Survey on the Evaluation of Generative Models in Music

ALEXANDER LERCH*, Music, Georgia Institute of Technology, Atlanta, United States
CLAIRE ARTHUR*, Music, Georgia Institute of Technology, Atlanta, United States
NICK BRYAN-KINNS*, Creative Computing Institute, University of the Arts, London, United Kingdom of Great Britain and Northern Ireland
COREY FORD*, Creative Computing Institute, University of the Arts, London, United Kingdom of Great Britain and Northern Ireland
QIANYI SUN*, Music, Georgia Institute of Technology, Atlanta, United States
ASHVALA VINAY*, NoneType Computing, Atlanta, United States

Research on generative systems in music has seen considerable attention and growth in recent years. A variety of attempts have been made to systematically evaluate such systems. We present an interdisciplinary review of the common evaluation targets, methodologies, and metrics for the evaluation of both system output and model use, covering subjective and objective approaches, qualitative and quantitative approaches, as well as empirical and computational methods. We examine the benefits and limitations of these approaches from a musicological, an engineering, and an HCI perspective.

CCS Concepts: • **General and reference → Surveys and overviews**; **Evaluation**; *Metrics*; • **Human-centered computing** → *HCI design and evaluation methods.*

Additional Key Words and Phrases: Music, Evaluation, Generative AI, Survey

1 Introduction

In recent years, advances in system architecture and training methodologies of generative systems in machine learning have led to increasingly powerful systems that have been applied to a variety of tasks and use cases. Particularly prominent are systems in natural language processing (e.g., [33]) and image models (e.g., [148, 215]). Generative systems for music, while not featured as prominently in the media, have also seen considerable progress. Although the concept of computer-generated music dates back to the mid-20th century [125, 246] and has been actively researched since then [59, 87], there has been a rapid increase in interest and publications in the past five years [57], mostly driven by advances in neural approaches. Inspired by Herremans et al. [123], we categorize generative music systems based on (i) architecture, (ii) output, and (iii) input or control.[1] A meta-review by Civit et al. [57] lists the prevalent *architectures* with decreasing number of occurrences as Recurrent Neural

[*] The first author conceptualized the paper and led the paper writing. The remaining co-authors contributed equally to the paper writing and are listed alphabetically.

[1] References will be limited to a select number of representative systems.

Authors' Contact Information: Alexander Lerch, Music, Georgia Institute of Technology, Atlanta, Georgia, United States; e-mail: alexander.lerch@gatech.edu; Claire Arthur, Music, Georgia Institute of Technology, Atlanta, Georgia, United States; e-mail: claire.arthur@gatech.edu; Nick Bryan-Kinns, Creative Computing Institute, University of the Arts, London, London, United Kingdom of Great Britain and Northern Ireland; e-mail: n.bryankinns@arts.ac.uk; Corey Ford, Creative Computing Institute, University of the Arts, London, England, United Kingdom of Great Britain and Northern Ireland; e-mail: c.j.ford@qmul.ac.uk; Qianyi Sun, Music, Georgia Institute of Technology, Atlanta, Georgia, United States; e-mail: qsun75@gatech.edu; Ashvala Vinay, NoneType Computing, Atlanta, Georgia, United States; e-mail: ashvala@gatech.edu.

ACM 1557-7341/2025/9-ART
https://doi.org/10.1145/3769106

Networks (RNNs) [114, 182], Feed-Forward Networks (FF) [73, 266], Variational Audio Encoders (VAEs) [70, 223], evolutionary algorithms [67, 235], and Transformer-based approaches [61, 72, 128], sometimes combined with Generative Adversarial Networks (GANs) [73, 266], with Diffusion networks [50, 85, 86, 168, 191, 227], or with Rectified Flows [164]. Details of many typical neural architectures for music generation can be found in Briot et al. [32].

A key distinction of *output* types of generative systems in music is whether the output is audio [2, 61, 71] or symbolic (e.g., MIDI, MusicXML, etc.) [128, 223]. The length of the output can also vary: it might range from as short as a single note or event in the case of audio synthesizers [81, 82, 153] over phrases [131, 189] to complete musical pieces [70, 157]. Furthermore, systems might generate a single-voiced melody [115, 266] or a polyphonic output with multiple voices [114, 122, 131, 196].

The *input* of a system depends very much on the design goals. A system might require no input at all [203], a few parameters for conditioning [51, 78], a text prompt [2, 61], a melody to be harmonized [114, 268], or a musical phrase to be continued [128, 131] or inpainted [8, 183, 205]. While this variety of approaches and the multitude of available studies imply rapid progress, this progress is hard to quantify, and there is evidence that the quality improvements might not be as dramatic as the number of publications suggests [270].

All generative systems pose challenges in terms of evaluation since a ground truth target, or unique “correct” reference result, does not usually exist. Systems targeting the generation of artistic output are particularly difficult to assess due to the subjectivity of aesthetic assessment. The assessment of music poses a unique set of challenges due to (i) its sequential yet highly structured form, (ii) the abstract musical language and the resulting unclear definition of content in music, (iii) the limited musical meaning of commonly-used music descriptors and the corresponding inadequacy to fully represent the multi-dimensionality of music and musical expression, (iv) the context-dependent interaction between expectation and surprise, and (v) the constant reinterpretation of musical ideas through music performance.

These challenges have led to a large variety in approaches to system evaluation with a multitude of different evaluation targets, methodologies, and metrics. Inter-study inconsistencies in evaluation make the comparison of research results essentially impossible. If results cannot be compared, do not sufficiently reflect the actual quality of a system, or have been acquired in very different settings, the notion of progress in this field becomes questionable, as we cannot measure progress without relevant, commonly used metrics. Despite these problems being recognized as important challenges [32, 165, 267], no general solutions have been proposed, and evaluation still seems to be largely neglected or treated as an afterthought. For instance, Civit et al. [57] provide a meta-review of generative music systems but only mention evaluation in passing. Zhao et al. [273] review prompt-based generative music systems but refer only to the evaluation of creativity as an unsolved challenge. While Bandi et al. [11] present a dedicated evaluation section in their extensive review of generative systems, music is unfortunately not discussed. The only two exceptions are Ji et al. [139] and Wang et al. [258], summarizing some objective and subjective approaches to evaluation.

Therefore, the goal of this article is to provide an accessible, interdisciplinary overview on current empirical and quantitative approaches to the evaluation of generative systems in music. Figure 1 gives a summary and accessible flow chart of the approaches presented. The article provides an in-depth discussion of evaluation targets and methodologies for the assessment of the output of generative systems as well as the user interaction with such methods (rather than other assessment targets such as sociological implications, etc.). In order to do so, we first introduce a comprehensive overview of the dimensions or targets to be evaluated in Sect. 2, followed by a description of methodologies and metrics to evaluation of system output and user interaction in Sects. 3 and 4, respectively. We conclude with a discussion on challenges and future directions in Sect. 5 and final remarks in Sect. 6.

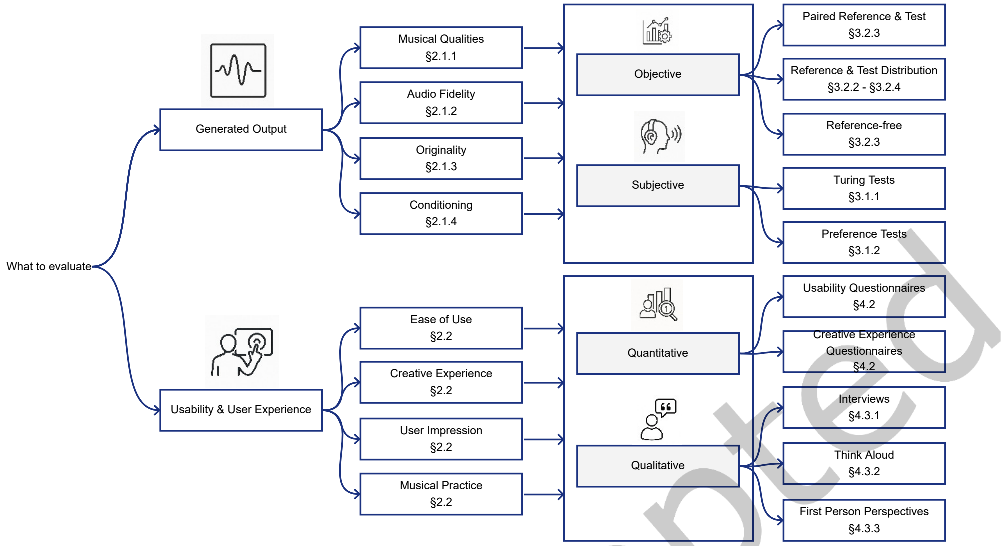

Fig. 1. Structure of the presented evaluation approaches and methodologies.

2 Evaluation targets

The main goals of evaluating a machine learning system are (i) determining whether the system works as intended and to what degree and (ii) comparing it (quantitatively) with other systems. These goals, however, can have many facets as an evaluation can focus on different targets. Wang et al. [259] group evaluation targets into "data-quality evaluation" and "property-controllability evaluation," the latter focusing on specific output properties that are implicitly or explicitly controlled. Pasquier et al. [204] lists the following aspects of a generative system to be assessed: quality, creativity, believability, complexity, robustness, and reliability. In this work, we propose to group the main evaluation targets into *system output* and *model use* while acknowledging that there are aspects of the model itself and the process of creation that could be subject of evaluation as well.

One of the most common assessment targets is the *system output.* Although the quality of the output is the arguably most intuitive criterion to evaluate, there are several aspects of quality ranging from artistic quality to perceptual audio quality, as well as confounding influences that make the evaluation of output quality a potentially challenging endeavor. To give an example of such confounding influences, we can easily imagine an inexperienced listener confusing the artistic quality of a piece of music with the audio quality of its rendition or the immersiveness of the recording if asked for quality. Other dimensions of the output to be evaluated beyond quality include the diversity and originality of the output or how well certain properties of the generated output (e.g., style, rhythmic complexity, or instrumentation) match expectations.

Given that music is a fundamental form of human creativity [154], it is also critical to evaluate *model use* — how easy to use, enjoyable, and engaging models are for people when they make music, whether they be hobbyists, music students, or professional musicians. A model may produce high quality output, but if it is unusable then it has little value for making music. Evaluating the use of models draws on research in Human-Computer Interaction (HCI) [212] to assess whether a model is, e.g., usable by people [190], or provokes surprise [43] or reflection [92].

2.1 System output evaluation

As mentioned above, the most common target for evaluating a generative system is its generated output; the main point of most generative systems is, after all, to produce high quality output or at least output that matches expectations. In this case, the system can be treated as a black box for evaluation as knowledge of internal processes of the system is unnecessary [58].

Different generative music systems may produce a variety of output formats, which in turn might require different evaluation methodologies and metrics. The output can vary from a single note or sound as generated by synthesizers, a single-voice melody, a polyphonic or multi-voiced musical snippet, to a complete piece of music conforming to structural and other musical expectations. Furthermore, the output may —in addition to the basic score-based information such as rhythm, harmony, melody— contain performance information such as tempo and micro-timing, expressive intonation, and dynamics.

The output can either be in an audio format such as PCM [195], or in a symbolic format such as MIDI [249] or MusicXML [105]. Note that while any symbolic output might or might not contain (partial) performance information, an audio signal as a "physical rendition of musical ideas" [160] automatically contains performance information.

The evaluation of the quality of these output signals can be as multi-faceted as different system design goals. This section aims at introducing the main directions of inquiry for the evaluation of aesthetic quality, the audio fidelity, the originality of the output, and its semantic relevance.

2.1.1 Aesthetic and musical qualities. Despite the fact that the assessment of aesthetic quality is possibly the most commonly stated evaluation target in the literature, it arguably is the most challenging to operationally define. While aesthetic quality (as applied to a musical artifact) is a catch-all phrase encompassing many attributes such as balance, complexity, novelty, etc. [29], it is most commonly evaluated only in the singular dimension of subjective preference, i.e., how much a listener "likes" a piece of music, and/or how "interesting" it is [68]. However, individual differences, particularly those that arise from the acquisition of musical expertise, are known to impact such aesthetic judgments [208].

In order to compare across studies, it is important that researchers explicitly operationally define the traits they aim to measure, not only for themselves, but to potential participants as well. For instance, Brattico and Jacobsen [29] mention the important distinction between *affective* responses (those that induce or modulate emotions or mood), hedonic responses (those that modulate reward; likes and dislikes), and aesthetic responses, which typically refer to inherent style-relevant attributes which lend the artifact beauty, elegance, or coherence. To complicate matters, in addition to style-relevant traits, the medium of creation (e.g., score versus audio), the cultural context (e.g., what is valued inside versus outside the culture), and the caliber or quality of the object or its execution are all important criteria that factor into the equation of evaluating aesthetic goals [97]. For instance, music generation that outputs a score or transcription may be evaluated on its adherence to various compositional norms, such as adherence to an appropriate vocabulary and grammar, the arrangement and organization of musical ideas, and the use of variation and repetition, to name a few [239]. On the other hand, the output rendered as an audio recording may be evaluated based on the execution of the performance in relation to parameters such as the authenticity or "humanness" of the performance, the expressivity or dynamicism, potentially in addition to factors related to the underlying composition itself. These criteria are highly multifaceted and context-dependent. For instance, the analysis and assessment of music performance is a research field in itself [149, 162, 163].

While the assessment of aesthetics typically involves human evaluation, there have been attempts at computational aesthetic evaluation. Galanter [97] gives an overview of such methods through 2012, noting that "computational aesthetic evaluation is an extremely difficult problem," and that it frequently "leads to deep philosophical waters regarding phenomenology and consciousness." Models that have been proposed and used in a musical context, have typically relied on evaluation by adherence to some set of statistical measures or proportions. Some AI models that generate an artistic work may be designed so as to implicitly include the goals of such fitness metrics, leading to increased diversity, for example. As pointed out by Galanter [97], however, "[c]reating evolutionary diversity and dynamics via artificial aesthetics foreign to our human sensibility is one thing. Appealing to human aesthetics is quite another." Other computational avenues for evaluation have included models based complexity — frequently drawing on Berlyne's theory of aesthetics, or Shannon information theory [97].

2.1.2 *Audio quality.*

The assessment of audio quality, sometimes also referred to as fidelity, is important for a variety of applications, including measuring the quality of audio codecs, the transmission quality of a channel, or the quality of recording or reproduction of audio equipment. Typical factors impacting the audio quality are non-linear processing such as distortion, changes in the spectral content such as bandwidth reduction, additive sources such as noise, and time-varying processes such as gain manipulation or spectro-temporal processing. When assessing quality in this context, the expectation is that the generated audio signal is free of artifacts and impairments that might negatively impact the human listening experience. The amount (e.g., very little noise vs. a lot of noise) of the impairment directly affects the perceived quality of a signal. The estimation of the quality of a signal is easiest in comparison to an identical but unimpaired reference signal.

A common way to estimate audio quality is through listening studies, with methodologies such as MUSHRA [136], which allows to compare the quality of multiple audio signals with respect to a reference signal. Simple objective measures for audio quality such as the Signal-to-Noise-Ratio only have limited perceptual meaning, but there exist objective methods which model perceptual qualities, e.g., PEAQ [134] and ViSQOL [126].

In addition, attempts have been made to develop reference-agnostic measures of audio quality. This reference-free approach is popular in speech quality assessment. For speech, a clear expectation on quality and intelligibility can be established making the availability of reference signals unnecessary or, at least, less important. In the case of music, however, the artistic and creative use of effects and heavily processed audio question any pre-conceived framework of quality criteria. Thus, the existence of a reference signal is usually deemed necessary for music, as an undesired quality impairment is not always obviously distinguishable from an impairment stemming from artistic intent. The generation of a distorted synthesized sound, for instance, can be perfectly desirable.

2.1.3 *Originality.*

The originality of the output is a common concern when evaluating a generative system. We interpret originality in three ways. First, originality is a concern with respect to *plagiarism*. Commonly, a generative system is expected to create novel output that does not replicate the training data. Second, the *diversity* of the model output should match the diversity of the training data in all relevant dimensions. This validates the effectiveness of the training. Last but not least, we may want to measure the *creativity* of the system output.

Plagiarism. Modern machine learning systems require a potentially massive amount of training data. In most music scenarios, the (partial) reproduction of memorized training data might be regulated by laws. Although systems are generally trained with the intent to generate novel artifacts, they might memorize individual training samples and reproduce them during inference [104, 253]. When that happens, claims of plagiarism can arise similar to when a human composer copies musical ideas from existing works. Examples for memorized output from generative systems in other fields include reproduced names and email addresses for language models [45, 129], and reproduced images for image generation models [44]. The analysis of plagiarism in music is more complicated, so cases are not often directly or easily identifiable. Therefore, most of the current discussion in music focuses on copyright [159] and the use of copyrighted data for training and whether this establishes an infringement of copyright [69, 245]. Currently, we see the first lawsuits around this topic reaching the courts

[151]. As such potential infringements and associated questions on liability are not necessarily easily addressed and answered, assessing the probability of a generative system reproducing memorized content can be part of an overall evaluation strategy.

