My research goal is to **build music technology that enables a broader set of users to engage with music on a deeper level**. Listening to music offers a degree of fulfillment in our day-to-day lives, but for most of us, our capacity to engage with it on a level beyond listening is stymied by the high level of expertise that existing tools require. To overcome this, my work involves (1) improving **machine learning** (ML) methods for generative modeling of music, audio, and other sequential data, and (2) building real-world interactive music systems which allow a broader audience—inclusive of non-musicians—to harness the power of resultant models. By pursuing new music technology, not only do we stand to push the boundaries of a treasured art form, but also to uncover general principles by which technology can empower human creativity.

With a broader audience in mind, a theme of my work is **bridging the expertise divide** in music. Regardless of our musical training, we all have profound musical *intuition*—evident in our ability to dance in time to live music, or hum the melody of our favorite song. Some of us may also desire to express this intuition externally by creating music. But existing tools (e.g., instruments, notation) require years of formal training, creating an *expertise divide* which segregates users into musicians and non-musicians, and prevents the latter from realizing their creative ambitions. Using ML, we can bridge this divide by learning to translate high-level intuition into low-level notes, sacrificing the fine-grained control that musicians enjoy but sidestepping the requirement of expertise. As a concrete example, a system I built called *Piano Genie* [11] uses novel generative modeling techniques to convert high-level input from novices into realistic piano improvisations,[1] and has been used by a professional rock band to blur the lines between audience and performers (Figure 1).

Another theme of my work involves **connecting music with other modalities** to create rich experiences. Music is inherently *multimodal*: not only does music itself manifest both as audio and notation, but it can also be combined with entirely different modalities to create distinct *experiences* with broad appeal—e.g., pairing music and movement yields dancing. Despite music's abstract nature, we have strong intuitions about these cross-modal connections—an unreasonable pairing may detract from the overall experience. My work uses ML to model these relationships, offering a bridge between modalities. For example, by modeling the relationship between music and the physical gestures found in rhythm games [8], I built *Beat Sage*, a live service which enables thousands of users per day to transform music into rich interactive content.[2]

While ML has the potential to unlock deeper interactions for a broader audience, realizing this potential requires new ML methodology and interdisciplinary coordination to overcome domain-specific challenges. One obstacle is that music audio is high-dimensional: even a four-minute song contains over ten million timesteps. Modeling this data demands new methodology—for example, my work was the first to show that adversarial learning can be used to generate audio [10], which required incorporating insights from signal processing. Another challenge lies in the multimodal nature of music: modeling cross-modal relationships requires coalescence with techniques designed for target modalities. Even if we can overcome these (and other) learning challenges, resultant models must be exposed through intuitive interfaces in order to be useful to users, necessitating synergy between ML and interaction design.

Below, I outline the two-pronged approach of my research. In Section 1, I describe my interdisciplinary work on building real-world interactive music systems for a broader audience. In Section 2, I discuss my parallel work on improving ML methods for generative modeling to enable new interactions in the long-term.

## 1 Building real-world interactive music systems for a broader audience

I build music technology for a broader audience by developing new ML methods to (1) bridge the expertise divide, and (2) facilitate multimodal interaction. I approach these goals holistically: not only do I develop new methodology, but I also build and deploy real-world interactive systems which allow users to directly benefit. It is my belief that the intrinsic value of music technology is determined not by the quantitative performance of its underlying methods, but instead by how the technology influences music culture in the hands of users.

[1] Piano Genie example: https://youtu.be/YRb0XAnUpIk, demo: https://chrisdonahue.com/piano-genie

[2] Beat Sage example: https://youtu.be/nDZ61cRBhzU?t=185, live service: https://beatsage.com

Consequently, deploying systems to real users is a part of the research process itself, functioning as the true evaluation of any new music technology, and providing signal to guide future work.

