Generating Nonlinear Models of Functional Connectivity from Functional Magnetic Resonance Imaging Data with Genetic Programming

James Alexander Hughes
St. Francis Xavier University
Mathematics, Statistics, and Computer Science
Antigonish, Nova Scotia, Canada
jhughes@stfx.ca

Mark Daley
University of Western Ontario
Computer Science Department
London, Ontario, Canada
mdaley2@uwo.ca

***Abstract*—The brain is a nonlinear computational system; however, most methods employed in finding functional connectivity models with functional magnetic resonance imaging (fMRI) data produce strictly linear models — models incapable of truly describing the underlying system.**

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 linear models, the nonlinear models contained fewer regions of interest and were never significantly worse when applied to data the models were fit to. Nonlinear models could generalize to unseen data from the same subject better than traditional linear models (intrasubject). Nonlinear models could not generalize to unseen data recorded from other subjects (intersubject) as well as the linear models, and reasons for this are discussed. This study presents the problem that many, manifestly different models in both operators and features, can effectively describe the system with acceptable metrics.

***Index Terms*—Computational Neuroscience; Functional Connectivity; Functional Magnetic Resonance Imaging; Genetic Programming; Symbolic Regression.**

I. Introduction

The brain is a provably nonlinear computational system[1]. Although the literature explicitly acknowledges this [1], [2], [3], [4], [5], [6], it is deemphasized or ignored, especially when working with *functional Magnetic Resonance Imaging* (fMRI) data. To better understand the brain as a computational system, researchers will create *functional connectivity* models of the brain — network relationship models of the statistical relationships between the spatially distributed regions of the brain during cognition. Despite being an intrinsically nonlinear system, almost all strategies for functional connectivity modeling use linear tools (Pearson Product-moment correlation coefficient, general linear model).

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, perhaps using a powerful method capable of describing the nonlinearities will help us understand the intricacies of the nonlinear dynamic complex system — the brain.

In addition to understanding the brain as a distributed natural information processing system, functional connectivity models have important clinical applications. These functional networks manifest differently in individuals with certain neural disorders, such as Alzheimer's [7] and schizophrenia [8], when compared to otherwise healthy control subjects. By improving model effectiveness with better modeling technologies, finer details and differences may be distilled, making it possible to differentiate between cohorts of interest.

Although nonlinear tools have been developed and studied, they remain underrepresented within the neuroscience literature. Friston *et al.* use nonlinear tools such as *Volterra series expansion* to study the balloon model [5], [9] and *dynamic causal modeling* to study *effective connectivity* — a modeling technique similar to functional connectivity which incorporates temporal dynamics [10]. Kruggel *et al.* used a form of nonlinear regression to model relationships between the hemodynamic response and stimulus conditions [11]. *Semi-parametric Volterra series* was used by Zhang *et al.* to find nonlinearities [12]. With *symbolic regression*, Allgaier *et al.* found novel nonlinear relationships within known networks in *resting state* fMRI data [13], [14]. Hughes & Daley used symbolic regression to develop nonlinear functional connectivity models of *task based* fMRI data. Not only did the nonlinear models fit their data better, but they contained fewer relationships when compared to linear models [15], [16]. Jackson *et al.* performed a similar analysis to that done by Hughes & Daley with independently gathered data. They found similar results and demonstrated that the nonlinear models did not overfit the data any more than typical linear methods [17].

In this work *Genetic Programming* (GP) will be used to perform *Symbolic Regression*. GP is a computational intelligence technique where, through a strategy based on the natural

[1]A human can simulate a Turing machine, therefore they are *at least* as powerful as a Turing machine — a nonlinear computational system.

process of evolution, the algorithm writes (*evolves*) its own programs to solve problems [18]. 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 [15], [16]; we develop nonlinear models of real fMRI data gathered from subjects performing a variety of tasks. These models provide interpretable functional connectivity network relationships between brain areas to ultimately give new insight into the underlying system. The motivation is *not* to create predictive models, but to develop *descriptive* models; the goal is the generation of interpretable functional connectivity model, not to collect the model's output. These objectives are ultimately one and the same since an accurate descriptive model will be capable of prediction, and we are using accuracy/prediction to measure quality, but the subtle difference is emphasized to frame the motivation for the creation of these models.

This work focuses on the application of GP to real data from a real world problem to make novel contributions to another fields. Although the fMRI data was recorded and preprocessed carefully to ensure the highest quality data possible, there are still many concessions that need to be made when working with this real data regardless of the mathematical techniques used for modeling. These are discussed within Section II.

Although this is based on the observations and contributions of previous work, we begin the analysis from scratch based on newly generated models. In this work we include additional subjects for greater insight and statistical significance. We also expand the analysis performed to include a deeper model validation investigation by applying models to unseen data recorded from the same subject (intrasubject generalization) and unseen data from subjects the models were not fit to (intersubject generalization).

