Relating Objective Complexity, Subjective Complexity and Beauty

Surabhi S Nath[1,2], Franziska Brändle[1], Eric Schulz[1], Peter Dayan[*,1,3], Aenne Brielmann[*1]

[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany

[2]Max Planck School of Cognition, Stephanstrasse 1a, Leipzig, Germany

[3]University of Tübingen, Tübingen, Germany

[*]Joint last authors

Author Note

Surabhi S Nath https://orcid.org/0000-0002-0569-4908

Peter Dayan https://orcid.org/0000-0003-3476-1839

Aenne Brielmann https://orcid.org/0000-0002-6742-2836

We have no conflict of interest to disclose.

This work was supported by the Max Planck Society, the Alexander von Humboldt Foundation, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—461354985.

Part of this paper is available online as part of the corresponding author's Masters thesis. This work has been updated and extended for inclusion in the current article.

Correspondence concerning this article should be addressed to Surabhi S Nath, AGPD, Max Planck Institute for Biological Cybernetics, Max-Planck-Ring 8, 72076 Tübingen, Germany. E-mail: surabhi.nath@tuebingen.mpg.de

Draft version 1.0, 04/03/2023

Abstract

The complexity of images critically influences our impression of them and our assessment of their beauty. However, there is no consensus on the best way to formalize an objective measure of complexity for images. Moreover, studies relating subjective assessments of complexity and beauty to potential objective measures are hampered by the use of hand-crafted stimuli from which it is hard to generalize. To tackle these issues, we generated 2D black-and-white patterns algorithmically using cellular automata, and collected the ratings of 80 participants of their subjective complexity and beauty. We then assessed the relationship between beauty and complexity ratings, and objective measures of complexity such as density, asymmetry, entropy, local spatial complexity, and (approximate) Kolmogorov complexity. We also introduced an "intricacy" measure that quantifies the number of components in a pattern using a graph-based approach. We found that a weighted combination of local spatial complexity and intricacy was an effective predictor ($R^2_{test} = 0.46$) of subjective complexity. This implies that people's complexity ratings depend on the number of distinct elements in the pattern along with the elements' local spatial distribution and therefore that global and local image features are integrated to determine complexity judgements. Furthermore, we found a positive linear relationship between beauty and complexity ratings, with a negative linear influence of disorder, namely asymmetry and entropy, and a negative interaction between the two (with a total predictive accuracy of $R^2_{test} = 0.64$). This implies that there is beauty in complexity as long as there is sufficient order in the form of low asymmetry and randomness. In addition, a moderated mediation analysis showed that subjective complexity mediates the influence of objective complexity on beauty at all levels of disorder. Lastly, we found some evidence for individual differences with different people displaying different degrees of preference towards intricacy (in their complexity assessments) and dislike of disorder (in their beauty assessments).

Keywords: empirical aesthetics, cellular automata, objective complexity, subjective complexity, beauty

1. Background and Introduction

What makes some geometric patterns such as Islamic tile designs more beautiful than others, such as QR-codes? Complexity is one factor that influences a pattern's beauty. Complexity and beauty greatly impact our sensory experiences. Researchers in the empirical aesthetics community have therefore tried to measure and quantify beauty and complexity. Understanding the relationship between subjective beauty, complexity and objective image features, and the relationship between the beauty and complexity assessments themselves has been a topic of great interest over the last several decades (McWhinnie, 1971; Machotka, 1980; Jacobsen, 2010; Van Geert & Wagemans, 2020; Nadal & Vartanian, 2021; Chamberlain, 2022). While progress has been made, common practices such as using handcrafted stimuli and measures have posed major challenges to encapsulating findings across studies.

Complexity: The complexity literature includes several attempts to arrive at an understanding of the basis of human subjective ratings of complexity. Although complete consensus has not been reached, two prominent streams of work emerge – one based on qualitative measures which focusses on identifying separable object features that contribute to experienced complexity, and the other based on quantitative measures, which focuses on the objects' statistical properties.

We will first discuss the first stream of work, comprising of a line of studies that has attempted to delineate relevant object features that contribute to complexity. These features include the density, number and variety of elements (such as vertices, lines, turns or sides), colours and variety of colours in the image (Chikhman et al., 2012; Friedenberg and Liby, 2016; Munsinger and Kessen, 1964; Tinio and Leder, 2009). The numbers of vertices, sides and independent elements in regular geometric stimuli have been shown to be good predictors of their subjective complexity (Arnoult, 1960; Attneave, 1957; Berlyne et al., 1968; Hall, 1969). The presence of symmetry has been noted to reduce the perceived complexity (Arnoult, 1960; Attneave, 1957; Day, 1967, 1968; Eisenman and Gellens, 1968), whereas broken symmetries increased perceived complexity (Gartus and Leder, 2013).

In an integrative contribution, Nadal (2007) proposed seven measures of complexity that related to subjective complexity judgements: unintelligibility of elements, disorganisation, number of elements, variety of elements, asymmetry, variety of colours and degree of three-dimensional appearance. Subjective complexity was reliably predicted using one or two of these seven dimensions, however, they varied according to the type of stimulus. While the number of elements was the most informative predictor, variety of colours and three-dimensional appearance had little influence on complexity judgements. These seven measures were further decomposed into three factors: (1) an elements factor comprising of number and variety of elements, (2) an organization factor, comprising of unintelligibility of elements and disorganisation, and (3) asymmetry, with each factor relating to different perceptual and cognitive processes.

Parallely, the works of Chipmann (1977), Chipmann and Mendelson (1979) proposed that subjective complexity was determined by two components – a quantity-based component derived from the number of vertices in the stimuli and a structure-based component based on the structural aspects like symmetry, repetitions etc. They also indicated that these two components might involve different cognitive mechanisms. Additional evidence for this dual processing theory came from Ichikawa (1985) who attributed different levels of processing to these two components – a lower level/primary processing for quantity-based features and a higher-level processing for the discovery of structure.

However, while these measures have been somewhat successful in explaining human assessments of complexity, they are largely evaluated by hand, and have been hard to systematize.

By contrast, the second stream of work has focused on more quantitative correlates of subjective complexity judgements based on statistical properties of the stimuli. Fred Attneave (1919-1991) was the first to apply information theory to quantify stimulus properties in the context of aesthetics . Shannon's information theory (Shannon, 1948) conceptualised "entropy" as a measure of the quantity of information potentially contained in a signal. This was later adopted as a measure of order and complexity in aesthetics by several authors (Arnheim, 1956, 1966; Bense, 1960, 1969; Moles, 1958; Schmidhuber, 2009). Low entropy implies low uncertainty, high predictability, high order, and less complexity, and vice versa. Shannon entropy also found its way into the creative fields of art (used to measure pattern complexity in the Kolam artform Tran et al. (2021)) and music (used to measure complexity in sequences of tones (Delplanque et al., 2019)). The compressibility of an image could be used as a measure of its complexity with less compressibility implying more complexity and vice-versa (Birkin, 2010; Donderi, 2006). Kolmogorov complexity is a similar quantification of data compression determined by the length of the shortest computer program that produces a desired output in any standard universal computer programming language, and is one of the most direct applications of algorithmic information theory to stimuli description (eg. Chikhman et al., 2012; Singh and Shukla, 2017). Snodgrass (1971) demonstrated the potential of information measures in predicting subjective complexity on black-and-white pixel patterns. Javid recently introduced a measure of spatial complexity based on the probabilistic spatial distribution of pairs of pixels and also examined the applicability of Kolmogorov complexity to evaluate the complexity of pixel patterns created using aesthetic automata (Javid, 2021). Corchs et al. (2016) use computational measures quantifying spatial, frequency and colour properties to predict complexity of real-world stimuli. Other modern studies have used computational image properties such as histogram of oriented gradients (HOG), Fourier slope and fractal dimension to quantify complexity (refer to Van Geert and Wagemans (2020) for a detailed review). These measures are objective and can be programmed; however, their ability to predict subjective complexity is highly dependent on the type and nature of stimuli (Chikhman et al., 2012; Marin & Leder, 2013).

As for the type and nature of stimuli used, most of the above-mentioned works use handcrafted stimuli from which is hard to generalize. As a result, it is difficult to evaluate the validity and reliability of the measures and accompanying findings to multiple families of stimuli. There is therefore the need for generalizable experimental stimuli and quantitative measures. In this work, we satisfy this requirement with systematically generated image classes and programmatic analysis measures.

Beauty: In contrast to work focussing directly on complexity, beauty, and aesthetic evaluations more generally (aesthetic preference/liking, pleasantness, pleasure), have been the subject of a rather larger body of studies, including many attempts to model the processes leading to aesthetic evaluations. Nevertheless, these two streams of work are closely coupled, as complexity has been considered an important contributor to aesthetic evaluations since the early days of empirical aesthetics (although most modern models of aesthetic value consider the role of complexity only implicitly (Iigaya et al., 2020; Brielmann and Dayan, 2022)).

Fechner's principle of "unitary connection" suggested that pleasant stimuli express a balance of complexity and order (Cupchik, 1986), and work by Birkhoff mathematically formulated an aesthetic measure (M) that varied positively with order (O) and negatively with complexity (C), or M = O / C for polygonal figures, vases, poetry and music (Birkhoff, 1933). However, Davis (1936) criticized the measure M as being inappropriate for empirical test. Several studies aimed at testing the applicability of M have yielded a high variance in correlations between actual rankings and those given by the formula (Harsh and Beebe-Center, 1939; Eysenck, 1941). Later experiments by Eysenck suggested an empirical formula for M which yielded much higher correlations with subject rankings. He suggested an approximation where aesthetic preference varied positively with *both* order and complexity, altering the equation to M = O × C (Eysenck, 1942, 1968).

Berlyne also formulated a relationship between complexity and aesthetic preference, via a more general inverted-U relationship between hedonic value and arousal potential (the "psychological strength" or the extent to which a stimulus is capable of raising arousal (Berlyne, 1967)) of a stimulus (Berlyne, 1960). Berlyne proposed three classes of variables that determined arousal potential, namely, psychophysical variables, ecological variables and collative variables (Berlyne, 1971). Collative variables include subjective novelty, complexity and surprise. This theory was further applied to the context of art, suggesting that the collative variable of subjective complexity is one of the most significant determinants of aesthetic preference. The implication was that there is a general preference for stimuli of intermediate complexity, as described by the inverted-U shape.

Several subsequent works have attempted to reproduce this inverted-U relationship between stimulus complexity and aesthetic preference. However, there is little evidence in support of this theory, but rather contradictions (Nadal et al., 2010). While several studies supported the importance of complexity in shaping aesthetic preferences (e.g., Jacobsen and Höfel (2002); Jacobsen et al. (2006); Tinio and Leder (2009)), the type and directionality of the relationship has not been conclusively settled to date (Nadal et al., 2010). Some studies concurred with the original inverted U-relationship (Lakhal et al., 2020; Chmiel and Schubert, 2017; Eisenman, 1967; Farley and Weinstock, 1980; Gordon and Gridley, 2013; Hekkert and Van Wieringen, 1990; Madison and Schiölde, 2017; Marin et al., 2016; Munsinger and Kessen, 1964; Nasar, 2002; Nicki, 1972; Vitz, 1966), while others suggested a linear relationship (Eysenck, 1941; Day, 1967; Heath et al., 2000; Javaheri Javid, 2019; Nicki and Moss, 1975; Osborne and Farley, 1970; Stamps III, 2002; Nicki and Gale, 1977; Taylor and Eisenman, 1964). Surprisingly, a few other studies reported either a non-inverted U (Adkins and Norman, 2016; Norman et al., 2010), or no relationship at all (Messinger, 1998).

The reasons for these discrepancies include theoretical and empirical challenges similar to those for complexity studies, along with other, unique factors. As mentioned previously, first, there are contrasting methods of defining, measuring and manipulating objective complexity (Nadal et al., 2010) and second, rather diverse, and frequently handcrafted stimuli have been used (Marin and Leder, 2013), hampering generalization. Third, there are substantial individual differences in aesthetic evaluation (e.g., Aitken (1974); Güçlütürk et al. (2016); Jacobsen and Höfel (2002); Tinio and Leder (2009)), with relationships achieved on an average over participants often being different to, or masking, participant-level relationships (Aitken, 1974).

Drawing on this diverse literature, our work has two components:

Firstly, we model the subjective complexity of a class of visual inputs via programmatic objective complexity measures of density, symmetry, entropy, number of components, and information in the stimuli, along with a novel "intricacy" measure quantifying the number of separable visual elements. To achieve this, we relinquish generality and naturalness in favour of reproducibility and rigour, by developing an algorithmically defined stimulus generator which is capable of systematically producing families of abstract patterns that span a range of subjective complexities, and values of our objective complexity measures.

Secondly, we inspect the relationship between ratings of the beauty and both the subjective and objective complexity of our stimuli. While we only record beauty evaluations, we consider preference, liking and pleasantness to refer to closely related constructs and expect our results to apply more generally across various measures of aesthetic evaluation (also supported by previous findings, see Marin et al., 2016).

2. Methods

2.1 Cellular Automata for Pattern Generation

We first require a reproducible way of creating visual stimuli spanning a suitable range of subjective complexities. For this, we require an algorithm that provides us a principled method for generating a diverse set of stimuli. Furthermore, we want the generated stimuli to be exactly reproducible by providing the algorithm a set of generation parameters. To satisfy the aforementioned criteria, we use Cellular Automata (CA) to generate pattern stimuli for our tasks. CA are iterative algorithms where cells on 1D, 2D or 3D spaces are assigned states as a function of the states of their neighbouring cells, based (conventionally) on deterministic rules.

They have previously been used to generate many forms of computer art (Adamatzky and Martínez, 2016; Wolfram et al., 2002).

