Laboratoire des
Systèmes
Perceptifs

CHAIRE
BEAUTÉ·S

Why do we Like (the) Music (we Like)? From Computational Neuroscience to Music Perception

Author : Guilhem Marion
Supervisor : Shihab Shamma
Jury : Barbara Tillmann (president)
Mounya Elhilali (reviewer)
Elvira Brattico (reviewer).

December 4, 2023

ACKNOWLEDGEMENTS

I would not have been able to produce this work alone without the help of many people, in many different ways. I would like first to thank Claire Pelofi who significantly shaped the direction of this thesis. She has been a close collaborator since the day we met and has been a constant ear for discussing all sorts of ideas. I would also like to thank Giovanni Di Liberto who co-supervised my internship in the lab four years ago and has been a close collaborator who taught me much about TRFs and EEG experiments! Many other colleagues contributed in myriad ways to make this thesis possible and will be specifically credited all along this document. They include Yves Boubenec, Rupesh Kumar Chillale, Flavien Feral, Jeffrey Boucher, Pierre Orhan, Yashish Maduwantha, Cong Han, Camille Barbarot, Henri Le Goff, Agathe Mangia, Michael Casey, Sean Paulsen, Giacomo Bignardi, Mergherita Malanchini, Mosing Miriam, and Gisèle Dambuyant.

I would also like to thank those with whom I interacted at the Laboratoire des Systèmes Perceptifs at ENS. They always inspired me with excellent and deep discussions, especially Jackson Graves, Léo Vernet, Alain de Cheveigné, Daniel Presnitzer, and Claudia Lunghi. I do not forget Pascal Mamassian, the Director of the lab, who was always there to support me, and Balkis Cadi, our incredible lab Administrator who always helped me with a friendly warmth. I really felt at home during my entire stay in this wonderful lab.

My PhD has been funded by the Chaire Beauté(s), a research chair, hosted by the PSL University, which is working on new contemporary definitions of beauty through multiple disciplines such as physics, neuroscience, philosophy, and sociology. I value greatly such cross-disciplinary approaches and was thrilled to play my part in this program. Moreover, this program allowed me to interact with remarkable researchers from all kinds of disciplines who considerably influenced my work and provided me with the necessary resources to build my own project in the best way possible. I would like to thank the entirety of the chair and especially Clélia Zernik, Justin Jaricot, and Anne-Lise Worms who have been so helpful and friendly to me in all their interactions.

This dissertation would also not be able to molt into its final form without the precious help of the reviewers. I would like to deeply thank Mounya Elhilali, Elvira Brattico, and Barbara Tillman for accepting to read this manuscript and evaluate this thesis.

Last but not least, I would like to thank Shihab Shamma, my supervisor who believed in me from the first day, even when my interests were unconventional, gave me all the freedom and means to explore the academic world in the best way possible, and were the best supervisor I could have dreamed of.

Contents

[1] Authors: Guilhem Marion, Giovanni Di Liberto, Shihab Shamma(Marion *et al.*, 2021)

[2] Authors: Giovanni Di Liberto, Guilhem Marion, Shihab Shamma(Di Liberto *et al.*, 2021)

[3]Authors: Shihab Shamma, Prachi Patel, Shoutik Mukherjee, Guilhem Marion, Bahar Khalighinejad, Cong Han, Jose Herrero, Stephan Bickel, Ashesh Mehta, Nima Mesgarani(Shamma *et al.*, 2020)

[4]Authors: Yashish M. Siriwardena, Guilhem Marion, Shihab Shamma(Siriwardena *et al.*, 2022)

[5] Authors: Guilhem Marion, Giovanni Di Liberto, Benjamin Gold, Shihab Shamma

[6] Authors: Guilhem Marion*, Amélie Picard*, Benjamin Gold, Benjamin Skerritt-Davis, Mounya Elhilali, Shihab Shamma

List of Tables

List of Figures

ABSTRACT

This PhD thesis explores musical perception through a multidisciplinary approach in the fields of neuroscience, psychology, and computer science, analyzing the link between music and culture. Experimental paradigms employing computational models of music, neuroimaging in humans and animals, electrophysiological recordings in ferrets, and behavioral analysis of human psychacoustics, have helped uncover the neural bases of many fascinating aspects of the musical experience.

We first explored experimentally the brain's ability to predict musical events while listening to extended melodic sequences. Recordings of these predictions correlated with expectations generated by computational models, highlighting the brain's ability to anticipate music. New innovative models of musical expectation were then formulated as a result; they include IDyOMpy and MusiREX. For example, IDyOMpy was used to demonstrate the strong relationship between brain responses during musical imagery and during natural moments of silence versus model-estimated probabilities, thus emphasizing the predictive nature of neural activity. Sensory-motor interactions were also explored through computational models inspired by the Mirror Network architecture, shedding light on their role in sensory-motor learning.

Musical enculturation in neural models of musical expectation was examined through various techniques, including human electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and behavioral recordings in humans as well as electrocorticography (ECoG), and functional ultrasound imaging (FUS) in a ferret model. We demonstrated that passive exposure to unfamiliar music enhanced the predictive abilities and pleasure, aligning with the so-called Wundt effect, a key element of contemporary literature on musical pleasure. In summary, this thesis presents a framework where musical expectations are learned through passive exposure and subsequently shape music enjoyment and individual preferences. Viewed from this perspective, these global mechanisms can also be interpreted as an evolutionary process for social bonding.

Finally, we expanded the investigation by considering the genetic and sociocultural factors impacting musical preferences. Our aim was to explore hereditary and non-shared environmental influences through a genetic study based on twin siblings. Cross-cultural investigations in Paris and Rome also provided sociocultural insights into musical preferences, contributing to a refined model of musical preferences including, social affiliation, genetics, and statistical passive learning.

In conclusion, this thesis advances our understanding of the neural mechanisms of musical perception, explores the impact of musical enculturation on musical enjoyment, and introduces

genetic and socio-cultural factors to understand their role in shaping musical preferences. It contributes to the neuroscience of music by uncovering the interplay between predictions, culture, and the musical experience.

RÉSUMÉ SUBSTENTIEL

Cette thèse de doctorat explore la perception musicale à travers une approche multidisciplinaire dans les domaines des neurosciences, de la psychologie et de l'informatique, en analysant le lien entre musique et culture. Des paradigmes expérimentaux utilisant des modèles informatiques de la musique, la neuro-imagerie chez l'homme et l'animal, des enregistrements électrophysiologiques chez le furet et l'analyse comportementale de la psychoacoustique humaine, ont permis de découvrir les bases neurales de nombreux aspects de l'expérience musicale.

J'essaie d'abord de démontrer que le cerveau emet des prédictions sur les notes de musique à venir dans le contexte de l'écoute de mélodies à travers deux expériences différentes. La première série d'expériences consiste en des enregistrements de musiciens professionnels écoutant et imaginant des chorals de Bach. Comme première observation de l'existence de prédictions musicales dans le cerveau, je montre que les réponses corticales enregistrées par electrocephalographie (EEG) présentent des amplitudes différentes pour différentes notes et que ces différences peuvent être expliquées par leurs probabilités calculées par un modèle statistique de la musique. Dans la seconde étude, j'ai démontré avec des collègues que les moments de silence naturel pendant ces chorales contiennent des réponses cérébrales qui ont la signature typique des signaux de prédiction (polarité négative) et que ces réponses ont également leur amplitude corrélée avec la probabilité de leurs attentes estimées.

Deux nouvelles versions de modèles informatiques d'attente musicale sont ensuite présentées. Elles sont basées sur les travaux de Marcus Pearce (suivant le cadre d'IDyOM(M. T. Pearce, 2005)), le modèle statistique de la musique le plus utilisé (plus de 300 études le citent), mais présentent des améliorations de mise en œuvre et de nouvelles fonctionnalités qui ont déjà été utilisées dans mes études. IDyOMpy est une formulation Python d'IDyOM qui, grâce à sa structure de code modulaire, permet des modifications faciles et l'ajout de nouvelles fonctionnalités. Par exemple, il a été utilisé pour calculer les probabilités d'entendre une note dans chaque moment de silence au sein de morceaux de musique naturels. Cette implémentation présente également une nouvelle façon de fusionner les statistiques recueillies à différentes échelles temporelles et permet de meilleures performances basées sur différentes mesures. MusiREX est une implémentation du modèle D-REX(Skerritt-Davis & Elhilali, 2018; 2019), formulé à l'origine dans le laboratoire LCAP de l'Université John Hopkins. Cette nouvelle version suit désormais la structure du modèle IDyOM (long et court terme, validation croisée/train-test, dépendances temporelles fixes) et travaille directement à partir de fichiers midi. Elle permet d'obtenir de meilleures performances que les deux précédentes implémentations d'IDyOM sur

différentes mesures et permet également d'utiliser des enregistrements audio au lieu de fichiers midi uniquement symboliques. Ces deux nouveaux modèles sont très efficaces pour rendre compte des réponses du cerveau humain. Par exemple, nous avons utilisé IDyOMpy pour montrer, pour la première fois, que les réponses corticales aux notes imaginées étaient corrélées avec la probabilité estimée par le modèle et nous avons utilisé ces corrélations pour construire un classificateur capable de détecter quelle chorale avait été imaginée par les participants. En outre, nous avons montré que la signature topographique de ces réponses présentait une polarité *inverse* par rapport aux réponses pendant l'écoute musicale de la même musique, tout comme pendant les moments de silence. Ces deux nouveaux résultats démontrent la nature prédictive de l'imagerie musicale dans le cerveau et offrent de nouvelles hypothèses sur la façon dont nous mémorisons le contenu musical. Une explication de ce phénomène est fournie par le cadre de traitement prédictif (Predictive Processing Framework) (Clark, 2013; K. J. Friston *et al.*, 2010). Ce cadre s'articule autour de l'idée que le cerveau développe un modèle du monde qui est utilisé pour prédire les entrées sensorielles et qui est continuellement mis à jour en comparant les stimuli prédits et réels (Barlow *et al.*, 1961). Ainsi, la perception émerge de l'interaction entre les entrées sensorielles (S) et les attentes ou prédictions internes (P). La comparaison entre les entrées sensorielles et leur prédiction produit une erreur de prédiction (PE, δ = S-P) qui, entre autres fonctions, permet la mise à jour du modèle de prédiction interne lui-même (Näätänen *et al.*, 2007). Par conséquent, la perception est un processus actif par lequel notre cerveau surveille en permanence les statistiques des informations sensorielles entrantes afin (i) d'apprendre et de mettre à jour un modèle interne des régularités du monde qui nous entoure ; et (ii) de prédire, sur la base de ce modèle, les entrées sensorielles entrantes afin de moduler leur encodage neuronal et de faciliter leur perception dans des conditions difficiles, par exemple lors de la restauration d'objets manquants ou bruyants.par exemple lors de la restauration de parties manquantes ou bruyantes d'un stimulus (Leonard *et al.*, 2016) ou en biaisant la perception d'images ou de sons ambigus (Brainard & Hurlbert, 2015; Pressnitzer *et al.*, 2018).

La perception musicale, en particulier, offre un paradigme éclairant pour explorer les mises en œuvre des principes de traitement prédictif en raison de ses régularités structurelles, temporelles, timbrales, mélodiques ou harmoniques (Koelsch *et al.*, 2019; M. A. Rohrmeier & Koelsch, 2012). Ainsi, contrairement au traitement des entrées sensorielles aléatoires et imprévisibles, le signal musical hautement structuré produit des prédictions concurrentes sur les événements à venir. Plus précisément, la régularité temporelle de la musique qui est présente à différentes échelles de temps conduit à des modèles récurrents de mélodie. Ainsi, la musique peut présenter des régularités dues à la répétition du même motif ainsi que des régularités qui peuvent être imprévisibles sur la base du seul contexte proximal, mais qui sont néanmoins conformes aux règles d'un style musical ou d'une culture particulière (Margulis, 2014).

Des preuves comportementales d'un traitement prédictif pendant l'écoute de la musique ont déjà été observées dans le passé en réponse à des stimuli artificiels contenant des événements plus ou moins attendus (par exemple, un accord de dominante se résolvant soit sur une tonique, soit sur une sixte napolitaine). Dans la conception expérimentale de l'amorçage, lorsqu'on demande aux auditeurs de détecter des écarts de timbre, des temps de réponse (TR) plus rapides ont été associés à des stimuli musicaux plus attendus (J. J. Bharucha & Stoeckig, 1987; Bigand & Pineau, 1997; Tillmann *et al.*, 2006; 2007). Il a également été démontré que la précision des performances s'améliorait avec la prévisibilité des notes (J. J. Bharucha & Stoeckig, 1987). Il est important de noter que les auditeurs sans formation musicale formelle ont montré des effets d'amorçage similaires à ceux des musiciens formés (J. J. Bharucha & Stoeckig, 1986; Bigand & Pineau, 1997; Tillmann *et al.*, 2006), ce qui confirme l'hypothèse selon laquelle le traitement prédictif pendant l'écoute de la musique ne nécessite pas de formation musicale formelle (Tillmann, Bharucha, & Bigand, 2000).

Néanmoins, les prédictions neurales déclenchées par la mémoire ne sont pas la seule forme de prédiction que l'on trouve dans la littérature. L'autre forme prédominante de prédiction est en effet considérée comme étant déclenchée par le système moteur. Une abondante littérature sur la parole montre que la parole cachée (ou l'imagerie mentale) affecte le cortex auditif (Y. Ding *et al.*, 2019; Tian & Poeppel, 2010; 2012; 2013; Whitford *et al.*, 2017), notamment sous la forme d'une copie d'efférence(Tian & Poeppel, 2010; 2012; 2013) utilisée pour calculer une erreur de prédiction(Ventura *et al.*, 2009) permettant un retour d'information auditif. Cette idée est également présente en dehors de la communauté de la parole, par exemple, les sons sémantiques liés à une action motrice suscitent une activation dans les aires motrices somatotopiques, tandis que les sons non sémantiques (tons purs) suscitent une activité uniquement dans les aires temporelles(Grisoni *et al.*, 2019). Mais, plus intéressant encore, cette idée est également présente dans les neurosciences de la musique, une étude ECoG a montré que la lecture silencieuse d'un piano électronique (son coupé) suscitait des activations auditives très similaires à celles induites par la lecture réelle des mêmes morceaux, démontrant que les mouvements moteurs peuvent moduler l'activité auditive(Martin *et al.*, 2017). Dans le sens inverse, il a été démontré que l'audition notationnelle (Brodsky *et al.*, 2008) (imagerie musicale entraînée par la lecture de partitions musicales) et l'écoute (Pruitt *et al.*, 2018) génèrent une excitation dissimulée des plis vocaux avec une signature neuronale similaire à celle observée pendant l'imagerie musicale (Zatorre *et al.*, 1996), démontrant la modulation de l'activité motrice par l'activité auditive. Enfin, une autre étude a demandé à des pianistes et clarinettistes professionnels de regarder des vidéos de musiciens professionnels jouant des morceaux connus sur leur instrument (piano ou clarinette). Certaines notes étaient décalées par rapport à la vidéo. Les notes mal assorties ont déclenché des ERP différents de ceux des autres notes, ce qui montre un réseau de prédiction clair entre les cortex moteur, visuel et auditif. (Mado Proverbio *et al.*,

2014). Dans ce chapitre, nous présenterons deux études portant sur de nouveaux modèles de calcul des prédictions neuronales sensori-motrices. La première étude s'inspirera des prédictions bidirectionnelles caractérisées entre les aires motrices et auditives pendant la parole et discutera de la raison évolutive d'un tel chemin direct entre les aires auditives et motrices en tant que chemin nécessaire à l'apprentissage de la production de sons à travers la présentation d'un nouveau modèle informatique pour l'apprentissage des interactions sensori-motrices. La seconde étude présentera un modèle informatique similaire pour l'apprentissage de la production sensori-motrice dans le cas de la musique.

Pour revenir à la théorie du codage prédictif, elle postule, qu'au delà d'emettre des predictions, que le cerveau apprend à predire et ceux à partir des diverses entrées sensorielles qui decrivent le monde extérieur. Sur la base de cette idée, je propose un cadre neurobiologique visant à élucider les différences culturelles dans les modèles neuronaux d'attente musicale. Nous avons conçu un ensemble d'expériences utilisant diverses techniques d'enregistrement, notamment l'électrophysiologie humaine (EEG), l'imagerie (IRMf) et les enregistrements comportementaux, ainsi que l'électrophysiologie corticale invasive (ECoG) et l'imagerie par ultrasons (FUS) dans le modèle animal du furet. Ces expériences s'articulent autour du concept d'apprentissage implicite des structures musicales (E. E. Hannon & Trehub, 2005b; Loui *et al.*, 2010). Comme ce mécanisme devrait influencer la manière dont les auditeurs prédisent et donc perçoivent la musique, nous l'appelons *Enculturation,* comme indiqué dans la littérature sur la cognition musicale comportementale(Demorest *et al.*, 2008; E. Hannon & Trainor, 2007; Haumann *et al.*, 2018; Morrison *et al.*, 2008; M. T. Pearce, 2018; van der Weij *et al.*, 2017; Wong *et al.*, 2009). Dans ces expériences, nous avons demandé à des participants occidentaux d'écouter de la musique traditionnelle chinoise non familière de la région de Shanxi (pendant que leur activité cérébrale était surveillée par EEG), et de signaler le plaisir qu'ils avaient ressenti pendant l'exposition. Ce test s'est déroulé en trois temps : avant et après une phase d'exposition à domicile, ainsi que deux mois après la phase d'exposition. Cette phase d'exposition à domicile portait soit sur de la musique chinoise non familière (groupe test), soit sur de la musique occidentale familière (groupe témoin). L'analyse a montré que les potentiels de réponse évoqués par la note présentaient des amplitudes réduites chez les participants exposés à la musique chinoise par rapport à ceux exposés à la musique occidentale. Ce modèle de résultats s'aligne parfaitement sur un modèle de corrélation, dans lequel l'IDyOMpy formé à la chanson chinoise exposée présente des corrélations accrues après l'exposition pour le groupe test, mais pas pour le groupe témoin. Il est important de noter que ces résultats sont validés par des enregistrements électrophysiologiques dans le cortex auditif d'un modèle animal de furet. Les deuxièmes expériences physiologiques ont utilisé des techniques d'imagerie par ultrasons pour obtenir des résultats préliminaires prometteurs dans le cerveau des furets. Bien que certaines de ces données soient encore préliminaires, elles ont stimulé les discussions sur

le rôle évolutif de la musique. En outre, les données comportementales humaines ont montré une augmentation du plaisir auto-déclaré dans le groupe testé, mais pas dans le groupe témoin, ce qui est cohérent avec la littérature existante sur la relation entre la prévisibilité et le plaisir musical(Droe, 2006; Martindale & Moore, 1989; Martindale *et al.*, 1990; Soley & Hannon, 2010). Ceci est également lié à des études récentes sur l'effet Wundt(Berlyne, 1971; Chmiel & Schubert, 2017) qui met en évidence une relation non linéaire en U inversé entre le plaisir musical et l'attente, suggérant qu'un niveau optimal de surprise génère un plaisir maximal(Cheung *et al.*, 2019; Gold, Pearce, *et al.*, 2019). La familiarisation passive avec une musique inconnue améliore donc les capacités de prédiction et permet de se rapprocher du plaisir optimal.

Tout le cadre analytique susmentionné de ma thèse présuppose que l'appréciation de la musique est principalement façonnée par des influences culturelles. Pour évaluer cette hypothèse de manière critique, nous avons mis en place des collaborations solides visant à mener une étude génétique avec des frères et sœurs jumeaux. Cette étude vise à élucider les composantes héréditaires des préférences musicales et à les juxtaposer à l'impact des facteurs provenant des expériences individuelles, souvent appelés "environnement non partagé". En outre, nous voulions voir dans quelle mesure ces facteurs pouvaient être expliqués par le statut socio-économique de nos participants, ajoutant ainsi une couche nuancée à notre exploration. En outre, nous nous sommes lancés dans une exploration des origines socioculturelles à multiples facettes des préférences musicales. Une enquête interculturelle a été menée à Paris et à Rome, impliquant des participants dans des expériences cognitives suivies d'entretiens sociologiques approfondis. Ces entretiens ont permis de recueillir un certain nombre de paramètres socioculturels. En nous appuyant sur ce riche ensemble de données, nous nous efforçons d'acquérir une compréhension globale des éléments constitutifs qui influencent les préférences musicales et leurs origines. Ce faisant, nous visons à affiner et à développer nos théories existantes concernant les fondements de l'appréciation de la musique.

Publication List

Already Published Papers

The music of Silence. Part I: Responses to Musical Imagery Encode Melodic Expectations and Acoustics, **Guilhem Marion**, Giovanni Di Liberto, Shihab Shamma, 2021, JNeurosci(Marion *et al.*, 2021).

The music of silence. Part II: Cortical Predictions during Silent Musical Intervals, Giovanni Di Liberto, **Guilhem Marion**, Shihab Shamma, 2021, JNeurosci(Di Liberto *et al.*, 2021).

Accurate Decoding of Imagined and Heard Melodies, Giovanni Di Liberto, **Guilhem Marion**, Shihab Shamma, 2021, Frontiers in Neuroscience(Liberto *et al.*, 2021).

The Mirrornet: Learning Audio Synthesizer Controls Inspired by Sensorimotor Interaction, Yashish M. Siriwardena, **Guilhem Marion**, Shihab Shamma, 2021(Siriwardena *et al.*, 2022).

Learning Speech Production and Perception through Sensorimotor Interactions, Shihab Shamma, Prachi Patel, Shoutik Mukherjee, **Guilhem Marion**, Bahar Khalighinejad, Cong Han, Jose Herrero, Stephan Bickel, Ashesh Mehta, Nima Mesgarani, 2020, Cerebral Cortex(Shamma *et al.*, 2020).

Under Review

Cross-cultural perspectives on the predictive coding perspective of music perception, Claire Pelofi*, **Guilhem Marion**, Giovanni Di Liberto, Pablo Ripollès, Shihab Shamma, under review (TICS)(Pelofi *et al.*, n.d.).

To be Submitted

IDyOMpy: a New Python Implementation for IDyOM, a Statistical Model of MusicalExpectations, **Guilhem Marion**, Giovanni Di Liberto, Benjamin Gold, Shihab Shamma, in prep.

Musi-Rex: a New Implementation of the D-Rex Model for Music Purposes, **Guilhem Marion***, Amélie Picard*, Benjamin Gold, Benjamin Skerritt-Davis, Mounya Elhilali, Shihab Shamma, in prep.

The precise division of tasks between authors and scientific contribution for each study is described in the *scientific contribution* section of each chapter.

1 INTRODUCTION

It always struck me to see that different people could like vastly different songs, and even perceive them in drastically different ways, to the point that two friends could use opposite words to describe the same piece of music. I therefore always wanted to study music perception but did not know where to start. As an undergraduate student in Biology and Computer Science, I designed formal and statistical models of music that were mainly used to generate new music *in the style of*. But later, my master's degree in musicology allowed me to understand better how musical expressions work and evolved across history, and also how they could differ between cultures. My master's thesis therefore evolved to reflect these interests, focusing on how musicology could make use of computational models of music. Soon after, I discovered that the same statistical models I worked with were also predictive of neural activity during music listening. It was a real turning point in my intellectual life. It bridged the gap between all my interests: Music, Computer Science, and Biology. I, therefore, decided to start my PhD around those ideas: how computational models of music could shed light on how the brain generates predictions about upcoming musical notes and how those predictions reflect the *culture* and the unique sensibility of listeners.

Today, I am very proud to present the results of this journey, even if very incomplete. This thesis is, I hope, like my academic education: multi-disciplinary. I consider this work to be centered on music cognition, influenced by and borrowing techniques from neuroscience, computer science, statistics, experimental psychology, sociology, and, more generally biology. For instance, I will present work based on brain recordings through electroencephalography (EEG) which is a non-invasive way of recording brain activity, self-reported musical pleasure, and even invasive recordings of electrical brain activity in ferrets listening to music. Those data will be analyzed using various techniques to pinpoint how the brain reacts to music and measure individual preferences for given songs. At the end of the dissertation, I will open up on two ongoing studies in genetics and the sociology of music to give a finer-grained view of individual differences.

First, I place this work into the framework of the predictive coding theory(Clark, 2013; K. J. Friston *et al.*, 2010) and will review the literature and present evidence that the brain is computing predictions about upcoming musical notes in the context of melody listening through two different experiments. Many studies already investigated musical predictions in the brain and, for instance, showed that harmonic violation generated specific responses(Koelsch, 2009; Koelsch & Mulder, 2002; Koelsch *et al.*, 2000; Leino *et al.*, 2007; Loui *et al.*, 2005; Saarinen *et al.*, 1992; Steinbeis *et al.*, 2006) known as the ERAN (Early Right Anterior Negativity) (Koelsch, 2009). This response has also been shown to continuously correlate with musical expectation as computed by statistical models (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Omigie, Pearce, *et al.*,

2019; Omigie *et al.*, 2013a). In the first set of experiments, I extend this literature with recordings of professional musicians listening to and imagining Bach's chorales. As new evidence for predictions in the brain, I demonstrate, with colleagues, that natural moments of silence during those chorales contain brain responses that have the typical signature of prediction signals (negative polarity) and that those responses also have their amplitude correlated with the probability of their expectations estimated. On the other hand, fMRI studies showed imagery brain responses partially overlap with listening responses (Bastepe-Gray *et al.*, 2020; Bunzeck *et al.*, 2005; Griffiths, 1999; A. R. Halpern, 2001; A. R. Halpern & Zatorre, 1999; A. R. Halpern *et al.*, 2004; Herholz *et al.*, 2012; Hubbard, 2013; Kraemer *et al.*, 2005; Lima *et al.*, 2015; Yoo *et al.*, 2001; Zatorre & Halpern, 2005; Zatorre *et al.*, 1996; Zhang *et al.*, 2017). However, the field was missing clear electrophysiological characterization of those responses. We showed that their dynamics were very related to those of during perception as they were of an almost perfect inverted polarity. We showed that it was possible to use imagery response to reconstruct listening responses, and inversely, very in line with the previous literature.

Neural predictions triggered by memory are not the only form of prediction to be found in the literature. The other predominant form of prediction is indeed thought to be triggered by the motor system. A substantial literature on speech shows that covert speech (or mental imagery) does affect the auditory cortex (Y. Ding *et al.*, 2019; Tian & Poeppel, 2010; 2012; 2013; Whitford *et al.*, 2017), especially in the form of an efference copy(Tian & Poeppel, 2010; 2012; 2013) that is used to compute a prediction error(Ventura *et al.*, 2009). That is why we complement our work with two studies on sensory-motor interactions which include two new computational models based on the *Mirror Network* architecture. This architecture is based on the findings that motor areas send a parallel internal neural copy of the speech signal to the auditory cortex – the *forward* prediction signal,(Hickok & Poeppel, 2007) along with an *inverse* mapping from the auditory to the motor areas during listening(Stephen *et al.*, 2004). This architecture replicates our findings in electrocorticography (ECoG) data and allows for a functional explanation of the role of those efference copies as a necessary mechanism for sensory-motor learning of speech and music production.

Computational models of music are widely used by the community whether in behavior(J. J. Bharucha & Stoeckig, 1986; Bigand & Pineau, 1997; Bigand *et al.*, 2001; Margulis, 2003; Margulis & Levine, 2006; Marmel *et al.*, 2008; 2010; Omigie, Pearce, & Stewart, 2012; Tillmann *et al.*, 2006), electrophysiology (Di Liberto, Pelofi, Bianco, *et al.*, 2020; A. R. Halpern *et al.*, 2017; Marion *et al.*, 2021; Omigie, Pearce, *et al.*, 2019; Omigie *et al.*, 2013a; M. T. Pearce *et al.*, 2010; Quiroga-Martinez, C. Hansen, *et al.*, 2020; Quiroga-Martinez, Hansen, *et al.*, 2020) and even fMRI(Cheung *et al.*, 2019) studies. However, IDyOM, the dominating model in the field, has intrinsic limitations: its modularity and its specificity to symbolic data (musical score, as opposed to audio recordings). Therefore, we present two new versions of computational models

of musical expectation. They are based on the work of Marcus Pearce (following the IDyOM framework(M. T. Pearce, 2005)) but present implementation improvements and new features that have been already used in my studies. IDyOMpy is a Python formulation of IDyOM, which thanks to its modular code structure, allows for easy modification and additions of new features. This implementation also presents a new way of merging the statistics gathered at different temporal scales and allows for better performances based on different measures. MusiREX is a re-implementation of the D-REX model(Skerritt-Davis & Elhilali, 2018; 2019), originally formulated in the LCAP Lab at John Hopkins University. This new version now follows the structure of the IDyOM model (long- and short-term, cross-validation/train-test, fixed temporal dependencies) and works directly from midi files. It allows better performances than the two previous IDyOM implementations on different measures and also allows for using real audio recordings instead of solely symbolic midi files. These two new models are quite effective at accounting for the responses of the human brain.

The predictive coding theory posits that the brain, in addition to sending predictions, also learns to predict based on the statistics of the sensory inputs, and incorporates them in an internal model of the external world even during *passive exposure*(Loui & Wessel, 2008; Loui *et al.*, 2006; 2010; M. Rohrmeier & Cross, 2009; 2013; M. Rohrmeier *et al.*, 2011). Based on this idea, I propose a neurobiological framework aimed at elucidating cultural differences in neural models of musical expectation. We designed an array of experiments utilizing various recording techniques, encompassing human electrophysiology (EEG), imaging (fMRI), and behavioral recordings, as well as invasive cortical electrophysiology (ECoG) and ultrasound imaging (FUS) in the ferret animal model. These experiments revolved around the concept of implicit learning of musical structures (E. E. Hannon & Trehub, 2005b; Loui *et al.*, 2010). Because this mechanism should shape the way listeners predict and therefore perceive music, we call this mechanism *Enculturation,* as referred to in the literature on behavioral music cognition(Demorest *et al.*, 2008; E. Hannon & Trainor, 2007; Haumann *et al.*, 2018; Morrison *et al.*, 2008; M. T. Pearce, 2018; van der Weij *et al.*, 2017; Wong *et al.*, 2009). In my experiments, Western participants were asked the listen to unfamiliar traditional Chinese music from the region of Shanxi (while their brain activity was monitored through EEG), and to report the pleasure they felt during the exposure. This testing occurred in three epochs: before and after an at-home exposure phase as well as 2 months after the exposure phase. This at-home exposure phase was either to Chinese unfamiliar (test group) music or to Western familiar (control group) music.

The analysis shows that note-evoked response potentials exhibited reduced amplitudes in participants exposed to Chinese music compared to those exposed to Western music. This pattern of results aligns seamlessly with a correlation model, wherein IDyOMpy trained on the exposed Chinese song exhibits enhanced correlations after exposure for the test group but not

for the control group. Importantly, these findings are validated by electrophysiological recordings in the auditory cortex of a ferret animal model. In addition, human behavioral data showed increased self-reported pleasure in the test group but not the control group, consistent with existing literature on the relationship between predictability and musical enjoyment(Droe, 2006; Martindale & Moore, 1989; Martindale *et al.*, 1990; Soley & Hannon, 2010). Those results are related to recent studies on the Wundt effect(Berlyne, 1971; Chmiel & Schubert, 2017) which highlights a non-linear inverted-U relationship between musical pleasure and expectation, suggesting that an intermediate optimal level of surprise generates maximal pleasure(Cheung *et al.*, 2019; Gold, Pearce, *et al.*, 2019). The passive familiarization with unfamiliar music, therefore, increases musical pleasure by enhancing predictive abilities and moving the cursor closer to the optimal pleasure.

All the above analytical framework in my thesis presupposes that musical enjoyment is predominantly shaped by cultural influences. Passive exposure to music would induce implicit learning (that we call enculturation) building an internal model of music. This internal model is also used to generate predictions about new incoming music. We call the degree to which a prediction is accurate *prediction error* which is known to relate to musical pleasure and activity in the dopaminergic regions in the form of an inverted U shape. Because the internal model of predictions is built through enculturation throughout our entire life to match the music of our environment, this environment, which is different for each individual, is the key element shaping our perception. Therefore, this work gives a clear link between cognition and sociology, revealing the neural underpinnings of the already studied mechanism of social reproduction.

To critically assess this assumption, we have initiated robust collaborations aimed at conducting a genetic study with Twin siblings. This study seeks to elucidate the hereditary components of musical preferences and juxtaposes them with the impact of factors originating from individual experiences, often termed the "non-shared environment". Additionally, we intended to see to which extent these factors could be explained by the socio-economic status of our participants, adding a nuanced layer to our exploration. Furthermore, we embarked on an exploration of the multifaceted sociocultural origins of musical preferences. A cross-cultural investigation was conducted in Paris and Rome, involving participants in cognitive experiments followed by in-depth sociological interviews. These interviews yielded a number of socio-cultural parameters. Leveraging this rich dataset, we endeavor to comprehensively understand the constituent elements influencing musical preferences and their origins. In doing so, we aim to refine and expand upon our existing theories regarding the underpinnings of musical enjoyment and their relationship with the enculturation mechanism.

Because of the extensive number of fields invoked in this thesis, we decided to create standalone general introductions and discussions for each chapter and avoided overloading this general introduction with too many references from very different fields. The reader can therefore

refer to the specific chapters for a detailed literature review and integration of the work in each field. The present PhD thesis is composed of 4 published papers (chapter 2), 2 ready-to-submit papers (chapter 3), 1 under-the-process of writing paper (chapter 4), and 2 newly started projects (chapter 5). Those details will be clearly stated at the beginning of each chapter and all the collaborators and their roles will be reported.

2 Evidence of Musical Predictions in the Brain

2.1 General Introduction to Musical Predictions in the Brain

While perception clearly involves the processing of sensory information, sensory inputs are not enough to give a full account of perceptual processes. Indeed, sensory stimuli (e.g., sound or image) may be perceived differently by different people (Brainard & Hurlbert, 2015; Pressnitzer *et al.*, 2018) or even by the same person but under different conditions (Chambers *et al.*, 2017; Pelofi *et al.*, 2017; Snyder *et al.*, 2015).

2.1.1 Predictive Coding Theory

An explanation of this phenomenon is provided by the Predictive Processing Framework (Clark, 2013; K. J. Friston *et al.*, 2010). This framework revolves around the idea that the brain develops a model of the world that is used to predict sensory inputs and is continuously updated by comparing predicted and actual stimuli (Barlow *et al.*, 1961). Thus, perception emerges from the interplay of sensory inputs (S) and internal expectations or predictions (P). The comparison between sensory inputs and their prediction produces a prediction error (PE, δ = S-P) which, among other functions, enables the update of the internal prediction model itself (Näätänen *et al.*, 2007). Therefore, perception is an active process through which our brain continuously monitors the statistics of incoming sensory information so as to (i) learn and update an internal model of the regularities in the world around us; and (ii) predict based on this model the incoming sensory input so as to modulate their neural encoding and facilitate their perception under challenging conditions, e.g. when restoring missing or noisy parts of a stimulus (Leonard *et al.*, 2016) or biasing the perception of images or sounds that are ambiguous (Brainard & Hurlbert, 2015; Pressnitzer *et al.*, 2018).

Music perception, in particular, offers an illuminating paradigm to explore implementations of Predictive Processing principles because of its structural, temporal, timbral, melodic, or harmonic regularities (Koelsch *et al.*, 2019; M. A. Rohrmeier & Koelsch, 2012). Thus, in contrast to the processing of random and unpredictable sensory inputs, the highly structured musical signal yields competing predictions about upcoming events. Specifically, the temporal regularity of music that is present at different time scales leads to recurrent patterns of melody. Hence, music can present regularities due to the repetition of the same pattern as well as regularities that may be unpredictable based solely on the proximal context, but nevertheless consistent with rules within a particular musical style or culture (Margulis, 2014).

The types of predictions one can make when listening to music can be further differentiated into the *prediction* itself and the *certainty* of the prediction (Koelsch *et al.*, 2019; M. T. Pearce &

Wiggins, 2006; Sohoglu & Chait, 2016). For instance, upon hearing a tonal chord progression, one can predict the next chord. In this process, a *prediction* about the content is made (what chord exactly is to be played next) but also an estimate of how *certain* this prediction is given the context. Hence, a set of irregular chords may also constitute an unpredictable context that will hinder the listener's ability to make accurate predictions on upcoming events (Bianco *et al.*, 2020; Hansen & Pearce, 2014). The certainty of prediction modulates the gain associated with the response error (K. Friston, 2009; Kanai *et al.*, 2015), even in the context of natural music listening (Hsu *et al.*, 2015) and therefore ultimately plays a role in how unexpected events are utilized to update the model of expectations (Hansen & Pearce, 2014; Koelsch *et al.*, 2019).

2.1.2 Behavioral and Neural Evidence

Behavioral evidence for predictive processing during music listening has been observed in response to artificial stimuli containing more or less expected events (e.g. a Dominant chord resolving either on a Tonic or on a Neapolitan sixth). In priming experimental design, when asking listeners to detect timbre deviants, faster Response Times (RTs) were associated with more expected musical stimuli (J. J. Bharucha & Stoeckig, 1987; Bigand & Pineau, 1997; Tillmann *et al.*, 2006; 2007). Performance accuracy was also shown to improve with note predictability (J. J. Bharucha & Stoeckig, 1987). Importantly, listeners with no formal musical training demonstrated similar priming effects as trained musicians (J. J. Bharucha & Stoeckig, 1986; Bigand & Pineau, 1997; Tillmann *et al.*, 2006), further supporting the hypothesis that predictive processing during music listening does not require formal musical training (Tillmann, Bharucha, & Bigand, 2000).

Numerous neuroimaging and neurophysiological studies have been conducted to highlight neural markers of predictive coding during music listening. Typically, these consisted of recordings of the neural responses in participants listening to musical events whose predictability was well-controlled and modulated. The earliest such measurements utilized irregular chord sequences comprising Neapolitan sixth chords which were harmonically distant from the harmonic context (Koelsch *et al.*, 2000; Leino *et al.*, 2007; Loui *et al.*, 2005). The irregular chords elicited Evoked Response Potentials (ERPs) very similar to the Mismatch Negativity response (MMN) (Saarinen *et al.*, 1992) and was referred to as the ERAN (Early Right Anterior Negativity). This "music-syntactic MMN" has a negative polarity, maximally observed over frontal right sensors, and a peak latency of about 150-180 ms, although longer latencies have also been observed (Koelsch & Mulder, 2002; Steinbeis *et al.*, 2006). As opposed to MMN, ERAN relies on musical syntax stored in long-term memory and acquired through life-long exposure to music, whereas MMN responses are based on regularities that are extracted online from the local auditory environment (Koelsch, 2009).

2.1.3 Motor Predictions

The literature also depicts motor predictions as a predominant form of prediction in the brain. Considerable research focusing on speech indicates a tangible impact of covert speech (or mental imagery) on the auditory cortex (Y. Ding *et al.*, 2019; Tian & Poeppel, 2010; 2012; 2013; Whitford *et al.*, 2017), especially in the form of an efference copy(Tian & Poeppel, 2010; 2012; 2013) making possible to computation of a prediction error(Ventura *et al.*, 2009) used for auditory-sensorimotor feedback. This idea is also present outside of the speech community, for instance, semantic sounds linked to a motor action elicit activation in somatotopic motor areas whereas non-semantic sounds (pure tones) elicit activity solely in the temporal areas(Grisoni *et al.*, 2019). But, more interestingly, this idea is also present in the neuroscience of music, one ECoG study showed that silent playing of an e-piano (sound turned off) elicited auditory activations very similar to those induced by the actual playing of the same pieces demonstrating that motor movements can modulate auditory activity(Martin *et al.*, 2017). In the reverse direction, notational audiation (Brodsky *et al.*, 2008) (musical imagery driven by reading music scores) and listening (Pruitt *et al.*, 2018) have been shown to generate covert excitation of the vocal folds with a neural signature similar to that observed during musical imagery (Zatorre *et al.*, 1996), demonstrating the modulation of the motor activity by the auditory activity. Finally, another study asked professional pianists and clarinetists to watch videos of professional musicians playing known pieces on their instruments (piano or clarinet). Some notes were mismatched with respect to the video. The mismatched notes elicited different ERPs than the other notes showing a clear prediction network between the motor, visual, and auditory cortex. (Mado Proverbio *et al.*, 2014)

2.1.4 Scientific Contribution

In this chapter, we will present 2 studies investigating silence in music and 2 studies investigating new computational models of sensory-motor neural predictions.

Musical imagery is the voluntary hearing of music internally without the need for physical action or acoustic stimulation. Previous fMRI studies have found shared areas of cortical activation for imagery and listening tasks, but also non-overlapping ones (see (Zatorre & Halpern, 2005) for a review). Still, the nature and functional role of such activation remains uncertain. That is why we decided to conduct a study in order to characterize the neural responses during musical imagery, to compare them to those during listening and to give a specific attention to the expectation mechanism in order to pinpoint to functional role and the nature of imaginary-induced response.

The second hypothesis of the Preditive Coding Theory (c.f. 2.1.1) claims that each neural

response should be composed of PE, δ = S-P with S the sensory signal and P the prediction signal. Following this idea, ubiquitous moments of silence in music should therefore contain a prediction signal consistent with the internal prediction of the listeners (δ = -P, as S = 0). Even if vigorous responses to silences have been observed across modalities when a sensory stimulus was strongly expected, for example corresponding to an omission during the rapid isochronous presentation of tones (Chennu *et al.*, 2016; Joutsiniemi & Hari, 1989; Simson *et al.*, 1976; Yabe *et al.*, 1997), this hypothesis has never been shown for natural silences in ecologically-valid music. We therefore decided to investigate this question in a second study.

From the results of those two studies, we designed a model for predictions in the human brain listening to music that slightly extends the strict frame of the predicting coding theory; this model will be discussed at the end of the chapter. We also designed a similar model for cross-modal predictions between the auditory and motor areas. We present two models which, consistently with the literature, use the idea of efference copy(Tian & Poeppel, 2010; 2012; 2013) for which the purpose of the backward efference copy (from motor to auditory) would be required to back-propagate the production error in order to learn to control the vocal tract. We then apply this same model to a music synthesizer to show its computational efficiency.

Those four sections are direct re-use of already published work. The author list as well as a reference to the published study are included at the beginning of each section. The articles have been slightly modified to fit this document. Especially, we removed the parts of study #3 which I did not directly participate in. I have designed the experimental protocol and collected the data of studies #1 and #2. I conducted the analysis, generated the figures, and wrote the manuscript of study #1. I implemented the statistical model of music required in study #2 to compute probabilities to have notes in moments of silence. Giovanni Di Liberto conducted the analyses, generated the figures, and wrote the manuscript. I implemented the MirrorNet in study #3 from the first implementation from Cong Han, conducted the analyses, and generated the figures related to the MirrorNet, I wrote the text relative to the MirrorNet, and Shihab Shamma wrote the rest of the text of the manuscript. I participated in the design of study #4, especially the choice of the synthesizer and its control through the MirrorNet. Yashish M. Siriwardena implemented this new version using my code from study #3, conducted the analyses, generated the figures, and wrote the manuscript. Shihab Shamma supervised the scientific process and proofread the manuscripts of all those studies.

2.2 The Music of Silence. Part I: Responses to Musical Imagery Encode Melodic Expectations and Acoustics[1]

2.2.1 Introduction

Musical imagery is the voluntary hearing of music internally without the need for physical action or acoustic stimulation. This ability is important in music creation (Godoy & Jorgensen, 2012), from composition and improvisation to mental practice (Bastepe-Gray *et al.*, 2020). One notable example is Robert Schumann's piano method, in which students are asked to reach the point of "hearing music from the page". But, what are the neural underpinnings of such musical imagery?

Previous fMRI studies have found shared areas of cortical activation for imagery and listening tasks, but also non-overlapping ones (see (Zatorre & Halpern, 2005) for a review). The shared activation was measured across several areas of the human cortex (Hubbard, 2013), specifically in the auditory belt areas (A. R. Halpern *et al.*, 2004; Herholz *et al.*, 2012; Kraemer *et al.*, 2005; Zatorre *et al.*, 1996), the association cortex (A. R. Halpern & Zatorre, 1999; Kraemer *et al.*, 2005), the prefrontal cortex (A. R. Halpern & Zatorre, 1999; Herholz *et al.*, 2012; Lima *et al.*, 2015) and Wernicke's area (Zhang *et al.*, 2017). Musical imagery also seems to engage motor areas (e.g. (A. R. Halpern, 2001; A. R. Halpern & Zatorre, 1999; Herholz *et al.*, 2012; Zhang *et al.*, 2017)), showing spatial activation patterns that are correlated with those measured during music production (Meister *et al.*, 2004; Miller *et al.*, 2010). Interestingly, there is only limited evidence for activation during musical imagery in the primary auditory cortex (e.g. (Bastepe-Gray *et al.*, 2020; Bunzeck *et al.*, 2005; Griffiths, 1999; A. R. Halpern *et al.*, 2004; Yoo *et al.*, 2001)), although this region is strongly activated during musical listening.

Although these previous studies provided detailed insights into which areas are active during musical imagery, the nature and functional role of such activation remains uncertain. One reason lies in the difficulty of studying the temporal dynamics of the underlying neural responses and processes with relatively slow fMRI measurements. A recent study using broadly distributed electrocorticography (ECoG) recordings has indicated that music listening and imagery activated shared cortical regions but with a latency of a reversed sequential order between the auditory and motor areas (Y. Ding *et al.*, 2019). Beyond this, a 2001 study using electroencephalography (EEG) showed that mental continuation of melodic fragment generated electrical responses correlated with the N100 topography during music listening and did

[1] Authors: Guilhem Marion, Giovanni Di Liberto, Shihab Shamma(Marion *et al.*, 2021)

not correlate with the topography during silences(Janata, 2001).

Part of the mystery of musical imagination stems from the fact that music is an elaborate symbolic system conveyed via complex acoustic signals, whose appreciation involves several hierarchical levels of processing. The foundations of such hierarchy depend on the processing of fundamental perceptual attributes, such as pitch, loudness, timbre, and space, which are extracted and represented at or before the primary auditory cortex (Janata, 2015; Koelsch & Siebel, 2005). Higher-order rules of grammar and engagement are then presumably implemented in secondary auditory areas and other associative regions (Cheung *et al.*, 2019; Di Liberto, Pelofi, Bianco, *et al.*, 2020; Zatorre & Salimpoor, 2013). These musical rules are related to how listeners interact and anticipate musical streams, in what is usually referred to as *melodic expectations*. Experimentally, such expectations are assumed to play a critical role in musical listening in relation to auditory memory (K. Agres *et al.*, 2018) and musical pleasure (Gold, Pearce, *et al.*, 2019; Zatorre & Salimpoor, 2013), and to interact with the reward system (Blood & Zatorre, 2001; Cheung *et al.*, 2019; Salimpoor *et al.*, 2011). However, it is unknown if these melodic expectations play any role during musical imagery, where they could be related to the ability to recall, create, and become emotionally engaged with the music generated within our own minds.

Melodic expectations can be quantified using statistical models trained on a musical corpus that summarizes the musical material listeners have been exposed to (Abdallah & Plumbley, 2009; Gillick *et al.*, 2010; M. T. Pearce, 2005; M. Rohrmeier, 2011), thus capturing listeners' perceptual judgments, musical reactions and expectations (C. Krumhansl *et al.*, 1999; 2000; M. T. Pearce, 2018). In our experiments, the musical corpus was a large repertoire of Western music that our participants were familiar with. Using these models of melodic structure, our experimental results suggest that imagery of naturalistic melodies (Bach chorals) elicits cortical responses to the imagined notes, exhibiting temporal dynamics and expectation modulations that are comparable to the neural responses recorded during music listening. We also find that the neural signal recorded in the imagery condition could be used to robustly identify the imagined melody with a single-trial classifier. A companion study (Di Liberto *et al.*, 2021) expands on these results to demonstrate that the ubiquitous short pauses and silent intervals in ongoing music elicit responses and melodic expectations remarkably similar to those seen during imagery. Furthermore, with the absence of simultaneous stimulus-driven (bottom-up) responses during silence, these two studies are able to attain direct evidence of the top-down predictive signals and processes critically involved in building musical expectations and culture.

2.2.2 Material and Methods

Participants and Data Acquisition

Twenty-one professional musicians or in training to become professional musicians (6 female; age: M=25, SD=5) participated in the EEG experiment. The sample size was consistent with a previous related study from our team (Di Liberto, Pelofi, Bianco, *et al.*, 2020). Each participant reported no history of hearing impairment or neurological disorder, provided written informed consent, and was paid for their participation. The study was undertaken in accordance with the Declaration of Helsinki and was approved by the CERES committee of Paris Descartes University (CERES 2013–11). The experiment was carried out in a single session for each participant. EEG data were recorded from 64 electrode positions, digitized at 2048 Hz using a BioSemi Active Two system as well as 3 extra electrodes placed on participants skin to record the activity of muscles of potential co-found (tongue, masseter, forearm fingers extensor). Audio stimuli were presented at a sampling rate of 44,100 Hz using a Genelec 8010 10w speaker and Python code for the presentation. Testing was carried out at École Normale Supérieure, in a dark soundproof room. Participants were asked to read the music scores fixed at the center of the desk during both imagery and listening conditions, however, they were instructed to minimize motor activities during the whole experiment. A SM58 microphone was placed in the booth in order to record participant sounds and make sure that they were not singing, taping, nor producing sounds during the experiment. The experimenter listened to those sounds online. Before the experiment, all participants took The Advanced Measures of Music Audiation (AMMA) online using the official website *giamusicassessment.com*.

A tactile metronome (Peterson Body Beat Vibe Clip) playing 100 bpm bars (each 2.4 s) was placed on the left ankle of the participants to provide them with a sensory cue to synchronize their imagination. The start of each trial (listening and imagery) was signaled by a short vibration on the vibro-tactile metronome device followed by a four-beat countdown. Notes closer than 500 ms from a metronome vibration were excluded in order to avoid potential contamination from the tactile stimulus. Experimental assessment showed that the metronome precision was within 5 ms, thus it did not impact our experiment. A constant lag was determined experimentally during the pilot experiments to compensate for perceptual auditory-tactile delays; the latency of 35 ms was determined and applied on all participants.

EEG Experimental Protocol

All participants were chosen to be very well-trained musicians and were all professionals or students at Conservatoire National Supérieur de Musique (CNSM) in Paris. They were given the musical score of the four stimuli in a one-page score and could practice on the piano for

about 35 minutes. The experimenter checked their practice and verified that there were no mistakes in the execution. After practice, participants were asked to sing the four pieces in the booth with the tactile metronome, the sound was recorded in order to check their accuracy offline.

The experiment consisted of a single session with 88 trials. For each condition (listening and imagery) each of the four melodies was repeated 11 times. Trials order was shuffled both in terms of musical pieces and conditions. In the listening condition, participants were asked to passively listen to the stimuli while reading the musical score. For the imagery condition, they were asked to imagine the melody in sync with the tactile metronome as precisely as they could. At the end of every four trials, a break was possible; participants were able to wait as long as they wanted before they continued with the experiment. A sheet of paper was available in the experimental booth, where participants were instructed to report trials where their imagination did not end with the metronome vibration, and therefore were performing the imagery task with incorrect synchronization. No participants reported any mistakes in that sense.

Synchronizing participants' imagination with stimuli is a challenging problem. Previous studies used the so-called *filling in* paradigm where participants are asked to fill an artificial blank introduced in the musical pieces using imagery (Cervantes Constantino & Simon, 2017; Y. Ding *et al.*, 2019; Kraemer *et al.*, 2005), which was not optimal for our experiment as it does not allow for imagery of long stimuli. Other studies displayed visual cues in karaoke-like fashion (Herholz *et al.*, 2012) or used dynamic pianoroll visuals of the stimuli (Zhang *et al.*, 2017). However, several studies have shown that, given the task of synchronizing movements with a discretely timed metronome (e.g., tapping a finger), humans have a striking advantage with auditory metronomes over visual ones (Jäncke *et al.*, 2000; Repp, 2005; Repp & Penel, 2004). In addition, a recent study showed that such an advantage is conserved with tactile metronomes (Ammirante *et al.*, 2016). We assumed that a tactile metronome was less likely to contaminate imagery responses than an auditory metronome because of the different sensory modalities. Therefore, we decided to use a tactile metronome even if some studies suggest that it can induce auditory responses (Ammirante *et al.*, 2016).

Stimuli

Four melodies from the corpus of Bach chorals were selected for this study (BWV 349, BWV 291, BWV354, BWV 271). All chorals use similar compositional principles: the composer takes a well-known melody from a Lutheran hymn (cantus firmus) and harmonizes three lower parts (alto, tenor and bass) accompanying the initial melody on soprano, these cantus firmi were usually written during the Renaissance era. Our analysis only uses monophonic melodies,

we therefore only use these cantus firmi as stimuli for our experiment, original keys were kept. The chosen melodies follow the same grammatical structures and show very similar melodic and rhythmic patterns. Participants were asked to listen to and imagine these stimuli at 100 bpm (about 30 seconds each). The audio versions were synthesized using a Fender Rhodes simulation software (Neo-Soul Keys). The onset-times and pitch values of the notes were extracted from the midi files that were precisely aligned with the audio versions presented during the experiment (see Figure 2.1).

Tools

IDyOM Information Dynamics Of Music (IDyOM) is a statistical model of musical expectation based on variable-order Markov chains (M. T. Pearce, 2005). This model allows for the quantitative estimation of the expectedness of a musical note, which have been shown to be physiologically valid by number of studies(K. Agres *et al.*, 2018; Di Liberto, Pelofi, Bianco, *et al.*, 2020; Egermann *et al.*, 2013; Omigie, Pearce, & Stewart, 2012; Omigie, Pearce, *et al.*, 2019; Song *et al.*, 2016). First, the model has been shown to correctly identify melodic expectation patterns in a consistent way with a musicological analysis (Meyer, 1973) of Schubert's *Octet for Strings and Winds* made by Leonard Meyer in 1973 (M. T. Pearce, 2018). The model also showed correlated expectation values with ones estimated from a behavioral experiment (Manzara *et al.*, 1992). IDyOM was able to account for approximately 63% of the variance in the mean uncertainty estimates reported by the original authors (M. T. Pearce, 2005). Finally, a recent study (Di Liberto, Pelofi, Bianco, *et al.*, 2020) showed that amplitude modulations in EEG and ECoG responses to monophonic music are correlated with the expectation values computed with IDyOM.

The IDyOM model is composed of two modules: a long-term model (LTM) that is pre-trained on a musical corpus (which did not include the stimuli presented in this experiment) in order to capture style-specific global patterns, and a short-term model (STM) that is trained on the preceding proximal context in the current piece to estimate expectedness based on local melodic sequences. Both modules use the same underlying method: Markov chains of different orders (n-grams as states) that can describe melodic patterns at various time scales. All the Markov chains are then aggregated into one model by merging all the probability distributions (M. T. Pearce, 2005). In our analysis, we use the IDyOMpy[2] model, which is an implementation of IDyOM where the Markov chains are combined through a weighting based on the entropy of the distributions from each order. The model was trained using note duration as well as note pitch. The joint distribution was then used to compute the unexpectedness (surprise) of events, which was quantified by means of the Information Content value (IC):

[2]https://github.com/GuiMarion/IDyOM

$$IC(x) = -log(P(X_t = x))$$

mTRF We used the mTRF toolbox[3] (Crosse, Di Liberto, & Lalor, 2016) to estimate the Temporal Response Functions (TRFs) describing the linear mapping of melodic features (onsets, expectation) into the EEG signal. This mapping was estimated for individual electrodes and was based on a convolutional kernel w including various time latencies between the music and the EEG signal:

$$\forall t, r(t,k) = (s * w_k)(t) + \varepsilon(t,k)$$

with t the time indices and k the electrodes and ε the residual response (unexplained noise).

The optimization problem is to find the vector w that minimizes this residual response ε using Ordinary Least Squares method over the vector w while considering a certain degree of regularization to prevent over-fitting by assuming a level of temporal smoothness (Ridge regularization). The optimal regularization parameter was identified at the individual subject level with an exhaustive search within the interval $[10^{-6}, 10]$ with a logarithmic step. The time-lag window [-300,900] ms was used to fit the TRF models. The main figures report weights for the reduced window [-100, 500], where the responses and effects of interest were hypothesized to emerge. This framework has been shown to be effective in assessing the EEG encoding of both low-level auditory features and higher-order auditory expectations (Broderick *et al.*, 2018; Daube *et al.*, 2019; Di Liberto *et al.*, 2015; Lalor & Foxe, 2010; O'Sullivan *et al.*, 2014).

Data Preprocessing

EEG data were analyzed offline using Matlab software. Signals were digitally filtered using Butterworth zero-phase filters (low- and high-pass filters of both order three and implemented with the function filtfilt) and down-sampled to 64 Hz. The main analysis was conducted on data filtered between 0.1 and 30 Hz. Results were also reproduced with the high-pass cut-off frequencies 0.01 and 1 Hz (Figure 2.5). Data were then re-referenced to the average of all 64 channels. EEG channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation.

Data Analysis

Previous studies showed that EEG responses to continuous melodies encode both the acoustic envelope(Di Liberto, Pelofi, Shamma, & de Cheveigné, 2020) and melodic expectations (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Omigie *et al.*, 2013a). The main aim of our study was to

[3]Downloadable at: https://github.com/mickcrosse/mTRF-Toolbox

investigate whether that encoding is conserved during musical imagery. To this end, we assessed the encoding of these features in the EEG signals by means of TRF forward modeling predictions.

The EEG signal was grouped in 88 trials (44 per condition). Each trial was associated with stimulus vectors representing acoustic onsets and melodic expectation:

Onsets vector: One-dimensional vector where the note onsets were marked by an impulse with value 1. All other time-point were assigned to zero;

Expectation vector: One-dimensional vector where the note onsets were marked by an impulse with value corresponding to the expectation value assigned to that note by IDyOM.

Onsets and Expectation Analyses Forward TRFs were fit and used to predict independently each channel of the EEG signal from the onsets and the expectation signal using leave-one-trial-out cross-validation. The correlation between the EEG signals and its prediction were computed for each channel separately resulting in scalp topographies used to assess the spatial activation. This signal (correlation of the feature of the signal of interest with each electrode) accounts for where the signal is computed and not where the amplitude is the strongest, as opposed to ERP topographic maps. Significance of the EEG prediction correlations was assessed by comparing the results with the ones for a null-model where parameters of interest were shuffled in our stimuli:

Onsets analysis: We shuffled the order of the trials, ensuring that the resulting shuffling does not produce matching stimulus-EEG pairs;

Expectation analysis: We shuffled the expectation values while preserving the onset times. This produced vectors with correct onset information but meaningless expectation values.

We ran 20 permutations for each analysis. Those distributions were used to assess significance both at the individual-subject and group levels. The group level significance was computed from the correlation gain distribution with respect to the null model (expectation models - null models or onset models - null models). We subtracted the null-model prediction correlations to the expectation/onsets model prediction correlations by keeping the participants order. Therefore, we got a distribution of 420 values ($21\,participants \cdot 20 shuffling = 420$). A control distribution was constructed by computing the difference between the null-model and other repetitions of itself (here $21\,participants \cdot 20\,shuffling \cdot 19\,different\,shuffling = 7980$). This distribution accounts for the variance of the prediction correlation with a mean of 0. We tested if the correlation gain was above the control distribution using a Wilcoxon sum rank test. Effect sizes were computed using the *common language effect size* between the expectation/onsets distribution gain and the control distribution. The common language effect size

was computed from the U statistic computed by the Wilcoxon sum rank test. The common language effect size is defined as

$$f = \frac{U}{n_1 \cdot n_2}$$

with n_1 and n_2 respectively the sizes of the two distributions (expectation gain and control distribution). As U indicates the number of pairs chosen in the two distributions that satisfy the hypothesis $(\#(i,j)|D1_i > D2_j)$, the common language effect size f therefore indicates the normalized number of pairs that satisfy the hypothesis ($100 \cdot f$ % of the pairs satisfy the hypothesis).

Cross-conditions analysis We assessed the consistency between imagery and listening responses by means of a cross-condition TRF approach. Specifically, TRF models were trained on one condition (e.g., listening) and evaluated on the other (e.g., imagery). The resulting EEG prediction correlations were examined to determine whether the two conditions elicited consistent EEG signals. Furthermore, we investigated whether simple transformations (polarity and latency shift) could explain possible differences between the two conditions. First, we applied a simple polarity inversion by multiplying the TRF kernels by -1. Second, we estimated a linear convolution mapping between the averaged listening responses and the averaged imagery signals (and vice versa) for $n-1$ participants. The learned mapping was then used to transform the listening response into imagery signal (and vice versa) in the left-out participant. The mTRF method was then used to fit subject-specific models on that left-out subject and to predict EEG signals based on the music onsets vectors. The resulting EEG prediction correlations indicate whether the cross-condition mapping is consistent across participants.

Short-term and long-term models An additional analysis was conducted to assess the relative contribution of the short- and long-term modules of IDyOM to the EEG encoding of melodic expectations. To do so, melodic expectation vectors were derived using the short- and long-term models separately. First, short-term model expectations were used to fit TRF models and predict the EEG. Then we used multivariate regression to predict the EEG when considering the two expectation vectors simultaneously (short-term and long-term). In this multivariate case, the null-model was derived by shuffling the values of the long-term expectation vector only. As such, this approach could assess if the long-term model expectations explain EEG variance that is not captured by the short-term expectations.

Decoding the Identity of Imagined Songs Classification was performed to decode the identity of a song from a single EEG trial. We devised a classification method using vote-boosting based on the prediction correlations computed from a forward TRF model trained on the left-out trials. Specifically, prediction correlations were calculated for each of the four pieces using,

separately, the onsets and the expectation vectors. This procedure produced 128 EEG prediction signals ($64 electrodes \cdot 2 features = 128$) for each piece. We then computed the correlation between the target EEG data and each predicted EEG signal *estimators*, leading to 128 correlation values for each of the four pieces. For each estimator, the piece with the highest correlation was chosen, providing one vote for that particular choice. The piece with most votes when considering all estimators was selected as the result of the classification. The methodology is illustrated in Figure 2.1.

2.2.3 Results

We recorded EEG signals (64-channel recording system) from twenty-one professional musicians as they imagined and listened to four monophonic Bach chorals (see Figure 2.1). In both conditions, participants wore a vibrotactile metronome on their left ankle, which allowed for precise synchronization during the imagery task (see **Material and Methods**). We first investigated the responses to the notes by regressing the EEG signals with a stimulus vector representing the note onsets at least 500 ms away from the metronome beats. Then, the melodic expectation for each note was estimated using a statistical model of musical structure (IDyOM) (M. T. Pearce, 2005) trained on a large corpus of Western melodies, supposed to mimic the musical culture of the listeners participating in this study (M. T. Pearce, 2018). We constructed the *expectation signal* as a sparse vector where time onsets of notes were modulated by the expectation value computed by the statistical model of music. As cortical EEG recordings during music listening have already been shown to encode this expectation signal (Di Liberto, Pelofi, Bianco, *et al.*, 2020), our analysis aimed to test the same hypothesis on the imagery condition and to compare the temporal activation between both conditions. The music stimuli, EEG data, and analysis codes are fully available upon request to the corresponding author.

2.2.3.1 Onsets Encoding

Temporal Response Functions (TRFs) describing the linear transformation of note-onsets to an EEG signal (0.1-30 Hz) were estimated for both conditions using lagged linear regression (mTRF-Toolbox (Crosse, Di Liberto, & Lalor, 2016)). EEG prediction correlations were derived on left-out portions of the data with cross-validation. The procedure was then repeated after the labels referring to the stimulus order were randomly shuffled (null-model; EEG_i was regressed with $stim_j$).

