A. Dataset Creation Details

The datasets were built by first searching the web for publicly-available MIDI files, resulting in ≈1.5 million unique files. We removed those that were identified as having a non-4/4 time signature and used the encoded tempo to determine bar boundaries, quantizing to 16 notes per bar (16th notes).

For the 2-bar (16-bar) drum patterns, we used a 2-bar (16-bar) sliding window (with a stride of 1 bar) to extract all unique drum sequences (channel 10) with at most a single bar of consecutive rests, resulting in 3.8 million (11.4 million) examples.

For 2-bar (16-bar) melodies, we used a 2-bar (16-bar) sliding window (with a stride of 1 bar) to extract all unique monophonic sequences with at most a single bar of consecutive rests, resulting in 28.0 million (19.5 million) unique examples.

For the trio data, we used a 16-bar sliding window (with a stride of 1 bar) to extract all unique sequences containing an instrument with a program number in the piano, chromatic percussion, organ, or guitar interval, [0, 31], one in the bass interval, [32, 39], and one that is a drum (channel 10), with at most a single bar of consecutive rests in any instrument. If there were multiple instruments in any of the three categories, we took the cross product to consider all possible combinations. This resulted in 9.4 million examples.

In all cases, we reserved a held-out evaluation set of examples which we use to report reconstruction accuracy, interpolation results, etc.

B. Lakh MIDI Dataset Results

For easier comparison, we also trained our 16-bar models on the publicly available Lakh MIDI Dataset (LMD) (Raffel, 2016), which makes up a subset of the our dataset described above. We extracted 3.7 million melodies, 4.6 million drum patterns, and 116 thousand trios from the full LMD. The models were trained with the same hyperparameters as were used for the full dataset.

We first evaluated the LMD-trained melody model on a subset of the full evaluation set made by excluding any examples in the LMD train set. We found less than a 1% difference in reconstruction accuracies between the LMD-trained and original model.

In Table 2 we report the reconstruction accuracies for all 3 16-bar models trained and evaluated on LMD. While the accuracies are slightly higher than Table 1, the same conclusions regarding the relative performance of the models hold.

	Teacher-Forcing		Sampling	
Model	Flat	Hierarchical	Flat	Hierarchical
16-bar Melody	0.952	**0.956**	0.685	**0.867**
16-bar Drum	0.937	**0.955**	0.794	**0.908**
Trio (Melody)	0.866	**0.868**	0.660	**0.760**
Trio (Bass)	0.906	**0.912**	0.651	**0.782**
Trio (Drums)	0.943	**0.946**	0.641	**0.895**

Table 2. Reconstruction accuracies for the Lakh MIDI Dataset calculated both with teacher-forcing (i.e., next-step prediction) and full sampling. All values are reported on a held-out test set. A softmax temperature of 1.0 was used in all cases, meaning we sampled directly from the logits.

C. Attribute Definitions

The following definitions were used to measure the amount of each attribute.

C Diatonic

The fraction of notes in the note sequence whose pitches lay in the diatonic scale on C (A-B-C-D-E-F-G, i.e., the "white keys").

Note Density

The number of note onsets in the sequence divided by the total length of the sequence measured in 16th note steps.

Average Interval

The mean absolute pitch interval between consecutive notes in a sequence.

16th Note Syncopation

The fraction of (16th note) quantized note onsets landing on an odd 16th note position (1-indexed) with no note onset at the previous 16th note position.

8th Note Syncopation

The fraction of (16th note) quantized note onsets landing on an odd 8th note position (1-indexed) with no note onset at either the previous 16th or 8th note positions.

D. Audio Samples

Synthesized audio for all examples here and in the main text can be found in the online supplement.[5]

[5] https://goo.gl/magenta/musicvae-examples

E. Additional Figures and Samples

Subsequent pages include additional figures, referenced from the main text.

