimental reference session), defined as

$$\min_{\mathbf{s}\in\text{trainig set}} |\widehat{\mathbf{s}} - \mathbf{s}| \quad , \tag{7.2}$$

where $|\mathbf{a} - \mathbf{b}|$ is the Hamming distance between the words $\mathbf{a}$ and $\mathbf{b}$. As shown in Fig. 7.2 (left), the generated patterns are diverse and do not merely reproduce the training set. We then tested how this distribution of Hamming distances compares to the one computed between batches of real data. We performed the same analysis comparing the second half of the experimental reference session to the first one (Fig. 7.2, center), obtaining a striking similarity to the distribution obtained from generated activity.

To assess the statistical consistency of the generated patterns, we computed the distribution of the log-likelihood (i.e., the energy of the inferred model, up to a constant) and compared it with the one computed on experimental recordings. Again, we observe a substantial similarity between the two distributions, see Fig. 7.2 (right panel).

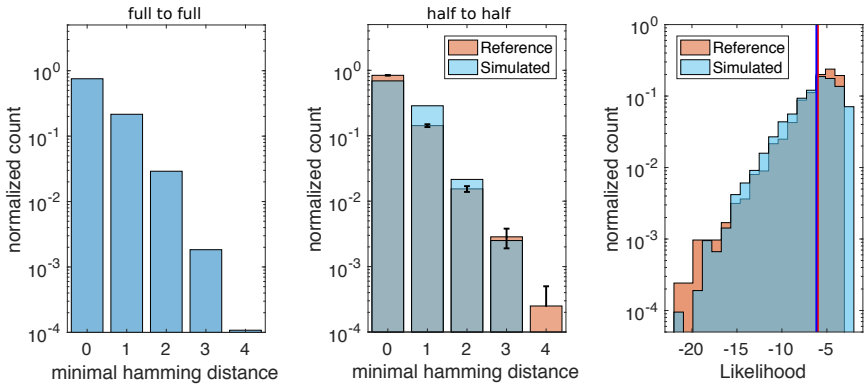


Figure 7.2: **Generated activity configurations are diverse and statistically consistent. Left**: distribution of minimal hamming distances between $T = 10,000$ generated patterns and the training (experimental) reference recordings. **Center**: same analysis but with $T = 1924$ generated patterns and the first half of the experimental reference recordings ($T = 2000$ patterns, in blue), superposed to the same quantity computed between patterns in the second half of the experimental reference recordings ($T = 1924$, shown in red). The error bars refer to standard error from the mean over $n = 10$ different training-test partitions of the real reference session. **Right**: distribution of log-likelihood in the generated (blue) and real (red) data.

Generated patterns are spatially selective, coding for localized positions and smooth trajectories in the environment

To assess the functionality of the generated activity in terms of spatial selectivity, we employed a naive Bayesian decoder for the position [204] for each pattern $\widehat{\mathbf{s}}$ of the generated session. The standard deviation σ of the spatial posterior $P(x, y \mid \widehat{\mathbf{s}})$, measured in *cm*, can be used as a proxy for the dispersion of the bump of activity. As shown in Fig. 7.3, the dispersion of the generated activity is comparable to the one obtained by neural patterns in the reference session. In contrast, patterns generated by an independent-site model show a sensibly-higher dispersion (see Fig.7.3, right panel), highlighting the importance of pairwise couplings in capturing and generalizing the functional (spatial, in this case) relation between the recorded neurons.

We then tested if the decoded position, defined as the probability mass center of the posterior, displays a continuity in time that is comparable to the one observed in the data. Since velocity is a dynamical property, we need to choose a scale factor between simulation time and real time in order to compare the results of the simulations with real data. The dynamics of the real neural system is driven by both the structural connectivity and external factors, such as visual inputs and the path integration, and is therefore linked to the running speed and behavior of the rat. Therefore, the definition of a precise relationship between real data and Monte Carlo simulations, which we

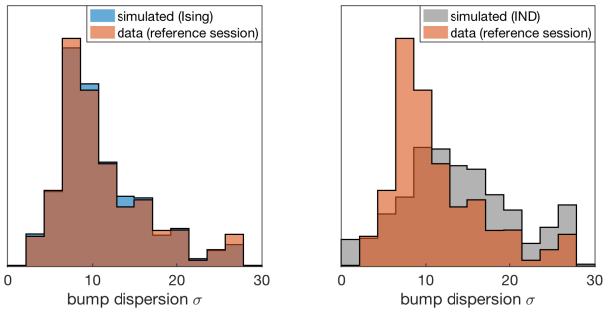


Figure 7.3: **Spatial selectivity of new generated activity patterns.** Distribution of bump dispersion σ, defined as the standard devation of the posterior of position inference and measured in cm, for patterns generated by the Ising model (left) and the independent-site model (right), compared to the one obtained from real data. The $\sim$ 20% (Ising) and $\sim$ 50% (IND) of the generated patterns are new, i.e., not present in the training batch.

might interpret more as spontaneous activities of the network, poses several challenges and is still a work in progress. For the following results, we choose the scale factor $\tau_{MC} = 5$, i.e., we perform 5 Metropolis step between two consecutive sampled activities. As shown in Fig. 7.4 (left panel), this choice is consistent in terms of average inferred spatial distance after a single time bin, that we interpret as the "velocity" of the simulated rat. Intuitively, if we break the temporal structure, by shuffling the activities in time, the distribution becomes comparable to the one expected from random points in a confined environment (Fig. 7.4, central panel). To assess the importance of the precise low-dimensional spatial relationship between neurons, we shuffled the neuron index of the activity before decoding the velocity from it. This procedure changes the spatial correlate of each neuron while keeping the time-correlation in the neural space, which is a function of the simulation time scale τ_{MC} that we have previously fitted. As reported in Fig. 7.4 (right panel), except for the zero-velocities (which are maintained independently on the neuron index), the velocities inferred from the shuffled activity are significantly higher than the ones retrieved with the correct activity-to-chart mapping.

Overall, these analyses suggest that the generated activity is confined to a manifold of a dimensionality comparable to the one where real neural patterns live and that this confinement is deeply associated to the spatial relations between neurons on the chart. An analysis of the effects of τ_{MC} on this and other dynamical observables is in progress and will be discussed in a future publication.

Stimulation of a single neuron causes neighbors to fire leading to a localized bump

In theoretical models, a property of the activity bump is that, during the navigation of a continuos-attractor manifold, it can be driven by external inputs to be positioned in a specific location on the chart. To test the response of the inferred Ising model to

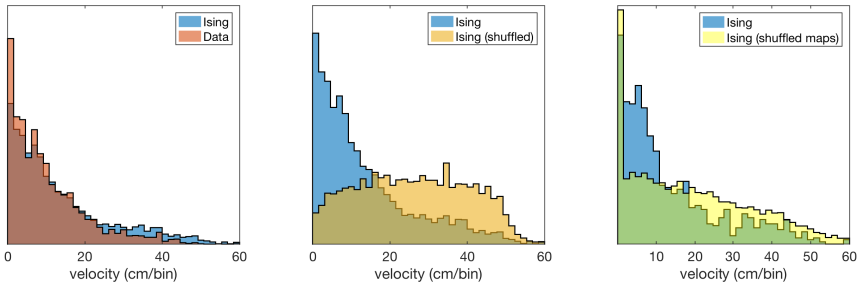


Figure 7.4: **Generated activity represents a continuous position in time. Left:** Distribution of velocities inferred from patterns generated by the Ising model, compared to the one obtained from real data, $T = 10,000$ generated time-bins. **Center:** same generated data of left panel compared with velocities obtained by breaking the time coherence (random shuffle in time). **Right:** Velocity inferred from generated data ($T = 1000$) compared to velocity inferred after a shuffle of the neuron index in the activity vectors. Data obtained with $n = 20$ different repetitions of the random shuffling. $\tau_{MC} = 5$.

positional stimulations, we performed a "pinning" test by forcing one specific neuron s_i to be always active, generating patterns from the conditioned distribution

$$P(s_1, s_2, \ldots, s_{i-1}, s_{i+1}, \ldots, s_N \mid s_i = 1) \quad . \tag{7.3}$$

Each neuron s_i is associated with a specific location (x_i, y_i) by the position of its place field. We analyzed the generated patterns by applying the same naive Bayesian decoder for the position, excluding the $i - th$ neuron from the procedure. In Fig. 7.5 (left) we show the $2D$ histogram of the decoded positions, normalized with the one obtained in un-pinned conditions, for 6 neurons located in different regions (red cross) of the squared environment. The distribution of positions decoded from the pinned activity is clusterized around the location of the stimulated neuron, often in a well-shaped unimodal form.

In Fig. 7.5 (right panel) we show the distribution of the distances between the decoded positions from pinned activity and the stimulated-cell position (x_i, y_i) for all "A-like" (i.e., non-silent in the recorded environment) neurons. A comparison with the same analysis done in un-pinned conditions (both analyses exclude cell i from the position-decoding procedure) shows a strong quantitative confirmation of the effect of the stimulation.

Therefore, the excitation of a neuron s_i biases the activity of the remaining cells to be localized around (x_i, y_i) effectively "pinning" the bump around the excited place field. This result gives additional support to the capability of the Ising model to reproduce a phenomenology that is compatible with the attractor picture, despite the strongly-reduced number of modelled functional units from the neural population (34 out of $\sim 10^5$).

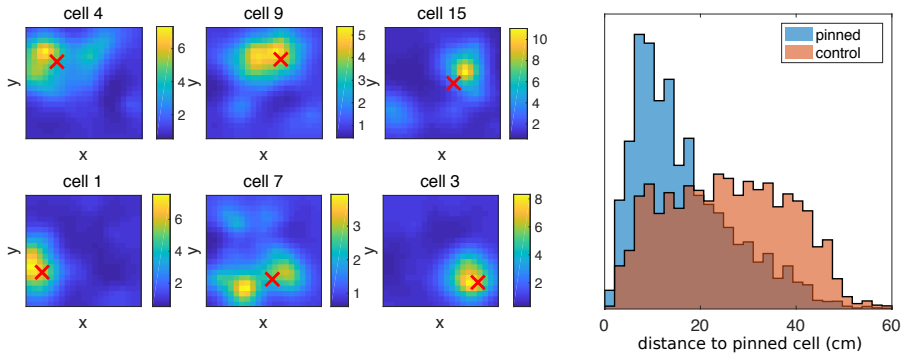


Figure 7.5: **Stimulation of a specific neuron causes neighboring neurons to fire leading to a localized bump of activity. Left:** Spatial distribution of inferred position from simulated neural activity in six different pinning conditions, compared with the inferred position of an un-pinned simulation. Color scale represents the ratio between the two. In all conditions (pinned and unpinned) the selected neuron (red X) was excluded from positional inference. **Right:** distribution of Euclidean distance from the decoded position in pinned (blue) and un-pinned (control, in red) conditions. Displayed data refers to the aggregated analysis of all "A" cells (1 to 16, 32, 33, 34). Each analysis was performed by concatenating five different simulations, each of length $T = 500$ time bins.

7.3 ATTRACTOR-LIKE BEHAVIOR OF THE INFERRED MODEL: TWO MAPS

Simulated activity is bimodally distributed and oscillates between the two maps

In the applications seen so far, an Ising model is inferred for each of the two environments, leading to map-specific couplings J_{ij}^{M}. However, in theoretical attractor models (as well as the real brain), multiple cognitive maps are stored in the same connectivity matrix. A natural question, therefore, concerns the capability of the inferred model of storing multiple attractor states in one single functional-connectivity matrix. To test this idea, we inferred a single Ising model from the concatenation of reference sessions recorded in CA3 from the exploration of two different environments [2, 143]. We then generated activity from the resulting model $P^{A+B}(\mathbf{s})$ and decoded the map by using the two single-map models $P^{A}(\mathbf{s})$ and $P^{B}(\mathbf{s})$ in the same Bayesian framework of [1] (see Chapters 5 and 6). Interestingly, we observed that the represented cognitive map of the simulated network oscillates, in time, between the two attractor states, never falling into mixed or uninformative states, see Fig.7.6. This oscillating phenomenology recalls the flickering behavior reported by Jezek et al. in the original experiment from which the reference sessions were recorded [143], interpreted as a stimuli-driven transition between the two memorized attractor states [2, 157, 206]. The typical persistence of the map in one single environment can be computed from the autocorrelation in time of a map index $m_t = 1$ if $\Delta\mathcal{L}_t > 0$ and 0 if $\Delta\mathcal{L}_t < 0$. The autocorrelation is exponential, with a typical time

$\tau_m \sim 10$ time bins. This value is not far from the one observed in real data during conflict-driven flickering transitions of the cognitive map, which we computed in Chapter 6 finding a value of ~ 6 time bins. However, the persistence of the simulated activity in one single map depends on the simulation time τ_{MC} and can vary from ~ 30 to ~ 5 in the range $\tau \in [1, 50]$. Importantly, the activity is time-by-time coherent with the retrieved map and represents a localized position on the corresponding chart, since the bump dispersion is comparable to the one obtained on real data, see Fig.7.6, bottom-right panel. For this comparison we used the teleportation session [143], where the cognitive map that changes in time and can be tracked by the log-likelihood difference [2]. The "depth" of the activity in one or the other attractor state at time t can be quantified by the delta-likelihood $\Delta\mathcal{L}_t$. If we visualize the histogram of $\Delta\mathcal{L}_t$ separated by the number of active neurons n, we observe a more accentuated polarization of the activity into the two states, with fewer mixed states, as we increase n (see Fig.7.6, bottom panel). This bimodality reflects the fact that low-energy mixed states are impossible when a high number of neurons is simultaneously active, suggesting an effective inhibition between the two attractors. As we will see below, this inhibition is encoded into effective negative couplings between the two sub-populations that encode for the two maps.

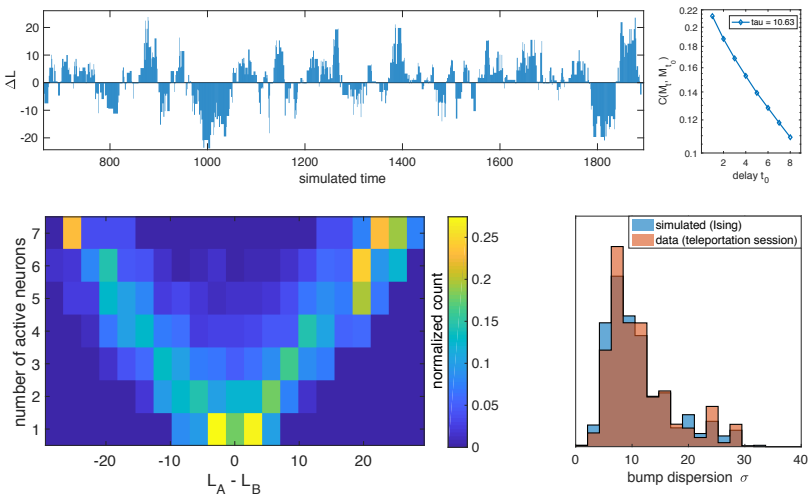


Figure 7.6: **Bimodal behavior of simulated activity in the two-maps-stored case. Top:** time course of the likelihood difference $\Delta\mathcal{L}_t$ between the cognitive maps in a simulated session. The recalled map oscillates between the two attractor states, with a persistence time of ~ 10 time bins. **Bottom:** histograms of $\Delta\mathcal{L}_t$ separated by the number of active neurons in the time bin $\mathbf{s}_t$. Each row is normalized. The spatial coherence of the activity is quantified by the bump dispersion in the chart decoded by $\Delta\mathcal{L}_t$, and is comparable to experimental data (teleportation session of [143]). Simulation $\tau_{MC} = 5$.

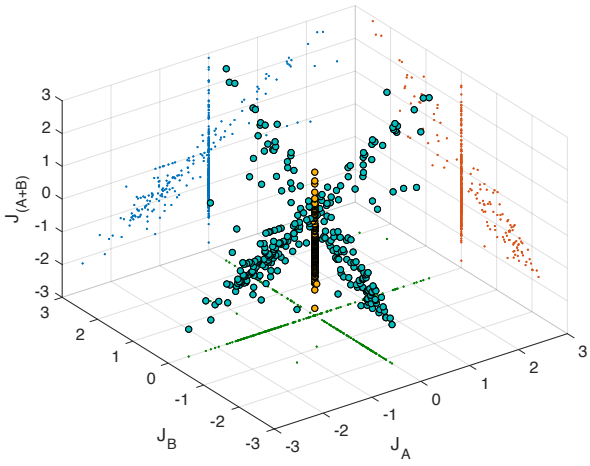


Figure 7.7: **Storing two maps in the connectivity matrix causes effective inhibition between the two sub-networks** Comparison between the network inferred by concatenation of reference sessions and the linear composition of the two networks independently inferred from the two reference sessions. The connectivity inferred from the concatenated session has an effective inhibition between the two sub-populations that is not present in the composed network, noticeable from the large number of points on the $(x, y) = 0$ line (highlighted in orange).

Storing two orthogonal maps in the connectivity matrix causes effective inhibition between the two sub-networks

The explanation of such oscillatory behavior lies in the connectivity structure of the inferred model. In Fig. 7.7 we compare the couplings obtained by the concatenated reference session, denominated as J_{ij}^{A+B} to the sum of the two couplings obtained by single reference sessions, $J_{ij}^{A} + J_{ij}^{B}$, which would follow from the linear-sum assumption of an Hopfield-like model [12, 152, 155]. In the analyzed CA3 data, the neurons display a significant map selectivity in their average firing rate, a phenomenon that is due to a combination of global remapping and pre-processing of the experimental recordings. Therefore, two neurons belonging to the sub-population "A" and "B", respectively, are rarely seen firing together. In the concatenation of the two reference session, this mutual exclusion is interpreted as a negative coupling in the inferred matrix. As the two populations are almost orthogonal, the result is an effective network composed of two subpopulations, each retaining their own structure of excitatory-inhibitory functional connections and connected with each other by inhibitory couplings (see Fig. 7.7).

Part III

INFER GLOBAL, PREDICT LOCAL: BIAS-VARIANCE TRADE-OFF IN PROTEIN FITNESS LANDSCAPE RECONSTRUCTION FROM SEQUENCE DATA

8 BACKGROUND

8.1 DIRECT-COUPLING ANALYSIS (DCA) FROM SEQUENCE DATA

Direct-coupling analysis (DCA) is an application of statistical inference to the modelling of protein sequences. It has recently gained interest in bioinformatics and genomics since it outperformed previous standard methods in predicting the 3D contact structure and the mutational landscape of proteins, both in real and synthetic data [43, 44, 56–58]. It has also been shown to provide meaningful biological predictions in relevant bioinformatics problems, such as protein-protein interaction [223, 224] or genome-wide analysis [225].

The DCA procedure consists in finding the interaction matrix that best reproduces the pairwise co-variation structure computed from a multiple-sequence-alignment (MSA) of a protein family. The resulting connectivity matrix represents the "easier explanation" of the empirical correlation matrix, which in turn can result from direct (i.e., between two sites that are directly connected, in a statistical sense) or indirect (i.e., via a common third node) correlations, see Fig. 8.1 for a pictorial example.

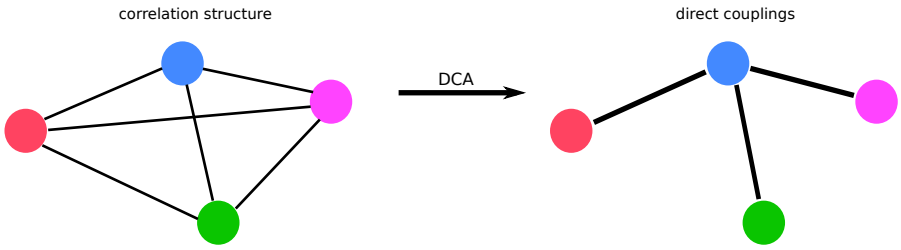


Figure 8.1: **Direct Coupling Analysis:** the correlation structure of the four nodes (colored circles) is fully connected (left panel). However, the correlations between the red, purple, and green node are *indirect*, since they are not caused by direct mutual connection but by an external common source (blue node, right panel). DCA methods are designed to retrieve this direct connectivity structure (right panel).

8.1.1 *The inverse Potts model*

The standard procedure to perform DCA is to solve the inverse Potts problem from sequence data in the MSA of the investigated protein family. The Potts model is an extension of the Ising model presented in Section 2.1 where each spin s_i can take A

different values, called *colors*. The case $A = 2$ retrieves the Ising model. Its Hamiltonian reads

$$\mathcal{H}^{Potts}(\mathbf{s}) = -\sum_i h_i(s_i) - \sum_{i<j} J_{ij}(s_i, s_j) \tag{8.1}$$

Where $i \in [1, N]$ indicates a site and $s_i \in [1, \dots, A]$ is the color on site i. Analogously to the Ising case, the Boltzmann-Gibbs measure of the Potts Hamiltonian is the max-entropy distribution that reproduces a set of observed correlations $p_{ij}(a, b)$ and conservations $p_i(a)$. The first step of DCA from sequence data is to compute the empirical correlations and conservations from the MSA of the relevant protein family:

$$p_i(a) := \frac{1}{B} \sum_{\mathbf{s} \in MSA} \delta(s_i, a) \tag{8.2}$$

$$p_{ij}(a, b) := \frac{1}{B} \sum_{\mathbf{s} \in MSA} \delta(s_i, a)\delta(s_j, b) \tag{8.3}$$

Then, approximate or exact methods (see Section 2.1) are used to retrieve the maximum-likelihood or maximum-a-posteriori set of Potts parameters $\mathbf{J}, \mathbf{h}$ that explain the observed sequence data. The resulting Potts Boltzmann-Gibbs measure

$$P(\mathbf{s}) = \frac{1}{\mathcal{Z}} e^{\sum_i h_i(s_i) + \sum_{i<j} J_{ij}(s_i, s_j)} \tag{8.4}$$

represents the probability for the sequence $\mathbf{s}$ to be part of the protein family, and the inferred Hamiltonian plays the role of "statistical energy". This probabilistic formulation allows for several applications [58, 226, 227]. For example, we can *score* a newly-observed sequence to predict to which protein family it belongs by using Bayesian hypothesis testing (homology detection), or we can *generate* new sequences that are likely to belong to a given family by sampling from the corresponding Boltzmann distribution.

8.1.2 *Contact prediction*

Homologous proteins, descending from a common ancestor, usually conserve the same tridimensional structure along evolution, despite showing high variability in their sequences of amino acids. This variability results from iterated random mutagenesis and natural selection. A random mutation is typically deleterious for the foldability of the protein since it might affect the physical compatibility of a chain site with its close-by residues in the folded conformation. However, this instability can sometimes be resolved by compensating mutations of residues that are in physical proximity (contact) with the mutated site. On an evolutionary scale, this causes a constraint in the variability of the involved sites, since they are forced to co-evolve to maintain the structural stability of the protein (see Fig. 8.2).

A physical contact, therefore, will cause a co-variation between amino acids on the two sites in contact, that can be observed as a correlation $p_{ij}(a, b)$ from the MSA of the

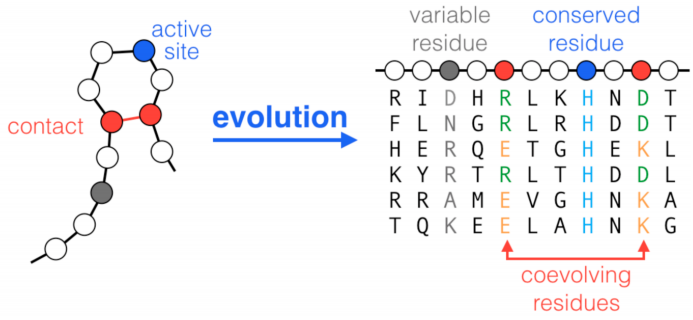


Figure 8.2: **Evolutionary constraints shaping the variability between homologous sequences:** While constraints on individual residues (e.g., active sites) lead to variable levels of aminoacid conservation, the conservation of contacts leads to the coevolution of structurally neighboring residues and therefore to correlations between columns in a multiple-sequence alignment of homologous proteins (here an artificial alignment is shown for illustration). Figure and caption from [227]

protein family. As we saw, the simple observation of a correlation does not imply a direct statistical connection, since it can be caused by a common connection to a third site. For this reason, correlation-based methods strived to take off despite being in the literature for several decades [228, 229].

DCA has originally been introduced to overcome this limitation, i.e., to retrieve the contact structure of a protein family from the correlation structure computed from an MSA of homologous sequences [57]. As shown by several works [57, 58, 230–232], a strong inferred Potts coupling $J_{ij}(a,b)$ (either positive or negative) constitutes a more reliable indication of a contact than the simple correlation $p_{ij}(a,b)$. Couples of sites can, therefore, be scored by computing the Frobenius norm of their coupling matrices $\mathbf{J}_{ij}$:

$$f_{ij} = \sqrt{\sum_{a,b} J_{ij}(a,b)^2} \tag{8.5}$$

As shown in [57, 58], the top-scoring couples are often good predictors of structural contacts.

8.1.3 *Fitness prediction*

Homologous proteins, i.e., belonging to the same family, are often assumed to share a common functionality related to their structure. If we define the fitness of a protein as the degree of performance of this function, an MSA can be considered as a collection of high-fitness proteins, since they have been sequenced from natural (alive, therefore successful) organisms. From a statistical physics point of view, an MSA is, therefore, a

low-temperature sampling of an unknown fitness landscape, whose shape we want to retrieve from the observed sequence data.

The inferred Potts Hamiltonian $\mathcal{H}^{Potts}(\mathbf{s})$ can be interpreted as a low-mode (pairwise) approximation of the unknown fitness landscape. We can therefore use it to predict the phenotypic effect of residue mutations by computing the difference of energy between the mutated and original sequence. If we choose the starting sequence $\mathbf{s}^0$ as the gauge of our Potts parameters (i.e., $J_{ij}(a,b) = 0 \; \forall a = s_i^0, \; \forall b = s_j^0$ and $h_i(a) = 0 \; \forall a = s_i^0$) we can compute the predicted fitness difference due to the single mutation $s_i^0 \to a$ as

$$\Delta\mathcal{H}_{ia}^{Potts} := \mathcal{H}^{Potts}(\mathbf{s}_{i\to a}^0) - \mathcal{H}^{Potts}(\mathbf{s}^0) = -h_i(a) \tag{8.6}$$

Likewise, we predict the epistatic effect of a double mutation $s_i^0 \to a, \; s_j^0 \to b$ as

$$\Delta\Delta\mathcal{H}_{ijab}^{Potts} := \mathcal{H}^{Potts}(\mathbf{s}_{i\to a, j\to b}^0) - \mathcal{H}^{Potts}(\mathbf{s}_{i\to a}^0) - \mathcal{H}^{Potts}(\mathbf{s}_{j\to b}^0) + \mathcal{H}^{Potts}(\mathbf{s}^0) = -J_{ij}(a,b) \tag{8.7}$$

In the $\mathbf{s}^0$ gauge, the DCA-inferred fields $h_i(a)$ represent therefore the statistical-energy differences of single mutations and the coupling $J_{ij}(a,b)$ are interpretable as the epistatic effects of double mutations. This approach has been shown to provide good results on real and synthetic protein datasets [7, 44, 233, 234]. Note that the fitness function is not expected to be a linear function of the statistical energy $\mathcal{H}^{Potts}$ [227], therefore the Spearman coefficient ρ between the experimental and predicted fitness difference (that does not assume a linear relationship between the correlated variables) is usually taken as performance measure of the inference [44, 234] (see for example Fig. 8.3, from [44]).

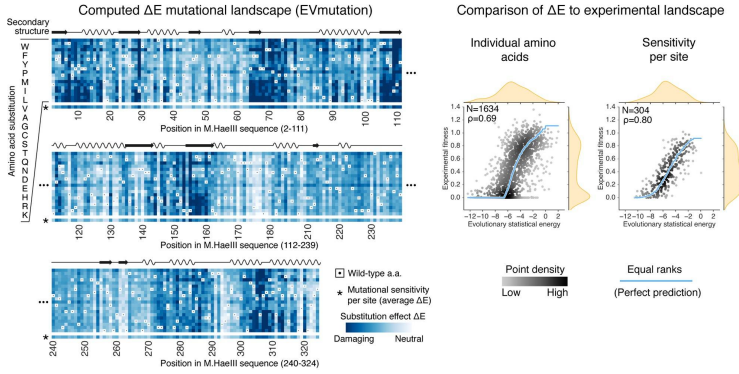


Figure 8.3: The computed $\Delta\mathcal{H}^{Potts}$ mutational landscape of the DNA methyltransferase M.HaeIII (left, color range from 5th percentile to 0) agrees quantitatively with experimental measurements of M.HaeIII fitness under selection by restriction enzyme cleavage (right, $\rho = 0.69$, $N = 1,634$; marginal distributions in orange). The average mutational sensitivity per position shows improved correlation beyond individual effects ($\rho = 0.80$, $N = 304$). Figure and caption adapted from [44]

8.1.4 *Open issues*

Despite their recent success, criticism has been raised against DCA techniques mainly due to their theoretical grounding in maximum-entropy arguments, whose suitability to describe co-evolution of protein sequences is controversial [235, 236]. One of the main arguments points to the fact that the "real" model, i.e., the fitness function whose landscape is explored by evolution, is not a "simple" exponential model of pairwise interaction between protein sites. The real fitness function of a protein results from a multitude of complex biological processes and constraints, clearly involving higher-order terms of interaction as well as degrees of freedom beyond the sequence of amino acids on the protein chain. The discrepancy between the inferred pairwise function and the real one, therefore, will cause a finite error even in the best case of infinite training data [235]. Another arguable point is that, by doing DCA, we explicitly choose to include only correlations and single-point averages in our theory - a choice that uniquely determines our statistical model through constrained maximization of the Shannon entropy [22] - while we typically have access to more information, namely the entire MSA and relevant metadata [236]. Finally, a practical matter is that the inferred pairwise model often suffers from over-parametrization since the number of training data points (sequences in the MSA $\sim 10^{3-5}$) is sensibly lower than the number of inferred parameters ($\sim 10^{5-7}$).

Some of these issues, namely the non-consistence of the inferred model with the real one and over-fitting in the typical under-sampled regime, are partly due to the fact that we are using a single Potts model, that has $O(N^2)$ degrees of freedom, to score the fitness of, in principle, any possible point of the immense sequence space. However, in the task of computing the single mutations landscape of a specific *wildtype* protein, we do not need such a general purpose, since we only have to retrieve the set of $N \times 20$ (N being the number of sites, 20 the number of possible mutations on a site, gap included) single-point mutational effects ΔE_{ia}.

In the next chapters, we will investigate theoretically and practically these issues on a toy model of protein folding, called Lattice Proteins, whose fitness is a high-order (non pairwise) function that represents the probability of a sequence to fold into a given structure. We will discuss how the complexity of the model used to infer the landscape and the training sequence data in the MSA have to be optimally adapted to make accurate single-point predictions, framing both these issues in the context of *bias-variance tradeoff*.

8.2 LATTICE PROTEINS

Lattice proteins are highly-simplified models for protein folding where all the possible folding structures are the non-intersecting walks that fill a squared or cubic lattice. Lattice Proteins display important behaviors that reproduce real-proteins features, such as efficient folding, non-trivial statistics, and the possibility of defining two proteins as homologous if they both are good folders for the same structure. At the same time, being a numerical model, they have the obvious advantage of being treatable by analytical methods and simulations. For these reasons, Lattice Proteins have been widely employed in the literature, for example as a model for protein folding [237–239], protein evolvability [240,241], or to benchmark statistical methods designed to retrieve the folding structure from observations of sequence data within the corresponding family [58].

