Development of a probabilistic domain-specific language for brain connectivity including heterogeneous knowledge representation

Gaston Ezequiel Zanitti

université
PARIS-SACLAY

Development of a probabilistic domain-specific language for brain connectivity including heterogeneous knowledge representation

Développement d'un langage dédié probabiliste pour la connectivité cérébrale incluant la représentation de connaissances hétérogènes

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 , Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Sciences du traitement du signal et des images
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **INRIA Saclay-Île-de-France** (Univesité Paris-Saclay, INRIA), sous la direction de **Demian WASSERMANN**, Directeur de recherche.

Thèse soutenue à Paris-Saclay, le 15 Mars 2023, par

Gaston Ezequiel ZANITTI

Composition du jury
Membres du jury avec voix délibérative

Pierre SENELLART Professeur, École normale supérieure - PSL	Président & Rapporteur
Daniel MARGULIES Directeur de recherche, CNRS, INCC	Rapporteur & Examinateur
Meghyn BIENVENU Directrice de recherche, CNRS, LaBRI	Examinatrice
Pierre BOURHIS Chargé de recherche, CNRS, CRIStAL	Examinateur
Maria Vanina MARTINEZ Chargée de recherche, IIIA-CSIC	Examinatrice

THESE DE DOCTORAT

NNT : 2023UPASG022

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Développement d'un langage dédié probabiliste pour la connectivité cérébrale incluant la représentation de connaissances hétérogènes

Mots clés: Datalog, Hypothèse de Monde Ouvert, Programmation Probabiliste, Query Answering, Méta-analyse, Neuro-imagerie.

Résumé: Grâce aux récents progrès technologiques, le chercheur en neurosciences dispose d'une quantité croissante de jeux de données pour étudier le cerveau. La multiplicité des travaux dédiés a également produit des ontologies encodant des connaissances à la pointe concernant les différentes aires, les schémas d'activation, les mots-clés associés aux études, etc. Il existe d'autre part une incertitude inhérente aux images cérébrales, du fait de la mise en correspondance entre voxels – ou pixels 3D – et points réels sur le cerveau de différents sujets. Malheureusement, à ce jour, aucun cadre unifié ne permet l'accès à cette mine de données hétérogènes avec l'incertitude associée, obligeant le chercheur à se tourner vers des outils ad hoc.

Dans cette étude, nous présentons NeuroLang, un langage probabiliste basé sur de la logique de premier ordre, comprenant des règles existentielles, de l'incertitude probabiliste, l'intégration d'ontologies reposant sur l'hypothèse du monde ouvert, ainsi que des mécanismes garantissant une réponse aux requêtes résolvables, même sur de très grandes bases de données. Nous soutenons que NeuroLang, par l'expressivité de son langage de requête, contribuera à grandement améliorer la recherche en neurosciences, en donnant notamment la possibilité d'intégrer de manière transparente des données hétérogènes, telles que des ontologies avec des atlas cérébraux probabilistes. Dans ce cas-ci, des domaines cognitifs – à la granularité fine – et des régions cérébrales seront associés via un ensemble de critères formels, favorisant ainsi la communication et la reproductibilité des résultats d'études sur les fonctions cérébrales. Aussi croyons-nous que NeuroLang est à même de se positionner en tête sur ces approches numériques qui visent à formaliser la recherche neuroscientifique à grande échelle via la programmation probabiliste et logique du premier ordre.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Development of a probabilistic domain-specific language for brain connectivity including heterogeneous knowledge representation

Keywords: Datalog, Open-world Assumption, Probabilistic Programming, Query Answering, Meta-Analysis, Neuroimaging

Abstract: Researchers in neuroscience have a growing number of datasets available to study the brain, which is made possible by recent technological advances. Given the extent to which the brain has been studied, there is also available ontological knowledge encoding the current state of the art regarding its different areas, activation patterns, keywords associated with studies, etc. Furthermore, there is inherent uncertainty associated with brain scans arising from the mapping between voxels—3D pixels—and actual points in different individual brains. Unfortunately, there is currently no unifying framework for accessing such collections of rich heterogeneous data under uncertainty, making it necessary for researchers to rely on ad hoc tools.

In this work we introduce NeuroLang, a probabilistic language based on first-order logic with existential rules, probabilistic uncertainty, ontologies integration under the open world assumption, and built-in mechanisms to guarantee tractable query answering over very large datasets. We propose that NeuroLang provides a substantial improvement to cognitive neuroscience research through the expressive power of its query language. We can leverage the ability of NeuroLang to seamlessly integrate useful heterogeneous data, such as ontologies and probabilistic brain atlases, to map fine-grained cognitive domains to brain regions through a set of formal criteria, promoting shareable and highly reproducible research on the domains of brain function. We believe that NeuroLang is well suited for leading computational approaches to formalize large-scale neuroscience research through probabilistic first-order logic programming.

A Luis y Claudia, mis gigantes.

A Stephanie, mi faro en este viaje.

Contents

1 - Résumé étendu en français

Grâce aux récents progrès technologiques, le chercheur en neurosciences dispose d'une quantité croissante de jeux de données pour étudier le cerveau. La multiplicité des travaux dédiés a également produit des ontologies encodant des connaissances à la pointe concernant les différentes aires, les schémas d'activation, les mots-clés associés aux études, etc. Il existe d'autre part une incertitude inhérente aux images cérébrales, du fait de la mise en correspondance entre voxels – ou pixels 3D – et points réels sur le cerveau de différents sujets. Malheureusement, à ce jour, aucun cadre unifié ne permet l'accès à cette mine de données hétérogènes avec l'incertitude associée, obligeant le chercheur à se tourner vers des outils ad hoc.

Dans cette thèse, nous présentons NeuroLang, un langage probabiliste basé sur de la logique de premier ordre, comprenant des règles existentielles, de l'incertitude probabiliste, l'intégration d'ontologies reposant sur l'hypothèse du monde ouvert, ainsi que des mécanismes garantissant une réponse aux requêtes résolvables, même sur de très grandes bases de données. Nous soutenons que NeuroLang, par l'expressivité de son langage de requête, contribuera à grandement améliorer la recherche en neurosciences, en donnant notamment la possibilité d'intégrer de manière transparente des données hétérogènes, telles que des ontologies avec des atlas cérébraux probabilistes. Dans ce cas-ci, des domaines cognitifs – à la granularité fine – et des régions cérébrales seront associés via un ensemble de critères formels, favorisant ainsi la communication et la reproductibilité des résultats d'études sur les fonctions cérébrales. Aussi croyons-nous que NeuroLang est à même de se positionner en tête sur ces approches numériques qui visent à formaliser la recherche neuroscientifique à grande échelle via la programmation probabiliste et logique du premier ordre.

Cette thèse est divisée en quatre parties : Background, Main Contributions,Real-World Use Cases in Neuroscience Research et Discussion. Nous présentons maintenant le plan de chaque chapitre :

Dans **Background**, nous introduisons certains des concepts nécessaires à la compréhension de la portée de cette thèse et des difficultés rencontrées. Il se compose de quatre sections : Dans la première section, nous nous plongerons dans la complexité du cerveau et les améliorations technologiques qui ont permis à la science d'avancer jusqu'à aujourd'hui, mais aussi dans les problèmes que cette avancée génère pour nous. Dans la deuxième section, nous introduirons l'idée de la programmation logique, nous expliquerons la syntaxe Datalog sur laquelle NeuroLang est basé et nous présenterons certaines des différences entre NeuroLang et Datalog, que le lecteur doit comprendre avant de naviguer à travers les exemples proposés dans ce manuscrit. Dans la troisième section, nous présenterons l'idée de la méta-analyse et pourquoi nous pensons qu'il s'agit d'un excellent cas d'utilisation

pour démontrer les caractéristiques les plus remarquables de NeuroLang. Enfin, dans la quatrième section, nous nous plongerons dans les ontologies et l'idée de l'hypothèse du monde ouvert, les avantages de leur intégration dans la réponse aux requêtes, et certaines des difficultés qui doivent être surmontées à de telles fins.

En **Main Contributions**, comme son nom l'indique, nous présenterons les deux contributions les plus importantes de cette thèse. Tout d'abord, nous ferons une présentation détaillée de l'architecture du système de réponse aux requêtes Neurolang, accompagnée de quelques exemples simples qui nous permettront de mettre en évidence certaines de ses caractéristiques les plus intéressantes. Ensuite, dans la deuxième contribution, nous nous plongerons dans les ontologies et l'hypothèse du monde ouvert pour présenter une expérience menée à l'aide de NeuroLang qui présente une caractérisation multi-niveaux des régions histologiques concernant les processus cognitifs.

Dans **Real-World Use Cases in Neuroscience Research** nous montrons deux applications du monde réel qui utilisent NeuroLang pour répondre à des questions d'actualité en neurosciences. Ces deux travaux montrent comment certaines des caractéristiques les plus pertinentes de NeuroLang sont adaptées au monde réel : l'intégration de connaissances hétérogènes et ontologiques dans le premier travail et les requêtes de ségrégation dans le second.

Enfin, dans la **Discussion**, nous passerons brièvement en revue les concepts les plus importants présentés au cours de cette thèse tout en introduisant quelques idées sur le développement futur de NeuroLang. Ces idées concernent des pistes de recherche que nous n'avons pas empruntées et qu'il serait intéressant d'explorer, mais aussi des idées personnelles sur des améliorations possibles qui pourraient être réalisées dans un futur proche.

2 - INTRODUCTION

2.1 . The Thinking Being

There is a point in the history of our planet that marks a before and after. An event that, perhaps imperceptible at that time, will transform the world in every corner: the emergence of the thinking being. This event triggers a fundamental paradigm shift; from this moment onwards, our capacity to generate questions will surpass in speed the capacity by which we can offer answers. A paradigm that will push humanity towards the path of studying, explaining, and attempting to predict the observable universe's social, artificial, and natural phenomena with nothing more than the pure intention of understanding it.

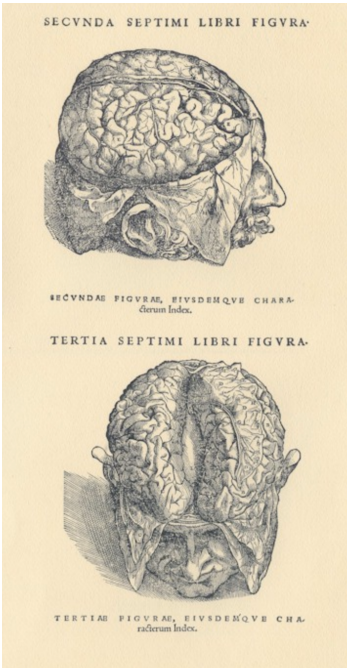

Figure 2.1: Vesalii, Andreae. De humani corporis fabrica libri septem

It is in order to guide and quantify the progress of this thinking being that what we know today as science arises; a set of systematic, verifiable knowledge that, based on observation, experience and rationalization, will allow our thinking being to organize its knowledge through methods, models, and theories in order to generate new structures of thought. New structures that, in turn, will allow this being to observe the world to which it belongs with a new perspective and to formulate new questions, thus giving rise to a cycle of knowledge that feeds back to the present day.

However, while it is true that our thinking being has taken advantage of this feedback loop of knowledge to reach places he could never have imagined, one destination still eludes him: his own thinking. Understanding the circuits and patterns of brain activity and how these give rise to the emergence of mental processes and behavior is, undoubtedly, one of the most critical questions in today's science in general and neuroscience in particular. In this way, the human brain of this being faces the titanic task of, stripped of all subjectivity, understanding itself. Fortunately, and for the benefit of our thinking being, the feedback cycle of sciences has not stopped.

2.2 . A Flood of Data

Technological advances in recent decades have allowed us to access information from the brain in vivo and in a non-intrusive manner with a resolution never imagined. One example of these advances is functional magnetic resonance imaging (fMRI), which uses a technique known as BOLD-contrast imaging (for Blood-Oxygen-Level Dependent contrast imaging) to produce an image that reflects local blood oxygen levels at each point in the brain [55]. The increase in oxygen level in an area is associated with an increase in neuronal activity due to the hemodynamic response generated by neurovascular coupling [34]. fMRI is a remarkable scientific breakthrough that has significantly contributed to the understanding of the human brain that we have today.

However, despite scientific and technological advances, neuroscience is at a turning point. On the one hand, recent studies show that current approaches to classifying brain areas, such as relative location, cell population type, or connectivity, are insufficient to characterize a cortical area and its function unequivocally, hindering its reproducibility and progress in neuroscience. On the other hand, the interest in unraveling the mysteries of the brain has given rise to a growing number of projects developed in the last decades that have produced large heterogeneous databases, including ontologies, multidimensional images, and demographic indicators, among others [72, 52, 33, 79]. As a consequence, neuroscience is faced with a flood of data that it needs to be able to harness in order to continue advancing in its titanic task.

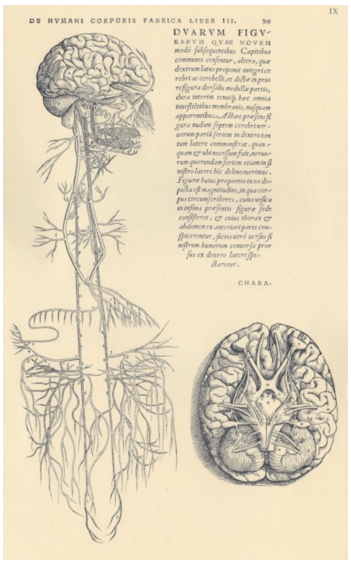

Figure 2.2: Vesalii, Andreae. De humani corporis fabrica libri septem

A powerful approach to synthesize neuroimaging results and data unbiasedly is meta-analysis [53]. Through a handful of meta-analysis tools developed over the past decades, researchers have been able to agree on results and derive latent patterns of brain-behavior mapping [33, 79]. However, most of the time, meta-analyses also require combining data from heterogeneous sources, such as brain activity patterns, brain atlases, textual terms, topic models, and formal cognitive ontologies [58]. However, commonly used meta-analysis tools are not keeping up with the rapid expansion and diversification of the field, impeding the search for domains of brain function.

Thus, neuroscience urgently needs a universal standard for specifying neuroanatomy and function: a way to allow neuroscientists to combine heterogeneous datasets and simultaneously specify tissue characteristics, relative location, known function, and topology of connections for the unambiguous identification of a given brain region.

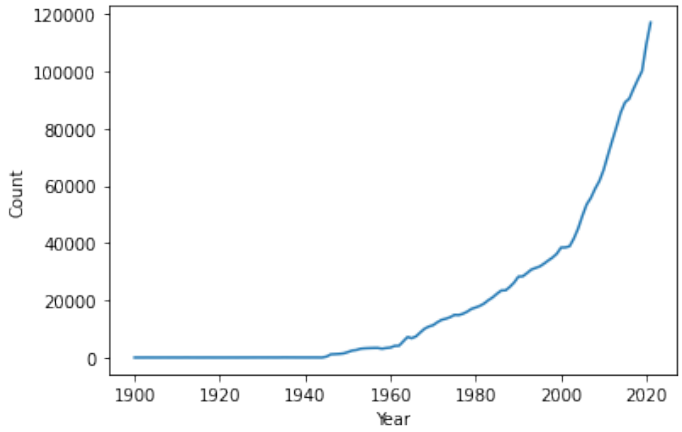

Figure 2.3: Studies mentioning the term brain from 1900 onwards, according to the PubMed.gov database.

2.3 . Need of a Unified Framework

Researchers in neuroscience have a growing number of datasets available to study the brain, which is made possible by recent technological advances [72, 52, 33, 79]. Given the extent to which the brain has been studied, there is also available ontological knowledge encoding the current state of the art regarding its different areas, activation patterns, keywords associated with studies, etc. Furthermore, there is inherent uncertainty associated with brain scans arising from the mapping between voxels—3D pixels—and actual points in different individual brains. Unfortunately, there is currently no unifying framework for accessing such collections of rich, heterogeneous data under uncertainty, making it necessary for researchers to rely on ad hoc tools.

In this work, we introduce NeuroLang, a probabilistic language based on first-order logic with existential rules, probabilistic uncertainty, ontologies integration under the open world assumption, and built-in mechanisms to guarantee tractable query answering over very large datasets. We propose that NeuroLang provides a substantial improvement to cognitive neuroscience research through the expressive power of its query language. We can leverage the ability of NeuroLang to seamlessly integrate valuable heterogeneous data, such as ontologies and probabilistic brain atlases, to map fine-grained cognitive domains to brain regions through a set of formal criteria, promoting shareable and highly reproducible research on the domains of brain function. We believe NeuroLang is well suited for leading computational approaches to formalize large-scale neuroscience research through probabilistic first-order logic programming. After presenting the language and its general query-answering architecture, we discuss real-world use cases showing how NeuroLang can be applied to practical scenarios.

2.4 . Organisation of this dissertation

This thesis is divided into four parts: Background, Main Contributions, Real-World Use Cases in Neuroscience Research and Discussion. We now present the outline of each chapter:

In **Background**, we introduce some of the concepts necessary to understand the scope of this dissertation and the difficulties encountered. It consists of four sections: In the first section, we will delve into the complexity of the brain and the technological improvements that allowed science to advance to the present day, but also into the problems that this advance is generating for us today. In the second section, we will introduce the idea of logic programming, explain the Datalog syntax on which NeuroLang is based and we will introduce some of the differences between NeuroLang and Datalog, which the reader needs to understand before navigating through the examples proposed in this manuscript. In the third section, we will present the idea of Meta-Analysis and why we believe it is an excellent use case to demonstrate the most outstanding features of NeuroLang. Finally, in the fourth section, we will delve into ontologies and the idea of the Open-world assumption, the advantages of integrating them in query answering, and some of the difficulties that need to be overcome for such purposes.

In **Main contributions** as the name suggests, we will present this thesis's two most important contributions. First, we will give a detailed presentation of the architecture of the Neurolang query answering system, accompanied by some simple examples that allow us to highlight some of its most exciting features. Then, in the second contribution, we will delve into ontologies and the open-world assumption to present an experiment conducted using NeuroLang that presents a multilevel characterization of histological regions concerning cognitive processes.

In **Real-World Use Cases in Neuroscience Research** we show two real-world applications that use NeuroLang to answer questions of current neuroscientific relevance. These two works show how some of the most relevant features of NeuroLang are adapted to the real world: the integration of heterogeneous and ontological knowledge in the first work and segregation queries in the second.

Finally, in **Discussion**, we will briefly review the most important concepts presented during this dissertation while introducing some ideas about the future development of NeuroLang. These ideas involve research avenues that we did not take and would be interesting to explore, but also personal ideas about possible improvements that could be carried out in the near future.

Part I

Background

3 - On brains and data

3.1 . Complexity of the brain

While the functionality carried out by most of the vital organs of the human being is unremarkable, the human brain sets us apart from the rest of life on our planet. The brain is a complicated tissue composed of approximately 86 billion neuronal cells, of which 16 billion form part of the cerebral cortex [39]. Each of these 16 billion neocortical neurons has an average of 7000 synaptic connections, giving a total of 150,000 to 180,000 kilometers of myelinated neuronal fibers in an average adult at the age of 20 [29]. In addition, there is another 85 billion non-neuronal cells [39]. Furthermore, let us not forget that all these cells and neural fibers are contained within an organ that weighs about a kilo and a half and is "enclosed" within the cave of bones we call the skull.

The brain, unlike other organs, is composed of exquisitely specialized substructures involved in motor, sensory and integrative functionalities [51]. While it is true that, given the abundance of cells and fibers, one would think that the brain has enough redundancy and plasticity to be able to lose a number of these without too many consequences, it is partly because of this hyperspecialization of different regions that it is not only the number of neurons that matters but also the region where the loss occurs. Donald O. Hebb, in his famous "The Organization of Behavior: A Neuropsychological Theory" [38] of 1949, echoed this situation:

> "The effect of a clear cut removal of cortex outside the speech area is often astonishingly small, at times no effect whatever can be found (Hebb, 1942a, 1945b). It is possible that there is always a loss of intelligence in aphasia, when the "speech area" is seriously damaged, but this does not, of course, explain why damage elsewhere should have no effect. It would be unreasonable to suppose that most of the cortex has nothing to do with intelligence, and there are, in fact, definite indications that this is not true. Intelligence must be affected by any large brain injury, yet sometimes it seems not to be."

In the face of this complexity, how do we then measure brain responses?

3.2 . The short history of how we measure the brain

The earliest written record of the anatomical description of a brain dates back to ancient Egypt. This document was written around 1600 BC and is known as the Edwin Smith papyrus (1822-1906), after the American farmer, antiquarian, and Egyptologist who discovered it. However, the contents of this manuscript

are believed to be a copy of a treatise written between 3000 - 2500 BC. The document contains 48 well-structured clinical traumatology cases and is devoted to civil construction injuries. In one of the cases, as a result of an open head injury, we can read the first known description of the history of the brain:

> "On examining the wound, one can touch the viscera of the skull, feeling ripples resembling the slagging of molten copper. Sometimes the brain beats under the fingers in the same way as the fontanelles of small children".

The subsequent advances are to be found in the Greek medical school founded in Alexandria during the 3rd century BC. There, Herophilus (325 - 280 BC) and Erasistratus (304 - 250 BC) were the first Greek physicians to perform systematic dissections of human cadavers. Unhappily, the work of both disappeared entirely with the destruction of the first library in Alexandria by Julius Caesar. We know about it through quotations from later authors, especially Galen. Furthermore, after the spread of Christianity during the Middle Ages, dissection was considered blasphemous, and its practice was forbidden.

In the Italian Renaissance (1450-1600), the brain's systematic anatomical and physiological studies were resumed because it became possible to dissect human cadavers. In 1543, the anatomist Andrew Vesalius (1514-1564) published De Humani Corporis Fabrica. The book is based on the lectures the author gave at the University of Padua, during which he carried out countless dissections of cadavers to illustrate his expositions. He presents a detailed examination of the organs and a complete structure of the human body.

With the dissemination of the first photographic process, the daguerreotype, in 1839 by Louis Daguerre (1787 - 1851), photography began to revolutionize the world. This new tool promised scientists a more objective representation than drawings and engravings. Thus, in 1873, Jules Bernard Luys (1828 - 1897) published the first photographic neuroanatomical atlas: Iconographie Photographique des Centres Nerveux. However, photography, despite its importance in various fields of science, was not widely used for the study of the brain in the decades following the publication of the first atlas.

During this time, the theory of the functional specialization of the brain began to emerge. The Viennese physician Franz Joseph Gall (1758-1828) proposed that the brain was the basis of the mind, that the mind was composed of different mental faculties, and that each mental faculty resided in a specific brain region. A heated debate on the localization of function in the brain had begun.

