Beyond Letters: Optimal Transport as a Model for Sub-Letter Orthographic Processing

Jack E. Taylor[1,2], Rasmus Sinn[1], Cosimo Iaia[1], and Christian J. Fiebach[1,3]

[1]Department of Psychology, Goethe University Frankfurt

[2]School of Psychology and Neuroscience, University of Glasgow

[3]Brain Imaging Center, Goethe University Frankfurt

Author Note

Jack E. Taylor https://orcid.org/0000-0003-4765-0118

Rasmus Sinn https://orcid.org/0009-0008-6058-506X

Cosimo Iaia https://orcid.org/0000-0002-4317-4540

Christian J. Fiebach https://orcid.org/0000-0003-0827-1721

Correspondence concerning this article should be addressed to Jack E. Taylor, Department of Psychology, Goethe University Frankfurt; Taylor@psych.uni-frankfurt.de

The authors are grateful to Maren Wehrheim and Domenic Bersch for comments on this manuscript.

This research was supported by a Fokus A|B Grant provided by Goethe University Frankfurt.

Authors report no conflict of interest.

Abstract

Letter processing plays a key role in visual word recognition. However, word recognition models typically overlook or greatly simplify early perceptual processes of letter recognition. We suggest that optimal transport theory may provide a computational framework for describing letter shape processing. We use representational similarity analysis to show that optimal transport cost (Wasserstein distance) between pairs of letters aligns with neural activity elicited by visually presented letters <225 ms after stimulus onset, outperforming an existing approach based on shape overlap. We additionally show that optimal transport can capture the emergence of geometric invariances (e.g., to position or size) observed in letter perception. Finally, we demonstrate that Wasserstein distance predicts neural activity similarly well to features from artificial networks trained to classify images and letters. However, whereas representations in artificial neural networks emerge in a computationally unconstrained manner, our proposal provides a computationally explicit route to modeling the earliest orthographic processes.

Keywords: orthography, letters, characters, visual word recognition, graphetics, optimal transport, RSA, EEG

Beyond Letters: Optimal Transport as a Model for Sub-Letter Orthographic Processing

Reading is one of the most important cultural inventions humans have developed, for which the majority of humans rely on visual perception and recognition of written words. Visual word recognition processes allow visually presented words to be mapped to corresponding phonological, lexical, and semantic representations. To enable this, our brains must first extract and process orthographic information, identifying, recognising, and combining strokes, letters, and word forms in visual input. It is well known that such processes are guided by higher-level orthographic knowledge. For instance, letters are more easily recognised in the context of real words than they are in pseudowords or nonwords (Cattell, 1886; Lally & Rastle, 2023; Reicher, 1969; Wheeler, 1970), and, when words are presented only briefly or at small visual angles, observers incorrectly report the presence of letters which are expected but are in fact absent (Jordan et al., 1999; Jordan et al., 2015; Pillsbury, 1897). Nevertheless, it is clear that bottom-up processing of component letters plays a substantial role in the recognition of visual words (Brossette et al., 2022; Grainger, 2008; Lally & Rastle, 2023; Pelli et al., 2003; Rosa et al., 2016). Indeed, neuroimaging evidence suggests that sensitivities to abstract letter identities, letter combinations, and then word forms peak successively over time, as information is passed on from more posterior, hierarchically 'earlier', to more anterior brain regions, along the ventral visual stream (Agrawal & Dehaene, 2023; Gutiérrez-Sigut et al., 2019; Lochy et al., 2018; Petit et al., 2006; Thesen et al., 2012; Vinckier et al., 2007; Woolnough et al., 2021).

In alphabetic scripts, letters thus provide a vital bridge between vision and language, and letter-shape processing can be conceptualised as the primary stage of orthographic processing. Despite this, the processes that comprise letter recognition have been largely overlooked in cognitive models of word recognition processes (Finkbeiner & Coltheart, 2009). When investigated experimentally, orthographic

processes in word recognition are typically captured by orthographic measures whose descriptions begin at the level of letter identities (or, graphemes). For instance, the effect of the regularity of a word's spelling on lexical decision performance, wherein more regular words are more easily recognised as real words, is captured by measures of orthographic neighbourhood density (Coltheart et al., 1977; Yarkoni et al., 2008), which describe how similar a given word's string of letters is to those of other words in the same language or script. However, these measures limit themselves to treating letters as functionally irreducible categories, neglecting sub-graphemic (i.e., *graphetic;* Meletis & Dürscheid, 2022; Rezec, 2013) information. For instance, Coltheart's *N* (Coltheart et al., 1977) quantifies a given word's orthographic neighbourhood density as the number of other words of equal length that share all but one letter. This implies that orthographic similarity between any two words is binary; they are either neighbours at a distance of one letter substitution, or not, irrespective of letter identities. Similarly, the generalisation of this approach provided by orthographic Levenshtein distance (Yarkoni et al., 2008) counts the minimum number of letter insertions, deletions, and substitutions required to convert one word into another. Such approaches overlook entirely the graphetic features from which letters are composed, and the contribution that (dis-)similarities among these low-level perceptual and orthographic features may make to word recognition processes.

Graphetic processing has also been greatly simplified, or overlooked entirely, in computational models of word recognition. As with measures of orthographic neighbourhood density, such models instead typically begin their descriptions at the level of letters' identities and their ordinal configurations within words (Norris, 2013). In one early and influential exception, Rumelhart and Siple (1974) pioneered an approach to modeling the processing of letters' composite strokes with an artificial monospace type font comprised of just 14 character segments of fixed position and dimensions. Via combinations of these 14 segments, all 26 Latin alphabetic characters can be approximated. The Rumelhart-Siple font enabled them to incorporate stroke-level

information into an interactive activation model of visual word recognition, implemented via an artificial neural network (ANN), without having to account for the shapes of real-world fonts. This innovation has remained a cornerstone in computational models of visual word recognition (e.g., Coltheart et al., 2001; Davis, 2010; Heilbron et al., 2020; McClelland & Rumelhart, 1981; Perry et al., 2007) that descend from the connectionist, interactive activation approach, and provides a framework that is extensible, supporting, for instance, descriptions of non-alphabetic scripts (Perfetti et al., 2005). And yet, models using variants of the Rumelhart-Siple font are, by definition, artificial, imposing synthetic structure and uniformity onto the diverse and heterogeneous typography of real-world letters. This limitation constrains the extent to which such models, and the orthographic similarities that can be extracted from them, reflect fully the true computations involved in orthographic processing.

More recent connectionist approaches include training ANNs to recognise real-world letters and words from raster-image representations (Agrawal & Dehaene, 2023; Chang et al., 2012; Hannagan et al., 2021). Rather than imposing artificial constraints on the input, graphetic representations in such ANNs are *emergent*, contingent on the model's architecture, training regime, training sets, and hyperparameters. In one recent and notable study, Agrawal and Dehaene (2023) trained the CORnet-Z neural network architecture, a convolutional neural network designed to reflect aspects of the primate ventral visual stream (Kubilius et al., 2018), on images of first, objects, and then, words. This study convincingly reproduced key effects observed in the so-called visual word form area, a region of the ventral occipitotemporal cortex (vOT) that has been consistently implicated in visual word recognition and orthographic processes (Dehaene & Cohen, 2011; Price, 2012; Woolnough et al., 2021). Agrawal and Dehaene showed that training the CORnet-Z architecture reproduced, in the late layers of the model thought to best capture particularly high-level ventral visual stream processing, the sensitivity to visual words observed in vOT, including its invariance to letter case, font,

position, and size. Agrawal and Dehaene also compared the model's representations to those of literate humans via a representational similarity analysis (RSA; Kriegeskorte et al., 2008) approach. By regressing out information that was already present in lower layers, and which was maintained and projected forwards into higher representations, they isolated the representational content that emerged in each layer. It was shown that the more low-level, retinotopic representations, in the CORnet-Z layer taken to reflect the primary visual cortex, predicted magnetoencephalographic (MEG) event-related fields elicited by strings of letters for a sustained period, from 60 ms until at least 500 ms post stimulus onset (although the infinite impulse response 1 Hz high-pass filter applied in preprocessing is likely to have distorted the component's onset somewhat; Rousselet, 2012; Tanner et al., 2015; VanRullen, 2011). A later layer in the model, taken to reflect subsequent processing in the inferotemporal cortex, predicted brain activity for a later sustained period, beginning at 140 ms post-stimulus. The onset of this latter period broadly aligns with that of the N1/N170 (or in MEG, M170) component commonly associated with orthographic processing (Bentin et al., 1999; Lin et al., 2011; Ling et al., 2019; Pleisch et al., 2019; Zweig & Pylkkänen, 2009) and which has a likely vOT source (Allison et al., 1994; Nobre et al., 1994; Pleisch et al., 2019; Woolnough et al., 2021). Overall, the hierarchical sequence of representations that emerged in CORnet-Z trained on images of words matched the progression observed in literate humans' brains, from lower-level, more retinotopic representations of letter shapes, to more abstract, invariant representations of letter identities encoded in ordinal positions in words. However, because representations in ANNs are emergent, rather than computationally constrained by a more interpretable model, it is inherently unclear how cognitive scientists should interpret the alignment between representations that emerge in ANNs like CORnet-Z and the neural representations of humans (Guest & Martin, 2023).

We propose that optimal transport theory may provide a framework for developing a model of early orthographic, graphetic processing, which avoids unrealistic constraints

on the orthographic inputs, yet is more computationally explicit and interpretable in its relationship to neural representations. Optimal transport refers to the efficient transportation of mass between two distributions, incurring a minimal transportation cost. For two points of equal mass in Euclidean space, the optimal transport plan between them can be trivially solved via their Euclidean distance. Optimal transport theory is primarily concerned with the generalisation of this problem to cases of many-to-many mapping, where we wish to efficiently transport mass between entire distributions of such points simultaneously (Peyré & Cuturi, 2020; Villani, 2009). Optimal transport solutions typically minimise *Wasserstein distance*, or "earth mover's distance", which represents an overall cost of transporting mass, calculated as the product of masses and the distances they must travel in the transport plan. Previous approaches to calculating distances between visual shapes and letters presented to participants have often used measures of visual overlap (e.g., Dunn-Rankin et al., 1968; Fischer-Baum et al., 2017; Humphreys et al., 1988; Laws & Gale, 2002; Ling et al., 2019; Op de Beeck et al., 2008; Qu et al., 2022; Sun et al., 2018), which can capture the extent to which letters share large components (e.g., strokes). Wasserstein distance permits a more elaborated description of shape similarity, representing a more global cost for the overall shape distance (**Supplementary Materials A**) and, when applied to letters, a more complete description of the (dis-)similarity structure among letter shapes. Accounting for global shape in this way may also better reflect the flexibility with which humans recognise the shapes of letters. For instance, literate humans routinely navigate a range of typefaces across which identical letter identities share global shape features, even though individual component features may not perfectly overlap. In the present study, we test this proposal with an RSA analysis of letter recognition processes.

The Present Study

In this preregistered study, we used RSA to examine the extent to which optimal transport Wasserstein distance aligns with patterns of neural activity elicited by letters,

comparing its performance to that of Jaccard distance (Gilbert, 1884; Jaccard, 1901), a common measure of shape overlap. Our aim was to test whether optimal transport theory provides a computationally explicit and interpretable model of the neural representation of letter shapes in humans. To this end, we exploit the high temporal resolution of electroencephalography (EEG) to constrain our hypothesis to a time window from 150-225 ms, matching the typical latency of the posterior N1 event-related potential (ERP) component associated with orthographic processing of letters and words (Bentin et al., 1999; Fraga-González et al., 2021; Lin et al., 2011; Ling et al., 2019; Madec et al., 2012; Petit et al., 2006; Pleisch et al., 2019; Rey et al., 2009; Wong et al., 2005; Zweig & Pylkkänen, 2009). We analysed patterns of neural activity elicited by letter stimuli in an alphabetic decision task (Grainger & Jacobs, 1991; Jacobs & Grainger, 1991), chosen to ensure that participants were attending to the shapes of letters.

We expected that the two measures of letter dissimilarity would share some information, and that both would align with neural representations of letters in the N1 window to some extent, but that Wasserstein distance should show greater representational alignment than Jaccard distance, as it provides a richer description of the global shape information relevant to perception of real-world letters. Specifically, we hypothesised (Hypothesis 1) that the correlation between the Normalised, Mass-Scaled Optimal Transport Wasserstein Distance matrix and neural representational dissimilarity matrices (RDMs) will be more likely to be greater than zero than it is to be less than zero, (Hypothesis 2a) that the correlation between the Normalised, Mass-Scaled Optimal Transport Wasserstein Distance matrix and neural RDMs will be higher than that between the Jaccard Distance matrix and neural RDMs, and (Hypothesis 2b) that the correlation between the Normalised, Mass-Scaled Optimal Transport Wasserstein Distance matrix and neural RDMs will uniquely explain more variance than the correlation between the Jaccard Distance matrix and neural RDMs, as assessed via partial correlations.

These hypotheses, as well as the data acquisition and analysis procedures, were

preregistered[1] on the Open Science Framework (https://osf.io/xwma2). The procedure and main analysis that we report here conform to the steps outlined in our preregistration, with the exception of two changes to the independent components analysis (ICA) used in the preprocessing pipeline. We outline those changes in the *Preprocessing* section, and explain them in more detail in **Supplementary Materials B**.