8.2.1 *The model*

Proteins are defined as sequences of $N = 27$ sites, each of which has one out of 20 possible amino acids (*residues*). Close-by sites in the structure are called contacts. Sites that are in contact interact with an interresidue energy that depends on their specific amino acids (a, b) through the so-called Miyazawa-Jernigan (MJ) matrix $\epsilon(a, b)$. Entries of this matrix have been estimated by analysis of the statistics of residue-residue contacts observed in a large dataset of crystallized proteins [242]. Given a fold F and a sequence of residues $\mathbf{s} = s_1, s_2, \ldots, s_N$, the *structural energy* is defined as the sum of all the MJ contact energies:

$$E(F, \mathbf{s}) = \sum_{(ij)\in C^F} \epsilon(s_i, s_j) \tag{8.8}$$

where C^F is the set of site couples that are in contact in the structure F. The idea behind the model is that a given sequence $\mathbf{s}$ explores the energy landscape, defined over all the folding possibilities, at thermal equilibrium. This allows us to write the probability for a specific sequence $\mathbf{s}$ to fold in a structure F as the corresponding Boltzmann distribution:

$$P_{nat}(F|\mathbf{s}) = \frac{e^{-E(F,\mathbf{s})}}{\sum_{F'} e^{-E(F',\mathbf{s})}} \tag{8.9}$$

$$= \frac{1}{1 + \sum_{F'\neq F} e^{-G_{F',F}(\mathbf{s})}} \tag{8.10}$$

where $G_{F,F'}(\mathbf{s}) := E(F', \mathbf{s}) - E(F, \mathbf{s})$ is energy gap between the two structures F,F'. The (8.10) exposes the central feature of the model: the probability of folding into a given structure F does indeed depend on the structural energy of the protein sequence, $E(F, \mathbf{s})$, but only through its difference with the energy of the other *competing* folds. The probability $P_{nat}(F|\mathbf{s})$ is therefore ruled by the *competition* between folds: the goodness of a sequence as a folder for a structure F depends on its specificity, not on its absolute

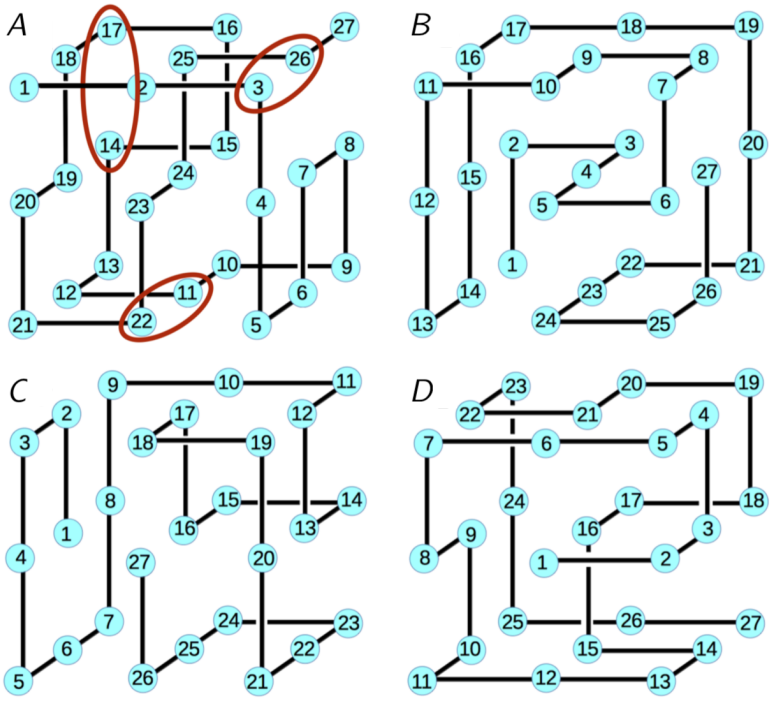


Figure 8.4: **Lattice proteins on the** $3 \times 3 \times 3$ **cubic lattice**: four specific folds, denominated A, B, C, and D, that have been used in the present work. Three contacts are highlighted on structure A.

structural energy. If a sequence $\mathbf{s}$ has a high probability of folding into the structure F we say that it is a *good folder* for that structure (a typical value is $P_{nat}(F|\mathbf{s}) > \theta = 0.99$ [58]).

We will hereafter focus on the case of $3 \times 3 \times 3$ cubic lattice, see Fig. 8.4. Each protein is a sequence of $N = 27$ residues, arranged in one of $N_F = 103,346$ possible cubic folds (estimated via enumeration, up to rotational and chirality symmetries [243]). The number of contacts is the same, $|C| = 28$, for every structure. Note that the chain contacts between two consecutive sites are not taken into account since they induce a constant factor (i.e., not fold-dependent) in the energy which is simplified in the (8.10).

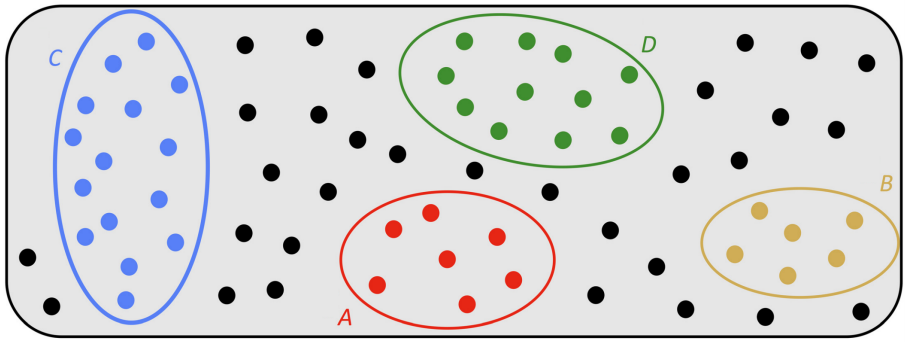


Figure 8.5: **Families as subsets of sequence space in lattice proteins:** Protein families, each corresponding to a particular structure $F = A, B, C, D$, represent portions of sequence space (colored blobs), in which all sequences (colored dots) fold into a unique conformation. Many sequences are expected to be non folding, and not to belong to any family (black dots). Figure and caption adapted from [58]

8.2.2 *Sampling an MSA from a Lattice Protein family*

As we saw, a sequence $\mathbf{s}$ is a good folder for the structure F_0 if it is highly specific, i.e., the structural energy $E(F, \mathbf{s})$ is low *only* for $F = F_0$. In other words, a sequence can be a good folder for (at most) one structure at the time. Therefore, we can partition the space of sequences into non-overlapping sets of good folders for each of the N_F structures (plus the set of those that do not fold at all). An outline of this idea is shown in Fig. 8.5. The set of good folders for the same structure defines a **family** of homologous proteins. The shared functionality within a family is, therefore, folding into the same 3D conformation.

The in-silico nature of Lattice Proteins allows for sampling a collection of good folders for a given structure, i.e., an MSA, under controlled conditions. As explained in [58][1], sequences belonging to the family of structure F can be collected via Metropolis Monte Carlo sampling of an effective Hamiltonian defined as

$$\mathcal{H}^{nat}(\mathbf{s}) := -\beta \log P_{nat}(F|\mathbf{s}) = \beta \log \left(1 + \sum_{F' \neq F} e^{-\Delta E_{F,F'}(\mathbf{s})}\right) \tag{8.11}$$

With a sampling temperature β that controls the average $P_{nat}(F|\mathbf{s})$ of the collected sequences (e.g. $\beta = 1000$ corresponds to a mean P_{nat} of c.a. 0.995). As we will see in the next chapters, we can exploit this controllability to study how the sampling conditions of the training MSA affects the performance of an inferred statistical model that aims to predict the mutational landscape of a reference sequence, called *wildtype*, or $\mathbf{s}^{wt}$. One of the control parameters we will investigate is the mean homology between the wildtype

1 To ensure consistency with the previous chapters we here use a notation that is different from the one in [58], where the sequences is noted as $\mathbf{A}$ and the structure as S.

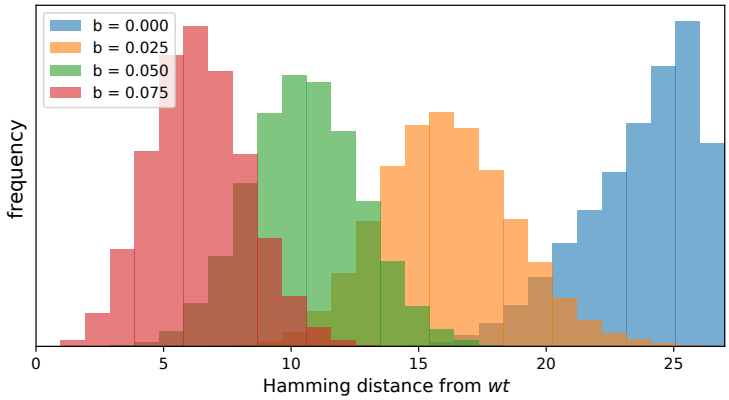


Figure 8.6: **Biased sampling of MSA in Lattice Proteins:** The four distributions of hamming distances from the wildtype are shown for four MSAs sampled with different biases b. Each MSA is sampled by Metropolis Monte Carlo with $\beta = 1000$, $T = 1000$ thermalization time steps between two collected sequences, $B = 10^5$ sequences.

and sequences in the sampled MSA. This parameter can be controlled by adding to the effective Hamiltonian a biasing term in the Hamming distance (i.e., the number of sites that have a different residue) to the wildtype [58], controlled by a parameter b

$$\mathcal{H}_b^{nat}(\mathbf{s}) := -\beta \log P_{nat}(F|\mathbf{s}) + \beta\, b\, d^H(\mathbf{s}, \mathbf{s}^{wt}) \tag{8.12}$$

where $d^H(\mathbf{s}, \mathbf{z}) := \frac{1}{N}\sum_i (1 - \delta(s_i, z_i))$ is the normalized Hamming distance between $\mathbf{s}$ and $\mathbf{z}$. In Fig. 8.6 we see an example of how the parameter b controls the mean Hamming distance of the MSA. The Metropolis acceptance rule for the single mutation $\mathbf{s}_{i\to a}$ (the i-th site of the protein $\mathbf{s}$ is mutated into amino acid a) then becomes

$$P(\text{accept } \mathbf{s}_{i\to a}) = \min\left[1,\ \left(\frac{1 + \sum_{F'\neq F} e^{-\Delta E_{F,F'}(\mathbf{s})}}{1 + \sum_{F'\neq F} e^{-\Delta E_{F,F'}(\mathbf{s}_{i\to a})}}\right)^{\beta} \cdot e^{\frac{\beta}{N} b(\delta_{a,s_i^{wt}} - \delta_{s_i,s_i^{wt}})}\right] \tag{8.13}$$

To ensure the independence of samples in the collection we add one sequence every T Metropolis Monte Carlo steps to the MSA. In the present work, as done in [58], we typically set $T = 1000$ and consider, without loss of generality, a fixed subset of $N_F' = 10,000$ structures.

8.3 OUTLINE OF THE FOLLOWING CHAPTERS

In the next chapters we will use Lattice Proteins to benchmark Potts-inference methods aimed at retrieving the single-point mutational landscape of a protein. The fitness function in Eq. 8.10 is not a pairwise model since it contains terms of higher order that model the competition between structures in the folding probability. Therefore, Lattice Proteins reproduce an essential issue discussed in Section 8.1, namely the non-consistence of the inferred (pairwise) and real model, making them a good compromise between controllability and complexity.

In chapter 9, we will study how the single-mutation landscape of Lattice Proteins depends on the fitness of the mutated sequence, and how this compares with the landscape predicted by the inferred Potts model.

In chapter 10 we will show that a sparse model that includes, as a prior, structural information outperforms the current standards in the common under-sampled regime, introducing an adaptive method that yields optimal performances in both the under-sampled and well-sampled cases.

Finally, in chapter 11, we will study how the predictive power of the inferred model depends on the training MSA through few simple descriptors: the number of sequences, that controls the variance of the Potts predictions, and the average Hamming distance to the mutated protein, which, as we will see, is related to the bias of the inferred Potts model. We will introduce a prescription, called *focusing procedure*, to choose the optimal training MSA for the specific task of inferring the single-point mutations fitness effects around one given sequence.

9 CHARACTERIZATION OF THE MUTATIONAL LANDSCAPE OF LATTICE PROTEINS

This chapter is composed of two parts. In the first part, we will show how the shape of the single-point mutational landscape of a Lattice Proteins, which is related to its evolvability, depends on its fitness, i.e., its folding probability P_{nat}. As we will see, a higher fitness corresponds to a smaller variance of the mutational landscape, an effect that is derivable from the mathematical definition of the fitness

$$\mathcal{H}^{nat}(\mathbf{s}) := -\beta \log P_{nat}(\mathbf{s}) \quad . \tag{9.1}$$

In the second part, we will investigate how the single-mutation landscape predicted by a Potts model, inferred from an MSA of homologous sequences, relates to the real one. We will see that in the range of non-deleterious mutations we can derive a linear relation with a slope that depends on the P_{nat} of the mutated sequence in the native structure and on the energy of competing structures, using the derivation of the pressure λ_{ij} obtained by Jacquin et al. [58].

9.1 DISPERSION OF THE MUTATIONAL LANDSCAPE DEPENDS ON THE FITNESS

An essential question in protein design concerns the evolvability of a sequence, i.e., its capacity to adapt to changing environmental pressures. During the evolutionary process, a protein explores its local fitness landscape by mutating one or more amino acids on the chain. The evolutionary paths, and consequently the evolvability of a protein, therefore crucially depends on the shape of the local mutational landscape.

To investigate the local landscape of Lattice Proteins, we computed the set of all possible single-mutation fitness effects $\left\{\Delta\mathcal{H}^{nat}_{ia}\right\}_{i,a}$ defined as

$$\Delta\mathcal{H}^{nat}_{ia} := -\beta \log P_{nat}(\mathbf{s}_{i\to a}) + \beta \log P_{nat}(\mathbf{s}) \quad , \tag{9.2}$$

where $\mathbf{s}$ is the starting sequence, the notation $\mathbf{s}_{i\to a}$ stands for the mutated $\mathbf{s}$ where the amino acid a is placed on the site i, and β is the MSA sampling temperature (see Eq. 8.11).

In Fig. 9.1, we show an example of the distribution of single-mutational effects $\Delta\mathcal{H}^{nat}$ on $n = 36$ Lattice Proteins with P_{nat} ranging from 0.985 to 0.999. Higher-fitness proteins are, as common sense suggests, surrounded by sequences with a lower fitness (almost all $\Delta\mathcal{H} > 0$ in panels belonging to the bottom row). However, the distribution of single-

mutational effects gets more squeezed as the fitness increases, an effect that we can quantify by computing the variance

$$\sigma^2(\Delta\mathcal{H}_{ia}^{nat}) = \frac{1}{27\cdot 19}\sum_{i,a}\left[\Delta\mathcal{H}_{ia}^{nat}\right]^2 - \left[\frac{1}{27\cdot 19}\sum_{i,a}\Delta\mathcal{H}_{ia}^{nat}\right]^2 \tag{9.3}$$

A systematic analysis of the variance of the single-mutational landscape for $n = 1000$ proteins is shown in Fig. 9.2. To avoid the dominating effect of very deleterious mutations (up to $\Delta\mathcal{H}^{nat} \sim 10^3$) the variance is computed within a threshold $\mathcal{H}^{nat} < 15$. As reported in the figure, the scaling of the standard deviation σ is linear in the P_{nat} of the mutated sequence, i.e.

$$\sigma\left[\Delta\mathcal{H}_{ia}^{nat}(\mathbf{s})\right] \propto \beta\left[1 - P_{nat}(\mathbf{s})\right] \tag{9.4}$$

This linear scaling can be explained, as shown below, by analytical development of the definition of the fitness of Lattice Proteins.

9.2 DERIVATION OF $\sigma \sim (1 - P_{nat})$ SCALING

Let's consider a structure F_0, defining a Lattice Protein family. The mutational landscape of the protein $\mathbf{s}$ is defined as the collection of single-point mutation effects on its effective Hamiltonian

$$\Delta\mathcal{H}_{ia}^{nat} := \mathcal{H}^{nat}(\mathbf{s}_{i\to a}) - \mathcal{H}^{nat}(\mathbf{s}) = \beta\log\left(\frac{1+\sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s}_{i\to a})}}{1+\sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}}\right) \tag{9.5}$$

Where $\mathbf{s}_{i\to a}$ is the sequence $\mathbf{s}$ with the residue a placed on site i. In order to investigate how the structure of the mutational landscape depends on the P_{nat} of the starting sequence, we will develop the (9.5) to isolate the role of P_{nat}. By plugging the expression of P_{nat} (8.9) in the (9.5), we can write

$$P_{nat}(\mathbf{s}_{i\to a}) = \left(1 + \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s}_{i\to a})}\right)^{-1} \tag{9.6}$$

$$= \left(1 + \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s}) - \Delta G_{F,F_0}(\mathbf{s}_{i\to a},\mathbf{s})}\right)^{-1} \tag{9.7}$$

$$\simeq \left(1 + \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}\left[1 - \Delta G_{F,F_0}(\mathbf{s}_{i\to a},\mathbf{s})\right]\right)^{-1} \tag{9.8}$$

$$= \left(\left[1 + \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}\right] - \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}\,\Delta G_{F,F_0}(\mathbf{s}_{i\to a},\mathbf{s})\right)^{-1} \tag{9.9}$$

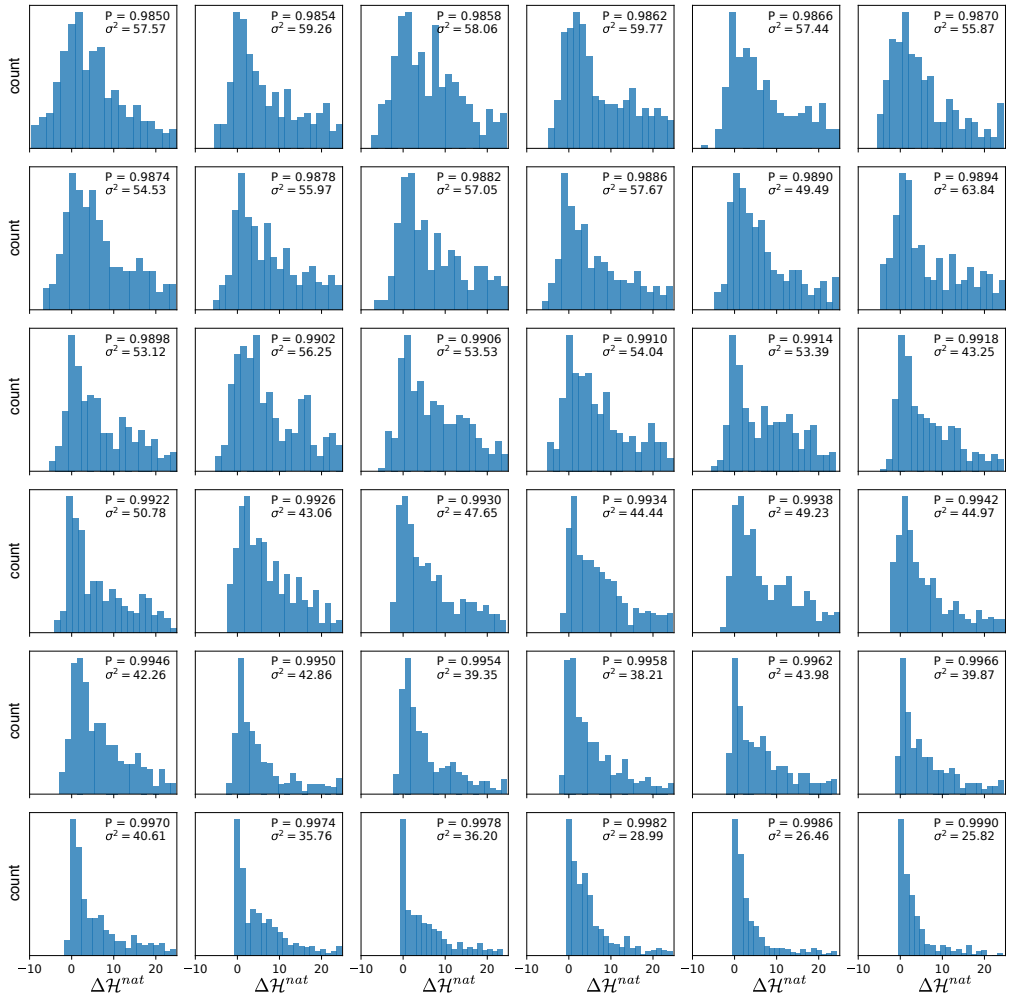


Figure 9.1: **Characterization of mutational landscape, dependence on P_{nat}:** Histogram of single-mutation landscape $\Delta\mathcal{H}^{nat}_{ia}(\mathbf{s})$ for $N = 36$ different sequences $\mathbf{s}$ of increasing P_{nat}. To avoid the dominating effect of mutations that cause very strong fitness drops (up to $\sim 10^3$) the variance is computed within the shown x range. The family-defining fold is the structure A of [58].

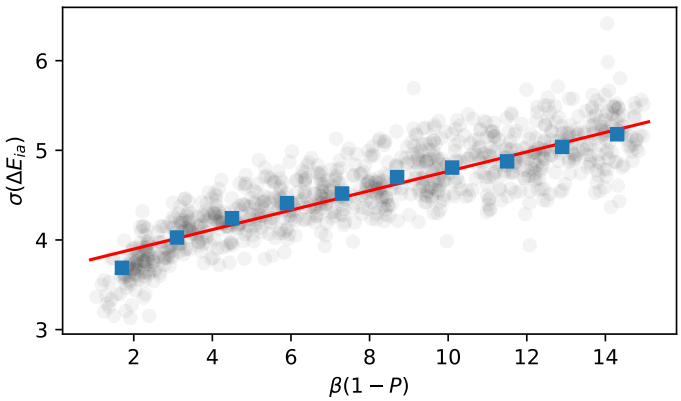


Figure 9.2: **Characterization of mutational landscape, dependence on P_{nat}:** Scatter of the standard deviation of the mutational landscape $\{\Delta\mathcal{H}_{ia}^{nat}\}_{i,a}$ for $n = 1000$ different Lattice Proteins with P_{nat} ranging from 0.985 to 0.999. To avoid the dominating effects of deleterious mutations and highlight the qualitative scaling the standard deviation is computed on $\mathcal{H}_{ia}^{nat} < 15$. In blue: average of y-values in 10 equally-spaced x intervals. In red is the best linear fit of the blue points. The family-defining fold is the structure A of [58].

where we assumed that the gap difference $\Delta G_{F,F_0}(\mathbf{s}_{i\to a}, \mathbf{s})$ is small, i.e., the mutated sequence is still a good folder for structure F_0. Note that we can write the gap difference in terms of the Miyazawa-Jernigan structural energy $E(\mathbf{s}, F)$:

$$\begin{aligned}\Delta G_{F,F_0}(\mathbf{s}_{i\to a}, \mathbf{s}) &:= G_{F,F_0}(\mathbf{s}_{i\to a}) - G_{F,F_0}(\mathbf{s}) \\ &= E(\mathbf{s}_{i\to a}, F) - E(\mathbf{s}_{i\to a}, F_0) - E(\mathbf{s}, F) + E(\mathbf{s}, F_0) \\ &= -\Delta E_{ia}^{F_0} + \Delta E_{ia}^{F}\end{aligned} \tag{9.10}$$

Where ΔE_{ia}^{F} is the structural-energy difference due to the single point mutation $s_i \to a$ in the fold F. By plugging this into the (9.9) we obtain

$$P_{nat}(\mathbf{s}_{i\to a}) \simeq \left(\left[1 + \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}\right] + \Delta E_{ia}^{F_0} \cdot \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})} + \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}\, \Delta E_{ia}^{F}\right)^{-1} \tag{9.11}$$

$$\simeq \left(\left[1 + \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}\right] + \Delta E_{ia}^{F_0} \cdot \sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}\right)^{-1} \tag{9.12}$$

where we assumed that the average over all the folds F of the structural energy difference ΔE_{ia}^{F} is zero. By using this into the (9.5) we obtain

$$\Delta\mathcal{H}_{ia}^{nat} \simeq \beta \log\left(\frac{\left[1+\sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}\right] + \Delta E_{ia}^{F_0}\cdot\sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}}{1+\sum_{F\neq F_0} e^{-G_{F,F_0}(\mathbf{s})}}\right) \tag{9.13}$$

$$= \beta \log\left(1+\Delta E_{ia}^{F_0}\cdot[1-P_{nat}(\mathbf{s})]\right) \tag{9.14}$$

Since we assumed $\mathbf{s}$ to be a good folder for the structure F_0, the term $[1-P_{nat}(\mathbf{s})]$ is small. We can therefore expand the logarithm to obtain the final form of our derivation:

$$\Delta\mathcal{H}_{ia}^{nat}(\mathbf{s}) \simeq \beta(1-P_{nat}(\mathbf{s}))\cdot\Delta E_{ia} \tag{9.15}$$

Where for simplicity we dropped the notation F^0. From Eq. 9.15 we retrieve the linear scaling of the standard deviation observed in Fig. 9.2 by assuming that ΔE_{ia}, the MJ energy difference due to a mutation, is a random variable (on the indexes i, a) equally distributed across different sequences. This assumption is, of course, a gross approximation, but qualitative explains the linear scaling observed in the analysis of Fig. 9.2.

9.3 RELATIONSHIP BETWEEN POTTS AND REAL LANDSCAPES DEPENDS ON THE FITNESS

As we saw in the previous chapter, several works in the literature have focused on retrieving the mutational landscape of a protein by inferring a Potts model from a collection of homologous sequences [44, 58, 234]. In the next chapters, we will use Lattice Proteins as a benchmark to understand how the performance of inference models depends on different sampling conditions. We will here make a preliminary characterization of the relationship between the inferred parameters and the real mutational landscape, integrating some results obtained by Jacquin et al. in [58].

An example of the comparison between inferred and real single-mutations landscape is shown in Fig. 9.3, where the x-scale has been set as linear for $x < 15$ and log for $x > 15$ to highlight the fact that some mutations are extremely deleterious for the real fitness ($P_{nat} < 0.3 \implies \Delta\mathcal{H}^{nat} \sim 10^3$), while the Potts prediction saturates at $\Delta\mathcal{H}^{Potts} \sim 15$.

This saturation is explainable from the fact that the model is inferred from a very-low-temperature sample of the probability of folding P_{nat} (see Eq. 8.11). The value we used for the sample, $\beta = 1000$, results in a MSA of homologous proteins with typical value of $P_{nat} \sim 0.995$ (see [58, 241] for a discussion on the role of β). The Potts model is a low-mode (pairwise) approximation of the rough and complex real fitness landscape of Lattice Proteins. The pairwise order of its interaction limits its resolution in the sequence space. In the process of solving the inverse Potts problem, we find the model that better adapts this smooth approximation to the typical sequence data in the MSA. It is therefore natural that a Potts model inferred from a collection of good-folders (typical $P_{nat} \sim 0.995$) is unable to quantitatively characterize very deleterious mutations ($P_{nat} < 0.5$). The value of these mutations is therefore mainly set by the regularization and other priors [244].

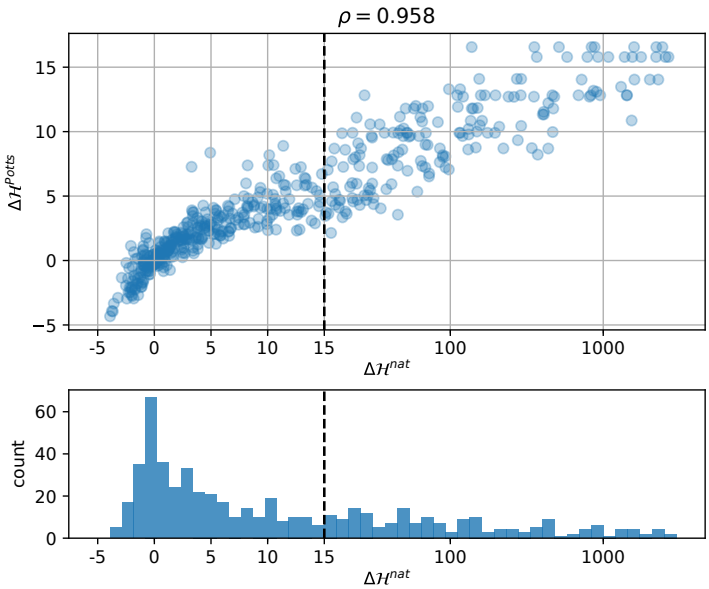


Figure 9.3: **Inference of the single-mutation landscape of Lattice Proteins** comparison between real and inferred single-mutation landscape in a case of good sample (number of training sequences $B = 10^5$). The protein family is the one corresponding to structure C in [58].

The non-linearity of the relationship between inferred and true fitness, as is the case for mutagenesis experiments on real proteins [44, 234], makes the standard Pearson correlation coefficient unfit to describe the performance of the predictions. As done in recent related works [44, 244], we will, therefore, use the Spearman coefficient ρ, i.e., the Pearson coefficient of the rank, as a measure for the performance of the inferred Potts model. As shown in Fig. 9.3, the Spearman coefficient is very high despite the strong non-linearity of deleterious mutations. In this and the next chapters, we will time by time restrict our analysis to a subset of mutations, usually the ones that keep the protein foldable, in order to highlight quantitative or qualitative features of the inferred mutational landscape.

From Fig. 9.3 we can see that the bulk of mutations that are in the range $\Delta\mathcal{H}^{nat} < 2 - 5$ are more linearly predicted by the Potts model, although with a slope $\neq 1$. By analyzing the slope of the bulk of non-deleterious mutations on several different sequences we see that the slope depends on the P_{nat} of the mutated protein, see Fig. 9.4. In particular, we observe that as P_{nat} increases, the slope of the bulk gets smaller and smaller. To explain this behavior, we combined the derivation of Eq. 9.15 with the equation for the pressure obtained by Jacquin et al. in [58].