Diversity. The generated output of a generative system is expected to show some variety, regardless of use of input prompts or conditioning. Generating only variations of the same piece, for example, is generally not desirable. Banar and Colton [9] point out the importance of measuring diversity of the output in various dimensions as the multitude of musical properties cannot be captured in a single dimension. For common machine learning approaches, the diversity of the generated output should match the diversity of the training set. The exact dimensions to measure diversity on, however, are not necessarily easy to define; they might, for instance, include genre diversity, metric and rhythmic diversity, melodic and harmonic diversity, or diversity of instrumentation and timbre. To complicate matters, Liu and Jin [169] show that modern audio synthesis systems often display a trade-off between output diversity and quality of the output.

Creativity. Creativity is a notoriously difficult to define concept. As pointed out by Jordanous [142], there are not even standard definitions of creativity within US or UK law, “despite the need to detect the presence of creativity for legal reasons.” Even more stubborn are the issues that arise in attempting to *assess* creativity, as pointed out by Rohrmeier [224]: “It is hard to assess something that one cannot define, and this reflects down to the difficulties in evaluating the success of models of general creativity without resorting to the ‘oracle’ of human evaluators.” While Rohrmeier is referring to the assessment of models of general creativity, the same issues arise in the assessment of creative artifacts or model outputs, regardless of the methods of creation [46]. Despite the fact that creativity is recognized as an ill-defined phenomenon, studies continue to attempt to evaluate it, as it remains a crucial component of artistic creation [156]. However, in many (if not most) modes of musical creation, one is bound by a set of rules or constraints [54]. Again, Rohrmeier [224] raises the question of creativity in the context of style imitation where such rules and constraints can, in some cases, be extreme: “What do we mean by ‘creativity’, and how do we relate novelty, innovation or transformation with the concept? For instance, are models ’creative’ that generate jazz lead sheets, chorales in Bach’s style, Indian tabla, or Balinese Gamelan? Is style replication ‘creative’?” As pointed out by Agres et al. [3], creativity *can* be evaluated in a limited sense in a highly constrained context —such as the harmonization of a melody in a strict chorale style— as the comparison between human and computational ability to solve a set of problems. This general logic can be extended to theoretically any aspect of creativity. However, even in only evaluating the artifact of a style-imitation task, this procedural or problem-solving definition becomes a slippery slope towards pure determinism, which has been argued to be the opposite of (or, at least, hindering to) creativity [224, 275]. In addition, many attempts to define creativity (e.g., [25, 60, 142, 222]) commonly include the notions of value (either aesthetic or utilitarian), the combination or connection of ideas or phenomena, and exploration and transformation within some conceptual space, all of which are missed in the constrained problem-solving definition of style imitation described above.

Nevertheless, numerous scholars in computational creativity and generative AI agree that, despite the inherent difficulties, assessing creativity —and having standard, scientific methods for doing so— is crucial for the field to grow and improve [142, 222].

2.1.4 Conditioning. A common approach to the evaluation of generative systems is to compare specific properties and characteristics of the generated output with the expectation. A system targeting, for instance, the generation of chorales should not output symphonic music, even if the generated music were aesthetically pleasing and original.

Some of the conditioning characteristics are implicitly defined through curating the training data prior to training the system. These could include, e.g., musical genre or style [261], pitch, sonic quality [80, 82], instrumentation, length, complexity, and mood [102]. Other characteristics can be controlled explicitly either through input conditioning [247] or regularization [207]. Examples for such characteristics are rhythmic complexity [206] and arousal [247].

Evaluating input conditioning is, in many cases, methodically relatively straight-forward, as the input specifies a target value that the output has to match. If the property can be quantified and measured from the generated output, it can be directly compared with the target value. Evaluating the characteristics implicitly specified through the input data can be more challenging as the number and relevancy of properties might not be known, drawing parallels to the evaluation of musical qualities introduced above. Any assessment poses the challenge of identifying a meaningful and complete set of descriptors indicative of the data characteristics to assess.

Thus, this direction of evaluation —focusing on individual output properties and whether they match the user input or training dataset properties— is mostly useful for the verification that the training process was successful. The evaluation of individual properties allows a very targeted quantitative validation of the semantic relevance of the output with respect to certain, usually narrowly defined characteristics. However, measuring and interpreting specific characteristics of the output as proxies for an overall assessment of the output is, at best, questionable. Any set of characteristics represents only a small subset of a possibly infinite number of musical characteristics, and thus can only give a snapshot of one facet of the generated output. It is particularly problematic if the measured property was explicitly used as a training target or loss function, cf. Goodhart's law: "When a measure becomes a target, it ceases to be a good measure" [106, 236].

A special, but recently very popular form of conditioning is the text prompt. Unlike the evaluation of other conditionings, measuring the (perceptual) alignment of the text prompt with the musical output is challenging due to the undefined structure and terminology or the prompt and the multitude of potentially impacted musical properties. Given these challenges, most evaluation strategies focus on global measures of fit [108, 130, 167].

2.2 Usability & user experience

Outputs from generative music systems are used and appreciated by people from musicians to audiences. In this section, we focus on evaluation of people's direct interaction with generative systems. Viewing this through a HCI [212] lens we refer to the people as users who interact with the generative system through a real-time user interface. Unlike evaluation of system outputs using listening tests described above, evaluation here is concerned with understanding the interaction between the user and the generative model — the human-in-the-loop. This aligns with recent Human-Centered AI discourse [198, 231] advocating for the use of HCI methods to evaluate and inform the design of AI systems which balance automation with human control. There are two main aspects of interacting with computers, and AI models specifically, that are usually evaluated: the *usability* and the *user experience*.

The assessment of *usability* asks how easily the generative system is to use [190]. Unlike evaluating the output of the generative system, usability is concerned with how easy the generative system is to control and to understand [6]. Usability evaluation is often best situated within the wider socio-technical system of use [76]. Within creative practice, the usability of the system will impact its use and uptake and whether the system is even accepted in a music making context. For example, a generative system may produce aesthetically pleasing outputs, but if it is not usable or controllable it will be less likely to be used in music making practice [171].

The assessment of *user experience* includes collecting subjective and experiential responses to using the generative system [194], focusing on evaluating the experience of interacting with the system. This may be, for example, an evaluation of people's affective responses to interaction with the system or an aesthetic evaluation of how the use of the AI relates to music making practice. A person's experience might include hyper-awareness, anxiety, or feelings of control over a situation [62]. It might also include feelings of confusion or creative failure when making music [118], or feelings of surprise when finding unexpected discoveries in AI-generated content [43]. In essence, the evaluation of user experience is about assessing people's subjective feelings of using a system.

It is important to note that usability and user experience are interrelated in complex ways — a system does not necessarily need to be usable to be enjoyable and rewarding to use (e.g., computer games purposefully introduce challenge and frustration to the user experience yet offer a rewarding user experience [24]). We see this, for instance, with traditional musical instruments which require years of time and effort to learn, yet this challenge becomes intrinsically rewarding [48].

3 Evaluation of system output

The previous Sect. 2 established *what* should be evaluated: the evaluation targets; Sections 3 and 4 review *how* these targets can be assessed, discussing methodology and metrics. For evaluating the generated output of a system, we first discuss subjective evaluation through listening experiments. Then, we describe objective approaches to assess the quality of the output.

3.1 Subjective evaluation

A large problem in the evaluation of system output is that, while one could argue that *style imitation* could be measured fairly objectively without the intervention of human opinion, nevertheless, the evaluation of the utility, aesthetics, creativity, or "humanness" is most aptly carried out by a human observer as these are all, essentially, value judgments [46, 228]. However, given that the definitions of these traits (utility, aesthetics, etc.) are subjective, largely personal, and subject to contextual information, it remains a significant challenge to design unbiased subjective metrics that would function in an "all purpose" manner.

The most typical method for evaluating music generation to date has been by asking listeners [267]. Unfortunately, there are no standardized approaches to the subjective evaluation for almost any Music Information Retrieval (MIR) task or product, including AI-generated music. The most common methods for assessing generative music include Turing-style tests —designed to evaluate whether the AI-generated music can pass as human-generated— or "preference tests," surveys or experiments assessing aesthetic quality, musicality, and originality. In the ensuing subsections we review these common approaches to subjective evaluation, including both the methodologies themselves as well as the implicit and explicit criteria being evaluated.

3.1.1 Turing tests. Turing-type tests —where a listener must identify whether a musical selection was made by a human or AI— remain one of the most common forms of subjective evaluation [120]. In its most basic form, the Turing test is a useful metric in the sense that the method is simple (i.e., typically binary forced selection), and that it offers a *theoretically* unbiased subjective evaluation by, in principle, implicitly evaluating a model's output according to a scale of "humanness." If a listener cannot tell apart machine from human-generated musical output, then the implication is that the machine is "at least as good" as a human. However, this logic only follows under the right conditions, which may not be met in small-scale, ad-hoc experiments. For example, what makes a Bach chorale different from another piece of music from the classical genre may be non-evident to a lay listener; similar arguments can be made for a jazz solo. In other words, recognition of the norms of a particular musical style typically takes at least a small degree of expertise [208]. Another consideration is the material used in the test itself. Since not *all* of the output material can be evaluated, only a small subset of the model output is used in the test. However, depending on how this material is selected, this selection may not adequately represent the model output overall. In addition, inferential statistical tests designed to test an alternative hypothesis against a null hypothesis, such as t-tests, are commonly used to evaluate the outcome of a Turing test. Yet typically the desired outcome is that the null hypothesis (i.e., no difference) is actually supported. This inappropriate use of such a test will "bias" towards supporting a null hypothesis is compounded by studies that rely on small sample sizes or exhibit high variance. The use of Turing tests in evaluating AI output has been criticized for a variety of reasons, most notably for its lack of sophistication, and for the tool being repurposed for something other than what it was intended for [121], which was as an evaluation of intelligence and not of aesthetics. It is important to note is that the Turing test conflates indistinguishability (between human and computer) with aesthetic or creative success. It is easily conceivable that a model trained on student input, for instance, would ask a listener to disambiguate between two samples both representing 'amateur.' In this case, a lack of ability to discriminate between the samples does not conclude that the output is of high aesthetic or creative value. Finally, an inherent problem with using a Turing test as a subjective evaluation metric is that it inherently over-rewards imitation over creativity [209]. Nevertheless, if used appropriately, the Turing test *can* be used to assess model output, provided that these considerations are taken into account and that the test is primarily used to support a conformity to a baseline standard of a specific musical style, rather than an indication of musically or aesthetically valued output.

3.1.2 Preference tests. Other forms of subjective evaluation involving human ratings inevitably fall under the broad category of "perceptual preference tests," where listeners rate their perception of various artifacts according to several criteria, such as aesthetic quality, creativity, musicality, fidelity, etc. As mentioned, however, since there are no standardized testing practices for subjective perceptual tests, the questions, the scales used, and the experimental designs all differ widely from study to study (e.g., Chu et al. [54], p.306). The most common criteria evaluated in preference tests include the overall quality, preference or enjoyment, stylistic appropriateness, complexity, coherence, aesthetic response or "interestingness," and musicality; most commonly evaluated on a Likert-type scale [166]. Despite the wide variety of criteria and survey designs, several scholars have nevertheless made commendable efforts towards breaking down these complex properties in a way that could help impose some common criteria or benchmarks in the design of subjective evaluation practices. For example, Chu et al. [54] reviewed 40 music generation studies that included a subjective evaluation component and reduced the various criteria into the eight categories *Overall*, *Melodiousness*, *Naturalness*, *Correctness*, *Structureness*, *Rhythmicity*, *Richness*, *Creativity.* Interestingly, the authors imply that *Creativity* was not included as a subjective criterion in any of the studies reviewed, but the authors added this eighth criterion as they felt that "prior research emphasize[d] the role of AI to boost human creativity in music composition." Unfortunately, however, as presented, these criteria appear difficult to isolate as, for example, the criterion "melodiousness" —defined by the question "are the music notes' relationships natural and harmonious?"— appears to overlap with the criterion of "naturalness," which asks, "how realistic is the sequence?"

The measurement of audio quality of a signal can be treated as a special case of preference tests with established procedures and methodologies. Audio coding is one of the main fields where the measurement of (perceptual) audio quality plays a crucial role and has driven the standardization of procedures to improve replicability of results. Thus, standards have been introduced that regulate not only the general methodology of listening tests but also number, selection and training of the listeners, selection of audio stimuli, properties of the reproduction equipment, as well as other factors such as room acoustics [135]. Two standards are most commonly followed for determining musical audio quality: ITU-R BS.1116 for high quality signals and ITU-R BS.1534 for medium quality signals. Both require a reference signal to be present, and rate the quality on a five point scale, although the scales are defined differently to accommodate for the different targets. In both cases, a high rating is indicative of higher audio quality. BS.1116 [135] is a so-called double blind, triple stimulus test with hidden reference, where two signals are presented alongside the reference signal and one of the two presented signals is a hidden reference signal. The listener then rates the two signals on the five point scale. BS.1534 [136], also referred to as MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA), follows a similar methodology, but adds an "anchor" signal that established a comparison point at an easily understandable and reproducible quality level. The methodology allows for multiple quality-impaired stimuli at the same time, and thus generally needs fewer participants to obtain statistically significant results than the more speech-focused Mean Opinion Score (MOS) methodology [137]. MOS methods are popular in speech methodology [137] and have been adopted for the field of generative audio in general [2]. Generally, MOS survey methodology is considered flexible, given that it is not always necessary to provide a reference signal. For instance, the Absolute Category Rating specification of the

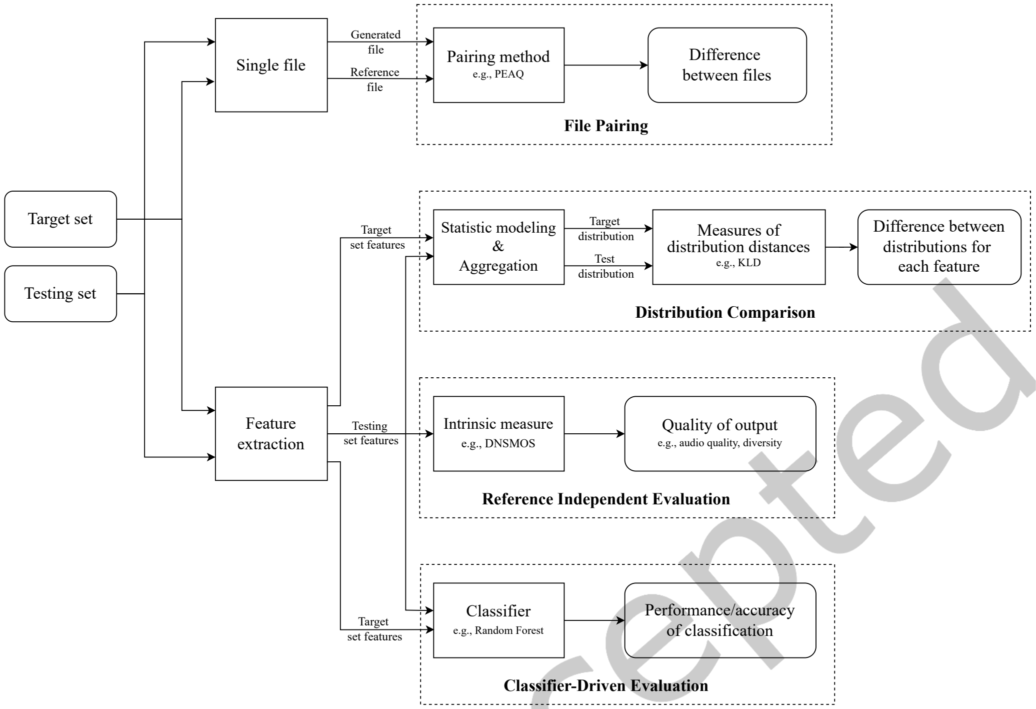

Fig. 2. Overview of the evaluation methodologies for generative music systems, including both reference-dependent and reference-independent approaches.

ITU-T P.800 standard [137] does not require a reference signal (as opposed to the Comparison Category Rating specification). The scale for rating can also vary between a five point scale and a seven point scale.

3.2 Objective evaluation

Aesthetics and the impression of artistic quality are inherently subjective, and thus hard —or even impossible— to approximate objectively. On the other hand, it has been argued that subjective results might not be trustworthy and that we must look beyond "human opinion for evaluation of computational creativity" [200]. Whichever side is taken, an evaluation based on computed objective metrics is not meaningless. For instance, Yang and Lerch [267] show that contemporary generative models can fail to properly model the distribution of even low-level musical properties such as pitch range and count, a result that has been substantiated for an extended set of properties by Banar and Colton [10]. Thus, metrics for such musical qualities can be helpful in analyzing how well the statistics of the generated output match the statistics of the training data. Furthermore, methodologically sound and properly executed listening studies cannot easily take place after each development iteration to measure progress. Taking into account the other advantages of objective metrics such as perfect reproducibility, objectivity, and scalability to large amounts of data, there is a need for objective measures to assess the output of generative systems. There is, however, a risk involved in applying objective metrics to music; Goguen [103] point out that "a common approach is to 'bracket' or exile the qualitative aspects, and concentrate attention on aspects that are reducible to scientific analysis."