Finally, we ran exploratory analyses to gain further insight into our findings. First, we evaluate the performance of forms of Wasserstein and Jaccard distance that are invariant to geometric transformations, like translation and rescaling, reflecting the invariance known to emerge during ventral-stream visual orthographic processing. Second, we compare the performance of Wasserstein and Jaccard distance to that of two ANN models, ResNet-50 1.5 (K. He et al., 2015) and CORnet-Z (Kubilius et al., 2018), trained on the ImageNet image dataset (ImageNet Large Scale Visual Recognition Challenge 2012 dataset; Russakovsky et al., 2015) and images of letters. ResNet-50 is a 50-layer-deep convolutional ANN, widely used in computer vision tasks. CORnet-Z, which has previously been employed as a model of human word recognition (Agrawal & Dehaene, 2023; Hannagan et al., 2021), is a shallow convolutional ANN with an architecture informed by the primate ventral visual stream.

We find support for our hypothesis that Wasserstein distance aligns better than Jaccard distance with patterns of neural activity elicited by letters. We also find that implementations of Wasserstein distance that are invariant to geometric transformations better align with later brain activity, suggesting that Wasserstein distance could be useful for modeling the transition from retinotopic to more categorical perception of letter shapes. Finally, we show that while Wasserstein distance is more computationally

[1] We have made three changes to the wording of the hypotheses from our preregistration: (1) We have removed the word "partial", as partial optimal transport is equivalent to more standard *balanced* optimal transport when the masses of two distributions are equal, and we preregistered that we would scale the masses of letters to be equal, (2) we have rewordered "Wasserstein Optimal Transport Distance" to "optimal transport Wasserstein distance", to match the formatting and wording used elsewhere in this manuscript, and (3) we have reworded "neural activity" and "matrices" to "RDMs", to match standard RSA terminology. These changes do not affect how the hypotheses are interpreted or tested.

constrained and interpretable than features derived from ANNs, the best performing variants of Wasserstein distance perform similarly to the best layers of ANNs in predicting brain activity elicited by letters. In sum, we suggest that optimal transport can provide a powerful yet interpretable framework for modeling graphetic processing of single letters.

Methods

Participants

On the basis of a pilot experiment, we decided to recruit a sample of at least 15 participants. Using a very similar preprocessing and a preliminary version of the analysis pipeline to that applied here, we detected Bayes Factors of >10,000 in favour of all our hypotheses in the pilot experiment, with a sample of five participants. To examine our expected sensitivity to the hypothesised effects, we conducted 500 simulations using estimates derived from the pilot experiment, assuming our pilot participants to be representative of any future sample. Simulating datasets from 15 participants showed that all simulations had Bayes Factors >15 in the simulated direction, representing strong evidence (Jeffreys Harold, 1961; Kass & Raftery, 1995), for all hypotheses. Moreover, the smallest Bayes Factor of any simulation was 84.1 in favour of H2a, and no simulated dataset at N=15 showed a Bayes Factor >15 in the opposite direction to that hypothesised for any hypothesis. Over 90% of simulations revealed Bayes Factors >1,000. We note that the five pilot participants were mostly aware of the purpose of the study. However, given the low-level nature of the effect, we did not expect this to have greatly influenced the pilot experiment. In addition to our minimum sample size of 15, we also specified a subsequent stopping rule for a scenario in which we estimated Bayes Factors which we would not have considered conclusive (i.e., <15) from our sample size of 15 participants (see preregistration for details).

The 15 participants (12 female, 3 male) ranged from 20 to 27 years of age (M=23.2, SD=2.7). All participants reported German as their first language, and reported that they had no proficiency in speaking or reading a language that uses a

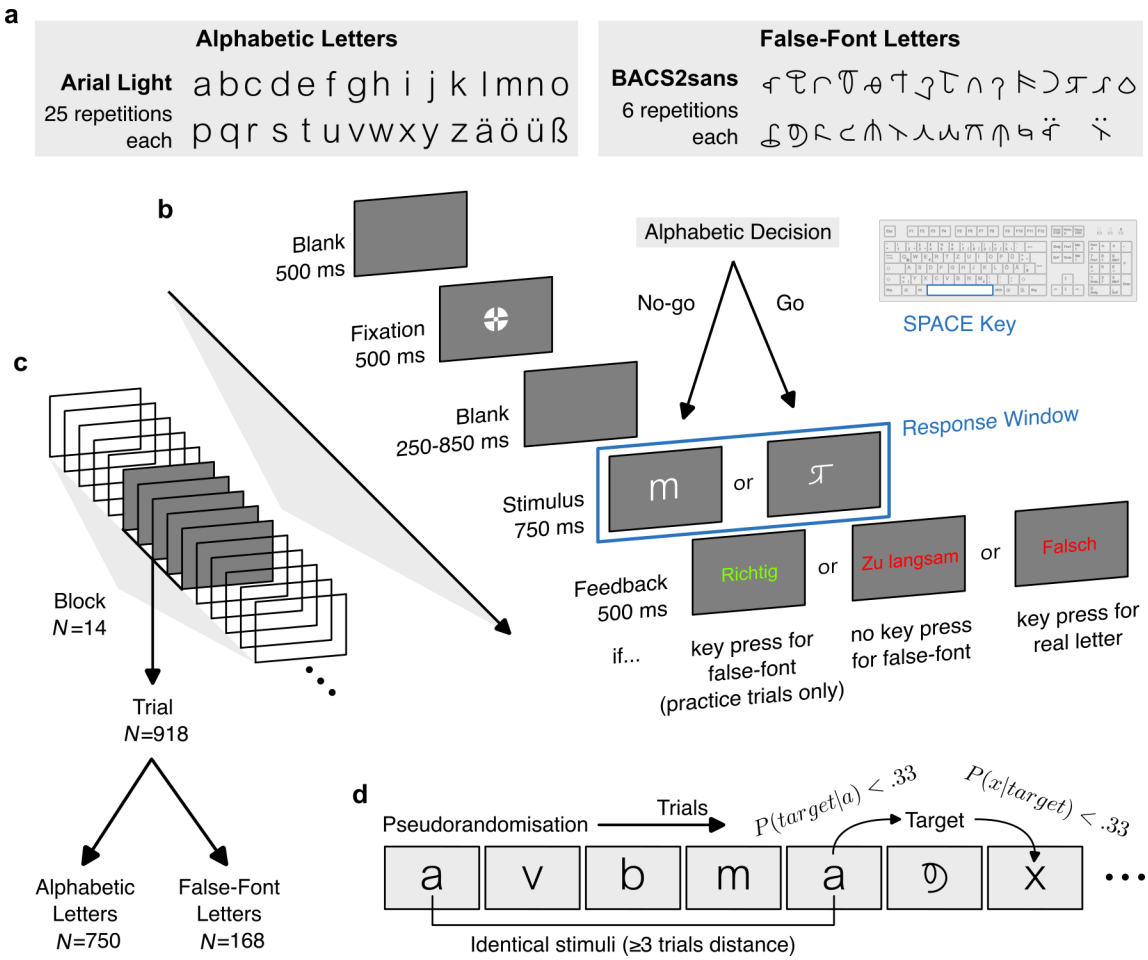
non-alphabetic script. Participants' handedness was assessed via the revised short form of the Edinburgh Handedness Inventory (Veale, 2014), with participants only permitted to take part if they scored a laterality quotient of at least +40 indicating right handedness. All participants completed education to at least level 4 in the European Qualifications Framework (i.e., in Germany, Abitur). No participants reported diagnosis of any disorder that would affect their reading, and all participants reported having normal or corrected-to-normal vision. Data from one participant were excluded from analyses due to a technical error that interrupted recording (see **Supplementary Materials C** for details). An additional participant was tested instead so that we still reached the required sample size. Participants were paid at a rate of €12 per hour, and ethical approval was provided by the Goethe University Psychology and Sport Science Ethics Committee (ethics ID: 2023-43).

Materials

Alphabetic letters comprised the 30 German letters in lower case (**Figure 1a**). These are the 26 alphabetic characters that German shares with English, plus the Umlaut and Eszett characters used in German: *ä*, *ö*, *ü*, and *ß*. Letter stimuli were presented in Arial Light font - a modified version of Arial with thinner strokes. This modified font was used to more closely match the stroke width of the false-font stimuli (explained below). This was done to prevent participants from using the width of characters' strokes to strategically discriminate between letter and false-font stimuli without recognising letters.

False-font stimuli comprised lower-case characters from the BACS-2 sans-serif font (Vidal et al., 2017, **Figure 1a**). This font is designed to resemble real sans-serif letters, matched on features including the numbers of strokes, junctions, and terminations. Although we presented 30 unique target stimuli, we only used 28 matched false-font characters. The characters *ß* and *ö* are excluded because for *ß* no equivalent in BACS-2 exists, while for *ö* we observed a display error in the rendered glyph. False-font equivalents of *ä* and *ü*, formed by combining the false-font equivalents of *a*

Figure 1
Alphabetic decision task paradigm.



(a) The alphabetic (left) and false-font (right) letters presented to participants. **(b)** Schematic visualisation of the trial sequence. **(c)** There were 918 trials in total, across 14 blocks, split into 750 alphabetic trials and 168 false-font letters. (**d**) The order of trials was pseudorandomised to avoid stimuli being repeated and to reduce the predictabilities of target or non-target identities. Image of German keyboard adapted (recoloured) from image by A.Illigen (https://commons.wikimedia.org/wiki/File:Tastatur_DE_hellgrau.svg; CC BY 4.0).

and *u* with umlaut diacritics, were displayed without error.

Procedure

On each trial (**Figure 1b**), participants were first presented with a blank screen for 500 ms, followed by the central "bullseye" fixation cross recommended by Thaler et al. (2013), for an additional 500 ms. There was then a blank screen for a jittered interval of

between 250 and 850 ms. After the blank screen, the stimulus was presented centrally for 750 ms, during which responses to false-font stimuli would be accepted as correct. Responses were given by participants pressing the space bar on a keyboard with the index finger of their right hand, when they detected a false font character. If participants responded to an alphabetic letter, they would be shown text in red reading "Falsch" (*"Incorrect"*) for 500 ms, before the next trial began. If participants failed to respond within 750 ms to a false-font character, they would be shown in red text reading "Zu langsam" (*"Too slow"*) for 500 ms, before the next trial began.

The 30 alphabetic letters were presented 25 times each, to each participant. Each of the 28 false-font characters were presented 6 times. This resulted in a total of 918 trials, split into 14 blocks (**Figure 1c**) of 66 trials each, except for the last block which comprised only 60 trials. Each block featured 10 - 14 trials presenting false-font characters, while all remaining trials were alphabetic letters. Trial order was determined for each participant pseudorandomly (**Figure 1d**), with the constraint that no letter identity would be repeated within 3 trials of its last presentation. In this constraint, true- and false-font counterparts are considered distinct letter identities. False-font characters were also presented pseudorandomly, with the constraint that, across the entire experimental session, the maximum transition probability of any alphabetic letter immediately preceding or following a false-font character was kept $<.33$.

At the start of the experiment there was a practice block comprising 17 alphabetic letters and 3 false-font characters. The practice block was otherwise identical to the 14 experimental blocks. If participants failed to respond within the 750 ms time limit to >1 false-font character in the practice block, they were asked to complete the practice block again. After successfully completing the practice trials, participants were asked to repeat, in their own words, the requirements of the task, to check that they understood the experiment. During the practice trials, participants additionally received feedback for correct responses, with text in green reading "Richtig" (*"Correct"*) for 500 ms, for every

false-font letter they correctly identified.

Participants were provided with self-paced breaks between blocks, during which they were presented with a summary of their performance in the block they just completed, showing: their average (median) response time in milliseconds and the number of mistakes they made together with a block-specific percentage (error rate).

Stimuli were presented at a distance of 103 cm (set via a chinrest), on a 1680$\times$1050 resolution display of dimensions 47.4$\times$29.6 cm, with a 60 Hz refresh rate. Letters and false-font stimuli were presented in white font on a grey-background screen, with the tallest characters' heights subtending 1.2°of visual angle.

Recording

EEG data were recorded from 64 active Ag/Ag-Cl electrodes (extended 10-20 system) using a BrainProducts actiCAP system. Data were recorded at a sample rate of 1000 Hz, with an online high-pass filter of .016 Hz. Electro-oculography (EOG) data were recorded from two pairs of flat bipolar electrodes, respectively comparing voltage differences between regions to the side of the outer canthi of each eye, and between regions directly below the participant's left eye and above the participant's left eyebrow. Electrode impedances were kept $<$15 kΩ, with any electrodes that failed to reach this threshold recorded for later removal and interpolation. Data were recorded in an electromagnetically shielded booth.

Triggers were delivered via parallel port to coincide with stimulus presentation. Calibration tests with photodiodes revealed a constant delay of 8 ms between trigger delivery and change in luminance for centrally presented stimuli. Triggers were additionally used to record any key presses.

Preprocessing

Preprocessing was implemented using MNE Python (Gramfort et al., 2013). First, channels that had been noted during the experiment as having high impedances ($\geq$15 kΩ) or being excessively noisy or flat-lining, had their voltage values interpolated (M=1.53

per participant; max=5 per participant). Interpolation was implemented using spherical spline models, estimated using per-participant head circumferences. Data were then re-referenced to the arithmetic mean of all EEG channels. We bandpass filtered data to between .1 and 40 Hz with a 4th order zero-phase Butterworth filter. Zero phase shift with 4th order was implemented by applying a 1st order filter in both directions, twice.