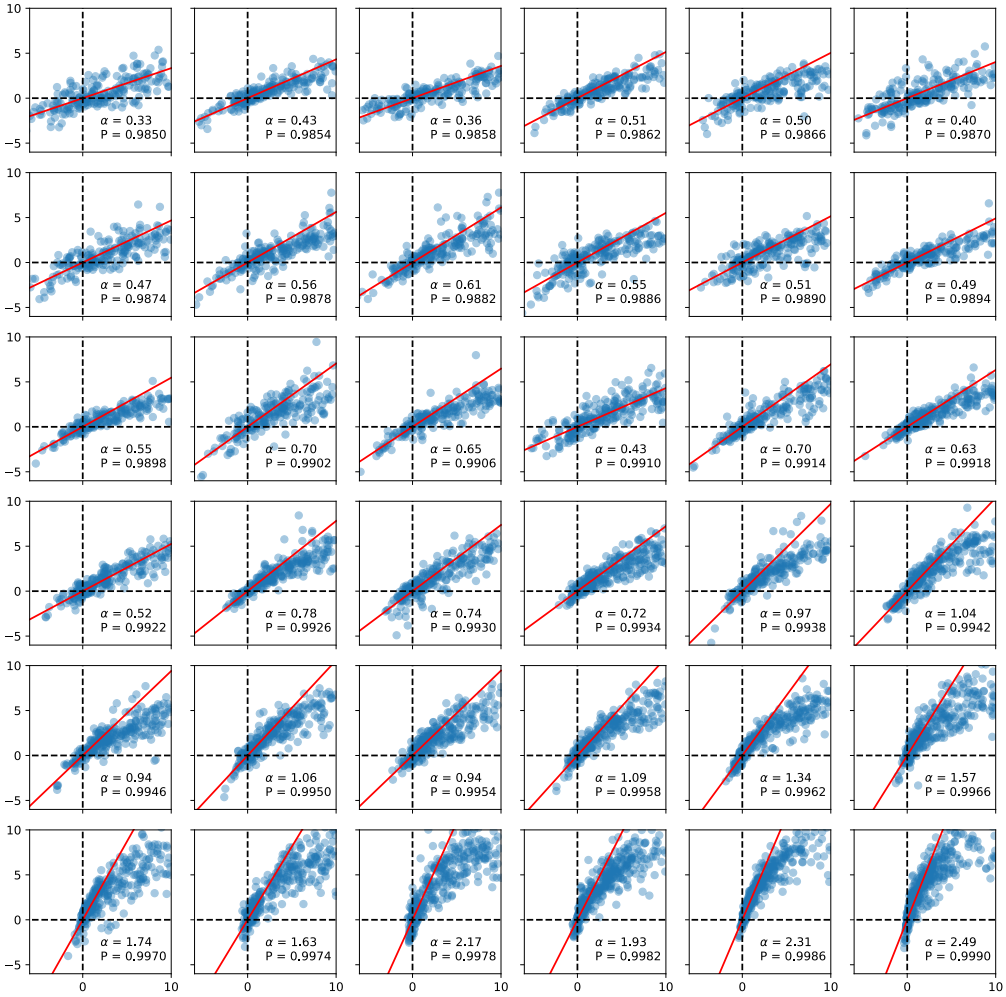


Figure 9.4: **Characterization of mutational landscape, dependence on P_{nat}:** Scatter plots of Potts (y axis) vs real (x axis) mutational landscape of $N = 36$ sequences. For coherence with the theoretical approximations, the analysis is performed on the bulk of non-deleterious mutations, i.e., $\mathcal{H}_{ia}^{nat} < \theta = 2$ (see also Fig. 9.4), and for good folders, i.e., $P_nat > 0.995$.

In [58] (Main text Eq. 1), Jacquin and collaborators derived an approximated equation that relates the inferred Potts couplings $J_{ij}(a,b)$ with the Miyazawa-Jerningan energy $\epsilon(a,b)$:

$$J_{ij}(a,b) \simeq -\lambda_{ij}\,\epsilon(a,b)\ , \tag{9.16}$$

where λ_{ij} is a pressure term that depends on the fold structure, on the sampling conditions, and on if the pairs (ij) are in contact or not in the competing structures. The Potts inferred fitness difference of a single mutation for a sequence $\mathbf{s}$ can be written as

$$\Delta\mathcal{H}_{ia}^{Potts}(\mathbf{s}) = -\left(h_i(a) - h_i(s_i)\right) - \sum_{j\neq i}\Big(J_{ij}(a,s_j) - J_{ij}(s_i,s_j)\Big) \tag{9.17}$$

By plugging the Eq. 9.16 into this definition, and ignoring the sub-leading field term we find

$$\begin{aligned}\Delta\mathcal{H}_{ia}^{Potts}(\mathbf{s}) &\sim -\sum_{j\neq i}\Big(J_{ij}(a,s_j) - J_{ij}(s_i,s_j)\Big) &(9.18)\\ &\simeq \sum_{i\neq j}\lambda_{ij}\Big(\epsilon(a,s_j) - \epsilon(s_i,s_j)\Big) &(9.19)\\ &\sim \bar{\lambda}\ \Delta E_{ia}(\mathbf{s}) &(9.20)\end{aligned}$$

Where we made the approximation $\lambda_{ij} \sim \bar{\lambda} =$ the average pressure computed on all the couples ij. By combining Eq. 9.20) with Eq. 9.15 we find the equation that relates the inferred mutational landscape with the real one:

$$\Delta\mathcal{H}_{ia}^{Potts}(\mathbf{s}) \ \simeq\ \frac{\bar{\lambda}}{\beta(1-P_{nat}(\mathbf{s}))}\,\Delta\mathcal{H}_{ia}^{P_{nat}}(\mathbf{s}) \tag{9.21}$$

To investigate the validity of this derivation, we inferred the fitness landscape of $n = 1000$ different sequences by retrieving a Potts model with PLM [44, 56], from an MSA of $B = 10^5$ homologous sequences. Results, reported in Fig. 9.5 show that, in a range of reasonable validity of the approximations ($P_{nat} > 0.995$, $\Delta H_{ia} < 2$), we indeed observe a linear relation between the slope of the bulk of non-deleterious mutations and the P_{nat} of the mutated sequence.

Moreover, we can estimate the parameter $\bar{\lambda}$ by best fit of the linear relationship in (9.21) on the $N = 1000$ points. Surprisingly, the retrieved values of $\bar{\lambda}$ are very similar to the ones reported in [58] (structure A: 1.94 compared to 2.00; structure C: 1.32 compared to 1.32). However, note that our estimation of $\bar{\lambda}$ changes with the threshold θ used in the definition of the bulk, and can vary up to ± 0.5 depending on the parameters of the analysis. Therefore, this result should be taken as a qualitative agreement and not as a precise quantitative estimation.

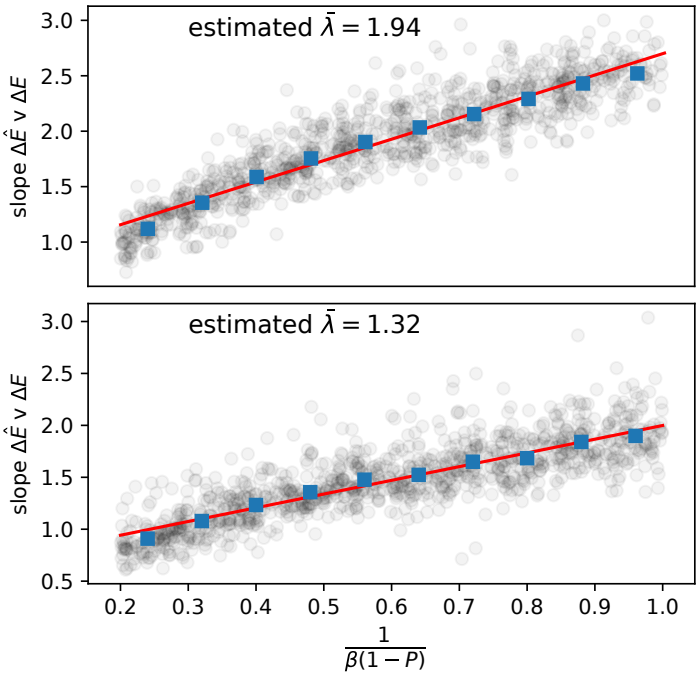


Figure 9.5: **Characterization of mutational landscape, dependence on P_{nat}:** Top: structure A; bottom: structure C. Scatter of the slope of Potts-vs-real mutational landscape against $\frac{1}{\beta(1-P_{nat}(\mathbf{s}))}$ for $N = 1000$ different sequences $\mathbf{s}$. To ensure coherence with the theoretical approximations the analysis is performed on non-deleterious mutations, i.e., $\Delta\mathcal{H}_{ia}^{nat} < 2$, and for good folders of the native structure, i.e., $P_nat > 0.995$. In blue: average of y-values in 10 equally-spaced x intervals. In red is the best linear fit of the blue squares.

10

SPARSE POTTS INFERENCE WITH STRUCTURAL PRIOR (SP-ACE) IMPROVES FITNESS PREDICTIONS IN UNDER-SAMPLED REGIME

In the previous chapter we studied how the Potts-inferred mutational landscape compares with the true one, defined through the log-P_{nat} difference between the mutated and the original sequence. The current standards to retrieve the Potts model from MSA data are pseudo-likelihood maximization (PLM) [44, 234] and adaptive cluster expansion (ACE) [227, 244]. The aim of this chapter is to show that to infer a fully-connected Potts model is rarely the optimal choice for typical applications where, contrarily to the case of synthetic data, the number of training data (the MSA) is limited by the available sequenced proteins for the relevant family.

10.1 INTRODUCTION

Balancing the complexity of the inferred model according to the amount of available data is a task of primary importance in statistical inference. As known from the statistics literature, if a model is over-parametrized it will yield predictions with a high *variance*, while an under-parametrization will cause *biased* predictions due to un-modeled degrees of freedom. In the context of modelling proteins from sequence data with the inverse Potts model, this problem is today of central importance since the number of inferred parameters systematically outnumbers the available sequences in the training MSA.

The number of parameters of a fully-connected Potts model is $P = 20^2 \cdot \frac{N(N+1)}{2} \simeq 200\, N^2$, N being the number of amino acids of the analyzed protein. To give an order of magnitude, the MSA used by Figliuzzi et al. [234] for predicting the mutational landscape of TEM-1 has $B \simeq 3000$ sequences and $N = 197$ sites, corresponding to $P \simeq 7 \cdot 10^6$ parameters. The number of available sequences in the 34 MSAs used by Hopf et al. [44] to predict as many mutagenesis experiments range from $B \sim 10^2$ to $B \sim 10^5$, while the number of parameters inferred for the corresponding fully-connected Potts models ($50 - 500$ sites) range from $P \sim 10^5$ to $P \sim 10^7$.

In the same work, Hopf and collaborators tested the predictive power of the coupled Potts model (called "epistatic" model) and of the independent Potts model (i.e., only average values are matched with $h_i(a)$). They reported a superior performance of the coupled Potts model on c.a. $2/3$ of the families. The remaining part showed an equal or better performance for the independent model. These results suggest that by varying the complexity of the model (the number of inferred parameters P) by interpolating between the independent and the fully-connected model, one could find an optimal point P^*, corresponding to the best bias-variance tradeoff of the statistical inference.

The classic way to reduce the number of parameters is to enforce the sparsity of the model by using a ℓ-1 regularization [32], which was shown to yield the correct topology

of a sparse interaction graph in the under-sampled regime. Practical applications usually employ an ℓ-2 norm, that corresponds to a Gaussian prior on the value of the inferred parameters [31, 44, 56, 234]. However, there is no global consensus on how to choose the optimal prior strength for a given number of training data points, since it is hard to precisely relate it to the effective number of inferred parameters.

Here, we test a sparse version of Potts inference, introduced in [47], where the interaction graph is enforced basing on the $3D$ contact structure of the analyzed Lattice. We show that this sparse method sensibly improves the performance of single-point mutations predictions, with respect to current standards, in the common under-sampled regime.

We then benchmark an unsupervised approach based on a sparse version of the cross-entropy minimization which only uses some prior knowledge on the degree of sparsity of the original interaction graph [244]. We show that the optimal performances are obtained when adapting the sparsity of the inferred graph to the number of available data points, i.e., the number of sequences in the training MSA, providing a rule of thumb to obtain optimal predictions without the need of fine-tuning prior parameters.

This work has been carried out in parallel on the analysis of real protein data done by F. Rizzato. Part of her results on mutagenesis experiments is presented here to highlight the generality of these approaches. The joint work will be included in a paper now at the draft stage [8].

10.2 RESULTS

cmap-ACE inference improves fitness predictions in the undersampled regime

In the case of Lattice Proteins, we know the real contact map of the $3D$ fold that defines the protein family. It is known that the strongest inferred Potts couplings, defined with the Frobenius norm of the interaction matrix $J_{ij}(a, b)$, are good predictors of the structural contacts in the fold [58]. Here we test the inverse idea, i.e., that the most informative functional couplings in the inferred Potts model are the ones between sites that are in contact. If this is true, we could exploit the structural information to infer only the most relevant parameters when only a small number of training points are available.

For the inference, we used a two-site cluster approximation limited to the sites that are in contact in the tridimensional structure of the protein [31]. This expansion is exact in the case of tree-like connectivity. Therefore, it is expected to be a good approximation in the case of very sparse connectivity, which is the case of Lattice Proteins (28 contacts, 27 sites). For simplicity, we have here re-written a single-purpose version of the cluster expansion code, see Methods and the upcoming publication for reference [8].

We first compared cmap-ACE against standard models in the literature, i.e., DCA via pseudo-likelihood minimization (PLM) [44, 56], ACE, and the independent Potts model (IND). As shown in Fig. 10.1, the structural Potts model sensibly outperforms the other models in the under-sampled regime ($B \leq 1000 \cdot N$).

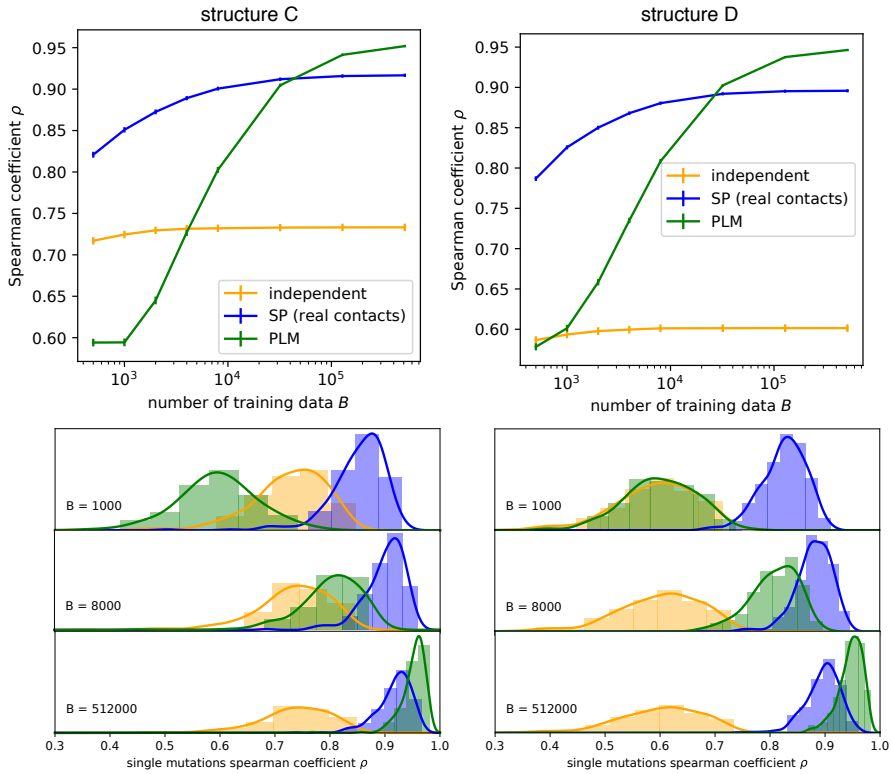


Figure 10.1: **Structural Potts compared to PLM and independent Potts model. Top:**performances, measured by the Spearman coefficient of the inferred and real mutational landscape, as a function of the number of sequences in the training MSA for the cmap-ACE (blue), PLM (green) and independent Potts model (orange). For each B the mean and standard error of the inference performance are computed on 256 different wildtypes. **Bottom** histograms of inference performances, computed on the 256 wildtypes, for the three compared models at three different values of B. Analysis performed on structure C and D in [58], on mutations with $\Delta\mathcal{H}^{nat} < \theta = 25$.

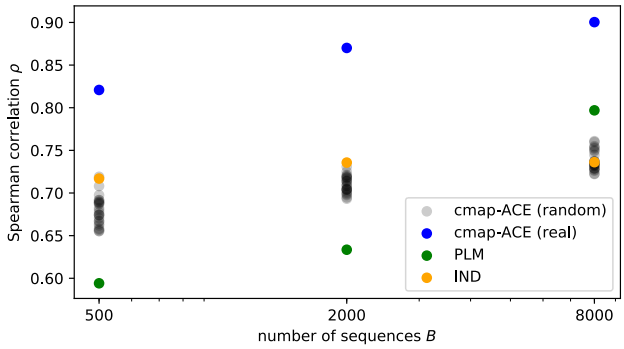


Figure 10.2: **Structural Potts: comparison between real and random contacts** 25 draws of 28 random contacts. Each point represents the average over $n = 64$ sequences drawn from the MSA of structure C.

We then tested how this improvement depends on the specific choice of the 28 real contacts, by performing the SP inference on $n = 25$ random sets of $K = 28$ different site couples (instead of the real contacts), for three different B in the under-sampled regime. Results, shown in Fig. 10.2, confirm that the parameters inferred on the connectivity graph defined by structural contacts are especially relevant for fitness predictions.

Adapting the complexity of the inferred model to the number of data: SP-ACE

The results above depend on the fact that we already know the contact structure of the protein. In practical applications, however, the 3D structure of the wildtype is rarely given. One possible procedure (that we are applying to real proteins [8]) is to infer the contact structure by classic DCA [56] and use the retrieved contact map as imposed topology for the sparse Potts inference. However, it is still unclear how to determine the precise number of structural contacts to infer from sequence data [227].

Here we show that, in the context of fitness inference from sequence data, the number of bounds in the graph of interaction should be adapted to the number of available data points. For the case of Lattice Proteins, it has been shown in [58] that the largest inferred couplings reflect the interactions on the site in contact on the native fold. However, additional non-zero inferred couplings might correspond to the constraint of not-folding in competing structures (negative design). Therefore, in the case of large available data, we should retrieve these additional couplings to refine our fitness estimations.

To test how to adapt the sparsity of the inferred graph to the number of data, we introduce a sparse model based on the adaptive cluster expansion approximation of the cross-entropy [45, 47, 244]. The sparsity of the inferred interaction graph is obtained by truncating the expansion when the total number of non-zero interactions $\mathbf{J}_{ij}$ reaches a

pre-determined number of contacts K, that is adjusted on the number (B) of available sequences in the MSA. Such sparse inference has been shown to be optimal when reconstructing Erdős-Rény models from configuration sampling data [244]. If the cross-entropy contribution is a good criterion for quantifying the relevance of inferred parameters, the resulting interactions are the most relevant for the functionality, either structural or biochemical, of the analyzed protein family. This method is referred to as structural-prior adaptive cluster expansion (SP-ACE).

We tested the performance of the SP-ACE model on Lattice Proteins from the structures C and D. As a rule of thumb, the routine is stopped when the number of 2-clusters reaches a threshold $K(B)$ such that the number of inferred parameters P equals the number of available sequences B, i.e.

$$K(B) = \frac{B - N \cdot \bar{Q}}{\bar{Q}^2} \quad , \tag{10.1}$$

where $\bar{Q}$ is the average number of colors and $\bar{Q}^2$ is the mean size of the J_{ij} matrix after compression of unseen colors, see also Methods for details. In Fig. 10.3 we compare the SP-ACE model for different B and different numbers of inferred contacts K. As expected in the case of the Lattice Proteins, where the 28 structural contacts strongly dominate the fitness function, the best performances are obtained for either $K = 28$ (when $K(B) < 28$, upper panels) or for $K(B)$ as expressed in Eq. 10.1, when greater than 28 (lower panels). This result confirms the hypothesis that the ACE routine automatically selects the most relevant couples of sites during the SP-ACE inference.

In Fig. 10.4 we show how the inference performance varies with B for the SP-ACE model with $K = K(B)$ (black line) and $K = max[N, K(B)]$ (SP-ACE-N, red line). As we expected from the results of the analysis shown in Fig. 10.3, the SP-ACE model is outperformed by the SP-ACE-N and cmap-ACE with the real contacts in the under-sampled regime ($B < 10^4$), still showing superior performances than both PLM and the independent models. When B is large ($B > 10^5$), however, SP-ACE and SP-ACE-N become equivalent to PLM in yielding the best performance. Therefore, SP-ACE-N is equivalent or superior to all models, including cmap-ACE with real contacts, in both the under-sampled and well-sampled regimes. Importantly, SP-ACE-N yields performances comparable to cmap-ACE without the need of knowing a-priori the contact map. Interestingly, the full ACE outperforms PLM in the sub-sampled regime, thanks to the cluster-selection routine that automatically retrieves a sparser model. The convergence of the SP-ACE inference, however, is much faster than the one of the full ACE inference (minutes-hours against hours-days, see [244] for more detailed benchmarks).

Adapting the complexity of the model to the number of data: real proteins

We will here present some early results of the analysis performed by Francesca Rizzato on four data sets that were analyzed in [234] and [44]. This part is a work in progress and will be included in a future publication [8].

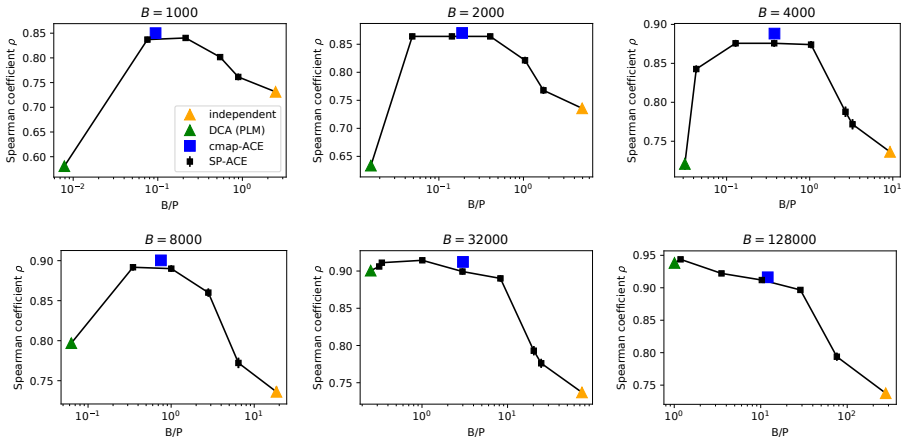


Figure 10.3: **Role of the number of bounds K in SP-ACE inference.** Black lines show the mean performance (over 256 wildtypes) of the SP-ACE model performed with different sparsities, i.e., a different input value of K. The SP-ACE is compared to the fully connected PLM (green triangle), the independent model (orange triangle) and the cmap-ACE model (blue square). Each model is positioned on the x-axis according to the number of inferred parameters related to the number of training data B. Structure C, 256 wildtypes, $\theta = 25$.

The model used for the analysis on real protein datasets is 2-site clusters approximation such as the one used for cmap-ACE inference (see methods and [31]), where the contact map is inferred by application of PLM-DCA [56] on the same MSA. For each data set four sparse models, with a variable number of inferred contacts, have been tested. Calling N the lenght of the protein, the sparsity was set such that the number of clusters K equals $0.5N, N, 2N, 3N$.

In Fig. 10.5 we compare these models with PLM and the independent model, showing that the performance has a maximum for one of the sparse structural models (leftmost points of each curve are the performance of PLM, rightmost refer to the independent model). For all four families, the performance peak (green circle) is in the vicinity of $P = B$, suggesting that the $P = B$ criterion could be a practical rule of thumb to approach the optimal inference in the subsampled regime.

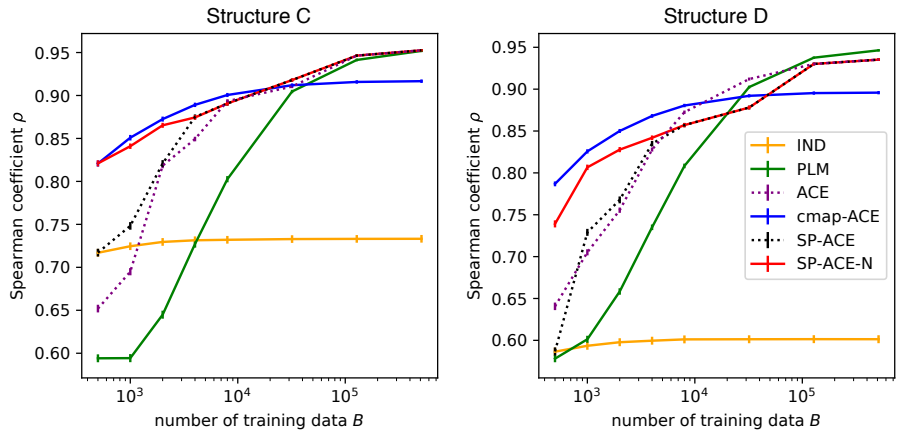


Figure 10.4: **Performance of SPACE and SPACE-N compared** Spearman coefficient of the inferred and real mutational landscape, as a function of the number of sequences in the training MSA, for the SP-ACE (black dotted line), ACE (purple dotted line), and SP-ACE-N (red line), compared to the three models reported in Fig. 10.1. For each B the mean and standard error of the inference performance are computed on 256 different wildtypes. $\theta = 25$.

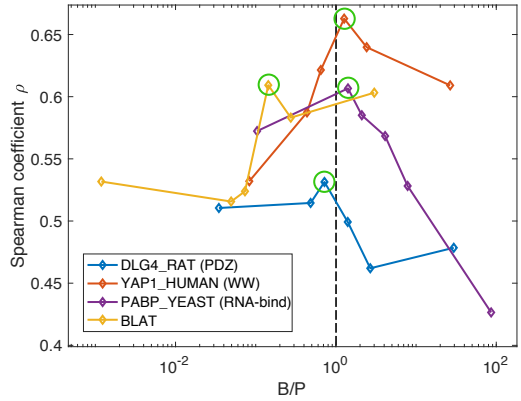


Figure 10.5: **Adapting the sparsity of the inferred Potts model: real data.** Performance, computed as the Spearman coefficient between inferred mutational landscape (using cmap-ACE) and the one estimated by mutagenesis experiments, as a function of the sparsity of the inferred model. Since true contacts are not known a priori, they are first estimated by DCA using PLM [57]. Performances of cmap-ACE are compared to the ones of fully-connected PLM (leftmost points) and independent model (rightmost points). The optimal sparsities are highlighted by green circles, the dashed black line represents the rule of thumb $B = P$. Analysis performed by Francesca Rizzato [8].

10.3 DISCUSSION

In this work, we tested the idea of adapting the sparsity of the inferred Potts model, aimed at retrieving the fitness landscape of a protein from sequence data, to the number of available training sequences in the MSA.

Deriving meaningful observables from the inferred Potts model in the under-sampled regime is highly non trivial, and the meaning of the inferred parameters are still object of debate in the DCA community [235, 236]. Therefore, reducing the number of inferred parameters is a problem of central importance for practical applications as well as theoretical interpretation of the inference results. The work here presented follows a number of propositions, in the literature, to tackle this problem. Examples include clustering multiple sites in one single variable [245], or pre-selection of the most relevant set of couplings [246, 247], an approach that has been proven successful in providing biologically-meaningful predictions on genome-wide modelling [225]. Another approach to the reduction of inferred parameters, proposed by members of our group, is to decrease the number of colors-per-site, a procedure called "color compression" [244].

We implemented an ad-hoc version of the 2-cluster approximated inference introduced by Barton et al. [47], previously carried from the ACE routine with the composition of the commands $-cmap$ and $-t$ 1. This routine is tested on Lattice Proteins, showing that adapting the sparsity on the known set of structural contacts leads to a significant improvement, in the under-sampled regime, with respect to standard plmDCA, however being limited when the number of data increases. We then tested an unsupervised method that adapts the sparsity following the rule of thumb P (number of paramters) $= B$ (number of data). As expected, this model is equivalent to plmDCA and ACE for large B, while yielding superior performances in the regime of low number of data. Overall the findings on Lattice Proteins suggest that one can adapt the sparsity of the model, including only the most informative parameters by imposing a hard constraint on the topology of the inferred interaction matrix $\mathbf{J}$. In the case of fitness predictions from sequence data of real protein families, one could, therefore, leverage on structural or functional information, for example, if some sites on the chain are known (or inferred [248]) to play a crucial role in the function of the protein, to choose to model only a selected subset of the interactions.

Therefore, we propose these approaches as unsupervised methods to obtain close-to-optimal predictions in both the under-sampled and well-sampled regime. This work has been carried in parallel with the analysis of real protein data sets from Francesca Rizzato, and will be included in a joint publication [8].

The presence of an optimum in the axis of model complexity is often referred to as "bias-variance tradeoff" in the statistics literature [249, 250]. In the next chapter, we will give a more precise description of how bias and variance control the performance of the Potts model in retrieving the fitness landscape of a protein.

10.4 METHODS

Contact-map cluster expansion (cmap-ACE) with known contacts

The inference is performed by considering the 2-cluster approximation of the inverse Potts model [31]. This approximation can be solved analytically in the case of non-regularized inference. In the present work we used a Bayesian regularization $\lambda_h = \frac{0.1}{B}$ for the field terms and $\lambda_J = \frac{1}{B}$ for the coupling terms, therefore the problem has to be solved numerically. The cross-entropy (log likelihood with a minus sign) of each two-site (ij) inverse problem is

$$\mathcal{S}_{ij} = \sum_{a,b=1}^{Q_i,Q_j} J_{ij}^{(ij)}(a,b) \cdot p_{ij}(a,b) + \sum_{a=1}^{Q_i} h_i^{(ij)}(a) p_i(a) + \sum_{b=1}^{Q_j} h_j^{(ij)}(b) p_j(b) \tag{10.2}$$

$$+ \log\left(1 + \sum_{a=1}^{Q_i} e^{h_i^{(ij)}(a)} + \sum_{b=1}^{Q_j} e^{h_j^{(ij)}(b)} + \sum_{a,b}^{Q_i,Q_j} e^{h_i^{(ij)}(a) + h_j^{(ij)}(b) + J_{ij}^{(ij)}(a,b)}\right) \tag{10.3}$$

$$+ \ \lambda_h \cdot \left(\sum_{a=1}^{Q_i} \left[h_i^{(ij)}(a)\right]^2 + \sum_{b=1}^{Q_j} \left[h_j^{(ij)}(a)\right]^2\right) + \lambda_J \cdot \left(\sum_{a,b=1}^{Q_i,Q_j} \left[J_{ij}^{(ij)}(a,b)\right]^2\right) \quad , \tag{10.4}$$

where Q_i and Q_j is the total number of un-compressed colors on sites i and j (see [226,244] for color-compression in the Potts inference). The gradient is

$$\nabla \mathcal{S}_{ij} = \left(\frac{\partial \mathcal{S}_{ij}}{\partial h_i^{(ij)}(1)}, \ldots, \frac{\partial \mathcal{S}_{ij}}{\partial h_i^{(ij)}(Q_i)}, \frac{\partial \mathcal{S}_{ij}}{\partial h_j^{(ij)}(1)}, \ldots, \frac{\partial \mathcal{S}_{ij}}{\partial h_j^{(ij)}(Q_j)}, \right. \tag{10.5}$$

$$\left. \frac{\partial \mathcal{S}_{ij}}{\partial J_{ij}^{(ij)}(1,1)}, \frac{\partial \mathcal{S}_{ij}}{\partial J_{ij}^{(ij)}(1,2)}, \ldots, \frac{\partial \mathcal{S}_{ij}}{\partial J_{ij}^{(ij)}(Q_i,Q_j)}\right) \tag{10.6}$$

The function is minimized for each couple of sites (i,j) in the contact map C, by Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (scipy implementation), and the final fields and coupling are built, following [45], as

$$h_i(a) = h_i^{ind}(a) + \sum_{j \in C_i} \Delta h_i^{(ij)}(a) \tag{10.7}$$

$$J_{ij}(a,b) = J_{ji}(b,a) = J_{ij}^{(ij)}(a,b) \tag{10.8}$$

Where C_i is the set of sites that are in contact with i, $\Delta h_i^{(ij)}(a) := h_i^{(ij)}(a) - h_i^{ind}(a)$, and $h^{ind} = \log(p_i(a) + p_0)$. The pseudo-count $p_0 = \frac{1}{B}$ is added to the 1-point averages to avoid divergent solutions.

Structural-prior adaptive cluster expansion (SP-ACE)

The method is a declination of the cross-entropy cluster expansion described in [33, 45]. Briefly, the cross-entropy of the inverse problem is expanded into all k-cluster contributions

$$S(\mathbf{h}, \mathbf{J}|MSA) = \sum_{k=1}^{N} \sum_{\Gamma \in \mathcal{C}_k} S_\Gamma \tag{10.9}$$

Where $\mathcal{C}_k$ is the set of all subsets of size k of the N sites. The inference process starts from clusters of size two and includes all those clusters whose contribution to the total cross-entropy is above a threshold θ_t. For larger k, only clusters composed of previously-included clusters are considered. At the next iteration, the threshold θ is lowered $\theta_{t+1} = \theta_t - \Delta\theta$, therefore more clusters are included in the expansion. The iteration stops when desired criterion of convergence (usually defined through the largest error ϵ_t in the moment-matching conditions) is matched by the proposed solution $(\mathbf{h}_t, \mathbf{J}_t)$.

