universität wien

# MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

## Predicting real-time fMRI-based neurofeedback performance


verfasst von / submitted by

Fabian Marvin Renz, BSc


angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2020 / Vienna 2020

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | UA 066 840 |
| Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: | Masterstudium Psychologie UG2002 |
| Betreut von / Supervisor: | Univ.-Prof. Dr. Frank Scharnowski, MSc |
| Mitbetreut von / Co-Supervisor: | Dr. David Steyrl |

# Abstract

Real-time fMRI-based neurofeedback is an emerging scientific and clinical tool that allows for learning to self-regulate brain activity. It has been shown to modulate behaviour in healthy individuals, and further has demonstrated the capacity to improve clinical symptoms in various patient populations. However, the performance in self-regulating neural activity varies considerably across studies and individuals. Consistent learning curve patterns, such as steadily rising regulation performances across runs, are rare. Here, we investigate whether neurofeedback regulation performances across runs are merely random or follow a predictable pattern. This is achieved by applying machine-learning (L1-regularized Linear Regression & Randomized Trees) to predict the regulation performance of a training run based on previous training run performances. Additionally, we included subject- and study-specific characteristics such as age, sex, instructions, trained brain regions, and the length of regulation blocks in our machine-learning models to investigate how these factors affect performance. For assessing the relevance of each feature, we applied permutation-based feature importance analyses to our trained models. To obtain results that generalize across the field of real-time fMRI neurofeedback, our analyses was conducted on a large and heterogeneous real-time fMRI neurofeedback dataset of 197 participants from 11 different studies that included healthy participants as well as patients, different ROIs, and diverse experimental designs. We were able to predict regulation performance significantly better than chance level. However, with median $R^2$ values of up to 0.26 a considerable part of variance remains unexplained. For the predictions, previous regulation performances were the most crucial features. Overall, we found that performance in neurofeedback training is not random but to some degree predictable. These results might help to develop a better understanding of how self-regulation of brain activity with neurofeedback is accomplished, thus allowing for more effective clinical and scientific use of this promising method. Considering increased availability of suitable data in the context of the Open Science movement, our data-driven approach might become a promising avenue for advancing our understanding and the applicability of neurofeedback.

# Inhaltsverzeichnis

# Introduction

## Neurofeedback

Biofeedback generally involves monitoring and using physiological information to teach patients or participants to modify specific physiological functions (McKee, 2008). There are many different modalities biofeedback can be based upon. This thesis focuses on neurofeedback. Neurofeedback is a specific form of biofeedback, enabling to learn the volitional control of one's own brain activity (Sitaram et al., 2017). Neurofeedback provides the subjects with an opportunity to observe changes in their neural activity in real-time, which can then be leveraged to achieve voluntary control over one's own neural activity. Ultimately, the self-learned control of neural activity can result in changes in behaviour, brain function and cognition (Weiskopf et al., 2004).

In 1969, Spilker and colleagues published the first neurofeedback studies in humans (Spilker et al., 1969). Their study design investigated the operant control of alpha waves from human electroencephalogram (EEG). However, although EEG has a good temporal resolution, it suffers from poor spatial resolution, making it difficult to precisely locate the origin of a neural signal, or target deep brain regions. Therefore, not all brain regions can be trained using EEG based neurofeedback. To overcome these limitations, neurofeedback training using functional Magnetic Resonance Imaging (fMRI) was established over the last years, allowing for a higher spatial resolution and full brain coverage (Weiskopf et al., 2004).

## Functional Magnetic Resonance Imaging

Functional MRI enables the acquisition of physiological brain data. The underlying idea behind fMRI is to infer brain activity by local changes in blood oxygenation. This is possible because fMRIs' signal intensity is affected by the proportion of oxygenated and deoxygenated blood. For obtaining a map of brain activity, fMRI relies on the so called blood-oxygen-level-dependent (BOLD) contrast (Ogawa, Lee, Kay, et al., 1990; Ogawa, Lee, Nayak, et al., 1990). The BOLD contrast is a marker for physiological events changing the blood oxygenation state of the brain (Ogawa, Lee, Kay, et al., 1990). Since 1990 much progress has been made in the field of fMRI. In 2001, Logothetis reported that the physiological basis of the BOLD signal reflects an indirect index of neural activity (Logothetis et al., 2001).

Increasing neural activity has been shown to coincide with increases in local blood flow, meaning that more oxygenated blood is supplied to the activated cortical region. Although the local oxygen consumption is increased, the increase in supply of oxygenated blood exceeds the consumption, and results in an overall increase of blood oxygenation. The fMRI signal can discriminate between more or less active brain areas by exploiting the different magnetic properties of oxygenated and deoxygenated haemoglobin (Hb) in erythrocytes. On the one hand, oxygenated Hb is diamagnetic, resulting in small, negative magnetic susceptibility, which dephases magnetized proton spins slowly. This eventually leads to an increase in BOLD signal, which is indirectly associated with elevated neural activity. On the other hand, deoxygenated Hb is paramagnetic, thereby it has a larger positive contribution to the net magnetic field, distorting the magnetic field, and dephasing magnetized proton spins fast, resulting in a decreased BOLD signal (Mark et al., 2015).

**Real-time functional Magnet Resonance Imaging-based neurofeedback**

In 2012, the first conference on real-time functional magnetic resonance imaging (rtfMRI) neurofeedback took place at the Swiss Federal Institute of Technology in Zurich (Sulzer et al., 2013). RtfMRI was first introduced in 1995 by Cox et al., as any process relying on functional information from a MRI scanner, where data analysis and visualization are performed simultaneously with data acquisition (Cox et al., 1995). Recently, rtfMRI neurofeedback is receiving increasing attention in the fields of cognitive and clinical neuroscience, as it allows the training of voluntary control over brain activity in any localized region of interest (ROI). The information flow in rtfMRI neurofeedback can be conceptualized as a closed loop system as illustrated in Figure 1 from Weiskopf et al. (2004). Participants are instructed to regulate local brain activity using cognitive strategies. While the participants perform the regulation, the corresponding brain activity is recorded using fMRI. The acquired data is pre-processed correcting for artefacts caused, for example, by head motion or respiration. In the subsequent step the difference between regulation and most often a baseline is calculated. Based on the analysed neural activity a feedback is provided to the participants, closing the loop. Most of the times feedback is presented visually as a thermometer display or a continuous scrolling curve representing brain activation (Sulzer et al., 2013). The signal can be provided continuously or in an intermittent fashion. For the latter, the feedback is only updated in certain time intervals, usually after a few seconds (Emmert et al., 2017). However, feedback has also been implemented in different ways like, for example, in virtual

reality settings, such as computer games (Goebel et al., 2004). It needs to be noted that rtfMRI based neurofeedback is always delayed for two reasons: The first is the time it takes to perform the preprocessing and statistical calculations which must happen in real-time and can be computationally expensive. The second reason is the haemodynamic coupling which introduces a delay between the neuronal activation and the BOLD signal changes (S.-G. Kim & Bandettini, 2010). The onset of signal increase is delayed by about 3 seconds, and the peak signal by roughly 6 seconds, depending on the brain region and the task. Advances in computer hardware allow to minimize any additional delays. However, further advances in processing speed will only be of limited value because the intrinsic haemodynamic delay does not change (Weiskopf et al., 2004). So, although it is referred to it as real-time feedback, there is always a delay of a few seconds, depending on the chosen equipment and performed analyses.

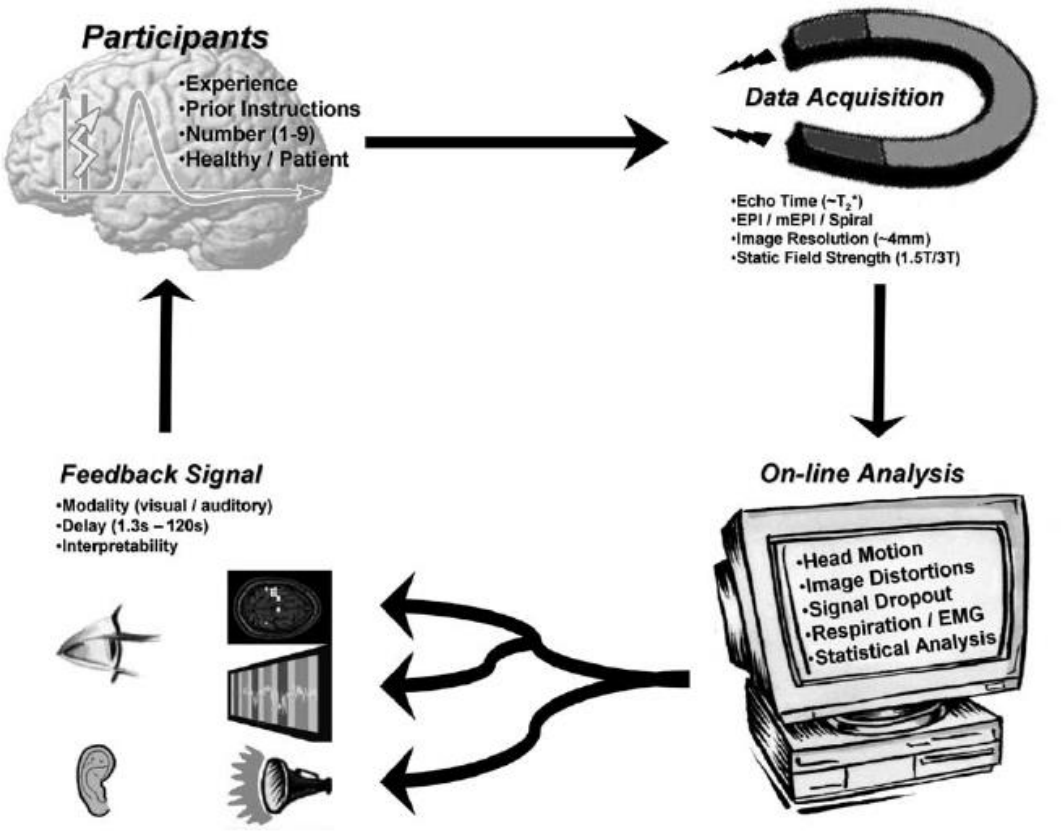



**Figure 1.** Image taken from Weiskopf et al. (2004). The set-up of a rtfMRI neurofeedback-based experiment. From the top-left the participants regulate their local brain activity.

Functional images are extracted from the MR scanner in real-time and analysed. The analysed data is presented to the participant for the purpose of neural self-regulation.

In the following rtfMRI-based neurofeedback study designs will be explained as well as target groups and different types of rtfMRI-based neurofeedback.

### *Experimental Design*

The experimental design in rtfMRI neurofeedback studies depends on the study objectives, which range from neurofeedback induced learning of neural self-regulation to improving clinical conditions (deCharms et al., 2005; Shibata et al., 2011). However, the majority of neurofeedback studies share a similar experimental framework as illustrated in Sulzer et al. (2013), which is comprised of 5 steps:

1. A definition of the physiological target area and the response measure. A region is defined either anatomically or a functional localizer is used to define the trained brain region and/or network.
2. Neurofeedback of the physiological target response and measurement of the subjects' performance. The participant is presented with feedback information on ongoing activity of the physiological target and aims to learn to control the activation using mental strategies. The feedback training can span from a few minutes or hours to several repeated sessions over multiple days.
3. Following the training comes the transfer: When participants have achieved successful regulation, they are tested to demonstrate whether they are able to maintain the skill of controlling brain activation or performing a task in the absence of feedback and/or in a different setting.
4. Experimental control group: studies either employ different control groups or within subject control conditions to control for confounds in learning, as well as behavioural and placebo effects.
5. Testing of behavioural effects: after the regulation training, it is tested whether there are specific behavioural effects. Typically, this is done in a pre- post-training comparison.

The majority of rtfMRI neurofeedback designs are using a block design for the regulation task. In this type of design, the participants are required to regulate the BOLD signal for

usually 15 – 30 seconds followed by a resting block with a similar duration. A single run usually consists of 3-6 blocks, while on a single experimental session 2 to 5 runs can take place.

***Target groups***

To this date various brain areas have been targeted using fMRI-based neurofeedback where the activity-based neurofeedback is the most widespread technique. Some examples include: the anterior cingulate cortex (deCharms et al., 2005; Emmert et al., 2014; Gröne et al., 2015; Guan et al., 2015; Li et al., 2013), dorsolateral prefrontal cortex (Sherwood et al., 2016), anterior insula (Yao et al., 2016), insula (Buyukturkoglu et al., 2013; Caria et al., 2007; Emmert et al., 2014; Frank et al., 2012; Zilverstand et al., 2015), amygdala (Nicholson et al., 2017; Young et al., 2014), auditory cortex (Emmert et al., 2017), motor cortex (Auer et al., 2015; Blefari et al., 2015; Buyukturkoglu et al., 2013; Marins et al., 2015; Scharnowski et al., 2015), visual cortex (Scharnowski et al., 2012) and the ventral tegmental area (MacInnes et al., 2016).

Recently, functional brain networks have also been successfully trained employing connectivity-informed neurofeedback in networks associated with motor control (Liew et al., 2016; Megumi et al., 2015), emotion regulation (Koush et al., 2015), default mode network (McDonald et al., 2017), attention (Koush et al., 2013), executive control (Spetter et al., 2017) and craving (D.-Y. Kim et al., 2015).

Next to improving behavioural and cognitive functions in healthy participants, rtfMRI neurofeedback has also been successfully applied in order to reduce clinical symptoms in neurological and psychiatric patient populations, such as borderline personality disorder (Paret et al., 2016), depression (Linden et al., 2012; Young et al., 2014), Huntington's disease (Papoutsi et al., 2018) or post-traumatic stress disorder (Nicholson et al., 2017).

***Real-time fMRI-based neurofeedback types***

There are different measures rtfMRI neurofeedback can be based upon. The most widely used option is activity-based neurofeedback. In this category the most common measures are the beta coefficients from the General Linear Model or Percent Signal Changes (PSC). Most of the times the feedback is calculated by the difference between the current regulation block and the preceding baseline in the same ROI. Another branch of

neurofeedback is concerned with training connectivity between two brain regions. For connectivity-based neurofeedback, the activity of two brain regions is simultaneously trained. Here the correlation between the respective activation of the brain regions is often used as a measure (Megumi et al., 2015). Another connectivity based neurofeedback measure is Dynamic Causal Modelling (DCM) (Koush et al., 2013; Watanabe et al., 2017). DCM requires defining hypotheses about the neural mechanisms underlying a fMRI measurement including the ROIs, connections between these ROIs as well as external inputs and context dependent manipulations of the network (Koush et al., 2013). In the next step Bayesian model comparison is used to discern which model explains the data best (Penny et al., 2004). Furthermore, DCM allows for estimating the model's parameters and thereby shedding light on the dynamic connectivity changes during an experiment. The 3rd group is multivariate pattern analysis (MVPA) feedback and Decoding neurofeedback (DecNef) (LaConte et al., 2007; Watanabe et al., 2017): Classic fMRI neurofeedback approaches in- or decrease the 1D amplitude of fMRI signals averaged across a ROI in the brain. However, MVPA feedback can change fMRI voxel patterns in the ROI rather than the mean amplitude of a ROI. Here, in the first step a decoder is constructed to classify a fMRI voxel pattern into one of different states for a participant in advance. Second, one of these states is selected as the target state for the following neurofeedback training. In the 3rd step, during each training trial the fMRI voxel pattern in the ROI of the participant is measured on a real-time basis, and the fMRI voxel pattern is input into the decoder, which then computes the likelihood of the targeted state. The feedback score which is provided to the participant is proportional to the likelihood of the targeted state. For a more detailed description of DecNef see Shibata et al. (2019).

**Learning in real-time fMRI-based neurofeedback**

RtfMRI-based neurofeedback has been successfully used in various settings as described in the previous sections. However, the underlying learning mechanisms that enable to gain control over one's own neural activity remain uncertain. Currently several different learning mechanisms are theorised to be engaged in neurofeedback learning (Sulzer et al., 2013).

On the psychological level one hypothesized learning mechanism which is supposedly crucially involved is associative learning. Associative learning describes the process of forming associations between behaviour and a stimulus. This comprises two sub forms:

classical and operant conditioning. The latter is often mentioned in the explanation of neurofeedback learning processes (Orndorff-Plunkett et al., 2017; Sulzer et al., 2013). In a nutshell, operant conditioning states that the probability of a physiological response is increased when a reinforcing stimulus follows that response, focusing on three components: 1. discriminative stimuli, 2. responses and 3. reinforcers.
In rtfMRI neurofeedback training one response can be reinforced in the presence of a discriminative stimulus but not in presence of another, for example when an upwards pointing arrow appears, but not when a downwards pointing one does. Following this learned association, the response probability will only increase for the first discriminative stimulus where the rtfMRI feedback of the brain activity functions as a reinforcing stimulus.

Next to investigating the psychological mechanisms, their corresponding physical instantiation on the hardware level is equally important when trying to understand the learning processes. Although not fully understood, the driving processes implementing associative learning are theorized to be a specific form of long-term potentiation, namely time dependent plasticity (Caporale & Dan, 2008; Sulzer et al., 2013). Long-term potentiation is the persistent strengthening of synaptic connections caused by recent patterns of activation, resulting in increased signal transmissions between two neurons (Cooke & Bliss, 2006). However, many questions remain open, like the fact that up to this point it cannot be determined whether the up-regulation of BOLD results in a neural excitation or inhibition (Gallistel & Matzel, 2013).

Summarising the excursion on learning in neurofeedback, it can be concluded that neither the psychological nor the biological level is fully understood. Progress in the understanding of the underlying learning mechanisms will be essential in the further, purposeful development of the field. For a more thorough discussion of learning in neurofeedback see Sherlin et al. (2011) and Sitaram et al. (2017).

**Predictability of regulation performance**

Various rtfMRI neurofeedback experiments report improvements in behaviour and the alleviation of clinical symptoms like for example Nicholson et al. (2017 or Young et al. (2014). However, the question arises whether we can observe the corresponding changes in neural activity. Figure 2 shows the PSCs of 120 participants over 4 training runs.

