

Bridging the Explanatory Gap between Neuronal Activity and Cognition by Multilevel Causal Modeling

Moritz Grosse-Wentrup, Akshey Kumar, Anja Meunier,
and Manuel Zimmer

March 9, 2023

Abstract

Explaining how neuronal activity gives rise to cognition arguably remains the most significant challenge in cognitive neuroscience. We introduce **neuro-cognitive multilevel causal modeling** (NC-MCM), a mathematical framework that bridges the explanatory gap between neuronal activity and cognition by construing **cognitive states as causally consistent abstractions of neuronal states**. Multilevel causal modeling allows us to consistently reason about the neuronal- and cognitive causes of behavior while maintaining a **physicalist (in contrast to a dualist)** position. We introduce an **algorithm for learning cognitive level causal models from neuronal activation patterns** and demonstrate its ability to learn cognitive states of the nematode *C. elegans* from calcium imaging data. Specifically, we show that the NC-MCM framework is able to derive **empirically testable predictions on causal relations between cognitive states** (and their neuronal representations) and behaviors in an automated, i.e., purely data-driven, fashion. This ability is essential for moving beyond the study of simple organisms to more complex biological systems whose complexity prohibits the development of hand-crafted computational models. We conclude the article by discussing future extensions and the philosophical implications of the NC-MCM framework.

1 Introduction

At least since the work of David Marr (1982) it is widely acknowledged that complex systems can be described at different levels, e.g., at the implementational, the algorithmic, and the computational level. Understanding the neuronal basis of cognition amounts to bridging the explanatory gap between the level of neuronal activity patterns and the level of cognitive states (Changeux and Dehaene, 1989). However, despite the long history of research into this problem, a mathematically rigorous framework for bridging this gap is still lacking.

The authors would like to thank the following colleagues (in alphabetical order) for their insightful comments on an earlier version of the manuscript: Jozsef Arato, Dirk Bernhardt-Walther, Mauricio Gonzalez Soto and Kevin Reuter.

Traditionally, observed statistical associations between single-unit neuronal activities and behaviors have formed the basis for identifying neuronal circuits and hand-crafting mechanistic models of their computations (cf. Borst (2014)). Due to increasing awareness of the importance of (potentially widely distributed) neuronal activity patterns for behavior and the difficulty in scaling up hand-crafted models to large-scale neuronal recordings, machine learning methods (also referred to as decoding- or multivariate pattern analysis (MVPA) models) are increasingly used to uncover relations between complex neuronal activity patterns, cognitive states, and behaviors (Norman et al., 2006; Mitchell et al., 2008; Pereira et al., 2009). However, the ability to decode behavior from complex neuronal activity patterns does not imply that we have revealed their representational contents (Ritchie et al., 2020). Under certain conditions, decoding models can be endowed with a causal interpretation, enabling experimentally testable predictions on causal relations between neuronal activity and behavior (Weichwald et al., 2015). Such causal models are not suitable though to study relationships between neuronal activity patterns and *cognition*, because causal relationships between these two levels would be irreconcilable with physicalism: A causal relation between two variables $X \rightarrow Y$ implies that manipulations of X can alter the probability of observing specific values of Y (Pearl, 2000). This relation is only possible if X and Y are two separate processes that are linked by a mechanism (cf. Spirtes (2009) for an in-depth discussion of the nature of causal variables). A causal relation between X and Y thus implies that it must be conceptually possible to intervene on Y without intervening on X , thereby breaking the mechanism that links the two variables. For instance, if X and Y represent two different neurons, one could construct an experimental setup that controls the membrane potential of Y , thereby eliminating any influence action potentials of X can have on the membrane potential of Y . If X and Y were to represent a neuronal- and a cognitive state, however, a causal relation of the form $X \rightarrow Y$ would imply that we can (at least conceptually) manipulate the cognitive state without changing the neuronal state that causes it. As such, causal relations between neuronal- and cognitive states would only be meaningful if we adopted a dualistic viewpoint in which neuronal- and cognitive states co-existed as independent physical processes. In the present work, we adhere to physicalism, i.e., we adopt the view that all mental phenomena are ultimately physical phenomena and hence reject the notion that neuronal activity causes cognition. Instead, we argue in the following that it is reasonable to consider relationships between neuronal activity and cognitive states as *constitutive* or, in more philosophical terms, that cognitive states supervene on neuronal states.

While we consider causal models unsuitable for representing relations between neuronal- and cognitive states, we consider them well-suited to represent relations within each level. We first consider the neuronal level. Building on the example in the previous paragraph, it is perfectly reasonable to consider two neurons as separate entities that can be manipulated individually and that are linked by a physical mechanism. For instance, the sentence *neuron X's action potential is a cause of neuron Y's membrane potential* is a meaningful causal statement, in the sense that it implies an empirically testable prediction that is

consistent with our physical understanding of neuronal processes: Manipulating the spike probability of neuron X , e.g., by injecting a current into neuron X , will alter the membrane potential of neuron Y . We argue that the same holds true for relations between cognitive states. We employ cognitive states to causally reason about our own as well as other people's mental processes. For instance, when we say *I am unhappy because I am bored*, we express the causal relation $boredom \rightarrow happiness$. This causal statement implies the empirically testable prediction that if we reduce boredom in a person we increase their probability of being happy. It is further consistent with the view that *boredom* and *happiness* are two separate processes that can be manipulated independently.

If causal models are well-suited to explain relations within the neuronal and cognitive levels but unsuitable for explaining relations between these two levels, which type of relations should hold between causal models on the neuronal- and causal models on the cognitive level? The solution we advocate here is causally consistent transformations between causal models on each level (Rubenstein et al., 2017). Causally consistent transformations are functional mappings between states and interventions of two causal models, chosen so that we can reason about the observational effects of causal interventions consistently across both models. For instance, assume the cognitive concepts *boredom* and *happiness* to have the two neuronal realizations \mathbf{x} and \mathbf{y} (with the bold notation indicating that \mathbf{x} and \mathbf{y} may represent elements of high-dimensional sets of neuronal activity patterns), with the causal relations $boredom \rightarrow happiness$ and $\mathbf{x} \rightarrow \mathbf{y}$. Intervening on \mathbf{x} , e.g., by electrical stimulation, would then be equivalent to modulating *boredom* and, via the causal effect \mathbf{x} exerts on \mathbf{y} , affect *happiness*. In this conceptual framework, which we term *multilevel causal modeling (MCM)*, neuronal cause-effect relations play out in parallel to cognitive cause-effect relations in a consistent manner that enables us to causally reason and explain observations interchangeably on each level. In the MCM framework, neuronal- and cognitive states thus do not co-exist as separate processes but are linked via functional mappings. In other words, neuronal states do not cause but are constitutive of cognitive states. The MCM framework thus provides a theoretical framework to bridge the explanatory gap between neuronal activity and cognition while maintaining a physicalist position.

When attempting to map neuronal- to cognitive states, we must consider that cognitive states are not universal but depend (among other things) on the cultural context (Quinn and Holland, 1987). When attempting to map neuronal- to cognitive states, one question thus seems of particular importance: What is the set of possible cognitive states? Is there for example a single cognitive state of being in pain, or should having a headache and feeling back pain be considered two different states? Do all individuals share the same sets of cognitive states? And if so, are the mappings from neuronal to cognitive level equivalent across subjects? The MCM framework does not assume any fixed set of cognitive states, and we do not argue for the existence of such a universal set. Instead, and in agreement with a recent argument by Krakauer et al. (2017) that neuroscience needs behavior, we let the behavioral context determine in a purely data-driven approach the relevant set of cognitive states.

We illustrate the relationship between cognition, neuronal activity, and behavior in the MCM framework with an example from physics. The temperature of a gas is determined by the average kinetic energy of its molecules. To reason about the conditions under which the gas ignites, both levels of description are equivalent, e.g., the two sentences *the gas will ignite if its temperature is raised by T degrees Celsius* and *the gas will ignite if the average kinetic energy of its molecules is increased by K Joules* are causally meaningful and consistent statements. Importantly, many kinetic energy configurations give rise to the same temperature. The gas temperature can not be altered, however, without also changing the kinetic energies. As such, the macroscopic concept of temperature is a causally consistent abstraction of the microscopic kinetic energy configuration with respect to the behavior of the gas. In the MCM framework, the analogies of temperature, kinetic energies, and ignition are cognitive state, neuronal states, and behaviors, in the sense that the macroscopic cognitive states are causally meaningful and consistent abstractions of the microscopic neuronal states that cause behavior.

In the following, we introduce the MCM framework in a mathematically rigorous fashion and show how causally consistent transformations between the neuronal- and the cognitive level can be learned from empirical data (Section 2). In Section 3, we illustrate the application of the MCM framework on calcium imaging data from the nematode *C. elegans*. We conclude our work in Section 4 by discussing future extensions and the philosophical implications of the MCM framework.

2 Multilevel causal modeling in cognitive neuroscience

We begin this section by discussing the construction of causal models on the neuronal level and elucidating how they are linked to behavior. We then consider the nature of cognitive states and causal models thereof, before explaining how causal models on the neuronal- and on the cognitive level can be linked via causally consistent transformations. We conclude the section by discussing how to learn MCMs from empirical data.

2.1 Causal models on the neuronal level

To construct causal models on the neuronal level, we first need to decide on a framework for causal modeling. A variety of causal modeling frameworks have been developed for and evaluated on neuronal data (Smith et al., 2011). To be applicable in the MCM framework, we require causal models that can make empirically testable predictions on the effects of experimental interventions. The framework of Causal Bayesian Networks (CBNs) (Pearl, 2000; Spirtes et al., 2000) fulfills this requirement and hence serves as the causal modeling framework in the present setting.

Causal relations in CBNs are modeled by structural causal models (SCMs):

Definition 1 (Structural Causal Model – SCM) A SCM is a triple $\mathcal{M}_X := \{\mathbf{X}, \mathbf{E}, \mathbf{F}\}$ with \mathbf{X} a set of N random variables endogenous to the model, \mathbf{E} a set of N exogenous noise variables, and \mathbf{F} a set of N functions defining each endogenous variable as a function of its direct causes (i.e., parents – $pa()$) and its corresponding exogenous noise variable, so that for each $i \in \{1, \dots, N\}$ we have $X_i := F_i(pa(X_i), E_i)$ where the F_i are chosen such that no variable is a (direct or indirect) cause of itself.

The joint probability distribution $P(\mathbf{E})$ over the exogenous noise variables induces a joint probability distribution $P(\mathbf{X})$ over the endogenous variables (via the pushforward measure). If the noise variables are mutually independent the SCM is *causally sufficient*, i.e., no unobserved confounders that influence multiple endogenous variables exist. In the present setting, each endogenous variable X_i represents the activity of one neuron. We note that in general \mathbf{X} and \mathbf{E} may represent continuous- as well as discrete variables, implying that $P(\mathbf{X})$ and $P(\mathbf{E})$ denote probability densities or distributions, respectively. In the following, we assume all variables to be discrete. We consider an extension of the MCM framework to continuous-valued random variables feasible but beyond the scope of the present work. We denote random variables by upper- and their realizations by lower-case letters, e.g., we write $P(X_i = x)$ to indicate the probability that the neuron represented by the random variable X_i takes on the value $x \in \mathcal{X}$.

Causal relations in SCMs are commonly depicted by directed acyclic graphs (DAGs), with an arrow drawn from node A into node B if the endogenous variable represented by node A is a parent of the endogenous variable represented by node B . An arrow from A into B indicates a direct causal influence of A on B , and a directed path from node A to node B indicates an indirect causal influence of A and B . We note that direct- and indirect causal relations are relative to the set of variables endogenous to the model and may change when dropping or adding nodes.

Knowledge of the DAG in combination with the joint probability distribution over all endogenous variables (represented by the nodes of the DAG) enables us to reason about the probabilistic effects of experimental interventions on variable subsets (Pearl, 2000). Experimental interventions are represented mathematically by the $do()$ -operator, e.g., $do(X_i = x)$ represents the experimental intervention of setting variable X_i to the value x . In empirical settings, the DAG and the joint probability distribution are usually not known and have to be inferred from a combination of experimental and observational data. We show in Section 2.3 though that knowledge of the causal model on the neuronal level is not a prerequisite for constructing a causally consistent mapping between the neuronal- and the cognitive level.

In their original form, SCMs model independent and identically distributed (i.i.d.) data, i.e., data without temporal structure. It is straight-forward, however, to extend SCMs to model dynamical neuronal systems by unrolling the endogenous variables across time, as exemplified in the middle row of Figure 1. Here, the set of random variables $\mathbf{X}[t_0]$ represents the global state of all N neu-

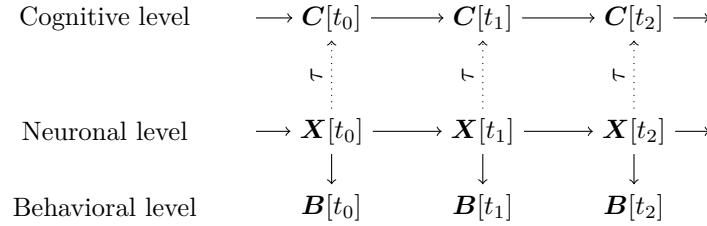


Figure 1: Relations between cognitive-, neuronal-, and behavioral states in MCM. Solid and dotted arrows denote causal and constitutive relations, respectively.

rons at time t_0 and the arrow from $\mathbf{X}[t_0]$ into $\mathbf{X}[t_1]$ indicates that the global neuronal state at time t_0 is a cause of the global neuronal state at time t_1 , in the sense that the state of each neuron at time t_1 is a function of a subset of the neuronal states at time t_0 and their respective exogenous noise term¹. More formally, we define the extension of SCMs to dynamic settings as follows:

Definition 2 (Dynamic Structural Causal Model – dSCM) A dynamic SCM (dSCM) is family of T triples $\mathcal{M}_{X[t]} := \{\mathbf{X}[t], \mathbf{E}[t], \mathbf{F}[t]\}$, indexed by $t \in \{1, \dots, T\}$, with $\mathbf{X}[t]$ a set of N random variables endogenous to the model, $\mathbf{E}[t]$ a set of N exogenous noise variables, and $\mathbf{F}[t]$ a set of N functions defining each endogenous variable as a function of its direct causes (i.e., parents – $pa()$) and its corresponding exogenous noise variable, so that for each $i \in \{1, \dots, N\}$ and $t \in \{1, \dots, T\}$ we have $X_i[t] := F_i[t](pa(X_i[t]), E_i[t])$, where $pa(X_i[t])$ may contain any endogenous variables prior to time t .

We note that we do not allow instantaneous causal relations in Definition 2, i.e., situations where $pa(X_i[t_1])$ may include other neurons at time t_1 , because our current understanding of physical processes posits that causes must precede effects. This constraint may have to be reconsidered when modeling dynamical systems that are observed at sampling rates slower than the system's dynamics.

If the exogenous noise terms are mutually independent and $pa(\mathbf{X}[t])$ only includes variables at the previous time step for all t , as in the example in Figure 1, the dSCM represents a first-order Markov process with transition probability distribution

$$T_{\mathbf{X}[t]} := P(\mathbf{X}[t] | \mathbf{X}[t-1], \mathbf{X}[t-2], \dots) = P(\mathbf{X}[t] | \mathbf{X}[t-1]). \quad (1)$$

Markov processes are of particular relevance in the MCM framework, because they define causally sufficient state-spaces, in the sense that the global state

¹We remark that this representation also allows feedback loops, e.g., as in $X_i[t] \rightarrow X_j[t+1] \rightarrow X_i[t+2]$

at any given point in time is causally sufficient for the probability distribution over the global state at the next time point. More formally, Markovianity implies that the exogenous noise terms at different time points are mutually statistically independent and hence (via the backdoor criterion, see Pearl (2000)) $P(\mathbf{X}[t]|\mathbf{X}[t-1]) = P(\mathbf{X}[t]|\text{do}(\mathbf{X}[t-1]))$ for all t , i.e., the observational-and interventional conditional distributions are identical. A counter example would be the case of an exogenous variable, e.g., a stimulus, affecting both $\mathbf{X}[t-2]$ and $\mathbf{X}[t]$. In this case, the Markov property would be violated ($\mathbf{X}[t-2] \not\perp\!\!\!\perp \mathbf{X}[t]|\mathbf{X}[t-1]$) and hence $P(\mathbf{X}[t]|\mathbf{X}[t-1]) \neq P(\mathbf{X}[t]|\text{do}(\mathbf{X}[t-1]))$). As we discuss in Section 2.3, empirical tests for Markovianity are essential to ensure that cognitive variables are causally meaningful.

In the following, we assume that the functional mappings between states as well as the probability distribution over their exogenous variables do not change over time, i.e., we assume the process to be time-homogeneous. We further assume that any process we consider has converged to a stationary probability distribution. These assumptions are not required from a theoretical perspective but greatly simplify empirical inference.

To model the causal effect of neuronal activity on behavior, we extend the DAG in Figure 1 by a behavioral state vector $\mathbf{B}[t]$ and let $\mathbf{X}[t] \rightarrow \mathbf{B}[t]$, i.e., we model the behavioral states at time t to be caused by (a subset of) the neuronal states at time t . We note that it is reasonable to consider relations between neuronal activity and behavior as causal, because we can in principle manipulate behavior independently of neuronal activity, e.g., we can fixate an animal and thereby prevent neuronal activity from causing an actual movement. We further note that we represent the behavioral states across time by measurement nodes, i.e., by nodes that have no causal effect on any other variables (Markham and Grosse-Wentrup, 2020). As such, the behavioral states do not form a Markov process and $\mathbf{B}[t] \not\perp\!\!\!\perp \mathbf{B}[t-2]|\mathbf{B}[t-1]$ due to the common effect of $\mathbf{X}[t-2]$ on both $\mathbf{B}[t-2]$ and $\mathbf{B}[t]$. We denote the extension of the dSCM by the behavioral state vector as $\mathcal{M}_{\mathbf{X}[t]}^{\mathbf{B}[t]}$.

To model feedback loops of the neuronal system with its environment, the DAG could be further extended by a state vector $\mathbf{S}[t]$ that represents stimuli, which could be influenced by the system's past behavior, e.g., $\mathbf{S}[t_0] \rightarrow \mathbf{X}[t_0] \rightarrow \mathbf{B}[t_0] \rightarrow \mathbf{S}[t_1]$. We leave such an extension for future work.

2.2 Bridging the neuronal- and the cognitive level

Before discussing how to bridge the neuronal- and the cognitive level, we must first consider the nature of cognitive states. Somewhat surprisingly, there exists no generally agreed upon definition of what a cognitive state is (cf. Prinz (2004), pp. 41 ff.). The broader term *cognition* is commonly used to denote *any kind of mental operation or structure that can be studied in precise terms* (Lakoff and Johnson, 1999). To arrive at a working definition that allows us to operationalize cognitive states for the MCM framework, we consider three refinements of this concept of cognition. First, we propose that a cognitive state

must be mathematically quantified to be *studied in precise terms*. Second, we consider an *operation* to refer to a causally meaningful process, in the sense that the operation represents a mechanism that translates a system from one state to another. And third, we consider the addendum *or structure* to indicate that any mental operation is supported by a physical substrate. These considerations lead us to the following working definition, which we will render mathematically precise towards the end of this section:

Definition 3 (Cognitive state) *A cognitive state is a causally meaningful abstraction of a neuronal state.*

To elucidate this definition, we revisit the physical analogy from the introduction. The temperature of a gas is a macroscopic *abstraction* of the microscopic states of the gas molecules, because there are infinitely many configurations of kinetic energies of the molecules in the gas that give rise to the same temperature. This relation, however, is asymmetric. It is impossible to change the temperature of a gas without altering the kinetic energies of its molecules. Accordingly, the temperature of a gas and its molecular configuration do not stand in a causal but in a constitutive relationship. The concept of temperature is a *causally meaningful* abstraction because it allows us to reason about the behavior of the gas under experimental interventions, e.g., as in the statement *the gas will ignite if we raise its temperature above a certain threshold*. These interventions are grounded in the microscopic level, i.e., the experimental intervention of raising the temperature is meaningful because there exist microscopic interventions (increasing kinetic energies) that are identical to raising the temperature. Definition 3 expresses the notion that cognitive- and neuronal states stand in a similar constitutive relationship. Consider the cognitive state of being bored. This state is causally meaningful, because, first, being bored can have an effect on other cognitive states, e.g., an experimental intervention that reduces boredom may have the causal effect of increasing happiness, and, second, being bored can alter our behavior, e.g., it may lead us to seek out certain (positively stimulating) behaviors. To generalize from this example, the first aspect of Definition 3, that cognitive states are *causally meaningful*, expresses our understanding that we employ cognitive states to reason about mental processes because we consider them to be causally effective. Conversely, we argue that any state that does not stand in a cause-effect relationship to another cognitive state or to a behavior is not helpful to reason about mental processes and should therefore be eliminated from a cognitive ontology. To illustrate the second aspect of Definition 3, that cognitive states are *abstractions of neuronal states*, we note that we can represent an organism's behavior at different levels of granularity. For instance, we may merely distinguish whether an animal is exploring or hunting in its environment, or we may further differentiate the individual movements an animal is performing during the exploration and the hunting processes. In the former case, we would consider all neuronal states that are equally likely to cause an exploratory vs. a hunting behavior as constituting one cognitive state. In the latter case, we would enlarge the cognitive

state space to be able to distinguish all sets of neuronal states that are equally likely to cause individual movements. In both cases, the cognitive state spaces serve as abstractions of neuronal activity patterns, albeit at different levels of granularity. As in the gas-temperature example, this relationship is asymmetric. There can be many neuronal states that give rise to the same cognitive state. It is impossible to change the cognitive state, however, without also altering the neuronal state. We emphasize that we do not consider the space of cognitive states as constant for a given model system or organism. Instead, we let the granularity at which an organism's behavior in an environmental context is represented determine the appropriate cognitive state space.