Figure 2.2 shows that the note-onset vector could predict the EEG signal better than chance in both conditions, demonstrating the robust encoding of note-onsets in the low-frequency EEG signal (Wilcoxon rank sum test between onsets gain and control distributions; listening:

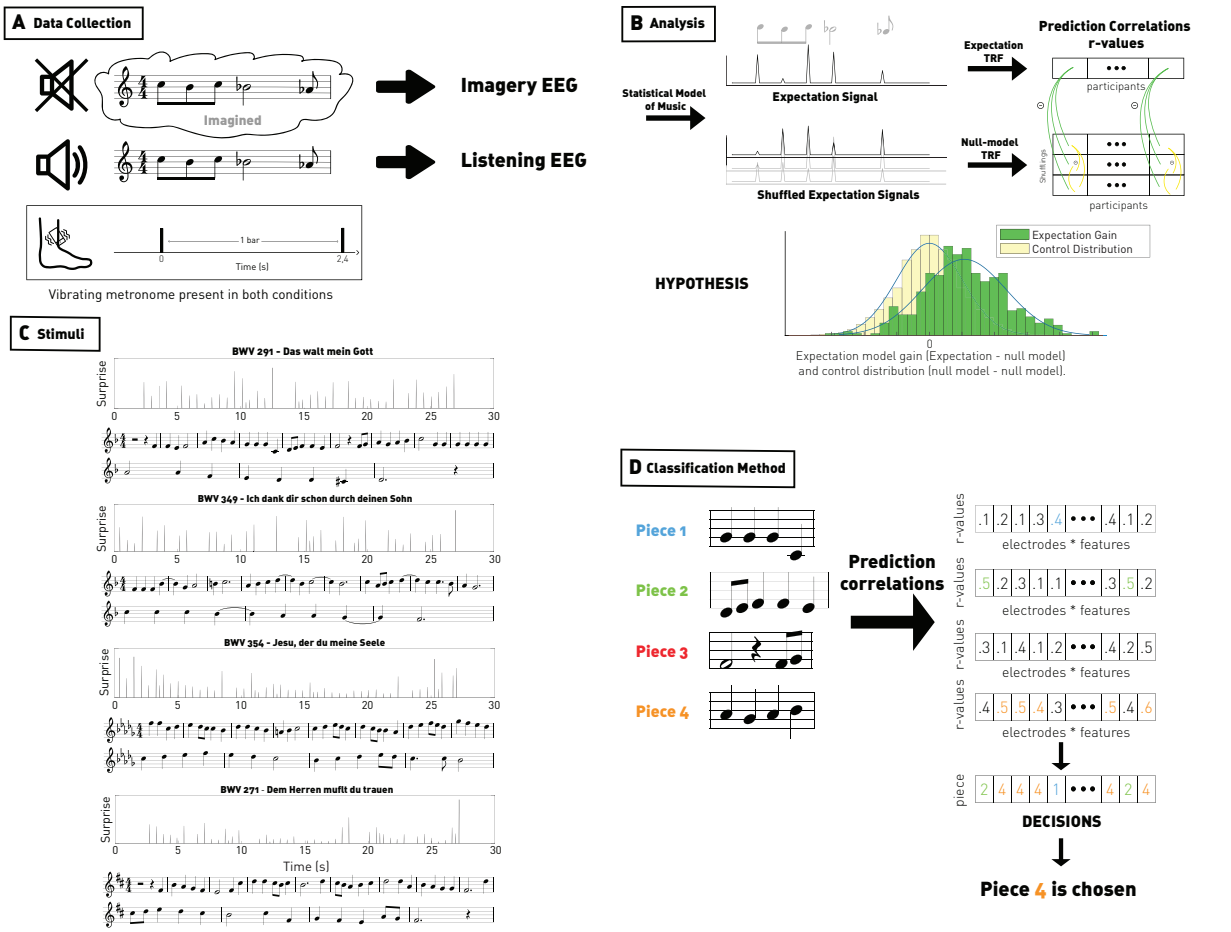


Figure 2.1. **Method Figure** (A) EEG signal was recorded from participants who listened to and imagined four monophonic Bach melodies. The musical bars were indicated using a vibrotactile metronome. (B) *Top-left panels:* Onset vectors amplitude-modulated according to a statistical model of musical expectations. Null-model distributions were derived by shuffling the expectation values while preserving the note onsets. *(Top-right)* Forward TRFs were estimated between the melody vectors and the EEG signal. EEG prediction correlations were derived based on the stimulus vectors and subtracted by the ones for the shuffled vectors, providing (Expectation gain; green), reflecting the EEG encoding of melodic expectations. A control distribution was derived by subtracting EEG prediction correlations between pairs of shuffled vectors (yellow). *Bottom* We hypothesized a positive shift in expectation gain (green distribution) relative to the control distribution (yellow distribution). (C) Stimuli. Musical scores and expectation vectors for each of the four Bach choral stimuli. Melodies were presented at 100 bpm (about 30 seconds each). The expectation signal was computed for each of the melodies using IDyOM. The information content value of each note (the negative log-likelihood) was used to modulate the note-onset values. Forward TRF models were then fit between the resulting vectors and the EEG signal. (D) Classification Method. We trained a TRF model with leave-one-out cross-validation and used this model to predict, from the 4 candidate pieces, the target EEG. We therefore, have *nb_electrodes * nb_features* prediction correlations. For each of these *estimators* we assess which piece maximizes the correlation and the final decision is the piece that occurs the most across electrodes and features.

$p = 8.4 \cdot 10^{-220}$ common language effect size $f = 0.98$; imagery: $p = 2.7 \cdot 10^{-209}$ common language effect size $f = 0.97$, see **Material and Methods**). The note-onset encoding was significant at the individual participant level (17/21, $p < 0.05$, FDR-corrected p-values extracted from the null-models distributions) and was most accurately encoded on central scalp areas, as previously shown in response to auditory experiments (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Di Liberto, Pelofi, Shamma, & de Cheveigné, 2020; Van Canneyt *et al.*, 2020). A significant ($p = 0.02$) correlation of $r = 0.3$ was measured between the topographies of the EEG prediction values for the two conditions (Pearson's correlation).

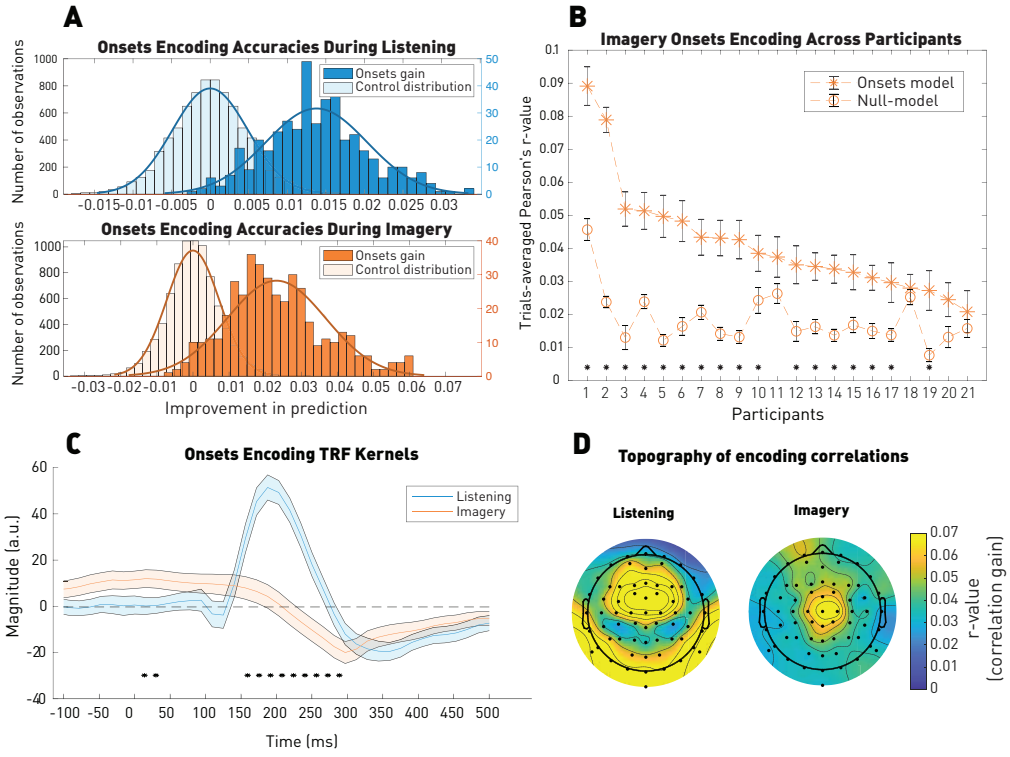


Figure 2.2. **Robust EEG Encoding of Note-Onsets during Imagery.** (A) EEG prediction correlations for the listening (top) and imagery (bottom). EEG prediction correlations were significantly above the control distribution in both conditions. Distributions illustrate the note-onsets correlation gain, adjusted relative to the null-model, as well as the control distribution. As for all the next figures, the left y-axis corresponds to the number of observations of the control distribution, and the right y-axis ones of the model of interest (here onsets gain). (B) EEG prediction correlations for the imagery condition for individual participants. Error bars show the standard error across the 44 trials and stars indicate significance ($p < 0.05$). (C) TRF kernels on Cz. Shaded areas indicate the standard error across participants (N=21) and significance between the two kernels computed by a permutation test ($p < 0.05$) is indicated by black stars. (D) Topography of the EEG predictions gain (onset model - null model). A significant ($p < 0.05$) correlation of $r = 0.3$ was measured between the topographies of the EEG prediction values for the two conditions (Pearson's correlation)

2.2.3.2 Cross-condition Analysis

In line with previous fMRI studies showing partly overlapping neural activation for auditory listening and imagery, we anticipated that a certain degree of similarity exists between the TRFs measured for the two tasks. Indeed, the TRF weights in Figure 2.2C provided us with a qualitative indication of whether the cortical dynamics for listening and imagery are different. Nevertheless, further quantitative assessment was conducted to determine the precise nature of the similarities between the two conditions and the consistency of such similarities across participants. One dominant difference between the two conditions is a time-shifted inverted polarity of the TRF dynamics. This effect of condition was quantitatively assessed by the cross-condition TRF analysis that follows (Figure 2.3).

First, we used the imagery TRF kernels to predict the listening EEG signal, and vice versa, the listening TRF kernels to predict the imagery EEG signal. As expected, these analyses did not produce EEG predictions that were significantly larger than the null-distribution (listening->imagery: $p = 0.83$; imagery->listening: $p = 10^{-19}$, with null-model > onsets-model), confirming that listening and imagery responses are different. Next, we predicted listening EEG responses from the imagery TRF kernels after a polarity inversion, leading to significant EEG predictions ($p = 10^{-46}$; (Figure 2.3), indicating that listening and imagery signals are inversely correlated. However, inverting the listening EEG responses did not lead to an adequate prediction of the EEG in the imagery condition ($p = 0.14$). Such an asymmetry in cross-condition predictions most likely stems from the large difference in the amplitude (and hence the SNR) between the two types of signals. Furthermore, it is also evident from Figure 2.3 that using only a simple polarity inversion is likely to be a sub-optimal description of the mapping between the TRFs in the two conditions. Therefore, we implemented a further refinement in characterizing the relationship between the two TRFs which included a linear mapping with a convolutional kernel as we describe next. In principle, the identification of such a reliable mapping would usher new ways to decode imagined melodies without the need for training imagery EEG data.

A linear mapping with a convolutional kernel was computed between the averaged listening responses and the averaged imagery responses for $n-1$ participants. We then applied the learned cross-condition mapping to estimate the imagery EEG signal of the left-out participant based on their listening responses and the note-onset vectors. This approach led to significant predictions ($p = 10^{-49}$) of the imagery EEG, confirming a reliable relationship between the listening and imagery responses (Figure 2.3). However, the EEG prediction correlations derived with this methodology were not larger than the ones from the cross-participants analysis ($p = 0.12$), where we directly used the averaged TRF kernels from $n-1$ participants on the left-out participant (see Figure 2.9). Using more complex nonlinear transformations between the two TRF kernels may lead to better performances and then allow the computation of the imagery

TRF kernels directly from the listening ones without having to measure imagery responses.

2.2.3.3 Encoding of Melodic Expectations

TRF models were computed to relate melodic expectations to the EEG signal. Expectations vectors were determined by modulating note-onset vectors according to the expectation values derived with the statistical model of melodic structure IDyOM (M. T. Pearce, 2005). Null-models were computed by shuffling the expectation values in the stimulus vectors while preserving the note-onset information. A null-distribution of EEG prediction correlations was then computed by running the TRF analysis on 20 shuffled versions of the expectation vectors per participant. The correlation gains achieved by using the expectation model (expectation - null model) were compared to the control distribution of "gains" determined based on the null models ($nullmodel_i - nullmodel_j$; see Figure 2.1).

Figure 2.4 shows that EEG prediction correlations were larger for the expectation signal than the null-model in both the listening and imagery conditions (Wilcoxon rank sum test; listening: $p = 4.2 \cdot 10^{-66}$, Common language effect size $f = 0.77$; imagery: $p = 3.4 \cdot 10^{-111}$, Common language effect size $f = 0.85$), with significance at the individual level for 12/21 participants ($p < 0.05$, FDR-corrected p-values extracted from the null-model distributions). We did not expect to observe within-subjects significance for all participants as each model was trained on one condition and therefore half of the data.

The shapes of the TRF kernels shown in Figure 2.5 were qualitatively similar to those depicted in Figure 2.2 when regressing the onsets signal. Interestingly, the effect of expectations (correlation gain) emerged on EEG channels that had little or no sensitivity to the unmodulated onsets, thus possibly reflecting different cortical generators for the EEG encoding of acoustics and expectations. In fact, the expectation gain emerged primarily in frontal scalp areas, which were previously linked with auditory expectations (Opitz *et al.*, 2002; Schönwiesner *et al.*, 2007; Tillmann *et al.*, 2003). Furthermore, the effect of expectation (correlation gain) had similar topographical distributions for the listening and imagery conditions (Pearson's correlation: $r = 0.9$.). This also suggests that the expectation signal is the same in both cases and originates from the same source. Figure 2.5 indicates that low frequencies (< 1 Hz) are important for expectation responses. However, even the analysis of the 1-30 Hz band displays significant encoding of expectation. Finally, the topographic distributions are similar for each frequency band, although somewhat weaker for 1-30 Hz.

Using the methodology above, we measured and compared the impact of the IDyOM short-term model, which relies on music patterns within a piece only, and the long-term model, which relies on music statistics derived from a large corpus of music not including the present piece (see the **Tools** section). First, we found that short-term expectations contribute significantly

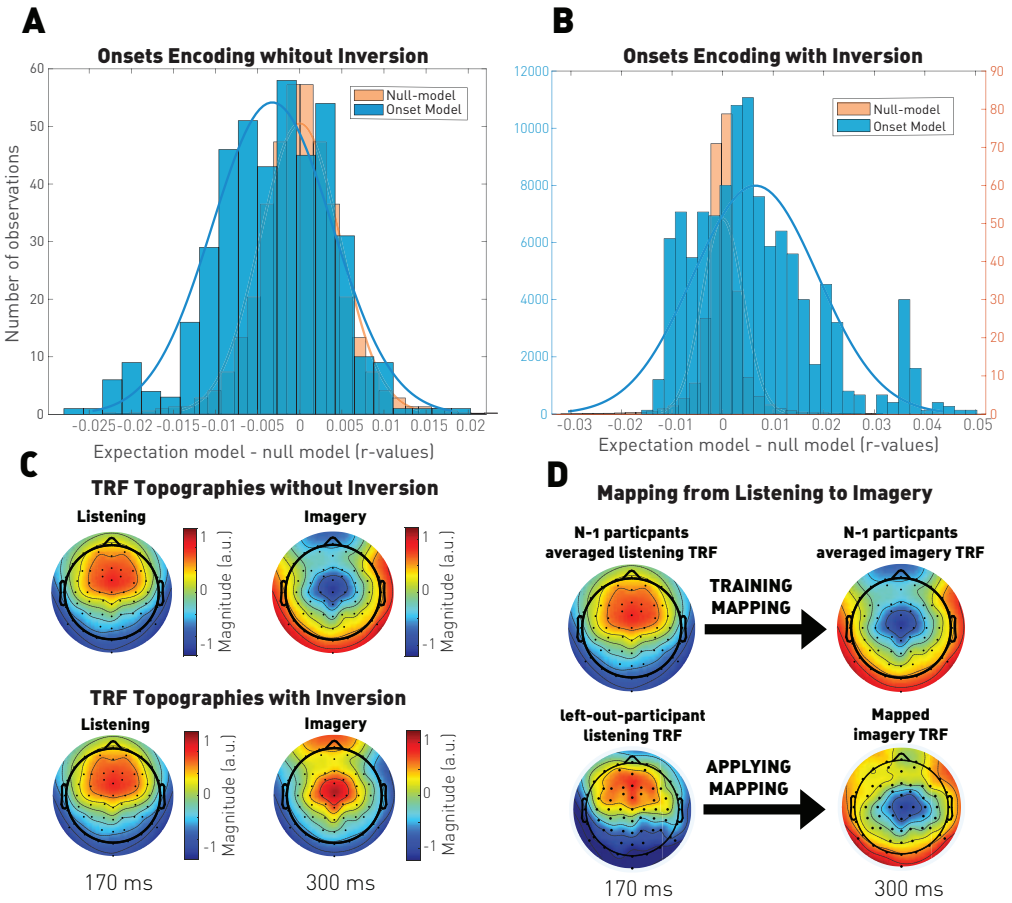


Figure 2.3. **Cross-Conditions Analysis.** TRF models fit on one condition and were evaluated on the other one to determine the consistency between conditions. (A) Distribution of the difference between the onsets model and the null-model prediction of the listening condition based on raw TRF kernels trained on the imagery condition. Significance was computed using a Wilcoxon rank sum test to assess that the distributions are above the control distribution. (B) Distribution of the difference between the onsets model and the null-model prediction of the listening condition based on inverted TRF kernels trained on the imagery condition. Significance was computed using a Wilcoxon rank sum test to assess that the distributions are above the control distribution ($p = 10^{-46}$). (C) TRF kernels topographies. The TRF kernels are normalized and extracted at the time where their Global Field Power was maximum to extract the latency where their responses were the most salient (170 ms for listening and 300 ms for imagery). We can observe a time-shifted inverted polarity of the responses that have been assessed in (B). We measured a significant ($p = 10^{-23}$) correlation of $r = 0.9$ between the listening and the imagery-inverted topographic maps. (D) A linear convolution mapping between the listening and imagery responses was learned, applied to individual listening responses, and resulted in significant predictions of the imagery EEG using the onsets ($p = 10^{-49}$).

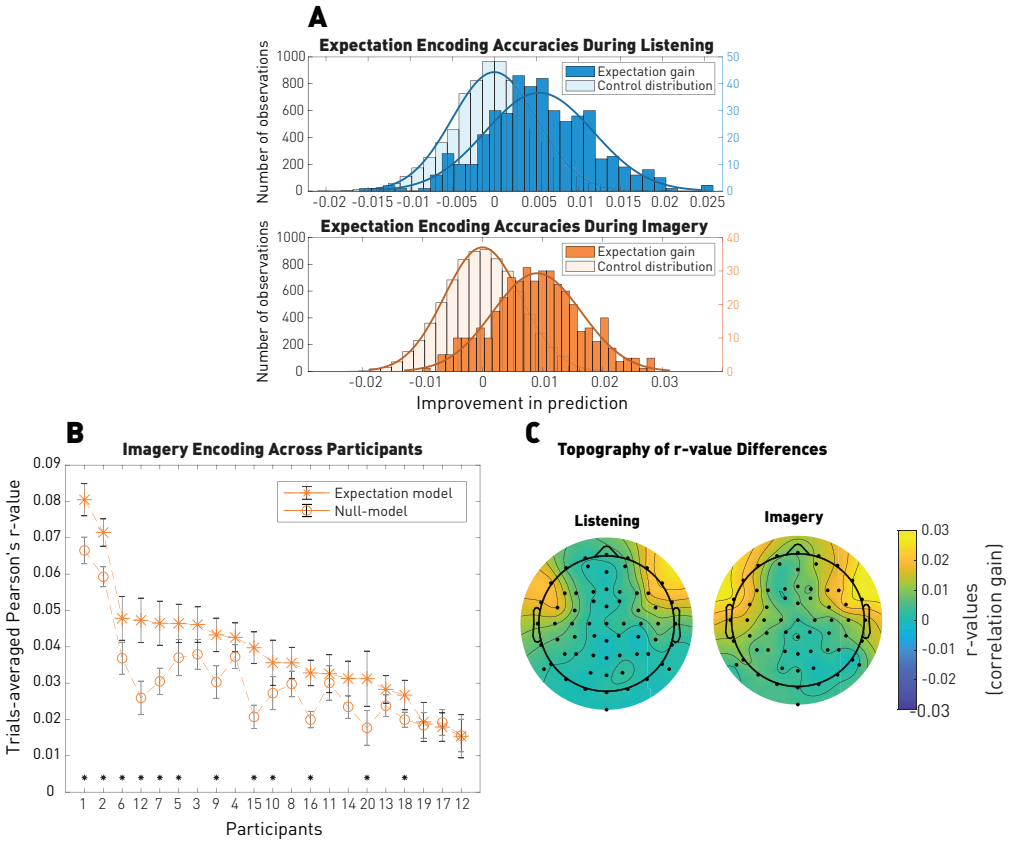


Figure 2.4. **Robust EEG Encoding of the Expectation Signal.** (A) EEG prediction correlations for the listening and imagery conditions using the expectation TRFs. EEG prediction correlations were significantly above chance in both conditions. (B) EEG prediction correlations at the individual participant level for the imagery condition. Error bars show the standard error across trials. Stars indicate significance ($p < 0.05$). (C) Topographies of the EEG predictions gain (expectation model - null model). Pearson's correlation between conditions: $r = 0.9$.

to the prediction of the EEG signals (listening: $p = 1.1 \cdot 10^{-107}$, Common language effect size: $f = 0.84$; imagery: $p = 6.9 \cdot 10^{-134}$, Common language effect size: $f = 0.88$), indicating that neural signals encode statistics based on the proximal melodic context. To examine and demonstrate that the long-term model is distinguishable and augments the expectation due to short-time-scale expectation features, we compared the expectations generated by a combined short-term + long-term model to one based on expectations from short-term + scrambled long-term processes (null-model). The resulting distributions shown in Figure 2.6 show a positive shift for the genuine models compared to the shuffled ones (listening: $p = 9.5 \cdot 10^{-64}$, Common language effect size: $f = 0.77$; imagery: $p = 1.2 \cdot 10^{-63}$, Common language effect size: $f = 0.77$). We then used the same analysis approach on the short-term expectations, showing that the short-term model captures information not explained by the long-term model (listening: $p = 3.7 \cdot 10^{-41}$, Common language effect size: $f = 0.72$; imagery: $p = 1.2 \cdot 10^{-77}$; Common language effect size $f = 0.79$). The topographical distributions of such contributions resemble

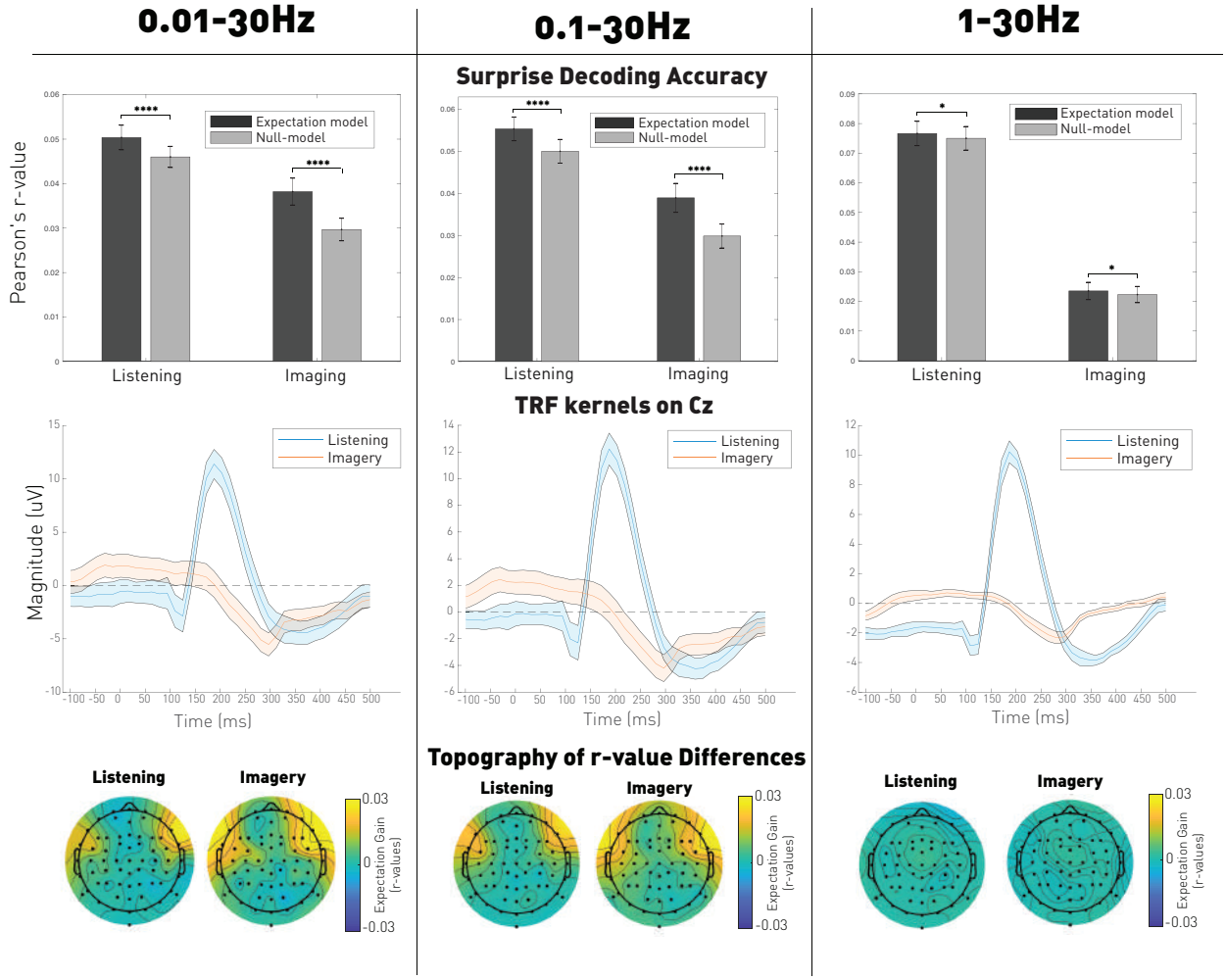


Figure 2.5. **EEG Encoding of the Expectation Signal by Frequency Bands** (0.01-30 Hz, 0.1-30 Hz, and 1-30 Hz). (top) Averaged prediction correlations for both the expectation model and null-models. Significance was computed using a Wilcoxon signed rank test paired by participants and averaged by trials and shuffling ($***: p < .0001, *: p < 0.05$). (middle) TRF kernels reflect the average neural response on Cz. Shaded error bars show the standard error across participants. (bottom) Topography of the prediction correlations gain (expectation model - null-model) over the electrodes.

those seen with the full expectation signal (the expectation values built by combining long- and short-term statistics; see Figure 2.4; short-term contribution: listening: $r = 0.67$, imagery: $r = 0.53$; long-term contribution: listening: $r = 0.75$, imagery: $r = 0.70$). Furthermore, similar topographic patterns were measured for the contributions of short- and long-term models (listening: $r = 0.64$; imagery: $r = 0.47$), suggesting that the neural activity explained by long- and short-term expectations originates in similar or overlapping brain areas.

Finally, we also examined the extent to which the correlation contribution due to the expectation signal is specifically related to the low-level features of the music signal (pitch, intervals, reversal in pitch direction, and duration). To do so, we compared the distribution of the correlations when regressing all these low-level features and the expectation signal on one side, compared to the distribution of the correlations computed when scrambling only the expectation vector (null-model). The difference between the two distributions shown in Figure 2.6 indicated that the expectation signal indeed contributed information beyond that due to the low-level features (listening: $p = 3.8 \cdot 10^{-155}$, Common language effect size: $f = 0.9$; imagery: $p = 7.9 \cdot 10^{-138}$, Common language effect size: $f = 0.89$). All these comparisons lead us to conclude that the long-term model, learned through exposure to a large corpus of music, is operable during both the listening and imagery conditions and in addition to the low-level musical features.

2.2.3.4 ERP Analysis

We conducted an ERP analysis by computing the average neural response in a window of [-100 ms, 500 ms] around the note-onsets at least 500 ms away from the metronome beats. The average power in the window of [-50 ms, 0 ms] was subtracted as a baseline. Significance between listening and imagery responses were computed using a permutation test from the values distributed by participants and topographic distributions were computed by plotting the response power over the scalp at specific time latencies. Finally, we also computed averaged responses for the 20% most expected and 20% less expected notes.

Figure 2.7.A shows that imagined notes elicit negative responses that are similar to the TRF kernels observed in Figure 2.4. In addition, notes in both listening and imagery conditions elicited stronger responses on the Cz-electrodes for notes related to low expectation (high surprise) as shown in Figure 2.7.B. This trend, even if not significant here, is consistent with the TRF analysis and in line with the literature (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Omigie *et al.*, 2013a). Finally, the topographic distribution of the ERP's in the two conditions is illustrated in Figure 2.7.C, highlighting the relative delay and inverted polarity of the imagery relative to listened responses.

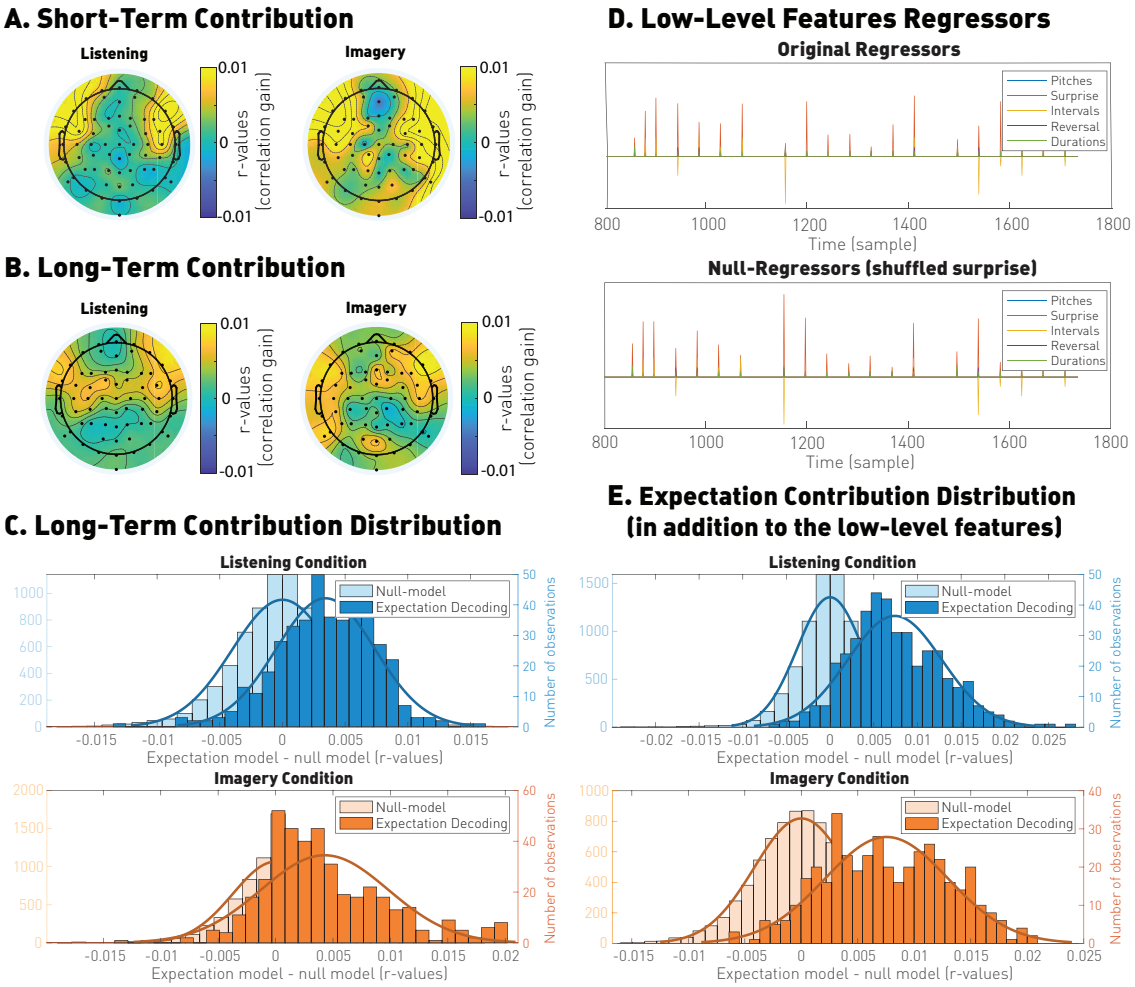


Figure 2.6. **Comparison of the short- and long-term and expectation and low-level features**. (A) Unique correlation contribution for short-term expectations. These values were calculated as the EEG prediction correlations with TRF models based on both long- and short-term expectations, minus the EEG correlations after shuffling the short-term expectation values. (B) Unique correlation contribution for long-term expectations. Correlation contribution of the long-term expectation model minus the EEG prediction correlations after shuffling the long-term expectation values. (C) Unique correlation contribution of the long-term model, showing that long-term expectations explain EEG variance that is not captured by long-term expectations. (D) TRF models were fit by combining low-level features (pitch, duration from the previous note, interval, reversal in pitch direction) were combined with the expectation vector. The null-model was derived by combining the same low-level features with a scrambled expectation vector. (E) The result of the TRF analysis shows that the expectation signal explains EEG variance that was not captured by the low-level features.

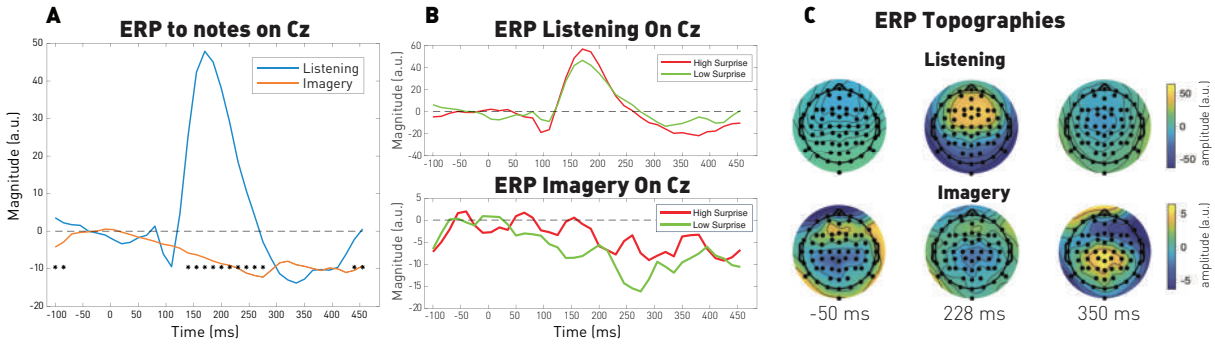


Figure 2.7. **ERP Analysis of Listened and Imagined Notes.** (**A**) Averaged responses for all notes. Significance between listening and imagery responses was computed using a permutation test from the values distributed by participants ($p < 0.05$) (**B**) Averaged responses for the 20% less and most expected notes in both listening (top) and imagery (bottom) conditions. (**C**) Participant-averaged topographic distributions from the ERP of all notes at least 500 ms away from the metronome.

2.2.3.5 Decoding Imagined Song Identity from the EEG

We tested whether the EEG encoding of note-onset and melodic expectation was sufficiently robust to reliably classify the song identity on single trials. To do so, EEG recordings were predicted using the TRF by regressing all four musical stimuli separately. The stimulus leading to the highest EEG prediction correlation was then selected for each trial (see **Material and Methods** section for more details). A null-model was computed by shuffling the songs in order to estimate the classification chance level.

Figure 2.8 shows significant classification accuracies, following the same trend, for each individual participant. Significance was computed using a Wilcoxon signed rank test paired by participants (listening: $p < 10^{-7}$, common language effect size $f = 1.0$; imagery: $p < 10^{-7}$, common language effect size $f = 1.0$). Note that statistical significance was determined based on the null-model performance rather than the theoretical chance level, which instead assumes infinite data-points (Combrisson & Jerbi, 2015).

2.2.3.6 Cross-Participants Analysis

In order to assess the variability in the neural responses across individuals, we used a leave-one-participant-out cross-validation technique. Specifically, average TRF models were trained on all participants but one, which was instead used for evaluation. The goal was to test whether the neural signals of individual participants were sufficiently consistent and synchronized between participants to allow for significant EEG predictions.

Figure 2.9 shows that the cross-participants analysis allowed for significant encoding of expectation. Significance was computed using a Wilcoxon rank sum test between expectation

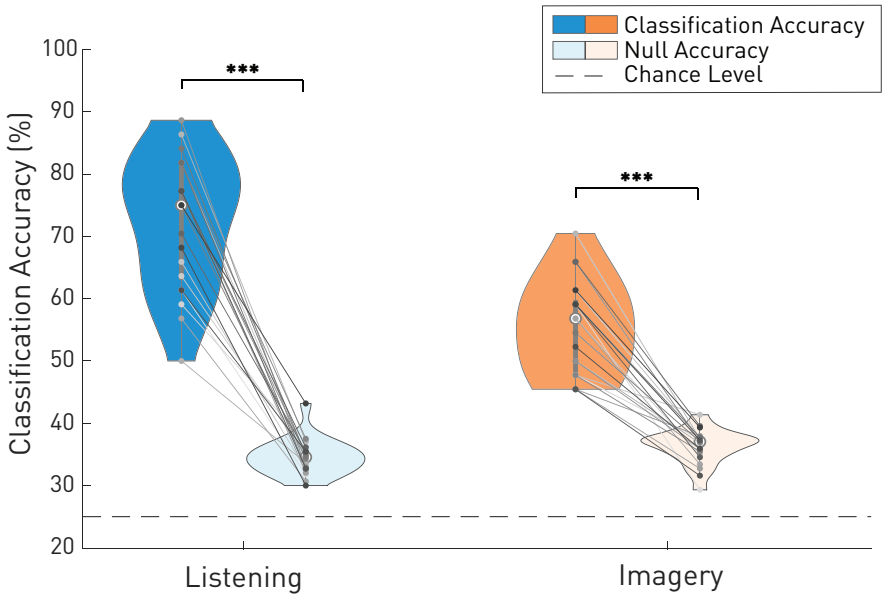


Figure 2.8. **Piece Classification Accuracy.** EEG predictions for note-onsets and melodic expectations were combined to determine which song was being listened to or imagined. The data are shown for each participant and indicate overall significance. The null-model was calculated from labels-shuffled data.

gain and control distributions (Listening: $p = 1.3 \cdot 10^{-108}$, common language effect size $f = 0.85$; Imagery: $p = 8.8 \cdot 10^{-68}$, common language effect size $f = 0.78$). Results were also significant on 11/21 individual participants for listening and 7/21 participants for imagery. Significance ($p < 0.05$) was assessed by comparing the probability of the observed expectation prediction correlation with the null-model distribution.

This analysis indicates that cortical responses were consistent between participants in both listening and imagery conditions, meaning that models can be trained and evaluated on different participants and that expectation encoding is shared between individuals within the same sociocultural environment (here professional classical musicians).

2.2.3.7 Comparison with Behavioral Audiation Measures

The literature is rich in behavioral measures of audiation capabilities (Gelding *et al.*, 2015; Gerhardstein, 2002; A. Halpern, 2015). We specified our analysis on one of these measures: the Gordon's Advanced Measure of Music Audiation (AMMA) designed by Edwin Gordon in 1989 to tackle audiation capabilities in musicians in order to tailor musical training and checked whether this test was correlated with the between-participant variability observed in our data.

Figure 2.10 shows that the onsets gain computed as the improvement of the onsets model with respect to its respective null-model (labels shuffled) does not significantly correlate with the AMMA audiation test. This finding suggests that the audiation capability as defined and

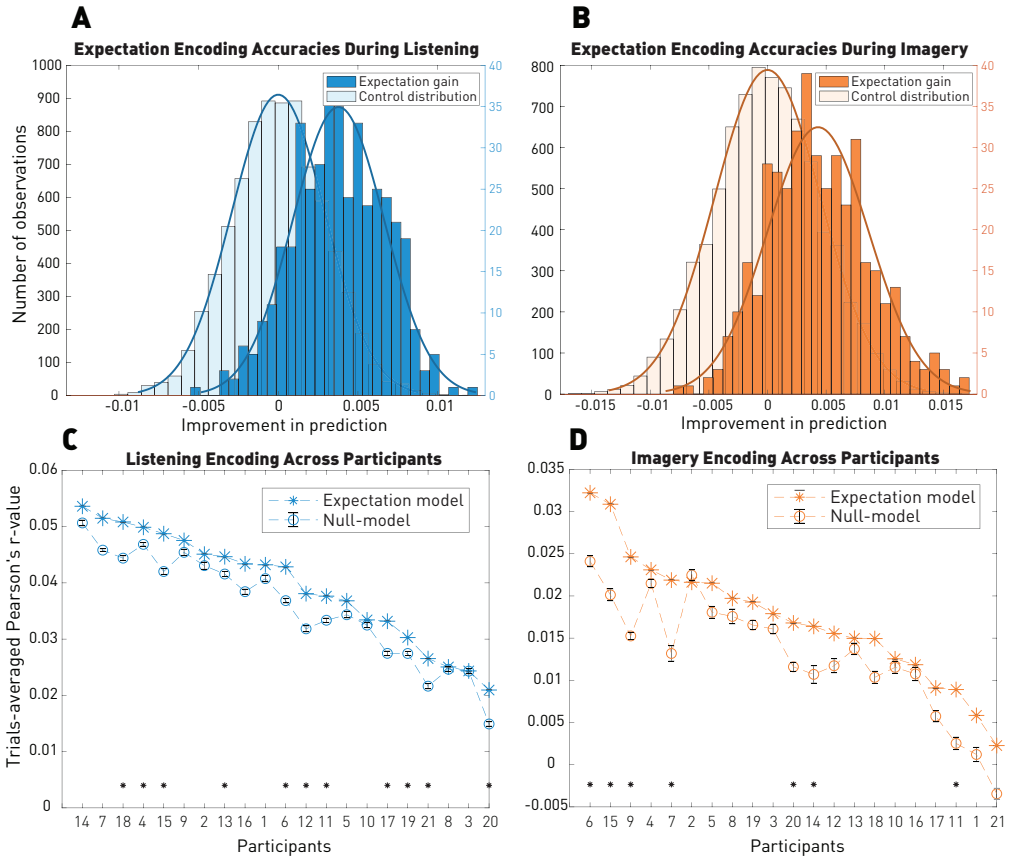


Figure 2.9. **Cross-participants analysis.** TRF models were fit by combining EEG data from all participants but one and evaluated on the left-out participant. (A) Distribution of expectation EEG prediction correlation gains (expectation - null model) during listening were significant when models were trained on different participants than the one of the evaluation. (B) Distribution of the expectation gain during imagery. The gain is conserved with models trained on different participants than the one of the evaluation.(C and D) Individual EEG prediction correlations for the listening (C) and imagery (D) conditions. Error bars for null-models indicate the standard error across shuffles. Stars indicate significance within participants ($*p < 0.05$).

measured by Gordon is something that is not reflected by the neural encoding of acoustics during imagery. A similar analysis based on the expectation gain instead of the acoustic gain has been conducted and resulted in similar results.

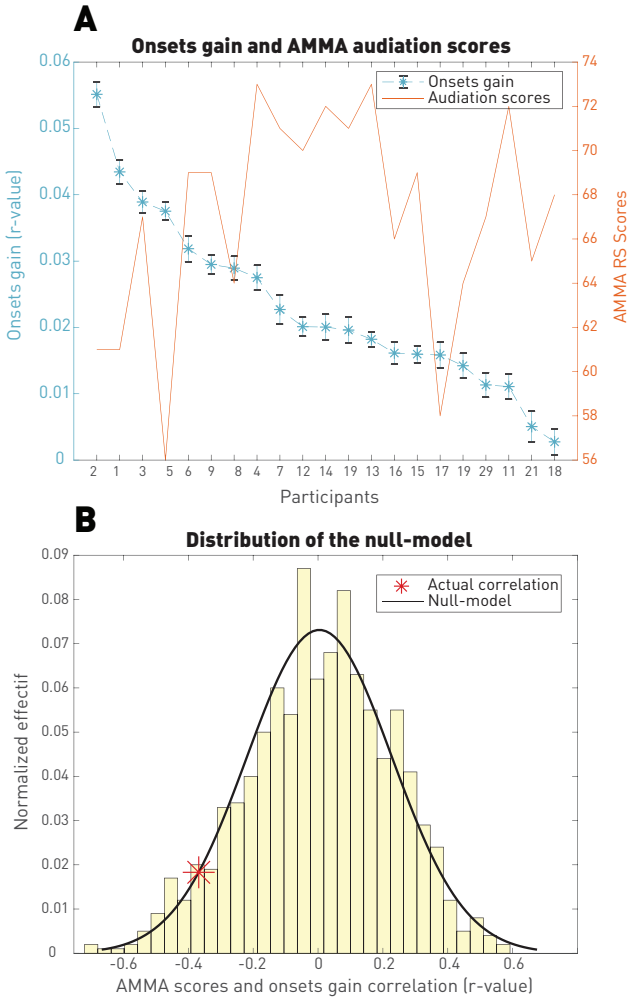


Figure 2.10. **Correlation of the Onset-Model Gain with the AMMA Audiation Test.** (A) Raw signals are shown in different axis. The Pearson's correlation computed on these two signals is $r = -0.36$. (B) This correlation is not significant as it resulted in a p-value $p > 0.05$ when looking at the null-distribution built by shuffling the order of participants. We therefore conclude that the AMMA audiation does not reflect the onsets gain.

2.2.4 Discussion

Neural responses recorded with EEG during musical imagery exhibited detailed temporal dynamics that reflected the effects of melodic expectations, and a TRF that is delayed and with an inverted polarity relative to that of responses exhibited during listening. The responses shared substantial characteristics across individual participants and were also strong and de-

tailed enough to be robustly and specifically associated with the musical pieces that the participants listened to or imagined.

This study demonstrates for the first time that melodic expectation mechanisms are as faithfully encoded during imagery as during musical listening. Electroencephalogram (EEG) responses to segments of music (and other auditory stimuli like speech) typically fluctuate based on the likelihood of hearing that particular sound within the ongoing sequence: the lower the probability (or unexpectedness), the more pronounced the EEG expectation response (Di Liberto, Pelofi, Bianco, *et al.*, 2020). Therefore, the discovery that imagined music undergoes similar modulation to heard music implies insights into the essence and function of musical expectation in shaping the perceptual markers of our cognitive processes. Comparable to language, these expectation mechanisms are employed to delineate musical phrases and discern grammatical elements that can later serve various cognitive purposes. This notion has been previously deliberated upon, and multiple studies have underscored the pivotal role of musical expectations as fundamental elements in diverse cognitive functions, ranging from memory processes (K. Agres *et al.*, 2018) to the experience of musical pleasure (Gold, Pearce, *et al.*, 2019). Notably, instances of expectations being met or unmet have been observed to influence brain activity in regions associated with the reward system (Cheung *et al.*, 2019), particularly in relation to emotional pleasure (Blood & Zatorre, 2001; Zatorre & Salimpoor, 2013), and the release of dopamine (Salimpoor *et al.*, 2011). Consequently, it is plausible that imagery elicits similar emotional responses and pleasure akin to those experienced during active musical listening due to the analogous encoding of melodic expectations in both scenarios. This elucidates why musical imagery serves as a versatile platform for music creation and holds considerable significance in the realm of music education. When Robert Schumann asked his students to arrive at the point of "hearing music from the page", he suggested that there exists individual variability in the vividness of imagery, which can be shaped and improved by practice. This ability can be assessed via behavioral measures (Gelding *et al.*, 2015; Gerhardstein, 2002; A. Halpern, 2015), and has also been shown to correlate with neural activity in fMRI (A. Halpern, 2015). In fact, it may also reflect language deficits as seen in children with *Specific Language Impairment* (SLI) who often exhibit significantly lower scores in behavioral musical imagery tests, suggesting shared neurodevelopmental deficits (Heaton *et al.*, 2018). Curiously, we did not find a significant correlation between the strength of the neural encoding of music and the participants' audiation scores from the widely-used Gordon's AMMA audiation test (see Figure 2.10). This can partially be explained by the weak SNR of the EEG signal, as well as by complex aptitudes that are not captured by the AMMA test. Therefore, we still lack an adequate demonstration of a link between our participants' ability to imagine and behavioral measures that can better indicate the cognitive underpinnings of the vividness of their imagery. By extension, the same lack of evidence applies to language deficits and their potential remediation

through musical training.

From a system's perspective, auditory imagery responses can be thought of as "predictive" responses, induced by top-down processes that normally model how an incoming stimulus is perceived in the brain, or the perceptual equivalent of the efference copy, often triggered by the motor system (Ventura *et al.*, 2009). This analogy has inspired numerous studies of auditory imagery in motor contexts as in covert speech, suggesting that imagined responses can be of a predictive motor nature (Y. Ding *et al.*, 2019; Tian & Poeppel, 2010; 2012; 2013; Whitford *et al.*, 2017). In musical imagery, rhythm, in particular, has been closely linked to the activity of the Supplementary Motor Areas (SMA) and pre-SMA (Bastepe-Gray *et al.*, 2020; Gelding *et al.*, 2019; A. R. Halpern, 2001; A. R. Halpern & Zatorre, 1999; Herholz *et al.*, 2012; Lima *et al.*, 2015; 2016; Meister *et al.*, 2004; Zatorre & Halpern, 2005), while notational audiation (Brodsky *et al.*, 2008) (musical imagery driven by reading music scores) and listening (Pruitt *et al.*, 2018) have been shown to generate covert excitation of the vocal folds with a neural signature similar to that observed during musical imagery (Zatorre *et al.*, 1996). This motor-imagery link also runs in reverse as demonstrated by an ECoG study that reveals strong auditory responses induced by *silent* playing of a keyboard (Martin *et al.*, 2017). In conclusion, it is evident that imagery may well be facilitated by the intimate links that exist between motor and sensory areas that are normally co-activated in task performance, e.g., vocal-tract and speech production (Shamma *et al.*, 2020), fingers and piano playing, and vision and reading. This also makes it difficult experimentally to disentangle the two sources of activity (Zatorre *et al.*, 2007) since auditory imagery may partially be affected by motor components (A. R. Halpern & Zatorre, 1999).

Regardless of their origins, imagery responses should be fully considered as top-down predictive signals, with the most striking evidence in our data being their inverted polarity relative to the listening responses. Such an inversion facilitates the comparison between bottom-up sensory activation and its top-down prediction by generating the "error" signal, long postulated in predictive coding theories to be the critical information that is propagated deep into the brain (Koster-Hale & Saxe, 2013; Rao & Ballard, 1999). This key observation is explored in detail in the companion study (Di Liberto *et al.*, 2021), which analyzed the EEG responses evoked during the pauses or short silences that are naturally interspersed within a musical score. These responses are analogous to imagery responses in that both lack direct stimuli to evoke them. The combined findings in the present work and the companion study provide a common framework that remarkably and seamlessly links listened and imagined music perception, and more broadly, sensory responses and their prediction in the brain.

2.3 The Music of silence. Part II: Cortical Predictions during Silent Musical Intervals [4]

2.3.1 Introduction

Silence is an essential component of our auditory experience, which serves important communicative functions by contributing to expectation, emphasis, and emotional expression. Here we investigate the neural encoding of silence with electroencephalography (EEG) and music stimuli.

That perception is underpinned by an interplay of sensory input and endogenous neural processes and has been a longstanding area for debate (Clark, 2016; den Ouden *et al.*, 2012; Heeger, 2017; Pouget *et al.*, 2013). Prediction theories (Spratling, 2017) posit that the brain continuously attempts to predict its upcoming sensory inputs, comparing (subtracting) them and hence deriving a prediction error (ϵ_{sur}) that is used to improve its internal (prediction) model of the world. A large body of research has found prediction effects in line with several neurophysiological phenomena, such as the magnitude modulation of sensory responses with their expectation (Kutas & Hillyard, 1980; 1984; Rabovsky *et al.*, 2018), where larger responses were measured for more unexpected inputs. In auditory neurophysiology, this prediction phenomenon has been extensively investigated using the responses evoked by sound stimuli (Friederici *et al.*, 1993; Kutas & Federmeier, 2011; Mars *et al.*, 2008; Seer *et al.*, 2016; Strauss *et al.*, 2013; Sutton *et al.*, 1965). A less common approach involves studying the predictions in the absence of the acoustic input, i.e., during silence, a strategy that potentially unveils neural predictive processing and its top-down mechanisms by decoupling it from the simultaneous bottom-up sensory inputs (Heilbron & Chait, 2018; Walsh *et al.*, 2020).

Vigorous responses to silences have been observed across modalities when a sensory stimulus was strongly expected, for example corresponding to an omission during the rapid isochronous presentation of tones (Chennu *et al.*, 2016; Joutsiniemi & Hari, 1989; Simson *et al.*, 1976; Yabe *et al.*, 1997). This finding demonstrated that unexpected silences can elicit robust neural responses that do not require a concurrent sensory input. However, silence has a much more pervasive presence in our auditory experience than what can be captured in the stimulus omission scenario, which is limited to silences occurring in place of highly expected stimuli. In fact, silence is a fundamental component of the rhythmic structure of music that can correspond to a wide range of expectation strengths. The regularities of music prompt our brain to build such expectations, which are accurately estimated by computational models of musical structure (M. T. Pearce, 2005), allowing us to assess the precise neural encoding of music expectations.

[4] Authors: Giovanni Di Liberto, Guilhem Marion, Shihab Shamma(Di Liberto *et al.*, 2021)

While such expectations have been shown to be encoded in the neural responses to notes in a melody during listening (Di Liberto, Pelofi, Bianco, *et al.*, 2020), little is known about the neural encoding of internally generated music.

Here we investigate the role of silence on the neural processing of music with EEG recorded as participants *listened to* or *performed mental imagery of* excerpts from Bach chorales. Endogenous and exogenous components of the neural signal are discerned by studying the comparison between listening and imagery conditions. According to prediction theories, the brain continuously builds predictions of upcoming music notes, with the prediction signal (P) appropriately modulated by the uncertainty of the prediction (Koelsch *et al.*, 2019). When subtracted from the sensory response (S), it produces a "surprise" or prediction error signal that is measurable with EEG ($\epsilon_{\text{sur}} = S - P$) (Grisoni *et al.*, 2019; Heilbron & Chait, 2018). In this study, we assumed "S" and "$-P$" to contribute to the EEG signal as two distinct additive components, where P mimics S and, conversely, "$-P$" has inverse polarity compared with S. Under that assumption, encountering silence when a note is plausible would correspond to a measurable EEG signal reflecting the neural prediction error signal "$-P$", which depends solely on the prediction signal P as $S = 0$, thus presenting the inverse polarity of the otherwise dominant sensory response (Fig. 1; see also (Bendixen *et al.*, 2009; Heilbron & Chait, 2018)). For these reasons, we hypothesized robust neural correlates to emerge in correspondence with the silent events of music, reflecting the prediction error "$-P$" and with magnitude changing with the expectation strengths.

The music imagery task allowed us to study the neural encoding of music silence further by investigating endogenous neural components in the absence of sensory responses. In an accompanying study (Marion *et al.*, 2021), we have shown robust neural activation corresponding to imagined notes, extending previous work on auditory imagery (A. R. Halpern & Zatorre, 1999; Kraemer *et al.*, 2005; Zhang *et al.*, 2017) by demonstrating that cortical signals encode melodic expectation during imagery. In the present work, conducted along with Giovanni Di Liberto and in line with prediction theories, we hypothesized that P is the main source of such neural activity since $S = 0$. As such, we anticipated a prediction signal ($-P$) to emerge in the EEG responses to both imagined notes and silent events, with inverse polarity relative to a sensory response. Finally, we anticipated the magnitude of the responses to silent events to reflect the precise expectation strengths of each music event, which were estimated by means of a computational model of melodic structure (M. T. Pearce, 2005), as it was demonstrated for music listening (Di Liberto, Pelofi, Bianco, *et al.*, 2020) and imagery (Marion *et al.*, 2021).

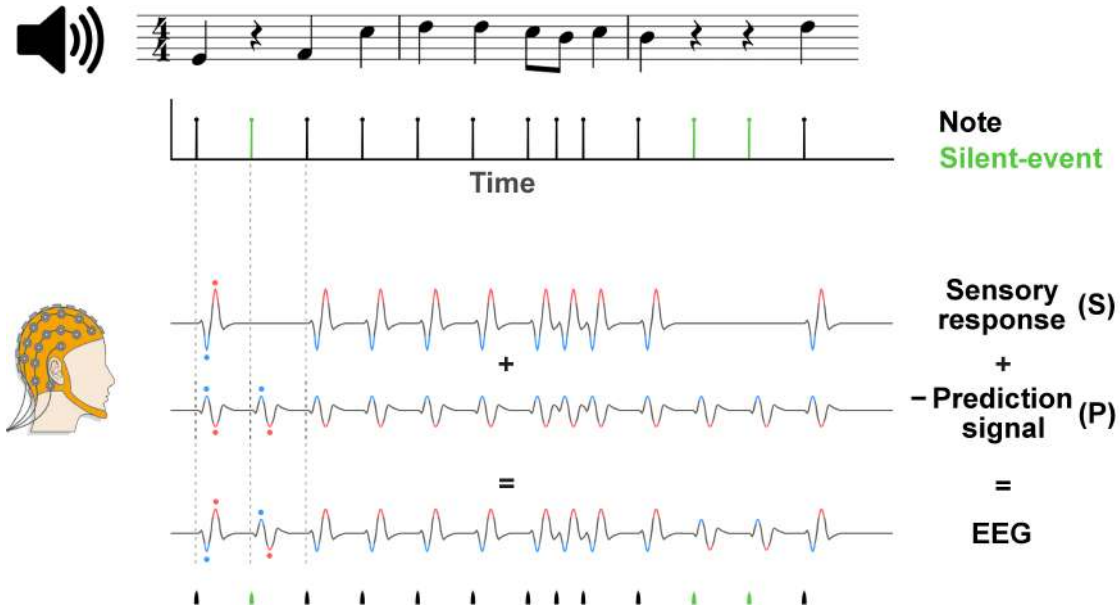


Figure 2.11. **Figure 1. Simplified predictive processing model** demonstrating the predictive processing hypothesis for the perception of melodies. Electroencephalography (EEG) signal recorded during monophonic music listening was hypothesized to reflect the linear combination of a *sensory evoked-response* (S) and a *neural prediction signal* (P). In line with the predictive processing framework, we modeled the EEG signal as a combination of the distinct components S and P; Specifically, as the subtraction S-P or, equivalently, S + (-P). Having defined P as a signal reflecting the attempt of our brain to predict the sensory stimulus, we posited P to emulate S (with $|S|>|P|$) and to have larger magnitude with stronger expectations (the expectation strengths are not included in this figure, for simplicity). As such, the S-P signal would become "-P" when a prediction is possible but no sensory stimulus is present (S=0), producing an overall EEG signal with inverse polarity compared with the response to a note. In other words, EEG responses with opposite polarities were expected for events with and without an input sound (see polarities for events marked in black and green in the figure). After selecting silent events as the instants where a note was plausible but did not occur (based on IDyOM, see **Methods**), the existence and precise dynamics of the prediction signal were assessed: 1) By comparing the responses to silent events during melody listening, where P could be measured in isolation as S=0; 2) By studying the neural processing of music during imagery, where P could be isolated as S=0 for both notes and silent-events; and 3) By separating S and P with a component analysis method.

2.3.2 Materials and Methods

2.3.2.1 EEG experiment 1

Data acquisition and experimental paradigm Twenty healthy subjects (10 females, aged between 23 and 42, M = 29) participated in the EEG experiment. Ten of them were highly trained musicians with a degree in music and at least ten years of experience, while the other participants had no musical background. Each subject reported no history of hearing impairment or neurological disorder, provided written informed consent, and was paid for their participation. The study was undertaken in accordance with the Declaration of Helsinki and was approved by the CERES committee of Paris Descartes University (CERES 2013-11). The experiment was carried out in a single session for each participant. EEG data were recorded from 64 electrode positions, digitized at 512 Hz using a BioSemi Active Two system. Audio stimuli were presented at a sampling rate of 44,100 Hz using Sennheiser HD650 headphones and Presentation software (*http://www.neurobs.com*). Testing was carried out at École Normale Supérieure, in a dark room, and subjects were instructed to maintain visual fixation on a crosshair centered on the screen and to minimize motor activities while music was presented.

Stimuli and procedure

Stimuli and procedure Monophonic MIDI versions of ten music pieces from Bach's monodic instrumental corpus were partitioned into short snippets of approximately 150 seconds. The selected melodies were originally extracted from violin (partita BWV 1001, presto; BWV 1002, allemande; BWV 1004, allemande and gigue; BWV 1006, loure and gavotte) and flute (partita BWV1013 allemande, corrente, sarabande, and bourrée angloise) scores and were synthesized by using piano sounds with MuseScore 2 software (MuseScore BVBA), each played with a fixed rate (between 47 and 140 bpm). This was done in order to reduce familiarity for the expert pianist participants while enhancing their neural response by using their preferred instrument timbre (Pantev *et al.*, 2001). Each 150s piece, corresponding to an EEG trial, was presented three times throughout the experiment, adding up to 30 trials that were presented in random order. At the end of each trial, participants were asked to report on their familiarity to the piece (from 1: unknown; to 7: know the piece very well). This rating could take into account both their familiarity with the piece at its first occurrence in the experiment, as well as the build-up of familiarity across repetitions. Participants reported repeated pieces as more familiar (paired t-test on the average familiarity ratings for all participants across repetitions: rep2 > rep1, $p = 6.9 \times 10^{-6}$; rep3 > rep2, p = 0.003, Bonferroni correction). No significant difference emerged between musicians and non-musicians on this account (two-sample t-test, p = 0.07, 0.16, 0.19

for repetitions 1, 2, and 3 respectively; (Di Liberto, Pelofi, Bianco, *et al.*, 2020)).

2.3.2.2 EEG experiment 2

Data acquisition and experimental paradigm Twenty-one healthy subjects (6 females, aged between 17 and 35, median = 25) participated in the EEG experiment. All participants were highly trained musicians with a degree in music. Each subject reported no history of hearing impairment or neurological disorder, provided written informed consent, and was paid for their participation. The study was undertaken in accordance with the Declaration of Helsinki and was approved by the CERES committee of Paris Descartes University (CERES 2013-11). The experiment was carried out in a single session for each participant. EEG data were recorded from 64 electrode positions and digitized at 2048 Hz using a BioSemi Active Two system. Three additional electrodes were placed on the upper midline of the neck, the jaw, and the right wrist to control for motor movements of the tongue, masseter muscle, and forearm fingers extensors respectively. Audio stimuli were presented at a sampling rate of 44,100 Hz using a Genelec 8010-10w loud speaker and custom Python code. Testing was carried out at École Normale Supérieure, in a dimmed room. Participants were instructed to minimize motor activities while performing the task.

The experiment consisted of 88 trials in which participants were asked to either listen or perform mental imagery of ~35 second melodies from a corpus of Bach chorales (see *stimuli and procedure*). The entire stimulus set consisted of four such melodies, with each melody being presented 11 times per condition (*listening* and *imagery*) over the duration of the experiment. The presentation order of the resulting 88 trials was randomized. Participants were asked to read the music scores placed at the center of the desk during both listening and imagery conditions. Participants were provided with the scores before the experiment and asked to become familiar with the melodies. This pre-exposure to the music material was planned to maximize the imagery performance. A tactile metronome (Peterson Body Beat Vibe Clip) marking the start of 100 bpm bars (each 2.4 s) was placed on the left ankle of all participants to allow them to perform the mental imagery task with high temporal precision. A constant lag of 35ms was determined during the pilot experiments based on the subjective report on the participants, who reported that the metronome with lag 0ms was not in sync with the music. That correction was applied for all participants with the same lag value. Neural data from 0 to 500 ms after each metronome onset were excluded from the main analyses in **Figures 2 and 3** to ensure that the results do not reflect tactile responses. The metronome responses were analyzed separately to assess the dynamics of the tactile response (see **Figure 3G**). Note that the EEG response to the metronome reflects a mixture of tactile and auditory responses in the listening condition.

Before the experiment, musical imagery skills (or audiation skills) were assessed for every subject with "The Advanced Measures of Music Audiation" test (AMMA; *https://giamusicassessment.com/*).

Stimuli and procedure Four melodies were selected from a monophonic MIDI corpus of Bach chorales (BWV 349, BWV 291, BWV354, BWV 271). All chorales use similar compositional principles: the composer takes a melody from a Lutheran hymn (*cantus firmus*) and harmonizes three lower parts (alto, tenor and bass) accompanying the initial melody on soprano. The monophonic version of those melodies consists of the *canti firmi*. Original keys were used. The four melodies are based on a common grammatical structure and show very similar melodic and rhythmic patterns. The audio stimuli were synthesized using a Fender Rhodes simulation software (Neo-Soul Keys) with 100 bpm, each corresponding to the start of a bar (every 2.4 seconds).

2.3.2.3 EEG data preprocessing

Neural data from both experiments were analysed offline using MATLAB software (The Mathworks Inc). EEG signals were digitally filtered between 1 and 30 Hz using a Butterworth zero-phase filter (low- and high-pass filters both with order 2 and implemented with the function *filtfilt*), and down-sampled to 64 Hz. EEG channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation. Channels were then re-referenced to the average of the 64 channels. The TRF weights did not qualitatively change when using high-pass filters down to 0.1 Hz. Low-frequencies below 1 Hz were crucial for the melodic expectations analysis in **Figure 5**, which was based on EEG data filtered between 0.1 and 30 Hz (see Marion et al., 2021 for more extensive analyses on the EEG frequency-band).

2.3.2.4 IDyOM

The Information Dynamics of Music model (IDyOM; (M. T. Pearce, 2005)) is a framework based on variable-order hidden Markov models. Given a note sequence of a melody, the probability distribution over every possible note continuation is estimated for every n-gram context up to a given length k (model order). The distributions for the various orders were combined according to an entropy-based weighting function (M. T. Pearce, 2005), Section 6.2). Here, we used an unbounded implementation of IDyOM that builds n-grams using contexts up to the size of each music piece. In addition, predictions were the result of a combination of long- and short-term models (LTM and STM respectively), which yields better estimates than either model alone. The LTM was the result of a pre-training on a large corpus of Western music that did not include the stimuli presented during the EEG experiment, thus simulating the statistical

knowledge of a listener that was implicitly acquired after a lifetime of exposure to music. The STM, on the other hand, is constructed online for each individual music piece that was used in the EEG experiment.

Our choice of IDyOM was motivated by the empirical support that Markov model-based frameworks received as a model of human melodic expectation (Omigie *et al.*, 2013b; M. T. Pearce & Wiggins, 2006; M. T. Pearce *et al.*, 2010; Quiroga-Martinez *et al.*, 2019). Furthermore, a previous study from our laboratory demonstrated robust coupling between the melodic expectations calculated with this configuration of IDyOM and cortical responses to music (Di Liberto, Pelofi, Bianco, *et al.*, 2020).

