Deep latent variable joint cognitive modeling of neural signals and human behavior

Khuong Vo[a,*], Qinhua Jenny Sun[b], Michael D. Nunez[c], Joachim Vandekerckhove[b,d], and Ramesh Srinivasan[b,e]

This is the author final copy of:
Vo, K., Sun, J. Q., Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2024). Deep latent variable joint cognitive modeling of neural signals and human behavior. *NeuroImage, 291,* 120559.

As the field of computational cognitive neuroscience continues to expand and generate new theories, there is a growing need for more advanced methods to test the hypothesis of brain-behavior relationships. Recent progress in Bayesian cognitive modeling has enabled the combination of neural and behavioral models into a single unifying framework. However, these approaches require manual feature extraction, and lack the capability to discover previously unknown neural features in more complex data. Consequently, this would hinder the expressiveness of the models. To address these challenges, we propose a Neurocognitive Variational Autoencoder (NCVA) to conjoin high-dimensional EEG with a cognitive model in both generative and predictive modeling analyses. Importantly, our NCVA enables both the prediction of EEG signals given behavioral data and the estimation of cognitive model parameters from EEG signals. This novel approach can allow for a more comprehensive understanding of the triplet relationship between behavior, brain activity, and cognitive processes.

EEG | Decision making | Neurocognitive model | Drift-diffusion model | Variational Bayes | Deep learning | Latent-variable models

Current approaches to understanding brain function emphasize the search for statistical relationships between human behavior and individual physiological measures (EEG, fMRI, fNIRS, etc.; e.g. Itthipuripat et al., 2019). Behavioral measures, such as accuracy and speed of responses, reflect latent cognitive processes that underlie decision making that are not observed directly and must be inferred by cognitive models (Lee & Wagenmakers, 2014). An ongoing challenge in computational cognitive neuroscience research is formulating the link between brain activity and latent cognitive processes. Here, we present a novel approach that allows a theoretical account of the cognitive process of decision-making, and artificial neural networks to estimate a joint latent space to link cognitive parameters to both neural signals and behavioral measures. This joint latent space model is a valuable new framework for computational cognitive neuroscience, allowing for new forms of inference and hypothesis generation.

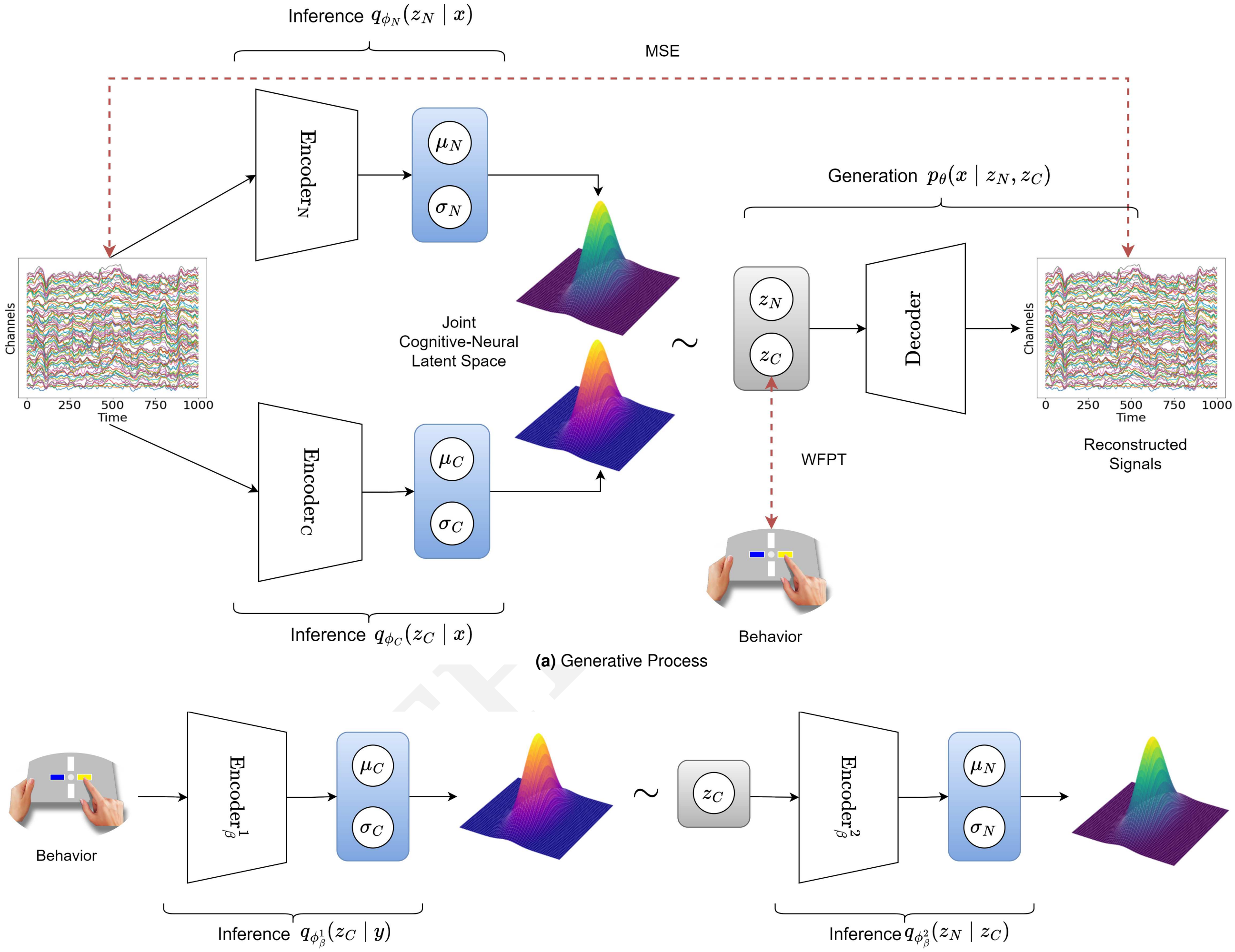
Previous work has focused on neurocognitive relationships between human neural data and behavioral data in decision-making tasks (Nunez et al., 2015, 2017, 2019; Lui et al., 2021; Turner et al., 2013, 2016). The hierarchical Bayesian models used in these projects make strong predictions about the relationships between brain activity and the speed of decision-making. These models typically make use of the drift-diffusion model (DDM; Ratcliff & McKoon, 2008), a widely-used cognitive model in decision-making, as their generative model of choice and reaction time data. To integrate neural signals, these models require knowledge of previously discovered features of the neural data (e.g., known functional signals in the cognitive neuroscience literature) that are then linked by prescribed (usually linear) relationships to the latent cognitive variables in a Bayesian hierarchical model. The resulting *neurocognitive* models test the relationship between neural signals and cognitive variables, and enhance the accuracy of predictions of behavior directly from brain signals (Turner et al., 2016; Nunez et al., 2017). This can be thought of as one domain of the larger field of *model-based cognitive neuroscience* (Forstmann & Wagenmakers, 2015).

A limitation of this approach is that we must know in advance which brain signals are possibly linked to cognitive functions. However, advances in frameworks and tools for neuroscience allow for the discovery of previously unknown neural features that we could use to explain latent cognitive variables. Ideally, such frameworks operate across observations, experimental manipulations, and individual differences. Deterministic models that leverage deep learning have been proposed for learning feature representation of EEG data to analyze and decode brain activity (Roy et al., 2019). As a notable example, Sun et al. (2022) have proposed a SincNet-based neural network that made use of EEG signals to learn the latent cognitive variables of the DDM on individual decisions. This approach identifies time windows of information processing and frequency bands that can be used to predict latent processes directly from EEG data as a trial-level association between neural features, choice, and response time.

