ESTIMATING MUSICAL SURPRISAL FROM AUDIO IN AUTOREGRESSIVE DIFFUSION MODEL NOISE SPACES

Mathias Rose Bjare[1] **Stefan Lattner**[2] **Gerhard Widmer**[1,3]
[1] Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2] Sony Computer Science Laboratories (CSL), Paris, France
[3] LIT AI Lab, Linz Institute of Technology, Austria
mathias.bjare@jku.at

arXiv:2508.05306v1 [cs.SD] 7 Aug 2025

ABSTRACT

Recently, the information content (IC) of predictions from a Generative Infinite-Vocabulary Transformer (GIVT) has been used to model musical expectancy and surprisal in audio. We investigate the effectiveness of such modelling using IC calculated with autoregressive diffusion models (ADMs). We empirically show that IC estimates of models based on two different diffusion ordinary differential equations (ODEs) describe diverse data better, in terms of negative log-likelihood, than a GIVT. We evaluate diffusion model IC's effectiveness in capturing surprisal aspects by examining two tasks: (1) capturing monophonic pitch surprisal, and (2) detecting segment boundaries in multi-track audio. In both tasks, the diffusion models match or exceed the performance of a GIVT. We hypothesize that the surprisal estimated at different diffusion process noise levels corresponds to the surprisal of music and audio features present at different audio granularities. Testing our hypothesis, we find that, for appropriate noise levels, the studied musical surprisal tasks' results improve. Code is provided on github.com/SonyCSLParis/audioic.

1. INTRODUCTION

Surprisal, estimated via *information content* (IC) or negative log-likelihood (NLL) of an autoregressive model, has been proposed as a proxy estimator for perceived musical surprise as experienced by human listeners [1–4]. With suitable models, the IC of musical events correlates with human surprise perception and complexity, including tonal and rhythmic aspects [5, 6]. Music analysis with IC enables quantitative information-theoretic hypotheses about music and music perception [7]. Furthermore, IC can serve as a conditioning signal for generative models [8–10]. Recently, [11] showed that surprisal modeling using IC calculated in the continuous audio latent space of *Music2Latent* [12] effectively models musical complexity, repetition reduction, and prediction of electroencephalogram (EEG) brain responses in human listeners. In [11], a GIVT model [13] is used that calculates IC using the likelihood of one-step predictions that are assumed to follow Gaussian mixture model (GMM) distributions with uncorrelated dimensions. However, this assumption may limit such models' effectiveness, given the nature of highly compressed continuous autoencoder representations, like Music2Latent.

Diffusion models have become powerful tools in generative AI, achieving state-of-the-art results in multiple domains, including music. A key advantage is that they do not rely on strong assumptions about how the data is distributed. Recent work [14] shows that by formulating diffusion processes as ODEs, a diffusion model can not only generate new samples but also estimate how likely (or "probable") any given data point is under the model. Remarkably, this likelihood estimate can be computed at different stages of the diffusion process, which correspond to varying levels of "noise" or abstraction in the data.

In this paper, we study the ability of diffusion models to estimate musical surprisal. We restrict our investigations to ADMs [15, 16], since musical surprisal is causal in time. We experiment with diffusion models trained on multi-track audio following the popular EDM [17] and the Rectified Flow [18] processes, which differ in how they noise details. We show that these models describe diverse music data better in terms of NLL than GIVT, consistent with audio fidelity results in generation tasks [16]. We evaluate the estimated IC's effectiveness in modelling aspects of musical surprisal in audio on two tasks that have been studied in the monophonic symbolic domain: capturing monophonic pitch surprisal, related to tonality understanding, and segment-boundary detection on multi-track audio, related to the information changes in music. We show that surprisal estimated using diffusion models captures pitch surprisal better than the GIVT model. Furthermore, we demonstrate that peaks in the surprisal function align with segment boundaries; however, additional peaks are found. Finally, we hypothesize that IC estimated at certain diffusion process noise levels can preserve the surprisal of higher-level audio features like pitch, while filtering out contributions to the IC of low-level features like timbre nuances. We support our hypothesis by showing that, for appropriate noise levels, the results of the studied musical surprisal tasks improve.

© Mathias Rose Bjare, Stefan Lattner, and Gerhard Widmer. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Mathias Rose Bjare, Stefan Lattner, and Gerhard Widmer, "Estimating Musical Surprisal from Audio in Autoregressive Diffusion Model Noise Spaces", in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

2. RELATED WORK

In the symbolic music domain, musical surprisal proxied by IC has most notably been studied with the variable order Markov-model IDyOM [2]. IDyOM modeling of human melodic expectation has been validated by numerous behavioral and neural studies [3, 19–23]. The model is, however, limited to monophonic symbolic music stimuli. In [10], the authors propose an IC-based technique for estimating surprisal in polyphonic symbolic music and show the IC to correlate with tonal and rhythmic complexity using solo piano performances.

In the audio domain, surprisal estimation typically relies on human-selected audio features. In [24], the IC of a D-REX [25, 26] model, calculated using Bayesian inference, is related to the magnetoencephalography (MEG) brain response of human participants. The Audio Oracle [27] analyzes surprisal using *information rate* calculated from self-similarities of audio features and identifies high surprisal at segment boundaries.

