TeaP2025, Goethe-University Frankfurt, FFM, DE | 12 March 2025

Representational Gradients of Emotion-relevant Musical Information in the Human Cerebral Cortex

Seung-Goo Kim[1], Tobias Overath[2], & Daniela Sammler[1]

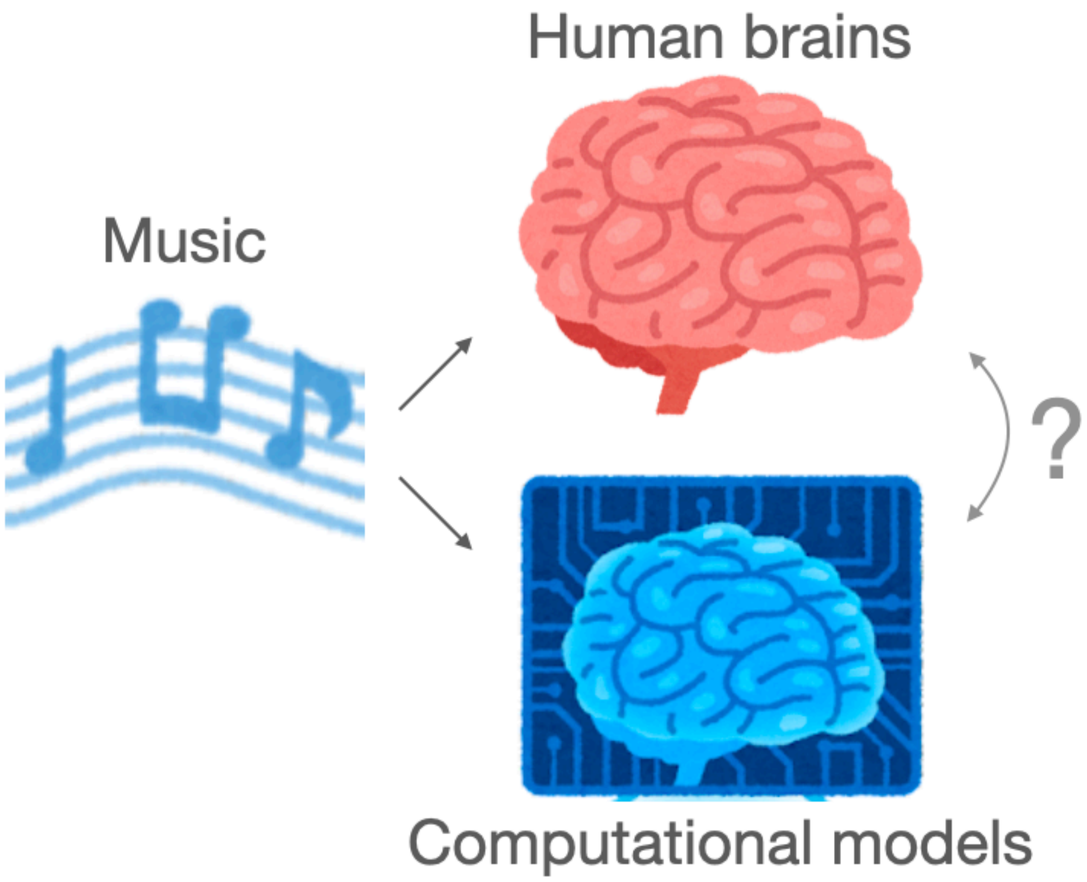
[1]**Research Group Neurocognition of Music and Language**
Max Planck Institute for Empirical Aesthetics, **Frankfurt am Main, Germany**
[2]Department of Psychology & Neuroscience, Duke University, US

https://voxpopulisphere.com/2015/11/01/audio-what-ancient-greek-music-actually-sounded-like/

How does music evoke emotions *via* the brain?

Music → Brain → Emotion

Research Questions



- **Q1**: How are ***felt* emotions** and **musical enjoyment** associated with neural activity over time?
- **Q2:** Would **increasingly abstract representations of music** in different layers of the CNN be encoded along the cortical gradient axis of abstraction?
- **Q3**: How do **layer-specific CNN embeddings** predict human behavioral ratings of musical emotions?

Kim, 2022. *Front. Neurosci.*
https://www.irasutoya.com/

Methods

CNN embedding for music emotion *recognition*

Potentially mid/high-level representation of music signal

An "deep audio semantic" model called "VGGish"

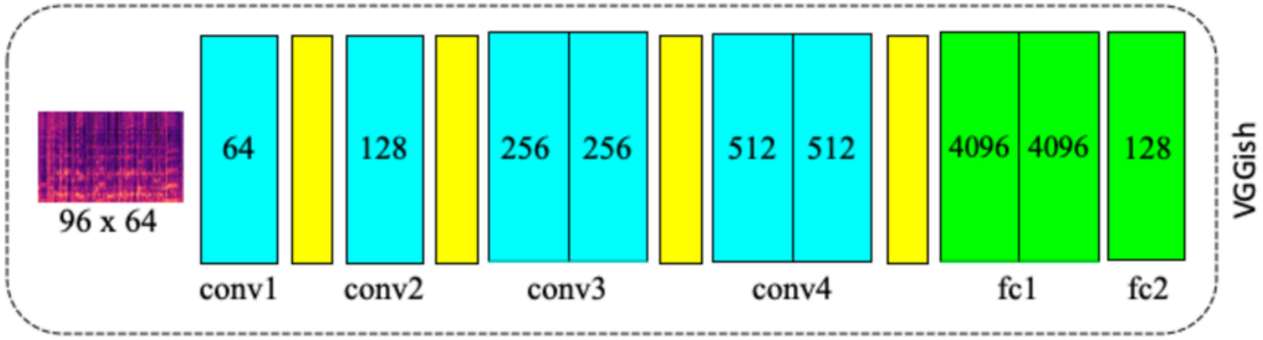
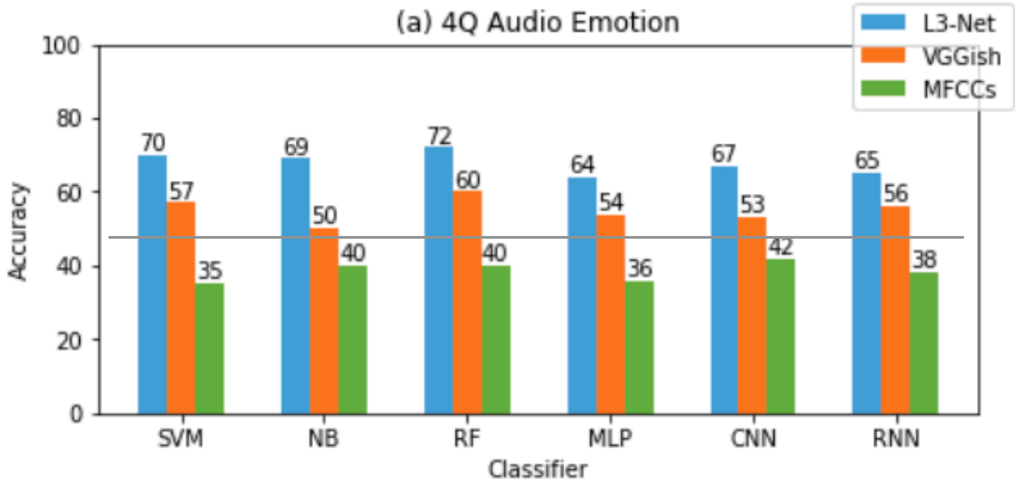


Diagram from Koh & Dubnov, 2021, *ACA;*
"VGGish" audio model is by Hershey et al. (Google), 2017, *ICCASP.*
Supervised learning (human annotations) on YouTube clips (5M hours, 30K tags)

4Q Audio Emotion Dataset: 255 music clips (30 s) for **Arousal-Valence quadrants**



Koh & Dubnov, 2021, *ACA*

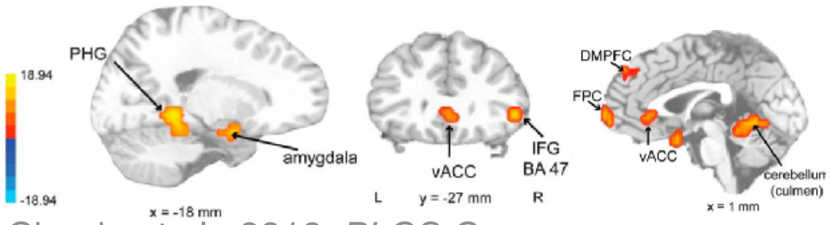
Deep audio semantic models carry more information related to expressed emotions than a traditional audio descriptor.