**Bridging the expertise divide.** We all have a degree of musical intuition, and some of us may want to express this intuition externally by creating music, but the presumption of expertise in existing tools creates a divide which prevents non-musicians from doing so. My work bridges this divide by training *controllable* generative models which map high-level gestures (user inputs) into music. Past work shows that ML can recognize gestures from sensor data [14], or generate music autonomously [17, 9, 4, 3], but these methods alone do not allow for controllable generation. In 2019, I built Piano Genie [11] which showed for the first time that generative models can learn a real-time and intuitive mapping from gestures to music, thereby allowing non-musicians a glimpse at musical improvisation. Specifically, I proposed a novel discrete autoencoding method as a mechanism for learning to decode novice improvisations on an eight-button miniature piano into realistic performances on a full 88-key piano. Using ML to map high-level inputs to low-level details is a general principle of aiding human creativity that goes beyond music—e.g., a subsequent system called *GauGAN* adopted a similar approach to translate high-level scene sketches into photorealistic images [21].

In addition to enabling non-musicians to improvise, generative models can also facilitate collaboration between musicians and non-musicians. In 2019, a professional rock band *The Flaming Lips* used Piano Genie to bring the audience into their concert. For this event, I modified Piano Genie to take in an additional conditioning signal: the musical content that the band was playing (specifically, the chords), thereby encouraging it to output appropriate notes. Then, the band gave the audience control over the eight buttons by mapping each button to a beach ball containing a remote sensor, such that Piano Genie would play a note out of the speakers each time a ball was struck (Figure 1).[3] **This concert serves as concrete evidence that ML can bridge the expertise divide in music**, enabling new interactions for a broader set of users.

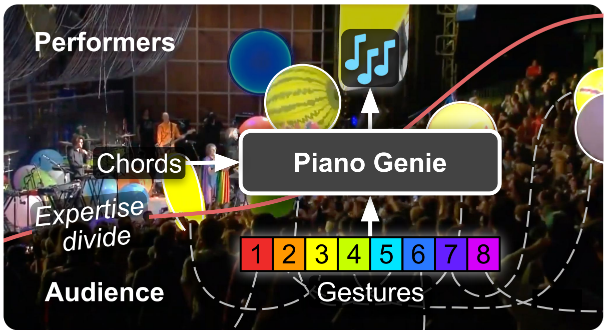

Figure 1: The Flaming Lips used my work to invite the audience into their show. The crowd struck beach balls to play the eight buttons of Piano Genie, which was also conditioned on the band's chords.

**Connecting music with other modalities.** Our appreciation of music can be heightened by combining it with other modalities. Certain unions can even give rise to new interactive experiences, ones which incorporate music listening but are entirely distinct sensations that appeal to a broad audience, e.g. dancing combines music and movement. These cross-modal relationships might seem abstract, but in reality we have strong intuitions about them, e.g., performing ballet to heavy metal would likely not be intuitive. Thus, building systems which facilitate multimodal experiences requires first understanding these intuitions.

My work uses ML to model such relationships, giving rise to rich interactions. In 2017, I demonstrated that ML can model the relationship between music audio and the physical gestures found in rhythm-based video games [8], unlocking the ability for users to play to any song of their choosing. To make this work widely accessible, I recently built a free, user-facing live service called Beat Sage which uses the same approach to create content for the popular virtual reality rhythm game *Beat Saber*. Beat Sage appeals to a broad audience—it can produce interactive content for any song and accommodates players of any skill level through configurable difficulty. Since launching in early 2020, Beat Sage has generated more than two million levels, and is still used daily by thousands of users. This uptake indicates that ML can indeed learn multimodal relationships that are intuitive to real users, and provides evidence that understanding human intuition across modalities is an important principle of technology which supports humans in creative domains.

[3] Additional context about The Flaming Lips' performance with Piano Genie: https://youtu.be/HGWkQP9lVPw

In addition to the connections that exist between music and other modalities, there is also a strong multimodal relationship *within* music. Namely, music manifests both as audio and *scores* (notation). To experts, these forms are deeply intertwined, as strongly connected as spoken and written language. Hence, modeling the relationship between audio and scores can close the gap between human and computer understanding of music. I recently built a system called *Sheet Sage* [7] which transcribes music audio into *lead sheets*: human-readable musical scores which depict the melody and harmony of the audio. This system builds on my recent work which demonstrates that large-scale pre-training of generative models is useful for transfer learning in music [1].[4] Melody is among the most salient aspects of human music perception, hence, Sheet Sage is a step towards human-like music understanding. Moreover, by transcribing any Western song into a human-readable score, my work helps a broad audience more quickly perform their favorite songs, in contrast to existing work [16] which can only transcribe particular domains like piano and cannot output scores.

