My research goal is to **develop and responsibly deploy generative AI for music and creativity, thereby unlocking and augmenting human creative potential**. Generative AI is poised to usher in a new paradigm of creative expression, one that both dramatically lowers the barrier to entry for creative pursuits while simultaneously presenting new avenues for artistic exploration. My work centers around music as a case study for generative AI and creativity, and involves (1) improving **machine learning** (ML) methods for controllable generative AI for music, audio, and other sequential data, and (2) deploying real-world interactive systems that allow a broader audience—inclusive of non-musicians—to harness generative music AI through intuitive controls.

With a broader set of users 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 to hum the melody of our favorite song. Some of us may also desire to exercise this intuition by creating music. But conventional tools for musical expression (e.g., instruments, notation, software) require years of training, forming an *expertise divide* that 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 control inputs into low-level notes, sacrificing the fine-grained control that musicians enjoy but dramatically increasing the accessibility of musical expression. As a concrete example, a system I built called *Piano Genie* [1] uses novel ML methods to convert high-level control from novices into realistic piano improvisations,[1] and has been used by a professional rock band to blur the lines between the audience and the performers (Figure 1).

Another theme of my work is **enabling richer forms of control for generative AI**. Natural language is increasingly regarded as a universal interface for generative AI, and there has been a corresponding explosion of recent interest in training models that convert text into music [2, 3, 4, 5]. For music generation, language is convenient for controlling global style (e.g., genre, instruments), but musicians and non-musicians alike would benefit from the addition of richer forms of control over time-varying attributes (e.g., notes, rhythms). For example, I recently proposed a system called *SingSong* [6] that generates audio accompaniments for input singing, allowing creators to use their voice as a form of control for music generation. SingSong was recently incorporated into Google's *Music AI Tools* and is being responsibly deployed to professional musicians for use in co-creation settings.[2] With a goal of better supporting realistic co-creation workflows both for music and language, I have separately proposed novel ML methods that enable large language models (LLMs) to fill in the blanks of creative works-in-progress [7]. This work has had broad impact on generative AI—a recent paper from OpenAI recommends training all LLMs with my proposed method by default [8].

While generative AI promises extraordinary creative potential for music, realizing this potential requires both new ML methodology and broad interdisciplinary perspectives. One domain-specific ML challenge is that music audio sequences are so high-dimensional that they cannot be tractably processed by standard sequence models like Transformers [9]. Modeling audio thus demands new methods: my work was the first to show that adversarial learning can be used to generate audio [10], a finding that underpins the audio tokenizers [11, 12] found in modern audio LLMs [13, 5]. Another challenge lies in understanding the multimodal nature of music—my work demonstrates that music *foundation models* [14] can help us model the relationship between music in its acoustic and symbolic forms [15, 16]. Moreover, while my past work centered around the *development* of generative music AI, my current and future work is shifting towards a more holistic view including its responsible *deployment*, which will involve coordination with diverse stakeholders including musicians, legal experts, and social scientists.

1 Improving machine learning methods for controllable generative AI

My work makes core contributions to ML methodology necessary to enable broadly capable and controllable generative AI for music, audio, and other sequential data like language. Of particular note are my contribu-

[1] Piano Genie example: https://youtu.be/YRb0XAnUpIk, demo: https://chrisdonahue.com/piano-genie
[2] Google's Music AI Tools: https://deepmind.google/discover/blog/transforming-the-future-of-music-creation

tions towards (1) methods for audio generation, (2) methods allowing LLMs to better support co-creation for language and music, and (3) the harnessing of foundation models for music understanding and control.

New ML methods for audio generation. My research has played a pivotal role in the development of modern ML methods for audio generation. The potential impact of audio generation extends far beyond music, encompassing areas such as speech synthesis and sound design. However, audio waveforms have proven to be empirically difficult to model with standard ML approaches. One challenge is that waveforms are extraordinarily high-dimensional: a single second of audio contains tens of thousands of individual audio samples. While Transformers [9] are broadly capable of modeling sequential data, their computational costs scale quadratically with sequence length, making them intractable for waveform modeling and exposing an Achilles' heel for ML methods. My work was the first to show that generative adversarial networks (GANs) [17] can be used to generate audio waveforms [10]. **My work on audio GANs continues to underpin audio tokenizers [13, 5], a key component of state-of-the-art hierarchical generative audio models based on LLMs [11, 5]**. While hierarchical approaches are necessary in the short term to achieve long-term consistency for audio generation with Transformers, my work has also recently proposed an alternative architecture [18] based on structured state space models [19]. Unlike Transformers, the computational costs for our proposed method scale linearly with sequence length, a concrete step towards a future of direct (non-hierarchical) modeling of audio waveforms.