Example: Expressive vs. Mechanical

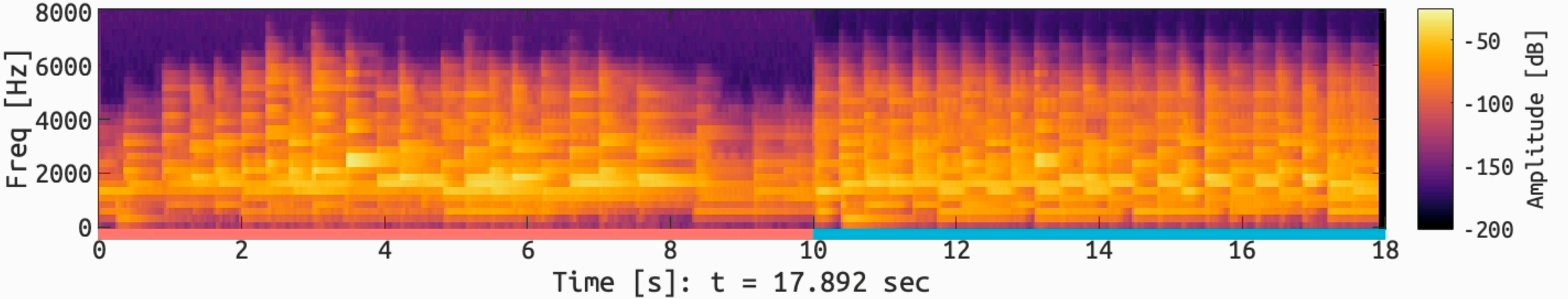
Chapin et al., 2010, PLOS ONE.

[EXPRESSIVE]: "Frédéric Chopin's Etude in E major, Op.10, No. 3 was performed by an undergraduate piano major (female, 22 years old) on a Kawai CA 950 digital piano"

Expressive > Mechanical



Chapin et al., 2010, *PLOS One*.

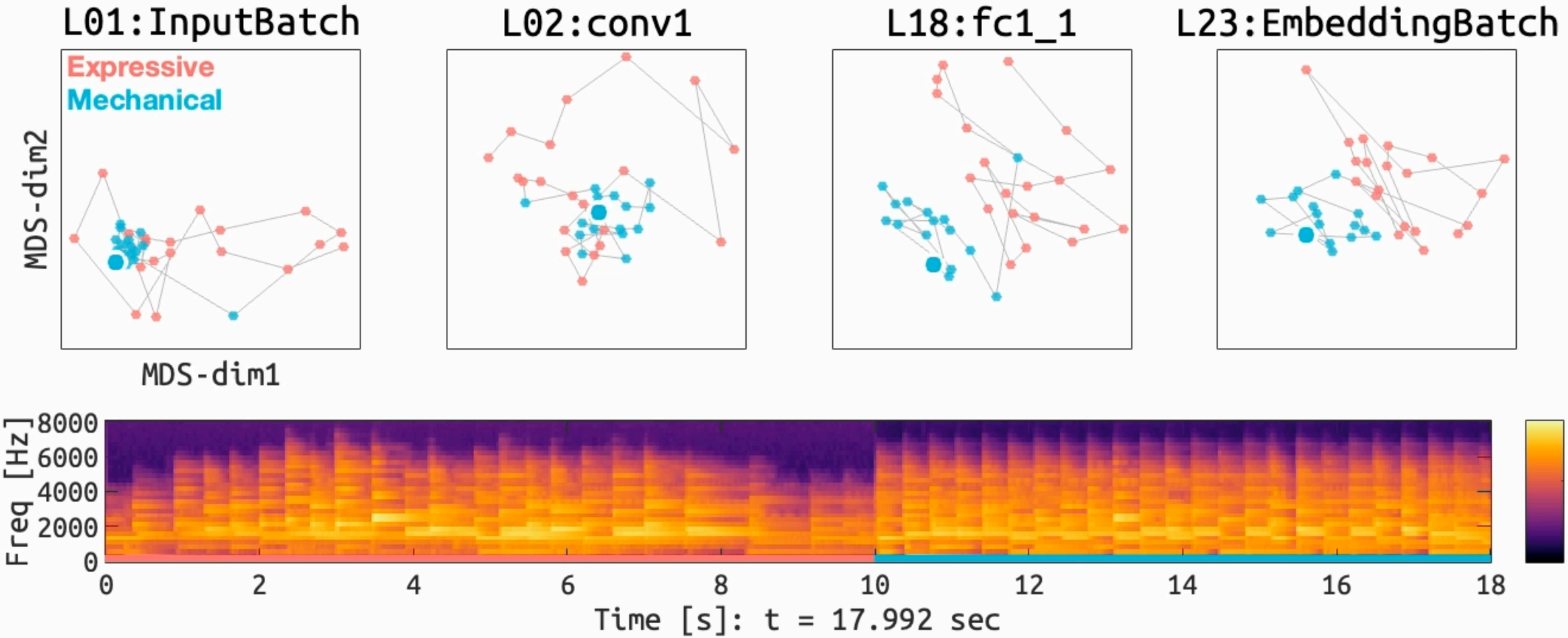


[MECHANICAL]: "The MIDI (Musical Instrument Digital Interface) onset velocity (key pressure) of each note (correlating with sound level) was set to 64 (range 0–128), and pedal information was eliminated."

Audio source: https://doi.org/10.1371/journal.pone.0013812

Example: Expressive vs. Mechanical

VGGish representations at various layers (from 1 to 23)



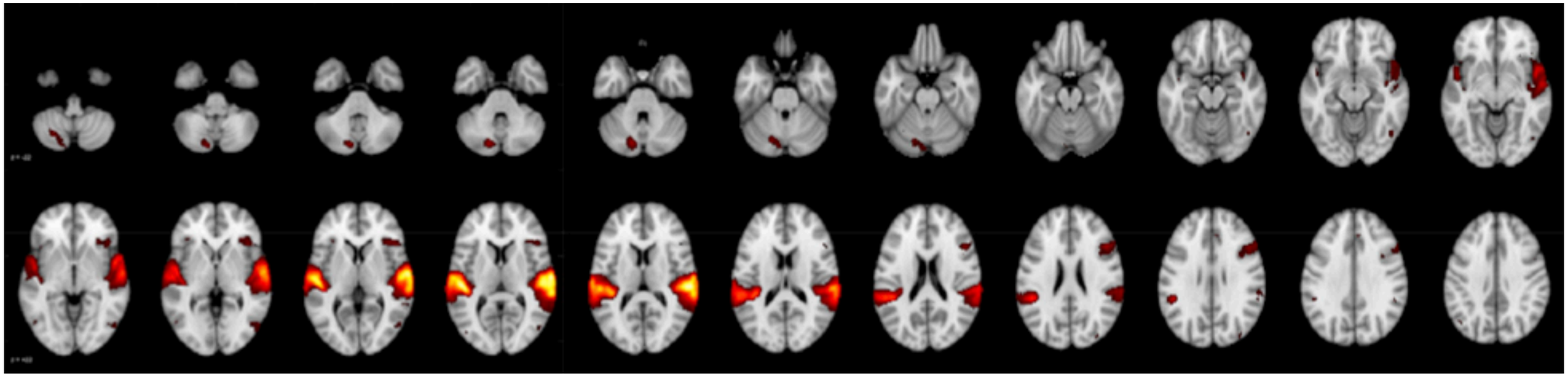
Audio source: https://doi.org/10.1371/journal.pone.0013812

Original study

Sachs et al., 2020, *NeuroImage*.

https://openneuro.org/datasets/ds003085

Inter-subject correlation during a "sad" piece of music: r ~ [0, 0.16], cluster-*P* < 0.05

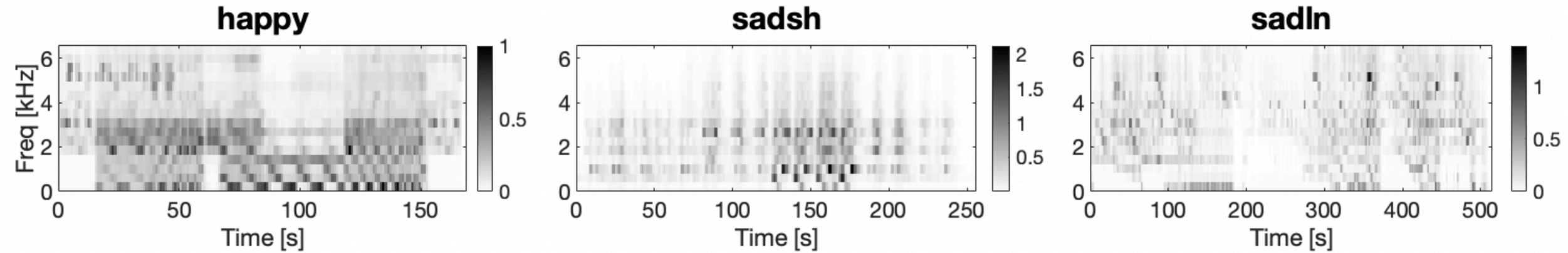


Sachs et al., 2020, *NeuroImage.*

Stimuli

Sachs et al., 2020, *NeuroImage*.

