eLife

Neuroscience

Revealing unexpected complex encoding but simple decoding mechanisms in motor cortex via separating behaviorally relevant neural signals

Yangang Li, Xinyun Zhu, Yu Qi, Yueming Wang

Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China • College of Computer Science and Technology, Zhejiang University, Hangzhou, China • The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China • Affliated Mental Health Center & Hangzhou Seventh People's Hospital and the MOE Frontier Science Center for Brain Science and Brain-machine Integration, Zhejiang University School of Medicine, Hangzhou, China • Zhejiang Brain-Computer Interface Institute, Hangzhou, China

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
February 2, 2024 (this version)

Reviewed preprint version 1
July 14, 2023

Posted to preprint server
May 9, 2023

Sent for peer review
May 8, 2023

Abstract

In motor cortex, behaviorally-relevant neural responses are entangled with irrelevant signals, which complicates the study of encoding and decoding mechanisms. It remains unclear whether behaviorally-irrelevant signals could conceal some critical truth. One solution is to accurately separate behaviorally-relevant and irrelevant signals, but this approach remains elusive due to the unknown ground truth of behaviorally-relevant signals. Therefore, we propose a framework to define, extract, and validate behaviorally-relevant signals. Analyzing separated signals in three monkeys performing different reaching tasks, we found neural responses previously considered useless encode rich behavioral information in complex nonlinear ways. These responses are critical for neuronal redundancy and reveal movement behaviors occupy a higher-dimensional neural space than previously expected. Surprisingly, when incorporating often-ignored neural dimensions, behavioral information can be decoded linearly as accurately as nonlinear decoding, suggesting linear readout is performed in motor cortex. Our findings prompt that separating behaviorally-relevant signals may help uncover more hidden cortical mechanisms.

eLife assessment

This study presents a **useful** method for the extraction of behaviour-related activity from neural population recordings based on a specific deep learning architecture - a variational autoencoder. However, the evidence supporting the scientific claims resulting from the application of this method is **incomplete** as the results may stem, in part, from its properties. The main limitations are: (1) benchmarking against comparable methods is limited; and (2) some observations may be a byproduct of their method, and may not constitute new scientific observations.

Introduction

Understanding how motor cortex encodes and decodes movement behaviors is a fundamental goal of neuroscience (***Kriegeskorte and Douglas, 2019***; ***Saxena and Cunningham, 2019***). Here, we define behaviors as behavioral variables of interest measured within a given task, such as arm kinematics during a motor control task; we employ terms like 'behaviorally-relevant' and 'behaviorally-irrelevant' only regarding such measured behavioral variables. However, achieving this goal faces significant challenges because behaviorally-relevant neural responses are entangled with behaviorally-irrelevant factors such as responses for other variables of no interest (***Fusi et al., 2016***; ***Rigotti et al., 2013***) and ongoing noise (***Azouz and Gray, 1999***; ***Faisal et al., 2008***). Generally, irrelevant signals would hinder the accurate investigation of the relationship between neural activity and movement behaviors. This raises concerns about whether irrelevant signals could conceal some critical facts about neural encoding and decoding mechanisms.

If the answer is yes, a natural question arises: what critical facts about neural encoding and decoding would irrelevant signals conceal? In terms of neural encoding, irrelevant signals may mask some small neural components, making their encoded information diffcult to detect (***Moreno-Bote et al., 2014***), thereby misleading us to neglect the role of these signals, leading to a partial understanding of neural mechanisms. For example, at the single-neuron level, weakly-tuned neurons are often assumed useless and not analyzed (***Georgopoulos et al., 1986***; ***Hochberg et al., 2012***; ***Wodlinger et al., 2014***; ***Inoue et al., 2018***); at the population level, neural signals composed of lower variance principal components (PCs) are typically treated as noise and discarded (***Churchland et al., 2012***; ***Gallego et al., 2018***, ***2020***; ***Cunningham and Yu, 2014***). So are these ignored signals truly useless or appear that way only due to being covered by irrelevant signals? And what's the role of these ignored signals? In terms of neural decoding, irrelevant signals would significantly complicate the information readout (***Pitkow et al., 2015***; ***Yang et al., 2021***), potentially hindering the discovery of the true readout mechanism of behaviorally-relevant responses. Specifically, in motor cortex, in what form (linear or nonlinear) downstream neurons read out behavioral information from neural responses is an open question. The linear readout is biologically plausible and widely used (***Geor- gopoulos et al., 1986***; ***Hochberg et al., 2012***; ***Wodlinger et al., 2014***), but recent studies (***Glaser et al., 2020***; ***Willsey et al., 2022***) demonstrate nonlinear readout outperforms linear readout. So which readout scheme is more likely to be used by the motor cortex? Whether irrelevant signals are the culprits for the performance gap? Unfortunately, all the above issues remain unclear.

One approach to address the above issues is to accurately separate behaviorally-relevant and irrelevant signals at the single-neuron level and then analyze noise-free behaviorally-relevant signals, which enables us to gain a more accurate and comprehensive understanding of the underlying neural mechanisms. However, this approach is hampered by the fact that the ground truth of behaviorally-relevant signals is unknown, which makes the definition, extraction, and validation of behaviorally-relevant signals a challenging task. As a result, methods of accurate separation remain elusive to date. Existing methods for extracting behaviorally-relevant patterns mainly focus on the latent population level (***Churchland et al., 2012***; ***Sani et al., 2021***; ***Pandarinath et al., 2018***; ***Hurwitz et al., 2021***; ***Yu et al., 2008***; ***Zhou and Wei, 2020***; ***Keshtkaran et al., 2022***) rather than the single-neuron level, and they extract neural activities based on assumptions about specific neural properties, such as linear or nonlinear dynamics (***Sani et al., 2021***; ***Pandarinath et al., 2018***). Although these methods have shown promising results, they fail to capture other parts of behaviorally-relevant neural activity that do not meet their assumptions, thereby providing an incomplete picture of behaviorally-relevant neural activity. To overcome these limitations and obtain accurate behaviorally-relevant signals at the single-neuron level, we propose a novel framework that defines, extracts, and validates behaviorally-relevant signals by simultaneously considering such signals' encoding (behaviorally-relevant signals should

be similar to raw signals to preserve the underlying neuronal properties) and decoding (behaviorally-relevant signals should contain behavioral information as much as possible) properties (see Methods and **Fig. 1**). This framework establishes a prerequisite foundation for the subsequent detailed analysis of neural mechanisms. Here, we conducted experiments using datasets recorded from the motor cortex of three monkeys performing different reaching tasks, where the behavioral variable is movement kinematics. After signal separation by our approach, we first explored how the presence of behaviorally-irrelevant signals affect the analysis of neural activity. We found that behaviorally-irrelevant signals account for a large amount of trial-to-trial neuronal variability, and are evenly distributed across the neural dimensions of behaviorally-relevant signals. Then we explored whether irrelevant signals conceal some facts of neural encoding and decoding. For neural encoding, irrelevant signals obscure the behavioral information encoded by neural responses, especially for neural responses with a large degree of nonlinearity. Surprisingly, neural responses that are usually ignored (weakly-tuned neurons and neural signals composed of small variance PCs) actually encode rich behavioral information in complex nonlinear ways. These responses underpin an unprecedented neuronal redundancy and reveal that movement behaviors are distributed in a higher-dimensional neural space than previously thought. In addition, we found that the integration of smaller and larger variance PCs results in a synergistic effect, allowing the smaller variance PC signals that cannot be linearly decoded to significantly enhance the linear decoding performance, particularly for finer speed control. This finding suggests that lower variance PC signals are involved in regulating precise motor control. For neural decoding, irrelevant signals complicate information readout. Strikingly, when uncovering small neural components obscured by irrelevant signals, linear decoders can achieve comparable decoding performance with nonlinear decoders, providing strong evidence for the presence of linear readout in motor cortex. Together, our findings reveal unexpected complex encoding but simple decoding mechanisms in the motor cortex. Finally, our study also has implications for developing accurate and robust brain-machine interfaces (BMIs) and, more generally, provides a powerful framework for separating behaviorally-relevant and irrelevant signals, which can be applied to other cortical data to uncover more neural mechanisms masked by behaviorally-irrelevant signals.

Results

Framework for defining, extracting, and validating behaviorally-relevant neural signals

What are behaviorally-relevant neural signals?

Since the ground truth of behaviorally-relevant signals is unknown, the definition of behaviorally-relevant signals is not well established yet. In this study, we propose that behaviorally-relevant signals should meet two requirements: (1) they should be similar to raw signals in order to preserve the underlying neuronal properties (encoding requirement), and (2) they should contain behavioral information as much as possible (decoding requirement). If the signals meet both of these requirements, we consider them to be effective behaviorally-relevant signals.

How to extract behaviorally-relevant signals?

One way to extract behaviorally-relevant signals is to use a distillation model to generate them from raw signals while considering the remaining signals as behaviorally-irrelevant. That is, we assume raw signals (**Fig. 1a**) are additively composed of behaviorally-relevant (**Fig. 1b**) and irrelevant (**Fig. 1d**) signals. However, due to the unknown ground truth of behaviorally-relevant signals, a key challenge for the model is to determine the optimal degree of similarity between the generated signals and raw signals. If the generated signals are too similar to raw signals, they may

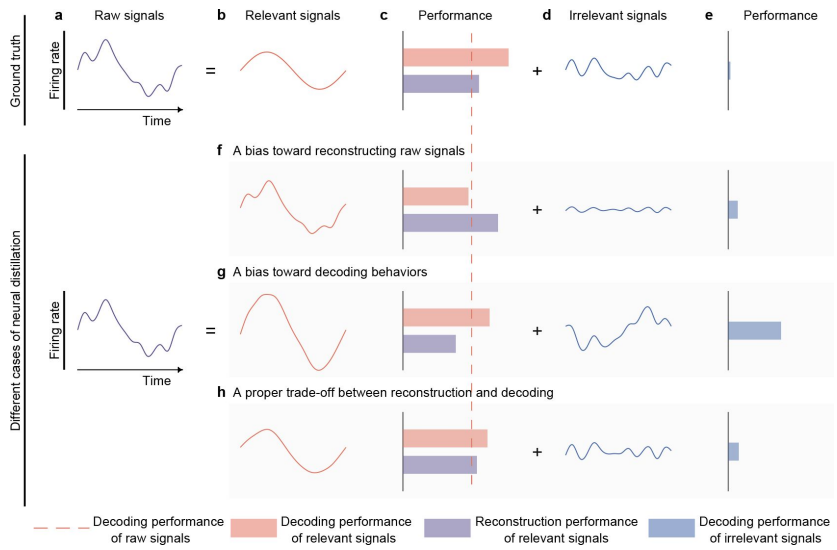


Figure 1.

Semantic illustration of extracting and validating behaviorally-relevant signals.

a-e The ideal decomposition of raw signals. **a**, The temporal neuronal activity of raw signals, where x-axis denotes time, and y-axis represents firing rate. Raw signals re decomposed to relevant (**b**) and irrelevant (**d**) signals. The red dotted line indicates the decoding performance of raw signals. The red and blue bars represent the decoding performance of relevant and irrelevant signals. The purple bar represents the reconstruction performance of relevant signals, which measures the neural similarity between generated signals and raw signals. The longer the bar, the larger the performance. The ground truth of relevant signals decode information perfectly (**c**, red bar) and is similar to raw signals to some extent (**c**, purple bar), and the ground truth of irrelevant signals contain little behavioral information (**e**, blue bar). **f-h**, Three different cases of behaviorally-relevant signals distillation. **f**, When the model is biased toward generating relevant signals that are similar to raw signals, it will achieve high reconstruction performance, but the decoding performance will suffer due to the inclusion of too many irrelevant signals. As it is diffcult for models to extract complete relevant signals, the residuals will also contain some behavioral information. **g**, When the model is biased toward generating signals that prioritize decoding over similarity to raw signals, it will achieve high decoding performance, but the reconstruction performance will be low. Meanwhile, the residuals will contain a significant amount of behavioral information. **h**, When the model balances the trade-off of decoding and reconstruction capabilities of relevant signals, both decoding and reconstruction performance will be good, and the residuals will only contain a little behavioral information.

contain a large amount of irrelevant information, which would hinder the exploration of neural mechanisms. Conversely, if the generated signals are too dissimilar to raw signals, they may lose behaviorally-relevant information, also hindering the exploration of neural mechanisms. To overcome this challenge, we exploited the trade-off between the similarity of generated signals to raw signals (encoding requirement) and their decoding performance of behaviors (decoding requirement) to extract effective behaviorally-relevant signals (for details, see Methods and **Fig. S1**). The core assumption of our model is that behaviorally-irrelevant signals are noise relative to behaviorally-relevant signals, and thereby irrelevant signals would degrade the decoding generalization of generated behaviorally-relevant signals. Generally, the distillation model is faced with three cases: a bias toward reconstructing raw signals (**Fig. 1f**), a bias toward decoding behaviors (**Fig. 1g**), and a proper trade-off between reconstruction and decoding (**Fig. 1h**). If the distillation model is biased toward extracting signals similar to raw signals, the distilled behaviorally-relevant signals will contain an excessive amount of behaviorally-irrelevant information, affecting the decoding generalization of these signals (**Fig. 1f**). If the model is biased toward extracting parsimonious signals that are discriminative for decoding, the distilled signals will not be similar enough to raw signals, and some redundant but useful signals will be left in the residuals (**Fig. 1g**), making irrelevant signals contain much behavioral information. Using face recognition as an example, if a model can accurately identify an individual using only the person's eyes (assuming these are the most useful features), other useful information such as the nose or mouth will be left in the residuals, which could also be used to identify the individual. Neither of these two cases is desirable because the former loses decoding performance, while the latter loses some useful neural signals, which are not conducive to our subsequent analysis of the relationship between behaviorally-relevant signals and behaviors. The behaviorally-relevant signals we want should be similar to raw signals and preserve the behavioral information maximally, which can be obtained by balancing the encoding and decoding properties of generated behaviorally-relevant signals (**Fig. 1h**).