An ICA model was fit to the EEG data to identify artefacts in the data, using the extended infomax method (Lee et al., 1999) to identify 32 components. In our preregistration, we specified that we would estimate 64 components, equal to the number of EEG channels in the data. However, this led to the ICA failing to identify clear artefacts in the data, such as large blink signals, possibly due to rank deficiency. Instead, estimating 32 ICA components resulted in clearly partitioned components. As a result, we opted to estimate 32 components. We provide more detail on this in (**Supplementary Materials B**).

The ICA was fit to periods of data that excluded between-block break periods, identified as periods of at least 3 seconds in duration during which no triggers were recorded, with an additional 1.1 seconds of padding after the last trigger for stimulus presentation before the break period, and .5 seconds of padding before the first trigger after the break period. The ICA was fit to a version of the data with a bandpass filter more appropriate for detecting artefactual components, using a higher high-pass cutoff of 1 Hz. In our preregistration, we specified that the low-pass cutoff for the ICA would be 40 Hz, to match the version of the data used in all subsequent analyses. However, we opted to instead use a 100 Hz low-pass cutoff, for comparability with the dataset on which the ICLabel classifier was originally trained (Pion-Tonachini et al., 2019), as recommended as best practice by the MNE-ICLabel library (A. Li et al., 2022). In (**Supplementary Materials B**) we explain and justify these changes in more detail. After fitting the ICA, artefactual components were identified using the ICLabel classifier (Pion-Tonachini et al., 2019), via the MNE-ICLabel Python library (A. Li et al., 2022). ICA components were

identified as capturing artefacts if the ICLabel classifier assigned them any label except "brain" or "other" with a predicted probability of >85%. These artefactual components were then removed from the main version of the data filtered to within .1 - 40 Hz.

Epochs were defined as periods from stimulus onset until 1000 ms post-onset, with a 200 ms pre-stimulus baseline period. To account for the delay between trigger delivery and stimulus presentation (see *Recording*), we adjusted epochs to begin 8 ms later, to coincide with actual stimulus onset. Epochs were excluded if the participant responded to no-go (i.e., letter) trials, either during stimulus presentation, or within 250 ms of either the epoch onset or offset (including the 200 ms baseline period). Epochs were also excluded if any channels were detected as flat signals, with a minimum peak-to-peak amplitude below .5 µV, or if any channels were detected as noisy, with a maximum peak-to-peak amplitude above 150 µV. If any channels were detected as noisy or flat in more than 50% of epochs using these criteria, then preprocessing would have been re-run from the start for this participant, with the problematic electrodes interpolated along with those marked as problematic during the recording. However, this was not necessary for any participant. Finally, we verified that all participants satisfied the criteria for inclusion in the analysis concerning minimal trial numbers, as specified in the *Participants* section.

At the end of this pipeline, 1.7% of all epochs were excluded, with a mean of 12.73 epochs (range: 0-85) excluded per participant. The number of remaining epochs per letter for each participant ranged from 18 to 25 (M=23.4) in the preprocessed dataset.

Main Analysis

The main analysis comprised an RSA (Kriegeskorte et al., 2008), in which RDMs were constructed for all pairs of alphabetic letters, from measures of graphemic similarity (Jaccard distance and Wasserstein distance), and from the average ERPs elicited by each letter for each participant. The correlation between these RDMs was calculated via rank (Spearman) correlation. We opted to use rank correlations, as it would allow us to

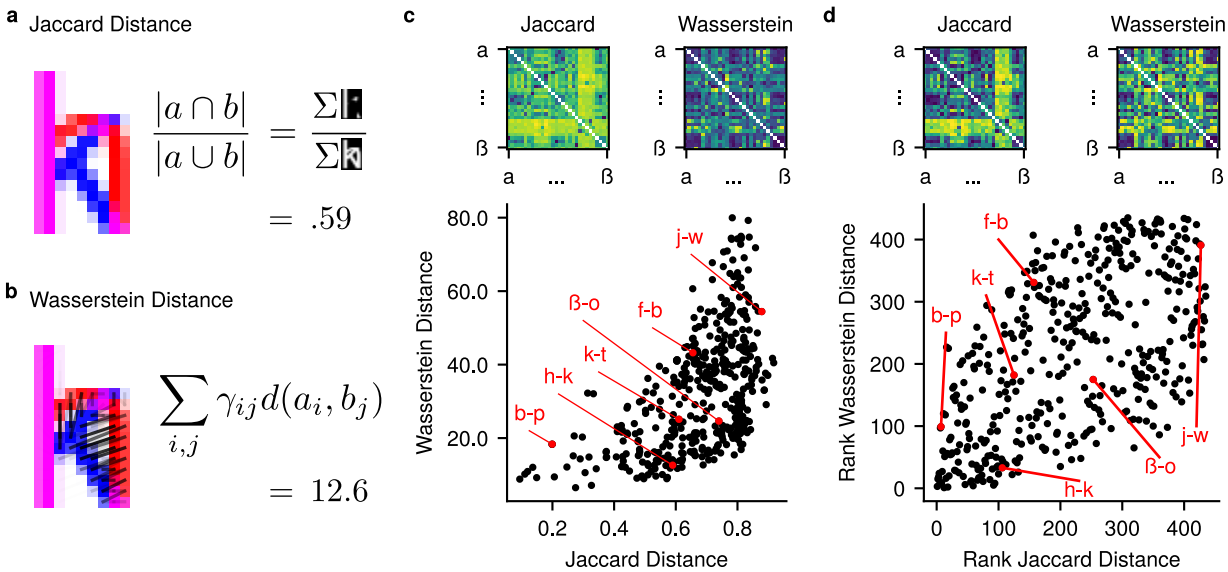
capture relationships without assuming a specific link function, or that the relationships are strictly linear.

Model RDMs

RDMs for graphetic similarity were calculated by first drawing the 30 alphabetic letters in 250-point Arial Light font via the Pillow fork (Clark, 2024) of the Python Image Library, and converting these to *numpy* (Harris et al., 2020) matrices. This produced matrices representing 8-bit raster images of the letters, in which the value of each element represented the intensity of a pixel value, normalised between 0 and 1 (i.e., the 8-bit intensity value divided by 255). The location of maximum overlap between the characters was then identified by means of two-dimensional cross-correlation. Specifically, the element at which peak cross-correlation was observed represented the location at which character a overlaps maximally with character b (Briechle & Hanebeck, 2001). Pairs of characters were then aligned on this basis, such that were the two matrices overlaid, the non-zero values would maximally overlap.

Jaccard and Wasserstein Distances were calculated for pairs of the resulting overlap-aligned matrices as described in **Supplementary Materials A**. Jaccard Distance was calculated as the intersection of the letters divided by their union (**Figure 2a**). Optimal Transport plans were solved via the network simplex algorithm implemented in the Python Optimal Transport library (Flamary et al., 2021), treating the pixel values as mass to be transported. Mass was scaled before solving the optimal transport plan, such that the two matrices each summed to one. The cost matrix, dictating the cost of transport between all pairs of pixels, was specified as the Euclidean distance. Wasserstein distance was calculated as the sum of the product of masses transported and the corresponding Euclidean distances (**Figure 2b**). **Figure 2c and 2d** depict the full model RDMs, the corresponding rank-transformed RDMs used in the RSA, and the relationships between them.

Figure 2
Jaccard and Wasserstein Distance RDMs.



(a, b) Calculation of Jaccard and Wasserstein distance between two example images of Arial *h* and *k* characters (a full description is provided in **Supplementary Materials A**). For visibility, the images shown are at much lower resolution than that used in the actual analysis. **(c)** The raw Jaccard and Wasserstein distances between pairs of characters, and the relationship between their lower triangles. **(d)** The relationship between the same RDMs, but with the distances ranked from 1 to 435, reflecting the rank correlation approach used in the RSA. The rank correlation between Jaccard and Wasserstein distance is ρ=.57. RDMs show dissimilarities for all pairs of basic alphabetic letters in the standard order, from *a* to *z*, followed by the German characters *ä*, *ö*, *ü*, and *ß*. The diagonal is excluded from the RDMs, but would always equal zero in the raw distance values. Example pairs of letters are highlighted to show how the measures differ.

Neural RDMs

Neural RDMs were calculated from ERPs using the *rsatoolbox* library - a Python port and update of the MATLAB RSA toolbox (Nili et al., 2014). Specifically, a neural RDM was constructed from the Pearson correlation distances between ERPs elicited by each letter, concatenating all average amplitudes elicited by each letter end-to-end. These vectors were calculated within the time window from 150 to 225 ms (providing, at the recorded sample rate of 1000 Hz, 75 samples for each electrode). Given the posterior scalp localisation of the previously reported effects and ERPs related to orthographic processes (see Introduction section), the neural RDMs were calculated

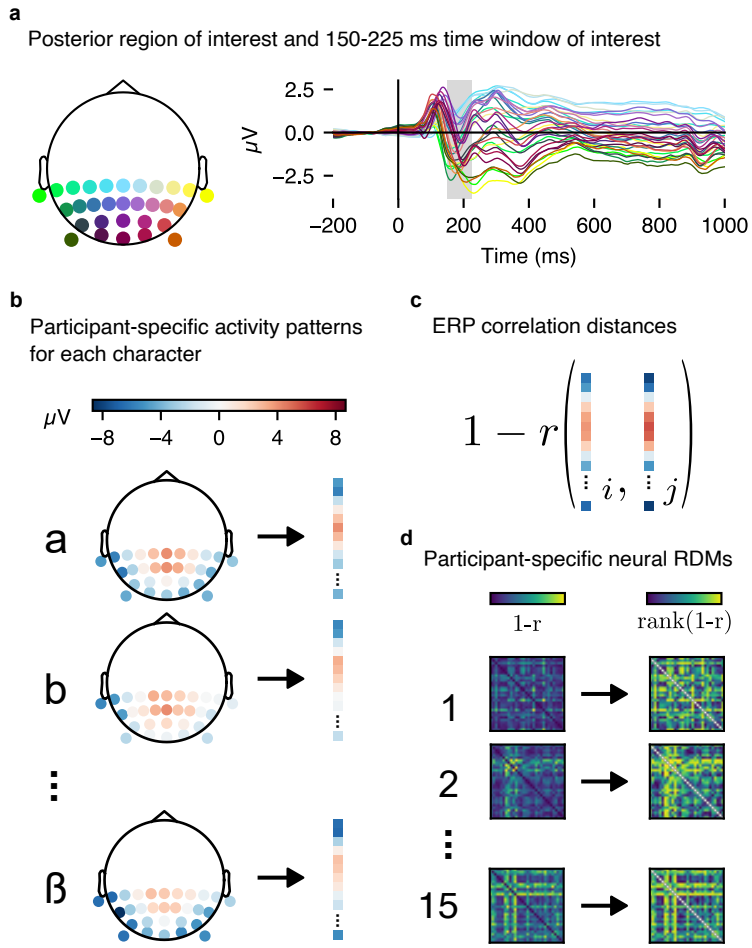
from 30 posterior electrode sites (**Figure 3a**). A neural RDM was constructed for each participant separately (**Figure 3b-d**), and rank transformed such that each value in the matrix was replaced by its rank, from 1 to 435, calculated from the lower triangle of the matrix. The noise ceiling for rank correlations with the EEG data was also calculated via the rsatoolbox library, with a leave-one-out procedure applied to participants' RDMs (Nili et al., 2014). Specifically, each leave-one-out iteration, an estimate of the lower bound of the noise ceiling was calculated as the correlation between the RDM of the left-out participant, and average of all participants' RDMs excluding the left-out participant. The upper bound, meanwhile, was calculated as the correlation between the left-out participant's RDM and the average RDM across participants *including* the left-out participant. The final noise ceiling bounds were then calculated as the average of these correlations across all iterations.

Correlation Estimation

We estimated the rank (Spearman's ρ) correlations between model and empirical neural RDMs (calculated from the full 150-225 ms period of interest; see previous section). The models were fit via *brms* (Bürkner, 2017), an interface to *STAN* (STAN Development Team, 2024). Full details on the models, sampling parameters, and prior specifications are provided in **Supplementary Materials E**. We additionally conducted a sensitivity analysis, which suggested that the observed pattern of effects was robust to even priors biased very strongly towards zero or one (**Supplementary Materials F**).

In addition to rank correlations, we calculated partial rank correlations, to quantify variation in the neural RDMs that is uniquely explained by Wasserstein Distance, with Jaccard Distance partialled out, and vice versa. Posterior distributions for partial correlations were yielded by converting correlation matrices from each posterior sample to partial correlation matrices, via the *corpcor* package for R (Schäfer et al., 2021).

Figure 3
Calculation of neural RDMs.



(a) We selected a region of interest (ROI) comprising posterior channels TP9, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, TP10, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, PO9, O1, Oz, O2, and PO10. The right panel shows average ERPs of all selected channels for alphabetic letters, with the time window of interest for the planned analysis (150-225 ms) highlighted in grey. **(b)** For each participant, we calculated per-character patterns of activity as a vector of the ROI channels' average ERP amplitudes for the given time window. Vectors from successive time points within the window were concatenated. **(c)** We calculated correlation distances between pairs of stimuli as 1 minus the Pearson's r correlation coefficient of the concatenated vectors. **(d)** We calculated correlation distance matrices in this manner for all participants, and rank-transformed the values to yield the rank RDM used in the RSA.