- **Happy** [2 min 48 sec]: Lullatone's "Race against the Sunset"
- **Sad-short** [4 min 16 sec]: Olafur Arnalds's "Frysta"
- **Sad-long** [8 min 35 sec]: Michael Kamen's "Discovery of the Camp"



Kim et al., *In prep*.

Participants & protocol

Sachs et al., 2020, *NeuroImage*.

- N = 40 (21 female, mean age = 24.1 ± 6.24 from LA)
  - Unfamiliar with 3 stimuli and reported "intended" emotions from 60-s excerpts

Passive listening with eyes open

<Not kids: just Japanese illustration of adults>

Rating with a slider

<Not kids: just Japanese illustration of adults>

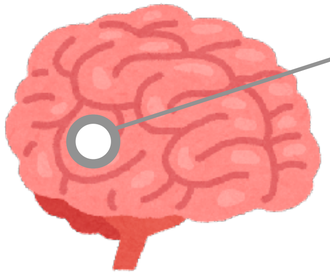
- "The intensity of felt sadness or happiness" (***Emotionality***)
- "The intensity of enjoyment" (***Enjoyment***)

https://www.irasutoya.com/

Encoding analysis

Training set

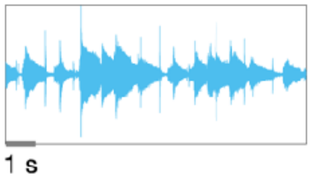
Neural responses



$\mathbf{y}_1$

Music #1

$\mathbf{X}_1$

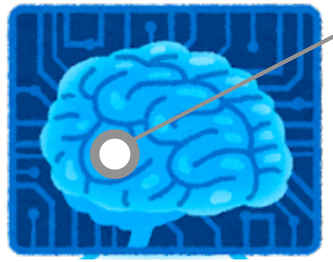


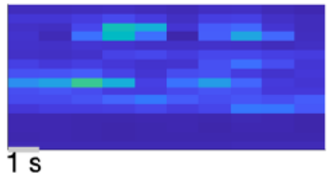
Weight fitting: $\hat{\mathbf{w}}_1 = \mathbf{Z}_1^{+}\mathbf{y}_1$

$\mathbf{y}_1 = \mathbf{Z}_1\mathbf{w}_1 + \epsilon$

Delays

$\mathbf{z}_1 = \mathscr{J}'(\mathbf{X}_1)$

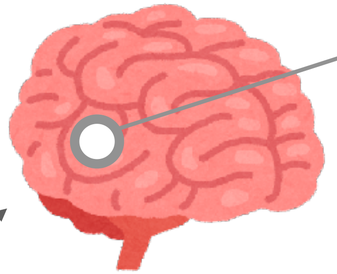




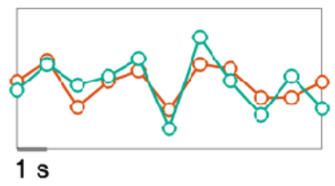
Model features

Test set

Neural responses

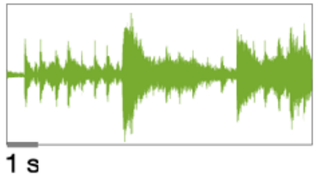


$\mathbf{y}_2$



Music #2

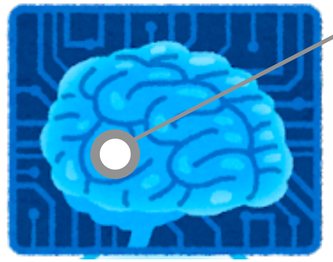
$\mathbf{X}_2$



Prediction: $\hat{\mathbf{y}}_2 = \mathbf{Z}_2\hat{\mathbf{w}}_1$

Delays

$\mathbf{z}_2 = \mathscr{J}'(\mathbf{X}_2)$



Model features

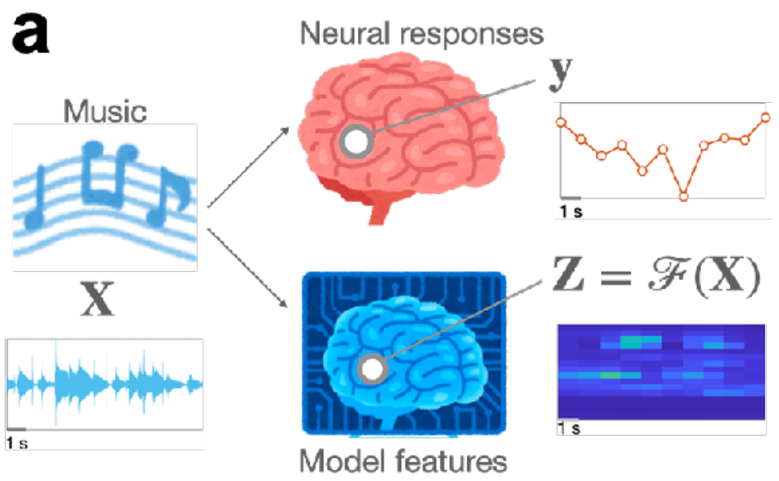
Kim, 2022. *Front. Neurosci.*

Layer-specific profile

How do we map VGGish-layer-specific encoding on the cortex?

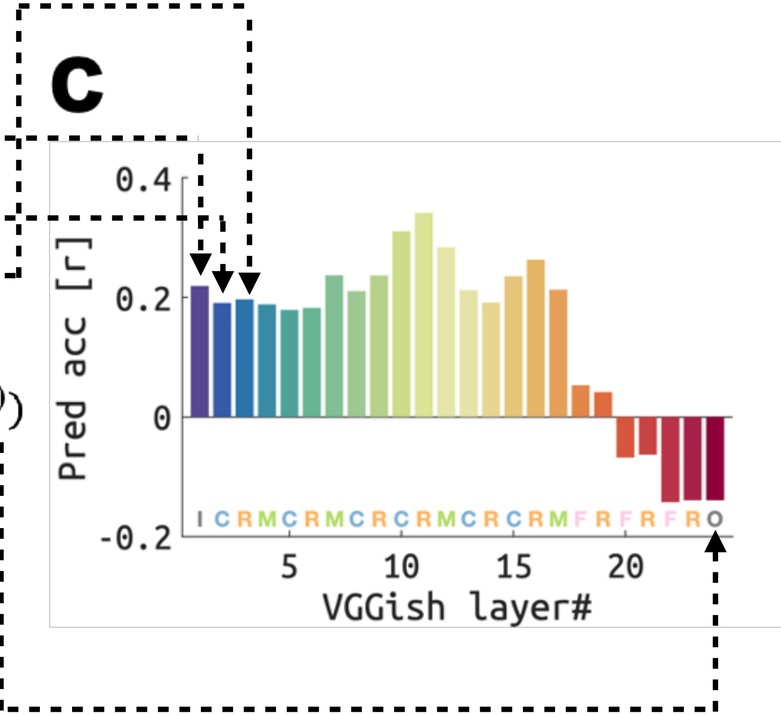
How good the "Layer X" is for this brain area?

A "profile" of layer-specific prediction accuracies



b

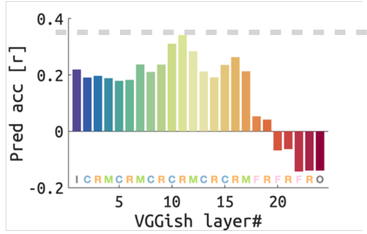
$$\mathbf{y}^{(1)} = \mathcal{F}_1(\mathbf{X}^{(1)})\mathbf{b}_1 + \varepsilon \rightarrow r_1 = \mathrm{corr}(\hat{\mathbf{y}}_1^{(2)}, \mathbf{y}^{(2)})$$
$$\mathbf{y}^{(1)} = \mathcal{F}_2(\mathbf{X}^{(1)})\mathbf{b}_2 + \varepsilon \rightarrow r_2 = \mathrm{corr}(\hat{\mathbf{y}}_2^{(2)}, \mathbf{y}^{(2)})$$
$$\mathbf{y}^{(1)} = \mathcal{F}_3(\mathbf{X}^{(1)})\mathbf{b}_3 + \varepsilon \rightarrow r_3 = \mathrm{corr}(\hat{\mathbf{y}}_3^{(2)}, \mathbf{y}^{(2)})$$
$$\vdots$$
$$\mathbf{y}^{(1)} = \mathcal{F}_{24}(\mathbf{X}^{(1)})\mathbf{b}_{24} + \varepsilon \rightarrow r_{24} = \mathrm{corr}(\hat{\mathbf{y}}_{24}^{(2)}, \mathbf{y}^{(2)})$$

