Information theoretic perspectives on en- and decoding in audition and vision

Christoph Daube

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Neuroscience and Psychology
College of Science and Engineering
University of Glasgow

January 2022

Abstract

In cognitive neuroscience, encoding and decoding models mathematically relate stimuli in the outside world to neuronal or behavioural responses. While both stimuli and responses can be multidimensional variables, these models are on their own limited to bivariate descriptions of correspondences. In order to assess the cognitive or neuroscientific significance of such correspondences, a key challenge is to set them in relation to other variables. This thesis uses information theory to contextualise encoding and decoding models in example cases of audition and vision. In the first example, encoding models based on a certain operationalisation of the stimulus are relativised by models based on other operationalisations of the same stimulus material that are conceptually simpler and shown to predict the same neuronal response variance. This highlights the ambiguity inherent in an individual model. In the second example, a methodological contribution is made to the problem of relating the bivariate dependency of stimuli and responses to the history of response components with high degrees of predictability. This perspective demonstrates that only a subset of all stimulus-correlated response variance can be expected to be genuinely caused by the stimulus, while another subset is the consequence of the response's own dynamics. In the third and final example, complex models are used to predict behavioural responses. Their predictions are grounded in experimentally controlled stimulus variance, such that interpretations of what the models predicted responses with are facilitated. Together, these three perspectives underscore the need to go beyond bivariate descriptions of correspondences in order to understand the process of perception.

Contents

List of Tables

List of Figures

Preface

0.1 Author's note

In late 2015 I applied for a PhD scholarship of the College of Science and Engineering at the University of Glasgow. For this application, I had to write a "research plan". Six years later, I am finally close to submitting my thesis, and stumble across this document. I decide to include it at the start of my thesis, as it documents the insecurities and the enthusiasm, the fears and the hopes as well as the ignorance and the understanding with which I went into this project.

When I ask myself what my younger self would have thought of this result, I cannot help but feel surprised of how many of the ideas mentioned in the plan indeed found their way into the final thesis.

Perhaps most striking seems the somewhat haphazardly cited work by Güçlü & van Gerven (2015), as I had no idea whatsoever that I would end up venturing into the modeling of visual processing with deep neural networks.

Another peculiar aspect is that the plan was centered on the idea of describing cortico-cortical connectivity with Transfer Entropy (TE) and "content-based connectivity measures" (Ince et al., 2015), which at the time were still under development. I indeed spent a great deal of my time on the maturely suggested simulations, something I often came back to from different stages of other projects. During those simulations however, I built up a great deal of skepticism towards the idea of cortico-cortical connectivity, especially with magnetoencephalography (MEG). This was probably best summarised in the work of Mehler & Kording (2018). The plans were thus changed to a restriction to cerebro-peripheral connectivity. That some of the TE simulations have still made it into a chapter of this thesis makes me very happy, not least since this chapter was finalised at the very end of the PhD.

A further mismatch is between the goal of "integrating auditory encoding into a broader, brain wide perspective" and the thesis' focus on bilateral auditory cortices. In the MEG passive story listening data I collected, I did not find clear evidence for robust systematic explainable variance outside of auditory cortices. This might on the one hand stem from a relatively conservative approach in identifying such regions (correlation of responses to a repeatedly presented chapter together with considerations of cross-talk and point-spread functions in MEG source space). On the other hand, this might stem from the task as such, which did not involve behaviour, and thus did not trigger a cascade of

brain activity from sensory to motor and perhaps frontal areas. Cascades like this are arguably on the spatial scale which the spatial resolution of MEG is best suited to exploit (Gwilliams & King, 2020).

A deep regret is that despite being aware of it when writing the research plan, we forgot to cite the in my opinion excellent work by Ding et al. (2016) in the chapter on linguistic vs acoustic representations. Its omission in the chapter makes it appear a little short-sighted – a reference to well-controlled experiments should have been part of its discussion section.

Nevertheless, the combination of MEG- and information theory expertise of Joachim and Robin, who would become my supervisors, the interest in abstraction, and the use of higher-order information theoretic concepts such as redundancy and synergy are indeed well represented in the final thesis.

I hope that you, dear reader, can enjoy some of it.

0.2 Research plan

I am fascinated by the seemingly infinite multitude of processes on any conceivable scale that result in human beings listening to their environment. For me, it is thrilling to discover patterns of neuronal activity which almost omnipresent phenomena such as verbal communication or music critically rely on.

In my studies, I was so far involved in several projects that were driven by an interest in basic science derived from this fascination. They mainly revolved around the role of the phase of low frequency EEG oscillations in primary auditory encoding. From Bachelor to Master studies, my investigations spanned the range from implementing and executing psychophysical and tACs experiments to fully designing, measuring and analyzing EEG studies.

For my Master's Thesis, I then decided to broaden my methodological scope by taking the chance of analyzing a previously acquired MEG dataset whose texture stimuli were similar to those of an undergraduate project of mine. Heavily inspired by work from the group of Joachim Gross (Park et al., 2015) I set up an analysis framework geared towards the identification of directed cortico-cortical communication using beamformer techniques and phase transfer entropy.

I would now like to continue to integrate auditory encoding into a broader, brain wide perspective. This relates to one of the perhaps oldest strands in the history of auditory neuroscience (Broca et al., 1861) and leads to the concept of processing streams (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009). Similarly to visual neuroscience, ventral and dorsal pathways have been proposed as an overarching principle explaining brain structure and functionality. However, these concepts so far mainly stem from fMRI studies with low temporal resolution. The current models are thus largely based on studies revealing rather static processing specializations (e.g. Obleser et al., 2007) and long range fiber tracts (Friederici, 2009). A transformation and spread of information in

space (cf. Güçlü & van Gerven, 2015) can only be indirectly inferred from these studies. An intuitive notion of a "stream" however critically relies on a direct demonstration of the dynamics of information representation.

During my PhD, I would like to contribute to filling this gap with a methodological approach that for me builds on the skills I have acquired during my Master's Thesis and for which expertise is available in Glasgow. A combination of beamformed MEG recordings (Gross et al., 2001) and content based connectivity measures (Ince, in prep.) seems ideally suited to improve our understanding of the human cortical framework that is built around primary encoding of the environment.

The project might start out with an assessment of the degree to which current models can be helpful in interpreting beamformed MEG data. An initial experimental design could be geared towards provoking a dissociation between ventral and dorsal pathways (cf. Saur et al., 2008). To start with, first analyses should capitalize on classic event related fields, oscillatory components and bidirectional measures of connectivity.

In a next step, it might then be possible to test the concept of forward and inverse mapping (Rauschecker & Scott, 2009). For the dorsal stream, it would be particularly exciting to investigate a proposed integration of efference copies and sensory signals in parietal cortex (see also Morillon et al., 2015). Regarding the ventral route, it would be interesting to see whether results obtained with microstimulation in macaque monkeys (Petkov et al., 2015) can prove their claimed significance. One would for example expect to find information stemming from primary or belt areas and transported via direct fiber connections to frontal cortex to later merge at their destination with outputs from anterior temporal networks.

Information theoretic measures such as redundancy and synergy or directed feature information seem of high relevance here. Their application to the above mentioned questions might well make simulations necessary to ensure they can capture the relevant effects.

A central question then is how to keep track of the domains the information is transformed into, especially under the limitation of neural mass signals recorded from outside the skull (Panzeri et al., 2015). A candidate concept that has repeatedly been proposed to be applicable to this problem is cross frequency coupling (Lakatos et al., 2005; Gross et al., 2013; Jiang et al., 2015). With a perspective on auditory functions that also covers music processing, it could finally be attractive to expand the approach developed so far correspondingly and use this concept to delineate more universal abstraction gradients allowing for hierarchical representations of information (Ding et al., 2016).

Acknowledgements

Thank you Robin. You have not only blown my mind, been as incredibly “sharp" (Kay, J.W., 2019, private conversation) as modest, whispered to computers, shared your no-nonsense scientific approach and its fruits, given me quasi-instantaneous and rich feedback, made me wonder where your other six arms are that allow you to be so prolific, helped me think through problems and saved my ass at ungodly hours time and time again, you have also been a moral compass, a motivator when I was down, a generous host, a considerate listener to my every woe, a competent consultant on all-encompassing aspects of life and a middleman to the highlights of British and internet quality humour.

Thank you Joachim. Thank you for making my PhD in Glasgow happen, for so much of your time, wisdom and patience, your interest in and encouragement to my science babysteps, your immediate and detailed feedback, for magically pushing the right buttons at the right time to make me feel respected, valued and motivated while selflessly drawing the attention away from your own gargantuan achievements, for the enthusiasm pervading anecdotes about you of any of your collaborators, and last but not least, for that unforgettable dinner with yourself, Christo and Jan-Mathijs. In these times, may you be surrounded by powerful angels similar to the one you have been for me.

Thank you Philippe. Thank you for providing me with perspectives and opportunities, for teaching me the historic context of cognitive science and its loops, for taking the time for long and challenging discussions while playing the 4D chess of running a department, for your pursuit of elegance and timelessness, for being an overwhelmingly industrious and understanding inspiration during the last two pandemic years, and for our chats about high magnitudes of rate of change of position on specially prepared surfaces which vehicles can use.

Thank you School of Neuroscience and Psychology, especially Oli, Christian, Bruno, Gregor, Gavin, Frances, Lindsay, John, Gary, Andrew, Sara, Christoph, Dale, Lisa, Rachael, Marc, Lucy, Yinan, Mimma, Roberto, Hame, Tuomas, Martina, Léo, Anne, Manuela, Victor, Chris, Merle, Miika, Tyler, Laura, Jack, Katarina, Jasper, Anna, Michele, Andrew, Jiayu, Yaocong, Kasia, Meng, Sander and Lukas.

Thank you scientific community, especially Jonas, Laurent, Jan-Mathijs, Molly, Jonathan, Tim, Giovanni, Edmund and anonymous reviewers.

Thank you Conrad, Gustav, Laura, Tobias, Lennart, Lea, Ronja, Lily, Moritz, Nadine, Benjamin, Carlos, Przemo, Natalia, Marek, Peejay, Veera, Joss and Sarah.

Thank you Aniol, Donata, Domitille, Herwig, Doris, Rita, Tantri, Ana, Kyle, Matt, Ewa, Felix, Max, Matt, Max, Gavin, Salvador, Ted, Katie, Alec, Emily, Flora, Paddy, Chuck, Florence, Rosie, Lewis, Sophia, Riley, Leilani, Sophie, Evie and Willow.

Thank you Anna, Jürgen, Jürgen, Eva, Johannes, Phil, Susanna, Josephine, David, Anne-Maria, Niklas, Theda, Leander, Clara, Georg, Assi, Dietrich, Vera and Karl and Hans.

This thesis was powered by Cottonrake's seeded rye bread.

Declaration

All work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

Introduction

1.1 Outline

This thesis deals with the question of how the outside world is related to human neuronal or behavioural responses. Unless one is interested in sensory neurons, which directly interact with the environment, answering this question entails assumptions about unobserved or “latent" phenomena that take place between the presentation of a stimulus and the response of downstream neurons or behaviour of interest. These assumptions are manifested in terms of models that specify which transformations of a stimulus are believed to be relevant to the responses, and that thus aim to bring light in the dark of our understanding of the processes that happen between the presentation of a stimulus and the elicitation of responses.

For example, we can consider a sound as it leaves a speaker. It exists in the form of pressure deviations at various intensities and frequencies that impinge on the eardrums before they get encoded in a frequency-specific way by the cochlea (McDermott, 2013). This brings the sound into the so-called spectrographic format, with comparably high agreement across the field. What happens to the sound afterwards, as neurons across multiple relay stations of the subcortical auditory pathway propagate it to the auditory cortex and beyond, is however heavily debated.

Likewise, we can think of light as it is reflected by an object in our visual field. It traverses the lens of the eye and excites the rods and cones of the retina, passes through multiple layers of further retinal neurons, travels through the optic nerve past the optic chiasm and lateral geniculate nucleus to finally arrive in the back of the brain, the so-called occipital cortex (Palmer, 1999). Although a lot of research has explored the cascade of neuronal firing that is triggered here and that eventually leads to behaviour, as of now we are nowhere near having “solved" human vision.

As a consequence, many models exist that compete to predict and explain neuronal or behavioural responses to sound or light. They all specify different transformations implying the relevance of different aspects of the stimulus for a certain type of responses. In order to achieve progress in the field of perceptual cognition and neuroscience, we thus need new approaches to adjudicate between the plethora of hypotheses about stimu-

lus transformations that have been and that will be developed. This thesis hopes to contribute to such progress by means of suggestions on how to relate models to one another, how to relate stimuli to responses and how to develop and exploit suitable experimental paradigms to obtain meaningful stimulus response pairings in the first place. A central common element of these suggestions is the adoption of a genuinely trivariate point of view on those problems (Figure 1.1). In this introduction, I will first generally outline each of these information theoretic perspectives on en- and decoding in audition and vision, and then place them in broader contexts of the terms constituting the thesis title.

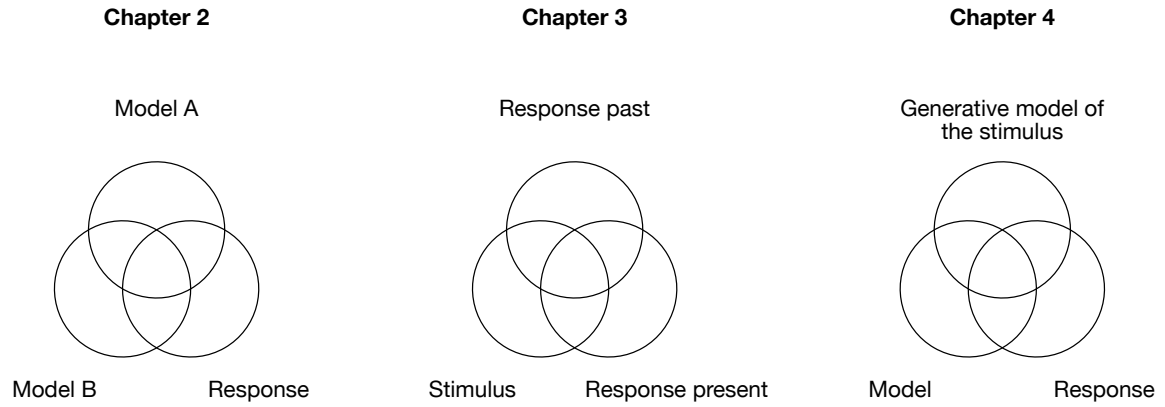

Figure 1.1: **Triple Venn diagrams illustrating concepts in chapters 2, 3 and 4.**

1.1.1 Going beyond pairwise tests of model performance

In chapter 2, we consider the general problem that the degree to which a model accurately predicts responses is as such insufficient to argue for or against the relevance of any of the assumptions manifested in the model to the system under study. In simple terms, this is because often, many other models can be constructed that achieve the same or an even better accuracy by virtue of other assumptions. From the perspective of philosophy of mind, this problem can be seen as an example of multiple realisability (Putnam, 1967).

We substantiate this point with a Magnetoencephalography (MEG) experiment in which humans listen to a continuous speech stimulus (Poeppel & Embick, 2005). We compare different models of varying complexity and demonstrate that even relatively simple models can reach prediction performances comparable to a given complex model. Moreover, we develop a framework to compare models not only with respect to their prediction performance, but also with respect to the amount of information about the responses of interest that they share with a competing model. This is important because it is easily conceivable that two models reach comparable performances by means of correctly predicting different parts of the responses, while failing to predict those parts that the respective competing model succeeds to predict. In our case however, we demonstrate that the simple model indeed succeeds in predicting the same parts of

the responses that the complex model predicts. Based on these findings, we make a case for parsimony as an often undervalued design principle for models of neuronal or behavioural responses. Identifying the simplest model that suffices to predict a given phenomenon is a promising goal to understand the phenomenon.

1.1.2 Predicting responses from stimuli and response history

Chapter 3 suggests an alternative to the paradigm of models that predict responses from stimuli alone. Specifically, we consider the perspective of Transfer Entropy (TE, Schreiber, 2000), where separately to the stimulus, the past of the response variable is considered for the prediction of a given present sample of the response. In cases where the responses are temporally correlated, it is possible that these “auto-correlations" between the response past and present can account for some portion of the information about the response available from the stimulus. TE aims to ignore such auto-information of the response when quantifying stimulus-reponse relationships. It is thus highly relevant for research questions pertaining to the neuronal prediction of upcoming stimulus material (Friston, 2005).

We focus on the issue of frequency-specific narrowband components of responses, which are ubiquitous in neuronal mass signals as recorded with MEG (Wang, 2010). Such band-limited components are by definition highly auto-correlated, but are usually seen as problematic in combination with TE. We develop an estimator that tackles such problems and subject it, together with classic estimators, to an extensive range of tests based on simple simulations. We finally explore the behaviour of our estimator on the same MEG data as studied in chapter 2, finding that measures of stimulus-response delay and interaction strength differ from those recovered by bivariate dependency measures.

1.1.3 Constraining the features which humans and models can use with experimental control

Chapter 4 revisits the issue of multiple realisability from chapter 2. Here, we make an additional suggestion of a constraint that could alleviate mappings of properties of a model to the system whose responses it tries to predict. The idea is that experiments should make use of generative models of stimuli to use the opportunity to decorrelate stimulus features of interest (Olman & Kersten, 2004). As concluded in chapter 2, if as in experimental designs involving naturalistic stimulus material this opportunity is not seized, it is harder to isolate effects of a stimulus feature of interest since many features will be confounded with one another. To the extent to which experimental control rules out confounds, it becomes possible to make causal statements regarding the relationship of stimulus features and responses.

This is of additional relevance when trying to interpret response predictions of com-

plex models. If we reduce the ambiguity about what stimulus features cause human responses, this also reduces ambiguity about what complex models can predict human responses with. With this approach, the chapter not only shows what kind of model of a set of complex candidate models can predict behavioural responses to face stimuli best, but also grounds its predictions in experimentally controlled shape features.

1.2 Information theory

In order to study relationships between the components of the individual chapters, this thesis makes use of data analysis strategies that are rooted in information theory. At its core, information theory is a set of formalisms to interpret probability distributions and relate them to one another. Together, these were originally conceived of as a “mathematical theory of communication" (Shannon, 1948; Cover & Thomas, 1991). The initial remit was in telecommunication, where messages of a given size were to be passed from a sender to a receiver along a noisy channel of a given capacity. Applications to neuroscience have both been criticised (de-Wit et al., 2016) and praised (Quiroga & Panzeri, 2009). In this thesis, we take a pragmatic stance and simply use information theory as a principled framework to quantify statistical dependencies between variables of interest. The following section will develop the information theoretic quantities relevant to this thesis.

1.2.1 Entropy

A central interpretation of probability within information theory is given by the definition of entropy, which can be described as the amount of uncertainty associated with a given probability distribution. For a discrete variable with a given number of possible states, it is defined as the product of the probability of a given state with the logarithm of the reciprocal of that probability (the latter factor can thus be rewritten as the negative logarithm of the probability), summed over all possible states. For a binary variable that can thus only assume one of two states, this results in a symmetric positive curve that peaks when both states of the variable are equiprobable (Figure 1.2). This corresponds to the uncertainty of the toss of a fair coin. When the logarithms are computed with respect to base 2, this defines the perhaps most popular unit in information theory, the bit.

The definition of entropy for a discrete variable can be generalised to continuous variables. This can be done by understanding that the summing operation together with the weighting of the logarithm of the reciprocal of the probability of a given state by its probability reflects the expected value over the logarithm of the reciprocal of the probability. For continuous variables, the expected value can be obtained by instead integrating the product of probability density multiplied by the logarithm of the reciprocal of the probability density over the support of the probability distribution.

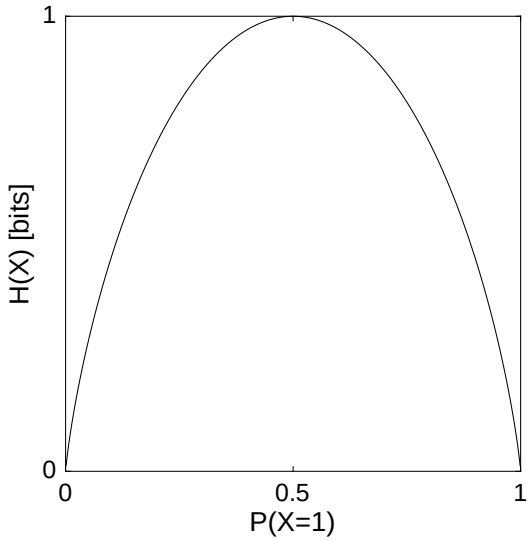

Figure 1.2: **Entropy of a binary variable X for all possible Bernoulli distributions.**

1.2.2 Mutual information

A fundamental quantity in information theory that justifies its relevance to problems of communication between a sender and a receiver is mutual information (MI). It can be computed by quantifying the joint entropy (using the joint probability distribution of the two variables) and subtracting it from the sum of the two marginal entropies. Graphically, it can thus be thought of as the set intersection of the entropies of two variables (Figure 1.3). In theory, it is a non-negative measure that is zero if and only if the two variables are statistically independent. As they become more dependent, the MI between them grows. In principle, it can thus be seen as a generalisation of the concept of linear correlation to arbitrary nonlinear relationships.

If we recall Figure 1.1, we however notice that the quantification of bivariate relationships is insufficient for an application within conceptualisations of the problems addressed in this thesis.

A widely known information theoretic quantity that is applicable to a trivariate system is conditional MI (Figure 1.3). In conditional MI, the relationship between two variables can be measured given that we already know a third variable. Often, this is described as “conditioning out" the third variable, since usually, the MI of two variables shrinks when conditioning it on a third variable (but see below). There are multiple ways to compute the MI between two variables A and B conditional on C. One possibility is to obtain it as the difference of two MI terms: We first construct a joint variable consisting of B and C and then compute the MI between this joint variable and A. From this, we subtract the MI between A and C.

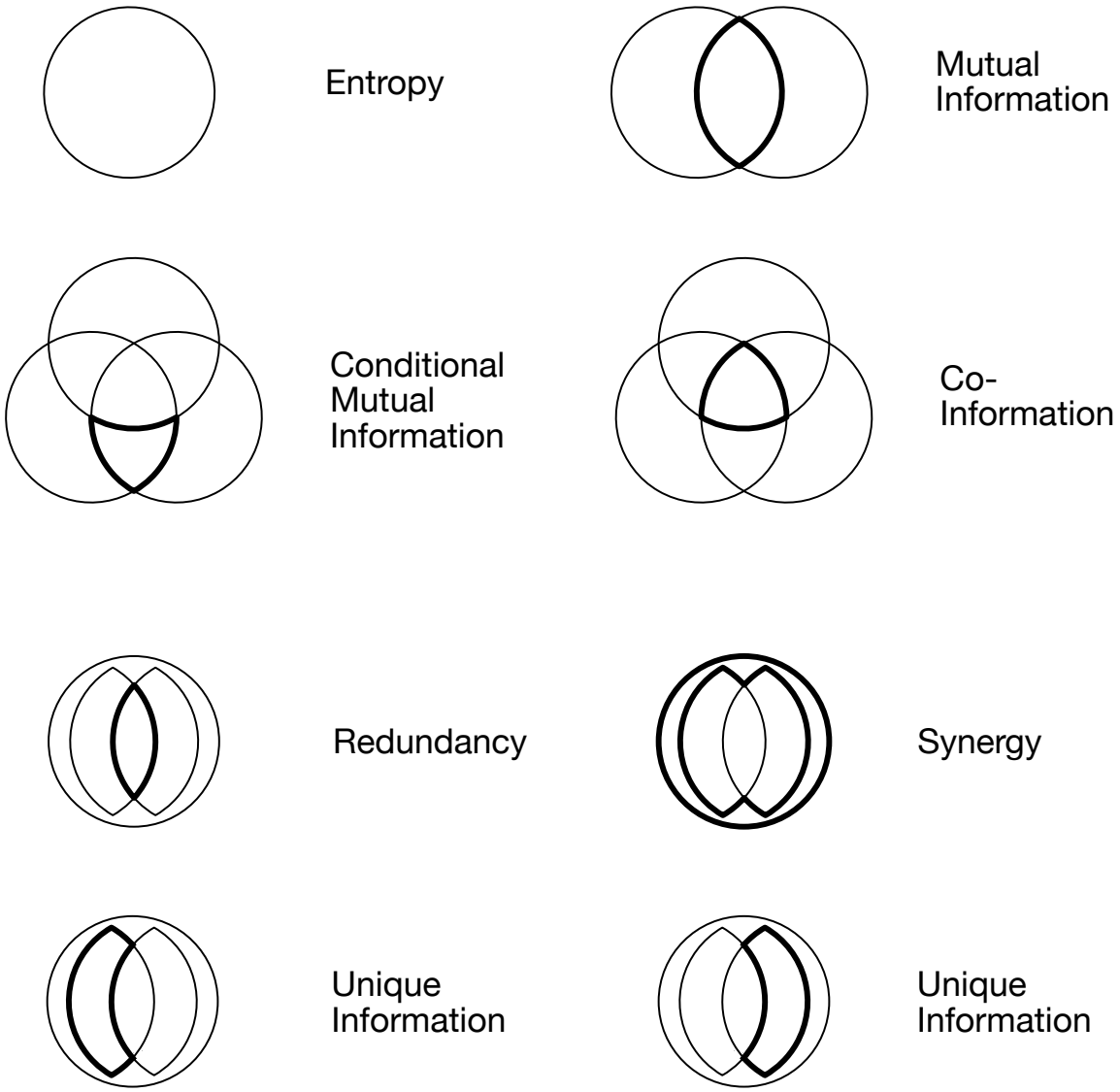

Figure 1.3: **Information theoretic quantities visualised as Venn diagrams.**

1.2.3 Co-information

If we are instead interested in the central triple set intersection, one option is to turn to interaction information (McGill, 1954). Here, the same idea that is used to relate entropies of two variables to one another is applied to two bivariate MI terms which share one variable (Figure 1.3). In this situation, it is often helpful to refer to the variable that appears in both MI terms as the “target", and the two variables that only appear in either of the MI terms as “sources". Note however that interaction information is symmetric with respect to its inputs, and the labels of source and target can thus be interchanged without changing the output interaction information. The triple intersection of three entropies can be obtained by subtracting joint MI of the two sources together with the target from the sum of two marginal MIs of each source with the target. By convention, this is referred to as negative interaction information or co-information. In cases where it is positive, i.e. in scenarios where the joint MI is smaller than the sum of the two marginal MI terms, this

indicates that the sources share the same information about the target. This is referred to as “redundant" information. In cases where it is instead negative, it indicates that there is information that can exclusively be obtained when jointly considering the two sources. This is referred to as “synergistic" information.

1.2.4 Partial information decomposition

Interestingly, this way of thinking about such higher-order interactions only recently got under closer scrutiny. In particular, it was pointed out that a problem of co-information is that it conflates both redundant and synergistic information into the same quantity (Williams & Beer, 2010). A problem with this is that it is theoretically conceivable that two sources at the same time both share information but also contribute synergistic information when considered together. These two quantities of synergy and redundancy would then cancel out in the net sum reflected by co-information. Therefore, the goal of partial information decomposition (PID, Williams & Beer, 2010) is to decompose the joint information of the sources about the target into four different “atoms": separate redundancy and synergy as well as unique information of each source (Figure 1.3). Note that this is generalisable to scenarios with more than two sources (Williams & Beer, 2010), leading to more than just four atoms. In this thesis however, only systems with two sources are considered. To achieve this decomposition, PID classically starts by quantifying redundancy. Once that is done, unique information can be obtained as the difference of MI and redundancy, and synergy can be obtained by subtracting redundancy and unique terms from the joint MI. How to obtain redundancy in the first place is however a matter of debate (Williams & Beer, 2010; Harder et al., 2013; Bertschinger et al., 2014; Ince, 2017a; James et al., 2018, 2019). In this thesis, the solution I_{ccs} provided by Ince (2017a) is used. For a detailed description, we here refer to the methods sections of the respective chapters. In brief, I_{ccs} resolves co-information on a pointwise level, such that terms that contribute to redundancy and synergy can be separated. To then obtain the global redundancy, the positive (i.e. redundant) pointwise co-information terms that co-incide with positive pointwise marginal and joint MI terms are summed.

From the perspective of PID, it can be explained how it is possible that conditioning bivariate MI on a third variable can not only lead to a decrease, but also an increase of conditional MI relative to bivariate MI. This is because conditional MI does not quantify unique information, but instead the sum of unique information and synergy. In cases of strong synergistic interactions of two sources, conditional MI can thus exceed the bivariate MI. This can also be understood from the perspective of co-information, i.e. the net sum of redundancy and synergy. Another way of obtaining it is as the difference of bivariate and conditional MI. Strong synergistic contributions will then manifest in negative co-information, meaning that conditional MI has to exceed bivariate MI.

1.2.5 Practical considerations

From a pragmatic data analysis perspective, it is hard to name conclusions that can exclusively be drawn with information theoretic tools and not with other statistical approaches. One practical advantage however is that information theory unifies many statistical tests under the common effect size of the bit (Ince et al., 2017). It thus becomes easily possible to meaningfully compare results across different research questions where one would classically deal with relatively incomparable values of t, F, χ^2 or R^2.

In theory, information theoretic approaches are sensitive to statistical dependencies manifested in all possible nonlinear effects and in that sense generalise linear models. With finite datasets, this can however only be realised to a limited degree depending on the estimator that is used to describe probability distributions. In this thesis, we follow the Gaussian copula MI approach by Ince et al. (2017). Here, continuous variables are transformed into standard normal variables. The desired information theoretic measures can then be calculated from closed-form expressions. Especially for higher-order information theoretic quantities, this is computationally efficient, but only preserves rank-based relationships. A similarly pragmatic alternative for cases where there is an interest in nonlinear effects is to use binning with a relatively low number of bins to avoid a combinatorial explosion when considering joint probability distributions for higher order information theoretic quantities.

1.3 Encoding and decoding models

In simple terms, encoding and decoding models (Dayan & Abbott, 2001; Friston, 2009) describe regression models that either predict the responses from the stimuli (encoding) or vice versa (decoding). As such, they are thus suited to address questions of correspondence (Baker et al., 2021). A classic approach to implement such models follows a two-stage procedure (Naselaris et al., 2011). Here, a hypothesis is first specified in the form of a nonlinear function of the stimulus. Examples of such “linearising feature spaces" in both vision and audition are Gabor features (Kay et al., 2008; Santoro et al., 2014), semantic features (Huth et al., 2012, 2016) or deep neural network (DNN) activations (Eickenberg et al., 2017; Kell et al., 2018). Once this is done, the hypothesis can then be tested with a linear regression mapping from the linearising feature space to the responses or vice versa.

1.3.1 Cross-validation

Irrespective of the direction, such models should be fit to data within a framework called cross-validation (Mosier, 1951; Hastie et al., 2009). This relates to a split of the available data into disjoint training- and testing sets. It is crucial since models usually come with

parameters which are optimised to maximise the match of predictions and observed data. This optimisation can result in overfitting to noise in the data which the model is trained on, which will impair that model's performance when it is used to predict new data. Such generalisation to new data is however what the scientific significance of a model critically hinges on (Ioannidis, 2005) – if it only works within the local conditions of a single experiment, it is hard if not impossible for other science to make use of it. Cross-validation is thus essentially a simulation where one pretends that a part of the available data was recorded in a follow-up replication experiment (Yarkoni & Westfall, 2017). It directly incentivises strategies to mitigate overfitting, which can be measured as the difference between a model's performance in the training- and testing sets.

Such regularisation strategies usually consist of decreasing a model's flexibility and thus lead to a decrease of a model's performance on the training set. If this is applied adequately, this will prevent the model from fitting the noise in the training set and thereby contribute to an increased performance on new data such as the testing set. In linear regression, arguably one of the simplest and yet ubiquitous machine learning models, such regularisation can be implemented by a zero-mean prior on the weights (Hoerl & Kennard, 1970). An alternative view on this is that it reflects a penalty term for large weights that is added to the loss function given by the error between observations and predictions. This penalty term can itself be adjusted and thus constitutes a hyperparameter of the model. The adjustment of such hyperparameters requires an additional splitting of the training set into validation- and genuine training sets (Varoquaux et al., 2017). The hyperparameter can then be chosen such that it maximises the model's performance on the validation set, leaving the test set untouched for a final assessment of the model's performance.

1.3.2 The question of the direction

Before setting up such a cross-validation procedure, one faces a choice of the direction. Should one opt for an "encoding" or "forward" model of the process, or should one invert this direction in favour of a "decoding" or "backward" model?

Encoding models (Naselaris et al., 2011) follow the arrow of causality during perception, and thus constitute a simulation of the process of perception. They are a way to assemble and interrelate all that is known or assumed about a perceptual process within one mathematical object. This can then be subjected to tests of generalisation to both stimuli of the same distribution as the training data but also to stimuli of different distributions. An individual researcher, or optimally groups of researchers such as labs or entire fields can use them to systematically reason about this model and attempt to continually improve it, rendering it a powerful tool for basic neuroscience. Ideally, such an approach should thus help neuroscientific inquiry to move beyond an era of isolated individual experiments and develop into an integrative endeavour (Schrimpf et al., 2020).

An important distinction is that of functional and mechanistic models (Kay, 2018). An

arbitrary encoding model that merely succeeds in predicting responses is a functional model, while only the inclusion of aspects of the biological implementation will make it mechanistic. For models of neuronal responses, mechanistic models eventually have to make concrete hypotheses about the computational significance of the responses themselves: Are these directly involved in the processing of information, or are they mere “exhaust fumes" of such activity (Jonas & Kording, 2017)?

Decoding models (Norman et al., 2006; Hebart & Baker, 2018) on the other hand invert the arrow of causality and are thus not models of the process of perception as such. Their prime application are brain-computer interfaces, where they can be used to solve the engineering problem of providing information about brain states to external devices such as hearing aids (Mirkovic et al., 2015; Fiedler et al., 2017; Geirnaert et al., 2021). When applied to neuronal responses, a possible neuroscientific interpretation is that of modeling what information about the stimulus neurons downstream from the recorded population could in principle read out from the recorded activity.

Similarly to the distinction between functional and mechanistic encoding models, a problem of this interpretation is that it is hard to compare the way neuroimaging interfaces with neuronal activity to how actual neurons interface with a given neuronal population. In all likelihood, neuroimaging will miss the lion’s share of the neuronal activity that effectively communicates with other neurons. It is however also well possible that neuroimaging is sensitive to activity that is invisible to downstream neurons. For the interpretations of both encoding and decoding models, it is thus important to keep in mind that they mediate between stimuli and observations made by the experimenter (de-Wit et al., 2016).

From a practical perspective, they are both suited to quantify a statistical relationship between stimuli and responses (Friston, 2009; Holdgraf et al., 2017; Hebart et al., 2020). It is further possible to convert them into one another by multiplication with stimulus- or response covariance matrices (Haufe et al., 2014; van Vliet & Salmelin, 2020).

It is sometimes argued that encoding models are exclusively suited for “complete functional characterisations" of a certain neuronal response of interest (Naselaris et al., 2011). The idea is that this is achieved by an encoding model that reaches the “noise ceiling" of the responses (defined by e.g. the correlation of responses to repeatedly presented stimuli) to broadly sampled naturalistic stimuli. If a given decoding model on the other hand allowed the perfect reconstruction of a given linearising feature space of the presented stimuli, this would not exclude that other untested feature spaces could be reconstructed as well. This argument however has three problems: firstly, the noise ceiling is not trivial to estimate, and rests on assumptions such as the invariance of responses to repeated stimuli despite commonly known effects such as habituation. Secondly, if an inverse function could perfectly restore the entire original stimulus from the perfectly decoded linearising feature space, this would reach the same footing as the encoding model affording a “complete functional characterisation" – correlated feature spaces could exist that could achieve the same. Thirdly, and most importantly, the argument crit-

ically hinges on the completeness of the sampling of the stimulus material. Seemingly trivial “out-of-distribution" stimuli not included in the training set of the encoding model can however lead to failures to predict the correct response (Szegedy et al., 2014; Barbu et al., 2019).

Depending on the available data, a given direction can however have statistical advantages (Kriegeskorte & Douglas, 2018b; Hebart et al., 2020). Classically, encoding and decoding models are implemented as a multiple linear regression (given a linearising feature space, see above). This means that covariances in stimuli or responses can be exploited to improve the statistical power of the approach. In the case of one-dimensional behavioural responses to multidimensional stimuli that are correlated on multiple dimensions, an encoding model is the optimal choice. Decoding models can then only be implemented as mass-univariate regressions, where the stimulus covariance cannot be leveraged. In typical neuroimaging applications with multiple response channels, decoding models are generally more sensitive. They can be implemented in a “mass-multivariate" fashion, where for each stimulus dimension, noise correlations across response channels can be exploited. Mass-multivariate encoding models can leverage the covariance of multiple stimulus dimensions, but will ignore correlations across response channels.

Ideally, these advantages can be combined into a single approach. That is, one attempts to find a linear combination of stimulus channels that best predicts a linear combination of response channels. This can for example be implemented as a canonical correlation analysis (Friston, 2009; de Cheveigné et al., 2018). An alternative approach suggested in chapter 2 is to parameterise a biophysically motivated function that provides a linear combination of response channels (i.e. a spatial filter with the parameters of position in source space and response channel covariance regularisation) and then optimise its parameters as hyperparameters of an encoding model aiming to predict the output of the linear combination of response channels.

Information that can be decoded, but which no forward process model can account for essentially highlights a gap in understanding. In order to study a given dataset, it can thus be advantageous to implement both directions: An encoding model will help to shift the focus on simulating the forward process of perception, and a decoding model can then serve as both a further characterisation of the observed process as well as a test of the forward model with respect to this characterisation. This approach is central to chapters 2 and 4.

1.4 Audition and vision

In this thesis, the tools as described above are applied to examples from two sensory modalities, audition and vision. Specifically, the focus lies on acoustic speech signals (and more specifically, non-invasive electrophysiological responses to them) and visual face signals (and more specifically, behavioural responses to them). The following sec-

tion will provide brief reviews of these topics with the goal of putting the respective chapters into a historical context.

1.4.1 Non-invasive electrophysiology of speech tracking in superior temporal gyrus

Speech is an acoustically rich stimulus full of spectrotemporal interdependencies. Healthy humans produce and understand it seemingly effortlessly in order to communicate information from one mind to another.

Its reflection in human electrophysiological signals spans a history of multiple decades (Wöstmann et al., 2017). Shortly after the first description of the electroencephalogram (EEG, Berger, 1929), auditory evoked potentials in response to simple pure tones were discovered (Davis, 1939). More systematic studies of variations of such averages of responses to multiple repeated presentations of the same stimulus with respect to various acoustic manipulations followed later (Davis et al., 1966), as did the use of speech sounds as stimulus material in EEG studies (Feldman & Goldstein, 1967; Roth et al., 1970; Wood et al., 1971). The localisation of the neuronal generators of such auditory evoked responses to bilateral auditory cortices was already possible based on EEG recordings of this early period (Vaughan Jr & Ritter, 1970). With the advent of Magnetoencephalography (MEG, Cohen, 1968), refinements of substantially increased spatial precision became possible (Näätänen & Picton, 1987). The dominant experimental paradigm for both MEG and EEG (MEEG) studies in this latter half of the twentieth century however remained the controlled psycholinguistic experiment with its analytical workhorse, the event-related potential (EEG) or field (MEG). For this research paradigm, scientists used either isolated subword components (Dorman, 1974; Aaltonen et al., 1987; Näätänen et al., 1997; Obleser et al., 2003), words (Bentin et al., 1993) or connected speech segments (McCallum et al., 1984; Friederici et al., 1993; Gross et al., 1998) as stimulus material.

This event-related approach contributed to an enormous wealth of experimental findings and remains an indispensable tool for the study of MEEG signatures of speech perception until the present day (Khalighinejad et al., 2017; Daube et al., 2019b; Gwilliams et al., 2020). However, when applied to connected speech, it requires a discretisation of the inherently continuous speech signal, and thereby comes with limitations on the hypotheses about perceptual processes it can serve to study. The last two decades have thus seen the rise of approaches relating continuous features of stimuli – most prominently the time-varying energy or “envelope" – and MEEG responses (Ahissar et al., 2001).

A popular narrative of this approach is that of band-limited response components referred to as “oscillations" (Luo & Poeppel, 2007; Giraud & Poeppel, 2012; Peelle & Davis, 2012; Gross et al., 2013; Doelling et al., 2014). Accordingly, such rhythmic activity is thought to be present in auditory cortices in the absence of auditory stimulation, to then

be temporally “entrained" by input signals, affording a segmentation of the continuous input into syllable-like units (Ahissar et al., 2001; Lakatos et al., 2005; Hyafil et al., 2015).

A contemporaneous approach of continuous analyses remains more agnostic towards mechanistic accounts of the responses. Here, either univariate cross-correlation (Abrams et al., 2008; Hertrich et al., 2012; Ince et al., 2017) or, more recently, multivariate temporal response function approaches (Lalor & Foxe, 2010; Ding & Simon, 2012; O’Sullivan et al., 2015; Crosse et al., 2016; Brodbeck & Simon, 2020) are used to construct en- or decoding models relating MEEG responses to stimulus features. This has opened the door to richer (Di Liberto et al., 2015) and more elaborate computational accounts (Brodbeck et al., 2018b; Donhauser & Baillet, 2020; Heilbron et al., 2021) of MEEG responses to speech.

In light of such developments, chapter 2 calls for sustained attention to not only more complex, but also simpler models. Neuroscientists will rightly assume great undiscovered complexity in MEEG responses to speech. Accordingly complex models should however always be subjected to severe tests against less complex alternatives. This holds especially for experiments relying on uncontrolled naturalistic stimulation, and is of direct interest to translational opportunities such as the application to hearing aids. Here, the decoding of attention to a speaker amongst a multitude of sound sources (Ding & Simon, 2012; O’Sullivan et al., 2015; Brodbeck et al., 2020) is hoped to be exploitable in order to selectively amplify the signal of interest for the wearer of the hearing aids based on EEG electrodes placed e.g. in the ear canal (Fiedler et al., 2017; Geirnaert et al., 2021). Such devices are limited in both computing power and energy consumption, and will therefore benefit from incentives to balance complex accounts with simpler explanations wherever possible (Kubilius, 2018).

Chapter 3 then takes a more methodological perspective on the problem. The vantage point here is the observation of band-limited components in responses to speech (Ding & Simon, 2014). A computational function of such oscillations could be to represent a hypothesis of upcoming stimulus input by virtue of its continuous alignment to the merely “quasi-rhyhtmic" temporal stimulus structure (Lakatos et al., 2005; Giraud & Poeppel, 2012; Lakatos et al., 2019). If such an “entrainment" of response components takes place, then an interesting challenge is to separate response components reflecting the prediction based on past input from response components reflecting the alignment, i.e. reactions to aspects of the stimulus that were not anticipated. Methodologically, this falls into the purview of Granger-causal approaches (Granger, 1969), or their information theoretic generalisation, transfer entropy (TE, itself classically implemented as conditional MI). Here, parts of the responses that are predictable from their own past are attempted to be omitted when estimating the stimulus-response relationship. Such approaches however are generally problematic when applied to band-limited signals (Florin et al., 2010; Barnett & Seth, 2011). The chapter aims to overcome these problems, test the proposal on simple simulations and explore the results it suggests when applied to real data from chapter 2.

Looking ahead, computational modeling of language and speech, methodological developments for mapping models to responses, medical applications and last but not least new recording techniques such as optically pumped magnetometers (Boto et al., 2018; De Lange et al., 2021) promise to keep the field exciting.

1.4.2 Visual perception of faces

As speech, faces are an important social stimulus that humans, unless they are living in isolation and without access to reflective surfaces or cameras, are confronted with on a daily basis. In light of the intensive exposure to and the high relevance of faces (Gauthier et al., 1999; Jack & Schyns, 2017), it is not surprising that entire brain areas have been ascribed the main purpose of processing this class of visual stimuli (Sergent et al., 1992; Kanwisher et al., 1997; Grill-Spector et al., 2004). In order to categorise faces on continua such as familiarity, age or sex, humans tend to be highly skilled in processing faces. Except for face-blind or “prosopagnostic" individuals (Bodamer, 1947), humans can detect even small differences in such high-dimensional visual objects with high accuracy, such that it has even been suggested to use parameterisations of them for data visualisation (Chernoff, 1973). It is for a similar reason that faces are an interesting class of stimuli for vision sciences: They are an example of a category for which it is comparably easy to construct generative models of stimuli in high dimensional pixel space by varying relatively few underlying dimensions.

To develop the significance of this, we will first consider a brief historic overview over a line of research that is concerned with extracting mental representations from experimental participants. In general, such experiments follow the idea of the encoding model as described above. Importantly however, stimuli are sampled from random distributions in order to characterise response biases.

These experiments go back to ideas of Wiener (1958), who had postulated that in order to characterise an unknown nonlinear system just by relating its inputs to its outputs, Gaussian white noise inputs would be optimally suited. The suggested procedure becomes intractable to characterise higher-order nonlinearities (Marmarelis & Naka, 1972; Franz & Schölkopf, 2005), but started a tradition in electrophysiology where impulse responses of single neurons to temporally decorrelated white noise stimuli were estimated with linear cross-correlation. This was originally referred to as “triggered correlation" (De Boer & Kuyper, 1968), since only performing the computation of the correlation at the onset of sparsely occurring spikes was more efficient than computing it over the whole response vector. Later, this was referred to as “reverse correlation", since in this procedure, one would go back in time to look up stimulus segments preceding the spike (Jones & Palmer, 1987; Dayan & Abbott, 2001; Ringach & Shapley, 2004).

Applying this principle to human behavioural experiments had its origins in the auditory domain (Ahumada Jr & Lovell, 1971), where it was implemented as a multiple regression encoding model. The application in vision experiments followed later (Abel

& Quick Jr, 1978; Ahumada Jr, 1996). Here it was mainly referred to as “classification images" (Murray, 2011), classically consisting of a structuring signal (e.g. a neutral face) on which pixel noise is overlaid. The classification image is then a visualisation of the statistical relationship between the stimuli at each stimulus dimension (e.g. pixel location) and the responses. Ignoring nuisance effects such as fatigue, the responses follow the task instructions of the experiment, which usually ask participants to rate the similarity of the stimuli to a category of interest. The assumption then is that the noise in stimuli that cause a given response contains patterns that match the participant’s personal mental representation of the concept of interest to some degree. In this way, the final classification image reflects an estimate of such mental representations. It is for example possible to present entirely unstructured pixel noise to participants, falsely inform them that on half of the trials the letter “S" will be shown which they are to detect, and obtain weights of encoding models that show the respective letter (Gosselin & Schyns, 2003). Importantly, details of the reconstructed letter such as its font are uniquely defined by the participant’s personal interpretation of the category that is abstractly defined in the task instructions. In principle, this approach is applicable to arbitrary categories, but has a strong tradition for face information such as the emotional expression, identity or ethnicity (Gosselin & Schyns, 2003; Mangini & Biederman, 2004; Dotsch et al., 2008). Within this tradition, the term “reverse correlation" is used for experiments characterising response biases from noise stimuli (Gosselin & Schyns, 2003), although an analysis of the eponymous temporal dimension is usually not considered.

An important element of these studies is the format which the noise is rendered in. Since a direct mapping of samples from a random distribution to pixel intensities of image stimuli can be seen as the simplest form of a generative model of visual stimuli, this problem brings us back to the beginning of the section. Pixel noise allows a high degree of freedom with respect to what can appear in the final classification image. However, it requires a substantial number of trials in order to reveal effects that surpass noise thresholds, since the perceived similarity of a given category of interest and a given pixel noise image is restricted to low levels. In analogy to this, the field of electrophysiology had found neurons that would only respond weakly to white noise stimulation, but vividly to naturalistic stimuli with more complex statistics (Rieke et al., 1995; Theunissen et al., 2000). When the experimenter has a strong prior about the object class of interest, it is possible to effectively leverage this prior in terms of a generative model that imposes a corresponding structure on samples from a random distribution (Chomsky, 1965; Grenander, 1994; Olman & Kersten, 2004; Jack & Schyns, 2017). This will constrain the kind of classification image (or “classification object", Olman & Kersten, 2004) that can be obtained: with a face prior, the reconstruction will always be a face as defined by a generative model of faces. To the degree to which such a prior is sensible, it will enhance the reconstruction of a mental representation by narrowing the sampling space for the experiment to fewer dimensions of higher relevance. As a result, more realistic reconstructions can be obtained, for example of dynamic emotional expressions (Jack

et al., 2012) or of the participant-specific memory of a given familiar person (Zhan et al., 2019a). Both of these examples rely on a generative model that renders pixel images of faces as a function of 3D shape and RGB texture parameters as well as viewing and lighting angles. An important corollary is that with such models of the latent causes of variance in pixels of images, it becomes possible to attribute variance in behavioural responses to the hypothesised causes of the images instead of only their pixels (Olman & Kersten, 2004). This becomes important when for example considering how entirely different pixel regions of an image carry the information of the same underlying face shape at different viewing angles. Attributing a system's outputs to such pixel regions in one viewing angle is of little help when trying to infer the output-relevant regions of an image of another viewing angle.

Chapter 4 re-examines the dataset recorded by Zhan et al. (2019a) mentioned above. In the chapter, a framework is developed to apply the vision scientific approach as described above to the question of forward encoding models of human vision. These models take on the daunting task of recreating central aspects of the mapping from the human retina to the latent mental spaces that eventually afford behaviour. Deep neural network models (DNNs, Fukushima, 1980; LeCun et al., 2015) are an interesting candidate, since they have been shown to solve end-to-end engineering challenges of computer vision with unprecedented performance scores. The central idea of the chapter is to subject various DNNs to the same controlled "face noise" that human participants had seen and rated, so that it becomes possible to compare DNNs and humans with respect to the same experimentally controlled causal structure of the stimuli. In this sense, it is attempting to move beyond the dominant approach in the current literature which evaluates encoding models of human vision (such as DNNs) by only seeking to establish a high prediction performance between uncontrolled inputs and outputs, and thus has no classical vision scientific grasp on what it is that the models are predicting outputs with.

Chapter 2

Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech

published as:
Daube, C., Ince, R.A.A. and Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology*, 29(12): 1924–1937

Author contributions:
C.D., R.A.A.I., and J.G. conceived of and designed the experiment.
C.D. collected and analyzed the data.
C.D. and R.A.A.I. contributed analytic tools.
C.D., R.A.A.I., and J.G. wrote and edited the manuscript.
J.G. acquired the financial support for the project leading to this manuscript.

2.1 Abstract

When we listen to speech, we have to make sense of a waveform of sound pressure. Hierarchical models of speech perception assume that, to extract semantic meaning, the signal is transformed into unknown, intermediate neuronal representations. Traditionally, studies of such intermediate representations are guided by linguistically defined concepts, such as phonemes. Here, we argue that in order to arrive at an unbiased understanding of the neuronal responses to speech, we should focus instead on representations obtained directly from the stimulus. We illustrate our view with a data-driven, information theoretic analysis of a dataset of 24 young, healthy humans who listened to a one-hour narrative while their magnetoencephalogram (MEG) was recorded. We find that two recent results, the improved performance of an encoding model in which annotated linguistic and acoustic features were combined, and the decoding of phoneme subgroups from phoneme-locked responses, can be explained by an encoding model that is based entirely on acoustic features. These acoustic features capitalise on acoustic edges and outperform Gabor-filtered spectrograms, which can explicitly describe the spectrotemporal characteristics of individual phonemes. By replicating our results in publicly available electroencephalography (EEG) data, we conclude that models of brain responses based on linguistic features can serve as excellent benchmarks. However, we believe that in order to further our understanding of human cortical responses to speech, we should also explore low-level and parsimonious explanations for apparent high-level phenomena.

2.2 Introduction

Speech perception is often conceptualised as a hierarchical process (Pisoni & Luce, 1987; DeWitt & Rauschecker, 2012). The human brain is assumed to extract semantic meaning from a highly dynamic sound pressure signal via a cascade of transformations that create increasingly abstract representations of speech. It is well established that perceived speech sounds are first decomposed into a spectrally resolved representation at the cochlea. Various structures along the subcortical auditory pathway are believed to then undertake further processing steps (Verhulst et al., 2018; Sitek et al., 2019). However, considerable uncertainty remains about exactly how sound is represented in the auditory cortex (Młynarski & McDermott, 2018).

One way to gain further insight into human speech processing is to employ encoding models. These models aim to predict the time-series of recorded neural data from the waveform of the presented stimulus. A popular framework for encoding models organises this in two steps (Naselaris et al., 2011; Holdgraf et al., 2017). In the first step, the stimulus material undergoes nonlinear transformations into various sets or spaces of features. These features capture hypotheses about the cortical computations that are performed on the input signal. In the second step, a linear mapping of these feature

spaces onto the neuronal responses is obtained to evaluate the utilised hypotheses in terms of out-of-sample linear prediction performance. In this way, data-rich, naturalistic listening conditions of a relatively long duration can be exploited, considerably improving a model's validity over isolated and artificial experimental paradigms (Theunissen & Elie, 2014; Hamilton & Huth, 2018). Recent results demonstrate the applicability of this approach across various neuroimaging modalities and research questions (Di Liberto et al., 2015; Huth et al., 2016; de Heer et al., 2017; Berezutskaya et al., 2017; Forte et al., 2017; Maddox & Lee, 2018; Kell et al., 2018; Brodbeck et al., 2018b,a; Biesmans et al., 2017; Broderick et al., 2018b).

A compelling finding obtained with this approach is that predictions of cortical responses as measured by EEG (Di Liberto et al., 2015) or functional magnetic resonance imaging (fMRI, de Heer et al., 2017) using acoustic feature spaces can be improved by additionally considering so-called articulatory feature spaces. The latter originate from the linguistic concept of representing a language with a set of minimal contrastive units, called phonemes. However, superior temporal regions are known to selectively respond to subgroups of phonemes rather than to individual phonemes (Mesgarani et al., 2014). Therefore, the full phoneme set is usually reduced by mapping each phoneme to its corresponding vocal gestures ("articulatory features"), such as the voicing, tongue position or place and manner of articulation. Recently, it was shown that these manners of articulation can also be decoded from EEG data time-locked to phoneme onsets in continuous speech stimuli (Khalighinejad et al., 2017). Encoding and decoding analyses based on articulatory feature spaces are thus interpreted as concordantly capturing a faculty called "pre-lexical abstraction" (Obleser & Eisner, 2008), i.e., a transformation of continuous physical properties of the waveform to speech-specific, categorical and invariant units of perception.

However, the transformation of speech stimuli into articulatory features comes with certain critical caveats. Most importantly, this representational format of speech is based on concepts that humans have agreed on to talk about language. And while a match of such linguistic constructs with physiological responses is conceivable, it is a potentially biased and specific hypothesis with a range of alternatives (Pisoni & Luce, 1987; Hasson et al., 2018; Massaro, 1974; Lotto & Holt, 2000). Moreover, the partly arbitrary mapping of phonemes to articulatory features provides a low degree of computational specification. As such, models that use articulatory features could be considered to be so-called 'oracle models', which rely on information that is not available to the individual's brain being modelled (Kriegeskorte & Douglas, 2018a).

Additionally, current implementations of this transformation rely on a semi-automated, forced alignment of a textual transcription to the sound wave of the stimulus material. While such alignment methods incorporate a high degree of computational sophistication, the task they solve is not a good model of the task that the listening brain faces. This compromises the usefulness of the intermediate representations generated by such alignments to serve as candidate features to predict brain responses, such that usually

only the final output is used. It thus remains unclear whether the level of complexity implied by the final articulatory features is actually necessary.

These caveats thus raise an important question. Can the gain in prediction performance that is reportedly provided by articulatory features be explained by alternative features that are based on computationally more specified, physiologically plausible and possibly less complex transformations of stimulus acoustics? The extent to which this were the case would indicate how much of the predictive gain that is provided by articulatory features is attributable to the generic, feed-forward processing of an acoustic stimulus that is not specific to speech processing. When choosing such acoustic feature spaces, one can proceed in different directions. One possibility is that in order to explain the same variance as models based on articulatory features, the characteristic spectrotemporal patterns that define the phoneme subgroups are needed. Correspondingly, one could extract such abstract information from the spectrogram with suitable filters. A physiologically inspired candidate feature space is the Gabor-filtered spectrogram, which interestingly improves the performance of automatic speech recognition (ASR) software when used as input features (Schädler et al., 2012). With this generic class of spectrotemporal kernels, one can describe several acoustic patterns that dissociate groups of phonemes. Examples include the spectral distance between formants, as captured by filters of different spectral modulation, and formant transitions, as captured by filters of joint spectrotemporal modulation. Although this feature space is long established in encoding and decoding models of the human and animal midbrain and auditory cortices (Holdgraf et al., 2017; Berezutskaya et al., 2017; Qiu et al., 2003; Pasley et al., 2012; Santoro et al., 2014, 2017; Norman-Haignere & McDermott, 2018; Schönwiesner & Zatorre, 2009), it has yet to be applied to magneto- and electroencephalography (MEEG) data.

Another possibility is that the performance boost provided by articulatory features is instead attributable to their correlation with simpler acoustic properties. It has repeatedly been observed that neuronal responses from bilateral superior temporal regions are particularly sensitive to acoustic edges (Prendergast et al., 2010; Hertrich et al., 2012; Gross et al., 2013; Doelling et al., 2014; Hamilton et al., 2018; Oganian & Chang, 2019). Features that extract these onsets from envelope representations via a half-wave rectification of the temporal gradient of time- varying energy have been used in several studies (Hertrich et al., 2012; Hambrook & Tata, 2014; Fiedler et al., 2018). Features that rely on the temporal gradient also capture the relationship of neighbouring time points, which contain information present in MEEG data across a range of different analyses (Ince et al., 2017). It is thus interesting to assess the degree to which the gain in prediction performance that is provided by articulatory features can be explained by such onset features.

In this study, we examined these two possible explanations by comparing the predictive power of different acoustic feature spaces to that of an annotated articulatory feature space. We performed these investigations on an MEG story dataset of one hour

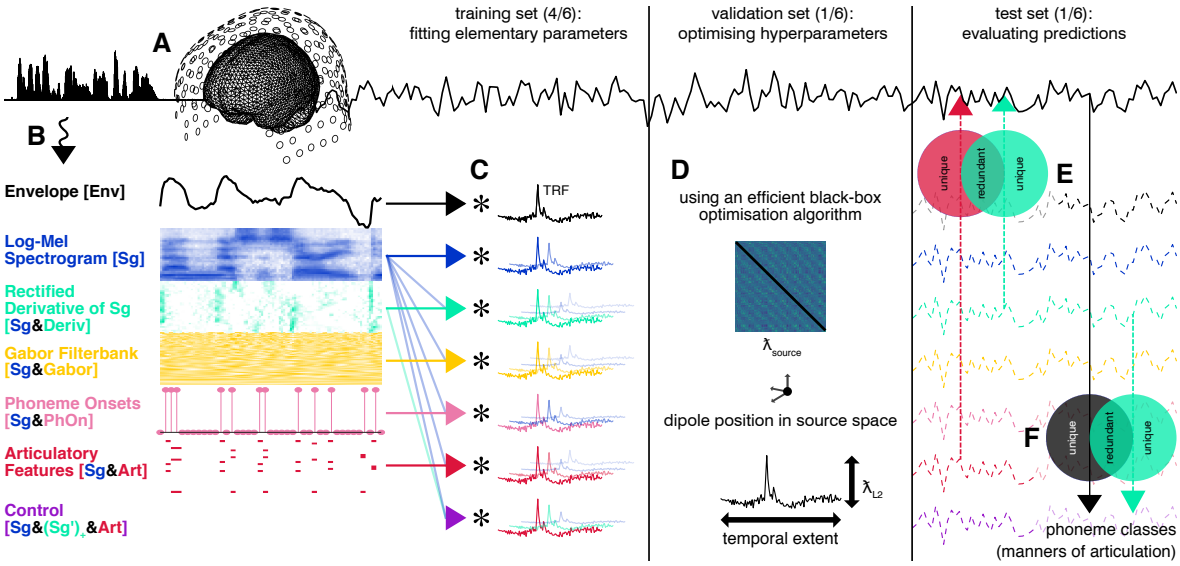

Figure 2.1: **Study concept and design.**
A Magnetoencephalography (MEG) data were recorded while participants ($n = 24$) listened to a story of 1 hour duration. **B** The speech waveform was then nonlinearly transformed into various feature spaces. **C** These feature spaces were used to predict neuronal responses using (multivariate) temporal response functions ((m)TRFs) in a nested cross-validation framework. The majority of the data were used to fit the (m)TRFs. **D** Hyper-parameters controlling the (m)TRFs (separately for each feature (sub-)space, hemisphere and participant: temporal extent and $L2$ regularisation) and the MEG source reconstruction (sensor covariance matrix regularisation and position of dipoles in source space) were optimised on separate validation data. **E** The predicted responses of the encoding model (dashed lines) were evaluated on unseen test data by asking to which degree a benchmark feature space that relied on articulatory features was redundant with competing, acoustic feature spaces using partial information decomposition (PID). **F** Additionally, four classes of phonemes were decoded from phoneme-locked observed and predicted MEG responses. PID was used to determine to which degree the predictions of the encoding models contained the same information about phoneme classes as the observed data.

duration per participant in a rigorous data-driven approach (see figure 2.1). A nested-cross validation framework (Varoquaux et al., 2017) was used to delegate the choice of model settings to a recent optimisation algorithm (Acerbi & Ma, 2017). We thus allowed encoding models based on different feature spaces the same chances to find optimal parameter combinations with a minimum of a-priori information, while minimising the risk of overfitting. We then applied partial information decomposition (PID, Ince, 2017a) to assess the degree to which the predictions of acoustic-feature-based models shared information about observed recordings with those of articulatory feature-based models, and to assess the degree to which these feature spaces contained unique predictive information. This flexible theoretic framework also allowed us to quantify to what extent the information about manners of articulation decodable from phoneme-evoked responses could be accounted for by the predictions of our encoding models. Lastly, since MEG and EEG data can reflect different neuronal processes (Destoky et al., 2019; Cohen & Cuffin, 1983), we performed similar analyses on a publicly available EEG story-listening

dataset (Broderick et al., 2018b). Using this approach, we found that apparent encoding and decoding signatures of high-level pre-lexical abstraction could be explained with simple low-level acoustic models.

2.3 Results

2.3.1 Speech tracking in bilateral auditory cortices

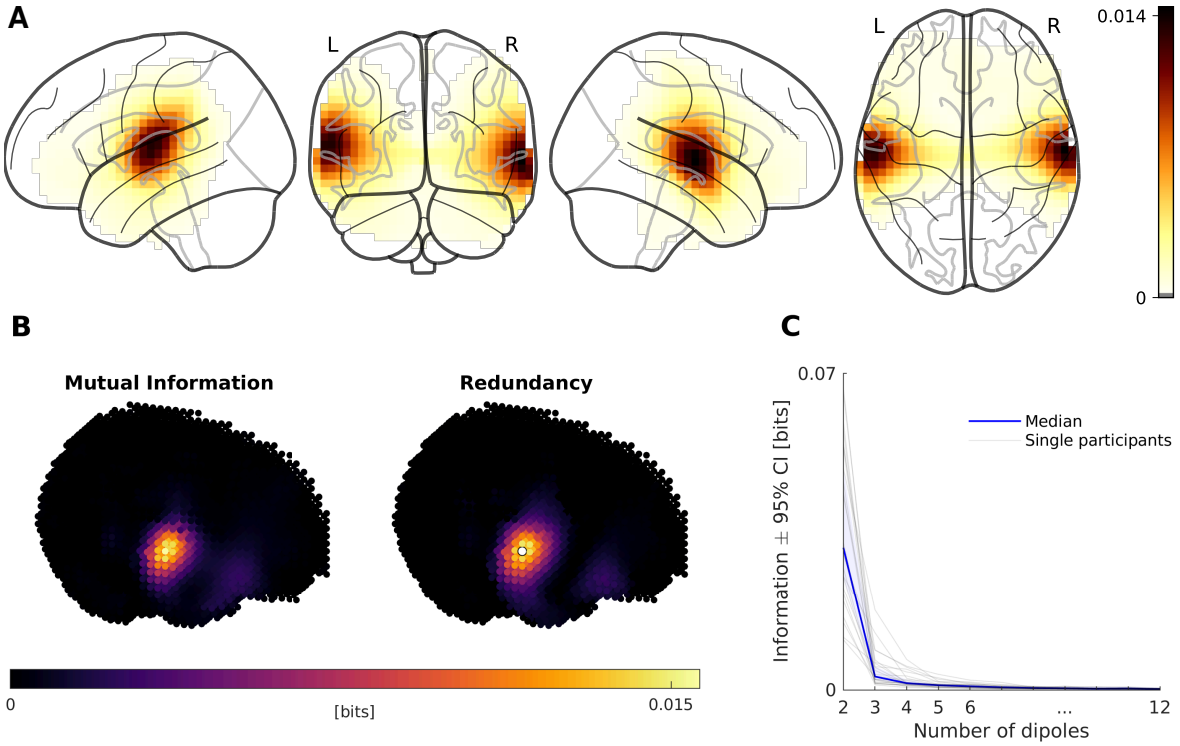

Figure 2.2: **Identification and characterisation of story-responsive regions in source space.**
A Grand average story responsivity (variance of source-reconstructed brain activity recorded during first presentation explained by activity recorded during second presentation of the last block). Each image shows different viewing angles on the same data. **B** (Left) Story responsivity using mutual information (MI). Plot shows MI of activity in the first repetition of the last block about activity in the second repetition of the last block. (Right) Shared information (redundancy) of activity at bilateral story-responsivity peaks in the first repetition and activity in the first repetition at each other grid point about activity at these other grid points in the second repetition. See video S1 for further explanation. Data from one exemplary participant are shown. **C** Unique information added by sources additional to the bilateral story-responsivity peaks. See also figure 2.3.

First, we characterised where in MEG source space we could find robust responses related to speech processing and also the spatial resolution that these responses could be studied at. To identify regions in source space where MEG responses were repeatably activated by the stimulus (“story responsive" regions, Honey et al., 2012; de Heer et al., 2017), we correlated source-localised, full-brain responses to one chapter of the story

to the responses to its repeated presentation. These correlations peaked in regions that agree with the typical localisation of the bilateral auditory cortices (ACs, figure 2.2A).