2.3.2.5 Music features

In the present study, we have assessed the coupling between the EEG data and various features of the music stimuli. The note onset-time information was extracted from the MIDI files and encoded into time-series marking with an impulse with value one all note onsets (**NT**), with length matching that of the corresponding music piece and with the same sampling frequency as the EEG data (**Fig. 2A**). We then used IDyOM to identify "silent-events", i.e., time instants without a note, but where a note could have plausibly occurred. IDyOM does not encode silent events explicitly, so we applied custom changes to the original Lisp code to extract the information of interest on the silent events without changing the way IDyOM operates. Specifically, for each note, with a quantization of $1/16^{\text{th}}$ of a bar, IDyOM was used to search for the time for the next most likely event. The search continued for progressively longer latencies until the model predicted a note with a high likelihood (>0.3). We called those instants "silent events". The procedure was repeated on the silent-event instants, to predict where the next note would occur by knowing that there was no note where the model had predicted. This information was then encoded into time-series marking with a unit impulse each silent-event onset (**SIL**). Experiment 1 had a total of 23514 notes and 5202 silent events. In Experiment 2, 1548 notes and 271 silent events were used to fit the TRF in each condition (listening and imagery). Note that such events co-occurring with the tactile metronome were excluded. **Figures 2C,D,** and **Figure 3E,F** report additional information on the distribution of notes and silent events in the two experiments.

In order to investigate the cortical processing of note and silence expectations, we estimated the *surprise* and *entropy* values for each individual note of a given music piece by using IDyOM. Given a note e_i, a note sequence $e_{1..n}$ that immediately precedes that note, and an alphabet E describing the possible onset-time values for the note, *surprise* $S(e_i|e_{1..i-1})$ refers to the inverse probability of occurrence of a particular note at a given position in the melody(MacKay, 2003; M. T. Pearce *et al.*, 2010):

$$S\left(e_i|e_{1..i-1}\right) = \log_2 \frac{1}{p\left(e_i|e_{1..i-1}\right)} \ .$$

The entropy in a given melodic context was defined as the Shannon entropy (Shannon, 1948) computed by averaging the surprise over all possible continuations of the note sequence, as described by E:

$$H\left(e_{1..i-1}\right) = \sum_{e \in E} p\left(e|e_{1..i-1}\right) S\left(e|e_{1..i-1}\right) \ .$$

In other words, the entropy provides an indication of the uncertainty of the upcoming music event given the preceding context.

IDyOM simulates implicit melodic learning by estimating the probability distribution of each upcoming note. This model can operate on multiple viewpoints, meaning that it can capture the distributions of various properties of music. Here, we focused on the *onset time* viewpoint. IDyOM generates predictions of upcoming music events based on what is learned, allowing the estimation of entropy values for the properties of interest. Each of these features was encoded into time series by using their values to modulate the amplitude of a note-onset vector. This resulted in four-time series: surprise and entropy of the onset time for *notes* (S_{NT} *and* H_{NT}) and *silences* (S_{SIL} and H_{SIL}).

2.3.2.6 Temporal response function analysis (TRF)

A system identification technique was used to compute the channel-specific music-EEG mapping for analyzing the EEG signals from both experiments. This method, here referred to as the temporal response function (TRF; (N. Ding *et al.*, 2014; Lalor *et al.*, 2009a)), uses a regularized linear regression (Crosse, Di Liberto, Bednar, & Lalor, 2016) to estimate a filter that optimally describes how the brain transforms a set of stimulus features into the corresponding neural response (forward model; **Fig. 2A**). Leave-one-out cross-validation (across trials) was used to assess how well the TRF models could predict unseen data while controlling for overfitting. Specifically, we implemented a nested-loop cross-validation, with the inner loop consisting of a leave-one-out cross-validation where TRF models were fit on the training fold and used to predict the EEG signal of the left-out trial. The purpose of the inner loop was to determine the optimal regularization parameter ($\lambda \in [10^{-9}, 10^{5}]$) by selecting the one maximizing the EEG prediction correlation (averaged across all electrodes and validation trials). The outer loop iterated over each left-out test trial, where the TRF model was fit on all other trials (using the optimal regularization parameter identified with the inner loop) and the quality of the model was quantified by calculating the Pearson's correlation between the preprocessed recorded signal and its prediction at each scalp electrode.

The interaction between stimulus and recorded brain responses is not instantaneous, in fact, a sound stimulus at time t_0 affects the brain signals for a certain time-window $[t_1, t_1+t_{win}]$, with $t_1 \geq 0$ and $t_{win} > 0$. The TRF takes this into account by including multiple time-lags between stimulus and neural signal, providing us with model weights that can be interpreted in both space (scalp topographies) and time (music-EEG latencies). The relative long inter-onset-interval (IOI) between music events (e.g., the most common note duration was 300 ms in experiment 2) could constitute a challenge for the TRF analysis, which may erroneously associate a neural response to a note *n* to the previous note *n*-1 because of the intrinsic regularity of music. To overcome this limitation, a broad time-lag window of -300–900 ms was selected to fit the TRF models, which enabled the regression model to more reliably distinguish the response to the current and neighboring events.

A univariate forward TRF analysis was used to assess the neural response to music notes and silent events. TRF models were fit for relating NT and SIL with the EEG signal from experiments 1 and 2. Note that note and silent-event TRFs were fit separately. The temporal dynamics of the neural response to music were then inferred from the TRF model weights for latencies that were considered physiologically plausible according to previous work (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Freitas *et al.*, 2018; Jagiello *et al.*, 2019), as shown in **Figures 2B**, **3C**, and **3D**. A multivariate TRF analysis was also conducted for Experiment 1 by combining NT and SIL, which allowed us to assess the neural signature corresponding to silent events while regressing the possible impact of the evoked responses to the preceding notes (**Figure 2E**).

In Experiment 2, a multivariate TRF analysis was also used to assess the cortical encoding of melodic surprise for *note* and *silence* events separately. Specifically, given either *note* or *silence* events, forward TRF models were fit by representing the stimulus as the concatenation of the corresponding 1) onset time-vector and 2) entropy time-vector; 3) and surprise time-vector. Then, the TRF analysis was repeated after shuffling the expectation values (entropy and surprise) values in the multivariate regressor. Specifically, a random permutation was applied to shuffle the entropy and surprise values of the events while preserving the onset time. This allowed for the comparison of the TRF models with shuffled modes with the same dimensionality but with meaningless melodic expectation values sequences (see the **Statistical analyses** subsection for additional details on the permutation analysis). The rationale was that the inclusion of melodic expectation information improves the EEG prediction correlations only if the EEG responses to music are modulated by such expectations, a phenomenon that was already shown for *notes* (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Marion *et al.*, 2021) but not for *silences*.

2.3.2.7 Multiway canonical correlation analysis (MCCA)

The TRF analysis has some limitations, such as working under the assumption of time-invariance of the neural responses to notes and silent events. That could be an issue because it is possible that the responses to silence change depending on its position (e.g., two consecutive silences). However, ERP analysis makes the same assumption and the high level of noise in the EEG hampers our ability to study questions on the raw data. We tackled this issue in Experiment 2 with multiway canonical correlation analysis (MCCA), a tool that merges EEG data across subjects to improve the SNR. MCCA is an extension of canonical correlation analysis (CCA; Hotelling, 1936) to the case of multiple (> 2) datasets. Given N multichannel datasets Y_i with size $T \times J_i$, $1 \leq i \leq N$ (time x channels), MCCA finds a linear transform W_i (sizes $J_i \times J_0$, where $J_0 < \min(J_i)$ for $1 \leq i \leq N$) that, when applied to the corresponding data matrices, aligns them to common coordinates and reveals shared patterns (de Cheveigné et al. 2018). These patterns can be derived by summing the transformed data matrices: $Y = \sum_{i=1}^{N} Y_i W_i$. The columns of matrix Y, which are mutually orthogonal, are referred to as summary components (SC) (de Cheveigné et al. 2018). The first components are signals that most strongly reflect the shared information across the several input datasets, thus minimizing subject-specific and channel-specific noise. Here, these datasets are EEG responses to the same task for 21 subjects. Note that EEG data were averaged across the 11 repetitions of each musical piece to improve the SNR before running the MCCA analysis.

This technique allows to extract a "consensus EEG signal" that is more reliable than that of any subject. This methodology is a better solution than averaging data across subjects which, in the absence of appropriate co-registration, leads to a loss of information because of topographical discrepancies. MCCA accommodates such discrepancies without the need for co-registration. Under the assumption that the EEG responses to music and music imagery share a similar time course within a homogeneous group of young adults, the MCCA procedure allows us to extract such common cortical signals from other, more variable aspects of the EEG signals, such as subject-specific noise. For this reason, our analysis focuses on the first N_{SC} summary components, which we can consider as spanning the most reliable EEG response to music and music imagery. N_{SC} was set to the number of components with the largest (5^{th} percentile) correlation with the original EEG data (N_{SC}=10 and N_{SC}=8 for the listening and imagery conditions respectively). De-noised EEG data were then calculated by inverting the MCCA mapping and projecting the N_{SC} summary components back to the subject-specific EEG channel space. The latter procedure allowed us to study the MCCA results in the same space as the TRF results (EEG channels) and to assess the robustness of the result across participants.

This last step was executed twice. First, de-noised EEG data were calculated by using only the first summary component which, intuitively, represents the strongest and most correlated

response across subjects: the sensory response. A second de-noised EEG dataset was calculated based on the remaining N_{SC}-1 components, which were expected to include the residual sensory response but, importantly, to encode the neural prediction signal. A time-locked average analysis was conducted on the two resulting signals, allowing us to derive an average response for notes and silent events for each of the signals (first component and the combination of the remaining N_{SC}-1 components) and for each condition (listening and imagery). Baseline correction was not applied for the time-locked average, as the MCCA procedure should have substantially reduced subject-specific noise (e.g., temporal drifts). As such, we were interested in assessing the exact average signals corresponding with notes and silent events, including possible non-zero activity before the event. This also allowed us to more clearly visualize the potential impact of previous notes on the average signal corresponding with notes or silent events. Differently from the TRF analysis, the time intervals corresponding to the metronome response were included in the MCCA procedure, allowing us to extract components related to the corresponding sensory response.

This analysis was conducted on EEG data that was filtered between 1 and 30 Hz. We also run the procedure by including frequencies down to 0.1 Hz. However, the separation between sensory and prediction components was not as clear-cut as in the 1-30 Hz case, as the sensory response contributed to the first several components.

2.3.2.8 Statistical analyses

Consistent statistical procedures were applied to the datasets from the two experiments.

Linear mixed model analyses were performed when testing for significant effects in the case of multiple factors over multiple groups. This statistical test was conducted when studying the TRF results in **Figures 3** and the MCCA results in **Figure 4**, to assess the effects of event-type (notes and silent-events) and condition (listening and imagery).

Pair-wise comparisons were assessed via the (non-parametric) Wilcoxon signed-rank test. Correction for multiple comparisons was applied where necessary via the false discovery rate (FDR) approach. In that case, the *q*-value is reported i.e. FDR adjuster *p*-value. This FDR-corrected Wilcoxon analysis was used when testing the TRF weights for significance in Experiment 1 by comparing each data-point of the TRF global field power with a baseline at latency zero. The same FDR-corrected analysis was also run when conducting a posthoc analysis on the TRF weights in Experiment 2 in **Figure 3**, again with a baseline at latency zero, and in the lateralization analysis in **Figure 5**.

A permutation procedure was used to test for a significant neural encoding of melodic expectations in Experiment 2 (**Figure 5**). That procedure consisted of re-running 100 times per participant the forward TRF procedure, each time after random shuffling of the expectation val-

ues, while preserving the timing information (see the **Temporal response function analysis** subsection for further details on the shuffling procedure). A null-distribution of the mean EEG prediction correlation across participants was estimated with bootstrap resampling to assess whether melodic expectations improved the EEG prediction correlations. The null-distribution was composed of N=10000 data-points, each derived by: selecting a random data-point per subject among the 100 shuffles; averaging the corresponding EEG prediction correlations across participants; repeating the procedure 10000 times. The uncorrected *p*-values are reported in this case, as several *p*-values were smaller than the sensitivity of the test ($p < 10^{-4}$). The 100 data-points per participant were also used to estimate a null-distribution to assess the significance for individual participants. Note that both the group- and individual-subject-level analyses were conducted after averaging the EEG prediction correlations across all electrodes.

2.3.3 Results

Neural data were recorded from participants as they alternately performed a music *listening* (Experiments 1 and 2) and a music *imagery* tasks (Experiment 2) based on monophonic piano melodies from Bach. IDyOMpy (c.f. 3.2) was used to identify *silent-events* i.e. silent instants where a note could have plausibly occurred (see **Methods**). Our analyses aimed at testing the hypothesis that an endogenous prediction signal emerges in correspondence with silent events. We parametrized the onset times of *notes* and *silent-events* in univariate vectors (NT and SIL respectively) and related them with the neural data by means of three distinct analysis procedures. In the listening task, the sensory response (S), which was present in NT but not SIL, was anticipated to account for most of the variance of the EEG responses to melodies. The residual non-sensory response was instead hypothesized to reflect top-down neural prediction signals (P). According to the predictive processing framework, P was expected to be measured in combination with the sensory response in correspondence with notes in the listening condition (S-P) and in isolation in correspondence with notes in the imagery condition and silent events in both conditions (S=0, -P).

2.3.3.1 Experiment 1: Robust cortical response to silence during music listening

In the first EEG experiment, twenty healthy participants were instructed to listen to ten monophonic piano excerpts from Bach's sonatas and partitas, each repeated three times and played in random order. The cortical responses to music were assessed by means of a multivariate linear regression framework known as the temporal response function (TRF), which takes into account the interactions and overlap between a succession of notes. Given a property of

interest of a sensory stimulus encoded in a time vector, the TRF estimates an optimal linear transformation of those vectors that minimizes the EEG prediction error (**Fig. 2A**). The TRF weights can then be interpreted to assess the spatiotemporal dynamics of the underlying neural system.

First, the cortical response to music *notes* was assessed by calculating the TRF between a time-vector marking note-onsets with value 1 (NT) and the corresponding EEG signal (1-30 Hz). Then, the same procedure was repeated when considering the onset-time of *silent-events* (SIL; **Fig. 2A**). The global field power of TRF_{NT} indicated that three components centered at about 80, 200, and 400 ms were significantly larger than the baseline at lag 0 ms of sound-EEG latency (FDR-corrected Wilcoxon tests, $q<0.001$; NT: significant effects for the time-latencies in the windows 62.5-250ms and 297-516ms). Instead, only two significant components emerged for TRF_{SIL} at 200 and 400 ms. The regression weights for TRF_{NT} and TRF_{SIL} are shown in **Figure 2B** for a representative electrode (FCz), with the corresponding topographies for all electrodes. Interestingly, *note* and *silent-event* responses showed inverse polarity, showing a large negative correlation between the two curves (r = -0.946, $p = 4.0*10^{-30}$) and leading to significant differences for the three components at 80, 200, and 400 ms (FDR-corrected Wilcoxon tests, q<0.001; SIL: significant effects for time-latencies in the windows 125-282ms and 359-484ms).

The large difference between the neural responses to *notes* and *silent-events* is likely due to the absence of auditory stimulation for music silence. As such, TRF_{SIL} was expected to reflect the effect of top-down predictions, which could include the prediction signal itself as well as the update of internal priors on the upcoming music event after detecting a silence. Indeed, the present design comes with a potential confound: TRF_{SIL} could be capturing an average late response to a previous note. While this risk is minimized by our choice of 10 music stimuli with various tempi, the majority of silent events occurred less than 300ms after a note (**Figure 2C,D**). To assess the likely interaction between silent events and the preceding notes, we conducted a multivariate forward TRF analysis where both the NT and SIL regressors were used to predict the EEG signal simultaneously. In this context, the NT vector could be seen as a nuisance regressor when studying TRF_{SIL} and vice versa. The result of this analysis (**Figure 2E**) indicates that the inclusion of NT as a nuisance regressor does not change the main TRF result (polarity inversion between TRF_{NT} and TRF_{SIL}, with a large negative correlation between the two curves: r -0.616, $p = 1.6*10^{-7}$). Furthermore, the dynamics of TRF_{SIL} did not change compared to the univariate analysis (Pearson's $r = 0.95$, $p = 3.3*10^{-31}$ between the TRF_{SIL} curves in the univariate and multivariate TRF analyses), despite a reduction in power over time-latencies (Wilcoxon tests, $p = 3.3*10^{-11}$), which reflects the expected smaller magnitude silent-event neural signals compared with evoked-responses to notes, an effect that is magnified in the multivariate model as the two neural responses are assessed simultaneously.

We designed a second experiment aiming at overcoming the limitations of Experiment 1. The following section describes Experiment 2, whose novel design based on musical imagery enables the isolation of endogenous neural signatures of both *notes* and *silent-events*.

2.3.3.2 Experiment 2: Cortical encoding of music silences during listening and imagery tasks

A second EEG experiment was conducted on twenty-one expert musicians. In the *listening condition* (**Fig. 3A**), participants were presented with four ~35 s piano melodies from Bach chorales. In the *imagery condition* (**Fig. 3B**), participants were instructed to imagine hearing the same melodies. Each piece was presented and imagined 11 times, for a total of 88 trials with random order. A vibrotactile metronome placed on the left ankle marked the beginning of 100 bpm measures (every 2.4 s) in both conditions, allowing participants to perform the auditory imagery task with high temporal precision. Therefore, the neural signal was expected to reflect sensory responses to auditory and tactile sensory inputs for the listening condition and to the tactile input only for imagery. Neural data within 500 ms from the metronome input were excluded from the TRF analyses that follow to ensure that the results do not reflect tactile responses.

First, we replicated the result from Experiment 1 by running a forward TRF analysis on the EEG data (1-30 Hz) for the listening condition. The TRF weights showed spatiotemporal dynamics consistent with the previous result, with inverse polarities for NT and SIL (**Fig. 3C**). Then, we run the same TRF procedure on the auditory imagery condition. While the investigation was conducted in a manner consistent with the previous experiment, the analyses largely focused on the EEG channel FCz, where the main effect (a polarity inversion in the listening condition) was expected based on the results from Experiment 1. A linear mixed model analysis indicated significant effects of condition (listening vs. imagery) and regressor (notes vs. silent-events) on the TRF weights, with a significant interaction effect between condition and regressor (the dependent variable was the average magnitude of the TRF component at FCz for the latencies 250 ± 100 ms, condition and regressor were the independent variables and subjects a random intercept; effect of 'regressor': *estimate* = -2538, *tStat* = -12.4, $p = 2.4*10^{-20}$; effect of 'condition': *estimate* = -2746, t = -11.4, $p = 1.7*10^{-18}$; interaction effect: *estimate* = 1307, t = 10.1, $p = 6.3*10^{-16}$). Post-hoc analysis on the individual TRFs indicated components larger than the baseline at latency zero for all conditions (FDR-corrected Wilcoxon tests, $q<0.01$; see **Methods**). **Figure 3C** shows significant TRF components at FCz in the listening condition. TRF traces for notes and silent-events were negatively correlated (r_{NT_LIST,SIL_LIST} = -0.60, $p = 7.0*10^{-5}$), thus replicating the result from Experiment 1. The result in **Figure 3D** indicates, as we showed in part 1 of this study (Marion et al., 2021), robust neural correlates of

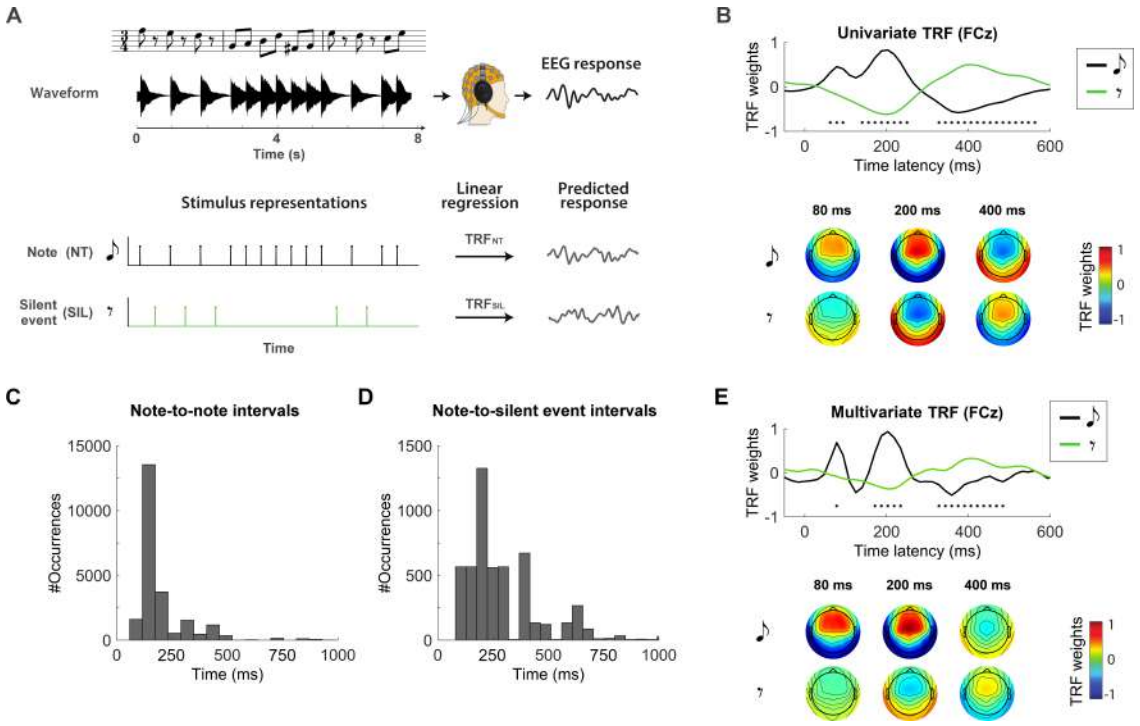


Figure 2.12. **Figure 2. Robust cortical response to silence during music listening.** (A) Experiment 1 setup. EEG signal was recorded as participants listened to monophonic piano music. Univariate vectors were defined that mark with value 1 the onset of either notes (**NT**) or silent events (**SIL**). A system identification procedure based on lagged linear regression was performed between each vector and the neural signal that minimizes the EEG prediction error. (B) The regression weights represent the temporal response function (TRF) describing the coupling of the EEG signal with notes (TRF_{NT}) and silent events (TRF_{SIL}). TRFs at the representative channel FCz are shown (top), revealing significant differences (FDR corrected Wilcoxon test, $*q < 0.001$) between the neural signature of note and silent-event due to inverted polarities, as clarified by the topographies of the TRF components (bottom). (C,D) The overall distribution of time-intervals between notes and between silent-event and the immediately preceding note. The y-axis indicates the number of occurrences for a given bin of time intervals when considering all trials. The data shows that a large number of silent events occurred less than 200ms after a note, implying that, in experiment 1, TRF_{SIL} could have potentially been affected by the late response to the previous note. (E) The analysis from panel B was re-run by using multivariate TRF models i.e., considering note and silent-event vectors simultaneously with multivariate lagged regression to account for possible interaction between the two. The figure shows the regression weights corresponding to the two regressors at the selected channel FCz, while the topographies show the regression weights. As for the univariate TRF result, significant differences were found between note and silent-event TRFs (FDR corrected Wilcoxon test, $*q < 0.001$). TRF_{NT} showed qualitatively more pronounced early TRF components.

auditory imagery in correspondence of notes. Crucially, the EEG dynamics of auditory imagery corresponding to *silent-events* showed shape and latencies comparable to those measured for imagery of *notes* ($r_{NT\_IMAG,SIL\_IMAG} = 0.89$, $p < 1.2*10^{-13}$), with the same polarity measured for *silent-events* in the listening condition ($r_{SIL\_LIST,SIL\_IMAG} = 0.57$, $p = 2.1*10^{-4}$, $r_{SIL\_LIST,NT\_IMAG} = 0.52$, $p = 7.4*10^{-4}$). Conversely, TRF_{NT} in the listening condition had inverse polarity, which was consistent with the polarity of tactile responses i.e. the only other sensory response in the EEG data (see the TRF result for the metronome-only vector in **Figure 3G**).

The result in **Figure 3D** indicates that the inverted cortical polarity measured for TRF_{NT} and TRF_{SIL} during music listening (**Figs. 2B** and **3C**) depends on the presence or absence of a sensory stimulus respectively, rather than a different encoding of *notes* and *silent-events per se*. In fact, that difference was not present in the imagery condition, where there was no auditory stimulation. This result is in line with a predictive processing account of auditory perception whereby the brain constantly attempts to predict sensory signals (**Fig. 1**). The analyses that follow aim to provide further support to this result by 1) disentangling sensory and prediction signals in both the listening and imagery conditions with a methodology that, differently from the TRF, does not use explicit knowledge of the position of notes and silent events; and by 2) assessing whether the prediction signal encodes the precise melodic expectation values as estimated by a computational model of musical structure.

2.3.3.3 Disentangling neural sensory responses and neural prediction signal

The TRF analysis showed robust *note* and *silent-event* encoding in both listening and imagery conditions. However that analysis is oblivious to the differences between responses to individual events. For example, neural responses change with the listener's expectation of a *note* based on the proximal music context (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Omigie *et al.*, 2013b). The next two sub-sections investigate the neural signature of individual music events across the time domain of a musical piece.

Investigating the cortical processing of individual music notes requires an approach that is effective despite the low SNR of EEG recordings. The TRF procedure in the previous sections summarizes information across the time domain, providing a summary neural trace for each participant representing the "typical" response to a note or a silent event. However, that approach does not provide us with a view at the level of individual events (notes and silences). To do so, we used multivariate canonical correlation analysis (MCCA), an approach that de-noises the EEG data by preserving components of the signal that are consistent across participants.

An MCCA analysis was run on EEG data from all participants simultaneously, preserving the first N_{SC} summary components (SC) with the largest inter-subject correlation (see **Methods**). This approach enables the investigation of neural data in the original EEG channel with

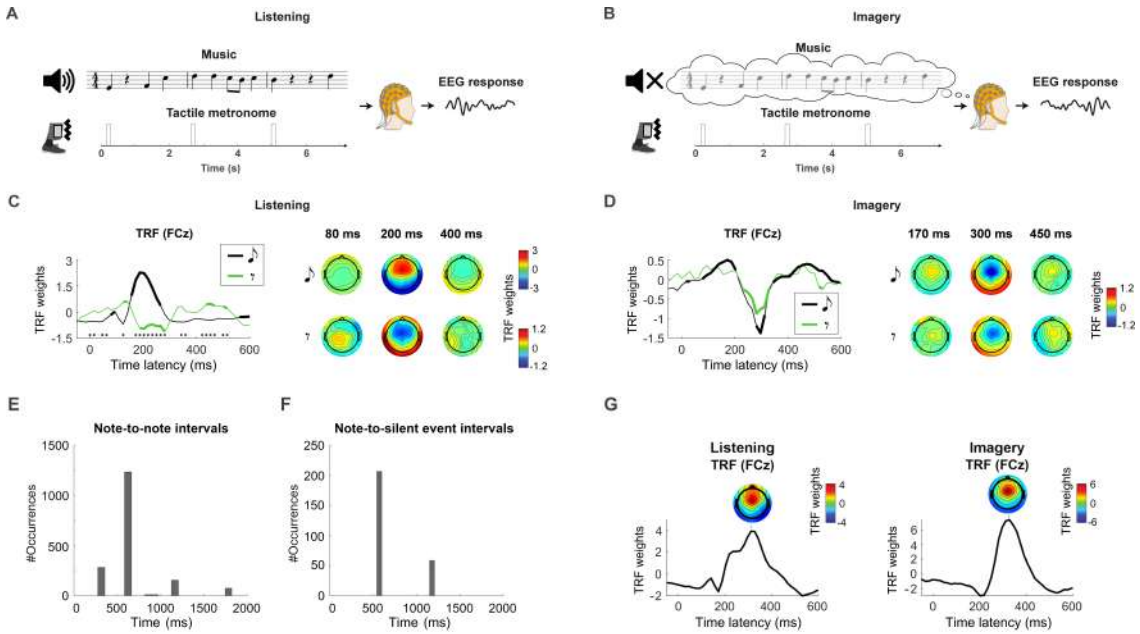


Figure 2.13. **Figure 3. Comparable cortical encoding of music silence and note during imagery.** (A,B) EEG signal were recorded as participants listened to and imagined piano melodies (Experiment 2). A vibrotactile metronome placed on the left ankle allowed for the precise execution of the auditory imagery task. (C) TRFs at the channel FCz (left) and topographies of the TRF at selected time-latencies (right) are reported for the listening condition. Thick lines indicate TRF weights that are larger than the baseline at latency zero (FDR corrected Wilcoxon sign-rank test, $q < 0.01$). Black asterisks indicate significant differences between NT and SIL (FDR corrected Wilcoxon sign-rank test, $q < 0.01$). (D) The TRF results is reported for the imagery condition, showing a significant component centered at ∼300 ms for both note and silent events with, as hypothesized, no significant difference between NT and SIL, which had the same polarity in this case. (E,F) The overall distribution of time intervals between notes and between silent events and the immediately preceding note in Experiment 2. The y-axis indicates the number of occurrences for a given bin of time intervals when considering all trials. (G) TRFs were fit for the listening and imagery conditions using a univariate stimulus regressor marking the metronome with unit impulses (and zero at all other time points). TRFs are shown at the EEG channel FCz. Topographies depicting the TRF weights at all channels are also shown at the peak of the dominant TRF component.

remarkably high SNR, allowing us to assess the neural signature of each individual event in a melody. SC_1 was expected to capture the sensory response, which is likely the strongest and most consistent signal across participants. As we had hypothesized, SC_1 showed strong neural activation corresponding to sensory events i.e. *notes* and *metronome* in the listening condition; and *metronome* only in the imagery condition (hypothesis in **Fig. 1** and result in **Fig. 4**). That result was visible both at the level of individual music events (**Fig. 4A,B**) and on the time-locked average signals (**Fig. 4C,D**). Next, the first sensory component (SC_1) was removed from the EEG data to study the residual N_{SC}–1 component, which was expected to capture neural predictions and, therefore, to be active in correspondence with both *notes* and *silent-events,* as depicted in **Figure 1**. The result in **Figure 4** confirms that hypothesis by showing neural activation for all music events, with negative components corresponding to both *notes* and *silent-events* between about 200 and 400 ms in the imagery conditions. A linear mixed model analysis confirmed such observations: Significant effects of condition (listening vs. imagery) and event-type (notes vs. silent-events) were measured on the time-locked averages for SC_1, with a significant interaction effect between condition and event-type (the dependent variable was the average magnitude of the time-locked average component at FCz for the latencies 250 ± 100 ms, condition and event-type were the independent variables and subjects a random intercept; effect of 'event-type': *estimate* = -6.1, *tStat* = -6.6, $p = 4.5*10^{-9}$; effect of 'condition': *estimate* = -4.3, t = -4.7, $p = 9.5*10^{-6}$; interaction effect: *estimate* = 3.0, t = 5.1, $p = 2.1*10^{-6}$). This result is in line with the interpretation of the first component as a sensory response signal, which is present only for notes in the listening condition. Significant effects were also measured on the residual N_{SC}–1 component for condition but not event-type nor the interaction between the two (effect of 'event-type': *estimate* = 3.6, *tStat* = 1.9, $p = 0.058$; effect of 'condition': *estimate* = 4.6, $t = 2.5$, $p = 0.016$; interaction effect: *estimate* = -1.9, t = -1.6, $p = 0.11$), which is in line with the interpretation of the residual N_{SC}–1 component as a prediction signal. Overall, this result is consistent with our initial hypothesis in **Figure 1**.

2.3.3.4 Cortical encoding of silence expectations during music listening and imagery

Recent studies indicated that low-frequency neural signals encode melodic expectations when participants listen to monophonic music (Di Liberto et al., 2020; Omigie et al., 2013). Specifically, melodic expectations modulate the magnitude of the auditory responses, with larger neural responses for less expected events. In line with those results and with the hypothesis that cortical responses to music reflect a combination of sensory and prediction signals (**Fig. 1**), we anticipated EEG responses to notes and silent events to be modulated by melodic expectations during both listening and imagery conditions. To test that, we first estimated the

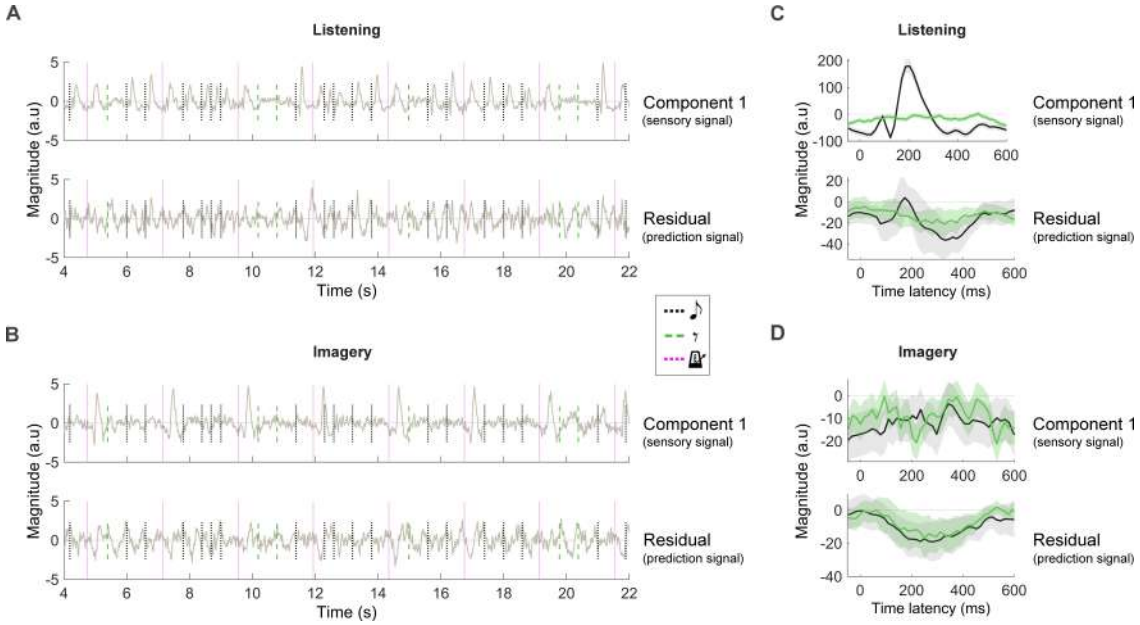


Figure 2.14. **Figure 4. Disentangling sensory and prediction neural signals with unsupervised correlation analysis.** Multiway canonical correlation analysis (MCCA) was used on all EEG data to identify components of the EEG signal that were consistent across subjects. N_{SC} summary components (SC) with the largest inter-subject correlation were preserved. The first SC represents the EEG response that is most correlated signal across subjects. Here, we hypothesized the first SC and the residual N_{SC}-1 SCs to capture *sensory* and *prediction* cortical signals respectively. (A,B) The first SC (top) and to the residual N_{SC}-1 SCs (bottom) were back-projected onto each participant's EEG channel space for each condition. The average signals at the EEG channel FCz were shown for a selected portion of "Melody 4" (brown lines). Vertical lines mark music events: notes (black dotted lines); silent events (green dashed lines); and vibrotactile metronome onset (purple dotted lines). Note that sensory responses could exist only for note and metronome in the listening condition and for metronome only in the imagery condition. (C,D) First SC (top) and the residual N_{SC}-1 SCs (bottom) at the EEG channel FCz after time-locked averaging to *note* and *silent-event* onsets. Shaded areas indicate the 95% confidence interval calculated across participants.

expectation of a note with IDyOM (M. T. Pearce, 2005), the model of melodic structure based on variable-order Markov chains that were also used to identify the silent events. Expectation values were calculated from the music score based on both a long-term model of Western music and short-term proximal information on the current piece. As a result, IDyOM provided us with measures of surprise and Shannon entropy of the onset-time of each upcoming note and silent event. Surprise informs us how unexpected was a note (or a silent event) at a given time-point while entropy indicates the uncertainty at a particular position in a melody before the musical note is observed. Each of these features was encoded into time series by using their values to modulate the amplitude of note and silent-event onset vectors. This resulted in four time-series: surprise for notes (S_{NT}) and silent-events (S_{SIL}), entropy for *notes* (H_{NT}) and *silent-events* (H_{SIL}). We then called EXP_{NT} and EXP_{SIL} the concatenation of the surprise, entropy, and onset vectors for notes and silent events respectively. Note that EXP_{NT} and EXP_{SIL} were calculated by using timing but not pitch information, as silent events do not have a pitch value.

Forward TRF models were fit to assess the coupling between low-frequency EEG (0.1–30 Hz) and the onset-time expectation vectors. Shuffled versions of the expectation vectors (N = 100 per subject), with surprise and entropy values randomly permuted while preserving the temporal information of the event onsets, were used as a baseline for the assessment of the expectation-EEG encoding. Both EXP_{NT} and EXP_{SIL} could predict the EEG better than their shuffled versions in both the listening and imagery conditions (EEG prediction correlation was averaged across all EEG channels; the mean across subjects was compared with a bootstrap resampling distribution of the mean across subjects derived from the shuffled data; N = 10000; $p < 10^{-4}$ for notes and silent-events in both conditions; see **Methods**; **Fig. 5A**). A significant EEG encoding of expectations was also measured at the individual subject level, with 12/21 and 10/21 subjects above chance level in the listening condition for note and silent-event TRFs respectively, and 10/21 and 17/21 subjects above chance level in the imagery condition (one-tailed permutation test, N = 100 permutations per subject per condition, $q < 0.05$, FDR-correction). While this effect of expectation was assessed on the average of all EEG channels, **Figure 5B** shows the topographical distribution of that effect (the contrast between EEG prediction accuracies for expectation and the 95th percentile of the shuffles). Similar but weaker effects were measured for EEG filtered between 1 and 30 Hz for all conditions but silent events in the imagined condition (EEG prediction correlation values were averaged across all channels; the mean value across subjects was compared with a bootstrap resampling distribution of the mean across subjects derived from the shuffled data; N = 10000; NT, listening: $p < 10^{-4}$; SIL, listening: $p = 0.021$; NT, imagery: $p = 0.009$; SIL, imagery: $p = 0.541$).

These results indicate a fine-grained encoding of melodic expectations in the cortical signals corresponding to music listening and imagery. We also tested whether the effect of onset-time expectations on the EEG prediction increase showed significant lateralization. We found a

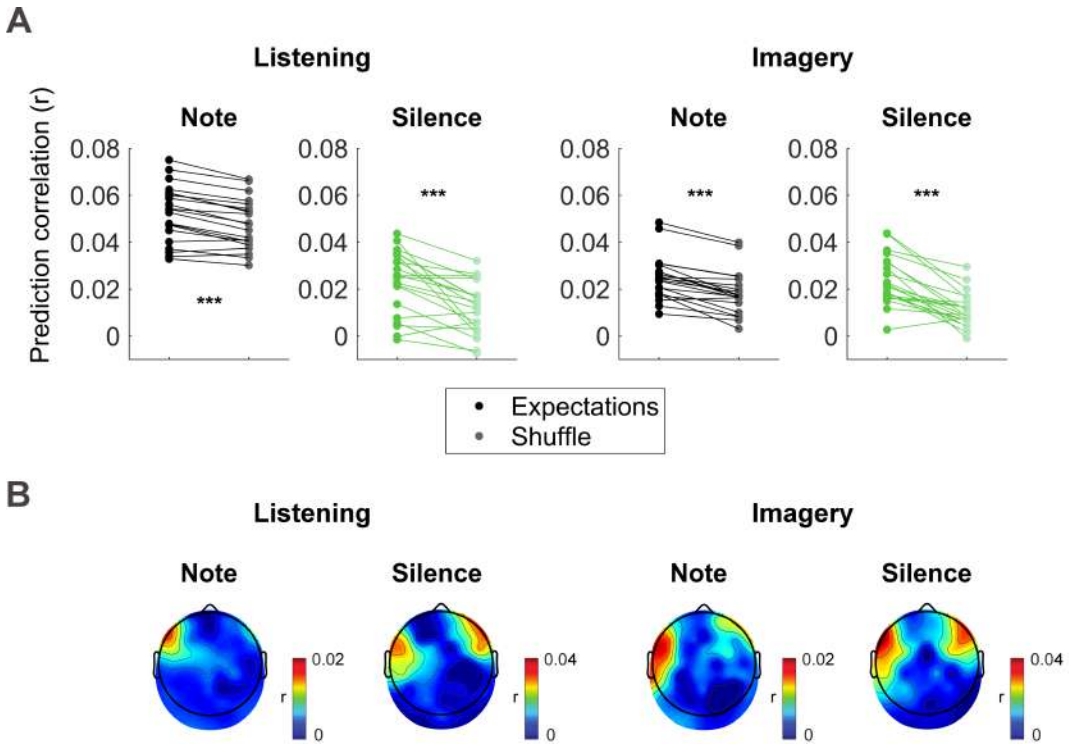


Figure 2.15. **Figure 5. Notes and silence expectation encoding in low-frequency EEG.** A multivariate TRF analysis was conducted to identify the linear transformation that best predicts low-frequency EEG data (0.1-30 Hz) based on a three-dimensional stimulus representation indicating, for either note or silent-events: event onset-time, entropy at that position, and surprise of that event. (A) EEG prediction correlations of the TRF using the note or silence expectation values estimated with IDyOM are compared to a null-model where the EEG prediction correlations were obtained with a TRF that was fit after a random shuffling of the expectation values (event onset-times were preserved). Results averaged across all electrodes are reported for both listening and imagery conditions. Each dot indicates the result for a single subject. Significant differences were measured for notes and silent events in both conditions (Permutation test, $***p < 10^{-4}$). (B) Topographical maps indicating the EEG prediction correlation increase (expectation minus null-model) at each EEG channel.

weak left-lateralization effect for notes in the listening condition that, however, did not survive correction for multiple comparisons (FDR-corrected Wilcoxon test, $q = 0.100$).

2.3.4 Discussion

Predictive processing explains rhythmic and melodic perception as a continuous attempt of our brain to anticipate the timing and identity of upcoming music events. While previous research investigated such predictive mechanisms indirectly by measuring how expectation modulates sensory responses, this study used neural measurements of music processing in the absence of sensory input. In particular, we combined, for the first time, low-frequency EEG measurements corresponding to silent music events during music listening with EEG signals

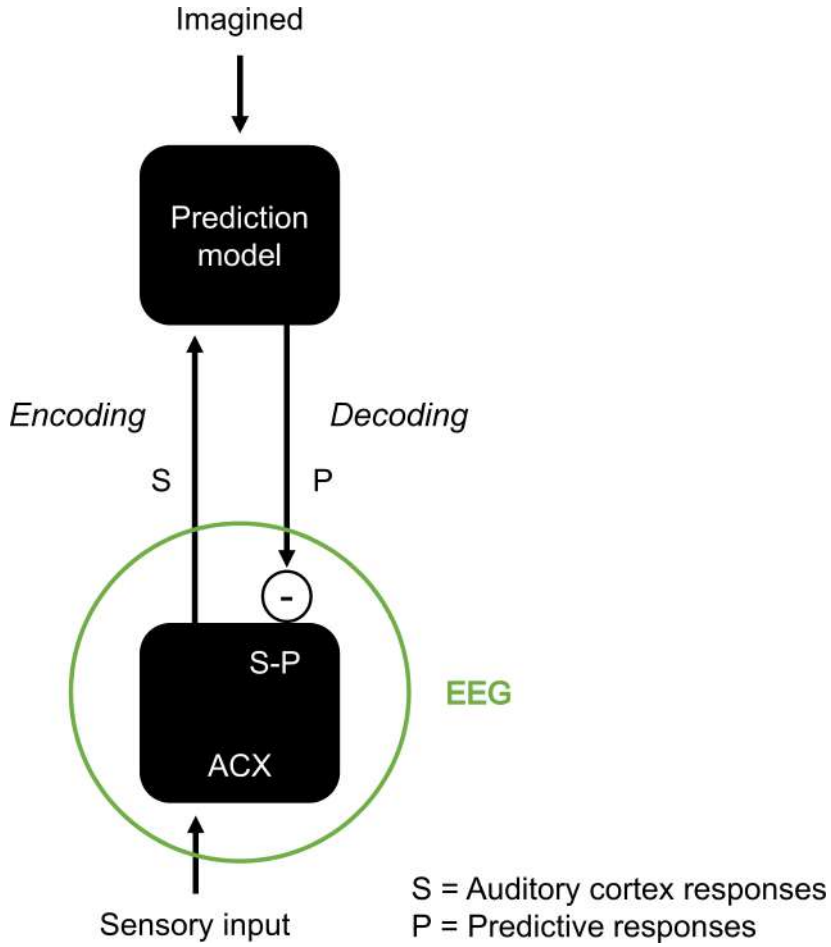


Figure 2.16. **Figure 6.** Computational model for how predictions influence neural signals corresponding to auditory listening and imagery. Auditory inputs elicit bottom-up sensory responses (S) through the auditory cortex (ACX). A prediction model generates a top-down prediction signal (P) that is more similar to S for more predictable sensory events. That prediction is compared with S, producing an error signal $_{sur}$ S-P. The EEG response is hypothesized to capture a combination of $_{sur}$ and S itself, meaning that some level of EEG activation is expected even when S is fully predictable (Margulis, 2014). When a sound is imagined, S = 0 and therefore $_{sur}$ -P, as for our hypothesis in Figure 1.

recorded during musical imagery, both in the context of natural melodies. In doing so, we could demonstrate that robust neural activity consistent with prediction error signals emerges during the meaningful silence of music and that such neural activity is modulated by the strength of the music expectations. We propose a unifying perspective on auditory predictions, where endogenous auditory predictions have a central role in music perception both during listening and imagery.

EEG encoding of music-silence reveals neural auditory predictions

Existing computational models of music structure, such as IDyOM, generally consider silences as "time intervals without notes" (SOA or IOI) (M. T. Pearce, 2005). However, melodies contain silent instants where a note could have plausibly occurred. The present study demon-

strates that the human brain encodes these music *silent-events,* suggesting that the physiological validity of those models can be augmented via an explicit account of silent events, rather than just an implicit encoding of that same information as in IDyOM. The finding that musical imagery elicits robust neural activity (Marion *et al.*, 2021; Tibo *et al.*, 2020) laid the foundations of the present study, providing us with an experiment that allows to discern endogenous neural processes from exogenous auditory perception mechanisms. Our results are summarized by the model in **Figure 6,** in line with the predictive processing framework. The neural encoding of sound and silence corresponds to S-P and, as such, to -P when there is no stimulus (S=0) i.e. silent events during listening and imagery, and notes during imagery. This result emerged both when using forward TRFs (Crosse, Di Liberto, Bednar, & Lalor, 2016) and MCCA (de Cheveigné *et al.*, 2019), two methods with different assumptions and rationale. One crucial difference between the two is that TRFs describe the neural responses for an individual subject with an impulse response, one for each stimulus feature-set (notes vs silence). Instead, MCCA does not require any explicit knowledge of the timing and identity of notes and silent events. This unsupervised approach combined the data from multiple subjects to extract neural components that were sensitive to individual events (note or silence) within a continuous music piece, with a remarkable signal-to-noise ratio. Crucially, the two methods converged to consistent results, revealing that silent events correspond to robust neural responses and that similar neural signals emerge during imagined notes and silent events. This internally generated "music of silence" is in line with a continuous attempt of the brain to predict upcoming plausible notes. Altogether, this study provides direct evidence for the duality of sensory and prediction signals posited by the predictive processing framework. Our results suggest that both listening and auditory imagery entail the transformation of external or imagined sounds by an internal "predictive model" that encodes our conceptions and expectations of the sound, which is then compared to the sensory stimulus - if present. The finding, which is captured by the model in **Figure 6**, is in line with previous fMRI and PET results on auditory imagery (A. R. Halpern & Zatorre, 1999; Kraemer *et al.*, 2005; Meister *et al.*, 2004; Zhang *et al.*, 2017). In fact, such studies showed robust neural activation in correspondence with auditory imagery, as we measured here with EEG. Crucially, our results linked the neural activation for auditory imagery to a general predictive mechanism that applies to both listening and imagery. Specifically, the model in **Figure 6** explains both imagery and silence activations as the result of the subtraction of sensory responses and prediction signals, leading to a change in response polarity when the sensory input is absent. Indeed, further work with multiple technologies (e.g., fMRI, EEG, ECoG) is needed to conclusively link our findings with studies based on hemodynamic measurements and test our model. One challenge is to clarify what exactly each neural measurement can capture within that model. EEG recordings provide us with macroscopic measurements that are likely to include a variety of neural components. While the evidence points to a strong

sensitivity to prediction errors (or surprise), there may be additional components that encode S and P separately.

While the music of silence allows to clearly separate the neural prediction signal from sensory responses, technologies with higher spatial resolution may be able to uncover more precise details on the neural implementation of this predictive mechanism. One unsolved question regards the possibility that prediction processes could be at the core of the ability to perform auditory imagery. Based on our finding that prediction processes explain a significant portion of EEG variance during auditory imagery and that imagined notes and silent events corresponded to similar neural activation, one possibility is that auditory imagery may rely on the same endogenous prediction mechanisms that are engaged during auditory listening, rather than involving a separate imagery process. Finally, additional research should be conducted to investigate possible links between our model and beat perception. In fact, the present design was optimized for the imagery task, thus working with relatively simple melodies. Experiments with a broader set of music stimuli are needed to tackle that question, for example by using syncopated music stimuli, which would allow for a more distinct separation of beat and notes (Tal *et al.*, 2017).

Silence neural signals are graded by expectations

The TRF analysis in **Figure 5** confirmed the hypothesis that low-frequency EEG responses to naturalistic music encode melodic expectations in correspondence of prospectively predictable silent events. The responses to silent events were shown to co-vary with the expectation strengths, which were drawn from a note onset-time statistical model (M. T. Pearce, 2005), as it was previously shown for music notes (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Omigie *et al.*, 2013b). These results are in line and go beyond previous measurements of the neural responses to sensory omissions, which focused on scenarios where strong expectations on the upcoming occurrence of a stimulus were built artificially (missing stimulus potentials - MSP; (Bendixen *et al.*, 2009)). Mismatch negativity responses (MMN) to omitted tones were measured for stimulus onset asynchronies (SOA) up to 150 ms (Yabe *et al.*, 1997), while studies with longer asynchronies, closer to those of the present study, were shown to elicit MSPs with a modality-specific (auditory) negativity at about 230ms and a modality independent (both auditory and visual) positivity at 465ms (Joutsiniemi & Hari, 1989; Simson *et al.*, 1976). Silent events in melodies differ from omissions in that they have a much lower probability of corresponding to a sound. Furthermore, omission cannot be predicted, while the participants of Experiment 2 were pre-exposed to the four melodies and, as such, silent events were not unexpected *per se*. In other words, the participants were certainly not "surprised" in the traditional sense when they encountered a silent event, as they had heard the melody before. Instead, our results are different from the "unexpectedness" investigated with sensory omission paradigms as they reveal prediction errors related to the processing of melodic structure based on the

melody statistics.

Further work should be conducted to directly explain the overlap and interaction of the two phenomena. Our finding contributes to that question by suggesting a unifying view linking MSP (omission response), expectation modulation of sensory responses (EM), and auditory imagery using naturalistic music listening. Our results suggest that the MSP negativity and EM are the results of the same prediction process. In addition to providing new direct evidence on the neural substrate of MSPs, we show that such responses can be measured when the music is internally generated (imagery). This result is in line with a view of the auditory system where predictions are simultaneously computed at multiple time scales (e.g. hierarchical predictive coding) and, crucially, where local expectations (at short time scales) are performed by our brain even in the presence of exact prior knowledge of the upcoming stimulus (e.g. repetition of a song; production or imagination of a song). In fact, the TRF analysis in **Figure 5** indicates a robust encoding of melodic expectations even though the stimuli were precisely known by the participants (only four repeated stimuli were presented and participants were exposed to the pieces before the start of the EEG experiment).

We contend that the present finding has implications for computational models of sensory perception. For example, neural signals have been modeled by focusing on evoked responses (e.g., (Doelling *et al.*, 2019; Ferezou & Deneux, 2017))), thus describing the neural signal as a sum of fixed-latency sensory evoked responses, while generally ignoring prediction processes. Instead, as highlighted by this study, prediction signals emerge in correspondence with both notes and silent events in the neural signal. As in **Figure 6**, evoked-response models could be extended by including such prediction mechanisms both in the presence and absence of a sensory event. The resulting model would describe the S and P duality and would explain the neural responses to music silence that were measured in the present study. We conclude that our brain considers silent events as temporally-precise and information-rich events that provide our brain with valuable information (namely that "a note was not present at a particular plausible time-point") contributing to the subsequent predictions. Our results may reflect a general property of sensory perception and, as such, we expect similar neural responses to emerge during meaningful silences in other auditory stimuli such as speech. Specifically, "expectation" signals similar to the predictive melodic expectations in music sequences, have been demonstrated in the neural responses to phoneme sequences, the fundamental units of speech (Brodbeck *et al.*, 2018; Di Liberto *et al.*, 2019). Therefore, we anticipate that future studies may reveal predictive responses that closely resemble those we identified in music "silences", but that would reflect the linguistic model of the listener, confirming that the findings of the current study are indicative of general auditory perception mechanisms.

In summary, the present study shows robust neural signatures of music silence, suggesting that silent events have great importance in the neural encoding of music. Furthermore, we

provide evidence that the encoding of silent events reflects a neural prediction signal, with results that are in line with the predictive processing framework.

2.4 Cross-Modal Predictions: Sensory Motor Predictions in Speech[5]

2.4.1 Introduction

Sensorimotor interactions have long been postulated as a fundamental ingredient of the performance of complex tasks engaging a perceptual system (visual, auditory, or somatosensory) and a concomitant suite of motor actions (reaching, speaking, lifting) (Georg *et al.*, 2012; Wolpert & Ghahramani, 2000). The conceptual motivations are anchored in control theory where rapid complex actions can benefit from fast sensory feedback to inform the controllers of the accuracy of the ongoing performance so as to maintain or correct its course (Daniel *et al.*, 1995; Roger & W, 1970; Wirthlin *et al.*, 2019). The same rationale and motivations also apply in purely sensory contexts where the balance between bottom-up stimulus representations and top-down predictions are postulated to play a key role in stimulus perception (Keller & Mrsic-Flogel, 2018).

The feedback may take the form of deviations (errors) between the sensory consequences of ideal target performance and its "prediction", which is computed by extrapolating a "*forward*" model of the motor plant. This is how accurate arm reaching is informed by visual and proprioceptive cues (S & M, 1997), and how the vocal tract exhibits smooth delivery and executes rapid corrections of speech from auditory feedback (G, 2012; John & Edward, 2015; Wirthlin *et al.*, 2019). This predictive function of sensorimotor interactions has even been postulated to apply in reverse, to explain how robust sensory perception can arise from observing motor action, e.g. the role of lip-reading in speech comprehension, or in the *Motor Theory of Speech* where acoustic features of speech are presumed to be transformed and encoded as articulatory commands (Alvin *et al.*, 1967; Dominic & Trevor, 2008) (L. Andrew *et al.*, 2009). Finally, these bidirectional sensorimotor interactions achieve their full generalization in the findings of the mirror-neuron responses (Anat *et al.*, 2018), which have claimed a causal role not only in all sensorimotor systems but also in accounts of social function and emotional relations (Marco, 2009). Predictably, these claims have provoked numerous detractions and debates that have served to enrich and deepen the understanding of these phenomena.

In order to characterize sensorimotor interactions in the human cortical speech system, we recorded and analyzed the sensorimotor neural interactions with ECoG in humans while they

[5]Authors: Shihab Shamma, Prachi Patel, Shoutik Mukherjee, Guilhem Marion, Bahar Khalighinejad, Cong Han, Jose Herrero, Stephan Bickel, Ashesh Mehta, Nima Mesgarani(Shamma *et al.*, 2020)

spoke, listened, or simulated speaking by moving their vocal tract without producing sound. The goal was to characterize more accurately the nature of the spectral or temporal representation of the auditory and motor cortical responses. We also used these responses to re-examine the basic computational architecture of the sensorimotor interactions with the aim of clarifying their functional role in action and perception. **Figure 2.17.A** illustrates the basic reciprocal sensorimotor projections as would typically be involved in speech production (John & Edward, 2015; Poeppel, 2014). Specifically, during speaking, motor areas control vocal-tract movements that generate a speech signal. It has also been proposed that the motor cortical areas send a parallel internal neural copy of the speech signal to the auditory cortex – the *forward* prediction signal, where it is compared to the responses induced by the incoming speech (Hickok & Poeppel, 2007). During listening to speech, an *inverse* mapping from the auditory to the motor areas would create a motor representation of the acoustic signals (Stephen *et al.*, 2004).

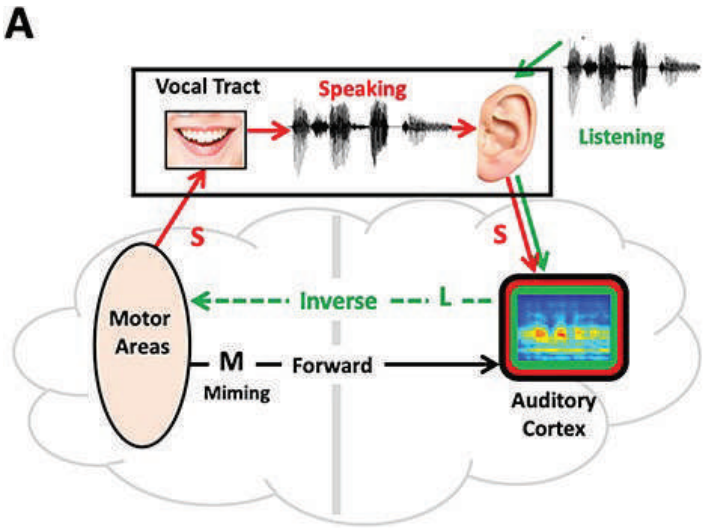


Figure 2.17. Schematic depicts the four types of recordings from all electrodes which are expected in each subject: Miming (M) responses are when a subject articulates the speech without any sound; Listening (L) responses are from the subject listening passively to the speech; Speaking (S) signals are recorded while subject articulates audibly the speech; Noise (N) are recordings of the background noise on the electrodes in silence. The schematic illustrates the postulated forward and inverse projections between the auditory and motor areas.

Because of this bidirectional flow of interactions between the auditory and motor responsive regions (**L** and **M** in **Figure 2.17.A**, we shall refer to this phenomenological network as the *Mirror Network* (or MirrorNet). We shall utilize this framework to explain how ECoG recordings revealed directly the spectrotemporal nature of the MirrorNet projections: the *forward* motor influences into the auditory cortex during silent speaking (or miming), the *inverse* auditory influences into the motor areas during listening, and finally the *bidirectional* influences during speech production. Two previous studies (C. Gregory *et al.*, 2014; Martin *et al.*, 2017) had

adopted experimental paradigms analogous to ours. However, the goals, analyses, and conclusions differ fundamentally from those of this study, although they are mutually consistent as we shall elaborate later.

The findings from our experiments confirmed the basic structure of the auditory-motor mirror network (**Figure 2.17.A**), and revealed that the responses of the *forward* and *inverse* projections are spectrotemporally rich enough to allow for relatively accurate representations of speech. The results also suggest that a key function of the sensorimotor interactions is to enable the brain to *learn* how to use the vocal tract for speech production, rather than simply to control its performance during speaking. In support of this idea, we developed a computational instantiation of this basic network and used it to train a speech synthesizer to produce speech from mere exposure to a corpus of speech data, thus demonstrating how complex actions like speaking or playing piano can be learned through auditory feedback and motor feedforward signals between the two cortical regions.

2.4.2 Results

2.4.2.1 Neural Data

First, we confirmed the functional projections postulated to exist in the network of sensorimotor interactions. Recordings during *silent* miming (**M**) revealed measurable responses in auditory responsive regions, confirming the influence of *forward* projections from the motor areas to the auditory-responsive cortex. During pure listening (**L**) without any motor actions, significant responses were also measured in the motor areas confirming the existence of an *inverse* projection. Finally, responses during speaking (**S**) were found to be, as previously reported, suppressed relative to the **M** and **L** responses in motor and auditory responsive regions, respectively.

Second, detailed analyses of signals carried by the *forward* and *inverse* projections revealed remarkable spectrotemporal specificity, sufficiently adequate to encode individual sentences. Thus, during a skilled task like speech production, we conjecture that these auditory-motor interactions modulate and control auditory and motor responses in detailed and meaningful ways so as to play a role in the learning and performance of the auditory-motor tasks.

In the experiments and analyses reported here, the *forward* and *inverse* activations (**M** in

auditory- and **L** in motor-electrodes) were small because they were measured in the absence of other background responses due to acoustic or motor stimuli. Consequently, to demonstrate the meaningful interpretation of these responses, we had to apply diverse methods, e.g., spectrogram reconstructions, STRF predictions, and correlation rankings, all with varying degrees of confidence. But, in the case of speaking, **S** responses in both auditory- and motor-electrodes are substantial, and they are strongly modulated by inputs projected from the counter regions. This was best demonstrated by the large changes between the various average STRFs, e.g., the changes from **L**-STRF to **S**-STRF to **M**-STRF.

Specifically, STRF changes revealed remarkably different dynamics and patterns of interactions depending on the task, which complement the interpretations gained from the direct response measurements. For instance, when speaking (**S**), relatively strong inhibitory influences are seen in the **S**-STRFs *preceding* at the onset of the responses. This timing seems to coincide with a *preceding* wave of responses on the **M**-STRFs. One possible interpretation of these patterns is that the early **M** activation reflects responses of local recipient inhibitory interneurons and that these in turn exert their inhibitory influences during speaking when the evoked auditory responses are sizable. This interpretation is also consistent with the fact that pure auditory **L** responses (which presumably supply no motor inputs) do not exhibit either of the preceding waves of activation in the **L**-STRFs. On the motor-electrodes, the situation is somewhat different, receiving an inhibitory wave preceding the **L** responses (**L**-STRFs), that roughly coincides with early activation of the **M** responses (**M**-STRF). The **S** responses which combine motor and auditory interactions are complex and less punctate, perhaps reflecting the local interactions between the **M** and **L** sources. All these details remain to be addressed in future analyses that would consider the timing of the interactions (e.g., (Einat & Riikka, 2018; C. Gregory *et al.*, 2014), and especially on individual localizable electrodes.

Third, the high spatiotemporal resolution of the ECoG allowed us to localize sources and destinations for the auditory-motor interactions and to reveal their relative timings. The results on the whole are consistent with findings from global imaging data with fMRI, EEG, and MEG. For instance, we found that the *forward* and *inverse* projections are largely between non-primary auditory responsive regions such as the *STG*, *PT*, versus *MTG, ITG* on the motor-side. Non-primary regions are known to be far more plastic and hence susceptible to the effects of behavioral engagement and learning from experience.

2.4.2.2 Sensorimotor interactions and learning in the Mirror Network

This experimental study was motivated by two aims concerning the nature of the *forward* and *inverse* projections of the conceptual network presented in **Figure 2.17**. The first sought to determine the spectral and temporal nature of the activity conveyed by these projections. The second was to explore, based on these findings, their functional significance, specifically in the context of speech production and perception, but more generally in enabling sensorimotor tasks. We briefly shed light on this second aim via a mathematical model and simulation of the Mirror Network, which brings out a potentially critical function of the *forward* projections, namely in learning the *inverse* maps needed for control and performance of sensorimotor tasks.

We begin with a redrawing of the network of **Figure 2.17,** by unfolding the *inverse* mapping from the *forward* as shown in **Figure 2.18,** referred to henceforth as the *MirrorNet*. Here the auditory cortex is depicted twice, as an input and as an output. This organization of the system is well-known in the neural network literature as an **Auto-Encoder**, where the input (responses in the auditory cortex) is mapped onto itself at the output, through two transformations: an Encoder to a latent (hidden) representation (the motor responsive region here), and then through a Decoder back to the output (auditory cortex). Normally, such auto-encoder networks are simply trained by requiring that the Encoder & Decoder projections be able to reproduce the input with minimum error. In doing so, the auto-encoder finds a new, possibly more compressed and efficient but equivalent, representation of the auditory input as activations in the hidden (motor) region, which can still be mapped back to the auditory representations.

In the sensorimotor literature, it has always been assumed that the *forward* predictive (or Decoder) projection from the motor to sensory areas serves to monitor task performance and to provide rapid feedback of errors to ensure accurate motor execution (Daniel & Zoubin, 2000)upport theoretically and experimentally in the purely sensory perceptual domain (Keller & Mrsic-Flogel, 2018). The formation of this projection in sensorimotor systems is conceptually straightforward in that it serves as a model of the motor-plant, and hence can be learned by minimizing the differences ($\boldsymbol{e_d}$) between the Decoder and vocal-tract outputs as illustrated in **Figure 2.18.A**.

The counter *inverse* projection (or Encoder) serves to map the sensory expectations and intentions into the necessary motor commands to reproduce them. What is probably less appreciated is how conceptually difficult it is to learn a functioning *inverse* projection, for without a large set of predetermined exemplars (training data) to associate sensory signals to the correct neural motor commands, one has to resort to trial-and-error approaches. In the world of

classifiers and neural networks, large amounts of training data are key to accurate performance and generalization to unseen data. But it is often difficult to acquire such training material. For example, in the case of controlling the vocal-tract, learning to pronounce words of a new language relies not on finding out what the motor commands *ideally* need to be (which is impossible!), but rather on listening to our pronunciation of the words and trying to map the perceived errors ($\boldsymbol{e_c}$) back to implicit corrections of the motor commands. In the *top panel* of **Figure 2.18.B**, we illustrate that this backward propagation of the error to the motor areas requires conceptually that we compute the inverse of the vocal-tract so as to translate the sensory errors into motor-command adjustments, which subsequently can be minimized by adjusting the *inverse* mapping. In general, computing the vocal-tract inverse is difficult if not impossible because of its complexity, nonlinearity, and our incomplete knowledge of its workings.