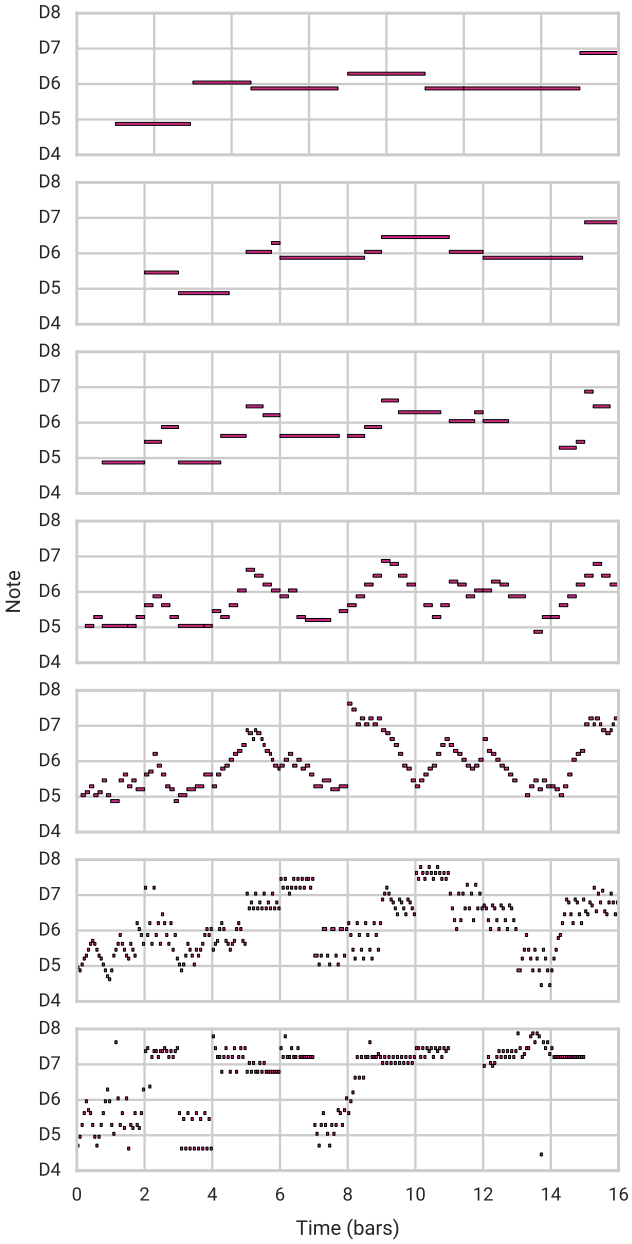

Figure 6. Varying the amount of the "Note Density" attribute vector. The amount varies from -1.5 to 1.5 in steps of 0.5, with the central sequence corresponding to no attribute vector. Audio for this example is available in the online supplement.[5]