Surprisal estimation using symbolic music or audio features faces two issues: the investigation is based on limited human-selected attributes, and a mismatch between what the computational model sees and what a human listener hears. Both cases potentially bias the investigation.

Therefore, most similar to our approach, [11] estimates surprisal in an audio representation that preserves all features of the original audio. The authors estimate IC using the likelihood of a GIVT model [13] and show that it can predict EEG responses to sung music. However, in contrast to ADMs, the GIVT model assumes that the next-step predictions follow a particular distribution, which may limit its predictive effectiveness.

Although unrelated to temporal surprisal, [28] uses the KL-divergence of a diffusion model to approximate the likelihood of 5-second monophonic music clips. This is used to reproduce the inverted U-shape relation between the total IC of music clips and listener preference presented in [7]. The model does not rely on audio features; however, it ignores causality and memory aspects of surprisal. In addition to IC, other measures of information have been proposed for the computational study of musical surprisal and expectation [29, 30]. These, however, are impractical to calculate for continuous autoregressive models and have been limitedly validated perceptually in the literature.

3. METHOD

3.1 Information Content Modelling

Estimation of causal IC in a discrete (symbolic music) domain can be achieved effectively with GPT-style one-step prediction modelling. In this case, IC is calculated from the prediction target's log-likelihood according to an explicit (softmax) probability mass function with logits from a multi-layer perceptron (MLP) that takes as inputs a context state summarized by a causal Transformer model [31]. As a result, the IC measures the likelihood of specific (musical) events. In contrast, we aim to estimate the IC of *continuous audio embeddings*, using the compressed representations of the *Music2Latent* autoencoder [12]. In this continuous case, the probability mass cannot be modelled. Consequently, in [11], a GIVT model [13] is used, which models the probability density of next-step predictions explicitly using a GMM with parameters from an MLP that takes as input a context state, summarized by a causal Transformer. In this work, we do not require explicit density modeling. Instead, it suffices to obtain IC by *log-likelihood point estimates* of the observations. To that end, we calculate such point estimates using *Autoregressive Diffusion Models (ADMs)* [15, 16]. Similar to GPT-style causal transformers, ADMs summarize the context of past observations into a context state. However, instead of using an MLP to transform the context states into softmax logits or GMM parameters, ADMs use the context states to condition small diffusion model MLPs to *generate* the next continuous state.

Estimating IC using a diffusion model requires the use of the Instantaneous Change of Variables formula [32]. This formula, as detailed below, computes the log-likelihood of data points $\mathbf{z}_0 \sim \pi_0$ (in our case, Music2Latent audio representations) that can be flown to a known analytic distribution π_1 according to an ODE $\frac{d}{dt}\mathbf{z}(t) = f(\mathbf{z}(t), t)$. Finding such ODEs is non-trivial, but it turns out that neural ODEs [32] derived from diffusion processes do exactly that: flow data samples to noise samples of known isotropic Gaussian distributions.

3.2 Instantaneous Change of Variables

For data points $z_0 \sim \pi_0$ flowing in time t according to $\frac{d}{dt}\mathbf{z}(t) = f(\mathbf{z}(t), t)$, [32] shows that the log-likelihood of points change according to another ODE: $\frac{d}{dt}\log p\left(\mathbf{z}\left(t\right)\right) = -\text{tr}\left(\frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}(t), t)\right)$, given some regularity conditions. Therefore, if $\mathbf{z}_0 \sim \pi_0$ is flown to $\mathbf{z}_1 \sim \pi_1$ and π_1 is known, we can evaluate $\log \pi_0(\mathbf{z}_0)$ by the sum of $\log \pi_1(\mathbf{z}_1)$ and the log-likelihood flow change from π_0 to π_1. Pratically, the two ODEs are combined into a system of equations and solved numerical from t_0 to t_1 given the initial conditions $\mathbf{z}(t_0) = \mathbf{z}_0, \log \pi_0(\mathbf{z}_0) - \log \pi_0(\mathbf{z}(t_0)) = 0$:

$$\int_{t_0}^{t_1} \underbrace{\begin{bmatrix} f(\mathbf{z}(t), t) \\ -\text{tr}\left(\frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}(t), t)\right) \end{bmatrix}}_{\text{dynamics}} dt = \underbrace{\begin{bmatrix} \mathbf{z}_1 \\ \log \pi_0(\mathbf{z}_0) - \log \pi_1(\mathbf{z}_1) \end{bmatrix}}_{\text{solutions}}. \tag{1}$$

We can now obtain $\log \pi_0(\mathbf{z}_0)$ by adding $\log \pi_1(\mathbf{z}_1)$ to the solution of the 2nd ODE. The former can be easily evaluated using the solution of the 1st ODE and the known π_1. Furthermore, [33] shows that Eq. 1 can be calculated efficiently with reverse-mode automatic differentiation using the Skilling-Hutchinson trace estimator [34, 35], which involves using n_r Monte Carlo runs with noise samples from a Rademacher distribution [35] to obtain an unbiased estimate of an expectation. For the approach to work, we therefore require finding ODEs that flow the data distribution π_0 to an analytic distribution π_1. In the following, we consider two diffusion model-based neural ODEs [32] learning such flows.

