7-10-2018 10:00 AM

Finding Nonlinear Relationships in Functional Magnetic Resonance Imaging Data with Genetic Programming

James Hughes, *The University of Western Ontario*

Supervisor: Mark Daley, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science

Abstract

The human brain is a complex, nonlinear dynamic chaotic system that is poorly understood. When faced with these difficult to understand systems, it is common to observe the system and develop models such that the underlying system might be deciphered. When observing neurological activity within the brain with functional magnetic resonance imaging (fMRI), it is common to develop linear models of functional connectivity; however, these models are incapable of describing the nonlinearities we know to exist within the system.

A genetic programming (GP) system was developed to perform symbolic regression on recorded fMRI data. Symbolic regression makes fewer assumptions than traditional linear tools and can describe nonlinearities within the system. Although GP is a powerful form of machine learning that has many drawbacks (computational cost, overfitting, stochastic), it may provide new insights into the underlying system being studied.

The contents of this thesis are presented in an integrated article format. For all articles, data from the Human Connectome Project were used.

In the first article, nonlinear models for 507 subjects performing a motor task were created. These nonlinear models generated by GP contained fewer ROI than what would be found with traditional, linear tools. It was found that the generated nonlinear models would not fit the data as well as the linear models; however, when compared to linear models containing a similar number of ROI, the nonlinear models performed better.

Ten subjects performing 7 tasks were studied in article two. After improvements to the GP system, the generated nonlinear models outperformed the linear models in many cases and were never significantly worse than the linear models.

Forty subjects performing 7 tasks were studied in article three. Newly generated nonlinear models were applied to unseen data from the same subject performing the same task (intra-subject generalization) and many nonlinear models generalized to unseen data better than the linear models. The nonlinear models were applied to unseen data from other subjects performing the same task (intersubject generalization) and were not capable of generalizing as well as the linear.

Keywords: Functional Magnetic Resonance Imaging, Nonlinear Relationships, Functional Connectivity, Symbolic Regression, Genetic Programming, Timeseries

Co-Authorship Statement

I would like to acknowledge Mark Daley, my supervisor as a coauthor on the integrated articles.

James Hughes wrote the genetic programming system, designed the experiments, executed the experiments, analyzed the results, and wrote the articles.

Acknowledgements

First, I would like to thank Mark Daley, my supervisor, for his leadership and support throughout this journey. His careful and meticulously well-crafted advice and guidance was fundamentally necessary for my success in both my graduate work, and in life.

I would like to thank Ethan Jackson, whom I have been studying computer science with for many years. Not only has his support helped me through my PhD, but also my Masters and Undergraduate degree. I'm not entirely sure I would have gotten through third year algorithms without our long study sessions in MC D205.

I would also like to thank the current and past members of the lab. I've always enjoyed talking about the research and sharing ideas. On multiple occasions, many critical ideas were brought up that greatly improved my work.

I thank the people within the Computer Science department here at the University of Western Ontario. Whether a student, teacher, peer, or administrator, many have played a large role throughout the past 4 years. I really do feel that I have grown a lot during this time, and it was those within my direct community that had the greatest impact.

Additionally, I thank those within the computer science department at Brock University, where I completed my Undergraduate and Masters degrees. It was the faculty members and friends I made there that sparked my excitement for computer science.

I would like to thank my family. Not only have they supported me throughout my studies, they have enabled me to be where I am today. They enabled me to enroll in a PhD, Masters, and Undergraduate degree. They supported me throughout high school. They instilled upon me my work ethic and curiosity. Ultimately, they made me into who I am today.

Above all, I thank Matea Drljepan. We have been pushing and supporting each other for the sake of our future together since 2008. Whether it was specific edit suggestions, or higher-level motivation, she has always patiently helped.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

The human brain is a nonlinear computational system. Although neuroscience literature explicitly acknowledges this [15, 18, 26, 38, 36, 122], it is commonly deemphasized or ignored, especially when working with functional magnetic resonance imaging (fMRI) data [18, 84]. When studying fMRI timeseries data to find *functional relationships* within the brain, it is common to exclusively use linear tools, such as Pearson product-moment correlation coefficient or the general linear model (GLM). However, these methods are not capable of describing what we know to be a nonlinear system as it lacks the power to truly model the underlying processes.

In spite of this inability to truly model the system, neuroscientific studies make meaningful contributions to the field with these linear methods [18]. However, one must wonder if a more expressive, nonlinear method capable of describing the underlying system would increase the analytical power and significance of results. It is in no way surprising that the nonlinear relationships are ignored and more complex tools are not used since it is exceptionally difficult to find nonlinear relationships, especially when working with large amounts of noisy high-dimensional data from a nonlinear, dynamic complex system.

1.1 fMRI Data

Magnetic resonance imaging (MRI) scanners harness magnetic fields and electromagnetic energy in a controlled way to capture localized information about physical properties of tissue within the brain[1]. More precisely, they capture information about the *spin-relaxation* properties of particles within the brain (this idea is discussed further in Chapter 3). MRI scanners capture the localized information which can then be represented in the form of voxels — three-dimensional analogues to two-dimensional pixels. Ultimately, the whole brain can be represented as a three-dimensional structure made up of voxels to give a static view of the underlying three-dimensional anatomy.

Functional MRI records the *blood oxygen level dependent* (BOLD) signal — a measurement of the relative oxygenation level of blood within tissue — which is used as a proxy for brain activation. These relative blood oxygen level variations occur since neurons do not store their own energy, and after activation, the vascular system must replenish the resources to the cerebral tissue. This process is called the *hemodynamic response* (HDR) and is a consequence

[1] MRI technology is not restricted to brain imaging, however neuroimaging is the focus of this thesis.

of metabolism. Although the BOLD signal is not actually brain activation, it can be used as a proxy and has been shown to strongly correlate with localized neural activity [95, 85, 84, 51].

The fMRI technology still captures images of the three-dimensional structure, however the information within the voxels is the BOLD signal.

Unlike MRI, which captures the three-dimensional anatomical information, fMRI captures the spatially localized BOLD signal from the three-dimensional brain. Since the moment-to-moment changes in neural activity are of interest, the scanner will take many three-dimensional snapshots of the brain over time such that the changes in the BOLD signal can be observed. The data being recorded is four-dimensional — the three-dimensional physical brain, over time (the additional dimension).

Often, subjects will be placed within an fMRI scanner and will be given some task, such as viewing images, playing a game, or moving a body part. By time-locking the task onsets with the observed BOLD signal, researchers try to determine which areas of the brain (voxels, or perhaps larger regions of interest (ROI)) are functionally related to the task being performed. This usage is sometimes called *task-based* fMRI analysis.

Sometimes subjects are placed within a scanner and are instructed to perform no task at all. In this scenario, the idea is to observe the spontaneous changes in the BOLD signal during rest instead of comparing the signal to what is expected to be seen when some task is being performed. This approach is called *resting-state* fMRI.

1.1.1 Graph Interpretation

Resting-state fMRI data is commonly used to develop functional connectivity models of the brain: if the BOLD signal within two areas of the brain (voxels, or ROI) appear to be moving together similarly in time, then they are said to be functionally connected. There are many ways one could measure area similarity over time, but the Pearson correlation coefficient is typically used to infer the functional connections (although in reality, all that can be said is that the two areas are linearly correlated).

By doing this, we can create a simple graph representation of the complex four-dimensional object. A simple graph is a collection of vertices/nodes connected together by edges. A vertex represents some entity, and an edge between two vertices represents some relationship between the entities. In this case, we treat the areas of the brain as vertices, and connect the vertices with an edge if the linear correlation score is above some predefined threshold. In this example, Pearson correlation was used to infer connectivities, however this is by no means a requirement.

These graphs provide static views of the synchronization of areas of the brain and greatly simplify the four-dimensional data. These models enable researchers to study the data in new ways; there are many well-defined graph theory metrics that are now available to the researchers [113]. For example, the topology of the graph can be analysed to study clinical questions, like if there are topological differences/similarities between individuals with certain neurological disorders [88, 17], or between adult and adolescent brain networks [31].

1.1.2 Human Connectome Project Data

The work is focused on studying task-based fMRI timeseries data to find nonlinear functional connectivities between ROI within the brain. Data is fit to a seed ROI and not an expected

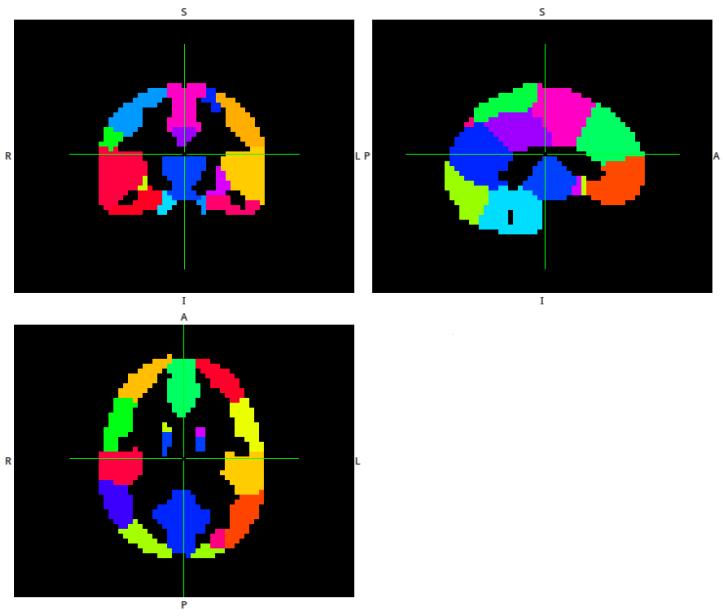


Figure 1.1: A snapshot of a brain when segmented into the 30 ROIs as seen in FSL view. Each colour represents a different region.

HDR.

The data selected for this analysis was obtained from the Human Connectome Project, WU-Minn Consortium, which can be found at `http://www.humanconnectome.org/`. The Human Connectome Project has an open database of a large collection of neuroimaging data, and as of April 2018, the database contains structural MRI, resting-state fMRI, diffusion imaging, and task-based fMRI data for roughly 1200 subjects, and Magnetoencephalography (MEG) data for resting-state and tasks on a subset of the participants.

The task-based fMRI data available from the Human Connectome Project include: Emotion, Gambling, Language, Motor, Relational, Social, and Working Memory. The actual tasks and number of subjects used varied throughout the project.

The fMRI timeseries data was segmented into meaningful ROIs (refer to Figure 1.1) with Craddock et al.'s *spatially constrained parcellation* [24]. Each voxel's activation within each ROI was averaged to determine the ROI's mean activation.

Although the fMRI timeseries data is a four-dimensional object (three-dimensional snapshots of a brain over time), it can easily be represented as a two-dimensional matrix of voxels over time if the three-dimensional physical space of the brain is flattened into one long vector, and time is left as the other dimension. Each entry in the matrix corresponds to the BOLD signal of a single voxel at a particular time point. Ultimately, the data can cleanly be represented in tabular format.

1.2 Neuroscientific Motivation

Neuroscientists will generate models of the brain to develop a better understanding of the underlying system. Having a high quality model of the brain can allow us to study the model

itself to discover properties about the complex system. For example, if we are interested in which regions of the brain are functionally connected, we may record resting-state or task-based fMRI data, fit a mathematical model to the data, and from the model determine which regions of the brain are related to one another, and in which way.

Typically the graphical models of functional connections are generated from resting-state fMRI data, however task-based fMRI data can also be used. By doing so, we can develop these functional connectivity models of the brain during certain tasks. Additionally, there is no need to restrict the model development to linear correlations.

For example, if we are interested in developing a model of functional connectivity, then we may perform some thresholding on the recorded data, and then develop a linear model of the data with the GLM. There are a couple of ways this could be done depending on the question being asked.

If we wanted to know how a given ROI X is functionally connected to all other ROIs, then we would calculate the Pearson product-moment correlation coefficients between the ROIs, perform some correction for multiple comparisons (typically *false discovery rate* (FDR) or *Bonferroni correction* (BC)), and remove statistically unrelated ROIs. Finally, the remaining ROIs are regressed to our ROI X and the beta weights (coefficients) can be used to indicate relatedness during the task.

This, along with the simpler linear correlation strategy described above, assumes that the system is linear, however we know that the human brain is a nonlinear system. This strategy makes many additional assumptions, including: the ROIs are fixed values as opposed to random variables, constant variance in the data, and the errors are independent.

Statistically unrelated ROIs are eliminated with thresholding to ensure that only meaningful ROIs are included in the resulting model; however, what does it mean for an ROI to be meaningfully related? The brain is a connected system being recorded at the same time under the same circumstances, and unsurprisingly many ROIs end up being highly correlated. After thresholding, a large number of ROIs will typically be left as *meaningful* — sometimes even all. Perhaps the whole brain is involved in the function of the task be studied, but this would seem unlikely.

Despite the assumptions described above, there are many reasons to prefer the traditional, simpler linear tools. Linear models are easy to generate, the tool is well understood, and the models are easy to interpret. More complex methodologies are susceptible to overfitting, the methods are harder to understand, the resulting models are difficult to interpret, and they typically have a much greater computational cost. However, despite these drawbacks, using a more complex method actually capable of describing the underlying nonlinear system may allow us to eliminate assumptions and develop a more accurate and descriptive model of the functional connectivities within brain.

1.3 Methods

Observing processes, developing models from data, and deriving natural rules, laws, and formalisms about a system is an intractable task that is difficult to automate. However, a promising approach is *genetic programming* (GP) [13, 103, 106]. GP is an optimization strategy based on the natural process of evolution that stochastically and iteratively writes its own programs

to learn how to solve a given problem [69]. GP is used in this work to automate the process of finding minimal and interpretable network relationships in a system for which we can only observe a recorded timeseries from the system's network's nodes. Namely, GP is used to find nonlinear functional relationships within fMRI timeseries data with symbolic regression.

Symbolic regression is a type of regression analysis that, in addition to parameter optimization, searches for model structure by performing feature selection and exploring the space of mathematical expressions. A GP system was developed for this work that was specifically designed for symbolic regression. The GP implementation was based on Schmidt et al.'s work, and incorporates improvements to increase performance [106]. Noteworthy improvements include a distributed population/island model, an acyclic graph representation [102], and fitness predictors [104, 105]. A summary of the method's improvements is provided here, but more details on evolutionary computation and GP can be found in Chapter 2.

Typical GP systems employ a tree based representation [76], however many popular non-tree based representations exist. In this work, an *acyclic graph representation* is used. The implemented representation has a lightweight array based encoding that avoids bloat, scales well, and can reuse subexpressions [102].

Computational costs of the evolutionary search is greatly reduced with the use of *fitness predictors*, which approximates the local search gradient [104, 105]. The high level idea is to evaluate each candidate solution on a small, but representative subset of the data being fit to. The subset of data is always changing such that it contains data points the current candidate solutions do not fit well; it focuses the search on areas of the space that need the most improvement. Fitness predictors were shown to lower computational cost, reduces overfitting, and improves results [105].

For much of the work, symbolic regression was used to develop nonlinear models of the brain, which can be interpreted as graph/network models of nonlinear functional relationships. This nonlinear regression is capable of describing the actual nonlinearities that must exist within the underlying system; nonlinear regression is strictly more powerful than linear regression in its descriptive power. Symbolic regression also performs feature selection, eliminating the need to manually perform thresholding.

These nonlinear models were compared to linear models developed with typical methods employed within the neuroscientific literature (GLM and the Pearson product-moment coefficient). A description of the methods used to develop the linear models are discussed within the integrated articles found in Chapters 4, 5, and 6. Each of the integrated articles found in Chapters 4, 5, and 6 also provide a description of the GP system and a summary of the system settings used for each project. Appendix B includes technical details about the GP system implemented (version number, number of classes, lines of code).

1.4 Nonlinear Analysis of fMRI data

Although it is overwhelmingly common to use linear tools when studying fMRI data, nonlinear tools have been used in some work. Friston et al. studied nonlinear responses in the BOLD signal with *Volterra series expansion* [36, 37]. Kruggel et al. used a nonlinear regression to study the timeseries of the BOLD signal to relate the dependency between the expected HDR shape and stimulus [77]. Friston et al. used *Dynamic Causal Modelling* to describe functional

connectivities between neuronal regions of the brain [35]. Zhang et al. used a semi-parametric model built around Volterra series to characterize BOLD signal and found deviations from the linear models and showed that their approach outperformed many existing methods [126]. *Symbolic regression*, a type of regression analysis, was used to describe nonlinear functional connectivities between known networks in *resting-state* fMRI data [2]. These works are discussed in more detail in Chapter 3.

The symbolic regression work done by Allgaier et al. in [2] is the most relevant to the work in this thesis as it also develops network models from nonlinear relationships found within fMRI data. Their work used resting-state data and focused on areas of the brain already known to exist within functional networks of interest. The work in this thesis searches for nonlinear relationships within task-based fMRI data within ROI throughout the whole brain. Additionally, the work contained within this thesis evaluates the models in different ways.

1.5 Contribution

A GP system incorporating a number of modular improvements was implemented and made publicly available. This GP system was developed for the purpose of finding nonlinear functional connectivities within fMRI data, however it is a specialized system for symbolic regression in general. Given the high dimensionality of the search space, the improvements incorporated into the system were required in order for the evolutionary search to complete in a reasonable amount of time. After the initial development of the system, numerous revisions were done over the past three years. Development of the GP system will continue for the foreseeable future.

While working within the limitations of the real task-based fMRI data available, a graph-based interpretation of the timeseries data was developed. Functional connectivities were modelled with a GP system specifically created for this project. These graph-based models of nonlinear relationships were found to be much more succinct (fewer relationships) when compared to models developed with conventional linear tools.

This thesis demonstrates a methodology that will enable the longer term goal of finding meaningful nonlinear functional relationships within fMRI data, interpreting the meaning of these relationships, and making contributions to the neuroscientific literature.

Three articles are presented in this work and are the natural progression of the project.

The first article presents the proof of concept by applying the GP system to data from a single task and exploring the differences between nonlinear and linear models of a network interpretation of fMRI data. In this article, data from 507 subjects were studied. It found the nonlinear models to contain fewer ROI than the linear models developed with typical linear methods, and the nonlinear models' ROIs were almost always subsets of the linear models' ROIs. It also found that the GP generated nonlinear models were not capable of fitting the recorded fMRI signal as well as the linear models.

The second article expands on the first by improving upon the GP system, broadening the analysis to additional tasks, and exploring the differences between the linear and nonlinear models in greater detail. In this article, data from 10 subjects were studied. After the improvements in the GP system, the nonlinear models were larger than those found in the first article, however they still had fewer ROI than the linear. In many cases, the nonlinear models were

now able to fit their data better than the linear models. There were many similarities in ROIs found between the model types, but the nonlinear models contained functional connectivities not found with linear tools.

The third article incorporates more subjects and a deeper analysis into the generalizability of the models to unseen data. In this article, data from 40 subjects were studied. Again, after improvements, the nonlinear models, on average, grew in size by a small amount over the previous work's nonlinear models, but still contained fewer ROI than the linear models. However, LASSO regression was included in the comparison and the linear models created with LASSO regularization were of comparable size to the nonlinear. The nonlinear models fit their data better than the linear models, and their intersubject and intrasubject generalizability was explored to determine if the nonlinear models were effective, and not overfitting.

Ultimately, the nonlinear models fit data better than traditional linear models, and were capable of generalizing to unseen data; however, the author very explicitly and clearly acknowledges the statistical biases and limitations of the current conclusions in Chapter 7. Methods for overcoming the limitations are discussed in Section 7.3 where future directions are presented.

1.6 Thesis Format

This thesis is presented in the *integrated-article* format. Chapter 2 provides a background and literature review on evolutionary computation and GP. Chapter 3 provides background and a brief literature review for the fMRI data and related neuroscientific works. Chapters 4, 5, and 6 are integrated articles from works completed during the duration of the author's PhD and are the natural progression of the overall project. Each of these chapters provide motivation, a small literature review, descriptions of the data used, and GP system implementation and settings details. Chapter 7 provides a discussion and concludes the work and includes possible future directions.

Appendix A contains the published extended abstract for the work in Chapter 4. Appendix B includes specific details on the GP implementation and execution.

Chapter 2

Evolutionary Computation and Literature Review

This chapter provides background information for *evolutionary computation* (EC), *genetic algorithms* (GA), and GP, along with a brief literature review of some algorithmic enhancements. This chapter is derived from the *Topics Survey/Proposal* written in December 2015 and presented June 2017. Brief details on the current implementation of the GP system can be found in Appendix B.

2.1 Genetic Algorithms

Evolutionary algorithms (EAs), a subcategory of evolutionary computation, are a population based *metaheuristic* — a high level algorithm designed to guide a problem space exploration — which search by simulating the process of biological evolution through a series of nature inspired operations: mutations, sexual reproduction, recombination, and natural selection.

Evolutionary algorithms developed over time with contributions by many researchers from around the world. These contributions began with simulation of artificial selection by many researchers, including Nils Barricelli in the late 1950s [10] and Alex Fraser in the 1960s [34]. Alan Turing even highlighted the parallels between a stochastic hypothetical "learning machine" and the natural process of evolution [120]. These ideas later developed into well defined evolutionary algorithms we use today, such as *Evolutionary Strategies*, *Evolutionary Programming*, and *Genetic Algorithms* (GAs) — a popular branch of EAs developed by John Holland in the mid 1970s [47]. These algorithms can typically be easily broken down into a few simple operations: *initialization*, *fitness evaluation*, *selection*, *genetic operators*, and *termination*.

Initialization involves generating a starting *population* (collection) of *chromosomes*, sometimes referred to as *candidate solutions*; a collection of potential solutions to a given problem. These candidate may be randomly generated, or seeded into the algorithm.

Fitness evaluation is the process of calculating how effective a given chromosome is at solving the problem the GA is being applied to. For example, if the GA was being applied to the travelling salesman problem[1], then the fitness could be the total Euclidean distance defined

[1] A common problem. Given a weighted connected graph, the goal is to visit all vertices while minimizing the total weight along all used edges.

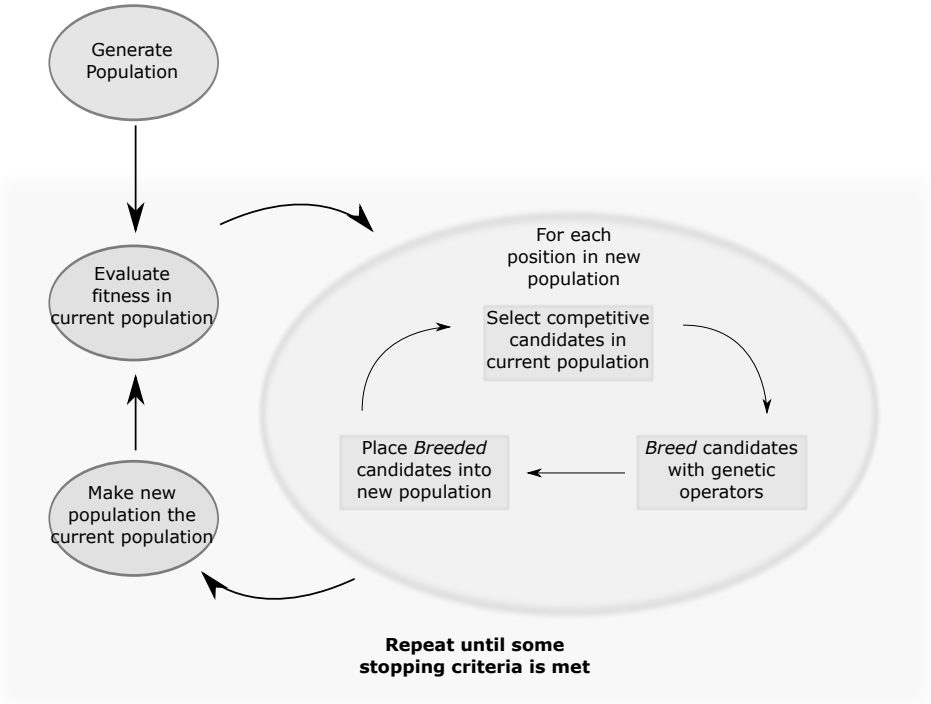


Figure 2.1: Example, high level description of a typical evolutionary algorithm.

by a chromosome's representation of a city ordering. In this case, the smaller the distance travelled, the *fitter* the chromosome. Calculating the fitness of the whole population is typically the bottleneck of the algorithm.

Selection is a method used to decide which chromosomes are to be propagated into the next *generation* (next round of evaluations, selection, and genetic operators). There are multiple ways one could go about doing this, however the main objective is to select some relatively *fit* chromosomes. One should avoid simply selecting the most fit individuals as this typically sends the search into a local optimum and the algorithm will converge too quickly. To discourage the GA from converging too quickly, one wants to encourage a good level of *genetic diversity* — a population containing chromosomes somewhat distinct from one another.

Genetic operators are the methods applied to selected chromosomes when propagated into the next generation's population. There are typically two genetic operators: *crossover* and *mutation*. Crossover is designed to be somewhat analogous to processes which occur during sexual reproduction; two reasonably fit chromosomes will *breed* and produce two offspring which are somewhat similar to both parent chromosomes. Mutations, unlike crossover, occur on a single chromosome at a time and will alter the chromosome slightly in some way.

This process repeats many times until some termination criteria is met. This could be after some number of generations, after the algorithm has converged, or after some fitness value is obtained. Figure 2.1 depicts the execution flow of a typical GA/EA.

Although this search is stochastic, it can produce high quality results to computationally intractable problems with minimal application/domain knowledge [44, 27, 41, 39]. EAs, being a type of *computational intelligence*, are ideal for problems which include uncertainty, are

potentially stochastic in nature, and have no other reasonable means of computational based problem solving. GAs, and its variations, have been applied to many applications. Engineering and design has been accomplished with the design of buildings to minimize energy use [119], structural design of commercial buildings [90], synthesis of the antennas for NASA's Space Technology 5 (ST5) mission [48, 86], NASA's deep space communication networks [45], electrical circuit design [73], robot programming [75], and robot design/manufacturing or robotic lifeforms [83, 127].

2.1.1 Modular Enhancements

One of the major advantages of genetic algorithms is the modular nature of the methods. It is easy to alter and add operators to the algorithm to better fit the application. Alterations and additions are frequently created and some have since become standard in the literature. Below are a collection of typical enhancements incorporated in genetic algorithms.

Representation

Representations have become a widely studied area within the field as there are numerous ways to represent any given problem and some representations may have some inherit advantages over others.

Classically, a genetic algorithm's representation of a candidate solution would be a string of 0s and 1s which would represent something meaningful with respect to the problem space. These 0s and 1s would be the *genotype* and would require some sort of translation into a *phenotype* (something to be evaluated by the *fitness function* — the function which calculates the candidate solution's fitness). With this binary representation John Holland introduced *Holland's schema theorem* for exponential increases in fitness over successive generations [47]. This theorem essentially demonstrates the power and usefulness of GAs.

This *indirect representation* is by no means a requirement. Over time it became easier to implement other representations and more *direct representations* — where no translation is required — became feasible. For example, when studying the travelling salesman problem, instead of a binary string, one could implement an ordered list of cities directly representing the order to visit each city. Although these more complex representations do not strictly align to Holland's schema theorem, they were shown early to be effective [42, 63].

Selection

Multiple selection algorithms for genetic algorithms exist and new ones are always being developed.

One could always select the best chromosomes to breed and populate the next generation, however this approach tends to cause the GA to converge quickly into a local optimum. Alternatively, one might implement a completely random selection, although this would eliminate the high *selection pressure* of the strong (relatively fitter) candidate solutions.

Typically an effective selection method would encourage the fittest candidate solutions to propagate while still avoiding early convergence.

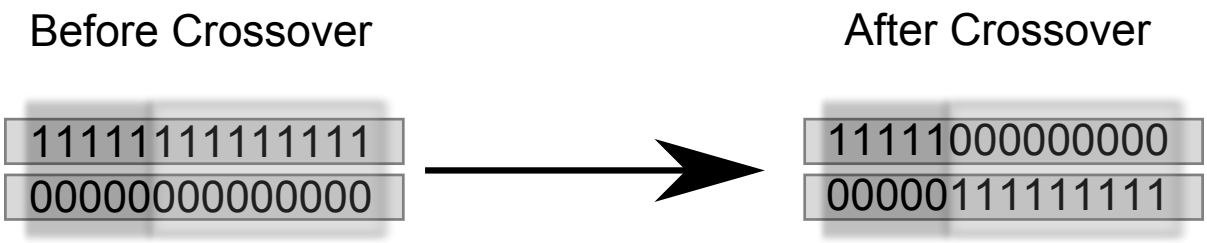


Figure 2.2: One point crossover example with a simple binary value representation. All values within the darker emphasised area are swapped between the two chromosomes. This figure also shows a simple binary representation.

Popular selection methods include *proportional selection* [47], *tournament selection* [43], and *linear ranking* [7]. These, and more, have well studied effects on selection pressure [6].

Elitism

Despite the fact that selection algorithms make an effort to avoid always selecting the best chromosomes, it has become standard to propagate the most fit chromosome (sometimes more than one) into the next generation to preserve the best known solution. The best known chromosome will always be monotonically non-decreasing over time; it cannot be destroyed by stochastic changes [8].

Genetic Operators

The genetic operators are how the GA explores new areas of the search space and exploits already known highly fit chromosomes. These operators area easily changed and tuned to appropriately align to the specific problem the GA is being applied to. There are a number common techniques for both operators, however, new techniques are always being developed to exploit the intricacies of specific problem spaces.

Common crossover techniques include *One-Point Crossover* (depicted in Figure 2.2) (every element after an index is swapped between two parent chromosomes), *Two-point crossover* (every element between two indices are swapped between two parent chromosomes), and *Uniform Crossover* (some number of indices are selected and all elements at these indices are swapped between the parent chromosomes). Some techniques are more destructive than others, and they all have their strengths and weaknesses.

