TYPE Review
PUBLISHED 20 September 2022
DOI 10.3389/fnins.2022.928841

On the encoding of natural music in computational models and human brains

Seung-Goo Kim*

Research Group Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

Check for updates

OPEN ACCESS

EDITED BY
Bruno L. Giordano,
UMR7289 Institut de Neurosciences
de la Timone (INT), France

REVIEWED BY
Christoph Daube,
University of Glasgow, United Kingdom
Tomoya Nakai,
INSERM U1028 Centre de Recherche
en Neurosciences de Lyon, France
Benjamin Morillon,
Institut National de la Santé et de la
Recherche Médicale (INSERM), France

*CORRESPONDENCE
Seung-Goo Kim
dr.seunggoo.kim@gmail.com

SPECIALTY SECTION
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience

RECEIVED 26 April 2022
ACCEPTED 15 August 2022
PUBLISHED 20 September 2022

CITATION
Kim S-G (2022) On the encoding of natural music in computational models and human brains. *Front. Neurosci.* 16:928841. doi: 10.3389/fnins.2022.928841

This article discusses recent developments and advances in the neuroscience of music to understand the nature of musical emotion. In particular, it highlights how system identification techniques and computational models of music have advanced our understanding of how the human brain processes the textures and structures of music and how the processed information evokes emotions. Musical models relate physical properties of stimuli to internal representations called features, and predictive models relate features to neural or behavioral responses and test their predictions against independent unseen data. The new frameworks do not require orthogonalized stimuli in controlled experiments to establish reproducible knowledge, which has opened up a new wave of naturalistic neuroscience. The current review focuses on how this trend has transformed the domain of the neuroscience of music.

KEYWORDS

system identification, naturalistic stimuli, computational model, natural music, artificial neural network, musical emotion

Introduction

Music is believed to have been a crucial part of all known societies from the very early days of the human species (Zatorre and Salimpoor, 2013). Bone flutes found near the Danube River in Germany suggest that the origin of music can be dated to about 40,000 years ago or more (Conard et al., 2009). Given that the emergence of *Homo sapiens* is believed to have emerged in Africa about 300,000 years ago (Hublin et al., 2017) and to have migrated from Africa to Eurasia around 60,000 years ago (Armitage et al., 2011), even earlier evidence of the musical traditions of humans may exist in Africa that is yet undiscovered (d'Errico et al., 2003). Furthermore, cross-cultural studies based on ethnographic texts and audio recordings provide empirical evidence that music appears in every society observed (e.g., Mehr et al., 2019); and this ubiquitous presence of music in human societies indicates music's significant functions for humans. Zatorre and Salimpoor (2013) suggested the reason for music's existence is that it allows for the expression and regulation of emotion and elicits pleasure. But the central question remains unsolved: How does music, a structured collection of abstract sounds, evoke such intensive emotions?

As neuroscientists understand it (Zatorre, 2005), music is processed by way of hierarchical pathways, potentially with feedbacks (Vuust et al., 2022), evoking emotions on multiple levels at different time scales (Juslin et al., 2013). According to the current understanding, waveforms of music are first transduced into neural activity throughout the auditory peripheral and central pathways, where spectral and temporal decompositions take place. The acoustic information is then believed to be transformed into musical events (e.g., notes, chords, beats). Musical structures (spectral and temporal relationships of musical events in the short and long terms; e.g., motifs, themes, tonalities, rhythms, tempi) can be recognized as statistical patterns depending on a listener's prior experience (e.g., enculturation, training) or concepts based on explicit knowledge. While explanations based on predictive processing have suggested important mechanisms (Zatorre and Salimpoor, 2013; Vuust et al., 2022), how other kinds of information of music transform into emotions remains largely to be discovered.

In many neuroscientific studies, stimuli were created as simple "models" (or approximations) of complex stimulations in real-world environments while parametrizing variables of interest and orthogonalizing nuisance variables. The orthogonalization of stimuli provides a simple model of the world and linearizes the assumed effects of the variable of interest. For instance, in our own previous study (Kim et al., 2017), we investigated the effect of dissonant harmony on evoked emotional responses and individual preferences using functional magnetic resonance imaging (fMRI). To this end, we created "dissonant versions" of 30-s excerpts taken from various instrumental musical pieces by transposing the original excerpts by dissonant intervals (major second upward, and diminished fifth downward) and mixing them all down. These "dissonant versions" constantly produced dissonant harmony regardless of the tonal structures in the original pieces. The altered audio clips certainly evoked "unpleasantness" (i.e., all participants rated the Pleasantness Scale lower) and decreased blood oxygen level-dependent (BOLD) responses in the auditory pathway and other brain regions as compared with the responses to the original pieces. Because the dissonant stimuli were created without altering other acoustic aspects such as loudness, beats, rhythms, phrases, and so on, the design was optimal for investigating the linear effect of consonance (or dissonance) without concerns about the multicollinearity of acoustics. One problem, however, was that the observed effect (i.e., "people disliked the dissonant versions") could not be generalized (i.e., "people dislike dissonant harmony"), because in real-world music such a dissonant harmony could nevertheless be perceived as "yet pleasant" when it is presented in different musical styles (Popescu et al., 2019). That is, although the effect of dissonant harmony was successfully found using the orthogonalized stimuli within the experiment, it remains unclear, unfortunately, how relevant the results are to our understanding of how harmony evokes various emotions beyond the experiment settings. Experimental approaches contrasting music vs. non-music stimuli can be seen as a "music-as-fixed-effect" fallacy, following the "language-as-fixed-effect" fallacy proposed by Clark (1973), who criticized the limited generalizability of simplistic, contrastive approaches in certain psycholinguistic research.

While not all controlled experiments suffer from limited validity, there might be difficulties stemming from their assumption that the human brain (or its behavior on average) is governed by simple, interpretable rules that can be discovered by cleverly isolated manipulations and can extrapolate to complex human behaviors (Nastase et al., 2020a). This misconception (or an arbitrary approximation) has been elegantly termed by Jolly and Chang (2019) as the "Flatland fallacy," after Edwin Abbot's famous short story (1884), which refers to a problem prevalent in psychology (and other cognitive sciences), wherein researchers misbelieve that "the parsimony offered by our low-dimensional theories reflects the reality of a much higher-dimensional problem" (Jolly and Chang (2019), p. 433). Whereas dimensionality reduction can be an efficient tool for many computational problems (i.e., "all models are wrong but some are useful" (Box, 1979), Jolly and Chang (2019) pointed out that the low-dimensional bias in cognitive sciences may be effected by "human" reasons (e.g., feelings of understanding, limitations of human cognitive capacity, cultural norms, communicating complexity) rather than computational reasons (e.g., predictive performance, computational cost). However, caution should also be taken against the humanly motivated high-dimensional bias (e.g., feelings of awe and excitement when met with an unprecedentedly large-scale artificial neural network, regardless of its efficiency).

To approach the complexity of real-world cognition and perception, naturalistic experiments are essential. The criticism against reductionism inherent in controlled experiments has a history in psychology (Brunswik, 1943; Gibson, 1978). In fact, even in neuroscience, the argument for naturalistic stimuli is not new (Barlow, 1961). One of the methodological arguments that has been discussed with regard to animal electrophysiological data (Rieke et al., 1995; Theunissen et al., 2000) has a striking resemblance to an assertion occurring in the recent discussions on human "naturalistic neuroimaging" (Sonkusare et al., 2019; Hamilton and Huth, 2020; Nastase et al., 2020a; Jääskeläinen et al., 2021): *the controlled stimulus may be too uninteresting for living animals, even for sensory neurons.* Moreover, the presumed linearity in sensory neurons may not hold given the non-linear responses to biologically salient stimuli (i.e., the sum of responses to subcomponents of a conspecific vocalization is smaller than the response to the whole vocalization in non-primary sensory neurons; Theunissen et al., 2000). Components that are uniquely responsive to speech and music (or their unique acoustic structures) found in human fMRI, electrocorticogram (ECoG), and electroencephalogram (EEG) data (Norman-Haignere et al., 2015, 2022; Zuk et al.,

2020) also suggest a strong degree of non-linearity even in the human auditory cortex[1].

The current review focuses on one of the distinguished recent advances in human neuroscience: the use of naturalistic stimuli supported by advanced computational models. Computational models provide us with not only physical descriptors but also information related to the underlying structures of natural images (Kay et al., 2008; Naselaris et al., 2009; Vu et al., 2011), natural videos (Nishimoto et al., 2011; Han et al., 2019), natural movies (Hasson et al., 2004, 2010; Hanke et al., 2014; Vodrahalli et al., 2018), and natural speech (Huth et al., 2012, 2016; Stephens et al., 2013; Mesgarani et al., 2014; Broderick et al., 2018; Nastase et al., 2020b), to name a few. In particular, this article attends to the recent investigations into the neural representation of natural music using computational models with a specific interest in musical emotion. For readers who are unfamiliar with model-based analyses, first predictive models, with mathematical descriptions, will be introduced in Section "Predictive modeling: From features to responses." Then, along with application examples, music models will be reviewed in Section "Music modeling: From stimuli to features." Finally, challenges and perspectives will be discussed in Section "Challenges and perspectives."

Predictive modeling: From features to responses

This section will guide readers from traditional linear models to non-linear models, highlighting how they relate to each other, in the context of predictive modeling. Please note that, in this section, we do not assume a specific neuroimaging modality. For the sake of discussion, let us assume that we have sufficient degrees of freedom in the temporal dimension given their respective inherent temporal dependencies and sampling rates. Exemplary applications to M/EEG and fMRI data, as will be provided throughout this section, demonstrate that a similar method can reveal different temporal scales of the brain activity of interest when applied to different data.

A model that predicts a response for a given stimulus based on an estimation how a stimulus is encoded in a system is called an encoding model. That is, the goal of an encoding model (see, e.g., Kay et al., 2008) can be understood as an identification of a transfer function that maps a given event (or a stimulus) in a physical space (e.g., time-points in audio signals) onto an evoked response in a measurement space (e.g., voxels in fMRI data or channels in M/EEG data). For instance,

$$\mathbf{y} = \mathcal{J}(\mathbf{X}) + \varepsilon \tag{1}$$

where $\mathbf{y}$ is a vector of evoked physiological responses (as a function of either stimuli or time-points) of a certain measurement unit (e.g., a voxel or a channel); $\mathbf{X}$ is a matrix that describes physical properties of stimuli in the same time scale as $\mathbf{y}$; and $\mathcal{J}(\cdot)$ is a transfer function (or a response function in the temporal domain) from stimuli to responses, which can be either linear or non-linear. The problem of estimating such transfer functions is traditionally known as *system identification* in various domains in engineering fields, including automatic control and signal processing (Zadeh, 1956; Keesman, 2011; Ljung et al., 2020). A popular approach to non-linearity in system identification, especially for naturalistic stimuli, is to approximate the system as a sequence of non-linear and linear transformations (Naselaris et al., 2011). That is, Equation 1 can be decomposed into a non-linear transform followed by a linear transform as:

$$\mathbf{y} = \mathcal{J}^{*}(\mathbf{X})\,\mathbf{w} + \varepsilon \tag{2}$$

where $\mathcal{J}^{*}(\cdot)$ is a non-linear function that maps stimuli (or time-points) from a physical space to a representational (or "feature") space and $\mathbf{w}$ is a vector of weights for a linear transform that maps stimuli (or time-points) from a representational space onto a neural measurement space. The non-linear function $\mathcal{J}^{*}(\cdot)$, which afterwards enables a linear mapping, is known as *linearization* (for a general overview, see Wu et al., 2006). That is, a linearized encoding model can still capture non-linearity while using a linear mapping between the assumed (or hypothesized) features and evoked neural responses. **Figure 1** illustrates a linearized encoding model.

For completeness, decoding models can also be seen as:

$$\mathbf{x} = \mathcal{K}^{*}(\mathbf{YV}) + \varepsilon \tag{3}$$

where $\mathbf{x}$ is a vector of the properties of stimuli with respect to a certain physical property; $\mathbf{Y}$ is a matrix of evoked physiological responses, with each column corresponds to a measurement unit; $\mathbf{V}$ is a matrix of linear decoding weights that maps neural responses to features (each column corresponds to a feature); and $\mathcal{K}^{*}(\cdot)$ is a non-linear function that maps features back to physical properties. Note that a decoding model can be converted from an encoding model based on Bayes' theorem, reflecting the prior of features (i.e., occurrence probability) in naturalistic stimuli (Naselaris et al., 2009). This is useful when the linearization function $\mathcal{J}^{*}(\cdot)$ is non-invertible (i.e., $\mathcal{K}^{*}(\cdot)$ cannot be found by the inverse of $\mathcal{J}^{*}(\cdot)$; e.g., the real absolute value function). See Naselaris et al. (2011) for a review of model-based decoding.

While the models need to be sufficiently flexible to capture the underlying transforms, the fitting (or learning) process is agonistic to the nature of variance (whether it is due to the underlying transforms or to independent noise). Therefore, a validation of an estimated model with unseen data is a crucial part of the system identification (Ljung, 2010). There are various

[1] There is a "third way" proposing controlled but naturalistic stimuli generated by sophisticated generative models such as generative adversarial networks (GANs). See Goetschalckx et al. (2021).

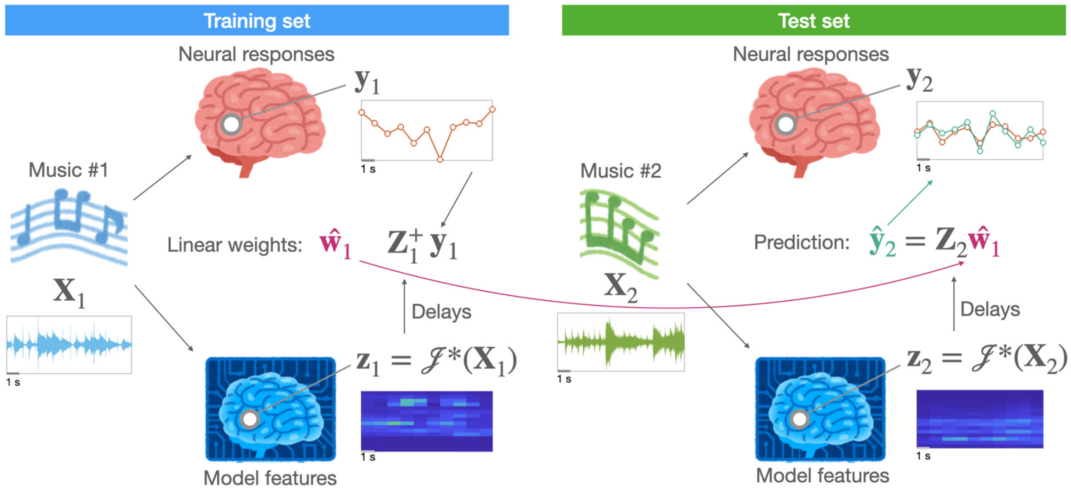
Training set
Test set
Neural responses
$\mathbf{y}_1$
1 s
Music #1
Linear weights: $\hat{\mathbf{w}}_1$ $\mathbf{Z}_1^+\mathbf{y}_1$
$\mathbf{X}_1$
Delays
$\mathbf{z}_1 = \mathcal{J}^*(\mathbf{X}_1)$
Model features
Neural responses
$\mathbf{y}_2$
Music #2
Prediction: $\hat{\mathbf{y}}_2 = \mathbf{Z}_2\hat{\mathbf{w}}_1$
$\mathbf{X}_2$
Delays
$\mathbf{z}_2 = \mathcal{J}^*(\mathbf{X}_2)$
Model features

FIGURE 1
A schematic of a linearized encoding model. A hypothesized neural representation in the brain of music (**X**) is modeled by a computation model [i.e., a linearization function $\mathcal{J}^*(\cdot)$] resulting in feature time-series (**z**). A linear relationship, described by a response function (**w**), between the delayed features (**Z**) and measured neural responses (**y**) is estimated using stimuli (e.g., music #1) in the training set. For independent stimuli (e.g., music #2) in the test set, the same computational model extracts features. Using them, the predictive performance of the hypothesized representation can be evaluated.

schemes of validation (Hastie et al., 2009), but the common idea is to test the generalizability of an estimated model, which is based on one set of data (*training set*), with another set of unseen data (*test set*) with independent and identically distributed noise (i.e., the *i.i.d.* assumption). In practice, the "unseen data" can be created by holding out some part of the data from the model estimation (i.e., cross-validation). The partitions of training sets vs. test sets can be half vs. half (split-half), k-1 parts vs. 1 part (k-fold), or all samples but one vs. the held-out one (leave-one-out). For hyperparameter optimization, the whole dataset can be divided into three parts: a training set, a validation set, and a test set that respectively comprise roughly 50, 25, and 25% of the whole set (Hastie et al., 2009). The *validation set* is so named because it is used to validate the *hyperparameters*[2]. In some cases, the training set can be split into two parts: an inner-training set and a hyperparameter validation set (nested cross-validation). Taking a larger training set (i.e., $k > 1$) would make a test set smaller for a given dataset, which would increase the sample variance in the test set. Therefore, selecting the CV scheme (i.e., determining k) is also a matter of bias-variance tradeoff. In general, 5–10-folds could result in more stable test accuracies than the leave-one-out scheme (Varoquaux et al., 2017). Critically, the partitions should be carefully designed to avoid *information leakage* between the training sets and test sets (Kaufman et al., 2012; Glaser et al., 2020). In particular, functional time series in neuroscience typically exhibits strong spatial and temporal dependencies at different scales. Moreover, intra-/inter-subject repetitions of stimuli are highly prevalent in many experiment designs, which could allow the repeated stimuli to introduce high similarity between seemingly independent time points or subjects. For example, if one randomly assigns the individual samples of EEG data to training sets or test sets, autocorrelated noise would leak into other partitions. Cross-validation is a method to test the reproducibility of a given model. Therefore, it has been argued that a statistical inference should be focused on predictive performance rather than on observed sample statistics for a reproducible science (Yarkoni and Westfall, 2017; Varoquaux and Poldrack, 2019).