The *MirrorNet* in **Figure 2.18.B** (*bottom panel*) solves this problem by adding a *forward* projection that parallels and serves as a model of the vocal-tract**.** The critical value of this "neural" projection is that it can readily provide a route for the $\boldsymbol{e_c}$ errors to backpropagate to the motor areas, and subsequently to train the *inverse* mapping. **Figure 2.18.C** illustrates a schematic of the resulting auto-encoder network, which like other neural networks, learns its connectivity by *backpropagating* the error (e.g., $\boldsymbol{e_c}$) through its "neural" pathways from stage to stage, adjusting the weights as the error proceeds backwards. This MirrorNet learns its Decoder weights by minimizing $\boldsymbol{e_d}$ as discussed earlier, but also learns its Encoder the same way, by backpropagating to minimize $\boldsymbol{e_c}$ *through* the Decoder neural pathway. Without this Decoder *forward* projection, the Encoder *inverse* mapping cannot be readily learned in this way since the error $\boldsymbol{e_c}$ has no route to propagate backwards through the motor-plant.

This leads us to the conclusion, that a crucial role played by the *forward* projection is to provide a pathway to learn the *inverse* mapping in an unsupervised way, and without any need for explicit motor training data. That is, by simply listening and uttering the words, the errors are automatically used to guide the vocal tract to reach its sensory target.

2.4.2.3 Simulating learning in the MirrorNet

A brief demonstration of "unsupervised" learning in the *MirrorNet* is provided here to illustrate the critical role of the *forward* projection in facilitating the learning of the *inverse* mapping. The network modules shown in **Figure 2.18.C** are implemented in *PyTorch* as a convolutional autoencoder to model the Encoder and Decoder pathways (see(Shamma *et al.*, 2020) for details). For the (input and output) auditory representations, we computed the audi-

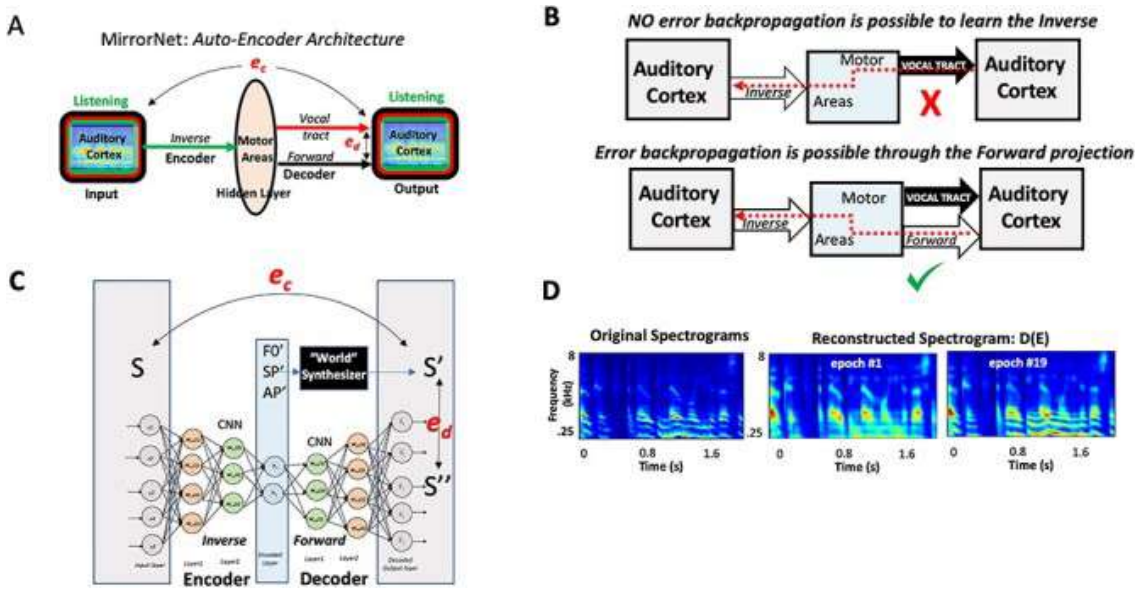


Figure 2.18. Simulating learning in the Mirror Network. (*A*). The overall layout of the sensorimotor interactions. It emphasizes the relative contributions of the inverse (Encoder) and forward (Decoder) projections between the auditory and motor areas. The overall network resembles a classic auto-encoder network that maps the auditory cortex activity onto itself through a hidden layer (motor regions), but with an additional non-neural motor-plant (vocal-tract) pathway that shares with the forward projection its motor input and auditory output. Two sources of error are available to train the neural pathways of the Encoder ($\mathbf{e_c}$) and Decoder ($\mathbf{e_d}$). (*B*) The critical role of the forward projection in providing a neural pathway for the ($\mathbf{e_c}$) error to backpropagate to the motor regions (hidden layers) so as to train the Encoder weights. (*C*) The MirrorNet implementation employs multiple layers of a convolutional neural network, and the "World" synthesizer as a simplified model of the vocal tract. (*D*) Training the MirrorNet results in progressive improvements in the reconstructed spectrograms projected through the sequence of Encoder–Decoder layers. The training is rather limited here involving only about 40 min of speech beyond the initialization with the random patterns.

tory spectrogram, a representation mimicking the cochlear outputs (Nima *et al.*, 2006; Taishih *et al.*, 2005). The vocal-tract model was simulated by the "World" synthesizer (Masanori *et al.*, 2016), a widely-used tractable vocoder model that takes three sets of input parameters as a function of time to synthesize a speech waveform: a spectral envelope function (SP), a pitch track (F0), and voicing/non-voicing indicator signals (AP). The goal of the *MirrorNet* here was to iteratively learn the Encoder weights (starting from random initial values) that map any (input) auditory spectrogram to the "motor" parameters that would both (1) reproduce the same spectrogram through the "World" synthesizer, and also (2) simultaneously regenerate it at the output of the Decoder projection, in which case both errors $\boldsymbol{e_d}$ and $\boldsymbol{e_c}$ are minimized.

The network was implemented with random initial weights for the Encoder and Decoder, and was fully trained using $<$ 60 minutes of speech. Two important procedures speeded up

and guided the learning of the correct mappings: (1) an initializing training epoch in which the network was briefly trained to minimize only $\boldsymbol{e}_d$ using *random* synthesizer-like parameters SP, AP, and F0. This epoch guided the Decoder to begin to reproduce the same type of output spectrograms as the synthesizer does, even if the input activations (in the hidden layer) were random. (2) Following the initialization step, speech spectrograms were used as auditory activations to minimize $\boldsymbol{e}_d$ and $\boldsymbol{e}_c$ alternately, i.e., with epochs in which only $\boldsymbol{e}_d$ is minimized while the Encoder is fixed, followed by epochs when the Decoder weights were fixed while the error $\boldsymbol{e}_c$ was backpropagated to compute the corresponding perturbations in the hidden layer, and subsequently make the necessary Encoder weight adjustments. These procedures succeeded in training the MirrorNet in an unsupervised manner and with normal speech material, thus demonstrating the utility of the *forward* pathway in learning the task of driving the synthesizer. **Figure 2.18.D** illustrates how reconstruction errors decreased over training epochs, and the evident improvement in the quality of the reconstructed speech spectrograms of an unseen sample sentence over time. Further technical details of constructing and training this neural network are given in (Shamma *et al.*, 2020).

Once the network was trained, it could readily *inverse*-map its sensory inputs (speech in this case) to the necessary parameters that drive the associated motor-plant (vocal-tract). Furthermore, the *forward* projection could still participate in its other commonly proposed predictive and control roles as a model of the motor-plant. The MirrorNet structure therefore is sufficiently general to serve as a model for analogous sensorimotor tasks requiring learning of a skilled performance, like playing a musical instrument, reading and writing, or training an autonomous vehicle to navigate traffic.

Speech production and comprehension

The experimental findings that justified the functional role of direct interactions between sensory percepts and motor acts are extremely diverse, beginning with the notion that a corollary discharge can function as a filter that suppresses self-generated sensory input allowing the animal to remain sensitive to external stimulation (Poulet *et al.*, 2006), to stabilize visual receptive fields by predicting saccade targets (Marc & Robert, 2002), to suppress auditory cortical activity during locomotion (Anders *et al.*, 2013; David *et al.*, 2014), or to facilitate vocal learning in birds (Georg & Hr, 2009; J *et al.*, 2008). Aside from the corollary discharge, or the *forward* projection common to all these examples, there are fundamental differences among them. For instance, all except for the last example, are due to instinctive processes that are not learned the way it is with the projections in birds learning a vocal repertoire. So, we shall distinguish and refer in our commentary here only to skillful continuous sensorimotor actions

requiring extensive practice such as the control of the vocal-tract in speech production or of the hand and fingers in musical playing. Hence, neither of these sensorimotor interactions are expected to exist with untrained motion or inappropriate sounds, as was demonstrated for speech and vocal tract production in (C. Gregory *et al.*, 2014).

At the phenomenological level that we adopt in this study, vocal learning in birds bears a close resemblance to the basic structure of human vocal-tract control and learning (**Fig.1A**). I physiological single-unit recordings in birds have unambiguously established the analog of the *forward* pathway, that it likely generates a detailed spectrotemporal representation of the stimulus which mimics that received from the ear during vocalizations (J *et al.*, 2008), and that this in turn would allow the bird to compare them and minimize the difference, and hence learn how to control its vocal source (Georg & Hr, 2009). Even the hypothesized induction of auditory responses with silent “chirping" seems to have been mentioned in passing many decades ago (Williams and Nottebohn, 1985)! All these details are reminiscent of the two directional projections and minimization of errors e_d and e_c depicted in **Figure 2.18**.

Speech production models vary considerably in their levels of description and detail. Some have focused on analytical formulations of the processes needed to control vocal-tract dynamics in speech production (Jason *et al.*, 2008; John & Edward, 2015; Parrell *et al.*, 2019). Others provided descriptions that encompass large regions of the brain combining both speech production and comprehension, and postulating specific bilateral neural substrates and connectivity patterns among them (C. Gregory *et al.*, 2014; Hickok & Poeppel, 2007; Poeppel, 2014; Poeppel *et al.*, 2012). Anatomically grounded accounts have also emerged from imaging experiments with fMRI and EEG that have emphasized the overall bidirectional flow of information across motor and sensory regions, and that have attempted to situate these processes within the overall flow of information from the auditory to the prefrontal cortex (Josef & Sophie, 2009; Lima *et al.*, 2016). The study by (C. Gregory *et al.*, 2014) comes closest to our experimental methodology in its recordings of responses in the **M, L,** and **S** conditions in similarly-defined auditory- and motor-electrodes. However, all their analyses had concentrated on the strong overt auditory and motor responses and the **S**-responses, and not as we do, on the covert activations due to the *forward*-and *inverse*-projections that are also evident in their data (e.g., their Fig. 2d displays weak *AUD* (green) and *PROD* (blue) responses during opposite conditions).

By contrast, the MirrorNet schematic that frames our experiments and motivates the data analyses is strictly phenomenological in flavor. Thus, while the postulated processes and interactions are biologically-plausible and supported by experimental evidence, the network model is largely agnostic with respect to the specific anatomical regions that source or receive the *forward* and *inverse* projections, or the biological implementations of the error signals, or how they might be backpropagated to adjust the weights and learn the projections. The network, however, makes specific predictions that intersect and potentially impact other

proposed formulations. For instance, the sensorimotor inputs into the auditory and motor cortical regions are evidently rapid, with dynamics that are commensurate with those of speech and the movements of the vocal tract. Furthermore, they are encoded in a manner consistent with the representational-domain of the recipient region, i.e., the *forward*-projection are auditory, and the *inverse*-projections are motor (**Fig. 6A**). The projections are also likely to be quite adaptive so as to learn (*forward*) and control (*inverse*) the specific structure of a person's vocal-tract (John & Michael, 2002). Hence, these properties are consistent the finding that the most auditory- and motor-electrodes implicated in the sensorimotor projections were localized in secondary (auditory) areas like the *STG* and *PT* (**Figs. 5**), and non-primary motor areas. These auditory responsive regions are highly adaptive, task-dependent, but are also spectrotemporally rich and agile to allow for reliable speech representation (Nima *et al.*, 2014), properties that are consistent with the MirrorNet requirements.

The framework of the MirrorNet is quite general and can serve many contexts outside of speech production and the vocal-tract. Any highly practiced actions associated with the reception or production of sensory signals would be served well by such a network as a means for controlling the motor-plant and learning its commands. For instance, sign-language and lip-reading are identical to speech production and perception in the context of the MirrorNet, but with visual and proprioceptive signals replacing the auditory, and hands, arms, or lips replacing the vocal-tract. Another example is playing the violin which involves extensive training of the fingers, arms, and postural musculature – the motor-plant – to produce the music. *Forward* projections must learn gradually with practice to model this motor-plant. Simultaneously, the *inverse* projection adapts to map the desired music into motor commands, and the learning thus proceeds by minimizing the two errors (**Fig. 6A**). Therefore, the MirrorNet structure predicts that these projections are highly specific to the skilled task that trained them, and hence their activations would not be recruited by inappropriate actions and sensory signals, as was demonstrated by the speech selectivity reported for vocal-tract activations (C. Gregory *et al.*, 2014).

In fact, MirrorNet interactions need not involve a motor task or motor-plant at all, but rather any constrained transformation that is not significantly amenable to adaptation. For instance, reading or sounding out a text is a transformation of a visual image (text) into corresponding sounds, often with complex rules of phonation (analogous to the complex rules of moving the vocal-tract) (Slowiaczek et al,1980). The *forward* projection would gradually learn the rules for mapping text to sounds, and in time, sound becomes an "imagined" output or the meaning of the text. The *inverse* mapping from the sound provides the image of the "expected" text – an imaginary writing task. These designations of course can be altered to describe learning to write or draw from a visual or an auditory image.

Therefore, the key idea common to all the above scenarios is an auto-encoder network

with *forward* and *inverse* mappings (**Fig. 6A**), which is the essence of the idea of the “mirror neurons". However, many extraneous issues have been appended to this network that are not an essential part of its function and that has led to numerous criticisms (L. Andrew *et al.*, 2009; H. Gregory, 2014). For instance, the *inverse* mapping has often been invoked as a realization of the “Motor Theory of Speech", the idea that speech perception occurs in the “motor domain" of the vocal-tract. Of course, nothing in the MirrorNet remotely suggests this. The purpose of the *inverse* mapping is not to perceive speech, but simply to control the vocal-tract. It is quite possible that speech perception and comprehension occurs in the auditory cortex, or in other derivative pathways, and this would still leave the MirrorNet architecture as an essential scheme for controlling the vocal-tract. Similar arguments apply to the role of the *forward* projection, which has been widely assumed to provide a predictive signal (the “efference copy") to facilitate control of motor performance (Daniel *et al.*, 1995), or to provide a sensory goal rather than a precise prediction (Caroline *et al.*, 2013). However, it is also possible to argue that this projection serves primarily as a route for the backpropagation of the error needed to learn the *inverse* mapping, without which it is difficult to control the vocal-tract. Therefore, the mirror neurons can serve an important function, but that does not need to include the “higher-level" cognitive tasks ascribed to them, from speech comprehension to empathy.

Finally, the architecture of the MirrorNet has been invoked in many perceptual contexts since it lends itself to many functional interpretations. One common case in point is as a substrate for *imagination*, i.e., sensory percepts devoid of external stimuli or actions without actual movements (Tian *et al.*, 2016). In the MirrorNet, the *forward* projection of a skilled pianist can recapitulate musical percepts by simply moving her fingers appropriately without actually producing a physical sound (Martin et al, 2018). In fact, as mentioned earlier, Martin’s study had already demonstrated that the “imagined" activity, which is experimentally similar to our **M** responses, exhibited detailed spectrotemporal structure much like the **L** responses. Similarly, the urge to dance or tap when listening to a beat or a melody can also be interpreted as commands injected from a trained *inverse* pathway into the appropriate motor areas. Such imagination can be recast as an expectation, anticipation, or prediction of sensory stimuli from a contextual memory or motor areas, and hence may serve a preparatory function (P. Andrew *et al.*, 2020). In fact, this view is consistent with Cogan et al.’s (2014) findings of sensorimotor transformations where auditory-responses were shaped by *subsequent*, hence expected vocal-tract actions. The MirrorNet, therefore, can be seen as a unifying architecture that can harmoniously organize diverse perceptual processes and sensorimotor tasks.

2.5 Cross-Modal Predictions: A New Computational Model for Sensory Motor Predictions in Music[6]

2.5.1 Introduction

Most organisms function by coordinating and integrating sensory signals with motor actions to survive and accomplish their desired tasks. For instance, visual and auditory signals guide animals to navigate their surroundings (Keller *et al.*, 2012; Wolpert & Ghahramani, 2000). Similarly, auditory and proprioceptive percepts are essential in skilled tasks like playing the piano or speaking. The difficulty of learning to perform these tasks is enormous. It stems from the fact that to control such actions, one needs harmoniously to close the loop between sensing and action. That is, it is necessary to map the desired sensory signals to the correct commands, which in turn produce exactly the desired sensory signals when executed.

But to learn the necessary mappings and interactions between the perception and action domains, standard Artificial Intelligence (AI) methodology typically relies on creating large databases that map the input sensory data to their corresponding actions, and then train intervening Deep Neural Networks (DNN) to associate the two domains (Fu *et al.*, 2019; Tai & Liu, 2016). Humans and animals however never learn complex tasks in this way. For instance, human infants learn to speak by first going through a "babbling" stage as they learn the "feel" or the range and limitations of their articulatory commands. They also listen carefully to the speech around them, initially implicitly learning it without necessarily producing any of it. When infants are ready to learn to speak, they utter incomplete malformed replica of the speech they hear. They also sense these errors (unsupervised) or are told about them (supervised) and proceed to adapt the articulatory commands to minimize the errors and slowly converge on the desired auditory signal. In other words, learning these complex sensorimotor mappings proceeds simultaneously and often in an unsupervised manner by listening and speaking all at once (Kuhl, 2004; Pagliarini *et al.*, 2021; Shamma *et al.*, 2020).

Motivated by such learning of complex sensorimotor tasks, a new autoencoder architecture, referred to as the "Mirror Network" (or MirrorNet) was recently proposed in Shamma et al. (Shamma *et al.*, 2020). The essence of this biologically motivated algorithm is the bidirectional flow of interactions ('forward' and 'inverse' mappings) between the auditory and motor responsive regions, coupled to the constraints imposed simultaneously by the actual motor plant to be controlled. In this study, conducted with Yashish Maduwantha, we extend and demonstrate the efficacy of the MirrorNet architecture in learning audio synthesizer controls/-parameters to synthesize a melody of notes using a commercial, widely available synthesizer

[6]Authors: Yashish M. Siriwardena, Guilhem Marion, Shihab Shamma(Siriwardena *et al.*, 2022)

(DIVA) developed by U-He[1].

MirrorNet is fundamentally different from the Differentiable Digital Signal Processing (DDSP) based models (Engel, Hantrakul, *et al.*, 2020; Engel, Swavely, *et al.*, 2020) which effectively learn a differentiable music synthesizer, whereas the goal of the MirrorNet is to learn controls to drive a given non-differentiable, off-the-shelf music synthesizer. Previous work with DNNs on determining music and speech synthesizer controls are all based on at least partially supervised techniques which often involve large databases of audio and control parameter pairs (order of 1000s) (Esling *et al.*, 2020; Georges *et al.*, 2021; Le Vaillant *et al.*, 2021; Yee-King *et al.*, 2018). Furthermore, previous efforts have mostly demonstrated the ability to compute the controls for single notes or single vowels for speech (Esling *et al.*, 2020; Saha & Fels, 2020). In this study, we propose an alternative approach model which is fundamentally unsupervised, in that it does not require matched pairs of input melodies and their corresponding control parameters. The proposed model can predict synthesizer controls for a melody composed of several notes demonstrating the scalability of the model for real-world applications. The true potential of the MirrorNet is further validated by showing how well it can predict synthesizer controls not only for DIVA-generated melodies but for other off-the-shelf synthesizer-generated melodies.

2.5.2 MirrorNet Model

2.5.2.1 Model Architecture

The MirrorNet was initially proposed as a model for learning to control the vocal tract and is based on an autoencoder architecture. The structure of this network is shown in Figure 2.19a (Shamma *et al.*, 2020), depicting the biological structures and experiments that motivated the network. The goal of the model is to learn two neural projections, an inverse mapping from the auditory representation to motor parameters (Encoder) and a forward mapping from the motor parameters to the auditory representation (Decoder). For simplicity, we use auditory spectrograms (Wang & Shamma, 1994) generated from the audio streams as the input and output representations, but other representations may prove more versatile (e.g., cortical representations (Chi *et al.*, 2005)). The "motor" parameters in this study are the parameters needed to synthesize the closest possible audio signals matching the inputs. The primary difference between this MirrorNet and the previously studied model in (Shamma *et al.*, 2020) is the use of the music synthesizer (DIVA) with its unique set of parameters.

As shown in Figure 2.19a, the MirrorNet model is optimized simultaneously with two loss functions namely the 'encoder loss'(e_c) and the 'decoder loss'(e_d). The encoder loss is the typical autoencoder loss - the Mean Squared Error (MSE) between the input auditory spectrogram and

[1]https://u-he.com/products/diva/

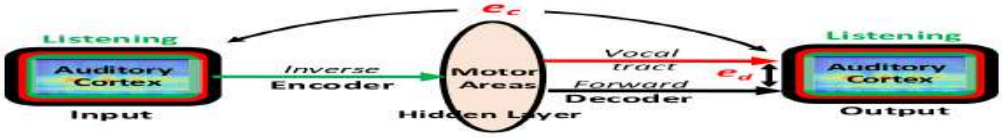
the reconstructed auditory spectrogram from the decoder (forward path). The decoder loss is the MSE between the auditory spectrograms generated by the DIVA (the motor plant path) and the decoder (forward path). It is the 'decoder loss' that constrains the latent space to converge to the expected control parameters while simultaneously reducing (e_c), and this is the key feature of the MirrorNet architecture.

Figure 2.19b shows the role of the 'forward' path in the model, namely to back-propagate the errors computed to learn the 'inverse' mapping and hence the control parameters. In general, directly computing a vocal tract or an audio synthesizer inverse is difficult if not impossible because of its complexity, nonlinearity, and our incomplete knowledge of its workings. The MirrorNet in Figure 2.19b (bottom panel) solves this problem by adding the forward projection that serves as a parallel, "neural" model of the vocal tract of the audio synthesizer, or any motor plant to be used. The critical importance of this "neural" projection is that it readily provides a route for the e_c errors to back-propagate to the motor areas (latent space), enabling the training of the inverse mapping (Encoder).

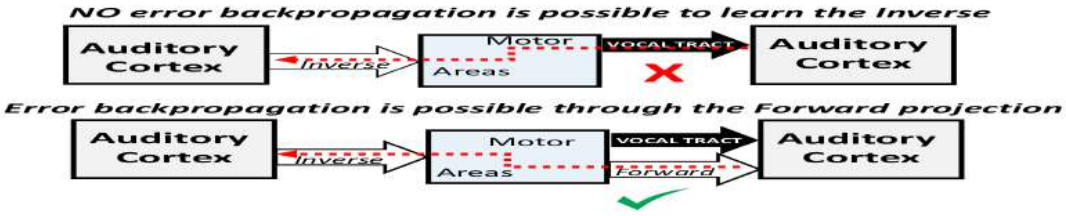
2.5.2.2 Model Implementation and Training

The MirrorNet for audio synthesizer control is implemented in PyTorch with 1-D convolutional (CNN) layers modeling both the encoder and decoder. The complete network is inspired by the multilayered Temporal Convolution Network (TCN) (Lea *et al.*, 2017). Figure 2.20 shows the complete DNN model architecture with its sub-modules used for pre/post-processing and dilated TCN. The pre/post-processing modules are symmetrically matched (C1≡C12, C2≡C11, C3≡C10) and have 128, 256, and 256 filters respectively with 1×1 kernels. d1, d2, and d3 dilated CNN layers have a kernel size of 3 with 1,4 and 16 dilation rates respectively. The CNN layers in the encoder and decoder are also symmetrically matched and the C4, C5, and C6 layers have 256, 128, and 7 filters respectively with 1×1 kernels. The latent space dimensions are chosen to match the number of parameters to be learned and the number of notes in each melodic segment. For example, to learn 7 controls of the DIVA synthesizer to generate a melodic segment of 5 notes, we use a latent space of (7×5) dimensions. Average pooling is done after C4, C5, and C6 layers (window sizes of 5, 5 and 2 respectively) while upsampling is done before C7, C8, and C9 layers (window sizes of 2, 5, and 5 respectively). The auditory spectrograms used as inputs (and outputs) of the model are of dimension (128×250). We use auditory spectrograms which have a logarithmic frequency scale, simply because they provide a unified multi-resolution representation of the spectral and temporal features likely critical in the perception of sound (Chi *et al.*, 2005; Wang & Shamma, 1994).

Unlike a regular autoencoder, the MirrorNet is trained in two alternating stages in each iteration. The decoder is trained first (to minimize e_d) for a chosen number of epochs. Then, the encoder is trained (to minimize e_c) for a given number of epochs and this alternation of



(a) MirrorNet: Autoencoder Architecture



(b) Role of the forward pass

Figure 2.19. MirrorNet Model Architecture for speech and the critical role of the forward projection (taken from *Learning Speech Production and Perception through Sensorimotor Interaction* by Shamma et al. in *Cerebral Cortex Communications*.)

training is continued until both losses converge to a minimum. Learning rates of 1e-2 and 1e-3 were used for the encoder and decoder networks, respectively. The best learning rates were determined based on a grid search testing all the combinations from [1e-2, 1e-3, 1e-4, 3e-4] for both the encoder and decoder which resulted in the lowest training errors at convergence. The two objective functions were optimized using the ADAM optimizer with an 'ExponentialLR' learning rate scheduler and a decay (gamma) of 0.5. All the models were trained using NVIDIA Quadro P6000 GPUs and on average the models converged after around 32 hours of training. For further implementation information on the network, the PyTorch project is publicly available in GitHub [2]. Sample audio reconstructions can also be found in the supporting web page hosted in the GitHub repository.

2.5.2.3 DIVA audio synthesizer

We use DIVA, an off-the-shelf commercial synthesizer as our audio synthesizer for the MirrorNet model. DIVA has almost all its parameters MIDI-controlled. A python library called RenderMan [3] is used to batch-generate audio files using a fixed set of parameters. We built a software layer with RenderMan to drive DIVA to synthesize a melody of notes by concatenating individual notes synthesized by DIVA. All the melodies used in this study are 2 seconds long and sampled at 44.1 kHz. The parameters are all continuous and normalized between [0,1]. Table 2.1 lists the set of parameters selected for the learning experiments with the MirrorNet, and the corresponding parameter labels from DIVA where applicable.

[2] https://github.com/Yashish92/MirrorNet-for-Audio-synthesizer-controls
[3] https://github.com/fedden/RenderMan

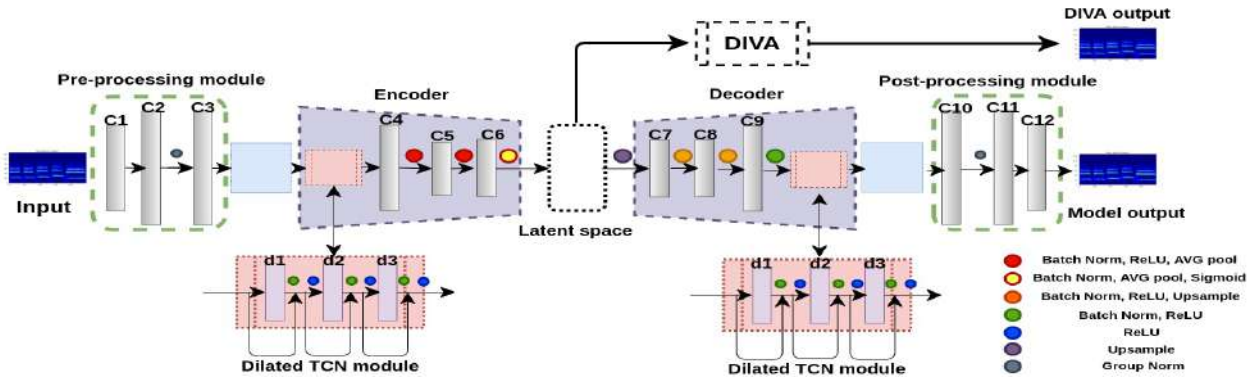


Figure 2.20. DNN architecture of the MirrorNet model. Here C1-C12 represent 1D-CNN layers where d1-d3 represent 1D dilated CNN layers.

Table 2.1. Set of Audio controls/parameters used. Here MIDI note and MIDI duration are parameters set in RenderMan library to drive the synthesizer patch.

Parameter Name	**DIVA preset**
MIDI note (Pitch)	-
MIDI duration	-
Volume	OSC : Volume2
Band pass filter (center frequency)	VCF1: Frequency
Filter Resonance	VCF1: Resonance
Envelope Attack	ENV1: Attack
Envelope Decay	ENV1: Decay
Vibrato Rate	LFO1: Rate
Vibrato Intensity	OSC : Vibrato
Vibrato Phase	LFO1: Phase

2.5.3 Experiments and Results

2.5.3.1 Learning DIVA parameters from melodies synthesized with the same set of parameters (set 1)

In this first experiment, we use 400 melodies (set 1) to train the MirrorNet and test with 80 melodies, all originally synthesized by DIVA. The advantage of this set of melodies is that we have its ground-truth parameter values, and hence we can assess how accurately the MirrorNet rediscovers them and reconstructs the melodies. Each melody contains 5 notes and is 2 seconds long. The train and test set of melodies were synthesized by randomly sampling a total of 7 parameters (the first 7 parameters in Table 2.1) using a defined range and keeping a pre-defined set of other parameters constant across all notes and melodies. The pre-defined set of parameters used for the experiments can be found in the GitHub repository of the project.

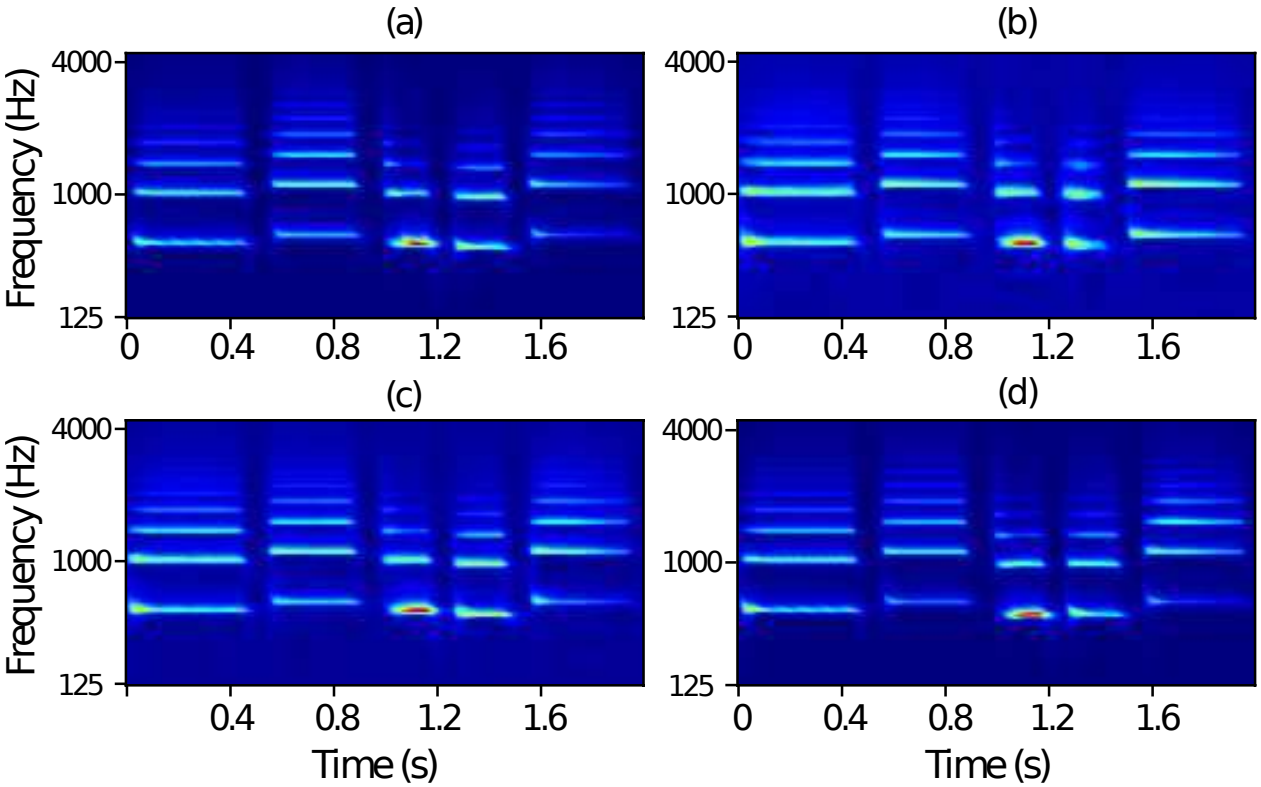


Figure 2.21. Auditory spectrograms from the model learned with DIVA synthesized melodies (set 1). (a) Input melody (b) Decoder output from true DIVA parameters (c) Final output from the decoder (d) DIVA output from the learned control parameters

Figure 2.21 depicts auditory spectrograms of a given melody at various stages in the fully-trained MirrorNet. The spectrogram (b) suggests how well the decoder has learned to generate an identical spectrogram to the one generated with DIVA for the exact same controls. The spectrogram (d) suggests how well predicted DIVA controls are from the encoder to synthesize an identical melody to the input.

We performed preliminary statistical tests to evaluate the robustness of the MirrorNet in predicting the 7 parameters. The plot in Figure 2.23.a validates that the predicted and ground truth parameters are significantly closer together than would result from a random set of values. A second test was performed to check how well the predictions of each parameter are compared to a random prediction. For that, we performed a Levene's test that confirmed that all parameter predictions were significantly better than chance. The plot in Figure 2.23b shows the parameter difference distributions for the test set. The distributions also suggest that critical parameters like pitch, bandpass filter, filter resonance, and duration are predicted with significant accuracy as volume and envelope attack parameters are predicted with comparatively lower accuracy.

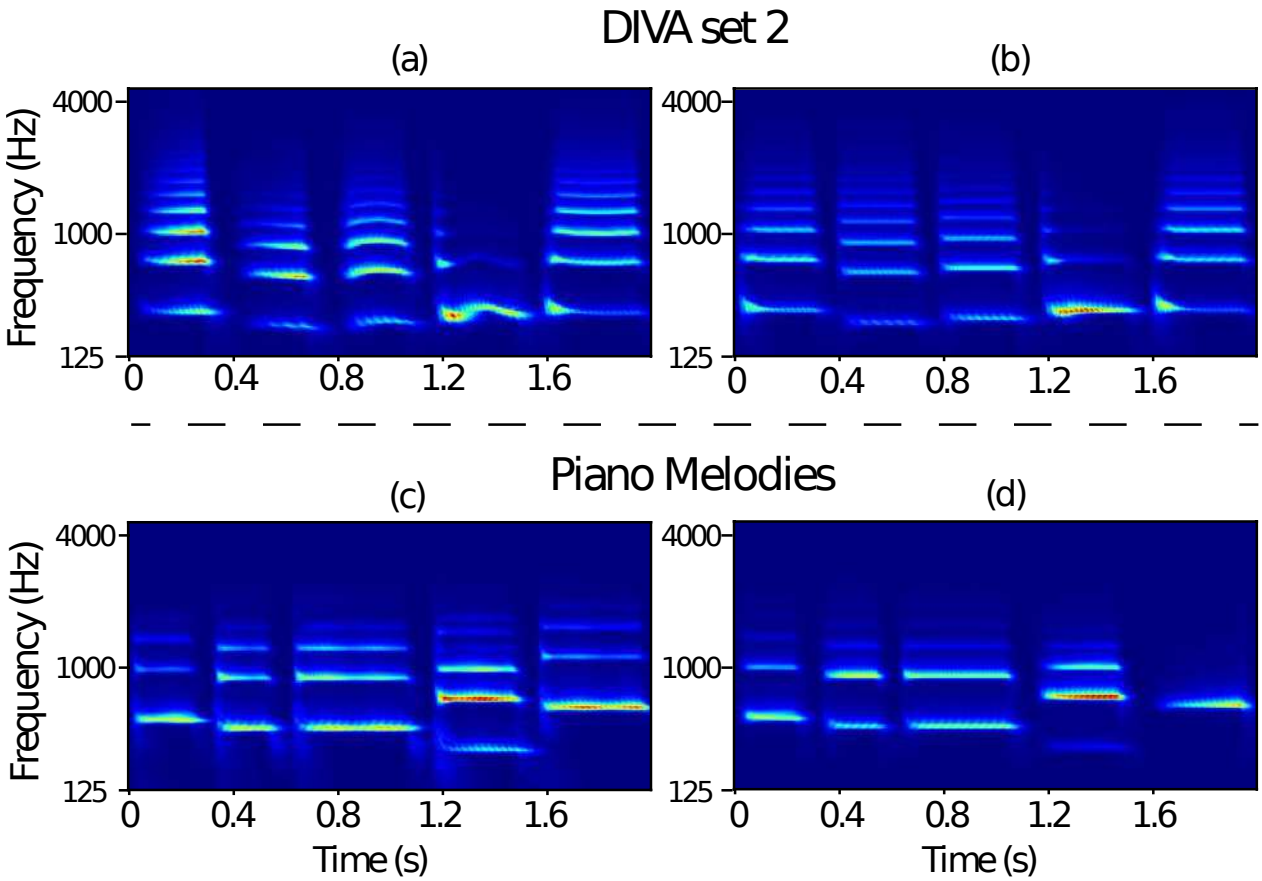


Figure 2.22. (Top panel) Auditory spectrograms from the model learned with DIVA synthesized melodies (set 2) (a) Input melody (b) DIVA output from the learned control parameters. (Bottom panel) Auditory spectrograms from the model learned with piano melodies. (c) Input melody (d) DIVA output from the learned control parameters.

2.5.3.2 Learning DIVA parameters from melodies synthesized with extra unknown DIVA parameters (set 2)

In this experiment, we use a train set of 400 and a test set of 80, both DIVA-generated melodies (set 2) which are synthesized in a similar fashion to set 1 except for the fact that they now use all the 10 parameters in Table 2.1. The MirrorNet is still trained to predict 7 parameters as in the previous experiment. The goal here is to demonstrate that the MirrorNet can approximate the input melodies even if they have additional sound/musical qualities that are impossible for the restricted set of 7 DIVA parameters to reproduce, e.g., vibrato in this case. The top panel in Figure 2.22 illustrates the original (vibrato) notes and the successfully regenerated melody captured with only 7 parameters (vibrato not included).

Table 2.2. Mean and variance of Mean Square Errors (MSE's) across multiple model training runs

Input melody type	Train/Test vs DIVA(learned)	Parameter-Train	Parameter-Test
DIVA melodies (set 1)	2.995±.21/3.596±.15	0.0666±.003	0.0671±.002
DIVA melodies (set 2)	6.380±.34/8.101±.20	0.0832±.007	0.0857±.004
Piano melodies	4.585±.25/4.751±.22	-	-

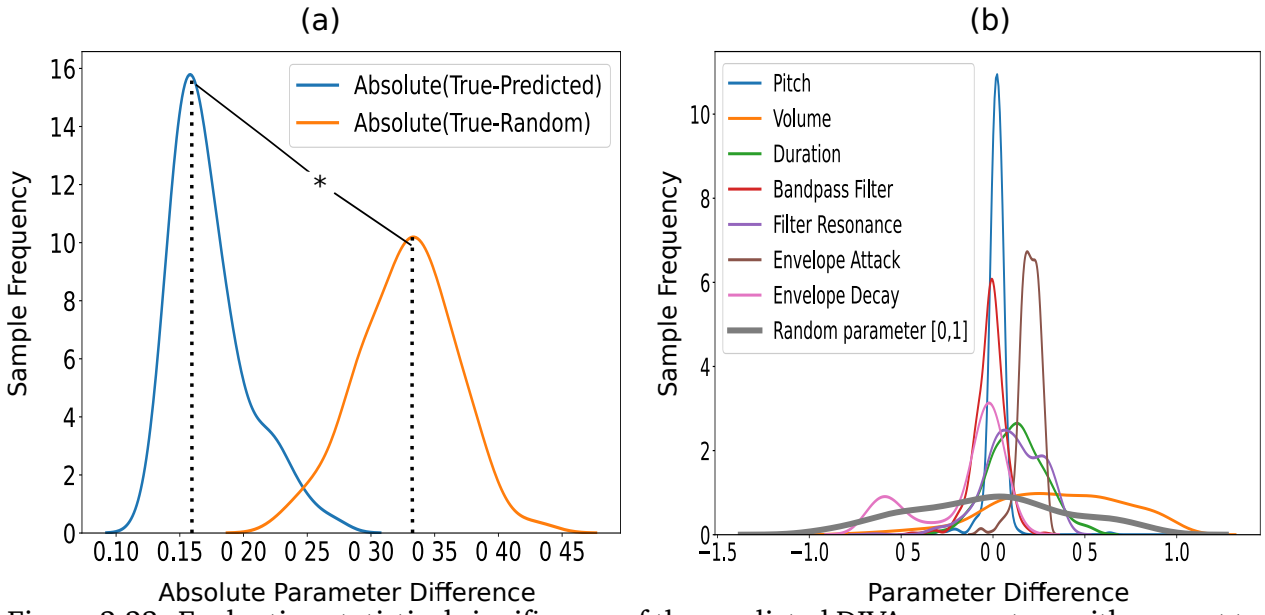


Figure 2.23. Evaluating statistical significance of the predicted DIVA parameters with respect to a set of random parameters on the test set (a) Distributions for absolute parameter differences across all parameters (b) Distributions of parameter differences (ground truth - predicted) for 7 parameters and the distribution for a random parameter difference (ground truth - random)

2.5.3.3 Learning DIVA parameters to synthesize melodies generated from other synthesizers

A fundamental advantage of the MirrorNet is its ability to discover the DIVA parameters corresponding to music generated by other sources and synthesizers by finding parameters that allow the DIVA output to be as close as possible, given the constraints of the number of parameters (here 7 are used), to the original input. The experiment utilized 400 5-note long piano melodies of 2 seconds that are synthesized by a Fender Rhodes digital imitation (Neo-Soul Keys generated through Kontakt 5). The network successfully reproduces accurate renditions of the piano music from unseen samples (test set of 80 samples) using the decoder/encoder mappings learned during the training. The bottom panel in Figure 2.22 shows such an example where the DIVA produces a melody that closely resembles the input piano melody.

2.5.4 Discussion

We described a MirrorNet model inspired by cortical sensorimotor interactions measured when humans speak or play a musical instrument (Shamma *et al.*, 2020). The first two experiments utilized DIVA-generated melodies for training, and this allowed us to evaluate the effectiveness of the MirrorNet given the ground truth parameters to compare against, e.g., to perform preliminary tests to validate the MirrorNet predictions of the synthesizer controls across all the training and test sets, as shown in Table 2.2. The MSE values for the test set compared to the train set in Table 2.2 also give an idea of how well the model generalizes for the unseen input melodies.

Taking the MirrorNet to the next level in the last experiment, we demonstrated how the MirrorNet could closely approximate a set of controls for DIVA to synthesize a set of piano melodies generated by a completely different synthesizer. This idea opens up a whole new area of applications in music synthesis as it describes a tool to find parameters for an arbitrary synthesizer that maximally approximate an arbitrary sound without being necessarily capable to exactly reproduce it (reproduce a violin using a guitar for instance). It should also be noted that this study only discusses results in synthesizing fixed duration melodies with a fixed number of notes, but it is a step in the right direction to synthesizing a piece of music which can have a variable number of notes in a fixed frame of audio.

The inspiration for the MirrorNet also comes from the area of computational neuroscience, especially to learning and predictive processing. Our brain is able to extract strong relations between sensory stimuli and their corresponding motor parameters that enable children to learn to speak by mere passive exposure to speech without any proper external teaching. In addition, after learning to control their own vocal tract, adults can, without any additional training, produce sounds they hear even if the acoustic target is not reachable by their specific vocal tract (case of experiments 2 and 3). However, the brain is able to find a set of motor parameters that approximate the target sound while being produced by the specific vocal tract. Such predictive mechanisms can also be seen in music production when humans learn how to play an instrument by mapping the auditory stimulation to the motor commands to a specific instrument. Even music perception relies on similar predictive pathways where high-order cortical areas constantly predict activation in the auditory cortices in order to modulate attention and emotions, for instance(Di Liberto, Pelofi, Bianco, *et al.*, 2020; Marion *et al.*, 2021).

Finally, from an engineering perspective, the MirrorNet can solve problems where it is hard to find a reasonable number of examples to train a regular feed-forward DNN network, or to learn from examples that may not be exactly similar to the motor-plant outputs, e.g., learning to synthesize a melody from naturally played music. We moreover believe that the MirrorNet can be generalized to design algorithms that can control motor plants such as self-driving vehicles given various sensory data.

2.5.5 Conclusion and Future Work

This study presents an autoencoder architecture inspired by sensorimotor interactions to discover and learn audio synthesizer controls. The work is novel in that the proposed MirrorNet can learn the necessary controls to produce a melody in a completely unsupervised way. It can also be potentially generalized to learn the controls for any motor-plant action from the sensory data associated with them. However, to realize all these potentials, many more advances are needed. For example, for the audio synthesizer controls explored here, it is necessary to scale up the current implementations to far more parameters that capture richer aspects of the sound (e.g., vibrato), to deploy more advanced and richer representations of the sound beyond the spectrograms, to devise more efficient and faster training paradigms, and finally to target the synthesis of continuous musical melodies which can have a variable number of notes.

2.6 General Discussion

Neural responses recorded with EEG during musical imagery exhibited detailed temporal dynamics that reflected the effects of melodic expectations, and a TRF that is delayed and with an inverted polarity relative to that of responses exhibited during listening. The responses shared substantial characteristics across individual participants and were also strong and detailed enough to be robustly and specifically associated with the musical pieces that the participants listened to or imagined.

This study demonstrates for the first time that melodic expectation mechanisms are as faithfully encoded during imagery as during musical listening. EEG responses to music (and other signals such as speech) segments are typically modulated by the probability of hearing that sound within the ongoing sequence: the less probable (unexpected) it is, the stronger the EEG expectation response (Di Liberto, Pelofi, Bianco, *et al.*, 2020). Therefore, the finding that imagined music is modulated similarly to listened music hints at the nature and role of musical expectation in setting the grammatical markers of our perception. Thus, as in speech, expectation mechanisms are utilized to parse the musical phrases and extract grammatical features to be used later for other purposes. This idea has already been discussed, and several studies have shown that musical expectations are used as primary features in other cognitive processes from memory (K. Agres *et al.*, 2018) to musical pleasure (Gold, Pearce, *et al.*, 2019). For instance, thwarted or fulfilled expectations have been shown to modulate activity in brain regions related to the reward system (Cheung *et al.*, 2019), specifically to emotional pleasure (Blood & Zatorre, 2001; Zatorre & Salimpoor, 2013) and dopamine release (Salimpoor *et al.*, 2011), discussed in more details in 4.1.0.2 and 5.1. Therefore, it is likely that imagery induces the same emotions and pleasure felt during musical listening because melodic expectations are encoded similarly

in both cases. This explains why musical imagery is a versatile place for music creation and plays a significant role in music education.

Viewed within a system framework, auditory imagery responses represent predictive responses prompted by higher-level cognitive processes that simulate the brain's perception of incoming stimuli. This mechanism can be compared to the perceptual equivalent of the efference copy, frequently initiated by the motor system (Ventura *et al.*, 2009). This conceptual analogy has spurred a multitude of investigations into auditory imagery within motor contexts, such as covert speech, suggesting that these imagined responses may possess a predictive motor nature (Y. Ding *et al.*, 2019; Tian & Poeppel, 2010; 2012; 2013; Whitford *et al.*, 2017).

In the domain of musical imagery, rhythm, specifically, has been intricately associated with the activation of the Supplementary Motor Areas (SMA) and pre-SMA (Bastepe-Gray *et al.*, 2020; Gelding *et al.*, 2019; A. R. Halpern, 2001; A. R. Halpern & Zatorre, 1999; Herholz *et al.*, 2012; Lima *et al.*, 2015; 2016; Meister *et al.*, 2004; Zatorre & Halpern, 2005). Moreover, notational audiation (musical imagery induced by reading musical scores) and passive listening have been demonstrated to elicit covert excitation of the vocal folds, exhibiting a neural signature akin to that observed during actual musical imagery (Brodsky *et al.*, 2008; Pruitt *et al.*, 2018; Zatorre *et al.*, 1996). The bidirectional relationship between motor and imagery processes is evident in an Electrocorticography (ECoG) study, which revealed robust auditory responses triggered by silent keystrokes on a keyboard (Martin *et al.*, 2017).

It becomes apparent that imagery processes are likely facilitated by the intricate connections between motor and sensory regions that typically co-activate during task performance, such as the coordination between vocal-tract activity and speech production, finger movements during piano playing, and visual processing during reading (Shamma *et al.*, 2020). However, these tight interconnections pose experimental challenges in disentangling the sources of neural activity (Zatorre *et al.*, 2007), given that auditory imagery may be partly influenced by motor components (A. R. Halpern & Zatorre, 1999). Irrespective of their origins, it is imperative to regard imagery responses as top-down predictive signals. The most striking evidence in our dataset is the polarity inversion relative to listening responses. This inversion facilitates the comparison between bottom-up sensory activation and its top-down prediction, thereby generating an "error" signal. Predictive coding theories have long postulated that this "error" signal is pivotal information that deeply permeates brain processing (Koster-Hale & Saxe, 2013; Rao & Ballard, 1999). This critical observation has been thoroughly investigated in detail in study #2.

Many studies have already shown that neural responses were modulated by expectation. For example, the amplitude of event-related potentials (ERPs) for model-predicted expected or unexpected notes can be compared. This approach has been successfully utilized with different methodologies, including EEG (Di Liberto, Pelofi, Bianco, *et al.*, 2020; A. R. Halpern *et al.*,

2017; Marion *et al.*, 2021; Omigie *et al.*, 2013a; M. T. Pearce *et al.*, 2010) , MEG (Quiroga-Martinez, C. Hansen, *et al.*, 2020; Quiroga-Martinez, Hansen, *et al.*, 2020), ECoG (Di Liberto, Pelofi, Bianco, *et al.*, 2020), and sEEG (Omigie, Pearce, *et al.*, 2019). Recent studies used time-continuous evaluation of expectation from the model to predict EEG data using *Temporal Response Function* decoding (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Di Liberto *et al.*, 2021; Marion *et al.*, 2021), allowing for precise correlation measurements between the recordings and the predicted EEG. We conducted such a TRF analysis and confirmed the hypothesis that low-frequency EEG responses to naturalistic music encode melodic expectations in correspondence of prospectively predictable silent events. The responses to silent events were shown to co-vary with the expectation strengths, which were drawn from a note onset-time statistical model (M. T. Pearce, 2005), as it was previously shown for music notes (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Omigie *et al.*, 2013b). These results are in line and go beyond previous measurements of the neural responses to sensory omissions, which focused on scenarios where strong expectations on the upcoming occurrence of a stimulus were built artificially (missing stimulus potentials - MSP; (Bendixen *et al.*, 2009)). This study shows a very clear evidence of the second hypothesis of the Predictive Coding Theory: probabilistic prediction hypothesis.

Cross-sensory predictions, which very possibly have links with the imagery responses as discussed earlier, were also investigated. Study #3, inspired by thereby characterized bi-directional predictions between motor and auditory areas during speech, defined a computational model for learning sensorimotor interactions. The architecture of the MirrorNet has been invoked in many perceptual contexts since it lends itself to many functional interpretations. One common case in point is as a substrate for *imagination*, i.e., sensory percepts devoid of external stimuli or actions without actual movements (Tian *et al.*, 2016). In the MirrorNet, the *forward* projection of a skilled pianist can recapitulate musical percepts by simply moving her fingers appropriately without actually producing a physical sound(Martin *et al.*, 2017). In fact, as mentioned earlier, Martin's study had already demonstrated that the "imagined" activity, which is experimentally similar to our **M** responses, exhibited detailed spectrotemporal structure much like the **L** responses. Similarly, the urge to dance or tap when listening to a beat or a melody can also be interpreted as commands injected from a trained *inverse* pathway into the appropriate motor areas. Such imagination can be recast as an expectation, anticipation, or prediction of sensory stimuli from a contextual memory or motor areas, and hence may serve a preparatory function (P. Andrew *et al.*, 2020). In fact, this view is consistent with(C. Gregory *et al.*, 2014) findings of sensorimotor transformations where auditory-responses were shaped by *subsequent*, hence expected vocal-tract actions. The MirrorNet, therefore, can be seen as a unifying architecture that can harmoniously organize diverse perceptual processes and sensorimotor tasks. In addition to his role in modelling perceptual processes and sensorimotor tasks, this model can also be seen as a neuromorphic engineering powerful algorithm. Indeed,

it sources from the fact that the vocal tract is non differentiable and proposes to model to way the brain solved this problem. However, learning inverse functions for non differentiable modules is a very complicated engineering problem that often requires very big amounts of labeled data to train an inverse model in a supervised way. The MirrorNet offers a way to solve this problem without any data in an unsupervised manner. A common case of application of such problem in music is for audio synthesizers. Music synthesizers are known for being extremely complex and complicated to the extent that even professional musicians can sometimes have a very hard time finding the right parameters of the synthesizer in order to generate a given sound. This can have industrial applications in the music industry when producers want to record extra takes of a synthesizer for which they have erased the presets or in order to quickly reproduce a sound from another recording, potentially using another synthesizer. Study #4 presents an autoencoder architecture inspired by sensorimotor interactions to discover and learn audio synthesizer controls. The work is novel in that the proposed MirrorNet can learn the necessary controls to produce a melody utilizing a given complex timbre in a completely unsupervised way. It can also be potentially generalized to learn the controls for any motor-plant action from the sensory data associated with them. However, to realize all these potentials, many more advances are needed. For example, for the audio synthesizer controls explored here, it is necessary to scale up the current implementations to more parameters that capture richer aspects of the sound (e.g., vibrato), to deploy more advanced and richer representations of the sound beyond the spectrograms, to devise more efficient and faster training paradigms, and finally to target the synthesis of continuous musical melodies which can have a variable number of notes.

To conclude, the work presented in the chapter enrich the field of neural predictions. It shows for the first time using electrophysiology that musical imagery induces neural responses of a predictive nature that are encoding musical expectation, even in the total absence of physical stimulation. We also showed that those same responses are the ones exhibited in natural moments of silence in ecologically valid music. This new finding is a very strong evidence toward the predictive coding theory, especially its second hypothesis. Finally, we built, based on recent findings about efference copies between the motor and auditory systems a computational model allowing for learning inverse functions of non differentiable modules. Especially, we have shown that we can solve the complex problem of finding the parameters for a complex synthesizer that produce a given sound.

3 New Statistical Models for Musical Expectation

3.1 General Introduction to Computational Models of Music Cognition

3.1.1 Presentation

The computational counterpart of musical predictions has received some recent highlights thanks to the development of computational models of musical structures that can be used to predict listeners' internal predictions for upcoming musical events. Specifically, this can be assessed by statistical models of music trained on a large corpus of musical pieces. They generate a probability distribution given a musical context and can be either generative or descriptive.

Generative models aim to produce music in a specific style by repeatedly sampling the probability distribution of the next note to produce an entirely new piece, while simultaneously respecting the structural grammar of the musical style.

Descriptive models are used to characterize the musical structures found in a particular piece of music. The modeled structures can be applied to novel musical samples in order to extract expectation values for each of the notes (i.e. how well the model predicted the note). Such models, supposed to predict listeners' expectations, have been able to account for, among others, melodic syntax (Margulis, 2003; M. T. Pearce, 2005), tension and resolution patterns (Farbood, 2012; Lerdahl & Krumhansl, 2007), harmonic and rhythmic structures(M. Rohrmeier, 2011), phrase boundaries (Lerdahl & Jackendoff, 1996), memory (K. Agres *et al.*, 2018), and arousal and valence (Egermann *et al.*, 2013; Sauvé *et al.*, 2018).

Architectures The architecture of such models is divided into two types: theory-based explicit rules *vs.* statistical learning of musical events (M. T. Pearce *et al.*, 2008).

Explicit, theory-based rules, also referred to as Gestalt-like rules, represent hard-coded principles of domain-general auditory perception applied to music processing (Temperley, 2008). For instance, the preference for small intervals may be rooted in Gestalt-like principles of general perception. Other explicit rules also include Western music-theoretic insights such as the idea that a melody should lie in a certain key (Lerdahl, 2004; Margulis, 2003; Narmour, 1990; Schellenberg, 1997).

These theory-based models fail to capture the fine-grained variations occurring between musical styles (Cenkerová & Parncutt, 2015) or musical cultures (M. T. Pearce, 2018). To bridge this gap, the statistical analysis of large musical corpora can be used to uncover embedded, non-explicit rules that vary from one style or culture to another. Statistical models can be formulated and applied to a musical corpus M with the aim of approximating the probability distribution P on a note n, given the context C consisting of a sequence of notes occurring before n. Thus,

the probability p of a given note can be written as (Eq.1):

$$P_M(n|C) = p \tag{3.1}$$

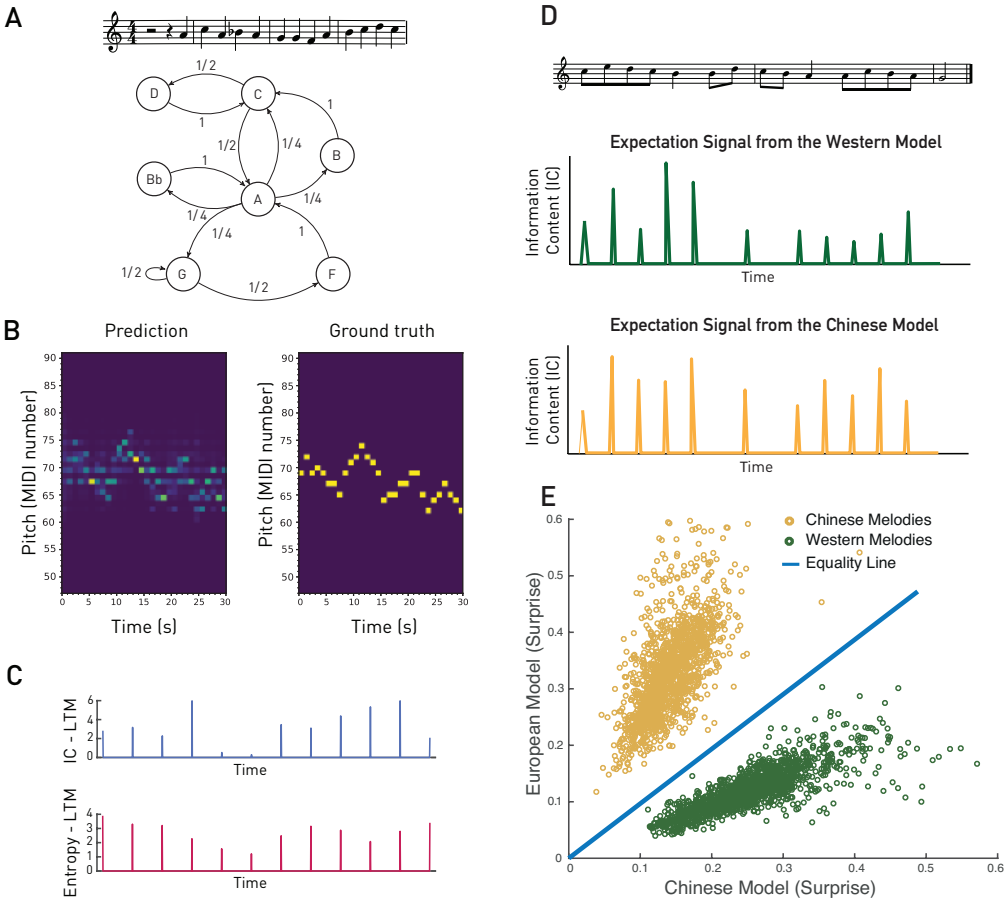


Figure 3.1. **A schematic representation of the IDyOM model. (A)** Graphical representation of the 1-order Markov-chain of the STM for the purple note on the melody *Als Jesus Christus in der Nacht* (BWV 265) by J. S. Bach. **(B)** (left panel) The predicted probability distribution for each upcoming note for the same melody. (right panel) The actual notes are used as a ground truth to compute the IC from the distribution. **(C)** The IC and Entropy for both the long-term model.

Recent advances in machine learning and especially in music generation have given rise to a new generation of models of music based on deep neural network architectures (Briot, 2021). They also aim at predicting a P_M distribution but using more complex learning functions. For instance, the DeepBach model (Hadjeres *et al.*, 2017) is designed to capture the syntax of Bach chorals using bidirectional LSTMs (long short-term memory). Another is the transformer model which was initially formulated for NLP translation machines (Vaswani *et al.*, 2018), but has been applied to music (C.-Z. A. Huang *et al.*, 2018) and is known to outperform previous models in

terms of music generation based on ratings from listeners. However, all these recent models are based on bi-directional (non-auto-regressive) information, meaning that musical events from the past and the future are used to compute the probability of a given event, making them inappropriate to model cognitive processes of music perception.

However, among the variety of models used to represent and estimate the probability distributions P_M, the most common of which are the models based on Markov chains (Abdallah & Plumbley, 2009; Ames, 1989; Gillick *et al.*, 2009; Manning & Schutze, 1999; M. Pearce & Wiggins, 2004; M. T. Pearce, 2005; Perruchet & Vinter, 1998; M. Rohrmeier & Cross, 2008). Markov chains applied to music are stochastic models describing the statistics of note sequences by collecting the probability of note transitions over k-order (as illustrated in fig.3.1.A in which 1-order statistics are collected) and estimating the probability of each note as a function of the preceding k-grams (a sequence of k notes). The IDyOM model, described by Marcus Pearce in his Ph.D. thesis(M. T. Pearce, 2005) dominates the field of music cognition. However, it has internal limitations: i) its implementation in the language Lisp makes it very hard to modify at ease in order to check for new cognitive hypotheses; ii) its Markov chains-based architecture (discrete and independent features) makes it only able to model symbolic data when audio data would allow for more ecologically-valid studies. We will discuss those limitations later.

3.1.2 Validation of Models

A computational approach to evaluate these models consists in assessing how well the model is able to generalize to unseen portions of the data set (theoretical evaluation). A usual way is to use the negative log-likelihood on testing data T, as described in Eq. 3.1.2. This technique allows to compare models on the same data in terms of computational generalization.

$$error = \sum_{n \in T} \frac{-log(P_M(n|C))}{|T|} \tag{3.2}$$

3.1.2.1 Neural and Behavioral Validation

However, evaluating these computational models against neural and behavioral evidence (experimental evaluation) contributes to a fine-grained characterization of the computational principles underlying musical enculturation. Yet, this is a challenging endeavor because of the multiple metrics used and the difference between the data sets they are trained and evaluated on.

For generative models, a type of evaluation consists of asking participants to report whether excerpts of music are taken from the training corpus (original data) or generated by the model

(Hadjeres *et al.*, 2017). A higher confusion between the two categories indicates a better performance of the model in mimicking human production. Singing (Carlsen, 1981; Fogel *et al.*, 2015; Sears *et al.*, 2018), rating (C. L. Krumhansl & Kessler, 1982; C. L. Krumhansl *et al.*, 2000) or guessing (Manzara *et al.*, 1992; M. T. Pearce, 2005) a probe tone in continuation of a priming melody can be used to approximate the probability distribution of the priming melody ending, which can be compared to the model's distribution (M. T. Pearce & Wiggins, 2006). Continuous ratings of arousal and valance were also demonstrated in evaluating a model's performance as they are known to correlate with the expectation of notes (Egermann *et al.*, 2013; Sauvé *et al.*, 2018). These measures are *subjective*, in the sense that it is based on participants' self-reports.

Alternative ways to evaluate models' performances exploit *objective* measures, either derived from behavioral or neural responses. These measures allow the experimenter to evaluate directly the signal of interest without the participants knowing what is being measured. For instance, memory for specific melodies is tied to how unexpected its content is (K. Agres *et al.*, 2018). In a similar vein, melodic priming can be used in paradigms that consist of collecting reaction times (RTs) on a timbre deviation task, relying on the correlation between RTs and the expectation of the pitch and rhythm (J. J. Bharucha & Stoeckig, 1986; Bigand & Pineau, 1997; Bigand *et al.*, 2001; Margulis, 2003; Margulis & Levine, 2006; Marmel *et al.*, 2008; 2010; Omigie, Pearce, & Stewart, 2012; Tillmann *et al.*, 2006). The efficacy of the model can then be evaluated by comparing the behavioral RTs with its computed expectations.

A recent paper took advantage of pupil dilation response to demonstrate that this physiological response was correlated to note expectation and modulated by different levels of uncertainty of predictions (Bianco *et al.*, 2019; Bianco *et al.*, 2020). Also, many studies have used neural data (EEG, MEG, ECoG) to assess the performance of descriptive models. For example, the amplitude of event-related potentials (ERPs) for model-predicted expected or unexpected notes can be compared. This approach has been successfully utilized with different methodologies, including EEG (Di Liberto, Pelofi, Bianco, *et al.*, 2020; A. R. Halpern *et al.*, 2017; Marion *et al.*, 2021; Omigie *et al.*, 2013a; M. T. Pearce *et al.*, 2010) , MEG (Quiroga-Martinez, C. Hansen, *et al.*, 2020; Quiroga-Martinez, Hansen, *et al.*, 2020), ECoG (Di Liberto, Pelofi, Bianco, *et al.*, 2020), and sEEG (Omigie, Pearce, *et al.*, 2019). Recent studies used time-continuous evaluation of expectation from the model to predict EEG data using *Temporal Response Function* decoding (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Di Liberto *et al.*, 2021; Marion *et al.*, 2021), allowing for precise correlation measurements between the recordings and the predicted EEG, as illustrated in Figure 3.2. These studies showed, for instance, that the expectation signal trained on Western music is encoded in the EEG recordings of Western listeners listening to and imagining Bach chorals (Marion *et al.*, 2021), and listening to partitas (Di Liberto, Pelofi, Bianco, *et al.*, 2020) and constructed spatial maps of the correlations. Also, they showed that both long- and short-term models are accurately encoded, justifying the structure of the IDyOM

model as well as the pertinence to the cultural information learned from the training corpus.

