Modelling Pitch Expectation in Melody: A Comparison of Variable-Order Markov and Transformer-Based Approaches

Ioannis Emmanouilidis
Student ID: 240955485
Supervisor: Dr Marcus Pearce
MSc. Sound and Music Computing
Queen Mary University of London

***Abstract*—Musical expectation is widely regarded as a key mechanism underlying the emotional impact of music. Past work has modelled expectation with rules-based approaches, such as Narmour's Implication–Realisation model, or statistical learning methods like the Information Dynamics of Music (IDyOM) which have shown stronger explanatory power. More recently, deep neural models such as the Music Transformer have demonstrated success in symbolic music generation, raising questions about their potential for modelling expectation. This study directly compared IDyOM and the Music Transformer in predicting probe-tone ratings from experiments by Cuddy and Lunney (1995) and Schellenberg (1996). IDyOM was tested with basic and derived viewpoint configurations, while Transformer probabilities were obtained from a model trained on MAESTRO and fine-tuned on the Monophonic Corpus of Complete Compositions. Regression analyses showed that the Transformer significantly outperformed IDyOM for short two-note contexts, whereas both models performed comparably for longer melodic contexts. These findings suggest that while the Music Transformer offers an advantage when higher-level structural information is absent, IDyOM remains a competitive model for longer melodic sequences, and their comparison helps render the Transformer's predictions more interpretable.**

***Index Terms*—musical expectation, statistical learning, deep learning, IDyOM, Music Transformer, pitch prediction**

I. Introduction and Related Work

Musical expectation is widely regarded as central to suggested explanations of the different affective states music has the capacity to elicit in listeners. A seminal—though by no means the first—account of the relationship between expectation and emotion was given by Leonard Meyer in his 1956 book *Emotion and Meaning in Music*, whose main thesis is that the confirmation, violation and delay of listeners' expectations constitute the primary source of music's emotive power (Meyer, 1956). In the decades since, composers, musicologists, psychologists, and cognitive scientists have speculated on, studied, and modeled musical expectation in order to uncover the mechanisms underlying its formation, understand the impact of its confirmation or violation on the listener, and ultimately detail its role in emotion induction. This is not to say that expectation is the only salient factor in generating different affective states during listening, with other contributors having been proposed and empirically supported, including rhythmic entrainment (the adaptation of one's internal body rhythms to music), aesthetic judgement, evaluative conditioning (the frequent pairing of a piece of music with other positive or negative stimuli), the evocation of visual imagery, and personal associations through episodic memory (Juslin et al., 2022). Nonetheless, with mounting experimental evidence, expectation endures as a key framework for explaining music's capacity to produce pleasure.

A. Sensory and Cognitive Processes in Musical Expectation

The pursuit of identifying valid theories of musical expectation has led to the creation and experimental testing of various rules-based, statistical and computational models. These exist on a sensory-cognitive continuum depending on the different representations they offer, with sensory models capturing continuity and variation in low-level acoustic properties on one end, and cognitive schemata that reflect higher-level learned musical knowledge—the kind that traditional music theory typically deals with—on the other (Collins et al., 2014). The proposed integration of sensory and cognitive processes in predictive coding for music is reflected in various theories of musical expectation, such as the five psychological response systems that make up David Huron's ITPRA theory (Huron, 2008). In this framework, the *reaction response* following an event is almost instantaneous and precedes thought and reflection, which in music relates to low-level acoustic information, such as the shock of a sudden, loud chord. On the other hand, the *prediction response*, which is positively valenced for an expected stimulus and negatively valenced for an unexpected one, might also encompass cognitive priming, such as in a V-I cadential resolution, while the *appraisal response* is even further along the spectrum, constituting an aesthetic judgement which calls on previously learned musical schemas, cultural knowledge, and personal taste.

Cheung et al. (2024) investigated the contrasting processes of sensory and cognitive priming and found that sensory and cognitive models provided independent, non-overlapping contributions to predicting listeners' subjective expectancy and pleasantness ratings for chords sampled from US Billboard pop songs, with cognitive expectations explaining over twice the variance in perceived surprise compared to sensory expectations, providing evidence for the larger relative importance of long-term representations of music structure over short-term sensory–acoustic information in musical expectancy. In light of this, an important clarification should be made: the description of predictive processes as lying on a sensory-cognitive continuum should not be interpreted as implying a common underlying mechanism underpinning all such processes; it is the existence of distinct mechanisms which is in fact supported by the independent and non-overlapping nature of the sensory and cognitive contributions.

B. Melodic expectation

One of the most studied aspects of expectation is melodic expectation, particularly in relation to pitch-based prediction and surprisal. Proposed factors shaping melodic expectation occupy different points along the sensory-cognitive spectrum. At the more sensory end are theories emphasising intervallic proximity and direction, while models which are based on cognitive processing emphasise abstract structural features, such as tonal-scale degree. Theories about the relative significance of such musical characteristics in shaping melodic expectation are usually grounded in two broader explanatory mechanisms, often presented as competing accounts: Gestalt-like rules and statistical learning.

C. Rules-based approaches to modelling pitch expectation

In books published in 1990 and 1992, Eugene Narmour, influenced by Meyer's application of Gestalt psychology principles in expectation, put forth a model of melodic expectation as an alternative to Schenkerian analysis. The Implication-Realisation (IR) Theory argued that intervals in a melody create implications for how the melody will continue, which are then either realised or violated by subsequent notes, and these implications are governed by two systems working simultaneously. The first is a top-down system involving musical schemas that have been internalised from prior exposure to music, such as tonality. The second consists of a collection of Gestalt-inspired rules: registral direction, intervallic difference, registral return, proximity, and closure. The system they make up is seen as being innate, reflecting universal properties of our perceptual systems, as opposed to being learned from experience (Narmour, 1990, 1992).