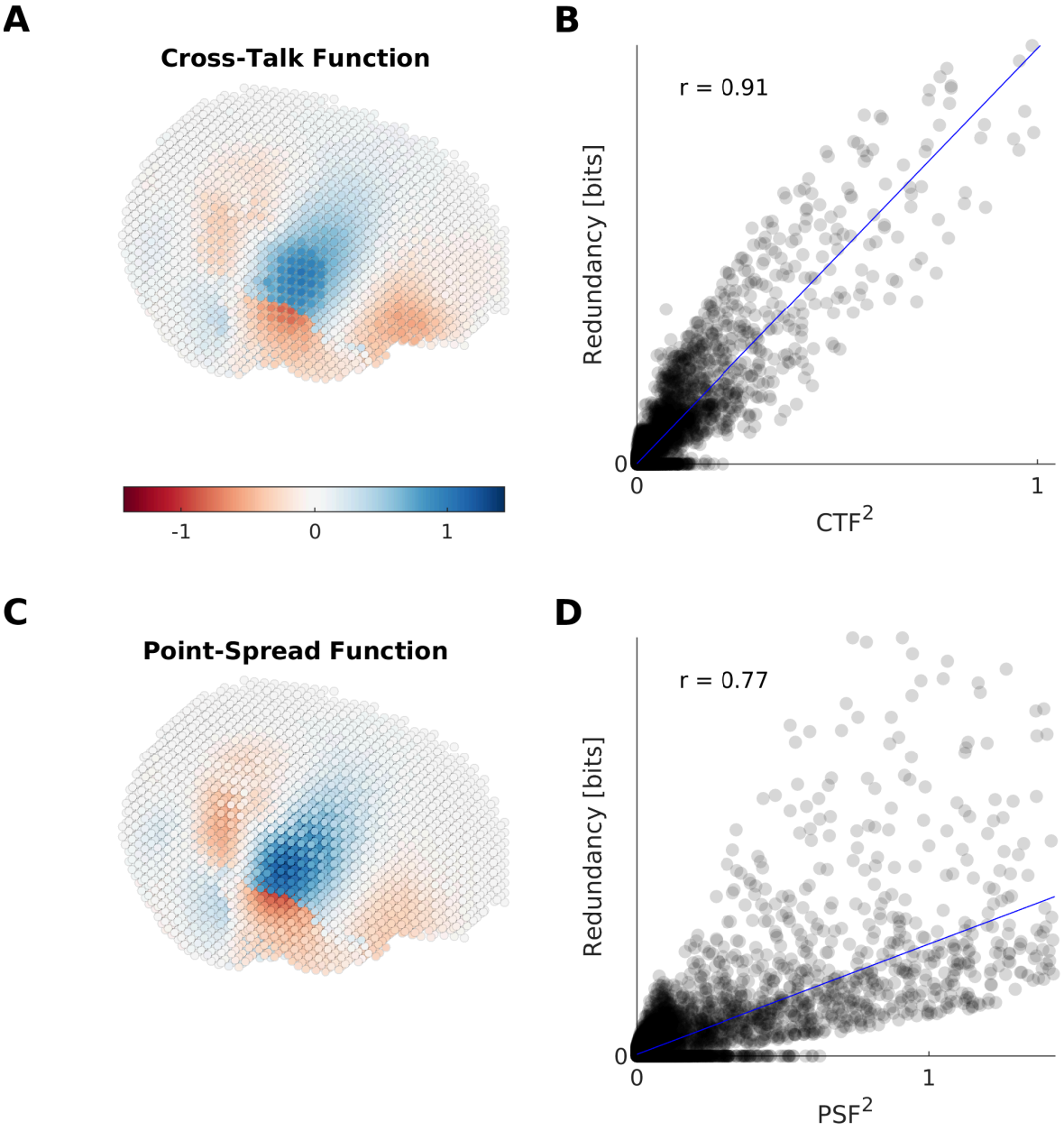

Figure 2.3: Caption on following page.

Instead of falling off sharply, the story responsivity decreased gradually with increasing distance from these peaks. However, we expected that querying activity from different locations within these story-responsive regions would yield highly similar (i.e. redundant) time series since the spatial resolution of MEG inverse solutions is inherently limited (Faharibozorg et al., 2018). To avoid an unnecessary computational burden for the later modelling, we therefore explored how much of the repeatable activity we could explain with dipoles at the two bilateral story responsivity peaks, and also how much we could explain by considering further dipoles at different locations. To do so, we implemented an iterative information theoretic approach based on PID (see video S1 and methods 2.5.2 for a detailed description). This approach revealed that indeed one source per hemisphere could account for most of the spatial spread of the story responsivity. The

Figure 2.3 *(previous page)*: **Redundancy is related to Cross-Talk and Point-Spread Functions (related to figure 2.2).**
A Cross-Talk Function in one exemplary participant (same as in figure 2B). It shows how activity at different grid points leaks into estimates of activity queried at grid point of interest, here at a right AC dipole. **B** Relationship of squared Cross-Talk Function (CTF) and Redundancy in exemplary participant. Correlation in top left reports Pearson correlation. The mean of this correlation (Fisher-Z transformed before averaging and retransformed after averaging) was 0.75 (right AC, range: $[0.26, 0.91]$) and 0.68 (left AC, range: $[-0.09, 0.94]$). **C** Point-Spread Function in one exemplary participant. It shows how activity at grid point of interest, here at a right AC dipole, leaks into estimates of activity queried at other positions in source space. **D** Relationship of squared Point-Spread Function (CTF) and Redundancy in exemplary participant. Correlation in top left reports Pearson correlation. The mean of this correlation (Fisher-Z transformed before averaging and retransformed after averaging) was 0.60 (right AC, range: $[0.16, 0.83]$) and 0.51 (left AC, range: $[-.10, .87]$).

individual maps of story-responsivity correlated highly with maps of redundancy (average Pearson correlation 0.89, range: $[0.80, 0.97]$, figure 2.2B). As such, the information that activity at additional grid points carried about the activity recorded during the second presentation of the same chapter largely overlapped with the information that could be obtained from activity at the bilateral peaks. Correspondingly, the amount of information contributed by sources additional to the bilateral peaks fell off in a characteristic L-shaped curve (figure 2.2D). This was largely attributable to measures of leakage of the spatial filters, such as their cross-talk and point-spread functions (see figure 2.3 for details).

Based on these results, we subsequently analysed one source location per hemisphere, since this single location could capture the repeatable signal that stems from the bilateral ACs. Note, however, that in the following modelling, the exact location of these two sources was not fixed, but instead was optimised independently for each tested feature space.

2.3.2 Predictive power of feature spaces

The main goal of this study was to compare the cross-validated performance of linear models that were trained to predict relevant parts of the MEG responses from different sets or "spaces" of features extracted from the speech stimulus. The central question we investigated was: to what degree can purely acoustic feature spaces achieve the performance of a benchmark feature space (namely, spectrograms and annotated articulatory features combined (Di Liberto et al., 2015)? Crucially, our modelling approach ensured that the settings of our models ("hyper-parameters") could flexibly adapt to each different feature (sub-)space, individual participant, and to each hemisphere (see methods for a detailed description). The hyper-parameters operated on the predictors, the model, and also the MEG responses such that, for example, the exact position of the dipole in source

space was optimised for each feature space (see table 2.1 for an overview over all feature spaces used in this study). This gave each feature space the same chances to optimally predict the MEG responses within our bilateral sources linear modelling framework. The performances of our models exhibited relatively large inter-participant variability and comparatively low variability across feature spaces (figure 2.4A).

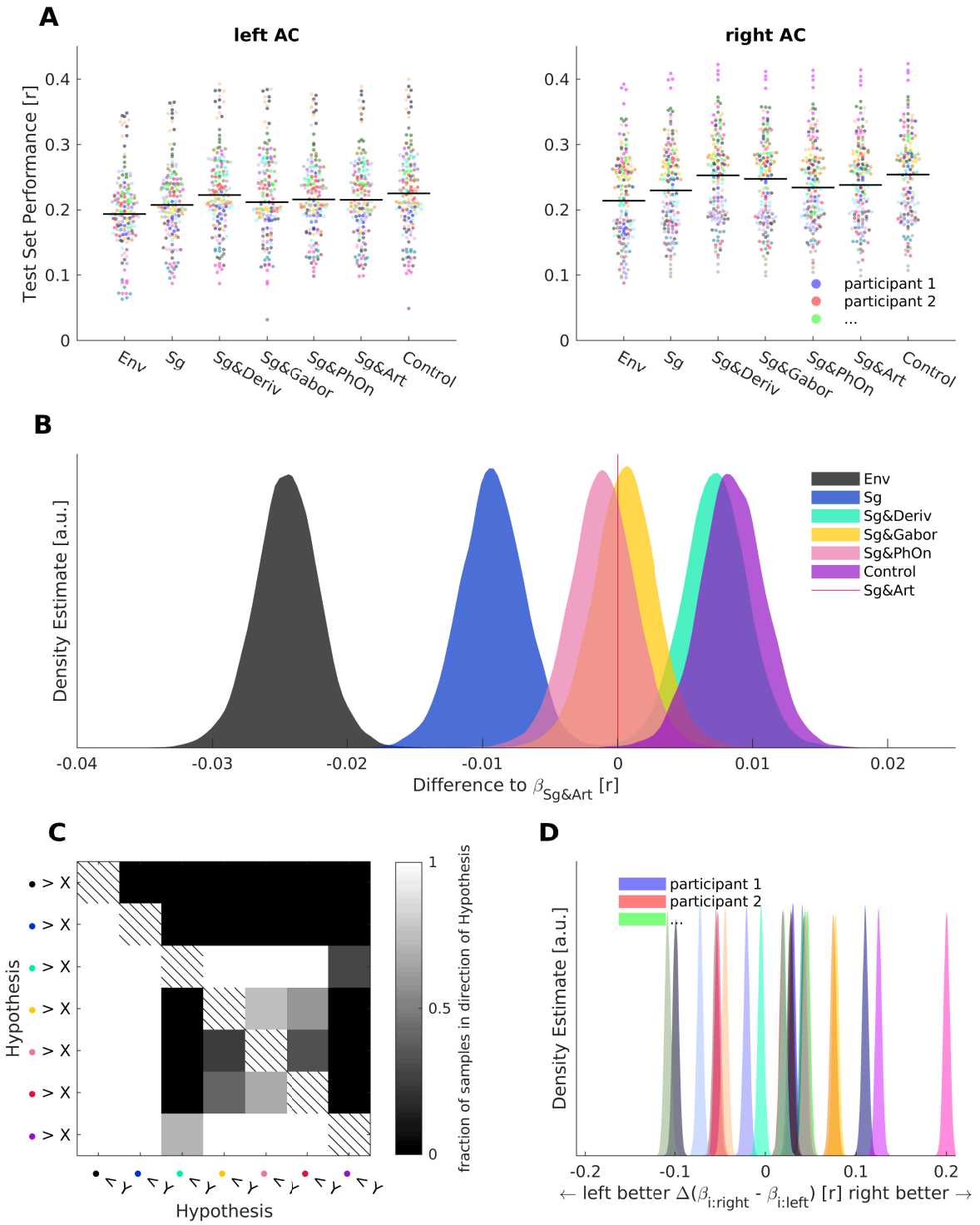

Figure 2.4: Caption on following page.

To focus on the systematic differences across the feature spaces, we used Bayesian hierarchical linear modelling (Bürkner, 2017) and separated the overall effects of different feature spaces from effects attributable to participants, hemispheres and cross-

Figure 2.4 *(previous page)*: **Evaluating the performance of different feature spaces.** **A** Raw test set performances in the left and right auditory cortex (AC) for models based on different feature spaces shown on the horizontal axis. See table 2.1 for an explanation of the feature spaces and their shorthand notations. Each colour codes for a single participant ($n = 24$), each dot is one test set. Pooled medians are indicated with black lines. **B** Samples from the posterior distribution of differences of beta estimates (competing feature spaces minus benchmark *Sg&Art* feature space, results left of the red line thus reflect that the *Sg&Art* feature space has a higher performance, results right of the red line indicate that the competing feature space has a higher performance). Feature spaces are colour coded as indicated. **C** Percent of samples in favour of hypotheses of differences of beta estimates between all feature spaces. Hypotheses are colour coded using the same colour mapping as in B, which corresponds to the bottom row and right column of the matrix shown here. **D** Samples from posterior distribution of differences of beta estimates of individual participants' right ACs minus left ACs. Colour mapping in D is the same as in A. See also Figures 2.5 and 2.6.

validation folds. We extracted the samples of the posterior distributions of the regression coefficients ("betas") of interest. We then subtracted the samples that referred to the benchmark feature space from those referring to the other competing feature spaces. From the resulting posteriors of differences (figure 2.4B), we could determine the fraction of samples above or below zero, i.e. in the direction of the corresponding hypotheses (f_{h_1}). We repeated this for all other possible comparisons between the feature spaces (figure 2.4C).

Initially, we were interested in whether we could replicate the previously reported increase in prediction performance when combining linguistically motivated articulatory features with spectrograms (*Sg&Art*, red vertical line, figure 2.4B) over spectrograms alone (*Sg*, blue) in our data. Indeed, we found a large fraction of samples of the posterior of differences in favour of a successful replication (mean of the difference in Pearson correlation $\Delta = 0.0093$, $f_{h_1} = 0.9994$). This allowed us to test whether various alternative feature spaces could achieve a similar gain in performance in order to investigate the origin of the improved prediction achieved using articulatory features.

We first investigated spectrotemporal Gabor patterns, which can be used to dissociate several phonemic groups (Schädler et al., 2012), because the articulatory feature space might have benefitted from describing responses that are specific to phoneme subgroups. In combination with the spectrogram, which directly accounted for time varying sound energy, this feature space (*Sg&Gb*, yellow) achieved a comparable gain in prediction performance over the spectrograms alone ($\Delta = 0.0098$, $f_{h_1} = 0.9994$). Its performance was on par with the benchmark feature space (*Sg&Art*), i.e. it was only negligibly better ($\Delta = 0.0006$, $f_{h_1} = 0.5960$). This feature space thus achieved a similar performance to that of the linguistically motivated feature space but did so without requiring linguistic concepts. Instead, it was physiologically motivated and computationally fully specified.

Shorthand	Name	Dimensionality	Description
Env	envelope	1	sum across channels of *Sg*
Sg	log-mel spectrogram	31	spectral decomposition of time-varying stimulus energy in 31 mel-spaced bands with logarithmic compressive nonlinearity
Sg&(Sg')$_+$	combination of *Sg* and half-wave rectified temporal derivatives of individual *Sg* channels	$31+31$	*Sg* and positive temporal rate of change of power in each channel of *Sg*
Sg&Gb	combination of *Sg* and Gabor-filtered *Sg*	$31+455$	*Sg* and decomposition of *Sg* according to spectral, temporal and joint spectrotemporal modulations
Sg&PhOn	combination of *Sg* and annotated phoneme onsets	$31+1$	*Sg* and unit impulses at the beginning of each annotated phoneme
Sg&Art	combination of *Sg* and articulatory features of each phoneme, “benchmark features"	$31+23$	*Sg* and 23 channels with unit impulses at the beginning of each phoneme characterised by the corresponding vocal gesture
Sg&(Sg')$_+$*&Art*	combination of *Sg&(Sg')*$_+$ and articulatory features of each phoneme	$31+31+23$	Control combination

Table 2.1: **Feature spaces.**

However, we also wanted to explore simpler models to determine the level of complexity that would be required to optimise prediction. Sound onsets offer a promising candidate for a neurally relevant, low-dimensional auditory feature (Hertrich et al., 2012; Hamilton et al., 2018; Oganian & Chang, 2019; Ince et al., 2017). As a first test of this hypothesis, we reduced the articulatory features to phoneme onsets (*Sg&PhOn*, pink). This model outperformed the spectrograms in a similar way to *Sg&Art* ($\Delta = 0.0081$, $f_{h_1} = 0.9983$), indicating that the performance increase obtained with articulatory features originates from the timings of the phoneme onsets, and not the identity of different phoneme subgroups.

The phoneme onsets were, however, still an abstracted representation of the stimulus resulting from transcription alignment, with an unclear relation to the original acoustics. One way to derive a signal representing sound onsets directly from speech acoustics is by half-wave rectification of the first derivative of the time-varying stimulus energy (Hertrich et al., 2012). This quantifies positive rates of change, i.e. increases in the stimulus amplitude. We found that spectrally resolving this energy using spectrograms (*Sg*, blue)

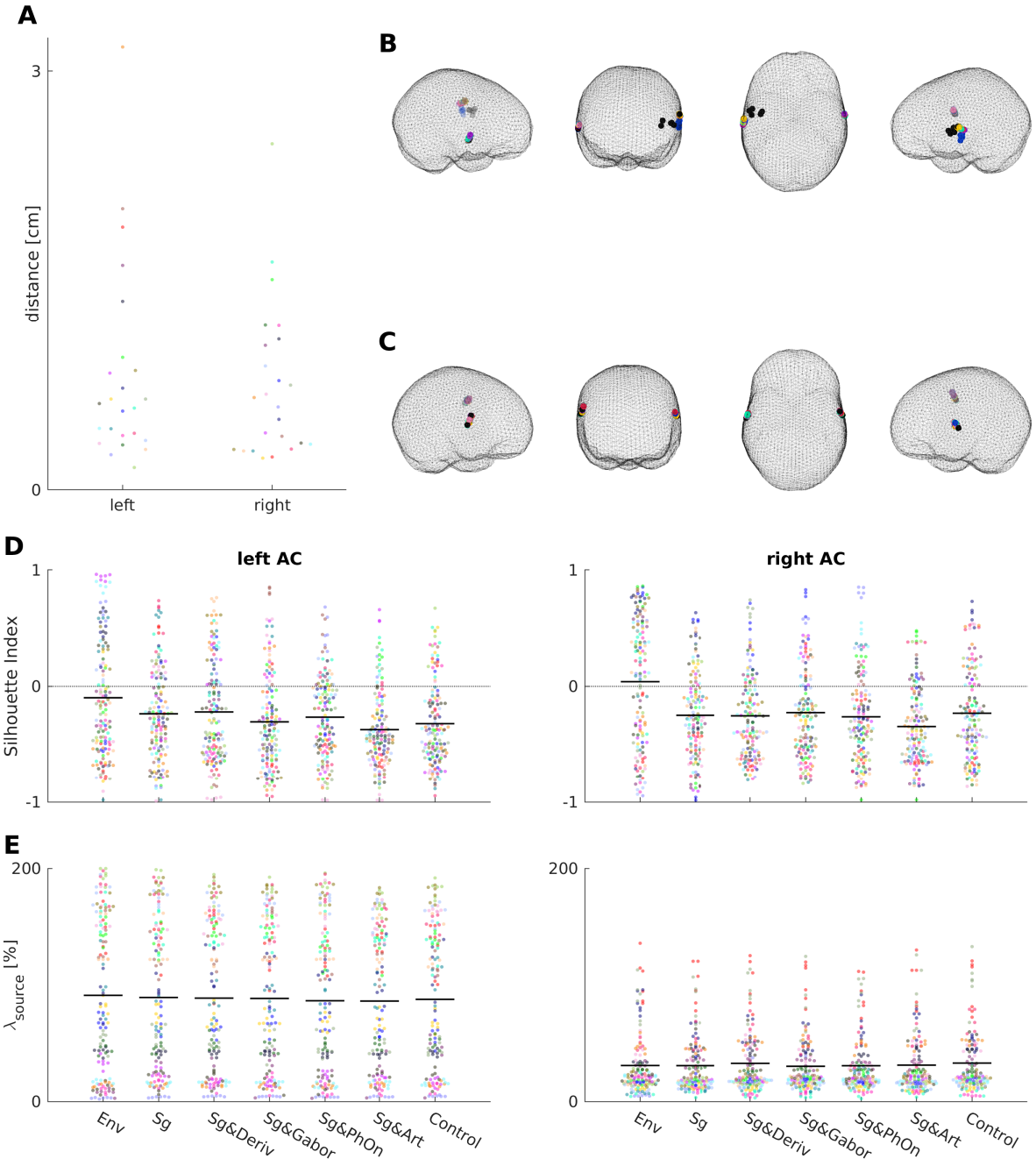

Figure 2.5: Caption on following page.

outperformed the broadband envelope (*Env*, black, $\Delta = 0.0152$, $f_{h_1} = 1$). We therefore computed the positive rate of change of energy of the individual channels of the spectrogram. Combined with the spectrogram features, this model (*Sg&(Sg')*$_+$, turquoise, (figure 2.4B) outperformed the benchmark feature space ($\Delta = 0.0073$, $f_{h_1} = 0.9972$). It also outperformed the combination of spectrograms and Gabor-filtered spectrograms ($\Delta = 0.0067$, $f_{h_1} = 0.9958$). Thus, a relatively simple acoustic feature space that focussed on acoustic edges not only equalled the benchmark but surpassed it. As a first test whether these best acoustic features could account for the same information as the articulatory features, we also tested a combination of them (*Sg&(Sg')*$_+$*&Art*, purple). The improvement of this combination of three feature subspaces over the best acoustic feature space was negligible ($\Delta = 0.0013$, $f_{h_1} = .0.7078$). This indicated that the articulatory

Figure 2.5 *(previous page)*: **Hyperparameter choices for source model optimisation (related to figure 2.4)**.
A Maximum distance between source positions used for test set prediction across all test sets and feature spaces. Each dot is one participant in the respective hemisphere, colour codes participants. **B** and **C** Positions of all test sets and feature spaces in meshes of individual brain volumes of two exemplary participants. **B** shows the participant with the largest maximum distance between chosen source locations and **C** the participant closest to the median of maximum distances between chosen source locations. Each dot is one test set, colour codes feature spaces. **D** Evaluation of spatial clustering of choices of source positions for different feature spaces using the Silhouette Index. Values close to 1 reflect that the optimisation procedure finds positions in source space that are highly similar within feature spaces but dissimilar across feature spaces, lower values reflect that the found positions are randomly arranged in source space. Each dot is one test set of one participant, colour codes participants. Black lines denote pooled means. **E** Choices of beamformer regularisation hyperparameter λ_{source} for each feature space. Each dot represents one test set of one participant, colour codes participants. Black lines denote pooled means.

features are not needed for an optimal prediction of the MEG responses.

We also explored the lateralisation of the performances by evaluating within-participant differences across hemispheres independent of feature spaces (figure 2.4D). We found that the posterior distributions of hemispheric beta differences were narrow for individual participants but exhibited a broad range of means within our sample. Some participants' responses were easier to explain in the left AC, others in the right AC, while for some there were no strong lateralisation effects.

Taken together, these results demonstrate that the gain in prediction performance obtained by combining articulatory features with spectrograms can be replicated in MEG data. However, a similar or even larger gain can be obtained by using algorithmically specified and generic acoustic features that capitalise on acoustic edges. Their performance in turn could not be improved by combining them with articulatory features. Next, we wanted to reveal in more detail how the precise information about the MEG predicted by the competing feature spaces was related to the information predicted by the benchmark articulatory features: were the similar levels of performance driven by the same or by different predictive information?

2.3.3 Shared and unique information of articulatory and acoustic features

Even if two models have the same predictive power, both higher than a reference model, each could offer improved performance based on different information (i.e. by better-predicting different periods of the speech signal) or the same information (i.e. by better-predicting the same periods of speech). The information theoretic PID framework (Ince, 2017a; Williams & Beer, 2010) provides a means to address this question (see meth-

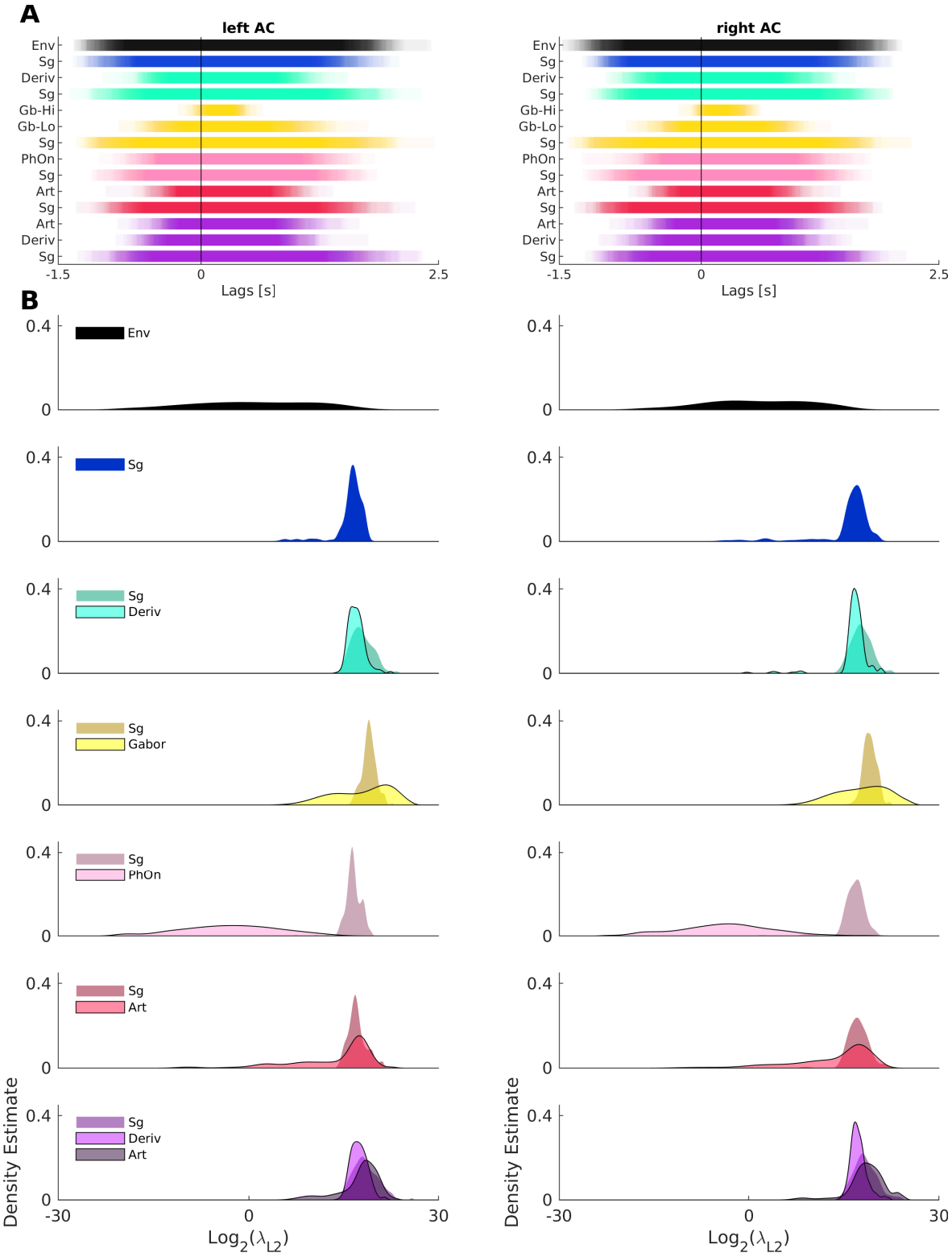

Figure 2.6: Caption on following page.

ods 2.5.5 for details). We used it to address two questions: 1) to which degree is the information carried by the acoustic- feature-based predictions shared (redundant) with that carried by predictions based on the benchmark articulatory features? And 2) to which degree do the predictions from each feature space contain unique information? If the benchmark features could be explained by the acoustic alternatives, then the results would be characterised by 1) a high degree of redundancy and 2) a low amount

Figure 2.6 *(previous page)*: **Hyperparameter choices of forward model optimisation differ systematically across feature (sub-)spaces (related to figure 2.4).**
A Choices of temporal extent hyperparameters for each feature (sub-)space. Shown are averages across inner folds used for test set predictions, pooled across participants. Values that were used in all cross-validated models of all participants are plotted as transparent bars ranging from $t_{Min} - t_{Max}$ such that the opacity codes for the number of participants and folds for which the temporal extent was chosen correspondingly. **B** Choices of $L2$ regularisation hyperparameters for each feature (sub-)space. Shown are distributions of choices averaged across inner folds used for test set predictions, pooled across participants. Colours code feature spaces.

of unique information left to the benchmark articulatory features. Such a finding would suggest that the two feature spaces predict the same parts of the response in the same way.

To investigate this question, we retrained all models with their source-space-related hyper-parameters fixed to the values that were found to be optimal for the benchmark articulatory features. We then considered separate, pairwise PIDs, where each acoustic feature space was compared to the benchmark articulatory feature space (figure 2.7). To make the resulting quantities more easily interpretable, we normalised the resulting redundant and unique information by the marginal Mutual Information (MI, Ince et al., 2017, a non-parametric measure of the relationship between variables) of the benchmark features and the observed MEG. We then statistically analysed these values using Bayesian hierarchical models similar to our analyses of the raw performances, focussing again on the regression coefficients that modelled the effects of feature spaces.

The acoustic features with the best prediction performance, *Sg&(Sg')*$_+$, were indeed also highly redundant with the benchmark articulatory features, reaching $\approx 100\%$ of the marginal MI provided by *Sg&Art* about the observed MEG (mean of the corresponding effect: 0.99, 95% credible interval (CI) $[0.98, 1.01]$). The same was the case when combining the best acoustic features with the articulatory features (*Sg&(Sg')*$_+$*&Art*, mean: 1.01, 95% CI: $[0.99, 1.02]$). Furthermore, we observed more unique information present in the acoustic feature space (mean: 0.07, 95% CI: $[0.06, 0.09]$) than in the benchmark articulatory feature space ($f_{h_1} = 1$), in which the unique information was distributed around 0 (mean: 0.01, 95% CI $[-0.01, 0.02]$). This means that all of the predictive information of the benchmark *Sg&Art* model was included in the predictions of the *Sg&(Sg')*$_+$ model. There was no unique information available in the *Sg&Art* prediction that a Bayesian optimal observer could not have extracted from the *Sg&(Sg')*$_+$ model.

Lastly, the information about the MEG responses only available from a joint consideration (i.e. synergy) of the benchmark articulatory features and the best acoustic features had a negligible effect size that was two orders of magnitude lower than that of the redundancy and failed to surpass a permutation-based noise threshold (see figure 2.11). These results agreed with the finding that a combination of the best acoustic feature spaces and the articulatory features did not have a better prediction performance than

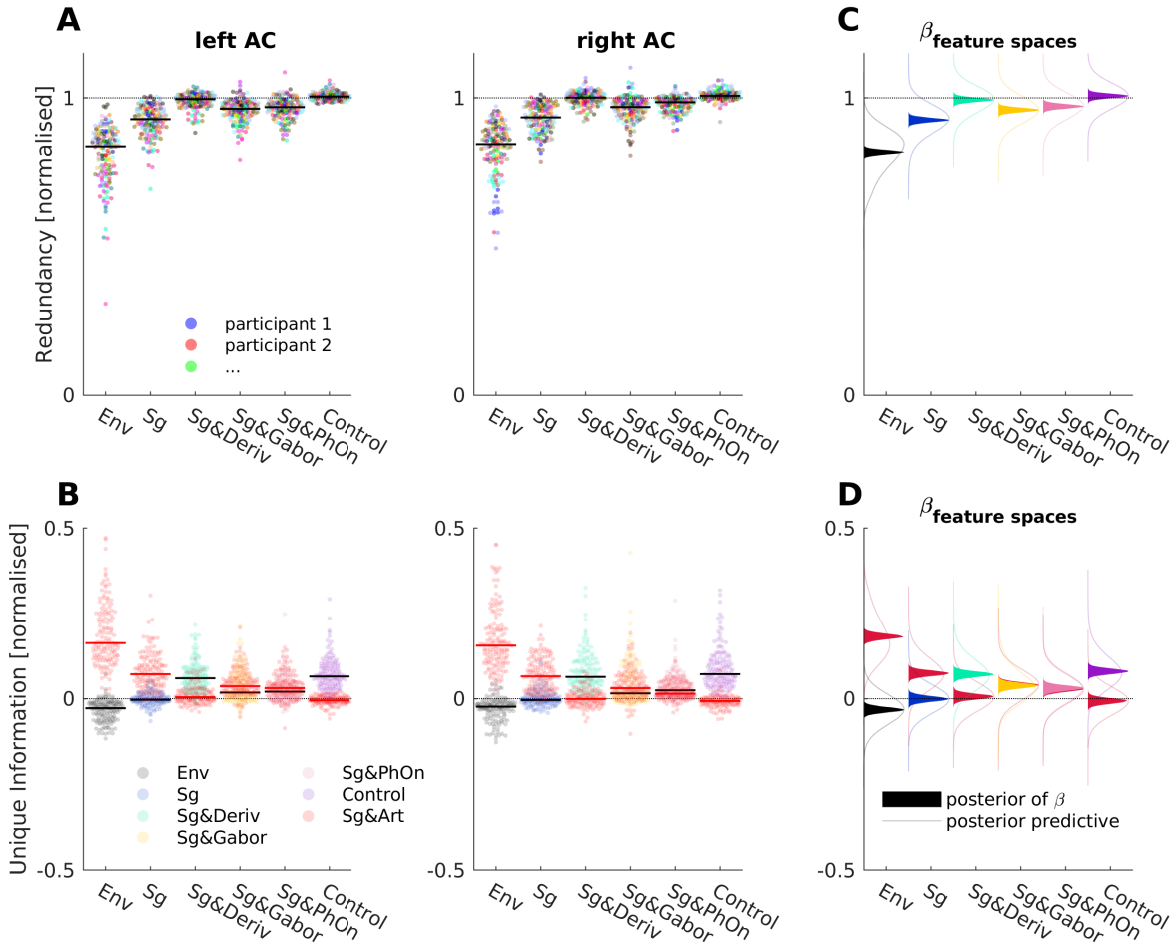

Figure 2.7: **Shared and unique contributions of articulatory and competing features.**
A Normalised redundancy in left and right auditory cortex (AC). Each colour codes for a single participant ($n = 24$). Each dot is one test set of one participant, black and red lines show pooled medians. **B** Normalised unique information of benchmark articulatory features and competing features in left and right AC. Colours code for a feature space, as shown. Each dot is one test set of one participant, black red lines show pooled medians. **C** and **D** Modelling of redundancy and unique information results, respectively. Filled areas show density estimates of posterior distributions of estimates of betas of feature spaces. Lines show density estimates of samples from posterior predictive distribution of the respective condition. Colour coding of feature spaces is the same as in B. See also figure 2.11.

the best acoustic features (see previous section).

A relatively high, normalised redundancy close to 100% was also achieved by *Sg&PhOn* (mean: 0.97, 95% CI: $[0.96, 0.99]$). In addition, *Sg&PhOn* provided a weak amount of unique information (mean: 0.03, 95% CI: $[0.01, 0.05]$) and left a very similar amount of unique information to the benchmark articulatory feature space (mean: 0.03, 95% CI: $[0.01, 0.04]$). The annotated onsets thus provide most of the information that the benchmark features provide about the observed MEG.

A very similar pattern was found for the second-best acoustic features, *Sg&Gb*. These features also achieved a relatively high redundancy (mean: 0.96, 95% CI: $[0.95, 0.97]$) but one that was lower than that of the best acoustic features ($f_{h_1} = 0.9988$).

Sg&Gb also provided a weak amount of unique information (mean: 0.04, 95% CI: $[0.02, 0.05]$) and left a very similar amount of unique information to the benchmark articulatory features (mean: 0.04, 95% CI: $[0.02, 0.06]$). We conclude that this high-dimensional acoustic feature space included both relevant and many irrelevant dimensions. The increase in the separability of the different spectrotemporal patterns that refer to different phoneme subgroups (Schädler et al., 2012) is thus less important than the sound energy patterns that are contained in the best acoustic feature space.

Finally, as expected from their comparably low prediction performances, the remaining feature spaces (*Env* and *Sg*) exhibited redundancies that were lower than that of the previously mentioned feature spaces (both $f_{h_1} = 1$). They also left considerable amounts of unique information to the benchmark feature space while providing no substantial positive unique information themselves (mean of *Env*: -0.03, 95% CI: $[-0.05, -0.02]$; mean of *Sg*: 0.00, 95% CI: $[-0.02, 0.02]$).

On a group level, all of these patterns were highly similar between left and right ACs.

Thus, the best acoustic features achieve their improved prediction performance over spectrograms alone by explaining the same parts of the responses that the benchmark articulatory features explain, and they additionally explain parts that the linguistic features do not while a joint consideration of both feature spaces does not add meaningful extra information.

2.3.4 Phoneme-evoked dynamics of observed and predicted time-series

As recently demonstrated, four manners of articulation of phonemes can be decoded from EEG data (Khalighinejad et al., 2017). We next assessed if this decoding was possible in our MEG data and the degree to which our encoding models could account for this phenomenon.

For this decoding analysis, we re-optimised the dipole position and sensor covariance matrix regularisation parameters of the spatial filters. We did this by using black-box optimisation, as before (Acerbi & Ma, 2017), only this time with respect to the MI between MEG data epoched to phoneme onsets and the manner of articulation of each phoneme (4 discrete phoneme classes were used: vowels, nasals, plosives and fricatives, see figure 2.13). The MI was calculated separately for each time point in the extracted phoneme epochs. For optimisation, we subsequently summed the MI across time points. In most cases, the positions found in this re-optimisation were very similar to those found before (figure 2.9A).

At the corresponding source locations, we found characteristic responses to the four manners of articulation (i.e. the four phoneme classes used, see figure 2.8A). We then retrained our encoding models based on all feature spaces with the source-level parameters fixed to the values found when optimising for MI between MEG data epoched to phoneme onsets and the manners of articulation. In the cross-validated predictions of

these retrained models, we observed phoneme-locked responses that were very similar to those obtained with observed MEG data (figure 2.8A, right).

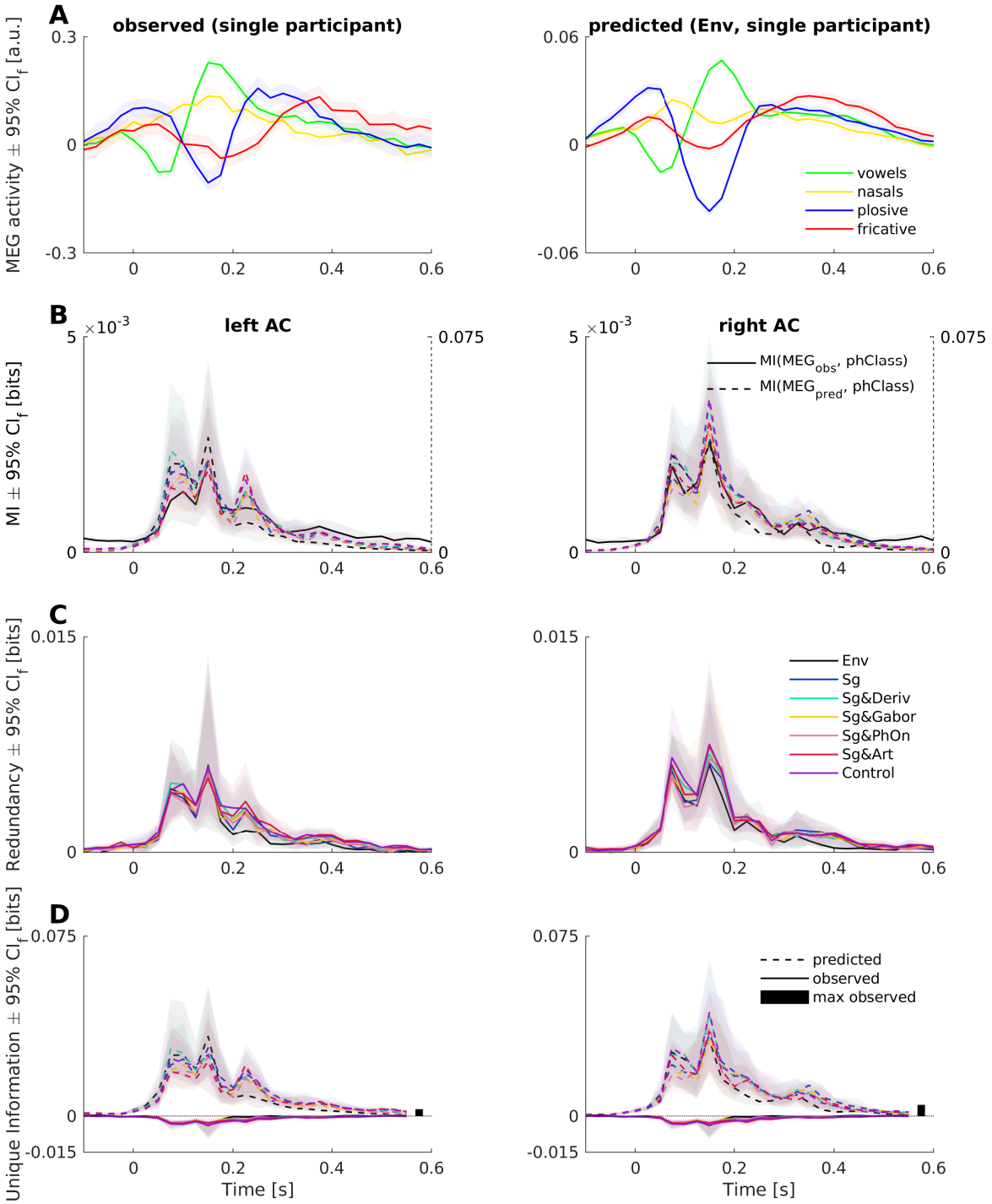

Figure 2.8: Caption on following page.

Correspondingly, we observed a sustained pattern of MI following the phoneme onsets in bilateral ACs for the observed data (figure 2.8B). We found very similar patterns of MI between manners of articulation and predicted phoneme-related fields, with values roughly an order of magnitude higher for the predictions. On average, this result pattern did not substantially differ between either of the two hemispheres or between the different feature spaces.

Together, our results thus show that the decoding of these manners of articulation

Figure 2.8 *(previous page)*: **Phoneme-related fields captured by model predictions.** **A** Phoneme-related fields of a single participant in (left) observed and (right) predicted MEG (from *Env* feature space). Colours code for 4 different phoneme classes that represent 4 manners of articulation. **B** MI of observed (solid lines, left y-axes) and predicted MEG (dashed and coloured, right y-axes) about the four phoneme categories in the left and right AC. Colour coding of feature sapces is the same as in C. **C** Redundancy from PID (amount of information that observed and predicted MEG share about the 4 manners of articulation). **D** Unique Information of observed (solid) and predicted MEG about the manners of articulation. Maximum information uniquely available from observed MEG across all participants, feature spaces, and time points are shown as black bars. Colour coding of feature spaces in C also applies to B and D. C and D show medians across all participants $\pm 95\%$ (frequentist) confidence intervals (CI_f), bootstrapped with $10,000$ samples. See also figure 2.9.

was replicable in the observed MEG data and in the MEG data predicted by our models.

To assess the amount of information that is shared by the observed and predicted time series about these manners of articulation, and the amount of information that is unique to them, we performed PIDs with observed and predicted time series as sources and the manners of articulation as targets. This analysis should reveal if the observed MEG contained information about these manners of articulation that is different from that obtained, for example, from the speech envelope when convolved with an encoding model temporal response function (TRF).

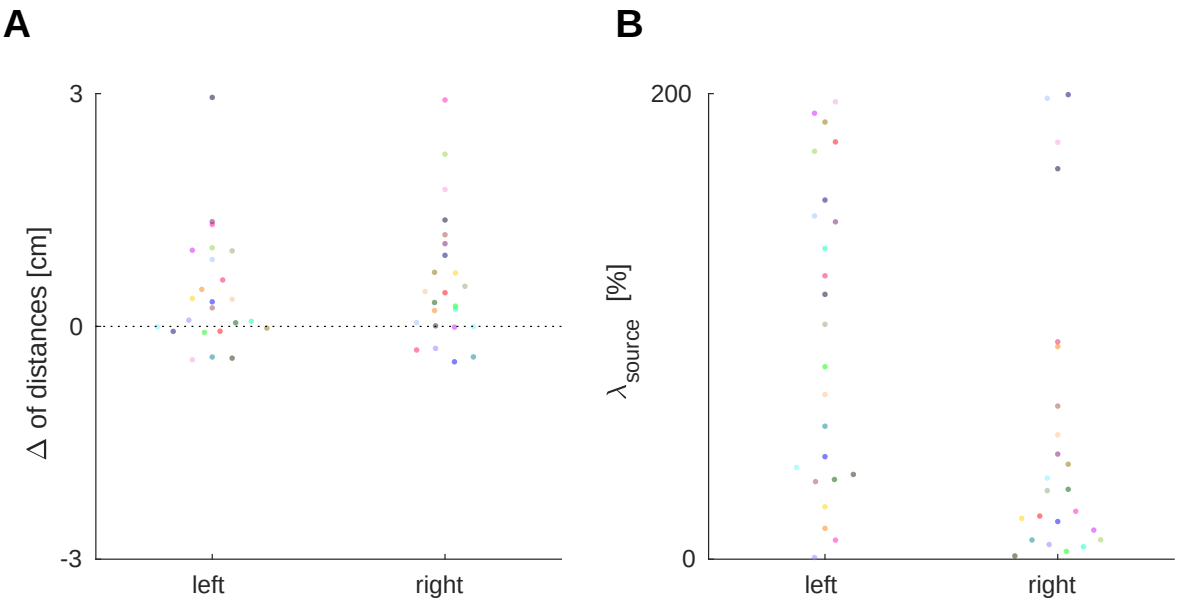

Figure 2.9: **Hyperparameter choices for phoneme related field (PRF) analysis (related to figure 2.8).**
A Maximal euclidean distances of source positions when optimised with regard to model performances across all test sets and feature spaces subtracted from maximal euclidean distances of source positions when positions found when optimising with regard to PRF MI are included. **B** Results of optimising sensor covariance regularisation parameter with regard to PRF MI. Colour in A and B codes participants.

The PIDs resulted in profiles of redundancy that closely resembled the marginal MI

profiles for both hemispheres and for all feature spaces alike (figure 2.8C). Most importantly, the information that was unique to the predicted MEG exhibited the same patterns (figure 2.8D, dashed lines), while the information unique to the observed MEG (solid lines) was negative, i.e. this information represented misinformation with respect to the predicted MEG source. This means that there were trials where an observer predicting phoneme classes optimally from the observed MEG would make a mistake (hence misinformation) that an observer of the predicted MEG would not make (hence unique to observed MEG; see methods 2.5.5 for more details on negative unique information). Thus, there was no relevant information about these manners of articulation present in the observed MEG that could not be retrieved from responses modelled with a convolution of any of our feature spaces with an encoding model filter. This pattern of results was also essentially the same for both hemispheres and for all feature spaces.

Taken together, these results demonstrate that models based on all of our feature spaces could fully account for the information about these four manners of articulation that was decodable from the observed MEG responses.

2.3.5 Replication using a publicly available EEG dataset

The original report of the effect of a performance gain provided by articulatory features over spectrograms alone was derived from EEG data (Di Liberto et al., 2015). Since MEG and EEG are sensitive to different sources (Cohen & Cuffin, 1983), it is possible that the MEG sensors we used here were blind to parts of the effect. We therefore investigated whether we could replicate our MEG results using EEG data. We analysed $n = 13$ participants for whom data with 128 channel recordings of approximately an hour are publicly available (Broderick et al., 2018b,a). On the stimulus side, we used the same analysis pipeline as for the MEG dataset. However, due to the higher noise level of the EEG data (Destoky et al., 2019), we did not try to fit the high- dimensional Gabor feature space. Instead, we concentrated on comparing the benchmark articulatory feature space to the lower dimensional acoustic feature spaces that had best explained the MEG data. We fitted cross-validated encoding models to the scalp-level EEG data and focussed our modelling on the 12 electrodes reported in the original publication (Di Liberto et al., 2015).

Using the same Bayesian modelling approach, results derived from the EEG data closely accorded with those derived from the MEG data (figure 2.10B). Our analysis replicated the gain in performance of the benchmark articulatory feature space compared to spectrograms alone ($\Delta = 0.0031$, $f_{h_1} = 0.9712$). We again found that the benchmark articulatory feature space was outperformed by the combination of spectrograms and their rectified temporal derivatives ($\Delta = 0.0045$, $f_{h_1} = 0.9963$). Also, we again found that combining the articulatory features with the best acoustic features only led to a negligible increase in performance ($\Delta = 0.0009$, $f_{h_1} = 0.7218$). In addition, as before, the benchmark articulatory feature space performance was not stronger than that of the spectrograms

and phoneme onsets combined ($\Delta = 0.0010$, $f_{h_1} = 0.2577$). Lastly, we found that all competing feature spaces outperformed the one dimensional envelope ($\Delta = [0.0128, 0.0213]$, all $f_{h_1} = 1$). These results thus show that – in terms of prediction performance – acoustic features outperform the more complex articulatory features, which perform on a par with features that only describe the phoneme timing.

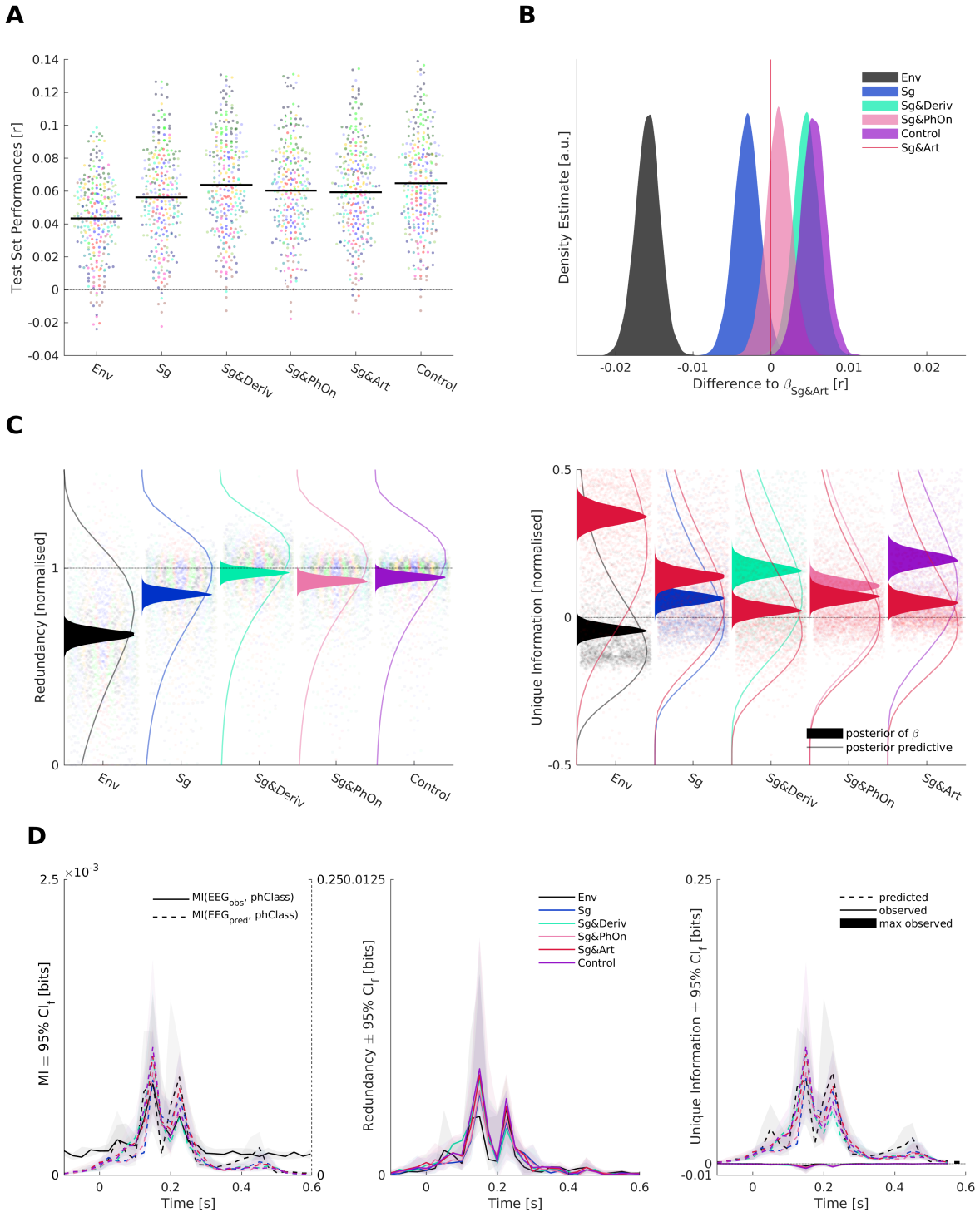

Figure 2.10: Caption on following page.

Note that when we replaced the log-mel spectrogram features chosen in the present study with a spectrogram more closely modelled after the one used in (Di Liberto et al., 2015), we obtained generally lower performances and also a different pattern of results.

Figure 2.10 *(previous page)*: **Analysis of EEG data.**
A Test set performances of forward models. Left: Each dot shows the performance in one test set averaged across electrodes. Colours code individual participants ($n = 13$), black lines show pooled medians. **B** Samples from posterior distribution of differences of beta estimates of competing feature spaces and the benchmark *Sg&Art* feature space. Colours code feature spaces. **C** PID results normalised by MI of predictions based on *Sg&Art* features and observed EEG signals. Each dot is one test set prediction of one participant and electrode. Samples from posterior distributions of effects of feature spaces are overlaid as filled areas, and posterior predictive distributions are shown as lines. Left: Redundancy of predictions based on benchmark articulatory features and competing feature spaces about observed EEG signals. Dot colours represent a participant, filled area and line colours represent feature spaces. Right: Unique information of benchmark articulatory features (red) and competing feature spaces about observed EEG signals. Colours of dots, filled areas and lines represent feature spaces. **D** Phoneme related potential analysis. Colours represent feature spaces, shaded areas denote 95% (frequentist) confidence intervals ($\mathrm{CI_f}$), bootstrapped with $10,000$ samples. All traces show averages across participants and electrodes. Left: MI of observed (solid black line, left y-axis of subplot) and predicted (dashed coloured lines, right y-axis of subplot) EEG about four manner of articulation phoneme classes ("phClass"). Middle: Redundancy – information shared by observed and predicted EEG from different feature spaces about phoneme classes. Right: Unique information of observed (solid lines) and predicted (dashed lines) EEG about phoneme classes. Maximum of information uniquely available from observed EEG across all participants, feature spaces and time points shown as black bar. See also figures 2.11 and 2.12.

Crucially, we found that this could be attributed to a compressive non-linearity as included in the log-mel spectrogram (see figure 2.11 for a more detailed explanation).

Taken together, these results further support the notion that simple and physiologically motivated transformations of the auditory stimulus can make important differences to the interpretation of more-complex annotated features.

Next, we considered the results of a PID analysis that assessed the degree to which the predictions of competing feature spaces shared information about the observed EEG responses with that of the benchmark feature space, and the degree to which they contributed unique information (figure 2.10C). We again found that the predictions based on the best acoustic feature space were highly redundant with predictions based on the benchmark articulatory features (mean of the corresponding effect: $0.9776; 95\%$ CI $[0.9358, 1.0167]$). The same was the case for the combination of the best acoustic features and the articulatory features (mean: 0.9527; 95% CI $[0.9104, 0.9945]$). We also again found that the unique information contributed by the benchmark articulatory features was close to 0 (mean of the corresponding effect: $0.0231, 95\%$ CI $[-0.0115, 0.0657]$), while the unique information contributed by the best acoustic feature space was weakly positive (mean of the corresponding effect: $0.1575, 95\%$ CI $[0.1121, 0.2109]$). Lastly, the amount of information only available when jointly considering the best acoustic features and the benchmark articulatory features (i.e., the syn-

ergy) was an order of magnitude lower than that of the redundancy and did not exceed noise thresholds (see figure 2.11), which agreed with the finding that combining the best acoustic features and the articulatory features did not lead to an improvement over the best acoustic features.

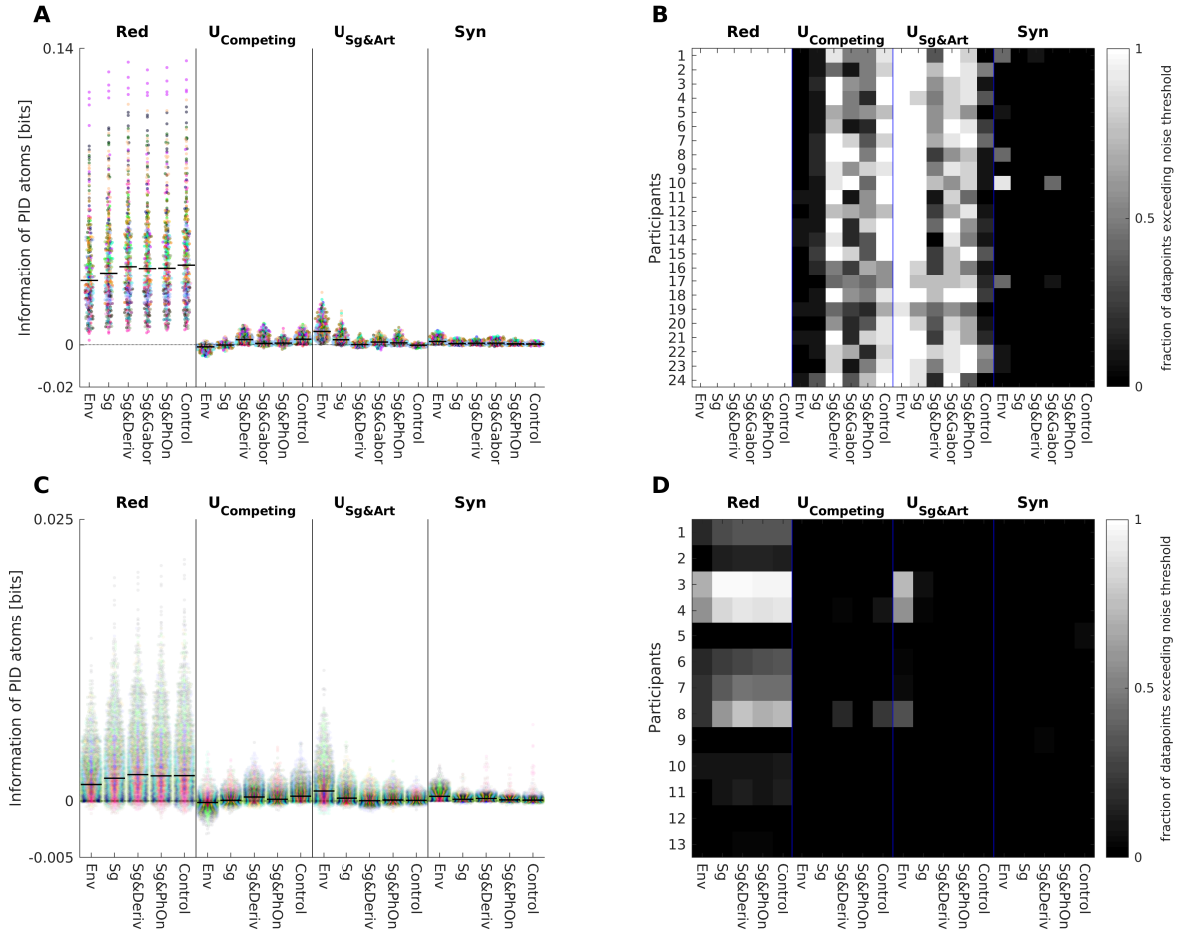

Figure 2.11: **Raw values and comparison to noise thresholds of PID in EEG and MEG (related to Figures 2.7 and 2.10).**
A Raw (unnormalised) PID values (Red: redundancy, U: unique information [of competing feature spaces and of benchmark articulatory feature space] and Syn: synergy) in MEG data from left and right AC. Each colour codes for a single participant, each dot is one test set. Pooled medians are indicated with black lines. **B** Comparison of PID values in MEG data to noise thresholds. Image plot shows the fraction of data points (sources, test sets) that exceeded the corresponding noise threshold in each participant and for each feature space and each PID atom. **C** Raw (unnormalised) PID values (redundancy, unique information of competing feature spaces, unique information of benchmark articulatory feature space and synergy) in EEG data from all 12 electrodes. Each colour codes for a single participant, each dot is one test set. Pooled medians are indicated with black lines. **D** Comparison of PID values in EEG data to noise thresholds. Image plot shows the fraction of data points (sources, test sets) that exceeded the corresponding noise threshold in each participant and for each feature space and each PID atom.

Similar to our results using MEG, the combination of spectrograms and phoneme onsets produced slightly lower levels of redundancy compared to the best acoustic model (mean: 0.9317; 95% CI $[0.8902, 0.9765]$), and even lower levels of redundancy were

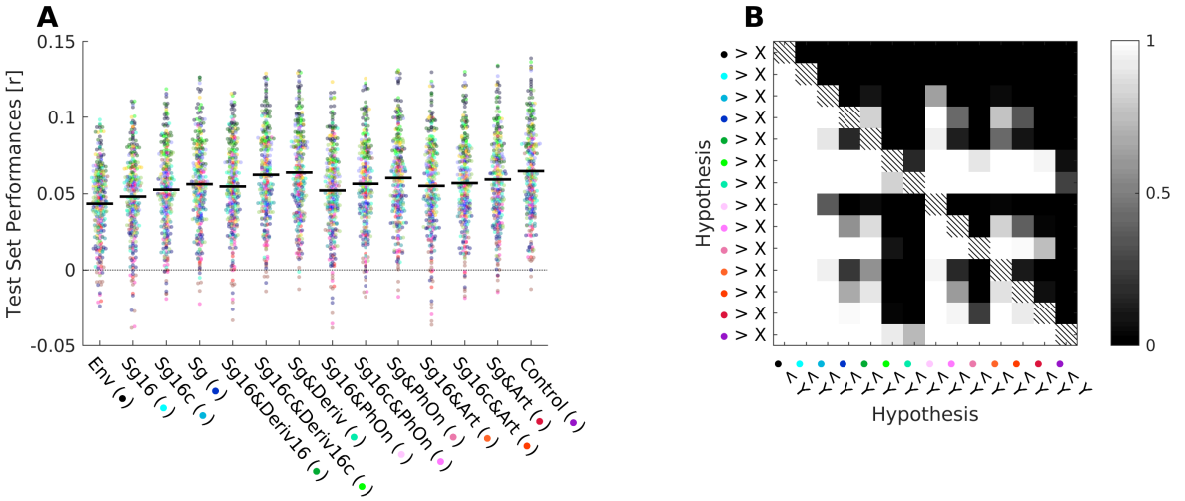

Figure 2.12: **Comparison of 16 channel spectrogram, 16 channel spectrogram with compressive nonlinearity and log-mel spectrogram in EEG data (related to figure 2.10).**
A Raw test set performances of feature spaces. Each dot is one test set of one participant, averaged across electrodes. Colour codes participants. Pooled medians overlaid. Colours in x-axis labelling refer to feature spaces. **B** Percent of samples in favour of hypotheses of differences of beta estimates between all feature spaces. Hypotheses are colour coded using the same colour mapping as in x-axis labelling of A. The performances obtained using Sg were in general higher than those obtained using $Sg16$, a 16-channel spectrogram modelled after the original study (Di Liberto et al., 2015). The combination of $Sg16\&Art$ failed to outperform Sg on its own ($f_{h_1} = 0.2318$). The overall pattern of performances was largely similar regardless of using Sg or $Sg16$ as the spectrogram. In contrast to the results produced using Sg however, we found that the $Sg16\&Art$ combination outperformed the $Sg16\&PhOn$ combination ($f_{h_1} = 0.9675$). To assess whether these differences were driven by the compressive linearity included in Sg, we additionally tested a version of $Sg16$ in which we raised its values to the power of 0.3 ("$Sg16c$"). Such nonlinearities are classically included in models of auditory processing, as early as the cochlea (Chi et al., 2005; Verhulst et al., 2018; Biesmans et al., 2017). This tweak indeed resulted in a pattern of performances that was largely similar to that obtained with Sg: The combination of $Sg\&Deriv$ did not clearly outperform the combination of $Sg16c\&Deriv16c$ ($f_{h_1} = 0.8191$), and combining $Sg16c\&Art$ was not better than combining $Sg16c\&PhOn$ ($f_{h_1} = 0.6097$).

obtained for spectrograms alone (mean: 0.8658; 95% CI $[0.8157, 0.9118]$), and for the envelope (mean: 0.6714; 95% CI $[0.5956, 0.7205]$).

Based on these results, we concluded that in both MEG and EEG data, the increased performance provided by benchmark articulatory features over spectrograms alone could be explained by a combination of spectrograms and their rectified temporal derivatives. This purely acoustic feature space achieved higher overall performance in predicting EEG responses. It did so by explaining the same information as the benchmark articulatory features. However, it also carried information that was not available from the predictions based on the benchmark articulatory features.

Finally, we also found a very similar pattern of results in an analysis of phoneme-evoked responses (figure 2.10D). The MI of the observed EEG time series and the four

phoneme classes was mostly shared with that of the predicted time series based on all feature spaces. The predicted time series could thus account for a substantial amount of positive unique information, while the observed EEG time series could only contribute negative unique information, i.e. misinformation. The observed EEG responses thus did not contain any more information about the manners of articulation than did the EEG response predictions based solely on the envelope.

2.4 Discussion

In this study, we set out to investigate to which degree signatures of "pre-lexical abstraction" in MEEG responses to speech can be explained with simpler, purely acoustic models. Our results suggest that care must be taken when interpreting the results of encoding or decoding models that consider higher-order constructs, such as the articulatory features of phonemes. We showed that the predictive information that can be derived from articulatory features is rooted in the timing information of these features rather than in a more-detailed characterisation of the phoneme. Similarly, the ability to reliably decode subgroups of phonemes from MEEG data can be explained by our simplest feature model, that is, it is a direct consequence of MEEG speech envelope tracking. It should therefore not be interpreted as evidence of more complex speech processing being reflected in the recorded signal. Based on these results, we argue here for the consideration of algorithmically interpretable and physiologically plausible models of sensory encoding, for which annotated feature spaces can nevertheless serve as excellent benchmarks.

An inevitable limitation of this study is that our results cannot ultimately prove the absence of explanatory power unique to the articulatory features. It is possible that analysis pipelines exist that could carve out parts of the responses such that the articulatory features could beat our best acoustic feature space. However, in our analyses, the articulatory features were given strong chances to predict response variance. And we could indeed replicate the originally reported effect of a performance gain over spectrograms alone, only to then find a more parsimonious explanation for this gain. Moreover, our findings suggest that if the articulatory features could better explain certain parts of the responses, these parts would account for a relatively small portion of the total response variance. Given the already small effect sizes, it would then be possible that additional and similarly simple transformations of the acoustics could compensate for possible articulatory advantages. The same holds true for recent demonstrations of more-sophisticated linguistic feature spaces (Broderick et al., 2018b; Brodbeck et al., 2018a). Essentially, this line of reasoning thus drives home our main point. Any invocation of exciting, high-level feature spaces will always entail the heavy burden of proof of the absence of simpler explanations (Sassenhagen, 2018). This should by no means discourage inspiring investigations from using such high-level feature spaces but it should encourage researchers to nevertheless continue to consider simpler explanations.