Nonlinear models were generated that fit their data and generalized to unseen data from the same subject better than the traditional linear models; however, the nonlinear models could not generalize to unseen data from different subjects as well. The authors discuss why the linear models may be better at intersubject generalization (fitting all data *well*, but no specific subject/task effectively) and the practicality and real utility of it from a neuroscientific perspective. We finish with remarks on the difficulty of model selection when presented with a collection of different models with similar error values.

II. Neuroscientific Data

The data studied in this work was obtained from the *Human Connectome Project, WU-Minn Consortium*[2] (HCP) — a large online database of neuroimaging data. Full details on the data and preprocessing pipelines are made available by the HCP. As of January 2019, 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.

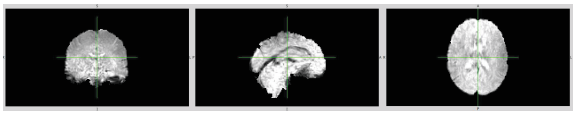

Fig. 1: A three-dimensional snapshot of the four-dimensional fMRI data. Each voxel in this snapshot contains its BOLD signal from a single time point. If the recording had t time points, then there would t snapshots for that recording.

We focus on the *task-based fMRI* recordings. This data are recorded from subjects performing a task while inside an fMRI scanner. The scanner records the changing blood oxygen level dependent (*BOLD*) signal — a measure of the relative oxygenation level of blood within tissue — over time which can be used to indicate functional activation. Although the exact nature of the BOLD signal is not well understood [3], it has been shown to be an effective proxy for brain activation [19], [20], [21]. Like many real world problems, the recorded data has many deficiencies. The BOLD signal is notoriously noisy, spatially diffused, has a relatively low temporal resolution, difficult/expensive to obtain (sparse), and lags behind actual neural activity.

The BOLD signal from the three-dimensional brain over time is recorded and represented as voxels (Figure 1). Voxels are three-dimensional analogues to two-dimensional pixels; just as a pixel in a gray scale image would contain the intensity value of the localized pixel, a voxel in the brain contains the localized BOLD signal.

This four-dimensional data (three-dimensional snapshots of the brain over time's one dimension) 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. The actual number of voxels depends on the resolution of fMRI scanner; modern hardware with a resolution on the order of *2-5*mm^3 can capture hundreds of thousands of voxels. Although each voxel's size is on the order of millimeters, it contain tens of thousands of neurons. Similarly, the number of time points depends on the hardware and overall duration of the experiment. Modern scanners are capable of capturing whole brain volumes at a frequency of *0.75Hz – 2Hz*.

The *seven* tasks performed by subjects for the 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/197s), and Working Memory (405 time points/292s). The temporal resolution of the scans were *720ms* per sample (*∼1.389Hz*).

Given the significant computational cost of developing the nonlinear models (discussed in Section IV-B), data from all tasks for *forty* of the 1200 subjects were arbitrarily chosen

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

and analyzed. Since fMRI data is difficult and expensive to obtain, it is common to have a very limited amount of data from the same subject performing the exact same task. This makes it difficult to test the effectiveness of functional connectivity models, regardless of the modeling technique used. Fortunately, 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 [21]). The phase encoding does not affect the data between the two separate scans, therefore, for simplicity the LR phase encoding data was used as the *training* data, and the RL was used as independent *testing* data (although the choice was arbitrary). Given the sparsity and few number of time points for the fMRI recordings, we are precluded from splitting the data further into a validation set as there would be far too few data points to fit to. The authors want to clearly acknowledge this limitation; however, the use of fitness predictors (discussed in Section IV-A) provides a limited pseudo validation set throughout evolution.

Data was *z-score* (standard score) normalized since data from the fMRIs were not already normalized between sessions. The data was segmented into 30 meaningful regions of interest (ROIs) with Craddock *et al.*'s *spatially constrained parcellation* [22]. 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 higher resolution is desirable; however, 30 is not out of line with other ROI based neuroscientific studies and allowed for the generation of models in a reasonable amount of time. It is expected that higher or lower resolutions would work similarly well in general (as shown in [17]). A high level overview of the ROIs can be found in Table I.