Cuddy and Lunney (1995) tested the salience of these rules in predicting human rating data using a probe-tone experimental paradigm. This provided experimental support for three of Narmour's five principles—intervallic difference, registral return, and proximity—as well as a modified version of registral direction.

Building on Narmour's work, Schellenberg (1997) developed a model that sought to simplify the IR model, while retaining predictive power. His version of the IR model distilled the framework down to just two rules—proximity and pitch reversal—while also incorporating key-profiles as a top-down tonal predictor. Schellenberg found that many predictors of the original IR model were strongly correlated with each other when predicting human probe-tone ratings, while the principles of pitch proximity and reversal which carried most of the explanatory weight in his revised model were considerably more orthogonal.

D. Statistical learning models of pitch expectation

Since the proposal of the IR model, its claim that melodic expectations arise from innate and universal bottom-up rules has been challenged by approaches grounded in statistical learning. In this alternative conception, musical expectation does not arise from hardwired rules, but rather patterns acquired through innate general-purpose cognitive learning mechanisms, which extract domain-specific representations (e.g., tonal hierarchies, interval probabilities) from exposure to a musical corpus (Pearce and Wiggins, 2006).

Subsequently, what appear to be universal constants in melodic structure are often more aptly explained as consequences of performance constraints rather than innate features of the auditory system. For instance, when a singer makes a large leap toward the edge of their vocal range, they are often compelled to return toward the middle of their range, where pitch production is more comfortable, and this constraint can give rise to the statistical predominance of pitch reversals without invoking hardwired, domain-specific rules.

A prevalent statistical-learning framework of this kind is the Information Dynamics of Music (IDyOM), originally formulated in Pearce (2005) as a computational framework for modelling melodic expectation using variable-order Markov models. IDyOM estimates the conditional probability of each possible next note given the preceding musical context based on the observed frequency of matching n-gram sequences encountered in the musical corpus the model has been trained on (a more detailed account of IDyOM's mechanics is provided in the Materials and Methods section to follow).

Pearce and Wiggins (2006) compared IDyOM with Schellenberg's two-factor IR model in predicting listeners' ratings of melodic continuations from three probe-tone experiments, including Schellenberg (1996) and Cuddy and Lunney (1995) (to be explored further in the current study). IDyOM was found to explain significantly more variance in the human ratings, especially when confronted with longer, more complex melodic stimuli. IDyOM was also shown to subsume the effects of proximity and reversal in its predictions, supporting the hypothesis that the inclusion of innate bottom-up rules is unnecessarily complex. This conclusion is reinforced by cross-cultural evidence showing that IDyOM has the capacity to capture Gestalt-like principles across diverse musical repertoires (Savage et al., 2015). In addition, IDyOM (in its ltm configuration) also significantly outperformed a Gestalt-style model based on central pitch tendency, pitch proximity, and key-profiles from Temperley (2008) at predicting MEG responses to each note-onset of musical stimuli (Kern et al., 2022).

Some contrasting results come from Morgan et al. (2019), which finds evidence for independent contributions by both Gestalt principles and IDyOM. However, the variance in the human data unexplained by IDyOM could simply reflect statistical regularities which were not picked up by IDyOM's chosen configuration, or could be a consequence of the use of a singing experimental paradigm, which might introduce biases and regularities stemming from singing ability and preferences.

E. Deep Neural Models of Expectation

Having cited evidence for the superior explanatory power of statistical learning methods for capturing melodic expectation compared to rule-based approaches, we can now turn to the question of how such models compare with more recent models rooted in deep neural networks.

One example of this is the PolyRNN model presented in (Robert et al., 2024). This is a recurrent neural network designed to model predictions across multiple, simultaneous information streams. Unlike many earlier models which focus on a single melodic line, PolyRNN can learn dependencies within and between multiple streams, thus enabling it to generate probabilistic predictions for polyphonic stimuli. Experiments showed that PolyRNN successfully predicted neural activity in primary and associative auditory regions around 200 ms after note onset, as measured with MEG during polyphonic music listening. Furthermore, a comparison between PolyRNN and IDyOM's capacity to predict neural responses in monophonic music found that PolyRNN was able to account for variance beyond that which was explained by IDyOM. The authors speculated that PolyRNN's advantage may stem from its training on polyphonic music, allowing it to better capture the expectations of listeners who, having primarily been enculturated in polyphonic music, meet certain intervallic events in monophonic stimuli with surprise. Beyond modelling brain responses, further work remains to be done to investigate whether PolyRNN also outperforms IDyOM in predicting results from behavioural experiments.

The striking success of recent Transformer architectures in speech recognition and prediction, together with the structural parallels between speech and music, raises questions about their potential for modelling musical expectation. One such model is the Music Transformer, a state-of-the-art Transformer which, in its original presentation (Huang et al., 2018), was shown to generate coherent polyphonic piano music by leveraging relative self-attention mechanisms that enable the model to take long-range and variable context into account. The Music Transformer generates a probability distribution for the musical characteristics (pitch, onset timing, duration, and dynamics) of the next musical event given a preceding sequence, and can thus be exploited for the purposes of expectation modelling. This approach was taken by Kern et al. (2022), where IDyOM ltm and Music Transformer's predictions of mEEG responses recorded while participants listened to monophonic classical compositions demonstrated a similarly strong ability of the two models to capture melodic expectancy, with no statistically significant difference between the two.