Similarly, the ability to decode high-level semantic or phonetic properties of speech from evoked neural data tantalisingly suggests that the measured neural response reflects high-level processing. However, in general it is extremely difficult to control properly for all possible low-level stimulus properties, which could confound the interpretation of the high-level feature decoding. Applying decoding analyses to the predictions of forward models as we suggest here provides one way to address this issue. If, as we find here, the high-level feature can indeed be decoded from the prediction of a forward model based on low-level stimulus features, it suggests that the decoding results should not be interpreted as strong evidence of high-level neural processing.

We used in our study a source reconstruction approach that used data derived from two dipoles in bilateral auditory cortices. Source-level MEG data in (Brodbeck et al., 2018a), for example, suggest that multiple, superior temporal sources related to speech processing are robustly separable. This could be explained by the difference in source localisation algorithms. Given the relatively coarse, spatial resolution of our source-level data, we chose not to focus our analysis on modelling activity reconstructed from multiple locations in source space. Instead, we invested our computational resources in a detailed analysis of responses from a single point per hemisphere that accounted for much of the speech-related variance. This allowed us to flexibly optimise analysis parameters specific to participants, hemispheres and feature (sub-)spaces. We believe that this data-driven approach to parameter settings (Hahn et al., 2018) marks an important step towards more-principled pipelines in neuroimaging (Bzdok & Yeo, 2017), and our approach was inspired by growing efforts to avoid MEG analysis parameter settings based on tradition (Woolrich et al., 2011; Engemann & Gramfort, 2015).

Since forward-encoding models promise to inform theories of neuronal computations, what are the potential implications of this study? The central question of interest concerns the origins of the response variance that is commonly explained by the best acoustic and articulatory benchmark features. However, interpreting the results of encoding and decoding models with regard to such a causal question is never trivial (Weichwald et al., 2015; Kriegeskorte & Douglas, 2018b). The feature spaces considered here reflect functional – not mechanistic – models (Kay, 2018) of varying predictive performance. What they essentially relate is the input of the waveform of a speech stimulus to the output of MEEG responses. These responses are far from reflecting the entire, drastically higher-dimensional cortical auditory representation of the stimulus. Against this backdrop, the fact that the envelope, a low-fidelity representation of the stimulus, could still account for most of the observed response variability, suggests that this part of brain activity might not so readily provide a window to arbitrary high-level cognitive processes.

Furthermore, an algorithmic consideration of our best acoustic feature space rather points to operations which occur relatively early in auditory processing. A spectral decomposition of compressed dynamic range is typically part of cochlear models (Verhulst et al., 2018; Chi et al., 2005). An additional temporal derivative and half-wave rectification might possibly be implemented by the various stations along the subcortical auditory

pathway. The question then is why cortical neuronal mass signals (Panzeri et al., 2015) are time-locked to this result of very early auditory processing, and whether these low-frequency cortical responses carry such information so that further cortical processes react to it. Deeper insights into this problem will also have to consider proxies to what downstream neurons are encoding, such as the final behavioural responses (Williams et al., 2007; de-Wit et al., 2016; Panzeri et al., 2017; Bouton et al., 2018; Keitel et al., 2018; Carlson et al., 2018; Brette, 2018).

Despite these caveats, it is interesting to speculate how the feature spaces considered here might reflect aspects of actual cortical computations. Unlike modern ASR systems that can, with limitations, understand a speaker's intention (Sarikaya et al., 2016), the mid-level feature spaces considered here are all far from this feat. Nevertheless, they can be interpreted as contributing to this goal. The information bottleneck framework (Tishby et al., 2000) for example suggests that feature spaces should allow information compression, i.e., gradual decreasing stimulus fidelity, while retaining relevant aspects of the input. The log-mel spectrograms allow to discard irrelevant spectral and dynamic ranges, and Gabor-filtering can do the same for spectrotemporal patterns relevant for ASR systems (Schädler et al., 2012). This decomposition seems to be especially beneficial for speech in noise, when features similar to the best acoustic feature space used here can be used to exploit the rapid amplitude dynamics in speech signals to the benefit of ASR systems (Kumar et al., 2011). It is thus conceivable that the predictive performance of this feature space could be rooted in a tuning of the auditory system to ubiquitous noisy listening environments. Hypotheses about the processing of speech in noise are, however, best examined in datasets that sample the stimulus space correspondingly (Fiedler et al., 2018; Giordano et al., 2016).

Another interesting observation is that the edges of these rapid amplitude dynamics coincide with transitions to the central vowels of syllables (Oganian & Chang, 2019). A rich literature is available on the interpretation of low-frequency signals as a signature of a chunking of the speech signal into syllable-like units (Hertrich et al., 2012; Gross et al., 2013; Doelling et al., 2014; Hyafil et al., 2015; Räsänen et al., 2018; Giraud & Poeppel, 2012; Ghitza, 2013). An eventual goal would however be to treat mid-level representations as less independent from the more-abstract aspects of speech understanding. Extracting the intermediate representations generated while embedding speech into fixed dimensional semantic vectors (Chung et al., 2018) could be a promising step towards an unbiased and context dependent description of speech signals.

2.4.1 Conclusion

In a data-driven approach, we have studied models that explain cortical neuronal responses as captured by source-localised MEG and sensor level EEG in a story- listening paradigm. Our results underscore that annotated linguistic feature spaces are useful tools to explore neuronal responses to speech and serve as excellent benchmarks. We

find their performance for explaining neuronal responses of high temporal resolution to be exceeded and explained by a simple low-level acoustic feature space that capitalises on spectrotemporal dynamics. Thus, we conclude that the consideration of parsimonious, algorithmically interpretable and physiologically plausible features will eventually lead to clearer explanations of observed neuronal responses.

2.5 Methods

2.5.1 Participants

24 healthy young participants (native speakers of English, 12 female, mean age 24.0 years, age range $[18, 35]$ years) agreed to take part in our experiment. They provided informed written consent and received a monetary compensation of £9 per hour. The study was approved by the College of Science and Engineering Ethics Committee at the University of Glasgow (application number: 300170024).

2.5.2 MEG recording, preprocessing and spatial filtering

MEG recording

Participants listened to a narrative of 55 minutes duration ("The Curious Case of Benjamin Button", public domain recording by Don W. Jenkins, librivox.org) while their brain activity was recorded with a 248 channel magnetometer MEG system (MAGNES 3600 WH, 4D Neuroimaging) at a sampling rate of 1017.25 Hz (first 10 participants) and 2034.51 Hz (last 14 participants). Prior to recording, we digitised each participant's headshape and attached five head position measurement coils to the left and right preauricular points as well as to three positions spread across the forehead. The session was split into 6 blocks of equal duration and additionally included a repetition of the last block. The last ten seconds of each block were repeated as a lead-in to the following block to allow listeners to pick up the story. Prior to and after each block, we measured the positions of the coils. If the movement of any of them exceeded 5 mm, we repeated the block. Playback of the story and trigger handling was done using PsychToolBox (Brainard, 1997), and sound was delivered via two MEG compatible Etymotic ER-30 insert earphones. After the recording, participants had to answer 18 multiple choice questions with 3 options each, where the number of correct options could vary between 1 and 3 per question. The questions referred to the entire story, covering three details per recording block. The average performance was $.95$ with a standard deviation of 0.05 and a range from 0.78 to 1.00.

MEG preprocessing

Most of our analyses were carried out within the MATLAB computing environment (v2016a, MathWorks, Natick, MA, USA) using several open-source toolboxes and custom code. Deviations from this are highlighted. Preprocessing was done using the fieldTrip toolbox (Oostenveld et al., 2011). Initially, we epoched the data according to the onsets of the full blocks including the ten seconds of lead-in. For noise cancellation, we subtracted the projection of the raw data on an orthogonal basis of the reference channels from the raw data. We manually removed and subsequently replaced artefactual channels with spherical spline interpolations of surrounding channels (mean number of artefactual channels per block: 3.07, standard deviation: 3.64; pooled across participants), replaced squid jumps with DC patches, filtered the signal with a fourth-order forward-reverse zero-phase butterworth high-pass filter with a cutoff-frequency of $.5$ Hz and downsampled the data to 125 Hz. We then excluded the lead-in parts from the blocks and performed Independent Component Analysis (ICA, runica algorithm) to identify and remove components reflecting eye and heart activity (mean number of components per block: 6.70, standard deviation: 5.01; pooled across participants) and further downsampled the data to 40 Hz.

MEG source space

We employed three different source modelling approaches for our analysis. Firstly, we aimed to identify regions in source space whose activity was in a repeatable relationship with our auditory stimulation ("story-responsive" regions, Honey et al., 2012; de Heer et al., 2017). Secondly, we wished to visualise these results on a group-level. Lastly, for our main intention of modelling the story-responsive regions, we designed a framework that would allow us to optimise parameters of our spatial filters as part of a cross-validation, similar to a recent proposal by Engemann & Gramfort (2015).

Volume conductor models For all three approaches, we obtained common volume conductor models. We first aligned individual T1-weighted anatomical MRI scans with the digitised headshapes using the iterative closest point algorithm. Then, we segmented the MRI scans and generated corrected-sphere volume conductor models (Nolte, 2003). We generated grids of points in individual volumes of 5 mm resolution. For group-level visualisation purposes, we also generated a grid with 5 mm point spacing in MNI space, and transformed this to individual spaces by applying the inverse of the transform of individual anatomies to MNI space.

Initial data exploration: identification and characterisation of story-responsive regions To identify story-responsive regions in MEG source space, we projected the time- domain sensor level data through rank-reduced linearly constrained minimum variance beamformer spatial filters (Van Veen et al., 1997) with the regularisation of the

sensor covariance matrices λ_{source} set to 5%, using the dipole orientation of maximal power. We correlated the responses to the last block with those to its repeated presentation within each participant to obtain maps of test-retest-R^2. We repeated this using the grids in MNI space we had warped into individual anatomies for a group-level visualisation using the *plot_glassbrain* function of the Python module Nilearn (Abraham et al., 2014).

We then explored how many dipoles would explain how much of the repeatable activity in story-responsive regions. It is known that due to the non-uniqueness of the inverse problem, the spatial resolution of MEG source reconstructions is inherently limited. Neighbouring grid points are thus often highly correlated, rendering analyses on a full grid highly redundant (Faharibozorg et al., 2018). To avoid such an unnecessary computational burden for our modelling, we used an information theoretic approach to characterise redundant and unique regions in source space.

First, we computed Mutual Information (MI, Ince et al., 2017) at each grid point in individual source spaces between activity in the first and the second repetition, essentially repeating the initial identification of story-responsive regions. Next, we applied the framework of PID (Ince, 2017a) to the data of repeated blocks in an iterative approach. PID aims to disentangle redundant, unique and synergistic contributions of two source variables about a target variable (see later section dedicated to PID for more details). As the first source variable, we here used the two-dimensional activity at bilateral grid points of individual peak story-responsivity during the first repetition. We then scanned the whole grid in parallel for both repetitions, using the activity recorded during the first repetition as the second source variable and the activity recorded during the second repetition as the target variable of PIDs (see video S1 for an intuitive visualisation). We were then interested in the resulting maps of redundancy and unique information. The former would allow us to infer to what degree other grid points with high story-responsivity shared their information about the repetition with the grid points of peak story-responsivity. The latter on the other hand would show us where information unexplainable by these two peaks could be found. After this first iteration, we added the grid point of peak unique information to the then three-dimensional first source variable in the PIDs and repeated the computation across the whole grid. We reran this approach for a total of ten iterations. Finally, we computed MI between the two-dimensional activity at bilateral peaks of story-responsivity in the first and the second repetition and compared this to the unique information found in each iteration of our iterative approach.

Optimisation of source space coordinates and sensor covariance regularisation In order not to unnecessarily spend computational resources, we wanted to limit our main endeavour of modelling MEG responses to parts of the signal which actually were in a systematic relationship with the stimulus. A straight-forward solution for a selection of these parts would have been to directly use the grid points identified as story-responsive using the test-retest correlation. However, since it is likely that participants

paid a lesser degree of attention to the more predictable repeated presentation of the last chapter, we could not rule out that the test-retest-R^2 maps would be biased towards low-level auditory processing. Furthermore, these maps could be influenced by differences in the position of the participant's head in the scanner as well as the amount of eye blinks and head movements. The peak test-retest points are thus not guaranteed to be the optimal locations for any given feature space model fit, tested over the whole experiment. Moreover, it was possible that different feature spaces would optimally predict distinct regions. Finally, we did not know a-priori what level of regularisation of the sensor covariance matrices would be ideal to capture the responses of interest in each individual dataset.

To account for all of these considerations in a data-driven manner, we treated the coordinates of regions of interest as well as the regularisation of sensor covariance matrices as hyperparameters of our model, which we optimised by means of a black-box optimisation algorithm. We kept the other specifications of the spatial filter design as described above. As initial coordinates, we used the maxima of test-retest-R^2 maps within each hemisphere. The boundaries of the coordinate hyperparameters were defined by the boundaries of the respective hemisphere of the individual brain volume which we shrunk by a factor of 0.99 for this purpose to avoid instabilities of the forward models close to their boundaries (table 2.2). In each iteration of the black-box optimisation, we then applied a given amount of regularisation to the precomputed sensor covariance matrix and computed the leadfield for a given vector $\bar{c} = (X, Y, Z)$ of coordinates in source space using the precomputed volume conductor model for each block. Since the orientation of the resulting dipoles was then arbitrary, i.e. possibly flipped across blocks, we estimated the mean axis of dipoles across blocks and changed the sign of the orientations of dipoles whose dot products with the orientation of the dipole closest to the mean axis were negative. We then recomputed the leadfields for these aligned dipole orientations. Finally, we projected the sensor level data through these spatial filters and z-scored them within each block to account for differences in mean amplitude across blocks.

2.5.3 Stimulus transformations

The speech stimulus was transformed into various feature spaces. We used the GBFB toolbox (Schädler et al., 2012) to obtain 31-channel Log-Mel-Spectrograms ("*Sg* ", ranging from $[124.1, 7284.1]$ Hz) and summed these across the spectral dimension to also obtain the amplitude envelope ("*Env* ").

Additionally, we filtered the spectrograms with a set of 455 2D Gabor filters ("*Gb*") of varying centre frequencies corresponding to those of the *Sg* as well as spectral modulation frequencies Ω (0, 2.9, 6, 12.2 and 25 Hz) and temporal modulation frequencies ω (0, 6.2, 9.9, 15.7 and 25 Hz). Notably, this implementation of the toolbox only considers a subset of all possible combinations of centre frequency as well as spectral and temporal

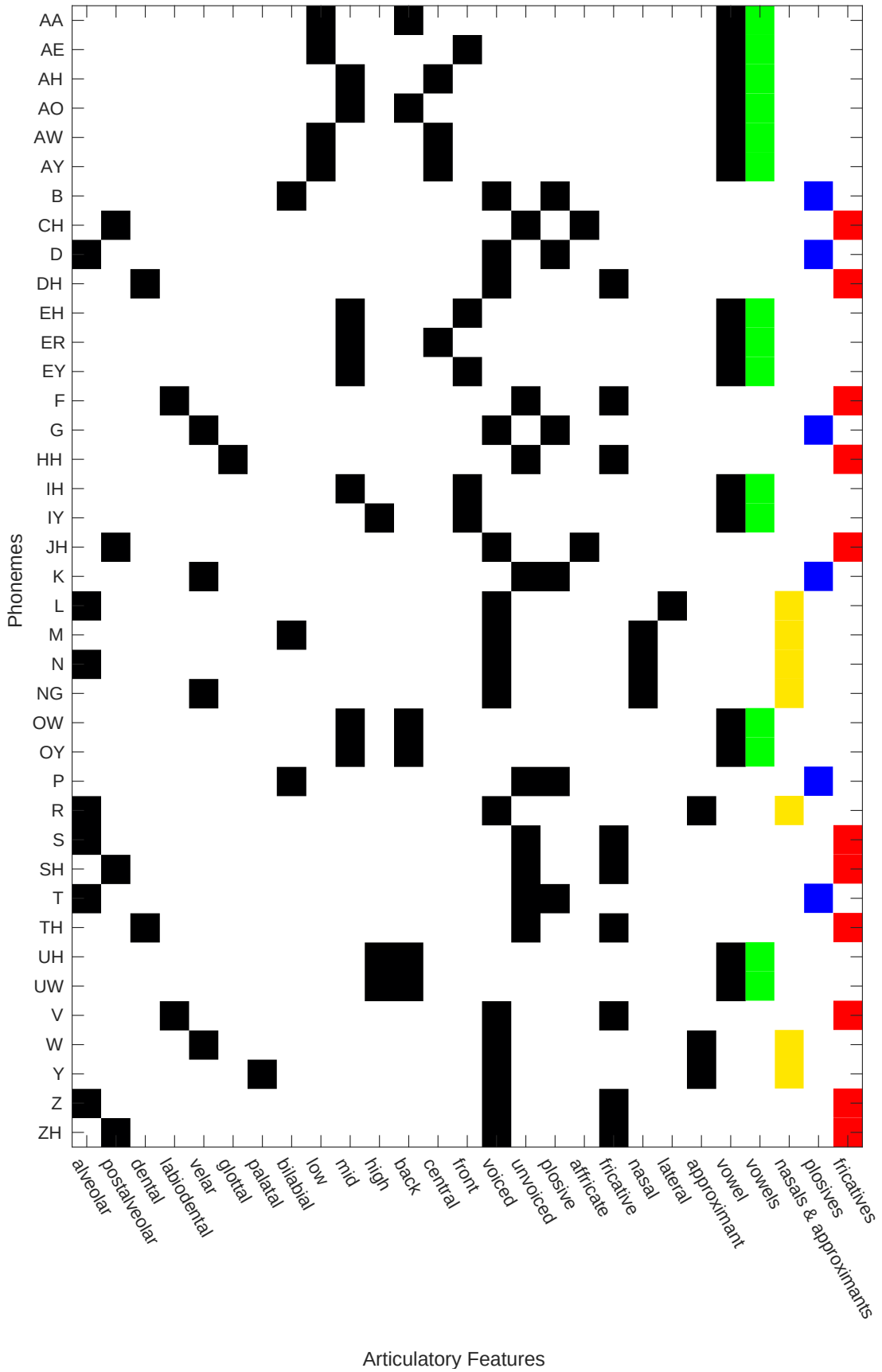

Figure 2.13: Caption on following page.

modulation frequencies to avoid overly redundant features. As a last acoustic feature space, we computed half-wave rectified first derivatives of the individual channels of the spectrograms "$(Sg')_+$", (Hertrich et al., 2012; Brodbeck et al., 2018a).

Figure 2.13 *(previous page)*: **Mapping of phonemes to articulatory features and manners of articulation (related to methods 2.5.3).**
Black and white part shows articulatory features used for forward modelling (Di Liberto et al., 2015), coloured part shows manners of articulation used for decoding (Khalighinejad et al., 2017).

To construct annotated feature spaces, we used the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008) to align the text material to the stimulus waveforms, providing us with onset times of phonemes comprising the text. These were manually corrected using Praat (Boersma, 2001) and subsequently transformed into a 23-dimensional binary articulatory feature space ("*Art*" de Heer et al., 2017). Figure figure 2.13 provides the mapping from each phoneme to the articulatory features. We generated 23 time-series of zeros at a sampling rate of 40 Hz and inserted unit impulses at the onset times of phonemes corresponding to the respective articulatory feature. Finally, we discarded the information about phoneme identity to obtain a one-dimensional binary feature space of phoneme onsets ("*PhOn*"). Our set F_{MEG} of employed feature spaces then consisted of the following combinations: F_{MEG} = {*Env*, *Sg*, *Sg&(Sg')*$_+$, *Sg&Gb*, *Sg&PhOn*, *Sg&Art*, *Sg&(Sg')*$_+$*&Art* }. We downsampled the acoustic feature spaces to 40 Hz and z-scored all feature spaces prior to modelling.

2.5.4 Mapping from stimulus to MEG

To perform a linear mapping from our feature spaces to the recorded MEG signals, we used ridge regression (Crosse et al., 2016) in a 6-fold nested cross-validation framework (Varoquaux et al., 2017). This allowed us to tune hyperparameters controlling the temporal extent and the amount of $L2$ regularisation of the ridge models as well as the amount of regularisation of the sensor covariance matrices and the coordinates of positions in source space for the beamformer spatial filters in the inner folds, yielding data-driven optimised models for each feature space, hemisphere and participant.

Linear model

The single-subject linear model we employed can be formulated in discrete time as:

$$\hat{r}_{\bar{c},\lambda_{source}}(t) = \sum_{\upsilon} \sum_{\tau=t_{Min}}^{t_{Max}} w(\upsilon,\tau) s(\upsilon, t-\tau) \tag{2.1}$$

Here, $\hat{r}$ denotes the neuronal response as obtained with a spatial filter with maximum gain at the vector $\bar{c}$ of coordinates (X,Y,Z) in source space and a regularisation of the sensor covariance matrix of λ_{source}. Further, s is a representation of the stimulus in a given feature space, possibly multidimensional with dimensions υ. Finally, w describes the filter weights across these dimensions and time lags τ ranging from t_{Min} to t_{Max}, where negative values refer to samples in the future of t and positive values refer to

samples in the past of t.

To obtain these filter weights, we used the following closed-form solution:

$$w = (S^T S + \lambda_{L2} I)^{-1} S^T r_{\bar{c},\lambda_{source}} \tag{2.2}$$

Here, S denotes the lagged time series of the stimulus representation, each column consisting of a particular combination of lags τ and feature dimensions υ, organised such that neighbouring feature dimensions populate neighbouring columns within groups of columns corresponding to time lags. The identity matrix I is multiplied with λ_{L2}, a hyperparameter adjusting the amount of $L2$ regularisation. Larger values of λ_{L2} force the resulting weights w closer to zero and thus reduce overfitting.

For the joint feature spaces consisting of multiple subspaces, the temporal extent and $L2$ regularisation was optimised individually for each subspace to obtain the best possible prediction performance. This meant that the matrix S was constructed as the columnwise concatenation of multiple submatrices with different numbers not only of feature dimensions υ but also of lags τ. Additionally, this meant that λ_{L2} here was a vector instead of a scalar, with as many elements as feature spaces in the joint space. Corresponding to the concatenation of S, different sections of the diagonal of the identity matrix were multiplied with the dedicated regularisation parameters of the corresponding subspace.

We used an additional regularisation for the *Gb* feature space. We had observed that feature dimensions belonging to the group of fastest temporal modulation frequencies ω had noisy and small filter weights at long absolute temporal lags. Based on this, we concluded that the temporal extents τ chosen for this feature space were essentially a compromise of long optimal τ for feature dimensions of slow ω ("*Gb-Low*") and short optimal τ for feature dimensions of fast ω ("*Gb-Hi*"). To remedy this problem, we assigned the usual τ to the group of slowest ω and added additional τ hyperparameters for the group of fastest ω. The τ of the central ω were then spaced proportionally to the mean auto-correlation times (ACT) of the corresponding groups of feature dimensions of this stimulus representation. We defined the ACT as the shortest lag where the normalised and absolute auto-correlation dropped below a value of $.05$. This allowed the optimisation algorithm to pick long τ for feature dimensions of slow ω and short τ for feature dimensions of fast ω.

Nested cross-validation and hyperparameter tuning

To make data-driven optimal choices for the range of lags τ defined by t_{Min} and t_{Max}, the amount of $L2$ regularisation λ_{L2}, the coordinates in source space as well as the amount of regularisation of the sensor covariance matrices λ_{source}, we used nested cross-validation. Specifically, this means that we split our stimulus and response data in six portions of equal durations. Two loops then subdivided the data into training, tuning and testing sets. In each iteration, an outer loop assigned each of the six portions to be the testing set.

Additionally, in each iteration of the outer loop, a full run of an inner loop was performed, assigning four portions to be the training set and the remaining portion to be the tuning set. This resulted in a total of 30 different assignments of portions to different sets. With this framework, we first picked a certain combination of hyperparameters and computed the corresponding weights w, the elementary parameters, using the training set. The resulting filters were convolved with the stimulus of the tuning set to obtain predictions $\hat{r}$ which we correlated with the observed responses r to obtain the tuning performance. This was repeated 200 times with different combinations of the hyperparameters.

These combinations were chosen by a black-box optimisation algorithm, Bayesian Adaptive Direct Search (Acerbi & Ma, 2017). BADS uses Gaussian Processes to construct a computationally cheap internal model of the multidimensional performance landscape using already available evidence and smoothness assumptions. As the computationally relatively costly linear models are evaluated across iterations, more evidence about the true performance landscape builds up which is used to update the internal model, i.e. assumptions about the smoothness and shape of the performance landscape at hyper-parameter combinations not yet evaluated. The internal model is used to update an acquisition function, whose maximum determines which combination of hyperparameters would be most informative to evaluate next in order to find the global optimum of the performance landscape. While this algorithm is not guaranteed to find the optimal combination, i.e. it is possible that it gets stuck in local optima, it has been shown to outperform other black-box optimisation algorithms on datasets typical for cognitive neuroscience (Acerbi & Ma, 2017). The values at which the hyperparameters were initiated as well as the ranges to which they were constrained are shown in table 2.2. Once all iterations of an inner loop were finished, we averaged the hyperparameter choices of all inner folds. We then retrained the elementary model parameters with stimulus and response data corresponding to these averaged hyperparameters on all five possible assignments of data portions to training sets in the current outer fold. We subsequently averaged the elementary parameters across inner folds and used the resulting weights to perform a prediction on the test set of the current outer fold. This was repeated for all outer folds to obtain a number of test set predictions corresponding to the number of outer folds.

	initial value	lower boundary	upper boundary
$t_{Min}[s]$	-0.2	-1.5	.5
$t_{Max}[s]$	0.8	0.2	2.5
$\log_{2(\lambda_{L2})}$	19	-30	30
$\lambda_s[\%]$	30	0	200
X_{source}	$X_{max(R^2)\|hemisphere}$	$min(X_{volume\|hemisphere})$	$max(X_{volume\|hemisphere})$
Y_{source}	$Y_{max(R^2)\|hemisphere}$	$min(Y_{volume\|hemisphere})$	$max(Y_{volume\|hemisphere})$
Z_{source}	$Z_{max(R^2)\|hemisphere}$	$min(Z_{volume\|hemisphere})$	$max(Z_{volume\|hemisphere})$

Table 2.2: **Initial values and boundaries for hyperparameters in BADS optimisation (related to related to methods 2.5.4).**

As the optimisation procedure was not guaranteed to find the optimal combinations of parameters, a crucial quality control of our approach was to check the amount of variance across parameter choices. High degrees would e.g. reflect that the optimisation algorithm would get stuck in local optima, or that the respective parameter was of minor importance for the model performance. Low degrees on the other hand would demonstrate that the black-box optimisation would converge on the same choice. For the positions in source space, we found the overall amount of variation to be rather small (figure 2.5A). In the worst case (figure 2.5B), the source locations were scattered within a range of 3.06 cm, the median of this range was 0.61 cm (figure 2.5C), only slightly above the amount we allowed the participants to move in the scanner. In the best case, the range was only 0.23 cm.

We were also interested if our optimisation would consistently pick distinct locations in source space for different feature spaces. To evaluate this, we computed the silhouette index. As a measure of the consistency of a clustering, it relates the similarity of data within a given class to the similarity of data outside of that given class and is bound between -1 and $+1$. For the optimised source positions of each outer fold o of the set of outer folds O and each feature space f of the set of feature spaces F, we computed the silhouette index $s(o_f)$ using the following formula:

$$s(o_f) = \frac{b(o_f) - a(o_f)}{max(a(o_f), b(o_f))} \tag{2.3}$$

Here, $a(o_f)$ denotes the average euclidean distance between the source position chosen in the outer fold o and the source positions chosen in $O \setminus o$ for that feature space f, while $b(o_f)$ refers to the minimum of average distances between the source position chosen in the outer fold o for feature space f and source positions chosen for all outer folds in O for all feature spaces in $F \setminus f$.

Across feature spaces and hemispheres, we found results that were mostly inconsistent across participants (figure 2.5D). Specifically, we observed participants for whom the assignment of chosen source positions to feature spaces was appropriate as reflected by silhouette indices close to 1, but also participants for whom this assignment was inappropriate as reflected by silhouette indices close to -1. In sum, on a group level and across feature spaces, there was no clear relationship between the choices of positions in source space and the feature space used to model the MEG responses. Overall, this suggests that while there was no direct and robust mapping of feature spaces to source positions, the optimisation of the source positions tended to converge on relatively small regions within a participant.

The choices of optimal hyperparameters for the beamformer spatial filter did not differ substantially across feature spaces (figure 2.5E). While we observed a relatively high degree of variance in optimal choices across participants, we found the choices to be relatively consistent within one hemisphere of a participant and across feature spaces as indicated by relatively high intra-class correlation coefficients across participants with

outer folds and feature spaces as different measurements for the left (0.96) and right (0.88) hemispheres. However, we observed a pronounced difference between left and right ACs, with a higher level of regularisation for the left AC. Here, in some cases the optimal values even bordered on the boundaries we chose for the hyperparameter, suggesting that in some cases, even higher values could have been optimal.

For the temporal extent, we found that this optimisation resulted in characteristic temporal extents for each feature (sub-)space (figure 2.6A). For example, for the combination of articulatory features and log-mel spectrograms the optimisation algorithm consistently found shorter temporal extents for the articulatory features than for the log-mel spectrogram. Pooled across participants, we observed very similar patterns in left and right ACs.

For the $L2$ regularisation, we again found that the optimisation found characteristic values to be optimal for each feature (sub-)space. Specifically, for lower dimensional feature (sub-)spaces the amount of $L2$ regularisation seemed to be less critical, yielding flat distributions. However, for higher- dimensional (sub-)spaces, a higher value of regularisation seemed to be beneficial (figure 2.6B). This was especially the case for the combination of articulatory features and the log-mel spectrogram, for which the distributions for the two subspaces clearly differ.

2.5.5 Model comparisons

Bayesian Hierarchical Modeling of performances

In an initial evaluation of the encoding models, we wanted to statistically compare the predictive performance from models using different feature spaces, obtained from multiple participants. Similar situations often arise in neuroimaging and are usually complicated by small raw effect sizes across conditions in the presence of much larger between subject variability. A promising way to address this is provided by hierarchical models, which allow to maintain sensitivity to effects of interest in these cases. To evaluate the model performances r in both hemispheres h for each outer fold b of all participants i and focus on the differences between the m different feature spaces f, we used a Bayesian hierarchical model with a zero intercept, participant-independent and participant-specific effects for each feature space as well as effects specific to each combination of participants and folds, participants and hemispheres as well as hemispheres and feature spaces. This allowed us to assess posterior distributions of the beta estimates of the means of each level of the categorical variable feature space. To implement this model, we used the brms package (Bürkner, 2017) within the R computing environment (R Core Team, 2013). Specifically, the chosen package implements a user-friendly interface to set up Bayesian hierarchical models using stan (Stan Development Team, 2020). We used Markov chain Monte- Carlo sampling with four chains of 4000 iterations each, 1000 of which were used for their warmup. The priors for standard deviation parameters were not changed from the default values, i.e. half-student-t distributions with 3 degrees of

freedom, while we used weakly informative normal priors with a mean of 0 and a variance of 10 for the effects of individual feature spaces. The model can be described with the following formula:

$$
\begin{aligned}
r_n &\sim \mathcal{N}(\mu_n, \sigma^2) \\
\sigma &\sim |t(3,0,10)| \\
\mu_n &\sim \beta_{i:f[n]} + \beta_{i:b[n]} + \beta_{i:h[n]} + \beta_{h:f[n]} + \beta_{f_{f_1}[n]} + \cdots + \beta_{f_{fm}[n]} \\
(\beta_{i:f[n]}, \beta_{i:b[n]}, \beta_{i:h[n]}, \beta_{h:f[n]}) &\sim \mathcal{N}(0, \sigma^2_{\beta_{int}}) \\
\sigma_{\beta_{int}} &\sim |t(3,0,10)| \\
\beta_{f_{f_1}[n]} &\sim \mathcal{N}(0,10) \\
&\vdots \\
\beta_{f_{fm}[n]} &\sim \mathcal{N}(0,10)
\end{aligned}
$$

To compare the resulting posterior distributions for several parameter combinations of interest, we evaluated the corresponding directed hypotheses using the brms package: $\beta_{f_a} - \beta_{f_b} > 0$, for all possible pairwise combinations of feature spaces, and obtained the ratio of samples of the posterior distributions of differences that were in line with the hypothesis.

Partial Information Decomposition

Besides directly comparing the raw predictive power of models across feature spaces, we were also interested in characterising the detailed structure of predictive information carried in the different feature spaces. Since we were particularly interested in discovering to what degree the contributions of the annotated feature spaces can be explained with contributions of acoustic feature spaces, we thus asked to what degree their predictions contained the same information about the observed MEG (redundancy, or shared information) or to what degree their contributions were distinct (unique information). In information theory, this is possible within the framework of Partial Information Decomposition (PID, Williams & Beer, 2010; Wibral et al., 2015). This can be seen as a further development of the concept of interaction information (McGill, 1954) or co-information (defined equivalently but with opposite sign). Considering the case where we have two source variables (for example test set predictions from different models, $\hat{r}_{M1}$ and $\hat{r}_{M2}$) and a single target variable (for example the observed test set MEG time course, r), co-Information can be thought of as the set intersection of the two source-target MI values (i.e. the predictive information common to the two considered models). It is calculated as the difference between the sum of the individual source-target MIs and the full joint

MI when considering both sources together:

$$CoI = MI(\hat{r}_{M1}, r) + MI(\hat{r}_{M2}, r) - MI([\hat{r}_{M1}, \hat{r}_{M2}], r) \tag{2.4}$$

If both sources provide the same information about the target then

$$CoI = MI(\hat{r}_{M1}, r) = MI(\hat{r}_{M2}, r) = MI([\hat{r}_{M1}, \hat{r}_{M2}], r) \tag{2.5}$$

which quantifies in this case fully redundant overlap in information content. However, it is possible that

$$MI([\hat{r}_{M1}, \hat{r}_{M2}], r) > MI(\hat{r}_{M1}, r) + MI(\hat{r}_{M2}, r) \tag{2.6}$$

This results in a negative value for co-information and this sort of super-additive predictive effect is termed synergy.

Crucially however, co-information measures only the difference between redundancy and synergy, i.e. a net effect (Williams & Beer, 2010). In the presence of equally strong synergistic and redundant contributions, co-information is zero. Therefore, co-information does not provide a way to quantify information provided uniquely by a single source.

The PID framework provides a solution to this problem. We used a recent implementation based on common change in surprisal (I_{CCS}, Ince, 2017a) which has previously been applied within a neuroimaging context (Park et al., 2018). The crucial step in a PID is to quantify redundancy, since once this is done, the other quantities (unique information and synergy) can then be inferred via a lattice structure (Williams & Beer, 2010). For the redundancy measure I_{CCS}, pointwise co-information is considered.

MI can be quantified at the pointwise level (i.e. at specific values of the underlying variables): MI is defined as the expectation of pointwise MI (PMI) over all values of both variables and is non-negative. PMI on the other hand is a signed quantity. When it is positive it indicates those two particular values of the considered variables are more likely to occur together than would be expected if the variables were independent. When it is negative, it indicates that those two particular values are less likely to co-occur than in the independent case. Positive PMI can be interpreted as redundant entropy, while negative PMI is synergistic entropy (Ince, 2017b). Negative PMI values have also been termed misinformation (Wibral et al., 2015), since they correspond to a case where a Bayes optimal gambler who was betting on the outcome of one variable based on observation of the other would actually do worse (on that particular observation) than if they ignored the observation.

In regression terms, negative PMI relates to values that, were they to occur in the data, would have large absolute residual from the regression line (i.e. deviate from the overall relationship), while positive PMI occurs for values that would be close to the regression line (i.e. following the overall relationship). Similarly, pointwise co-information can be considered as quantifying the set theoretic intersection of PMI values from two

sources. Two conditions have to be fulfilled in order for a pointwise co-information term to contribute to I_{CCS} redundancy: (I) both sources have PMI about the target with the same sign and (II) the pointwise co-information of these three variables is of the same sign as the two PMI values. This allows to quantify pointwise contributions of the sources about the target which can be unambiguously interpreted as redundant or overlapping contributions. A crucial advantage of this redundancy measure as opposed to other PID implementations is that it measures the overlap at the pointwise level and therefore can be interpreted as a within sample measure of redundant prediction, directly linked to the decoding interpretation of MI. This is essential for the comparison of predictive models as we consider here, for which redundancy measures which ascribe redundancy to sources even when they predict the target on disjoint sets of samples would be inappropriate (Ince, 2017a).

This implementation of PID does not provide a non-negative decomposition. For example, negative unique information values are possible and they reflect a situation where there are pointwise misinformation terms that are unique to one source-target relationship (Ince, 2017a, see Table 7). In our application, negative unique information means there are time periods where one model mis-predicts, i.e. that combination of model prediction and MEG values is less likely to occur than if the model and prediction were shifted randomly, while the second model does not. In other words, there is a time window where that model is uniquely unhelpful for predicting the MEG signal, even though, of course, on average over time, it does have predictive value. In cases where there is negative unique information in the predictions of one model whose marginal MI about the MEG values is being used to normalise the redundancy values, it is therefore possible to obtain normalised redundancy ratios > 1.

We here performed PIDs for each combination of outer fold predictions of the annotated feature space with those of the acoustic feature spaces as sources and the recorded MEG as targets. Critically, we retrained all models with fixed hyperparameters of regularisation of sensor covariance matrices and coordinates in source space to those previously chosen as optimal in the inner folds when training the model based on the *Sg&Art* feature space. This way, we gave the *Sg&Art* feature space the best chances to achieve maximal unique information. To compute the respective information theoretic quantities with these continuous variables, we transformed the variables to be standard normal while preserving rank relationships by calculating the empirical cumulative density function (CDF) value at each data point and applying the inverse standard normal CDF (Ince et al., 2017) prior to running I_{CCS} PIDs for Gaussian variables via Monte Carlo integration (Ince, 2017a). To interpret the raw values of the PIDs, we divided them by the marginal MI of the benchmark articulatory feature space. The normalised redundancy then represents the proportion of the predictive information of the benchmark model which is available also from the tested acoustic feature model.

To evaluate the results across folds, hemispheres, participants and feature spaces, we used Bayesian models similar to those used for the evaluation of the performances.

The corresponding model can be described as follows:

$$
\begin{aligned}
\frac{red_n}{MI_n} &\sim \mathcal{N}(\mu_n, \sigma^2) \\
\sigma &\sim |t(3,0,10)| \\
\mu_n &\sim \beta_{i:f[n]} + \beta_{i:b[n]} + \beta_{i:h[n]} + \beta_{h:f[n]} + \beta_{f_1[n]} + \cdots + \beta_{f_{m-1}[n]} \\
(\beta_{i:f[n]}, \beta_{i:b[n]}, \beta_{i:h[n]}, \beta_{h:f[n]}) &\sim \mathcal{N}(0, \sigma^2_{\beta_{int}}) \\
\sigma_{\beta_{int}} &\sim |t(3,0,10)| \\
\beta_{f_1[n]} &\sim \mathcal{N}(0,10) \\
&\vdots \\
\beta_{f_{m-1}[n]} &\sim \mathcal{N}(0,10)
\end{aligned}
$$

For the ratios of unique information, we concatenated the unique information of both competing sources x and y in all comparisons to a single response variable and changed the modelling approach to include predictors for unique information of both sources in all $m-1$ comparisons.

$$
\begin{aligned}
\frac{unq_n}{MI_n} &\sim \mathcal{N}(\mu_n, \sigma^2) \\
\sigma &\sim |t(3,0,10)| \\
\mu_n &\sim \beta_{i:f[n]} + \beta_{i:b[n]} + \beta_{i:h[n]} + \beta_{h:f[n]} + \\
&\quad \beta_{unqx_{f_1}[n]} + \cdots + \beta_{unqx_{f_{m-1}}[n]} + \beta_{unqy_{f_1}[n]} + \cdots + \beta_{unqy_{f_{m-1}}[n]} \\
(\beta_{i:f[n]}, \beta_{i:b[n]}, \beta_{i:h[n]}, \beta_{h:f[n]}) &\sim \mathcal{N}(0, \sigma^2_{\beta_{int}}) \\
\sigma_{\beta_{int}} &\sim |t(3,0,10)| \\
\beta_{unqx_{f_1}[n]} &\sim \mathcal{N}(0,10) \\
&\vdots \\
\beta_{unqx_{m-1}[n]} &\sim \mathcal{N}(0,10) \\
\beta_{unqy_{f_1}[n]} &\sim \mathcal{N}(0,10) \\
&\vdots \\
\beta_{unqy_{f_{m-1}}[n]} &\sim \mathcal{N}(0,10)
\end{aligned}
$$

The resulting values of synergy were very low. We thus wanted to assess to which degree the observed synergy could only be obtained with intact predictions from the benchmark articulatory features, or to which degree it could also be observed when the benchmark's predictions were randomly permuted. We performed circular shifts of the

predictions based on the *Sg&Art* features by a random number of samples, where the random number was constrained to be at least 200 samples and maximally the number of available samples minus 200 samples to avoid temporal autocorrelation. We computed PIDs of 1000 of these permutations. We then defined noise thresholds as the 95th percentile of the 1000 maximum values found in permutations across feature spaces, sources and outer fold test sets and calculated the fraction of data points (outer fold test sets, sources) within each participant and feature space. To also compare unique information of both sources and redundancy values to such noise thresholds, we repeated this process, shuffling predictions based on the *Sg&Art* features for thresholds for the information unique to predictions based on the *Sg&Art* features and shuffling observed MEG time series for redundancy and information unique to predictions based on the competing feature spaces.

Phoneme-evoked dynamics

A recent study reported that epoching EEG recordings from a story-listening paradigm according to the onsets of phonemes allowed the decoding of four classes of phonemes, so-called manners of articulation, from the resulting event-related potentials (Khalighinejad et al., 2017). We aimed to firstly replicate this finding with our MEG data and secondly assess to which degree our linear encoding models could account for this phenomenon.

We computed "Phoneme-Related Fields" (PRFs) using the 34562 phoneme presentations we had previously identified in our stimulus material. For this, we mapped the set of phonemes to manners of articulation as specified by Khalighinejad et al. (2017, see Figure S6 for a mapping table): Plosives, fricatives, nasals and vowels. We then epoched the continuous MEG data for a time range from -0.1s – $+0.6$s around phoneme onsets, binned it across epochs for each time point using four equipopulated bins and computed mutual information between the MEG data and the four manners of articulation.

To ensure that we would capture the maximum effect of the MI, we delegated the choice of source positions for the left and right hemispheres as well as sensor covariance regularisations to the BADS algorithm similarly as before (figure 2.9). However, this time we optimised the source model parameters with respect to the sum of MI of observed MEG data about the phoneme classes across time points. We then retrained our encoding models with the source model parameters fixed to these choices. To assess the results of this optimisation, we recalculated the maximum distance metric used in the assessment of the chosen source positions during our modelling, this time also including the positions found for optimal phoneme class decoding and plotted the difference to the previously obtained maximum differences (figure 2.9A). The results reflected that still, all positions lay in STG, while for some participants, the positions found to be optimal for the PRF analysis were different from those obtained during the modelling.

Subsequently, we performed the same PRF analysis on the outer fold predictions of each feature space. We were then interested in the redundant and unique contributions of observed and predicted MEG to the MI about manners of articulation. We

thus performed PIDs with observed and predicted PRFs as sources and the manners of articulation as the target, separately for each feature space, yielding phoneme-related redundancy as well as unique profiles.

2.5.6 Analysis of EEG dataset

To assess to which degree our main findings would generalise from our MEG to EEG data, we also performed an analysis of an openly available EEG story listening dataset (Broderick et al., 2018a). This dataset is part of the data on which the effect of a gain of prediction performance of the combination of spectrograms and articulatory features over spectrograms alone was originally reported.

EEG preprocessing

We analysed the 128 channel EEG recordings of a duration of 1 hour and 29 seconds of 13 participants. They had been acquired in 20 blocks of approximately equal duration at a sampling rate of 512 Hz using a BioSemi ActiveTwo system and downsampled to 128 Hz. We rereferenced the data to the average of two additional mastoid reference channels, spline interpolated noisy channels identified by visual inspection (mean number of noisy channels: 3.29 standard deviation: 4.19, pooled across participants), applied a fourth order forward-reverse butterworth high-pass filter with a cutoff frequency of $.5$ Hz and attenuated strong transient artefacts identified by visual inspection with a hamming window to have an absolute amplitude of 90% of the maximum of the absolute clean signal. Next, we z-scored individual blocks and winsorized the time series by replacing remaining artefacts with an amplitude stronger than ± 3 standard deviations by ± 3 and concatenated the individual blocks to single datasets. We then found unmixing matrices using the *runica* ICA algorithm. We identified artefactual components reflecting eye or heart activity and backprojected the unmixed data using mixing matrices where the artefactual components were removed. Finally, we downsampled the data to a sampling rate of 40 Hz.

Stimulus Transformations

In general, we reused the same pipeline to generate non-linear transformations of the stimulus as we had used for the stimulus of our MEG dataset. However, due to the high noise level of the EEG data, we decided to omit the high-dimensional Gabor feature space and focussed on assessing if the acoustic feature space found to explain the performance gain of the benchmark articulatory features over spectrograms alone in the MEG dataset could do so in the EEG data as well. Additionally, we were interested in more faithfully reproducing the original results (Di Liberto et al., 2015), where a spectrogram different from the log-mel spectrogram employed here had been applied. To do so, we generated a bank of 16 fourth order zero-phase butterworth bandpass filters

with mel-spaced centre frequencies (250, 402, 577, 780, 1015, 1288, 1605, 1971, 2396, 2888, 3459, 4121, 4888, 5777, 6807 and 8001 Hz), where the cutoff frequencies were defined as half of the distances to the neighbouring centre frequencies. The absolute values of the Hilbert transform of the output of these filters served as an approximation to the spectrogram used in the original publication ("*Sg16*"). Moreover, we were interested to which degree possible differences between the performances achieved with this spectrogram compared to our log-mel spectrogram were attributable to a compressive nonlinearity (Biesmans et al., 2017) included in the latter. We therefore generated an additional spectrogram ("*Sg16c*") where we raised the values of *Sg16* to the power of .3. This gave us a set of feature spaces $F_{EEG} = \{$*Env*, *Sg16*, *Sg16c*, *Sg*, *Sg16&(Sg16')*$_+$, *Sg16c&(Sg16c')*$_+$, *Sg&(Sg')*$_+$, *Sg16&PhOn*, *Sg16c&PhOn*, *Sg&PhOn*, *Sg16&Art*, *Sg16c&Art*, *Sg&Art*, *Sg&(Sg')*$_+$*&Art* $\}$.

Forward Modelling

To keep the results comparable to the original publication, we performed ridge regressions to model responses at the 12 electrodes whose performances were reported in the main result of the original publication (B28, B29, B30, C3, C4, C5, D3, D4, D5, D10, D11, D12) using the function "mTRFcrossval.m" from the mTRF toolbox (Crosse et al., 2016). However, we implemented a small change that allowed us to do a nested crossvalidation to tune the regularisation hyperparameter λ_{L2}. We trained models on 18 of the 20 available blocks, picked the λ_{L2} that resulted in the best prediction performance on a validation block and evaluated the test performance on the remaining block. This procedure was rotated such that each block served as the test set once. We specified the range of λ_{L2} values as $\{0.1k \mid k\varepsilon[-25...60]\}$, over which the function performed an exhaustive grid search where the extreme values were never chosen. For the parameters of temporal extent, we used the same values as in the original publication, i.e. $t_{Min} = -0.1$ and $t_{Max} = 0.4$ seconds.

Model Comparisons

The model comparisons employed here were largely the same as for the MEG data. To evaluate the test set prediction performances of the forward models, we used the same Bayesian modelling approach as we had used for the analysis of the MEG data. We also performed the same PID analysis with model predictions as sources and observed EEG time-series as targets and evaluated its performances with the same Bayesian models as we had used for the MEG analysis. However, due to the higher noise level, we only considered data points where the MI of predictions based on the benchmark articulatory feature space and observed time-series surpassed a noise threshold defined as the 95th percentile of MI values obtained from time shifted permutations, corrected across electrodes using maximum statistics. Additionally, to account for the skewed distributions of the ratios of PID quantities normalised by the marginal MI of the predictions based on

the benchmark articulatory feature space and the observed MEG, we used a log-normal response family for the Bayesian modelling and considered the posterior distributions of the medians of the effects of interest. We repeated the computation of noise thresholds as described for the MEG data. Finally, we performed the same analysis of phoneme evoked responses on the set of electrodes used for the modelling as we had performed on the MEG data.

2.6 Acknowledgments

CD is funded by the College of Science and Engineering at the University of Glasgow; JG received support from the Wellcome Trust (UK; 098433). We thank Moritz Boos, Jan-Mathijs Schoffelen, Dale Barr and Christoph Scheepers for helpful discussions.

Chapter 3

A whitening approach for Transfer Entropy permits the application to narrow-band signals

published as:
Daube, C., Gross, J. and Ince, R.A.A. (2022). A whitening approach for Transfer Entropy permits the application to narrow-band signals. *arxiv*, CC-BY license.

Author Contributions:
C.D., J.G. and R.A.A.I. conceived of and designed the experiment.
C.D. collected and analyzed the data.
C.D. and R.A.A.I. contributed analytic tools.
C.D. wrote the manuscript.
C.D., J.G. and R.A.A.I. edited the manuscript.

3.1 Abstract

Transfer Entropy, a generalisation of Granger Causality, promises to measure "information transfer" from a source to a target signal by ignoring self-predictability of a target signal when quantifying the source-target relationship. A simple example for signals with such self-predictability are narrowband signals. These are both thought to be intrinsically generated by the brain as well as commonly dealt with in analyses of brain signals, where band-pass filters are used to separate responses from noise. However, the use of Transfer Entropy is usually discouraged in such cases. We simulate simplistic examples where we confirm the failure of classic implementations of Transfer Entropy when applied to narrow-band signals, as made evident by a flawed recovery of effect sizes and interaction delays. We propose an alternative approach based on a whitening of the input signals before computing a bivariate measure of directional time-lagged dependency. This approach solves the problems found in the simple simulated systems. Finally, we explore the behaviour of our measure when applied to delta and theta response components in Magnetoencephalography (MEG) responses to continuous speech. The small effects that our measure attributes to a directed interaction from the stimulus to the neuronal responses are stronger in the theta than in the delta band. This suggests that the delta band reflects a more predictive coupling, while the theta band is stronger involved in bottom-up, reactive processing. Taken together, we hope to increase the interest in directed perspectives on frequency-specific dependencies.

3.2 Introduction

Over the last decades, the description of statistical dependencies in cerebro-cerebral and cerebro-peripheral pairs of time series has witnessed a surge of interest (Bassett & Bullmore, 2006; Brookes et al., 2011; Naselaris et al., 2011; Crosse et al., 2016; Mell et al., 2021; Gross et al., 2021). In these fields, the general idea is to gain insight into the workings of the brain by either studying how time series of neuronal activity relate to other neuronal activity or to external signals such as auditory or visual stimuli as well as the activity of other organs.

As a consequence, countless methodological approaches have been suggested to mathematically quantify these dependencies (Bastos & Schoffelen, 2016). Some of these ideas specifically aim at the description of directed interactions, for example by using measures of the so-called "Granger-causal" (Granger, 1969) family, or their generalisation to nonlinear relationships, Transfer Entropy (Schreiber, 2000; Barnett et al., 2009). In both of these measures, the main idea is to quantify a directional dependency by first assessing to what degree a target time-series can be predicted from itself and secondly assessing to what degree this auto-prediction can be improved upon with the assumed source time-series. In its classic formulation, TE implements this by way of conditioning the mutual information (MI) between source and target on an operationali-

sation of the target past. It has been suggested that this warrants the capacity to correctly estimate not only "predictive information transfer" between source and target time-series, but also the recovery of the true underlying interaction delay (Wibral et al., 2013). The precise estimation of such quantities is of high interest for the research programmes not only of functional connectivity, but of cognitive neuroscience in general.

However, in the arguably simplest and in many applications ubiquitous case of self-predictable or auto-correlated time-series, namely narrowband time-series as obtained e.g. when applying band-pass filters, TE fails to deliver intuitively comprehensible results. The application of TE in such cases has therefore repeatedly been discouraged (Florin et al., 2010; Barnett & Seth, 2011).

Such applications however are of potentially high interest, given that frequency specific interactions are at the core of popular hypotheses about cerebro-peripheral (Giraud & Poeppel, 2012; Donhauser & Baillet, 2020) and cerebro-cerebral (Schnitzler & Gross, 2005; Fries, 2015; Michalareas et al., 2016; Schoffelen et al., 2017) interactions. There is an abundance of evidence for intrinsically auto-correlated or band-limited parts of neuronal activity ("oscillations"), whose presence in neuronal recordings (Wang, 2010; Donoghue et al., 2020) should accordingly impede the use of TE even without the use of analysis filters. Moreover, in light of the usually low signal-to-noise ratio (SNR) of many recording modalities, particularly of non-invasive neuroimaging, the isolation of band-limited activity via spectral filtering is a pervasive strategy to achieve acceptable sensitivity and specificity. The main idea of TE thus turns out to be an empty promise for many real-world applications where auto-correlations are indeed clearly visible.

Suggestions to implement TE mainly differ in their approaches to remove the self-predictability of the target signal from the quantification of the effect. While it has been argued previously that simplistic approaches relying on a target past operationalisation consisting of a single delay are insufficient (Wibral et al., 2013), such approaches remain popular. This might imply that the literature is missing more intuitive demonstrations of the shortcomings of such approaches. Furthermore, it also has been shown previously that prominent proposals relying on multidimensional embeddings of the target time series fail in scenarios of narrow-band effects (Wollstadt et al., 2017). While suggestions exist that try to overcome this problem by means of constructing frequency-specific surrogate data (Pinzuti et al., 2020) or state-space models (Faes et al., 2017), more intuitive explanations of these failure cases are arguably still lacking. Such an intuitive understanding should pave the way towards both more widespread awareness as well as simple fixes of these issues.

One interesting use case for measures that overcome the problems outlined above is the heavily studied phenomenon of speech envelope tracking as observed in magneto- and electroencephalography (henceforth MEEG to denote both modalities) recordings (Ahissar et al., 2001; Hertrich et al., 2012; Gross et al., 2013; Ding & Simon, 2012; O'Sullivan et al., 2015; Di Liberto et al., 2015; Wöstmann et al., 2017; Brodbeck et al., 2018a; Daube et al., 2019b; Obleser & Kayser, 2019; Zan et al., 2020; Donhauser &

Baillet, 2020). In short, the low-frequency portion of MEEG signals is reliably related to the time varying energy of the speech signal at a certain delay. In noninvasive recordings, speech envelope tracking is heavily studied in the canonical delta and theta bands (Ding & Simon, 2014). At higher frequencies, the effect usually fails to robustly exceed noise thresholds (but see Kulasingham et al., 2020). Limiting the analysis to the delta and theta frequency bands is thus an efficient means to achieve stronger effects. Theories exist that link this phenomenon to frequency specific mass-neuronal processes of debated algorithmic significance (Giraud & Poeppel, 2012; Hyafil et al., 2015).

From the perspective of directed connectivity, an interesting question is to what degree the resemblance of MEEG signals and the speech envelope at a given time t can be explained from the MEEG signal's own past, and to what degree it can only be explained when additionally considering the stimulus. Given the auto-correlated nature of both neuronal processes and the speech envelope in the relevant spectral range, it is in theory possible for a system to predict the upcoming speech envelope (and, as made evident by the progress of past and current machine learning approaches, also richer parts of the speech signal Chung et al., 2020; Lakhotia et al., 2021). According to popular theories of brain function, such predictive coding is also of high utility for a biological system (Rao & Ballard, 1999; Friston, 2005). The extent to which speech tracking as quantified by undirected measures such as delayed MI would decrease when using suitable directed measures would highlight the extent to which the heavily studied tracking might in fact reflect predictive rather than reactive processing. This would add to accounts that characterise low-frequency oscillations as the deliberate effort of biological systems to be in a state of optimal neuronal excitability (Lakatos et al., 2008; Henry & Obleser, 2012), such that metabolically costly states of high encoding fidelity co-occur with the relevant parts of the stimulus (Jones & Boltz, 1989; Schroeder & Lakatos, 2009; Kayser et al., 2015; Młynarski & Hermundstad, 2018).

Here, we consider a simple simulated system of band-limited delayed bivariate interactions with a clear and intuitively comprehensible ground truth spectral range and delay. We implement delayed MI, two classic algorithms to estimate TE as well as a novel, whitening based estimator ("Directed Information based on conditional entropy", "DI_{ce}") within the Gaussian copula MI framework (gcmi, Ince et al., 2017) and extensively test these measures with simulations. Finally, we explore the behaviour of DI_{ce} in an MEG dataset of continuous speech listening. We find that estimates of the delay between the stimulus and the response as well as the recovered interaction strength in the delta and theta bands differ from those recovered by the bivariate delayed MI.

3.3 Results

3.3.1 An intuitive overview over multiple measures

A first goal of this study was to provide an overview of various implementations of TE in a simple and intuitively understandable example case. To do so, we simulated coupled systems with a known ground truth spectral interaction range and delay. These consisted of 4 - 8 Hz narrow-band filtered Gaussian white noise to obtain a band-limited source signal, which was delayed in time by .15 s to obtain a target signal. We added Gaussian white noise to these signals to mimic the noisy measurement process in neuroimaging. For such a simplistic but intuitive system, a suitable measure of TE should peak at the ground truth simulated delay and spectral range. Further, for such highly auto-correlated signals, TE should yield a highly reduced effect size in comparison to undirected measures such as delayed MI: a potential source signal can only add little information if the target signal is already highly predictable from its own recent history.

Figure 3.1 shows the results of applying delayed MI, classical TE estimators TE_{1D} and TE_{SPO} as well as our proposal, DI_{ce} (see below for more detailed explanations of each individual measure) to the same time series, both without and with applying an analysis filter.

A first observation is that in the present implementation with gcmi (Ince et al., 2017), all measures yield increased effect sizes when a suitable analysis filter is applied. This, potentially trivially (but see Pinzuti et al., 2020), demonstrates that the effect size and consequently the sensitivity can benefit from the application of analysis filters. It thus underscores the utility of directed connectivity measures that behave robustly when applied to narrow-band signals. Further, the pass-band of the effect can be found by applying analysis filters of varying centre frequencies: all measures return the highest effect sizes across analysis bands within the pass-band of the simulated effect.

Secondly, we observe that TE_{1D} fails to return a reduced effect size in comparison to the undirected delayed MI, even when applied to highly self-predictable signals as simulated here. This happens for both broadband and filtered analysis scenarios. TE_{1D} is a classical implementation of TE (Besserve et al., 2010; Lobier et al., 2014; Ince et al., 2015; Park et al., 2015; Giordano et al., 2016; Morillon & Baillet, 2017) that operationalises the target past by means of using only one delay of the target time series (more specifically, the same delay as that of the source signal relative to the target present when scanning across delays). For the ground truth delay simulated here, this single delay fails to capture most of the self-predictability of the target signal, resulting in an overestimation of the directed effect.

A third measure, TE_{SPO} (Wibral et al., 2013), yields the anticipated strong reduction in effect size by roughly an order of magnitude in comparison to MI. It achieves this by using a more effective handle on the target past: a multi-dimensional “non-uniform” (Vlachos & Kugiumtzis, 2010; Faes et al., 2011) embedding optimised for self-prediction. As opposed to TE_{1D}, this essentially consists of multiple delays that are independent

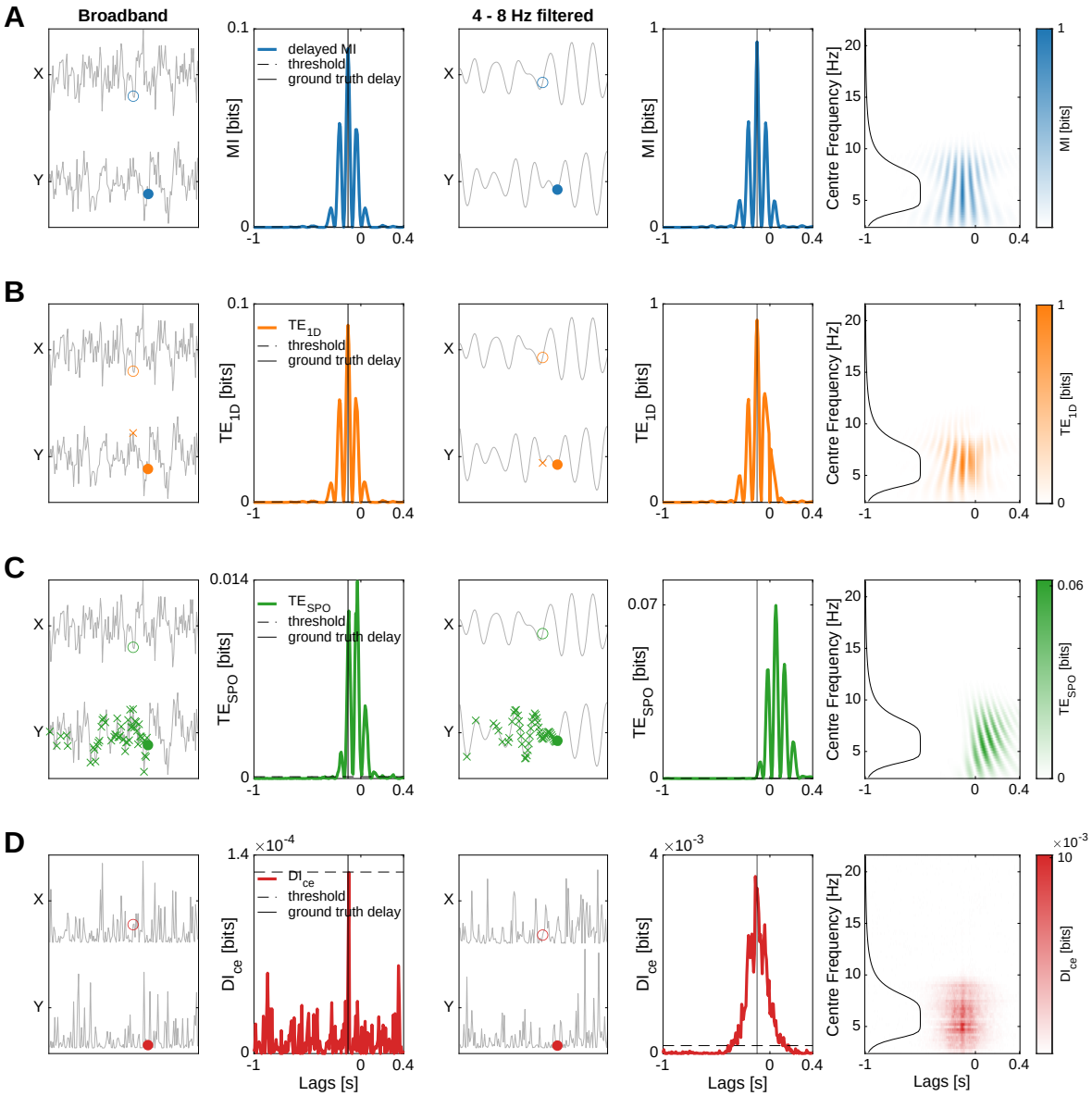

Figure 3.1: **Comparison of delay profiles of various undirected and directed dependency measures in a simplistic simulation scenario.**
The simulation here consists of a 4 - 8 Hz bandpass filtered white noise source signal that is temporally delayed to obtain the target. White noise is added to both source and target to model measurement noise. First and middle columns show time domain sections of the analysed time series, where a filled circle denotes a target present sample, an empty circle denotes a source past sample at a delay corresponding to the simulated effect, and crosses denote the respective target past samples. 2nd and 4th columns show delay profiles of the respectives measures relative to the ground truth interaction delay and noise thresholds. 5th column shows spectrotemporal maps, where the frequency response of the filter used to generate the ground truth effect is overlaid as a black curve. **A** Delayed Mutual Information correctly recovers the ground truth interaction delay. **B** TE_{1D} also correctly recovers the interaction delay, but measures an interaction effect that is highly similar to delayed MI. **C** TE_{SPO} recovers an interaction delay different from the ground truth, even more so when the analysis is performed on the filtered signal. **D** DI_{ce} only finds a super-threshold effect with correct recovery of the interaction delay when an analysis filter is used, but with a strongly reduced effect size in comparison to other measures.

of the analysis delay in the scanning procedure. However, it fails to recover the ground truth interaction delay simulated in the coupled systems (Wollstadt et al., 2017), in both broadband and filtered analysis settings. Given this failure in a simplistic problem setting that is ubiquitous in neuroscientific datasets, it is unclear to which degree TE_{SPO} can practically live up to its promise of correctly quantifying "predictive information transfer" as well as the interaction delay.

Finally, the last measure considered in and proposed by this study, DI_{ce}, solves both of these problems: It returns the smallest effect sizes in this comparison, even to the degree that it fails to detect the effect in a broadband analysis setting. However, when a suitable analysis filter is applied, it returns a delay profile with a super-threshold peak at the simulated delay. It achieves this in a two-step approach (Haugh, 1976): First, both source and target signals are transformed into time series of surprisal, i.e. the sample-wise entropy of each time point conditional on the same non-uniform embedding used in TE_{SPO}. Secondly, delayed MI is computed for these whitened time series.

Taken together, this first analysis provides an overview of four different connectivity measures in an intuitive and simplistic simulation setting, illustrating how DI_{ce} succeeds in returning intuitive results where TE_{1D} and TE_{SPO} fail.

3.3.2 Synergy of source and target past about target present distorts conditional MI based TE

We wanted to investigate possible explanations for the counterintuitive results returned by our implementation of TE_{SPO}. To do so, we considered a perspective on TE offered by partial information decomposition (PID). PID is an information theoretic approach to study trivariate relationships (Williams & Beer, 2010; Ince, 2017a). The central goal in PID is to quantify shared information (redundancy) between two source variables about a third target variable. Further, PID aims to measure unique information of both source variables as well as synergistic information that is only obtainable when jointly considering both sources. A key insight of PID relevant to TE is that the basis of TE, conditional MI, is the sum of two "atoms" of PID: unique information and synergy (James et al., 2016). In other words, conditioning a bivariate relationship on a third variable will remove redundant information, but will not deliver solely unique information. It has been pointed out that this conflation of unique and synergistic information defies interpretations of TE as measuring "information flow" (James et al., 2016). By relying on conditional MI, TE measures not just unique information of the source about the target present ignoring the target past, but instead conflates this with synergistic information stemming from interactions of the source and the target past. The resulting quantity is thus not "localisable" to the source (James et al., 2016).

