Article:
Lepa, Steffen, Herzog, Martin, Steffens, Jochen et al. (2020) A computational model for predicting perceived musical expression in branding scenarios. Journal of New Music Research. ISSN: 0929-8215

https://doi.org/10.1080/09298215.2020.1778041

Author's postprint:

Lepa, S., Herzog, M., Steffens, J., Schoenrock, A., Egermann, H. (in press). A Computational Model for Predicting Perceived Music Expression in Branding Scenarios. *Journal of New Music Research.*

A Computational Model for Predicting Perceived Musical Expression in Branding Scenarios

Steffen Lepa[a]*, Martin Herzog[a], Jochen Steffens[a], Andreas Schoenrock[a], and Hauke Egermann[b]

[a]*Audio Communication Group, Technische Universität Berlin, Germany;*
[b]*York Music Psychology Group, University of York, UK*

*Dr. Steffen Lepa, Audio Communication Group, Technische Universität Berlin
Einsteinufer 17c, 10587 Berlin, Germany
Phone: +49 (0)30 - 314 - 29313
Mail: steffen.lepa@tu-berlin.de *(corresponding author)*

Martin Herzog, Audio Communication Group, Technische Universität Berlin
Einsteinufer 17c, 10587 Berlin, Germany
Phone: +49 (0)30 - 314 - 29093
Mail: herzog@tu-berlin.de

Prof. Dr. Jochen Steffens, Institute of Sound and Vibration Engineering, University of Applied Sciences Düsseldorf
Münsterstraße 156, 40476 Düsseldorf
Phone: +49 (0)211 - 4351 - 9001
Mail: jochen.steffens@hs-duesseldorf.de

Andreas Schoenrock, Audio Communication Group, Technische Universität Berlin
Max-Beer-Str. 25, 10119 Berlin, Germany
Phone: +49 (0) 177 2555 262
Mail: info@studio-schoenrock.de

Dr. Hauke Egermann, York Music Psychology Group (Director), Music Science and Technology Research Cluster, Department of Music, University of York
Heslington, York, UK
YO10 5DD, UK
Phone: +44 - 1904 - 324303
Mail: hauke.egermann@york.ac.uk

A Computational Model for Predicting Perceived Musical Expression in Branding Scenarios

The article describes the development of a computational model predicting listener-perceived expressions of music in branding contexts. To address the 'semantic gap' between audio signals and complex brand identities, population-representative ground truth from multi-national online listening experiments was combined with machine learning of music branding expert knowledge, and audio signal analysis toolbox outputs. A mixture of random forest and traditional regression models is able to predict average ratings of perceived brand image on four dimensions of the employed GMBI music branding inventory. Resulting cross-validated prediction accuracy (R^2) was Arousal: 61%, Valence: 44%, Authenticity: 55%, and Timeliness: 74%. Audio descriptors for rhythm, instrumentation, and musical style contributed most to prediction. Adaptive sub-models for different marketing target groups further increase prediction accuracy.

Keywords: music information retrieval, listener modelling, recommendation systems, audio branding, machine learning

Introduction

Until now, commercial algorithmic music indexing and recommendation systems have predominantly focused on predicting consumers' musical *preferences* and choices when listening to music. In this way, they help listeners to navigate the boundless digital music archives currently available and let them discover new titles and artists for enhancing their musical enjoyment in everyday life. Another important segment of commercial music exploitation is the use of existing music tracks (typically: pop songs, dance tracks and 'hits' from the classical repertoire) as a means for brand communication. This practice is often called *music branding* (Müllensiefen & Baker, 2015) and forms an important part of *audio branding* (sometimes also called *sonic branding* or *sound branding*), referring to the strategically-planned employment of sounds and music in advertising, public relations, product design, and at the point of

sale (Jackson, 2003; Kilian, 2009; Gustafsson, 2015; Egermann, 2019). Music branding, as a new type of exploitation of musico-cultural assets, contributes a growing number of revenue shares in the overall music industry today. However, finding 'suitable' music for communicating specific brand aims is a challenging task, given the sheer amount of music available suitable for branding purposes and the lack of appropriate metadata. Hence, audio branding agencies, as well as music labels and specialised stock music providers, would benefit greatly from algorithmic software tools that help to identify music with the 'correct message' from their digital archives which often hold millions of titles. The current contribution documents major outcomes of a publicly funded European research project that forms the statistical basis of an algorithmic solution for automatic indexing of digital music files in terms of brand communication goals[1]. Hence, the resulting computational model is thought to feed a new type of business-to-business (B2B) music recommendation service for branding purposes.

Music as a communicative means in advertising and branding

Early empirical works on the beneficial effects of popular music in advertising have typically emphasised classical conditioning (Gorn, 1982), symbolic consumption (Larsen et al., 2010), as well as attention and memory effects (Allan, 2006) as the main acting principles of music employed in advertisements, arguably, with mixed results. For in these studies, music was typically treated as an abstract, symbolic stimulus that may increase the persuasiveness of an advertisement message; but, however, as a stimulus that had no proper "message" in itself. Most newer works on music in brand

[1] Preliminary short papers about specific aspects of this project were already presented at ISMIR 2016 (Herzog et al., 2016) and ESCOM 2017 (Herzog et al., 2017).

communication take a radically different approach, insofar as they literally conceive of music as a language. They predominantly focus on so-called *congruity effects* (or *musical fit*) in order to explain the (in-) effectiveness of music in branding, adverts and at the point of sale (see North et al., 2016 for an extensive overview). The notion of "fit" specifically refers to the listener-perceived *semantic congruence* between the communicative meaning of a certain piece of music and the communicator-intended *identity* of a certain brand, product, or service (MacInnis & Park, 1991, Zander, 2006). In the relevant marketing literature, the semantic content of a *brand identity* is typically conceived of as the combination of *brand personality* and *brand values* (Burmann, Jost-Benz, & Riley, 2009; Chernatony, 1999; Nandan, 2005). Both sub-dimensions of a brand's identity can be expressed linguistically by adjectives. While brand personality refers to a set of human personality traits associated with a brand, such as *responsible, active, emotional, aggressive,* or *simple* (Geuens, Weijters, & De Wulf, 2009), brand identity typically also encompasses abstract human values, such as *powerful, aesthetic, benevolent, ecological, healthy, hedonist, stimulating,* or *traditional* (Gaus et. al., 2010).

The empirical output of more than two decades of music congruity research in brand communication can be summarised as follows: If the music employed as part of a branding strategy is in itself able to communicate similar values and traits as the brand, product or service, it will lead to significantly increased brand awareness, improved persuasive effects of advertising measures, and finally, also to a more enjoyable customer experience at the point of sale (North et al., 2016). Based on a seminal study by MacInnis and Park (1991), this general finding has since then been robustly validated in numerous follow-up studies (e.g. Hung, 2000; North et al. 2004; Oakes & North, 2006). Therefore, the current study draws on the theoretical concept of *musical congruity* in order to understand music's communicative role in the branding process.

Musical communication according to music psychology

How can musical congruity be achieved in branding scenarios? To understand music's specific ability to communicate brand identities, it is useful to draw on the functionalist approach to communication (Brunswik, 1952), as this has already been found to be helpful for understanding musical communication within the field of music psychology (see Juslin, 2000, for a theoretical introduction and an empirical example). The main notion of this approach is that, similar to a non-verbal language, musical meaning is conveyed by evoking fuzzy (though collectively shared) emotional and semantic associations in the listeners based on a number of partially redundant acoustic *cues* contained in the sounding musical material. According to the theory, these cues form music's actual "vocabulary", which is acquired as informal knowledge during music socialisation; hence, music performers and listeners stemming from a similar musical culture are in result able to "understand" each other (Juslin, 2000). As prototypical musical cues, Juslin (2000) refers to music and sound parameters such as *tempo, loudness, spectrum, articulation, mode and measure*, as well as sounding features of the musical structure unfolding in time (*employed scales, functional harmonics*), and finally, to the *semantic content of the lyrics*. Extending from this original notion, it can be argued, that also other easily-recognisable sound features of popular music such as *genre, style*, and *production sound* should form additional acoustical cues that also convey an easily understandable "message", even for musical laymen (Tagg, 2013).

Accordingly, a marketing literature review by Oakes (2007) found that manipulation of *mood, genre, score, lyrics, tempo,* and *timbre* of the specific music employed in adverts and branding resulted in significantly different musical congruity effects.

Formalisation of music branding and the need for an algorithmic solution

Based on the depicted theories and empirical findings on musical congruity as the basis of successful music branding, it is possible to formalise *music branding* (Müllensiefen & Baker, 2015) as a profession: Specialised music consultants working for audio branding agencies have to translate the attributes of a given brand identity (brand personality and brand values, as specified by marketing strategists) into *fitting* musical cues that are able to express this identity, such as *melody, instrumentation, genre, rhythm, sound* (see to Figure 1). This fit is essential in order to evoke the desired semantic congruity between a brand identity and the selected music in listeners, resulting in the desired "brand image"). Thereto, branding consultants have to rely on their practical experience with musical meaning attribution from the perspective of different audiences towards different types of music in differing contexts. In other words, they apply their musico-cultural knowledge about the contextual meaning of musical cues. Then, in a second step, the consultants have to identify single music tracks or assemble playlists (from their own archive or from specialised stock music providers) conforming to the corresponding musical attributes. Finally, audio branding agencies also develop a specific strategy of how to practically employ the selected music in a specific branding campaign and sell this concept together with the rights to use the created/chosen music to their customers (see Bronner & Hirt, 2009 for an overview on the general challenges of audio branding practice).