3.3 Probability Flow ODEs

In [17, 36], the authors define a diffusion noise (forward) process by a stochastic differential equation (SDE) that flow the data distribution π_0 to a (known) Gaussian distribution π_1 in time $t_0 \rightarrow t_1$. The dynamics of the SDE on data points $\mathbf{z}(t_0)$ flowing to $\mathbf{z}(t)$ can effectively be described by:

$$p_{t_0,t}\left(\mathbf{z}(t)|\mathbf{z}(t_0)\right) = \mathcal{N}\left(\mathbf{z}(t); \mathbf{z}(t_0)s(t), s(t)^2\sigma(t)^2\mathbf{I}\right), \tag{2}$$

where σ is a noise scale and s a contraction chosen such that $p_{t_0,t_0}\left(\mathbf{z}(t_0)|\mathbf{z}(t_0)\right) = \pi_0$ and $p_{t_0,t_1}\left(\mathbf{z}(t_1)|\mathbf{z}(t_0)\right) = \pi_1 \approx \mathcal{N}\left(\mathbf{z}(t_1); 0, \sigma_{\max}^2\mathbf{I}\right)$. Remarkably, the SDEs can be translated to deterministic processes (probability flow ODEs) that equivalently flow π_0 to π_1 given by:

$$\frac{d}{dt}\mathbf{z}(t) = s(t)^2\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{z}}\log p\left(\frac{\mathbf{z}(t)}{s(t)}; \sigma(t)\right) - \frac{\dot{s}(t)}{s(t)}\mathbf{z}(t). \tag{3}$$

These ODEs, thus, fulfil the requirements of Section 3.2. Eq. 3 can be turned into a neural ODE by learning $\nabla_{\mathbf{z}}\log p$ with a neural network (see Section 4.2) using score matching. In Section 4, we use the EDM initialization of Eq. 3 from [17], where $s(t) = 1$, $\sigma(t) = t$, and the process flows in time $t_0 = 0.002$ to $t_1 = 80$. This model will be referred to as EDM in our experiments.

3.4 Rectified Flow

Rectified Flow (RFF) [18] defines a process that flow the data distribution to a standard Gaussian distribution ($\pi_1 = \mathcal{N}\left(\mathbf{z}(t_1); 0, \mathbf{I}\right)$) by following straight lines as much as possible. Formally, given the ODE: $\frac{d}{dt}\mathbf{z}(t) = v(\mathbf{z}(t), t)dt$ a, RFF between π_0 and π_1 is learned by the minimization:

$$\min_v \int_0^1 \mathbb{E}\left[\left\|(\mathbf{z}_1 - \mathbf{z}_0) - v(\mathbf{z}_t, t)\right\|^2\right] dt, \tag{4}$$

where $\mathbf{z}_0 \sim \pi_0, \mathbf{z}_1 \sim \pi_1$ and $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$ for $t \in [0, 1]$. $\mathbf{z}_t$ is therefore a point on the straight line connecting $\mathbf{z}_0, \mathbf{z}_1$. Substituting v with a neural network (see Section 4.2) in the ODE, we get a neural ODE. The weights are learned using sample estimates of Eq. 4 as a loss.

3.5 Likelihood Estimations in Noise Space Continua

In addition to estimating likelihoods of the data distribution π_0 using the framework described above, we can also compute likelihoods at various noise levels along the noise/data continuum (that are traversed by varying t—either in the perturbation kernel of Eq. 2 for probability flow ODEs, or in $\mathbf{z}_t$ for RFF). In both cases, rather than solving the ODE from t_0 to t_1, we solve it from the noise level t down to t_1.

It is noted that both processes introduce noise gradually into the data. As a result, high-detail information is removed first, while lower-detail information is retained at lower noise levels, and then also lost as the noise increases. Therefore, in Section 4.5 we hypothesize that the IC extracted at moderate noise levels captures the surprisal of certain lower-detail musical features—such as pitch—while filtering out the contributions of higher-detail features, like subtle timbral nuances.

Given a test example existing in π_0, there are three natural ways to obtain a "noised" version of the data at a given noise level t: (1) sampling from the noise process, (2) using the expected value of the noise process, or (3) solving the ODE from t_0 to t. We discard option (1) because it yields a stochastic estimate, and option (3) because we found that option (2)—using the expected value—produces better results in practice.

4. EXPERIMENTS AND RESULTS

4.1 Data

For training and evaluating our models, we use the following audio datasets and encode them into Music2Latent representations, using the public checkpoint of [12]. For model training, we use a dataset consisting of 150,000 CC licenced full-length mixed-source MP3 files from Jamendo (JAM) [37], which we split into 125k, 12.5k, and 12.5k examples for training, validating, and testing purposes, respectively. For experiments involving monophonic singing voices, we use a private dataset for fine-tuning our models, comprising vocal stems from 20k songs. For our experiments with monophonic pitch, we use a synthetic dataset (SYN) of 49 Irish folk tunes from *The Session* dataset [38], synthesized at constant velocity with diverse SoundFont-based instruments according to the midi-programs: *"Pad 1 (new age)"*, *"Synth Voice"*, *"Acoustic Guitar (nylon)"*, *"Acoustic Grand Piano"* and *"Trumpet"* for a total of 245 examples. Additionally, for each melody, the IC of IDyOM is computed (see [39] for details). Furthermore, we use the dataset of [40] (VOC), consisting of 18 recorded vocal melodies paired with IDyOM IC estimates of the transcribed melodies. For our experiment on segment boundary detection, we use the Salami dataset (SAL) [41], containing 1310 audio files, having segment annotations from one or two human annotators. The annotations are hierarchical and include the following levels: functional, uppercase, and lowercase, describing global structure with semantic segment labels, global structure, and local structures, respectively. We use all datasets to evaluate model effectiveness using NLL.

