Find the Neurons, Control the Feature? Selection Strategies for Network Bending Audio Models

Louis McCallum*
CCI
University of the Arts, London
London, UK
l.mccallum@arts.ac.uk

Mick Grierson
CCI
University of the Arts, London
London, UK
m.grierson@arts.ac.uk

Abstract

Neural audio synthesisers like RAVE enable high-quality generation, but designing effective control interfaces requires understanding where musical features are encoded and whether those encodings can be manipulated. We present evidence that these questions yield different answers: neurons that "know about" a feature are not necessarily the neurons that control it. Through analysis of three RAVE models, we identify functionally-coherent neuron groups that encode BPM more strongly than individual layers. However, manipulation experiments across 288,000 samples find little compelling evidence for a relationship between control and encoding (all $|\rho| \leq 0.16$, n.s.). Encoding strength instead predicts effect magnitude, with the direction feature-dependent. We find strongly BPM-encoding groups produce larger effects (cross-layer clusters $\rho = +0.24$; layer-wise $\rho = +0.50$), while strongly pitch-encoding layers produce smaller effects (layer-wise $\rho = -0.28$). Scaling whole layers achieves superior control, emerging as the best method across over $6\times$ as many feature/section combinations as clustering, whereas clusters provide nearly $4\times$ better efficiency and stronger magnitude.

1 INTRODUCTION

Recent advances in neural audio synthesis have enabled real-time, high-fidelity music generation through models like RAVE (Caillon and Esling, 2021), EnCodec (Défossez et al., 2022), and SoundStream (Zeghidour et al., 2021). These variational autoencoder (VAE)-based architectures learn compressed latent representations that can reconstruct audio with perceptual quality rivalling traditional synthesis methods. For musical performance, these models offer exciting possibilities, however, understanding how to effectively manipulate their internal representations for creative control remains largely underinvestigated.

A natural approach is to first identify where musical features are encoded in the network, then target those locations for manipulation. If we know which decoder layers or neurons represent pitch, tempo, or timbre, we could design control interfaces that directly modify those representations. This interpretability-first strategy has proven effective in computer vision (Bau et al., 2019; Härkönen et al., 2020), where identifying feature-encoding neurons enables targeted manipulation.

However, this approach rests on the assumption that understanding where features are encoded helps us to control them. We test this assumption through systematic analysis of RAVE decoder activations, followed by large-scale manipulation experiments.

This paper addresses three research questions:

RQ1: Does interpretability analysis (identifying feature encodings) predict which neurons to target for manipulation?

RQ2: What neuron selection strategies provide effective control over audio features in neural synthesiser network bending?

RQ3: How do controllability, efficiency, and robustness trade off across selection methods and audio features?

We investigate these questions through two complementary studies. Study 1 examines feature encoding through layer-wise and cross-layer cluster analysis of three RAVE models (trained on strings, drums, and vocals) across four stimulus types. We identify where pitch and tempo are encoded by correlating feature change with activations and test whether cross-layer clustering can reveal distributed encodings that single-layer analysis misses. This provides grounding for **RQ1**.

Study 2 tests whether these encodings translate into effective manipulation targets through network bending experiments across 288,000 rendered samples, comparing layer-wise selection, cross-layer clustering, and size-matched random baselines to inform **RQ2**. To address **RQ3**, we measure controllability as how predictable the effect of bending is on a given audio feature, magnitude as the size of the effect in relation to user input and efficiency as the number of neurons needed to be bent make that change.

We contribute findings that:

- **Circuit coherence trumps neuron targeting.** Layer-wise selection provides strongest controllability for most features, suggesting that maintaining consistent relationships among co-activated neurons matters more than identifying feature-encoding subsets.
- **Encoding does not predict control, and its relationship with magnitude depends on feature** For both selection methods, encoding strength shows no meaningful correlation with controllability (all $|\rho| \leq 0.16$, n.s.). It does correlate with effect magnitude, but the sign

*code available at https://github.com/Louismac/network-bending-clusters/

varies: strongly BPM-encoding groups produce larger effects (cross-layer clustering $\rho = +0.24$, $p < 0.01$; layer-wise $\rho = +0.50$, $p < 0.001$), whereas strongly pitch-encoding layers produce smaller effects (layer-wise $\rho = -0.28$, $p < 0.01$). This suggests correlation-based methods identify neurons that carry feature information without being reliable control points.

- **Controllability and magnitude are largely independent.** The cluster-based methods provide higher magnitude effects across features and intervention points while layer-wise selection provides higher controllability. Practitioners must choose: reliable control or maximal effect, but rarely both from the same subset.
- **Controllability and efficiency trade off.** Sparse cluster-based interventions achieve greater effect per neuron in late layers but less predictable control than dense layer-wise manipulation. Learned clusters identify high-leverage neurons, but this targeting does not translate to superior control.
- **Structured methods sacrifice robustness of effect size.** Controllability degrades similarly for all methods under distribution shift (Δ=-0.03 to -0.04), leaving layer-wise ahead in both conditions. Magnitude, however, degrades in proportion to how much structure a method assumes: cross-layer clustering loses 15% of its effect size out-of-distribution, random matched 9%, and layer-wise none.