Definition: Cellular Automata

A 2D CA, C, is specified by a quadruple $\langle L, S, N, f \rangle$ where:

1. L is a $P \times Q$ grid with cells (i, j), $1 \leq i \leq P$, $1 \leq j \leq Q$
2. $S = \{c_1, c_2, \dots, c_k\}$ are the potential states of each cell $(i, j) \in L$. Therefore, each cell (i, j) has a state at time t denoted by $s^t_{(i,j)} \in S$
3. $N_{(i,j)} = \{(i_1, j_1), (i_2, j_2), \dots, (i_N, j_N)\}$ is the neighbourhood of cell (i, j) which can either be von Neumann/5-cell ($N = 5$) or Moore/9-cell ($N = 9$) neighbourhoods, where N is the neighbourhood size (Packard and Wolfram, 1985). N is the set of all $N_{(i,j)}$ over all cells (i, j) in the grid.
4. f is the state-transition function which computes the state of cell (i, j) at the $t+1 = s^{t+1}_{(i,j)}$ as a function of the states of the cells in its neighbourhood. Hence, $s^{t+1}_{(i,j)} = f(s^t_{(i_1,j_1)}, s^t_{(i_2,j_2)}, \dots, s^t_{(i_N,j_N)})$, where $(i_1, j_1), (i_2, j_2), \dots, (i_N, j_N) \in N_{(i,j)}$

Starting form an initial configuration of cells at time $t = 0$, the CA is iteratively updated over T timesteps.

We use binary 2D CA (k = 2). Our state space is $S = \{0,1\}$ (corresponding to the colours white and black). Taking inspiration from pixel-art sketches, we set our grid size to 15 × 15 ($P = Q = 15$). We consider simple rules and initial configurations. We use two classes of rules, conventionally called "totalistic" (Tot) or "outer-totalistic" (Otot) (Refer to Appendix I for specific details of the generation algorithm). For our initial configurations (ICs), based on work by Javid (Javaheri Javid, 2019), we use either a single central cell (IC = 1), a small disordered central region (IC = 2) or a fully random starting grid configuration (IC = 3) (Figure 1a). We limit the number of iterations to T = 40, and add every 5th pattern to the stimuli set. Therefore, each distinct set of algorithm parameter values produces 8 patterns.

We use 51 rules in total with differing combinations of neighbourhood size (5-cell/9-cell neighbourhood), rule code, rule type (totalistic/outer-totalistic) and initial configuration (IC 1, 2 or 3) resulting in a total of nearly 400 patterns. Since pattern evolutions may enter oscillating configurations, some of the generated patterns are identical. We remove such patterns from the stimuli set. Figure 1b shows an example pattern evolution based on a specified rule and Figure 2 shows some of the produced patterns. We also attempted to avoid any recognizable semantic content in our patterns, since complexity perception is largely influenced by familiarity (Forsythe et al., 2008) which could confound our findings.

Figure 1

Cellular automata (A) initial configurations (ICs): (i) IC = 1, (ii) IC = 2, (iii) IC = 3, (B) example rule (shown in box) along with example pattern updates for the central square where the outer rim of cells is kept constant

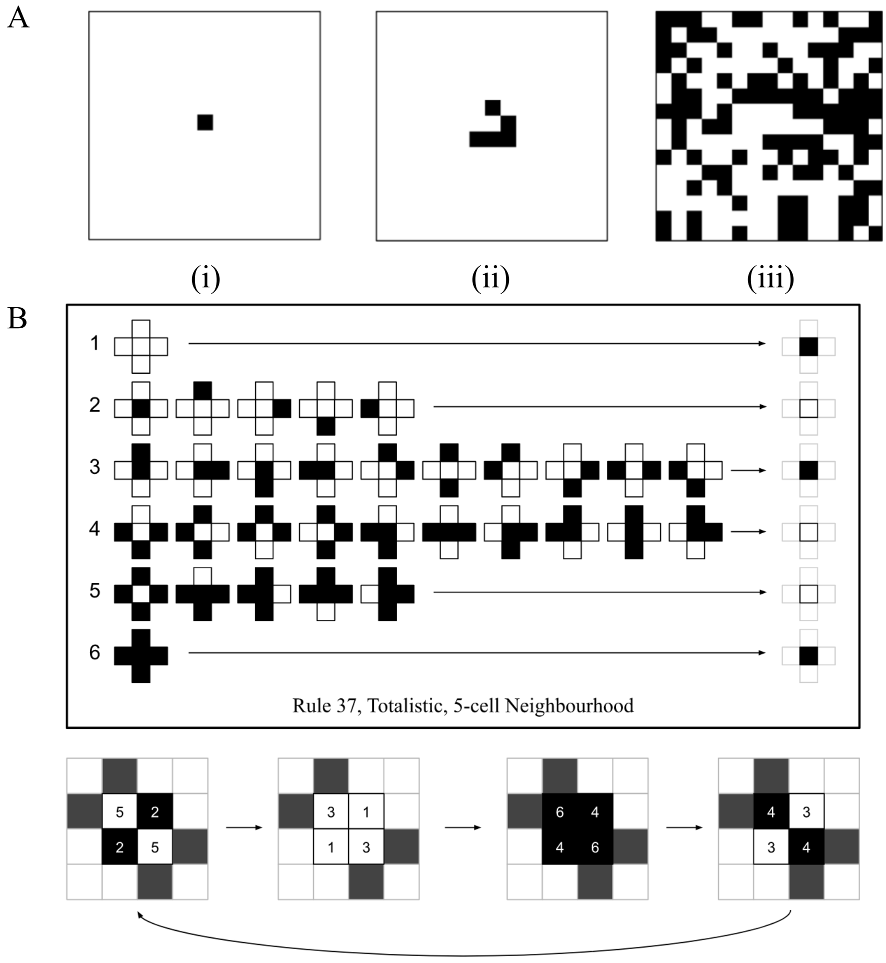
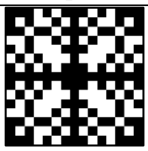
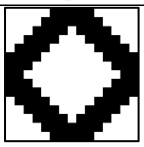
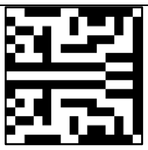
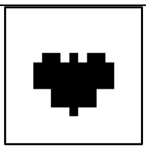
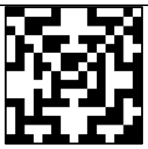
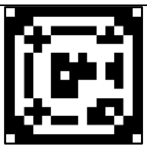
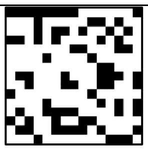
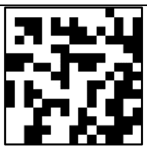
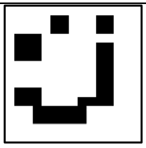
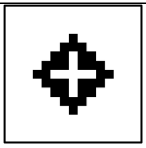
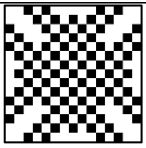
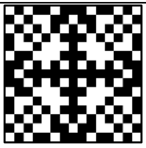


Figure 2

Cellular automata generated patterns with (rule code, IC, rule type, N, iteration)

Category	Sub-category	Example 1	Example 2
Symmetry	Full symmetry (about horizontal, vertical, diagonal and rotational axes)	 (467, 1, Otot, 5, 15)	 (510, 1, Otot, 5, 10)
	Unidirectional symmetry (about either horizontal or vertical axes)	 (736, 2, Otot, 9, 40)	 (256746, 2, Otot, 9, 10)
	Partial asymmetry	 (699054, 2, Otot, 9, 8)	 (93737, 2, Otot, 9, 5)
	Full asymmetry		

		(93737, 2, Otot, 9, 15)	(3276, 3, Otot, 9, 35)
Number of Components	Low	 (24, 3, Tot, 5, 35)	 (478, 1, Otot, 5, 5)
	High	 (452, 1, Otot, 5, 25)	 (469, 1, Otot, 5, 25)

CA provide a principled and structured method for pattern generation since it is fully specified by ⟨*L, S, N, f*⟩, we can deterministically reproduce all our patterns with algorithm parameters of state space, grid size, neighbourhood size, state-transition function, along with initial configuration and timestep. Moreover, from Figure 2, we see that the produced patterns are diverse and vary across multiple dimensions (while some of this diversity is hard to predict at the outset, some is pre-determinable for example, IC = 1 results in fully symmetric patterns while partial asymmetry patterns are obtained using IC = 2).

2.2 Computational Measures for Pattern Quantification

We defined six measures that could potentially explain subjective complexity ratings: density, entropy, local spatial complexity, (approximate) Kolmogorov complexity, intricacy, and symmetry; these are described below. We chose these six measures since they are commonly studied in the literature as potential determinants of subjective complexity (Friedenberg and Liby, 2016; Fan et al., 2022; Nadal, 2007; Attneave, 1957; Gartus and Leder, 2017; Damiano et al., 2021; Snodgrass, 1971; Arnheim, 1956, 1966; Bense, 1960, 1969; Moles, 1958; Schmidhuber, 2009; Javid, 2016; Rigau, 2008; Chikhman et al., 2012; Singh and Shukla, 2017; Silva, 2021). We do not consider other popular measures such as number of vertices or edges since as remarked in the previous section, they are hard to define for our CA patterns.

1. Density: Density is defined as the proportion of black pixels in the pattern.

2. Entropy: Entropy assesses the Shannon entropy of the density of black/white pixels. However, since entropy does not take spatial arrangement into consideration, we compute density entropies at all scales and average them (Eq. 1) (Huber, 2011).

$$-\frac{1}{ns}\sum_{s=1}^{ns}\frac{1}{nw_s}\sum_{w=1}^{nw_s} P(b)_{s,w}log_2P(b)_{s,w} + P(w)_{s,w}log_2P(w)_{s,w} \qquad (1)$$

where *ns* is the number of scales (in our case, ns = *P* = *Q* = 15), nw_s is the number of sliding windows at scale *s* (defined with overlap nw_s = (15 – *s* + 1)^2), $P(b)_{s,w}$ and $P(w)_{s,w}$ are the proportions of black and white pixels in the sliding window *w* at scale *s* respectively. $P(w)_{s,w} = 1 - P(b)_{s,w}$.

3. Local Spatial Complexity: Local Spatial Complexity (LSC) is defined as the mean information gain of pixels having homogeneous (same colour) or heterogeneous (different colour) neighbouring pixels (Javaheri Javid, 2019). This measure takes the *local* probabilistic spatial distribution of pixels into consideration. The average spatial complexity across 8 directions of pixel-neighbour pairs is evaluated (Eq. 2). However, this measure is implemented only across one scale giving it the name *local* spatial complexity.

$$LSC = \frac{1}{8}\sum_{d=1}^{8} \bar{G}_d = -\frac{1}{8}\sum_{d=1}^{8}\sum_{s_1=1}^{2}\sum_{s_2=1}^{2} P(s_1, s_2)_d \log_2 P(s_1|s_2)_d \qquad (2)$$

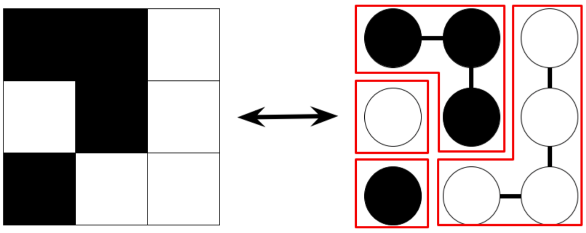
Here, d denotes the direction. s_1, s_2 denote the state combination (black-black, black-white, white-black or white-white) under consideration. $P(s_1|s_2)_d$ is the proportion that a pixel pair (along direction *d*) has states (s_1, s_2) . $P(s_1|s_2)_d$ is the probability that a pixel has state s_1 given its neighbouring pixel (along direction d) has state s_2. $\bar{G}_d$ is the mean information gain across all state combinations for direction *d*.

4. Kolmogorov Complexity: Kolmogorov complexity (KC) is a measure of algorithmic complexity based on algorithmic information theory. It is defined as the length of the shortest computer program that produces the desired pattern. Kolmogorov complexity is uncomputable, and methods to compute it only estimate an upper bound. Some such methods are the LZ78 universal compression algorithm (Ziv and Lempel, 1978) and the Block Decomposition Method (Zenil et al., 2018). In this work, we use the Block Decomposition Method.

5. Intricacy: To quantify the number of elements in a pattern, we introduce an intricacy measure using a graph-based approach. The pattern is encoded as a graph with each pixel as a node. Edges are added between neighbouring pixels of the same colour. We considered up, down, right and left (non-diagonal) adjacent pixels as valid neighbours. Depth first search is performed to count the number of connected components which is used as our intricacy measure. This value is the sum of black components and white components in the pattern. This procedure is shown in Figure 3. The range of intricacy values is 1 to 225.

Figure 3

Intricacy computation for an example pattern



Note. The graph on right is constructed from the pattern on left. Red boxes indicate connected components. Here, intricacy = 4

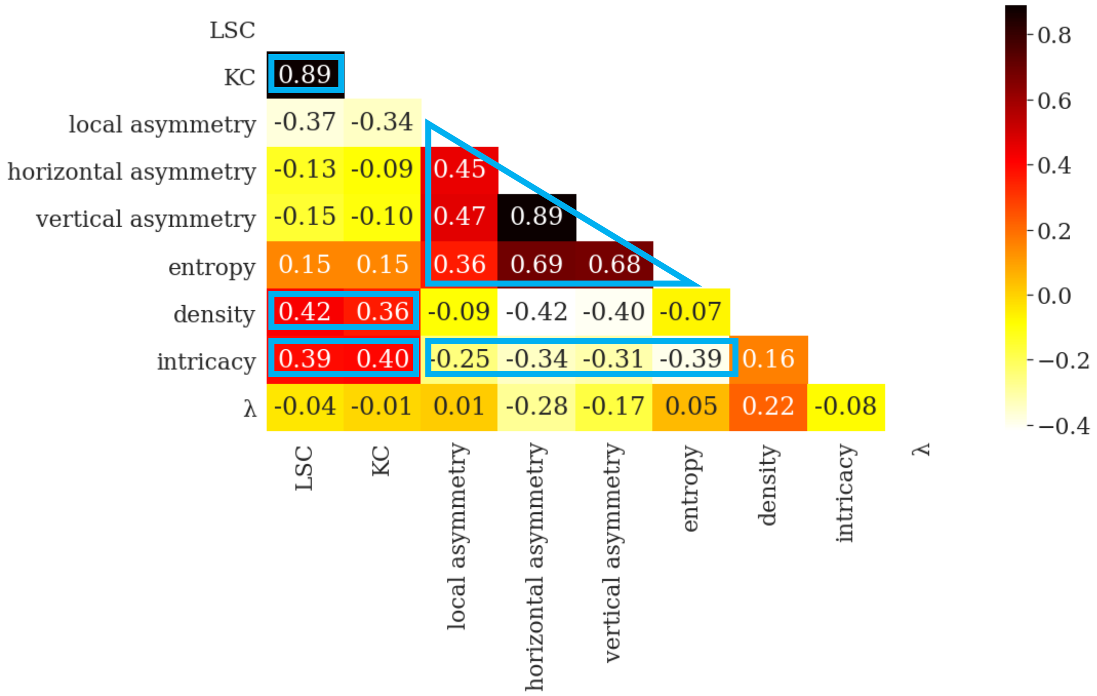
6. Symmetry: We use three measures to capture the global and local symmetry in the patterns. For global symmetry, we restrict ourselves to the horizontal and vertical directions as they are most readily perceived by humans (Giannouli, 2013). Horizontal asymmetry (Hasymm) and vertical asymmetry (Vasymm) were computed as a percentage of mismatches in the horizontal and vertical directions respectively. Our third symmetry measure, local asymmetry was computed as the average difference in mean information gains as specified by the LSC, along 4 directions.