How to validate behaviorally-relevant signals?

To validate the effectiveness of the distilled signals, we proposed three criteria. The first criterion is that the decoding performance of the behaviorally-relevant signals (red bar, **Fig.1**) should surpass that of raw signals (the red dotted line, **Fig.1**). Since decoding models, such as deep neural networks, are more prone to overfit noisy raw signals than behaviorally-relevant signals, the distilled signals should demonstrate better decoding generalization than the raw signals. The second criterion is that the behaviorally-irrelevant signals should contain minimal behavioral information (blue bar, **Fig.1**). This criterion can assess whether the distilled signals maximally preserve behavioral information from the opposite perspective and effectively exclude undesirable cases, such as over-generated and under-generated signals. Specifically, in the case of over-generation, suppose $z = x + y$, where z, x, and y represent raw, relevant, and irrelevant signals, respectively. If the distilled relevant signals $\hat{x}$ are added extra signals m which do not exist in the real behaviorally-relevant signals, that is, $\hat{x} = x + m$, then the corresponding residuals yA will be equal to the ideal irrelevant signals y plus the negative extra signals $-m$, namely, $yA = y - m$, thus the residuals yA contain the amount of information preserved by negative extra signals $-m$. Similarly, in the case of under-generation, if the distilled behaviorally-relevant signals are incomplete and lose some useful information, this lost information will also be reflected in the residuals. In these cases, the distilled signals are not suitable for analysis. The third criterion is that the distilled behaviorally-relevant signals should be similar to raw signals to maintain essential neuronal properties (purple bar, **Fig.1**). If the distilled signals do not resemble raw signals, they fail to retain the fundamental characteristics of raw signals, which are not qualified for subsequent analysis. Overall, if the distilled signals satisfy the above three criteria, we consider the distilled signals to be effective.

d-VAE extracts effective behaviorally-relevant signals

To demonstrate the effectiveness of our model (d-VAE) in extracting behaviorally-relevant signals, we conducted experiments on the synthetic dataset where the ground truth of relevant and irrelevant signals are already known (see Methods) and three benchmark datasets with different paradigms (**Fig. 2a,e,i**; see Methods for details), and compared d-VAE with four other distillation models, including pi-VAE (***Zhou and Wei, 2020***), PSID (***Sani et al., 2021***), TNDM (***Hurwitz et al., 2021***) and LFADS (***Pandarinath et al., 2018***). Specifically, we first applied these distillation models to raw signals to obtain the distilled behaviorally-relevant signals, considering the residuals as behaviorally-irrelevant signals. We then evaluated the decoding $\mathbf{R}^2$ between the predicted velocity and actual velocity of the two partition signals using a linear Kalman filter (KF) and a nonlinear artificial neural network (ANN) and measured the neural similarity between behaviorally-relevant and raw signals. Besides, we compared the decoding performance of behaviorally-relevant signals extracted by d-VAE and behaviorally-relevant embeddings extracted by CEBRA (a non-generative method which outperforms pi-VAE) (***Schneider et al., 2023***).

Overall, d-VAE successfully extracts effective behaviorally-relevant signals that meet the three criteria outlined above on both synthetic (**Supplementary Fig. S2**) and real data (**Fig. 2**). On the synthetic data (**Supplementary Fig. S2**), results show that d-VAE can strike an effective balance between the reconstruction and decoding performance of generated signals to extract effective relevant signals that are similar to the ground truth relevant signals, meanwhile removing effective irrelevant signals that resemble the ground truth irrelevant signals (**Supplementary Fig. S2 a-g**), and outperforms other distillation models (**Supplementary Fig. S2 h-k**). On the real data, specifically in the obstacle avoidance task (**Fig. 2a**), the monkey is required to move the ball from the start point (yellow) to the target point (blue) without hitting the obstacle. For the decoding performance of behaviorally-relevant signals (**Fig. 2b**), the signals distilled by d-VAE outperform the raw signals (purple bars with slash lines) and the signals distilled by all other distillation models (PSID, pink; pi-VAE, green; TNDM, blue; and LFADS, light green) with the KF as well as the ANN. For the decoding performance of behaviorally-irrelevant signals (**Fig. 2c**), behaviorally-irrelevant signals obtained by d-VAE achieves the lowest decoding performance compared with behaviorally-irrelevant signals obtained by other approaches. Therefore, the combination of dVAE's highest decoding performance for behaviorally-relevant signals and lowest decoding performance for behaviorally-irrelevant signals demonstrate its superior ability to extract behaviorally-relevant signals from noisy signals. For the neural similarity between behaviorally-relevant and raw signals (**Fig. 2d**), the distilled signals obtained by d-VAE achieve the highest performance among competitors ($P < 0.01$, Wilcoxon rank-sum test). Similar results were obtained for the center-out task (**Fig. 2e-h**) and the self-paced reaching task (**Fig. 2i-l**), indicating the consistency of d-VAE's distillation ability across a range of motor tasks. To provide a more intuitive illustration of the similarity between raw and distilled signals, we displayed the firing rate of neuronal activity in five trials under the same condition (**Fig. 2m**), and results clearly show that the firing pattern of distilled signals is similar to the corresponding raw signals. Finally, we compared d-VAE with CEBRA on the four datasets; results show that d-VAE achieves comparable performance with CEBRA using the ANN decoder but outperforms CE-BRA with the KF decoder (**Supplementary Fig. S3**), demonstrating that d-VAE can effectively extract behavioral information.

In summary, d-VAE distills effective behaviorally-relevant signals that preserve behavioral information maximally and are similar to raw signals. Meanwhile, the behaviorally-irrelevant signals discarded by d-VAE contain a little behavioral information. Therefore, these signals are reliable for exploring the encoding and decoding mechanisms of relevant signals.

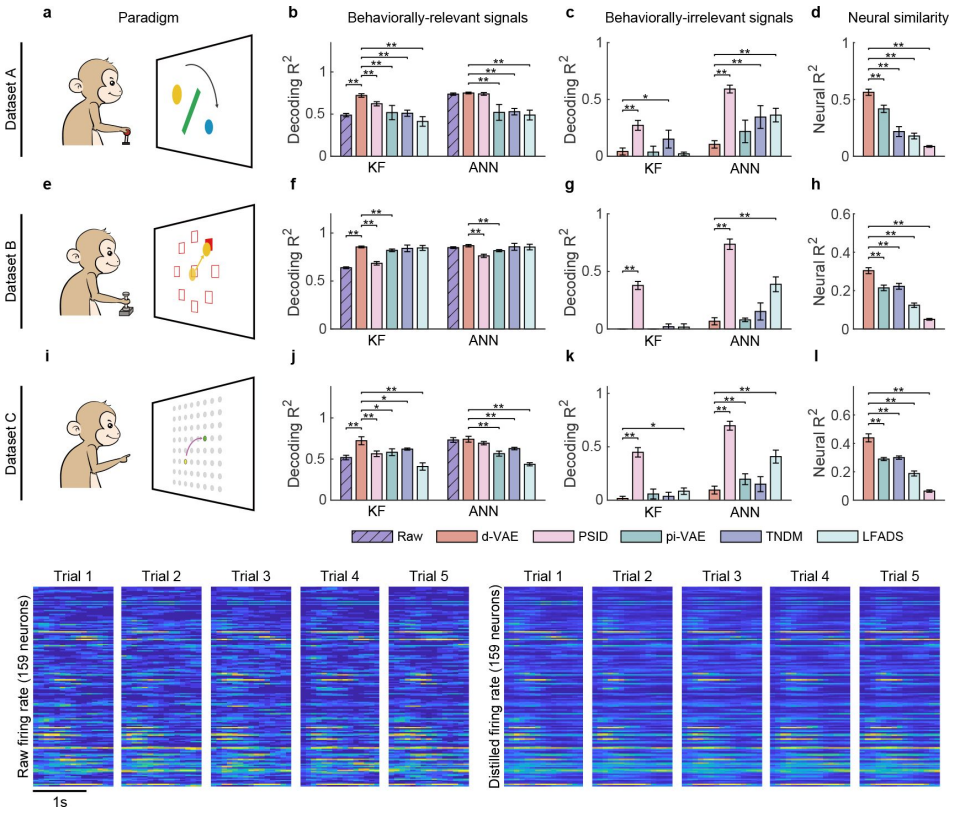


Figure 2.

Evaluation of separated signals.

a, The obstacle avoidance paradigm **b**, The decoding $\mathbf{R}^2$ between true velocity and predicted velocity of raw signals (purple bars with slash lines) and behaviorally-relevant signals obtained by d-VAE (red), PSID (pink), pi-VAE (green), TNDM (blue), and LFADS (light green) on dataset A. Error bars denote mean ± standard deviation (s.d.) across five cross-validation folds. Asterisks represent significance of Wilcoxon rank-sum test with $^{*}P < 0.05$, $^{**}P < 0.01$. **c**, Same as **b**, but for behaviorally-irrelevant signals obtained by five different methods. **d**, The neural similarity ($\mathbf{R}^2$) between raw signals and behaviorally-relevant signals extracted by d-VAE, PSID, pi-VAE, TNDM, and LFADS. Error bars represent mean ± s.d. across five cross-validation folds. Asterisks indicate significance of Wilcoxon rank-sum test with $^{**}P < 0.01$. **e-h** and **i-l**, Same as **a-d**, but for dataset B with the center-out paradigm (**e**) and dataset C with the self-paced reaching paradigm (**i**). **m**, The firing rates of raw signals and distilled signals obtained by d-VAE in five held-out trials under the same condition of dataset B.

How do behaviorally-irrelevant signals affect the analysis of neural activity at the single-neuron level?

Following signal separation, we first explored how behaviorally-irrelevant signals affect the analysis of neural activity at the single-neuron level. Specifically, we examined the effect of irrelevant signals on two critical properties of neuronal activity: the preferred direction (PD) (*Georgopoulos et al., 1986*) and trial-to-trial variability. Our objective was to know how irrelevant signals affect the PD of neurons and whether irrelevant signals contribute significantly to neuronal variability.

To explore how irrelevant signals affect the PD of neurons, we first calculated the PD of both raw and distilled signals separately and then quantified the PD deviation by the angle difference between these two signals. Results show that the PD deviation increases as the neuronal $\mathbf{R}^2$ (*Collinger et al., 2013*; *Wodlinger et al., 2014*) decreases (red curve, **Fig. 3a,e** and **Supplementary Fig. S4a**); neurons with larger $\mathbf{R}^2$ (strongly linear-tuned neurons) exhibit stable PDs with signal distillation (see example PDs in the inset), while neurons with smaller $\mathbf{R}^2$s (weakly linear-tuned neurons) show a larger PD deviation. These results indicate that irrelevant signals have a small effect on strongly-tuned neurons but a large effect on weakly-tuned neurons. One possible reason for the larger PD deviation in weakly-tuned neurons is that they have a lower degree of linear encoding but a higher degree of nonlinear encoding, and highly nonlinear structures are more susceptible to interference from irrelevant signals (*Nogueira et al., 2023*). Moreover, after filtering out the behaviorally-irrelevant signals, the cosine tuning fit ($\mathbf{R}^2$) of neurons increases ($P < 10^{-20}$, Wilcoxon signed-rank test; **Fig. 3b,f** and **Supplementary Fig. S4b**), indicating that irrelevant signals reduce the neurons' tuning expression. Notably, even after removing the interference of irrelevant signals, the $\mathbf{R}^2$ of neurons remains relatively low. These results demonstrate that the linear encoding model only explains a small fraction of neural responses, and neuronal activity encodes behavioral information in complex nonlinear ways.