We turned to data from the same simulated system as analysed in the previous section. In a first step, we considered co-information (McGill, 1954) of the source, the target present and the target past. It can be obtained by subtracting conditional MI from bivari-

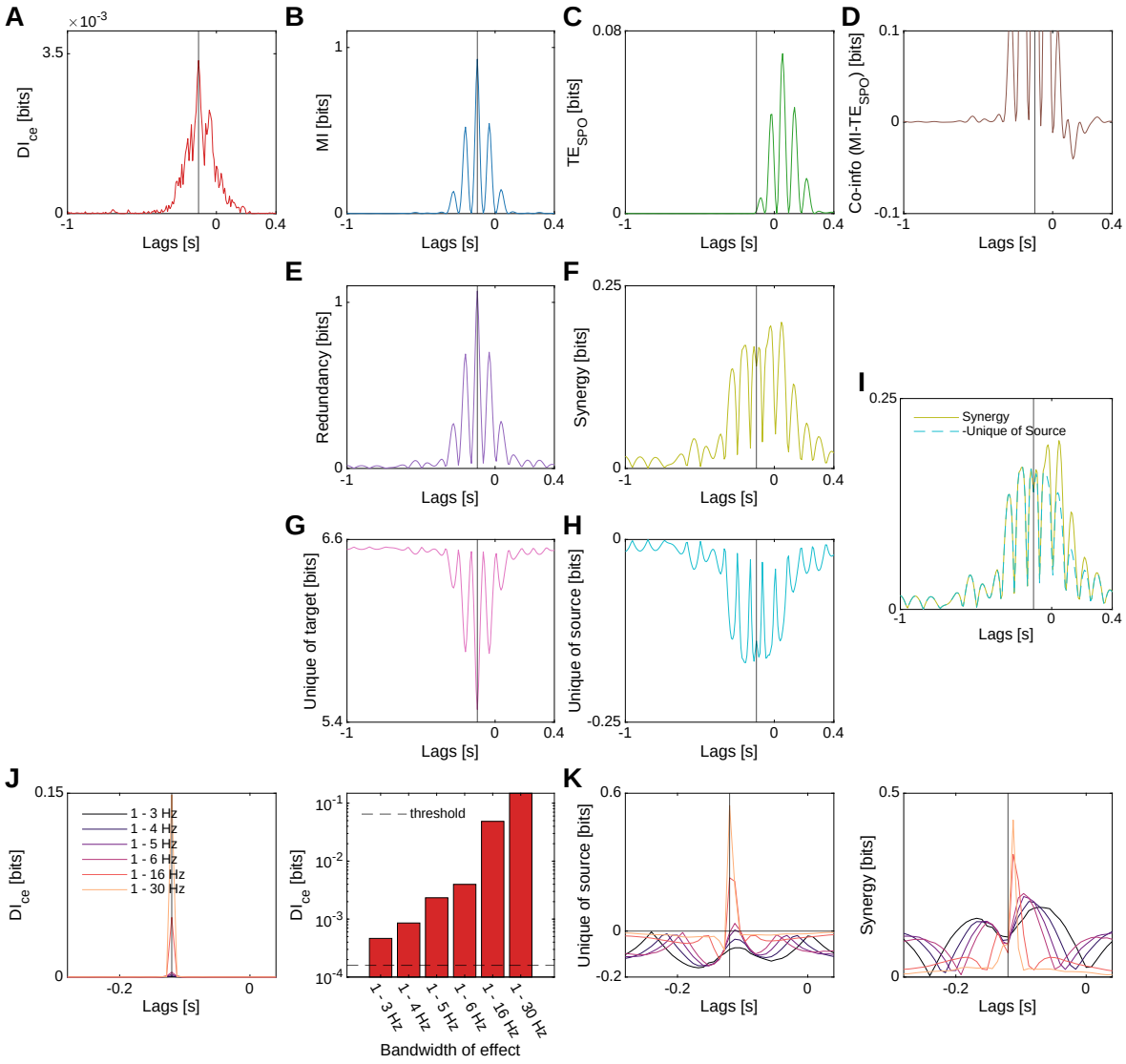

Figure 3.2: **Partial information decomposition perspective on TE.**
A Delay profile of DI_{ce} for reference. **B** Delay profile of delayed MI. **C** Delay profile of TE_{SPO}, which is based on conditional MI and can, according to PID, be seen as the sum of synergy (F) and unique information of the source (H). **D** Delay profile of co-information (abbreviated as “co-info”), which quantifies the triple set intersection in classic Venn-diagram conceptualisations of three variable systems. According to PID, it is the net sum of synergy (F) and redundancy (E). Axis limits chosen to highlight the negative (net synergistic) portion at positive delays. **E - H** Delay profiles of the PID atoms redundancy, synergy and unique information (of target and source). **I** Delay profiles of synergy and sign-flipped unique information of the source, highlighting a surplus of synergy at positive delays. **J** Delay profiles of DI_{ce} when simulating data with different ground truth effect bandwidths. DI_{ce} recovers the correct delay at all bandwidths (left) and recovers an increasing effect size as a function of bandwidth. It exceeds the noise threshold in all cases. **K** Delay profiles of unique information of the source (left) and synergy (right) for the same effect bandwidths as in J. For narrow bandwidths, unique information of the source fails to become positive and recovers the wrong delay. Further, the sum of unique information of the source and synergy is dominated by synergy, which features a greater delay estimation error than unique information.

ate MI, and thus quantifies the triple set intersection in classic Venn diagram conceptualisations. It can take on positive and negative values, where positive values denote redundancy and negative values denote synergy. It is important to note that from a PID perspective, the conceptually simpler (and less controversial) co-information conflates redundancy and synergy to a single net quantity, which PID aims to decompose into pure redundancy and synergy. Figure 3.2D shows negative co-information values at later delays, demonstrating that there is indeed a net synergy of the source and the target past about the target present.

Next, we applied PID using an implementation based on common change in surprisal (I_{ccs} Ince, 2017a). We find that TE_{SPO} (figure 3.2C), as it is based on conditional MI, can be decomposed into largely identical profiles of synergy (figure 3.2F) and unique information of the source (figure 3.2H) differing only in their sign. The positive net conditional MI (i.e. TE_{SPO}) however turns out to entirely stem from a surplus of synergy at later delays that is not cancelled out by unique information of the source (figure 3.2I). In other words, in this simplistic example case, TE_{SPO} measures a synergistic effect.

As a consequence of the PID perspective on conditional MI based TE implementations, it has been proposed to use the unique information of the source as a more appropriate measure in order to avoid the quantification of synergistic effects (Barrett, 2015). However, in our example case, this quantity is negative across the entire range of considered interaction delays (figure 3.2H), meaning that from considering the source on its own, predictions of the target would become worse. This can be seen as the result of two factors: firstly, in this simulation, the source signal is noisy, and secondly, the simulated effect has a very narrow spectral range and thus highly limited degrees of freedom. As these factors come together, the efficient operationalisation of the target past makes it impossible to improve on its prediction of the target present.

We thus considered similar simulations with varying bandwidths of the ground truth effect, and verified that at broader bandwidths, the unique information of the source becomes positive (figure 3.2K). Of note, the delay recovered by unique information of the source generally has lower deviations from the ground truth than what is recovered by synergy, but this deviation is still non-zero at narrower effect bandwidths. We further find that with increasing bandwidths of the effect, synergy increases as well and peaks at delays closer to the ground truth delay. Crucially however, across all simulated bandwidths, our proposed measure, DI_{ce}, finds super-threshold effects at the correct delay (figure 3.2J).

Taken together, we have shown that the delay mis-estimation problems of TE_{SPO}, when applied to narrow-band signals, stem from the known conflation of unique and synergistic effects in conditional MI based TE implementations. In this example, the effect quantified by TE_{SPO} is in fact dominated by synergy. We further show that a possible alternative, namely unique information of the source as obtained from PID, has both a lower sensitivity and a worse performance in recovering the ground truth interaction delay than our proposed conditional MI-free measure DI_{ce}.

3.3.3 Varying the ground truth interaction delay

After these initial intuitive and single sample based demonstrations of problems of two classic algorithms to quantify TE as well as a suggestion to explain their origin, we were interested in more thoroughly testing the characteristics of the measures. To do so, we performed extensive analyses of the simplistic coupled systems that measured the problems over repeated samples. For each of these samples, we computed the delay profiles of the set of four measures and evaluated the recovered interaction delay as well as the recovered effect size as obtained from the peaks across the delay profile of a given repetition.

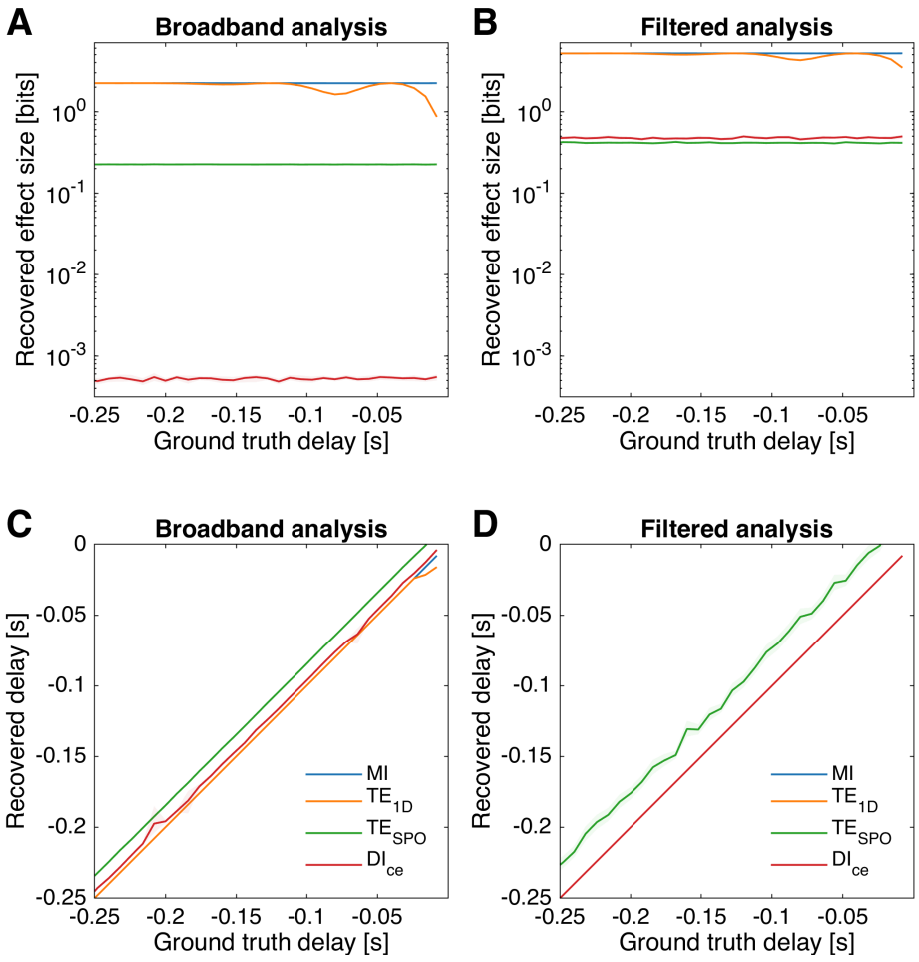

Figure 3.3: **Simulation with varying ground truth delay.**
Basic setup of the simulation is the same as in figure 3.1, however, here, the ground truth interaction delay is varied, and 100 repetitions are sampled. Moreover, a higher SNR is used. Plots show the median across repetitions, shaded regions indicate bootstrapped 95% confidence intervals. **A** Recovered effect size when the data are analysed without a filter. TE_{1D} suffers from recovering a systematically varying effect size across different ground truth delays despite a constant simulated interaction strength. **B** Same as A, but applying an analysis filter. **C** Recovered delays when data are analysed without a filter. TE_{SPO} underestimates the true interaction delay. **D** Same as C, but applying an analysis filter.

In a first simulation, we were interested in the characteristics of the measures across a range of simulated ground truth interaction delays. In principle, a suitable measure of

TE should recover the same constant effect size at varying simulated interaction delays when all other factors are kept constant. In figure 3.3, the results demonstrate that this is indeed the case for all measures in both broadband and filtered analysis scenarios except for TE_{1D}. This measure exhibits a systematic variation in the recovered effect size across different simulated interaction delays reminiscent of the filter ringing of the auto-correlation profile of the target signal. The failure here again highlights the problems of operationalising the target past with a single delay that varies as the analysis delay is scanned (Wibral et al., 2013). In case of very short ground truth delays, the MI at the peak of the delay profile is conditioned on the same short delay of the target variable. Since narrow-band signals have high auto-correlation at such short delays, this leads to a strong reduction of conditional MI vs MI. As the ground truth delay increases, the MI is conditioned on longer delays, where the auto-correlation of the target variable wanes and waxes and thus leaves a correspondingly varying conditional MI. This can lead to potential interpretational pitfalls when for example the results of TE_{1D} obtained in two different experimental conditions are compared. If these conditions simply differ in delay, this will lead to different recovered effect sizes and could thus be falsely interpreted as differences in directed dependency.

Further, the results reiterate that TE_{SPO}, while featuring a constant effect across different ground truth interaction delays, recovers a flawed estimate of the interaction delay. DI_{ce} on the other hand behaves favourably by returning a constant effect size at the correct interaction delay. However, due to its drastically reduced effect size in this band-limited interaction, its estimates are noisier.

3.3.4 Varying the signal-to-noise ratio in source and target signals

In a second simulation, we were interested in comparing the four measures when faced with increasingly noisy signals. While the ideal measure should be highly sensitive to a present effect and recover it at the correct delay, ignoring self-predictable parts of dependencies should inevitably reduce the effect size.

We indeed found that DI_{ce} recovered the smallest effects and hit the noise floor at the lowest noise amplitude in comparison to the other measures (figure 3.4). In a filtered analysis scenario, the recovered effects were in general higher than those in a broadband analysis scenario. Interestingly, for some noise amplitudes, DI_{ce} recovered stronger effects than TE_{SPO}. At very low noise amplitudes, all measures succeeded in recovering the correct delay irrespective of whether filters were applied in the analysis or not. However, as noise increased, especially TE_{SPO} failed in recovering correct delay estimates. While this was especially problematic for filtered analyses, other measures had lower delay estimation errors at higher noise amplitudes due to the gain in sensitivity afforded by filtering.

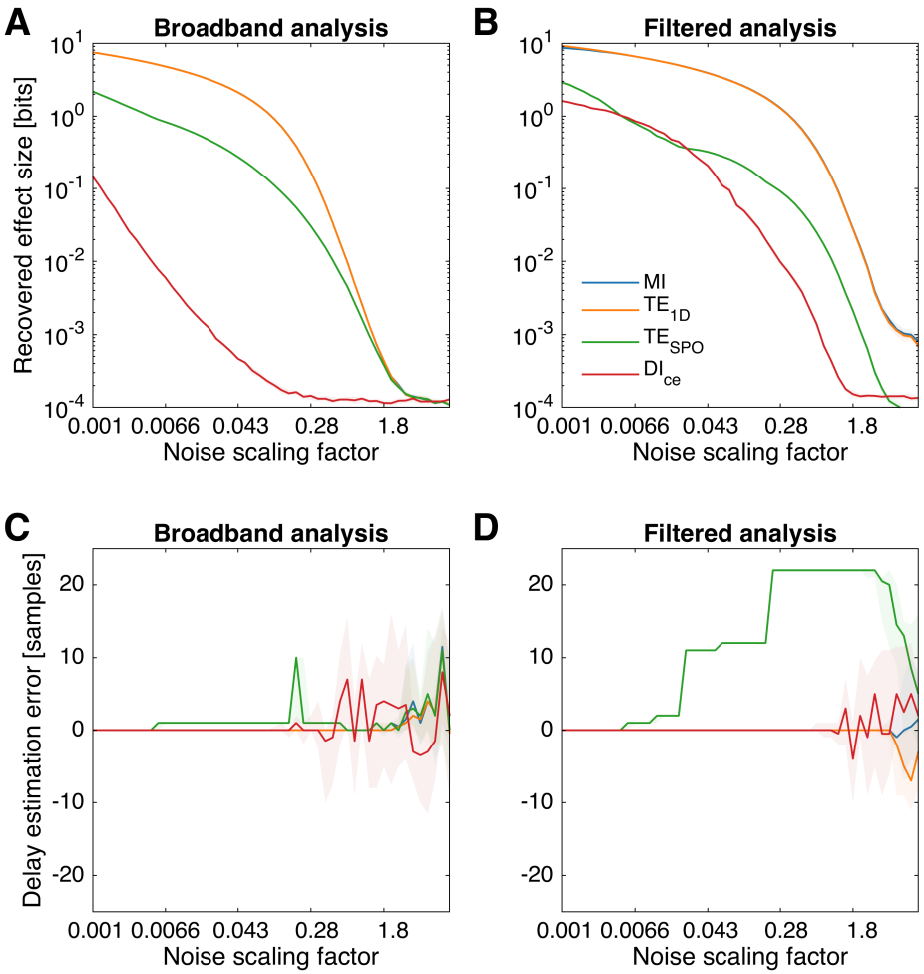

Figure 3.4: **Simulation with varying signal-to-noise ratios.**
Basic setup of the simulation is the same as in figure 3.1, however, here, the amplitude of noise added to source and target varies (but the same amount is added to source and target), and 100 repetitions are sampled. Plots show the median across 100 repetitions, shaded regions denote bootstrapped 95% confidence intervals. **A** In a broadband analysis scenario, all measures recover decreasing effect sizes with increasing noise amplitudes. DI_{ce} recovers the smallest effect sizes and reaches the noise floor the earliest. **B** In a filtered analysis scenario, the recovered effect sizes are in general higher for all measures. **C** The recovered delays deteriorate as the noise increases. In a broadband analysis scenario, all measures correctly recover the ground truth interaction delay at low noise amplitudes. **D** In a filtered analysis scenario, TE_{SPO} has higher delay reconstruction errors, while the delays recovered by the other measures benefit from the increased effect size.

3.3.5 Varying the signal-to-noise ratio independently in source and target signals

In a third step, we were interested in assessing how the four measures would react to asymmetric variations of the SNR (Bastos & Schoffelen, 2016). The ideal measure of TE should recover an invariant interaction delay and an effect size that symmetrically decreases as noise of increasing amplitude is added to either source or target signals.

Figure 3.5 shows how the set of four measures behaved. All measures succeeded in returning symmetric decreases of the recovered effect sizes when there was more noise

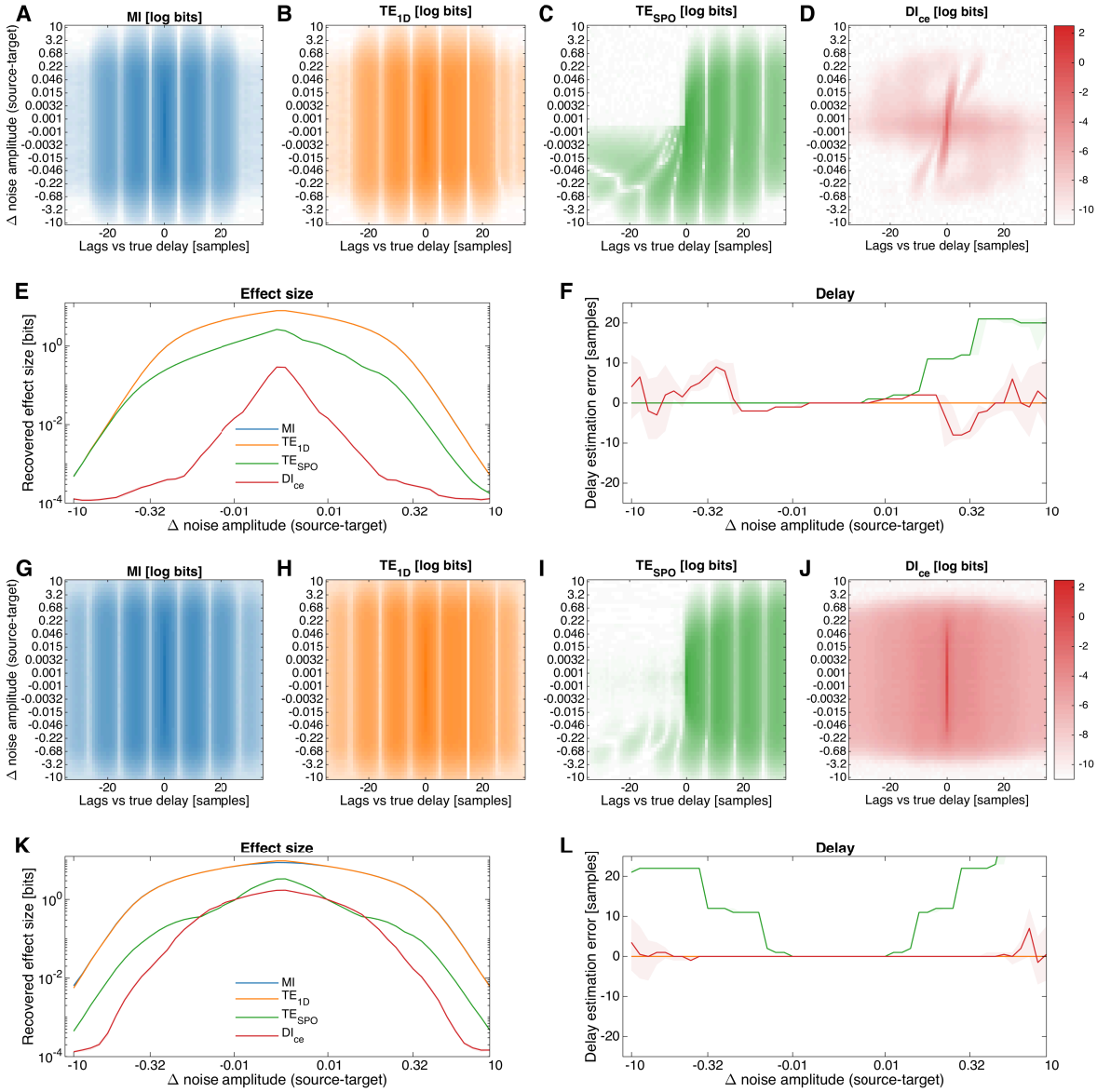

Figure 3.5: Caption on following page.

in either source or target signals, irrespective of whether analysis filters were applied or not. However, only the undirected delayed MI as well as the practically undirected TE_{1D} (see figure 3.1) recovered interaction delays that were unaffected in these settings. For TE_{SPO}, an increasing SNR imbalance with noisier source signals led to stronger biases in the estimated interaction delay for broadband analyses. For analyses where filters were applied, these biases grew stronger for both noisier source and noisier target signals. DI_{ce} exhibited biases in the broadband analyses, but performed favourably when analyses filters were applied. This analysis additionally corroborates an argument of caution when interpreting TE results (Wollstadt et al., 2017): when there are asymmetric changes in e.g. measurement noise across conditions, this does not automatically imply a change of coupling across these conditions, but only reflects the sensitivity of TE to such measurement noise.

Figure 3.5 *(previous page)*: **Simulation with signal-to-noise ratios varying independently in source and target signals.**
Basic setup of the simulation is the same as in figure 3.1, however, here, the amount of noise added to source or target signals is varied, and 100 repetitions are sampled. **A - D** show single trial delay profiles in a broadband analysis scenario for various measures, corrected for the true interaction delay across different noise amplitudes (negative delta on the y-axis refers to higher noise amplitude in the target signal, positive delta on the y-axis refers to a higher noise amplitude in the source signal). The ideal measure should always return the highest values at a lag of 0 relative to the true delay. Only MI and TE_{1D} succeed (but do not quantify Granger-causal information). **E** and **F** show the same as A – D, but repeated for 100 trials. Plots show the median across trials, shaded regions denote bootstrapped 95% confidence intervals. The maximum effect size across lags for all measures symmetrically decays as the noise increases in either source or target signals. The delay misestimation however behaves asymmetrically, especially for TE_{SPO}, which recovers wrong delays when the source is noisier than the target. **G – J** The same as A – D, but within a filtered analysis scenario. Here, DI_{ce} performs favourably. **K** and **L** show the same as E and F, but within a filtered analysis scenario.

3.3.6 Varying the bandwidths of the simulated effect as well as of the analysis filter

In our fourth simulation, we reasoned that a characteristic effect that a suitable measure of TE should show is a varying recovered effect size as a function of the bandwidth of the ground truth effect. Specifically, the theoretical extreme case of a sinusoidal source signal and its phase-shifted copy as a target signal should lead to zero TE. Such signals have no degrees of freedom and are perfectly predictable from themselves. Additionally considering potential source signals can thus not add any information. As one turns to signals with increasingly broad passbands, these degrees of freedom increase. Consequently, a higher directed effect should be quantifiable.

When testing this with the set of four measures (figure 3.6), we found that all measures monotonically increased in effect size as the bandwidth of the effect increased. However, in a broadband analysis, this increase spanned 2.5 orders of magnitude for DI_{ce}, which thus had the strongest relationship between its recovered effect size and increased degrees of freedom of the input signals. This was however also driven by the ground truth effect covering a larger part of the entire frequency spectrum. In a filtered analysis scenario where filter parameters were chosen to isolate the effect from the noise, only TE_{SPO} and DI_{ce} showed increases in effect size with broader effect bandwidths. The recovered delays were, as shown in previous analyses, unaffected for all measures except for TE_{SPO}. The latter exhibited stronger biases in the recovered delay for narrower simulated effect bandwidths.

In theory, applying analysis filters that match the ground truth effect should return a stronger effect size than analysis filters that are too broad or too narrow. Too broad analysis filters should fail to reduce the impact of noise in neighbouring frequencies, while too

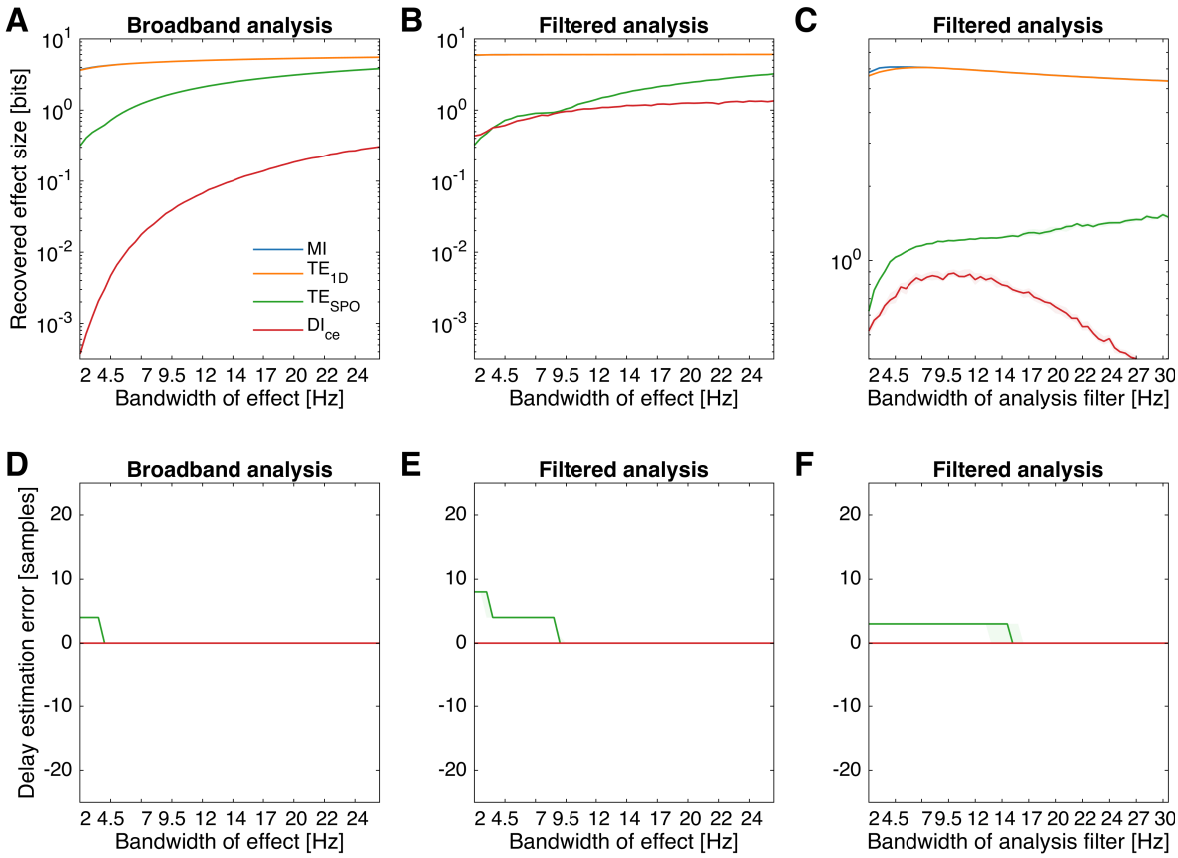

Figure 3.6: **Simulation with varying bandwidths of the transmitted signal.** Basic setup of the simulation is the same as in figure 3.1, however, here, the ground truth bandwidth of the effect is varied, and 100 repetitions are sampled. Plots show the median across 100 repetitions, shaded regions denote bootstrapped 95% confidence intervals. **A** The recovered effect size increases as the bandwidth of the effect grows. This increase however spans the most orders of magnitude for DI_{ce}. **B** In a filtered analysis scenario, the effect sizes are stronger than in a broadband analysis scenario. **C** When varying the bandwidth of the filter used in the analysis, DI_{ce} returns the peak effect size when the analysis filter matches the ground truth effect. All other measures fail to do so, TE_{SPO} even increases monotonically as the analysis filter passband grows. **D - F** report the interaction delays recovered in the same simulations as shown in A - C. The previously reported misestimation of the interaction delay of TE_{SPO} is limited to narrow band effects.

narrow analysis filters should ignore variance of the effect of interest and thus reduce the sensitivity. When applying analysis filters of varying width to a simulated coupled system of fixed ground truth effect bandwidth, DI_{ce} showed the clearest peak close to the ground truth effect (bandwidth of 10 Hz, figure 3.6C) across analysis filter widths. TE_{SPO} on the other hand exhibited a monotonic increase in effect size as the analysis filter was broadened, which would thus result in mis-estimations of the interaction bandwidth. The undirected delayed MI and the practically undirected TE_{1D} only showed comparably weak variance across different analysis filters.

3.3.7 Varying the centre frequency of the simulated effect

In a last simulation, we were interested in assessing the stability of the set of four measures when the centre frequency was varied. A suitable measure of TE should recover effect sizes and interaction delays with no variation under different ground truth centre frequencies, and it should find the strongest effects at the ground truth frequency.

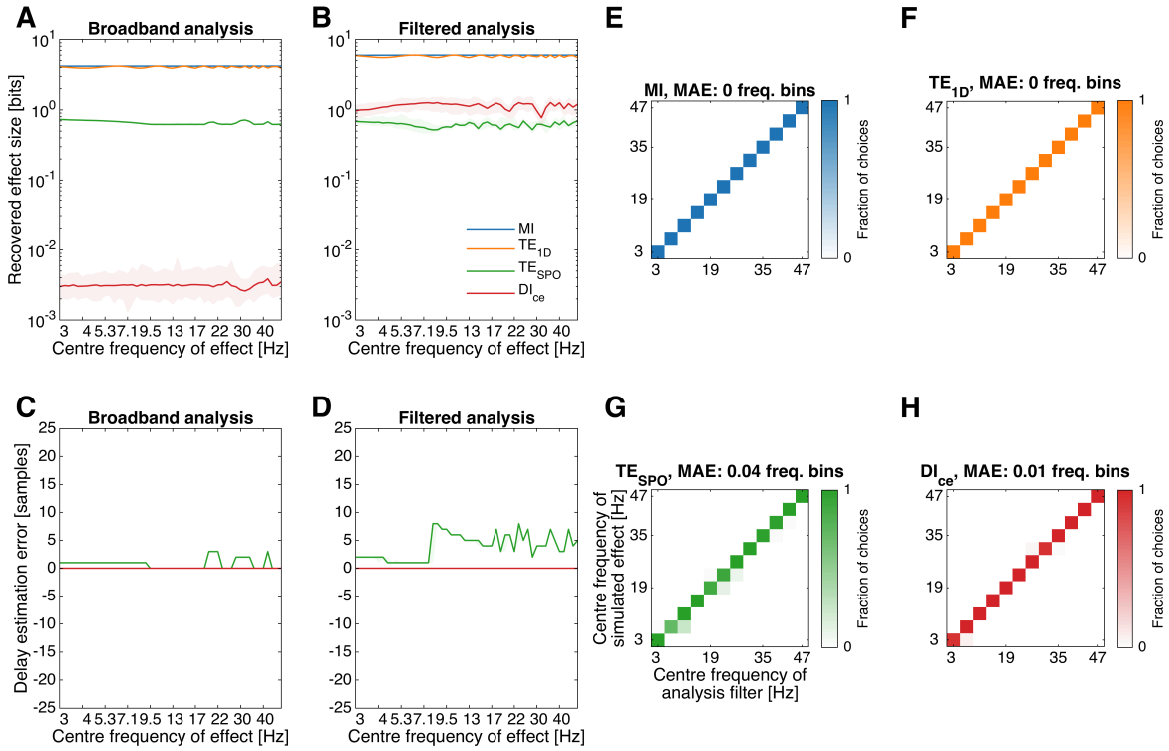

Figure 3.7: **Simulation with varying centre frequencies of the effect.** Basic setup of the simulation is the same as in figure 3.1, however, here, the centre frequency of the effect is varied, and 100 repetitions are sampled. Line plots show median across 100 repetitions, shaded regions denote bootstrapped 95% confidence intervals. **A** In a broadband analysis scenario, all measures recover effect sizes that are highly constant across centre frequencies. TE_{1D} however exhibits a slight ringing, TE_{SPO} and DI_{ce} recover noisier estimates. **B** The same holds for a filtered analysis scenario, where DI_{ce} and TE_{SPO} exhibit stronger variations. **C** and **D** As shown in other simulations, TE_{SPO} suffers from delay estimation problems. These vary especially strong across centre frequencies for filtered analysis scenarios. **E - H** Confusion matrices obtained when scanning data with different ground truth interaction frequencies with a bank of analysis filters (choosing the frequency with the maximum across analysis filters). All measures generally succeed in returning peak effects at the true interaction frequency. TE_{SPO} has the highest error.

We found that all measures considered here recovered effect sizes that were indeed relatively stable across different ground truth centre frequencies in both broadband and filtered analysis scenarios (figure 3.7). TE_{1D} and DI_{ce} however did exhibit slight biases, especially for the combination of DI_{ce} and higher simulated centre frequencies. All measures except TE_{SPO} recovered unbiased estimates of the interaction delay for all tested centre frequencies. TE_{SPO} on the other hand exhibited a marked variation of the recovered interaction delay across the different centre frequencies.

Finally, when analysing the simulated data with filters across different bands, all measures mostly peaked when analysis filters matched the ground truth frequency. For MI and TE_{1D}, the mean absolute error (MAE) across frequency bands was 0 frequency bins, for DI_{ce} it was .01 frequency bins and for TE_{SPO}, the MAE was strongest with .04 frequency bins. Identifying the correct spectral band thus did not seem to pose a particular problem for any of the considered measures.

3.3.8 Studying low-frequency MEG speech envelope tracking with directed measures

Finally, we wanted to explore the behaviour of our measure on real data. To do so, we turned to a dataset of $n = 24$ participants who listened to a continuous narrative of 1 hour duration while their MEG was recorded (Daube et al., 2019b). Examples of cerebro-peripheral coupling such as this present a good testbed for frequency-specific measures of directed connectivity, because the ground truth direction of the effect is known. Further, a plethora of studies has examined the relationship of MEEG responses to the time varying energy, or "amplitude envelope", of the speech stimulus in the delta and theta bands (Ding & Simon, 2014). However, it is usually studied using bivariate, undirected measures of connectivity that do not consider the self-predictability of the MEG responses. The degree to which such bivariate dependencies could be accounted for by auto-regressive models of the neuronal response signal could in principle reflect the degree to which bivariate speech tracking is a signature of predictive rather than reactive bottom-up processing. On the other hand, effects found by directed measures of dependency (under the assumption of a given auto-regressive model) would be stronger evidence of reactive, bottom-up processing of unpredictable parts of the stimulus input.

In most participants, we found spectra of delayed MI (see figure 3.8B and figure 3.9) that were suggestive of two spectral components involved in speech tracking which we will refer to as delta and theta bands (note that our functional definition of the theta band had higher upper cutoffs than the canonical theta band, which is commonly defined between 4 to 8 Hz, Klimesch, 1999; Wang, 2010). We found super-threshold delayed MI in both left and right auditory cortices of all participants in both delta and theta bands (figure 3.8C). This was stronger in delta than in theta frequency bands (fraction of samples in favour of hypothesis $f_{h_1} = 1$, figure 3.8E, see also figure 3.11). With DI_{ce} however, we found only weak effects that barely exceeded the noise thresholds in the delta bands, while effects in the theta bands exceeded noise thresholds with only 1 exception in the left and 2 exceptions in the right hemisphere (figure 3.8C). This constitutes strong evidence for a robust population-level prevalence of the effect (Ince et al., 2021). These theta DI_{ce} effects were generally much stronger than in the delta band ($f_{h_1} = 1$; figure 3.8E, see also figure 3.8E, see also figure 3.11). These results translated into a difference of ratios of DI divided by MI between the delta and theta frequency bands, where we found higher ratios in the theta than in the delta band in both left and right auditory

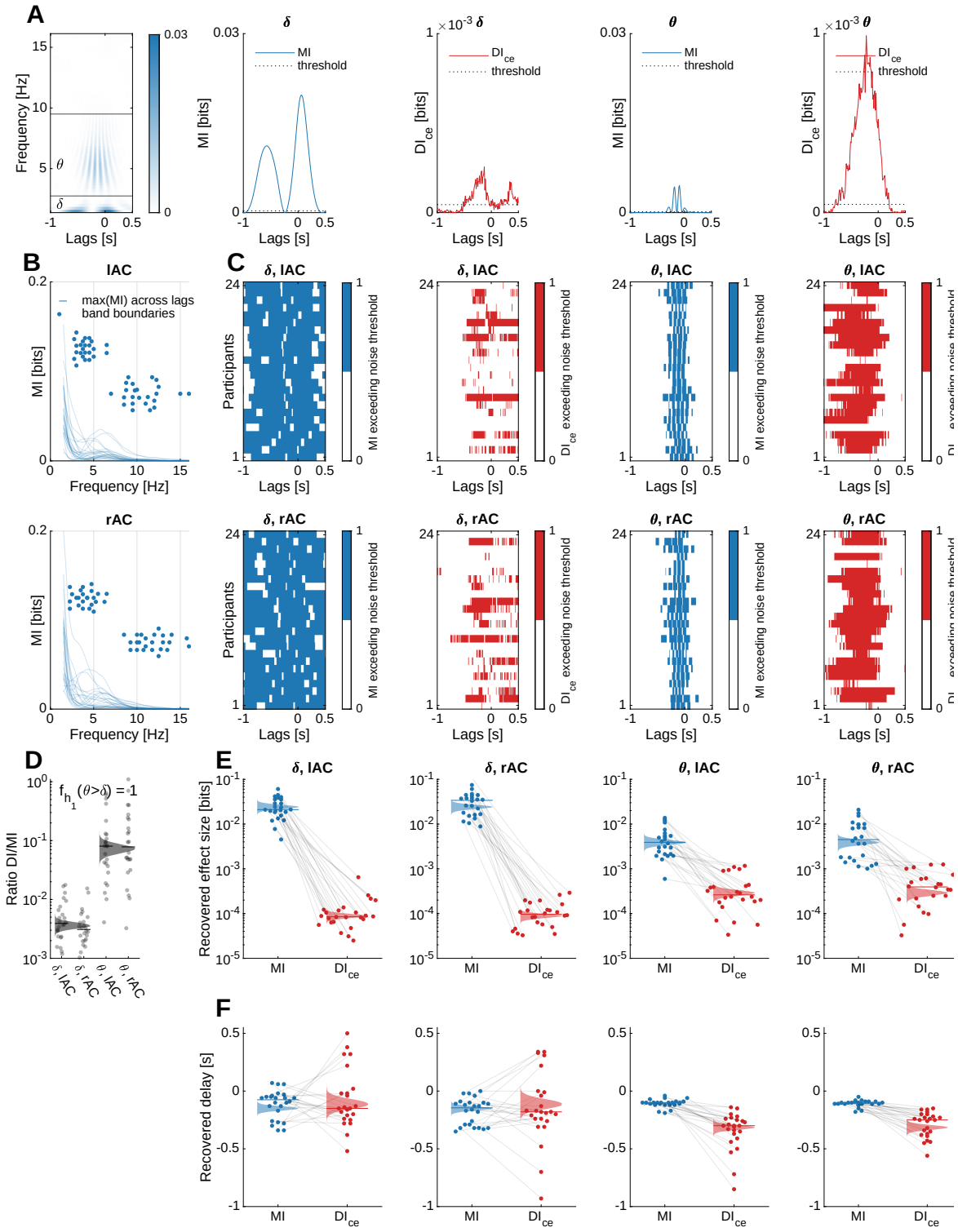

Figure 3.8: Caption on following page.

cortices ($f_{h_1} = 1$; figure 3.8D, see also figure 3.8E, see also figure 3.11). This suggests that the bivariate speech tracking in the theta band could consist of more bottom-up and reactive processing than the bivariate speech tracking in the delta band. We had found in simulation analyses that the effect size recovered by DI_{ce} grows as a function of increasing bandwidth (see figure 3.6). Since our individualised definitions of delta and theta bands led to wider theta than delta analysis filters, we repeated this analysis while constraining the analysis filter bandwidth to be the same in delta and theta bands. We

Figure 3.8 *(previous page)*: **Results obtained from source level MEG recordings (left and right auditory cortices, l and r ACs) obtained during continuous speech listening (n=24).**
A Results for a typical participant. Plots show spectrotemporally resolved delayed MI with individual boundaries of delta and theta bands overlaid as black horizontal lines (leftmost plot) as well as delay profile of delayed MI and DI_{ce} in delta and theta bands, with noise thresholds overlaid as black dotted lines. **B** Spectral profiles of delayed MI for all participants in left (top) and right (bottom) ACs. Points denote upper pass-band cut-off frequencies of the delta (upper scatter) and theta (lower scatter) bands. **C** Binarised delayed MI as well as DI_{ce} delay profiles for all participants in delta and theta bands in left and right ACs. While delayed MI in both bands and DI_{ce} in the theta bands cross the noise threshold in most cases, DI_{ce} in the delta band is often close to or below the noise threshold. **D** Ratios of DI/MI in delta and theta bands of left and right ACs for each individual participant, medians overlaid as black lines, density estimates from posterior distributions overlaid as transparent shapes. Theta band bivariate tracking can be less explained by auto-prediction of the MEG signal than delta band tracking. **E** Effect sizes underlying the computation of the ratio in D in delta and theta bands of the left and right ACs in each individual participant. Medians are overlaid as coloured lines, density estimates from posterior distributions overlaid as transparent shapes. **F** Delays recovered by delayed MI and DI_{ce} in the delta and theta bands of left and right ACs. Since DI_{ce} fails to cross the noise threshold in many cases for the delta band, these delay estimates are uninterpretable. In the theta band, DI_{ce} recovers a longer delay than delayed MI. Also see Figures 3.9 – 3.11.

found that the difference in DI over MI ratios of the delta vs the theta bands shrunk, but persisted under this constraint ($f_{h_1} = 1$; see figures 3.10 and 3.11). We could thus rule out that analysis bandwidth alone could explain the difference in DI/MI ratios in delta vs theta bands.

Given the overall relatively narrow passbands of both delta and theta bands, the measurable directed effects were in general up to several orders of magnitude lower than the undirected effects, suggesting that predictive processing could indeed make up the lion's share of delta and theta envelope tracking.

Lastly, we considered the delays recovered by delayed MI and DI_{ce} in delta and theta bands (figure 3.8F). For the delta band, the delays recovered by DI_{ce} spanned a wide range, even reaching into regions suggestive of the MEG signal preceding the speech signal for some participants. However, since the effect sizes in the delta band were in many cases close to or below the noise threshold, these recovered delays were uninterpretable. For the theta band however, effect sizes were robust in most cases. Interestingly, the delays recovered by DI_{ce} (lAC: median across participants of -.43s, rAC: median of -.40s) were longer than those recovered by delayed MI (lAC: median of -.10s, rAC: median of -.10s; $f_{h_1} = 1$), suggesting a slower bottom-up, reactive processing.

Taken together, we take our results as aligning with models that characterise the lion's share of activity of auditory cortices visible in MEG at a given point in time as reflecting predictions of the stimulus at a short delay. Such predictions can be formed on the basis of an integration of reactive, bottom-up processing of stimulus input at longer delays with

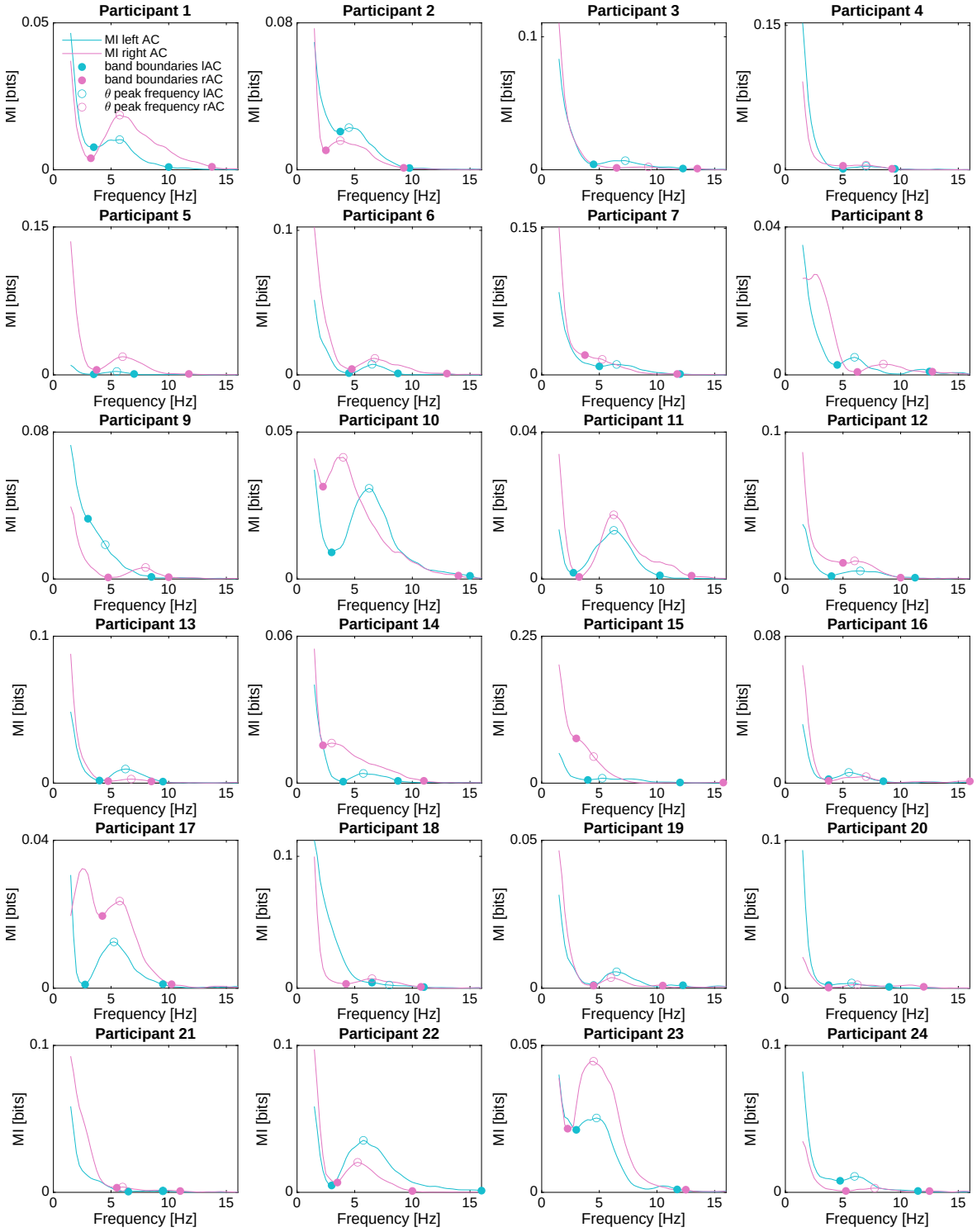

Figure 3.9: **Spectra of MI for left and right auditory cortices for each individual participant (related to figure 3.8).**
Individual boundaries of delta and theta bands are overlaid as filled circles, theta peak frequencies (used for figure 3.10) are overlaid as empty circles.

an internal language model.

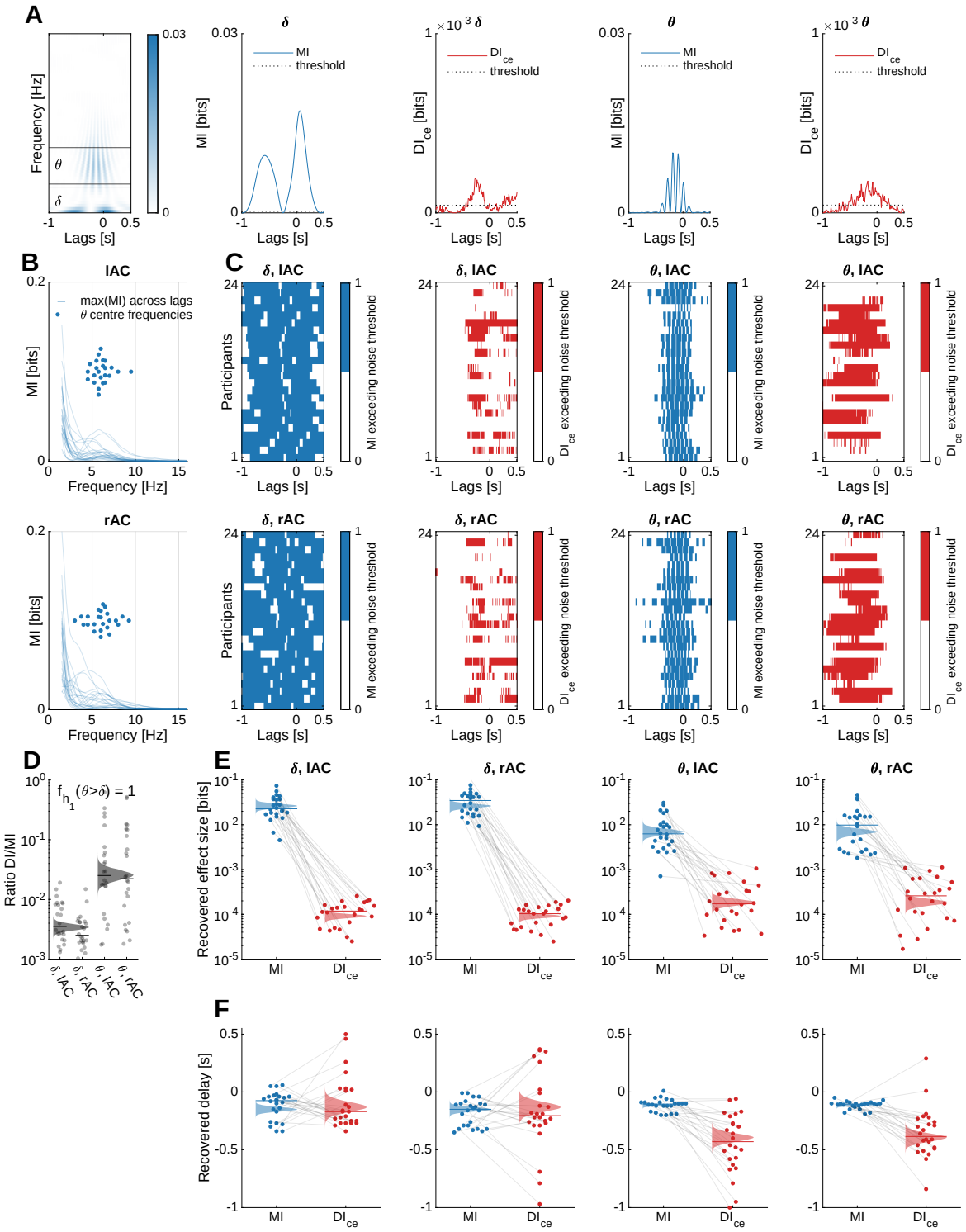

Figure 3.10: Caption on following page.

3.4 Discussion

In this study, we have addressed the problem of quantifying band-limited directed interactions in bivariate sets of source and target signals. With a simplistic and intuitive simulation setup, we have highlighted shortcomings of common estimators of TE when facing this problem. Our proposed alternative DI_{ce}, relying on a type of temporal whitening of the source and target signals, overcomes these shortcomings. With an array of simulations, we extensively characterised DI_{ce} in relation to undirected MI as well as

Figure 3.10 *(previous page)*: **Results obtained from source level MEG recordings with constrained frequencies (related to figure 3.8).**
Same as figure 3.8, but computed with constrained bandwidths of analysis filters. Delta band was fixed by means of a low-pass filter common across all participants with a cutoff frequency of 3.5Hz. Theta frequency was defined as a 3Hz wide band centered on individual theta centre frequencies (see figure 3.9). The increased DI/MI ratio found with individualised band definitions (where often theta was defined as a wider band than delta) was robust to the constrained analysis bandwidths, making it unlikely that bandwidth alone can fully account for the difference in DI/MI ratio between delta and theta bands. Also see figure 3.11.

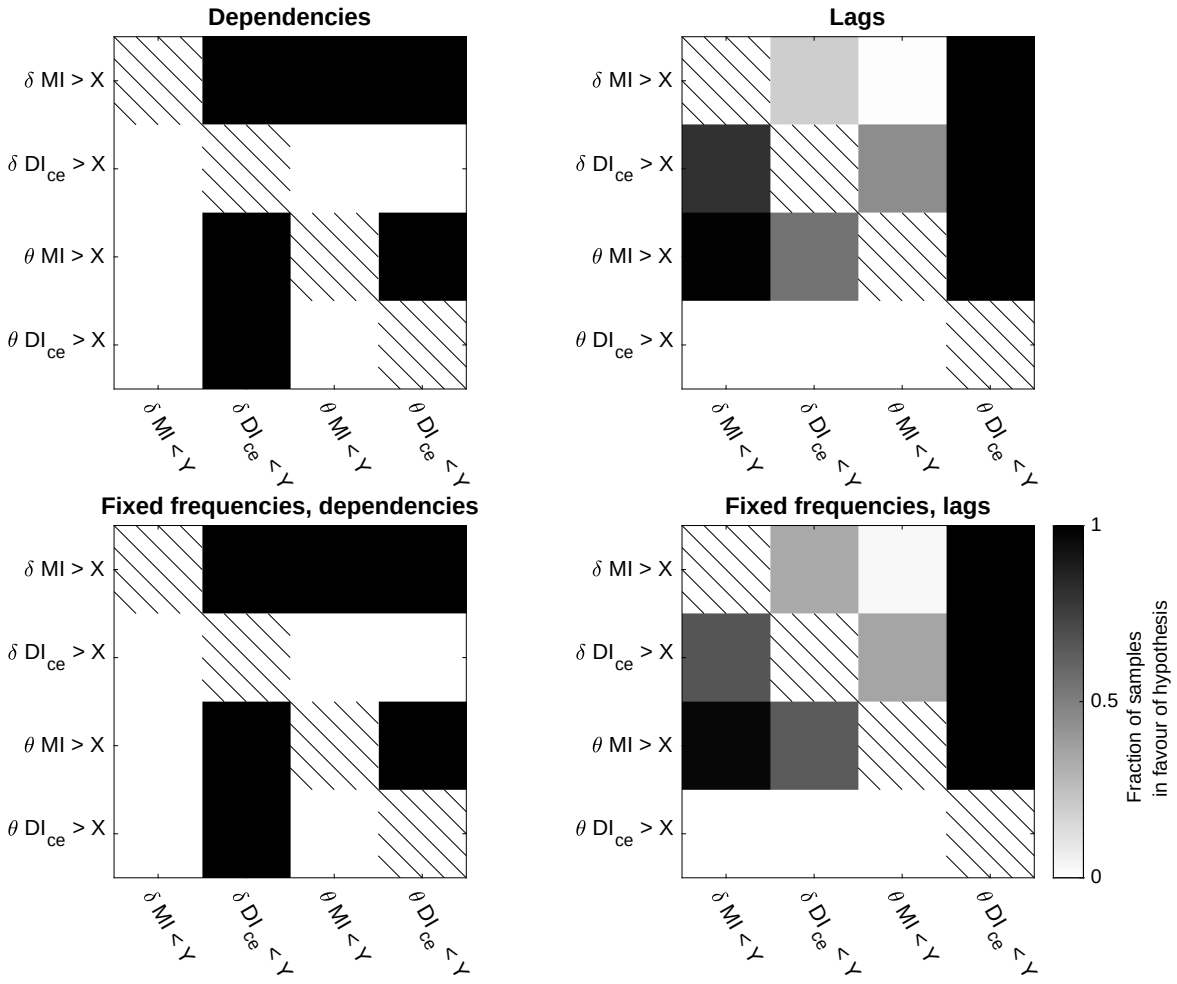

Figure 3.11: **Comparisons of posterior distributions of main effects of combinations of frequency bands and dependency measures from Bayesian linear modeling of the raw dependency and lag estimates (related to figure 3.8).**
In each cell in the matrices, the greyscale colour denotes the fraction of samples of the combination of frequency band and dependency measure referenced on the y-axis that is larger than the combination referenced on the x-axis (testing a hypothesis). Top row reports results corresponding to figure 3.8, bottom row reports results corresponding to figure 3.10.

common estimators of TE. Lastly, we turned to a dataset of continuous speech listening in MEG to study the speech envelope tracking specific to delta and theta bands from the perspective of a directed measure. We found that in such narrow bands with essentially

low degrees of freedom, the measurable directed effects are very small in comparison to undirected effects. Moreover, with DI_{ce}, we found that theta speech tracking has a stronger directed part than delta speech tracking, and that the directed theta speech tracking has a longer delay and a broader temporal profile than the undirected tracking. The directed effects could potentially reflect a purely bottom-up processing stage at a later delay. On the basis of this reactive processing, the auditory cortical system could then generate predictions about the upcoming speech acoustics, which the undirected effects could be a signature of.

Many of the problems and perspectives highlighted in this study are not new to the literature. We are for example not the first to point out problems of TE estimators relying on only a single sample of the target past signal at a delay equivalent to that of the source to the target at a given scanned interaction delay (Wibral et al., 2013). However, we hope that our intuitive demonstrations contribute to a more widespread appreciation of the high similarity of the resulting "TE"$_{1D}$ and simple delayed MI as well as variations of recovered effect sizes in the presence of different ground truth interaction delays. Likewise, the delay misestimation problems of more sophisticated estimators such as TE_{SPO} when faced with narrow-band effects have been reported before (Wollstadt et al., 2017). In the same way, the PID perspective on conditional MI based TE and the implication of including synergistic interactions has been developed previously (James et al., 2016). Here, we found that from this PID perspective on the TE problem, the delay mis-estimations of TE_{SPO} could mostly be attributed to such synergistic interactions of the source and the target that TE estimators relying on conditional MI pick up on.

An intuitive solution to fix this would be to instead consider the unique information of the source about the target, and thus effectively ignore the synergistic contributions included in conditional MI (Barrett, 2015). However, we found that this was neither particularly sensitive in detecting effects nor accurate in recovering the ground truth delays, albeit with much less error than synergy-driven conditional MI based estimators. Our proposed estimator, DI_{ce}, circumnavigates these issues by not relying on conditional MI in the first place, but by instead converting the source and target time series into time series of sample wise conditional entropy. This effectively frees the signals from predictable parts, ensuring that a subsequently computed delayed MI cannot be affected by autocorrelation. Applying this whitening to the target signal only would be analogous to the basic idea of TE, but would introduce asymmetric temporal distortions in the target relative to the source, resulting in erroneous delay estimates. Again, the general idea of such a two-staged approach is not new as such (Haugh, 1976; Cliff et al., 2021). However, we have here developed it from a simple cross-correlation of residuals (Haugh, 1976), which only takes into account the mean of a predicted time series, to the delayed MI of sample wise conditional entropy, which instead relates a given sample to the mean and variance of a predicted distribution. On the basis of our simulations, we put forward that it does offer a useful perspective on the problem of estimating directed interactions of band-limited processes at the correct delay.

In the light of various careful considerations of the pitfalls associated with this goal (Florin et al., 2010; Barnett & Seth, 2011), we thus hope that we can contribute to a more widespread adoption of directed perspectives on band-limited dependencies. Within neuroscience, such genuine narrow-band signal components are arguably not in all places where aperiodic signals have been straight-jacketed into oscillations by means of band-pass filters (Donoghue et al., 2020; Gerster et al., 2021). In fact, depending on the algorithm used to define oscillatory components, even the delta component identified by us is potentially attributable to the aperiodic part of the MEG spectrum. In general, oscillatory components are nevertheless indisputably widespread (Wang, 2010). It is in such situations of high auto-correlation where the consideration of self-predictability should have the most obvious and strongest effects, and where the use of measures like TE is thus of potentially high interest. After all, frequency-specific oscillatory components are at the heart of popular conceptualisations of interactions of neuronal populations (Schnitzler & Gross, 2005; Fries, 2015). We conjecture that estimators of directed interactions that return counterintuitive results in such circumstances as simplistically simulated may not be well matched to the interpretations typically applied. This problem has been recognised elsewhere, and (Pinzuti et al., 2020) propose a clever permutation scheme to address this issue. Here, we instead suggest prioritise the interpretability of the directly measured effect size.

In this spirit, our measure is in principle a versatile tool applicable to arbitrary (neuro-)scientific questions. We however subscribe to the view that as such, mechanistic interpretations of cerebro-cerebral dependencies are prohibitively hard in many cases (Mehler & Kording, 2018). Given the highly incomplete picture of the entire neuronal activity that is accessible with any given neuroimaging modality, it is virtually impossible to rule out that a given dependency stems from a third unobservable region affecting the supposed source and target. Directed dependency measures per se thus never warrant the inference of causality. This problem is however alleviated in cerebro-peripheral settings, where the dependency of neuroimaging signals on external signals is studied (Gross et al., 2021).

The question of causality is thus arguably less controversial in our example case of the dependence of MEG signals on a continuous acoustic speech stimulus. According to the burgeoning field of predictive coding however, the brain appears to predict the upcoming speech input (Brodbeck et al., 2018a; Donhauser & Baillet, 2020; Heilbron et al., 2021), rendering temporal relations of time series of assumed cause and effect less trivial. Our suggestions regarding these problems are mostly of an indirect nature, arguing that the part of undirected dependencies that directed measures deliberately ignore could reflect predictive processing, while what directed measures quantify should reflect reactive, bottom-up processing. It is important to note however that simply because the part of a neural response that we can observe with a given neuroimaging modality is predictable from its own past does not automatically imply that it is in fact being predicted by the brain (de-Wit et al., 2016). Response components identified with

impulse response functions as measured with autocorrelations, temporal response functions (Crosse et al., 2016) or here delayed MI have been suggested to be analogous to the well-studied components of auditory event-related potentials (Lalor et al., 2009). These do occur in response to unpredictable events (Ritter et al., 1968), implying that the underlying generators are not at all restricted to prediction. It has however also been suggested that the low-frequency potential "entrains" in response to sustained stimuli such as continuous speech, especially when they are (quasi-)rhythmic (Lakatos et al., 2008, 2019). Viewing ERPs in response to unpredictable events through this lens, they could manifest the launch or "reset" of such predictions for subsequent events (Sayers et al., 1974; Makeig et al., 2002). Such entrainment would serve to align phases of low and high neuronal excitability to relevant parts of the stimulus, such that crucial stimulus information could be encoded with high fidelity. In this sense, the low-frequency potential is indeed often seen as a neuronal signature of a prediction about the upcoming stimulus. A further interesting detail of our results in this respect is the distribution of lags recovered by delayed MI (i.e. the timing of the peak of the delay profile). This was relatively broad in the delta frequency band, with the delays of some participants even suggesting the stimulus to follow the response, a hallmark of prediction that has been observed and interpreted similarly before in EEG delta-band responses to continuous speech (Etard & Reichenbach, 2019). Interestingly, such slow delta tracking signatures disappear when speech of a language foreign to the participant is listened to (Ding et al., 2016), which is to be expected if it is seen as a predictive component relying on an internal language model.

Under this perspective, the reactive, bottom-up part of the coupling carved out by DI_{ce} would then essentially be driven by adjustments of the predictions in reaction to unpredictable parts in the stimulus. An important constraint of our proposed approach however follows from the choice of the model used to whiten the signals. Our choice of a non-uniform embedding to predict a signal's present state from its own past is rooted in the history of TE, and its suitability to model an actual hypothesised biological acoustic model is disputable. In principle, there is nothing that prohibits the combination of our framework with more powerful autoregressive models such as recurrent neural networks. That the auditory system is actually entertaining correspondingly complex multi-level predictions of the upcoming input is an increasingly popular perspective (Donhauser & Baillet, 2020; Koskinen et al., 2020; Heilbron et al., 2021; Jain et al., 2021; Schmitt et al., 2021; Caucheteux et al., 2021). This would then effectively further shrink the effect sizes that our proposed DI_{ce} could discover, assigning an even smaller portion of variance to unambiguously bottom-up reactive processing than found here.

Lastly, it is interesting to compare our approach to a recently increasingly popular analysis strategy in the field of predictive coding. This strategy usually builds on the framework of encoding models (Naselaris et al., 2011), where first a so-called linearising feature space (a nonlinear transformation of the stimulus) is identified and then linearly mapped onto the brain response. For questions of predictive coding, researchers make

use of powerful predictive models of a stimulus class to derive measures of surprisal (and uncertainty) associated with a given part of the stimulus given its preceding context to obtain the linearising feature space (Brodbeck et al., 2018a; Donhauser & Baillet, 2020; Koskinen et al., 2020). Our transformation of the signal of time varying energy of the speech stimulus into a time-series of conditional entropy can be seen as a simple version thereof. Within the standard approach, it is then however uncommon to perform a similar operation on the neuronal response, which is central to our proposed measure. The standard approach makes sense from the viewpoint of the hypothesis that the neuroimaging signal mainly reflects a prediction error. The conceptualisation of neuronal signals as reflecting prediction errors however decreases in its appeal to the degree to which the neuronal signal is predictable from itself. How could the brain be surprised if it already knew it was going to be surprised? We thus hypothesise that studies of prediction errors could increase their specificity by including a whitening of the target signals akin to what we suggest here. In theory, this should allow a more detailed characterisation of processes related to prediction errors.

