Is it that simple? Linear mapping models in cognitive neuroscience

Anna A. Ivanova,[1,2] Martin Schrimpf,[1,2,5] Stefano Anzellotti,[3] Noga Zaslavsky,[1,5] Evelina Fedorenko,[1,2] and Leyla Isik[4]

[1]Department of Brain and Cognitive Sciences, MIT
[2]McGovern Institute for Brain Research, MIT
[3]Department of Psychology, Boston College
[4]Department of Cognitive Science, Johns Hopkins University
[5]Center for Brains, Minds and Machines, MIT

Abstract

Advances in cognitive neuroscience are often accompanied by an increased complexity in the methods we use to uncover new aspects of brain function. Recently, many studies have started to use large feature sets to predict and interpret brain activity patterns. Of crucial importance in this paradigm is the mapping model, which defines the space of possible relationships between the features and neural data. Until recently, most encoding and decoding studies have used *linear* mapping models. However, some researchers have argued that the space of linear mappings is overly constrained and advocated for the use of more flexible *nonlinear* mapping models. Here, we discuss the choice of a mapping model in the context of three overarching goals: predictive accuracy, interpretability, and biological plausibility. We show that, contrary to popular intuition, these goals do not map cleanly onto the linear/nonlinear divide. Moreover, we argue that, instead of categorically treating the mapping models as linear or nonlinear, we should instead aim to estimate the complexity of these models. We show that, in most cases, complexity provides a more accurate reflection of restrictions imposed by various research goals and outline several complexity metrics that can be used to effectively evaluate mapping models.

1. Introduction

In recent decades, neuroscientists have witnessed a massive increase in the amount of available data, as well as in the computational power of tools we can apply to the data. As a result, today we can leverage huge datasets to build powerful models of brain activity. In this era of new opportunities, it is important to be mindful of conceptual choices we have to make before modeling our data. The goal of this paper is to discuss one such choice: a mapping model that relates features of interest to brain responses.

When studying a brain circuit, area, or network, it is often useful to formulate and test hypotheses about **features** that elicit a response in the relevant neural units (a single cell, a population of neurons, a brain area, etc.). The features can be stimulus-based (**Figure 1A**), behavior-based (**Figure 1B**), or based on responses in other neural units, within the same brain or a brain of another individual (**Figure 1C**). The exact source of the features may vary: common sources include human annotations (e.g., “faces”), empirical measurements (e.g., behavioral responses), and outputs of a computational model (e.g., a vector of responses to each image in layer 4 of a deep neural network (DNN); **Figure 1D**).

To relate a set of features to brain data, we need to establish a **mapping** between them. A perfect feature set would fully explain neural activity, such that there would be a 1:1 mapping between a the feature set and neural activity in a given neuron/electrode/voxel[1]. However, cognitive neuroscience today is far from producing perfect features. One limitation is our incomplete understanding of the computational mechanisms underlying brain function. Another limitation is the data we use, which typically provide an indirect and/or noisy measure of neural activity. These limitations mean that most feature sets cannot be perfectly mapped onto the neural data; instead, some level of fitting is required.

A mapping model is a model trained to map between the features and neural data. In this paper, we discuss both *encoding* mapping models, i.e. models that map from the feature model to the neural variable, and *decoding* mapping models, i.e. models that map from the neural variable to the feature model (see **Figure 1**). Others have discussed the relative merits of the two approaches (Holdgraf et al., 2017; King et al., 2018; Kriegeskorte & Douglas, 2019; Naselaris et al., 2011); our arguments in this paper apply to both mapping directions, unless specified otherwise. Note also the difference between a mapping model, which is trained directly on brain data, and a computational model that aims to mimic brain function but does not necessarily use neural data (**Figure 2**).

[1] Note that the neural data being fitted is not necessarily the neural recording itself: researchers may choose to predict the average firing rate, power in a particular frequency band, or beta coefficients from the general linear model (GLM) of fMRI responses (King et al., 2018).