One significant challenge for the work of music consultants when creating a music branding strategy is the sheer breadth of online music archives combined with a lack of brand-relevant metadata describing the tracks in these archives in a proper way for the music branding task. Even experienced senior music consultants are typically not able to oversee the attributes of music existing in their own archive, much less those

from music available in other archives, not to mention the attributes of the breadth of new music released every day.

Anticipating this problem, music consulting agencies, record labels and providers of stock music archives for advertisements have begun to tag the contents of their music archives in terms of *genre, style, mood, tempo*, and *instrumentation*. However, the taxonomy behind these tags, as well as the tagging itself, is often inconsistent. Moreover, the available metadata are rarely extensive enough to provide satisfactory results for search requests originating from the complex structure of a given brand identity. This challenge, which forms a practical obstacle for small European audio branding agencies to take part in a global music exploitation market, gave rise to a publicly funded, comprehensive research and development project from which we present selected findings in the current paper.

The objective of our research was to develop an algorithmic solution for predicting perceived musical expression in branding scenarios (as depicted in the theoretical framework above) based on social research methods, knowledge from music psychology and employing existing music information retrieval (MIR) techniques. Importantly, we did not aim at substituting the music consultants' work with an algorithm, but rather to provide them with a practical tool that helps with preselecting a range of suitable "fitting" music for their everyday work and to thereby empower them to focus on final decisions that truly demand their (human) expertise.

- place Figure 1 here -

The General Music Branding Inventory (GMBI)

For algorithmic modelling of the translation process described in the theoretical part above and depicted in Figure 1, it is first necessary to identify the relevant semantic

elements of any given brand identity that can successfully be encoded into musical cues and also be successfully decoded by typical music listeners into a brand image. A further challenge for this task is that consumers from different social milieus and cultures and with different musical expertise might draw on different linguistic terms to describe perceived musical expression. Moreover, the descriptive terminology of consumers may only partly overlap with that of branding experts. Addressing these challenges systematically, a four-dimensional model of *musical expression in branding contexts* measurable by the *General Music Branding Inventory* (GMBI) was presented by Herzog and colleagues (Herzog, Lepa, Steffens, et al., 2018; Steffens et al., 2018). The underlying questionnaire GMBI_22 consists of 22 adjectives and was developed empirically following results from an audio branding expert focus group and a marketing expert survey (Herzog, Lepa, Schönrock, et al., 2017; Herzog et al. 2020). Within two pilot studies, a word list representing the central elements of a brand identity that can also be expressed through music, was generated (the Music Branding Expert Terminology – MBET). In the next step, the MBET list underwent comprehensive listening tests with a large number of consumers from different countries who were presented with a large range of music titles. Resulting ratings were analysed with Exploratory Factor Analysis (Fabrigar et al., 1999) using orthogonal Crawford-Ferguson Rotation (Browne, 2001) and the obtained factor solution was optimised by stepwise item deletion based on modification indices and with the aim to achieve language invariance (Steenkamp & Baumgartner, 1998). The result was a condensed list of 22 questionnaire items (GMBI_22) which operationalises a four-dimensional parametric *musical expression space.* These dimensions are able to capture the most important aspects of contemporary typical brand identities that can be communicated with popular music (see Table 1): We find two *emotion expression* dimensions (*Arousal*

/ *Valence*), as well as two *brand value* dimensions (*Authenticity / Timeliness*) which together represent the essence of music branding communication.

- place Table 1 here -

Study Aim: Predicting perceived brand-relevant musical expression using MIR features

Since dimensions of perceived musical expression for branding contexts have been successfully formalised by the GMBI_22 instrument, the aim of the current study was to develop a computational model that is able to predict the GMBI scores of any given piece of music based on *music information retrieval* (MIR) techniques. In terms of machine learning, this study aimed to solve a *regression problem*, as this approach has been shown to empirically perform better than a discrete classification approach when it comes to predicting higher-order human responses to music such as emotions (Yang et al., 2017).

However, *perceived expression of music in branding contexts* is not to be conceived of as an inherent quality of audio files, but may partly lie 'in the ear of the beholder'. Prior studies dealing with the semantic expression of music (Bonneville-Roussy, Rentfrow, Xu, & Potter, 2013; Shevy, 2008) demonstrate that members of different social milieus, countries, gender and generations tend to attribute slightly different semantic expressions to the very same musical pieces. Further, there is virtually no valid 'Big Data' information basis in terms of existing MIR datasets or exploitable user transactions on existing online music platforms, that would deliver ground truth data on perceived semantic expression of a large number of music tracks. Hence, a *knowledge-based recommendation approach* (Burke, 2000) drawing on

manually acquired ground truth by means of a large-scale online survey appeared as the single realistic option for developing a valid prediction model for perceived musical expression.

Research questions and summary of research design

In summary, the aim of the current contribution was to develop a computational model, which can predict perceived musical expression for the branding context, taking into account social differences with regard to the perception of musical meaning. By drawing on an experimental online survey approach, we aimed on answering the following three research questions:

(1) To which degree can we predict perceived musical expression (as measured by the GMBI_22) in the context of music branding?
(2) What is the explanatory power of different kinds of audio descriptors regarding perceived musical expression?
(3) To what extent can we increase the prediction accuracy for perceived musical expression, when modelling inter-individual differences, represented by typical marketing target groups?

To approach these research questions, we created ground-truth data by conducting two multi-national online listening experiments. In the course of the experiments, 10,144 European listeners rated the perceived fit between GMBI items and 549 presented musical excerpts. Furthermore, we extracted 487 different audio features from the same excerpts, drawing on up-to-date audio signal analysis and music information retrieval techniques, including a number of high-level music descriptors that had been developed before by employing supervised machine learning of music branding expert knowledge (describing e.g. *genre* or *instrumentation* of a track).

Combining acquired ground truth with the MIR features, we then applied two different machine learning methods (*hierarchical stepwise regression* and *random forest regression*) in order to test which model family would perform best in predicting the GMBI scores and ultimately identified the best method for each GMBI factor dimension. Then, we analysed resulting models with regard to the explanatory contribution of different audio descriptor blocks. Finally, to approach the problem of modelling inter-individual differences in perceived musical expression, we developed adaptive model variants for typical marketing target groups, aiming to further increase the predictive accuracy of the overall computational model.

Materials and methods

In this section, we describe the development of a computational model for predicting perceived musical expression in branding scenarios. We initially describe the composition of the music stimulus set chosen for the listening experiment and the prediction model development. Following this, we present the methodology of the online listening experiments leading to a large data set of listener ratings and resulting in factor scores based on the GMBI_22 instrument. Afterwards, we present the development and extraction of audio descriptors for the computational prediction models, which include machine learning of branding expert music knowledge on one hand and the application of available MIR toolboxes on the other. Finally, we describe the statistical development of the final computational prediction models. Figure 2 provides a graphical overview of the methodological steps described in this chapter.

- place figure 2 here -

Music stimulus set for listening experiments and prediction models

All music recordings used in the presented study stem from the library of the collaborating audio branding agency *HearDis* containing approximately 100,000 music pieces. Many of the tracks are well-known popular music titles from the past decades, extended by various dance music tracks and some "hits" from the classical repertoire. The library was organised in ten different musical *genres* (*Blues, Classical, Dance, Folk, Hip Hop, Jazz, Pop, Rock, Soul/Funk,* and *World Music*) and 61 musical *styles* (sub-genres, e.g. *Fusion Jazz*, see Table 2 for a complete list). In addition to *genre* and *style* adherence, branding experts of the agency also tagged the pieces with additional information on dominant *instrumentation* (13 classes, plus 5 classes representing the *existence* and *gender* of *vocal* parts), as well as in terms of dominant *production timbre*, with the 6 mutually exclusive tags *hard, soft, warm, cold, bright,* and *dark*.

For the online listening experiments, a sub-sample of 549 tracks was manually selected by the experts, with nine representative tracks for each musical *style*. The choice of tracks reflects extensive discussions and agreement among six of the cooperating agency's professional music consultants. For each style, it comprises of the nine pieces that were deemed to best represent the complete musical spectrum of the respective style. After an agreement had been reached for all styles, in a second step, an independent group of three further audio branding experts verified the plausibility of each track per style, leading to further optimisation of the final selection.

Subsequently, excerpts of approximately 30 seconds were taken from each digital audio file, comprising the first transition from verse to chorus of the tracks. The aim of this step was to provide suitable stimuli for the planned online listening experiments, ensuring that participants would be able to rate multiple tracks in a reasonable amount of time, thereby employing a well-established economical practice in

music psychology research. Afterwards, the resulting files underwent a perceptual loudness adjustment: Based on a reference track representing the mean loudness of the complete music stimulus set, the level of all other tracks was corrected individually by a mastering engineer, since a mere automatic loudness adjustment is typically not enough to accommodate for differing production schemes concerning loudness and dynamics which are found with music from different decades and styles. Finally, each track excerpt received a smooth fade-in and fade-out and was then MP3-encoded (Stereo, 320 Kbit/s) for the online listening experiments.