4.2 Architecture and Training Details

For the diffusion models, we use a standard 12-layer causal Transformer backbone with Pre-Layer Normalization similar to [42], rotary positional embeddings [43], and FlashAttention [44], and for the diffusion MLP, we follow the architecture of [16]. For the GIVT model, we follow the architecture presented in [11]. All models are trained with a maximum sequence length of 4600, corresponding to approximately 7 minutes of audio and a batch size of 8 sequences, resulting in an effective batch size of up to $\sim$1 hour of music. We use Adam optimization for 270k steps with a learning rate of $3 \cdot 10^{-4}$ for the diffusion models and 10^{-4} for GIVT, using a cosine schedule with a warmup of 1800 steps. For our experiments involving monophonic

	n_r	1	2	4	8	16
S-MAE	EDM	.109	.078	.057	.043	.033
	RFF	.109	.079	.057	.043	.033
	tol	1	.1	.01	.001	1e-4
Q-MAE	EDM	.085	.078	.076	.076	.076
	RFF	.076	.076	.076	.076	.076
	tol	1	.1	.01	.001	1e-4
Q-ME	EDM	-.044	-.018	-.004	.000	.000
	RFF	.000	-.001	.000	.000	.000

Table 1. Approximation errors of the likelihood estimation, indicated with the Skilling-Hutchinson (S) and quantization (Q) mean average error (MAE) and the quantization mean error (ME). The error is reported with respect to references of $n_r = 32$ runs and a tolerance of $tol = 10^{-4}$ for the two error types, respectively. The results are normalized to the mean absolute NLL of the references.

singing voices, we additionally fine-tune each model on the dataset mentioned in Section 4.1 until convergence.

4.3 ODE-based Likelihood Approximation Errors

The likelihood estimation from ODE-based diffusion models is affected by two types of approximation errors: the discretization error of the ODEs and the Skilling-Hutchinson trace estimator (see Section 3.2). In both cases, the approximation error can be controlled by trading off speed. We therefore perform initial experiments on 500 examples from our validation dataset to determine a suitable trade-off. For the former, similarly to [33, 36], the error is controlled using the Runge-Kutta 5(4) [45] method. For the latter, the unbiased approximation can be made arbitrarily small using enough Monte Carlo runs n_r. We use the scipy Runge-Kutta implementation [46] with standard parameters except for setting the tolerances to $atol = rtol = tol = 10^{-3}$ and compute the mean absolute error (MAE) of the difference between NLL calculated with different n_r and NLL of a reference calculated with a very large number of runs ($n_r = 32$).

To relate the MAE to the scale of the NLL, we divide it by the reference's average absolute NLL and report the resulting measure as S-MAE in Table 1. For all n_r, we found that the average error is small for both models. Even when $n_r = 1$, the error is 0.109 of the average NLL. This demonstrates that it is possible to obtain a coarse estimate of a sample's NLL with minimal computational overhead compared to traditional diffusion model generation. We identify $n_r = 4$ as a good balance and fix it for further experiments.

For determining tolerance parameters tol, we similarly compare NLL calculated with different values of tol to a reference of $tol = 10^{-5}$ and report this as Q-MAE in Table 1. Furthermore, we investigate the bias by plotting the mean error (ME), normalized to the average NLL, and report it as Q-ME in Table 1. We find that for $tol \leq 0.1$ and 0.01 for EDM and RFF, the absolute error does not improve compared to the reference. Comparing the bias of the error, we find that while RFF is unbiased, EDM has a negative bias for $tol > 0.001$. The better performance of RFF is likely due to the straight flows imposed by the method, which allow the solver to take larger steps. Thus, we take $tol = 0.001$, such that the relative MAE is 0.057.

	JAM	SAL	VOC	SYN
GIVT	0.925	1.053	1.182	0.981
EDM	0.707	**0.829**	**0.823**	**0.642**
RFF	**0.699**	**0.829**	0.831	0.656

Table 2. Comparison of model NLLs in the Music2Latent on different datasets reported in bits/dimension.

4.4 Predictive Efficiency Comparision

Similar to previous work in density estimation models [33, 36, 47], we compare the model's predictive effectiveness (how well the models predict diverse audio data) using the average NLL reported in bits/dimension (mean negative $\log_2$-likelihood/dimension). Since all compared models estimate likelihoods in the fixed coordinate system of the Music2Latent codec, we can compare the NLL in that space. We emphasize that our reported results are, therefore, not directly comparable with those described in [11] as it uses a different version of Music2Latent.