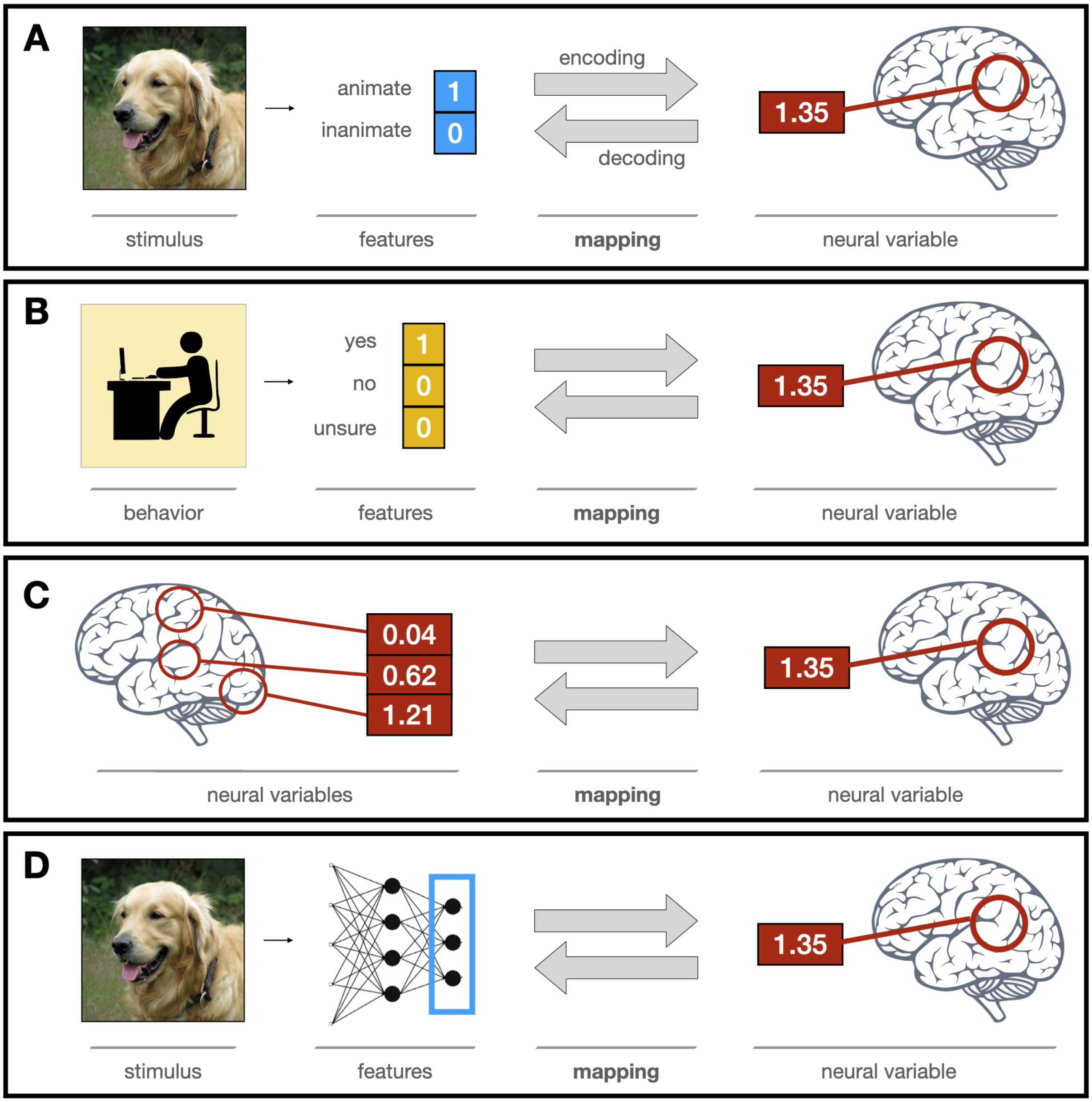

***Figure 1**. The encoding/decoding modeling framework in cognitive neuroscience. (**A**) Studies investigating the effect of external stimuli on brain activity start with the stimulus, extract its features of interest, and use a mapping model to establish the mapping between these features and a neural variable extracted from the data recorded during/after stimulus presentation. (**B**) In other studies, researchers extract features associated with participants' behavior and map those onto the neural variable recorded before/during this behavior. (**C**) Yet another class of studies describes the mapping between activity in different brain regions. (**D**) In recent years, more and more studies replace hand-crafted features, like those shown in (A), with large feature vectors derived from computational models, such as neural networks.*

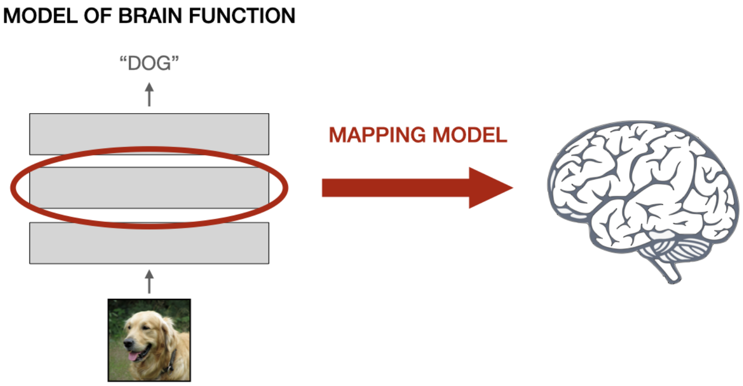

***Figure 2**. The distinction between a model of brain function and a mapping model. A model of brain function aims to mimic the brain, but does not directly fit to neural data. Our focus is on mapping models, which directly link a feature set to a neural variable. The mapping model depicted here uses features derived from a model of brain function (like in **Figure 1D**).*

Mapping models have many properties, but the most common distinction is drawn between (A) a *linear mapping model* (such as linear regression) and (B) a *nonlinear mapping model* (such as a neural network). The key question we consider here is how to decide which mapping model is most appropriate for a particular research goal.

2. The controversy

Recent advances in machine learning (ML) have had a large effect on cognitive neuroscience (Bzdok et al., 2017). Whereas traditional research approaches use relatively simple mappings to relate a handful of features to a single neural variable (e.g., "animate/inanimate" vs. the average BOLD response in a brain region of interest), ML-based mapping models can operate on high-dimensional feature vectors and can learn the mapping between features and data without presupposing the exact nature of that mapping. The concurrent increase in the size of available datasets (e.g., Chang et al., 2019; Majaj et al., 2015; Schoffelen et al., 2019) has enabled researchers to train large-scale mapping models without overfitting them.

A number of studies have leveraged the power of ML methods to build flexible nonlinear mapping models and use them to identify neural correlates of brain disorders (e.g., Hasanzadeh et al., 2019; Kazemi & Houghten, 2018; Kim et al., 2016; Leming et al., 2020) and behavioral traits (e.g., Kumar et al., 2019; Morioka et al., 2020; Xiao et al., 2019). Yet the vast majority of cognitive neuroscience studies use linear mapping models (such as linear regression), resulting in a gap between different neuroscience subfields.

Three primary justifications have been proposed for using linear mapping models:

1. Linear mapping models facilitate comparison of *predictive accuracy* across feature sets (e.g., Caucheteux & King, 2020; Jain & Huth, 2018; Schrimpf et al., 2018; Yamins et al., 2014).
2. Linear mapping models estimate weights for individual features, making the mapping more *interpretable* (e.g., Anderson et al., 2017; Lee Masson & Isik, 2021; Naselaris et al., 2011; Sudre et al., 2012; cf. Haufe et al., 2014; Kriegeskorte & Douglas, 2019).
3. Linear mapping models are more *biologically plausible*: they approximate readout by a downstream area and can therefore indicate what information is available to the rest of the brain (e.g., Kamitani & Tong, 2005; Kriegeskorte, 2011).

In the following section, we critically review arguments for linear vs. nonlinear mapping models in the context of these and other research goals, with the aim of establishing a general framework for specifying the mapping model in various research scenarios. We show that each of the three broad criteria used to evaluate mapping models — predictive accuracy, interpretability and biological plausibility — can be broken down into several different research goals, each of which places its own set of constraints on the mapping model.

3. What do we want from our mapping models?

To pick the best model, we first need to specify the goal that we are trying to achieve (Kording et al., 2020). This goal can be described as a set of particular desiderata for the model. The most common desiderata for models in cognitive neuroscience are predictive accuracy, interpretability, and biological plausibility. Here, we discuss what each of these desiderata might mean in the context of mapping models.

3.1. Predictive accuracy

A necessary condition for a successful scientific model is its ability to explain past results, as well as to predict future (or held-out) data. In neuroscience, as in other fields, scientific progress is driven by researchers generating new hypotheses, testing their predictions against experimental results, and focusing on the most accurate hypotheses. In the encoding/decoding framework, a hypothesis can be operationalized as a set of features derived from a computational model (e.g., a neural network layer; Yamins et al., 2014) or from behavioral ratings (e.g., Anderson et al., 2017).