Online listening experiments

In order to generate ground truth on the branding-relevant perceived musical expression of the 549 chosen music track excerpts, we conducted comprehensive online listening experiments within three European countries. Specifically, two consecutive experimental survey waves were realised with the support of commercial online-access panel providers which systematically recruited participants according to requested quotas and provided participants with a monetary compensation. During the first wave, 183 music excerpts (three from each style) were presented to 3,485 listeners from the UK, Spain and Germany, with the sample containing an equal distribution of members from each country, gender, educational background (ISCED 0-2, 3-4, 5-8) and age group (18-34, 35-51, 52-68).

Each participant in the first wave rated four randomly chosen music excerpts by means of the GMBI_22 questionnaire. During the second wave of listening experiments, 366 music excerpts (six from each style) were presented to 6,659 listeners from the UK, Spain and Germany, with the sample again containing an equal number of residents of each country, but population-representative relative shares for each gender, educational background and age group (with fully “crossed” quotas, meaning that e.g.

also the quota of gender within each age group in each education group in each country was representative to corresponding population shares). Each participant in the second wave rated six randomly chosen excerpts by means of the GMBI_22 questionnaire.

The listening experiments' procedure always started with the collection of participants' socio-demographic information and a short sound test for calibrating audio playback volume. In the second wave, a short initial questionnaire with 38 Likert items measuring SINUS meta milieu membership was additionally administered, in order to represent typical marketing target groups beyond socio-demographics (SINUS meta milieus are a well-established multi-lingual commercial operationalisation of lifestyle-groups in international marketing; see Homma & Ulktzhöffer, 1990 for a theoretical introduction; see SINUS, 2017 for an overview of the current version of the instrument which clusters consumers into nine groups called "meta milieus", the labels of the resulting nine meta milieus are provided in the bottom nine rows of Table 5).

After the initial questions, the first 30s music excerpt was played, followed by the instruction to rate the subjectively perceived degree of fit between the music and the 22 adjective items of the GMBI_22 questionnaire, which were presented in a random order, using a 6-point scale for the ratings. In the UK, GMBI_22 items were presented in English, in Germany in German, and in Spain in Spanish.

After the first trial, the subsequent track excerpts were presented in exactly the same way (3 further excerpts per person in wave 1, and 5 further excerpts per person in wave 2). Random music excerpt selection was programmed for both waves in a way to enforce equal playback probability for each track within the 54 groups formed by combinations of all socio-demographic variables. In total, each online experiment took about 15-30 minutes and ended with a short questionnaire asking for participants' musical preferences and the audio playback set up they used.

Development and extraction of audio descriptors

The audio descriptors used as predictors in the computational prediction model developed in this paper are derived from two different sources: Machine learning of branding expert music knowledge and existing MIR toolboxes.

Machine learning of branding expert music knowledge

In order to include branding experts' music knowledge in our planned prediction model, we first applied supervised learning of all the tags contained in the collaborating audio branding company's music archive (*genre, style, instrumentation, vocals_existing, vocals_gender, production timbre*, see Table 2). For this purpose, 17,163 representative full music tracks were chosen from the library, to represent at least 100 tracks of each *style* and all possible combinations of descriptor tags. Note that none of the tracks utilised in the subsequent listening experiments in wave 1 or 2 were part of this procedural step. For each of the six families of expert tags, a machine-learning (ML) model was trained. We employed the *IRCAM_classification* meta-framework (Burred & Peeters, 2009; Peeters et al., 2015), which allowed us to train a ML classifier given a set of exemplary music tracks belonging to a given tag.

Using the provided training sets, and after exclusion of the few tracks with ambivalent tagging, we found that all classifications could be realised as single-label classifications. This means that a given music track belongs only one tag within a given tag family as opposed to multi-label classifications where a given track can belong to several tags within a given tag family simultaneously. The following feature-based ML approach is implemented in IRCAM_classification:

(1) extracting a large set of audio features using the *ircamdescriptor* software library (Peeters, 2004)

(2) modelling their behaviour over time using AR-vector models, Modulation Spectrum and/or Universal Background-Model / GMM-Supervectors
(3) performing feature space projection using Principal Component Analysis (PCA) and/or Linear Discriminant Analysis (LDA)
(4) performing supervised training of the final classifier using Support Vector Machine (SVM).

It should be noted that *IRCAM_classification* is a meta-framework which automatically finds the best combinations of parameters for a given task (discriminating between tags with a given tag family). For this reason and due to matters of space, we do not provide the specific values of e.g. identified SVM kernel parameters for each classifier in the results section; however, we do document the final classification accuracy benchmarks (see Table 2). The six ML classification models resulting from this procedure (*vocals_gender, vocals_existing, instrumentation, genre, style, production timbre*) were finally applied to the 549 music tracks selected for the online listening experiments. Each track was then characterised by its membership probabilities concerning each tag, of each tag family. This led to a set of 95 machine learning-based descriptors (the sum of all tag classes, see Table 2), representing the individual tag probabilities to be used later as input for the prediction model of perceived musical expression.

Extraction of further audio descriptors using existing MIR toolboxes

To gather further meaningful audio and music descriptors for the prediction model, an extensive signal analysis of the 549 music tracks was conducted, mainly drawing on existing MIR software toolboxes. The resulting set of content descriptors relates either to musical characteristics (such as the tempo or key) or to global sound characteristics

(such as the frequency bandwidth of the audio signal). To create a comprehensive musical description, we employed *IRCAM_beat* (Peeters, 2006b, 2011; Peeters & Papadopoulos, 2011) in order to represent rhythm and tempo (9 descriptors), *IRCAM_keymode* (Peeters, 2006a) to represent mode and key (12 descriptors), as well as *IRCAM_chord* (Papadopoulos & Peeters, 2011) in an adapted version (Steffens et al., 2017) that is able to represent typical chord successions and functional harmonics (13 descriptors).

Moreover, we utilised *IRCAM_descriptor* (Peeters, 2004) to represent the overall sound of the music track in terms of e.g. sinusoidal components, roughness or mean energy in specific frequency bands (42 descriptors). Finally, the *IRCAM_timbre_toolbox* (Peeters, Giordano, Susini, Misdariis, & McAdams, 2011) was employed to gather 316 further descriptors suitable to represent production specific audio features, e.g. frequency band limitations typical for certain decades of pop music.

Since the computational model to be developed was thought to later feed a fully-automatic recommender system that does not need any user intervention, we analysed full audio tracks and drew on the toolboxes' default options only. In summary, we gathered 392 audio and music descriptors (e.g. timbre, mode, tempo) and 95 machine learning-based descriptors (e.g. genre, style and instrumentation, see previous paragraph) for the prediction model.

Computational prediction model development

In the previous sections we have described the construction and characteristics of the variables used for the computational prediction model (see Figure 3). In the following, we document the statistical procedures taken to answer the three research questions.

- Place Figure 3 here -

Data pre-processing

After initial data cleaning, factor scores of the four GMBI dimensions *Arousal, Valence, Authenticity* and *Timeliness* were calculated based on the ratings obtained in the listening experiments of both waves. This was done by employing robust maximum likelihood estimation (MLR) of factor scores (Fabrigar et al., 1999) using the statistical software package MPlus 6 (Muthén & Muthén, 2010) and specifying the GMBI_22 factor measurement model (see Table 1) which draws on the ESEM-approach with target rotation (Asparouhov & Muthén, 2009). Initially, we performed a language invariance test (Steenkamp & Baumgartner, 1998), resulting in *scalar invariance*, then we accordingly fitted the final multiple-group ESEM factor model (three groups representing the three language versions of the questionnaire) constraining factor loadings and item intercepts to be equal across groups and factor inter-correlations to zero, resulting in a good measurement model fit of $X^2=21092.550$; $df=627$; $p<0.01$; $RMSEA=0.043$; $CFI=0.959$; $SRMR=0.030$ based on n=53344 observations. During estimation of factor model and scores, the clustered structure of data (repeated measurements within individuals, two different waves with different cluster sizes) was addressed by using a robust sandwich estimator procedure implemented for such scenarios in MPlus.

Subsequently, we determined arithmetic means of resulting factor scores for each of the 549 track excerpts across the whole sample of participants (based on about 80-110 ratings per track). These 'track-based' factor scores were then merged with scores of the 487 audio and music descriptors (resulting from ML=machine learning of

music branding expert knowledge, as well as IRCAM=existing IRCAM toolboxes), constituting a reduced dataset, henceforth denoted as *population sample*.

In the same way, we calculated mean GMBI factor scores for 29 relevant marketing target groups formed by two-way-interactions of socio-demographic variables, as well as for the nine SINUS meta milieus (see column 1 in Table 5 for their labels). Resulting mean GMBI factor scores drew on approximately 10 to 60 ratings per track and were again merged with the scores of the 487 audio descriptors. The resulting 38 datasets are henceforth denoted as *target group samples*.

As the GMBI_22 factors were orthogonal by design and thus uncorrelated, four separate regression problems had to be solved for the whole *population sample*, as well as for the 38 separate *target group samples*. For each of the required partial models, all 487 (mostly metric) predictor variables in terms of audio descriptors were potentially useful. To address this combined feature selection and prediction problem, a stratified 9-fold cross-validation procedure was performed with the population sample in order to develop the *population models*: Therefore, we split the dataset by assigning the first eight tracks of each style (488 observations) to a *training dataset* and the one remaining track per style (61 observations) to a *test dataset*. In the next fold, we repeated this procedure, now leaving out the second track of each style for the test dataset, etc.