3.2.1 Methodology. The approaches to the objective evaluation of music generation systems are diverse, as depicted in Fig. 2. Objective evaluation methodology for generative systems typically involves either estimating a score that reflects a specific quality characteristic or using algorithms to compute differences between generated outputs and target benchmarks. In both symbolic and audio generation, the methodologies can be broadly categorized into several distinct types, each serving a different purpose in the evaluation framework. These categories include: (i) the reconstruction error or match to reference signal, (ii) a comparison of distributions, (iii) reference-agnostic tests, and (iv) classifier-driven evaluation.

File pairing & reconstruction. This method compares the generated output directly to a matching reference. Typical use cases are masked music generation (also referred to as music in-painting) [205] or measuring the reconstruction error of auto-encoder setups. Thus, this approach allows for direct comparison between the generated output and a known reference but only under the limiting assumption that there is *exactly one* correct output. While this assumption allows for a mostly straight-forward assessment approach given a metric (e.g., the Mean Squared Error (MSE) or an edit distances, the Signal-to-Noise Ratio (SNR), or psycho-acoustically motivated metrics such as PEAQ [134]), it is not suited to many typical evaluation scenarios for generative systems where a unique correct output does not exist.

Comparison of distributions. An alternative to comparing files by pairs is to compute distributions of certain characteristics or descriptors across many files to compare a testing (generated) and target (training) set. It is known that generative models are primarily trained to learn characteristics over a training set and generate output replicating those characteristics. Thus, the expectation is for the test distribution to be similar to the target distribution of various descriptors. The target distribution is typically human-composed music.

In statistics, a popular distribution distance used to compute differences between distributions is the Kullback-Leibler Divergence (KL-Divergence), which computes the difference between two distributions over the same sample space. The distance measurement is non-symmetric, i.e., the distance from A to B does not generally equal the distance from B to A. In symbolic music evaluation, statistical measures had already been proposed by Collins [58]. The KL-divergence is, e.g., used by Yang and Lerch [267] to evaluate the difference between two distributions in various descriptor dimensions. In addition to KL-divergence, they propose the usage of "Overlapping Area" as an indicator of distribution similarity that is —unlike the KL-Divergence— both symmetric and bounded. Another commonly used distribution distance is the Wasserstein distance between two multi-variate Gaussians. Recent work by Guo et al. [111] proposes calculating a measure of statistical significance between the training and generated distribution. All these distance metrics can be calculated from a variety of descriptors, ranging from basic synbolic score features such as pitch histograms to learned embeddings.

While a high similarity of distributions for various descriptors is desirable, it does not necessarily allow for a conclusive assessment. On the one hand, it depends on how musically meaningful the descriptor itself is. On the other hand, a system could achieve perfect scores by simply replicating the training set, thus leading to favoring conformity over novelty [32, 123].

Measures without reference. This method involves evaluating the quality of generated music without comparing it to a reference or ground truth. Instead, various metrics are employed to assess aspects such as diversity, novelty, coherence, or aesthetic appeal within the generated set itself, thus providing intrinsic (i.e., non-intrusive) evaluations of the generated music, independent of any external reference. In many cases, these approaches either measure relative attributes changes of the output without clear anchor point (e.g., increasing novelty beyond a certain point will decrease output quality) or by establishing a framework of clearly defined quality standards that are generally true (e.g., intelligibility for speech signals). For example, Chu et al. [55] evaluated repetitiveness within the same corpus of 100 generated songs by segmenting each generated melody into two-bar units and then comparing these segments without the use of an external reference set. Complementing these intrinsic evaluations, Yuan et al. [271] developed MusicTheoryBench, a set of college-level music theory and composition questions designed to test LLM-based symbolic generation systems' music knowledge and music reasoning.

When evaluating speech quality, no-reference methods are usually preferred. For instance, Manocha et al. [181] find that methods assessing *audio similarity* are not optimal surrogates for assessing speech quality and recommend no-reference methods instead. These methods attempt to estimate audio quality for a signal directly by predicting how humans might rate a signal.

Classification-driven evaluation. A number of evaluation strategies have been proposed to evaluate the quality of generated output through a classifier, i.e., a machine learning system categorizing input into pre-defined classes. On the one hand, a classification model can be trained to distinguish between human-composed and generated music. The performance of the classifier in accurately distinguishing between real and generated music serves as an (inverse) measure of the quality of the generated compositions. On the other hand, classifier performance can be used as a proxy quality measure in specific domains. A genre classifier, for instance, can be used to assess whether the generated music adheres to the desired style [34, 140] or is similar to the target style [83]. In generative audio tasks, classifiers are directly used to measure whether the generated signals are capable of matching the labels assigned to signals in the training set. In the case of NSynth [82], they use a timbral classifier to evaluate whether the generated signals could be trivially classified as being one of the timbre labels in the dataset. To augment these approaches, Godwin et al. [102] introduced an innovative variation by incorporating generated samples into the training dataset for the classifier and investigated how such augmentation affects the classifier's performance, shifting the focus from merely comparing the classification of generated outputs with those of the training data to a more dynamic assessment with adaptability.

Classifiers can also be used as the basis for the Inception Score metric [192, 225], which is used to measure "sample diversity," i.e., if the model is capable of generating signals that match the label distribution of the testing dataset. Banar and Colton [9] review classification-based methods and argue that while these methods might be useful for post-hoc quality evaluation, the embedding spaces utilized by the classifiers are not optimized in terms of the distances between different classes. This lack of optimization can result in suboptimal performance of these models in distinguishing various types of musical content and is of particular concern if the embedding space is used for a distance-based metric.

3.2.2 Aesthetic and musical qualities.

As pointed out above, objectively assessing *aesthetic quality* presents many challenges due to its subjective nature, especially the challenge of how any objective measure could directly reflect aesthetic quality. Juslin et al. [144] define aesthetic judgment as the assessment process through which the value of a piece of music as "art" is determined based on subjective criteria such as novelty, expressivity, and beauty, which relate to both the form and content of the artwork. They state that aesthetic judgments result from psycho-physical interactions between the music's objective properties and the subjective impressions of the evaluator. Therefore, there are no absolute or universal criteria for aesthetic value, as aesthetic norms are subject to change over time within society. Kalonaris and Jordanous [145] substantiated this perspective by noting that many experiments and theories on musical aesthetics —such as concepts of beauty, pleasantness, and well-formedness— are heavily reliant on context-specific and often arbitrary assumptions about the nature of music, and are thus hard to generalize. Despite this known variability, attempts towards developing objective evaluation metrics of aesthetic quality have been made, drawing on both traditional music theory and empirical aesthetics principles.

Kalonaris and Jordanous [145] categorize the measures of computational aesthetics in music into several distinct types, each grounded in different theoretical frameworks. These include (i) information and complexity-based aesthetic measures, such as entropy, which quantify the order and predictability within musical compositions, (ii) geometric measures, which assess aesthetic value by analyzing the statistical distribution of musical elements, (iii) psychological measures, such as Gestalt principles of grouping, to understand how listeners perceive and interpret musical structures as a whole rather than merely sums of parts, and (iv) biologically inspired measures, which explore the social, genetic, and evolutionary algorithms.

Notably, evidence has been presented that many musical parameters (e.g., pitch, dynamics) follow a Zipfian (power law) distribution and that adherence in musical composition to such distributions is more aesthetically pleasing [97]. Zipf's law [276] suggests that the frequency of a symbol within a piece of music should be inversely related to its rank in terms of frequency of use. Computational aesthetic measures relying on predictions from Zipfian distributions have been successfully used to predict human aesthetic ratings such as "pleasantness" for both music and visual art [97, 178, 179]. It should be noted that the concepts of musical "pleasantness" and overall aesthetics or "appreciation" [208] might not be necessarily synonymous, as the latter implies a broader range of emotional and intellectual responses. Another automated computational approach reviewed by Galanter [97] include using various simulation models based on chromosome behavior that compute a single weighted metric, an "evolutionary fitness score," that rewards some desired behavior or trait.

Berlyne's theory of "New Empirical Aesthetics" [19] provides another theoretical framework for understanding aesthetic quality. According to this theory, the appreciation of an artwork correlates with its complexity and its ability to stimulate arousal, with the audience's liking having an inverted-U relationship with "arousal potential." In a practical setting, the measurement of subjective aesthetic quality often includes beauty, groove, originality, complexity, expression, emotion, sound quality, prototypicality, message, and skill [143].

Metrics targeting *musical qualities* are commonly distribution-based, meaning that the similarity between a generated and a target distribution is measured for a specific musical descriptor or property. Typically, these descriptors are low-level representations of music [139] and some of these features and distributions might be more musically relevant than others. For example, pitch distributions are commonly applied to MIDI numbers or pitch values (e.g., A4) as opposed to much more meaningful pitch or chord distributions that take key information into account, such as scale degree (the position of a pitch class in a tonal context) or Roman numeral (the function of a chord given its context). Still, basic features such as pitch difference and the ratio of in-key pitches or entropy of pitch and chords [260] continue to be employed in recent works and are sometimes referred to as "music theory evaluation" [111]. Over time, many descriptors for musical qualities have been introduced [185]; Yang and Lerch [267] proposed a set of simple metrics separately targeting tonal and rhythmic content. Similarly, Garcia-Valencia et al. [98] utilized the concepts proposed by Tymoczko [252] to assess qualities like the smoothness of melody transitions, pitch diversity, and the occurrence of local notes to provide an analysis of melodic structure with rhythmic features. Dervakos et al. [68] introduced a framework based on the basic consonance aspects of melodies that allows the construction of a variety of metrics. They used this framework to construct four heuristic properties, polyphonicity, used-pitch-classes per bar, total number of pitches, and the total number of pitch classes, to address fundamental musical properties that are reliable and interpretable. These examples show the variety of descriptors that have been used to model musical properties. Table 1 provides an overview of common descriptors proposed in the literature.

While there have been many calls for more musically meaningful and relevant objective metrics, many of these descriptors and resulting metrics are highly specialized to certain styles, genres, or musical content. For example, Wu and Yang [263] employed specific metrics such as grooving and chord progression to evaluate generated jazz performances. They observed that erratic usage of pitch classes, inconsistent grooving pattern and chord progression, and the absence of repetitive structures contribute to the quality gap between the generated and human-composed jazz samples. Although these metrics appear to be pertinent to jazz, they might not fully capture or evaluate the qualities of music produced in other genres. This specialization of metrics may restrict their broader application across different musical styles and genres and implies that other musical styles might require the design of style-specific descriptors as well.

Furthermore, scholars have pointed out the importance of higher level features such as form and repetition in contributing to the improvement of generative music systems (e.g., [63]). However, specific metrics to measure

Table 1. Previously used low-level descriptors used to describe musical qualities.

Type	Unit	Name	Description
Pitch [73, 139, 186, 267]	Note	Total Used Pitches, Pitch Class Histograms	number of distinct pitches or pitch classes (often accompanied by entropy)
		In-Key Note Frequency, Scale consistency	fraction of notes adhering to a scale
		Pitch Range, Tone Span	span between the highest and lowest pitches (usually in semitones)
		Consecutive Pitch Repetition	frequency of repetitions of a pitch
		Pitch Interval Average, N-gram	interval between consecutive notes (usually in semitones)
		Pitch Class Transition	frequency and type of transition between pitch classes
	Sequence	Empty Bars	number or ratio of bars w/o musical onsets
		Sequence Repetition	number of repetitions of short sequences
		Rote Memorization Frequency	frequency of which the generated sample reproduces exact sequences from the training corpus
		Frequency of Pitch Change	how often pitches change
		Pitch Variations	number of distinct pitches
		Voice Motion	how voices or melodic lines move relative to each other
Rhythm [139, 186, 263, 267]	Note	Total Used Notes	total number of notes
		Note Length Histogram	distribution of note lengths
		Qualified Note Length Frequency	frequency of note durations
		Average Inter-onset Interval	average time between the onset of consecutive notes
		Note Length Transition	frequency and type of transition between consecutive note lengths
	Sequence	Rhythmic Similarity between Measures	rhythmic similarity between different measures
		Rhythm Variations	number of distinct note durations
		Off-beat Recovery Frequency	how frequently the model can recover back onto the beat after being forced to be off beat
Harmony [83, 139, 263]	Chord	Chord Duration	length of time a chord is held
		Chord Content	pitch composition of a chord
		Chord Vocabulary	e.g., number of used chords, chord coverage, number of repeated chords
		Tonal Distance Average, Histogram	average tonal distance between pairs of adjacent chords
		Chord Tone to Non-Chord Tone Ratio	in-chord vs. non-chord tones
	Overall	Melody-Chord Tonal Distance	average tonal distance between each melody note and its corresponding chord
		Progression Irregularity	degree of difference in chord progressions between a sample and templates
		Polyphonicity	frequency of simultaneously played pitches
		Dissonance	dissonance level of onsets based on their periodicity

these higher-level features have not yet been proposed. Instead, metrics such as the Fréchet Distance (see Sect. 3.2.3) utilizing trained embeddings with opaque or unknown musical meaning (VGGish [124], CLAP [264], etc.) have been increasingly applied to music evaluation [56, 109, 130, 218].

3.2.3 Audio quality. For any generative audio system, assessing the audio quality of the outputs is integral to understand whether the generated signals can be considered high-quality by human listeners. Simple error measurements such as the MSE, the Mean Absolute Error (MAE), or the SNR have been long shown to be ineffective as "quality" measurements [147], so that listening studies are often considered to be the ultimate way to assess audio quality. Regardless, objective metrics for measuring audio quality have been proposed.

In many cases, where generative systems are trained to synthesize and match the inputs as closely as possible, differences in the presence of artifacts, distortion, noise and bandwidth are measured. This usually assumes that the cleanest reproduction of the signal is free of artifacts, distortion and noise while maintaining an identical frequency bandwidth to the original signal. If the reference signal is available, it is feasible to use metrics that rely on file pairing. Popular reference-based objective methods are (1) *Perceptual Evaluation of Audio Quality* (PEAQ) [134], a method that uses a psycho-acoustic model to compute features representing the difference between a reference and a test signal, (2) *Signal-to-Noise Ratio* (SNR), *Signal-to-Distortion* Ratio (SDR) and *Signal-to-Artifact* Ratio (SAR), which compute the difference in noise levels, distortion levels, and artifact levels between paired signals, respectively, and, (3) Deep Perceptual Audio Metric (DPAM) [180], a metric that is trained to estimate perceptual quality similarity between paired signals.

Many of the contemporary approaches to the problem use the aforementioned distribution comparison metrics. By and large, these methods extract neural representations or embeddings for both the target and testing sets. The primary argument for the usage of distribution comparison methods hinges on the notion that minimal divergence between the two distributions is indicative of the two sets being roughly identical in terms of quality. The most popular distribution divergence metric used in evaluating generative quality is the Fréchet Audio Distance (FAD) [150], which computes the difference between the two embedding distributions using a Wasserstein distance on VGGish embeddings [124], and more recently on a variety of other embeddings, such as PANNs [152], CLAP [264], and Encodec [77] (cf. [108, 109]). The perceptual relevance of FAD has, however, been questioned for both audio quality evaluation [255] and for music evaluation [130]. Gui et al. [109] propose a variation of the FAD for audio-based music quality measurement called FAD$^\infty$. FAD$^\infty$ is an extension of FAD that mitigates the sample-size bias by estimating the behavior of the metric as if it had an infinite number of samples by using Quasi Monte-Carlo integrals. Gui et al. [109] report that —given the right embeddings and mitigating sample size bias— FAD$^\infty$ can be a useful indicator of quality, a result that is intended to assuage concerns raised by previous results [255]. An alternative to FAD is to use Maximum-Mean Discrepancy (MMD) [21, 192], which is a two-sample test computed over the kernel embedding of the training and test data. For more details on the mechanics of MMD, we refer the reader to Gretton et al. [107]. Recent work in the image domain has shown that MMD with CLIP embeddings is a better approximator of generative image quality than Fréchet distances [138]. Chung et al. [56] propose to adapt the MMD to the Kernel Audio Distance (KAD) to replace FAD.

Distribution comparison methods measure if the generator is capable of modeling the underlying distribution that it was trained to generate. However, the perceptual relevance of the result largely depends on the selected feature / embedding space [7, 255]. Unbounded, distance-based metrics also lack interpretability in the sense that even a statistically significant difference in a metric does not necessarily imply a perceptually significant difference.

Other notable approaches to estimating audio quality include DNSMOS [217] NDB/k [221] and ViSQOL [126], which are predominantly speech quality focused approaches. ViSQOL and DNSMOS are speech quality estimation methods trained using results from large-scale MOS listening studies. ViSQOL is a paired metric that uses spectrogram patches and DNSMOS is a no-reference metric that uses a model trained to predict what a human might rate a signal. NDB/k is a metric originally proposed for images that has been used to evaluate some generative systems such as Diffwave [153]. It clusters the features of the training set using k-means clustering and Voronoi cell partitioning. To compute the metric, the number of statistically different bins or cells are computed. Recent work has shown that in speech, similarity is not a reliable proxy for quality and no-reference metrics, such as the ones mentioned above, might be better at estimating quality [180].

Table 2 shows an overview of commonly used objective metrics for audio quality assessment.