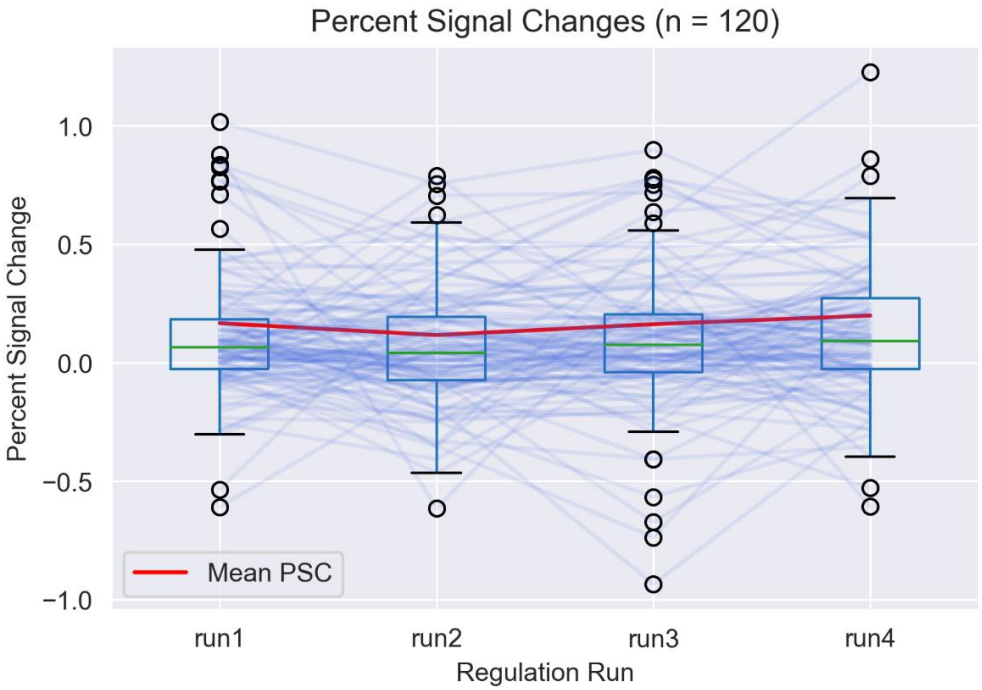



**Figure 2.** For 120 participants the Percent Signal Change values are plotted over 4 runs. Each individual blue line corresponds to one participants performance over the 4 runs.

Assuming improved self-regulation through neurofeedback training, a positive slope across training runs would be expected. However, visual inspection of the learning curves does not indicate such a trend, i.e., a regression line with positive slope. Hence, the question: Does regulation performance in rtfMRI-based neurofeedback training follow a systematic, predictable pattern? This thesis aims at answering this question by investigating whether it is possible to predict regulation performance measured in PSC in rtfMRI neurofeedback based on previous regulation performance and other study- and participant specific characteristics.

***Hypothesis:***

H0: Neurofeedback performance as indicated by PSC is random and does not reveal systematic patterns that can be predicted above chance level.

H1: Neurofeedback performance as indicated by PSC is not random but follows patterns that can be predicted above chance level.

In the following sections the precise approach will be explained for testing the hypotheses and assessing whether rtfMRI-based neurofeedback is predictable and induces systematic changes.

# Methods

## Overview

The approach this thesis takes is a meta-analytic one, including the data of several studies from the field of rtfMRI based neurofeedback. In order to test the hypothesis, we train statistical models, trying to predict the regulation performance of a given run, measured in PSC, based on previous regulation performance. For example, this would mean predicting the regulation performance of the participants in the fourth run based on the regulation performance in the first, second and 3rd runs. In a subsequent step next to the previous performance we incorporate additional features. The analysis then includes both the previous regulation performance, as well as study and participant specific information. As part of this thesis the 3rd, 4th, 5th, and 6th run are predicted. The 1st and 2nd are not predicted since there are too few features to base the prediction on. Furthermore, we did not predict runs after the 6th one, because there are too little available data, since only few participants from the studies included in the meta-analysis participated in more than 6 neurofeedback runs.

After training the models, the same models are trained on randomly shuffled data, to obtain the chance level prediction, in order to then compare the performances of the trained models to their chance level equivalent. The model's performances are captured by their explained variance as indicated by the coefficient of determination ($R^2$). To compare the different models to the corresponding chance-level predictions bootstrap tests are used. Furthermore, the individual importance of features and their contributions to the overall prediction using permutation-based feature importance are analysed.

The following sections report on the included data and the features, the used machine learning techniques, as well as the process of training the models.

## Data description

The data included in the meta-analysis were contributed from various labs all over the world and comprise the largest dataset available in rtfMRI neurofeedback. Since required data could not be extracted just form publications exclusively, suitable studies were identified via the real-time fMRI neurofeedback mailing list and by contacting authors directly by Amelie Haugg who provided the data. For increased generalizability, we did not limit this study to a

specific participant cohort or a target ROI. However, the employed statistical methods required the dataset to be homogenous with regard to the measure indicating the regulation performance. Therefore, only activity-based neurofeedback studies, as indicated by PSC, could be included in the analysis. Moreover, all included data measured one ROI and its respective Up/Downregulation compared to a baseline. The PSC of downregulation were inverted before the statistical analysis. For a detailed listing of the included studies see the corresponding Table 1.

| **Author** | **Participants** | **ROIs** |
|---|---|---|
| Emmert et al. (2017) | tinnitus (N=14) | auditory cortex |
| Kim et al. (2015) | tobacco use disorder (N=7) | Anterior cingulate cortex (ACC), medial prefrontal cortex (mPFC), orbitofrontal cortex (OFC) |
| MacInnes et al. (2016) | healthy (N=19) | Ventral tegmental area (VTA) |
| Papoutsi et al. (2018) | Huntington's disease (N=10) | Supplementary motor area (SMA) |
| Scharnowski et al. (2012) | healthy (N=10) | visual cortex |
| Shuxia et al. (2016) | healthy (N=18) | anterior insula |
| Young et al. (2017) | depression (N=18) | amygdala |
| Hellrung et al. (2018) | healthy (N=49) | amygdala |
| Hellrung (unpublished) | healthy (N=11) | insula |
| Papoutsi et al. (2018) | Huntington's disease (N=8) | SMA |
| Marxen et al. (2016) | healthy (N=32) | amygdala |

**Table 1**. Summary of included studies for the predictions of the real-time functional magnet resonance imaging-based neurofeedback performance.

The respective ROIs are illustrated in Figure 3. Overall, the analysed dataset contains 197 participants from a total of 11 studies. All participants completed at least 3 neurofeedback training runs. Some of the participants had to be excluded from the analysis due to individual missing values. Furthermore, due to missing additional information the data from the unpublished study was incorporated in the PSC-based predictions, but could not be used in the analysis incorporating previous performance and the additional information about study and participant specific characteristics. The available data was analysed with regard to outliers using isolation forests. Identified outliers were excluded from the predictions (see section Isolation Forests).

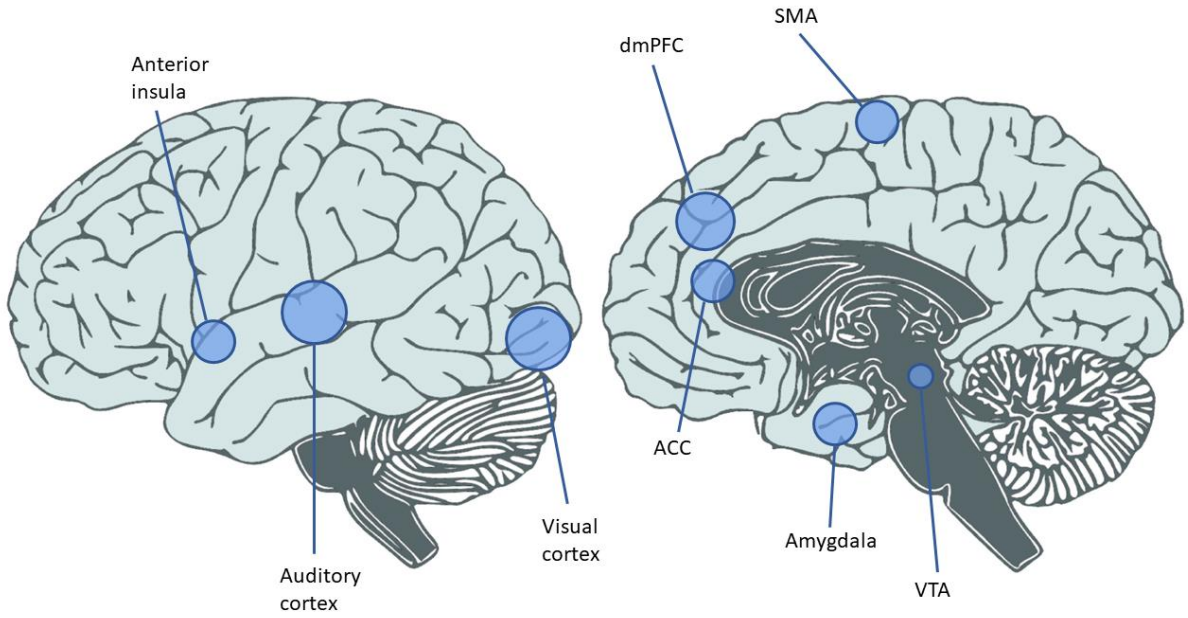



**Figure 3.** Illustration of the regions of interest from the included studies.

Authors contributing the data were asked to provide one value indicating PSC per neurofeedback run determining neurofeedback success. Next to the PSC values of the individual runs, participant and study specific information were included. These additional variables were added as features in the analysis:

- Age
- Sex (Male/Female
- Patient/Healthy

- Runs on one or several days
- Feedback type (continuous or intermittent)
- One or multiple clusters
- Functional localizer (yes/no)
- Precise instructions/No precise instructions
- Length on training run
- Length of regulation blocks

In the prediction two machine learning models were always trained on the same data, namely a Lasso and a randomized trees model. For each run, the models were trained first on the previous regulation performance and second on the previous regulation performance combined with the additional information.

**Machine Learning models & cross validation**

A seminal definition of machine learning was given by Mitchell in 1997: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with E" (Mitchell, 1997). Breaking down this very technical definition: The experience E corresponds, in our case, to data based on which our constructed program learns to solve a task T. Here, the data is the previous run performance and the additional features listed in the previous section, while the task is a regression task, namely the prediction of the participants regulation performance. A performance measure is a technical term from the field of Artificial Intelligence, representing the evaluation of the program or agents performance (Russell & Norvig, 2016). In our case the performance measure is the increase of explained variance as represented by the coefficient of determination $R^2$.

In the following sections k-fold cross validation will be explained followed by two regression methods in Lasso regularised Linear Regression and Randomized Trees based on "An Introduction to Statistical Learning" (James et al., 2013). K-fold cross validation is a resampling procedure employed to evaluate machine learning models on limited data samples. It ensures that the model does not overfit the data used for training the model but allows for a realistic estimation of the models' performance to new, unseen cases.

Furthermore, a machine learning technique for identifying outliers – Isolation Forests – will be described. The identified outliers were excluded from the analysis to form a representative sample. Finally, the concept of permutation-based feature importance will be explained, to introduce the technique used to determine each features importance for the made predictions.

All reported analyses were conducted using Python 3.7 (Python Software Foundation, https://www.python.org/) and the library Scikit-learn (Pedregosa et al., 2011).

***k - fold cross - validation***

A common problem in machine learning is that models always have to be trained on a sample, due to resource limitations and the lack of possibility to have an example for every alternative of a given problem. Although the ideal sample would be representative for the population, the models still always face the challenge of generalising to new, so far unseen data. If the models are exclusively trained on the collected sample not paying any attention to this problem, our models suffer from overfitting. Overfitting describes the case of our model fitting the individual characteristics of our sample well, including unusual cases but not generalizing to new data, resulting in poor performance. James et al. (2013) address different techniques how to tackle this problem. We choose a k-fold cross validation approach. Here the available set of observations are randomly divided into k folds of approximately equal size. The first fold is used as a validation or test set while the remaining k-1 folds are used to train the model. This procedure is repeated k times, where each time a different fold is used as a test set. Each time the mean squared error is computed and, eventually, the mean over all computed mean squared errors is used as the overall model performance. Because of the dependence on the assignment of individual observations to the k-folds, there is still some variability in the result. Therefore, the just described method of splitting up the dataset and iterating over the folds is often repeated several times to account for the variability.

For finding the optimal hyperparameters and obtaining a good estimate of our models' performance on new data, we use two nested cross validation loops. The outer loop is a 10-times 10-fold cross validation loop. The 9 folds used as a training set from the outer loop are used for the 10-times 5-fold inner loop, where the data is split over again. In this inner loop the hyperparameters are optimized using randomized search from a specified

distribution. The exact optimized hyperparameters are described in the following sections. After having identified the best value for the hyperparameters in the inner cross validation loop, the entire dataset of the inner loop is used to fit the model using the specified hyperparameter value. In the next step the left-out testing fold of the outer cross validation is used to identify the model's performance. This completes one iteration of the outer cross validation loop. This process of searching for the best hyperparameters is then repeated a total of 100 times until the outer cross validation loop is completed and the model fully trained.

***Lasso regularised linear regression***

For understanding Lasso regularised linear regression the explanation starts with linear regression, followed by multiple linear regression to eventually understand the regularized version following the explanation from James et al. (2013). Simple linear regression is straight forward: A quantitative response Y is predicted on the basis of a single predictor variable X, while assuming a linear relationship. This linear relation can be written as

$$Y \approx \beta_0 + \beta_1 X + \varepsilon. \quad (1)$$

$\beta_0$ and $\beta_1$ are unknown constants, representing the intercept and the slop in the linear model. After using our training data to obtain the estimates for these coefficients we can predict values for Y based on X. The values for $\beta_0$ and $\beta_1$ are unknown in practice. For obtaining estimations we estimate them based on our data. Let

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \quad (2)$$

be a set of n estimation pairs. We want to choose the values for $\hat{\beta}_0$ and $\hat{\beta}_1$in a way that the differences between our fitted line and the training data is minimal. The most commonly used approach is to minimize the least squares criterion, which is achieved by minimizing the sum of squared errors as illustrated in Figure 4.

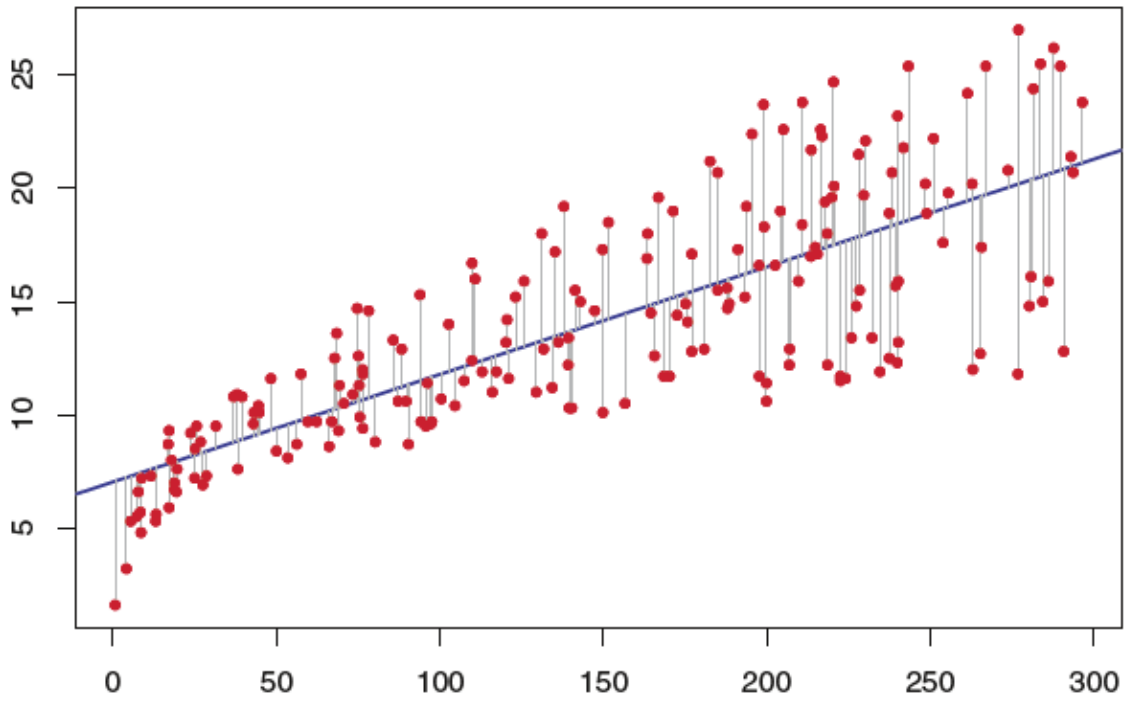



**Figure 4**. Illustration of the residuals between the data point and the fitted linear regression line. Taken from (James et al., 2013).

The difference between the actual values corresponding to the red dots and the predicted ones indicated by the line represents the residuals which can be calculated via

$$e_i = y_i - \hat{y}_i. \tag{3}$$

For each data point there is one residual which can be thought of as the error of the prediction for that given data point. The residual sum of squares (RSS) is calculated by

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2, \tag{4}$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \tag{5}$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. This can be achieved calculating

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \quad (6)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \quad (7)$$

where $\bar{y}$ and $\bar{x}$ are the sample means.

Multiple linear regression follows the same principles but instead of predicting Y only based on X we add further features to our prediction. Including a total of p predictors our multiple linear regression model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (8)$$

We can now in the same way as for linear regression estimate our parameters with the difference that we now have to estimate p coefficients based on the number of features we use for our prediction. For estimating the parameters, we again apply the same least squares approach, minimizing the sum of squared residuals:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_1 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \quad (10)$$

Having explained linear regression and its expansion to multiple linear regression, the only remaining step is adding the regularization term. Here, the Lasso regularized regression will be explained as one of the most common techniques next to Ridge Regression. Lasso regression estimates the coefficients in a very similar way to the least squares approach. The only difference is the addition of a regularization term to the RSS. In particular the β-coefficients are chosen to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \quad (11)$$

$\lambda$ is a tuning- or also called hyperparameter which is to be determined separately, often in a cross-validation fashion. The RSS searches for the fit which minimizes the residuals as

before. The second term, the so-called lasso penalty is shrinking the coefficients towards zero trying to make them small, and thereby also naturally performing a feature selection, since the predictors with a coefficient of 0 do not contribute to the prediction. The tuning parameter $\lambda$ serves the overall scaling of the regularization term where a $\lambda$ = 0 is equal to a classic multiple linear regression. $\lambda$ in our case is optimized inside the inner 5-times 10-fold cross validation loop. Using randomized search, we select 250 values in the value range 0.001 to 1000 on each iteration, aiming at finding the optimal value for the hyperparameter $\lambda$.

***Randomized trees***

This section will briefly explain the ideas behind decision trees, as they are used for regression tasks, followed by an explanation of random forests to eventually explain the modification in randomized trees. Predictions from regression trees for a specific observation are typically made after a series of splits and eventually using the mean or mode of the training observations in the region which it belongs to. The splitting rules used to segment the predictor space can be summarized in a tree, therefore these types of approaches are generally referred to as decision trees. Figure 5 modified from James et al. (2013) illustrates such a decision tree. In line with the tree analogy, the regions R1 to R5 are referred to as leaves, while the split points leading to the leaves are internal nodes. The tree originates at the so-called root creating branches to the leaves.