Finally, we applied z-standardisation on the numeric variables of the training datasets first and afterwards on test datasets, both based on determined training dataset scales (mean and variance estimations of predictor variables). This resulted in nine different, style-representative training datasets and nine disjunct holdout datasets for testing, which we then used for a later 9-fold cross-validation with resulting model R^2s being the average across all nine folds.

For the development and testing of the 38 *target-group-specific sub-models*, we drew on wave 1 data (183 observations) as *holdout* and wave 2 data (366 observations) as *training sample*. Note that a folding procedure was not deemed feasible in this procedural step due to low sample size. Similar to the population sample, we first applied z-standardisation on training datasets and then on the respective holdout datasets based on previously determined training dataset scales (mean and variance estimations of predictor variables). Figure 4 depicts the resulting training and test datasets used for the development and selection of prediction models for perceived musical expression and for target-group-specific sub-models as described in the following sections.

- Place Figure 4 here –

Training and selection of final regression models

In order to address research question 1, we trained prediction models for the four factors (*Arousal, Valence, Authenticity* and *Timeliness*) based on the nine different training datasets derived from the *population sample* as depicted in Figure 4. We tested two different model types, *hierarchical stepwise regression* and *random forest regression*. The rationale for this was to compare a rather traditional social science modelling approach that is based on linearity (stepwise regression) with a modern machine learning approach (random forest regression) that can handle non-linearity and complex interactions (Strobl, Malley, & Tutz, 2009). Random forests were estimated using the *cforest* function of the 'party'-package for the statistical software environment R (Hothorn, Hornik, Strobl, & Zeileis, 2019) while hierarchical stepwise regression was performed using the statistical software package IBM SPSS 25, drawing on the

regression function. For each model family, we first tuned hyper-parameters (see results section below for details) with the whole sample from both waves in a grid-like fashion, drawing on the *Arousal* scores and taking R^2 as an optimisation criterion. Hierarchical stepwise regression models were realised by entering predictor variables in a block-wise fashion (the blocks where either composed by toolbox origin or machine learning descriptor group, see Table 4 column 1 for a list of all predictor blocks). Then, we performed a stepwise variable selection procedure (forward/backward-method) within each block. During the course of initial hyper-parameter tuning, we also compared every possible order of functional variable blocks, since, due to the hierarchical nature of linear regression analysis, this could affect estimation results. As a selection criterion for the final model family to choose for each of the four dependent variables, we compared averaged R^2 across all nine CV-folds resulting from using either *hierarchical stepwise regression* or *random forest regression*. R^2 was always calculated by dividing the explained sums of squares by the total sums of squares throughout the whole study. After completing model family selection, we trained the chosen model variant again, now drawing on the whole *population sample* dataset encompassing both waves, in order to increase the informational basis for the final models.

Estimation of audio descriptor block importance for the final models

In order to address research question 2, we calculated incremental R^2 for each predictor block of the hierarchical stepwise regression solution. We drew on the respective model family previously selected for each of the four musical expression factors (see Table 4). In order to achieve maximum comparability, in spite of overlapping explanatory potential of predictors, we used the same order of blocks for the random forests as we had established for the hierarchical stepwise regression. We did this to get an estimate for the importance of different types of audio descriptors in the final prediction models,

drawing on the full population sample.

Training and selection of target-group-specific sub-models

Finally, to address research question 3, we calculated separate models for the 38 *target group samples,* now drawing on the individual training datasets derived from the 38 target group samples (see Figure 4). For estimating the target group sub-models, we always employed the same model type and hyper-parameters that had been found to be best for the *population sample* (*hierarchical stepwise regression* for *Arousal* and *Timeliness, random forest regression* for *Valence* and *Authenticity*). When resulting predictive R^2 values for the *holdout sample* fell below the R^2 acquired with the population model, we discarded the target-group-specific model. However, in cases where the fit was better than the R^2 reachable with the population model, an adaptive model for the respective target group was trained, now drawing on the full target group sample (training and test data).

Results

Results of machine learning of branding expert knowledge

In the following, we document the final results of the machine learning of branding expert knowledge which was realised with the *IRCAM_classifcation* software framework.

- place Table 2 here -

The machine learning of the various classifiers led to very robust results (see Table 2). Classification of three out of six tag categories (*genre*, *style*, and

vocals_existing) was accomplished with over 90% accuracy. Recognition of *instrumentation* (81% accuracy) and *production_timbre* turned out to be more difficult (82% accuracy). Finally, retrieval of the three tag categories of *vocals_gender* was most difficult (76% accuracy).

Hence, machine learning of branding expert resulted in reliable automatic higher-order music classifiers that we could employ to produce high-level music descriptors for the sample of 549 tracks used in the listening experiment. In this way, we obtained probability values from each of the classifiers for each track which complemented the other lower-order descriptors stemming from existing audio analysis toolboxes.

Multivariate computational prediction model for musical expression

In the following, we document the four final partial prediction models for musical expression (population models) we gained by applying the machine learning procedures described in the method section to the listening test dataset. Results firstly apply to hyper-parameters identified for the two different ML approaches under investigation. For *hierarchical stepwise regression*, we determined a forward-backward approach with *p_in=.05* and *p_out=.10* to be the solution leading to highest R^2 by employing grid-based hyper-parameter optimisation. Then as the last "hyper-parameter" (in a more qualitative sense), we also compared every possible order of functional variable blocks (which were roughly composed by toolbox origin / tag type). The best order in terms of highest resulting R^2 turned out to be (from first to last): *IRCAM_beat, IRCAM_keymode, IRCAM_chord, ML_Instrumentation, ML_Musical_style, ML_Musical_genre, IRCAM_descriptor, ML_Production_timbre, ML_Branding_suitability, IRCAM_timbre_toolbox.*

- place Table 3 here -

Training R^2 obtained for these model variants for each of the four dependent variables and R^2 for the 9-fold-cross validation procedure are documented in Table 3. Results indicate that *random forest regression (RFR)* provided the best solution for *Valence* and *Authenticity*, with *hierarchical stepwise regression (HSR)* resulting in lower R^2, both for training and CV results. For *Arousal* and *Timeliness*, however, *hierarchical stepwise regression* clearly performed better. Resulting R^2 from training the four selected population model variants with the full sample are *Arousal (HSR)* = 84%, *Valence (RFR)* = 77%, *Authenticity (RFR)* = 83%, *Timeliness (HSR)* = 85%. Note that these values are probably overestimating true future performance; however, they reveal the explanatory potential of the model, which is further expanded in the following paragraph.

Explanatory power of specific audio descriptor blocks

The estimated incremental R^2s for the single predictor blocks of the two hierarchical stepwise regression models and the two random forest models (see Table 4) provide a clear picture concerning predictor importance in the finalised full prediction model for musical expression in branding contexts: For the dimensions *Arousal* and *Timeliness*, the audio descriptors for *instrumentation* and musical *style* resulting from machine learning of expert tags as well as the audio descriptors from *IRCAM_beat* describing *rhythm* and *tempo* of music represent the most important types of predictors. *Harmonic* descriptors of music tracks stemming from *IRCAM_keymode* and *IRCAM_chord* as well as *production sound* descriptors (stemming from *IRCAM_descriptor* and *IRCAM_timbre_toolbox*) play a minor role.

- place Table 4 here -

In contrast, for the two dimensions *Valence* and *Authenticity*, *rhythmic* aspects of music as measured by *IRCAM_beat* play a considerably more important role, followed again by machine-learned *instrumentation* descriptors. Musical *style,* however, only plays a minor role for predicting perceived musical expression in these two dimensions. Here, *IRCAM_timbre_toolbox* and *IRCAM_descriptor* together were able to predict only a lesser amount of variance, followed by *harmonics* as grasped by *IRCAM_chord*, which all appear to be of minor importance. Moreover, descriptors of *IRCAM_keymode* led to a decrease of R^2 in the selected random forest models selected for *Valence* and *Authenticity.*

Finally, a general result across all GMBI dimensions is that *production timbre* descriptors, *IRCAM_keymode*, as well as *musical genres* do not substantially contribute to the prediction of perceived musical expression in branding context (the more fine-grained musical *styles* partly do), while *rhythmic aspects* of music*,* as well as the more fine-grained expert knowledge about dominant *instrumentation* gathered through supervised machine learning, appears to be essential.

Testing for improved prediction accuracy of target-group-specific sub-models

For the target-group-specific prediction models, we employed the same model types as were selected as *population models*, but trained them with the training datasets of the respective target group sample. As R^2 results of this procedure demonstrate (see Table 5), target-group-specific adaptive models turned out to be advantageous in the majority of cases (87 out of 152), For the remaining 65 cases, the *population model* proved equal or better in predicting GMBI factor scores for target groups. For some combinations of

target groups and GMBI factors, the *target group models* were especially beneficial to predict perceived musical meaning. This applies to predictions for the target group 'Sensation-oriented' (for all factors, except *Authenticity*) as well as to the SINUS meta milieu 'Adaptive navigators', where a substantial increase due to target group-specific prediction modelling was achieved for *Valence (+0.15)* and *Authenticity (+0.17)*.