This work aims to develop a deep probabilistic method for linking neural data from EEG to the latent parameters of

[a]Department of Computer Science, University of California, Irvine; [b]Department of Cognitive Sciences, University of California, Irvine; [c]Psychological Methods, University of Amsterdam; [d]Department of Statistics, University of California, Irvine; [e]Department of Biomedical Engineering, University of California, Irvine
[*]Correspondence concerning this article should be addressed to Khuong Vo (khuongav@uci.edu).

This work was supported by National Science Foundation grants #1658303, #1850849, #2051186, and #2126976.



(a) Generative Process

(b) Regularized Discriminative Process

Fig. 1. The Neurocognitive VAE. After the generative process (a) learns the joint latent neurocognitive variables (Section), the regularized discriminative process (b) retrofits its hierarchical latent space to the joint latent space (Section). Inference networks q and Generation networks p contain neural network parameters θ and ϕ. Black arrows: flows of operations. Red arrows: loss functions. MSE and WFPT stand for Mean Squared Error and Wiener First Passage Time, respectively. The heatmaps represent the probability distributions in the latent spaces. Plasma color maps are for the drift-diffusion variables ($z_C \in \mathbb{R}^3$), while greenery color maps are for residual neural variables ($z_N \in \mathbb{R}^{32}$). Blue blocks contain μ and σ, which are the parameters of the multivariate Gaussian latent spaces. Gray blocks contain z sampled ($\sim$) from the distributions. The variables x and y represent EEG signals and choice-RTs, respectively. Each trapezoid represents a different convolutional neural network (see Table 2 for detailed architectures).

a cognitive model. The innovation of our work lies in the use of a theoretical account of the cognitive process. This theoretical account drives the analysis of neural and behavioral measures. The framework allows for one-step, joint inference on integrative neurocognitive models that map EEG and behavior into a joint latent space. Uniquely, this new approach has the potential to allow us to generate task-relevant EEG signals from behavioral data, and *predict modulation of EEG signals by cognitive model parameters*. By combining the exploratory potential of modern latent variable methods with the theoretical appeal of human-interpretable cognitive model parameters, the proposed technique can be used to make predictions of brain signals and cognitive parameters in future experiments to test neurocognitive theories.

Neurocognitive Variational Autoencoders

Generative EEG Modeling with VAEs. Consider first a data set $\mathcal{P} \overset{\text{def}}{=} \{\mathcal{D}_1, \dots, \mathcal{D}_M\}$ containing M subjects, where each subject $\mathcal{D}_m \overset{\text{def}}{=} \{\mathbf{x}_1, \dots, \mathbf{x}_I\}$ consists of I trials $\mathbf{x}_i \in \mathbb{R}^{C \times T}$ that are EEG signals of C channels by T time samples. Throughout the paper, the subscript m is omitted when we refer to only one subject or when it is clear from the context.

For each subject m, we aim to learn an EEG generative process with a latent-variable model comprising of a fixed Gaussian prior over latent variables $p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the identity covariance matrix, and a parametric non-linear Gaussian likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$. The learning process finds θ such that the Kullback-Leibler (KL) divergence is minimized between the true data generating distribution

$p_{\mathcal{D}}$ and the model p_θ:

$$\begin{aligned} &\arg\min_\theta \mathrm{KL}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_\theta(\mathbf{x})\right) \\ &= \arg\max_\theta \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right] \end{aligned} \quad [1]$$

where $p_\theta(\mathbf{x}) = \int_{\mathcal{Z}} p_\theta(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z}$ is the likelihood of data point $\mathbf{x}$, approximated by averaging over the latent $\mathbf{z}$.

Nevertheless, estimating $p_\theta(\mathbf{x})$ is typically intractable. This issue can be mitigated by introducing a parametric inference model $q_\phi(\mathbf{z} \mid \mathbf{x})$ to construct a variational evidence lower bound on the log-likelihood $\log p_\theta(\mathbf{x})$ as follows:

$$\begin{aligned} &\mathcal{L}(\mathbf{x}; \theta, \phi) \\ &\quad \stackrel{\text{def}}{=} \log p_\theta(\mathbf{x}) - \mathrm{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z} \mid \mathbf{x})\right) \\ &\quad = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - \mathrm{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})\right) \end{aligned} \quad [2]$$

Taking the likelihood model $p_\theta(\mathbf{x} \mid \mathbf{z})$ to be a decoder and the inference model $q_\phi(\mathbf{z} \mid \mathbf{x})$ to be an encoder, a variational autoencoder (VAE; Kingma & Welling, 2013; Sohn et al., 2015) considers this objective from a deep probabilistic autoencoder perspective. Here, θ and ϕ are neural network parameters, and learning takes place via stochastic gradient ascent using unbiased estimates of $\nabla_{\theta,\phi} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(\mathbf{x}_i; \theta, \phi\right)$.

In the following sections, we extend the traditional VAE to create the Neurocognitive VAE (NCVA) (Figure 1). This model allows us to model a joint distribution of neural and behavioral data. Instead of a training technique that encourages disentanglement, as in β-VAE (Higgins et al., 2016), NCVA imposes restrictions on latent space by using a cognitive model that provides interpretability and controllable generation.

Disentangled Cognitive Latent Space of EEG. Now consider the data $\mathcal{D}_m \stackrel{\text{def}}{=} \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I)\}$, consisting, on the one hand, of N trials of the EEG data $\mathbf{x}_i$ and, on the other hand, of the corresponding choice response times (choice-RT) $\mathbf{y}_i$. Both $\mathbf{x}_i$ and $\mathbf{y}_i$ are associated with a context vector $\mathbf{c}_i$ (where the applicable context might be an experimental condition; say, noise conditions $\mathbf{c}_i$). For mathematical simplicity, the context vector $\mathbf{c}$ is not mentioned when we refer to one of the data modalities.

Crucially, we propose a generative model with two sources of variation: $\mathbf{z}_C$, which is cognitively specific, and $\mathbf{z}_N$, which captures any residual neural variations left in $\mathbf{x}$. We assume the approximate posterior $q_\phi(\mathbf{z}_N, \mathbf{z}_C \mid \mathbf{x})$ has the following fully factorized form:

$$\begin{aligned} &q_\phi\left(\mathbf{z}_N, \mathbf{z}_C \mid \mathbf{x}\right) = q_{\phi_N}\left(\mathbf{z}_N \mid \mathbf{x}\right) q_{\phi_C}\left(\mathbf{z}_C \mid \mathbf{x}\right) \\ &\quad q_{\phi_N}\left(\mathbf{z}_N \mid \mathbf{x}\right) = \mathcal{N}\left(\mathbf{z}_N \mid \boldsymbol{\mu}_{\phi_N}(\mathbf{x}), \text{diag}\left(\boldsymbol{\sigma}^2_{\phi_N}(\mathbf{x})\right)\right) \\ &\quad q_{\phi_C}\left(\mathbf{z}_C \mid \mathbf{x}\right) = \mathcal{N}\left(\mathbf{z}_C \mid \boldsymbol{\mu}_{\phi_D}(\mathbf{x}), \text{diag}\left(\boldsymbol{\sigma}^2_{\phi_D}(\mathbf{x})\right)\right) \end{aligned} \quad [3]$$

A Gaussian prior over latent variables $p(\mathbf{z}_C)$ can be chosen for each subject. We use subject priors obtained from a Bayesian hierarchical fitting of a DDM using the Markov chain Monte Carlo (MCMC) (Nunez et al., 2019).