Along with these computed metrics, we added the generation algorithm parameters to the set of potential predictors including neighbourhood size (N), rule type (totalistic (T)/outer-totalistic (O)), iteration number and IC. Furthermore, based on evidence for λ (defined as the percentage of all the entries in a rule table which map to non-zero states – in our case, 1s) being an important indicator of pattern complexity for 1D CA (Langton, 1986; Li et al., 1990), we added it as a predictor as well. λ is calculated as the number of 1s in the binary representation of the rule code. Refer to Appendix II for two additional measures we implemented along with some examples of the computed measures on various patterns.

Figure 4 shows the Pearson correlations between the programmed measures. We see a high positive correlation between LSC and KC (r = 0.89, p < 0.01, 95% CI = [0.88, 0.89]). This is consistent with Javid (2019). Intricacy is positively correlated with LSC (r = 0.39, p < 0.01, CI = [0.37, 0.42]) and KC (r = 0.40, p < 0.01, CI = [0.38, 0.42]), and negatively correlated with asymmetry measures and entropy. Density is correlated with LSC (r = 0.42, p < 0.01, CI = [0.4, 0.45]), KC (r = 0.36, p < 0.01, CI = [0.33, 0.38]). The asymmetry measures are positively correlated with each other and with entropy. Specifically, the large correlation between horizontal and vertical asymmetry is driven by the choice of rules (totalistic and outer-totalistic rules always produce patterns with bidirectional symmetry when using IC 1, and largely produce fully asymmetric patterns using IC 2 and 3). These correlation values helped us exclude variables from the predictor set, for example, we did not use both LSC and KC in the same model simultaneously, and combined the three asymmetry measures into a single mean asymmetry measure.

Figure 4

Pearson correlations between the measures



Note. The colour scale represents the strength of correlation (r), the correlations marked in blue are the correlations of interest in our work

2.3 Pattern Rating Experiment

For obtaining human ratings on the CA patterns, we programmed an online behavioural experiment where participants were recruited to view and rate the beauty and complexity of the patterns.

I) Design

We asked each participant in our experiment to rate the beauty and complexity of the patterns as they perceived them. We did not provide any definitions of complexity or beauty in order to elicit people’s unbiased opinion. However, to set the prior over the possible types and variety of patterns, we showed 12 example patterns in 2 groups of 6 patterns, where each group comprised of sufficient visual diversity. We randomized the order of the example patterns to avoid biasing participant ratings. We recorded the complexity and beauty ratings using two slider bars ranging from 0 to 100. Both slider bars were shown below the pattern at the same time. They were labelled only at their ends as either “High/Low Complexity” for the complexity rating slider and “High/Low Beauty” for the beauty rating slider, with no intermediate marking. This was done to encourage participants to rate evenly over the whole range of possible ratings. We also did not set any default value on the sliders to avoid influencing the participant ratings. We programmed the experiment in JavaScript using jsPsych (De Leeuw, 2015).

II) Stimuli

From our set of nearly 400 CA generated patterns, we selected 216 and divided them into 4 sets of 54 patterns each. The patterns in each set were picked such that they span the range of complexity values as quantified by our LSC, density and intricacy measures. The sets were manually balanced to contain visually similar (but non-identical) patterns. Participants were assigned one of the 4 sets in serial order (participant I received set (i%4)+1). For each participant, we randomly selected 6 patterns (out of the total 54) and showed them twice to get repeated measures. We ensured no two repeats occurred together in the pattern sequence. Following this design, each participant rated 60 patterns.

III) Procedure

The experiment opened with a welcome screen, followed by a consent form and data protection form. An overview page provided the task instructions. Participants then proceeded to the ratings where they were shown a pattern at the top of the screen with 2 sliders positioned below the pattern. The two sliders recorded responses to the two questions – "How complex is this pattern?" and "How beautiful is this pattern?". The order of questions remained the same across trials and across participants.

In addition to the 60 pattern ratings, each participant also encountered 2 or 3 attention checks where the pattern contained an overlaid text "read the questions below" and the slider questions were modified to read "place the slider head at the extreme right". The slider endpoint labels also changed from "Low/High Complexity", "Low/High Beauty" to "Left/Right". These two staged checks served as both a compliance and attention check. Following the ratings, participant demographic data on gender, age and nationality were recorded. We used the Vienna Art Interest and Art Knowledge questionnaire to record each participant's level of art training (Specker et al., 2020). Finally, we included some open-ended questions about rating behaviour. These questions asked the participants to indicate the strategies they used to rate complexity and beauty, and the patterns they found most complex and beautiful. We required participants to answer all questions. In all, the study took less than 20 minutes to complete.

IV) Participants

80 participants from Prolific (50 female, 29 male, 1 other; mean age = 32.3, min age = 18, max age = 79; 20 participants per set) took part in the study. 2 participants failed all attention checks and were removed from the analysis leaving us with 78 participants. All participants were based in the United States, were fluent English speakers and had not previously participated in the experiment. The average study completion time was 14.35 minutes. Each participant was paid £3.50. All experiments were approved by the ethics committee of the University of Tübingen.

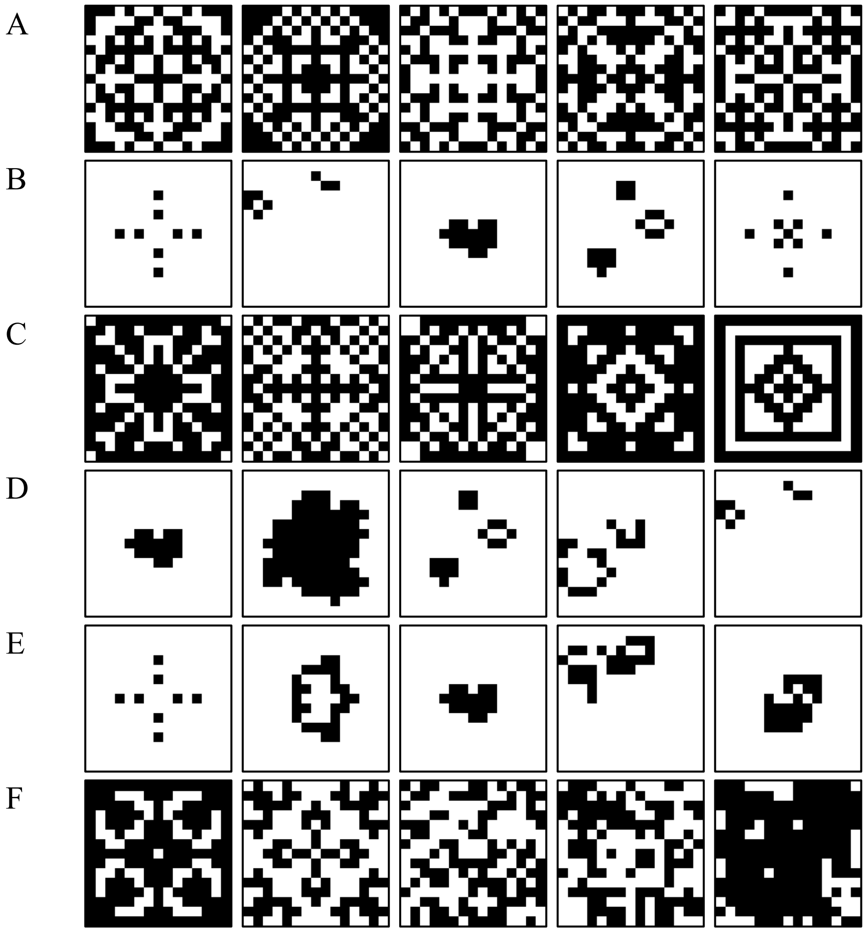
V) Analysis

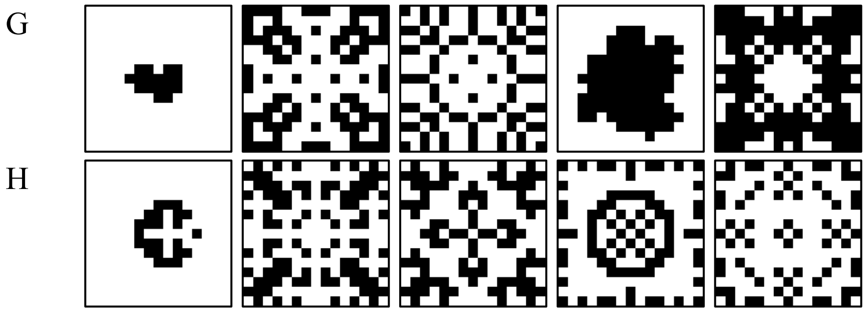
Having acquired the complexity and beauty ratings, we (1) sought a combined computational measure that could suitably predict the subject-specific complexity ratings across the population, and (2) determine the relationship between beauty and complexity ratings.

Complexity and beauty ratings were found to be balanced across sets, no significant sequential effects (trends, autocorrelation between participant ratings) were observed, and participant repeated responses were consistent (refer to Appendix II for details). In reporting their strategies, participants indicated that pattern intricacy (participants spontaneously used this word), number of elements, their position and arrangement, density and the ability to replicate a pattern influenced their complexity ratings while symmetry and "intuition" influenced their beauty ratings. Figures 5a and 5b show the patterns with the highest and lowest average complexity ratings and Figures 5c and 5d show the patterns with the highest and lowest average beauty ratings. We also studied variance in ratings per pattern and categorized patterns into "high agreement" (low variance) or "low agreement" (high variance) (Figures 5e and 5f for complexity and 5g and 5h for beauty).

Figure 5

Pattern visualizations. (A) Patterns with highest average complexity ratings, (B) Patterns with lowest average complexity ratings, (C) Patterns with highest average beauty ratings, (D) Patterns with lowest average beauty ratings, (E) Patterns with highest agreement in complexity ratings, (F) Patterns with lowest agreement in complexity ratings, (G) Patterns with highest agreement in beauty ratings, (H) Patterns with lowest agreement in beauty ratings





Metric for Complexity

We fit regression models on participant complexity ratings using R (Version 4.1.1, library lme4 (Bates et al., 2015), function lmer()). We used an incremental, bottom up approach (*i.e.*, starting from the predictors that had high linear correlations with our dependent variable and adding predictors one by one to our models (Table 1)), to arrive at the objective complexity metric that best predicted subjective complexity. High linear correlations between the ratings and the measures justify the use of linear models; hence we fit linear mixed effects models. We performed cross validation by splitting our data into 3 stratified folds. Each test fold had 20 randomly sampled ratings from every participant and each training fold had the remaining 40 ratings from every participant. The seed was set to 20 for all our experiments. We z-scored all the predictor variables. We also created additional variables by squaring each predictor to check for quadratic effects. The three asymmetry measures were combined into one mean asymmetry measure because of their high correlations. We evaluated our model using 3-fold cross validation and reported the average Akaike information criterion (AIC), Bayesian information criterion (BIC) and the average R^2 values across the three folds. We also evaluated Root Mean Square Error (RMSE) values on train and test sets. In addition, we checked for multicollinearity using Variance Inflation Factors (VIF). We experimented with random intercepts, random slopes, quadratic and interaction effects in our models. The best model was defined as the one that had significant predictors and resulted in a low BIC and high R^2 while ensuring VIFs did not exceed 5. If a model achieved higher R^2 at the expense of higher BIC, we prioritized model simplicity.

Relationship between Beauty and Complexity Ratings

Finally, to investigate the relationship between beauty and complexity ratings, we used a similar analysis to that above. We fit linear mixed effects models on the beauty ratings but now including the complexity ratings as a predictor. The objective complexity measures that best described the complexity ratings from the previous analysis were combined into one "objective complexity" measure. Since mean asymmetry and entropy were highly correlated, we combined them into one "disorder" measure. We also systematically examined the relationship between objective complexity, subjective complexity and beauty and disorder using moderated mediation analysis (using PROCESS and mediation libraries (Hayes, 2017; Tingley et al., 2014)).

3. Results

3.1 Metric for Complexity

Table 1 provides a summary of our main models and their performance (averaged over 3 cross validation folds). Our dependent variable, complexity ratings, are abbreviated as CR. We use the Wilkinson-Rogers notation to report our models (Wilkinson and Rogers, 1973).