2 RELATED WORK

2.1 Neural Audio Synthesis

Recent neural audio compression models learn compact latent representations enabling high-quality generation. SoundStream (Zeghidour et al., 2021) introduced residual vector quantisation, EnCodec (Défossez et al., 2022) improved perceptual quality, and RAVE (Caillon and Esling, 2021)—the focus of our analysis—achieves real-time synthesis through variational autoencoders with adversarial training. While reconstruction quality is impressive, internal representations remain underexplored.

Controllable generation has been pursued through structured latent spaces: DDSP (Engel et al., 2020) combines learned representations with differentiable signal processing; Music Transformer (Huang et al., 2018) and Jukebox (Dhariwal et al., 2020) use attention and hierarchical VQ-VAE; Universal Music Translation (Mor et al., 2018) enables style transfer via decoder manipulation. These approaches primarily operate in latent space rather than decoder activations. Our work examines how decoders encode features internally, enabling fine-grained control through targeted neuron manipulation.

2.2 Network Interpretability

Grouping neurons by functional similarity aids understanding of network organisation. SVCCA (Raghu et al., 2017) compares layer representations via canonical correlation, while Neural Activation Patterns (Bäuerle et al., 2022) groups inputs by activation profiles. Feature visualisation (Olah et al., 2017), deep visualisation (Yosinski et al., 2015), and network dissection (Bau et al., 2017) have established activation-based interpretation as central to interpretability research.

Work has formalised activation manipulation as causal intervention. GAN Dissection (Bau et al., 2019) identifies causal units through neuron clamping; causal mediation analysis (Vig et al., 2020) tests pathways in language models; Geiger et al. (Geiger et al., 2021) provide formal frameworks for causal abstractions. We adopt this perspective, using manipulation experiments to test whether correlation-identified neurons causally control musical features.

2.3 Creative Control and Network Bending

Latent space navigation enables creative control: GANSpace (Härkönen et al., 2020) discovers interpretable directions via PCA, InterFaceGAN (Shen et al., 2020) identifies attribute boundaries, and StyleGAN2 (Karras et al., 2020) demonstrates natural disentanglement. These operate on latent codes rather than internal activations.

Network bending—intentionally manipulating network internals for creative effect—has gained traction in visual art (Broad et al., 2021) but remains exploratory in audio (McCallum and Yee-King, 2020). Our work provides a principled foundation by identifying feature-encoding neurons and testing whether targeted manipulation achieves predicted effects.

3 STUDY 1: IDENTIFYING FEATURE-ENCODING NEURON SUBSETS

3.1 Overview

RAVE models are well-suited for investigating neural audio synthesis due to their comparative architectural simplicity (aiding activation analysis) and real-time execution (enabling expressive network bending). While these models converge best on homogeneous datasets with focused timbres, they can successfully train on polyphonic, multi-instrumental, arrhythmic, and even non-musical sounds. This specialisation makes broad generalisations challenging. To balance focused internal validation against reliable musical properties and externally valid real-world sounds, we analyse three models across four datasets, yielding 12 combinations that span natural and synthetic sounds, in- and out-of-distribution inputs, and a breadth of pitches and tempos.

First, we examine per-layer correlations between decoder activations and musical features, identifying baseline encoding strengths and testing for architecturally consistent encoding across models. Second, we perform cross-layer clustering to determine whether functionally-related neurons spanning multiple layers encode features more coherently than individual layers. In doing so, we are addressing the limitation that per-layer analysis may not be able to identify distributed functional groupings.

3.2 Method

3.2.1 Models and Stimuli

We analysed three RAVE models (44.1kHz, 128 latent dimensions, 28 decoder layers):

- **Strings**: Solo string instruments (violin, cello, viola) across various keys and BPMs
- **Drums**: Drum loops and percussion across various BPMs
- **Vocals**: Isolated female vocals from popular music (source-separated)

All models used identical architectures but different training data, allowing separation of shared architectural functions from domain-specific adaptation.

We created four stimulus types: (1) in-distribution natural audio from each model's training set, (2) out-of-distribution natural audio (training samples from other models), (3) synthetic pulse trains (4-second bursts at 60–180 BPM), and (4) synthetic sine tones (pure tones across 24–96 MIDI notes).

3.2.2 Feature Extraction

We extracted features at time scales matching their acoustic properties. This includes continuous fundamental frequency from a 100ms window of each 4-second segment, capturing monophonic single-note events while avoiding polyphonic content. Where accurate metadata was not available, we also conducted BPM estimation via librosa requiring longer windows to capture multiple beats and establish periodic patterns.

3.2.3 Per-Layer Correlation Analysis

For each audio segment, we: (1) encoded audio through RAVE, capturing activations at each decoder layer; (2) averaged activations spatially and temporally to obtain per-layer feature vectors; (3) computed Spearman correlations between layer activations and acoustic features; and (4) repeated across all stimuli to obtain correlation distributions. Throughout, we report absolute correlation coefficients $|\rho|$, as encoding strength is relevant regardless of direction.

3.2.4 Cross-Layer Clustering Analysis

Per-layer analysis treats each layer independently, potentially missing functionally-related neurons spanning multiple layers. To address this, we performed hierarchical cross-layer clustering.

We divided the 28 decoder layers into three sections based on architectural depth: early (layers 0–7), middle (layers 8–14), and late (layers 15+). This sectioning reflects both architectural structure (early layers process compressed latents, late layers generate waveforms) and our per-layer findings.

Within each section we applied K-means clustering (k=6) based on activation patterns across the audio corpus, grouping neurons with similar response profiles.