We learn the generative model by maximizing the lower bound on $\log p_\theta(\mathbf{x}, \mathbf{y})$ as:

$$\begin{aligned} &\mathcal{L}(\mathbf{x}, \mathbf{y}; \theta, \phi_N, \phi_C) \\ &\ = \mathbb{E}_{q_\phi(\mathbf{z}_N, \mathbf{z}_C|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z}_N, \mathbf{z}_C) + \log p(\mathbf{y} \mid \mathbf{z}_C)\right] \\ &\quad - \mathrm{KL}\left(q_{\phi_N}(\mathbf{z}_N \mid \mathbf{x}) \| p(\mathbf{z}_N)\right) \\ &\quad - \mathrm{KL}\left(q_{\phi_C}(\mathbf{z}_C \mid \mathbf{x}) \| p(\mathbf{z_C})\right) \end{aligned} \quad [4]$$

where $p_\theta(\mathbf{x} \mid \mathbf{z}_N, \mathbf{z}_C) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_\theta(\mathbf{z}_N, \mathbf{z}_C), \mathbf{I})$ and $p(\mathbf{y}|\mathbf{z}_C)$ can be any neurocognitive likelihood. This work applies the Wiener First Passage Time distribution (WFPT; Navarro & Fuss, 2009) corresponding to the lower boundary:

$$\begin{aligned} &p(\mathbf{y}|\mathbf{z}_C) \\ &\ = \ \text{Wiener}\ (\text{RT} \mid \alpha, \tau, \delta) \\ &\ = \frac{\pi}{\alpha^2} e^{-\frac{1}{2}\left(\alpha\delta + \delta^2 (RT - \tau)\right)} \\ &\quad \times \sum_{k=1}^{+\infty} \left[k \sin\left(\frac{\pi k}{2}\right) e^{-\frac{k^2 \pi^2}{2\alpha^2}(RT - \tau)}\right] \end{aligned} \quad [5]$$

The probability at the upper boundary is obtained by setting $\delta' = -\delta$. $\mathbf{z}_C$ comprises of three parameters including drift rate δ, boundary α, non-decision time (ndt) τ. The bias towards correct or incorrect responses is fixed at 0.5, that is, the starting point is always unbiased.

The joint inference is performed using only EEG $\mathbf{x}$ to ensure that encoder θ_C would learn to extract neural features that are tailored to cognitive parameters, without relying on choice-RT $\mathbf{y}$. This has the advantage of providing more accurate trial-level parameter estimates that are associated with the EEG data.

Note that the dimension of the cognitive space is significantly lower than that of the residual neural space. This facilitates the representation of the variation in neural signals only through flexible $\mathbf{z}_N$. Maximizing the likelihood of observing neural signals does not guarantee decoder θ utilizing $\mathbf{z}_C$ to output $\mathbf{x}$. In the next section, we present an approach to capture the correlation between behavior and cognition, as well as the mapping of the variability of behavior and cognition to neural signals.

Structured EEG Modeling from Behavior. Here, we propose a discriminative model regularized by the generative model learned in the previous section. We aim to discriminatively learn the distribution of the cognitive parameters conditioned on behaviors, and the distribution of the neural latent variables conditioned on cognitive parameters. The joint latent space inferred from the behavior can be factorized into the two-level latent space as follows:

$$q_{\phi_B}\left(\mathbf{z}_N, \mathbf{z}_C \mid \mathbf{y}_i\right) = q_{\phi_B^2}\left(\mathbf{z}_N \mid \mathbf{z}_C\right) q_{\phi_B^1}\left(\mathbf{z}_C \mid \mathbf{y}_i\right) \quad [6]$$

Inspired by Suzuki et al. (2016), we learn the following approximations, w.r.t parameter ϕ_B^1:

$$\mathbb{E}_{p_{\mathcal{D}}}\left[\mathrm{KL}\left(q_{\phi_C}(\mathbf{z}_C \mid \mathbf{x}) \mid q_{\phi_B^1}(\mathbf{z}_C \mid \mathbf{y})\right)\right] \quad [7]$$

and w.r.t parameter ϕ_B^2:

$$\mathbb{E}_{p_{\mathcal{D}}}\left[\mathrm{KL}\left(q_{\phi_N}(\mathbf{z}_N \mid \mathbf{x}) \mid q_{\phi_B^2}(\mathbf{z}_N \mid \mathbf{z}_C)\right)\right] \qquad [8]$$

By decomposing the KL divergences as in Hoffman & Johnson (2016); Vedantam et al. (2017), we effectively minimize $\mathrm{KL}\left(q^{\text{avg}}_{\phi_C}(\mathbf{z}_C \mid \mathbf{x}) \mid q_{\phi_B^1}(\mathbf{z}_C \mid \mathbf{y})\right)$ and $\mathrm{KL}\left(q^{\text{avg}}_{\phi_N}(\mathbf{z}_N \mid \mathbf{x}) \mid q_{\phi_B^2}(\mathbf{z}_N \mid \mathbf{z}_C)\right)$, where $q^{\text{avg}}_{\phi}(\mathbf{z} \mid \mathbf{x}) = \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}\left[q_\phi(\mathbf{z} \mid \mathbf{x})\right]$. As there is little posterior uncertainty once conditioned on an EEG signal $\mathbf{x}_i$, the approximations are close to the average posterior induced by each of the EEG $\mathbf{x}_i$ associated with similar $\mathbf{y}$.

Having fit both the generative and discriminative models, we can now explore the three-way relationship between behavior, brain activity, and cognitive processes.