F. The Current Study

The current study aims to further investigate whether the Music Transformer's greater architectural sophistication might translate into a superior capacity to model melodic expectation compared with IDyOM, or if the similar performance reported in Kern et al. (2022) will be replicated. In the current approach, a few notable deviations from the aforementioned model comparison have been made. Firstly, the IDyOM predictions are no longer limited by only applying the `cpitch` viewpoint (which tracks MIDI pitch number), but are assisted by incorporating regularities of higher-level cognitive features, such as pitch-class, scale-degree, and information about the first note of the piece and of each bar. Secondly, due to the absence of direct comparisons between IDyOM and neural models on modelling behavioural data, the present work will examine the ability of IDyOM and the Music Transformer to accurately predict ratings from the probe-tone experiments of Cuddy and Lunney (1995) and Schellenberg (1996) (it is acknowledged that previous work exists on modelling ratings from these experiments with IDyOM in Pearce and Wiggins (2006), but their use here facilitates comparison with the Music Transformer).

The following section details the materials and procedures used in these two experiments, and a detailed explanation of the mechanics of IDyOM and the Music Transformer, as well as the statistical analysis which was carried out to evaluate their performance, before we turn to the results and subsequent conclusions of this study.

II. Materials and Methods

A. Cuddy and Lunney (1995)

Participants: Twenty-four undergraduate students from Queen's University, participated. Half of the sample (nine females, three males) were musically trained, defined as holding at least Grade IX certification from the Royal Conservatory of Music or being enrolled in the second year or higher of the Bachelor of Music programme. Their mean age was 20.8 years (range: 19–22), with an average of 13.2 years of formal training. The remaining twelve participants (ten females, two males) were untrained listeners, with a mean age of 19.4 years and an average of 1.9 years of musical training.

Stimuli: The stimuli consisted of eight implicative intervals derived from Narmour's (1990) implication–realisation model: two small intervals (major second, minor third) and two large intervals (major sixth, minor seventh), each presented in ascending and descending form. To discourage the formation of a top-down sense of tonality across trials, two versions of the stimuli were created: one for which the second tone of the interval was C4, and one for which the second tone was the tonally distant F#4, the high unrelatedness of which is known from experimentally-derived key profiles (Krumhansl, 1990). Each interval was presented with a long–short rhythmic pattern (1.2 s followed by 0.4 s) and a strong–weak accentuation, intended to establish a 4/4 meter. For each implicative interval, all twenty-five chromatic continuations within the range of one octave above to one octave below the second tone (C3–C5 or F#3–F#5) were tested. Each continuation tone lasted 0.8 s and was equal in amplitude to the second tone of the interval. Stimuli were synthesized with Cakewalk Professional for Windows and realised on a Yamaha TX-802 synthesizer with a wood-piano timbre.

Procedure: On each trial, participants heard a two-tone implicative interval followed by a continuation tone, and rated how well the continuation completed the melodic beginning on a seven-point Likert scale (1= extremely poor continuation; 7= extremely good continuation). A short practice phase with one interval and five continuations familiarised participants with the task. The main experiment comprised eight blocks of 27 trials each. Each block presented one implicative interval with all 25 possible continuations in randomised order, preceded by two practice trials that were excluded from analysis. Block order was randomised per participant, with the restriction that no two consecutive blocks began or ended on the same pitch. Each participant completed 200 experimental trials in total, half of which ended on C4, with the other half ending on F#4.

B. Schellenberg (1996), Experiment 1

Participants: Twenty members of the Cornell University community took part. Listeners were classified by musical background: limited training ($n = 10$; < 6 years of training and no musical activity in the past 2 years) and moderate training ($n = 10$; ≥ 6 years of training and regular musical activity in the past 2 years).

Stimuli: Eight notated fragments from British folk songs (four major, four minor) served as contexts (see Fig. 1). Each fragment ended with an implicative (unclosed) interval meeting five criteria: (1) the second tone had equal or shorter duration than the first; (2) it was less tonally stable in the established key; (3) it fell on a metrically weaker beat; (4) it did not occur at the last or penultimate position of the phrase; and (5) it occurred 16–21 tones from phrase onset (to equate overall duration). Small implicative intervals were 2 or 3 semitones; large intervals were 9 or 10 semitones, each in ascending and descending versions. All fragments were rendered with a piano timbre on a Yamaha TX816 FM tone generator, at a natural tempo and with subtle metrical accents to clarify meter.

Procedure: Listeners rated how well a single continuation tone followed the fragment on each trial, using a 7-point scale (1= extremely bad continuation; 7= extremely good). Instructions emphasised continuation rather than completion and encouraged use of the full scale. A practice phase used one non-test fragment with eight practice trials in total. The test session comprised eight blocks (one per fragment). In each block, listeners rated 15 diatonic test tones (all scale members within ± 1 octave of the fragment's final tone), presented once each in a new random order per listener; block order was separately randomised. The total workload was 120 ratings (15×8 fragments).

TABLE I: IDyOM configurations (from Pearce (2025, Ch. 3)).

Configuration	Description
STM	The short-term model: learns incrementally each musical piece from an initially empty state
LTM	The long-term model: learns from a corpus of music and, after this initial training, generates predictions for subsequently presented music without further learning
LTM+	The long-term model configured to continue learning after its initial training
BOTH	The STM combined with the LTM
BOTH+	The STM combined with the LTM+

C. IDyOM

The Information Dynamics of Music (IDyOM) model has undergone several revisions since its introduction. The account here is of the most recent version and follows its description in Pearce (2025).

General model description: IDyOM is a probabilistic model of musical expectation which uses statistical learning to acquire and process internal representations of the syntactic structure of a musical corpus. IDyOM uses the learned statistical regularities in this training corpus in combination with similar short-term knowledge learned from the piece at hand to create a prediction of the musical characteristics of the next event.