Table 1

Summary of the main models of complexity ratings

S. no.	Model	Significance	AIC	BIC	R^2	RMSE	
						Train	Test
1	CR ~ 1 + 1|Participant		8545.4	8563.6	0.13	0.91	0.93
2	CR ~ 1 + 1|Participant + 1|Set		8546.3	8570.5	0.13	0.91	0.93
3	CR ~ LSC + 1|Participant	LSC*	7551.9	7576.1	0.37	0.77	0.79
4	CR ~ LSC + LSC|Participant	LSC*	7491.2	7527.5	0.39	0.75	0.78
5	CR ~ KC + 1|Participant	Intercept* KC*	7703.8	7728.0	0.34	0.79	0.81
6	CR ~ LSC + asymm + 1|Participant	LSC* asymm*	7525.3	7555.5	0.37	0.77	0.78
7	CR ~ LSC + entropy + 1|Participant	LSC* entropy*	7443.8	7474.0	0.39	0.76	0.78
8	CR ~ LSC + density + 1|Participant	LSC* density*	7551.3	7581.5	0.37	0.77	0.79
9	CR ~ LSC + intricacy + 1|Participant	LSC* intricacy*	7285.2	7315.4	0.42	0.74	0.76
10	CR ~ LSC + intricacy + LSC:intricacy + 1|Participant	LSC* intricacy* LSC:intricacy*	7276.8	7313.0	0.43	0.74	0.75
11	CR ~ LSC + intricacy + LSC|Participant	LSC* intricacy*	7219.8	7262.2	0.44	0.71	0.74
12	**CR ~ LSC + intricacy + intricacy|Participant**	**LSC*** **intricacy***	**7166.3**	**7208.6**	**0.46**	**0.70**	**0.73**
13	CR ~ LSC + intricacy + LSC:intricacy + intricacy|Participant	LSC* intricacy*	7163.6	7212.0	0.46	0.70	0.73
14	CR ~ LSC^2 + intricacy^2 + 1|Participant	Intercept* LSC^2* intricacy^2*	8100.3	8130.5	0.25	0.84	0.86

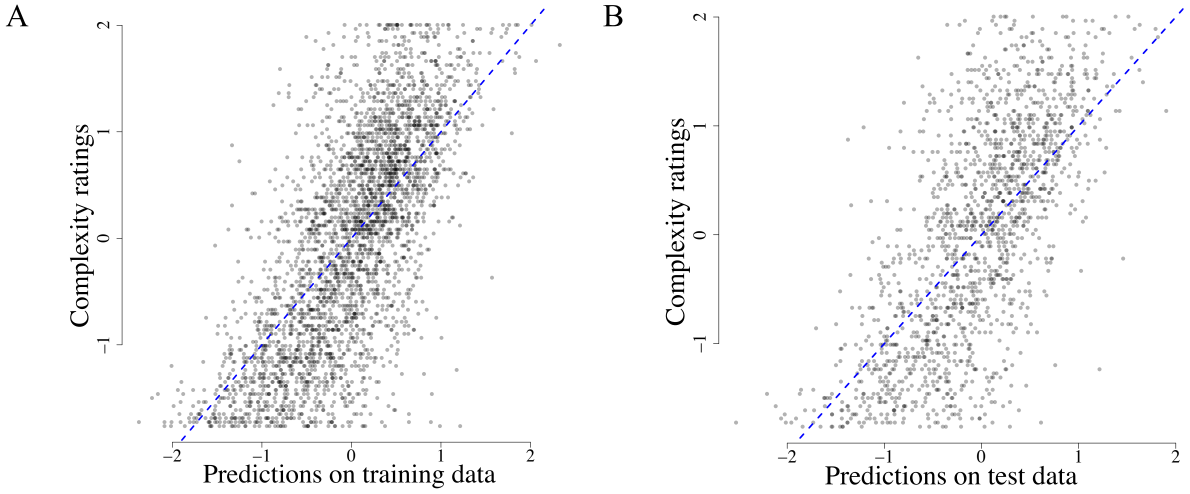
15	CR ~ LSC + LSC^2 + intricacy + intricacy^2 + 1|Participant	LSC* LSC^2* intricacy* intricacy^2*	7263.7	7306.0	0.43	0.73	0.75

Note. CR=complexity ratings, LSC=local spatial complexity, KC=Kolmogorov complexity; * indicates p < 0.05. Bold indicates best model.

Our analysis shows that subjective complexity can be predicted by a positive linear combination of LSC and intricacy measures with random slopes of intricacy and a random intercept of participant (Table 1, row 12, $R^2 = 0.46$, AIC = 7166.3, BIC = 7208.6). While the AIC for the model with an interaction effect is lower (Table 1, row 13), the BIC is higher implying possible overfitting, and so we preferred the simpler model (Table 1, row 12). Figure 6 shows the plot of predictions versus ground truth on the train and test set from cross-validation fold 1.

Figure 6

Model performance on (A) training and (B) test data from fold 1 for complexity ratings



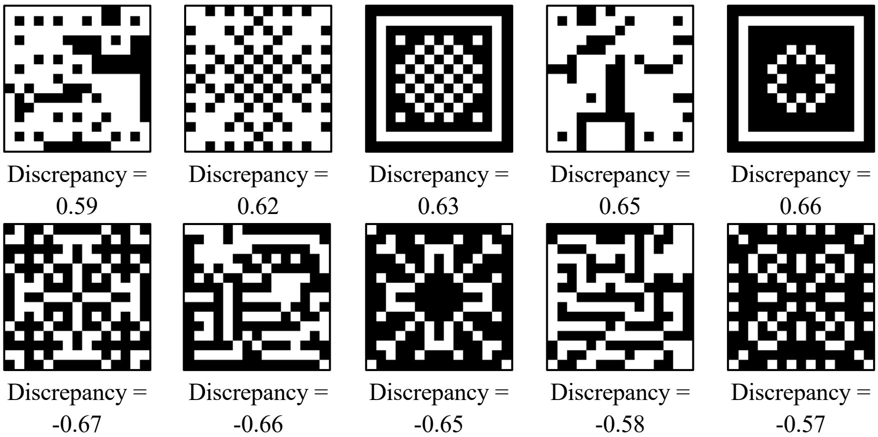
Note. y-axis displays z-scored complexity ratings and x-axis display their corresponding model predictions. Blue dashed line represents y = x.

Thus, we found evidence that a weighted combination of spatial complexity and intricacy measures can reliably explain a substantial fraction of human subjective complexity ratings. This implies that people's complexity judgments depend on the number of distinct visual elements in the pattern (captured by intricacy) along with their local spatial distribution (captured by LSC). The slopes for both quantities are positive indicating that a larger LSC and a larger intricacy are associated with higher complexity and vice versa. Since intricacy evaluates the number of connected components in the whole pattern, it is a global feature. On the other hand, LSC looks at pairwise pixel distributions at one scale and hence it is a local feature. Hence, people integrate local and global pattern features to arrive at their complexity estimates. Furthermore, the random slope of intricacy implies that intricacy influences the complexity assessments of different participants to different degrees.

We also visualized the patterns for which the difference between objective complexity measure and subject ratings was largest on average across all participants (Figure 7). We see that the measure overweighs complexity in some patterns where the computed intricacy measure is high (top row in Figure 7. Refer to Appendix II for additional analysis using an 8-neighbourhood variant of intricacy), and underweighs complexity in some of the high density and (partially) asymmetric patterns (bottom row in Figure 7).

Figure 7

Patterns with largest (z-scored) underestimation of complexity by the model (row 1, increasing in magnitude from right to left) and largest (z-scored) overestimation (row 2, increasing from left to right)



Discrepancy = 0.59

Discrepancy = 0.62

Discrepancy = 0.63

Discrepancy = 0.65

Discrepancy = 0.66

Discrepancy = -0.67

Discrepancy = -0.66

Discrepancy = -0.65

Discrepancy = -0.58

Discrepancy = -0.57

Note. Discrepancy values denote the average z-scored difference between objective complexity measure and subjective ratings for the pattern across all subjects.

3.2 Relationship between Beauty and Complexity Ratings

Table 2 summarises our main models and their performance (averaged over 3 cross validation folds from data split 1). Our dependent variable, beauty ratings, are abbreviated as BR.

Table 2

Summary of models of beauty ratings

S. no.	Model	Significance	AIC	BIC	R^2	RMSE	
						Train	Test
		(a) With CR as a predictor					
1	BR ~ CR + 1|Participant	CR*	8046.3	8070.5	0.25	0.84	0.86
2	BR ~ CR + disorder + 1|Participant	CR* disorder*	6449.7	6479.9	0.55	0.64	0.66
3	BR ~ CR + disorder + disorder|Participant	CR* disorder*	5985.0	6027.3	0.62	0.57	0.61
4	**BR ~ CR + disorder + CR:disorder + disorder|Participant**	**CR*** **disorder*** **CR:disorder***	**5918.2**	**5966.6**	**0.64**	**0.57**	**0.60**

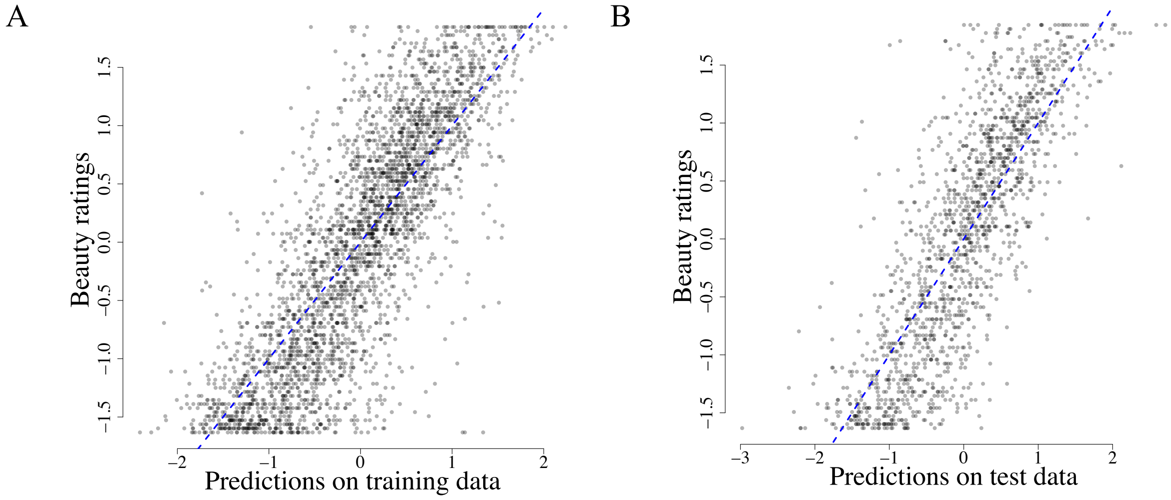
5	BR ~ CR^2 + disorder + 1|Participant	disorder*	6871.3	6901.6	0.49	0.69	0.71
6	BR ~ CR + CR^2 + disorder + 1|Participant	CR* disorder*	6458.1	6494.3	0.55	0.64	0.66
(b) Without CR as a predictor							
7	BR ~ LSC + intricacy + disorder + 1|Participant	LSC* intricacy* disorder*	6686.9	6723.2	0.52	0.67	0.69
8	BR ~ LSC + intricacy + disorder + disorder|Participant	LSC* intricacy* disorder*	6186.3	6234.6	0.60	0.59	0.62
9	BR ~ LSC + intricacy + disorder + (LSC + intricacy):disorder + disorder|Participant	LSC* disorder* LSC:disorder* intricacy:disorder*	6088.4	6148.8	0.62	0.58	0.61

Note. BR=beauty ratings, CR=complexity ratings, LSC=local spatial complexity; * indicates $p < 0.05$. Bold indicates best model.

We found that the beauty ratings are well predicted by complexity ratings, disorder and their interaction, along with a random intercept of participant and a random slope of disorder (Table 2, row 4, $R^2 = 0.64$). Figure 8 shows the plot of predictions versus ground truth on a train and test set from cross-validation fold 1.

Figure 8

Model performance on (A) training and (B) test data from fold 1 for beauty ratings.

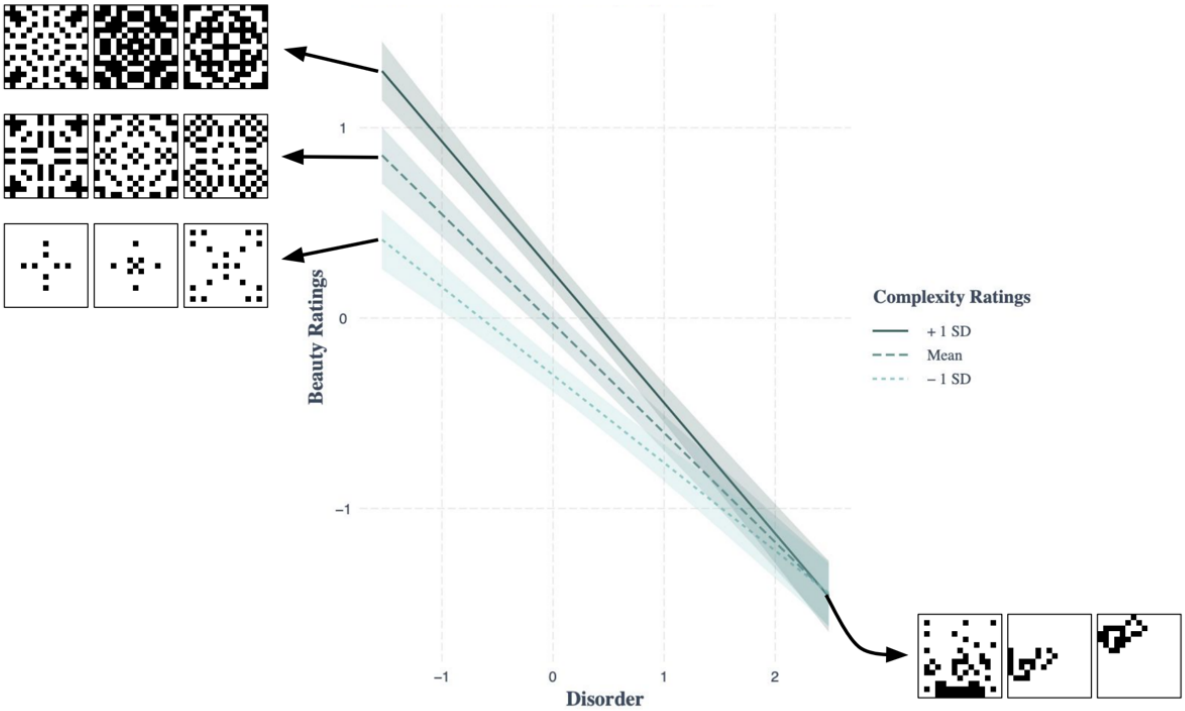


Note. Blue dashed line represents y = x.

Beauty ratings correlated positively with complexity ratings but negatively with disorder (mean asymmetry and entropy). Studying the interaction effect (Figure 9) suggested that high complexity is considered beautiful as long as the amount of disorder is low; and when the disorder is high, beauty is low irrespective of complexity. In other words, people prefer complexity (or find it more beautiful) while ensuring order. Further, the random slope of disorder implies that different people may have a different degree of perceived dislike towards disorder. Adding a quadratic effect of complexity to the model however, did not lead to a significant performance enhancement (Table 2, rows 5-6).