We now formalize the relations discussed above. In analogy to causal models on the neuronal level (cf. Section 2.1), we represent the cognitive state of a model system or organism by a cognitive dSCM $\mathcal{M}_{C[t]} := \{\mathbf{C}[t], \mathbf{E}[t], \mathbf{F}[t]\}$, with $\mathbf{C}[t]$ the cognitive state vector (note that $\mathbf{E}[t]$ and $\mathbf{F}[t]$ are specific to each model and not shared across neuronal- and cognitive dSCMs). To link the the neuronal- and the cognitive dSCM in a causally consistent manner, we require the two models to be *behaviorally* and *dynamically* causally consistent:

Definition 4 (Behavioral Causal Consistency – BCC) Let $\mathcal{M}_{X[t]}^{B[t]}$ a neuronal dSCM in a behavioral context and $\mathcal{M}_{C[t]}$ a cognitive dSCM. Denote the state space of $X[t]$ and $C[t]$ by \mathcal{X} and \mathcal{C} , respectively. We call the triple $\{\mathcal{M}_{X[t]}^{B[t]}, \mathcal{M}_{C[t]}, \tau\}$ behaviorally causally consistent if $\tau : \mathcal{X} \mapsto \mathcal{C}$ is a surjective mapping such that for all $b \in \mathcal{B}$ and for all $x \in \mathcal{X}, c \in \mathcal{C}$ with $c = \tau(x)$ we have

$$P(\mathbf{B}[t] = b | do(X[t] = x)) = P(\mathbf{B}[t] = b | do(C[t] = c)). \quad (2)$$

If a neuronal- and a cognitive SCM are behaviorally causally consistent, every experimental intervention on the neuronal level has a matching intervention on the cognitive level, in the sense that both interventions lead to the same probability distribution over the behaviors. The surjectivity of τ ensures, first, that the cognitive level is an abstraction of the neuronal level, i.e., many neuronal states may map to the same cognitive state yet distinct cognitive states can only be linked to distinct neuronal states, and, second, that every cognitive state has at least one neuronal representation. Intuitively, the cognitive SCM constitutes a lossless compression of all information in the neuronal SCM that is causally relevant for a given set of behaviors. As such, behavioral causal consistency allows us reason interchangeably about the causes of behaviors on the neuronal- and on the cognitive level. Importantly, behavioral causal consistency endows cognitive states with a causal meaning, in the sense that cognitive interventions, for which it may be unclear how the intervention can be experimentally implemented (e.g., increase happiness), can be translated into equivalent interventions on the neuronal level, for which a well-defined experimental procedure is available (e.g., stimulate a set of neurons).

Definition 5 (Dynamic Causal Consistency – DCC) We call a triple $\{\mathcal{M}_{X[t]}^{B[t]}, \mathcal{M}_{C[t]}, \tau\}$ dynamically causally consistent if for all pairs of cognitive

states $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$

$$P(\mathbf{C}[t+1] = \mathbf{c}' | do(\mathbf{X}[t] = \mathbf{x}_1)) = P(\mathbf{C}[t+1] = \mathbf{c}' | do(\mathbf{X}[t] = \mathbf{x}_2)) \quad (3)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \tau^{-1}(\mathbf{c})$ where $\tau^{-1}(\mathbf{c})$ is the pre-image of \mathbf{c} .

In this case, we define $do(\mathbf{C}[t] = \mathbf{c}) := do(\mathbf{X}[t] = \mathbf{x} | \mathbf{x} \in \tau^{-1}(\mathbf{c}))$ and have

$$P(\mathbf{C}[t+1] = \mathbf{c}' | do(\mathbf{C}[t] = \mathbf{c})) = P(\mathbf{C}[t+1] = \mathbf{c}' | do(\mathbf{X}[t] = \mathbf{x})) \quad (4)$$

for all $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$ and $\mathbf{x} \in \tau^{-1}(\mathbf{c})$.

While behavioral causal consistency defines consistency of neuronal- and cognitive states with respect to behavior, dynamic causal consistency defines consistency with respect to the dynamics of the neuronal system. Intuitively, the cognitive SCM constitutes a lossless compression of all information in the neuronal SCM that is causally relevant for the dynamics of the neuronal system. In analogy to behavioral causal consistency, dynamic causal consistency grounds causal relations between cognitive states in the neuronal level, e.g., statements such as *reducing boredom is likely to lead to increased happiness* can be translated into the neuronal-level statement *inducing a neuronal activity pattern \mathbf{x}' that represents low boredom, e.g., by electrical stimulation, is likely to lead to increased happiness*.

We are now in a position to formally introduce the primary contribution of this work:

Definition 6 (Neuro-Cognitive Multilevel Causal Model – NC-MCM)

We call a triple $\{\mathcal{M}_{\mathbf{X}[t]}^{\mathbf{B}[t]}, \mathcal{M}_{\mathbf{C}[t]}, \tau\}$ a neuro-cognitive multilevel causal model (NC-MCM) if it is behaviorally and dynamically causally consistent.

We remark that we have chosen τ not to depend on time, because we consider a constant mapping between levels more desirable. It would be straight-forward to extend the definition to time-varying mappings, but this generalization would substantially complicate learning the mapping from empirical data. We further note that in contrast to Rubenstein et al. (2017) we only consider experimental interventions on the full neuronal- and cognitive state vectors at each time step and leave the extension of NC-MCMs to interventions on subsets of state variables for future work.

We illustrate the concept of a NC-MCM on the following toy example.

Example 1 Consider a neuronal dSCM $\mathcal{M}_{\mathbf{X}[t]}^{\mathbf{B}[t]}$ with one endogenous variable ($N = |\mathbf{X}| = 1$) and discrete state space $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$, where each state represents one out of a total of four neuronal activity patterns. Further, assume there are three distinct behaviors $\mathcal{B} = \{b_1 \text{ (e.g., move)}, b_2 \text{ (e.g., feed)}, b_3 \text{ (e.g., rest)}\}$ with conditional probabilities given the neuronal state

$P(\mathbf{B} \mathbf{X})$	x_1	x_2	x_3	x_4
b_1	0.5	0.5	0.1	0.1
b_2	0.4	0.4	0.1	0.1
b_3	0.1	0.1	0.8	0.8

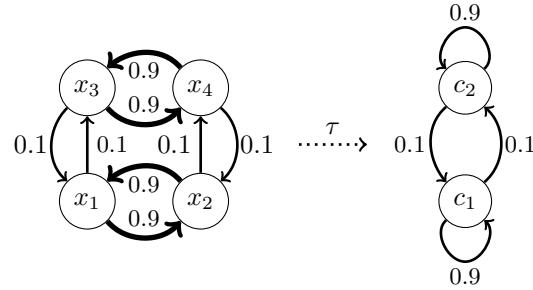


Figure 2: State-space diagram with transition probabilities for Example 1.

Note that the conditional probability distribution over the three behaviors is identical for $\{x_1, x_2\}$ and for $\{x_3, x_4\}$. As such, one may interpret the system as having two distinct macro-states, $\{x_1, x_2\}$ and $\{x_3, x_4\}$, with each macro-state subserving a particular behavioral pattern that is expressed by a constant probability distribution over the three behaviors. We may thus define a mapping

$$\tau(\mathbf{X}) = \begin{cases} c_1 & \text{if } \mathbf{X} \in \{x_1, x_2\} \\ c_2 & \text{if } \mathbf{X} \in \{x_3, x_4\} \end{cases}. \quad (5)$$

These two macro-states form the cognitive state space $\mathcal{C} = \{c_1, c_2\}$, where we may choose to name the two cognitive states foraging and recuperating, respectively. Next, assume the neuronal dSCM has the steady-state transition probability distribution $T_{\mathbf{X}}$ shown in the left-hand side of Figure 2. The transition matrix $T_{\mathbf{X}}$ and the mapping τ induce a cognitive dSCM with the transition probabilities shown in the right-hand side of Figure 2. The triple $\{\mathcal{M}_{\mathbf{X}[t]}^{B[t]}, \mathcal{M}_{\mathcal{C}[t]}, \tau\}$ is behaviorally causally consistent, because the cognitive macro-states condense all information in the neuronal micro-states that are causally relevant for the set of behaviors, i.e., for every intervention on the neuronal micro-level there is an equivalent intervention on the cognitive macro-level that entails the same probabilities over the behaviors (cf. Definition 4). The triple is further dynamically causally consistent, because the probabilities of remaining in the current or progressing to another cognitive state are identical for each neuronal state that maps to the same cognitive state. Because the triple is behaviorally and dynamically causally consistent, it fulfills the criteria for a NC-MCM. As such, the cognitive dSCM is a causally meaningful abstraction of the underlying neuronal dSCM, because causal statements on the effects of current cognitive states on behavior and on future cognitive states are grounded in the neuronal level, e.g., the cognitive-level statement ‘the animal is moving around and feeding because it is foraging’ can be translated into the neuronal-level statement ‘the animal is moving around and feeding because it is in neuronal state x_1 or x_2 ’. We note that behavioral- does not imply dynamic causal consistency, e.g., introducing an asymmetry in the neuronal-level transition probabilities in Figures 2 would break

dynamics- but not behavioral causal consistency.

We remark that the toy example above is constructed to illustrate the idea that in the NC-MCM framework redundancies in the neuronal dynamics and their relation to behaviors are exploited to build a cognitive-level model that preserves all causally relevant information. In practice, the challenge is to learn which (potentially widely-distributed and complex) neuronal activity patterns are redundant with respect to a behavioral context *and* with respect to the neuronal dynamics, a problem that we turn to in the next section.

2.3 Learning neuro-cognitive multilevel causal models

A NC-MCM model is specified by the triple $\{\mathcal{M}_{\mathbf{X}[t]}^{B[t]}, \mathcal{M}_{C[t]}, \tau\}$, a neuronal-level dSCM in a behavioral context, a cognitive-level dSCM, and a causally consistent mapping between the two. A learning problem arises when any subset of these three components is not fully specified and needs to be inferred from data. In this section, we consider the case where we have access to a set of samples $\mathcal{S} = \{\mathbf{X}[t], \mathbf{B}[t]\}_{t=1}^N$, generated by some unknown neuronal dSCM in a behavioral context, and our goal is to learn a mapping τ that induces a (behaviorally and dynamically) causally consistent cognitive dSCM. We consider this setting of particular interest, because it amounts to learning the cognitive states that a neuronal system has developed in a specific behavioral context from empirical data. Other learning problems are briefly discussed in Section 4.

Learning a mapping τ that gives rise to a causally consistent cognitive dSCM proceeds in two steps. The first step is to learn a causal model between neuronal states and behaviors. This model then forms the basis for learning the mapping τ that induces a behaviorally and dynamically consistent cognitive dSCM. In this section, we discuss these steps from a conceptual perspective. A particular instantiation of a pipeline that learns a causally consistent cognitive dSCM from empirical data \mathcal{S} is presented in Section 3.2.

Learning a causal model that relates neuronal activity to behavior amounts to learning the interventional distribution $P(\mathbf{B}[t]|\text{do}(\mathbf{X}[t]))$. The gold-standard to identify interventional distributions is by experimentation, i.e., by repeatedly setting $\mathbf{X}[t]$ to random values via an intervention and observing the induced behaviors. However, large-scale neuronal stimulation with concurrent recordings of neuronal activity and behavior remains a challenge (Yazdan-Shahmorad et al., 2016). Alternatively, we may attempt to learn the interventional distribution from observational data only. A variety of causal inference algorithms have been developed for this purpose (Pearl, 2000; Spirtes et al., 2000; Peters et al., 2017) and applied to neuronal data (Smith et al., 2011; Weichwald et al., 2015; Grosse-Wentrup et al., 2016). However, causal structure learning from observational data is also a hard problem, the computational complexity of which typically grows exponentially in the number of variables. We thus follow a third approach in which we merely attempt to learn an observational prediction model $P(\mathbf{B}[t]|\mathbf{X}[t])$ from the set of samples \mathcal{S} . We discuss the implications of

substituting an interventional model for an observational one at the end of this section.

After learning the **interventional distribution between neuronal activity and behavior** (or an observational approximation thereof), the second step is to learn the actual mapping τ . Learning this mapping again proceeds in two steps. The first step is to construct a **behaviorally consistent mapping between neuronal activity and cognitive states**. The second step is to test empirically whether the **induced cognitive dSCM is also dynamically causally consistent**. We note that designing a one-step algorithm that directly constructs a mapping that is guaranteed to be behaviorally and dynamically consistent would be preferable but is beyond the scope of the present work.

To construct a behaviorally consistent mapping, we need to find subsets of the neuronal feature space, i.e., partitions, for which $P(\mathbf{B}[t]|\mathbf{X}[t])$ is (approximately) constant. One way to solve this problem is to first learn a model that predicts the probability of each behavior as a function of the neuronal state, e.g., using a logistic regression model, a neural network model, or any other suitable modeling approach. In a second step, the predicted probabilities can then be clustered, e.g., using k-means or any other preferred clustering algorithm. The inferred clusters define a partition of the neuronal feature space with (approximately) constant conditional probabilities over the behaviors, i.e., the mapping τ . This mapping then induces a cognitive dSCM that is by construction **behaviorally consistent, with the number of cognitive states equal to the number of clusters**.

Because behavioral does not imply dynamic causal consistency, we additionally need to test whether the learned mapping τ also induces a **dynamically causally consistent cognitive dSCM**. To do so, we first note that our definition of dynamic causal consistency (Def. 5) is identical to condition (3) in Theorem 1 of Burke and Rosenblatt (1958), with the exception that former and the latter are based on interventional- and observational transition probabilities, respectively. Accordingly, we term condition (3) in Burke and Rosenblatt (1958) *dynamic observational consistency*. We can then invoke Theorem 4 of Burke and Rosenblatt (1958) to conclude that **Markovianity of the cognitive dSCM is sufficient for dynamic observational consistency**. Finally, we recall from Section 2.1 that **Markovianity of a dSCM implies that the observational- and interventional transition probabilities coincide**. As such, Markovianity of the cognitive dSCM is also sufficient for dynamic causal consistency.

To empirically test a cognitive dSCM for Markovianity, we test the **null-hypothesis $H_0 : C[t-1] \perp\!\!\!\perp C[t+1] | C[t]$** . If we find sufficient evidence against Markovianity, we reject the null-hypothesis and conclude that the cognitive dSCM is not dynamic causally consistent. Otherwise, we accept the null-hypothesis and conclude that **the cognitive dSCM induced by τ is behaviorally and dynamically causally consistent to the data-generating neuronal dSCM**. We note that if we find a cognitive dSCM not to be dynamically causally consistent, we can vary the number of clusters to tune the granularity of the cognitive dSCM and repeat the test for Markovianity.

If we base the procedure described above on the **observational distribution**

$P(\mathbf{B}[t]|\mathbf{X}[t])$ instead of on the interventional distribution $P(\mathbf{B}[t]|\text{do}\{\mathbf{X}[t]\})$, the cognitive dSCM is not guaranteed to be behaviorally causally consistent. However, the *causal coarsening theorem* states that a causal partitioning is almost always a coarsening of an observational partitioning (Chalupka et al., 2015, 2017). As such, a cognitive dSCM that is behaviorally causally consistent can be obtained from a cognitive dSCM learned on observational data by fusing subsets of cognitive states. While experimental interventions may be required to identify which cognitive states cause behaviors with identical probabilities and thus should be fused, the number of required interventions is on the order of the number of cognitive states and not on the order of (potentially orders of magnitude larger) number of neuronal states. As such, learning a cognitive dSCM on observational data first and reducing the number of cognitive states by experimental interventions afterwards is experimentally more tractable than directly constructing the interventional distribution $P(\mathbf{B}[t]|\text{do}(\mathbf{X}[t]))$ on the neuronal level. If even a reduced number of experimental interventions is not feasible, we may have to accept that a cognitive dSCM learned from observational data can only generate causal hypotheses on the relations of neuronal activities and behavioral patterns. To distinguish these models from those learned on interventional data, we subsequently differentiate between behavioral *observational consistency* and behavioral *causal consistency*.

3 Multilevel causal modeling in *C. elegans*

In this section, we demonstrate the application and utility of the NC-MCM framework on calcium imaging data recorded in the nematode *C. elegans*. After introducing the data in Section 3.1, we show how to learn cognitive dSCMs that are behaviorally observationally consistent and test them for dynamic causal consistency (Section 3.2). In Section 3.3, we demonstrate how the NC-MCM models uncover behavioral motifs that are supported by distinct neuronal activation patterns.

C. elegans is an ideal model system to demonstrate the utility of the NC-MCM framework for multiple reasons. It has only 302 neurons, about a third of which can be simultaneously recorded by Ca^{2+} imaging in an individual animal at single-cell resolution (Prevedel et al., 2014). Due to the stereotypical nature of the nematode’s neuronal system, many of these neurons can be identified and thus compared across animals. *C. elegans* further has a small behavioral repertoire of roughly seven distinct movement patterns. This moderate complexity, in combination with extensive prior knowledge on relations between neuronal activity and behavior, allows us to demonstrate the neuropysiological plausibility of the insights derived from the NC-MCM framework, paving the way for its application to more complex model systems.

We remark that learning a causally consistent cognitive dSCM is model-agnostic, in the sense that each of the computational steps in Section 2.3 can be accomplished by different models or algorithms. In the subsequent analyses, we follow the guideline that simple models should be preferred over complex ones

if they provide sufficiently accurate results. The code to reproduce all results is available at [INSERT GIT LINK HERE].

3.1 The data

The data set we use has been recorded by Kato et al. (2015) and is available at <https://osf.io/2395t/>. We use the data subset that has been recorded without external stimulation. It consists of data from five worms with $107 - 131$ neurons recorded in each individual worm for a period 18 minutes at a sampling rate of approximately 2.85 Hz. We subsequently refer to a sample of the calcium traces as the neuronal state vector $\mathbf{X}[t]$, which represents the state of all neurons recorded in one animal at time t .

Even though all five worms were immobilized during the recordings, established equivalences between the activity of individual neurons and the worms' behavior were used to label each neuronal state vector with the behavior that would have been concurrently observed in a non-immobilized worm (Kato et al., 2015). The behavioral labels assigned to the neuronal states by Kato et al. (2015) are $\mathcal{B} = \{\text{reversal1 } (\text{rev1}), \text{ reversal2 } (\text{rev2}), \text{ slowing } (\text{slow}), \text{ ventral turn } (\text{vt}), \text{ sustained reversal } (\text{reversus}), \text{ forward movement } (\text{fwd}), \text{ dorsal turn } (\text{dt})\}$. Neuronal states for which the behavioral label could not be identified are labeled as *nostate*.