Event representation and viewpoints: IDyOM takes as input discrete, symbolic representations of music, with each event (note or chord) having a set of basic attributes, or *basic viewpoints*, (e.g., `pitch`, `onset`, `duration`). From these, IDyOM generates and tracks additional *derived viewpoints*, such as `interval` (semitone distance to the previous pitch), `contour` (up/same/down), `scale-degree` (relative to key signature), and temporal viewpoints (`ioi`, `ioi-ratio`, `metrical-level`). Viewpoints tracked may also be linked as Cartesian products; for example, `interval`$\otimes$`scale-degree` captures joint melodic–tonal structure. During prediction, one or more source viewpoints are used to predict one or more basic target viewpoints (which will be pitch in the present work, but can extend to also include onset). IDyOM's flexibility in being able to track two musical attributes in separate models whose probabilistic output is subsequently combined in the final prediction, or as a linked viewpoint which tracks the two features concurrently, enables the testing of theories which propose that the cognitive representation attributes in question rely on separate or joint mechanisms. The regularities present among different viewpoints are tracked through the building of variable-order n-gram models (using a compressed suffix tree).

Model configurations: IDyOM combines information stored across two distinct components to make probabilistic predictions for a given stimulus: the *long-term model* (LTM) which reflects the stylistic traits of the training corpus, and the *short-term model* (STM) which is initially empty and incrementally picks up regularities of the piece being processed. The existence of these two models is rooted in psychology, as the long-term model is intended to capture schematic stylistic knowledge about music acquired through a lifetime of listening, while the short-term model reflects dynamic learning of the patterns within an individual piece of music. The contributions of each model are weighed and combined into a single prediction, making up a BOTH model; however the IDyOM framework also offers the option to do prediction based on only the LTM or STM, or the LTM+ and BOTH+ variants (see Table I). When using BOTH, the maximum context length taken into account for each prediction corresponds to the longest suffix of the past that has actually been seen in training (in LTM and in STM, separately). If the full recent context hasn't occurred before, the model backs off by dropping the earliest symbol until it finds a seen context, with a smoothing process shaping how the probability distributions for orders up to that point are defined and blended (see Pearce (2025, chap. 2) for further details).

Implementation in the present work: I used IDyOM to predict `cpitch` at the probe position for the two sets of stimuli already described. The BOTH configuration was used for all predictions, without k-fold cross-validation (k parameter set to 1). Following common practice in the literature (Pearce and Wiggins, 2006), the LTM was pre-trained on a corpus consisting of 152 Canadian folk songs and ballads (Creighton, 1966), 185 of the chorale melodies harmonized by J. S. Bach (Riemenschneider, 1941), and 566 German folk songs (dataset fink) from the Essen Folk Song Collection (Schaffrath, 1992, 1994, 1995). This collection of music will hereafter be referred to as the "IDyOM corpus".

For modelling human data from Cuddy and Lunney (1995) only `cpitch` was used as the source viewpoint. The justification for this is that the short (two-note) context of the stimuli does not permit the inference of higher-level representations (modelled by derived viewpoints), such as the scale-degree of the notes. The longer stimuli from Schellenberg (1996) allowed for the exploration of more adventurous viewpoint choices. `((cpint cpintfref))`—a linked viewpoint combining chromatic pitch interval and chromatic interval from tonic—was investigated, as well as `(cpintfib cpintfip (cpint cpintfref) (cpitch ioi))`, combining probability distributions from four predictors capturing chromatic interval from the first note in the bar and piece, the aforementioned linked viewpoint, and another linked viewpoint with MIDI pitch number and onset time. The latter was chosen due to previous findings suggesting it is the optimal viewpoint combination for modelling this dataset (Pearce and Wiggins, 2006).

From the resulting probability distributions, the appropriate 25 chromatic tones were chosen to be mapped onto the human ratings from the Cuddy and Lunney dataset, and the 15 relevant diatonic pitches were selected for each stimulus in the Schellenberg dataset. The information content, IC, given by $h = -\log_2 p$, was calculated for each of these extracted probabilities.

D. Music Transformer

General model outline: The Music Transformer introduced by Huang et al. (2018) is based on prior work by Vaswani et al. (2017) on the original Transformer, which found that a sequence model based on self-attention could achieve strong performance in generative tasks that required long-range coherence. Music has multiple dimensions (such as timing and pitch) for which relative differences are often at least as important as absolute ones. Thus, the original Transformer, which relied on representations of absolute positions, was modified to consider representations of the relative positions, or distances between sequence elements by incorporating relative self-attention, first introduced in Shaw et al. (2018). A further alteration to this mechanism was made via an algorithmic improvement which greatly reduced the memory requirements of the system from $O(L^2D)$ to $O(LD)$, enabling it to parsimoniously process and train on long musical sequences.

The resulting Music Transformer ingests symbolically represented music tokenised into a stream of events containing dynamics, pitch, time-shift and note-off information, making up a vocabulary consisting of 128 `NOTE_ON` events, 128 `NOTE_OFF` events, 100 `TIME_SHIFT` tokens (allowing expressive timing at 10 ms), and 32 `SET_VELOCITY` bins for dynamics (Oore et al., 2020). It then processes the left-context with stacked decoder layers using masked multi-head relative self-attention and at each step produces a softmax over the aforementioned vocabulary.

A piano roll visualisation of how the Music Transformer uses relative-self attention to identify suitable continuations using information from prior musical events is shown in Figure 1. The notes highlighted in are those that receive the highest softmax probability, the thickness of the lines corresponds to the weight of the softmax probability, and the different colours indicate different attention heads. This example calls for a prediction on a piece with a recurring up-and-down contour at a local melodic apex. The model's attention to these details is evident through the prioritisation of other past melodic peaks as precedent for this prediction, as well as the consideration of nearby voice-leading and left-hand harmony.