Linear model

Many readers will be familiar with a multiple linear regression that models linear effects of experimental factors such as:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon \qquad (4)$$

where $\mathbf{y}$ is a $n \times 1$ vector of neural responses (of a certain measurement unit; e.g., a voxel) to a given set of stimuli, a $n \times p$ matrix $\mathbf{X}$ describes the variables (or properties) of interest of the stimuli over columns, a $n \times q$ matrix $\mathbf{Z}$ describes the nuisance variables over columns, and are $p \times 1$ and $q \times 1$ vectors of unknown coefficients, respectively, and ϵ is a $n \times 1$ vector of Gaussian noise. Equation 4 can be solved using the ordinary least squares (OLS) method when the $\mathbf{X}$ and $\mathbf{Z}$ are orthogonally created by experimenters. If we replace $\mathbf{y}$ with a matrix of

[2] In other literature (e.g., Varoquaux et al., 2017),, a "validation set" can refer to what we call a "test set" inasmuch as it validates *model weights*.

multiple units (e.g., channels, voxels) and $\boldsymbol{\epsilon}$ with a matrix of uncorrelated Gaussian noise (i.e., without modeling inter-unit dependency), the linear model is called the general linear model (GLM), which is simply a set of identical univariate models applied to multiple units (i.e., "massive univariate testing" for large-scale units; Friston et al., 1994a). For the sake of simplicity, response variables in the following are written as vectors (i.e., univariate models). However, note that they can be easily extended to their massive-univariate equivalences by concatenating response and error terms.

In case sample-wise physiological measures are directly related to time-varying stimulus properties, a proper transfer function (or a response function in the temporal domain) from neural activity to physiological measurements needs to be determined. While such a function is known to spatially and temporally vary in non-trivial ways (Aguirre et al., 1998; Handwerker et al., 2004; Badillo et al., 2013; Taylor et al., 2018), if we assume such a transfer function h, Equation 4 can be rewritten (Friston et al., 1994b) as:

$$\mathbf{y} = h * [\,\mathbf{X}\beta \quad \mathbf{Z}\gamma\,] + \varepsilon \tag{5}$$

where * denotes a convolution in the temporal domain and ε is a Gaussian noise with serial-correlation, which is typically non-zero in most of non-invasive measures. Note that all vectors and matrices are now defined for each sample (i.e., time-point) of the physiological measurement (i.e., the number of rows is the number of samples, t). Also, note that both $\mathbf{X}$ and $\mathbf{Z}$ describe stimulus properties so that they can be concatenated to apply the same convolution. Equation 5 can be solved using a variant (due to the serial-correlation) of OLS such as weighted least squares (WLS) (Friston et al., 2006; Poldrack et al., 2011).

Note that a linear model with OLS can also be used as a predictive model and be cross-validated. As long as the variables are orthogonal (or minimally intercorrelated), the OLS is an unbiased estimator for the training set. However, the estimates may not apply to the test set because the OLS will also fit the noise, together with the signal, in the training set. Regularization (see Section "Regularized linear methods") can be useful in such cases. When a linear model is used as a predictive model, the inference will be on whether the prediction accuracy (e.g., Pearson correlation between prediction and observation) is above chance level as opposed to whether the estimated contrast (e.g., a difference between condition A vs. condition B) is above chance level.

Reverse correlation

While Equation 5 can still model delayed processes between neural activity and measurement by a fixed physiological transfer function h, it cannot flexibly model delayed processes between stimulus and neural activity. A system identification method—also known as *reverse correlation* or *triggered correlation*, for the analysis averages stimuli based on response as opposed to averaging responses based on stimuli)—uses autocorrelation (or autocovariance) of the stimuli and the cross-correlation (or cross-covariance) between the stimuli and responses to capture the delayed responses in a response function (i.e., a linear filter) under the assumption of the system (i.e., a linear time-invariant [LTI] system). This was introduced in electrophysiology as a receptive field mapping technique by presenting white noise (instead of many narrow-band filtered signals) and estimating the frequency selectivity of individual neurons (Boer and Kuyper, 1968).

When we assume that a set of physical properties of stimuli that are relevant to the neural system of interest (i.e., often called *features*) $\left\{\boldsymbol{f}_1,\boldsymbol{f}_2,\ldots,\boldsymbol{f}_p\right\}$ are given—such as, e.g., narrow-band filtered acoustic energy of the presented white noise)—then Equation 5 can be rewritten as:

$$\mathbf{y} = \begin{bmatrix} h_1 * \boldsymbol{f}_1\beta_1 & h_2 * \boldsymbol{f}_2\beta_2 & \cdots & h_p * \boldsymbol{f}_p\beta_p \end{bmatrix} + \varepsilon$$

where h_i is a feature-specific convolutional kernel, or a transfer function, for the i-th feature; $\boldsymbol{f}_i$ is a $n \times 1$ vector of the i-th feature; and the effect size β_i is now simply a signed amplitude of the h_i, which is in fact redundant. Therefore, we can further simplify:

$$\mathbf{y} = \begin{bmatrix} h_1 * \boldsymbol{f}_1 & h_2 * \boldsymbol{f}_2 & \cdots & h_p * \boldsymbol{f}_p \end{bmatrix} + \varepsilon \tag{6}$$

In the case of discrete signals, the convolution above (Equation 6) can be rewritten as a multiplication of delayed features $\mathbf{G}$ and finite impulse response (FIR) functions $\mathbf{w}$ defined over finite delays $\left\{l_1, l_2, \ldots, l_d\right\}$ as:

$$\mathbf{y} = \mathbf{G}\mathbf{w} + \varepsilon \tag{7}$$

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_2 & \ldots & \mathbf{F}_p \end{bmatrix} \in \mathbb{R}^{n \times pd},$$

$$\mathbf{F}_{\mathrm{i}} = \begin{bmatrix} f_i(t_1 - l_1) & \cdots & f_i(t_1 - l_d) \\ \vdots & \ddots & \vdots \\ f_i(t_n - l_1) & \cdots & f_i(t_n - l_d) \end{bmatrix} \in \mathbb{R}^{n \times d},$$

i.e., $[\mathbf{F}_{\mathrm{i}}]_{j,k} = f_i(t_j - l_k)$, $f_i(t)$ is the element of the i-th feature vector at the timepoint t, $\mathbf{y}$ is a $n \times 1$ vector of neural responses from the timepoint t_1 to t_n, $\mathbf{w} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{\mathrm{p}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{pd \times 1}$, and $\boldsymbol{u}_i$ is a $d \times 1$ vector of a discrete response function to estimate. If the delays $\left\{l_1, l_2, \ldots, l_d\right\}$ are adjacent to each other in the sample space (i.e., $l_{j+1} - l_j = t_{i+1} - t_i$), $\mathbf{F}_{\mathrm{i}}$ is a $n \times d$ rectangular Toeplitz matrix (i.e., $t_i - l_j = t_{i+1} - l_{j+1}$). Note that the FIR model (Equation 7) has been used to fit a physiological response function (e.g., a hemodynamic function) to the data in controlled experiments (Henson et al., 2001). In reverse correlation, the FIR function is used to further include the transform from stimulus to neural activity in addition to the transform from neural activity to non-invasive measurement (e.g., fMRI image intensity or scalp-EEG potential).

If $\mathbf{G}$ is fully ranked—i.e., features are not correlated, such as white noise—then we can use the OLS to estimate $\boldsymbol{w}$:

$$\widehat{\mathbf{w}} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{y} = \mathbf{G}^{+}\mathbf{y} \tag{8}$$

where the hat operator $\widehat{\cdot}$ denotes an estimation; $\cdot^{\mathrm{T}}$ denotes a matrix transposition; $\cdot^{-1}$ denotes a matrix inversion; and $\cdot^{+}$ denotes Moore-Penrose pseudoinversion[3]. Note that the OLS solution minimizes the prediction error (i.e., loss function) on the training set itself given as:

$$\mathrm{L}\left(\mathbf{w};\mathbf{y}\right) = \|\,(\mathbf{y}-\mathbf{G}\mathbf{w})^{\mathrm{T}}(\mathbf{y}-\mathbf{G}\mathbf{w})\,\|_2 = \|\,\mathbf{y}-\mathbf{G}\mathbf{w}\,\|_2^2$$

For a discrete timeseries $\mathbf{a} \in \mathbb{R}^n$, an autocovariance matrix for d lags is given by $\mathbf{A}^{\mathrm{T}}\mathbf{A} \in \mathbb{R}^{d\times d}$ where $\mathbf{A}$ is a $n \times d$ Toeplitz matrix. With another timeseries $\mathbf{b} \in \mathbb{R}^n$, a cross-covariance matrix of $\mathbf{a}$ and $\mathbf{b}$ is given by $\mathbf{A}^{\mathrm{T}}\mathbf{B} \in \mathbb{R}^{d\times p}$ where $\mathbf{A}$ and $\mathbf{B}$ are $m \times d$ and $m \times p$ Toeplitz matrices, respectively. Therefore, Equation 8 can be rewritten as:

$$\widehat{\mathbf{w}} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{y} = \mathbf{C}_{\mathbf{GG}}{}^{-1}\mathbf{c}_{\mathbf{Gy}} \tag{9}$$

where $\mathbf{C}_{\mathbf{GG}}$ is a $pd \times pd$ autocovariance matrix and $\mathbf{c}_{\mathbf{Gy}}$ is a $pd \times 1$ cross-covariance vector since $\mathbf{y}$ is not delayed. Note that this expression highlights the fact that the method *decorrelates* the stimulus-response cross-covariance with the autocovariance of the stimulus itself. If the features do not have any autocovariance structures—for example, if a single predictor is given by white noise—then the sample autocovariance matrix of the predictors can be very close to the identity matrix: $\mathbf{C}_{\mathbf{GG}} \approx \mathbf{I}$.

Note that the reverse correlation method itself does not require regularization when the stimulus is well-behaving Gaussian. Therefore, for linear systems, or for systems that can be well approximated by linear models, the OLS solution is sufficient. However, this is not the case in many real-world systems including the human brain (and even the auditory neurons as discussed earlier).

Regularized linear methods

In general, regularization serves two purposes: one is to avoid overfitting, even for a model with a single predictor, by penalizing overly complex models; and the other is to deal with strong multicollinearity present in the predictors (i.e., an ill-posed inverse problem). Note that even for a single-feature model (i.e., $p = 1$ in Equation 6), multicollinearity may exist across delayed features (i.e., columns of $\mathbf{G}$ in Equation 7) if serial-correlation is present in the feature. Now that the $\mathbf{C}_{\mathbf{GG}}$ in Equation 10 can be very different from $\mathbf{I}$ and non-invertible, it is necessary to introduce regularization to make it invertible. Tikhonov regularization (Tikhonov, 1943; Tikhonov et al., 1995) is a general solution with a regularization matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{pd\times pd}$:

$$\widehat{\mathbf{w}}^{*}(\boldsymbol{\Lambda}) = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G} + \boldsymbol{\Lambda}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{y} \tag{10}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}$ when the loss function to minimize is defined with the L2 norm penalty as:

$$\mathrm{L}\left(\mathbf{w};\mathbf{y},\boldsymbol{\Gamma}\right) = \|\,\mathbf{y}-\mathbf{G}\mathbf{w}\,\|_2^2 + \|\,\boldsymbol{\Gamma}\mathbf{w}\,\|_2^2$$

Ridge regression is a special case of Tikhonov regularization where $\boldsymbol{\Lambda} = \lambda\mathbf{I}$ and λ is a regularization scalar (Hoerl and Kennard, 1970):

$$\widehat{\mathbf{w}}^{*}(\lambda) = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G} + \lambda\mathbf{I}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{y} = (\mathbf{C}_{\mathbf{GG}} + \lambda\mathbf{I})^{-1}\mathbf{c}_{\mathbf{Gy}} \tag{11}$$

Note that the ridge solution (and its prediction performance) is a function of the regularization. The regularization controls the flexibility of the model, which impacts the bias (i.e., the expected distance between "true" parameters and estimates *across multiple experiments*) and variance (i.e., the spread of estimates across multiple experiments) of the solutions in opposite ways. For example, estimates from an extremely rigid model, such as the 0-th order model that always returns a constant, will be highly biased but have no variance. In other words, it will be wrong but in a very consistent fashion. On the other extreme, a flexible model will be minimally biased (i.e., accurate on average), but largely varied. That is, it can be sometimes very accurate, but it can also be very wrong, depending on the realization of random noise. The "best" regularization depends on the "true" structure of the system (Ljung et al., 2020), which is unknown. Therefore, in practice, the regularization is "optimized" to balance the tradeoff between bias and variance for specific datasets. Several optimization methods have been used for fMRI and M/EEG data acquired while listening to natural speech and music: e.g., ridge tracing (Santoro et al., 2014; Moerel et al., 2018), bootstrapping (Huth et al., 2016), and nested cross-validation (Daube et al., 2019).

[3] In Penrose (1955), Corollary 1 states that the general solution of a linear model $\mathbf{y} = \mathbf{X}\mathbf{b}$ is $\hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y} + \left(\mathbf{I} - \mathbf{X}^{+}\mathbf{X}\right)\boldsymbol{a}$ where $\boldsymbol{a}$ is arbitrary (notations are rewritten for consistency with the current article). For real-valued, full-ranked $\mathbf{X}$, $\mathbf{X}^{+} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. Thus, $\mathbf{X}^{+}\mathbf{X} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$, leaving only the first term of the solution: $\hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y}$.