For each cluster, we computed the mean activation across all member neurons, then correlated this cluster-level trajectory with musical features. This differs fundamentally from per-layer analysis. Rather than asking "does layer X correlate with pitch?", we ask "does this group of functionally-similar neurons spanning multiple layers correlate with pitch?"

3.3 Results

3.3.1 Overall Encoding Strength

We found that for pitch and BPM (the two features measured), we found Pitch had a mean $|\rho|$=0.29, whereas BPM came in with with a higher mean $|\rho|$=0.32.

3.3.2 Synthetic vs Natural Audio

When directly comparing synthetic and natural audio for $|\rho|$ using layer-matched paired Wilcoxon signed rank tests, synthetic stimuli were encoded far more strongly than natural audio across all measures for pitch (100% of layer-matched pairs saw rank-biserial r ≥ 0.999; all $p < 0.001$)

3.3.3 Architecturally Consistent Encoding

Given our intention is to find localised intervention points based on encoding strength, we first examined whether the encoding strength of each layer was consistent across models and audio types. We computed Kendall's coefficient of concordance (W) over the rank ordering of layers by $|\rho|$, treating each model–dataset combination as an independent rater. Significance was assessed via a permutation test (5,000 permutations, shuffling each rater's ranking independently) and pooling all raters ($N = 9$ for Pitch Hz, $N = 12$ for BPM) across 28 layers, we found significant agreement for both properties. Pitch Hz encoding strength showed strong concordance ($W = 0.32$, permutation $p = 0.0002$), while BPM encoding strength showed weaker but still significant concordance ($W = 0.14$, $p = 0.014$). These results indicate that which layers encode pitch is considerably more consistent across models and datasets than it is for tempo.

We also investigated whether specific decoder layers consistently encode strongly features across models. We did this by finding the top-5 strongest encoding layers across them for each dataset. For temporal encoding, `net.7` showed consistency only for drum datasets and `net.19` for vocal datasets. For pitch, `net.19` also showed commonality amongst the string dataset. No single layer consistently encoded either feature across across all models and audio types.

3.3.4 Feature Encoding Enhancement

For comparing clusters, we pick the best cluster per section (after filtering out clusters that are less than 5 neurons) and compare this against the best layer in the corresponding section. Filtering prevents best selection picking a small cluster that may have high correlation by chance.

For each cell, feature, and measure, we identified the best-performing cross-layer cluster and the best-performing single layer within that section, and computed their difference. We tested across cells with a Wilcoxon signed-rank test. For natural audio, only tempo (BPM) showed a reliable cluster advantage for mean $|\rho|$ (N=27 $\Delta + 11\%$, $r = +0.67$, Fig 1).

3.3.5 Hierarchical Distribution of Top Clusters

We identified which sections contained the top-10 highest-correlating clusters for each feature. Both showed a balanced bimodal pattern. For Pitch Hz, the middle and late section dominated with 8/10 top clusters, and BPM demonstrated an early skew (8/10 in early and middle sections).

3.4 Discussion

3.4.1 Architectural Consistency and Content-Dependent Processing

Our layer-level analysis suggests that neither pitch nor tempo is consistently encoded in one particular site. Examining the top-5 strongest encoding layers across all three models, consistency emerged only within stimulus types rather than across them. That the same layer responds well to different features for different audio sets points to encoding that is content-dependent rather than feature-

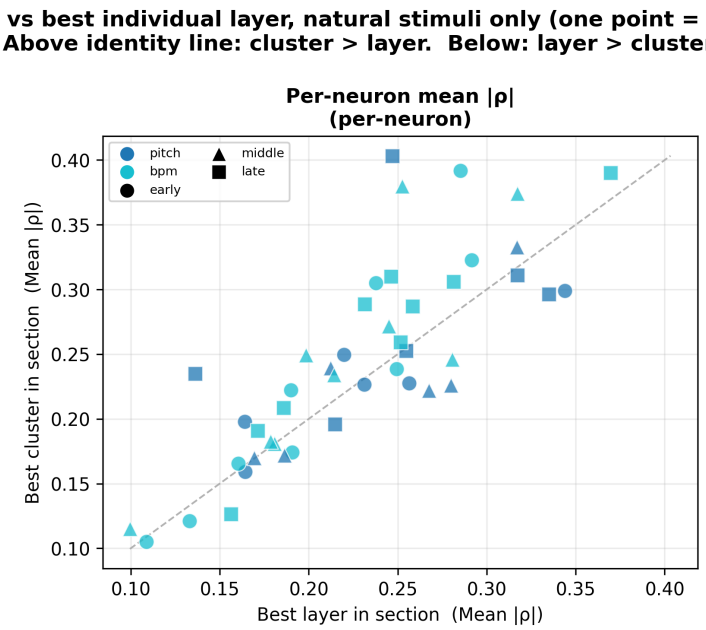

Figure 1: Delta of best layer vs best cluster per section

dedicated. Beyond this, the two features differ in how reproducible their layer-wise profiles are. Pitch encoding strength showed strong rank concordance across model–dataset pairs whereas tempo was markedly less concordant despite comparable overall encoding strength.