Application in expectation modelling: For the purpose of expectation modelling, the next-token probability distribution given by the Music Transformer at each step is taken to indicate the model's expectation of suitable continuations for the upcoming musical event, which is to be compared to probe-tone human ratings and IDyOM predictions. In Kern et al., predictions for all pitch-related `NOTE_ON` tokens were made using the entire left-context token stream (including tokens providing information other than pitch) to simulate the note-by-note prediction needed to match their experimental paradigm. In the present work, this approach was taken for ingested MIDI stimuli with arbitrary added probe-tones, and the probability distribution output for the probe-tone was renormalised over the NOTE_ON token subset. As with IDyOM, the relevant MIDI notes for each experiment were

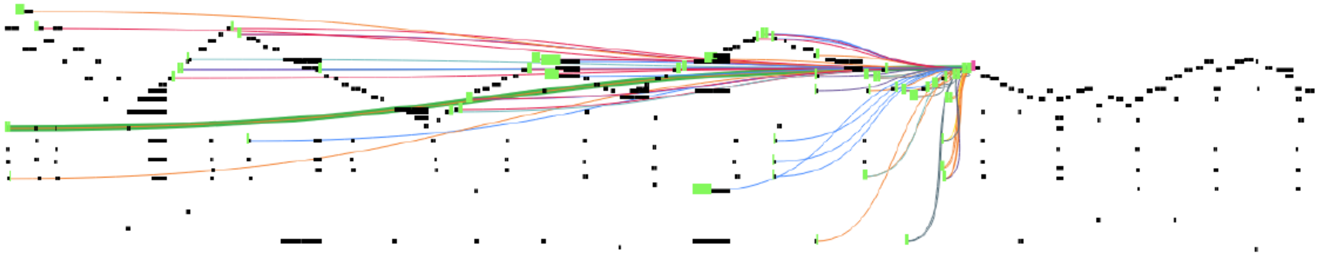

Fig. 1: A piano-roll visualisation of which musical events most strongly affect a prediction made by the Music Transformer from Huang et al. (2018)

extracted and their corresponding IC values were calculated.

Training: Initial training of the Music Transformer was carried out on the MAESTRO dataset (Hawthorne et al., 2018) for 150 epochs. The chosen training parameters matched those used in Kern et al. (2022) (learning rate = 0.1, batch size = 2, number of layers = 6, number of attention heads = 6, dropout rate = 0.1), which were the parameters selected by Huang et al. (2018) as optimal for realistic music generation. As in Huang et al., cross-entropy loss was used to monitor training progress. This choice of loss function has the effect of finding weights that minimise the upcoming notes' surprisal. The weights from epoch 144 were chosen, as these gave the lowest validation loss encountered during training, avoiding significant overfitting - by comparison epoch 150 was chosen in Kern et al, with the similar selection allowing for cross-study comparison. The fine-tuning approach also matched that of the latter study, training on the Monophonic Corpus of Complete Compositions (MCCC) (https://osf.io/dg7ms/) for 40 epochs, with weights from epoch 33 being chosen for subsequent use in probability distribution estimation (epoch 21 was selected in Kern et al.).

E. Statistical Analysis of Predictions

Ordinary least squares (OLS) regression analyses were carried out to judge the fit between model predictions for the relevant extracted pitches (taken to be the independent variable) and human ratings (dependent variable). Both R^2 and r (Pearson correlation coefficient) were obtained; the former quantified the variance explained, while the latter was used in Steiger's test (see later step). For regressions on the Cuddy and Lunney (1995) human responses, the ratings of each MIDI probe-tone option by musicians and non-musicians were merged by taking the average of each pair, since the authors found no significant statistical difference between the predictions of the two groups.

Further linear regression analyses were carried out for each set of ratings by using the predictions from both IDyOM and the Music Transformer as predictors for the human ratings. The resulting R^2 value was used to calculate:

$$R^2_{\text{two predictors}} - R^2_{\text{IDyOM}} = V_{\text{Transf}},$$

where V_{Transf} stands for variance in human ratings uniquely explained by the Transformer (and vice versa for obtaining the unique variance accounted for by IDyOM).

Moreover, in order to determine whether the difference between the correlations for IDyOM–human and Transformer–human was significant, Steiger's test for dependent correlations with one variable in common was used ($\alpha = 0.05$, two-tailed) (Steiger, 1980), via the online tool available at http://comparingcorrelations.org/ (Diedenhofen and Cocor, 2015). To facilitate this, a regression analysis with the Transformer predictions as the predictor and IDyOM predictions as the target was carried out, and the sample size for each experiment was taken to be the number of stimuli multiplied by the number of rated probe-tones per stimulus ($n = 200$ for Cuddy and Lunney and $n = 120$ for Schellenberg).

III. Results

A. Cuddy & Lunney (1995)

Table II shows the regression results for the Cuddy & Lunney experiment. When considered separately, both IDyOM and the Music Transformer significantly predicted human probe-tone ratings, with the Transformer accounting for a larger proportion of variance ($R^2 = 0.559$) than IDyOM ($R^2 = 0.367$). The two-predictor regression indicated that, taken together, the models explained 58.1% of the variance in the human data. Examining the regression coefficients suggests that both predictors made significant contributions, with the Transformer's slope larger in magnitude than that of IDyOM.