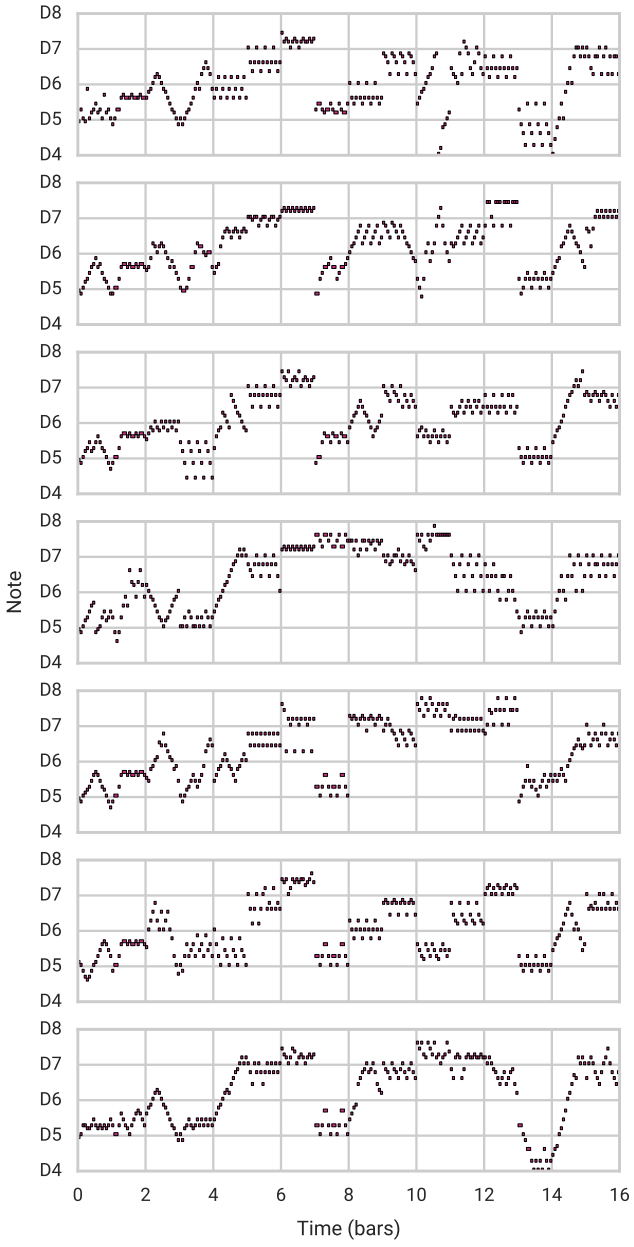

Figure 7. Additional resamplings of the same latent code (corresponding to the second-to-the-bottom in Fig. 6). While semantically similar, the specific notes vary due to the sampling in the autoregressive decoder. Audio for this example is available in the online supplement.[5]

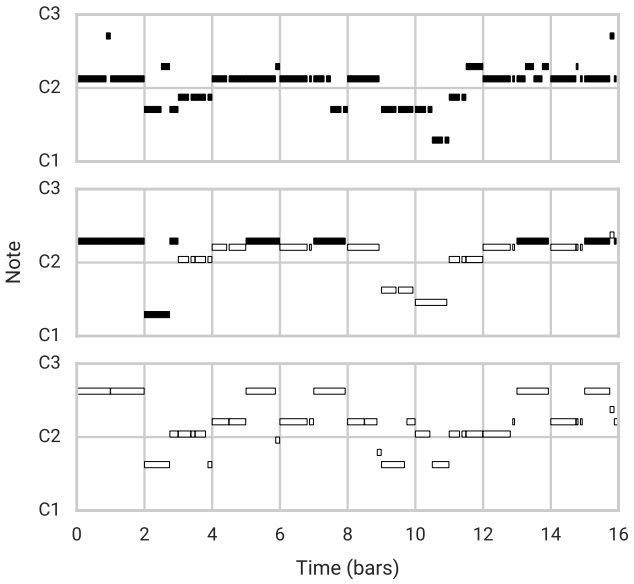

Figure 8. Subtracting (top) and adding (bottom) the "C Diatonic" attribute vector from the note sequence in the middle. For ease of interpretation, notes in the C diatonic scale are shown in white and notes outside the scale are shown in black. Audio for this example is available in the online supplement.[5]

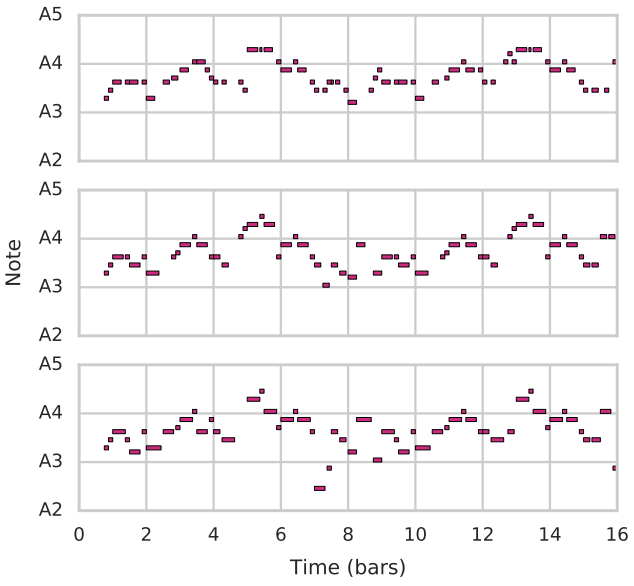

Figure 9. Subtracting (top) and adding (bottom) the "Average Interval" attribute vector from the note sequence shown in the middle. Audio for this example is available in the online supplement.[5]