Here we employed the criterion of convergence introduced in [244]: we monitor the number of 2-clusters included in the expansion at each iteration, and we stop the process at when this number reaches a pre-determined number K. The corresponding iteration is called t_K. Then we analyze the profile of the largest error of the moment-matching condition ϵ_t (4th column of ACE output from https://github.com/johnbarton/ACE) for all $t < t_K$, and we take the threshold θ_{t^*} that corresponds to the lowest error, i.e., $t^* = argmin_t\, \epsilon_t$. The corresponding Potts model $(\mathbf{h}_{t^*}, \mathbf{J}_{t^*})$ is then returned. This corresponds to the best ACE solution conditioned to have a topology of interaction with K or fewer bounds.

The number of two-clusters K is chosen such that the total number of inferred parameters P is equal to the number of available training points. In the case of compressed-color Potts inference [244] we restrict the inference to Q_i colors for each site i, which correspond to the number of Potts states a such that $p_i(a) > p_{cut}$. In the present work we use $p_{cut} = 0$, i.e., we set the value of unseen colors to a Bayesian prior value $p_{unseen} = B^{-1}$. As a consequence, the total number of inferred fields is $\bar{Q} \cdot N$, with $\bar{Q} := \frac{1}{N} \sum_i Q_i$. Each pairwise interaction of the included K yields a number of parameters that depend on the values of Q_i. To estimate a-priori this value we computed the average coupling matrix size as $\bar{Q^2} := \frac{2}{N(N-1)} \sum_{i<j} Q_i Q_j$. The number of interactions is therefore set as

$$P = K \cdot \bar{Q^2} + N \cdot \bar{Q} \implies K(B) = \frac{B - N \cdot \bar{Q}}{\bar{Q^2}} \tag{10.10}$$

Note that both $\bar{Q}$ and $\bar{Q^2}$ can be computed directly from the MSA. In the SP-ACE-N version we exploit the fact that the dominant interactions in Lattice Proteins are of the order of the number of sites (28 contacts for 27 sites) and use

$$K = max\left(N, \frac{B - N \cdot \bar{Q}}{\bar{Q^2}}\right) \tag{10.11}$$

Finally, if $K \geq \frac{N(N-1)}{2}$ we use the normal ACE routine [47] until convergence.

11

THE FOCUSING PROCEDURE: SELECT THE OPTIMAL TRAINING MSA FOR FITNESS PREDICTIONS

11.1 INTRODUCTION

Retrieving the mutational landscape of a protein is a problem of fundamental importance in bioinformatics and in biology in general. In bioengineering, characterizing the local fitness landscape of a specific sequence can guide the de-novo design of highly-functional mutants of a known genotype; in vaccine-antibiotics design, knowledge of the mutational landscape of a protein is of fundamental importance due to its tight relation to the evolvability (therefore the escape capability) of the host specimen from external, drug-induced, pressure [251, 252]. It has recently been shown that the *max-entropy* statistical model inferred from the alignment of homologous sequences (MSA), called Potts model in Statistical Physics, outperforms standard bioinformatics and biophysics methods in predicting the single-point mutation effects observed in mutagenesis experiments [44, 234, 252].

Performances, however, with Spearman coefficients typically ranging from $\rho \sim 0.4$ to $\rho \sim 0.7$ in the best cases, are far from being optimal. The max-entropy approach indeed suffers of a number of limitations [235, 236]. Among others, two key issues that affect the performance of the retrieval of single-mutation effects are

(a) the Potts model is a low-order approximation of the complex and rough fitness landscape explored by evolution, from which sequences included in the training MSA are drawn. This inconsistency, therefore, leads to **biased** estimations, due to un-modelled high-order couplings and unknown degrees of freedom.

(b) the number of available sequences in the training MSA is usually much lower than the number of inferred parameters. The over-parametrization of the statistical model causes a high **variance** of its predictions, affecting the retrieval performance.

In this chapter, we will use Lattice Proteins as a model to investigate both these issues in the context of bias-variance tradeoff [249, 250, 253]. We will show that, given one MSA, the performance of the model inferred from it can be predicted in terms of two simple descriptors: the number of sequences B, which is connected to the scaling of the variance (b), and the average Hamming distance D from the wildtype, which is related to the bias (a) of the inferred model. We will derive a scaling law for the performance as a linear sum of two terms that are computable a priori from the sequence data in linear time. From this derivation, we will show that one can improve the performance of the inference model by choosing a subset of the MSA that is optimal in terms of bias-variance tradeoff. We will finally present a procedure, called **focusing**, that retrieves the optimal training set from a starting MSA.

11.2 BIAS-VARIANCE TRADEOFF IN INDEPENDENT-POTTS INFERENCE

As pointed above, the probability of observing a sequence in nature is clearly not a simple exponential model truncated to pairwise interaction terms. Even in the reasonable, but not obvious, approximation that this probability is related to a fitness function $\mathcal{F}$ that depends on the sequence only, i.e. (in the exponential formulation)

$$P(\mathbf{s}) = \frac{e^{\mathcal{F}(\mathbf{s})}}{\mathcal{Z}} \quad , \tag{11.1}$$

this latter will certainly contain terms of any order of interaction. Being defined over a discrete set (all sequences of length N), we can always write $\mathcal{F}(\mathbf{s})$ as a sum of all possible interaction terms for all the possible combinations of amino acids, times a delta function that is one only if the sequence $\mathbf{s}$ displays that specific combination on those specific sites. By defining $\theta_{\boldsymbol{\pi}}(\mathbf{a})$ as the fitness interaction term that the amino acids $\mathbf{a} = a_1, a_2, \dots$ would yield if placed on sites $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$, we can therefore expand $\mathcal{F}$ as:

$$\mathcal{F}(\mathbf{s}) \; = \sum_{\boldsymbol{\pi} \in \mathcal{P}(N)} \; \sum_{\mathbf{a}:|\mathbf{a}|=|\boldsymbol{\pi}|} \theta_{\boldsymbol{\pi}}(\mathbf{a}) \; \delta(\mathbf{s}_{\boldsymbol{\pi}}, \mathbf{a}) \tag{11.2}$$

Where $\mathcal{P}(N)$ is the set of all possible combinations of sites of any size from 1 to N, $\delta(\mathbf{s}_{\boldsymbol{\pi}}, \mathbf{a}) := \delta(s_{\pi_1}, a_1) \cdot \delta(s_{\pi_2}, a_2) \cdots$, and $|\mathbf{a}|$ is the number of elements in the vector $\mathbf{a}$. If we separate the sum into the n-wise interaction terms, we can write it as

$$\mathcal{F}(\mathbf{s}) \; = \sum_{n=1}^{N} \; \sum_{\boldsymbol{\pi}:|\boldsymbol{\pi}|=n} \; \sum_{\mathbf{a}:|\mathbf{a}|=n} \theta_{\boldsymbol{\pi}}(\mathbf{a}) \; \delta(\mathbf{s}_{\boldsymbol{\pi}}, \mathbf{a}) \tag{11.3}$$

If we truncate the sum to pairwise terms ($|\boldsymbol{\pi}| \leq 2$), we retrieve a (minus) Potts energy function, where the values of $\boldsymbol{\pi}$ can only take single indexes and pairs, i.e.

$$\boldsymbol{\pi} \in \{1, 2, \dots, N, (1,2), (1,3), \dots, (1,N), \dots, (N-1, N)\} \tag{11.4}$$

and the parameters θ are fields and couplings, i.e.

$$\begin{cases} \theta_i(a) = h_i(a) \\ \theta_{ij}(ab) = J_{ij}(a,b) \end{cases} \tag{11.5}$$

As for the Potts Hamiltonian, the function $\mathcal{F}$ is over-parametrized. We therefore need to choose a gauge sequence $\mathbf{s}^0$ such that all multi-body interaction terms $\theta_{\boldsymbol{\pi}}(\mathbf{a})$ that contain at least one amino acid a_i equal to the gauge sequence on the site π_i, i.e. $s^0_{\pi_i} = a_i$, are 0. This ensures that $\mathcal{F}(\mathbf{s}_0) = 0$. In what follows we will conveniently gauge our theory on the wildtype sequence (the one whose mutational landscape we want to predict), i.e. $\mathbf{s}^0 = \mathbf{s}^{WT}$.

Let's now suppose that, in an idealized situation, we are able to sample a large collection of natural sequences $\{\mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^B\}$, which all differ from the wildtype $\mathbf{s}^{WT}$ by only one single mutation. Assuming that all the sequences are unbiasedly sampled from the same fitness function, the characterization of the single-mutation landscape of the wildtype would be, in this case, straightforward. Indeed it would only require to compute the statistics of amino acids on single sites, since nature itself provided us with a good sample of the local landscape region that we want to characterize. Formally, we would invert the single-point statistics

$$p_i(a) = \frac{1}{B} \sum_{k=1}^{B} \delta(s_i^k, a) \quad , \tag{11.6}$$

to infer an independent Potts model. Since we chose the wildtype as the gauge of the fitness expansion in (11.2), we will have

$$\mathcal{F}(\mathbf{s}^0_{i\to a}) = \theta_i(a) \implies \frac{p_i(a)}{p_i(s_i^0)} = \frac{\text{Prob}(\mathbf{s}^{wt}_{i\to a})}{\text{Prob}(\mathbf{s}^0)} = e^{\theta_i(a)} \quad . \tag{11.7}$$

Therefore, irrespectively of the complexity of the real fitness function $\mathcal{F}$, the independent Potts field $h_i(a)$ computed from the single-site statistics will correctly retrieve the single-point fitness difference:

$$-\Delta\mathcal{H}^{Potts}_{ia} = h_i(a) = \log \frac{p_i(a)}{p_i(s_i^{wt})} = \theta_i(a) \quad . \tag{11.8}$$

In other words, in this idealized scenario there is *no background effect* on the single mutations, and therefore we can directly access an unbiased estimation of the local landscape of the wildtype.

This is of course not true in the still-idealized but less-convenient scenario where only sequences at distance two are available. In this case, the single-point statistic $p_i(a)$ would not be sufficient to retrieve the fitness effect of the mutation $i \to a$, since the amino acid on the "second" mutated site will interact via pairwise epistasis, biasing our single-point estimation. The fitness of a sequence at distance two from the gauge is

$$\mathcal{F}(\mathbf{s}^0_{i\to a, j\to b}) = \theta_i(a) + \theta_j(b) + \theta_{ij}(ab) \tag{11.9}$$

which leads to a single-point statistic

$$\frac{p_i(a)}{p_i(s_i^{wt})} \simeq \frac{\sum_{\mathbf{s} \text{ at } D=2} P(\mathbf{s})\ \delta(s_i, a)}{\sum_{\mathbf{s} \text{ at } D=2} P(\mathbf{s})\ \delta(s_i, s_i^{wt})} \tag{11.10}$$

$$= \frac{\sum_{j\neq i} \sum_{b} e^{\theta_i(a)+\theta_j(b)+\theta_{ij}(a,b)}}{\sum_{j,k\neq i} \sum_{b,c} e^{\theta_j(b)+\theta_k(c)+\theta_{jk}(b,c)}} \tag{11.11}$$

$$:= e^{\theta_i(a)} X_{i,a} \tag{11.12}$$

where $X_{i,a} := \frac{\sum_{j\neq i}\sum_{b} e^{\theta_j(b)+\theta_{ij}(a,b)}}{\sum_{j,k\neq i}\sum_{b,c} e^{\theta_j(b)+\theta_k(c)+\theta_{jk}(b,c)}}$ depends on the specific mutation we are inferring. Therefore, when we infer our independent Potts parameter we will obtain a biased estimation

$$-\Delta\mathcal{H}_{ia}^{Potts} = h_i(a) = \log\frac{p_i(a)}{p_i(s_i^{wt})} = \theta_i(a) + \log X_{i,a} \neq \theta_i(a) \tag{11.13}$$

This argument can be generalized to any distance greater than two, and consequently to any order of epistasis: the larger the number of mutations, i.e., the Hamming distance, the more important is the bias induced by high-order background effects to single-points estimations. In an idealized scenario, therefore, the best choice to characterize single-point mutations would be to build an MSA with only sequences at distance $D = 1$ and infer an independent Potts model from it. Some single-point mutagenesis experiments indeed reproduce a controlled version of this scenario. They yield unbiased estimations from single-point statistics since the fitness effect is retrieved from observations of the survival rate of organisms all of which express the same protein except for one single mutation.

The performance ρ scales with the descriptors B and D

Thanks to the controllability of the training MSA of Lattice Proteins we can test the idea put forward in the previous section, i.e., that the mean Hamming distance D affects the performance of the inferred model due to its relation with unmodelled high-order terms. We performed a systematic analysis by sampling several MSAs with controlled descriptors D and B and tested the retrieval performance ρ of an independent Potts model inferred on these training data. To obtain a fine-controlled distance D, we started from four main-MSAs sampled with different values of $b = 0.0, 0.025, 0.050, 0.075$ in Eq. 8.13 (controlling the distance), following the method described in [58]. We then built a training MSA by mixing sequences from the four main MSAs with an iterative procedure that stops when the sample reaches the desired mean Hamming distance D and the number of sequences B.

In Fig. 11.1 (left panel) we show 9 examples of the comparison between retrieved and real single-point mutational landscape in as many different sampling conditions of the training MSA. We observe that the retrieval performance, measured by the Spearman correlation coefficient ρ, depends on the two descriptors D (left panel, y axis) and B (left panel, x axis). Intuitively, a larger number of training data corresponds to more accurate inference. Interestingly, we also observe that models trained on MSAs with smaller mean Hamming distance D yield the best performances, confirming the role of this descriptor in affecting the inference precision. A systematic analysis of several combinations of the descriptors, reported in Fig. 11.1 (right panel), shows that the performance ρ depends almost linearly from both of them. In the next section, we will try to explain this linear dependence by framing the inference problem in the context of bias-variance tradeoff.

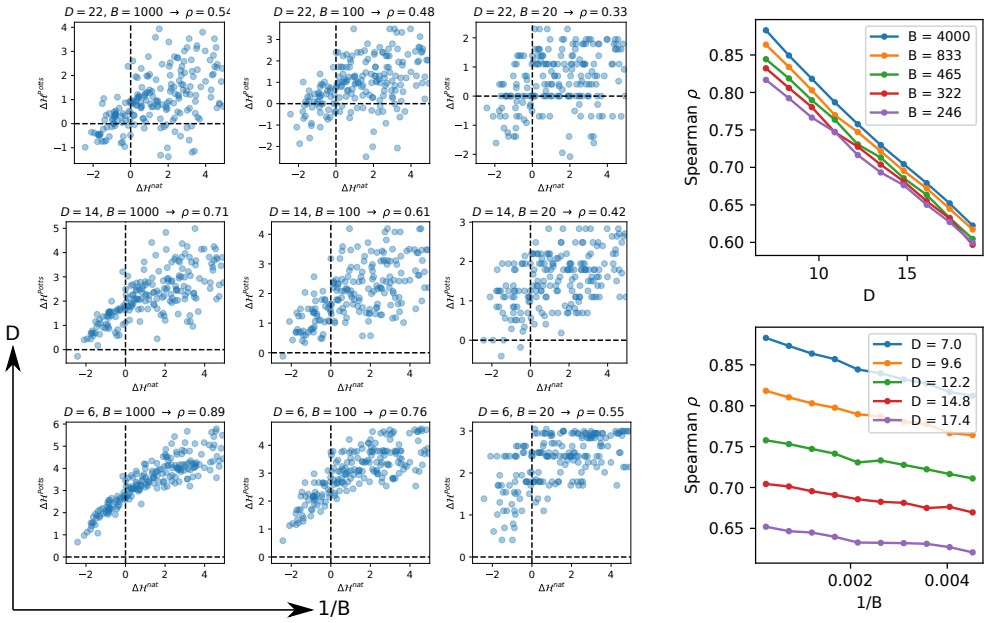


Figure 11.1: **Linear dependence on B and D of the performance ρ. Left**: 9 examples of the comparison between independent-Potts-inferred and real single-mutation landscape of a *wildtype* sequence in different sampling conditions of the training data. A dependence on the number of data (B) and the mean hamming distance of sequences in the MSA from the wildtype is reported. **Right**: the performance ρ (Spearman coefficient of inferred and real mutational landscape) is inversely proportional, almost linearly, to the value of the two descriptors D and $1/B$. Each point is the average ρ obtained on $n = 10$ different MSAs with fixed descriptors.

Computation of bias and variance

It is known from the statistics literature that the mean squared error (MSE) of an estimator $\widehat{Y}$, for a variable Y, can be divided into a linear sum of two terms: the squared *bias* (μ^2) and the *variance* (σ^2) of the estimator, plus a term that does not depend on the estimator [249, 253]

$$\underbrace{\langle\left(\widehat{Y}-Y\right)^2\rangle}_{MSE} = \underbrace{\left(\langle\widehat{Y}-Y\rangle\right)^2}_{\mu^2:=bias^2} + \underbrace{\left(\langle\widehat{Y}^2\rangle-\langle\widehat{Y}\rangle^2\right)}_{\sigma^2:=variance} \tag{11.14}$$

where $\langle . \rangle$ is the average over the probability distribution of the sample data from which the inference is performed. By recalling our variables $\widehat{Y} = \Delta\mathcal{H}_{ia}^{Potts}$ and $Y = \Delta\mathcal{H}_{ia}^{nat}$, the bias and variance of each estimator (for each single-point mutation i, a) are therefore defined as

$$\mu_{ia}^2 := \left(\langle\Delta\mathcal{H}_{ia}^{Potts} - \Delta\mathcal{H}_{ia}^{nat}\rangle\right)^2 \tag{11.15}$$

$$\sigma_{ia}^2 := \left(\langle\left[\Delta\mathcal{H}_{ia}^{Potts}\right]^2\rangle - \langle\Delta\mathcal{H}_{ia}^{Potts}\rangle^2\right) \quad . \tag{11.16}$$

Our goal is to investigate the role of the descriptors B and D in controlling the performance of the inference through their relation with the variance and the bias, respectively. The averages in Eq.s 11.15 and 11.16 are therefore computed on several (typically $n = 10$) MSAs with different sequences and identical descriptors B, D.

To connect the bias and variance to the performance of the inference, i.e., the Spearman correlation coefficient ρ of the inferred-vs-real single mutational landscape, we define two global measures of bias and variance that account for all single-point mutations by averaging over the index (i, a):

$$\mu^2 := \langle\mu_{ia}^2\rangle_{ia|\theta} - \langle\mu_{ia}\rangle_{ia|\theta}^2 \tag{11.17}$$

$$\sigma^2 := \langle\sigma_{ia}^2\rangle_{ia|\theta} \quad , \tag{11.18}$$

where the notation $\langle\rangle_{ia|\theta}$ stands for the average over all the single-point mutations such that $\Delta\mathcal{H}_{ia}^{nat} < \theta$. As we saw in the previous chapters, some mutations cause such a negative effect on the fitness that are never observed, therefore we can use only prior information to estimate their value from data. To provide a precise quantitative description of the dependence of bias and variance from the sampling conditions we restricted our analysis, without loss of generality, to the least-noisy mutations by setting $\theta = 5$. Since the Spearman correlation coefficient ρ is unaffected by a global shift of the single points $\Delta\mathcal{H}_{ia}^{Potts}$, in Eq. 11.17 we defined a measure of global bias that disregards the global displacement (the mean bias $\langle\mu_{ia}\rangle_{ia}^2$) of the inferred mutational effects.

Bias and variance control the inference performance ρ

In the case of Lattice Proteins, knowing the real single-mutation landscape, we can estimate these quantities for each performed inference and correlate them with the performance ρ. We numerically estimated the global bias and variance for $n = 400$ different combinations of the two descriptors, $B \in [200, 4000]$ and $D \in [7, 20]$. We then correlated the Spearman coefficient, averaged over the $n = 10$ MSAs with equal descriptors, with their linear sum $\mu^2 + \sigma^2$. Results, reported in Fig. 11.2 (a,b), show a perfect linear correlation ($R^2 \simeq 1$). This confirms that the formulations given in Eq.s 11.17 and 11.18 are the correct terms that control the fitness performance of the inferred independent Potts model. This results holds as well for different protein families (see Appendix).

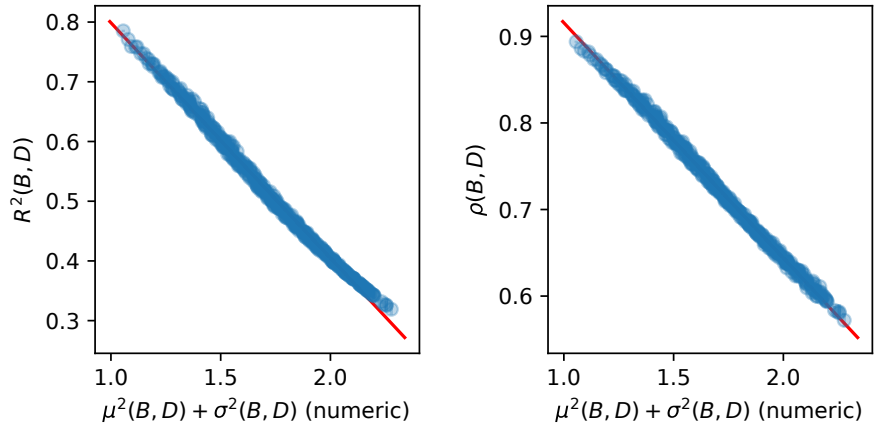


Figure 11.2: **Independent Potts: Inference performance ρ scales with numerical bias and variance** Comparison between the performance of the Independent Potts model, quantified by the Spearman correlation coefficient of inferred and real mutational landscapes, with the linear sum of numerical bias and numerical variance. Each point is obtained by averaging over $n = 10$ different MSAs with same descriptors B and D, which vary in the intervals $B \in [200, 4000]$ (equally spaced $\frac{1}{B}$ values) and $D \in [7, 20]$ (equally spaced values). Structure A, $\theta = 5$, 400 points.

Analytical estimation of bias and variance in terms of the descriptors B,D

Considered the scaling observed in Fig. 11.2, we could be tempted to explain the linear relationship between each of the two descriptors, B^{-1} and D, and the inference performance ρ, observed in Fig. 11.1, as a consequence of bias and variance terms that linearly

depend on the descriptors, i.e., $\mu^2 \propto D$ and $\sigma^2 \propto B^{-1}$. Here we will show, by analytical development of a simple case, that this is indeed the case.

VARIANCE An approximate expression for the variance of the inferred independent Potts model can be computed by error propagation from the uncertainty of single-point estimations. A derivation can be found in [47], Supplementary Material. The variance reads

$$\widehat{\sigma}^2_{ia} = \frac{1}{B}\left(\frac{1-p_i(a)}{p_i(a)} + \frac{1-p_i(s_i^{wt})}{p_i(s_i^{wt})}\right) \quad , \tag{11.19}$$

from which we immediately retrieve the scaling $1/B$, observed in Fig. 11.1, for the global "theoretical" variance:

$$\widehat{\sigma}^2 \;:=\; \frac{1}{B}\cdot\mathcal{V}^{ind} \;:=\; \frac{1}{B}\left\langle \frac{1-p_i(a)}{p_i(a)} + \frac{1-p_i(s_i^{wt})}{p_i(s_i^{wt})} \right\rangle_{ia} \quad . \tag{11.20}$$

The term $\mathcal{V}^{ind}$ depends on the sequences in the MSA but is reasonably regular for MSAs drawn from the same protein family. Therefore, the dominant factor for the variance, at least in the case of the independent model, is the scaling B^{-1}. The full formulation is nevertheless necessary to obtain a reliable quantitative estimation of the variance.

BIAS Estimating the bias of a statistical model is generally complicated since it involves, by definition, un-modelled degrees of freedom from the "real", unknown, underlying model. The reasoning put forward in the previous section suggests that the Hamming distance could be taken as a correlate for the bias of single-point statistics, since the background effect of high-order interactions is reduced when the distance is lowered. We will here show how this intuition can be formalized in the case in which the "real" fitness function is a pairwise Potts energy, providing an example of how the Hamming distance directly relates to the inference bias of an independent Potts model.

In the case of a probability distribution resulting from a Potts Hamiltonian with parameters $J_{ij}(a,b)$, $h_i(a)$, i.e. $P_{potts}(\mathbf{s}) = \frac{1}{\mathcal{Z}}\exp\left(\sum_i h_i(s_i) + \sum_{i<j} J_{ij}(s_i,s_j)\right)$, the single-point average $p_i(a)$ can be written as

$$p_i(a) = \left\langle \frac{e^{h_i(a)+\sum_j J_{ij}(a,s_j)}}{\sum_b e^{h_i(b)+\sum_j J_{ij}(b,s_j)}} \right\rangle_{P_{Potts}(\mathbf{s})} \quad , \tag{11.21}$$

which is is called the Callan identity. If we infer an independent Potts in the wildtype gauge from these single-point averages, the inferred field $\widehat{h}_i(a)$ will be

$$\widehat{h}_i(a) = \log\frac{p_i(a)}{p_i(s_i^{wt})} = \log\left\langle \frac{e^{h_i(a)+\sum_j J_{ij}(a,s_j)}}{\sum_b e^{h_i(b)+\sum_j J_{ij}(b,s_j)}} \right\rangle_{P_{Potts}(\mathbf{s})} - \log\left\langle \frac{1}{\sum_b e^{h_i(b)+\sum_j J_{ij}(b,s_j)}} \right\rangle_{P_{Potts}(\mathbf{s})} \quad , \tag{11.22}$$

where $J_{ij}(s_i^{wt}, b) = 0 \ \forall b$. Since the averages $\langle\rangle_{P_{Potts}(\mathbf{s})}$ are difficult to compute analytically, we need to resort to approximations. A first possibility is to make the *annealed* approximation, i.e. $\log\langle\cdot\rangle \simeq \langle\log\cdot\rangle$. When applied to the (11.22), it yields

$$\widehat{h}_i(a) \simeq h_i(a) + \left\langle \sum_j J_{ij}(a, s_j) \right\rangle_{P_{Potts}(\mathbf{s})} \quad . \tag{11.23}$$

Therefore, the bias of the estimator for the single mutation is

$$\widehat{\mu}_i(a) := h_i(a) - \widehat{h}_i(a) = \left\langle \sum_j J_{ij}(a, s_j) \right\rangle_{P_{Potts}(\mathbf{s})} \quad . \tag{11.24}$$

Let's now assume that we have a collection of B sequences $\left\{\mathbf{s}^k\right\}_{k=1}^{B}$ i.i.d. sampled from $P_{Potts}(\mathbf{s})$ (i.e. our MSA). If B is large enough, we can assume that sample averages are good estimations of the distribution averages $\langle\rangle_{P_{Potts}(\mathbf{s})}$, therefore

$$\widehat{\mu}_i(a) = \left\langle \sum_j J_{ij}(a, s_j) \right\rangle_{P_{Potts}(\mathbf{s})} \simeq \frac{1}{B} \sum_{k=1}^{B} \sum_{j=1}^{L} J_{ij}(a, s_j^k) \quad . \tag{11.25}$$

Since we gauged the model on the wildtype sequence, each $J_{ij}(a, s_j^k)$ where $s_j^k = s_j^{wt}$ is zero. Therefore, the RHS in (11.25) is a sum of $B \times D$ terms, where D is the average hamming distance of the sequences in our collection from the wildtype. If we make the simplified assumption that each $J_{ij}(a, b)$ is sampled from a random distribution, i.e.

$$J_{ij}(a, b) \sim \mathcal{N}(\bar{J}, J_0) \tag{11.26}$$

we finally find our *ansatz* expression of the bias as

$$\widehat{\mu}^2 = var_{ia}\ \widehat{\mu}_i(a) \simeq J_0 \cdot D \tag{11.27}$$

from which we retrieve the linear scaling with D observed in Fig. 11.1. The coefficient J_0 depends on the variance of high-order (pairwise, in this case) interactions. Therefore, its value has to be estimated for each different protein family, see below for further discussion.

in Fig. 11.3, we compared the theoretical estimations of bias $\widehat{\mu}^2$ and variance $\widehat{\sigma}^2$ with their numerical counterparts in the same $n = 400$ different combinations of B and D of the analysis shown in Fig. 11.2. The theoretical variance, defined in Eq. 11.20, correlates well and almost linearly with the numerical one, although with a pre-factor $\neq 1$. This pre-factor is likely due the theoretical approximation of the variance, derived by assuming an independent model as the probability distributions that rule the $p_i(a)$. Therefore we have the relation

$$\widehat{\sigma}^2 = \alpha\ \sigma^2 \quad , \tag{11.28}$$

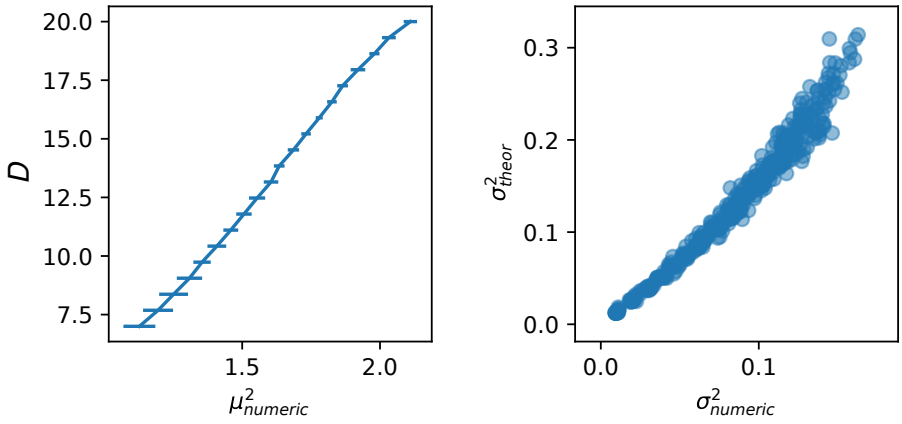


Figure 11.3: **Independent Potts: Theoretical estimations of bias and variance correlate with the numerical counterparts.** Comparison between theoretical estimations of bias μ^2 and variance σ^2 with the numerical ones. Error bars in the left panel refer to mean and standard error computed over $n = 20$ cases at fixed D. Same procedure and data of Fig. 11.2

with $\alpha \sim [1.5, 2]$ for the three structures here considered (see Appendix).

Surprisingly, we also find that the numerical bias scales well, and almost linearly, with the average hamming distance D, confirming the ansatz of Eq. 11.27. By linear fit of this dependence we can infer the value of J_0 in Eq. 11.26, reported in Tab. 11.1. From the derivation of the linear dependence in Eq. 11.27 the J_0 encodes the variance of the $J_{ij}(a,b)$ couplings. We can therefore try to give an estimate for this value by inferring a fully-connected Potts model (in the wildtype gauge) from a very large MSA, then computing

$$\widehat{J}_0 \;=\; \frac{1}{nc}\sum_{i,a|\theta}\sum_{j,b}\widehat{J}_{ij}(a,b)^2 \;-\; \left(\frac{1}{nc}\sum_{i,a|\theta}\sum_{j,b}\widehat{J}_{ij}(a,b)\right)^2 \quad , \tag{11.29}$$

where $\widehat{J}_{ij}(a,b)$ are the inferred couplings (using ACE fully-connected inference) and $nc := \sum_{i,a|\theta}\sum_{j,b}\delta(a \neq s_i^{wt})\delta(b \neq s_j^{wt})$. As shown in Table 11.1, the Potts-estimated $\widehat{J}_0$ is well retrieved for the structure C, which is the one whose fitness function mostly resembles a pairwise model [58], while is over-estimated, still in the same order of magnitude, for structure A and D.

structure	J_0 (linear fit)	$\widehat{J_0}$ from ACE Potts inference
A	0.071	0.140
C	0.077	0.071
D	0.063	0.142

Table 11.1: **Bias term** J_0: comparison between the bias factor J_0 of Eq. 11.26 inferred from linear fit of numeric bias and Hamming distance (first column) and the factor computed from inferred fully-connected Potts model from a large MSA ($B = 625,000$) from Eq. 11.29.