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

TABLE I: 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

III. Neuroscientific Motivation

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

Almost all task based fMRI studies employ linear methods to generate models. These methods include *Pearson product-moment correlation coefficient* and the *Generalized Linear Model* (GLM). The typical strategy used to develop a functional connectivity model is as follows. If one wanted to derive how a given *ROI X* was functionally connected to all other ROIs, one would **(1)** calculate the correlation between ROI X's timeseries to all other ROIs and **(2)** perform some correction for multiple comparisons (*false discovery rate* (FDR) or *Bonferroni correction* (BC)). **(3)** Statistically unrelated ROIs are eliminated and the **(4)** remaining ROIs will be used as regressors in our linear regression to ROI X. **(5)** The resulting functional connectivity model and beta weights will be used to indicate which areas of the brain are functionally related during a task, and to what extent. In other words, the linear equation generated is simply analyzed to determine which ROIs it contains and what their beta weights are.

These methods assume that the underlying system is linear; however, we know this to be incorrect. 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 a part 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 simple models; complex models tend to overfit, are hard to interpret, and typically have greater computational costs to generate. But, perhaps by using a more powerful method we can find more accurate and descriptive models of the brain. There are many approaches one could employ to find nonlinearities within the data. One could perform linear regression of many nonlinear basis functions to obtain very low errors, but these would overfit significantly, and how does one select such basis

functions? An artificial neural network would likely fit the data well; however, it would be difficult to interpret the resulting model. Here we use symbolic regression to replace steps **1 – 4** discussed above because it performs feature selection, is at least as expressive as linear regression, eliminates many of the assumptions the linear methods make, and it produces a symbolic model that can be interpreted in a similar way to what is already done with the linear models (step **5**).

IV. Methods

A. Genetic Programming Implementation

A GP system based on Schmidt *et al.*'s work was developed [23]. This system is specialized for symbolic regression and includes many improvements to increase performance and speed. Although many ideas are incorporated into the system, noteworthy ones include fitness predictors [24], [25] and an acyclic graph representation [26]. This GP system has also been shown to be robust to noise [27]. These ideas are summarized below, but full descriptions are available from their respective sources. The implementation of the GP system has been made available online [28].

Fitness predictors reduce the computational cost of the search by approximating the local search gradient [24], [25]. 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 may benefit by focusing on that area. This subset of data is always *evolving* throughout the search as the data points required to create the disagreement between candidate solutions will depend on the current population. Given that only a small and dynamic subset of data is being fit to at a given time, it provides a pseudo/simulated automatic validation throughout evolution. Although there are similar techniques [29], 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 improve results.

Figure 2 provides a high level view of the execution of the search. There are two major routines: *Solutions* and *Predictors*. Solutions executes like a standard evolutionary algorithm with subpopulations evolving independently and recombining periodically until some stopping criteria is met. Predictors select the fitness predictors to be used to calculate the fitness values of the evolving candidate solutions.

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 [26]. Figure 3 shows an example of the phenotype implemented in the system used and demonstrates how it represents an acyclic graph. In our implementation, array indexes 0 and 1 must be terminals (literal or variable). The last index in the array is the root of the tree. Each element in the array will reference some number of lower indexes: 2 if the element is a binary operator, 1 if it is a unary operator, and 0 if it is a terminal. Some elements are non-coding genes (denoted as *?*) that do not impact the phenotype. It has been shown that these non-coding genes can function as some *vestigial memory* and become expressed later in evolution with positive effects [26].

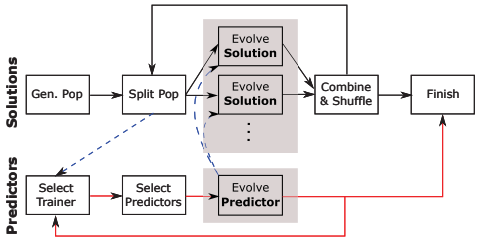

Fig. 2: High level view of the algorithm execution. Two major sections: *Solutions* depicts the flow of evolving candidate solutions; *Predictors* shows the flow of the evolving fitness predictors. Dashed lines denote communication between the solution and predictor routines. Evolving candidate solutions use the current predictors to evaluate their fitness.

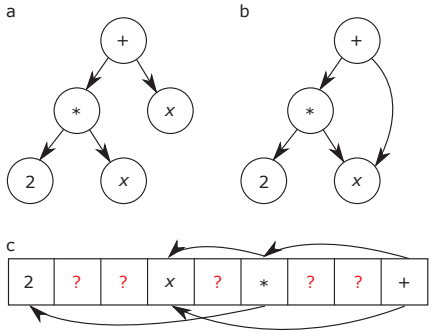

Fig. 3: a, b, and c all represent the expression $2 \cdot x + x$. a represents the typical tree-based representation. b shows how the tree based structure could be represented as an acyclic graph. c shows how the acyclic graph can be represented as an array. It is important to note that this representation will not simplify the expression to $3x$.

B. Genetic Programming Settings

The system parameters used are presented in Table II. These values were determined through preliminary tests; however, no significant parameter sweep was performed.