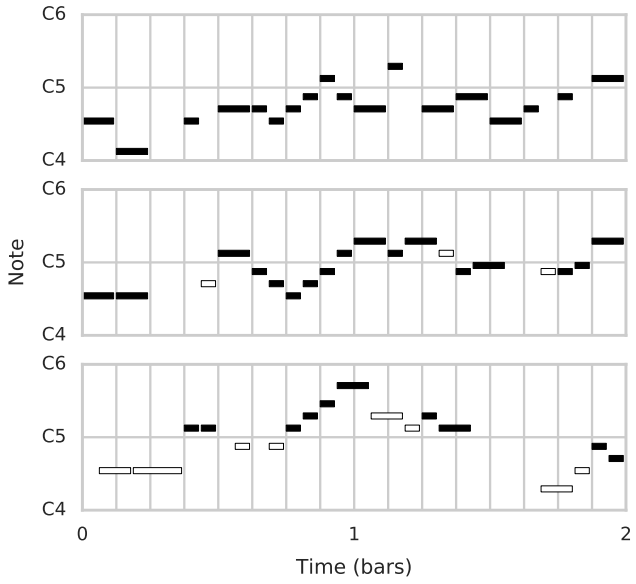

Figure 10. Subtracting (top) and adding (bottom) the "16th Note Syncopation" attribute vector from the note sequence in the middle. For ease of interpretation, only the first 2 of each sequence's 16 bars are shown. Vertical lines indicate 8th note boundaries. White and black indicate syncopated and non-syncopated notes, respectively. Audio for this example is available in the online supplement.[5]

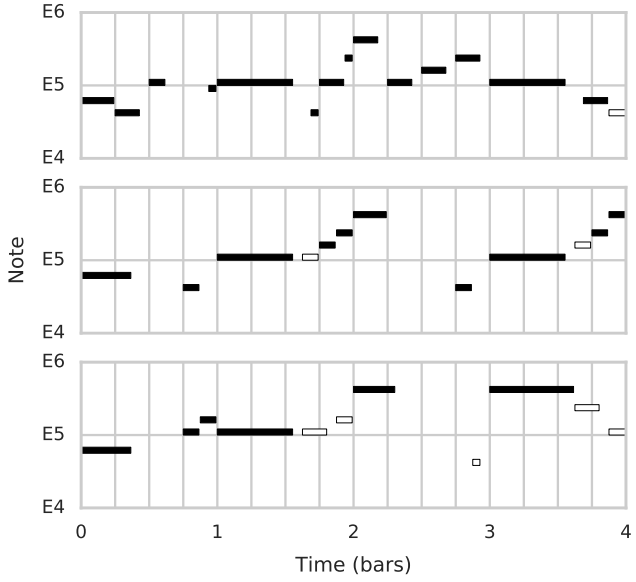

Figure 11. Subtracting (top) and adding (bottom) the "8th Note Syncopation" attribute vector from the note sequence in the middle. For ease of interpretation, only the first 4 of each sequence's 16 bars are shown. Vertical lines indicate quarter note boundaries. White and black indicate syncopated and non-syncopated notes, respectively. Audio for this example is available in the online supplement.[5]