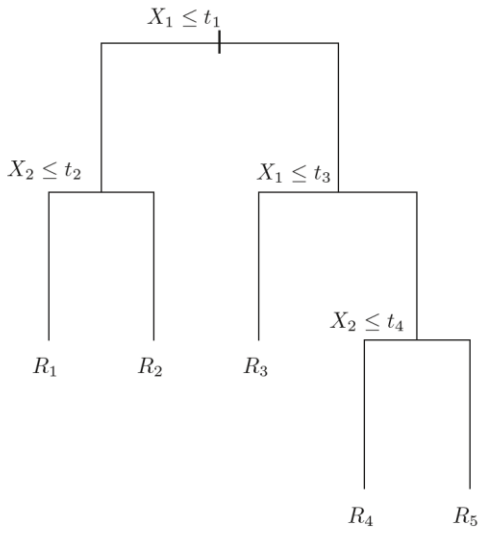



**Figure 5.** Modified from James et al. (2013). Visual illustration of a decision tree resulting in 5 distinct regions.

The building process of a tree follows two steps:

1. The predictor space (X), that is the set of possible values for our p predictors, is divided into J distinct non-overlapping regions $R_1, \dots, R_J$.
2. Then the same prediction is made for every observation, falling into region $R_J$, which is simply the mean of the response values for the training observations in $R_J$.

In other words, using the decision tree the prediction is made by stepping through the decision tree and following down a branch according to the splitting points to one of the J regions. The prediction for that given observation is the value of that leave which is determined by the examples in the training phase of the decision tree. The open question is how are the regions constructed and the splitting points determined? The goal is to find the regions $R_1, \dots, R_J$ that minimize the RSS given by

$$\sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \hat{y}_{R_j})^2, \tag{12}$$

where $\hat{y}_{R_j}$ is the mean over the training examples in the J-th region. Since it is impossible to consider every possible partitioning of the feature space, we have to take a top-down, greedy approach, known as recursive splitting. Here we step through the tree starting at the root and at each splitting node choose the given best split minimizing the local RSS at that given node. Thereby at each split two new regions are created which are then again split over again into two new regions each in the subsequent step. This process is repeated until no more splits are possible because each node only contains one example, or a specific stopping criterion is reached, for example a minimum of 5 observations in each node. One problem that decision trees suffer from is high variance, meaning that the decision trees overfit to the data and work very well on the trained dataset, but do not generalize well to new examples.

One way to solve that problem is the random forest algorithm which builds upon individual decision trees (Breiman, 2001). It takes advantage of a method called bagging. Since every individual tree tends to suffer from high variance, a solution is to build several decision trees and average their output. For doing so repeatedly a bootstrap sample is drawn from the data set with the same size as the initial dataset. From each bootstrap

sample a decision tree is built. In a next step the output of the built trees is averaged to obtain the final prediction. Furthermore, random forest implements another method that decorrelates the individual trees which eventually after the averaging improves the performance of the model: at each splitting node only a random subset of predictors is used to compute the optimal local split. This results in more diverse trees, which in combination with the averaging over many decision trees improves the performance of the random forest.

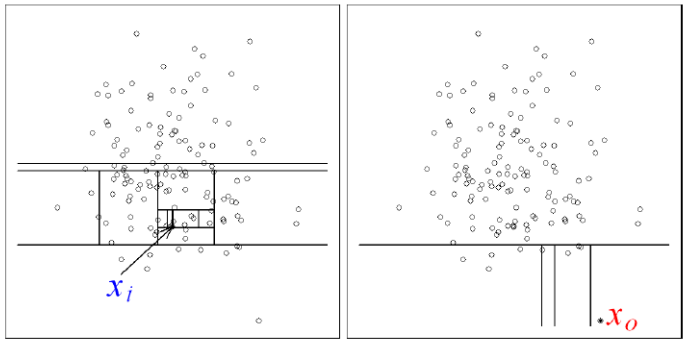
Building on the idea of a random forest a modified version was developed, called the randomized trees (Geurts et al., 2006). There are two main differences between a random forest and a randomized tree. However, the basic principle of building many different decision trees that are averaged at the end remains the same. The first difference is that rather than drawing bootstrap samples from the dataset for each built tree, all available data is used. The second difference are the split points. In random forests a subset of all available predictors is chosen and then the best value minimizing the RSS chosen as a splitting criterion, whereas in the case of a randomized tree all available predictors are used for each split. For each predictor, a random value is chosen and evaluated. The predictor with the chosen random value minimizing the RSS the best is chosen as the splitting criterion. Hence, the trees built using the randomized tree method are very diverse and decorrelated which is desirable since the subsequent averaging over very diverse trees tends to result in the best overall model.

One hyperparameter that was optimized was the maximum number of leaves each tree is allowed to grow in order to control for tree complexity. Here, similar to the lasso regression, we search for the ideal parameters in the inner cross validation loop trying 85 different values using randomized search with an upper threshold of a maximum of 256 leaves on each built tree.

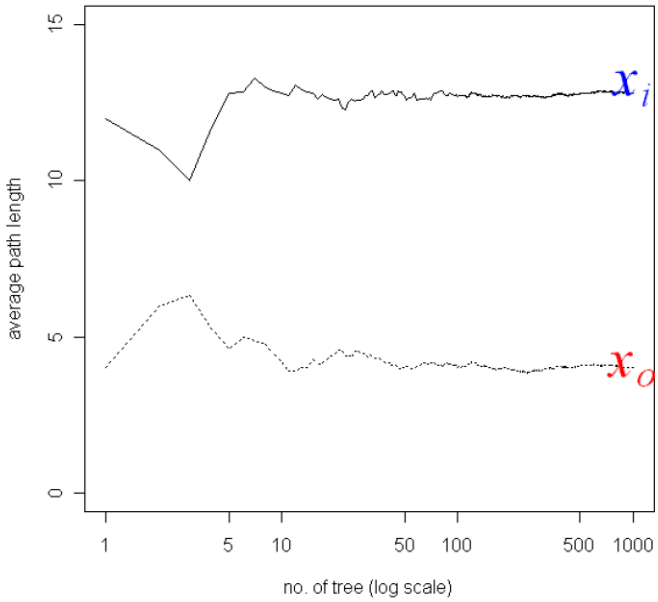
***Isolation forest***

Isolation forests are a tree-based method just like randomized trees. However, they serve a different purpose. Rather than solving regression tasks they can be applied for detecting anomalies (Liu et al., 2008). Isolation forests exploit two properties of outliers to identify them. The first is that they are minorities, consisting of fewer instances. The second is that they have attribute values that are different to those of the normal instances. These

properties make them very susceptible to isolation. Thus, the anomalies are isolated closer to the root of a decision tree, whereas normal datapoints are isolated deeper in the tree. Similar to random forests and randomized trees, also in the case of isolation forests many trees are built resembling a forest of random trees. When these trees collectively produce shorter path lengths for some particular datapoints they are highly likely to be anomalies. For a more thorough explanation see Liu et al. (2008).




(a) Isolating $x_i$ (b) Isolating $x_o$




(c) Average path lengths converge

**Figure 6.** taken from Liu et al. (2008). (a) and (b) illustrate the different path lengths in order to isolate a given datapoint. (c) shows the convergence of path lengths with increasing number of trees.

Figure 6 (a) shows the isolation of a datapoint which is no outlier in comparison to the isolation of an anomaly in shown in Figure 6 (b). Figure 6 (c) shows the averaged path lengths of the forest. With increasing tree number, the path length converges to a stable

value. In the here presented example the anomaly requires 4 splits to be isolated while the normal datapoint is isolated after 12 splits (Liu et al., 2008).

In the present study for each predicted run the PSC data was analysed using isolation forests, excluding the outliers before fitting the models. On average about 10% of the data were identified as outliers and excluded before the training of the models.

***Bootstrap test***

In order to assess the statistical significance of the difference between trained models and chance level, bootstrap tests were conducted. For doing so, models were trained on randomly shuffled data first, to generate 100 $R^2$ values of the model predicting at chance. This distribution is now compared to the distribution of 100 $R^2$ values generated by the trained models on the correct data.

For comparing the distributions, the bootstrap test implements the following steps:

1. The difference in mean based on the to sample distributions was computed ("Mean Difference 1").
2. In the subsequent step the two distributions are merged to one joint distribution. Under the H0 the two distributions are the same.
3. Now two bootstrap samples are drawn with replacement from the joint distribution of equal size to the initial two distributions. In our case 100 values.
4. Now the mean difference between the two drawn distributions is calculated, here named "Mean Difference bootstrapped".
5. The steps 3 and 4 are repeated very often generating many different Mean Difference bootstrapped values. We generated a total of 20000 Mean Difference bootstrapped values per bootstrap test.
6. Based on the ratio of how often Mean Difference 1 is larger than Mean Difference 2 a p-value can be calculated and compared to a significance level.

For each prediction, a bootstrap test was calculated using a significance level of $\alpha = 0.05$.

## Feature importance

### *Permutation based feature importance*

When inputting several features into a machine learning model an interesting question is which features are actually important for the prediction and which do not yield valuable information for the predictions. One approach to analyse and interpret the contributions of the individual variables used as predictors was introduced by Altmann et al. (2010). They propose a permutation-based feature importance measure. The permutation-based feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled. By shuffling, the relationship between the feature and the target is broken up. Hence, the drop--in model score, here the $R^2$ scores, is indicative of how much the model depends on the specific feature. Therefore, in an iterative fashion one feature is shuffled at a time, then the model is trained over again, and the corresponding drop observed, allowing for an assessment of the model's overall dependence on each feature. One big advantage of the permutation-based feature importance approach is that it is model agnostic, meaning that it is not specific to one type of model (like a Lasso regularised linear regression or randomized trees), but can be compared across models. However, it needs a substantial amount of data to clearly identify the most important features and their importance precisely. Another characteristic which is important to note is that correlated features represent a problem in the permutation-based feature importance approach, since when one feature is shuffled, the information is still available through the correlated feature. In the results section for each reported model the most important feature is reported as well as its proportional reduction of $R^2$.

### *Model based importance*

In order to gain a better understanding of the individual contributions from each feature next to approaches like permutation-based feature importance, the Lasso regularised linear regression and randomized trees models can be inspected themselves.

For the lasso regularised linear regression the model's weights were additionally analysed for each included feature. For each model the four most important predictors are reported and their weights in form of the median over the 100 models trained in the cross-validation. The individual weights indicate the importance of the corresponding feature and can be directly interpreted since the features were standardised before inputting them into

the model. The weight for each feature indicates the increase along the y-axis for one unit increase along the x-axis.

In case of the analysis of the randomized trees models, instead of the model coefficients for each feature, the relative reduction of the root mean squared error compared to the total reduction in root mean squared error was analysed and ranked. The most important predictor reduces the root mean squared error the most. In the results section the four most important predictors are listed, and their relative reduction of the root mean squared error as the median value over the 100 models trained in the cross-validation.

It has to be noted that one disadvantage of the model inspections is that they are not model agnostic like the permutation-based feature importance. Hence, it is not possible to compare the importance rankings directly across different model types.

# Results

**Prediction results based on previous run performance**

In the following, the model performance summaries will be reported. Table 2 and Table 3 contain the model summaries for the Lasso and the randomized trees models based on the PSC values only. For each run the number of available participants is listed as well as the average and median $R^2$ values over the 100 models trained in the cross-validation. Furthermore, the $R^2$ values for the models trained on the randomly shuffled data are reported to obtain the chance level predictions. Additionally, *p*-values were calculated comparing the model fits to the chance level predictions and to the trivial predictors using bootstrap tests.

**Table 2.** Summary of the Lasso models trained on the Percent Signal Change data.

| Predicted Run, (n) | Avg. $R^2$ (std), Median $R^2$ | p-value to trivial predictor | Chance Level Avg. $R^2$ (std) | p-value to chance level |
|---|---|---|---|---|
| Run 3, (170) | 0.12 (0.23), 0.14 | < .001* | -0.1 (0.12) | <.001* |
| Run 4, (110) | 0.12 (0.49), 0.26 | .003* | -0.13(0.2) | <.001* |
| Run 5, (66) | -0.02 (0.53), 0.11 | .634 | -0.36 (0.57) | <.001* |
| Run 6, (66) | -0.06 (0.7), 0.16 | .806 | -0.36(0.68) | .003* |

Avg. $R^2$ and Median are calculated over the 100 models fitted in the outer cross validation. The comparison to the trivial predictor tests whether the prediction is significantly better than always predicting the mean value. The Chance Level comparison is based on fitting the model to the randomly shuffled data. (n) = number of included participants. (std) = standard deviation. *$p < .05$.

The Lasso regularised regression and the randomized trees show similar performances. Both models always show significantly better predictions than chance levels

for all runs ($p < 0.05$). Furthermore, both models showed significantly better performances than the trivial predictor, corresponding to predicting using the mean value, for the 3rd and 4th run with a median $R^2$ up to 0.26. However, neither the lasso regularised regression nor the randomized trees models were able to predict the $5^{th}$ and $6^{th}$ run significantly better than the trivial predictor. For more detailed information on each model see the Appendix for plots showing the models predictions plotted against the true values in the data (Figure7-14).

**Table 3.** Summary of the Randomized Trees models trained on the Percent Signal Change data.

| Predicted Run, (n) | Avg. $R^2$ (std), Median $R^2$ | p-value to trivial predictor | Chance Level Avg. $R^2$ (std) | p-value to chance level |
|---|---|---|---|---|
| Run 3 (170) | 0.12 (0.29), 0.16 | <0.001* | -0.1 (0.17) | <.001* |
| Run 4, (110) | 0.1 (0.43), 0.22 | .008* | -0.17 (0.19) | <.001* |
| Run 5, (66) | -0.14 (0.86), 0.06 | .936 | -0.33 (0.46) | .04* |
| Run 6, (66) | 0.02 (0.55), 0.16 | .338 | -0.57 (1.01) | <.001* |

Avg. $R^2$ and Median are calculated over the 100 models fitted in the outer cross validation. The comparison to the trivial predictor tests whether the prediction is significantly better than always predicting the mean value. The Chance Level comparison is based on fitting the model to the randomly shuffled data. (n) = number of included participants. (std) = standard deviation. *$p < .05$.

**Prediction results based on all available information**

The table 4 and table 5 summarise the lasso regularised regression and randomized trees models based on all available information which include study and participant specific information next to the PSC values. Similarly, to table 2 and table 3 the mean and median $R^2$ values are reported for the trained models and the chance level predictions. Additionally,

the *p*-values for bootstrap tests are reported comparing the trained models to the chance level prediction and the trivial predictor.

**Table 4.** Summary of the Lasso models on all available information.

| Predicted Run, (n) | Avg. $R^2$ (std), Median $R^2$ | p-value to trivial predictor | Chance Level Avg. $R^2$ (std) | p-value to chance level |
|---|---|---|---|---|
| Run 3, (160) | 0.03 (0.23), 0.07 | .128 | -0.06 (0.12) | .0012* |
| Run 4, (99) | 0.07 (0.56), 0.18 | 0.099 | -0.22 (0.36) | <.001* |
| Run 5, (55) | -0.21 (0.69), -0.02 | .997 | -0.62 (1.3) | .008* |
| Run 6, (55) | -0.54 (1.47), -0.07 | .999 | -0.54 (0.94) | .971 |

Avg. $R^2$ and Median are calculated over the 100 models fitted in the outer cross validation. The comparison to the trivial predictor tests whether the prediction is significantly better than always predicting the mean value. The Chance Level comparison is based on fitting the model to the randomly shuffled data. (n) = number of included participants. (std) = standard deviation. *$p < .05$.

Both models can predict significantly better than chance level for all runs except the sixth run. However, for the comparison to the trivial predictor only one prediction was significantly better, namely the 3rd run predicted by the randomized trees model. All other runs for both models remained not significant. A detailed description on the models' individual performances can be found in the Appendix figures 15-22.

**Table 5.** Summary of the Randomized Trees models trained on all available information.

| Predicted Run, (n) | Avg. $R^2$ (std), Median $R^2$ | p-value to trivial predictor | Chance Level Avg. $R^2$ (std) | p-value to chance level |
|---|---|---|---|---|
| Run 3, (160) | 0.04 (0.24), 0.01 | 0.029* | -0.21 (0.25) | <.001* |
| Run 4, (99) | -0.11 (0.65), 0.04 | 0.944 | -0.38 (0.55) | .0024* |
| Run 5, (55) | -0.42 (1.0), -0.06 | 1.0 | -0.9 (1.65) | .017* |
| Run 6, (55) | -0.31 (1.23), -0.1 | .985 | -0.74 (3.71) | .251 |

Avg. $R^2$ and Median are calculated over the 100 models fitted in the outer cross validation. The comparison to the trivial predictor tests whether the prediction is significantly better than always predicting the mean value. The Chance Level comparison is based on fitting the model to the randomly shuffled data. (n) = number of included participants. (std) = standard deviation. *$p < .05$.

**Feature importance analysis**

***Permutation based feature importance analysis***

For analysing the importance of each feature permutation-based feature importance was used on the models including all available information. Since the overall sample sizes of our models are low, the permutation-based feature importance results show large variability over the 100 models trained in the cross-validation. The full rankings for each model can be found in the Appendix. Here, in Table 6 and Table 7, for each run the most important predictor is listed and the median importance for each run's prediction. Since a clear most important predictor could not always be determined, some fields are left blank corresponding to no predictor having an importance above the arbitrary threshold of 0.05.

**Table 6.** Permutation based feature importance for Lasso regularised linear regression

| Predicted Run | Ranked 1st, Median |
|---|---|
| Run 3 | Run 2, 0.38 |
| Run 4 | Run 3, 0.22 |
| Run 5 | Run 4, 0.05 |
| Run 6 | Run 1, 0.10 |

The most important predictor for each lasso regularised linear regression model according to permutation-based feature importance and the corresponding median importance.

**Table 7.** Permutation based feature importance for randomized trees

| Predicted Run | Ranked 1st, Median |
|---|---|
| Run 3 | Run 2, 0.2 |
| Run 4 | - |
| Run 5 | - |
| Run 6 | Run 5, 0.057 |

The most important predictor for each randomized trees model according to permutation-based feature importance and the corresponding median importance. Empty cells indicate that all predictor values were below 0.05.

As listed in Table 6 and Table 7, the past runs always were the most important predictors and in 5 cases the most recent run was the most crucial predictor. However, permutation-based feature importance suffers from the small sample, which makes clearly identifying the importance for each predictor difficult. Hence, it was not possible to identify one most important predictor for the randomized trees models predicting run 4 and run 5.

For a better understanding of the individual contributions additionally a model-based analysis was conducted and is reported in the following chapter.

***Model importance Lasso regularised linear regression***

In the case of lasso regularised linear regression the coefficients can be directly interpreted as the increase along the y-axis with one unit of increase along the standardized x-axis. Table 8 lists the coefficients for the 4 most important predictors for each model trained.