Furthermore, a notable increase in accuracy could be observed for *Valence* predictions for the three target groups 'UK' (+0.11), 'Spain' (+0.09) and 'Germany' (+0.18), whereas almost no positive or negative difference was measured across all four GMBI factors for the three different age cohorts 'age 18-34', 'age 35-51' and 'age 52-68'.

The only two target groups exhibiting an increase across all four musical meaning factors were 'Spain' and 'Spain, female'. Note that the mean prediction accuracy of the four *population models* for the complete 'population' (see Table 5, row 1), which was taken as a baseline here, is lower than our main prediction results (see Table 3), in which 488 tracks were used for model training, compared to only 366 tracks in the analysis documented here. Note further that the strongest target group heterogeneities were in general observed for *Valence*, hence this factor also benefited most from the adaptive approach.

- place Table 5 here -

Discussion

In the present paper, we have documented the development of a ground truth-based, computational prediction model for perceived musical expression in the branding context, which will be turned into a publicly available fully-automatic B2B music

recommendation system addressing the needs of audio branding agencies and online music libraries in the near future. Given the statistical results obtained, the model is able to predict branding-relevant musical expression of popular music tracks as measured by the GMBI (in three different languages) with a high accuracy ranging between 44-74%.

Specifically, our final models will be able to predict the *Arousal* and *Timeliness* dimensions of musical expression as measured by the GMBI_22 with an accuracy somewhere in between 61-74%, while the *Valence* and *Authenticity* dimensions may be predicted with an accuracy of somewhere in between 44-55%. Interestingly, *random forest regression* models displayed their well-known advantages in grasping non-linearities and complex interactions only for the *Valence* and *Authenticity* dimensions of branding-relevant musical expression. This might be explained by empirical findings from music psychology, that musical communication cues often seem to work in a linear-additive fashion (Eerola, Friberg, & Bresin, 2013).

While there is a lack of prior machine learning studies concerning the two brand value dimensions (*Authenticity* and *Timeliness*) that could be compared to our findings, our results for the two emotional expression dimensions (*Arousal* and *Valence*) perform quite well compared to prior studies in this area. Equivalent studies also drew on a regression approach, musically diverse stimuli and the Arousal/Valence model to predict perceived musical emotion (Leman, Vermeulen, De Voogdt, Moelants, & Lesaffre, 2005; Yang et al., 2008; Tuomas Eerola, Lartillot, & Toiviainen, 2009; Han, Ho, Dannenberg, & Hwang, 2009; Schmidt, Turnbull, & Kim, 2010; Gingras, Marin, & Fitch, 2014; Saari et al., 2016). To the best of our knowledge, better results were only achieved by Eerola et al. (2009), in terms of Arousal R^2=77%, Valence R^2=70% with 5-fold CV, who only drew on a comparably narrower repertoire of 360 film music excerpts. Any other results of the above-quoted studies clearly fall below what we

present in this study in terms of predictive power and/or size of the music sample. Additionally, it is important to note that, since our prediction models are based on regression logics, even slightly biased predictions may still be expected to form sufficiently reasonable estimations of a music track's correct place in the four-dimensional musical expression space mapped by GMBI_22. Given that our developed prediction models will be implemented in a fully automatic recommender system that will not incorporate any form of preceding user or expert tags, our results appear to be very satisfying.

Furthermore, analysis of the most important explanatory predictor blocks that we calculated (Table 4) demonstrate that machine-learned branding expert music knowledge and audio descriptors from existing signal analysis tool boxes both contribute approximately equal weight to the models' prediction accuracy, partly differing in size between the GMBI dimensions. It thus appears to be the branding experts' implicit knowledge about musical styles and instrumentation together with easily derivable rhythm and tempo of music tracks that are decisive for a good prediction of perceived musical expression in branding contexts.

In addition, our study approached the challenge of target-group-specific musical meaning attribution. For the 29 different socio-demographic target groups and nine SINUS consumer milieus in three European countries, the prediction accuracy could be increased in 87 out of 152 cases by drawing on adaptive sub-models. These gains will also improve the prediction results of a fully automatic recommender system in development. While the overall gain is arguably not large across all tested sub-models in terms of the answer to research question 3, we still think that in applied scenarios where it is important to address specific target groups, the additional benefits of up to 23% in R^2 for some of the groups will be considered substantial.

Our analysis of target-group specificity further brought about interesting heterogeneities in terms of consumers' attribution of musical expression. Principally, we found differences in the *Valence* model's prediction performance across the three countries UK, Germany, and Spain. This finding is in line with prior research on culture-dependent influences on music listening behaviour (Pichl, Zangerle, Specht, & Schedl, 2017) and valence responses to music (Egermann, Fernando, Chuen, & McAdams, 2015). Therefore, a general recommendation to future developers of recommendation systems is to always include culturally adaptive modelling when addressing perceived emotions or semantics of music.

Limitations

One limitation of the approach to music branding recommendation presented in the present study is typical for ground-truth based prediction models employed in knowledge-based recommender systems (Burke, 2000): It is an open empirical question, to what degree the perceived branding-relevant musical expression of popular music pieces may underlie changes over time and across new listener generations. Hence, to keep up flexibility towards possible future changes in perceived music semantics, the latent music listener knowledge contained in the prediction models demands systematic updates through future online listening experiments providing the necessary ground truth in the forthcoming years.

A second limitation is the still pending real-world evaluation of the algorithmic solution for music branding that was developed by the study depicted in this paper. While results of the 9-fold cross-validation already demonstrate an expectable performance accuracy of our final models with "unseen" music titles, it will nevertheless be necessary to validate its actual performance in real music branding campaigns.

Conclusions and outlook

As a next developmental step, we will implement the final prediction models depicted in this study into a licensable software library providing users of digital music archives with the functionality to index any given piece of music with a valid score for each of the four GMBI dimensions and the underlying GMBI items. In this way, the branding-relevant expressive content of a music track can be estimated automatically. As a result, users of this system will be able to easily search their music archive for music tracks that fit best to a given brand identity.

A first operational scenario and validation test bed for the GMBI prediction model will be a commercial software tool to be presented soon by our project partner HearDis. It automatically generates playlists based on any given brand identity and target group profile. These playlists are then used to feed an in-store music player application, which can be used by retail stores interested in music branding at the point of sale. To this end, an additional *brand filter* software module was developed that uses the GMBI factor loading matrix and the estimated GMBI item reliabilities as input. Salespersons will have the ability to define a set of GMBI factor and/or item values that represent their intended brand characteristics best. The system will thus be able to generate music playlists with complex search constraints adapted to very specific marketing tasks and target groups. For any track in a music library that has been previously indexed by the *prediction module* described in this paper, it is then possible to calculate the Euclidian distance to the given search vector. After entering possible additional constraints in terms of genre, tempo and audio quality; for example, a playlist containing a requested number of least distant tracks adhering to given search constraints, will be returned.

Additional software modules allow for seamless streaming playback of the playlist at the point of sale and additionally include management of artists' playback royalties. One of our EU project partners is the clothing retail company *Piacenza* that will perform an initial validation of the system in their shops. Since the public-private partnership research project depicted throughout this paper is based on public funding, results of this real-world evaluation will be published in the public domain. Future works by our project group will simultaneously evaluate the possible benefits of employing the prediction model presented here for addressing basic musicological research questions regarding musical meaning attribution.

Acknowledgement

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 688122. We would like to express our gratitude to Geoffroy Peeters and his team from project partner *IRCAM* for their contributions to the machine learning parts of this study. Similarly, we thank our project partner *HearDis* for contributing the stimulus material and tag knowledge, as well as our partners from *Integral Markt- und Meinungsforschung* to contribute with their SINUS meta milieu scales.

References

Allan, D. (2006). Effects of Popular Music in Advertising on Attention and Memory. *Journal of Advertising Research*, 46(4), 434–444. doi: 10.2501/S0021849906060491

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397–438. doi: 10.1080/10705510903008204

Bonneville-Roussy, A., Rentfrow, P. J., Xu, M. K., & Potter, J. (2013). Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *Journal of Personality and Social Psychology*, *105*(4), 703–717. doi: 10.1037/a0033770

Bronner, K., & Hirt, R. (Eds.). (2009). *Audio Branding: Brands, Sound and Communication*. Baden-Baden: Nomos.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*(1), 111–150.

Brunswik, E. (1952). *The Conceptual Framework of Psychology*. Chicago: University of Chicago Press.

Burke, R. (2000). Knowledge-based recommender systems. *Encyclopedia of library and information systems*, *69*(Supplement 32).

Burmann, C., Jost-Benz, M., & Riley, N. (2009). Towards an identity-based brand equity model. *Journal of Business Research, 62*(3), 390–397. doi: 10.1016/j.jbusres.2008.06.009

Burred, J. J., & Peeters, G. (2009). An Adaptive System for Music Classification and Tagging. *Proceedings of the LAS – Learning the Semantics of Audio Signals, Graz, Austria, 2009*.

Chernatony, L. de. (1999). Brand Management Through Narrowing the Gap Between Brand Identity and Brand Reputation. *Journal of Marketing Management, 15*(1–3), 157–179. doi: 10.1362/026725799784870432

Eerola, Tuomas, Friberg, A., & Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.00487

Eerola, Tuomas, Lartillot, O., & Toiviainen, P. (2009). *Prediction of Multidimensional Emotion Ratings in Music from Audio using Multivariate Regression Models*. Presented at the 10th International Society for Music Information Retrieval Conference.