3.2 Learning a cognitive model of *C. elegans*

To learn an observational model that relates neuronal activity patterns to behavior in each individual animal, i.e., to estimate $P(\mathcal{B}|\mathbf{X})$, we trained a set of logistic regression models to discriminate between all pairs of behaviors, resulting in 28 models per animal. We chose logistic regression models for this step because they directly estimate the probability of a behavior as a function of the neuronal state. Each classifier was trained on all recorded neurons of each individual animal, excluding only those neurons that were used to identify the behavioral labels (AVAL, AVAR, SMDDR, SMDDL, SMDVR, SMDVL, RIBR, and RIBL). Classification accuracy was evaluated via majority voting in combination with 10-fold cross-validation using random data splits. This evaluation resulted in a group-average decoding accuracy of 92.2%, with decoding results between 89.4% and 95.6% across individual worms, indicating that the linear logistic regression models provide a good characterization of the observational relationships between neuronal activities and behaviors. For subsequent modeling steps, the regression models were trained on all data without cross-validation. We then used these models to compute the conditional probabilities each pairwise classifier assigned to their respective behaviors as a function of the neuronal state, resulting in a 28-dimensional vector of behavioral probabilities for each neuronal state, which we subsequently term the behavioral probability trajectories. We note that the logistic regression models estimate observational- and not interventional probabilities (cf. Section 2.3).

To learn the mapping τ , we employed k -means clustering in the 28-dimensional space of probability estimates to assign neuronal states to clusters with (approximately) constant conditional behavioral probabilities (cf. Definition 4). We varied the number of clusters for k -means between two and 20 and re-ran k -means 100 times for each k with random initial seeds. For each k and run, we thereby obtained an assignment of every neuronal state to one out of k cognitive states. For each worm and number of cognitive states, we then computed the final cognitive state trajectory $C[t]$ by aggregating the 100 clustering runs via consensus clustering based on majority voting (Xanthopoulos, 2014).

Figure 3 illustrates the steps to learn a cognitive state trajectory on the data of the first worm with five cognitive states (all trajectories are projected to their first two principal components for visualization): The neuronal state trajectories together with the behavioral labels (A) are used to estimate the behavioral probability trajectories (B). These trajectories are then clustered to assign each element of the behavioral probability trajectories to a cognitive state (C). Re-projecting the cognitive labels onto the neuronal state trajectories gives the neuro-cognitive state trajectories (D).

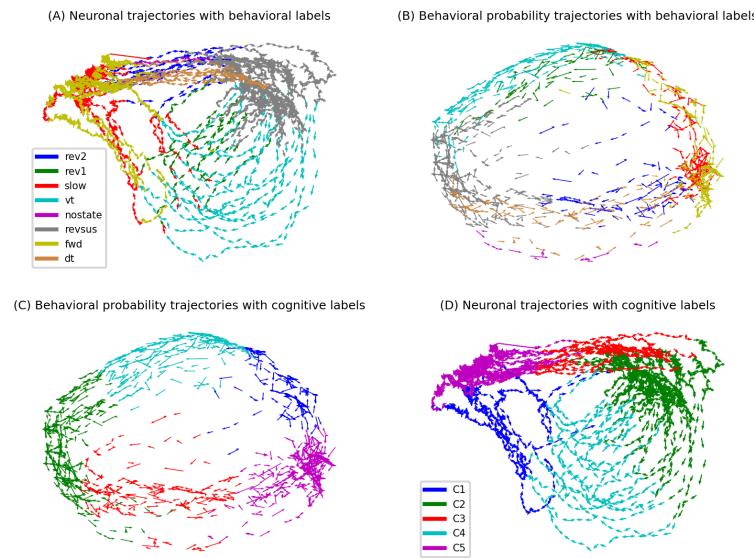


Figure 3: [TBD]

We tested the cognitive state trajectory of each worm and number of cognitive states for Markovianity by, first, estimating the first- and second-order

cognitive state transition probability matrices $P(\mathbf{C}[t]|\mathbf{C}[t-1])$ and $P(\mathbf{C}[t]|\mathbf{C}[t-1], \mathbf{C}[t-2])$, second, computing the total variance of $P(\mathbf{C}[t]|\mathbf{C}[t-1])$ across all states of $\mathbf{C}[t-2]$ (whose expected value for a Markov process is zero), and third, computing the same total variance for one thousand simulated Markovian cognitive state trajectories with state transition probability matrix $P(\mathbf{C}[t]|\mathbf{C}[t-1])$. This enabled us to estimate the p -value under the null-hypothesis of a Markovian cognitive dSCM as the frequency at which the simulated total variance exceeded the observed one.

To aggregate the cognitive state trajectories across individual worms into one group-level model, we matched (for each number of cognitive states) the cluster labels across all worms for which the null-hypothesis of Markovianity was not rejected at $\alpha = 0.1$ ². We did so by, first, computing the average neuronal state vector for each worm and cognitive state and, second, matching the cognitive states across worms according to the minimum mean squared error between averaged neuronal state vectors. The matched group-level cognitive state trajectories $\mathbf{C}[t]$ then enabled us to estimate the state-transition probability matrix T_C of the cognitive dSCM for each number of cognitive states. For each number of cognitive states, we finally tested the group-level cognitive model for Markovianity (using the same test and α -level employed for the individual worms).

We remark that the cognitive dSCM of each animal is behaviorally observationally consistent by construction and dynamically causally consistent if the cognitive dSCM is Markovian. Because we built the cognitive dSCM on the observational distribution $P(\mathbf{B}|\mathbf{X})$ rather than on the interventional distribution $P(\mathbf{B}|\text{do}(\mathbf{X}))$, which is not available in the present setting, behavioral causal consistency is not guaranteed (cf. Section 2.3). In its present form, the cognitive dSCM thus only supports causal statements about the dynamics of the cognitive dSCM but not about its relation to behavior. However, its observational behavioral consistency can be used to derive causal hypotheses that guide the design of experimental studies to establish causal relations between neuronal states and behaviors via interventions.

3.3 Interpreting the cognitive model of *C. elegans*

[MANUEL, PLEASE FILL IN YOUR PART ON THE NEUROBIOLOGICAL INTERPRETATION IN THIS SECTION].

We begin this section by discussing the original behavioral transition probabilities observed in *C. elegans*. This discussion establishes a baseline against which we can assess the benefits of analyzing *C. elegans* in the NC-MCM framework.

Figure 4 shows a graph representation of the behavioral transition probability matrix, using the behavioral data from all five worms. Here, each node represents one behavioral state. The sizes of each node and outgoing edge represent

²We remark again that Markovianity can not be verified. We can only fail to find evidence against Markovianity and thus accept the null-hypothesis.

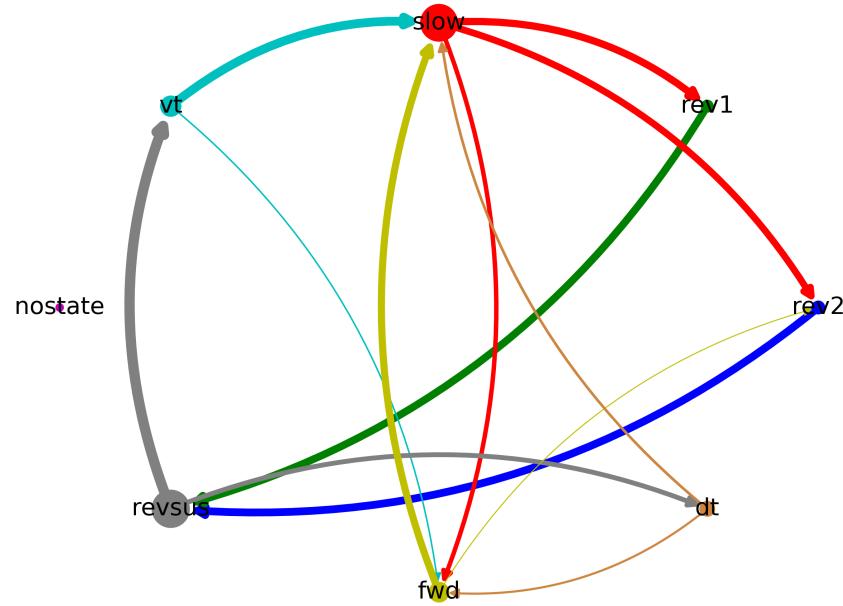


Figure 4: Graph representation of the behavioral transition probabilities. Links that make up less than 0.05% of the total transition probabilities are omitted to reduce clutter (see appendix for the full graph).

the probabilities of a worm being in a particular behavioral state at any point in time and transitioning from this particular state to another state, respectively (edges indicating the probabilities of staying in a particular behavioral state, i.e., edges from one node to itself, are not drawn to reduce clutter). Figure 4 appears to indicate a central role of the slowing (*slow*) behavior. In particular, alternating slowing and forward (*fwd*) movements appear to form a behavioral motif. This motif is interrupted during a slowing behavior by one out of two reversal movements, *rev1* or *rev2*, both of which are followed by a sustained reversal movement (*reversus*). A sustained reversal is either followed by a ventral- (*vt*) or a dorsal turn (*dt*), both of which segue back to the slowing-forward motif.

While Figure 4 thus reveals distinct structure in the behavioral dynamics, we show in the following that it conflates distinct behavioral motifs. This is due to the fact that the behavioral dynamics do not form a Markov chain (the statistical test described in Section 3.2 rejects the null-hypothesis of Markovianity at $p = 0.035$). Accordingly, behavioral transitions do not only depend on the current behavioral state but also on past behaviors, e.g., the probability that a slowing movement is followed by a forward movement is higher when the

slowing movement is preceded by a dorsal turn than when it is preceded by a ventral turn. In contrast, the NC-MCM framework can reveal multiple distinct behavioral motifs despite overlapping behavioral states because it abstracts the relationships between neuronal dynamics and behaviors into a Markovian cognitive dSCM.

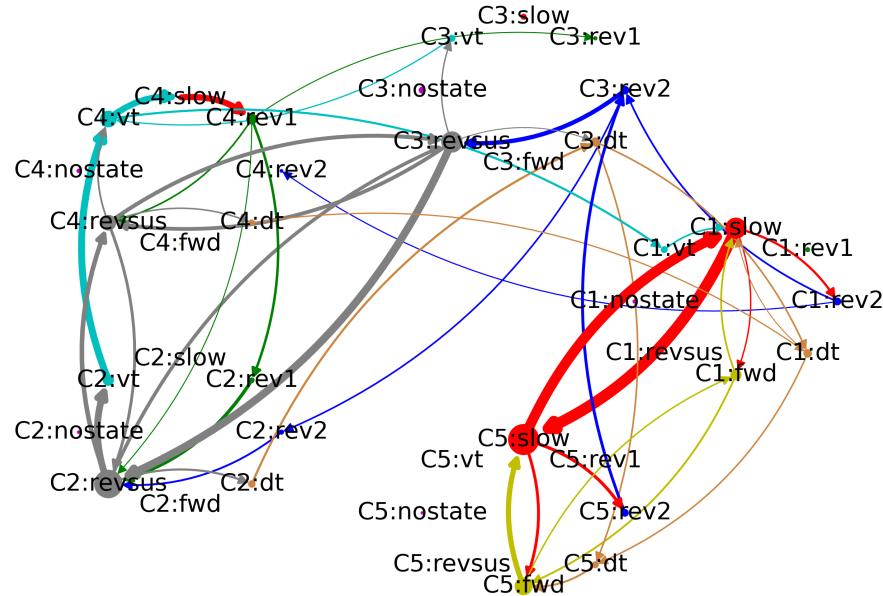


Figure 5: Graph representation of the cognitive-behavioral transition probabilities for the cognitive dSCM with five cognitive states. Links that make up less than 0.05% of the total transition probabilities are omitted to reduce clutter (see appendix for the full graph).

Figure 5 shows the cognitive dSCM for five cognitive states, a model for which the individual dSCMs of all worms as well the group-level dSCM do not exhibit evidence against Markovianity, i.e., for all models the p -value under the null-hypothesis of Markovianity is larger than $\alpha = 0.1$ (cognitive dSCMs for other numbers of cognitive states are shown in the appendix). In contrast to Figure 4, this graph separates the behavioral states according to the cognitive states in which they occur, with the five cognitive states drawn counter-clockwise. This separation reveals three distinct behavioral motifs. The first motif is the alternating forward-slowing movement, which is supported by cognitive states C1 and C5. The second behavioral motif is a sustained reversal movement, which is supported by cognitive states C2–C4. These two motifs

are linked by cognitive state C3: *C. elegans* predominantly initiates and ends the forward-slowing motif through a dorsal turn (*dt*) and *rev2* movement in cognitive state C3, respectively. These two major behavioral motifs are complemented by a third motif that consists of a quick succession of a ventral turn, a slowing, and type one reversal movement (C2:reversus → C2:vt → C4:vt → C4:slow → C4:rev1 → C2:reversus).

In contrast to the behavioral transition probabilities in Figure 4, the cognitive dSCM thus reveals that the slowing behavior is embedded in two behavioral motifs that are supported by distinct cognitive states (C1 and C5 vs. C4). Indeed, comparing the mean neuronal activations during slowing movements in cognitive states C1 and C5 with those in cognitive state C4, shown in the first column and third row of Figure 6, reveals that these apparently similar behavioral states are supported by distinct neuronal activation patterns. The average neuronal activation pattern of the slowing behavior in C4 is more similar to the neuronal activation patterns of the *rev1* and *vt* movements in C4 than to that of the slowing behavior in states C1 and C5. Analogously, the two different reversal movements *rev1* and *rev2*, both of which follow a slowing movement and are followed by a sustained reversal and thus appear redundant in Figure 4, are part of distinct behavioral motifs. While the *rev1* behavior marks the transition from the brief ventral turn–slowing–reversal motif in C4 back to the sustained reversal motif, the *rev2* behavior marks the transition in C3 from the forward–slowing motif (supported by C1 and C5) to the sustained reversal motif (in C2–C4). We thus see how the NC-MCM framework is able to assign seemingly similar behaviors into distinct behavioral motifs based on their neuronal representations.

While the behavioral dynamics in Figure 4 only support observational interpretations, e.g., such as in *a slowing movement is often followed by a forward movement*, the dynamic causal consistency property of the cognitive dSCM in Figure 6 allows causal interpretations of the state transitions. For instance, C3 is the cognitive state that causes a slowing movement in C1 or C5 to be followed by a sustained reversal movement in C2 or C3. And C2 is the cognitive state that predominantly initiates the *vt* → *slow* → *rev1* motif. Further, the (observational) behavioral consistency property allows us to map the cognitive states back to the neuronal states that they are composed of. For instance, the cognitive dSCM makes the causal prediction that setting the neuronal states of *C. elegans* to any of the patterns observed in state C4:slow (cf. Figure 12 in the appendix) will lead to a slowing behavior that is embedded in the ventral turn–slowing–reversal motif. In contrast, the model predicts that setting the neuronal states of *C. elegans* to any of the patterns observed in state C5:slow will lead to a slowing behavior that is embedded in the forward–slowing motif.

3.4 Summary

In this section, we have demonstrated how to learn a cognitive dSCM of the nematode *C. elegans*. Specifically, we have shown how the NC-MCM framework in combination with standard machine learning techniques can compress all be-



Figure 6: Group-averaged neuronal activations (for those neurons that have been identified in every worm) within each cognitive-behavioral state. Shaded regions indicate one standard deviation. Individual neuronal patterns for each cognitive-behavioral state are shown in Figure 12 in the appendix.

haviorally and dynamically relevant information in the neuronal states into a set of high-level, cognitive variables. We have then demonstrated how to test the resulting cognitive-level model for Markovianity which, via the dynamic causal consistency property, allows us to interpret the cognitive dynamics causally. Further, we demonstrated how the behavioral consistency property allows us to ground cognitive-level statements in the neuronal level, enabling us to derive causal predictions on the relations of neuronal activity patterns and behavior. The resulting cognitive dSCM revealed three distinct behavioral motifs in *C. elegans* and shed new light on their neuronal representations. As such, the present section serves as a justification and motivation for using the NC-MCM framework to learn cognitive models in an automated, data-driven fashion in larger organisms where the high dimensionality of neuronal states and the complexity of behavioral patterns prohibits the use of hand-crafted mechanistic models.

4 Discussion

The framework of NC-MCM provides a formal and mathematically rigorous framework to bridge the explanatory gap between neuronal activity and cognition. This bridging is achieved by construing cognitive states as abstractions of neuronal states that are causally consistent with respect to a set of behavioral patterns (behavioral causal consistency) and with respect to the system’s dynamics (dynamic causal consistency). As such, a NC-MCM enables us to causally reason about a system’s dynamics and behavior on the cognitive level while grounding all causal statements in the system’s neuronal states.

Dynamic causal consistency is achieved by constructing a Markovian representation of the observed dynamics. Markovianity is essential here, because it guarantees a causally sufficient description of the dynamics of a physical system. Consider a simple pendulum, whose complete physical state at any given point in time is given by its position and velocity. As such, setting the pendulum’s position and velocity by an external intervention fully determines its future behavior. Based on observational data only, we can check that the position and velocity provide a full characterization of the pendulum’s state by testing for Markovianity. This would reveal that the past positions and velocities do not provide additional information on the pendulum’s future states if the current position and velocity are known. Conversely, only observing the position of the pendulum does not allow us to determine its future state, e.g., because it could be swinging forward or backwards. However, we can reconstruct the full state space by also considering past positions from which we can then infer its velocity. In other words, the current position is not a Markovian representation of the pendulum’s dynamics and thus does not provide a causally sufficient representation of the physical system. Returning to the NC-MCM framework, the dynamic causal consistency property guarantees that we have found a complete, causally sufficient characterization of the neuronal dynamics that are relevant for a given behavioral context. We note that learning such a representation from empirical data is not guaranteed to succeed, e.g., because the set of observed

neurons may not provide a full characterization of the organisms physical state with respect to a given behavioral context.

In contrast to other frameworks that model cognition and its relation to neuronal dynamics, e.g., ACT-R (Anderson, 2013; Fincham et al., 2002) and atlases of cognition (Poldrack and Yarkoni, 2016; Varoquaux et al., 2018), a cognitive model in the NC-MCM framework does not incorporate prior assumptions on the structure of cognition, e.g., by positing information-processing modules (cf. Ritter et al. (2019)) or by assuming a cognitive ontology (Varoquaux et al., 2018). Rather, a cognitive model in the NC-MCM framework is learned bottom-up from the system’s neuronal dynamics in combination with a behavioral context. In the following, we discuss in which sense such a model provides an understanding of the neuronal system it models.

On a practical level, the NC-MCM framework can be interpreted as a data compression and visualization method. By eliminating redundant information in the neuronal states with respect to the behavioral context, irrelevant information is discarded and meaningful relationships emerge in an intuitively interpretable fashion. In this view, the NC-MCM framework is a tool for cognitive neuroscientists to analyze complex neuronal data.

On a more fundamental level, we first note that the term *understanding* is not well defined and as such subjective, e.g., what qualifies as an explanation that leads to understanding by one person may be considered an insufficient explanation by another person. We adopt the viewpoint of Papineau (1998) that the deepest level of understanding is that of identities, e.g., if we learn that the morning star and the evening star are in fact the same star, i.e., Venus, it does not make sense anymore to ask why Venus is Venus. In analogy, the NC-MCM framework, first, describes an equivalence relation between neuronal states with respect to a set of behaviors and, second, establishes identities between the neuronal equivalence classes and cognitive states. This raises the question what a cognitive state is or what it should be. We have defined a cognitive state as a causally meaningful abstraction of a neuronal state, implying that a cognitive model captures all relevant causal relations between neuronal states and behaviors (behavioral causal consistency) and enables us to causally reason about its dynamics (dynamic causal consistency). We argue that this definition of a cognitive state is intuitively plausible because it is in line with how we use cognitive states in everyday life. Faced with the daunting task of understanding the complex neuronal systems that support our own as well as other people’s behavior, we have developed a (sometimes more and sometimes less adequate) set of concepts which we employ to reason about our own and other people’s mental processes in a given behavioral context.