**Table 8.** Model importance analysis for the Lasso regularised linear regression trained on all available information.

| Predicted Run | Ranked 1, Median | Ranked 2, Median | Ranked 3, Median | Ranked 4, Median |
|---|---|---|---|---|
| Run 3 | Run 2, 0.27 | Run 1, 0.09 | Length of neurofeedback run, 0.07 | Patient or Healthy, 0.06 |
| Run 4 | Run 3, 0.33 | Run 2, 0.14 | Patient or Healthy, 0,06 | Length of neurofeedback run, 0.04 |
| Run 5 | Run 4, 0.27 | Up regulation, 0.19 | Run 3, 0.14 | Run 2, 0.04 |
| Run 6 | Run 2, 0.31 | Run 5, 0.18 | Run 4, 0.08 | Run 3, 0.07 |

The 4 most important predictors according to their model weights are ranked and their corresponding median values over the 100 models trained in the cross validation.

Overall, the results are similar to the permutation-based feature importance results, in the preceding runs being important predictors where in 3 out of 4 cases the most recent run was the highest ranked predictor. Additionally, the length of the neurofeedback run as well as the participant coming from a healthy or a clinical population were ranked as important predictors. All rankings for each model can be found in the Appendix.

### *Model importance randomized trees*

Just like the coefficients in lasso regularised linear regression, in randomized trees the features can be analysed by observing the normalized root mean squared error reductions each feature contributes. Table 9 summarises the most important predictors for each run

and their relative root mean squared error reduction as the median over the 100 models trained in the cross-validation.

**Table 9.** Model importance analysis for the Randomized Trees model trained on all available information.

| Predicted Run | Ranked 1, Median | Ranked 2, Median | Ranked 3, Median | Ranked 4, Median |
|---|---|---|---|---|
| Run 3 | Run 2, 0.44 | Patient or Healthy, 0.16 | Run 1, 0.12 | One or more days, 0.09 |
| Run 4 | Run 3, 0.31 | Patient or Healthy, 0.21 | Run 2, 0.21 | Run 1, 0.04 |
| Run 5 | Run 4, 0.23 | Part of sensory system, 0.19 | Run 3, 0.15 | Run 2, 0.13 |
| Run 6 | Run 2, 0.31 | Run 5, 0.3 | Run 4, 0.14 | Run 3, 0.08 |

The 4 most important predictors based on the relative root mean squared error reduction are ranked and their corresponding median values over the 100 models trained in the cross validation.

In line with the previously reported results, the past runs are the most important predictors and in 3 out of 4 cases the most recent run reduced the relative root mean squared error the most. Additionally, it can be noted that for the prediction of run 3 and run 4 again the information about a participant coming from a clinical or heathy population was important information and allowed for a considerable amount of root mean squared error reduction. The entire ranking for each model can be found in the Appendix (Figures 25-38).

# Discussion

RtfMRI-based neurofeedback is an emerging scientific tool allowing to learn to self-regulate neural activity. It has been applied to various fields showing promising results as described in the chapter on target groups. However, the question arises whether we can observe corresponding patterns in neural activity that go along these behavioural changes. Therefore, this Thesis investigated the predictability of neurofeedback training performance as measured by PSC.

**Predictability of regulation performance**

Overall, our results show that neurofeedback performance can be predicted above chance level as shown by both model types trained on PSC values. All models trained on PSC values predicted significantly better than chance level. However, the incorporation of additional information did not yield improvements in the amount of explained variance. Overall, the models trained on all available information show a similar pattern with all runs showing above chance level prediction performance except the 6$^{th}$. In any case, only the 3$^{rd}$ and 4$^{th}$ run could be predicted significantly better than the trivial predictor with the highest median $R^2$ value being 0.26. This points out that although our models are able to predict above chance level and in some cases better than the mean, a decent amount of PSC variance remains unexplained. Regarding the trained models, the effect of the decreasing sample size with the increasing number of runs, used as target, has to be pointed out. For predicting the 3$^{rd}$ run the data of up to 170 participants were available while for the 6$^{th}$ run only 66, corresponding to less than 40% of the data compared to run 3. This must be considered when interpreting our results, since less data leads to overall worse model performance. One reason for this is the fewer examples the model can be trained on, which will not cover the full variable value range. In order to assess the predictability of training performance for the 6$^{th}$ run and even later runs we would require additional data. Hence, one expansion for future research is the analysis of later runs on an ideally further growing data repository. Here we would expect the predictions of the 6$^{th}$ and later runs to work similarly well to the earlier runs assuming equally large datasets. Perhaps even an increased performance could be possible, since every participant would have had more training runs and therefore contributes more datapoints overall, helping the fitting of statistical models.

**Feature importance**

Next to assessing the predictability of rtfMRI-based neurofeedback training itself, the analysis of the individual importance of features was a crucial part of our data analysis. One approach we took was permutation-based feature importance. Overall, the relatively small available sample made it difficult to clearly determine each features importance with permutation importance methods since their relevance varies considerably throughout the models trained in the cross-validation. The full rankings can be found in the Appendix (Figures 25-38). However, it can be concluded, that the previous runs' PSC value seems to carry the most information in predicting the regulation performance. This can be expected since the BOLD response tends to be slow and is therefore autocorrelated (Worsley et al., 1995). Therefore, going in line with the signal being autocorrelated the directly preceding run carries the most information for the predictions.

In addition to the permutation-based feature importance, a model-based feature analysis was conducted for the trained lasso regularised linear regression and the randomized trees models. These results support the conclusion that previous runs yield the most information for predicting future performance. For all predictions, past runs were the most important predictor and in 75 percent of the cases the directly preceding run carried the most valuable information. The results from the feature importance analysis are in line with the models trained on PSC values in that they are not performing worse than the models including additional information.

Inspecting the models predicting the 3$^{rd}$ and 4$^{th}$ run with the highest $R^2$ values, it can be concluded that for these models the length of the neurofeedback training run was a valuable information, as well as the information about the participants being healthy or from a clinical population.

**Strengths, limitations, and future directions**

Our dataset is a joint dataset comprised out of various studies. The only restrictions we applied were that the feedback measure had to be a PSC value of a fixed region or set of regions, from a regulation period in comparison to a baseline period. Hence, our dataset mixes healthy participants with patients as well as different ROIs and experimental designs. Overall, all studies aim at inducing systematic changes to neural activity and train the ability to voluntarily control brain activity. Therefore, the overall goal of the studies is comparable. Nevertheless, it is hard to point out to what extend one of the many variables might affect

the comparability of the included data. It will be a very interesting challenge for future research to generate a better understanding of the relevant factors affecting training performance. A recent systematic review conducted by Thibault et al. (2017) also maps out this heterogeneity in the field from the differences in experimental designs, over feedback techniques to the wide spanning range of outcomes: one the one hand, studies in which participants fail to obtain control of their neural activity, and, on the other hand, studies showing great success with participants being able to control their brain activity and showing concurrent behavioural changes. Here, an important aspect they point out is the likely bias in the literature overrepresenting successful studies and unsuccessful studies not being published, and therefore not included in many studies as it is the problem in many fields (Ioannidis, 2005; Thibault et al., 2017).

However, as the field matures, and further data is collected we will hopefully continue to see the trend of sharing data, allowing more comprehensive meta analytic approaches like ours. Additionally, finding a way to make data from unsuccessful studies accessible would enable further insights, allowing for a more accurate depiction of rtfMRI-based neurofeedback as a field. Furthermore, endeavours like the CRED-nf checklist will help a lot in incorporating new data and assess its comparability (Ros et al., 2019). Here the main objective is to introduce a consensus-driven checklist in the field, aimed at improving reporting and experimental design standards.

Next to the already mentioned future research directions, other possible extensions of this Thesis will be worth pursuing. On the one hand, as mentioned before, an expansion of the available data with new studies would allow for more precise results. On the other hand, also the models themselves could be optimized by looking further into hyperparameter tuning or the training of different model types. Furthermore, next to training different models on the same data, other datasets should be analysed like different activity-based measures for instance the beta-coefficients or connectivity-based neurofeedback studies. This could either confirm our present results or raise new questions if one of the feedback measures is better predictable than another.

The last branch of research following up from here is an expansion on the briefly touched upon topic of learning in neurofeedback. A better understanding of the psychological and neurobiological underpinnings could inform the experimental designs and

trained targets in future neurofeedback studies. Here open questions are the ideal length of the neurofeedback runs, the number of runs to learn to gain control and the kind of feedback to name a few. Progress here could increase the efficacy of neurofeedback and allow for a better understanding of the underlying processes in neurofeedback.

## Conclusion

Our study aimed at assessing the predictability of neurofeedback performance, measured by PSC in a large dataset incorporating data from various studies, allowing for a generalisation across the field of neurofeedback. We were able to show that training performance can be predicted and shows systematic patterns. However, going in line with the large variability in regulation performance, our best models were only able to explain about a fourth of the overall variance. We hope that our might help to optimize the experimental designs in neurofeedback studies and increase the efficacy of neurofeedback training.

# References


Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Auer, T., Schweizer, R., & Frahm, J. (2015). Training Efficiency and Transfer Success in an Extended Real-Time Functional MRI Neurofeedback Training of the Somatomotor Cortex of Healthy Subjects. *Frontiers in Human Neuroscience*, *9*(October), 547. https://doi.org/10.3389/fnhum.2015.00547

Blefari, M. L., Sulzer, J., Hepp-Reymond, M.-C., Kollias, S., & Gassert, R. (2015). Improvement in precision grip force control with self-modulation of primary motor cortex during motor imagery. *Frontiers in Behavioral Neuroscience*, *9*(February), 1–11. https://doi.org/10.3389/fnbeh.2015.00018

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buyukturkoglu, K., Rana, M., Ruiz, S., Hackley, S. A., Soekadar, S. R., Birbaumer, N., & Sitaram, R. (2013). Volitional regulation of the supplementary motor area with fMRI-BCI neurofeedback in Parkinson's disease: A pilot study. *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 677–681. https://doi.org/10.1109/NER.2013.6696025

Caporale, N., & Dan, Y. (2008). Spike timing-dependent plasticity: A Hebbian learning rule. *Annual Review of Neuroscience*, *31*, 25–46. https://doi.org/10.1146/annurev.neuro.31.060407.125639

Caria, A., Veit, R., Sitaram, R., Lotze, M., Weiskopf, N., Grodd, W., & Birbaumer, N. (2007). Regulation of anterior insular cortex activity using real-time fMRI. *NeuroImage*, *35*(3), 1238–1246. https://doi.org/10.1016/j.neuroimage.2007.01.018

Cooke, S. F., & Bliss, T. V. P. (2006). Plasticity in the human central nervous system. *Brain*, *129*(7), 1659–1673. https://doi.org/10.1093/brain/awl082

Cox, R. W., Jesmanowicz, A., & Hyde, J. S. (1995). Real-Time Functional Magnetic Resonance Imaging. *Magnetic Resonance in Medicine*, *33*(2), 230–236. https://doi.org/10.1002/mrm.1910330213

deCharms, R. C., Maeda, F., Glover, G. H., Ludlow, D., Pauly, J. M., Soneji, D., Gabrieli, J. D. E., & Mackey, S. C. (2005). Control over brain activation and pain learned by using real-time functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(51), 18626. https://doi.org/10.1073/pnas.0505210102

Emmert, K., Breimhorst, M., Bauermann, T., Birklein, F., Van De Ville, D., & Haller, S. (2014). Comparison of anterior cingulate vs. Insular cortex as targets for real-time fMRI regulation during pain stimulation. *Frontiers in Behavioral Neuroscience*, *8*, 350. https://doi.org/10.3389/fnbeh.2014.00350

Emmert, K., Kopel, R., Koush, Y., Maire, R., Senn, P., Van De Ville, D., & Haller, S. (2017). Continuous vs. Intermittent neurofeedback to regulate auditory cortex activity of tinnitus patients using real-time fMRI - A pilot study. *NeuroImage : Clinical*, *14*, 97–104. https://doi.org/10.1016/j.nicl.2016.12.023

Frank, S., Lee, S., Preissl, H., Schultes, B., Birbaumer, N., & Veit, R. (2012). The obese brain athlete: Self-regulation of the anterior insula in adiposity. *PLoS ONE*, *7*(8), 3–8. https://doi.org/10.1371/journal.pone.0042570

Gallistel, C. R., & Matzel, L. D. (2013). The Neuroscience of Learning: Beyond the Hebbian Synapse. *Annual Review of Psychology*, *64*(1), 169–200. https://doi.org/10.1146/annurev-psych-113011-143807

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely Randomized Trees. *Machine Learning*, *63*, 3–42. https://doi.org/10.1007/s10994-006-6226-1

Goebel, R., Sorger, B., Kaiser, J., Birbaumer, N., & Weiskopf, N. (2004). BOLD brain pong: Self-regulation of local brain activity during synchronously scanned, interacting subjects. *34th Annual Meeting of the Society for Neuroscience*.

Gröne, M., Dyck, M., Koush, Y., Bergert, S., Mathiak, K. A., Alawi, E. M., Elliott, M., & Mathiak, K. (2015). Upregulation of the rostral anterior cingulate cortex can alter the perception of

emotions: FMRI-based neurofeedback at 3 and 7 T. *Brain Topography*, *28*(2), 197–207. https://doi.org/10.1007/s10548-014-0384-4

Guan, M., Li, L., Tong, L., Zhang, Y., Zheng, D., Yan, B., & Shi, D. (2015). Self-regulation of rACC activation in patients with postherpetic neuralgia: A preliminary study using real-time fMRI neurofeedback. *PLoS One*, *10*(4), e0123675.

Hellrung, L., Dietrich, A., Hollmann, M., Pleger, B., Kalberlah, C., Roggenhofer, E., Villringer, A., & Horstmann, A. (2018). Intermittent compared to continuous real-time fMRI neurofeedback boosts control over amygdala activation. *NeuroImage*, *166*, 198–208. https://doi.org/10.1016/j.neuroimage.2017.10.031

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos Med*, *2*(8), e124.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kim, D.-Y., Yoo, S.-S., Tegethoff, M., Meinlschmidt, G., & Lee, J.-H. (2015). The Inclusion of Functional Connectivity Information into fMRI-based Neurofeedback Improves Its efficacy in the Reduction of Cigarette Cravings. *Journal of Cognitive Neuroscience*, *27*(8), 1552–1572. https://doi.org/10.1162/jocn

Kim, S.-G., & Bandettini, P. A. (2010). Principles of Functional MRI. In S. H. Faro & F. B. Mohamed (Eds.), *BOLD fMRI: A Guide to Functional Imaging for Neuroscientists* (pp. 3–22). Springer. https://doi.org/10.1007/978-1-4419-1329-6_1

Koush, Y., Meskaldji, D.-E., Pichon, S., Rey, G., Rieger, S. W., Linden, D. E. J., Van De Ville, D., Vuilleumier, P., & Scharnowski, F. (2015). Learning Control Over Emotion Networks Through Connectivity-Based Neurofeedback. *Cerebral Cortex*, bhv311. https://doi.org/10.1093/cercor/bhv311

Koush, Y., Rosa, M. J., Robineau, F., Heinen, K., W. Rieger, S., Weiskopf, N., Vuilleumier, P., Van De Ville, D., & Scharnowski, F. (2013). Connectivity-based neurofeedback: Dynamic causal modeling for real-time fMRI. *NeuroImage*, *81*, 422–430. https://doi.org/10.1016/j.neuroimage.2013.05.010

LaConte, S. M., Peltier, S. J., & Hu, X. P. (2007). Real-time fMRI using brain-state classification. *Human Brain Mapping*, *28*(10), 1033–1044. https://doi.org/10.1002/hbm.20326

Li, X., Hartwell, K. J., Borckardt, J., Prisciandaro, J. J., Saladin, M. E., Morgan, P. S., Johnson, K. A., Lematty, T., Brady, K. T., & George, M. S. (2013). Volitional reduction of anterior cingulate cortex activity produces decreased cue craving in smoking cessation: A preliminary real-time fMRI study. *Addiction Biology*, *18*(4), 739–748. https://doi.org/10.1111/j.1369-1600.2012.00449.x

Liew, S. L., Rana, M., Cornelsen, S., Fortunato De Barros Filho, M., Birbaumer, N., Sitaram, R., Cohen, L. G., & Soekadar, S. R. (2016). Improving Motor Corticothalamic Communication after Stroke Using Real-Time fMRI Connectivity-Based Neurofeedback. *Neurorehabilitation and Neural Repair*, *30*(7), 671–675. https://doi.org/10.1177/1545968315619699

Linden, D. E. J., Habes, I., Johnston, S. J., Linden, S., Tatineni, R., Subramanian, L., Sorger, B., Healy, D., & Goebel, R. (2012). Real-Time Self-Regulation of Emotion Networks in Patients with Depression. *PLOS ONE*, *7*(6), e38115. https://doi.org/10.1371/journal.pone.0038115

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157. https://doi.org/10.1038/35084005

MacInnes, J. J., Dickerson, K. C., Chen, N., & Adcock, R. A. (2016). Cognitive Neurostimulation: Learning to Volitionally Sustain Ventral Tegmental Area Activation. *Neuron*, *89*(6), 1331–1342. https://doi.org/10.1016/j.neuron.2016.02.002

Marins, T. F., Rodrigues, E. C., Engel, A., Hoefle, S., Basílio, R., Lent, R., Moll, J., & Tovar-Moll, F. (2015). Enhancing Motor Network Activity Using Real-Time Functional MRI Neurofeedback of Left Premotor Cortex. *Frontiers in Behavioral Neuroscience*, *9*(December), 1–12. https://doi.org/10.3389/fnbeh.2015.00341

Mark, C. I., Mazerolle, E. L., & Chen, J. J. (2015). Metabolic and vascular origins of the BOLD effect: Implications for imaging pathology and resting-state brain function. *Journal of Magnetic Resonance Imaging: JMRI*, *42*(2), 231–246. https://doi.org/10.1002/jmri.24786

Marxen, M., Jacob, M. J., Müller, D. K., Posse, S., Ackley, E., Hellrung, L., Riedel, P., Bender, S., Epple, R., & Smolka, M. N. (2016). Amygdala Regulation Following fMRI-Neurofeedback without Instructed Strategies. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00183

McDonald, A. R., Muraskin, J., Dam, N. T. V., Froehlich, C., Puccio, B., Pellman, J., Bauer, C. C. C., Akeyson, A., Breland, M. M., Calhoun, V. D., Carter, S., Chang, T. P., Gessner, C., Gianonne, A., Giavasis, S., Glass, J., Homann, S., King, M., Kramer, M., … Craddock, R. C. (2017). The real-time fMRI neurofeedback based stratification of Default Network Regulation Neuroimaging data repository. *NeuroImage*, *146*, 157–170. https://doi.org/10.1016/j.neuroimage.2016.10.048

McKee, M. (2008). Biofeedback: An overview in the context of heart-brain medicine. *Cleveland Clinic Journal of Medicine*, *75 Suppl 2*, S31-4. https://doi.org/10.3949/ccjm.75.Suppl_2.S31

Megumi, F., Yamashita, A., Kawato, M., & Imamizu, H. (2015). Functional MRI neurofeedback training on connectivity between two regions induces long-lasting changes in intrinsic functional network. *Frontiers in Human Neuroscience*, *9*(March). https://doi.org/10.3389/fnhum.2015.00160

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Nicholson, A. A., Rabellino, D., Densmore, M., Frewen, P. A., Paret, C., Kluetsch, R., Schmahl, C., Théberge, J., Neufeld, R. W. J., McKinnon, M. C., Reiss, J., Jetly, R., & Lanius, R. A. (2017). The neurobiology of emotion regulation in posttraumatic stress disorder: Amygdala downregulation via real-time fMRI neurofeedback. *Human Brain Mapping*, *38*(1), 541–560. https://doi.org/10.1002/hbm.23402

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, *87*(24), 9868–9872. https://doi.org/10.1073/pnas.87.24.9868

Ogawa, S., Lee, T. M., Nayak, A., & Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, *14*(1), 68–78. https://doi.org/10.1002/mrm.1910140108

Orndorff-Plunkett, F., Singh, F., Aragón, O. R., & Pineda, J. A. (2017). Assessing the Effectiveness of Neurofeedback Training in the Context of Clinical and Social Neuroscience. *Brain Sciences*, *7*(8). https://doi.org/10.3390/brainsci7080095

Papoutsi, M., Weiskopf, N., Langbehn, D., Reilmann, R., Rees, G., & Tabrizi, S. J. (2018). Stimulating neural plasticity with real-time fMRI neurofeedback in Huntington's disease: A proof of concept study. *Human Brain Mapping*, *39*(3), 1339–1353. https://doi.org/10.1002/hbm.23921

Paret, C., Kluetsch, R., Zaehringer, J., Ruf, M., Demirakca, T., Bohus, M., Ende, G., & Schmahl, C. (2016). Alterations of amygdala-prefrontal connectivity with real-time fMRI neurofeedback in BPD patients. *Social Cognitive and Affective Neuroscience*, *11*(6), 952–960. https://doi.org/10.1093/scan/nsw016

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage*, *22*(3), 1157–1172. https://doi.org/10.1016/j.neuroimage.2004.03.026

Ros, T., Enriquez-Geppert, S., Zotev, V., Young, K. D., Wood, G., Whitfield-Gabrieli, S., Wan, F., Vuilleumier, P., Vialatte, F., Van De Ville, D., & others. (2019). Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist). *Brain*.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Malaysia: Pearson Education Limited.