Taken together, this study makes a threefold contribution to the problem of quantifying band-limited directed dependencies: We offer simplistic yet intuitive and accessible simulations, propose our own estimator based on an information theoretic pre-whitening as well as provide an application to the case of speech tracking in MEG in delta and theta bands. With this, we hope to spark further interest in directed perspectives on frequency-specific interactions.

3.5 Methods

3.5.1 Estimation of information theoretic quantities

We used the Gaussian Copula Mutual Information framework (gcmi, Ince et al., 2017). Here, the basic idea is to transform variables into standard normals to then apply closed-form expressions for information theoretic quantities of Gaussians. These can be derived as follows. The entropy H of a variable X is given by the expected value of the surprisal:

$$H(X) = E(h(x)) \tag{3.1}$$

For a Gaussian continuous variable X, this corresponds to the "global" differential entropy:

$$H(X) = \int_{-\infty}^{\infty} f(x)h(x)dx \tag{3.2}$$

Here, the local or "sample wise" contribution, i.e. the surprisal (also called "information content", "Shannon information" or "self-information"), is thus given by:

$$h(x) = -\log(f(x)) \tag{3.3}$$

For a Gaussian continuous variable of k dimensions, this is equal to to the negative log-likelihood:

$$h(x) = -\left(-\frac{1}{2}\left[\log(|\Sigma|) + (x-\mu)'\Sigma^{-1}(x-\mu) + k * \log(2\pi)\right]\right) \tag{3.4}$$

$H(X)$ can then be simplified to:

$$H(X) = \frac{k}{2} + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log(|\Sigma|) \tag{3.5}$$

We can then compute the mutual information between the variables X and Y (where [] denotes concatenation to obtain a joint variable) follows:

$$MI(X,Y) = H(X) + H(Y) - H([XY]) \tag{3.6}$$

This pragmatic estimator comes at the cost of not being able to quantify all possible nonlinear effects, as information theoretic measures should do in theory. Instead, it quantifies relationships described by a Gaussian copula. However, it avoids the loss of information and incompatibility with multidimensional variables as well as higher-order information theoretic quantities incurred by binning, a similarly pragmatic choice. Further, it is computationally cheaper and has a higher sensitivity for rank relationships than more sophisticated estimators such as nearest neighbour approaches, which can resolve nonlinear effects, but can in practice only be computed on vast amounts of samples from variables of relatively limited dimensionality. In situations where nonlinear effects are of interest and resolvable when using a binned estimator with a realistically low amount of bins (such that each bin of the joint distribution is sufficiently sampled), such estimators are a pragmatic complement to gcmi.

3.5.2 Delayed mutual information

The simplest information theoretic approach of estimating a dependency between source and target variables is computing the delayed mutual information (MI). It is the information theoretic equivalent of cross-correlation, where versions of the source- and target variables with a range of lags to each other are generated to then compute the dependency measure at each lag and obtain a delay profile of the relationship. In principle, the lagging of the source and target variables to each other can be implemented either way, by lagging the source and fixing the target or vice versa. To simplify concrete explanations that follow, we here choose an implementation where the source is lagged while the target is kept fixed. We thus define delayed MI between time series X and Y at an interaction delay δ_a as:

$$\text{delayed } MI(X,Y)_{\delta_a} = MI\left(X_{\delta_a}, Y_0\right) \tag{3.7}$$

Here, X_{δ_a} denotes the lagged source with negative values of δ_a corresponding to

delays where the source precedes the target and positive values corresponding to delays where the source follows the target. We trim the target present Y_0 to be of an equal number of samples as the lagged source.

3.5.3 Transfer entropy

Originally, transfer entropy (TE) has been described as a conditional MI (Schreiber, 2000). Instead of computing the simple delayed MI at each lag, the idea is to condition this MI on an operationalisation of the target past. In principle, the resulting quantity is thereby supposed to quantify to what extent the target signal can be explained by the source signal over and above to the extent to which the target signal can be explained by its own past. This idea is equivalent to that of Granger causality (Granger, 1969) in the case of linear effects observed in Gaussian variables (Barnett et al., 2009). Transfer Entropy based on conditional MI can be expressed as a difference of two MI terms:

$$TE(X,Y)_{\delta_a} = MI\left(X_{\delta_a}, [Y_- Y_0]\right) - MI\left(X_{\delta_a}, Y_-\right) \tag{3.8}$$

A critical problem is the concrete operationalisation of the target past Y_-.

One-dimensional embedding

A simple form of operationalising the target past is to simply mirror the delay of the delay scanning procedure of the delayed MI (Besserve et al., 2010; Lobier et al., 2014; Park et al., 2015; Ince et al., 2015; Giordano et al., 2016; Morillon & Baillet, 2017). When a given interaction delay is considered in diagonal TE, the delayed MI between X and Y is conditioned on a version of the target to which the same lag is applied as that which is applied to the source. We refer to this one-dimensional embedding as "$\text{TE}_{1\text{D}}$". With Y_- thus corresponding to Y_{δ_a}, we obtain:

$$TE_{1\text{D}}(X,Y)_{\delta_a} = MI\left(X_{\delta_a}, \left[Y_{\delta_a} Y_0\right]\right) - MI\left(X_{\delta_a}, Y_{\delta_a}\right) \tag{3.9}$$

Multidimensional embedding approaches

The one-dimensional method of operationalising the target past is simple to implement, but hard to motivate. When considering that the core goal of quantifying TE is to estimate to what degree the source variable contains information about the target that is not available from the target past itself, it becomes obvious that the assumption of capturing the target's autoinformation with a single delay that varies as a function of the considered interaction delay is daunting (Wibral et al., 2013). Multiple target past time points could jointly influence the target present, and there is no obvious reason why the target past should be varied as a function of the interaction delay considered. A source variable could have information about the target variable at a different delay than the delay at which the target past informs its own future in the same way as the source. A

more sensible approach to describe the target past is thus to consider a multidimensional representation or "embedding" (Takens, 1981) that is invariant to the interaction delay considered between source and target in the delay scanning procedure. A classic idea to find such an embedding of the target past is parameterised by two parameters: A spacing parameter and a dimensionality parameter (Ragwitz & Kantz, 2002). This means that the assumption is that the information the target past contains about its present can be captured with a number of equally spaced delays relative to the present. A further development of embedding methods relaxes the assumption that the embedding delays must be uniformly spaced (Vlachos & Kugiumtzis, 2010; Faes et al., 2011). It is conceivable that a lot of information about the target present can be found in a combination of many, densely clustered delays at very short time scales as well as only few unevenly spaced delays at longer time scales, or vice versa. To account for this, non-uniform embedding procedures adopt an iterative approach to generate a target past embedding. Concretely, a search space of candidate delays is defined by the maximal delay that is assumed to have an influence on the target present. In the first iteration, the delay within this search space is chosen that maximises the MI of the target variable at this delay relative to the target present. In subsequent iterations, a conditional MI between all remaining candidate delays and the target present conditioned on the already chosen delays is computed. If desired, it can be tested whether the found maximal conditional MI surpasses a defined noise threshold. In that case, the iterative procedure can be stopped once the found maximal conditional MI no longer exceeds the threshold. Alternatively, a pragmatic hyperparameter of maximal iterations can be defined, which avoids the computationally costly permutation testing. This procedure thus has two hyperparameters: the length of a search space as well as a noise threshold (in the form of a significance level alpha) or alternatively a fixed embedding dimensionality. In our study, we fixed these to a search space of 1 second and a fixed embedding dimensionality of 50 (simulations) or 25 (MEG data). Conditional MI based TE estimators that employ such multidimensional target past operationalisations have been described to be "self prediction optimal" and are thus referred to as $\mathrm{TE}_{\mathrm{SPO}}$ (Wibral et al., 2013).

$$TE_{SPO}(X,Y)_{\delta_a} = MI\left(X_{\delta_a}, [Y_{emb}Y_0]\right) - MI\left(X_{\delta_a}, Y_{emb}\right) \tag{3.10}$$

Here, Y_{emb} denotes a multidimensional embedding of the target past.

Whitening approaches

The suggestion we make here is to tackle the TE estimation problem with a whitening approach. The basic idea is to use a non-uniform multidimensional target past embedding as a model to derive a prediction about the target present state. Analogous to (Haugh, 1976), this approach could then be used to subtract the prediction from the observed target time series and compute delayed MI on the residuals. This however ignores the uncertainty associated with each prediction. We thus instead compute a time series

of sample wise conditional entropy of each sample in the observed target time series given its past embedding. We obtain the conditional entropy of each sample x given its non-uniform embedding x_{emb} by subtracting the marginal entropy of x_{emb} from the joint entropy of x and x_{emb}:

$$h(x \mid x_{emb}) = h([x x_{emb}]) - h(x_{emb}) \tag{3.11}$$

The resulting time series is thus effectively temporally decorrelated, or "whitened". We define DI_{ce} between the source X and the target Y at an analysis delay δ_a as:

$$DI_{ce}(X,Y)_{\delta a} = MI\left(h(x \mid x_{emb})_{\delta_a}, h(y \mid y_{emb})\right) \tag{3.12}$$

Any resulting quantifiable delayed MI between the whitened source and the target can thus no longer stem from information that the target carries about itself. To avoid asymmetrical temporal distortions caused by the temporal whitening, we apply the whitening to both source and target time series.

3.5.4 Partial information decomposition

In Partial information decomposition (PID, Williams & Beer, 2010), systems of three or more variables are considered. In situations where there are two sources and one target variable, it is the goal to quantify the "redundant" amount of information that the two sources share about the target as well as the "synergistic" information about the target that is only available when considering the two sources jointly as well as the unique information about the target that is only available from one of the sources but not the other. From the lens of this formalism, conditional MI can be seen as the sum of two "atoms" of PID: Unique information and synergy. TE based on conditional MI is thus the sum of not only unique information of the source signal about the target present but also synergy of the source signal and the target past about the target present.

PID has its origins in co-information (McGill, 1954), which is computed as a triple set intersection.

$$\mathrm{CoI}\left(X_{\delta_a}, Y_{emb}, Y_0\right) = MI\left(X_{\delta_a}, Y_0\right) + MI\left(Y_{emb}, Y_0\right) - MI\left(\left[X_{\delta_a} Y_{emb}\right], Y_0\right) \tag{3.13}$$

Here, negative values correspond to synergistic information, and positive values correspond to redundant information. From the lens of PID however, these are mere net sums of the "pure" redundancy and synergy that PID aims to resolve. To do so, the essential first step is to define redundancy. Here, we turned to an implementation based on "common change in surprisal" (I_{ccs} Ince, 2017a). This starts at the observation that local MI can be seen as the positive or negative change in surprisal of a given value when

another is observed:

$$\mathrm{mi}(x,y) = \Delta_y h(x) = h(x) - h(x \mid y) \tag{3.14}$$

We can then also consider equally positive or negative co-information at the local level:

$$\mathrm{coi}\left(x_{\delta_a}, y_{emb}, y_0\right) = \mathrm{mi}\left(x_{\delta_a}, y_0\right) + mi\left(y_{emb}, y_0\right) + mi\left(\left[x_{\delta_a} y_{emb}\right], y_0\right) \tag{3.15}$$

Each of such co-information terms can then either contribute to net synergy or net redundancy. The key idea in I_{ccs} is then to consider only those positive (redundant) terms of local $coi(x_{\delta_a}, y_{emb}, y_0)$ which coincide with positive terms of $mi(x_{\delta_a}, y_0)$, $mi(y_{emb}, y_0)$ and $mi([x_{\delta_a} y_{emb}], y_0)$. Thus, only co-information terms are counted that represent a commonly shared change in surprisal. The resulting global redundancy can then be used to infer the other PID atoms according to a lattice structure (Williams & Beer, 2010):

$$\begin{aligned} \text{unique } \left(X_{\delta_a}\right) &= MI\left(X_{\delta_a}, Y_0\right) - \text{ redundancy} \\ \text{unique } \left(Y_{emb}\right) &= MI\left(Y_{emb}, Y_0\right) - \text{ redundancy} \\ \text{synergy } &= MI\left(\left[X_{\delta_a} Y_{emb}\right], Y_0\right) - \text{ redundancy } - \text{ unique } \left(X_{\delta_a}\right) - \text{ unique } \left(Y_{emb}\right) \end{aligned} \tag{3.16}$$

3.5.5 Noise thresholds

To establish whether a given effect size exceeded a level that would be expected of data not containing the effect of interest, we considered the $95th$ percentile of distributions of effects obtained from 1000 permutations. For this, we performed circular shifts of the source variable by a random number of samples and then recomputed the estimators of interest. We constrained this random amount with a minimum and a maximum, such that the permuted data could not include instances where potential effects of the observed data would end up within the range of delays of interest. In cases where multiple comparisons were made, we corrected the noise threshold by means of the family wise error rate, that is, by considering the maximum across all conditions within a given permutation and subsequently applying the $95th$ percentile of the resulting distribution for all conditions.

3.5.6 Simulations

The basic approach to simulating narrow-band time series was to firstly sample Gaussian white noise and subsequently apply band-pass filters to it ($3rd$ order butterworth, forward-only). Such signals M were then circularly shifted by a ground truth interaction delay. Finally, source and target time series X and Y were generated by adding independent white noise of varying amplitude to M. The following specific parameters were used for the individual simulations and analyses:

Figure	Number of samples	Noise amplitude source	Noise amplitude target	Effect filter [Hz]	Ground truth delay [s]	Analysis filter [Hz]	Lags [s]
1	75000	.5	.25	6 ± 2	$-.12$	6 ± 2	$-.8$ - $+.32$
2	75000	.5	.25	6 ± 2	$-.12$	6 ± 2	$-.8$ - $+.32$
3	40000	.05	.025	6 ± 2	$-.08$ - $-.4$	6 ± 2	.66 - $+.4$
4	40000	.001 - 10	.001- 10	6 ± 2	$-.12$	6 ± 2	$-.4$ - $+.2$
5	40000	0 - 10	0 - 10	6 ± 2	$-.12$	6 ± 2	$-.4$ - $+.2$
6.1	40000	.01	.01	15 ± 1 - 13.25	$-.12$	15 ± 1 - 13.25	$-.4$ - $+.2$
6.2	40000	.01	.01	20 ± 5	$-.12$	20 ± 1 - 15	$-.4$ - $+.2$
7.1	40000	.01	.01	3 - 50 ± 2	$-.12$	3 - 50 ± 2	$-.4$ - $+.2$
7.2	40000	.01	.01	3 - 47 ± 2	$-.12$	3 - 47 ± 2	$-.4$ - $+.2$
8	330000	*na*	*na*	*na*	*na*	1.5 - 16 ± 1	-1 - .5

Table 3.1: **Parameter settings in simulations.** Figure 6.1 refers to panels A, B, D and E, Figure 6.2 refers to panels C and F. Figure 7.1 refers to panels A - D, Figure 7.2 refers to panels E - H. Effect and analysis filters are specified in terms of the passbands as defined by centre frequencies $\pm$ bandwidths. *na* denotes “not available".

3.5.7 MEG data and analyses

The MEG data in this study has been recorded and analysed before. For details concerning the participants, the experimental design, the recording procedures as well as the preprocessing, please refer to (Daube et al., 2019b). In brief, 24 participants had listened to an audiobook of 55 minutes duration (in 6 blocks of equal duration) while their MEG (MAGNES 3600 WH, 4D Neuroimaging, 248 magnetometers) had been recorded at a sampling rate of 1017.25 Hz (first 10 participants) or 2034.51 Hz (last 14 participants). We applied the same pre-processing steps as in the original study. This included interpolation of artifactual channels, replacement of squid jumps with DC patches, 4th order zero-phase high-pass filter of .5 Hz as well as independent component analysis for removal of eye and heart activity (Daube et al., 2019b). Here, we then downsampled the data to a sampling rate of 100Hz.

To estimate activity from bilateral auditory cortices (ACs), we re-used linearly constrained minimum variance beamformer spatial filters (Van Veen et al., 1997) as in (Daube et al., 2019b). These had been optimised within a nested cross-validation to return responses that would be maximally correlated with linear predictions of the re-

sponses based on a combination of log-mel spectrograms and their temporal derivative. This correlation had been maximised with respect to position- and regularisation hyperparameters per participant, hemisphere and fold. Here, we averaged these hyperparameters across folds to then extract time series of activity from left and right ACs.

To define individual delta and theta frequency ranges, we considered spectra of delayed MI between AC activity and the speech envelope. To compute these spectra, we applied $3rd$ order butterworth forward filters to both AC activity and the envelope. These had centre frequencies increasing from 1.5 to 15 Hz in steps of $.25$ Hz, and cutoff frequencies were defined as bands of ± 1 Hz width around the centre frequencies. We computed delayed MI for a range of lags from -1 to .5s and used the maximum across lags within each frequency. We then searched for troughs in the resulting spectra and used the trough corresponding to the lowest frequency as the boundary between delta and theta. In 4 out of 48 cases, this returned boundaries between delta and theta that surpassed 6.5 Hz, which we considered as unlikely a priori (Klimesch, 1999; Wang, 2010). Therefore, in these cases we instead searched for peaks of the spectral gradient and used the peak corresponding to the lowest centre frequency. In practise, this corresponded to parts of the spectra where the initial decrease plateaued. To define an upper boundary of the theta component, we used the highest frequency at which the spectrum surpassed the noise threshold. This resulted in theta components with a broader bandwidth compared to the delta components. Since we had found the bandwidth of an effect to be correlated with the effect size recovered by DI_{ce}, we also repeated analyses of delayed MI and DI_{ce} with fixed bandwidths for delta and theta. For this, we defined the delta component to be below 3.5 Hz and the theta component to be centered on a theta peak frequency, around which we defined bandpass filters of 3 Hz width. The theta peak frequencies were defined as the maximum of the MI spectra within the theta frequency band as defined previously. In 4 out of 48 cases, this corresponded to the upper boundary of the delta band. In these cases, we defined the theta centre frequency to be 1.5 Hz above the delta-theta boundary.

To extract time series corresponding to the passbands as defined above, we used $3rd$ order forward butterworth filters, which we applied to both the envelope and the AC activity. We then subjected the resulting time series to the computation of delayed MI and DI_{ce}. We statistically evaluated the nth observation of recovered log-transformed effect sizes and the recovered delays r at the maxima across lags of each participant p, hemisphere h, frequency band b and measure m by fitting Bayesian linear distributional models as implemented in the brms package (Bürkner, 2017). These models can be

summarised with the following formula:

$$
\begin{aligned}
r_{[n]} &\sim N\left(\eta_{\mu[n]}, \exp\left(\eta_{\sigma[n]}\right)\right) \\
\eta_{\mu[n]} &\sim \beta_{\mu_{p[n]}} + \beta_{\mu_{p:h[n]}} + \beta_{\mu_{m:b[n]}} \\
\left(\beta_{\mu_{p[n]}}, \beta_{\mu_{p:h[n]}}, \beta_{\mu_{m:b[n]}}\right) &\sim N(0, v) \\
\eta_{\sigma[n]} &\sim \beta_{\sigma_{h[n]}]} + \beta_{\sigma_{m:b[n]}} \\
\left(\beta_{\sigma_{h[n]}]}, \beta_{\sigma_{m:b[n]}}\right) &\sim N(0, 10)
\end{aligned}
\tag{3.17}
$$

Chapter 4

Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity

published as:
Daube, C., Xu, T., Zhan, J., Webb, A. Ince, R.A.A., Garrod, O.G.B. and Schyns, P.G. (2021). Grounding deep neural network predictions of human categorization behaviour in understandable functional features: The case of face identity. *Patterns*, 2(10).

Permission to reproduce this article has been granted by Dr Sarah Callaghan, Editor-in-Chief at *Patterns*, Cell Press.

Author contributions:
C.D., J.Z., O.G.B.G., and P.G.S. designed the research.
C.D. and T.X. developed the DNN models.
O.G.B.G. and P.G.S. developed the GMF.
R.A.A.I. developed the implementation of PID.
J.Z. recorded the data.
C.D. analyzed the data.
C.D. and P.G.S. drafted the manuscript.
C.D., J.Z., A.W., R.A.A.I., and P.G.S. revised the manuscript.
P.G.S., O.G.B.G., and R.A.A.I. supervised the project.
P.G.S. acquired the financial support for the project leading to this publication.

4.1 Abstract

Deep neural networks (DNNs) can resolve real-world categorization tasks with apparent human-level performance. However, true equivalence of behavioral performance between humans and their DNN models requires that their internal mechanisms process equivalent features of the stimulus. To develop such feature equivalence, our methodology leveraged an interpretable and experimentally controlled generative model of the stimuli (realistic three-dimensional textured faces). Humans rated the similarity of randomly generated faces to four familiar identities. We predicted these similarity ratings from the activations of five DNNs trained with different optimization objectives. Using information theoretic redundancy, reverse correlation, and the testing of generalization gradients, we show that DNN predictions of human behavior improve because their shape and texture features overlap with those that subsume human behavior. Thus, we must equate the functional features that subsume the behavioral performances of the brain and its models before comparing where, when, and how these features are processed.

4.2 Introduction

Visual categorization is the pervasive process that transforms retinal input into a representation that is used for higher-level cognition, such as for memory, language, reasoning, and decision. For example, to guide adaptive behaviors we routinely categorize faces as being relatively happy, aged, or familiar, using different visual features. A long-standing challenge in the field of cognitive science is therefore to understand the categorization function which selectively uses stimulus features to enable flexible behavior (Schyns et al., 1998; DiCarlo & Cox, 2007; Nestor et al., 2020).

From a computational standpoint, this challenge is often framed as understanding the encoding function (Naselaris et al., 2011) that maps high-dimensional, highly variable input images to the lower-dimensional representational space of features that serve behavior. Deep neural networks (DNNs) have recently become the model of choice to implement this encoding function. Two key properties justify the popularity of DNNs: first, they can solve complex, end-to-end (e.g., image-to-behavior) tasks by gradually compressing real-world images over their hierarchical layers into highly informative lower-dimensional representations. Second, evidence suggests that the activations of DNN models share certain similarities with the sensory hierarchies in the brain, strengthening their plausibility (Yamins et al., 2014; Eickenberg et al., 2017; Kell et al., 2018; Kubilius et al., 2019; Kietzmann et al., 2019; Zhuang et al., 2021). Such findings underlie the surge of renewed research at the intersection between computational models, neuroscience, and cognitive science (Kriegeskorte & Douglas, 2018a).

However, there is ample and mounting evidence that DNNs do not yet categorize like humans. Arguably, the most striking evidence comes from adversarial examples, whereby a change in the stimulus imperceptible to humans can counter-intuitively

change its categorization in a DNN (Szegedy et al., 2014) and vice versa (Jacobsen et al., 2019). Even deceptively simple visual discrimination tasks reveal clear inconsistencies in the comparison between humans and state-of-the-art models (Rajalingham et al., 2018). Furthermore, when tested with photos of everyday objects taken from unusual perspectives, DNNs trained on common databases of naturalistic images decrease in test-set performance in ways humans do not (Barbu et al., 2019). In sum, although DNNs can achieve human-like performance on some defined tasks, they often do so via different mechanisms that process stimulus features different from those of humans (Geirhos et al., 2020; Golan et al., 2020).

These results suggest that successful predictions of human behavioral (or neural) responses with DNN models are not sufficient to fully evaluate their similarity, a classic argument on the shortcomings of similarity in cognitive science (Medin et al., 1993; Edelman, 1995). In fact, we already know that similar behaviors in a task can originate from two human participants processing different features (Schyns & Rodet, 1997). Generalizing to the comparison of a human and their DNN model, consider the example whereby both categorize a given picture as a horse. Should we conclude that they processed the same features? Not if the DNN learned to use the incidental horse-specific watermarks from the image database (Lapuschkin et al., 2019). This simple example illustrates both the general importance of attributing behavior to the processing of specific features, and the long-standing challenge of doing so, especially given the dense and unknown correlative structure of real-world stimuli (Schyns et al., 2003). From an information-processing standpoint, we should know what stimulus information (i.e., features) the brain and its DNN models process, before comparing where, when, and how they do so (Marr, 2010; Krakauer et al., 2017). Otherwise, we risk studying the processing of different features without being aware of the problem (cf. watermark example above). Thus, to realize the potential of DNNs as information-processing models of human cognition (Kay, 2018), we need to first take a step back and demonstrate that similar behavior in a task is grounded in the same stimulus features—i.e., more specifically, in similar functional features: those stimulus features that influence the behavioral output of the considered system (Schyns et al., 1998). When such functional feature equivalence is established, we can meaningfully compare where, when, and how the processing of these same functional features is reduced with equivalent (or different) algorithmic-implementation-level mechanisms in humans and their models.

To develop such equivalence of functional features, we explicitly modeled stimulus information with an interpretable generative model of faces (GMF, Zhan et al., 2019a). The GMF allows parametric experimental control over complex realistic face stimuli in terms of their three-dimensional (3D) shape and two-dimensional (2D) RGB texture. As illustrated in figure 4.1, a candidate DNN model is typically evaluated on how it predicts human responses, by computing the bivariate relationship between human responses and DNN predictions. Here, we further constrained this evaluation by relating human behavioral responses and their DNN predictions to the same set of experimentally con-

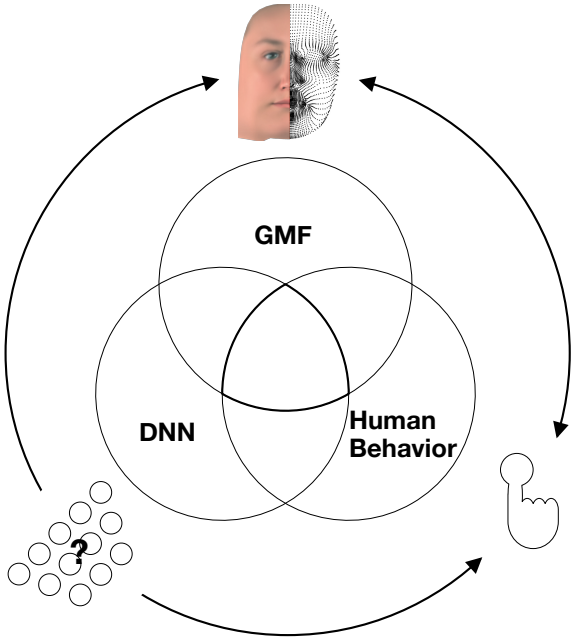

Figure 4.1: **Trivariate relationship to understand the functional features of DNN models that predict human behaviour.**
In general, complex visual inputs are processed in an unknown way in the brain and its DNN models to produce behavior. DNNs (schematized as layers of neurons) can predict human behavior and can in principle be used to facilitate our understanding of the inaccessible information-processing mechanisms of the brain. However, nonlinear transformations of information in DNNs complicate our understanding, in turn limiting our understanding of the mechanistic causes of DNN predictions (and human behavior). To address this issue of interpretability, we used a generative model of realistic faces (GMF) to control the high-level stimulus information (3D shape and RGB texture). The Venn diagram illustrates the logic of our approach. Human behavior and its DNN model predictions are both referred to in the same stimulus model: (1) the GMF features that underlie human behavior; (2) the GMF features that underlie DNN predictions of human behavior. The question then becomes: are these GMF features equivalent? That is, do the two intersections intersect (Schyns et al., 2020)? We quantify GMF feature overlap with information theoretic redundancy (Ince, 2017a; Daube et al., 2019a)–i.e., as the information that GMF features and the activations of the embedded layers of DNN models provide about human behavior. In doing so, we assess the functional feature equivalence of individual human participants and their DNN models in relation to a specific model of the stimulus and behavioral task. See figure 4.3 for a detailed overview of the analysis pipeline. Our results develop why such feature equivalence enhances our understanding of the information-processing mechanisms underlying behavior in the human brain and its DNN models. Experiment on human participants was conducted by Jiayu Zhan.

trolled GMF features. Conceptually, this is represented as the triple intersection in figure 4.1, where the pairwise intersections <GMF features; human> and <GMF features; DNN predictions> comprise the functional face features that subsume human responses and their DNN models. The triple intersection further tests whether the same responses in the two systems arise from the same face features, on the same trials. We then compared how each candidate DNN model represents these face features to predict human behavior and reconstructed the internal face representations of humans and their DNN

models with reverse correlation (Murray, 2011). Lastly, and importantly, we used our generative model to compare the generalization gradients of humans and DNNs to typical out-of-distribution stimuli (i.e., generalizations to changes of face pose, age, and sex to create siblings with family resemblance). With this approach, we ranked models not only according to their surface similarity of predicted human behavior but also according to the deeper similarity of the underlying functional features that subsume behavioral performance.

4.3 Results

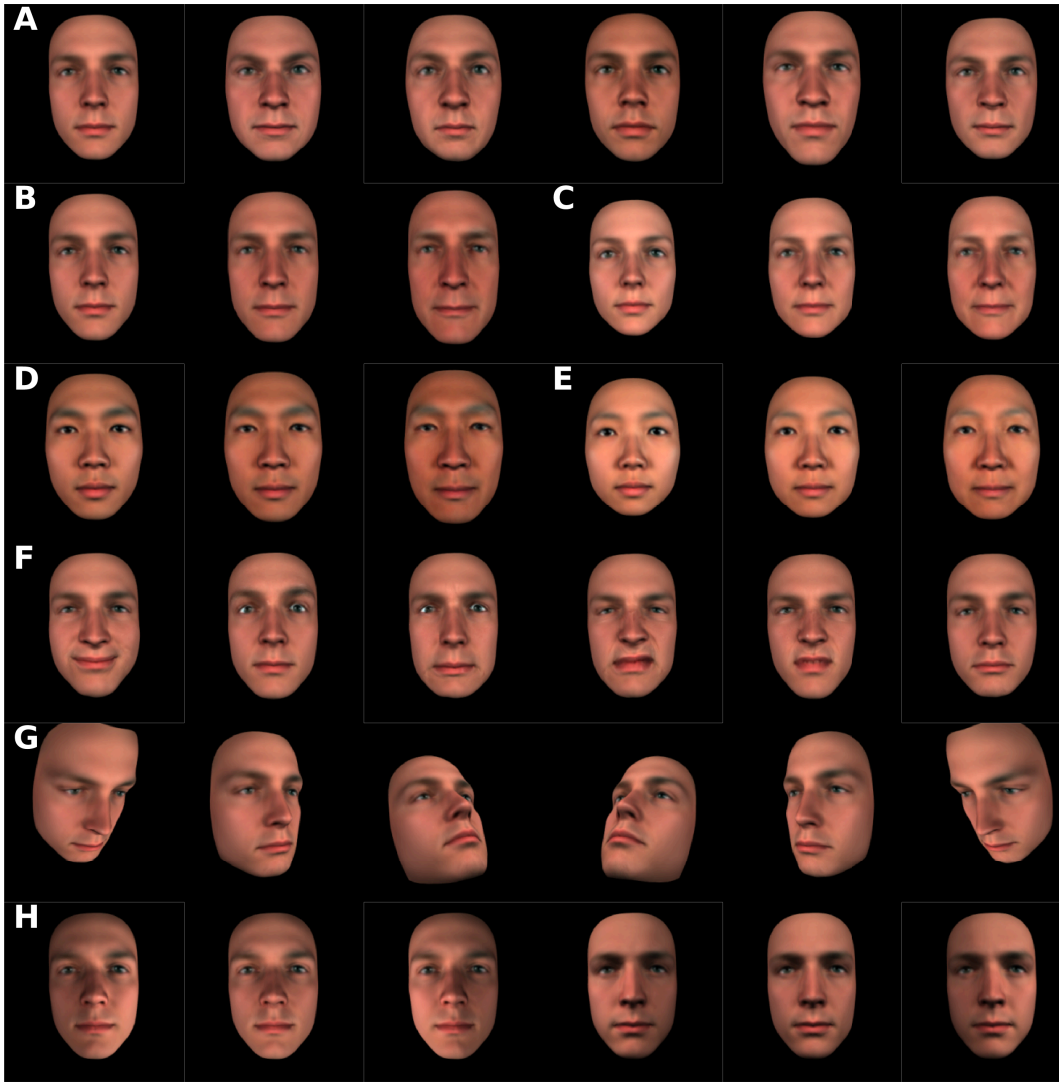

Figure 4.2: **Demonstration of GMF variations used for training set of DNNs (related to figure 4.3).**
A Six different example identities. **B** First identity from A rendered in three different ages. **C – E** Same as in B, but rendered with different sex and ethnicity. **F** First identity from A rendered with 6 additional expressions. **G** First identity from A rendered with different viewing angles. **H** First identity from A rendered with different lighting angles. Experiment on human participants was conducted by Jiayu Zhan.

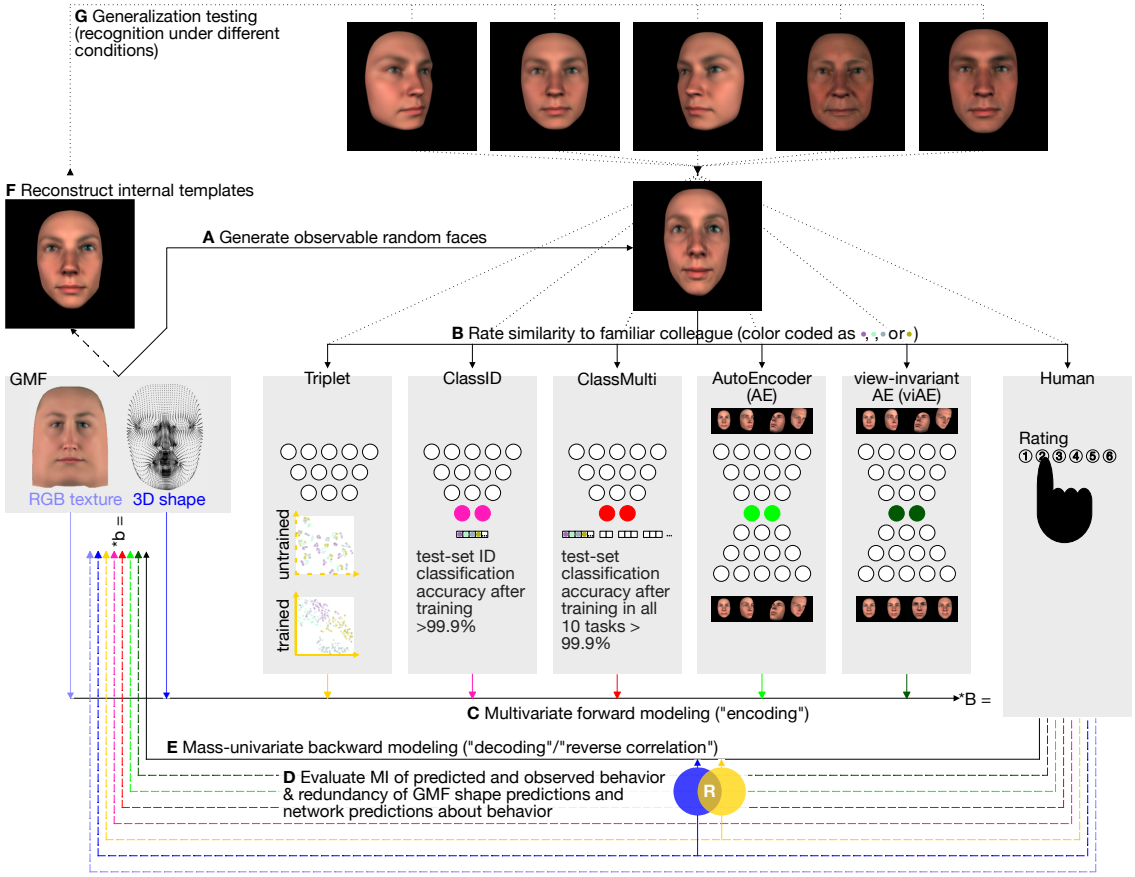

Figure 4.3: **Study overview.**
We seek to establish the GMF feature equivalence between humans and their DNN models. **A** We used the GMF to synthesize random faces (3D shape and RGB texture). **B** We asked humans to rate the similarity of these synthesized faces to the faces of four familiar colleagues (symbolized by purple, light-blue, gray, and olive dots). **C** Linear multivariate forward models predicted human responses (denoted by the multiplication with linear weights B) from GMF shape and texture features and DNN activations (DNN architectures are schematized with white circles symbolizing neurons, embedding layers are colored; scatterplots for Triplet network show two-dimensional t-stochastic neighborhood embeddings (Van der Maaten & Hinton, 2008) of the embedding layer when activated with 81 different combinations of viewing and lighting angles per colleague). As a baseline model, we also included the first 512 components of a principal components analysis on the pixel images ("pixelPCA," not shown here). **D** We then evaluated shared information between human behavior, DNN predictions from embedded activations, and GMF features using partial information decomposition (Ince, 2017a). Here, the Venn diagram shows the mutual information (MI) between human responses and their predictions based on the GMF shape features (blue circle) or based on the Triplet model (yellow circle). The overlapping region denotes redundancy (R). **E–G** We performed reverse correlation E to reconstruct internal templates F of the familiar colleague faces from human and model predicted behavior. Lastly, we amplified either the task-relevant versus task-irrelevant features of the four colleagues (identified in E) and rendered these faces in five different generalization conditions G that humans and DNNs had to identify. See also figure 4.2. Experiment on human participants was conducted by Jiayu Zhan.

We used a generative model that parameterizes faces in terms of their 3D shape and 2D RGB texture (GMF; experimental procedures 4.5.1) to control the synthesis of 3

million 2D face images that varied in identity, sex, age, ethnicity, emotion, lighting, and viewing angles (see figure 4.2 for a demonstration; see experimental procedures 4.5.4 for details). We used these images to train five DNNs that shared a common ResNet (He et al., 2015) encoder architecture but differed in their optimization objectives.

The five DNNs were as follows (see figure 4.3 for their schematic architectures and performances): (1) a triplet loss network (Schroff et al., 2015) that learned to place images of the same (versus different) identity at short (versus long) Euclidean distances on its final layer; (2) a classification network (Xu et al., 2018) that learned to classify $2,004$ identities ($2,000$ random faces, plus four faces familiar to our participants as work colleagues, "ClassID"); (3) another classification network that learned to classify $2,004$ identities plus six other factors of variation of the generative model ("ClassMulti"); (4) an autoencoder (AE) (Ballard, 1987) that learned to reconstruct all input images; and (5) a view-invariant autoencoder (viAE) (Zhu et al., 2013) that learned to reconstruct the frontal face image of each identity irrespective of the pose of the input.

We used these five DNNs to model the behavior of each of $n = 14$ individual human participants who resolved a face familiarity experiment (see experimental procedures 4.5.2). In this experiment, participants were asked to rate, from memory, the similarity of random face stimuli generated by the GMF (figure 4.3A) to four familiar identities (see experimental procedures 4.5.3). On each of $1,800$ trials, each participant was presented six random faces. They were asked to first choose the face most similar to a target identity and then rate this similarity on a 6-point scale.

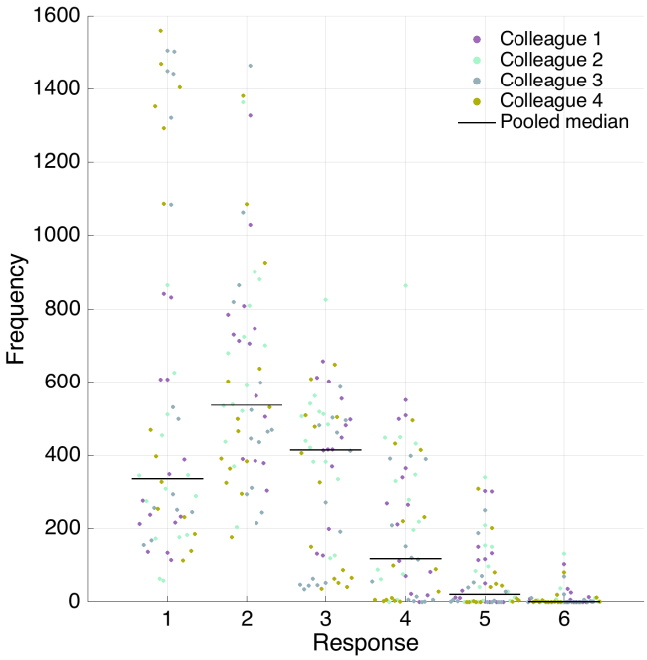

Figure 4.4: **Distribution of rating responses in the human reverse correlation experiment (related to figure 4.5).**
1 codes for low similarity, 6 codes for highest similarity of stimulus to familiar target identity. Each data point represents the combination of one participant and one target familiar identity. Experiment on human participants was conducted by Jiayu Zhan.

Importantly for our modeling, we propagated these 2D images through the five DNNs and then used the activations of their respective layer of maximum compression (i.e., the “embedding layer”) for the subsequent analyses detailed below. To assess functional feature equivalence between human participants and the DNN models, we proceeded in four stages (see figure 4.3 for an overview of our pipeline). First, we used the representations of the experimental stimuli on the DNNs’ embedding layers to predict the corresponding behavior of humans in the experiment (figure 4.3C and D). We did so using linear models to restrict the assessment to explicit representations (Naselaris et al., 2011). We call this first stage of seeking to equate human and DNN behavior “forward modeling". In a second stage, we analyzed the face features represented on the DNN embedding layers that predict human behavior. In a third stage (figure 4.3E and F), we used reverse correlation to reconstruct and compare these categorization features between humans and their DNN models. Lastly, in a fourth stage (figure 4.3G), we compared the generalization performances of humans and DNNs under new testing conditions of face viewing angles, sex, or age that did not appear in the data used to fit the forward models.

Previewing the results of the DNN models tested, the viAE afforded the best predictions of human behavior. These could be attributed to the shape features of the GMF, which also subsumed human behavior. That is, the surface similarity of behavioral performance was grounded in a deeper similarity of functional face features. Of the DNN models tested, the viAE model was therefore the most functionally similar to humans.

4.3.1 Forward modeling of human behavior using DNN activations

To evaluate how accurately the compressed stimulus representations on the DNNs’ embedding layers predicted the face similarity ratings (on a 6-point rating scale, see figure 4.4) of human participants, we activated their embedding layers with the $1,800$ 2D face stimuli rated in terms of similarity to each target identity in the human experiment. We then used these activations to linearly predict (see experimental procedures 4.5.5) the corresponding human ratings in a nested cross-validation (Varoquaux et al., 2017). We compared DNN performances with three additional benchmark models that also linearly predicted human behavior. The first model used on each trial the objective 3D shape parameters of the GMF that define the identity of each face stimulus (rather than the face image); the second one used instead the GMF texture parameters (cf. Figures 4.1 and 4.1, and 3D shape and 2D RGB texture). Finally, the third model was a simpler architecture that linearly predicted human behavior from the first 512 components of a principal components analysis (PCA) of all stimulus images (“pixelPCA”). For each model, we evaluated predictions of human behavior with two information theoretic quantities (figure 4.5A and B). With mutual information (MI), we quantified the strength of the relationship between the observed human and DNN predicted similarity ratings (figure 4.5A and B, y-axes). Importantly, we also used redundancy (from partial information decomposition)

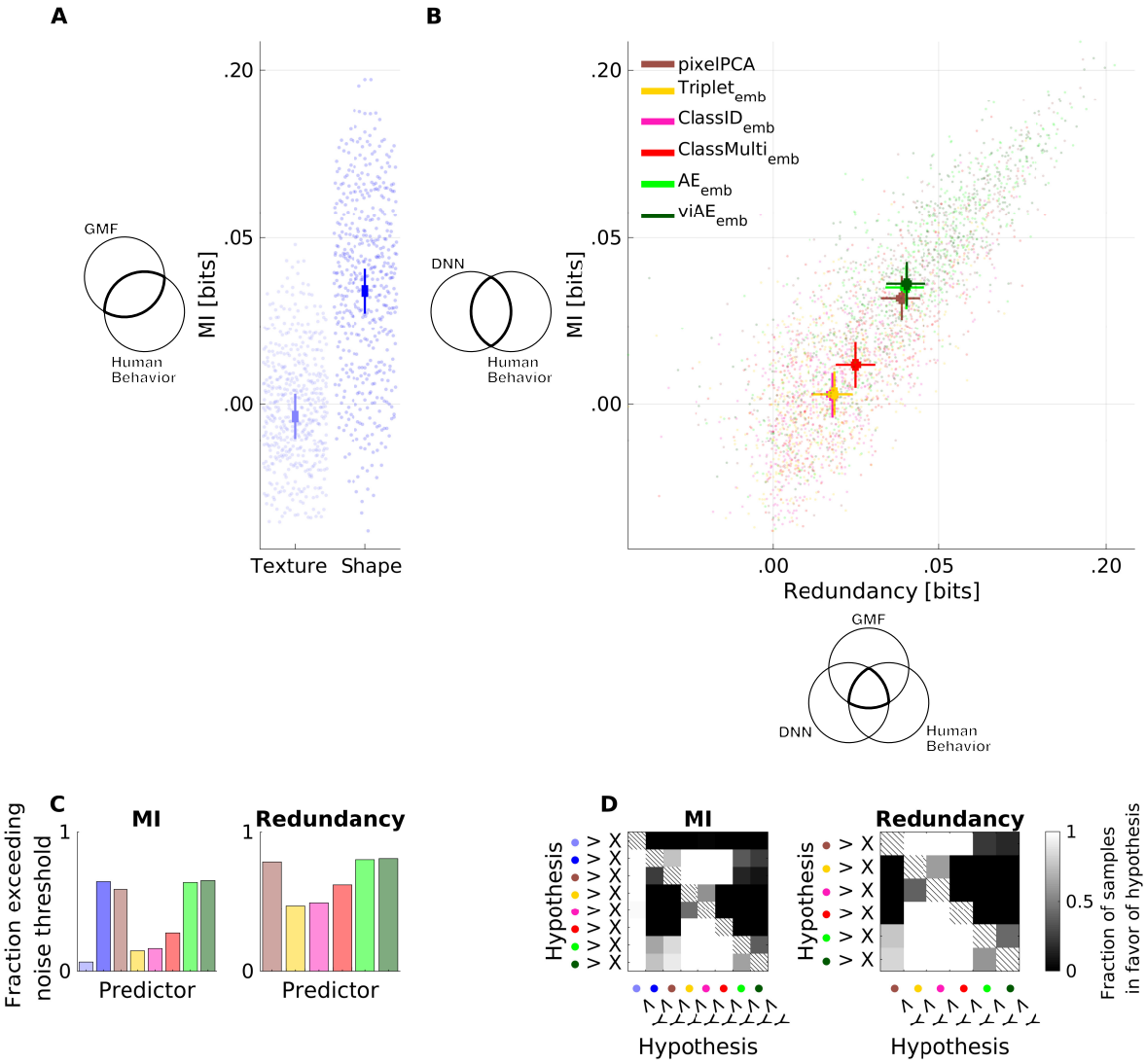

Figure 4.5: Caption on following page.

to evaluate the triple set intersection of figure 4.1, which quantifies the overlap between predictions from DNN models and predictions from GMF shape parameter models (figure 4.5B, x-axes). This overlap indicates the extent to which the DNN embedding layers and the GMF shape parameters both predict the same human behaviors on the same trials. With Bayesian linear models (Bürkner, 2017), we then statistically compared the bivariate relationships (i.e., MI) and overlaps (i.e., redundancy) of different GMF parameters and DNN embedding layers with each other.

Of all models, the viAE best predicted human behavior (see figure 4.5B), closely followed by the AE, with a performance level similar to that of the GMF shape parameters (fraction of samples of posterior in favor of viAE over shape: $f_{h_1} = 0.7536$; AE $>$ shape: $f_{h_1} = 0.6457$; $f_{h_1} = 0$ for all other networks versus shape). Surprisingly, the simple pixelPCA came close to the complex AEs (with the AE only narrowly beating pixelPCA, $f_{h_1} = 0.8582$, figure 4.5B). Critically, as model predictions increased in accuracy, they also increased in overlap (i.e., redundancy) with the GMF shape parameters (figure 4.5B), implying that single-trial behavior across systems (i.e., humans, viAE, and pixelPCA) could be attributed to these same specific parameters of 3D face shape–i.e.,

Figure 4.5 *(previous page)*: **Relationship among GMF features, DNN activations, and observed behavior.**
A Mutual information (MI) between human behavior and test-set predictions from GMF features. **B** y-axis: MI between human behavior and test-set DNN predictions; x-axis: redundant information about human behavior that is shared between DNN predictions and GMF shape feature predictions. These plots show that DNN prediction performance of human behavior increases on the y-axis when the DNN embedding layers represent the same shape features as humans. Each data point in A and B represents the combination of one test set, one participant, and one familiar identity. Overlaid lines reflect the 95% (bold) and 50% (light) highest posterior density intervals (HPDIs) of the corresponding main effects of predictor spaces from Bayesian linear models fitted to the MI and redundancy values. **C** Fractions of MI and redundancy data points exceeding noise threshold (95th percentile of MI and redundancy distributions obtained from trial-shuffled data). **D** Comparisons of the posterior distributions of the main effects for all predictor spaces from Bayesian linear modeling of the raw performances. For each pair in the matrices, the grayscale color map shows the fraction of samples of the predictor space color coded on the y-axis that is larger than the predictor space color coded on the x-axis (testing a hypothesis). Colors in C and D correspond to those in A and B. See also Figures 4.4–4.11 and figure 4.27. Experiment on human participants was conducted by Jiayu Zhan.

under these conditions they used the same functional face features to achieve the same behaviors.

Furthermore, we validated this overlap in shape parameters by showing that a model using jointly (vi)AE activations and GMF shape parameters (versus (vi)AE activations on their own) did not improve prediction of human behavior (see Figures 4.7 and 4.11 for additional candidate models, including combinations of the predictor spaces reported here, weighted and unweighted Euclidean distances, variational AEs, and decision neuron activities; see figure 4.8 for the same comparison using Kendall's tau as an evaluation metric; see figures 4.9 and 4.10 for a model comparison on the across-participant average). Note that the performances of these models could not be reached when predicting the behavior of participants with the behavior of other participants (see Figures 4.6–4.8). This means that participants behaved in systematically idiosyncratic ways.

In sum, in our first stage to assess functional equivalence between humans and their DNN models, we built forward models that predicted human behavior from the DNNs' embedding layers. The embedding layer of the (vi)AE won. We further showed that better predictions of human behavior from the embedding layers of DNNs were caused by their increased representation of the 3D face features that predict human behavior. However, a simple PCA of the pixel images performed competitively. At this stage, we know that better predictions of human behavior are caused by better representations of the 3D shape features that humans use for behavior. Next, we characterized what these 3D features are.

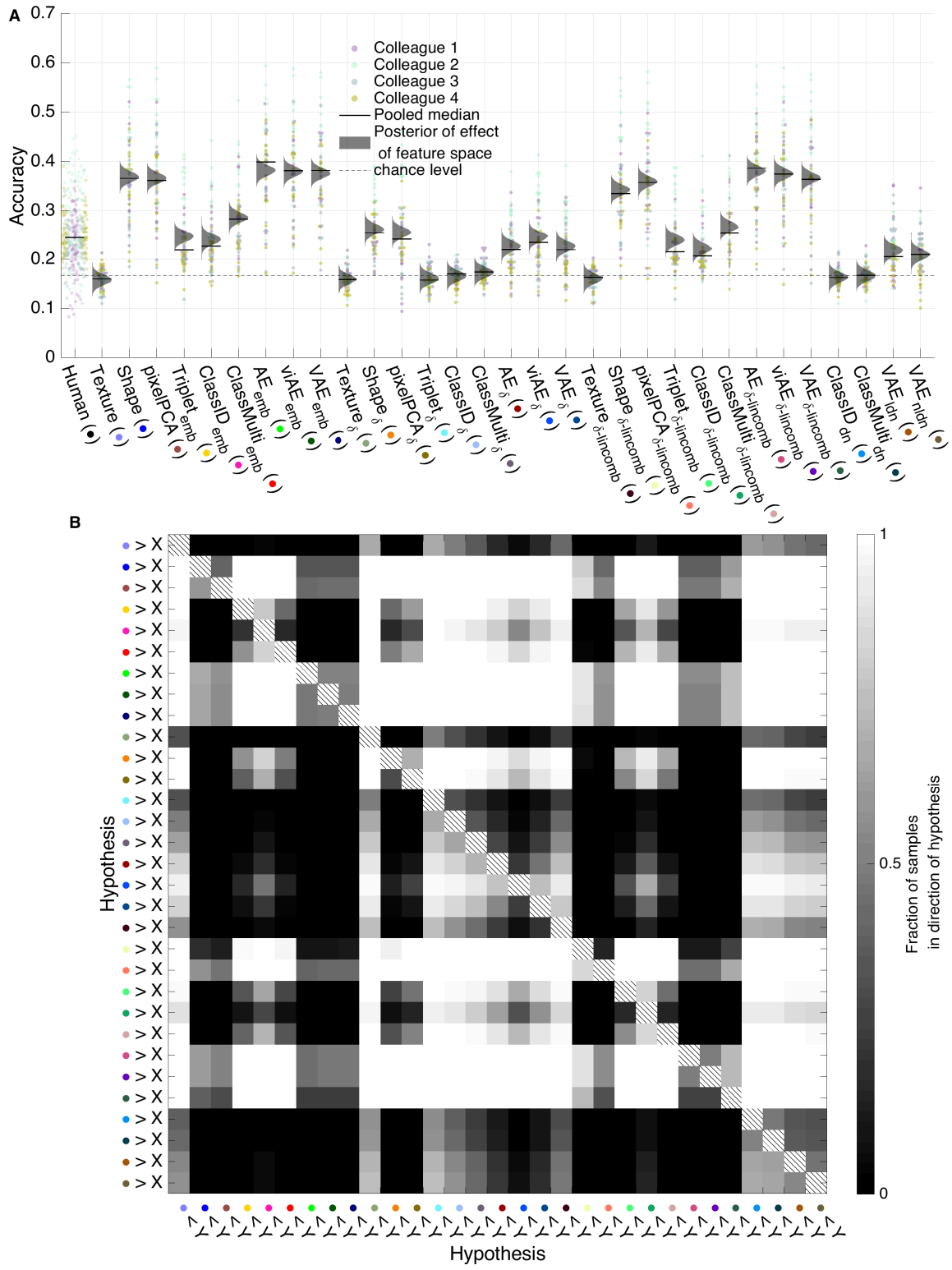

Figure 4.6: Caption on following page.

Figure 4.6 *(previous page)*: **Accuracy of forward models in predicting choice behavior (related to figure 4.5)**.
A Choice accuracy. On each trial, humans were presented with an array of 6 different random faces. They were asked to choose the one that most resembled the respective target colleague prior to reporting the perceived similarity on a 6-point rating scale. On each trial, the forward models "chose" the face of the array of 6 that had the highest rating among all faces of the array. The panel shows how well each model's choices matched the choices of the human participants. Pairwise matches of human participants with each other are displayed for reference. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all forward models from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the system color coded on the y-axis that is larger than the system color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

Figure 4.7 *(next page)*: **Bivariate evaluations of a larger set of encoding models (related to figure 4.5)**.
A Mutual Information (MI) of observed behavior and test-set predictions from GMF features and various functions of DNN activations as well as human participants predicting other human participants (pairwise comparisons). Models include variational autoencoders ("VAE", (Kingma & Welling, 2014)), VAEs with regularization ("β-VAE", (Higgins et al., 2016)), euclidean distances of representations of the ground truth colleagues and the respective trials ("δ"), weighted euclidean distances ("δ-lincomb"), pre-softmax decision neuron activity ("logits") of the respective colleagues of ClassID and ClassMulti networks ("dn") as well as of ID classifiers trained on top of frozen VAE encoder networks (linear, "VAE$_{ldn}$", and with 2 rectified fully connected layers of 512 neurons ("VAE$_{nldn}$"). **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

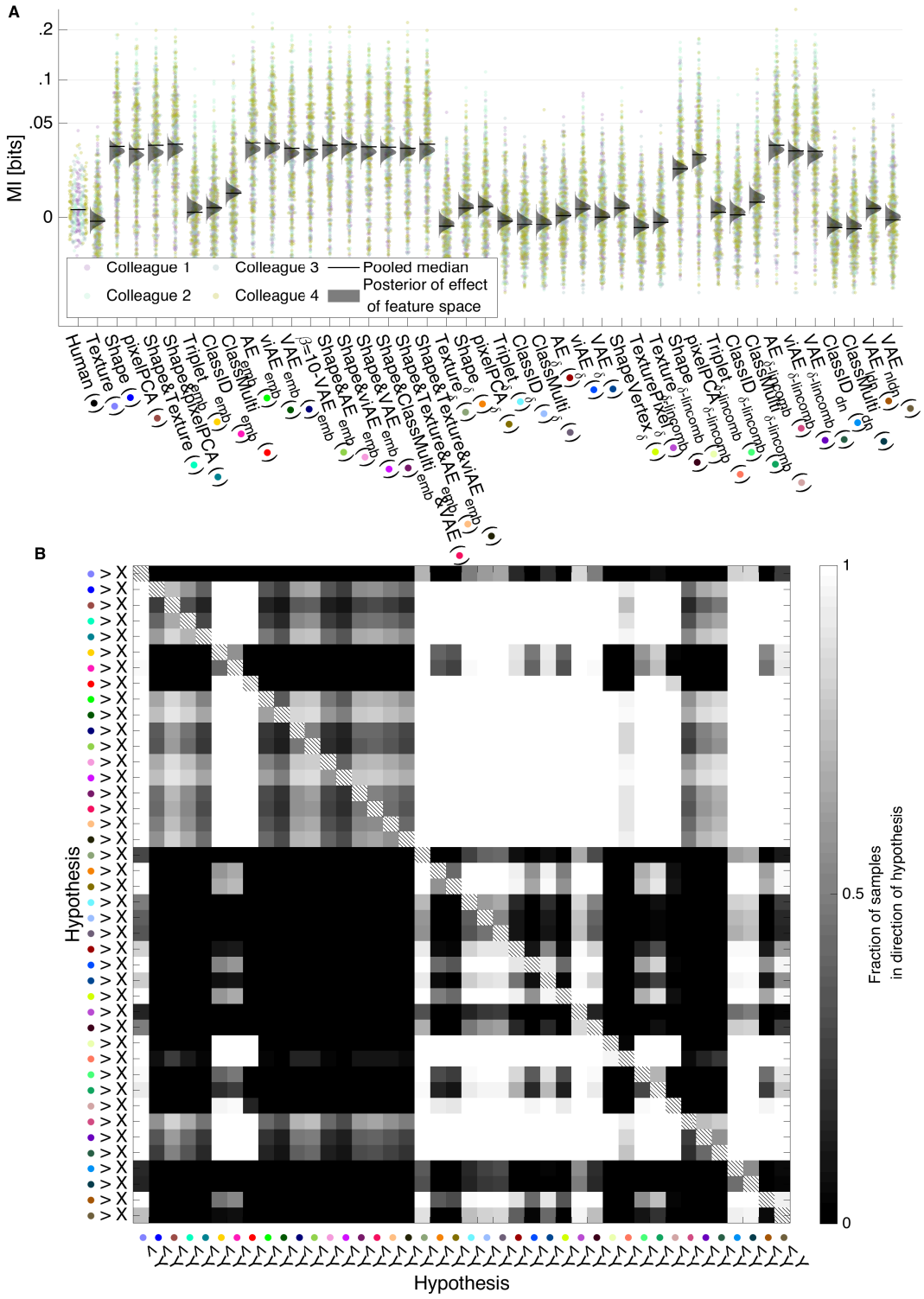

Figure 4.7: Caption on previous page.