Infilling with LLMs for co-creation. LLMs increasingly constitute a general purpose methodological foundation for generative modeling of broad modalities [14] including music [20, 21, 22, 3, 5, 23]. However, LLMs suffer from a key drawback in human-AI co-creation settings both for music and language: by default, they generate left-to-right, taking into account past context (the *prefix*) when generating the *middle* but ignoring any future context (the *suffix*). To support non-sequential workflows, creators would prefer models capable of *infilling*, i.e., ones which can fill in gaps of information to connect a prefix and suffix. My work [7] proposes a simple method to imbue language models with this capability based on reordering training examples from (prefix, middle, suffix) to (prefix, suffix, middle). By mixing in reordered examples with standard ones, my work shows that we can add the capability to infill without sacrificing performance on standard left-to-right generation. This simple idea has had broad impact—**a recent paper from OpenAI recommends training all LLMs with my proposed method by default [8]**. One empirical challenge with this method in the domain of music is that the suffix can often be sufficiently long as to evict the prefix from the LLM's limited context window. To sidestep this issue, my work proposed a novel method [23] that allows LLMs to consider some fixed portion of the suffix (as opposed to the entire suffix), enabling rich infilling capabilities in the domain of symbolic music.

Harnessing music foundation models. My recent work leverages music foundation models for music information retrieval (MIR) and controllable generation. The use of foundation models, models pre-trained on large quantities of data at scale and then adapted to different tasks, increasingly constitutes a paradigm shift in ML methodology across research areas. My work was the first to propose adopting this paradigm for MIR research. Specifically, my work shows that Jukebox [22], a generative model of music audio pre-trained on 1m songs, learns powerful representations useful for a variety of downstream MIR tasks such as genre detection and emotion recognition [15]—this paper was a runner-up for best paper at ISMIR 2021 (top three papers out of over 200 submissions). I later proposed a method for using these representations in time-varying prediction tasks, dramatically improving the state-of-the-art for pop music transcription [16] and moving the needle on our ability to reason about music across its dual acoustic and symbolic forms. The limited availability of labeled data has long been a bottleneck to progress in MIR, and leveraging foundation models represents a promising path towards broad performance improvements. In addition to music understanding, MIR models can also be used to create paired (control, audio) data for training controllable generative models [6]. Along this line, my student recently proposed an efficient strategy for adapting generative music foundation models to have richer forms of control, using control data extracted by MIR models [24].

2 Deploying real-world generative music AI systems for a broader audience

It is my belief that the intrinsic value of any new creative technology is determined not by the quantitative performance of its underlying methods, but instead by the degree to which the technology engenders creativity in the hands of users. With this belief in mind, my work takes a holistic view beyond developing methods and involves building and deploying generative music AI systems in the real world. Here I provide an overview of instances where my research has directly translated into real-world impact.

Bridging the expertise divide. Musical creativity is inaccessible to most of us due to the high barrier to entry of conventional tools, creating an *expertise divide* between musicians and non-musicians. My work bridges this divide by developing generative AI that maps high-level, novice-friendly control inputs into music. One example of this is my work on Piano Genie [1], which showed for the first time that generative AI can map intuitive control to music in real-time, 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. A more recent example is my work on SingSong [6], an LLM-based system that takes singing as input and generates corresponding audio accompaniments in broad styles. SingSong unlocks rich expressive capabilities, allowing musicians and non-musicians alike to create music using only their voice. This system has been incorporated into Google DeepMind's Music AI Tools, and is currently being explored by major recording artists in real-world creative settings (see footnote on first page).