From a Bayesian perspective, the Tikhonov regularization can be seen as multivariate normal priors on the "distribution of weights" (Nunez-Elizalde et al., 2019) as:

$$\mathbf{w} \sim \mathcal{N}_{pd}\left(0, \lambda^{-2}\boldsymbol{\Sigma}\right) \tag{12}$$

where $\mathcal{N}_{pd}$ is a pd-dimensional multivariate normal distribution and $\Sigma \in \mathbb{R}^{pd \times pd}$ is the positive definite prior covariance matrix, and λ is a scalar regularization parameter. The ridge regularization can be seen as a special case of spherical priors (i.e., $\boldsymbol{\Sigma} = \mathbf{I}$) whereas other forms of regularization can be seen as non-spherical priors (Nunez-Elizalde et al., 2019). The maximum likelihood solution of the problem can be found in a closed form as a Tikhonov solution (Nunez-Elizalde et al., 2019):

$$\widehat{\mathbf{w}}_T(\lambda, \boldsymbol{\Gamma}) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

where $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}$. This is equal to a ridge solution when $\boldsymbol{\Gamma} = \mathbf{I}$ (thus, $\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma} = \mathbf{I} = \boldsymbol{\Sigma}$). The Tikhonov solution can be simplified (Nunez-Elizalde et al., 2019) by a linear transform such that:

$$\mathbf{A} = \mathbf{X}\,\boldsymbol{\Gamma}^{-\mathbf{1}}$$

Then, a ridge solution with $\mathbf{A}$ is given as:

$$\widehat{\mathbf{w}}_A(\lambda) = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2\mathbf{I}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}$$

The estimates can be projected back into the original space, which finally gives us the Tikhonov solution:

$$\widehat{\mathbf{w}}_T(\lambda, \boldsymbol{\Gamma}) = \boldsymbol{\Gamma}^{-1}\widehat{\mathbf{w}}_A(\lambda)$$

Now, the prior matrix $\boldsymbol{\Sigma}$ (or the inverse transform matrix $\boldsymbol{\Gamma}$) can be based on stimulus models (e.g., a semantic embedding), physiological models (e.g., hemodynamic response function [HRF]), or appreciation of different scales of features, i.e., independently regularizing features or feature spaces, because the globally optimal regularization could be suboptimal for individual features (i.e., "banded ridge"; Nunez-Elizalde et al., 2019). In fact, the last usage is widely known as "multi-penalty ridge." It was first proposed in the original publication of ridge regression (Hoerl and Kennard, 1970) in sections 5 (p. 63) and 7 (p. 65), where a regularization parameter ("k_i") is found for each column of the orthogonalized design matrix (i.e., canonical variates). The multi-penalty ridge can be seen as a general case of ridge where the Tikhonov regularization matrix (Equation 10) is given as:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1\mathbf{I}_1 & \cdots & \mathbf{0}_k \\ \vdots & \ddots & \vdots \\ \mathbf{0}_1 & \cdots & \lambda_1\mathbf{I}_k \end{bmatrix}$$

where λ_i is a scalar hyperparameter for the i-th feature space, $\mathbf{I}_i \in \mathbb{R}^{p_i d \times p_i d}$ is an identity matrix for the i-th feature space with p_i as the number of columns of the i-th feature space, and $\mathbf{0}_i \in \mathbb{R}^{p_i d \times p_i d}$ is a zero square matrix. Recent studies optimized a regularization parameter for each feature space to perform model comparisons without *over-regularization*, i.e., suboptimal regularization for specific features; see, e.g., Daube et al. (2019) and Sohoglu and Davis (2020).

However, the optimization of p hyperparameter (i.e., determination of regularization parameters) via grid-search would have a complexity that is proportional to j^k for j grid points and k hyperparameters, which rapidly makes the optimization intractable. There are fast algorithms for multi-penalty ridge problems that are inspired by the original formulation (Hoerl and Kennard, 1970) and where the design matrix is first orthogonalized to reduce the number of necessary hyperparameters [see, e.g., van de Wiel et al. (2021) for applications on the "large p, small n" genomic data].

Besides the ridge penalty, other types of penalty terms are also commonly used such as lasso (i.e., L1 norm penalty; Tibshirani, 1996), which has been used for naturalistic speech MEG data (Brodbeck et al., 2018a,b), and elastic net (i.e., both of L1 and L2 penalties; Zou and Hastie, 2005).

Non-linear kernel methods

So far, we have discussed linear methods. However, there are various methods for handling non-linearity, even in the field of traditional machine learning and statistical learning. Kernel-based machine learning methods, such as the *support vector machine* (SVM; Boser et al., 1992) can be seen as a linearization of features in the sense that it provides non-linear transformation of original predictors into a high-dimensional feature space where a linear fit can be useful (Bishop and Nasrabadi, 2006). The idea is that some non-linear-looking problems can be seen as linear in a higher-dimensional space. Let such a mapping from a lower-dimensional original predictor space to a higher-dimensional feature space (i.e., *features space mapping*) be $\phi : \mathbb{R}^p \to \mathbb{R}^q$ where $p < q$. Equation 6 then can be defined in the *feature* space:

$$\mathbf{y} = \left[\ \phi\left(\boldsymbol{g}_1\right)^{\mathrm{T}}\ \phi\left(\boldsymbol{g}_2\right)^{\mathrm{T}}\ \ldots\ \phi\left(\boldsymbol{g}_n\right)^{\mathrm{T}}\ \right]^{\mathrm{T}}\mathbf{w} + \varepsilon \tag{13}$$

where $\boldsymbol{g}_i$ is the i-th row vector in the matrix $\mathbf{G} \in \mathbb{R}^{n \times pd}$ from Equation 7.

Now the problem becomes how we can find the feature space mapping in unknown high (theoretically infinite) dimensions. Fortunately, instead of explicitly finding this mapping to unknown dimensions, it has been shown that the prediction can nonetheless be made without knowing the mapping itself, but rather with a kernel function k, which is an inner product of transformed features: $k\left(\boldsymbol{g}_1, \boldsymbol{g}_2\right) = \left\langle \phi\left(\boldsymbol{g}_1\right), \phi\left(\boldsymbol{g}_2\right)\right\rangle$ (Bishop and Nasrabadi, 2006). A prediction on the new datapoint $\boldsymbol{g}_x$ with a regularization λ can be given without explicitly knowing the mapping $\phi(\cdot)$ as:

$$\hat{\mathbf{y}}\left(\boldsymbol{g}_x\right) = \phi\left(\boldsymbol{g}_x\right)\hat{\mathbf{w}} = k\left(\boldsymbol{g}_x\right)\left(\mathbf{K}_{\mathrm{t}} + \lambda\mathbf{I}_n\right)^{-1}\boldsymbol{y}_t \tag{14}$$

where the Gram matrix $[\mathbf{K}_{\mathrm{t}}]_{i,j} = k(\boldsymbol{g}_i, \boldsymbol{g}_j)$ is defined over all n datapoints in the training set, $\mathbf{I}_n \in \mathbb{R}^{n\times n}$, $\boldsymbol{y}_t = \left[\, y_1 \; y_2 \; \cdots \; y_n \,\right]^{\mathrm{T}}$. This substitution (also known as *kernel trick*) of the unknown mapping with the Gram matrix that is defined between datapoints makes the problem tractable. The kernel function can be constructed with non-linear basis functions such as the p-th order polynomial with a constant term c: $k\left(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}}\right) = \left(\mathbf{x}_{\mathrm{i}}^{\mathrm{T}}\mathbf{x}_{\mathrm{j}} + c\right)^p$, Gaussian basis function: $k\left(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}}\right) = \exp\left(-\parallel \mathbf{x}_{\mathrm{i}} - \mathbf{x}_{\mathrm{j}} \parallel^2 / 2\sigma^2\right)$, or the radial basis function: $k\left(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}}\right) = \exp(-\gamma \parallel \mathbf{x}_{\mathrm{i}} - \mathbf{x}_{\mathrm{j}} \parallel^2)$ (Bishop and Nasrabadi, 2006). The hyperparameters of the kernels can be validated through cross-validation in practice (Chu et al., 2011). Kernel regression (alibeit with linear kernels) has also been used for auditory encoding models (De Angelis et al., 2018; Erb et al., 2019; Rutten et al., 2019).

Related to the kernel trick, *representational similarity analysis* (RSA) compares a kernel of the brain with a kernel of a reference model (Kriegeskorte et al., 2008). RSA does not attempt to explicitly identify the transfer functions of a system (i.e., the first-order isomorphism between the physical properties and representations in the brain), but it can query whether systems share similar non-linear mappings of the identical set of stimuli (i.e., the second-order isomorphism between representations in different systems), which establishes important foundations for understanding the human brain as a non-linear system.

Neural network methods

Recently, with increased computational capacity under Moore's law and large-scale (i.e., petabytes) data (Sun et al., 2017), modern artificial neural network (ANN) models have outperformed traditional models (including traditional ANNs) and are reaching human-level performances in many tasks such as semantic visual recognition (Donahue et al., 2014), language generation (Floridi and Chiriatti, 2020), and even a specific scientific discovery activity (Jumper et al., 2021). In general, modern ANN models, also known as deep neural networks (DNNs), have multiple ("deep") layers of units (e.g., perceptrons) and include iterative adaptive processes, which allow a model to update (or "learn") its parameters (or weights) based on the binary labels (or continuous values) given by humans (supervised learning) or real data (unsupervised learning). A perceptron can be seen as a linear binary classifier (Bishop and Nasrabadi, 2006) as:

$$f\left(\mathbf{x}_i\right) = \mathrm{sign}(\mathbf{x}_i\mathbf{w}) \tag{15}$$

where $\mathrm{sign}\,(\mathrm{a}) = \begin{cases} +1 \;,\; a \geq 0 \\ -1 \;,\; a < 0 \end{cases}$, $\mathbf{x}_i$ is a $1 \times p$ vector of features of the i-th instance (e.g., a stimulus), $\mathbf{w}$ is a $p \times 1$ vector of weights. As seen earlier (Equation 13), a non-linear transformation $\phi(\cdot)$ can be introduced:

$$f\left(\mathbf{x}_i\right) = \mathrm{sign}(\phi(\mathbf{x}_i)\mathbf{w}) \tag{16}$$

By combining such simple units, a network can be very flexible. For example, a prediction from a two-layer network can be expressed (Bishop and Nasrabadi, 2006) as:

$$\hat{y}\left(\mathbf{x}_i, \boldsymbol{\omega}\right) = \sigma\left(h\left(\mathbf{x}_i\mathbf{w}^{(1)}\right)\mathbf{w}^{(2)}\right) \tag{17}$$

where $\sigma(\cdot)$ is a sigmoid function, $h(\cdot)$ is a hyperbolic tangent function, $\mathbf{w}^{(i)}$ is a vector of weights at the i-th layer, and $\boldsymbol{\omega} = \begin{bmatrix} \mathbf{w}^{(1)} \\ \mathbf{w}^{(2)} \end{bmatrix}$ is a vector of all weights. (i.e., a set of all weight vectors). The loss function of this network can be defined as a sum-of-squares error function:

$$\mathrm{L}\left(\boldsymbol{\omega}; \mathcal{M}\right) = \sum_{\mathbf{x}_i \in \mathcal{M}} \parallel \hat{y}\left(\mathbf{x}_i, \boldsymbol{\omega}\right) - y_{\mathrm{i}} \parallel^2 \tag{18}$$

where $\mathcal{M} = \{\mathbf{x}_1, \; \mathbf{x}_2, \; \ldots, \; \mathbf{x}_n\}$ is a training set and y_i is the true response (or label) of x_{i}. Given the non-linearity, the loss function cannot be analytically minimized. However, the loss function can still guide the model to adjust weights to reduce errors for the next example. This process is called backpropagation (Amari, 1967; Werbos, 1974, 1994). One of the widely used approaches is called *gradient descent* optimization (Bishop and Nasrabadi, 2006). The idea is that, even without knowing the loss function analytically, one can still empirically minimize it by perturbating each weight and figuring out which direction in the weight space would decrease, or at least not increase, the loss function. More formally, one can compute partial derivative with respect to changes of individual weight for the i-th data point: $\frac{\partial \mathrm{L}(\boldsymbol{\omega}; \mathrm{x}_i)}{\partial w_j^{(i)}}$ where $w_j^{(i)}$ is the j-th weight in the i-th layer. A vector of partial derivatives is called a *gradient*. For a function f that maps a n-dimensional real vector $\mathbf{x} = \left[\, x_1 \; x_2 \; \cdots \; x_n \,\right]$ to a real scalar (i.e., $f : \mathbb{R}^n \rightarrow \mathbb{R}$), the gradient of f is defined as: $\nabla f\left(\mathbf{x}\right) := \left[\, \frac{\partial f}{\partial x_1} \; \frac{\partial f}{\partial x_2} \; \cdots \; \frac{\partial f}{\partial x_n} \,\right]$. Thus, a gradient is a "direction" vector (in the weight space) that maximizes a given function. Therefore, we would like to update our weight by subtracting the gradient but by a small magnitude. That is, given the i-th training data point, the next weight can be expressed as:

$$\boldsymbol{\omega}_{i+1} = \; \boldsymbol{\omega}_i - \gamma\nabla\mathrm{L}\left(\boldsymbol{\omega}_i; \mathbf{x}_i\right) \tag{19}$$

where $\boldsymbol{\omega}_{i+1}$ is the weights learned from the i-th data point, γ is a control parameter called a *learning rate*.

DNN models learn weights in various forms of architectures consisting of multiple well-known structures with several practical modifications. For example, a convolutional neural network (CNN; LeCun et al., 1989) is a multilayer architecture that includes convolutions in the input space (e.g., with respect to horizontal and vertical axes in 2-D images; with respect to time dimension in the audio waveform) to exploit a topographical organization (i.e., local dependency) of the data. This can be seen as a strong prior (i.e., "geometric knowledge about the task"; LeCun et al., 1989, p. 550) that completely disconnects some connections (Goodfellow et al., 2016).

Another fundamental architecture known as a recurrent neural network (RNN) (Rumelhart et al., 1986) was developed to learn structures in sequential data such as language (e.g., word sequences). In this network, a node gets inputs from a node at a previous time point as well as the current stimulus at the current time point, returns outputs for the current time point, and feeds an input to a node at the next time point.

Recently, DNN models have been widely used in finding a non-linear relationship in human neuroimaging data. However, an improvement by using non-linear models over linear models requires a high SNR and/or a very large sample size as Gaussian noise can linearize the decision boundary (Schulz et al., 2020). Except for a few consortia (e.g., UK Biobank, Human Connectome Project), large-scale functional data, especially with naturalistic stimuli, are scarce in comparison to behavioral data (e.g., crowdsourced tagging data for millions of songs). Therefore, in many applications for naturalistic stimuli, DNN models are trained to replicate large-scale human behavioral data, then models' representations (i.e., linearized features) are related to smaller-scale human neural data via regularized linear models (Agrawal et al., 2014; Güçlü and van Gerven, 2015; Güçlü et al., 2016; Caucheteux and King, 2022) or representational similarity analysis (Khaligh-Razavi and Kriegeskorte, 2014; Kell et al., 2018). In particular, CNN models mimicked human auditory behaviors (e.g., pitch [F0] perception, word recognition, musical genre recognition) and neural responses (Kell et al., 2018; Schulz et al., 2020), arguing for a representational gradient across the superior temporal gyrus/sulcus (Güçlü et al., 2016). While the CNN is considered to be one of the greatest achievements of neuromorphic engineering (i.e., a system that is inspired by the hierarchical structure of the sensory system of brains, performing perceptual tasks at a near-human-level) and has shown a partial convergence with neural representations in various sensory modalities (Agrawal et al., 2014; Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015; Kell et al., 2018), their implication (i.e., whether they can be accepted as evidence for a mechanical model of a human sensory system) is still under debate (see Section "Challenges and perspectives").