B. Schellenberg (1996)

The results for the Schellenberg dataset are summarised in Table III. Three IDyOM variants with different viewpoint selections were used, as already mentioned. Variant 1 had the basic (`cpitch`) as its source viewpoint, Variant 2 used `((cpint cpintfref))`, and Variant 3 used `(cpintfib cpintfip (cpint cpintfref) (cpitch ioi))`. The 3 IDyOM variants and the Music Transformer again predicted human ratings significantly. The amount of explained variance varied across variants, with single-predictor R^2 values for IDyOM ranging from 0.554 to 0.729, compared to 0.664 for the Transformer across all configurations. When both models were included simultaneously, explained variance increased to between 71.4% and 78.6%, indicating that the two models provide complementary information. Regression coefficients were again significant in

TABLE II: Regression results for the Cuddy & Lunney (1995) dataset. Reported are the Pearson correlation r (for single-predictor regressions) or the multiple correlation $R = \sqrt{R^2}$ (for the two-predictor regression), the proportion of explained variance (R^2), and unstandardised regression coefficients (slope β) with standard errors.

Model	r / R	R^2	β (SE)	p
IDyOM only	−0.606	0.367	−0.224 (0.021)	< .001
Transformer only	−0.748	0.559	−0.289 (0.018)	< .001
Both predictors	0.762	0.581	−0.074 (0.023); −0.238 (0.024)	< .001; < .001

Note. For both predictors, β and p are written in the order IDyOM; Transformer.

TABLE III: Regression results for the Schellenberg (1996) dataset across three IDyOM configurations (IDyOM 1–3). Reported are the Pearson correlation r (for single-predictor regressions) or the multiple correlation $R = \sqrt{R^2}$ (for the two-predictor regression), the proportion of explained variance (R^2), and unstandardised regression coefficients (slope β) with standard errors.

IDyOM variant	Model	r / R	R^2	β (SE)	p
IDyOM 1					
	IDyOM only	−0.824	0.679	−0.330 (0.021)	< .001
	Transformer only	−0.815	0.664	−0.320 (0.021)	< .001
	Both predictors	0.884	0.782	−0.198 (0.025); −0.181 (0.024)	< .001; < .001
IDyOM 2					
	IDyOM only	−0.744	0.554	−0.311 (0.026)	< .001
	Transformer only	−0.815	0.664	−0.320 (0.021)	< .001
	Both predictors	0.845	0.714	−0.135 (0.030); −0.228 (0.028)	< .001; < .001
IDyOM 3					
	IDyOM only	−0.854	0.729	−0.345 (0.019)	< .001
	Transformer only	−0.815	0.664	−0.320 (0.021)	< .001
	Both predictors	0.887	0.786	−0.225 (0.028); −0.150 (0.027)	< .001; < .001

Note. For both predictors, β and p are written in the order IDyOM; Transformer.

TABLE IV: Unique variance partitioning results. Reported are R^2 from the two-predictor regression (IDyOM + Transformer), and the unique variance contributions of each model ($R^2_{\text{both}} - R^2_{\text{other}}$).

Experiment	IDyOM Variant	R^2 (both)	Unique IDyOM	Unique Transformer
C & L	1	0.581	0.022	0.214
Schell	1	0.782	0.118	0.103
Schell	2	0.714	0.050	0.160
Schell	3	0.786	0.122	0.057

TABLE V: Results of Steiger's test comparing IDyOM–human vs. Transformer–human correlations (I stands for IDyOM, T for Music Transformer, and H for human ratings). Reported are the three correlations, the test statistic Z, and associated p-value.

Experiment	r(I,H)	r(T,H)	r(T,I)	Z ; p
C & L	-0.606	-0.748	0.658	3.5796 ; 0.003
Schell (IDyOM 1)	-0.824	-0.815	0.719	0.2685 ; 0.7883
Schell (IDyOM 2)	-0.744	-0.815	0.724	-1.8755 ; 0.0607
Schell (IDyOM 3)	-0.854	-0.815	0.779	1.3385 ; 0.1807

all cases, with relative magnitudes differing across IDyOM variants.

C. Unique Variance Partitioning

To measure the unique variance accounted for by each model, we compared the R^2 of the two-predictor regression with those of the single-predictor models (Table IV). For Cuddy & Lunney, the Transformer uniquely explained 21.4% of variance in the ratings, compared to only 2.2% uniquely explained by IDyOM. For Schellenberg, the unique contributions depended on the IDyOM configuration. In the optimal viewpoint (IDyOM 3), IDyOM accounted for 12.2% unique variance, while the Transformer accounted for 5.7%. IDyOM 1 also explained slightly more unique variance than the Transformer but by a much smaller margin. By contrast, the Transformer picked up more distinct variance than IDyOM 2.

D. Comparison of Correlations (Steiger's Test)

Finally, the results of Steiger's test, used to determine whether the correlation between IDyOM and human ratings differed significantly from that between the Transformer and human ratings, are shown in Table V. In the Cuddy & Lunney dataset, the difference was significant ($Z = 3.58$, $p = .003$), with the Transformer outperforming IDyOM. In the Schellenberg dataset none of the comparisons showed significant difference, with the results for IDyOM 2 coming the closest to significance ($p = .061$) in favour of the Transformer. This suggests that, while the Transformer had a clear advantage in the Cuddy & Lunney dataset, its performance relative to IDyOM in the Schellenberg dataset depended more on the specific IDyOM viewpoint selection made, but without significant differences in any of the three cases.

IV. Discussion

The preceding analyses provide evidence that both models capture important aspects of melodic expectation, with some notable differences in their strengths and unique contributions when predicting probe-tone data.

A. Cuddy & Lunney (1995)

For the Cuddy & Lunney dataset, the Transformer substantially outperformed IDyOM. While both models explained significant variance in human ratings, the Transformer uniquely accounted for more than 20% of variance beyond IDyOM, whereas IDyOM's unique contribution was negligible. Steiger's test confirmed that the Transformer–human correlation was significantly stronger than the IDyOM–human correlation. We can therefore conclude that the Transformer might generally be better-suited to modelling expectation with very short contexts which do not leave space for the inference of higher-order musical representations, such as a tonal key.