Figure 9

Visualization of the interaction effect between complexity ratings and disorder



Note. Plot shows beauty ratings *vs* disorder for different bins of complexity ratings. Example patterns for each case are shown in the left and right panels

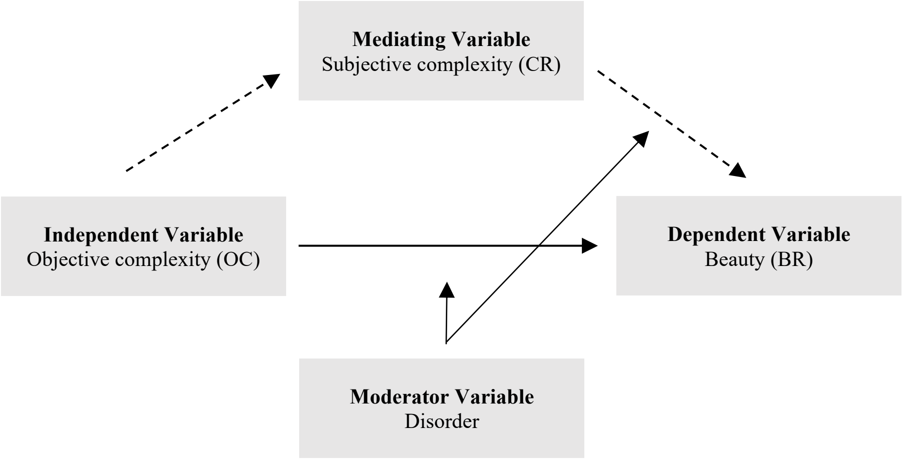
In sum, across both analyses, we observe a dissociation between the factors influencing perceived complexity and beauty. While subjective complexity can be explained by an integration of objective measures that encode local and global image features, beauty is explained by a contrast between complexity and order.

3.3 Relationship between Objective Complexity, Subjective Complexity and Beauty – Moderated Mediation Analysis

We also attempted to find an objective model of beauty, without the explicit use of the complexity ratings. Our previous analysis implies that beauty is well explained by a combination of purely objective measures, namely – LSC, intricacy (which together predict subjective complexity), and disorder. This model explains 62% of the variance in ratings (From Table 2, row 8). We examine the relationship between objective complexity, disorder, subjective complexity and beauty more formally using moderated mediation analysis.

Figure 10 shows the underlying moderated mediation structure of the variables of interest from our data. Objective complexity (OC) was evaluated as the weighted combination of LSC and intricacy, where the weights were obtained from the regression coefficients (including random effects) of the best performing model (Table 1, row 12) fit on complexity ratings from cross validation fold 1. We performed moderated mediation analysis using PROCESS macro model 15 in R. Due to limited data and to avoid overfitting, we only used the average model description (*i.e.*, excluding random effects) for our analysis. We fit the PROCESS model on data from cross-validation data fold 2.

Figure 10
Mediation structure



In line with our finding from Figure 9, there was a significant conditional direct effect of objective complexity on beauty at low disorder, but a non-significant effect at high disorder. There was a significant conditional indirect effect of objective complexity on beauty. However, the index of moderated mediation was found to be non-significant. This implies that the interaction effect with disorder did not influence the mediation between objective complexity, subjective complexity and beauty. We studied these direct and indirect effects more closely using regressions now including random effects while ignoring the interaction effect with disorder. The OC computation here excludes the random effects since the random effects structure in the model includes subject specific slope and intercepts. We use the cross-validation data fold 1 to fit these models. Table 3 shows the results from the regressions.

Table 3
Regressions to study mediation

S. no.	Model	Significance	OC slope coefficient
1	BR ~ OC + disorder + disorder|Participant	OC* disorder*	0.31
2	CR ~ OC + intricacy|Participant	OC*	0.98
3	BR ~ CR + OC + disorder + disorder|Participant	CR* disorder*	0.03

Note. BR=beauty ratings, OC=objective complexity, CR=complexity ratings

Table 4

Results of mediation analysis

	Estimate	95% CI	Significance
ACME	0.27	[0.22, 0.32]	Significant
ADE	0.03	[-0.03, 0.11]	Not significant

Note. ACME=average causal mediation effect, ADE=average direct effect

From Table 3, we find that OC is no longer a significant predictor in the presence of CR, and the slope coefficient of OC is largely reduced (Table 3, row 3, compared to row 1). To check if this mediation effect is significant, we use the mediate() function in R under the mediation library. Table 4 summarises the results for Average Causal Mediation Effect (ACME) and the Average Direct Effect (ADE). We find a significant average causal mediation effect. This means there is a significant indirect effect of objective complexity on beauty that goes through the mediator subjective complexity. Further, there is a non-significant direct effect of objective complexity on beauty. Together, we can conclude that the effect of objective complexity on beauty is mediated by subjective complexity. This implies that subjective complexity can supply useful information towards the prediction of beauty over and above what can be explained by objective complexity.

4. Discussion

A large body of work has attempted to assess subjective complexity and study its relationship with beauty, but has been subjected to a fair share of contradictions and an overall lack of consensus. The incomplete agreement about the relationship between beauty and complexity in the literature is difficult to resolve due to the predominant use of hand-crafted stimuli and measures which fail to generalize, and are hard to systematize. To address these challenges, we stepped back from the difficulties of using natural scenes and non-programmatic measures to create a foundation for future investigations based on a very simple class of patterns which admits algorithmically transparent measures – we used cellular automata, which provided us a systematic method of algorithmically generating diverse families of patterns.

Metric for Complexity

To relate participant ratings of complexity on these patterns to objective measures, we programmed six computational complexity measures including density, entropy, local spatial complexity, Kolmogorov complexity, and local and global asymmetry. These measures have been considered frequently in past studies (Friedenberg and Liby, 2016; Fan et al., 2022; Nadal, 2007; Attneave, 1957; Gartus and Leder, 2017; Damiano et al., 2021; Snodgrass, 1971; Arnheim, 1956, 1966; Bense, 1960, 1969; Moles, 1958; Schmidhuber, 2009; Javid, 2016; Rigau, 2008; Chikhman et al., 2012; Singh and Shukla, 2017; Silva, 2021). We also introduced a novel intricacy measure which quantified the number of visual elements in a pattern using a graph-based approach. While many previous works have discussed the role of the number of components (Berlyne et al., 1968; Hall, 1969; Roberts, 2007) as important factors governing complexity judgement, they were quantified by hand and to the best of our knowledge, we are the first to implement a computational measure, intricacy, quantifying this factor.

Using linear mixed effects regression, we found that a positive weighted combination of spatial complexity and intricacy (including a random slope of intricacy and a random intercept of participant) was an effective predictor ($R^2 = 0.46$) of subjective complexity ratings. This result is consistent with existing results suggesting the number of elements and their spatial arrangement are good predictors of subjective complexity (Berlyne, 1960; Berlyne et al., 1968; Nadal, 2007;). Moreover, the result is also in line with the literature suggesting that two aspects of processing may be involved in complexity perception – a quantity-based component focussing on the number of visual features and a structure-based component focussing on the distribution and organization of visual features (Chipman, 1977; Ichikawa, 1985). Moreover, local spatial complexity (LSC) is a local property averaged over the entire pattern, whereas intricacy is computed at a global level using a graphical representation of the entire pattern. Therefore, these measures add complementary information which are suitably integrated to give rise to human complexity evaluations.

Contrary to prior work (Arnoult, 1960; Attneave, 1957; Day, 1967; Eisenman and Gellens, 1968; Marin and Leder, 2013; Redies and Brachmann, 2017), neither symmetry nor entropy related to subjective complexity. This could be explained by the simplified black-and-white pixel nature of our stimuli as opposed to natural scenes, or the significant correlations between these measures and intricacy. Further, we observed a disparity between participant strategy responses and our metric – participants indicated that their ratings depended on their ability to create or replicate the pattern (which could be seen as a direct link to algorithmic complexity), while our approximate Kolmogorov complexity (KC) was not predictive of subjective complexity as per our model. This could again be due to the large underlying correlations of KC, asymmetry and entropy with LSC or intricacy, in turn masking their effect.

Relationship between Beauty and Complexity Ratings

In contrast, however, asymmetry and entropy did relate to beauty judgements. Beauty ratings correlated positively with subjective complexity and negatively with asymmetry and entropy. As a result, this work has lent support for a monotonic relationship between beauty and complexity. This goes against the inverted-U like dependence proposed by Berlyne. However, one reason for this, as stated above and by others (Krupinski and Locher, 1988; Nicki, Lee, and Moss, 1981; Stamps III, 2002), could be that our stimuli are so simple in nature as to lie in the lower quantiles of complexity. If true, then we would only expect to be able to reproduce the first half of the inverted-U curve, and as a result could not falsify a linear relationship.

A weighted combination of complexity ratings, disorder (itself a weighted combination of asymmetry and entropy) and their interaction (along with random slopes for disorder and random intercepts for participants) effectively modelled beauty ratings. Since symmetry and entropy quantify order, this concurs with the work proposing that beauty lies along a balance between order and complexity, albeit with a complementary relationship (as opposed to their being partial opposites, see also Van Geert and Wagemans, 2020). For example, Arnheim (1966) stated: “Complexity without order produces confusion. Order without complexity causes boredom”. However, based on our linear model (with interactions), we cannot lend support for Birkhoff’s (1933) proposed M = O / C relationship, or Eysenck’s (1941, 1942) M = O × C relationship. The interaction effect we found emphasizes the relative influence of order and complexity on beauty – complexity is beautiful, as long as the degree of disorder is low, and at high disorder, beauty is consistently low. This result is at odds with Van Geert and Wageman’s (2021) suggestion that a balance between order and complexity involves no interaction. They, however used real world images of neatly organized compositions and recorded fascination and soothingness judgements in a 2-choice task. The contradiction between their and our findings underlines the impeding challenge of comparing results across varying stimuli types and task designs and reiterates that the added value of the interaction term can depend on the stimuli and the specific operationalizations of order and complexity used (Van Geert and Wagemans, 2020).

Relationship between Objective Complexity, Subjective Complexity and Beauty

While a combination of pure objective measures of LSC, intricacy, and disorder (entropy and asymmetry) was able to explain 62% of the variance in beauty ratings, formal analysis of the relationship between subjective complexity, objective complexity and beauty using moderated mediation analysis revealed that subjective complexity mediates the influence of objective complexity on beauty at all levels of disorder. This indicates that subjective complexity encodes information beyond what is expressed in terms of objective complexity measures. This is why some views criticize such methods attempting to quantify subjective complexity in objective terms. Heckhausen (1964) argued that relating subjective complexity to simple visual properties of stimuli as done by information theory approaches is insufficient. He claimed that the subjective complexity does not solely depend on the complexity of the stimulus but also on the way it is perceived. Attneave (1957) also suggested that people’s perception of complexity is not a mere reflection of the visual stimuli. This explanation aligns with Gestalt philosophies of “perceptual organisation” often summarised as “whole is greater than sum of the parts”. This need was also highlighted by Berlyne who had claimed that complexity was a property of both the physical stimulus properties and the processes within the subject. Last but not least, one must clarify here that no causal implications can be made from this moderated mediation analysis and the relationship between subjective and objective quantities observed here are purely correlational.

Limitations and Future Directions

Our methods have limitations. Since the beauty rating was recorded along with the complexity rating, there might have been an anchoring effect which could have yielded a spurious correlation. Further, while we expect our complexity metric to generalize across basic transformations of the patterns, for example increased size, added colours, rotations or occlusions (Van Lier et al., 1994), explicit tests for this have not been addressed in this article and would be a part of future work. However, one might argue that our patterns, even if they generalize in the above mentioned ways, do not represent real-world stimuli for several reasons: (1) their statistics are very different from those of natural scenes, (2) they are overly simplistic, allowing for only limited colours in a restricted grid size and being devoid of overt semantic content, and (3) the nature of these pixel patterns makes it hard to define several popular measures of complexity such as number of vertices, edges, lines, or curves. Such a choice of stimuli therefore renders some of the prevalent contradictions irrelevant because the measures they employ do not apply to our stimuli. This raises the concern as to whether the complexity measure we arrived at remains correct in more ecologically valid settings. While the definitions of LSC and intricacy can formally be extended to larger stimuli with more colours easily, it will be necessary to study explicitly how predictive they are of subjective complexity for richer stimuli. Modern methods such as diffusion models for producing photorealistic images could be used as programmatic generators of images of potentially varying subjective complexity. Equally, convolutional neural networks could be used as feature extractors in place of our manually defined complexity measures (for example, Iigaya et al., 2020). Using such methods may, however, come at the expense of losing interpretability. We found consistent ratings of complexity and beauty within-participant, and a large amount of variation between-participants which was explained by random effects – we saw a large performance gain from adding a random slope of intricacy in our complexity models and disorder in our beauty models per participant. A more thorough analysis of individual differences is another important target for future work.

Conclusion

Our work showcases the usefulness of computational models to understand the link between assessments of complexity and beauty. We hope that it can motivate other researchers to look at aesthetic evaluations using a computational lens. Ultimately, we believe that computational models can increase both the reproducibility and generalisability of the field of empirical aesthetics more generally.

References

Adamatzky, A., & Martínez, G. J. (Eds.). (2016). *Designing beauty: the art of cellular automata* (Vol. 20). Springer.

Adkins, O. C., & Norman, J. F. (2016). The visual aesthetics of snowflakes. *Perception*, *45*(11), 1304-1319.

Aitken, P. P. (1974). Judgments of pleasingness and interestingness as functions of visual complexity. *Journal of Experimental Psychology*, *103*(2), 240.

Arnheim, R. (1956). *Art and visual perception: A psychology of the creative eye*. Univ of California Press.

Arnheim, R. (1966). Towards a psychology of art/entropy and art an essay on disorder and order. *The Regents of the University of California*, *160*.

Arnoult, M. D. (1960). Prediction of perceptual responses from structural characteristics of the stimulus. *Perceptual and Motor Skills*, *11*(3), 261-268.

Attneave, F. (1957). Physical determinants of the judged complexity of shapes. *Journal of experimental Psychology*, *53*(4), 221.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bense, M. (1960). Programmierung des Schönen Allgemeine Texttheorie Und Textästhetik.

Bense, M. (1969). *Kleine abstrakte ästhetik*. Edition Rot.

Berlyne, D. E. (1960). Conflict, arousal, and curiosity.

Berlyne, D. E. (1967). Arousal and reinforcement. In *Nebraska symposium on motivation*. University of Nebraska Press.