In addition to anatomy, another way to study the brain is through imaging (or recording) its functions, such as electrical or metabolic activity. In the case of electrical activity, its study in the nervous system of animals during the last quarter of the 19th century opened the door to the development of electroencephalography. With this technique, we can map electrical activity in the cerebral cortex at rest

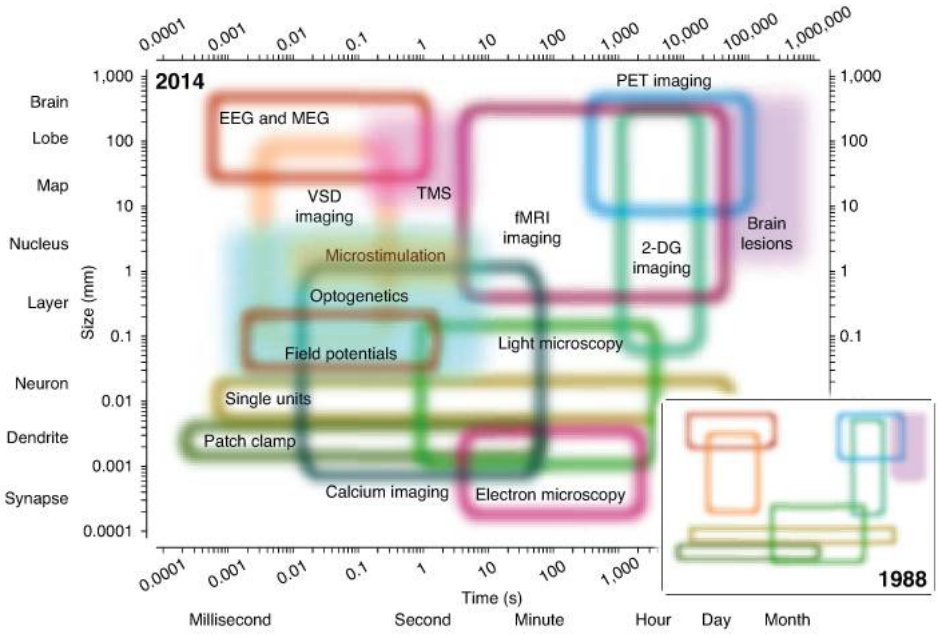

Figure 3.1: The spatiotemporal domain of neuroscience and the main methods available for studying the nervous system in 2014. Each colored region represents the useful domain of spatial and temporal resolution for one method available for the study of the brain. Open regions represent measurement techniques; filled regions, perturbation techniques. Image from Sejnowski et. al. [67]

and in response to a stimulus. The first human study was published in 1929 by the neuropsychiatrist Hans Berger (1873 - 1941).

On the other hand, functional magnetic resonance imaging (fMRI) is one way to analyze the brain's metabolic responses. fMRI is based on a technique known as Blood-Oxygen-Level Dependent (BOLD) contrast imaging [55]. Neuronal signaling processes in the brain, including the formation and propagation of action potentials, resulting in a local increase in energy requirements, which in turn leads to an increase in the oxygen concentration in the area [34]. Functional resonance imaging allows us to measure variations in oxygenated and deoxygenated hemoglobin, and to assume a proportional relationship of this variation with its corresponding activation in the brain.

3.3 . Data unification

It has been a long time since, around 1600 BC, humans wrote the first anatomical description of the brain on papyrus. We have been collecting information about the brain for thousands of years in increasingly complex ways. Now, neuroscience is in the era of big data and open science, with thousands of articles published

annually [53], enabled by large-scale brain mapping initiatives [72, 52, 33, 79]. This new era offers enlightenment and new insights, but it can also be a bane if it obscures, obstructs and overwhelms due to the lack of standards for collaboration or the impossibility of combining heterogeneous data quickly and without human error.

To exploit the full potential of this data, there must be ways to standardize, integrate and synthesize diverse types of data from different levels of analysis and across laboratories. The first step in solving this problem, through a cultural change in the way data sharing between laboratories is done, is already happening in the open with, for example, multimodal data sharing repositories [50, 35]. However, for a multidisciplinary field like neuroscience, a universal framework for heterogeneous data integration and analysis is crucial for bridging the gap between methods and data. For this purpose, we harness the expressive power of NeuroLang, a unified domain-specific language for functional neuroanatomy based on probabilistic first-order logic. With NeuroLang, it is possible to formally express hypotheses, synthesize results, and integrate data from various heterogeneous sources, opening doors to questions that were previously hard to define and answer.

4 - Datalog and Logic programming

Logic programming emerged in the 1970s from debates concerning procedural versus declarative representations of knowledge in artificial intelligence. The driving force behind logic programming is the idea that a single formalism suffices for both logic and computation. Therefore, logic programming offers a slightly different paradigm for computation: computation is logical deduction.

Languages based on the logic programming paradigm are non-procedural, because their programs are more concerned with a formal formulation of the problem than the description of how to solve them: they internally use evaluation engines to derive the solution. This means that the programmer only specifies the facts and rules that define the problem, and the system automatically derives the steps to solve it. This declarative approach allows for a more concise and readable representation of the problem, as well as better adaptability and maintainability of the code. This makes logic programming languages an excellent tool for expressing hypotheses.

One of the prominent exponents of this paradigm is Prolog. Prolog evolved out of research at the University of Aix-Marseille in the early 70's. Alain Colmerauer and Phillipe Roussel, both of University of Aix-Marseille, collaborated with Robert Kowalski of the University of Edinburgh to create the underlying design of Prolog as we know it today. The basic building block behind Prolog's syntax are Horn clauses.

Horn clauses are named after the logician Alfred Horn. A Horn clause logic program is a set of sentences (or clauses) each of which can be written in the form:

$$A_0 \leftarrow A_1 \wedge \cdots \wedge A_n \text{ where } n \geq 0 \tag{4.1}$$

Each A_i is an atomic formula of the form $p(t_1, \ldots, t_m)$, where p is a predicate symbol and the t_i are terms. Each term is either a constant symbol, a variable or a function symbol. Every variable occurring in a clause is universally quantified, and its scope is the clause in which the variable occurs. The backward arrow $\leftarrow$ is read as "if", and $\wedge$ as "and". The atom A_0 is called the conclusion (or head) of the clause, and the conjunction $A_1 \wedge \cdots \wedge A_n$ is the body of the clause. The atoms $A_1, \ldots, A_n$ in the body are called conditions. If n = 0, then the body is equivalent to true, and the clause $A_0 \leftarrow true$ is abbreviated to A_0 and is called a fact. Otherwise if $n \neq 0$, the clause is called a rule.

Submitting a query means asking Prolog to prove that the statement(s) implied by the question can be made true as long as the correct variable instantiations are made. The search for such proof is often referred to as goal execution. Each predicate in the query constitutes a (sub)goal, which Prolog tries to satisfy one after the other.

With their ability to use predicates to relate objects with one another, logic programs can naturally express relational databases and queries. This is in fact one of the main applications of logic programming. Datalog is perhaps the most famous language for expressing and querying logic databases. Datalog and Prolog are two closely related programming languages that are based on the first-order predicate logic. Both languages provide a declarative and logical approach to computation, and they are widely used for a variety of applications, such as knowledge representation, data integration, and deductive reasoning. However, there are some key differences between Datalog and Prolog, which include:

Syntax and semantics: Datalog is a subset of Prolog, which means that every Datalog program is also a valid Prolog program, but not vice versa. Datalog has a simpler and more restricted syntax and semantics than Prolog, which makes it easier to learn and understand, but also less expressive and flexible.

Evaluation strategy: Datalog and Prolog use different evaluation strategies to derive the solutions of a given program. Datalog usually uses a bottom-up approach, where the rules are applied to the facts in the database to derive new facts, whereas Prolog uses a top-down approach, where the facts and rules are used to satisfy the goals in a query [2].

Decidability and complexity: Datalog is a decidable language, which means that every Datalog program has a finite and polynomial-time evaluation, whereas Prolog is a Turing-complete language, which means that some Prolog programs may not halt or may require exponential time to evaluate. Therefore Datalog programs are more predictable and efficient than Prolog programs, but they are also less powerful and versatile. This difference can be seen in Prolog's ability to contain function symbols in its literals, something that is not supported in Datalog. This allows to generate an infinite domain from a finite set of symbols, and thus simulate an infinite Turing tape. On the other hand, a Datalog program is bounded in size and its queries are solvable in polynomial time [2].

Overall, the main differences between Datalog and Prolog lie in their syntax and semantics, evaluation strategy, and decidability and complexity. These differences can affect the expressiveness, efficiency, and predictability of the two languages, and they can influence the choice of which language to use for a given application. In our case, we decided to use Datalog as the main skeleton and influence of NeuroLang. We will see, mainly in the contributions of this thesis, that we can extend or limit the expressive power of Datalog according to our needs (allowing recursion, negation, etc when nedded). However, the decidability of the language is a fundamental feature for our use cases that we cannot afford to avoid.

4.1 . Datalog syntax

Now we will give a small introduction to the Neurolang syntax, marking differences and some syntactic sugars to facilitate its writing, with respect to the

original Datalog syntax. For a deeper dive into datalog syntax, we ask the reader to refer to *Foundations of databases* [2], which is the reference material on which this chapter is based.

As we mentioned before, Datalog is based on the idea of Horn clauses, therefore we can define a Datalog rule as an expression of the form:

$$A_0(u_0) \leftarrow A_1(u_1), \ldots, A_n(u_n) \tag{4.2}$$

where $n \geq 0$, $A_0, \ldots A_n$ are relation names and $u_0, \ldots u_n$ are free tuples of appropriate arities. Each variable occurring in u_0, must also be occurring in at least one of $u_1, \ldots u_n$. The expression A_0 is called the *head* of the rule, while $A_1, \ldots, A_n$ forms the body. A datalog program is a finite set of Datalog rules. A *positive literal* is an atom, i.e $A_1(u_1)$; and a *negative literal* is the negation of one, i.e, $\neg A_1(u_1)$.

A formula of the form $A_0(u_0), \ldots, A_n(u_n) \leftarrow B_0(u_0), \ldots, B_n(u_n)$ is called a *clause*. However, and this marks the first difference with our implementation, a clause with more than one literal in the head, is not allowed in Neurolang. Clauses with a single literal are called *definite clauses*. In the case that no variable occurs in our clause, it is called a *ground clause*.

Moreover, since Neurolang is a probabilistic language, there are two fundamental differences in the syntax that we must emphasize.

1) Given a *probabilistic atom* of the form $\mathbf{a} : p$, where p is a real number in the interval $[0, 1]$ and $\mathbf{a}$ is an atom with a predicate, Neurolang has the ability to define *probability encoding rules* (PERs), which we will explain in more depth in the following chapters, but which basically allows us to include the probability of the atom, as another element of its tuple, allowing its subsequent manipulation.

2) The ability to define conditional probabilities in a simple way using the double backslash ("//") as syntactic sugar.

We can see an example of both features in Listing 4.1 where given a probabilis-

Listing 4.1: PERs and conditional probability

```
ProbMap(i, j, k, PROB) :-
    Activation(i, j, k)
    // TermAssociation("emotion").
```

tic atom *Activation* and a deterministic atom *TermAssociation*, we can calculate the conditional probability *ProbMap* of an *Activation* in the voxel (i, j, k) given an association with the term "emotion" (*TermAssociation("emotion")*), where the reserved word *PROB* in the head of the rule tells Neurolang that we want to extract the probability of each result atom as a new parameter in *ProbMap*. A more detailed explanation of how PERs work will be given in the next chapters.

Other forms of syntactic sugar present in Neurolang include: 3) The use of @ as a reserved word to describe expressions defining the probabilities of an atom. The use of this feature can be seen in Listing 4.2 where we define the probability of a voxel (i, j, k) based on the modified version of ALE proposed by Eickhoff et. al. [30].

Listing 4.2: Example of the use of @

```
Activation(i, j, k) @ max(
        (exp(-(1 / 2) * (d / sigma) ** 2)
        / ((2 * pi) ** (3 / 2) * sigma ** 3))
        * (4 ** 3)
    ) :- StimTypeAuditory(bmapID, expID),
    BrainMap(bmapID, expID, ..., ..., minSubj, i1, j1, k1),
    Voxels(i, j, k),
    (d == FocusCoactivates(i, j, k, i1, j1, k1)),
    (sigma == sigmaGivenSubjects(minSubj))
```

This modification is based on the idea of using between-subjects and between-templates variance to estimate the size of the modeled Gaussian from which to compute the corresponding FWHM.

4) The use of *exists* as a reserved word, mainly associated with segregation queries, that can be used in a variety of scenarios that require the explicit definition of an existential. Listing 4.3 presents an example of the use of the reserved word *exists* in the context of a segregation query that seeks to obtain only those studies that are associated with the left bin and not the right bin. More details on this experiment are presented in section 10.

Listing 4.3: Example of the use of the reserved word 'exists'

```
OnlyLeftBinActive(bin, study) :- LeftBinActive(bin, study),
    ~exists(bin2;
        Bin(bin2), Study(study), RightBinActive(bin2, study)
    )
```

5 - Meta-analysis as a Use-case

Meta-analysis tools are examples of central neuroscience use cases requiring the combination of heterogeneous datasets. Given the increasing availability of information describing cognitive structures and processes, Meta-analysis constitutes a fertile ground to show how current knowledge representation advancements can combine heterogeneous datasets, pushing forward neuroimaging research. Meta-Analysis is a set of techniques used to combine a finite number of published articles, which often disagree, to infer consensus-based findings [59]. Combining data may improve statistical power when there are several studies on a specific question, but each one of them is largely under-powered or has not been designed to address that research question [40]. While the inappropriate use of these techniques can lead to statistical errors and results can be misleading, some general rules can help solve these problems [53]. This allows us to have a tool that can provide an accurate and robust estimate after a systematic and rigorous integration of the available evidence.

A popular method of performing a meta-analysis of neuroimage data is Coordinate-based Meta-Analysis (CBMA), which tests for consistent activation of the same anatomical regions. CBMA aims to find results that indicate replicable effects across studies. By analyzing multiple studies simultaneously, those results replicated across at least some can be identified and assumed to be relevant. CBMA databases are then built by combining the extracted coordinates of reported peak activations and a set of terms from neuroimaging studies.

Currently, there are three widely used CBMA methods: ALE, KDA, and MKDA [64]. Briefly, ALE, or activation likelihood estimation [70] generates "likelihood" maps for each activation focus by placing a 3D Gaussian density with full width at half maximum (FWHM) specified at the focus location with the idea that activation foci are more accurately viewed not as single points, but as localization probability distributions centered at the given coordinates; these maps are then combined to create a whole-brain map assigning each voxel within the brain a value equal to the probability that at least one of the points in the dataset actually lay within the voxel. KDA, or kernel density analysis [76], also treats each focus independently but instead uses a spherical kernel and a simple addition rule to produce a map showing the number of foci within a given radius. MKDA, or multilevel KDA [77] where the binary maps is based on the proportion of studies that activate in a region rather than the number of peaks, showing where there are one or more foci within a given radius; these binary maps of studies are then averaged, giving the proportion of studies that have any focus within a given radius from a voxel.

At present, some of the current standard tools that implements variations of this CBMA methods are Neurosynth [79], NeuroQuery [28], and BrainMap [46], which harness automatically extracted as well as manually-curated information

present across neuroscientific articles. Briefly, these tools interpret each article as an independent sample of *neuroscientific knowledge* and then develop query systems centered on study subset selection and posterior probabilistic inference on such subsets. For instance, selecting all studies mentioning "fear" and inferring the most common areas of the brain reported as active—i.e., deferentially oxygenated—in such studies. In these tools, queries select a subset of around 15k full-text articles reporting involvement of several brain locations each and a brain tessellation of 300k cubes, or voxels, then infer commonalities across these articles through maximum likelihood estimations combined with spatial information smoothing. Such queries can express questions like "*Where do articles reporting the term 'emotion' show activations?*", or "*Which terms associated with cognitive processes are most likely associated with articles reporting activations in the amygdala?*". Finally, after the inferential tasks, the obtained probabilities are manipulated and aggregated to frame results into the frequentist language neuroscientists commonly use to communicate the significance of their results [79, 65]. These meta-analyses are performed in under 30 seconds on a regular laptop computer—however, these tools are limited in terms of the expressivity of their associated query languages.

Neurosynth combines text mining, meta-analysis, and machine learning techniques to generate probabilistic mappings relating text-mined terms with activations in the human brain. While NeuroSynth has proven to be of great value to the neuroscience community, the language used to infer these relationships is based on propositional logic, which can limit the expressiveness of its query system. This limitation excludes, for instance, the use of existential quantifiers and negation, forbidding queries such as "*What are the terms most probably mentioned in articles reporting activations in the parietal lobe and no other brain region*", which we dub *segregation queries*. Another example is BrainMap, which has a hand-curated dataset of great precision and an ontology for structuring all this knowledge and annotating the articles. Nonetheless, Brainmap's query system is also based on propositional logic. It only allows selecting terms mentioned in articles knowing them in advance, which again cannot express segregation queries or harness the full information of neuroscience ontologies—such as CogAt [58]—that use open knowledge.

While recent advances in automated meta-analysis techniques are mostly centered on better representing spatial correlations [65], to the best of our knowledge, none have formally addressed expressivity limitations of query languages and the feasibility of a more expressive resolution. Breaching the expressivity limitations of current approaches and handling heterogeneous data requires tackling several issues: handling noisiness in neuroimaging data and conclusions reported across studies calls for a unifying formalism with probabilistic modeling capabilities; being able to leverage ontological information modeled under the open world assumption; finally, performance cannot be ignored since the amount of information needed to model the human brain is considerable. In short, we need to design a logic-based

language capable of:

- dealing with existentially quantified variables.
- performing negation and aggregation
- performing probabilistic inference
- post-processing inferred probabilities
- dealing with neuroimaging databases having, at least, a similar performance to current meta-analytic tools

Our main proposal in this work is the development of a subset of Datalog+/– extended with probabilistic semantics, aggregation, and negation, focused on meta-analytic applications. Such an approach allows us to have a language based on first-order logic with negation and existential ($FO^{\neg\exists}$), enabling more complex queries such as segregation queries or manipulation of information under the open-world assumption. In all, we produce a language able to express the full breadth of the pipeline needed for meta-analytic applications: from data preprocessing to probabilistic modeling and inference, and finally, the post-processing of probabilistic results into images and reports that are easily interpretable in terms of current reporting used in neuroscience publications. Our main contribution is the introduction and evaluation of NeuroLang, a probabilistic language based on Datalog+/– developed to express and solve rich logic-based queries meeting the functional requirements of neuroimaging meta-analyses.

6 - Ontologies and the Open World Assumption

Among the variety of data structures that we can find to represent neuroscientific knowledge, there is one that will be of particular interest to us because of the advantages it presents but also because of the challenges it implies: knowledge graphs, also know as ontologies.

Ontologies are logical theories that formalize domain-specific knowledge and consist of a formal representation of information as a set of concepts and the relationships between instance of them, thereby making them available for machine processing. This way of structuring knowledge hierarchically through relationships gives ontologies a high expressive power. As a consequence, an incipient interest in the scientific community to develop techniques that allow the combination of traditional databases with ontological databases, has arisen with a particular focus on use the knowledge provided by the ontology to answer queries over an incomplete database under the open world semantics. This technique where an extensional database $\mathcal{D}$ is combined with an ontology Σ, and the input conjunctive query is not evaluated against the database $\mathcal{D}$ in the traditional way, but against the logical theory $\mathcal{D} \cup \Sigma$, is known as *ontology-mediated query answering (OMQA)* [13]. Some of the possible use cases for this approach include:

- The enrichment of incomplete data sources with the expert knowledge provided by an ontology in order to obtain a more complete set of answers. For example, an ontology that associates cognitive processes with specific brain regions could be used to improve the results provided by a meta-analytic database where, auditory results for example, also tend to present activations in the motor cortex produced by the fact that the patient is usually asked to press a button or raise their hand to report hearing a certain sound.

- The enrichment of data schemas (i.e., the relationship symbols used in the representation of data) with additional symbols to be used in queries. For example, in the case of terms that refer to cognitive processes such as *pain* and *nociception*, we could have an ontology that expands the schema of our data, adding information about the synonymy of these terms and that they can be used interchangeably, allowing us to nourish our query results with more information.

- The use of ontologies as data integrators, where an ontology can be used to provide a unified view of different datasets. For example, an ontology that combine regions and sub-regions of different atlases of the human brain through the use of a unified set of entities.

However, while queries are typically specified as unions of conjunctive queries, the languages by which ontologies are usually defined are *description logics (DLs)* [6], a family of knowledge representation languages, which is a subset of first-order logic [63]. It's interesting to note that DLs form the logical basis of the Web Ontology Language (OWL) [66] and its revision OWL2, standardised by the W3C, one of the most widely used languages for the definition of ontologies. Nevertheless, the complete OWL language (called OWL Full to distinguish it from the subsets) was designed to have maximum expressiveness, but without computational guarantees, which has led to extensive research to find language fragments that are capable of guaranteeing this property. In seeking to overcome this situation, several lightened versions of DL have been proposed, which guarantee decidability but also a polynomial time response for conjunctive queries, in relation to the complexity of the data. Some of these versions include $\mathcal{EL}$[5] and the members of the DL-Lite family[17]: DL-*Lite*$_R$, DL-*Lite*$_F$ and DL-*Lite*$_A$. These languages are tractable fragments of OWL and, actually, the language DL-*Lite*$_R$ forms the OWL 2 QL profile of OWL 2.

DLs are equipped with formal semantics that allows humans and computers to exchange information unambiguously and make possible the creation of reasoning systems capable of inferring additional information from the facts explicitly stated in an ontology. Leveraging the fact that they are based on a subset of first-order logic, a DL ontology doesn't describe a particular state of the world but instead consists of a set of rules called axioms (also known as ABox statements), each of which must be true in the world described. These axioms usually capture only partial knowledge of the situation described by the ontology; hence there may be many states of the world that are consistent with the ontology.

Having a way of expressing ontologies with emphasis on reasoning as a core principle, that guarantee decidability and polynomial time response for conjunctive queries was the necessary stepping stone for the development of ontologies to take off. In the field of neuroscience, the interest in ontologies is reflected in the emergence of projects such as Cognitive Atlas (CogAt) [58], Foundational Model of Anatomy (FMA) [62], Cognitive Paradigm Ontology (CogPO) [71], among others. The growing literature available on the human brain, and its natural division into regions/processes, makes ontologies a suitable structure for the storage of this information. Unfortunately, there is currently no unifying framework for accessing such collections of rich heterogeneous data under uncertainty, making it necessary for researchers to rely on ad-hoc tools and despite all their potential benefits, the use of ontologies is still underappreciated in the area [59]. With this in mind, we decided to, in the context of NeuroLang, offer a tool that allows neuroscientists to integrate ontologies, enabling them to use the expert knowledge embedded in these data structures, while providing a simple way to integrate this information with the inherent uncertainty associated with for example, brain scans, arising from the mapping between voxels and actual points in different individual brains.

However, ontological knowledge is interpreted under the open-world assumption (OWA), which entailed that the model is a representation of partial knowledge about a domain. This means that the facts asserted in the model are not assumed to be complete; if a statement cannot be inferred as true or false about an object, it's assumed to be unknown [2]. For this reason, the open world assumption makes conclusions depend not only on the information contained in the knowledge base, but also on unknown/missing and plausible information. This means that using the information contained therein we can conclude that a certain assertion (query) holds but we cannot simply deny it if there is insufficient evidence. This is because information that could make the conclusion to be false might not be present in the current knowledge base. As a consequence, it's a necessary condition that any attempt to reason on an ontological knowledge database must take this characteristic into account. The open-world assumption helps to solve the problem of data incompleteness, by allowing inference of new facts from the constraints proposed by the ontology.