3.4.2 Robustness to Acoustic Complexity

Natural audio encoding despite acoustic complexity demonstrates robust feature representations under realistic conditions. While synthetic stimuli provided significant improvements, natural audio maintained meaningful correlations. The larger synthetic advantage for pitch suggests pitch-based manipulations may degrade more substantially with complex natural audio, while tempo-based network bending may generalise more reliably across diverse content.

4 STUDY 2: TARGETING NEURON SUBSETS FOR NETWORK BENDING

4.1 Overview

Study 1 identified that RAVE decoders encode pitch and tempo, as well as results that suggest cross-layer clusters can achieve stronger tempo encoding than individual layers. These findings suggest natural manipulation targets, however, correlation with a feature does not guarantee causal control over that feature.

To investigate this, Study 2 tests whether identified encoding locations translate into effective manipulation targets through large-scale network bending experiments.

This study differs from Study 1 in three ways. First, we expand from two features (pitch, tempo) to ten, adding spectral and timbral properties (brightness, noisiness, spectral width) that musicians typically wish to control. Pitch and tempo make good targets for encoding analysis because they can be systematically varied in synthetic stimuli with known ground-truth values. Spectral features lack equivalent natural scales but represent precisely the characteristics musicians wish to manipulate. For example, it is more challenging to construct a dataset that linearly increases in "noisiness" while holding other properties constant. Second, we use exclusively natural audio, as synthetic stimuli are useful for isolating encodings but represent unrealistic use cases for musical performance. Third, our success metrics shift from encoding strength (correlation between activations and feature values) to predictability, magnitude and efficiency of intervention.

4.2 Method

4.2.1 Dataset

We evaluated three neuron subset selection methods for their ability to provide reliable, bidirectional control over audio features through activation scaling. Cross-layer clustering groups neurons by activation pattern similarity cross-layers within each network section (early, middle, late), producing 6 clusters per section (18 total). The random matched baseline uses random neuron selection matching the exact sizes of cross-layer clusters, drawn from the same pool of available neurons. This controls the size of the cluster while testing whether activation-based grouping provides benefits beyond random selection. Layer-wise selection takes all neurons within individual layers, providing 28 distinct subsets corresponding to each decoder layer. When needed, they are also grouped into early, middle and late using the same mapping.

The comparison between cross-layer clustering and random matched is not a complete baseline, as we used information from the clustering process to determine random cluster sizes. However, given the potential effect of neuron count on outcomes, size-matching was necessary to isolate the contribution of informed clustering versus random chance.

We rendered 100 samples from each natural audio dataset (strings, drums, vocals) through each of the 3 models, applying activation scaling at levels of 0, 0.25, 0.5, 1 (original), 2, and 4. This produced 288,000 samples (100 x 3 x 3 x 5 = 4,500 per subset; (4,500 x 18) + (4,500 x 18) + (4,500 x 28) = 288,000).

4.2.2 Features

We evaluated across 10 audio features: spectral centroid (brightness), spectral rolloff (high-frequency content), spectral bandwidth (spectral width), RMS energy (loudness), RMS energy coefficient of variation (loudness stability), zero-crossing rate (noisiness), tempo, median pitch, and pitch stability. Pitch was extracted only from vocal and strings model outputs, as drum models do not reliably produce pitched audio. Features were computed over entire 4-second clips (rather than the 100ms windows used for encoding analysis) to capture larger-scale temporal effects of bending.

4.2.3 Metrics

For each neuron subset, we computed controllability as the Spearman correlation between log scale factor and the resulting feature change (delta from the unmodified decoded audio). Using log scale factor ensures symmetric treatment of amplification and attenuation (e.g., doubling and halving activations have equal weight). High $|\rho|$ indicates predictable bidirectional control: positive ρ means scaling up increases the feature, while negative ρ indicates an inverted but equally controllable relationship.

We also calculated the magnitude of effect to determine which selection methods produce the greatest change in audio features, and efficiency to determine how many neurons are required to achieve that change. Magnitude is computed as the mean absolute feature change per unit log scale factor. Efficiency is magnitude divided by the number of neurons modified. Both metrics are standardised by feature standard deviation to permit comparison across audio features.

For creative control, practitioners need only one effective subset per feature. We therefore report the highest metric achieved by any subset within each method/section combination. To avoid using the same data to both select the best-performing subset and evaluate its effectiveness, we first identify the best subset for each feature using 70% of available data, then validate with the remaining 30%.

In summary, controllability measures predictability of direction, magnitude measures size of effect regardless of direction, and efficiency measures effect per unit cost (neurons modified).

Standard inferential statistics are challenging when comparing audio features as they are derived from the same source (in this case an audio file). We provide ranking based statistics where the methods are compared across audio features and intervention points and the number of wins reported.

4.3 Results

4.3.1 Size of Subsets

The subsets selected as best-of-K (and the average across clusters in general) are higher for the cross-layer methods than whole layers (688 vs 418, 561 vs 361). Because cross-layer clusters pool neurons across all layers within a section, they are not bounded by any single layer's width and can therefore be larger than the widest individual layer. The random matched baseline is size-matched to the clusters, so cluster-versus-random comparisons remain controlled for neuron count even where cluster-versus-layer comparisons are not.

To test whether subset size explains method differences, we correlated the number of neurons in each subset with controllability for each feature. Low values suggest no relationship with mean $|\rho|$=0.137 across all selection methods (layer-wise $|\rho|$=0.11, cluster $|\rho|$=0.13, random matched $|\rho|$=0.18).