We note that the granularity of the behavioral context, which is determined by the observer of a system, influences the complexity of the cognitive abstractions. For instance, only distinguishing between forward and backward movements of *C. elegans* can be expected to lead to more redundancies in the neuronal states (and hence fewer cognitive states) than labeling *C. elegans*’ movements according to angles between multiple body segments.

On a more abstract philosophical level, we remark that the NC-MCM frame-

work offers a potential resolution to the problems of causal overdetermination and downward causation in the philosophy of mind (Woodward, 2020). By showing how a neuronal- and a cognitive level model can be constructed that are causally consistent, i.e., that allow us to interchangeably argue about a system’s dynamics and behavior on both levels, we hope that the NC-MCM framework will stimulate discussions on how to reconcile dualistic with physicalistic accounts of mental causation (Melnik, 2003; Menzies, 2003).

Returning to more practically relevant interpretations of the NC-MCM framework, we note that (in contrast to hand-crafted mechanistic models commonly employed in computational neuroscience) the NC-MCM framework does not require mechanistic models of the neuronal-level dynamics to learn (observationally) behaviorally and dynamically causally consistent cognitive-level models. While a mechanistic model may be useful to derive the interventional distribution $P(\mathbf{B}[t]|\text{do}(\mathbf{X}[t]))$ from which a causally behaviorally consistent model can be constructed, machine learning models in combination with a number of experimental interventions on the order of the number of cognitive states are in principle sufficient for learning $P(\mathbf{B}[t]|\text{do}(\mathbf{X}[t]))$ (cf. Section 2.3). Being able to leverage state-of-the-art machine learning algorithms for modeling the relations between neuronal activity patterns and behavior is particularly appealing when attempting to scale up NC-MCMs from small systems such as *C. elegans*, where hand-crafted mechanistic models may be feasible, to more complex organisms consisting of potentially hundred thousands or even millions of neurons. While detailed mechanistic models may continue to provide important insights into the computations of individual circuits, we conjecture that causally consistent NC-MCMs will be more useful in the domain of cognitive neural engineering, where we would like to know how to experimentally intervene on the neuronal level of large-scale organisms to treat cognitive disorders.

To conclude the article, we discuss some extensions of NC-MCMs for future work. First, we recall that we have chosen a rather simple machine learning pipeline in Section 3 to illustrate how to learn a NC-MCM. More complex computational models of *C. elegans* are available and could be leveraged for learning NC-MCMs (Brennan and Proekt, 2019). In general, leveraging the power of state-of-the-art artificial intelligence algorithms, e.g., as being developed within the framework of causal representation learning (Schölkopf et al., 2021), is probably required to learn NC-MCMs in more complex organisms. Second, we consider the extension to multiple behavioral contexts. In Section 3, we have considered a set of behavioral states that are mutually exclusive, e.g., the worm can either crawl forward or backward but not execute both actions simultaneously. In more complex organisms, we may consider multiple behavioral contexts that are not mutually exclusive, e.g., optomotor- and swimming (bout) actions in larval zebrafish. We could then learn a NC-MCM for each behavioral context and study causal interactions between the cognitive states of each NC-MCM. Naturally, this approach could be scaled-up to an arbitrary number of behavioral contexts, potentially giving rise to a causally meaningful and neuronally grounded approach to studying cognition in complex organisms. Finally, we remark that we have focused in this work on learning cognitive states

from neuronal data based on behavioral contexts. We could also consider other learning problems, e.g., settings in which sets of neuronal and cognitive states are already given and where we wish to determine whether there exists a behavioral context that gives rise to a NC-MCM. Such learning problems may be helpful to study whether already established cognitive ontologies, e.g., in psychology, are grounded in neuronal activity or, conversely, whether certain established cognitive concepts should be refined or excluded from scientific discourse. Independently of which particular learning problem we are interested in, the NC-MCM framework constitutes a rich and theoretically principled approach to bridging the explanatory gaps between neuronal activity patterns, cognitive states, and behaviors.

References

- Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.
- Borst, A. (2014). Fly visual course control: behaviour, algorithms and circuits. *Nature Reviews Neuroscience*, 15(9):590–599.
- Brennan, C. and Proekt, A. (2019). A quantitative model of conserved macroscopic dynamics predicts future motor commands. *eLife*, 8:e46814.
- Burke, C. and Rosenblatt, M. (1958). A markovian function of a markov chain. *The Annals of Mathematical Statistics*, 29(4):1112–1122.
- Chalupka, K., Eberhardt, F., and Perona, P. (2017). Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164.
- Chalupka, K., Perona, P., and Eberhardt, F. (2015). Visual causal feature learning. In *Proceedings of the Thirty-First Annual Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 181–190.
- Changeux, J.-P. and Dehaene, S. (1989). Neuronal models of cognitive functions. *Cognition*, 33(1-2):63–109.
- Fincham, J. M., Carter, C. S., Van Veen, V., Stenger, V. A., and Anderson, J. R. (2002). Neural mechanisms of planning: a computational analysis using event-related fmri. *Proceedings of the National Academy of Sciences*, 99(5):3346–3351.
- Gray, J. M., Hill, J. J., and Bargmann, C. I. (2005). A circuit for navigation in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences*, 102(9):3184–3191.
- Grosse-Wentrup, M., Janzing, D., Siegel, M., and Schölkopf, B. (2016). Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage*, 125:825–833.

- Kato, S., Kaplan, H. S., Schrödel, T., Skora, S., Lindsay, T. H., Yemini, E., Lockery, S., and Zimmer, M. (2015). Global brain dynamics embed the motor command sequence of *caenorhabditis elegans*. *Cell*, 163(3):656–669.
- Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3):480–490.
- Lakoff, G. and Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*, volume 640. Basic books New York.
- Markham, A. and Grosse-Wentrup, M. (2020). Measurement dependence inducing latent causal models. In *Proceedings of the Thirty-Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2020)*.
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.
- Melnyk, A. (2003). *A physicalist manifesto: Thoroughly modern materialism*. Cambridge University Press.
- Menzies, P. (2003). The causal efficacy of mental states. *Physicalism and Mental Causation*, Exeter: Imprint Academic, pages 195–224.
- Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.
- Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, 10(9):424–430.
- Papineau, D. (1998). Mind the gap. *Philosophical Perspectives*, 12:373–388.
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage*, 45(1):S199–S209.
- Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.
- Poldrack, R. and Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annual Review of Psychology*, 67:587–612.
- Prevedel, R., Yoon, Y.-G., Hoffmann, M., Pak, N., Wetzstein, G., Kato, S., Schrödel, T., Raskar, R., Zimmer, M., Boyden, E. S., and Alipasha, V. (2014). Simultaneous whole-animal 3d imaging of neuronal activity using light-field microscopy. *Nature Methods*, 11(7):727–730.

- Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford University Press.
- Quinn, N. and Holland, D. (1987). Culture and cognition. *Cultural models in language and thought*, 1.
- Ritchie, J. B., Kaplan, D. M., and Klein, C. (2020). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*.
- Ritter, F. E., Tehranchi, F., and Oury, J. D. (2019). Act-r: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3):e1488.
- Rubenstein, P., Weichwald, S., Bongers, S., Mooij, J., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency of structural equation models. In *Proceedings of the Thirty-Third Annual Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, page ID11.
- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.
- Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., Ramsey, J., and Woolrich, M. (2011). Network modelling methods for fMRI. *NeuroImage*, 54(2):875–891.
- Spirites, P. (2009). Variable definition and causal inference. Technical report.
- Spirites, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press.
- Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J.-B., and Thirion, B. (2018). Atlases of cognition with large-scale human brain mapping. *PLoS computational biology*, 14(11):e1006565.
- Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59.
- Woodward, J. (2020). Causal complexity, conditional independence, and downward causation. *Philosophy of Science*, 87(5):857–867.
- Xanthopoulos, P. (2014). A review on consensus clustering methods. In *Optimization in Science and Engineering*, pages 553–566. Springer.
- Yazdan-Shahmorad, A., Diaz-Botia, C., Hanson, T. L., Kharazia, V., Ledochowitsch, P., Mahabiz, M. M., and Sabes, P. N. (2016). A large-scale interface for optogenetic stimulation and recording in nonhuman primates. *Neuron*, 89(5):927–939.

Appendix

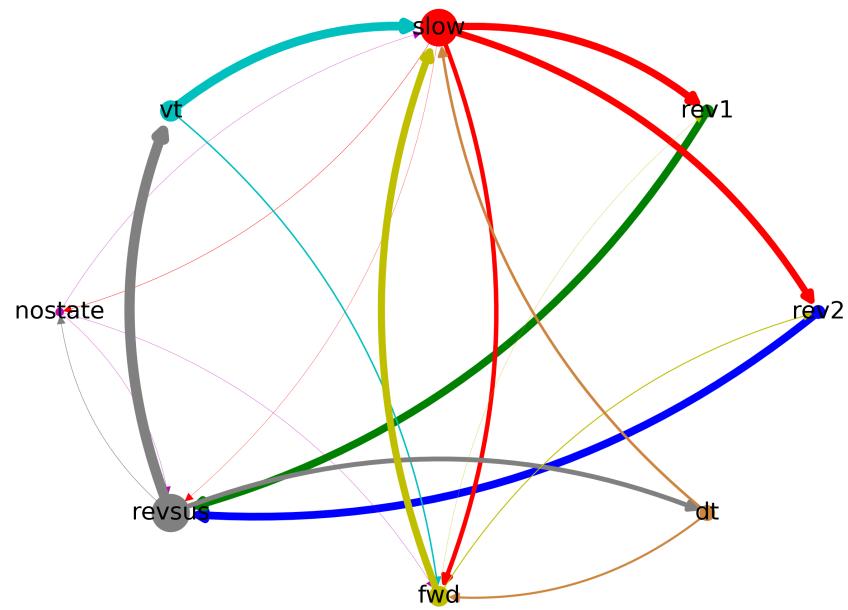


Figure 7: [TBD]

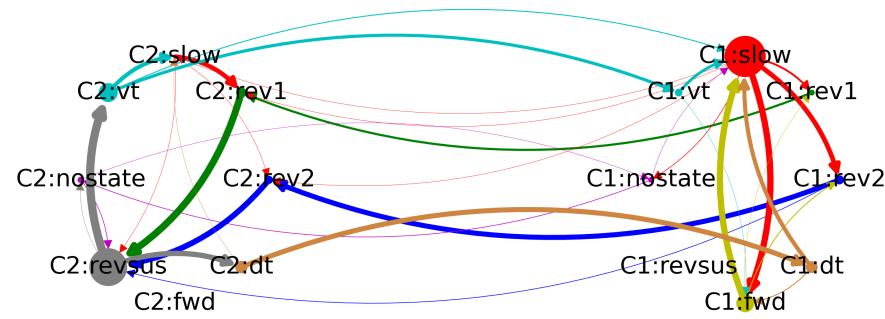


Figure 8: [TBD]

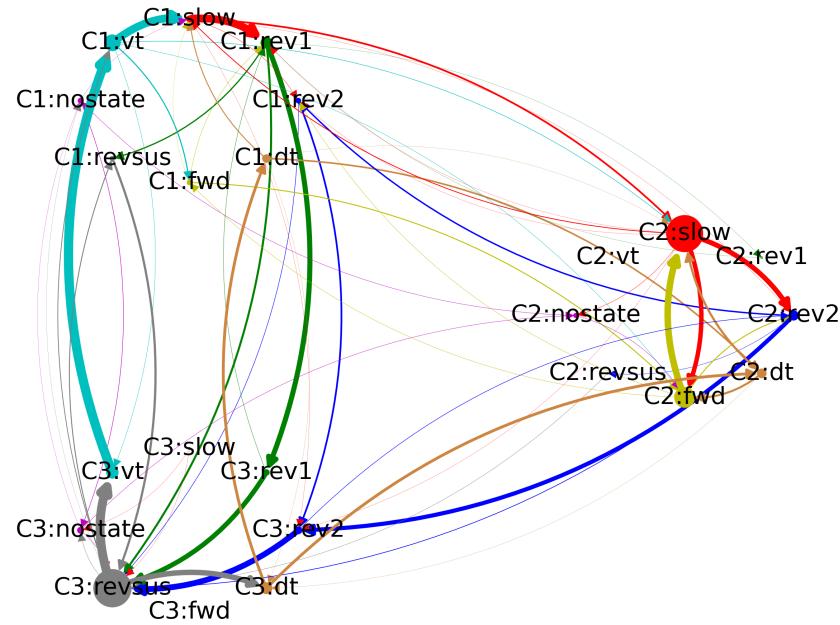


Figure 9: [TBD]

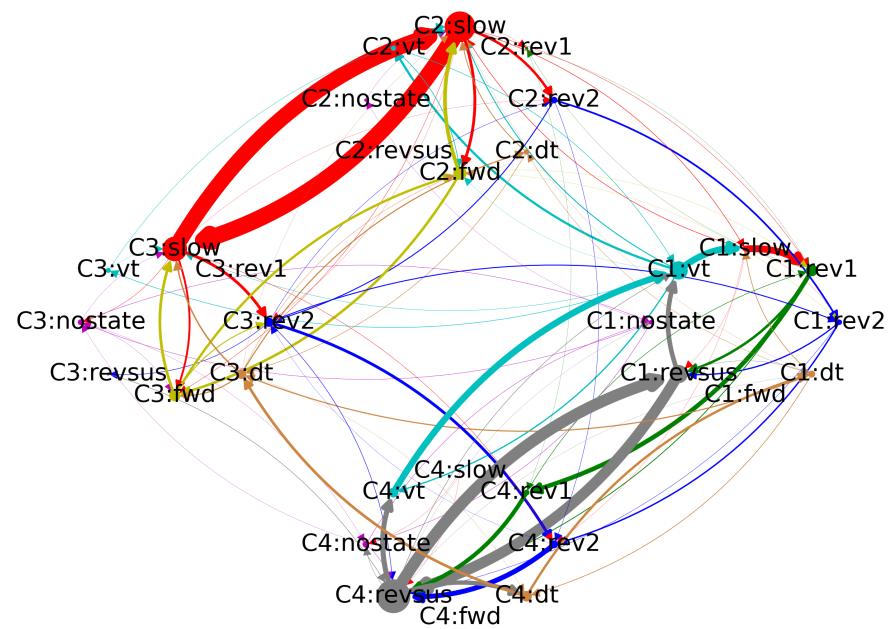


Figure 10: [TBD]

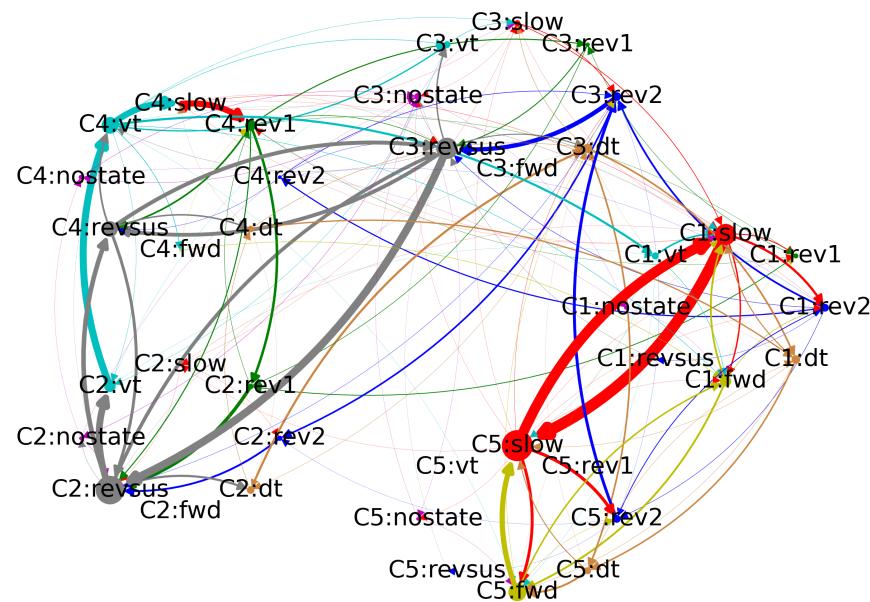


Figure 11: [TBD]

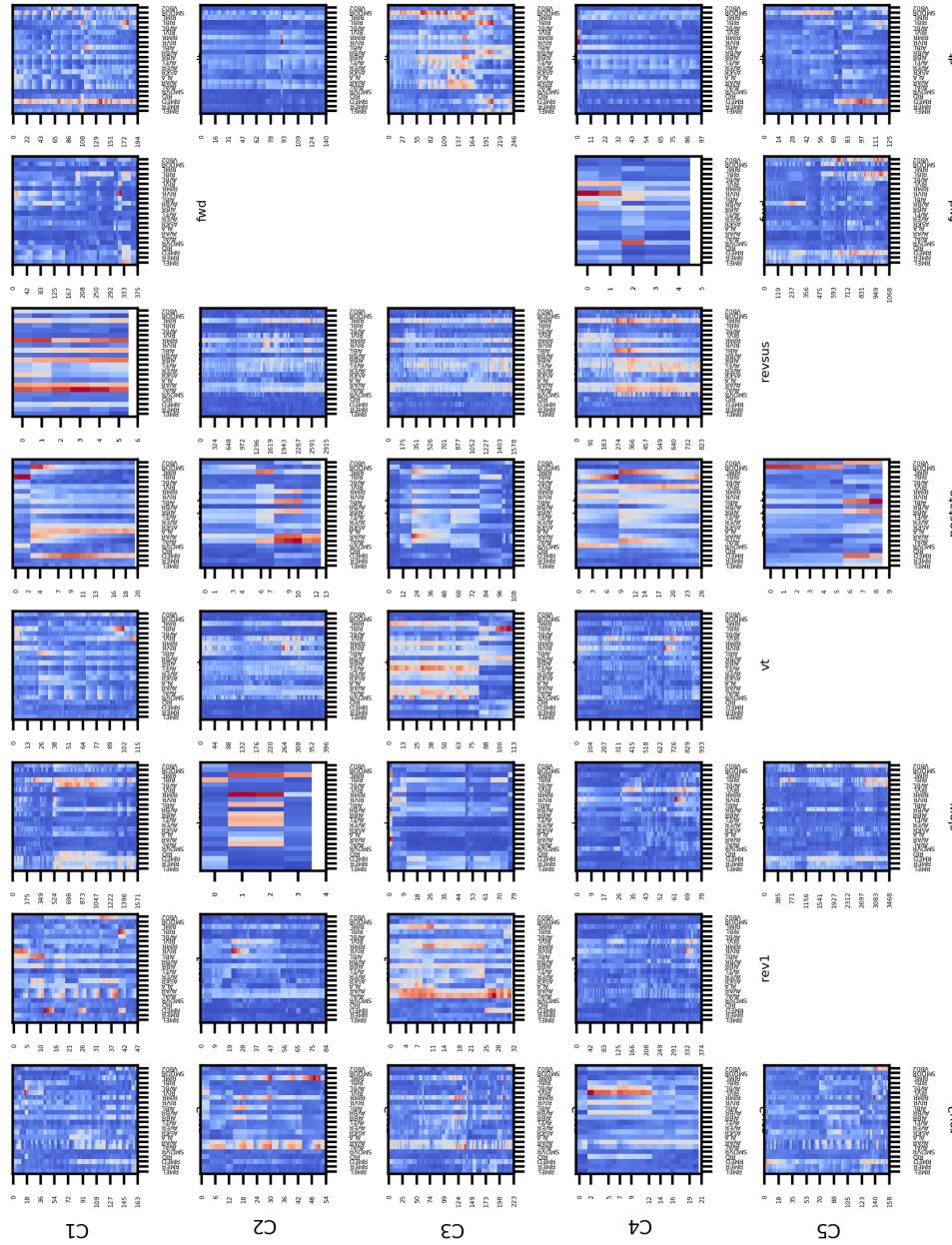


Figure 12: [TBD]