In the context of OWL, the most popular language for defining ontologies, open knowledge can be present, for example, under a property known as *someValuesFrom*. The inclusion of this constraint in our program results in the creation of a rule with an existential in the head by which information is defined under the open-world assumption [63]. Let's see what this restriction means [66]:

> "A restriction containing a *someValuesFrom* constraint defines a class of individuals **x** for which there is at least one **y** such that the pair (x,y) is an instance of P."

The following example, which belongs to the Foundational Model of Anatomy (FMA) [62] ontology, defines a class (the *Left Superior Frontal Gyrus*) of individuals which have at least one member under the property of being *RegionalPartOf* of the *Left Frontal Lobe* class:

```
<owl:Class rdf:ID="Left Superior Frontal Gyrus">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#RegionalPartOf" />
      <owl:someValuesFrom>
        <owl:Class rdf:about="Left Frontal Lobe" />
      </owl:someValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

The *someValuesFrom* constraint is analogous to the existential quantifier of first order logic (FOL) - for each instance of the class that is being defined, there exists at least one value for *P* that fulfills the constraint. This would translate into

the following FOL rule:

$$\forall x\,\text{Left Superior Frontal Gyrus}(x) \rightarrow \exists y\,\text{RegionalPartOf}(x, y) \wedge \text{Left Frontal Lobe}(y) \in \Sigma_1^{\text{FMA}}. \tag{6.1}$$

When considering X and Y as voxels, this constraint states that for every voxel belonging to the *Left Superior Frontal Gyrus*, there is at least one voxel in the *Left Frontal Lobe* such that both satisfy the *RegionalPartOf* property. In other words, this is how FMA tells us that the left superior frontal gyrus, and hence its sub-regions, belong to the left frontal lobe without explicitly stating what those sub-regions are.

Consequently, if we want to be able to solve queries containing rules under the open-world assumption, we must be able to infer results from rules that have existentials in their heads. This shows us how the choice of Datalog+/– as the backbone on which NeuroLang is based, is a wise choice when solving queries under the open-world assumption due to its ability to deal with existentially quantified variables.

As we mentioned before, one of the most important approach to query answering over ontologies is via rewriting the input ontology (and query) into a new set of axioms that are expressed in logics for which scalable query answering algorithms exist [12]. Taking into account that most ontologies are currently written in the OWL2 tractable language, that OWL2 is equivalent to the DL-Lite families of Descriptions Logics, and that Cali et al. [19] prove that the Linear set of datalog rules extended with existential qualifiers is more expressive than the description logic DL-$Lite_R$, we decided to implement the XRewriter algorithm proposed by Gottlob et. al. [36]. XRewriter is designed as a practical rewriting algorithm for linear and sticky TGDs, sufficient syntactic conditions to guarantee first order rewritability of CQ answering. Also, XRewriter is based on backward-chaining resolution, this means that the algorithm uses the TGDs as rewriting rules, with the aim of simulating, independently from the extensional database, the chase derivations which are responsible for the generation of the image of the input query. This approach is much more efficient than a forward substitution procedure because of the fact that during the rewriting process, we only explore the part of the chase which is needed in order to entail the query.

A more detailed description of NeuroLang's architecture and how ontologies are integrated into its resolution engine is presented in the first contribution of this work, in the following chapter.

Part II

Main Contributions

7 - Scalable Query Answering under Uncertainty to Neuroscientific Ontological Knowledge

Abstract In this chapter, we present one of the main contributions of this dissertation: we will precisely define the architecture of NeuroLang and we will present a series of examples based on real-world use cases in neuroscientific research, specifically applied to solving meta-analysis questions. This work was a collaboration with Yamil Soto, Valentin Iovene, Maria Vanina Martinez, Ricardo O.Rodriguez, Gerardo I. Simari and Demian Wassermann. For more information, please refer to the original paper [80].

7.1 . Basic Probabilistic Ontological Model

In this section, we recall the basics on relational databases, conjunctive queries, Datalog, and ontology-mediated query answering (including tuple-generating dependencies and negative constraints), all based on a probabilistic extension with a corresponding query answering semantics.

We assume an infinite universe of *(data) constants* Δ, an infinite set of *(labeled) nulls* Δ_N (used as "fresh" Skolem terms) that are placeholders for unknown values, and an infinite set of variables $\mathcal{V}$. Different constants represent different values (*i.e., unique name assumption*), while different nulls may represent the same value. Sequences of $k \geq 0$ variables, namely $X_1, \ldots, X_k$, are denoted by $\mathbf{X}$.

Furthermore, we assume a *relational schema* $\mathcal{R}$, which is a finite set of *predicate symbols*, we also allow built-in predicates (with finite extensions) and equality. As expected, a *term* t is a constant, null, or variable. An *atomic formula* (or *atom*) $\mathbf{a}$ has the form $p(t_1, \ldots, t_n)$, where p is an n-ary predicate, and $t_1, \ldots, t_n$ are terms. We denote with $\mathcal{F}$ the set of all ground atoms built from $\mathcal{R}$ and Δ. A negated atom is of the form $\neg a$ where a is an atom. We assume that $\mathcal{R} = \mathcal{R}_D \cup \mathcal{R}_P$, with $\mathcal{R}_D \cap \mathcal{R}_P = \emptyset$, containing predicates that refer to deterministic and probabilistic events, respectively.

A *database instance* D for a relational schema $\mathcal{R}_D$ is a (possibly infinite) set of atoms with predicates from $\mathcal{R}_D$ and arguments from Δ. On the other hand, let a *probabilistic atom* be of the form $\mathbf{a} : p$, where p is a real number in the interval $[0, 1]$ and $\mathbf{a}$ is an atom with a predicate from $\mathcal{R}_P$. We do not allow negation in probabilistic atoms.

A *probabilistic constraint* c has the form

$$\mathbf{a}_1 : p_1 \mid \ldots \mid \mathbf{a}_k : p_k,$$

where $k > 0$, each $\mathbf{a}_i : p_i$ is a probabilistic atom, and $\sum p_i \leq 1$. If the p_i's in a

probabilistic constraint do not sum to 1, then there exists also the possibility that none of them happen. The probability of this complementary event is $1 - \sum p_i$. Given a probabilistic constraint $c = \mathbf{a}_1 : p_1 \mid \ldots \mid \mathbf{a}_k : p_k$, we will make use of the notation *atoms*$(c) = \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$. We will also denote the probability of any atom $\mathbf{a}$ with $p(\mathbf{a})$. We have that $p(\mathbf{a}_i) = p_i$ whenever $\mathbf{a}_i : p_i$ belongs to a probabilistic constraint c.

Given a set of probabilistic constraints C, note that each ground atom can only appear in one constraint in C. From a practical point of view, this assumption restricts the number of possible worlds by limiting the potential combinations. Vennekens et. al.[73, Eq. 5] propose more complex semantics where this assumption is relaxed. This approach is similar to *probabilistic databases* [69] where each tuple comes from a general probability distribution over tuples and inexistence is one of the options. This allows to incorporate beliefs about the likelihood of tuples and cell values.

Example 1 *Consider the following database instance D and a set of probabilistic constraints C (recall that t_i atoms cannot appear in C).*

$$D = \{t_1(a), t_1(c), t_2(a), t_2(b)\}$$

$$C = \left\{ \begin{array}{lll} c_1 = s(a,b) & : 0.3 & \\ c_2 = s(b,c) & : 0.7 & \\ c_3 = r(b) & : 0.4 \quad \mid \quad r(c) & : 0.1 \end{array} \right\} \tag{7.1}$$

Tuple Generating Dependencies Given a relational schema $\mathcal{R}$, a *tuple-generating dependency (TGD)* σ is a first-order formula of the form:

$$\forall \mathbf{X} \forall \mathbf{Y}\, \Phi(\mathbf{X}, \mathbf{Y}) \rightarrow \exists \mathbf{Z}\, \Psi(\mathbf{X}, \mathbf{Z}),$$

where $\Phi(\mathbf{X}, \mathbf{Y})$ and $\Psi(\mathbf{X}, \mathbf{Z})$ are conjunctions of atoms over $\mathcal{R}$ (without nulls), called the *body* and the *head* of σ, denoted $body(\sigma)$ and $head(\sigma)$, respectively. Such σ is satisfied in a database D for $\mathcal{R}$ if and only if, whenever there exists a homomorphism h that maps the atoms of $\Phi(\mathbf{X}, \mathbf{Y})$ to atoms of D, there exists an extension h' of h that maps the atoms of $\Psi(\mathbf{X}, \mathbf{Z})$ to atoms of D. All sets of TGDs are finite here and we assume that every TGD has a single atom in its head. Furthermore, we say that a TGD σ is *full* whenever there are no existential variables in the head. Let's extend our example further:

Example 2 *Based on Example 1 we can add the following set of rules:*

$$\begin{aligned} \Sigma = \{&\forall X\ t_1(X) \rightarrow \exists Z\ o(X, Z), \\ &\forall X \forall Y\ t_2(X) \wedge o(X, Y) \rightarrow t(X), \\ &\forall X \forall Y\ s(X, Y) \wedge r(Y) \rightarrow w(X, Y)\} \end{aligned}$$

$$A = \{\forall X \forall W\ v(X, W) \rightarrow u(X, \min(W))\}$$

TGDs can be extended to allow negation—in this work we allow semi-positive Datalog [2] in the case that a rewriting is needed and stratified negation [2] otherwise. Furthermore, as shown by the rule in set A in the previous example, we extend the language so aggregation functions can be used in the head of full TGDs [2]. As we see in the following section, we restrict the syntax of this type of rules so that neither negation nor recursion is allowed.

Definition 1 *A* probabilistic ontology $\mathcal{O} = (D, C, \Sigma)$ *consists of a database instance* D*, a set* C *of probabilistic constraints, and a set* Σ *of arbitrary TGDs.*

Note that a database instance can be thought of as a set of probabilistic constraints with only probabilistic atoms, each one annotated with probability 1. Furthermore, the structure (D, Σ) corresponds to a knowledge base with existential rules as defined in [19], whenever rules in Σ do not involve atoms that appear in probabilistic constraints.

Semantics We take the notion of possible world (or interpretation) of a probabilistic ontology as a subset of $\mathcal{F}$ and we denote with Ω the set of all possible worlds. Each possible world $\omega \in \Omega$ satisfies the following property:

$$\forall F \in \mathcal{F} : \omega \models F \text{ iff } F \in \omega; \quad \text{otherwise } \omega \models \neg F$$

This means that ω is a complete interpretation of every element of $\mathcal{F}$. The usual semantics of a classical Datalog program P is the least Herbrand model that contains exactly all ground facts in P plus every ground atom inferred from it, i.e. the intersection of all worlds that satisfy P.

However, in the probabilistic case, we need to consider a generalization of this semantics so that every ground fact is associated with a probability value. According to this idea, we are going to take the models of a set of non-probabilistic ontologies, induced by total choices, so that they all share the same TGDs but the corresponding database instances differ. As mentioned before, in our approach, we have two ways of associating probability with facts. In the first one, a fact corresponds to a Boolean random variable that is true with probability p and false with probability $1 - p$. In the second, we interpret facts as multi-valued random variables instead of binary ones. We use probabilistic constraints to represent both and assume that the facts within the same constraint are mutually exclusive events, whereas facts in different constraints are mutually independent events. According to this idea, we give the following definition:

Definition 2 *Given a probabilistic ontology* $\mathcal{O} = (D, C, \Sigma)$*, for each* $1 \leq j \leq |C|$: $c^j = \mathbf{a}_1^j : p_1^j \mid \ldots \mid \mathbf{a}_k^j : p_k^j$*, with* $c^j \in C$*, we have:*

$$\textit{choices}(c^j) = \{\mathbf{a}_i^j \mid 1 \leq i \leq k\} \cup \{\bot_{c^j}\}.$$

For each $b = \mathbf{a}_i^j \in c^j$, *we have* $p(b) = p_i^j$ *and* $p(\bot_{c^j}) = 1 - \sum_{1 \leq i \leq k} p_i^j$. *The set of total choices for* $\mathcal{O}$ *is defined as* $\mathit{total\_choices}(C) =$

$$\{[b_1, \ldots, b_l] \mid l = |C|, 1 \leq j \leq |C| : b_j \in \mathit{choices}(c^j)\}$$

The probability of a particular total choice $\lambda \in \mathit{total\_choices}(C)$ *is defined as* $p(\lambda) = \prod_{1 \leq j \leq l}^{[b_1,\ldots,b_l] \in \lambda} p(b_j)$. *We use notation* $\mathit{atoms}(\lambda) = \{b_j \neq \bot_{c^j} \mid 1 \leq j \leq l : [b_1, ..., b_l] \in \lambda\}$ *and* $\mathit{atoms}(C) = \bigcup_{\lambda \in \mathit{total\_choices}(C)} \mathit{atoms}(\lambda)$.

Definition 3 *Let* ω *and* λ *be a possible world and a total choice, respectively. Then, we will say that* ω *satisfies* λ, *denoted* $\omega \models \lambda$, *if and only if* $\mathit{atoms}(\lambda) \subseteq \omega$. *Also,* $\|\lambda\|$ *will denote the set of possible worlds of a total choice, i.e.* $\|\lambda\| = \{\omega \in \Omega \mid \omega \models \lambda\}$.

Example 3 *The set of all total choices for probabilistic ontology* (D, C, Σ) *from Examples 1 and 2 is the following:*

$$\begin{array}{llllll}
\lambda_1 & = [s(a,b), & s(b,c), & r(b)] & p(\lambda_1) & = 0.084 \\
\lambda_2 & = [s(a,b), & \bot_{c_2}, & r(b)] & p(\lambda_2) & = 0.036 \\
\lambda_3 & = [\bot_{c_1}, & s(b,c), & r(b)] & p(\lambda_3) & = 0.196 \\
\lambda_4 & = [\bot_{c_1}, & \bot_{c_2}, & r(b)] & p(\lambda_4) & = 0.084 \\
\lambda_5 & = [s(a,b), & s(b,c), & r(c)] & p(\lambda_5) & = 0.021 \\
\lambda_6 & = [s(a,b), & \bot_{c_2}, & r(c)] & p(\lambda_6) & = 0.009 \\
\lambda_7 & = [\bot_{c_1}, & s(b,c), & r(c)] & p(\lambda_7) & = 0.049 \\
\lambda_8 & = [\bot_{c_1}, & \bot_{c_2}, & r(c)] & p(\lambda_8) & = 0.021 \\
\lambda_9 & = [s(a,b), & s(b,c), & \bot_{c_3}] & p(\lambda_9) & = 0.105 \\
\lambda_{10} & = [s(a,b), & \bot_{c_2}, & \bot_{c_3}] & p(\lambda_{10}) & = 0.045 \\
\lambda_{11} & = [\bot_{c_1}, & s(b,c), & \bot_{c_3}] & p(\lambda_{11}) & = 0.245 \\
\lambda_{12} & = [\bot_{c_1}, & \bot_{c_2}, & \bot_{c_3}] & p(\lambda_{12}) & = 0.105
\end{array}$$

It is easy to see that $\mathit{total\_choices}(C)$ defines a partition on Ω by using the following equivalence relation on $\Omega \times \Omega$: $\omega \equiv \omega'$ if and only if $\forall \lambda \in \mathit{total\_choices}(C) : \omega \models \lambda \Leftrightarrow \omega' \models \lambda$.

We define the semantics of a probabilistic ontology based on the semantics of a classical ontology with existential rules (TGDs). Intuitively, each total choice induces a classical (i.e., non-probabilistic) ontology.

Definition 4 *Let* $\mathcal{O} = (D, C, \Sigma)$, *be a probabilistic ontology, and let* λ *be a total choice of* C. *Then, the (non-probabilistic) ontology induced by* $\lambda = [b_1, \ldots b_l]$ *is defined as* $O_\lambda = (D_\lambda, \Sigma)$, *with* $D_\lambda = D \cup \{b_1, \ldots b_l\}$.

Example 4 *Based on the total choices from Example 3 and probabilistic ontology* $\mathcal{O} = (D, C, \Sigma,)$, *each* λ_i *with* $1 \leq i \leq 12$, *induces a non-probabilistic ontology* $\mathcal{O}_{\lambda_)} = (D_{\lambda_i}, \Sigma)$ *where* $D_{\lambda_i} = \mathcal{D} \cup \{b_1, \ldots, b_l\}$ *with* $b_k \in \lambda_i$ *and* $b_k \neq \bot_{c_j}$ *for every* $c_j \in C$.

We now recall the notions of satisfaction and entailment in classical ontologies from Cali et. al. [19]. We first introduce the definition of answers to conjunctive queries for an instance database, on which the others are based.

A *conjunctive query (CQ)* over a relational schema $\mathcal{R}$ has the form $Q(\mathbf{X}) = \exists\Phi(\mathbf{X}, \mathbf{Y})$, where $\Phi(\mathbf{X}, \mathbf{Y})$ is a conjunction of atoms (possibly equalities, but not inequalities) with the variables $\mathbf{X}$ and $\mathbf{Y}$, and possibly constants, but without nulls. A Boolean CQ (BCQ) over $\mathcal{R}$ is a CQ of the form $Q()$.

Answers to CQs and BCQs are defined via *homomorphisms*, which we recall are mappings $\mu\colon \Delta \cup \Delta_N \cup \mathcal{V} \to \Delta \cup \Delta_N \cup \mathcal{V}$ such that (i) $c \in \Delta$ implies $\mu(c) = c$, (ii) $c \in \Delta_N$ implies $\mu(c) \in \Delta \cup \Delta_N$, and (iii) μ is naturally extended to atoms, sets of atoms, and conjunctions of atoms.

Definition 5 *Given a database instance D and a CQ $Q(\mathbf{X}) = \exists\mathbf{Y}\, \Phi(\mathbf{X}, \mathbf{Y})$, the set of answers for $Q(\mathbf{X})$ in D, denoted $Q(D)$, is the set of all tuples t over Δ such that there exists a homomorphism $\mu\colon \mathbf{X} \cup \mathbf{Y} \to \Delta \cup \Delta_N$ with $\mu(X) = t$ and $\mu(\Phi(\mathbf{X}, \mathbf{Y})) \subseteq D$. The answer to a BCQ Q is* Yes, *denoted $D \models Q$ iff $Q(D) \neq \emptyset$.*

As we can have several completions of D that satisfy the TGDs in Σ we need to consider the set of all possible models of an ontology (D, Σ).

Definition 6 *Given an ontology (D, Σ), the set of* models, *denoted $mods(D, \Sigma)$, is the set of all (possibly infinite) databases B such that (i) $D \subset B$, and (ii) every $\sigma \in \Sigma$ is satisfied in B.*

Note that each B in the above definition can be considered as a possible world under the closed world assumption, i.e. every tuple that does not appear in B is false. It is important to recall that for full TGDs (pure Datalog rules), an ontology (D, Σ) has a unique least model [2].

The definitions for query answering and entailment for deterministic ontologies are as follows:

Definition 7 *Given an ontology (D, Σ) and a CQ $Q(\mathbf{X}) = \exists\mathbf{Y}\, \Phi(\mathbf{X}, \mathbf{Y})$, the set of answers for Q in (D, Σ), denoted $ans(Q, D, \Sigma)$, is the set of all tuples t over Δ such that $t \in Q(B)$ for every $B \in mods(D, \Sigma)$. Furthermore, the answer to a BCQ Q over D given Σ exists, denoted $(D, \Sigma) \models Q$, if and only if $ans(Q, D, \Sigma) \neq \emptyset$.*

With these definitions in place, we can now turn to probabilistic ontologies. The probability of a conjunction of ground atoms is defined as follows.

Definition 8 *Let $\mathcal{O}$ be a probabilistic ontology, and Φ be a conjunction of ground atoms built from predicates in $\mathcal{R}$. The probability that Φ holds in $\mathcal{O}$, denoted $Pr^{\mathcal{O}}(\Phi)$, is the sum of the probabilities of all total choices λ such that $(D_\lambda, \Sigma) \models \Phi$; that is, $Pr^{\mathcal{O}}(\Phi) = \sum_{(D_\lambda, \Sigma) \models \Phi}^{\lambda \in total\_choice(C)} p(\lambda)$.*

At this point, it is interesting to remark the connection between our approach and the one considered by Riguzzi et. al. [60]. The Logic Programs with Annotated Disjunctions (LPADs) mentioned in their paper make an implicit treatment of mutually exclusive facts, whereas our approach does it explicitly. In fact, LPADs are more expressive than our language since they use non-Horn clauses. In addition, they use well-founded semantics in order to deal with negation as failure. Both aspects have a computational cost that we wish to avoid.

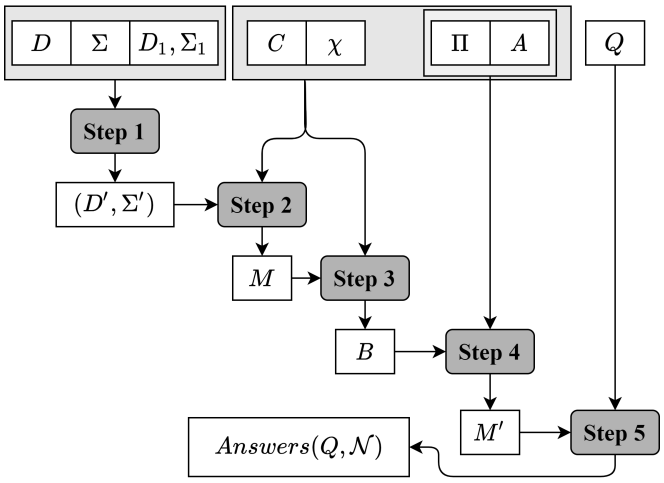

Figure 7.1: Overview of the NeuroLangQA algorithm. Step numbers refer to those described in Algorithm 1

Semantics for Probabilistic Query Answering Probabilistic answers to CQs are defined as pairs of tuple and probability such that the probability adds all the probabilities of the choices for which the tuple is a classical answer to the query in the deterministic ontology that the choice induces. Formally, this is:

Definition 9 *The set of all* probabilistic answers *to a CQ* $Q(\mathbf{X})$ *over a probabilistic ontology* $\mathcal{O} = (D, C, \Sigma)$*, denoted with* $ans(Q, D, C, \Sigma)$*, or* $ans(Q, \mathcal{O})$*, is a set of pairs* (t, p_t) *where* t *is a tuple over* Δ *and* $p_t = \sum_{t \in ans(Q, D_\lambda, \Sigma)}^{\lambda \in total\_choice(C)} p(\lambda)$.

Observations If a probabilistic ontology $\mathcal{O} = (D, C, \Sigma)$ is such that C is empty, then the semantics for (B)CQs as defined above coincides with that for classical ontologies [19].

Note that query answering under general TGDs for non-probabilistic ontologies is undecidable [11], even when the schema and TGDs are fixed [18]. The two problems of CQ and BCQ evaluation under TGDs are LOGSPACE-equivalent [31, 27]. As mentioned above, in the non-probabilistic case, for arbitrary full TGDs there exists exactly one minimal model [2] over which Q is evaluated. Furthermore, it

has been shown that for full TGDs CQ evaluation can be done in polynomial time in data complexity (*i.e., assuming σ and Q fixed*) [24].

7.2 . NeuroLang Programs

In addition to our model, we assume the existence of a separate schema $\mathcal{T}$, the target schema, that defines the language by means of which users of NeuroLang can query about the probability of certain events. Predicates in $\mathcal{T}$ have a distinguished term in the n-th position (for n-ary predicates) reserved exclusively for real numbers in the interval $[0, 1]$; i.e., for any predicate $p \in \mathcal{T}$, atoms of the form $p(a_1, \ldots, a_n)$ are such that $a_1, \ldots, a_{n-1}$ are variables or constants from Δ, while a_n is a variable or a constant from $[0, 1]$. Below we show an example of how this language is used.

A NeuroLang program $\mathcal{N}$ is comprised of the following components:

- D, Σ: where D is a set of ground atoms from $\mathcal{R}_D$, and Σ is a set of full TGDs that only use atoms from $\mathcal{R}_D$ and can have recursion and stratified negation. In the scenario where rewriting must be applied, Sigma has to be restricted to the non-recursive and semi-positive case to ensure the correctness of the rewriting using XRewriter.
- (D_1, Σ_1): a classical ontology, where D_1 is a set of ground atoms from $\mathcal{R}_D$, Σ_1 is a set of TGDs that belong to the Sticky fragment [15], and the bodies and heads are atoms built from predicates in $\mathcal{R}_D$.
- C: a set of probabilistic constraints only involving atoms from $\mathcal{R}_P$.
- χ: a set of full TGDs, whose bodies and heads may contain atoms from $\mathcal{R}_D \cup \mathcal{R}_P$. Neither negation nor recursion is allowed in this set of rules.
- Π: a set of *probability encoding rules* (PERs) with the following form:

  $$\sigma^* : \forall \mathbf{X} \forall \mathbf{Y} (\Phi(\mathbf{X}, \mathbf{Y})) \rightarrow \psi(\mathbf{X}, \rho_X)$$

  where Φ is a conjunction of atoms from $\mathcal{R}_D \cup \mathcal{R}_P$, ψ is an atom in $\mathcal{T}$ and ρ_X is the distinguished term that in this case must be a variable (ranging over the reals in $[0, 1]$).
- A: a set of rules of the form

  $$\forall \mathbf{X} \forall \mathbf{Y} (\Phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \rightarrow \psi(\mathbf{X}, \mathit{agg}(\mathbf{Z})) \tag{7.2}$$

  where Φ is a conjunction of atoms in $\mathcal{R}_D \cup \mathcal{T}$, ψ is an atom in $\mathcal{T}$ and *agg* is an aggregation function (e.g., sum, count, avg, etc.). Neither negation nor recursion is allowed in these rules.

Informally, the above sets together provide the following functionalities:

(i) Σ, Σ_1, $\mathcal{C}$, and χ are used by the probabilistic inference mechanism, which applies ontological rules and ultimately associates probabilities to atoms (following the semantics described in Section 2);

(ii) Π incorporates probabilities as values inside atoms; and

(iii) rules in A manipulate these probabilities via aggregation functions to present them as requested by the user.

Algorithm 1: NeuroLangQA

Input : NeuroLang program $\mathcal{N} = (D, \Sigma, (C, \chi), (D_1, \Sigma_1), \Pi, A)$ and query $Q(\mathbf{X}) = \exists \mathbf{Y} \Phi(\mathbf{X}, \mathbf{Y})$
Output: $ans(Q(\mathbf{X}), \mathcal{N})$
Step 1: Obtain database instance D' and set of full TGDs Σ' such that $D^{'} = D \cup D_1$ and $\Sigma^{'}$ is the rewriting of Σ with respect to Σ_1.
Step 2:
2a: Let *Aux* be the set of TGDs in $\Sigma^{'}$ whose bodies do not depend on $C \cup \chi \cup \Pi$.
2b: Let M the set of ground atoms a such that $(D', Aux \cup A) \models a$
Step 3:
B:= $\emptyset$
foreach *PER* $\pi \in \Pi$ **do**
    `// Rule bodies are taken as queries`
    Let $Q_\pi(\mathbf{X}) = body(\pi)$
    `// Obtain probability values`
    `// associated with each query answer`
    probAnsPairs:= $ans(Q_\pi(\mathbf{X}), (M, C, \chi))$
    foreach $(t, p) \in$ *probAnsPairs* **do**
        `// Add query answers and PER`
        `// heads to set` B
        Let h' be the instantiation of $head(\pi)$ with values from (t, p)
        B:= $B \cup \{h', Q_\pi(t)\}$
    end
end
Step 4: Let M' the set of ground atoms a such that $(B, (\Sigma' - Aux) \cup A) \models a$
Step 5: Return $ans(Q(\mathbf{X}), \mathcal{N})$ computed from atoms in set M'.

Note that PERs are full TGDs that will be used to translate from a source schema to a target one, in the same spirit as source-to-target TGDs for data exchange [31]. Effectively, they reify the probability of an atom, given by the semantics, as a term in a new atom that can be further manipulated by other rules. For instance, a set of probabilistic constraints $C = \{s(a, b) : 0.3\}$ will be reified by the PER $\forall X \forall Y s(X, Y) \rightarrow t(X, Y, \rho_X)$ as $\{t(a, b, 0.3)\}$. On the other hand, for rules in A we incorporate functional symbols *agg* to the distinguished

term in ψ to indicate that its value takes the result of applying the function *agg* to all ρ_X that satisfy the body of the rule. Note that users here can define arbitrary rules that manipulate probabilities by means of aggregation functions. It's defined as a post-processing step that builds a view as defined by the user issuing the query. Therefore, it's the user's responsibility that the handling of the probabilities obtained in the previous steps complies with the laws of probability. We extend notation *body* and *head* used for TGDs to all types of rules defined in this section. The following is a simple example of query answering using PERs.

Example 5 *Consider the following NeuroLang program $\mathcal{N}$. We add a set of PERs and rules with aggregations.*

$$
\begin{aligned}
D_1 &= \{t_1(a), t_1(c)\},\\
\Sigma_1 &= \{\forall X t_1(X) \rightarrow \exists Z\ o(X, Z)\},\\
D &= \{t_2(a), t_2(b)\},\\
\Sigma &= \{\forall X \forall Y\ t_2(X) \wedge o(X, Y) \rightarrow t(X)\},\\
C &= \left\{ \begin{array}{llllll} s(a,b) & : 0.3 & & & & \\ s(b,c) & : 0.7 & & & & \\ r(b) & : 0.4 & | & r(c) & : 0.1 & \end{array} \right\},\\
\chi &= \{\forall X \forall Y\ s(X, Y) \wedge r(Y) \rightarrow w(X, Y)\}\},\\
\Pi &= \{\forall X \forall Y\ w(X, Y) \rightarrow v(X, \rho_X)\},\\
A &= \{\forall X \forall W\ v(X, W) \rightarrow u(\min(W))\},\\
Q_1(X, P) &= v(X, P), t(X),\\
Q_2(X, P) &= v(X, P), u(P).
\end{aligned}
$$

Now, the partition of possible worlds used to compute queries Q_1 and Q_2 is the following (excluding atoms from D and (D_1, Σ_1) for clarity, and including probabilities):

$$
\left\{ \begin{array}{llllll}
\{s(a,b) & s(b,c) & w(a,b) & r(b) & t(a)\} & : 0.084 \\
\{s(a,b) & & w(a,b) & r(b) & t(a)\} & : 0.036 \\
\{ & s(b,c) & & r(b) & t(a)\} & : 0.196 \\
\{ & & & r(b) & t(a)\} & : 0.084 \\
\{s(a,b) & s(b,c) & w(b,c) & r(c) & t(a)\} & : 0.021 \\
\{s(a,b) & & & r(c) & t(a)\} & : 0.009 \\
\{ & s(b,c) & w(b,c) & r(c) & t(a)\} & : 0.049 \\
\{ & & & r(c) & t(a)\} & : 0.021 \\
\{s(a,b) & s(b,c) & & & t(a)\} & : 0.105 \\
\{s(a,b) & & & & t(a)\} & : 0.045 \\
\{ & s(b,c) & & & t(a)\} & : 0.245 \\
\{ & & & & t(a)\} & : 0.105
\end{array} \right\}
$$

Answering Q_1, Q_2 leads to the target schema solution $\{v(a, 0.12),\ v(b, 0.07),\ u(0.07)\}$. *Hence, the resulting answer set is* $\{Q_1(a, 0.12), Q_2(b, 0.07)\}$.

Query Answering in NeuroLang A *NeuroLang query* Q is any conjunction of atoms in $\mathcal{R}_D \cup \mathcal{T}$, such that atoms in $\mathcal{T}$ have as distinguished term a variable; these variables will be instantiated with the probability of certain events as computed by the inference mechanism. Algorithm 1 describes the pseudocode for answering queries in the NeuroLang framework— Figure 7.1 provides a high-level view of the main steps involved in this process, where inputs are as defined above.

There are two steps in which NeuroLangQA makes external calls. First, in Step 1 the rewriting of Σ w.r.t. Σ_1 is done by means of the XRewrite algorithm developed in [36] for rewriting queries with respect to the Sticky fragment of existential rules (also known as Datalog+/–). Note that here the algorithm is used to rewrite every appearance of heads of rules in Σ_1 in the bodies of rules in Σ, yielding a potentially larger set of full TGDs (rules without existentials in the head).

Then, Step 3 derives the probabilities associated with atoms. This is done by dynamically choosing the best algorithm for the job: if π is liftable according to [23], then lifted query answering is applied and, based solely on syntatic analysis of queries, a set of rules that derive an algebraic expression to computes the probability of the query is derived; otherwise, the query is said to be *non-liftable*, and has been proven to have a #P-hard complexity, then the query is compiled to an SDD representation and model counting is applied [75]. Both cases are implemented in relational algebra with provenance [68].

Provenance (or data provenance) is the idea of associating a column of extra information to our data to answer meta-questions related to the output of our data. For example, what operations gave rise to our results, or where did the final data come from. In particular, we use Semiring provenance[37] that has been shown to generalize previous formalisms using a clean mathematical framework.

Given a fixed semiring ($\mathbb{K}$, $\mathbb{O}$, $\mathbb{1}$, $\bigoplus$, $\bigotimes$) where $\mathbb{O}$ and $\mathbb{1}$ are two distinguished elements along with two binary operations: $\bigoplus$, an associative and commutative operator with identity $\mathbb{O}$ and $\bigotimes$, an associative and commutative operator with identity $\mathbb{1}$, and based on Senellart[68] and defining *prov[A]* as the provenance column in the Relational Algebra set extended with Provenance (RAP) A, and *non_prov[A]* as the set of the columns in the RAP set A without the provenance column, we can define for the operations selection, projection, union and cross product, how each operation affects each of these columns: I) Selection and Renaming do not affect provenance annotations II) in the bag semantics, Projection does not affect provenance annotations, but duplicate elimination $\bigoplus$-es the annotation of merged tuples. Therefore, in the case of duplicate rows, and assuming a

Γ function that aggregate duplicate tuples and do nothing with single tuples, we have:

$$\frac{\pi_{cols}(RAP[A])}{RAP_{\pi_{cols}(\text{non\_prov[A]} \bigcup \{prov=\Gamma(prov[A],\bigoplus)\}}[A]} \tag{7.3}$$

III) In the union of tuples, the provenance annotations are $\bigoplus$-ed, giving us:

$$\frac{RAP[A] \bigcup RAP[B]}{RAP_{\pi[\text{non\_prov[A]} \bigcup \text{non\_prov[B]} \bigcup \{\text{prov=prov[A]} \bigoplus \text{prov[B]}\}}[A \bigcup B]} \tag{7.4}$$

IV) Finally, in a cross product, provenance annotations of tuples combined, are $\bigotimes$-ed, so we have

$$\frac{RAP[A] \bowtie RAP[B]}{RAP_{\pi[\text{non\_prov[A]} \bigcap \text{non\_prov[B]} \bigcup \{\text{prov=prov[A]} \bigotimes \text{prov[B]}\}}[A \bowtie B]} \tag{7.5}$$

Note also that up to Step 3 we can guarantee the correctness of the semantics of NeuroLangQA, i.e., the probabilities associated with atoms in set B correspond to the probability with which they are entailed in the probabilistic ontology. However, since after this step users can manipulate the probabilities of atoms through aggregation functions provided in A, it cannot be guaranteed that this relationship holds in the next steps, so users have the responsibility of making a sound use of such values. This manipulation is intentionally incorporated to increase the expressive power of the languages; similar additions occur in other languages, like Prolog. This feature is useful in our application case allowing, for instance, to aggregate probabilistic values into voxel overlays (cf. Section 7.3.1), or select the 95th percentile top probabilities of a result set (cf. Section 7.3.2).

The final step of the algorithm returns the answers to query Q as the set of all tuples t built from Δ such that there exists a homomorphism μ where $\mu(\mathbf{X}) = t$ and $\mu(\Phi(\mathbf{X}, \mathbf{Y})) \in M'$.

Correctness of NeuroLangQA We now discuss the correctness of NeuroLangQA algorithm with respect to the probabilistic semantics described in Section 7.1. Without loss of generality, we assume a query of the form

$$Q(\mathbf{X}, \rho_{\mathbf{X}}) = \Phi(\mathbf{X}) \wedge \psi_i(\mathbf{X}, \rho_{\mathbf{X}}),$$

where $\Phi(\mathbf{X})$ is a conjunction of atoms in $\mathcal{R}_D$ and $\psi_i(\mathbf{X}, \rho_{\mathbf{X}})$ is an atom in $\mathcal{T}$.

The result of Step 1 in NeuroLangQA is a special case of a probabilistic ontology (D', Σ'), where Σ' is a set of full TGDs that may contain semi-positive Datalog negation and no-recursion. In the scenario in which step one is not applied because Σ_1 is empty, Σ' is a set of full TGDs that may contain stratified negation and recursion. Furthermore, Step 2a removes from Σ' all rules that depend on $C \cup \chi \cup \Phi$ [9]. Therefore, M computed in Step 2b is unique as neither probabilistic atoms, nor existential rules are involved. Step 3 now considers the probabilistic ontology defined by $\mathcal{O} = (M, C \cup C', \chi)$. Note that atoms in M materialize ontology (D', Aux) and they will hold in every possible world for probabilistic ontology $\mathcal{O}$.

Recall that the purpose of PERs is to incorporate the probability of an atom as an additional term—Step 3 does precisely that: for each PER π, it computes the probability of all ground instantiations of *body*(π) that are entailed by $\mathcal{O}$. For each such instantiation t, set B contains the instantiation itself ($Q_\pi(t)$) and the head of π instantiated by values in t and an extra position with value $Pr^{\mathcal{O}}$(*body*(π)(t)).

Finally, Step 4 considers a deterministic ontology comprised by B (a set of ground atoms) and the set of full TGDs $(\Sigma' - Aux) \cup A$; M' contains all ground atoms that are entailed by such ontology. As in the case of M, M' is unique since neither existential rules nor probabilistic atoms are involved.

Therefore, we can conclude that—by construction—the results computed by the NeuroLangQA algorithm are correct with respect to the probabilistic semantics defined in Section 7.1 up to Step 3. This means that the probabilities associated with atoms in B correspond to the probability with which they are entailed by the probabilistic ontology. The final two steps simply follow the user-specified rules for establishing personalized views, which may manipulate probability values in an arbitrary fashion. With the framework in place, in the following we show how it can be applied in practice.

7.3 . Examples based on Real-World Use Cases in Neuroscience Research

In this section, we illustrate via concrete examples several use cases that appear in real-world tasks carried out by neuroscience researchers. Since all of our analyses are based on meta-analytic components, we first give a brief description of the Neurosynth database we use in our examples. Where extra data is used, it will be clarified in each particular case. The Neurosynth database is composed of 3.228×10^3 terms, 1.4370×10^4 studies (*SelectedStudy*), and 3.3593×10^4 voxels; but this information would not be useful without associations, so we also have $1.049\,299 \times 10^6$ terms reported as present in studies (*TermInStudy*) and $5.078\,91 \times 10^5$ voxels reported as active (*FocusReported*), also with their respective study. Finally, there are 112 brain regions from Destrieux's atlas [26] associated with brain coordinates through the *VoxelByRegionDestrieux* relation. These data give rise to the following extensional databases:

$$D = \left\{ \begin{array}{l} \textit{TermInStudy}(\text{“emotion"}, s_1), \\ \vdots \\ \textit{TermInStudy}(\text{“pain"}, s_{120}), \\ \textit{FocusReported}(5, -5, 3, s_1), \\ \vdots \\ \textit{FocusReported}(-10, 5, 1, s_{25}), \\ \textit{VoxelByRegionDestrieux}(15, 47, 16, \\ \quad \textit{'l\_g\_and\_s\_frontomargin'}), \\ \vdots \\ \textit{VoxelByRegionDestrieux}(16, 46, 15, \\ \quad \textit{'l\_g\_and\_s\_frontomargin'}), \end{array} \right\}$$

$$C = \left\{ \begin{array}{l} \textit{SelectedStudy}(s_i) : \frac{1}{\#studies} \\ \textit{FocusCoactivates}(\ 5, -5,\ 3,\quad 5, -5,\ 3) : 1 \\ \vdots \\ \textit{FocusCoactivates}(\ 5, -5,\ 3,\quad -10,\ \ 5,\ 1) : \\ \qquad (2\pi 2)^{-3/2} \exp\left(-\frac{1}{2} \frac{\|(5,-5,3)-(-10,5,1)\|^2}{2^2}\right) \end{array} \right\}$$

where *FocusCoactivates* represents spatial uncertainty in foci reporting, as they encode that the probability that two foci co-activate is mediated by their distance as measured by a 3D Gaussian law with standard deviation 2mm. This dataset has approximately 5 million atoms. Furthermore, the CogAt ontology [58] is composed of 5.6807×10^4 rules. In the following, examples are written in extended Datalog syntax, as in our implemented tool[1]. We base our examples on versions 1.4.0 of IOBC, 0.3.1 of CogAt, and the Destrieux 2009 atlas [25] provided by Nilearn software package v0.7.0 [32]. In addition, both the software code and other examples can be found on the official NeuroLang repository[1]. While the examples presented in this section are simple demonstrations of some of the features of Neurolang in sections 9 and 10, we present use cases on the use of NeuroLang for real-world neuroscience research.

7.3.1 . Forward inference

In this task, we wish to assess the probability of a voxel being reported as active in a study given that the word “emotion” is present in the specific study. The corresponding Neurolang program can be seen in listing 7.1.

Note that in order to represent this knowledge we only need the expressive power of full TGDs (no existential rules are needed). In fig. 7.2 we see that the most important reported activations are concentrated in the amygdala, the region most related to emotions, as generally accepted in the neuroscience field.

[1] https://neurolang.github.io/

Listing 7.1: Forward inference