These results suggest that the performance ρ of the independent Potts model could be deduced directly from sequence data of the training MSA, up a single parameter J_0', as a function of the linear sum of theoretical bias and variance:

$$\rho(MSA) \equiv \rho\Big(\underbrace{J_0' \cdot D}_{\text{th. bias } \widehat{\mu}^2} + \underbrace{\frac{1}{B} \cdot \mathcal{V}^{ind}}_{\text{th. variance } \widehat{\sigma}^2} \Big) \tag{11.30}$$

Where $J_0' := \alpha\ J_0$ includes the pre-factor in front of the theoretical variance with respect to the numeric one. If the scaling law of Eq. 11.30 is valid, we can retrieve the parameter J_0' by supervised maximization of the correlation coefficient between the performance ρ and the sum of theoretical bias and variance for many different training MSAs, see Fig. 11.5. Once the parameter J_0' is inferred, the scaling law is confirmed with astonishing precision ($R^2 \sim 1$), see Fig. 11.4.

If represented as a 2-dimensional function where one independent variable is the mean hamming distance ($\widehat{\mu}^2 \sim D$) and the other is the inverse of the number of sequences of the training MSA ($\widehat{\sigma}^2 \sim \frac{1}{B}$), the scaling of the performance $\rho(B, D)$ draws a plane in the 3D space, corresponding to linear and parallel level curves in the 2D projection, whose slope depends on the ratio between J_0' and the mean rescaled variance $\mathcal{V}^{ind}$. We call this representation the **bias-variance diagram**, see Fig. 11.6. As we will see in the next section, this representation is helpful in visualizing the "focusing procedure" as a trajectory on the bias-variance plane.

11.3 THE "FOCUSING" PROCEDURE: INDEPENDENT MODEL

The results above have been obtained by generating several MSAs with varying descriptors B, D, and analyzing how these latter control the performance of the inferred independent Potts model. However, this scenario is different from real protein data sets, were we start from a given MSA of the relevant protein family, generated by alignment of sampled sequences, and infer our model from it.

Taking the full MSA as a training set is rarely the optimal choice. In fact, by simply knowing the value of one parameter, i.e. J_0' in the (11.30), we could in principle score any subset of the given MSA and then choose the subset msa^* that minimizes the bias-variance tradeoff:

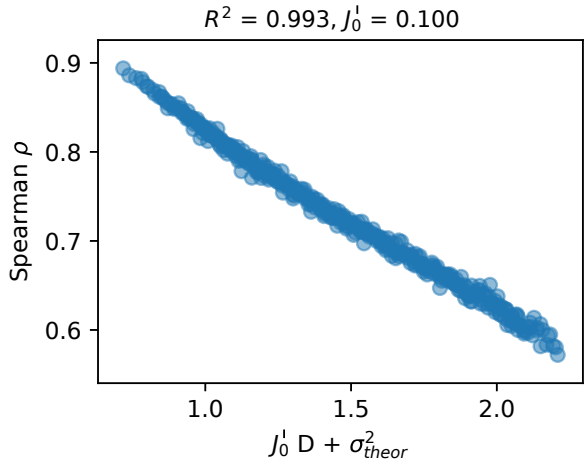


Figure 11.4: **Independent Potts: Scaling law for the inference performance**. **Top:** Comparison between the linear sum of theoretical bias and variance, estimated from the sequence data in the MSA, and the performance of the inferred Potts model. Same data of Fig. 11.2.

$$msa^* = \text{argmin}\left(J_0' \cdot D(msa) + \frac{1}{B(msa)} \cdot \mathcal{V}(msa)\right) \tag{11.31}$$

that is, the one that is predicted to yield the best performance. The number of possible subsets of a given MSA is astronomical, therefore the minimization can not be done by enumeration. Luckily, the definition of our descriptors D and B allows for an iterative scheme that explores an optimal trajectory in the bias-variance diagram: the **"focusing" procedure**.

Starting from a given MSA with descriptors B and D, we remove all the sequences that have an hamming distance from our wildtype greater than a cutoff d_0. Doing so, we reduce the bias, since we are reducing the average distance ($D \uparrow_{d_0}$), and increase the variance, since we are taking into account a smaller number of training sequences ($\frac{1}{B} \downarrow_{d_0}$). For each $d_0 \in [1, N]$ we can estimate the the tradeoff $\widehat{\mu}^2(d_0) + \widehat{\sigma}^2(d_0)$ of the subsampled MSA, and choose the optimal cutoff as the one that minimizes it. This succession of decreasing cutoffs draws a trajectory on the bias-variance diagram. Note that, given a cutoff d_0, there is no subset of the starting MSA that has the same number of sequences $B(d_0)$ and a mean hamming distance smaller than $D(d_0)$. Therefore, the focusing procedure is guaranteed to reach the maximum performance that we can achieve by binary inclusion/exclusion of sequences.

In Fig. 11.7 we show an example of this trajectory, starting from an MSA with $B = 3000$ and $D = 20$, and iteratively decreasing the cutoff from $d_0 = 27$ to $d_0 = 5$. For each cutoff

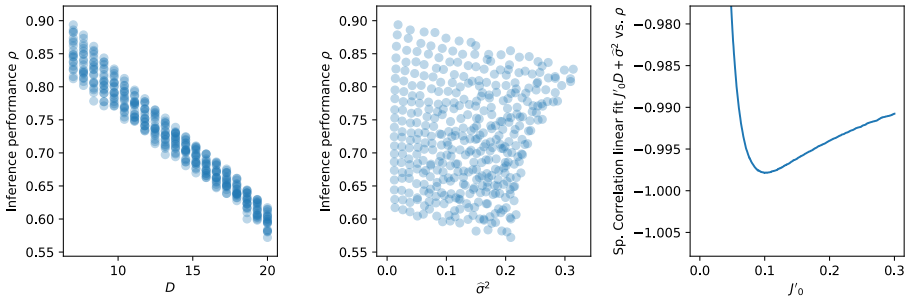


Figure 11.5: **Independent Potts: Scaling law for the inference performance**. Individual contributions of bias and variance to the performance ρ and fit of the parameter J_0'. Same data of Fig. 11.2.

d_0 we draw a point in the bias-variance diagram (in its linear-scaling approximation of Eq. 11.30) by estimating the bias and variance of the resulting MSA from Eq.s 11.27 and 11.20. The resulting trajectory is a curve that starts at the top-left corner, i.e. low variance and high bias, and ends in the low-right one. The predicted optimal cutoff is the one corresponding to the closer point to the low-left region. For each cutoff d_0 we compute the performance ρ of the inferred independent Potts model. As shown in Fig. 11.7, the performance reaches a maximum around $d_0 = 10$ (red dot), which indeed corresponds to the distance that minimizes the tradeoff (blue triangle). We then tested this procedure on $n = 64$ starting MSAs with different B (equally spaced from 500 to 4000) and D (integers from 12 to 20). The result, reported in Fig. 11.8, shows that the optimal performance is reached in almost all cases, as expected.

We finally tested if the tradeoff of (11.31) is predictive for the optimal cutoff for sequences different than the wildtype used to fit J_0' in the scaling law. To do that, we repeated the procedure of Fig. 11.7 for $n = 128$ different sequences, sampled with $p_{nat} > 0.990$ for the same family of the original wildtype. For each sequence we therefore computed the performance ρ_{pred}, at the predicted optimal cutoff by (11.30), the supervised optimal one, ρ_{opt}, and the one obtained from the full MSA, ρ_{nocut}. As reported in Fig. 11.9, this systematic analysis shows that the value of J_0' is indeed predictive for all sequences belonging to the same family.

Therefore, J_0' is, as suggested by the theoretical derivation in (11.26), a feature of the protein family. This is a crucial result, since from the knowledge of only one mutational landscape (e.g. from one mutagenesis experiment), we can retrieve the best value of J_0' by changing systematically the training MSA and testing the scaling law (11.30) as reported in this section, and then use it to predict the optimal MSA for any other sequence in the same family.

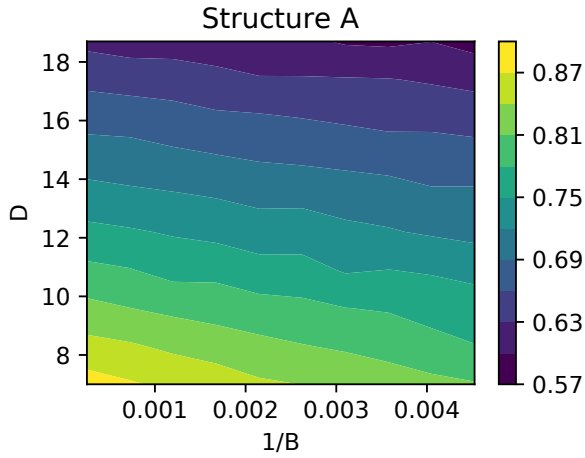


Figure 11.6: **Independent Potts: bias-variance diagram.** performance of the independent Potts inference represented in color scale as a function of $\frac{1}{B}$ and D. As a consequence of the linear scaling law in (11.30) curve levels are almost-parallel straight lines. Same data of Fig. 11.2.

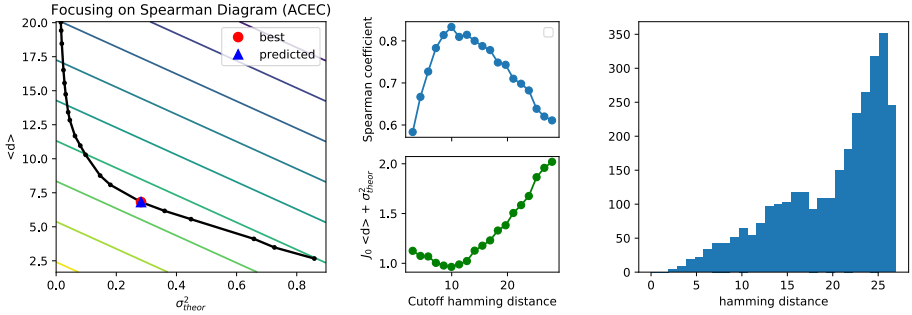


Figure 11.7: **Independent Potts model: focusing on the bias-variance diagram.** Each cutoff d_0 yields a different MSA, whose bias and variance are quantified by the theoretical estimations. This couple of values represents a point in the bias-variance diagram. The succession of points obtained by decreasing cutoff values draws a curve from the top-left to the low-right corner of the diagram. For each cutoff the performance and the bias-variance tradeoff $\mathcal{T}(d_0) = J'_0 \cdot D(d_0) + \frac{1}{B(d_0)} \cdot \mathcal{V}(d_0)$ are computed (central panel). Starting MSA with $B = 3000$, $D = 20$; distribution of hamming distances from the wildtype of the starting MSA is shown in the right panel. Same wildtype used to retrieve J'_0.

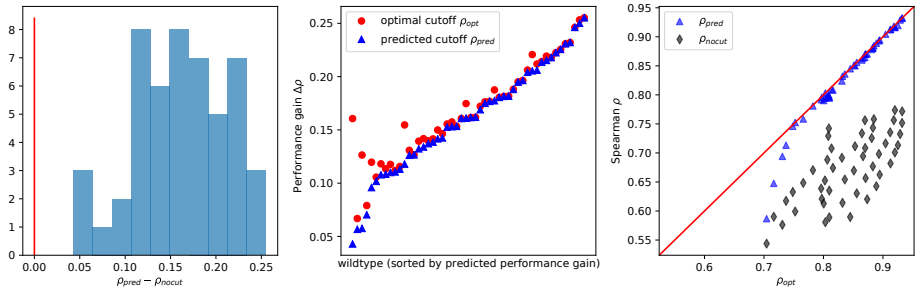


Figure 11.8: **Independent Potts model: systematic analysis of focusing on different MSAs.** Same analysis of Fig. 11.7 performed on $n = 50$ different starting MSA with $B = 500, 1000, 2000, 4000, 8000$, $D = 12, 13, \ldots, 21$, same wildtype used for the fit of J'_0. For each MSA we define the predicted optimal cutoff as the one that minimizes the bias-variance tradeoff $\hat{\mu}^2(d_0) + \hat{\sigma}^2(d_0)$. The performance of the independent Potts model inferred at the predicted optimal cutoff is called ρ_{pred}. This is compared to the performance computed at the supervised optimal cutoff ρ_{opt}. The performance gain $\Delta\rho$ is computed with respect to the performance of the full MSA ρ_{nocut}.

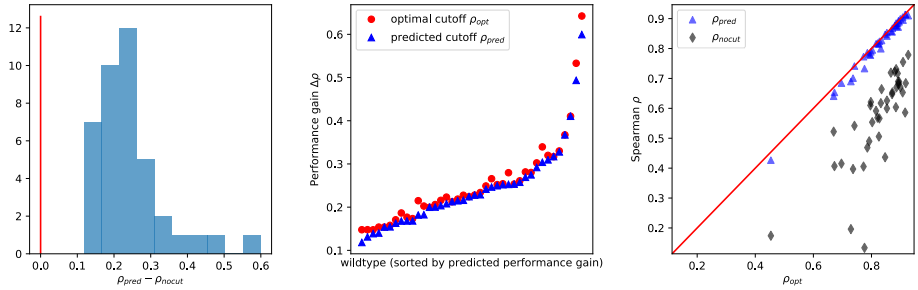


Figure 11.9: **Independent Potts model: systematic analysis of focusing on different wildtypes** Analysis on $n = 40$ different sequences. For each sequence the starting MSA is sampled with $B = 4000$, $D = 18$.

11.4 BIAS-VARIANCE TRADEOFF IN STRUCTURAL-POTTS INFERENCE

The results obtained in the case of the independent Potts model are useful to get an intuition on how the descriptors B and D contribute to the performance of the inference by controlling, respectively, the variance and the bias of the inferred model. However, the standard model for fitness prediction is the *epistatic* Potts model, i.e., the one involving pairwise interactions $J_{ij}(a, b)$ between residues. As we saw in the previous chapter, the fully-connected model is very often over-parametrized with respect to the number of available data points [8], which can be interpreted as a very high variance. Therefore, a fully-connected model will rarely improve its performance by further increasing its

variance, i.e., by focusing the training set around a specific wildtype. This is however not the case for the sparse Structural-Potts model (cmap-ACE) presented in the previous chapter since the number of parameters is sensibly lower than in the fully-connected one. In this section, we will extend the analysis done for the independent model to the cmap-Potts model, i.e., the sparse Potts model whose couplings are inferred only on the graph of interaction defined by the contact map of the folded protein.

Theoretical bias and variance in the pairwise Potts model

Analytical results, in the case of the coupled Potts model, are much more involved with respect to the independent one. The variance can be expressed by adapting the approximated derivation of [47] to a small number of couplings interacting on the contact structure. This approximation can then be checked by comparison with the numerical computation of the variance, that is model-independent, as defined in Eq. 11.18. By defining n_{C_i} as the number of sites in contact with the site i, the variance of the cmap-Potts model reads

$$\widehat{\sigma}^2_{ia} = \frac{1}{B}\left[n_{C_i}\left(\frac{1-p_i(a)}{p_i(a)} + \frac{1-p_i(s_i^{wt})}{p_i(s_i^{wt})}\right) + \sum_{j\in C_i}\left(\frac{1-p_{ij}(a\ s_j^{wt})}{p_{ij}(a\ s_j^{wt})} + \frac{1-p_{ij}(s_i^{wt}s_j^{wt})}{p_{ij}(s_i^{wt}s_j^{wt})}\right)\right] \quad . \tag{11.32}$$

Again, the global variance is defined as per Eq. 11.18. For what concerns the bias, we need to resort to an ansatz, that can be numerically verified a-posteriori. As we saw for the independent model, the pairwise interactions $J_{ij}(a,b)$ of a coupled Potts model have a biasing effect, on single point statistics, which is proportional to the average number of mutations, i.e., to the mean hamming distance from the gauge sequence (the wildtype). Informally, we could extend the argument for higher-order interactions: three-wise couplings will realistically have a similar biasing effect on averages and correlations, and the number of biasing terms in the fitness function would be again proportional (although not linearly) to the mean hamming distance. Moreover, in the present case of a very sparse coupled model, the biasing effect of un-modeled couplings (i.e., the one put to zero from the sparsity prior) will be not so different from the bias described in the independent case. We will therefore assume, as an ansatz, that the linear term is the dominant one:

$$\widehat{\mu}^2 = J_0^{coupled} \cdot D \quad . \tag{11.33}$$

For the sake of notation we will use $J_0 \equiv J_0^{coupled}$.

Numerical vs. theoretical bias and variance

The approximated formula for the variance of the cmap-Potts model in Eq. 11.32 and our assumption for the bias in Eq. 11.33 can be verified by comparing them with their

numerical counterparts. Similarly to what we did for the independent Potts model, we generated several MSAs, changing the descriptors B and D, and computed the numerical and theoretical bias and variance. As shown in Fig. 11.10, the theoretical estimations of bias and variance linearly correlate with their numerical counterparts. Importantly, the variance is estimated with a prefactor ~ 1, which confirms the validity of the approximated form of Eq. 11.32. Therefore, in the case of the coupled model we have $J_0' = \alpha J_0 = J_0$.

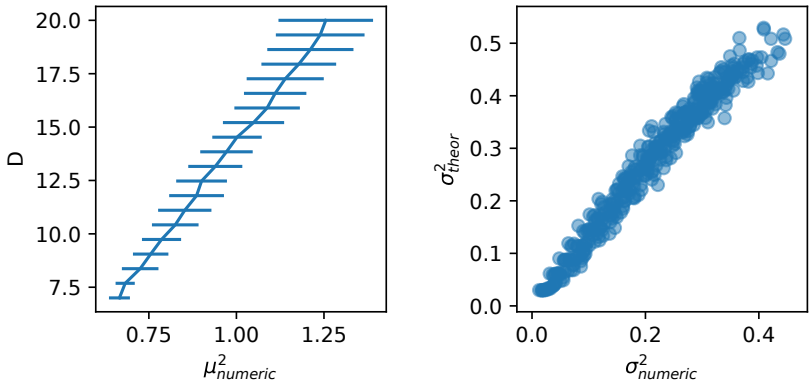


Figure 11.10: **Cmap Potts: Theoretical estimations of bias and variance correlate with the numerical counterparts.** Same procedure used in the analysis of the independent Potts model, see Fig. 11.2. Structure A, $\theta = 5$, 400 points.

We then validated the scaling law, as done in the independent case, by finding the best J_0 that maximizes the correlation between the linear sum of theoretical bias and variance $J_0 D + \widehat{\sigma}^2$ and the performance of the cmap-Potts model ρ (Fig. 11.12). Despite the considerable approximations involved in the theoretical derivations of bias and variance, the scaling law is still confirmed with astonishing precision ($R = -0.991$), see Fig. 11.11.

Drawing the corresponding bias-variance diagram, shown in Fig. 11.13, we again observe fairly-parallel and fairly-straight level curves. Note that, while in the independent case the tradeoff was substantially dominated by the bias, here the variance is a strong factor in determining the performance. This is shown by the steepest curves in the right panel in Fig. 11.13, and from the detailed analysis of the fit of J_0 shown in Fig. 11.12.

11.5 THE "FOCUSING" PROCEDURE: CMAP-ACE POTTS MODEL

We then tested the focusing procedure in the case of the inference on the known interaction graph ($cmap - ACE$ defined in the previous chapter) of the known 28 contacts on

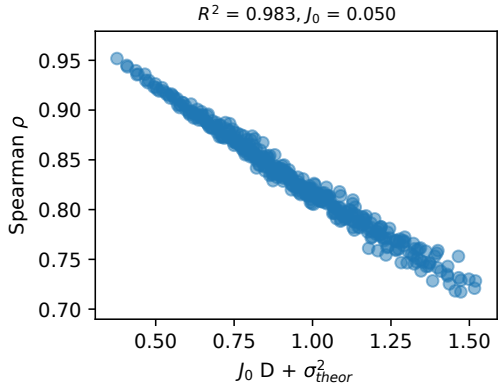


Figure 11.11: **Cmap Potts: Scaling law for the inference performance** Comparison between the theoretically-estimated bias-variance tradeoff and the inference performance. Same values of B and D of Fig. 11.4. Structure A, $\theta = 5$, 400 points.

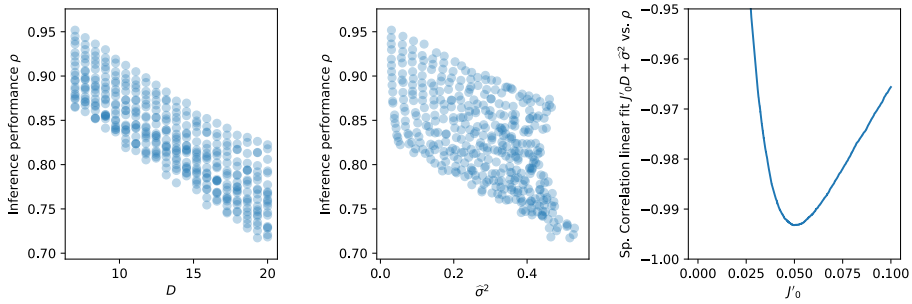


Figure 11.12: **Cmap Potts: Scaling law for the inference performance**. Individual contributions of bias and variance to the performance ρ and fit of the parameter J_0. Same data of Fig. 11.11

the structure. We repeated the analysis that we performed for the independent model in Fig. 11.8, for $n = 40$ different starting MSAs with different B and D. Results, reported in Fig. 11.15, show that the optimal cutoff predicted by minimizing the bias-variance tradeoff retrieves the optimal performance in almost all cases. An example of the trajectory drawn by the focusing procedure on the bias-variance diagram is shown in Fig. 11.14.

We finally tested whether the inferred value of J_0 is again predictive for the best cutoff in the analysis of the mutational landscape of different sequences. We performed the focusing procedure on $n = 40$ new wildtype sequences, with starting MSA with $B = 8000$ and $D = 18$. Results, reported in Fig. 11.16, show that minimizing the bias-variance tradeoff where J_0 has been inferred from the analysis of one wildtype sequence

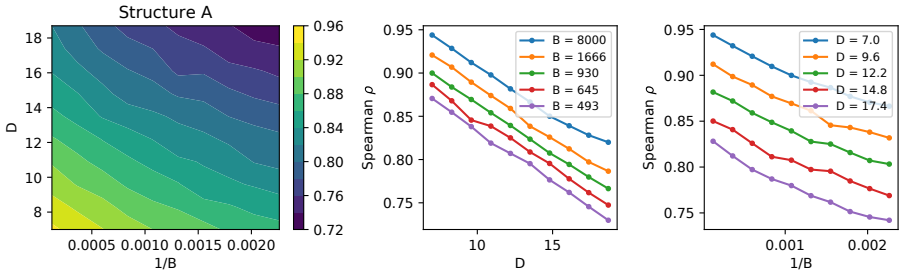


Figure 11.13: **Cmap Potts: bias-variance diagram** Same procedure of Fig. 11.6. Structure A, $\theta = 5$, 400 points.

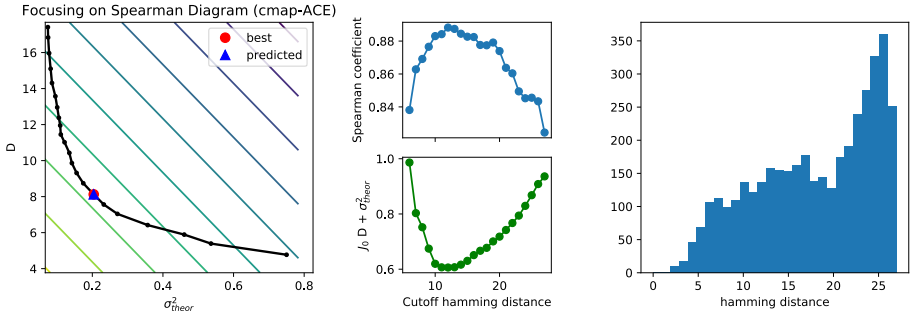


Figure 11.14: **Cmap Potts model: trajectory on the bias-variance diagram** At each cutoff a cmap-Potts model is inferred, its performance is tested and compare with the computed bias-variance tradeoff. The predicted cutoff, i.e. the one that minimizes the tradeoff (blue triangle), is indeed the optimal one (red dot). Starting MSA with $B = 4000$, $D = 18$.

is predictive for the best cutoff for any sequence belonging to the same family, as we saw for the independent Potts case.

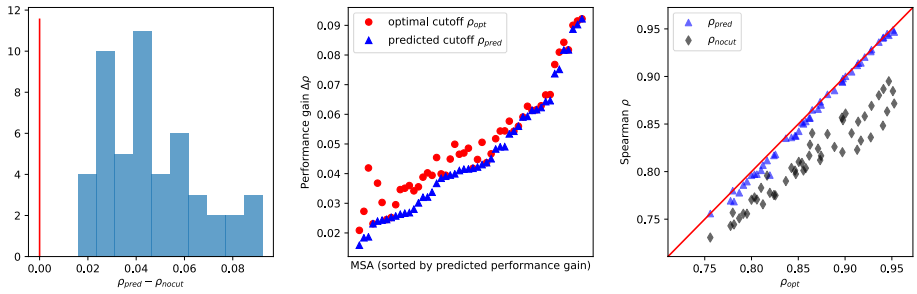


Figure 11.15: **Cmap Potts model: systematic analysis of focusing on diverse MSAs.** Focusing performed on $n = 50$ different starting MSA with $B = 500, 1000, 2000, 4000, 8000$, $D = 12, 13, \ldots, 21$, same wildtype used for the fit of J_0. For each MSA we define the predicted optimal cutoff as the one that minimizes the bias-variance tradeoff $\widehat{\mu}^2(d_0) + \widehat{\sigma}^2(d_0)$. The performance of the independent Potts model inferred at the predicted optimal cutoff is called ρ_{pred}. This is compared to the performance computed at the supervised optimal cutoff ρ_{opt}. The performance gain $\Delta\rho$ is computed with respect to the performance of the full MSA ρ_{nocut}.

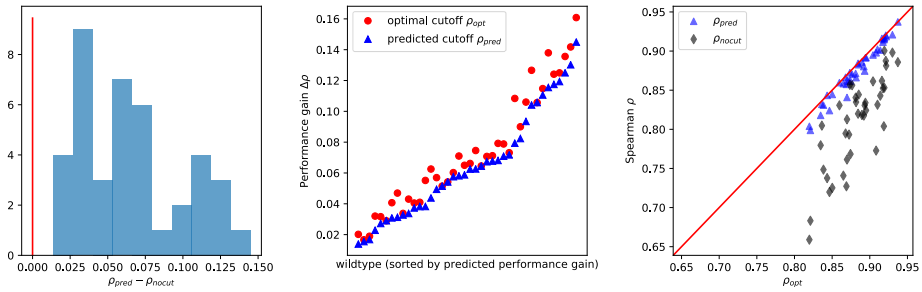


Figure 11.16: **Cmap Potts model: systematic analysis of focusing on different wildtype sequences.** Analysis on mutational landscape of 25 different sequences from the same the family (A) of the original wildtype used for the analysis above. For each sequence, a starting MSA with $D = 18$ and $B = 8000$ is generated, and the procedure of Fig. 11.14 is repeated to compare the predicted, optimal, and full-MSA performances.

11.6 SCALING LAW IN REAL PROTEIN DATASETS

We here report some early results obtained in collaboration with Francesca Rizzato on the four protein datasets analyzed in the previous chapter. We took the initial MSA and mutagenesis fitness experiments used in [44] to test how the prediction Spearman coefficient of the independent Potts model scales with the descriptors B, D. We varied the mean hamming distance D from $0.4 \times N$ to N (number of sites), and the number of sequences from $B = 100$ to the full starting MSA. Results, reported in Fig. 11.17, show a striking confirmation of the scaling law in all the four cases (PDZ: $\rho = -0.90$, RNA-bind: $\rho = -0.94$, WW: $\rho = -0.79$, BLAT: $\rho = -0.94$). The study of bias and variance on real protein datasets is a work in progress and will be included in a future publication [7].

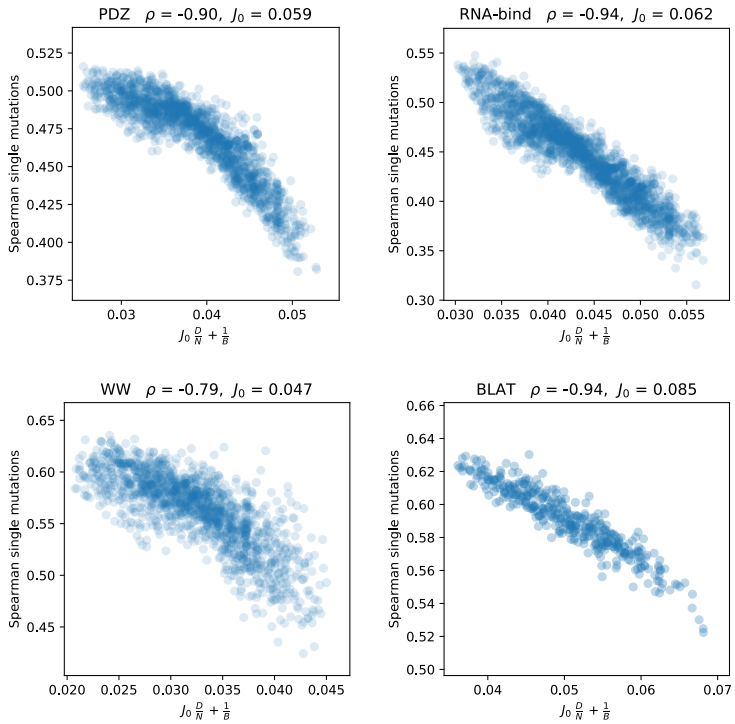


Figure 11.17: **Bias-variance Scaling law in real protein data sets to predict mutagenesis experiments** The performance ρ of the inferred independent Potts model on four protein data sets, compared to the linear composition of two terms: the scaling of the variance $1/B$ and the mean Hamming distance from the wildtype, normalized on the number of protein sites N. Each point corresponds to a different training MSA. The scaling parameter J_0 is retrieved by minimizing the Spearman correlation between the linear composition of bias-variance terms and the performance ρ.

11.7 DISCUSSION

In this work, we investigated how the training sample data (MSA) can be optimally chosen to improve the performance of max-entropy models aimed at retrieving the local (single-point) mutational landscape of a specific wildtype. We framed the problem in terms of bias and variance of the estimated parameters and tested our reasoning on in-silico data generated from a Lattice-Protein model. We showed that the performance of the estimated model scales with the linear sum of bias and variance, both for the independent-Potts and the sparse-Potts model (cmap-ACE, introduced in the previous chapter). We then provided a theoretical argument for the proportional relationship between the bias of the inferred model and the mean hamming distance of sequences in the MSA from the wildtype, called D. Similarly, the variance is known from previous works [47] to scale with the inverse number of sequences, i.e., B^{-1}. We provided evidence of the performance of the inferred model to scale linearly with the two descriptors, both in the independent and in the cmap-Potts model.