Crossover was a simple one-point crossover. The strategy for mutation was to randomly select a gene and replace it with a randomly chosen operator/terminal from the language. 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 (non-coding genes).

TABLE II: Parameter settings for GP System. The last 4 settings are specific to the improvements discussed in IV-A.

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

Evolution concludes once a predefined number of generations have occurred.

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. Although these operations can frequently be observed in nature and there is no reason that they cannot be found within the data, we make no assertion that they necessarily exist within brain function as no reference for such a thing exists.

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.

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.

C. Experimental Methods

Forty 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 each subject and task. For linear regression, *six* different ways of generating models were investigated: (1) fitting all ROI; (2) performing FDR and thresholding then fitting; (3) performing BC and thresholding then fitting; (4) fitting all ROIs with LASSO regression; (5) performing FDR and thresholding then fitting with LASSO regression; (6) performing BC and thresholding then fitting with LASSO regression. *Least absolute shrinkage and selection operator* (LASSO) regression is 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 ROIs compared to linear regression [16], so it is of interest to compare symbolic regression to a linear method with similarly succinct models (LASSO).

The authors emphasize that they are comparing *six* deterministic linear methods to *one* nondeterministic nonlinear method. Given the stochastic nature of the modeling technique, it is necessary to create many models for each subject/task combination. In the neuroscience literature it is common to generate a single linear model for each subject/task. We make clear that comparing 100 nonlinear models to six introduces bias into the analysis and take special care to remind the reader of this where appropriate.

For both linear and symbolic regression, an ROI known to be involved with the task was chosen to be the dependent variable (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 dependent variable for the emotion, gambling, language, motor, relational, social, and working memory tasks were ROIs 7, 2, 12, 21, 28, 3, and 21 respectively.

V. Results and Discussion

A. Model Effectiveness

Table III 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. We also include a Mann-Whitney U test's p-value obtained by comparing the nonlinear and linear models' distributions of mean absolute errors (MAEs) from all subjects performing the same task. The results show that nonlinear models are comparable to the linear when applied to the data they were fit to. The only linear modeling strategy capable of consistently outperforming the nonlinear models was when all ROIs were used in regular linear regression; however, they were never significantly better. Additionally, this linear strategy is not employed in the literature as these models would likely overfit and provide no neuroscientific insight since they used all features (a model containing all ROIs would indicate that every ROI is functionally related to all other ROIs for that subject/task).

Figure 4 presents a *p-value transition* plot. This plot was generated by comparing the top nonlinear models' errors from all subjects 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 one linearly correlated ROI (to the ROI chosen to be the dependent variable). More ROIs are added in the order of absolute correlation score until all ROIs are included (the last row). The average number of ROIs (over all subjects) in a linear model with BC and FDR is written on the plot along with the average number of ROIs in all (100) nonlinear models generated for each subject (NL-A-) and the average number of ROIs in the top nonlinear model for each subject (NL-T-).

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

Although the models are symbolic, they can be fairly intricate. The authors do not suggest taking a nonlinear model

TABLE III: 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 MAEs of the nonlinear models to the respective linear model.

	Nonlinear		BC LASSO			FDR LASSO			BC			FDR			ALL LASSO			ALL		
	Mdn	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

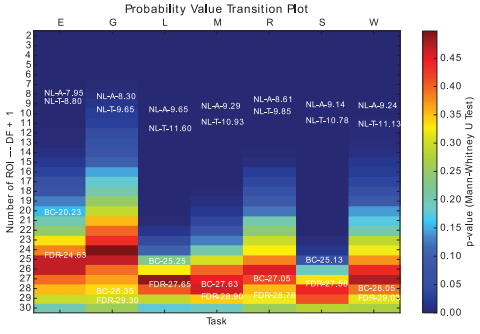

Fig. 4: Comparing linear and nonlinear models' MAEs (averaged over all subject) as the number of ROIs used to create the linear model increases. ROIs were added to the linear models in the order of their absolute correlation score. The number of ROIs in the nonlinear models was fixed.

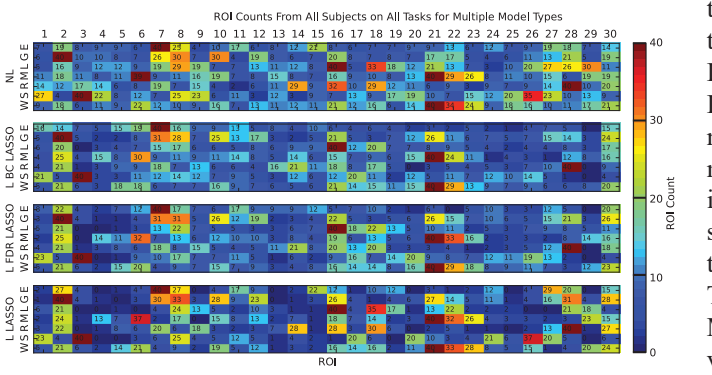

Fig. 5: Number of subjects for each ROI (column) that appeared in the top model for each task (row). Counts for the nonlinear (NL) and LASSO generated linear (L) models are presented. 40 is maximum. Note that the ROI corresponding to the dependent was in all models.