Music modeling: From stimuli to features

This section discusses recent endeavors in modeling the structure of music. From the perspective of encoding models, a computational music model can be seen as a hypothesized linearization and an inference on the model performance (and additional contributions of individual sets of features) would be equivalent to effect testing in controlled experiments. Models at various levels and their applications to natural music will be reviewed.

Auditory models

Models based on psychophysics and electrophysiology have been developed to simulate the neural activity of the auditory pathway (Chi et al., 1999; Klein et al., 2000). Among various formulations, a MATLAB implementation, namely Neural Systems Laboratory (NSL) tools[4] (Chi et al., 2005) was created based on electrophysiological findings. In the first stage, the *auditory spectrogram* is computed, which is a time-frequency representation of given sound signals. The basilar membrane filter bank is modeled by bandpass filters. The outputs are further processed accounting for various non-linear transforms through the auditory nerves. Then a lateral inhibition in the cochlear nucleus is simulated by the first-order derivative across frequency channels. This cascade model can be described as follows:

(1) cochlear filter bank:

$$y_c\left(t,f\right) = s\left(t\right)^{*} h\left(t;f\right)$$

where $y_c\left(t,f\right)$ is a cochlear output at time t and frequency channel f (often referred to as [simulated] *cochleogram*), $s\left(t\right)$ is a signal at time t, * is a convolution operator in the temporal domain, $h\left(t;f\right)$ is a response function of the f-th frequency channel (i.e., a membrane filter),

(2) auditory nerve:

$$y_a\left(t,f\right) = g\left(\partial_t y_c\left(t,f\right)\right)^{*} w\left(t\right),$$

where $y_a(t,f)$ is an auditory-nerve output, $g\left(\cdot\right)$ is a non-linear compression (i.e., gain), ∂_t denotes partial differential with respect to time as a high-pass filter, $w\left(t\right)$ is a low-pass filter that mimics the decrease of phase-locking above 2 kHz,

(3) cochlear nucleus:

$$y_l\left(t,f\right) = \max\left(\partial_f y_a\left(t,f\right), 0\right)$$

where $y_l(t,f)$ is a cochlear-nucleus output, ∂_f denotes partial differential with respect to frequency to mimic the lateral

[4] http://nsl.isr.umd.edu/downloads.html

inhibition in the cochlear nucleus, $\max(\cdot, 0)$ is a half-wave positive rectifier,

(4) and, finally, midbrain:

$$y_m\left(t,f\right) = y_l\left(t,f\right)^{*}\mu(t;\tau) \tag{20}$$

where $y_m\left(t,f\right)$ is the midbrain (final) output, and $\mu(t;\tau)$ is a short-time ($\tau = 8$ ms) integration window (for further loss of phase-locking in the midbrain).

Then, in the second stage, the *cortical representation* is computed by a 2-D convolution of an spectrotemporal respective field (STRF) filter bank and the auditory spectrogram. For a specific "cell" that is sensitive to a specific combination of spectral and temporal modulations (i.e., "spatial" components in the spectrogram) and a direction (downward or upward), the cortical representation z is given as:

$$z_{c\Downarrow(\Uparrow)}\left(t,f;\,\omega_c,\Omega_c,\theta_c,\ \phi_c\right) = y_m\left(t,f\right)\otimes \mathrm{STRF}_{c\Downarrow(\Uparrow)} \tag{21}$$

where $c\Downarrow(\Uparrow)$ denotes a downward (or upward) cell c, ω_c is the temporal modulation rate (i.e., ripple velocity; in Hz) of the cell c, $\boldsymbol{\Omega}_c$ is the spectral modulation scale (i.e., ripple density; in cycles/octave), θ_c is the rate phase, ϕ_c is the scale phase, and $\otimes$ denotes 2-D convolution. $\mathrm{STRF}_{c\Downarrow(\Uparrow)}$ is defined as the real part of the product of two complex functions describing temporal modulations (ripples along the time axis in the auditory spectrogram) and spectral modulations (ripples along the frequency axis) as:

$$\mathrm{STRF}_{c\Downarrow(\Uparrow)} = \mathcal{R}\left\{h_{IRT}^{(*)}\left(t;\omega_c,\theta_c\right)\cdot h_{IRS}\left(t;\Omega_c,\phi_c\right)\right\}$$

where $\mathcal{R}\{\cdot\}$ denotes the real part, h_{IRS} and h_{IRT} denotes complex impulse response functions for temporal modulation and spectral modulation, respectively, and the superscripted $*$ denotes the complex conjugate, which allows for differentiation of downward and upward tone sweeps. Because the third and fourth quadrants can be constructed using complex conjugates of the STRFs in the first two quadrants, we only consider those two quadrants (1st: positive rate, positive scale, downward; 2nd: negative rate, positive scale, upward). Further details can be found in Chi et al. (2005).

While this model has been widely used for various short (1–2 s) naturalistic audio clips and speech (Moerel et al., 2013; Santoro et al., 2014; Khalighinejad et al., 2019; Sohoglu and Davis, 2020), applications to naturalistic music remain relatively sparse: ECoG data over the left frontotemporal areas while a highly trained pianist playing 2-min classical pieces with and without auditory feedback (Martin et al., 2017), ECoG data over bilateral frontotemporal regions from 29 patients while listening to a 3-min song with vocals (Bellier et al., 2022), and whole-brain fMRI data while listening to five-hundred-forty 15-s excerpts from the GTZAN (G. Tzanetakis and P. Cook) Musical Genre Dataset (Nakai et al., 2021). These studies consistently demonstrated that the auditory models (the auditory and cortical representations) successfully predicted neural responses to natural music in the primary and non-primary auditory cortices. A preliminary report suggested redundant information is encoded in non-auditory regions such as sensory-motor cortices and the inferior frontal gyrus (Bellier et al., 2022).

Music information retrieval models

Music information retrieval (MIR) is an interdisciplinary field that has emerged with the arrival of the electronic music distribution systems (e.g., *Napster* in 1999, *iTunes* in 2001, and *Spotify* in 2006) (Aucouturier and Bigand, 2012). Therefore, its main focus is to develop technologies that are useful for such services including searching, organizing, accessing, and processing digitized music signals and related data. Compared to music psychology, which focuses more on processes via simplistic examples, the MIR focuses more on the "end-to-end" results (i.e., linking physical characterizations of music signals and population behaviors, possibly in the music market) (Aucouturier and Bigand, 2012). Nonetheless, the MIR field has discovered multiple acoustic features that are predictive of "perceived emotions" in music (Yang and Chen, 2011), which are not necessarily experienced by listeners, but are recognizable. These findings motivated psychologists and neuroscientists to investigate whether this functional (end-to-end) relationship implies anything for the internal processes in listeners, even at a population level, because this may have relevance to "experienced emotions."

Traditional MIR models, as opposed to recent ANNs, tend to have a couple hundred (counting all subcomponents) "hand-crafted" features. Some of the features in MIRtoolbox (Lartillot and Toiviainen, 2007) are briefly explained here for a discrete signed signal $s(x)$ of t timepoints:

- Root-mean-square (RMS) envelope: $E = \sqrt{\frac{1}{t}\sum_{x=1}^{t} s\left(x\right)^2}$ for global energy, $E(w) = \sqrt{\frac{1}{n}\sum_{x=1}^{n} w\left(x\right)^2}$ with a windowed signal w(x) of n ($< t$) timepoints for local energy.

- Zero-crossing rate: z = $\frac{1}{t-1}\sum_{x=1}^{t-1}\mathcal{J}\left(s\left(x\right)s\left(x-1\right)\right)$ where $\mathcal{J}\left(y\right) = \begin{cases} 1, & y < 0 \\ 0, & y \geq 0 \end{cases}$, which can also be locally calculated for a given windowed signal. In a simplistic case (e.g., a sine wave), this can be used to estimate the fundamental frequency. However, more generally, it describes some aspects of timbral quality rather than pitch.

- Spectrogram: $\mathbf{S}\left(w,f\right) = \boldsymbol{f}\left(k\right)H(k;f)$ with $\boldsymbol{f}$(k) the magnitude of a discrete Fourier transform (i.e., fast Fourier transform) of a given windowed signal $\boldsymbol{f}\left(k\right) = \left|\sum_{m=0}^{n-1} w\left(x\right)\exp(-2\pi ikm/n)\right|,\ k = 0,..,\ n/2$ and $H(k;f)$ is a transfer function of a given filter bank for a characteristic frequency in a linear scale (Hz). Various filter banks based on behavioral psychophysical experiments have been used: e.g., Mel-scale and Bark-scale. The logarithm of magnitude is more often used. This describes acoustic energy decomposed in frequency bands.

- Cepstral coefficient: $\mathbf{C}\left(w,k\right)=\sum_{f=0}^{p-1}\mathbf{S}\left(w,f\right)\cos\left[\frac{\pi}{n}\left(f+\frac{1}{2}\right)k\right],\quad k=0,\ldots,p-1$ for a spectrogram defined over p frequency bands. Because in most of natural stimuli, spectrograms have high spectral dependency (magnitudes in adjacent channels are similar), discrete cosine transform is used to further compress and orthogonalize the spectrogram.

- Spectral flux: $L\left(w_i,w_{i-1};\mathbf{S}\right)=\|\mathbf{S}(w_i)-\mathbf{S}(w_{i-1})\|_2$ with $\mathbf{S}(w_i)$ a spectrum (or a cepstrum) for the i-th window. It could be more sensitive to frequency-specific energy changes compared to the RMS. Given that natural musical pieces could have widely various spectra (i.e., unnormalized), in some cases, this metric could mainly reflect the spectral density.

- Spectral centroid: $N\left(w;S\right)=\sum_{f=0}^{F-1}K(f)\mathbf{S}(w,f)/\sum f=0^{F-1}\mathbf{S}(w,f)$ for a given filter bank S that constructs the spectrogram $\mathbf{S}(w,f)$ over p frequency bins with characteristic frequencies $K(f)$. That is, a weighted average of characteristic frequencies $K(f)$ of where the weights are normalized magnitudes.

- Key clarity: $K\left(\mathbf{S}\right)=\max_{i\in\mathcal{K}}\operatorname{corr}\left(\psi_i,\theta(\mathbf{S})\right)$ where $\mathcal{K}$ is a set of all 12 major and 12 minor keys, ψ_i is the tonal stability profile of the i-th key (Krumhansl and Shepard, 1979), and $\theta\left(\cdot\right)$ is a chromagram of a given spectrogram (a power spectrum for 12 pitch classes based on the standard 440-Hz tuning). The key similarity (i.e., $\operatorname{corr}\left(\psi_i,\theta(\mathbf{S})\right)$) can be used to find a most possible key. The maximal correlation with any key is used as a measure of the key clarity.

- Pulse clarity: $P\left(E\right)=\max_{l\in L}\sum_{x=0}^{n-1}E(x)E(x-l)$ where L is a set of lags. That is, maximal autocorrelation of the envelope at any lag is used as a measure of pulse clarity.

It should be noted once again that the primary goal of the MIR features is to describe audio contents at a low computational cost (e.g., real-time computation for automatic music identification services such as Shazam), rather than to describe psychological correlates of acoustic properties (Aucouturier and Bigand, 2012). Therefore, the names of the MIR features are only to serve practical purposes (i.e., they make it easier for human users to remember than numerical indices) and are at best suggestive, but do not necessarily allow for psychological interpretations. One example could be "key clarity." Because it is a maximal correlation with any possible key, it rather describes how clear a key is to an algorithm based on Krumhansl's profiles and cross-correlation than how it sounds to general human listeners. This discrepancy may be negligible when estimating a key based on a chromagram averaged across a whole excerpt as done in the original algorithm (Gómez, 2006). However, when the metric is calculated for short (1–3 s) frames, the discrepancy can be non-trivial. In principle, any clearly presented triad can have a high "key clarity" value even if the chord is distant from the dominant key along the circle of fifths. That is, even if a chord is tonally unstable, disturbing the overall tonality (e.g., Db major triad in C major key; i.e., the famous Neapolitan chord), its "key clarity" could be as high as C major triad in C major key (i.e., tonic). A similar discrepancy could exist for other metrics, such as "pulse clarity," when they are computed for short frames. Thus, readers who wish to better understand the nature of MIR features are strongly encouraged to study the documents provided by the developers of respective implementations. Only for intuitive illustrations, readers can find exemplary audio clips with minimal or maximal values of the listed MIRtoolbox features from 985 intact songs (Sturm, 2013) in GTZAN Musical Genre Dataset (Tzanetakis and Cook, 2002)[5] in **Supplementary Files 1, 2**.

One of the most exciting characteristics of the MIR models lies in the open-source principle: many well-maintained packages are freely available online: e.g., librosa (McFee et al., 2015)[6], MIRtoolbox (Lartillot and Toiviainen, 2007)[7], Essentia (Bogdanov et al., 2013)[8], and more[9]. This has allowed for a rapid adaptation of MIR features in predicting neural responses to natural music in EEG data (Cong et al., 2012, 2013; Sturm et al., 2015, 2017; Stober, 2017; Kaneshiro et al., 2020; Wang et al., 2020; Leahy et al., 2021), intracranial EEG data (Sturm et al., 2014; Omigie et al., 2020), and fMRI data (Alluri et al., 2012, 2013; Toiviainen et al., 2014; Casey, 2017; Hoefle et al., 2018). These studies consistently revealed that the MIR features extract relevant information that is predictive of ongoing neural activity during naturalistic music listening. Some features seem to be more reliable than others in predicting fMRI signals. In particular, when a lasso regression with a fixed canonical hemodynamic function was used as a predictive model (Alluri et al., 2012; Burunat et al., 2016), short-term features that are based on a 25-ms window (mostly describing spectral contents and their short-term dynamics) showed greater reliability than long-term features that are based on 3-s window ("key clarity" and "pulse clarity") (Burunat et al., 2016). The "long-term" measures attempted to capture higher-level perceptions, such as tonal center or meter recognition. While the long-term features could still be useful for differentiating musical pieces (thereby decoding perceived emotions and musical genres from music signals), the studies show that localizing the neural correlates of musical percepts in time can be difficult. In a recent study (Leahy et al., 2021), differential encoding of meters and beats (e.g., for a 4/4 time signature, strong-weak-middle-weak vs. four beats without accents) was detected in human EEG data using an automated beat-tracking algorithm (McFee

[5] https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

[6] https://librosa.org

[7] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox

[8] http://essentia.upf.edu/

[9] https://www.ismir.net/resources/software-tools/

et al., 2015). This suggests that an improvement of linearization may lead to applications of encoding models beyond low-level sensory processing.

It would also be worth mentioning, although it slightly deviates from the main focus of the current review, that the MIR models can be used to objectively describe acoustic features of natural musical stimuli in *controlled* experiments (see, e.g., Whitehead and Armony, 2018). Given that the MIR features would be more sensitive than traditionally used aggregated metrics such as overall RMS, loudness, and spectra, the MIR models can be used as tools to control or match nuisance variability in stimuli. Alternatively, biological models, more explicit modeling of the neural activation throughout the auditory pathway, can also be used to match global statistics (e.g., the first four moments [mean, variance, skewness, and kurtosis] of the cortical representations) (Norman-Haignere and McDermott, 2018).