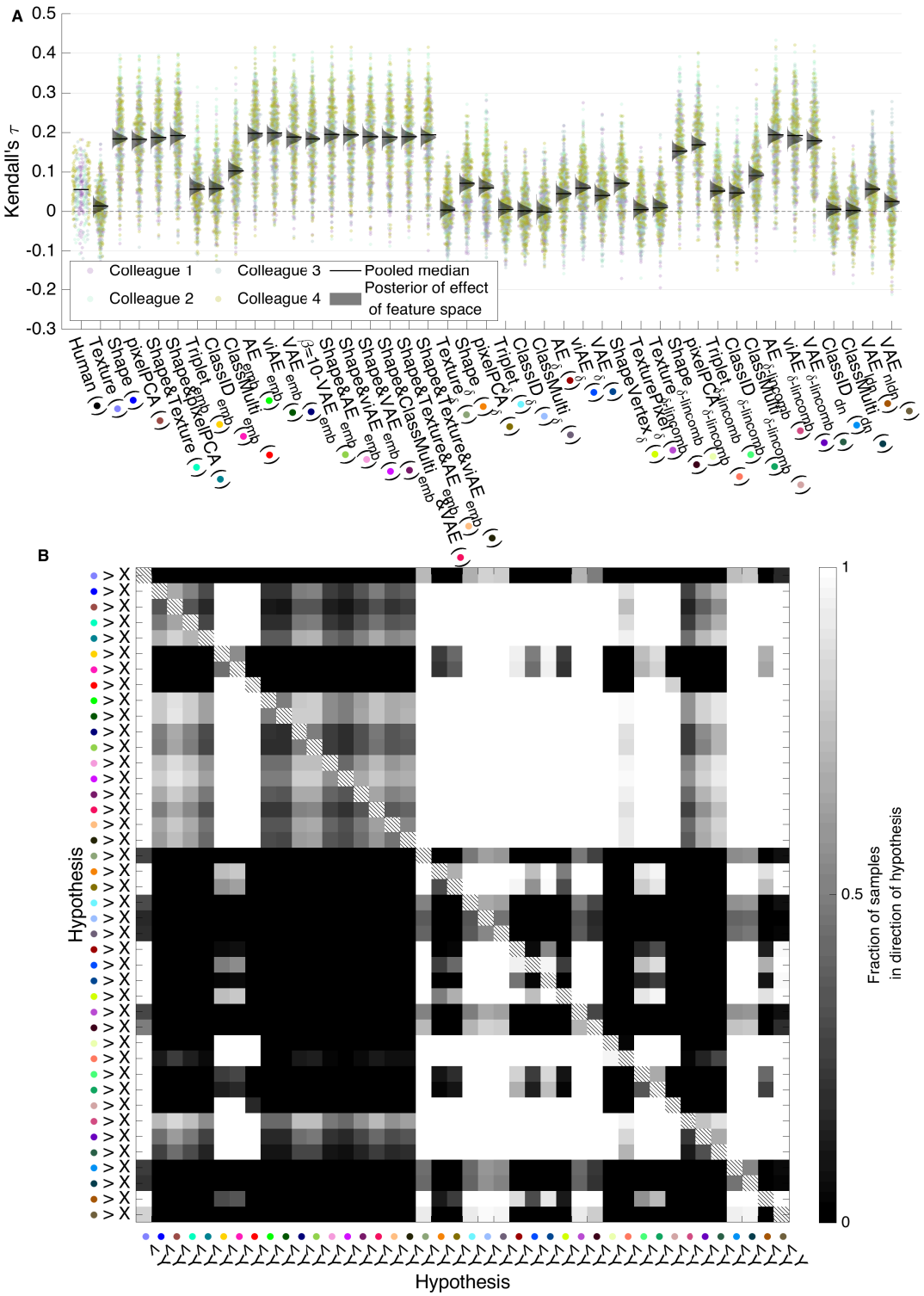

Figure 4.8: **Bivariate evaluations of a larger set of encoding models (related to figure 4.5)**.
A Kendall's τ of observed behavior and test-set predictions from GMF features and DNN activations as well as human participants predicting other human participants (pairwise comparisons). Except for the different metric, the analysis of this figure is identical to figure 4.7. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

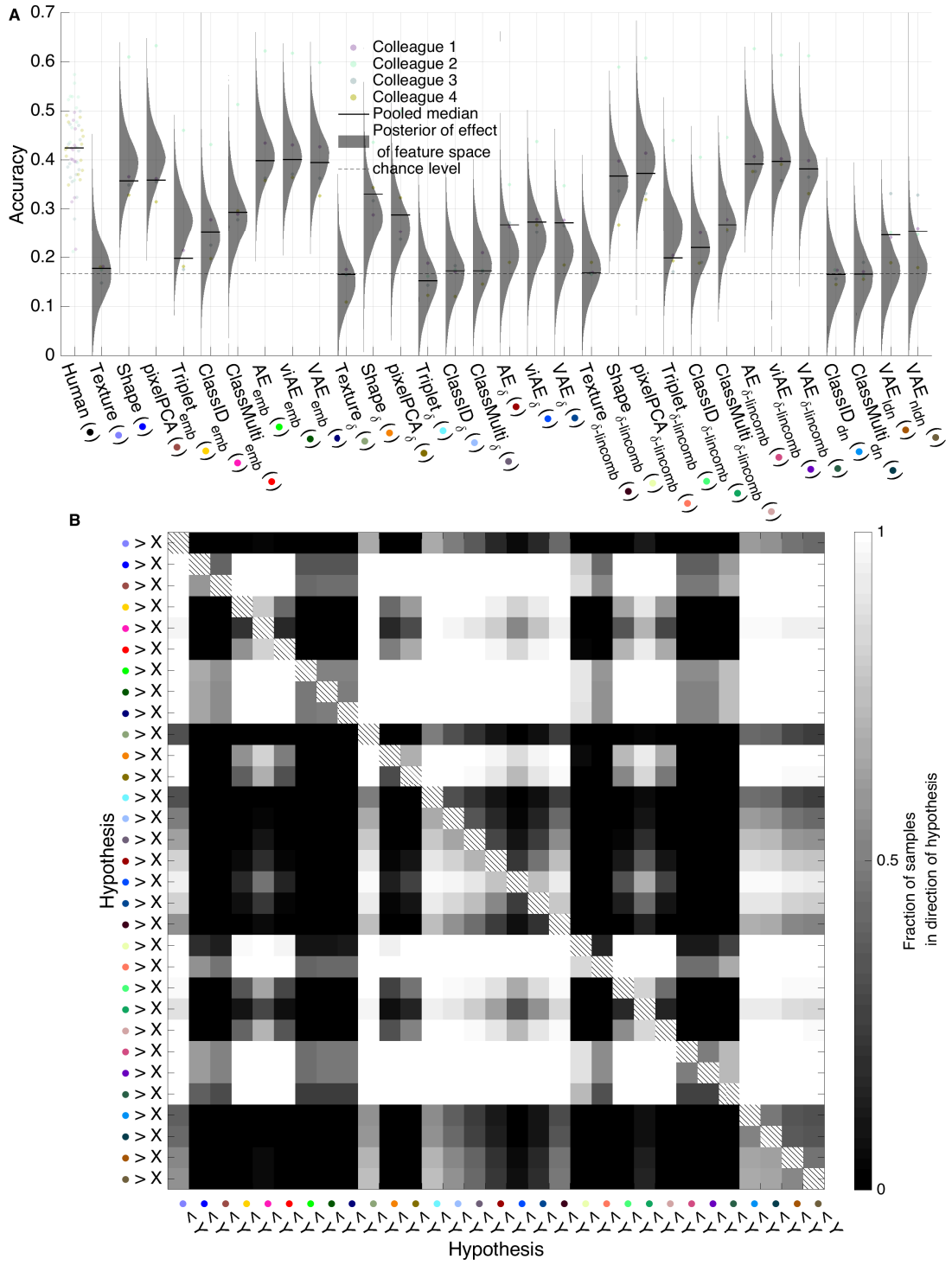

Figure 4.9: **Accuracy of forward models in predicting choice behavior consensus across participants (related to figure 4.5)**.
A Choice accuracy. Instead of predicting the behavior of individual human participants as in figure 4.6, here, for each panel of 6 faces per trial, the option chosen by the highest number of participants was used to represent the consensus across participants. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

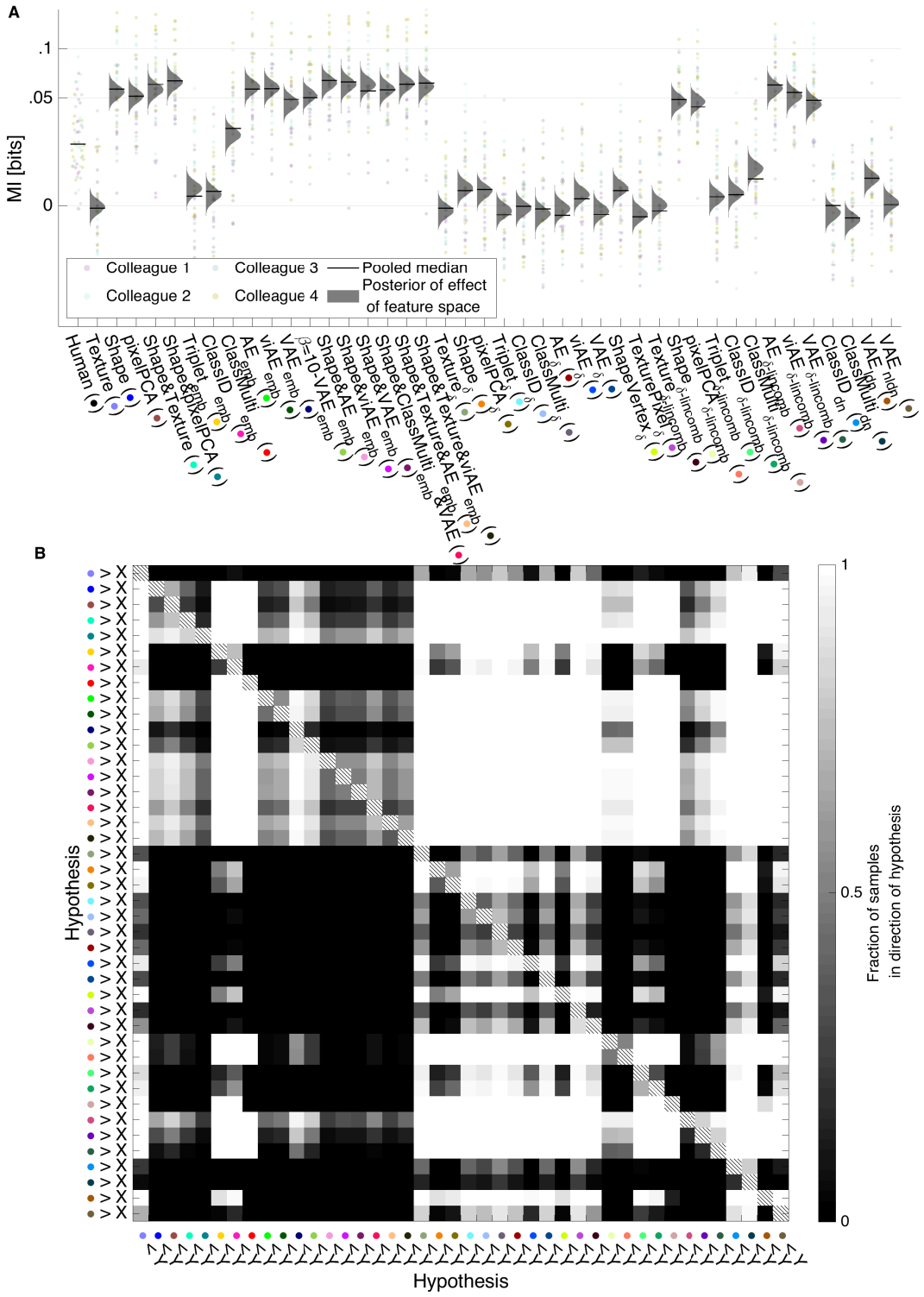

Figure 4.10: **Bivariate evaluations of a larger set of encoding models on ratings averaged across participants (related to figure 4.5).**
A Mutual Information of averaged behavior and test-set predictions from GMF features and DNN activations as well as human participants predicting other human participants. Except for the different predictee, the analysis of this figure is identical to figure 4.7. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

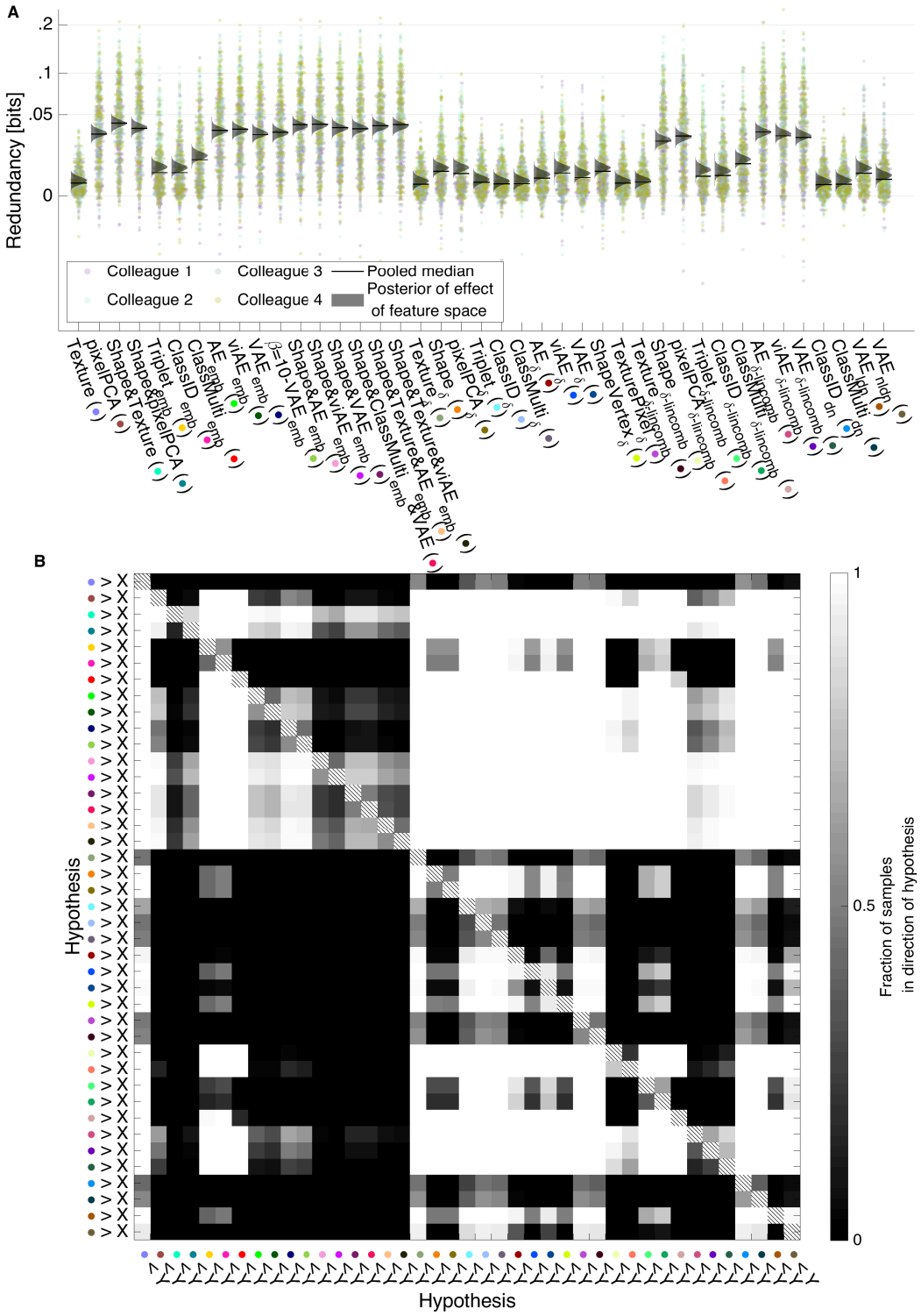

Figure 4.11: **Redundancy with shape of a larger set of encoding models (related to figure 4.5)**.
A Redundant information about human behavior that is shared between model predictions and GMF shape feature predictions. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw redundancies. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

4.3.2 Embedded face-shape features that predict human behavior

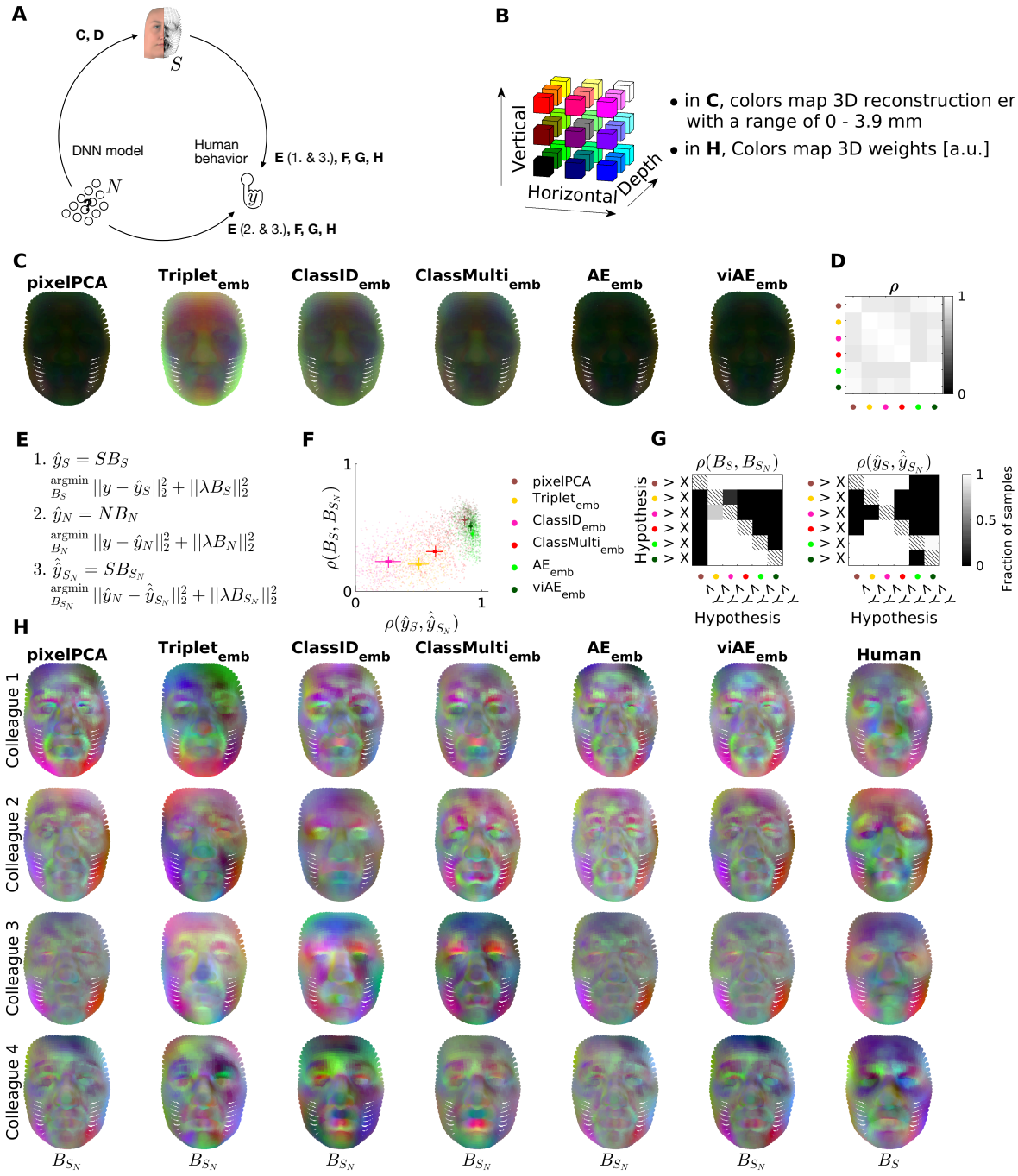

Figure 4.12: Caption on following page.

The viAE learned to represent on its embedding layer, from 2D images, the face-shape features that provide the best per-trial prediction of human behavior. Here, we establish: (1) how the DNNs represent these face-shape features on their embedding layers; and (2) how each feature impacts behavioral prediction in the forward models discussed in stage 1 above. We did not analyze the GMF texture features further because they could not predict human behavior (see figure 4.5).

Figure 4.12 *(previous page)*: **DNN representations of face-shape features for the forward linear models of human behavior.**
A Schema of the analyses. **B** Legend for 3D color codes in C and H. **C** Linear readout of face-shape features from the embedding layers of the five DNNs, where readout fidelity of GMF parameters is plotted per face vertex as the mean absolute error (MAE, averaged across a large set of test faces). Higher fidelity (lower MAE) of (vi)AE activations (compared with other DNNs) shows they better represent GMF shape features. **D** Correlation matrix of error patterns across DNNs. Colored dots on x and y axes represent each DNN model (see F for a legend). Correlating the MAE patterns from C across models reveals a high similarity of errors across models: vertices that are difficult to decode from Triplet activity are also difficult to decode from viAE activity. **E** Simulating DNN predictions of observed human behavior with GMF shape features using re-predictions. First, we estimate B_S, the shape receptive fields (SRFs) that predict human behavior from GMF shape features. Second, we estimate B_N, the weights that predict human behavior from DNN activations. Third, we estimate B_{S_N} , the SRFs that predict DNN predictions of human behavior from GMF shape features. **F** Aggregated SRF results from all participants and target familiar colleagues. x-axis: correlations between original DNN predictions of human behavior and the simulated predictions; y-axis: correlations between the human SRFs with DNN SRFs. The ideal DNN model should be located in the top right corner. The (vi)AE comes closest to this location. Each dot is one test set of one participant in one target familiar colleague condition. Overlaid crosses denote 95% (bold) and 50% (light) HPDIs of main effects of feature spaces from Bayesian linear models of the raw results. **G** Comparisons of the posterior distributions of main effects of the models from Bayesian linear modeling of the results in F. **H** Weight profiles of forward models (SRFs) plotted on 3D scatter of vertices. From the left, simulated shape weights of each DNN forward model (see main text, schematic in A, and equations in E for explanations) and weights of the direct GMF shape forward model of human responses. Plots show results from a typical participant with the lowest average difference from the six pooled group medians in F. Color coding in D, F, and G is the same. Experiment on human participants was conducted by Jiayu Zhan.

Face-shape features represented on the embedding layers of DNNs

To reveal these face-shape features, we built linear decoding models. These used the embedding layer activations to predict the positions of individual 3D vertices (see experimental procedures 4.5.6). We then evaluated the fidelity of their reconstructions with the Euclidean distance between the linearly predicted and the objective 3D face vertex positions. Fidelity increased from the Triplet to the two classifier networks, to the (vi)AE (which had the lowest error, see figure 4.12C). The pixelPCA achieved a similarly low error, and all models shared a common type of reconstruction errors (figure 4.12D) which misrepresented the depth of the peripheral and nasal face regions.

Patterns of face-shape features that predict behavior in the DNN forward models

To better understand the shape features that the aforementioned forward models used to predict human behavior, we examined their linear weights (see experimental proce-

dures 4.5.5). The forward GMF shape model weights directly relate a 3D shape space to human behavior. Thus, their weights form an interpretable face-space pattern that modulates behavior–i.e., a "shape receptive field" (SRF), see figure 4.12H (rightmost column). In contrast, the forward models based on the DNN relate (i.e., linearly weigh) DNN activations, not GMF shape parameters, to human behavior. Thus, we used an indirect approach to interpret these weights. We built auxiliary forward models that simulated (i.e., linearly re-predicted, figure 4.12E) the DNN predictions of human behavior, but this time using the GMF shape parameters instead of the embedding layers. This produced interpretable SRFs (figure 4.12H) with which we could therefore understand which shape features are (or are not) represented on the DNN embedding layers to predict human behavior. Specifically, we reasoned that DNN activations and GMF features would similarly predict behavior if: (1) both shared the same SRF; and (2) predictions from DNN activations were similar to their simulations based on GMF features. Our analyses revealed that the (vi)AE best satisfied these two conditions (figure 4.12F and G). PixelPCA features were again close to the performance of the best DNN models (figure 4.12F). In this second stage to assess functional feature equivalence, we identified, at the level of individual 3D face vertices, the shape features that DNNs represent to predict (cf. "forward modeling of human behavior using DNN activations") human behavior. Of all five DNNs, we found that the (vi)AE represents face-shape vertices most faithfully, leading to the most accurate predictions of human behavior. However, the simpler pixelPCA used apparently very similar features.

4.3.3 Decoding the shape features with reverse correlation

So far, we have assessed the functional equivalence between human behavior and DNN-based forward models in two stages: we have quantified to what degree the DNN model predictions of human behavior are attributable to GMF face-shape parameters (in stage 1), and we have characterized how the DNN models used specific patterns of face-shape parameters to predict behavior (in stage 2). In this third stage, we use the behavior observed in humans and predicted by DNN models to reconstruct, visualize, and compare the actual 3D shape features of the target faces represented in both humans and their DNN models. To run the human experiments (Zhan et al., 2019a) with the DNN models, we proceeded in three steps (see experimental procedures 4.5.7). First, we used the forward models described in stage 1 to predict human behavior in response to all face stimuli of the human experiment (6*3*1,800 = 10,800 face stimuli per familiar target face, Zhan et al., 2019a). On each trial, the forward models "chose" the face stimulus with the highest predicted rating from an array of 6 (see figure 4.6). This resulted in $1,800$ chosen faces and their corresponding similarity rating predictions. Second, for each model and participant, we regressed (mass univariately) the GMF parameters of the chosen faces on the corresponding ratings to derive a slope and intercept per GMF shape and texture parameter. Third, we multiplied these slopes by individual "amplification values"

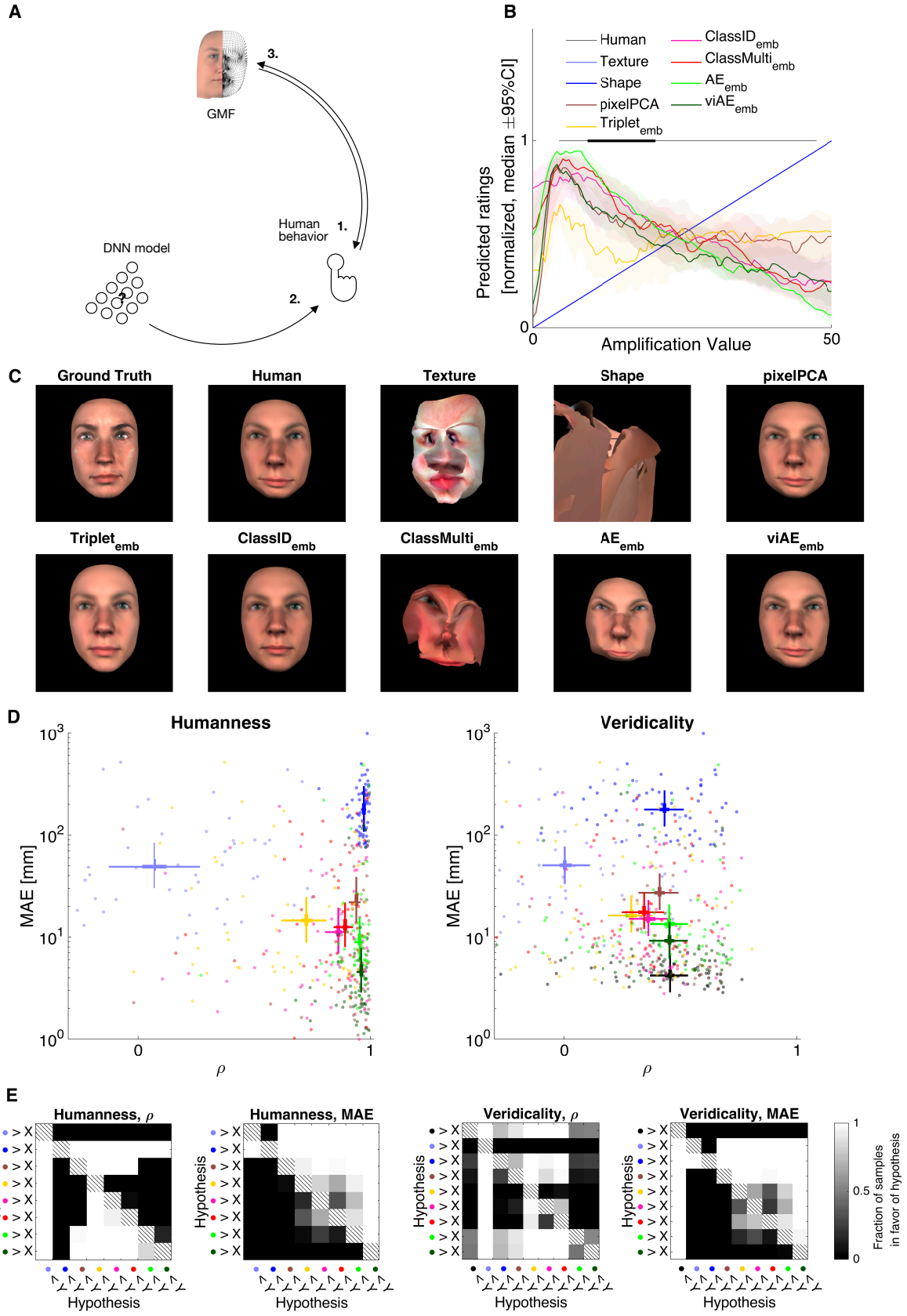

Figure 4.13: Caption on following page.

that maximized the behavioral responses (figure 4.13B). The results were faces whose functional features elicited a high similarity rating in the DNN models (figure 4.13C), analogous to faces that elicited high similarity ratings in each human participant, as in the original study (Zhan et al., 2019a). We then compared the functional face features

Figure 4.13 *(previous page)*: **Internal templates reconstructed from human behavior and its model predictions**
A Schema of analysis. We predicted human behavior from GMF features (1.) and DNN activations (2.). With mass-univariate regression, we predicted each individual GMF feature from human behavior and its DNN predictions (3.). **B** Amplification tuning curves. We presented the reverse correlated templates amplified at different levels to each model. Solid lines denote pooled median across participants and colleagues, shaded regions denote 95% (frequentist) confidence intervals. Black lines at the top denote 95% (bold) and 50% (light) highest density estimates of human amplification values. The linear GMF shape and texture forward models predicted monotonically increasing responses for higher amplification levels. Other models peaked at a given amplification level. See Figure S9 for amplification tuning responses of a broader range of models. **C** Comparison of rendered faces. Panels show ground truth face of one exemplary target familiar colleague captured with a face scanner (top left) and reconstructions of the face features from human behavior and its DNN predictions for one typical participant (i.e., closest to the pooled group medians shown in D). Figure 4.19 presents the three other familiar colleagues. **D** Evaluation of correspondence of humans and model templates ("humanness," left) and the relation of templates to ground truth faces ("veridicality," right). The x axis shows Pearson correlation of the 3D features projected onto a single inward-outward direction; the y axis shows the mean absolute error (MAE) of the 3D features. Each dot corresponds to a single participant in a specific target familiar colleague condition. Crosses denote 95% (bold) and 50% (light) HPDIs for each system from Bayesian linear modeling of the results. **E** Comparison of main effects of systems in Bayesian linear models of the results in (D). See also figures 4.14–4.19 and 4.27. Experiment on human participants was conducted by Jiayu Zhan.

of human participants and their DNN models (figure 4.13D, left). We also computed how veridical these human and DNN features were to the ground truth faces of familiar colleagues (figure 4.13D, right).

How human-like are DNN features?

The viAE had the most human-like features, with the lowest mean absolute error (MAE, figure 4.13D, left, y-axis; comparison with second best DNN model, AE $>$ viAE: $f_{h_1} = 0.9943$) and a correlation with human features similar to that of the AE (figure 4.13D, left, x-axis; viAE $>$ AE: $f_{h_1} = 0.8489$). All DNN models had a lower MAE than the simple pixelPCA model (all DNNs $<$ pixelPCA: $f_{h_1} > 0.9492$), but only the (vi)AE had a better correlation with human features (AE and viAE $>$ pixelPCA: both $f_{h_1} > 0.9729$).

How veridical are DNN and human features?

viAE features were closest to the veridicality of human features to the ground truth 3D faces, with the lowest MAE (figure 4.13D, right, y-axis; second best DNN model AE $>$ viAE: $f_{h_1} = 0.9558$; viAE $>$ human: $f_{h_1} = 0.9996$) and a correlation comparable with that of the AE. All DNN models had a lower MAE than the simple pixelPCA model (all DNNs

$<$ pixelPCA: all $f_{h_1} > 0.9732$), but only the (vi)AE had a better correlation with the ground truth face identity features (AE and viAE $>$ pixelPCA: both $f_{h_1} > 0.8842$).

In sum, this analysis compared the internal representations of the target faces in human participants and their DNN models, and all with the ground truth 3D shapes of the target identities. These comparisons, supported by intuitive visualizations, revealed that the viAE had internal feature representations that best matched the internal representations of humans.

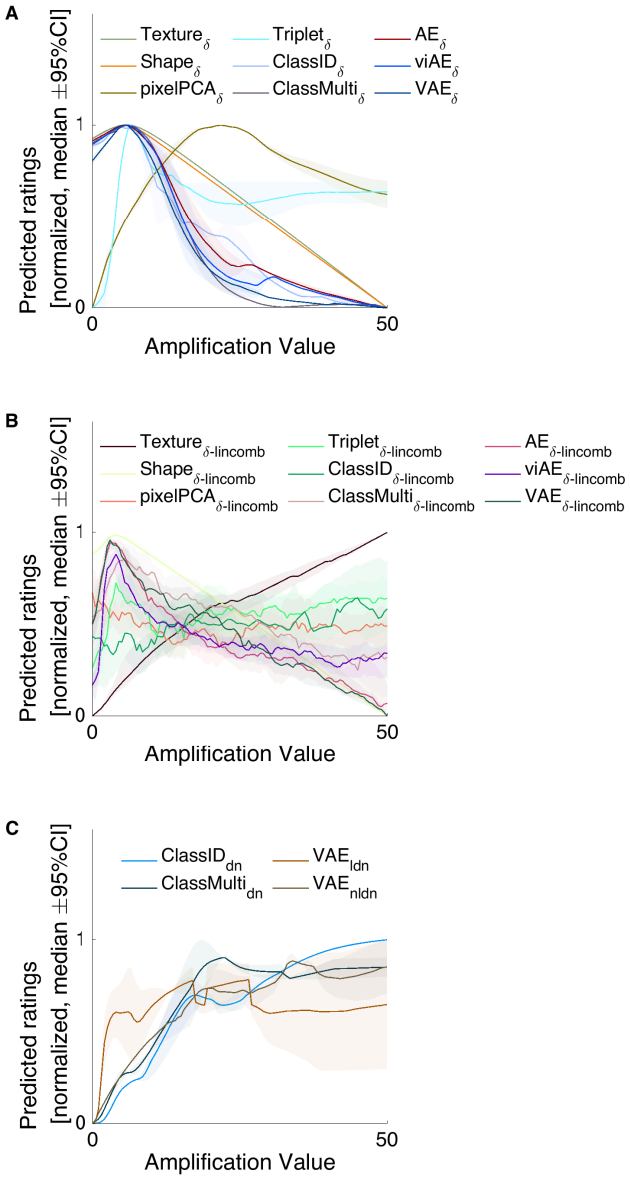

Figure 4.14: **Amplification tuning responses of additional encoding models (related to figure 4.13)**.
A Amplification tuning responses of euclidean distances ("δ") of templates amplified at different levels and ground truth representations of the target colleagues. Solid lines denote the pooled median across participants and target colleagues, shaded regions denote 95% (frequentist) confidence intervals bootstrapped using $10,000$ samples. **B** Same as in A, but showing amplification tuning responses of linearly weighted euclidean distances instead ("δ-lincomb"). **C** Same as in A, but showing amplification tuning responses of pre-softmax decision neuron activities ("logits") of respective target colleagues instead ("dn"). Experiment on human participants was conducted by Jiayu Zhan.

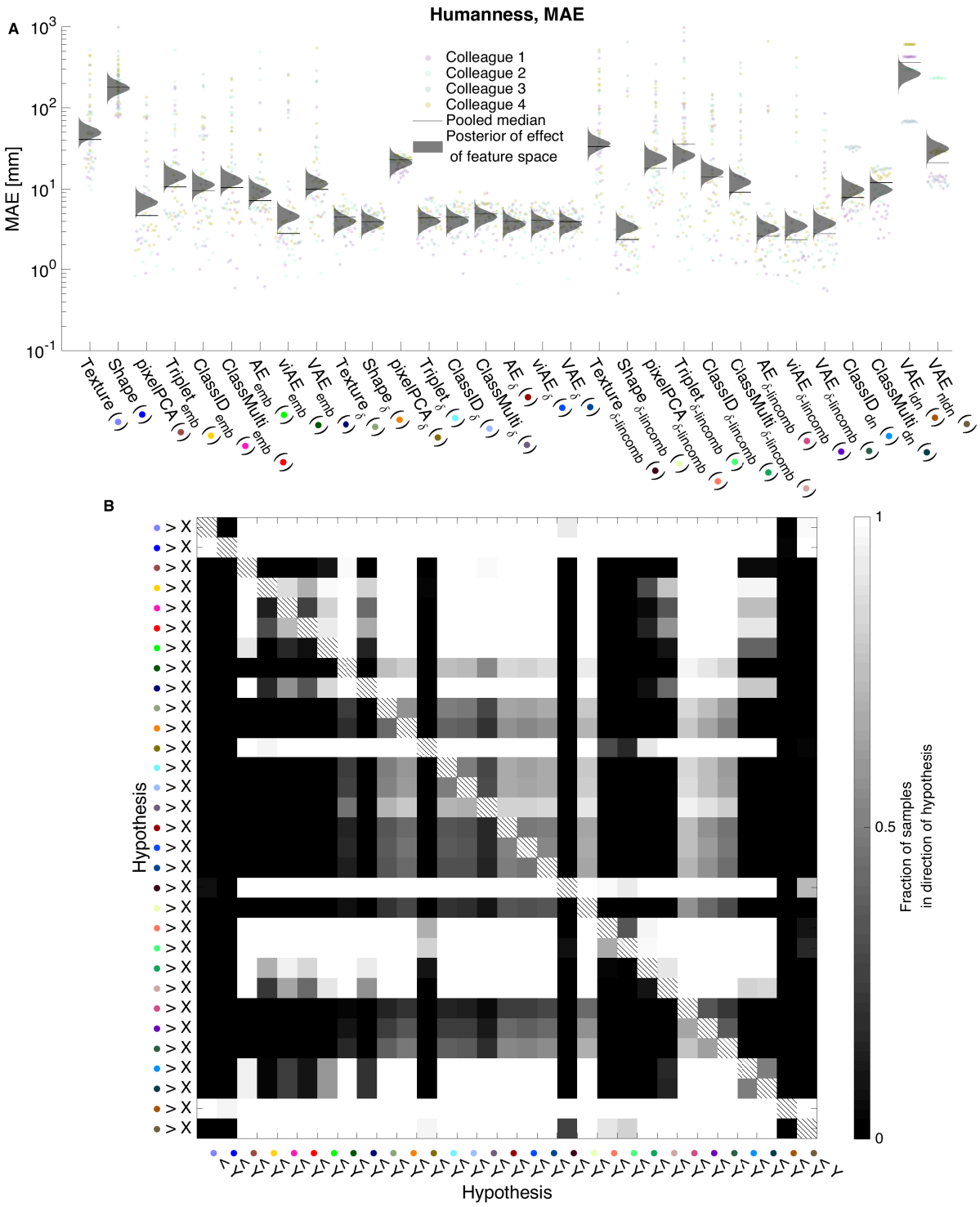

Figure 4.15: **Evaluation of the mean absolute error between reverse-correlated faces of humans and reverse-correlated faces of models for a larger set of encoding models (related to figure 4.13).**
A Mean absolute error (MAE, computed as the euclidean distances in 3D space averaged across vertices) of reverse correlated templates of the models and those of humans. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

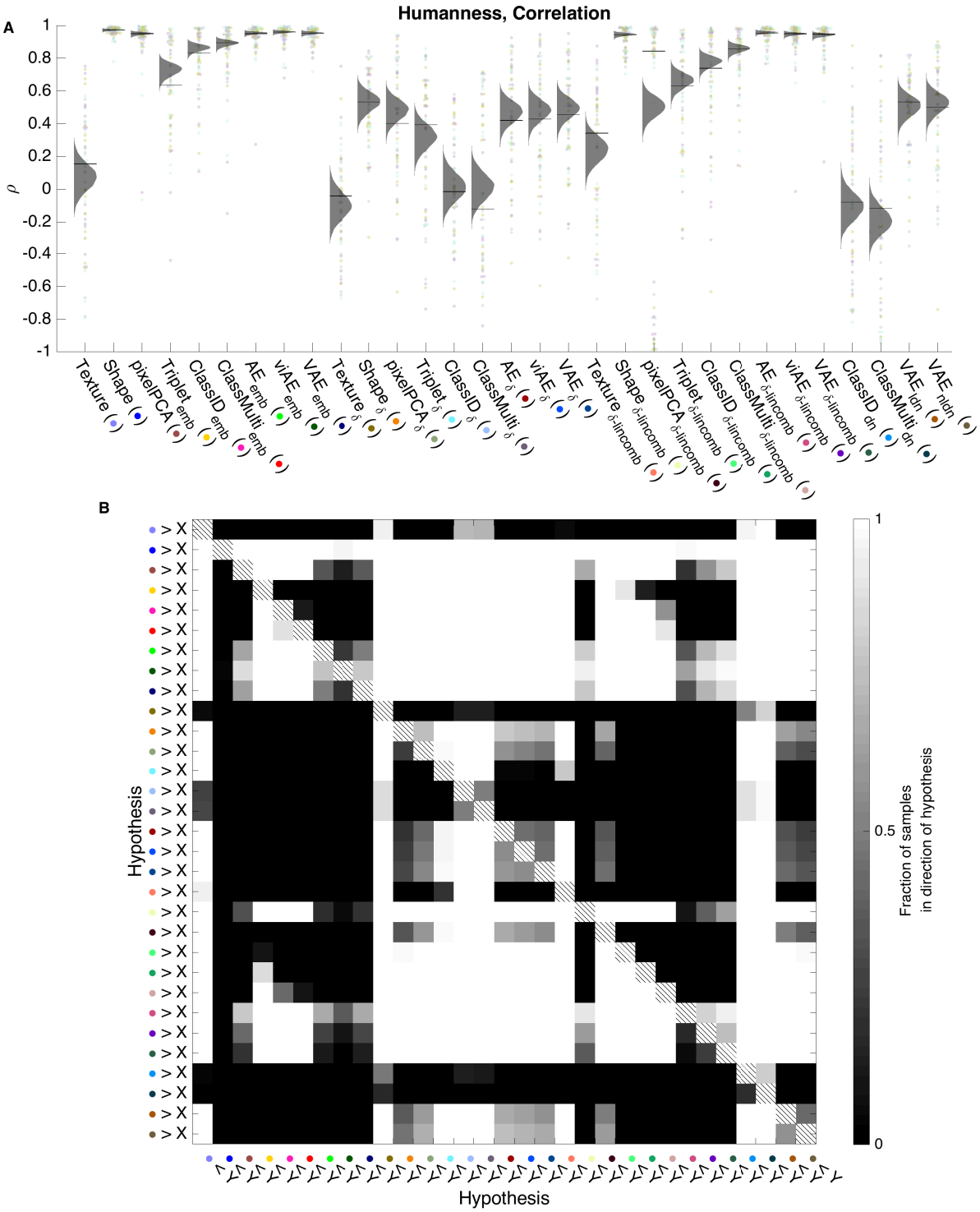

Figure 4.16: **Evaluation of the Pearson correlation between reverse-correlated faces of humans and reverse-correlated faces of models for a larger set of encoding models (related to figure 4.13).**
A Pearson correlation (computed with vectors of 3D vertices projected on a single inward-outward dimension) of reverse correlated templates of the models and those of humans. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

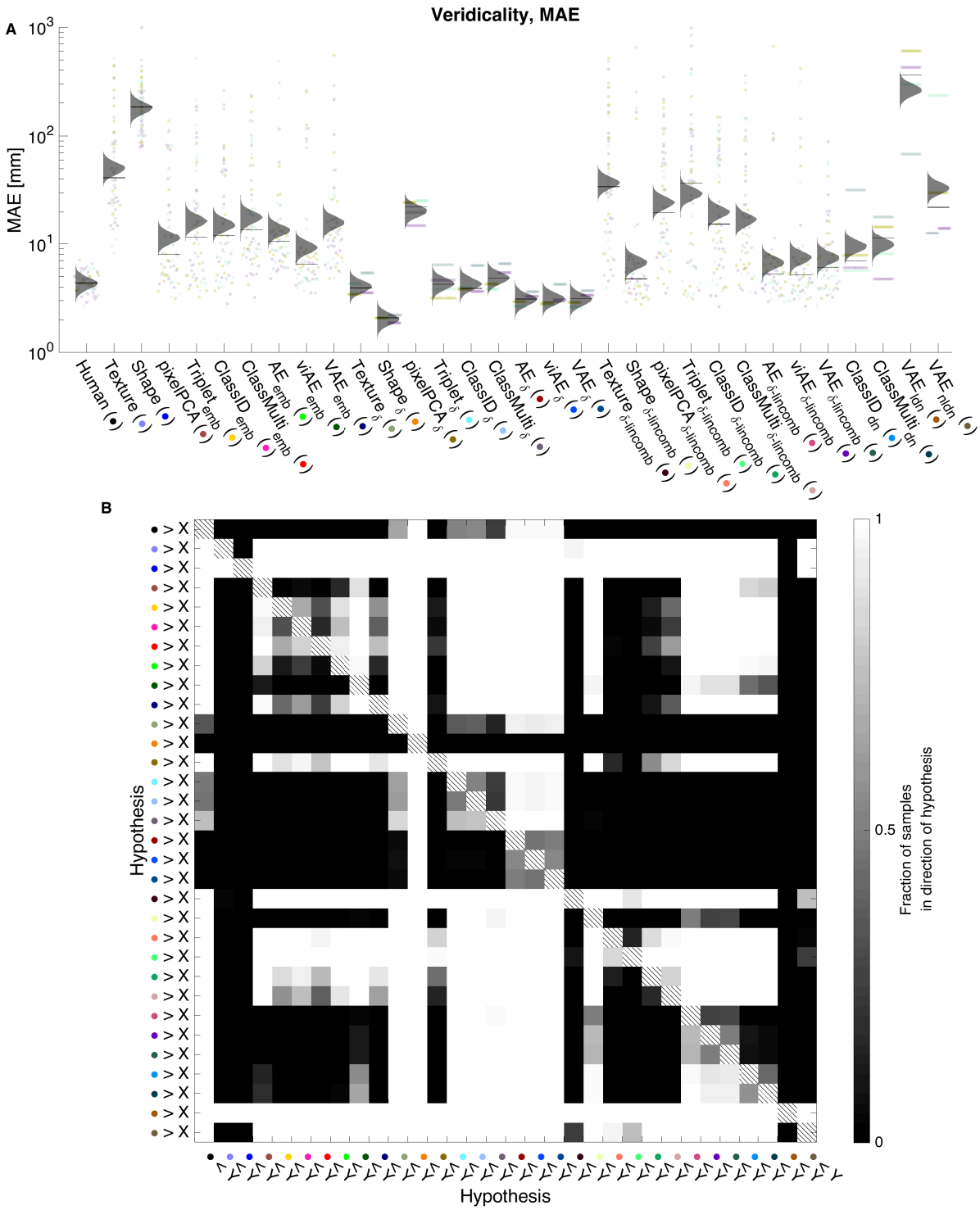

Figure 4.17: **Evaluation of the mean absolute Error between reverse-correlated faces of humans and models and the ground truth face shapes for a larger set of encoding models (related to figure 4.13).**
A Mean absolute error (MAE, computed as the euclidean distances in 3D space averaged across vertices) of reverse correlated templates of the models and ground truth 3D shape of the target colleagues as captured with a 3D camera array. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

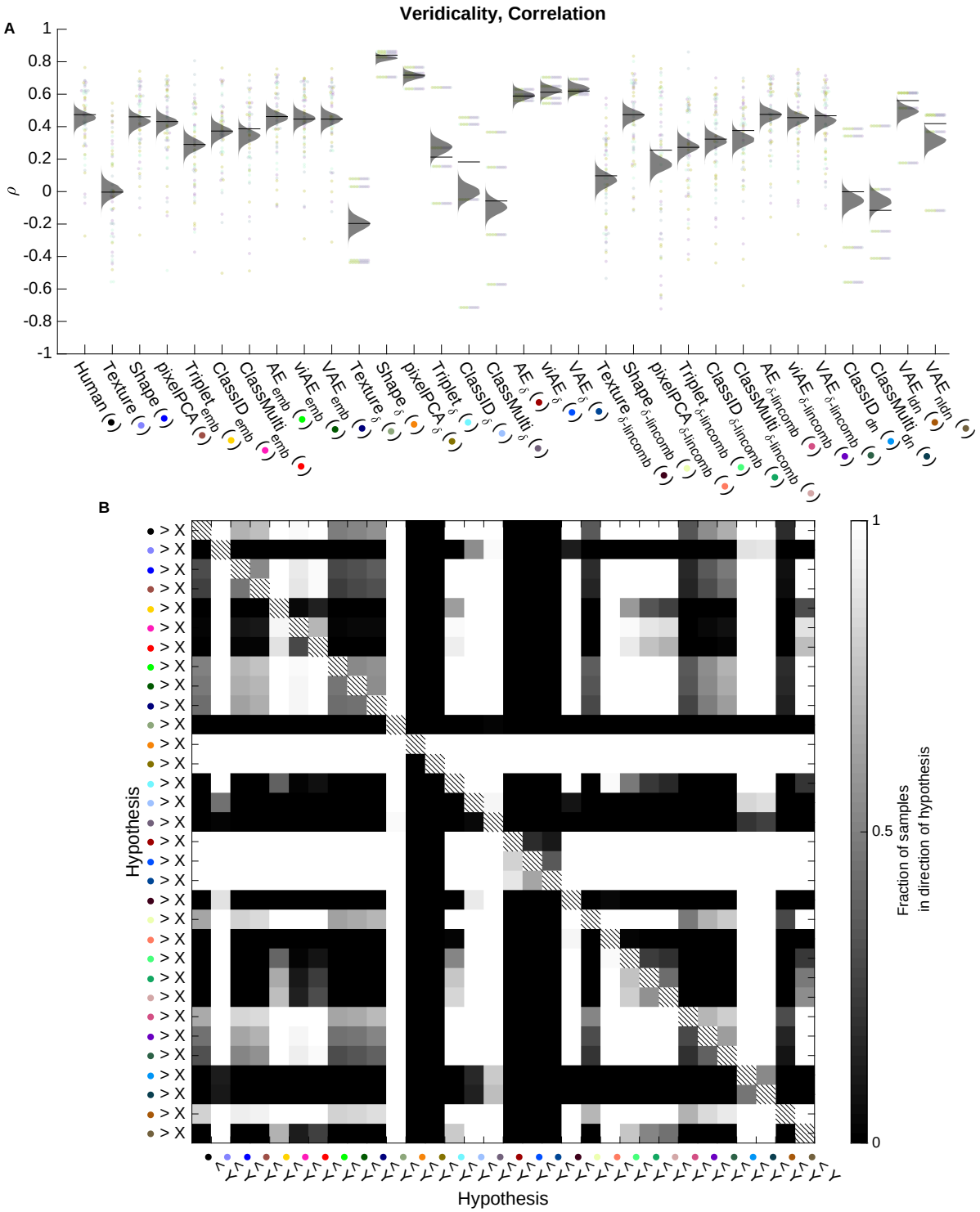

Figure 4.18: **Evaluation of the Pearson correlation between reverse-correlated faces of humans and models and the ground truth face shapes for a larger set of encoding models (related to figure 4.13).**
A Pearson correlation (computed with vectors of 3D vertices projected on a single inward-outward dimension) of reverse correlated templates of the models and those of humans. See figure 4.7 for explanation of the model shorthands. **B** Comparisons of the posterior distributions of the main effects for all systems from Bayesian linear modeling of the raw performances. For each pair in the matrices, the color gradient reflects the fraction of samples of the forward model color coded on the y-axis that is larger than the forward model color coded on the x-axis. See x-axis labels for color legend. Experiment on human participants was conducted by Jiayu Zhan.

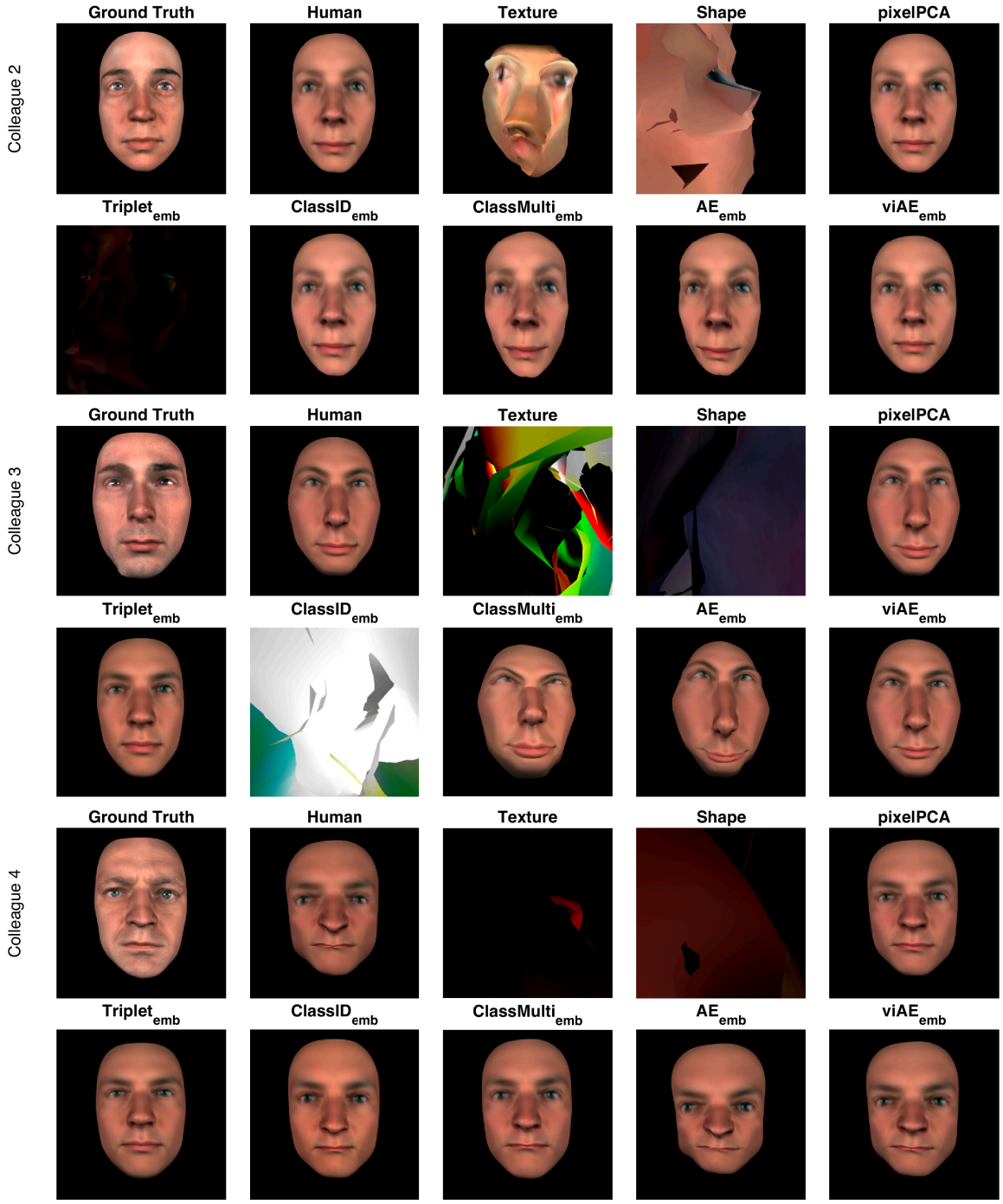

Figure 4.19: **Renderings of reverse-correlated templates of the three remaining colleagues of exemplary participant (related to figure 4.13).**
Comparison of rendered faces for one exemplary target colleague. Top left panel in each block of two rows shows ground truth face of one target colleague as captured with a 3D camera array. Following panels show reconstructions of the face features from human observed and predicted behavior for one typical participant (i.e. closest to the pooled group medians shown in figure 4.12D). Experiment on human participants was conducted by Jiayu Zhan.

4.3.4 Generalization testing

A crucial test of models of human behavior is their generalization to conditions that differ from the distribution of the training data. We performed such out-of-distribution testing

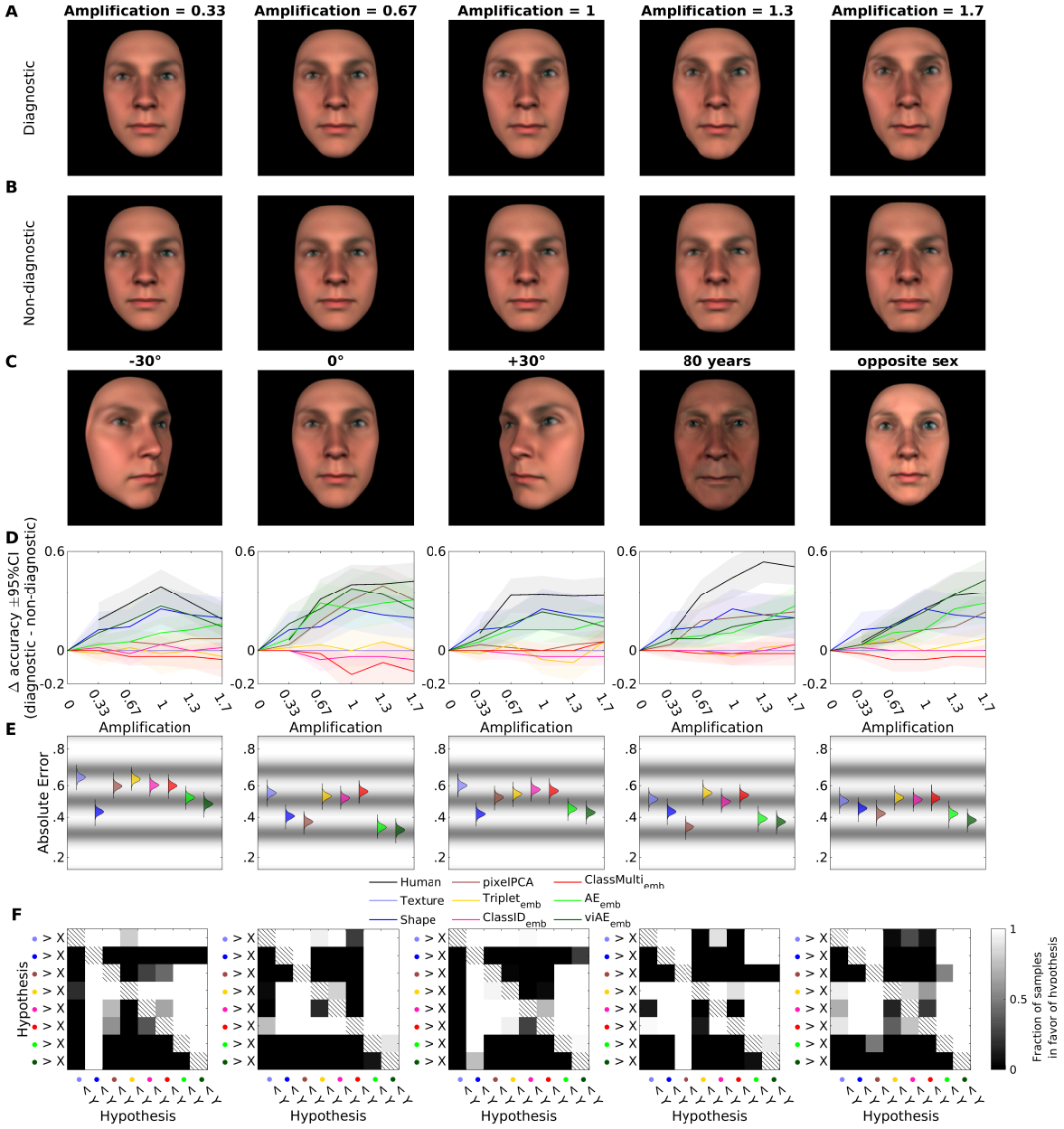

Figure 4.20: Caption on following page.

in five different tasks (Zhan et al., 2019a), using the GMF to change the viewing angle, the age (to 80 years), and the sex (to the opposite sex) of the target familiar face (figure 4.20C). Importantly, we did so while also selectively amplifying functional face features that were expected (figure 4.20A) or not expected (figure 4.20B) to cause the identification of each familiar face (based on reverse correlation, see experimental procedures 4.5.3). Using these new stimuli, we compared the generalization performance of a new group of $n = 12$ human validators and the DNN models. On each trial, validators responded by selecting the familiar identity that was most similar to the face stimulus (or used a fifth option when the stimulus was not similar to any familiar face). For each face stimulus, we predicted the human similarity ratings using the forward models fitted to each of the 14 participants and four familiar faces as described in stage 1 above, and chose the faces that yielded the highest predicted rating. We then compared the absolute error of the model choice accuracies with the human choice accuracies. The viAE best

Figure 4.20 *(previous page)*: **Generalization testing.**
A Example stimuli for the task-relevant condition in the 0° viewing angle condition of one familiar colleague. Using a group model, each face feature of each familiar identity was classified as being either task relevant or task irrelevant for human identification. Versions of each colleague were then created whereby the task-relevant (versus -irrelevant) features were amplified at different levels, while the remaining features were defined as those of the average face. **B** Example stimuli for the task-irrelevant condition in the 0° viewing angle condition of the same target familiar identity as in A. **C** Renderings of the task-relevant face amplified at a level of 1.3 for five different generalization conditions. **D** Difference of choice accuracy between the task-relevant and -irrelevant conditions. Positive values denote a higher accuracy when task-relevant features were amplified. **E** Posterior distributions of main effects of feature spaces when modeling absolute error (relative to human behavior) with Bayesian linear models. Gray bandings denote density estimates of thresholds separating the five possible different error values (human accuracies are averaged across five ratings of the same item). **F** Comparison of the posterior distributions in E. For each pair in the matrices, the color gradient reflects the fraction of samples of the feature space color coded on the y-axis $>$ the predictor space color coded on the x-axis. See also Figures 4.21–4.27.
Experiment on human participants was conducted by Jiayu Zhan.

matched human identification performance, which both increased when the functional features were amplified in the stimulus (figure 4.20D–F). The viAE had only a slightly smaller error compared with the AE for the frontal view (viAE $<$ AE: $f_{h_1} = 0.8958$), but a better view invariance with a clearly smaller error for the -30° (viAE $<$ AE: $f_{h_1} = 0.9995$) and $+30^\circ$ views (viAE $<$ AE: $f_{h_1} = 0.9696$). Only the GMF shape feature model came close to the (vi)AE (and was better than both AEs at -30°, both $f_{h_1} = 1$, and $+30^\circ$ both $f_{h_1} > 0.7656$). However, recall that the GMF is a non-image-computable "ground truth" 3D model whose input is not affected by 2D image projection. Critically, the simple pixelPCA model did not generalize well to viewpoint changes (viAE and AE $<$ pixelPCA: $f_{h_1} = 1$) except in the age generalization task, where it had a slightly lower error than the second best viAE (pixelPCA $<$ viAE: $f_{h_1} = 0.9940$). In the opposite sex task, the viAE again had the lowest error (viAE $<$ second best AE: $f_{h_1} = 1$). Whereas previous analyses suggested that a model as simple as the pixelPCA could explain human responses, more comprehensive tests of the typical generalization gradients of face identity demonstrated that such a conclusion was unwarranted. Thus, rigorous comparative tests of typical generalization gradients are required to properly assess human visual categorization in relation to their DNN models.

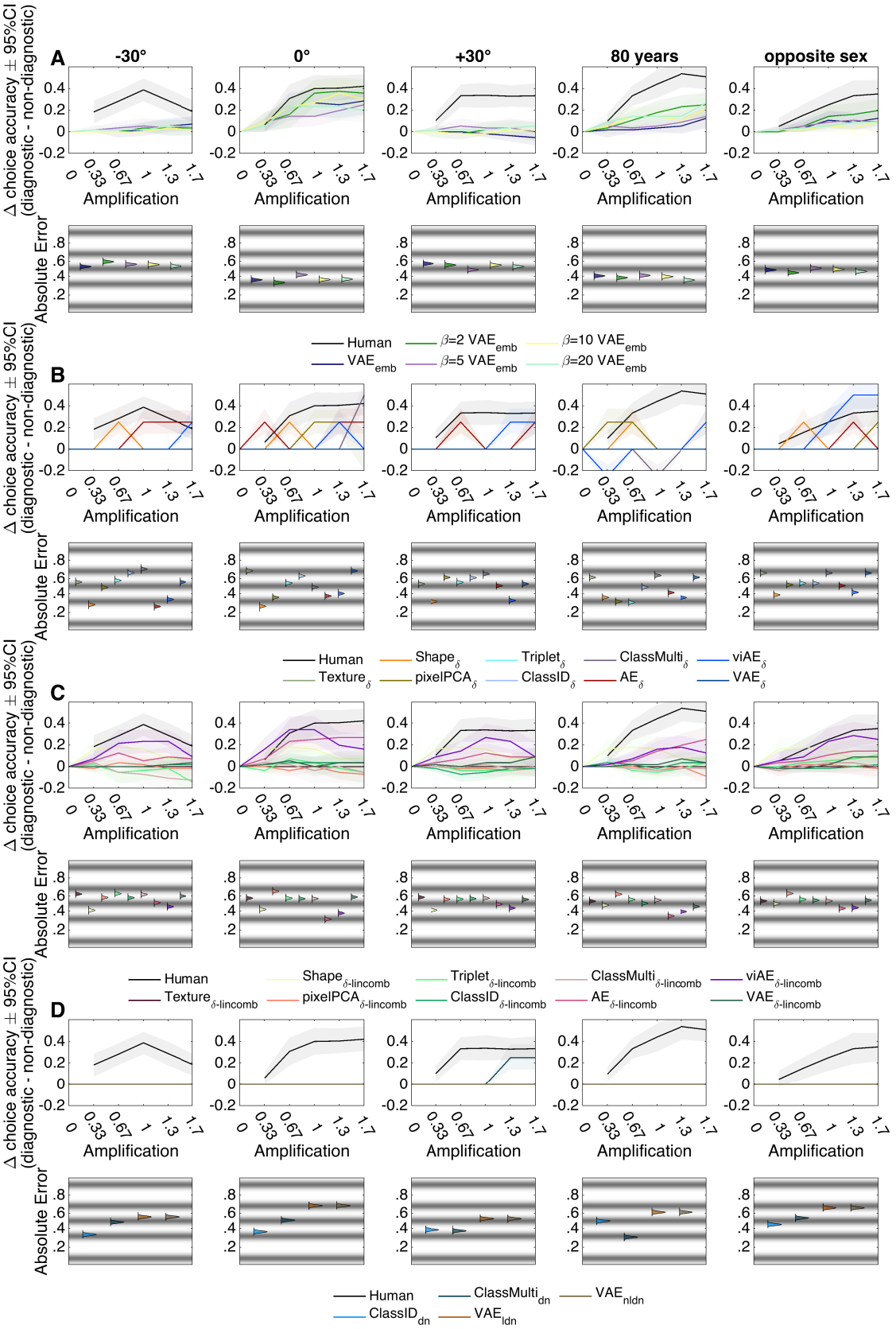

Figure 4.21: Caption on following page.

Figure 4.21 *(previous page)*: **Generalization testing of a larger set of encoding models (related to figure 4.20)**.
A Generalization testing for VAE models with various degrees of regularization. None yield a factorization of the latent spaces that disentangles viewing angle from other factors. Top row shows difference of choice accuracy between the diagnostic and non-diagnostic conditions. Positive values denote a higher accuracy when diagnostic features were amplified. Bottom row shows posterior distributions of main effects of feature spaces when modeling absolute error vs humans with Bayesian linear model. Grey bandings denote density estimates of thresholds separating the five different error values possible (human accuracies are averaged across five ratings of the same item). **B** – **D** show the same as in **A**, but for different forward models. See figure 4.7 for explanation of the model shorthands. Experiment on human participants was conducted by Jiayu Zhan.

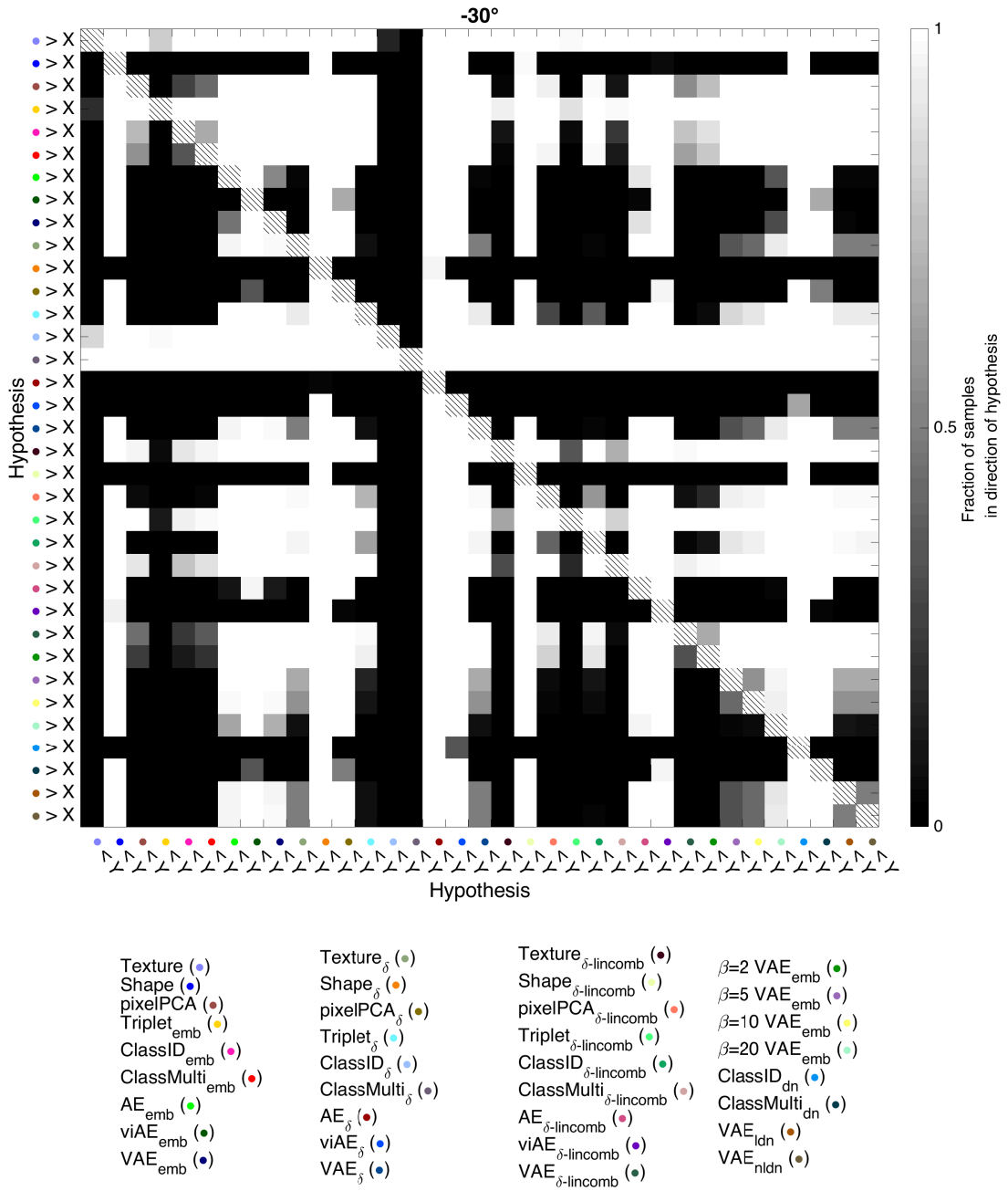

Figure 4.22: **Comparison of posterior distributions for larger set of forward models in -30° viewing angle generalization (related to figure 4.20).** Comparison of the posterior distributions of the leftmost column in figure 4.21. For each pair in the matrices, the color gradient reflects the fraction of samples of the feature space color coded on the y-axis that is larger than the predictor space color coded on the x-axis. See figure 4.7 for explanation of the model shorthands. Experiment on human participants was conducted by Jiayu Zhan.