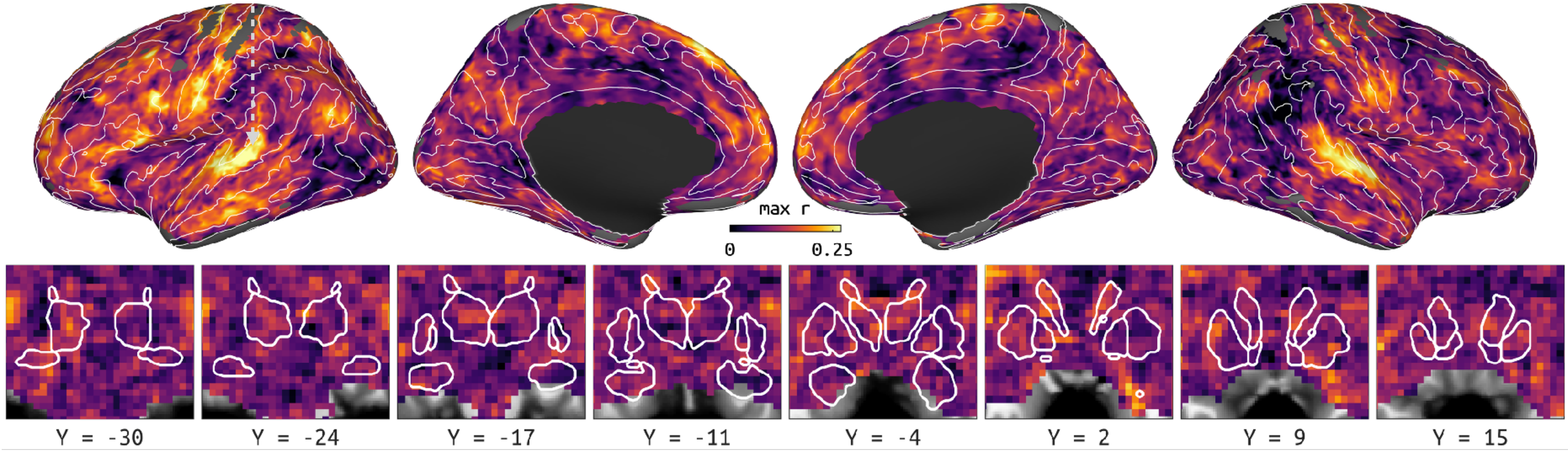


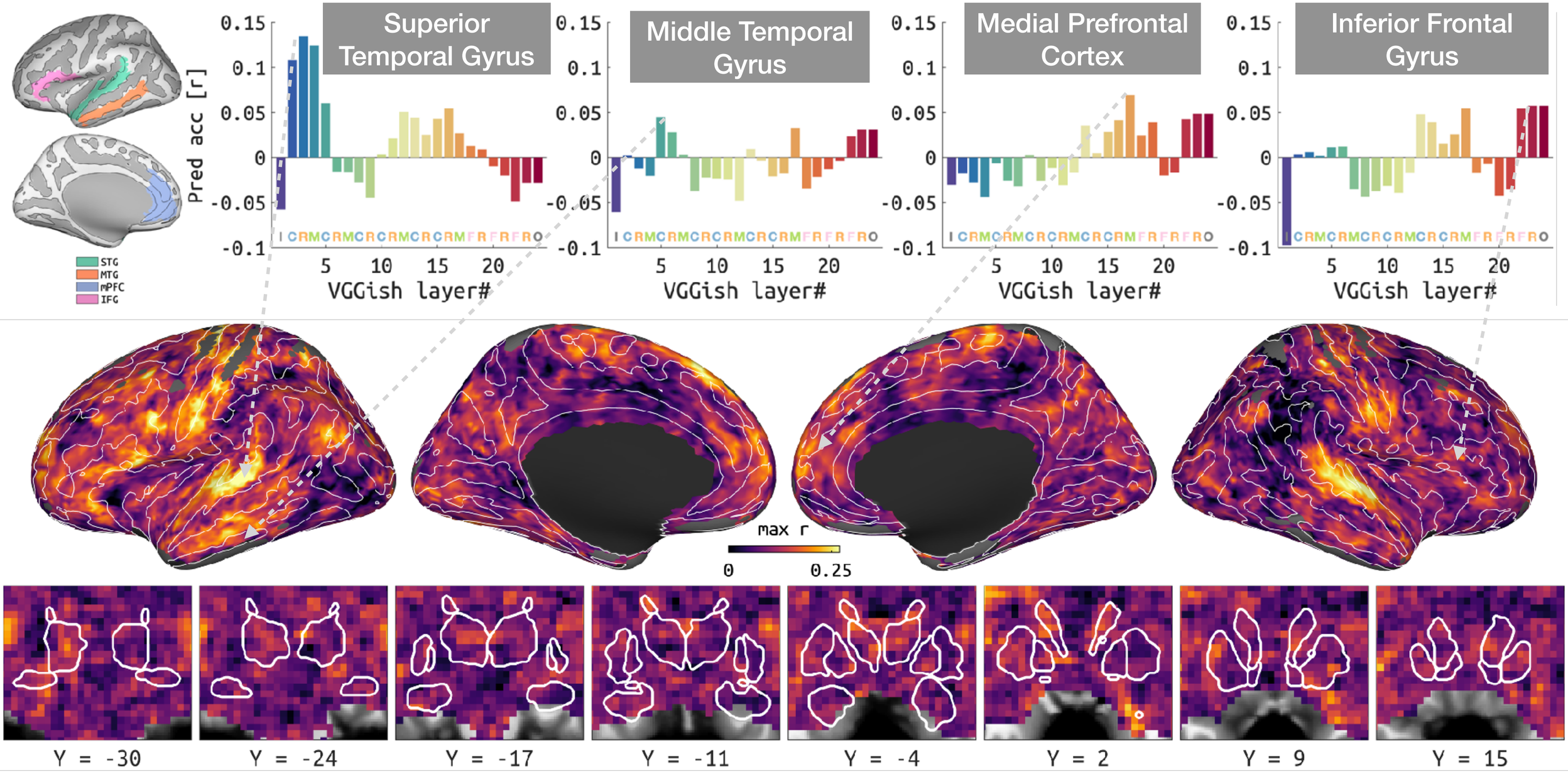
Results

Maximal prediction accuracy 🎶→🧠

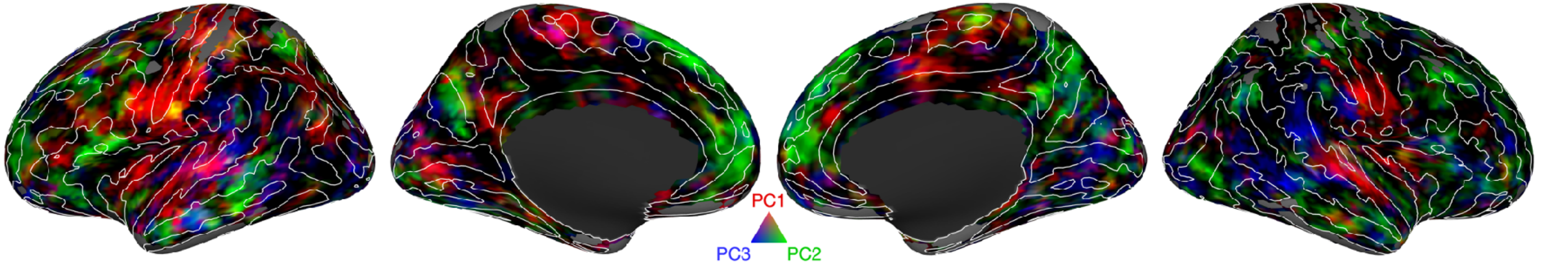
c

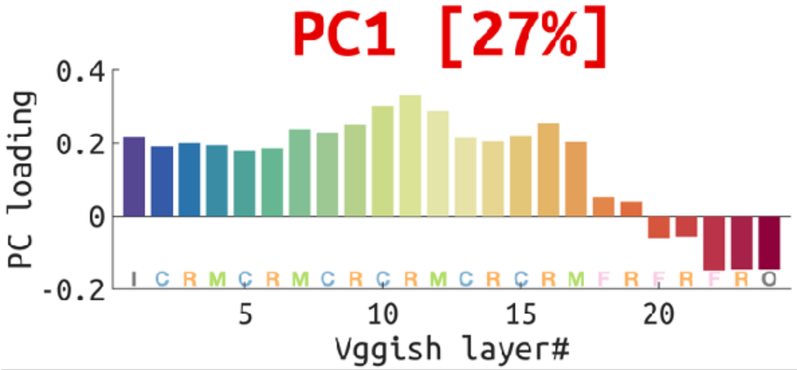


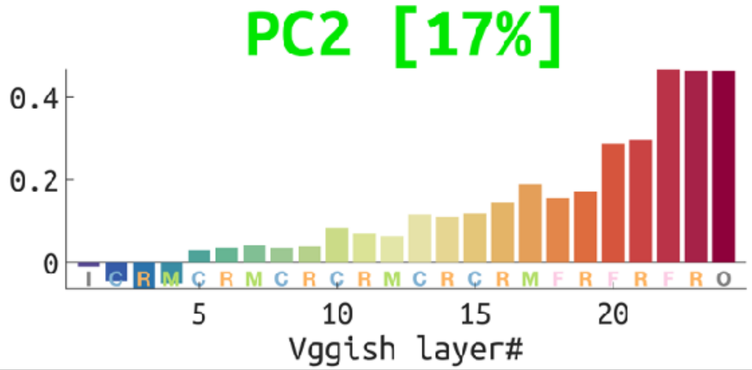


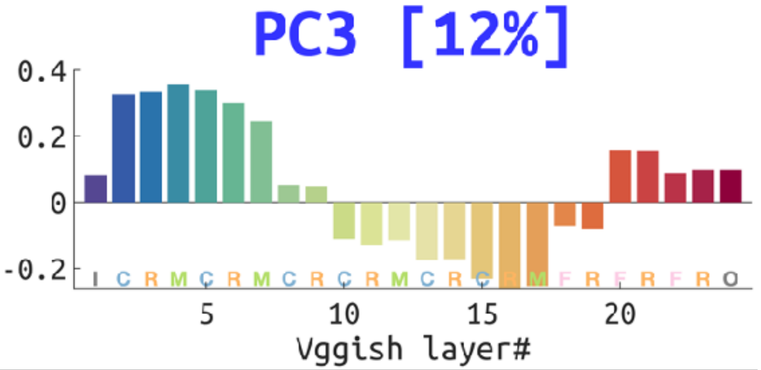

STG
MTG
mPFC
IFG
Pred acc [r]
Superior Temporal Gyrus