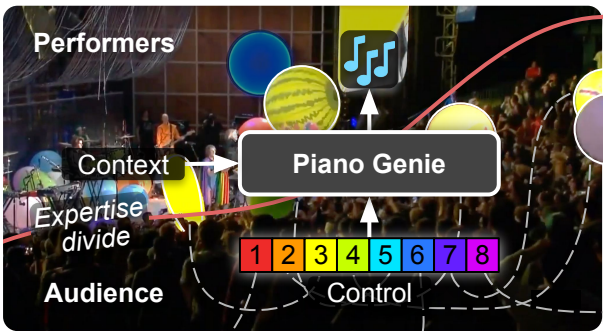

Figure 1: The Flaming Lips used my work to invite the audience into their show. The crowd struck beach balls to control the 8 buttons of Piano Genie, which was also conditioned on performance context.

In addition to enabling non-musicians to create, generative AI 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 context of the band's performance (specifically, the chords), thereby steering 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.

Connecting music with other modalities. Music is inherently *multimodal*: music itself can be acoustic or symbolic, and music can also be paired with other modalities in a *cross-modal* fashion. Such cross-modal pairings can heighten our appreciation of both music and its companion modality, e.g., music can increase our appreciation of films and vice versa. However, while cross-modal relationships may seem abstract, in reality we have strong intuitions about them, and an unreasonable pairing may detract from the experience. My work uses ML to model cross-modal relationships, giving rise to rich interactions. In 2017, I demonstrated that ML can model the relationship between music and the physical movements found in rhythm-based video games [25]. To make this work widely accessible, I built a free live service called Beat Sage which uses the same approach to create content for the popular VR rhythm game *Beat Saber*.[4] Beat Sage appeals to a broad audience—it can produce interactive content for any song and accommodates players of any skill level. Since 2020, Beat Sage has generated over four million levels, and is still used daily by thousands of users.

[3] Additional context about The Flaming Lips' performance with Piano Genie: https://magenta.tensorflow.org/fruitgenie
[4] Beat Sage example: https://youtu.be/nDZ61cRBhzU?t=185, live service: https://beatsage.com

3 Future directions

My future research will broadly explore the intersection of generative AI and creativity, with music as a primary (but not an exclusive) focus. Additionally, my research will adapt with respect to both an ongoing paradigm shift in ML and the emerging societal impact of generative AI. Research in generative AI increasingly revolves around the pre-training, adaptation, and deployment of foundation models [14]. My future work will focus on closing the gap between the raw potential of foundation models and real-world creative needs. Within music, generative AI is just now beginning to have a substantive impact. In the past year we have seen glimpses of extraordinary generative capabilities [2, 3, 6, 5] alongside troubling use cases posing serious ethical questions [26]. To help ensure that generative AI has a positive impact on society, my work will increasingly focus on responsible development and deployment of models in creative contexts.

Generalized control. Current music foundation models offer limited forms of control, hindering their usefulness to creators. As a research community, we are currently adding individual controls to models in a "whack-a-mole" fashion, gathering paired (control, music) data and using supervised learning to generate music from e.g., text [3], melody [5], and singing [6]. My future work will pursue a more general paradigm of control reminiscent of in-context learning in LLMs [27], one where creators can use prompting to control models with a mixture of suitable modalities, e.g., "generate me a `<Recording>` of `<Score Image>` in the style of `<Another Recording>` and aligned with `<Vacation Video>`". Concrete progress towards this goal is possible at academic compute scales by, for example, (1) codifying this goal into instruction tuning datasets [28], or (2) proposing new methods to combine unimodal models into multimodal ones. This thrust is broadly applicable beyond music, as multimodal prompting remains a nascent research direction.

Composable outputs. In addition to the input side, music foundation models also fall short of real-world needs on the output side. A key issue is that models mostly output audio *mixtures*, which can be thought of as the "compiled" version of the underlying musical "source code" (e.g., symbolic notation, individual instrument recordings). Consequently, mixtures constitute a creative dead end as they cannot be easily manipulated or refined. My future work will seek to improve the *composability* of outputs, either through direct generation of the "source code", or by learning to "decompile" mixtures (thereby improving MIR). Additionally, current models suffer from poor audio fidelity and weak global consistency. My work will also focus on increasing quality, e.g., by improving audio tokenizers and exploring new architectures that can support long sequences, and on improving the reproducibility of evaluation for music and audio generation.