Egermann, H. (in press). Creating a Sounding Image: Psychological Aspects of Audio Branding. In M. Grimshaw, M. Walther-Hansen, & M. Knakkergaard (Eds.), *The Oxford Handbook of Sound and Imagination*. Oxford: Oxford University Press.

Egermann, H., Fernando, N., Chuen, L., & McAdams, S. (2015). Music induces universal emotion-related psychophysiological responses: comparing Canadian listeners to Congolese Pygmies. *Frontiers in Psychology*, *5*, 1341.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods, 4*(3), 272–299.

Gaus, H., Jahn, S., Kiessling, T., & Drengner, J. (2010). How to Measure Brand Values? In M. C. Campbell, J. Inman, & R. Pieters (Eds.), *NA - Advances in*

Consumer Research Volume 37 (pp. 697–698). Duluth (MN), USA: Association for Consumer Research.

Geuens, M., Weijters, B., & De Wulf, K. (2009). A new measure of brand personality. *International Journal of Research in Marketing, 26*(2), 97–107. doi: 10.1016/j.ijresmar.2008.12.002

Gingras, B., Marin, M. M., & Fitch, W. T. (2014). Beyond Intensity: Spectral Features Effectively Predict Music-Induced Subjective Arousal. *Quarterly Journal of Experimental Psychology*, *67*(7), 1428–1446. doi: 10.1080/17470218.2013.863954

Gorn, G. J. (1982). The Effects of Music In Advertising On Choice Behavior: A Classical Conditioning Approach. *Journal of Marketing,* 46(1), 94–101.

Gustafsson, C. (2015). Sonic branding: A consumer-oriented literature review. *Journal of Brand Management*, *22*(1), 20–37. doi: 10.1057/bm.2015.5

Han, B., Ho, S., Dannenberg, R., & Hwang, E. (2009). SMERS: Music Emotion Recognition Using Support Vector Regression. *Computer Science Department*. Retrieved from http://repository.cmu.edu/compsci/514

Herzog, M., Lepa, S., & Egermann, H. (2016). Towards Automatic Music Recommendation for Audio Branding Scenarios. Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR), New York. doi: 10.14279/depositonce-5983

Herzog, M., Lepa, S., Steffens, J., Schönrock, A., & Egermann, H. (2017). Predicting Musical Meaning in Audio Branding Scenarios. Proceedings of the 25th Anniversary Conference of the European Society for the Cognitive Sciences of Music (ESCOM). doi: 10.14279/depositonce-5984

Herzog, M., Lepa, S., Steffens, J., Egermann, H., & Schönrock, A. (2018). How do Musical Means of Expression affect the Perception of Musical Meaning? *ICMPC/ESCOM*. Presented at the ICMPC15/ESCOM10, July 26th 2018, Graz.

Herzog, M., Lepa, S., Egermann, H., Schoenrock, A., & Steffens, J. (2020). Towards a common terminology for music branding campaigns. *Journal of Marketing Management, 36*(1–2), 176–209. doi: 10.1080/0267257X.2020.1713856

Homma, N., & Ulktzhöffer, J. (1990). The internationalisation of Everyday-Life-Research: Markets and milieus. *Marketing and Research Today*, *18*, 197–207.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2019). party: A Laboratory for Recursive Partytioning (Version 1.3-3). Retrieved from https://CRAN.R-project.org/package=party

Hung, K. (2000). Narrative Music in Congruent and Incongruent TV Advertising. *Journal of Advertising, 29*(1), 25–34. doi: 10.1080/00913367.2000.10673601

Jackson, D. M. (2003). *Sonic branding: an introduction* (P. Fulberg, Hrsg.). Basingstoke: Palgrave Macmillan.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. Journal of Experimental Psychology: *Human Perception and Performance, 26*(6), 1797–1813.

Kilian, K. (2009). From brand identity to audio branding. In K. Bronner & R. Hirt (Hrsg.), *Audio Branding: Brands, Sound and Communication* (S. 35–48). Baden-Baden: Nomos.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J.A., and Turnbull, D. (2010). Music emotion recognition: A state of the art review. *Proc. ISMIR*, 255–266.

Larsen, G., Lawson, R., & Todd, S. (2010). The symbolic consumption of music. *Journal of Marketing Management*, 26(7/8), 671–685. doi: 10.1080/0267257X.2010.481865

Leman, M., Vermeulen, V., De Voogdt, L., Moelants, D., & Lesaffre, M. (2005). Prediction of Musical Affect Using a Combination of Acoustic Structural Cues. *Journal of New Music Research*, *34*(1), 39. doi: 10.1080/09298210500123978

MacInnis, D. J., & Park, C. W. (1991). The Differential Role of Characteristics of Music on High- and Low-Involvement Consumers' Processing of Ads. *Journal of Consumer Research*, *18*(2), 161–173. doi: 10.1086/209249

Müllensiefen, D., & Baker, D. J. (2015). Music, Brands, & Advertising: Testing What Works. In K. Bronner, C. Ringe, & R. Hirt (Hrsg.), *Audio Branding Academy Yearbook 2014/2015* (S. 31–51). Baden-Baden: Nomos.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's Guide. Statistical Analysis with Latent Variables. Sixth Edition*. Los Angeles, CA (USA): Muthén & Muthén.

Nandan, S. (2005). An exploration of the brand identity–brand image linkage: A communications perspective. *Brand Management, 12*(4), 264–278. doi: 10.1057/palgrave.bm.2540222

North, A. C., Mackenzie, L. C., Law, R. M., & Hargreaves, D. J. (2004). The Effects of Musical and Voice "Fit" on Responses to Advertisements. *Journal of Applied Social Psychology*, *34*(8), 1675–1708. doi: 10.1111/j.1559-1816.2004.tb02793.x

North, A. C., Sheridan, L. P., & Areni, C. S. (2016). Music Congruity Effects on Product Memory, Perception, and Choice. *Journal of Retailing, 92*(1), 83–95. doi: 10.1016/j.jretai.2015.06.001

Oakes, S. (2007). Evaluating Empirical Research into Music in Advertising: A Congruity Perspective. *Journal of Advertising Research, 47*(1), 38–50. doi: 10.2501/S0021849907070055

Oakes, S., & North, A. (2006). The impact of background musical tempo and timbre congruity upon ad content recall and affective response. *Applied Cognitive Psychology, 20*, 505–520. doi:10.1002/acp.1199

Papadopoulos, H., & Peeters, G. (2011). Joint Estimation of Chords and Downbeats From an Audio Signal. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 138–152. doi: 10.1109/TASL.2010.2045236

Peeters, Geoffroy, Cornu, F., Doukhan, D., Marchetto, E., Mignot, R., Perros, K., & Regnier, L. (2015). When audio features reach machine learning. *International Conference on Machine Learning - Workshop on "Machine Learning for Music Discovery"*. Presented at Lille, France. Retrieved from https://hal.archives-ouvertes.fr/hal-01254057

Peeters, Geoffroy. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project [online]*. Retrieved from IRCAM website: http://recherche.ircam.fr /anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf

Peeters, Geoffroy. (2006a). Chroma-based estimation of musical key from audio-signal analysis. *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*. Retrieved from https://hal.archives-ouvertes.fr/hal-01106203

Peeters, Geoffroy. (2006b). Template-based estimation of time-varying tempo. *EURASIP Journal on Applied Signal Processing*, *2007*, article ID 67215. doi: 10.1155/2007/67215

Peeters, Geoffroy. (2011). Spectral and Temporal Periodicity Representations of Rhythm for the Automatic Classification of Music Audio Signal. *IEEE*

Transactions on Audio, Speech, and Language Processing, *19*(5), 1242–1252. doi: 10.1109/TASL.2010.2089452

Peeters, Geoffroy, Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902–2916.

Peeters, Geoffroy, & Papadopoulos, H. (2011). Simultaneous Beat and Downbeat-Tracking Using a Probabilistic Framework: Theory and Large-Scale Evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(6), 1754–1769. doi: 10.1109/TASL.2010.2098869

Pichl, M., Zangerle, E., Specht, G., & Schedl, M. (2017). Mining Culture-Specific Music Listening Behavior from Social Media Data. *2017 IEEE International Symposium on Multimedia (ISM)*, 208–215. doi: 10.1109/ISM.2017.35

Saari, P., Fazekas, G., Eerola, T., Barthet, M., Lartillot, O., & Sandler, M. (2016). Genre-Adaptive Semantic Computing and Audio-Based Modelling for Music Mood Annotation. *IEEE Transactions on Affective Computing*, *7*(2), 122–135. doi: 10.1109/TAFFC.2015.2462841

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010). *Feature Selection for Content-Based, Time-Varying Musical Emotion Regression*. Proceedings of the MIR'10 (pp. 267–273), Philadelphia, PA (USA).

Shevy, M. (2008). Music genre as cognitive schema: Extramusical associations with country and hip-hop music. Psychology of Music, 36(4), 477–498. doi: 10.1177/0305735608089384

SINUS. (2017). *SINUS Meta-Milieus. Customization all over the world*. Retrieved from https://www.sinus-institut.de/fileadmin/user_data/sinus-institut/Dokumente/downloadcenter/Sinus_Meta_Milieus/Working_with_Sinus-Meta-Milieus.pdf - last accessed 11/25/2019

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*(1), 78–90. doi: 10.1086/209528

Steffens, J., Lepa, S., Herzog, M., Schönrock, A., & Egermann, H. (2018). „Bridging the Semantic Gap" – Kann der semantische Ausdruck von Musik mithilfe von akustischen Signaleigenschaften vorhersagt werden? *Fortschritte der Akustik*. Proceedings of the DAGA-Meeting 2018, München.