Middle Temporal Gyrus
Medial Prefrontal Cortex
Inferior Frontal Gyrus
VGGish layer#
max r
0
0.25
Y = -30
Y = -24
Y = -17
Y = -11
Y = -4
Y = 2
Y = 9
Y = 15

Topography of CNN-layer-specific encoding 🎶→🧠



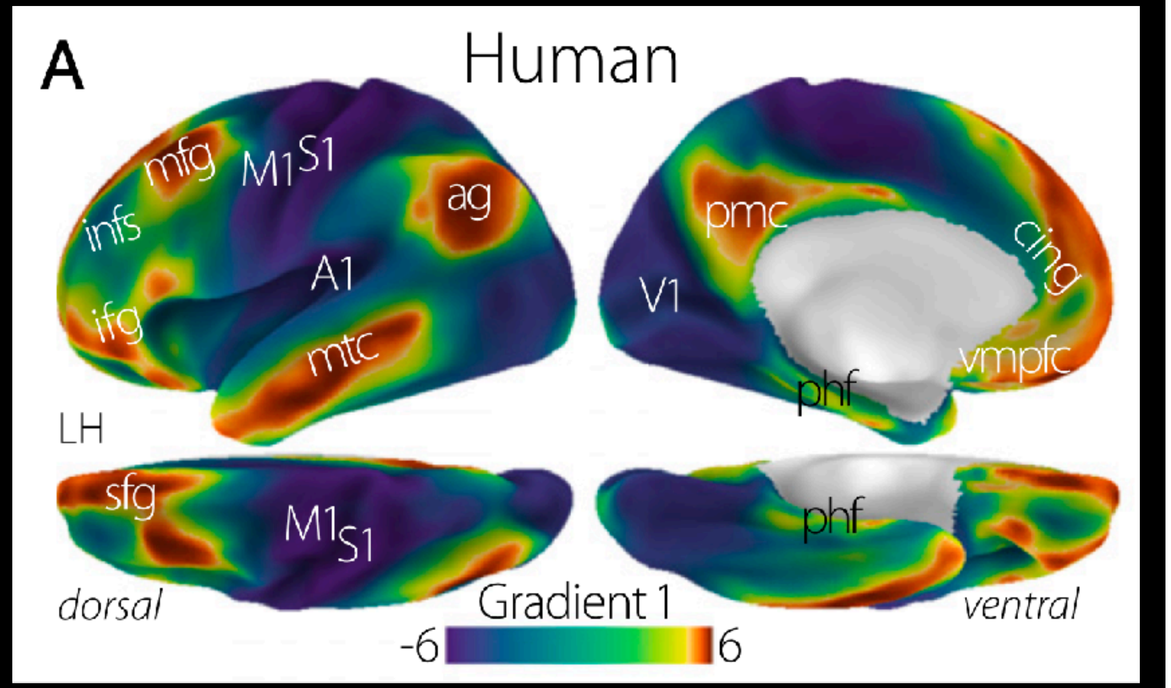




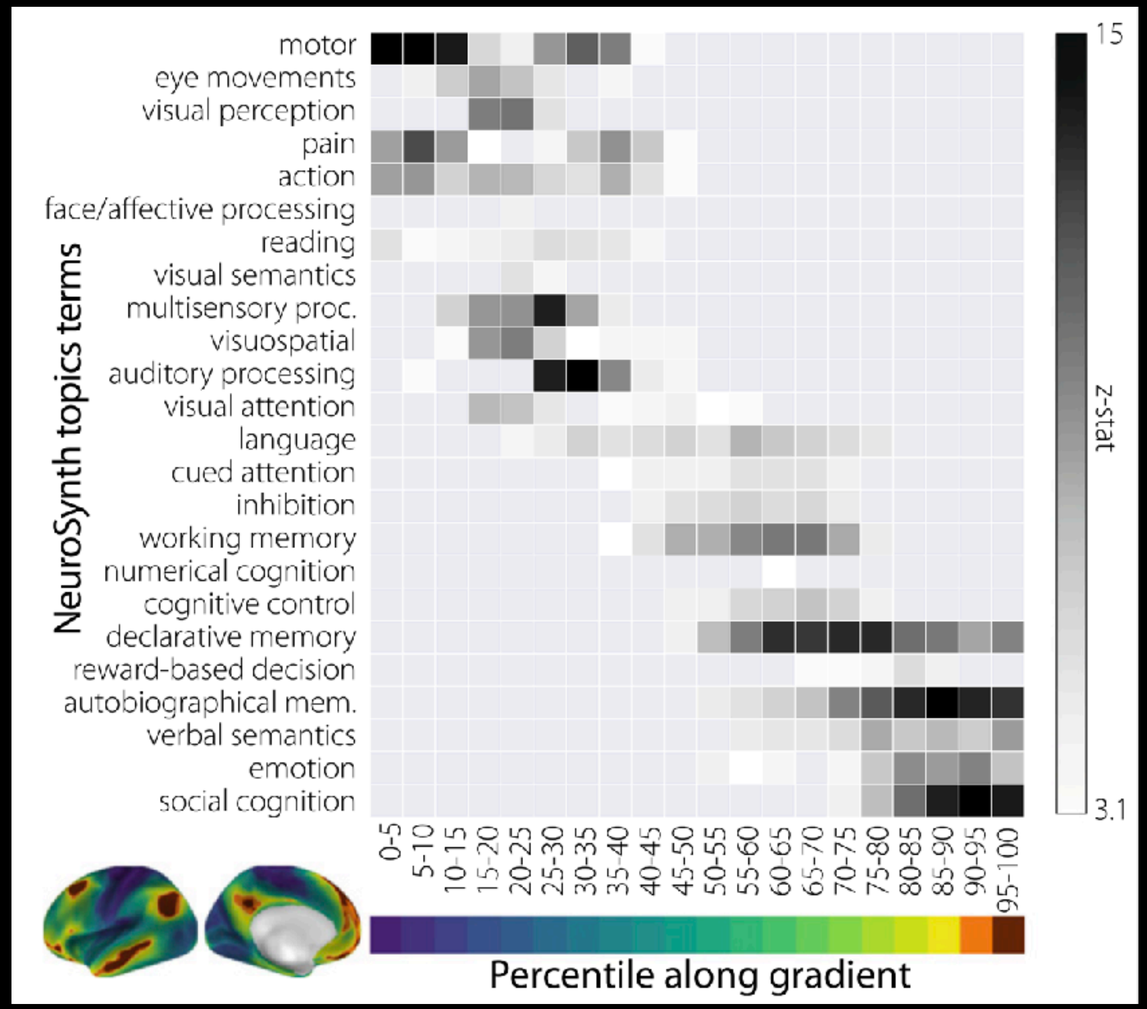


What is already known about the cortical topography of information abstraction?