Mutation only occurs on one chromosome at a time. Common mutation techniques include a single/multi-point mutation (select random indices and replace them with new values from the set of available values), exchange mutation (swap two or more elements), and updates (altering real value elements with an increment/decrement).

More complex genetic operators may become necessary for certain problems. For example, some problems may require variable length chromosomes or in the case of the travelling salesman problem, the uniqueness of elements may need to be preserved. In these cases, special genetic operators will be required to accommodate these requirements. Many of these unique operators are reviewed and studied in [80].

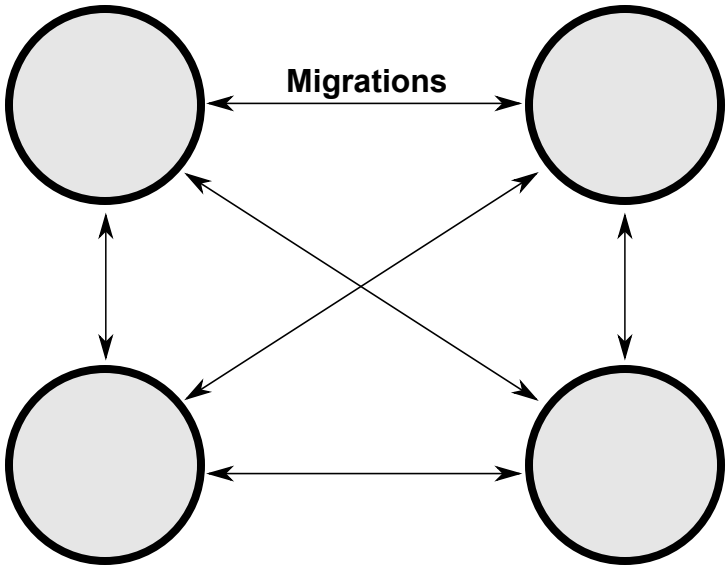


Figure 2.3: In this example, each circle represents a separate population (4 in this case) which evolve independently from one another. After some number of generations, chromosomes from each population have the opportunity to *migrate* to other populations. This particular figure shows allowable migrations between all populations, however this is not a requirement.

Distributed Populations

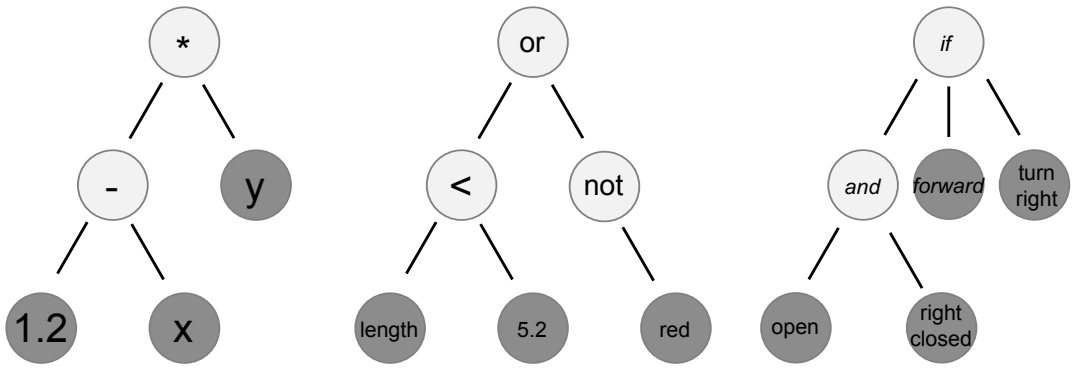
Distributing the search by dividing a population into multiple sub-populations has become popular. This method attempts to simulate *punctuated equilibria* and *allopatric speciation*, or simply, encouraging genetic diversity over the *whole* population by allowing the *sub*-populations to traverse the search space along their own trajectories. This idea is sometimes called the *island model*.

The general idea is to break the population down into sub-populations and execute a GA on each of the sub-populations with periodic information transfer between them. Multiple versions of these distributed systems exist. Figure 2.3 demonstrates a case with four sub-populations that is completely connected; information can be transferred, or *migrated*, between any of these sub-populations.

The idea of distributing the search has existed for some time. Booker notes in his Doctoral Dissertation [14]:

> Two separate populations are used rather than one large one so that the learning algorithms can benefit from having classifiers already separated into gross functional "niches."

However these distributed searches became popular later with various works [23, 94, 97, 118] including [116] which demonstrates that these distributed GAs typically have faster evaluation, faster convergence, and better results over a single population alternative.



(a) Function. Represents a mathematical expression, namely, this represents $(1.2 + v1) * v2$.

(b) Binary classification Program. Represents some binary classification, namely, (LENGTH < 5.2) | *not* RED

(c) Represents a program, namely, *if* (OPEN & RIGHT CLOSER) *then* FORWARD *else* TURN RIGHT.

Figure 2.4: Three example *programs* represented in a tree-structure.

Fitness Approximation

The fitness evaluation of each chromosome is, in most cases, the most computationally expensive portion of the evolutionary search.

It can be advantageous to use a method which can *quickly* approximate the fitness of a chromosome or a collection of chromosomes. There are a number of approaches to this within the literature, and many are not restricted to just evolutionary searchers. Some popular approaches include sub-sampling the data, *fitness inheritance* (inherited fitness values from parent chromosomes) [107], *fitness imitation* (cluster chromosomes and evaluate only representative chromosomes) [68, 66], and *partial evaluation* (a combination of fitness inheritance and imitation) [99]. These, and other techniques are reviewed in [66].

As stated in [105], these techniques are beneficial as they can reduce the complexity of the problem, eliminate the need for an explicit fitness function (some problems don't have an explicit evaluation method), reduces the concerns of a noisy fitness function, smooths the fitness landscape (reduces the number of local optima), and promotes genetic diversity.

2.2 Genetic Programming

As interesting and creative representations for GAs became more popular, a tree structure representing computer programs was implemented [25]. John R. Koza expanded upon the idea of using a tree structure representation and ultimately developed the field of *Genetic Programming* (GP); using evolutionary search to explore the space of functions/computer programs [69, 70, 71, 72, 75, 73, 74].

Figure 2.4 demonstrates three small functions/programs which could easily be represented with a tree structure. All leaf nodes (dark nodes) are terminals and non-leaf nodes (light nodes) are operators of some kind. The fitness evaluation method in GP would be some measure of how effective the function/program is at solving the given problem. For example, if the function described in Figure 2.4a was being used to perform some mathematical regression/modelling

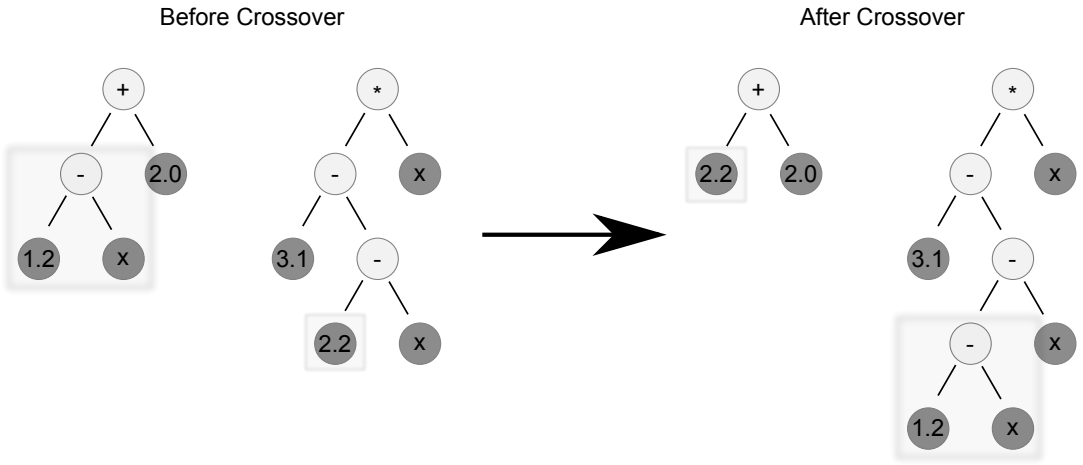


Figure 2.5: Example of a one point crossover operation between two tree-structure chromosomes.

— *symbolic regression*[2] — then the fitness may be the *mean squared error* calculated when applying the expression to data. If Figure 2.4b was some decision tree for binary classification, the fitness could be the percent accuracy. If the program in Figure 2.4c is describing how a robot should manoeuvre to solve a maze, then the fitness may how close the robot got to the exit [3].

Special genetic operators are also required for these tree structures. A common approach would be to select sub-trees from each chromosome and swap them. Figure 2.5 demonstrates this sub-tree exchange. This is a very similar technique to one-point crossover, a common crossover technique with basic GAs. Mutation could be a single point mutation (select a node and change it), or an exchange mutation (select two nodes and swap them within the same chromosome). Similar to GAs, new genetic operators are developed for GP constantly.

The operators and operands that are used in the representation are defined by a *language* (basis functions). Languages are selected for the problem being solved. If one was performing symbolic regression then an appropriate language may be $+$, $-$, $*$, $/$, *log*, *exp*, *variables*, and *floating point number constants*. If a decision tree was to be developed, a more appropriate language may be the logical operators, real numbers, and the variables. Note that in the latter example there are multiple types (Boolean and Numerals). A GP system with multiple types (*typed GP*) has additional requirements on the genetic operators as they need to preserve node return types.

GP is in no way limited to a tree structure. *Linear Genetic Programming* is an alternative which treats the representation as a sequence of instructions from an imperative, or machine language. This differs from the more functional implementation of the tree structure. Other noteworthy representations exist, including graph based representations (more on this in Section 2.2.1).

Notable early applications of GP are reviewed in [71, 72, 74], and include quantum com-

[2] A type of regression analysis often performed with GP that searches for the whole model (operators, coefficients, structure, feature selection) as opposed to just coefficients.

[3] There is no suggestion that these would be effective candidate solutions.

puting [9, 112, 109, 110, 111], robot programming [3, 87], bioinformatics [71], engineering, and circuit design [71].

2.2.1 Acyclic Graph Representation

An *acyclic graph representation* for symbolic regression was studied and compared to the traditional tree structure by Schmidt et al. in [102]. Other graph encodings (either explicit or implicit) for GP have existed for some time [98]. One of the most popular is *Cartesian* GP [93, 91, 92], however, Schmidt et al.'s gives some unique advantages (although this is entirely implementation dependent).

Figure 2.6 presents a comparison of a tree representation and an acyclic graph representation. These structures represents the following equation: $(1.23 - x) + \sin((1.23 - x) \cdot y \cdot e^x)$, where x and y are variables.

This representation, when compared to the tree representation, scales better, has a lightweight array encoding, and avoids bloat — the tendency of evolved programs to grow arbitrarily large without significant improvement in fitness [98]. Additionally, it can easily reuse possibly important sub-expressions and can maintain vestigial information within the encoding which may resurface effectively in future generations.

It was also noticed that this acyclic graph representation converges slower (although this may be considered an advantage) and is susceptible to deleterious crossovers [102]. However, their results strongly demonstrate the benefits of this representation.

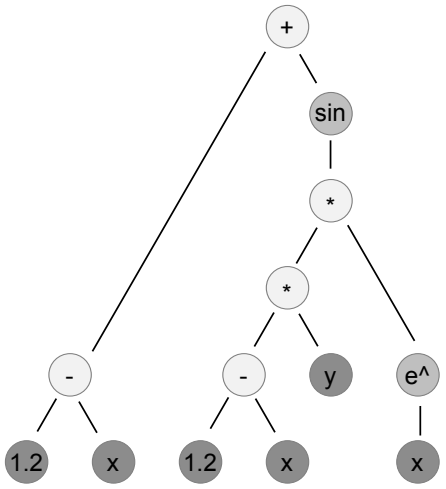
2.2.2 Fitness Predictors

Fitness Predictors, a fitness approximation approach, were studied by Schmidt el al. in [104, 105] and it was demonstrated that they can reduce computational cost by approximating the local search gradient.

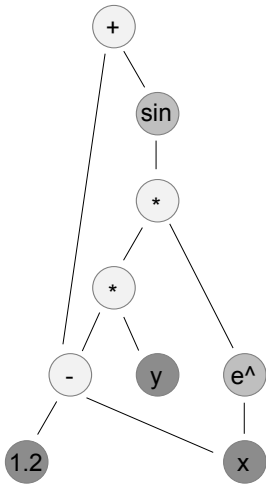
The candidate solutions are evaluated with an adaptive subset of the data as opposed to all data. If the subset of data can sufficiently describe the whole, then it can be used as an effective approximation of fitness. Using a subset in the realm of 10% the size of the whole data set can greatly reduce the number of evaluations required to determine fitness values. This value is parameterized and os typically determined empirically with preliminary testing.

The fitness predictors are adapted by evolving alongside the candidate solutions, and the fitness predictors' fitness value is determined by a measure of how well it can approximate the whole data set. Additionally, this method also attempts to select the data points in a way which creates a large variance in the fitness of candidate solutions on fitness predictors through the use of *fitness trainers*. In other words, the subset of data points are selected in a way to focus the search on areas of the search space where the candidate solutions are less effective. Figure 2.7 presents an overview of how these fitness predictors would evolve alongside the evolutionary search.

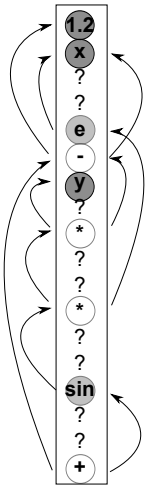
The fitness of the candidate solutions are evaluated using only the current top fitness predictor, however since these populations are evolving in parallel, these predictors are always changing. This allows for a highly dynamic search which can focus on areas of the search space needing the most improvement.



(a) A mathematical expression using a tree representation. 13 are used to represent this expression. Every part of this expression must be explicitly represented with a node.



(b) A mathematical expression using an acyclic graph representation. 9 nodes were used to represent this expression. Notice how the sub-expression $(1.23 - x)$ is easily reused.



(c) Example encoding of the acyclic graph representation. Root is at the bottom, and all children must exist above the parent. '?' represent currently unused information within the structure.

Figure 2.6: Figures 2.6a and 2.6b both represent the same expression: $(1.23 - x) + \sin((1.23 - x) \cdot y \cdot e^x)$. Figure 2.6c shows a possible encoding for an acyclic graph with an array.

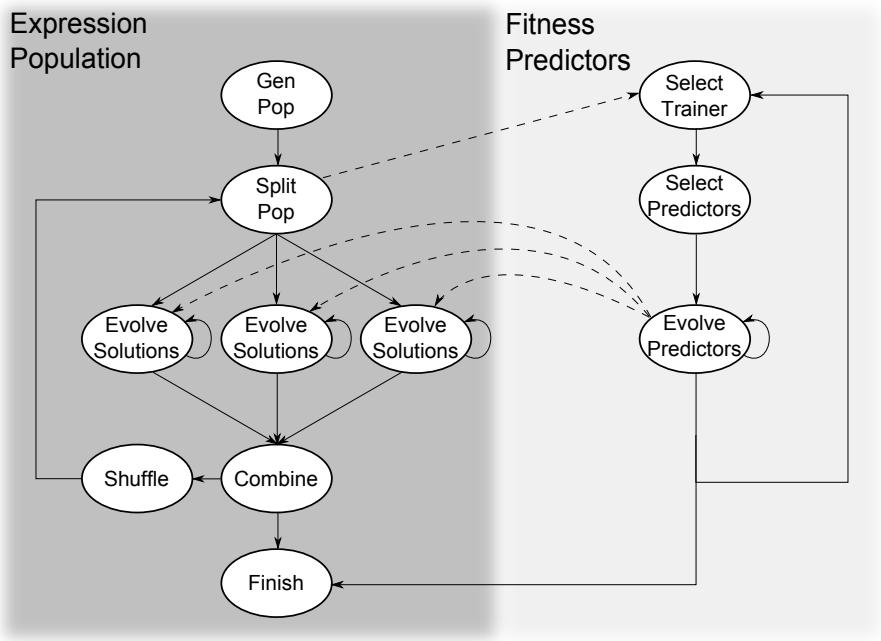


Figure 2.7: High level overview of a GP system implementation with fitness predictors evolving in parallel. This particular example contains multiple sub-populations.

In addition to the performance enhancement of a reduction in computation cost through using a small subset of data points for evaluation, fitness predictors have been shown to produce better quality results by reducing overfitting. Overfitting is curbed since evolution is always based on the fitness of an ever evolving subset of data points; overfitting becomes difficult when the target points are always changing.

It has also been shown that symbolic regression performs better when allowed to focus on key features (subsets of data) as opposed to the whole data set [104, 105].

2.3 Genetic Programming Implementation

The papers in Chapters 4, 5, and 6 contain brief descriptions of the GP implementation and parameter settings used. In summary, the GP system built for this work incorporates multiple enhancements to improve the search capability of the evolutionary algorithm and were ultimately required in order to effectively traverse the large space in a reasonable amount of time. These enhancements include: elitism, an acyclic graph representation [102], fitness predictors [104, 105], and parallel evolution of subpopulations. Figure 2.7 provides a high level overview of the algorithmic flow of the implemented GP system. A GitHub repository of the GP system can be found at `https://github.com/jameshughes89/jGP` [52].

Chapter 3

Functional Magnetic Resonance Imaging Data and Literature Review

MRI is a technology which harnesses magnetic fields to create images of anatomy and physiological processes within a body. MRI uses nuclear magnetic resonance in a controlled way to generate 1 – 5mm^3 *voxels* — three-dimensional volume elements analogous to a two-dimensional pixel — containing information about the spin-relaxation properties of atomic particles within the voxels. This information can be used to distinguish tissue types and properties [16, 26].

MRI works by aligning protons within a body with a very strong magnetic field, applying electromagnetic (EM) energy at a resonance frequency such that specific atomic particles absorb it, and then recording the resulting particle activity [50]. Whether the MRI technology is being used to generate structural images of the brain or the functional moment-to-moment changes within the brain, the high level idea is the same. For the interested reader, a thorough discussion of the underlying technology and phenomenon can be found in [16].

A large static magnetic field is used to align a small, but not insignificant number of protons (the atomic nuclei within the hydrogen atoms in water). Typically, the magnetic field used for MRI is created by passing a current through a coil of superconducting wire. Modern scanners can generate and maintain static magnetic fields between 1.5 – 11T for humans (realistically, it is common to see 1.5T and 3T scanners), and up to 24T for animals. The earth's magnetic field is on the order of 0.0001T (between 25 – 65μT).

The static magnetic field does not create any magnetic resonance signal, but the application of resonant EM radiation to the aligned protons and resulting reaction does. EM radiation (photons) tuned to a specific frequency is applied to the body within the scanner such that some protons absorb the energy and enter an excited state. The specific frequency is selected to be the resonant frequency for the target particle, typically hydrogen nuclei. When the application of the EM radiation is stopped, the excited protons will eventually return to align with the magnetic field, and in doing so, will release *energy over time* that can be recorded.

The way this is measured will record different signals that may be suited for structural or functional imaging. The important part is that with the application of EM radiation, a collection of particles will react a predictable way, and the results can be measured to indicate blood oxygen levels. With the clever use of controlled spatial variations in the magnetic field strength, the recorded signal can be spatially localized. Localized variations in the blood oxygen levels

are of interest as they can be used as a proxy for brain activation.

Functional magnetic resonance imaging (fMRI) is a neuroimaging modality used to measure functional brain activity with the BOLD signal. The BOLD signal is a measure of the relative *(de)*oxygenation level of blood within tissue resulting from an increase in blood flow to cerebral tissue which correlates with neural activation (a result of the HDR) [79, 95, 50]. This is believed to happen because neurons do not store their own energy and oxygen and must depend on the vascular system to replenish resources.

The actual nature of the BOLD signal is not entirely understood, and it should be noted that fMRI is in reality measuring a phenomenon that lags behind electrical recordings of neural actively by a few seconds and is spatially diffused; the *surrogate signal* is based on to the blood flow of surrounding tissue of recent activity. However, it has been firmly demonstrated that this signal is strongly linked with the underlying neural activity [85], but ultimately there are physical and biological limitations to the signal which are consistently under-represented, many of which are reviewed in [5, 46, 84].

fMRI data is four-dimensional; it contains the three-dimensional *anatomical space* along with the changes in activation over *time*. Depending on the technology, the anatomical space is measured from $1 - 5\text{mm}^3$ and changes in time are sampled every 0.5 – 3s; modern scanners are capable of capturing at a frequency of 0.75 – 2Hz. Although the resolution each voxel is on the order of millimetres, each voxel contains tens of thousands of neurons. For this reason they can be thought of as a "mesoscale" representation; it lies between the microscale of neurons and the macroscale of brain lobes.

fMRI is particularly popular as it is non-invasive and has relatively high spatial resolution when compared to other imagining technologies. fMRI allows researchers to ask which brain regions are involved in tasks/stimulus, how they relate to one another, and how they communicate. This technique has been used to ask many interesting behavioural, physiological (functional and structural), clinical questions.

In task-based fMRI, tasks or stimulus are presented to a subject and the corresponding measured signal (BOLD) is compared to the expected HDR [1, 30]; what we expect a measured signal to look like if it were responding to some presented stimulus. Areas of the brain (voxels or other ROI) whose signal corresponds to the expected HDR is said to have been activated by the task/stimulus [18, 96, 101, 49]. Any timeseries metric could be used for comparing the BOLD signal to the expected HDR, however the GLM (general linear model) is very common. The general linear model is a linear model of the form $Y = BX + U$, where Y is a *matrix* of dependent variables, X is a matrix of independent variables, B is a matrix of parameters (to be found), and U is a matrix of errors/residuals. Effectively, it is a generalization of multiple linear regressions for many dependent variables.

Figure 3.1 depicts an HDR function; if a voxel/ROI's measured BOLD signal were to linearly correlate with this spiking event then it is said to be activated by the stimulus.

In resting-state fMRI subjects are placed in the scanner and are *not* presented with an explicit task or stimulus [12, 33]. Here, instead of comparing the measured BOLD signal with the expected HDR, voxels/ROIs are compared to *each other* with some timeseries metric. Similar to the task-based studies which use linear tools, a linear *Pearson product-moment correlation coefficient* is commonly used. These studies analyse spontaneous changes in the measured BOLD signal. One particular area of interest is *default mode networks* (DMN); highly correlated brain regions in the absence of stimulus whose thought to be involved with, although not

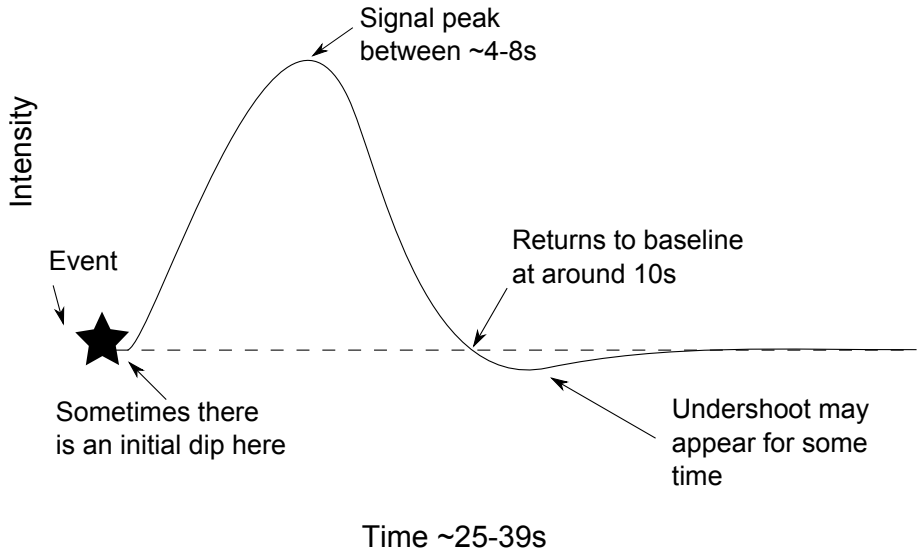


Figure 3.1: Hemodynamic Response Function [15]. After an event/neural spike, the relative deoxygenated blood levels increases (sometimes with an initial dip before the increase) and after roughly 10 seconds, levels returns to close to baseline.

limited to, self-referencing, self-memory, thinking of others, moral reasoning, social and moral reasoning, remembering the past, and planning the future [4]. Additionally, these resting-state networks also have significant clinical implications with respect to better understanding of mental illnesses such as Alzheimer's, and schizophrenia [17, 88, 100, 113].

3.1 Graph Theory

When studying these brain relationships, neuroscientists began simplifying their data and analysis by reducing the four-dimensional fMRI data into a static graph — a set of vertices and edges — representing functional or structural connectivities. With this, graph properties, such as vertex degree or distance, become easy to study.

Interesting early discoveries with this graph approach include network motifs [115], and the prevalence of *small world properties* within these connectivities [11]; graphs with densely connected clusters and a small number of connections between clusters [124].

This approach has become increasingly popular and many of the methods can be viewed in these review papers [19, 28, 114].

3.1.1 Discovering Relationships

For both task and resting-state fMRI it is common to use linear tools (GLM, Pearson product-moment correlation coefficients) to discover underlying relationships. Once these relationships are measured some thresholding method is then used to determine which relationships are significant. Popular thresholding approaches in the literature are based on topological expectations and controlling the FDR (both of which have their statistical and implicational limitations), although alternative approaches with different properties do exist, such as tools using

random matrix theory [26].

It is interesting how well linear models can describe the relationships within the system, given that the brain is a nonlinear computing system as it is Turing-complete. Perhaps a significant portion of the relationships are linear; there has been work done with these linear tools and the majority of the meaningful relationships appear to be linear [108]. Noise has also been demonstrated to obscure potentially important nonlinearities [29]. Additionally, it has been noted that the BOLD response is a nonlinear integrator [121, 123, 82].

3.2 Previous Work on Nonlinear Relationships

Below is a collection of works by various authors exploring nonlinearities within fMRI data. This collection is by no means exhaustive and each individual work includes a literature review of similar research. The last work in this section is the most relevant to this thesis as it is the most similar; it studies the brain from a network/graph perspective and attempts to describe the network relationships using symbolic regression (see Section 2.2).

Buxton et al. describe the *Balloon model*, a nonlinear input-output model of blood flow and oxygenation changes where input is blood flow and output is the BOLD signal [20, 22]. The simplistic biophysical model of the HDR was used in a finger tapping task-based fMRI study to capture essential features of the BOLD signal.

Nonlinear models for the BOLD signal were studied with *Volterra series expansion* by Friston et al. [36, 37]. They develop a nonlinear model of the BOLD signal using Volterra series expansion, a model-independent method capable of modelling the behaviours or any nonlinear time invariant dynamic system [36]. They show that the Balloon model can account for nonlinearities in event-related responses. They also describe a nonlinear dynamic model of the relationship between synaptic activity and fMRI signals. This model incorporates the Balloon model and is characterized in terms of it's Volterra kernels. They argue that the kernel parameters are biologically plausible and are sufficient to account for a number of nonlinearities in the data.

Deneux and Faugeras studied variations of the balloon models (such as those discussed in [22, 21, 37]) and physiological plausible models and their use in fMRI data analysis. They suggest that their models better describe the BOLD response when compared to linear tools, but are comparable when being applied to noisy data [29].

Kruggel et al. used fMRI data recorded from an event related item recognition experimental design [77]. After preprocessing, the authors used linear regression to find areas of functional activation within the data to select ROIs. Once the areas of interest were selected, they used nonlinear regression to quantify the relationships between the stimulus and the BOLD signal's shape. The nonlinear regression used in the work was developed by Kruggel and von Cramon for modelling nonlinearities within fMRI data (described in [78]). They note that the success of a nonlinear analysis of the data is dependent on a well thought out collection of model equations and conclude that their presented approach achieves a *finer* description of the fMRI experiments and hope that it will lead to new insight into cognitive neuroscience.

Friston et al. develop a *dynamic* causal model for nonlinear input-output system and used it to analyze experimental inputs/stimulus and the fMRI measured responses [35]. The model uses bilinear parameters for modelling, however they state that no such restriction is required.

The work was extended by Stephan et al. to perform a more general nonlinear dynamic causal model [117]. More up-to-date information regarding the project can be found at: `http://www.scholarpedia.org/article/Dynamic_causal_modelling` [89]. The extended nonlinear dynamic causal model was capable of distinguishing nonlinear and bilinear processes when applied to synthetic fMRI data. The models were also applied to real fMRI data gathered from a motion task to analyze nonlinearities, namely, *gating* — the response of a neuron to activity is dependent on the history of inputs from other neurons.

Wager et al. show nonlinear effects in fMRI BOLD signal when a rapid event-related experimental design is used (1s apart) [123]. Their interest was in nonlinearities introduced by stimulus history and they developed a low-dimensional parametrization of nonlinearities in response magnitude, time to peek, and response onset time. The authors demonstrate that their model is more accurate and reasonably consistent across the brain. They argue the importance of accounting for nonlinearities when focused on subject specific analysis relative to a group analysis since inaccurate linear models of the nonlinear phenomenon for the group could create biases when applied across participants.

Zhang et al. used a nonlinear semi-parametric model built around Volterra series to characterize measured BOLD signal and found deviations from the linear models [126]. They applied the method to real fMRI data from a *monetary incentive delay* experiment [67] and showed that their approach outperformed many existing methods. They acknowledge the difficulty in selecting the number of parameters for describing nonlinearities with their approach, and therefore they limit the number of functional bases (free parameters).

In 2015 Nicholas Allgaier used symbolic regression as a means to discover nonlinear relationships within resting-state fMRI data [2]. This particular work is the most relevant to this thesis as it is using the same underlying technique to discover nonlinearities and studies them as a network; however, there are important distinctions[1]. The authors studied known networks within resting-state data to develop nonlinear models. 52, $9mm^3$ ROIs were selected based on the DMN. It was found that their nonlinear terms generated in the models better account for variance when compared to traditional linear tools. It was also found that the most common relations modelled corresponded to known *intrinsic connectivity networks*. Similar work was also done by Icke et al. in [61] which hybridized GP with deterministic techniques with success. They also suggest that symbolic regression alone has too many shortcomings to be effective for modelling nonlinearities in fMRI data. Other unpublished works from the same lab also analyze some small sample task-based studies.