```
TermAssociation(t) :- SelectedStudy(s),
    TermInStudy(t, s).

Activation(i, j, k) :- SelectedStudy(s),
    FocusReported(i1, j1, k1, s),
    FocusCoactivates(i, j, k, i1, j1, k1).

% Probability Encoding Rule: PROB is
% used to encode probability as defined in
% Section 3. The // operator is
% syntactic sugar for conditional
% probability as P(A|B) = P(A,B) / P(B).
ProbMap(i, j, k, PROB) :-
    Activation(i, j, k)
    // TermAssociation("emotion").

% Aggregation to build a single image with
% the probability p in each position
% within the top 95% of probability
Percentile_95(compute_percentile(p, 95)) :-
    ProbMap(i, j, k, p).

ProbabilityImage(create_region_overlay(i, j, k, p)) :-
    ProbMap(i, j, k, p),
    Percentile_95(p95), p > p95

Ans(x) :- ProbabilityImage(x)
```

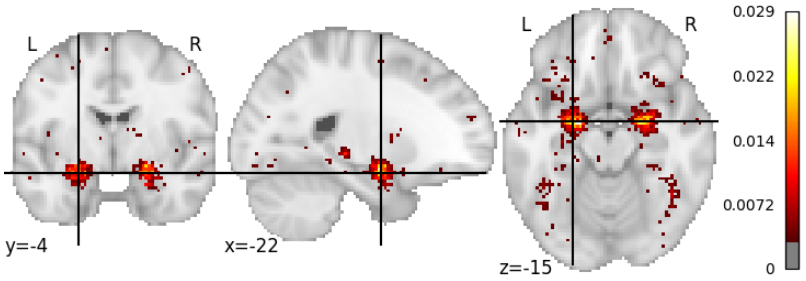

Figure 7.2: Resulting thresholded brain image from the NeuroLang use case showing that foci in the amygdala are most probably reported if a study includes the word "emotion". As expected, the main area shown corresponds to the amygdala [51].

7.3.2 . Segregation reverse inference query

This example shows how we can use negation and existentials to express specificity. We pick the terms present in the CogAt ontology that are mentioned in studies reporting activations within the short insural gyri. Listing 7.2 presents the corresponding Neurolang program associated with this example.

Listing 7.2: Segregation reverse inference query. See the description in section 7.3.2.