Steffens, J., Lepa, S., Herzog, M., Schönrock, A., Peeters, G., & Egermann, H. (2017). *High-level chord features extracted from audio can predict perceived musical expression*. Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), Suzhou, China.

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods*, *14*(4), 323–348. doi: 10.1037/a0016973

Tagg, P. (2013). *Music's meanings: a modern musicology for non-musos*. New York: Mass Media Music Scholar's Press

Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., & Chen, H. H. (2008). Toward Multi-modal Music Emotion Classification. In Y.-M. R. Huang, C. Xu, K.-S. Cheng, J.-F. K. Yang, M. N. S. Swamy, S. Li, & J.-W. Ding (Hrsg.), *Advances in Multimedia Information Processing - PCM 2008* (Bd. 5353, pp. 70–79). doi: 10.1007/978-3-540-89796-5_8

Zander, F. (2006). Musical influences in advertising: How music modifies first impressions of product endorsers and brands. *Psychology of Music, 34*, 465–480. doi: 10.1177/0305735606067158

Tables

Table 1. GMBI_22 questionnaire instrument - factor solution and measurement model for perceived musical expression in branding contexts

Item / Factor	Arousal (-)	Valence	Authenticity	Timeliness
relaxing	**0.782**	0.185	0.311	0.102
soft	**0.740**	0.181	0.250	0.066
chilled	**0.723**	0.140	0.233	0.152
warm	**0.581**	0.473	0.395	0.050
loving	**0.566**	0.360	0.421	0.085
happy	0.210	**0.781**	0.264	0.179
bright	0.187	**0.706**	0.315	0.243
playful	0.160	**0.664**	0.281	0.238
friendly	0.422	**0.648**	0.358	0.113
authentic	0.270	0.334	**0.656**	0.144
honest	0.372	0.324	**0.649**	0.113
detailed	0.288	0.228	**0.632**	0.248
intellectual	0.356	0.083	**0.631**	0.261
trustworthy	0.409	0.347	**0.605**	0.151
creative	0.223	0.314	**0.590**	0.381
passionate	0.315	0.346	**0.578**	0.155
natural	0.463	0.350	**0.540**	0.053
modern	0.132	0.214	0.090	**0.804**
futuristic	0.088	0.035	0.162	**0.688**
young	0.111	0.347	0.082	**0.677**
contemporary	0.257	0.204	0.249	**0.591**
innovative	0.207	0.216	0.482	**0.559**

Note. Coefficients are standardised item weights, values >.5 set in **bold,** factors are orthogonal, polarity of *Arousal* is inversely interpreted due to item formulations

Table 2. Results of machine learning of branding expert music knowledge

Classifier (Tag family)	**Class labels (Tags)**	**No of classes**	**Accuracy**	**Recall**	**F1 score**
genre	Blues, Classical, Dance, Folk, Hip Hop, Jazz, Pop, Rock, Soul/Funk, World Music	10	0.92	0.62	0.62
style	*Style tags are provided below this table**	61	0.98	0.45	0.45
instrumentation	Acapella, Acoustic-Guitar, Brass, Choir, Electric-Guitar, Live Drums, Orchestral, Percussions, Piano, Speech, Strings, Synthetic Drums, Whistle	13	0.81	0.42	0.42
vocals_existing	yes, no	2	0.92	0.92	0.92
vocals_gender	male, female, mixed	3	0.76	0.63	0.63
production timbre	hard, soft, warm, cold, bright, dark	6	0.82	0.46	0.46

*Class labels of style classifier: Afro, Ambient, AOR, Asian, Balearic, Balkan, Blues, Boogaloo, Boogie, Bossa-Nova, Broken-Beats, Calypso, Chanson, Classical-Jazz, Classic-Rock, Contemporary-Classical, Contemporary-Folk, Country, Dancehall, Deep-House, Disco, Downbeat, Dream-Pop, Drum & Bass, Dubstep, Easy-Listening, EDM, Electro, Electro-Pop, Electro-Rock, Flamenco, Folkloric, Funk, Fusion-Jazz, Hip-Hop, Historical-Classical, House, Indie-Dance, Indie-Pop, Indie-Rock, Krautrock, Latin, Mainstream, Northern-Soul, Nu-Jazz, Oriental, Progressive-Rock, Punk, R&B, Rare-Groove, Reggae, Reggaeton, Rock & Roll, Samba, Schlager, Smooth-Jazz, Soul, Tango, Tech-House, Traditional-Folk, UK-Funky

Table 3. Machine learning results for prediction model selection (training set vs. cross-validation)

Variable	Model type	R^2 (training)	R^2 (9-fold CV)
Arousal	Hierarchical stepwise regression	**.87**	**.61**
	Random forest regression	.83	.60
Valence	Hierarchical stepwise regression	.73	.38
	Random forest regression	**.80**	**.44**
Authenticity	Hierarchical stepwise regression	.79	.54
	Random forest regression	**.86**	**.55**
Timeliness	Hierarchical stepwise regression	**.85**	**.74**
	Random forest regression	.85	.66

Note: coefficients of finally selected model types set in **bold**

Table 4. Relative explanatory potential of predictor blocks obtained with the hierarchical stepwise regression (HSR) and random forest regression (RFR) approach as selected for the four dependent variables (based on population sample)

Predictor block	**Arousal R^2 (HSR)**	**Valence R^2 (RFR)**	**Authenticity R^2 (RFR)**	**Timeliness R^2 (HSR)**
IRCAM beat	.18	.51	.63	.32
IRCAM keymode	.03	-.12	-.13	.01
IRCAM chord (adapted)	.04	.04	.04	.02
ML Instrumentation	.24	.21	.19	.21
ML Musical style	.23	.06	.04	.25
ML Musical genre	-	.01	.01	-
IRCAM descriptor	.07	.04	.02	.02
ML Production timbre	.01	-	-	-
IRCAM timbre toolbox	.04	.02	.03	.02
Σ R^2	.84	.77	.83	.85

Table 5. Increase of prediction accuracy (R^2) per target group when using specific target group models instead of the population model for predicting GMBI factor scores; number of underlying consumer ratings for each target group given in parentheses

Target group	**Arousal**		**Valence**		**Authenticity**		**Timeliness**	
	R^2 pop. model	R^2 incr.	R^2 pop. model	R^2 incr.	R^2 pop. model	R^2 incr.	R^2 pop. model	R^2 incr.
population (53,344)	0.68	0	0.39	0	0.45	0	0.69	0
male (26,667)	0.63	no gain	0.37	no gain	0.47	no gain	0.62	0.01
female (26,677)	0.62	0.01	0.35	0.03	0.37	no gain	0.69	no gain
UK (17,512)	0.63	0.01	0.19	0.11	0.39	0.05	0.64	no gain
Spain (17,677)	0.63	0.05	0.41	0.09	0.13	0.01	0.47	0.11
Germany (18,155)	0.50	0.04	0.06	0.18	0.45	no gain	0.64	no gain
age 52-68 (18,074)	0.59	0.02	0.31	no gain	0.45	no gain	0.57	0.05
age 35-51(17,805)	0.63	no gain	0.38	no gain	0.23	0.06	0.64	no gain
age 18-34 (17,465)	0.63	no gain	0.32	0.04	0.39	0.03	0.67	0.02
male, age 52-68 (9,196)	0.57	no gain	0.26	no gain	0.44	no gain	0.53	no gain
female, age 52-68 (8,878)	0.51	0.03	0.27	0	0.35	no gain	0.48	0.07
male, age 35-51 (8,941)	0.56	no gain	0.36	no gain	0.19	0.10	0.53	0.02
female, age 35-51 (8,864)	0.39	0.01	0.05	0.10	0.06	0.02	0.45	no gain
male, age 18-34 (8,530)	0.34	0.12	0.22	0	0.32	0.04	0.53	0.02
female, age 18-34 (8,935)	0.58	no gain	0.29	0.03	0.28	0.02	0.64	0.01
UK, male (8,676)	0.57	no gain	0.16	0.06	0.40	0.07	0.53	0.02
UK, female (8,836)	0.49	0.10	0.16	0.15	0.29	0.01	0.61	no gain
Spain, male (8,913)	0.56	0.01	0.33	0.05	0.11	no gain	0.42	0.06
Spain, female (8,764)	0.56	0.06	0.35	0.08	0.09	0.05	0.38	0.11
Germany, male (9,078)	0.39	0.08	0.07	0.16	0.44	0	0.54	no gain
Germany, female (9,077)	0.45	no gain	0.03	0.14	0.33	no gain	0.59	no gain
UK, age 52-68 (6,064)	0.36	0.12	0.19	0.07	0.34	0	0.54	no gain
UK, age 35-51 (5,855)	0.54	0.03	0.11	0.04	0.15	0.07	0.53	no gain
UK, age 18-34 (5,593)	0.57	no gain	0.11	0.15	0.23	0.1	0.47	no gain
Spain, age 52-68 (5,892)	0.54	0.05	0.27	0.02	0.31	no gain	0.23	0.14
Spain, age 35-51 (5,934)	0.54	no gain	0.32	0.06	-0.02	0.12	0.42	no gain
Spain, age 18-34 (5,851)	0.50	0.01	0.33	0.07	-0.18	0.20	0.44	0.06
Germany, age 52-68 (6,118)	0.45	no gain	-0.03	0.23	0.30	no gain	0.42	no gain
Germany, age 35-51 (6,016)	0.36	0	0.02	0.11	0.24	0.01	0.54	no gain
Germany, age 18-34 (6,021)	0.30	0.08	0.09	0.05	0.39	no gain	0.55	no gain
Established (4,044)	0.51	no gain	0.09	0.02	0.35	0.05	0.50	no gain
Intellectuals (3,912)	0.50	no gain	0.20	0.05	0.25	0.05	0.35	no gain
Performers (3,996)	0.45	no gain	0.13	0.08	0.29	0.12	0.31	no gain
Cosmopolitan Avantgarde (2,838)	0.44	no gain	0.33	0.09	0.34	0.04	0.55	no gain
Adaptive Navigators (4,344)	0.54	no gain	0.00	0.15	0.12	0.17	0.40	no gain
Modern Mainstream (5,766)	0.53	0.02	0.30	no gain	0.18	0.01	0.42	no gain
Traditionalists (3,954)	0.59	no gain	0.30	0	0.15	0.10	0.49	no gain
Consumer Materialists (4,014)	0.37	0.06	0.26	0.06	0.05	0.22	0.26	no gain
Sensation-Oriented (6,804)	0.20	0.14	0.16	0.16	0.06	no gain	0.04	0.19