A common way to measure the predictive accuracy of a set of features is to use a mapping model that will estimate the best link between the features and the training set of the neural data. The mapping — fit on the training set — can then be used to predict responses in a held-out test set, after which we can evaluate those predictions by, e.g., correlating them with actual data. These correlations are often normalized by an estimate of the reliability of the data (a “ceiling”) to yield an estimate of *explained variance* (e.g., Cadieu et al., 2014; Kell et al., 2018; Schrimpf et al., 2018; Yamins et al., 2014).

This prediction-oriented framework can be used to achieve multiple goals, only some of which require imposing specific constraints on the mapping model.

3.1.1. Compare feature sets. Model accuracy (e.g., explained variance) can be used to compare competing feature sets to figure out which of them best reflects neural responses (Schrimpf et al., 2020). Such comparison-oriented studies tend to use linear mappings in order to minimize the number of additional computations performed on the features. Indeed, using a powerful non-linear mapping model could wash out important differences across feature sets. For example, if the goal is to determine whether activity in inferior temporal cortex is better predicted by an early or a late layer of a convolutional neural network, we should use a mapping model with a limited expressive power; otherwise, the mapping model will be able to transform features from an early layer into features from a late layer, eliminating meaningful differences between them. Thus, feature comparison studies often benefit from linear mapping models.

3.1.2. Test feature decodability. Another research question a neuroscientist might ask is “do neural data Y contain information about features X?” A more applied version of this question is “can I predict features X based on neural data Y?” In this scenario, the goal is to find a mapping that allows us to reach above-chance decoding accuracy[2]. If this is the primary goal, then we should not put restrictions on the space of possible mappings — all that matters is the mapping model’s performance on held-out data. For instance, if a study aims to determine whether certain behavioral traits (X) can be predicted from neural data (Y), it does not need to limit its scope of possible mappings as long as the resulting mapping model performs successfully on unseen data. Similarly, a study that finds information about an imagined visual scene (X) in primary visual cortex (Y) may provide a valuable contribution to the field even if it uses highly unconstrained nonlinear mappings. All in all, for studies in this category, the main objective is for the mapping model to achieve the highest possible predictive accuracy (or, for applied research, to reach a certain accuracy threshold) regardless of (non)linearity.

3.1.3. Build accurate models of brain data. Finally, some researchers are trying to build accurate models of the brain that can *replace experimental data* or, at least, reduce the need for experiments by running studies in silico (e.g., Jain et al., 2020; Kell et al., 2018; Yamins et al., 2014). The main criterion for such models is their predictive accuracy, but they need to clear a very high accuracy bar (ideally at the level of the noise ceiling). The best way to build these in silico brains might be to train large powerful mapping models on large amounts of neural data. In

[2] Some authors refer to predictive accuracy as a measure of mutual information between features X and neural data Y (e.g., Kriegeskorte, 2011), but this relationship is not always straightforward. In general, estimating mutual information is a hard problem (Paninski, 2003), although some ML-based approaches can provide a useful approximation (e.g., Belghazi et al., 2018; Xu et al., 2020).

this scenario, there is no theoretical justification for a linear mapping constraint because the primary goal is maximizing predictive accuracy.

3.2. Interpretability

Once we find a mapping that achieves sufficiently high predictive accuracy, we often want to interpret it. Which features contribute the most to neural activity? Do neurons/electrodes/voxels respond to single features or exhibit mixed selectivity? How does the mapping relate to other models or theories of brain function?

The traditional view is that linear mappings are easier to interpret than non-linear mappings (Naselaris et al., 2011). However, the goal of building interpretable models is ultimately complicated by the fact that there is no clear-cut definition for interpretability. Below, we discuss three definitions of interpretability, ranging from strictest to loosest, and show that interpretability does not always require a linear mapping model. Importantly, in each of these cases, interpretability places restrictions not only on the mapping model, but also on the features that can be used to yield meaningful interpretations.

3.2.1. Examine individual features. Traditionally, many cognitive neuroscientists have believed that interpreting a neural signal requires identifying a set of words to describe its function (e.g., Desimone et al., 1984; Kanwisher et al., 1997). In this scenario, a useful model of brain activity has features that can be described using one or a few words ("faces", "vertical lines", etc.) and a linear mapping from these features to neural data. We consider this to be the strictest definition of interpretability because it places the strongest constraints on both the features (which have to be nameable) and the mapping models (which have to be linear). With a linear mapping model, the regression weights can be interpreted as a relative measure of contribution to the neural activity (though this is not always straightforward in cases where features span different values or suffer from multicollinearity).

Interpretable features have played a crucial role in understanding brain function (Kanwisher, 2010). However, nameability of features may be an overly restrictive metric of interpretability as it limits our understanding to a vocabulary that is heavily biased by a priori hypotheses and may not include words for the concepts we actually need (Buzsáki, 2019). For instance, recent work has shown that neurons typically described as "face-responsive" respond more strongly to artificial images produced by DNNs than to natural images described by the word "face" (Ponce et al., 2019), suggesting that simple verbal features cannot provide a full account of neural activity. As a result, many researchers have started to use higher-dimensional sets of features, a move that has introduced new definitions of what it means for a mapping to be interpretable.

3.2.2. Test representational geometry. A looser definition of interpretability that has become popular in the last decade is the use of high-dimensional feature vectors that are linearly mapped to a neural variable. These features may be produced by humans, such as rating properties of words (Binder et al., 2016), or by computational models, such as semantic embeddings (e.g., Mitchell et al., 2008; Pereira et al., 2018) or computer vision features (e.g., Kay et al., 2008; Yamins et al., 2014). When using large-scale feature sets, we cannot always interpret the weights of a linear mapping model in the same way as we did with nameable features. If individual features within a set cannot be labeled (e.g., in the case of DNN layer activations), examining them one by one has a limited potential to inform our intuition (Kay, 2018). However, we can examine the feature set as a whole, asking: do features X, generated by a known process, accurately describe the space of neural responses Y? Thus, the feature set becomes a new unit of interpretation, and the linearity restriction is placed primarily to preserve the overall geometry of the feature space. For instance, the finding that convolutional neural networks and the ventral visual stream produce similar representational spaces (Yamins et al., 2014) allows us to infer that both processes are subject to similar optimization constraints (Richards et al., 2019). That said, mapping models that probe the representational geometry of the neural response space do not have to be linear, as long as they correspond to a well-specified hypothesis about the relationship between features and data.