4.3.2 Controllability

Layer-wise selection achieved the highest controllability (as defined in Section 4.2.3) in 22 of 30 feature/section combinations on held-out test data, followed by cross-layer clustering (3 wins) and random matched (0 wins), with 5 ties. All three methods achieved strong controllability ($|\rho| > 0.8$) for loudness and loudness stability. Figure 2 presents the full comparison across features and sections with wins compared across selection methods in Table 1

Looking at mean controllability across all audio features, layer-wise selection outperforms both other methods across all sections (early: 0.45 vs 0.28/0.31; middle: 0.48 vs 0.36/0.33; late: 0.54 vs 0.37/0.37) and demonstrates best performance in late layers. Cross-layer clustering shows relatively consistent performance across the network, matching or outperforming random matched, particularly in late layers.

Loudness performs well for all methods in all sections ($|\rho| > 0.80$). Brightness and spectral width perform well in late sections for layer-wise selection against cross-layer methods. Tempo is poorly controlled in all sections for all methods ($|\rho| = 0.01$–0.08), remaining challenging even for layer-wise approaches. Pitch shows a different pattern where cross-layer clustering outperforms layer-wise selection in middle ($|\rho| = 0.30$ vs 0.17) and late layers ($|\rho| = 0.17$ vs 0.09), suggesting pitch control may benefit from targeted neuron selection.

4.3.3 Magnitude of Effect

Table 1 presents effect magnitude for subsets selected by highest magnitude on training data, then evaluated on held-out test data. Random matched produces the largest mean magnitude effects across all audio features (0.85), followed closely by cross-layer clustering (0.82), with layer-wise selection lower (0.66).

Table 1: Effect magnitude and controllability by method, for subsets selected by magnitude and controllability. Evaluated on held-out test data.

Metric	**Layer-wise**	**Random**	**Cross-Layer**
Selected by Best Controllability			
Control. ($\|\rho\|$)	0.49	0.30	0.31
Wins (Control.)	22	0	3
Selected by Best Magnitude			
Mag.	0.66	0.85	0.82
Wins (Mag.)	4	14	7

Indeed, when subsets are selected for magnitude, the two sparse methods head the win counts (random matched 14 wins, cross-layer clustering 7 wins, layer-wise 4). This is clearly demonstrated in the middle and right plots in Fig 3 where the majority of points land below the identity line.

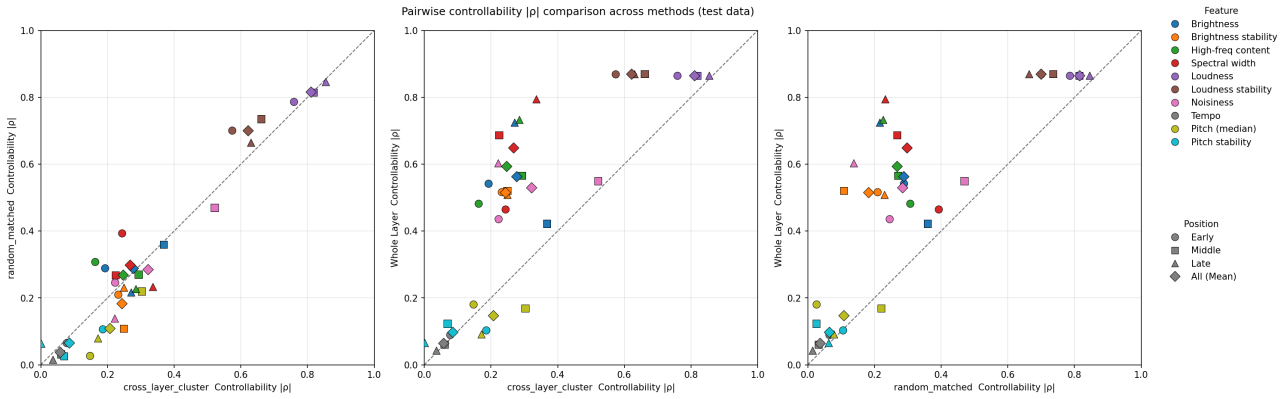

Figure 2: Best-of-K controllability ($|\rho|$) by feature and network section for each selection method, evaluated on held-out test data.

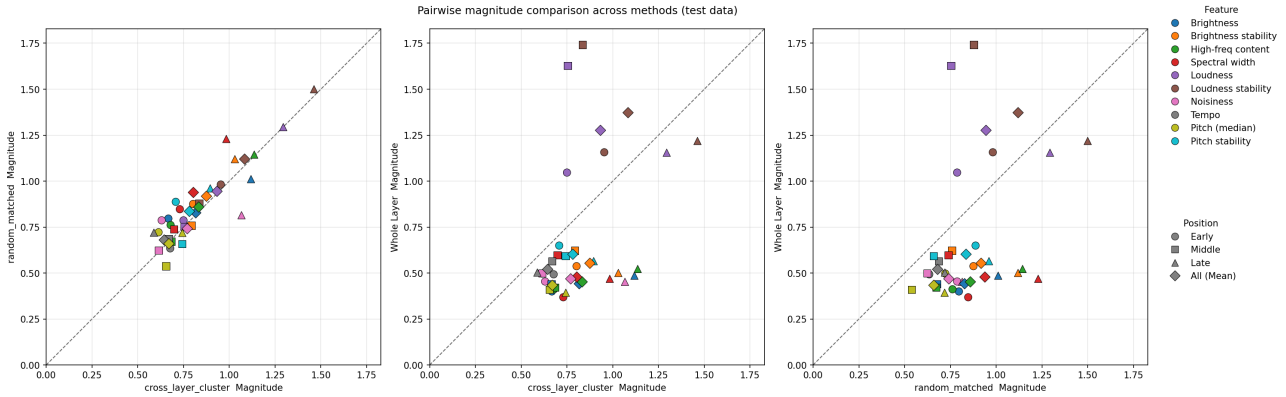

Figure 3: Best-of-K magnitude by feature and network section for each selection method, evaluated on held-out test data.