3.3 Details on Data Used

As discussed in Chapter 1, task-based fMRI timeseries data was obtained from the Human Connectome Project and segmented into 30 ROIs. Figure 1.1 provides a view of the ROIs and Table 3.2 names the neuroanatomical regions of the 30 ROIs.

Tasks performed for the Human Connectome Project's task-based fMRI data include: Emotion, Gambling, Language, Motor, Relational, Social, and Working Memory.

Table 3.1 provides an example of how the data can be represented simply in tabular format.

[1] James Hughes would like to emphasize that this work was not reported on until after the James' project began.

Table 3.1: Data excerpt from subject 100307 performing the Emotion task after z-score normalization. This table demonstrates the tabular representation of the fMRI timeseries data. ROIs 6 through 29, and time points 10 through 175 were excluded to conserve space.

Time Point	**ROI 1**	**ROI 2**	**ROI 3**	**ROI 4**	**ROI 5**	**...**	**ROI 30**
1	1.22	3.36	1.01	1.49	5.81	...	3.79
2	-0.89	2.00	0.09	-0.36	2.47	...	2.34
3	-2.03	0.46	-0.11	-1.36	0.98	...	1.10
4	-2.33	0.49	0.72	-0.51	0.19	...	-0.64
5	-1.49	1.30	1.13	-0.79	-0.12	...	0.26
6	-1.32	0.37	1.05	-0.98	0.13	...	0.96
7	-0.97	0.06	1.11	0.05	0.30	...	-1.09
8	-0.35	-0.01	1.84	1.50	1.30	...	-1.35
9	0.26	0.00	2.01	1.05	0.28	...	-0.77
...	...	...	...	...	...	...	...
176	0.31	-0.30	0.35	-0.10	0.27	...	0.28

General preprocessing for the fMRI data was implemented in *FMRIB Software Library* (FSL) and FreeSurfer [40, 32, 65, 64]. The tfMRI data, which was processed with *FMRIB's Expert Analysis Tool* (FEAT) [125], was selected for use in this study for convenience and because our goal was to describe nonlinear functional relationships in fMRI data.

Additional preprocessing was attempted on the data in an attempt to reduce noise in the signal; high pass filtering with a Fourier Transform was performed with cutoff values ranging between 1 – 100s (0.01 – 1Hz) using FSL's fslmaths band pass filtering [65]. Preliminary tests however demonstrated that this filtering provided no noticeable impact on the results. Interestingly, using a heavy filter like those common in resting-state fMRI studies significantly hurt results. Additionally, any filtering of the results risks eliminating actual meaningful signal from the data.

A github repository with a dump of preprocessing scripts used can be found at `https://github.com/jameshughes89/NonlinaerFMRIpreprocessing`.

Additional information about the data used in each work are provided within their respective chapters (Chapters 4, 5, 6).

Table 3.2: Region of interest number and corresponding neuroanatomical region. This table provides a frame for the resolution of the brain segmentation.

Region of Interest #	Description
1	Visual (V1)
2	Insula/Medial Temporal (MT)
3	Cuneus
4	Posterior Ventral Temporal
5	Memory
6	Prefrontal Cortex (PFC)
7	Temporal Pole/Amygdala
8	Auditory (Middle/Lateral Temporal)
9	Intraparietal
10	Insula/Medial Temporal (MT)
11	Cerebellar
12	Thalamys/Midbrain
13	Intraparietal/Calculations
14	Prefontal/Orbitalfrontal Cortex (OFC)
15	Temporal Pole/Amygdala
16	Language Associated Prefrontal Cortex
17	Fusiform/Ventral Temporal
18	Prefrontal Cortex (PFC)
19	Lateral Occipital
20	Auditory (Middle/Lateral Temporal)
21	Medial Frontal/M1 area
22	Somatosensory/Premotor (M1/S1)
23	Somatosensory/Premotor (M1/S1)
24	Fusiform/Ventral Temporal
25	Lateral Occipital
26	Cingulate
27	Medial Orbitalfrontal Cortex (OFC)
28	Prefontal/Orbitalfrontal Cortex (OFC)
29	Language Associated Prefrontal Cortex
30	Anterior Cingulate Cortex (ACC) & Prefontal

Chapter 4

Paper 1

This paper was submitted to *Association for Computing Machinery*'s (ACM) *Genetic and Evolutionary Computation Conference* (GECCO) 2016. This paper was accepted as an extended abstract and published in the conference's companion proceedings. The full 8 pages submitted are included within this chapter. The published extended abstract can be found in Appendix A. References contained within this article are numbered according to the article's bibliography.

Finding Nonlinear Relationships in fMRI Time Series with Symbolic Regression

James Alexander Hughes
Computer Science, Brain and Mind Institute
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3K7
jhughe54@uwo.ca

Mark Daley
Computer Science, Biology, Statistics & Actuarial Science, Brain and Mind Institute
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3K7
mdaley2@uwo.ca

ABSTRACT

The brain is an intrinsically nonlinear system, yet the dominant methods used to generate network models of functional connectivity from fMRI data use linear methods (e.g., the Pearson product-moment coefficient). Although these approaches have been used successfully, they are limited in that they can find only linear relations within a system we know to be nonlinear.

This study employs a highly specialized genetic programming system which incorporates multiple enhancements to perform symbolic regression, a type of regression analysis that searches for declarative mathematical expressions to describe relationships in observed data.

Publicly available fMRI data from the Human Connectome Project was segmented into meaningful regions of interest and highly nonlinear mathematical expressions describing functional connectivity were generated with symbolic regression. These nonlinear expressions exceed the explanatory power of traditional linear models and allow for more accurate probing of the underlying physiological connectivities.

CCS Concepts

•Computing methodologies → Genetic programming; *Modeling methodologies;* **•Applied computing → Systems biology;**

Keywords

Symbolic regression; Computational neuroscience; Functional magnetic resonance imaging; Modeling nonlinear relationships.

1. INTRODUCTION

The literature in the field of neuroscience explicitly acknowledges the existence of nonlinear relationships in brain function [3, 4, 7, 12, 10, 27], but it is common to treat them as a footnote or ignore them altogether [4, 19]. Linear tools, such as the General Linear Model (GLM) or the Pearson product-moment coefficient, are used, almost exclusively, to model functional magnetic resonance imaging (fMRI) time series data. However, it would ultimately be improper to use a linear method to observe what we *know* to be nonlinear phenomenon as it lacks the power to truly model the underlying processes.

Despite this, neuroscientific studies are able to make powerful contributions with the limited linear model [4]. However, one wonders if a more expressive, nonlinear method for modeling functional relationships — that must exist within the space — would increase the analytical power and significance of results. Nevertheless, it is not surprising that the nonlinear relationships are ignored; discovering underlying nonlinearities is an exceptionally non-trivial task, especially when working with large amounts of high-dimensional data.

Nonlinear tools have been used in analysis of fMRI time series, but still remain underutilized. Friston et al. used *Volterra series expansion* to study nonlinear responses and showed that nonlinearities in the *Balloon* model (a physiological model) sufficiently describe hemodynamic refractoriness and other nonlinearities in fMRI [10, 11]. A form of *Nonlinear regression* was implemented by Kruggel et al. for modeling dependencies between hemodynamic response and stimulation conditions [18]. *Dynamic Causal Modeling* was done by Friston et al. to describe effective connectivity [9]. In 2014, Zhang et al. succeeded in using a *Semi-parametric* Volterra series based analysis to find deviations from linear assumptions [30]. In 2015, *Symbolic Regression* was used to describe nonlinear relationships within known networks in *resting state* fMRI data [1].

The procedure of observing data, detecting natural laws, and discovering their corresponding formalisms is an intractable task that has been difficult to automate effectively [2, 23, 26]. A promising approach to searching for mathematical expressions describing data is a machine learning technique called *genetic programming* (GP), a tool that can perform *symbolic regression* — a type of regression analysis. GP is a machine learning technique which iteratively writes and updates its own programs to independently learn how to solve a given problem based on the principles of the natural process of evolution [17].

In this work GP will be implemented to automate the process of discovering minimal and interpretable network

GECCO '16 July 20-24, 2016, Denver, CO, USA

ACM ISBN 978-1-4503-4323-7/16/07.
DOI: http://dx.doi.org/10.1145/2908961.2909021

relationships in the behavior of a system for which we can observe only time series derived from a network's nodes: task based fMRI time series data. No assumptions or prior knowledge will be applied to the system.

The GP system built for the purpose of this work incorporates multiple enhancements to improve the search capability of the evolutionary algorithm and were ultimately required in order to effectively traverse the large space in a reasonable amount of time. These enhancements include: an acyclic graph representation [22], fitness predictors [24, 25], and parallel evolution of subpopulations.

2. DATA

The data selected for this analysis was obtained from the Human Connectome Project, WU-Minn Consortium[1]. The data contains structural MRI, resting state fMRI (rfMRI), diffusion imaging (dMRI), and task-evoked fMRI (tfMRI) for roughly 500 subjects and Magnetoencephalography (MEG) data for resting state and tasks on a subset of the participants.

This work focuses on task based fMRI time series data which is obtained by placing subjects into fMRI scanners and instructing them to perform some task. This technology harnesses nuclear magnetic resonance to generate data containing information which can ultimately be used as proxy for functional activation within particular brain voxels [7].

The actual information being recorded by the fMRI is the blood oxygen level dependent (*BOLD*) signal — a measure of the relative oxygenation level of blood within tissue. This change is the result of an increase in blood flow to cerebral tissue which correlates with neural activation (a result of the *hemodynamic response* (HDR)) [21, 14]. Although it has been demonstrated that the modulation of blood flow to tissue is strongly linked with the actual underlying functional activity [20], it is important to note that the actual nature of the BOLD signal is still under investigation [7]. Additionally, the BOLD signal lags behind electrical recordings of neural actively by a few seconds and is spatially diffused — the signal is based on to the bloodflow of surrounding tissue of recent activity. fMRI signal also has a very high signal to noise ratio.

The fMRI time series data can be visualized as a two dimensional matrix of voxels (three-dimensional analogues to two-dimensional pixels). One flattened dimension represents all voxels in the three dimensional physical space and the other dimension represents time points; each entry in the matrix corresponds to the BOLD signal intensity of a single voxel at a particular time point. The actual number of voxels depends on the resolution of fMRI scanner. For example, modern hardware with a resolution of $1\text{-}5mm^3$ can capture hundreds of thousands of voxels. Similarly, the number of time points depends on the hardware and overall duration of the experiment; how long a subject was in the fMRI. Modern scanners are capable of capturing at a frequency of *0.75Hz - 2Hz*.

Although the resolution of each voxel is on the order of millimeters, each voxel contains tens of thousands of neurons. For this reason they can be thought of as a "mesoscale" representation; it lies between the microscale of neurons and the macroscale of brain lobes.

General preprocessing for the fMRI data was implemented in FSL and FreeSurfer [13, 8, 16, 15]. The tfMRI data, which was processed with *FMRIB's Expert Analysis Tool* (FEAT) [28], was selected for use in this study for convenience and because our goal was to describe nonlinear functional relationships in fMRI data.

Tasks performed for the Human Connectome Project's tfMRI data include: Working Memory, Gambling, Motor, Language, Social Cognition, Relational Processing, and Emotion Processing. It was decided to focus on the Motor task as it is a clean and simple task that is highly studied and already well understood with linear tools. A total of 509 subjects were imaged for the Motor task, but only 507 were studied in this work as two were missing data.

The Motor task was adapted from studies performed by Buckner et al. and Yeo et al. [5, 29]. Visual cues were presented to subjects which instructed them to either tap their left or right finger, squeeze their left or right foot, or move their tongue. Tasks were divided into *blocks* and each block lasted 12 seconds and included 10 movements of respective body part. 13 blocks were included in each run: 2 left finger, 2 right finger, 2 left foot, 2 right foot, 2 tongue, and 3 fixation blocks of 15 seconds each. A 3 second cue was presented before each block. This task had a total run duration of 3:34 and contained 284 frames per run (including pre- and post-task data) and was performed twice for each subject, one for each phase encoding. The temporal resolution was *720ms*; a whole brain volume was captured at a rate of roughly *1.389Hz*.

Additional preprocessing was attempted on the data in an attempt to reduce noise in the signal; high pass filtering was performed with values ranging between $1 - 100s$. Preliminary tests however demonstrated that this filtering provided no noticeable impact on the results. Interestingly, using a heavy filter like those common in resting state fMRI studies significantly hurt results. Additionally, any filtering of the results risks eliminating actual meaningful signal from the data.

The fMRI time series data was segmented into meaningful regions of interest (ROIs) (refer to Figure 1) with Craddock et al.'s *spatially constrained parcellation* [6]. Each voxel's activation within each ROI was averaged to determine the respective ROI's mean activation. A variety of resolutions were explored and it was determined that using *30 ROIs* consistently resulted in meaningful expressions. Using fewer than 30 ROIs appeared to not provide enough resolution for the GP system to find substantive expressions, and using anything greater than 30 ROIs overloaded the GP system; current hardware's computational power limits the search space of the GP system. Refer to Table 1 for details on the neuroanatomy of the 30 ROIs.

Ultimately the data was represented in a two dimensional matrix with 30 columns (ROI average activation) and 284 rows (time points).

3. GENETIC PROGRAMMING IMPLEMENTATION

This specific GP implementation is motivated by Schmidt et al.'s work [26], is extremely specialized for symbolic regression, and incorporates modular improvements which significantly increase the performance of the system. Some of these improvements including parallel evolution of subpopulations, fitness predictors [24, 25], and an acyclic graph rep-

[1]http://www.humanconnectome.org/

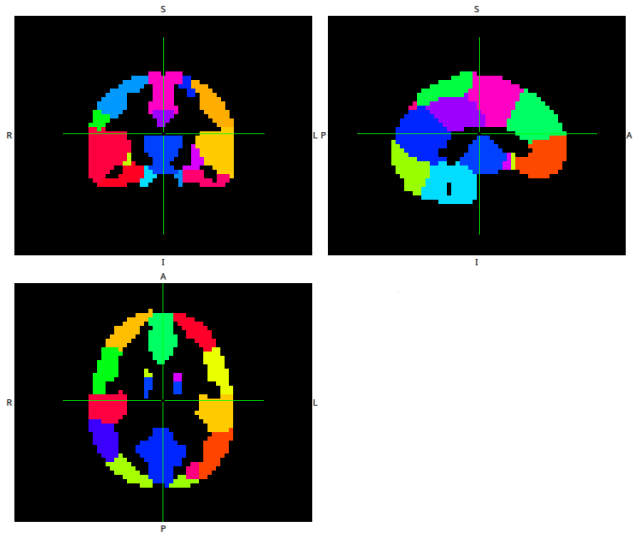


Figure 1: A snapshot of a brain when segmented into the 30 ROIs. Each colour represents a different region.

ROI #	Description
1	Visual (V1)
2	Insula/Medial Temporal (MT)
3	Cuneus
4	Posterior Ventral Temporal
5	Memory
6	Prefrontal Cortex (PFC)
7	Temporal Pole/Amygdala
8	Auditory (Middle/Lateral Temporal)
9	Intraparietal
10	Insula/Medial Temporal (MT)
11	Cerebellar
12	Thalamys/Midbrain
13	Intraparietal/Calculations
14	Prefontal/Orbitalfrontal Cortex (OFC)
15	Temporal Pole/Amygdala
16	Language Associated Prefrontal Cortex
17	Fusiform/Ventral Temporal
18	Prefrontal Cortex (PFC)
19	Lateral Occipital
20	Auditory (Middle/Lateral Temporal)
21	Medial Frontal/M1 area
22	Somatosensory/Premotor (M1/S1)
23	Somatosensory/Premotor (M1/S1)
24	Fusiform/Ventral Temporal
25	Lateral Occipital
26	Cingulate
27	Medial Orbitalfrontal Cortex (OFC)
28	Prefontal/Orbitalfrontal Cortex (OFC)
29	Language Associated Prefrontal Cortex
30	Anterior Cingulate Cortex (ACC) & Prefontal

Table 1: Neroanatomical regions and their corresponding segmented regions of interest. This list provides a frame for the resolution of the currently attempt to model functional activity with symbolic regression.

resentation [22]. With these improvements, this intractable problem becomes much more manageable.

Multiple subpopulations are evolved separately from one another to encourage a well diversified set of candidate solutions. Periodically these subpopulations are combined, shuffled, and redistributed into subpopulations again. This procedure attempts to simulate *punctuated equilibria* and *allopatric speciation*, or simply, encourage genetic diversity over the whole population by allowing the subpopulations to traverse the search space along their own trajectories.

3.1 Fitness Predictors

Fitness predictors have been demonstrated to reduce computational cost by approximating the local search gradient [24, 25].

Candidate solutions are evaluated with an adaptive subset of the data as opposed to all data points. If the subset of data can sufficiently describe the whole, then it can be used as an effective approximation of fitness. Using a subset in the realm of 10% the size of the whole data set can reduce the number of evaluations required to determine fitness values.

The fitness predictors are evolved alongside the candidate solutions, and the fitness predictors's *fitness* value is determined by a measure of how well it can approximate the whole data set. Additionally, this method ultimately also attempts to select the data points in a way which creates a large variance in the *fitness* of fitness predictors through the use of *fitness trainers*. In other words, the subset of data points are selected in a way to focus the search on areas of the search space where the candidate solutions need the most improvement.

The fitness of candidate solutions are evaluated using *only* the subset of data points within the current top fitness predictor, however since these populations are evolving in parallel, these predictors are always changing. This allows for a highly dynamic search which can focus on areas of the search space needing the most improvement.

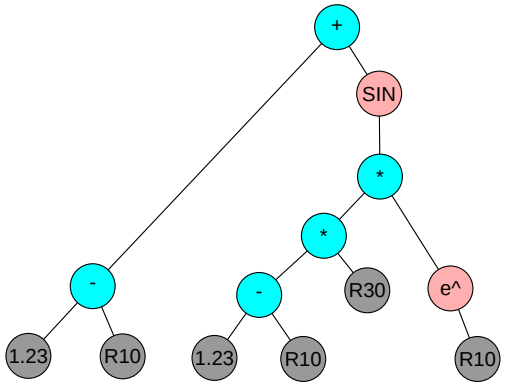
In addition to the performance enhancement of a reduction in computation cost, fitness predictors have been shown to produce better quality results by reducing overfitting. Overfitting is curbed since evolution is always based on the fitness of an ever evolving subset of data points; overfitting becomes difficult when the target points are always being adapted [25].

It has also been shown that symbolic regression performs better when allowed to focus on key features (subsets of data) as opposed to the whole data set [25].

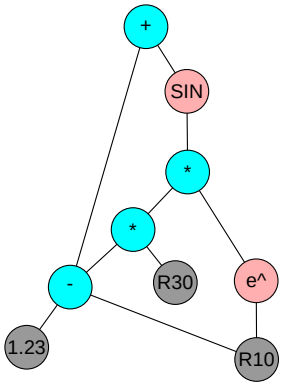
3.2 Acyclic Graph Representation

Bloat can be reduced with the use of fitness predictors [25], however, in this work an acyclic graph representation has been implemented which can also bias the algorithm away from bloated, overfit expressions [22].

Traditional systems typically represent expressions in tree structures where leaf nodes are terminals (constants or variables) and internal nodes are operators. In this work an *acyclic graph representation* is used. Figure 2 presents a comparison of a tree representation and an acyclic graph representation. These structures both represent the following equation: $R21 = (1.23 - R10) + Sin((1.23 - R10) * R30 * e^{R10})$, where $R10$, $R21$, and $R30$ are variables representing the values of regions of interest 10, 21, and 30 over the time series. Even with this simple example, notice how the acyclic graph



(a) An example mathematical expression using a typical tree representation. 13 nodes were used to represent this expression with this tree representation. Every part of this expression needs to be explicitly represented with a individual node.



(b) An example mathematical expression using an acyclic graph representation. 9 nodes were used to represent this expression with this acyclic graph representation. Notice how the sub-expression $(1.23 - R10)$ is easily reused.

Figure 2: Figures 2a and 2b both represent the same expression: $(1.23 - R10) + Sin((1.23 - R10) * R30 * e^{R10})$, however, Figure 2b was able to represent the same information as Figure 2a with less resources. In this example, *blue* nodes represent binary operators, *red* nodes represent unary operators, and *grey* nodes represent a terminals;Rx signifies a variable (region of interest in this case) and a number signifies a constant.

representation is more *succinct* relative to the tree structure.

Motivation for selecting the acyclic graph representation for this work was, first, the representation scales well and avoids bloat, and second, it can reuse possibly important subexpressions effectively [22].

3.3 Execution of Evolutionary Search

An overview of the execution flow of the whole GP system is presented in Figure 3. The left side of Figure 3 depicts the execution of the evolution of the population of candidate solutions (mathematical expressions in this particular case) while the right side depicts the execution of the evolution of the fitness predictors (in this case, a collection of subsets of data points from the fMRI time series).

After a population of candidate solutions is generated a set of fitness trainers and fitness predictors are generated to be evolved. Following this, the whole population of candidate solutions is split into an arbitrary number of subpopulations to be evolved in parallel.

Evolution of the subpopulations occurs in parallel, however, evaluation of the subpopulations is calculated with the subset of data points within the most fit fitness predictor, which is always changing as it is evolving at the same time as the subpopulations. The motivation for this is described in Section 3.1.

After a predefined number of iterations of evolution, the subpopulations are recombined, and if some stopping criteria is met, the execution completes. If execution is not complete, the combined subpopulations are shuffled and will ultimately split once again into many subpopulations. Just as it was done earlier, the fitness trainers and predictors are updated based on the current whole population before it is split into subpopulations.

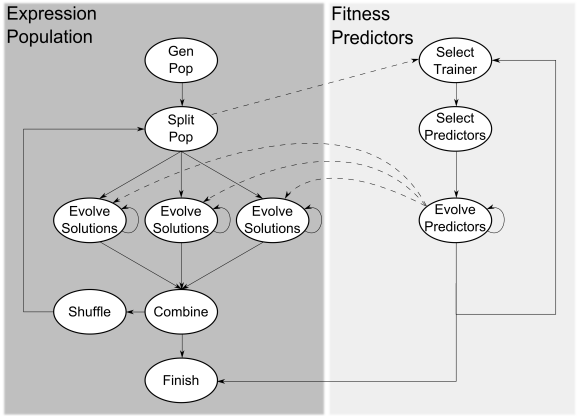


Figure 3: High level structure of this symbolic regression implementation. The left side demonstrates the evolution of the expressions while the right side depicts the evolution of fitness predictors. This example shows only three subpopulations evolving in parallel.

4. EXPERIMENTAL METHODS

Typical task based fMRI studies employ linear methods, such as Pearson product-moment correlation and the GLM to analyze their results. For example, if one wants to derive

how ROI 21 is related to the other ROIs, then they would use some thresholding technique to eliminate statistically unrelated ROIs (based on correlation) and then employ a GLM to fit the other ROIs to ROI 21. These methods make the large assumption that the system is describable with only linear operators.

For symbolic regression, it was required to have some value over the time series that the evolved expressions fits to. For the purpose of the motor task, *ROI 21* was selected as it is the ROI that contained the *primary motor cortex*. All ROIs, with the exception of ROI 21, were fed into the GP system and acted as variables over the time series and the expressions evolved to equate these variables to ROI 21. Although these values are being equated to ROI 21, the actual expressions generated can be thought as of more how the ROIs relate to one another over the time series.

All *507* subjects for which a complete set of data existed from the Human Connectome Project were used for the purpose of this work. For each of these subjects the search for expressions describing the data was executed *100* times in an effort to improve significance of obtained results. A total of *50,700* executions of evolutionary search were performed.

The language/basis functions used for the experiments included unary and binary linear and nonlinear operators. These operators include: $+$, $-$, $*$, $/$, e, abs, ln, sin, cos, and tan. It should be emphasized here that using symbolic regression to model fMRI time series does make one assumption: it is assumed that the language (basis functions) provided to the system is sufficient to describe the data.

Mean Squared Error was used for the fitness function.

No rigorous parameter tuning was done since preliminary results demonstrated that fine tuning the large number of system parameters provided no meaningful improvement over the large number of runs per subject.

7 individual populations of *25* candidate solutions were evolved in parallel. The choice of *7* populations was because the evolutionary search was being executed on *8* core systems, and with the addition of fitness predictors running on a single core, a total of 8 threads were effectively utilized: 1 thread per core. *10,000,000* generations were done with shuffles of the populations and updates of fitness trainers occurring every *1,000* generations; a total of *10,000* shuffles and updates to fitness predictors were done over the course of the *10,000,000* generations. This all results in a total of *1,750,000,000* mating events. Although these values are excessive and a reduction by orders of magnitude has little impact on the models, the goal is to find the best possible nonlinear model. Even after apparent convergence, any marginal improvement may be important for describing the underlying complex system.

The maximum number of operators/operands in the acyclic graph representation for a candidate solution was set to *40*, however the actual number operators and operands in the expression represented can be higher since subexpressions can be reused.

The crossover and mutation rates were set to *80%* and *50%* respectively. *Two* mutations were possible per candidate solution for each propagation to the next generation.

The number of fitness trainers was *8*, the number of fitness predictors was *10*, and the number of data points per fitness predictor was set to *10%* the size of the total number of data points within the data, which was *28* for this particular task as there were *284* data points in total for each subject.

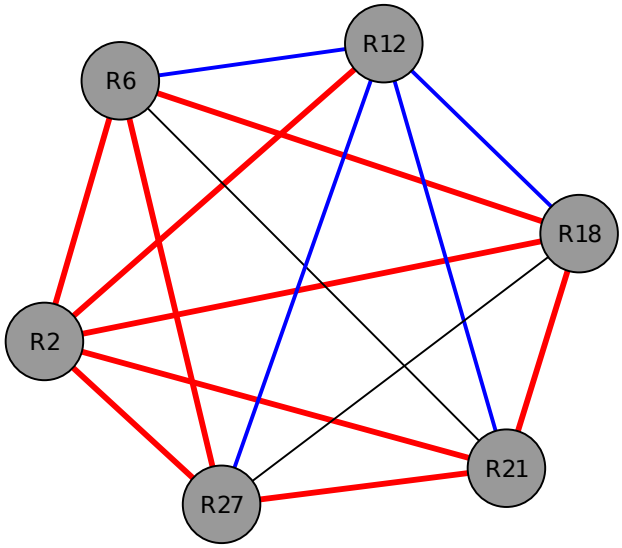


Figure 4: Representation of relationships between regions of interest for a single generated expression. Red lines represent nonlinear relationships, blue lines represent nonlinear and linear relationships, and black lines represent strictly linear relationships. This particular example corresponds to the equation: $R21 = R12 - sin(11.97 * (18.30 - R12)) - (0.42 * |(R12 - R18) * R27|)/(R6 - tan(R2))$ which had a absolute average error from the measured signal of 12.4.

Keeping in mind the stochastic nature of the algorithm, each run of the evolutionary search using this GP system takes between *1-4* hours when running with 8 cores on an *IBM System x iDataPlex dx360 M3* node with 2 quad-core *Intel Nehalem (Xeon 5540)* processors running at 2.53GHz.

5. RESULTS AND DISCUSSION

5.1 Individual Models

The GP system was run 100 times for each subject to find numerous high quality models. Highly nonlinear expressions are found in top models on all subjects. Figure 4 depicts the relationships between ROIs in the top model from the very first subject studied. Notice how most relationships in this model are nonlinear.

To demonstrate that there is meaningful information in the data being studied, and that symbolic regression finds meaningful relationships, it was applied to a random Gaussian brain. This random brain was generated to have similar signal intensities to the real subjects. For this reason, the brain was generated with a mean of *10,000* and standard deviation of *2,000*. Models for this random brain would unsurprisingly equate ROI 21 with some value near *10,000*, such as $e^{9.21}$, and would have an absolute error of roughly *2000*.

Figure 5 shows a time series of one subject's recorded signal alongside two models describing the signal, one found with the nonlinear tool, and the other with linear regression after thresholding ROIs with a 95% false discovery rate (FDR). This figure clearly shows that both the nonlinear and linear models fit the data very well with minimal error. The mean absolute errors over the whole time series on this

one subject for the nonlinear model and linear models are *10.82* and *8.85* respectively.

The interesting observation here is that the nonlinear search found highly nonlinear models that could not be found with the linear tool; however, the models generated with the linear tool fit the data better.

5.2 Analyzing All Models

Figures 6 and 7 show the percentage of times each ROI appeared in models on average over all generated models on every subject.

Figure 6 shows the percentage of times ROIs appeared in the nonlinear models versus the number of times ROIs appear when using a linear tool after thresholding with 95% FDR. Even with the popular thresholding technique, almost every ROI is always related to ROI 21. The mean absolute time series errors for the top nonlinear models and the thresholded linear models on their respective subjects averaged over every subjects were calculated. These values were roughly *16.68* with a standard deviation of *3.51* for the nonlinear models and *11.79* with a standard deviation of *1.11* for the linear models. A Mann-Whitney U test provides a p-value of $3.08 * 10^{-133}$, which clearly demonstrates that the linear models, although only slightly better than the nonlinear, are in fact *fitting the recorded signal better*.

Even with a popular thresholding method almost every ROI is linearly related to ROI 21. This ultimately generates a set of models created with high degrees of freedom that fit the data well, however they provide minimal insight and are difficult to interpret.

On average, a nonlinear model contained slightly less than 4 ROIs (3 when excluding ROI 21). Figure 7 shows a comparison only when the top 3 linearly related ROIs from the linear models (4 when including ROI 21) were counted. By observing this figure one can see that the same ROIs appear at very similar rates for both the linear and nonlinear tools. The mean absolute time series error of the linear models generated with roughly the same number of ROIs was calculated to be approximately *19.16* with a standard deviation of *5.08*. A Mann-Whitney U test comparing the 4 ROIs models provided a p-value of $8.56 * 10^{-19}$; the nonlinear models were significantly better. In fact, it was not until the linear models were given the top *8 ROIs* that there was no more statistical difference. Linear models only performed better than the 4 ROIs nonlinear models once they received *10 or more ROIs* (with a p-value of $1.34 * 10^{-3}$); it took at least 10 ROIs for a linear model to fit the recorded signal better than a nonlinear model containing only 4.