3.2.4 *Originality.*

Originality is often considered a key characteristic of a generative system, however, only few quantitative measures have been proposed despite calls for work on, e.g., quantification of output diversity [84]. As mentioned above, we understand originality to be comprised of *diversity*, *novelty / plagiarism*, and *creativity*.

Table 2. Commonly used metrics and distances for evaluation of audio quality. In the range column, ↓ indicates that lower is better and vice versa.

Type	Metric	Domain	Range
Paired reference & test	PEAQ	psycho-acoustic model	-4 to 0 ↑
	Mean Squared Error	waveform or spectral	0 to ∞ ↓
	Mean Absolute Error	waveform or spectral	0 to ∞ ↓
	DPAM	trained from listening study	0 to ∞ ↓
	SI-SDR	waveform	0 to ∞ ↓
	SNR	waveform	0 to ∞ ↓
	ViSQOL	trained from listening study	1 to 5 ↓
Ref. & test distribution	NDB/k	Deep Learning Images	0 to ∞ ↓
	Fréchet Audio Distance	VGGish, CLAP, etc.	0 to ∞ ↓
	Kernel Distance	trained from labels	0 to ∞ ↓
	Inception Scores	trained from labels	0 to ∞ ↑
No reference	DNSMOS	trained from listening study	0 to 5 ↑

While *diversity* is implicitly evaluated by commonly used metrics for image generation such as the Inception Score and the Fréchet Inception Distance, they do not differentiate the diversity aspect from the quality/fidelity aspect. Other quantifiable metrics on diversity have been pre-dominantly proposed for GAN-based generative systems [26]. However, as Gulrajani et al. [110] point out, many of the existing metrics can be tricked by the model simply memorizing the training data. Thus, they propose so-called neural network divergences (NNDs), measured through the loss of a neural network trained to distinguish between generated and training data, and show that NNDs can serve as meaningful measures of diversity. Alaa et al. [4] propose to measure diversity using a metric they refer to as β-recall, measuring the fraction of real samples covered by the most typical generated samples. In the context of music, Yin et al. [269] propose the "Originality Score" to quantify how original a set of musical pieces is and compare the originality scores between training and generated data to understand if the originality of the generated data matches expectations.

Sturm et al. [242] note that for music —unlike for text— automated *plagiarism* detection does not reliably work, and worse, no clear standards exist about what type and amount of alterations make a piece of music novel as opposed to plagiarized. Thus, approaches such as the Authenticity metric proposed by Alaa et al. [4], estimating the probability of a generated sample being copied from the training data, are only of limited use. Still, approaches to objectively measure plagiarism have been proposed, usually framed as a music similarity task. Most of the seminal work can be grouped into (i) melodic similarity measures based on symbolic input and (ii) general audio-based similarity measures. To measure *symbolic melodic similarity*, a variety of metrics have been proposed. Most of them are based on some form of sequence similarity [47, 65, 66, 188, 201, 262] or vector-based similarity measures hand-designed [177] or trained as end-to-end systems [173, 202]. Some of these algorithms have been tested against real-world court decisions, however, the test set sizes are necessarily small and external validity is hard to verify. *Audio-based similarity* [161] is a multi-dimensional problem at the risk of confounding dimensions of score similarity (melody, harmony, rhythm, etc.) with performance similarity (tempo, playing techniques, etc., but particularly timbre). Systems for audio-based plagiarism detection have been proposed to use Non-Negative Matrix Factorization to decompose the audio [64], MFCC vector-based representations of audio [244], similarity measures inspired by audio fingerprinting approaches [27, 175], or utilizing pre-trained embeddings for similarity measurement [15]. Commonly, however, plagiarism is understood in terms of score similarity (with the very prominent exception of sampling, where audio is copied and mixed into a new musical artifact [113]). A study into the contributing factors of court cases on plagiarism has been presented by Yuan et al. [272].

The evaluation of *creativity* poses an unsolved problem. Jordanous [142] notes the longstanding lack of attention to evaluation and lack of evaluation standards in the computational creativity community due to “difficulties in defining what it means for a computer to be creative; indeed, there is no consensus on this for human creativity, let alone its computational equivalent.” Indeed, in a survey of 75 papers on computational creativity, the author found that in one third of the papers, ‘creativity’ was not mentioned, and that only one third of the papers actually attempted to evaluate the creativity of their systems. Moreover, “[o]ccurrences of creativity evaluation by people outside the system implementation team were rare.” Judging from the scarcity of proposed evaluation standards since the publication of this paper —despite the fact that it has been over a decade since its publication and that creative AI is presently in its heyday— it seems that this issue continues to persist. As mentioned above, one can distinguish between the evaluation of a computational system itself, and such a system’s output (i.e., “the product/process debate” [142]). However, since the output is restricted or constrained by the system, it can be helpful to conceptualize a combined methodology for the evaluation of both the system (in terms of its creative potential) and its output. Jordanous [142] presents such a framework, referred to as “SPECS” (Standardised Procedure for Evaluating Creative Systems). The SPECS method —which is not only well recommended [200], but unlike other proposed models has actually been used in practice [197, 211]— asks the creator/evaluator to adhere to a specific working definition of creativity that includes two or more “components” that are evaluated independently according to some standard, such as skill, novelty, value, etc., and compared against an appropriate system/output. These evaluations may be quantitative or qualitative, however, in this case, both still involve significant investment of time and resources into evaluation with human subjects. While other models and frameworks have been proposed for attempting to theoretically quantitatively assess creative output, such at the FACE and IDEA models [209], to our knowledge, no adopted standardized practices or procedures currently exist for assessing creativity without human intervention.

4 Evaluation of usability & user experience

In this section we describe Human-Computer Interaction (HCI) methods and techniques that have been used to evaluate user interaction with generative music models. These methods and techniques share aims with Human-Centered AI [198, 231] to research, design, and evaluate AI systems from a human-centered or user-first perspective. Below, we first introduce HCI methodologies for evaluating AI music systems. We then describe data collection methods for understanding the user experience of generative AI including quantitative and qualitative approaches. Table 3 closes this section by summarizing key HCI evaluation approaches discussed. It is important to note that there are no de-facto standards defining which data collection methods are used for which evaluation methodologies. Instead, selection of data collection and methodology is based on best practice in the field and individual HCI practitioner’s skills and experience.

4.1 Methodology

HCI research has traditionally focused on functional aspects of user interaction such as the usability of a system [190], whilst later waves of HCI placed more focus on subjective qualities of users’ experience when interacting with computers [40]. Approaches to HCI evaluation of generative music systems draw on both functional (usability) and experiential (user experience) forms of HCI evaluation. Broadly, evaluation methods and can be split into controlled experiments which are more objective and typically suited to exploring the usability of a system, and more subjective and ecologically valid (meaning applicable to real-world practice) approaches which are more suited to studying experiential aspects of generative music systems [38].

4.1.1 Controlled experiments. Controlled experiments place participants in distraction free environments such as a research lab, where they are set a number of musical tasks to complete with an AI system in a constrained amount of time. In this setting, typically two or more versions of an AI system are used to allow comparison between versions, often referred to as A-B testing. The tasks undertaken might be open-ended (e.g., write a piece of music) or more specific (e.g., harmonize a given melody) depending on the model and the evaluation goals. For example, Suh et al. [243] and Louie et al. [171] asked participants to compose music for a fictional character from a game, while Frid et al. [96] asked participants to create music for a video based on an example song. In Louie et al. [171], two versions of a music making interface for harmonizing melodies were tested —one with and one without AI-steering tools— to allow for comparison of the effect of the AI. In this case the tool without AI features can be referred to as a baseline. Alternatively, different forms of interaction and participation might be tested with the same system, e.g., Addessi et al. [1] compared their AI music improvisation system when used by individual children vs. groups of children.

4.1.2 Online evaluation. Online evaluation settings can be helpful for evaluating generative AI systems across a large sample size. For example, Ben-Tal et al. [17] evaluated a version of the *FolkRNN* generative AI model hosted online, where they were able to examine how people generated content with FolkRNN serendipitously and how they tweaked values to modify and curate outputs. Typically, generative AI systems are either hosted online for user interaction or available as a download to users. Audiences across the globe can then be reached using survey platforms such as Prolific,[2] and filtered for characteristics such as nationality or technical skills. Whilst online settings allow for large numbers of participants, they lack the rigor of controlled experiments.

4.1.3 Exploratory studies. Exploratory studies emphasize evaluating user experience and collecting subjective feedback. They typically involve more open-ended tasks than controlled experiments or online evaluations. Exploratory studies can, for example, take place in a controlled lab setting which is more typical of comparative studies, yet use open-ended tasks that allow music making to occur in a more natural way. For example, Bougueng Tchemeube et al. [28] tasked people with exploring their generative AI interface and collected structured questionnaire measures to capture aspects of people's user experience, but do not make strict comparisons between interface designs. Thus, balance can be struck between controlled and exploratory evaluations, such as giving open-ended tasks in controlled study settings to test generative AI in a way that is closer to real-world music making (e.g., [171]), or to give structure to data collection in real-world settings (e.g., [93]).

4.1.4 'In-the-wild' studies. Research-in-the-wild [18, 49] approaches contrast controlled experiments to evaluate generative AI models in their real-world places of use, possibly over extended periods of time. For example, in ethnographic approaches [18, 49] the researcher takes observations or field notes, or collects data on patterns of behavior that people have naturally exhibited while making music. For AI music, this type of approach has been used in, e.g., the international AI music song contest [127]. The researchers identified how developers and musicians collaborated to create music , for example by preferring to curate AI generated content instead of (re-)developing their AI tools. Across the HCI studies on generative music, there are several examples of ethnographic-inspired observations being collected. Fiebrink et al. [90], for instance, "recorded text minutes of [composer's] activities, discussion topics, and specific questions, problem reports, and feature requests" for seminars on their Wekinator [88] system. Bryan-Kinns et al. [37] used first-person accounts of music making with a generative AI system over several months to understand how it was appropriated into music making practice.

4.2 User data collection: Quantitative

The primary quantitative method for evaluating generative AI systems is the use of *questionnaires*. These are used to quantify both subjective feelings of a system's usability as well as more experiential aspects. Typically, these questionnaires use a Likert-type scale [166] to measure user agreement with statements on a scale (e.g., 1–5).

[2] https://www.prolific.com/, last accessed: Jun 25, 2024

Several standard questionnaire measures exist to evaluate the usability of technology. Common examples are the NASA Task Load Index [117] and the Standard System Usability Scale [12]. For AI music interfaces, we found several evaluations (e.g., [132]) using the Cognitive Dimensions of Notations questionnaire [23], which assess several cognitive qualities of the interface such as whether the system has many hidden elements of represents information in a diffuse way.

User experience-oriented scales include the User Experience Questionnaire [158] and the User Engagement Scale [199]. The most influential questionnaire with respect to creative technology within the last 10 years is arguably the Creativity Support Index (CSI) [52]: a questionnaire designed to test the capacity of a tool to support creativity, offering factors for several important aspects of the creative user experience including: focused attention, enjoyment and collaboration. For AI music, however, we found surprisingly few examples of AI music user studies that have adopted the CSI — Bougueng Tchemeube et al. [28] a notable exception. We instead observe many examples where researchers have chosen to define their own questionnaires to explore constructs for the AI systems that are not captured in current standardized scales. For instance, Louie et al. [171] invented their own questions to capture a person's feelings of agency. Ford and Bryan-Kinns [92] identified reflection as an aspect missing from the CSI whilst being an important factor in AI music making [91, 93]. Alongside established measures of creativity support and usability, Bougueng Tchemeube et al. [28] also added questions central to human-AI interaction on trust, perceived authorship, and flexibility. This use of researcher defined questionnaires reflects the dominance of measures of engagement and usability in assessing AI interaction in creativity-related HCI research [219, 220].

Questionnaires are more often used in usability-focused evaluation methodologies such as controlled experiments and less frequently used as part of more experiential evaluations such as in-the-wild studies. HCI researchers also use questionnaires to establish characteristics of their sample under study. For example, the Goldsmith's Musical Sophistication Index [187] offers a standardized metric for musical expertise, helping to identify whether the users of a generative AI system under study have above or below average musical skills.

4.3 User data collection: Qualitative

In addition to quantitative data collection researchers use qualitative data collection to gain greater insight into users' feelings, motivations, and reflections when using a generative AI system. It is important to note that triangulation across different data collection approaches (also referred to as mixed methods) is crucial — using both qualitative and quantitative data can help to demonstrate which features of the user experience are improved by AI or not, as well as offering insights into why this might be so [38].

4.3.1 Interviews. Interviews are a common technique used in HCI, often to give insights into users' thoughts and feelings on their interaction. They can be structured, semi-structured or fully open-ended [30]. For generative music user studies, we found that semi-structured and unstructured approaches were common, with workshops or group interviews used for need-finding studies [89, 93, 96, 171, 243]. There is no standard set of questions used in interviews for evaluating interaction with generative AI models, nor standard analysis approaches. Results tended to be reported thematically following a process such as Thematic Analysis [31] or using more experimental approaches as in Fiebrink and Sonami [89], who transcribed their interviews verbatim when reflecting on their extensive experience on AI music. Generative AI studies in music are yet to explore qualitative analysis methods emerging in more modern HCI paradigms [95], e.g., [213, 229], which might capture qualities from interviews that Thematic Analysis does not.

4.3.2 Think-aloud. Several studies [96, 132, 171, 172, 243] have applied the HCI "think-aloud" method to gain insight into how users interact when making music with a generative AI model. In the think-aloud method participants are asked to describe their thought process while performing their task, e.g., while making music

Table 3. Overview of Human-Computer Interaction evaluation approaches.

Quantitative Methods: Questionnaires			
	Acronym	**Captures**	**Reference**
Goldsmith's Musical Sophistication Index	GMSI	Musical expertise	[187]
NASA Task Load Index	NASA TLX	How complex a task is perceived to be	[117]
Standard System Usability Scale	SUS	The usability of a user interface	[12]
Cognitive Dimensions of Notations questionnaire	CD	The usability of a user interface	[23]
User Experience Questionnaire	UEQ	The usability and experiential aspects of a user interface	[158]
User Engagement Scale	UES	The experiential aspects of a user interface	[199]
Creativity Support Index	CSI	How well a user interface supports creative work	[52]
Reflection in Creative Experience	RiCE	Types of reflection in creative contexts	[92]
Qualitative Methods			
	Examples	**Captures**	**Reference**
Interviews: Structured, semi-structured, or open-ended	[89, 93, 96, 171, 243]	Insights into users' thoughts, motivations, and feelings on their interaction	[30]
Think-aloud	[96, 132, 171, 172, 243]	Users' thought processes whilst using a user interface	[212]
Video-cued recall	[41]	Users' post-hoc thoughts about using a user interface	[5]
Autoethnographies	[193, 238]	A researcher's subjective and personal reflections on their musical practice and use of technology	[216]
First-person accounts	[93, 240]	Rich descriptions of users' personal reflections on using AI models	[174]

with an AI tool [212]. Whilst this method can give detailed insight into participant's cognitive process, it can distract users, meaning that certain aspects of the creative user experience such as flow states cannot then be investigated [62]. It is also impractical for certain music practices such as live improvisation. An alternative approach is to perform the think-aloud retrospectively [5] with participants describing a recording of their composition practice (sometimes referred to as video-cued recall [41]) — this approach is under-utilized in the literature for generative music.

4.3.3 First-person perspectives. The use of methods such as questionnaires and interviews described above is borne from a psychology-driven epistemological stance: to identify generalizable models of how people interact with technology. Approaches inspired by Arts and Humanities offer insights into the individuality and subjectivity of how artists have interpreted their use of technology [42]. We identified an increasing trend to report on the use of AI from a first-person perspective [17, 37, 93, 193, 238, 240], publishing perspectives on how individuals have been able to use and incorporate models into their music-making. In some cases, these are autoethnographies [193, 238] where a researcher reflects on their own practice by means of capturing data over a long time period. Other examples show collections of first-person accounts [93, 240]. Sturm et al. [241]'s proposal for a field of AI music studies engages with these methods to explore ways to more meaningfully and critically engage with the broader communities in social sciences and humanities. In contrast to questionnaires, first-person perspectives are more frequently used in user experience-focused evaluations such as in-the-wild studies and less frequently, if at all, in controlled experiments.

4.4 Other HCI evaluation approaches

There is a wide variety of other evaluation methods used in HCI beyond those we found are most often used when evaluating AI music systems, as described in this section. Usability metrics such as task completion rates, task time or the number of errors, have been explored to evaluate the usability of a computer music system [257]. However, these are not prominent in generative AI music user studies where music making tends to have no clearly defined goal. "Wizard of Oz" was an early popular approach to user studies of human-AI interaction more broadly, where users interact with a user interface whilst a researcher provides feedback through the interface in lieu of an AI. For example, Thelle and Fiebrink [251] tested participant's reactions to researchers who performed piano phrases live, acting as an AI system. Similar examples have been tested with more complex generative music programming languages [16]. We also did not find many examples of using physiological measures such as heart rate or eye-tracking to explore people's interaction with music AI tools. In other arts-based HCI studies, physiological measures have been used as proxies of aspects of the user experience such as anxiety or boredom, indicated by participants' heart rate or skin conductance [176]. This could be an open area for further research.