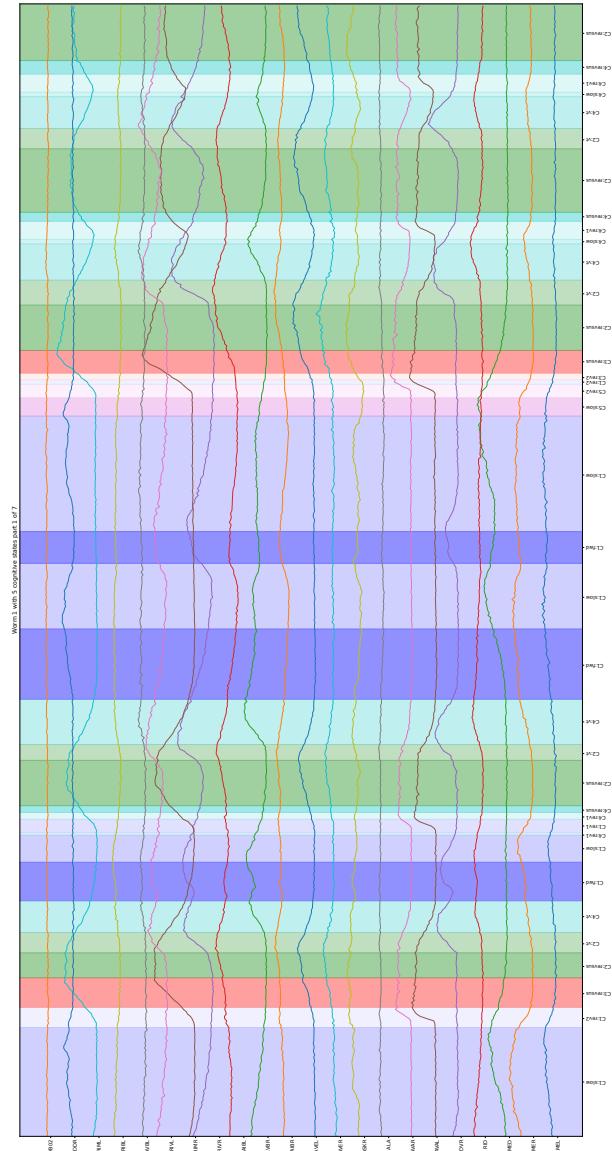


Figure 13: [TBD]

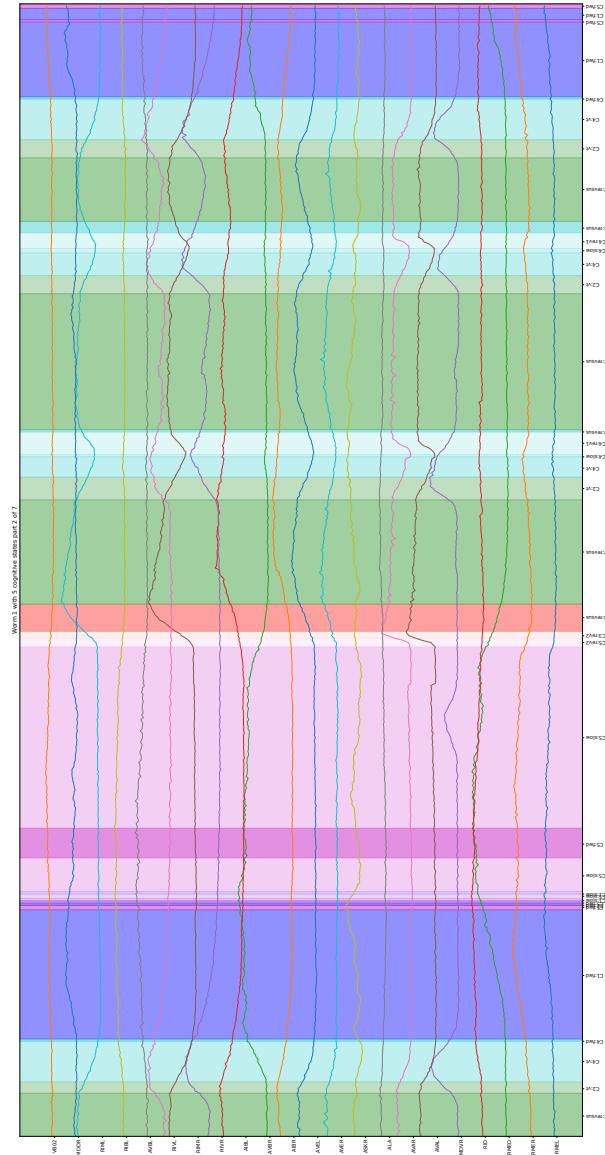


Figure 14: [TBD]

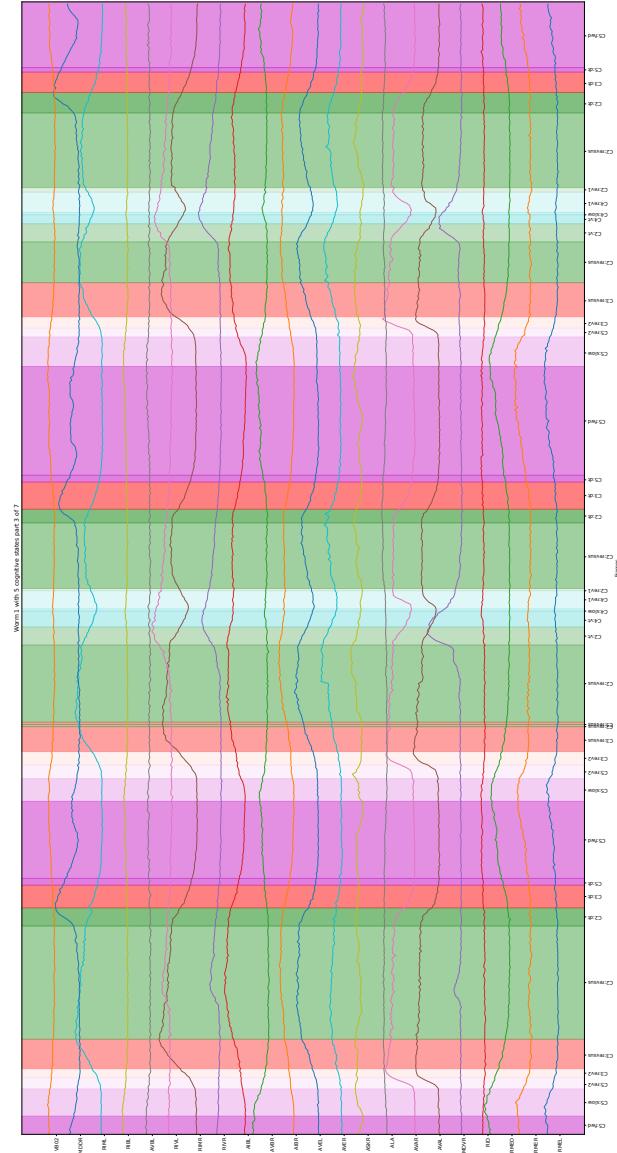


Figure 15: [TBD]

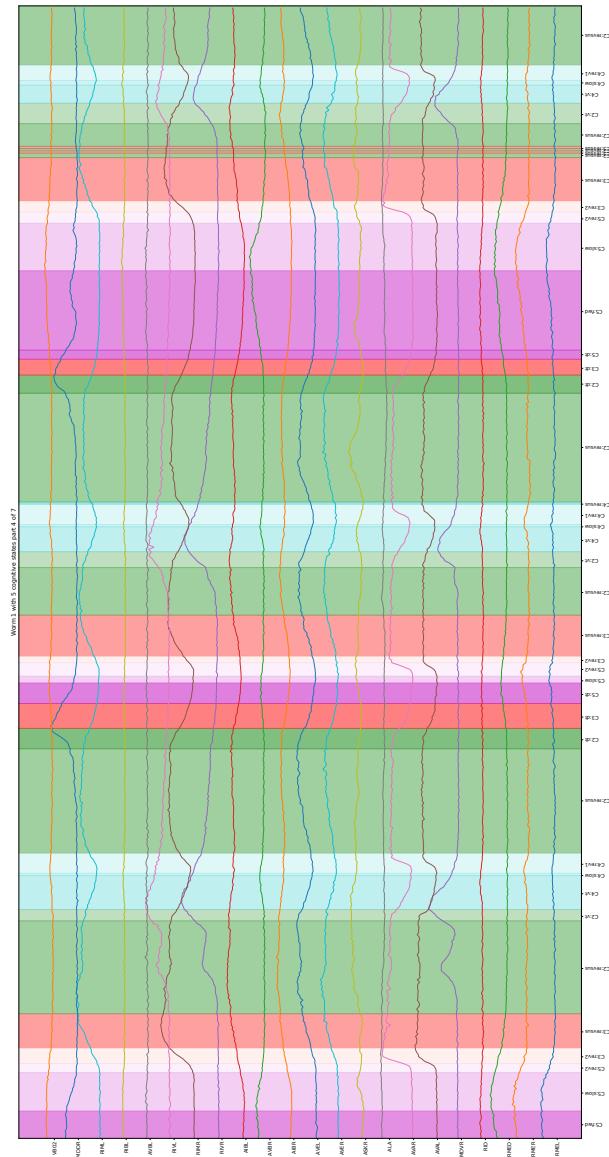


Figure 16: [TBD]

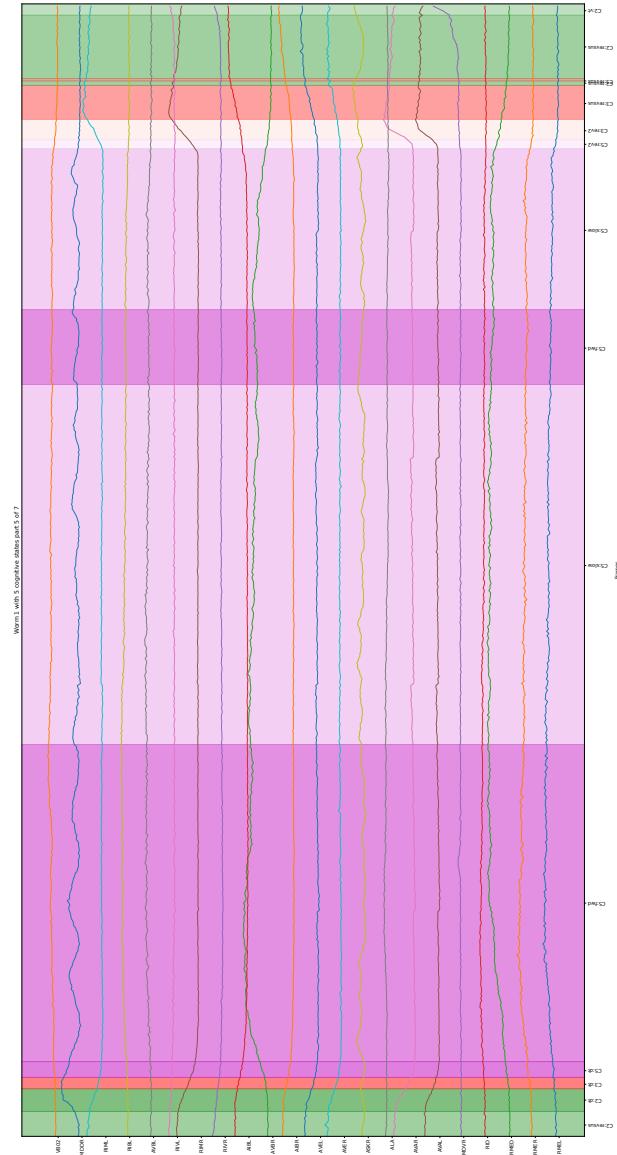


Figure 17: [TBD]

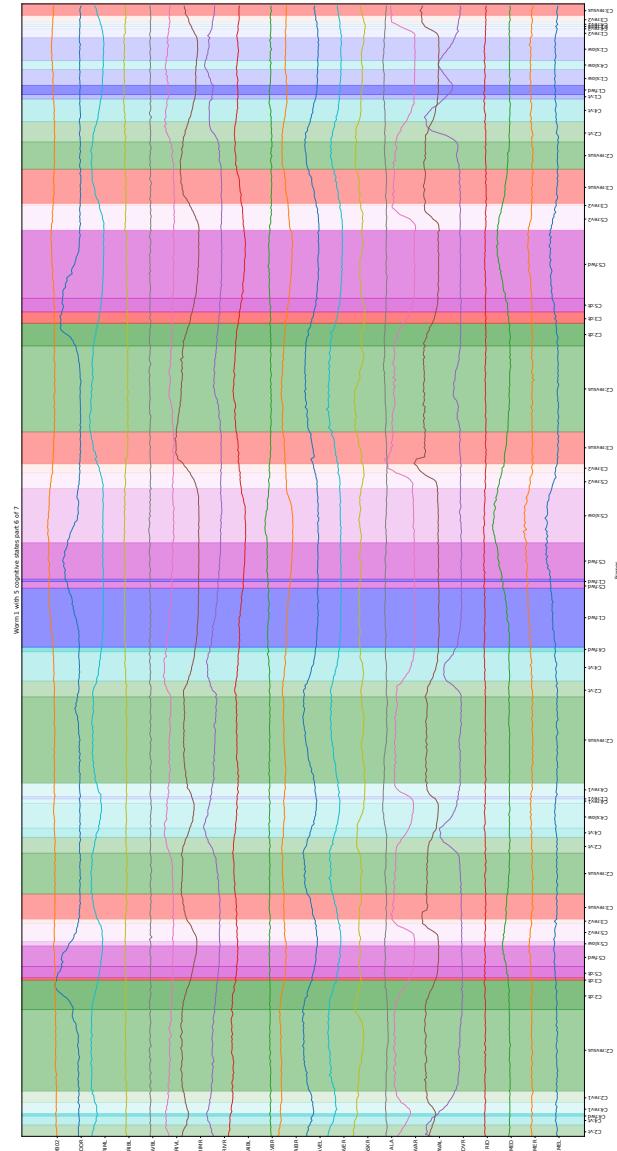


Figure 18: [TBD]

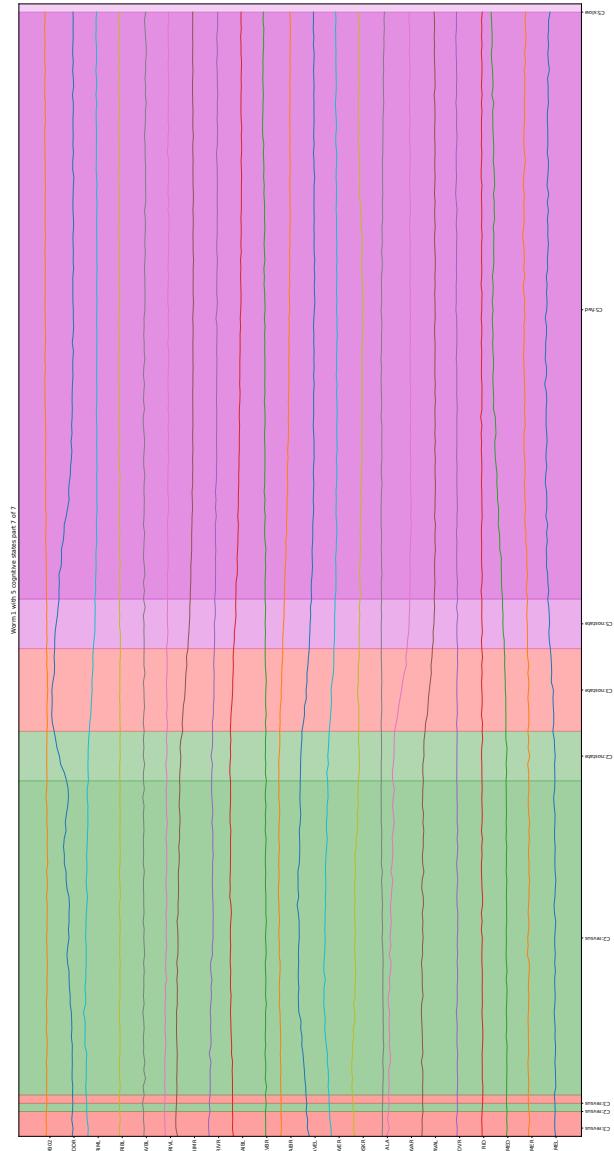


Figure 19: [TBD]

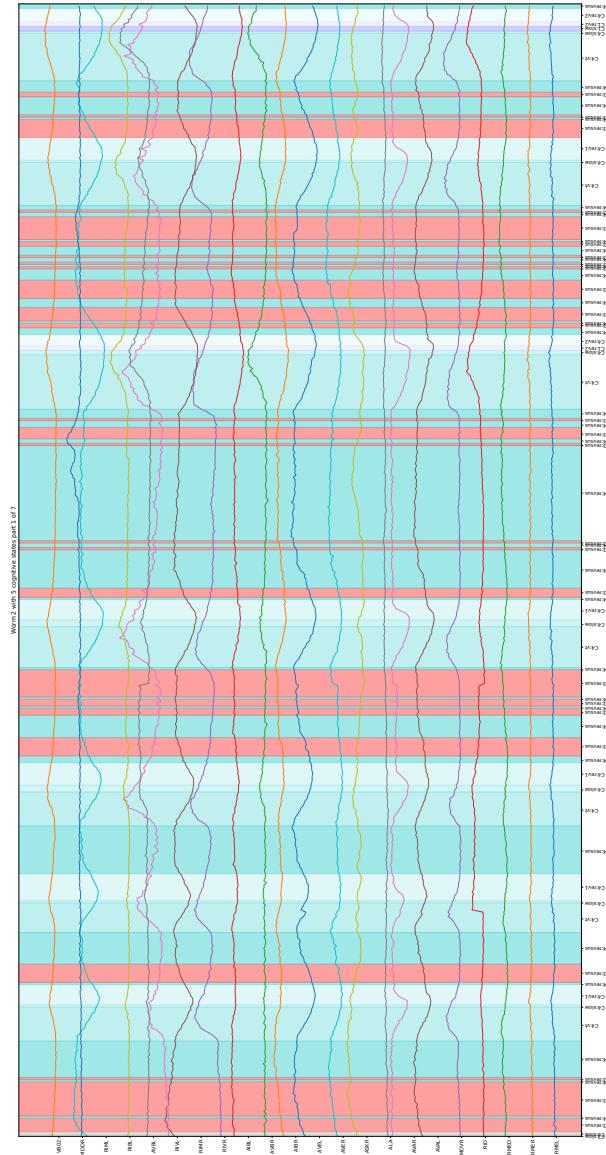


Figure 20: [TBD]

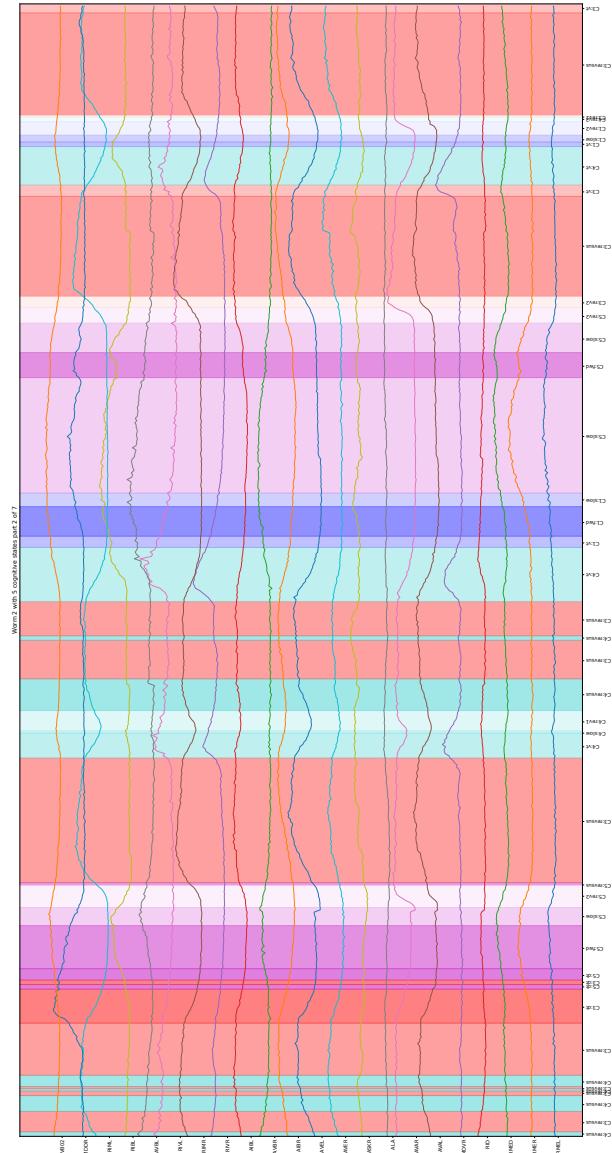


Figure 21: [TBD]