6. CONCLUSIONS AND FUTURE WORK

A highly specialized GP designed for symbolic regression was implemented to search for nonlinear relationships in a dynamic complex nonlinear system: the human brain. Task based fMRI data was modeled and numerous nonlinear relationships were found that would otherwise not be discovered with conventional tools.

When compared to linear models generated with all ROIs available after a typical thresholding technique, the nonlinear models, although close, could not fit the signal as well. Unfortunately however, these linear models would typically contain more than 25 ROIs and would be difficult to interpret and provide minimal insight into understanding the underlying processes. Alternatively, nonlinear models were more succinct. On average, with just 4 ROIs, a nonlinear model could fit the recorded signals better than linear models using 8; even with more information (ROIs), linear models could not describe the data as clearly.

This work can be continued in multiple directions. Current work is being done to explore how well a model generated for some subject can fit data from another subject. Results have been generated for unnormalized data with inconsistent results as there are numerous linear translations in the data across different ROIs. However there are few cases where other subjects models can fit the data very well. Modeling will be performed on z-score normalized data in order to evaluate how models can generalize across subjects.

Data from other tasks, such as gambling, emotional processing, social cognition, and relational processing will be analyzed to determine if the descriptive power of the nonlinearities is persistent across an array of task complexities.

Performing symbolic regression on other neuroimaging modalities where nonlinearities are observed, such as EEG or MEG, could allow for a comparison to the models generated on fMRI data.

7. ACKNOWLEDGMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Computations were performed on the General Purpose Cluster supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

8. REFERENCES

[1] N. Allgaier. Reverse engineering the human brain: An evolutionary computation approach to the analysis of fmri. 2015.

[2] J. Bongard and H. Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007.

[3] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *The journal of neuroscience*, 16(13):4207–4221, 1996.

[4] R. Buckner and T. Braver. Event-related functional mri. In P. Bandettini and C. Moonen, editors, *Functional MRI*, chapter 36, pages 441–452. Springer-Verlag.

[5] R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. Yeo. The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(5):2322–2345, 2011.

[6] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. A whole brain fmri atlas

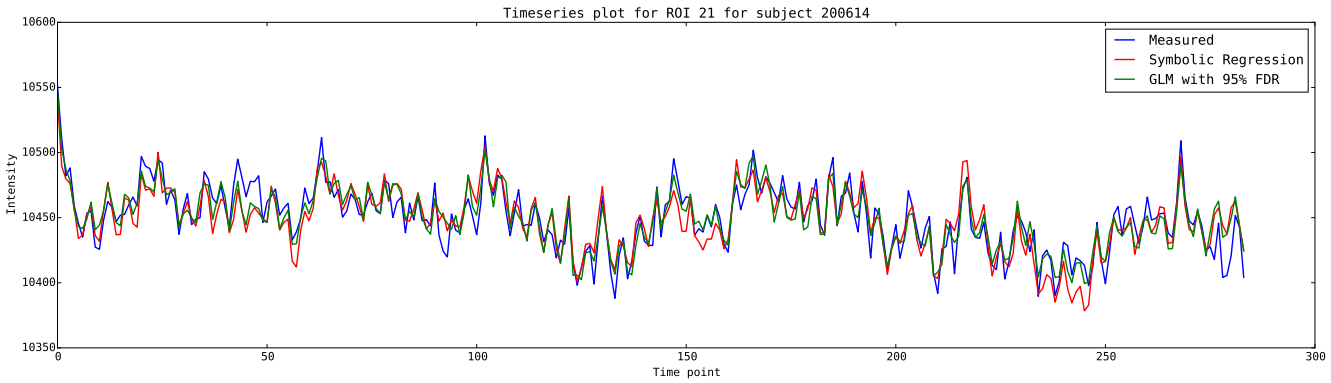


Figure 5: Time series of ROI 21's signal compared to the generated nonlinear and linear models. It is clearly depicted that both models can fit the data very well over the whole time series. An interesting observation is that the models are closer to one another than to the recorded signal.

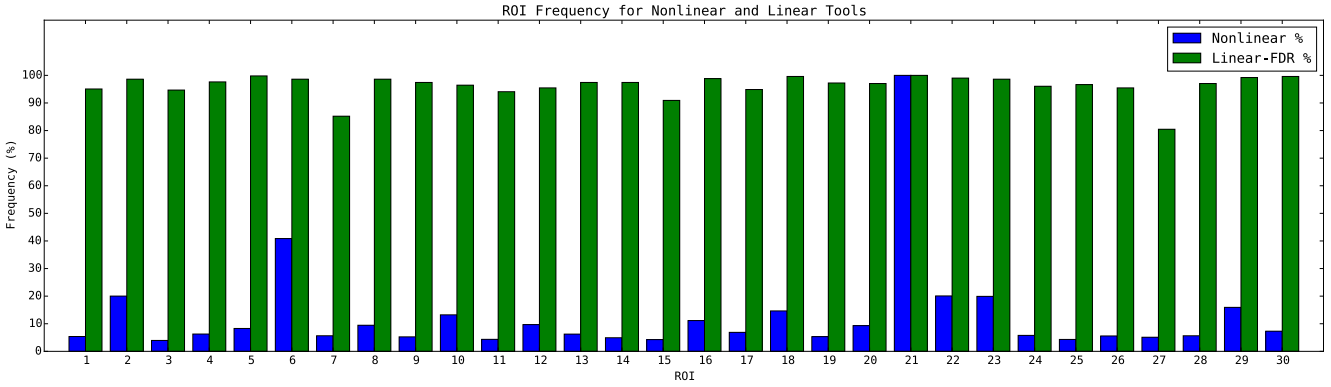


Figure 6: ROI count averaged and compared to False Discovery Rate with 95%. Almost all ROIs are always related to ROI 21 with this popular neuroimaging thresholding technique.

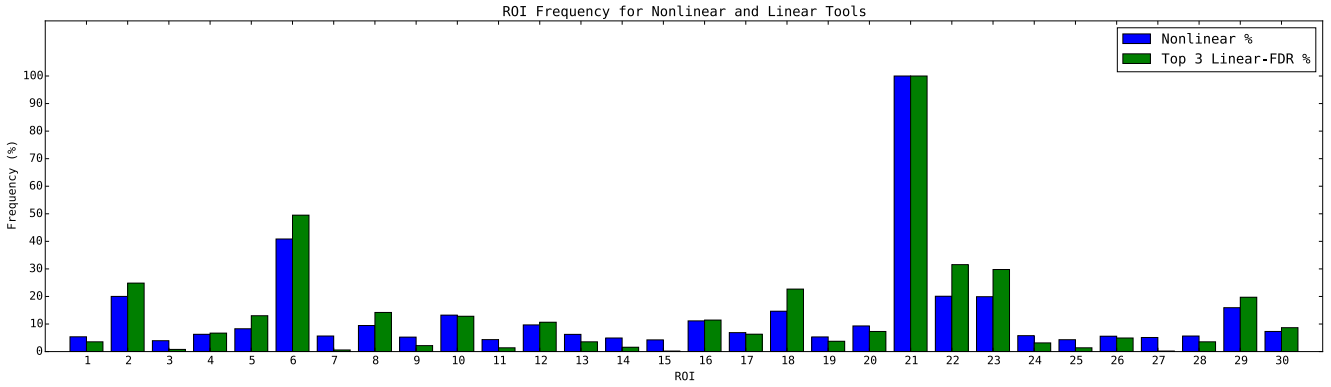


Figure 7: ROI count averaged and compared to top 3 of the False Discovery Rate with 95%. Note that the average number of ROI that appeared in a nonlinear model over all models on all subjects is slightly less than 3 (4 when counting ROI 21). Because of this, the comparison with the top 3 ROI (4 when counting ROI 21) in the linear model will not exactly match.

generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

[7] M. Daley. An invitation to the study of brain networks, with some statistical analysis of thresholding techniques. In *Discrete and Topological Models in Molecular Biology*, pages 85–107. Springer, 2014.

[8] B. Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[9] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[10] K. J. Friston, O. Josephs, G. Rees, and R. Turner. Nonlinear event-related responses in fmri. *Magnetic resonance in medicine*, 39(1):41–52, 1998.

[11] K. J. Friston, A. Mechelli, R. Turner, and C. J. Price. Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics. *Neuroimage*, 12(4):466–477, 2000.

[12] K. J. Friston, C. Price, P. Fletcher, C. Moore, R. Frackowiak, and R. Dolan. The trouble with cognitive subtraction. *Neuroimage*, 4(2):97–104, 1996.

[13] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[14] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, second edition, 2009.

[15] M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[16] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. FSL. *Neuroimage*, 62(2):782–790, 2012.

[17] J. R. Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.

[18] F. Kruggel, S. Zysset, and D. Y. von Cramon. Nonlinear regression of functional mri data: an item recognition task study. *Neuroimage*, 12(2):173–183, 2000.

[19] N. K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

[20] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–157, 2001.

[21] S. Ogawa, D. W. Tank, R. Menon, J. M. Ellermann, S. G. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

[22] M. Schmidt and H. Lipson. Comparison of tree and graph encodings as function of problem complexity. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1674–1679. ACM, 2007.

[23] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.

[24] M. D. Schmidt and H. Lipson. Coevolving fitness models for accelerating evolution and reducing evaluations. In *Genetic Programming Theory and Practice IV*, pages 113–130. Springer, 2007.

[25] M. D. Schmidt and H. Lipson. Coevolution of fitness predictors. *Evolutionary Computation, IEEE Transactions on*, 12(6):736–749, 2008.

[26] M. D. Schmidt, R. R. Vallabhajosyula, J. W. Jenkins, J. E. Hood, A. S. Soni, J. P. Wikswo, and H. Lipson. Automated refinement and inference of analytical models for metabolic networks. *Physical biology*, 8(5):055011, 2011.

[27] A. L. Vazquez and D. C. Noll. Nonlinear aspects of the bold response in functional mri. *Neuroimage*, 7(2):108–118, 1998.

[28] M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith. Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, 14(6):1370–1386, 2001.

[29] B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–1165, 2011.

[30] T. Zhang, F. Li, M. Z. Gonzalez, E. L. Maresh, and J. A. Coan. A semi-parametric nonlinear model for event-related fmri. *Neuroimage*, 97:178–187, 2014.

Chapter 5

Paper 2

This paper was submitted to *Association for Computing Machinery*'s (ACM) *Genetic and Evolutionary Computation Conference* (GECCO) 2017 and was accepted for presentation and publication in the conference proceedings [57]. An abstract based on this work was submitted and accepted to the *11th Annual Canadian Association of Neuroscience Meeting 2017* [56]. References contained within this article are numbered according to the article's bibliography.

Within this article it stresses that the goal is to find descriptive models and not predictive models. This is a subtle point that is intended only to emphasize the motivation of generating these models. Ultimately, assuming the models are accurate, they could be used for prediction.

The article states "Although it is possible the whole brain is involved with the task meaningfully, it would seem unlikely that the relationships are truly significant.", which is not an empirically demonstrated fact, and is a current discussion within connectomics.

It also states that symbolic regression is model free. Within the field this is not uncommon phrasing, however it is not truly model free as there are a number of constraints the GP system must work with (language/basis functions, independent variables).

Searching for Nonlinear Relationships in fMRI Data with Symbolic Regression

James Alexander Hughes
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3M1
jhughe54@uwo.ca

Mark Daley
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3M1
mdaley2@uwo.ca

ABSTRACT

The vast majority of methods employed in the analysis of functional Magnetic Resonance Imaging (fMRI) produce exclusively linear models; however, it is clear that linear models cannot fully describe a system with the observed behavioral complexity of the human brain — an intrinsically nonlinear system. By using tools embracing the possibility of modeling the underlying nonlinear system we may uncover meaningful undiscovered relationships which further our understanding of the brain.

We employ genetic programming, an artificial intelligence technique, to perform symbolic regression for the discovery of nonlinear models better suited to capturing the complexities of a high dimensional dynamic system: the human brain.

fMRI data for multiple subjects performing different tasks were segmented into regions of interest and nonlinear models were generated which effectively described the system succinctly. The nonlinear models contained undiscovered relationships and selected different sets of regions of interest than traditional tools, which leads to more accurate understanding of the functional networks.

CCS CONCEPTS

•Computing methodologies → Genetic programming; *Modeling methodologies;* **•Applied computing → Systems biology;**

KEYWORDS

Symbolic regression; Computational neuroscience; Functional magnetic resonance imaging; Nonlinear Modelling.

ACM Reference format:
James Alexander Hughes and Mark Daley. 2017. Searching for Nonlinear Relationships in fMRI Data with Symbolic Regression. In *Proceedings of GECCO '17, Berlin, Germany, July 15-19, 2017,* 8 pages.
DOI: http://dx.doi.org/10.1145/3071178.3071209

1 INTRODUCTION

The human brain is a manifestly nonlinear system; however, despite the prevalence of literature acknowledging the existence of these nonlinearities [4, 5, 7, 9, 11, 26], the bulk of the cognitive neuroscience literature continues to almost exclusively employ linear tools to model functional Magnetic Resonance Imaging (fMRI) data (e.g., the Pearson product-moment coefficient, general linear model) [5, 18]. This approach, while effective in some ways, robs us of the ability to find the nonlinearities which we know exist within the system.

Despite this, neuroscientific studies are able to make powerful contributions with the limited linear model [5]. However, a question naturally arises: would a more expressive, nonlinear method for modeling functional relationships — that *must* exist within the space — increase the analytical power and significance of results? Nevertheless, it is not surprising that the nonlinear relationships are ignored. Discovering underlying nonlinearities is an exceptionally non-trivial task, especially when studying large amounts of high-dimensional data.

Nonlinear tools have been applied to fMRI time series, but still remain remarkably underutilized. Friston et al. used *Volterra series expansion* to study nonlinear responses and showed that nonlinearities in the *Balloon* model (a physiological model) sufficiently describe hemodynamic refractoriness and other nonlinearities [9, 10]. A form of *nonlinear regression* was implemented by Kruggel et al. for modeling dependencies between hemodynamic response and stimulation conditions [17]. *Dynamic Causal Modeling* was done by Friston et al. to describe effective connectivity [8]. In 2014, Zhang et al. succeeded in using a *Semi-parametric Volterra series* based analysis to find deviations from linear assumptions [27]. Icke et al. employed a *hybrid deterministic regression/genetic programming* approach in 2014 to study *resting state* fMRI data [14]. In 2015, *Symbolic Regression* was used by Allgaier et al. to describe nonlinear relationships within known networks in *resting state* fMRI data and reported novel nonlinearities [1, 2]. In 2016, Hughes and Daley used *Symbolic Regression* to discover novel nonlinear relationships within *task based* fMRI data (a motor task) and found nonlinear models to be more succinct [13].

The procedure of observing data, detecting natural laws, and discovering their corresponding formalisms is an intractable task that has been difficult to automate effectively [3, 22, 25]. A promising approach to searching for mathematical expressions describing data is *genetic programming* (GP), an evolutionary algorithm that can perform *symbolic regression* — a type of regression analysis. GP is a technique that iteratively writes and updates its own programs to independently learn how to solve a given problem based on the principles of the natural process of evolution [16].

In this work GP will be employed to automate the process of discovering interpretable network relationships in the behavior of

GECCO '17, Berlin, Germany

978-1-4503-4920-8/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3071178.3071209

a system for which we can observe only time series derived from a network's nodes: *task based* fMRI time series data. No assumptions or prior knowledge of linearity or how the system interacts with itself is used. The goal is *not* to find predictive models, but descriptive ones. The brain is a very high dimensional system and technological limitations allow for only a small number of data points. It would be unreasonable to expect the creation of a true predictive model with so little data whether using traditional linear regression, or something as rigorous as symbolic regression. The current goal is to generate descriptive representations of the systems that can be used to further our understanding of the functional network relationships within the brain.

2 DATA

Data selected for analysis was obtained from the *Human Connectome Project, WU-Minn Consortium*[1]. The current iteration of the data contains structural MRI, resting state fMRI (rfMRI), diffusion imaging (dMRI), and task-evoked fMRI (tfMRI) for roughly 900 subjects and Magnetoencephalography (MEG) data for resting state and tasks on a subset of participants.

This study focuses on *task based* fMRI time series data obtained by placing subjects into fMRI scanners (Figure 1) and having them perform some task, such as a movement, memory, or gambling task. This technology harnesses nuclear magnetic resonance to generate data which can be used as *proxy* for functional activation within the brain [7]. The actual information being recorded by fMRI is the blood oxygen level dependent (*BOLD*) signal — a measure of the relative oxygenation level of blood within tissue. This change is the result of an increase in blood flow to cerebral tissue which correlates with neural activation (a result of the *hemodynamic response* (HDR)) [12, 20]. Although it has been demonstrated that the modulation of blood flow to tissue is strongly linked with the actual underlying functional activity [19], it is important to note that the precise nature of the BOLD signal is still under investigation [7]. Additionally, the BOLD signal has a low signal to noise ratio, lags behind electrical recordings of neural actively by a few seconds, and is spatially diffused — the signal is based on the bloodflow to tissue surrounding recent activity.

fMRI time series data can be represented as a two dimensional matrix of voxels (three-dimensional analogues to two-dimensional pixels). One dimension represents all voxels in the three dimensional physical space flattened into one long vector, and the other dimension represents time points; each entry in the matrix corresponds to the BOLD signal intensity of a single voxel at a particular time point. The actual number of voxels depends on the resolution of fMRI scanner; modern hardware with a resolution on the order of *2-5mm*^{3} can capture hundreds of thousands of voxels. Similarly, the number of time points depends on the hardware and overall duration of the experiment (how long a subject was in the fMRI). Modern scanners are capable of capturing whole brain volumes at a frequency of *0.75Hz - 2Hz*.

Although the resolution of each voxel is on the order of millimeters, each voxel contains tens of thousands of neurons. For this reason they can be thought of as a "mesoscale" representation; it lies between the microscale of neurons and the macroscale of brain lobes.

Figure 1: Siemens 3T Magnetom Prisma functional Magnetic Resonance Imaging scanner located in the Robarts Research Institute at the University of Western Ontario.

Tasks performed for the Human Connectome Project's tfMRI data include: Emotion Processing (176 time points), Gambling (253 time points), Language (316 time points), Motor (284 time points), Relational Processing (232 time points), Social Cognition (274 time points), and Working Memory (405 time points). fMRI time series data for all tasks was recorded at a temporal resolution of *720ms* per sample; a whole brain volume was captured at roughly *1.389Hz*.

The data was *z-score* normalized and segmented into meaningful regions of interest (ROIs) (refer to Figure 2) with Craddock et al.'s *spatially constrained parcellation* [6]. Each voxel's activation within each ROI was averaged to determine the respective ROI's mean BOLD signal intensity. A variety of resolutions were explored and it was determined that using *30 ROIs* consistently resulted in high quality models, although it is expected that higher or lower resolutions would work similarly well in general. Table 1 provides details on the neuroanatomy of the 30 ROIs.

Ultimately the data was represented as a two dimensional matrix with 30 columns (ROI average BOLD intensity) and t rows, where t is the total number of time points for the specific task.

3 NEUROSCIENTIFIC MOTIVATION

The motivation for generating a model of the brain is to better understand the system. If one wants to study the execution of some brain function, they will generate a model based on appropriate recorded data which can be used to interpret the physiological phenomenon by revealing which areas of the brain are involved, and to what extent.

Task based fMRI studies employ linear methods to generate their models, such as *Pearson product-moment correlation* and the *Generalized Linear Model* (GLM).

For example, if one wants to derive how *ROI 21* is related to the other ROIs, the first step would be to calculate the correlations in the data. Once the correlation coefficients are calculated, correction for multiple comparisons is performed with a method such as

[1] http://www.humanconnectome.org/

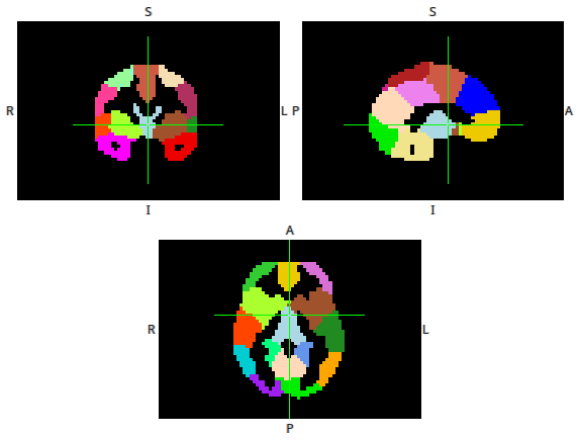


Figure 2: A snapshot of a brain when segmented into the 30 ROIs. Each color represents a different region.

Table 1: Neroanatomical regions and their corresponding segmented regions of interest. This list provides a frame for the resolution of the segmentation of the brain.

ROI #	Description
1	Visual (V1)
2	Insula/Medial Temporal (MT)
3	Cuneus
4	Posterior Ventral Temporal
5	Memory
6	Prefrontal Cortex (PFC)
7	Temporal Pole/Amygdala
8	Auditory (Middle/Lateral Temporal)
9	Intraparietal
10	Insula/Medial Temporal (MT)
11	Cerebellar
12	Thalamys/Midbrain
13	Intraparietal/Calculations
14	Prefontal/Orbitalfrontal Cortex (OFC)
15	Temporal Pole/Amygdala
16	Language Associated Prefrontal Cortex
17	Fusiform/Ventral Temporal
18	Prefrontal Cortex (PFC)
19	Lateral Occipital
20	Auditory (Middle/Lateral Temporal)
21	Medial Frontal/M1 area
22	Somatosensory/Premotor (M1/S1)
23	Somatosensory/Premotor (M1/S1)
24	Fusiform/Ventral Temporal
25	Lateral Occipital
26	Cingulate
27	Medial Orbitalfrontal Cortex (OFC)
28	Prefontal/Orbitalfrontal Cortex (OFC)
29	Language Associated Prefrontal Cortex
30	Anterior Cingulate Cortex (ACC) & Prefontal

false discovery rate or *Bonferroni correction*. Thresholding is then used to eliminate statistically unrelated ROIs; any number of ROIs could be used in the creation of the linear models, however it is ideal to use only *meaningful* ROIs. Here lies one of the problems — what does it mean for an ROI to be *meaningful*? Because each ROI is a part of one larger connected system that was measured at the same time under the same circumstances with the same environmental noise factors, many of the ROIs tend to have reasonably high correlation coefficients. Even after thresholding, one is typically left with a large number of ROIs being statistically related — sometimes even all ROIs. Although it is possible the whole brain is involved with the task meaningfully, it would seem unlikely that the relationships are truly significant.

Once the set of ROIs for the linear model are selected with thresholding, linear regression is used to derive the model by fitting the remaining ROIs to ROI 21.

These methods make many incorrect assumptions: the system is linear, ROIs are treated like fixed values as opposed to random variables (weak exogeneity), constant variance in the data, independence of errors, and a lack of multicollinearity.

By using a different method, such as symbolic regression, which is at least as powerful as linear regression and capable of eliminating many of the assumptions linear regression makes, it may be possible to develop more accurate and descriptive models of the complex system. Although the computational cost of employing symbolic regression is extensive, it is well worth the price if the generated models do help in the discovery of the true underlying phenomenon.

4 GENETIC PROGRAMMING IMPLEMENTATION

The GP implementation used was inspired by Schmidt et al.'s work [25]. It is specialized for symbolic regression and incorporates modular improvements which significantly increase the performance of the search. Some of these improvements include parallel evolution of subpopulations, fitness predictors [23, 24], and an acyclic graph representation [21].

Multiple subpopulations are evolved separately to encourage a well diversified set of candidate solutions. Periodically these subpopulations are combined, shuffled, and redistributed into new subpopulations. This procedure encourages genetic diversity over the whole population by allowing the subpopulations to traverse the search space along separate trajectories.

Fitness predictors have been demonstrated to reduce computational cost by approximating the local search gradient [23, 24]. Candidate solutions are evaluated with a subset of data that is both representative of the whole dataset, and creates a large variance among candidate solutions' fitness. In other words, a subset of data points are selected in a way to emphasize the evolution on relatively small areas of the search space where candidate solutions need the most improvement. There are many successful similar techniques [15], however this particular approach was selected as it has many benefits, such as reducing computational cost, reducing overfitting, and focuses the search on key features, an idea shown to significantly improve symbolic regression [24].

Traditional systems typically represent expressions in tree structures where leaf nodes are terminals (constants or variables) and internal nodes are operators. In this work an *acyclic graph representation* is used. Figure 3 presents an example of an acyclic graph representation with an array encoding. Other graph representations

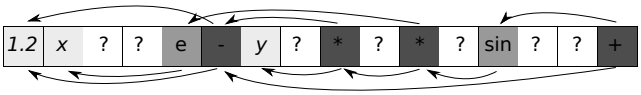


Figure 3: An array encoding for the expressions $(1.23 - x) + Sin((1.23 - x) * y * e^x)$. Sub-expressions can be referenced multiple times by any number of operators in a higher index. '?' represent information not expressed in the phenotype, however they may contain *vestigial* sub-expressions [21].

exist in the literature, however the acyclic graph representation is unique, and motivations for selecting this representation were: first, the encoding is lightweight, second, the representation scales well and avoids bloat, and third, it can reuse important subexpressions effectively [21].

4.1 Execution of Evolutionary Search

An overview of the execution flow of the whole GP system is presented in Figure 4 — full details can be found in Schmidt et al.'s work [25]. The left side of Figure 4 depicts the evolution of the population of candidate solutions (mathematical expressions in this particular case) while the right side depicts the execution of the evolution of fitness predictors (in this case, a collection of subsets of data points from the recorded task based fMRI time series).

For initialization, and after a population of candidate solutions is generated, a set of fitness predictors and fitness trainers (a subset of candidate solutions used to evaluate the quality of the fitness predictors [24]) are generated. Following this, the whole population of candidate solutions is split into an arbitrary number of subpopulations to be evolved in parallel.

Evolution of the subpopulations occurs in parallel, however, evaluation of the subpopulations is calculated with the subset of data points from the fitness predictors, which is always changing as it is evolving alongside the subpopulations.

After a predefined number of iterations the subpopulations are recombined, and if some stopping criteria is met, the execution completes. If execution is not complete, the combined subpopulations are shuffled and split once again into subpopulations. Just as it was done earlier, the fitness trainers and predictors are updated based on the current whole population before it is split into subpopulations.

4.2 Experimental Methods

For symbolic regression, it is required to have some value over the time series that the evolved expressions fits to. This value is chosen to be an ROI that is already known to be involved in the specific tasks. For example, *ROI 21* was selected as the left hand side of the equation for the motor task as it is the ROI containing the *primary motor cortex*. The left hand sides of the equations for the emotion, gambling, language, motor, relational, social, and working memory tasks were ROIs 7, 2, 12, 21, 28, 3, and 21 respectively. All remaining ROIs were given to the GP system and acted as variables over the time series and the expressions evolved to equate these variables to the ROI on the left hand side of the equation. Although these values are being equated to a specific ROI, the actual expressions generated can be thought of as a network relationship.

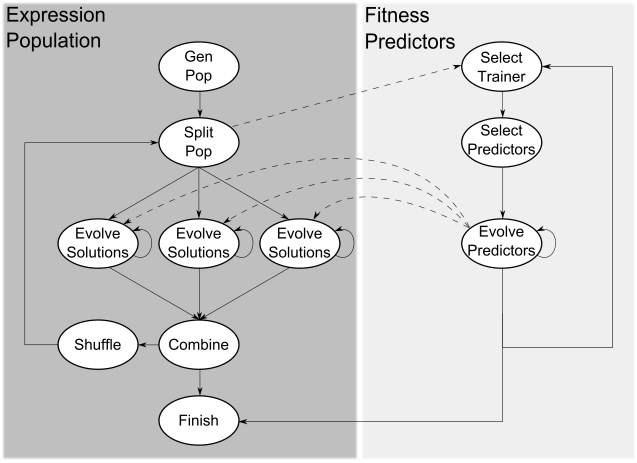


Figure 4: High level structure of this symbolic regression implementation. The left side demonstrates the evolution of the expressions while the right side depicts the evolution of fitness predictors. This example shows only three subpopulations evolving in parallel.

Ten subjects for which a complete set of data existed for all *7* tasks from the Human Connectome Project were studied. For each of these subjects, the search for expressions describing the data was executed *100* times in an effort to improve the significance of results. A total of *7000* evolutionary searches were performed.

The language/basis functions used for the experiments included unary and binary linear and nonlinear operators that can be observed in nature. These operators include: $+$, $-$, $*$, $/$, e, abs, sin, cos, and tan. The inclusion of the arithmetic operators allow the search to be at least as powerful as linear regression. e allows for exponentiation, and absolute value introduces point nonlinearities. Trigonometric operators are included as they introduce simple nonlinearities, and we know from Fourier that any periodic function can be expressed as a sum of sine waves.

It should be emphasized here that using symbolic regression to model fMRI time series does require the acceptance of one assumption: it is assumed that the language provided to the system is sufficient to describe the data.

Mean Squared Error was used for the fitness function.

7 individual populations of *101* candidate solutions were evolved in parallel. The choice of *7* populations was because the evolutionary search was being executed on *8* core systems, and with the addition of fitness predictors evolving on a single core, a total of 8 threads were effectively utilized: 1 thread per core. *10,000,000* generations were done with shuffles of the subpopulations and updates of fitness trainers occurring every *1,000* generations. This all results in a total of *7,070,000,000* mating events for every model. It should be noted that these values are excessive and a reduction by orders of magnitude has little impact on the models, however it is crucial to note that the goal is to find the best possible representative explanation of the underlying nonlinear system to further our understanding of the brain — predictive models are *not* the goal.