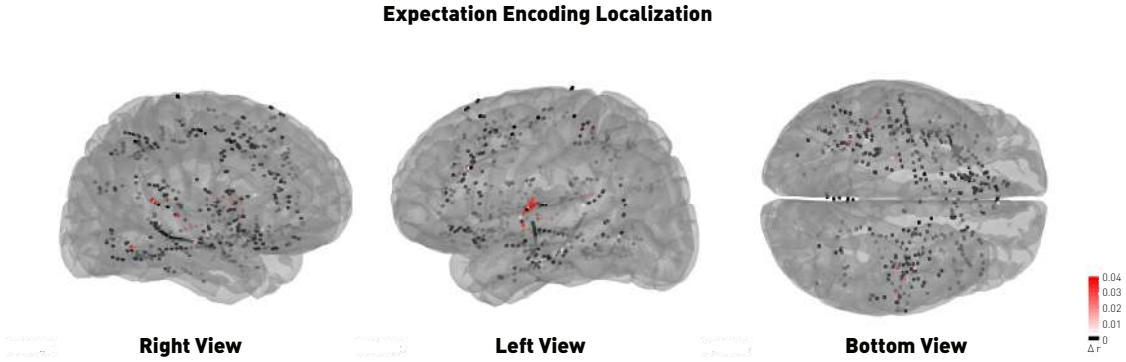


Figure 3.2. **Musical expectations are encoded in the brain.** In Di Liberto et al., 2019, the authors established that the relative surprise of musical events, as predicted by the IDyOM model (M. T. Pearce, 2005), was encoded by cortical activity recorded by EEG sensors, and especially with ECoG electrodes implanted in auditory regions typically involved in high order auditory processes (Di Liberto, Pelofi, Bianco, *et al.*, 2020). A kernel to describe the mapping of the continuous musical surprise in each sensor recording was trained and optimized to assess the musical encoding accuracy by correlating predicted to actual neurophysiological measurements. The figure shows all the electrodes implanted (2 patients). Black dots refer to electrodes non-responsive to music, the others are colored according to the electrode signal correlation with the expectation signal predicted by IDyOM.

3.1.2.2 Measuring Distance Between Musical Cultures

Finally, computational models of musical structure can be useful to quantify the stylistic or cultural distance between two musical corpora. As aforementioned, musical structures vary from culture to culture (Reck, 1977; Stevens, 2004). Pitch, rhythm, or meter features follow norms carved by centuries of tradition and practice (Arom, 2004; Fracile, 2003; M. T. Pearce & Wiggins, 2006). The extent to which extent two musical systems resemble or differ from each other is an important question, as it may predict how easily listeners can learn unfamiliar musical systems (Thaut *et al.*, 2018). Such a distance between two musical styles can be defined as "the degree to which the music of any two cultures differ in the statistical patterns of pitch and rhythm, and it will predict how well a person from one of the cultures can process the music of the other" (Demorest & Morrison, 2016). This is illustrated in Figure3.1.E which simulates the cultural distance between two corpora of Western and Chinese music, replicating the results reported in (M. T. Pearce, 2018). The corpora were made up of melodies ranging from 45 to 60 seconds in duration. From each corpus, a number of melodies were selected that had a comparable overall number of notes. An expectation signal for each melody was predicted from IDyOM trained on the Chinese or Western corpora (Figure 3.1.D).

3.1.3 Limitations of Current Models and Scientific Contribution

Statistical models of music are a key aspect of the contemporary challenges in the field of music cognition. However, the field is currently dominated by IDyOM which has two major limitations.

The first one is that its only implementation is in the Lisp programming language. This language becomes quite old and is less and less used, especially, it is not the language of predilection of our community. This makes it rather hard to use for colleagues who are not very comfortable with it or with the command line. One of the results is that almost no one can modify to source code of IDyOM to test new hypotheses or to use it for different purposes. It is for those reasons that we think it is very important for the community to publicly release of Python re-implementation of the IDyOM model. In this chapter, we will present our new implementation of IDyOM in Python, give a clear and deep comparison of its performances concerning the original Lisp implementation, and present new features and behavior of the model that were used in recent articles in the field and made possible thanks to this new implementation.

The second limitation of the IDyOM model is that is it solely symbolic making it suited to musical scores and not to audio recordings. This is a very important limitation as the corpora of symbolic data we have access to are limited and do not allow us to train models for specific communities, and more importantly individual listeners. Also, it constrains our experiments to be based on musical scores rather than actual recorded musical performances. We therefore have to synthesize the musical scores using computer software to guarantee the alignment between the stimuli heard by the participants and the input of the model; which considerably decreases the ecological validity of our experiments. That is why we decided to work on a new model based on continuous Bayesian inferences to statistically model music through spectrograms instead of symbolic musical scores.

This chapter will present those two new models: IDyOMpy and MusiREX. I designed and implemented IDyOMpy based on the original architecture of IDyOM, the information was extracted from Pearce's PhD thesis (M. T. Pearce, 2005). The project has been facilitated by Giovanni Di Liberto, Shihab Shamma, and Benjamin Gold who prepared the analysis using his self-reported pleasure data; they will be authors in the future article. Amélie Picard implemented the MusiREX model, and wrote the description of the model (section 3.3.2.3) under my supervision during her internship at LSP, based on the D-REX model originally designed by Benjamin Skerritt-Davis and Mounya Elhilali (N. Huang & Elhilali, 2017). The project has been facilitated by and Benjamin Gold who prepared the analysis using his self-reported pleasure data and will be reviewed by Benjamin Skerritt-Davis, Mounya Elhilali, Shihab Shamma; they all will be authors in the future article. I conducted all the analyses and generated all the

figures presented in the chapter (except for explicit mention). The presented sections are still unpublished but present a first version of future articles that will be slightly modified and soon sent to publishers.

3.2 IDyOMpy: a New Python Implementation for IDyOM, a Statistical Model of Musical Expectations [1]

3.2.1 Introduction

During the 1950's, the music critic Leonard Meyer advanced the idea that musical predictions were at the core of music perception(Meyer, 1956). The development of the Predictive Coding Framework(Clark, 2013; K. J. Friston *et al.*, 2010) has since further elaborated this idea and provided computational formulations for its implementation(Koelsch *et al.*, 2019; M. A. Rohrmeier & Koelsch, 2012; Vuust, Heggli, *et al.*, 2022a). This framework revolves around the notion that the brain learns a model of the world that is continuously used to predict sensory inputs. Perception, therefore, becomes an encounter between sensory inputs and their predictions(Keller & Mrsic-Flogel, 2018) generating a *prediction error* that is exploited to update the model(Näätänen *et al.*, 2007). This theory rests on two main hypotheses: (1) The Statistical Learning Hypothesis which states that the brain needs to learn and update an internal model of the environment's regularities; (2) The Probabilistic Prediction Hypothesis which postulates that predictions of the sensory inputs are based on the same internal model so as to modulate their neural encoding and facilitate their perception.

A large number of studies are currently investigating predictions in music(M. T. Pearce, 2018; M. A. Rohrmeier & Koelsch, 2012; Vuust, Heggli, *et al.*, 2022a; Witten *et al.*, 1994), speech(Norris *et al.*, 2016; Poeppel, 2012), vision(Enns & Lleras, 2008; Kimura, 2012), touch(Kilteni & Ehrsson, 2017; Schubotz, 2007), and even smell(Zelano *et al.*, 2011), many using computational models to account for human cognition(C. L. Krumhansl *et al.*, 2000; Nixon & Tomaschek, 2021) or cortical activity(Broderick *et al.*, 2018; Di Liberto, Pelofi, Bianco, *et al.*, 2020; Marion *et al.*, 2021). Models for speech are particularly varied and widespread, and include complex DNN implementations (Brown *et al.*, 2020; Mikolov *et al.*, 2013; Vaswani *et al.*, 2017) that are presumed to reflect different facets of human cognition(Caucheteux & King, 2022; Goldstein *et al.*, 2022). The community of music cognition has embraced this approach(M. T. Pearce & Wiggins, 2012) and already demonstrated its two hypotheses by demonstrating that explicit(Corrigall *et al.*, 2022; Fogel *et al.*, 2015; Morgan *et al.*, 2019; M. T. Pearce & Wiggins, 2006; Sears *et al.*, 2018) and implicit(Bianco *et al.*, 2019; Bianco *et al.*, 2020; Corrigall

[1] Authors: Guilhem Marion, Giovanni Di Liberto, Benjamin Gold, Shihab Shamma

et al., 2022; Di Liberto, Pelofi, Bianco, *et al.*, 2020; A. R. Halpern *et al.*, 2017; Marion *et al.*, 2021; Omigie, Pearce, *et al.*, 2019; Omigie, Pearce, & Stewart, 2012; Omigie *et al.*, 2013a; M. T. Pearce *et al.*, 2010; Politimou *et al.*, 2021; Quiroga-Martinez, C. Hansen, *et al.*, 2020; Quiroga-Martinez, Hansen, *et al.*, 2020) predictions correlate well with the probability of musical events in the listeners' culture(C. L. Krumhansl *et al.*, 2000). Prediction signals have even been measured during moments of musical silence which correlated well with the probability of the absent note(Di Liberto *et al.*, 2021). Moreover, it has also been determined that passive exposure to unfamiliar music engenders statistical learning that is consistent with the music heard. For instance, passive exposure to Eastern music (chosen because of its uncommon time signatures) facilitates in young children and adults the detection of violations in new musical excerpts with similar time signatures (E. E. Hannon & Trehub, 2005b). Another study replicated this phenomenon for pitch with adult listeners who gained superior abilities to predict the next note in melodies sampled from random musical grammar after being passively exposed to different melodies sampled from the same musical grammar(Loui *et al.*, 2010).

In general, predictions have accounted for many other facets of music cognition such as memory(K. Agres *et al.*, 2018), emotions(Sauvé *et al.*, 2018), pleasure(Gold, Mas-Herrero, *et al.*, 2019), reward(Cheung *et al.*, 2019) making it a rich framework for future musical studies (Pelofi *et al.*, n.d.; Vuust, Heggli, *et al.*, 2022). Compared to speech, however, the modeling of music cognition has been dominated by a single powerful model: IDyOM(M. T. Pearce, 2005), which has been used in almost every study of musical prediction and cited in over 300 articles. This model, however, is implemented in Lisp making it difficult to use and modify to test new cognitive hypotheses.

Here we propose a Python implementation of the IDyOM model with improvements such as an alternate technique for merging different Markov chains' orders. We also propose new features that have been used to explore new ideas about the brain, e.g., a model for computing the probability of having melodic notes during silent intervals, and a model that monitors learning during training. Finally, we provide a specific quantitative comparison with the original Lisp implementation using both theoretical (based on generalization error) and cognitive (based on EEG decoding and self-reported data) measures. We demonstrate that this new implementation replicates the original Lisp implementation and improves on some of its findings.

3.2.2 Implementation

Information Dynamics Of Music (IDyOM) is a statistical model of melodic progressions created by Marcus Pearce and published in 2005(M. T. Pearce, 2005). The model computes how expected (by means of *information content* and *entropy*) a note is in a given context after a training phase on a corpus of melodies.

3.2.2.1 Architecture

The model is based on variable-order Markov chains and is composed of two parts: a long-term model (LTM) which is pre-trained on a musical corpus, and a short-term model (STM) which is trained on the current evaluated piece in order to catch repeating structures within the song. Both the LTM and STM models rest on the same architecture, but the data they are trained on are different. An important limitation of Markov chains is that they are discrete models, making them unsuitable to work on continuous data such as raw audio waveforms or spectrograms. The use of IDyOM is therefore limited to symbolic musical scores.

Variable Order Markov Chains A Markov chain describes a memoryless[2] process which means that any event is only a function of the previous one. Formally, for $\forall i, X_i$ sequential random variables,

$$P(X_k = x | X_{k-1}, X_{k-2}, ..., X_0) = P(X_k = x | X_{k-1})$$

Let Σ be the set of all possible notes, referred top as the *alphabet*, borrowing the term from formal languages. $P : \Sigma^2 \rightarrow [0,1]$ is a function for the probabilities of transitions from note to note. Such a model can be expressed as an $n * n$ matrix or a graph $G = (V, E)$ where V (vertexes) is the set of notes, and E (edges) indicate the transition probabilities. This model is known as a *first-order Markov Chain*.

The fig. 3.3.A illustrates a simple example of a graph representation of a first-order Markov chain for music. It expresses the statistical model representing the beginning of the melody of *Au Clair de la Lune* (fig. 3.3.B).

Because of the highly structured nature of music, it reasonable to assume that note probabilities would depend on more than one prior note. Musical sentences are often constructed over a large number of previous notes and thus show long-term dependencies. By using n-grams as the alphabet of the Markov chain, it is still possible to use the Markov model and include long-term dependencies.

An n-gram is a combination of n elements of the alphabet Σ. For instance, if the alphabet is $\Sigma = \{a, b\}$, all the 2-grams are $\{aa, ab, ba, bb\}$. Formally, it is an element of the Cartesian product of the original set of states (in our case Σ). For instance, 2-gram $\in \Sigma \times \Sigma$, 3-gram $\in \Sigma \times \Sigma \times \Sigma$, ..., and so,

$$n\text{-gram} \in \prod_{k=1}^{n} \Sigma$$

By using n-grams as elements of S, the set of states of our Markov Chain, we can define the transition probability between n-long words ω, $\forall n$,

[2] The next event only depends on the value of the current event. No memory is stored.

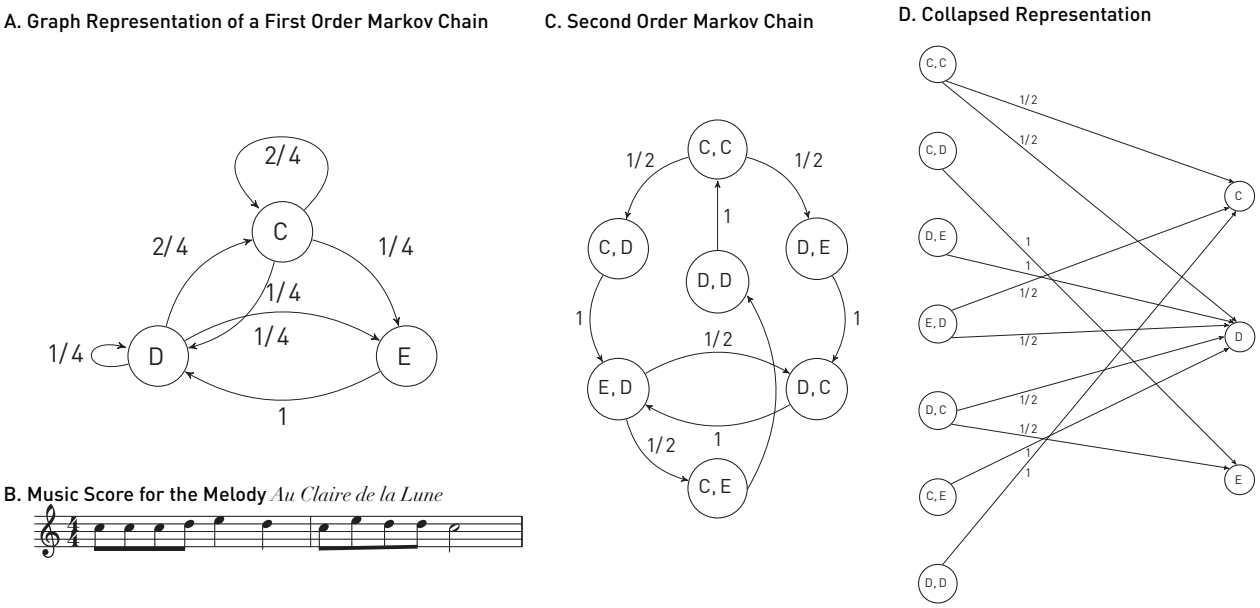


Figure 3.3. (**A**) Markov chains of order 1 and (**B**) order 2 corresponding to **(C)** the score of the beginning of the melody *Au Claire de la Lune*. (**D**) A collapsed representation of order 2 illustrating the flow of all subsequent states collapsed into single notes. Therefore, the context states are seen on the left and the target (subsequent) notes on the right. This is a simplified version of the Markov chains that is more suited for predicting the next note (as opposed to the next sequence) and is used in the implementation of IDyOM to simplify the computations.

$$P(X_{k:k+n} = \omega | X_{k-n:k})$$

Fig. 3.3.C shows a graph representation of the 2-order Markov chain on the melody *Au Clair de la Lune*, where the number of states hugely increase, and more data are needed to accurately train the model, but now facilitating the representation of more complex structures. For computational reasons, we can collapse the graph by summing across all words starting with the same note and therefore get the probability to observe a given single note after a given context (c.f. fig. 3.3.C):

$$P(X_k = x | X_{k-n:k}) = \sum_{\{\omega\}|\omega_0=x} P(X_{k:k+n} = \omega | X_{k-n:k})$$

Variable-order Markov chains have the flexibility to use n-grams of different lengths and to dynamically adapt the utility of each order. The ability to embed n-long temporal dependencies allows for modeling melodic sentences.

Merging Different Orders (Lisp Implementation) It is generally difficult to merge all distributions (one per order) into a single one. In the original IDyOM, the Prediction by Partial Matching (PPM) algorithm is used to approximate the final $P(X_k = x | X_{k-n:k})$. PPM(Cleary & Witten, 1984) is a data compression scheme in which the central component is an algorithm for

performing back-off smoothing of n-gram distributions. This model is usually referred to as the order-minus-one model and allows for the prediction of events that have yet to be encountered.

The original IDyOM uses the following definition:

$$P(X_k = x|X_{k-n:k}) = \alpha(x|X_{k-n:k}) + \gamma(X_{k-n:k}) \cdot P(X_k = x|X_{k-n+1:k})$$

The functions $\alpha()$ and $\gamma()$ are computed using the PPM algorithm(M. T. Pearce, 2005)(see (Moffat, 1990) for the original method). Note that $P(X_k = x|X_{k-n:k})$ corresponds to the Markov chain of order $n-1$ (here estimated with PPM). By iterating recursively, we encounter all orders and assign a weight to each probability distribution. The following method is used:

$$\gamma(X_{k-n:k}) = \frac{t(X_{k-n:k})}{\#X_{k-n:k} + t(X_{k-n:k})}, \text{ and,}$$
$$\alpha(x|X_{k-n:k}) = \frac{\#X_{k-n:k} \cdot x}{\#X_{k-n:k} + t(X_{k-n:k})}$$

Where $t(C)$ denotes the total number of symbols of Σ that have occurred with non-zero frequency in context C. This method allows one to account for the *diversity* of distributions. Thus, a distribution that only encountered a few n-grams will be less represented than a distribution that saw all the alphabet.

Merging Different Orders: A New Implementation The PPM algorithm computes an approximation of large distributions and guarantees some important properties of the approximated distribution (such as that they approximately sum to 1). This allows for fast computation but results in sub-optimal results. Therefore, instead of using the PPM algorithm to merge the different orders of the Markov chains, we propose to use an arithmetic mean weighted by the inverse of the relative entropies of the distributions. We denote by RE_i the relative entropy of the probability distribution given by the context $X_{k-i:k}$ corresponding to the i^{th}-order model.

$$P(X_k = z|X_{k-n:k}) = \frac{\sum_{i=1}^{n} P(X_k = x|X_{k-i:k}) \cdot RE_i^{-1}}{\sum_{i=1}^{n} RE_i^{-1}}$$

The relative entropy is the Shannon entropy normalized by the maximal entropy of the distribution (defined by the number of elements in the support of the distribution). It allows the weights to be comparable between orders. As the higher the order the more states are represented the entropy is then artificially higher, normalizing by the maximal entropy to account for this problem. Entropy is defined later in 2.5

$$RE(X) = E(X)/Emax(X)$$

The maximal entropy is defined by the entropy of the uniform distribution that shares the same support (number of states):

$$Emax(X) = -\sum_{n} 1/n \cdot log_2(1/n)$$

This method allows better cross-validated predictions over the training set.

The Short-Term Model The short-term model consists exactly of the same computational model as the long-term model described before but is not trained on a corpus. It is trained during the testing phase, therefore, it only takes into account the very local grammar of the tested piece. It is useful for accounting for local structures and repetitions within the pieces that do affect the predictions but do not come from a long-term statistical learning process(Conklin, 1990) (key, modulations, theme repetitions, ...). The probability distributions of the short-term model and the long-term model are merged using the arithmetic mean weighted by the inverse relative entropies of the models (as described above for the different orders), b is an additional parameter that allows for sharpening or smoothing of the final distribution[3]:

$$P(X_k = x) = \frac{E_{LTM}^{-b} \cdot P_{LTM}(X_k = x) + E_{STM}^{-b} \cdot P_{STM}(X_k = x))}{E_{LTM}^{-b} + E_{STM}^{-b}}$$

Entropy Approximation A straightforward implementation would directly compute the entropy of the long- or short-term models so as to merge them. However, since the entropy was already computed in order to merge the Markov chains' orders, we already know them for the distributions from which the long-term and short-term models are drawn.

$$P(Z = z|C) = \frac{\sum_{i=1}^{n} P(Z = z|C_i) \cdot E_i^{-1}}{\sum_{i=1}^{n} E_i^{-1}}$$

Therefore, it is useful to find a way to compute the entropy of P only from E_i. One possible approach is to use the mean of the self-weighted entropies which proved to be a good approximation that reduced computation times by a factor of 5:

$$E = \sum_{i}^{n} E_i \cdot E_i^{-1} / \sum_{i}^{n} E_i^{-1}$$

[3] In our implementation we use $b = 1$.

We ran the entire set of analyses presented above to compare the versions using the approximation *versus* the actual computations of the entropies, and found no significant differences. Nevertheless, the online implementation provided includes both options.

3.2.2.2 Viewpoints

Music evolves across at least 5 dimensions: Pitch, duration, timbre, intensity, and spatialization. IDyOM assumes that those dimensions are independent when computing their joint product. While the dimensions most often considered with IDyOM are pitch and duration of the notes, any other feature can be included in the model as long as it is discrete.

$$P(X_k = x) = P(Pitch_k = x_{pitch}) \cdot P(Duration_k = x_{duration})$$

X_k is a valid probability distribution (sums to 1) if $Pitch_k$ and $Duration_k$ are. With P and D, respectively the sets of all pitches and durations:

$$\sum_{x \in \Sigma} P(X_k = x) = \sum_{p_i \in P} \sum_{d_i \in D} P(Pitch_k = p_i) \cdot P(Duration_k = d_i)$$
$$\sum_{x \in \Sigma} P(X_k = x) = \sum_{p_i \in P} P(Pitch_k = p_i) \cdot \sum_{d_i \in D} P(Duration_k = d_i)$$
$$\sum_{x \in \Sigma} P(X_k = x) = 1$$

3.2.2.3 Training

Transition probabilities are learned from a corpus of melodies. We compute the frequencies (counts) over all random variables and use them as probabilities[4].

$$P(X_k = x | X_{k-n:k}) = \frac{\#X_{k-n} \ldots X_{k-1} X_k}{\#X_{k-n} \ldots X_{k-1}} = \frac{\#X_{k-n:k} \cdot x}{\#X_{k-n:k}}$$

3.2.2.4 Features Computed from the Models

Information Content The negative log-likelihood of a note x, referred to as *information content* (IC), represents how well the model predicts it given the context $X_{k-n:k}$. This computation is numerically stable with an interpretation in terms of compressibility, or of measuring information. For instance, events with high information content are difficult to compress as they occur rarely, one can therefore say that they contain a lot of information. This metric has been shown to provide good measures for psychological interpretations of perceptual data (Attneave, 1954; Chater & Vitányi, 2003).

$$IC(x | X_{k-n:k}) = -log_2(P(X_k = x | X_{k-n:k}))$$

[4]We use $\#\omega$ as the number of occurrences of the word ω in the corpus and $\cdot$ as the concatenation operator. Therefore, $\#X_{k-n:k} \cdot x$ denotes the count of words starting with $X_{k-n:k}$ and ending with x in the whole corpus.

Entropy The *entropy* provides an approximation of the uncertainty given a context C. In information theory, this measure evaluates the amount of information contained in a signal (and not for an event, as the IC). In the case of a probability distribution, it reflects the flatness of the distribution given by the model to estimate the confidence of the prediction. If all outcomes are equiprobable (the model cannot gather any information), the entropy will be maximum and the prediction will be highly uncertain. If one outcome has a probability 1 and all others 0, the entropy will be minimum ($E = 0$) and the prediction is certain. For instance, the first note of a melody is very uncertain as almost all notes are equiprobable (high entropy), whereas, the next note during a repeated sequence is very certain as it is very likely to be the one we heard during the previous repetitions.

$$E_M(C) = \sum_n P_M(n|C) \cdot log_2(P_M(n|C))$$

3.2.3 Methods For Evaluating Model Performance

In order to compare our new implementation with the previous Lisp version, we define a few metrics we will run with both implementations. We first present *theoretical measures* assessing how well each model generalizes to unseen data, then assess *cognitive measures* through the decoding of EEG recordings of participants listening to music, and finally, we correlate the results with behavioral data.

3.2.3.1 Generalization Errors

A common computational approach to evaluate the different implementations consists of assessing how well the model generalizes to unseen portions of the dataset (theoretical evaluation), using the negative log-likelihood with testing data T as described in Eq. 3.2.3.1. This technique allows us to compare models trained on the same data in terms of computational generalization[5].

$$error = \sum_{n \in T} \frac{-log(P_M(n|C))}{|T|} \tag{3.3}$$

Using the average negative log-likelihood over unseen data is based on the idea that notes in an unseen score (underlined by the same distribution, i.e., same musical genre) should have in mean (because of the law of large numbers) a greater probability than the ones that did not appear. Since the probability distribution must sum to 1, a more accurate distribution should

[5]This method only works if the two models compute the IC in similar domains. For instance, this method cannot be used to compare discrete and continuous models.

generate large probability (low negative log-likelihood) on the notes of the score. Machine Learning methodologies often use negative log-likelihood to evaluate their models(C. A. Huang *et al.*, 2018).

To this end, we used three homogeneous datasets of melodies: Bach chorals, traditional Chinese melodies from the region of Shanxi, and a large database of Western folk melodies. All were sampled from the Essen Folk Songs database[6]. We used k-fold cross-validation by dividing each dataset into 5 folds. We trained a model on 4 of them and evaluated the remaining one. We then computed the average negative log-likelihood for each song and compared them between models.

3.2.3.2 Cultural Distance

IDyOM has been shown to be a good model for musical enculturation as it allows modeling of cultural distances(M. T. Pearce, 2018). Therefore, one way to assess the accuracy of a model is through the extent to which it can differentiate melodies taken from different cultures. Here, we train 2 models: one on Bach chorals and one on traditional music from the region of Shanxi. We use both test/train and cross-validation to compute the average generalization error for every excerpt according to both models. We then construct a scatter plot where the x and y axis is the generalization error for, respectively, the Shanxi and the Bach models, where each point is a music piece. A bad model would collapse all pieces on the equality line failing to separate the two cultures, whereas an excellent model draws the 2 groups apart on either side of the equality line and thus classifies the two cultures well.

To quantify the extent to which the two cultures are separated we defined three measures:

Inter-cultural distance (interCD) represents the average euclidean distance between each point of the first culture and each point of the second culture. A value of 0 means that all points collapse, the bigger the value the further the two cultures are in the model space.

Intra-cultural distance (intraCD) represents how close the pieces are withing a culture, it is a proxy for the variability in generalization error and the stability of the model. Small values mean more stable model (less variance).

Clustering index $= \frac{interCD}{intraCD(A)/2+intraCD(B)/2}$ combines both inter- and intra-cultural distances into a composite measure that tells to which extend it is easy to classify the two cultures.

[6]http://www.esac-data.org/

3.2.3.3 EEG Decoding

IDyOM has been widely used in studies of the psychology and neuroscience of music, especially recently in decoding of EEG recordings that allow for a physiological benchmarking of the model. To this end, we used data from two recent studies (Study #1(Di Liberto, Pelofi, Bianco, *et al.*, 2020) and Study #2(Marion *et al.*, 2021)) that employed IDyOM to decode the EEG data. We compared the results of the analyses using the two implementations of the IDyOM model. Both experiments used a Biosemi Active Two 64-electrodes System and were digitally filtered between 1 and 8 Hz using a Butterworth zero-phase filter (low- and high-pass filters both with order 2 and implemented with the function filtfilt) , and down-sampled to 64 Hz for Study #1 and with high-pass filters down to 0.1 Hz and low-pass filters up to 30 Hz for Study #2. EEG channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation. All channels were then re-referenced to the average of the two mastoid channels for Study #1 and using global re-referencing for Study #2 study. The stimuli were composed of 10 Bach partitas for Study #1 and 4 Bach chorals for Study #2.

The analysis was conducted in a similar fashion as in the original studies by estimating temporal response functions (TRFs)(N. Ding & Simon, 2012; Lalor *et al.*, 2009b) with the mTRF-Toolbox(Crosse, Liberto, & Lalor, 2016). This de-convolution method (implemented as a lagged linear regression) was used to regress the IC signal computed by both implementations of IDyOM with the pre-processed EEG recordings using cross-evaluation. Pearson's correlation was computed between the predicted and original EEG signals. As the predicted EEG signal was only constructed from the IC signal from IDyOM, the correlation measures the resemblance between the IC and the EEG recording. An IC signal that is more accurately matching human perception is expected to generate larger EEG prediction correlations, providing us with a tool for estimating the physiological validity of each model.

3.2.3.4 Behavioral Preference

A recent study(Gold, Pearce, *et al.*, 2019) showed that the Entropy from the IDyOM model could explain 19% of the variance of 44 participants' behavioral liking measured by means of a 7-item Likert scale on 57 stimuli. Stimuli reported to be familiar to the participants were excluded from the analysis. There was a significant Wundt (quadratic correlation, a.k.a. inverted-U shape) effect between the liking ratings for the songs and the mean duration-weighted Entropy of the same songs. We, therefore, used these data as a way to estimate the validity of the entropies computed by our model. To do so, we replicated the results of this study on the same data but using our model trained on the same corpus. We then compared the r^2 (explained variance) using both models. In order to compute the significance of the difference between

the two models, we computed the distribution for each model using a Bootstrap method. We computed the r^2 of the sub-sampled data (80% sampled from both participants and songs) 5000 times using the same indexes for each model. We then computed the difference distribution and computed the p-value for it being inferior or equal to 0. This p-value is reported in the result section. We also report the individual p-values computed during the correlations.

3.2.4 Results

3.2.4.1 Information Content

We first used the generalization error (c.f. 3.2.3.1) to compare the models on different datasets. We found that the new Python version significantly outperformed the previous implementation in all three datasets: traditional Chinese music from Shanxi, Bach chorals, and a large Western database (Fig. 3.4.A). We also used our new feature *Training Monitoring* (c.f. 3.2.5.2) to compare the trace of the generalization error over the course of the training. We observed that the final point of the Lisp implementation is reached with fewer data for IDyOMpy (Fig. 3.4.B). Finally, we correlated the raw IC for each note of each Bach choral between the two models. We found a relatively strong, correlation of $r = 0.7$ indicating that the two models are consistent but not identical.

We then plotted the cultural distance between traditional Chinese music from Shanxi and Bach chorals for the two models (Fig.3.5.A). The IDyOMpy outperformed the IDyOM Lisp for the inter-cultural distance in that it separated better the two cultures (Table 1). However, the results also showed different intra-cultural distances depending on the corpus. However, the overall clustering index was better for IDyOMpy demonstrating an overall superior performance for musical cultural classifications.

	Inter-Cultural Distance	Intra-Cultural Distance on A	Intra-Cultural Distance on B	Clustering Index
IDyOM Lisp	1.3924	0.99461	1.0726	1.3471
IDyOMpy	1.7914	1.169	0.97733	1.6693

Table 3.1. **Cultural Classification Metrics for Both Models.** The metrics are defined in 3.2.

Finally, we used the mTRF toolbox to predict EEG recordings of participants listening to Western music (in two different studies, c.f. 3.2.3.3) from the IC signal computed with the two models. We found no significant difference in the accuracy. However, we should note that the EEG recordings are extremely noisy signals and it is likely that subtle differences in the IC's would not result in significant differences in EEG predictions.

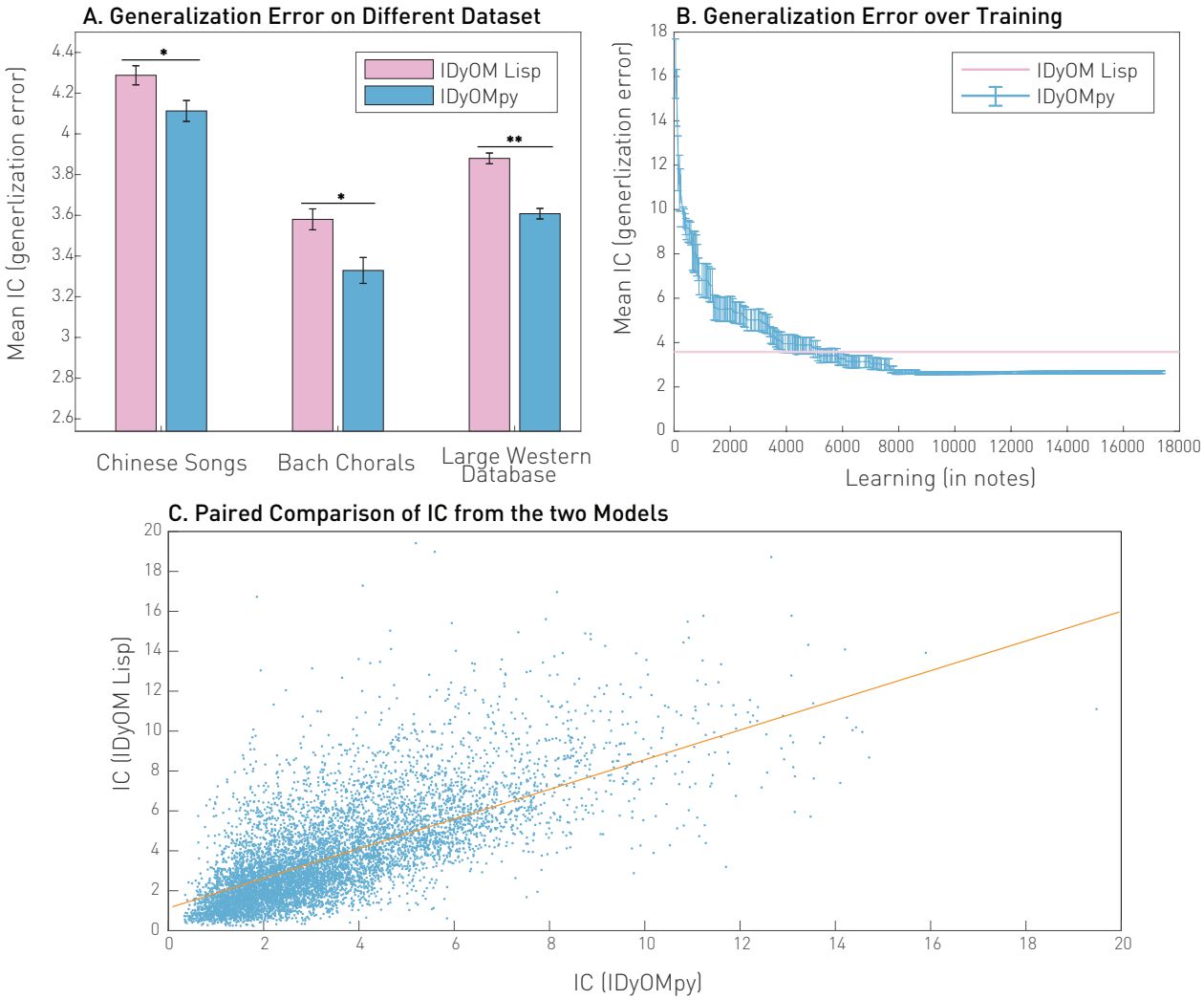


Figure 3.4. **Comparison of the Generalization Errors.** **A**: Average generalization error for different datasets. Significance: $* : p < 10^{-4}$; $** : p < 10^{-23}$. **B**; Generalization error over the course of the training of the model. **C**: Correlation of the IC for each note. Pearson's $r = 0.7$

3.2.4.2 Entropy

To compare the Entropies computed by both models, we first correlated the raw estimates from the two models for each note of each Bach choral. We found a relatively weak correlation of r=0.3 (Fig.3.6.C) indicating that the two models compute the entropy differently. We then used data from (Gold et. al., 2019) in order to assess which model explains the most variance of the behavioral liking rating (c.f. 3.2.3 for method). We found that the new implementation explains 22% of the variance compared to 19% due to the Lisp version. This difference was significant and resulted in a p-value < 0.0001. This result leads us to conclude that even if the two models compute Entropy somewhat differently, they both replicate results from (Gold et. al., 2019) and that IDyOMpy even outperforms the Lisp implementation in terms of variance explained giving it a cognitive validation of the Entropy computations.

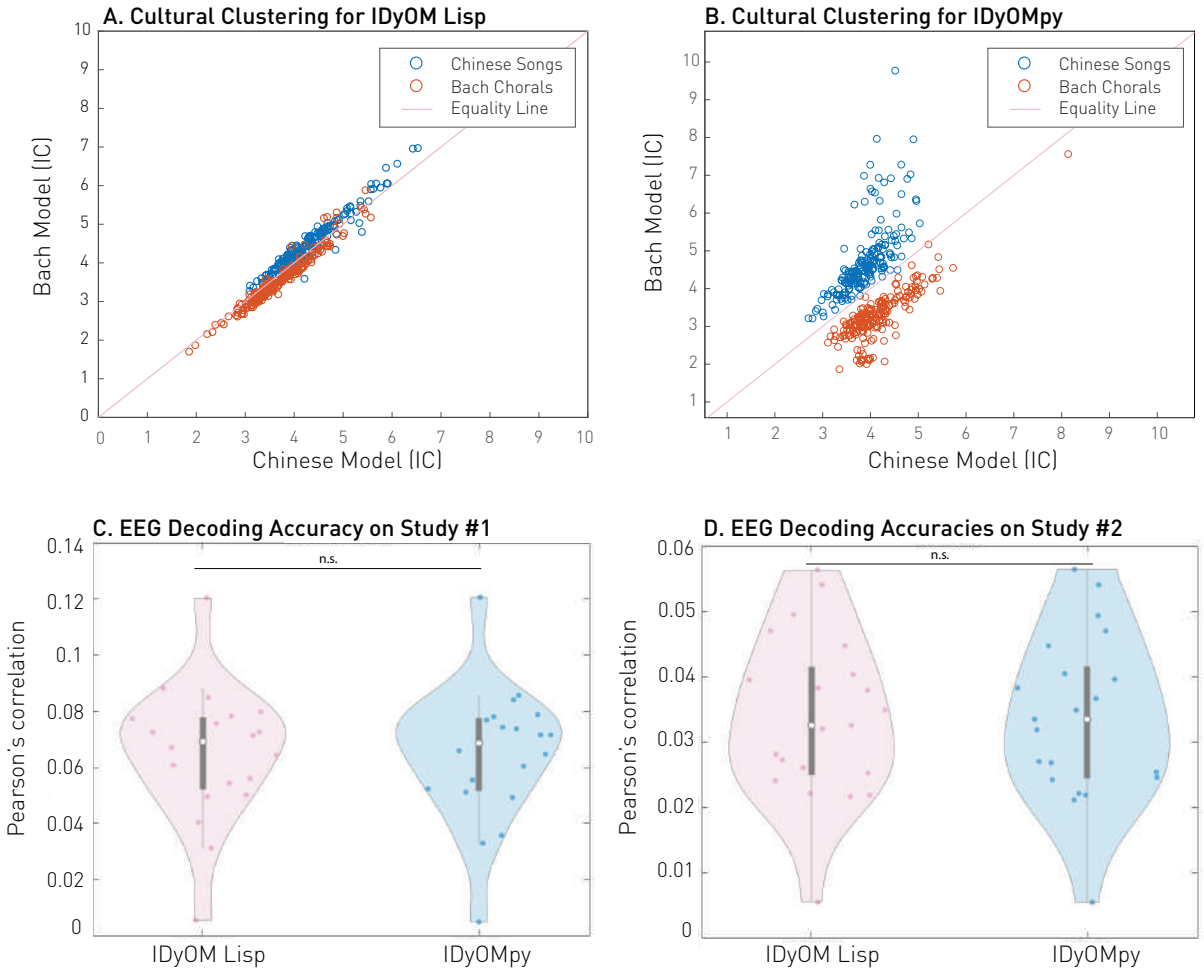


Figure 3.5. **Accuracies for cultural clustering and EEG decoding. A & B**: We plotted the piece-averaged IC for both a model trained on Shanxi traditional music (Chinese model) and a model trained on Bach chorals (Bach model) for both the Lisp and IDyOM implementations. We see that IDyOMpy outperforms the Lisp version in terms of cultural clustering. **C & D**: We used the mTRF toolbox to encode the IC from each model (IDyOM Lisp and IDyOMpy) trained on the same large Western database into EEG recordings of participants listening to Western music (not in the training dataset). We did not observe any significant difference between the models.

3.2.5 New Features

3.2.5.1 Missing Notes Detection

A recent study showed that it is possible to decode predictions from EEG recordings in intervals of musical silences during which the IDyOM model estimated a high probability of having a note. Moreover, the amplitude of those neural responses was correlated with the probabilities computed by the model(Di Liberto *et al.*, 2021). This analysis is replicated here using a new feature from the IDyOMpy implementation: the *missing notes detection feature*.

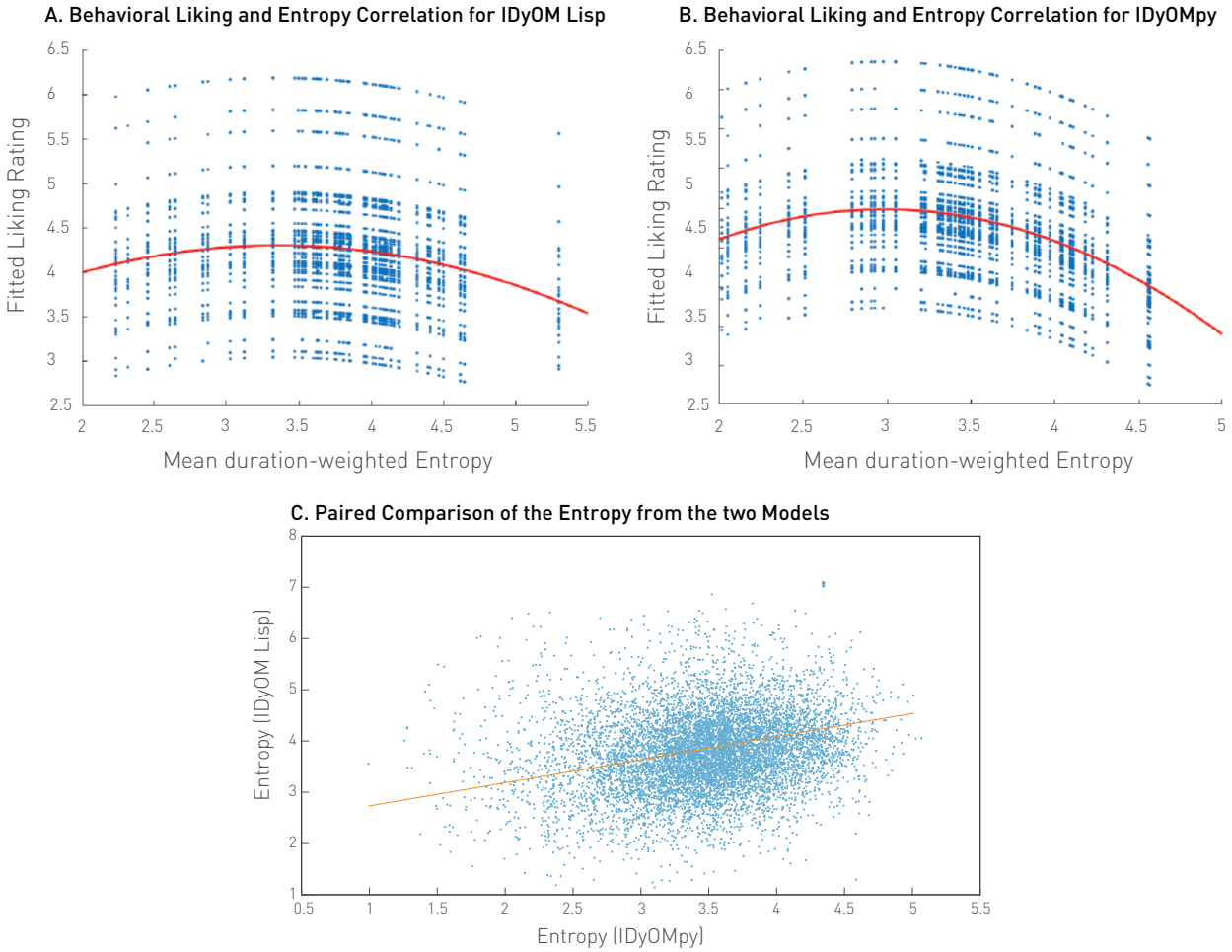


Figure 3.6. **Comparison and Validation of the Entropy. A & B**: Correlation of the Entropy from respectively IDyOM Lisp and IDyOMpy with the self-reported liking ratings from (Gold et. al., 2019). IDyOM Lisp explained 19% ($p = 0.005$) of the variance while IDyOMpy explained a significantly higher proportion of 22% ($p < 0.001$). **C** : Correlation of the Entropy for each note. Pearson's $r = 0.3$

This feature only uses the duration viewpoint and computes the probability distribution over the duration of each note. Therefore, we can compute the probability to have played a note during the natural silences between notes. Figure 3.7 shows examples of four Bach chorals ran with this feature.

3.2.5.2 Training Monitoring

Another new feature is the *Training Monitoring*. It allows monitoring of the training of the model. Therefore, one can assess the amount of data needed for model convergence. Also, since it is possible to initialize the model with another dataset, this feature is a good way to compare inter- and intra-variability between corpora. Figure 3.8 demonstrates results from two datasets of traditional Chinese music versus a large corpus of Western music. Finally,

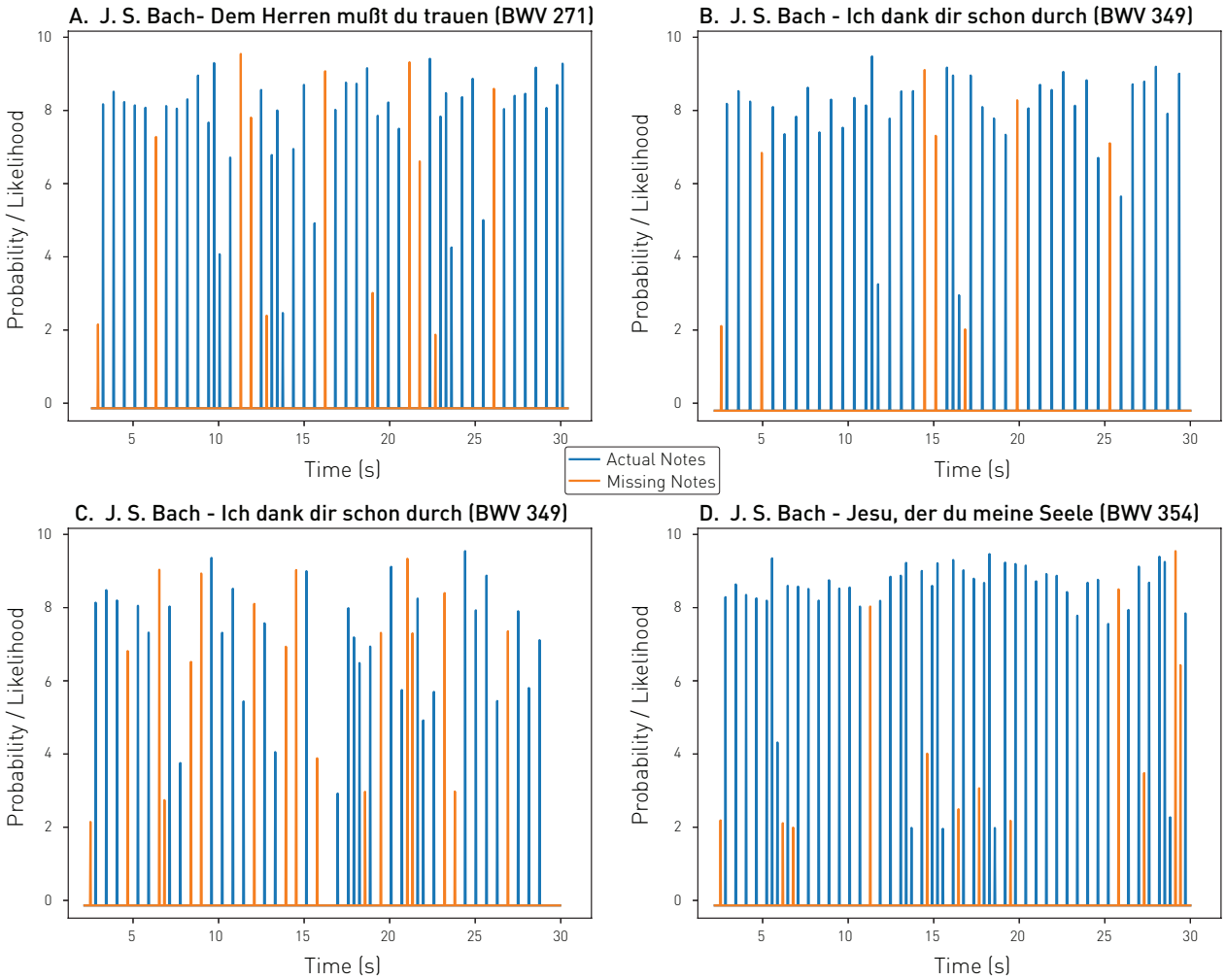


Figure 3.7. **Missing Notes Detection**: The model was run along the *duration* dimension only in order to compute the probability of having a note at each time step. A threshold probability of .2 was applied. Blue lines show the likelihood of the actual notes of the melody, and orange lines show the probabilities of the probable notes during silences detected by the model.

this monitoring can serve as a model for learning new music and musical enculturation as it simulates the learning of an unfamiliar musical grammar on top of an already familiar one. We used 2-fold cross-validation to compute the generalization errors, and it is possible to choose the number of pieces to test on. Note that the results may noticeably change depending on how the two sets are chosen (randomly done here). It is therefore recommended to compute the learning trace several times with different partitions of the data and take the average as the final trace.

3.2.6 Discussion

We have presented in this report IDyOMpy, a new implementation of the IDyOM using Python. This implementation differs in the way that the different Markov chains (for each

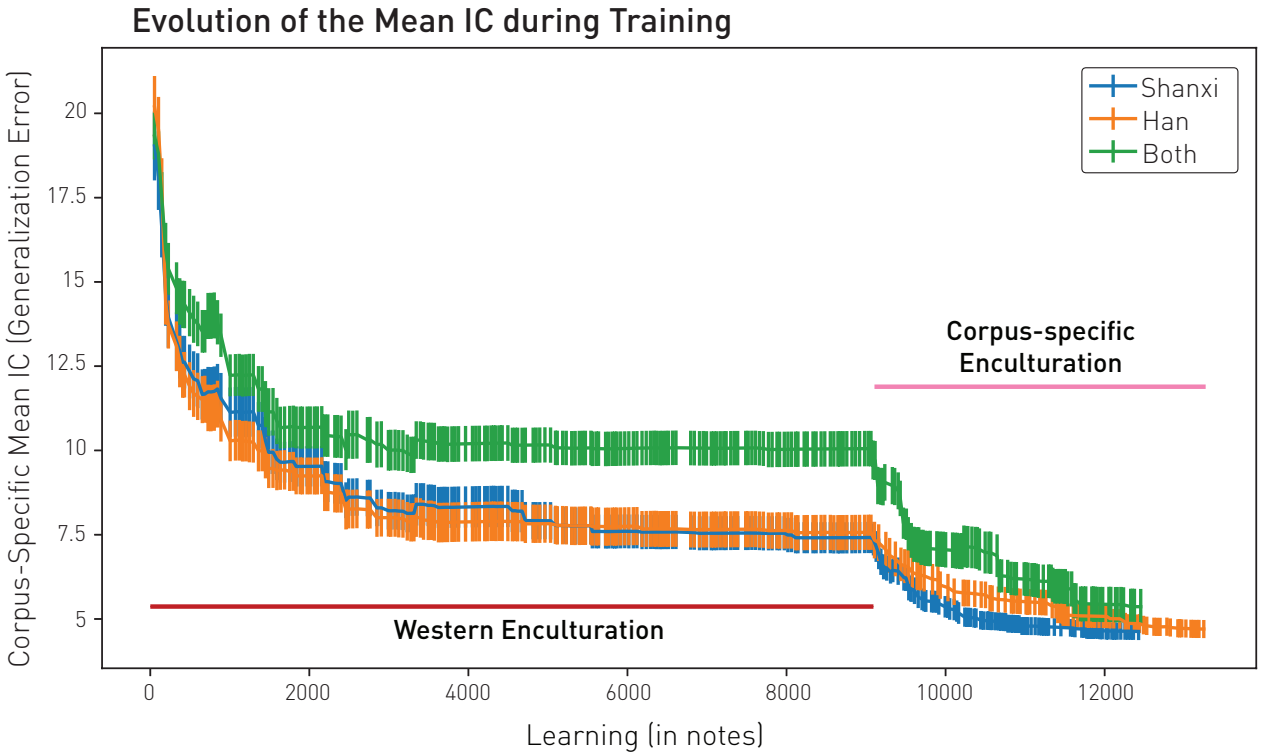


Figure 3.8. **Training Monitoring**: The model has been initially trained on a large database of Western melodies (Western Enculturation phase), then the model had been trained on 3 different corpora (Shanxi, Han, or a mix, all traditional Chinese music). Each line represents the generalization error on each specific corpus and shows when each model finished converging. How deep the line goes during the Western Enculturation indicates how much the Western corpus can account for Chinese grammar. If the corpus contains more variability (as the mix dataset) the line will be higher. How deep the line goes during the Corpus-Specific Enculturation indicates how much variability each specific corpus contains (aka how easily it is to generalize). Finally, the distance between the convergence plateau of the Western Enculturation and the Corpus-Specific Enculturation (in mean IC) indicates to which extent training the model with each specific corpus changed the model and therefore is a good proxy for how similar the corpora are with the initialization corpus.

order) are merged using an entropy-weighted average and not the PPM algorithm as in the Lisp version (c.f. 3.2.2.1). We also propose a way to approximate the entropy that reduces the computation time by at least a factor of 4 and does not significantly affect the results discussed in this study.

This new implementation generates overall comparable or superior results and allows for significant future improvements. We first showed that it performs better in terms of generalization errors, the amount of training data needed to converge the model, and cultural classifications. Additionally, we showed that the IC computed from the two models were relatively close ($r = 0.7$, c.f. Fig3.4) and resulted in comparable results for two EEG decoding experiments (c.f. Fig 3.5) thus confirming their physiological consistency.

Finally, we showed that, even if the entropies only weakly correlate between the two models ($r = 0.3$), IDyOMpy generates results with better correlation with the behavioral data of self-reported liking ratings(c.f. Fig 3.6), thus providing a cognitive validation of the model's outcomes. In addition, we presented two original new features (missing notes detection and training monitoring, c.f. Section 3.2.5).

Finally, to summarize, this Python implementation is generally easier to use and can be readily installed on any computer. But more significantly, it permits quick modifications as demonstrated by the two new features. We, therefore, believe that IDyOMpy will be a valuable tool of high interest to the community and will facilitate rapid progress in the field of computational music cognition.

3.3 Musi-Rex: a New Implementation of the D-Rex Model for Music Purposes [7]

3.3.1 Introduction

Predictions are often at the core of studies in music cognition. The Predictive Coding Framework(Clark, 2013; K. J. Friston *et al.*, 2010) is built on a theoretical basis (Koelsch *et al.*, 2019; M. A. Rohrmeier & Koelsch, 2012; Vuust, Heggli, *et al.*, 2022a) postulating that the brain builds a model of the world and then uses it to predict incoming sensory inputs. Perception would then emerge at the encounter between the sensory inputs and their predictions(Keller & Mrsic-Flogel, 2018), potentially generating a prediction mismatch (or "error") that is used to update the model(Näätänen *et al.*, 2007). Evidence for this *prediction* and subsequent adjustments and *learning* has been experimentally validated. For instance, explicit(Corrigall *et al.*, 2022; Fogel *et al.*, 2015; Morgan *et al.*, 2019; M. T. Pearce & Wiggins, 2006; Sears *et al.*, 2018) and implicit(Bianco *et al.*, 2019; Bianco *et al.*, 2020; Corrigall *et al.*, 2022; Di Liberto, Pelofi, Bianco, *et al.*, 2020; A. R. Halpern *et al.*, 2017; Marion *et al.*, 2021; Omigie, Pearce, *et al.*, 2019; Omigie, Pearce, & Stewart, 2012; Omigie *et al.*, 2013a; M. T. Pearce *et al.*, 2010; Politimou *et al.*, 2021; Quiroga-Martinez, C. Hansen, *et al.*, 2020; Quiroga-Martinez, Hansen, *et al.*, 2020) predictions have been shown to correlate well with the probability of musical events in the culture of the participants. Furthermore, cross-cultural studies have demonstrated that such predictions are consistent with the culture of the listeners (C. L. Krumhansl *et al.*, 2000). One study even discovered prediction signals at moments of silence in music that were correlated with the probability to have a note during those silences(Di Liberto *et al.*, 2021). Studies have also shown that even *passive* exposure to unfamiliar music engenders statistical learning that is consistent with the exposed music. For instance, passive exposure to Eastern music (chosen because of its unfamiliar time signatures) facilitates in young children and adults from Western musical cultures, the detection of violations in novel musical excerpts from the Eastern corpus(E. E. Hannon & Trehub, 2005b). Another study replicated this phenomenon for pitch. Adults gained better abilities to predict the next note in melodies sampled from random musical grammar after being passively exposed to different melodies sampled from the same musical grammar(Loui *et al.*, 2010).

Most such studies have used the widely available IDyOM model(M. T. Pearce, 2005) or its Python implementation (cite IDyOMpy). These models are based on variable-order Markov chains and have intrinsic limitations such as the assumption of independence between the

[7] Authors: Guilhem Marion*, Amélie Picard*, Benjamin Gold, Benjamin Skerritt-Davis, Mounya Elhilali, Shihab Shamma

musical dimensions (pitch, duration, intensity, timbre, space) and the discrete and symbolic nature of their input. An alternate approach is based on a Bayesian formulation known as the D-Rex model(N. Huang & Elhilali, 2017). It is reformulated here to allow for dependence between the dimensions and continuous inputs such as, eventually, audio spectrograms. Our implementation here is intended to behave in a similar manner to the IDyOM model, e.g., by considering the different dimensions of music (called viewpoints in the IDyOM model), by having short-term and long-term aspects, and by allowing for both k-fold cross-evaluation and train-test cycles.

We elaborate next on this model's Bayesian implementation and compare it to the results from the two previous IDyOM implementations (M. T. Pearce, 2005) through raw correlations of the Information Content (IC) and Entropy, cultural classification, EEG encoding, and correlation with behavioral data.

We shall begin by presenting the details of the new "D-Rex" implementation, referred to henceforth as the MusiRex model, followed by the results of the comparisons.

3.3.2 D-Rex and MusiRex

3.3.2.1 Definitions and Notations

We begin by highlighting four parameters to take into consideration:

- The current event: $x(t)$
- The past events that we assume the current event is dependent on:
  $x(t-1:t-D) = x(t-1), \ldots, x(t-D+1)$
- The context, e.g., the beginning of the piece: $\mathscr{C}$.
  We do not keep in memory the whole $x(0), x(1), \ldots, x(t-1)$
- The priors, that represent the culture and the musical environment of interest: π

At each iteration, we define a probability measure $\mathbb{P}_{\pi,\mathscr{C},x(t-1:t-D)}$ which indicates what we are expecting to hear. Then we compute $\mathbb{P}_{\pi,\mathscr{C},x(t-1:t-D)}(x(t))$:

$$\mathbb{P}(x) = \sum_i \alpha_i . P_i[x(t)|x(t-1:t-D)] \tag{3.4}$$

where the coefficients α_i and the probability functions P_i depend on π and $\mathscr{C}$, but not on $x(t-1:t-D)$

This finally gives the Information Content : $S(t) = -\log(\mathbb{P}(x(t)))$.

3.3.2.2 Training: priors vs context

The model functions by training on a musical dataset. Thus in equation 3.4, the probability distribution $\mathbb{P}(x)$ is decomposed into a sum of simple probability distributions P_i. Each term corresponds to a "learned context" $\mathscr{C}_i \in \pi$.

Training For each musical piece used in the training, we start with only one context. $\mathbb{P}(x(t)) = P_0(x(t)|x(t-1:t-D))$ where P_0 is a probability measure set with default parameters.

Then, at each time t, we:

- create a new context i that begins at time t, with current parameters;
- update the parameters of the older context, as we describe later;
- modify the coefficients α_i in order to increase the weight of the contexts that give the highest probability;
- save the current parameters; which will serve as the parameters of the contexts ending at time t.

In the end, we have multiple distributions P_i for each context i corresponding to an interval $[t_1, t_2]$ of the music piece.

Treatment of the priors We repeat this training with several music pieces and get a large set of probability distributions P_i. Some are very similar and hence can be collapsed together to simplify and reduce the set. Finally, we delete the distributions that have small weights due to the a_i-s).

Testing The resulting prior π is the set of distributions $\{P_i\}_i$ learned from the training corpus. We now consider the P_i-s as fixed during the listening. By contrast, the weight coefficients α_i are adapted during the listening, with the idea being that the listener tries to infer which of the learned situations fit the ongoing context $\mathscr{C}$. For example, in classical occidental music, there are major and minor modes, and a listener familiar with such classical music would start with $\alpha_{\text{Major}} = \frac{1}{2}$ and $\alpha_{\text{minor}} = \frac{1}{2}$. If the musical piece begins in the major mode, then α_{Major} increases; If, however, it changes to a minor mode at some time, α_{Major} would then decrease again.

3.3.2.3 Functioning

Each musical piece is composed of a sequence of notes with a few parameters for each note. Thus, $x(t)$ is a finite-dimensional vector:

$$x(t) = (x_1(t), \ldots, x_n(t))$$

with, for example, x_1 is its pitch, x_2 is duration, and x_3 the sound volume. The core implementation of MusiREX permits a dependence among these dimensions. However, in this particular implementation, where we seek to match the IDyOM implementation for comparison purposes, we shall assume the dimensions to be independent and hence can be treated separately such that we consider only mono-dimensional data.

Parameters of the P_i-s The P_i-s (see equation 3.4) are multi-dimensional Gaussian mixtures (*gmm*) with the following parameters:

- $k_i \in \mathbb{N}^*$ the number of components in the *gmm*
- $n_i \in (\mathbb{N})^{k_i}$ the number of samples used previously to compute the parameters σ and μ
- $\sigma_i \in (\mathscr{M}_{D,D}(\mathbb{R}))^{k_i}$ the covariance matrices for every Gaussian of the *gmm*
- $\mu_i \in (\mathbb{R}^D)^{k_i}$ the mean vectors for each gaussian in the *gmm*
- $sp_i \in [0,1]^{k_i}$ the weights of each component of the *gmm*

The probability measure therefore becomes:

$$P_i(\vec{x}) = \sum_{j=1}^{k_i} sp_i^j . P_i^j(\vec{x}) \tag{3.5}$$

where

$$P_i^j(\vec{x}) = \frac{(\frac{n_i^j+1}{2})!/(\frac{n_i^j}{2})!}{\sqrt{n\pi \det(\sigma_i^j)}} \left(1 + \frac{1}{n_i^j} < \vec{x} - \vec{\mu}_i^j | (\sigma_i^j)^{-1} | \vec{x} - \vec{\mu}_i^j > \right)^{-\frac{n_i^j - 1}{2}} \tag{3.6}$$

and $\vec{x} = x(t : t-D) = (x(t), \ldots, x(t-D+1))$

What is actually computed is $P_i(\vec{x}(t : t-D) | \vec{x}(t-1 : t-D))$, the conditional probability to have $x(t)$, since we already know $x(t-1), \ldots, x(t-D+1)$.

Updating the parameters After listening to $x(t)$, we update (or not, cf later) the parameters of the P_i-s as follows. First, the distribution is a Gaussian mixture, and hence the first step is to find the component of the gmm to which $x(t)$ belongs. Let $j_0 := \text{argmax}_j(P_i^j(x(t)))$ the index of the Gaussian where $x(t)$ belongs.