To investigate whether irrelevant signals significantly contribute to neuronal variability, we compared the neuronal variability (measured with the Fano factor (*Churchland et al., 2010*), FF) of relevant and raw signals. Results show that the condition-averaged FF of each neuron of distilled signals is lower than that of raw signals ($P < 10^{-20}$, Wilcoxon signed-rank test; **Fig. 3c,g**), and the mean (broken line) and median FFs of all neurons under different conditions are also significantly lower than those of raw signals ($P < 0.01$, Wilcoxon signed-rank test; **Fig. 3d,h**), indicating that irrelevant signals significantly contribute to neuronal variability. We then visualized the single-trial neuronal activity of example neurons under different conditions (**Fig. 3i** and **Supplementary Fig. S5**).

Results demonstrate that the patterns of relevant signals are more consistent and stable across different trials than raw signals, and the firing activity of irrelevant signals varies randomly. These results indicate that irrelevant signals significantly contribute to neuronal variability, and eliminating the interference of irrelevant signals enables us observe the changes in neural pattern more accurately.

How do behaviorally-irrelevant signals affect the analysis of neural activity at the population level?

The neural population structure is an essential characteristic of neural activity. Here, we examined how behaviorally-irrelevant signals affect the analysis of neural activity at the population level, including four aspects: (1) the population properties of relevant and irrelevant signals, (2) the subspace overlap relationship between the two signal components, (3) how the two partitions contribute to raw signals, and (4) the difference in population properties between raw and distilled signals.

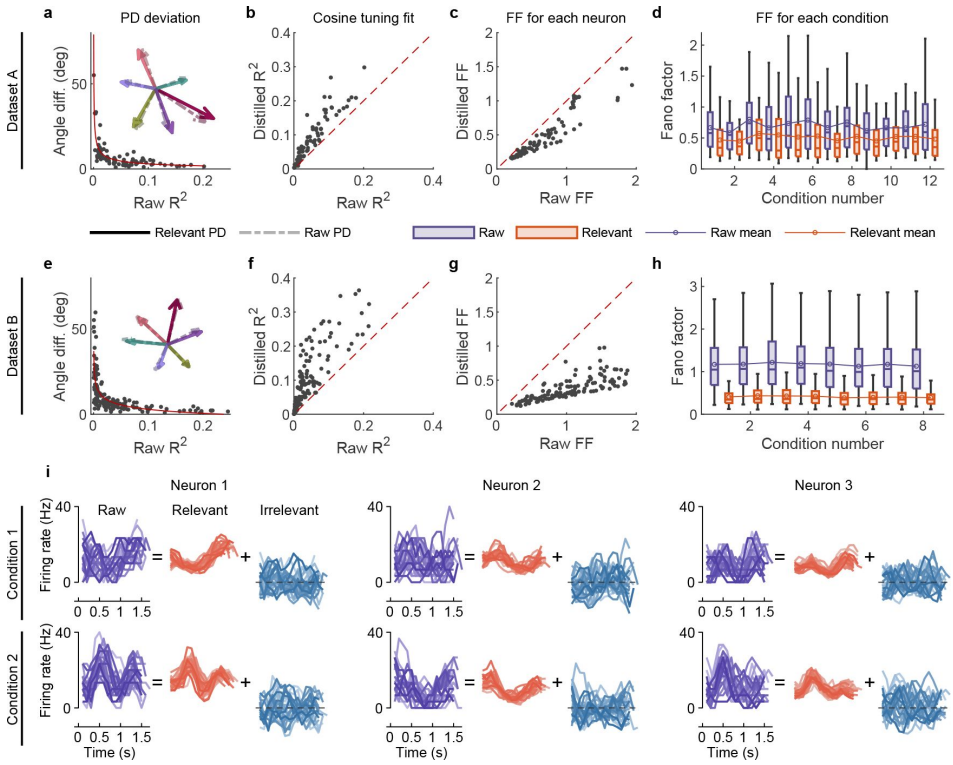


Figure 3.

The effect of irrelevant signals on analyzing neural activity at the single-neuron level.

a, The angle difference (AD) of preferred direction (PD) between raw and distilled signals as a function of the $\mathbf{R}^2$ of raw signals on datasets A. When employing $\mathbf{R}^2$ to characterize neurons, it indicates the extent to which neuronal activity is explained by the linear encoding model. Smaller $\mathbf{R}^2$ neurons have a lower capacity for linearly tuning (encoding) behaviors, while larger $\mathbf{R}^2$ neurons have a higher capacity for linearly tuning (encoding) behaviors. Each black point represents a neuron (n = 90). The red curve is the fitting curve between $\mathbf{R}^2$ and AD. Five example larger $\mathbf{R}^2$ neurons' PDs are shown in the inset plot, where the solid and dotted line arrows represent the PDs of relevant and raw signals, respectively. **b**, Comparison of the cosine tuning fit ($\mathbf{R}^2$) before and after distillation of single neurons (black points), where the x-axis and y-axis represent neurons' $\mathbf{R}^2$ of raw and distilled signals, respectively. **c**, Comparison of neurons' Fano factor (FF) averaged across conditions of raw (x-axis) and distilled (y-axis) signals, where FF is used to measure the neuronal variability of different trials in the same condition. **d**, Boxplots of raw (purple) and distilled (red) signals under different conditions for all neurons (12 conditions). Boxplots represent medians (lines), quartiles (boxes), and whiskers extending to ± 1.5 times the interquartile range. The broken lines represent the mean FF across all neurons. **e-h**, Same as **a-d**, but for dataset B (n=159, 8 conditions). **i**, Example of three neurons' raw firing activity decomposed into behaviorally-relevant and irrelevant parts using all trials under two conditions (2 of 8 directions) in held-out test sets of dataset B.

To explore the population properties of relevant and irrelevant signals, we separately applied principal component analysis (PCA) on each partition to obtain the corresponding cumulative variance curve in a descending variance order. Our results show that the primary subspace (capturing the top 90% variance) of relevant signals (thick red line, Fig.4a,e and **Supplementary Fig. S6a**) is only explained by a few dimensions (7, 13, and 9 for each dataset), indicating that the primary part of behaviorally-relevant signals exists in a low-dimensional subspace. In contrast, the primary subspace of irrelevant signals (thick blue line, **Fig. 4b,f** and **Supplementary Fig. S6b**) requires more dimensions (46, 81, and 59). The variance distribution of behaviorally-irrelevant signals across dimensions (thick blue line, **Fig. 4b,f** and **Supplementary Fig. S6b**) is more even than behaviorally-relevant signals (thick red line, **Fig. 4a,e** and **Supplementary Fig. S6a**) but not as uniform as Gaussian noise .N (**0, I**) (thin gray line, Fig.4b,f and **Supplementary Fig. S6b**), indicating that irrelevant signals are not pure noise but rather bear some significant structure, which may represent information from other irrelevant tasks.

To investigate the subspace overlap between relevant and irrelevant signals, we calculated how many variances of irrelevant signals can be captured by relevant PCs by projecting irrelevant signals onto relevant PCs and vice versa (***Elsayed et al., 2016***; ***Rouse and Schieber, 2018***; ***Jiang et al., 2020***)(see Methods). We found that the variance of irrelevant signals increases relatively uniformly over relevant PCs (blue thin line, **Fig. 4a,e** and **Supplementary Fig. S6a**), like random noise's variance accumulation explained by relevant PCs (gray thin line, **Fig. 4a,e** and **Supplementary Fig. S6a**); similar results are observed for relevant signals explained by irrelevant PCs (red thin line, **Fig. 4b,f** and **Supplementary Fig. S6b**). These results indicate that relevant PCs can not match the informative dimensions of irrelevant signals and vice versa, suggesting the dimensions of behaviorally-relevant and irrelevant signals are unrelated. It is worth mentioning that the signals obtained by pi-VAE are in contrast to our findings. Its results show that a few relevant PCs can explain a considerable variance of irrelevant signals (thin red line, **Supplementary Fig. S7b,f,j**), which indicates that the relevant and irrelevant PCs are closely related. The possible reason is that the pi-VAE leaks a lot of information into the irrelevant signals. Notably, **Fig. 4a,e** and **Supplementary Fig. S6a** show that irrelevant signals got by d-VAE projecting on behaviorally-relevant primary subspace only capture a little variance information (9%, 12%, and 9%), indicating that the primary subspace of behaviorally-relevant signals is nearly orthogonal to irrelevant space.

To investigate the composition of raw signals by the two partitions, we performed PCA on raw neural signals to obtain raw PCs, and then projected the relevant and irrelevant signals onto these PCs to assess the proportion of variance of raw signals explained by each partition. First, we analyzed the overall proportion of relevant and irrelevant signals that constitute the raw signals (the inset pie plot, **Fig. 4c,g** and **Supplementary Fig. S6c**). The variance of the raw signals is composed of three parts: the variance of relevant signals, the variance of irrelevant signals, and the correlation between relevant and irrelevant signals (see Methods). The results demonstrate that the irrelevant signals account for a large proportion of raw signals, suggesting the motor cortex encodes considerable information that is not related to the measured behaviors. In addition, there is only a weak correlation between relevant and irrelevant signals, implying that behaviorally-relevant and irrelevant signals are nearly uncorrelated.

We then examined the proportions of relevant and irrelevant signals in each PC of raw signals. We found that relevant signals (red) occupy the dominant proportions in the larger variance raw PCs (before the PC marked with a red triangle), while irrelevant signals (blue) occupy the dominant proportions in the smaller variance raw PCs (after the PC marked with a red triangle) (**Fig. 4c,g** and **Supplementary Fig. S6c**). Similar results are observed in the accumulation of each raw PC (**Fig. 4d,h** and **Supplementary Fig. S6d**). Specifically, the results show that the variance accumulation of raw signals (purple line) in larger variance PCs is mainly contributed by relevant signals (red line), while irrelevant signals (blue line) primarily contribute to the lower variance

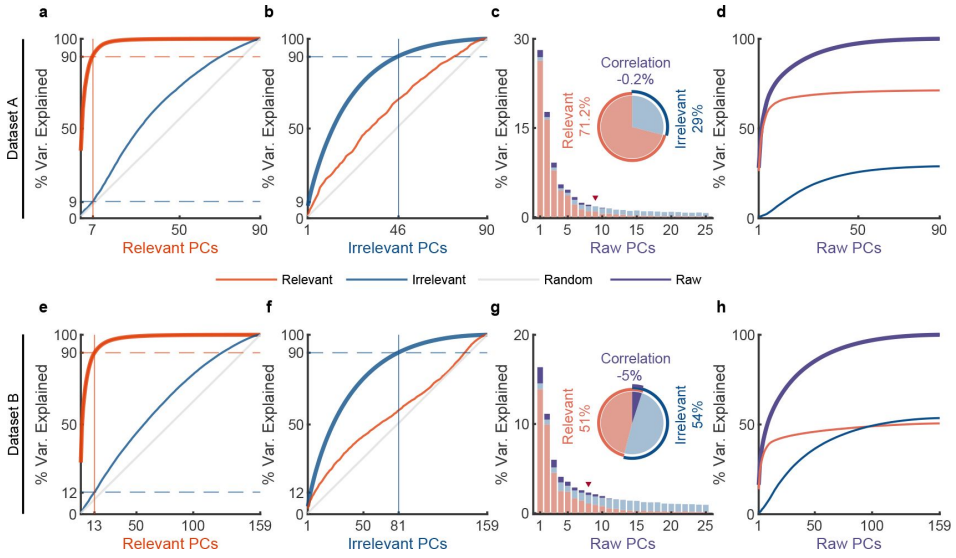


Figure 4.

The effect of irrelevant signals on analyzing neural activity at the population level.

a,**b**, PCA is separately applied on relevant and irrelevant signals to get relevant PCs and irrelevant PCs. The thick lines represent the cumulative variance explained for the signals on which PCA has been performed, while the thin lines represent the variance explained by those PCs for other signals. Red, blue, and gray colors indicate relevant signals, irrelevant signals, and random Gaussian noise .N (**0, I**) (for chance level) where the mean vector is zero and the covariance matrix is the identity matrix. The horizontal lines represent the percentage of variance explained. The vertical lines indicate the number of dimensions accounted for 90% of the variance in behaviorally-relevant (left) and irrelevant (right) signals. For convenience, we defined the principal component subspace describing the top 90% variance as the primary subspace and the subspace capturing the last 10% variance as the secondary subspace. The cumulative variance explained for behaviorally-relevant (**a**) and irrelevant (**b**) signals got by d-VAE on dataset A. **c**,**d**, PCA is applied on raw signals to get raw PCs. **c**, The bar plot shows the composition of each raw PC. The inset pie plot shows the overall proportion of raw signals, where red, blue, and purple colors indicate relevant signals, irrelevant signals, and the correlation between relevant and relevant signals. The PC marked with a red triangle indicates the last PC where the variance of relevant signals is greater than or equal to that of irrelevant signals. **d**, The cumulative variance explained by raw PCs for different signals, where the thick line represents the cumulative variance explained for raw signals (purple), while the thin line represents the variance explained for relevant (red) and irrelevant (blue) signals. **e-h**, Same as **a-d**, but for dataset B.