and expecting it to be an exact representation of the functional connectivities. It is not the specific operators found in the model, but the presence of ROIs and the fact that they are related in some nonlinear way that is of interest. Figure 5 shows how often each ROI appeared in the 40 subjects' models for each task. The first matrix shows the results for the top nonlinear models and the bottom three show the results from three linear. Although all matrices are similar, the one for the nonlinear is the most distinct. Not only are the nonlinear models effective, but they contain somewhat different ROIs. For example, ROI 27 (Medial Orbitalfrontal) appeared in many nonlinear models of the language task, but was not found in many linear models — perhaps this is an important functional relationship that has been missed by traditional tools. Only the LASSO models are shown since the regular linear regression models typically contained nearly all (if not all) ROIs.

B. Intersubject Generalizability

Although it has been observed in the literature that there is a large intersubject variability in network models [30], it is still of interest to test our models' ability to generalize to other subjects. Figure 6 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 MAEs were then averaged over all subjects performing the same task. Only the LASSO linear models are included as the others did not generalize to other subjects as well. The diagonals are of particular interest as they show how well, on average, models for a 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 be noted that the LASSO models also fit all other task's data well. Perhaps these LASSO models are not as capable of describing task specific nuances. It can be seen that the nonlinear models' accuracy matrix is very dependent on the task, and although the matrix is of model MAEs when applied to different data, it provides some visualization of which task's models can fit independent tasks' data similarly well (somewhat of a similarity matrix). Take note of the similarity between the Motor and Working Memory Task. Both tasks used ROI 21 as the dependent variable. Perhaps the similarity is a consequence of the choice of independent variable, or maybe there is some functional similarities between these tasks.

C. Intrasubject Generalizability

If we take the top nonlinear model and the 6 linear models and apply them to unseen *test* data from the same subject and task, we can compare the resulting MAEs 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 during each task (eg. moving hand, foot, tongue) were different. Fortunately, this should not matter since the models are temporally independent. Figure 7 plots the training and testing errors against each other. Unsurprisingly, nearly all points for all models are above the $y = x$ line, indicating that the models obtain better errors on the data they were fit to.

Given the stochastic nature of the evolutionary search, a total of 100 nonlinear models were generated for each subject and task combination for statistical power and 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.

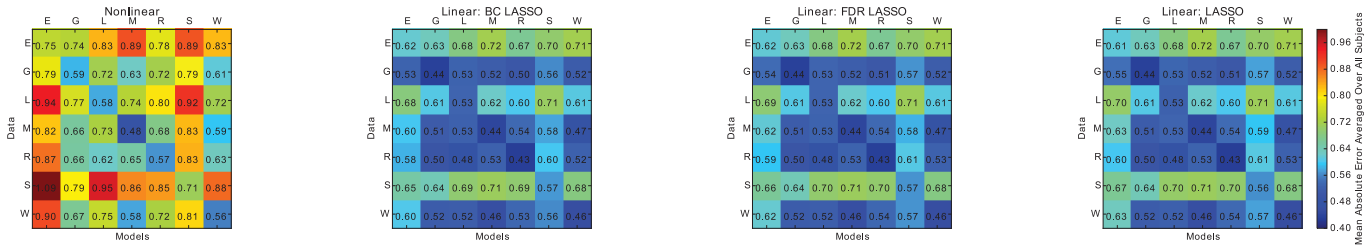

Fig. 6: Matrices showing the MAE values obtained by applying every task/subject combination's models to all other datasets and averaged over all subjects performing the same task. The diagonal provides a means of quantifying intersubject generalization; if all subject's models on the same task can fit all other subject's data from that task similarly well, then the models are capable of generalizing between subjects.

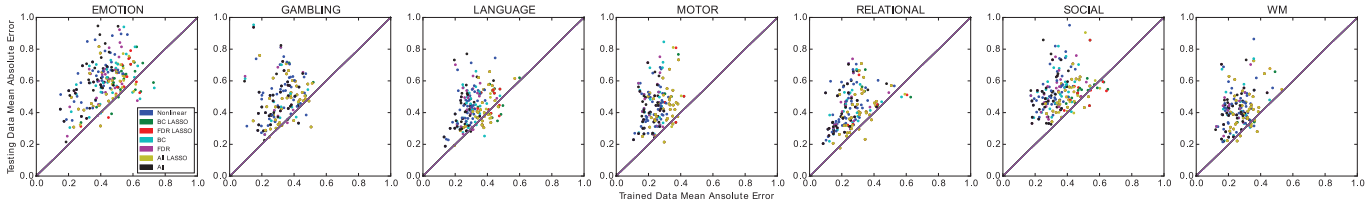

Fig. 7: Scatter plot comparing the training and testing MAEs for all models. For the nonlinear model, the top model on the training data was compared to it's error when applied to unseen data. The difference between the training and testing errors averaged over all subjects in tasks for each model are: 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.