Together, these results suggest that the predictive power of the inferred max-entropy model can be deduced by simple observables computable in linear/polynomial time from the training MSA, i.e., the mean Hamming distance D and the variance $\sigma \propto B^{-1}$. We finally introduced a method, called "focusing" that yields the optimal subset of a given MSA, demonstrating its efficacy and generality (generalization to different wildtypes) in both the independent and the cmap-Potts model on in-silico data.

We provided some early evidence of the same scaling law (performance as a function of a linear composition of estimated bias and variance) on four real protein datasets. These results are part of a more comprehensive analysis performed in our group, in parallel to the work presented here, by Francesca Rizzato. Results of in-silico and real data analysis will be published alongside in a future paper [7].

The application of a pairwise model on real data poses a series of challenging problems, such as the strongly-limited amount of data and a non-equilibrium biased sample of sequences in the training MSA [235, 236]. While coupled epistatic models have been demonstrated to provide superior predictions in several cases, some mutagenesis experiments were best retrieved by an independent model [44]. While we might be confident in forecasting a bright future of abundant data for bioinformatic analyses, at the present moment there still are strong limitations due to the low amount of training data with respect of the complexity of the model [8]. Therefore, the independent-site Potts model is still, in some cases, the best compromise.

The suitability of max entropy models to provide biological predictions such as the effects of single mutations is still an object of alive discussion in the statistical physics and bioinformatics community [235, 236]. For this reason, we think that our results provide a step forward in the understanding of which factors determine the performance of these methods, approaching the problem by both theoretical and numerical analysis, as well as providing practical prescriptions to improve the accuracy of the predictions.

Part IV

CONCLUSIONS

12

DISCUSSION AND PERSPECTIVES

12.1 OUTLINE

This thesis presents results obtained on two diverse biological systems by a complementary approach that includes both top-down modelling (abstraction to observables) and bottom-up statistical-inference methods (observables to abstraction). This bidirectional approach has, as we think the diversity of systems studied here might suggest, a wide range of applicability.

Our modelling approach is based on sampling equilibrium configurations from the Boltzmann-Gibbs measure of statistical-physics systems. To model the neural activity encoding for self-location during navigation within a memorized cognitive map we employed a continuous attractor neural network (CANN) [12, 154, 155]. We proposed a CANN subject to inputs from the external world and path integration, and showed that a conflict between the two positionally-coherent sources of input could explain the metastable "flickering" oscillations reported in the CA3 region by Jezek et al. in the "teleportation" experiment [143] (Chapter 6). To model the fitness landscape of proteins, we employed the so-called "Lattice Proteins", a model for the structural energy and competition between different structures in protein folding [238, 239]. Due to the competition terms between folding structures, Lattice Proteins are complex enough to reproduce non-trivial feature of real proteins, while at the same time being treatable by analytical and numerical means (Chapter 9). We leveraged the controllability of Lattice Proteins to control the sampling conditions of multiple-sequence alignments (MSA) used to train an inference model aimed at the retrieval of the fitness landscape of a specific protein (wildtype).

Our modelling was complemented with the application of Bayesian-inference methods that allowed us to infer a statistical description of experimental and synthetic data. In Chapter 5 and 4, we showed how the expressed attractor state in the hippocampal network can be decoded, from neural activity, by the application of the inverse Ising model to experimental and simulated neural recordings. We then applied the Ising decoder to real CA3 data from the teleportation experiment. Since the Ising decoder, contrarily to the standard methods, does not rely on the position of the animal to retrieve the cognitive map, we were able to decode the precision of the positional representation and the flickering oscillations of the cognitive state independently. We, therefore, were able to show that the position of the animal is coherently represented even during fast oscillations of the contextual variable, with a precision that is slightly (yet measurably) affected by the mismatch between the two positional inputs, as predicted by the CANN model.

In chapters 10 and 11, we showed that the local fitness landscape of a protein could be retrieved by applying the inverse Potts model to a collection of sequences sampled from the relevant family. We investigated how prior information regarding structural features of the protein could be used to enforce sparsity on the most relevant variables. The adaptive cluster expansion of the Potts model (ACE) revealed as a natural method to retrieve the most informative parameters in an unsupervised way, thanks to its iterative procedure that discards those clusters that less contribute to the log-likelihood of the model. We then investigated the factors that primarily determine the precision of the inferred Potts model, as it represents a low-order approximation of the "real" rough fitness landscape, in the task of retrieving the local mutational landscape of a specific wildtype. We showed that we could improve the predictive power of the inferred model by giving up its generality and "focusing" on the local region of the sequence space that surrounds the wildtype.

12.2 METHODOLOGY AND FUTURE RESEARCH

The application of statistical physics modelling and max-entropy inference to the complex phenomenology of biological systems raises several methodological questions, which range from the strong simplifications made in theoretical models to the arbitrariness of means and correlations as constraints included in the max-entropy inference. The founding roots of the physicist' approach to foreign fields, including but not limited to biology, have been thoroughly discussed in the literature, and it would be an utter simplification to attempt here a comprehensive outline of a discussion that spans several decades of scientific research. The reader who is interested in deep methodological considerations is invited to the seminal books of Amit [153] and Jaynes [17], both physicists by formation and fathers of, respectively, the theory of attractor neural network and the max-entropy argument in Bayesian inference. At the same time, we think that there are a few points, undoubtedly less profound and more practical, that deserve to be mentioned, especially concerning how the assumptions that underlie our inference approach relate to the specific biological systems investigated here.

One of the most arguable features of the Ising model inferred from activities of a neural population is that it does not explicitly account for a dynamics in time, assuming that each observed neural pattern is independently sampled from equilibrium conditions. In fact, a time-shuffle of the neural patterns used to train the inference would leave the inferred model unchanged. One way to account for this unrealistic equilibrium assumption is to be careful in choosing the time bin used in the discretization procedure of the raw spiking data. For example, in our work on CA3 recordings, we used a discretization in theta cycles. The theta rhythm is thought to represent a "refresh time" of neural activity; therefore each theta cycle could be though, in a first approximation, as a natural binning time. In the benchmark on CA1 data (Chapter 5) we indeed showed that the Ising decoder loses some performance points in favor of a rate-based model on very short timescales ($\lesssim$ 20 ms, see Fig. 5.3).

Despite the lack of explicit time dependence, the Metropolis sampling of the Ising Gibbs-Boltzmann distribution provides a sort of coherence in time due to the Markov-chain exploration of the energy landscape of the model. Interestingly, in some recent results concerning the analysis presented in Chapter 7, and not reported in this manuscript, we observed that the Ising model inferred from hippocampal activity and used to generate new patterns can reproduce a trajectory that is coherent in time, in the sense that it reproduces a diffusive process on the chart corresponding to the cognitive map. These results are in agreement with recent findings obtained by F. Stella [254] in the rodent hippocampus, on activities sampled in sharp-wave ripples during sleep. Neural patterns sampled during sleep are more realistically comparable to an "equilibrium-like" spontaneous sampling of the energy landscape defined by the post-learning neural connectivity. A future line of research, therefore, will be to directly assess these results on similar "spontaneous" data.

The max-entropy graphical Ising model, in our view, represents a well-grounded and well-interpretable approach to retrieve the direct functional connectivity between neurons. Today, it is a vivid research field that includes development of new methods [255] as well as application to retinal [51–53], cortical [173, 174, 256] and hippocampal [1, 2, 257] regions. The retrieved functional network encodes both for structural constraints as well as for statistical properties induced by the collective dynamics of the neural population. Therefore, it would be interesting to assess whether it could discriminate between health and condition in diseases that are related to a dysfunction of the network behavior of the neural population, such as Alzheimer disease, schizophrenia, and epilepsy.

As discussed in Chapter 11, the fact that the rugged high-order fitness landscape of a protein could be described by a low-order (pairwise) Potts model is not trivial. In facts, despite outperforming previous methods [44, 227, 234], performances of max-entropy models are far from being perfect. A possible approach would be to include higher order terms in the inferred model. However, this generalization raises the number of parameters dramatically, and several approaches have been proposed to overcome this problem by, for example, enforcing a sparse high-order connectivity or by machine learning techniques that unsupervisedly retrieve a sparse representation, such as restricted Boltzmann machines [258]. This last line of research has been carried in our group by J. Tubiana and has been proven capable of predicting meaningful biological features on protein domains Kunitz and WW and a chaperone protein Hsp70 [259]. Following a complementary direction, we showed how the precision of the predictions for single-point mutations could be improved by giving up generality and focusing on a local region of the landscape around the wildtype. In order to predict the optimal focusing threshold, we fitted a parameter J_0 that encodes for the high order un-modelled biasing terms of the real fitness function. A possible future generalization of the method could be to retrieve the value of this parameter by comparing multiple mutagenesis experiments around proteins from the same family. By computing how the local mutational landscapes de-correlate with the Hamming distance between the different wildtypes, we can estimate the magnitude/variance of high order terms, possibly integrating this information in the focusing procedure.

Part V

APPENDIX

APPENDIX - CHAPTER 6

A

A.1 EFFECTIVE TWO-STATE MODEL FOR HIPPOCAMPAL CANN ACTIVITY

We show below how the two-state model pictured in Main Text, Fig. 3B, can be derived from the definition of the microscopic CANN model, see Main Text, Methods. The dynamical evolution of the CANN ensures that the log probability of a configuration of activity $\mathbf{s} = \{s_i\}$ is given by [158]

$$L(\mathbf{s}) = \sum_{i<j} J_{ij}\, s_i\, s_j + \sum_i \left(h_i^V(\mathbf{r}) + h_i^{PI}(\mathbf{r})\right) s_i + L^* \ , \tag{A.1}$$

where $\mathbf{r}$ if the rodent position and L^* is a constant term such that the sum of the probabilities e^L over all 2^N configurations $\mathbf{s}$ is normalized to unity.

For simplicity, we consider that the bump of activity consists of $a \times N$ active neurons $s_i = 1$ with place-field centers as close as possible in a map, say, $m = A$, where a is the fraction of active neurons in any time bin. This corresponds to the limiting case of zero neural noise, $\beta \to \infty$ [160]; calculation of effective potentials at finite β is much more involved and requires the use of sophisticated statistical physics techniques able to take into account the fluctuations of neural activities, see [12].

Let us define r_{bump} as the radius of the bump, i.e. the maximal distance in environment m between the rodent position $\mathbf{r}$ and the place-field centers r_i^m of active neurons. We have

$$a\, N = \frac{\pi\, r_{max}^2}{\delta^2} \ , \tag{A.2}$$

where δ^2 is the elementary portion of surface per place cell, defined as the total area of the environment, L^2, over the number of place cells, N. We thus obtain the expression of the bump radius as a function of the activity,

$$r_{max}(a) = L\sqrt{\frac{a}{\pi}} \ . \tag{A.3}$$

Contributions to log-likelihood due to inputs.

We assume that the CANN has activity localized in map m, and that the whole system is in the conflicting phase, with $PI = A$ and $V = B$. The contributions to L due to the visual (V) and path-integrator (PI) inputs reads, up to quadratic terms in a,

$$L_{input} = \gamma \sum_i s_i\, \phi\left(\mathbf{r}_i^M - \mathbf{r}\right) \ , \tag{A.4}$$

where $\gamma = \gamma_{PI}$ if $m = A$ and $\gamma = \gamma_V$ if $m = B$. According to the definition of the bump radius r_{max}, we have

$$L_{input} = \gamma \int_0^{r_{max}} \frac{d\mathbf{r}}{\delta^2}\, \phi(\mathbf{r}) = \frac{\gamma}{\sigma^2} \int_0^{r_{max}} dr\, r\, e^{-r^2/(2\sigma^2)} = \gamma \left(1 - e^{-r_{max}^2/(2\sigma^2)}\right) . \tag{A.5}$$

Contributions to log-likelihood due to recurrent connections.

We now consider the contribution $L_{recurrent}$ to the log-likelihood coming from the recurrent connection in the CANN. The coupling J_{ij} between neurons i and j is the sum of one interaction specific to map A and another one specific to map B, see Eqn. (9) in Main Text, Methods. Assuming again that the bump of activity is localized in map $m = A$, we neglect the contribution to L due to the interaction specific to map B. This simplifying approximation amounts to an error of the order of a^2, see [160] for more details. We obtain

$$\begin{aligned} L_{recurrent} &= \frac{1}{2}\,\gamma_J \int_0^{r_{max}} \frac{d\mathbf{r}}{\delta^2} \int_0^{r_{max}} \frac{d\mathbf{r}'}{\delta^2}\, \phi(\mathbf{r} - \mathbf{r}') = \frac{\gamma_J\, N}{4\pi\, L^2\, \sigma^2} \int_0^{r_{max}} \int_0^{r_{max}} d\mathbf{r}\, d\mathbf{r}'\, e^{-(\mathbf{r}-\mathbf{r}')^2/(2\sigma^2)} \\ &= \gamma_J\, \frac{2\pi\, N\, \sigma^2}{L^2} \int_0^{\left(\frac{r_{max}}{\sigma}\right)^2} du\, I_0(u) \int_u^{\left(\frac{r_{max}}{\sigma}\right)^2} dv\, \frac{e^{-v}}{\sqrt{\left(\frac{v}{u}\right)^2 - 1}} , \end{aligned} \tag{A.6}$$

where I_0 is the first kind modified Bessel function of zero order.

Case of mixed state.

Assume now that a fraction α of the bump is localized in map $m = A$ and the remaining fraction, $1 - \alpha$, is localized in map B. The log-likelihood of this mixed state is obtained by summing the expressions of the log-likelihood in state A above with activity $a \to \alpha\, a$ and of the log-likelihood in state B above with activity $a \to (1 - \alpha)\, a$. The result is shown in Supplementary Fig. A. As indicated in Main Text, the amplitude γ_J of the recurrent connections controls the depth of the well separating the two complete bump states (all A or all B), while the ratio γ_{PI}/γ_V controls the asymmetry of the log-likelihood profile and favors one of the two states.

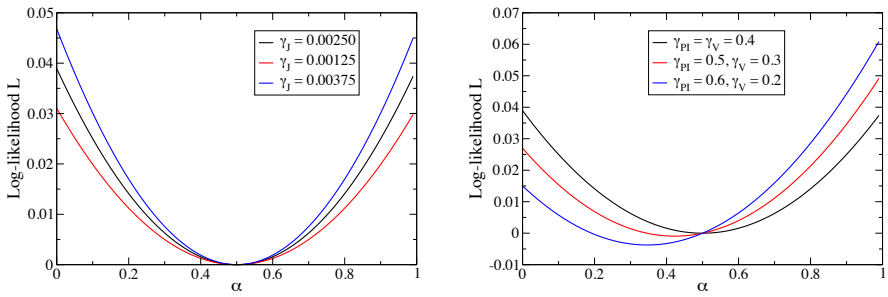
A.2 EFFECTS OF PARAMETERS ON THE MODEL PROPERTIES

The CANN model is defined up to a set of parameters:

(a) the level of neural noise in the simulated activity, β; higher β corresponding to lower noise. This parameter is formally equivalent to the inverse temperature in the Monte Carlo simulation;

(b) the strength of the recurrent connectivity, γ_J;

(c) the strength of the two inputs, γ_V and γ_{PI};



Supplementary Figure A. Log-likelihood of a mixed state with a fraction α of bump in state A and a fraction $1-\alpha$ in state B. An additive constant, independent of α, is introduced such that $L=0$ for $\alpha=0.5$. Top: symmetric case $\gamma_{PI}=\gamma_V=0.4$ for three values of γ_J, showing how the intensity of recurrent connections control the depth in log-likelihood of the mixed state. Bottom: Asymmetric case with $\gamma_J=0.0025$ and for three values of $\gamma_{PI}>\gamma_V$. In all cases, $\sigma/L=0.125$, $a=0.1$.

(d) the spread of place fields and positional inputs, σ;

(e) the number of neurons, N.

(f) the mean activity (fraction of active neurons at any time), a.

A fully-detailed analysis of the response of the system to the each of these parameters is beyond the scope of this paper, and previous works have fully characterized the behavior of the model in the absence of positional inputs [12, 158, 160]. Hereafter, we show how some of these parameters control the dynamical properties of the *flickering* of the cognitive map and the *ability to navigate*, i.e. the correct positioning of the bump of activity in the position defined by the V and PI inputs. These two quantities are indeed observable in the CA3 electrophysiology data, through the map and position-decoding analysis. A characterization of their parametric dependence in the model is therefore a necessary step to a correct quantitative modelling.

For this reason, we will here divide the parameters into two classes:

- the *structural* parameters, N, σ, a. The number of neurons N was varied from a few hundreds to a few thoousands in simulations. To keep the contributions to the total in put $H_{i,t}$ acting on neuron i at time t independent of N, we scale the recurrent connection strength γ_J as $1/N$, see Eqn. (9) in Main Text. This ensures that the sum of local inputs over all active neurons due to these connections has a finite, fixed value as N grows. This is why we will compare below the value of $\gamma_J \times N$ to the other input strengths, γ_V and γ_{PI}. In addition, we have fixed the average linear size of place fields to $\sigma/L \sim 0.125$, which sets the average area occupied by a place field to $2\pi(\sigma/L)^2 \simeq 10\%$ of the environment total area, a value comparable to experimental findings [260]. The average activity (in a time bin) was fixed to

$a = 10\%$ throughout our simulations to match the values fixed in previous works focusing on the same model in the absence of inputs, see discussions in [12, 160].

- the *control* parameters (γ_J, γ_V, γ_{PI}, β), that have a predictable influence on the behaviors we are interested in. Note that the four control parameters are redundant, as the properties of the model depend only on $(\beta \times \gamma_J, \beta \times \gamma_{PI}, \beta \times \gamma_V)$; we may therefore fix one of them and let the other three vary. We now study how the model properties depend on the values of these parameters.

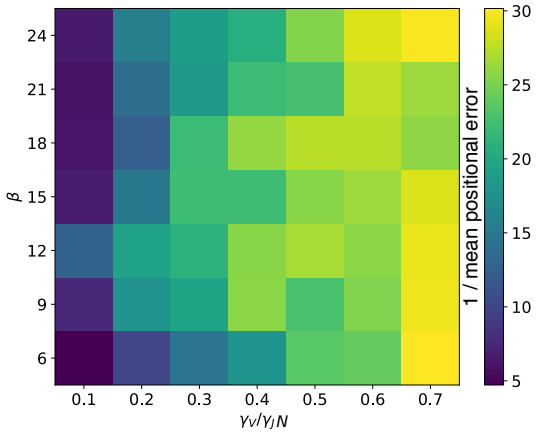
Navigation of the environment

The model is explicitly designed to mimic the representation of self-location in the hippocampal network under the influence of positional inputs. A natural question is how the values of parameters influence the capability of the model to actually represent the correct position in a single map, that is, the correct centering of the neural bump around the input position. Consider the case of coherent inputs at a certain time t, i.e. PI and V point to the same position $\mathbf{r}_t$ in the same map , and let us assume that the bump is correctly centered around $\mathbf{r}_t$. As the input position changes in the next simulated time bin $t + \Delta t$, PI and V will try to activate place cells corresponding to a shifted location, effectively pushing the bump to $\mathbf{r}_{t+\Delta t}$. If the positional input is too weak compared to the recurrent network connections are too strong, the bump will fail to update to the new position, being trapped by the strong connection with the active cells at position $\mathbf{r}_t$. Similarly, a very high value of β, i.e. a low neural noise, would have the effect of enhancing the roughness of the energy landscape, in the positional space, and of trapping the bump and impairing its motion. As a consequence, the model would lose the ability to correctly navigate the environment. Conversely, a very low value of β would result in the inability of the model to condensate the bump of activity [12], therefore losing any notion of represented position.

The inverse of the mean positional error ϵ_t can be used as a proxy for the navigation ability, and is shown in Supplementary Fig. B as a function of β and of the relative strength $\gamma_V/(N\gamma_J)$ (in the balanced case $\gamma_V = \gamma_{PI}$). The navigable region (yellow) has a triangular shape that widens with higher values of the input strength, meaning that the temperature has to be fine tuned for low values of $\gamma_{V/PI}$, while it can take a wider set of values in the presence of strong inputs.

Flickering of the cognitive map

As discussed above, the system acts as an effective two-state model when the two inputs are put into conflict, i.e. point to the same position in different cognitive maps. The transition of the bump from one map to the other happens stochastically, and its dynamical properties are controlled by the parameters of the model. Characterizing this dynamics in the simulated test experiment is rather involved, since the positional inputs move at a variable speed (we use the recorded trajectory of the real rat as input) and a



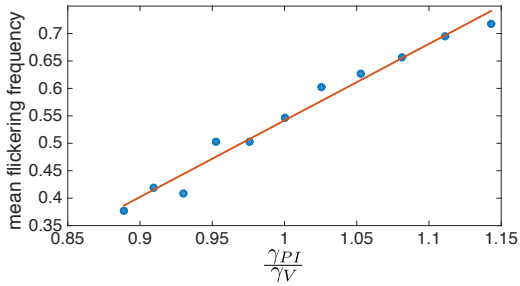
Supplementary Figure B. Navigation of the environment. Dependence of the inverse mean positional error (in cm^{-1}) on the control parameters: β and the relative strength between network connectivity and positional inputs. $\gamma_V = \gamma_{PI}$, $N = 400$, $a = 0.1$.

fast change of the positional input can facilitate the evaporation of the bump from one map, increasing the transition rate between maps. We here analyze the dependence of two data-testable quantities on the parameters. The first is the statistics of permanence in the visual-cue associated map or in the PI-associated map during the conflicting phase, as a function of the relative strength between the two inputs, shown in Supplementary Fig. C. We see that a ration γ_V / γ_{PI} close to 1 results in a mean fraction of flickers (MFF) close to 0.55. This value, slightly different from the expected 0.5 is due to the inertia of the bump that, for few bins after the teleportation, tends to stay in the PI-associated map. Since each simulation is carried for a finite number of time bins after the teleportation (600), this discrepancy is explained as a consequence of the finite-time simulated for each trial.

Next we analyze a dynamical quantity, i.e. the mean sojourn time of the activity in one of the two maps, given a balanced value of $\gamma_{PI} = \gamma_V$, see Methods for the definition of the sojourn time. This quantity is directly proportional to the height of the barrier described in the two-state approximation, which is controlled by the network connectivity strength γ_J and the parameter β. High noise (small β) or weak connections (low γ_J) is expected to enhance the probabiity of crossing the barrier easy, and to make the sojourn times low. This statement is confirmed by the results shown in the diagram in Supplementary Fig. D.

Putting together the results reported in Supplementary Figs. B and D, we see that the model reproduces the dynamical properties of the observed data, while at the same time keeping an accurate representation of the input position, for a range of parameters in

the center of the diagrams. In particular, we have chosen, for the simulations reported in the Main Text, $\beta = 15$ and $\gamma_V = \gamma_{PI} = 0.4$, with no need for fine tuning these two parameters. Indeed any choice in the range $\beta \in [15, 30]$ and $\gamma_V = \gamma_{PI} \in [0.3, 0.6]$ would qualitatively reproduce the behavior observed in data in terms of flickering of the cognitive map and precision in the positional encoding.



Supplementary Figure C. Dependence of the mean frequency of flickers (MFF) in simulated data upon model parameters. The MFF is defined as the fraction of time bins, during the conflicting phase, in which the hippocampal representation m differs from light cues. Simulations were performed using the real trajectory of the rat, with $N = 400$ neurons, $\gamma_W \sim \infty$ to hold the conflicting state to a fixed amount of time bins (600), $\beta = 15$.

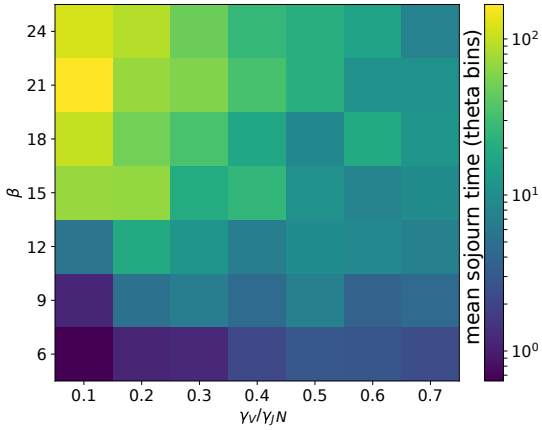
A.3 RELATIONSHIP BETWEEN SOJOURN TIME AND CORRELATION TIME

The correlation time τ_0 (see Main Text Methods and Main Text Fig. 2B & 4C) is related to the sojourn times of the neural bump in the cognitive maps, defined as a sequence of contiguous theta bins that are all decoded in the same map. The relationship between correlation and sojourn time can be established by assuming a Markovian dynamics for the 2-state model ($m = A$ or $m = B$) evolving in discrete time. The dynamics is determined by the map transition probabilities from one time bin to the next:

$$\begin{cases} p_{A\to A} = e^{-1/\tau_A} \\ p_{A\to B} = 1 - e^{-1/\tau_A} \\ p_{B\to B} = e^{-1/\tau_B} \\ p_{B\to A} = 1 - e^{-1/\tau_B} \end{cases} \tag{A.7}$$

where τ_A and τ_B are the mean sojourn times in, respectively, map A and B. A straightforward calculation shows that the time correlation $C(\tau)$ between the map state at times t and $t + \tau$ is decreases exponentially with the delay τ only, with an average time equal to

$$\tau_0 = -\left[\log\left(e^{-1/\tau_A} + e^{-1/\tau_B} - 1\right)\right]^{-1}. \tag{A.8}$$

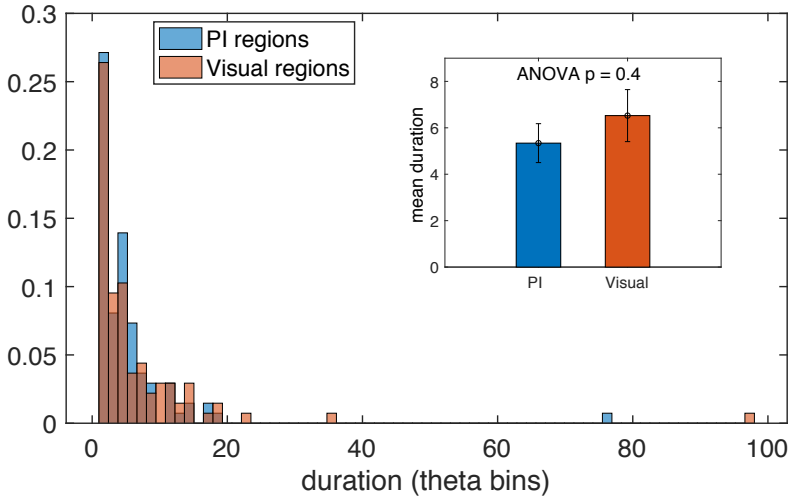


Supplementary Figure D. Flickering behavior: Dependence of the mean sojourn time, defined as the number of consecutive theta bin in which the bump is condensed in the same map, from the control parameters: β and the relative strength between network connectivity and positional inputs. $\gamma_V = \gamma_{PI}$, $N = 400$, $a = 0.1$.

Hence, the correlation time τ_0 is approximately given by the smaller mean sojourn time among τ_A and τ_B. The distribution of sojourn times in each map for experimental CA3 data [143] is shown in Supplementary Fig. E.

A.4 INFERENCE OF PATH-INTEGRATOR REALIGNMENT TIMES - DISCUSSION ON PARAMETERS p_0 AND p_e

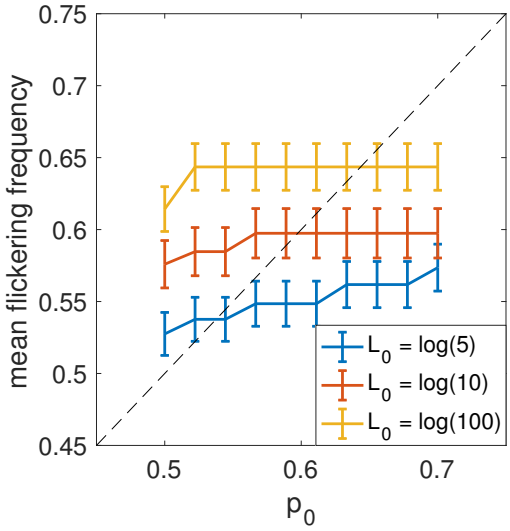
To identify the realignment times of the PI we first introduce a simple probabilistic model for the hippocampal representation to be incoherent with the light-cue conditions (flickering time bin) as a function of time elapsed after the switch (see Main Text Methods). This procedure needs an input value for p_0, the probability of flickering during the conflicting phase, whose consistency can be checked a posteriori by computing the mean flickering frequency in the conflicting phase. In Supplementary Fig. F we show that the a-posteriori average flickering frequency remains remarkably stable, around the self-consistent choice ~ 0.6, for any input value of p_0, with a slight dependence on the chosen L_0 threshold. The same value is observed in the model when the strength of PI and V projections are set to similar values ($\gamma_{PI}/\gamma_V \in [0.95, 1.05]$), see Supplementary Fig. C.



Supplementary Figure E. Sojourn times of hippocampal activity in the two cognitive maps during the conflicting phase for CA3 recordings. Regions of consecutive theta bins whose decoded representation disagree with the external light conditions are marked as "PI" regions. Vicevesa, if they agree with light conditions, they are marked as "Visual". Results obtained after application of our map decoder to the recorded CA3 data of [143]. As shown in the bar plot and from the ANOVA comparison, the permanence times in the two maps during the conflicting phase have roughly the same distribution. Results obtained with map-decoding threshold $L_0 = 2.3$.

A.5 INDEPENDENCE OF FREQUENCY OF FLICKERS FROM DELAY AFTER LIGHT SWITCH: PARAMETERS p_0 AND p_e AND L_0

As pointed in the main text, the constant flickering frequency hypothesis $H_{constant}$ is extremely more likely ($\Delta\ell \sim 150$) than the decaying model H_{decay}. The result is robust against changes in the parameters, see Supplementary Fig. F. For instance, we obtain $\Delta\ell \simeq 170$ and $\Delta\ell \simeq 60$ when the flickering identification is done based on, respectively, a less ($L_0 = \log 2$) and more ($L_0 = \log 100$) restrictive criterion. Similarly, the log-likelihood difference between the two hypothesis remains very large and positive if we change the constant-rate model p_0 value, e.g. $\Delta\ell \simeq 160$ for $p_0 = 0.4$, $\Delta\ell \simeq 120, 33$ for $p_e = 0.1, 0.001$, or if we extend the definition of H_{decay} up to 30 seconds after the switch (instead of 15 seconds): $\Delta\ell \simeq 125$. In many teleportation events the flickering-dense area is indeed too short (too few flickers) or too long (too many flickers occurring far away the teleportation time) to be explained by the H_{decay} hypothesis, see Supplementary Fig. G.