PCs. These results demonstrate that irrelevant signals have a small effect on larger variance raw PCs but a large effect on smaller variance raw PCs. This finding eliminates the concern that irrelevant signals would significantly affect the top few PCs of raw signals and thus produce inaccurate conclusions. To further validate this finding, we used the top six PCs as jPCA (***Churchland et al., 2012***) did to examine the rotational dynamics of distilled and raw signals (**Fig. S8**). Results show that the rotational dynamics of distilled signals are similar to those of raw signals.

Finally, to directly compare the population properties of raw and relevant signals, we plotted the cumulative variance curves of raw and relevant signals (**Supplementary Fig. S9**). Results (upper left corner curves, **Supplementary Fig. S9**) show that the cumulative variance curve of relevant signals (red line) accumulates faster than that of raw signals (purple line) in the preceding larger variance PCs, indicating that the variance of the relevant signal is more concentrated in the larger variance PCs than that of raw signals. Furthermore, we found that the dimensionality of primary subspace of raw signals (26, 64, and 45 for datasets A, B, and C) is significantly higher than that of behaviorally-relevant signals (7, 13, and 9), indicating that using raw signals to estimate the neural dimensionality of behaviors leads to an overestimation.

Distilled behaviorally-relevant signals uncover that smaller R^2 neurons encode rich behavioral information in complex nonlinear ways

The results presented above regarding PDs (**Fig. 3** and **Fig. S4**) demonstrate that irrelevant signals significantly impact smaller $\mathbf{R}^2$ neurons and weakly impact larger $\mathbf{R}^2$ neurons. Under the interference of irrelevant signals, it is diffcult to explore the amount of behavioral information in neuronal activity. Given that we have accurately separated the behaviorally-relevant and irrelevant signals, we explored whether irrelevant signals would mask some encoded information of neuronal activity, especially for smaller $\mathbf{R}^2$ neurons.

To answer the question, we divided the neurons into two groups of smaller $\mathbf{R}^2$ ($\mathbf{R}^2$ <= 0.03) and larger $\mathbf{R}^2$ ($\mathbf{R}^2$ > 0.03), and then used decoding models to assess how much information is encoded in raw and distilled signals. As shown in **Fig. 5a**, for the smaller $\mathbf{R}^2$ neuron group, both KF and ANN decode behavioral information poorly on raw signals, but achieve high decoding performance using relevant signals. Specifically, the KF decoder (left plot, **Fig. 5a**) improves the decoding $\mathbf{R}^2$ significantly from 0.044 to 0.616 (improves by about 1300%; $P < 0.01$, Wilcoxon rank-sum test) after signal distillation; the ANN decoder (right plot, **Fig. 5a**) improves from 0.374 to 0.753 (improves by about 100%; $P < 0.01$, Wilcoxon rank-sum test). For the larger $\mathbf{R}^2$ neuron group, the decoding performance of relevant signals with ANN does not improve much compared with the decoding performance of raw signals, but the decoding performance of relevant signals with KF is significantly better than that of raw signals ($P < 0.01$, Wilcoxon rank-sum test). Similar results are obtained with datasets B (**Fig. 5d**) and C (**Supplementary Fig. S10d**). These results indicate that irrelevant signals mask behavioral information encoded by neuronal populations, especially for smaller $\mathbf{R}^2$ neurons with a higher degree of nonlinearity, and that smaller $\mathbf{R}^2$ neurons actually encode rich behavioral information.

The fact that the smaller $\mathbf{R}^2$ neurons encode rich information seems unexpected, and interestingly, we cannot obtain this rich information solely by distilling smaller $\mathbf{R}^2$ neurons. This observation gives rise to two alternative scenarios. The first is that larger $\mathbf{R}^2$ neurons introduce additional signals to smaller $\mathbf{R}^2$ neurons, which they do not inherently possess, resulting in an excessive amount of behavioral information within the smaller $\mathbf{R}^2$ neurons. The second is that the smaller $\mathbf{R}^2$ neurons inherently possess a substantial amount of information, and larger $\mathbf{R}^2$ neurons utilize their neural activity, which is correlated with that of small $\mathbf{R}^2$ neurons, to aid in restoring the small $\mathbf{R}^2$ neurons' original appearance; this process is analogous to image denoising, where damaged

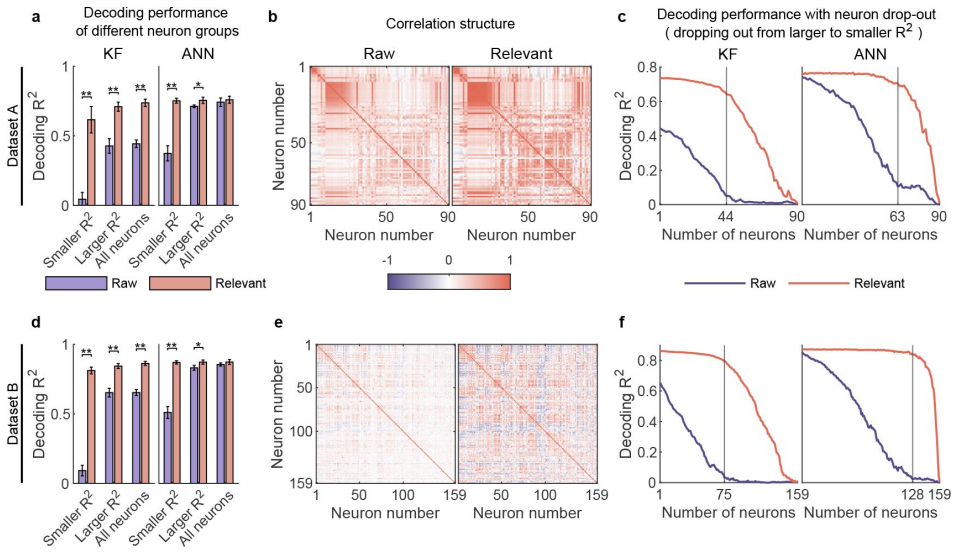


Figure 5.

Smaller R^2 neurons encode rich behavioral information in complex nonlinear ways.

a, The comparison of decoding performance between raw (purple) and distilled signals (red) on dataset A with different neuron groups, including smaller $\mathbf{R}^2$ neuron ($\mathbf{R}^2$ <= 0.03), larger $\mathbf{R}^2$ neuron ($\mathbf{R}^2$ > 0.03), and all neurons. Error bars indicate mean ± s.d. across five cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{*}P < 0.05$, $^{**}P < 0.01$. **b**, The correlation matrix of all neurons of raw (left) and behaviorally-relevant (right) signals on dataset A. Neurons are ordered to highlight correlation structure (details in Methods). **c**, The decoding performance of KF (left) and ANN (right) with neurons dropped out from larger to smaller $\mathbf{R}^2$ on dataset A. The vertical gray line indicates the number of dropped neurons at which raw and behaviorally-relevant signals have the greatest performance difference. **d-f**, Same as **a-c**, for dataset B.

noisy pixels necessitate the assistance of their correlated, clean neighboring pixels to recover their original appearance. We initially tested the first scenario and found it to be unsupported for two key reasons. Firstly, our model enforces that distilled neuronal activity closely resembles the corresponding original neuronal activity, effectively preventing the generation of arbitrarily shaped neuronal activity, such as that of other neurons. As shown in **Fig. 3i** and **Supplementary Fig. S5**, our distilled relevant neuronal activity exhibits a high degree of similarity to the corresponding raw neuronal activity. To assess whether the distilled neurons exhibit the highest similarity to the corresponding raw neurons, we compared the neural similarity ($\mathbf{R}^2$) of each distilled neuron to all raw neurons. The results indicate that 78/90 (87%, dataset A), 153/159 (96%, dataset B), and 91/91 (100%, dataset C) distilled neurons are most similar to the corresponding neurons. The remaining distilled neurons rank among the top four in similarity to the corresponding neurons, further confirming the close resemblance of distilled neuronal activity to the corresponding raw neuronal activity. Secondly, as we emphasized in the section on validating behaviorally-relevant signals with the second criterion, if this large amount of information is compensated by other neurons, the residuals should also contain a large amount of information. However, as illustrated in **Fig. 2c**, g, and k, the residuals contain only a little information. Therefore, based on these two reasons, the first scenario is rejected. Then we tested the second scenario. To verify this scenario, we conducted experiments using synthetic data with known ground truth (see Methods). In this dataset, small $\mathbf{R}^2$ neurons inherently contained a substantial amount of information but were obscured by noise, making them undecodable. We aimed to assess whether d-VAE could recover the lost information and restore the damaged neuronal activity. The results demonstrate that, with the assistance of large $\mathbf{R}^2$ neurons, d-VAE effectively recovers a significant amount of information that are obscured by noise (**Fig. S11a**). Additionally, the distilled signals exhibit a remarkable improvement in neural similarity to the ground truth compared to the raw signals ($P < 0.01$, Wilcoxon rank-sum test; **Fig. S11b**). Therefore, these results support the second scenario and collectively confirm that smaller $\mathbf{R}^2$ neurons indeed contain rich behavioral information, and this finding is not a by-product of d-VAE.

Given that both smaller and larger $\mathbf{R}^2$ neurons encode rich behavioral information, it is worth noting that the sum of the decoding performance of smaller $\mathbf{R}^2$ neurons and larger $\mathbf{R}^2$ neurons is significantly greater than that of all neurons for relevant signals (red bar, **Fig. 5a,d** and **Supplementary Fig. S10a**), demonstrating that movement parameters are encoded very redundantly in neuronal population. In contrast, we can not find this degree of neural redundancy in raw signals (purple bar, **Fig. 5a,d** and **Supplementary Fig. S10a**) because the encoded information of smaller $\mathbf{R}^2$ neurons are masked by irrelevant signals. Therefore, these smaller $\mathbf{R}^2$ neurons, which are usually ignored, are actually useful and play a critical role in supporting neural redundancy. Generally, cortical redundancy can arise from neuronal correlations, which are critical for revealing certain aspects of neural circuit organization (***Yatsenko et al., 2015***). Accordingly, we visualized the ordered correlation matrix of neurons (see Methods) for both raw and relevant signals (**Fig. 5b,e** and **Supplementary Fig. S10b**) and found that the neuronal correlation of relevant signals is stronger than that of raw signals. These results demonstrate that irrelevant signals weaken the neuronal correlation, which may hinder the accurate investigation of neural circuit organization.

Considering the rich redundancy and strong correlation of neuronal activity, we wondered whether the neuronal population could utilize redundant information from other neurons to exhibit robustness under the perturbation of neuronal destruction. To investigate this question, we evaluated the decoding performance of dropping out neurons from larger $\mathbf{R}^2$ to smaller $\mathbf{R}^2$ on raw and relevant signals. The results (**Fig. 5c,f** and **Supplementary Fig. S10c**) show that the decoding performance of the KF and ANN on raw signals (purple line) decreases steadily before the number of neurons marked (vertical gray line), and the remaining smaller $\mathbf{R}^2$ neurons decode behavioral information poorly. In contrast, even if many neurons are lost, the decoding performance of KF and ANN on relevant signals (red line) maintains high accuracy. This finding indicates that behaviorally-relevant signals are robust to the disturbance of neuron drop-out, and

smaller $\mathbf{R}^2$ neurons play a critical role in compensating for the failure of larger $\mathbf{R}^2$ neurons. In contrast, this robustness can not be observed in raw signals because irrelevant signals mask neurons' information and weaken their correlation. Notably, the ANN outperforms the KF when only smaller $\mathbf{R}^2$ neurons are left (**Fig. 5c,f** and **Supplementary Fig. S10c**), suggesting that smaller $\mathbf{R}^2$ neurons can fully exploit their nonlinear ability to cope with large-scale neuronal destruction.

Distilled behaviorally-relevant signals uncover that signals composed of smaller variance PCs encode rich behavioral information in complex nonlinear ways

The results presented above regarding subspace overlap (**Fig. 4** and **Fig. S6**) show that irrelevant signals have a small impact on larger variance PCs but dominate smaller variance PCs. Therefore, we aimed to investigate whether irrelevant signals would mask some encoded information of neural population, especially signals composed of smaller variance PCs.