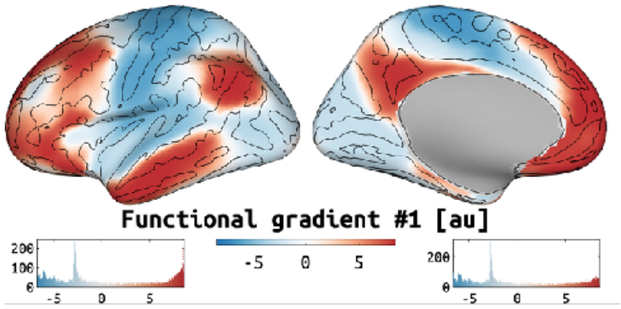
Margulies et al., 2016, PNAS



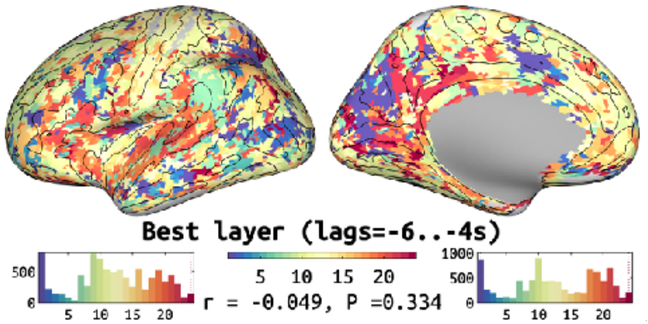
Topography found from resting-state functional connectivity