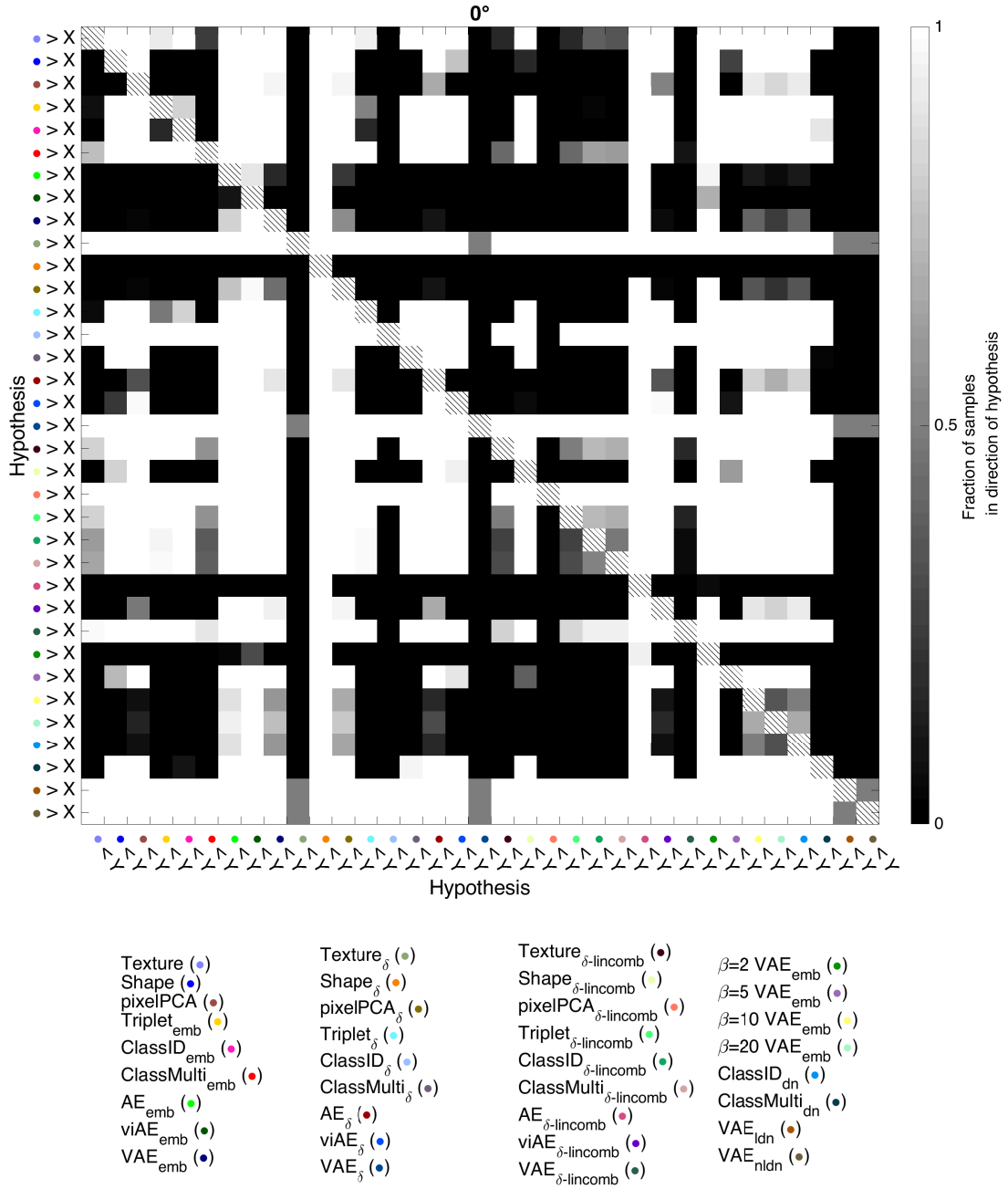

Figure 4.23: **Comparison of posterior distributions for larger set of forward models in 0° viewing angle generalization (related to figure 4.20).** Comparison of the posterior distributions of the second column in figure 4.21. For each pair in the matrices, the color gradient reflects the fraction of samples of the feature space color coded on the y-axis that is larger than the predictor space color coded on the x-axis. See figure 4.7 for explanation of the model shorthands. Experiment on human participants was conducted by Jiayu Zhan.

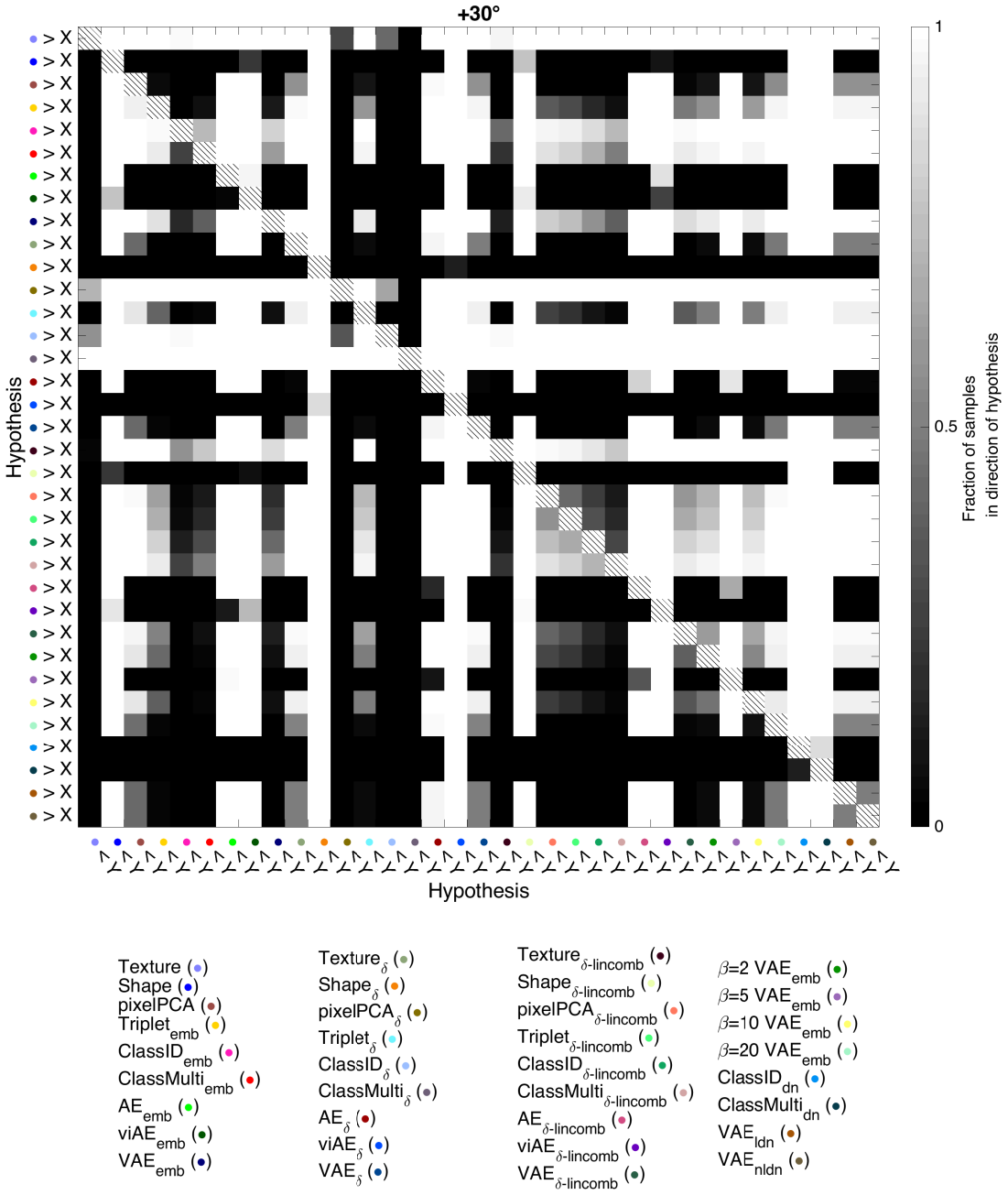

Figure 4.24: **Comparison of posterior distributions for larger set of forward models in +30° viewing angle generalization (related to figure 4.20).** Comparison of the posterior distributions of the middle column in figure 4.21. For each pair in the matrices, the color gradient reflects the fraction of samples of the feature space color coded on the y-axis that is larger than the predictor space color coded on the x-axis. See figure 4.7 for explanation of the model shorthands. Experiment on human participants was conducted by Jiayu Zhan.

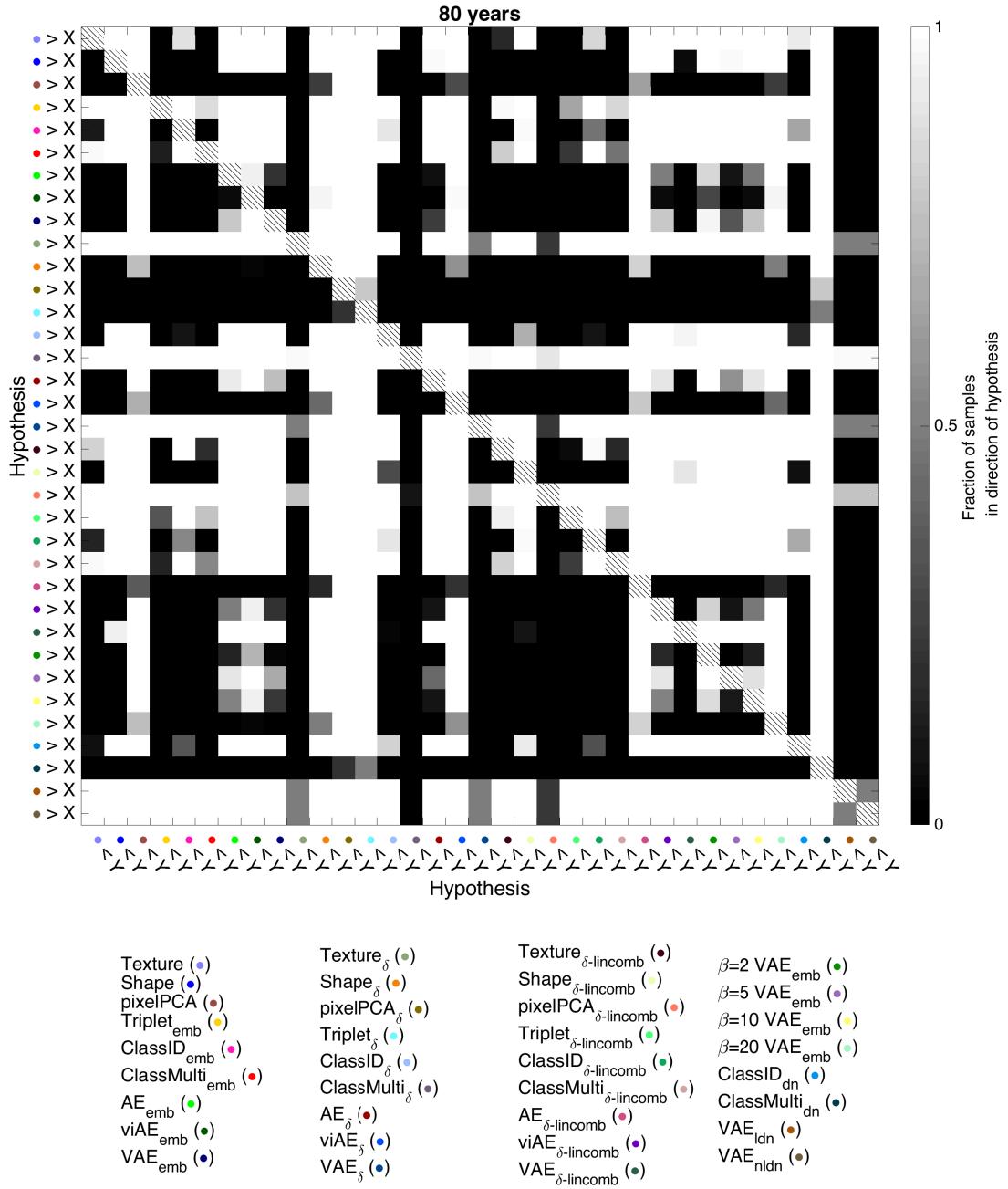

Figure 4.25: **Comparison of posterior distributions for larger set of forward models in 80 years generalization (related to figure 4.20).**
Comparison of the posterior distributions of the fourth column in figure 4.21. For each pair in the matrices, the color gradient reflects the fraction of samples of the feature space color coded on the y-axis that is larger than the predictor space color coded on the x-axis. See figure 4.7 for explanation of the model shorthands. Experiment on human participants was conducted by Jiayu Zhan.

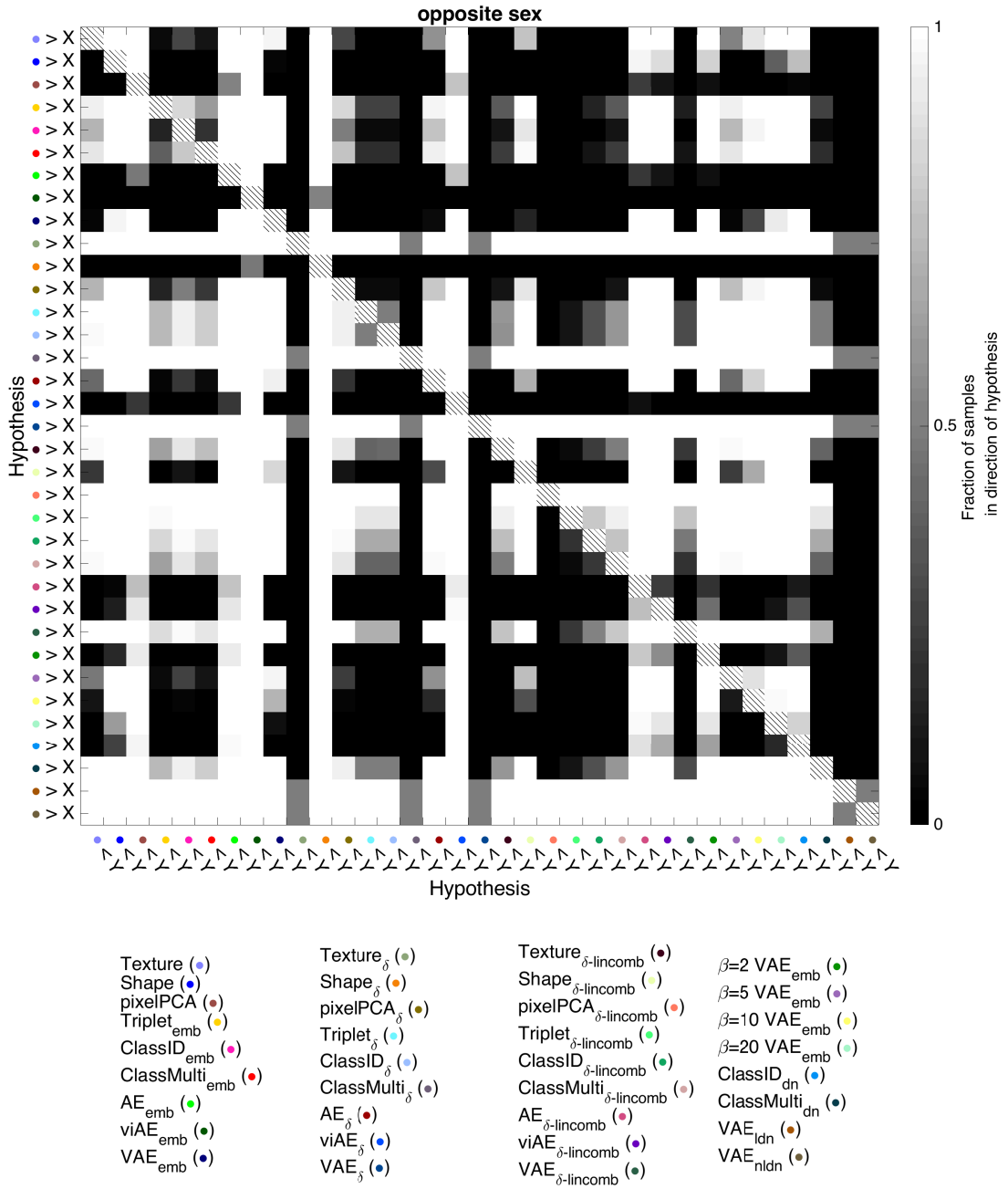

Figure 4.26: **Comparison of posterior distributions for larger set of forward models in opposite sex generalization (related to figure 4.20)**.
Comparison of the posterior distributions of the rightmost column in figure 4.21. For each pair in the matrices, the color gradient reflects the fraction of samples of the feature space color coded on the y-axis that is larger than the predictor space color coded on the x-axis. See figure 4.7 for explanation of the model shorthands. Experiment on human participants was conducted by Jiayu Zhan.

4.4 Discussion

In this study, we sought to address the long-standing problem of interpreting the information processing performed by DNN models so as to ground their predictions of human behavior in interpretable functional stimulus features. Key to achieving this was our use of a generative model to control stimulus information (3D face shape and RGB texture). We trained five DNN models with different objectives, following which we activated the

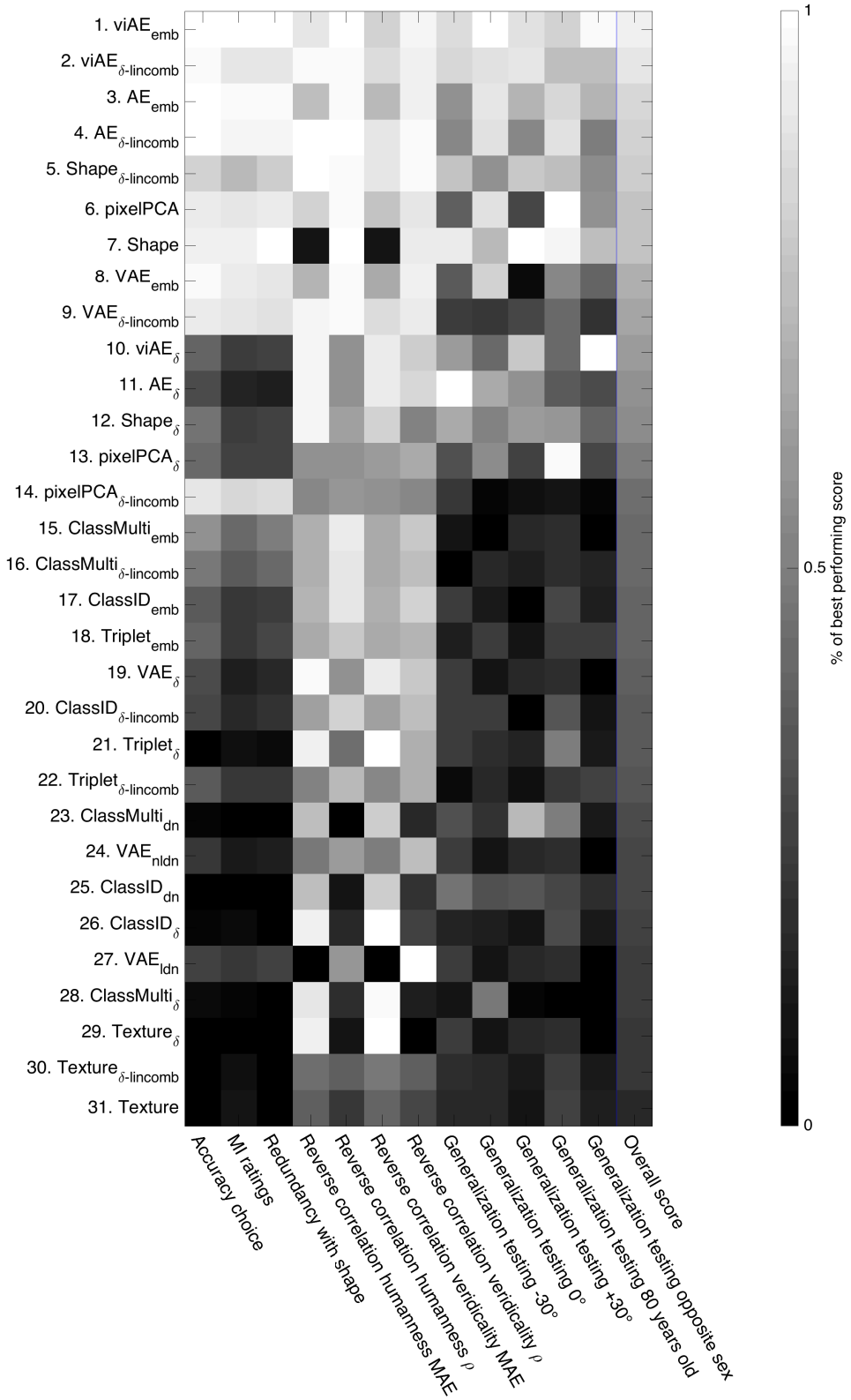

Figure 4.27: Caption on following page.

DNNs' embedding layers with the face stimuli of a human experiment (in which participants were asked, based on their memory, to assess the similarity of random faces to the faces of four familiar colleagues). We then used these activations to fit forward models that predicted human behavior. Of the tested models, (vi)AE embeddings best predicted human behavior, because these embeddings represented the human-relevant 3D shape of familiar faces with the highest fidelity. Next, we reconstructed the face fea-

Figure 4.27 *(previous page)*: **Ranking of extended set of models (related to Figures 4.5, 4.13 and 4.20)**.
We integrated the results of the models in all comparisons except for the re-prediction analysis reported in figure 4.12 (which is only applicable to linear combination forward models). Redundancy of the shape model with itself is not computable and was thus manually set to the best possible score. Scores in veridicality of reverse correlation were defined as the absolute difference to the veridicality achieved by humans. Scores in generalization testing (absolute error to human behavior) were additionally penalized for a low delta in accuracy of diagnostic and non-diagnostic stimuli. Performances of models (maxima a posteriori of Bayesian linear models) were normalized within comparisons to range from 0 (worst considered model) to 1 (best considered model). Scores were summed across comparisons and divided by the number of comparisons for the overall score. See figure 4.7 for explanation of the model shorthands.
Experiment on human participants was conducted by Jiayu Zhan.

tures represented in the embeddings that impact the behavioral predictions. The 3D reconstructions demonstrated that the viAE models and humans used the most similar functional features for behavior. Lastly, we found that the viAE best matched human generalization performance in a range of five different out-of-distribution changes of the stimuli (testing several viewing angles, older age, and opposite sex versions of the four colleagues).

Together, our approach (cf. figure 4.1) and analyses suggests a more stringent test of functional feature equivalence between human responses and their DNN models beyond the simple equivalence of responses to uncontrolled naturalistic stimuli. Such deeper functional features equivalence enables the mechanistic interpretations of the processing of these same features across the layers of the human brain and its DNN models. However, as shown in psychophysics, exhaustively testing the generalization gradients of human visual categorization is difficult because it requires not only modeling behavioral (or neuronal) responses but also the real-world (and artificial) dimensions of variations of the stimulus categories under consideration.

4.4.1 Why focus on functional equivalence?

A key finding that motivates usage of DNNs as models of the human brain is that their activations predict behavioral and neural responses to novel real-world stimuli better than any other model. However, it remains unclear whether these surface similarities between humans and DNNs imply a deeper similarity of the underlying information-processing mechanisms (Saxe et al., 2021). Real-world stimuli comprise multiple unknown correlations of undefined features. It is generally unknown which of these features DNNs use, leading to unpredictable out-of-distribution generalizations. Consequently, it is difficult to assess the featural competence of the model that predicts the behavioral or neural responses. Surprisingly simple feature alternatives (“feature fallacy" Diedrichsen, 2020; Daube et al., 2019b) could explain such surface similarities (Lapuschkin et al., 2019).

Relatedly, extensive testing of the generalization gradients of humans and DNNs is required to reveal algorithmic intricacies that would otherwise remain hidden, leading to failure with out-of-distribution exemplars.

Marr's framework offers a solution to these problems (Marr, 2010): we should constrain the similarity of complex information-processing mechanisms at the abstract computational level of their functional goals of seeking specific information to resolve a task. Our methodology sought to assess whether the human participants and their DNN models processed similar functional face features in a face identity task where features are defined within a generative model of the stimulus. Once functional equivalence is established, we can turn to the algorithmic-implementation levels of Marr's analysis. That is, we can seek to understand where, when, and how detailed mechanisms of the occipitoventral hierarchy, and suitably constrained DNN architectures (e.g., with two communicating hemispheres, properties of contralateral followed by bilateral representations, and so forth) process the same functional features of face identity, using a model of the stimulus (Schyns et al., 2009). Such algorithmic-implementation-level explorations could then consider estimates of the algorithmic complexity of the task (Chaitin, 1975) to regularize explanations of model predictions to be as simple as possible (Geirhos et al., 2020; Morgan, 2018; Buckner, 2019; Kubilius, 2018). We see the deeper functional equivalence of the information processed as a necessary prerequisite to surface comparisons of network activations or behaviors in a task.

4.4.2 Hypothesis-driven research using generative models

The idea of using generative models in psychophysics and vision research is not new (Olman & Kersten, 2004; Greene et al., 2014; Lescroart & Gallant, 2019; Jack & Schyns, 2017). It arose from the recognition by synthesis framework (Grenander, 1994; Yuille & Kersten, 2006), itself an offspring of Chomsky's generative grammars (Chomsky, 1965). Explicit experimental hypotheses are directly tied to the parameterization of stimuli by generative models and vice versa. For example, we explicitly tested that a parameterization of faces in terms of their 3D shape and RGB texture could mediate human and DNN behavior in the task (Zhan et al., 2019a; Yildirim et al., 2020). Our study thereby contributes to the debate about the degree to which convolutional DNNs can make use of shape information in images (Xu et al., 2018; Kubilius et al., 2016; Baker et al., 2018; Geirhos et al., 2019; Brendel & Bethge, 2019; Hermann & Kornblith, 2019; Doerig et al., 2020). In this context, the exact structure of the information represented in the human brain remains an empirical question. The veridical representation implied by computer graphics models (Yildirim et al., 2020; Chang et al., 2021; Jozwik et al., 2021) is one hypothesis. Other specific ideas about face, object, and scene representations must and will be tested with different designs of generative models, including DNNs (VanRullen & Reddy, 2019; Bashivan et al., 2019; Ponce et al., 2019). The ideal generative model for the encoding function of visual categorization would "simply" be the inverse of the

function implemented by the biological networks of the brain. Such an inverse would provide the control to experiment with each stage of the brain's algorithm of the stimulus processing for visual categorizations. In the absence of such an ideal, we must develop alternative generative models to test alternative hypotheses of the brain's encoding function for categorization. Modern systems such as generative adversarial networks (Karras et al., 2020) and derivatives of the classical variational autoencoders (VAEs) such as vector-quantized VAEs (van den Oord et al., 2018b; Razavi et al., 2019) and nouveau VAEs (Vahdat & Kautz, 2020) which can be trained on large, naturalistic face databases, can synthesize tantalizingly realistic faces, complete with hair, opening up an interesting avenue for future research and applications (Suchow et al., 2018; Bontrager et al., 2018; Todorov et al., 2020; Goetschalckx et al., 2021; Peterson et al., 2021a). However, understanding and disentangling their latent spaces remains challenging (Mathieu et al., 2019; Schölkopf et al., 2021).

4.4.3 viAE wins

Among the tested DNNs and across the multiple tests, the viAE provided the best face-shape representations to predict human behavior. With the notable exception of the generalization testing, the simple nonlinear pixelPCA model came close to this performance. This speaks to a model of human familiar face perception whereby the goal of feedforward processing is a view-invariant but holistic representation of the visual input. Interestingly, the Triplet, ClassID, and ClassMulti built up to this performance level (cf. Figures 3, 4, and 5). This suggests that the latent space learned to reconstruct an entire image of the input ((vi)AE) is approximated by the latent space learned when performing multiple stimulus categorizations (recall that ClassMulti learned all the categorical factors of the GMF), whereas simpler cost functions (Triplet and ClassID) yielded less informative latent spaces. Their discriminative goals can be solved with shortcuts (Geirhos et al., 2020) relying on a few isolated features, which are not sufficient to generalize as humans do (Hoel, 2021). This aligns with previous findings that multi-task learning (Scholte et al., 2018; Standley et al., 2020; Mao et al., 2020) and generative models (Schott et al., 2018) enhance robustness against adversarial attacks and best predict behavior under severe testing (Golan et al., 2020). In relation to faces as a broad category, future research could systematically study the number and combinatorics of categorizations (e.g., identity, sex, age, ethnicity, facial expressions) and rendering factors (e.g., illumination, pose, occlusions) that would be required to enhance the latent spaces to match (or surpass) the predictiveness of behavior of the latent space of the viAE, also across varying levels of familiarity (Blauch et al., 2020). Note that our specific viAE model remained imperfect in its prediction of human behavior and functional similarity of features. Its architecture did not incorporate many well-known characteristics of the human visual hierarchy, including temporal, recurrent9 processing (e.g., with multiple fixations (Fabius et al., 2019) due to foveated and para-foveated image resolution (Friston et al., 2012)), contralateral,

hemispherically split representations of the input, transfer of visual features across hemispheres (Ince et al., 2016), and integration in the ventral pathway (Zhan et al., 2019b), among others. An algorithmic-implementation-level explanation of the functional features learned by the viAE should be part of future research.

4.4.4 Constraints on the comparison of models with human behavior

Our modeling explicitly fitted regressions of multivariate features on unidimensional behavior (Naselaris et al., 2011). Our attempts to directly (parameter-free) extract one-dimensional predictions of human behavior from DNNs failed (4.7). Whereas models might exist to solve this problem more efficiently (Golan et al., 2020; Schott et al., 2018), an obstacle remains in that the human task is subjective: we do not expect the behavior of a given participant to perfectly predict that of another (see figures 4.6 and 4.7, (although representations tend to converge across participants Zhan et al., 2019a; Smith et al., 2012)). Participants can have their own internal representations of each target colleague (Schyns et al., 1998; Smith et al., 2012), which is impossible to predict without considering data from individual participants. From that perspective, learning an abstracted feature representation that still allows prediction of individual behavior is an attractive compromise. We implemented such a weighting, either directly as a linear combination of GMF features and DNN activations, or as a linear combination of feature- or activation-wise distances of stimuli to model representations of the target identities. For the image-computable models, these approaches did not lead to strong differences. Arbitrating between such computational accounts of human categorization behavior thus remains a question for future research (Smith & Sloman, 1994; Griffiths et al., 2017; Chang & Tsao, 2017). The interpretability of DNNs is now an important research topic. Sophisticated methods exist to visualize the stimulus features that cause the activation of a network node, such as deconvolution (Zeiler & Fergus, 2013), class activation maps (Zhou et al., 2015), activation maximization (Erhan et al., 2009; Simonyan et al., 2014; Olah et al., 2017, 2018, 2020), locally linear receptive fields (Keshishian et al., 2020), or layer-wise relevance propagation (Lapuschkin et al., 2019; Bach et al., 2015; Montavon et al., 2018). These methods usually rely on the noise-free accessibility of the activations, which is not possible with humans, making these methods unsuitable to compare humans with their DNN models. This is a significant hindrance to developing a human-like artificial intelligence, which requires resolving the challenge of designing experiments and analyses that enable inferences about the hidden representations of both humans and models (Funke et al., 2021; Thoret et al., 2021).

4.4.5 Conclusion

We have developed an example of how we can extend mechanistic interpretations of DNN predictions of human responses, in which we progress beyond surface predictions to a functional equivalence of the features that affect behavior. We did so by controlling complex stimulus features via an interpretable generative model. The limits of what we can predict about human behavior may be defined by the limits of current computer vision models. However, within these limits, the proportion that we can meaningfully understand is defined by the ever increasing capacities of interpretable generative models of stimulus material (Gan et al., 2020). Databases of natural images will only take us so far. Hence, we believe that future research attention should be distributed on the gamut between discriminative models to do the tasks, and generative models of the stimulus to understand what these models do.

4.5 Methods

4.5.1 Generative model of 3D faces

The generative model of 3D faces decomposes the shape and texture components of a database of 357 faces, captured with a 3D face-capture system (Dimensional Imaging Ltd., 2021) to enable their controlled recombination (Zhan et al., 2019a). For this study, two variations of the database were created: one excluding the faces of two female target colleagues and another excluding the faces of two male target colleagues. Each of the two database subsets then consists of a $[355 \cdot (4,735 \cdot 3)]$ $(N \cdot (vertices \cdot XYZ))$ shape matrix S and 5 $[355 \cdot (800/2^i \cdot 600/2^i \cdot 3))]$ $(N \cdot (X/2^{band} \cdot Y/2^{band} \cdot RGB))$ texture matrices T_i for bands $i = 0, ..., 4$ of a Gaussian pyramid model.

For each of the two database subsets, the modeling is achieved in two steps. In the first step, two separate general linear models are used to estimate the linear parameters of a constant term as well as sex, age, ethnicity (coded using two dummy variables), and their two- and three-way interactions. This is done with a $[355 \cdot 12]$ design matrix X describing the predictor values, a $[12 \cdot (4,735 \cdot 3)]$ matrix A_S describing the shape coefficients, and $[12 \cdot (800/2^i \cdot 600/2^i \cdot 3)]$ matrices A_{T_i} describing the texture coefficients:

$$S = XA_S + E_S \tag{4.1}$$

$$T_i = XA_{T_i} + E_{T_i} \tag{4.2}$$

Here, E_S $[355 \cdot (4,735 \cdot 3)]$ and E_{T_i} $[355 \cdot (800 \cdot 600 \cdot 3)]$ are the model residuals for shape and texture, respectively. A_S and A_{T_i} are estimated using least-squares linear regression.

In the second step, the residual components E_S and E_{T_i} are then isolated by removing the linear effects of ethnicity, sex, and age as well as their interactions from S and

T_i. Next, singular value decomposition (SVD, using MATLAB's economy-sized decomposition) is performed to orthogonally decompose the shape and texture residuals:

$$U_S S_S V_S^T = E_S \tag{4.3}$$

$$U_{T_i} S_{T_i} V_{T_i} = E_{T_i} \tag{4.4}$$

The matrices U_S $[(4,735 \cdot 3) \cdot 355]$ and U_{T_i} $[(800/2^i \cdot 600/2^i \cdot 3) \cdot 355$ for each i of $i = 0,...,4$ spatial frequency bands$]$ can thus be used to project randomly sampled shape or texture identity vectors into vertex or pixel space, respectively.

Any single face can then be considered as a linear combination of two parts: a basic "prototype face" defined by its factors of sex, age, and ethnicity and a specific individual variation on that prototype defined by its unique component weights. Once we know these two parts of the individual face, e.g., by random sampling, we are free to change one or the other, producing for example the same individual at a variety of different ages. This can then be rendered to an observable image with a desired viewing and lighting angle.

4.5.2 Participants

Ratings of random faces

To obtain behavioral data from humans (Zhan et al., 2019a), we recruited seven male and seven female white Caucasian participants aged 25.86 ± 2.26 years (mean $\pm$ SD).

Generalization testing

For a second validation experiment, 12 separate participants (7 white Caucasian female and 1 East Asian females, 5 white Caucasian males aged 28.25 ± 4.11 years $[$mean $\pm$ SD$]$) were recruited (Zhan et al., 2019a). In both experiments, all participants had been working at the Institute of Neuroscience and Psychology at the University of Glasgow for at least 6 months and were thus familiar with the target faces. All participants had normal or corrected-to-normal vision, without a self-reported history or symptoms of synesthesia, and/or any psychological, psychiatric, or neurological condition that affects face processing (e.g., depression, autism spectrum disorder, or prosopagnosia). They gave written informed consent and received UK£6 per hour for their participation. The University of Glasgow College of Science and Engineering Ethics Committee provided ethical approval for both experiments.

4.5.3 Experiments

Ratings of random faces

Four sets of $10,800$ random faces were generated, one for each of the four target colleagues. Two sets of random faces were created using the GMF that was built with the database that excluded the two female target colleagues. The other two sets of random faces were created using the GMF built with the database that excluded the two male target colleagues. The demographic variables were fixed (sex, age, and ethnicity) to those of the target colleagues. The resulting faces were rendered at frontal viewing and lighting angles. For each participant and target colleague, the generated faces were randomly gathered into $1,800$ groups of $2 \cdot 3$ arrays, which were superimposed on a black background. In a given trial, these face arrays were shown on a computer screen in a dimly lit room while the participant's head was placed on a chin rest at a 76 cm viewing distance from the image, such that each face subtended an average of $9.5° \cdot 6.4°$ of visual angle. Participants were instructed to choose the face of the array that most resembled that of the target colleague by pressing the corresponding button on a keyboard. The screen then changed to display the instruction to rank the chosen face with respect to its similarity to the target colleague on a 6-point rating scale, ranging from 1 ("not similar") to 6 ("highly similar"). These trials were split into four sets of 90 blocks of 20 trials each, resulting in a total of $7,200$ trials that all participants completed over several days (Zhan et al., 2019a).

Generalization testing

For each target colleague, 50 new 3D face stimuli were generated. These comprised the combinations of two levels of diagnosticity at five levels of amplification, which were each rendered in five different generalization conditions. Each of these factors will be explained in the following. In the original analysis (Zhan et al., 2019a), the mass-univariate reconstructions from observed human behavior (see "reverse correlation" below) had been referenced to reconstructions from $1,000$ permuted versions of the responses (using the same amplification values). For each vertex, the Euclidean distance of the chance reconstruction to the categorical average had been signed according to whether it was located inside or outside of the categorical average and averaged across permutations ("chance distance"). This was repeated using the ground truth target colleague shape ("ground truth distance") as well as the human-reconstructed shape ("human-reconstructed distance"). If the absolute difference of the chance distance and the ground truth distance was larger than the absolute difference of the human-reconstructed distance and the ground truth distance, the vertex was classified as "faithful." This had resulted in a $4,735 \cdot (14 * 4)$ binary matrix which had then been decomposed into matrices W $[4,735 \cdot 8]$ and H $[8 \cdot 56]$ (each column corresponding to a combination of a participant and a target colleague) using non-negative matrix factorization. Any of the eight component columns in W had been classified as contributing to a group representation of the

target colleagues if the median of the loadings H across participants surpassed a threshold value of 0.1. The "diagnostic component" C_D of each target colleague had then been defined as the maximum value on that vertex across components considered to load on the respective target colleague representation. After construction, C_D had then been normalized by its maximum value. Its "non-diagnostic" complement C_N was then defined as $C_N = 1 - C_D$. Taken together, the vectors C_D and C_N could now be interpreted as reflecting to what degree each vertex contributed to the faithful representation of each target colleague across the group of participants. These diagnostic and non-diagnostic components could then be used to construct 3D faces containing varying levels of either diagnostic (F_D) or non- diagnostic (F_N) shape information:

$$F_D = G \cdot C_D \cdot \alpha + XA_S(1 - C_D \cdot \alpha) \tag{4.5}$$

$$F_N = G \cdot C_N \cdot \alpha + XA_S(1 - C_N \cdot \alpha) \tag{4.6}$$

Here, G reflects the ground truth representation of the respective colleagues recorded with the 3D camera array, α reflects an amplification value that was set to one of five levels (0.33, 0.67, 1, 1.33, 1.67), and X describes the sex, ethnicity, age, and interaction values that describe the respective colleague such that XA_S represents the categorical average (see "generative model of 3D faces"). Each of these ten faces per target colleague were rendered at the viewing angles -30°, 0°, and $+30^\circ$ as well as with their age factor set to 80 years and a swapped sex factor. The 12 validation participants completed three sessions (3 viewpoints, age, and sex) in a random order, with one session per day. On a given trial, the validators saw a central fixation cross for 1 s, followed by a face stimulus on a black background for 500 ms. They were then asked to classify the seen face as showing one of the four target colleagues (or their siblings or parents in the age and sex conditions) or "other" if they could not identify the face as accurately and quickly as possible. Between each trial, a fixation cross was shown for a duration of 2 s. Each stimulus was shown five times in a randomized order. In the viewpoint session, validators completed 15 blocks of 41 trials; in the age and sex sessions, validators completed 5 blocks of 44 trials. This yielded accuracies of either 0, 0.2, 0.4, 0.6, 0.8, or 1 for each of the 10 stimuli per target colleague.

4.5.4 Networks

Training and testing of the networks was performed in Python 3.6.8 using keras 2.2.4 (Chollet & others, 2015) with a tensorflow 1.14.0 backend (Abadi et al., 2016). All networks shared the same training and testing sets and were constructed using the same encoder module. All models were trained using three data augmentation methods (random shifts in width and range by 5% as well as random zooms with a range of 10%).

Training and testing sets

The networks were trained on observable images generated by the GMF. We created 500 random identity residuals and combined them with the four combinations of two sexes (male and female) and two types of ethnicity (Western Caucasian and East Asian). To these, we added the four target colleagues, resulting in a total of $2,004$ identities. We rendered these at three different ages (25, 45, and 65 years), seven different kinds of emotion (happy, surprise, fear, disgust, anger, sadness, neutral), and three different horizontal and vertical viewing and lighting angles ($-30°, 0°, +30°$), resulting in $3,408,804$ images at a resolution of $224 \cdot 224$ RGB pixels. The four colleagues were rendered with two versions of the GMF built on face database subsets that excluded the two target colleagues of the same sex. Fifty percent of the $2,000$ random identities were rendered with one of these two GMFs. This dataset had first been generated for experiments not including the data from the human experiment. The version of the GMF that had been used to generate the stimuli for the human experiment had slight differences (rescaling of the data from the face database and different range of random coefficients). To allow for effortless generalization to the slightly different statistics of the stimuli that had been generated for the human experiment, we rendered all $3,408,804$ images twice, once with each of the two versions, effectively resulting in a further data augmentation. For the purpose of training, development, and testing, the dataset of $6,817,608$ images was split into a training set containing 80% of the images, and into a development and test set each containing 10% of the images.

Encoder module

We used a ResNet architecture to encode the pixel space images into a low-dimensional feature space (He et al., 2015). The $224 \cdot 224$ RGB images were first padded with three rows and columns of zeros, then convolved with 64 $7 \cdot 7$ filters with a stride of 2, batch normalized, subjected to a rectifying linear unit (ReLU) nonlinearity, max-pooled in groups of $3 \cdot 3$, and propagated through four blocks with skip connections, across which an increasing number of $3 \cdot 3$ filters was used (64, 128, 256, and 512), with a default stride of 1 in the first block and a stride of 2 in the remaining three blocks. In each skip block, the input was first convolved with the corresponding filters and default stride, then batch normalized and subjected to a ReLU function, then convolved with filters corresponding to the current block, however with a stride of 1, batch normalized and then added to a branch of the input that was only convolved with a $1 \cdot 1$ filter with default stride and batch normalized. The resulting activation was again subjected to an ReLU nonlinearity. After four of these blocks, an average pooling on groups of $7 \cdot 7$ was applied.

Triplet

We used SymTriplet loss (Zhang et al., 2017; Codella, 2020), a version of the triplet loss function (“Face-Net" Schroff et al., 2015). To do so, we connected the encoder module to

a dense mapping from the encoder output to a layer of 64 neurons. We then fed triplets of images to this encoder, consisting of an "anchor," a "positive," and a "negative," where the anchor and positive were random images constrained to be of the same identity while the negative was an image constrained to be of a different identity. The loss function then relates these three images in the 64-dimensional embedding space such that large Euclidean distances between anchor and positive, and short distances between anchor and negative, are penalized, as are short distances between positive and negative images. When training the parameters of this network, this yields a function that places samples of the same identity close to each other in the embedding space. The triplet loss network was trained with stochastic gradient descent with an initial learning rate of 10^{-3} until no more improvements were observed, and fine-tuned with a learning rate of 10^{-5} until no more improvements were observed.

ClassID

Here, we connected the encoder module to a flattening operation and performed a dense mapping to $2,004$ identity classes. We performed a softmax activation and applied a cross-entropy loss to train this classifier (Xu et al., 2018). We trained the ClassID network with a cyclical learning rate (Smith, 2017) that cycled between a learning rate of 10^{-6} and 0.3.

ClassMulti

This network was the same as the ClassID network; however, it classified not only the $2,004$ identity classes but also all other factors of variation that were part of the generation: the 500 identity residuals, the two sexes, the two ethnicities, the three ages, and the seven emotional expressions, as well as the three vertical and horizontal viewing and lightning angles. For each of these extra classification tasks, a separate dense mapping from the shared flattened encoder output was added to the architecture (Xu et al., 2018). We trained the ClassMulti network with a cyclical learning rate(Smith, 2017) that cycled between a learning rate of 10^{-6} and 0.3.

Autoencoder

For this architecture, we connected the encoder module to two branches, each consisting of a convolution with 512 $1 \cdot 1$ filters and a global average pooling operation. This was then connected to a decoder module, which up-sampled the 512-D vector back into the original $224 \cdot 224$ RGB image space. To do so, we used an existing decoder ("Darknet" decoder Graves, 2020). In brief, this decoder upsamples the spatial dimension gradually from a size of 1 to 7 and then in five steps that each double the spatial resolution to reach the resolution of the final image. Between these upsampling steps, the sample is fed through sets of blocks of convolution, batch normalization, and ReLU with the number

of filters alternating between $1,024$ and 512 in the first set of five blocks, between 256 and 512 in the second set of five blocks, between 256 and 128 in the third set of three blocks, between 128 and 64 in the fourth set of three blocks, staying at 64 in the fifth set of one block, and alternating between 32 and 64 in the last set of two blocks. The filter size in all of these blocks alternated between $3 \cdot 3$ and $1 \cdot 1$. Finally, the $224 \cdot 224 \cdot 64$ tensor was convolved with three filters of size $1 \cdot 1$ and passed through a $tanh$ nonlinearity.

The loss function used to optimize the parameters of this network is the classic reconstruction loss of an AE, operationalized as the MAE of the input image and the reconstruction in pixel space. We trained the AE using the Adam optimizer(Kingma & Ba, 2017) with an initial learning rate of 10^{-3} until no further improvements were observed.

View-invariant autoencoder

This network shared its architecture and training regime with the AE; however, we changed the input-output pairing during training. Instead of optimizing the parameters to reconstruct the unchanged input, the goal of the viAE was to reconstruct a frontalized view, independent of the pose of the input, while keeping all other factors of variation constant. This resulted in a more view- invariant representation in the bottleneck layer compared with the AE (Zhu et al., 2013).

Variational autoencoder

For this architecture (Kingma & Welling, 2014), we connected the encoder module to two branches, each consisting of a convolution with 512 $1 \cdot 1$ filters and a global average pooling operation. These were fed into a sampling layer as mean and variance inputs, transforming an input into a sample from a 512-D Gaussian with specified mean and diagonal covariance matrix.

This sample was then fed into the same decoder module as described for the AE and viAE above.

The loss function used to optimize the parameters of this network is the sum of two parts: The first is the reconstruction loss of a classic autoencoder, for which we used the MAE between the reconstruction and the original image. The second part is the Kullback-Leibler divergence measured between the multivariate normal distribution characterized by the mean and variance vectors passed into the sampling layer and the prior, a centered, uncorrelated, and isotropic multivariate normal distribution. The second part can be seen as a regularization that effectively leads to a continuous latent space. As it has been reported that weighing the second part of the loss function stronger than the first part can improve the disentanglement of the resulting latent space ("β-VAE" Higgins et al., 2016) we also repeated the training with several values of the regularization parameter β. However, this did not substantially change the latent space that we obtained.

We also trained two additional identity classifiers that used the frozen weights of the ($\beta = 1$)-VAE. The first directly connected the VAE encoder to a dense linear mapping to

$2,004$ identity classes. The second first passed through two blocks of fully connected layers of 512 neurons that were batch normalized and passed through an ReLU nonlinearity before the dense linear mapping to identity. In both cases, a $softmax$ activation function was applied and the resulting networks were trained with a cross-entropy loss function. All models shared the training regime of the AE and viAE models as described above.

4.5.5 Forward models

We were interested in comparing the degree to which various sets or "spaces" of predictors describing the rated stimuli were linearly relatable to the human behavioral responses. To do so in a way that minimizes the quantification of just overfitting, we linearly regressed the ratings on a range of different descriptors extracted from the random faces presented on each trial in a cross-validation framework.

The predictor spaces we used for this (each consisting of multiple predictor channels) were the texture and shape components of the single trials, as provided by the GMF, as well as the activations of the networks on their "embedding layers," as obtained from forward passes of the stimuli through the networks. Specifically, we used the 512-dimensional, pre-decision layers of the classifiers (ClassID and ClassMulti), the 64-dimensional final layer of the triplet loss network, and the 512-dimensional bottleneck layer of the AE, viAE, and VAE. We then also propagated images of the four target colleagues as recorded with the 3D capture system, fit by the GMF, and rendered with frontal viewing and lighting angles through the four networks, and computed the Euclidean distances on the embedding layers between the random faces of each trial and these ground truth images. We extended this by computing the channel-wise distances of each feature space and using them as an input to the regression described below to obtained weighted Euclidean distances. Additionally, we extracted the pre-$softmax$ activity ("logits") of the decision neurons trained to provide the logits for the four target colleagues in the final layer of the classifier networks (ClassID and ClassMulti, as well as the linear and nonlinear VAE classifiers). Since we were interested in assessing to what degree the GMF shape and texture features and various embedding layer activations provided the same or different information about the behavioral responses, we also considered models with joint predictor spaces consisting of the two subspaces of shape features and AE, viAE, or VAE activations as well as the three subspaces of shape features, texture features, and AE, viAE, or VAE activations. Lastly, to assess the extent to which a simple linear PCA could extract useful predictors from the images, we performed an SVD on the nonzero channels, a subset of the training images used for the DNNs. Performing SVD on the entire set of training images used for the DNNs would have been computationally infeasible. The subset we used consisted of $18,000$ RGB images of all $2,000$ identities rendered at nine different viewing angles, limiting emotion expression to the neutral condition and lighting angles to frontal angles. The first 512

dimensions could account for 99.5976% of variance in the training set. We projected the experimental stimuli onto these for further analyses.

We performed the regression separately for each participant and target colleague in a nested cross-validation procedure (Varoquaux et al., 2017). This allowed us to continuously tune the amount of $L2$ regularization necessary to account for correlated predictor channels and avoid excessive overfitting using Bayesian adaptive direct search (BADS Acerbi & Ma, 2017) a black-box optimization tool (see Daube et al., 2019b, for a comparable approach). Specifically, we divided the $1,800$ trials per participant into folds of 200 consecutive trials each and, in each of nine outer folds, assigned one of the resulting blocks to the testing set and eight to the development set. Then, within each of the nine outer folds, we performed eight inner folds, where one of the eight blocks of the development set was assigned to be the validation set and seven were assigned to the training set. In each of the eight inner folds, we fitted an $L2$ regularized linear regression ("ridge regression") using the closed form solution:

$$B = (X^T X + R)^{-1} X^T y \tag{4.7}$$

where B denotes the weights, y denotes the $n \cdot 1$ vector of corresponding human responses, R describes a regularization matrix, and X denotes the matrix of trials $n\cdot$ predictors M, where

$$M = \sum_{s=1}^{o} m_s \tag{4.8}$$

such that o denotes the number of combined predictor subspaces and ms describes the number of predictor channels in the sth subspace. In the cases where the features were combinations of multiple feature subspaces, i.e., where $o > 1$, we used a dedicated amount of $L2$ regularization for each subspace. This avoids using a common regularization term for all subspaces, which can result in solutions that compromise the need for high and low regularization in different subspaces, which fails to optimally extract the predictive power of the joint space. The regularization matrix R can then be described as

$$R = diag(\lambda_{1_1}, ..., \lambda_{m_1}, \lambda_{1_2}, ..., \lambda_{m_2}, ..., \lambda_{1_o}, ..., \lambda_{m_o}) \tag{4.9}$$

where λ_{c_s} describes the amount of $L2$ regularization for channel c of predictor subspace s, which is constant for all c in one s. For each predictor subspace, λ_{c_s} thus was one hyperparameter that we initialized at a value of 2^{17} and optimized in BADS with a maximum of 200 iterations, where the search space was constrained within the interval $[2^{-30}, 2^{30}]$. The objective function that this optimization maximized was Kendall's τ, as measured between predicted and observed responses of the inner fold validation set. We used the median of the optimal λ_{c_s} across all inner folds and retrained a model on the entire development set to then evaluate it on the unseen outer fold.

This yielded sets of 200 predicted responses for each test set of the nine outer folds. We evaluated them using two information theoretic measures: MI and redundancy, both computed using binning with three equipopulated bins (Ince et al., 2017). We computed bivariate MI with Miller-Madow bias correction between the predictions of each forward model and the observed human responses. We also computed redundancy, using a recent implementation of partial information decomposition (PID), I_{ccs} (Ince, 2017a). When there are two source variables and one target variable, PID aims to disentangle the amount of information the two sources share about the target (redundancy), the amount of information each source has on its own (unique information), and the amount of information that is only available when considering both sources. In our case, we were interested in quantifying how much information the predictions derived from DNN-based forward models shared with the predictions derived from GMF shape features about observed human behavior. To assess whether the amount of MI and redundancy exceeded chance level, we repeated the nested cross-validation procedure 100 times for each combination of participant and target colleague, each time shuffling the trials. From these surrogate data, we estimated null distributions of MI and redundancy and defined a noise threshold within each participant and target colleague condition as the 95^{th} percentile of MI and redundancy measured in these surrogate data. We counted the number of test folds of all participants and colleagues that exceeded this noise threshold and report this as a fraction relative to all data points.

To then assess whether different predictor spaces gave rise to different levels of MI and redundancy in the presence of high between-subject variance, we employed Bayesian linear models as implemented in the brms package (Bürkner, 2017), which provides a user-friendly interface for R(R Core Team, 2013) to such models using Stan (Stan Development Team, 2020). Specifically, we had performances (MI and redundancy) for each of the nine outer folds b for each combination of target colleague j, participant i, and all predictor spaces f_1 to f_q. The factor of interest were the predictor spaces f. We used Hamiltonian Monte-Carlo sampling with four chains of $4,000$ iterations each, $1,000$ of which were used for their warm-up. The priors for standard deviation parameters were not changed from their default values, i.e., half-Student-t distributions with three degrees of freedom, while we used weakly informative normal priors with a mean of 0 and a variance of 10 for the effects of individual predictor spaces. Specifically, we modeled the log-transformed and thus roughly normal distributed MI and redundancy

as performances k with the following model:

$$\begin{aligned} k_n &\sim N(\mu_n, \sigma^2) \\ \sigma &\sim |t(3,0,10)| \\ \mu_n &\sim \beta_{i:f[n]} + \beta_{i:b[n]} + \beta_{i:j[n]} + \beta_{f_1[n]} + \ldots + \beta_{f_q[n]} \\ (\beta_{i:f}, \beta_{i:b}, \beta_{i:j}) &\sim N(0, \sigma^2_{\beta_{int}}) \\ \sigma^2_{\beta_{int}} &\sim |t(3,0,10)| \\ \beta_{f_1}, \ldots, \beta_{f_q}) &\sim N(0,10) \end{aligned} \tag{4.10}$$

To compare the resulting posterior distributions of the parameters of interest, we evaluated the corresponding hypotheses using the brms package – $\beta_{f_a} - \beta_{f_b} > 0$ for all possible pairwise combinations of predictor spaces – and obtained the proportion of samples of the posterior distributions of differences that were in favor of the corresponding hypotheses.

As well as the predictions, the forward models also produced weights that linearly related predictors to predicted responses. We were interested in examining these weights to learn how individual shape features were used in the forward models. For the forward models, predicting responses from shape features was directly possible: the weights B_S mapped GMF shape features to responses and could thus be interpreted as the "shape receptive field." However, to be able to compare these weights on the vertex level, we used a differently scaled version of the shape features. This was obtained by multiplying the $4,735 \cdot 3$ z-scored 3D vertex level shape features with the pseudoinverse of the matrix of left-singular vectors U_S from the SVD performed on the identity residuals of the 3D vertex features of the face database (see "generative model of 3D faces"). This 355-dimensional representation of the shape features performed virtually identically to the unscaled version in the forward modeling. For visualization, we could then project the weights B_S from the 355D PCA component space into the $4,735 * 3$D vertex space, where the absolute values could be coded in RGB space. This resulted in a map that indicated how the random faces at each vertex affected the response predictions in the three spatial dimensions.

The weight maps B_N that form the forward models that relate DNN activations to responses were less simple to study in this shape space, since they mapped the less interpretable network activations, not GMF shape features, to behavioral responses. To interpret these models in vertex space, we re-predicted ("simulated") the response predictions $\hat{y}$ derived from DNN features with the GMF shape features to obtain re-predictions $\hat{\hat{y}}$ as well as weights B_{S_N}. We reasoned that response predictions of the ideal DNN model should be perfectly predictable by the shape features and that the corresponding simulated shape weights B_{S_N} should be identical to the original shape weights B_S in this case. We thus correlated the simulated response predictions with the DNN response predictions, as well as the simulated shape weights with the original shape weights for each test fold in each participant for each target colleague condition.

4.5.6 Decoding shape information from embedding layers

To understand what shape information is available on the embedding layers of the networks, independently of human behavior, we trained linear models that decoded GMF shape PCA components from embedding layer activations in response to images of faces. We used a cross-validation framework on the full set of stimuli, consisting of $43,200$ RGB images and their corresponding GMF shape PCA components, using a random set of 80% of the images for training, a further 10% for tuning, and the remaining 10% for testing. Specifically, we trained mass-multivariate $L2$ regularized regressions, separately predicting each GMF shape component from all neurons of the DNN embedding layers. Similar to the approach taken for the forward models, we tuned the $L2$ regularization using BADS to maximize the prediction performance on the tuning set. We then projected all predicted GMF shape PCA components into vertex space and, at each vertex, assessed the Euclidean distance between the original GMF shape model and the predictions from the DNN embedding layers.

4.5.7 Reverse correlation

To reconstruct internal templates of the target colleagues' faces under the GMF, we performed a mass-univariate linear mapping from the observed behavior of the human participants to each GMF shape and texture feature.

We repeated this with the choice behavior and rating behavior predicted by the forward models to compare these forward models, human observed behavior, and the ground truth shape information of the target colleagues as captured by our 3D camera array.

We performed the linear regressions of variation in the shape vertices and texture pixels of the random stimuli on the ratings of the images chosen by the human participants and their forward models based on GMF features, as well as DNN and PCA activations. This was done separately for each vertex and spatial dimension, as well as for each pixel and RGB dimension. In principle, this is equivalent to inverting the weights of the forward model (Haufe et al., 2014; van Vliet & Salmelin, 2020). However, to match the procedure in Zhan et al. (2019a), we re-estimated these parameters per vertex and pixel using the MATLAB function "robustfit".

Each of the $v = 1, ..., 4735 \cdot 3$ shape vertex positions s was thus modeled as

$$s_v = b_{0_v} + b_{1_v} \cdot r \tag{4.11}$$

and each of the $p = 1, ..., 800 \cdot 600 \cdot 3$ texture pixel RGB values t was modeled as

$$t_p = b_{0_p} + b_{1_p} \cdot r \tag{4.12}$$

Here, r are the vectors of observed or predicted responses, b_0 is an intercept term, and b_1 is a slope term.

In the original experiment, new faces were then generated by multiplying the slopes obtained from the regressions with different "amplification values". The resulting faces had then been presented to the participants to titrate the "amplification" of the weights that would result in the highest perceptual similarity of the reconstructed face for each participant. An amplification of 0 here corresponds to the shape or texture feature being reconstructed as a function of the intercept term only. This corresponds to the shape or texture feature resulting from the average of the faces chosen from the array of six faces in the first stage of each trial.

We repeated this for the forward models by storing the shape and texture components and by rendering observable images of faces corresponding to amplification values ranging from 0 to 50 (the same range used to titrate the human reconstructions) in steps of 0.5. We then computed forward model predictions from GMF shape and texture features, and propagated the observable images through encoding models based on DNNs. This resulted in responses of all systems across the range of amplification values. We chose the peak of each curve and reconstructed the internal templates corresponding to the shape and texture components at these peaks.

We rendered the corresponding internal templates as intuitively visualizable faces. We also considered the explicit descriptions in vertex space to compare templates from humans and templates from forward models among each other, and with the ground truth face shape from the target colleagues. To evaluate the "humanness" of the forward models, we computed the Euclidean distances and correlations from the internal templates of the forward models with the internal templates of the humans. To also evaluate the "veridicality," we computed the Euclidean distances and correlations from the ground truth target colleagues with the internal templates from the forward models and the human participants.

This resulted in Euclidean distances and correlations for each target colleague condition j and human participant i (observed and predicted by different predictor spaces f). We then log-transformed the Euclidean distances and Fisher z-transformed the correlations to obtain evaluation measures e and modeled them with Bayesian hierarchical models similar to the ones used to model the prediction performances of the forward models:

$$\begin{aligned} e_n &\sim N(\mu_n, \sigma^2) \\ \sigma &\sim |t(3,0,10)| \\ \mu_n &\sim \beta_{i:f[n]} + \beta_{i:j[n]} + \beta_{f_1[n]} + \ldots + \beta_{f_q[n]} \\ (\beta_{i:f}, \beta_{i:j}) &\sim N(0, \sigma^2_{\beta_{int}}) \\ \sigma^2_{\beta_{\text{int}}} &\sim |t(3,0,10)| \\ (\beta_{f_1}, \ldots, \beta_{f_q}) &\sim N(0,10) \end{aligned} \tag{4.13}$$

To compare the resulting posterior distributions of the parameters of interest, we evaluated the corresponding hypotheses using the brms package – $\beta_{f_a} - \beta_{f_b} > 0$ for all possi-

ble pairwise combinations of predictor spaces – and obtained the proportion of samples of the posterior distributions of differences that were in favor of the corresponding hypotheses. Prior to visualization, we back-transformed the posterior distributions of the log Euclidean distances with an exponential and the posterior distributions of correlations with the inverse Fisher z-transformation.

4.5.8 Generalization testing

The models of human behavior had been trained and tested under the same conditions. To also test how they would perform under data from a different distribution, we re-used data from a validation experiment originally conducted by Zhan et al. (2019a).

We propagated the 50 stimulus images per target colleague (combinations of two levels of diagnosticity at five levels of amplification, which were each rendered in five different generalization conditions, see "experiments – generalization testing") through each of the model systems under consideration and extracted the rating predictions for each of the 14 participants of the first experiment for each of the four colleagues from each of the four correspondingly fitted forward models. Next, we normalized the predictions to values between 0 and 1 within target colleagues to eliminate possible biases from participants rating the random stimuli of the first experiment higher for one target colleague than for others. We then used the maximum predicted rating across all target colleagues for a given stimulus as the choice of the respective system. The predictions for each of the 14 participants of the first experiment were compared with the behavior of each of the 12 additional participants of the second experiment.

Since all systems were deterministic, the resulting accuracy values for the systems were thus binary (this was different for the human responses, since each stimulus had been shown to the validators five times; see "experiments—generalization testing").

We analyzed the data by first computing the absolute difference of human and model accuracies and then subjecting the resulting absolute errors to a Bayesian linear model. Since the model accuracies could only take one of six different values (from 0 to 1 in steps of 0.2), we used an ordinal model. To do so, we used a cumulative model assuming a normally distributed latent variable as implemented in brms (Bürkner & Vuorre, 2019). Concretely, we modeled the probability of a model accuracy a of model type f predicting behavior in task g of participant i for target colleague j and validated by validator k to fall into category t given the linear predictor η as:

$$\Pr(a = t \mid \eta) = F(\tau_t - \eta) - F(\tau_{t-1} - \eta) \tag{4.14}$$

where F is a cumulative distribution function, τ_t is one of $T = 5$ different thresholds that partition the standard Gaussian continuous latent variable $\tilde{a}$ into $T + 1$ categories,

and η describes $\tilde{a}$ corresponding to the following model:

$$\begin{aligned} \tau_t &\sim t(3,0,10) \\ \tilde{a}_n &\sim N(\mu_n,1) \\ \mu_n &\sim \beta_{f:g[n]} + \beta_{i:j:k[n]} \\ (\beta_{f:g},\beta_{i:j:k}) &\sim N(0,\sigma^2_{\beta_{int}}) \\ \sigma^2_{\beta_{\text{int}}} &\sim |t(3,0,10)| \\ (\beta_{f_1},\ldots,\beta_{f_q}) &\sim N(0,10) \end{aligned} \tag{4.15}$$

To compare the resulting posterior distributions of the parameters of interest, we evaluated the corresponding hypotheses using the brms package ($\beta_{f_a:g_x} - \beta_{f_b:g_x} > 0$ for all possible pairwise combinations of model types within each task), and obtained the proportion of samples of the posterior distributions of differences that were in favor of the corresponding hypotheses.

Chapter 5

General discussion

This thesis consists of three suggestions on how to contextualise bivariate encoding and decoding models with higher-order information theory. In chapter 2, this was done by relativising the correlation between a set of stimulus features and responses as achievable by other and potentially even simpler sets of features. Chapter 3 then looked at a new idea on how to consider the response's own past as a variable to explain the stimulus-response correlation. Finally, chapter 4 added experimental control over the stimulus features to constrain what features models could possibly predict responses with. We thus extended descriptions of mere correspondence between the stimuli and responses to trivariate relationships. These extensions directly echo the demand for stricter considerations of the concept of mental representations, which requires more than just correlations between features of the outside world and responses (Baker et al., 2021).

The rest of this discussion will highlight differences and commonalities between the approaches taken in the individual chapters. While chapters 2 and 3 considered neuronal responses to naturalistic acoustic stimuli as recorded with MEG, chapter 4 considered behavioural responses to controlled visual stimuli. Leaving comparisons of visual and auditory perception aside, this provides interesting tensions:

Should we continue to do research on naturalistic paradigms, or should we abandon them in favour of experimental control? What insights do we really get in return for the efforts we undertake to record neuroimaging data, which are certainly enormous in comparison to relatively easily obtainable behavioural data?

All chapters on the other hand subscribe to a “data-driven" approach. Since many studies credit themselves with this label, and since it is not always clear what it is supposed to entail, the final discussion section will flesh out what it means within the context of this thesis.

5.1 Naturalistic versus controlled paradigms

At the start of this PhD, the little experience in cognitive neuroscience I could draw from consisted of several undergraduate projects which I believed all suffered most from two problems: small sample sizes and experimental designs full of assumptions that strongly limited opportunities to carry out analyses of a deliberately exploratory nature. Under the impression of the replication crisis (Ioannidis, 2005) and the increasingly popular and powerful machine learning models of the 2010s, I grew an interest in larger datasets on which these types of models could be leveraged to get to more robust conclusions. At the time, this was (and arguably still is at the time of the submission of this thesis) a popular best bet on what would open up perspectives on questions of perceptual processes and mental representations that had hitherto been unexplored (Bzdok & Yeo, 2017). Especially studies of the type as vigorously promoted by the Gallant lab (Huth et al., 2012, 2016; de Heer et al., 2017; Hamilton & Huth, 2018) seemed to epitomise this approach, arguing for paradigms of passive viewing or listening of naturalistic stimuli that would yield large amounts of data in comparably short experimental time. Moreover, such paradigms would keep participants entertained by a task that is a voluntarily chosen leisure activity for many people. This would minimise detrimental effects of frustration and fatigue, which are commonly observed in more classical psychophysics paradigms. Reasoning about the perceptual processes underlying the neuronal responses would be formalised in computational models competing to explain the most response variance, and thereby move beyond the seemingly speculative box-and-arrow models I had been taught during my undergraduate studies. The models could be explored after data collection, with seemingly little limitation. In comparison, constructing an experimental paradigm from prior assumptions to then go through the arduous process of data collection, only to conclude that the initial assumptions were nonsensical to start with, seemed to be an unjustifiably risky and cumbersome process for an early career researcher.