To answer the question, we compared the decoding performance of raw and distilled signals with different raw PC groups. Specifically, we firstly divided the raw PCs into two groups, that is, smaller variance PCs and larger variance PCs, defined by ratio of relevant to irrelevant signals in the raw PCs (the red triangle, see **Fig. 4c,g** and **Supplementary Fig. S6c**). Then we projected raw and distilled signals onto these two PC groups and got the corresponding signals. Results show that, for the smaller variance PC group, both KF and ANN achieve much better performance on distilled signals than raw signals ($P < 0.01$, Wilcoxon rank-sum test, for ANN), whereas for the larger variance PC group, the decoding performance of relevant signals does not improve a lot compared with the decoding performance of raw signals (see **Fig. 6a,d** and **Supplementary Fig. S10a**). These results demonstrate that irrelevant signals mask the behavioral information encoded by different PC groups, especially for signals composed of smaller variance PCs (smaller variance PC signals), and smaller variance PC signals actually encode rich behavioral information.

The above results are based on raw PCs. However, raw PCs are biased by irrelevant signals and thus cannot faithfully reflect the characteristics of relevant signals. As we have successfully separated the behaviorally-relevant signals, we aimed to explore how behavioral information of distilled signals is distributed across relevant PCs. To do so, we used decoding models to evaluated the amount of behavioral information contained in cumulative PCs of relevant signals (using raw signals as a comparison). The cumulative variance explained by PCs in descending and ascending order of variance and the dimensionality corresponding to the top 90% variance signals (called primary signals) and the last 10% variance signals (called secondary signals) are shown in **Supplementary Fig. S9**.

Here we first investigated secondary signals' decoding ability solely by accumulating PCs from smaller to larger variance. The results show that, for relevant signals, KF can hardly decode behavioral information solely using secondary signals (red line; left plot, **Fig. 6b,e** and **Supplementary Fig. S10e**), but ANN can decode rich information (red line; right plot, **Fig. 6b,e** and **Supplementary Fig. S10e**). These results indicate that smaller variance PC signals encode rich information in complex nonlinear ways. In contrast, when using raw signals composed of the same number of dimensions as the secondary signals (purple line, **Fig. 6b,e** and **Supplementary Fig. S10e**), the amount of information identified by ANN is significantly smaller than that of relevant secondary signals ($P < 0.01$, Wilcoxon rank-sum test). These results demonstrate that signals composed of these neural dimensions actually encode rich behavioral information, and irrelevant signals make them seem insignificant, indicating that behavioral information is distributed in a higher-dimensional subspace than expected from raw signals.

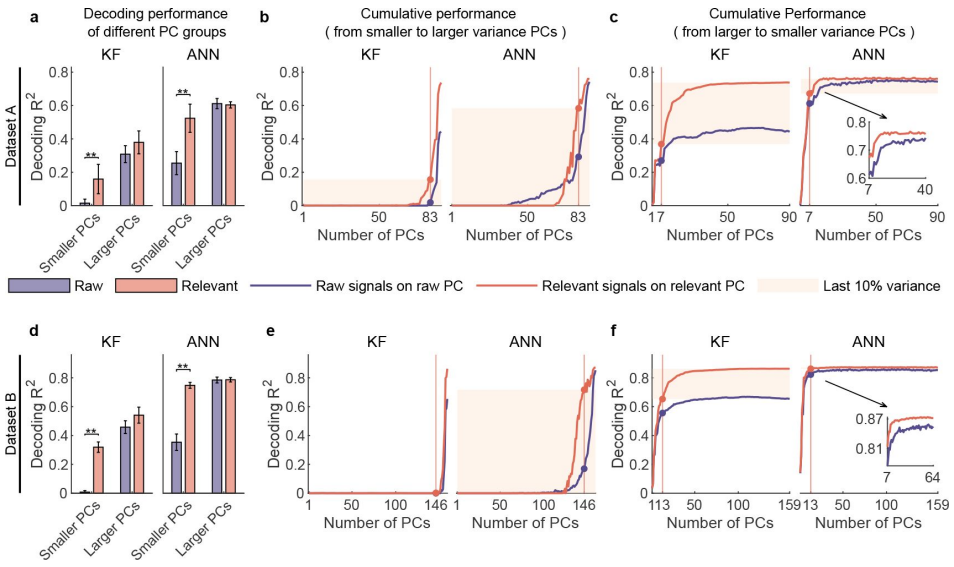


Figure 6.

Signals composed of smaller variance PCs encode rich behavioral information in complex nonlinear ways.

a, The comparison of decoding performance between raw (purple) and distilled signals (red) composed of different raw PC groups, including smaller variance PCs (the proportion of irrelevant signals that make up raw PCs is higher than that of relevant signals), larger variance PCs (the proportion of irrelevant signals is lower than that of relevant ones) on dataset A. Error bars indicate mean ± s.d. across five cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{*}P < 0.05$, $^{**}P < 0.01$. **b**, The cumulative decoding performance of signals composed of cumulative PCs that are ordered from smaller to larger variance using KF (left) and ANN (right) on dataset A. The red patches indicate the decoding ability of the last 10% variance of relevant signals. **c**, The cumulative decoding performance of signals composed of cumulative PCs that are ordered from larger to smaller variance using KF (left) and ANN (right) on dataset A. The red patches indicate the decoding gain of the last 10% variance signals of relevant signals superimposing on their top 90% variance signals. The inset shows the partially enlarged plot for view clearly. **d-f**, Same as **a-c**, but for dataset B.

We then investigated the effect of superimposing secondary signals on primary signals by accumulating PCs from larger to lower variance. The results (**Fig. 6c,f**) and **Supplementary Fig. S10f**) show that secondary signals improve the decoding performance of ANN a little but improve the decoding performance of KF a lot. The discrepancy between the two decoders reflects their different abilities to utilize the information within the signal. KF cannot use the nonlinear information in primary signals as effectively as ANN can and thus requires secondary signals to improve decoding performance. Notably, the improvement of KF using secondary signals is significantly higher than using raw signals composed of the same number of dimensions as the secondary signals ($P < 0.01$, Wilcoxon rank-sum test). These results demonstrate that these smaller variance PC signals actually encode behavioral information, and irrelevant signals make their encoded information diffcult to be linearly decoded, suggesting that behavioral information exists in a higher-dimensional sub-space than anticipated from raw signals. Interestingly, we can find that although secondary signals nonlinearly encode behavioral information and are decoded poorly by linear decoders, they considerably improve KF performance by superimposing on primary signals (**Fig. 6c,f**) and **Supplementary Fig. S10f**) left); and the sum of the sole decoding performance of primary and secondary signals is lower than the decoding performance of full signals. These results indicate that the combination of smaller and larger variance PCs produces a synergy effect (***Narayanan et al., 2005***) enabling secondary signals that cannot be linearly decoded to improve the linear decoding performance.

Finally, considering the substantial enhancement in KF decoding performance when superimposing the secondary signals on the primary ones, we explored which aspect of movement parameters was improved. In BMIs, directional control has achieved great success (***Georgopoulos et al., 1986***; ***Hochberg et al., 2012***), but precise speed control, especially at lower speeds such as hold or stop, has always been challenging (***Wodlinger et al., 2014***; ***Inoue et al., 2018***). Thus we hypothesized that these signals might improve the lower-speed velocity. To test this, we divided samples into lower-speed and higher-speed regions and assessed which region improved the most by superimposing the secondary signals (see details in Methods). After superimposing the secondary signals, the absolute improvement ratio of the lower-speed region is significantly higher than that of the higher-speed region ($P < 0.05$, Wilcoxon rank-sum test; Supplementary Fig.S12a,b,c). Furthermore, we visualized the relative improvement ratio of five example trials for the two regions, and the results (**Supplementary Fig. S12d**) demonstrate that secondary signals significantly improve the estimation of lower speed. These results demonstrate that the secondary signals enhance the lower-speed control, suggesting that smaller variance PC signals are involved in regulating precise motor control.

Distilled behaviorally-relevant signals suggest that motor cortex may use a linear readout mechanism to generate movement behaviors

Understanding the readout mechanism of the motor cortex is crucial for both neuroscience and neural engineering, which remains unclear. By filtering out the interference of behaviorally-irrelevant signals, we found a stunning result: the linear decoder KF achieves comparable performance to that of the nonlinear decoder ANN ($P = 0.10$, 0.15, and 0.55 for datasets A, B, and C, Wilcoxon rank-sum test; **Fig. 2b,f,g**). Considering the decoding performance and model complexity (the simplicity principle, also called Occam's razor), movement behaviors are more likely to be generated by the linear readout, suggesting linear readout is performed in the motor cortex.

Given the significant improvement in linear decoding performance, one might doubt that it is our distillation model that makes signals that are inherently nonlinearly decodable become linearly decodable. In practice, this situation does not hold for two reasons. Firstly, if such a scenario were to occur, given that the signals extracted by d-VAE exhibit significantly different linear decoding performance compared to the ground truth, this difference would be reflected at the signal level,

and theoretically, this difference should be substantial enough to be recognizable, rather than so minor as to be imperceptible. Consequently, this would result in lower neural similarity between these signals and the ground truth, as well as raw signals, and a substantial amount of uncharacterized signals of the ground truth signals would be left within the residuals, leading to the residuals containing a considerable amount of information. In this scenario, these signals would be effectively excluded by our criteria. To clarify it, let's consider an example. Suppose $z = x + y = n^2 + y$, where z, x, y, and n represent raw signals, relevant signals, irrelevant signals, and latent variables, respectively. If the distilled relevant signals are $xA = n$, then these signals can be linear decoded accurately. However, the corresponding irrelevant signals are $n^2 - n + z$; thus, irrelevant signals will have much information, and these extracted relevant signals will not be selected. Secondly, our synthetic experiments offer additional evidence supporting the conclusion that d-VAE does not make inherently nonlinearly decodable signals become linearly decodable ones. As depicted in **Fig. S11c**, there exists a significant performance gap between KF and ANN when decoding the ground truth signals of smaller $\mathbf{R}^2$ neurons ($P < 0.01$, Wilcoxon rank-sum test). KF exhibits notably low performance, leaving substantial room for compensation by d-VAE. However, following processing by d-VAE, KF's performance of distilled signals fails to surpass its already low ground truth performance and remains significantly inferior to ANN's performance ($P < 0.01$, Wilcoxon rank-sum test). These results collectively confirm that our approach does not convert signals that are inherently nonlinearly decodable into linearly decodable ones.

In summary, these findings demonstrate that behaviorally-irrelevant signals significantly complicate the readout of behavioral information and provide compelling evidence supporting the notion that the motor cortex uses a linear readout mechanism to generate movement behaviors.

Discussion

In this study, we proposed a new perspective for studying neural mechanisms, namely, using separated accurate behaviorally-relevant signals instead of raw signals; and we provided a novel distillation framework to define, extract, and validate behaviorally-relevant signals. By separating behaviorally-relevant and irrelevant signals, we found that neural responses previously considered useless encode rich behavioral information in complex nonlinear ways, and they play an important role in neural encoding and decoding. Furthermore, we found that linear decoders can achieve comparable performance to that of nonlinear decoders, providing compelling evidence for the presence of linear readout in the motor cortex. Overall, our results reveal unexpected complex encoding but simple decoding mechanisms in the motor cortex.

Signal separation by d-VAE

Behaviorally-relevant patterns can be extracted either at single-neuron or latent neural population levels. In our study, we focused on the single-neuron level, aiming to preserve the underlying properties of individual neurons. By maintaining the properties of each neuron, researchers can investigate how the neuronal population performs when one of the neurons is destroyed. This kind of analysis is particularly useful in closed-loop stimulation experiments that use electrophysiological (***Sun et al., 2022***) or optogenetic (***Zhang et al., 2022***) interventions. Furthermore, behaviorally-relevant signals also allow for population-level analysis and provide clean benchmark signals to test and compare the variance capture ability of different hypothesis-driven models.

At the single-neuron level, it is common practice to use trial-averaged neuronal responses of the same task parameters to analyze neural mechanisms (***Kobak et al., 2016***; ***Rouse and Schieber, 2018***). However, trial averaging sacrifices single-trial information, thereby providing an incomplete characterization of neural activity. Furthermore, trial-averaged responses still contain a significant amount of behaviorally-irrelevant signals caused by uninstructed movements (***Musall***

et al., 2019), which can lead to a contaminated version of behaviorally-relevant signals. In contrast, our model is capable of extracting clean behaviorally-relevant neural activity for every single trial. At the latent population level, existing latent variable models (*Sani et al., 2021*; *Pandarinath et al., 2018*; *Yu et al., 2008*; *Zhou and Wei, 2020*) focus on modeling some specific properties of latent population representations, such as linear or nonlinear dynamics (*Sani et al., 2021*; *Pandarinath et al., 2018*; *Churchland et al., 2012*), temporal smoothness (*Yu et al., 2008*), and interpretability (*Zhou and Wei, 2020*). Since these models make restrictive assumptions involving characterizing specific neural properties, they fail to capture other parts of behaviorally-relevant signals that do not meet their assumptions, providing no guarantee that the generated signals preserve behavioral information maximally. In contrast, our objective is to extract accurate behaviorally-relevant signals that closely approximate the ground truth relevant signals as much as possible. To ensure this, we deliberately impose constraints on the model, ensuring that it generates signals that retain neuronal properties while preserving behavioral information to the highest degree possible. Notably, the pivotal operation of striking a balance between the reconstruction and decoding performance of generated signals to extract relevant signals is a distinctive feature absent in other models. At the population level, dimensionality reduction methods aided by task parameters (*Kobak et al., 2016*; *Schneider et al., 2023*) are another important way to discover the latent neural embeddings relevant to task parameters, which may provide new insight into neural representations. In contrast with this class of methods, our model focuses on the signal level, not the latent embedding level.