```
OntologyTerms(t) :- hasTopConcept(u, c), label(u, t)
FilteredTerms(s, t) :- TermInStudy(s, t), OntologyTerms(t)
RegionActivated(s, r) :- VoxelByRegionDestrieux(i, j, k, r),
    FocusReported(i, j, k, s).
SegregatedStudies(s) :- RegionActivated(s, r),
    (DestrieuxLabels(r, 'l_g_insular_short') |
    DestrieuxLabels(r, 'l_g_insular_short')),
    ~exists(r2, RegionActivated(s, r2), r != r2)
TermProbability(t, PROB) :- FilteredTerm(s, t)
    // (SegregatedStudies(s)
      SelectedStudy(s)).
Percentile_95(compute_percentile(p, 95)) :- TermProbability(t, p).
Ans(t, p) :- TermProbability(t, p), Percentile_95(p95), p > p95.
```

Processing took 42.45 seconds. Results are shown in table 7.1.

term	prob
anxiety	0.097819

Table 7.1: Terms, within the 95th percentile, mentioned in our segregation query in section 7.3.2. Shows that studies presenting activations only related to the short insula gyrus tend to be associated with anxiety.

7.3.3 . Variance in primary neuroimaging data

In this example, we demonstrate how it's possible, by implementing techniques developed and validated by the scientific community, to account for variance in primary neuroimaging data. In particular, our example focuses on one of the most common algorithms for coordinate-based meta-analyses: activation likelihood estimation, ALE [70, 47]. We will perform a meta-analysis using the modified version of ALE proposed by Eickhoff et. al .[30]. This modification is based on the idea of using between-subjects and between-templates variance to estimate the size of the modeled Gaussian from which to compute the corresponding FWHM.

For this purpose, we will use the BrainMap database [46], composed of 3,112 publications totaling 15,256 experiments, which provides us with information on the number of subjects present in each experiment. Our program will use three different atoms from this database: StimMod, which relates each experiment to its stimulus modality, StimType, which does the same with the type of stimulus, and finally BrainMap, composed of the list of publications and experiments included in the database, the number of participants, and the reported activations. Based on empirical measurements made in 2009 by the BrainMap team, we can calculate the FWHM that includes variation between-subjects and between-templates with the following formula. Given N, the number of subjects in the experiment, the formula is defined as

$$\text{FWHM}(N) : \sqrt{\pi \ln(2) \left(5.7^2 + \frac{11.6^2}{N}\right)} \tag{7.6}$$

The calculation includes the square root of the inverse of the user-specified number of subjects. For our example, we used 142 studies related to an auditory stimulus modality among one of the following types: "Vocal Sounds", "Nonvocal Sounds", "Sounds (Environmental)", or "Nonverbal Vocal Sounds". Listing 7.3 presents the program used to calculate the modified ALE. The rule defining the *Activation* atom uses syntactic sugar to define an expression that assigns probabilities to each of the possible values based on the formula for a three-dimensional Gaussian distribution defined in Laird et. al. [47]. Based on algorithm 1, this rule adds a new probabilistic relation *Activation* to C where the probability is computed according to an expression that can only contain elements of the rule body belonging to D or Σ, or constants. The variable '*d*' is the Euclidean distance

between both points (i, j, k) and $(i1, j1, k1)$. Function *sigmaGivenSubjects* calculates, given the number of subjects who participated in the experiments reported by BrainMap, the formula defined in eq. (7.6). Finally, '*resolution*' is a constant that defines the resolution of the brain image used in the experiment. At the same time, we will present results using the classical ALE variant as a reference, with an FWHM value manually selected of 9. The code for this program can be found at Listing 7.4.

Listing 7.3: Program code that computes the modified version of ALE

```
StimTypeAuditory(bmapID, expID) :- StimMod('auditory',bmapID,expID),
        StimType('vocal sounds', bmapID, expID)
        ⋮
StimTypeAuditory(bmapID, expID) :- StimMod('auditory',bmapID,expID),
        StimType('nonverbal vocal sounds', bmapID, expID)


Activation(i, j, k) @ max(
        (exp(-(1 / 2) * (d / sigma) ** 2)
        / ((2 * pi) ** (3 / 2) * sigma ** 3))
        * (resolution ** 3)
    ) :- StimTypeAuditory(bmapID, expID),
    BrainMap(bmapID, expID, ..., ..., minSubj, i1, j1, k1),
    Voxels(i, j, k)
    (d == FocusCoactivates(i, j, k, i1, j1, k1)),
    (sigma == sigmaGivenSubjects(minSubj))

Ans(i, j, k, PROB) :- Activation(i, j, k, p)
```

Figure 7.5 presents a comparison of the results of both algorithms. The ALE scores for each voxel in the reference space were calculated and filtered using the 95th percentile of the modified ALE as the threshold for comparison.

Figure 7.3 shows a more accurate selection of voxels than Figure 7.4 concerning the expected results for an auditory stimulus modality. This is because the modified version of ALE allows us to weigh each voxel according to the number of subjects that participated in the experiment. Though Figure 7.4 tends to show expected results, it is unable to capture the variance and relies on each experiment present in the BrainMap database with the same "weight", leading to noisier results.

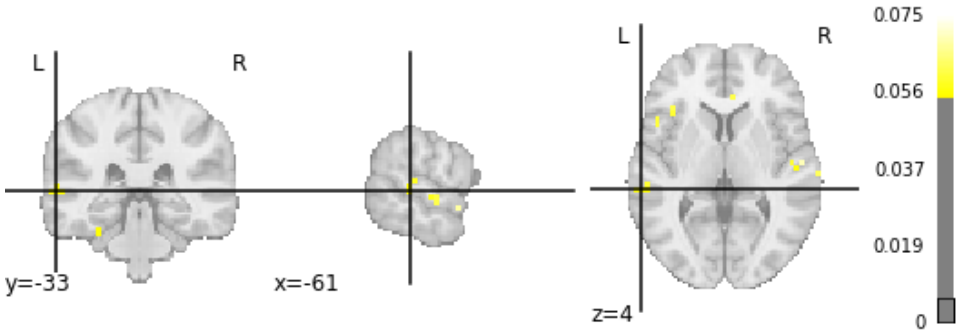

Figure 7.3: Modified ALE, accounting for between-subjects and between-templates variance

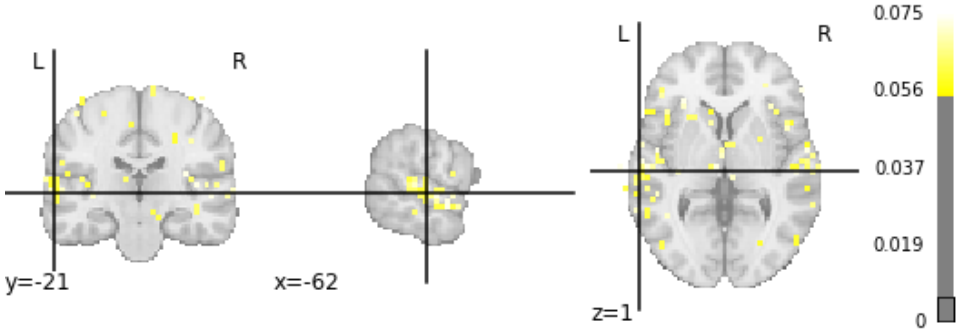

Figure 7.4: Classic ALE, with FWHM = 9

Figure 7.5: Comparison of results between ALE [70, 47] and Modified ALE [30] Figure 7.3 shows a more accurate selection of voxels than Figure 7.4 concerning the expected results for an auditory stimulus modality. This is because the modified version of ALE allows us to weigh each voxel according to the number of subjects that participated in the experiment. Figure 7.4 is unable to capture the variance and relies on each experiment present in the BrainMap database with the same weight, leading to noisier results.

Listing 7.4: Program code that computes the classic version of ALE

```
StimTypeAuditory(bmapID, expID) :- StimMod('auditory',bmapID,expID),
        StimType('vocal sounds', bmapID, expID)
        :
StimTypeAuditory(bmapID, expID) :- StimMod('auditory',bmapID,expID),
        StimType('nonverbal vocal sounds', bmapID, expID)


Activation(i, j, k) @ max(
        (exp(-(1 / 2) * (d / 9) ** 2)
        / ((2 * pi) ** (3 / 2) * 9 ** 3))
        * (resolution ** 3)
    ) :- StimTypeAuditory(bmapID, expID),
    BrainMap(bmapID, expID, ..., ..., minSubj, i1, j1, k1),
    Voxels(i, j, k),
    (d == FocusCoactivates(i, j, k, i1, j1, k1))

Ans(i, j, k, PROB) :- Activation(i, j, k)
```

8 - Neuroscientific Ontological Knowledge in the context of NeuroLang

Abstract In this chapter, we present the second main contribution of this dissertation. We will delve deeper into the world of ontologies, expanding on what has already been introduced in chapter 6. In particular, we will look at some advantages of using ontologies, such as the ability to resolve queries under the open-world assumption or the possibility of using the hierarchical information of ontologies to improve our results. Finally, we will present a real-world use case in neuroscience research in which, by combining ontologies and other heterogeneous databases such as the Julich atlas or the NeuroSynth meta-analysis database, we are able to provide a multilevel-characterization of the different histological regions in the atlas with respect to cognitive processes.

8.1 . Solving queries under the Open World Assumption

As we said in the chapter 6, Ontologies are primarily based on a family of formal knowledge representation languages known as Descriptions Logics (DL), which are a subset of first-order logic [63]. As DLs have been designed to deal with the problem of incomplete information, instead of making assumptions to specify a particular interpretation fully, the semantics considers all possible situations in which the axioms are satisfied. Since it keeps unspecified information open, this feature is called open-world assumption (OWA) [2]. This assumption is also implied by the fact that most DLs are fragments of first-order logic, which also adheres to the OWA. Therefore, the lack of a given assertion or fact does not imply whether the statement is true or false: it is not known. Thus, the open-world assumption requires considering of existentially quantified variables to denote unknown individuals (as in eq. 6.1). This property allows us to answer queries over incomplete data by employing the implicit information the ontology provides. But at the same time, this often leads to more complex reasoning since the open-world assumption encodes a possibly infinite set of elements that comply with the statements described by the ontology. When this infinite expansion takes place, reasoning about ontologies under the open world assumption becomes intractable.

A most promising approach to deal with the infinite expansion where answering queries including knowledge under the open-world assumption are Rewriting algorithms [36, 12]. This class of algorithms allows us to rewrite our ontological query in first-order logic so that the domain of our database with which it interacts is preserved. In particular, these algorithms usually restrict the type of queries that can be rewritten, avoiding an infinite expansion of possible results.

Our implementation is based on XRewriter, an algorithm proposed by Gottlob

et al. [36] that guarantees tractability for ontologies that belong to the DL-Lite family [16]. DL-Lite is a member of Description Logics (DL), a family of knowledge representation languages used to define ontologies. DL-Lite is specifically tailored to capture basic ontology languages while keeping polynomial data complexity (polynomial time) for query answering. Fortunately, as far as we know, DL-Lite is the adopted language in neuroinformatics for constructing ontologies.

To test Neurolang's ability to solve queries based on open knowledge, we will use the NeuroSynth meta-analysis database and the Cognitive Atlas ontology (CogAt). The Cognitive Atlas is a collaborative knowledge-building project that aims to develop an ontology that characterizes the state of current thought in cognitive science. It defines a set of mental concepts along with a set of mental tasks and the measurement relations between those classes [58]. Our approach will then use terms present in the NeuroSynth database to associate studies with the cognitive processes proposed by CogAt. This approach allows us to bridge both databases, using the information structured in the CogAt ontology to ask questions that result in a list of studies and their corresponding activation maps. As we mentioned, ontological knowledge is interpreted under the open-world assumption, which entails that the model represents partial knowledge about a domain. Therefore, we will focus on solving queries based on some ontology constraints defined as *someValuesFrom*. These constraints usually represent open-world knowledge and have the common characteristic of being defined by using existentially quantified variables in the rule's consequent (or head) as the one defined in equation 6.1.

In particular, we aim at inferring the brain activation pattern associated with *visual awareness*. This process is encoded in the CogAt ontology, but is not found in the Neurosynth database neither as a term nor a topic. Therefore, we need a way to link the studies and activations related to this process, given that it is not explicitly mentioned in the database. CogAt helps us in solving this problem: a Tuple-generating dependency (TGDs or existential rules) specifies that *spatial attention* is a sub-process of *visual awareness*. Which, expressed as a Datalog+/- rule in CogAt's TGD set, is:

$$\forall X \text{SpatialAttention}(X) \rightarrow \ \exists Y \text{PartOf}(X, Y) \wedge \text{VisualAwareness}(Y) \in \Sigma_1^{\text{CogAt}}, \tag{8.1}$$

which has an existentially-quantified variable Y in the head, representing open-world knowledge.

We seamlessly harness the open knowledge presented in CogAt to analyze activations related to visual awareness using to NeuroLang's built-in capabilities: we write a program (see Listing 8.1) to obtain all studies that, while not mentioning visual awareness, mention terms which, according to CogAt, imply that the cognitive process is involved. Importantly, we achieve this by combining an automatically-produced literature database with an expert-produced ontology. The resulting activations are presented in figure 8.1.

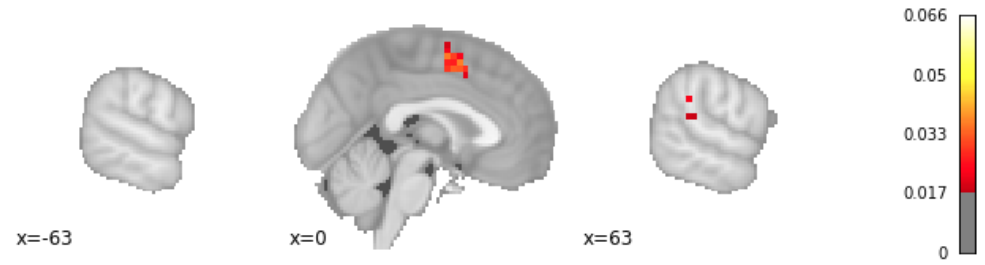

Figure 8.1: The resultant brain activation map corresponding to *'spatial attention'*, obtained through the resolution of a forward inference query under the open-world assumption. The activation map is thresholded at the 95th percentile

Listing 8.1: Open world assumption.

```
Entity(t, s) :- TermInStudy(t, s)
OpenWorldStudies(s) :- PartOf(t, s), VisualAwareness(s).
ProbMap(x, y, z, PROB) :- FocusReported(x, y, z, s) //
    (SelectedStudy(s) & OpenWorldStudies(s))
ProbabilityImage(create_region_overlay(x, y, z, p)) :-
    ProbMap(x, y, z, p)
```

8.2 . Retrieving synonyms via the hierarchical structure of the ontology

One of the main characteristics that differentiate first-order logic from propositional logic is the impossibility of the latter to define quantified variables. Although this limitation allows the queries of languages based on propositional logic to be solved in a computationally efficient manner, but at the same time, it imposes strong limitations in terms of expressiveness. This limitation is due to the absence of quantifiers that allow defining relations, which makes it necessary to explicitly establish each of the variables on which we want to operate with our query. Thus, this limitation requires that we know in advance each of the terms of interest, which makes the queries increasingly complex as the number of elements increases.

To overcome this limitation, we can take advantage of domain-specific logical theories formalized within ontologies. Since ontologies are formal representations of knowledge based on a set of concepts and the relationships between them, we could leverage first-order logic to take advantage of these relationships and increase the domain covered by our query. With this approach, we can use the relationships between terms already defined within the ontology to link the terms in the query. We thus use the expert knowledge described by the ontologies in terms of defining

the relationships. We do not need to know which terms are related to each other in advance.

We will show how we can take advantage of NeuroLang's ability to process ontologies and solve complex queries. We will introduce an example that asks about a series of terms and their related synonyms without the need to know them beforehand or define them specifically in our query. For this purpose, we will write a query using NeuroLang that reproduces Neurosynth's forward inference for the term *'pain'*. Unlike NeuroSynth, we will take advantage of NeuroLang's first-order language to automatically obtain all the synonyms of the term, from the CogAt ontology, without the need to know them explicitly beforehand. In particular, we will focus on a query that, making use of the structured knowledge specified in the IOBC ontology and the possibility of manipulating existentially quantified variables of our language, obtains the *'altLabel'* property of the entities related to *'Pain'*. Then, we will use the obtained result and the Neurosynth database to infer the most commonly activated voxels in studies where these terms are significantly present. To achieve this, we will get the conditional probability of each of the voxels to be activated, given that one or more of the selected terms was mentioned in the study. Finally, we filter out the resultant activation map at the 95th percentile of voxel weights.

In this example, we show how we can leverage the ontological knowledge provided by the International Organization for Biological Control (IOBC) [45] to perform an analysis that includes terms related to our main term (*noxious* and *nociceptive* related to *pain*, in this example) without knowing them beforehand, enriching our results automatically.

Listing 8.2: Retrieving information from related terms via the hierarchical structure of the ontology

```
RelatedTerm(term) :- term == "pain".
RelatedTerm(term) :- label(pain_entity, "pain"),
    related(pain_entity, subclass),
    altLabel(subclass, term).
FilteredBySynonym(t, s) :- TermInStudy(t, s), RelatedTerm(t).
Result(i, j, k, PROB) :- FocusReported(i, j, k, s) // (SelectedStudy(s),
    FilteredBySynonym(t, s)).
Percentile_95(compute_percentile(p, 95)) :- Result(i, j, k, p).
VoxelActivationImg(create_region_overlay(i, j, k, p)) :-
    Result(i, j, k, p),
    Percentile_95(p95), p > p95.
ans(img) :- VoxelActivationImg(img).
```

Fig. 8.2 provides a view of the results obtained from this example (see Listing 8.2). In this case, the activations of Noxious and Nociceptive were also automati-

cally included in the result.

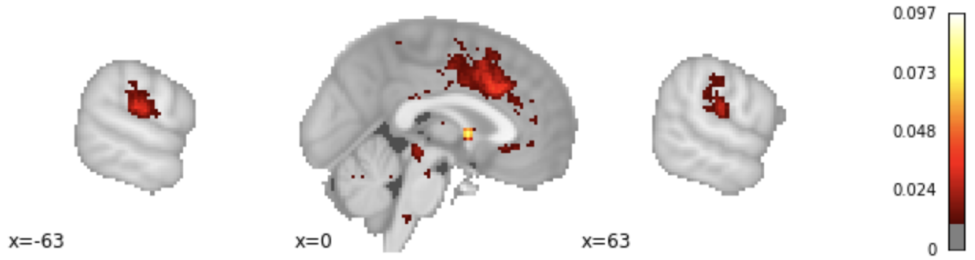

Figure 8.2: The resultant brain activation map associated with *'pain'* and its related terms (*'noxious'* and *'nociceptive'*) defined by the IOBC ontology. The results show that the dorsal anterior cingulate cortex and parietal regions are consistently active in studies mentioning "pain" or any related term

8.3 . Multilevel characterization of brain regions through large-scale reverse inference with heterogeneous data sources

Finally, we will show how to integrate some of the ideas discussed in previous examples to model a meta-analysis experiment that links regions defined based on histology with a set of hand-curated cognitive processes.

We will use NeuroLang to solve a set of queries that infer the probability of a term being mentioned in a study, given that a particular region of the brain is activated and the likelihood of the same term being mentioned given that the region is not activated. A brain region is considered active in a specific study if at least one of the reported coordinates matches one in the selected area.

Terms and studies were obtained from the Neurosynth database. The acquired terms were then filtered based on the cognitive process defined in the CogAt ontology to remove noise irrelevant to our analysis. In addition, we use CogAt's hierarchical organization of cognitive processes, which organizes those terms into general categories known as "top concepts". We also generated 150 samples randomly, allocating 70% of the studies. Julich's Brain Atlas was used to partition the brain into regions based on cytoarchitecture.

As the last step, the Bayes Factor of both hypotheses was computed, and terms with a value less than 3.16 were filtered out. Those values were used to analyze the preponderance of different cognitive processes in cytoarchitectonic areas. In the case of top concepts, these appear if at least one term belonging to the top concept passes the threshold, so values lower than 3.16 may be present.

It is essential to highlight that all this procedure, from filtering the Neurosynth's terms to the calculation and later filtering of the Bayes Factor, was carried out within the same NeuroLang program, allowing its easy reproduction and subsequent updating in the case of new data. The code of this program can be found in

Listing 8.3

In this experiment, we perform a large-scale reverse inference using NeuroLang and heterogeneous data: the Julich probabilistic histological atlas and an ontology, CogAt. We aim to construct a knowledge tree linking histology to cognitive ontology for a richer, multilevel characterization of brain regions. The importance of this analysis is that it formalizes reverse inference by combining probabilistic inference, fine-grained atlases, and detailed cognitive ontologies in a universal language. Although reverse inference may still be imperfect at this point, as more data becomes available and the CogAt ontology develops further, the exact same queries can be used to update the results. This analysis could provide new insights into which ontological entities are biologically realized and which are not and how closely-related conceptual structures differ in their neural substrates. Furthermore, this analysis could be extended to include network connectivity estimations for an unequivocal characterization of brain regions.

Listing 8.3: Multilevel characterization of brain regions

```
ActivationsJulich(i, j, k, regionId, study) :-
    Activations(study, i, j, k), JulichBrain(p, i, j, k, regionId)

ActiveRegion(study, regionId) :-
    ActivationsJulich(i, j, k, regionId, study)

NonActiveREgions(study, regionId) :-
    ~ActiveRegion(study, regionId), Studies(study),
    JulichAtlas(regionId, name, hemis)

TermProb(term, fold, PROB) :-
    FilteredTerms(study, term) //
        (ActiveRegion(study, regionId), Folds(study, fold),
        Studies(study))

NegTermProb(term, fold, PROB) :-
    FilteredTerms(study, term) //
        (NonActiveRegion(study, regionId), Folds(study, fold),
        Studies(study))