5 Challenges and future work

Despite a large number of previously proposed approaches, a multitude of challenges remain unsolved in the evaluation of generative systems in music. Given the nature of the task, it is unclear if generalizable satisfactory solutions can ever be found for some of these challenges.

5.1 Validity

Although subjective evaluation is often considered the most meaningful way of evaluating the output of generative systems, the results cannot be automatically assumed to be robust or reliable. Subjective evaluation of system output (Sect. 3.1) usually relies on survey approaches [120], much as quantitative user data collection relies on questionnaires (Sect. 4.2). However, designing surveys and questionnaires is non-trivial — the creation of batteries and psychometrics form an entire subfield of psychology [101, 226]. Moreover, existing questionnaires often fail to capture experiential aspects of human-AI interaction, such as users' sense of agency. A considerable number of researchers working on generative music systems lack the background, skill and/or resources to successfully carry out such surveys with valid, reliable, and replicable results [267].

In addition, there is a known bias in people's perception against AI-generated music [230]. As such, tests highlighting human vs. machine authorship may introduce bias in the results. With respect to Turing tests, Hernández-Orallo [121] point out that a known validity issue with Turing tests is that the outcomes cannot disambiguate whether the model was a good imitator, or the human was a poor judge.

Given these potential validity issues with subjective evaluation of system output, objective evaluation remains a viable choice to complement listening studies. With objective evaluation, however, there are other validity concerns. These concerns often start with the data and its characteristics: Is the sample size sufficient? Do the data reflect the targeted homogeneity or heterogeneity? Are there confounding characteristics in the data that complicate drawing conclusions? Another concern is the validity of the chosen metrics — do they meaningfully represent the evaluation target, and are observed differences in evaluation results perceptually significant? Even if some metrics prove to be relevant, individual metrics or criteria provide sufficient breadth for comprehensive evaluation; Theis et al. [250] rightly note that "Good performance with respect to one criterion (...) need not imply good performance with respect to the other criteria." Note that even when targeting very specific criteria (e.g., complexity), subjective impressions might be better predictors of responses than objective measures [116]. For that reason, it might make more sense to use subjective impressions of criteria (rated by the listener) as predictors of the overall judgments of aesthetic value [144].

For evaluating user interaction with generative systems using HCI approaches, the main challenge is to balance the ecological validity of the evaluation (how realistic is the study setting) and the generalizability of the results of the study. For example, “in-the-wild” studies give in-depth insights into how generative music is used in real-world places of music making such as performances on stage or music making at home, but the findings are tied to the study’s cultural context and individual musicians making it hard to generalize from the results. Certain protocols such as “think-aloud” also affect ecological validity because subjects tend not to speak aloud about their thought processes when making music. Likewise, the “Wizard of Oz” protocol has poor ecological validity as understanding how people interact with pretend AI tools can be different to how these systems work once actually deployed.

Thus, both internal and external validity remain core challenges of evaluating generative music systems. This is true for both subjective and objective approaches.

5.2 Perceptual and musical relevance of objective metrics

Revisiting the objective metrics introduced in Sect. 3, it can be observed that the most common metrics compare training data characteristics with characteristics of the generated data in one way or the other. A major differentiation between such metrics is the space and the dimensions in which different metrics approach such a comparison. On the one hand, learned embeddings such as VGGish [124] and a variety of other embeddings have been utilized for the FAD or related metrics, on the other hand there are custom-designed low-level statistical descriptors such as pitch range, pitch class histograms, etc. [267]. This has considerable impact on the interpretability of the descriptors; while a learned descriptor such as VGGish cannot be interpreted directly, custom-designed descriptors tend to be more interpretable. However, high interpretability does neither mean that the descriptor is relevant for assessment nor that it is perceptually or musically meaningful.

Perceptual studies of specific descriptors are necessary to understand their relevance and meaning. Simply finding a difference between two set of data with respect to one descriptor does not automatically mean that these data are different from a perceptual point of view. Recent studies on the suitability of learned embeddings for evaluation seem to focus on overall relevance for summary aesthetic judgments without considering interpretability or musical meaning [108, 130].

Furthermore, even descriptors known for their perceptual validity can be more or less meaningful depending on context and scenario. For instance, it has been demonstrated that people are incredibly sensitive to the statistical distribution of pitches in a piece of music, and can even learn new musical systems and grammars based on such statistical inference (e.g., [133, 155, 170, 248]). However, in general, not all musical representations are equal, and this can have a sizeable impact in the perceptual relevance of any given feature or metric. For instance, most music is tonal meaning that it is (at least temporarily) in a given key or mode and has a stable tonic. For such tonal music, listeners are very sensitive to notes that are outside of the key (i.e., ’wrong notes’) [75, 155, 214]. As such, measuring the statistical distribution of pitches in relation to that tonic (as scale degrees or musical intervals from a tonic) carries a different musical and perceptual relevance compared to the distribution of all pitch classes measured in a tonic-agnostic way. Thus, it is not only necessary to validate whether certain descriptors have perceptual relevance per se, but also in what (e.g., tonal or stylistic) context they are extracted. But even given a set of relevant descriptors we can only guess how exhaustive this set is. At the very least, the number and type of relevant descriptors is genre-dependent, and is quite possibly indefinite.

5.3 Reproducibility

Reproducibility has long been identified as a problem in the machine learning community [233]. For software-based technologies and approaches, the pure description of research in a paper is increasingly considered insufficient and the publication of well-documented open-source code has been identified as one important part of a solution [184, 254]. In addition, as most machine learning is data-driven, understanding the training data and the test data is crucial. Unlike other systems for other machine learning tasks, generative systems often use massive amounts of unlabeled (and potentially unpublished) data of potentially unclear origin and with unclear data curation approaches, meaning that the training of such systems cannot be reproduced by unaffiliated parties.

For generative systems, we introduce the following levels of reproducibility with increasing level of transparency: (i) *publication of an academic text*, describing the method and approaches, (ii) *publication of all raw results*, including the generated music in order to reproduce the result-based conclusions, (iii) *publication of the generative system* itself (e.g., through an API) to allow reproducing the results with a clear documentation of the system prompts and parameter settings from the study, (iv) *publication of documented source code* of the pre-trained system in order to allow in-depth understanding of architectural details and parameters not published otherwise, (v) *publication of training source code* for the generative system to share details on data processing and training methodologies, (vi) *publication of training data statistics* to improve transparency around data distribution and characteristics, possible bias, and other details, (vii) *publication of training data* and source code for data pre-processing and curation, and (viii) *publication of data acquisition and curation strategies* to be transparent about potential bias, data licenses, and fair data use. While we acknowledge that constraints exist that do not always allow for publication of every single detail, we call for full transparency as the goal of any scientific work in this area to the extent possible.

In addition, HCI studies of generative systems require the publication of user data collected such as questionnaire results, interview transcripts, music generated, and recordings of human interaction with generative system if the studies are themselves to be reproducible by other researchers. There are substantial privacy and practical challenges to making such data available and shareable, not least the lack of any standards for sharing user study data in HCI to date.

5.4 Need for de-facto evaluation standards

As described above, there is currently a multitude of evaluation methodologies and metrics across studies in the field that prevents results from being comparable to each other. This means that the capabilities and shortcomings of systems are not consistently assessed, and no meaningful conclusions regarding the progress of the field can be drawn. Clearly, there is a need for de-facto evaluation standards (compare also Xiong et al. [265], Zhou [274]). More specifically, standards are needed with respect to (i) assessment targets, (ii) evaluation methodology, (iii) commonly used, publicly available reference (test) data sets, and (iv) agreement on a base set of evaluation metrics that allows for future extension with additional metrics to avoid metric overfitting. Even an imperfect set of metrics can help a field moving forward, as the continued use of BSSEval metrics SDR, SAR, and SIR [256] for source separation systems shows — despite widely known shortcomings [79, 94, 112]. These metrics might complement HCI approaches where there are no de-facto standards for evaluation in general. Whilst "mixed-methods" is a current methodological trend, HCI studies are designed to respond to the goals of the evaluation, drawing from both quantitative and qualitative approaches. As such, de-facto standards and benchmarks might not be as meaningful in HCI evaluations where it is more important to understand the features of the evaluation technique used and the user data collected.

6 Conclusion

We presented an overview of the state-of-the-art in evaluating generative systems in music from the perspective of both the system output and the usability of the system. We categorized different evaluation goals and targets, as well as corresponding methodologies and metrics, and concluded that the current state of system assessment makes it difficult to generalize results, compare state of the art systems objectively, and measure progress in the field. The main challenges identified include (i) the perceptual and musical meaningfulness of current evaluation metrics, (ii) the internal and external validity of common experimental setups, and (iii) the lack of reproducibility, emphasizing the need for de-facto evaluation standards adopted by the research community. While the direct assessment of aesthetics might "remain out of reach in our lifetime and perhaps forever" [97], there is a need for methodologies and metrics that give us at least a glimpse into some aspects of quality.

In addition to the evaluation targets presented above, there exist other evaluation targets of interest. Of note could be targets that can be summarized under the umbrella of Responsible AI [210]. These include, e.g., (i) *explainability*: the increasing complexity of machine learning systems puts forth questions with respect to usage and deployment of these systems [146], including in the arts [35], however, standardized evaluation strategies do not exist [14]; (ii) *bias*: bias of machine learning systems is a known problem [13, 39] and can —in the case of generative music systems— lead to marginalization of non-mainstream musical styles [36]; (iii) *ethical use of data*: ethical data acquisition, guiding principles, and transparency on data content and curation are crucial for the holistic evaluation of a machine learning system [141], as the call for "dataset audit cards" [22] and formalized datasheets for datasets [100, 234] emphasize; (iv) *resource use* [20, 53, 237]: the high energy consumption of today's models can be linked to environmental impacts; the reporting of carbon emissions [119] or the relation of energy consumption to the subjective output quality of generative music systems [74] could promote energy-responsible research.

Given the complexity and open-endedness of the task, one should not forget about other ways of assessing or engaging in a dialogue with generative systems. Musicology has a long history of engaging critically with new pieces and forms of music, and traditional modes of assessment should not be discarded as invalid approaches to analyzing and evaluating music, although —as Sturm et al. [241] point out— the large scale generation of music creates new challenges for these approaches. Artistic inquiry is another form of assessing system output that can create societal awareness. For instance, artists have a history of exposing bias and discrimination in generative AI systems [99, 232]. These artistic discourses offer a lens through which to explore future values and metrics of evaluation beyond the state of the art surveyed in this paper.

Furthermore, it became clear in writing this paper that for addressing these challenges interdisciplinarity is a necessity, as an exhaustive system evaluation requires expertise not only in the field of machine learning, but also in music theory and musicology, psychology, human computer interaction, and possibly others. In our view, this is especially true for generative systems for music as music is a fundamental form of human creativity, social interaction, and intangible cultural heritage which itself has defied evaluation for millennia.

Acknowledgments

Corey Ford was a PhD student at Queen Mary University of London's UKRI Centre for Doctoral Training in AI and Music at the inception of this work (grant number EP/S022694/1).

References

[1] A. R. Addessi, L. Ferrari, and F. Carugati. 2015. The Flow Grid: A Technique for Observing and Measuring Emotional State in Children Interacting with a Flow Machine. *J. of New Music Res. (JNMR)* 44, 2 (2015), 129–144. doi:10.1080/09298215.2014.991738

[2] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank. 2023. MusicLM: Generating Music From Text. doi:10.48550/arXiv.2301.11325

[3] K. Agres, J. Forth, and G. A. Wiggins. 2016. Evaluation of Musical Creativity and Musical Metacreation Systems. *Computers in Entertainment* 14, 3 (2016), 1–33. doi:10.1145/2967506

[4] A. Alaa, B. Van Breugel, E. S. Saveliev, and M. van der Schaar. 2022. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. In *Proc. ICML*. PMLR, Baltimore.

[5] T. Alshammari, O. Alhadreti, and P. Mayhew. 2015. When To Ask Participants To Think Aloud: A Comparative Study of Concurrent and Retrospective Think-Aloud Methods. *Int. J. of Human Computer Interaction* 6, 3 (2015), 48–64.

[6] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proc. CHI*. ACM, Glasgow. doi:10.1145/3290605.3300233

[7] I. Ananthabhotla, S. Ewert, and J. A. Paradiso. 2019. Towards a Perceptual Loss: Using a Neural Network Codec Approximation as a Loss for Generative Audio Models. In *Proc. ACMMM*. ACM, Nice. doi:10.1145/3343031.3351148

[8] M. Araneda-Hernandez, F. Bravo-Marquez, D. Parra, and R. F. Cádiz. 2023. MUSIB: musical score inpainting benchmark. *EURASIP J. on Audio, Speech, and Music Process.* 2023, 1 (2023), 19. doi:10.1186/s13636-023-00279-6

[9] B. Banar and S. Colton. 2022. A Quality-Diversity-based Evaluation Strategy for Symbolic Music Generation. In *Proc. ML Evaluation Standards Workshop at ICLR*. Online.

[10] B. Banar and S. Colton. 2022. A Systematic Evaluation of GPT-2-Based Music Generation. In *Artificial Intelligence in Music, Sound, Art and Design*. Springer, Cham, 19–35. doi:10.1007/978-3-031-03789-4_2

[11] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi. 2023. The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet* 15, 8 (2023), 260. doi:10.3390/fi15080260

[12] A. Bangor, P. T. Kortum, and J. T. Miller. 2008. An Empirical Evaluation of the System Usability Scale. *Int. J. of Human–Computer Interaction* 24, 6 (2008), 574–594. doi:10.1080/10447310802205776

[13] J. Barnett. 2023. The Ethical Implications of Generative Audio Models: A Systematic Literature Review. In *Proc. AAAI/ACM Conf. on AI, Ethics, and Soc. (AIES)*. ACM, New York. doi:10.1145/3600211.3604686

[14] R. Batlle-Roca, E. Gómez, W. Liao, X. Serra, and Y. Mitsufuji. 2023. Transparency in Music-Generative AI: A Systematic Literature Review. doi:10.21203/rs.3.rs-3708077/v1

[15] R. Batlle-Roca, W.-H. Liao, X. Serra, Y. Mitsufuji, and E. Gómez. 2024. Towards Assessing Data Replication in Music Generation with Music Similarity Metrics on Raw Audio. doi:10.48550/arXiv.2407.14364

[16] M. Bellingham. 2022. *Choosers: A Visual Programming Language for Nondeterministic Music Composition by Non-Programmers.* Ph. D. Dissertation. The Open University, Milton Keynes.

[17] O. Ben-Tal, M. T. Harris, and B. L. T. Sturm. 2021. How Music AI Is Useful: Engagements with Composers, Performers and Audiences. *Leonardo* 54, 5 (2021), 510–516. doi:10.1162/leon_a_01959

[18] S. Benford, C. Greenhalgh, A. Crabtree, M. Flintham, B. Walker, J. Marshall, B. Koleva, S. Rennick Egglestone, G. Giannachi, M. Adams, N. Tandavanitj, and J. Row Farr. 2013. Performance-Led Research in the Wild. *ACM Trans. on Computer-Human Interaction* 20, 3 (2013). doi:10.1145/2491500.2491502

[19] D. E. Berlyne. 1971. *Aesthetics and psychobiology.* Appleton-Century-Crofts, New York.

[20] A. Berthelot, E. Caron, M. Jay, and L. Lefèvre. 2024. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. In *Proc. CIRP Conf. on Life Cycle Engineering.*

[21] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. 2018. Demystifying MMD GANs. In *Proc. ICLR*. Vancouver.

[22] A. Birhane and V. U. Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *Proc. WACV*. IEEE. doi:10.1109/WACV48630.2021.00158

[23] A. F. Blackwell and T. R. G. Green. 2000. A Cognitive Dimensions Questionnaire Optimised for Users. In *Proc. Workshop on Philosophy of Programming Interest Group (PPIG).*

[24] M. A. Blythe, K. Overbeeke, A. F. Monk, and P. C. Wright. 2004. *Funology: From Usability to Enjoyment* (1 ed.). Springer, Dordrecht.

[25] M. A. Boden. 1998. Creativity and artificial intelligence. *Artificial Intelligence* 103, 1 (1998), 347–356. doi:10.1016/S0004-3702(98)00055-1

[26] A. Borji. 2022. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding* 215 (2022), 103329. doi:10.1016/j.cviu.2021.103329

[27] N. Borkar, S. Patre, R. S. Khalsa, R. Kawale, and P. Chakurkar. 2021. Music Plagiarism Detection using Audio Fingerprinting and Segment Matching. In *Proc. STCR*. doi:10.1109/STCR51658.2021.9587927

[28] R. Bougueng Tchemeube, J. Ens, C. Plut, P. Pasquier, M. Safi, Y. Grabit, and J.-B. Rolland. 2023. Evaluating Human-AI Interaction via Usability, User Experience and Acceptance Measures for MMM-C: A Creative AI System for Music Composition. In *Proc. IJCAI*. Macao. doi:10.24963/ijcai.2023/640

[29] E. Brattico and T. Jacobsen. 2009. Subjective Appraisal of Music. *Annals New York Acad. of Sciences* 1169, 1 (2009), 308–317. doi:10.1111/j.1749-6632.2009.04843.x

[30] V. Braun and V. Clarke. 2013. *Successful Qualitative Research: A Practical Guide for Beginners.* SAGE Publications Ltd.