Experiments

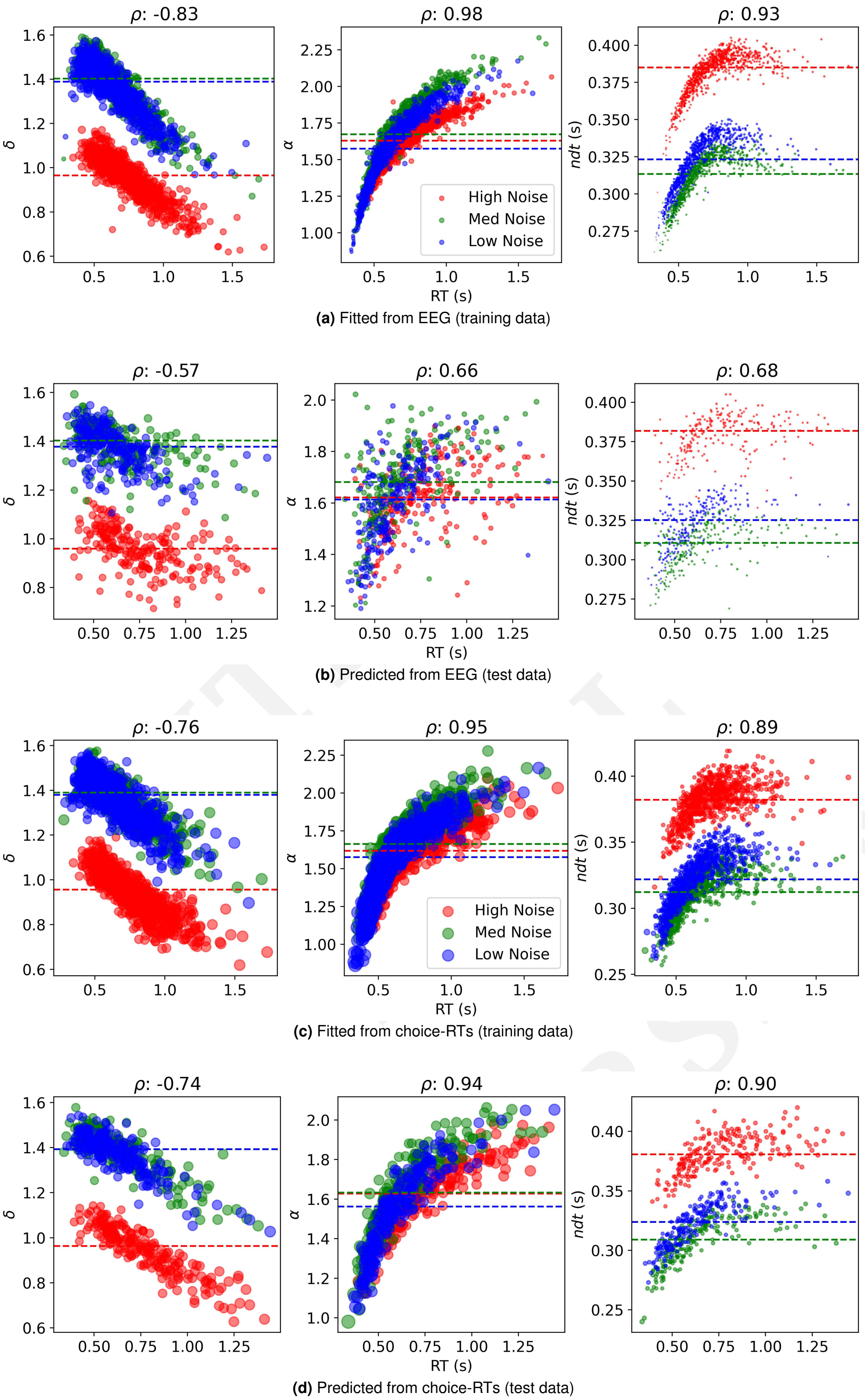
EEG and Behavioral Datasets. We used behavioral and EEG data collected while participants performed a two-alternative forced-choice task where they had to decide whether a Gabor patch presented with added dynamic noise is higher or lower spatial frequency (for details, see Experiment 2 by Nunez et al., 2019). Task difficulty was manipulated by adding spatial white noise to manipulate the quality of the perceptual evidence available to make the discrimination. The signal and the noise flickered at 40 and 30 Hz frequencies, respectively. 4 participants performed the task in blocks of trials at 3 added noise levels (low, medium, and high). Each subject performed approximately 3000 trials over 7 experimental sessions, while 128 channels of EEG and behavioral data were recorded. The independent component analysis (ICA)-based artifact rejection method was used on EEG data to remove eyeblinks, electrical noise, and muscle artifacts. A subset of 98 EEG channels were selected, excluding channels located in the outer ring. EEG data were bandpass filtered to 1 to 45 Hz in the frequency domain and then downsampled from 1000 Hz to 250 Hz in the time domain prior to data analysis. The data for each subject were divided into 80% for training and validation and the remaining 20% for testing.

Table 1. Comparison of the sum of Wiener negative log-likelihood ($-\sum \log$ Wiener $(\mathbf{RT}_i \mid \boldsymbol{\omega}_i)$) of four subjects on the test sets. $\bar{\boldsymbol{\omega}}$ represents the median fitted cognitive parameters from the training set.

Subjects	$\boldsymbol{\omega}_i^{\text{test}}$	$\bar{\boldsymbol{\omega}}^{\text{train}}$
s1	-0.018	0.212
s2	-0.244	0.159
s3	0.264	0.735
s4	0.031	0.230

Results. To validate the neurocognitive modeling approach, we first examine the trial-by-trial variability of the parameters within each subject and the generalization of the model to unseen data. Figures 2a and 2c show the trial-by-trial correlations between estimated DDM posteriors and observed choice-RTs in the training data from neural signals and behavior, respectively. Spearman correlations between fitted drift rates (δ) and choice-RTs are negatively strong. At the same time, there are strong positive correlations between boundaries (α) and choice-RTs, as well as between non-decision time and choice-RTs. The estimates in NCVA are regularized by the subject priors obtained from a Bayesian hierarchical fitting of a DDM using MCMC Nunez et al. (2019). The model was individually fitted for each subject using choice-RT and accuracy only and accounted for between-condition variability within subjects. Clear clusters of drift rates and non-decision-time estimates depending on the noise conditions can be seen, though boundary estimates are highly overlapped. It is worth noting that uncertainties in the estimates can be inspected from the figures through the posterior covariance. Understandably, the uncertainties in the estimations from choice-RTs are significantly higher than from EEG signals, which agree with the theoretical derivations in Section . Figures 2b and 2d also demonstrate a satisfactory generalization to unseen data. The drift rates positively correlate with choice-RTs, whereas the boundaries and non-decision time negatively correlate with choice-RTs. The model successfully learns to extract the neural features that account for the choice-RT variability at each trial. To evaluate whether obtaining trial estimates of cognitive parameters improved the model of choice and choice-RT data, Table 1 presents the Wiener likelihood test for the neurocognitive generalization ability to unseen data. The results show that the use of single-trial predictions of cognitive parameters $\boldsymbol{\omega}_i$ provides higher likelihood than the median estimates $\bar{\boldsymbol{\omega}}$ fitted from the training data. This implies that single-trial estimates better account for new data compared to median estimates.

Figure 3a shows the average of signals generated by the neurocognitive autoencoder when given a set of approximately 800 test choice-RTs compared to the average of actual signals associated with the same choice-RTs. At the selected electrodes, the window of interest is 100 ms pre-stimulus to 500 ms post-stimulus, which captures the N200 waveform. The generated and original signals appear visually similar in the timing and amplitudes of the peaks and troughs. Figures 3b, 3c, and 3d depict the trial-averaged frequency spectra and corresponding ERP waveforms of the reconstructed signals. Regarding the frequency spectra, the most important features are the 40 and 30 Hz peaks, which correspond to the flicker frequency of the signal (Gabor patch) and spatial white noise, respectively. Interestingly, the generative model learns to structure output the steady-state visually evoked potentials (SSVEPs) that occur in response to a visual stimulus flickering at different frequencies, even though it was never explicitly encoded in the model. Moreover, in the low noise condition (b), the 30 Hz peak is large and the 40 Hz is small, while in the high noise condition (d), the 30 Hz peak is reduced and the



(a) Fitted from EEG (training data)

(b) Predicted from EEG (test data)

(c) Fitted from choice-RTs (training data)

(d) Predicted from choice-RTs (test data)

Fig. 2. Drift-diffusion single-trial parameter estimations from correct responses of subject s1. The parameters are constrained by the subject priors resulting from a Bayesian MCMC modeling (without EEG data). Scatter plots illustrate the relationship between the parameters and the observed choice-RTs for each trial. The top two rows are posterior inferences from neural signals, while the bottom two are from behaviors. The left column shows the drift-rate (δ) estimates, the middle column shows boundary (α) estimates, and the right column presents non-decision time (ndt) estimates. The correlations between the choice-RTs and the inferred DDM parameters are consistent with what is expected. On top of each panel are the Spearman correlation coefficients (ρ). The covariances of the inferred parameters are indicated by circles, which correspond to contours having one standard deviation. For clarity, each circle is magnified 300 times.

40 Hz peak is enhanced. In terms of ERP waveforms, the model captures the relationships of the N200 peak latencies with respect to the additive noise conditions. Higher additive white noise in the stimulus effectively increases the latency and decreases the amplitude of the N200. We focus on the N200 signal because the original study (Nunez et al., 2019) found strong relationships between N200 latency and choice-RT, and thus the N200 is a good validation of our model. These prove the convergence of the model in optimizing the lower bound of the conditional likelihood mapping from behavioral data to EEG features, which effectively encodes differences in the stimuli presented to the subjects in the latent variable space.