If $P_i^{j_0}(x(t)) < \beta$ (where β is a fixed threshold), then $x(t)$ doesn't belong to any of the gmm components, and thus we have to create a new component. In this case:

$$k_i \leftarrow k_i + 1 \tag{3.7}$$
$$n_i^{k_i+1} \leftarrow D \tag{3.8}$$
$$\sigma_i^{k_i+1} \leftarrow \sigma_{\text{default}} \tag{3.9}$$
$$\vec{\mu}_i^{k_i+1} \leftarrow \vec{x} \tag{3.10}$$
$$sp_i^{k_i+1} \leftarrow 1 \tag{3.11}$$

By contrast, no new component is needed if $x(t)$ already belongs to the mixture :

$$k_i \leftarrow k_i \tag{3.12}$$
$$n_i^{j_0} \leftarrow n_i^{j_0} \tag{3.13}$$
$$\sigma_i^{j_0} \leftarrow (1-w)\sigma_i^{j_0} + w.(z^2 I_D + (1+w)\delta\vec{x} \otimes \delta\vec{x}) \tag{3.14}$$
$$\vec{\mu}_i^{j_0} \leftarrow \vec{\mu}_i^{j_0} + w.\delta\vec{x} \tag{3.15}$$

where $\vec{x} = x(t : t-D) = (x(t), \ldots, x(t-D+1))$ and $\delta\vec{x} = \vec{x} - \vec{\mu}_i^j$

z is the noise parameter (which can be set to 0) and w is the weight that we give to the new observation

3.3.2.4 Spectro-REX model

Because this new implementation allows for working with dependant and continuous dimensions, the most striking application is to use spectrograms instead of symbolic (midi) representations of the music.

This model tries to deal with a more complete representation of sound: a spectrogram. In this case, we have

$$x(t) = (x_f(t))_{0<f\leq F_v/2} \tag{3.16}$$

$x(t)$ is a continuous-dimensional vector. In practice, we have to represent it with only a finite number of dimensions. But these dimensions are absolutely not independent. As we can see the spectrogram as a pitch distribution. So we can use the gmm distribution as defined previously, which we call the "expected distribution" in order to approximate the spectral distribution at each time point and compare it to $x(t)$ the "real distribution" to compute the prediction error. In this case, we, therefore, do not compute an Information Content or Entropy but the L2 distance between the predicted and actual spectral distribution. This computation also takes a long time compared to the symbolic version. We publish this implementation anyway

and do not show extensive benchmarks (EEG or behavioral validations) and only present the cultural distance as a proof-of-concept of the use of this model on audio data. We believe that the community will be able to take this implementation to the next level and provide sufficient validation.

3.3.2.5 Musical Dimensions

Music can be characterized across at least 5 dimensions (called viewpoints in the IDyOM framework): Pitch, duration, timbre, intensity, and spatialization. MusiRex readily operates on the axes of pitch and note durations (extracted from the midi files) and assumes their independence. Therefore, the final probability distribution is the joint product of the distributions of these two dimensions:

$$P(Z = z) = P(Pitch = z_{pitch}) \cdot P(Duration = z_{duration})$$

Output Features

Information Content The negative log-likelihood of a note n, referred to as *information content* (IC), represents how well the model predicts it given a context C. This computation is numerically stable with an interpretation in terms of compressibility which has been shown to correspond well to psychological interpretations of the perceptual data (Attneave, 1954; Chater & Vitányi, 2003). Eq. 3.3.2.5 relates the information content to the probability of an event as follows:

$$IC_M(n|C) = -log_2(P_M(n|C)) \tag{3.17}$$

Entropy *Entropy* provides an approximation of the uncertainty in the prediction. In information theory, this measure also reflects the information content in a signal. In the case of a probability distribution, it is associated with the flatness of the distribution generated by the model as an estimate of the confidence of the prediction. On one extreme, if all outcomes are equiprobable (the model has no information), the entropy is maximum and the prediction is highly uncertain. On the other extreme, if one outcome has a probability of 1 and all others 0, the entropy is minimum and the prediction is certain. Eq. 3.3.2.5 relates this entropy ($E = 0$) to a given probability distribution:

$$E_M(C) = \sum_n P_M(n|C) \cdot log_2(P_M(n|C)) \tag{3.18}$$

3.3.3 Methods For Benchmarking the Model

In order to compare this MusiREX to previous models, we define a few metrics that we will run on all three implementations (i.e., MusiREX, IDyOM, and IDyOMpy). We first present *theoretical measures* assessing how well each model generalizes to unseen data as well as *cognitive measures* through decoding EEG recordings from participants listening to music and correlation with behavioral data.

A general theoretical measure often used to benchmark statistical models is to compare the generalization errors (e.g., average error) computed on the same unseen data(C. A. Huang *et al.*, 2018). However, MusiREX is a *continuous* model, as opposed to the *discrete* implementations of IDyOM. Because the domain of the definition of the errors is completely different, comparing the generalization errors is impossible. However, because cultural distance is based on ratios of IC, it is a measure that we can use to compare across the models.

3.3.3.1 Cultural Distance

IDyOM has been shown to be a reasonable model for musical enculturation allowing for the assessment of cultural distances(M. T. Pearce, 2018). One way to measure the accuracy of a model is through the extent to which it can differentiate melodies taken from different cultures. Consequently, we shall train the models on two musical corpora: one is the Bach chorals and the other is traditional music from the Chinese region of Shanxi. We use test/train and cross-validation to compute the average generalization errors for all excerpts in both models. We then construct a scatter plot where the x and y axes are the generalization error for, respectively, the Shanxi and the Bach models, and where each point is a musical excerpt. A "bad" model would collapse all pieces on the equality line failing to separate the two cultures, whereas an excellent model would draw the 2 groups apart on either side of the line and thus classifying the two cultures.

To quantify the extent to which the two cultures are separated we defined three measures:

Inter-cultural distance (interCD) represents the average Euclidean distance between each point of the first culture and each point of the second culture. A value of 0 means that all points collapse, the bigger the value the further the two cultures are in the model space.

Intra-cultural distance (intraCD) represents how close the pieces are within a culture, it is a proxy for the variability in generalization error and the stability of the model. Small values mean a more stable model (less variance).

Clustering index $= \frac{interCD}{intraCD(A)/2+intraCD(B)/2}$ combines both inter- and intra-cultural distances into a composite measure that tells to which extent it is easy to classify the two cultures.

3.3.3.2 EEG Decoding

IDyOM has been widely used in studies of the psychology and neuroscience of music, especially recently in decoding EEG recordings that allow for a physiological benchmarking of the model. Therefore, we used data from two recent studies (Study #1(Di Liberto, Pelofi, Bianco, *et al.*, 2020) and Study #2(Marion *et al.*, 2021)) that employed IDyOM to decode such EEG data. We compared the results of the analyses using the three models of interest. The experiments used a Biosemi Active Two 64-electrodes System, where the recordings for Study #1 were digitally filtered between 1 and 8 Hz using a 2nd-order Butterworth zero-phase filter and down-sampled to 64 Hz, and bandpassed between 0.1 Hz - 30 Hz for Study #2. EEG channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation. All channels were then re-referenced to the average of the two mastoid channels for Study #1 and using global re-referencing for Study #2. The stimuli were 10 Bach partitas in Study #1 and 4 Bach chorals in Study #2.

The analysis was conducted in a similar fashion as in the original studies by estimating temporal response functions (TRFs)(N. Ding & Simon, 2012; Lalor *et al.*, 2009b) with the mTRF-Toolbox(Crosse, Liberto, & Lalor, 2016). This de-convolution method (implemented as a lagged linear regression) was used to regress the IC signal computed by all models with the pre-processed EEG recordings using cross-evaluation. Pearson's correlation was computed between the predicted and original EEG signals. Since the predicted EEG signal was constructed from the IC signals from the models, the correlation then estimates the resemblance between the IC and the EEG recording. An IC signal that more accurately matches human perception is expected to generate better EEG predictions (and hence correlations), thus providing a direct way to estimate the physiological relevance of each model.

3.3.3.3 Behavioral Preference

A recent study(Gold, Pearce, *et al.*, 2019) demonstrated that the Entropy from the IDyOM model could explain 19% of the variance of 44 participants' behavioral liking measured by means of a 7-item Likert scale on 57 stimuli. Stimuli reported to be familiar to the participants were excluded from the analysis. There was a significant Wundt (quadratic correlation, a.k.a. inverted-U shape) effect between the preference ratings for the songs and the mean duration-weighted Entropy of the same songs. We, therefore, used these data as a way to estimate the validity of the entropy computed by our model. To do so, we replicated the results of this study on the same data but using our model trained on the same corpus. We then compared the

explained variance r^2 using both models. In order to compute the significance of the difference between them, we computed the distribution for each model using the Bootstrap method. We then estimated the r^2 of the sub-sampled data (80% sampled from both participants and songs) 5000 times using the same indices for each model. We then computed the difference distribution and computed the p-value for its inferior or equal to 0. These p-values are reported below in the **Results** section.

3.3.4 Results

We first compared the raw IC and Entropy from all models and computed their Pearson correlation as in Figure 1. We found weak correlations with both IDyOM LIsp and IDyOMpy (respectively $r = 0.45$ and $r = 0.43$) for the IC, and similarly weak correlations for the Entropy (respectively $r = 0.4$ and $r = 0.1$). Those results suggest that MusiREX computes the IC and the Entropy in a different way from the IDyOM implementations, potentially suggesting new insights into music cognition. To test this proposition, we computed a further set of measures to determine the extent to which MusiREX can replicate previous results generated with IDyOM.

	Inter-Cultural Distance	Intra Cultural Distance on A	Intra Cultural Distance on B	Clustering Index
IDyOM Lisp	1.8245	1.3033	1.4054	1.3471
IDyOMpy	2.0474	1.3361	1.117	1.6693
MusiRex	2.4846	0.92071	0.79838	2.8906
IDyOMpy (10 pieces)	1.7518	1.3438	1.7433	1.1349
Spectro-REX	0.97958	0.87511	0.35285	1.5955
Spectro-REX (timbre)	1.6212	1.1798	0.4102	2.0392

Table 3.2. **Cultural Classification Metrics for All Models.** The metrics are defined in 3.2.

We plotted the cultural distance between the Chinese music and Bach chorals for the various models(**Fig. 2**. MusiREX separated or clustered the two corpora better than both IDyOM and IDyOMpy, regarding the inter-cultural and intra-cultural distances (Table 1), demonstrating that it is an excellent model for such assessment among different datasets. We also used the spectral version of D-REX on piano-synthesized versions of 10 pieces from the Bach and Chinese datasets. We unfortunately were not able to run the analysis on the full dataset because of computational time (it took about 2 days to run for only 10 pieces). We ran 3 experiments, we computed the cultural distance using IDyOMpy using only 10 pieces per dataset; we computed the cultural distance on Spectro-REX on the exact same audio-generated pieces; finally, we audio-generated the 10 Bach pieces but using two different instruments (acoustic and electric pianos). The results (c.f. Fig. 3.11) show that SPECTRO-REX, in addition to providing an excellent timbre separation (no overlap between the datasets), also provides a better separation

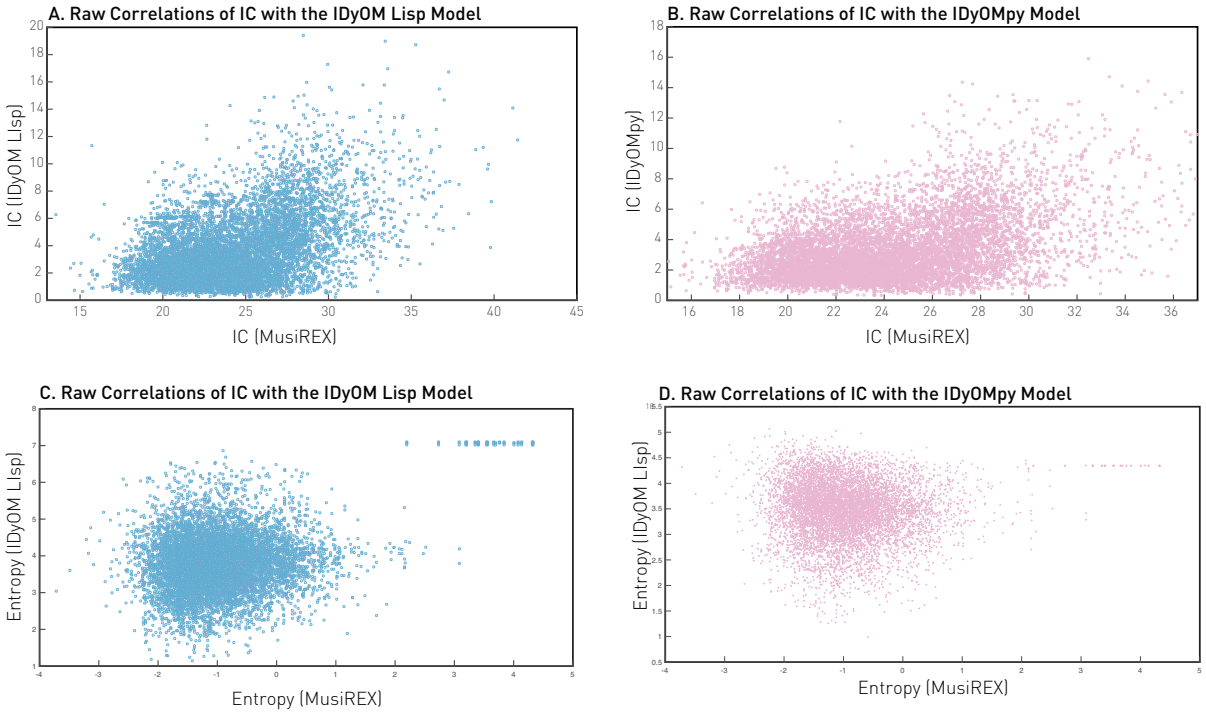


Figure 3.9. **Raw Comparisons of IC and Entropy with IDyOM and IDyOMpy. A & B**: The raw correlation of the IC produced a correlation of $r = 0.45$ with IDyOM (A) and $r = 0.43$ for IDyOMpy (B). **C & D**: The raw correlation of the Entropy was $r = 0.4$ with IDyOM and $r = 0.1$ for IDyOMpy.

than IDyOMpy between the 10 Bach and Chinese songs. This is striking evidence that this model could pretend to be the next generation of statistical models of music. Still, it requires improvement in order to be run on full datasets and intensive benchmarking as we provide here for Musi-REX.

We used the mTRF toolbox to predict EEG recordings of participants listening to Western music (in two studies, c.f. 3.3.3.2) from the IC signal computed with the three models. There were no significant differences in the accuracy for Study #1. However, we found a significantly enhanced accuracy for MusiREX over the two other implementations of IDyOM in Study #2, confirming the physiological utility of MusiREX.

To test a cognitive measure of the Entropy computed from all models, we used data from (Gold et. al., 2019) to determine the variance-explained of the behavioral liking ratings (c.f. 3.3.3.2 for method) by each model. MusiREX explained 19% of the variance, exactly the same as with the Lisp implementation of IDyOM. Both however were significantly lower than IDyOMpy (22%, **Fig. 4.A**). The significance was computed using the bootstrap method to generate the distributions of the explained variances (**Fig. 4.B**) which were found to be significantly different between MusiREX and IDyOMpy ($p < 0.0001$) but not between MusiREX and IDyOM Lisp ($p = 0.2$). This suggests that even if MusiREX produces different values (e.g., the shape of the correlation with the behavioral data is different), it is still as predictive of self-reported

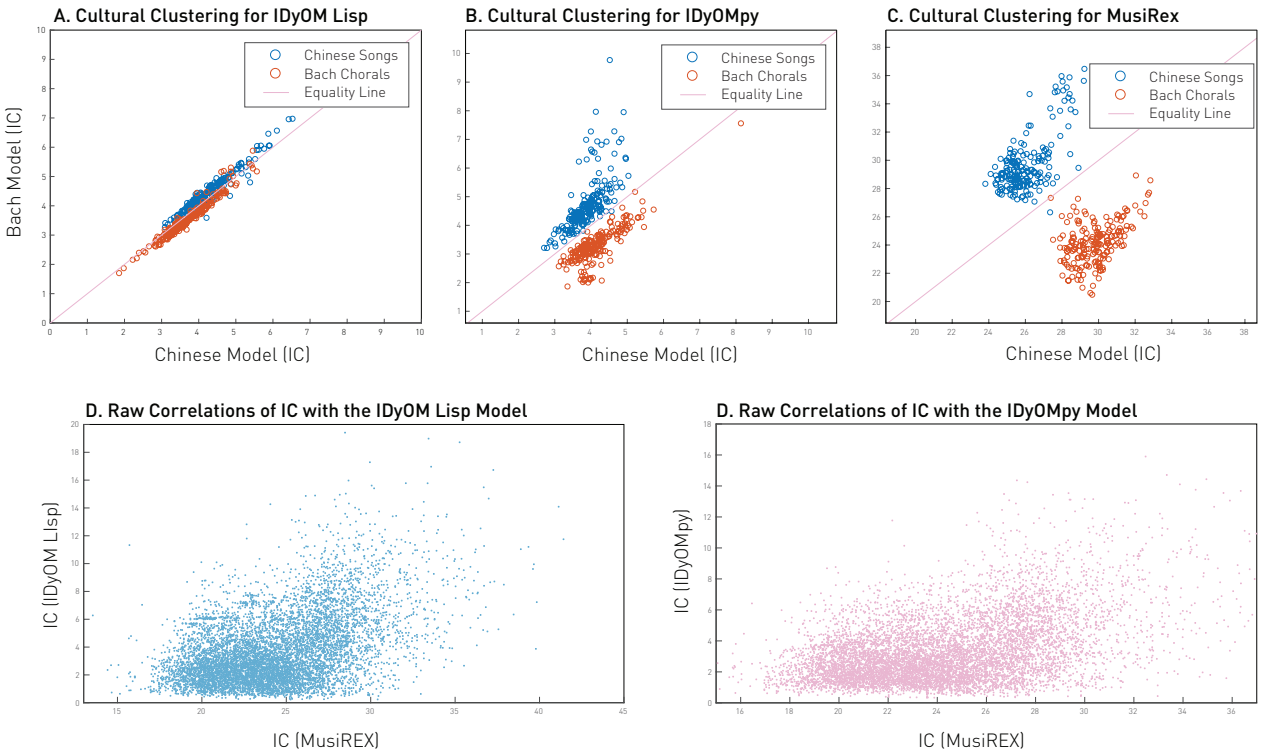


Figure 3.10. **Cultural Distances.** Excerpt-averaged ICs for models trained on traditional Chinese music (Chinese model) and on Bach chorals (Bach model). MusiREX outperforms the other models (c.f. Table 3.2 for precise metrics).

pleasure as the original implementation of IDyOM.

3.3.5 Discussion

We have presented MusiREX, a Bayesian-based musical model that can capture the *continuous* information flow of such signals along multiple dimensions, learn from corpora of musical excerpts, then provide predictions based on the long- and short-time contexts of the signal, and hence can be used in a diverse range of experimental and theoretical investigations. The results presented here in comparison with previous IDyOM implementations exhibit relatively different trends suggesting that MusiREX's Bayesian nature allows for computing different statistics. Moreover, MusiREX demonstrated better cultural classification than both previous IDyOM implementations, showing that it can be reliably used as a model for musical culture in cross-cultural studies. Finally, in analyses of EEG recordings, MusiREX performed similar or superior decoding of EEG recordings of participants listening to music, thus providing a physiological validation. MusiREX, however, is unique in its ability to process continuous audio data and consider multiple interdependent dimensions, both being the biggest limitations of IDyOM models. We believe that offering here the MusiREX implementation will enhance the use of statistical modeling in music cognition and cognitive neuroscience of music, as well as generate new

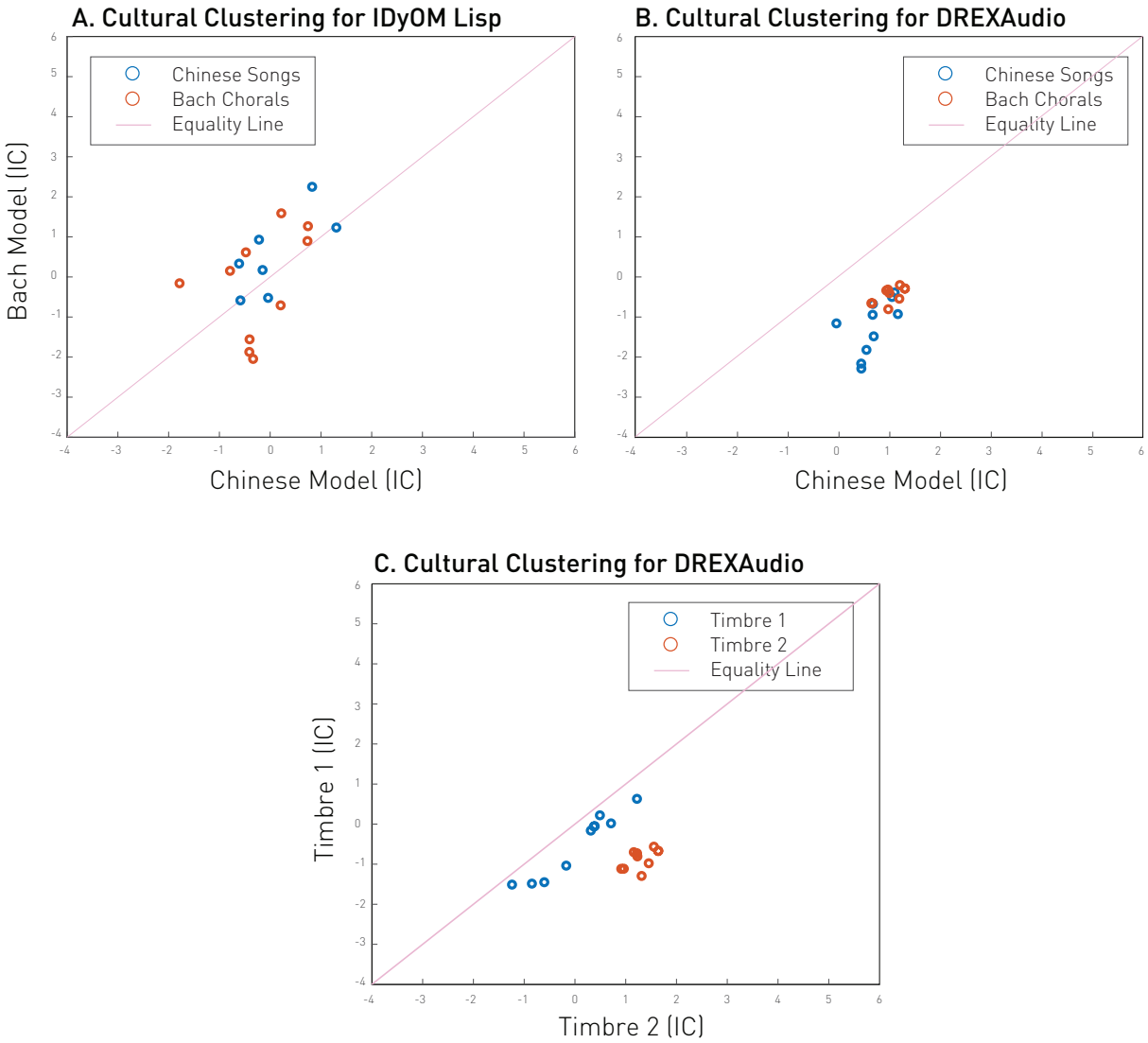


Figure 3.11. **Cultural Distances for Spectro-REX. A:** Excerpt-averaged ICs for models trained on traditional Chinese music (Chinese model) and on Bach chorals (Bach model) using the midi versions and IDyOMpy (control). **B:** Excerpt-averaged ICs for models trained on traditional Chinese music (Chinese model) and on Bach chorals (Bach model) using the audio versions and Spectro-REX. **C:** Excerpt-averaged ICs for models trained on Bach chorals played on an acoustic piano (timbre 1) and on Bach chorals played on an electric piano (timbre 2) and Spectro-REX. Spectr-REX outperforms the symbolic IDyOMpy but shows an irregular trend (the Chinese model is always better than the Western one) and shows very excellent separation for the timbers (c.f. Table3.2 for precise metrics).

creative uses in the field.

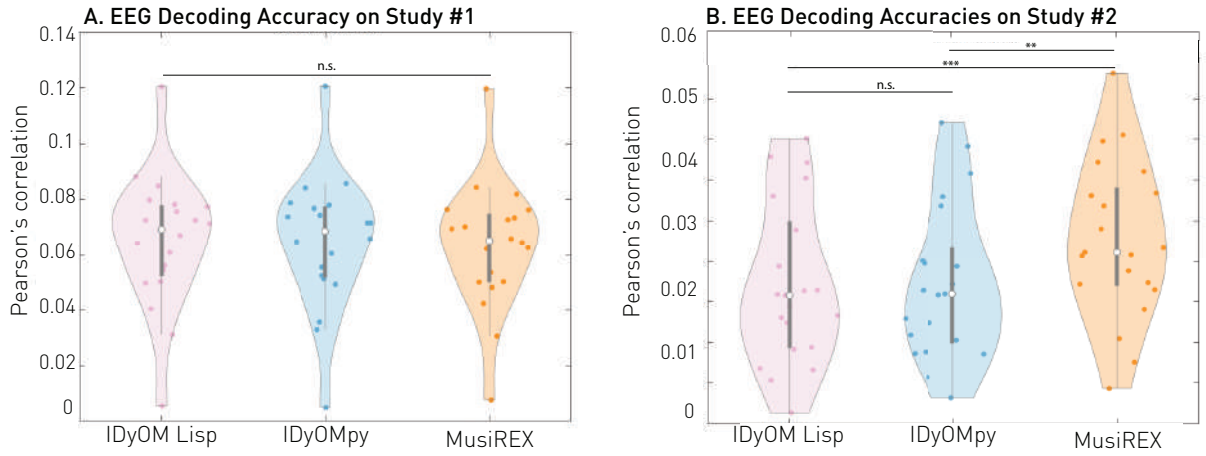


Figure 3.12. **EEG Decoding Accuracies.** MusiREX significantly outperforms both IDyOM ($p < 10^{-18}$) and IDyOMpy ($p < 10^{-12}$) in terms of EEG decoding accuracies.

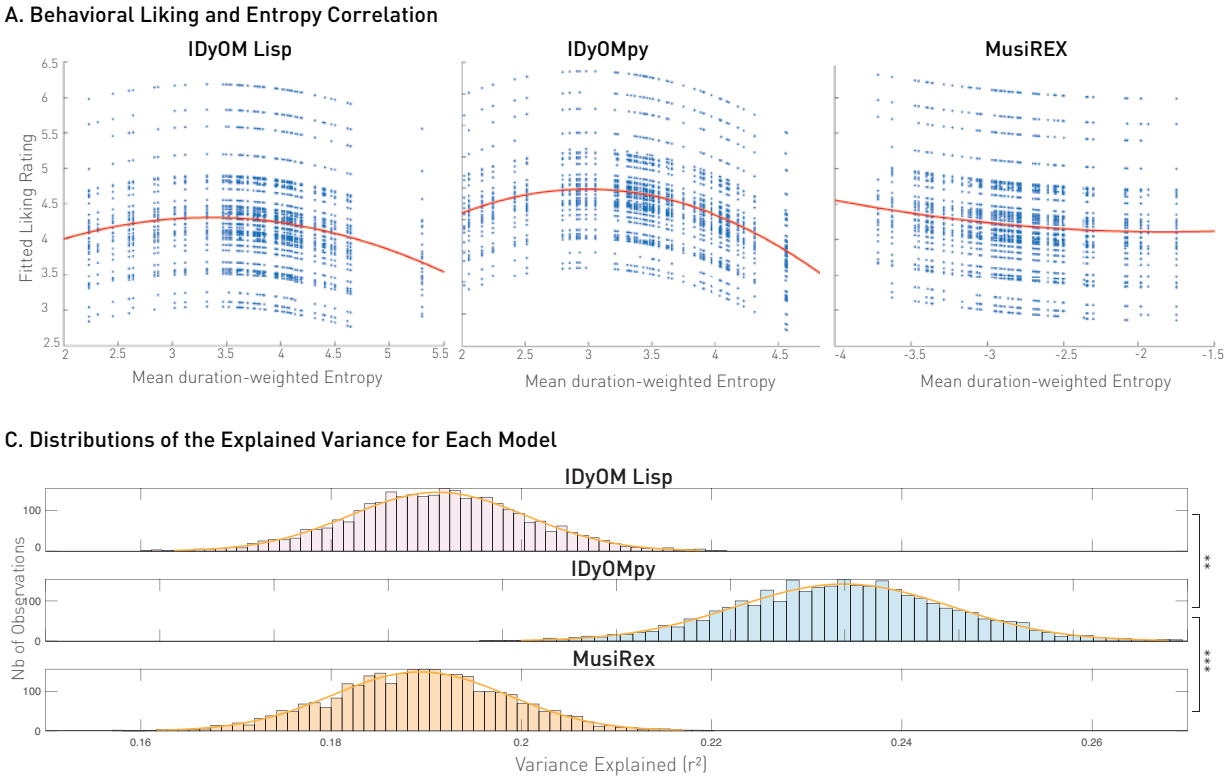


Figure 3.13. **Entropy and Self-Reported Pleasure Correlation.** **A**: Correlation of the Entropy from all models with the self-reported liking ratings from (Gold et. al., 2019). **B**: MusiREX and IDyOM Lisp explained 19% of the variance while IDyOMpy explained a significantly higher proportion of 22% ($p < 0.0001$).

3.4 General Discussion

This section described two new statistical models of music along with their physiological and behavioral validations. We isolated IDyOM as the most used model of music in the field of

music cognition and explicit its two main limitations which are the inaccessibility of the Lisp programming language in our community as well as the fact that it is only suited for symbolic data (c.f. 3.1.3). We have then presented two new models that overcome those limitations.

First, we have presented IDyOMpy, a new implementation of the IDyOM using Python. This implementation differs in the way that the different Markov chains (for each order) are merged using an entropy-weighted average and not the PPM algorithm as in the Lisp version (c.f. 3.2.3.1). We also propose a way to approximate the entropy that reduces the computation time by at least a factor of 4 and does not significantly affect the results discussed in this study. This new implementation generates overall comparable or superior results shown by means of theoretical measures (c.f. 3.1.2) or physiological measures. In addition, we presented two original new features (missing notes detection and training monitoring, c.f. 3.2.5). Finally, to summarize, this Python implementation is generally easier to use and can be readily installed on any computer. But more significantly, it permits quick modifications as demonstrated by the two new features.

We also have presented MusiREX, a Bayesian-based musical model that can capture the continuous information flow of such signals along multiple dimensions, learn from corpora of musical excerpts. It, as IDyOM, provides predictions based on the long- and short-term models (c.f. 3.2.2.1), and hence can be used in a diverse range of experimental and theoretical investigations. The results presented here in comparison with previous IDyOM implementations exhibit relatively different trends suggesting that MusiREX's Bayesian nature allows for computing different statistics. Moreover, MusiREX demonstrated better cultural classification than both previous IDyOM implementations, showing that it can be reliably used as a model for musical culture, and especially, in cross-cultural studies. Finally, in analyses of EEG recordings, MusiREX performed similar or superior decoding of EEG recordings of participants listening to music, thus providing a physiological validation. MusiREX, however, is unique in its ability to process continuous audio data and consider multiple interdependent dimensions, both being the biggest limitations of IDyOM models. We, however, hit a computational limitation about continuous audio data. Because of the high dimensional nature of those data, the algorithm requires a lot of resources in terms of computational power which restricts its use to small datasets. We are in the process of optimizing the computations in GPU to benefit from the intense parallelization of graphical computing in order to reduce the computations time.

Because those two new models allow for overcoming the previous limitations of IDyOM we believe that those tools will be valuable to the community and will facilitate rapid progress in the field of computational music cognition. Still, we think that we need more energy dedicated to those models to make them more accurate, publish more cognitive validations, and make them more usable by the community.

4 The Neural Underpinnings of Musical Enculturation and its Link to Musical Preferences

4.1 General Introduction to Musical Enculturation: Learning to Enjoy Music

4.1.0.1 Learning to Predict

An essential aspect of the human condition is their evolving environment, which raises the question of how individuals' cognitive systems cope with this constant change. Whether considering the experience of moving to another country and being immersed in an altogether radically new culture, or more subtle changes such as seeking exposure to a new language, a musical style, or a type of cuisine, human beings are constantly challenged with new experiences. Because human cultures are carved by norms and conventions, novel exposure to an estranged culture induces a type of learning that is often described as implicit: When exposed to a set of stimuli constrained by unspoken rules, cognitive systems build up a mental representation of the underlying grammar. The learning of these grammars speaks for how much enculturation is continually occurring. This type of learning undoubtedly constitutes one of the essential aspects of human cognition (Fiser & Aslin, 2001; 2002; Perruchet & Pacton, 2006).

Music cognition offers a uniquely compelling opportunity to investigate the computational and neural basis of enculturation: music is ubiquitous, produced and enjoyed in all known human cultures (Nettl, 2015; Reck, 1977; Stevens, 2004), displays varying structural norms across cultures (Castellano *et al.*, 1984; Fourer *et al.*, 2014; C. L. Krumhansl *et al.*, 2000; Polak *et al.*, 2018), and humans spontaneously seek musical experiences. It has been shown that children naturally learn the music of their own culture, much as they learn their native tongue (Snow, 1972), through a mixture of active engagement and passive exposure and without requiring explicit instructions (Campbell, 2010; Henrich, 2008). In fact, responses indicative of one's own musical systems can be measured as early as 12 months of age (E. E. Hannon & Trehub, 2005a; 2005b; Lynch & Eilers, 1992; Schellenberg & Trehub, 1999) reflecting varied structural aspects of the musical culture, such as consonance (Trainor *et al.*, 2002) and rhythms (Trehub & Hannon, 2006). Much like sensitivity dynamics to native language (Kuhl, 2004), children exhibit reduced sensitivity to out-of-culture structural features (E. E. Hannon & Trehub, 2005a; Lynch & Eilers, 1992) and enhanced sensitivity to native ones (Politimou *et al.*, 2021). Critically, enculturation to novel musical systems persists at later stages of life, although it may rely on partly different cognitive mechanisms (J. S. Johnson & Newport, 1989). Nevertheless, it has been shown that adults can learn novel and entirely unfamiliar new musical systems by mere passive exposure (Loui & Wessel, 2008; Loui *et al.*, 2006; 2010), which can even subsequently modulate tonal expectations (C. L. Krumhansl *et al.*, 2000; Oram *et al.*, 1995).

Listening to music is an inherently active cognitive process, as we predict upcoming musical

events in terms of how we might generate it ourselves (Fogel *et al.*, 2015) (c.f. chapter 2). This presupposes the existence of an internal model of the musical syntactic structure that guides the listener's expectations –or *predictions*– of the next notes. This is evidenced by studies using priming paradigms and response times (RTs) to investigate the relative predictability of specific musical notes or chords found that RTs were correlated with expectation –low expectation yielding slower RTs– for both musicians and non-musicians (J. J. Bharucha & Stoeckig, 1987; Bigand & Pineau, 1997; Tillmann *et al.*, 2006; 2007). Such behavioral evidence suggests that musical enculturation essentially implicates implicit learning of the statistics of a musical corpus (Bigand & Poulin-Charronnat, 2006; M. Rohrmeier & Rebuschat, 2012; M. Rohrmeier *et al.*, 2011). Studies indeed show that listeners are able to acquire the regularities of new artificial musical systems by just being exposed to them (Loui & Wessel, 2008; Loui *et al.*, 2006; 2010; M. Rohrmeier & Cross, 2009; 2013; M. Rohrmeier *et al.*, 2011) and that the same ability also holds for musical systems with microtonal tuning (Leung & Dean, 2018) and out-of-culture musical systems (M. Rohrmeier & Widdess, 2017). A recent study reported that adult participants were even able to learn the regularities of two artificial musical grammars by just listening to them (Guillemin & Tillmann, 2021), a learning process that is known to be preserved in *amusic* listeners (Omigie & Stewart, 2011), suggesting that it exploits a domain-general mechanism. Cross-cultural studies on statistical properties of pitch sequences have also consistently revealed culture-specific responses to one's own musical culture across varied musical cultures, both with musicians: German, American, and Hungarian (Unyk & Carlsen, 1987), Finnish and Western (C. L. Krumhansl *et al.*, 1999), and North Sami Yoiks and Western (Eerola *et al.*, 2009; C. L. Krumhansl *et al.*, 2000), and non-musicians: Western and Balinese (Kessler *et al.*, 1984), and Chinese and American (C. L. Krumhansl, 1995) listeners.

Therefore, there is ample and consistent evidence that exposure to the evolving musical environment triggers changes in response to the newly acquired musical systems and that adjusted computational models could model those changes(Tillmann, Bharucha, & Bigand, 2000).

4.1.0.2 Learning to Enjoy

Music theorists have postulated that making predictions and experiencing pleasure during music listening go hand in hand (Meyer, 1956). This translates into a non-linear inverted U-shape distribution, meaning that too simplistic or too complex musical excerpts are associated with less musical enjoyment, in line with the Wundt effect (Berlyne, 1971; Chmiel & Schubert, 2017; Huron, 2006), as illustrated in Figure 4.2. Supporting this, two recent studies applying IDyOM (Cheung *et al.*, 2019; Gold, Pearce, *et al.*, 2019) show a non-linear relationship between the computational modeling of musical expectations and self-reports of musical pleasure.

A large body of evidence has demonstrated that dopaminergic regions (the ventral stria-

tum and caudate in the basal ganglia and the substantia nigra/ventral tegmental area in the midbrain) are activated by pleasurable music (Blood & Zatorre, 2001; Ferreri *et al.*, 2019; Koelsch, 2020; Mas-Herrero *et al.*, 2021; Salimpoor *et al.*, 2011). Consequently, recent models of musical pleasure have attempted to bind these two threads together, suggesting that musical pleasure emerges from the increased activation of these dopaminergic and reward-related regions (Salimpoor *et al.*, 2015).

Recent findings also suggest that dopaminergic neurons projecting to the ventral striatum encode a reward prediction error driven by musical stimuli (Gold, Mas-Herrero, *et al.*, 2019). This study used a decision-making task in which participants learned which cues led to more probable endings. Critically, the authors reported that the prediction error generated by unexpected events correlated with brain activity in the ventral striatum. Another study, supporting the same hypothesis (Shany *et al.*, 2019), showed that self-reports of musical expectation (using subjective surprise behavioral ratings C. L. Krumhansl, 1997) correlated with self-reports of musical pleasure and importantly, also with increased activity in dopaminergic and reward-related structures (including the ventral striatum). On the other hand, a recent study provides a more nuanced view and calls for a different role of the ventral striatum in the network responsible for musical pleasure. Using IDyOM to compute musical expectations in sequences of chords, Cheung and colleagues argue that the ventral striatum plays an ancillary role in the generation of the pleasurable experience (Cheung *et al.*, 2019). According to this study, the dopaminergic ventral striatum encodes the degree of uncertainty and modulates attention deployment in the amygdala, hippocampus and the auditory cortex, the critical regions that encode musical predictions.

Even if the community is unable to agree on the neural implementation of the prediction-to-pleasure relationship, it is clear that there is a relationship between i) the statistical structures of the musical stimuli, ii) the prediction error, and iii) the musical self-reported pleasure. We can therefore draw a clear red line between musical enculturation and musical enjoyment through this story: Our brain models a prediction model of the musical environment based on the music we have been exposed to. This prediction model is later used to predict the next incoming music. The prediction error, which assesses the extent to which our inner model of consistent with the new incoming stimuli, is encoded in the auditory cortex (c.f. chapter 2) and probably sent to the reward areas possibly in the form of the uncertainty of the events in the case of the ventral striatum. The activation of those neurons generates musical pleasure that could be self-reported by listeners in the form of an inverted U shape. It is therefore our internal model of musical predictions that shapes our perception toward liking certain songs and not others. This is why we like to call musical enculturation the process of learning to enjoy music.

4.1.0.3 Limitations and Scientific Contribution

However, the community lacks evidence for a within-subject link between enculturation and self-reported pleasure. Other questions are crucial for the understanding of the building of musical preferences such as the persistence of the enculturation mechanisms, and the associated neural markers.

We, therefore, propose here a set of projects that investigate the neural underpinnings of Musical Enculturation. Those projects are divided into two panels: i) an electro-physiology panel using EEG in humans and ECoG in ferrets and ii) a brain imaging panel using fMRI in humans and FUS in ferrets. Having data from both ferrets and humans allows having a better and more precise view of the neural mechanisms (using invasive recordings in ferrets, which is usually not possible in humans) supporting musical enculturation.

As some studies of this project are still ongoing and unpublished this chapter presents the already collected data and preliminary analyses. I designed the experimental protocols along with the persons who collected the data (credited in each section) and Claire Pelofi and I conducted the analysis of the data and the generation of the figures (except with explicit mention). Shihab Shamma supervised the scientific process and will proofread the manuscripts of all those studies.

4.1.1 EEG and Self-Reported Pleasure in Humans (recorded by Guilhem Marion & Camille Barbarot at ENS, Paris)

4.1.1.1 Method

Figure 4.1. Schematic presentation of the EEG experiment.

We recruited 34 French participants living in Paris (mostly non-musicians) and divided them into two groups: a control group (15 participants) and a test group (19 participants). The experiment consisted of 3 EEG recordings and 1 at-home exposure phase.

Recording #1 Participants come to the lab and listen to about 30 minutes of unfamiliar Chinese music from the region of Shanxi. Their brain is recorded by the EEG system during the entire experiment. After each song, they are asked to self-report the amount of pleasure they feel while listening to the song. They are able to take a break. At the end of the

recording, they are randomly assigned to the control or test group and are given access to a lab-made streaming platform in which we can monitor what they can listen to.

Exposure phase Each participant is asked to listen to at least 30 minutes each day for 2 weeks. The test group has access to unknown songs (not present in the recording #1) of Chinese music from the region of Shanxi (same genre as during the recordings) played on a Guzheng (Chinese instrument that was also used for the EEG stimuli). The control group has access to Bach chorals (assumed to follow the same musical structures as the participants have been exposed to throughout their life) played on the Guzheng. Therefore, we control for the behavior of listening to music on our streaming platform and the timbre of the songs. The only difference between the control and the test group is the musical structures present in the songs they are exposed to.

Recording #2 After 2 weeks of exposure, participants come back to the lab and undergo the same procedure as for recording #1. We add 15 minutes of new songs (not played during recording #1 nor in the exposure) to check for the potential remembering of the songs already played during recording #1.

Resting phase Participants (of both groups) do not do anything for 2 months and can pursue their normal behavior of music listening.

Recording #3 After the 2-month resting phase, participants come back to the lab and undergo the same procedure as for recording #2. We add 15 minutes of new songs (not played during recording #1, nor recording #2, nor the exposure) to check for the potential remembering of the songs already played during recording #1 or recording #2.

A schematic description of the experiment is given in Figure 4.1.

4.1.1.2 Results

Self-reported Pleasure Ratings Music theorists have postulated that making predictions and experiencing pleasure during music listening go hand in hand (Meyer, 1956). This translates into a non-linear inverted U-shape distribution, meaning that too simplistic or too complex musical excerpts are associated with less musical enjoyment, in line with the Wundt effect (Berlyne, 1971; Chmiel & Schubert, 2017; Huron, 2006), as illustrated in Figure 4.2. Supporting this, two recent studies applying IDyOM (Cheung *et al.*, 2019; Gold, Pearce, *et al.*, 2019) show a non-linear relationship between the computational modeling of musical expectations and self-reports of musical pleasure.

It is therefore assumed that while rating new unfamiliar music participants will fall into the right side of the inverted-U shape and therefore won't feel the maximal pleasure. However,

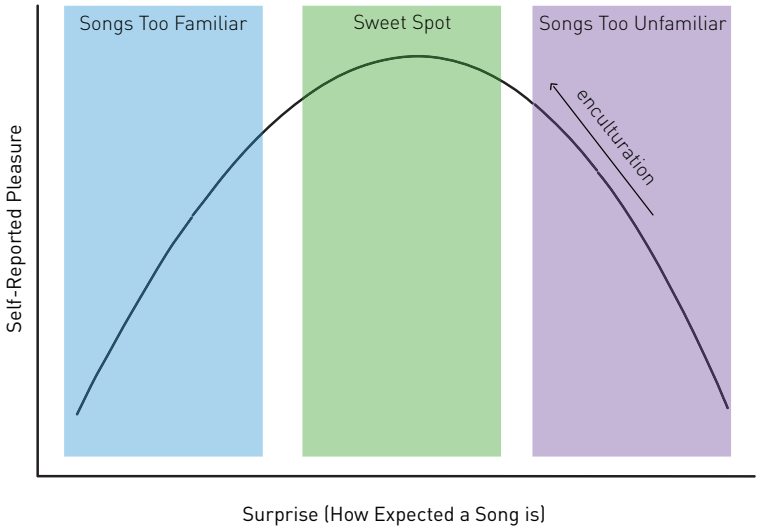


Figure 4.2. According to the Wundt effect, an intermediate level of predictability generates the maximal self-reported pleasure.

over enculturation, those participants will get more familiar with and will learn to predict the music better. We then hypothesized to see a drift of the musical pleasure toward the center of the inverted-U shape and therefore see an increase of the self-reported pleasure overexposure for the test group and not in the control group.

Figure 4.3 shows the change in self-reported pleasure between i) the first session and the second session and ii) the second and third sessions. Between the two first sessions, the test group was exposed to Chinese music (and therefore is expected to get more familiar with it), and the control group to Western music, both played on the Guzheng, a Chinese instrument.

We observe that there is a significant increase in the pleasure ratings for the test group which reflects the left shift in the inverted-U shape. We also observe a decrease in pleasure ratings for the control group which probably reflects the habituation of the timbre of the Guzheng with Western music. Participants were then more likely to predict the music as Western and were worse at predicting the Chinese notes after the exposure to Bach.

The right panel shows that the opposite effect was present after the resting period. This is nice evidence of the decay of the previously shown learning.

Topographic Change Over Time Many studies investigating the neural underpinnings of musical expectation showed that expectedness was encoded in the ERP amplitude around 200ms from the note onset with greater amplitude for unexpected notes(Di Liberto, Pelofi, Bianco, *et al.*, 2020; Di Liberto *et al.*, 2021; Lee *et al.*, 2019; Marion *et al.*, 2021; Omigie, Pearce, *et al.*, 2019; Omigie *et al.*, 2013a). An enculturation process should therefore affect those neural responses.

Figure 4.4 shows that both control and test participants have their ERP amplitude at 200ms

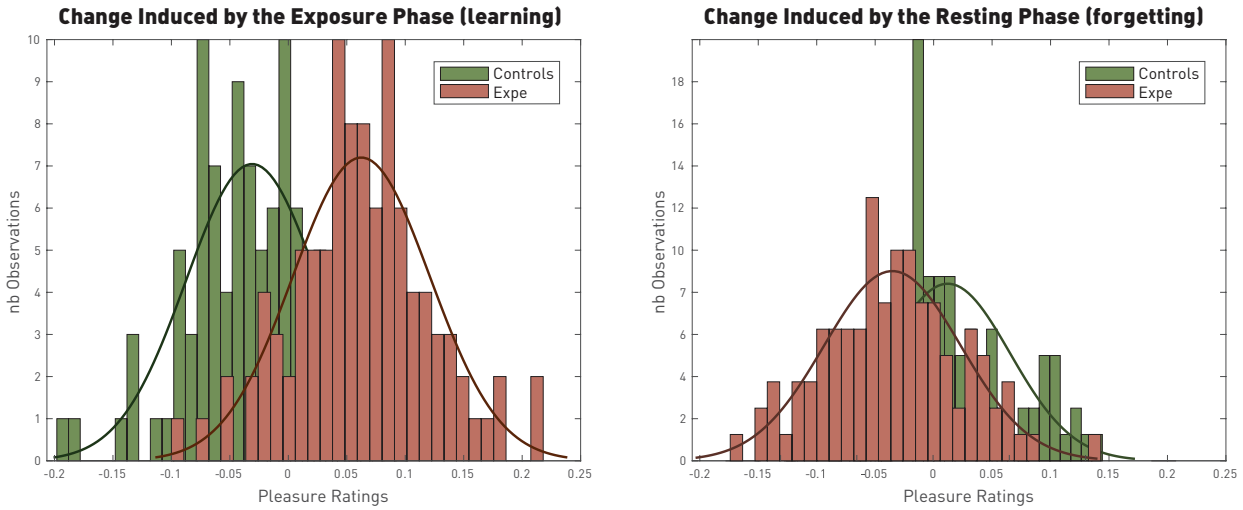


Figure 4.3. Change in pleasure ratings (after - before). An increase (mean above 0) means that participants increased their liking of the pieces. Those figures show the change induced by the exposure phase (left panel) and the resting phase (right panel) and show, respectively, the learning of a new musical grammar, and its decay after 2 months of no exposure. We can see that the effect of the exposure in the test group induced an increase in the pleasure ratings that have been partly erased after the resting phase, which is clear evidence of the learning of the new musical grammar and its decay.

affected by the exposure. This effect in the control group is explained by an increase in the SNR. Indeed, familiarity with the experiment (stress, ...) generally improves the SNR and generates a bigger response at the maximal time-latency of the neural response. However, this change at 200ms was significantly different between the two groups and the control group had a greater response than the test group, which is consistent with the previous literature about musical expectation.

The bottom panel shows that the resting phase has no effect on the control group which supports the hypothesis of SNR improvement for the change in the ERP after exposure. Moreover, the resting phase has a very clear effect in the test group which showed an increase in the ERP at 200ms. This is consistent with a decay of the learning as worse predictions generate bigger ERP responses.

Expectation Modeling Using IDyOM IDyOM, a statistical model of musical grammars (M. T. Pearce, 2005; 2018) has been used to model the EEG responses of Western listeners (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Di Liberto *et al.*, 2021; Marion *et al.*, 2021; Omigie, Pearce, *et al.*, 2019; Omigie *et al.*, 2013a) and showed a linear correlation between the amplitude of the ERP responses at 200ms and the Information Content[1] (IC) of the notes.

[1]The negative log-likelihood of a note x, referred to as *information content*, represents how well the model predicted it given a context $X_{k-n:k}$. This computation is numerically stable with an interpretation in terms of compressibility, the science of measuring information. For instance, events with high information content means

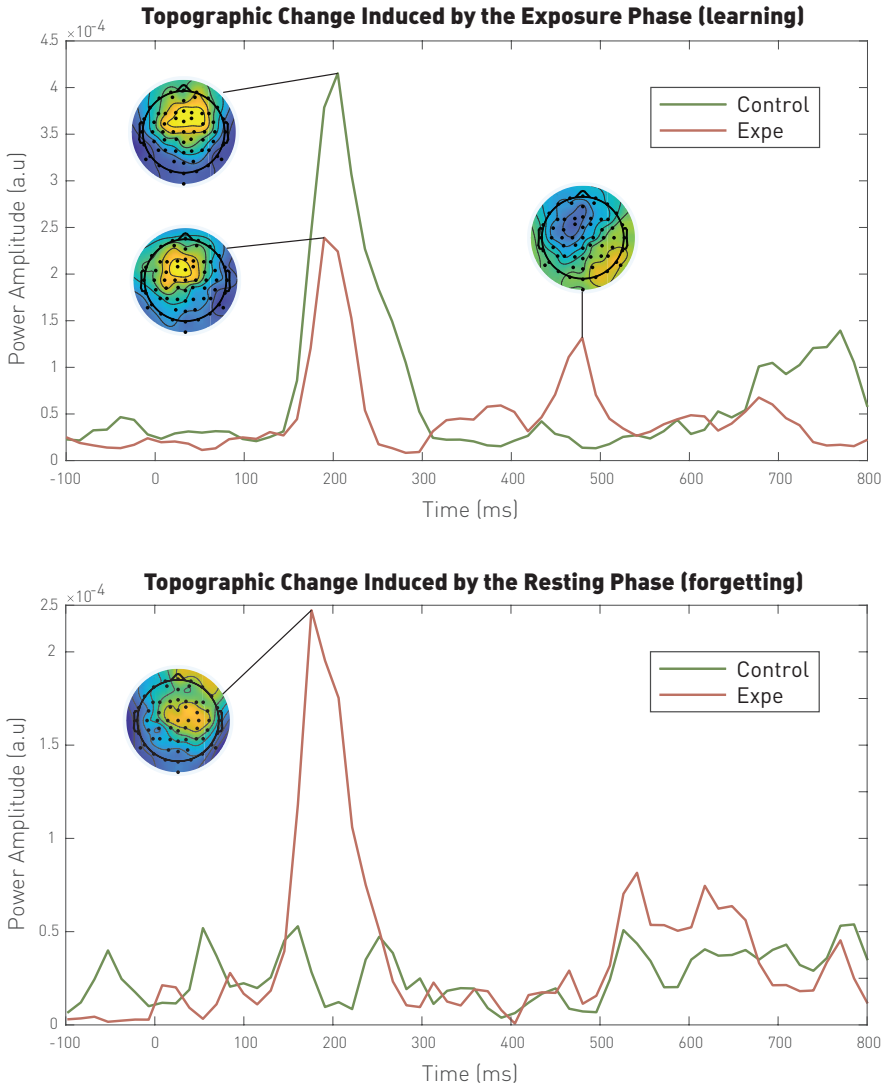


Figure 4.4. Change in pleasure ratings (after - before). An increase (mean above 0) means that participants increased their liking of the pieces. Those figures show the change induced by the exposure phase (left panel) and the resting phase (right panel) and show, respectively, the learning of a new musical grammar, and its decay after 2 months of no exposure. We can see that the effect of the exposure in the test group induced an increase in the pleasure ratings that have been partly erased after the resting phase, which is clear evidence of the learning of the new musical grammar and its decay.

This correlation can be done using regression methods between the IC signal and the EEG signal. Here we used the mTRF toolbox to compute a forward-lagged ridge regression between the IC computed with IDyOM trained on the exposed Chinese music (using the long-term model and a maximal order of 20 notes). We compared the correlations for each group before and after exposure.

are hard to compress as they occur rarely, one can therefore say that they contain a lot of information. that has been shown to provide good measures for psychological interpretations of perceptual data (Attneave, 1954; Chater & Vitányi, 2003).

Figure 4.5 shows a significantly different behavior between the two groups in terms of correlation change after exposure. We clearly see an increase in the correlation for the test group, which goes in the direction of the hypothesis.

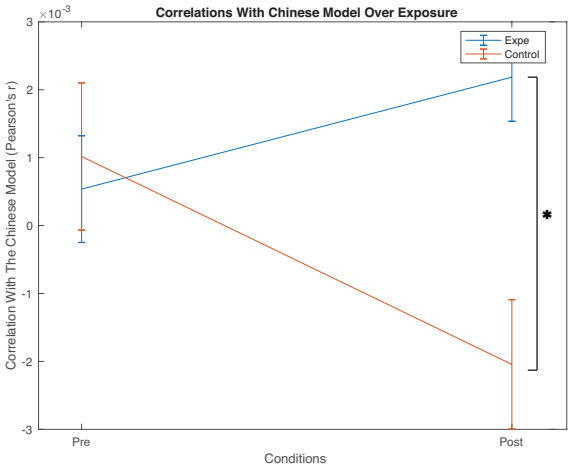


Figure 4.5

Disclaimer: This analysis is still ongoing, the results need to be double-checked and more analysis has to be done. For instance, we will run this analysis for the last sessions to see the decay of the learning. Then it would be interesting to check if all orders (temporal dependencies) are learned or if statistics are only learned to a certain order.

4.1.1.3 Discussion

We showed clear evidence of statistical learning engendered by passive exposure to unfamiliar music by means of behavior (pleasure ratings), neural (ERP amplitude), and neurocomputational modeling (using IDyOM).

4.1.2 ECoG Recordings in Ferrets (recorded by Rupesh Kumar Chillale at UMD)

4.1.2.1 Methods

We reproduced the protocol of the EEG experiments in ferrets.

Two ferrets were recorded for this experiment using exposure to Bach chorals instead of Chinese music. Two sessions were conducted before and after a 1-month exposure.

Recording #1 The ferrets have been recorded while played 30 minutes of Bach chorals by means of ECoG electrodes placed in the auditory cortex.

Exposure phase The test ferret has been exposed to Bach chorals (different pieces than for the recordings) 2h a day, 5 days a week for 5 weeks. The control ferret has been undergoing psycho-physic experiments manipulating sounds similar to music.

Recording #2 All ferrets were recorded again on the pieces used in recording # 1. They were also recorded while listening to a shuffled version of the stimuli. Those stimuli kept their first-order statistics (scale, duration set, and tempo) and got shuffled higher-order statistics. This serves to check whether the test ferrets were able to process better those high-order statistics than the control ferrets.

4.1.2.2 Results

We replicated the analyses from the EEG experiment. Figure 4.6 summarizes those analyses. Panel A shows the electrodes-averaged power ERP. We can see that the post-recordings have a smaller amplitude than the pre-recordings for the test ferrets but not for the control ferret. Panel C shows the same analysis as Figure 4.4 for the EEG. We computed the activity change over time for the note-ERPs. The test ferret shows a clear negativity at about 150ms (which is comparable to the responses at 200ms for the EEG) that is not present for the control ferret. Finally, panel B shows the analysis using the IDyOM model, we see similar behavior as for Figure 4.5 in EEG, which is that the exposure induced an increase in the correlation with the statistical model of the exposed music for the test ferret but not for the control ferret.

4.1.2.3 Discussion

It seems that ferrets can learn the complex structure of music in a similar way to humans. However, we still want to investigate what order of the statistics of the music the ferret is able to learn.

4.1.3 Intracranial Electrodes and Single Cell recordings in Ferrets (recorded by Flavien Feral & Pierre Orhan at ENS, Paris)

We recorded data from 3 ferrets using deep electrodes in the auditory cortex. These recordings will inform us of the internal mechanisms at the cell level. Those data have not been analyzed yet.

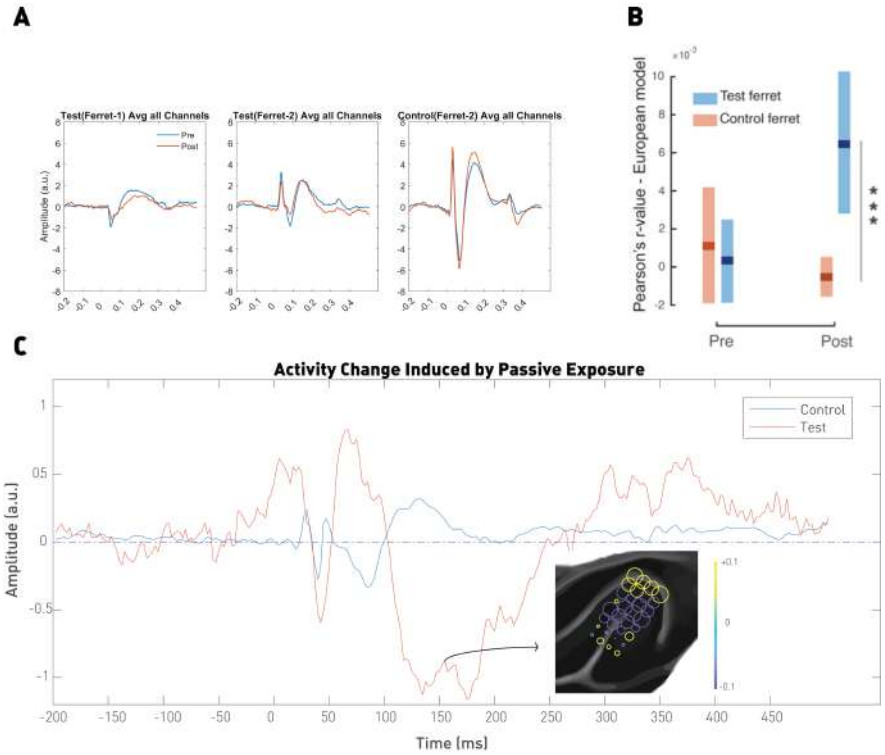


Figure 4.6. Results of the analyses on ECoG recordings on ferrets. Panel A shows that the raw note-ERP power amplitude is higher before than after exposure for the test ferrets but not for the control ferrets. Panel B shows that the statistical model of the exposed music (Western music) increased after exposure for the test ferrets but not for the control ferrets. Finally, panel C shows that there are greater changes at 150ms in the ERPs for the test ferrets but not for the control ferret.

4.1.4 Imaging Section, with Emphasis on Reward and Pleasure

As shown in Figure 4.3 of the EEG experiment, enculturation is followed by an increase in the self-reported pleasure of musical pieces that share the same musical grammar. On the other hand, the literature on musical reward suggests that the mesolimbic striatum is highly involved in behavioral musical pleasure, musical predictions, and learning(Zatorre & Salimpoor, 2013). In addition, surprise and uncertainty from statistical models of music predict self-reported musical pleasure as well as activity in the nucleus accumbens (Nacc) and the amygdala (Cheung *et al.*, 2019). It is therefore easy to hypothesis that when the predictions are modified, the pleasure will, as well as the activity in the NAcc. However, the spatial resolution of EEG does not permit this analysis. Therefore, we sought to complement our previous findings by measuring the effects of musical enculturation via the same Bach and Shanxi music clips with fMRI rather than EEG. The high spatial resolution of fMRI would then permit the direct analysis of NAcc. Finally, using, in addition, a ferret animal model will shed light on the enigma of why humans are the only social mammal that socially gathers around music. A huge corpus of literature in

music cognition raises the hypothesis that the evolutionary argument of music in humans is social bonding(Kathios & Loui, 2022; Savage *et al.*, 2021; Stupacher *et al.*, 2020; Trehub *et al.*, 2015). The reason why other mammals did not develop their sociability around music is still an open question. One answer could be that the involvement of the reward system in music enculturation was not developed enough. We here propose a group of experiments utilizing musical enculturation and imaging techniques in humans and ferrets to tailor those questions.

Note that we went for a different experimental protocol than for the electro-physiology panel. Here the subjects are exposed sequentially and recorded on two corpora. Therefore, there is no need for a control group as we have control within each subject.

4.1.5 FUS Recordings in Ferrets' NAcc (recorded by Jeffrey Boucher at ENS, Paris, France)

4.1.5.1 Methods

One ferret has been implanted with a FUS window in the NAcc. The protocol consists of 3 recordings and two exposure sessions on Bach chorals and Chinese music from the region of Shanxi.

Recording #1 The ferret has been recorded while played 30 minutes of Bach chorals and 30 minutes of Bach in shuffled orders.

Exposure # 1 The test ferret has been exposed to Bach chorals (different pieces than for the recordings) 2h a day, 5 days a week for 5 weeks.

Recording #2 The ferret has been recorded while played 30 minutes of Bach chorals and 30 minutes of Bach in shuffled orders, the same pieces as during recording # 1.

Exposure # 2 The test ferret has been exposed to Chinese Shanxi music (different pieces than for the recordings) 2h a day, 5 days a week for 5 weeks.

Recording #3 The ferret has been recorded while played 30 minutes of Bach chorals and 30 minutes of Bach in shuffled orders, the same pieces as during recording # 1 et #2.

4.1.5.2 Results

We only did preliminary analyses. We gave nutrical (treat for ferrets) to the ferret at regular intervals and recorded. We then computed the evoked response to the nutrical and to the musical notes. We did find a significant response (above noise) for the nutrical but not for the

musical notes (even after exposure). This hints that ferrets are not rewarded when listening to music, even after learning the structure. This means that the prediction error is not sent to the reward system as it is for humans. However, we are doing more extensive analyses to double-check this result.

4.1.6 fMRI in Humans (recorded by Sean Paulsen & Michael Casey at Dartmouth, USA

4.1.6.1 Methods

We replicated the protocol of the FUS experiment which consists of 3 recordings and two exposure sessions on Bach chorals and Chinese music from the region of Shanxi. The experiment is conducted with 10 participants, the data collection is still ongoing.

Recording #1 Participants have been recorded while played 30 minutes of Bach chorals and 30 minutes of Shanxi music in shuffled orders. They had to self-report how much pleasure they felt while listening to each song.

Exposure # 1 The participants are asked to listen to at least 30 minutes of music every day for 2 weeks on the same online streaming platform as for the EEG experiment. Bach chorals are played on this platform (but exposure #1 and #2 are interchanged every two participants).

Recording #2 Participants have been recorded while played 30 minutes of Bach chorals and 30 minutes of Bach in shuffled orders, the same pieces as for recording # 1.

Exposure # 2 The participants are asked to listen to at least 30 minutes of music every day for 2 weeks on the same online streaming platform as for the EEG experiment. Chinese music from the region of Shanxi is played on this platform (but exposure #1 and #2 are interchanged every two participants).

Recording #3 Participants have been recorded while played 30 minutes of Bach chorals and 30 minutes of Bach in shuffled orders, the same pieces as for recording # 1 and # 2.

Participants gave their written informed consent for each scan in accordance with the Institutional Review Board at Dartmouth College. They completed a brief questionnaire to determine eligibility. All participants responded that they had actively listened to Western classical music for more than 5 years of their life, and Chinese folk music for 0 years. All were thus deemed eligible. Upon arrival for each scan, the participants filled out a screening form to

confirm they could be scanned safely. They were each compensated $60 USD after the second session.

Each scan consisted of 8 runs. Each run began with two TRs and then consisted of four "blocks," which themselves consisted of four "trials." The design of a single trial is shown in Figure 4.7. All trials in a given block are the same style, resulting in 48 trials for each style per scan. There is no time between trials. A randomized jitter value between 4 and 7.5 seconds is assigned to the beginning of each trial to decouple the evoked response from elapsed time and prevent a consistent expectation of music starting. The parameters were 1mm3 voxels and a 1.5s TR.