Even after apparent convergence, any marginal improvement may be important for describing the underlying complex system.

The maximum number of *unique* operators/operands in the acyclic graph representation for a candidate solution was set to *40*, however the actual number of operators and operands in the expression can be higher as subexpressions are reused.

The crossover and mutation rates were set to *80%* and *10%* respectively. *Two* mutations were possible per candidate solution for each propagation to the next generation. The mutation rate was set high as the encoding (Figure 3) makes it so mutations may have no effect on the phenotype.

The number of fitness trainers was *8*, the number of fitness predictors was *20*, and the number of data points per fitness predictor was set to *25%* the size of the total number of data points.

Keeping in mind the stochastic nature of the algorithm and the varying amounts of data between each task, each run of the evolutionary search using this GP system takes between *5* and *24* hours (in the most extreme cases) when running with 8 cores on an *IBM System x iDataPlex dx360 M3* node with 2 quad-core *Intel Nehalem (Xeon 5540)* processors running at 2.53GHz.

5 RESULTS AND DISCUSSION

Symbolic regression is model free: it selects which ROIs to fit the data with and how they relate to one another. The evolutionary search selected many fewer ROIs than any of the typical correction for multiple comparison and thresholding techniques would for the linear models, demonstrating that nonlinear models are able to describe the system with much less information.

For a thorough comparison to linear models, multiple techniques were used to correct for multiple comparisons before thresholding ROIs. This was done for the linear models as their effectiveness depends on how many ROI are used to fit the data, and different techniques can result in a different set of ROIs. Two popular techniques within the neuroscience literature, Bonferroni Correction, which is known to be conservative (higher false negative rate), and False Discovery Rate (with familywise error rate $\alpha = 0.05$) were employed. In almost all cases, Bonferroni Correction would result in fewer ROIs than False Discovery Rate. Two other ideas were also included: forcing linear models to have the same number of ROIs as the nonlinear models (typically between 7 to 10 ROIs), and allowing linear models to fit the data with all ROIs.

Table 2 provides a summary of the mean absolute error of the models averaged over all subjects per task and the corresponding standard deviations. A Mann-Whitney U test p-value is included comparing the ten subjects' corresponding linear models' mean absolute error values to the nonlinear models' (first columns). Note that unlike the linear models, the nonlinear models are generated stochastically so only the top nonlinear model for each subject's 100 models is used in the comparison. Linear models improve as they are allowed to fit the data with more ROIs; however, even with all 30 ROI, *the linear models never perform significantly better than the nonlinear.*

It is possible that although the nonlinear models contain fewer ROIs, and fit the data well, they may not truly be meaningful or novel. Additionally, nonlinear models — despite using much less information than linear models to fit the data — may still overfit the recorded fMRI signal.

Figure 5 shows how the Mann-Whitney U test's probability value between the linear and nonlinear models' mean absolute error distributions change as the number of ROIs used in the linear model increase. This novel representation of the data was created to better understand how the number of ROIs impacts the quality of the linear models. Columns represent tasks, rows show the number of the top linearly correlated ROIs used in the linear model (including the left hand side of the equation — the first row with the number 2 uses only the top 1 linearly correlated ROI to fit the data), and color represents the probability value calculated comparing the fixed, top performing nonlinear models for each subject to the linear models fit with the corresponding row's number of ROIs. The white text on the plot show the average number of ROIs for all 100 nonlinear models generated per subject (*NL-A-*), the average number in the top nonlinear models per subject (*NL-T-*), how many when Bonferroni Correction was used before thresholding selected (*BC-*), and how many with False Discovery Rate (*FDR-*).

Figure 5 unsurprisingly shows that, in general, linear models improve as the number of ROIs used to fit the data increases. However, the nonlinear models had smaller mean absolute error values than the linear models up until the linear models were fit with at least 20 ROIs; *more than twice as many ROIs were needed by the linear models to explain the underlying system as well as the nonlinear.* In each column, the reddest area shows where the models were nearly indistinguishable, and any point below shows when linear models obtained better mean absolute errors. Although the linear models, once given enough ROIs, had better average mean absolute error values, they were *never significantly better.* The closest linear models were to being significantly better was when all ROIs were used to fit the language task's data (*p-value of 0.07*). Refer to the last column in Table 2 for the p-values comparing nonlinear models to linear models generated with all ROIs.

Figure 5 also shows that the number of ROIs required for linear models to become comparable to nonlinear models differs between task, suggesting the level of nonlinearity in the functional connections is dependent on the task being performed by the system. Although this phenomenon seems rather obvious, it would be unaccounted for with traditional tools. It is not possible to demonstrate this variation without a tool that can infer model structure, such as symbolic regression.

Figure 6 shows a comparison of a nonlinear model along with two linear models against the actual recorded signal for the ROI being fit to (ROI 7 for the Emotion task). The nonlinear model, which contained *only 8 ROIs* (including ROI 7 — the left hand side of the equation) in this example, fit the data best with a mean absolute error of *0.14.* When a linear model contained only 8 ROIs (the same number the nonlinear model used) it was only able to fit the data with a mean absolute error of *0.27.* Even when the linear model was given all ROIs, it was only able to achieve a mean absolute error of *0.15.* In this particular case, using Bonferroni correction resulted in no ROI being eliminated (same model as using all ROI) and False Discovery Rate resulted in 28 ROI being fit to the data and achieved a mean absolute error of *0.16* (both not shown).

Models generated with symbolic regression were highly nonlinear and could not possibly be created with linear regression. Not

Table 2: summary of the top nonlinear models and linear models with different correction for multiple comparison and thresholding techniques. *MAE* is the averaged mean absolute error over all subjects for each task and the probability value (*p-val.*) was calculated with a Mann-Whitney U test between the nonlinear models and the respective column's linear model.

Task	Nonlinear		Same # ROIs			Bonferroni Correction			False Discovery Rate			All ROIs		
	MAE	Std.	MAE	Std.	p-val.	MAE	Std.	p-val.	MAE	Std.	p-val.	MAE	Std.	p-val.
EMOTION	0.34	0.11	0.42	0.12	0.06	0.36	0.14	0.45	0.33	0.12	0.40	0.31	0.10	0.17
GAMBLING	0.31	0.08	0.36	0.09	0.09	0.30	0.08	0.43	0.30	0.08	0.37	0.30	0.08	0.34
LANGUAGE	0.29	0.06	0.39	0.09	0.00	0.29	0.06	0.26	0.28	0.06	0.14	0.27	0.05	0.07
MOTOR	0.24	0.05	0.32	0.05	0.00	0.23	0.05	0.48	0.23	0.04	0.37	0.22	0.04	0.15
RELATIONAL	0.23	0.08	0.31	0.10	0.04	0.22	0.07	0.37	0.21	0.07	0.31	0.21	0.07	0.29
SOCIAL	0.30	0.05	0.42	0.09	0.01	0.32	0.08	0.31	0.30	0.06	0.48	0.28	0.05	0.21
WM	0.27	0.10	0.32	0.10	0.06	0.26	0.10	0.24	0.25	0.09	0.24	0.25	0.09	0.21

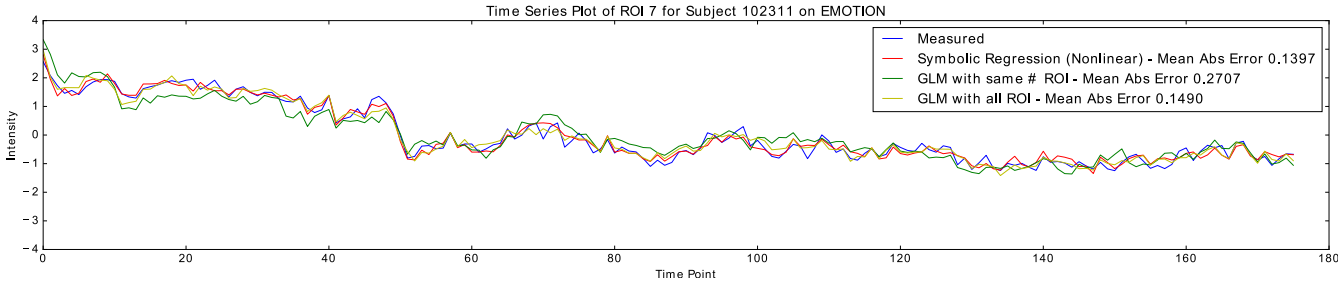


Figure 6: Nonlinear and Linear models expected ROI intensity value compared to the measures signal.

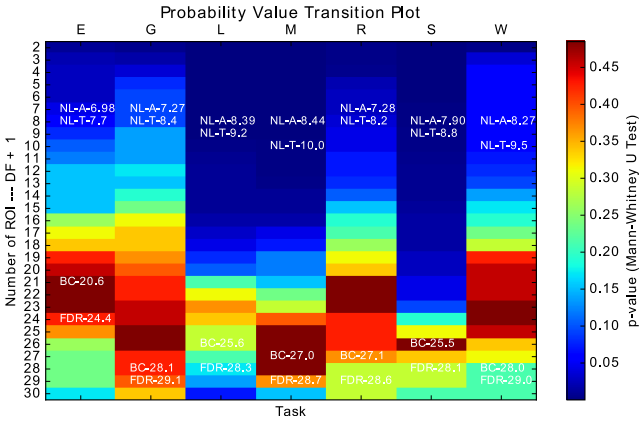


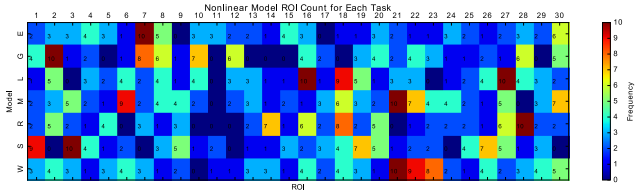
Figure 5: Probability value transition plot between the linear and nonlinear models' mean absolute errors (averaged over all subjects per task) as the number of the top linearly correlated ROIs used to fit the data with linear regression is increased. The number of ROIs used in the nonlinear models was fixed as the number of ROIs linear models used were increased.

only were the relationships between the ROIs novel, but the ROIs selected by symbolic regression would sometimes differ from the top correlated ROIs. Figure 7 shows which ROIs are most important for the respective models and Figure 7c emphasizes where the ROIs differ by showing the top linearly correlated ROIs (which would be found in the linear models) minus the ROIs selected by symbolic regression. For example, ROI 27 was in all nonlinear models for the language task, however Figure 7b shows ROI 27 being one of the least linearly correlated ROIs. These differences demonstrate that linear correlation may be inadequate for accurately modeling functional connectivity. Just because an ROI is, or is not, highly correlated, does not mean that it is, or is not, meaningful in the function of the underlying system; high correlation between ROI time series does not imply the existence of a true functional relationship between them.

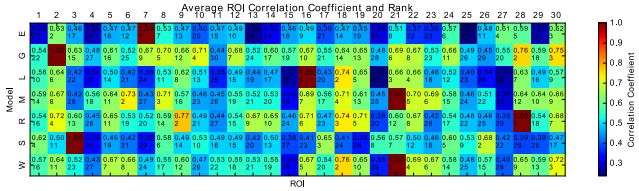
A similar argument could be made against symbolic regression: just because an ROI is, or is not, selected by the evolutionary search, does not mean that it is, or is not, meaningful. However symbolic regression is at least as powerful as linear regression; symbolic regression can exploit any linear relationships within the data that traditional linear tools are limited to, while also capable of discovering more complex nonlinear relationships. Although it is possible the ROIs in the nonlinear models may not truly be meaningful, they are at least selected with fewer assumptions, namely, without the significant incorrect assumption that the system is linear.

Symbolic regression was executed with the only constraint being that the data could be modeled sufficiently with the language (mathematical operators) provided to the GP system. This assumption can negatively affect the models as the language may not be descriptive enough to accurately model the functional networks, however the language used by symbolic regression is more expressive than the limited operators used for linear regression — which were also included in the language.

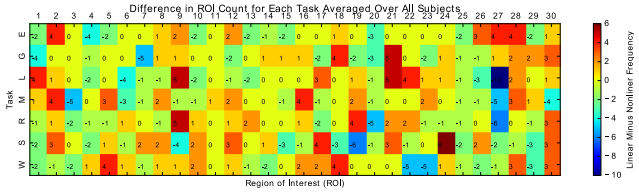
It is possible that the data recorded by fMRI is not truly representative of the underlying system, which could result in nonlinear models not accurately describing the functional connectivities. However this limitation of the hardware and our imperfect ability



(a) Number of times every ROI appeared in top models for each subject on all tasks. The ROI being fit to (left hand side of equation) is forced to be in the expression. For example, the Motor task was fitting to ROI 21, therefore ROI 21 was in all models for the Motor task.



(b) Every ROI's average absolute correlation coefficient's for each subject on all tasks along with that ROIs rank. The ROI being fit to (left hand side of equation) will have a correlation coefficient of 1.



(c) Difference between ROI counts when linear models were forced to have the same number of ROIs as nonlinear models (*linear - nonlinear*). For example, if a nonlinear model contained 7 ROI, the top 7 linearly correlated ROI would be used. Highly negative values mean nonlinear models were more likely to select that ROI, and highly positive values were when nonlinear models were less likely to use the ROIs.

Figure 7: Representations of where the linear and nonlinear models disagree on what ROIs are important in describing the system.

to measure functional brain activity affects all data driven techniques. Additionally, if fMRI is only capable of extracting linearities from a system we know to be nonlinear, then the technology itself is insufficient for the understanding the system under study.

As the tools being used become more sophisticated and able to describe the data in more complex ways, overfitting becomes a concern. Although nonlinear models only used between 7 and 10 ROIs to fit the data, and linear models used up to 30 (markedly more degrees of freedom), it is important to keep in mind that symbolic regression is a powerful tool susceptible to overfitting.

Figure 8 shows three error matrices with similar values. The generated models for each subject and task were applied to every other subject performing every other task. The mean absolute error was recorded and the error was then averaged over all subjects within the same task. The nonlinear models, as shown by the averaged mean absolute errors in th left matrix generalize reasonably well across all subjects performing the same task. The left matrix also has a similar set of results as the linear models (The center and right matrices), showing that the nonlinear models generalize similarly well to other subjects as the linear models. Despite being more complex, these current results provide some evidence that the nonlinear models are not overfitting the data significantly, or at least, not overfitting the data any more than linear models.

6 CONCLUSIONS AND FUTURE WORK

A highly specialized GP system designed for symbolic regression was implemented to search for nonlinear relationships in a dynamic complex nonlinear system: the human brain. Task based fMRI data was effectively modeled with symbolic regression, and not only were novel nonlinear relationships found, but when compared to linear tools, a different and much smaller set of ROIs were selected as meaningful. These relationships and ROIs could not possibly be discovered with conventional tools — findings that provide new insight into the underlying system. Despite the parametric complexity of symbolic regression, results suggest that significant overfitting was unlikely.

Nonlinear models described the system well, and performed significantly better than linear models in many cases; however, as linear models were given more ROIs to fit the data, their effectiveness surpassed nonlinear models — *although never significantly.*

Nonlinear models required markedly fewer ROIs to describe the system than linear models, which contained nearly all ROIs. "Verbose models" are difficult to interpret in a meaningful way, so succinct models may lead to better insight into the system being modeled. Additionally, we know that linear correlation is not an effective means for describing a system we know to be nonlinear.

As one should expect, our results demonstrated that the level of nonlinearity required to describe the functional networks within the system depended on the task being performed by the subject.

This work can be continued in multiple directions. Current work is being done to explore additional subjects to expand the significance of the findings. Further work is currently being done to better understand the nonlinear models and overfitting. Although the goal of this work is not to generate predictive models, exploration into creating such models can be done with the goal of finding more general models. Performing symbolic regression on other neuroimaging modalities where nonlinearities are commonly observed, such as electroencephalogram or Magnetoencephalography, would allow for a comparison to the novel nonlinear models generated for fMRI data.

The investigation of what neuroscientific questions can now be answered with these novel discoveries, and the deployment of a symbolic regression tool for neuroscientists are current priorities.

7 ACKNOWLEDGMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Computations were performed on the General Purpose Cluster supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

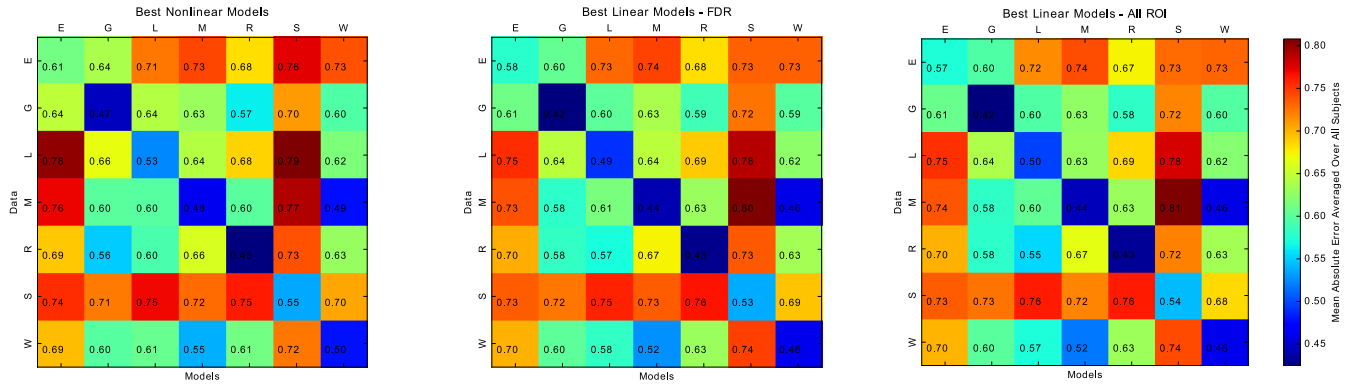


Figure 8: From left to right, the best mean absolute error values averaged over all subjects when performing the same task for nonlinear models, linear models generated with false discovery rate, and linear models generated with all ROIs respectively. For example, row *M* and column *L* contains the averaged mean absolute error values of the models for all subjects fit to the language task, but applied to the motor task's data.

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada.

Computations were made on the supercomputer Guillimin from McGill University and Mammouth Parallèle from Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

REFERENCES

[1] Nicholas Allgaier. 2015. Reverse Engineering the Human Brain: An Evolutionary Computation Approach to the Analysis of fMRI. (2015).

[2] Nicholas Allgaier, Tobias Banaschewski, Gareth Barker, Arun LW Bokde, Josh C Bongard, Uli Bromberg, Christian Büchel, Anna Cattrell, Patricia J Conrod, Christopher M Danforth, and others. 2015. Nonlinear functional mapping of the human brain. *arXiv preprint arXiv:1510.03765* (2015).

[3] Josh Bongard and Hod Lipson. 2007. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* 104, 24 (2007), 9943–9948.

[4] Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. *The journal of neuroscience* 16, 13 (1996), 4207–4221.

[5] R.L. Buckner and T.S. Braver. Event-Related Functional MRI. In *Functional MRI*, P. Bandettini and C. Moonen (Eds.). Springer-Verlag, Chapter 36, 441–452.

[6] R Cameron Craddock, G Andrew James, Paul E Holtzheimer, Xiaoping P Hu, and Helen S Mayberg. 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human brain mapping* 33, 8 (2012), 1914–1928.

[7] Mark Daley. 2014. An Invitation to the Study of Brain Networks, with Some Statistical Analysis of Thresholding Techniques. In *Discrete and Topological Models in Molecular Biology*. Springer, 85–107.

[8] Karl J Friston, Lee Harrison, and Will Penny. 2003. Dynamic causal modelling. *Neuroimage* 19, 4 (2003), 1273–1302.

[9] Karl J Friston, Oliver Josephs, Geraint Rees, and Robert Turner. 1998. Nonlinear event-related responses in fMRI. *Magnetic resonance in medicine* 39, 1 (1998), 41–52.

[10] Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price. 2000. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *Neuroimage* 12, 4 (2000), 466–477.

[11] Karl J Friston, CJ Price, Paul Fletcher, C Moore, RSJ Frackowiak, and RJ Dolan. 1996. The trouble with cognitive subtraction. *Neuroimage* 4, 2 (1996), 97–104.

[12] Scott A Huettel, Allen W Song, and Gregory McCarthy. 2009. *Functional magnetic resonance imaging* (second ed.). Vol. 1. Sinauer Associates Sunderland, MA.

[13] James Alexander Hughes and Mark Daley. 2016. Finding Nonlinear Relationships in fMRI Time Series with Symbolic Regression. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*. ACM, 101–102.

[14] Ilknur Icke, Nicholas A Allgaier, Christopher M Danforth, Robert A Whelan, Hugh P Garavan, and Joshua C Bongard. 2014. A deterministic and symbolic regression hybrid applied to resting-state fmri data. In *Genetic Programming Theory and Practice XI*. Springer, 155–173.

[15] Yaochu Jin. 2005. A comprehensive survey of fitness approximation in evolutionary computation. *Soft computing* 9, 1 (2005), 3–12.

[16] John R Koza. 1992. *Genetic programming: on the programming of computers by means of natural selection*. Vol. 1. MIT press.

[17] F Kruggel, Stefan Zysset, and D Yves von Cramon. 2000. Nonlinear regression of functional MRI data: an item recognition task study. *Neuroimage* 12, 2 (2000), 173–183.

[18] Nikos K Logothetis. 2008. What we can do and what we cannot do with fMRI. *Nature* 453, 7197 (2008), 869–878.

[19] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 6843 (2001), 150–157.

[20] Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. 1992. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences* 89, 13 (1992), 5951–5955.

[21] Michael Schmidt and Hod Lipson. 2007. Comparison of tree and graph encodings as function of problem complexity. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*. ACM, 1674–1679.

[22] Michael Schmidt and Hod Lipson. 2009. Distilling free-form natural laws from experimental data. *science* 324, 5923 (2009), 81–85.

[23] Michael D Schmidt and Hod Lipson. 2007. Coevolving fitness models for accelerating evolution and reducing evaluations. In *Genetic Programming Theory and Practice IV*. Springer, 113–130.

[24] Michael D Schmidt and Hod Lipson. 2008. Coevolution of fitness predictors. *Evolutionary Computation, IEEE Transactions on* 12, 6 (2008), 736–749.

[25] Michael D Schmidt, Ravishankar R Vallabhajosyula, Jerry W Jenkins, Jonathan E Hood, Abhishek S Soni, John P Wikswo, and Hod Lipson. 2011. Automated refinement and inference of analytical models for metabolic networks. *Physical biology* 8, 5 (2011), 055011.

[26] Alberto L Vazquez and Douglas C Noll. 1998. Nonlinear aspects of the BOLD response in functional MRI. *Neuroimage* 7, 2 (1998), 108–118.

[27] Tingting Zhang, Fan Li, Marlen Z Gonzalez, Erin L Maresh, and James A Coan. 2014. A semi-parametric nonlinear model for event-related fMRI. *Neuroimage* 97 (2014), 178–187.

Chapter 6

Paper 3

This paper was submitted to *Association for Computing Machinery*'s (ACM) *Genetic and Evolutionary Computation Conference* (GECCO) 2018. An abstract based on this work was submitted and accepted to the *12th Annual Canadian Association of Neuroscience Meeting 2018* [58]. References contained within this article are numbered according to the article's bibliography.

Similar to Article 2, the article states "It is possible that the entire brain is involved in the task being studied, but this seems unlikely.", which is not an empirically demonstrated fact, and is a current discussion within connectomics.

Generalizability of Nonlinear Models of Functional Connectivity

James Alexander Hughes
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3M1
jhughe54@uwo.ca

Mark Daley
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3M1
mdaley2@uwo.ca

ABSTRACT

The brain is an intrinsically nonlinear system; however, most methods employed in studying functional magnetic resonance imaging data (fMRI) produce strictly linear models — models incapable of truly describing the underlying system.

In this work, genetic programming is used to develop nonlinear models of functional connectivity from fMRI data. The study builds on previous work and observes that nonlinear models contain relationships not found by traditional linear methods. When compared to traditional methods, nonlinear models are more succinct and are never significantly worse when applied to data the models were fit to. It was also observed that the nonlinear models could not generalize to other subjects as well, but would generalize to unseen data from the same subject better than the linear models. This study presents the problem that many, manifestly different models in both operators and features, can effectively describe the system with acceptable metrics.

CCS CONCEPTS

•**Computing methodologies** → **Genetic programming;** *Modeling methodologies;* •**Applied computing** → **Systems biology;**

KEYWORDS

Symbolic regression; Computational neuroscience; Functional magnetic resonance imaging; Nonlinear Modelling.

ACM Reference format:
James Alexander Hughes and Mark Daley. 2018. Generalizability of Nonlinear Models of Functional Connectivity. In *Proceedings of GECCO '18, Kyoto, Japan, July 15-19, 2018,* 8 pages.
DOI: somenumber

1 INTRODUCTION

Almost all methods typically used for modeling functional Magnetic Resonance Imaging (fMRI) data produce models with linear tools (Pearson Product-moment correlation coefficient, general linear model) despite the brain being a nonlinear system and the literature acknowledging the existence of nonlinearities [3, 4, 6, 8, 10, 24]. Although the field has made many contributions using linear tools, they may not be powerful enough to truly capture the underlying nonlinearities that are known to exist within the system.

The benefit of using linear tools is that they, and the models they produce, are easily understood; often, simpler tools and models are desirable. Finding nonlinearities is a non-trivial task, especially when faced with large amounts of high-dimensional data. Sophisticated nonlinear tools introduce more degrees of freedom, are more computationally expensive, and in many cases, produce hard to interpret models. However, sometimes these sacrifices are required to understand the intricacies of a nonlinear dynamic complex system, such as the brain.

Although nonlinear tools have been developed and studied, they remain under-represented within the neuroscience literature. Friston et al. use *Volterra series expansion* to study the balloon model [8, 9] and *dynamic causal modelling* to study effective connectivity [7]. Kruggel et al. used a form of *nonlinear regression* to model relationships between the hemodynamic response and stimulus conditions [17]. *Semi-parametric Volterra series* was used by Zhang et al. to find nonlinearities in fMRI data [25]. With *symbolic regression*, Allgaier et al. found novel nonlinear relationships within known networks in *resting state* fMRI data [1, 2]. Hughes & Daley used *symbolic regression* to develop nonlinear models of *task based* fMRI data and found that, when compared to linear models, they were more succinct and fit the data better [12, 13].

In this work *Genetic Programming* (GP) will be used to perform *Symbolic Regression.* GP is an artificial intelligence technique where, through a strategy based on the natural process of evolution, the algorithm writes (*evolves*) its own programs to solve problems [16]. Symbolic regression is a regression technique that not only searches for coefficients, but also for the structure of the model. This allows for a more powerful regression capable of finding nonlinear relationships with fewer assumptions, when compared to typical linear regression. Since we are using GP for symbolic regression, the *programs* being written by GP are *mathematical expressions.*

We build upon the work of Hughes & Daley [12, 13]; we develop nonlinear models of fMRI data gathered from real subjects performing a variety of tasks. These models provide interpretable functional network relationships between brain areas to ultimately give new insight into the underlying system. The goal is *not* to create predictive models, but to develop *descriptive* models; the goal is the generation of interpretable model, not to collect the model's output. We expand upon the previous work by including additional subjects for greater insight and statistical significance. We also expand the analysis by applying unseen data from the same subject performing the same task to the developed models.

High quality nonlinear models were generated that generalized to unseen form the same subject, however the nonlinear models could not generalize to unseen data from different subjects as well

GECCO '18, Kyoto, Japan

no clue yet. . . $15.00
DOI: somenumber

as the traditional linear models. We finish with remarks on the difficulty of model selection when presented with a collection of different models with similar error values.

2 NEUROSCIENTIFIC DATA

The data studied in this work was obtained from the *Human Connectome Project, WU-Minn Consortium*[1] (HCP) — a large, noteworthy publicly available bank of neuroscientific data. As of April 2017, the project contains data from 1200 subjects. Although a variety of types of data is included in the dataset, we focus on *task based fMRI* recordings. This data is recorded from subjects performing a task while inside fMRI scanners. The fMRI scanner records the blood oxygen level dependent (*BOLD*) signal — a measure of the relative oxygenation level of blood within tissue — which can be used as an effective *proxy* for functional activation [11, 18, 19]; however, the precise nature of the BOLD signal is not well understood [6]. The BOLD signal is notoriously noisy, spatially diffused, and lags behind actual neural activity.

The BOLD signal from the three-dimensional brain over time is recorded and represented as voxels (three-dimensional analogues to two-dimensional pixels). This four dimensional data (Figure 1) can be represented as a two dimensional matrix of voxels by flattening the three-dimensional physical space into one long vector and treating time as the second dimension. Each entry in the two-dimensional matrix corresponds to the BOLD signal intensity of a single voxel at some time point. Although each voxel's size is on the order of millimetres, they contain tens of thousands of neurons.

The *seven* tasks performed by subjects for the HCP's fMRI recordings include: Emotion Processing (176 time points/127s), Gambling (253 time points)/182s), Language (316 time points/228s), Motor (284 time points/204s), Relational Processing (232 time points/167s), Social Cognition (274 time points/197), and Working Memory (405 time points/292s). The temporal resolution of the scans were *720ms* per sample.

Data from all tasks for *forty* subjects were analyzed in this work. Each subject had two separate recordings for each task (one left-to-right (LR) phase encoding direction, and one right-to-left (RL) — the direction of applied gradient required for fMRI data acquisition [11]). The LR phase encoding data was used as the *training* data, and the RL was used as independent *testing* data.

Data was *z-score* normalized and segmented into 30 regions of interest (ROIs) with Craddock et al.'s *spatially constrained parcellation* [5]. Each ROI's value is the mean BOLD signal from all voxels within it. Multiple resolutions were explored and *30 ROIs* consistently produced high quality models. A high level overview of the ROIs can be found in Table 1.

After preprocessing, the data was represented as a two dimensional matrix of 30 columns of ROI average BOLD intensities and t rows, where t is the number of time points for a given task.