Results

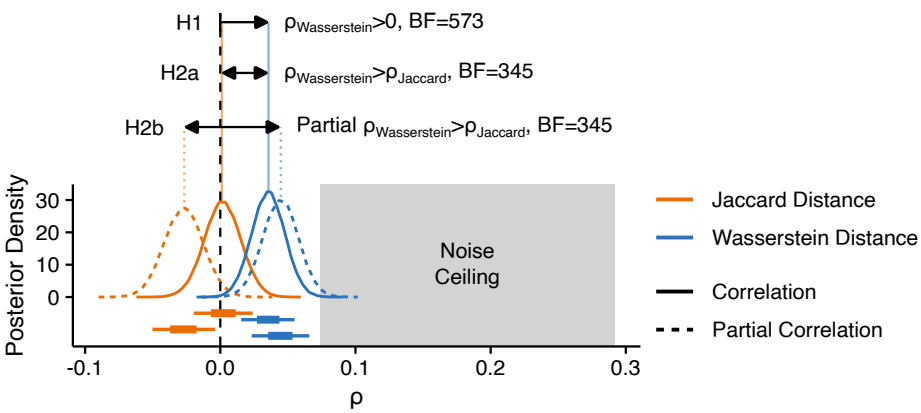
Planned Analysis

Posterior estimates (**Figure 4**) revealed a positive correlation for Wasserstein distance (median=.036, 89% HDI = [.016, .055]), while the correlation for Jaccard

distance was close to zero (median=.002, 89% HDI = [-.02 .024]). Partial correlations further showed that Wasserstein distance uniquely explained more variation in neural RDMs (median=.045, 89% HDI = [.023, .066]). In contrast, the partial correlation between Jaccard Distance and neural RDMs was negative (median=-.027, 89% HDI = [-.050, -.004]), showing that once the variation shared by Jaccard distance and Wasserstein distance was discounted, Jaccard Distance was in fact systematically *anti*-correlated with neural RDMs.

Figure 4

Planned analysis results.



Posterior distributions for correlations with neural RDMs for posterior electrodes in the 150-225 ms window, for Jaccard (*orange*) and Wasserstein (*blue*) distance. Correlation (*solid lines*) and partial correlation estimates (*dashed lines*) are shown separately. Box and whisker plots below the densities depict the extents of the 50% (*box*) and 89% (*whiskers*) highest density intervals (HDIs) of the correlation estimates. The grey region to the right depicts the extents of the noise ceiling.

To quantify support for our hypotheses, we calculated Bayes Factors as ratios of the number of posterior samples consistent, to the number of samples inconsistent, with our hypotheses, after observing the data. This is appropriate because our symmetric priors resulted in prior odds equal to 1 for all hypotheses, such that the ratios between posterior odds and between the Bayes Factors are equivalent (Schmalz et al., 2023). We note that the Bayes Factors we use are not tests of point hypotheses, comparing $P(\theta \neq 0 \mid y)$ to $P(\theta = 0 \mid y)$. Instead, we test informative hypotheses (Hoijtink et al., 2008) in a directional manner, comparing $P(\theta > 0 \mid y)$ to $P(\theta < 0 \mid y)$. Were a posterior

distribution for a parameter to be perfectly centred on 0, then there would be an equal number of posterior samples for both $P(\theta > 0 \mid y)$ and $P(\theta < 0 \mid y)$, resulting in a Bayes Factor of 1 exactly. A posterior distribution in the opposite direction to our hypothesis, meanwhile, would result in a Bayes Factor of $<$1.

In support of Hypothesis 1, we found strong evidence that the correlation between Wasserstein Distance and neural RDMs is greater than zero (*BF*=573). For H2a and H2b[2], we also found strong evidence that the brain-model correlation was higher for Wasserstein Distance than for Jaccard Distance (*BF*=345).

As a control analysis, we tested whether our results could have been confounded by effects of letter size, letter frequency, or phonological representations of letters. This analysis revealed a similar pattern of results for Jaccard and Wasserstein distance to that detailed above (**Supplementary Materials D**).

Exploratory Analyses

In addition to the planned analysis, we performed exploratory analyses delineating the time-course of brain-model alignments, implementing different forms of geometric invariance in Wasserstein and Jaccard distance, and examining how well ANNs trained on image and letter recognition align with the same neural data. We fit Bayesian models with similar specifications to those in the planned analysis (see **Supplementary Materials E** for details).

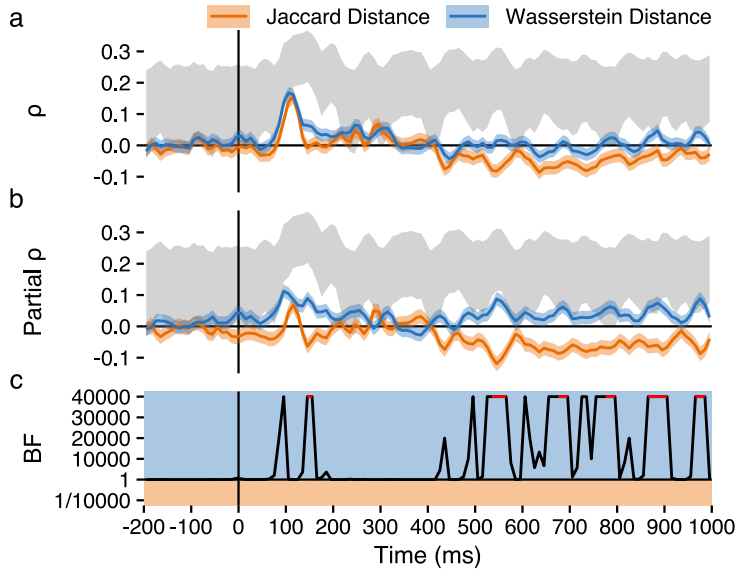
Time-Course Analysis

To more fully describe the dynamics of brain-model correlations, we fit time-resolved models to the full time-course of data, in bins of 10 ms. At the recorded sample rate of 1000 Hz, these correspond to non-overlapping windows of 10 samples each. As in the planned analysis, observed amplitudes from successive samples in a window were concatenated end-to-end.

[2] The Bayes Factors for for H2a and H2b are here identical, because transforming correlations to partial correlations preserves the rank order of the correlation matrix.

Results (**Figure 5a**) showed an early peak in the correlation between brain and model RDMs at around 100 ms after letter presentation, with a similar peak correlation for Wasserstein (peak median ρ=.17 at 100-110 ms) and Jaccard distance (peak median ρ=.16 at 110-120 ms). Partial correlations (**Figure 4b**) showed, however, that Wasserstein distance uniquely explained more variation in the neural RDMs, with the partial correlation peaking at ρ=.14 (90-100 ms) for Wasserstein distance, but only at ρ=.06 (110-120 ms) for Jaccard distance. Of particular interest, there was a sustained period from around 150 to 225 ms, corresponding to the period of interest from our main analysis, during which the correlation was larger for Wasserstein distance, and the partial correlation for Jaccard Distance was negative.

Figure 5
Time-resolved results.



Exploratory analyses of time-resolved estimates calculated in 10 ms windows, showing point-wise medians and 89% HDIs for (**a**) rank RDM correlations, and (**b**) partial rank RDM correlations. (**c**) Time-resolved Bayes factors quantifying evidence for $\rho_{Wasserstein} > \rho_{Jaccard}$ (Bayes factors for non-partial and partial correlations are identical). For periods in red, all posterior samples supported the hypothesis.

Surprisingly, after 400 ms, there was a sustained period during which the correlation between neural RDMs and Jaccard distance was consistently negative. This suggests that more dissimilar patterns of neural activity were elicited between letters

whose shapes overlapped more.

Analyses of Geometric Invariance

A robust feature of the processing of visual categories across the ventral visual stream is the emergence of invariance to category-irrelevant manipulations (Grill-Spector et al., 1999; Ito et al., 1995; Rust & DiCarlo, 2010). This also applies to ventral visual processes involved in letter and word recognition, with representations becoming increasingly abstracted from visual input as information is passed along a posterior-to-anterior spatial gradient over time (Petit et al., 2006; Vinckier et al., 2007; Woolnough et al., 2021). It is across this gradient that invariance to variables like letter case (Dehaene et al., 2001; Polk & Farah, 2002) and script (Krafnick et al., 2016; Zhou et al., 2019) emerge. Similarly, letter and word processing also show a degree of invariance to *geometric* transformations - particularly to differences in retinotopic location (i.e., *translation*; Cohen et al., 2000, 2002; Dehaene et al., 2004; Dufau et al., 2008; Hannagan & Grainger, 2013; Taylor et al., 2019) and visual angle (i.e., *scale*; Chauncey et al., 2008; Eddy & Holcomb, 2009; Legge & Bigelow, 2011; Rajalingham et al., 2020). An additional geometric transformation of interest in letter perception is *rotation*, to which orthographic processes show a degree of tolerance (Fernández-López et al., 2022; Fernández-López et al., 2023; Kim & Straková, 2012; Perea et al., 2018; Yoshihara et al., 2024).

To characterise the relevance of geometric invariance to letter representations, we calculated forms of Wasserstein and Jaccard distance that are invariant to different combinations of translation, scale, and rotation. For this, we used the Nelder-Mead method of non-linear optimisation (Nelder & Mead, 1965) to identify the geometric parameters that produce the minimal Wasserstein (Figure **6a**) and Jaccard (Figure **6b**) distances between pairs of letters, varying one letter's parameters while keeping the other constant. To avoid identifying local minima, we ran the algorithm with five starting values for each optimised dimension, spaced equidistantly from the lower to the upper

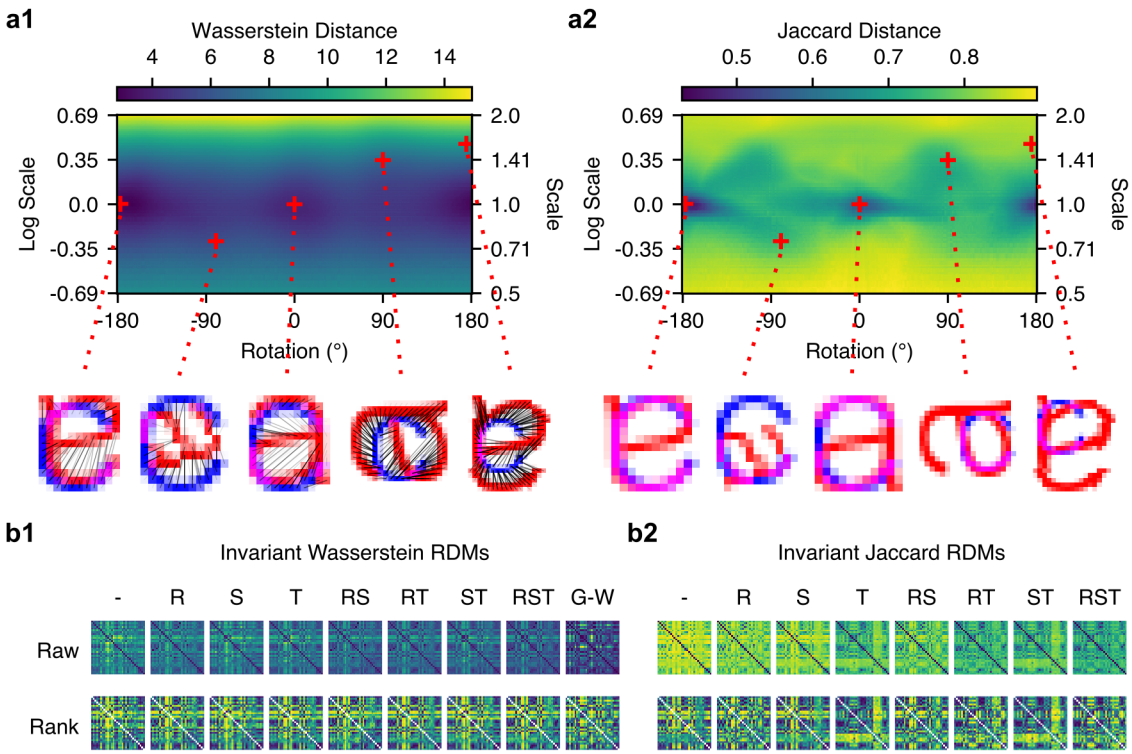
bounds of the dimension. Translation was bounded between $-x$ and $2x$, where x is the width or height of the letter image, and zero is at the top-left origin of the image. To reduce computational intensiveness, the optimisation of translation was implemented for Jaccard distance by finding the maximum of the padded images' two-dimensional linear cross-correlation matrices, which provides a much more efficient method for maximising overlap (Briechle & Hanebeck, 2001). Scale was optimised on a log scale, bounded between $log(0.5)$ and $log(2)$ (i.e., between half and twice the original size, respectively). Finally, rotation was optimised without any explicit bounds, but with starting values spaced between -180 and 180. For feasibility, letter images were drawn at 75-point font, rather than the 250-point font used in the planned analysis.