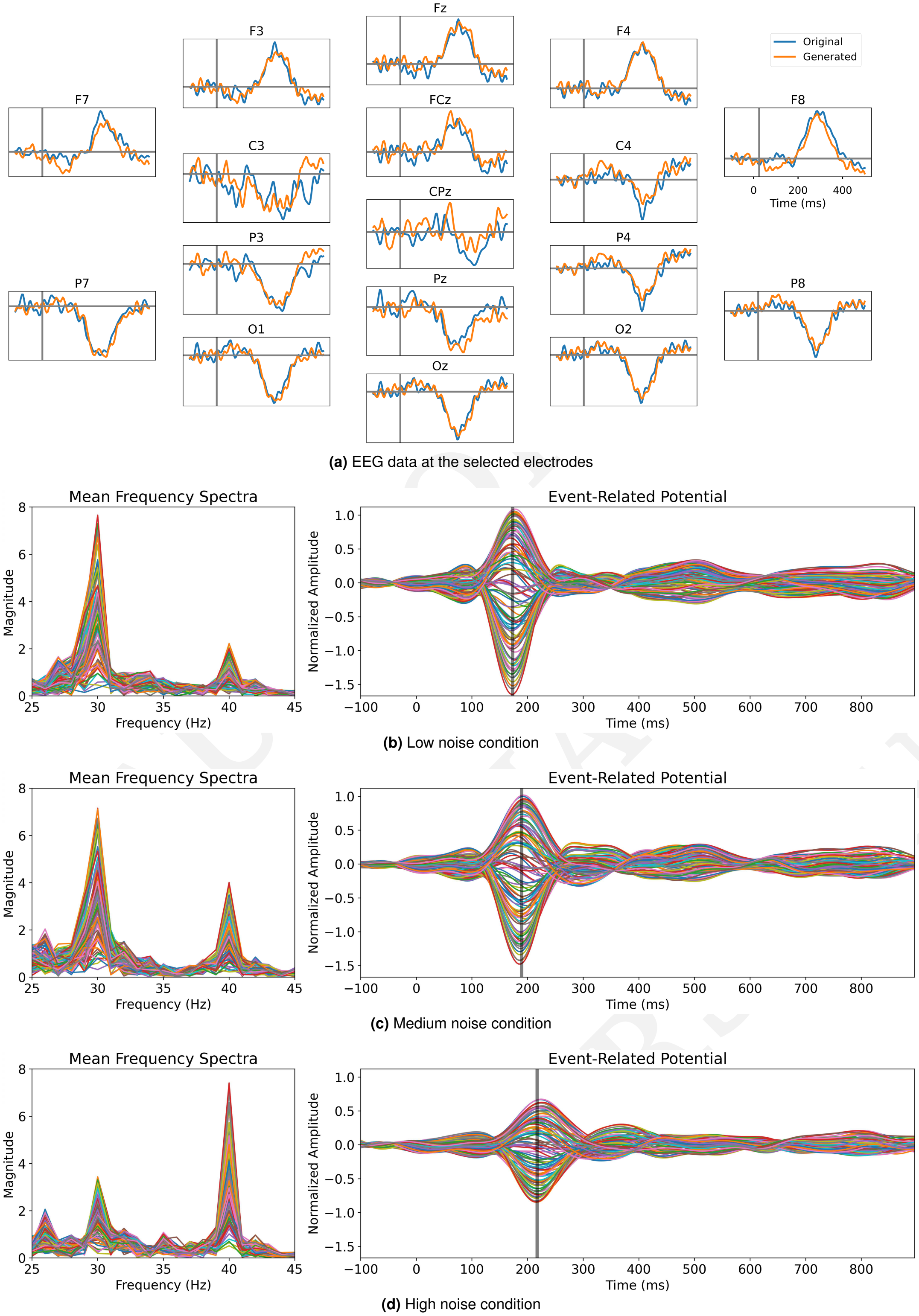
In addition to evaluating traditional ERP estimates (trial-averaged), we also assess the single-trial ERP estimate (channel-averaged). To increase the signal-to-noise ratio to better detect the N200, the first singular-value decomposition (SVD) component obtained from the ERP response is taken as a channel weighting function. More details of the SVD method can be seen at (Nunez et al., 2019). Figure 4 shows the performance of the model in learning the N200 feature in each trial. As shown in Figure 4, the distributions of the single-trial N200 peak latencies, as well as the amplitudes calculated from the generated signals, closely match those of the original signals at three different noise levels. The peak amplitude distribution is somewhat broader than the original data's generated distribution. Importantly, the model can generate the variability of the N200 latency with the experimental manipulation of low, medium, and high noise, systematically increasing the N200 latency in the generated signals.

Figure 5 represents the sensitivity analysis of the choice-RT and drift-diffusion parameters regardless of the noise conditions. In the left column, we examine the sensitivity of the neural signals generated by the choice-RTs. We can see similar patterns across subjects where the increases in choice-RTs lead to significant declines in the 30 Hz and the rises of the N200 latencies. This confirms the minimization approach of the KL divergence between the latent spaces inferred from the behavioral data and the neural signals. Power at 40 Hz reflecting the neural response to the noise also changes according to the choice-RTs, though the pattern is not as strong as the subjects suppressed the noise signal in all conditions.

One of the powerful tools for exploring the relationship between cognitive processes is to examine the sensitivity of neural signals to cognitive parameters. The middle and right columns of Figure 5 depict the effect of *hypothetical* modulations of drift rates and non-decision time on the generated neural signals. The results show that our model reveals the intricate interactions between cognitive parameters and neural signals, which is consistent with prior discoveries in the cognitive modeling literature. As the non-decision time is faster, the N200 latencies are shorter, and the 30 Hz peaks are larger. Accordingly, the amplitudes of the N200 peaks are more prominent, though not shown in the figures for clarity. The same interactions are observed with the increase in drift rates, representing evidence accumulation. Again, the effects on 40 Hz peaks are weaker and depend on the subjects. We did not observe the effects of the boundary separation (caution) on the neural signals. The effect can be reversed with slower non-decision times and lower drift rates. The strongest effects can be seen when both parameters influence neural signals. This demonstrates the effectiveness of the designs of the hierarchical latent variables inferred from choice-RTs and the disentangled latent space produced by the EEG data.