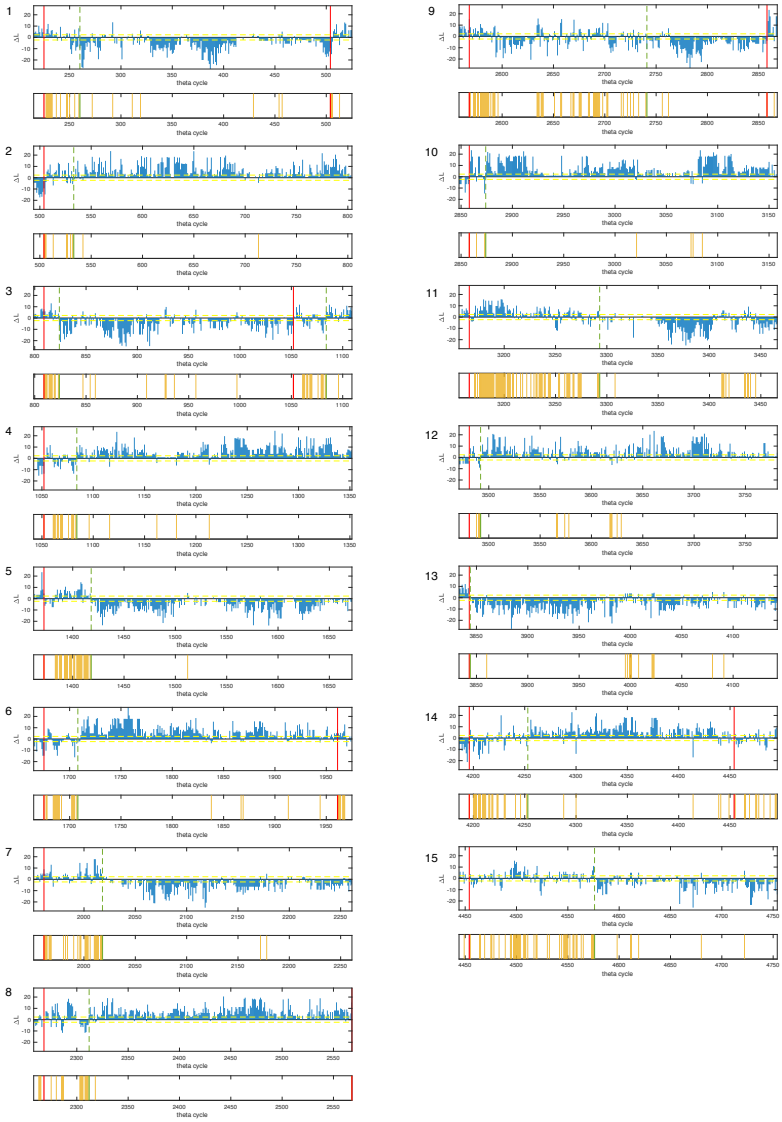


Supplementary Figure F. Dependence of mean frequency of flickers in recorded data upon threshold L_0 and upon the mean frequency p_0 used in the inference of PI-realignment time. A self-consistent value of p_0 is identified, for each value of map-decoding statistical threshold L_0, as the intersection between the corresponding curve and the $y = x$ line (Methods). Simulations performed with the same parameters described in Fig. 2, main text.

A.6 ASSESSMENT OF PERFORMANCES OF MAP DECODER

Our map decoder (based on the inference of an Ising model for each cognitive map) does not use any information about the current rodent position. Its performance can be assessed against the correlation-based decoders used in [143] (which compares the activity $\mathbf{s}_t$ to the expected activities in maps A and B at the given rat position) by means of the classical binary-classifier theory [191–194]. The Ising model was shown, first on retinal ganglion cell recordings, and, more recently, on prefrontal cortex [173, 174] and hippocampal data [1, 257], to provide a good approximation for the distribution of population activity configurations. The performance in the decoding task has been shown to be superior to rate-based decoders on CA1 data [1].

The standard tool used to compute the performance of binary decoders is the Receiver Operating Characteristic (ROC) diagram [191]. This diagram is drawn by computing the true positive rate (TPR) and false positive rate (FPR) as a function of different thresholding values, and plotting the resulting curve in the TPR-FPR plane. These quantities can be defined in the context of our map decoder as follows: for each theta bin t the decoder outputs a value $\Delta\mathcal{L}(t)$, which is then interpreted as referring to map A or B depending on its value compared to a moving threshold Θ. Note that this is slightly different from



Supplementary Figure G. Decoded maps as a function of time for all 15 **light-switch events in the test session**. Same analysis in main text Fig. 1C, which was restricted to a single light switch. PI-realignment times are marked with a dashed green line.

the map-decoding method reported in Main Text, since it does not allow undecoded statistically-not-significant bins.

$$m_t = \begin{cases} A & if \;\; \Delta\mathcal{L}(t) > \Theta \,, \\ B & if \;\; \Delta\mathcal{L}(t) < \Theta \,. \end{cases} \tag{A.9}$$

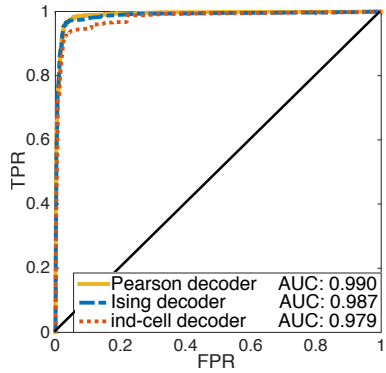
To match the vocables used in the ROC framework we will arbitrarily follow the convention that the output is *positive* if the map is decoded to be A, and *negative* if the map is predicted to be B. Doing so, a True Positive is defined as a correctly-decoded environment A (with respect to the light conditions: $m_t = A =$ light cues), while a True Negative will be a correctly-decoded environment B. The final observable (area under the ROC curve) is symmetrical under the inversion of this convention, which is summarized in Table A.1. The decoding capability is finally assessed by applying the decoder to two "constant" test sessions, where the environment is constantly set to A and B, respectively. Assuming that the neural representation is stable under fixed light conditions, we can compute the TPR and FPR of the decoder by counting how many theta bins are correctly and falsely decoded in the two reference sessions. For a specific value of the threshold Θ, this corresponds to a point in the FPR-TPR plane. By varying this value we then draw the curve as the succession of the corresponding TPR-FPR values. The standard quantitative measure of the decoding performances is the Area Under the Curve (AUC) of the ROC diagram [191]. According to this measure, the ideal decoder has AUC $= 1$, while random guessing would give AUC $= 0.5$. All the decoders, tested on constant test sessions, i.e. where no teleportation is performed, show very high performances, see Supplementary Fig. H. Note, in addition, that our functional-network based decoder is robust against the presence of correlations between the maps: it shows much better performance than correlation-based methods for CA1 recordings, where maps are much less orthogonal than in CA3 [1].

decoder output	A	B	A	B
light conditions	A	A	B	B
denomination	True Positive	False Negative	False Positive	True Negative

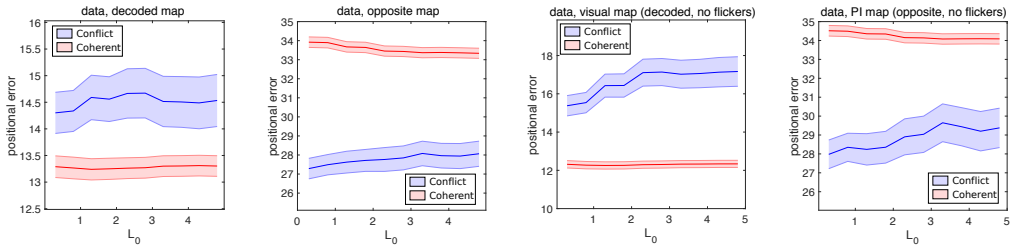
Table A.1: Denominations used for the four possible events, depending on the output of the decoder and on the environment-defining cue. The cue is not changed throughout the reference session.

A.7 DEPENDENCE OF POSITIONAL-ERROR ANALYSIS WITH L_0

The significance of positional-error analysis as a function of the threshold L_0 is shown in Supplementary Fig. I.



Supplementary Figure H. ROC curves for our Ising-model map decoder (blue) compared to independent-cell decoder (specific case of Ising model, with zero couplings J, red dotted curve) and the Pearson-correlation based decoder of [143], which used the true position of the rodent.



Supplementary Figure I. Significance of positional-error analysis as a function L_0 Mean positional error (line) and standard error (shaded area) for CA3 data reported in main text Figure 6 A (position inferred according to the decoded map), Figure 6 B (position inferred according to the opposite map) and Figure 6 C,D (position inferred according to the decoded/opposite map without considering flickering events), for a large range of values of the threshold L_0 of the map decoder used in the conflict period identification; higher values of L_0 correspond to shorter decoded conflict periods.

APPENDIX - CHAPTER 11

B

B.1 BIAS-VARIANCE DIAGRAM OF FAMILIES OF STRUCTURES C AND D

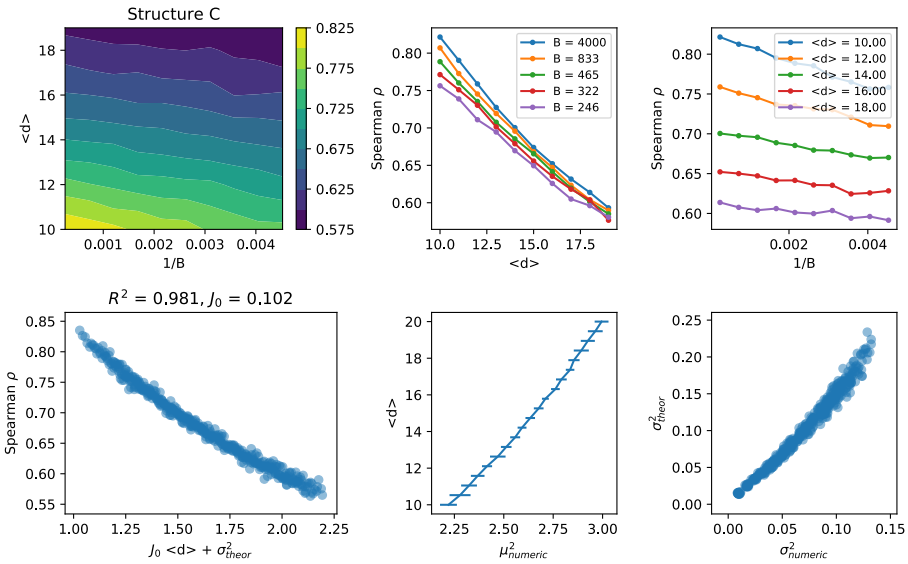


Figure B.1: **Bias-variance diagram for structure C** $\theta = 5$, 400 points.

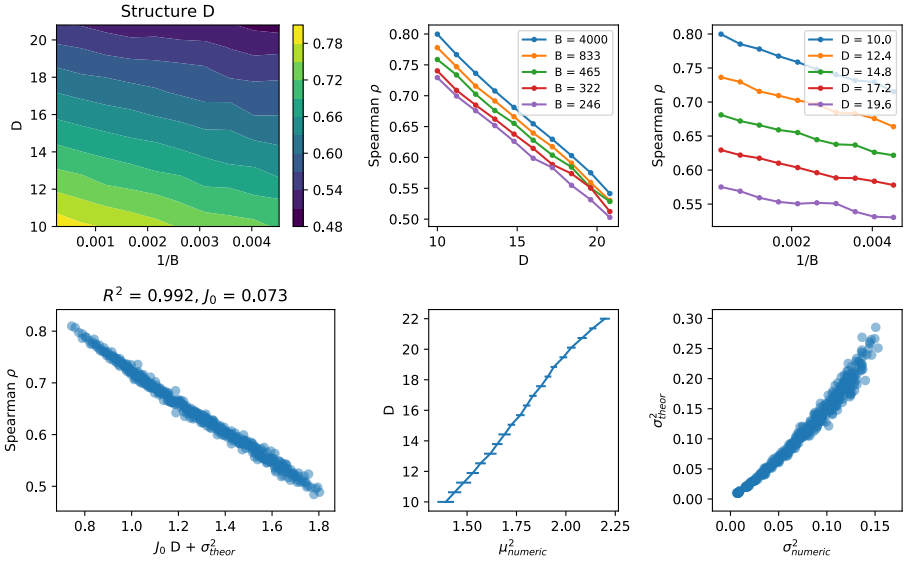


Figure B.2: **Bias-variance diagram for structure D** $\theta = 5$, 400 points.

BIBLIOGRAPHY

[1] L. Posani, S. Cocco, K. Ježek, and R. Monasson, "Functional connectivity models for decoding of spatial representations from hippocampal ca1 recordings," Journal of Computational Neuroscience, pp. 1–17, 2017.

[2] L. Posani, S. Cocco, and R. Monasson, "Integration and multiplexing of positional and contextual information by the hippocampal network," PLoS computational biology, vol. 14, no. 8, p. e1006320, 2018.

[3] L. Posani, S. Cocco, K. Jezek, and R. Monasson, "Position is coherently represented during flickering instabilities of place-cell cognitive maps in the hippocampus," in BMC neuroscience - 26th Annual Computational Neuroscience Meeting (CNS* 2017), vol. 18, p. 58, BioMed Central, 2017.

[4] S. Cocco, R. Monasson, L. Posani, and G. Tavoni, "Functional networks from inverse modeling of neural population activity," Current Opinion in Systems Biology, 2017.

[5] S. Cocco, R. Monasson, L. Posani, S. Rosay, and J. Tubiana, "Statistical physics and representations in real and artificial neural networks," Physica A: Statistical Mechanics and its Applications, p. doi/10.1016/j.physa.2017.11.153, 2017.

[6] L. Posani, , S. Cocco, and R. Monasson, "Pairwise models inferred from hippocampal activity generate neural configurations typical of single or multiple low-dimensional attractors," Draft, 2018.

[7] L. Posani, F. Rizzato, R. Monasson, and S. Cocco, "Infer global, predict local: Bias-variance trade-off in protein fitness landscape reconstruction from sequence data," Draft, 2018.

[8] L. Posani, F. Rizzato, R. Monasson, and S. Cocco, "Improved performance of dca fitness predictions with sparsity prior from structural information," Draft, 2018.

[9] J. W. Gibbs, Elementary principles in statistical mechanics. Courier Corporation, 2014.

[10] H. Poincaré, "Sur le problème des trois corps et les équations de la dynamique," Acta mathematica, vol. 13, no. 1, pp. A3–A270, 1890.

[11] J. L. Lebowitz and O. Penrose, "Rigorous treatment of the van der waals-maxwell theory of the liquid-vapor transition," J. Math. Phys., vol. 7, p. 98, 1966.

[12] R. Monasson and S. Rosay, "Transitions between spatial attractors in place-cell models," Physical review letters, vol. 115, no. 9, p. 098101, 2015.

[13] K. Sharp and F. Matschinsky, "Translation of ludwig boltzmann's paper "on the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium" sitzungberichte der kaiserlichen akademie der wissenschaften. mathematisch-naturwissen classe. abt. ii, lxxvi 1877, pp 373-435 (wien. ber. 1877, 76: 373-435). reprinted in wiss. abhandlungen, vol. ii, reprint 42, p. 164-223, barth, leipzig, 1909," Entropy, vol. 17, no. 4, pp. 1971–2009, 2015.

[14] M. Campisi and D. H. Kobe, "Derivation of the boltzmann principle," American Journal of Physics, vol. 78, no. 6, pp. 608–615, 2010.

[15] E. T. Jaynes, "Gibbs vs boltzmann entropies," American Journal of Physics, vol. 33, no. 5, pp. 391–398, 1965.

[16] E. T. Jaynes, "Foundations of probability theory and statistical mechanics," in Delaware seminar in the foundations of physics, pp. 77–101, Springer, 1967.

[17] E. T. Jaynes, Probability theory: The logic of science. Cambridge university press, 2003.

[18] R. T. Cox, "Probability, frequency and reasonable expectation," American journal of physics, vol. 14, no. 1, pp. 1–13, 1946.

[19] G. Polya, Mathematics and Plausible Reasoning: Patterns of plausible inference, vol. 2. Princeton University Press, 1990.

[20] J. Bernoulli, Ars conjectandi. Impensis Thurnisiorum, fratrum, 1713.

[21] D. J. MacKay, Information theory, inference and learning algorithms. Cambridge university press, 2003.

[22] E. T. Jaynes, "Information theory and statistical mechanics," Physical review, vol. 106, no. 4, p. 620, 1957.

[23] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," IEEE Transactions on information theory, vol. 26, no. 1, pp. 26–37, 1980.

[24] E. T. Jaynes, "Information theory and statistical mechanics. ii," Physical review, vol. 108, no. 2, p. 171, 1957.

[25] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback, "Bayesian inference of phylogeny and its impact on evolutionary biology," science, vol. 294, no. 5550, pp. 2310–2314, 2001.

[26] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to bayesian network inference for generating causal networks from observational biological data," Bioinformatics, vol. 20, no. 18, pp. 3594–3603, 2004.

[27] D. J. Wilkinson, "Bayesian methods in bioinformatics and computational systems biology," Briefings in bioinformatics, vol. 8, no. 2, pp. 109–116, 2007.

[28] E. Ising, "Beitrag zur theorie des ferromagnetismus," Zeitschrift für Physik, vol. 31, no. 1, pp. 253–258, 1925.

[29] L. Onsager, "Crystal statistics. i. a two-dimensional model with an order-disorder transition," Physical Review, vol. 65, no. 3-4, p. 117, 1944.

[30] M. J. Wainwright, M. I. Jordan, et al., "Graphical models, exponential families, and variational inference," Foundations and Trends® in Machine Learning, vol. 1, no. 1–2, pp. 1–305, 2008.

[31] J. P. Barton, S. Cocco, E. De Leonardis, and R. Monasson, "Large pseudocounts and l 2-norm penalties are necessary for the mean-field inference of ising and potts models," Physical Review E, vol. 90, no. 1, p. 012132, 2014.

[32] P. Ravikumar, M. J. Wainwright, J. D. Lafferty, et al., "High-dimensional ising model selection using l1-regularized logistic regression," The Annals of Statistics, vol. 38, no. 3, pp. 1287–1319, 2010.

[33] S. Cocco and R. Monasson, "Adaptive cluster expansion for the inverse ising problem: convergence, algorithm and tests," J. Stat. Phys, vol. 147, no. 2, pp. 252–314, 2012.

[34] H. C. Nguyen, R. Zecchina, and J. Berg, "Inverse statistical problems: from the inverse ising problem to data science," Advances in Physics, vol. 66, no. 3, pp. 197–261, 2017.

[35] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," Cognitive science, vol. 9, no. 1, pp. 147–169, 1985.

[36] M. Opper and D. Saad, Advanced mean field methods: Theory and practice. MIT press, 2001.

[37] T. Tanaka, "Information geometry of mean-field approximation," Neural Computation, vol. 12, no. 8, pp. 1951–1968, 2000.

[38] A. Pelizzola, "Cluster variation method in statistical physics and probabilistic graphical models," Journal of Physics A: Mathematical and General, vol. 38, no. 33, p. R309, 2005.

[39] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," IEEE Transactions on information theory, vol. 51, no. 7, pp. 2282–2312, 2005.

[40] M. Mezard and A. Montanari, Information, physics, and computation. Oxford University Press, 2009.

[41] S. Geman and C. Graffigne, "Markov random field image models and their applications to computer vision," in Proceedings of the international congress of mathematicians, vol. 1, p. 2, Berkeley, CA, 1986.

[42] J. D. Kalbfleisch, "Pseudo-likelihood," Encyclopedia of Biostatistics, vol. 6, 2005.

[43] M. Ekeberg, T. Hartonen, and E. Aurell, "Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences," Journal of Computational Physics, vol. 276, pp. 341–356, 2014.

[44] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. Schärfe, M. Springer, C. Sander, and D. S. Marks, "Mutation effects predicted from sequence co-variation," Nature biotechnology, vol. 35, no. 2, p. 128, 2017.

[45] S. Cocco and R. Monasson, "Adaptive cluster expansion for inferring boltzmann machines with noisy data," Physical review letters, vol. 106, no. 9, p. 090601, 2011.

[46] J. P. Barton and S. Cocco, "Ising models for neural activity inferred via selective cluster expansion: Structural and coding properties," J Stat Mech, p. P03002, 2013.

[47] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco, "Ace: adaptive cluster expansion for maximum entropy graphical model inference," Bioinformatics, vol. 32, no. 20, pp. 3089–3097, 2016.

[48] I. H. Stevenson, J. M. Rebesco, L. E. Miller, and K. P. Körding, "Inferring functional connections between neurons," Current opinion in neurobiology, vol. 18, no. 6, pp. 582–588, 2008.

[49] S. Cocco, R. Monasson, L. Posani, and G. Tavoni, "Functional networks from inverse modeling of neural population activity," in pubblication on Curr Opinion in Systems Biology, 2017.

[50] S. Cocco, S. Leibler, and R. Monasson, "Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods," Proceedings of the National Academy of Sciences, vol. 106, no. 33, pp. 14058–14062, 2009.

[51] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, "Weak pairwise correlations imply strongly correlated network states in a neural population," Nature, vol. 440, no. 7087, pp. 1007–1012, 2006.

[52] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," Nature, vol. 454, no. 7207, pp. 995–999, 2008.

[53] G. Tkačik, J. S. Prentice, V. Balasubramanian, and E. Schneidman, "Optimal population coding by noisy spiking neurons," Proceedings of the National Academy of Sciences, vol. 107, no. 32, pp. 14419–14424, 2010.

[54] G. Tavoni, U. Ferrari, F. P. Battaglia, S. Cocco, and R. Monasson, "Inferred model of the prefrontal cortex activity unveils cell assemblies and memory replay," bioRxiv, p. 028316, 2015.

[55] G. Tavoni, S. Cocco, and R. Monasson, "Neural assemblies revealed by inferred connectivity-based models of prefrontal cortex recordings," Journal of Computational Neuroscience, vol. 41, pp. 269–293, 2016.

[56] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, "Improved contact prediction in proteins: using pseudolikelihoods to infer potts models," Physical Review E, vol. 87, no. 1, p. 012707, 2013.

[57] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," Proceedings of the National Academy of Sciences, vol. 108, no. 49, pp. E1293–E1301, 2011.

[58] H. Jacquin, A. Gilson, E. Shakhnovich, S. Cocco, and R. Monasson, "Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models," PLoS computational biology, vol. 12, no. 5, p. e1004889, 2016.

[59] M. Gell-Mann, "What is complexity?," in Complexity and industrial clusters, pp. 13–24, Springer, 2002.

[60] J. Ladyman, J. Lambert, and K. Wiesner, "What is a complex system?," European Journal for Philosophy of Science, vol. 3, no. 1, pp. 33–67, 2013.

[61] F. Cucker, S. Smale, et al., "Emergent behavior in flocks," IEEE Transactions on automatic control, vol. 52, no. 5, pp. 852–862, 2007.

[62] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, et al., "Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study," Proceedings of the national academy of sciences, vol. 105, no. 4, pp. 1232–1237, 2008.

[63] J. K. Parrish, S. V. Viscido, and D. Grunbaum, "Self-organized fish schools: an examination of emergent properties," The biological bulletin, vol. 202, no. 3, pp. 296–305, 2002.

[64] M. Moussaïd, D. Helbing, and G. Theraulaz, "How simple rules determine pedestrian behavior and crowd disasters," Proceedings of the National Academy of Sciences, vol. 108, no. 17, pp. 6884–6888, 2011.

[65] C. Sparrow, The Lorenz equations: bifurcations, chaos, and strange attractors, vol. 41. Springer Science & Business Media, 2012.

[66] M. Marsili, I. Mastromatteo, and Y. Roudi, "On sampling and modeling complex systems," Journal of Statistical Mechanics: Theory and Experiment, vol. 2013, no. 09, p. P09003, 2013.

[67] A. Haimovici and M. Marsili, "Criticality of mostly informative samples: a bayesian model selection approach," Journal of Statistical Mechanics: Theory and Experiment, vol. 2015, no. 10, p. P10013, 2015.

[68] L. D. Landau, "On the theory of phase transitions," Ukr. J. Phys., vol. 11, pp. 19–32, 1937.

[69] H. E. Stanley, Phase transitions and critical phenomena. Clarendon Press, Oxford, 1971.

[70] K. Sokolowski and J. G. Corbin, "Wired for behaviors: from development to function of innate limbic system circuitry," Frontiers in molecular neuroscience, vol. 5, p. 55, 2012.

[71] W. B. Scoville and B. Milner, "Loss of recent memory after bilateral hippocampal lesions," Journal of neurology, neurosurgery, and psychiatry, vol. 20, no. 1, p. 11, 1957.

[72] S. Zola-Morgan, L. R. Squire, and D. Amaral, "Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field ca1 of the hippocampus," Journal of Neuroscience, vol. 6, no. 10, pp. 2950–2967, 1986.

[73] P. Andersen, R. Morris, D. Amaral, J. O'Keefe, and T. Bliss, The hippocampus book. Oxford university press, 2007.

[74] D. Virley, R. M. Ridley, J. D. Sinden, T. R. Kershaw, S. Harland, T. Rashid, S. French, P. Sowinski, J. A. Gray, P. L. Lantos, et al., "Primary ca1 and conditionally immortal mhp36 cell grafts restore conditional discrimination learning and recall in marmosets after excitotoxic lesions of the hippocampal ca1 field," Brain, vol. 122, no. 12, pp. 2321–2335, 1999.

[75] E. C. Tolman, "Cognitive maps in rats and men.," Psychological review, vol. 55, no. 4, p. 189, 1948.

[76] E. C. Tolman, B. F. Ritchie, and D. Kalish, "Studies in spatial learning. i. orientation and the short-cut.," Journal of experimental psychology, vol. 36, no. 1, p. 13, 1946.

[77] R. G. Morris, "Spatial localization does not require the presence of local cues," Learning and motivation, vol. 12, no. 2, pp. 239–260, 1981.

[78] R. Morris, P. Garrud, J. a. Rawlins, and J. O'Keefe, "Place navigation impaired in rats with hippocampal lesions," Nature, vol. 297, no. 5868, p. 681, 1982.

[79] J. O'Keefe and J. Dostrovsky, "The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat.," Brain research, 1971.

[80] J. O'Keefe, "Place units in the hippocampus of the freely moving rat," Experimental neurology, vol. 51, no. 1, pp. 78–109, 1976.

[81] M. A. Wilson and B. L. McNaughton, "Dynamics of the hippocampal ensemble code for space," Science, vol. 261, no. 5124, pp. 1055–1058, 1993.

[82] L. Thompson and P. Best, "Place cells and silent cells in the hippocampus of freely-behaving rats," Journal of Neuroscience, vol. 9, no. 7, pp. 2382–2390, 1989.

[83] J. K. Leutgeb, S. Leutgeb, A. Treves, R. Meyer, C. A. Barnes, B. L. McNaughton, M.-B. Moser, and E. I. Moser, "Progressive transformation of hippocampal neuronal representations in "morphed" environments," Neuron, vol. 48, no. 2, pp. 345–358, 2005.

[84] R. U. Muller and J. L. Kubie, "The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells," Journal of Neuroscience, vol. 7, no. 7, pp. 1951–1968, 1987.

[85] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," Nature, vol. 436, no. 7052, pp. 801–806, 2005.

[86] J. J. Knierim, H. S. Kudrimoti, and B. L. McNaughton, "Place cells, head direction cells, and the learning of landmark stability," Journal of Neuroscience, vol. 15, no. 3, pp. 1648–1659, 1995.

[87] T. A. Allen, D. M. Salz, S. McKenzie, and N. J. Fortin, "Nonspatial sequence coding in ca1 neurons," Journal of Neuroscience, vol. 36, no. 5, pp. 1547–1563, 2016.

[88] D. M. Smith and S. J. Mizumori, "Hippocampal place cells, context, and episodic memory," Hippocampus, vol. 16, no. 9, pp. 716–729, 2006.

[89] J. O'keefe and D. Conway, "Hippocampal place units in the freely moving rat: why they fire where they fire," Experimental Brain Research, vol. 31, no. 4, pp. 573–590, 1978.

[90] K. J. Jeffery and M. I. Anderson, "Dissociation of the geometric and contextual influences on place cells," Hippocampus, vol. 13, no. 7, pp. 868–872, 2003.

[91] S. Leutgeb, J. K. Leutgeb, A. Treves, M.-B. Moser, and E. I. Moser, "Distinct ensemble codes in hippocampal areas ca3 and ca1," Science, vol. 305, no. 5688, pp. 1295–1298, 2004.

[92] J. O'keefe and L. Nadel, The hippocampus as a cognitive map. Oxford: Clarendon Press, 1978.

[93] L. L. Colgin, E. I. Moser, and M.-B. Moser, "Understanding memory through hippocampal remapping," Trends in neurosciences, vol. 31, no. 9, pp. 469–477, 2008.

[94] S. Leutgeb, J. K. Leutgeb, C. A. Barnes, E. I. Moser, B. L. McNaughton, and M.-B. Moser, "Independent codes for spatial and episodic memory in hippocampal neuronal ensembles," Science, vol. 309, no. 5734, pp. 619–623, 2005.

[95] J. Ranck Jr, "Head direction cells in the deep layer of dorsal presubiculum in freely moving rats," in Society of Neuroscience Abstract, vol. 10, p. 599, 1984.

[96] J. S. Taube, "The head direction signal: origins and sensory-motor integration," Annu. Rev. Neurosci., vol. 30, pp. 181–207, 2007.

[97] F. Sargolini, M. Fyhn, T. Hafting, B. L. McNaughton, M. P. Witter, M.-B. Moser, and E. I. Moser, "Conjunctive representation of position, direction, and velocity in entorhinal cortex," Science, vol. 312, no. 5774, pp. 758–762, 2006.

[98] S. Mizumori and J. Williams, "Directionally selective mnemonic properties of neurons in the lateral dorsal nucleus of the thalamus of rats," Journal of Neuroscience, vol. 13, no. 9, pp. 4015–4028, 1993.

[99] P. E. Sharp and K. Koester, "Lesions of the mammillary body region severely disrupt the cortical head direction, but not place cell signal," Hippocampus, vol. 18, no. 8, pp. 766–784, 2008.

[100] L. L. Chen, L.-H. Lin, E. J. Green, C. A. Barnes, and B. L. McNaughton, "Head-direction cells in the rat posterior cortex," Experimental Brain Research, vol. 101, no. 1, pp. 8–23, 1994.

[101] S. I. Wiener, "Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task," Journal of Neuroscience, vol. 13, no. 9, pp. 3802–3817, 1993.

[102] N. M. Van Strien, N. Cappaert, and M. P. Witter, "The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network," Nature Reviews Neuroscience, vol. 10, no. 4, p. 272, 2009.

[103] J. S. Taube, R. U. Muller, and J. B. Ranck, "Head-direction cells recorded from the postsubiculum in freely moving rats. ii. effects of environmental manipulations," Journal of Neuroscience, vol. 10, no. 2, pp. 436–447, 1990.

[104] E. J. Golob and J. S. Taube, "Head direction cells in rats with hippocampal or overlying neocortical lesions: evidence for impaired angular path integration," Journal of Neuroscience, vol. 19, no. 16, pp. 7198–7211, 1999.

[105] P.-Y. Jacob, G. Casali, L. Spieser, H. Page, D. Overington, and K. Jeffery, "An independent, landmark-dominated head-direction signal in dysgranular retrosplenial cortex," Nature neuroscience, vol. 20, no. 2, p. 173, 2017.

[106] M. Fyhn, S. Molden, M. P. Witter, E. I. Moser, and M.-B. Moser, "Spatial representation in the entorhinal cortex," Science, vol. 305, no. 5688, pp. 1258–1264, 2004.

[107] V. H. Brun, T. Solstad, K. B. Kjelstrup, M. Fyhn, M. P. Witter, E. I. Moser, and M.-B. Moser, "Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex," Hippocampus, vol. 18, no. 12, pp. 1200–1212, 2008.

[108] M. Fyhn, T. Hafting, A. Treves, M.-B. Moser, and E. I. Moser, "Hippocampal remapping and grid realignment in entorhinal cortex," Nature, vol. 446, no. 7132, pp. 190–194, 2007.

[109] B. L. McNaughton, F. P. Battaglia, O. Jensen, E. I. Moser, and M.-B. Moser, "Path integration and the neural basis of the'cognitive map'," Nature Reviews Neuroscience, vol. 7, no. 8, pp. 663–678, 2006.

[110] J. O'keefe and N. Burgess, "Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells," Hippocampus, vol. 15, no. 7, pp. 853–866, 2005.