Scharnowski, F., Hutton, C., Josephs, O., Weiskopf, N., & Rees, G. (2012). Improving Visual Perception through Neurofeedback. *Journal of Neuroscience*, *32*(49), 17830–17841. https://doi.org/10.1523/JNEUROSCI.6334-11.2012

Scharnowski, F., Veit, R., Zopf, R., Studer, P., Bock, S., Diedrichsen, J., Goebel, R., Mathiak, K., Birbaumer, N., & Weiskopf, N. (2015). Manipulating motor performance and memory through real-time fMRI neurofeedback. *Biological Psychology*, *108*, 85–97. https://doi.org/10.1016/j.biopsycho.2015.03.009

Sherlin, L. H., Arns, M., Lubar, J., Heinrich, H., Kerson, C., Strehl, U., & Sterman, M. B. (2011). Neurofeedback and Basic Learning Theory: Implications for Research and Practice. *Journal of Neurotherapy*, *15*(4), 292–304. https://doi.org/10.1080/10874208.2011.623089

Sherwood, M. S., Kane, J. H., Weisend, M. P., & Parker, J. G. (2016). Enhanced control of dorsolateral prefrontal cortex neurophysiology with real-time functional magnetic resonance imaging (rt-fMRI) neurofeedback training and working memory practice. *NeuroImage*, *124*, 214–223. https://doi.org/10.1016/j.neuroimage.2015.08.074

Shibata, K., Lisi, G., Cortese, A., Watanabe, T., Sasaki, Y., & Kawato, M. (2019). Toward a comprehensive understanding of the neural mechanisms of decoded neurofeedback. *NeuroImage*, *188*, 539–556. https://doi.org/10.1016/j.neuroimage.2018.12.022

Shibata, K., Watanabe, T., Sasaki, Y., & Kawato, M. (2011). Perceptual Learning Incepted by Decoded fMRI Neurofeedback Without Stimulus Presentation. *Science*, *334*(6061), 1413–1415. https://doi.org/10.1126/science.1212003

Sitaram, R., Ros, T., Stoeckel, L., Haller, S., Scharnowski, F., Lewis-Peacock, J., Weiskopf, N., Blefari, M. L., Rana, M., Oblak, E., Birbaumer, N., & Sulzer, J. (2017). Closed-loop brain training: The science of neurofeedback. *Nature Reviews Neuroscience*, *18*(2), 86–100. https://doi.org/10.1038/nrn.2016.164

Spetter, M. S., Malekshahi, R., Birbaumer, N., Lührs, M., van der Veer, A. H., Scheffler, K., Spuckti, S., Preissl, H., Veit, R., & Hallschmid, M. (2017). Volitional regulation of brain responses to food stimuli in overweight and obese subjects: A real-time fMRI feedback study. *Appetite*, *112*, 188–195. https://doi.org/10.1016/j.appet.2017.01.032

Spilker, B., Kamiya, J., Callaway, E., & Yeager, C. L. (1969). Visual Evoked Responses in Subjects Trained to Control Alpha Rhythms. *Psychophysiology*, *5*(6), 683–695. https://doi.org/10.1111/j.1469-8986.1969.tb02871.x

Sulzer, J., Haller, S., Scharnowski, F., Weiskopf, N., Birbaumer, N., Blefari, M. L., Bruehl, A. B., Cohen, L. G., deCharms, R. C., Gassert, R., Goebel, R., Herwig, U., LaConte, S., Linden, D., Luft, A., Seifritz, E., & Sitaram, R. (2013). Real-time fMRI neurofeedback: Progress and challenges. *NeuroImage*, *76*, 386–399. https://doi.org/10.1016/j.neuroimage.2013.03.033

Thibault, R., Macpherson, A., Lifshitz, M., Roth, R., & Raz, A. (2017). Neurofeedback with fMRI: A Critical Systematic Review. *NeuroImage*, *172*. https://doi.org/10.1016/j.neuroimage.2017.12.071

Watanabe, T., Sasaki, Y., Shibata, K., & Kawato, M. (2017). Advances in fMRI Real-Time Neurofeedback. *Trends in Cognitive Sciences*, *21*(12), 997–1010. https://doi.org/10.1016/j.tics.2017.09.010

Weiskopf, N., Scharnowski, F., Veit, R., Goebel, R., Birbaumer, N., & Mathiak, K. (2004). Self-regulation of local brain activity using real-time functional magnetic resonance imaging (fMRI). *Journal of Physiology-Paris*, *98*(4–6), 357–373. https://doi.org/10.1016/j.jphysparis.2005.09.019

Worsley, K. J., Friston, K. J., & others. (1995). Analysis of fMRI time-series revisited—Again. *Neuroimage*, *2*(3), 173–181.

Yao, S., Becker, B., Geng, Y., Zhao, Z., Xu, X., Zhao, W., Ren, P., & Kendrick, K. M. (2016). Voluntary control of anterior insula and its functional connections is feedback-independent and increases pain empathy. *NeuroImage*, *130*, 230–240. https://doi.org/10.1016/j.neuroimage.2016.02.035

Young, K. D., Zotev, V., Phillips, R., Misaki, M., Yuan, H., Drevets, W. C., & Bodurka, J. (2014). Real-time FMRI neurofeedback training of amygdala activity in patients with major depressive disorder. *PloS One*, *9*(2), e88785. https://doi.org/10.1371/journal.pone.0088785

Zilverstand, A., Sorger, B., Sarkheil, P., & Goebel, R. (2015). FMRI neurofeedback facilitates anxiety regulation in females with spider phobia. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00148

# Appendix

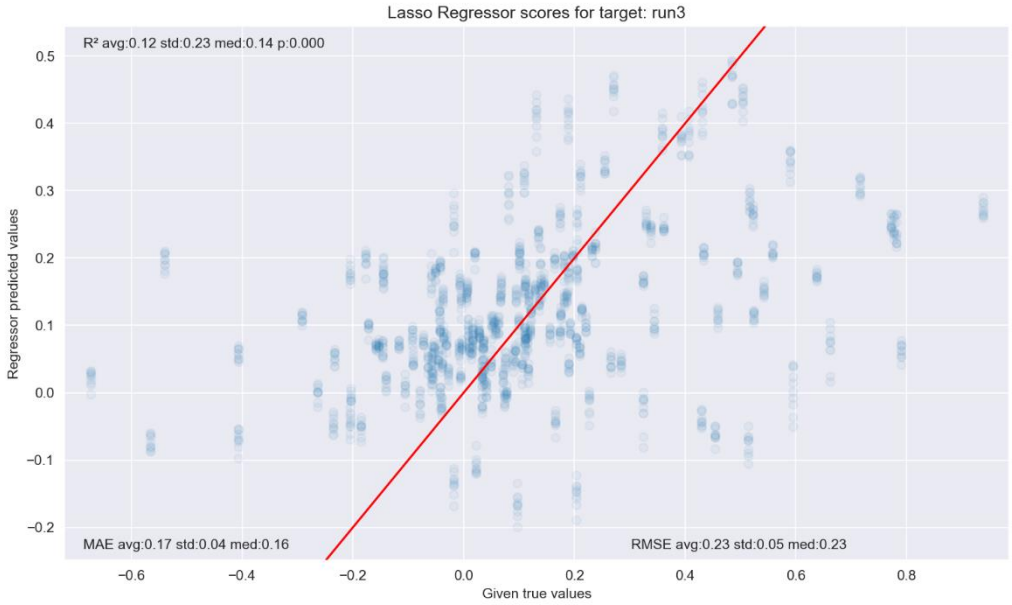
**Deutsche Zusammenfassung**

Echtzeit funktionelle magnetresonanztomographiebasiertes Neurofeedback ist ein neues wissenschaftliches und klinisches Werkzeug, welches erlaubt die eigene Gehirnaktivität zu regulieren. Es ermöglich das Verhalten von gesunden Teilnehmerinnen zu beeinflussen als auch die Behandlung von klinischen Symptomen in Patientenpopulationen. Jedoch variiert die Fähigkeit der Regulation stark über Studien und TeilnehmerInnen. Konsistente Lernkurven wie beispielsweise stetige Anstiege sind selten. Unsere Studie untersucht ob die Neurofeedbackregulationsleistung über einzelne Runs zufällig variiert oder ob vorhersagbare Muster vorhanden sind. Unsere Modelle beinhalten zudem subjekt- und studienspezifische Merkmale wie beispielsweise Alter, Geschlecht, Instruktionen, trainierte Gehirnareale oder die Länge der einzelnen Runs, um zu untersuchen wie diese Faktoren die Leistung beeinflussen. Unter Verwendung von maschinellem Lernen (L1 regularisierte lineare Regression und Randomized Trees) sagen wir die Leistung in einem bestimmten Run vorher auf Basis der vorangehenden Regulationsleistungen. Zur Analyse des Beitrags der einzelnen Prädiktoren wurde die Methode der permutationsbasierten Prädiktorenrelevanz (Permutation-based feature importance) angewendet sowie die trainierten Modelle selbst analysiert. Unsere Analysen wurden auf einem Datensatz von 197 Probanden durchgeführt (aus 11 Studien), aus welchen wir schlussfolgern, dass die Regulationsleistung stets signifikant besser als das Zufallsniveau vorhergesagt werden kann. Jedoch bleibt mit Median $R^2$ werten von bis zu 0.26 stets ein großer Teil der Varianz in unseren Daten unerklärt. Die Analyse der Relevanz der Prädiktoren ergab, dass die vorhergehenden Runs die relevantesten Prädiktoren waren und der vorangehende Run oftmals der wichtigste Prädiktor. Insgesamt zeigen unsere Ergebnisse, dass Neurofeedbackregulationsleistung überzufällig gut vorhersagbar ist und systematischen Mustern folgt. Diese Ergebnisse können dazu beitragen die Selbstregulationsfähigkeit in Neurofeedbacktraining besser zu verstehen und damit die Anwendungsmöglichkeiten von Neurofeedback vergrößern. Im Rahmen der Open Science Bewegung könnte ein solcher datenbasierter Ansatz ein vielversprechender Weg sein für ein tiefergehendes Verständnis von Neurofeedback.

## Figures

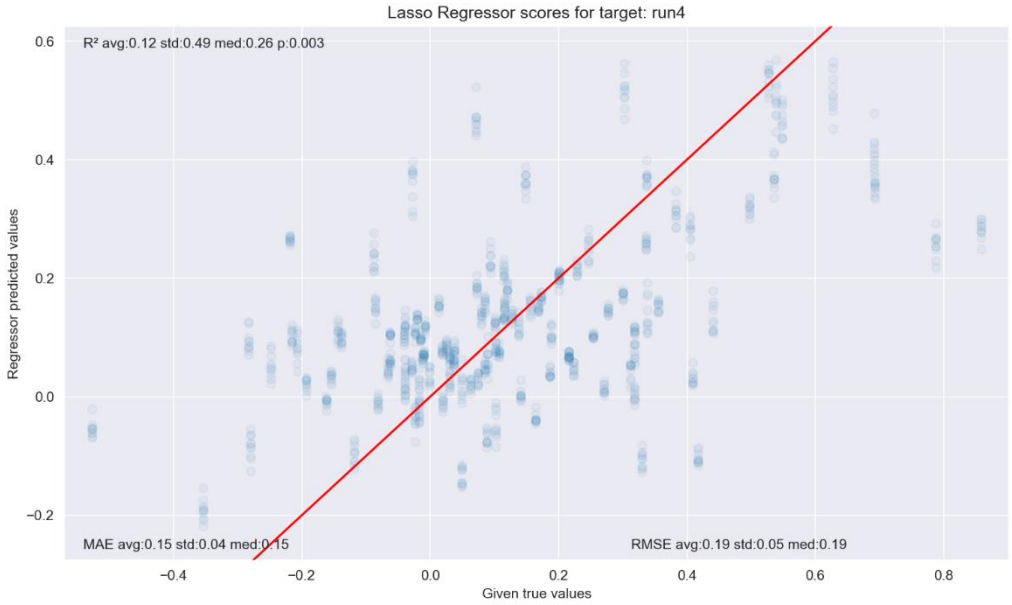
### *Models trained on PSC values only*

Figure 7 Lasso regularised linear regression predicting run 3



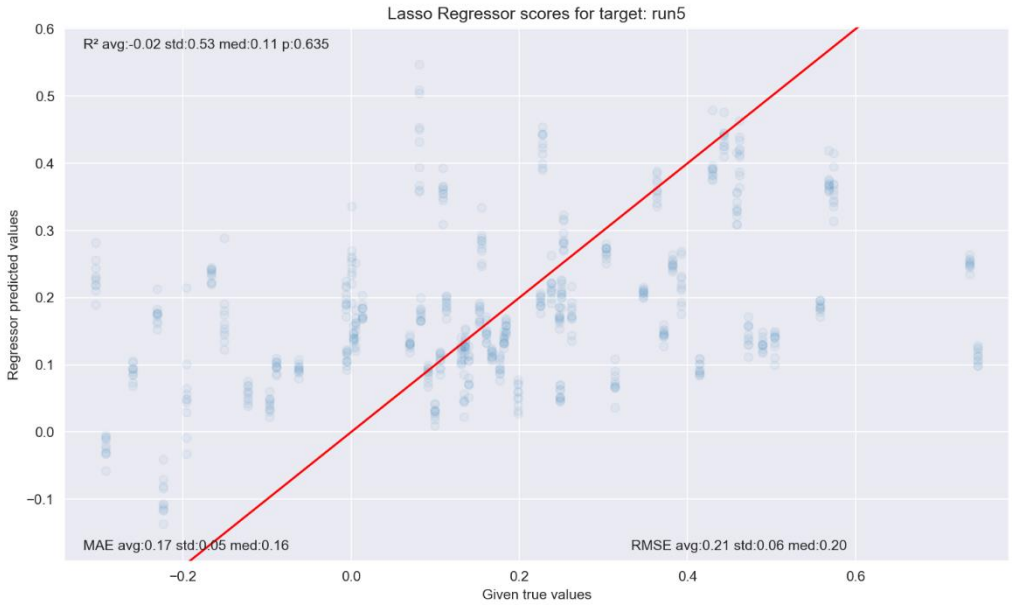

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 8 Lasso regularised linear regression predicting run 4



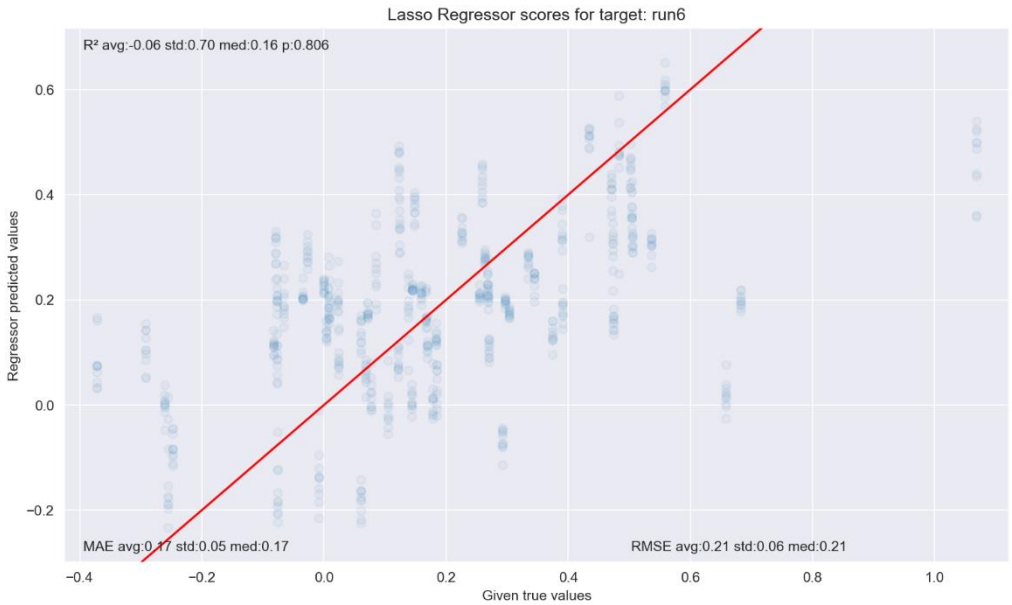

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 9 Lasso regularised linear regression predicting run 5