Berlyne, D. E., Ogilvie, J. C., & Parham, L. C. (1968). The dimensionality of visual complexity, interestingness, and pleasingness. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *22*(5), 376.

Berlyne, D. E. (1973). Aesthetics and psychobiology. *Journal of Aesthetics and Art Criticism*, *31*(4).

Birkhoff, G. D. (1933). Aesthetic Measure.

Birkin, G. (2010). *Aesthetic complexity: practice and perception in art & design*. Nottingham Trent University (United Kingdom).

Brielmann, A. A., & Dayan, P. (2022). A computational model of aesthetic value. *Psychological Review*.

Chamberlain, R. (2022). The interplay of objective and subjective factors in empirical aesthetics. In *Human Perception of Visual Information* (pp. 115-132). Springer, Cham.

Chikhman, V., Bondarko, V., Danilova, M., Goluzina, A., & Shelepin, Y. (2012). Complexity of images: Experimental and computational estimates compared. *Perception*, *41*(6), 631-647.

Chipman, S. F. (1977). Complexity and structure in visual patterns. *Journal of Experimental Psychology: General*, *106*(3), 269.

Chipman, S. F., & Mendelson, M. J. (1979). Influence of six types of visual structure on complexity judgments in children and adults. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(2), 365.

Chmiel, A., & Schubert, E. (2017). Back to the inverted-U for music preference: A review of the literature. *Psychology of Music*, *45*(6), 886-909.

Corchs, S. E., Ciocca, G., Bricolo, E., & Gasparini, F. (2016). Predicting complexity perception of real world images. *PloS one*, *11*(6), e0157986.

Cupchik, G. C. (1986). A decade after Berlyne: New directions in experimental aesthetics. *Poetics*, *15*(4-6), 345-369.

Damiano, C., Wilder, J., Zhou, E. Y., Walther, D. B., & Wagemans, J. (2021). The role of local and global symmetry in pleasure, interest, and complexity judgments of natural scenes. *Psychology of Aesthetics, Creativity, and the Arts*.

Davis, R. C. (1936). An evaluation and test of Birkhoff's aesthetic measure formula. *The Journal of General Psychology*, *15*(2), 231-240.

Day, H. Y. (1967). Evaluations of subjective complexity, pleasingness and interestingness for a series of random polygons varying in complexity. *Perception & Psychophysics*, *2*(7), 281-286.

Day, H. (1968). The importance of symmetry and complexity in the evaluation of complexity, interest and pleasingness. *Psychonomic Science*, *10*(10), 339-340.

Delplanque, J., De Loof, E., Janssens, C., & Verguts, T. (2019). The sound of beauty: How complexity determines aesthetic preference. *Acta Psychologica*, *192*, 146-152.

Donderi, D. C. (2006). An information theory analysis of visual complexity and dissimilarity. *Perception*, *35*(6), 823-835.

Eisenman, R. (1967). Complexity-simplicity: I. Preference for symmetry and rejection of complexity. *Psychonomic Science*, *8*(4), 169-170.

Eisenman, R., & Gellens, H. K. (1968). Preferences for complexity-simplicity and symmetry-asymmetry. *Perceptual and Motor Skills*, *26*(3), 888-890.

Eysenck, H. J. (1941). The empirical determination of an aesthetic formula. *Psychological Review*, *48*(1), 83.

Eysenck, H. J. (1942). The experimental study of the'good Gestalt'—a new approach. *Psychological Review*, *49*(4), 344.

Eysenck, H. J. (1968). An experimental study of aesthetic preference for polygonal figures. *The Journal of General Psychology*, *79*(1), 3-17.

Fan, Z. B., Li, Y. N., Zhang, K., Yu, J., & Huang, M. L. (2022). Measuring and evaluating the visual complexity of Chinese ink paintings. *The Computer Journal*, *65*(8), 1964-1976.

Farley, F. H., & Weinstock, C. A. (1980). Experimental aesthetics: Children's complexity preference in original art and photoreproductions. *Bulletin of the Psychonomic Society*, *15*(3), 194-196.

Forsythe, A., Mulhern, G., & Sawey, M. (2008). Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior research methods*, *40*(1), 116-129.

Friedenberg, J., & Liby, B. (2016). Perceived beauty of random texture patterns: A preference for complexity. *Acta psychologica*, *168*, 41-49.

Gartus, A., & Leder, H. (2013). The small step toward asymmetry: Aesthetic judgment of broken symmetries. *i-Perception*, *4*(5), 361-364.
Gartus, A., & Leder, H. (2017). Predicting perceived visual complexity of abstract patterns using computational measures: The influence of mirror symmetry on complexity perception. *PloS one*, *12*(11), e0185276.

Giannouli, V. (2013). Visual symmetry perception. *Encephalos*, *50*, 31-42.

Gordon, J., & Gridley, M. C. (2013). Musical preferences as a function of stimulus complexity of piano jazz. *Creativity Research Journal*, *25*(1), 143-146.

Güçlütürk, Y., Jacobs, R. H., & Lier, R. V. (2016). Liking versus complexity: Decomposing the inverted U-curve. *Frontiers in Human Neuroscience*, *10*, 112.

Hall, A. C. (1969). Measures of the complexity of random black and white and coloured stimuli. *Perceptual and motor skills*, *29*(3), 773-774.

Harsh, C. M., Beebe-Center, J. G., & Beebe-Center, R. (1939). Further evidence regarding preferential judgment of polygonal forms. *The Journal of Psychology*, *7*(2), 343-350.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Heath, T., Smith, S. G., & Lim, B. (2000). Tall buildings and the urban skyline: The effect of visual complexity on preferences. *Environment and behavior*, *32*(4), 541-556.

Heckhausen, H. (1964). Complexity in perception: Phenomenal criteria and information theoretic calculus—A note on D. Berlynes" Complexity Effects.".

Hekkert, P., & Van Wieringen, P. C. (1990). Complexity and prototypicality as determinants of the appraisal of cubist paintings. *British journal of psychology*, *81*(4), 483-495.

Ichikawa, S. (1985). Quantitative and structural factors in the judgment of pattern complexity. *Perception & psychophysics*, *38*(2), 101-109.

Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K., & O'Doherty, J. P. (2020). Aesthetic preference for art emerges from a weighted integration over hierarchically structured visual features in the brain. *BioRxiv*.

Jacobsen, T. (2010). Beauty and the brain: culture, history and individual differences in aesthetic appreciation. *Journal of anatomy*, *216*(2), 184-191.

Jacobsen, T., & Höfel, L. E. A. (2002). Aesthetic judgments of novel graphic patterns: Analyses of individual judgments. *Perceptual and motor skills*, *95*(3), 755-766.

Jacobsen, T., Schubotz, R. I., Höfel, L., & Cramon, D. Y. V. (2006). Brain correlates of aesthetic judgment of beauty. *Neuroimage*, *29*(1), 276-285.

Javaheri Javid, M. A., Blackwell, T., Zimmer, R., & Al-Rifaie, M. M. (2016, March). Correlation between human aesthetic judgement and spatial complexity measure. In *International Conference on Computational Intelligence in Music, Sound, Art and Design* (pp. 79-91). Springer, Cham.

Javaheri Javid, M. A. (2019). *Aesthetic Automata: Synthesis and Simulation of Aesthetic Behaviour in Cellular Automata* (Doctoral dissertation, Goldsmiths, University of London).

Javaheri Javid, M. A. (2021, April). Aesthetic Evaluation of Cellular Automata Configurations Using Spatial Complexity and Kolmogorov Complexity. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)* (pp. 147-160). Springer, Cham.

Krupinski, E., & Locher, P. (1988). Skin conductance and aesthetic evaluative responses to nonrepresentational works of art varying in symmetry. *Bulletin of the Psychonomic Society*, *26*(4), 355-358.

Lakhal, S., Darmon, A., Bouchaud, J. P., & Benzaquen, M. (2020). Beauty and structural complexity. *Physical Review Research*, *2*(2), 022058.

Langton, C. G. (1986). Studying artificial life with cellular automata. *Physica D: Nonlinear Phenomena*, *22*(1-3), 120-149.

Li, W., Packard, N. H., & Langton, C. G. (1990). Transition phenomena in cellular automata rule space. *Physica D: Nonlinear Phenomena*, *45*(1-3), 77-94.

Machotka, P. (1980). Daniel Berlyne's contributions to empirical aesthetics. *Motivation and Emotion*, *4*(2), 113-121.

Madison, G., & Schiölde, G. (2017). Repeated listening increases the liking for music regardless of its complexity: Implications for the appreciation and aesthetics of music. *Frontiers in neuroscience*, *11*, 147.

Marin, M. M., & Leder, H. (2013). Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music. *PloS one*, *8*(8), e72412.

Marin, M. M., Lampatz, A., Wandl, M., & Leder, H. (2016). Berlyne revisited: Evidence for the multifaceted nature of hedonic tone in the appreciation of paintings and music. *Frontiers in human neuroscience*, *10*, 536.

McWhinnie, H. J. (1971). A review of selected aspects of empirical aesthetics III. *Journal of Aesthetic Education*, *5*(4), 115-126.

Messinger, S. M. (1998). Pleasure and complexity: Berlyne revisited. *The Journal of Psychology*, *132*(5), 558-560.

Moles, A. A. (1958). Theorie de linformation et perception esthetique.

Munsinger, H., & Kessen, W. (1964). Uncertainty, structure, and preference. *Psychological Monographs: General and Applied*, *78*(9), 1.

Nadal, M., & Vartanian, O. (2021). Empirical Aesthetics: An overview.

Nadal, M., Munar, E., Marty, G., & Cela-Conde, C. J. (2010). Visual complexity and beauty appreciation: Explaining the divergence of results. *Empirical Studies of the Arts*, *28*(2), 173-191.

Nasar, J. L. (2002). What design for a presidential library? Complexity, typicality, order, and historical significance. *Empirical Studies of the Arts*, *20*(1), 83-99.

Nicki, R. M. (1972). Arousal increment and degree of complexity as incentive. *British Journal of Psychology*, *63*(2), 165-171.

Nicki, R. M., & Moss, V. (1975). Preference for non-representational art as a function of various measures of complexity. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *29*(3), 237.

Nicki, R. M., & Gale, A. (1977). EEG, measures of complexity, and preference for nonrepresentational works of art. *Perception*, *6*(3), 281-286.

Nicki, R. M., Lee, P. L., & Moss, V. (1981). Ambiguity, cubist works of art, and preference. *Acta Psychologica*, *49*(1), 27-41.

Norman, J. F., Beers, A., & Phillips, F. (2010). Fechner's Aesthetics Revisited. *Seeing and Perceiving*, *23*(3), 263-271.

Osborne, J. W., & Farley, F. H. (1970). The relationship between aesthetic preference and visual complexity in absract art. *Psychonomic Science*, *19*(2), 69-70.

Redies, C., Brachmann, A., & Wagemans, J. (2017). High entropy of edge orientations characterizes visual artworks from diverse cultural backgrounds. *Vision Research*, *133*, 130-144.

Rigau, J., Feixas, M., & Sbert, M. (2008). Informational aesthetics measures. *IEEE computer graphics and applications*, *28*(2), 24-34.

Roberts, M. N. (2007). Complexity and aesthetic preference for diverse visual stimuli. *Doctoral), Universitat de les Illes Balears, Palma, Spain*.

Schmidhuber, J. (2009). Art & science as by-products of the search for novel patterns, or data compressible in unknown yet learnable ways. *Multiple ways to design research. Research cases that reshape the design discipline, Swiss Design Network-Et al. Edizioni*, 98-112.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Silva, J. M., Pratas, D., Antunes, R., Matos, S., & Pinho, A. J. (2021). Automatic analysis of artistic paintings using information-based measures. *Pattern Recognition*, *114*, 107864.

Singh, S., & Shukla, D. (2017). A review on various measures for finding image complexity. *International Journal of Scientific Research Engineering & Technology. ISSN*, 2278-0882.

Snodgrass, J. G. (1971). Objective and subjective complexity measures for a new population of patterns. *Perception & Psychophysics*, *10*(4), 217-224.

Stamps III, A. E. (2002). Entropy, visual diversity, and preference. *The Journal of general psychology*, *129*(3), 300-320.

Sun, Z., & Firestone, C. (2022). Beautiful on the inside: Aesthetic preferences and the skeletal complexity of shapes. *Perception*, *1*, 15.

Taylor, R. E., & Eisenman, R. (1964). Perception and production of complexity by creative art students. *The Journal of Psychology*, *57*(1), 239-242.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis.

Tinio, P. P., & Leder, H. (2009). Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization. *Acta psychologica*, *130*(3), 241-250.

Tran, N. H., Waring, T., Atmaca, S., & Beheim, B. A. (2021). Entropy trade-offs in artistic design: A case study of Tamil kolam. *Evolutionary Human Sciences*, *3*.

Van Geert, E., & Wagemans, J. (2020). Order, complexity, and aesthetic appreciation. *Psychology of aesthetics, creativity, and the arts*, *14*(2), 135.

Van Geert, E., & Wagemans, J. (2021). Order, complexity, and aesthetic preferences for neatly organized compositions. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(3), 484.

Van Lier, R., Van Der Helm, P., & Leeuwenberg, E. (1994). Integrating global and local aspects of visual occlusion. *Perception*, *23*(8), 883-903.

Vitz, P. C. (1966). Preference for different amounts of visual complexity. *Behavioral science*, *11*(2), 105-114.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *22*(3), 392-399.

William, A. Huber (2011). Measuring entropy/ information/ patterns of a 2d binary matrix.

Wolfram, S. (2002). A new kind of science, Wolfram Media, Inc. *Champaign, IL.*

Zenil, H., Hernández-Orozco, S., Kiani, N. A., Soler-Toscano, F., Rueda-Toicen, A., & Tegnér, J. (2018). A decomposition method for global evaluation of shannon entropy and local estimations of algorithmic complexity. *Entropy*, *20*(8), 605.

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory*, *24*(5), 530-536.