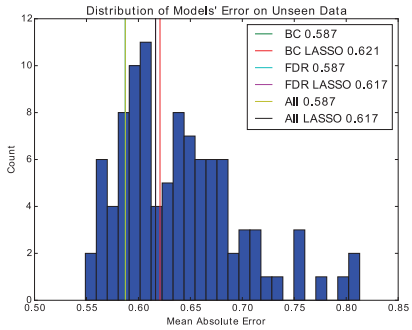

Fig. 8: Distribution of MAE values when applying all 100 nonlinear models to unseen data from the same subject performing the same task. Vertical lines correspond to the MAEs obtained by linear models.

Figure 8 shows the distribution of MAEs from the 100 models along with vertical lines indicating the MAEs from the 6 linear models fit to the same data as the nonlinear and applied to the same unseen data. From this one example we can see that some number of the 100 nonlinear models performed better than the best linear.

Distributions like Figure 8 can be generated for each of the 280 subject and task combination. Figure 9 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 further away from this line the greater the difference between the nonlinear and linear model's error. 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 IV shows the average difference between the models

TABLE IV: Average difference between the best nonlinear and linear models' MAEs when the respective column's model was best. The values are averaged over all subjects performing the same task. Eg: 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

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.

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 available. Although these results are still meaningful, this limitation is important to keep in mind when interpreting these results.

Figure 9 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 models obtained a lower error than the best linear model. Figure 10 shows a distribution of how many nonlinear models were better than the best linear model for all subjects (if such models exist). Simply, when referring to Figure 8, it would be the number of nonlinear models to the left of the leftmost (smallest) linear model's error. These numbers were collected for all subjects and the distributions are plotted for each task. For many subjects, numerous nonlinear models generalized to

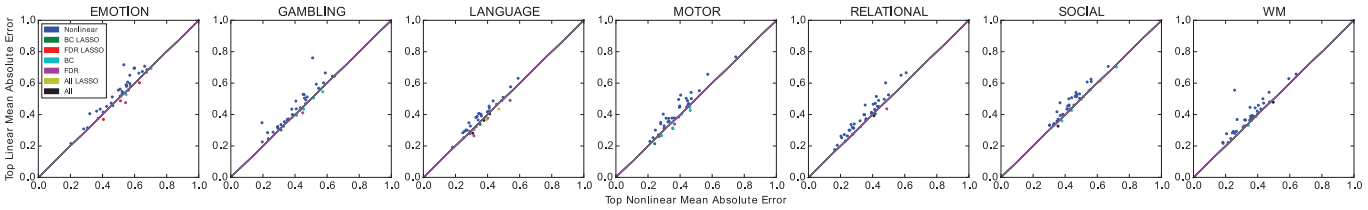

Fig. 9: Scatter plot comparing the smallest MAE from the 100 nonlinear models when applied to unseen data versus the best of the 6 linear models. Points above the $y = x$ line indicate that the nonlinear model was best. Points below indicate that a linear model was best. Color indicates method for model generation.

TABLE V: Mann-Whitney U test p-values for comparing the originally selected nonlinear model for intersubject generalization (lowest error on training) to the model selected based on its performance at intrasubject generalizability (lowest error on intrasubject testing) — denoted by *pseudo*.

Task #	*pseudo* vs org.	*pseudo* vs LASSO ALL
Emotion	2.05e-01	5.97e-57
Gambling	1.22e-03	4.36e-23
Language	7.36e-07	7.71e-03
Motor	4.45e-07	4.54e-35
Relational	2.40e-03	3.70e-16
Social	6.56e-22	2.59e-11
W. Memory	7.34e-08	5.52e-13

unseen data better than the best linear, 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 than linear models.

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 still apply the top model on unseen data from the same subject to this pseudo third dataset. Similar to Figure 6, Figure 11 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 intersubject generalization. However, the better generalizing nonlinear models were still significantly worse than the best linear models. Table V contains the relevant p-values.

D. Model Selection Problem

The purpose of generating these models is to find a descriptive model that can provide insight into the underlying system. 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 others, 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 provided some consensus on which ROIs were meaningful, then we could use that information to develop our functional connectivity network. Figure 12 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 5, 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 12. 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 corresponds to observations about intersubject variability [30] and 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 ROIs), 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.