3.2.3. Describe the feature set. The loosest definition of interpretability is the ability to describe the set of features that was used to train the mapping model (e.g. “phonological features”). In this scenario, we make no assumptions about a particular representational geometry of these features (such as linear separability). The lack of specific assumptions about the form of the feature-data mapping means that constraints on the mapping model are not strictly necessary — all we need is an epistemologically satisfying description of the features. If a mapping model achieves good predictivity, we can say that a given set of features is reflected in the neural signal. In contrast, if a powerful mapping model trained on a large set of data achieves poor predictivity, it provides strong evidence that a given feature set is not represented in the neural data. Under this definition, any mapping model is interpretable as long as we can describe the set of features that it uses.

3.3. Biological plausibility

In addition to prediction accuracy and interpretability-related considerations, biological plausibility can also be a factor in deciding on the space of acceptable feature-brain mappings. We discuss two goals related to biological plausibility: simulating linear readout and accounting for physiological mechanisms affecting measurement.

3.3.1. Simulate linear readout. One of the main arguments put forth in favor of linear mappings is the claim that they approximate the linear readout performed by a putative downstream brain area (Kamitani & Tong, 2005; Kriegeskorte, 2011). Under this view, the mapping model approximates the transmission of the features to a hypothetical information consumer. The linear readout requirement often serves as a proxy for feature usability: if the features can be extracted with a linear mapping model, it means that they do not require extensive computations in order to be used downstream.

The ability to use features of interest in downstream computations is indeed an important consideration. However, there are reasons to be cautious about the linear readout requirement. First, some models operate on neural data collected from multiple recording sites rather than a single neural population/region, making subsequent linear readout biologically implausible. For instance, decoding models that use whole-brain data, such as M/EEG, have no downstream region that could ‘read out’ information from all over the brain — the only entity performing readout is the observer. Second, linear readout might not be an accurate characterization of the decoding mechanisms used by downstream areas to extract information from the brain region of interest. In fact, unlike linear models that can pool across all measured neurons or voxels in the region of interest, readout in biological neural systems is likely to be both sparse (e.g., Barak et al., 2013; Barlow, 1969; Olshausen & Field, 2004; Vinje & Gallant, 2000) and nonlinear (e.g., Ghazanfar & Nicolelis, 1997; Gidon et al., 2020; Hodgkin & Huxley, 1939; Shamir & Sompolinsky, 2004). Third, linear regression is a fairly arbitrary threshold to draw for mechanistic plausibility. Even a relatively ‘constrained’ linear classifier can read out many features from the data, many of them biologically implausible (e.g., voxel-level ‘biases’ that allow orientation decoding in V1 using fMRI; Ritchie et al., 2019). In sum, unconstrained linear mapping models (or linear mapping models constrained by weight distribution among many features, like ridge regression) may be both overly limiting because they do not account for possible nonlinear computations and overly greedy because they might leverage information in a way that real neurons do not.

Is there a better mapping model that accounts for possible nonlinear computations during readout without being overly broad? One possible approach is to introduce parsimony constraints on the feature space of the models (Kukačka et al., 2017). Introducing the sparsity constraint (i.e., allowing the mapping model to access only a limited number of neurons) could increase the biological plausibility of putative readout (Yoshida & Ohki, 2020). However, in the context of measurements that collapse across large numbers of neurons (i.e. most measurements in cognitive neuroscience), the sparsity constraint might be impossible to enforce, as a single voxel or electrode already combines signal from a large number of neurons. More broadly, evaluating the biological plausibility of decoding is difficult, as readout might differ across brain regions of interest (Anzellotti & Coutanche, 2018), and the current understanding of the details of readout mechanisms is still limited. Future progress in research on readout mechanisms will be key to evaluate the plausibility of different assumptions about readout in a more principled manner.

3.3.2. Incorporate physiological mechanisms affecting measurement. When brain recordings are known to be nonlinear transformations of underlying neural activity (e.g. fMRI, in which BOLD responses are related to neural responses via the HRF; Friston et al., 2000), knowledge about the nonlinear relationship between the neural responses and the measurements can (and often should) be explicitly incorporated into the mapping. Failing to do so might privilege feature sets that incorporate properties of the measurement over feature sets that more accurately reflect the neural representations encoded in a brain region.

In some cases, the nonlinearity can be incorporated when deriving the neural variable before the model is fitted (e.g. the beta weights for fMRI or the power in a given frequency range for EEG/MEG/ECoG). However, this is not always the best approach. For fMRI, the shape of the relationship between neural activity and BOLD responses varies across different subjects and brain regions, and even across different voxels (Ekstrom, 2021; Handwerker et al., 2004). This implies that the frequently used strategy of convolving feature responses with a standard HRF that is fixed across voxels and participants has its limitations, and that mapping models might benefit from integrating nonlinear estimation of the HRF shape within a family of functions motivated by physiological data (see, for instance, Pedregosa et al., 2015; Shain et al., 2020). Region- and voxel-specific HRFs can be set by estimating a relatively small number of parameters; thus, they would require only a small increase in model complexity. For M/EEG, the recorded signal is a combination of both inhibitory and excitatory signals; thus, treating it as a straightforward linear combination is not always possible (Hansen et al., 2010). Linear mapping models often overlook the complexities of neuroimaging signals, sacrificing biological plausibility as a result.

To summarize thus far, different research goals place very different constraints on the mapping model. A particular goal might require choosing a linear mapping model, adding additional restrictions to that model, using a particular class of nonlinear models, or imposing no a priori restrictions whatsoever.

4. Practical considerations

The criteria outlined above are primarily based on theoretic considerations: which mapping model has the properties that allow us to achieve a particular goal? However, another important consideration is practical feasibility: do we have enough data to accurately estimate the mapping? Will the noise in our data lead certain mapping models to fail?