Although we made every effort, our model is still not able to perfectly extract behaviorally-relevant neural signals, resulting in a small amount of behavioral information leakage in the residuals. Nevertheless, the signals distilled by our model are reliable, and the minor imperfections do not affect the conclusions drawn from our analysis. In the future, better models can be developed to extract behaviorally-relevant signals more accurately, such as incorporating multiple time-step information (*Pandarinath et al., 2018*; *Sani et al., 2021*; *Hurwitz et al., 2021*) and contrastive learning (*Schneider et al., 2023*) or metric learning (*Li et al., 2021*) techniques into models.

Implications for analyzing neural activity by separation

Studying neural mechanisms through noisy signals is akin to looking at flowers in a fog, which is diffcult to discern the truth. Thus, removing the interference of irrelevant signals is necessary and beneficial for analyzing neural activity, whether at the single-neuron level or population level.

At the single-neuron level, trial-to-trial neuronal variability poses a significant challenge to identifying the actual neuronal pattern changes. The variability can arise from various sources, including meaningless noise (*Faisal et al., 2008*), meaningful but behaviorally-irrelevant neural processes (*Musall et al., 2019*), and intrinsic components of neural encoding (*Walker et al., 2020*). However, it is still unclear to what extent each source contributes to the variability (*Faisal et al., 2008*). By separating behaviorally-relevant and irrelevant parts, we could roughly determine the extent to which these two parts contribute to the variability and explore which type of variability these two parts may contain. Our results demonstrate that behaviorally-irrelevant signals are a significant contributor to variability, which may include both meaningless noise and meaningful but behaviorally-irrelevant signals as behaviorally-irrelevant signals are not pure noise and may carry some structures (thick blue line, **Fig. 4b,f** and **Supplementary Fig. S6b**). Notably, behaviorally-relevant signals also exhibit some variability, which may arise from intrinsic components of neural encoding and provide the neural basis for motor learning (*Dhawale et al., 2017*). Moreover, eliminating the variability caused by irrelevant signals enables us to better observe and compare actual neuronal pattern changes and may facilitate the study of learning mechanisms (*Sadtler et al., 2014*; *Hennig et al., 2021*).

At the population level, determining the neural dimensionality of specific behaviors from raw signals is challenging since it is diffcult to discern how many variances correspond to irrelevant signals, which often depend heavily on signal quality. A previous study (***Altan et al., 2021***) demonstrated, through simulation experiments involving different levels of noise, that such noise makes methods overestimate the neural dimensionality. Our results, consistent with theirs, indicate that using raw signals which includes many irrelevant signals will cause an overestimation of the neural dimensionality (**Supplementary Fig. S9**). These findings highlight the need to filter out irrelevant signals when estimating the neural dimensionality. It is important to note that the dimensionality of neural subspaces encoding behavioral information cannot be accurately determined by the variance they capture. Estimating the dimensionality accurately requires considering the amount of behavioral information encoded by signals distributed within these subspaces. This is because the variance of neural signals does not directly reflect the amount of behavioral information they encode. For example, behaviorally-irrelevant signals occupy some neural subspaces without containing behavioral information. Furthermore, this perspective of signal separation has broader implications for other studies. For instance, researchers can isolate neural signals corresponding to different behaviors and explore their shared and exclusive patterns to uncover underlying common and unique mechanisms of different behaviors (***Gallego et al., 2018***).

Implications for exploring neural mechanisms by separation

At the single-neuron level, previous studies (***Carmena et al., 2005***; ***Narayanan et al., 2005***) have shown that neuronal ensembles redundantly encode movement behaviors in the motor cortex. However, our results reveal a significantly higher level of redundancy than previously reported. Specifically, prior studies found that the decoding performance steadily declines as neurons drop out, which is consistent with our results drawn from raw signals. In contrast, our results show that decoders maintain high performance on distilled signals even when many neurons drop out. Our findings reinforce the idea that movement behavior is redundantly encoded in the motor cortex and demonstrate that the brain is robust enough to tolerate large-scale neuronal destruction while maintaining brain function (***Alstott et al., 2009***).

At the population level, previous studies have proposed that motor control is achieved through low-dimensional neural manifolds (***Cunningham and Yu, 2014***; ***Gallego et al., 2017***). However, our results challenge this idea by showing that signals composed of smaller variance PCs nonlinearly encode a significant amount of behavioral information. This suggests that behavioral information is distributed in a higher-dimensional neural space than previously thought. Interestingly, although smaller variance PC signals nonlinearly encode behavioral information, their behavioral information can be linearly decoded by superimposing them onto larger variance PC signals. This result is consistent with the finding that nonlinear mixed selectivity can yield high dimensional neural responses and thus allow linear readout of behavioral information by downstream neurons (***Rigotti et al., 2013***; ***Fusi et al., 2016***). Moreover, we found that smaller variance PC signals can improve precise motor control, such as lower speed control. Analogously, recent studies have found that smaller variance PCs of hand postures are task-dependent and relate to the precise and complex postures (***Yan et al., 2020***). These findings suggest that neural signals composed of lower variance PCs are involved in the regulation of precise motor control.

In the motor cortex, in what form downstream neurons read out behavioral information is still an open question. Previous studies have shown that nonlinear readout is superior to linear read-out on raw signals (***Naufel et al., 2019***; ***Glaser et al., 2020***; ***Willsey et al., 2022***). However, by filtering out the interference of behaviorally-irrelevant signals, our study found that accurate decoding performance can be achieved through linear readout, suggesting that the motor cortex may perform linear readout to generate movement behaviors. Similar findings have been reported in other cortices, such as the inferotemporal cortex (IT) (***Majaj et al., 2015***), perirhinal cortex (PRH) (***Pagan et al., 2013***), and somatosensory cortex (***Nogueira et al., 2023***), supporting the idea of linear readout as a general principle in the brain. However, further

experiments are needed to verify this hypothesis across a wider range of cortical regions. While neurons encode significant behavioral information nonlinearly, our study demonstrates that when neuronal assemblies with diverse firing patterns collaborate to decode information, they do not necessarily need to fully exert their nonlinear ability. Instead, a simple linear combination of these assemblies can accomplish the decoding task. This finding suggests that the complexity of encoding mechanisms underlies the simplicity of decoding mechanisms.

With regard to studying decoding mechanisms, recent studies (***Pitkow et al., 2015***; ***Yang et al., 2021***) have focused on investigating how the brain decodes task information in the presence of noise. Unlike their works, we focus on studying the decoding mechanism after removing the noise (irrelevant signals). Although our research perspectives differ, our results support their idea that the brain needs nonlinear operations to suppress noise interference (***Yang et al., 2021***). Since the brain contains both relevant and irrelevant signals, it is reasonable first to separate the relevant signals and then investigate their inherent encoding and decoding mechanisms. Theoretically, the process of removing irrelevant signals should not be considered part of the inherent encoding and decoding mechanisms of the relevant signals. To illustrate this concept, consider a communication system as an example: the signal that the sender transmits has its own decryption rules. However, during transmission, noise is added to the signal. Upon reception, the receiver must first denoise the signal before decrypting it. The denoising operation should not be considered as part of the signal's inherent decryption rules. Additionally, investigating how the brain filters out noise or separates behaviorally-irrelevant factors is a promising research direction (***Schneider et al., 2018***). In this regard, distilled behaviorally-relevant signals can serve as reliable reference signals to facilitate this study. Furthermore, our study reveals that irrelevant signals are the most critical factor affecting accurate and robust decoding, and achieving accurate and robust linear decoding requires weak neural responses. These findings have two important implications for developing accurate and robust BMIs: designing preprocessing filtering algorithms or developing decoding algorithms that include filtering out behaviorally-irrelevant signals, and paying attention to the role of weak neural responses in motor control. More generally, our study provides a powerful framework for separating behaviorally-relevant and irrelevant signals, which can be applied to other cortical data to uncover more hidden neural mechanisms.

Methods

Dataset and preprocessing

Three datasets with different paradigms are employed, including obstacle avoidance task dataset (***Wang et al., 2015***), center-out reaching task dataset (***Dyer et al., 2017***), and self-paced reaching task dataset (***O'Doherty et al., 2017***).

The first dataset (dataset A) is the obstacle avoidance dataset. An adult male Rhesus monkey was trained to use the joystick to move the computer cursor to bypass the obstacle and reach the target. Neural data were recorded from the monkey's upper limb area of the dorsal premotor (PMd) using a 96-electrode Utah array (Blackrock Microsystems Inc., USA). Multi-unit activity (MUA) signals are used in the present study. The corresponding behavioral data (velocity) were simultaneously collected. There are two days of data (20140106 and 20140107), and each day contains 171 trials on average. All animal handling procedures were authorized by the Animal Care Committee at Zhejiang University, China, and conducted following the Guide for Care and Use of Laboratory Animals (China Ministry of Health).

The second dataset (dataset B) is publicly available and provided by Kording Lab (***Dyer et al., 2017***). The monkey was trained to complete two-dimensional 8-direction center-out reaching tasks. We used two days of data from subject C (20161007 and 20161011). Each day contains 190

trials on average. Neural data are spike-sorted PMd signals. The behavioral data were simultaneously collected in instantaneous velocity.

The third dataset (dataset C) is publicly available and provided by Sabes Lab (Zenodo dataset) (***O'Doherty et al., 2017***). An adult male Rhesus monkey was trained to finish self-paced reaching tasks within an 8-by-8 square grid. There are no inter-trial intervals during the experiment. Neural data were recorded from the monkey's primary motor cortex (M1) area with a 96-channel silicon microelectrode array. The neural data are the multi-unit activity (MUA) signals. Hand velocity was obtained from the position through a discrete derivative. The recording period for the data (20170124 01) is about 10 minutes.

For all datasets, the neural signals were binned by a 100ms sliding window without overlap. As a preprocess, we smoothed the neural signals using a moving average filter with three bins. We excluded some electrode activities with low mean firing rates (<0.5 Hz mean firing rates across all bins) and did not perform any other pre-selection to select neurons. For the computation of the Fano factor, we chose twelve and fourteen points as the thresholds of trial length for datasets A and B, respectively; trials with a length less than the threshold were discarded (discard about 7% and 2% trials for datasets A and B), trials longer than the threshold were truncated to threshold length from the starting point. Since dataset C has no trial information, FF is not calculated for this dataset. For the analysis of datasets A and B, we selected one day of these two datasets for analysis (20140107 for dataset A and 20161011 for dataset B).

The synthetic dataset

The synthetic dataset is used to demonstrate that d-VAE can extract effective behaviorally-relevant signals that are similar to the ground truth signals. The specific process of generating synthetic data is as follows. First, we randomly selected nine larger $\mathbf{R}^2$ neurons from neurons that $\mathbf{R}^2$ is greater than 0.1, and three smaller $\mathbf{R}^2$ neurons from neurons that $\mathbf{R}^2$ is lower than 0.01 of dataset B (20161011). Second, we used deep neural networks to learn the encoding model between movement kinematics (movement velocity of dataset B) and neural signals using one fold train data. The details of the networks are demonstrated as follows. The networks use two hidden layer multilayer perceptron (MLP) with 500 and 500 hidden units. The activation function is ReLU. A SoftPlus activation function follows the last layer of the networks. The reconstruction loss is the Poisson likelihood function. After learning the encoding model, we used the learned encoding model to generate the ground truth of behaviorally-relevant signals from all kinematics data of dataset B. Then we added white Gaussian noise to the behaviorally-relevant signals such that the noisy signals have a signal-to-noise ratio (SNR) of 7dB. After adding noise, the $\mathbf{R}^2$ of the three smaller $\mathbf{R}^2$ neurons is lower than 0.03. We regarded the noisy signals as raw signals and the added Gaussian noise as behaviorally-irrelevant signals. Finally, we separated the synthetic data into five folds for cross-validation model evaluation.