3 NEUROSCIENTIFIC MOTIVATION

Neuroscientists generate models of the brain to better understand the underlying system. If we generate effective models, we can

[1] http://www.humanconnectome.org/

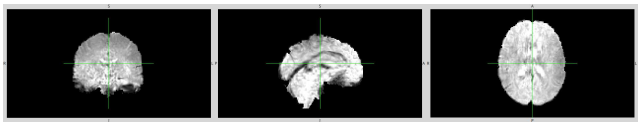


Figure 1: A three-dimensional snapshot of the four-dimensional fMRI data. The voxels in this brain contain the BOLD signal from a single time point.

Table 1: Region of interest number and corresponding neuroanatomical region. This table provides a frame for the resolution of the brain segmentation.

Region of Interest #	Description
1	Visual (V1)
2	Insula/Medial Temporal (MT)
3	Cuneus
4	Posterior Ventral Temporal
5	Memory
6	Prefrontal Cortex (PFC)
7	Temporal Pole/Amygdala
8	Auditory (Middle/Lateral Temporal)
9	Intraparietal
10	Insula/Medial Temporal (MT)
11	Cerebellar
12	Thalamys/Midbrain
13	Intraparietal/Calculations
14	Prefontal/Orbitalfrontal Cortex (OFC)
15	Temporal Pole/Amygdala
16	Language Associated Prefrontal Cortex
17	Fusiform/Ventral Temporal
18	Prefrontal Cortex (PFC)
19	Lateral Occipital
20	Auditory (Middle/Lateral Temporal)
21	Medial Frontal/M1 area
22	Somatosensory/Premotor (M1/S1)
23	Somatosensory/Premotor (M1/S1)
24	Fusiform/Ventral Temporal
25	Lateral Occipital
26	Cingulate
27	Medial Orbitalfrontal Cortex (OFC)
28	Prefontal/Orbitalfrontal Cortex (OFC)
29	Language Associated Prefrontal Cortex
30	Anterior Cingulate Cortex (ACC) & Prefontal

study the models to discover which areas of the brain are *functionally connected*. Although error values can indicate model accuracy, the model itself is of interest, not the output of the models.

Typical task based fMRI studies employ linear methods to generate models. These methods include *Pearson product-moment correlation coefficient* and the *Generalized Linear Model* (GLM). For example, if one wanted to derive how a given *ROI X* was functionally connected to all other ROIs, one would correlate the timeseries from each ROI and do some correction for multiple comparisons (*false discovery rate* (FDR) or *Bonferroni correction* (BC)). Statistically unrelated ROIs are eliminated and the remaining ROIs will be used as regressors in our linear regression to ROI X. The resulting model and beta weights will be used to indicate which areas of the brain are functionally related during a task, and to what extent.

These methods assume that the underlying system is linear, however we know this to be incorrect — the human brain is a nonlinear system. It also treats ROIs as fixed values as opposed to random variables (weak exogeneity). Other assumptions include: constant variance in the data, independence of errors, a lack of multicollinearity, and that the residuals are not autocorrelated.

Thresholding is done to eliminate statistically unrelated ROIs before regression as one would only want to include *meaningful*

ROIs as regressors. However, what does it mean for an ROI to be meaningfully related? All ROIs are apart of a larger, connected system being recorded at the same time, under the same circumstances, in the same environment susceptible to the same noise factors. Ultimately, many ROIs are highly correlated, and even after thresholding, one is typically left with a large number of ROIs being statistically related (sometimes even all). It is possible that the entire brain is involved in the task being studied, but this seems unlikely.

Despite the drawbacks, there are many reasons to use the simple models; complex models tend to overfit data, are hard to interpret, and typically have a much greater computational cost. But, perhaps using a different method, such as symbolic regression, we can develop more accurate and descriptive models of the brain. Symbolic regression is at least as powerful as linear regression, will eliminate many of the assumption the linear methods make, and will perform feature selection.

4 METHODS

4.1 Genetic Programming Implementation

The GP implementation used in this work is based on Schmidt et al.'s work, is specialized for symbolic regression, and incorporates improvements to increase performance [23]. Although many ideas are incorporated into the system, noteworthy ones include fitness predictors [21, 22] and an acyclic graph representation [20]. These ideas are summarized below, but full descriptions are available from their respective sources.

Fitness predictors reduce the computational cost of the search by approximating the local search gradient [21, 22]. Chromosomes are only evaluated on a representative subset of data that emphasizes the search on areas of the space candidate solutions disagree the most — if the population has no consensus on an area, then the search might benefit by focusing on that area. This subset of data is always *evolving* throughout the evolutionary search as the data points required to create the disagreement between candidate solutions will depend on the current population. Although there are similar techniques [15], this method was selected since it not only lowers computational cost by reducing the number of data points needed for evaluation, but it has also been shown to reduce overfitting, focus on key features, and improves results [22].

Unlike typical GP systems which represent the candidate solutions as trees, an *acyclic graph representation* is used in this work as it has a lightweight encoding, scales well, avoids bloat, and has the ability to easily reuse subexpressions. Many graph based encodings exist in the literature, but the implementation described by Schmidt et al. was used for the above reasons [20].

4.2 Genetic Programming Settings

The system parameters used are presented in Table 2.

The mutation rate was set high as a mutation may have no change on the phenotype due to the nature of the acyclic graph encoding.

The language was selected to be at least as powerful as linear regression (arithmetic operators), and to have nonlinear operators: absolute value for point nonlinearity, e for exponentiation, and trigonometric operators since any periodic function can be expressed as a sum of sine waves.

Table 2: Parameter settings for GP System. The last 4 settings are specific to the improvements discussed in 4.1.

Elitism	1
Population	101/subpopulation (707 total)
Subpopulations	7
Migrations	10,000
Generations	1,000 *per* migration (10,000,000 total)
Crossover	80%
Mutation	10% (x2 chances)
Fitness Metric	Mean Squared Error: $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$
Language	$+, -, *, /, exp, abs, sin, cos, tan$
Trainers	8
Predictors	20
Predictor Pop. Size	25% of Dataset
Max # Graph Nodes	140

The choice of *7* subpopulations was because the evolutionary search was performed on systems with *8* core processors, and with the addition of fitness predictors evolving on a single core, a total of 8 threads were effectively utilized.

A total of *7,070,000,000* mating events could occur for every model. These values are excessive by orders of magnitude, however any marginal improvement may result in a better description of the underlying system; predictive models are *not* the goal, high quality descriptive models are.

Given the stochastic nature of the search and the varying amounts of data in each task, each execution of the search took between *24* and *124* hours (in the most extreme cases) when running with 8 cores on an *IBM System x iDataPlex dx360 M3* node with 2 quad-core *Intel Nehalem (Xeon 5540)* processors running at 2.53GHz.

4.3 Experimental Methods

Fourty subjects with data from all seven tasks were studied (280 datasets total). For symbolic regression, to improve the significance and quality of results, *100* models were generated. For linear regression, six different ways of generating models were investigated: fitting all ROI, performing FDR and thresholding then fitting, performing BC and thresholding then fitting, fitting all ROI with LASSO regression, performing FDR and thresholding then fitting with LASSO regression, performing BC and thresholding then fitting with LASSO regression. *Least absolute shrinkage and selection operator* (LASSO) regression is already used in some of the neuroscience literature, and it typically generates smaller models compared to typical linear regression. Previous work found that symbolic regression selected very few ROI compared to linear regression [13], so it is of interest to compare symbolic regression to a linear method with similarly succinct models.

For both linear and symbolic regression, an ROI known to be involved with the task was chosen to be the dependent variable (left hand side of the equation (y)) and all other ROIs are used as the regressors (X). For example, ROI 21 was selected as the dependent variable for the motor task as it is the ROI containing the primary motor cortex. The left hand sides of the equations for the emotion, gambling, language, motor, relational, social, and working memory tasks were ROIs 7, 2, 12, 21, 28, 3, and 21 respectively.

5 RESULTS AND DISCUSSION

5.1 Model Effectiveness

Table 3 contains summary statistics for the top models for each subject on all tasks. Although 100 nonlinear models were generated for each subject and task, only the top performing model was analyzed here. Additionally, a Mann-Whitney U test's p-value obtained by comparing the mean absolute errors of the nonlinear and the respective linear model is also included. This table is similar to previous work's [13], however the presented values were calculated on LASSO models and over more subjects. The results in this table reinforces previous observations that nonlinear models are comparable to linear when applied to the data they were fit to. The model type obtaining the best results were those fit to all ROI with regular linear regression, however they were never significantly better than nonlinear models. Additionally, these models likely overfit the data and would not provide any neuroscientific insight since they used all features.

Figure 2 presents the *p-value transition* plot. This plot was generated by comparing the top nonlinear models' errors from all subject to the errors from a linear model fit with increasingly more ROIs. The p-value is represented as color and each column corresponds to different tasks. The first row compares the nonlinear models to linear models fit with the top 1 linearly correlated ROI (to the ROI on the left hand side of the equation). More ROIs are added in the order of absolute correlation score until all ROI are included (the last row). The average number of ROI (over all subjects) in a linear model with BC and FDR is written on the plot along with the average number of ROI in all (100) nonlinear models generated for each subject (NL-A-) and the average number of ROI in the top nonlinear model for each subject (NL-T-).

This plot shows that the average number of ROI in the nonlinear model is much lesser than those generated with linear regression. It also shows that nonlinear models are significantly better than linear models fit with few, top correlated ROI, but as the number increases, the difference disappears. The last row corresponds to the last column in Table 3 where we see that the best linear models are not significantly better than nonlinear. This plot does not include the LASSO models as the ROI in those models are not determined based on correlation scores. The number of ROI in the linear models generated with LASSO was typically between 7 – 11 and is much more comparable to the number of ROI in the nonlinear models, however, as seen in Table 3, the LASSO models typically have worse mean absolute errors than the nonlinear models.

It is not the specific operators found in the model, but the presence of the ROI, and the fact that they are related in some nonlinear way, that is of interest. Figure 3 shows how often each ROI appeared in the 40 subjects's models for each task. The first matrix shows the results for the top nonlinear models and bottom three shows the results from three linear models. Although all these four matrices are similar, the one for the nonlinear models is the most different. Not only are the nonlinear models effective, but they appear to present somewhat different ROIs. Similar observations were noted in [13]. The other linear regression models' ROIs were not shown as they typically had nearly all, or all ROI present.

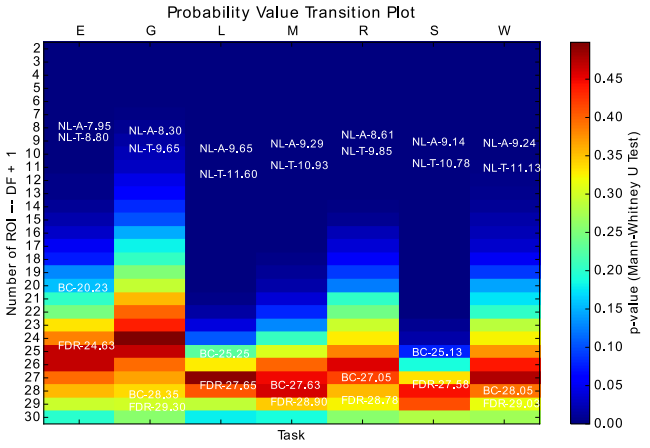


Figure 2: p-value transition plot comparing linear and nonlinear models' mean absolute errors (averaged over all subject) as the number of ROIs used to create the linear model increases. ROIs were added to the linear models in the order of their absolute correlation score. The number of ROIs in the nonlinear models was fixed.

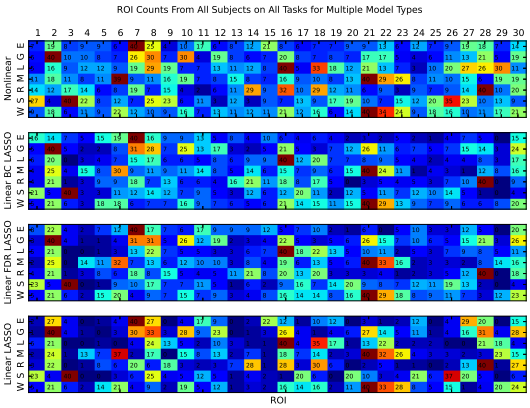


Figure 3: Number of subjects for each ROI (column) that appeared in the top model for each task (row). Counts for the nonlinear and LASSO generated linear models are presented. The other liner models were excluded as they typically contained nearly all (or all) ROIs. 40 is maximum. Note that the ROI corresponding to the left hand side of the equation was in all models.

5.2 Intersubject Generalizability

Figure 4 contains matrices showing how well models generalize to unseen data from different subjects. The matrices were generated by applying models from all subjects and tasks to every other subject and task's data. The mean absolute errors were then averaged over all subjects performing the same task. The matrices for the other linear models were not included as they did not generalize to other subjects as well as the LASSO models. The diagonals are of particular interest as they show how well, on average, models for a

Table 3: Summary statistics (median and in interquartile range (IQR)) for all generated models along with probability values obtained with a Mann-Whitney U test when comparing the mean absolute errors of the nonlinear models to the respective linear model.

	Nonlinear		BC LASSO			FDR LASSO			BC			FDR			ALL LASSO			ALL		
	Median	IQR	Mdn	IQR	p-Val	Mdn	IQR	p-Val	Mdn	IQR	p-Val	Mdn	IQR	p-Val	Mdn	IQR	p-Val	Mdn	IQR	p-Val
EMOTION	0.39	±0.06	0.49	±0.08	1.08e-04	0.47	±0.07	2.81e-04	0.41	±0.09	1.29e-01	0.39	±0.08	4.11e-01	0.47	±0.07	5.44e-04	0.37	±0.06	2.00e-01
GAMBLING	0.32	±0.06	0.37	±0.07	1.58e-02	0.36	±0.07	1.65e-02	0.31	±0.06	3.17e-01	0.3	±0.06	2.77e-01	0.36	±0.07	1.65e-02	0.3	±0.06	2.58e-01
LANGUAGE	0.28	±0.03	0.39	±0.04	4.02e-10	0.38	±0.04	4.28e-10	0.28	±0.04	2.19e-01	0.27	±0.03	4.87e-01	0.38	±0.04	6.92e-10	0.26	±0.03	1.79e-01
MOTOR	0.23	±0.04	0.32	±0.05	8.94e-08	0.32	±0.05	9.41e-08	0.23	±0.05	4.90e-01	0.23	±0.05	3.52e-01	0.32	±0.05	1.16e-07	0.23	±0.04	1.89e-01
RELATIONAL	0.23	±0.05	0.31	±0.05	2.50e-05	0.31	±0.05	2.71e-05	0.22	±0.06	4.41e-01	0.22	±0.05	3.20e-01	0.31	±0.05	2.82e-05	0.22	±0.05	2.58e-01
SOCIAL	0.3	±0.04	0.44	±0.07	5.12e-10	0.42	±0.06	6.52e-10	0.33	±0.06	5.56e-02	0.31	±0.05	3.38e-01	0.42	±0.06	9.89e-10	0.29	±0.04	2.80e-01
WM	0.26	±0.05	0.31	±0.06	3.35e-04	0.31	±0.06	3.72e-04	0.25	±0.05	4.18e-01	0.25	±0.05	3.59e-01	0.31	±0.06	3.72e-04	0.25	±0.05	2.71e-01

specific task can fit data from other subjects performing the same task. The three linear models generalize to other subjects similarly well, and significantly better than the nonlinear models. However, it should noted that the LASSO models also fit all other task's data well. Perhaps these LASSO models are not as capable at describing task specific nuances. This problem can be seen in the non-LASSO linear models (not shown), but to a far lesser extent.

5.3 Intrasubject Generalizability

If we take the top models of each type (1 nonlinear and 6 linear) and apply them to unseen *test* data from the same subject and task, we can compare the resulting mean absolute error values and use the difference as a way to understand overfitting. Although the task was the same in the unseen data, the order in which the subtasks were done (ex: moving hand, foot, tongue) were different. Fortunately, this should not matter since the models are temporally independent. Figure 5 plots the training and testing errors against each other. Unsurprisingly, nearly all the points are above the $y = x$ line, indicating that the models obtain better errors on the data they were fit to. The difference between the training and testing errors averaged over all subjects in tasks for each model are: NL – 0.20, BC LASSO – 0.10, BC – 0.10, FDR LASSO – 0.18, FDR – 0.19, All LASSO – 0.11, and ALL – 0.21.

A total of 100 nonlinear models were generated for each subject and task combination for statistical power and, because of the stochastic nature of the search, to increase our chance of obtaining high quality models. We apply these 100 models, which should all be reasonably effective, to the unseen testing data from the same subject performing the same task. Figure 6 shows the distribution of mean absolute errors from the 100 models along with vertical lines indicating the mean absolute errors from the 6 linear models fit to the same data as the nonlinear and applied to the same unseen data. From this example we can see that some number of the 100 nonlinear models performed better than the best.

Distributions like Figure 6 can be generated for each of the 280 subject and task combination. Figure 7 was generated by plotting the best nonlinear model's error (left most error from the respective distribution) against the best linear model's. Each point on these plots corresponds to a different subject. Points above the $y = x$ line indicate that a nonlinear model was best at generalizing to unseen data from the same subject. The overwhelming majority of these points are above this line, suggesting that, in general, a nonlinear model can generalize to unseen data from the same subject better than the linear. Table 4 shows the average difference between the models when the respective column's model type was best. Not only were more nonlinear models better, but when they were better, they were better by more than when linear models were better.

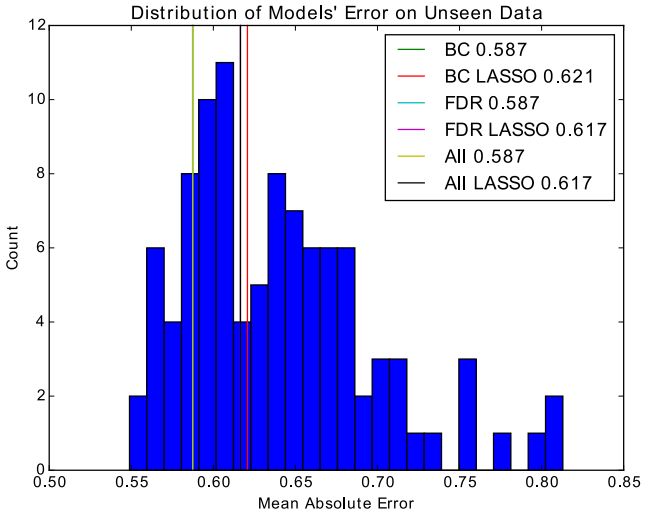


Figure 6: Distribution of mean absolute error values when applying all 100 nonlinear models to unseen data from the same subject performing the same task. Vertical lines correspond to the mean absolute errors obtained by linear models.

The authors want to make very clear that they acknowledge the bias being introduced in this section; we have 100 nonlinear models to choose from and only 6 linear to choose from. The only way to confirm the generalizability of any of these models is to apply the selected models to new unseen data. Unfortunately, a third set of data for each subject and task is not available and this confirmation is not currently possible. This limitation is important to keep in mind when interpreting these results.

Figure 7 only compares the best nonlinear model found when applied to unseen data. However, for each subject, it is likely that more than just one of the 100 nonlinear model performed better than the best linear model. Figure 8 shows a distribution of how many nonlinear models were better than the best linear model for all subjects (if such models exist). For many subjects, numerous nonlinear models generalized to unseen data better than the best linear model, suggesting that these nonlinear models are meaningful and, while still acknowledging the bias, perhaps more capable of generalizing to unseen data from the same subject better than linear models.

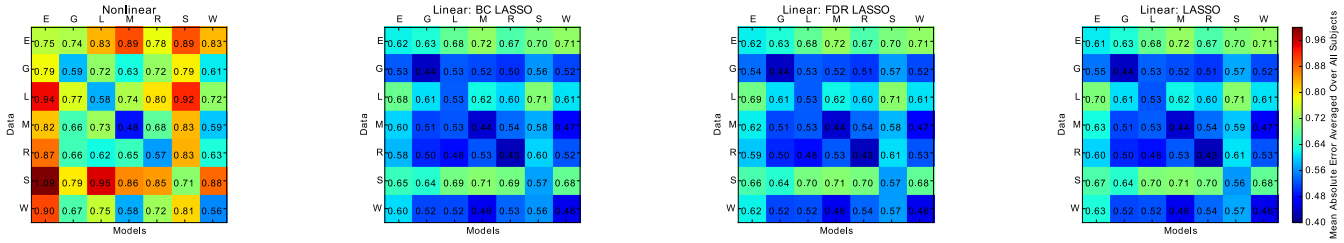


Figure 4: Matrices showing the mean absolute error values obtained by applying every task/subject combination's models to all other datasets and averaged over all subjects performing the same task. The diagonal provides a means of quantifying intersubject generalization; if all subject's models on the same task can fit all other subject's data from that task similarly well, then the models are capable of generalizing between subjects.

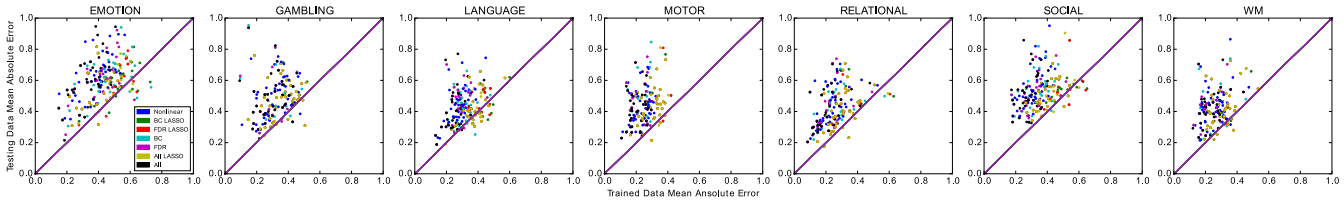


Figure 5: Scatterplot comparing the training and testing mean absolute errors for all models. For the nonlinear model, the top model on the training data was compared to it's error when applied to unseen data.

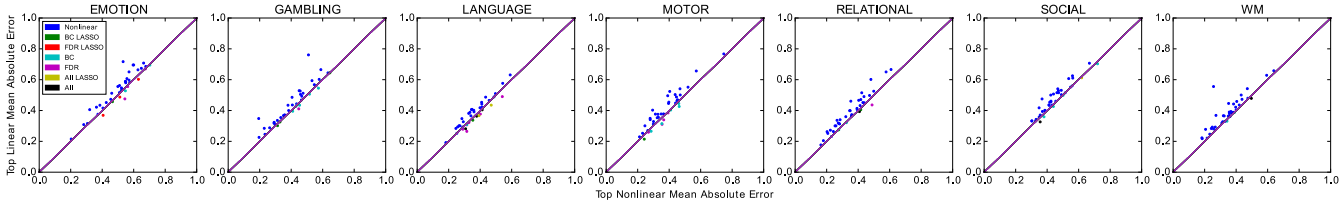


Figure 7: Scatterplot comparing the smallest mean absolute error from the 100 nonlinear models when applied to unseen data versus the best of the 6 linear models. Points above the $y = x$ line indicate that the nonlinear model was best. Points below indicate that a linear model was best. Color indicates method for model generation.

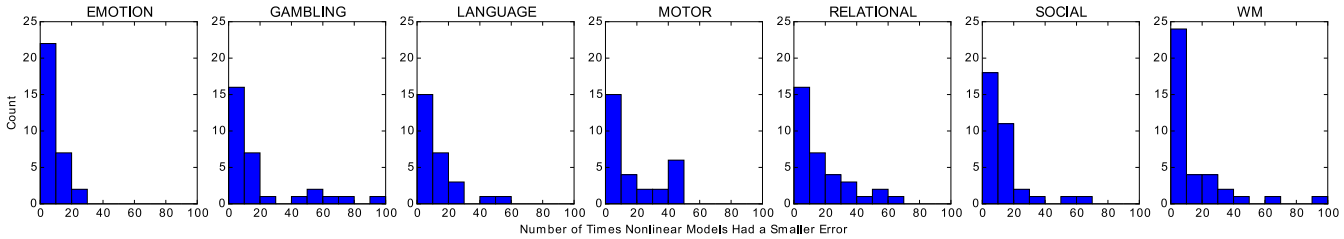


Figure 8: For each subject, the number of the 100 nonlinear models generated that were better than the best linear model when applied to unseen data was calculated and the distributions were plotted. Bins (x-axis) represent the number of nonlinear models better than the best linear. The bin height (y-axis) corresponds to the number of subjects.

Although we unfortunately do not have a third set of data for each subject, we can use the other subjects' data from the same task as a *pseudo* third dataset. Understanding that the data is not obtained from the same random variable, we can apply the top model on unseen data from the same subject to this pseudo third dataset. Similar to figure 4, Figure 9 shows how well the best same subject generalizing models fit all other subject's data. When comparing the matrices for the nonlinear models, with the exception of the emotion and motor task, we observe a significant improvement in between subject generalization. However, the better generalizing nonlinear models were still significantly worse than the best linear models at generalizing to other subjects.

Table 4: Average difference between the best nonlinear and linear models' mean absolute errors when the respective column's model was best. The values are averaged over all subjects performing the same task. Ex: for the emotion task, when nonlinear models were better than linear, they were on average better by 0.041.

Task #	Nonlinear Better	Linear Better
Emotion	0.041	0.023
Gambling	0.044	0.013
Language	0.031	0.022
Motor	0.045	0.022
Relational	0.043	0.014
Social	0.034	0.015
W. Memory	0.039	0.010

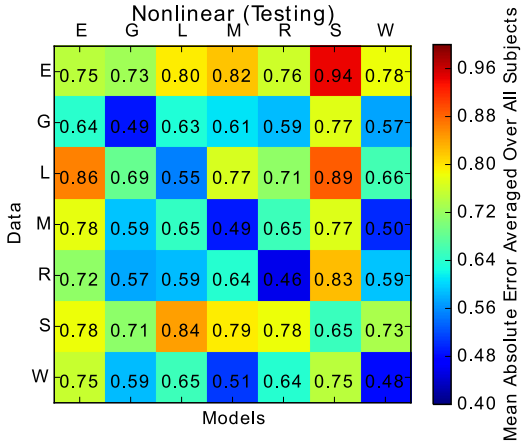


Figure 9: Similar to Figure 4, this matrix shows the mean absolute error values obtained by applying *the best model on the unseen data* from every subject/task to all other datasets and averaged over all subjects performing the same task.

5.4 Model Selection Problem

The purpose of generating these models is to find a descriptive model that can provide insight into the underlying system — the brain. Since we have no actual target, we use the error values to indicate model quality. Here in lies a significant problem. We have a collection of high quality models, both linear and nonlinear. Although some have smaller errors than other, and since the error can only be used as a proxy for model *correctness*, any small differences in error should not be taken as meaningful. How can one select a model, or decipher meaning from the collection of models?

Perhaps if the collection of models generated provide some consensus on which ROI were meaningful, then we could use that information to develop our functional connectivity network. Figure 10 shows how often each ROI (column) appeared in the 100 models generated for each subject (row) on each task. Although these matrices are similar to those found in Figure 3, the ROI counts corresponds to how often they appeared in all 100 models, not how often they appeared in the top models for each subject. There are two main observations to be made from Figure 10.

First, there is no overwhelming consistency of ROIs between the subjects. There are some ROIs that appear to be more prevalent in all subjects' models than others, but it would be difficult to draw strong conclusions from this. This inconsistency could explain why the nonlinear models do not generalize between subjects as well as the linear models. This is also interesting since, given the resolution of the brain being studied (30 ROI), one would expect some level of consistency. It is difficult to conclude why this inconsistency would happen. It could be the result of low quality models, noisy data, or that there really is this much of a difference in the functional connectivity networks between these subjects.

The second observation is that when focusing on specific subjects (rows), there is again, in general, no overwhelming consistency in which ROIs are prevalent in all models generated for each subject and task. These differences are also difficult to account for. It could be the result of low quality models or that more than one ROI can explain the same phenomenon. One could try to develop subject specific functional connectivity networks from this information, but this would likely require arbitrary thresholding.

We have generated seemingly high quality models based on error values, but how can one select a single model from the collection and expect it to be representative of the underlying system? A more concerning question is: how can one generate a single linear model and expect it to be representative of the underlying system?

6 CONCLUSIONS AND FUTURE WORK

Nonlinear models of functional connectivity of human brains were generated with symbolic regression. These models were built from real fMRI data obtained from the Human Connectome Project. The nonlinear models were found to be more succinct than many linear models. The nonlinear models had different ROI than the linear and contained nonlinear relationships — something not possible with traditional linear tools.

Nonlinear models obtained low error values, were better than certain linear models, and were never significantly worse than the best linear models when applied to data they were fit.

The nonlinear models were unable to fit unseen data from other subjects as well as the traditional linear models. However, the linear models were capable of fitting other subject's data from unrelated tasks well.

Many nonlinear models fit unseen data from the same subject better than linear models, however the analysis did introduce bias and would require additional data for confirmation. Unfortunately additional data is not available and is a common limitation in neuroscience. In an attempt to simulate additional unseen data, the more general nonlinear models were applied to data from different subjects. These nonlinear models significantly improved the between subject generalizability of the nonlinear models, but the linear models still generalized between subjects better.

This work presents the problem of model selection. In the end, a large collection of seemingly high quality nonlinear and linear models (based on acceptable error metrics) was obtained. These models had similarities, but had no strong consensus on ROI and relationship type. Since the goal is to discover the underlying functional connectivities and not to find the model with the lowest error, it is difficult to intelligently select any model, whether linear

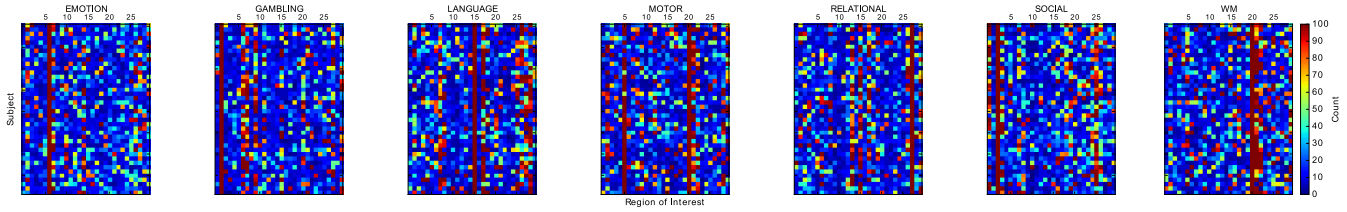


Figure 10: Matrices showing the number of times (color) each ROI (column) appeared in the 100 nonlinear models generated for each subject (row) on each task. Note that the ROI corresponding to the left hand side of the equation was in all models.

or nonlinear. At the very least however, it would seem better to have a collection of models rather than a linear single model — something GP and symbolic regression delivers.