4.3.4 Encoding vs Controllability and Magnitude

Table 2 presents correlations between encoding strength (activation-feature correlation from Study 1) and manipulation effectiveness (controllability and magnitude from Study 2), computed separately for each selection method.

For cross-layer clustering, encoding strength shows no significant relationship with controllability for either feature. Encoding strength is positively correlated with magnitude for BPM (ρ=+0.23, p<0.01), while the pitch relationship is not significant.

For layer-wise selection, encoding strength again shows no significant relationship with controllability. The magnitude relationships are strong but feature-dependent and opposite in sign: BPM shows a strong positive relationship (ρ=+0.502, p<0.001), while pitch shows a strong negative relationship (ρ=−0.28, p<0.01), indicating that layers which more strongly encode pitch produce smaller effects when manipulated.

4.3.5 Efficiency

Table 3 presents efficiency results for the subsets selected by controllability. High efficiency numbers mean a big effect per neuron scaled. Late layers demonstrate substantially higher efficiency (0.011–0.047) compared to early (0.001) and middle layers (0.002–0.003). Cross-layer clustering achieves the highest mean efficiency (0.016) with the gap between methods is driven almost entirely by the late layers, where the cluster-based methods concentrate their effect in far fewer neurons than whole layers.

4.3.6 In-Distribution vs Out-of-Distribution

Show in Table 4, we examined controllability and magnitude for in-distribution audio (e.g., strings into strings model) versus out-of-distribution audio (e.g., strings into drums model). Controllability degrades modestly and by a similar amount for all three methods ($\Delta = -0.03$ to -0.04), with layer-wise retaining its advantage in both conditions (0.51 in-distribution, 0.48 out-of-distribution).

Magnitude behaves differently. Whilst both cross-layer methods have higher magnitude than layer-wise in both conditions, the gain achieved on in-distribution audio is considerable in comparison (-0.01 .vs. +0.15). Between the two methods, in-distribution, cross-layer clustering clearly dominates effect magnitude (15 wins vs 9 for random matched and 3 for layer-wise), but this advantage reverses under distribution shift, where random matched leads (16 wins) and clustering falls to 7. The domain-specific structure captured by clustering therefore appears to benefit effect size rather than predictability, and only for inputs resembling the training distribution.

4.4 Discussion

Several findings reveal unexpected patterns that inform both our understanding of RAVE's internal organisation and practical approaches to network bending.

4.4.1 Encoding Strength as a Predictor of Manipulation Success

A clear result to discuss is the dissociation between encoding strength and controllability. Study 1 identified

Table 2: Correlation between encoding strength (from Study 1) and manipulation effectiveness (Study 2), separated by method. Encoding→Control tests whether neurons that strongly encode a feature also provide controllable manipulation. Encoding→Magnitude tests whether they produce larger effects. *p<0.05, **p<0.01, ***p<0.001.

			Encoding→Control		Encoding→Magnitude	
Method	**Feature**	**n**	r	p	r	p
Cross-layer cluster	Pitch	72	−0.01	0.88	−0.09	0.44
	BPM	162	+0.04	0.44	+0.23	0.002**
Layer-wise	Pitch	112	−0.16	0.08	−0.28	0.003**
	BPM	252	−0.10	0.10	+0.502	<0.001***

Table 3: Efficiency ($|\text{effect}|/\sigma$ per neuron) by method and network position, for subsets selected by controllability on held-out test data.

Method	**Early**	**Middle**	**Late**	**Mean**
Cross-layer cluster	0.001	0.002	0.047	0.016
Random matched	0.001	0.002	0.012	0.005
Layer-wise	0.001	0.003	0.011	0.005

Table 4: Distribution shift effects

Method	**ID $\lvert\rho\rvert$**	**OOD $\lvert\rho\rvert$**	**Δ**
Layer-wise	0.51	0.48	−0.03
Random matched	0.35	0.31	−0.04
Cross-layer	0.42	0.38	−0.04
Method	**ID Mag.**	**OOD Mag.**	**Δ**
Layer-wise	0.66	0.67	+0.01
Random matched	0.97	0.88	−0.09
Cross-layer	0.99	0.84	−0.15

cross-layer clusters that encode tempo more strongly than individual layers, yet these clusters provide virtually no tempo control ($|\rho| < 0.1$).

For both selection methods, encoding strength fails to predict controllability in any strong or consistent way. Encoding strength does, however, predict effect magnitude, and here the two features diverge sharply. For tempo, stronger encoding predicts larger effects but for pitch under layer-wise selection, the relationship reverses: layers that more strongly encode pitch produce smaller effects. In neither case does stronger encoding translate into predictable directional control. This dissociation suggests that encoding strength identifies neurons that carry feature information without being reliable control points; strong encoders may sit in redundant pathways the network can compensate around (yielding smaller effects, as for pitch) or in positions that move a feature substantially but inconsistently (as for tempo).