We find that *Music2Latent* encodes silence into a small region of its latent space, causing the model to assign extremely low IC values to silent frames (since IC is unbounded from below for densities). Consequently, these low values downweight the average NLL calculations without improving the models' predictive capabilities. To address this, we remove leading and trailing silence from the audio before computing NLL. Similarly, for each dataset, we discard the IC values at time steps that fall within the 1% most extreme IC values (across any model). We present the models' average NLL in Table 2.

We find that the diffusion models have much lower NLL than the GIVT model, and as such, model the one-step prediction densities more accurately. This is consistent with the findings of [16] for audio fidelity in a generative task. Comparing the NLL values of the EDM and RFF models reveals no clear winner. Interestingly, we find that the NLL of diffusion models on the SYN monophonic dataset, which is dissimilar to the training distribution JAM, is lowest. Using an information-theoretic interpretation, the low NLL indicates that the SYN dataset is less surprising regarding timbre and melody, and therefore has lower IC.

4.5 Pitch Surprisal in Noise Space Continua

In [12], it is noted that a small MLP can predict pitches from the Music2Latent representations with high accuracy. Thus, it is reasonable to hypothesize that pitch is embedded in coarse structures in the Music2Latent representations. We, therefore, investigate to what extent our IC can explain pitch surprisal and whether IC estimates at different noise levels describe pitch surprisal to a greater extent.

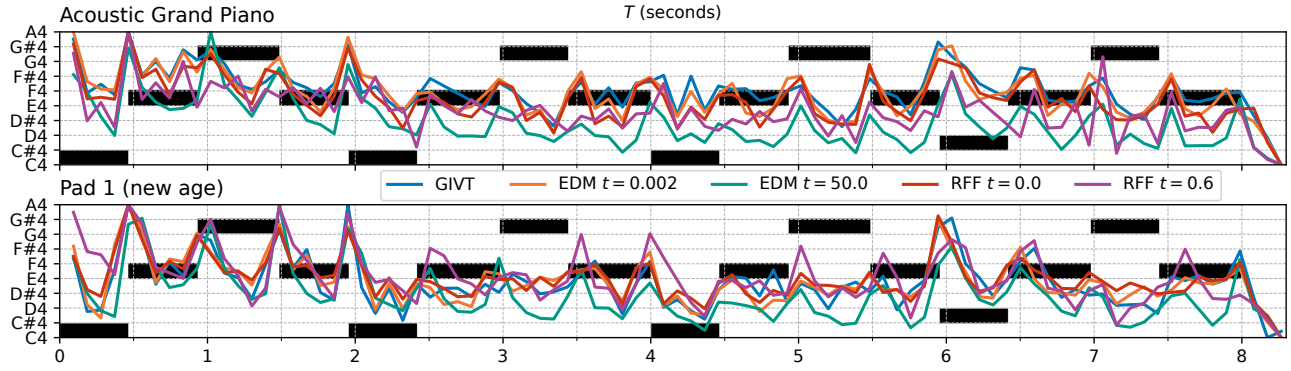

Figure 1. Piano roll of simple melody and IC of EDM and RFF models calculated from SoundFont audio synthesized at constant note-velocity with "Acoustic Grand Piano" (top) and "Pad 1 (new age)" (bottom) instruments. The IC is shown at noise levels $t = 0.002, 0.0$ (corresponding to fully unnoised data), and $t = 50, 0.6$ (corresponding to mid-process values) for EDM, RFF, respectively. The IC is affinely transformed to a min/max-value of $0/1$ for visualization.

4.5.1 IC Surprisal Qualitative

In Fig. 1, we plot a simple melody along with IC estimates on SoundFont audio synthesized with two different instruments. The melody is composed of a C major arpeggio repeated four times. On the fourth occurrence at $T = 6$, the pattern is modified with the out-of-tonality pitch $C^{\#}$. When predicting E at $T = 2.5s$ and forward, there is a partial match between the current context and the context seen $2s$ earlier. That is, E and the following notes are expected at time $T = 2.5s$ and forward, which is reflected in the lower IC across all models in the investigated noise levels, compared to the ICs at times $T - 2s$. At $T = 6s$, the surprising out-of-tonality note prompts a big peak in IC estimates across all models and noise levels. Finally, at time $T = 8s$, the melody is surprisingly abruptly terminated, reflected in a smaller final peak in IC estimates across all models. Comparing the IC estimates for the different timbre, we find that the mid-process IC estimates at $t = 50.0, 0.6$, are more similar across timbre than the estimates of fully-denoised data at $t = 0.002, 0.0$ for EDM and RFF, respectively. This is evident, for example, by the lesser variation in IC between note onsets within $T = [0.5, 1.0]$, which is especially visible for the EDM model. This suggests that the IC estimates at moderate noise levels capture pitch surprisal rather than timbre variations more effectively than estimates from fully unnoised data, which we investigate in the following.