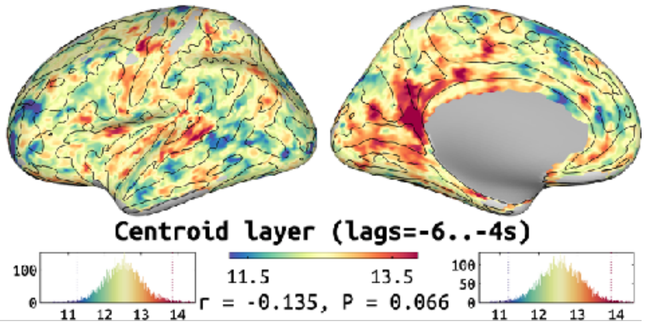


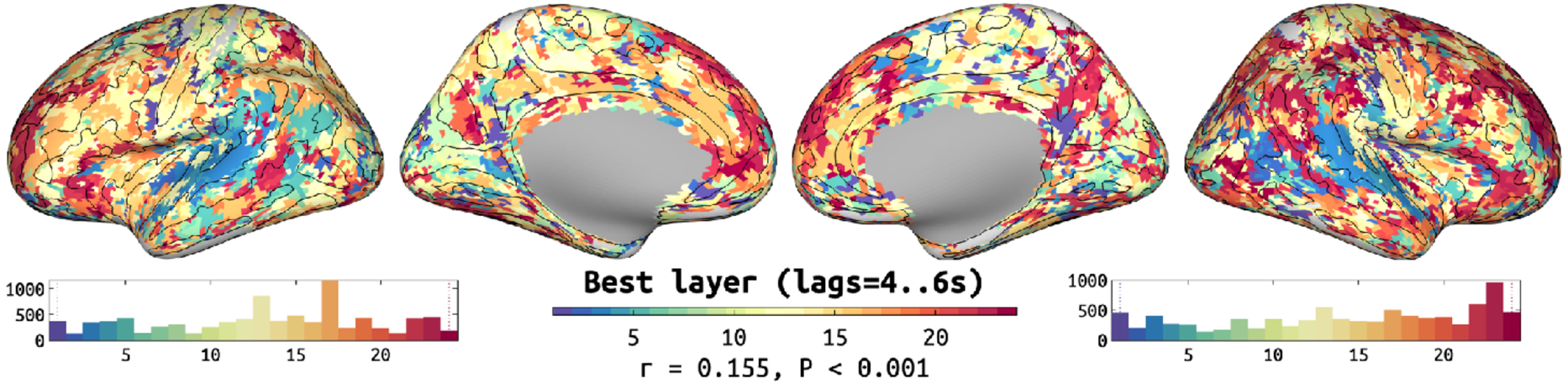
RSFC G1 vs. the "best" CNN-layer map

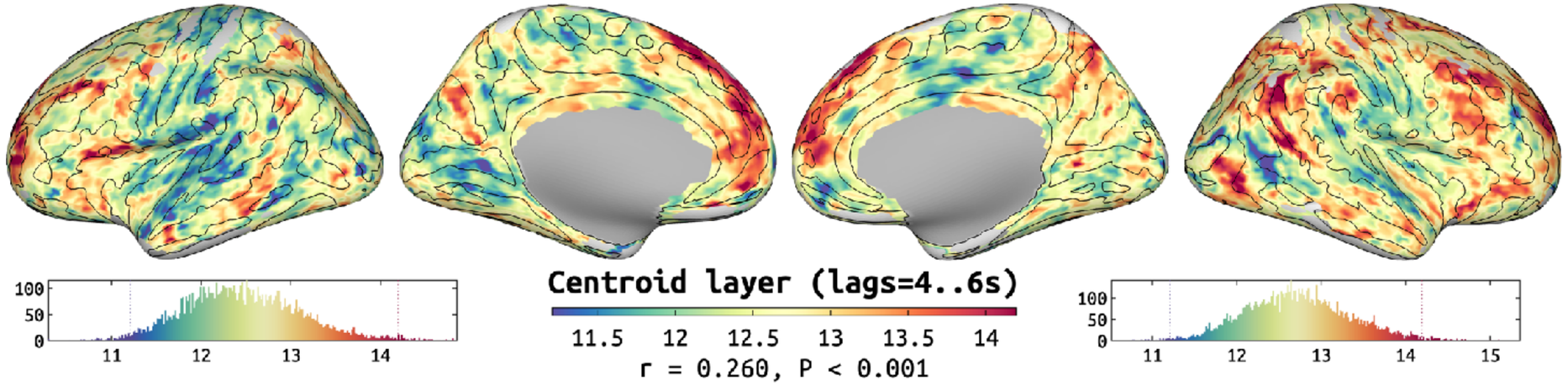


Is it "*just*" because of fMRI noise structures? 🧐

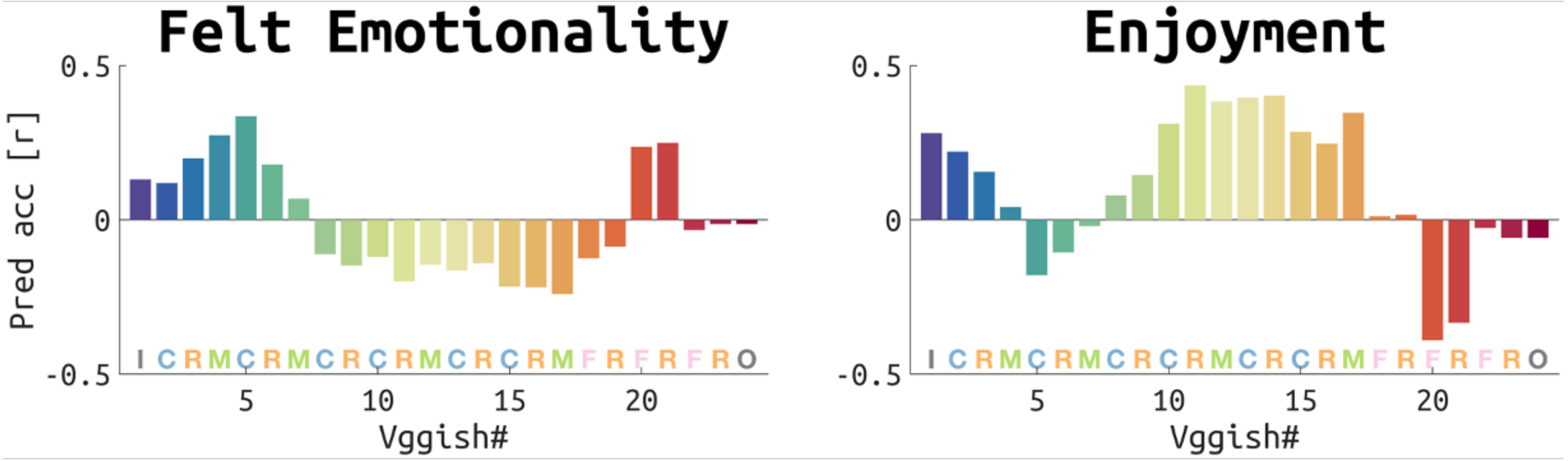








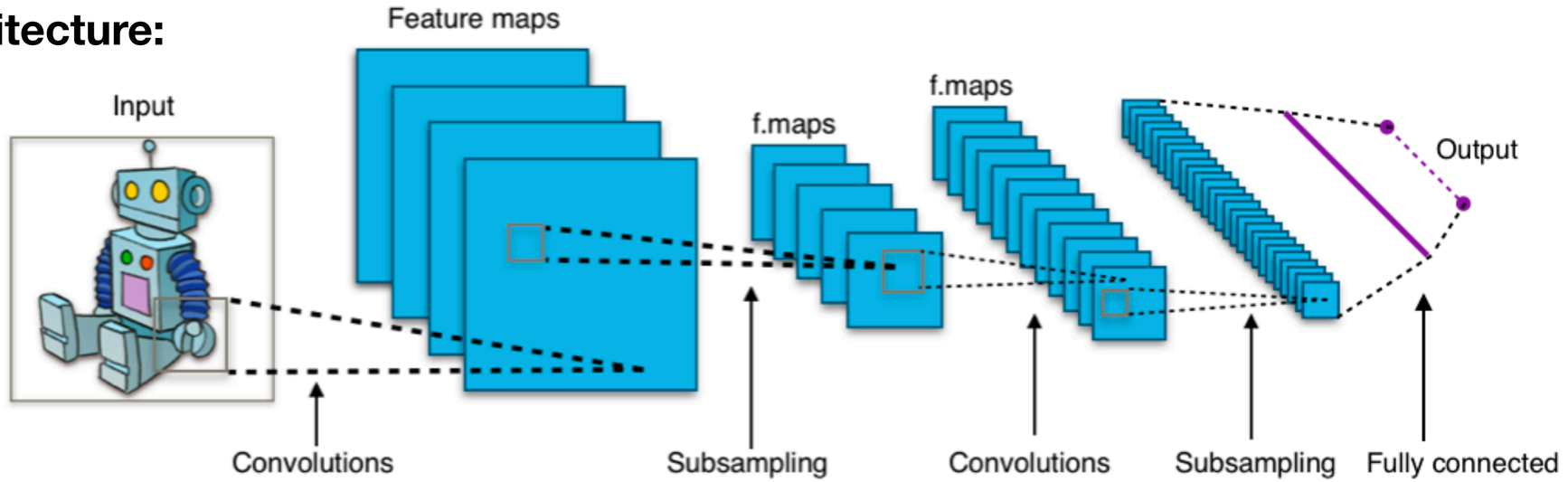
Distinctive patterns of Emotionality and Enjoyment 😀😢→🧠



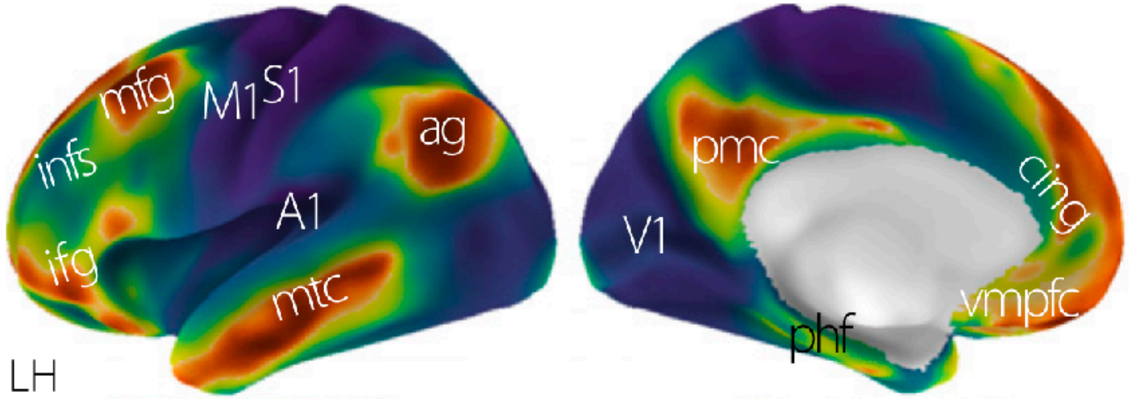
Discussion

Abstraction of sounds

A CNN architecture:

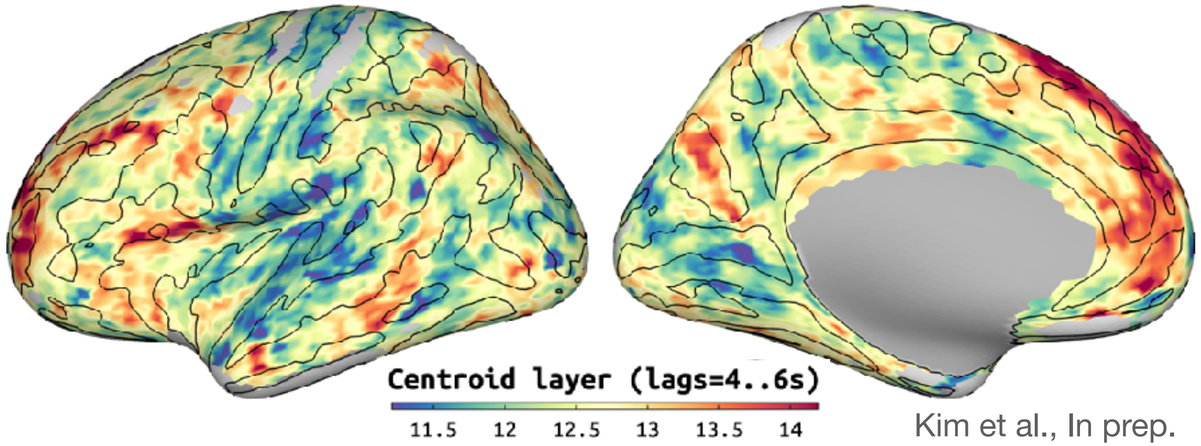


Abstraction in the human cerebral cortex:



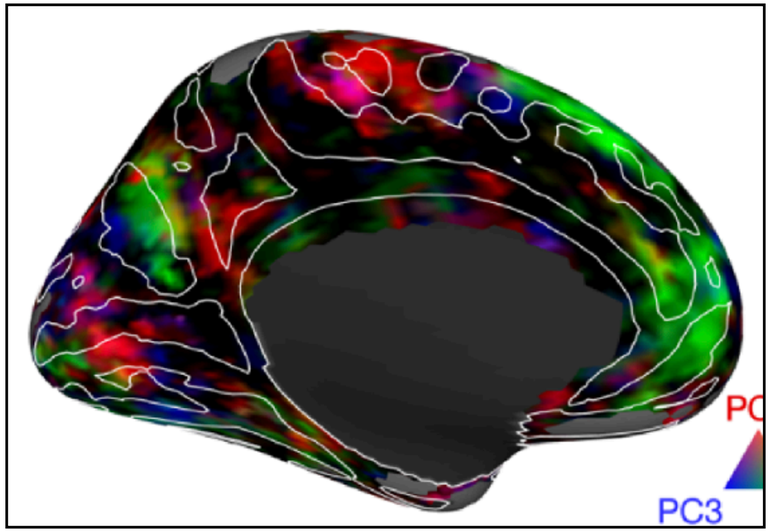
Marguleius et al., 2016, PNAS

Musical abstraction in the human cerebral cortex:

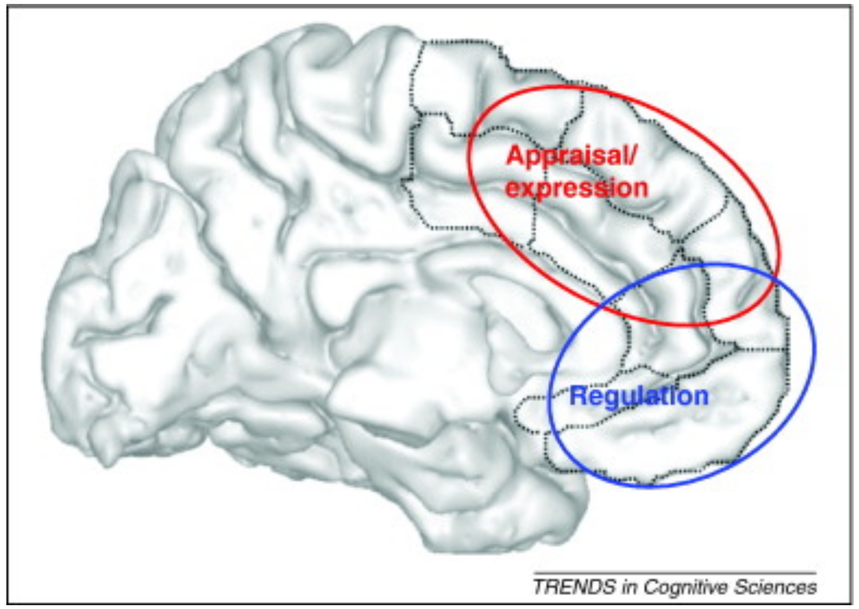


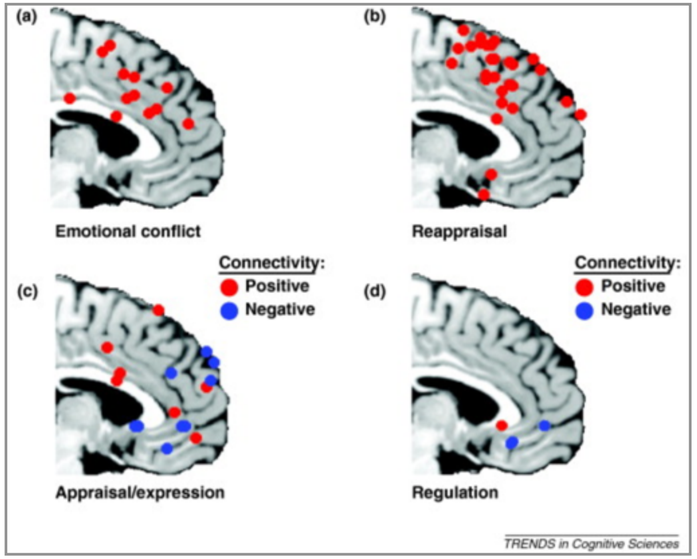
Kim et al., In prep.

mPFC and emotional processing



Kim et al., In prep.

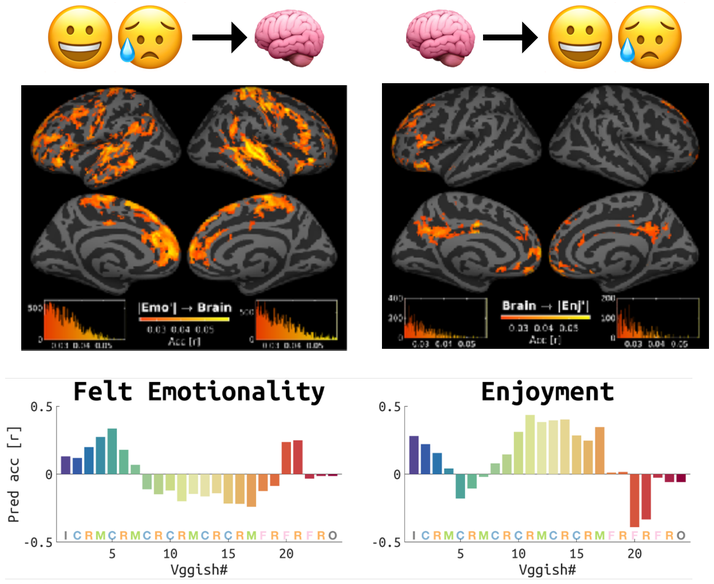




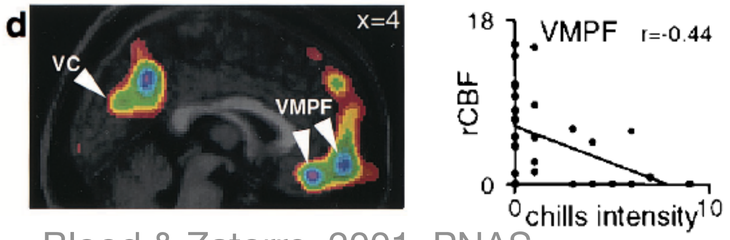
Etkin, Egner, Kalisch, 2010, *Trends in Cognitive Sciences.*

Audio semantic model changes were encoded in the mPFC, which showed a sensitivity to musical structures ("boundaries").

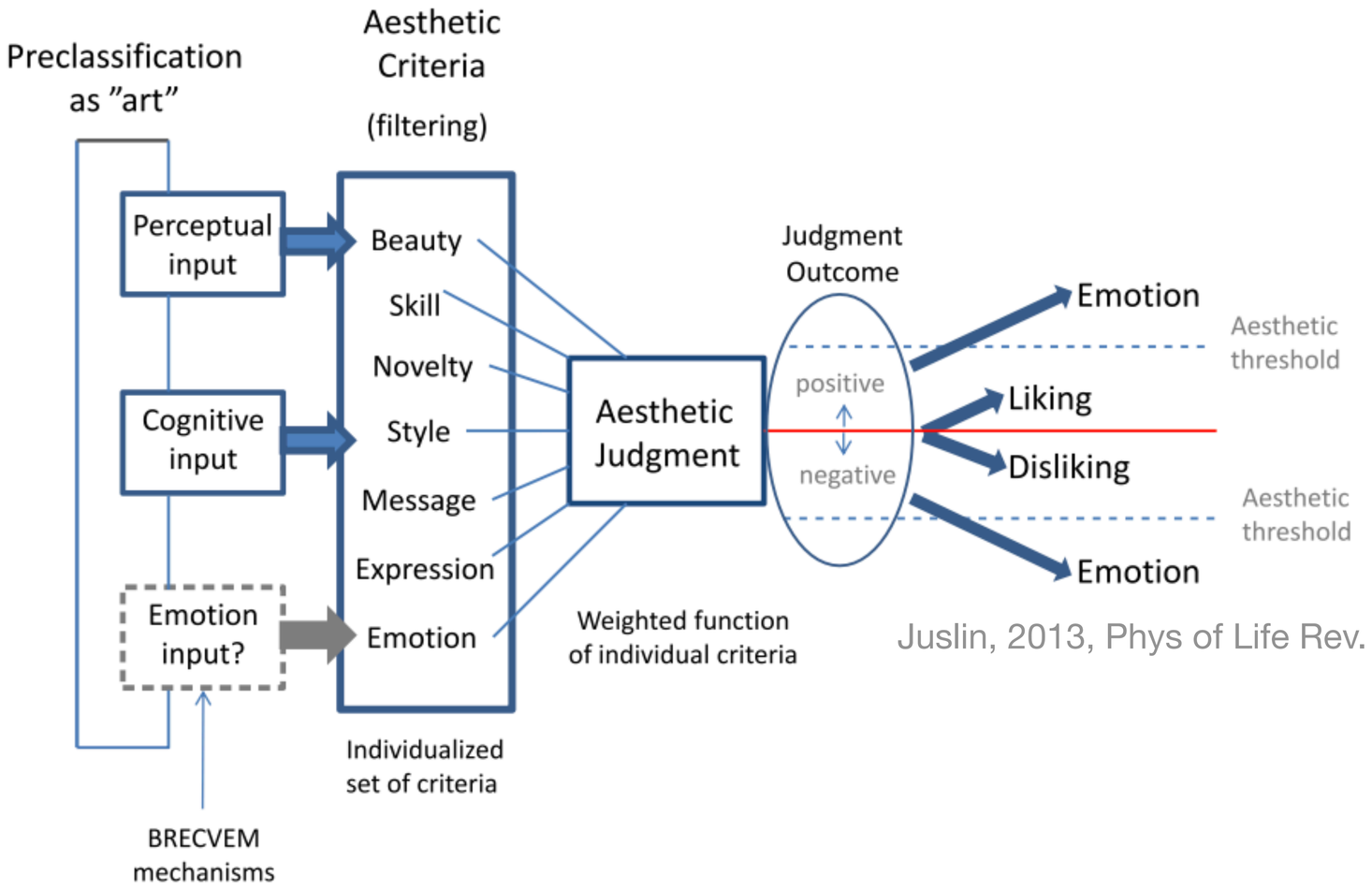
Different encoding of emotionality & enjoyment



Kim et al., In prep.



Blood & Zatorre, 2001, PNAS



Juslin, 2013, Phys of Life Rev.

vmPFC activity was followed by Enjoyment rating changes.

Conclusions

- Audio CNN embeddings are sensitive to information that is relevant for emotional responses, **beyond low-level audio features**.
- The abstraction in CNN shows high similarity to **the known cortical topography of information abstraction**, as well as relevance to behavioral ratings of emotional responses.
- Two continuous ratings (*Emotionality* and *Enjoyment*) were differentially encoded in the brain and predicted by different CNN layers, potentially reflecting **distinct mechanisms** of ***felt* emotions** and **aesthetic judgements**.

Thank you for your attention! (and time for discussion🤓)

https://seunggookim.github.io/

Dr. Daniela Sammler
MPI-EA, Frankfurt, Germany

Dr. Tobias Overath
Duke University, NC, USA

Dr. Tom H. Fritz
MPI-CBS, Leipzig, Germany

https://www.aesthetics.mpg.de/en/research/research-group-neurocognition-of-music-and-language.html