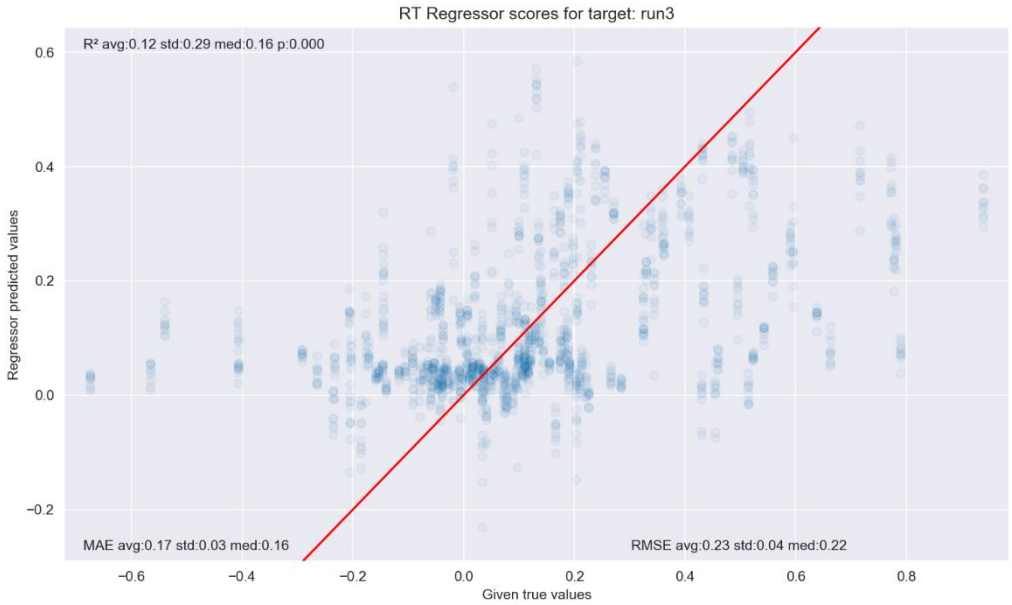

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 10 Lasso regularised linear regression predicting run 6




$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 11 Randomized Trees predicting run 3

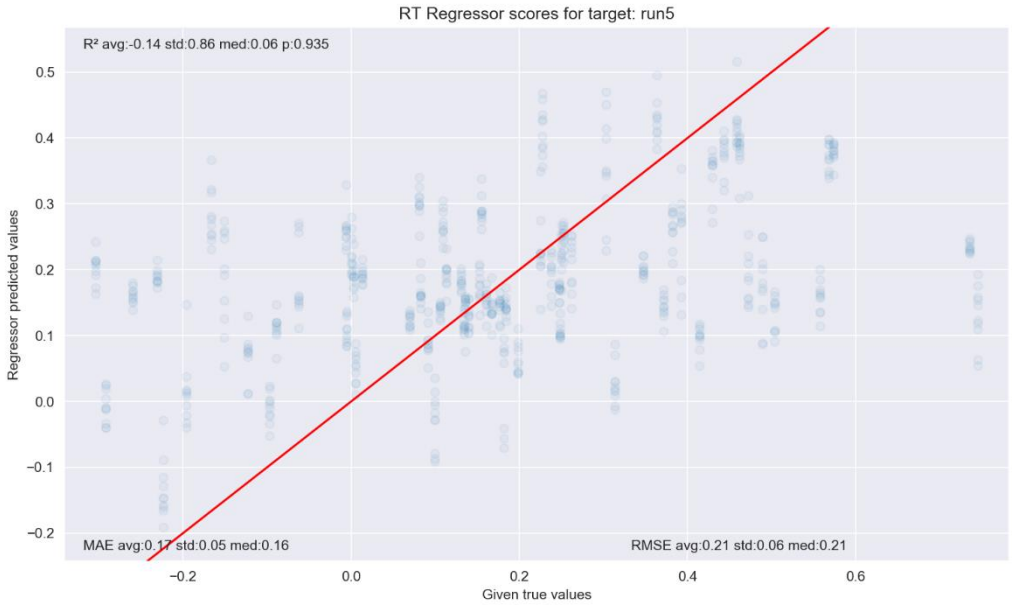



$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 12 Randomized Trees predicting run 4

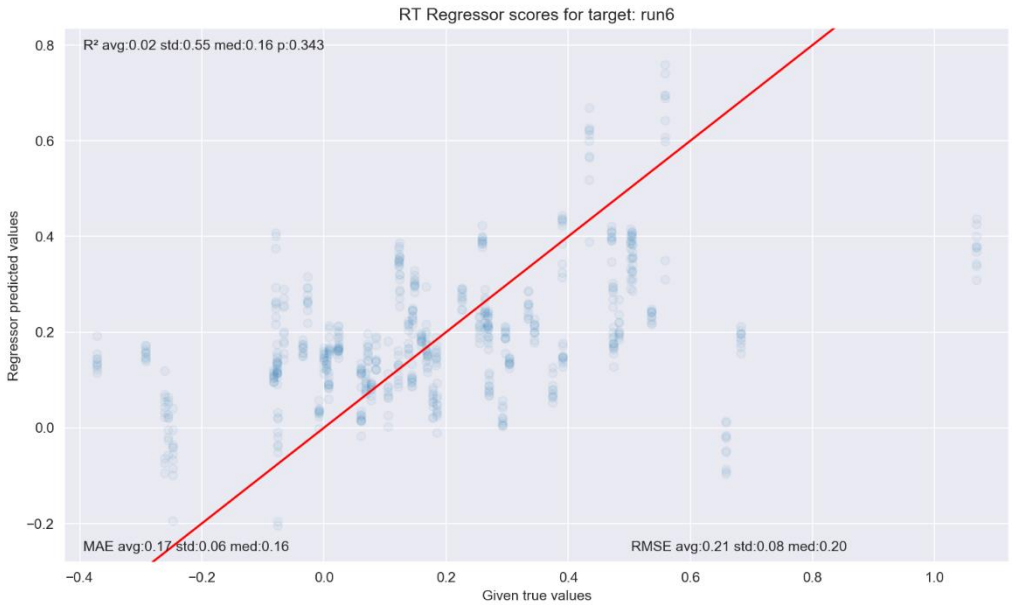

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 13 Randomized Trees predicting run 5




$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

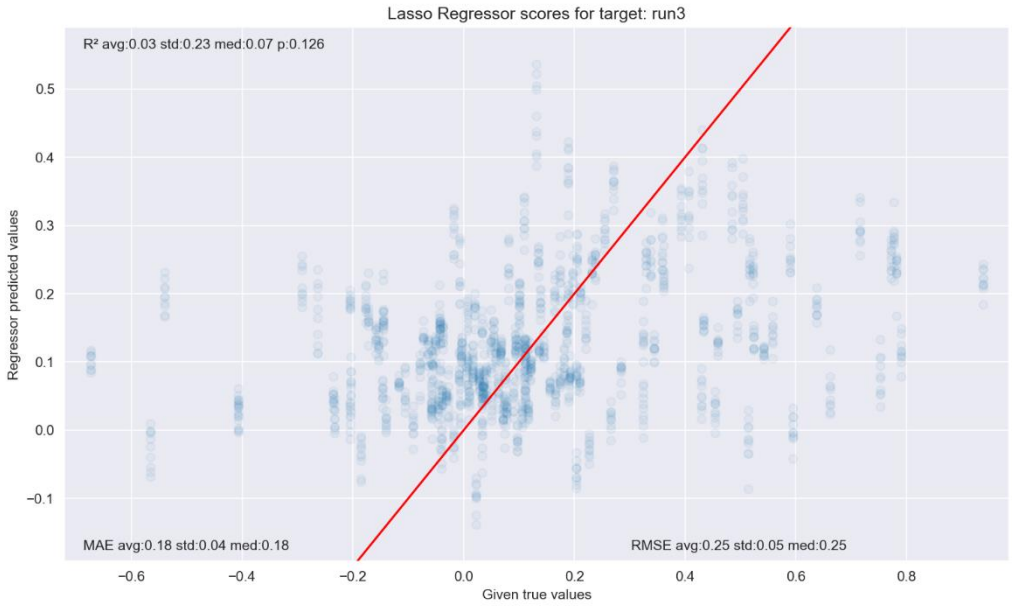
Figure 14 Randomized Trees predicting run 6




$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

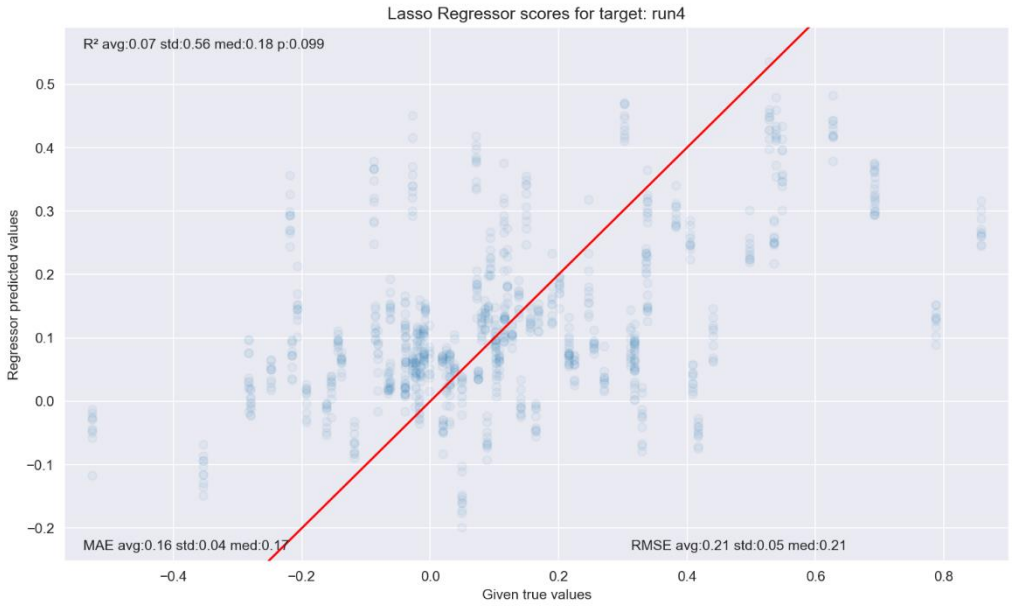
***Models trained on all available information***

Figure 15 Lasso regularised linear regression predicting run 3



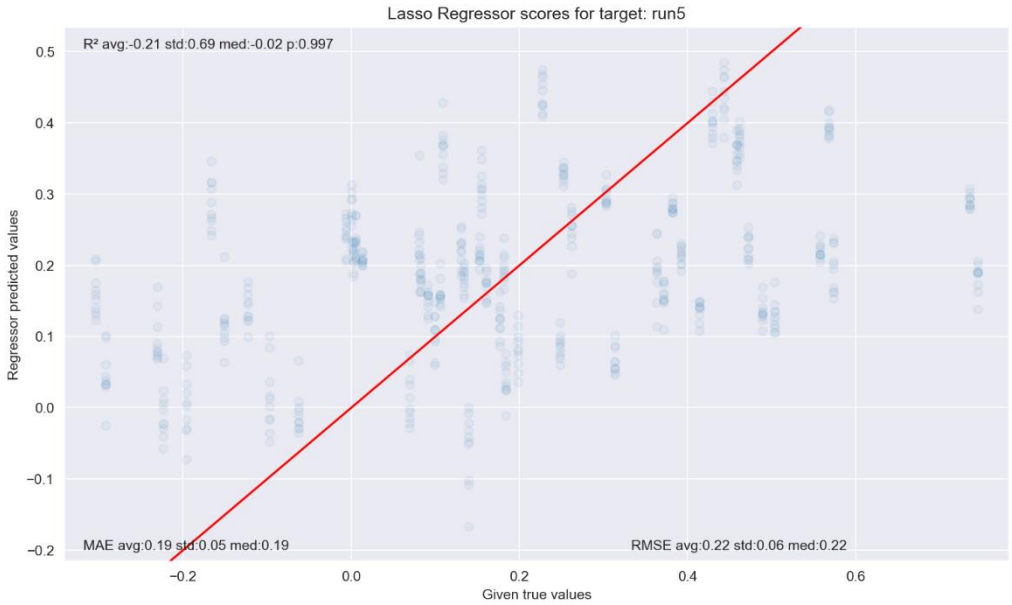

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 16 Lasso regularised linear regression predicting run 4



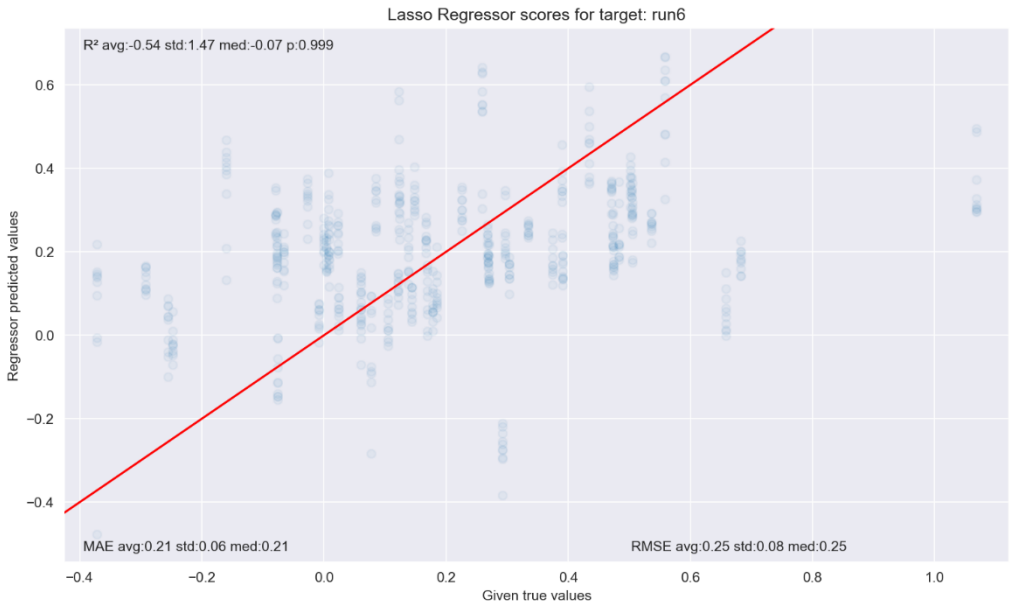

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 17 Lasso regularised linear regression predicting run 5



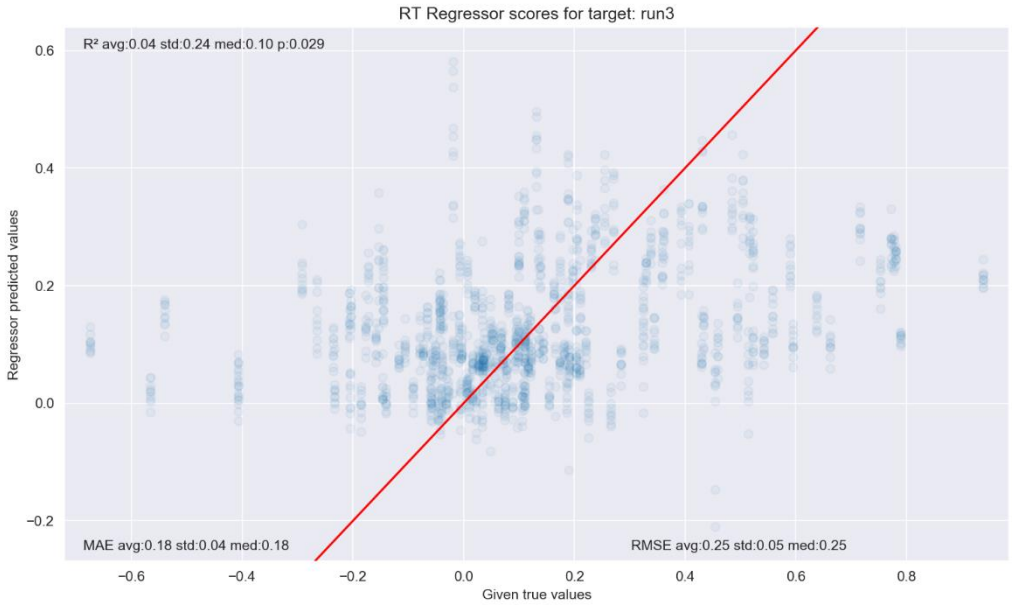

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 18 Lasso regularised linear regression predicting run 6



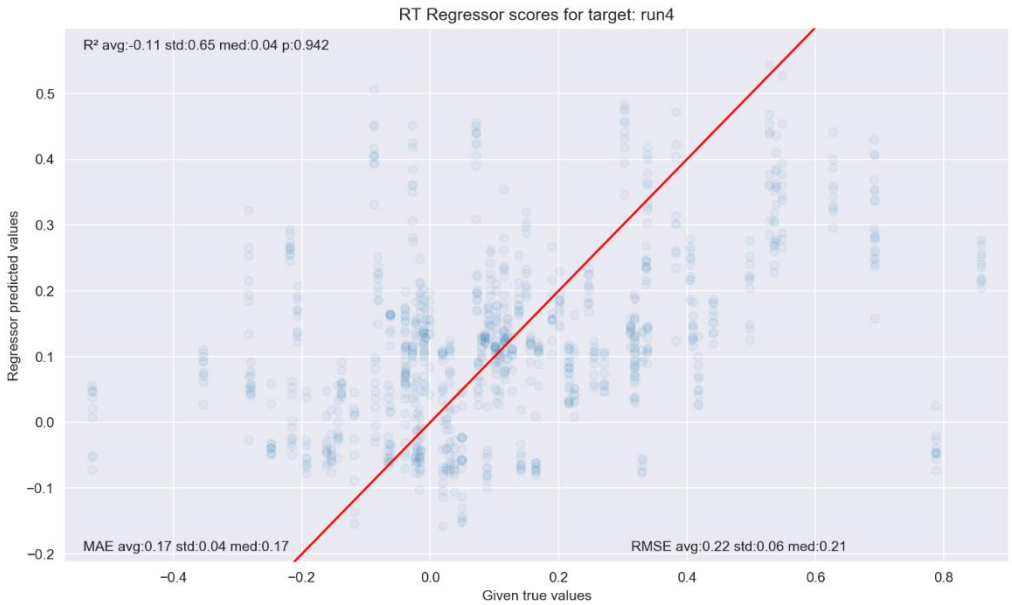

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 19 Randomized Trees predicting run 3