[111] K. Yoon, M. A. Buice, C. Barry, R. Hayman, N. Burgess, and I. R. Fiete, "Specific evidence of low-dimensional continuous attractor dynamics in grid cells," Nature neuroscience, vol. 16, no. 8, pp. 1077–1084, 2013.

[112] M. Müller and R. Wehner, "Path integration in desert ants, cataglyphis fortis," Proceedings of the National Academy of Sciences, vol. 85, no. 14, pp. 5287–5290, 1988.

[113] K. Von Frisch, "The dance language and orientation of bees.," 1967.

[114] P. Moller and P. Görner, "Homing by path integration in the spider agelena labyrinthica clerck," Journal of Comparative Physiology A, vol. 174, no. 2, pp. 221–229, 1994.

[115] U. von Saint Paul, "Do geese use path integration for walking home?," in Avian navigation, pp. 298–307, Springer, 1982.

[116] M.-L. Mittelstaedt and H. Mittelstaedt, "Homing by path integration in a mammal," Naturwissenschaften, vol. 67, no. 11, pp. 566–567, 1980.

[117] M.-L. Mittelstaedt and S. Glasauer, "Idiothetic navigation in gerbils and humans," Zool. Jb. Physiol, vol. 95, no. 427-435, 1991.

[118] J. O'Keefe and A. Speakman, "Single unit activity in the rat hippocampus during a spatial memory task," Experimental brain research, vol. 68, no. 1, pp. 1–27, 1987.

[119] J. O'keefe, N. Burgess, et al., "Geometric determinants of the place fields of hippocampal neurons," Nature, vol. 381, no. 6581, pp. 425–428, 1996.

[120] T. Hartley, N. Burgess, C. Lever, F. Cacucci, and J. O'Keefe, "Modeling place fields in terms of the cortical inputs to the hippocampus," Hippocampus, vol. 10, no. 4, pp. 369–379, 2000.

[121] C. Lever, T. Wills, F. Cacucci, N. Burgess, and J. O'keefe, "Long-term plasticity in hippocampal place-cell representation of environmental geometry," Nature, vol. 416, no. 6876, pp. 90–94, 2002.

[122] C. Lever, S. Burton, A. Jeewajee, J. O'Keefe, and N. Burgess, "Boundary vector cells in the subiculum of the hippocampal formation," Journal of Neuroscience, vol. 29, no. 31, pp. 9771–9777, 2009.

[123] C. Barry and N. Burgess, "Neural mechanisms of self-location," Current Biology, vol. 24, no. 8, pp. R330–R339, 2014.

[124] G. Chen, J. A. King, N. Burgess, and J. O'Keefe, "How vision and movement combine in the hippocampal place code," Proceedings of the National Academy of Sciences, vol. 110, no. 1, pp. 378–383, 2013.

[125] P. Ravassard, A. Kees, B. Willers, D. Ho, D. Aharoni, J. Cushman, Z. M. Aghajan, and M. R. Mehta, "Multisensory control of hippocampal spatiotemporal selectivity," Science, vol. 340, no. 6138, pp. 1342–1346, 2013.

[126] K. M. Gothard, W. E. Skaggs, and B. L. McNaughton, "Dynamics of mismatch correction in the hippocampal ensemble code for space: interaction between path integration and environmental cues," Journal of Neuroscience, vol. 16, no. 24, pp. 8027–8040, 1996.

[127] P. Andersen, T. Bliss, and K. K. Skrede, "Lamellar organization of hippocampal excitatory pathways," Experimental Brain Research, vol. 13, no. 2, pp. 222–238, 1971.

[128] P. Andersen, "Organization of hippocampal neurons and their interconnections," in The hippocampus, pp. 155–175, Springer, 1975.

[129] D. Bush, C. Barry, and N. Burgess, "What do grid cells contribute to place cell firing?," Trends in neurosciences, vol. 37, no. 3, pp. 136–145, 2014.

[130] O. V. Haas, J. Henke, C. Leibold, and K. Thurley, "Modality-specific subpopulations of place fields coexist in the hippocampus," Cerebral Cortex, 2018.

[131] M. C. Fuhs and D. S. Touretzky, "A spin glass model of path integration in rat medial entorhinal cortex," Journal of Neuroscience, vol. 26, no. 16, pp. 4266–4276, 2006.

[132] E. T. Rolls, S. M. Stringer, and T. Elliot, "Entorhinal cortex grid cells can map to hippocampal place cells by competitive learning," Network: Computation in Neural Systems, vol. 17, no. 4, pp. 447–465, 2006.

[133] T. Solstad, E. I. Moser, and G. T. Einevoll, "From grid cells to place cells: a mathematical model," Hippocampus, vol. 16, no. 12, pp. 1026–1031, 2006.

[134] R. M. Hayman and K. J. Jeffery, "How heterogeneous place cell responding arises from homogeneous grids—a contextual gating hypothesis," Hippocampus, vol. 18, no. 12, pp. 1301–1313, 2008.

[135] B. Si and A. Treves, "The role of competitive learning in the generation of dg fields from ec inputs," Cognitive neurodynamics, vol. 3, no. 2, pp. 177–187, 2009.

[136] R. F. Langston, J. A. Ainge, J. J. Couey, C. B. Canto, T. L. Bjerknes, M. P. Witter, E. I. Moser, and M.-B. Moser, "Development of the spatial representation system in the rat," Science, vol. 328, no. 5985, pp. 1576–1580, 2010.

[137] T. J. Wills, F. Cacucci, N. Burgess, and J. O'Keefe, "Development of the hippocampal cognitive map in preweanling rats," Science, vol. 328, no. 5985, pp. 1573–1576, 2010.

[138] C. Barry, L. L. Ginzberg, J. O'Keefe, and N. Burgess, "Grid cell firing patterns signal environmental novelty by expansion," Proceedings of the National Academy of Sciences, vol. 109, no. 43, pp. 17687–17692, 2012.

[139] M. Brandon, J. Koenig, M. Hasselmo, J. Leutgeb, and S. Leutgeb, "Septal inactivation eliminates grid cell spatial periodicity and causes instability of hippocampal place cells in novel environments," in Soc. Neurosci. Abstr, vol. 203, 2012.

[140] J. Koenig, A. N. Linder, J. K. Leutgeb, and S. Leutgeb, "The spatial periodicity of grid cells is not sustained during reduced theta oscillations," Science, vol. 332, no. 6029, pp. 592–595, 2011.

[141] E. Kelemen and A. A. Fenton, "Dynamic grouping of hippocampal neural activity during cognitive control of two spatial frames," PLoS biology, vol. 8, no. 6, p. e1000403, 2010.

[142] W. A. Phillips and W. Singer, "In search of common foundations for cortical computation," Behavioral and Brain Sciences, vol. 20, no. 4, pp. 657–683, 1997.

[143] K. Jezek, E. J. Henriksen, A. Treves, E. I. Moser, and M.-B. Moser, "Theta-paced flickering between place-cell maps in the hippocampus," Nature, vol. 478, no. 7368, pp. 246–249, 2011.

[144] D. O. Hebb, The organization of behavior: A neuropsychological theory. Psychology Press, 2005.

[145] T. V. Bliss and T. Lømo, "Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path," The Journal of physiology, vol. 232, no. 2, pp. 331–356, 1973.

[146] R. C. Malenka, "The long-term potential of ltp," Nature Reviews Neuroscience, vol. 4, no. 11, p. 923, 2003.

[147] T. V. Bliss and G. L. Collingridge, "A synaptic model of memory: long-term potentiation in the hippocampus," Nature, vol. 361, no. 6407, p. 31, 1993.

[148] R. Morris, "Synaptic plasticity and learning: selective impairment of learning rats and blockade of long-term potentiation in vivo by the n-methyl-d-aspartate receptor antagonist ap5," Journal of Neuroscience, vol. 9, no. 9, pp. 3040–3057, 1989.

[149] S. J. Martin, P. D. Grimwood, and R. G. Morris, "Synaptic plasticity and memory: an evaluation of the hypothesis," Annual review of neuroscience, vol. 23, no. 1, pp. 649–711, 2000.

[150] S. Davis, S. Butcher, and R. Morris, "The nmda receptor antagonist d-2-amino-5-phosphonopentanoate (d-ap5) impairs spatial learning and ltp in vivo at intracerebral concentrations comparable to those that block ltp in vitro," Journal of Neuroscience, vol. 12, no. 1, pp. 21–34, 1992.

[151] G. Riedel, J. Micheau, A. Lam, E. Roloff, S. J. Martin, H. Bridge, L. De Hoz, B. Poeschel, J. McCulloch, and R. G. Morris, "Reversible neural inactivation reveals hippocampal participation in several memory processes," Nature neuroscience, vol. 2, no. 10, p. 898, 1999.

[152] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proceedings of the national academy of sciences, vol. 79, no. 8, pp. 2554–2558, 1982.

[153] D. J. Amit, Modeling brain function: The world of attractor neural networks. Cambridge University Press, 1992.

[154] M. Tsodyks and T. Sejnowski, "Associative memory and hippocampal place cells," International journal of neural systems, vol. 6, pp. 81–86, 1995.

[155] A. Samsonovich and B. L. McNaughton, "Path integration and cognitive mapping in a continuous attractor neural network model," The Journal of neuroscience, vol. 17, no. 15, pp. 5900–5920, 1997.

[156] F. P. Battaglia and A. Treves, "Attractor neural networks storing multiple space representations: a model for hippocampal place fields," Physical Review E, vol. 58, no. 6, p. 7738, 1998.

[157] F. Stella and A. Treves, "Associative memory storage and retrieval: involvement of theta oscillations in hippocampal information processing," Neural plasticity, vol. 2011, 2011.

[158] R. Monasson and S. Rosay, "Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Collective motion of the activity," Physical Review E, vol. 89, no. 3, p. 032803, 2014.

[159] R. Monasson and S. Rosay, "Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Dynamics within one map (ii)," 2013.

[160] R. Monasson and S. Rosay, "Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Phase diagram," Physical Review E, vol. 87, no. 6, p. 062813, 2013.

[161] R. Ben-Yishai, R. L. Bar-Or, and H. Sompolinsky, "Theory of orientation tuning in visual cortex.," Proceedings of the National Academy of Sciences, vol. 92, no. 9, pp. 3844–3848, 1995.

[162] P. E. Latham, S. Deneve, and A. Pouget, "Optimal computation with attractor networks," Journal of Physiology-Paris, vol. 97, no. 4-6, pp. 683–694, 2003.

[163] H. S. Seung, "How the brain keeps the eyes still," Proceedings of the National Academy of Sciences, vol. 93, no. 23, pp. 13339–13344, 1996.

[164] S. Cannon and D. Robinson, "Loss of the neural integrator of the oculomotor system from brain stem lesions in monkey," Journal of neurophysiology, vol. 57, no. 5, pp. 1383–1409, 1987.

[165] R. Romo, C. D. Brody, A. Hernández, and L. Lemus, "Neuronal correlates of parametric working memory in the prefrontal cortex," Nature, vol. 399, no. 6735, p. 470, 1999.

[166] P. Miller, C. D. Brody, R. Romo, and X.-J. Wang, "A recurrent network model of somatosensory parametric working memory in the prefrontal cortex," Cerebral Cortex, vol. 13, no. 11, pp. 1208–1218, 2003.

[167] K. Zhang, "Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory," Journal of Neuroscience, vol. 16, no. 6, pp. 2112–2126, 1996.

[168] H. T. Blair and P. E. Sharp, "Anticipatory head direction signals in anterior thalamus: evidence for a thalamocortical circuit that integrates angular head motion to compute head direction," Journal of Neuroscience, vol. 15, no. 9, pp. 6260–6270, 1995.

[169] Y. Burak and I. R. Fiete, "Accurate path integration in continuous attractor network models of grid cells," PLoS computational biology, vol. 5, no. 2, p. e1000291, 2009.

[170] S. S. Kim, H. Rouault, S. Druckmann, and V. Jayaraman, "Ring attractor dynamics in the drosophila central brain," Science, vol. 356, no. 6340, pp. 849–853, 2017.

[171] K. Wimmer, D. Q. Nykamp, C. Constantinidis, and A. Compte, "Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory," Nature neuroscience, vol. 17, no. 3, pp. 431–439, 2014.

[172] T. J. Wills, C. Lever, F. Cacucci, N. Burgess, and J. O'Keefe, "Attractor dynamics in the hippocampal representation of the local environment," Science, vol. 308, no. 5723, pp. 873–876, 2005.

[173] G. Tavoni, S. Cocco, and R. Monasson, "Neural assemblies revealed by inferred connectivity-based models of prefrontal cortex recordings," Journal of computational neuroscience, vol. 41, no. 3, pp. 269–293, 2016.

[174] G. Tavoni, U. Ferrari, F. P. Battaglia, S. Cocco, and R. Monasson, "Functional coupling networks inferred from prefrontal cortex activity show experience-related effective plasticity," Network Neuroscience, vol. 1, no. 3, pp. 275–301, 2017.

[175] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, "The importance of mixed selectivity in complex cognitive tasks," Nature, vol. 497, no. 7451, pp. 585–590, 2013.

[176] K. Zhang, I. Ginzburg, B. L. McNaughton, and T. J. Sejnowski, "Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells," Journal of neurophysiology, vol. 79, no. 2, pp. 1017–1044, 1998.

[177] J. Barton, E. De Leonardis, A. Coucke, and S. Cocco, "Ace: adaptive cluster expansion for maximum entropy graphical model inference," Bioinformatics, 2016.

[178] M. Okatan, M. A. Wilson, and E. N. Brown, "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity," Neural computation, vol. 17, no. 9, pp. 1927–1961, 2005.

[179] L. L. Colgin, S. Leutgeb, K. Jezek, J. K. Leutgeb, E. I. Moser, B. L. McNaughton, and M.-B. Moser, "Attractor-map versus autoassociation based attractor dynamics in the hippocampal network," Journal of neurophysiology, vol. 104, no. 1, pp. 35–50, 2010.

[180] L. Posani, S. Cocco, K. Jezek, and R. Monasson, "Position is coherently represented during flickering instabilities of place-cell cognitive maps in the hippocampus," 26th Annual Computational Neuroscience Meeting (CNS 2017), 2017.

[181] S. Lin and D. Gervasoni, "Defining global brain states using multielectrode field potential recordings," in Methods for Neural Ensemble Recordings (N. MAL, ed.), CRC Press/Taylor and Francis, 2008.

[182] R. Monasson and S. Rosay, "Transitions between spatial attractors in place-cell models," Physical Review Letters, vol. 115, p. 09810, 2015.

[183] W. Truccolo, U. Eden, M. Fellows, J. Donoghue, and E. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble,and extrinsic covariate effects," J. Neurophysiol., vol. 93, pp. 1071–89, 2005.

[184] A. Fenton and R. Muller, "Place cell discharge is extremely vairiable during individual passes of the rat through the firing field," Proc. Natl. Acad. Sci. USA, vol. 95, pp. 3182–3187, 1998.

[185] V. A. Makarov, F. Panetsos, and O. de Feo, "A method for determining neural connectivity and inferring the under- lying network dynamics using extracellular spike recordings," J. Neurosci. Methods, vol. 244, p. 165, 2005.

[186] R. Monasson and S. Cocco, "Fast inference of interactions in assemblies of stochastic integrate-and-fire neurons from spike recordings," J. Comp. Neurosci., vol. 31, pp. 199–227, 2011.

[187] S. Koyama and L. Paninski, "Efficient computation of the maximum a posteriori path and parameter estimation in integrate-and-fire and more general state-space models," J. Comp. Neurosci., vol. 29, p. 89, 2010.

[188] J. J. Knierim, "Neural representations of location outside the hippocampus," Learning & Memory, vol. 13, no. 4, pp. 405–415, 2006.

[189] A. Treves and E. T. Rolls, "Computational analysis of the role of the hippocampus in memory," Hippocampus, vol. 4, no. 3, pp. 374–391, 1994.

[190] D. Garcia, "Robust smoothing of gridded data in one and higher dimensions with missing values," Computational statistics & data analysis, vol. 54, no. 4, pp. 1167–1178, 2010.

[191] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," Radiology, vol. 143, no. 1, pp. 29–36, 1982.

[192] C. E. Metz, "Basic principles of roc analysis," in Seminars in nuclear medicine, vol. 8, pp. 283–298, Elsevier, 1978.

[193] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," Pattern recognition, vol. 30, no. 7, pp. 1145–1159, 1997.

[194] M. J. Pencina, R. B. D'Agostino, R. B. D'Agostino, and R. S. Vasan, "Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond," Statistics in medicine, vol. 27, no. 2, p. 157, 2008.

[195] J. R. Manns and H. Eichenbaum, "A cognitive map for object memory in the hippocampus," Learning & Memory, vol. 16, no. 10, pp. 616–624, 2009.

[196] M. I. Anderson and K. J. Jeffery, "Heterogeneous modulation of place cell firing by changes in context," Journal of Neuroscience, vol. 23, no. 26, pp. 8827–8835, 2003.

[197] E. Bostock, R. U. Muller, and J. L. Kubie, "Experience-dependent modifications of hippocampal place cell firing," Hippocampus, vol. 1, no. 2, pp. 193–205, 1991.

[198] E. T. Rolls, "An attractor network in the hippocampus: theory and neurophysiology," Learning & Memory, vol. 14, no. 11, pp. 714–731, 2007.

[199] S. J. Guzman, A. Schlögl, M. Frotscher, and P. Jonas, "Synaptic mechanisms of pattern completion in the hippocampal ca3 network," Science, vol. 353, pp. 1117–1123, 2016.

[200] S.-i. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," Biological cybernetics, vol. 27, no. 2, pp. 77–87, 1977.

[201] J. J. Hopfield, "Neurodynamics of mental exploration," Proceedings of the National Academy of Sciences, vol. 107, no. 4, pp. 1648–1653, 2010.

[202] M. Geva-Sagiv, S. Romani, L. Las, and N. Ulanovsky, "Hippocampal global remapping for different sensory modalities in flying bats," Nature neuroscience, vol. 19, no. 7, p. 952, 2016.

[203] O. Haas, J. Henke, C. Lebold, and K. Thurley, "Distinct subpopulations of locomotion- and vision-induced place fields in the hippocampus," Submitted, vol. 16, no. 24, pp. 8027–8040, 2019.

[204] E. N. Brown, L. M. Frank, D. Tang, M. C. Quirk, and M. A. Wilson, "A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells," The Journal of Neuroscience, vol. 18, no. 18, pp. 7411–7425, 1998.

[205] E. Schneidmann, M. J. Berry, R. Segev, and W. Bialek, "Weak pairwise correlations imply strongly correlated network states in a population," Nature, vol. 440, pp. 1007–1012, 2006.

[206] S. Mark, S. Romani, K. Jezek, and M. Tsodyks, "Theta-paced flickering between place-cell maps in the hippocampus: A model based on short-term synaptic plasticity," Hippocampus, vol. 27, 2017.

[207] G. B. Keller, T. Bonhoeffer, and M. Hübener, "Sensorimotor mismatch signals in primary visual cortex of the behaving mouse," Neuron, vol. 74, no. 5, pp. 809–815, 2012.

[208] M. C. Fuhs and D. S. Touretzky, "A spin glass model of path integration in rat medial entorhinal cortex," Journal of Neuroscience, vol. 26, no. 16, pp. 4266–4276, 2006.

[209] C. Schmidt-Hieber, G. Toleikyte, L. Aitchison, A. Roth, B. A. Clark, T. Branco, and M. Häusser, "Active dendritic integration as a mechanism for robust and precise grid cell firing," Nature neuroscience, vol. 20, no. 8, p. 1114, 2017.

[210] F. Stella, E. Cerasti, B. Si, K. Jezek, and A. Treves, "Self-organization of multiple spatial and context memories in the hippocampus," Neuroscience & Biobehavioral Reviews, vol. 36, no. 7, pp. 1609–1625, 2012.

[211] L. L. Colgin, T. Denninger, M. Fyhn, T. Hafting, T. Bonnevie, O. Jensen, M.-B. Moser, and E. I. Moser, "Frequency of gamma oscillations routes flow of information in the hippocampus," Nature, vol. 462, no. 7271, p. 353, 2009.

[212] K. M. Igarashi, L. Lu, L. L. Colgin, M.-B. Moser, and E. I. Moser, "Coordination of entorhinal–hippocampal ensemble activity during associative learning," Nature, vol. 510, no. 7503, p. 143, 2014.

[213] M. Gil, M. Ancau, M. I. Schlesiger, A. Neitz, K. Allen, R. J. De Marco, and H. Monyer, "Impaired path integration in mice with disrupted grid cell firing," Nature neuroscience, vol. 21, no. 1, p. 81, 2018.

[214] F. Savelli, J. Luck, and J. J. Knierim, "Framing of grid cells within and beyond navigation boundaries," eLife, vol. 6, 2017.

[215] J. D. Monaco and L. F. Abbott, "Modular realignment of entorhinal grid cell activity as a basis for hippocampal remapping," Journal of Neuroscience, vol. 31, no. 25, pp. 9414–9425, 2011.

[216] T. Bonnevie, B. Dunn, M. Fyhn, T. Hafting, D. Derdikman, J. L. Kubie, Y. Roudi, E. I. Moser, and M.-B. Moser, "Grid cells require excitatory drive from the hippocampus," Nature neuroscience, vol. 16, no. 3, pp. 309–317, 2013.

[217] K. Hardcastle, S. Ganguli, and L. M. Giocomo, "Environmental boundaries as an error correction mechanism for grid cells," Neuron, vol. 86, no. 3, pp. 827–839, 2015.

[218] D. J. MacKay, Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003.

[219] H. Akaike, "A new look at the statistical model identification," IEEE transactions on automatic control, vol. 19, no. 6, pp. 716–723, 1974.

[220] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," IEEE Transactions on pattern analysis and machine intelligence, no. 6, pp. 721–741, 1984.

[221] W. Truccolo, L. R. Hochberg, and J. P. Donoghue, "Collective dynamics in human and monkey sensorimotor cortex: predicting single neuron spikes," Nature neuroscience, vol. 13, no. 1, p. 105, 2010.

[222] J. Humplik and G. Tkačik, "Probabilistic models for neural populations that naturally capture global coupling and criticality," PLoS computational biology, vol. 13, no. 9, p. e1005763, 2017.

[223] T. Gueudré, C. Baldassi, M. Zamparo, M. Weigt, and A. Pagnani, "Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis," Proceedings of the National Academy of Sciences, vol. 113, no. 43, pp. 12186–12191, 2016.

[224] C. Feinauer, H. Szurmant, M. Weigt, and A. Pagnani, "Inter-protein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the trp operon," PloS one, vol. 11, no. 2, p. e0149166, 2016.

[225] M. J. Skwark, N. J. Croucher, S. Puranen, C. Chewapreecha, M. Pesonen, Y. Y. Xu, P. Turner, S. R. Harris, S. B. Beres, J. M. Musser, et al., "Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis," PLoS genetics, vol. 13, no. 2, p. e1006508, 2017.

[226] A. Coucke, High dimensional inference with correlated data: statistical modeling of protein sequences beyond structural prediction. PhD thesis, PSL Research University, 2016.

[227] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, "Inverse statistical physics of protein sequences: a key issues review," Reports on Progress in Physics, vol. 81, no. 3, p. 032601, 2018.

[228] U. Göbel, C. Sander, R. Schneider, and A. Valencia, "Correlated mutations and residue contacts in proteins," Proteins: Structure, Function, and Bioinformatics, vol. 18, no. 4, pp. 309–317, 1994.

[229] E. Neher, "How frequent are correlated changes in families of protein sequences?," Proceedings of the National Academy of Sciences, vol. 91, no. 1, pp. 98–102, 1994.

[230] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, "Genomics-aided structure prediction," Proceedings of the National Academy of Sciences, vol. 109, no. 26, pp. 10340–10345, 2012.

[231] T. Nugent and D. T. Jones, "Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis," Proceedings of the National Academy of Sciences, vol. 109, no. 24, pp. E1540–E1547, 2012.

[232] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, "Three-dimensional structures of membrane proteins from genomic sequencing," Cell, vol. 149, no. 7, pp. 1607–1621, 2012.

[233] A. Lapedes, B. Giraud, and C. Jarzynski, "Using sequence alignments to predict protein structure and stability with high accuracy," arXiv preprint arXiv:1207.2484, 2012.

[234] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, "Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1," Molecular biology and evolution, vol. 33, no. 1, pp. 268–280, 2015.

[235] E. Aurell, "The maximum entropy fallacy redux?," PLoS computational biology, vol. 12, no. 5, p. e1004777, 2016.

[236] E. van Nimwegen, "Inferring contacting residues within and between proteins: what do the probabilities mean?," PLoS computational biology, vol. 12, no. 5, p. e1004726, 2016.

[237] H. Li, R. Helling, C. Tang, and N. Wingreen, "Emergence of preferred structures in a simple model of protein folding," Science, vol. 273, no. 5275, pp. 666–669, 1996.

[238] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," Macromolecules, vol. 22, no. 10, pp. 3986–3997, 1989.

[239] E. Shakhnovich, M. Karplus, et al., "How does a protein fold?," nature, vol. 369, no. 6477, p. 248, 1994.

[240] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, "Protein stability promotes evolvability," Proceedings of the National Academy of Sciences, vol. 103, no. 15, pp. 5869–5874, 2006.

[241] J. P. Barton, A. K. Chakraborty, S. Cocco, H. Jacquin, and R. Monasson, "On the entropy of protein families," Journal of Statistical Physics, vol. 162, no. 5, pp. 1267–1293, 2016.

[242] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation," Macromolecules, vol. 18, no. 3, pp. 534–552, 1985.

[243] E. Shakhnovich and A. Gutin, "Enumeration of all compact conformations of copolymers with random sequence of links," The Journal of Chemical Physics, vol. 93, no. 8, pp. 5967–5971, 1990.

[244] F. Rizzato, A. Coucke, J. P. Barton, E. de Leonardis, J. Tubiana, R. Monasson, and S. Cocco, "Benchmarking inference of graphical models on erdoos-renyi graphs: how to reduce the number of parameters with no information loss," Draft, 2018.

[245] C.-Y. Gao, H.-J. Zhou, and E. Aurell, "Correlation-compressed direct-coupling analysis," Physical Review E, vol. 98, no. 3, p. 032407, 2018.

[246] S. Puranen, M. Pesonen, J. Pensar, Y. Y. Xu, J. A. Lees, S. D. Bentley, N. J. Croucher, and J. Corander, "Superdca for genome-wide epistasis analysis," Microbial genomics, vol. 4, no. 6, 2018.

[247] N. Bulso, M. Marsili, and Y. Roudi, "Sparse model selection in the highly under-sampled regime," Journal of Statistical Mechanics: Theory and Experiment, vol. 2016, no. 9, p. 093404, 2016.

[248] S. Grigolon, S. Franz, and M. Marsili, "Identifying relevant positions in proteins by critical variable selection," Molecular BioSystems, vol. 12, no. 7, pp. 2147–2158, 2016.

[249] J. H. Friedman, "On bias, variance, 0/1—loss, and the curse-of-dimensionality," Data mining and knowledge discovery, vol. 1, no. 1, pp. 55–77, 1997.

[250] F. Cucker and S. Smale, "Best choices for regularization parameters in learning theory: on the bias-variance problem," Foundations of computational Mathematics, vol. 2, no. 4, pp. 413–428, 2002.

[251] L. Asti, G. Uguzzoni, P. Marcatili, and A. Pagnani, "Maximum-entropy models of sequenced immune repertoires predict antigen-antibody affinity," PLoS computational biology, vol. 12, no. 4, p. e1004870, 2016.

[252] J. K. Mann, J. P. Barton, A. L. Ferguson, S. Omarjee, B. D. Walker, A. Chakraborty, and T. Ndung'u, "The fitness landscape of hiv-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing," PLoS computational biology, vol. 10, no. 8, p. e1003776, 2014.

[253] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," Neural computation, vol. 4, no. 1, pp. 1–58, 1992.

[254] "Hippocampal reactivations of open environments are described by brownian diffusion." 2018.

[255] U. Ferrari, S. Deny, M. Chalk, G. Tkačik, O. Marre, and T. Mora, "Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons," Physical Review E, vol. 98, no. 4, p. 042410, 2018.

[256] T.-A. Nghiem, B. Telenczuk, O. Marre, A. Destexhe, and U. Ferrari, "Maximum-entropy models reveal the excitatory and inhibitory correlation structures in cortical neuronal activity," Physical Review E, vol. 98, no. 1, p. 012402, 2018.

[257] L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, "Collective behavior of place and non-place neurons in the hippocampal network," Neuron, vol. 96, no. 2, pp. 1178–1191, 2017.

[258] J. Tubiana and R. Monasson, "Emergence of compositional representations in restricted boltzmann machines," Physical review letters, vol. 118, no. 13, p. 138301, 2017.

[259] J. Tubiana, S. Cocco, and R. Monasson, "Learning protein constitutive motifs from sequence data," arXiv preprint arXiv:1803.08718, 2018.

[260] S. A. Hollup, S. Morlen, J. G. Donnett, M.-B. Moser, and E. I. Moser, "Accumulation of hippocampal place fields at the goal location in an annular watermaze task," The Journal of Neuroscience, vol. 21, pp. 1635–1644, 2001.

RÉSUMÉ

L'avènement récent des procédures expérimentales à haut débit a ouvert une nouvelle ère pour l'étude quantitative des systèmes biologiques. De nos jours, les enregistrements d'électrophysiologie et l'imagerie du calcium permettent l'enregistrement simultané in vivo de centaines à des milliers de neurones. Parallèlement, grâce à des procédures de séquençage automatisées, les bibliothèques de protéines fonctionnelles connues ont été étendues de milliers à des millions en quelques années seulement. L'abondance actuelle de données biologiques ouvre une nouvelle série de défis aux théoriciens. Des méthodes d'analyse précises et transparentes sont nécessaires pour traiter cette quantité massive de données brutes en observables significatifs. Parallèlement, l'observation simultanée d'un grand nombre d'unités en interaction permet de développer et de valider des modèles théoriques visant à la compréhension mécanistique du comportement collectif des systèmes biologiques. Dans ce manuscrit, nous proposons une approche de ces défis basée sur des méthodes et des modèles issus de la physique statistique, en développent et appliquant ces méthodes au problèmes issu de la neuroscience et de la bio-informatique : l'étude de la mémoire spatiale dans le réseau hippocampique, et la reconstruction du paysage adaptatif local d'une protéine.

MOTS CLÉS

Inférence, Physique Statistique, Neuroscience, Hippocampe, Protéines, Bio-informatique, Modèle d'Ising.

ABSTRACT

The recent advent of high-throughput experimental procedures has opened a new era for the quantitative study of biological systems. Today, electrophysiology recordings and calcium imaging allow for the in vivo simultaneous recording of hundreds to thousands of neurons. In parallel, thanks to automated sequencing procedures, the libraries of known functional proteins expanded from thousands to millions in just a few years. This current abundance of biological data opens a new series of challenges for theoreticians. Accurate and transparent analysis methods are needed to process this massive amount of raw data into meaningful observables. Concurrently, the simultaneous observation of a large number of interacting units enables the development and validation of theoretical models aimed at the mechanistic understanding of the collective behavior of biological systems. In this manuscript, we propose an approach to both these challenges based on methods and models from statistical physics. We present an application of these methods to problems from neuroscience and bioinformatics, focusing on (1) the spatial memory and navigation task in the hippocampal loop and (2) the reconstruction of the fitness landscape of proteins from homologous sequence data.

KEYWORDS

Bayesian Inference, Statistical Physics, Ising Model, Neuroscience, Hippocampus, Attractor Neural Network, Proteins, Fitness Landscape.