Determining how much data is required for fitting a particular mapping model has critical implications for experimental design (the number of trials/data points per participant, the number of repetitions per stimulus, etc.). In general, the fewer constraints are placed on the mapping model, the more data will be needed to converge on a good mapping. This relationship can be estimated using standard validation methods by, for instance, taking a large dataset and evaluating the mapping model's predictive accuracy on left-out test data while gradually increasing the size of the training dataset. However, few studies report such analyses (and in some cases large-enough datasets may still be lacking). One exception is a line of fMRI studies that aim to determine the best mapping model for linking interregional functional correlations and behavioral/demographic traits. The results of these studies are mixed: some report a marked advantage of nonlinear mapping models over linear ones (Bertolero & Bassett, 2020), whereas others report that linear mapping models perform equally well even when the training set includes several thousand brain images (He et al., 2020; Schulz et al., 2020). Thus, the field would greatly benefit from further systematic examinations of the influence of dataset size (and other experimental design properties) on the performance of a particular mapping model type.

Even with infinite data, certain measurement properties might force us to use a particular mapping class. For instance, Nozari et al. (2020) show that fMRI resting state dynamics are best modeled with linear mappings and suggest that fMRI's inevitable spatiotemporal signal averaging might be to blame (although see Anzellotti et al., 2017, for contrary evidence). In sum, even after establishing theoretical desiderata for the mapping model, we need to conduct rigorous empirical tests to determine which model class will achieve good predictive accuracy given the amount and quality of available data.

5. Moving forward: estimating model complexity

Instead of focusing exclusively on the linear/nonlinear dichotomy, we propose to view the choice of the mapping model in the context of a broader notion of model complexity. Complexity lies at the heart of most desiderata discussed above. Arbitrarily complex models are less likely to generalize, leading to decreased predictive accuracy on test data; they can be harder to interpret; and they are less likely to match computations in biological circuits. All these considerations support the idea of an Occam's razor, whereby one should strive for the simplest mapping that will achieve the desired goal.

5.1 The role of complexity in mapping model selection

We argue that, most of the time, the debate over linear/nonlinear models is, in reality, a debate over how complex the mapping model should be. For instance, the use of linear mappings when comparing feature sets is dictated by the desire to reduce the amount of computation performed on the features — in other words, the desire to reduce mapping model complexity. Similarly, the linear readout requirement is introduced to ensure that the features are usable by downstream brain regions. However, a simple nonlinear mapping might make the features equally usable (given the nonlinear nature of most brain computations). Thus, we suggest reframing the linear/nonlinear debate in terms of model complexity.

Figure 3 shows the research goals discussed above together with the mapping model types that are traditionally used to achieve these goals, as well as our proposal to shift from the linear/nonlinear dichotomy to explicit estimates of model complexity. Note that this diagram depicts theoretical, a priori criteria for restricting mapping model complexity; practical considerations might impose additional constraints to achieve better predictivity (see **Section 4**).

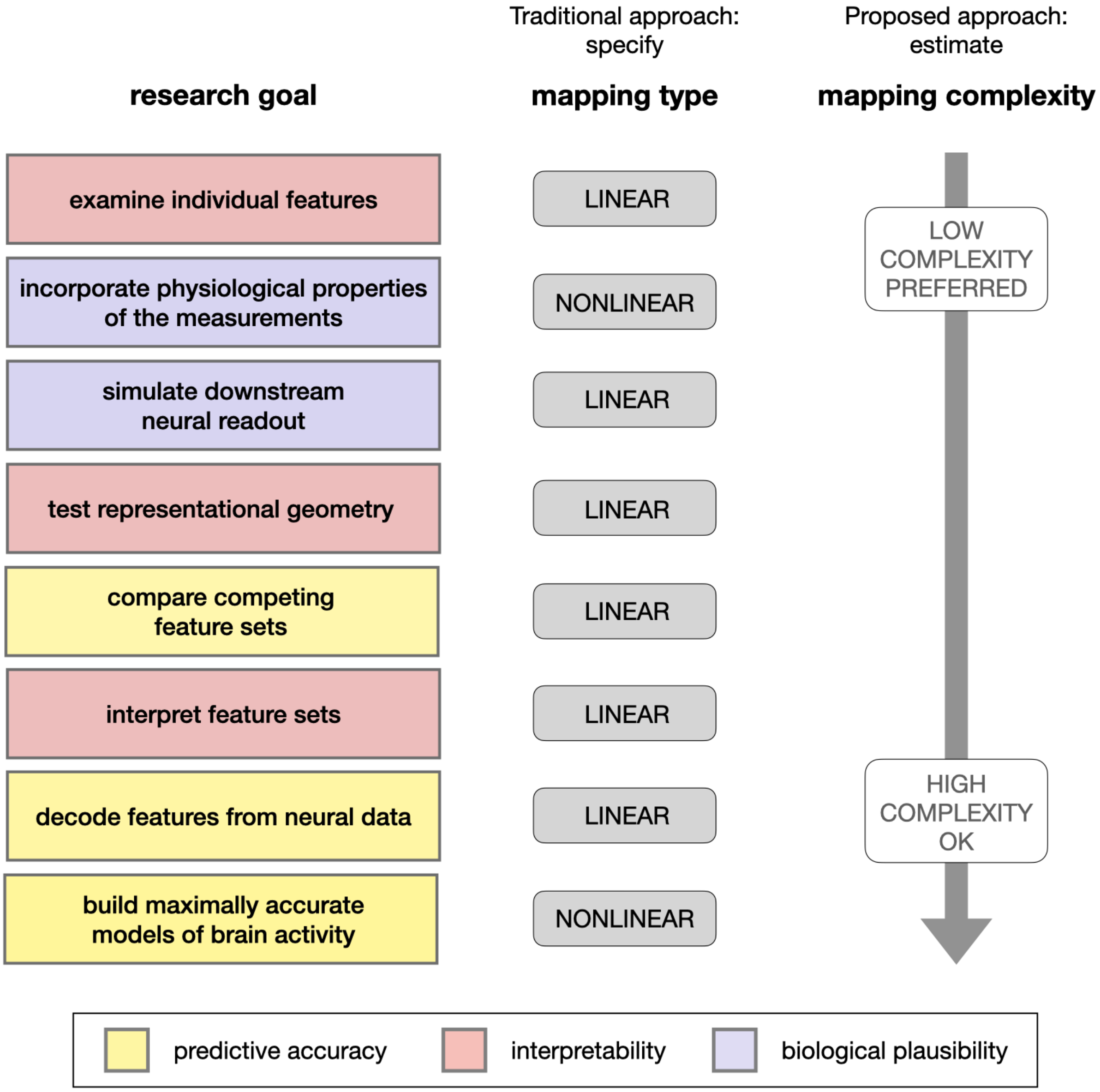
Traditional approach: specify
Proposed approach: estimate
research goal
mapping type
mapping complexity
examine individual features
LINEAR
LOW COMPLEXITY PREFERRED
incorporate physiological properties of the measurements
NONLINEAR
simulate downstream neural readout
LINEAR
test representational geometry
LINEAR
compare competing feature sets
LINEAR
interpret feature sets
LINEAR
decode features from neural data
LINEAR
HIGH COMPLEXITY OK
build maximally accurate models of brain activity
NONLINEAR
predictive accuracy
interpretability
biological plausibility