Another usage of the MIR model is an automatic annotation of musical events to perform analyses comparing neural responses among different events or correlating an aggregated metric of neural responses with the extracted features: e.g., event-related potential in EEG data (Poikonen et al., 2016a,b), dynamic functional connectivity analysis in fMRI data (Singer et al., 2016; Toiviainen et al., 2020), and inter-subject synchronization in fMRI data (Trost et al., 2015; Sachs et al., 2020). However, while traditional methods are computationally efficient and readily available, their performance needs to be taken with caution. In a recent EEG study (Haumann et al., 2021), the MIRtoolbox missed 41.6–45.0% (based on either RMS or spectral flux) of perceivable onsets that were manually detected by an expert rater (i.e., a musicologist), which deteriorated the following EEG analyses based on the automatically extracted onsets. Recent models based on neural networks such as Madmom (Eyben et al., 2010)[10] are known to outperform traditional onset extraction models (∼7% error rate when tested on datasets including music with percussive sessions), which encourages researchers not to be restricted by the traditional models.

More recently, pretrained DNN models have been used to extract their embeddings in new MIR research (Lee and Nam, 2017; Grekow, 2021; Grollmisch et al., 2021). In particular, VGGish [Visual Geometry Group-ish] (Hershey et al., 2017) and Open-L3 [Look, Listen, and Learn more] (Cramer et al., 2019) are CNNs that were developed to generate text labels for given short (∼1 s) audio signals. Both models are pre-trained on large-scale video data (i.e., 60 million AudioSet clips and 8 million YouTube clips) exploiting the correspondence between image and audio data in video sources. In a sense, the networks effectively learn the second-order isomorphism between the image frames and the audio spectrograms. The possibility of a transfer learning of these networks to MIR tasks has been investigated. Koh and Dubnov (2021) extracted audio embeddings using the VGGish and Open-L3 models, then created shallow classifiers (e.g., SVM, Naïve Bayes, and Random Forest) to decode emotional classes (e.g., four quadrants on the Arousal-Valence space or six emotional categories). The CNN embeddings outperformed (32–88% in decoding accuracy) the conventional MIR descriptor (e.g., Mel-Frequency Cepstral Coefficient [MFCC] as a baseline; 31–46%), demonstrating their relevance to music emotion recognition at the excerpt level. In visual domain, the CNN embeddings of images have been found to be related to affective ratings and fMRI responses in univariate and multivariate fashions (Kragel et al., 2019; Horikawa et al., 2020; Koide-Majima et al., 2020), suggesting distributed representations of emotion-specific features (i.e., high-order statistical descriptors of physical properties that are differentially associated with diverse emotions) in the human cortical networks (Sievers et al., 2021). Taken together, DNN embeddings of music are expected to serve as effective predictors for the MIR tasks and the neural encoding analysis.

Computational musicological models

In computational musicology, music is often modeled as a sequence of symbols (e.g., a sequence of notes forms a melody of one part, a sequence of chords forms harmonic progressions and tonality). While this approach ignores multiple "unscored" variability in music signals including timbre, dynamics, and tempo, which are known to be very relevant to emotional responses and associated neural activity (Chapin et al., 2010; Bresin and Friberg, 2011; Trochidis and Bigand, 2013), this approach enables *scalable* analyses on musical structures (Rohrmeier and Cross, 2008; Moss et al., 2019; Rohrmeier, 2020; Hentschel et al., 2021). That is, once symbolic representations are collected, an analysis can be scaled up to a large volume of corpora using computers, a task that would take decades or more for human experts (musicologists) to complete. Moreover, neuroscientific studies based on this approach have investigated how the musical structures form anticipations in spectral and temporal domains in listeners' minds and how they evoke emotional responses via suspended fulfilment or betrayal of such anticipations (for a review of the "predictive coding of music" model, see Vuust et al., 2022).

One successful model, called Information Dynamics Of Music (IDyOM; Pearce, 2005), is a variant of the n-gram model (Shannon, 1948) based on a combinatory n-gram model called Prediction by Partial Match (PPM; Cleary and Witten, 1984). For a given sequence with k events $s = \{e_1, e_2, \ldots, e_k\}$ where e_i denotes the i-th event in the sequence and all discrete events are from a finite set ($e_i \in \mathcal{M}$) and a sub-sequence from the i-th

[10] https://madmom.readthedocs.io/

event to the j-th event is denoted as $s_i^j = \{e_i, e_{i+1}, \ldots, e_j\}$, the conditional probability to observe an event e_i after observing a preceding sequence s_1^{i-1} can be approximated by the (n-1)-th order Markov (i.e., n-gram) model, whose maximum likelihood (ML) estimate is given as:

$$\Pr\left(e_i|s_1^{i-1}\right) \approx \widehat{\Pr}_n\left(e_i|s_1^{i-1}\right) = \begin{cases} 1/\left|\mathcal{M}\right|, & c\left(e_i|s_{(i-n)+1}^{i-1}\right) = 0 \\ \dfrac{c\left(e_i|s_{(i-n)+1}^{i-1}\right)}{\sum_{d\in\mathcal{M}} c\left(d|s_{(i-n)+1}^{i-1}\right)}, & \textit{otherwise} \end{cases} \tag{22}$$

where $|\cdot|$ denotes the number of elements of a set (i.e., cardinality); $c\,(e|s)$ is the number of counts (in the training sets or corpora) of an event e after a sequence s from a given training set; and d is any event from the finite set $\mathcal{M}$. When the transition appears for the first time, a fixed probability based on the size of the set $\mathcal{M}$ can be defined (Pearce, 2005). Put differently, the model estimates the probability by counting transitions in the training set. To allow the model to flexibly learn musical styles across compositions and the local context within each composition, long-term models (counting transitions only across corpora) and short-term models (counting transitions only within the current composition) were combined (Pearce, 2005; Harrison and Pearce, 2018) as in the PPM model (Cleary and Witten, 1984). A Common LISP implementation of the model is available online[11]. This particular model has been developed to predict the pitch and duration of coming notes in monopoly melodies. This allows us to compute the uncertainty of the context (i.e., entropy) and the negative log likelihood of a certain event (i.e., the change of entropy, also known as information content or surprisal) and has successfully predicted neural data (EEG and ECoG from different participants) while listening to MIDI-generated piano melodies extracted from J. S. Bach's Chorales via encoding models (Di Liberto et al., 2020) in line with a previously reported association between conditional probability and evoked neural responses demonstrated with brief orthogonalized stimuli (Koelsch and Jentschke, 2010; Kim et al., 2011). A behavioral study using MIDI-generated flute melodies from classical compositions revealed that the information content had an inverted U-shaped effect (i.e., the "*Wundt*" effect; Wundt, 1874) on mean liking (i.e., an intermediate level of surprisal was preferred over extreme levels), and this effect was modulated by the uncertainty of contexts (Gold et al., 2019). Using a more generalized variant of PPM with memory decay over time, a similar antisymmetric pattern of behavioral responses (preferences for low-uncertainty/high-surprisal or for high-uncertainty/low-surprisal pairs) was reported in an fMRI experiment using chord sequences extracted from McGill Billboard corpus (Cheung et al., 2019), where the interaction between the uncertainty and information content was parametrically localized in clusters over the amygdala/hippocampal complex and medial auditory cortices.

A more commonly used model in symbolic analyses (Mor et al., 2020) is a hidden Markov model (HMM; Baum and Petrie, 1966). The HMM models conditional probabilities between latent (non-observable) states rather than between surface (observable) states, which allows for non-local dependencies and underlying (non-observable) structures in musical compositions to be explicitly expressed (Pearce and Rohrmeier, 2018). Various kinds of HMM models have been used to model different musical structures including melody, rhythm, and harmony (Raphael and Stoddard, 2004; Mavromatis, 2009). The HMM models showed greater prediction performance than n-gram models for certain musical structures (e.g., chord progressions in a jazz corpus; Rohrmeier and Graepel, 2012). However, the application of the HMM in related studies has been done mainly as a classifier that decodes entire musical pieces or genres from EEG signals (Kaur et al., 2017; Ntalampiras and Potamitis, 2019) rather than as a predictive model of musical structures.

Generative neural network models

It has been argued that a prominent method for understanding the statistical structures of natural data is to create a system that can synthesize data *de novo* (Odena et al., 2017). Even if Richard Feynman's famous dictum ("What I cannot create, I do not understand") is true, only its contrapositive ("What I understand, I can create") is also true, not its inversion ("What I can create, I understand"). That is, creating synthetic data would be necessary, but not sufficient, for understanding data. Having said that, various DNN models for music generation have been developed both in the audio and symbolic domains demonstrating the improvements in building such a synthetic system. While the successful performance of such models (e.g., synthetic speech and music that are physically and perceptually similar to real data) does not entail that the model "understands" the natural structures, it has drawn great attention in various fields (see Section "Interpretations of high-dimensional models").

As an example of a symbolic CNN model, MidiNet is a modified deep convolutional generative adversarial network (Yang et al., 2017), where a generator network G generates artificial data to "fool" a discriminator network D, which distinguishes real data from generated data. For a given $t \times p$ binary matrix $\mathbf{X}$ that encodes onsets of notes over p pitch classes and t time steps (quantized beats) and a random noise vector

[11] http://mtpearce.github.io/idyom/

z, the objective function of the generative adversarial networks (GANs) is given as:

$$\min_{G}\max_{D} V(D, G) = \mathbb{E}_{\mathbf{X}\sim p_{data}(\mathbf{X})}\log D(\mathbf{X}) \\ + \mathbb{E}_{\mathbf{z}\sim p_{z}(\mathbf{z})}\log(1 - D(G(\boldsymbol{z}))) \quad (23)$$

where $\mathbf{X} \sim p_{data}(\mathbf{X})$ denotes the sampling from real data and $\mathbf{z} \sim p_z(\mathbf{z})$ denotes the sampling from a random distribution. The output of the generator $G(\cdot) = \hat{\mathbf{X}}$ is a generated "score" with a random input, and the output of the discriminator $D(\cdot)$ is in [0, 1] such that 1 for real data and 0 for generated data. Over iterations, the G learns weights that minimize this function (e.g., $D(G(\boldsymbol{z})) \approx 1$) while the D learns weights that maximizes it (e.g., $D(G(\boldsymbol{z})) \approx 0$). Further details including stabilization and conditioning can be found in the publication (Yang et al., 2017).

Amongst symbolic RNN models, Melody RNN by Google Magenta[12] and Folk-RNN[13] are well known. Both use long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997) to model non-local dependencies. In particular, Folk-RNN is an RNN with three LSTM layers of 512 units (Sturm et al., 2016a), of which a vast number of parameters (over 5.5 million) were trained over 23,000 transcriptions of traditional tunes from Ireland and the UK (with over 4 million tokens [discrete classes of music notation including meter, mode, pitch, and duration]).

In the audio domain, WaveNet[14] by DeepMind is a CNN model that operates on individual audio samples (i.e., amplitude at each time-point) at 16 kHz (Oord et al., 2016). Unlike image-CNN models, all connections between layers were causal (i.e., only nodes that process previous and current, but not following, time steps are connected to a node in a higher layer), so that the temporal order of the underlying structures in the audio data can be preserved.

Another well-known architecture for generative models is a variational autoencoder (VAE; Kingma and Welling, 2013), which introduced a variational Bayesian approach to "non-linear principal component analysis" (Kramer, 1991). The VAE models comprise an encoder, which finds efficient latent representations that are continuous (i.e., being able to be interpolated) and interpretable (i.e., the Euclidian distance between two classes in the latent space reflects "semantic" distance between two classes), and a decoder that generates new data from samples in the latent space. MusicVAE by Google Magenta is a symbolic model (Roberts et al., 2018) and Jukebox[15] by OpenAI is an audio model (Dhariwal et al., 2020). In general, multiple time scales or hierarchical structures are used to capture non-local dependency in musical structures. In particular, Jukebox first learns latent representations of audio samples at three temporal scales (i.e., a sequence of 8, 32, 128 audio samples at 44,100 Hz) using VAE, then quantizes learned patterns into discrete tokens. Then, a transformer is trained on the tokens with a context covering ∼23 s at the longest temporal scale. In a recent study (Castellon et al., 2021), the middle layer of the transformer (4,800 features), which describes the 23-s audio patterns, showed a better performance in predicting arousal and valence ratings (66.9%) as compared to other models including a MFCC model (37.2%) and a CNN model (58.5%).

Recently, VAE models were shown to decode highly realistic face images from fMRI data (VanRullen and Reddy, 2019; Dado et al., 2022). In those studies, latent representations of face images (i.e., a high-dimensional vector representing an image) were extracted using the encoders of the VAE models, then these were used as predictors for fMRI responses (as a weighted linear sum of the vectors) to the corresponding face images. Using this linear model, latent representations were predicted for unseen images using fMRI responses. Finally, using the decoders of the VAE models, face images were reconstructed from the estimated latent representations. These experiments suggest a high similarity in information structures that the generative models and the brain extract from natural images. However, whether this can be generalized to the music domain, in particular in relation to musical emotion, remains to be investigated.

Challenges and perspectives

Stimuli for predictive modeling

So far, we have discussed why naturalistic stimuli are needed. To summarize, (1) the non-linearity of the neural system renders responses to controlled stimuli weak, (2) the external validity of a controlled experiment can be highly limited when applied to a non-linear system, and (3) recent developments of computational models of natural data provide testable models of non-linear transforms. However, there are also clear disadvantages of natural stimuli for experiments, which keep researchers inclined to orthogonalized stimuli (Hamilton and Huth, 2020): (1) multicollinearity, (2) over/under-representation, and (3) domain-specificity. As shown in Equation 9, the estimate of an encoding model reflects the serial-correlation of features and the multicollinearity among features, which would be minimized when white noise is used as a stimulus. In natural music, similarly to many other natural stimuli, various properties are often highly correlated (e.g., the strong correlation between pitch and onset density in a Western corpus; Broze and Huron, 2013); many features

[12] https://github.com/magenta/magenta/tree/main/magenta/models/melody_rnn

[13] https://github.com/IraKorshunova/folk-rnn

[14] https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio

[15] https://openai.com/blog/jukebox/

follow the power-law distributions (e.g., pitch, chord transition, and timbre in classical corpora, contemporary Western popular music, and cross-cultural corpora of folk songs; Serrà et al., 2012; Mehr et al., 2019; Moss et al., 2019); and the patterns of multicollinearity are variable across musical styles, cultures, and time (Broze and Huron, 2013; White, 2014; Pearce, 2018). Furthermore, a small (n = 10–20) set of exemplary stimuli typically used in neuroimaging experiments could over-represent certain covariance patterns that are different from the population of natural stimuli.

A brute-force approach to mitigating this problem would be to sample a massive set of stimuli. For example, the Natural Scenes Dataset (Allen et al., 2022) comprises ~38 h/subject of 7-T fMRI data collected over ~40 sessions watching 10,000 pictures of natural scenes with three repetitions, resulting in a total of 70,566 unique pictures from 8 subjects. Interestingly, but unsurprisingly, the DNN models predicted brain responses worse than did a traditional model (i.e., Gabor wavelet) with small samples (e.g., when trained on < 1,000 pictures with three repetitions) in some subjects, but were far better with more samples (e.g., > 3,000 pictures with three repetitions). If such massive-stimulus (i.e., "deeply-sampled") high-quality datasets with naturalistic music are shared as open-source resources, it would foster rapid advances in the field (Poldrack and Gorgolewski, 2014).

While the large-scale stimuli data are necessary to investigate how the biases and variance in small sets impact estimates, there can be specific cases where only limited stimuli can be used (e.g., pediatric or elderly populations, epileptic patients undergoing neurosurgery). In such cases, a stimuli selection can be made in order to reduce multicollinearity and alleviate over/under-representation. Insofar as such a selection could degrade the ecological validity of the experiment to some extent, an optimal tradeoff should be carefully determined.