4.5.2 IC Surprisal Quantitative

To quantitatively validate whether the IC computed in audio can explain pitch surprisal, we would require a ground truth, which does not exist. As a proxy, we use the surprisal (IC) values predicted by the perceptually validated pitch expectancy model IDyOM (see Section 2), which operates in the symbolic domain. We conduct our experiment with the SYN and VOC datasets. We extract the IDyOM IC of each note pitch in the symbolic datasets, and pair these with IC values calculated from our models on the audio datasets. We align the latter to the notes by identifying the two Music2Latent frames that contain the note onset and calculating their average IC. Since we expect monotonic, rather than linear, relations in the IC pairs, we compare the paired estimates using Spearman's rank correlation and report the results in Table 3. We find all correlations to be significant at a 5%-significance level and, except for GIVT, positive. Comparing the GIVT to the diffusion models estimating IC of unnoised data ($t = 0.002, 0.0$ for EDM and RFF, respectively), we find that the correlations are higher for SYN and similar for the VOC dataset. In all cases, except for the RFF-VOC, we find the highest correlations using estimations with $t \neq t_0$, i.e., when IC is estimated using the noised data. For SYN, we see the highest correlations for high noise scale values $t = 50.0, 0.6$ (compared to the fully noised noise values $t = 80.0, 1.0$) for EDM and RFF, respectively. In particular, we find that EDM at noise level $t = 50$ is overall mostly correlated with the IC of IDyOM, which supports the findings in Section 4.5.1 for similar data, where this value shows the smoothest surprisal curves, with clearly defined peaks around the note onsets. For the VOC dataset, the highest correlations occur at lower noise levels, and overall, the correlations are less pronounced than in the SYN dataset. This may be because singers are less precise when changing pitch, often using portamento to glide into a note. As a result, peaks in IC during note changes may not be driven solely by pitch

GIVT	SYN	-.031				
	VOC	.147				
	t	2e-3	10.0	20.0	50.0	60.0
EDM	SYN	.137	.048	.190	**.264**	.255
	VOC	.135	**.213**	.206	.138	.106
	t	0.0	0.1	0.5	0.6	0.7
RFF	SYN	.134	.189	.216	**.218**	.189
	VOC	**.137**	.133	.069	.048	.030

Table 3. Spearman's correlation between IC calculated on (symbolic) melodies using IDyOM and IC calculated at different noise levels with EDM and RFF.

GIVT			.380		
t	2e-3	10.0	20.0	50.0	60.0
EDM	0.385	0.517	**.525**	0.522	.518
t	0.0	0.1	0.5	0.6	0.7
RFF	0.391	0.307	.385	**.429**	.402

Table 4. Spearman's correlation between IC estimations sharing the same note material, but with different timbre.

shifts, but also by other cues, such as emphasis at the onset (e.g., volume changes), vowel transitions, or the presence of plosives. Masking such potentially subtle characteristics with noise may explain the observed correlation reduction.

4.5.3 Noise Space Continua Timbre Invariance

As shown above for SYN, the IC is more correlated with pitch surprisal at intermediate noise levels, where the fine details in the audio embeddings have been removed. We investigate to what extent this can be explained by a higher invariance to timbre irrelevant to pitch surprisal. We, therefore, investigate if IC estimated on music that contains the same note content but with different timbre is more similar at the noise levels studied above. Specifically, we use SYN and investigate Spearman's correlation between IC of all combinations of pairs of synthetics sharing the same note material (but having different timbre), and report the results in Table 4. The results are all significant at a 5% significance level. For unnoised data, the correlations are similar for the diffusion models and GIVT. However, for EDM especially, the correlations increase for noised data. We find high correlations for noise level values $t = 50, 0.6$ for EDM and RFF, respectively, which have the highest correlation with IDyOM (See table 3), supporting the notion that these noise levels are more invariant to timbre.

4.6 IC for Unsupervised Segment Boundary Detection

In the symbolic domain, IC has been used as a novelty function for segment boundary detection [48–50]. Therefore, we investigate whether big changes in IC extracted from audio also coincide with segment boundaries. We conduct an experiment where we predict Salami lowercase segment boundaries using the most significant peaks extracted from an IC novelty function. The novelty function is constructed by smoothing our IC curves with a Gaussian kernel with standard deviation $\sigma = 5$, and differencing the smoothed series. Using the off-the-shelf Röder peak picking algorithm [51] with standard parameters, we report, in Table 5, precision, recall, and F_1-score on predictions that are accurate within ± 0.5 seconds of the annotations. Generally, precision values are substantially lower than recall, implying that the IC novelty curves tend to have extra peaks not attributed to segment boundaries. For the GIVT, and the IC estimated with diffusion models on unnoised data, we find the F_1 scores to be similar. For RFF, and to a lesser extent EDM, precision and recall increase with the noise level. This shows that the IC estimated at a coarser level aligns better with the segment boundaries.

GIVT	prec		.158		
	rec		.309		
	F_1		.209		
EDM	t	2e-3	17.6	40.0	60.0
	prec	.159	.162	.169	**.178**
	rec	.286	.311	.324	**.345**
	F_1	.204	.213	.222	**.235**
RFF	t	0.0	0.25	0.50	0.70
	prec	.159	.163	.179	**.198**
	rec	.287	.342	.380	**.416**
	F_1	.205	.221	.243	**.268**

Table 5. Precision, recall and F_1 scores of Salami lowercase ± 0.5 seconds boundary detection.