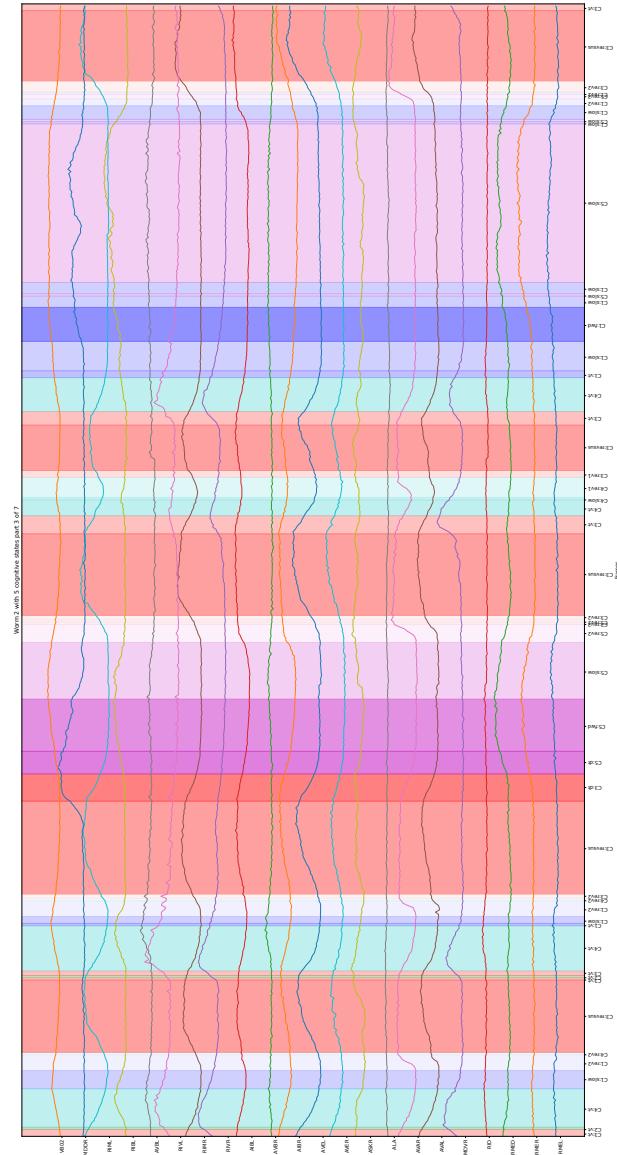


Figure 22: [TBD]

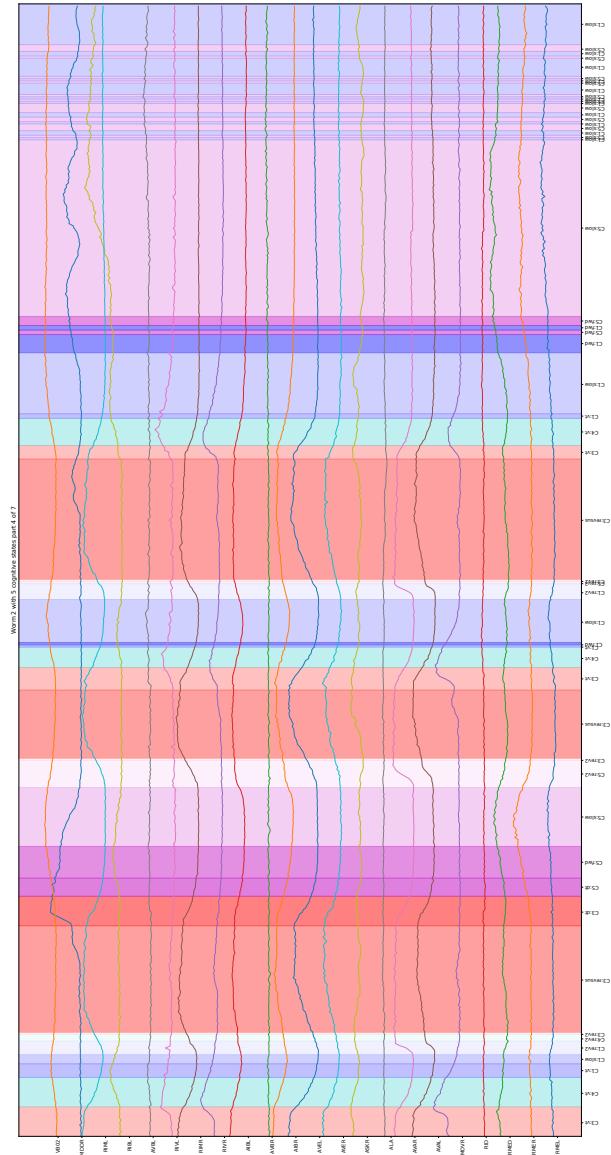


Figure 23: [TBD]

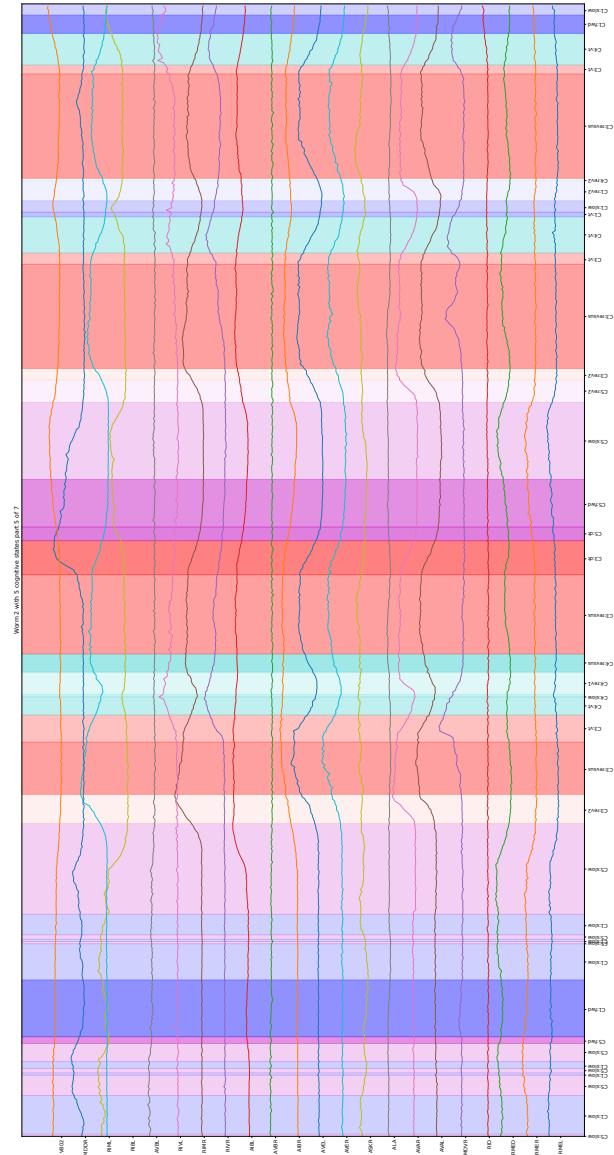


Figure 24: [TBD]

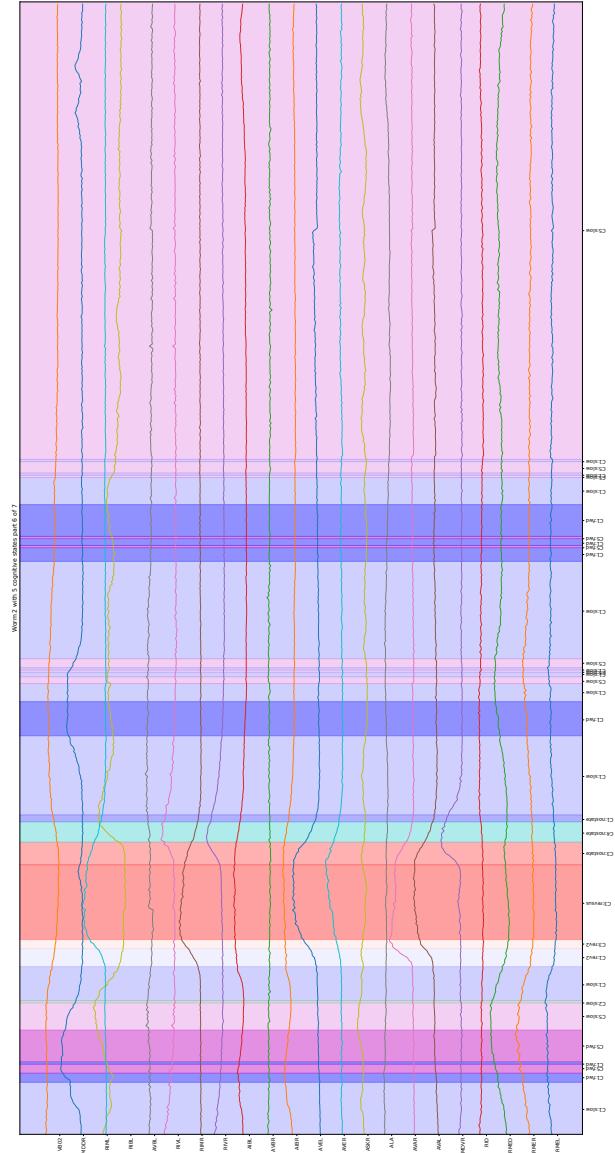


Figure 25: [TBD]

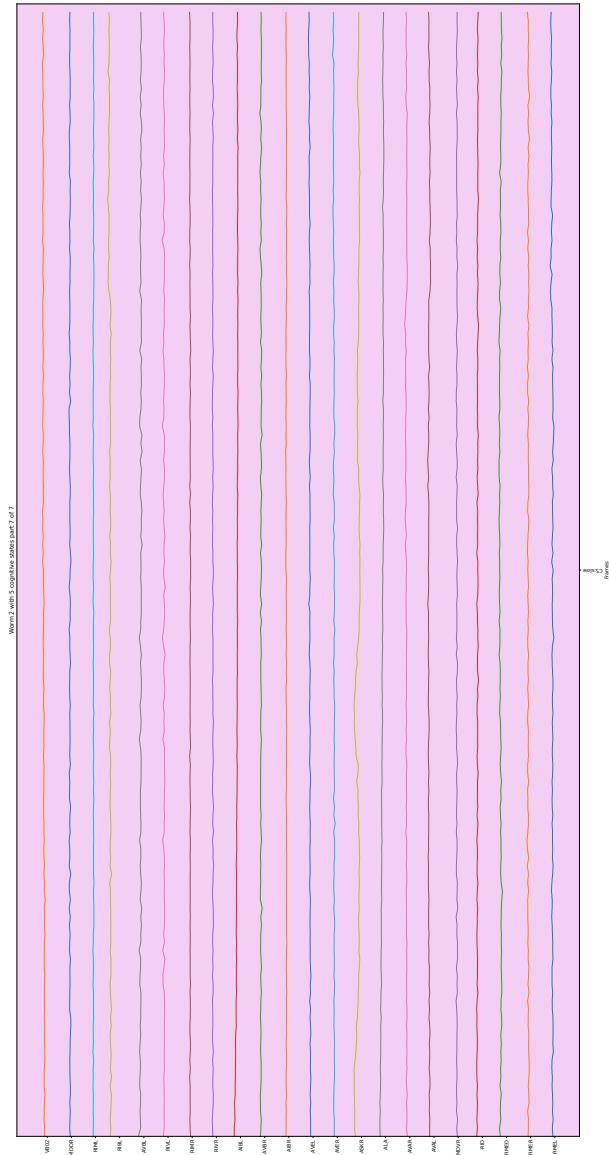


Figure 26: [TBD]

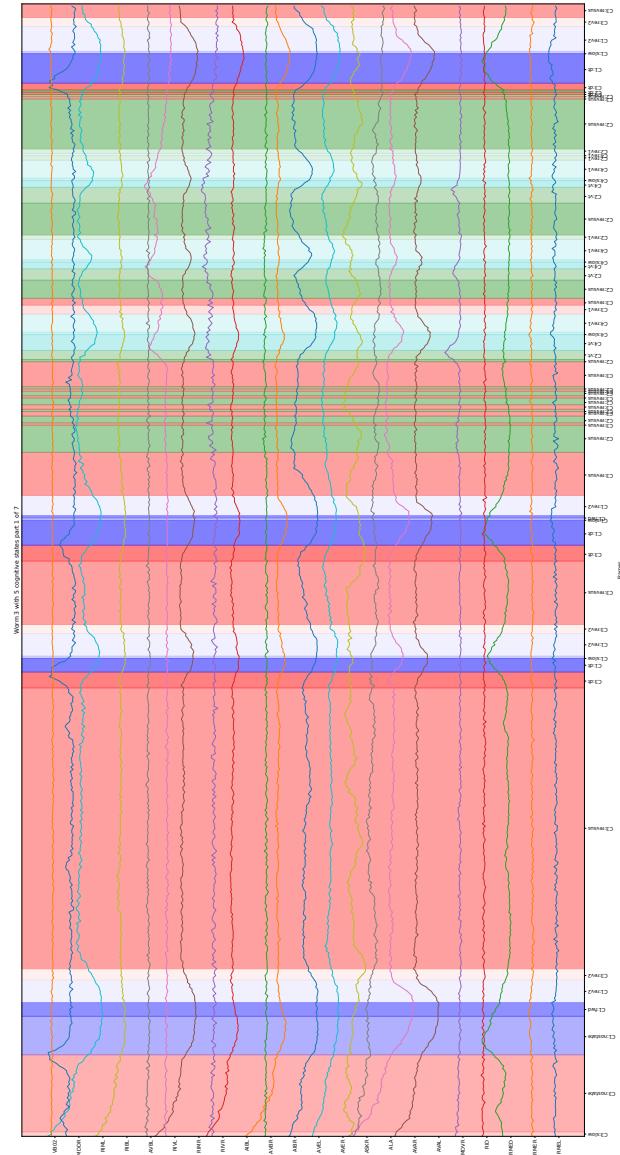


Figure 27: [TBD]

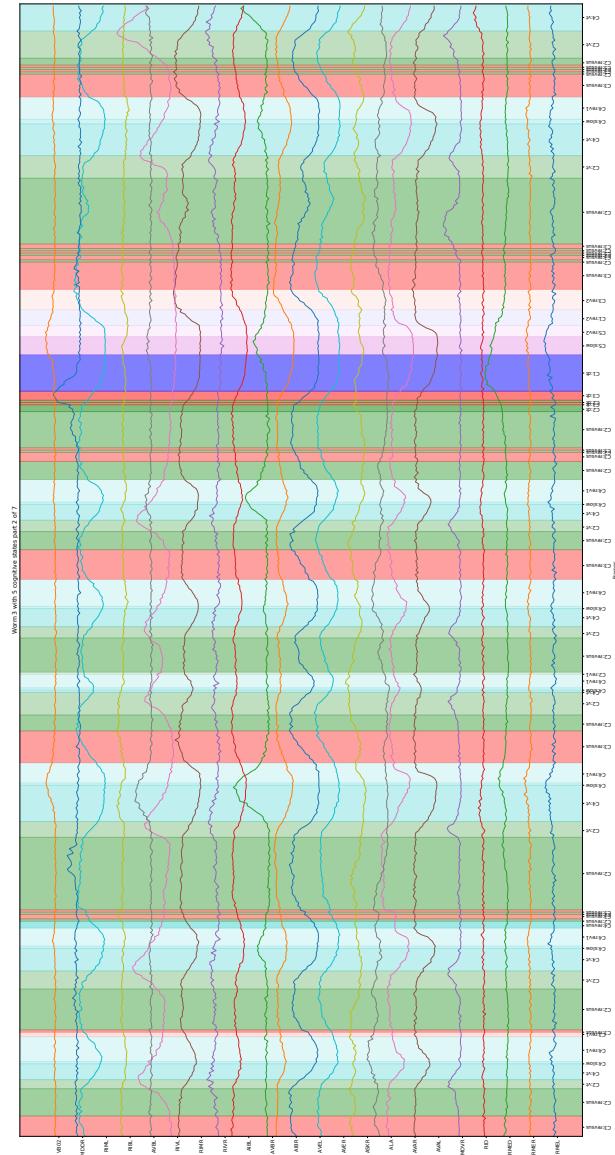


Figure 28: [TBD]

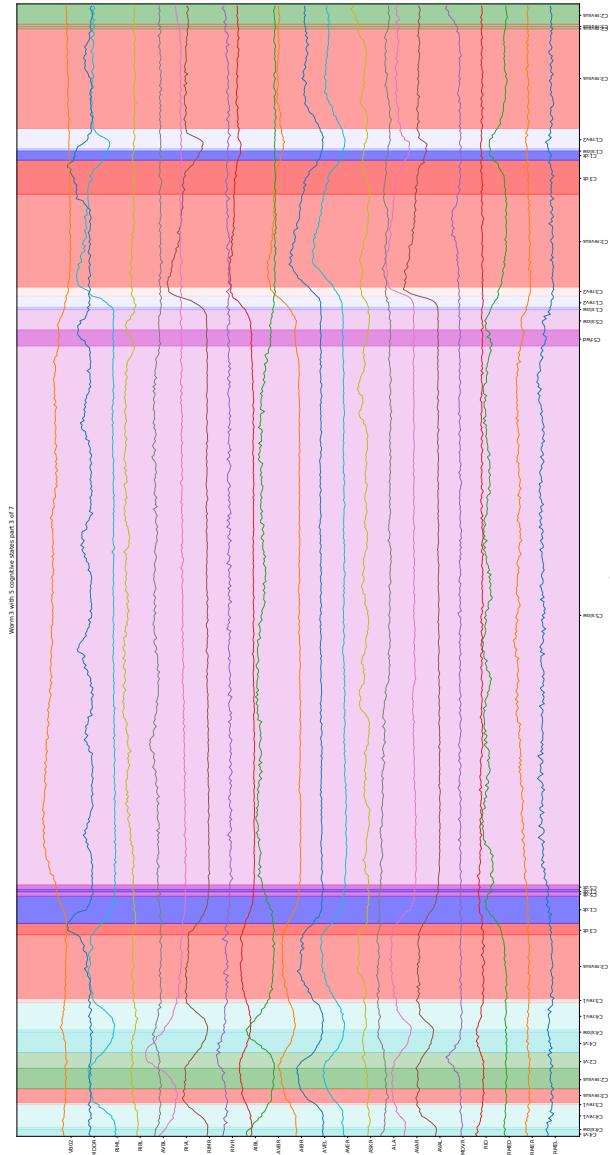


Figure 29: [TBD]

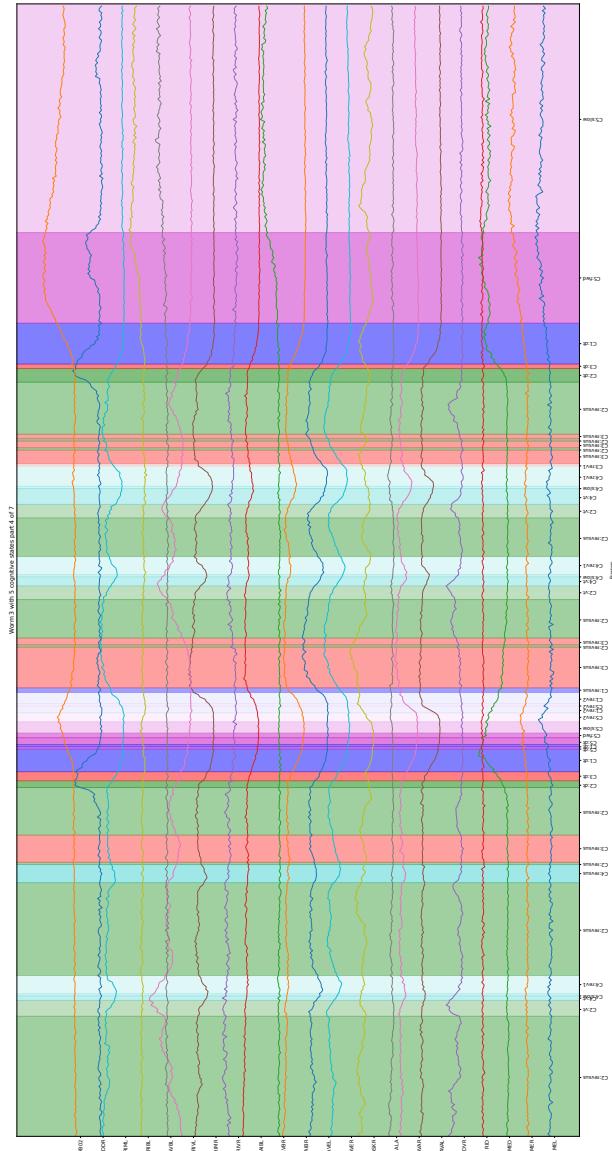


Figure 30: [TBD]