Appendix I – Cellular Automata Generation

The number of potential state-transition update functions for a 2D CA with n states and an N-cell neighbourhood is n^{n^N}. Hence, for our 2D binary CA with 5-cell and 9-cell neighbourhoods, the number of possible state-transition update functions are $2^{2^5} = 2^{32} = 4 \times 10^9$ and $2^{2^9} = 2^{512} = 10^{152}$ respectively. Moreover, the total number of initial configurations for an n-state P × Q grid is $n^{\mathrm{P} \times \mathrm{Q}}$. In our binary 15 × 15 grid, this would be $2^{15 \times 15} = 2^{225}$ possible initial configurations. Since these are very large spaces, we consider simplified versions of rules (Wolfram, 1983). The rules are set such that the state of a cell depends only on the sum of the states of cells in its neighbourhood (SCN). Such rules can be of two types: “totalistic” (tot) where the state of the cell (i, j) at time t+1 depends only on SCN at time t, or “outer-totalistic” (Otot) where the state of the cell (i, j) at time t+1 depends on both SCN and the value of cell (i, j) at time t. These rules can be expressed as decimal “rule codes” given by $\sum_{i=0}^{N}(f(i) \times 2^{i})$ for totalistic rules and $\sum_{i=0}^{N}\sum_{j=0}^{1}(f(i,j) \times 2^{2i+j})$ for outer-totalistic rules.

It is a well-known fact that unlike 1D/elementary CA, the limiting behaviour of 2D CA is undecidable. Patterns might terminate, oscillate, tend towards full randomness or tend towards order depending on the combination of algorithm parameters. It is also not possible to reverse-engineer rules that result in particular types of pattern outputs. Therefore, we referred to work that has elucidated some rules with their evolving behaviours (Packard and Wolfram, 1985; Wolfram, 2002) while selecting our set of rules and fixing the number of iterations based on our grid size to T = 40.

The generation script was coded in Python v3.8.8. For each cell, either the 5-cell or a 9-cell neighbourhood was determined and the sum of the SCN was computed. The grid was wrapped around such that the cells on the boundary considered cells on the opposite boundary as neighbours. The updated state of the cell was obtained as a function of the computed sum as indicated by the rule code read in binary. The pseudocode is given here.

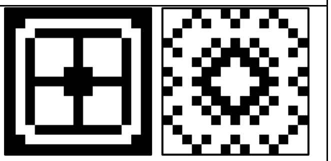
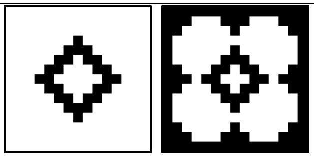
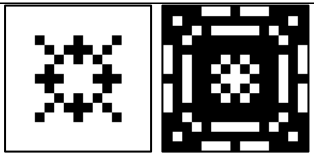
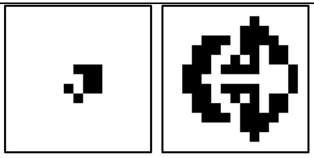
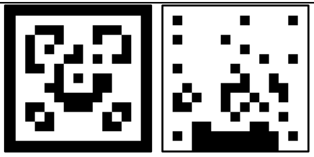
ALGORITHM: Cellular Automata Pattern Generation

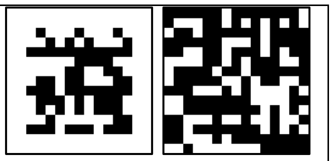
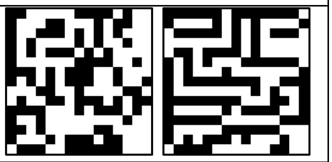
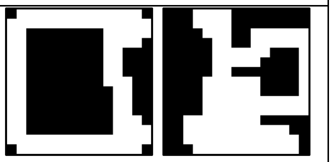
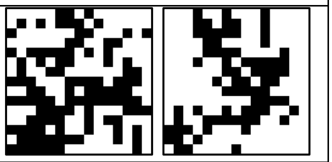
```
Input:
    Rule code: integer,
    N: integer,
    Grid with IC, G: 2D array,
    Rule type (tot/Otot): string,
    T: integer
Output: 8 patterns (one every 5th iteration)
Bin_rule_code ← binary form of decimal rule code
Powers ← list of indices where bin_rule_code = 1 when read in reverse
For timestep t = 1 to T
    G’ ← Copy of G
    For each cell (i, j) in G’
        If N = 5
```

```
            Neigh ← list of neighbourhood cells = [(i,j), (i+1,j), (i,j+1), (i-1,j), (i,j-1)]
        Else if N = 9
            Neigh ← list of neighbourhood cells = [(i,j), (i+1,j), (i,j+1), (i-1,j), (i,j-1),(i-1,j-1),(i-1,j+1),(i+1,j-1),(i+1,j+1)]
        End
        Neigh_sum ← sum of the states of the cells in Neigh
        If rule type is “tot”
            New_state ← 1 if Neigh_sum in Powers, else 0
        Else if rule type is “Otot”
            New_state ← 1 if 2 ×Neigh_sum + G[i, j] in Powers, else 0
        End
        G’[i,j] = New_state
    End
    G ← Copy of G’
    If t % 5 = 0
        Save G
End
```

Some example rules used by us along with the produced patterns are listed in Table AI.1.

Table AI.1. Rules used for cellular automata

S. no.	Rule code	Neighbourhood size	Rule type (Tot/Otot)	IC	Pattern (iteration 5, 20)
1	451	5	Otot	1	
2	510	5	Otot	1	
3	15822	9	Otot	1	
4	736	9	Otot	2	
5	85507	5	Otot	2	

6	15822	9	Otot	2	
7	736	9	Otot	3	
8	196623	9	Otot	3	
9	52	5	Tot	3	

Appendix II – Objective Complexity Measures

In addition to the complexity measure mentioned in Section 2.2, we implemented two other computational complexity measures: an 8-neighbourhood version of our intricacy measure and a hierarchical quad-tree measure. The motivation for and computation of these measures are presented here.

1. **8-neighbourhood intricacy measure**: In Figure 7, we presented some of the patterns for which our complexity measure is most erroneous. Complexity is overestimated in some patterns with high evaluated intricacy – potentially because diagonal relationships do not contribute to connectedness. To tackle this, we implemented an 8-neighbourhood version of intricacy wherein the graphs considers all 8 neighbours. An edge is inserted between two neighbours if they are the same colour. This measure would in turn result in a lower value of intricacy compared to the 4-neighbourhood version. Figure AII.1 illustrates its computation of the 4-nieghbourhood *vs.* the 8-neighbourhood intricacy.

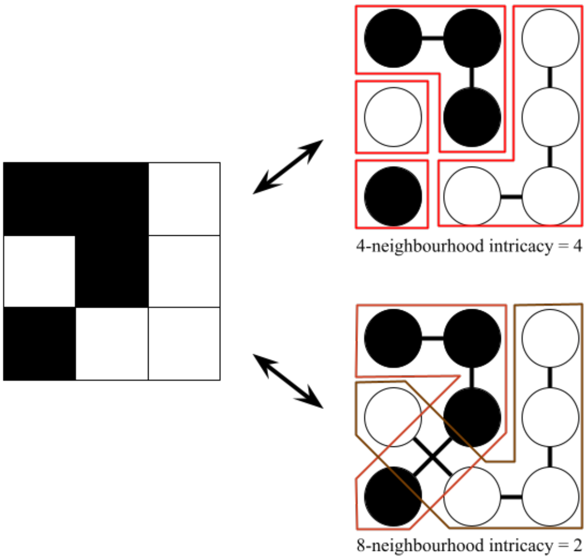


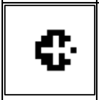
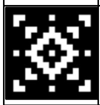
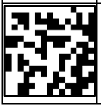
Figure AII.1 4-neighbourhood and 8-nieghbourhood intricacy computation for an example pattern. The graphs on right are constructed from the pattern on left. Boxes indicate connected components. Here, 4-neighbourhood intricacy = 4 and 8-nieghbourhood intricacy = 2.

2. **Quadtree:** Since most of our measures (except entropy) are either local or global, we programmed a hierarchical measure of complexity, namely quadtrees. Quadtrees are hierarchical data structures and are commonly used for representing images (Finkel and Bentley, 1974). A quadtree is based on the principle of recursive decomposition where the pattern is repeatedly divided into four sub-patterns. The condition for subdivision is based on the number of distinct states in the graph. If all cells are of the same colour, the algorithm stops, if not, the graph is subdivided into four sub-graphs and the same rule is recursively applied to each sub-graph. The number of times the graph undergoes

subdivisions is evaluated as a measure of complexity. This measure however is not ideal for our pattern with grid size 15 × 15 where the width and height are not powers of 4 and as a result the four sub-patterns are of unequal dimensions. Moreover, the measure was found to be highly correlated with LSC (r = 0.88, $p < 0.01$, CI = [0.87, 0.88]).

Appendix III, section 2 reports the performance of these measures in predicting subjective complexity and Table AII.1 shows some example patterns along with the corresponding computed measures.

AII.1 Example patterns along with computed measures

Pattern	Density	Entropy	LSC	KC	Asymmetry (horizontal, vertical, local)	Intricacy, 4-neigh	Intricacy, 8-neigh	Quadtree	N	Rule type	IC	λ
	0.10	1.97	0.43	134.2	58.3, 0, 0	3	3	22	9	Otot	2	4
	0.74	2.16	0.81	257.8	0, 0, 0	16	16	54	5	Otot	1	8
	0.45	2.76	0.96	278.7	53.3, 46.6, 0.004	30	8	65	9	Otot	3	6

Appendix III – Supplementary Analysis

1. Correlations between measures and ratings

From Figure AIII.1 we see that human complexity ratings are positively correlated with LSC (r = 0.48, p < 0.01, CI = [0.37, 0.42]), KC (r = 0.46, p < 0.01, CI = [0.46, 0.51]), quadtree (r = 0.53, p < 0.01, CI = [0.5, 0.55]) and intricacy (r = 0.40, p < 0.01, CI = [0.38, 0.43]) measures. The ratings were not highly correlated with asymmetry or entropy. On the other hand, the beauty ratings were highly negatively correlated with the three asymmetry measures (r = -0.37, p < 0.01, CI = [-0.39, -0.34]; r = -0.55, p < 0.01, CI = [-0.57, -0.52]; r = -0.56, p < 0.01, CI = [-0.58, -0.54]) and entropy (r = -0.51, p < 0.01, CI = [-0.53, -0.49]) and positively with complexity ratings (r = 0.4, p < 0.01, CI = [0.37, 0.42]).

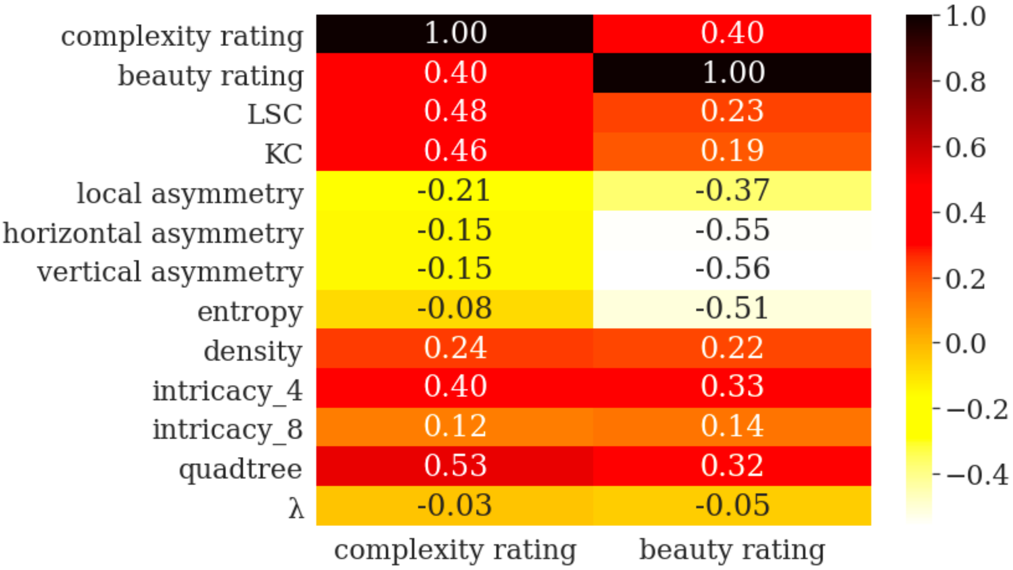


Figure AIII.1 Correlation between ratings and computational measures

2. Mixed Effects Regressions for predicting Complexity Ratings with 8-neighbourhood Intricacy and Quadtree

Table AIII.1 summarizes models of complexity ratings involving 8-neighbourhood intricacy and quadtree. Comparing with Table 1, we see that although quadtree achieves good performance as a predictor alone, there are negligible gains when we introduce intricacy into the expression. Further, the slight increase in BIC when doing so indicates that the model nearly overfits the data. Further, the 8-neighbourhood intricacy is unable to explain much variance in the ratings compared to the 4-neighbourhood version implying that the 4-neighbourhood intricacy is a more superior predictor of complexity ratings. Therefore, the model we propose in Section 3.1 comprising of LSC and (4-neighbourhood) intricacy along with random slope of intricacy and random intercept of participant is our best predictor for complexity ratings.

Table AIII.1: Summary of models of complexity ratings

S. no.	Model	Significance	AIC	BIC	R^2	RMSE	
						Train	Test
1	CR ~ quadtree + 1|Participant	quadtree*	7317.1	7341.3	0.41	0.74	0.76
2	CR ~ quadtree + quadtree|Participant	quadtree*	7199.9	7236.2	0.45	0.71	0.74
3	CR ~ intricacy_8 + 1|Participant	intricacy_8*	8493.1	8517.2	0.14	0.90	0.92
4	CR ~ quadtree + intricacy_4 + 1|Participant	quadtree* intricacy_4*	7311.8	7342.0	0.42	0.74	0.76
5	CR ~ quadtree + intricacy_4 + quadtree|Participant	quadtree* intricacy_4*	7194.4	7236.7	0.45	0.71	0.74
6	CR ~ quadtree + intricacy_4 + intricacy_4|Participant	quadtree* intricacy_4*	7191.2	7233.5	0.45	0.71	0.74

3. Test for trends, autocorrelation, and consistency of repeated measures in the ratings

To test for trends and autocorrelation (at lag 1) in the data, trial number and previous rating were respectively added as a predictor of complexity and beauty ratings. Table AIII.2 (a) and (b) report the performance of these models.