Figure 4.7. The design of each trial during scanning. A randomized jitter value between 4 and 7.5 seconds is assigned to the beginning of each trial to decouple the evoked response from elapsed time and prevent a consistent expectation of music starting. The compensation lag is calculated such that the Pleasure Rating prompt appears after 39s, although this prompt only appears at the end of each block. Each participant's functional data consists of 8 runs, each of which had 4 blocks with 3 trials in each block. Each block was either all Bach or all Shanxi. Half of the blocks for each participant were Bach and the other half Shanxi. The arrangement of blocks was randomized for each participant. The two sessions for each participant had identical stimuli presentation. Figure taken from Sean Paulsen's PhD thesis

4.1.6.2 Preliminary Analyses

Self-reported pleasure We first plan to replicate the analysis we did in the EEG experiment using self-reported pleasure. We plan two analyses, both are based on the idea that we have access to the self-reported pleasure ratings. Therefore, we can align them for each session, and compute the difference so we have the change induced by exposure #1 and exposure #2. Then, for each participant, we have as many numbers as songs for the first exposure and for the second exposure. Now we can divide those numbers into two groups: Bach and Chinese. So you have 4 sets for each participant:

- BachSTIM-firstExposure
- ChineseSTIM-firstExposure

- BachSTIM-secondExposure
- ChineseSTIM-secondExposure

Now you have to flag which ones are congruent. For instance, if the participants were exposed to Bach in the first exposure and Chinese in the second exposure, it will be:

- BachSTIM-firstExposure (congruent)
- ChineseSTIM-firstExposure (incongruent)
- BachSTIM-secondExposure (incongruent)
- ChineseSTIM-secondExposure (congruent)

It is now just a matter of ordering those values. Two interesting analyses could be done (cf Figure 4.8):

Tracing the learning Concatenate all the values over all the participants whether the stimuli are Chinese or Bach and if they are congruent or incongruent. Then, we'll have 4 sets, it's just a matter of plotting them. In Figure 4.3 I chose to plot the raw distribution and to fit a normal distribution on them, but this is not necessary.

Effect of the Exposure Order This is to check whether the effect is different for the first or second exposure. So now instead of concatenating them according to whether the Stimuli are Chinese or Bach, just do it whether they are first or second exposure.

We expect to see a significant effect or congruent/incongruent but no significant effect or first/second exposure. Figure 4.8 shows the expected results.

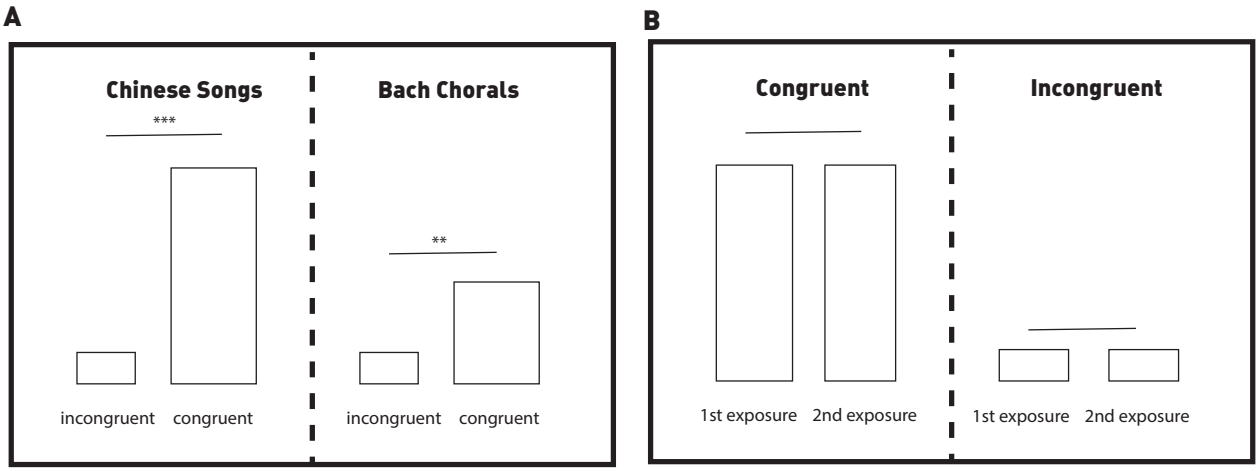


Figure 4.8. Expected results for the self-reported pleasure analysis.

fMRI Contrast Analysis We want to know what areas were affected by the exposure. Given one matrix (time * voxels) for each stimulus/participant/session. We want to compute the difference between each session, so we can see the effect for the first session and the second session, for each participant. It is a matter of computing the difference between those matrices for each stimulus and then averaging across stimuli (or keeping the variance for later). Then those difference maps (because we kept the voxel dimension) can be congruent or incongruent and we want to assess that the differences are greater for the congruent dimension. We can compute which voxels are significantly different and look at those maps for the congruent and incongruent conditions (it would be great to also see it for Chinese and Bach). We hypothesize to see differences in the incongruent condition as well. If the areas are overlapping between congruent and incongruent we can look at the difference between congruent and incongruent and look at this map (after FDR-corrected p-value thresholding).

Anatomical fMRI We also plan to check whether the anatomical scans during the resting state before the experiment (maps of white/grey matter) are affected by the exposure. We have 3 scans for each participant, we can compute the contrast and see what areas changed when the participant was exposed to Chinese and when the participant was exposed to Bach. We should see the auditory cortex and the pre-frontal areas (where the predictions are supposed to be sent from c.f. chapter 3), but not the reward system (where only the prediction error should be encoded), which should be affected by exposure only on functional scans.

4.1.7 General Discussion

We showed clear evidence of statistical learning produced by passive exposure to unfamiliar music using behavior (pleasure ratings), neural (ERP amplitude), and neurocomputational modeling (using IDyOMpy). Those findings are in line with the previous literature showing that computational models of music are good models for enculturation(M. T. Pearce, 2018; Tillmann, Bharucha, & Bigand, 2000) and that passive exposure to unfamiliar music does induce changes in the way humans predict musical events(Loui, 2012).

Also, it is the first time a study shows that enculturation also increases the musical pleasure felt while listening to music following the same structure, without being the same pieces, of the exposed music. This is a new and very strong evidence for the already demonstrated relationship between musical predictions and pleasure, and in a more general way, musical enjoyment, as discussed in the introduction.

Showing that this hypothesis is true is strong evidence that musical enjoyment is defined culturally through musical enculturation and that this process occur throughout our entire life. Still, there is no evidence of the Wundt effect for musical pleasure and predictions in non-

Western populations. It is therefore possible that this effect is, as others (c.f. chapter 5), socio-culturally defined and represents a cultural way to extract aesthetic value from music in Western societies. It is indeed important to replicate such experiments in non-Western populations and newborns to rule out this hypothesis.

On the other hand, we showed for the first time a physiological neural adaptation to passive exposure to unfamiliar music. This is the first neural evidence of the first hypothesis of the Predictive Coding Theory for Music, as defined by Pearce(M. T. Pearce, 2018): The statistical learning hypothesis. This hypothesis claims that the brain is always updating an internal statistical model of the music of our own culture. As shown in the introduction of the chapter, behavioral evidence has been discovered since the 90s, however, the field was missing a clear neural validation of this hypothesis. This consolidates an entire part of the field of music cognition that, starting in the 50s(Meyer, 1956), claimed that musical expectations were a cognitive root for perception (c.f. section 2.1). This thesis gave more neural validations of the first Predictive Coding hypothesis (Probabilistic Prediction Hypothesis) by showing predictive signals during moments of natural silences in ecologically valid music. We, therefore, think that we have been taking part in the field effort to push this theory to the front of the stage.

It is, however, evident that some of our work is still in progress (analysis of fMRI and FUS data). We hypothesize that fMRI will be strongly affected by the enculturation, especially in the NAcc. This hypothesis is directly derived from the literature about musical pleasure and its neural roots in the relationship with musical predictions(Cheung *et al.*, 2019; Zatorre & Salimpoor, 2013). Also, because we saw an increase in self-reported musical pleasure and that, in the case of familiar music, hemodynamic activity in the NAcc is associated with increasing self-reported pleasure(Zatorre & Salimpoor, 2013), we have strong reasons to believe in the hypothesis. However, it is very unclear whether such a mechanism will be shared with the ferrets. Indeed, the Wundt has only been validated in Western populations (using familiar and unfamiliar music)(Chmiel & Schubert, 2017) and, as discussed before it is unclear whether it is from a social construct of the aesthetic experience or a physiological root of learning(Ripollés *et al.*, 2014; 2016; 2018). The idea that learning and pleasure often co-occur(Ripollés *et al.*, 2014; 2016; 2018) and that, in the case of music, pleasure occurs in cases where a statistical model would get an efficient optimization (Gold, Mas-Herrero, *et al.*, 2019) is an argument for an evolutionary ancient root of the link between those two mechanisms. Discussing the cross-species validity of this mechanism would be able to rule out this question and could explain why human societies massively socialize around music whereas other mammals such as ferrets don't.

5 What Drives Musical Preferences?

5.1 General Introduction To Musical Preferences

It is evident that musical preferences and more generally music perception and listening behaviors differ between individuals. Spotify can nowadays provide databases to investigate those questions and studies show drastic differences between individuals and a great amount of this variance can be explained by socio-cultural and demographic factors. First, listeners prefer music from their own culture over music from other cultures, as shown through ethnicity (Appleton, 1971; Fung, 1993; Killian, 1990; LeBlanc, 1979; Meadows, 1970), even through non-explicit identification(May, 1985; McCrary & Gauthier, 1995; Teo, 2005), geographical variables(Mellander *et al.*, 2018), Spotify data(Thomas, 2017) and cross-cultural studies(H. Lee *et al.*, 2021). Beyond cultural factors, two models have been presented to explain the relationship between socioeconomic background and music preferences within a single culture. First, the model of Cultural Legitimacy was presented and validated in France(Bourdieu, 1979) and in the USA(Baumann, 1958; Schuessler, 1948) before the '80s. This model claims that higher social classes consume largely music considered more sophisticated or associated with a higher cultural value (such as classical and contemporary music and opera) and reject other forms of popular music as a legitimacy mechanism for affiliation to their social group. However, since the 90s another model called Omnivore/Univore has been presented by Peterson(Peterson, 1992) and empirically validated by studies in France (Coulangeon, 2005), USA(Peterson & Kern, 1996; Peterson & Simkus, 1992; White, 2001) and in the Netherlands(Van Eijck, 2001). This alternative hypothesis claims that social classes are differentiated by the variety in their music preferences rather than in a specific genre, with a trend of higher class listening to a broader range of musical genres. Studies from (White, 2001) using data between 1982 and 1997 in the USA show a temporal trend of a replacement of the Cultural Legitimacy model by the Omnivore/Univore model and that "cultural exclusivity is no longer valued as it may have been in the past and is more often a sign of ignorance rather than status". However, it has been still shown in France in 2003 that higher classes, even if principally distinguished by their omnivoreness, also tend to listen to more art music (classical and opera principally) than lower social classes(Coulangeon, 2005), showing that even if the most explanatory model switched from Cultural Legitimacy to Omnivore/Univore, the two of them still cohabit.

The idea that individuals within given social groups have similar musical preferences could be explained by the *social agreement hypothesis*. It has indeed been shown that listeners change their musical preferences to agree with peers(Alpert, 1982; Furman & Duke, 1988) or avoid disagreement(Furman & Duke, 1988; Inglefield, 1968; 1972) especially when the pairs represent an authority figure in term of musical taste(Alpert, 1982; Inglefield, 1972; Tanner, 1976). This idea follows the general idea that sociocultural familiarity increases music preferences(Droe, 2006) and elicits stronger emotional responses(Ritossa & Rickard, 2004).

Familiarity can be seen through different processes. First, listeners tend to prefer musical content that is already known as shown by explicit short-term repetitions of excerpts of classical music, including both tonal and atonal compositions(Gilliland & Moore, 1924; Margulis, 2013; Mull & Hennessy, 1957), jazz music(Verveer *et al.*, 1933), Korean music(M. K. Johnson *et al.*, 1985), Pakistani music(Heingartner & Hall, 1974), or by long-term exposure to specific pieces(Martindale & Moore, 1989; Martindale *et al.*, 1990). It can also be defined using the idea of predictability. Listeners, when presented with new musical content, would prefer pieces containing more statistical patterns shared with their own musical culture. For instance, American infants preferred (measured using looking-time) Western meters over Balkan meters whereas Turkish infants, familiar with both Western and Balkan meters, demonstrated no preferences(Soley & Hannon, 2010). Memory has even been demonstrated to correlate with familiarity cross-culturally as memory accuracy for Western melodies was better for Western participants than Turkish participants and the opposite; in addition, Chinese melodies (as controls) generated very bad memory accuracy for both groups(Demorest *et al.*, 2008). Both previous examples have been replicated and confirmed using prediction hypotheses with melodies containing statistics less consistent with Western culture (as assessed by statistical models); They engendered less memory accuracy in Western listeners (K. Agres *et al.*, 2018; K. R. Agres *et al.*, 2013). Furthermore, western infants were able to detect metric irregularities in Western-metered musical pieces, and also in Balkan-metered pieces but only after a week-long of at-home exposure to Balkan music(E. E. Hannon & Trehub, 2005a; 2005b).

In 1971, Berlyne proposed a model of such mechanisms by suggesting that an inverted U-shape (Wundt curve) would relate familiarity and preference(Berlyne, 1971). He argued that the reward system is activated by increasing arousal with exposure to certain patterns, but over time, the aversion system opposes this activation, leading to the increasing dominance of the aversion system as arousal continues to increase due to subsequent exposure. This effect has been validated by many studies (review in (Chmiel & Schubert, 2017)) and a study using fMRI showed that familiar music activates emotion-related limbic and paralimbic regions as well as the reward circuitry to a greater extent than unfamiliar music(Pereira *et al.*, 2011). Recently, two studies using a statistical model of music (IDyOM) showed a Wundt curve between the computational modeling of musical expectations and self-reports of musical pleasure(Cheung *et al.*, 2019; Gold, Pearce, *et al.*, 2019) as well as modulation of the activity in the reward-related region of the ventral striatum induced by the expectation (Gold, Mas-Herrero, *et al.*, 2019; C. L. Krumhansl, 1997; Shany *et al.*, 2019) or the uncertainty(Cheung *et al.*, 2019) of the musical events. This literature seems to show a clear line between cultural environments and musical preferences. The most striking limitation of those studies is that they are almost all based on questionnaires of self-reported musical genre preferences that are very likely to be biased and shaped by socio-cultural affiliation and agreement and not a direct observation of

music cognition. We propose that there are three forms of musical preferences: i) the music we say we like, measured by self-report of genre preferences, we think it is very influenced by genre and socio-cultural affiliation; ii) the music we actually listen to, measured by listening habits, e.g. through Spotify data which we believe is influenced by both cognition and socio-cultural affiliation; finally iii) the music we actually like as measured by cognitive experiments on ratings (or objective measures) on unknown music pieces, which we believe is mainly influenced by the prediction familiarity described before. It is clear that those three forms of musical preferences are correlated and intertwined, however, we posit that they result in different results and that their comparison could allow us to better understand the different factors influencing music perception.

From the literature, it is not clear that sociocultural factors could entirely explain the variance of musical preferences, meaning that there probably exist individual differences caused by the unique life experiences and physiology of the individuals. Personality being the first candidate for such differences, Rentfrow and Gosling (2003)(Rentfrow & Gosling, 2003) used the STOMP (Short Test of Musical Preferences) and observed that individuals with a preference for *Reflective* and *Complex styles* (such as classical and jazz) tend to have higher levels of *Openness to Experience* (as measured using the Big Five Personality test), perceive themselves as more intelligent, and have lower self-perceived athleticism. On the other hand, those who prefer *Rhythmic* and *Energetic* styles (such as electronic/dance and hip-hop) are more likely to exhibit higher levels of *Extraversion* and *Openness to Experience*, and perceive themselves as more athletic. However, a meta-analysis (combining 263,196 participants) on the relationship between personality traits and music preferences(Schäfer & Mehlhorn, 2017) finds that while there were small correlations for particular styles and traits, particularly between openness to experience and soft rock R'n'B, Classical and avant-garde, personality traits have limited predictive power in explaining interindividual differences in music preferences. Physiology could also be an explanation, but the few studies that have studied the relationship between heart rate and preference for certain tempi have exhibited contradictory results. Two studies investigated the genetics components of musical preferences (shown to be very small for face preferences(Germine *et al.*, 2015) but *not* absent) generated contradictory results on non-direct observations of musical preferences (through cultural preferences(Faust, 1974) and choice for instruments(Mosing & Ullén, 2018)).

5.1.1 Scientific Contribution

This is why we propose in this chapter a new study aiming at a direct comparison between the genetic and socio-economic components of musical preferences through a Twin study involving 30k twins from Sweden and the UK. Because this study will explicitly compare the

amount of variance explained by heredity and by social classes it is a good way to rule out some of the hypotheses about the vectors of the variance in terms of musical preferences and enrich the literature, mainly sociological, with physiological findings.

We will then present a new paradigm for studying the sociology of musical preferences through an original protocol consisting of i) a cognitive experiment; ii) a questionnaire about the socio-cultural environment and genre preferences; and iii) a semi-conducted interview about the music listening habits and present a cross-cultural validation of it. This study aims at replicating the previous findings in sociology but using other measures for musical preferences than solely questionnaires about musical genres. Indeed, as discussed earlier, we think that musical preferences could be measured by different means and that those different measures could reflect different aspects of the socialization around music. This new study will be able to uncover those different measures in a multi-site study in Paris and Rome.

Finally, taken together, these studies will compare multiple socio-cultural and genetic variables in a regression fashion and will be able to compare what exact variables explain the variance. This could result in decisive arguments for the enculturation theory that we are drawing here, as internalized socialization (also called *habitus* in sociology) could be seen as a passive learning, social reproduction, or even classical reproduction mechanism. Therefore, the scientific question we are asking here is whether the cognitive process of enculturation is actually the same as the social reproduction mechanism defined by sociology and whether it explains more variance than genetics.

This chapter will present two new studies: a genetic Twin Study and a sociology study. I decided to design the genetic study from the lack in the literature. I built collaborations with Miriam Mosing, Frederic Ullén, and Margherita Melanchini to have access to Twin cohorts. We found a collaboration where they gave me access to already collected data in England and Sweden (datasets are presented later). I designed and wrote the following description of the project and will be conducting the analyses, probably with the help of Giacomo Bignardi, PhD student working with Miriam Mosing. I designed the sociology project for similar reasons discussed in this introduction. I have been helped by very informative discussions with Gisèle Dambuyant, a French sociologist. The project has been supported by the École Française de Rome which funds projects in social sciences and mobility grants in Rome. The presented documents can be considered as pre-registration (even if not published yet) documents for those studies that will contain, at least, the persons cited before as authors. Shihab Shamma supervised the scientific process and will proofread the manuscripts of all those studies.

5.2 Genetic Components of Musical Preferences: A Twin Study

5.2.1 Introduction

Multiple studies in the field of sociology showed that the socio-economic environment shapes musical preferences. The initial model elucidating this association is the Cultural Legitimacy model, first expounded and substantiated in France by Bourdieu (Bourdieu, 1979), and earlier in the USA (Baumann, 1958; Schuessler, 1948). This model posits that higher social strata predominantly engage with music perceived as more sophisticated or possessing greater cultural value (such as classical, contemporary music, and opera) while dismissing other forms of popular music. This behavior serves as a legitimacy mechanism for social affiliation with their respective social stratum. However, from the 1990s onward, an alternative model known as Omnivore/Univore, pioneered by Peterson (Peterson, 1992), gained prominence and was empirically supported by studies conducted in France (Coulangeon, 2005), the USA (Peterson & Kern, 1996; Peterson & Simkus, 1992; White, 2001), and the Netherlands (Van Eijck, 2001). This alternative hypothesis suggests that social classes are delineated not by a preference for specific musical genres but by the diversity of their music preferences, with a discernible trend among higher social classes indicating a broader range of musical genre consumption. Investigations conducted between 1982 and 1997 in the USA (White, 2001) demonstrate a temporal shift from the Cultural Legitimacy model to the Omnivore/Univore model. This shift signifies that the once highly regarded cultural exclusivity is no longer esteemed as it might have been previously and is more frequently perceived as indicative of ignorance rather than social status. Nonetheless, a study in France conducted in 2003 (Coulangeon, 2005) revealed that even though higher social classes primarily distinguished themselves through omnivorous musical tastes, they still tended to gravitate towards more art music, predominantly classical and opera, in comparison to lower social classes. This observation implies that despite the transition from the Cultural Legitimacy model to the Omnivore/Univore model as the more explanatory framework, both models persist concurrently.

The idea that individuals within given social groups have similar musical preferences could be explained by the *social agreement hypothesis*. It has indeed been shown that listeners change their musical preferences to agree(Alpert, 1982; Furman & Duke, 1988) or avoid disagreement(Furman & Duke, 1988; Inglefield, 1968; 1972) with peers especially when they represent an authority figure in term of musical taste(Alpert, 1982; Inglefield, 1972; Tanner, 1976). This idea follows the general idea that sociocultural familiarity increases music preferences(Droe, 2006) and elicits stronger emotional responses(Ritossa & Rickard, 2004). This hypothesis,

which is the most present in the literature of sociology(Bourdieu, 1979; Coulangeon, 2017; White, 2001) and psychology(M. T. Pearce, 2018; Pelofi *et al.*, n.d.; Vuust, Heggli, *et al.*, 2022b; Vuust, Heggli, *et al.*, 2022) of music proposes music preferences as a purely cultural and acquired norm. However, there is no clear evidence that innate components in musical preferences do not exist.

For instance, personality seems to play a role in music preferences (as seen in the previous section) (Rentfrow & Gosling, 2003; Schäfer & Mehlhorn, 2017) and personality traits from the Big Five, consistently with the rest of the literature (Koenig, 2020; Zwir *et al.*, 2020), have been shown to be substantially heritable and explain 40–60% of the variance(Power & Pluess, 2015), it suggests the possibility that music preferences could also have a significant genetic component. Moreover, social reproduction, defined by Bourideu as *habitus* has been recently argued to have potential genetic components, especially as highbrow/lowbrow taste (perceived as high- or low-cultural value for higher social classes), participation (a person's involvement in activities that provide interaction with others in society or the community) and cultural omnivoreness, core elements of the social reproduction theory, have been shown to be notably heritable. Especially, participation in highbrow, lowbrow, and popular culture has been found to be, respectively, 47-70%, 59-67%, 48-61% heritable and 39-50% for self-reported music omnivoureness(Jæger & Møllegaard, 2022).

To date, only two studies investigated the genetic components of musical preferences (shown to account for 20% in face preferences(Germine *et al.*, 2015)) which generated contradictory results through unconventional measures of musical preferences. The first (Faust, 1974) conducted a questionnaire about many aspects of personal preferences including like/dislike questions about classical instrumental, classical vocal, jazz, folk, and pop, and did not find any significant genetic effect. A second more recent study investigated the specialization of musicians for a specific instrument and musical genre. It found a significant effect of genetics(Mosing & Ullén, 2018) showing that musical genre preference for music production is heritable but not necessarily for music perception. Another interesting result from the sociology of music found that women liked a wider range of styles, especially "serious" ones(Hargreaves *et al.*, 1995). Another study observed that females consistently showed more positive attitudes toward music than males, particularly at younger ages(Crowther & Durkin, 1982). However, it is important to consider the influence of culture, age, and country when interpreting gender-related preferences, as highlighted by (LeBlanc *et al.*, 1999)

We, therefore, think that it would be interesting for the community of music cognition to have a study on musical preferences and musical omnivorouness comparing the respective effects of socioeconomic factors (as defined by income and occupation category) and genetics. Moreover, testing the genetic effects on sex differences would allow us to make hypotheses on gender differences in musical omnivourness.

The Twin methodology consists of conducting a behavior measurement (here musical preferences) on identical (MZ) and fraternal twins (DZ), which respectively share 100% or about 50% of their genetics but which both share the same environment (e.g., uterine environment, parenting style, education, wealth, culture, community). Therefore, a trait that is purely heritable would be identical for MZ twins but potentially different for DZ twins. Therefore, comparing the correlation on the behavioral traits between MZ and DZ pairs allows us to conclude whether the trait is largely driven by genetics (MZ pairs would be more correlated than DZ pairs), by shared environment (both correlations are high but not different), or by non-shared environment (correlations are both low and comparable). The common model to estimate heritability in twin studies is known as the ACE model (or its non-additive genes version, ADE)(Zyphur *et al.*, 2013). Twin studies using this model estimate how much the variation in a phenotype is due to additive genetic effects (A), the common environment (C) or non-additive genes (D), and the unique, random, environment (E). Structural equation modeling (SEM) partitions the variance of a phenotype into these three components using maximum likelihood methods.

We propose five analyzes in order to assess to which extent musical preferences is heritable, the amount of variance that can be exclusively explained by genetics and socio-economics and shared familial background, and whether the gender differences could be supported by genetics or socio-cultural gender norms. We will, therefore, i) compute the ACE analysis for 19 musical genres, ii) compare the computed heredity with its lay-estimate (heritability estimated by non-specialists), iii) test whether heredity is moderated by age, iv) test whether heredity differs by gender, and finally v) conduct a multilevel modeling that includes socioeconomic data. All five analyses will be conducted on two different datasets: the TEDS' *18 Year Fashion Food and Music Preferences* and the *Music Preference Questionnaire* on the Swedish Cohort.

5.2.2 Data

5.2.2.1 TEDS

Study participants were twins from the Twin Early Development Study (TEDS)(Haworth *et al.*, 2013), a birth cohort of 16,810 families with twins born in England and Wales from 1994–1996. TEDS was previously shown to be reasonably representative of the general population(Haworth *et al.*, 2013). Requests to complete the online food preference questionnaires were sent out to a subsample (3166 pairs; n = 6332 individuals) by letter and e-mail during the year of their 18^{th} birthday. Subjects were offered a £10 voucher to complete the survey, resulting in 3155 individual twins who consented to participate. This breakdown is representative of typical monozygotic/dizygotic proportions observed in twin populations.

Date of birth, sex, and socioeconomic information (income, job, and education level of

parents when twins were 16 years old) were collected in the baseline questionnaire. Zygosity had previously been collected by using a parental report questionnaire completed in early childhood. DNA analysis has shown the questionnaire to be .95% accurate(Price *et al.*, 2012); uncertain zygosity was determined from DNA.

The questionnaire contained questions about Fashion, Food, and Music Preferences. This dataset includes a few questions about musical preferences, including ratings for 10 musical genres that have never been used for any published study. The only two studies based on those data are about drink (Smith *et al.*, 2017)and food preferences (Smith *et al.*, 2016). The 10 different musical genres were supplemented by examples of artists:

- How much do you like listening to Pop music (for example, music by Lady Gaga, Katy Perry, Justin Bieber or Taylor Swift)?
- How much do you like listening to Hip Hop / Rap music (for example, music by Macklemore, Kanye West, Jay-Z and Eminem)?
- How much do you like listening to R&B / Soul music (for example, music by Beyonce, Amy Winehouse, Adele and Iggy Azalea)?
- How much do you like listening to Rock music (for example, music by Linkin Park, Muse, Arctic Monkeys, and Red Hot Chilli Peppers)?
- How much do you like listening to Metal music (for example, music by Metallica, ACDC, Slipknot, Avenged Sevenfold and System of a Down)?
- How much do you like listening to Dance / Electronic music (for example, music by Skrillex, Daft Punk, David Guetta and Calvin Harris)?
- How much do you like listening to Alternative / Indie music (for example, music by Lana del Rey, alt-J, Bastille and Kings of Leon)?
- How much do you like listening to Jazz music (for example, music by Ella Fitzgerald, Ray Charles, Louis Armstrong and Norah Jones)?
- How much do you like listening to Classical music (for example, music by Mozart, Bach, Vivaldi and Beethoven)?
- How much do you like listening to Folk music (for example, music by Mumford and Sons, The Lumineers, Of Monsters and Men and Noah and the Whale)?

Participants were asked to to rate how much they enjoy listening to music of each broad genres using a 10-items Likert scale labeled as: 1='Not at all'; 2-9=numbered; 10='A lot'.

5.2.2.2 Sweden Cohort

Swedish Twin Registry STAGE cohort includes $\tilde{3}$2,000 adult twins born between 1959 and 1985. A web-based data collection 'Humans making music' in 2010/2011 was initiated by Fredrik Ullén and followed up in 2021 as 'Humans making music 2.0' (HUMMUS). They contain a rich array of questionnaires about musical engagement, music-related behavior and musical aptitude tests.

Those questions contain a 7-item Likert rating for the following 19 musical genres: **Classical Music, Opera, Jazz, Blues, Reggae, Funk, Pop, Gospel, Rock, Soul, Metal, Electronic Music (including techno and house), Hip-hop/rap, Indie/alternative rock, Dance Music, Latin, Country, Sweden Traditional Music, and Folk/World Music**.

The questionnaire also contains the self-reported number of hours of music listening per week and socioeconomic questions including level of education, and occupation.

5.2.3 Analyses

5.2.3.1 ACE for Music Preferences

Intra-class correlations (ICCs) will be calculated for each musical genre as well as for the overall musical preference (using multidimensional correlation over all musical genres) scores for MZ and DZ pairs to provide an indication of the pattern of similarity for the two types of twins. Maximum Likelihood Structural Equation Modelling (MLSEM) will be used to derive precise estimates of A, C and E (with 95% confidence intervals and goodness-of-fit statistics) based on the expected structure of the variance-covariance matrices for MZs and DZs.

This analysis will provide us with values for ACE for each musical genre and for the overall musical preferences. Those values will be compared to ACE of other traits such as preferences for food, color, and fashion.

5.2.3.2 Does Heredity Correspond to its Lay Estimate?

An independent data collection about music perception contained a question about the estimate of the heritability of musical preferences. We plan to compare the heritability computed using the ACE model with the lay estimates computed an another cohort of participants. We therefore will compare those values with other famous traits such as political beliefs, depression, ADHT, and height, c.f. Fig 5.1 for an estimated figure.

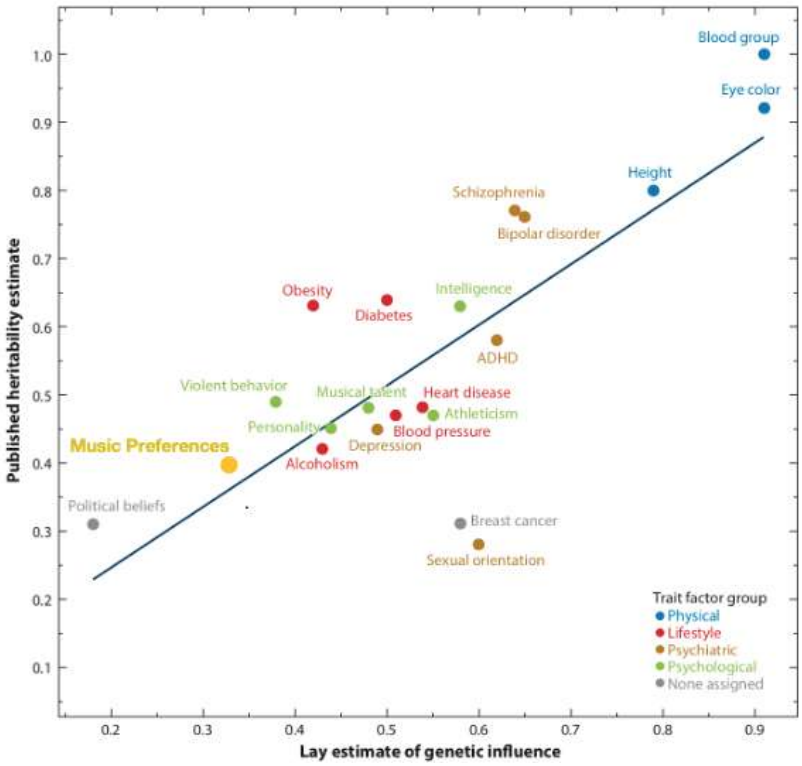


Figure 5.1. Comparison of the musical preference heredity and its lay estimate with those for different known traits.

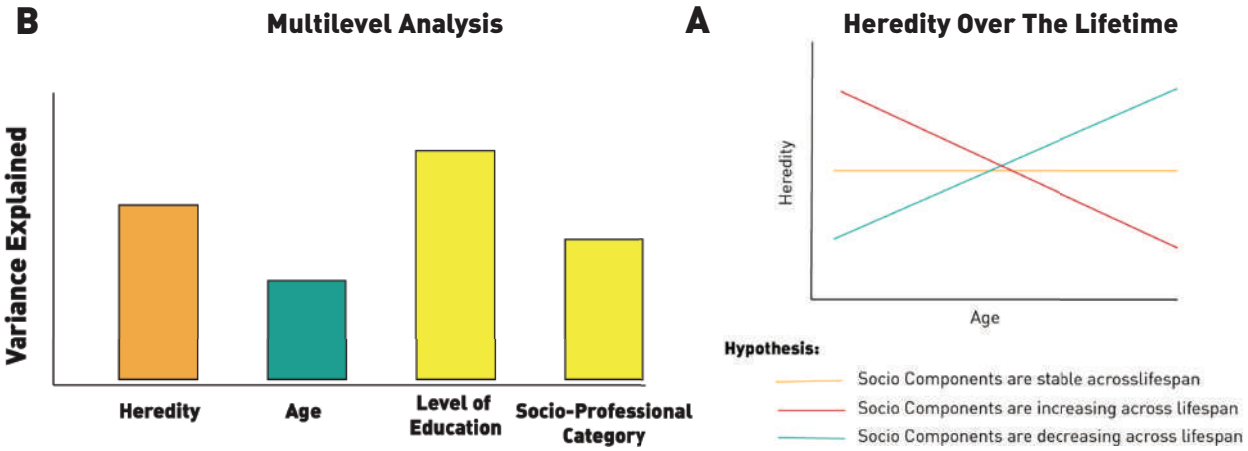


Figure 5.2. **A** Hypotheses about the effect of the socio-cultural environment on music preferences over the lifetime. **B** Hypothesis the exclusive variance explained for socioeconomic and genetic factors.

5.2.3.3 Does the Effect of Environment Change Throughout Life?

In order to see whether the genetic components of musical preferences evolve during life we plan to compare the A, C and E parameters over subgroups of different age. This analysis will allow us to check the hypothesis that the non-shared environment (E) explains more and more variance over age. This hypothesis rests on the social agreement hypothesis that states that individuals change their musical preferences to agree(Alpert, 1982; Furman & Duke, 1988) or avoid disagreement(Furman & Duke, 1988; Inglefield, 1968; 1972) within a social group, in order to strengthen their sociocultural affiliation(Bourdieu, 1979). Once a moderator fitted between age and E, we can see whether the E computed from the TEDS (participants aged of 18 yo) is consistent with the regression. See Fig 2.A for a hypothesis illustration.

5.2.3.4 Heredity By Gender

If twin correlations suggest sex differences, sex-limitation models will be fitted for the musical genres, as well as for the general musical preferences. These models test whether the magnitude of A, C, and E differ for males and females (quantitative sex differences), and whether the genetic and environmental influences are the same or different for males and females (qualitative sex differences)(Neale & Cardon, 1992). This analysis will allow to conclude whether the gender differences in music preferences observed in sociology studies are due to genetic or sociocultural norms.

5.2.3.5 Multilevel Twin Modeling Including Socioeconomic

Twin modeling is commonly addressed within the context of structural equation modeling (SEM)(Rijsdijk & Sham, 2002) as a one-level model where the family serves as the primary sampling unit. However, the analysis of twin data can also be approached from the perspective of multilevel models (MLMs). A classic illustration involves children as level 1 units clustered within classes at level 2, which in turn are clustered within schools at level 3(Sellström & Bremberg, 2006). Detailed insights into the CTM within the MLM framework, along with illustrative code and various extensions, are provided by (Hunter, 2020)(Hunter, 2021). Previous studies have successfully employed this approach(Tamimy *et al.*, 2021). Notably, music preferences have been found to exhibit a strong influence from socioeconomic factors, and these factors are also interconnected with genetics through social reproduction. As such, conducting a multilevel analysis would enable the disentanglement of socioeconomic and genetic influences, yielding a precise estimate of the proportion of variance exclusively attributed to genetics and socioeconomic factors. For an overview of the hypothesis, please refer to Figure 2.B.

5.3 Sociology of Music: A Cross-Cultural Study on Musical Preferences

5.3.1 Introduction

Cognitive sciences typically do not address individual differences; however, I believe it is essential not to disregard them. The main objective of this project is to identify and quantify these differences to understand their origin using different measurement techniques in order to also see the difference between self-reported preferences, objective measurements of preferences through music listening. To achieve this, I decided to conduct a cognitive experiment on music perception, followed by a semi-directed sociological interview and a questionnaire to identify participants' environment and social backgrounds. By conducting these in different locations where populations are socially homogeneous, it will be possible to quantify and compare inter-individual and inter-group differences and link them to sociological variables identified during the interview/questionnaire. The chosen locations are Zalib Circolo Arci in Rome and La REcyclerie in Paris. These two places are similar structures (socio-cultural centers) that attract a diverse audience, making sociological intervention feasible. This document will outline the details of the study and planned analyses.

5.3.2 Locations

5.3.2.1 Zalib Circolo Arci - Rome - Italy

The Italian Arci network is a social mission association created after World War II, centered around solidarity and anti-fascism values. Currently, this association counts 1,115,002 participants spread across 4,867 local associations that foster social bonds through cultural, social, and political events. With a membership fee of 8 euros per year to access all network associations, Arci stands out as a significant contributor to social cohesion in Italy. Zalib has been part of this network since 2023. Originally centered around a bookstore and promoting reading among people under 30, this venue is located in the Trastevere neighborhood of Rome. It now serves as a hub for cultural dissemination, hosting book presentations, concerts, stand-up comedy (in Italian or English), film screenings, and roundtable discussions. It also serves as a workspace and meeting place, allowing people to work alone or in groups without the obligation to consume. Thus, Zalib plays a role in cultural promotion and social bonding in the Trastevere neighborhood and more broadly in the center of Rome.

I engaged with 50 individuals (with an average age of 23,12), including 27 males, 22 fe-

males, and 1 non-binary person, out of which 15 were musicians. Among them, 10 lived in neighborhoods close to Zalib (Monteverde, Pratti, Trastevere), 4 in central Rome, and 9 in southern Rome. Only 4 lived in the eastern neighborhoods of Rome, which can be explained by the fact that many Centro Sociale and Circolo Arci are located in East Rome. Zalib thus emerges as a socio-cultural center for the central and western neighborhoods of Rome. Generally, the participants' demographic data from Zalib reveal individuals with a very high level of education (34 participants were fluent in at least 2 languages, 48 had graduated (average level of education of 3.8) and 29 were students at the time of the interview). However, they show relatively low incomes (median of 10k per year), likely due to their young age (median age of 24 years). Nevertheless, it can be assumed that they also possess significant economic assets, as 30 of the participants have at least one parent in socio-professional category 3 (executives and higher intellectual professions).

5.3.2.2 La REcyclerie - Paris - France

Multiple places were considered as location in Paris, such as *La parole Errante* (Montreuil), *La flèche d'or* (Paris), La cité Fertile (Pantin), *le 6B* (Saint-Denis), *le T-KAWA* (Paris). It was very difficult to find a place that had a similar social role as Zalib Circolo Arci. We finally decided to go for *La REcyclerie*, situated in Paris in the 18th district, for its situation in inner Paris (as Zalib in Rome), and its cultural, political, and ecological engagement.

The REcyclerie presents itself as a third place of exchange rooted in the values of sustainable collaboration and empowerment. According to the founders, a third place is a space a space offering a thousand and one activities and possibilities for a wide range of needs and audiences. The REcyclerie may be a workspace, but it also has a cultural cultural, educational, leisure, restoration, relaxation, repair, learning, experimentation, etc. The site is intended to be very friendly and open.

However, La REcyclerie, as opposed to Zalib is a private structure that owns the space through another company (Sinny & Ooko) and is not funded by the state, they also have private partnerships such as Veolia and Black& Deker. A direct consequence of this status is that prices are quite high for an alternative place (7,5€ for the cheapest pint of beer). Even if La REcyclerie does not require consumption to stay in the space, this makes the place frequented by higher social classes in a popular neighborhood of Paris.

I engaged with 50 individuals (with an average age of 29,8), including 16 males, 32 females, and 2 non-binary people, out of which 20 were musicians. Among them, 18 lived in the 18th district, 10 in inner Paris (including 5 in the center), and 17 in Ile de France, including 12 in considered wealthy suburbs (e.g. Levallois-Perret, Saint Germain en Laye, Issy-les-Moulineaux) and 5 from considered popular suburbs (Clignancourt, Noisy le grand, Choisy le

roi). La REcyclerie thus emerges as a socio-cultural center hosting individuals from relatively wealthy neighborhoods in a popular one. Generally, the participants' demographic data reveal individuals with a very high level of education (48 graduated with an average level of education of 3.6 years and 23 were students at the time of the interview). However, they show relatively low incomes (median of 17k per year) which is way lower than the average salary in Paris (33k). Nevertheless, it can be assumed that they also possess significant economic assets, as 25 of the participants have at least one parent in socio-professional category 3 (executives and higher intellectual professions).

5.3.2.3 Comparison

It seems that both Zalib and La REcyclerie both host individuals with high levels of education: 48/50 went to University for both places with an average level of education of 3.6 and 3.8. Both populations also had incomes way lower than the average of the city and parents belonging to high socio-professional categories. Those demographics hint at a population with low personal economic resources but probably medium/high economic resources from their parents and high cultural assets both from home parenting and higher education. However, the first striking difference is the proportion of male/female, even, if the proportion of non-binary individuals were comparable (2% and 4%) and similar to the one of the global population (between 0.5% and 5% depending on studies). Zalib has a proportion very close to parity (27M/22F), La REcyclerie exhibits a larger amount of females (16M/32F). Finally, while Zalib is definitely an important place for the neighborhood of Trastevere and does not host individuals from the eastern part of Rome (because other sociocultural places are situated there), La REcyclerie seems to host people coming from center areas of Paris (5) and from outside Paris (17). In conclusion, the two places seem to host similar populations (high level of education and low incomes) but because of the different situations about socio-cultural places in Paris and Rome, the hosts come from different geographical places.

5.3.3 Experimental Procedure

The sociological intervention involved approaching isolated individuals (groups of up to 3) at Zalib and asking them to participate in a music perception psychology experiment. The experiment's introduction was pre-written and standardized, presenting a 30-minute music psychology experience that would be compensated with 10 euros upon completion. All interactions were conducted in Italian. If participants agreed to participate, they were asked to sign a consent form outlining the procedure and their right to withdraw from the experience at any time. Compensation was funded by the Laboratory of Perceptual Systems at the École Normale

Supérieure, and the experiment adhered to the ethical guidelines of the CERES 2013-11 committee of the University of Paris Descartes. The intervention comprised three stages: cognitive experiment, questionnaire, and semi-directed interview. The cognitive experiment involved 80 musical excerpts of 20 seconds each (described in the following section). Participants were required to rate their enjoyment of the excerpts on a -3 to 3 Likert scale and indicate the emotions associated with the excerpts using emojis on a sad/happy scale (see appendix for images). They could also indicate if they were familiar with an excerpt, allowing its exclusion from analysis. This experiment precisely measures musical preferences and emotional responses during music listening, which are the two most important factors in music perception. This data enables us to compare music perception precisely across different social groups. The program for conducting the experiment was developed by me using Python with the Psychopy platform. Subsequently, participants were asked to complete an online questionnaire, starting with demographic information (age, gender, place of residence, etc.), followed by specific questions about their music listening conditions (location, duration, context, etc.). Finally, they answered a questionnaire about their musical preferences, rating 19 music genres from -3 to +3. This step allows us to compare our results with previous sociological studies and to compare directly (through music listening) and indirectly (through questionnaires without music listening) measured musical preferences. Lastly, participants were given the option to contribute an additional 15 minutes for a sociological interview about their relationship with music. This interview was conducted with 10 participants and could take various forms based on the participants' responses. However, it typically began with general questions about why they enjoyed listening to music, in which situations, and whether they listened to different music based on situations and people they were with. The interview could then delve into their connection between music and memory (reminiscence, Proust's Madeleine effect), their connection between sociability and music, or their relationship between individuality and music.

5.3.4 Stimuli

I compiled 80 musical excerpts from various sources, including every noise, YouTube, SoundCloud, and Spotify. The approach was to include a diverse range of musical grammar (structures modulating timbre, melody, harmony, and rhythm) to capture subtle differences between participants. To achieve this, I aimed to select pieces from different countries, musical genres, and eras (see appendix for the list), with special attention to choosing examples that participants were unlikely to be familiar with (there is a significant perceptual bias when one is already familiar with a piece). After initial analyses, very few stimuli were recognized by participants. Finally, an algorithm was developed to randomly choose 20-second passages, which were used in the experiment.

5.3.5 Analyses for How to Measure Musical Preferences: A Comparative Study

We proposed in section 5.1 that there are three forms of musical preferences: by the genre they self-report they prefer, by the genre they actually listen to (through Spotify data), and by the music they actually prefer when probed with music they don't know. We hypothesize that those three proxies over musical preferences measure a mix of different mechanisms: sociocultural affiliation and individual music cognition (mainly thought of through familiarity). We propose to study the differences between the two measures made through questionnaires of preferred musical genres and self-reported preferences while listening to musical excerpts.

All our 80 stimuli will be categorized into the 17 genres of the questionnaire by a panel of professionals (some working at Radio France, the French public radio). All the analyses will be conducted on high-dimensional data (all the genres/stimuli). Therefore, we will use multidimensional generalizations of the correlation and explained variance equations. Basically, they consist of averaging the results across the dimensions (there is no obvious way of weighing them). In the case of the correlation, we concatenate all dimensions into a single vector and run the usual Pearson's correlation. For the explained variance, when the variables are numeric, we just use the generalized correlation formula and take the square of it (as in the unidimensional case). In the case of categorical variables, the multi-dimensional variance is computed as the sum of the variance over all the dimensions. To compute the explained variance, we compute the ratio of the between-groups sum squares ($\sum_d \sum_i (\overline{x}_{..d} - \overline{x}_{i..})^2$, for each group i and dimension d) and the total sum squares ($\sum_d \sum_{i,j} (\overline{x}_{...} - x_{ijd})^2$, for each group i, group sample j and dimension d).

The first analysis will compare the raw preference correlations between genres as measured by means of the questionnaire and by means of the audio stimuli. Big differences in this direct comparison will suggest that the two measures capture different components (Fig.5.3.A). We check whether this difference allows to explain the individual differences by computing the distance between each participant by means of the two methods (questionnaire and audio stimuli). We therefore end up with 2 distance vectors for which we can compute the correlation. This correlation tells how much the structure of the individual differences resemble between the two methods. To assess whether one method contains more information than the other, we conduct a regression analysis and compute how much of the variance of the two methods can be explained by the data of the other method (Fig.5.3.B). We hypothesize that a significantly higher amount of the variance of the questionnaire can be explained by the audio stimuli than inversely, meaning that audio stimuli result in richer and finer-grained measures of musical preferences. We can now ask the question of whether the variance of the two measures can

be explained by different socio-cultural variables, meaning that in addition to being different, they are affected by different sociological mechanisms. To this end, we compute the relative explained variance for several socio-cultural variables and for the two methods. The relative explained variable is computed using multi-variable linear regression from all the socio-cultural variables to the preferences measures and from all the variables but the one of interest in order to compute the relative contribution of this variable. Then, in order to compare all the variables even if they have different numerical substrates, we run a null model where we shuffle the variable of interest and look at the improvement of the model with respect to the null models (Fig.5.3.D). We hypothesize that the questionnaire measures will be more correlated to socially affiliating variables such as gender and socio-professional categories than cultural variables such as the city of living and level of education. Finally, for methodological interest for future studies, we run a PCA-like analysis where we sort the audio stimuli by their exclusive contribution to the total variance of the 80 stimuli. Then, we compute the amount of variance explained by the first n stimuli, the function is, by definition, monotonic and should reach 0 when all 80 stimuli are included(Fig.5.3.C). To compute the exclusive contribution of each stimulus, we run a PCA and weigh the explained variance of each principal component with the absolute value of the coefficient for each variable. We then sum the resulting values over all the principal components and sort the contribution of each variable. Then, we run a multi-variable linear model from the stimuli 1:n to the entire set of 80 stimuli. The explained variance is computed using the previously defined high-dimensional Pearson's correlation, and the square of it (r^2) gives the explained variance. Before our experiment is composed of a large set of various musical stimuli, this method will allow to replicate this experiment with similar populations using a lot less stimuli, and, therefore, significantly shorten the duration of the experiment which is crucial in sociologically valid situations.

5.3.6 Analyses for What Sociocultural Variables Explain Musical Preferences and Emotions

Here and for the rest of this section we will use the socio-professional categories as an example, however, we will run those analyses for different variables. Also, because we asked the participants to self-report the valence of the stimuli on a sad/happy scale, we have a proxy over the emotional responses of the participants, which, in addition to the preference, allows to study the perception of the participants.

The first analysis (Fig.5.4.A) will divide the entire set of participants into groups according to the given variable of interest (here socio-professional categories as an example) and will show the raw preferences for the different musical genres (by means of both measures previ-

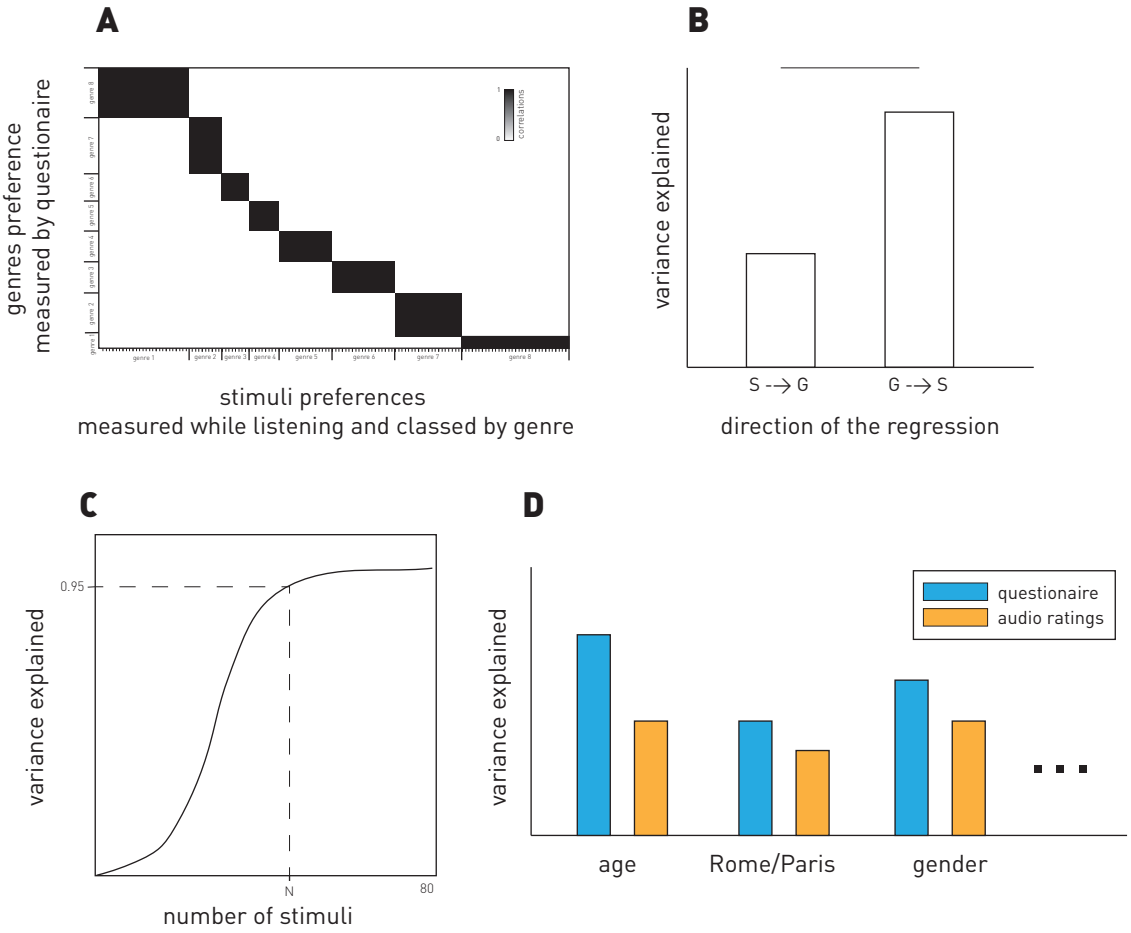


Figure 5.3. **Planned analyzes about the difference between the two measures of musical preferences.** **A** Cross-correlations between genre preferences measured by the questionnaire or by self-reported preferences on audio stimuli. **B** Variance explained by the regression from the stimuli-measured preferences to the questionnaire-measured preferences, and inversely. **C** We defined the optimal number of audio stimuli in order to explain 95% of the total variance. **D** We compare the relative variance explained by socio-cultural and demographic variables for the two preference measures.

ously compared) as well as the omnivoreness. The omnivoreness is computed as the flatness of the musical preferences function, therefore, we use the Shannon entropy over the sum-normalized preferences by genres. This is a good way to replicate the findings by (Coulangeon, 2005) that the French population is divided by omnivoreness over socio-professional categories and that higher social classes also listen to more culturally-valued genres such as classical music and opera. We also hope to discover other trends, that will be, hopefully, different between Rome and Paris, over other variables. We want to assess whether some variables are more explicative of the preferences than others and if the relative contribution of the variables is similar between musical preferences and emotional responses. To this end, we conduct a similar analysis to in Fig.5.3.D but add the emotional responses and will interpret it in relationship to the other analyzes. In addition to this analysis, we propose in Fig.5.4.B a graphical representation

of the individual differences for each variable. The scatter plot represents all the participants plotted on the two first principal components of the data. The color and size do the points will show variance over the given variable of interest. We can, therefore, assess whether a given variable correlates to the data in a linear or non-linear way. We can also visualize distances between participants in the perceptual space in relationship to the socio-cultural variables.

This study will give a new and innovative view on the sociology of music in the two places in Paris and Rome through qualitative and quantitative data of music perception and fine-grained and top-level statistical analyzes.

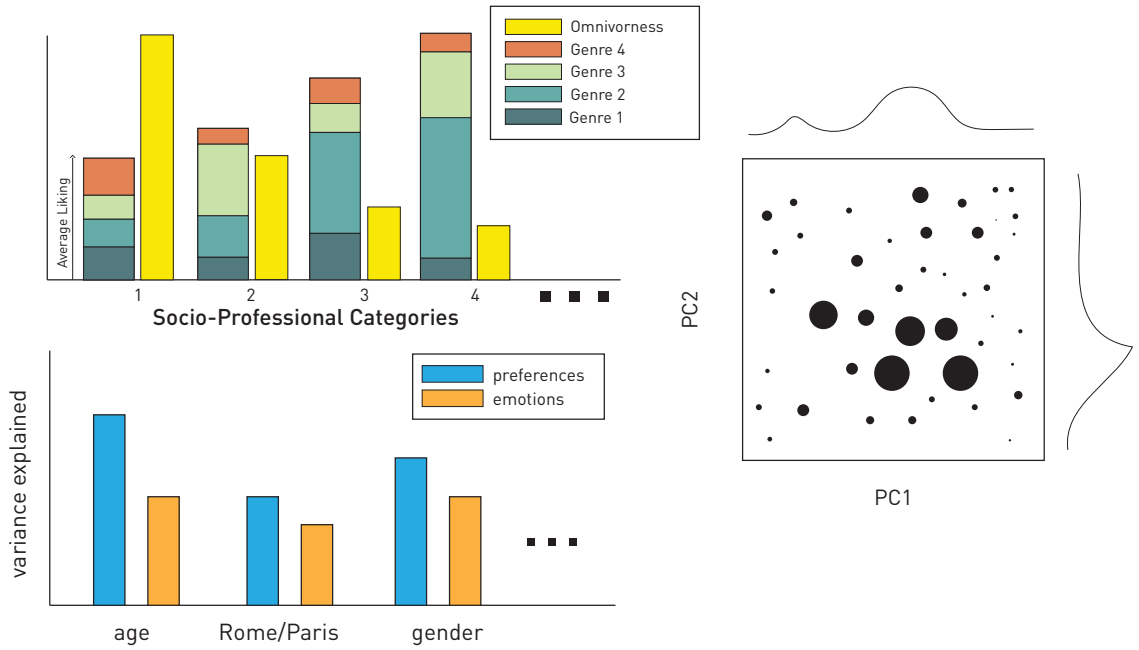


Figure 5.4. **Planned analyses about the comparison of the variables explaining musical perception. A** Comparison of the raw genre preferences and omnivorness by socio-professional categories.**B** 1-principal component map of socio-professional categories (given by the size of the points). **C** Comparison of the exclusive and relative explained variance between variables for both the musical preferences and the emotional responses.

6 Discussion and Conclusions

In this thesis, we covered different aspects of music perception, putting ourselves in the framework of the predictive coding theory (Clark, 2013; K. J. Friston *et al.*, 2010) that claims that the brain emits sensory predictions about upcoming events based on the statistics of the external environment.

The literature was already rich in examples of brain responses to unfamiliar musical events (Koelsch, 2009; Koelsch & Mulder, 2002; Koelsch *et al.*, 2000; Leino *et al.*, 2007; Loui *et al.*, 2005; Saarinen *et al.*, 1992; Steinbeis *et al.*, 2006), even using continuous measures of expectation through statistical models of music (Di Liberto, Pelofi, Bianco, *et al.*, 2020; Omigie, Pearce, *et al.*, 2019). We first explored this question in more detail. Specifically, we isolated those predictions by using a statistical model of music to predict the moments of natural silences in ecologically valid music. We showed that those moments induced statistically significant brain responses of an inverse polarity with respect to listening responses and that those responses were also correlated with the expectation of having a note in those moments (the more likely the note the more negative the response). This finding is strong and new evidence for a predictive mechanism during music perception that tries to ultimately cancel the sensory responses. Such a mechanism has been shown to have a role in facilitating perception, for instance, by aiding in restoring missing or noisy parts of a stimulus (Leonard *et al.*, 2016; McClelland & Rumelhart, 1988) or biasing ambiguous perception (Brainard & Hurlbert, 2015; Pressnitzer *et al.*, 2018). We also showed that similar responses are found during musical imagery, the action of mentally hearing music without any physical stimulation. Those responses already demonstrated to spatially overlap with listening responses through fMRI studies (Bastepe-Gray *et al.*, 2020; Bunzeck *et al.*, 2005; Griffiths, 1999; A. R. Halpern, 2001; A. R. Halpern & Zatorre, 1999; A. R. Halpern *et al.*, 2004; Herholz *et al.*, 2012; Hubbard, 2013; Kraemer *et al.*, 2005; Lima *et al.*, 2015; Yoo *et al.*, 2001; Zatorre & Halpern, 2005; Zatorre *et al.*, 1996; Zhang *et al.*, 2017). However, the field was missing clear electrophysiological characterization of those responses. We showed that their dynamics were very related to those of during perception as they were of an almost perfect inverted polarity. We showed that it was possible to use imagery response to reconstruct listening responses, and inversely, very in line with the previous literature.

This predictive coding framework offers a versatile theoretical base for computational modeling as statistical models generate predictions after being updated with training data. Such models have been extensively used by the community for behavior(J. J. Bharucha & Stoeckig, 1986; Bigand & Pineau, 1997; Bigand *et al.*, 2001; Margulis, 2003; Margulis & Levine, 2006; Marmel *et al.*, 2008; 2010; Omigie, Pearce, & Stewart, 2012; Tillmann *et al.*, 2006), electrophysiology (Di Liberto, Pelofi, Bianco, *et al.*, 2020; A. R. Halpern *et al.*, 2017; Marion

et al., 2021; Omigie, Pearce, *et al.*, 2019; Omigie *et al.*, 2013a; M. T. Pearce *et al.*, 2010; Quiroga-Martinez, C. Hansen, *et al.*, 2020; Quiroga-Martinez, Hansen, *et al.*, 2020) and even fMRI(Cheung *et al.*, 2019). Since those models are useful in music cognition, we developed and presented several updated and enhanced versions. A first set was developed for musical expectations based on the IDyOM architecture(M. T. Pearce, 2005) and has proven to outperform the previous implementation of IDyOM on certain measures, especially the theoretical modeling of musical culture, which makes them appropriate models for future cross-cultural studies on music cognition. The second set of models pinpoints cross-sensory predictions and is based on the literature on cross-modal predictions between the motor and auditory cortex (Brodsky *et al.*, 2008; Y. Ding *et al.*, 2019; Grisoni *et al.*, 2019; Mado Proverbio *et al.*, 2014; Tian & Poeppel, 2010; 2012; 2013; Ventura *et al.*, 2009; Whitford *et al.*, 2017; Zatorre *et al.*, 1996). This literature proposes that efference copies are sent between those two areas.

An explanation for such a system is learning motor controls. Indeed, speaking requires building a mapping from an auditory representation to a motor representation. However, the feedback given from the vocal tract back to the ears is a physical path that does not allow it to backpropagate the errors naturally. Building an alternate neural (motor/auditory) pathway allows the system to do so. We call this architecture the Mirror Network and hypothesize that it could solve problems that would require computing the inverse of a complex and non-differentiable physical component. Such mapping has also been proposed in music production and perception(Martin *et al.*, 2017). We presented here two implementations of the Mirror Network: one for speech and one for music. In addition to showing that simulations of those models allow for learning non-differentiable modules, we also show that they have strong engineering applications as they allow to solve complex engineering problems very efficiently (without labeled data for instance).

Finally, there has remained much uncertainty as to how musical expectations are formed. Studies have shown that participants from different cultures form different predictions, consistent with statistical models trained on the music of their own culture(C. Krumhansl *et al.*, 1999). It was also known that passive exposure to structured pitch sequences induced predictions that were consistent with the statistical structures of those sequences(Loui *et al.*, 2010). However, it was not clear what neural mechanisms were underpinning this adaptation, the lasting of this effect, and its relationship to musical enjoyment. We showed in this thesis that the same components of the EEG responses that are modulated by musical expectations were affected by the passive exposure to unfamiliar music and that those changes in the amplitude were consistent with statistical models trained on the exposed music. The accompanied increase in self-reported pleasure allowed us to also bridge the gap between recent findings showing that the relationship between expectation and musical pleasure follows an inverted U shape (also known as the Wundt effect)(Berlyne, 1971; Chmiel & Schubert, 2017; Huron, 2006) that

shows that musical pleasure has its optimum for an intermediate-level of surprise(Cheung *et al.*, 2019; Gold, Pearce, *et al.*, 2019). Our finding that an increase in familiarity pushes toward the optimum of the curve is therefore consistent with this previous literature.

Those three chapters kept together draw a very clear line between predictions, culture, and enjoyment: i) an internal statistical model of musical structures always sends predictions about upcoming events, ii) we like the music that is coherent with those predictions but also not too predictable. And finally, iii) being exposed to new music updates our model and makes our musical preferences evolve. This theory seems to corroborate nicely previous findings in music cognition and the sociology of musical preferences. In addition, it gives a clear evolutionary argument for cortical specification for music in humans. Indeed, what the Wundt effect proposes is that we like music that challenges our internal model without injecting too much noise into it. This is a computational argument for inclusion and stability: we are willing to include the music of others in our own culture as long as it does not break the stability of the system. As this mechanism is implemented in every listener, the collective behavior is to update models by including the music of all individuals making the global preferences evolve toward a more representative type of music. Since humans socialize and experience rituals around music, we can see this neural mechanism as an evolutionary mechanism for social cohesion.

In this thesis, we also discussed the limitations of this theory and proposed two new studies aiming at a better overview of this phenomenon. The first study is a genetic study based on twin modeling that could shed light on the respective effects of the environment and genetics on individual musical preferences. The second study proposed to examine cross-cultural sociology data to see the effects of sociocultural affiliation on musical preferences and discuss what could be acquired by other means than mere passive exposure to music.

That said, more work is needed to have a better and sounder overview of this phenomenon. First, most of the studies are conducted on western populations making it hard to conclude whether the observed mechanisms are due to social constructs or innate physiological processes. For instance, there is currently no cross-cultural validation of the Wundt effect(Chmiel & Schubert, 2017). The only study conducted on non-Western populations showed a linear (and not inverted U-shaped) relationship between familiarity and preferences(Chmiel & Schubert, 2017; Teo *et al.*, 2008). As Western populations have recently transitioned from a Cultural Legitimacy paradigm (each social class consumes its exclusive music, and higher social classes consume exclusively more complex genres of music) to an omnivorous paradigm (higher social classes are not distinguished by specific genres but by the variety of music they like), it is thus possible that the Wundt effect would also be a social construct that the Western ideology deeply imbues in our brains by the argument that intermediate levels of complexity result in superior aesthetic experiences. Ethical cross-cultural studies and experiments in newborns will help answer this question and give sound arguments about the universality of certain cognitive

mechanisms.

We have proposed here that musical preferences could be shaped not only by the statistics we are exposed to but also by the way we value aesthetic experiences. I, therefore, think that it is also important to include sociology in music cognition as some cognitive phenomena can be entirely shaped by the environment. In addition to giving very good insights into the balance of power and the role of social values in the aesthetic experience, it could also suggest new methods that could be seen as micro cross-cultural studies. Populations within a city, or between different non-remote geographic areas are, indeed, very heterogeneous, and many sociological variables such as gender, socio-economics, and education level are known to explain a lot of the variance of musical preferences. Finally, the interplay of music cognition and culture roots itself at the core of the nature/nurture debate. I also think that a collaboration with genetic and developmental studies, in addition to ethical cross-cultural studies, could shed light on those questions.

BIBLIOGRAPHY

Abdallah, S., & Plumbley, M. (2009). Information dynamics: Patterns of expectation and surprise in the perception of music. *Connection Science*, *21*(2-3), 89–117.

Agres, K., Abdallah, S., & Pearce, M. (2018). Information-theoretic properties of auditory sequences dynamically influence expectation and memory. *Cognitive Science*, *42*(1), 43–76.

Agres, K. R., Abdallah, S. M., & Pearce, M. T. (2013). An information-theoretic account of musical expectation and memory. *Cognitive Science*, *35*.

Alpert, J. (1982). The effect of disc jockey, peer, and music teacher approval of music on music selection and preference. *Journal of Research in Music Education*, *30*(3), 173–186.

Alvin, L., F, C., & D, S. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Ames, C. (1989). The markov process as a compositional model: A survey and tutorial. *Leonardo*, *22*(2), 175–187.

Ammirante, P., Patel, A. D., & Russo, F. A. (2016). Synchronizing to auditory and tactile metronomes: A test of the auditory-motor enhancement hypothesis. *Psychonomic Bulletin & Review*, *23*(6), 1882–1890.