Further, 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 regarding typical practices in the neuroscience literature is: how can one generate a single linear model and expect it to be representative of the underlying system?

VI. Conclusions and Future Work

Nonlinear models of functional connectivity 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 contain fewer relationships than many linear models. The nonlinear models had different ROIs than the linear and contained nonlinear relationships — something not possible with traditional linear tools. The authors do caution the reader from expecting the nonlinear models as generated by GP to be exact models of the underlying phenomenon and only recommend extracting neuroscientific meaning from the presence of ROIs and the type of relationship (linear vs. nonlinear).

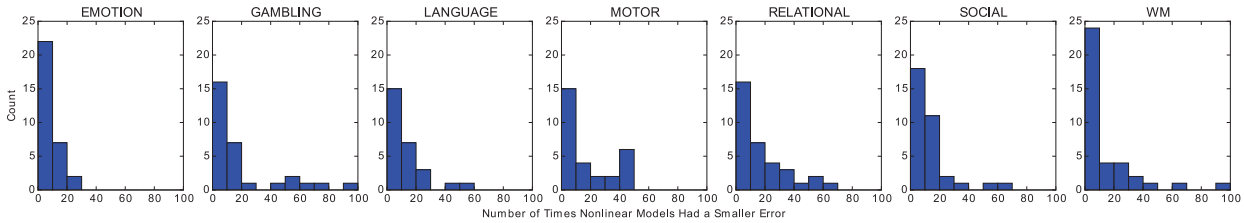

Fig. 10: For each subject, the number of the 100 nonlinear models generated that were better than the best linear model when applied to unseen data was calculated and the distributions were plotted. Bins (x-axis) represent the number of nonlinear models better than the best linear. The bin height (y-axis) corresponds to the number of subjects.

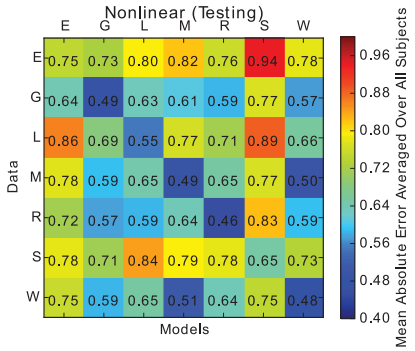

Fig. 11: Similar to Figure 6, this matrix shows the MAE values obtained by applying *the best model on the unseen data* from every subject/task to all other datasets and averaged over all subjects performing the same task.

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

The nonlinear models were unable to fit unseen data from other subjects as well as the traditional linear models (intersubject generalization). However, the linear models were capable of fitting other subject's data from unrelated tasks well. This may be a consequence of the linear models not fitting task specific nuances, but more general properties of the fMRI data. Additionally, the neuroscience literature already acknowledges that it is difficult to create effective intersubject generalizing models given the variability in functional connectivity models between subjects [30].

Although our motivation is descriptive models, testing model predictive ability allows us to evaluate model correctness. Many nonlinear models fit unseen data from the same subject better than linear models (intrasubject generalization); 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 intersubject generalizability of the nonlinear models, but the linear models still generalized between subjects better.

This work presents the problem of model selection, which is an idea related to complex systems in general. 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 or nonlinear. At the very least however, it would seem better to have a collection of models rather than a single linear model — something GP delivers.

This idea is related to a phenomenon seen and informally discussed in deep learning. When creating deep networks with a large set of parameters, it would be nearly impossible to find the *optimal* layout. However, by increasing the size of the search space through deep networks, we seemingly create many sufficiently *good* layouts that surprisingly generalize very well. In other words, it seems that by increasing the complexity of the search space, we increase the number of *good enough* models.

Further, it relates to the idea of *biological degeneracy* — multiple independent systems performing the same function under certain conditions [31], [30]. Functional degeneracy has been observed in neural populations [32], [30], and it is believed to play a significant role in the robustness and evolvability of complex systems in general. Although it is perhaps optimistic to suggest that our nonlinear models are different as a consequence of degeneracy, we emphasize it here to highlight a question: is the search for a single model fundamentally flawed?

Although similar work has shown that the nonlinear models do not overfit data any more than the linear [17], further investigation into generalizability is required. It is necessary to obtain additional data with multiple recordings from each subject performing each task. This would allow for a training, validation, and testing analysis to eliminate bias. Effect size should be analysed in addition to the statistical significance reported here. Performing various important measures on the the ROIs would improve the analysis over the simple feature count used in this work. An investigation into model consensus (Figure 12) could yield stronger evidence of functional connectivities. This could be achieved with some methods of thresholding, filtering, and data smoothing. Building functional connectivity models from multiple subjects' data may yield better intersubject generalization.