Table AIII.2 (a) Summary of models of complexity ratings. CR = complexity ratings, LSC = local spatial complexity, prevCR = previous complexity rating (from previous trial)

S. no.	Model	Significance	AIC	BIC	R^2	RMSE	
						Train	Test
1	CR ~ trial + 1|Participant		8550.4	8574.6	0.13	0.91	0.93
2	CR ~ LSC + intricacy_4 + trial + intricacy_4|Participant	LSC* intricacy_4*	7168.4	7216.7	0.46	0.70	0.73
3	CR ~ prevCR + 1|Participant	prevCR*	8520.4	8556.6	0.13	0.90	0.92
4	CR ~ LSC + intricacy_4 + prevCR + intricacy_4|Participant	LSC* intricacy_4* prevCR*	7137.3	7185.6	0.46	0.70	0.73

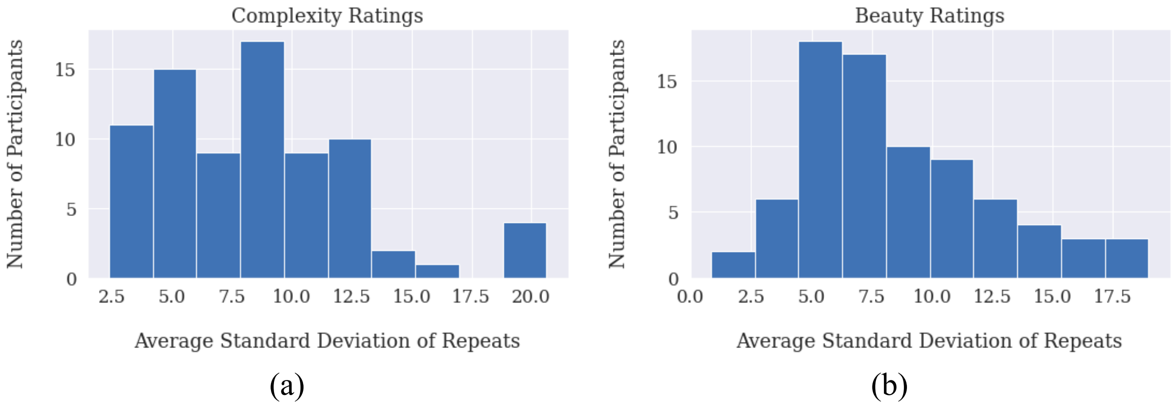
(b) Summary of models of beauty ratings. BR = beauty ratings, OC = objective complexity

S. no.	Model	Significance	AIC	BIC	R^2	RMSE	
						Train	Test

1	BR ~ trial + 1|Participant		8550.4	8574.5	0.13	0.91	0.93
2	BR ~ CR + disorder + CR:disorder + trial + disorder|Participant	CR* disorder* CR:disorder *	5948.3	6002.7	0.65	0.57	0.59
3	BR ~ prevBR + 1|Participant	prevBR*	8532.1	8568.4	0.13	0.91	0.93
4	BR ~ CR + disorder + CR:disorder + prevBR + disorder|Participant	CR* disorder* CR:disorder * prevBR*	5925.4	5979.8	0.65	0.57	0.59

For predicting complexity ratings, we find that trial number is not a significant predictor, indicating there are no significant trends in our data – as would be expected in case of boredom, or over-familiarity. We see that pervious complexity rating is a significant predictor, however, it does not enhance performance in conjunction with LSC and intricacy. Similarly for predicting beauty ratings, trial is not significant and previous beauty rating, though significant, does not enhance performance beyond our best model reported in Section 3.2.

Further, to test for consistency of repeated responses, we plot participant mean standard deviation in repeated measures (Figure AIII.2 (a) for complexity ratings and (b) for beauty ratings) to find that participants are largely consistent with their ratings.



(a) (b)

Figure AIII.2: Histogram of participant mean standard deviation in repeated measures for (a) complexity ratings and (b) beauty ratings.

4. Average complexity ratings vs average beauty ratings (colour coded by symmetry and entropy)

A plot of average beauty ratings per pattern across all participants versus average complexity ratings per pattern across all participants neatly underlines the role of asymmetry and entropy in beauty assessment (Figure AIII.3a,b). In Figure AIII.3a, the linear relationship between beauty and complexity is evident but we see two modes in the distribution (a Gaussian Mixture Model with 2 components was able to fit the average ratings well, Appendix III.5). The degree of symmetry is successfully able to explain this bimodality with fully symmetric patterns being rated higher on average than unidirectional symmetric (semi-symmetric) patterns, which are themselves rated higher on average than fully asymmetric (non symmetric) patterns. Further, Figure AIII.3b indicates that patterns with high entropy were rated as low beauty whereas patterns with low entropy were rated as high beauty unless the pattern was rated very low complexity.

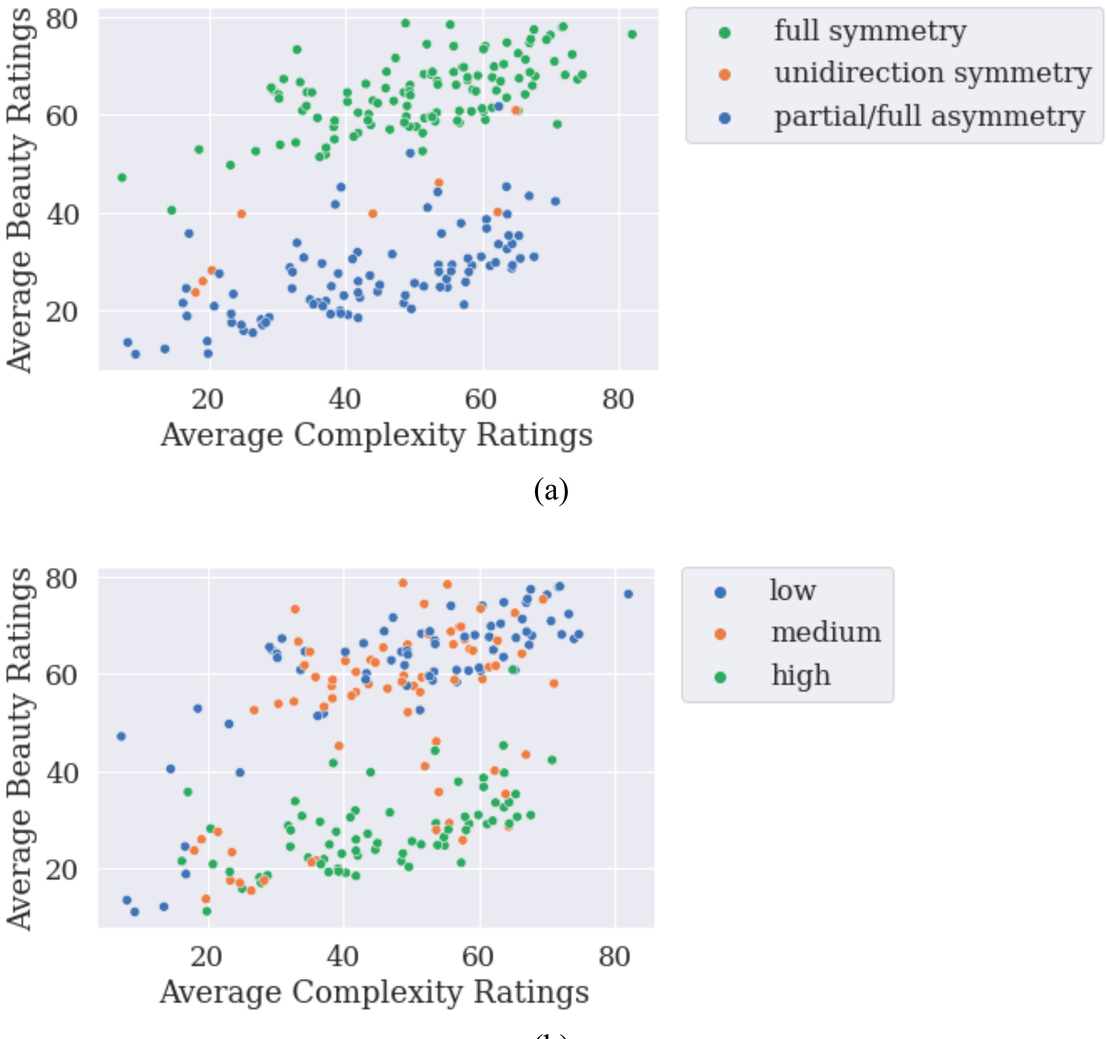


Figure AIII.3: Average complexity ratings vs average beauty ratings per pattern across all participants labelled according to (a) degree of symmetry in the pattern (as defined in Table 1), (b) level of entropy of the pattern (tertile split of entropy)

5. Gaussian Mixture Model to fit average complexity ratings versus average beauty ratings

As seen above, the plot of average complexity ratings versus average beauty ratings reflects a bimodal distribution. We therefore used the gmr package in Python to perform Gaussian Mixture Regression (GMR) with 2 modes on the average ratings. In GMR, first the joint distribution *p(x, y)* is learnt, where in this case *x* refers to average beauty ratings and *y* refers to average complexity ratings. Then, the conditional *p(y | x)* is computed to make predictions. GMR with 2 components is able to fit the data well (Figure AIII.4) with an RMSE of 12.0 on a held-out test set containing 40 data points.

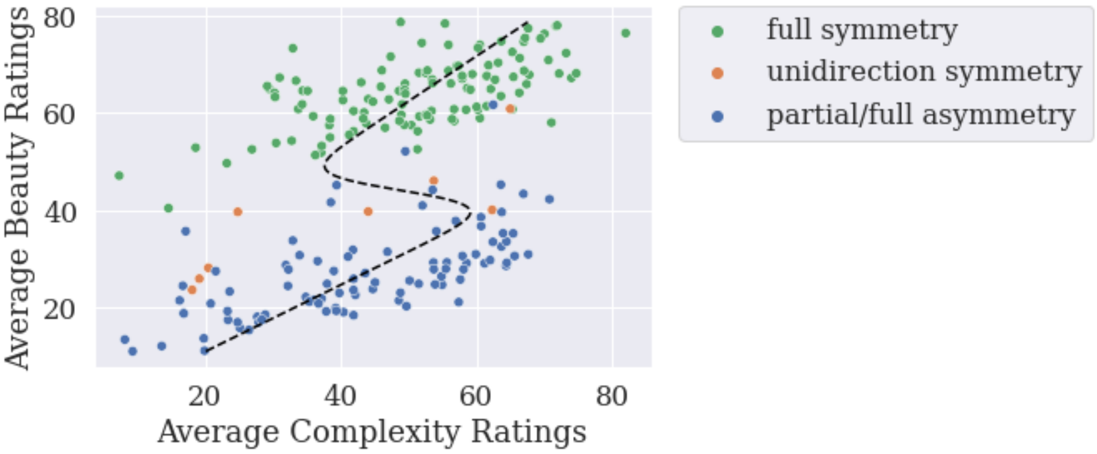
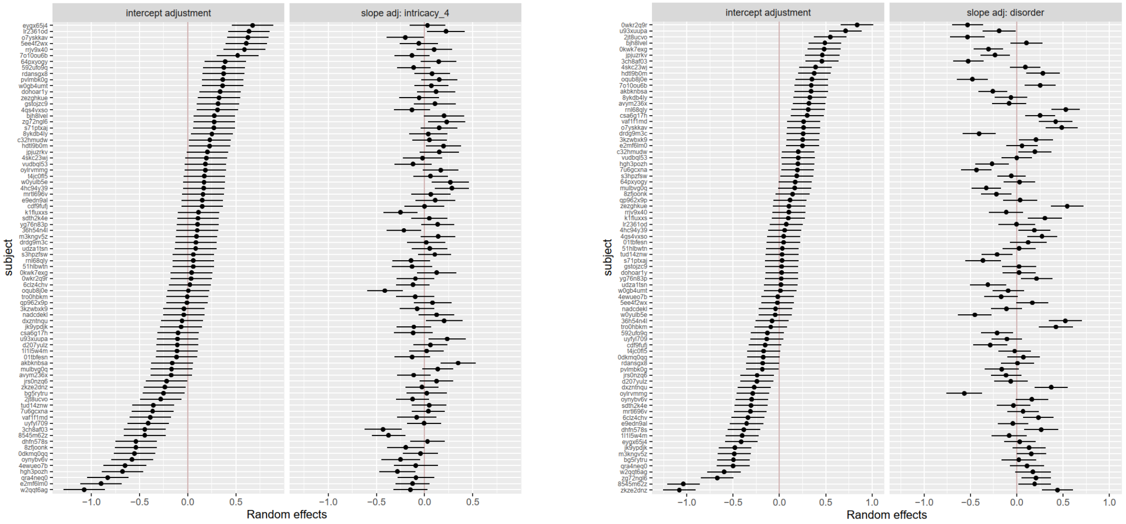


Figure AIII.4: GMM fit on average ratings. The dashed line represents the model fit by plotting the prediction of average complexity ratings against a continuous range of average beauty ratings.

6. <u>Check for Individual Differences</u>

Figure AIII.4 displays a plot of random effects for the best performing model of complexity ratings (Figure AIII.5a) and beauty ratings (Figure AIII.5b). The plots show that random intercepts for participant have a higher variance than random slopes for either intricacy or disorder. The variation in random intercepts of participants show that while some people rate complexity to be high on average, some others rate them to be low, and similarly for beauty. Moreover, the random slopes indicate that people perceive the impact of intricacy and disorder on complexity and beauty respectively to different degrees. The variation in the random slope of disorder is larger than the variation of the random slope of intricacy. However, this could be since disorder is an aggregate of asymmetry and entropy, where asymmetry is itself an aggregate of local, horizontal and vertical asymmetry, and hence is less stable across participants. More thorough analysis of individual differences is intended to be a part of our future work.



(a) Random intercept of participant and random slope of intricacy

(b) Random intercept of participant and random slope of disorder

Figure AIII.5: Plot of random effects from our best performing (a) complexity model and (b) beauty model

References

Finkel, R. A., & Bentley, J. L. (1974). Quad trees a data structure for retrieval on composite keys. *Acta informatica*, *4*(1), 1-9.

Packard, N. H., & Wolfram, S. (1985). Two-dimensional cellular automata. *Journal of Statistical physics*, *38*(5), 901-946.

Wolfram, S. (1983). Statistical mechanics of cellular automata. *Reviews of modern physics*, *55*(3), 601.