[31] V. Braun and V. Clarke. 2019. Reflecting on Reflexive Thematic Analysis. *Qualitative Res. in Sport, Exercise and Health* 11, 4 (2019), 589–597. doi:10.1080/2159676X.2019.1628806

[32] J.-P. Briot, G. Hadjeres, and F. Pachet. 2020. *Deep Learning Techniques for Music Generation.* Springer, Cham.

[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language Models are Few-Shot Learners. In *Proc. NeurIPS*, Vol. 33. Online.

[34] G. Brunner, Y. Wang, R. Wattenhofer, and S. Zhao. 2018. Symbolic Music Genre Transfer with CycleGAN. In *Proc. ICTAI*. doi:10.1109/ICTAI.2018.00123

[35] N. Bryan-Kinns. 2024. Reflections on Explainable AI for the Arts (XAIxArts). *Interactions* 31, 1 (2024), 43–47. doi:10.1145/3636457

[36] N. Bryan-Kinns, B. Banar, C. Ford, C. N. Reed, Y. Zhang, and J. Armitage. 2024. Explainable AI and Music. In *Artificial Intelligence for Art Creation and Understanding* (1st ed.), L. Mou (Ed.). CRC Press. doi:10.1201/9781003406273

[37] N. Bryan-Kinns, A. Noel-Hirst, and C. Ford. 2024. Using Incongruous Genres to Explore Music Making with AI Generated Content. In *Proc. of Creativity and Cognition (C&C)*. ACM, Chicago. doi:10.1145/3635636.3656198

[38] N. Bryan-Kinns and C. N. Reed. 2023. A Guide to Evaluating the Experience of Media and Arts Technology. In *Creating Digitally: Shifting Boundaries: Arts and Technologies—Contemporary Applications and Concepts*, Anthony L. Brooks (Ed.). Springer, Cham, 267–300.

[39] N. Bryan-Kinns, B. Zhang, S. Zhao, and B. Banar. 2024. Exploring Variational Auto-encoder Architectures, Configurations, and Datasets for Generative Music Explainable AI. *Machine Intelligence Res.* 21, 1 (2024), 29–45. doi:10.1007/s11633-023-1457-1

[40] S. Bødker. 2015. Third-Wave HCI, 10 Years Later—Participation and Sharing. *Interactions* 22, 5 (2015), 24–31.

[41] L. Candy, S. Amitani, and Z. Bilda. 2006. Practice-Led Strategies for Interactive Art Research. *CoDesign* 2, 4 (2006), 209–223. doi:10.1080/15710880601007994

[42] L. Candy and E. Edmonds. 2018. Practice-Based Research in the Creative Arts: Foundations and Futures from the Front Line. *Leonardo* 51, 1 (2018), 63–69. doi:10.1162/LEON_a_01471

[43] B. Caramiaux and S. Fdili Alaoui. 2022. "Explorers of Unknown Planets": Practices and Politics of Artificial Intelligence in Visual Arts. In *Proc. ACMHCI*, Vol. 6. doi:10.1145/3555578

[44] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. 2023. Extracting Training Data from Diffusion Models. In *Proc. USENIX Security Symp. (USENIX Security)*.

[45] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel. 2021. Extracting Training Data from Large Language Models. In *Proc. USENIX Security Symp. (USENIX Security)*.

[46] F. Carnovalini and A. Rodà. 2020. Computational Creativity and Music Generation Systems: An Introduction to the State of the Art. *Frontiers in Artificial Intelligence* 3 (2020). doi:10.3389/frai.2020.00014

[47] R. J. S. Cason and D. Müllensiefen. 2012. Singing from the same sheet: computational melodic similarity measurement and copyright law. *Int. Review of Law, Computers & Technology* 26, 1 (2012), 25–36. doi:10.1080/13600869.2012.646786

[48] R. Chaffin and A. F. Lemieux. 2004. General perspectives on achieving musical excellence. In *Musical Excellence: Strategies and Techniques to Enhance Performance*, Aaron Williamon (Ed.). Oxford University Press, 19–39.

[49] A. Chamberlain, A. Crabtree, T. Rodden, M. Jones, and Y. Rogers. 2012. Research in the Wild: Understanding 'In the Wild' Approaches to Design and Development. In *Proc. DIS*. ACM, New York. doi:10.1145/2317956.2318078

[50] K. Chen, Y. Wu, H. Liu, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov. 2024. MusicLDM: Enhancing Novelty in text-to-music Generation Using Beat-Synchronous mixup Strategies. In *Proc. ICASSP*. doi:10.1109/ICASSP48485.2024.10447265

[51] K. Chen, G. Xia, and S. Dubnov. 2020. Continuous Melody Generation via Disentangled Short-Term Representations and Structural Conditions. In *Proc. ICSC*. IEEE. doi:10.1109/ICSC.2020.00025

[52] E. Cherry and C. Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. on Computer-Human Interaction* 21, 4 (2014). doi:10.1145/2617588

[53] A. A. Chien, L. Lin, H. Nguyen, V. Rao, T. Sharma, and R. Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *Proc. Workshop on Sustainable Computer Systems (HotCarbon)*. ACM, New York. doi:10.1145/3604930.3605705

[54] H. Chu, J. Kim, S. Kim, H. Lim, H. Lee, S. Jin, J. Lee, T. Kim, and S. Ko. 2022. An Empirical Study on How People Perceive AI-generated Music. In *Proc. CIKM*. ACM, Atlanta. doi:10.1145/3511808.3557235

[55] H. Chu, R. Urtasun, and S. Fidler. 2016. Song From PI: A Musically Plausible Network for Pop Music Generation. In *Proc. ICLR*. San Juan.

[56] Y. Chung, P. Eu, J. Lee, K. Choi, J Nam, and B. Sangbae Chon. 2025. KAD: No More FAD! An Effective and Efficient Evaluation Metric for Audio Generation. https://arxiv.org/abs/2502.15602

[57] M. Civit, J. Civit-Masot, F. Cuadrado, and M. J. Escalona. 2022. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applic.* 209 (2022), 118190. doi:10.1016/j.eswa.2022.118190

[58] N. Collins. 2008. The Analysis of Generative Music Programs. *Organised Sound* 13, 3 (2008), 237–248. doi:10.1017/S1355771808000332

[59] D. Cope. 2000. *The algorithmic composer.* Number 16 in The computer music and digital audio series. A-R Ed, Middleton.

[60] D. Cope. 2005. *Computer Models of Musical Creativity.* MIT Press, Cambridge.

[61] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Defossez. 2023. Simple and Controllable Music Generation. In *Proc. NeurIPS*, Vol. 36. New Orleans.

[62] M. Csíkszentmihályi. 1990. *Flow: The Psychology of Optimal Experience.* Harper Collins, New York.

[63] S. Dai, H. Yu, and R. B. Dannenberg. 2022. What is missing in deep music generation? A study of repetition and structure in popular music. In *Proc. ISMIR*. Bengaluru.

[64] S. De, I. Roy, T. Prabhakar, K. Suneja, S. Chaudhuri, R. Singh, and B. Raj. 2012. Plagiarism detection in polyphonic music using monaural signal separation. In *Proc. INTERSPEECH*. ISCA, Portland. doi:10.21437/Interspeech.2012-476

[65] R. De Prisco, A. Esposito, N. Lettieri, D. Malandrino, D. Pirozzi, G. Zaccagnino, and R. Zaccagnino. 2017. Music Plagiarism at a Glance: Metrics of Similarity and Visualizations. In *Proc. IV*. doi:10.1109/iV.2017.49

[66] R. De Prisco, D. Malandrino, G. Zaccagnino, and R. Zaccagnino. 2017. A computational intelligence text-based detection system of music plagiarism. In *Proc. ICSAI*. doi:10.1109/ICSAI.2017.8248347

[67] R. De Prisco, G. Zaccagnino, and R. Zaccagnino. 2020. EvoComposer: An Evolutionary Algorithm for 4-Voice Music Compositions. *Evolutionary Computation* 28, 3 (2020), 489–530. doi:10.1162/evco_a_00265

[68] E. Dervakos, G. Filandrianos, and . Stamou. 2021. Heuristics for Evaluation of AI Generated Music. In *Proc. ICPR*. doi:10.1109/ICPR48806.2021.9413310

[69] D. R. Desai and M. Riedl. 2024. Between Copyright and Computer Science: The Law and Ethics of Generative AI. doi:10.48550/arXiv.2403.14653

[70] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. 2020. Jukebox: A Generative Model for Music. doi:10.48550/arXiv.2005.00341

[71] S. Dieleman, A. van den Oord, and K. Simonyan. 2018. The challenge of realistic music generation: modelling raw audio at scale. In *Proc. NeurIPS*, Vol. 31. Barcelona.

[72] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. McAuley. 2019. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. In *Proc. ISMIR*. Delft.

[73] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang. 2018. MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. In *Proc. AAAI*, Vol. 32. doi:10.1609/aaai.v32i1.11312

[74] C. Douwes, G. Bindi, A. Caillon, P. Esling, and J.-P. Briot. 2023. Is Quality Enough? Integrating Energy Consumption in a Large-Scale Evaluation of Neural Audio Synthesis Models. In *Proc. ICASSP*. IEEE, Rhodes. doi:10.1109/ICASSP49357.2023.10096975

[75] W. Jay Dowling. 1991. Pitch structure. In *Representing Musical Structure*, P. Howell, R. West, and I. Cross (Eds.). London: Academic Press, 33–57.

[76] H. B.-L. Duh, G. C. B. Tan, and V. H.-H. Chen. 2006. Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. In *Proc. MobileHCI*. ACM, Helsinki. doi:10.1145/1152215.1152254

[77] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. 2023. High Fidelity Neural Audio Compression. *Trans. on Machine Learning Res.* (2023). https://openreview.net/forum?id=ivCd8z8zR2

[78] L. Eisenbeiser. 2020. Latent Walking Techniques for Conditioning GAN-Generated Music. In *Proc. IEEE Annual Ubiquitous Comput., Electronics & Mobile Comm. Conf. (UEMCON)*. IEEE. doi:10.1109/UEMCON51285.2020.9298149

[79] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. 2011. Subjective and Objective Quality Assessment of Audio Source Separation. *IEEE Trans. on Audio, Speech, and Language Process.* 19, 7 (2011), 2046–2057. doi:10.1109/TASL.2011.2109381

[80] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and . Roberts. 2019. GANSynth: Adversarial Neural Audio Synthesis. In *Proc. ICLR*. New Orleans.

[81] J. Engel, L. Hantrakul, C. Gu, and A. Roberts. 2020. DDSP: Differentiable Digital Signal Processing. In *Proc. ICLR*. Addis Ababa.

[82] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In *Proc. ICML*. PMLR, Sydney.

[83] J. Ens and P. Pasquier. 2019. Quantifying Musical Style: Ranking Symbolic Music based on Similarity to a Style. In *Proc. ISMIR*. Delft.

[84] Z. Epstein, A. Hertzmann, and The Investigators of Human Creativity. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111. doi:10.1126/science.adh4451

[85] Z. Evans, C. J. Carr, J. Taylor, S. H. Hawley, and J. Pons. 2024. Fast Timing-Conditioned Latent Audio Diffusion. In *Proc. ICML)*. https://openreview.net/forum?id=jOlO8t1xdx

[86] Z. Evans, J. D. Parker, C. J. Carr, Z. Zukowski, J. Taylor, and J. Pons. 2025. Stable Audio Open. In *Proc. ICASSP*. doi:10.1109/ICASSP49660.2025.10888461

[87] J. D. Fernandez and F. Vico. 2013. AI Methods in Algorithmic Composition: A Comprehensive Survey. *J. of Artificial Intelligence Res.* 48 (2013), 513–582. doi:10.1613/jair.3908

[88] R. Fiebrink, P. R. Cook, and D. Trueman. 2011. Human Model Evaluation in Interactive Supervised Learning. In *Proc. CHI*. ACM, Vancouver. doi:10.1145/1978942.1978965

[89] R. Fiebrink and L. Sonami. 2020. Reflections on Eight Years of Instrument Creation with Machine Learning. In *Proc. NIME*. Birmingham City University, Birmingham. doi:10.5281/zenodo.4813334

[90] R. Fiebrink, D. Trueman, C. Britt, M. Nagai, K. Kaczmarek, M. Early, M. R. Daniel, A. Hege, and P. Cook. 2010. Toward Understanding Human-Computer Interaction in Composing the Instrument. In *Proc. ICMC*. ICMA.

[91] C. Ford and N. Bryan-Kinns. 2022. Speculating on Reflection and People's Music Co-Creation with AI. In *Proc. of Gen. AI and HCI Workshop, CHI*. ACM.

[92] C. Ford and N. Bryan-Kinns. 2023. Towards a Reflection in Creative Experience Questionnaire. In *Proc. CHI*. ACM, Hamburg. doi:10.1145/3544548.3581077

[93] C. Ford, A. Noel-Hirst, S. Cardinale, J. Loth, P. Sarmento, E. Wilson, L. Wolstanholme, K. Worrall, and N. Bryan-Kinns. 2024. Reflection Across AI-based Music Composition. In *Creativity and Cognition (C&C '24)*. ACM, Chicago. doi:10.1145/3635636.3656185

[94] B. Fox, A. Sabin, B. Pardo, and A. Zopf. 2007. Modeling Perceptual Similarity of Audio Signals for Blind Source Separation Evaluation. In *Proc. ICA*. Springer, Berlin. doi:10.1007/978-3-540-74494-8_57

[95] C. Frauenberger. 2019. Entanglement HCI The Next Wave? *ACM Trans. on Computer-Human Interaction* 27, 1 (2019). doi:10.1145/3364998

[96] E. Frid, C. Gomes, and Z. Jin. 2020. Music Creation by Example. In *Proc. CHI*. ACM, Honolulu. doi:10.1145/3313831.3376514

[97] P. Galanter. 2012. Computational Aesthetic Evaluation: Past and Future. In *Computers and Creativity*, Jon McCormack and Mark d'Inverno (Eds.). Springer, Berlin, 255–293.

[98] S. Garcia-Valencia, A. Betancourt, and J. G. Lalinde-Pulido. 2021. A framework to compare music generative models using automatic evaluation metrics extended to rhythm. doi:10.48550/arXiv.2101.07669

[99] N. R. Gaskins. 2022. Interrogating AI Bias through Digital Art.

[100] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III, and K. Crawford. 2021. Datasheets for Datasets. *Communications ACM* 64, 12 (2021), 86–92.

[101] H. Gehlbach and M. E. Brinkworth. 2011. Measure Twice, Cut down Error: A Process for Enhancing the Validity of Survey Scales. *Review of General Psychology* 15, 4 (2011), 380–387. doi:10.1037/a0025704

[102] T. Godwin, G. Rizos, A. Baird, N. D. Al Futaisi, V. Brisse, and B. W. Schuller. 2021. Evaluating Deep Music Generation Methods Using Data Augmentation. In *Proc. MMSP*. IEEE. doi:10.1109/MMSP53017.2021.9733502

[103] J. Goguen. 2004. Musical Qualia, Context, Time and Emotion. *J. of Consciousness Studies* 11, 3-4 (2004), 117–147.

[104] A. Golda, K. Mekonen, A. Pandey, V. Singh, A.and Hassija, V. Chamola, and B. Sikdar. 2024. Privacy and Security Concerns in Generative AI: A Comprehensive Survey. *IEEE Access* 12 (2024), 48126–48144. doi:10.1109/ACCESS.2024.3381611

[105] M. Good. 2001. MusicXML: An Internet-Friendly Format for Sheet Music. In *Proc. XML Conf.* Orlando.

[106] C. Goodhart. 2015. Goodhart's Law. In *The Encyclopedia of Central Banking*, L.-P. Rochon and S. Rossi (Eds.). Edward Elgar Publishing, 29–33.

[107] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. 2012. A kernel two-sample test. *JMLR* 13, 25 (2012), 723–773.

[108] F. Grötschla, A. Solak, L. A. Lanzendörfer, and R. Wattenhofer. 2025. Benchmarking Music Generation Models and Metrics via Human Preference Studies. In *Proc. ICASSP*. doi:10.1109/ICASSP49660.2025.10887745

[109] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou. 2024. Adapting Frechet Audio Distance for Generative Music Evaluation. In *Proc. ICASSP*. IEEE, Seoul.

[110] I. Gulrajani, C. Raffel, and L. Metz. 2019. Towards GAN Benchmarks Which Require Generalization. In *Proc. ICLR*. New Orleans.