Neural measurements

Encoding and decoding models can suffer from excessive noise in data, which is typically very high in most of the non-invasive measurements of neural signals. For instance, the fractional signal change induced by neuronal activity is estimated to be 1–2% at 3 T (Uludağ et al., 2009) and a simulation study found the SNR of M/EEG signals between –30 and –20 dB (Goldenholz et al., 2009). Especially for long, complex naturalistic stimuli, it can be difficult (or could violate assumptions such as non-familiarity of presented stimuli) to repeat the identical stimuli many times, which is a commonly used denoising technique (e.g., event-related potentials), whereby many trials with identical stimuli are averaged to cancel out non-stimulus-locked activities. Alternatively, there are denoising techniques that have been used for resting-state fMRI data, which is also, in a sense, single-trial data. The methods for suppressing non-physiological noise (e.g., spin-history artifacts due to head motions) and non-neural noise (e.g., fluctuations due to cardiac pulses and respirations) in fMRI data (see Caballero-Gaudes and Reynolds, 2017 for a review) include: RETROICOR (image-based retrospective correction; Hu et al., 1995), CompCor (component based noise correction; Behzadi et al., 2007), and ICA-AROMA (ICA-based automatic removal of motion artifacts; Pruim et al., 2015); these methods have been widely used for resting-state data and are applicable for naturalistic experiments as well. Task-based denoising techniques, such as GLMdenoise (Kay et al., 2013) and GLMsingle (Prince et al., 2021), that extract principal components that are not related to the experiment design has been used for encoding models with naturalistic stimuli and found to be beneficial for multivariate pattern analysis (Charest et al., 2018). Recently, a DNN-based interpolation method that reconstructs fMRI volumes while removing independent noise has been developed (Lecoq et al., 2021).

Multi-echo fMRI sequence (Posse et al., 1999) with a dedicated denoising technique (i.e., ME-ICA, multi-echo imaging with spatial independent component analysis; Kundu et al., 2013) has been suggested to separate BOLD signals (i.e., physiological) from other signals (e.g., non-physiological artifacts such as head motion, thermal noise from subjects and electronics, device imperfection) by exploiting a linear dependency of the BOLD effect on echo times. Although multi-echo fMRI requires a larger voxel and/or a longer time of repetition (TR) than the standard fMRI sequence for multiple readouts [e.g., ~4-mm iso-voxel and TR of 2 s in multi-echo fMRI (Kundu et al., 2013) as compared to 2-mm iso-voxel and TR of 2 s in conventional single-echo fMRI, i.e., ~8 times larger in volume], simultaneous multi-slice (also known as multi-band) acceleration techniques are expected to make the spatial and temporal resolutions of the multi-echo fMRI comparable to single-echo fMRI (Kundu et al., 2017).

Besides the lab-based neural measurements that require gigantic machines such as fMRI and MEG, wearable and portable EEG systems have been developed and already used in various naturalistic paradigms. For instance, wireless EEGs were used for hyper-scanning two pianists performing a piano duet (Zamm et al., 2020); and wireless systems were used to simultaneously collect physiological activity (e.g., electrocardiogram, facial muscle electromyogram, respiration, heart rates) from whole audiences (~40 participants per concert) attending live string quintet performances (Czepiel et al., 2021; Merrill et al., 2021; Tschacher et al., 2021). In particular, the feasibility of a wireless in-ear EEG system has drawn considerable attention (Looney et al., 2014; Bleichner and Emkes, 2020; Nithya and Ramesh, 2020). A built-in EEG system in everyday devices (e.g., wireless in-ear headphones) might open a new possibility of collecting neural data from millions of people while they listen to their favorite music in their day-to-day lives.

Interpretations of high-dimensional models

It may be broadly accepted that DNN models can serve, with caution, as functional models of some aspects of human cognition (notions such as ‘a DNN model processes similar information as humans do for certain tasks for some aspects’). But whether they can serve as a mechanistic model, even for a particular domain, (ideas such as ‘weights of a certain layer in a DNN model correspond to effective connectivity between neurons in a human brain’) remains under debate (Kay, 2018). In fact, this point has created heated discussions in the field of cognitive neurosciences (Kriegeskorte, 2015; Kay, 2018; Cichy and Kaiser, 2019; Kell and McDermott, 2019; Kriegeskorte and Golan, 2019; Lindsay, 2021; Pulvermüller et al., 2021). In particular, it has been pointed out that evidence and counterevidence should be examined in an unbiased fashion (Guest and Martin, 2021).

A promising approach has been suggested that involves investigating, instead of the first-order isomorphism between the physical properties of an object and its representation in a system, the second-order isomorphism between representations of multiple objects in multiple systems (Kriegeskorte et al., 2008). Investigating the representations in systems is consistent with efforts to “understand” (or interpret) the high-dimensional models. Various techniques for the interpretation of the DNN models have also been vigorously discussed and developed (Sturm et al., 2016b; Montavon et al., 2018; Sturm, 2018; Keshishian et al., 2020). In particular, Kay (2018) summarized three practical approaches for deepening our understanding of high-dimensional models: (1) we can observe the model’s behaviors to stimuli (i.e., mapping to the lower, intuitive space), (2) we can manipulate the models (i.e., perturbing parameters and observing performance chances), and (3) we can model a model (i.e., creating a simpler form that approximates the model’s simulated behaviors). Efforts to understand how models process musical information would be critical to deepening our intuitions as to how the human brain processes musical information.

Conclusion

The current review discussed predictive models used for encoding and decoding analyses and music models that capture acoustics and underlying structures. In particular, predictive models have introduced a reproducible form of cognitive neuroscience and computational models have provided us with quantitative metrics that can be compared with neural representations of natural music. Novel data with large-scale stimuli and high-dimensional models are expected to allow us to better handle the non-linearity of the musical brain.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

The Max Planck Society supported this work.

Acknowledgments

The author thanks Daniela Sammler, Tobias Overath, Federico de Martino, Seung-Choel Baek, and Mariëlle Baelemans for insightful discussions that inspired this work and for their inputs that improved the earlier versions of the manuscript. The author also thanks William Martin for proof-reading and copy-editing and all reviewers for their constructive comments that helped improving the rigor and readability of the manuscript. The author acknowledges the free image sources from www.irasutoya.com (Takashi Mifune), which reserves the copyrights of the source illustrations.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.928841/full#supplementary-material

References

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J. L. (2014). Pixels to voxels: Modeling visual representation in the human brain. *arXiv* [Preprint].

Aguirre, G. K., Zarahn, E., and D'esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *Neuroimage* 8, 360–369. doi: 10.1006/nimg.1998.0369

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* 25, 116–126. doi: 10.1038/s41593-021-00962-x

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., and Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *Neuroimage* 59, 3677–3689. doi: 10.1016/j.neuroimage.2011.11.019

Alluri, V., Toiviainen, P., Lund, T. E., Wallentin, M., Vuust, P., Nandi, A. K., et al. (2013). From vivaldi to beatles and back: Predicting lateralized brain responses to music. *Neuroimage* 83, 627–636. doi: 10.1016/j.neuroimage.2013.06.064

Amari, S. (1967). A Theory of Adaptive Pattern Classifiers", in *IEEE Transactions on Electronic Computers,* (New York, NY: IEEE)16, 299–307 doi: 10.1109/PGEC.1967.264666

Armitage, S. J., Jasim, S. A., Marks, A. E., Parker, A. G., Usik, V. I., and Uerpmann, H.-P. (2011). The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. *Science* 331, 453–456. doi: 10.1126/science.1199113

Aucouturier, J.-J., and Bigand, E. (2012). "Mel Cepstrum & Ann Ova: The Difficult Dialog Between MIR and Music Cognition," in *13th International Society for Music Information Retrieval Conference*. (New York, NY: ACM Digital Library), 397–402.

Badillo, S., Vincent, T., and Ciuciu, P. (2013). Group-level impacts of within- and between-subject hemodynamic variability in fMRI. *Neuroimage* 82, 433–448. doi: 10.1016/j.neuroimage.2013.05.100

Barlow, H. B. (1961). *Possible Principles Underlying the Transformation of Sensory Messages Sensory Communication*. Cambridge: MIT Press, 217–234.

Baum, L. E., and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Ann. Math. Stat.* 37, 1554–1563. doi: 10.1214/aoms/1177699147

Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007). A component based noise correction method (compcor) for bold and perfusion based fmri. *Neuroimage* 37, 90–101. doi: 10.1016/j.neuroimage.2007.04.042

Bellier, L., Llorens, A., Marciano, D., Schalk, G., Brunner, P., Knight, R. T., et al. (2022). Encoding and decoding analysis of music perception using intracranial EEG. *bioRxiv* [Preprint]. doi: 10.1101/2022.01.27.478085

Bishop, C. M., and Nasrabadi, N. M. (2006). *Pattern Recognition and Machine Learning*. Germany: Springer.

Bleichner, M. G., and Emkes, R. (2020). Building an ear-EEG system by hacking a commercial neck speaker and a commercial EEG amplifier to record brain activity beyond the lab. *J. Open Hardware* 4:5. doi: 10.5334/joh.25

Boer, E. D., and Kuyper, P. (1968). Triggered Correlation. *IEEE Trans. Biomed. Eng.* 15, 169–179. doi: 10.1109/TBME.1968.4502561

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., et al. (2013). "Essentia: An audio analysis library for music information retrieval," in *14th Conference of the International Society for Music Information Retrieval (ISMIR)*, eds A. Britto, F. Gouyon, and S. Dixon (Brazil: International Society for Music Information Retrieval).

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. (New York, NY: ACM Digital Library), 144–152. doi: 10.1145/130385.130401

Box, G. E. P. (1979). "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics*, eds R. L. Launer and G. N. Wilkinson (Amsterdam: Elsevier), 201–236. doi: 10.1016/B978-0-12-438150-6.50018-2

Bresin, R., and Friberg, A. (2011). Emotion rendering in music: Range and characteristic values of seven musical variables. *Cortex* 47, 1068–1081. doi: 10.1016/j.cortex.2011.05.009

Brodbeck, C., Hong, L. E., and Simon, J. Z. (2018a). Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol.* 28, 3976–3983.e5. doi: 10.1016/j.cub.2018.10.042

Brodbeck, C., Presacco, A., and Simon, J. Z. (2018b). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage* 172, 162–174. doi: 10.1016/j.neuroimage.2018.01.042

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* 28, 803–809.e3. doi: 10.1016/j.cub.2018.01.080

Broze, Y., and Huron, D. (2013). Is Higher Music Faster? Pitch–Speed Relationships in Western Compositions. *Music Percept.* 31, 19–31. doi: 10.1525/mp.2013.31.1.19

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychol. Rev.* 50:255. doi: 10.1037/h0060889

Burunat, I., Toiviainen, P., Alluri, V., Bogert, B., Ristaniemi, T., Sams, M., et al. (2016). The reliability of continuous brain responses during naturalistic listening to music. *Neuroimage* 124, 224–231. doi: 10.1016/j.neuroimage.2015.09.005

Caballero-Gaudes, C., and Reynolds, R. C. (2017). Methods for cleaning the BOLD fMRI signal. *NeuroImage* 154, 128–149. doi: 10.1016/j.neuroimage.2016.12.018

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* 10:e1003963. doi: 10.1371/journal.pcbi.1003963

Casey, M. A. (2017). Music of the 7Ts: Predicting and Decoding Multivoxel fMRI Responses with Acoustic, Schematic, and Categorical Music Features. *Front. Psychol.* 8:1179. doi: 10.3389/fpsyg.2017.01179

Castellon, R., Donahue, C., and Liang, P. (2021). Codified audio language modeling learns useful representations for music information retrieval. *arXiv* [Preprint].

Caucheteux, C., and King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Commun. Biol.* 5:134. doi: 10.1038/s42003-022-03036-1

Chapin, H., Jantzen, K., Scott Kelso, J., Steinberg, F., and Large, E. (2010). Dynamic emotional and neural responses to music depend on performance expression and listener experience. *PLoS One* 5:e13812. doi: 10.1371/journal.pone.0013812

Charest, I., Kriegeskorte, N., and Kay, K. N. (2018). GLMdenoise improves multivariate pattern analysis of fMRI data. *NeuroImage* 183, 606–616. doi: 10.1016/j.neuroimage.2018.08.064

Cheung, V. K. M., Harrison, P. M. C., Meyer, L., Pearce, M. T., Haynes, J. D., and Koelsch, S. (2019). Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Curr. Biol.* 29, 4084–4092.e4. doi: 10.1016/j.cub.2019.09.067

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.* 106, 2719–2732. doi: 10.1121/1.428100

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887–906. doi: 10.1121/1.1945807

Chu, C., Ni, Y., Tan, G., Saunders, C. J., and Ashburner, J. (2011). Kernel regression for fMRI pattern prediction. *NeuroImage* 56, 662–673. doi: 10.1016/j.neuroimage.2010.03.058

Cichy, R. M., and Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends Cogn. Sci.* 23, 305–317. doi: 10.1016/j.tics.2019.01.009

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *J. Verb. Learn. Verb. Behav.* 12, 335–359. doi: 10.1016/S0022-5371(73)80014-3

Cleary, J., and Witten, I. (1984). Data Compression Using Adaptive Coding and Partial String Matching. *IEEE Trans. Commun.* 32, 396–402. doi: 10.1109/TCOM.1984.1096090

Conard, N. J., Malina, M., and Munzel, S. C. (2009). New flutes document the earliest musical tradition in southwestern germany. *Nature* 460, 737–740. doi: 10.1038/nature08169

Cong, F., Alluri, V., Nandi, A. K., Toiviainen, P., Fa, R., Abu-Jamous, B., et al. (2013). Linking Brain Responses to Naturalistic Music Through Analysis of Ongoing EEG and Stimulus Features. *IEEE Trans. Multimedia* 15, 1060–1069. doi: 10.1109/TMM.2013.2253452

Cong, F., Phan, A. H., Zhao, Q., Nandi, A. K., Alluri, V., Toiviainen, P., et al. (2012). "Analysis of ongoing EEG elicited by natural music stimuli using nonnegative tensor factorization," in *2012 Proceedings of the 20th European Signal Processing Conference*. (Bucharest: IEEE)

Cramer, J., Wu, H.-H., Salamon, J., and Bello, J. P. (2019). "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing.* (New York, NY: IEEE), 3852–3856. doi: 10.1109/ICASSP.2019.8682475

Czepiel, A., Fink, L. K., Fink, L. T., Wald-Fuhrmann, M., Tröndle, M., and Merrill, J. (2021). Synchrony in the periphery: Inter-subject correlation of physiological responses during live music concerts. *Sci. Rep.* 11:22457. doi: 10.1038/s41598-021-00492-3

Dado, T., Güçlütürk, Y., Ambrogioni, L., Ras, G., Bosch, S., Van Gerven, M., et al. (2022). Hyperrealistic neural decoding for reconstructing faces from fMRI activations via the GAN latent space. *Sci. Rep.* 12:141. doi: 10.1038/s41598-021-03938-w

Daube, C., Ince, R. A. A., and Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Curr. Biol.* 29, 1924–1937.e9. doi: 10.1016/j.cub.2019.04.067

De Angelis, V., De Martino, F., Moerel, M., Santoro, R., Hausfeld, L., and Formisano, E. (2018). Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds. *NeuroImage* 180, 291–300. doi: 10.1016/j.neuroimage.2017.11.020

d'Errico, F., Henshilwood, C., Lawson, G., Vanhaeren, M., Tillier, A.-M., Soressi, M., et al. (2003). Archaeological Evidence for the Emergence of Language Symbolism, and Music–An Alternative Multidisciplinary Perspective. *J. World Prehistory* 17, 1–70. doi: 10.1023/A:1023980201043

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv* [Preprint].

Di Liberto, G. M., Pelofi, C., Patel, R. B., Mehta, P., Herrero, A. D., Cheveigné, A., et al. (2020). Cortical encoding of melodic expectations in human temporal cortex. *eLife* 9:e51784. doi: 10.7554/eLife.51784

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2014). "Decaf: A deep convolutional activation feature for generic visual recognition," in *International Conference on Machine Learning*. (New York, NY: ACM Digital Library), 647–655.