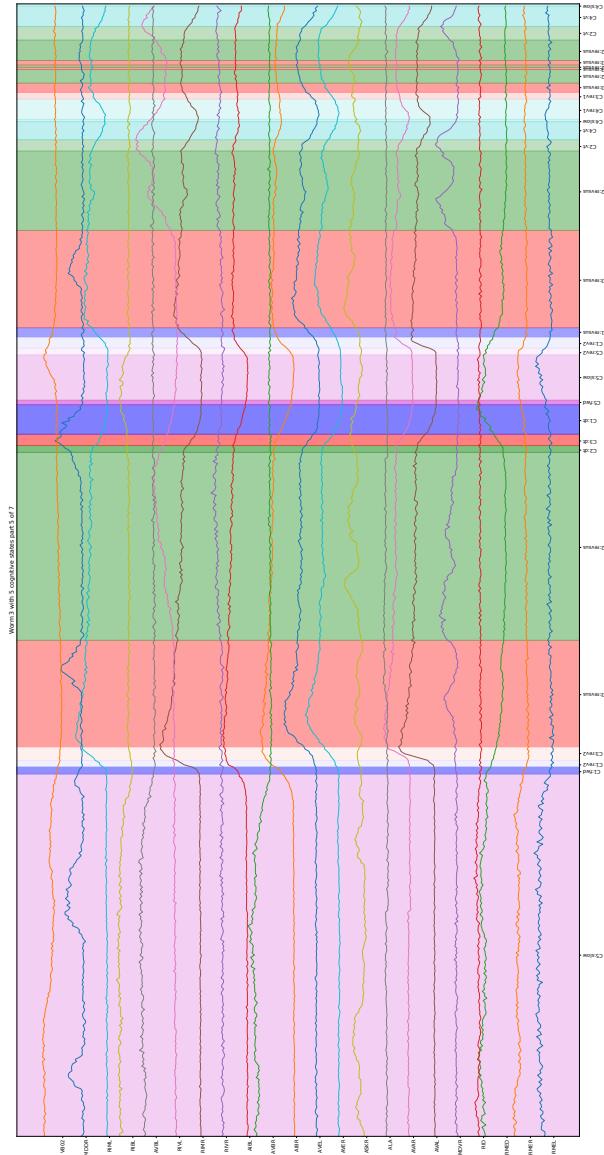


Figure 31: [TBD]

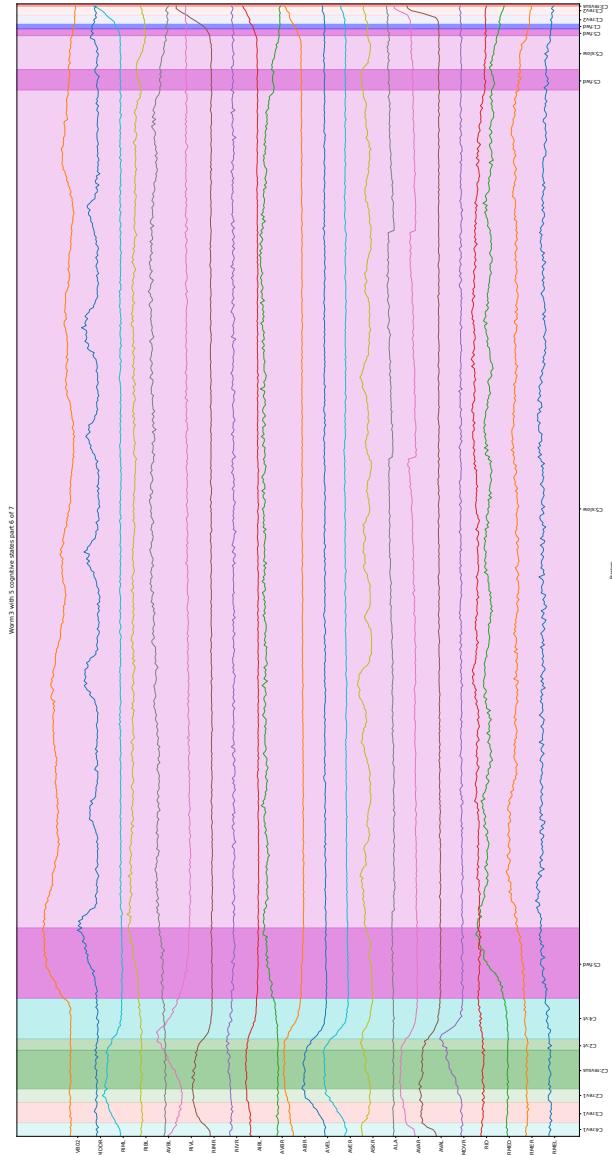


Figure 32: [TBD]

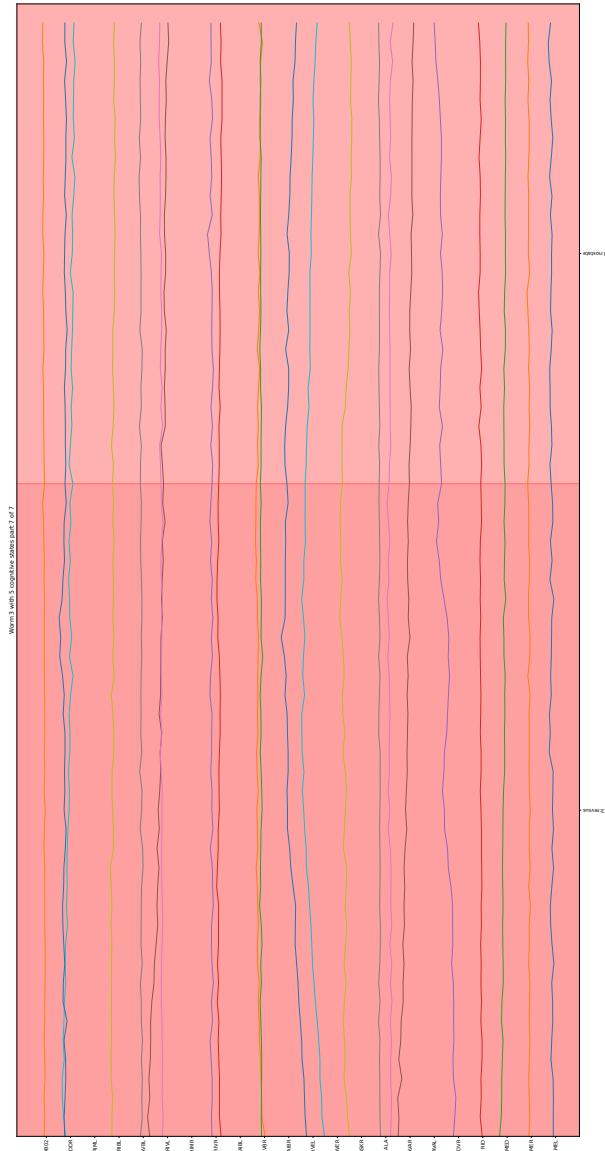


Figure 33: [TBD]

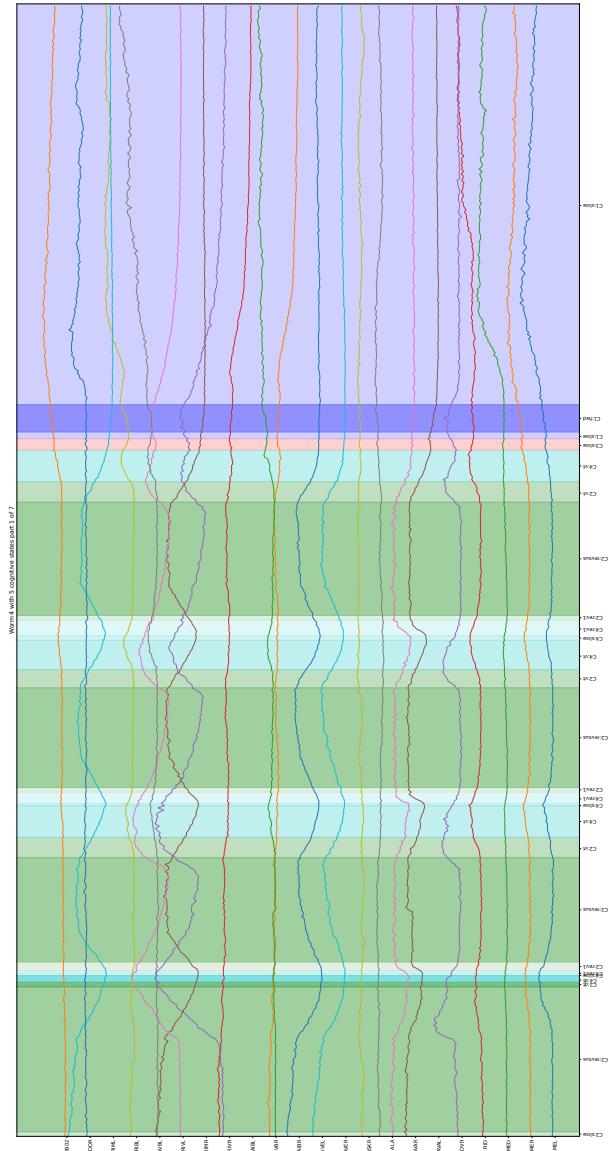


Figure 34: [TBD]

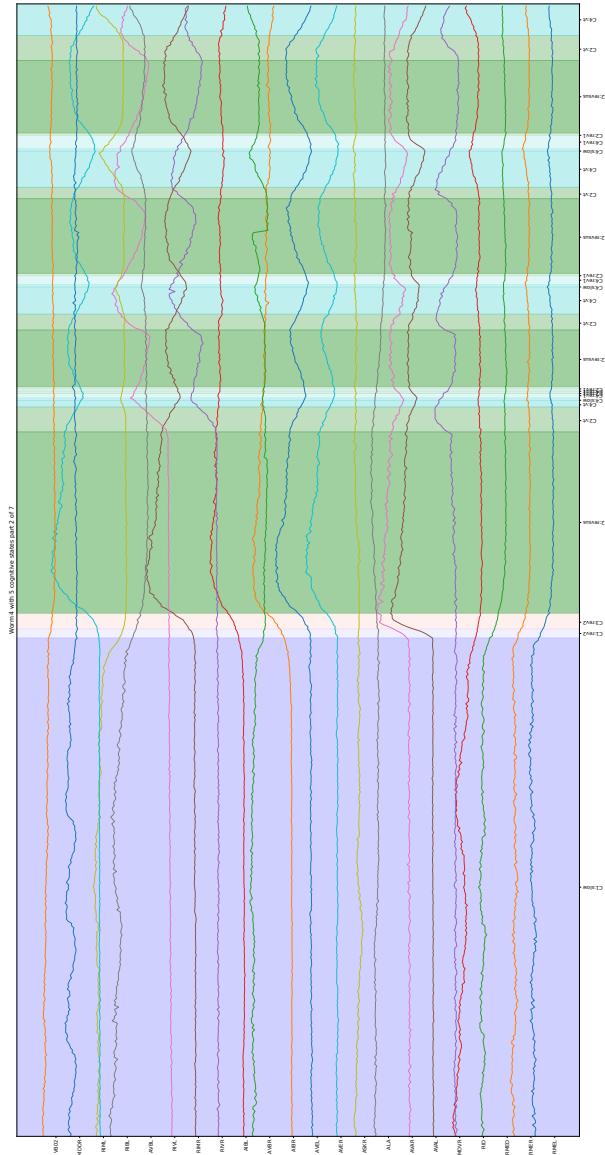


Figure 35: [TBD]

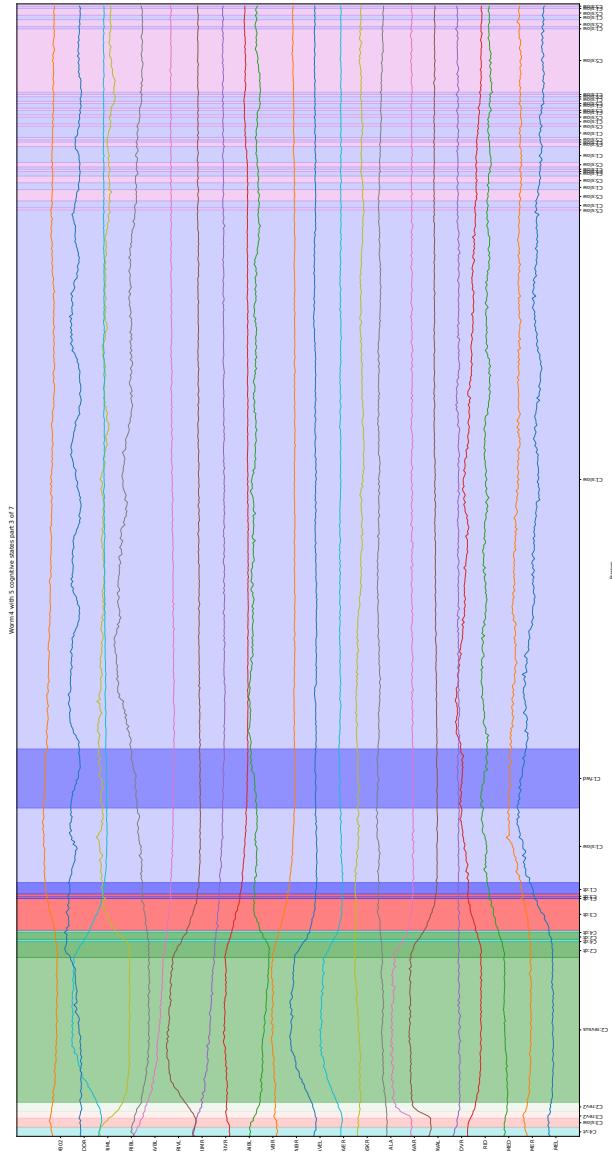


Figure 36: [TBD]

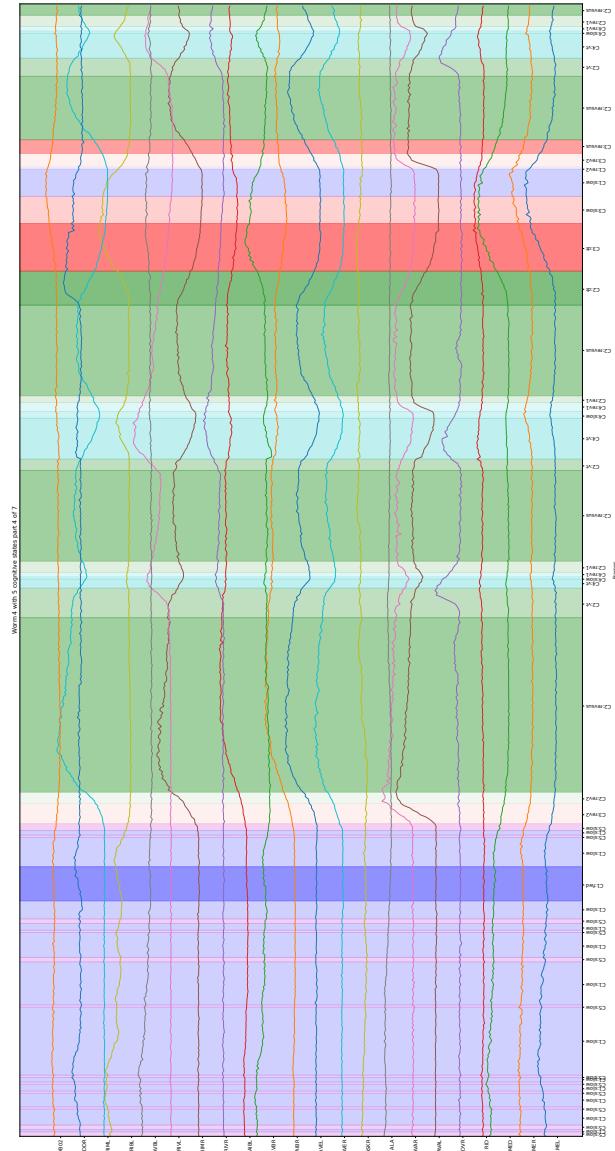


Figure 37: [TBD]

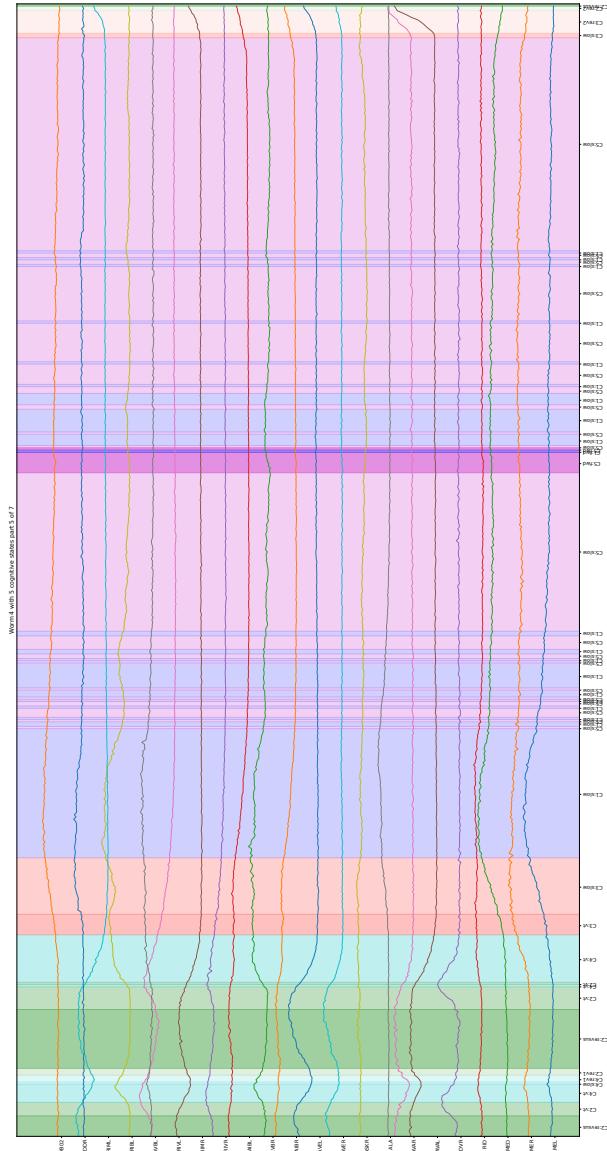


Figure 38: [TBD]