Figure 3*. Different research goals are currently being collapsed into the "linear/nonlinear" dichotomy (L - linear, NL - nonlinear), but in fact correspond to different degrees of mapping model complexity. Note that the ordering of research goals along the complexity continuum is approximate and shown primarily for illustration purposes.*

It turns out that the goals within each of the three broad categories — predictive accuracy, interpretability, and biological plausibility — impose very different constraints on the complexity of the mapping model. Further, these constraints are often more graded than the linear/nonlinear distinction:

- Interpreting individual features is easier when the mapping is not only linear, but also sparse, so that each neuron can be described with only a few features. Reframing the mapping model choice in terms of complexity allows us to pick out simple mappings *within* the class of linear mapping models, thus facilitating interpretation.
- Satisfying biological constraints, such as accounting for physiological properties of the measurement or simulating neural readout, may require a certain degree of nonlinearity but these nonlinearities are often well-defined and can keep overall model complexity relatively low.
- Testing whether a feature set accurately captures the representational geometry of neural responses requires the mapping model to reflect that geometry. Thus, the complexity of the mapping model depends primarily on the hypothesis being tested.
- Comparing and/or interpreting feature sets is possible even when the mapping is nonlinear, as long as we can compare the mappings using a metric that incorporates both predictive accuracy and model complexity.
- Decoding features from neural data and building accurate encoding models of the brain does not require placing any theory-based restrictions on the mapping model (although such restrictions might improve performance in practice).

5.2 Complexity measures

How can we estimate the complexity of our mapping models? To date, many studies have focused primarily on a binary distinction in which linear models are "simple" and nonlinear models are "complex". However, as discussed above, this distinction is often overly simplistic. Here, we review several measures of mapping model complexity that are commonly used in the ML literature and may serve as an alternative to the linear/nonlinear dichotomy.

5.2.1. Number of free parameters. A very common approach to measuring complexity is by taking the number of free parameters in the model. In this approach, each model class receives a penalty that corresponds to its number of parameters, such that classes with more parameters have a larger penalty. In order to justify the use of additional parameters, the model needs to achieve substantial performance improvement compared to simpler models. This tradeoff is often implemented using Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which reward models for good predictive performance but penalize them for the number of parameters. Note, however, that this approach often fails to capture distinctions that seem intuitively relevant. For instance, a linear and a nonlinear model with the same number of parameters would have equal complexity in this view, even though the latter often has a greater expressive power. Another example is a sparse mapping model that only has non-zero weights for 1-2 features vs. a dense model that places non-zero weights on, say, 500 features: if the initial feature vector size is the same, then these models will have the same number of parameters and therefore equivalent complexity under this metric.

Benefits: easy to estimate.
Limitations: does not always reflect relevant complexity distinctions; sensitive to model architecture.
Sample use case: choosing among multiple well-performing mapping models with the goal of maximizing predictive accuracy on a new (untested) dataset.

5.2.2. Minimum description length. Another common approach to measuring model complexity is based on the idea of minimum description length (MDL; Rissanen, 1978). This approach typically assumes an encoding function over a class of models, and the complexity of each model within the class is determined by the length of the model's encoding. The encoding function essentially serves as a prior over the model class: more probable mapping models would be assigned shorter descriptions (see also Diedrichsen & Kriegeskorte, 2017; Wu et al., 2006, for a discussion of the relationship between priors and regularization constraints). The MDL approach can overcome some limitations of complexity measures based solely on the number of free parameters by exploiting correlations between parameters to achieve a shorter description length. For instance, under this scheme, sparse models would have a shorter description length and would therefore be considered less complex. The main limitation of an MDL-based metric is that it requires specifying a mapping model class, as well as an encoding scheme for mappings within that class. Thus, if there is no natural prior over the set of mappings we wish to compare, an architecture-free complexity measure may be preferred[3].

Benefits: captures many relevant distinctions; less arbitrary than the number of parameters.
Limitations: requires specifying a mapping model class and a prior over mappings in that class.
Sample use case: comparing competing feature sets without presupposing linear separability of these features.

5.2.3. Sample complexity. Finally, a more practice-oriented metric is sample complexity. Loosely speaking, the sample complexity of a model class is a function that determines the minimal number of training samples that are required in order to achieve desired model performance. It is not always straightforward to compute this function a-priori; however, it can be assessed empirically by computing learning curves, i.e., the achieved level of predictive accuracy on a test set as a function of the number of training samples (see **Section 4**). Estimates

[3] For example, certain informational measures (e.g., Bialek et al., 2001; Gilad-Bachrach et al., 2003) can be used to measure the complexity of the statistical relationship between the inputs and outputs of the mapping model (e.g., features and predicted neural data) regardless of a particular architecture or model class, and in some cases may also capture the complexity of non-parametric generative models.

of sample complexity are vital for understanding whether a given model failed because the underlying hypothesis was wrong or because the dataset was too small to achieve a proper fitting.

Benefits: immediate practical application.
Limitations: an indirect measure of model complexity.
Sample use case: estimating the amount of data required to achieve good predictivity for a given mapping model.

Overall, instead of defaulting to linear models, we propose incorporating the estimate of mapping model complexity into the general evaluation framework of encoding/decoding models. This estimate can be used in different ways depending on the research goal. For instance, for feature comparison, if two feature sets produce equally accurate mapping models, the feature set corresponding to a simpler mapping model represents a better fit to neural data. For estimates of potential downstream readout, instead of limiting ourselves to linear functions, we can consider a range of possible mappings, where simpler mappings reflect a higher probability that these features are used downstream. Thus, estimates of model complexity can serve as a powerful complement to predictive accuracy when evaluating mapping model performance.