[111] Y. Guo, Y. Liu, T. Zhou, L. Xu, and Q. Zhang. 2023. An automatic music generation and evaluation method based on transfer learning. *PLOS ONE* 18, 5 (2023), e0283103. doi:10.1371/journal.pone.0283103

[112] U. Gupta, E. Moore II, and A. Lerch. 2015. On the Perceptual Relevance of Objective Source Separation Measures for Singing Voice Separation. In *Proc. WASPAA*. IEEE, New Paltz. doi:10.1109/WASPAA.2015.7336923

[113] S. Gururani and A. Lerch. 2017. Automatic Sample Detection in Polyphonic Music. In *Proc. ISMIR*. Suzhou. doi:10.5281/zenodo.1418331

[114] G. Hadjeres, F. Pachet, and F. Nielsen. 2017. DeepBach: a Steerable Model for Bach Chorales Generation. In *Proc. ICML*. PMLR, Sydney.

[115] S. H. Hakimi, N. Bhonker, and R. El-Yaniv. 2020. BebopNet: Depp Neural Models for Personalized Jazz Improvisations. In *Proc. ISMIR*. Online.

[116] David Hargreaves. 2010. Experimental aesthetics and liking for music. In *The Handbook of Music and Emotion:*, P. N. Juslin and J. A. Sloboda (Eds.). Oxford University Press, 515–546.

[117] S. G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proc. Human Factors and Ergonomics Soc. Annual Meeting* 50, 9 (2006), 904–908. doi:10.1177/154193120605000909

[118] A. Hazzard, C. Greenhalgh, M. Kallionpaa, S. Benford, A. Veinberg, Z. Kanga, and A. McPherson. 2019. Failing with Style: Designing for Aesthetic Failure in Interactive Performance. In *Proc. CHI*. ACM, Glasgow. doi:10.1145/3290605.3300260

[119] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau. 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *JMLR* 21, 248 (2020), 1–43.

[120] C. Hernandez-Olivan, J. A. Puyuelo, and J. R. Beltran. 2022. Subjective Evaluation of Deep Learning Models for Symbolic Music Composition. In *Workshop on Gen. AI and HCI at CHI*. ACM.

[121] J. Hernández-Orallo. 2020. Twenty Years Beyond the Turing Test: Moving Beyond the Human Judges Too. *Minds and Machines* 30, 4 (2020), 533–562. doi:10.1007/s11023-020-09549-0

[122] D. Herremans and E. Chew. 2019. MorpheuS: Generating Structured Music with Constrained Patterns and Tension. *IEEE Trans. on Affective Comput.* 10, 4 (2019), 510–523. doi:10.1109/TAFFC.2017.2737984

[123] D. Herremans, C.-H. Chuan, and E. Chew. 2017. A Functional Taxonomy of Music Generation Systems. *ACM Comput. Surveys* 50, 5 (2017), 69:1–69:30. doi:10.1145/3108242

[124] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *Proc. ICASSP*. IEEE, New Orleans. doi:10.1109/ICASSP.2017.7952132

[125] L. A. Hiller and L. M. Isaacson. 1959. *Experimental Music: Composition with an Electronic Computer.* McGraw-Hill.

[126] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte. 2015. ViSQOL: an objective speech quality model. *EURASIP J. on Audio, Speech, and Music Process.* 2015, 1 (2015), 13. doi:10.1186/s13636-015-0054-9

[127] C.-Z. A. Huang, H. V. Koops, E. Newton-Rex, M. Dinculescu, and C. J. Cai. 2020. AI Song Contest: Human-AI Co-Creation in Songwriting. In *Proc. ISMIR.* Montréal.

[128] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. 2019. Music Transformer: Generating Music with Long-Term Structure. In *Proc. ICLR.* New Orleans.

[129] J. Huang, H. Shao, and K. C.-C. Chang. 2022. Are Large Pre-Trained Language Models Leaking Your Personal Information?. In *Findings Assoc. for Computational Linguistics (EMNLP).* Abu Dhabi.

[130] Y. Huang, Z. Novack, K. Saito, J. Shi, S. Watanabe, Y. Mitsufuji, J. Thickstun, and C. Donahue. 2025. Aligning Text-to-Music Evaluation with Human Preferences. doi:10.48550/arXiv.2503.16669

[131] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions. In *Proc. ACMMM.* ACM, New York. doi:10.1145/3394171.3413671

[132] S. Hunt. 2021. *Empirical Studies in End-User Computer-Generated Music Composition Systems.* PhD Thesis. University of the West of England, Bristol.

[133] D. Huron. 2006. *Sweet Anticipation: Music and the Psychology of Expectation.* MIT Press, Cambridge.

[134] ITU-R. 2006. BS.1387:2006: Method for objective measurements of perceived audio quality.

[135] ITU-R. 2015. BS.1116-3 : Methods for the subjective assessment of small impairments in audio systems.

[136] ITU-R. 2015. BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems.

[137] ITU-T. 1996. P.800 : Methods for subjective determination of transmission quality.

[138] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar. 2023. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In *Proc. CVPR.* Seattle.

[139] S. Ji, X. Yang, and J. Luo. 2023. A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges. *ACM Comput. Surveys* 56, 1 (2023), 7:1–7:39. doi:10.1145/3597493

[140] C. Jin, Y. Tie, Y. Bai, X. Lv, and S. Liu. 2020. A Style-Specific Music Composition Neural Network. *Neural Process. Letters* 52, 3 (2020), 1893–1912. doi:10.1007/s11063-020-10241-8

[141] E. S. Jo and T. Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proc. FAT.* ACM, New York. doi:10.1145/3351095.3372829

[142] A. Jordanous. 2012. A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative. *Cognitive Computation* 4, 3 (2012), 246–279. doi:10.1007/s12559-012-9156-1

[143] P. N. Juslin. 2013. From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Phys. of Life Reviews* 10, 3 (2013), 235–266. doi:10.1016/j.plrev.2013.05.008

[144] P. N. Juslin, E. Ingmar, and J. Danielsson. 2023. Aesthetic judgments of music: Reliability, consistency, criteria, self-insight, and expertise. *Psychology of Aesthetics, Creativity, and the Arts* 17, 2 (2023), 193–210. doi:10.1037/aca0000403

[145] S. Kalonaris and A. Jordanous. 2018. Computational Music Aesthetics: a survey and some thoughts. In *Proc. CSMC.* Dublin.

[146] S. Kambhampati. 2022. Changing the nature of AI research. *Communications ACM* 65, 9 (2022), 8–9. doi:10.1145/3546954

[147] M. Karjalainen. 1985. A new auditory model for the evaluation of sound quality of audio systems. In *Proc. ICASSP*, Vol. 10. IEEE, Tampa. doi:10.1109/ICASSP.1985.1168376

[148] T. Karras, S. Laine, and T. Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proc. CVPR.* IEEE, Seoul.

[149] H. Katayose, M. Hashida, G. De Poli, and K. Hirata. 2012. On Evaluating Systems for Generating Expressive Music Performance: the Rencon Experience. *J. of New Music Res.* 41, 4 (2012), 299–310. doi:10.1080/09298215.2012.745579

[150] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi. 2019. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Proc. INTERSPEECH.* ISCA, Graz. doi:10.21437/Interspeech.2019-2219

[151] M. Kinsella. 2024. Time to Face the Music: A.I. Music Copyright Infringement Battle Makes It to Court.

[152] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Trans. on Audio, Speech, and Language Process.* 28 (2020), 2880–2894. doi:10.1109/TASLP.2020.3030497

[153] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. 2020. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *Proc. ICLR.* Addis Ababa.

[154] A. Kozbelt. 2019. Evolutionary Approaches to Creativity. In *The Cambridge Handbook of Creativity*, J. C. Kaufman and R. J. Sternberg (Eds.). Cambridge University Press, 109–131.

[155] C. L. Krumhansl. 2000. Tonality Induction: A Statistical Approach Applied Cross-Culturally. *Music Perception* 17, 4 (2000), 461–479. doi:10.2307/40285829

[156] D. Kvak. 2022. Towards Evaluation of Autonomously Generated Musical Compositions: A Comprehensive Survey. doi:10.48550/arXiv.2204.04756

[157] M. W. Y. Lam, Q. Tian, T. Li, Z. Yin, S. Feng, M. Tu, Y. Ji, R. Xia, M. Ma, X. Song, J. Chen, W. Yuping, and Y. Wang. 2023. Efficient Neural Music Generation. In *Proc. NeurIPS*, Vol. 36. New Orleans.

[158] B. Laugwitz, T. Held, and M. Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *Proc. Symp. on HCI and Usability for Education and Work*. Springer, Berlin.

[159] E. Lee. 2024. AI and the Sound of Music. *The Yale Law J. Forum* 134 (2024). https://www.yalelawjournal.org/forum/ai-and-the-sound-of-music

[160] A. Lerch. 2009. *Software-Based Extraction of Objective Parameters from Music Performances.* GRIN Verlag, München.

[161] A. Lerch. 2023. *An Introduction to Audio Content Analysis: Music Information Retrieval Tasks and Applic.* (2 ed.). Wiley-IEEE Press, Hoboken.

[162] A. Lerch, C. Arthur, A. Pati, and S. Gururani. 2019. Music Performance Analysis: A Survey. In *Proc. ISMIR*. Delft.

[163] A. Lerch, C. Arthur, K. A. Pati, and S. Gururani. 2020. An Interdisciplinary Review of Music Performance Analysis. *TISMIR* 3, 1 (2020), 221–245. doi:10.5334/tismir.53

[164] S. Li, K. Kallidromitis, A. Gokul, Z. Liao, Y. Kato, K. Kozuka, and A. Grover. 2025. OmniFlow: Any-to-any Generation with Multi-Modal Rectified Flows. In *Proc. CVPR*.

[165] J. Liang. 2023. Harmonizing minds and machines: survey on transformative power of machine learning in music. *Frontiers in Neurorobotics* 17 (2023), 1267561. doi:10.3389/fnbot.2023.1267561

[166] R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 22 140 (1932), 55.

[167] C. Liu, H. Wang, J. Zhao, S. Zhao, H. Bu, X. Xu, J. Zhou, H. Sun, and Y. Qin. 2025. MusicEval: A Generative Music Dataset with Expert Ratings for Automatic Text-to-Music Evaluation. In *Proc. ICASSP*. doi:10.1109/ICASSP49660.2025.10890307

[168] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley. 2023. AudioLDM: Text-to-audio Generation with Latent Diffusion Models. In *Proc. ICML)*, Vol. 202. PMLR. https://proceedings.mlr.press/v202/liu23f.html

[169] Y. Liu and C. Jin. 2024. Impact on quality and diversity from integrating a reconstruction loss into neural audio synthesis. In *Proc. Meeting Acoustical Soc. of America*, Vol. 52. Sydney. doi:10.1121/2.0001871

[170] P. Loui, B. M. Kubit, Y. Ou, and E. H. Margulis. 2023. Imaginings from an unfamiliar world: Narrative engagement with a new musical system. *Psychology of Aesthetics, Creativity, and the Arts* (2023). doi:10.1037/aca0000629

[171] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proc. CHI*. ACM, Honolulu. doi:10.1145/3313831.3376739

[172] R. Louie, J. Engel, and C.-Z. A. Huang. 2022. Expressive Communication: Evaluating Developments in Generative Models and Steering Interfaces for Music Creation. In *Proc. IUI*. ACM, Helsinki. doi:10.1145/3490099.3511159

[173] T. Lu, C.-M. Geist, J. Melechovsky, A. Roy, and D. Herremans. 2025. MelodySim: Measuring Melody-aware Music Similarity for Plagiarism Detection. doi:10.48550/arXiv.2505.20979

[174] A. Lucero, A. Desjardins, C. Neustaedter, K. Höök, M. Hassenzahl, and M. E. Cecchinato. 2019. A Sample of One: First-Person Research Methods in HCI. In *Companion Publication of the 2019 Designing Interactive Systems Conf. (DIS)*. ACM, New York, NY, USA. doi:10.1145/3301019.3319996

[175] A. López-García, B. Martínez-Rodríguez, and V. Liern. 2022. A Proposal to Compare the Similarity Between Musical Products. One More Step for Automated Plagiarism Detection?. In *Mathematics and Computation in Music*. Springer, Cham, 192–204. doi:10.1007/978-3-031-07015-0_16

[176] M. Maier, D. Elsner, C. Marouane, M. Zehnle, and C. Fuchs. 2019. DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks. In *Proc. IJCAI*. Macao. doi:10.24963/ijcai.2019/196

[177] D. Malandrino, R. De Prisco, M. Ianulardo, and R. Zaccagnino. 2022. An adaptive meta-heuristic for music plagiarism detection based on text similarity and clustering. *Data Mining and Knowledge Discovery* 36, 4 (2022), 1301–1334. doi:10.1007/s10618-022-00835-2

[178] B. Manaris, P. Machado, C. McCauley, J. Romero, and D. Krehbiel. 2005. Developing Fitness Functions for Pleasant Music: Zipf's Law and Interactive Evolution Systems. In *Applic. of Evolutionary Comput. (Lecture Notes in Computer Science)*. Springer, Berlin, 498–507. doi:10.1007/978-3-540-32003-6_50

[179] B. Manaris, D. Vaughan, C. Wagner, J. Romero, and R. B. Davis. 2003. Evolutionary Music and the Zipf-Mandelbrot Law: Developing Fitness Functions for Pleasant Music. In *Applic. of Evolutionary Comput. (Lecture Notes in Computer Science)*. Springer, Berlin, 522–534. doi:10.1007/3-540-36605-9_48

[180] P. Manocha, Z. Jin, and A. Finkelstein. 2022. Audio Similarity is Unreliable as a Proxy for Audio Quality. In *Proc. INTERSPEECH*. ISCA, Incheon.

[181] P. Manocha, B. Xu, and A. Kumar. 2021. NORESQA: A Framework for Speech Quality Assessment using Non-Matching References. In *Proc. NeurIPS*. Online.

[182] H. H. Mao, T. Shin, and G. Cottrell. 2018. DeepJ: Style-Specific Music Generation. In *Proc. ICSC*. IEEE. doi:10.1109/ICSC.2018.00077

[183] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin. 2021. GACELA: A Generative Adversarial Context Encoder for Long Audio Inpainting of Music. *IEEE J. of Selected Topics in Signal Process.* 15, 1 (2021), 120–131. doi:10.1109/JSTSP.2020.3037506

[184] B. McFee, J. W. Kim, M. Cartwright, J. Salamon, R. M. Bittner, and J. P. Bello. 2019. Open-Source Practices for Music Signal Processing Research: Recommendations for Transparent, Sustainable, and Reproducible Audio Research. *IEEE Signal Process. Mag.* 36, 1 (2019), 128–137. doi:10.1109/MSP.2018.2875349

[185] D. Meredith (Ed.). 2016. *Computational Music Analysis*. Springer, Cham.

[186] O. Mogren. 2016. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. In *Proc. Constructive Machine Learning Workshop, NIPS*.

[187] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart. 2014. The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLOS ONE* 9, 2 (2014), 1–23. doi:10.1371/journal.pone.0089642

[188] D. Müllensiefen and M. Pendzich. 2009. Court decisions on music plagiarism and the predictive value of similarity algorithms. *Musicae Scientiae* 13, 1_suppl (2009), 257–295. doi:10.1177/102986490901300111

[189] D. Naruse, T. Takahata, Y. Mukuta, and T. Harada. 2022. Pop Music Generation with Controllable Phrase Lengths. In *Proc. ISMIR*. Bengaluru.

[190] J. Nielsen. 1994. *Usability Engineering*. Morgan Kaufmann, San Francisco.

[191] Z. Ning, H. Chen, Y. Jiang, C. Hao, G. Ma, S. Wang, J. Yao, and L. Xie. 2025. DiffRhythm: Blazingly Fast and Embarrassingly Simple End-to-End Full-Length Song Generation with Latent Diffusion. doi:10.48550/arXiv.2503.01183

[192] J. Nistal, S. Lattner, and G. Richard. 2020. Comparing Representations for Audio Synthesis Using Generative Adversarial Networks. In *Proc. EUSIPCO*. doi:10.23919/Eusipco47968.2020.9287799

[193] A. Noel-Hirst and N. Bryan-Kinns. 2023. An Autoethnographic Exploration of XAI in Algorithmic Composition. In *Proc. Int. Workshop on Explainable AI for the Arts (XAIxArts) at ACM Creativity and Cognition*.