Answer(term, fold, bayesFactor) :-
    TermProb(term, fold, prob), NegTermProb(term, fold, negProb),
    bayesFactor == (prob / negProb)
```

To infer specific structure-function associations, we estimate the evidence in favor of the presence of a general (top) concept or a particular term given activation

in a region compared to when given no region activation. Terms and top concepts are organized using the hierarchy provided by CogAt, which categorizes sets of different cognitive terms that appear in articles under broad categories known as top concepts (Emotion, Language, etc.). We use the Bayes factor (BF), which represents the ratio of the posterior probability of one hypothesis to that of another, as a measure of the strength of evidence in favor of the association. A BF > 1 suggests evidence for association, whereas BF > 3 is generally regarded as an indicator of considerable evidence [43]. Although we use all the regions defined by the Julich atlas in the reverse inference, we discuss only two regions of interest for brevity: hippocampus and deep cerebellar nuclei. The results are summarized in figures 8.4, 8.5, 8.7 and 8.8 using circular trees.

Figures 8.4 and 8.5 depict the reverse inference results for hippocampus. The hippocampus is a sub-cortical brain region crucial for long-term memory encoding and retrieval as well as involved in emotion-related processing. The Julich atlas (version 2.9) divides the hippocampus into ten sub-regions in both brain hemispheres. We observe that most hippocampus sub-regions are associated with the top concept "Learning and Memory", especially with its constituent terms: autobiographical memory, episodic memory, and consolidation. Likewise, the term "navigation" found under the top concept "Perception" is also associated with all hippocampus sub-regions except the hippocampal amygdala transition area (HATA). These results recapitulate the long-standing view of hippocampus's major role in memory processes. Nonetheless, the results also reveal fine differences among hippocampal sub-regions, even across hemispheres. For instance, most left hippocampus regions are associated with "semantic memory "under the top concept "Language", while only the "Presubiculum", "Parasubiculum", and "CA3" in the right hippocampus are. This result is consistent with the dominance of language/semantic processing in the left brain hemisphere. Furthermore, the HATA seems to be the one region strongly linked to terms of "Emotion", such as "valence" and "facial expression", in both hemispheres. Note that a region's associations with singular terms are much stronger than with top concepts, as the latter represent overly broad behavioral domains that are not specific to any brain region.

Figures 8.7 and 8.8 show the reverse inference results for the deep cerebellar nuclei (DCNs). DCNs are grey matter clusters embedded in the white matter of the cerebellum. They are the sole output of the cerebellum and form part of the closed-loop system connected to the sensorimotor, association, and limbic cortices. The Julich atlas defines three DCNs: fastigial (FN), interposed (IN), and dentate (DN). The cerebellum has been widely viewed as a sensorimotor region, but this view has drastically changed in the past few decades. The cerebellum is now believed to be involved in as many functions as the cerebral cortex. This is evident in the results of the present reverse inference meta-analysis. For each nucleus, in addition to associations with terms under "Action" and "Perception", we observe stronger associations with terms of "Language", "Learning and Memory"

and "Attention". One nucleus, the IN in the left cerebellum, is also associated with "maintenance" under "Executive-Cognitive Control" and "pain" under "Emotion". Nevertheless, caution must be taken when interpreting these results, as studies from the 1990s and early 2000s often discarded the cerebellum from their imaging experiments. Yet, the results are consistent with recent views of functional diversity in the cerebellum beyond sensorimotor function.

Future generation neuroscience demands a universal standard to synthesize the great loads of information into structured knowledge. In this work, we have proposed that NeuroLang, a unified formal language for functional neuroanatomy, is well suited for leading computational approaches to formalize large-scale neuroscience research, including meta-analysis, through probabilistic first-order logic programming. NeuroLang can seamlessly integrate heterogeneous data and enable mapping cognitive domains to brain regions through a set of formal criteria. We harness the expressiveness of NeuroLang to overcome the limitations of standard meta-analysis tools, promoting elaborate and highly reproducible research on the domains of brain function. Towards achieving this goal, we provide use-case examples that demonstrate how NeuroLang combines useful data sources, such as ontologies and brain atlases, and allows us to express and solve meta-analytic queries using declarative and compact programs.

Commonly used meta-analysis tools are limited in their formal expressivity as they mainly rely on propositional logic semantics to query databases. Propositional logic can at best be used to study a limited number of easy-to-define concepts at a time, as it only deals with specific facts that carry a truth value. However, when faced with a question that requires inferring associations among *a variable* number of instances of objects such as behavioral domains, tasks, and brain regions, while combining external data, propositional logic is inefficient. This is because it entails that each instance of the objects be explicitly declared in a meta-analysis and already be defined in the database at hand. Effectively, such a meta-analysis is not scalable as it's not possible to write queries containing existentially quantified variables ranging over all instances of objects and formally represent multi-level associations among them. Therefore, the breadth and specificity of queries that can be expressed and solved with current tools are limited, impeding progress in the field.

As our examples demonstrate, NeuroLang expands the scope neuroimaging meta-analysis through its ability to combine and elaborately query heterogeneous data in a single unifying framework. This allows us to make use of human-curated databases, for instance, and take full advantage of knowledge models created by experts in the field. A common way of modeling expert knowledge is ontologies, which are high-level conceptual organizers of information. An ontology in cognitive neuroscience presents a hierarchical schema for systematically fitting mental functions and cognitive tasks together based on collaborative knowledge building. Significant advances have been made in the past two decades to provide

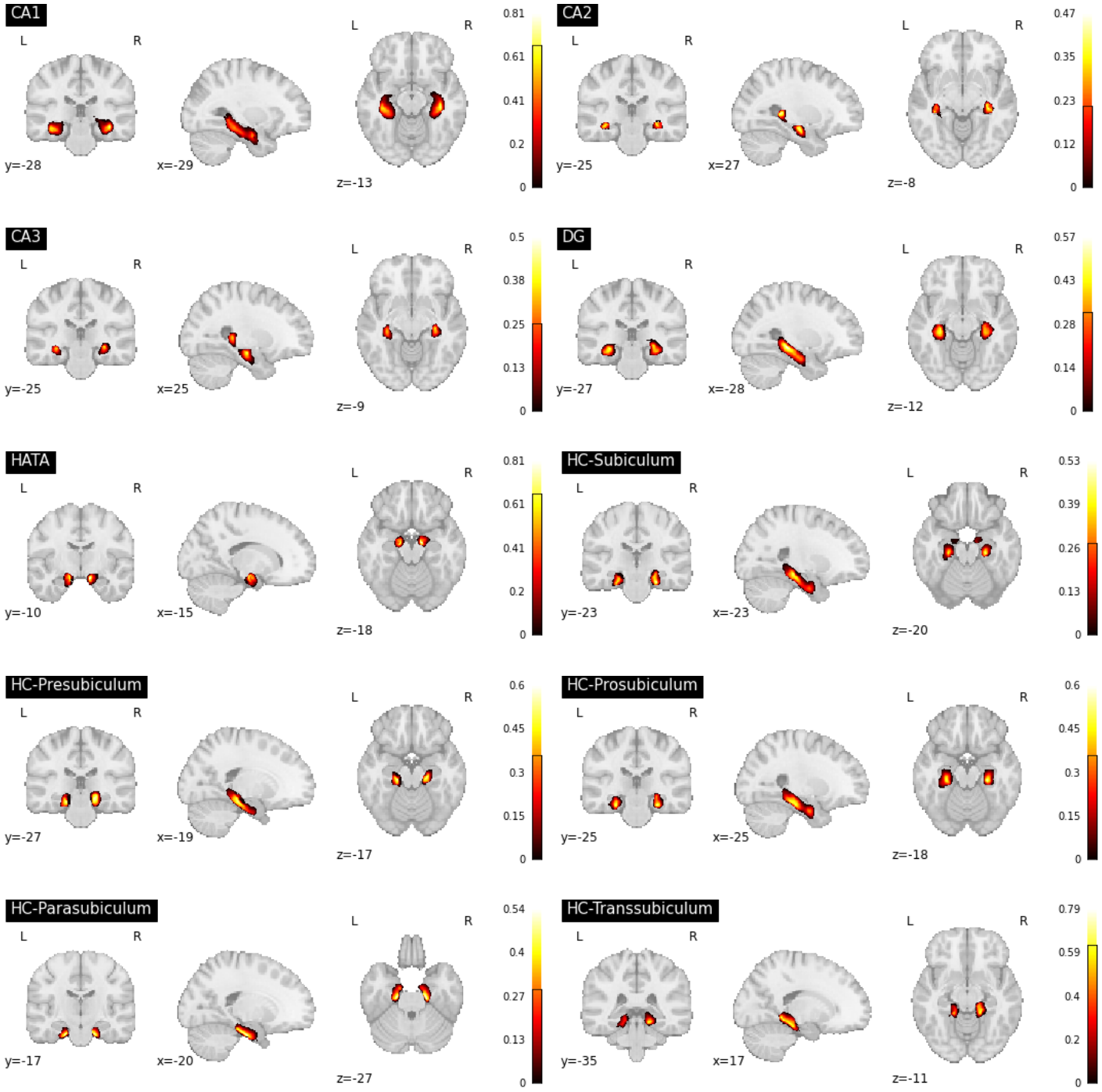

Figure 8.3: Probabilistic maps of the hippocampal subregions defined by Julich's brain atlas used in the reverse inference analysis [3].

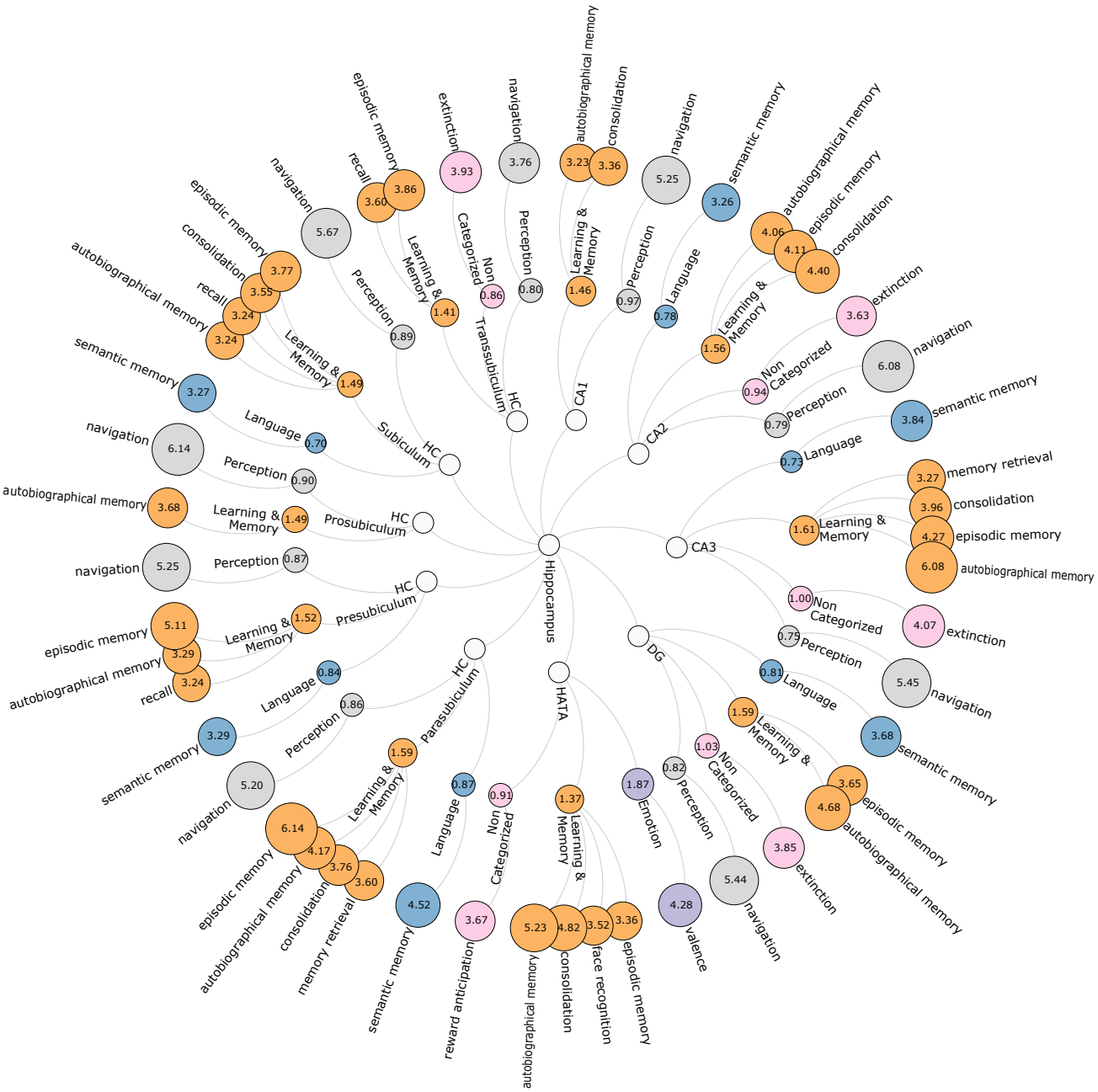

Figure 8.4: Reverse inference results for left hippocampus. Estimation of the amount of evidence in favor of the presence of a general (top) concept or a particular term given activation in a region compared to when given no region activation. We use the Bayes factor (BF), which represents the ratio of the posterior probability of one hypothesis to that of another, as a measure of strength of evidence in favor of association.

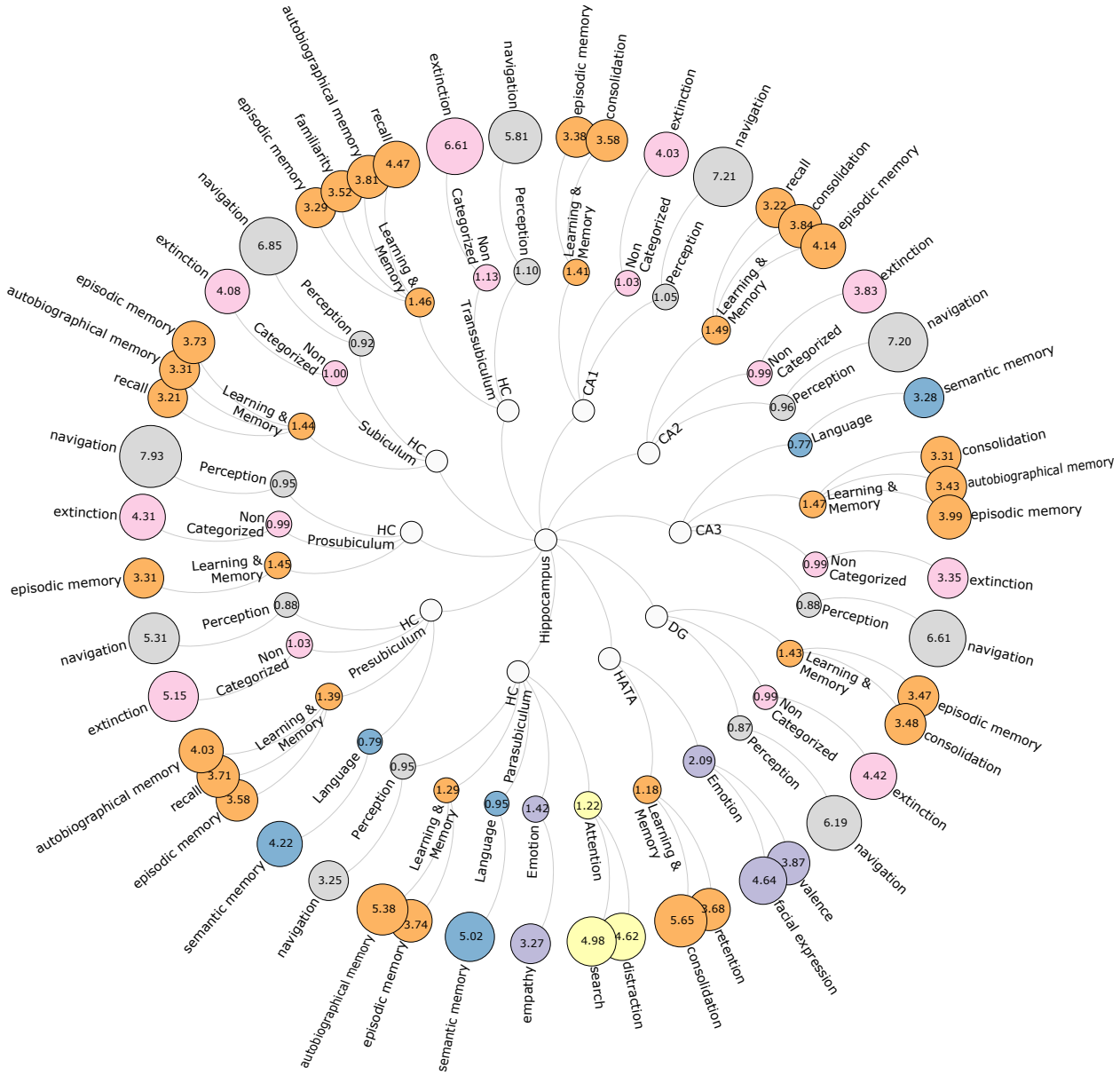

Figure 8.5: Reverse inference results for right hippocampus. Estimation of the amount of evidence in favor of the presence of a general (top) concept or a particular term given activation in a region compared to when given no region activation. We use the Bayes factor (BF), which represents the ratio of the posterior probability of one hypothesis to that of another, as a measure of strength of evidence in favor of association.

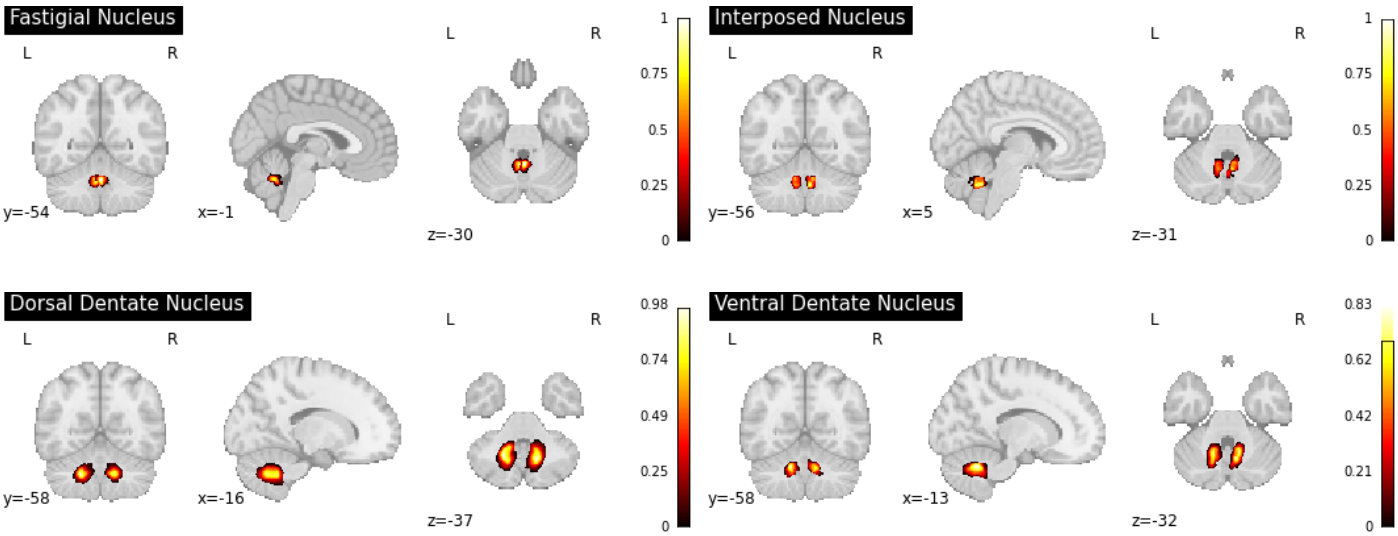

Figure 8.6: Probabilistic maps of the Cerebellum subregions defined by Julich's brain atlas [3].

knowledge bases for neuroscience, such as the Cognitive Atlas (CogAt) [58] or Cognitive Paradigm Ontology (CogPo) [71]. However, computationally leveraging the hierarchical structure of these ontologies, preserving all its expressiveness, in a well-grounded way that links it with empirical evidence from neuroimaging to drive new discoveries remains an open challenge.

Integrating ontologies and leveraging their hierarchical structure can be readily performed with NeuroLang to solve queries under the open-world assumption (OWA). This assumption allows us to infer the neural correlates of *Visual Awareness*, for instance, through its sub-parts defined using CogAt, without it being explicitly mentioned in the Neurosynth database. Thus, by leveraging NeuroLang's expressivity and the CogAt's knowledge base we can make inferences not only on the information contained in a meta-analytic database, but also on missing, but plausible, information. In contrast, standard meta-analysis tools solve queries under the closed-world assumption, where the truth is entirely present in the data. Using Neurosynth, for instance, it's not possible to query terms not present in the database nor retrieve a set of related terms without explicitly declaring every single one of them. Using Neurosynth, for instance, it's not possible to automatically retrieve the set of terms related to *Pain* as we need explicit prior knowledge of each one of them. BrainMap, on the other hand, includes an ontology, CogPo, which explains how the database is organized. However, since BrainMap's query system is also based on propositional logic, it only grants access to either overly broad behavioral domains or singular sub-domains, separately. Therefore, it necessitates that we have prior knowledge of and explicitly declare all terms of interest for meta-analysis. Conversely, through first-order logic semantics in NeuroLang, fine-

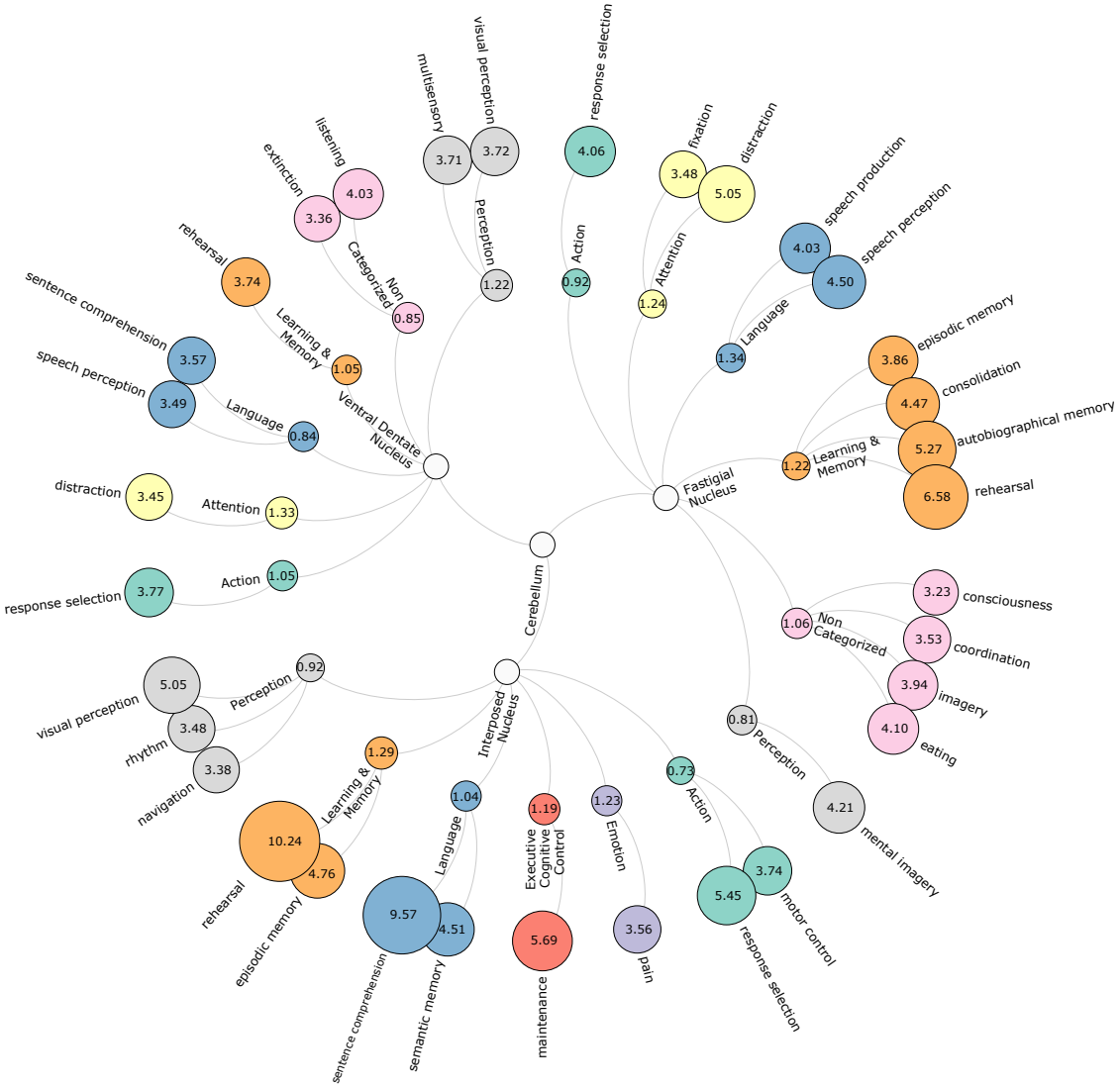

Figure 8.7: Reverse inference results for left cerebellum. Estimation of the amount of evidence in favor of the presence of a general (top) concept or a particular term given activation in a region compared to when given no region activation. We use the Bayes factor (BF), which represents the ratio of the posterior probability of one hypothesis to that of another, as a measure of strength of evidence in favor of association.

Figure 8.8: Reverse inference results for right cerebellum. Estimation of the amount of evidence in favor of the presence of a general (top) concept or a particular term given activation in a region compared to when given no region activation. We use the Bayes factor (BF), which represents the ratio of the posterior probability of one hypothesis to that of another, as a measure of strength of evidence in favor of association.

grained hierarchical relations from *any* ontology can be automatically leveraged, expanding the scope of meta-analysis beyond the limits of databases and sparing us from the need to explicitly declare the entire domain of terms of interest.

NeuroLang's probabilistic logic semantics enable inferring specific structure-function associations through *formal and declarative* reverse inference. In many neuroscience textbooks and articles, the functions of brain structures are mainly inferred through qualitative assessments of findings [57]. This kind of informal reverse inference is not deductively valid and can be misleading. A reverse inference is only truly informative if it quantifies the ratio of a region's process-specific activation to the overall likelihood of its activation with other processes [57]. Fortunately, with the presence of large meta-analytic databases, a wide range of brain states can be compiled and quantitatively contrasted to infer the level of specificity in function to structure mappings. And with automated machine learning methods, such as natural language processing, it is easy to mine the literature for useful data like terms and activation patterns. However, performing large-scale reverse inference on hundreds of regions and terms/topics can be challenging with existing approaches. For example, a recent large-scale effort to infer an ontology from empirical evidence has revealed fine-grained mappings from terms of mental function to brain circuits [10]. The authors of this study conclude that this computational ontology captures scientific knowledge of human brain function better than human-curated ontologies. As groundbreaking as the results are, the methodology of the study is seemingly very sophisticated and potentially strenuous to reproduce independently, including exhaustive analysis steps, a brain atlas, multiple formal ontologies, large databases, and a substantial level of analytical flexibility. On the other hand, our reverse inference example, although not as developed, uses probabilistic first-order logic semantics, such as quantified variables and existential quantifiers, allowing reverse inference to be compactly expressed and executed, while combining heterogeneous data. Thus, NeuroLang's approach can reduce the complexity of analysis pipelines to a set of formally defined questions. In this sense, a user has only to worry about describing their desired results rather than explicitly declaring each step of the meta-analysis.

The reverse inference analysis presented herein provides an architecture to systematically merge fine-grained cognitive ontologies with multiscale brain atlases through neurobiological evidence. This serves to mitigate confirmation and reification biases by bringing insights into which expertly denoted distinctions in the mental process are biologically realized and which are not [10, 48]. Moreover, the results of this analysis may even be leveraged as a reference for future hypothesis generation. Importantly, however, this analysis as performed with NeuroLang presents a step in a long process of formal knowledge compilation in cognitive neuroscience. That is, in the long-run, the evidence we infer to link brain regions to ontologically organized terms can be updated, consolidated, or discarded as more data becomes available. Thus, reverse inference in this case serves as a discovery

strategy for "inference to the best explanation" given the available data [57]. For instance, we present results on the deep cerebellar nuclei linking them to terms related to disparate behavioral domains. However, the cerebellum in general is under-represented in the literature, being entirely excluded from neuroimaging experiments in older studies. This may have yielded amplified or reduced functional associations for the DCNs that do not reflect the true extent of their functional associations. Nonetheless, as more studies (that hopefully include the cerebellum) bring new results, these associations can be automatically updated using the same NeuroLang program. Alongside new results, as ontologies and brain atlases are further updated, finer structure–function mappings can be directly and systematically inferred atop existing knowledge.

8.3.1 . B2RIO

An open source tool based on the multilevel characterisation presented in this section is freely available for use. Following the methodology presented in this experiment, B2RIO[1] allows neuroscientists to obtain, from a mask of the human brain, a list of the cognitive processes associated with it, evaluated with respect to their specificity using the same hypotheses presented in this experiment.

[1] https://github.com/NeuroLang/B2RIO

Part III

Real-World Use Cases in Neuroscience Research

9 - Foundational number sense training gains are predicted by hippocampal–parietal circuits

In this chapter, we present some exciting real world experiments developed using NeuroLang which take advantage of some of its most attractive features, such as segregation queries or integrating ontologies and heterogeneous databases. In this particular section, we present the work led by Chang et al. [21] that uses the ontology integration capabilities of NeuroLang. This work is part of a collaboration through which we conducted a reverse meta-analysis to provide evidence of the association between hippocampal-parietal functional circuitry and learning and related functions, confirming the results obtained in the study.

9.1 . Summary of the work

The development of mathematical skills in early childhood relies on number sense, the foundational ability to discriminate between quantities. Number sense in early childhood is predictive of academic and professional success, and deficits in number sense are thought to underlie lifelong impairments in mathematical abilities. Despite its importance, the brain circuit mechanisms that support number sense learning remain poorly understood. In this work, a theoretically motivated training program to determine brain circuit mechanisms underlying foundational number sense learning in female and male elementary school-aged children (ages 7-10) was designed. The four-week integrative number sense training program gradually strengthened the understanding of the relations between symbolic (Arabic numerals) and non-symbolic (sets of items) representations of quantity. The study founds that the number sense training program improved symbolic quantity discrimination ability in children across a wide range of math abilities, including those with learning difficulties.

9.2 . NeuroLang's contribution

In this study, we use NeuroLang to perform a reverse inference meta-analysis of inter-regional co-activations across 14,371 fMRI studies and 89 cognitive functions to confirm a reliable role for hippocampal-intraparietal-sulcus circuits in learning. The study identifies a canonical hippocampal–parietal circuit for learning which plays a foundational role in children's cognitive skill acquisition. Findings provide important insights into neurobiological circuit markers of individual differences in children's learning and delineate a robust target for effective cognitive interventions.

We used reverse meta-analysis to estimate the probability that a term related

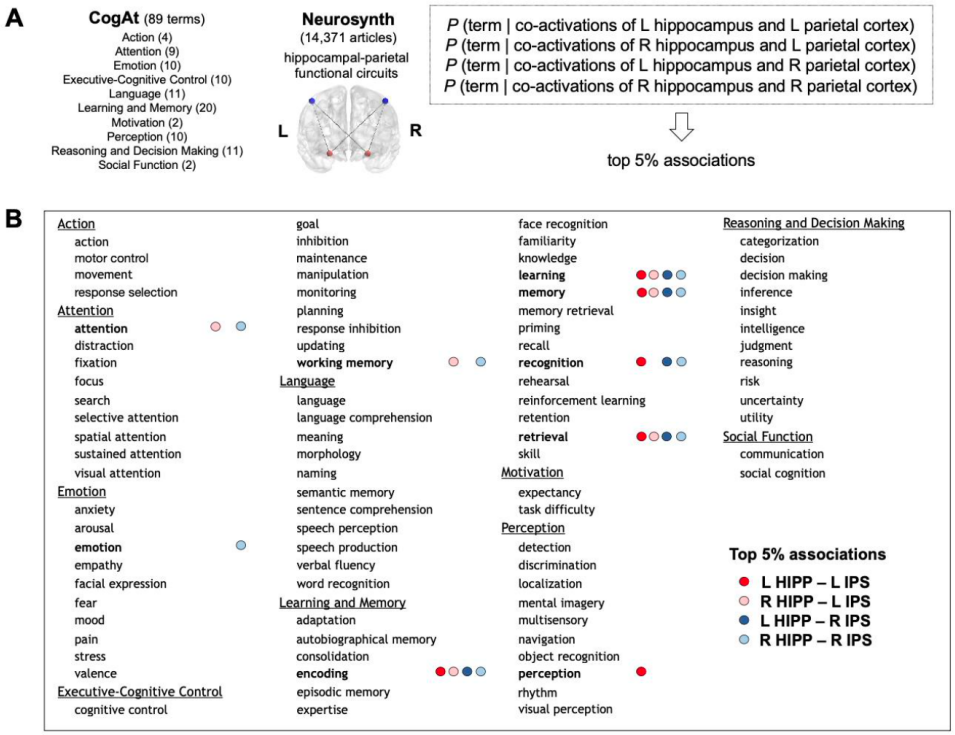

Figure 9.1: Reverse meta-analysis of 14,371 fMRI studies and cognitive functions reveals a significant association between hippocampal–parietal functional circuits and learning. **A**. A reverse meta-analysis was performed to map hippocampal–parietal functional circuits identified in the current study to cognitive functions (see Methods for details). **B**. Top 5% cognitive functions mentioned in published articles where co-activations of the left or right hippocampus (HIPP) and the left or right intraparietal cortex (IPS) are reported. L = left, R = right.

to a cognitive function was mentioned in an fMRI study under the condition that both regions to be analyzed, the hippocampus and the parietal cortex, were also reported to be active in the study for either hemisphere. In other words, given L, a cognitive function-related term defined in CogAt, and S, a study in the NeuroSynth database, we calculate:

$$P(\text{term L is mention in Study S} \mid \text{S reports activations in the left/right hippocampus and in the left/right parietal cortex}) \tag{9.1}$$

To estimate reverse meta-analysis probabilities, we followed the following steps. First, the probability of a term being present in a study was encoded by thresholding the term frequency-inverse document frequency (TF – IDF) value of the term being present at 10^{-3}, in agreement with NeuroSynth's implementation [78]. Second, we considered the probability of a region being reported in a given study as directly proportional to the number of activations within the regions being mentioned in the study. Third, terms were filtered using the CogAt [58] ontology to ensure that only those relating to cognitive processes (89 terms listed in Figure 9.1 B) were taken into account. To assess the stability of our estimations, we computed the confidence interval of our reverse meta-analysis probability estimations; we split the 14,371 studies into 20 equal folds, maximizing the measurements for estimation. Finally, the top 5% probable terms were considered to be sufficient evidence for associations with analyzed circuits. Our analysis resulted in the selection of 25 out of 356 associations (4 circuits, 89 terms), which represented above the 95 percentile of probable term mentions for studies where hippocampal parietal circuits were reported.

Our results from the reverse meta-analysis show that co-activations of both the left and right hippocampus and IPS are significantly associated with the term *learning* as well as related terms like *encoding*, *memory*, and *retrieval*. These meta-analytic findings from a large set of fMRI studies expand on findings from our training study and provide converging evidence for a strong association between hippocampal–parietal functional circuitry and learning and related functions.

10 - Functional gradients in the human lateral prefrontal cortex revealed by a comprehensive coordinate-based meta-analysis

Following the previous section, in this case, we present a paper led by Abdallah et al. [1] that adopts NeuroLang to infer the organizational principles of the lateral prefrontal cortex (LPFC) through meta-analysis.

Besides providing a small summary of the meta-analysis performed in this work, we will analyze two of the queries used since they present another exciting feature of NeuroLang that we want to highlight in this work, which is only possible because Neurolang implements a first-order logic resolution engine: Segregation queries.

10.1 . Summary of the work

The lateral prefrontal cortex (LPFC) supports a wide variety of cognitive processes considered hallmark features of the human brain. Understanding the functional organization of the LPFC is thus crucial to studying adaptive human behavior. Yet, the overarching organizing principles of the LPFC are still actively debated. Recently, large-scale attempts have been made to map the entire LPFC using conventional and meta-analytical approaches. So far, these mappings have lacked specificity due to the limited breadth and complexity of the queries that widely used tools can express and solve. In this study, a novel approach to expressive neuroimaging meta-analysis based on NeuroLang is adopted to infer the organizing principles of LPFC from thousands of studies with greater specificity.

The versatility of the LPFC suggests that it is far from being a unitary brain structure, and several hypotheses about its organization have emerged. For example, one hypothesis, which arises from the domain of abstraction and hierarchical cognitive control, proposes a rostrocaudal gradient in the LPFC [4, 7, 8, 14, 22, 44, 41, 54, 56]. In this spatial layout, caudal regions respond to immediate sensory stimuli, medial regions select actions based on an prevailing context, and rostral regions integrate concrete representations into more abstract rules to enable top-down temporal control of behaviour. The authors propose that the multitude of hypothesis on the LPFC organization is mainly due to the diversity of protocols and researchers' degrees of freedom across studies and therefore, it remains unclear to what extent the functional boundaries derived from each individual study correspond to the gross organization of the LPFC. On the other hand, regarding meta-analysis studies, the study proposes the same limitations we have discussed throughout this dissertation: commonly used tools must be more expressive to represent complex hypotheses of specific functional associations in the LPFC.

Finally, the study proposes to overcome these limitations by using NeuroLang.

More specifically by performing a meta-analysis on 14371 articles available in the NeuroSynth database together with a gradient-mapping approach to identify the organising principles in the LPFC.

10.2 . Segregation queries

The last analysis of the study leads by Abdallah[1] aims to characterize the two hemispheres of the LPFC in terms of specific topic associations in a gradient-like manner. In order to infer the variation of coactivation patterns along the primary gradient in the LPFC, they first create regions of interest from successive twenty-percentile gradient bins (i.e., five quintile bins) in the right and left LPFC. Then, they infer specific structure-function associations by estimating the extent to which a spatially-localized activation along the principal gradient in the LPFC predicts a Neurosynth topic's presence in a study.

For this purpose, they write a NeuroLang program that solves segregation queries between hemispheres. Here, they infer the probabilities "that a topic is present in a study given activation in a quintile bin in the right (respectively left) LPFC and there exists no reported activation in the entire left (respectively right) LPFC".

Segregation queries infer the probability that a topic is present in a study given spatially constrained activation within a range of quintile bins and the simultaneous absence of activation outside this range within the same hemisphere. Concurrently, a segregation query infers the probability of the opposite event: a topic is present, given that no activation within the quintile bins range or activation outside the range exists. The log odds ratio of these two hypotheses gives us a measure of evidence in favor of the association between a topic and patterns of activity along the principal gradient. The NeuroLang program that infers specific structure-function associations in the left LPFC using segregation queries is presented in Listing 10.1

Finally, they also write a program that applies inter-hemispheric segregation queries to infer the probability "that a topic is present given activation in a right lpfc quintile bin and there exists no activation in the entire left lpfc". The neurolang program that infers hemisphere-specific topic-bin associations is presented in Listing 10.2

The results of this analysis can be seen in Figure 10.1. Apart from the implications of these results, we need to highlight the ease with which NeuroLang segregation queries can obtain these results and the clarity that the queries offer for understanding the hypothesis analyzed. A procedure that might otherwise be hidden or obfuscated in a text description.

Segregation queries are one of the most relevant features of NeuroLang and its first-order logic-based resolution engine. Given their simplicity and NeuroLang's ability to combine heterogeneous data, these queries can open the door to answer-

Listing 10.1: Infering specific structure-function associations in the left LPFC