6. Conclusion

The encoding/decoding framework in modern cognitive neuroscience has provided many valuable insights. However, in some cases, the field has been held back by its excessive reliance on linear mappings between features and brain activity. Here, we have described various research goals that should be taken into account when specifying a mapping model. Contrary to popular belief, few of these goals require the use of linear mapping models. Instead, some do not require placing *any* constraints on the mapping model, some require placing specific *nonlinear* constraints, and some use linearity simply as a proxy for reducing model complexity. We therefore propose to explicitly include estimates of model complexity when evaluating mapping models. Incorporating such estimates can help the field overcome its reliance on linear mappings and discover a richer space of accurate, simple, biologically plausible predictors of brain activity.

Acknowledgements

This paper is part of the Generative Adversarial Collaboration (GAC) initiative organized by the Computational Cognitive Conference board. We thank the GAC organizers, especially Megan Peters and Gunnar Blohm, for their invaluable help with this initiative. Many of the ideas discussed in this work arose during the GAC workshop in October 2020 (https://www.youtube.com/watch?v=UI5KclR71IE&list=PLNWftEg2R4s5iObSUvPXhnDJvyN

bs4PnM&index=2). We thank the invited workshop speakers — Kohitij Kar, Mariya Toneva, Laura Gwilliams, Jean-Rémi King, Martin Hebart, and Anna Schapiro, — as well as workshop participants, for their ideas, comments, questions and suggestions. We also thank the reviewers who provided comments on our GAC proposal in August 2020 (the reviews are available at https://openreview.net/forum?id=-o0dOwashib). MS was supported by a Takeda Fellowship and the SRC Semiconductor Research Corporation. NZ was supported by a BCS Fellowship in Computation. EF was supported by R01 awards DC016607 and DC016950.

References

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Aguilar, M., Wang, X., Doko, D., & Raizada, R. D. S. (2017). Predicting Neural Activity Patterns Associated with Sentences Using a Neurobiologically Motivated Model of Semantic Representation. *Cerebral Cortex*, *27*(9), 4379–4395. https://doi.org/10.1093/cercor/bhw240

Anzellotti, S., & Coutanche, M. N. (2018). Beyond Functional Connectivity: Investigating Networks of Multivariate Representations. *Trends in Cognitive Sciences*, *22*(3), 258–269. https://doi.org/10.1016/j.tics.2017.12.002

Anzellotti, S., Fedorenko, E., Kell, A. J. E., Caramazza, A., & Saxe, R. (2017). Measuring and Modeling Nonlinear Interactions Between Brain Regions with fMRI. *BioRxiv*, 074856. https://doi.org/10.1101/074856

Barak, O., Rigotti, M., & Fusi, S. (2013). The Sparseness of Mixed Selectivity Neurons Controls the Generalization–Discrimination Trade-Off. *Journal of Neuroscience*, *33*(9), 3844–3856. https://doi.org/10.1523/JNEUROSCI.2753-12.2013

Barlow, H. (1969). Trigger features, adaptation and economy of impulses. In *Information Processing in the Nervous System* (pp. 209–230). Springer.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, R. D. (2018). MINE: Mutual Information Neural Estimation. *ArXiv:1801.04062*. https://arxiv.org/abs/1801.04062v4

Bertolero, M. A., & Bassett, D. S. (2020). Deep Neural Networks Carve the Brain at its Joints. *ArXiv:2002.08891 [Physics, q-Bio]*. http://arxiv.org/abs/2002.08891

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, *13*(11), 2409–2463. https://doi.org/10.1162/089976601753195969

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3–4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

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

Bzdok, D., Varoquaux, G., & Thirion, B. (2017). Neuroimaging Research: From Null-Hypothesis

Falsification to Out-of-Sample Generalization. *Educational and Psychological Measurement*, *77*(5), 868–880. https://doi.org/10.1177/0013164416667982

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Caucheteux, C., & King, J.-R. (2020). Language processing in brains and deep neural networks: Computational convergence and its limits. *BioRxiv*, 2020.07.03.186288. https://doi.org/10.1101/2020.07.03.186288

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, *6*(1), 49. https://doi.org/10.1038/s41597-019-0052-3

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*(8), 2051–2062. https://doi.org/10.1523/JNEUROSCI.04-08-02051.1984

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, *13*(4), e1005508. https://doi.org/10.1371/journal.pcbi.1005508

Ekstrom, A. D. (2021). Regional variation in neurovascular coupling and why we still lack a Rosetta Stone. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1815), 20190634. https://doi.org/10.1098/rstb.2019.0634

Friston, K. J., Mechelli, A., Turner, R., & Price, C. J. (2000). Nonlinear responses in fMRI: The Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, *12*(4), 466–477. https://doi.org/10.1006/nimg.2000.0630

Ghazanfar, A. A., & Nicolelis, M. A. (1997). Nonlinear processing of tactile information in the thalamocortical loop. *Journal of Neurophysiology*, *78*(1), 506–510. https://doi.org/10.1152/jn.1997.78.1.506

Gidon, A., Zolnik, T. A., Fidzinski, P., Bolduan, F., Papoutsi, A., Poirazi, P., Holtkamp, M., Vida, I., & Larkum, M. E. (2020). Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, *367*(6473), 83–87. https://doi.org/10.1126/science.aax6239

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2003). An Information Theoretic Tradeoff between Complexity and Accuracy. In B. Schölkopf & M. K. Warmuth (Eds.), *Learning Theory and Kernel Machines* (pp. 595–609). Springer. https://doi.org/10.1007/978-3-540-45167-9_43

Handwerker, D. A., Ollinger, J. M., & D’Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, *21*(4), 1639–1651. https://doi.org/10.1016/j.neuroimage.2003.11.029

Hansen, P. C., Kringelbach, M. L., & Salmelin, R. (Eds.). (2010). *MEG: An introduction to methods* (pp. xii, 436). Oxford University Press.

https://doi.org/10.1093/acprof:oso/9780195307238.001.0001

Hasanzadeh, F., Mohebbi, M., & Rostami, R. (2019). Prediction of rTMS treatment response in major depressive disorder using machine learning techniques and nonlinear features of EEG signal. *Journal of Affective Disorders*, *256*, 132–142. https://doi.org/10.1016/j.jad.2019.05.070

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. https://doi.org/10.1016/j.neuroimage.2013.10.067

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, *206*, 116276. https://doi.org/10.1016/j.neuroimage.2019.116276