We note that in our main, preregistered analysis, we aligned characters using two-dimensional cross-correlation before calculating model RDMs. This means that the preregistered Jaccard distance measure was already translation-invariant using its optimal location, matching matrix *T*, in **Figure 6b2**, whereas the Wasserstein distance measure in the preregistered analysis was approximately translation-optimal.

As an additional comparison, we calculated Gromov-Wasserstein distance between pairs of letter images (**Figure 6b1**). This measure generalises Wasserstein distances to the problem of transporting masses between distributions that do not exist in the same metric space (Mémoli, 2011). Applied to images, Gromov-Wasserstein distance can be understood as quantifying the distances between *configurations* of mass in each image (Peyré & Cuturi, 2020). This makes Gromov-Wasserstein distance invariant to affine transformations including translation, scale, and rotation, but also less sensitive to nonlinear or non-affine transformations which preserve the configuration of mass. We calculated Gromov-Wasserstein distance via the Python Optimal Transport library (Flamary et al., 2021).

Finally, we were interested in the early peak in brain-model correlation observed at around 100 ms. As an exploratory comparison to the preregistered 150-225 ms period

Figure 6
RDMs calculated with nonlinear optimisation of geometric parameters.



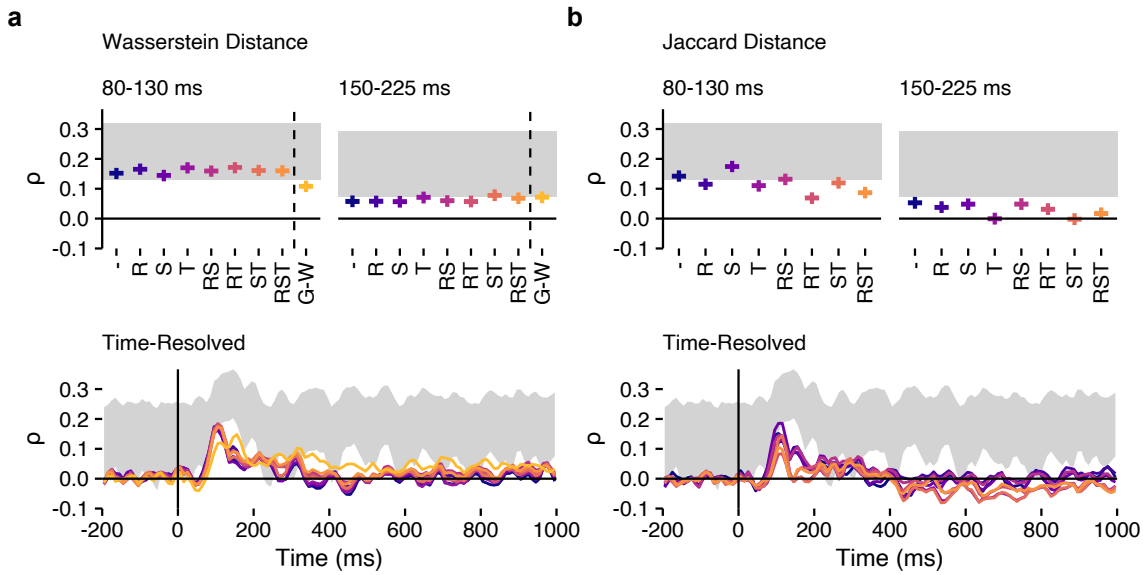
Wasserstein and Jaccard distance RDMs were calculated with invariance to different combinations of geometric transformations. **(a1)** An example space, over which Wasserstein distance was optimised for the pair of letters, *a* and *c*. (**a2**) Jaccard distances in the same space. Translation (vertical and horizontal) was additionally optimised, via nonlinear optimisation for Wasserstein distance, and via peak two-dimensional cross correlation for Jaccard distance. Five locations are highlighted, in lower resolution than used in calculating the actual RDMs, to illustrate the differences between the spaces for Wasserstein and Jaccard distance. The example locations are identical in scale and rotation, but show differences in optimal translation. (**b1, b2**) RDMs for Wasserstein and Jaccard distance that are invariant to different combinations of geometric transformations, where values are calculated as the minima of per-letter spaces depicted in **a1** and **a2**. (-=default positions, *R*=rotation, *S*=scale, *T*=translation, *RS*=rotation and scale, etc.) The Gromov-Wasserstein RDM (G-W) is also depicted for comparison, though it was not calculated using the nonlinear optimisation method.

of interest, we examined a second, earlier period of 80-130 ms, chosen to capture this peak and reflect a typical P100 latency.

Results showed that for Wasserstein distance (**Figure 7a**), invariance to geometric transformations slightly improved alignment with neural RDMs during the

150-225 ms period of interest. For Jaccard distance (**Figure 7b**), meanwhile, geometric invariances altered the correlation, but did not seem to improve it. In the 80-130 ms window, the brain-model correlation was very high for Wasserstein distance, though the most invariant measure, Gromov-Wasserstein distance (G-W), showed the lowest correlation. For Jaccard distance, there was more variability between the geometric variants, with scale-optimised Jaccard distance performing the best, though geometric invariances did not consistently improve the brain-model correlation.

Figure 7
RSA results for the transformation-invariant RDMs.



Spearman rank correlation estimates for transformation-invariant measures are shown for (**a**) Wasserstein and (**b**) Jaccard distance. Two windows are highlighted: an exploratory 80-130 ms window, and the original 150-225 ms period of interest. Horizontal lines depict medians of posteriors, while vertical whiskers show the extents of 89% HDIs. The lower panels show time-resolved estimates (medians of posteriors), calculated with 10 ms windows, as in the main analysis. Grey regions depict the noise ceiling, and colours are consistent across panels. As in **Figure 6**, model labels reflect the geometric transformations that have been optimised (-=default positions, *R*=rotation, *S*=scale, *T*=translation, *RS*=rotation and scale, etc.), while *G-W* refers to Gromov-Wasserstein distance.

It is notable that the RDM for Jaccard distance that was optimised for translation only (*T*, in **Figure 6b2**), which was the method used for the preregistered analysis, performed the worst of all the Jaccard distance variants compared, suggesting that the

variant used in the preregistered analysis may not have been representative of the other variants possible. Nevertheless, this exploratory analysis suggests that Wasserstein distance consistently outperforms Jaccard distance in predicting neural dissimilarities. Indeed, the variant of Wasserstein distance for which translation was optimised via nonlinear optimisation of the Wasserstein metric also performed somewhat better than the cross-correlation-aligned variant of Wasserstein distance used in the preregistered analysis.

The most salient difference was observed between the transformation-optimised variants of Wasserstein distance, and Gromov-Wasserstein distance. Specifically, Gromov-Wasserstein performed somewhat worse than the Wasserstein distance measures in predicting early activity at around 100 ms (see **Figure 7a** upper panel), but performed much better in predicting neural dissimilarities in later periods, with a marked peak at around 160 ms, and prolonged period of better performance throughout the epoch (see **Figure 7a** lower panel). This is an interesting finding, and is consistent with the more invariant correspondences between letter shapes, captured by Gromov-Wasserstein distance, being processed later.

Comparison to ANNs

A common application of RSA is to examine the representational alignment between humans and ANNs (e.g., Caucheteux & King, 2022; Cichy et al., 2016; T. He et al., 2022; Xu & Vaziri-Pashkam, 2021). In contrast to computationally constrained measures like Wasserstein and Jaccard distance, estimates of representational similarity extracted from ANNs are largely unconstrained, instead emerging from higher-level computational goals like image classification. To compare the performance of Wasserstein distance to that of ANNs, we trained two ANN architectures on image and letter classification: ResNet-50 1.5 (K. He et al., 2015) and CORnet-Z (Kubilius et al., 2018). For both ANN architectures, we tested three variants, trained to classify images of letters in Arial Light font, images from ImageNet (Russakovsky et al., 2015), or on both

letters and ImageNet images concurrently. Full details of ANN training are provided in **Supplementary Materials G**.

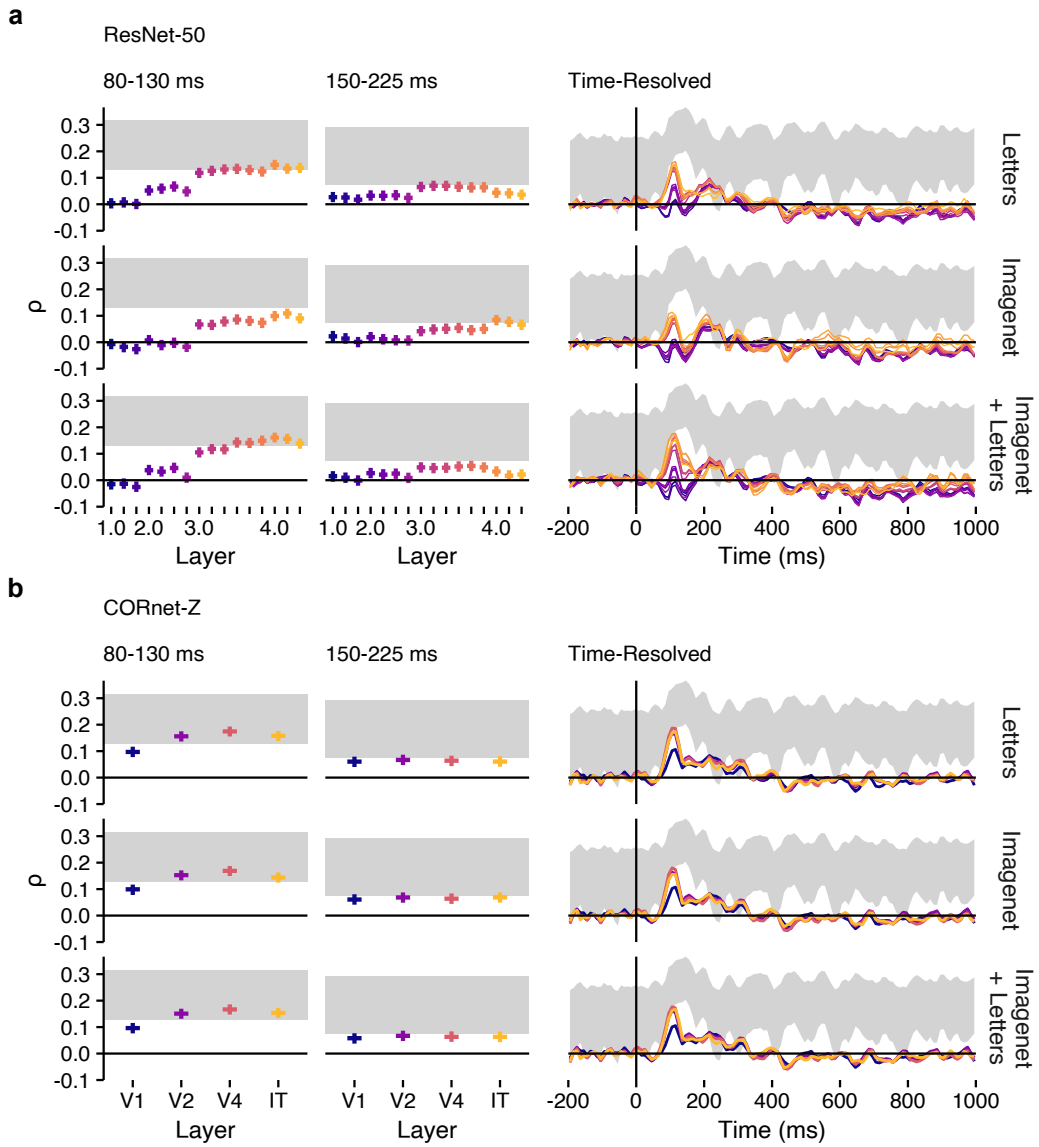
RDMs were calculated for ANN features by providing as input images of the letter stimuli, then extracting features via hooks attached to modules in the ANNs as implemented in the *torchextractor* library (https://github.com/antoinebrl/torchextractor). Brain-model rank correlations were estimated via multivariate Bayesian models, as described in other sections.

Results (**Figure 8**) showed that model RDMs calculated from activation patterns of the best-performing layers of ResNet-50 and CORnet-Z performed similarly to Wasserstein distance in predicting neural RDMs (compare **Figure 7**). The analysis also revealed that ResNet-50 and CORnet-Z activations aligned similarly well with patterns of neural activity, though the brain-model alignment observed for layers CORnet-Z varied much less by differences in training materials than did the alignment observed for ResNet-50. Finally, both ResNet-50 and CORnet-Z showed a pattern in which the layers which aligned best with neural activity were generally deep in the model, but not final.

Discussion

Investigations into visual word recognition have typically overlooked low-level orthographic (i.e, graphetic) processes (Finkbeiner & Coltheart, 2009). In this study we investigated whether Wasserstein distance, a metric from optimal transport theory, aligns with similarities in neural activity elicited by alphabetic letters in early, graphetic processes. We demonstrated that the measure outperforms Jaccard distance, a description of simple overlap between shapes which has been applied in previous models of graphetic similarities (e.g., Dunn-Rankin et al., 1968; Fischer-Baum et al., 2017; Ling et al., 2019; Qu et al., 2022; Sun et al., 2018). We additionally suggest that calculating forms of Wasserstein (and Jaccard) distance that are invariant to different geometric transformations could be valuable as a tool for investigating geometric invariances that are known to be critical in human letter and word perception. Finally, we

Figure 8
RSA results for the ANNs.