It should also be noted that the existence of very minimal context in this case means that there is a smaller scope for accurate prediction of human ratings, reflected by the lower proportion of variance in human data accounted for by both models compared to the Schellenberg data. This is in line with prior work in modelling note-by-note prediction, which showed increased uncertainty (or entropy) by both computational models and humans when predicting the opening few notes of a piece, where context is limited (Huang et al., 2018) (Pearce and Wiggins, 2006, Experiment 3). In that light, the conclusion that IDyOM's performance is worse might be somewhat softened by the dataset's limited predictive potential, as model performance here may be partly bounded by the nature of the stimuli themselves rather than solely by differences in modelling capacity.

B. Schellenberg (1996)

In contrast, results for the Schellenberg dataset revealed a more nuanced picture. Both models made significant contributions in all cases, but IDyOM configured to track chromatic pitch interval and chromatic interval from tonic was revealed to be weaker than the other three models, even the very simple IDyOM 1 which was limited to tracking MIDI pitch numbers. One possible explanation for this is that these attributes might be perceived through separate mechanisms, which in psychological terms would mean that our sense of whether or not a note fits the underlying key is derived independently from our sense of the relative pitch of that note to the previous one. Investigating the performance of two models capturing these attributes separately—`(cpint cpintref)` instead of `((cpint cpintref))` in IDyOM terms—whose distributions are combined to generate IDyOM's final output would help shine a light on this question.

The strongest performing IDyOM variant was the one which reflected the optimal viewpoint selection for this dataset uncovered in previous work, as already mentioned. Yet, the lack of significance in its slightly superior performance compared to the Music Transformer suggests that both models capture expectation almost equally well in this case, with most of their explanatory power being overlapping. This suggests that the Transformer's representation of the musical structure is heavily influenced by the features that make up this IDyOM viewpoint, including pitch relations to the first note of the bar, as well as tonal key and note timing information.

C. Theoretical Implications

Taken together, the results showed that the Music Transformer's sophisticated modelling utilising relative self-attention did not improve on the statistical learning approach provided by IDyOM's variable-order Markov modelling when it came to capturing pitch expectations in longer melodic contexts, but with some advantage of arguably limited broader theoretical significance for shorter contexts. The partly non-overlapping contributions made by the two models resonate with broader cognitive theories in which expectation is shaped both by local statistical regularities (well modelled by IDyOM) and by more global structural constraints (potentially better captured by deep learning models such as the Transformer). The Transformer's strong performance, comparable to that of a well-established statistical learning model, highlights the potential value of further exploring its capabilities in the context of modelling musical expectation.

D. Limitations and Future Work

A number of limitations and directions for future work can be highlighted. Firstly, the evaluation was restricted to two behavioural datasets; the generalisability of these findings to other corpora remains to be established. A natural first extension would be to compare the models' performance on note-by-note expectation modelling, for example by following Experiment 3 in Pearce and Wiggins (2006), which tested expectation in Bach chorale melodies. This would serve to further comment on the findings of Kern et al. (2022), which also used a note-by-note experimental paradigm, but in this case modelling neural responses.

A subsequent possible extension would be to further explore different choices of IDyOM viewpoints. This might be a fruitful endeavour for two reasons. First, testing more different viewpoint combinations can unveil more about the structural representations which most strongly inform listeners' expectations. In addition, comparisons between the Music Transformer and IDyOM with different viewpoint selections can help render the Transformer's predictions more explainable. By testing different viewpoint combinations and quantifying the overlapping variance with the Transformer in each case, more and less relevant structural representations can be uncovered. In particular, complex model combinations can be stripped down to simpler single or linked viewpoints to narrow down on the significance of each individual component; for example, following the current study, predictions can be carried out using each of IDyOM 3's individual four components in isolation and compared with the Transformer each time to check for their respective overlapping variances.

Furthermore, in the current study, IDyOM was trained on the same corpus as in most previous investigations by Pearce and others. However, we might also want to attempt to get as close as possible to having IDyOM and the Music Transformer trained on the same dataset, to simulate a common musical enculturation across the two models. Kern et al. (2022) made meaningful progress towards that end by training IDyOM on the MCCC. This approach was also attempted here but ran into an insurmountable obstacle: the range of pitches required to model probe-tone ratings for the Cuddy and Lunney and Schellenberg stimuli stretched beyond the range of pitches present in the MCCC. This is a problem because IDyOM, by design, is not able to generate predictions for pitches not present in its training corpus (it is likely that Kern et al. overcame this by choosing a dataset within the acceptable melodic range, a task that would have been made easier by their chosen method of note-by-note rating, which, unlike in the case the probe-tone experiments considered here, does not necessitate an additional octave on either side of the stimulis' range to be included in the set of tones for which probabilistic predictions are generated). Thus, to ensure a closer comparison with the findings by Kern et al., future research could seek behavioural data collected for stimuli whose notes are all present in the MCCC. An alternative extension would be to use the IDyOM corpus as the fine-tuning dataset for the Music Transformer. This would not offer as direct a comparison with the work of Kern et al., but it would nevertheless be a valuable exploration of the two models. Another proposed investigation that might appear natural would be to train IDyOM and the Music Transformer on the same corpus, without any prior training of the latter on the MAESTRO dataset. While the simulation of identical prior listening histories between the two models might seem appealing, it is likely infeasible due to the very large size of datasets known to be required for the sufficient training of transformers, and would likely result in poor performance from the Transformer.