Anat, P., Jennifer, S., Edward, C., Jack, L., Josef, P., & Robert, K. (2018). Mirroring in the human brain: Deciphering the spatial-temporal patterns of the human mirror neuron system. *Cerebral Cortex*, *28*, 1039–1048.

Anders, N., David, S., Jun, T., Katsuyasu, S., Fan, W., & Richard, M. (2013). A circuit for motor cortical modulation of auditory cortical activity. *Journal of Neuroscience*, *33*, 14342–14353.

Andrew, L., S, G., Lori, H., & Holt. (2009). Reflections on mirror neurons and speech perception. *Trends in cognitive sciences*, *13*, 110–114.

Andrew, P., Jason, A., Laurentius, H., Elisha, M., & Alex, M. (2020). Layer-specific contributions to imagined and executed hand movements in human primary motor cortex. *Current Biology*.

Appleton, C. R. (1971). The comparative preferential response of black and white college students to black and white folk and popular musical styles. *Unpublished doctoral dissertation, New York University*.

Arom, S. (2004). *African polyphony and polyrhythm: Musical structure and methodology*. Cambridge university press.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, *61 3*, 183–93.

Barlow, H. B., *et al.* (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*(01).

Bastepe-Gray, S. E., Acer, N., Gumus, K. Z., Gray, J. F., & Degirmencioglu, L. (2020). Not all imagery is created equal: A functional magnetic resonance imaging study of internally driven and symbol driven musical performance imagery. *Journal of Chemical Neuroanatomy*, *104*, 101748.

Baumann, V. H. (1958). *Socio-economic status and the music preferences of teenagers* [Doctoral dissertation, University of Southern California].

Bendixen, A., Schröger, E., & Winkler, I. (2009). I heard that coming: Event-related potential evidence for stimulus-driven prediction in the auditory system. *The Journal of Neuroscience*, *29*(27), 8447–8451.

Berlyne, D. (1971). *Aesthetics and psychobiology*. Appleton-Century-Crofts.

Bharucha, J. J., & Stoeckig, K. (1987). Priming of chords: Spreading activation or overlapping frequency spectra? *Perception & Psychophysics*, *41*(6), 519–524.

Bharucha, J. J., & Stoeckig, K. (1986). Reaction time and musical expectancy: Priming of chords with no partials in common. *The Journal of the Acoustical Society of America*, *80*(S1), S87–S87.

Bianco, R., Gold, B., Johnson, A., & Penhune, V. (2019). Music predictability and liking enhance pupil dilation and promote motor learning in non-musicians. *Scientific reports*, *9*(1), 1–12.

Bianco, R., Ptasczynski, L. E., & Omigie, D. (2020). Pupil responses to pitch deviants reflect predictability of melodic sequences. *Brain and Cognition*, *138*, 103621.

Bigand, E., & Pineau, M. (1997). Global context effects on musical expectancy. *Perception & Psychophysics*, *59*(7), 1098–1107.

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? a review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100–130.

Bigand, E., Tillmann, B., Poulin, B., D'Adamo, D. A., & Madurell, F. (2001). The effect of harmonic context on phoneme monitoring in vocal music. *Cognition*, *81*(1), B11–B20.

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences*, *98*(20), 11818–11823.

Bourdieu, P. (1979). *La distinction. critique sociale du jugement*. Éd. de Minuit.

Brainard, D. H., & Hurlbert, A. C. (2015). Colour vision: Understanding thedress. *Current Biology*, *25*(13), R551–R554.

Briot, J.-P. (2021). From artificial neural networks to deep learning for music generation: History, concepts and trends. *Neural Computing and Applications*, *33*(1), 39–65.

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983.e5.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.e3.

Brodsky, W., Kessler, Y., Rubinstein, B.-S., Ginsborg, J., & Henik, A. (2008). The mental representation of music notation: Notational audiation. *Journal of experimental psychology. Human perception and performance*, *34*, 427–45.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners.

Bunzeck, N., Wuestenberg, T., Lutz, K., Heinze, H.-J., & Jancke, L. (2005). Scanning silence: Mental imagery of complex sounds. *NeuroImage*, *26*(4), 1119–1127.

Campbell, P. S. (2010). *Songs in their heads: Music and its meaning in children's lives*. Oxford University Press.

Carlsen, J. C. (1981). Some factors which influence melodic expectancy. *Psychomusicology: A Journal of Research in Music Cognition*, *1*(1), 12.
Caroline, N., Srikantan, N., & John, H. (2013). What does motor efference copy represent? evidence from speech production. *The Journal of Neuroscience*, 16110–16116.
Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of north india. *Journal of Experimental Psychology: General*, *113*(3), 394.
Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134.
Cenkerová, Z., & Parncutt, R. (2015). Style-dependency of melodic expectation: Changing the rules in real time. *Music Perception: An Interdisciplinary Journal*, *33*(1), 110–128.
Cervantes Constantino, F., & Simon, J. Z. (2017). Dynamic cortical representations of perceptual filling-in for missing acoustic rhythm. *Scientific Reports*, *7*(1), 17536.
Chambers, C., Akram, S., Adam, V., Pelofi, C., Sahani, M., Shamma, S., & Pressnitzer, D. (2017). Prior context in audition informs binding and shapes simple features. *Nature Communications*, *8*(1).
Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19–22.
Chennu, S., Noreika, V., Gueorguiev, D., Shtyrov, Y., Bekinschtein, T. A., & Henson, R. (2016). Silent expectations: Dynamic causal modeling of cortical prediction and attention to sounds that weren't. *The Journal of Neuroscience*, *36*(32), 8305–8316.
Cheung, V. K., Harrison, P. M., Meyer, L., Pearce, M. T., Haynes, J.-D., & Koelsch, S. (2019). Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology*, *29*(23), 4084–4092.
Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*.
Chmiel, A., & Schubert, E. (2017). Back to the inverted-u for music preference: A review of the literature. *Psychology of Music*, *45*(6), 886–909.
Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.
Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.
Cleary, J., & Witten, I. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, *32*(4), 396–402.
Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy [Cutting-edge EEG Methods]. *Journal of Neuroscience Methods*, *250*, 126–136.
Conklin, D. (1990). *Prediction and entropy of music*. University of Calgary.
Corrigall, K. A., Tillmann, B., & Schellenberg, E. G. (2022). Measuring children's harmonic knowledge with implicit and explicit tests. *Music Perception: An Interdisciplinary Journal*, *39*(4), 361–370.
Coulangeon, P. (2005). Social stratification of musical tastes : Questioning the cultural legitimacy model. *Revue française de sociologie*, *46*, 123.
Coulangeon, P. (2017). Cultural openness as an emerging form of cultural capital in contemporary france. *Cultural Sociology*, *11*(2), 145–164.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*.

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers of Human Neuroscience*, *10*(604).

Crosse, M. J., Liberto, G. M. D., & Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers of Human Neuroscience*, *10*(604).

Crowther, R. D., & Durkin, K. (1982). Sex-and age-related differences in the musical behaviour, interests and attitudes towards music of 232 secondary school students. *Educational Studies*, *8*(2), 131–139.

Daniel, W., & Zoubin, G. (2000). Computational principles of movement neuroscience. *Nature*, *3*, 1212–1217.

Daniel, W., Zoubin, G., & Michael, J. (1995). An internal model for sensorimotor integration. *Science*, *269*, 1880–1882.

Daube, C., Ince, R. A., & Gross, J. (2019). Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, *29*(12), 1924–1937.e9.

David, S., Anders, N., & Richard, M. (2014). A synaptic and circuit basis for corollary discharge in the auditory cortex. *Nature*, *513*, 189–194.

de Cheveigné, A., Di Liberto, G. M., Arzounian, D., Wong, D. D., Hjortkjær, J., Fuglsang, S., & Parra, L. C. (2019). Multiway canonical correlation analysis of brain data. *NeuroImage*, *186*, 728–740.

Demorest, S. M., & Morrison, S. J. (2016). Quantifying culture: The cultural distance hypothesis of melodic expectancy. *The Oxford handbook of cultural neuroscience*, 183.

Demorest, S. M., Morrison, S. J., Jungbluth, D., & Beken, M. N. (2008). Lost in translation: An enculturation effect in music memory performance. *Music Perception*, *25*(3), 213–223.

den Ouden, H. E., Kok, P., & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*, 548.

Di Liberto, G. M., Marion, G., & Shamma, S. A. (2021). The music of silence. part ii: Music listening induces imagery responses. *Journal of Neuroscience*.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465.

Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., de Cheveigné, A., Shamma, S., & Mesgarani, N. (2020). Cortical encoding of melodic expectations in human temporal cortex. *eLife*, *9*, e51784.

Di Liberto, G. M., Pelofi, C., Shamma, S., & de Cheveigné, A. (2020). Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoustical Science and Technology*, *41*(1), 361–364.

Di Liberto, G. M., Wong, D., Melnik, G. A., & de Cheveigné, A. (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *NeuroImage*, *196*, 237–247.

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46.

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening [PMID: 21975452]. *Journal of Neurophysiology*, *107*(1), 78–89.

Ding, Y., Zhang, Y., Zhou, W., Ling, Z., Huang, J., Hong, B., & Wang, X. (2019). Neural correlates of music listening and recall in the human brain. *Journal of Neuroscience*, *39*(41), 8112–8123.

Doelling, K. B., Assaneo, M., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences*, *116*(21), 10113–10121.

Dominic, M., & Trevor, C. (2008). The motor theory of speech perception revisited. *Psychonomic bulletin & review15*, 453–457.

Droe, K. (2006). Music preference and music education: A review of literature. *Update: Applications of Research in Music Education*, *24*(2), 23–32.

Eerola, T., Louhivuori, J., & Lebaka, E. (2009). Expectancy in sami yoiks revisited: The role of data-driven and schema-driven knowledge in the formation of melodic expectations. *Musicae Scientiae*, *13*(2), 231–272.

Egermann, H., Pearce, M. T., Wiggins, G. A., & McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(3), 533–553.

Einat, L., & Riikka, M. (2018). An interactive model of auditory-motor speech perception. *Brain and language*, *187*, 33–40.

Engel, J., Hantrakul, L., Gu, C., & Roberts, A. (2020). Ddsp: Differentiable digital signal processing.

Engel, J., Swavely, R., Hantrakul, L., Roberts, A., & Hawthorne, C. (2020). Self-supervised pitch detection by inverse audio synthesis.

Enns, J. T., & Lleras, A. (2008). What's next? new evidence for prediction in human vision. *Trends in Cognitive Sciences*, *12*(9), 327–333.

Esling, P., Masuda, N., Bardet, A., Despres, R., & Chemla-Romeu-Santos, A. (2020). Flow synthesizer: Universal audio synthesizer control with normalizing flows. *Applied Sciences*, *10*(1).

Farbood, M. M. (2012). A parametric, temporal model of musical tension. *Music Perception*, *29*(4), 387–428.

Faust, J. (1974). A twin study of personal preferences. *Journal of Biosocial Science*, *6*(1), 75–91.

Ferezou, I., & Deneux, T. (2017). Review: How do spontaneous and sensory-evoked activities interact? *Neurophotonics*, *4*(3), 031221.

Ferreri, L., Mas-Herrero, E., Zatorre, R. J., Ripollés, P., Gomez-Andres, A., Alicart, H., Olivé, G., Marco-Pallarés, J., Antonijoan, R. M., Valle, M., *et al.* (2019). Dopamine modulates the reward experiences elicited by music. *Proceedings of the National Academy of Sciences*, *116*(9), 3793–3798.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, *12*(6), 499–504.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826.

Fogel, A. R., Rosenberg, J. C., Lehman, F. M., Kuperberg, G. R., & Patel, A. D. (2015). Studying musical and linguistic prediction in comparable ways: The melodic cloze probability method. *Frontiers in psychology*, *6*, 1718.

Fourer, D., Rouas, J.-L., Hanna, P., & Robine, M. (2014). Automatic timbre classification of ethnomusicological audio recordings. *International Society for Music Information Retrieval Conference (ISMIR 2014)*.

Fracile, N. (2003). The aksak rhythm, a distinctive feature of the balkan folklore. *Studia Musicologica Academiae Scientiarum Hungaricae*, *44*(1-2), 191–204.

Freitas, C., Manzato, E., Burini, A., Taylor, M. J., Lerch, J. P., & Anagnostou, E. (2018). Neural correlates of familiarity in music listening: A systematic review and a neuroimaging meta-analysis. *Frontiers in Neuroscience*.

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

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in cognitive sciences*, *13*(7), 293–301.

Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological cybernetics*, *102*(3), 227–260.

Fu, Y., Jha, D. K., Zhang, Z., Yuan, Z., & Ray, A. (2019). Neural network-based learning from demonstration of an autonomous ground robot. *Machines*, *7*(2).

Fung, C. V. (1993). A review of studies on non-western music preference. *Update: Applications of Research in Music Education*, *12*(1), 26–32.

Furman, C. E., & Duke, R. A. (1988). Effect of majority consensus on preferences for recorded orchestral and popular music. *Journal of Research in Music Education*, *36*(4), 220–231.

G, H. (2012). Computational neuroanatomy of speech production. *Nat. Rev. Neurosci*, *13*.

Gelding, R. W., Thompson, W. F., & Johnson, B. W. (2019). Musical imagery depends upon coordination of auditory and sensorimotor brain activity. *Scientific Reports*, *9*(1), 16823.

Gelding, R. W., Thompson, W. F., & Johnson, B. W. (2015). The pitch imagery arrow task: Effects of musical training, vividness, and mental control. *PloS one*, *10*(3), e0121809–e0121809.

Georg, K., & Hr, H. R. (2009). Neural processing of auditory feedback during vocal practice in a songbird. *Nature*, *457*, 187–190.

Georg, K., Tobias, B., & Mark, H. (2012). Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, *74*, 809–815.

Georges, M.-A., Girin, L., Schwartz, J.-L., & Hueber, T. (2021). Learning Robust Speech Representation with an Articulatory-Regularized Variational Autoencoder. *Proceedings Interspeech 2021*, 3345–3349.

Gerhardstein, R. C. (2002). The historical roots and development of audiation: A process for musical understanding. *In Hanley, B. Goolsby, T.W. (Eds.) Musical understanding: Perspectives in theory and practice*.

Germine, L., Russell, R., Bronstad, P. M., Blokland, G. A., Smoller, J. W., Kwok, H., Anthony, S. E., Nakayama, K., Rhodes, G., & Wilmer, J. B. (2015). Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Current Biology*, *25*(20), 2684–2689.

Gillick, J., Tang, K., & Keller, R. (2010). Learning jazz grammars. *Computer Music Journal*, *34*, 56–66.

Gillick, J., Tang, K., & Keller, R. M. (2009). Learning jazz grammars.

Gilliland, A. R., & Moore, H. T. (1924). The immediate and long-time effects of classical and popular phonograph selections. *Journal of Applied Psychology*, *8*(3), 309–323.

Godoy, R., & Jorgensen, H. (2012). *Musical imagery*. Taylor & Francis.
Gold, B. P., Mas-Herrero, E., Zeighami, Y., Benovoy, M., Dagher, A., & Zatorre, R. J. (2019). Musical reward prediction errors engage the nucleus accumbens and motivate learning. *Proceedings of the National Academy of Sciences*, *116*(8), 3310–3315.
Gold, B. P., Pearce, M. T., Mas-Herrero, E., Dagher, A., & Zatorre, R. J. (2019). Predictability and uncertainty in the pleasure of music: A reward for learning? *Journal of Neuroscience*, *39*(47), 9397–9409.
Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.
Gregory, C., Chad, T. T., Werner, C., Orrin, D., Bijan, D., & Pesaran. (2014). Sensory-motor transformations for speech occur bilaterally. *Nature*, *507*, 94–98.
Gregory, H. (2014). The myth of mirror neurons: The real neuroscience of communication and cognition.
Griffiths, T. D. (1999). Human complex sound analysis*. *Clinical Science*, *96*(3), 231–234.
Grisoni, L., Mohr, B., & Pulvermüller, F. (2019). Prediction mechanisms in motor and auditory areas and their role in sound perception and language understanding. *NeuroImage*, *199*, 206–216.
Guillemin, C., & Tillmann, B. (2021). Implicit learning of two artificial grammars. *Cognitive Processing*, *22*(1), 141–150.
Hadjeres, G., Pachet, F., & Nielsen, F. (2017). Deepbach: A steerable model for bach chorales generation. *International Conference on Machine Learning*, 1362–1371.
Halpern, A. (2015). Differences in auditory imagery self-report predict neural and behavioral outcomes. *Psychomusicology: Music, Mind and Brain*, *25*, 37–47.
Halpern, A. R. (2001). Cerebral substrates of musical imagery. *Annals of the New York Academy of Sciences*, *930*(1), 179–192.
Halpern, A. R., & Zatorre, R. J. (1999). When That Tune Runs Through Your Head: A PET Investigation of Auditory Imagery for Familiar Melodies. *Cerebral Cortex*, *9*(7), 697–704.
Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*(9), 1281–1292.
Halpern, A. R., Zioga, I., Shankleman, M., Lindsen, J., Pearce, M. T., & Bhattacharya, J. (2017). That note sounds wrong! age-related effects in processing of musical expectation. *Brain and cognition*, *113*, 1–9.
Hannon, E., & Trainor, L. (2007). Music acquisition: Effects of enculturation and formal training on development. *Trends in cognitive sciences*, *11*, 466–72.
Hannon, E. E., & Trehub, S. E. (2005a). Metrical categories in infancy and adulthood. *Psychological science*, *16*(1), 48–55.
Hannon, E. E., & Trehub, S. E. (2005b). Tuning in to musical rhythms: Infants learn more readily than adults. *Proceedings of the National Academy of Sciences*, *102*(35), 12639–12643.
Hansen, N. C., & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in psychology*, *5*, 1052.

Hargreaves, D. J., Comber, C., & Colley, A. (1995). Effects of age, gender, and training on musical preferences of british secondary school students. *Journal of Research in Music Education*, *43*(3), 242–250.

Haumann, N. T., Vuust, P., Bertelsen, F., & Garza-Villarreal, E. A. (2018). Influence of musical enculturation on brain responses to metric deviants. *Frontiers in neuroscience*, *12*, 218.

Haworth, C. M. A., Davis, O. S. P., & Plomin, R. (2013). Twins early development study (teds): A genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Research and Human Genetics*, *16*, 117–125.

Heaton, P., Tsang, W. F., Jakubowski, K., Mullensiefen, D., & Allen, R. (2018). Discriminating autism and language impairment and specific language impairment through acuity of musical imagery. *Research in Developmental Disabilities*, *80*, 52–63.

Heeger, D. J. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences*, *114*(18), 1773–1782.

Heilbron, M., & Chait, M. (2018). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*, *389*, 54–73.

Heingartner, A., & Hall, J. V. (1974). Affective consequences in adults and children of repeated exposure to auditory stimuli. *Journal of Personality and Social Psychology*, *29*(6), 719.

Henrich, J. (2008). A cultural species. *Explaining culture scientifically*, 184–210.

Herholz, S. C., Halpern, A. R., & Zatorre, R. J. (2012). Neuronal correlates of perception, imagery, and memory for familiar tunes [PMID: 22360595]. *Journal of Cognitive Neuroscience*, *24*(6), 1382–1397.

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

Hsu, Y.-F., Le Bars, S., Hämäläinen, J. A., & Waszak, F. (2015). Distinctive representation of mispredicted and unpredicted prediction errors in human electroencephalography. *Journal of Neuroscience*, *35*(43), 14653–14660.

Huang, C. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Hawthorne, C., Dai, A. M., Hoffman, M. D., & Eck, D. (2018). An improved relative self-attention mechanism for transformer with application to music generation. *CoRR*, *abs/1809.04281*.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music transformer. *arXiv preprint arXiv:1809.04281*.

Huang, N., & Elhilali, M. (2017). Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America*, *141*, 2163–2176.

Hubbard, T. (2013). Multisensory imagery. In L. R. Lacey S. (Ed.). Springer, New York, NY.

Hunter, M. D. (2021). Multilevel modeling in classical twin and modern molecular behavior genetics. *Behavior Genetics*, *51*, 301–318.

Huron, D. (2006). Sweet anticipation.

Inglefield, H. G. (1968). *The relationship of selected personality variables to conformity behavior reflected in the musical preferences of adolescents when exposed to peer group leader influences* [Doctoral dissertation, The Ohio State University].

Inglefield, H. G. (1972). Conformity behavior reflected in the musical preference of adolescents. *Contributions to Music Education*, 56–67.

J, P., S, P., S, N., & R, M. (2008). Precise auditory-vocal mirroring in neurons for learned vocal communication. *Nature*, *451*, 305–310.

Jæger, M., & Møllegaard, S. (2022). Where do cultural tastes come from?: Genes, environments, or experiences [Publisher Copyright: Copyright: c 2022 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited.]. *Sociological Science*, *9*, 252–274.

Jagiello, R., Pomper, U., Yoneya, M., Zhao, S., & Chait, M. (2019). Rapid brain responses to familiar vs. unfamiliar music – an eeg and pupillometry study. *Scientific Reports*, *9*, 1–13.

Janata, P. (2001). Brain electrical activity evoked by mental formation of auditory expectations and images. *Brain topography*, *13*, 169–93.

Janata, P. (2015). Chapter 11 - neural basis of music perception. In M. J. Aminoff, F. Boller, & D. F. Swaab (Eds.), *The human auditory system* (pp. 187–205, Vol. 129). Elsevier.

Jäncke, L., Loose, R., Lutz, K., Specht, K., & Shah, N. (2000). Cortical activations during paced finger-tapping applying visual and auditory pacing stimuli. *Cognitive Brain Research*, *10*(1), 51–66.

Jason, T., Kevin, R., & Frank, G. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, *39*, 1429–1443.

John, H., & Edward, C. (2015). The cortical computations underlying feedback control in vocal production. *Current opinion in neurobiology*, *33*, 174–181.

John, H., & Michael, J. (2002). Sensorimotor adaptation of speech i. *Journal of Speech, Language, and Hearing Research*.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive psychology*, *21*(1), 60–99.

Johnson, M. K., Kim, J. K., & Risse, G. (1985). Do alcoholic korsakoff's syndrome patients acquire affective reactions? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(1), 22.

Josef, R., & Sophie, S. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature neuroscience*, *12*, 718–724.

Joutsiniemi, S.-L., & Hari, R. (1989). Omissions of auditory stimuli may activate frontal cortex. *European Journal of Neuroscience*, *1*(6), 524–528.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140169.

Kathios, N., & Loui, P. (2022). *Evolutionary Studies in Imaginative Culture*, *6*(1), 33–38.

Keller, G. B., Bonhoeffer, T., & Hübener, M. (2012). Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, *74*(5), 809–815.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435.

Kessler, E. J., Hansen, C., & Shepard, R. N. (1984). Tonal schemata in the perception of music in bali and in the west. *Music Perception*, *2*(2), 131–165.

Killian, J. N. (1990). Effect of model characteristics on musical preference of junior high students. *Journal of Research in Music Education*, *38*(2), 115–123.

Kilteni, K., & Ehrsson, H. H. (2017). Sensorimotor predictions and tool use: Hand-held tools attenuate self-touch. *Cognition*, *165*, 1–9.

Kimura, M. (2012). Visual mismatch negativity and unintentional temporal-context-based prediction in vision. *International Journal of Psychophysiology*, *83*(2), 144–155.
Koelsch, S. (2009). Music-syntactic processing and auditory memory: Similarities and differences between eran and mmn. *Psychophysiology*, *46*(1), 179–190.
Koelsch, S. (2020). A coordinate-based meta-analysis of music-evoked emotions. *NeuroImage*, *223*, 117350.
Koelsch, S., Gunter, T., Friederici, A. D., & Schröger, E. (2000). Brain indices of music processing:"nonmusicians" are musical. *Journal of Cognitive Neuroscience*, *12*(3), 520–541.
Koelsch, S., & Mulder, J. (2002). Electric brain responses to inappropriate harmonies during listening to expressive music. *Clinical Neurophysiology*, *113*(6), 862–869.
Koelsch, S., & Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences*, *9*(12), 578–584.
Koelsch, S., Vuust, P., & Friston, K. (2019). Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences*, *23*(1), 63–77.
Koenig, L. B. (2020). Twin studies in personality research. In *The wiley encyclopedia of personality and individual differences* (1st ed., pp. 415–419). Wiley.
Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron*, *79*(5), 836–848.
Kraemer, D. J. M., Macrae, C. N., Green, A. E., & Kelley, W. M. (2005). Sound of silence activates auditory cortex. *Nature*, *434*(7030), 158–158.
Krumhansl, C., Louhivuori, J., Toiviainen, P., Järvinen, T., & Eerola, T. (1999). Melodic expectation in finnish spiritual folk hymns: Convergence of statistical, behavioral, and computational approaches. *Music Perception*, *17*, 151–195.
Krumhansl, C., Toivanen, P., Eerola, T., Toiviainen, P., Järvinen, T., & Louhivuori, J. (2000). Cross-cultural music cognition: Cognitive methodology applied to north sami yoiks. *Cognition*, *76*, 13–58.
Krumhansl, C. L. (1995). Effects of musical context on similarity and expectancy. *Systematische Musikwissenschaft*, *3*(2), 211–250.
Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *51*(4), 336.
Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological review*, *89*(4), 334.
Krumhansl, C. L., Louhivuori, J., Toiviainen, P., Järvinen, T., & Eerola, T. (1999). Melodic expectation in finnish spiritual folk hymns: Convergence of statistical, behavioral, and computational approaches. *Music Perception*, *17*(2), 151–195.
Krumhansl, C. L., Toivanen, P., Eerola, T., Toiviainen, P., Järvinen, T., & Louhivuori, J. (2000). Cross-cultural music cognition: Cognitive methodology applied to north sami yoiks. *Cognition*, *76*(1), 13–58.
Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831–843.
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, *62*, 621–647.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163.

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*(1), 189–193.

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009a). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology*, *102*(1), 349–359.

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009b). Resolving precise temporal processing properties of the auditory system using continuous stimuli [PMID: 19439675]. *Journal of Neurophysiology*, *102*(1), 349–359.

Le Vaillant, G., Dutoit, T., & Dekeyser, S. (2021). Improving synthesizer programming from variational autoencoders latent space. *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)*.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

LeBlanc, A., Jin, Y. C., Stamou, L., & McCrary, J. (1999). Effect of age, country, and gender on music listening preferences. *Bulletin of the Council for Research in Music Education*, 72–76.

LeBlanc, A. (1979). Generic style music preferences of fifth-grade students. *Journal of Research in Music Education*, *27*(4), 255–270.

Lee, H., Höger, F., Schönwiesner, M., Park, M., & Jacoby, N. (2021). Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms. *CoRR*, *abs/2108.00768*.

Lee, Jung, & Loui, P. (2019). Attention modulates electrophysiological responses to simultaneous music and language syntax processing. *Brain Sciences*, *9*, 305.

Leino, S., Brattico, E., Tervaniemi, M., & Vuust, P. (2007). Representation of harmony rules in the human brain: Further evidence from event-related potentials. *Brain research*, *1142*, 169–177.

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature communications*, *7*(1), 1–9.

Lerdahl, F. (2004). *Tonal pitch space*. Oxford University Press.

Lerdahl, F., & Jackendoff, R. S. (1996). *A generative theory of tonal music, reissue, with a new preface*. MIT press.

Lerdahl, F., & Krumhansl, C. L. (2007). Modeling tonal tension. *Music Perception*, *24*(4), 329–366.

Leung, Y., & Dean, R. T. (2018). Learning unfamiliar pitch intervals: A novel paradigm for demonstrating the learning of statistical associations between musical pitches. *PloS one*, *13*(8), e0203026.

Liberto, G. M. D., Marion, G., & Shamma, S. A. (2021). Accurate decoding of imagined and heard melodies. *Frontiers in Neuroscience*, *15*.

Lima, C. F., Krishnan, S., & Scott, S. K. (2016). Roles of supplementary motor areas in auditory processing and auditory imagery. *Trends in Neurosciences*, *39*(8), 527–542.

Lima, C. F., Lavan, N., Evans, S., Agnew, Z., Halpern, A. R., Shanmugalingam, P., Meekings, S., Boebinger, D., Ostarek, M., McGettigan, C., Warren, J. E., & Scott, S. K. (2015).

Feel the Noise: Relating Individual Differences in Auditory Imagery to the Structure and Function of Sensorimotor Systems. *Cerebral Cortex*, *25*(11), 4638–4650.
Loui, P. (2012). Learning and liking of melody and harmony: Further studies in artificial grammar learning. *Topics in Cognitive Science*, *4*(4), 554–567.
Loui, P., Grent, T., Torpey, D., Woldorff, M., *et al.* (2005). Effects of attention on the neural processing of harmonic syntax in western music. *Cognitive Brain Research*, *25*(3), 678–687.
Loui, P., & Wessel, D. (2008). Learning and liking an artificial musical system: Effects of set size and repeated exposure. *Musicae Scientiae*, *12*(2), 207–230.
Loui, P., Wessel, D., & Kam, C. H. (2006). Acquiring new musical grammars: A statistical learning approach.
Loui, P., Wessel, D. L., & Kam, C. L. H. (2010). Humans rapidly learn grammatical structure in a new musical scale. *Music perception*, *27*(5), 377–388.
Lynch, M. P., & Eilers, R. E. (1992). A study of perceptual development for musical tuning. *Perception & Psychophysics*, *52*(6), 599–608.
MacKay, D. (2003, January). *Informationtheory, inference, and learning algorithms* (Vol. 50).
Mado Proverbio, A., Calbi, M., Manfredi, M., & Zani, A. (2014). Audio-visuomotor processing in the musician's brain: An erp study on professional violinists and clarinetists. *Scientific Reports*, *4*(1), 5866.
Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
Manzara, L. C., Witten, I. H., & James, M. (1992). On the entropy of music: An experiment with bach chorale melodies. *Leonardo Music Journal*, *2*(1), 81–88.
Marc, S., & Robert, W. (2002). A pathway in primate brain for internal monitoring of movements. *Science*, *296*, 1480–1482.
Marco, I. (2009). Imitation, empathy, and mirror neurons. *Annual review of psychology*, *60*.
Margulis, E. H. (2003). Melodic expectation: A discussion and model.
Margulis, E. H. (2013). Aesthetic responses to repetition in unfamiliar music. *Empirical Studies of the Arts*, *31*(1), 45–57.
Margulis, E. H. (2014). *On repeat, on repeat*. Oxford University Press.
Margulis, E. H., & Levine, W. H. (2006). Timbre priming effects and expectation in melody. *Journal of New Music Research*, *35*(2), 175–182.
Marion, G., Di Liberto, G. M., & Shamma, S. A. (2021). The music of silence: Part i: Responses to musical imagery encode melodic expectations and acoustics. *Journal of Neuroscience*, *41*(35), 7435–7448.
Marmel, F., Tillmann, B., & Delbé, C. (2010). Priming in melody perception: Tracking down the strength of cognitive expectations. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 1016.
Marmel, F., Tillmann, B., & Dowling, W. J. (2008). Tonal expectations influence pitch perception. *Perception & Psychophysics*, *70*(5), 841–852.
Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., & Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *The Journal of Neuroscience*, *28*(50), 12539–12545.

Martin, S., Mikutta, C., Leonard, M. K., Hungate, D., Koelsch, S., Shamma, S., Chang, E. F., Millán, J. d. R., Knight, R. T., & Pasley, B. N. (2017). Neural Encoding of Auditory Features during Music Perception and Imagery. *Cerebral Cortex*, *28*(12), 4222–4233.

Martindale, C., & Moore, K. (1989). Relationship of musical preference to collative, ecological, and psychophysical variables. *Music Perception*, *6*(4), 431–445.

Martindale, C., Moore, K., & Borkum, J. (1990). Aesthetic preference: Anomalous findings for berlyne's psychobiological theory. *The American Journal of Psychology*, 53–80.

Masanori, M., Fumiya, Y., & Kenji, O. (2016). World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, *99*, 1877–1884.

Mas-Herrero, E., Maini, L., Sescousse, G., & Zatorre, R. J. (2021). Common and distinct neural correlates of music and food-induced pleasure: A coordinate-based meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *123*, 61–71.

May, W. V. (1985). Musical style preferences and aural discrimination skills of primary grade school children. *Journal of Research in Music Education*, *33*(1), 7–22.

McClelland, J. L., & Rumelhart, D. E. (1988). An interactive activation model of context effects in letter perception: Part 1.: An account of basic findings.

McCrary, J., & Gauthier, D. (1995). The effects of performers' ethnic identities on preadolescents' music preferences. *Update: Applications of Research in Music Education*, *14*(1), 20–22.

Meadows, E. S. (1970). *The relationship of music preference to certain cultural determiners* [Doctoral dissertation, Michigan State University. Department of Music].

Meister, I., Krings, T., Foltys, H., Boroojerdi, B., Müller, M., Töpper, R., & Thron, A. (2004). Playing piano in the mind—an fmri study on music imagery and performance in pianists. *Cognitive Brain Research*, *19*(3), 219–228.

Mellander, C., Florida, R., Rentfrow, P. J., & Potter, J. (2018). The geography of music preferences. *Journal of Cultural Economics*, *42*, 593–618.

Meyer, L. (1956). *Emotion and meaning in music*. University of Chicago Press.

Meyer, L. (1973). *Explaining music: Essays and explorations*. Berkeley, CA: University of California Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

Miller, K. J., Schalk, G., Fetz, E. E., den Nijs, M., Ojemann, J. G., & Rao, R. P. N. (2010). Cortical activity during motor execution, motor imagery, and imagery-based online feedback. *Proceedings of the National Academy of Sciences*, *107*(9), 4430–4435.

Moffat, A. (1990). Implementing the ppm data compression scheme. *IEEE Transactions on Communications*, *38*(11), 1917–1921.

Morgan, E., Fogel, A., Nair, A., & Patel, A. D. (2019). Statistical learning and gestalt-like principles predict melodic expectations. *Cognition*, *189*, 23–34.

Morrison, S., Demorest, S., & Stambaugh, L. (2008). Enculturation effects in music cognition: The role of age and music complexity. *Journal of Research in Music Education*, *56*.

Mosing, M. A., & Ullén, F. (2018). Genetic influences on musical specialization: A twin study on choice of instrument and music genre. *Annals of the New York Academy of Sciences*, *1423*(1), 427–434.

Mull, R. M., & Hennessy, J. F. (1957). The effect of familiarity on aesthetic preference for music. *Journal of General Psychology*, *57*(2), 203–212.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (mmn) in basic research of central auditory processing: A review. *Clinical neurophysiology*, *118*(12), 2544–2590.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model.* University of Chicago Press.

Neale, M., & Cardon, L. (1992). *Methodology for genetic studies of twins and families*. Springer Science & Business Media.

Nettl, B. (2015). *The study of ethnomusicology: Thirty-three discussions*. University of Illinois Press.

Nima, M., Connie, C., Keith, J., & Edward, C. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*, 1006–1010.

Nima, M., Malcolm, S., & Shihab, S. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 920–930.

Nixon, J. S., & Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition*, *212*, 104697.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18.

Omigie, D., Pearce, M., Lehongre, K., Hasboun, D., Navarro, V., Adam, C., & Samson, S. (2019). Intracranial recordings and computational modeling of music reveal the time course of prediction error signaling in frontal and temporal cortices. *Journal of Cognitive Neuroscience*, *31*(6), 855–873.

Omigie, D., Pearce, M., & Stewart, L. (2012). Tracking of pitch probabilities in congenital amusia. *Neuropsychologia*, *50*, 1483–93.

Omigie, D., Pearce, M. T., Lehongre, K., Hasboun, D., Navarro, V., Adam, C., & Samson, S. (2019). Intracranial recordings and computational modelling of music reveal the time-course of prediction error signaling in frontal and temporal cortices. *Journal of Cognitive Neuroscience*, *31*(6), 855–873.

Omigie, D., Pearce, M. T., & Stewart, L. (2012). Tracking of pitch probabilities in congenital amusia. *Neuropsychologia*, *50*(7), 1483–1493.

Omigie, D., Pearce, M. T., Williamson, V. J., & Stewart, L. (2013a). Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia*, *51*(9), 1749–1762.

Omigie, D., Pearce, M. T., Williamson, V. J., & Stewart, L. (2013b). Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia*, *51*(9), 1749–1762.

Omigie, D., & Stewart, L. (2011). Preserved statistical learning of tonal and linguistic material in congenital amusia. *Frontiers in Psychology*, *2*, 109.

Opitz, B., Rinne, T., Mecklinger, A., von Cramon, D., & Schröger, E. (2002). Differential contribution of frontal and temporal cortices to auditory change detection: Fmri and erp results. *NeuroImage*, *15*(1), 167–174.

Oram, N., Cuddy, L. L., & Oram, N. (1995). Responsiveness of western adults to pitch-distributional information in melodic sequences. *Psychological Research*, *57*(2), 103–118.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral Cortex*.

Pagliarini, S., Leblois, A., & Hinaut, X. (2021). Canary Vocal Sensorimotor Model with RNN Decoder and Low-dimensional GAN Generator. *2021 IEEE International Conference on Development and Learning (ICDL)*, 1–8.

Pantev, C., Roberts, L. E., Schulz, M., Engelien, A., & Ross, B. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *NeuroReport*, *12*(1), 169–174.

Parrell, Adam, B., Gregory, L., Thomas, C., & Quatieri. (2019). Current models of speech motor control: A control-theoretic overview of architectures and properties. *The Journal of the Acoustical Society of America*, *145*, 1456–1481.

Pearce, M., & Wiggins, G. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*(4), 367–385.

Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* (Publication No. 4) [Doctoral dissertation, City University London].

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*(1), 378–395.

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2008). A comparison of statistical and rule-based models of melodic segmentation. *ISMIR*, 89–94.

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*(1), 302–313.

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, *23*(5), 377–405.

Pearce, M. T., & Wiggins, G. A. (2012). Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science*, *4*(4), 625–652.

Pelofi, C., Marion, G., Di Liberto, G., Ripolles, P., & Shamma, S. (n.d.). Cross-cultural perspectives on the predictive coding perspective of music perception. *In prep*.

Pelofi, C., De Gardelle, V., Egré, P., & Pressnitzer, D. (2017). Interindividual variability in auditory scene analysis revealed by confidence judgements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160107.

Pereira, C. S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S. L., & Brattico, E. (2011). Music and emotions in the brain: Familiarity matters. *PloS one*, *6*(11), e27241.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in cognitive sciences*, *10*(5), 233–238.

Perruchet, P., & Vinter, A. (1998). Parser: A model for word segmentation. *Journal of memory and language*, *39*(2), 246–263.

Peterson, R. A. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics*, *21*(4), 243–258.

Peterson, R. A., & Kern, R. M. (1996). Changing highbrow taste: From snob to omnivore. *American Sociological Review*, *61*(5), 900–907.

Peterson, R. A., & Simkus, A. (1992). How musical tastes mark occupational status groups. *Cultivating differences: Symbolic boundaries and the making of inequality*, *152*.

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, *29*(1-2), 34–55.

Poeppel, D. (2014). The neuroanatomic and neurophysiological infrastructure for speech and language. *Current Opinion in Neurobiology*, *28*, 142–149.
Poeppel, D., Emmorey, K., Hickok, G., & PylkkÃ€nen, L. (2012). Towards a new neurobiology of language. *Journal of Neuroscience*, *32*, 14125–14131.
Polak, R., Jacoby, N., Fischinger, T., Goldberg, D., Holzapfel, A., & London, J. (2018). Rhythmic prototypes across cultures: A comparative study of tapping synchronization. *Music Perception: An Interdisciplinary Journal*, *36*(1), 1–23.
Politimou, N., Douglass-Kirk, P., Pearce, M., Stewart, L., & Franco, F. (2021). Melodic expectations in 5-and 6-year-old children. *Journal of Experimental Child Psychology*, *203*, 105020.
Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*.
Poulet, F, J., & Berthold, H. (2006). The cellular basis of a corollary discharge. *Science*, *311*, 518–522.
Power, R. A., & Pluess, M. (2015). Heritability estimates of the big five personality traits based on common genetic variants. *Translational Psychiatry*, *5*, e604.
Pressnitzer, D., J., G., C., C., V., d., & P., E. (2018). Auditory perception: Laurel and yanny together at last. *Current Biology*, *28*(13), R739–R741.
Price, T., Freeman, B., Craig, I., Petrill, S., Ebersole, L., & Plomin, R. (2012). Infant zygosity can be assigned by parental report questionnaire data. *Twin Research*, *3*, 129–133.
Pruitt, T., Halpern, A., & Pfordresher, P. (2018). Covert singing in anticipatory auditory imagery. *Psychophysiology*, *56*.
Quiroga-Martinez, D. R., Hansen, N. C., Højlund, A., Pearce, M. T., Brattico, E., & Vuust, P. (2019). Reduced prediction error responses in high-as compared to low-uncertainty musical contexts. *Cortex*, *120*, 181–200.
Quiroga-Martinez, D. R., C. Hansen, N., Højlund, A., Pearce, M., Brattico, E., & Vuust, P. (2020). Musical prediction error responses similarly reduced by predictive uncertainty in musicians and non-musicians. *European Journal of Neuroscience*, *51*(11), 2250–2269.
Quiroga-Martinez, D. R., Hansen, N. C., Højlund, A., Pearce, M., Brattico, E., & Vuust, P. (2020). Decomposing neural responses to melodic surprise in musicians and non-musicians: Evidence for a hierarchy of predictions in the auditory system. *NeuroImage*, *215*, 116816.
Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the n400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*, 693–705.
Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.
Reck, D. B. (1977). *Music of the whole earth*. Macmillan Reference USA.
Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of personality and social psychology*, *84*(6), 1236.
Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin & Review*, *12*(6), 969–992.
Repp, B. H., & Penel, A. (2004). Rhythmic movement is attracted more strongly to auditory than to visual rhythms. *Psychological Research*, *68*(4), 252–270.

Rijsdijk, F. V., & Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. *Briefings in Bioinformatics*, *3*(2), 119–133.

Ripollés, P., Ferreri, L., Mas-Herrero, E., Alicart, H., Gómez-Andrés, A., Marco-Pallares, J., Antonijoan, R. M., Noesselt, T., Valle, M., Riba, J., & Rodriguez-Fornells, A. (2018). Intrinsically regulated learning is modulated by synaptic dopamine signaling (V. Murty, J. I. Gold, V. Murty, & B. Larsen, Eds.). *eLife*, *7*, e38113.

Ripollés, P., Marco-Pallarés, J., Alicart, H., Tempelmann, C., Rodríguez-Fornells, A., & Noesselt, T. (2016). Intrinsic monitoring of learning success facilitates memory encoding via the activation of the SN/VTA-Hippocampal loop (V. Murty, Ed.). *eLife*, *5*, e17441.

Ripollés, P., Marco-Pallarés, J., Hielscher, U., Mestres-Missé, A., Tempelmann, C., Heinze, H.-J., Rodríguez-Fornells, A., & Noesselt, T. (2014). The role of reward in word learning and its implications for language acquisition. *Current Biology*, *24*(21), 2606–2611.

Ritossa, D. A., & Rickard, N. S. (2004). The relative utility of 'pleasantness' and 'liking' dimensions in predicting the emotions expressed by music. *Psychology of Music*, *32*(1), 5–22.

Roger, C., & W, A. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, *1*, 89–97.

Rohrmeier, M. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music*, *5*(1), 35–53.

Rohrmeier, M., & Cross, I. (2008). Statistical properties of tonal harmony in bach's chorales. *Hokkaido University Sapporo, Japan*.

Rohrmeier, M., & Cross, I. (2009). Tacit tonality-implicit learning of context-free harmonic structure.

Rohrmeier, M., & Cross, I. (2013). Artificial grammar learning of melody is constrained by melodic inconsistency: Narmour's principles affect melodic learning. *PloS one*, *8*(7), e66174.

Rohrmeier, M., & Rebuschat, P. (2012). Implicit learning and acquisition of music. *Topics in Cognitive Science*, *4*(4), 525–553.

Rohrmeier, M., Rebuschat, P., & Cross, I. (2011). Incidental and online learning of melodic structure. *Consciousness and cognition*, *20*(2), 214–222.

Rohrmeier, M., & Widdess, R. (2017). Incidental learning of melodic structure of north indian music. *Cognitive science*, *41*(5), 1299–1327.

Rohrmeier, M. A., & Koelsch, S. (2012). Predictive information processing in music cognition. a critical review. *International Journal of Psychophysiology*, *83*(2), 164–175.

S, J., & M, H. (1997). Visual control of hand action. *Trends Cogn. Sci*, *1*, 310–317.

Saarinen, J., Paavilainen, P., Schöger, E., Tervaniemi, M., & Näätänen, R. (1992). Representation of abstract attributes of auditory stimuli in the human brain. *NeuroReport*, *3*(12), 1149–1151.

Saha, P., & Fels, S. (2020). Learning Joint Articulatory-Acoustic Representations with Normalizing Flows. *Proceedings Interspeech 2020*, 3196–3200.

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience*, *14*(2), 257–262.

Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., & McIntosh, A. R. (2015). Predictions and the brain: How musical sounds become rewarding. *Trends in cognitive sciences*, *19*(2), 86–91.

Sauvé, S., Sayed, A., Dean, R., & Pearce, M. (2018). Effects of pitch and timing expectancy on musical emotion. *Psychomusicology: Music, Mind, and Brain*, *28*.

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Music as a coevolved system for social bonding. *Behavioral and Brain Sciences*, *44*.

Schäfer, T., & Mehlhorn, C. (2017). Can personality traits predict musical style preferences? a meta-analysis. *Personality and Individual Differences*, *116*, 265–273.

Schellenberg, E. G. (1997). Simplifying the implication-realization model of melodic expectancy. *Music Perception*, *14*(3), 295–318.

Schellenberg, E. G., & Trehub, S. E. (1999). Culture-general and culture-specific factors in the discrimination of melodies. *Journal of experimental child psychology*, *74*(2), 107–127.

Schönwiesner, M., Novitski, N., Pakarinen, S., Carlson, S., Tervaniemi, M., & Näätänen, R. (2007). Heschl's gyrus, posterior superior temporal gyrus, and mid-ventrolateral prefrontal cortex have different roles in the detection of acoustic changes [PMID: 17182905]. *Journal of Neurophysiology*, *97*(3), 2075–2082.

Schubotz, R. I. (2007). Prediction of external events with our motor system: Towards a new framework. *Trends in Cognitive Sciences*, *11*(5), 211–218.

Schuessler, K. F. (1948). Social background and musical taste. *American Sociological Review*, *13*(3), 330–335.

Sears, D. R., Pearce, M. T., Caplin, W. E., & McAdams, S. (2018). Simulating melodic and harmonic expectations for tonal cadences using probabilistic models. *Journal of New Music Research*, *47*(1), 29–52.

Seer, C., Lange, F., Boos, M., Dengler, R., & Kopp, B. (2016). Prior probabilities modulate cortical surprise responses: A study of event-related potentials. *Brain and Cognition*, *106*, 78–89.

Sellström, E., & Bremberg, S. (2006). Is there a "school effect" on pupil outcomes? a review of multilevel studies. *Journal of Epidemiology & Community Health*, *60*(2), 149–155.

Shamma, S., Patel, P., Mukherjee, S., Marion, G., Khalighinejad, B., Han, C., Herrero, J., Bickel, S., Mehta, A., & Mesgarani, N. (2020). Learning Speech Production and Perception through Sensorimotor Interactions [tgaa091]. *Cerebral Cortex Communications*, *2*(1).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shany, O., Singer, N., Gold, B. P., Jacoby, N., Tarrasch, R., Hendler, T., & Granot, R. (2019). Surprise-related activation in the nucleus accumbens interacts with music-induced pleasantness. *Social cognitive and affective neuroscience*, *14*(4), 459–470.

Simson, R., Vaughan Jr, H. G., & Walter, R. (1976). The scalp topography of potentials associated with missing visual or auditory stimuli. *Electroencephalography and Clinical Neurophysiology*, *40*(1), 33–42.

Siriwardena, Y. M., Marion, G., & Shamma, S. (2022). The mirrornet: Learning audio synthesizer controls inspired by sensorimotor interaction. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Skerritt-Davis, B., & Elhilali, M. (2018). Detecting change in stochastic sound sequences. *PLOS Computational Biology*, *14*(5), 1–24.

Skerritt-Davis, B., & Elhilali, M. (2019). A model for statistical regularity extraction from dynamic sounds. *Acta acustica united with acustica : the journal of the European Acoustics Association (EEIG)*, *105*(1), 1–4.

Smith, A. D., Fildes, A., Cooke, L., Herle, M., Shakeshaft, N., Plomin, R., & Llewellyn, C. (2016). Genetic and environmental influences on food preferences in adolescence12. *The American Journal of Clinical Nutrition*, *104*(2), 446–453.

Smith, A. D., Fildes, A., Forwood, S., Cooke, L., & Llewellyn, C. (2017). The individual environment, not the family is the most important influence on preferences for common non-alcoholic beverages in adolescence. *Scientific Reports*, *7*(1), 16822.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, 549–565.

Snyder, J. S., Schwiedrzik, C. M., Vitela, A. D., & Melloni, L. (2015). How previous experience shapes perception in different sensory modalities. *Frontiers in human neuroscience*, *9*, 594.

Sohoglu, E., & Chait, M. (2016). Detecting and representing predictable structure during auditory scene analysis. *Elife*, *5*, e19113.

Soley, G., & Hannon, E. (2010). Infants prefer the musical meter of their own culture: A cross-cultural comparison. *Developmental psychology*, *46*, 286–92.

Song, Y., Dixon, S., Pearce, M. T., & Halpern, A. R. (2016). Perceived and induced emotion responses to popular music: Categorical and dimensional models. *Music Perception: An Interdisciplinary Journal*, *33*(4), 472–492.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92–97.

Steinbeis, N., Koelsch, S., & Sloboda, J. A. (2006). The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. *Journal of cognitive neuroscience*, *18*(8), 1380–1393.

Stephen, W., Martin, A. P. S., Marco, S., & Iacoboni. (2004). Listening to speech activates motor areas involved in speech production. *Nature neuroscience*, *7*.

Stevens, C. (2004). Cross-cultural studies of musical pitch and time. *Acoustical science and technology*, *25*(6), 433–438.

Strauss, A., Kotz, S. A., & Obleser, J. (2013). Narrowed expectancies under degraded speech: Revisiting the n400. *Journal of Cognitive Neuroscience*, *25*(9), 1383–1395.

Stupacher, J., Witek, M., Vuoskoski, J., & Vuust, P. (2020). Cultural familiarity and individual musical taste differently affect social bonding when moving to music. *Scientific Reports*, *10*, 10015.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187–1188.

Tai, L., & Liu, M. (2016). Deep-learning in mobile robotics - from perception to control systems: A survey on why and why not. *CoRR*, *abs/1612.07139*.

Taishih, C., Powen, R., & Shihab, S. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*, 887–906.

Tal, I., Large, E. W., Rabinovitch, E., Wei, Y., Schroeder, C. E., Poeppel, D., & Golumbic, E. Z. (2017). Neural entrainment to the beat: The "missing-pulse" phenomenon. *Journal of Neuroscience*, *37*, 6331–6341.

Tamimy, Z., Kevenaar, S. T., Hottenga, J. J., Hunter, M. D., de Zeeuw, E. L., Neale, M. C., van Beijsterveldt, C. E. M., Dolan, C. V., van Bergen, E., & Boomsma, D. I. (2021). Multilevel twin models: Geographical region as a third level variable. *Behavior Genetics*, *51*(3), 319–330.

Tanner, F. D. (1976). *The effect of disc jockey approval of music and peer approval of music on music selection* [Doctoral dissertation, Columbia University].
Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science*, *32*(2), 418–444.
Teo, T. (2005). Relationship of selected listener variables and musical preference of young students in singapore. *Music Education Research*, *7*(3), 349–362.
Teo, T., Hargreaves, D., & Lee, J. (2008). Musical preference, identification, and familiarity: A multicultural comparison of secondary students from singapore and the united kingdom. *Journal of Research in Music Education*, *56*, 18–32.
Thaut, M. H., Hodges, D. A., Morrison, S. J., Demorest, S. M., & Pearce, M. T. (2018). Cultural distance: A computational approach to exploring cultural influences on music cognition. In *The oxford handbook of music and the brain* (pp. 41–65). Oxford University Press.
Thomas, K. (2017). Sounds of disadvantage: Musical taste and the origins of ethnic difference. *Poetics*, *60*, 29–47.
Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, *1*, 166.
Tian, X., & Poeppel, D. (2012). Mental imagery of speech: Linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, *6*, 314.
Tian, X., & Poeppel, D. (2013). The effect of imagination on stimulation: The functional specificity of efference copies in speech processing [PMID: 23469885]. *Journal of Cognitive Neuroscience*, *25*(7), 1020–1036.
Tian, X., Zarate, J., & Poeppel, D. (2016). Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, *77*, 1–12.
Tibo, M., Geirnaert, S., & Bertrand, A. (2020). Eeg-based decoding and recognition of imagined music. *bioRxiv*, *2020.09.30.320176*.
Tillmann, B., Bharucha, J., & Bigand, E. (2000). Implicit learning of tonality: A self-organized approach. psychol rev 107, 885-913. *Psychological review*, *107*, 885–913.
Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, *107*(4), 885.
Tillmann, B., Bigand, E., Escoffier, N., & Lalitte, P. (2006). The influence of musical relatedness on timbre discrimination. *European Journal of Cognitive Psychology*, *18*(03), 343–358.
Tillmann, B., Janata, P., & Bharucha, J. J. (2003). Activation of the inferior frontal cortex in musical priming. *Cognitive Brain Research*, *16*(2), 145–161.
Tillmann, B., Peretz, I., Bigand, E., & Gosselin, N. (2007). Harmonic priming in an amusic patient: The power of implicit tasks. *Cognitive Neuropsychology*, *24*(6), 603–622.
Trainor, L. J., Tsang, C. D., & Cheung, V. H. (2002). Preference for sensory consonance in 2-and 4-month-old infants. *Music Perception*, *20*(2), 187–194.
Trehub, S. E., Becker, J., & Morley, I. (2015). Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664), 20140096.
Trehub, S. E., & Hannon, E. E. (2006). Infant music perception: Domain-general or domain-specific mechanisms? *Cognition*, *100*(1), 73–99.
Unyk, A. M., & Carlsen, J. C. (1987). The influence of expectancy on melodic perception. *Psychomusicology: A Journal of Research in Music Cognition*, *7*(1), 3.
Van Canneyt, J., Wouters, J., & Francart, T. (2020). Enhanced neural tracking of the fundamental frequency of the voice. *bioRxiv*.

van der Weij, B., Pearce, M. T., & Honing, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in psychology*, *8*, 824.

Van Eijck, K. (2001). Social differentiation in musical taste patterns. *Social forces*, *79*(3), 1163–1185.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., *et al.* (2018). Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need.

Ventura, M. I., Nagarajan, S. S., & Houde, J. F. (2009). Speech target modulates speaking induced suppression in auditory cortex. *BMC Neuroscience*, *10*(1), 58.

Verveer, E. M., Barry, H., & Bousfield, W. A. (1933). Change in affectivity with repetition. *American Journal of Psychology*, *45*, 130–134.

Vuust, P., Heggli, O., Friston, K., & Kringelbach, M. (2022a). Music in the brain. *Nature Reviews Neuroscience*, *23*.

Vuust, P., Heggli, O., Friston, K., & Kringelbach, M. (2022b). Reply to 'towards a cross-cultural framework for predictive coding of music'. *Nature Reviews Neuroscience*, *23*.

Vuust, P., Heggli, O. A., Friston, K. J., & Kringelbach, M. L. (2022). Music in the brain. *Nature Reviews Neuroscience*, *23*(5), 287–305.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*.

Wang, K., & Shamma, S. (1994). Self-normalization and noise-robustness in early auditory representations. *IEEE Transactions on Speech and Audio Processing*, *2*(3), 421–435.

White, C. G. (2001). *The effects of class, age, gender and race on musical preferences: An examination of the omnivore/univore framework* [Doctoral dissertation, Virginia Tech].

Whitford, T. J., Jack, B. N., Pearson, D., Griffiths, O., Luque, D., Harris, A. W., Spencer, K. M., & Le Pelley, M. E. (2017). Neurophysiological evidence of efference copies to inner speech. *eLife*, *6*, e28197.

Wirthlin, M., Chang, E. F., KnÃ¶rnschild, M., Krubitzer, L. A., Mello, C. V., Miller, C. T., Pfenning, A. R., Vernes, S. C., Tchernichovski, O., & Yartsev, M. M. (2019). A modular approach to vocal learning: Disentangling the diversity of a complex behavioral trait. *Neuron*, *104*(1), 87–99.

Witten, I. H., Manzara, L. C., & Conklin, D. (1994). Comparing human and computational models of music prediction. *Computer Music Journal*, *18*(1), 70–80.

Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3 suppl. 1*, 1212–1217.

Wong, P. C., Roy, A. K., & Margulis, E. H. (2009). Bimusicalism: The implicit dual enculturation of cognitive and affective systems. *Music Perception*, *27*(2), 81–88.

Yabe, H., Tervaniemi, M., Reinikainen, K., & Näätänen, R. (1997). Temporal window of integration revealed by mmn to sound omission. *Neuroreport*, *8*, 1971–1974.

Yee-King, M. J., Fedden, L., & d'Inverno, M. (2018). Automatic programming of vst sound synthesizers using deep networks and other techniques. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(2), 150–159.

Yoo, S.-S., Lee, C. U., & Choi, B. G. (2001). Human brain mapping of auditory imagery: Event-related functional mri study. *NeuroReport*, *12*(14).

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory–motor interactions in music perception and production. *Nature Reviews Neuroscience*, *8*(7), 547–558.

Zatorre, R. J., & Halpern, A. R. (2005). Mental concerts: Musical imagery and auditory cortex. *Neuron*, *47*(1), 9–12.

Zatorre, R. J., Halpern, A. R., Perry, D. W., Meyer, E., & Evans, A. C. (1996). Hearing in the mind's ear: A pet investigation of musical imagery and perception [PMID: 23972234]. *Journal of Cognitive Neuroscience*, *8*(1), 29–46.

Zatorre, R. J., & Salimpoor, V. N. (2013). From perception to pleasure: Music and its neural substrates. *Proceedings of the National Academy of Sciences*, *110*(Supplement 2), 10430–10437.

Zelano, C., Mohanty, A., & Gottfried, J. A. (2011). Olfactory predictive codes and stimulus templates in piriform cortex. *Neuron*, *72*(1), 178–187.

Zhang, Y., Chen, G., Wen, H., Lu, K.-H., & Liu, Z. (2017). Musical imagery involves wernicke's area in bilateral and anti-correlated network interactions in musicians [cited By 7]. *Scientific Reports*, *7*(1), 1–13.

Zwir, I., Arnedo, J., Del-Val, C., & et al. (2020). Uncovering the complex genetics of human character. *Molecular Psychiatry*, *25*(10), 2295–2312.

Zyphur, M. J., Zhang, Z., Barsky, A. P., & Li, W.-D. (2013). An ace in the hole: Twin family models for applied behavioral genetics research. *The Leadership Quarterly*, *24*(4), 572–594.

Appendix A: Stimuli for the Study on Sociology of Music

Artist	Song Name	Genre	Year of Production	Country of Production
Fuyumi Sakamoto	Muhou Ichidai Iri	Traditional	2014	Japan
Abdou Gambetta	Allah Ghaleb	Rai	2023	Algeria
Adela Jens	Nuits	chanson	2023	France
Adriana Franco	Luz da Nossa Afeição	Fado Antigo	2013	Portugal
Airelle Besson & Nelson Veras	Pouki Pouki	Jazz	2014	France
Al Ruzafa	Samai Hijaz Kurdi	Trad Arabo-Andalus	2007	Spain
Antent	Magnolia	hardwave	2022	Russia
Antonella Colaianni mezzo	(Pergolesi) Stabat Mater	baroque	1730	Italy
Barrut	L'èrba dagram	Trad Occitana	2018	France
Black Bear Creek	Round Dance	Powwow (trad indignous)	2008	Canada
Bloco Não Serve Mestre	Não Serve Mestre Eu Sou	bloco	2018	Brasil
Willie Bobo	Evil Ways Willie Bobo	afro-cuban percussion	2003	Cuba
CAROLE PELÉ	Courir	french trap	2022	France
Cao Jianguo	Imperial court music in Tang dynasty (Part I, Prelude of Danc	Traditional	2007	China
Carlina Lara	Be Pemontonay (She Looks Like a Tree with Flowers)	musica indigena latinoamericana	2001	Venezuela
Chico Mann	Dilo Como Yo (Te Están Llamando)	latin afrobeat	2012	USA
Chiddy Bang	Wonderful Feeling	indie pop rap	2023	USA
Clyde Davenport & W.L. Gregory	Cumberland Gap	appalachian folk	2021	USA
Constantin Moscovici	Sen Gelmez Oldun	moldovan population	2017	Azerbaijan
Dan Kye	Actually	indie jazz	2020	England
David Darling	Lugu Lugu Kanlbi	bow population	2021	USA
David Herrero	The Present, the Future (Extended Mix)	latin tech house	2020	Spain
Diamante Negro	Club Caribe	spanish noise population	2020	Spain
Post Nebbia	Oltre la soglia	contemporary jazz	2022	Italy
Dr. Dipali Bhatt, Foram Maheta	Yamunaji Ni Stuti	gujarati garba	2001	England
Prospero Cauciello	Duetto No. 2 for 2 Mandolins in G Minor I. Andante	baroque classical music	2014	italy
Floyd Lee	When You Break A Young Girls Heart	blues	2001	USA
Gacha Empega	Adieu paure carnavas	musica occitana	1998	France
George Braith	BOOP BOP BING BASH	soul jazz	2002	USA
Gesleir	Tout va bien	french indie pop	2020	France
Grentperez	Absence Of You	bedroom r&b	2022	Australia
Grupo Sotz'il	Chinimital	musica indigena latinoamericana	2017	Guatemala
Gyudmed Tantric Monastery	Taking Refuge in Three Gems; Invitation of the Green Tara; Pu	tibetan mantra	2002	Nepal
Hands of Elohim	Never Alone	christian rock	2017	USA
Isabella Zirilli	Buio freddo	pop	2022	Italy
Joe Armon-Jones	Midnite Oil (Sparkzzz)	indie jazz	2017	England
Joe Perrino	Giovanottello	musica sarda	2014	Italy
Kapela Harnasie	Idzie Dysc	goralski	2012	Poland
Kelyan Muller, DJ High	Con Cuidao	latin viral pop	2020	France
LEXO	Tagounite	french rap	2020	France
La Fabi	Viva la Tierra	nuevo flamenco	2018	Spain
Lavender Fields	Sleepy Tiger	shamanic	2020	USA
Leslie Medina	Particulière	french pop	2023	France
Banis	Tubaru Bai	population flamenco	2010	Spain
Masanka Sankayi + Kasai Allstars	Wa Muluendu	musique traditionnelle congolai	2005	Congo
Maulidi And Musical Party	Mume Ni Moshi Wa Koko	taarab	2006	Kenya
Mauresca	Per la montanha	musica occitana	2011	France
Nhii	Feather feat. Pippermint (Iorie & Madmotormiquel Remix)	ethnotronica	2020	USA
Nicolas Fraissinet	La fée clopette	french folk pop	2008	France
Niuri Te sule & Pinzani Pizzica la '	Ballataranta	tarantella	2012	Italy
Braja	Bungkam	indonesian experimental	2019	Indonesia
Oh, Weatherly	Here Tonight	pop punk	2018	USA
Olga Tira	Inima Mamei	moldovan pop	2018	Moldova
Oskar Easterfield	Funkadelic	nordic house	2018	Norvegia
Pibes Ran	Me escapé de la granja	cumbia chilena	2017	Chili
Pura Fé	Mohomoneh	indigenous folk	2015	USA
Rovere	astronauta	italian indie pop	2022	Italy
Seyoum Gèbrèyès	Muziqa muziqa	funk ethiopan soul	2004	Ethiopy
Shoming Bouboul Akwel	Tambola	comptine africaine	2020	France
Sirrus	Frill Ride	gamecore	2012	England
Suamenlejjona	Penkist sata	finnish metal	2019	Finland
Surf Curse	Freaks	surf punk	2021	USA
Sverdlovsk Philharmonic	(Sergei Prokofiev) Piano Concerto No. 2 in G Minor, Op. 16 II	classical orchestra	1990	England
The Nazgûl	Shelobs Lair	ambient soundscapes	1975	Spain
Tony Chasseur	Souri ba mwen	zouk	2015	France
Trio Maravilla	Plena De San Anton	bomba y plena	1954	Colombia
Tristan	Talking in Technicolour	psytrance	2014	England
Twelve Thousand Days	Thistles	deep neofolk	2012	England
VILDA	Vildaluodda	finnish folk	2019	Finland
Vetcho Lolas	Shamakuana	zouglou	2004	Germany
Jillian Dawn	Watered Down	folk rock	2022	USA
Xacobe Martînez Antelo Quintet	Paï	galician jazz	2006	Spain
Kim Sawol	Someone	korean indie	2018	Korea
arthrn	malaimé	indie electronic pop	2015	France
centomilacarie	Strappami la pelle a morsi	emo trap italiana	2022	Italy
В'ячеслав Кукоба	Гадом буду не забуду	ukrainian folk pop	2005	Ukrain
אנסמבל יעלה	"יונת חזוקים"	israeli folk	2020	Israel
اغنية تخاصمنى تصالحنى- فيلم البس عشا	حسن الرداد ايمي سمير غانم محمود الليثى بوسى	egyptian traditional	2016	Egypt
太合音樂 Taihe Music	窒息樂隊 Suffocated【黎明之下】HD 高清官方完整版 M\	metal	2019	Taiwan
山田ギャル神宮	MyStory	japan pop	2021	Japan