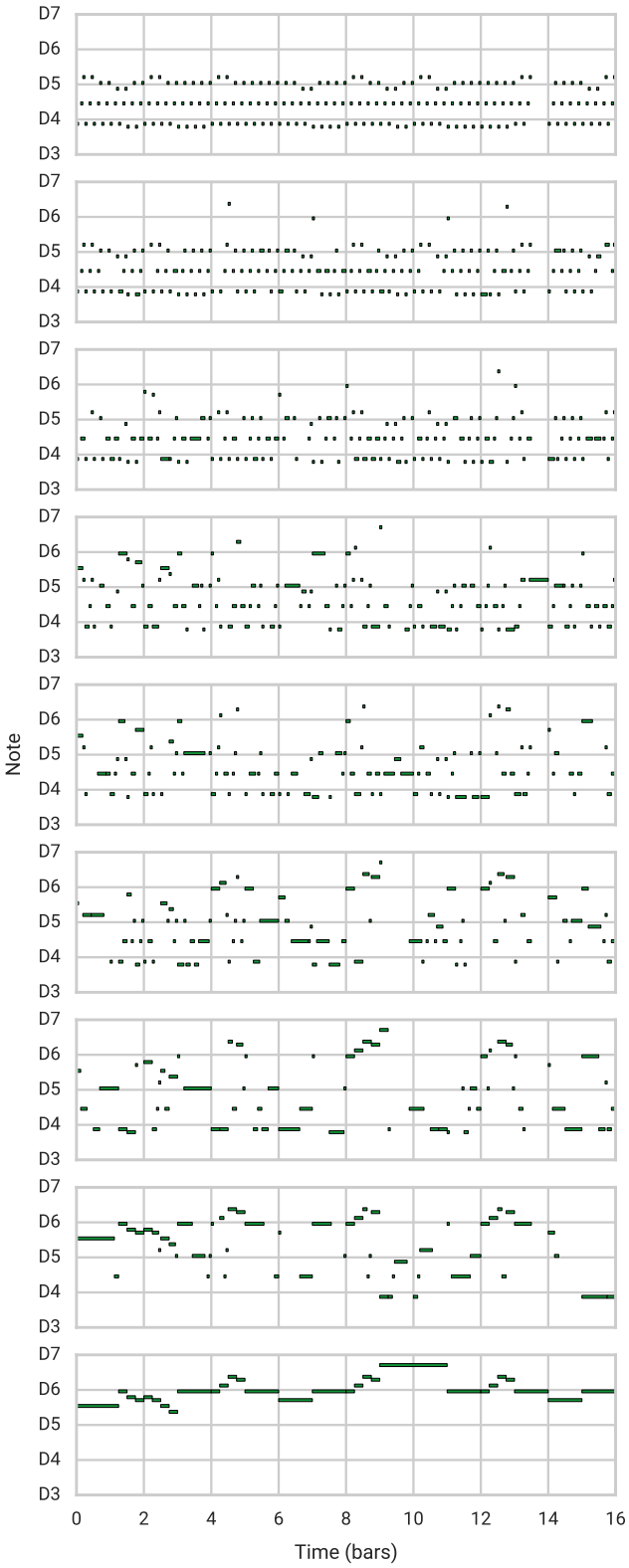

Figure 12. Interpolating between the top and bottom sequence in data space. Audio for this example is available in the online supplement.[5]