Another important point to mention is that both Kern et al. (2022) and Robert et al. (2024) cite IDyOM's inability to model polyphonic music when explaining their decision to compare it with their respective deep neural models on monophonic data. However, the latest version of IDyOM is able to ingest simultaneous musical events and derive harmonic representations, as detailed in Pearce (2025). The predictions of neural models for appropriate stimuli can therefore be compared to those of IDyOM, leaving open the possibility to gain new insights into how polyphonic music is represented in the brain.

Finally, an ideal goal for future work would be to employ more realistic musical stimuli incorporating variable dynamics, expressive timing, diverse timbres, and ultimately audio signals, to help us develop a deeper understanding of expectation in naturalistic listening. In particular, this would allow for the incorporation of lower-level acoustic features, thus accounting for a bigger portion of the aforementioned sensory-cognitive continuum and resulting in a richer picture of musical expectation, due to the unique contributions which have been shown to come from sensory processes. While such extensions would be beyond the scope of IDyOM in its current form, the representational capabilities of transformers could make them adept at handling more naturalistic musical input.

References

Cheung, V. K. M., Harrison, P. M. C., Koelsch, S., Pearce, M. T., Friederici, A. D. and Meyer, L. (2024), 'Cognitive and sensory expectations independently shape musical expectancy and pleasure', *Philosophical Transactions of the Royal Society B: Biological Sciences* **379**(1895), 20220420.

Collins, T., Tillmann, B., Barrett, F. S., Delbé, C. and Janata, P. (2014), 'A combined model of sensory and cognitive representations underlying tonal expectations in music: From audio signals to behavior', *Psychological Review* **121**(1), 33–65.

Creighton, H. (1966), *Songs and Ballads from Nova Scotia*, Dover Publications, New York.

Cuddy, L. L. and Lunney, C. A. (1995), 'Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity', *Perception & Psychophysics* **57**(4), 451–462.

Diedenhofen, B. and Cocor, J. M. (2015), 'A comprehensive solution for the statistical comparison of correlations., 2015, 10, e0121945', *DOI: https://doi.org/10.1371/journal. pone* **121945**.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J. and Eck, D. (2018), 'Enabling factorized piano music modeling and generation with the maestro dataset', *arXiv preprint arXiv:1810.12247* .

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', *arXiv preprint arXiv:1809.04281* .

Huron, D. (2008), *Sweet anticipation: Music and the psychology of expectation*, MIT press.

Juslin, P. N., Sakka, L. S., Barradas, G. T. and Lartillot, O. (2022), 'Emotions, mechanisms, and individual differences in music listening: A stratified random sampling approach', *Music Perception* **40**(1), 55–86.

Kern, P., Heilbron, M., de Lange, F. P. and Spaak, E. (2022), 'Cortical activity during naturalistic music listening reflects short-range predictions based on long-term experience', *Elife* **11**, e80935.

Krumhansl, C. L. (1990), 'Tonal hierarchies and rare intervals in music cognition', *Music Perception* **7**(3), 309–324.

Meyer, L. B. (1956), *Emotion and Meaning in Music*, University of Chicago Press, Chicago.

Morgan, E., Fogel, A., Nair, A. and Patel, A. D. (2019), 'Statistical learning and gestalt-like principles predict melodic expectations', *Cognition* **189**, 23–34.

Narmour, E. (1990), *The analysis and cognition of basic melodic structures: The implication-realization model.*, University of Chicago Press.

Narmour, E. (1992), *The analysis and cognition of melodic complexity: The implication-realization model*, University of Chicago Press.

Oore, S., Simon, I., Dieleman, S., Eck, D. and Simonyan, K. (2020), 'This time with feeling: Learning expressive musical performance', *Neural Computing and Applications* **32**(4), 955–967.

Pearce, M. T. (2005), The construction and evaluation of statistical models of melodic structure in music perception and composition, PhD thesis, City University London.

Pearce, M. T. (2025), *Learning to Listen, Listening to Learn: Music Perception and the Psychology of Enculturation*, Oxford University Press.

Pearce, M. T. and Wiggins, G. A. (2006), 'Expectation in melody: The influence of context and learning', *Music Perception* **23**(5), 377–405.

Riemenschneider, A. (1941), *371 Harmonized Chorales and 69 Chorale Melodies with Figured Bass*, G. Schirmer, New York.

Robert, P., Van Cang, M. P., Mercier, M., Trébuchon, A., Bartolomei, F., Arnal, L. H., Morillon, B. and Doelling, K. (2024), 'Multi-stream predictions in human auditory cortex during natural music listening', *bioRxiv* pp. 2024–11.

Savage, P. E., Brown, S., Sakai, E. and Currie, T. E. (2015), 'Statistical universals reveal the structures and functions of human music', *Proceedings of the National Academy of Sciences* **112**(29), 8987–8992.

Schaffrath, H. (1992), 'The ESAC databases and MAPPET software', *Computing in Musicology* **8**.

Schaffrath, H. (1994), 'The ESAC electronic songbooks', *Computing in Musicology* **9**.

Schaffrath, H. (1995), The essen folksong collection in kern format, Technical report, Center for Computer Assisted Research in the Humanities (CCARH), Menlo Park, CA.

Schellenberg, E. G. (1996), 'Expectancy in melody: Tests of the implication-realization model', *Cognition* **58**(1), 75–125.

Schellenberg, E. G. (1997), 'Simplifying the implication-realization model of melodic expectancy', *Music Perception* **14**(3), 295–318.

Shaw, P., Uszkoreit, J. and Vaswani, A. (2018), 'Self-attention with relative position representations', *arXiv preprint arXiv:1803.02155* .

Steiger, J. H. (1980), 'Tests for comparing elements of a correlation matrix.', *Psychological bulletin* **87**(2), 245.

Temperley, D. (2008), 'A probabilistic model of melody perception', *Cognitive Science* **32**(2), 418–444.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.