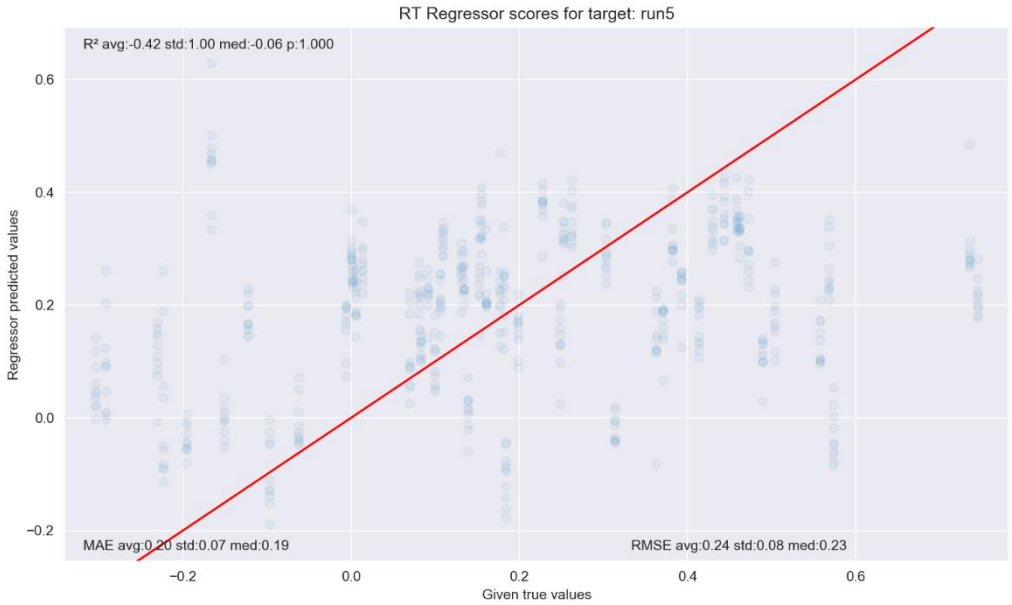

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 20 Randomized Trees predicting run 4



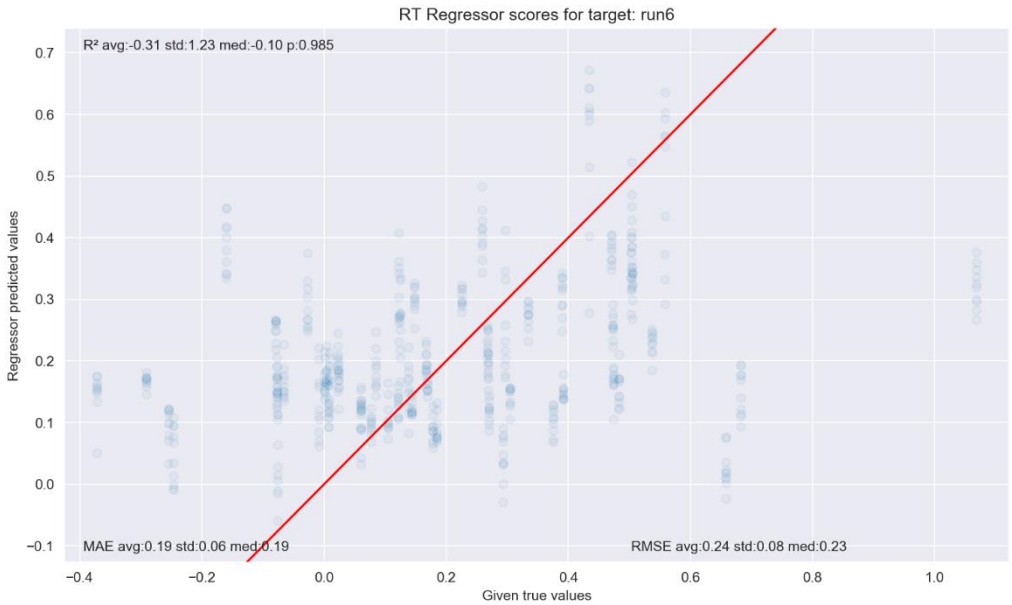

$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 21 Randomized Trees predicting run 5




$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

Figure 22 Randomized Trees predicting run 6




$R^2$ = coefficient of determination. MAE = mean average error. Std = Standard deviation. Med = median. RMSE = Root mean squared error.

***Feature importance analysis***

Feature importance rankings for regularised linear regression predicting run 3

Figure 23

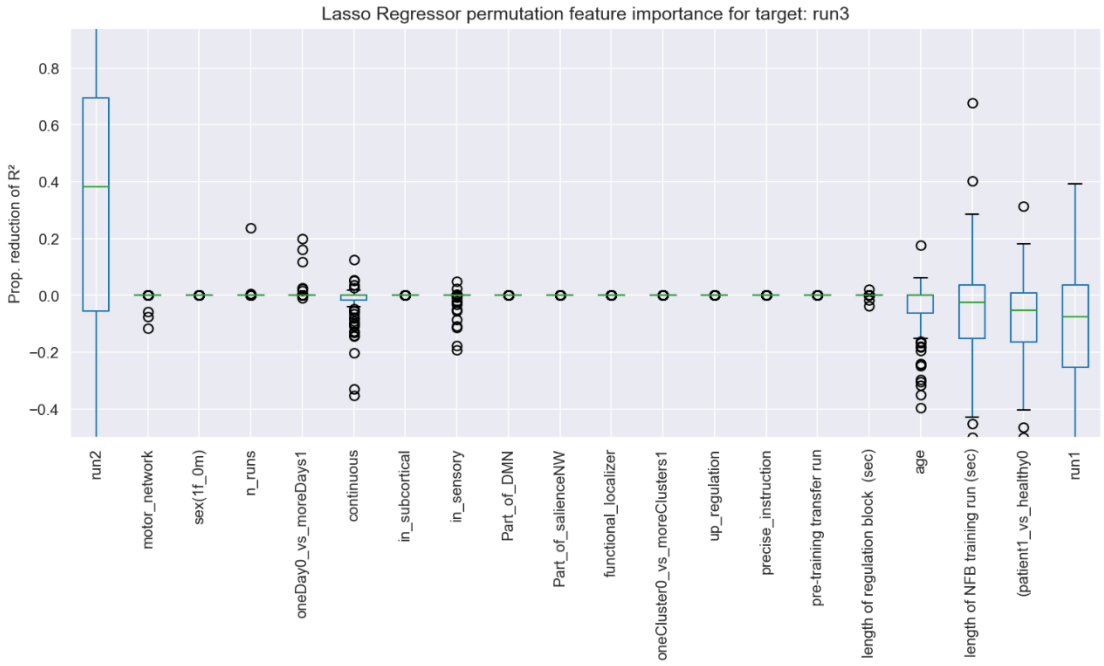
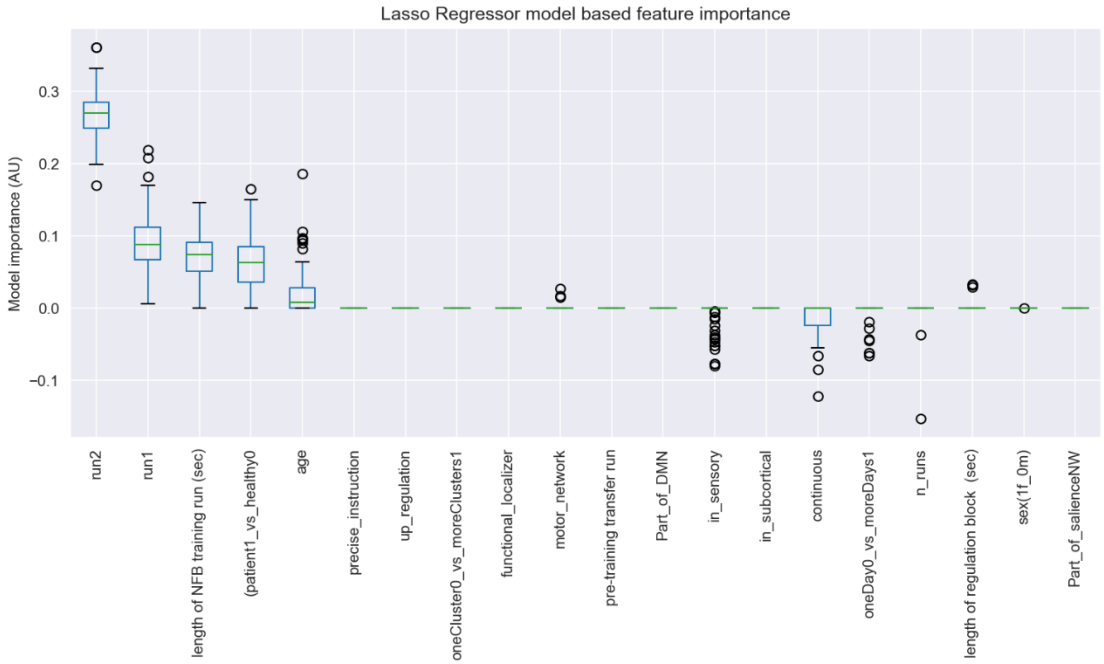



Figure 24



Feature importance rankings for regularised linear regression predicting run 4

Figure 25

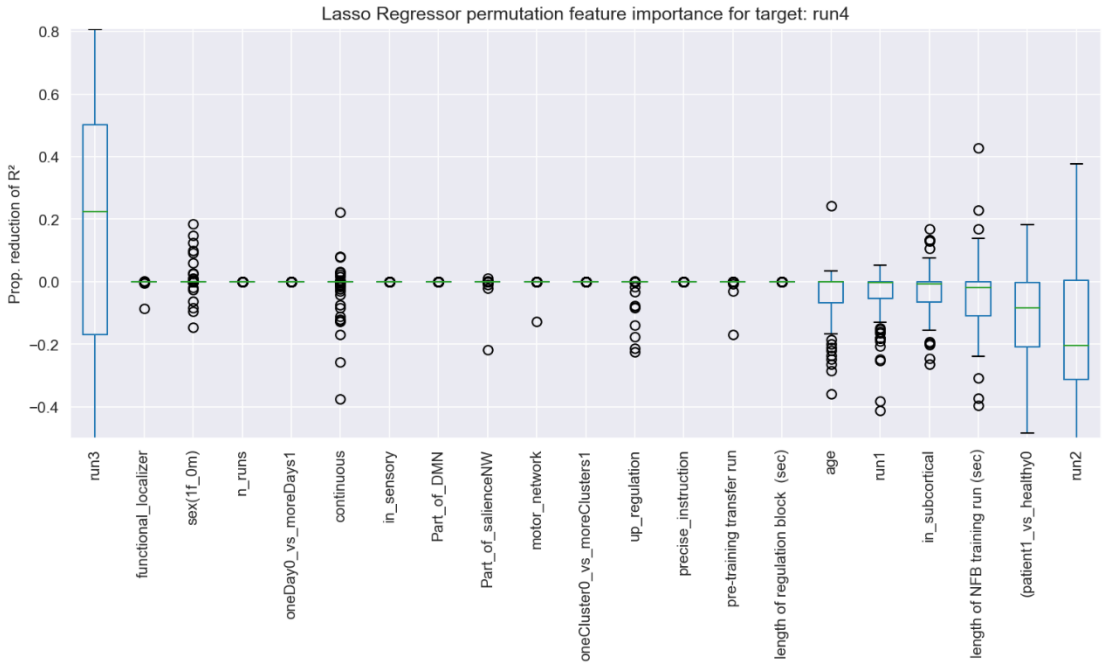
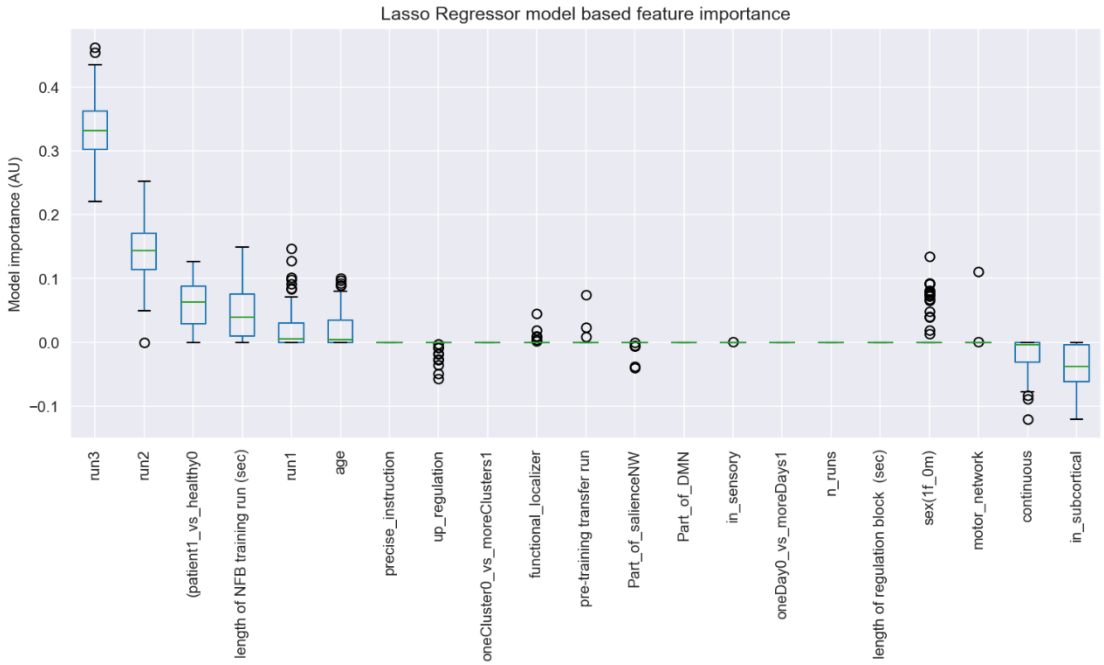



Figure 26



Feature importance rankings for regularised linear regression predicting run 5

Figure 27

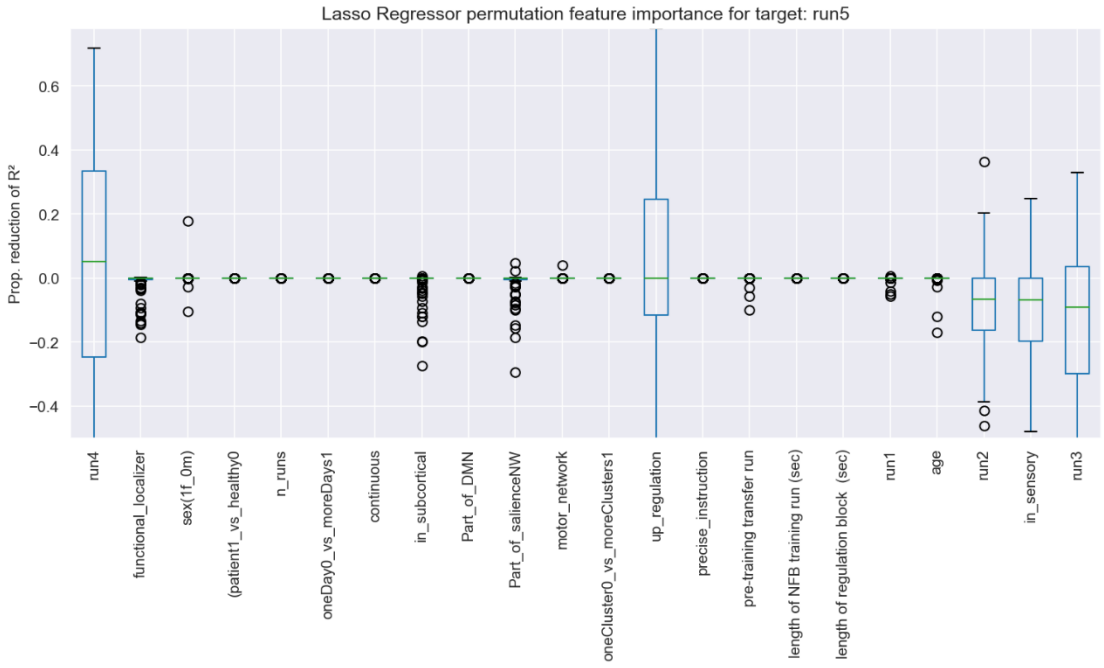
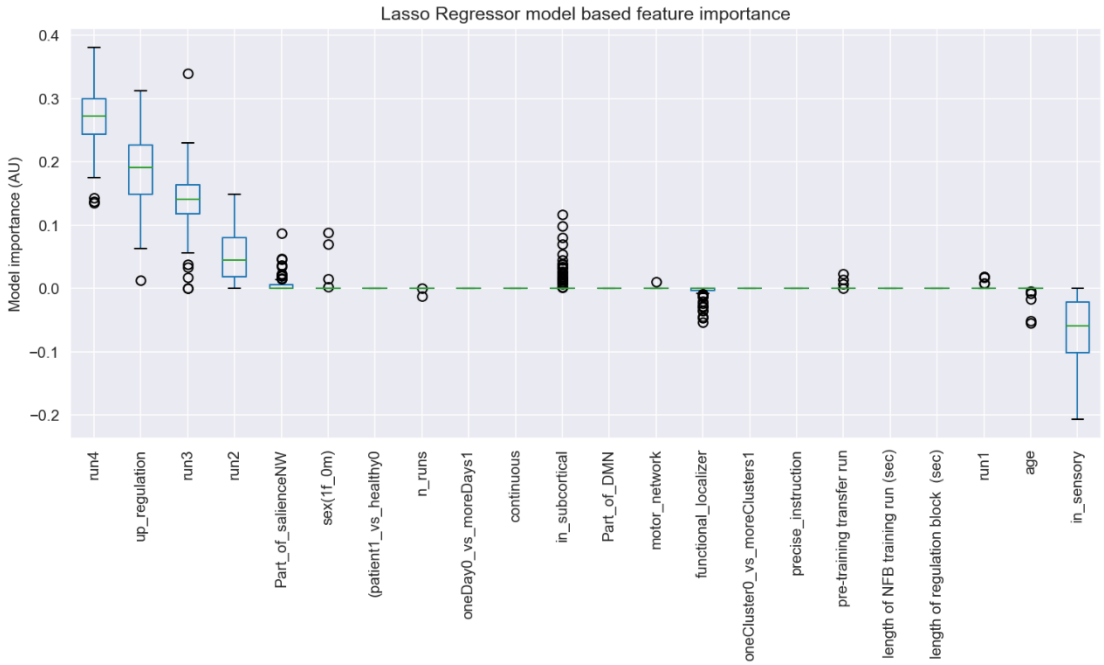