Conclusion

In this work, we proposed a joint behavioral and EEG modeling approach driven by a cognitive model of decision making. The experimental results demonstrate the effectiveness of our Neurocognitive VAE in simultaneously modeling high-dimensional EEG signals and low-dimensional behavioral data. Remarkably, the model learns essential task-relevant neural features, e.g. N200 peaks and SSVEP, without explicit specification in the optimization objective. Furthermore, the model captures how these features modulate behavior, specifically discovering relationships between brain activity and behavior consistent with other models based on prior knowledge. This suggests that the Neurocognitive VAE helps uncover neural signals linked to behavioral data by mapping to a structured latent space. Compared to the aforementioned published joint models (Nunez et al., 2015, 2017, 2019; Lui et al., 2021; Turner et al., 2013, 2016), our end-to-end model is capable of inferring task-relevant EEG features from behavior without prior knowledge of which features to optimize. The structured latent space allows the learning of behavioral variability to drive the EEG data generation process, leading to the prediction of the structure of EEG features in relation to the stimuli used in the experiments (N200 and SSVEP) and the behavioral performance (choice-RT). In addition, the model allows us to directly map the variability of cognitive parameters to neural signals, allowing for theoretical predictions that guide future experimental studies. It should be noted that our framework does not serve to refine the functional form of process-oriented computational models. Instead, it presumes a set of fixed assumptions; in the DDM, a constant drift rate and boundary separation within trials. Importantly, our framework can be generalized to encompass any other neural measures combined with any cognitive model to explain behavior, provided that the cognitive model expresses a closed-form likelihood of behavioral data. Importantly, by parameterizing the likelihood by a deep neural network receiving neural data as input, trial-level parameter inferences are made possible. In this research, we assume a DDM posterior with a diagonal covariance matrix. This could lead to an overestimation of the variance of the marginal posteriors if the true posterior has dependencies. It would be beneficial to investigate the use of a full covariance matrix as an alternative. It is



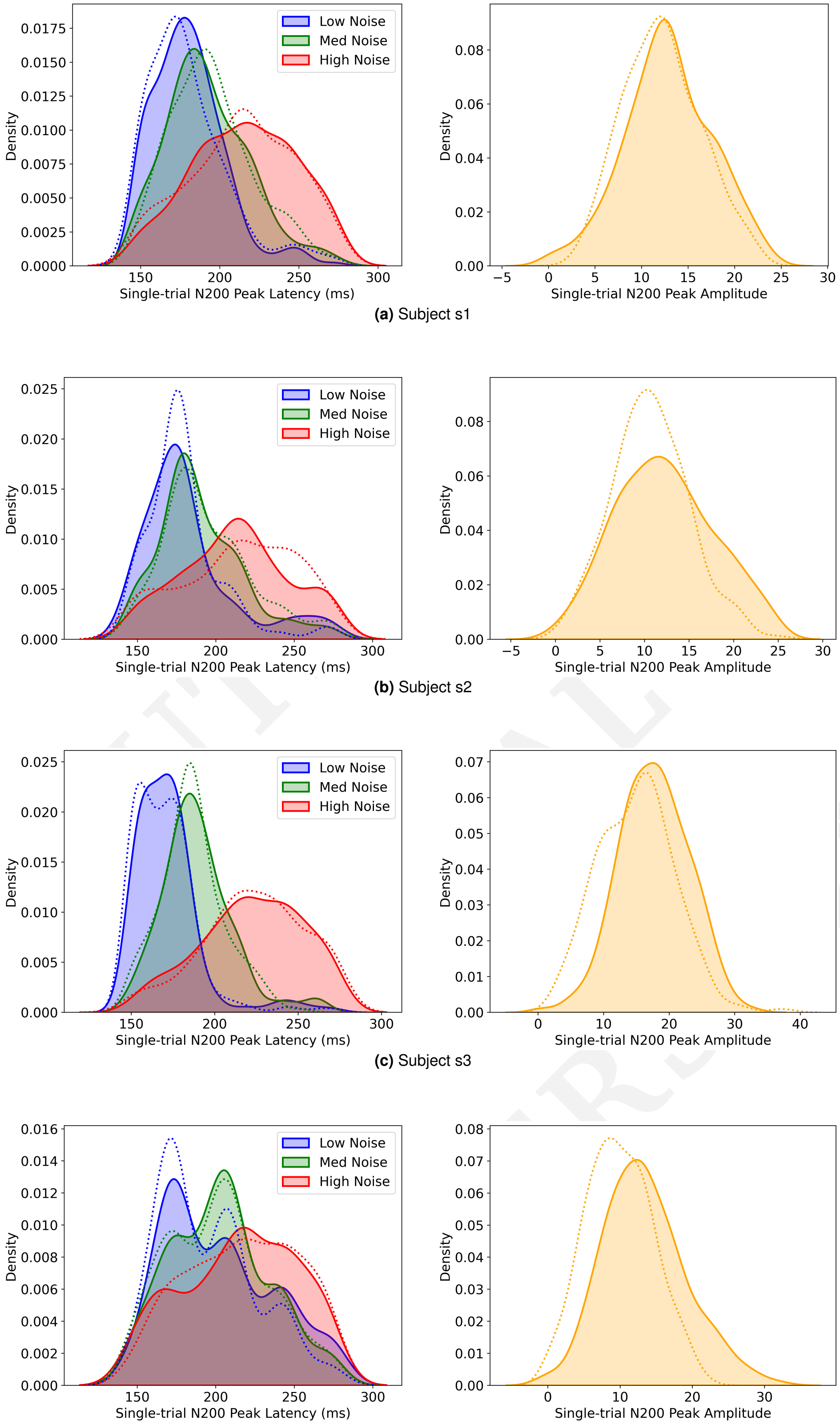
(a) EEG data at the selected electrodes

(b) Low noise condition

(c) Medium noise condition

(d) High noise condition

Fig. 3. Performance of the model in reconstructing 98 EEG channels of subject s1 by averaging $\approx$ 800 predicted EEG trials from $\approx$ 800 choice-RTs in the test set. Time point zero denotes the time point of stimulus onset. The first row displays the original (blue) and generated (orange) trial-averaged EEG data at the pooled electrodes. The x-axis denotes the time in milliseconds from stimulus onset, and the y-axis denotes the signal amplitude. The second, third, and fourth rows are (left) frequency spectra and (right) EEG signals averaged over all test choice-RT trials ($\approx 800/3$ per condition). The signals on the right are low-pass filtered at 15 Hz for clarity of N200 peaks. Each colored line corresponds to one reconstructed EEG channel. In low-noise conditions, the spectra show a strong peak at the Gabor flicker frequency of 30 Hz, and the ERP waveform shows a shorter N200 latency and larger peak amplitude. Under high-noise conditions, the spectra show a strong peak at the noise flicker frequency of 40 Hz, and the ERP waveform shows a longer N200 latency and a smaller peak amplitude.