5. CONCLUSION AND DISCUSSION

We investigated ADMs' ability to estimate musical surprisal and found that EDM and RFF diffusion models more effectively describe music data than a GITV in terms of NLL. We evaluated the diffusion models IC's effectiveness in capturing monophonic pitch surprisal and found that these capture pitch surprisal better than a GIVT. Furthermore, we found that IC estimates of noised data increase correlation with pitch surprisal, and showed that this coincides with these estimates being more invariant to timbre. Furthermore, we showed that peaks in a novelty function derived from IC coincide with Salami lowercase segment boundaries; however, the function has additional peaks. Finally, using the IC estimated in noise space improves the segment boundary predictions regarding precision and recall. As such, diffusion models surpass GIVT models in surprisal estimation and offer additional estimates that can capture aspects important to musical surprisal.

Similarly to [11], we estimate surprisal with IC in Music2Latent representations, so their findings on musical complexity, repetition reduction, and EEG prediction are likely to extend to diffusion-based IC. This should be validated in future work and extended with other perceptual validating experiments on diverse music. Furthermore, our investigation of noise levels relevant to pitch surprisal could be extended to consider other perceptual features and their entanglement in different data representations. For instance, the IC calculated at suitable (high) noise levels in mel-spectrograms or constant-Q transformed representations may give estimations of surprisal that correlate more with pitch surprisal. Also, the exploratory investigation of optimal noise levels could be automated using a methodology similar to [52], by monitoring performance degradations of a classifier/regressor model trained to predict the feature using variably noised inputs. Finally, our pitch surprisal analysis measured IC of coarse-grained structures, but our framework also allows studying surprising changes in fine-grained structures. This might, for instance, be relevant for analyzing timbre changes or singing techniques.

6. ACKNOWLEDGMENTS

The work leading to these results was conducted in a collaboration between JKU and Sony Computer Science Laboratories Paris under a research agreement. The first and third author also acknowledge support by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement 101019375 (*"Whither Music?"*).

7. REFERENCES

[1] L. B. Meyer, "Meaning in music and information theory," *The Journal of Aesthetics and Art Criticism*, vol. 15, no. 4, pp. 412–424, 1957.

[2] D. Conklin and I. H. Witten, "Multiple viewpoint systems for music prediction," *Journal of New Music Research*, vol. 24, no. 1, pp. 51–73, 1995.

[3] M. Pearce, "The construction and evaluation of statistical models of melodic structure in music perception and composition," Ph.D. dissertation, Department of Computing, City University, London, UK, 2005.

[4] M. T. Pearce and G. A. Wiggins, "Auditory expectation: The information dynamics of music perception and cognition," *Top. Cogn. Sci.*, vol. 4, no. 4, pp. 625–652, 2012.

[5] S. A. Sauvé and M. T. Pearce, "Information-theoretic modeling of perceived musical complexity," *Music Perception: An Interdisciplinary Journal*, vol. 37, no. 2, pp. 165–178, 2019.

[6] M. R. Bjare, S. Lattner, and G. Widmer, "Exploring sampling techniques for generating melodies with a transformer language model," in *ISMIR*, 2023, pp. 810–816.

[7] B. P. Gold, M. T. Pearce, E. Mas-Herrero, A. Dagher, and R. J. Zatorre, "Predictability and uncertainty in the pleasure of music: a reward for learning?" *Journal of Neuroscience*, vol. 39, no. 47, pp. 9397–9409, 2019.

[8] C.-i. Wang and S. Dubnov, "Guided music synthesis with variable markov oracle," in *AAAI*, vol. 10, no. 5, 2014, pp. 55–62.

[9] T. Collins, R. Laney, A. Willis, and P. H. Garthwaite, "Developing and evaluating computational models of musical style," *AI EDAM*, vol. 30, no. 1, pp. 16–43, 2016.

[10] M. R. Bjare, S. Lattner, and G. Widmer, "Controlling surprisal in music generation via information content curve matching," in *ISMIR*, 2024.

[11] M. R. Bjare, G. Cantisani, S. Lattner, and G. Widmer, "Estimating musical surprisal in audio," in *ICASSP*, 2025.

[12] M. Pasini, S. Lattner, and G. Fazekas, "Music2latent: Consistency autoencoders for latent audio compression," in *ISMIR*, 2024.

[13] M. Tschannen, C. Eastwood, and F. Mentzer, "GIVT: generative infinite-vocabulary transformers," *CoRR*, vol. abs/2312.02116, 2023.

[14] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *ICML*, vol. 202, 2023, pp. 32 211–32 252.

[15] T. Li, Y. Tian, H. Li, M. Deng, and K. He, "Autoregressive image generation without vector quantization," in *NeurIPS*, 2024.

[16] M. Pasini, J. Nistal, S. Lattner, and G. Fazekas, "Continuous autoregressive models with noise augmentation avoid error accumulation," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[17] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *NeurIPS*, 2022.

[18] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *ICLR*, 2023.

[19] M. T. Pearce, M. H. Ruiz, S. Kapasi, G. A. Wiggins, and J. Bhattacharya, "Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation," *NeuroImage*, vol. 50, no. 1, pp. 302–313, 2010.