In retrospect, I still believe that these arguments have kept much of their relevance. Having at least one computational model that robustly explains response variance as a function of complex naturalistic stimulus input is a non-trivial achievement, and a stepping stone for lively discussions about alternatives that are simpler, explain more variance, generalise to more conditions or satisfy more biological or cognitive constraints. Broadly sampled naturalistic stimulus material is a crucial starting point for such experiments, as it will yield the strongest neuronal responses (Rieke et al., 1995) leading to the best generalisation within the behaviourally relevant stimulus regime (Theunissen et al., 2000; David et al., 2004). This approach is currently experiencing a surge of publications that succeed in explaining ever more response variance by leveraging ever more sophisticated computational models (Donhauser & Baillet, 2020; Jain et al., 2021; Schmitt et al., 2021; Caucheteux et al., 2021). Furthermore, it has has paved the way towards applied research on hearing aids (O'Sullivan et al., 2015; Mirkovic et al., 2015; Fiedler et al., 2017; Geirnaert et al., 2021), cementing the relevance of the modeling of

(neuronal) responses to naturalistic stimuli as an intrinsically interesting basic science endeavour.

And yet, during the process of chapter 2, I started to think that the grass of naturalistic stimuli had just been greener by virtue of being on the other side. While it is possible to explore a wide range of hypotheses in such datasets, it is hard to then rule out confounds. This is especially true for “oracle models" (Kriegeskorte & Douglas, 2018a) that do not provide end-to-end stimulus-computable hypotheses, where for example simpler models that suggest lower-level processing could predict the same variance. The problem particularly applies to a growing body of work whose models rely on text alignment procedures in order to subsequently leverage ideas from natural language processing (see e.g. Millet & King, 2021, for a notable exception employing models directly operating on the speech sounds). Ideally, the field should move towards the use of models that arrive at the targeted phonetic and semantic feature spaces without implausible processing steps requiring a textual annotation (van den Oord et al., 2018a; Chung et al., 2020; Lakhotia et al., 2021). Testing their successive hierarchies could provide a clearer idea of the required representational complexity for a given type of responses, such that the risk of an overestimation could be mitigated.

Experiments that explicitly control for low- or higher level (e.g. linguistic, Taylor et al., 2020) confounds however are indispendable when arguing for high-level processing (Ding et al., 2016). To the extent to which such controlled experiments additionally include generalisation conditions that manipulate isolated stimulus dimensions (Zhan et al., 2019a), they can increase their value for model comparisons as in chapter 4. Controlled experiments can further consider specific non-natural stimuli that are known to be characteristic for a given system or set of systems of interest. These include human perceptual illusions, adversarial examples (Szegedy et al., 2014; Jacobsen et al., 2019), model metamers (Feather et al., 2019), controversial stimuli (Golan et al., 2020) and stimuli that are specifically optimised to maximise a given neuronal response (Ponce et al., 2019). Such positions in stimulus space are of high value for model comparisons, but cannot be included in naturalistic experiments by definition. As the development of generative models progresses (Gan et al., 2020; Daube et al., 2021; Goetschalckx et al., 2021; de Melo et al., 2021), controlled experimentation on the other hand loses its necessary synonymy with unnatural stimulus statistics and the entailed impoverished neuronal and behavioural responses (Rieke et al., 1995; Theunissen et al., 2000; Jack & Schyns, 2017; Hamilton et al., 2018).

Taken together, both naturalistic and controlled experiments have been and will be important components of cognitive neuroscience in their own right (see e.g. Oganian & Chang, 2019, for a hopefully trendsetting combination of the two). Broadly sampled naturalistic experiments are well suited to explore ideas and build intuitions for models that cover the behaviourally most relevant region of stimulus space. Subsequent rigorous testing both inside and outside of this region along isolated dimensions is then necessary to identify models with a better generalisation performance. Fields as well as

individual researchers will run into danger when prematurely moving on to confirmatory research, and likewise when refusing to mature by clinging to exploratory approaches (Tukey, 1980).

5.2 Behavioural versus neuronal responses

A question that is in practise – but not in principle (Hebart et al., 2020) – correlated with the distinction of naturalistic and controlled experiments is that of neuronal and behavioural data.

A central tenet of neuroscience as such is that all behaviour and conscious experience critically rely on the activity and interactions of the approximately 86 billion (Azevedo et al., 2009) neurons in the human brain. Studying them promises to move beyond a limitation to the input and output of a system (Chomsky, 1959) and to identify the physical substrate of the complex processing taking place in between – to learn about the mechanisms giving rise to cognition (Boone & Piccinini, 2016). The abandonment of behavioural responses can then be motivated from the perspective of maximising the amount of unavoidably noisy data recordable from neural activity in expensive experimental time.

Consequently, the classic passive listening (or viewing) paradigm either reports no behavioural testing at all (e.g. Huth et al., 2016; Di Liberto et al., 2015; Brodbeck et al., 2018a) or only very coarse checks of understanding and attention (e.g. Daube et al., 2019b; Donhauser & Baillet, 2020; Heilbron et al., 2021; Gwilliams et al., 2020). As described in the previous section, the resulting datasets are well-suited for exploratory and basic research, with the hearing-impaired population being an increasingly realistic potential benefactor.

Another motivation for datasets of isolated neuronal responses is that healthy human participants usually “cannot help" but follow the content of passive perceptual paradigms. Despite the absence of an explicitly defined behavioural task, this approach constrains brain activity to a degree that correlations across participants even beyond low-level sensory areas can be found (Hasson et al., 2004). When scrambling the stimulus material at different linguistic scales reaching from words to paragraphs, these intersubject correlations reveal a hierarchy from the low-level sensory areas to higher-order parietal and frontal regions (Lerner et al., 2011). Such analyses thus aim to realise the dream of studying the brain “from the inside out" (Buzsáki, 2019), i.e., the dream of unveiling of neuroscientific phenomena under a minimisation of “philosophical constraints" (Buzsáki, 2020).

And yet, it is notoriously impossible to escape philosophy, theory and prior assumptions (Poeppel & Adolfi, 2020; Gershman, 2021). An overly narrow interpretation or glorification of the “epistemological primacy of hardware" (Poeppel & Adolfi, 2020) runs in danger of discarding the relevance of behavioural data (that is, discarding the field of psychology). This would imply a great loss for neuroscience (Krakauer et al., 2017;

Musall et al., 2019; Niv, 2021). The ultimate purpose of the brain is to allow an organism to behave adaptively in its environment. Such behaviour can be recorded with virtually zero measurement noise (i.e. the only “noise" is that of behaviour itself that cannot be explained by available predictors). Given a defined perceptual query, the behavioural response is interpretable as the participant’s decision, integrating all neuronal processes leading to it in specifically that way that characterises how this participant or organism under study attempts to behave adaptively – that is, all the read-outs of neurons involved in this particular response are weighted and summed up meaningfully.

Any quantification of neuronal processes on the other hand will always be compromised in three important ways: Firstly, it will be mixed with measurement noise. Secondly, it will be a subset of the entire neuronal response, i.e. even with the most sophisticated synchronous combination of multiple neuroimaging modalities, it is currently inevitable to miss unquantifiable large parts of the processes of interest. Thirdly, the information contained in recordings of neuronal activity cannot readily be assumed to have any relevance for downstream neuronal processes or the final goal, adaptive behaviour (de-Wit et al., 2016; Baker et al., 2021). It is admittedly questionable to which degree an organ that has undergone such high evolutionary pressure for energy efficiency as the human brain does not itself exploit (i.e., read out) processes so energy intensive that they can be noninvasively recorded. However, the way we record neuronal activity is only in extremely rare and isolated experimental cases how other neurons “listen” to this neuronal activity.

These three problems are all non-trivially handicapping the interpretation of isolated neuronal recordings for the algorithmic understanding of neuronal processes (see chapter 2). For such interpretations, strong assumptions about the noise free unmeasurable part of the entire population response and its integration across subprocesses are necessary. An interesting case of isolated neuronal data are the rare opportunities of invasive recordings, especially when paired with direct electrical stimulation (Hamilton et al., 2021; Veit et al., 2021). However, even then it is the combination with behavioural responses (such as the report of intelligibility of speech signals under electrical stimulation of different regions, Hamilton et al., 2021) that affords the strongest conclusions. In this spirit, another exciting and more broadly applicable proposal is that of analysing paired neuronal and behavioural experimental data with a focus on neuronal response components to stimuli that are redundant with behavioural responses (Williams et al., 2007; Panzeri et al., 2017; Bouton et al., 2018; Keitel et al., 2018; Zhan et al., 2019b; Schyns et al., 2020). This can be combined with stimulus-computable models that provide hypotheses of nonlinear transformations between latent stimulus representations and thus potentially increase the sensitivity to the processing of more abstracted information.

In sum, neuronal and behavioural data can be used to answer different representational questions (Baker et al., 2021). With purely neuronal datasets, one is limited to interpretations of correspondences between stimuli and responses. As done in chapter 2, this can be extended to the relationship of multiple correspondences to one another.

It can also be extended by relation to the response's own history as in chapter 3, which can point to further algorithmic characterisation of the responses. Questions concerning the representation of information in a stricter sense however require behavioural data as in chapter 4. To link such representations to the neuronal underpinnings, a combination of neuronal and behavioural data is suited best.

5.3 Data-driven approaches

During my PhD, there were many debates in which it was tried to create two camps: one being on the "big data" side, and one being on the "big theory" side (e.g. Cognitive Neuroscience Society, 2018). The availability of ever growing datasets and ever more powerful computer hardware has led to a high popularity of "data-driven" approaches (Bzdok & Yeo, 2017; Bzdok, 2017; Buzsáki, 2019, 2020; Peterson et al., 2021b). In reaction to this, highlighting arguments against "big data" in isolation is similarly in vogue (Poeppel & Adolfi, 2020; Gershman, 2021). In general, I share the common opinion that forcing a "data versus theory" dichotomy is not helpful to achieve the necessary balance of both (Sejnowski et al., 2014; Frégnac, 2017). However, I feel provoked to defend the "data-driven" stance I took up for large parts of the thesis against what I think often are straw-man arguments. It is not at all synonymous with the belief that hypotheses must be avoided, that data alone will paint a picture of interpretable scientific results obviating the need for biasing theory or that domain knowledge is to be ignored.

Instead, I suggest to understand it as a proxy for a careful updating of beliefs in light of data. This entails the belief that the process of a rigorous mathematical formalisation of hypotheses is a chance (but not a guarantee) to become aware of their assumptions and shortcomings. Further, it contends that adaptive parameterisations can lay the foundations for abstraction from ideas in order to further increase their reproducibility and generalisability. Its position with respect to domain knowledge is that it is to be questioned in order to effectively exploit it and its associated uncertainty for the question at hand.

The perhaps simplest example of this in this thesis is arguably provided in chapter 3. Here, a relevant subproblem was to predict samples of a time-series from their own past. A popular choice to implement this was to use a single delay. The problems caused by this could be improved with a more flexible implementation relying on a multidimensional and non-uniform parameterisation (Vlachos & Kugiumtzis, 2010; Faes et al., 2011; Wibral et al., 2013).

Both chapters 2 and 4 made use of a more flexible parameterisation of the classic ridge regression (Hoerl & Kennard, 1970) to predict responses from stimuli. Here, a hyperparameter λ is used to reduce overfitting by adjusting the cost associated with large weights. The optimal choice of this hyperparameter is affected by the statistics of the predictor. When multiple sets of predictors with different statistics are combined, usually a compromise is made by choosing the λ that is best for the average predictor.

This was improved by instead using independent λs for each set of predictors, such that each could contribute its maximal predictive power (Daube et al., 2018). This approach has since been independently developed by other labs (under the name “banded ridge regression" Nunez-Elizalde et al., 2019), and is increasingly adopted (e.g. Sohoglu & Davis, 2020; Dupré la Tour et al., 2021).

As chapter 2 focussed on the analysis of neuroimaging data, it entailed many more parameter settings. A classic case in MEG is the regularisation of the sensor covariance matrix for the ill-posed inverse problem. For a class of spatial filters called beamformers, the choice of 5% is suggested in a tutorial of a popular analysis toolbox (Oostenveld et al., 2011) and as a likely consequence has reached high popularity. In many cases, this heuristic is probably not a bad choice. However, a data driven approach suggests to treat this as a flexible parameter that can be optimised with respect to a concrete goal (Woolrich et al., 2011; Engemann & Gramfort, 2015). As a result, we obtained values low enough to prevent a mixing of source estimates with unrelated sources and high enough to provide a robust, noise suppressive spatial filter. This optimised the performance of the respective encoding models predicting the filter output for individual participants and even individual points in source space. The values we found were often considerably higher than the usually chosen 5%, exhibiting considerable stability within a hemisphere of a given participant and considerable variance across participants.

Another example was the temporal extent of the encoding models in the same chapter. This parameter determines which temporal segment of the stimulus is considered to predict a given sample of the response. Our optimised parameters revealed characteristic ranges for different predictor sets, showcasing again an adaptive flexibility across and within participants. An adaptive parameterisation of such an interpretable parameter is especially interesting in scenarios where multiple regions of interest with similar signal-to-noise ratios can be identified, such that it becomes possible to compare the “temporal receptive windows" of these regions (Lerner et al., 2011).

Most of the above examples critically relied on an efficient black-box optimisation algorithm (Acerbi & Ma, 2017), allowing to optimise parameters without differentiation. It constructs a computationally cheap model of the function it is consigned to optimise and then samples outputs from the actual black-box function in places that lead to the greatest reduction in model uncertainty. This entails two benefits over the commonly chosen alternative of a grid search: it treats hyperparameters as continuous, and it can handle a (limited) amount of multiple hyperparameters much more efficiently, i.e. without a combinatorial explosion of running time.

Traditionally, such optimisations are seen as resulting in an increase of the “researcher degrees of freedom”. Usually, they are thus discouraged as they increase the danger of overfitting and easily compromise the reproducibility of results (Nichols et al., 2017). However, the use of a robust nested cross-validation framework (Varoquaux et al., 2017) should effectively limit this danger. Instead, such optimisation efforts thus contribute to fairer model comparisons, where each model’s chances are maximised

by the parameter settings that make it perform best. They make it possible to evolve from parameter settings that are traditional in the field to potentially surprising optimal choices.

Taken together, data-driven science means rejecting “quick and dirty solutions" to questions of parameter settings by reverting to traditional folklore of the field or cherry picked values. Such pragmatism is certainly sensible when initially exploring datasets. When attempting to rigorously compare models however, we empirical scientists should attempt to allow for an integration of our hypotheses with the “bottom-up sensory input" we face when looking at data.

5.4 Conclusion

By using a principled information theoretic framework, this thesis contextualised the bivariate relationship between neuronal or behavioural responses to aspects of the outside world. These contexts were provided by multiple operationalisations of the outside world with different complexity, the response’s own history and an experimentally controlled “world" that constrained what models could predict responses with. Such genuinely multivariate perspectives in cognitive computational neuroscience break up its confinement to the study of simple correspondence problems, and are thus an important tool to realise the promise of cognitive computational neuroscience to identify and characterise the mechanisms underlying cognition.

Bibliography

Aaltonen, O., Niemi, P., Nyrke, T., & Tuhkanen, M. (1987). Event-related brain potentials and the perception of a phonetic continuum. *Biological Psychology*, 24(3):197–207.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.

Abel, L. A. & Quick Jr, R. (1978). Wiener analysis of grating contrast judgments. *Vision research*, 18(8):1031–1039.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *frontiers in Neuroinformatics*, 8(14).

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *Journal of Neuroscience*, 28(15):3958–3965.

Acerbi, L. & Ma, W. J. (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. *Advances in Neural Information Processing Systems*, 30:1834–1844.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, 98(23):13367–13372.

Ahumada Jr, A. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception*, 25(1_suppl):2–2.

Ahumada Jr, A. & Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, 49(6B):1751–1756.

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., Filho, W. J., Lent, R., & Herculano-Houzel, S. (2009). Equal numbers of neuronal and

nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513(5):532–541.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140. doi: 10.1371/journal.pone.0130140. URL https://dx.plos.org/10.1371/journal.pone.0130140.

Baker, B., Lansdell, B., & Kording, K. (2021). A Philosophical Understanding of Representation for Neuroscience. *arXiv preprint arXiv:2102.06592*.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12):e1006613. doi: 10.1371/journal.pcbi.1006613. URL http://dx.plos.org/10.1371/journal.pcbi.1006613.

Ballard, D. H. (1987). Modular learning in neural networks. In *Proceedings of the sixth National conference on Artificial intelligence - Volume 1*, AAAI'87, pages 279–284, Seattle, Washington. AAAI Press.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., & Katz, B. (2019). ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *arXiv*, page 11.

Barnett, L., Barrett, A. B., & Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical review letters*, 103(23):238701.

Barnett, L. & Seth, A. K. (2011). Behaviour of Granger causality under filtering: theoretical invariance and practical application. *Journal of neuroscience methods*, 201(2):404–419.

Barrett, A. B. (2015). Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E*, 91(5):052802.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436. doi: 10.1126/science.aav9436. URL https://www.sciencemag.org/lookup/doi/10.1126/science.aav9436.

Bassett, D. S. & Bullmore, E. (2006). Small-world brain networks. *The neuroscientist*, 12(6):512–523.

Bastos, A. M. & Schoffelen, J.-M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, 9:175.

Bentin, S., Kutas, M., & Hillyard, S. A. (1993). Electrophysiological evidence for task effects on semantic priming in auditory word processing. *Psychophysiology*, 30(2):161–169.

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A. J., & Ramsey, N. F. (2017). Neural Tuning to Low-Level Features of Speech throughout the Perisylvian Cortex. *The Journal of Neuroscience*, 37(33):7906–7920.

Berger, H. (1929). Über das Elektroenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 87(1):527–570.

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., & Ay, N. (2014). Quantifying unique information. *Entropy*, 16(4):2161–2183.

Besserve, M., Schölkopf, B., Logothetis, N. K., & Panzeri, S. (2010). Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis. *Journal of computational neuroscience*, 29(3):547–566.

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2017). Auditory-Inspred Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(5):402–412.

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2020). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition*, page 104341. doi: 10.1016/j.cognition.2020.104341. URL http://www.sciencedirect.com/science/article/pii/S0010027720301608.

Bodamer, J. (1947). Die Prosop-Agnosie. *Archiv für Psychiatrie und Nervenkrankheiten*, 179(1):6–53.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5:341–345.

Bontrager, P., Lin, W., Togelius, J., & Risi, S. (2018). Deep Interactive Evolution. *arXiv:1801.08230 [cs]*. URL http://arxiv.org/abs/1801.08230.

Boone, W. & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese*, 193(5):1509–1534.

Boto, E., Holmes, N., Leggett, J., Roberts, G., Shah, V., Meyer, S. S., Muñoz, L. D., Mullinger, K. J., Tierney, T. M., Bestmann, S., et al. (2018). Moving magnetoencephalography towards real-world applications with a wearable system. *Nature*, 555(7698):657–661.

Bouton, S., Chambon, V., Tyrand, R., Guggisberg, A. G., Seeck, M., Karkar, S., van de Ville, D., & Giraud, A. L. (2018). Focal versus distributed temporal cortex activity for speech sound category assignment. *Proceedings of the National Academy of Sciences of the United States of America*, 115:E1299–E1308.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10:433–436.

Brendel, W. & Bethge, M. (2019). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. *arXiv:1904.00760 [cs, stat]*. URL http://arxiv.org/abs/1904.00760. arXiv: 1904.00760.

Brette, R. (2018). Is coding a relevant metaphor for the brain? *bioRxiv*, 168237.

Broca, P. et al. (1861). Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique*, 6:330–57.

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018a). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, 28:3976–3983.

Brodbeck, C., Jiao, A., Hong, L. E., & Simon, J. Z. (2020). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. *PLoS biology*, 18(10):e3000883.

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018b). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage*, 172:162–174.

Brodbeck, C. & Simon, J. Z. (2020). Continuous speech processing. *Current Opinion in Physiology*.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018a). Data from: Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural Narrative Speech. *Dryad Digial Repository*.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018b). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural Narrative Speech. *Current Biology*, 28:803–809.

Brookes, M. J., Woolrich, M., Luckhoo, H., Price, D., Hale, J. R., Stephenson, M. C., Barnes, G. R., Smith, S. M., & Morris, P. G. (2011). Investigating the electrophysiological basis of resting state networks using magnetoencephalography. *Proceedings of the National Academy of Sciences*, 108(40):16783–16788.

Buckner, C. (2019). The Comparative Psychology of Artificial Intelligences. *philsci-archive.pitt.edu/16034/*. URL http://philsci-archive.pitt.edu/16034/. Type: Preprint.

Bürkner, P. & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):77–101. URL https://journals.sagepub.com/doi/full/10.1177/2515245918823199.

Bürkner, P. C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1).

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press.

Buzsáki, G. (2020). The brain–cognitive behavior problem: a retrospective. *Eneuro*, 7(4).

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Frontiers in neuroscience*, 11:543.

Bzdok, D. & Yeo, B. T. T. (2017). Inference in the age of big data: Future perspectives on neuroscience. *NeuroImage*, 155:549–564.

Carlson, T., Goodard, E., Kaplan, D. M., Klein, C., & Ritchie, J. B. (2018). Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *NeuroImage*, 180:88–100.

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Long-range and hierarchical language predictions in brains and algorithms. *arXiv preprint arXiv:2111.14232*.

Chaitin, G. J. (1975). A Theory of Program Size Formally Identical to Information Theory. *J. Assoc. Comput. Mach*, 22:329–340.

Chang, L., Egger, B., Vetter, T., & Tsao, D. Y. (2021). Explaining face representation in the primate brain using different computational models. *Current Biology*, 31(13):2785–2795.e4. doi: 10.1016/j.cub.2021.04.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982221005273.

Chang, L. & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, 169(6):1013–1028.e14. doi: 10.1016/j.cell.2017.05.011. URL https://linkinghub.elsevier.com/retrieve/pii/S009286741730538X.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association*, 68(342):361–368.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906. doi: 10.1121/1.1945807. URL https://asa.scitation.org/doi/abs/10.1121/1.1945807. Publisher: Acoustical Society of America.

Chollet, F. & others (2015). https://github.com/fchollet/keras.

Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. *Language*, 35(1):26–58.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*, volume 11. MIT press.

Chung, Y.-A., Tang, H., & Glass, J. (2020). Vector-quantized autoregressive predictive coding. *arXiv preprint arXiv:2005.08392*.

Chung, Y.-A., Weng, W. H., Tong, S., & Glass, J. (2018). Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces. *arXiv*, 1805.07467.

Cliff, O. M., Novelli, L., Fulcher, B. D., Shine, J. M., & Lizier, J. T. (2021). Assessing the significance of directed and multivariate measures of linear dependence between time series. *Physical Review Research*, 3(1):013145.

Codella, N. (2020). github.com/noelcodella/tripletloss-keras-tensorflow.

Cognitive Neuroscience Society (2018). Big Theory versus Big Data, CNS 2018 : DEBATE Moderated by David Poeppel.

Cohen, D. (1968). Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. *Science*, 161(3843):784–786.

Cohen, D. & Cuffin, B. N. (1983). Demonstration of useful differences between magnetoencephalogram and electroencephalogram. *Electroencephalography and clinical Neurophysiology*, 56:38–51.

Cover, T. & Thomas, J. (1991). *Elements of Information Theory*. Wiley New York.

Crosse, M. J., DiLiberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *frontiers in Human Neuroscience*, 10(604).

Daube, C., Giordano, B. L., Schyns, P., & Ince, R. (2019a). Quantitatively comparing predictive models with the Partial Information Decomposition. In *2019 Conference on Cognitive Computational Neuroscience*, Berlin, Germany. Cognitive Computational Neuroscience.

Daube, C., Gross, J., & Ince, R. A. A. (2022). A whitening approach for Transfer Entropy permits the application to narrow-band signals. *arXiv:2201.02461*.

Daube, C., Ince, R. A. A., & Gross, J. (2018). Phoneme-level processing in low-frequency cortical responses to speech explained by acoustic features. In *2018 Conference on Cognitive Computational Neuroscience*, Philadelphia, PA, USA. Cognitive Computational Neuroscience.

Daube, C., Ince, R. A. A., & Gross, J. (2019b). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology*, 29(12):1924–1937.e9. doi: 10.1016/j.cub.2019.04.067. URL https://www.sciencedirect.com/science/article/pii/S0960982219304968.

Daube, C., Xu, T., Zhan, J., Webb, A., Ince, R. A., Garrod, O. G., & Schyns, P. G. (2021). Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns*, 2(10):100348.

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience*, 24(31):6991–7006.

Davis, H., Mast, T., Yoshie, N., & Zerlin, S. (1966). The slow response of the human cortex to auditory stimuli: recovery process. *Electroencephalography and clinical neurophysiology*, 21(2):105–113.

Davis, P. A. (1939). Effects of acoustic stimuli on the waking human brain. *Journal of neurophysiology*, 2(6):494–499.

Dayan, P. & Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT Press.

De Boer, E. & Kuyper, P. (1968). Triggered correlation. *IEEE Transactions on Biomedical Engineering*, 15(3):169–179.

de Cheveigné, A., Wong, D. D. E., Di Liberto, G. M., Hjortkjær, J., Slaney, M., & Lalor, E. C. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage*, 172:206–216.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience*, 37:6539–6557.

De Lange, P., Boto, E., Holmes, N., Hill, R. M., Bowtell, R., Wens, V., De Tiège, X., Brookes, M. J., & Bourguignon, M. (2021). Measuring the cortical tracking of speech with optically-pumped magnetometers. *NeuroImage*, 233:117969.

de Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., & Hodgins, J. (2021). Next-generation deep learning based on simulators and synthetic data. *Trends in cognitive sciences*.

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Reviews*, 23:1415—1428.

Destoky, F., M, P., Bertels, J., Verhasselt, M., N, C., van der Ghinst M, V, W., X, D. T., & Bourguignon, M. (2019). Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope. *NeuroImage*, 184:201–213.

DeWitt, I. & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, 109:E505–E514.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, 25:2457–2465.

DiCarlo, J. J. & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341. doi: 10.1016/j.tics.2007.06.010. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661307001593.

Diedrichsen, J. (2020). Representational Models And the Feature Fallacy. In Poeppel, D., Mangun, G. R., & Gazzaniga, M. S., editors, *The Cognitive Neurosciences*. MIT Press, 6th edition.

Dimensional Imaging Ltd. (2021). di4d.com.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1):158–164.

Ding, N. & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11854–11859.

Ding, N. & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience*, 8:311.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, 85:761–768.

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLOS Computational Biology*, 16(7):e1008017. doi: 10.1371/journal.pcbi.1008017. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008017.

Donhauser, P. W. & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, 105(2):385–393.

Donoghue, T., Haller, M., Peterson, E. J., Varma, P., Sebastian, P., Gao, R., Noto, T., Lara, A. H., Wallis, J. D., Knight, R. T., et al. (2020). Parameterizing neural power spectra into periodic and aperiodic components. *Nature neuroscience*, 23(12):1655–1665.

Dorman, M. F. (1974). Auditory evoked potential correlates of speech sound discrimination. *Perception & Psychophysics*, 15(2):215–220.

Dotsch, R., Wigboldus, D. H., Langner, O., & van Knippenberg, A. (2008). Ethnic outgroup faces are biased in the prejudiced mind. *Psychological science*, 19(10):978–980.

Dupré la Tour, T., Lu, M., Eickenberg, M., & Gallant, J. L. (2021). A finer mapping of convolutional neural network layers to the visual cortex. In *SVRHM 2021 Workshop @ NeurIPS*.

Edelman, S. (1995). Representation, similarity, and the chorus of prototypes. *Minds and Machines*, 5(1):45–68. doi: 10.1007/BF00974189. URL https://doi.org/10.1007/BF00974189.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194. doi: 10.1016/j.neuroimage.2016.10.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811916305481.

Engemann, D. A. & Gramfort, A. (2015). Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage*, 108:328–342.

Erhan, D., Bengio, Y., Courville, A., Vincent, P., & Box, P. O. (2009). Visualizing Higher-Layer Features of a Deep Network. *University of Montreal*, 1341.

Etard, O. & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience*, 39(29):5750–5759.

Fabius, J. H., Fracasso, A., Nijboer, T. C. W., & Stigchel, S. V. d. (2019). Time course of spatiotopic updating across saccades. *Proceedings of the National Academy of Sciences*, 116(6):2027–2032. doi: 10.1073/pnas.1812210116. URL https://www.pnas.org/content/116/6/2027.

Faes, L., Nollo, G., & Porta, A. (2011). Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. *Physical Review E*, 83(5):051112.

Faes, L., Nollo, G., Stramaglia, S., & Marinazzo, D. (2017). Multiscale granger causality. *Physical Review E*, 96(4):042150.

Faharibozorg, S.-R., Henson, R. N., & Hauk, O. (2018). Adaptive cortical parcellations for source reconstructed EEG/MEG connectomes. *NeuroImage*, 169:23–45.

Feather, J., Durango, A., Gonzalez, R., & McDermott, J. (2019). Metamers of neural networks reveal divergence from human perceptual systems. In *NeurIPS*, pages 10078–10089.

Feldman, R. M. & Goldstein, R. (1967). Averaged evoked responses to synthetic syntax sentences (3s). *Journal of speech and hearing research*, 10(4):689–696.

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., & Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of Neural Engineering*, 14.

Fiedler, L., Wöstmann, M., Herbst, S., & Obleser, J. (2018). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *bioRxiv*, http://dx.doi.org/10.1101/238642.

Florin, E., Gross, J., Pfeifer, J., Fink, G. R., & Timmermann, L. (2010). The effect of filtering on Granger causality based multivariate causality measures. *Neuroimage*, 50(2):577–588.

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife*, 6(e27203).

Franz, M. O. & Schölkopf, B. (2005). Implicit Wiener series for higher-order image analysis. In *Advances in neural information processing systems*, pages 465–472.

Frégnac, Y. (2017). Big data and the industrialization of neuroscience: A safe roadmap for understanding the brain? *Science*, 358(6362):470–477.

Friederici, A. D. (2009). Pathways to language: fiber tracts in the human brain. *Trends in cognitive sciences*, 13(4):175–181.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive brain research*, 1(3):183–192.

Fries, P. (2015). Rhythms for cognition: communication through coherence. *Neuron*, 88(1):220–235.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

Friston, K., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as Hypotheses: Saccades as Experiments. *Frontiers in Psychology*, 3. doi: 10.3389/fpsyg.2012.00151. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00151/full.

Friston, K. J. (2009). Modalities, modes, and models in functional neuroimaging. *Science*, 326(5951):399–403.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202. doi: 10.1007/BF00344251. URL http://link.springer.com/10.1007/BF00344251.

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S. A., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3). doi: 10.1167/jov.21.3.16. URL https://jov.arvojournals.org/article.aspx?articleid=2772393. Publisher: The Association for Research in Vision and Ophthalmology.

Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., Kim, K., Wang, E., Mrowca, D., Lingelbach,

M., Curtis, A., Feigelis, K., Bear, D. M., Gutfreund, D., Cox, D., DiCarlo, J. J., McDermott, J., Tenenbaum, J. B., & Yamins, D. L. K. (2020). ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation. *arXiv:2007.04954 [cs]*. URL http://arxiv.org/abs/2007.04954.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature neuroscience*, 2(6):568–573.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. *arXiv:2004.07780 [cs, q-bio]*. URL http://arxiv.org/abs/2004.07780.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231 [cs, q-bio, stat]*. URL http://arxiv.org/abs/1811.12231.

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigne, A., Lalor, E., Meyer, B. T., Miran, S., Francart, T., & Bertrand, A. (2021). Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, 38(4):89–102.

Gershman, S. J. (2021). Just looking: The innocent eye in neuroscience. *Neuron*.

Gerster, M., Waterstraat, G., Litvak, V., Lehnertz, K., Schnitzler, A., Florin, E., Curio, G., & Nikulin, V. (2021). Separating neural oscillations from aperiodic 1/f activity: challenges and recommendations. *bioRxiv*.

Ghitza, O. (2013). The theta-syllable: a unit of speech information defined by cortical function. *frontiers in Psychology*, 4(138).

Giordano, B. L., Ince, R. A. A., Gross, J., Schyns, P. G., Panzeri, S., & Kayser, C. (2016). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. *eLife*, 6:e24763(DOI: 10.7554/eLife.24763).

Giraud, A. L. & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, 15:511–517.

Goetschalckx, L., Andonian, A., & Wagemans, J. (2021). Generative adversarial networks unlock new methods for cognitive science. *Trends in Cognitive Sciences*, 25(9):788–801. doi: 10.1016/j.tics.2021.06.006.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337. doi: 10.1073/pnas.1912334117. URL https://www.pnas.org/content/117/47/29330.

Gosselin, F. & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological science*, 14(5):505–509.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Graves, A. (2020). alecGraves/BVAE-tf.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2014). Visual Noise from Natural Scene Statistics Reveals Human Scene Category Representations. *arXiv:1411.5331 [cs]*. URL http://arxiv.org/abs/1411.5331.

Grenander, U. (1994). *General Pattern Theory: A Mathematical Study of Regular Structures*. Oxford Mathematical Monographs. Oxford University Press, Oxford, New York.

Griffiths, D. W., Blunden, A. G., & Little, D. R. (2017). 12 - Logical-Rule Based Models of Categorization: Using Systems Factorial Technology to Understand Feature and Dimensional Processing. In Little, D. R., Altieri, N., Fifić, M., & Yang, C.-T., editors, *Systems Factorial Technology*. Academic Press, San Diego.

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature neuroscience*, 7(5):555–562.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P. G., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, 11(12).

Gross, J., Kluger, D. S., Abbasi, O., Chalas, N., Steingräber, N., Daube, C., & Schoffelen, J.-M. (2021). Comparison of undirected frequency-domain connectivity measures for cerebro-peripheral analysis. *NeuroImage*, 245:118660.

Gross, J., Kujala, J., Hämäläinen, M., Timmermann, L., Schnitzler, A., & Salmelin, R. (2001). Dynamic imaging of coherent sources: studying neural interactions in the human brain. *Proceedings of the National Academy of Sciences*, 98(2):694–699.

Gross, J., Loannides, A. A., Dammers, J., Maeß, B., Friederici, A. D., & Müller-Gärtner, H.-W. (1998). Magnetic field tomography analysis of continuous speech. *Brain Topography*, 10(4):273–281.

Güçlü, Umut, U. & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5023-14.2015.

Gwilliams, L. & King, J.-R. (2020). Recurrent processes support a cascade of hierarchical decisions. *eLife*, 9:e56603.

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2020). Neural dynamics of phoneme sequences: Position-invariant code for content and order. *bioRxiv*, pages 2020–04.

Hahn, T., Emden, D., Grotegerd, D., Kaehler, C., Leenings, R., & Winter, N. (2018). https://www.photon-ai.com/ A Python-based Hyperparameter Optimization Toolbox for Neural Networks designed to accelerate and simplify the construction, training, and evaluation of machine learning models.

Hambrook, D. A. & Tata, M. S. (2014). Theta-band phase tracking in the two-talker problem. *Brain & Language*, 135:52–56.

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, 28:1–12.

Hamilton, L. S. & Huth, A. G. (2018). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*.

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell*, 184(18):4626–4639.

Harder, M., Salge, C., & Polani, D. (2013). Bivariate measure of redundant information. *Physical Review E*, 87(1):012130.

Hasson, U., Egidi, G., M, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, 180:135–157.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110. doi: 10.1016/j.neuroimage.2013.10.067. URL https://www.sciencedirect.com/science/article/pii/S1053811913010914.

Haugh, L. D. (1976). Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach. *Journal of the American Statistical Association*, 71(354):378–385.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*. URL http://arxiv.org/abs/1512.03385.

Hebart, M. N. & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, 180:4–18.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2021). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*, pages 2020–12.

Henry, M. J. & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences*, 109(49):20095–20100.

Hermann, K. L. & Kornblith, S. (2019). Exploring the Origins and Prevalence of Texture Bias in Convolutional Neural Networks. *arXiv:1911.09071 [cs, q-bio]*. URL http://arxiv.org/abs/1911.09071.

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., & Ackermann, H. (2012). Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology*, 49:322–334.

Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2016). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR*. URL https://openreview.net/forum?id=Sy2fzU9gl.

Hoel, E. (2021). The overfitted brain: Dreams evolved to assist generalization. *Patterns*, 2(5):100244. doi: 10.1016/j.patter.2021.100244. URL https://www.sciencedirect.com/science/article/pii/S2666389921000647.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and Decoding Models in Cognitive Electrophysiology. *frontiers in Systems Neuroscience*, 11(61).

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., Rubin, N., Heeger, D. J., & Hasson, U. (2012). Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron*, 76:423–434.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.

Hyafil, A., Fontolan, L., Kabdebon, C., Boris, G., & Giraud, A. L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *eLife*, 4(e06213).

Ince, R. (2017a). Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy*, 19(7):318. doi: 10.3390/e19070318. URL http://www.mdpi.com/1099-4300/19/7/318.

Ince, R. A., Van Rijsbergen, N. J., Thut, G., Rousselet, G. A., Gross, J., Panzeri, S., & Schyns, P. G. (2015). Tracing the flow of perceptual features in an algorithmic brain network. *Scientific reports*, 5(1):1–17.

Ince, R. A. A. (2017b). The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv*, 1702.01591.

Ince, R. A. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2017). A Statistical Framework for Neuroimaging Data Analysis Based on Mutual Information Estimated via a Gaussian Copula. *Human Brain Mapping*, 38(3):1541–1573.

Ince, R. A. A., Jaworska, K., Gross, J., Panzeri, S., van Rijsbergen, N. J., Rousselet, G. A., & Schyns, P. G. (2016). The Deceptively Simple N170 Reflects Network Information Processing Mechanisms Involving Visual Feature Coding and Transfer Across Hemispheres. *Cerebral Cortex*, 26(11):4123–4135. doi: 10.1093/cercor/bhw196. URL https://academic.oup.com/cercor/article/26/11/4123/2374065.

Ince, R. A. A., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of population prevalence. *eLife*, 10(e62461). doi: 10.7554/eLife.62461. URL https://elifesciences.org/articles/62461.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244.

Jack, R. E. & Schyns, P. G. (2017). Toward a Social Psychophysics of Face Communication. *Annual Review of Psychology*, 68(1):269–297.

doi: 10.1146/annurev-psych-010416-044242. URL https://doi.org/10.1146/annurev-psych-010416-044242.

Jacobsen, J.-H., Behrmann, J., Zemel, R., & Bethge, M. (2019). Excessive Invariance Causes Adversarial Vulnerability. *arXiv:1811.00401 [cs, stat]*. URL http://arxiv.org/abs/1811.00401.

Jain, S., Mahto, S., Turek, J. S., Vo, V. A., LeBel, A., & Huth, A. G. (2021). Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. *bioRxiv*, pages 2020–10.

James, R. G., Barnett, N., & Crutchfield, J. P. (2016). Information flows? A critique of transfer entropies. *Physical review letters*, 116(23):238701.

James, R. G., Emenheiser, J., & Crutchfield, J. P. (2018). Unique information via dependency constraints. *Journal of Physics A: Mathematical and Theoretical*, 52(1):014002.

James, R. G., Emenheiser, J., & Crutchfield, J. P. (2019). Unique information and secret key agreement. *Entropy*, 21(1):12.

Jiang, H., Bahramisharif, A., van Gerven, M. A., & Jensen, O. (2015). Measuring directionality between neuronal oscillations of different frequencies. *NeuroImage*, 118:359–367.

Jonas, E. & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLoS computational biology*, 13(1):e1005268.

Jones, J. P. & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1187–1211.

Jones, M. R. & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological review*, 96(3):459.

Jozwik, K. M., O'Keeffe, J., Storrs, K. R., & Kriegeskorte, N. (2021). Face dissimilarity judgements are predicted by representational distance in deep neural networks and principal-component face space. *bioRxiv*, page 2021.04.09.438859. doi: 10.1101/2021.04.09.438859. URL https://www.biorxiv.org/content/10.1101/2021.04.09.438859v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN. *arXiv:1912.04958 [cs, eess, stat]*. URL http://arxiv.org/abs/1912.04958.

Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, 180:101–109.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.

Kayser, C., Wilson, C., Safaai, H., Sakata, S., & Panzeri, S. (2015). Rhythmic auditory cortex activity at multiple timescales shapes stimulus–response gain and background firing. *Journal of Neuroscience*, 35(20):7750–7762.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, 16(e2004473).

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16. doi: 10.1016/j.neuron.2018.03.044. URL https://www.cell.com/neuron/abstract/S0896-6273(18)30250-2. Publisher: Elsevier.

Keshishian, M., Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2020). Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *eLife*, 9:e53445. doi: 10.7554/eLife.53445. URL https://doi.org/10.7554/eLife.53445. Publisher: eLife Sciences Publications, Ltd.

Khalighinejad, B., da Silva, G. C., & Mesgarani, N. (2017). Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *The Journal of Neuroscience*, 37(8):2176–2185.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863. doi: 10.1073/pnas.1905544116. URL https://www.pnas.org/content/116/43/21854.

Kingma, D. P. & Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. URL http://arxiv.org/abs/1412.6980.

Kingma, D. P. & Welling, M. (2014). Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*. URL http://arxiv.org/abs/1312.6114.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(2-3):169–195.

Koskinen, M., Kurimo, M., Gross, J., Hyvärinen, A., & Hari, R. (2020). Brain activity reflects the predictability of word sequences in listened continuous speech. *NeuroImage*, 219:116936.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, 93(3):480–490. doi: 10.1016/j.neuron.2016.12.041. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627316310406.

Kriegeskorte, N. & Douglas, P. K. (2018a). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160. URL http://www.nature.com/articles/s41593-018-0210-5.

Kriegeskorte, N. & Douglas, P. K. (2018b). Interpreting Encoding and Decoding Models. *arXiv*, 1812.00278.

Kubilius, J. (2018). Predict, then simplify. *NeuroImage*, 180:110–111. doi: 10.1016/j.neuroimage.2017.12.006. URL http://www.sciencedirect.com/science/article/pii/S1053811917310212.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLOS Computational Biology*, 12(4):e1004896. doi: 10.1371/journal.pcbi.1004896. URL http://dx.plos.org/10.1371/journal.pcbi.1004896.

Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. *arXiv:1909.06161 [cs, eess, q-bio]*. URL http://arxiv.org/abs/1909.06161.

Kulasingham, J. P., Brodbeck, C., Presacco, A., Kuchinsky, S. E., Anderson, S., & Simon, J. Z. (2020). High gamma cortical processing of continuous speech in younger and older listeners. *Neuroimage*, 222:117291.

Kumar, K., Kanwoo, C. J., & Stern, R. M. (2011). Delta-spectral-cepstral coefficients for robust speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic. IEEE.

Lakatos, P., Gross, J., & Thut, G. (2019). A new unifying account of the roles of neuronal entrainment. *Current Biology*, 29(18):R890–R905.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *science*, 320(5872):110–113.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, 94(3):1904–1911.

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., et al. (2021). Generative spoken language modeling from raw audio. *arXiv preprint arXiv:2102.01192*.

Lalor, E. C. & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European journal of neuroscience*, 31(1):189–193.

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of neurophysiology*, 102(1):349–359.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096. doi: 10.1038/s41467-019-08987-4. URL http://www.nature.com/articles/s41467-019-08987-4.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539. Number: 7553 Publisher: Nature Publishing Group.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *The Journal of Neuroscience*, 31(8):2906–2915.

Lescroart, M. D. & Gallant, J. L. (2019). Human Scene-Selective Areas Represent 3D Configurations of Surfaces. *Neuron*, 101(1):178–192.e7. doi: 10.1016/j.neuron.2018.11.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627318309954.

Lobier, M., Siebenhühner, F., Palva, S., & Palva, J. M. (2014). Phase transfer entropy: a novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. *Neuroimage*, 85:853–872.

Lotto, A. J. & Holt, L. L. (2000). The illusion of the phoneme. In Billings, S. J., editor, *The Panels*, volume 35, pages 191–204. Chicago Linguistic Society.

Luo, H. & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–1010.

Maddox, R. K. & Lee, A. K. C. (2018). Auditory Brainstem Responses to Continuous Natural Speech in Human Listeners. *eNeuro*, 5(1). doi: 10.1523/ENEURO.0441-17.2018. URL https://www.eneuro.org/content/5/1/ENEURO.0441-17.2018. Publisher: Society for Neuroscience Section: Methods/New Tools.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science*, 295(5555):690–694.

Mangini, M. C. & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, 28(2):209–226.

Mao, C., Gupta, A., Nitin, V., Ray, B., Song, S., Yang, J., & Vondrick, C. (2020). Multitask Learning Strengthens Adversarial Robustness. *arXiv:2007.07236 [cs]*. URL http://arxiv.org/abs/2007.07236.

Marmarelis, P. Z. & Naka, K.-I. (1972). White-noise analysis of a neuron chain: an application of the Wiener theory. *Science*, 175(4027):1276–1278.

Marr, D. (2010). *Vision: a computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge, Mass.

Massaro, D. W. (1974). Perceptual Units in Speech Recognition. *Journal of Experimental Psychlogy*, 102(2):199–208.

Mathieu, E., Rainforth, T., Siddharth, N., & Teh, Y. W. (2019). Disentangling Disentanglement in Variational Autoencoders. *arXiv:1812.02833 [cs, stat]*. URL http://arxiv.org/abs/1812.02833.

McCallum, W., Farmer, S., & Pocock, P. (1984). The effects of physical and semantic incongruites on auditory event-related potentials. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 59(6):477–488.

McDermott, J. H. (2013). Audition. In Ochsner, K. N. & Kosslyn, S., editors, *The Oxford Handbook of Cognitive Neuroscience, Volume 1: Core Topics*, volume 1. Oxford University Press.

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19:97–116.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2):254–278. doi: 10.1037/0033-295X.100.2.254.

Mehler, D. M. A. & Kording, K. P. (2018). The lure of causal statements: Rampant misinference of causality in estimated connectivity. *arXiv e-prints*, pages arXiv–1812.

Mell, M. M., St-Yves, G., & Naselaris, T. (2021). Voxel-to-voxel predictive models reveal unexpected structure in unexplained variance. *NeuroImage*, page 118266.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 343:1006–1010.

Michalareas, G., Vezoli, J., Van Pelt, S., Schoffelen, J.-M., Kennedy, H., & Fries, P. (2016). Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. *Neuron*, 89(2):384–397.

Millet, J. & King, J.-R. (2021). Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *arXiv preprint arXiv:2103.01032*.

Mirkovic, B., Debener, S., Jaeger, M., & De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *Journal of Neural Engineering*, 12.

Młynarski, W. & McDermott, J. H. (2018). Learning Midlevel Auditory Codes from Natural Sound Statistics. *Neural Computation*, 30(3):631–669.

Młynarski, W. F. & Hermundstad, A. M. (2018). Adaptive coding for dynamic sensory inference. *Elife*, 7:e32055.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15. doi: 10.1016/j.dsp.2017.10.011. URL https://linkinghub.elsevier.com/retrieve/pii/S1051200417302385.

Morgan, C. L. (2018). *An Introduction to Comparative Psychology*. Forgotten Books, Place of publication not identified.

Morillon, B. & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences*, 114(42):E8913–E8921.

Morillon, B., Hackett, T. A., Kajikawa, Y., & Schroeder, C. E. (2015). Predictive motor control of sensory dynamics in auditory active sensing. *Current Opinion in Neurobiology*, 31:230–238.

Mosier, C. I. (1951). I. Problems and designs of cross-validation 1. *Educational and Psychological Measurement*, 11(1):5–11.

Murray, R. F. (2011). Classification images: A review. *Journal of Vision*, 11(5):2–2. doi: 10.1167/11.5.2. URL https://jov.arvojournals.org/article.aspx?articleid=2191849.

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22(10):1677–1686.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385(6615):432–434.

Näätänen, R. & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410. doi: 10.1016/j.neuroimage.2010.07.073. URL http://www.sciencedirect.com/science/article/pii/S1053811910010657.

Nestor, A., Lee, A. C. H., Plaut, D. C., & Behrmann, M. (2020). The Face of Image Reconstruction: Progress, Pitfalls, Prospects. *Trends in Cognitive Sciences*, 24(9):747–759. doi: 10.1016/j.tics.2020.06.006. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(20)30147-9.

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M. P., Poldrack, R. A., Poline, J.-B., Proal, E., Thirion, B., Van Essen, D. C., White, T., & Yeo, B. T. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, 20(3):299–303.

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*.

Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. *Physics in Medicine & Biology*, 48(22):3637–3652.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–430.

Norman-Haignere, S. V. & McDermott, J. H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biology*, 16(e2005127).

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, 197:482–492.

Obleser, J. & Eisner, F. (2008). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, 13(14–19).

Obleser, J., Elbert, T., Lahiri, A., & Eulitz, C. (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cognitive Brain Research*, 15(3):207–213.

Obleser, J. & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in cognitive sciences*, 23(11):913–926.

Obleser, J., Zimmermann, J., Van Meter, J., & Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cerebral Cortex*, 17(10):2251–2257.

Oganian, Y. & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science advances*, 5(11):eaay6279.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, 2(11):e7. doi: 10.23915/distill.00007. URL https://distill.pub/2017/feature-visualization.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3):e10. doi: 10.23915/distill.00010. URL https://distill.pub/2018/building-blocks.

Olman, C. & Kersten, D. (2004). Classification objects, ideal observers & generative models. *Cognitive Science*, 28(2):227–239.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 156869.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral cortex*, 25(7):1697–1706.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT press.

Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., & Fellin, T. (2017). Cracking the Neural Code for Sensory Perception by Combining Statistics, Intervention and Behaviour. *Neuron*, 93:491–507.

Panzeri, S., Macke, J. H., Gross, J., & Kayser, C. (2015). Neural population coding: combining insights from microscopic and mass signals. *Trends in Cognitive Sciences*, 19:162–172.

Park, H., Ince, R. A., Schyns, P. G., Thut, G., & Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Current Biology*, 25(12):1649–1653.

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2018). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS Biology*, 16(8):e2006558.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., & Chang, E. F. (2012). Reconstructing Speech from Human Auditory Cortex. *PLoS Biology*, 10(1).

Peelle, J. E. & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in psychology*, 3:320.

Peterson, J., Uddenberg, S., Griffiths, T., Todorov, A., & Suchow, J. (2021a). Capturing and modifying the perceived traits of all possible faces. *PsyArxiv*. doi: 10.31234/osf.io/brzfy. URL https://psyarxiv.com/brzfy/. type: article.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021b). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214.

Petkov, C. I., Kikuchi, Y., Milne, A. E., Mishkin, M., Rauschecker, J. P., & Logothetis, N. K. (2015). Different forms of effective connectivity in primate frontotemporal pathways. *Nature Communications*, 6:6000.

Pinzuti, E., Wollstadt, P., Gutknecht, A., Tüscher, O., & Wibral, M. (2020). Measuring spectrally-resolved information transfer. *PLoS Computational Biology*, 16(12):e1008526.

Pisoni, D. B. & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, 25:21–52.

Poeppel, D. & Adolfi, F. (2020). Against the epistemological primacy of the hardware: The brain from inside out, turned upside down. *Eneuro*, 7(4).

Poeppel, D. & Embick, D. (2005). Defining the relation between linguistics and neuroscience. In Cutler, A., editor, *Twenty-first century psycholinguistics: Four cornerstones*, pages 103–120. Hillside, NJ: Lawrence Erlbaum.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009.e10. doi: 10.1016/j.cell.2019.04.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419303915.

Prendergast, G., Johnson, S. R., & Green, G. G. R. (2010). Temporal dynamics of sinusoidal and non-sinusoidal amplitude modulation. *European Journal of Neuroscience*, 32:1599–1607.

Putnam, H. (1967). Psychological predicates. *Art, mind, and religion*, 1:37–48.

Qiu, A., Schreiner, C. E., & Escabi, M. A. (2003). Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. *Journal of Neurophysiology*, 90:456–476.

Quiroga, R. Q. & Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, 10(3):173–185.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ragwitz, M. & Kantz, H. (2002). Markov models from data by simple nonlinear time series predictors in delay embedding spaces. *Physical Review E*, 65(5):056201.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience*, 38(33):7255–7269. doi: 10.1523/JNEUROSCI.0388-18.2018. URL http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0388-18.2018.

Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171:130–150.

Rauschecker, J. P. & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6):718–724.

Razavi, A., Oord, A. v. d., & Vinyals, O. (2019). Generating Diverse High-Fidelity Images with VQ-VAE-2. *arXiv:1906.00446 [cs, stat]*. URL http://arxiv.org/abs/1906.00446.

Rieke, F., Bodnar, D., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 262(1365):259–265.

Ringach, D. & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, 28(2):147–166.

Ritter, W., Vaughan Jr, H. G., & Costa, L. D. (1968). Orienting and habituation to auditory stimuli: a study of short terms changes in average evoked responses. *Electroencephalography and clinical Neurophysiology*, 25(6):550–556.

Roth, W. T., Kopell, B. S., & Bertozzi, P. E. (1970). The effect of attention on the average evoked response to speech sounds. *Electroencephalography and clinical Neurophysiology*, 29(1):38–46.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*, 10(1).

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., & Formisano, E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 114(18):4799–8804.

Sarikaya, R., Crook, P. A., Marin, A., Jeong, M., Robichaud, J. P., Celikyilmaz, A., Kim, Y. B., Rochette, A., Z, K. O., & Liu, X. e. a. (2016). An overview of end-to-end language understanding and dialog management for personal digital assistants. *IEEE Spoken Language Techonology Workshop (SLT)*, pages 391–397.

Sassenhagen, J. (2018). How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression. *Language, Cognition and Neuroscience*, 34:474–490.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., et al. (2008). Ventral and dorsal pathways for language. *Proceedings of the national academy of Sciences*, 105(46):18035–18040.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67. doi: 10.1038/s41583-020-00395-8. URL https://www.nature.com/articles/s41583-020-00395-8. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Learning algorithms;Network models Subject_term_id: learning-algorithms;network-models.

Sayers, B. M., Beagley, H. A., & Henshall, W. R. (1974). The mechanism of auditory evoked EEG responses. *Nature*, 247(5441):481–483.

Schädler, M.-R., Meyer, B. T., & Kollmeier, B. (2012). Spectro-temporal modulation subspace-spanning filterbank features for robust automatic speech recognition. *Journal of the Acoustical Society of America*, 131:4134–4151.

Schmitt, L.-M., Erb, J., Tune, S., Rysop, A., Hartwigsen, G., & Obleser, J. (2021). Predicting speech from a cortical hierarchy of event-based time scales. *Science Advances*, 7(49).

Schnitzler, A. & Gross, J. (2005). Normal and pathological oscillatory communication in the brain. *Nature reviews neuroscience*, 6(4):285–296.

Schoffelen, J.-M., Hultén, A., Lam, N., Marquand, A. F., Uddén, J., & Hagoort, P. (2017). Frequency-specific directed interactions in the human brain network for language. *Proceedings of the National Academy of Sciences*, 114(30):8083–8088.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward Causal Representation Learning. *Proceedings of the IEEE*,

109(5):612–634. doi: 10.1109/JPROC.2021.3058954. Conference Name: Proceedings of the IEEE.

Scholte, H. S., Losch, M. M., Ramakrishnan, K., de Haan, E. H., & Bohte, S. M. (2018). Visual pathways from the perspective of cost functions and multi-task deep neural networks. *Cortex*, 98:249–261. doi: 10.1016/j.cortex.2017.09.019. URL https://linkinghub.elsevier.com/retrieve/pii/S0010945217303258.

Schönwiesner, M. & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14611–14616).

Schott, L., Rauber, J., Bethge, M., & Brendel, W. (2018). Towards the first adversarially robust neural network model on MNIST. *arXiv:1805.09190 [cs]*. URL http://arxiv.org/abs/1805.09190.

Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, 85(2):461.

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3):413–423. doi: 10.1016/j.neuron.2020.07.040. URL https://linkinghub.elsevier.com/retrieve/pii/S089662732030605X.

Schroeder, C. E. & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in neurosciences*, 32(1):9–18.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823. doi: 10.1109/CVPR.2015.7298682. URL http://arxiv.org/abs/1503.03832.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, 21(1):1–17. doi: 10.1017/S0140525X98000107. URL https://www.cambridge.org/core/product/identifier/S0140525X98000107/type/journal_article.

Schyns, P. G., Gosselin, F., & Smith, M. L. (2009). Information processing algorithms in the brain. *Trends in Cognitive Sciences*, 13(1):20–26. doi: 10.1016/j.tics.2008.09.008.

Schyns, P. G., Jentzsch, I., Johnson, M., Schweinberger, S. R., & Gosselin, F. (2003). A principled method for determining the functionality of brain responses. *Neuroreport*, 14(13):1665–1669. doi: 10.1097/00001756-200309150-00002.

Schyns, P. G. & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(3):681–696. doi: 10.1037/0278-7393.23.3.681. Place: US Publisher: American Psychological Association.

Schyns, P. G., Zhan, J., Jack, R. E., & Ince, R. A. A. (2020). Revealing the information contents of memory within the stimulus information representation framework. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1799):20190705. doi: 10.1098/rstb.2019.0705. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2019.0705.

Sejnowski, T. J., Churchland, P. S., & Movshon, J. A. (2014). Putting big data to good use in neuroscience. *Nature neuroscience*, 17(11):1440–1441.

Sergent, J., Ohta, S., & Macdonald, B. (1992). Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain*, 115(1):15–36.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]*. URL http://arxiv.org/abs/1312.6034.

Sitek, K. R., Gulban, O. F., Calabrese, E., Johnson, G. A., Gosh, S. S., & Martino, F. D. (2019). Mapping the human subcortical auditory system using histology, post mortem MRI and in vivo MRI at 7T. *bioRxiv*, 568139.

Smith, E. E. & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, 22(4):377–386. doi: 10.3758/BF03200864. URL https://doi.org/10.3758/BF03200864.

Smith, L. N. (2017). Cyclical Learning Rates for Training Neural Networks. *arXiv:1506.01186 [cs]*. URL http://arxiv.org/abs/1506.01186.

Smith, M. L., Gosselin, F., & Schyns, P. G. (2012). Measuring Internal Representations from Behavioral and Brain Data. *Current Biology*, 22(3):191–196. doi: 10.1016/j.cub.2011.11.061. URL http://www.sciencedirect.com/science/article/pii/S0960982211013947.

Sohoglu, E. & Davis, M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *Elife*, 9:e58077.

Stan Development Team (2020). RStan: the R interface to Stan.

Standley, T., Zamir, A. R., Chen, D., Guibas, L., Malik, J., & Savarese, S. (2020). Which Tasks Should Be Learned Together in Multi-task Learning? *arXiv:1905.07553 [cs]*. URL http://arxiv.org/abs/1905.07553.

Suchow, J. W., Peterson, J. C., & Griffiths, T. L. (2018). Learning a face space for experiments on human identity. *arXiv:1805.07653 [cs]*. URL http://arxiv.org/abs/1805.07653.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*. URL http://arxiv.org/abs/1312.6199.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer.

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior research methods*, 52(6):2372–2382.

Theunissen, F. E. & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, 15:355–366.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, 20(6):2315–2331.

Thoret, E., Andrillon, T., Léger, D., & Pressnitzer, D. (2021). Probing machine-learning classifiers using noise, bubbles, and reverse correlation. *Journal of Neuroscience Methods*, 362:109297. doi: 10.1016/j.jneumeth.2021.109297. URL https://www.sciencedirect.com/science/article/pii/S0165027021002326.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv*, 0004057.

Todorov, A., Uddenberg, S., Peterson, J., Griffiths, T., & Suchow, J. (2020). Data-Driven, Photorealistic Social Face-Trait Encoding, Prediction, and Manipulation Using Deep Neural Networks. *Princeton University*. URL https://collaborate.princeton.edu/en/publications/data-driven-photorealistic-social-face-trait-encoding-prediction-.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25.

Vahdat, A. & Kautz, J. (2020). NVAE: A Deep Hierarchical Variational Autoencoder. *arXiv:2007.03898 [cs, stat]*. URL http://arxiv.org/abs/2007.03898.

van den Oord, A., Li, Y., & Vinyals, O. (2018a). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2018b). Neural Discrete Representation Learning. *arXiv:1711.00937 [cs]*. URL http://arxiv.org/abs/1711.00937.

Van der Maaten, L. v. d. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Van Veen, B. D., van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of Brain Electrical Activity via Linearly Constrained Minimum Variance Spatial Filtering. *IEEE Transactions on Biomedical Engineering*, 44(9):867–880.

van Vliet, M. & Salmelin, R. (2020). Post-hoc modification of linear models: Combining machine learning with domain information to make solid inferences from noisy data. *NeuroImage*, 204:116221. doi: 10.1016/j.neuroimage.2019.116221.

VanRullen, R. & Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology*, 2(1):193. doi: 10.1038/s42003-019-0438-y. URL http://www.nature.com/articles/s42003-019-0438-y.

Varoquaux, G., Raamana, P. R., Engeman, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats and guidelines. *NeuroImage*, 145:166–179.

Vaughan Jr, H. G. & Ritter, W. (1970). The sources of auditory evoked responses recorded from the human scalp. *Electroencephalography and clinical neurophysiology*, 28(4):360–367.

Veit, M. J., Kucyi, A., Hu, W., Zhang, C., Zhao, B., Guo, Z., Yang, B., Sava-Segal, C., Perry, C., Zhang, J., et al. (2021). Temporal order of signal propagation within and across intrinsic brain networks. *Proceedings of the National Academy of Sciences*, 118(48).

Verhulst, S., Altoè, A., & Vasilikov, V. (2018). Computational modeling of the human auditory periphery: Auditory nerve responses, evoked potentials and hearing loss. *Hearing Research*, 360:55–75.

Vlachos, I. & Kugiumtzis, D. (2010). Nonuniform state-space reconstruction and coupling detection. *Physical Review E*, 82(1):016207.

Wang, X.-J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological reviews*, 90(3):1195–1268.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59.

Wibral, M., Lizier, J. T., & Priesemann, V. (2015). Bits from brains for biologically inspired computing. *frontiers in Robotics and AI*, 2(5).

Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., Lizier, J. T., & Vicente, R. (2013). Measuring information-transfer delays. *PloS one*, 8(2):e55809.

Wiener, N. (1958). *Nonlinear problems in random theory*. The MIT Press.

Williams, M. A., Dang, S., & Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nature Neuroscience*, 10:685–686.

Williams, P. L. & Beer, R. D. (2010). Nonnegative Decomposition of Multivariate Information. *arXiv*, 1004.2515.

Wollstadt, P., Sellers, K. K., Rudelt, L., Priesemann, V., Hutt, A., Fröhlich, F., & Wibral, M. (2017). Breakdown of local information processing may underlie isoflurane anesthesia effects. *PLoS computational biology*, 13(6):e1005511.

Wood, C. C., Goff, W. R., & Day, R. S. (1971). Auditory evoked potentials during speech perception. *Science*, 173(4003):1248–1251.

Woolrich, M., Hunt, L., Groves, A., & Barnes, G. (2011). MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. *NeuroImage*, 57:1466–1479.

Wöstmann, M., Fiedler, L., & Obleser, J. (2017). Tracking the signal, cracking the code: Speech and speech comprehension in non-invasive human electrophysiology. *Language, Cognition and Neuroscience*, 32(7):855–869.

Xu, T., Zhan, J., Garrod, O. G. B., Torr, P. H. S., Zhu, S.-C., Ince, R. A. A., & Schyns, P. G. (2018). Deeper Interpretability of Deep Networks. *arXiv:1811.07807 [cs]*. URL http://arxiv.org/abs/1811.07807.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624. doi: 10.1073/pnas.1403112111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1403112111.

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, 6(10):eaax5979. doi: 10.1126/sciadv.aax5979. URL https://advances.sciencemag.org/content/6/10/eaax5979.

Yuan, J. & Liberman, A. M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(3878).

Yuille, A. & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308. doi: 10.1016/j.tics.2006.05.002. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661306001264.

Zan, P., Presacco, A., Anderson, S., & Simon, J. Z. (2020). Exaggerated cortical representation of speech in older listeners: mutual information analysis. *Journal of Neurophysiology*, 124(4):1152–1164.

Zeiler, M. D. & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]*. URL http://arxiv.org/abs/1311.2901.

Zhan, J., Garrod, O. G. B., van Rijsbergen, N., & Schyns, P. G. (2019a). Modelling face memory reveals task-generalizable representations. *Nature Human Behaviour*, 3(8):817–826. doi: 10.1038/s41562-019-0625-3. URL http://www.nature.com/articles/s41562-019-0625-3.

Zhan, J., Ince, R. A., van Rijsbergen, N., & Schyns, P. G. (2019b). Dynamic Construction of Reduced Representations in the Brain for Perceptual Decision Behavior. *Current Biology*, 29(2):319–326.e4. doi: 10.1016/j.cub.2018.11.049. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982218315483.

Zhang, S., Huang, J.-B., Lim, J., Gong, Y., Wang, J., Ahuja, N., & Yang, M.-H. (2017). Tracking Persons-of-Interest via Unsupervised Representation Adaptation. *arXiv:1710.02139 [cs]*. URL http://arxiv.org/abs/1710.02139.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Learning Deep Features for Discriminative Localization. *arXiv:1512.04150 [cs]*. URL http://arxiv.org/abs/1512.04150.

Zhu, Z., Luo, P., Wang, X., & Tang, X. (2013). Deep Learning Identity-Preserving Face Space. In *2013 IEEE International Conference on Computer Vision*, pages 113–120, Sydney, Australia. IEEE.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3). doi: 10.1073/pnas.2014196118. URL https://www.pnas.org/content/118/3/e2014196118.

Statement of Originality

University of Glasgow
College of Science and Engineering

Name: Christoph Daube
Registration Number:

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Statement of conjoint work
I confirm that Chapter 2 was jointly authored with Robin A. A. Ince and Joachim Gross and I contributed $> 50\%$ of this work.
I confirm that Chapter 3 was jointly authored with Joachim Gross and Robin A. A. Ince and I contributed $> 50\%$ of this work.
I confirm that Chapter 4 was jointly authored with Tian Xu, Jiayu Zhan, Andrew Webb, Robin A. A. Ince, Oliver G. B.Garrod and Philippe G. Schyns and I contributed $> 50\%$ of this work.

Name: Christoph Daube
Signature:
Date: 12.01.2022