Spearman rank correlation estimates for module feature similarities of (**a**) ResNet-50 and (**b**) CORnet-Z, trained on (top to bottom) letters, the Imagenet dataset, and both. The results for the two periods and time-course are calculated and shown as in **Figure 7**.

show that Wasserstein distance performs comparably, in predicting patterns of neural activity, to less computationally explicit representations that emerge in deep artificial neural network architectures.

We believe that a major reason for previous investigations overlooking graphetic-level processes has been the lack of a computationally explicit framework

capable of capturing the graphetic features of letters in naturalistic typographies. We argue that optimal transport theory can provide such a framework. Importantly, however, we do not claim that early orthographic processing in the brain involves algorithms that solve optimal transport problems. Rather, optimal transport between images of letters *aligns* with distances between neural representations that capture the same information. At present, we remain agnostic as to the algorithms that may describe how the brain builds and processes those representations. Just as measures like Coltheart's *N* and OLD20 have been used to establish empirical orthographic findings that computational models of visual word recognition should be able to recreate (Andrews, 1997; Norris, 2013), measures like Wasserstein distance may be used to establish graphetic-level findings that computational models of cognitive and neural processes should be able to explain.

An advantage of applying a framework like optimal transport theory to establish such effects is that it is computationally transparent and explicit. This contrasts with the more opaque and emergent representations that can be extracted from ANNs, which we showed to perform similarly to (but not better than) Wasserstein distance in predicting patterns of neural activity elicited by letters. To be clear, we believe that ANNs can be informative tools in cognitive science, including the cognitive neuroscience of visual word recognition. ANNs provide an avenue, via alterations to model architecture and training, to formulate and test hypotheses about the nature of representations that may emerge for naturalistic stimuli. In particular, methods like RSA allow researchers to examine the alignment (Sucholutsky et al., 2023) between the representational content of ANNs, and representations reflected in behaviour (e.g., Daube et al., 2021; Marjieh et al., 2022; Rajalingham et al., 2018) or neural activity (e.g., Antonello & Huth, 2024; Caucheteux & King, 2022; Cichy et al., 2016). Researchers can attempt to infer from such representational alignment that the model and neural processes also align in computation. However, an enduring obstacle for all computational models of cognitive

processes is that of multiple realisability (Putnam, 1967), which is to say that two systems that align in representational content need not align in how that representation is realised - i.e., at the level of the neural implementation or cognitive mechanisms. As Guest and Martin (2023) argue, this means that it is a logical fallacy to infer from alignment alone that ANNs implement the same computations or algorithms as the brain. This concern is not restricted to ANNs, and can be applied to inferences about human cognition made on the basis of any model, including an optimal transport description of letter similarities. Nevertheless, one route to making more valid inferences may be to constrain models to ensure that, "...models resemble phenomena because they indeed somehow capture the essence of the phenomenon" (Guest & Martin, 2023, p. 219).

Certainly, imposing theory- and biology-informed constraints on the architecture of ANNs can improve the quality of inference and understanding of cognitive and neural mechanisms that such models permit (Kubilius et al., 2018; Mano et al., 2021; Pang et al., 2021; Pogoncheff et al., 2023; Schyns et al., 2022; Sucholutsky et al., 2023). However, to borrow from Marr's (1982) levels of analysis, ANNs constrained at the level of implementation can nevertheless remain unconstrained at the levels of computation and algorithm. Such ANNs are trained with a high-level computational goal (e.g., classification) yet researchers commonly extract and interpret features from intermediate layers, where relevant representations are emergent, rather than explicitly constrained. While model depth in ANNs trained on visual object classification does somewhat correspond to depth in the ventral visual stream (Cichy et al., 2016), the lower and intermediate model layers typically explain the most variance in neural activity, while later layers show much poorer brain-model alignment (Xu & Vaziri-Pashkam, 2021). This is a finding we replicated to some extent, with features extracted from deep, but not final, layers aligning best with neural data. A parallel can be drawn with large language model ANNs trained on next-word prediction: such models show the best alignment with neural representations of language at layers of an intermediate model depth of 60-80%, while

later layers that process information more relevant to the model's overarching computational goal show much poorer brain-model alignment (Antonello & Huth, 2024). Rather than constraining implementation or algorithm, the optimal transport description of letter similarities which we have suggested is constrained at the level of computation; it could be instantiated by multiple algorithms or implementations.

Indeed, we believe that a range of model architectures could reproduce the effect of graphetic similarity that we identified. For instance, an ANN-like model, in which letter strokes and shapes are processed by hierarchically layered feature detectors (e.g., Dehaene et al., 2005; Grainger, 2008), which show a tuning-curve response determined by their input's distance from preferred parameters, could be expected to reproduce our finding, as Wasserstein distance between letters would likely align with distances from such tuning-curve preferences for letter detectors. Conversely, a more symbolic language-of-thought model of letter shape processing (e.g., like the proof-of-concept model implemented by Sablé-Meyer et al., 2022), in which letter shapes are encoded as minimum-length expressions in an abstracted language of geometry, could also be expected to align with Wasserstein distance between letters, as letters with more similar geometry should also be expected to possess similar specifications in such a language. Our contribution is to propose that optimal transport could serve as a model-agnostic computational framework, that can describe general principles of graphetic processing that such models share.

Such models raise the question, however, of whether graphetic distances should be calculated between pixel-based representations, as applied in our analysis, or from stroke-based representations. Both connectionist feature detectors and language-of-thought primitives may be expected to encode features of letter shapes in a manner that is invariant to stroke thickness. Indeed, for arbitrary shapes, distances between topological skeletons align well with subjective judgements of similarity (Ayzenberg & Lourenco, 2019; Lowet et al., 2018) and neural correlates of shape

processing in the visual cortex (Ayzenberg et al., 2022). We suggest that the optimal transport framework we have applied here could be similarly applied to stroke-based representations of letters. For instance, Wasserstein distance could be calculated from stroke-based representations, as extracted or described by existing algorithms for processing typography (e.g., Berio et al., 2022; Jakubiak et al., 2006; Knuth, 1979). We believe that comparisons between pixel- and stroke-based representations may be an interesting avenue for future work on computational descriptions of graphetic similarity. Nevertheless, a stroke-based approach would likely perform similarly well, in describing letters, to the pixel-based approach that we have applied here. This is because, unlike the shapes employed in studies on general shape processing, letters are moderately uniform in their width, such that descriptions of stroke-based representations are likely to correlate highly with descriptions of pixel-based representations.

A more general question that may be raised about the calculation of graphetic distances concerns the relevance of graphetics to visual *word* recognition. On the one hand, it has been acknowledged that whole-word shape can aid word recognition in visual scripts (Allen & Emerson, 1991; Healy & Cunningham, 1992; Lavidor et al., 2001; Perea & Rosa, 2002; Simon et al., 2007), and it is indeed possible to calculate graphetic distances for images of whole words. However, recognising words on the basis of their shape alone can be understood as an inefficient strategy, as it would require learning representations of word shapes for all words known to the reader, across variations in typography like letter case and font (Grainger, 2008). Rather than using whole-word shapes, readers are generally thought to use a more computationally efficient (Grainger, 2008, although see Pelli et al., 2003) strategy of identifying component letters via *their* shapes and features, before subsequently processing words as ordered strings of abstract letter identities (Dehaene et al., 2005; Norris, 2013). In this case, graphetic features of component letters should show downstream consequences on word recognition. Indeed, effects of visual similarities of letter substitutions have been

demonstrated in orthographic priming (Marcet & Perea, 2017, 2018), and, in some cases, in unprimed lexical decision behaviour (Perea et al., 2022, 2024; Rocabado et al., 2024). Similarly, using a framework like optimal transport to model graphetics may provide an avenue to better understand established downstream effects of typography on word recognition (e.g., Kuchinke et al., 2014; Macaya & Perea, 2014; Nazir et al., 1998).

Beyond alphabetic orthographies, graphetic descriptions of visual words are also relevant to modeling word recognition in more logographic scripts, in which single glyphs or characters can represent entire words. This relates more broadly to the overreliance of visual word recognition research on English and alphabetic orthographies, such that models and theories of orthographic processing show poor generalisation to other languages and scripts (X. Li et al., 2022; Share, 2008). This is despite a wealth of evidence for shared neural mechanisms of orthographic processing across scripts (Bai et al., 2011; Bolger et al., 2005; Krafnick et al., 2016; Taylor et al., 2019). A key strength of the approach to calculating graphetic similarities which we have suggested here is that it could be applied to any visual orthographic script. Given that all visual orthographies begin with a visual stage of processing, a robust framework for describing graphetic features of real-world characters could contribute to the goal of identifying universal features and mechanisms of visual word recognition processes.

Even beyond orthography, an optimal transport description may provide insights into visual shape processing more generally. Just as with letters, visual similarity is often calculated for arbitrary visual shapes via measures that capture information about overlap (e.g., Dunn-Rankin et al., 1968; Humphreys et al., 1988; Laws & Gale, 2002; Op de Beeck et al., 2008). It is possible that, as with letters, such investigations into the effects of visual shape similarity could benefit from the more nuanced and sensitive descriptions provided by optimal transport theory. Visual shape processing also shares with orthographic processing the emergence, across the ventral visual stream, of a degree of invariance in recognition to variations in geometric transformations (Goodale &

Milner, 1992; Isik et al., 2014; Ito et al., 1995; Rust & DiCarlo, 2010). As a result, such research may also benefit from the invariant representations captured by descriptions like Gromov-Wasserstein distance. In sum, we view an optimal transport framework as a promising avenue for describing and investigating general visual shape processing.

A final point that we wish to mention is that our analyses have assumed that all features of a letter are equally important. However, in both letter (Fiset et al., 2008, 2009) and word (Rosa et al., 2016) recognition, the features that make up letters have been shown to be differentially informative. The optimal transport framework we have suggested is in fact capable of accommodating such findings. First, we note that regions which best differentiate characters, and which may thus be more informative about letter identity, are likely to be those which have the greatest distance from features of other characters. For instance, the informativeness of stroke terminations (Fiset et al., 2008; Rosa et al., 2016) may be captured, in part, by the manner in which stroke terminations will be a larger distance from the image centre, where more letters are likely to have features. In this way, the present implementation of Wasserstein distance may already capture some information about the relevance of different letter regions. Second, further differences between the perceptual informativeness of different regions of letters could be captured by distributing the letters' masses differentially across the image space, upweighting the contribution of more informative letter regions to the overall Wasserstein distance between letters.

In sum, we provide evidence that Wasserstein distance between letters can provide a computationally explicit measure of graphetic distance between letters which aligns with the distance between corresponding neural representations. While the graphetics of real-world characters has been largely overlooked in visual word recognition research, we believe that using a framework like optimal transport to explicitly quantify aspects of letter and word shapes can provide key insights into orthographic processing, and possibly into visual shape processing more generally.

Preregistration

The experiment, preprocessing, and planned analysis were preregistered on the Open Science Framework: https://osf.io/xwma2

Data Availability

The EEG and corresponding behavioural data are shared openly in BIDS (Pernet et al., 2019) format via OpenNeuro: https://openneuro.org/datasets/ds005594

Code Availability

Scripts, outputs, and a Docker container to reproduce analyses, including preprocessing, RDM calculation, and model fitting, are shared in a GIN repository: https://gin.g-node.org/JackEdTaylor/lettersim-ot-rsa

Contributions

Contributions are described according to the Contributor Role Taxonomy (CRediT).