Tempo, in particular, shows a robust encoding–magnitude relationship but weak controllability across all methods, reinforcing the view that tempo emerges from distributed computation that neither method can effectively isolate into predictable control.

4.4.2 Controllability and Magnitude

When subsets are selected for controllability, layer-wise selection dominates controllability (22 wins); when selected for magnitude, the sparse-and-random methods dominate magnitude (random matched 14 wins, cross-layer clustering 7 wins). This confirms that "how much can I change something" and "can I predictably control the direction of change" are largely independent properties. Practitioners must choose: reliable control (layer-wise) or maximal effect (clustering or random matched), but rarely both from the same subset.

4.4.3 Clustering

Cross-layer clustering provides a little advantage over size-matched random selection in controllability. However, when the objective is maximal magnitude, random matched matches or exceeds clustering, indicating that the effect of clustering is less important than selecting groups across layers.

The comparison against layer-wise selection shows less control but larger effects. This suggests the learned clusters successfully identify which neurons matter but fail to capture complete functional circuits. When scaling entire layers, the intervention includes not only feature-relevant neurons but also their operational context: co-activated neurons, lateral connections, and regulatory feedback mechanisms that together produce coherent, predictable feature modulation. Sparse, cross-layer interventions change the identified neurons while leaving their context unchanged, creating inconsistency between modified and unmodified neighbours that may explain reduced input-output predictability despite targeting the "correct" neurons.

4.4.4 Efficiency

The results reveal a fundamental tradeoff between controllability and efficiency. Layer-wise selection provides the strongest feature control across most combinations, potentially because entire layers represent coherent functional units in the RAVE architecture. However, this comes at substantial cost. In the late layers, where efficiency is highest, cross-layer clustering is roughly 4× more efficient than layer-wise selection (0.047 vs 0.011), indicating that manipulating all neurons in a layer produces only marginally more effect than targeting a small subset. Our results suggest that selecting specific neurons captures most of the available leverage, particularly those in later layers.

4.4.5 Variance in Audio Features

Different acoustic features show distinct controllability profiles that inform practical steering strategies. Loudness exhibits both low magnitude variation and high controllability across all methods and distribution conditions, suggesting relatively linear, accessible encoding where small interventions produce proportional, predictable effects regardless of selection strategy. Loudness is an easy, reliable target.

Noisiness and high-frequency content show the inverse pattern: large magnitude effects are achievable, but directional control is limited. These features may be encoded in more distributed, nonlinear representations where targeted perturbations produce dramatic but unpredictable cascading effects. Tempo presents a unique case with high magnitude potential but weak controllability across all methods, and remains challenging. This result potentially requires greater scrutiny as it may be that distortions caused by bending effect the reliability of downstream tempo extraction. Pitch shows a different pattern where cross-layer clustering outperforms layer-wise selection in middle and late layers, suggesting pitch may be encoded in more localisable neuron populations amenable to targeted intervention.

4.4.6 Bias-Variance Tradeoff in Effect Size

The in-distribution versus out-of-distribution analysis reveals a bias-variance tradeoff, but in effect size rather than in control. Controllability degrades by a similar, modest amount for all three methods ($\Delta = -0.03$ to -0.04), and the ordering is unchanged: layer-wise leads both in-distribution (0.51) and out-of-distribution (0.48). Whatever makes layer-wise more controllable is therefore not a robustness property; it holds equally in both conditions.

Magnitude tells the opposite story. The two sparse methods achieve far larger effects than layer-wise in-distribution (0.99 and 0.97 vs 0.66), but this advantage is domain-dependent. Cross-layer clustering loses 15% of its effect size under distribution shift ($\Delta = -0.15$), random matched loses 9%, and layer-wise loses nothing ($\Delta = +0.01$).

This gradient is informative. Clustering learns structured groupings that capture domain-specific organisation, such as how brightness is encoded in string timbres, and those groupings deliver the largest effects precisely when the input matches the domain they were learned from. When it does not, the learned structure no longer corresponds to how features are organised for the new timbre, and the advantage is less. Random matched, which shares clustering's size distribution but none of its structure, degrades roughly half as much and this is consistent with the interpretation that the extra degradation is attributable to the learned structure itself rather than to sparsity or subset size.

Layer-wise intervention makes no assumptions about feature organisation. By scaling all neurons in a dense location uniformly, it neither exploits nor depends upon domain-specific structure. This is a higher-bias approach that trades effect size for stability, and it is the only method whose magnitude is unaffected by distribution shift.

For practitioners, the implication is that the choice of selection method depends on both objective and expected inputs. Layer-wise intervention provides the most reliable control in either condition. Cross-layer clustering offers the largest effects, but only for inputs resembling the training domain.

4.4.7 Network Bending Applications

For practical network bending applications, these findings suggest method selection should consider both the objective and expected inputs. Applications requiring precise, graded control should favour layer-wise intervention despite lower efficiency. Applications prioritising dramatic transformation with less predictability concern may benefit from cross-layer clustering or random matched selection. When steering will be applied to diverse or unpredictable inputs, layer-wise intervention provides more reliable control through its robustness to distribution shift.