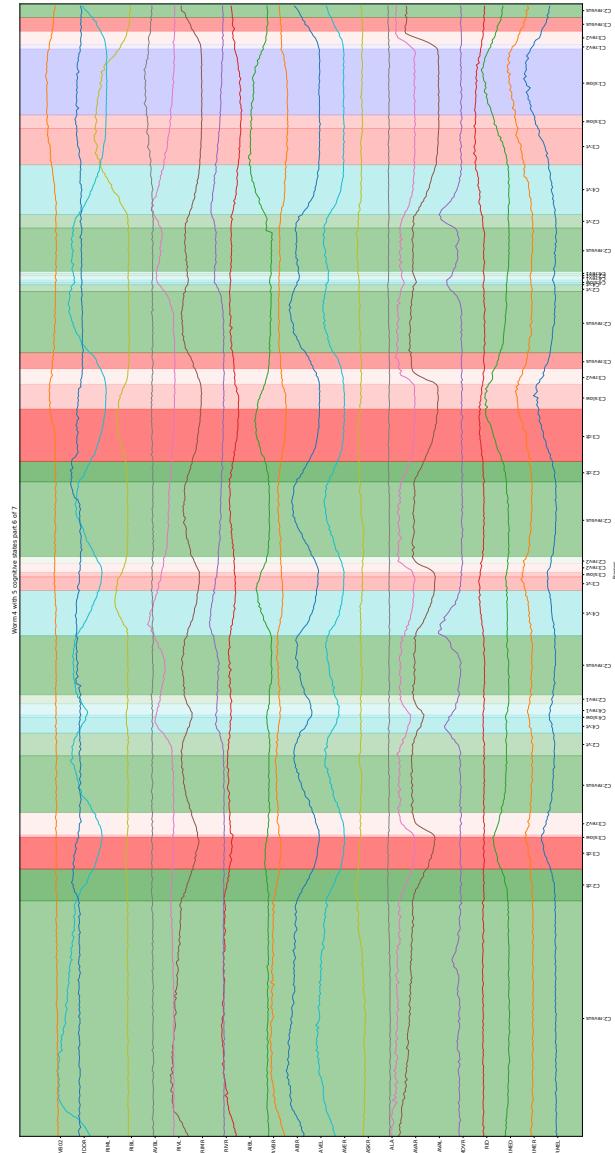


Figure 39: [TBD]

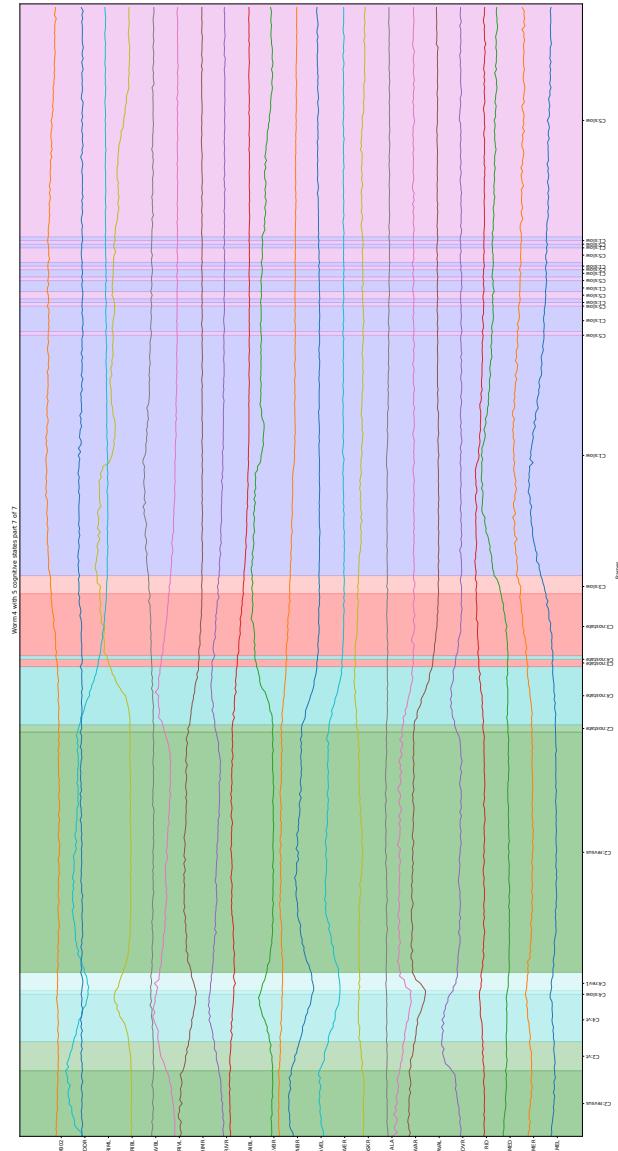


Figure 40: [TBD]

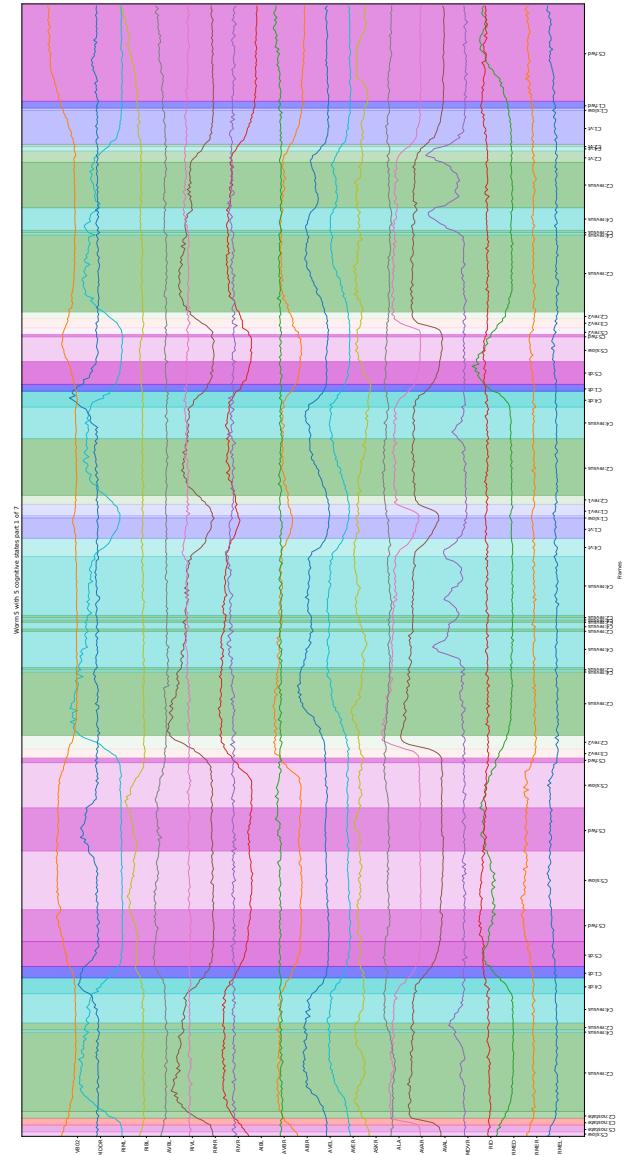


Figure 41: [TBD]

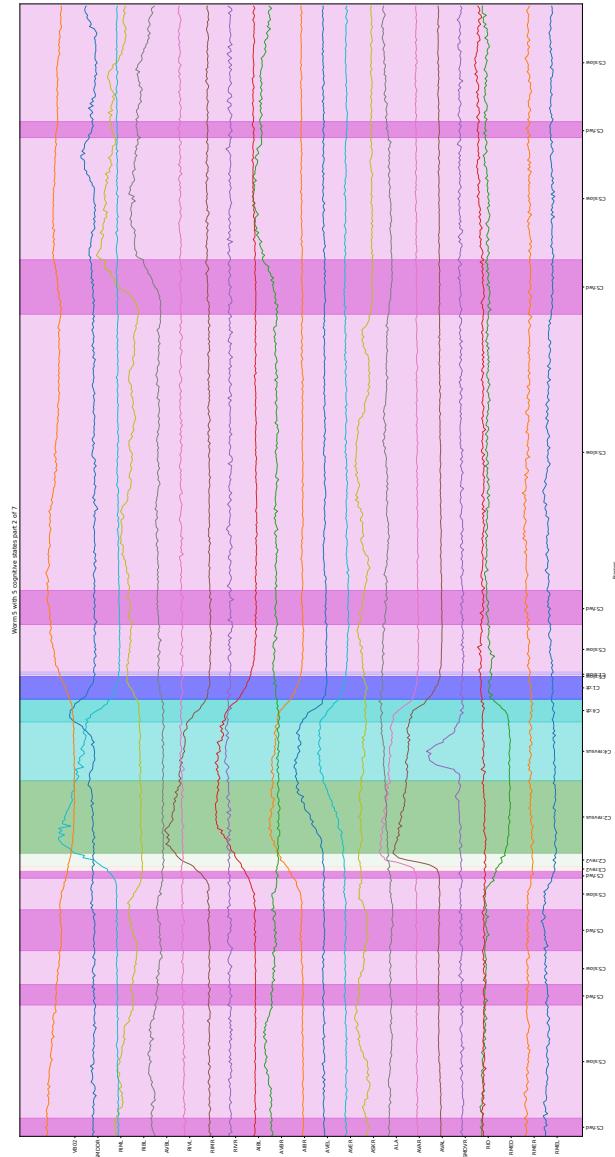


Figure 42: [TBD]

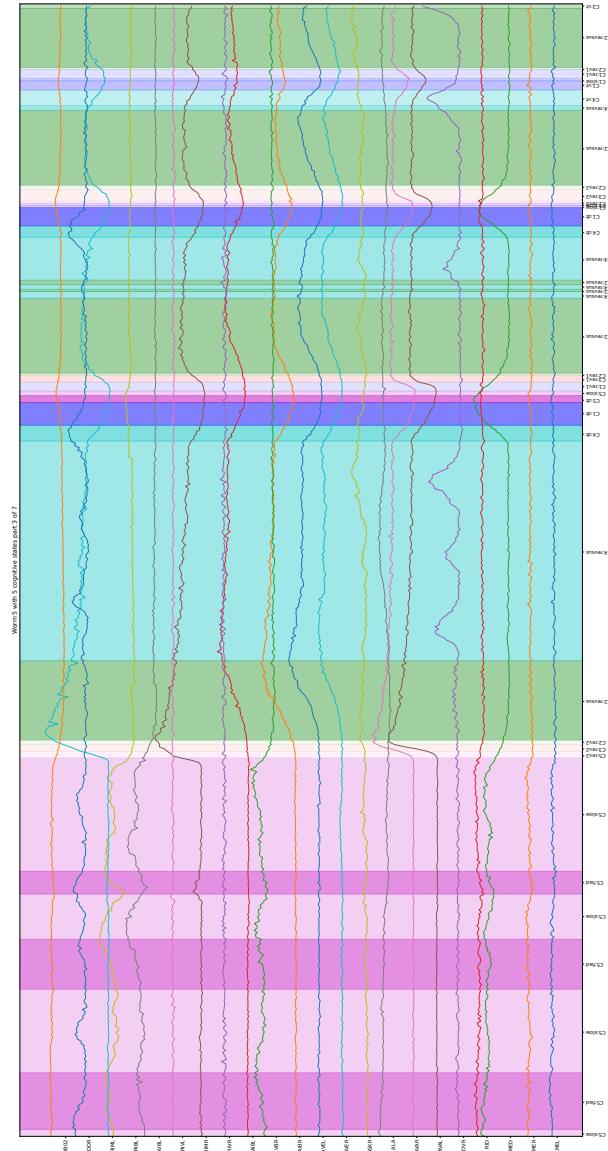


Figure 43: [TBD]

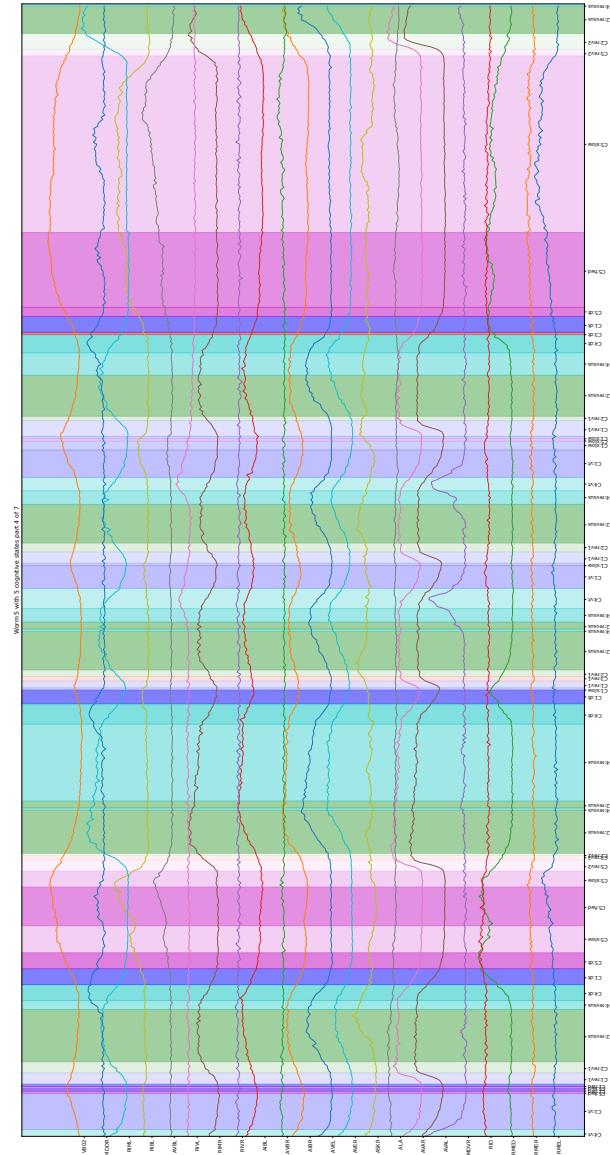


Figure 44: [TBD]

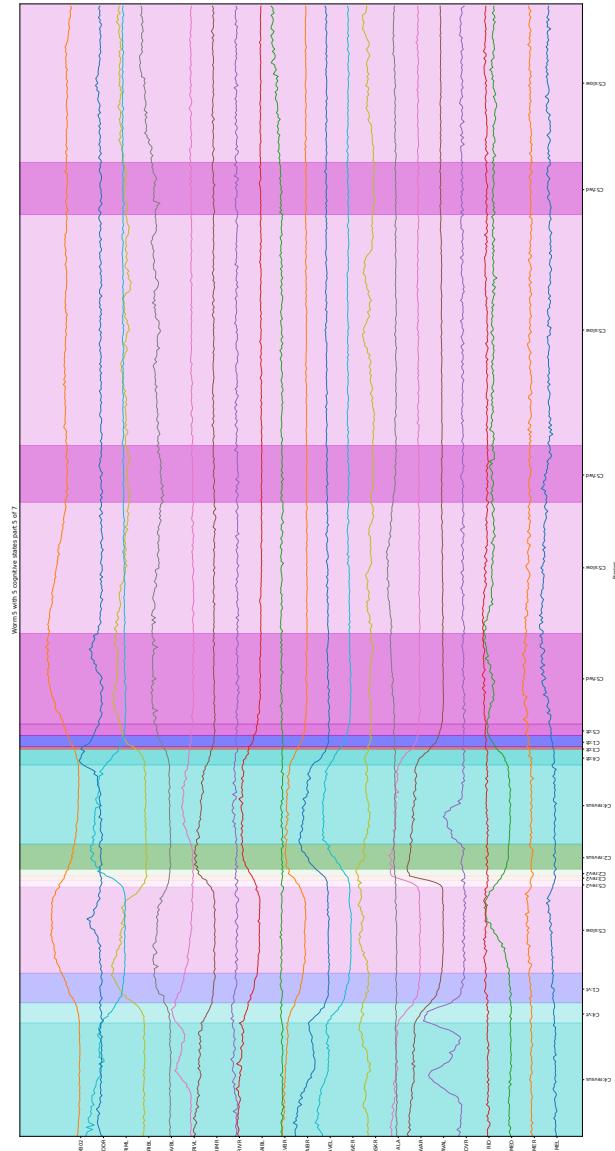


Figure 45: [TBD]

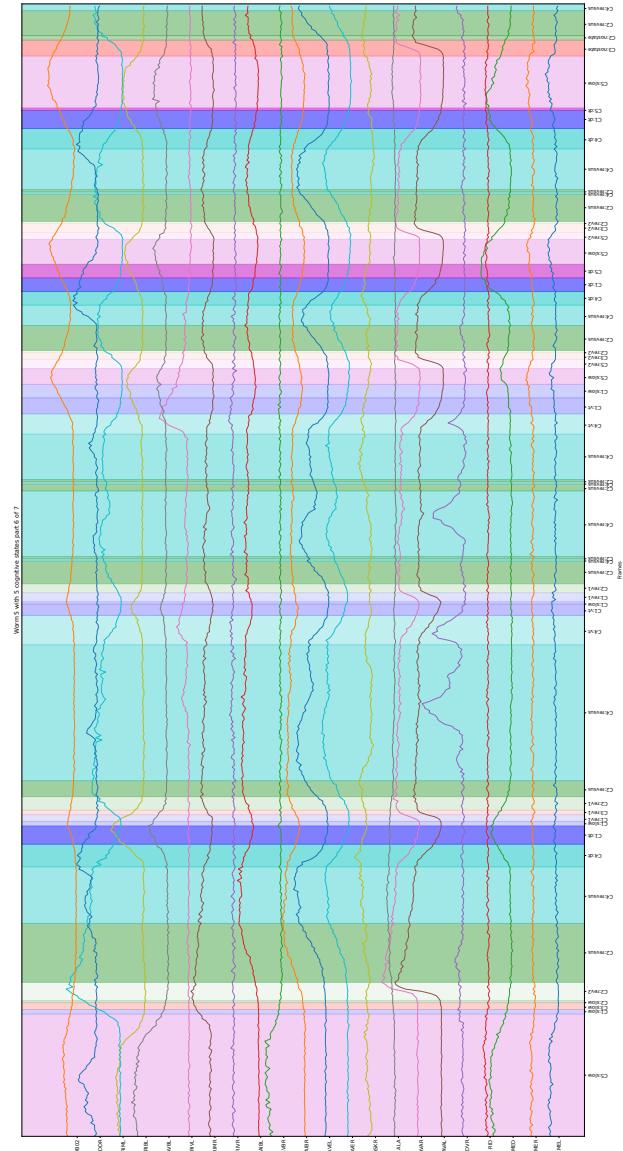


Figure 46: [TBD]

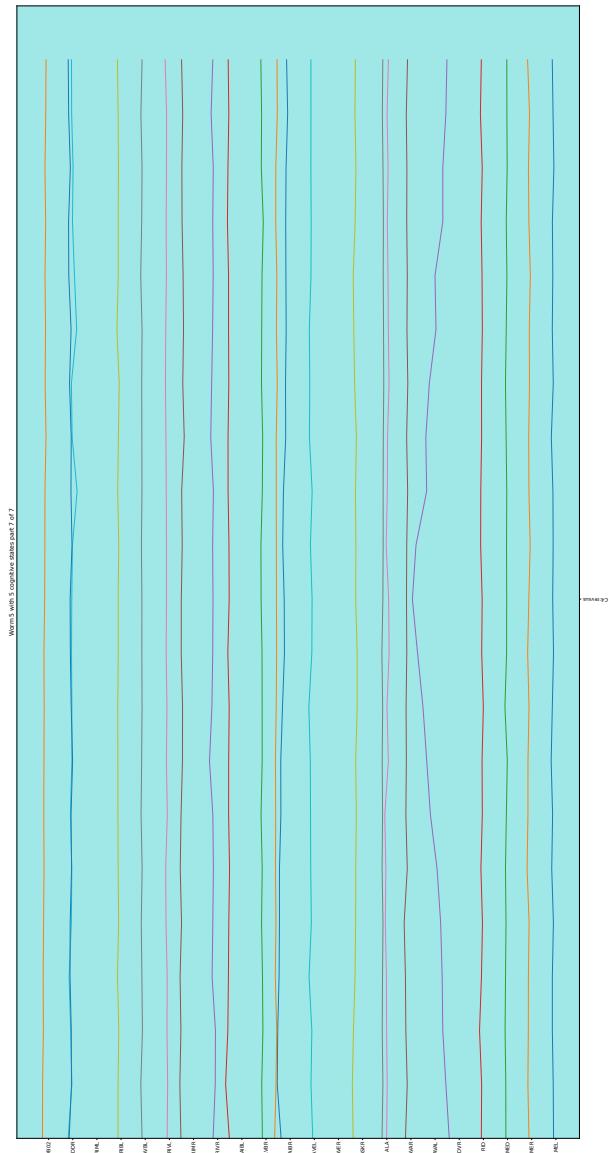


Figure 47: [TBD]