JET: Conceptualisation, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Software, Validation, Visualisation, Writing (original draft)

RS: Data curation, Formal analysis, Investigation, Methodology, Software, Visualisation, Writing (review & editing)

CI: Formal analysis, Software, Writing (review & editing)

CJF: Conceptualisation, Methodology, Supervision, Writing (review & editing)

References

Agrawal, A., & Dehaene, S. (2023, November 7). Dissecting the neuronal mechanisms of invariant word recognition. https://doi.org/10.1101/2023.11.06.565812

Allen, P. A., & Emerson, P. L. (1991). Holism revisited: Evidence for parallel independent word-level and letter-level processors during word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 489–511. https://doi.org/10.1037/0096-1523.17.2.489

Allison, T., Mccarthy, G., Nobre, A., Puce, A., & Belger, A. (1994). Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral Cortex*, *4*(5), 544–554. https://doi.org/10.1093/cercor/4.5.544

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461. https://doi.org/10.3758/BF03214334

Antonello, R., & Huth, A. (2024). Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, *5*(1), 64–79. https://doi.org/10.1162/nol_a_00087

Ayzenberg, V., Kamps, F. S., Dilks, D. D., & Lourenco, S. F. (2022). Skeletal representations of shape in the human visual cortex. *Neuropsychologia*, *164*, 108092. https://doi.org/10.1016/j.neuropsychologia.2021.108092

Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, *9*(1), 9359. https://doi.org/10.1038/s41598-019-45268-y

Bai, J., Shi, J., Jiang, Y., He, S., & Weng, X. (2011). Chinese and korean characters engage the same visual word form area in proficient early chinese-korean bilinguals. *PLOS ONE*, *6*(7), e22765. https://doi.org/10.1371/journal.pone.0022765

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: Time course and scalp distribution. *Journal of Cognitive Neuroscience*, *11*(3), 235–260. https://doi.org/10.1162/089892999563373

Berio, D., Leymarie, F. F., Asente, P., & Echevarria, J. (2022). StrokeStyles: Stroke-based segmentation and stylization of fonts. *ACM Transactions on Graphics*, *41*(3), 28:1–28:21. https://doi.org/10.1145/3505246

Bolger, D. J., Perfetti, C. A., & Schneider, W. (2005). Cross-cultural effect on the brain revisited: Universal structures plus writing system variation [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.20124]. *Human Brain Mapping*, *25*(1), 92–104. https://doi.org/10.1002/hbm.20124

Briechle, K., & Hanebeck, U. D. (2001). Template matching using fast normalized cross correlation. *Optical Pattern Recognition XII*, *4387*, 95–102. https://doi.org/10.1117/12.421129

Brossette, B., Grainger, J., Lété, B., & Dufau, S. (2022). On the relations between letter, word, and sentence-level processing during reading. *Scientific Reports*, *12*(1), 17735. https://doi.org/10.1038/s41598-022-22587-1

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, *11*(41), 63–65. https://www.jstor.org/stable/2247157

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 1–10. https://doi.org/10.1038/s42003-022-03036-1

Chang, Y.-N., Furber, S., & Welbourne, S. (2012). Modelling normal and impaired letter recognition: Implications for understanding pure alexic reading. *Neuropsychologia*, *50*(12), 2773–2788. https://doi.org/10.1016/j.neuropsychologia.2012.07.031

Chauncey, K., Holcomb, P. J., & Grainger, J. (2008). Effects of stimulus font and size on masked repetition priming. *Language and cognitive processes*, *23*(1), 183–200. https://doi.org/10.1080/01690960701579839

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Clark, A. (2024). Pillow (PIL fork) documentation. https://pillow.readthedocs.io/en/stable/reference/

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, *123*, 291–307. https://doi.org/https://doi.org/10.1093/brain/123.2.291

Cohen, L., Lehéricy, S., Chochon, F., Lemer, C., Rivaud, S., & Dehaene, S. (2002). Language-specific tuning of visual cortex? functional properties of the visual word form area. *Brain*, *125*(5), 1054–1069. https://doi.org/10.1093/brain/awf094

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Lawrence Erlbaum.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. https://doi.org/10.1037/0033-295x.108.1.204

Daube, C., Xu, T., Zhan, J., Webb, A., Ince, R. A., Garrod, O. G., & Schyns, P. G. (2021). Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns*, *2*(10), 100348. https://doi.org/10.1016/j.patter.2021.100348

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, *117*(3), 713–758. https://doi.org/10.1037/a0019738

Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J. B., Bihan, D. L., & Cohen, L. (2004). Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychological Science*, *15*(5), 307–313. https://doi.org/10.1111/j.0956-7976.2004.00674.x

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J., & Riviere, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature neuroscience*, *4*(7), 752–8. https://doi.org/10.1038/89551

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262. https://doi.org/10.1016/j.tics.2011.04.003

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, *9*(7), 335–341. https://doi.org/10.1016/j.tics.2005.05.004

Dufau, S., Grainger, J., & Holcomb, P. J. (2008). An ERP investigation of location invariance in masked repetition priming. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(2), 222–228. https://doi.org/10.3758/CABN.8.2.222

Dunn-Rankin, P., Leton, D. A., & Shelton, V. F. (1968). Congruency factors related to visual confusion of english letters. *Perceptual and Motor Skills*, *26*(2), 659–666. https://doi.org/10.2466/pms.1968.26.2.659

Eddy, M. D., & Holcomb, P. J. (2009). Electrophysiological evidence for size invariance in masked picture repetition priming. *Brain and Cognition*, *71*(3), 397–409. https://doi.org/10.1016/j.bandc.2009.05.006

Fernández-López, M., Perea, M., & Vergara-Martínez, M. (2022). On the time course of the tolerance of letter detectors to rotations: A masked priming ERP investigation.

Neuropsychologia, *172*, 108259. https://doi.org/10.1016/j.neuropsychologia.2022.108259

Fernández-López, M., Gómez, P., & Perea, M. (2023). Letter rotations: Through the magnifying glass and what evidence found there. *Language, Cognition and Neuroscience*, *38*(2), 127–138. https://doi.org/10.1080/23273798.2022.2093390

Finkbeiner, M., & Coltheart, M. (2009). Letter recognition: From perception to representation. *Cognitive Neuropsychology*, *26*(1), 1–6. https://doi.org/10.1080/02643290902905294

Fischer-Baum, S., Bruggemann, D., Gallego, I. F., Li, D. S. P., & Tamez, E. R. (2017). Decoding levels of representation in reading: A representational similarity approach. *Cortex*, *90*, 88–102. https://doi.org/10.1016/j.cortex.2017.02.017

Fiset, D., Blais, C., Arguin, M., Tadros, K., Éthier-Majcher, C., Bub, D., & Gosselin, F. (2009). The spatio-temporal dynamics of visual letter recognition. *Cognitive Neuropsychology*, *26*(1), 23–35. https://doi.org/10.1080/02643290802421160

Fiset, D., Blais, C., Éthier-Majcher, C., Arguin, M., Bub, D., & Gosselin, F. (2008). Features for identification of uppercase and lowercase letters. *Psychological Science*, *19*(11), 1161–1168. https://doi.org/10.1111/j.1467-9280.2008.02218.x

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T. H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., … Vayer, T. (2021). POT: Python optimal transport. *Journal of Machine Learning Research*, *22*(78), 1–8. Retrieved February 1, 2024, from http://jmlr.org/papers/v22/20-451.html

Fraga-González, G., Pleisch, G., Di Pietro, S. V., Neuenschwander, J., Walitza, S., Brandeis, D., Karipidis, I. I., & Brem, S. (2021). The rise and fall of rapid occipito-temporal sensitivity to letters: Transient specialization through elementary

school. *Developmental Cognitive Neuroscience*, *49*, 100958. https://doi.org/10.1016/j.dcn.2021.100958

Gilbert, G. K. (1884). Finley's tornado predictions. *American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches of Study (1884-1896)*, *1*(5), 166. Retrieved June 3, 2024, from https://www.proquest.com/docview/124374084/abstract/7CB940B1A8FC4999PQ/1

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Grainger, J. (2008). Cracking the orthographic code: An introduction. *Language and Cognitive Processes*, *23*(1), 1–35. https://doi.org/10.1080/01690960701578013

Grainger, J., & Jacobs, A. M. (1991). Masked constituent letter priming in an alphabetic decision task. *European Journal of Cognitive Psychology*, *3*(4), 413–434. https://doi.org/10.1080/09541449108406237

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience*, *7*, 1–13. https://doi.org/10.3389/fnins.2013.00267

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*(1), 187–203. https://doi.org/10.1016/s0896-6273(00)80832-6

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*(2), 213–227. https://doi.org/10.1007/s42113-022-00166-x

Gutiérrez-Sigut, E., Marcet, A., & Perea, M. (2019). Tracking the time course of letter visual-similarity effects during word recognition: A masked priming ERP

investigation. *Cognitive, Affective and Behavioral Neuroscience*, *19*(4), 966–984. https://doi.org/10.3758/s13415-019-00696-1

Hannagan, T., Agrawal, A., Cohen, L., & Dehaene, S. (2021). Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(46), 1–12. https://doi.org/10.1073/pnas.2104779118

Hannagan, T., & Grainger, J. (2013). The lazy visual word form area: Computational insights into location-sensitivity. *PLoS Computational Biology*, *9*(10), 1–12. https://doi.org/10.1371/journal.pcbi.1003250

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). Deep residual learning for image recognition. https://doi.org/10.48550/arXiv.1512.03385

He, T., Boudewyn, M. A., Kiat, J. E., Sagae, K., & Luck, S. J. (2022). Neural correlates of word representation vectors in natural language processing models: Evidence from representational similarity analysis of event-related brain potentials. *Psychophysiology*, *59*(3), e13976. https://doi.org/10.1111/psyp.13976

Healy, A. F., & Cunningham, T. F. (1992). A developmental evaluation of the role of word shape in word recognition. *Memory & Cognition*, *20*(2), 141–150. https://doi.org/10.3758/bf03197163

Heilbron, M., Richter, D., Ekman, M., Hagoort, P., & de Lange, F. P. (2020). Word contexts enhance the neural representation of individual letters in early visual

cortex. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-019-13996-4

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.). (2008). *Bayesian evaluation of informative hypotheses*. Springer. https://doi.org/10.1007/978-0-387-09612-4

Humphreys, G. W., Riddoch, M. J., & Quinlan, P. T. (1988). Cascade processes in picture identification. *Cognitive Neuropsychology*, *5*(1), 67–104. https://doi.org/10.1080/02643298808252927

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. https://doi.org/10.1152/jn.00394.2013

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, *73*(1), 218–226. https://doi.org/10.1152/jn.1995.73.1.218

Jaccard, P. (1901). Ètude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Sociètè Vaudoise des Sciences Naturelles*, *37*(142), 547–579. https://doi.org/10.5169/seals-266450

Jacobs, A. M., & Grainger, J. (1991). Automatic letter priming in an alphabetic decision task. *Perception & Psychophysics*, *49*(1), 43–52. https://doi.org/10.3758/BF03211615

Jakubiak, E. J., Perry, R. N., & Frisken, S. F. (2006). An improved representation for stroke-based fonts. *ACM SIGGRAPH 2006 Sketches*, 137–es. https://doi.org/10.1145/1179849.1180020

Jeffreys Harold. (1961). *Theory of probability* (3rd). Oxford University Press.

Jordan, T. R., Thomas, S. M., & Scott-Brown, K. C. (1999). The illusory-letters phenomenon: An illustration of graphemic restoration in visual word recognition. *Perception*, *28*(11), 1413–1416. https://doi.org/10.1068/p2919

Jordan, T. R., Sheen, M., AlJassmi, M. A., & Paterson, K. B. (2015). A new demonstration of the illusory letters phenomenon: Graphemic restoration in arabic word perception. *Perception*, *44*(2), 215–218. https://doi.org/10.1068/p7885

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kim, A. E., & Straková, J. (2012). Concurrent effects of lexical status and letter-rotation during early stage visual word recognition: Evidence from ERPs. *Brain Research*, *1468*, 52–62. https://doi.org/10.1016/j.brainres.2012.04.008

Knuth, D. E. (1979). Mathematical typography. *Bulletin of the American Mathematical Society*, *1*(2), 337–372. https://doi.org/10.1090/S0273-0979-1979-14598-1

Krafnick, A. J., Tan, L.-H., Flowers, D. L., Luetje, M. M., Napoliello, E. M., Siok, W.-T., Perfetti, C., & Eden, G. F. (2016). Chinese character and english word processing in children's ventral occipitotemporal cortex: fMRI evidence for script invariance. *NeuroImage*, *133*, 302–312. https://doi.org/10.1016/j.neuroimage.2016.03.021

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 1–28. https://doi.org/10.3389/neuro.06.004.2008

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018, September 4). CORnet: Modeling the neural mechanisms of core object recognition. https://doi.org/10.1101/408385

Kuchinke, L., Krause, B., Fritsch, N., & Briesemeister, B. B. (2014). A familiar font drives early emotional effects in word recognition. *Brain and Language*, *137*, 142–147. https://doi.org/10.1016/j.bandl.2014.08.007

Lally, C., & Rastle, K. (2023). Orthographic and feature-level contributions to letter identification. *Quarterly Journal of Experimental Psychology*, *76*(5), 1111–1119. https://doi.org/10.1177/17470218221106155

Lavidor, M., Babkoff, H., & Faust, M. (2001). Analysis of standard and non-standard visual word format in the two hemispheres. *Neuropsychologia*, *39*(4), 430–439. https://doi.org/10.1016/s0028-3932(00)00125-1

Laws, K. R., & Gale, T. M. (2002). Category-specific naming and the visual characteristics of line drawn stimuli. *Cortex*, *38*(1), 7–21. https://doi.org/10.1016/S0010-9452(08)70635-X

Lee, T. W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, *11*(2), 417–441. https://doi.org/10.1162/089976699300016719

Legge, G. E., & Bigelow, C. A. (2011). Does print size matter for reading? a review of findings from vision science and typography. *Journal of Vision*, *11*(5), 8. https://doi.org/10.1167/11.5.8

Li, A., Feitelberg, J., Saini, A. P., Höchenberger, R., & Scheltienne, M. (2022). MNE-ICALabel: Automatically annotating ICA components with ICLabel in python. *Journal of Open Source Software*, *7*(76), 4484. https://doi.org/10.21105/joss.04484

Li, X., Huang, L., Yao, P., & Hyönä, J. (2022). Universal and specific reading mechanisms across different writing systems. *Nature Reviews Psychology*, *1*(3), 133–144. https://doi.org/10.1038/s44159-022-00022-6

Lin, S. E., Chen, H. C., Zhao, J., Li, S., He, S., & Weng, X. C. (2011). Left-lateralized n170 response to unpronounceable pseudo but not false chinese characters - the key role of orthography. *Neuroscience*, *190*, 200–206. https://doi.org/10.1016/j.neuroscience.2011.05.071

Ling, S., Lee, A. C. H., Armstrong, B. C., & Nestor, A. (2019). How are visual words represented? insights from EEG-based visual word decoding, feature derivation

and image reconstruction. *Human Brain Mapping*, *40*(17), 5056–5068. https://doi.org/10.1002/hbm.24757