Figure 28



Feature importance rankings for regularised linear regression predicting run 6

Figure 29

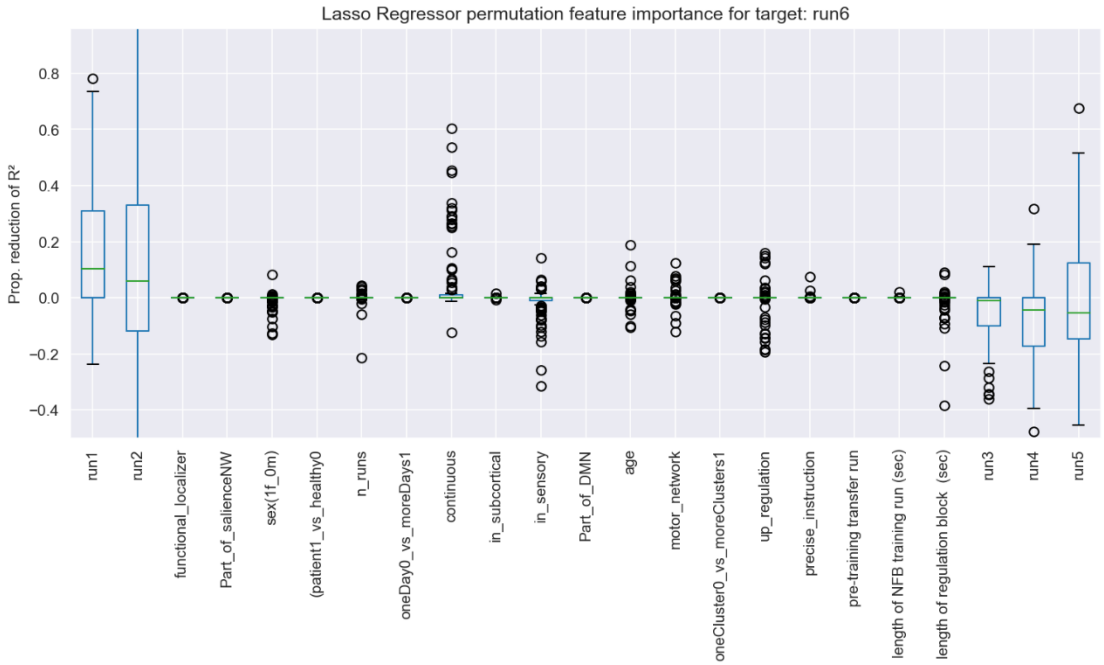
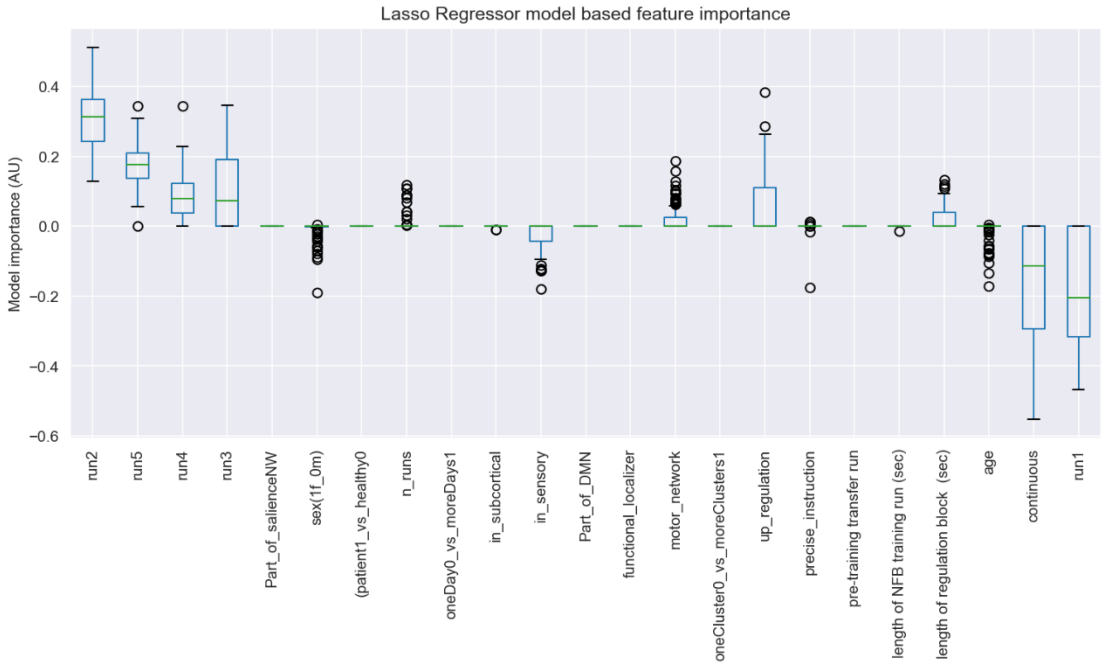



Figure 30



Feature importance rankings for randomized trees predicting run 3

Figure 31

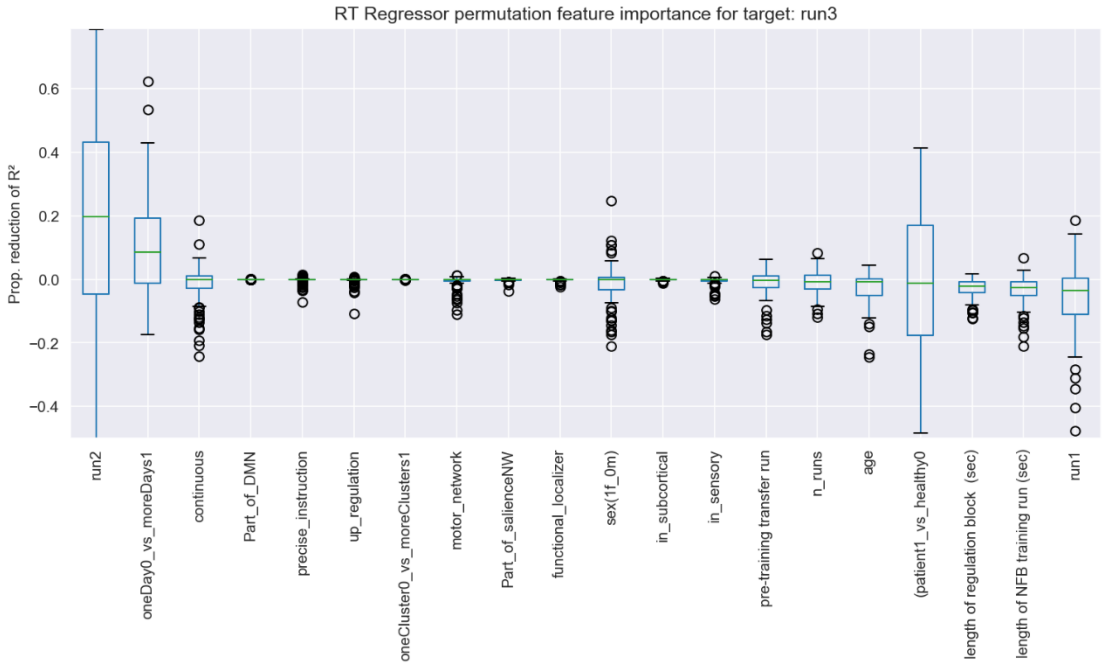
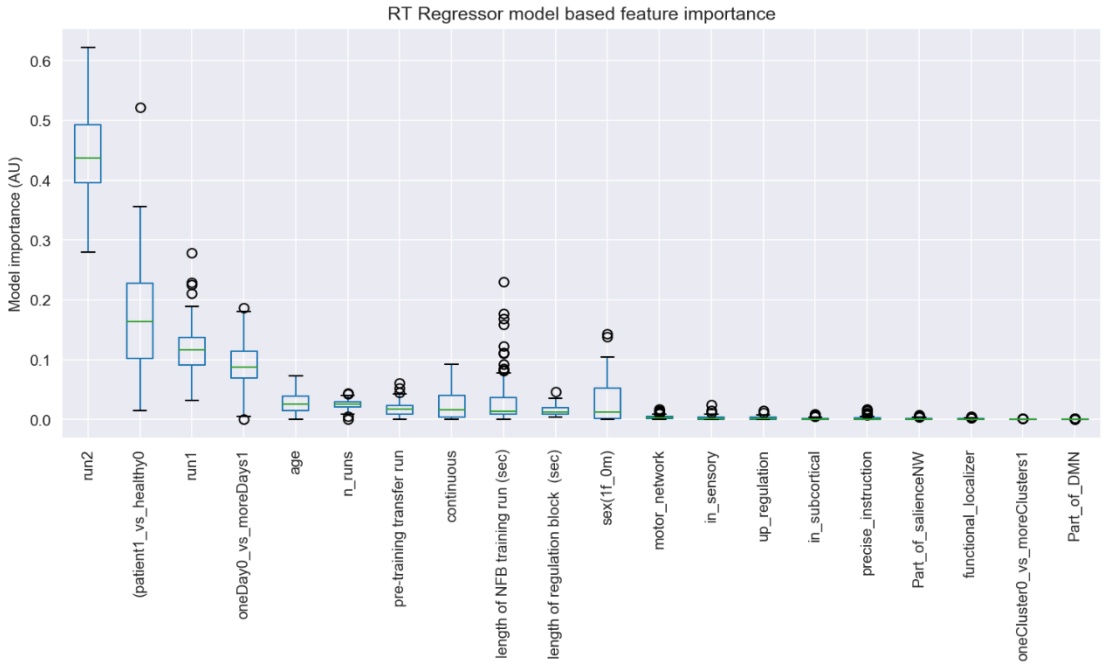


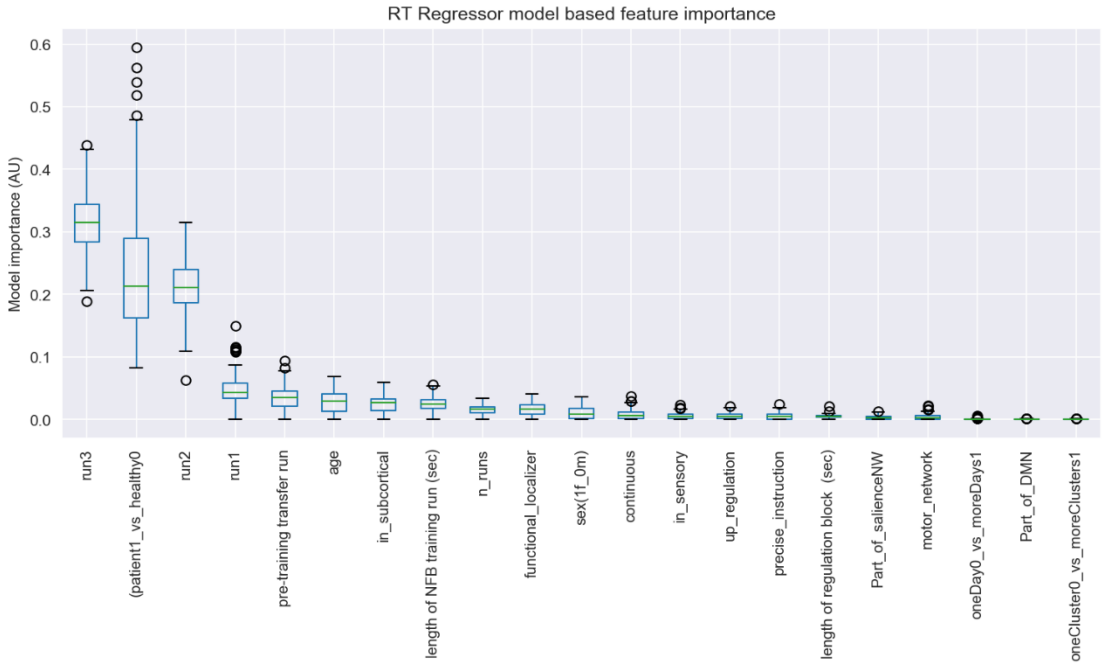

Figure 32



Feature importance rankings for randomized trees predicting run 4

Figure 33


Figure 34



Feature importance rankings for randomized trees predicting run 5

Figure 35

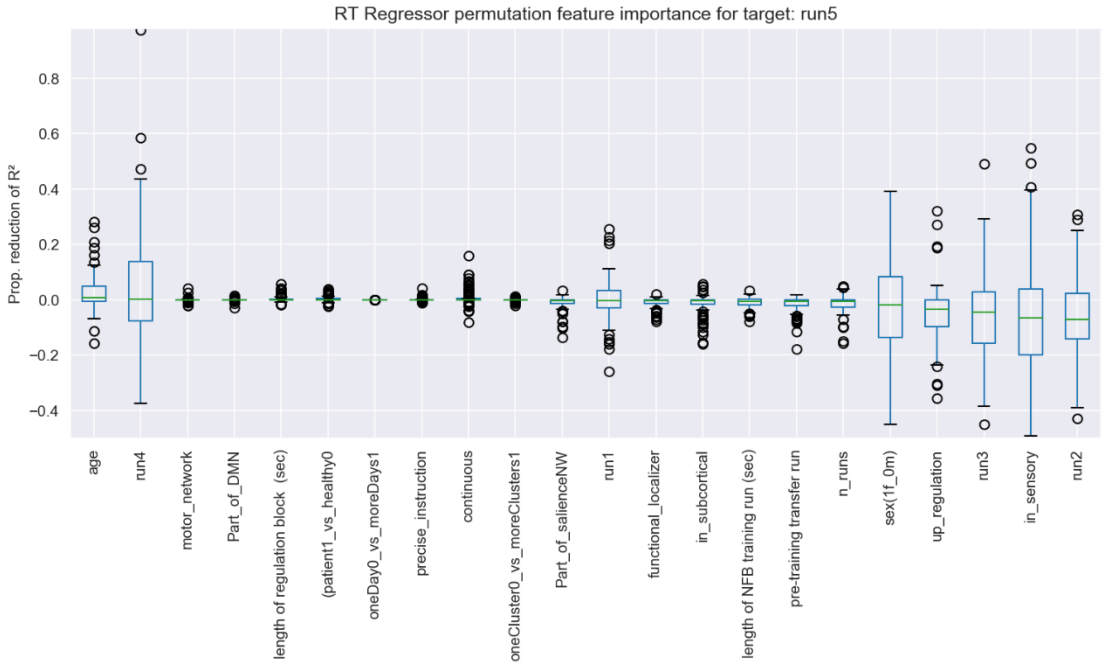
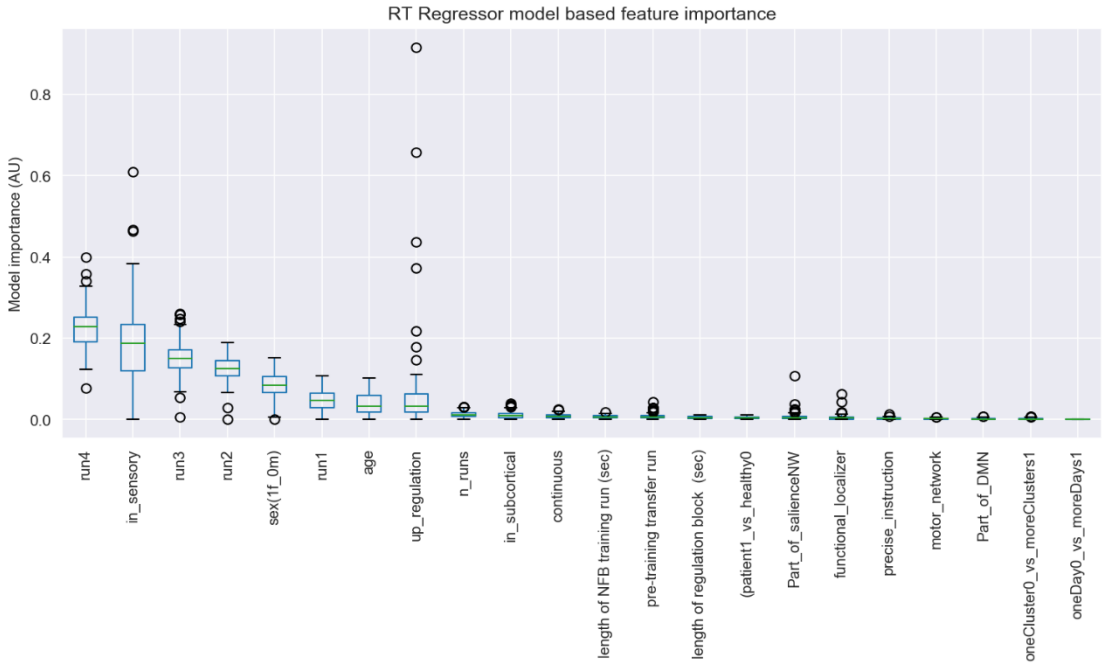


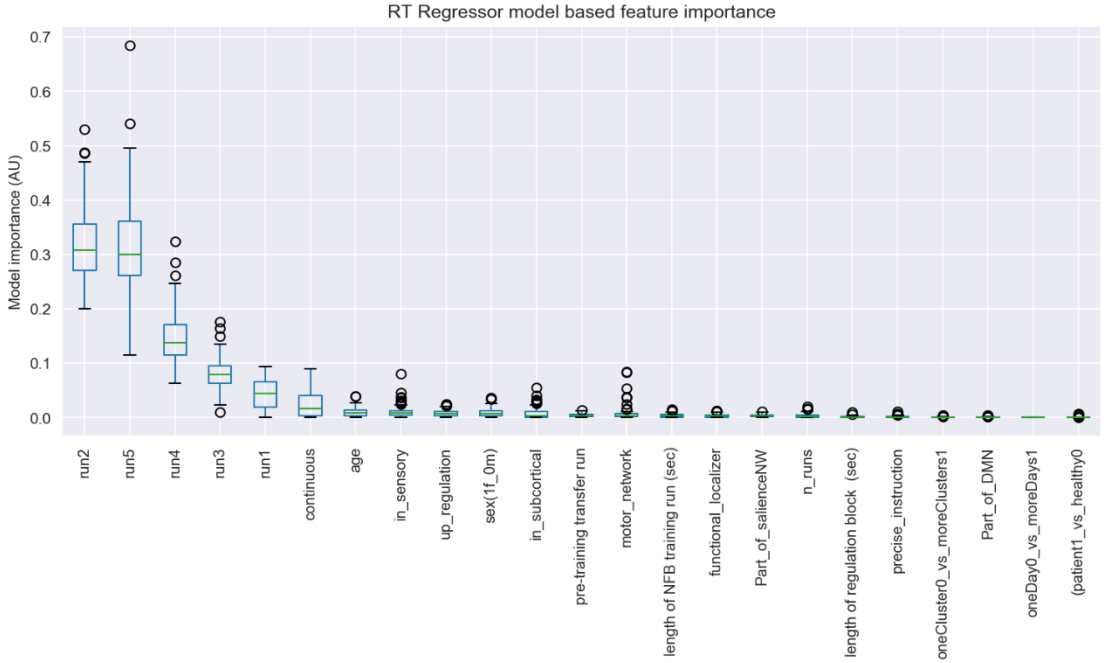

Figure 36



Feature importance rankings for randomized trees predicting run 6

Figure 37


Figure 38



**Acknowledgments**

I would like to thank my family and friends for always being supportive. Your critical feedback, proof reading, and mental support helped me a lot throughout the past months. Furthermore, I would like to thank Professor Frank Scharnowski for allowing me to join his group, giving me the opportunity to work on this exciting project, and enabling me to present my work. Lastly, I would like to thank David Steyrl for guiding me through this project. Your patience, advice, and expertise made the last months one of the most educational experiences I had.

I very much hope to be able to further collaborate with both of you on future projects in the years to come.