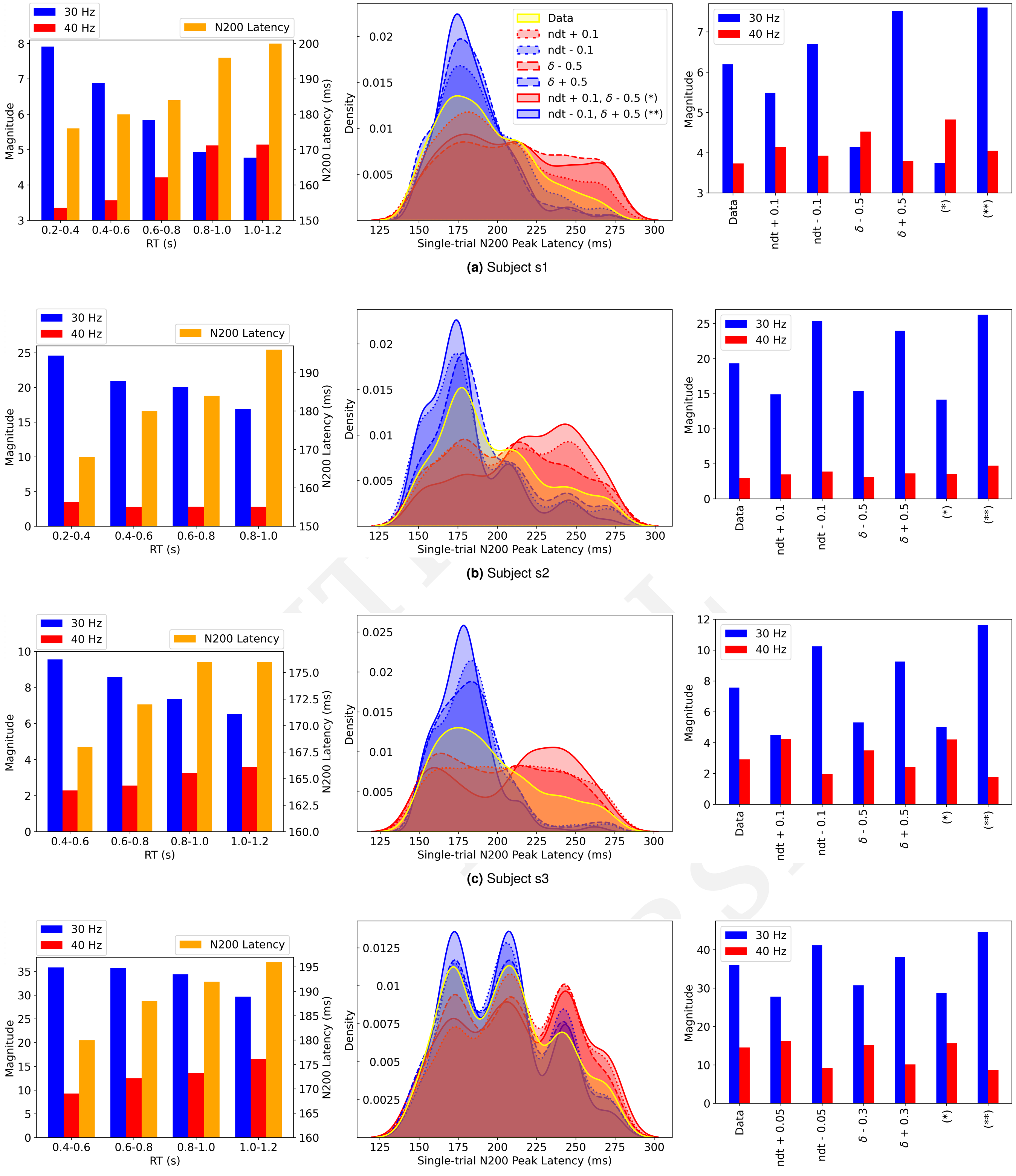
(a) Subject s1

(b) Subject s2

(c) Subject s3

(d) Subject s4

Fig. 4. Performance of the model in reconstructing single-trial N200 peaks from choice-RTs in four subjects. The dotted lines are references to the original data. The distributions of (left) single-trial N200 peak latencies across three noise conditions and (right) the N200 peak amplitude statistics are shown. Single-trial observations of the peak latency of N200 are found using the SVD method (Nunez et al., 2019) for each subject and noise condition.



(a) Subject s1

(b) Subject s2

(c) Subject s3

(d) Subject s4

Fig. 5. Sensitivity analysis of choice-RTs and latent drift-diffusion parameters on EEG signal generation in four subjects. The left column presents the effects of choice-RTs on the output neural signals. The blue bars represent the power at 30 Hz, while the red bars represent the power at 40 Hz. The orange bars show the N200 latencies. The middle column shows the changes in the single-trial N200 distribution w.r.t to *hypothetical* changes in the cognitive parameters. The yellow distribution represents the reference data, while the blue and red ones correspond to modified parameter settings that decrease or increase the N200 latencies, respectively. The modification in subject s4 (ndt ± 0.05, δ± 0.3) is different from other subjects. The right column characterizes the changes in 30 Hz and 40 Hz peaks w.r.t to the changes in the same cognitive parameters.

important to mention that our validation process focused on correct responses. Due to the low number of incorrect responses compared to correct ones, we lack confidence in interpreting the results in this study for the incorrect trials, although the direction of the trial-level parameter fits was consistent with the results for correct trials. We anticipate future research to explore strategies to address the class imbalance problem in deep learning models (Johnson & Khoshgoftaar, 2019). Further work with a larger dataset is needed to demonstrate that we can extend the model to new individuals. In principle, this would potentially allow us to predict brain activity in clinical populations with known behavioral differences.

Acknowledgements

This research was supported by grants #1658303, #1850849, #2051186, and #2126976 from the United States (US) National Science Foundation (NSF). We thank Erik Sudderth for his insight into the NVCA. All reviewers are appreciated for their constructive suggestions and comments.

Ethics Statement

All participants gave written informed consent, and all data was collected at the University of California, Irvine with approval from the Institutional Review Board.

Data and Code Availability Statement

The dataset analyzed during the current study is available on `https://zenodo.org/record/8381751`, and the implementation of the model is in the following repository `https://github.com/khuongav/neurocognitive_vae`.

Author Contributions

Khuong Vo: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Jenny Quinhua Sun**: Data curation, Methodology, Visualization, Writing - review & editing. **Michael D. Nunez**: Data curation, Writing - review & editing. **Joachim Vandekerckhove**: Conceptualization, Formal analysis, Methodology, Writing - review & editing. **Ramesh Srinivasan**: Conceptualization, Formal analysis, Methodology, Writing - original draft, Writing - review & editing.

Appendix

Neural Network Architectures and Training Hyperparameters. The inferential and generative processes are parameterized by deep neural networks, as shown by the flows in Figure 1. Table 2 details the architectures of the five networks. The input EEG signals are of size 98 x 250 (1 second of data of 98 channels at 250 Hz). The feature extraction layers in the EEG and cognitive encoders are similar to Vo et al. (2022). All the feature maps have 128 channels. Leaky ReLU (lReLU) activation functions are applied to all layers, with a slope of 0.1 to stimulate easier gradient flow. Batch normalizations (BN) (Ioffe & Szegedy, 2015) are used in each convolutional layer of the encoders and decoders. Self-attention layers (Zhang et al., 2019) are applied in the encoders and decoders to better account for long-range relationships in time series. $\mathbf{c}$ are noise condition embeddings as one-hot vectors (size 3). The size of $\mathbf{z}_N$ is set at 32 as increasing the dimension did not lead to any improvement in performance on a validation set.

In Equation (4), the term $\log p(\mathbf{y} \mid \mathbf{z}_C)$ is weighted by $\lambda = 2$ to scale up the likelihood of low-dimensional behavior. The KL terms are weighted by $\beta = 20$. The KL terms are normalized to balance the KL divergence loss and the reconstruction loss. Please refer to Sections 4.2 and A6 of (Higgins et al., 2016) for further information. The optimization of $q_{\phi_C}(\mathbf{z}_C \mid \mathbf{x})$ is divided into two stages. We first optimize the network w.r.t drift rate δ and boundary α, while non-decision time τ is set to $0.93 \cdot RT_{min}$ for each subject, approximating the results of the Bayesian MCMC modeling Nunez et al. (2019). Having trained ϕ_C for δ and α, we can proceed to train only the last fully connected layer that predicts τ. This procedure is to circumvent the difficulty of simultaneously optimizing the network for the boundary and the non-decision time on the experimental data. We used Adam (Kingma & Ba, 2014) for optimizations, with a learning rate of 5e-4 and exponential decay rates β_1 = 0.9 and β_2 = 0.999.

Simulation Studies. We assessed our ability to recover true non-decision time (NDT) and drift rate by simulating response time data and EEG signals. Response time data were simulated from a drift-diffusion model with trial-to-trial variability in NDT and evidence accumulation rate (i.e., drift rate). To simulate EEG signals with a known relationship with DDM parameters, we specifically focused on N200 due to the significant associations between N200 latency and NDT reported by Nunez et al. (2019). In our new experiments, we additionally observed a substantial relationship between drift rate and N200 latency, which we included in the simulation. Boundary separation was not included in the simulation, as we did not find any neural correlates of variability in boundary separation, and those are usually only found in tasks with trial-level accuracy feedback (Cavanagh & Frank, 2014; Nunez et al., 2024).