```
LeftBinActive(bin, study) :- LeftBinVoxel(bin, x, y, z),
    PeakReported(x2, y2, z2, study),
    distance == EUCLIDEAN(x, y, z, x2, y2, z2), distance < 3

SegregationRule(bin1, bin2, study) :-
    LeftBinActive(bin1, study), LeftBinActive(bin2, study),
    (bin2 >= bin1), ~exists(bin3;
        Bin(bin3), (bin3 < bin1 | bin3 > bin2), Study(study),
        LeftBinActive(bin3, study)
    )

NoSegregationRule(bin1, bin2, study) :-
    Study(study), Bin(bin1), Bin(bin2),
    ~SegregationRule(bin1, bin2, study)

TopicPresentGivenSegregationRule(topic, bin1, bin2, PROB) :-
    TopicInStudy(topic, study) //
    (SegregationRule(bin1, bin2, study),
    SelectedStudy(study))

TopicPresentGivenNoSegregationRule(topic, bin1, bin2, PROB) :-
    TopicInStudy(topic, study) //
    (NoSegregationRule(bin1, bin2, study),
    SelectedStudy(study))

TopicAssociationMatrix(topic, bin1, bin2, LOR) :-
    TopicPresentGivenSegregationRule(topic, bin1 , bin2, p1),
    TopicPresentGivenNoSegregationRule(topic, bin1, bin2, p0),
    LOR == log10((p1/(1 - p1))/(p0/(1 - p0)))

ans(topic, bin1, bin2, LOR) :-
    TopicAssociationMatrix(topic, bin1, bin2, LOR)
```

Listing 10.2: Infering hemisphere-specific topic-bin associations

```
LeftBinActive(bin, study) :- LeftBinVoxel(bin, x, y, z),
    PeakReported(x2, y2, z2, study),
    distance == EUCLIDEAN(x, y, z, x2, y2, z2), distance < 3

RightBinActive(bin, study) :- RightBinVoxel(bin, x, y, z),
    PeakReported(x2, y2, z2, study),
    distance == EUCLIDEAN(x, y, z, x2, y2, z2), distance < 3

OnlyLeftBinActive(bin, study) :- LeftBinActive(bin, study),
    ~exists(bin2;
        Bin(bin2), Study(study), RightBinActive(bin2, study)
    )

OnlyRightBinActive(bin, study) :- RightBinActive(bin, study),
    ~exists(bin2;
        Bin(bin2), Study(study), LeftBinActive(bin2, study)
    )

TopicPresentGivenOnlyLeftBinActive(topic, bin, PROB) :-
    TopicInStudy(topic, study) //
    (OnlyLeftBinActive(bin, study), SelectedStudy(study))

TopicPresentGivenOnlyRightBinActive(topic, bin, PROB) :-
    TopicInStudy(topic, study) //
    (OnlyRightBinActive(bin, study), SelectedStudy(study))

InterHemisphereTopicBinAssociation(topic, bin, LOR) :-
    TopicPresentGivenOnlyRightBinActive(topic, bin, p1),
    TopicPresentGivenOnlyLeftBinActive(topic, bin, p2),
    LOR == log10((p1/(1-p1))/(p2/(1-p2)))