[194] D. Norman and J. Nielsen. 2016. *The Definition of User Experience (UX)*. Technical Report. https://www.nngroup.com/articles/definition-user-experience/

[195] B.M. Oliver, J.R. Pierce, and C.E. Shannon. 1948. The Philosophy of PCM. *Proc. IRE* 36, 11 (1948). doi:10.1109/JRPROC.1948.231941

[196] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan. 2020. This time with feeling: learning expressive musical performance. *Neural Comput. and Applic.* 32, 4 (2020), 955–967. doi:10.1007/s00521-018-3758-9

[197] F. Ostermann, I. Vatolkin, and G. Rudolph. 2021. Evaluating Creativity in Automatic Reactive Accompaniment of Jazz Improvisation. *TISMIR* 4, 1 (2021), 210–222. doi:10.5334/tismir.90

[198] O. Ozmen Garibay, B. Winslow, S. Andolina, M. Antona, A. Bodenschatz, C. Coursaris, G. Falco, S. M. Fiore, I. Garibay, K. Grieman, J. C. Havens, M. Jirotka, H. Kacorri, W. Karwowski, J. Kider, J. Konstan, S. Koon, M. Lopez-Gonzalez, I. Maifeld-Carucci, S. McGregor, G. Salvendy, B. Shneiderman, C. Stephanidis, C. Strobel, C. Ten Holter, and W. Xu. 2023. Six Human-Centered Artificial Intelligence Grand Challenges. *Int. J. of Human–Computer Interaction* 39, 3 (2023), 391–437. doi:10.1080/10447318.2022.2153320

[199] H. L. O'Brien, P. Cairns, and M. Hall. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *Int. J. of Human-Computer Studies* 112 (2018), 28–39. doi:10.1016/j.ijhcs.2018.01.004

[200] M. O'Neill and R. Loughran. 2017. Limitations from Assumptions in Generative Music Evaluation. *J. of Creative Music Systems* 2, 1 (2017). doi:10.5920/JCMS.2017.12

[201] J.-I. Park, S.-W. Kim, and M. Shin. 2005. Music Plagiarism Detection Using Melody Databases. In *Proc. KES*, Vol. 3683. Springer Berlin Heidelberg, Berlin. doi:10.1007/11553939_98

[202] K. Park, S. Baek, J. Jeon, and Y.-S. Jeong. 2022. Music Plagiarism Detection Based on Siamese CNN. *Human-centric Comput. and Information Sciences* 12, 0 (2022), 502–511. doi:10.22967/HCIS.2022.12.038

[203] M. Pasini and J. Schlüter. 2022. Musika! fast infinite waveform music generation. In *Proc. ISMIR*. Bengaluru.

[204] P. Pasquier, A. Eigenfeldt, O. Bown, and S. Dubnov. 2017. An Introduction to Musical Metacreation. *Computers in Entertainment* 14, 2 (2017), 2:1–2:14. doi:10.1145/2930672

[205] A. Pati, A. Lerch, and G. Hadjeres. 2019. Learning to Traverse Latent Spaces for Musical Score Inpainting. In *Proc. ISMIR*. Delft.

[206] K A. Pati and A. Lerch. 2019. Latent Space Regularization for Explicit Control of Musical Attributes. In *ICML Machine Learning for Music Discovery Workshop (ML4MD), Extended Abstract*. Los Angeles.

[207] K. A. Pati and A. Lerch. 2020. Attribute-based Regularization for Latent Spaces of Variational Auto-Encoders. *Neural Comput. and Applic.* (2020). doi:10.1007/s00521-020-05270-2

[208] M. T. Pearce. 2015. Effects of Expertise on the Cognitive and Neural Processes Involved in Musical Appreciation. In *Art, Aesthetics, and the Brain*, J. P. Huston, M. Nadal, F. Mora, L. F. Agnati, and C. J. C. Conde (Eds.). Oxford University Press. doi:10.1093/acprof:oso/9780199670000.003.0016

[209] A. Pease and S. Colton. 2011. On Impact and Evaluation in Computational Creativity: A Discussion of the Turing Test and an Alternative Proposal. In *Proc. AISB*.

[210] A. M. Piskopani, A. Chamberlain, and C. Ten Holter. 2023. Responsible AI and the Arts: The Ethical and Legal Implications of AI in the Arts and Creative Industries. In *Proc. Int. Symp. on Trustworthy Autonomous Systems (TAS)*. ACM, Edinburgh. doi:10.1145/3597512.3597528

[211] J. A. Plucker and M. C. Makel. 2010. Assessment of Creativity. In *The Cambridge Handbook of Creativity* (1 ed.), J. C. Kaufman and R. J. Sternberg (Eds.). Cambridge University Press, 48–73.

[212] J. Preece, H. Sharp, and Y. Rogers. 2015. *Interaction Design: Beyond Human-Computer Interaction.* Wiley.

[213] N. Rajcic, M. T. Llano Rodriguez, and J. McCormack. 2024. Towards a Diffractive Analysis of Prompt-Based Generative AI. In *Proc. CHI.* ACM, Honolulu. doi:10.1145/3613904.3641971

[214] R. Raman, K. Herndon, and W. J. Dowling. 2016. Effects of Familiarity, Key Membership, and Interval Size on Perceiving Wrong Notes in Melodies. In *Proc. ICMPC.* San Francisco.

[215] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proc. ICML.* PMLR, Vienna.

[216] Amon Rapp. 2018. *Autoethnography in Human-Computer Interaction: Theory and Practice.* Springer, Cham, 25–42. doi:10.1007/978-3-319-73374-6_3

[217] C. K. A. Reddy, V. Gopal, and R. Cutler. 2021. DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors. In *Proc. ICASSP.* Toronto.

[218] J. Retkowski, J. Stępniak, and M. Modrzejewski. 2025. Frechet Music Distance: A Metric For Generative Symbolic Music Evaluation. doi:10.48550/arXiv.2412.07948

[219] S. Rhys Cox, H. Bøjer Djernæs, and N. van Berkel. 2025. Beyond Productivity: Rethinking the Impact of Creativity Support Tools. In *Proc. Conf. on Creativity and Cognition (CC).* ACM, New York. doi:10.1145/3698061.3726924

[220] S. Rhys Cox, H. Bøjer Djernæs, and N. van Berkel. 2025. Reflecting Human Values in XAI: Emotional and Reflective Benefits in Creativity Support Tools. In *Proc. Explainable AI for the Arts Workshop 2025 (XAIxArts).* ACM, New York.

[221] Eitan Richardson and Yair Weiss. 2018. On GANs and GMMs. In *Proc. NeurIPS*, Vol. 31. Montreal.

[222] G. Ritchie. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds and Machines* 17, 1 (2007), 67–99. doi:10.1007/s11023-007-9066-2

[223] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck. 2018. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *Proc. ICML.* PMLR, Stockholm.

[224] M. Rohrmeier. 2022. On Creativity, Music's AI Completeness, and Four Challenges for Artificial Musical Creativity. *TISMIR* 5, 1 (2022), 50–66. doi:10.5334/tismir.104

[225] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. 2016. Improved Techniques for Training GANs. In *Proc. NeurIPS.* Barcelona.

[226] D. M. Sanbonmatsu, E. H. Cooley, and J. E. Butner. 2021. The Impact of Complexity on Methods and Findings in Psychological Science. *Frontiers in Psychology* 11 (2021), 580111. doi:10.3389/fpsyg.2020.580111

[227] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf. 2024. Moûsai: Efficient Text-to-Music Diffusion Models. In *Proc. ACL.* ACL, Bangkok, Thailand. doi:10.18653/v1/2024.acl-long.437

[228] E. Schubert. 2021. Creativity Is Optimal Novelty and Maximal Positive Affect: A New Definition Based on the Spreading Activation Model. *Frontiers in Neuroscience* 15 (2021).

[229] H. Scurto, B. Caramiaux, and F. Bevilacqua. 2021. Prototyping Machine Learning Through Diffractive Art Practice. In *Proc. DIS.* ACM, New York. doi:10.1145/3461778.3462163

[230] D. B. Shank, C. Stefanik, C. Stuhlsatz, K. Kacirek, and A. M. Belfi. 2023. AI composer bias: Listeners like music less when they think it was composed by an AI. *J. of Exp. Psychology: Appl.* 29, 3 (2023), 676.

[231] B. Shneiderman. 2022. *Human-Centered AI.* Oxford University Press.

[232] Z. Small. 2023. Black Artists Say A.I. Shows Bias, With Algorithms Erasing Their History. *New York Times* (2023).

[233] S. Sonnenburg, M. L. Braun, C. S. Ong, S. Bengio, L. Bottou, G. Holmes, and Y. LeCun. 2007. The Need for Open Source Software in Machine Learning. In *Proc. Conf. on Human-Computer Interaction with Mobile Devices and Services.*

[234] R. Srinivasan, E. Denton, J. Famularo, N. Rostamzadeh, F. Diaz, and B. Coleman. 2021. Artsheets for Art Datasets. In *Proc. NeurIPS.* Online.

[235] B. Stoltz and A. Aravind. 2019. MU_PSYC: Music Psychology Enriched Genetic Algorithm. In *Proc. IEEE Congress on Evolutionary Computation (CEC).* IEEE. doi:10.1109/CEC.2019.8790099

[236] M. Strathern. 1997. 'Improving ratings': audit in the British University system. *Europ. Review* 5, 3 (1997), 305–321. doi:10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4

[237] E. Strubell, A. Ganesh, and A. McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proc. ACL.*

[238] B. Sturm. 2022. Generative AI Helps One Express Things for Which They May Not Have Expressions (Yet). In *Proc. of Gen. AI and HCI Workshop, CHI.* ACM.

[239] B. L. T. Sturm and O. Ben-Tal. 2017. Taking the Models back to Music Practice: Evaluating Generative Transcription Models built using Deep Learning. *J. of Creative Music Systems* 2 (2017), 32–60. doi:10.5920/JCMS.2017.09

[240] B. L. T. Sturm, O. Ben-Tal, Ú. Monaghan, N. Collins, D. Herremans, E. Chew, G. Hadjeres, E. Deruty, and F. Pachet. 2019. Machine Learning Research That Matters for Music Creation: A Case Study. *J. of New Music Res. (JNMR)* 48, 1 (2019), 36–55. doi:10.1080/

09298215.2018.1515233

[241] Bob L T Sturm, Ken Déguernel, Rujing S Huang, André Holzapfel, Oliver Bown, Nick Collins, Jonathan Sterne, Laura C Vila, Luca Casini, David C Dalmazzo, Eric Drott, and Oded Ben-Tal. 2024. MusAIcology: AI Music and the Need for a New Kind of Music Studies. doi:10.31235/osf.io/9pz4x

[242] B. L. T. Sturm, M. Iglesias, O. Ben-Tal, M. Miron, and E. Gómez. 2019. Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis. *Arts* 8, 3 (2019), 115. doi:10.3390/arts8030115

[243] M. Suh, E. Youngblom, M. Terry, and C. J. Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proc. CHI*. ACM, New York. doi:10.1145/3411764.3445219

[244] K. Suneja and M. Bansal. 2015. Comparison of time series similarity measures for plagiarism detection in music. In *Proc. IEEE India Conf. (INDICON)*. IEEE. doi:10.1109/INDICON.2015.7443304

[245] E. Sunray. 2021. Sounds of Science: Copyright Infringement in AI Music Generator Outputs. *Catholic University J. of Law and Technology* 29, 2 (2021), 185–218.

[246] M. Supper. 2001. A Few Remarks on Algorithmic Composition. *Computer Music J.* 25, 1 (2001), 48–53.

[247] H. H. Tan and D. Herremans. 2020. Music FaderNets: Controllable Music Generation Based On High-Level Features via Low-Level Feature Modelling. In *Proc. ISMIR*. Online.

[248] D. Temperley. 2004. *The cognition of basic musical structures* (1. paperback ed ed.). MIT Press, Cambridge.

[249] The MIDI Assoc. 2023. MIDI 2.0 Specification Overview.

[250] L. Theis, A. v. d. Oord, and M. Bethge. 2016. A note on the evaluation of generative models. In *Proc. ICLR*. San Juan.

[251] N. J. W. Thelle and R. Fiebrink. 2022. How Do Musicians Experience Jamming With a Co-Creative “AI”?. In *Proc. NeurIPS*. New Orleans.

[252] D. Tymoczko. 2011. *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice.* Oxford University Press, New York.

[253] G. van den Burg and C. Williams. 2021. On Memorization in Probabilistic Deep Generative Models. In *Proc. NeurIPS*, Vol. 34. Online.

[254] P. Vandewalle, J. Kovacevic, and M. Vetterli. 2009. Reproducible research in signal processing. *IEEE Signal Process. Mag.* 26, 3 (2009), 37–47. doi:10.1109/MSP.2009.932122

[255] A. Vinay and A. Lerch. 2022. Evaluating Generative Audio Systems and their Metrics. In *Proc. ISMIR*. Bangalore.

[256] E. Vincent, R. Gribonval, and C. Fevotte. 2006. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech, and Language Process.* 14, 4 (2006), 1462–1469. doi:10.1109/TSA.2005.858005

[257] M. Wanderley and N. Orio. 2002. Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI. *Computer Music J.* 26, 3 (2002), 62–76.

[258] L. Wang, Z. Zhao, H. Liu, J. Pang, Y. Qin, and Q. Wu. 2024. A review of intelligent music generation systems. *Neural Comput. and Applic.* 36, 12 (2024), 6381–6401. doi:10.1007/s00521-024-09418-2

[259] S. Wang, Y. Du, X. Guo, B. Pan, Z. Qin, and L. Zhao. 2024. Controllable Data Generation by Deep Learning: A Review. *ACM Comput. Surveys* 56, 9 (2024), 228:1–228:38. doi:10.1145/3648609

[260] S. Wang, Y. Tie, X. Li, X. Wang, and L. Qi. 2023. Intelligence Evaluation of Music Composition Based on Music Knowledge. In *Advanced Intelligent Comput. Technology and Applic.* Springer Nature, Singapore, 373–384. doi:10.1007/978-981-99-4761-4_32

[261] W. Wang, J. Li, Y. Li, and X. Xing. 2024. Style-conditioned music generation with Transformer-GANs. *Frontiers of Information Technology & Electronic Engineering* 25, 1 (2024), 106–120. doi:10.1631/FITEE.2300359

[262] A. Wolf and D. Müllensiefen. 2011. The perception of similarity in court cases of melodic plagiarism and a review of measures of melodic similarity. In *Proc. SysMus*. Cologne.

[263] S.-L. Wu and Y.-H. Yang. 2020. The Jazz Transformer on the Front Line: Exploring the Shortcomings of AI-composed Music through Quantitative Measures. In *Proc. ISMIR*. Montréal.

[264] Y. Wu, K. Chen, T. Zhang, Y.. Hui, T. Berg-Kirkpatrick, and S. Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *Proc. ICASSP*. IEEE. doi:10.1109/ICASSP49357.2023.10095969

[265] Z. Xiong, W. Wang, J. Yu, Y. Lin, and Z. Wang. 2023. A Comprehensive Survey for Evaluation Methodologies of AI-Generated Music. (2023).

[266] L.-C. Yang, S.-Y. Chou, and Y.-H. Y. 2017. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation. In *Proc. ISMIR*. Suzhou.

[267] L.-C. Yang and A. Lerch. 2020. On the Evaluation of Generative Models in Music. *Neural Comput. and Applic.* 32 (2020), 4773–4784. doi:10.1007/s00521-018-3849-7

[268] Y.-C. Yeh, W.-Y. Hsiao, S. Fukayama, T. Kitahara, B. Genchel, H.-M. Liu, H.-W. Dong, Y. Chen, T. Leong, and Y.-H. Yang. 2021. Automatic melody harmonization with triad chords: A comparative study. *J. of New Music Res. (JNMR)* (2021).

[269] Z. Yin, F. Reuben, S. Stepney, and T. Collins. 2021. “A Good Algorithm Does Not Steal – It Imitates”: The Originality Report as a Means of Measuring When a Music Generation Algorithm Copies Too Much. In *Artificial Intelligence in Music, Sound, Art and Design*. Springer, Cham, 360–375. doi:10.1007/978-3-030-72914-1_24

[270] Z. Yin, F. Reuben, S. Stepney, and T. Collins. 2023. Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation. *Machine Learning* 112, 5 (2023), 1785–1822. doi:10.1007/s10994-023-06309-w

[271] R. Yuan, H. Lin, Y. Wang, Z. Tian, S. Wu, T. Shen, G. Zhang, Y. Wu, C. Liu, Z. Zhou, Z. Ma, L. Xue, Z. Wang, Q. Liu, T. Zheng, Y. Li, Y. Ma, Y. Liang, X. Chi, R. Liu, Z. Wang, P. Li, J. Wu, C. Lin, Q. Liu, T. Jiang, W. Huang, W. Chen, E. Benetos, J. Fu, G. Xia, R. Dannenberg, W. Xue, S. Kang, and Y. Guo. 2024. ChatMusician: Understanding and Generating Music Intrinsically with LLM. doi:10.48550/arXiv.2402.16153

[272] Y. Yuan, C. Cronin, D. Müllensiefen, S. Fujii, and P. E. Savage. 2023. Perceptual and automated estimates of infringement in 40 music copyright cases. *TISMIR* 6, 1 (2023), 117–134. doi:10.5334/tismir.151

[273] Y. Zhao, M. Yang, Y. Lin, X. Zhang, F. Shi, Z. Wang, J. Ding, and H. Ning. 2025. AI-Enabled Text-to-Music Generation: A Comprehensive Review of Methods, Frameworks, and Future Directions. *Electronics* 14, 6 (2025), 1197. doi:10.3390/electronics14061197

[274] X. Zhou. 2023. Analysis of Evaluation in Artificial Intelligence Music. *J. of Artificial Intelligence Practice* 6, 8 (2023), 6–11. doi:10.23977/jaip.2023.060802

[275] D. Zimmermann. 1996. Creativity versus Determinism: Cognitive Science and Music Theory as Touchstones of Automatic Music Composition. In *Proc. ICMC*. ICMA, Hong Kong.

[276] G. K. Zipf. 1965. *Human behavior and the principle of least effort; an introduction to human ecology.* Hafner, New York.

Received 26 June 2024; revised 8 August 2025; accepted 12 September 2025