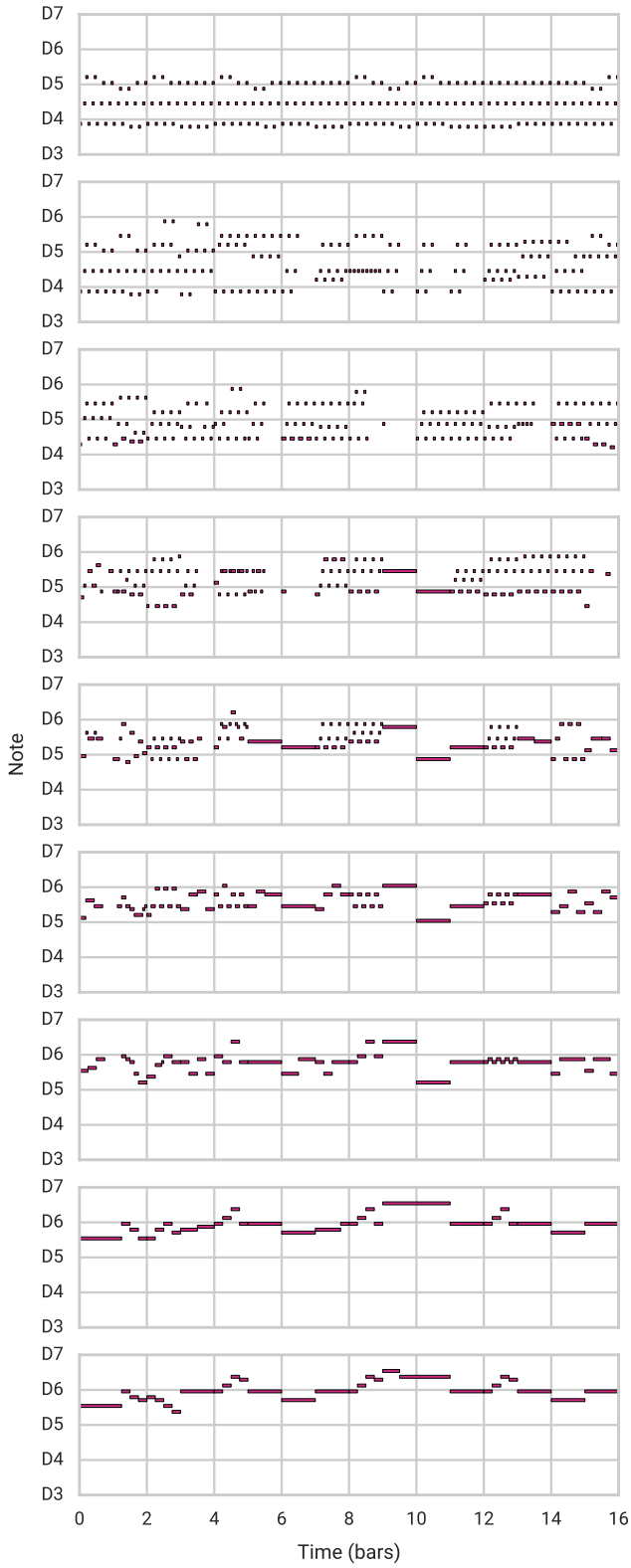

Figure 13. Interpolating between the top and bottom sequence (same as Fig. 12) in MusicVAE's latent space. Audio for this example is available in the online supplement.[5]

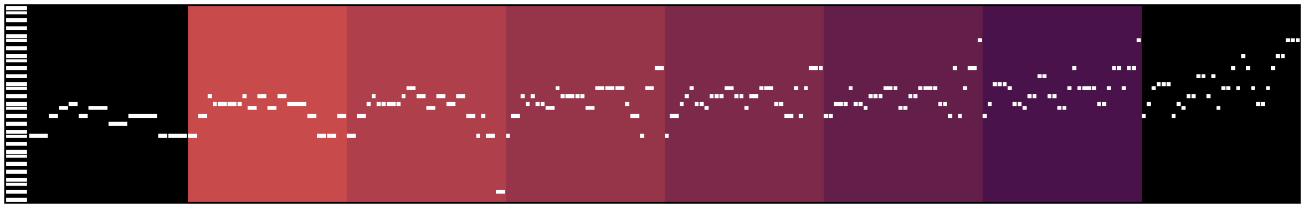

Figure 14. Example interpolation in the 2-bar melody MusicVAE latent space. Vertical axis is pitch (from A_3 to C_8) and horizontal axis is time. We sampled 6 interpolated sequences between two test-set sequences on the left and right ends. Each 2-bar sample is shown with a different background color. Audio of an extended, 13-step interpolation between these sequences is available in the online supplement.[5]