To enable a better investigation into same subject generalizability, it is necessary to obtain additional, different data with multiple sets of data from each subject performing each task. This would allow for a training, validation and testing analysis to eliminate bias. A deeper investigation into model consensus (Figure 10) could yield stronger evidence of functional connectivities. This could be achieved with some methods of thresholding, filtering, and data smoothing.

This work regressed models to ROIs within the system. It is possible to generate models fit to a design matrix, where the left hand side of the equation is some prediction of functional activity. This approach would work towards the same conclusions, but from another direction. Further, this strategy could be used to fit *resting state fMRI* data (fMRI data gathered while the subject was performing no task) to an unrelated design matrix. Testing GP's ability to fit a model to a system that has no relationship could provide insight into symbolic regression's propensity to overfit this type of data. This work has been started by in [14].

7 ACKNOWLEDGMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

This research was enabled in part by support provided by Compute Ontario (computeontario.ca), Calcul Québec (calculquebec.ca), Westgrid (westgrid.ca), and Compute Canada (computecanada.ca).

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

REFERENCES

[1] Nicholas Allgaier. 2015. Reverse Engineering the Human Brain: An Evolutionary Computation Approach to the Analysis of fMRI. (2015).

[2] Nicholas Allgaier, Tobias Banaschewski, Gareth Barker, Arun LW Bokde, Josh C Bongard, Uli Bromberg, Christian Büchel, Anna Cattrell, Patricia J Conrod, Christopher M Danforth, and others. 2015. Nonlinear functional mapping of the human brain. *arXiv preprint arXiv:1510.03765* (2015).

[3] Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. *The journal of neuroscience* 16, 13 (1996), 4207–4221.

[4] R.L. Buckner and T.S. Braver. Event-Related Functional MRI. In *Functional MRI*, P. Bandettini and C. Moonen (Eds.). Springer-Verlag, Chapter 36, 441–452.

[5] R Cameron Craddock, G Andrew James, Paul E Holtzheimer, Xiaoping P Hu, and Helen S Mayberg. 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human brain mapping* 33, 8 (2012), 1914–1928.

[6] Mark Daley. 2014. An Invitation to the Study of Brain Networks, with Some Statistical Analysis of Thresholding Techniques. In *Discrete and Topological Models in Molecular Biology*. Springer, 85–107.

[7] Karl J Friston, Lee Harrison, and Will Penny. 2003. Dynamic causal modelling. *Neuroimage* 19, 4 (2003), 1273–1302.

[8] Karl J Friston, Oliver Josephs, Geraint Rees, and Robert Turner. 1998. Nonlinear event-related responses in fMRI. *Magnetic resonance in medicine* 39, 1 (1998), 41–52.

[9] Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price. 2000. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *Neuroimage* 12, 4 (2000), 466–477.

[10] Karl J Friston, CJ Price, Paul Fletcher, C Moore, RSJ Frackowiak, and RJ Dolan. 1996. The trouble with cognitive subtraction. *Neuroimage* 4, 2 (1996), 97–104.

[11] Scott A Huettel, Allen W Song, and Gregory McCarthy. 2009. *Functional magnetic resonance imaging* (second ed.). Vol. 1. Sinauer Associates Sunderland, MA.

[12] James Alexander Hughes and Mark Daley. 2016. Finding Nonlinear Relationships in fMRI Time Series with Symbolic Regression. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*. ACM, 101–102.

[13] James Alexander Hughes and Mark Daley. 2017. Searching for nonlinear relationships in fMRI data with symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 1129–1136.

[14] Ethan Jackson, James Alexander Hughes, and Mark Daley. 2018. On the Generalizability of Linear and Non-Linear Region of Interest-Based Multivariate Regression Models for fMRI Data. *arXiv preprint arXiv:1802.02423* (2018).

[15] Yaochu Jin. 2005. A comprehensive survey of fitness approximation in evolutionary computation. *Soft computing* 9, 1 (2005), 3–12.

[16] John R Koza. 1992. *Genetic programming: on the programming of computers by means of natural selection*. Vol. 1. MIT press.

[17] F Kruggel, Stefan Zysset, and D Yves von Cramon. 2000. Nonlinear regression of functional MRI data: an item recognition task study. *Neuroimage* 12, 2 (2000), 173–183.

[18] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 6843 (2001), 150–157.

[19] Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. 1992. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences* 89, 13 (1992), 5951–5955.

[20] Michael Schmidt and Hod Lipson. 2007. Comparison of tree and graph encodings as function of problem complexity. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*. ACM, 1674–1679.

[21] Michael D Schmidt and Hod Lipson. 2007. Coevolving fitness models for accelerating evolution and reducing evaluations. In *Genetic Programming Theory and Practice IV*. Springer, 113–130.

[22] Michael D Schmidt and Hod Lipson. 2008. Coevolution of fitness predictors. *Evolutionary Computation, IEEE Transactions on* 12, 6 (2008), 736–749.

[23] Michael D Schmidt, Ravishankar R Vallabhajosyula, Jerry W Jenkins, Jonathan E Hood, Abhishek S Soni, John P Wikswo, and Hod Lipson. 2011. Automated refinement and inference of analytical models for metabolic networks. *Physical biology* 8, 5 (2011), 055011.

[24] Alberto L Vazquez and Douglas C Noll. 1998. Nonlinear aspects of the BOLD response in functional MRI. *Neuroimage* 7, 2 (1998), 108–118.

[25] Tingting Zhang, Fan Li, Marlen Z Gonzalez, Erin L Maresh, and James A Coan. 2014. A semi-parametric nonlinear model for event-related fMRI. *Neuroimage* 97 (2014), 178–187.

Chapter 7

Conclusions and Future Directions

7.1 Genetic Programming System

Although the majority of this thesis concerns itself with the application of GP to fMRI data for the purpose of finding nonlinear functional connectivities, the development of the GP system was a significant part of the contribution. The system was based on Schmidt et al.'s work [103, 106] and it incorporated a number of enhancements, some of which are fairly standard within the field of evolutionary computation (elitism, distributed populations/island model), and others were more application specific (representation, fitness predictors). These enhancements enabled the success of the project as they improved results and runtimes when compared to a more basic GP implementation.

The development of version 1 took months, and multiple subsequent versions were developed throughout the course of the project as new results would guide the development of the system. The system is currently in version 9, and will continue to be developed. A GitHub repository of the GP system can be found at `https://github.com/jameshughes89/jGP` [52].

The GP system proved to be highly effective in both quality of results and runtimes. The success of the GP system led to the application of it to additional side projects: modelling human walking data [53, 54], and developing a predictive model of intracranial pressure [59, 60]. The GP system has also been used by Jackson et al. for a project related to the contents of this thesis — analysing nonlinear models of fMRI data [62].

Despite being a form of computational intelligence, which is known to be highly susceptible to overfitting, it was demonstrated that the specific implementation was capable of finding generalizable results, as demonstrated in [53, 59, 54, 62].

7.2 Application: Nonlinear Models of fMRI Data

The GP system specifically designed to perform symbolic regression on fMRI timeseries data was successful in finding nonlinear relationships within the data recorded from the dynamic nonlinear system: the human brain. All fMRI data used was obtained from the Human Connectome Project's database, a popular publicly available collection of neuroimaging data of a large number of subjects performing a variety of tasks. Originally, data from a single task

(motor) from all subjects that were available (507) were modelled; however, the most recent developments in the project studied all seven tasks available from only forty subjects.

Symbolic regression, a data driven form of regression analysis, found nonlinear relationships within the data. The models of data were *highly* nonlinear; models contained nonlinear operators and variables were combined in ways that did not follow the law of superposition (see Figure 4 in Chapter 4 (article 1) for an example). These nonlinear models could not have been found with the typical linear tools used within the field. Although it is possible to derive a linear combination of nonlinear basis functions with linear regression, the nonlinear basis functions would have to be selected beforehand. Unlike linear regression, one of the major benefits of using symbolic regression is that it allows the user to reduce the number of prior assumptions about the space; model structure is completely derived by the search.

A major goal of this work was to produce *descriptive* models of the system. For some instances, particularly in the application of artificial intelligence and machine learning within computer science, models are developed for their *predictive* capabilities. For example, developing an image classifier where the model outputs a classification. In this long term project, although model output (error) is used to indicate effectiveness, it is the models themselves that are of interest, and not the model output.

Unlike many other forms of computational intelligence and machine learning, symbolic regression produces symbolic models that can be relatively easily interpreted. For example, artificial neural networks might model the system well, but it would be difficult to derive meaning from them.

The author wants to stress that they are not suggesting to interpret every relationship and operator within the symbolic models literally, but to interpret the models at the level of ROIs/features and if their relationships are linear or nonlinear (see Figure 4 in Chapter 4 (article 1)). For example, if a model was $ROI_{21} = \sin(ROI_5)$, one should take away that ROI_{21} and ROI_5 are related in some nonlinear way.

After interpreting the models, it was observed that although the nonlinear models and linear models found many of the same relationships, the nonlinear models also found novel relationships not contained within the linear models. It was also noted that the similarity/differences between the linear and nonlinear models was task dependent. Given that a collection of nonlinear models was created for each subject/task combination, it's possible to look at the models as a whole group. When doing so, there is some consistency with respect to ROI relationships, but there is no consensus among the nonlinear models.

Not only did the nonlinear models contain nonlinear relationships, but they contained many fewer relationships. Symbolic regression is capable of performing feature selection and also searches for a model structure. The nonlinear models of functional connectivity were consistently more succinct (fewer ROIs) when compared to linear models with thresholding. Although it depends on the task, nonlinear models would typically contain fewer than 10 ROIs while linear regression with thresholding would have well over 20 ROIs (sometimes all). Although it is possible that the entire brain is meaningfully functionally connected during a task, perhaps these more succinct models are more representative of relationships of interest (the project in its current state makes no rigorous attempt to justify this hypothesis). When LASSO regularization was used, the number of features included in the linear and nonlinear models were comparable.

Although the linear and nonlinear models being similar provides face validity that the non-

linear models are actually capable of describing functional connectivities found with already accepted methods, it is their differences that are of neuroscientific interest. However, this project in its current state makes no attempt to justify the meaning behind these differences.

7.2.1 Error Values

In many cases the nonlinear models were better at fitting to the data they were fit to (had smaller mean absolute errors) when compared to linear models, and they were never significantly worse than the best linear models.

When compared to linear models, the nonlinear models were not capable of generalizing to unseen data from different subjects as well (intersubject generalization). However, it was observed that the linear models generated with LASSO regularization were capable of generalizing to unseen data from other subjects on *unrelated tasks*. It is hypothesized that perhaps the linear models were not capable of extracting task specific nuances within the data.

All 100 nonlinear models generated for each subject and task were applied to unseen data from the same subject (intrasubject generalization). In the majority of instances, many of the 100 nonlinear models generalized to the unseen data better than the best of 6 linear models. Having 100 nonlinear models to choose from introduces bias into the analysis, and it would be required a third unseen testing set of data. Unfortunately, no such data exists from the Human Connectome Project as the database only contains two independent scans for each subject performing each task.

As a means of having a *pseudo* testing set of unseen data, the top model of the 100 being applied to unseen data from the same subject was applied to unseen data from different subjects. Effectively, another analysis of intersubject generalization was done, but the best of the 100 models at intrasubject generalization was used as opposed to the best of the 100 models on the data the models were fit to. If the best intrasubject generalizing model can generalize to unseen data from other subjects better, then we can show how general the nonlinear models are. In most cases the top nonlinear model at intrasubject generalization was capable of generalizing to unseen data from other subjects significantly better than the top nonlinear models found when applied to the data they were fit to; however, they were still significantly worse at intersubject generalization when compared to linear models.

7.2.2 Model Selection Problem

Ultimately, in its current state, this project presents a *model selection problem*. Many high quality linear and nonlinear models are produced for noisy timeseries data recorded from a high-dimensional nonlinear chaotic dynamic system. Many models had relatively low errors, whether it was applied to data the models were fit to, unseen data from the same subject, or unseen data from other subjects. However, the goal is to develop a meaningful model of the underlying system and to learn about the functional connectivities within the brain.

The author feels that this is one of the major conclusions of the work in its current state. When presented with a large number of similarly high quality models, how does one intelligently select the appropriate descriptive model? Perhaps the best method is to study the models as a group in an attempt to discover some agreement between ROIs and relationship types, but

currently there is no consensus in the collection of models. Fortunately, symbolic regression provides a means of developing a collection of models that can be studied in this way.

7.3 Future Work

The GP system will continue to be developed. Small incremental changes are inevitable as continued work always presents results that may suggest improvements. Larger changes will also be included as the field of evolutionary computation is always growing and many improvements tend to be modular in nature and are easily incorporated into existing systems. An example of this could be *novelty search*, a strategy which encourages novelty in the population [81]. Another major change that is of particular interest is to enable the system to develop temporally dependent models since the system itself if temporally dependent; the current state of the system depends on previous states. Additionally, as discussed in Section 3.2, much of the current work studying nonlinearities within the data are concerned with temporal nonlinearities.

Gathering our own data, or having access to similarly controlled data with 3 independent recordings for each subject would allow for a much better analysis of intrasubject generalization as it will provide a training, testing, and validation dataset.

The methods are not restricted to fMRI data, and it would be interesting to model other neuroimaging modalities where nonlinearities are commonly observed, such as electroencephalogram or MEG.

Currently the data was studied at a resolution of 30 ROIs. Although there was original motivation in selecting only 30 ROIs, Jackson et al. used the same GP system to develop models of fMRI data at a resolution of 50 ROIs [62]. By using only 30 ROIs, important features in the data may have been lost and *smoothed* out. Future work on this project should increase the number of ROIs from 30 in hopes of creating higher quality models.

The models developed for this project were regressed to ROIs within the system. This is not a requirement and it is possible to regress the data to an expected HDR function instead (although this may remove some nonlinearlities as doing so is effectively a low-pass filter). This approach would still work towards developing nonlinear models of the system, but from a different direction. This could also allow more specific questions to be asked, such as which ROI is functionally connected and related to specific stimuli. This work has already been conducted by Jackson et al. [62].

As a means of studying if the nonlinear models are overfitting the data, symbolic regression can be used in an attempt to fit resting-state fMRI data to unrelated expected hemodynamic response functions. If symbolic regression is able to develop models of unrelated data, then it would indicate that the models are not effective and are finding non-existent relationships. This investigation has already been started by Jackson et al. and currently symbolic regression was incapable of fitting resting-state data to an unrelated expected hemodynamic response function. This does eliminate some concerns of overfitting [62].

One of the major strengths of symbolic regression is that it performs feature selection, and perhaps this conclusion is one of the most significant contributions of this thesis. With Compute Canada resources, it is currently computationally feasible to develop linear models

of all permutations of a reasonably small number of ROIs[1]. The resulting models could be studied in a similar way to the nonlinear models if provided with enough training, validation, and testing data. If the nonlinear models are still capable of describing the system better than the large number of linear models with varying features, then it could indicate that symbolic regression's benefit is more than just feature selection[2].

Symbolic regression was used to develop nonlinear models of fMRI data, and the models were analysed and compared to linear models; however, there is currently no attempt to justify the nonlinear models from a neuroscientific perspective. For example, why do certain ROIs appear in the nonlinear models, but not in the linear models? Is there any reasonable neuroscientific justification for this? Although there is still much work to be done in the analysis of the methodology, working with neuroscientists to develop a meaningful understanding of the resulting models is a priority.

[1] At the very least, it is possible to do a Monte Carlo method.

[2] However, the author predicts that the nonlinear models will not outperform optimal models of all permutations of ROIs.

Bibliography

[1] GK Aguirre, JA Detre, E Zarahn, and DC Alsop. Experimental design and the relative sensitivity of bold and perfusion fmri. *Neuroimage*, 15(3):488–500, 2002.

[2] Nicholas Allgaier. Reverse engineering the human brain: An evolutionary computation approach to the analysis of fmri. 2015.

[3] David Andre and Astro Teller. Evolving team darwin united. In *RoboCup-98: Robot soccer world cup II*, pages 346–351. Springer, 1999.

[4] Jessica R Andrews-Hanna. The brain's default network and its adaptive role in internal mentation. *The Neuroscientist*, 18(3):251–270, 2012.

[5] Owen J Arthurs and Simon Boniface. How well do we understand the neural origins of the fmri bold signal? *TRENDS in Neurosciences*, 25(1):27–31, 2002.

[6] Thomas Bäck. Selective pressure in evolutionary algorithms: A characterization of selection mechanisms. In *Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on*, pages 57–62. IEEE, 1994.

[7] James Edward Baker. Adaptive selection methods for genetic algorithms. In *Proceedings of an International Conference on Genetic Algorithms and their applications*, pages 101–111. Hillsdale, New Jersey, 1985.

[8] Shumeet Baluja and Rich Caruana. Removing the genetics from the standard genetic algorithm. In *Machine Learning: Proceedings of the Twelfth International Conference*, pages 38–46, 1995.

[9] Howard Barnum, Herbert J Bernstein, and Lee Spector. Quantum circuits for or and and of ors. *Journal of Physics A: Mathematical and General*, 33(45):8047, 2000.

[10] Nils Aall Barricelli. Symbiogenetic evolution processes realized by artificial methods. *Methodos*, 9(35-36):143–182, 1957.

[11] Danielle Smith Bassett and ED Bullmore. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006.

[12] Bharat B Biswal. Resting state fmri: a personal history. *Neuroimage*, 62(2):938–944, 2012.

[13] Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007.

[14] Lashon Booker. Intelligent behavior as a adaptation to the task environment. *Doctoral Dissertation, Department of Computer and Communication Sciences, University of Michigan, Ann Arbor*, 1982.

[15] Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *The journal of neuroscience*, 16(13):4207–4221, 1996.

[16] Robert W Brown, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[17] Randy L Buckner, Jorge Sepulcre, Tanveer Talukdar, Fenna M Krienen, Hesheng Liu, Trey Hedden, Jessica R Andrews-Hanna, Reisa A Sperling, and Keith A Johnson. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to alzheimer's disease. *The Journal of Neuroscience*, 29(6):1860–1873, 2009.

[18] R.L. Buckner and T.S. Braver. Event-related functional mri. In P. Bandettini and C. Moonen, editors, *Functional MRI*, chapter 36, pages 441–452. Springer-Verlag.

[19] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[20] Richard B Buxton and Lawrence R Frank. A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *Journal of Cerebral Blood Flow & Metabolism*, 17(1):64–72, 1997.

[21] Richard B Buxton, Kâmil Uludağ, David J Dubowitz, and Thomas T Liu. Modeling the hemodynamic response to brain activation. *Neuroimage*, 23:S220–S233, 2004.

[22] Richard B Buxton, Eric C Wong, and Lawrence R Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic resonance in medicine*, 39(6):855–864, 1998.

[23] James P Cohoon, Shailesh U Hegde, Worthy N Martin, and D Richards. Punctuated equilibria: a parallel genetic algorithm. In *Genetic algorithms and their applications: proceedings of the second International Conference on Genetic Algorithms: July 28-31, 1987 at the Massachusetts Institute of Technology, Cambridge, MA*. Hillsdale, NJ: L. Erlhaum Associates, 1987., 1987.

[24] R Cameron Craddock, G Andrew James, Paul E Holtzheimer, Xiaoping P Hu, and Helen S Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

[25] Nichael Lynn Cramer. A representation for the adaptive generation of simple sequential programs. In *Proceedings of the First International Conference on Genetic Algorithms*, pages 183–187, 1985.

[26] Mark Daley. An invitation to the study of brain networks, with some statistical analysis of thresholding techniques. In *Discrete and Topological Models in Molecular Biology*, pages 85–107. Springer, 2014.

[27] Kenneth De Jong. Learning with genetic algorithms: An overview. *Machine learning*, 3(2-3):121–138, 1988.

[28] Gustavo Deco, Viktor K Jirsa, and Anthony R McIntosh. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience*, 12(1):43–56, 2011.

[29] Thomas Deneux and Olivier Faugeras. Using nonlinear models in fmri data analysis: model selection and activation detection. *NeuroImage*, 32(4):1669–1689, 2006.

[30] David I Donaldson and Randy L Buckner. Effective paradigm design. In *IN P. JEZZARD (ED.), FUNCTIONAL MRI*. Citeseer, Functional magnetic resonance imaging of the brain: Methods for neuroscience. Oxford: Oxford University Press, 2001.

[31] Nico UF Dosenbach, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361, 2010.

[32] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[33] Michael D Fox and Marcus E Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9):700–711, 2007.

[34] Alex Fraser, Donald Burnell, et al. Computer models in genetics. *Computer models in genetics.*, 1970.

[35] Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[36] Karl J Friston, Oliver Josephs, Geraint Rees, and Robert Turner. Nonlinear event-related responses in fmri. *Magnetic resonance in medicine*, 39(1):41–52, 1998.

[37] Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price. Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477, 2000.

[38] Karl J Friston, CJ Price, Paul Fletcher, C Moore, RSJ Frackowiak, and RJ Dolan. The trouble with cognitive subtraction. *Neuroimage*, 4(2):97–104, 1996.

[39] Mitsuo Gen and Runwei Cheng. *Genetic algorithms and engineering optimization*, volume 7. John Wiley & Sons, 2000.

[40] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[41] David E Golberg. Genetic algorithms in search, optimization, and machine learning. *Addion wesley*, 1989:102, 1989.

[42] David E Goldberg. The theory of virtual alphabets. In *Parallel problem solving from nature*, pages 13–22. Springer, 1991.

[43] David E Goldberg, Bradley Korb, and Kalyanmoy Deb. Messy genetic algorithms: Motivation, analysis, and first results. *Complex systems*, 3(5):493–530, 1989.

[44] John J Grefenstette. Optimization of control parameters for genetic algorithms. *IEEE Transactions on systems, man, and cybernetics*, 16(1):122–128, 1986.

[45] Alexandre Guillaume, Seugnwon Lee, Yeou-Fang Wang, Hua Zheng, Robert Hovden, Savio Chau, Yu-Wen Tung, and Richard J Terrile. Deep space network scheduling using evolutionary computational methods. In *Aerospace Conference, 2007 IEEE*, pages 1–6. IEEE, 2007.

[46] David J Heeger and David Ress. What does fmri tell us about neuronal activity? *Nature Reviews Neuroscience*, 3(2):142–151, 2002.

[47] John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[48] Gregory S Hornby, Al Globus, Derek S Linden, and Jason D Lohn. Automated antenna design with evolutionary algorithms. In *AIAA Space*, pages 19–21, 2006.

[49] Scott A Huettel. Event-related fmri in cognition. *Neuroimage*, 62(2):1152–1156, 2012.

[50] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

[51] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, second edition, 2009.

[52] James Alexander Hughes. jGP. https://github.com/jameshughes89/jGP, March 2015. Accessed: April 4, 2018.

[53] James Alexander Hughes, Joseph Alexander Brown, and Adil Mehmood Khan. Smartphone gait fingerprinting models via genetic programming. In *Evolutionary Computation (CEC), 2016 IEEE Congress on*, pages 408–415. IEEE, 2016.

[54] James Alexander Hughes, Joseph Alexander Brown, Adil Mehmood Khan, Asad Masood Khattak, and Mark Daley. Analysis of symbolic models of biometric data and their use for action and user identification. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2018 IEEE Conference on*, pages FIX–ME. IEEE, 2018.

[55] James Alexander Hughes and Mark Daley. Finding nonlinear relationships in fmri time series with symbolic regression. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*, pages 101–102. ACM, 2016.

[56] James Alexander Hughes and Mark Daley. Nonlinear model of a nonlinear system: An alternative view of fmri modelling (poster). In *11th Annual Canadian Neuroscience Meeting*. Canadian Association of Neuroscience (CAN), May 2017.

[57] James Alexander Hughes and Mark Daley. Searching for nonlinear relationships in fmri data with symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1129–1136. ACM, 2017.

[58] James Alexander Hughes and Mark Daley. Some title (poster). In *12th Annual Canadian Neuroscience Meeting*. Canadian Association of Neuroscience (CAN), May 2018.

[59] James Alexander Hughes, Ethan C Jackson, and Mark Daley. Modelling intracranial pressure with noninvasive physiological measures. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on*, pages 1–8. IEEE, 2017.

[60] James Alexander Hughes, Ethan C Jackson, and Mark Daley. Modelling intracranial pressure with noninvasive measures and genetic programming (poster). In *London Health Research Day 2018*, May 2018.

[61] Ilknur Icke, Nicholas A Allgaier, Christopher M Danforth, Robert A Whelan, Hugh P Garavan, and Joshua C Bongard. A deterministic and symbolic regression hybrid applied to resting-state fmri data. In *Genetic Programming Theory and Practice XI*, pages 155–173. Springer, 2014.

[62] Ethan C Jackson, James Alexander Hughes, and Mark Daley. On the generalizability of linear and non-linear region of interest-based multivariate regression models for fmri data. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2018 IEEE Conference on*, pages FIX–ME. IEEE, 2018.

[63] Cezary Z Janikow and Zbigniew Michalewicz. An experimental comparison of binary and floating point representations in genetic algorithms. In *ICGA*, pages 31–36, 1991.

[64] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[65] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. FSL. *Neuroimage*, 62(2):782–790, 2012.

[66] Yaochu Jin. A comprehensive survey of fitness approximation in evolutionary computation. *Soft computing*, 9(1):3–12, 2005.

[67] Brian Knutson, Andrew Westdorp, Erica Kaiser, and Daniel Hommer. Fmri visualization of brain activity during a monetary incentive delay task. *Neuroimage*, 12(1):20–27, 2000.

[68] Srinivas Kodiyalam, Somanath Nagendra, and Joel DeStefano. Composite sandwich structure optimization with application to satellite components. *AIAA journal*, 34(3):614–621, 1996.

[69] John R Koza. *Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems*. Stanford University, Department of Computer Science, 1990.

[70] John R Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.

[71] John R Koza. *Genetic programming II: automatic discovery of reusable programs*. MIT press, 1994.

[72] John R Koza. *Genetic programming III: Darwinian invention and problem solving*, volume 3. Morgan Kaufmann, 1999.

[73] John R Koza, David Andre, Forrest H Bennett III, and Martin A Keane. Use of automatically defined functions and architecture-altering operations in automated circuit synthesis with genetic programming. In *Proceedings of the First Annual Conference on Genetic Programming*, pages 132–140. MIT Press, 1996.

[74] John R Koza, Martin A Keane, Matthew J Streeter, William Mydlowec, Jessen Yu, and Guido Lanza. *Genetic programming IV: Routine human-competitive machine intelligence*, volume 5. Springer Science & Business Media, 2006.

[75] John R Koza and James P Rice. Automatic programming of robots using genetic programming. In *AAAI*, volume 92, pages 194–207, 1992.

[76] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[77] F Kruggel, Stefan Zysset, and D Yves von Cramon. Nonlinear regression of functional mri data: an item recognition task study. *Neuroimage*, 12(2):173–183, 2000.

[78] Frithjof Kruggel and D Yves von Cramon. Modeling the hemodynamic response in single-trial functional mri experiments. *Magnetic Resonance in Medicine*, 42(4):787–797, 1999.

[79] Kenneth K Kwong, John W Belliveau, David A Chesler, Inna E Goldberg, Robert M Weisskoff, Brigitte P Poncelet, David N Kennedy, Bernice E Hoppel, Mark S Cohen,

and Robert Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679, 1992.

[80] Pedro Larranaga, Cindy MH Kuijpers, Roberto H Murga, and Yosu Yurramendi. Learning bayesian network structures by searching for the best ordering with genetic algorithms. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 26(4):487–493, 1996.

[81] Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

[82] Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. Modeling the hemodynamic response function in fmri: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198, 2009.

[83] Hod Lipson and Jordan B Pollack. Automatic design and manufacture of robotic lifeforms. *Nature*, 406(6799):974–978, 2000.

[84] Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

[85] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–157, 2001.

[86] Jason D Lohn, Gregory S Hornby, and Derek S Linden. An evolved antenna for deployment on nasa's space technology 5 mission. In *Genetic Programming Theory and Practice II*, pages 301–315. Springer, 2005.

[87] Sean Luke et al. Genetic programming produced competitive soccer softbot teams for robocup97. *Genetic Programming*, 1998:214–222, 1998.

[88] Mary-Ellen Lynall, Danielle S Bassett, Robert Kerwin, Peter J McKenna, Manfred Kitzbichler, Ulrich Muller, and Ed Bullmore. Functional connectivity and brain networks in schizophrenia. *The Journal of Neuroscience*, 30(28):9477–9487, 2010.

[89] Andre C. Marreiros. Dynamic causal modeling. http://www.scholarpedia.org/article/Dynamic_causal_modelling, February 2010. Accessed: April 10, 2018.

[90] John Christopher Miles, GM Sisk, and Carolynne Jane Moore. The conceptual design of commercial buildings using a genetic algorithm. *Computers & Structures*, 79(17):1583–1592, 2001.

[91] Julian F Miller. An empirical study of the efficiency of learning boolean functions using a cartesian genetic programming approach. In *Proceedings of the Genetic and Evolutionary Computation Conference*, volume 2, pages 1135–1142, 1999.

[92] Julian F Miller. *Cartesian genetic programming*. Springer, 2011.

[93] Julian F Miller, Peter Thomson, and Terence Fogarty. Designing electronic circuits using evolutionary algorithms. arithmetic circuits: A case study, 1997.