Finally, cluster-based manipulation in decoder activation space offers complementary advantages to traditional latent space control. Latent manipulation produces global, holistic changes affecting the entire output through a high-dimensional control surface. Activation-space manipulation enables targeted, feature-specific changes that can potentially address one aspect (pitch, brightness) without globally disrupting generation, through a lower-dimensional control surface (18 clusters versus potentially hundreds of latent dimensions). For interface designers, the 18-dimensional cluster-based approach (6 clusters $\times$ 3 sections) offers a tractable alternative to high-dimensional latent manipulation, though with the tradeoffs in controllability documented above.

5 LIMITATIONS

Our findings may not generalise beyond RAVE's VAE architecture. Diffusion models and transformer architectures may exhibit fundamentally different encoding patterns.

That cross-layer clustering identifies tempo encoding neurons, but no subsets show any meaningful tempo control requires further investigation. This divergence suggests tempo may be an emergent property of layer-level computation rather than a manipulable localised representation, but alternative explanations (e.g., clustering artifacts, scale factor dynamics) cannot be ruled out without further investigation.

We evaluated controllability using extracted audio features rather than human perception. This gives an advantage of scale in that we were able to analyse a far greater number of combinations of sounds, models and manipulations than human participants would be able to. However, a listening study would verify whether statistically significant feature changes correspond to perceptually meaningful differences.

We only investigated one approach (scaling) amongst many possibles for network bending. Its naturally linear parameterisation allowed for a good evaluation of controllability analogous to a traditional user interface, but the activations of the groups of neurons identified could be *bent* in many more ways.

6 CONCLUSION

We presented the first comprehensive layer-wise and cross-layer cluster analysis of spectro-temporal feature encoding in RAVE audio synthesis models, examining both where features are represented (interpretability) and how effectively they can be manipulated (controllability).

Cross-layer clustering revealed functionally coherent neuron groups achieving stronger feature correlation than individual layers for tempo.

When considering **RQ1**, network bending experiments with 288,000 rendered samples revealed that encoding strength does not straightforwardly predict manipulation effectiveness. These experiments also provided insight for **RQ2** as we show layer-wise selection achieved highest controllability for most features, while cross-layer clustering provided superior efficiency and magnitude.

As response to **RQ3**, we highlight this divergence between encoding and controllability as it implies that interpretability analysis and controllability testing are complementary but non-redundant evaluation stages for practitioners. When computational cost is acceptable and maximal control is needed, layer-wise manipulation provides the most reliable results. When efficiency matters, for example in real-time applications or when minimising artifacts, cross-layer approaches identify targeted neuron subsets.

Future work should explore: (1) causal validation through ablation of specific identified clusters, (2) extension to timbre and dynamics features, (3) cross-architecture comparisons with diffusion and GAN-based models, (4) perceptual validation of manipulation effects through listening studies, and (5) real-time neural synthesiser interfaces that leverage both encoding insights and controllability findings. The distinction between understanding feature encoding and achieving feature control will likely remain relevant as neural audio synthesis architectures continue to evolve.

7 ETHICAL STATEMENT

This research was conducted within the scope of the authors' academic responsibilities and received no external funding. The authors declare no conflicts of interest.

No human participants were involved in this study, and no ethical approval was required.

All music used for training and evaluation was sourced in accordance with applicable licensing terms. Specifically, string and drum datasets were either synthesised or obtained from licensable music libraries under their respective terms of use. Vocal datasets comprise commercially available recordings; however, the trained models are used exclusively for internal analysis of activations and outputs rather than for the generation of novel musical works. Neither the model weights nor any generated outputs are made publicly available, mitigating concerns around unauthorised reproduction or derivative use of copyrighted material.

All open-source software libraries used in this work are employed in accordance with their respective licences and are attributed accordingly. The authors acknowledge the broader ethical considerations surrounding generative audio models, including their potential for misuse in voice cloning or the unauthorised replication of artistic style. While this work is analytical rather than generative in its aims, we recognise the importance of responsible development practices in this space.

REFERENCES

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, pages 6541–6549.

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. (2019). Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*.

Broad, T., Leymarie, F. F., and Grierson, M. (2021). Network bending: Expressive manipulation of deep generative models. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, pages 20–36. Springer.

Bäuerle, A., Jönsson, D., and Ropinski, T. (2022). Neural activation patterns (naps): Visual explainability of learned concepts.

Caillon, A. and Esling, P. (2021). Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. (2022). High fidelity neural audio compression. *Transactions on Machine Learning Research*.

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

Engel, J., Hantrakul, L., Gu, C., and Roberts, A. (2020). Ddsp: Differentiable digital signal processing. In *ICLR*.

Geiger, A., Lu, H., Icard, T., and Potts, C. (2021). Causal abstractions of neural networks. volume 34, pages 9574–9586.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. (2020). Ganspace: Discovering interpretable gan controls. volume 33, pages 9841–9850.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., and Eck, D. (2018). Music transformer.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

McCallum, L. and Yee-King, M. (2020). Network bending neural vocoders. In *Proceedings NeurIPS 2020 Machine Learning for Creativity and Design Workshop*.

Mor, N., Wolf, L., Polyak, A., and Taigman, Y. (2018). A universal music translation network.

Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

Shen, Y., Gu, J., Tang, X., and Zhou, B. (2020). Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. volume 33, pages 12388–12401.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. In *ICML Deep Learning Workshop*.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. (2021). Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.