## 2 Improving machine learning methods for generative modeling

In addition to building interactive music systems that are realistic in the short-term, I work in parallel on improving ML methods for generative modeling to lay the foundation for new interactions in the long term. Because music incorporates many modalities, my research seeks to improve generative modeling for several types of data in two broad categories: (1) *acoustic* data like music audio and speech, and (2) *symbolic* data like music scores and language.

**Acoustic data.** Audio generation has numerous potential applications not only in music but also in speech processing and sound design, but audio waveforms have proven to be empirically difficult to model with standard ML approaches. One challenge is that they are extraordinarily high-dimensional: a single second of audio contains tens of thousands of individual audio *samples*. Hence, naive adoption of standard sequence models like Transformers [22]—i.e., treating each sample as a sequence timestep—would only allow for modeling fractions of a second of audio due to memory constraints. Other work has proposed specialized architectures which do scale to longer durations [20, 2], but generating from these models is inefficient—they produce audio one sample at a time, and can take minutes to generate a few seconds of audio.

My work was the first to show that generative adversarial networks (GANs) [15] can be used to generate audio waveforms [10]. GANs are a natural fit for audio generation because they can efficiently synthesize an entire waveform in parallel. However, I found empirically that insights from signal processing were needed to match the audio fidelity and semantic coherence of slower methods. For example, I observed that modeling audio *spectrograms*—time-frequency representations extracted through signal processing—resulted in improved coherence but worse fidelity compared to modeling raw audio. In a later paper [13], I demonstrated that modeling a different spectrogram formulation originally engineered for audio manipulation [5] yields efficient generation with competitive fidelity and coherence. Others have since built on top of my work to develop production-grade audio generation systems [18, 12], which now power commercial applications.[5]

**Symbolic data.** Music manifests both as acoustic waveforms and symbolic scores—my work involves generative modeling of both modalities. Like language, symbolic scores can be represented as sequences of discrete data. Hence, my work on generative modeling of scores incorporates cutting-edge techniques from natural language processing (NLP). For example, my work demonstrates that Transformers [22] can be useful for generating multi-instrumental music [9]. I also work directly on NLP, not only because of the methodological synergy with modeling scores, but also because music often incorporates language in the form of lyrics. Conventional strategies for language generation only consider past context—my work on generation via *infilling* [6] overcomes this key limitation by considering both past and future context, i.e., filling in the blanks in incomplete text. I explored creative domains like lyrics and stories in this work, where infilling may be particularly useful for connecting the dots in disjoint raw material provided by users.

[4] This work was a **best paper runner-up at ISMIR 2021** (top three papers out of 200+ submissions).

[5] Audio GANs now power audio synthesis in *Descript* (link)—a popular podcasting tool—and emerging tools in music (link).

## 3 Future directions

My work thus far demonstrates that it is possible to build systems that enable a broader audience to engage more deeply with music. Looking forward, I envision a paradigm of *universal music expression*, where anyone with any level of expertise can express their musical intuition with ease. In such a paradigm, non-musicians could create or manipulate music as easily as they can currently compose or edit text, while musicians could enjoy seamless high- and low-level control. Towards this goal, my future work will make progress on several directions: (1) "decompiling" music, (2) improving generative models, and (3) adapting control mechanisms to user preferences. Along the way, I will continue releasing real-world systems that directly benefit users. In parallel, I plan to explore broad questions about creativity through the lens of music.

**Decompiling music.** Music audio can be thought of as a "compiled" version of its underlying "source code", i.e., the score (rhythm, notes, lyrics) and sound characteristics (instruments, timbre) which gave rise to the audio. The ability to "decompile" a waveform back into these components would have numerous benefits. In the short-term, it would aid music understanding and informatics. In the long-term, it could enable preference-tailored music education, content-based recommendations, or create training data for *factorized generation*. Factorized music generation—e.g., first generating a score and then generating audio given the score—may be a crucial component of tools for universal music expression, as it mirrors the workflows of human experts. By transcribing melody and harmony, my work on Sheet Sage is an important step towards music decompilation, but many attributes (e.g., rhythm, timbre) remain locked in opaque waveforms.