[20] G. M. Di Liberto, C. Pelofi, R. Bianco, P. Patel, A. D. Mehta, J. L. Herrero, A. De Cheveigné, S. Shamma, and N. Mesgarani, "Cortical encoding of melodic expectations in human temporal cortex," *Elife*, vol. 9, p. e51784, 2020.

[21] N. C. Hansen and M. T. Pearce, "Predictive uncertainty in auditory sequence processing," *Frontiers in psychology*, vol. 5, p. 1052, 2014.

[22] R. Bianco, L. E. Ptasczynski, and D. Omigie, "Pupil responses to pitch deviants reflect predictability of melodic sequences," *Brain and Cognition*, vol. 138, p. 103621, 2020.

[23] T. Moldwin, O. Schwartz, and E. S. Sussman, "Statistical learning of melodic patterns influences the brain's response to wrong notes," *Journal of cognitive neuroscience*, vol. 29, no. 12, pp. 2114–2122, 2017.

[24] E. Abrams, E. M. Vidal, C. Pelofi, and P. Ripollés, "Retrieving musical information from neural data: how cognitive features enrich acoustic ones." in *ISMIR*, 2022, pp. 160–168.

[25] B. Skerritt-Davis and M. Elhilali, "Detecting change in stochastic sound sequences," *PLoS Comput. Biol.*, vol. 14, no. 5, 2018.

[26] B. Skerritt-Davis and M. Elhilali, "A model for statistical regularity extraction from dynamic sounds," *Acta Acustica united with Acustica*, vol. 105, no. 1, pp. 1–4, 2019.

[27] S. Dubnov, G. Assayag, and A. Cont, "Audio oracle: A new algorithm for fast learning of audio structures," in *ICMC*, 2007, pp. 224–227.

[28] N. L. Masclef and T. A. Keller, "Deep generative models of music expectation," *NeurIPS ML for Audio Workshop 2023*, 2023.

[29] S. Abdallah and M. Plumbley, "Information dynamics: patterns of expectation and surprise in the perception of music," *Connection Science*, vol. 21, no. 2-3, pp. 89–117, 2009.

[30] S. Dubnov, "Deep music information dynamics," *arXiv preprint arXiv:2102.01133*, 2021.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[32] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," in *NeurIPS*, 2018, pp. 6572–6583.

[33] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, "FFJORD: free-form continuous dynamics for scalable reversible generative models," in *ICLR*, 2019.

[34] J. Skilling, "The eigenvalues of mega-dimensional matrices," *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*, pp. 455–466, 1989.

[35] M. F. Hutchinson, "A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines," *Communications in Statistics-Simulation and Computation*, vol. 18, no. 3, pp. 1059–1076, 1989.

[36] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*, 2021.

[37] Jamendo, "Jamendo Music," https://www.jamendo.com.

[38] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, "Music transcription modelling and composition using deep learning," in *Proceedings of the Conference on Computer Simulation of Musical Creativity*, Huddersfield,UK, 2016.

[39] M. R. Bjare, S. Lattner, and G. Widmer, "Differentiable short-term models for efficient online learning and prediction in monophonic music," *Trans. Int. Soc. Music. Inf. Retr.*, vol. 5, no. 1, p. 190, 2022.

[40] G. Cantisani, A. Chalehchaleh, G. Di Liberto, and S. Shamma, "Investigating the cortical tracking of speech and music with sung speech," in *INTERSPEECH*. ISCA, 2023, pp. 5157–5161.

[41] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. D. Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations," in *ISMIR*, 2011, pp. 555–560.

[42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[43] J. Su, M. H. M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[44] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," in *NeurIPS*, 2022.

[45] J. R. Dormand and P. J. Prince, "A family of embedded runge-kutta formulae," *Journal of computational and applied mathematics*, vol. 6, no. 1, pp. 19–26, 1980.

[46] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[47] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, "Flow++: Improving flow-based generative models with variational dequantization and architecture design," in *ICML*, vol. 97, 2019, pp. 2722–2730.

[48] M. T. Pearce, D. Müllensiefen, and G. A. Wiggins, "Melodic grouping in music information retrieval: New methods and applications," in *Advances in Music Information Retrieval*, 2010, vol. 274, pp. 364–388.

[49] S. Lattner, M. Grachten, K. Agres, and C. E. C. Chacón, "Probabilistic segmentation of musical sequences using restricted boltzmann machines," in *MCM*, vol. 9110, 2015, pp. 323–334.

[50] S. Lattner, C. E. C. Chacón, and M. Grachten, "Pseudo-supervised training improves unsupervised melody segmentation," in *IJCAI*. AAAI Press, 2015, pp. 2459–2465.

[51] M. Müller and F. Zalkow, "libfmp: A python package for fundamentals of music processing," *J. Open Source Softw.*, vol. 6, no. 63, p. 3326, 2021. [Online]. Available: https://doi.org/10.21105/joss.03326

[52] G. Daras, A. Rodriguez-Munoz, A. Klivans, A. Torralba, and C. Daskalakis, "Ambient diffusion omni: Training good models with bad data," *arXiv preprint arXiv:2506.10038*, 2025.