Distill-VAE (d-VAE)

Notations

$\mathbf{x} \in \mathbb{R}^n$ denotes raw neural signals. $\mathbf{x}_r \in \mathbb{R}^n$ represents behaviorally-relevant signals. $\mathbf{x}_i = \mathbf{x} - \mathbf{x}_r \in \mathbb{R}^n$ represents behaviorally-irrelevant signals. $\mathbf{z} \in \mathbb{R}^d$ denotes the latent neural representations. $\mathbf{z}_{prior} \in \mathbb{R}^d$ denotes the prior latent neural representations. $\mathbf{y} \in \mathbb{R}^k$ represents kinematics. $f\ \mathbb{R}^n\ \mathbb{R}^d$ represents the inference model (encoder) of VAE. $g\ \mathbb{R}^d\ \mathbb{R}^n$ represents the generative model (decoder) of VAE. $m\ \mathbb{R}^k\ \mathbb{R}^d$ represents the mapping from kinematics to prior latent representations. $h\ \mathbb{R}^d\ \mathbb{R}^k$ represents an affne mapping from latent representations to kinematics.

d-VAE is a generative model based on variational autoencoders (VAEs) (***Kingma and Welling, 2013***), specially designed to extract behaviorally-relevant signals from raw signals. The generative model of d-VAE is

$$p_\theta(\mathbf{x} \mid \mathbf{y}) = \int_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z} \mid \mathbf{y}) d\mathbf{z} = \int_{\mathbf{z}} p_m(\mathbf{z} \mid \mathbf{y}) p_g(\mathbf{x} \mid \mathbf{z}) d\mathbf{z}, \tag{1}$$

where $p_m(\mathbf{z} \mid \mathbf{y})$ denotes the conditional prior distribution of latent variables given the kinematics parameterized by feedforward neural networks m, $p_g(\mathbf{x} \mid \mathbf{z})$ denotes the conditional prior distribution of raw signals given the latent variables parameterized by feedforward neural networks g, $p_\theta(\mathbf{x}, \mathbf{z} \mid \mathbf{y})$ represents the joint distribution of raw signals and latent variables given the kinematics parameterized by parameters $\theta = (m, g)$, and $p_\theta(\mathbf{x} \mid \mathbf{y})$ is the marginal distribution of raw signals parameterized by parameters θ.

To learn the model, we need to maximize the evidence lower bound (ELBO) of $p_\theta(\mathbf{x} \mid \mathbf{y})$:

$$\mathcal{L}_{ELBO}(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})} \left[p_\theta(\mathbf{x} \mid \mathbf{z}) \right] - D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \mid p(\mathbf{z} \mid \mathbf{y})) \leq \log p_\theta(\mathbf{x} \mid \mathbf{y}), \tag{2}$$

where the first term of right hand side of Eq. 2 is called reconstruction term, the second term is called regularization term, $q_\varphi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ denotes inference model parameterized by parameters φ, and $D_{KL}(\cdot \mid \cdot)$ denotes the Kullback-Leibler divergence. In d-VAE, we set $q_\varphi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) = q_f(\mathbf{z} \mid \mathbf{x})$, where f is the inference model parameterized by feedforward neural networks. Because during the test stage, we cannot obtain the ground truth of kinematics, and we need to use only raw signals to extract relevant signals. Note that d-VAE aims to extract behaviorally-relevant signals from raw signals, not generate signals that are too similar to raw signals. Therefore, we modified the objective loss function based on $\mathcal{L}_{ELBO}(\mathbf{x})$ (see Eq. 8).

To distill behaviorally-relevant neural signals, d-VAE utilizes the trade-off between the decoding and reconstruction abilities of generated behaviorally-relevant signals $\mathbf{x}_r$. The basic assumption is generated behaviorally-relevant signals that contain behaviorally-irrelevant signals harms their decoding ability. Our approach for distilling behaviorally-relevant signals consists of three steps, including identifying latent representations $\mathbf{z}$, generating behaviorally-relevant signals $\mathbf{x}_r$, and decoding behaviorally-relevant signals $\mathbf{x}_r$.

Identfiying latent representations

Identifying latent representations containing behaviorally-relevant information is the crucial part because latent representations influences the subsequent generation. Effective representations are more likely to generate proper behaviorally-relevant neural signals $\mathbf{x}_r$. d-VAE identifies latent representations with inference model f, that is, $[\boldsymbol{\mu}; \boldsymbol{\sigma}^2] = f(\mathbf{x})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ denote the mean and variance of latent representations; thus the posterior distribution is $q_f(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mathbf{z} \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Then we guide latent representations containing behavioral information through an affne map $h \in^d \in^k$ under the loss $\mathcal{L}_{dec1}$,

$$\mathcal{L}_{dec1} = \mathrm{MSE}(h(\boldsymbol{\mu}), \mathbf{y}), \tag{3}$$

where $\mathrm{MSE}(\cdot, \cdot)$ denotes mean squared loss. In other words, we encourage latent representations to decode kinematics to distill behaviorally-relevant information. Here we sample from the approximation posterior $\mathbf{z} \sim q_f(\mathbf{z} \mathrm{I} \mathbf{x})$ using $\mathbf{z} \sim \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ denotes element-wise product. This sampling strategy is known as the reparameterization trick.

Generating behaviorally-relevant signals

After sampling latent representations $\mathbf{z}$, we send latent representations to the generative model g to generate behaviorally-relevant neural signals $\mathbf{x}_r$, that is, $\mathbf{x}_r = g(\mathbf{z})$. We use following loss to make behaviorally-relevant signals reconstruct raw signals as much as possible:

$$\mathcal{L}_{rec} = \text{Poisson}(\mathbf{x}_r, \mathbf{x}), \tag{4}$$

where Poisson(·, ·) denotes Poisson negative log likelihood loss. It is important to note that optimizing the generation of behaviorally-relevant signals to accurately reconstruct noisy raw signals may result in the inclusion of many behaviorally-irrelevant signals in the generated signals, which deviates from our initial goal of extracting behaviorally-relevant signals. In the following subsection, we will introduce how to avoid generating behaviorally-irrelevant signals.

Decoding behaviorally-relevant signals

As mentioned above, if the generation of behaviorally-relevant signals $\mathbf{x}_r$ is only guided by $\mathcal{L}_{rec}$, generated signals may contain more behaviorally-irrelevant signals. To avoid generated signals containing behaviorally-irrelevant signals, we introduce decoding loss $\mathcal{L}_{dec2}$ to constrain $\mathbf{x}_r$ to decode behavioral information. The basic assumption is that behaviorally-irrelevant signals act like noise for decoding behavioral information and are detrimental to decoding. Thus, there is a trade-off between neural reconstruction and decoding ability of $\mathbf{x}_r$.: the more behaviorally-irrelevant signals $\mathbf{x}_r$ contains, the more decoding performance $\mathbf{x}_r$ loses. Then we send the $\mathbf{x_r}$ to the encoder f and obtain the mean and variance of latent representations, that is $[\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r^2] = f(\mathbf{x}_r)$. The decoding loss $\mathcal{L}_{dec2}$ is as follows:

$$\mathcal{L}_{dec2} = \text{MSE}(h(\boldsymbol{\mu}_r), \mathbf{y}). \tag{5}$$

We use the same networks f and h for $\mathbf{x}_r$ and $\mathbf{x}$ in our experiment, because $\mathbf{x}_r$ can act as data augmentation and make f distill robust representations without increasing model parameters. In addition, we combine the two decoding loss as one loss:

$$\mathcal{L}_{dec} = \frac{1}{2}(\mathcal{L}_{dec1} + \mathcal{L}_{dec2}). \tag{6}$$

Learning the prior distribution with behavioral information

The prior distribution of latent representation is crucial because inappropriate prior assumptions can degrade latent representations and generated neural signals. Vanilla VAE uses a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to regularize the space of latent representation $\mathbf{z}$. However, in neuroscience, the distribution of latent representations is unknown and may exceed the scope of Gaussian. Therefore we adopt neural networks m to learn the prior distribution with kinematics $\mathbf{y}$, that is, $[\boldsymbol{\mu}_{prior}, \boldsymbol{\sigma}_{prior}^2] = m(\mathbf{y})$ and thus $p_m(\mathbf{z} \mid \mathbf{y}) = \mathcal{N}(\mathbf{z}_{prior} \mid \boldsymbol{\mu}_{prior}, \boldsymbol{\sigma}_{prior}^2)$. The prior distribution $p_m(\mathbf{z} \mid \mathbf{y})$ and approximation posterior distribution $q_f(\mathbf{z} \mid \mathbf{x})$ are aligned by the Kullback–Leibler divergence:

$$\mathcal{L}_{KL} = D_{KL}(p_m(\mathbf{z} \mid \mathbf{y}) \mid q_f(\mathbf{z} \mid \mathbf{x})) = D_{KL}(\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \mid \mathcal{N}(\mathbf{z}_{prior} \mid \boldsymbol{\mu}_{prior}, \boldsymbol{\sigma}_{prior}^2)). \tag{7}$$

Since $\mathbf{z}_{prior}$ and $\mathbf{z}$ are aligned and the generative network g models the relationship between $\mathbf{z}$ and $\mathbf{x}_r$, this is equivalent to indirectly establishing a neural encoding model from $\mathbf{y}$ to $\mathbf{x}_r$. Thus, we can observe the change of $\mathbf{z}_{prior}$ and $\mathbf{x}_r$ by changing $\mathbf{y}$ and can better understand the encoding mechanism of neural signals.

End-to-end optimization

d-VAE is optimized in an end-to-end manner under the following loss:

$$\mathcal{L} = \mathcal{L}_{rec} + \beta \mathcal{L}_{KL} + \alpha \mathcal{L}_{dec}, \tag{8}$$

where β and α are hyperparameters, β is used to adjust the weight of KL divergence, and α determines the trade-off between reconstruction loss $\mathcal{L}_{rec}$ and decoding loss $\mathcal{L}_{dec}$. Given that the ground truth of latent variable distribution is unknown, even a learned prior distribution might not accurately reflect the true distribution. We found the pronounced impact of the KL divergence would prove detrimental to the decoding and reconstruction performance. As a result, we opt to reduce the weight of the KL divergence term. Even so, KL divergence can still effectively align the distribution of latent variables with the distribution of prior latent variables (see **Fig. S13**).

In the training stage, we feed raw neural signals into the inference network f to get latent representation $\mathbf{z}$, which is regularized by $\mathbf{z}_{prior}$ coming from kinematics $\mathbf{y}$ and network m. Then we use the mean of $\mathbf{z}$, that is $\boldsymbol{\mu}$, to decode kinematics $\mathbf{y}$ by affne layer h and send $\mathbf{z}$ to the generative networks g to generate neural signals $\mathbf{x}_r$. To ensure that $\mathbf{x}_r$ preserves decoding ability, we send $\mathbf{x}_r$ to f and h to decode $\mathbf{y}$. The whole model is trained in an end-to-end manner under the guidance of total loss. Once the model has been trained, we can feed raw neural signals to it to obtain behaviorally-relevant neural signals $\mathbf{x}_r$, and we can aslo generate behaviorally-relevant neural signals using the prior distribution of $\mathbf{z}_{prior}$.

The identifiability of d-VAE

Recent studies (***Khemakhem et al., 2020***) show that to obtain the identifiability of latent variables, we should either restrict the structures of generating models or introduce some additional constraints on the distribution of latent variables and show that the additive noise models are identifiable under certain assumptions. pi-VAE (***Zhou and Wei, 2020***) has proved that pi-VAE is identifiable under some assumptions. Since d-VAE is a straightforward extension of pi-VAE, it also uses auxiliary variables (kinematics) to constrain the latent variables as pi-VAE did; with the same assumptions, d-VAE is also identifiable.

Differences from pi-VAE

pi-VAE lacks the decoding constraint on latent variables (Eq. 3) and the decoding constraint on generated signals (Eq. 5).

Cross-validation evaluation of models

For each model, we use the five-fold cross-validation manner to assess performance. Specifically, we divide the data into five equal folds. In each experiment, we take one fold for testing and use the rest for training (taking three folds as the training set and one as the validation set). The reported performance is averaged over test sets of five experiments. The validation sets are used to choose hyperparameters based on the averaged performance of five-fold validation data. To avoid overfitting, we apply the early stopping strategy. Specifically, we assess the criteria (loss for training distillation methods, $\mathbf{R}^2$ for training ANN and KF) on the validation set every epoch in the training process. The model is saved when the model achieves better validation performance than the earlier epochs. If the model cannot increase by 1% of the best performance previously obtained within 30 epochs, we stop the training process.

The strategy for selecting effective behaviorally-relevant signals

As previously mentioned, the hyperparameter α of d-VAE plays a crucial role in balancing the trade-off between reconstruction and decoding loss. Once the appropriate value of α is determined, we can use this value to obtain accurate behaviorally-relevant signals for subsequent analysis. To determine the optimal value of α, we first enumerated different values of α to guide d-VAE in distilling the behaviorally-relevant signals. Next, we used ANN to evaluate the decoding $\mathbf{R}^2$ of behaviorally-relevant (D_{re}) and irrelevant (D_{ir}) signals generated by each α value. Finally, we selected the α value with the criteria formula $0.75 \times D_{re} + 0.25 \times (1 - D_{ir})$. The α value that obtained the highest criteria score on the validation set was selected as the optimal value.