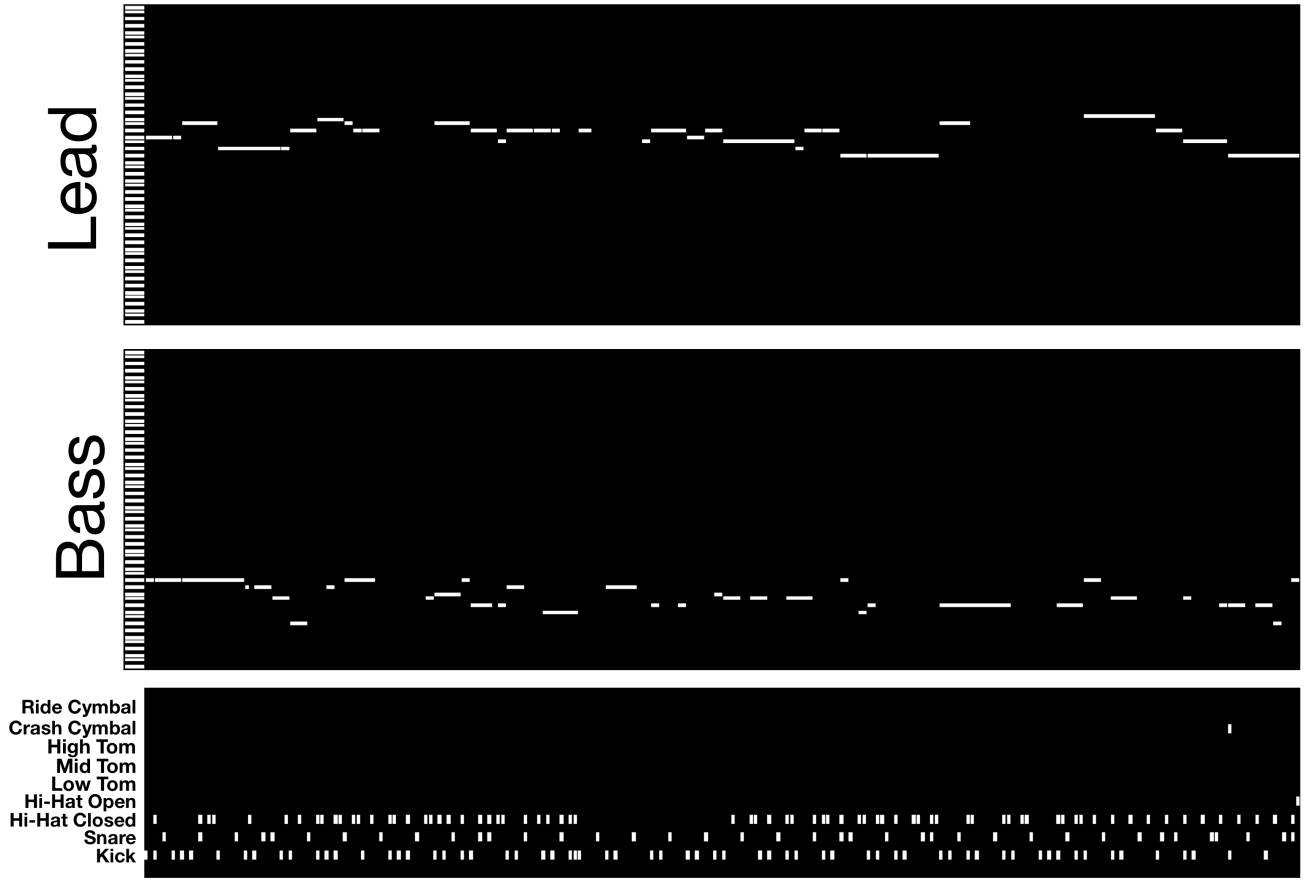

Figure 15. Selected example 16-bar trio sample generated by MusicVAE. Audio for this and other samples is available in the online supplement.[5]

.