Lochy, A., Jacques, C., Maillard, L., Colnat-Coulbois, S., Rossion, B., & Jonas, J. (2018). Selective visual representation of letters and words in the left ventral occipito-temporal cortex with intracerebral recordings. *Proceedings of the National Academy of Sciences*, *115*(32), E7595–E7604. https://doi.org/10.1073/pnas.1718987115

Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception & Psychophysics*, *80*(5), 1278–1289. https://doi.org/10.3758/s13414-017-1457-8

Macaya, M., & Perea, M. (2014). Does bold emphasis facilitate the process of visual-word recognition? *The Spanish Journal of Psychology*, *17*, E2. https://doi.org/10.1017/sjp.2014.2

Madec, S., Rey, A., Dufau, S., Klein, M., & Grainger, J. (2012). The time course of visual letter perception. *Journal of Cognitive Neuroscience*, *24*(7), 1645–1655. https://doi.org/10.1162/jocn_a_00178

Mano, O., Creamer, M. S., Badwan, B. A., & Clark, D. A. (2021). Predicting individual neuron responses with anatomically constrained task optimization. *Current Biology*, *31*(18), 4062–4075.e4. https://doi.org/10.1016/j.cub.2021.06.090

Marcet, A., & Perea, M. (2017). Is nevtral NEUTRAL? visual similarity effects in the early phases of written-word recognition. *Psychonomic Bulletin and Review*, *24*(4), 1180–1185. https://doi.org/10.3758/s13423-016-1180-9

Marcet, A., & Perea, M. (2018). Can i order a burger at rnacdonalds.com? visual similarity effects of multi-letter combinations at the early stages of word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(5), 699–706. https://doi.org/10.1037/xlm0000477

Marjieh, R., Sucholutsky, I., Sumers, T. R., Jacoby, N., & Griffiths, T. L. (2022, February 9). Predicting human similarity judgments using large language models. https://doi.org/10.48550/arXiv.2202.04728

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman; Company. https://doi.org/10.7551/mitpress/9780262514620.001.0001

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. an acccount of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Meletis, D., & Dürscheid, C. (2022). *Writing systems and their use: An overview of grapholinguistics*. De Gruyter. Retrieved June 2, 2024, from https://doi.org/10.1515/9783110757835

Mémoli, F. (2011). Gromovwasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, *11*(4), 417–487. https://doi.org/10.1007/s10208-011-9093-5

Nazir, T. A., Jacobs, A. M., & O'Regan, J. K. (1998). Letter legibility and visual word recognition. *Memory & Cognition*, *26*(4), 810–821. https://doi.org/10.3758/BF03211400

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLOS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Nobre, A. C., Allison, T., & McCarthy, G. (1994). Word recognition in the human inferior temporal lobe. *372*, 260–263. https://doi.org/10.1038/372260a0

Norris, D. (2013). Models of visual word recognition. *Trends in Cognitive Sciences*, *17*(10), 517–524. https://doi.org/10.1016/j.tics.2013.08.003

Op de Beeck, H. P., Torfs, K., & Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience*, *28*(40), 10111–10123. https://doi.org/10.1523/JNEUROSCI.2511-08.2008

Pang, Z., OMay, C. B., Choksi, B., & VanRullen, R. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks*, *144*, 164–175. https://doi.org/10.1016/j.neunet.2021.08.024

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature*, *423*(6941), 752–756. https://doi.org/10.1038/nature01516

Perea, M., Baciero, A., Labusch, M., Fernández-López, M., & Marcet, A. (2022). Are brand names special words? letter visual-similarity affects the identification of brand names, but not common words. *British Journal of Psychology*, *113*(3), 835–852. https://doi.org/10.1111/bjop.12557

Perea, M., Labusch, M., Fernández-López, M., Marcet, A., Gutierrez-Sigut, E., & Gómez, P. (2024). One more trip to barcetona: On the special status of visual similarity effects in city names. *Psychological Research*, *88*(1), 271–283. https://doi.org/10.1007/s00426-023-01839-3

Perea, M., Marcet, A., & Fernández-López, M. (2018). Does letter rotation slow down orthographic processing in word recognition? *Psychonomic Bulletin & Review*, *25*(6), 2295–2300. https://doi.org/10.3758/s13423-017-1428-z

Perea, M., & Rosa, E. (2002). Does whole-word shape play a role in visual word recognition? *Perception & Psychophysics*, *64*(5), 785–794. https://doi.org/10.3758/BF03194745

Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: Some implications of research on chinese for general theories of reading. *Psychological Review*, *112*(1), 43–59. https://doi.org/10.1037/0033-295X.112.1.43

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., & Oostenveld, R. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, *6*(1), 103. https://doi.org/10.1038/s41597-019-0104-8

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, *114*(2), 273–315. https://doi.org/10.1037/0033-295X.114.2.273

Petit, J.-P., Midgley, K. J., Holcomb, P. J., & Grainger, J. (2006). On the time course of letter perception: A masked priming ERP investigation. *Psychonomic Bulletin & Review*, *13*(4), 674–681. https://doi.org/10.3758/BF03193980

Peyré, G., & Cuturi, M. (2020, March 18). Computational optimal transport. https://doi.org/10.48550/arXiv.1803.00567

Pillsbury, W. B. (1897). A study in apperception. *The American Journal of Psychology*, *8*(3), 315–393. https://doi.org/10.2307/1411485

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Pleisch, G., Karipidis, I. I., Brem, A., Röthlisberger, M., Roth, A., Brandeis, D., Walitza, S., & Brem, S. (2019). Simultaneous EEG and fMRI reveals stronger sensitivity to orthographic strings in the left occipito-temporal cortex of typical versus poor beginning readers. *Developmental Cognitive Neuroscience*, *40*, 1–13. https://doi.org/10.1016/j.dcn.2019.100717

Pogoncheff, G., Granley, J., & Beyeler, M. (2023). Explaining v1 properties with a biologically constrained deep learning architecture. *Advances in Neural Information Processing Systems*, *36*, 13908–13930. https://doi.org/10.48550/arXiv.2305.11275

Polk, T. A., & Farah, M. (2002). Functional MRI evidence for an abstract, not perceptual, word-form area. *Journal of Experimental Psychology: General*, *131*(1), 65–72. https://doi.org/10.1037/0096-3445.131.1.65

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062

Putnam, H. (1967). Psychological predicates. In W. H. Captain (Ed.), *Art, mind, and religion* (pp. 37–48, Vol. 1). University of Pittsburgh Press.

Qu, J., Pang, Y., Liu, X., Cao, Y., Huang, C., & Mei, L. (2022). Task modulates the orthographic and phonological representations in the bilateral ventral occipitotemporal cortex. *Brain Imaging and Behavior*, 1–13. https://doi.org/10.1007/s11682-022-00641-w

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269. https://doi.org/10.1523/JNEUROSCI.0388-18.2018

Rajalingham, R., Kar, K., Sanghavi, S., Dehaene, S., & DiCarlo, J. J. (2020). The inferior temporal cortex is a potential cortical precursor of orthographic processing in untrained monkeys. *Nature Communications*, *11*(1), 3886. https://doi.org/10.1038/s41467-020-17714-3

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, *81*(2), 275–280. https://doi.org/10.1037/h0027768

Rey, A., Dufau, S., Massol, S., & Grainger, J. (2009). Testing computational models of letter perception with item-level event-related potentials. *Cognitive Neuropsychology*, *26*(1), 7–22. https://doi.org/10.1080/09541440802176300

Rezec, O. (2013). Ein differenzierteres strukturmodell des deutschen schriftsystems. *Linguistische Berichte*, *2013*(234), 227–254. https://doi.org/10.46771/2366077500234_4

Rocabado, F., Labusch, M., Perea, M., & Duñabeitia, J. A. (2024). Dissociating the effects of visual similarity for brand names and common words. *Journal of Cognition*, *7*(1). https://doi.org/10.5334/joc.397

Rosa, E., Perea, M., & Enneson, P. (2016). The role of letter features in visual-word recognition: Evidence from a delayed segment technique. *Acta Psychologica*, *169*, 133–142. https://doi.org/10.1016/j.actpsy.2016.05.016

Rousselet, G. A. (2012). Does filtering preclude us from studying ERP time-courses? *Frontiers in Psychology*, *3*, 1–9. https://doi.org/10.3389/fpsyg.2012.00131

Rumelhart, D. E., & Siple, P. (1974). Process of recognizing tachistoscopically presented words. *Psychological Review*, *81*(2), 99–118. https://doi.org/10.1037/h0036117

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015, January 29). ImageNet large scale visual recognition challenge. https://doi.org/10.48550/arXiv.1409.0575

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. https://doi.org/10.1523/JNEUROSCI.0179-10.2010

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, *139*, 101527. https://doi.org/10.1016/j.cogpsych.2022.101527

Schäfer, J., Opgen-Rhein, R., Zuber, V., Ahdesmaki, M., Silva, A. P. D., & Strimmer, K. (2021). *Corpcor: Efficient estimation of covariance and (partial) correlation*. https://CRAN.R-project.org/package=corpcor

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a bayes factor? *Psychological Methods*, *28*(3), 705–718. https://doi.org/10.1037/met0000421

Schyns, P. G., Snoek, L., & Daube, C. (2022). Degrees of algorithmic equivalence between the brain and its DNN models. *Trends in Cognitive Sciences*, *26*(12), 1090–1102. https://doi.org/10.1016/j.tics.2022.09.003

Share, D. L. (2008). On the anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, *134*(4), 584–615. https://doi.org/10.1037/0033-2909.134.4.584

Simon, G., Petit, L., Bernard, C., & Rebaï, M. (2007). N170 ERPs could represent a logographic processing strategy in visual word recognition. *Behavioral and Brain Functions*, *3*(1), 21. https://doi.org/10.1186/1744-9081-3-21

STAN Development Team. (2024). Stan modeling language users guide and reference manual, 2.34. https://mc-stan.org

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., Tenenbaum, J. B., Collins, K. M., Hermann, K. L., Oktar, K., Greff, K., Hebart, M. N., Jacoby, N., Zhang, Q., Marjieh, R., … Griffiths, T. L. (2023, November 2). Getting aligned on representational alignment. https://doi.org/10.48550/arXiv.2310.13018

Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (CLD): A large-scale lexical database for simplified mandarin chinese. *Behavior Research Methods*, *50*(6), 2606–2629. https://doi.org/10.3758/s13428-018-1038-3

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*(8), 997–1009. https://doi.org/10.1111/psyp.12437

Taylor, J. S., Davis, M. H., & Rastle, K. (2019). Mapping visual symbols onto spoken language along the ventral visual stream. *Proceedings of the National Academy of*

Sciences of the United States of America, *116*(36), 17723–17728. https://doi.org/10.1073/pnas.1818575116

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? the effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

Thesen, T., McDonald, C. R., Carlson, C., Doyle, W., Cash, S., Sherfey, J., Felsovalyi, O., Girard, H., Barr, W., Devinsky, O., Kuzniecky, R., & Halgren, E. (2012). Sequential then interactive processing of letters and words in the left fusiform gyrus. *Nature Communications*, *3*, 1–8. https://doi.org/10.1038/ncomms2220

VanRullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Frontiers in Psychology*, *2*, 1–6. https://doi.org/10.3389/fpsyg.2011.00365

Veale, J. F. (2014). Edinburgh handedness inventory - short form: A revised version based on confirmatory factor analysis. *Laterality*, *19*(2), 164–177. https://doi.org/10.1080/1357650X.2013.783045

Vidal, C., Content, A., & Chetail, F. (2017). BACS: The brussels artificial character sets for studies in cognitive psychology and neuroscience. *Behavior Research Methods*, *49*(6), 2093–2112. https://doi.org/10.3758/s13428-016-0844-8

Villani, C. (2009). *Optimal transport: Old and new* (Vol. 338). Springer. https://doi.org/10.1007/978-3-540-71050-9

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, *55*(1), 143–156. https://doi.org/10.1016/j.neuron.2007.05.031

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology*, *1*(1), 59–85. https://doi.org/10.1016/0010-0285(70)90005-8

Wong, A. C. N., Gauthier, I., Woroch, B., DeBuse, C., & Curran, T. (2005). An early electrophysiological response associated with expertise in letter perception. *Cognitive, Affective & Behavioral Neuroscience*, *5*(3), 306–318. https://doi.org/10.3758/cabn.5.3.306

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2021). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, *5*(3), 389–398. https://doi.org/10.1038/s41562-020-00982-w

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, *12*(1), 2065. https://doi.org/10.1038/s41467-021-22244-7

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond coltheart's n: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Yoshihara, M., Nakayama, M., Junyi, X., & Hino, Y. (2024). Does rotation eliminate masked priming effects for japanese kanji words? *Cognition*, *246*, 105759. https://doi.org/10.1016/j.cognition.2024.105759

Zhou, Z., Vilis, T., & Strother, L. (2019). Functionally separable font-invariant and font-sensitive neural populations in occipitotemporal cortex. *Journal of Cognitive Neuroscience*, *31*(7), 1018–1029. https://doi.org/10.1162/jocn_a_01408

Zweig, E., & Pylkkänen, L. (2009). A visual m170 effect of morphological complexity. *Language and Cognitive Processes*, *24*(3), 412–439. https://doi.org/10.1080/01690960802180420