To simulate single-trial EEG signals, we shifted the true averaged ERP waveform based on each sample of trial-level NDT, using a linear regression slope of 1, as in Nunez et al. (2019). EEG noise was obtained from the original data, using independently sampled segments that did not include responses to stimuli. The resulting ERP and EEG waveforms were then combined to generate artificial EEG signals for each trial that carried the N200 latency information and was associated with choice and response time.

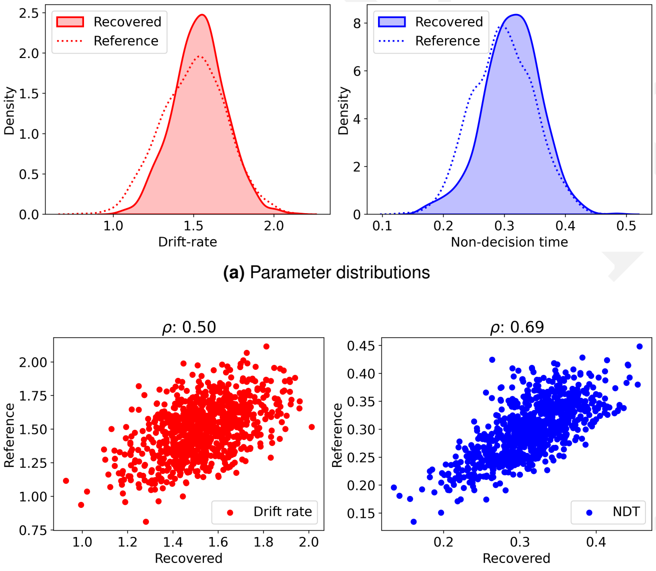
It is evident from the results in Figure 6 that the model can

Table 2. Neural network parametrization

Encoder$_N$ $- q_{\phi_N}(\mathbf{z}_N \mid \mathbf{x})$ **maps EEG signals to neural latents**		**Encoder$_C$** $- q_{\phi_C}(\mathbf{z}_C \mid \mathbf{x})$ **maps EEG signals to cognitive latents**	**Decoder -** $p_\theta(\mathbf{x} \mid \mathbf{z}_N, \mathbf{z}_C)$ **reconstructs EEG signals**
Dropout(0.3) Conv 1, lReLU, 128 x 250		Conv 1, lReLU, 128 x 250	Get $\mathbf{z}_C$ Linear 128, lReLU
Conv 6, BN, lReLU Conv 6, Stride 2, BN, lReLU	X 2	Conv 6, BN, lReLU, Dropout(0.7) Conv 6, Stride 2, BN, lReLU, Dropout(0.7)	Linear 32, lReLU Concat $\mathbf{z}_N$, $\mathbf{c}$
Self Attention		Self Attention	Conv Transp 8, Stride 4, 512 Channels, BN, lReLU
Conv 6, BN, lReLU Conv 6, Stride 2, BN, lReLU	X 2	Conv 6, BN, lReLU, Dropout(0.7) Conv 6, Stride 2, BN, lReLU, Dropout(0.7)	Conv Transp 8, Stride 4, 256 Channels, BN, lReLU Self Attention
Reshape 2048, Concat $\mathbf{c}$ Linear 32 (mean $\mathbf{z}_N$) Linear 32 (logvar $\mathbf{z}_N$)		Reshape 2048, Concat $\mathbf{c}$ Linear 1 (mean δ), Linear 1 (logvar δ) Linear 1, Softplus (mean α) Linear 1 (logvar α) Linear 1, Softplus (mean ndt) Linear 1 (logvar ndt)	Conv Transp 6, Stride 3, 128 Channels, BN, lReLU Conv Transp 6, Stride 3, 128 Channels, BN, lReLU Self Attention Conv Transp 10, Stride 2, 98 Channels

Encoder$^2_\beta$ $- q^2_\beta(\mathbf{z}_N \mid \mathbf{z}_C)$ **maps cognitive latents to neural latents**		**Encoder$^1_\beta$** $- q^1_\beta(\mathbf{z}_C \mid \mathbf{y}_i)$ **maps behaviors to cognitive latents**	
Linear 128, lReLU Linear 128, lReLU Concat $\mathbf{c}$ Linear 64		Linear 128, lReLU Linear 128, lReLU Concat $\mathbf{c}$ Linear 6	

accurately recover the original distributions of trial-specific parameters. In particular, the generating and recovered distributions strongly overlap, and the correlation plots indicate that our single-trial estimates of cognitive parameters exhibit good correlations with the reference parameters.



(a) Parameter distributions

(b) Parameter correlations

Fig. 6. Drift-diffusion parameter estimates from neural signals in a simulation of trial-level choice RTs and EEG signals. The top panels show the overlap between the recovered and the original distributions of trial-specific drift-rate and NDT. The reference values for the drift rate and NDT are drawn from the normal distributions $\mathcal{N}(1.5, 0.2)$ and $\mathcal{N}(0.3, 0.05)$, respectively. The bottom scatter plots illustrate the relationship between the recovered parameters and the original parameters each trial. ρ are the Spearman correlation coefficients.

References

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*, *18*(8), 414–421.

Forstmann, B. U., & Wagenmakers, E.-J. (Eds.). (2015). *An Introduction to Model-Based Cognitive Neuroscience*. New York, NY: Springer New York. doi:

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations.*

Hoffman, M. D., & Johnson, M. J. (2016). Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in advances in approximate bayesian inference, nips* (Vol. 1).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Itthipuripat, S., Sprague, T. C., & Serences, J. T. (2019). Functional MRI and EEG index complementary attentional modulations. *Journal of Neuroscience*, *39*(31), 6162–6179.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1–54.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lui, K. K., Nunez, M. D., Cassidy, J. M., Vandekerckhove, J., Cramer, S. C., & Srinivasan, R. (2021). Timing of readiness potentials reflect a decision-making process in the human brain. *Computational Brain & Behavior*, *4*(3), 264–283.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in wiener diffusion models. *Journal of mathematical psychology*, *53*(4), 222–230.

Nunez, M. D., Fernandez, K., Srinivasan, R., & Vandekerckhove, J. (2024). A tutorial on fitting joint models of m/eeg and behavior to understand cognition. *PsyArXiv*. doi:

Nunez, M. D., Gosai, A., Vandekerckhove, J., & Srinivasan, R. (2019). The latency of a visual evoked potential tracks the onset of decision making. *Neuroimage*, *197*, 93–108.

Nunez, M. D., Srinivasan, R., & Vandekerckhove, J. (2015). Individual differences in attention influence perceptual decision making. *Frontiers in psychology*, *8*, 18.

Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of mathematical psychology*, *76*, 117–130.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, *16*(5), 051001.

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, *28*.

Sun, Q. J., Vo, K., Lui, K., Nunez, M., Vandekerckhove, J., & Srinivasan, R. (2022). Decision sincnet: Neurocognitive models of decision making that predict cognitive processes from neural signals. In *2022 international joint conference on neural networks (ijcnn)* (pp. 1–9).

Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206.

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage*, *128*, 96–115.

Vedantam, R., Fischer, I., Huang, J., & Murphy, K. (2017). Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*.

Vo, K., Vishwanath, M., Srinivasan, R., Dutt, N., & Cao, H. (2022). Composing graphical models with generative adversarial networks for EEG signal modeling. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1231–1235).

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354–7363).