Figures

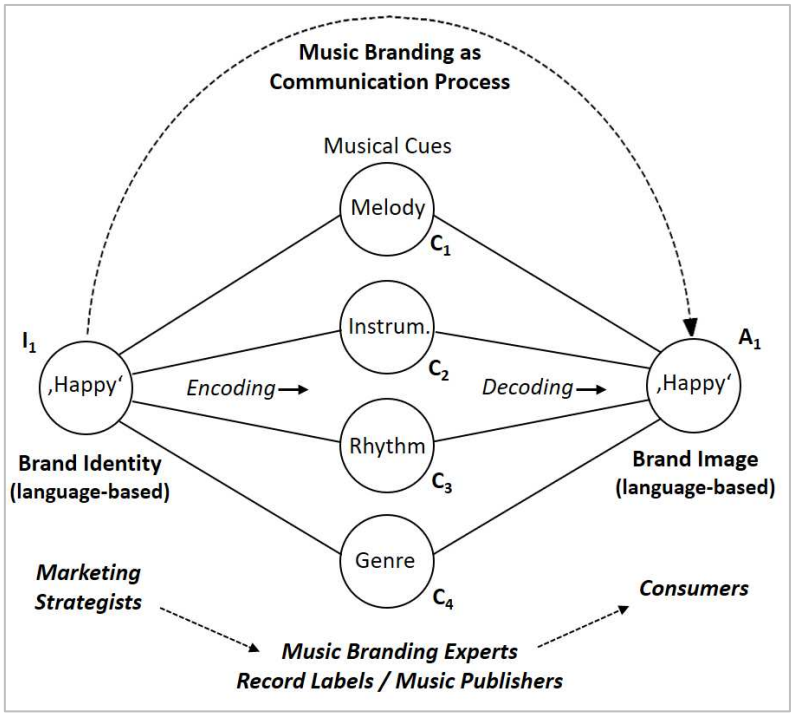

Figure 1. Music branding as formalised communication process

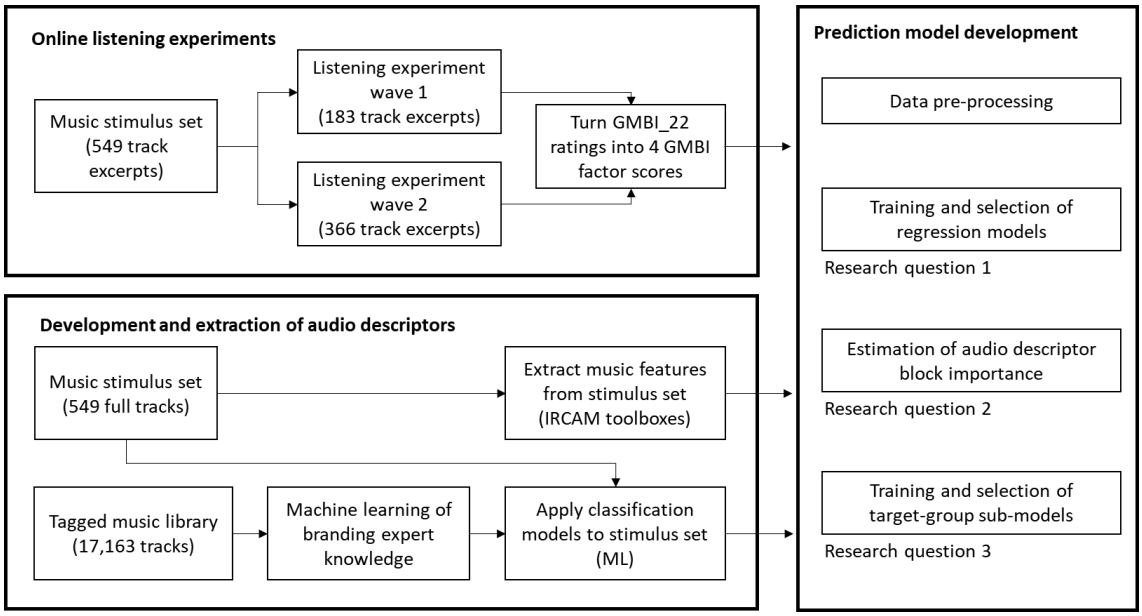

Figure 2. Schematic overview of the methodological steps taken in this paper

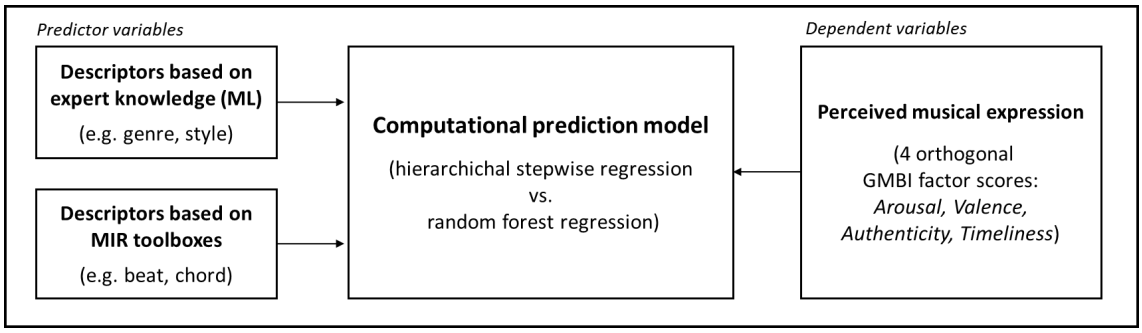

Figure 3. Schematic overview of the variables used to estimate the final prediction model

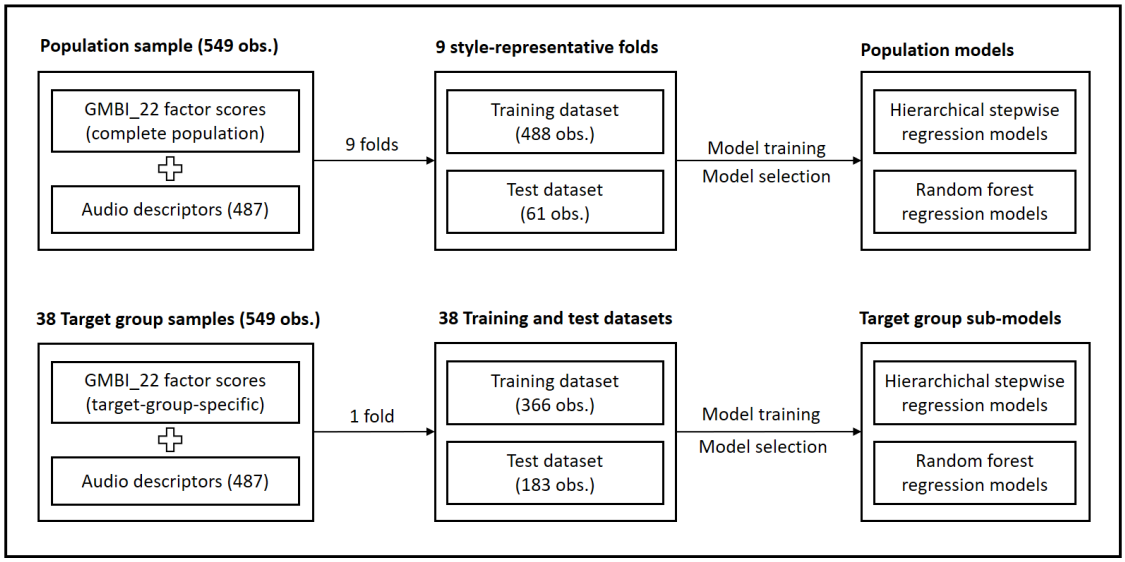

Figure 4. Schematic overview of the training and test datasets, as well as the ML prediction procedures employed in the study