Creative co-pilots. Generative AI tools like GitHub Co-pilot have transformed the everyday workflows of programmers, but creators have yet to see analogous benefits. My future work will seek to bring generative AI into real-world creative workflows, including but not limited to music. This will involve collaborating with human-computer interaction researchers to explore effective co-creation workflows across user skill levels, optimizing not only for productivity but also for creative satisfaction. Once we have co-pilots deployed in the real world, rich research directions emerge around the improvement and personalization of models from human interaction data. Especially in the context of music, this will also involve systems research to enable users to run large models on device and in real-time, addressing potential latency and privacy concerns.

Generative music AI and society. Generative AI is starting to have a real impact on music as a creative, economic, and cultural activity. My future work will continue to focus on the development of generative AI but also broaden to help ensure that this technology is deployed in a manner benefiting society. With a societal lens in mind, *provenance* will be a particular technical focus of my future work—how do we understand the relationship between a model's outputs and its training data? An improved understanding of provenance could be a critical part of the ethical and policy conversations around generative AI as a whole. More broadly, I plan to collaborate with diverse stakeholders (e.g., musicians, cultural representatives, educators, legal scholars, social scientists) to more effectively anticipate societal risks posed by generative AI and identify scenarios where technical interventions can potentially improve outcomes.

References

[1] **Chris Donahue**, Ian Simon, and Sander Dieleman. Piano Genie. In *International Conference on Intelligent User Interfaces (IUI)*, 2019.

[2] Seth Forsgren and Hayk Martiros. Riffusion - Stable diffusion for real-time music generation, 2022.

[3] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. MusicLM: Generating music from text. *arXiv:2301.11325*, 2023.

[4] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning (ICML)*, 2023.

[5] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[6] **Chris Donahue**, Antoine Caillon, Adam Roberts, Ethan Manilow, Philippe Esling, Andrea Agostinelli, Mauro Verzetti, Ian Simon, Olivier Pietquin, Neil Zeghidour, et al. SingSong: Generating musical accompaniments from singing. *arXiv:2301.12662*, 2023.

[7] **Chris Donahue**, Mina Lee, and Percy Liang. Enabling language models to fill in the blanks. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[8] Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle. *arXiv:2207.14255*, 2022.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[10] **Chris Donahue**, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[11] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[12] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

[13] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. AudioLM: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[14] Rishi Bommasani, ... many other authors ..., **Chris Donahue**, ... many other authors ..., and Percy Liang. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

[15] Rodrigo Castellon*, **Chris Donahue***, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[16] **Chris Donahue**, John Thickstun, and Percy Liang. Melody transcription via generative pre-training. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[18] Karan Goel, Albert Gu, **Chris Donahue**, and Christopher Ré. It's raw! audio generation with state-space models. In *International Conference on Machine Learning (ICML)*, 2022.

[19] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

[20] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music Transformer: Generating music with long-term structure. In *International Conference on Learning Representations (ICLR)*, 2019.

[21] **Chris Donahue**, Huanru Henry Mao, Yiting Ethan Li, Garrison W Cottrell, and Julian McAuley. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[22] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv:2005.00341*, 2020.

[23] John Thickstun, David Hall, **Chris Donahue**, and Percy Liang. Anticipatory music transformer. *arXiv:2306.08620*, 2023.

[24] Shih-Lun Wu, **Chris Donahue**, Shinji Watanabe, and Nicholas J Bryan. Music ControlNet: Multiple time-varying controls for music generation. *arXiv:2311.07069*, 2023.

[25] **Chris Donahue**, Zachary C Lipton, and Julian McAuley. Dance Dance Convolution. In *International Conference on Machine Learning (ICML)*, 2017.

[26] Michael Feffer, Zachary C. Lipton, and **Chris Donahue**. DeepDrake ft. BTS-GAN and TayloRVC: An exploratory analysis of musical deepfakes and hosting platforms. In *Workshop on Human-Centered Music Information Retrieval (HCMIR)*, 2023.

[27] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[28] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022.

* indicates equal contribution.