VII. 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

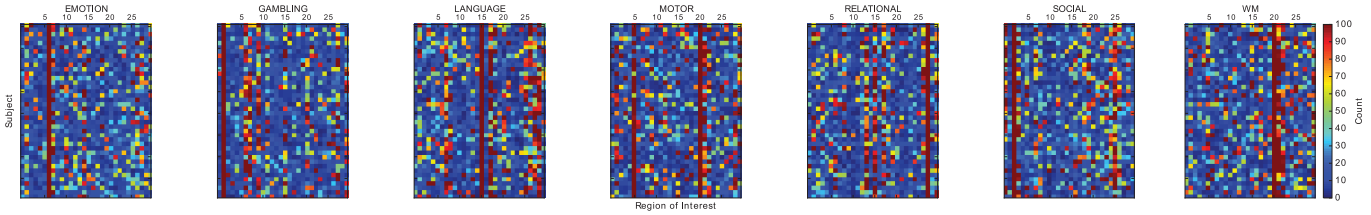

Fig. 12: Matrices showing the number of times (color) each ROI (column) appeared in the 100 nonlinear models generated for each subject (row) on each task. Note that the ROI corresponding to the left hand side of the equation was in all models.

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] 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*, vol. 16, no. 13, pp. 4207–4221, 1996.

[2] R.L. Buckner and T.S. Braver, "Event-related functional mri", in *Functional MRI*, P. Bandettini and C. Moonen, Eds., chapter 36, pp. 441–452. Springer-Verlag.

[3] 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*, pp. 85–107. Springer, 2014.

[4] Karl J Friston, CJ Price, Paul Fletcher, C Moore, RSJ Frackowiak, and RJ Dolan, "The trouble with cognitive subtraction", *Neuroimage*, vol. 4, no. 2, pp. 97–104, 1996.

[5] Karl J Friston, Oliver Josephs, Geraint Rees, and Robert Turner, "Nonlinear event-related responses in fmri", *Magnetic resonance in medicine*, vol. 39, no. 1, pp. 41–52, 1998.

[6] A L Vazquez and D C Noll, "Nonlinear aspects of the bold response in functional mri", *Neuroimage*, vol. 7, no. 2, pp. 108–118, 1998.

[7] 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", *Journal of neuroscience*, vol. 29, no. 6, pp. 1860–1873, 2009.

[8] 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", *Journal of Neuroscience*, vol. 30, no. 28, pp. 9477–9487, 2010.

[9] Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price, "Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics", *Neuroimage*, vol. 12, no. 4, pp. 466–477, 2000.

[10] Karl J Friston, Lee Harrison, and Will Penny, "Dynamic causal modelling", *Neuroimage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[11] F Kruggel, Stefan Zysset, and D Yves von Cramon, "Nonlinear regression of functional mri data: an item recognition task study", *Neuroimage*, vol. 12, no. 2, pp. 173–183, 2000.

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

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

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

[15] 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*. ACM, 2016, pp. 101–102.

[16] 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*. ACM, 2017, pp. 1129–1136.

[17] E. C. Jackson, J. A. Hughes, and M. Daley, "On the generalizability of linear and non-linear region of interest-based multivariate regression models for fmri data", in *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, May 2018, pp. 1–8.

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

[19] 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*, vol. 89, no. 13, pp. 5951–5955, 1992.

[20] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann, "Neurophysiological investigation of the basis of the fmri signal", *Nature*, vol. 412, no. 6843, pp. 150–157, 2001.

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

[22] 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*, vol. 33, no. 8, pp. 1914–1928, 2012.

[23] 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*, vol. 8, no. 5, pp. 055011, 2011.

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

[25] Michael D Schmidt and Hod Lipson, "Coevolution of fitness predictors", *Evolutionary Computation, IEEE Transactions on*, vol. 12, no. 6, pp. 736–749, 2008.

[26] 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*. ACM, 2007, pp. 1674–1679.

[27] 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*. IEEE, 2016, pp. 408–415.

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

[29] Y Jin, "A comprehensive survey of fitness approximation in evolutionary computation", *Soft computing*, vol. 9, no. 1, pp. 3–12, 2005.

[30] Olaf Sporns, *Networks of the Brain*, MIT press, 2010.

[31] Gerald M Edelman and Joseph A Gally, "Degeneracy and complexity in biological systems", *Proceedings of the National Academy of Sciences*, vol. 98, no. 24, pp. 13763–13768, 2001.

[32] Giulio Tononi, Olaf Sporns, and Gerald M Edelman, "Measures of degeneracy and redundancy in biological networks", *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 3257–3262, 1999.