Hodgkin, A. L., & Huxley, A. F. (1939). Action Potentials Recorded from Inside a Nerve Fibre. *Nature*, *144*(3651), 710–711. https://doi.org/10.1038/144710a0

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*. https://doi.org/10.3389/fnsys.2017.00061

Jain, S., & Huth, A. G. (2018). Incorporating Context into Language Encoding Models for fMRI. *BioRxiv*, 327601. https://doi.org/10.1101/327601

Jain, S., Vo, V. A., Mahto, S., LeBel, A., Turek, J. S., & Huth, A. G. (2020). Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. *BioRxiv*, 2020.10.02.324392. https://doi.org/10.1101/2020.10.02.324392

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170. https://doi.org/10.1073/pnas.1005062107

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. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, *180*(Pt A), 101–109. https://doi.org/10.1016/j.neuroimage.2017.08.016

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. https://doi.org/10.1038/nature06713

Kazemi, Y., & Houghten, S. (2018). A deep learning pipeline to classify different stages of Alzheimer's disease from fMRI data. *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 1–8. https://doi.org/10.1109/CIBCB.2018.8404980

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. https://doi.org/10.1016/j.neuron.2018.03.044

Kim, J., Calhoun, V. D., Shim, E., & Lee, J.-H. (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage*, *124*, 127–146. https://doi.org/10.1016/j.neuroimage.2015.05.018

King, J.-R., Gwilliams, L., Holdgraf, C., Sassenhagen, J., Barachant, A., Engemann, D., Larson, E., & Gramfort, A. (2018). *Encoding and Decoding Neuronal Dynamics: Methodological Framework to Uncover the Algorithms of Cognition*. https://hal.archives-ouvertes.fr/hal-01848442

Kording, K. P., Blohm, G., Schrater, P., & Kay, K. (2020). Appreciating the variety of goals in computational neuroscience. *Neurons, Behavior, Data Analysis, and Theory*, *3*(6). http://arxiv.org/abs/2002.03211

Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *NeuroImage*, *56*(2), 411–421. https://doi.org/10.1016/j.neuroimage.2011.01.061

Kriegeskorte, N., & Douglas, P. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167–179. https://doi.org/10.1016/j.conb.2019.04.002

Kukačka, J., Golkov, V., & Cremers, D. (2017). Regularization for Deep Learning: A Taxonomy. *ArXiv:1710.10686 [Cs, Stat]*. http://arxiv.org/abs/1710.10686

Kumar, S., Yoo, K., Rosenberg, M. D., Scheinost, D., Constable, R. T., Zhang, S., Li, C.-S. R., & Chun, M. M. (2019). An information network flow approach for measuring functional connectivity and predicting behavior. *Brain and Behavior*, *9*(8), e01346. https://doi.org/10.1002/brb3.1346

Lee Masson, H., & Isik, L. (2021). Functional selectivity for naturalistic social interaction perception in the human superior temporal sulcus. *BioRxiv*, 2021.03.26.437258. https://doi.org/10.1101/2021.03.26.437258

Leming, M., Górriz, J. M., & Suckling, J. (2020). Ensemble Deep Learning on Large, Mixed-Site fMRI Datasets in Autism and Other Tasks. *International Journal of Neural Systems*, *30*(07), 2050012. https://doi.org/10.1142/S0129065720500124

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, *35*(39), 13402–13418. https://doi.org/10.1523/JNEUROSCI.5181-14.2015

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, *320*(5880), 1191–1195.

https://doi.org/10.1126/science.1152876

Morioka, H., Calhoun, V., & Hyvärinen, A. (2020). Nonlinear ICA of fMRI reveals primitive temporal structures linked to rest, task, and behavioral traits. *NeuroImage*, *218*, 116989. https://doi.org/10.1016/j.neuroimage.2020.116989

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nozari, E., Stiso, J., Caciagli, L., Cornblath, E. J., He, X., Bertolero, M. A., Mahadevan, A. S., Pappas, G. J., & Bassett, D. S. (2020). Is the brain macroscopically linear? A system identification of resting state dynamics. *ArXiv:2012.12351 [Cs, Eess, Math, q-Bio]*. http://arxiv.org/abs/2012.12351

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487. https://doi.org/10.1016/j.conb.2004.07.007

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*(6), 1191–1253. https://doi.org/10.1162/089976603321780272

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., & Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, *104*, 209–220. https://doi.org/10.1016/j.neuroimage.2014.09.060

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 1–13. https://doi.org/10.1038/s41467-018-03068-4

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. https://doi.org/10.1016/j.cell.2019.04.005

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471. https://doi.org/10.1016/0005-1098(78)90005-5

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *The British Journal for the Philosophy of Science*, *70*(2), 581–607. https://doi.org/10.1093/bjps/axx023

Schoffelen, J.-M., Oostenveld, R., Lam, N. H. L., Uddén, J., Hultén, A., & Hagoort, P. (2019). A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, *6*(1), 17. https://doi.org/10.1038/s41597-019-0020-y

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*, 407007. https://doi.org/10.1101/407007

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, *108*(3), 413–423. https://doi.org/10.1016/j.neuron.2020.07.040

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., & Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, *11*(1), 4238. https://doi.org/10.1038/s41467-020-18037-z

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). FMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Shamir, M., & Sompolinsky, H. (2004). Nonlinear Population Codes. *Neural Computation*, *16*(6), 1105–1136. https://doi.org/10.1162/089976604773717559

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, *62*(1), 451–463. https://doi.org/10.1016/j.neuroimage.2012.04.048

Vinje, W. E., & Gallant, J. L. (2000). Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science*, *287*(5456), 1273–1276. https://doi.org/10.1126/science.287.5456.1273

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*(1), 477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024

Xiao, L., Stephen, J. M., Wilson, T. W., Calhoun, V. D., & Wang, Y.-P. (2019). Alternating Diffusion Map Based Fusion of Multimodal Brain Connectivity Networks for IQ Prediction. *IEEE Transactions on Biomedical Engineering*, *66*(8), 2140–2151. https://doi.org/10.1109/TBME.2018.2884129

Xu, Y., Zhao, S., Song, J., Stewart, R., & Ermon, S. (2020). A Theory of Usable Information Under Computational Constraints. *ArXiv:2002.10689*. https://arxiv.org/abs/2002.10689v1

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 of the United States of America*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yoshida, T., & Ohki, K. (2020). Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, *11*(1), 872. https://doi.org/10.1038/s41467-020-14645-x