**High-fidelity, long-range, multimodal generative models.** Despite decades of research, we have barely scratched the surface of generative modeling of music. State-of-the-art models of music audio [4, 3] are *low-fidelity*—they are rife with noisy artifacts—preventing them from being deployed in music production tools where the bar for sound quality is high. Models of music scores [17, 9] fail to learn *long-range* phenomena like song structure and repetition. Past work has yet to confront the duality of audio and scores: existing models are trained myopically on one or the other. Moreover, we have limited understanding of the high-level relationships between music and other modalities like lyrics and video. Hence, generative models of music must overcome a multitude of remaining obstacles before they can meaningfully aid human expression.

**Intuitive creation via adaptive control.** Intuitive control mechanisms are critical for facilitating universal music expression. My past work on Piano Genie [11] enables non-musicians to improvise, but this system learned a one-size-fits-all mapping from a particular controller to a single instrument—a user has no recourse if they dislike the learned mapping. Instead, it would be preferable to have systems which can *adapt* to a user's control preferences. An ideal system would allow the use of *any* control mechanism to play *any* instrument, automatically providing an intuitive mapping that the user could then refine to taste. This idea of adaptive control connects to emerging trends in robotics [19]—I envision this line of work growing beyond music, into other areas where controlling complex processes with simple interfaces is desirable.

**Music as a test bed for studying creativity.** Research in music technology naturally gives rise to broad questions about human creativity and the role that technology might play therein. How can we build models which extrapolate as remarkably as humans do, e.g., ones which can invent entirely new genres? How might models and interfaces expand the breadth of music which humans create, rather than homogenizing it? In what ways is musical creativity similar to or distinct from creativity in other domains? While ML has made remarkable progress in grounded domains like language and vision, we are just beginning to make progress in more abstract domains—music can serve as a rich test bed for exploring hypotheses about creativity.

**Conclusion.** Centuries of music technology have ingrained a dichotomy of musicians and non-musicians. Blurring these lines requires new ML methodology to handle the idiosyncrasies and multimodal nature of music. My future work seeks to allow anyone to experience the boundless satisfaction of music expression.

## References

[1] R. Castellon*, C. Donahue*, and P. Liang. Codified audio language modeling learns useful representations for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[2] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv:1904.10509*, 2019.

[3] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. Jukebox: A generative model for music. *arXiv:2005.00341*, 2020.

[4] S. Dieleman, A. v. d. Oord, and K. Simonyan. The challenge of realistic music generation: modelling raw audio at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[5] M. Dolson. The phase vocoder: A tutorial. *Computer Music Journal (CMJ)*, 1986.

[6] C. Donahue, M. Lee, and P. Liang. Enabling language models to fill in the blanks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[7] C. Donahue and P. Liang. Sheet Sage: Lead sheets from music audio. In *Late-breaking Demos at the International Society for Music Information Retrieval Conference (ISMIR LBD)*, 2021.

[8] C. Donahue, Z. C. Lipton, and J. McAuley. Dance Dance Convolution. In *International Conference on Machine Learning (ICML)*, 2017.

[9] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. McAuley. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[10] C. Donahue, J. McAuley, and M. Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[11] C. Donahue, I. Simon, and S. Dieleman. Piano Genie. In *International Conference on Intelligent User Interfaces (IUI)*, 2019.

[12] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan. End-to-end adversarial text-to-speech. In *International Conference on Learning Representations (ICLR)*, 2021.

[13] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts. GANSynth: Adversarial neural audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[14] R. Fiebrink, D. Trueman, P. R. Cook, et al. A meta-instrument for interactive, on-the-fly machine learning. In *International Conference on New Interfaces for Musical Expression (NIME)*, 2009.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[16] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel. Sequence-to-sequence piano transcription with transformers. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[17] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. Music Transformer: Generating music with long-term structure. In *International Conference on Learning Representations (ICLR)*, 2019.

[18] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville. MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[19] D. P. Losey, K. Srinivasan, A. Mandlekar, A. Garg, and D. Sadigh. Controlling assistive robots with learned latent actions. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[20] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv:1609.03499*, 2016.

[21] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

* indicates equal contribution.