Erb, J., Armendariz, M., De Martino, F., Goebel, R., Vanduffel, W., and Formisano, E. (2019). Homology and specificity of natural sound-encoding in human and monkey auditory cortex. *Cereb. cortex* 29, 3636–3650. doi: 10.1093/cercor/bhy243

Eyben, F., Böck, S., Schuller, B., and Graves, A. (2010). "Universal onset detection with bidirectional long-short term memory neural networks," in *Proc. 11th Intern. Soc. for Music Information Retrieval Conference, ISMIR*. (Germany: Technische Universität München). 589–594.

Floridi, L., and Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* 30, 681–694. doi: 10.1007/s11023-020-09548-1

Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., and Penny, W. D. (2006). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Burlington: Elsevier.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1994a). Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402

Friston, K. J., Jezzard, P., and Turner, R. (1994b). Analysis of functional MRI time-series. *Hum. Brain Mapp.* 1, 153–171. doi: 10.1002/hbm.460010207

Gibson, J. J. (1978). The ecological approach to the visual perception of pictures. *Leonardo* 11, 227–235. doi: 10.2307/1574154

Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., and Kording, K. P. (2020). Machine Learning for Neural Decoding. *eneuro* 7, ENEURO.0506–19.2020. doi: 10.1523/ENEURO.0506-19.2020

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

Gold, B. P., Pearce, M. T., Mas-Herrero, E., Dagher, A., and Zatorre, R. J. (2019). Predictability and Uncertainty in the Pleasure of Music: A Reward for Learning? *J. Neurosci.* 39, 9397–9409. doi: 10.1523/JNEUROSCI.0428-19.2019

Goldenholz, D. M., Ahlfors, S. P., Hämäläinen, M. S., Sharon, D., Ishitobi, M., Vaina, L. M., et al. (2009). Mapping the signal-to-noise-ratios of cortical sources in magnetoencephalography and electroencephalography. *Hum. Brain Mapp.* 30, 1077–1086. doi: 10.1002/hbm.20571

Gómez, E. (2006). Tonal description of polyphonic audio for music content processing. *Informs J. Comput.* 18, 294–304. doi: 10.1287/ijoc.1040.0126

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge: MIT press.

Grekow, J. (2021). Music emotion recognition using recurrent neural networks and pretrained models. *J. Intell. Inf. Syst.* 57, 531–546. doi: 10.1007/s10844-021-00658-5

Grollmisch, S., Cano, E., Kehling, C., and Taenzer, M. (2021). "Analyzing the potential of pre-trained embeddings for audio classification tasks," in *2020 28th European Signal Processing Conference*. (New York: IEEE), 790–794. doi: 10.23919/Eusipco47968.2020.9287743

Güçlü, U., Thielen, J., Hanke, M., and Van Gerven, M. (2016). "Brains on beats," in *Proceedings of the 30th International Conference on Neural Information Processing Systems.* (New York, NY: Curran Associates Inc). 29, 2109–2117.

Güçlü, U., and van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Guest, O., and Martin, A. E. (2021). On logical inference over brains, behaviour, and artificial neural networks. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/tbmcg

Hamilton, L. S., and Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Lang. Cogn. Neurosci.* 35, 573–582. doi: 10.1080/23273798.2018.1499946

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., Fu, D., et al. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage* 198, 125–136. doi: 10.1016/j.neuroimage.2019.05.039

Handwerker, D. A., Ollinger, J. M., and D'esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage* 21, 1639–1651. doi: 10.1016/j.neuroimage.2003.11.029

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., et al. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Sci. Data* 1, 1–18. doi: 10.1038/sdata.2014.3

Harrison, P., and Pearce, M. (2018). "Dissociating sensory and cognitive theories of harmony perception through computational modeling," in *Proceedings of ICMPC15/ESCOM10*, eds R. Parncutt and S. Sattmann (Graz: University of Graz). 194–199. doi: 10.31234/osf.io/wgjyv

Hasson, U., Malach, R., and Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends Cogn. Sci.* 14, 40–48. doi: 10.1016/j.tics.2009.10.011

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science* 303, 1634–1640. doi: 10.1126/science.1089506

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Germany: Springer Science and Business Media. doi: 10.1007/978-0-387-84858-7

Haumann, N. T., Lumaca, M., Kliuchko, M., Santacruz, J. L., Vuust, P., and Brattico, E. (2021). Extracting human cortical responses to sound onsets and acoustic feature changes in real music, and their relation to event rate. *Brain Res.* 1754:147248. doi: 10.1016/j.brainres.2020.147248

Henson, R., Rugg, M. D., and Friston, K. J. (2001). The choice of basis functions in event-related fMRI. *NeuroImage* 13, 149–149. doi: 10.1016/S1053-8119(01)91492-2

Hentschel, J., Neuwirth, M., and Rohrmeier, M. (2021). The annotated mozart sonatas: Score, harmony, and cadence. *Trans. Int. Soc. Music Inform. Retrieval* 4, 67–80. doi: 10.5334/tismir.63

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (New York, NY: IEEE), 131–135. doi: 10.1109/ICASSP.2017.7952132

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hoefle, S., Engel, A., Basilio, R., Alluri, V., Toiviainen, P., Cagy, M., et al. (2018). Identifying musical pieces from fMRI data using encoding and decoding models. *Sci. Rep.* 8:2266. doi: 10.1038/s41598-018-20732-3

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

Horikawa, T., Cowen, A. S., Keltner, D., and Kamitani, Y. (2020). The Neural Representation of Visually Evoked Emotion Is High-Dimensional, Categorical, and Distributed across Transmodal Brain Regions. *iScience* 23:101060. doi: 10.1016/j.isci.2020.101060

Hu, X., Le, T. H., Parrish, T., and Erhard, P. (1995). Retrospective estimation and correction of physiological fluctuation in functional MRI. *Magn. Reason. Med.* 34, 201–212. doi: 10.1002/mrm.1910340211

Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., et al. (2017). New fossils from Jebel Irhoud. Morocco and the pan-African origin of Homo sapiens. *Nature* 546, 289–292. doi: 10.1038/nature22336

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458. doi: 10.1038/nature17637

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Jääskeläinen, I. P., Sams, M., Glerean, E., and Ahveninen, J. (2021). Movies and narratives as naturalistic stimuli in neuroimaging. *NeuroImage* 224:117445. doi: 10.1016/j.neuroimage.2020.117445

Jolly, E., and Chang, L. J. (2019). The Flatland Fallacy: Moving Beyond Low–Dimensional Thinking. *Topics Cogn. Sci.* 11, 433–454. doi: 10.1111/tops.12404

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2

Juslin, P. N., Harmat, L., and Eerola, T. (2013). What makes music emotionally significant? *Exploring the underlying mechanisms*. *Psychol. Music* 42, 599–623. doi: 10.1177/0305735613484548

Kaneshiro, B., Nguyen, D. T., Norcia, A. M., Dmochowski, J. P., and Berger, J. (2020). Natural music evokes correlated eeg responses reflecting temporal structure and beat. *NeuroImage* 214:116559. doi: 10.1016/j.neuroimage.2020.116559

Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in Data Mining: Formulation, Detection, and Avoidance. *Acm Trans. Knowl. Discov.* 6, 1–12. doi: 10.1145/2382577.2382579

Kaur, B., Singh, D., and Roy, P. P. (2017). A Novel framework of EEG-based user identification by analyzing music-listening behavior. *Multimed. Tools. Appl.* 76, 25581–25602. doi: 10.1007/s11042-016-4232-2

Kay, K., Rokem, A., Winawer, J., Dougherty, R., and Wandell, B. (2013). GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* 7:247. doi: 10.3389/fnins.2013.00247

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

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

Keesman, K. J. (2011). *System Identification: An Introduction*. London: Springer. doi: 10.1007/978-0-85729-522-4

Kell, A. J. E., and McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Curr. Opin. Neurol.* 55, 121–132. doi: 10.1016/j.conb.2019.02.003

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., and Mcdermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* 98, 630–644.e16. doi: 10.1016/j.neuron.2018.03.044

Keshishian, M., Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., and 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

Khalighinejad, B., Herrero, J. L., Mehta, A. D., and Mesgarani, N. (2019). Adaptation of the human auditory cortex to changing background noise. *Nat. Commun.* 10:2509. doi: 10.1038/s41467-019-10611-4

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10:e1003915. doi: 10.1371/journal.pcbi.1003915

Kim, S.-G., Kim, J. S., and Chung, C. K. (2011). The effect of conditional probability of chord progression on brain response: An meg study. *PLoS One* 6:e17337. doi: 10.1371/journal.pone.0017337

Kim, S.-G., Lepsien, J., Fritz, T. H., Mildner, T., and Mueller, K. (2017). Dissonance encoding in human inferior colliculus covaries with individual differences in dislike of dissonant music. *Sci. Rep.* 7:5726. doi: 10.1038/s41598-017-06105-2

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv* [Preprint].

Klein, D. J., Depireux, D. A., Simon, J. Z., and Shamma, S. A. (2000). Robust Spectrotemporal Reverse Correlation for the Auditory System: Optimizing Stimulus Design. *J. Comput. Neurosci.* 9, 85–111. doi: 10.1023/A:1008990412183

Koelsch, S., and Jentschke, S. (2010). Differences in electric brain responses to melodies and chords. *J. Cogn. Neurosci.* 22, 2251–2262. doi: 10.1162/jocn.2009.21338

Koh, E., and Dubnov, S. (2021). Comparison and analysis of deep audio embeddings for music emotion recognition. *arXiv* [Preprint].

Koide-Majima, N., Nakai, T., and Nishimoto, S. (2020). Distinct dimensions of emotion in the human brain and their representation on the cortical surface. *NeuroImage* 222:117258. doi: 10.1016/j.neuroimage.2020.117258

Kragel, P. A., Reddan, M. C., Labar, K. S., and Wager, T. D. (2019). Emotion schemas are embedded in the human visual system. *Sci. Adv.* 5:eaaw4358. doi: 10.1126/sciadv.aaw4358

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* 37, 233–243. doi: 10.1002/aic.690370209

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu. Rev. Vis. Sci.* 1, 417–446. doi: 10.1146/annurev-vision-082114-035447

Kriegeskorte, N., and Golan, T. (2019). Neural network models and deep learning. *Curr. Biol.* 29, R231–R236. doi: 10.1016/j.cub.2019.02.034

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Krumhansl, C. L., and Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *J. Exp. Psychol. Hum. Percept. Perform.* 5:579. doi: 10.1037/0096-1523.5.4.579

Kundu, P., Brenowitz, N. D., Voon, V., Worbe, Y., Vértes, P. E., Inati, S. J., et al. (2013). Integrated strategy for improving functional connectivity mapping using multiecho fmri. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16187–16192. doi: 10.1073/pnas.1301725110

Kundu, P., Voon, V., Balchandani, P., Lombardo, M. V., Poser, B. A., and Bandettini, P. A. (2017). Multi-echo fMRI: A review of applications in fMRI denoising and analysis of BOLD signals. *NeuroImage* 154, 59–80. doi: 10.1016/j.neuroimage.2017.03.033

Lartillot, O., and Toiviainen, P. (2007). A matlab toolbox for musical feature extraction from audio. *Proc. Int. Conf. Digital Audio Effects*. 2007, 237–244.

Leahy, J., Kim, S.-G., Wan, J., and Overath, T. (2021). An Analytical Framework of Tonal and Rhythmic Hierarchy in Natural Music Using the Multivariate Temporal Response Function. *Front. Neurosci.* 15:894. doi: 10.3389/fnins.2021.665767

Lecoq, J., Oliver, M., Siegle, J. H., Orlova, N., Ledochowitsch, P., and Koch, C. (2021). Removing independent noise in systems neuroscience data using DeepInterpolation. *Nat. Methods* 18, 1401–1408. doi: 10.1038/s41592-021-01285-2

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

Lee, J., and Nam, J. (2017). Multi-Level and Multi-Scale Feature Aggregation Using Pretrained Convolutional Neural Networks for Music Auto-Tagging. *IEEE Signal Process. Lett.* 24, 1208–1212. doi: 10.1109/LSP.2017.2713830

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *J. Cogn. Neurosci.* 33, 2017–2031. doi: 10.1162/jocn_a_01544

Ljung, L. (2010). Perspectives on system identification. *Annu. Rev. Control* 34, 1–12. doi: 10.1016/j.arcontrol.2009.12.001

Ljung, L., Chen, T., and Mu, B. (2020). A shift in paradigm for system identification. *Int. J. Control* 93, 173–180. doi: 10.1080/00207179.2019.1578407

Looney, D., Kidmose, P., and Mandic, D. P. (2014). “"Ear-EEG: User-Centered and Wearable BCI,” in *Brain-Computer Interface Research: A State-of-the-Art Summary -2*, eds C. Guger, B. Allison, and E. C. Leuthardt (Berlin: Springer), 41–50. doi: 10.1007/978-3-642-54707-2_5

Martin, S., Mikutta, C., Leonard, M. K., Hungate, D., Koelsch, S., Shamma, S., et al. (2017). Neural Encoding of Auditory Features during Music Perception and Imagery. *Cereb. Cortex* 28, 4222–4233. doi: 10.1093/cercor/bhx277

Mavromatis, P. (2009). “HMM Analysis of Musical Structure: Identification of Latent Variables Through Topology-Sensitive Model Selection,” in *International Conference on Mathematics and Computation in Music*, (Germany: Springer), 205–217. doi: 10.1007/978-3-642-02394-1_19

McFee, B., Raffel, C., Liang, D., Ellis, D. P., Mcvicar, M., Battenberg, E., et al. (2015). “Librosa: Audio and music signal analysis in python,” in *Proceedings of the*

14th Python in Science Conference, (New York, NY: University Music and Audio Research Laboratory). 18–25. doi: 10.25080/Majora-7b98e3ed-003

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., et al. (2019). Universality and diversity in human song. *Science* 366:eaax0868. doi: 10.1126/science.aax0868

Merrill, J., Czepiel, A., Fink, L. T., Toelle, J., and Wald-Fuhrmann, M. (2021). The aesthetic experience of live concerts: Self-reports and psychophysiology. *Psychol. Aesthet. Creat. Arts* doi: 10.1037/aca0000390 [Epub ahead of print].