[94] Heinz Mühlenbein, M Schomisch, and Joachim Born. The parallel genetic algorithm as function optimizer. *Parallel computing*, 17(6):619–632, 1991.

[95] Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

[96] Steven E Petersen and Joseph W Dubis. The mixed block/event-related design. *Neuroimage*, 62(2):1177–1184, 2012.

[97] Chrisila B Pettey, Michael R Leuze, and John J Grefenstette. Parallel genetic algorithm. In *Genetic algorithms and their applications: proceedings of the second International Conference on Genetic Algorithms: July 28-31, 1987 at the Massachusetts Institute of Technology, Cambridge, MA*, 1987.

[98] Riccardo Poli, William B Langdon, Nicholas F McPhee, and John R Koza. *A field guide to genetic programming*. Lulu. com, 2008.

[99] Alberto A Ochoa Rodriguez and Marta R Soto Ortiz. Partial evaluation in genetic algorithms. In *Proc. Int. Conf. Ind. Eng. Appl. Artif. Intell. Expert Syst*, pages 217–222, 1997.

[100] Cristina Rosazza and Ludovico Minati. Resting-state brain networks: literature review and clinical applications. *Neurological Sciences*, 32(5):773–785, 2011.

[101] Bruce R Rosen, Randy L Buckner, and Anders M Dale. Event-related functional mri: past, present, and future. *Proceedings of the National Academy of Sciences*, 95(3):773–780, 1998.

[102] Michael Schmidt and Hod Lipson. Comparison of tree and graph encodings as function of problem complexity. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1674–1679. ACM, 2007.

[103] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.

[104] Michael D Schmidt and Hod Lipson. Coevolving fitness models for accelerating evolution and reducing evaluations. In *Genetic Programming Theory and Practice IV*, pages 113–130. Springer, 2007.

[105] Michael D Schmidt and Hod Lipson. Coevolution of fitness predictors. *Evolutionary Computation, IEEE Transactions on*, 12(6):736–749, 2008.

[106] Michael D Schmidt, Ravishankar R Vallabhajosyula, Jerry W Jenkins, Jonathan E Hood, Abhishek S Soni, John P Wikswo, and Hod Lipson. Automated refinement and inference of analytical models for metabolic networks. *Physical biology*, 8(5):055011, 2011.

[107] Robert E Smith, Bruce A Dike, and SA Stegmann. Fitness inheritance in genetic algorithms. In *Proceedings of the 1995 ACM symposium on Applied computing*, pages 345–350. ACM, 1995.

[108] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

[109] Lee Spector, Howard Barnum, Herbert J Bernstein, and N Swamy. Genetic programming for quantum computers. *Genetic Programming*, pages 365–373, 1998.

[110] Lee Spector, Howard Barnum, Herbert J Bernstein, and Nikhil Swamy. Quantum computing applications of genetic programming. In *Advances in genetic programming*, pages 135–160. MIT Press, 1999.

[111] Lee Spector, Howard Barnum, Herbert J Bernstein, Nikhil Swamy, et al. Finding a better-than-classical quantum and/or algorithm using genetic programming. In *Proceedings of the Congress on Evolutionary Computation*, volume 3, pages 2239–2246, 1999.

[112] Lee Spector and Herbert J Bernstein. Communication capacities of some quantum gates, discovered in part through genetic programming (with additional figures from the qcmc 2002 poster). In *In Proc. 6th Int. Conf. Quantum Communication, Measurement, and Computing (QCMC*, 2003.

[113] Olaf Sporns. *Networks of the Brain*. MIT press, 2011.

[114] Olaf Sporns. Contributions and challenges for network models in cognitive neuroscience. *Nature neuroscience*, 17(5):652–660, 2014.

[115] Olaf Sporns and Rolf Kötter. Motifs in brain networks. *PLoS Biol*, 2(11):e369, 2004.

[116] Timothy Starkweather, Darrell Whitley, and Keith Mathias. *Optimization using distributed genetic algorithms*. Springer, 1991.

[117] Klaas Enno Stephan, Lars Kasper, Lee M Harrison, Jean Daunizeau, Hanneke EM den Ouden, Michael Breakspear, and Karl J Friston. Nonlinear dynamic causal models for fmri. *Neuroimage*, 42(2):649–662, 2008.

[118] Reiko Tanese. Parallel genetic algorithm for a hypercube. In *Genetic algorithms and their applications: proceedings of the second International Conference on Genetic Algorithms: July 28-31, 1987 at the Massachusetts Institute of Technology, Cambridge, MA*. Hillsdale, NJ: L. Erlhaum Associates, 1987., 1987.

[119] Daniel Tuhus-Dubrow and Moncef Krarti. Genetic-algorithm based approach to optimize building envelope design for residential buildings. *Building and environment*, 45(7):1574–1581, 2010.

[120] Alan M Turing. Computing machinery and intelligence. *Mind*, pages 433–460, 1950.

[121] Alberto L Vazquez, Eric R Cohen, Vikas Gulani, Luis Hernandez-Garcia, Ying Zheng, Gregory R Lee, Seong-Gi Kim, James B Grotberg, and Douglas C Noll. Vascular dynamics and bold fmri: Cbf level effects and analysis considerations. *Neuroimage*, 32(4):1642–1655, 2006.

[122] Alberto L Vazquez and Douglas C Noll. Nonlinear aspects of the bold response in functional mri. *Neuroimage*, 7(2):108–118, 1998.

[123] Tor D Wager, Alberto Vazquez, Luis Hernandez, and Douglas C Noll. Accounting for nonlinear bold effects in fmri: parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage*, 25(1):206–218, 2005.

[124] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[125] Mark W Woolrich, Brian D Ripley, Michael Brady, and Stephen M Smith. Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, 14(6):1370–1386, 2001.

[126] Tingting Zhang, Fan Li, Marlen Z Gonzalez, Erin L Maresh, and James A Coan. A semi-parametric nonlinear model for event-related fmri. *Neuroimage*, 97:178–187, 2014.

[127] Victor Zykov, Efstathios Mytilinaios, Bryant Adams, and Hod Lipson. Robotics: Self-reproducing machines. *Nature*, 435(7039):163–164, 2005.

Appendix A

Published Work for Paper 1

This appendix includes the published 2 page extended abstract for *Finding Nonlinear Relationships in fMRI Time Series with Symbolic Regression* [55] along with the corresponding presented poster. The submitted work was presented in Chapter 4.

Finding Nonlinear Relationships in fMRI Time Series with Symbolic Regression

James Alexander Hughes
Computer Science, Brain and Mind Institute
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3K7
jhughe54@uwo.ca

Mark Daley
Computer Science, Brain and Mind Institute
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3K7
mdaley2@uwo.ca

ABSTRACT

The brain is an intrinsically nonlinear system, yet the dominant methods used to generate network models of functional connectivity from fMRI data use linear methods. Although these approaches have been used successfully, they are limited in that they can find only linear relations within a system we know to be nonlinear.

This study employs a highly specialized genetic programming system which incorporates multiple enhancements to perform symbolic regression, a type of regression analysis that searches for declarative mathematical expressions to describe relationships in observed data.

Publicly available fMRI data from the Human Connectome Project were segmented into meaningful regions of interest and highly nonlinear mathematical expressions describing functional connectivity were generated. These nonlinear expressions exceed the explanatory power of traditional linear models and allow for more accurate investigation of the underlying physiological connectivities.

Keywords

Symbolic regression; Computational neuroscience; Functional magnetic resonance imaging; Nonlinear relationships

1. INTRODUCTION

Literature in the field of neuroscience explicitly acknowledges the existence of nonlinear relationships in brain function [1, 3], but it is common to treat them as a footnote or ignore them altogether [2, 3]. Linear tools, such as the General Linear Model (GLM) or the Pearson product-moment coefficient are used, almost exclusively, to model functional magnetic resonance imaging (fMRI) time series. Despite this, neuroscientific studies are able to make contributions with limited linear model [1]; however, it would ultimately be improper to use linear methods to observe what we *know* to be nonlinear phenomenon as it lacks the power to truly model the underlying processes. It is not surprising that the nonlinear relationships are ignored; discovering underlying nonlinearities is an exceptionally non-trivial task, especially when working with large amounts of high-dimensional data.

In this work Genetic Programming (GP) is implemented to automate the discovery of minimal and interpretable network relationships in the behavior of a system for which we can observe only time series derived from a network's nodes: task based fMRI time series data. No prior knowledge or assumptions are applied to the system, such as linearity or how the system interacts with itself.

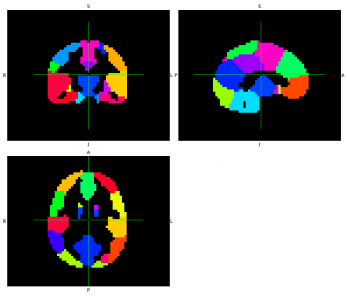


Figure 1: Snapshot of a brain segmented into the 30 ROIs. Each color represents a different region.

2. EXPERIMENTAL METHODS

The task based fMRI time series data selected was of a Motor task and was obtained from the Human Connectome Project, WU-Minn Consortium[1]. This four-dimensional data (three-dimensional brain over time) was collected into *30 spatial regions of interest* (ROIs) (Figure 1) for the time series of *284 time points*, and can be represented as a two-dimensional matrix of *30* columns with *284* rows.

This specific GP implementation is motivated by Schmidt et al.'s work [6], is extremely specialized for symbolic regression, and incorporates modular improvements which significantly increase performance. These improvements including parallel evolution of subpopulations, fitness predictors [5], and an acyclic graph representation [4].

For symbolic regression, it was required to have some value over the time series that evolved expressions fits to. For the purpose of this motor task, *ROI 21* was selected for the left hand side of the equation as it is the ROI that contains the *primary motor cortex*. *100* models for all *507* subjects available were generated.

[1]http://www.humanconnectome.org/

GECCO '16 July 20-24, 2016, Denver, CO, USA

ACM ISBN 978-1-4503-4323-7/16/07.
DOI: http://dx.doi.org/10.1145/2908961.2909021

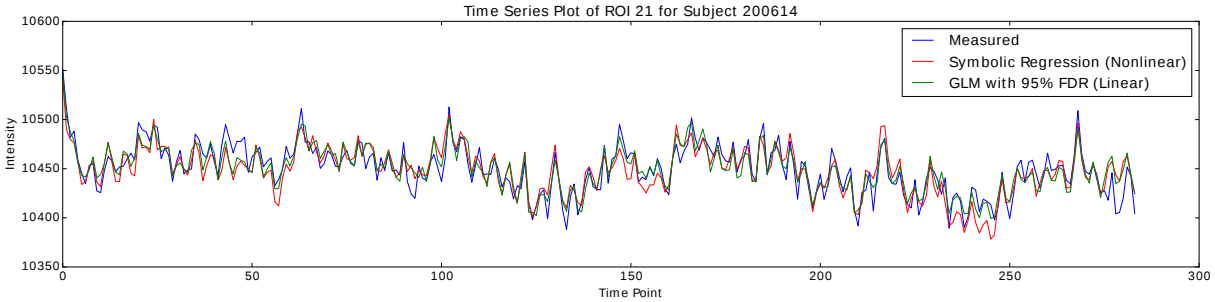


Figure 2: Time series of ROI 21's signal compared to the generated nonlinear and linear models.

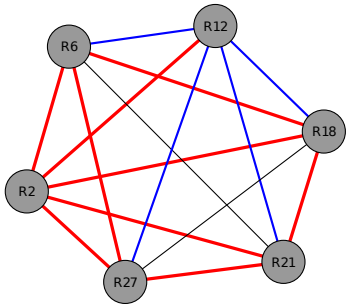


Figure 3: Relationships between ROIs for a single generated *nonlinear model*. Red represents nonlinear relationships *between ROIs*, blue represents nonlinear and linear relationships, and black is strictly linear. This particular example corresponds to the equation: $R21 = R12 - sin(11.97 * (18.30 - R12)) - (0.42 * |(R12 - R18) * R27|)/(R6 - tan(R2))$.

3. RESULTS AND CONCLUSIONS

With Pearson product-moment correlation and false discovery rate (FDR) thresholding (typical linear methods), almost every ROI is linearly related to ROI 21 (on average, 28 ROIs were related to ROI 21 per subject). Performing linear regression with this many ROIs generates models with high degrees of freedom that fit the data well, but provide minimal insight and are difficult to interpret.

Figure 2 shows a time series of one subject's recorded signal alongside two models describing the signal — one found with the nonlinear tool (Figure 3), the other with linear regression after thresholding ROIs with a 95% FDR. The *mean absolute error* over the time series for the top nonlinear models and the thresholded linear models were averaged over all subject. These values were roughly *16.68 (sd = 3.51)* and *11.79 (sd = 1.11)* respectively. Although both models fit the data well, a Mann-Whitney U test (U-test) provides a p-value of $3.08 * 10^{-133}$, which demonstrates that the linear models fit the recorded signal better.

On average, a nonlinear model contained fewer than 4 ROIs (3 when excluding ROI 21). The mean absolute time series error of the linear models generated with the top 4 correlated ROIs — which were typically the same ROIs as those found with GP — was calculated to be approximately *19.16* (sd = *5.08*). A U-test comparing the 4 ROIs models provided a p-value of $8.56 * 10^{-19}$; *the nonlinear models were significantly better*. In fact, it was not until the linear models were given the top *8 ROIs* that there was no more statistical difference. Linear models only performed better than the nonlinear models with 4 ROIs once they received *10 or more ROIs* (U-test p-value of $1.34 * 10^{-3}$); it took at least 10 ROIs for a linear model to fit the recorded signal better than a nonlinear model containing only 4.

When compared to linear models generated with all ROIs available after a typical thresholding technique, nonlinear models, although close, could not fit the signal as well. However, these linear models would typically contain more than 28 ROIs and would be difficult to interpret and provide minimal insight into understanding the underlying processes. Nonlinear models, in contrast, were more succinct and describe nonlinear relationships that would otherwise *not be discovered with conventional tools*. On average, with just 4 ROIs, a nonlinear model could fit the recorded signals better than linear models using 8; even with more information (ROIs), linear models could not describe the data as clearly.

4. ACKNOWLEDGMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). Computations were enabled by the SciNet HPC Consortium. Data were provided by the Human Connectome Project, WU-Minn Consortium.

5. REFERENCES

[1] R. Buckner and T. Braver. Event-related functional mri. In P. Bandettini and C. Moonen, editors, *Functional MRI*, chapter 36, pages 441–452. Springer-Verlag.

[2] M. Daley. An invitation to the study of brain networks, with some statistical analysis of thresholding techniques. In *Discrete and Topological Models in Molecular Biology*, pages 85–107. Springer, 2014.

[3] N. K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

[4] M. Schmidt and H. Lipson. Comparison of tree and graph encodings as function of problem complexity. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1674–1679. ACM, 2007.

[5] M. D. Schmidt and H. Lipson. Coevolution of fitness predictors. *Evolutionary Computation, IEEE Transactions on*, 12(6):736–749, 2008.

[6] M. D. Schmidt, R. R. Vallabhajosyula, J. W. Jenkins, J. E. Hood, A. S. Soni, J. P. Wikswo, and H. Lipson. Automated refinement and inference of analytical models for metabolic networks. *Physical biology*, 8(5):055011, 2011.

Appendix B

Genetic Programming System Details

B.1 Brief Version History

The genetic programming (GP) system created for this work was based on work done by Hod Lipson's research group [13, 103, 106]. It was implemented in *Java* version 8 with no additional libraries beyond *java.io* and *java.util*. A *C++* version was implemented, but provided no significant speed up and has since been deprecated.

The system is called *jGPvX* (java genetic programming) where *X* is the version number[1]. It is currently in its 9th version and is continuously being developed and improved. The first iteration was completed in *March 2015* but has since changed significantly.

This first version was very basic and did not include any noteworthy improvements. Subsequent version incorporated improvements such as the island model and fitness predictors (discussed in Chapter 2). Various bugs were also worked out throughout development.

Version 4 threaded the system such that the distributed populations (islands) ran on separate cores. This greatly decreased runtimes as the fitness evaluation was parallelized.

Typically the system spends more than *95%* of the CPU time evaluating the fitness of the candidate solutions. The fitness evaluation is performed with tail end recursion. Version 5 attempted to eliminate the recursive evaluation with a stack, however this yielded no improvement as Java's complier already optimizes tail end recursion effectively.

Around the beginning of the second phase of the project when the runtimes increased significantly, in order to not hit computing resources runtime limits, save states were incorporated into the system. After a parametrized number of generations, the whole population is saved in a serialized file to be reloaded into a subsequent run of the search. This enabled longer runtimes by allowing the user to simply submit a new job continuing a search from the saved state. Additionally, if a search failed because of some system issue with the computing resource (memory error, power outage), the search could continue from a saved state, as opposed to starting from generation 0.

Version 9 has *15 classes* and a total of roughly *2200* lines of code with typical white space. A GitHub repository of the current GP system can be found at `https://github.com/jameshughes89/jGP` [52].

Figures 2.1 and 2.7 give a high level overview of how the evolutionary search is executed.

[1]This is a working title as jGP is already in use for an implementation of genetic programming in java

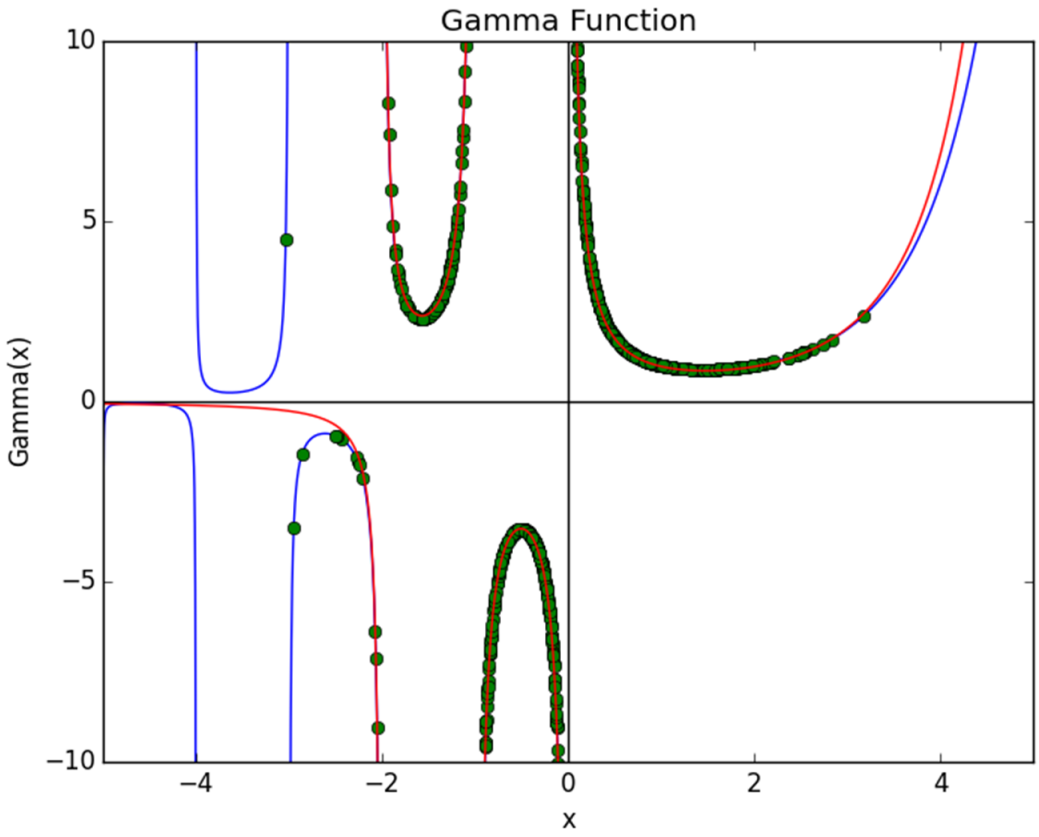


Figure B.1: Gamma function and approximation of the gamma function where $-3 < x < 3$. Blue is the gamma function, the green points are the $(x, \Gamma(x))$ pairs provided to the GP system, and red is the model derived by the GP system.

B.2 Early Test

During preliminary tests of the system, data from simple functions were provided to the system. These functions would be simple linear and nonlinear expressions. Unfortunately the functions provided are now lost, but would have been similar to: $x + \sin(y\dot{z}) + 5$. This data was provided to me by Mark Daley (supervisor), however Mark never told me what the functions were and would only confirm if the GP system found the correct function. On all but one, the GP system found exact, or near exact functions (near exact would happen when real numbers were introduced).

The only exception to finding exact or near exact functions was when the GP system was provided data from the gamma function. Hundreds of $(x, \Gamma(x))$ pairs where $-3 < x < 3$ were given to the system and the system developed the model $y = \frac{3.59+e^x}{5.43-2x+5.43+(5.43-x)x} + \frac{5.43-x}{5.43308-x+(5.43308-x)x)x}$, where y is an approximation of $\Gamma(x)$. Figure B.1 plots the gamma function, the data points provided to the GP system, and the resulting model. The GP system was not capable of deriving the actual gamma function based on the provided basis functions, however it developed an effective approximation of it where $-3 < x < 3$.

B.3 Resources, System Settings, and Runtimes

For the majority of the project, the system was executed on the general purpose cluster (GPC) at *SciNet*, a high performance computing system. The GPC consists of 2780 nodes with at least 16GB of memory per node of 8 Intel cores. Keeping in mind the stochastic nature of the algorithm and that the run times are dependent on the specific problem being solved, each run of the evolutionary search using this GP system searching for the nonlinear relationships takes between *1-4* hours when running with 8 cores (one node) on an *IBM System x iDataPlex dx360 M3* node with 2 quad-core *Intel Nehalem (Xeon 5540)* processors running at 2.53GHz. The system was also executed on *Graham*, *Guillimin*, *Mammouth Parallèle 2 (MP2)*, *Orca*, and *Cedar*.

Runtimes differed significantly throughout the course of the project for a variety of reasons: the search itself is stochastic so each execution would take different amounts of time, the computing resources were not the same throughout the project, and the tasks being studied had a different number of time points (176 – 405).

The system parameters also affected runtimes in four ways. First, changing the probability of a mating event occurring would affect the runtime. Second, the population size, number of generations, and number of island migrations affected the number of potential mating events by orders of magnitude. Third, the number of data points in the fitness predictors would affect the time to evaluate fitness values. Forth, the maximum size of an acyclic graph would alter the size of the models, and larger models take longer to evaluate.

The system settings for each paper are discussed in detail in their respective chapters. For the first phase/paper, a total of $1,750,000,000$ potential mating events were allowed and the runtimes were between 1 – 4 hours per model on Compute Canada resources. In the second phase, $7,070,000,000$ potential mating events were allowed with runtimes taking between 5 – 24 hours. For the third phase, the same number of mating events were allowed, but the number of data points included in the fitness predictors, and the maximum size of the models was increased, resulting in runtimes between 24 – 124 hours.

Although the incorporated enhancements improved the search, they had interesting impacts on the runtime. For example, although fitness predictors reduced the number of data points to be evaluated for fitness evaluation, which would reduce the time required to evaluate a candidate solution, it would also typically slow down the speed of convergence. The evolutionary search could run for many more generations and find much better solutions, which would in turn increased the total runtime.

Total runtime for the GP system on Compute Canada resources is reported in Table B.1. These numbers reflect the total amount of work on all phases of the project to date.

The only system setting of note that seems unusual is the high mutation rate. This was a consequence of the representation/genotype/candidate solution encoding. Due to the nature of the genotype, mutations might make alterations to the candidate solution that do not actually manifest in an actual alteration in the phenotype. For this reason, the mutation rate was set higher.

Table B.1: Total CPU usage in core years for the project over a 4 year period. All project GP system executions were done on Compute Canada resources.

Year	CPU Usage (core years)
2015	238.45
2016	536.54
2017	768.32
2018	8.26
Total	1551.58

Curriculum Vitae

Name:	James Hughes
Post-Secondary Education and Degrees:	Brock University St. Catharines, ON 2008 - 2012 BSc (Hon) Brock University St. Catharines, ON 2013 - 2014 MSc University of Western Ontario London, ON 2014 - 2018 PhD
Related Work Experience:	Contract Professor/Limited Duties Instructor The University of Western Ontario & Brock University 2014 - 2018 Teaching Assistant The University of Western Ontario & Brock University 2012 - 2015 Business Analyst/Quality Assurance Ministry of Transportation — Ontario 2010 - 2011

Publications:

Theses

Hughes, J. (2014). A study of ordered gene problems featuring DNA error correction and DNA fragment assembly with a variety of heuristics, genetic algorithm variations, and dynamic representations.

Hughes, J. (2012).Recentering, Reanchoring & Restarting an Evolutionary Algorithm.

Refereed Conference Publications

Jackson, E. C., **Hughes, J. A.**, & Daley, M. (2018). On the Generalizability of Linear and Non-Linear Region of Interest-Based Multivariate Regression Models for fMRI Data. *Accepted to IEEE's 2018 Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB).*

Houghten, S., Collins, T. K., **Hughes, J. A.**, & Brown, J. A. (2018). Edit Metric Decoding: Return of the Side Effect Machines. *Accepted to IEEE's 2018 Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB).*

Hughes, J. A., Brown, J. A., Khan, A. M., Khattak, A. M., & Daley, M. (2018). Analysis of Symbolic Models of Biometric Data and their use for User and Task Identification. *Accepted to IEEE's 2018 Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB).*

Hughes, J. A., Jackson, E. C., & Daley, M. (2017, August). Modelling intracranial pressure with noninvasive physiological measures. *In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on* (pp. 1-8). IEEE.

Jackson, E. C., **Hughes, J. A.**, Daley, M., & Winter, M. (2017, August). An algebraic generalization for graph and tensor-based neural networks. *In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on* (pp. 1-8). IEEE.

Hughes, J. A., & Daley, M. (2017, July). Searching for nonlinear relationships in fMRI data with symbolic regression. *In Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 1129-1136). ACM.

Hughes, J. A., & Daley, M. (2016, July). Finding Nonlinear Relationships in fMRI Time Series with Symbolic Regression. *In Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion* (pp. 101-102). ACM.

Hughes, J. A., Brown, J. A., & Khan, A. M. (2016, July). Smartphone gait fingerprinting models via genetic programming. *In Evolutionary Computation (CEC), 2016 IEEE Congress on* (pp. 408-415). IEEE.

Hughes, J., Houghten, S., Mallen-Fullerton, G. M., & Ashlock, D. (2014, May). Recentering and restarting genetic algorithm variations for DNA fragment assembly. *In Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on* (pp. 1-8). IEEE.

Hughes, J., Houghten, S., & Ashlock, D. (2013, August). Recentering, reanchoring & restarting an evolutionary algorithm. *In Nature and Biologically Inspired Computing (NaBIC), 2013 World Congress on* (pp. 76-83). IEEE.

Hughes, J., Brown, J. A., Houghten, S., & Ashlock, D. (2013, June). Edit metric decoding: Representation strikes back. *In Evolutionary Computation (CEC), 2013 IEEE Congress on* (pp. 229-236). IEEE.

Journal Papers

Hughes, J. A., Houghten, S., & Ashlock, D. (2016). Restarting and recentering genetic algorithm variations for DNA fragment assembly: The necessity of a multi-strategy approach. *Biosystems*, 150, 35-45.

Hughes, J. A., Houghten, S., & Ashlock, D. (2014). Recentering and restarting a genetic algorithm using a generative representation for an ordered gene problem 1. *International journal of hybrid intelligent systems*, 11(4), 257-271.

Mallén-Fullerton, G. M., **Hughes, J. A.**, Houghten, S., & Fernàndez-Anaya, G. (2013). Benchmark datasets for the DNA fragment assembly problem. *International Journal of Bio-Inspired Computation*, 5(6), 384-394.

Book Chapters

Hughes, J. A., Houghten, S., & Ashlock, D. (2017). Permutation Problems, Genetic Algorithms, and Dynamic Representations. *In Nature-Inspired Computing and Optimization* (pp. 123-149). Springer, Cham.

Posters

Hughes, J. A., Jackson, E. C., & Daley, M. (2018) Modelling Intracranial Pressure with Non-invasive Measures and Genetic Programming. *At: London ON, Conference: London Health Research Day (2018)*: LHRD 2018, Accepted

Hughes, J. A., & Daley, M. (2018) On the Generalizability of Nonlinear Models of fMRI Data and the True Model Selection Problem. *At: Vancouver, Conference: 12th Annual Canadian Neuroscience Meeting (2018)*: CAN 2018, Accepted

Hughes, J. A., & Daley, M. (2017) Nonlinear Model of a Nonlinear System: An Alternative view of fMRI Modelling. *At: Montreal, Conference: 11th Annual Canadian Neuroscience Meeting (2017)*: CAN 2017, DOI: 10.13140/RG.2.2.17508.58246

Hughes, J. A., & Daley, M. (2016) Finding Nonlinear Relationships in fMRI Time Series. *At: Denver Colorado, Conference (2016)*: GECCO 2016, DOI: 10.13140/RG.2.1.4489.1128

Honours and Awards:

Compute Canada Resource Allocation
2017-2018

NSERC PGS D
2016-2018

Doctoral Excellence Research Award
2016-2018

Queen Elizabeth II Scholarship (Declined)
2016-2017

Ontario Graduate Scholarship
2015-2016

Ontario Graduate Scholarship
2014-2015

IEEE Computational Intelligence Society CIBCB Best Student Paper Award
2014

IEEE Computational Intelligence Society Travel Grant
2014

Dr. Raymond and Mrs. Sachi Moriyama Graduate Fellowship
2014

Jack M. Miller Excellence in Research Award
2014

W. D. Hatch Memorial Scholarship
2014

Goldsmith-Wyatt Mathematics & Science Scholarship
2013

Dean of Graduate Studies Brock University Research Fellowship
2013

Dean of Graduate Studies Excellence Award
2013

Dean of Graduate Studies Entrance Scholarship
2013

NSERC USRA
2012