ans(topic, bin, LOR) :-
    InterHemisphereTopicBinAssociation(topic, bin,LOR)
```

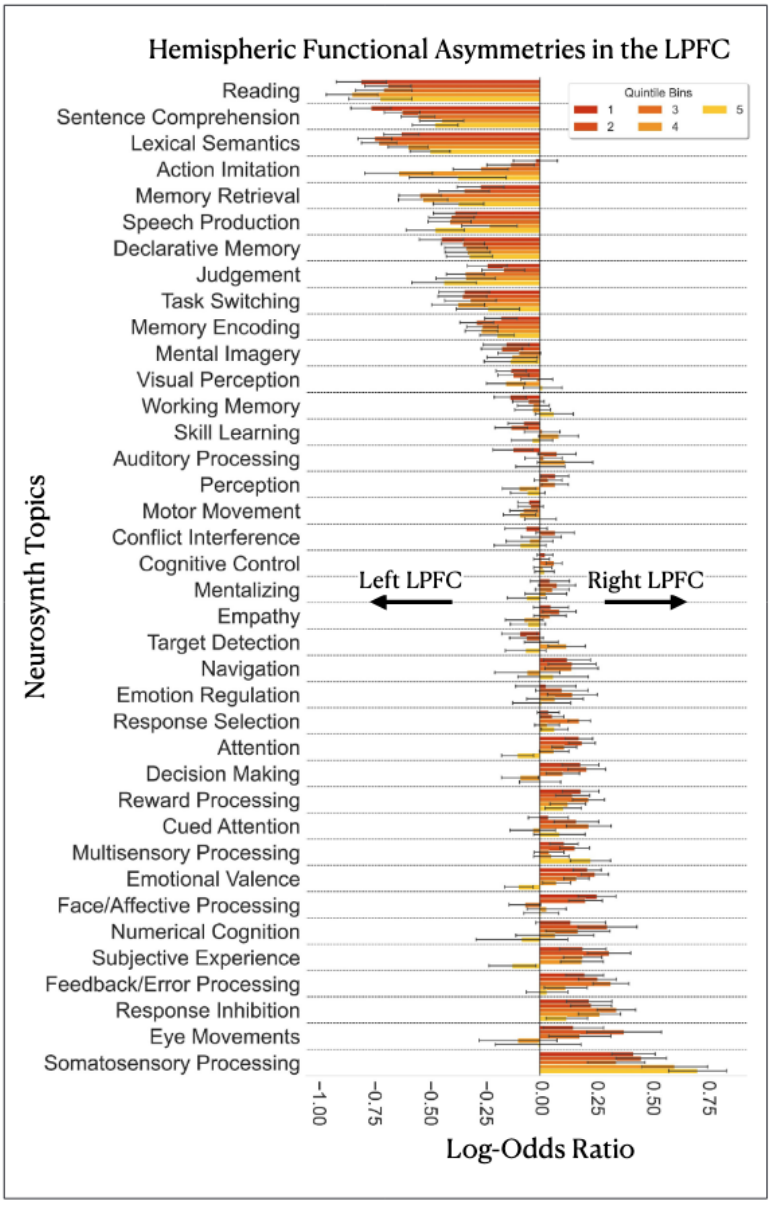

Figure 10.1: Gradient-based mapping of hemispheric asymmetries in the LPFC. Meta-analysis of inter-hemispheric asymmetries reasserts the left hemispheric dominance of language and memory and the right-hemispheric dominance of inhibition and sensory processing/monitoring in the LPFC. Positive log-odds ratios indicate evidence in favor of right-hemispheric preference of a topic in a given bin, whereas negative values indicate evidence in favor of left-hemispheric preference of a topic in a given bin. Error bars represent the 95% confidence intervals estimated from 5000 re-runs of the meta-analysis on random sub-samples of the Neurosynth dataset. Each random sub- sample comprises 60% of the studies of the original dataset (around 8623 studies). Topics are ordered from most-left dominant to most- right dominant based on the average of the log-odds ratio values over the five quintile bins.

ing hard-to-formulate questions.

Part IV

DISCUSSION

11 - Discussion

In this dissertation, we presented a fragment of probabilistic Datalog+/– enriched with negation and aggregation, along with a scalable query resolution algorithm coined NeuroLang. In addition, we present a series of practical applications that demonstrate its potential and some of the most exciting advantages provided by this tool, such as the integration of knowledge modelled under the open world assumption in chapter 8 or the use of segregations as in chapter 10.

Several different approaches to probabilistic Datalog+/– semantics and query resolution exist [36, 20]. Nonetheless, these do not incorporate aggregation, and the possibility of manipulating the probabilistic query results within the same language. These two features, as shown by our use-case analysis in Section 7.3, are fundamental traits required to provide a probabilistic logic programming language that can encode neuroimaging meta-analysis applications end-to-end.

The possibility of manipulating probabilities within the language comes at a great expense. After our PERs are computed, in Step 4 of Algorithm 1, our language allows handling probabilities as a standard float column. While this allows for analyses required by our target applications, it calls for disciplined programming from the user such that the manipulation of probabilities remains sound. Nonetheless, this gives our language great power; for instance, we can build probabilistic brain images, through aggregation, as shown in Section 7.3.1; and compute the probability differences between two events, which we show in Section 7.3.2.

All these features allow us to go beyond current tools in meta analyses whose queries are based in propositional logic [79, 46] and harness the full power of the $\text{FO}^{\neg\exists}$ fragment, as well as open-world semantics, to express meta-analysis tasks in a sound, disciplined, and declarative manner. Furthermore, by using a lifted query processing approach when possible (see Algorithm 1, Step 3), we are able to process current meta-analytic datasets enriched with ontologies that are of considerable size, as described in chapter 8. While it is true that there are other works that make possible the resolution of Datalog+/– queries ([20, 42]), the definition of the problem we wish to solve makes it necessary to have a framework capable of solving probabilistic choice and handling deterministic open-world knowledge. Moreover, we are not aware of any practical implementation of the mentioned works, beyond the provided theory. It's important to highlight that NeuroLang limits the representation of probabilistic atoms as mutually exclusive events or mutually independent events. We are aware of this limitation and of several advances in the field that overcome this limitation, such as MarkoViews [42]. Moreover, reasoning over fine-grained ontologies can be computationally intractable, and as we want to solve queries with at least a comparable performance as existing tools, it's necessary to make constraining decisions. In particular, we base our ontological query answering on XRewriter, an algorithm that derives knowledge from ontologies

proposed by Gottlob et. al. [36], providing practical and effective query resolutions. However, XRewriter only guarantees tractability for ontologies that belong to the DL-Lite family [36]. DL-Lite is a member of Description Logics (DL), a family of knowledge representation languages that can be used to build ontologies. The focus of the DL-Lite sub-family is to guarantee completeness (all newly defined relations are guaranteed to be computed) and decidability (all computations will finish in finite time) with maximum expressiveness. Fortunately, as far as we know, DL-Lite is the adopted language in neuroinformatics for constructing ontologies, and hence the impact of our decision should be currently negligible.

At a practical side, the multilevel characterization of brain regions through large-scale reverse inference with heterogeneous data sources presented in chapter 8 shows that we can not only make use of ontological knowledge to feed our queries as in example 8.2, but we can also embed our results in the hierarchy proposed by the ontology and obtain qualitatively more interesting results. On the other hand, we present two practical applications of neuroscientific interest in chapter 9 and 10 that, besides of being two important work with interesting results, allow us to highlight some of the most exciting features of NeuroLang, such as segregation queries. These experiments show that NeuroLang is more than just a project based on theoretical ideas without practical applications. They show a real need for an integrative framework that can solve queries based on first-order logic combining heterogeneous data. NeuroLang is not only used in-house in our lab, but is a mature and efficient tool that is starting to permeate other labs with their own projects and ideas. To conclude, we have shown that neuroimaging meta-analytic applications are an excellent real-world application for a language such as probabilistic Datalog+/−. By using probabilistic semantics that have recently converged from different probabilistic logic and open-world language approaches [60, 20, 73], with open-world semantics [19, 36, 20], and query resolution approaches [23, 20, 74], we have produced a language that is ready to be used in neuroimaging applications.

Finally, and on a personal note, I think NeuroLang is a great tool, it has a lot of potential and I hope that one day it will get the recognition it deserves. My personal hope is that NeuroLang will be the cornerstone of a future full of tools that enable unified formal interactions with large neuroscience databases, promoting openness and reproducibility in the field.

12 - Future improvements and beyond

In this section, I gather some ideas of possible improvements for the future of NeuroLang. Some of them are already materialising thanks to the work of other members of the team, while others are personal ideas or shortcomings that could not be addressed during my thesis.

12.1 . English controlled language

While it is true that programming skills are becoming more and more important, especially in research and even more so if large amounts of data consumption is required, our intention was always to provide a tool for neuroscientists, who are not necessarily experts in programming. Furthermore, we must take into account that logic programming languages, such as the one used in NeuroLang, are not widely used in general. That is why, in order to decrease the learning curve that the use of NeuroLang may impose on some people, and in order to offer a tool as accessible as possible to the average researcher, we decided and are currently working on an interface that provides NeuroLang with a controlled English-based language.

We believe this is a fundamental necessity if we want NeuroLang to go beyond the doors of our lab and be adopted by the community.

12.2 . Learning architecture

NeuroLang currently allows probabilistic inference over data but not learning to train machine learning models, and hence predictive models are currently not possible to implement. But given their wide and effective usage in discriminating cognitive functions and in deriving ontologies from empirical data [48], a learning architecture in NeuroLang is planned in the future.

12.3 . Performance improvements through parallelisation

While it is true that NeuroLang can match the performance of most of the meta-analysis tools against which it was compared during this work, there are still some bottlenecks that can be reduced. Especially since many of the optimisations made are at a theoretical level, for example, the use of optimisations to make a relational algebra execution plan more efficient. However, in my personal opinion, there are still optimisations to be made at the engineering level. For example, the parallelisation of part of its execution. I personally believe that data loading is one of the areas that would have the greatest benefit in relation to the cost of its implementation.

12.4 . Probabilistic ontologies

Modeling the real world requires the ability to be able to represent uncertainty. While not addressed during this work, probabilistic ontologies can also be used to model the uncertainty present in our data. While there has been significant progress in this area [61, 49], we decided to leave this feature out of NeuroLang due to the scarcity of probabilistic ontologies to take advantage of in the field of neuroscience. If this situation change and probabilistic ontologies gain popularity, extending NeuroLang's ontology engine with the ability to manipulate this data could be an exciting aspect of NeuroLang's future.

12.5 . Σ expressivity during rewriting using XRewriter

Since, in order to guarantee the correctness of NeurolangQA, we had to limit the expressivity of Σ in the case that rewriting by XRewriter was applied, we consider that it is of full interest for the future of NeurolangQA to study the possible expansion of Sigma during rewriting to know if we can guarantee its correctness in the case of stratified negation.

13 - List of publications

13.1 . Publications

- G. Zanitti, Y. Soto, V. Iovene, M. Martinez, R. Rodriguez, G. Simari, D. Wassermann - **Scalable Query Answering under Uncertainty to Neuroscientific Ontological Knowledge: The NeuroLang Approach**. Neuroinformatics (2022).

- G. Zanitti, M. Abdallah, D. Wassermann - **Multilevel characterization of brain regions through large-scale reverse inference with heterogeneous data sources**. To be submitted.

13.2 . Conferences

- G. Zanitti, M. Abdallah, D. Wassermann - **Towards Heterogeneous Data Integration: The NeuroLang Approach**. Organization for Human Brain Mapping (2022)

- G. Zanitti, V. Iovene, D. Wassermann - **Verifying ontological knowledge through meta-analysis: Study cases of Pain and Consciousness**. Organization for Human Brain Mapping (2021).

- G. Gallardo, G. Zanitti, A. Anwander, M. Higger, S. Bouix, S. Deslauriers-Gauthier, D. Wassermann - **Lesion-robust white-matter bundle identification through diffusion driven label fusion**. International Society for Magnetic Resonance in Medicine (2020)

13.3 . Collaborations

- M. Abdallah, V. Iovene, G. Zanitti, D. Wassermann - **Meta-analysis of the functional neuroimaging literature with probabilistic logic programming**. Scientific Reports (2022)

- M. Abdallah, G. Zanitti, V. Iovene, D. Wassermann - **Functional gradients in the human lateral prefrontal cortex revealed by a comprehensive coordinate-based meta-analysis**. eLife (2022)

- H. Chang, L. Chen, Y. Zhang, Y. Xie, C. de Los Angeles, E. Adair, G. Zanitti, D. Wassermann, M. Rosenberg-Lee, V. Menon - **Foundational number sense training gains are predicted by hippocampal–parietal circuits**. Journal of Neuroscience (2022)

- E. Levitis et al. - **Centering inclusivity in the design of online conferences — An OHBM–Open Science perspective**. GigaScience (2021)
- V. Iovene, G. Zanitti, D. Wassermann - **Complex coordinate-based meta-analysis with probabilistic programming** - AAAI (2021)
- A. Machlouzarides-Shalit, N. Makris, G. Zanitti, V. Iovene, G. M. Lemaitre, G. Favelier, D. Wassermann - **NeuroLang: representing neuroanatomy with sulcus-specific queries**. Organization of Human Brain Mapping (2020)

Bibliography

[1] Majd Abdallah, Gaston E Zanitti, Valentin Iovene, and Demian Wassermann. Functional gradients in the human lateral prefrontal cortex revealed by a comprehensive coordinate-based meta-analysis. *eLife*, 11:e76926, September 2022.

[2] S. Abiteboul, Richard Hull, and Victor Vianu. *Foundations of databases*. Addison-Wesley, Reading, Mass, 1995.

[3] Katrin Amunts, Hartmut Mohlberg, Sebastian Bludau, and Karl Zilles. Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture. *Science*, 369(6506):988–992, August 2020.

[4] C. Azuar, P. Reyes, A. Slachevsky, E. Volle, S. Kinkingnehun, F. Kouneiher, E. Bravo, B. Dubois, E. Koechlin, and R. Levy. Testing the model of caudo-rostral organization of cognitive control in the human with frontal lesions. *NeuroImage*, 84:1053–1060, January 2014.

[5] Franz Baader and TU Dresden. Least Common Subsumers and Most Specific Concepts in a Description Logic with Existential Restrictions and Terminological Cycles. page 6.

[6] Franz Baader, Ian Horrocks, and Ulrike Sattler. Chapter 3 Description Logics. In *Foundations of Artificial Intelligence*, volume 3, pages 135–179. Elsevier, 2008.

[7] David Badre. Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5):193–200, May 2008.

[8] David Badre and Mark D'Esposito. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, 10(9):659–669, September 2009.

[9] Jean-François Baget, Michel Leclère, Marie-Laure Mugnier, and Eric Salvat. On Rules with Existential Variables: Walking the Decidability Line. *Artificial Intelligence*, 175(9-10):1620, March 2011.

[10] Elizabeth Beam, Christopher Potts, Russell A. Poldrack, and Amit Etkin. A data-driven framework for mapping domains of human neurobiology. *Nature Neuroscience*, pages 1–12, November 2021. Bandiera_abtest: a Cg_type: Nature Research Journals Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cognitive neuroscience;Computational neuroscience;Diseases of the nervous system;Neural circuits;Scientific

community Subject_term_id: cognitive-neuroscience;computational-neuroscience;diseases-of-the-nervous-system;neural-circuit;scientific-community.

[11] C. Beeri and M. Y. Vardi. The implication problem for data dependencies. In Shimon Even and Oded Kariv, editors, *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 73–85, Berlin, Heidelberg, 1981. Springer.

[12] Meghyn Bienvenu. Ontology-mediated query answering: Harnessing knowledge to get more from data. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 4058–4061. AAAI Press, 2016.

[13] Meghyn Bienvenu and Magdalena Ortiz. Ontology-Mediated Query Answering with Data-Tractable Description Logics. In Wolfgang Faber and Adrian Paschke, editors, *Reasoning Web. Web Logic Rules*, volume 9203, pages 218–307. Springer International Publishing, Cham, 2015. Series Title: Lecture Notes in Computer Science.

[14] Matthew M. Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, 12(5):201–208, May 2008.

[15] Andrea Cali', Georg Gottlob, and Andreas Pieris. Towards more expressive ontology languages: The query answering problem. *Artificial Intelligence*, 193:87–128, December 2012.

[16] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. DL-Lite: Tractable Description Logics for Ontologies. volume 2, pages 602–607, January 2005.

[17] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tractable Reasoning and Efficient Query Answering in Description Logics: The DL-Lite Family. *Journal of Automated Reasoning*, 39(3):385–429, October 2007.

[18] A. Calì, G. Gottlob, and M. Kifer. Taming the Infinite Chase: Query Answering under Expressive Relational Constraints. *Journal of Artificial Intelligence Research*, 48:115–174, October 2013.

[19] Andrea Calì, Georg Gottlob, and Thomas Lukasiewicz. A general Datalog-based framework for tractable query answering over ontologies. *Journal of Web Semantics*, 14:57–83, July 2012.

[20] İsmail İlkan Ceylan, Adnan Darwiche, and Guy Van den Broeck. Open-world probabilistic databases: Semantics, algorithms, complexity. *Artificial Intelligence*, 295:103474, June 2021.

[21] Hyesang Chang, Lang Chen, Yuan Zhang, Ye Xie, Carlo de Los Angeles, Emma Adair, Gaston Zanitti, Demian Wassermann, Miriam Rosenberg-Lee, and Vinod Menon. Foundational Number Sense Training Gains Are Predicted by Hippocampal–Parietal Circuits. *Journal of Neuroscience*, 42(19):4000–4015, May 2022. Publisher: Society for Neuroscience Section: Research Articles.

[22] Kalina Christoff and John D. E. Gabrieli. The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, 28(2):168–186, June 2000.

[23] Nilesh Dalvi and Dan Suciu. The dichotomy of probabilistic inference for unions of conjunctive queries. *Journal of the ACM*, 59(6):30:1–30:87, January 2013.

[24] Evgeny Dantsin, Thomas Eiter, Georg Gottlob, and Andrei Voronkov. Complexity and expressive power of logic programming. *ACM Computing Surveys*, 33(3):374–425, September 2001.

[25] Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. A sulcal depth-based anatomical parcellation of the cerebral cortex. *Neuroimage*, 47, July 2009.

[26] Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, October 2010.

[27] Alin Deutsch, Alan Nash, and Jeff Remmel. The chase revisited. In *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '08*, page 149, Vancouver, Canada, 2008. ACM Press.

[28] Jérôme Dockès, Russell A Poldrack, Romain Primet, Hande Gözükan, Tal Yarkoni, Fabian Suchanek, Bertrand Thirion, and Gaël Varoquaux. NeuroQuery, comprehensive meta-analysis of human brain mapping. *eLife*, 9:e53385, March 2020.

[29] David A. Drachman. Do we have brain to spare? *Neurology*, 64(12):2004–2005, June 2005. Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Editorials.

[30] Simon B. Eickhoff, Angela R. Laird, Christian Grefkes, Ling E. Wang, Karl Zilles, and Peter T. Fox. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, 30(9):2907–2926, September 2009.

[31] Ronald Fagin, Phokion G. Kolaitis, Renée J. Miller, and Lucian Popa. Data exchange: semantics and query answering. *Theoretical Computer Science*, 336(1):89–124, May 2005.

[32] Bruce Fischl, André van der Kouwe, Christophe Destrieux, Eric Halgren, Florent Ségonne, David H. Salat, Evelina Busa, Larry J. Seidman, Jill Goldstein, David Kennedy, Verne Caviness, Nikos Makris, Bruce Rosen, and Anders M. Dale. Automatically parcellating the human cerebral cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, 14(1):11–22, January 2004.

[33] Peter T. Fox, Angela R. Laird, Sarabeth P. Fox, P. Mickle Fox, Angela M. Uecker, Michelle Crank, Sandra F. Koenig, and Jack L. Lancaster. BrainMap taxonomy of experimental design: description and evaluation. *Human Brain Mapping*, 25(1):185–198, May 2005.

[34] Gary H. Glover. Overview of Functional Magnetic Resonance Imaging. *Neurosurgery clinics of North America*, 22(2):133–139, April 2011.

[35] Krzysztof J. Gorgolewski, Gael Varoquaux, Gabriel Rivera, Yannick Schwarz, Satrajit S. Ghosh, Camille Maumet, Vanessa V. Sochat, Thomas E. Nichols, Russell A. Poldrack, Jean-Baptiste Poline, Tal Yarkoni, and Daniel S. Margulies. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, 9(8), April 2015. Publisher: Frontiers.

[36] Georg Gottlob, Giorgio Orsi, and Andreas Pieris. Query Rewriting and Optimization for Ontological Databases. *ACM Transactions on Database Systems*, 39(3):1–46, October 2014.

[37] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '07*, page 31, Beijing, China, 2007. ACM Press.

[38] D. O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press, April 2005. Google-Books-ID: ddB4AgAAQBAJ.

[39] Suzana Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3, 2009.

[40] Julien I.E. Hoffman. Meta-analysis. In *Biostatistics for Medical and Biomedical Practitioners*, pages 645–653. Elsevier, 2015.

[41] Hyeon-Ae Jeon and Angela D. Friederici. Two principles of organization in the prefrontal cortex are cognitive hierarchy and degree of automaticity. *Nature Communications*, 4(1):2041, October 2013.

[42] Abhay Jha and Dan Suciu. Probabilistic Databases with MarkoViews. *arXiv:1208.0079 [cs]*, July 2012. arXiv: 1208.0079.

[43] C. Keysers, V. Gazzola, and E. Wagenmakers. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, 2020.

[44] Etienne Koechlin, Chrystèle Ody, and Frédérique Kouneiher. The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*, 302(5648):1181–1185, November 2003.

[45] Tatsuya Kushida. Interlinking Ontology for Biological Concepts - Basic chemistry research field - Classes | NCBO BioPortal.

[46] Angela R Laird, Simon B Eickhoff, P Mickle Fox, Angela M Uecker, Kimberly L Ray, Juan J Saenz, D Reese McKay, Danilo Bzdok, Robert W Laird, Jennifer L Robinson, Jessica A Turner, Peter E Turkeltaub, Jack L Lancaster, and Peter T Fox. The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Research Notes*, 4:349, September 2011.

[47] Angela R. Laird, P. Mickle Fox, Cathy J. Price, David C. Glahn, Angela M. Uecker, Jack L. Lancaster, Peter E. Turkeltaub, Peter Kochunov, and Peter T. Fox. ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping*, 25(1):155–164, May 2005.

[48] Agatha Lenartowicz, Donald J. Kalar, Eliza Congdon, and Russell A. Poldrack. Towards an Ontology of Cognitive Control. *Topics in Cognitive Science*, 2(4):678–692, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-8765.2010.01100.x.

[49] Thomas Lukasiewicz and Umberto Straccia. Managing uncertainty and vagueness in description logics for the Semantic Web. *Journal of Web Semantics*, 6(4):291–308, November 2008.

[50] Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncavles, Anita Jwa, and Russell Poldrack. The OpenNeuro resource for sharing of neuroscience data. *eLife*, 10:e71774, October 2021. Publisher: eLife Sciences Publications, Ltd.

[51] M. Mesulam. From sensation to cognition. *Brain*, 121(6):1013–1052, June 1998.

[52] Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, David L. Thomas, Essa Yacoub, Junqian Xu, Andreas J. Bartsch, Saad Jbabdi, Stamatios N.

Sotiropoulos, Jesper L. R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W. Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M. Matthews, and Stephen M. Smith. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536, November 2016. Number: 11 Publisher: Nature Publishing Group.

[53] Veronika I. Müller, Edna C. Cieslik, Angela R. Laird, Peter T. Fox, Joaquim Radua, David Mataix-Cols, Christopher R. Tench, Tal Yarkoni, Thomas E. Nichols, Peter E. Turkeltaub, Tor D. Wager, and Simon B. Eickhoff. Ten simple rules for neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, 84:151–161, January 2018.

[54] Derek Evan Nee and Mark D'Esposito. The hierarchical organization of the lateral prefrontal cortex. *eLife*, 5:e12112, March 2016.

[55] S Ogawa, T M Lee, A R Kay, and D W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24):9868–9872, December 1990.

[56] Michael Petrides. Lateral prefrontal cortex: architectonic and functional organization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):781–795, April 2005.

[57] Russell A. Poldrack. Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10(2):59–63, February 2006.

[58] Russell A. Poldrack, Aniket Kittur, Donald Kalar, Eric Miller, Christian Seppa, Yolanda Gil, D. Stott Parker, Fred W. Sabb, and Robert M. Bilder. The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. *Frontiers in Neuroinformatics*, 5, 2011.

[59] Russell A. Poldrack and Tal Yarkoni. From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annual review of psychology*, 67:587–612, January 2016.

[60] Fabrizio Riguzzi. ALLPAD: approximate learning of logic programs with annotated disjunctions. *Machine Learning*, 70(2):207–223, March 2008.

[61] Fabrizio Riguzzi, Elena Bellodi, Evelina Lamma, and Riccardo Zese. Reasoning with Probabilistic Ontologies. page 7.

[62] Cornelius Rosse and Jose Mejino. The Foundational Model of Anatomy Ontology. *Anatomy Ontologies for Bioinformatics: Principles and Practice*, 6, January 2008.

[63] Sebastian Rudolph. Foundations of Description Logics. In Axel Polleres, Claudia d'Amato, Marcelo Arenas, Siegfried Handschuh, Paula Kroner, Sascha Ossowski, and Peter Patel-Schneider, editors, *Reasoning Web. Semantic Technologies for the Web of Data: 7th International Summer School 2011, Galway, Ireland, August 23-27, 2011, Tutorial Lectures*, Lecture Notes in Computer Science, pages 76–136. Springer, Berlin, Heidelberg, 2011.

[64] Gholamreza Salimi-Khorshidi, Stephen M. Smith, John R. Keltner, Tor D. Wager, and Thomas E. Nichols. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, 45(3):810–823, April 2009.

[65] Pantelis Samartsidis, Silvia Montagna, Timothy D. Johnson, and Thomas E. Nichols. The Coordinate-Based Meta-Analysis of Neuroimaging Data. *Statistical Science*, 32(4):580–599, November 2017.

[66] Sean, Bechhofer, Frank, van Harmelen, Jim, Hendler, Ian, Horrocks, Deborah L, McGuinness, Peter F, Patel-Schneider, and Lynn Andrea, Stein. OWL Web Ontology Language Reference.

[67] Terrence J Sejnowski, Patricia S Churchland, and J Anthony Movshon. Putting big data to good use in neuroscience. *Nature neuroscience*, 17(11):1440–1441, November 2014.

[68] Pierre Senellart. Provenance and Probabilities in Relational Databases: From Theory to Practice. *ACM SIGMOD Record*, 46, December 2017.

[69] Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. Probabilistic Databases. *Synthesis Lectures on Data Management*, 3(2):1–180, May 2011.

[70] Peter E. Turkeltaub, Guinevere F. Eden, Karen M. Jones, and Thomas A. Zeffiro. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *NeuroImage*, 16(3 Pt 1):765–780, July 2002.

[71] Jessica A. Turner and Angela R. Laird. The Cognitive Paradigm Ontology: Design and Application. *Neuroinformatics*, 10(1):57–66, January 2012.

[72] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E. J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79, October 2013.

[73] Joost Vennekens, Marc Denecker, and Maurice Bruynooghe. CP-logic: A language of causal probabilistic events and its relation to logic programming. *Theory and Practice of Logic Programming*, 9(3):245–308, May 2009. Publisher: Cambridge University Press.

[74] Jonas Vlasselaer, Angelika Kimmig, Anton Dries, Wannes Meert, and Luc De Raedt. Knowledge Compilation and Weighted Model Counting for Inference in Probabilistic Logic Programs. In *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence Beyond NP*, page 6, 2016.

[75] Jonas Vlasselaer, Joris Renkens, Guy Van den Broeck, and Luc De Raedt. Compiling probabilistic logic programs into sentential decision diagrams. In *Proceedings Workshop on Probabilistic Logic Programming (PLP)*, pages 1–10, July 2014.

[76] Tor D. Wager, John Jonides, and Susan Reading. Neuroimaging studies of shifting attention: a meta-analysis. *NeuroImage*, 22(4):1679–1693, August 2004.

[77] Tor D. Wager, Martin Lindquist, and Lauren Kaplan. Meta-analysis of functional neuroimaging data: current and future directions. *Social Cognitive and Affective Neuroscience*, 2(2):150–158, June 2007.

[78] Tal Yarkoni. Neurosynth core tools v0.3.1, May 2014.

[79] Tal Yarkoni, Russell A. Poldrack, Thomas E. Nichols, David C. Van Essen, and Tor D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670, August 2011.

[80] Gaston E. Zanitti, Yamil Soto, Valentin Iovene, Maria Vanina Martinez, Ricardo O. Rodriguez, Gerardo I. Simari, and Demian Wassermann. Scalable Query Answering Under Uncertainty to Neuroscientific Ontological Knowledge: The NeuroLang Approach. *Neuroinformatics*, November 2022.