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Moerel, M., De Martino, F., Kemper, V. G., Schmitter, S., Vu, A. T., Uðurbil, K., et al. (2018). Sensitivity and specificity considerations for fmri encoding, decoding, and mapping of auditory cortex at ultra-high field. *Neuroimage* 164, 18–31. doi: 10.1016/j.neuroimage.2017.03.063

Moerel, M., De Martino, F., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., et al. (2013). Processing of natural sounds: Characterization of multipeak spectral tuning in human auditory cortex. *J. Neurosci.* 33, 11888–11898. doi: 10.1523/JNEUROSCI.5306-12.2013

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Process.* 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Mor, B., Garhwal, S., and Kumar, A. (2020). A Systematic Literature Review on Computational Musicology. *Arch. Comput. Methods Eng.* 27, 923–937. doi: 10.1007/s11831-019-09337-9

Moss, F. C., Neuwirth, M., Harasim, D., and Rohrmeier, M. (2019). Statistical characteristics of tonal harmony: A corpus study of Beethoven's string quartets. *PLoS One* 14:e0217242. doi: 10.1371/journal.pone.0217242

Nakai, T., Koide-Majima, N., and Nishimoto, S. (2021). Correspondence of categorical and feature-based representations of music in the human brain. *Brain Behav.* 11:e01936. doi: 10.1002/brb3.1936

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage* 56, 400–410. doi: 10.1016/j.neuroimage.2010.07.073

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron* 63, 902–915. doi: 10.1016/j.neuron.2009.09.006

Nastase, S. A., Goldstein, A., and Hasson, U. (2020a). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage* 222:117254. doi: 10.1016/j.neuroimage.2020.117254

Nastase, S. A., Liu, Y.-F., Hillman, H., Norman, K. A., and Hasson, U. (2020b). Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage* 217:116865. doi: 10.1016/j.neuroimage.2020.116865

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* 21, 1641–1646. doi: 10.1016/j.cub.2011.08.031

Nithya, V., and Ramesh, G. P. (2020). "Wireless EAR EEG Signal Analysis with Stationary Wavelet Transform for Co Channel Interference in Schizophrenia Diagnosis," in *Recent Trends and Advances in Artificial Intelligence and Internet of Things*, eds V. E. Balas, R. Kumar, and R. Srivastava (Cham: Springer International Publishing), 253–265. doi: 10.1007/978-3-030-32644-9_27

Norman-Haignere, S., Kanwisher, N. G., and McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* 88, 1281–1296. doi: 10.1016/j.neuron.2015.11.035

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., Mcdermott, J. H., et al. (2022). A neural population selective for song in human auditory cortex. *Curr. Biol.* 32, 1470–1484.e12. doi: 10.1016/j.cub.2022.01.069

Norman-Haignere, S. V., and McDermott, J. H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biol.* 16:e2005127. doi: 10.1371/journal.pbio.2005127

Ntalampiras, S., and Potamitis, I. (2019). A Statistical Inference Framework for Understanding Music-Related Brain Activity. *IEEE J. Select. Topics Signal Process.* 13, 275–284. doi: 10.1109/JSTSP.2019.2905431

Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage* 197, 482–492. doi: 10.1016/j.neuroimage.2019.04.012

Odena, A., Olah, C., and Shlens, J. (2017). "Conditional Image Synthesis with Auxiliary Classifier GANs," in *Proceedings of the 34th International Conference on Machine Learning*, eds P. Doina and T. Yee Whye (New York, NY: ACM Digital Library).

Omigie, D., Lehongre, K., Navarro, V., Adam, C., and Samson, S. (2020). Neuro-oscillatory tracking of low- and high-level musico-acoustic features during naturalistic music listening: Insights from an intracranial electroencephalography study Psychomusicology. *Music Mind Brain* 30, 37–51. doi: 10.1037/pmu0000249

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). Wavenet: A generative model for raw audio. *arXiv* [Preprint].

Pearce, M., and Rohrmeier, M. (2018). "Musical Syntax II: Empirical Perspectives," in *Springer Handbook of Systematic Musicology*, ed. R. Bader (Berlin: Springer), 487–505. doi: 10.1007/978-3-662-55004-5_26

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*, Ph.D thesis, Islington: City University London.

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Ann. N Y. Acad. Sci.* 1423:378. doi: 10.1111/nyas.13654

Penrose, R. (1955). A generalized inverse for matrices. *Math. Proc. Camb. Philos. Soc* . 51, 406–413. doi: 10.1017/S0305004100030401

Poikonen, H., Alluri, V., Brattico, E., Lartillot, O., Tervaniemi, M., and Huotilainen, M. (2016a). Event-related brain responses while listening to entire pieces of music. *Neuroscience* 312, 58–73. doi: 10.1016/j.neuroscience.2015.10.061

Poikonen, H., Toiviainen, P., and Tervaniemi, M. (2016b). Early auditory processing in musicians and dancers during a contemporary dance piece. *Sci. Rep.* 6:33056. doi: 10.1038/srep33056

Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: Data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi: 10.1038/nn.3818

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of Functional Mri Data Analysis*. Cambridge: Cambridge: University Press. doi: 10.1017/CBO9780511895029

Popescu, T., Neuser, M. P., Neuwirth, M., Bravo, F., Mende, W., Boneh, O., et al. (2019). The pleasantness of sensory dissonance is mediated by musical style and expertise. *Sci. Rep.* 9:1070. doi: 10.1038/s41598-018-35873-8

Posse, S., Wiese, S., Gembris, D., Mathiak, K., Kessler, C., Grosse Ruyken, M. L., et al. (1999). Enhancement of BOLD-contrast sensitivity by single-shot multi-echo functional MR imaging. *Magn. Reson. Med.* 42, 87–97. doi: 10.1002/(SICI)1522-2594(199907)42:1<87::AID-MRM13>3.0.CO;2-O

Prince, J. S., Pyles, J. A., Tarr, M. J., and Kay, K. N. (2021). GLMsingle: A turnkey solution for accurate single-trial fMRI response estimates. *J. Vision* 21, 2831–2831. doi: 10.1167/jov.21.9.2831

Pruim, R. H., Mennes, M., Van Rooij, D., Llera, A., Buitelaar, J. K., and Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage* 112, 267–277. doi: 10.1016/j.neuroimage.2015.02.064

Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., and Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nat. Rev. Neurosci.* 22, 488–502. doi: 10.1038/s41583-021-00473-5

Raphael, C., and Stoddard, J. (2004). Functional Harmonic Analysis Using Probabilistic Models. *Comput. Music J.* 28, 45–52. doi: 10.1162/0148926041790676

Rieke, F., Bodnar, D. A., and Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. Lond. B Biol. Sci.* 262, 259–265. doi: 10.1098/rspb.1995.0204

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., and Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. *arXiv* [Preprint]. doi: 10.48550/arXiv.1803.05428

Rohrmeier, M. (2020). The syntax of jazz harmony: Diatonic tonality, phrase structure, and form. *Music Theory Anal.* 7, 1–63. doi: 10.11116/MTA.7.1.1

Rohrmeier, M., and Cross, I. (2008). "Statistical properties of tonal harmony in bach's chorales," in *Proceedings of the 10th International Conference on Music Perception and Cognition*, (Japan: Hokkaido University Sapporo), 619–627.

Rohrmeier, M., and Graepel, T. (2012). "Comparing feature-based models of harmony," in *Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval*, (London: Springer), 357–370.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., and Golestani, N. (2019). Cortical encoding of speech enhances task-relevant acoustic information. *Nat. Hum. Behav.* 3, 974–987. doi: 10.1038/s41562-019-0648-9

Sachs, M. E., Habibi, A., Damasio, A., and Kaplan, J. T. (2020). Dynamic intersubject neural synchronization reflects affective responses to sad music. *NeuroImage* 218:116512. doi: 10.1016/j.neuroimage.2019.116512

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., et al. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Computat. Biol.* 10:e1003412. doi: 10.1371/journal.pcbi.1003412

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., et al. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat. Commun.* 11:4238. doi: 10.1038/s41467-020-18037-z

Serrà, J., Corral, Á, Boguñá, M., Haro, M., and Arcos, J. L. (2012). Measuring the Evolution of Contemporary Western Popular Music. *Sci. Rep.* 2:521. doi: 10.1038/srep00521

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Sievers, B., Parkinson, C., Kohler, P. J., Hughes, J. M., Fogelson, S. V., and Wheatley, T. (2021). Visual and auditory brain areas share a representational structure that supports emotion perception. *Curr. Biol.* 31, 5192–5203.e4. doi: 10.1016/j.cub.2021.09.043

Singer, N., Jacoby, N., Lin, T., Raz, G., Shpigelman, L., Gilam, G., et al. (2016). Common modulation of limbic network activation underlies musical emotions as they unfold. *NeuroImage* 141, 517–529. doi: 10.1016/j.neuroimage.2016.07.002

Sohoglu, E., and Davis, M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *eLife* 9:e58077. doi: 10.7554/eLife.58077

Sonkusare, S., Breakspear, M., and Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends Cogn. Sci.* 23, 699–714. doi: 10.1016/j.tics.2019.05.004

Stephens, G. J., Honey, C. J., and Hasson, U. (2013). A place for time: The spatiotemporal structure of neural dynamics during natural audition. *J. Neurophysiol.* 110, 2019–2026. doi: 10.1152/jn.00268.2013

Stober, S. (2017). Toward studying music cognition with information retrieval techniques: Lessons learned from the openmiir initiative. *Front. Psychol.* 8:1255. doi: 10.3389/fpsyg.2017.01255

Sturm, B. (2018). "What do these 5,599,881 parameters mean?: An analysis of a specific LSTM music transcription model, starting with the 70,281 parameters of its softmax layer," in *International Conference on Computational Creativity*. (London UK: Centre for Digital Music Queen Mary University).

Sturm, B., Santos, J. F., Ben-Tal, O., and Korshunova, I. (2016a). Music transcription modelling and composition using deep learning. *arXiv* [Preprint].

Sturm, I., Lapuschkin, S., Samek, W., and Müller, K.-R. (2016b). Interpretable deep neural networks for single-trial EEG classification. *J. Neurosci. Methods* 274, 141–145. doi: 10.1016/j.jneumeth.2016.10.008

Sturm, B. L. (2013). The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv* [Preprint].

Sturm, I., Blankertz, B., and Curio, G. (2017). Multivariate EEG analysis reveals neural correlates for the differential perception of chord progressions. *Psychomusicol. Music Mind Brain* 27:281. doi: 10.1037/pmu0000196

Sturm, I., Blankertz, B., Potes, C., Schalk, G., and Curio, G. (2014). ECoG high gamma activity reveals distinct cortical representations of lyrics passages, harmonic and timbre-related changes in a rock song. *Front. Hum. Neurosci.* 8:798. doi: 10.3389/fnhum.2014.00798

Sturm, I., Dähne, S., Blankertz, B., and Curio, G. (2015). Multi-variate eeg analysis as a novel tool to examine brain responses to naturalistic music stimuli. *PLoS One* 10:e0141281. doi: 10.1371/journal.pone.0141281

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE International Conference on Computer Vision*, (New York, NY: IEEE), 843–852. doi: 10.1109/ICCV.2017.97

Taylor, A. J., Kim, J. H., and Ress, D. (2018). Characterization of the hemodynamic response function across the majority of human cerebral cortex. *NeuroImage* 173, 322–331. doi: 10.1016/j.neuroimage.2018.02.061

Theunissen, F. E., Sen, K., and Doupe, A. J. (2000). Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds. *J. Neurosci.* 20, 2315–2331. doi: 10.1523/JNEUROSCI.20-06-02315.2000

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tikhonov, A. N. (1943). On the stability of inverse problems. *Proc. USSR Acad. Sci.* 39, 195–198.

Tikhonov, A. N., Goncharsky, A., Stepanov, V., and Yagola, A. G. (1995). *Numerical Methods for the Solution of Ill-Posed Problems*. Germany: Springer Science and Business Media. doi: 10.1007/978-94-015-8480-7

Toiviainen, P., Alluri, V., Brattico, E., Wallentin, M., and Vuust, P. (2014). Capturing the musical brain with Lasso: Dynamic decoding of musical features from fMRI data. *NeuroImage* 88, 170–180. doi: 10.1016/j.neuroimage.2013.11.017

Toiviainen, P., Burunat, I., Brattico, E., Vuust, P., and Alluri, V. (2020). The chronnectome of musical beat. *Neuroimage* 216:116191. doi: 10.1016/j.neuroimage.2019.116191

Trochidis, K., and Bigand, E. (2013). Investigation of the Effect of Mode and Tempo on Emotional Responses to Music Using EEG Power Asymmetry. *J. Psychophysiol.* 27, 142–148. doi: 10.1027/0269-8803/a000099

Trost, W., Frühholz, S., Cochrane, T., Cojan, Y., and Vuilleumier, P. (2015). Temporal dynamics of musical emotions examined through intersubject synchrony of brain activity. *Soc. Cogn.Affect. Neurosci.* 10, 1705–1721. doi: 10.1093/scan/nsv060

Tschacher, W., Greenwood, S., Egermann, H., Wald-Fuhrmann, M., Czepiel, A., Tröndle, M., et al. (2021). Physiological synchrony in audiences of live concerts. *Psychol. Aesthet. Creat. Arts* doi: 10.1037/aca0000431 [Epub ahead of print].

Tzanetakis, G., and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Trans. Speech Audio process.* 10, 293–302. doi: 10.1109/TSA.2002.800560

Uludağ, K., Müller-Bierl, B., and Uğurbil, K. (2009). An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging. *NeuroImage* 48, 150–165. doi: 10.1016/j.neuroimage.2009.05.051

van de Wiel, M. A., Van Nee, M. M., and Rauschenberger, A. (2021). Fast Cross-validation for Multi-penalty High-dimensional Ridge Regression. *J. Comput. Graphical Stat.* 30, 835–847. doi: 10.1080/10618600.2021.1904962

VanRullen, R., and Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Commun. Biol.* 2:193. doi: 10.1038/s42003-019-0438-y

Varoquaux, G., and Poldrack, R. A. (2019). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Curr. Opin. Neurobiol.* 55, 1–6. doi: 10.1016/j.conb.2018.11.002

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage* 145, 166–179. doi: 10.1016/j.neuroimage.2016.10.038

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., et al. (2018). Mapping between fMRI responses to movies and their natural language annotations. *NeuroImage* 180, 223–231. doi: 10.1016/j.neuroimage.2017.06.042

Vu, V. Q., Ravikumar, P., Naselaris, T., Kay, K. N., Gallant, J. L., and Yu, B. (2011). Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models. *Ann. Appl. Stat.* 5:1159. doi: 10.1214/11-AOAS476

Vuust, P., Heggli, O. A., Friston, K. J., and Kringelbach, M. L. (2022). Music in the brain. *Nat. Rev. Neurosci.* 23, 287–305. doi: 10.1038/s41583-022-00578-5

Wang, X., Liu, W., Toiviainen, P., Ristaniemi, T., and Cong, F. (2020). Group analysis of ongoing EEG data based on fast double-coupled nonnegative tensor decomposition. *J. Neurosci. Methods* 330:108502. doi: 10.1016/j.jneumeth.2019.108502

Werbos, P. (1974). *Beyond Regression New Tools for Prediction and Analysis in the Behavioral Sciences*, Ph.D thesis, Washington: Harvard University

Werbos, P. J. (1994). *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. Hoboken: John Wiley and Sons.

White, C. W. (2014). Changing Styles, Changing Corpora, Changing Tonal Models. *Music Percept.* 31, 244–253. doi: 10.1525/mp.2014.31.3.244

Whitehead, J. C., and Armony, J. L. (2018). Singing in the brain: Neural representation of music and voice as revealed by fMRI. *Hum. Brain Mapp.* 39, 4913–4924. doi: 10.1002/hbm.24333

Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* 29, 477–505. doi: 10.1146/annurev.neuro.29.051605.113024

Wundt, W. M. (1874). *Principles of Physiological Psychology*. Germany: Wilhelm Engelmann.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and Dicarlo, J. J. (2014).). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yang, L.-C., Chou, S.-Y., and Yang, Y.-H. (2017). Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv* [Preprint].

Yang, Y.-H., and Chen, H. H. (2011). *Music Emotion Recognition*. Florida: CRC Press. doi: 10.1201/b10731

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Zadeh, L. (1956). On the Identification Problem. *IRE Trans. Circuit Theory* 3, 277–281. doi: 10.1109/TCT.1956.1086328

Zamm, A., Debener, S., Konvalinka, I., Sebanz, N., and Knoblich, G. (2020). The sound of silence: An EEG study of how musicians time pauses in individual and joint music performance. *Soc. Cogn. Affect. Neurosci.* 16, 31–42. doi: 10.1093/scan/nsaa096

Zatorre, R. (2005). Music, the food of neuroscience? *Nature* 434, 312–315. doi: 10.1038/434312a

Zatorre, R. J., and Salimpoor, V. N. (2013). From perception to pleasure: Music and its neural substrates. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10430–10437. doi: 10.1073/pnas.1301228110

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Zuk, N. J., Teoh, E. S., and Lalor, E. C. (2020). EEG-based classification of natural sounds reveals specialized responses to speech and music. *NeuroImage* 210:116558. doi: 10.1016/j.neuroimage.2020.116558