Note that we did not use neural similarity between behaviorally-relevant and raw signals as a criterion for selecting behaviorally-relevant signals. This is because determining the threshold for neural similarity is challenging. However, not using similarity as a criterion does not affect the selection of suitable signals because the decoding performance of behaviorally-irrelevant signals can indirectly reflect the degree of similarity between the generated behaviorally-relevant signals and the raw signals. Specifically, if the generated behaviorally-relevant signals are dissimilar to the raw signals, the behaviorally-irrelevant signals will contain many useful signals. In other words, when the neural similarity between behaviorally-relevant and raw signals is low, the decoding performance of behaviorally-irrelevant signals is high. Therefore, the decoding performance of irrelevant signals is a reasonable alternative to the neural similarity.

For other generative models, we iterate through a range of hyperparameters, generating the corresponding behaviorally-relevant neural signals, and subsequently evaluate these signals using ANN. The hyperparameter associated with the signals that exhibit the highest ANN decoding performance is then selected. In other words, the signals corresponding to this particular hyperparameter are chosen as the selected behaviorally-relevant neural signals.

Implementation details for methods

All the VAE-based models use the Poisson observation function. The details of different methods are demonstrated as follows:

- d-VAE. The encoder f of d-VAE uses two hidden layer multilayer perceptron (MLP) with 300 and 100 hidden units. The activation function of the hidden layers is ReLu. The dimensionality of the latent variable is set to 50. The decoder g of d-VAE is symmetric with the encoder. The last layer of the decoder is followed by a SoftPlus activation function. The prior networks m use one hidden layer MLP with 300 units. The $f3$ is set to 0.001. The a is set to 0.3, 0.4, 0.7, and 0.9 for datasets A, B, and C. We perform a grid search for a in {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}, and $f3$ in {0.001, 0.01, 0.1, 1}, and latent variable dimension in {10, 20, 50}. For the synthetic dataset A and B experiments, the $f3$ and the latent variable dimension are directly set to 0.001 and 50. For synthetic dataset the a is set to 0.9, which is grid searched in {0.01–0.09, 0.1–0.9, 1–9}, with intervals of 0.01, 0.1, and 1, respectively.
- pi-VAE. The original paper utilizes label information (as shown in Eq. 6) to approximate the posterior of the latent variable and performs Monte Carlo sampling for decoding during the test stage. However, in our signal generation setting, it is inappropriate to use label information (kinematics) for extracting behaviorally-relevant signals during the test stage. As a result, we modified the model to exclude the use of label information in approximating the posterior. The encoder, decoder, and prior networks of our pi-VAE are kept the same as those in d-VAE.
- LFADS. The hidden units of the encoder for the generator's initial conditions, the controller, the generator are set to 200, 200, 200, and 100 for datasets A, B, and C and the synthetic dataset. The dimensionality of latent factor is set to 50 for all datasets. We perform a grid search for hidden units in {100, 200}, and latent factor dimensions in {10, 20, 50}. The dimensionality of inferred inputs is set to 1. The Poisson likelihood function is used. The

training strategy follows the practice of the original paper. For datasets A and B, a trial length of eighteen is set. Trials with lengths below the threshold are zero-padded, while trials exceeding the threshold are truncated to the threshold length from their starting point. In dataset A, there are several trials with lengths considerably longer than that of most trials. We found that padding all trials with zeros to reach the maximum length (32) led to poor performance. Consequently, we chose a trial length of eighteen, effectively encompassing the durations of most trials and leading to the removal of approximately 9% of samples. For dataset B (center-out), the trial lengths are relatively consistent with small variation, and the maximum length across all trials is eighteen. For dataset C, we set the trial length as ten because we observed the video of this paradigm and found that the time for completing a single trial was approximately one second. The segments are not overlapped.

- TNDM. The hidden units of the encoder for the generator's initial conditions, the controller, and the generator are set to 64, 64, 100, and 64 for datasets A, B, and C and the synthetic dataset. The dimensionality of relevant latent factors is set to 50, 25, 50, and 50 for datasets A, B, and C and the synthetic dataset. We set the dimensionality of irrelevant latent factors as the same as that of relevant latent factors. The behavior weight is set to 5, 5, 0.2, 0.5 for datasets A, B, and C and the synthetic dataset. We perform a grid search for relevant latent factor dimensionality in {10, 25, 50}, the hidden units in {64, 100, 200}, and the behavior weight in {0.2, 0.5, 1, 5, 10}. The Poisson likelihood function is used. The other hyperparameter setting and the training strategy follow the practice of the original paper. TNDM uses the same trial length and data as LFADS.
- PSID. PSID uses several (horizon size) past neural data to predict behavior at the current time without using neural observations at the current time. For a fair comparison, we let PSID see current neural observations by shifting the neural data one sample into the past relative to the behavior data. We perform a grid search for the horizon hyperparameter in {2, 3, 4, 5, 6, 7}. Due to the relevant latent dimension should be lower than the horizon times the dimensionality of behavior variables (two-dimensional velocity in this paper), we just set the relevant latent dimension as the maximum. The horizon number of datasets A, B, C, and synthetic datasets is 7, 6, 6 and 5, respectively. And thus the latent variable dimension of datasets A, B, and C and the synthetic dataset is 14, 12, 12 and 10, respectively.
- CEBRA. We use CEBRA-behavior mode. The learning rate is set to 0.003, 0.003, 0.001, and 0.005 for datasets A, B, and C and synthetic dataset. The time offset is set to 15, 15, 10 and 15. The output dimension is set to 10 for all datasets. The number of iterations is set to 5000. The temperature is set to 1, and the temperature mode is set to auto. The batch size is set to 512. We perform a grid search for the learning rate in {0.0001, 0.0005, 0.001, 0.003, 0.005}, the time offset in {5, 10, 15, 20}, and the output dimension in {5, 10, 15, 20, 50}. The optimal hyperparameter is selected based on the lowest averaged loss of five-fold training data.
- ANN. ANN has two hidden layers with 300 and 100 hidden units. The activation function of the hidden layers is ReLu.
- KF. The matrix parameters of observation and state transition process are optimized with the least square algorithm. KF is a linear-Gaussian state-space model designed to provide an optimal estimate of the current state (kinematics in this paper). It does so by considering both the current measurement observations (neural signals in this paper) and the previous state estimate. The KF operates in a recursive and iterative manner, continually updating its state estimate as new observations become available.

Percentage of explained variance captured in a subspace

We applied PCA to behaviorally-relevant and irrelevant signals to get relevant PCs and irrelevant PCs. Then we used the percentage variance captured (also called alignment index) to quantify how many variances of irrelevant signals can be captured by relevant PCs by projecting irrelevant

signals onto the subspace composed of some relevant PCs and vice versa. The percentage of variance captured is:

$$A = \frac{\mathrm{Tr}\left(\mathbf{D}^{\mathrm{T}}\mathbf{C}\mathbf{D}\right)}{\mathrm{Tr}\left(\mathbf{C}\right)} \times 100\%, \tag{9}$$

where $\mathbf{D}\ \mathbb{R} \in^{N\times d}$ is the top d principal components (PCs) of relevant signals (irrelevant signals). $\mathbf{C}\ \mathbb{R} \in^{N\times N}$ is the covariance matrix of irrelevant signals (relevant signals), and $\mathrm{Tr}(\mathbf{C}) = \sum_{i=1}^{N} \lambda_i$, where λ_i is the i^{th} largest eigenvalue for $\mathbf{C}$. The percentage variance is a quantity between 0% and 100%.

The composition of raw signals' variance

Suppose $\mathbf{x}\ \mathbb{R} \in^{1\times T}$, $\mathbf{y}\ \mathbb{R} \in^{1\times T}$, and $\mathbf{z}\ \mathbb{R} \in^{1\times T}$ are the random variables for a single neuron of behaviorally-relevant signals, behaviorally-irrelevant, and raw signals, where $\mathbf{z} = \mathbf{x} + \mathbf{y}$, T denotes the number of samples. The composition of raw signals' variance is as follows:

$$\begin{aligned}\mathrm{Var}(\mathbf{z}) &= \mathrm{Var}(\mathbf{x}+\mathbf{y}) = \mathbb{E}\left\{[\mathbf{x}-\mathbb{E}(\mathbf{x})]+[\mathbf{y}-\mathbb{E}(\mathbf{y})]\right\}^2 \\ &= \mathbb{E}[\mathbf{x}-\mathbb{E}(\mathbf{x})]^2 + \mathbb{E}[\mathbf{y}-\mathbb{E}(\mathbf{y})]^2 + 2\mathbb{E}[\mathbf{x}-\mathbb{E}(\mathbf{x})][\mathbf{y}-\mathbb{E}(\mathbf{y})] \\ &= \mathrm{Var}(\mathbf{x}) + \mathrm{Var}(\mathbf{y}) + 2\mathrm{Cov}(\mathbf{x},\mathbf{y}),\end{aligned} \tag{10}$$

where $\mathbb{E}$, Var, and Cov denote expectation, variance, and covariance, respectively. Thus, the variance of raw signals is composed by the variance of relevant signals Var(**x**), the variance of irrelevant signals Var(**y**), and the correlation between relevant and irrelevant signals 2Cov(**x, y**). If there are neurons, calculate the composition of each neuron separately and then add it up to get the total composition of raw signals.

Reordered correlation matrix of neurons

The correlation matrix is reordered with a simple group method. The order of reordered neurons is determined from raw neural signals, which is then used for behaviorally-relevant signals.

The steps of the group method are as follows:

Step 1: We get the correlation matrix in original neuron order and set a list A that contains all neuron numbers and an empty list B.

Step 2: We select the neuron with the row number of the largest value of the correlation matrix except for the diagonal line and choose nine neurons in list A that are most similar to the selected neuron. We selected the value nine because it offers a good visualization of neuron clusters.

Step 3: Remove these selected neurons from list A, and add these selected neurons in descending order of correlation value in list B.

Step 4: Repeat steps 2 and 3 until list A is empty.

The improvement ratio of lower and higher speed regions

Split lower and higher speed regions

Since the speed ranges of different datasets are different, it is hard to determine a common speed threshold to split lower and higher speed regions. Here we used the squared error as a criterion to split the two speed regions. And for the convenience of calculating the absolute improvement ratio, we need a unified benchmark for comparison. Therefore, we use half of the total squared error as the threshold. Specifically, first, we calculated all samples' total squared error (E_p) between actual velocity and predicted velocity obtained by primary signals only. Then we enumerated the speed from 1 to 50 with a step of 0.1 and calculated the total squared error of selected samples whose speed is less than the enumerated speed. Once the total squared error of

selected samples is greater than or equal to the half total squared error of all samples ($0.5E_p$), the enumerated speed is set as the speed threshold. The samples whose speed is less than or equal to the speed threshold belong to lower speed regions, and those whose speed is greater than the speed threshold belong to higher speed regions. The squared error of the lower speed part (E_p^{low}) is approximately equal to that of the higher one E_p^{high}, that is, $E_p^{low} \approx E_p^{high} \approx 0.5E_p$ (the difference is negligible).

The absolute improvement ratio

After splitting the speed regions, we calculated the improvement of the two regions by superimposing secondary signals to primary signals, and got the squared error of lower E_{p+s}^{low} and higher E_{p+s}^{high} regions. Then we calculated the absolute improvement ratio (AIR):

$$AIR_{low} = -\frac{E_{p+s}^{low} - E_p^{low}}{E_p^{low}} \approx -\frac{E_{p+s}^{low} - 0.5E_p}{0.5E_p} \tag{11}$$

$$AIR_{high} = -\frac{E_{p+s}^{high} - E_p^{high}}{E_p^{high}} \approx -\frac{E_{p+s}^{high} - 0.5E_p}{0.5E_p} \tag{12}$$

Since the two regions refer to a common standard ($0.5E_p$), the improvement ratio of the two regions can be directly compared, and that's why we call it the absolute improvement ratio.

The relative improvement ratio

The relative improvement ratio (RIR) measures the improvement ratio of each sample relative to itself before and after superimposing secondary signals. The relative improvement ratio is computed as follows:

$$RIR^i = -\frac{E_{p+s}^i - E_p^i}{E_p^i}, \tag{13}$$

where i denotes the ith sample of test data.

Data availability

The datasets are available for research purposes from the corresponding author on reasonable request.

Code availability

The code will be publically available in github (see *https://github.com/eric0li/d-VAE*).

Acknowledgements

We would like to thank Yuxiao Yang, and Huaqin Sun for valuable discussions. This work was partly supported by grants from the National Key R&D Program of China (2018YFA0701400), Key R&D Program of Zhejiang (no. 2022C03011), the Chuanqi Research and Development Center of Zhejiang University, the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-002), the Fundamental Research Funds for the Central Universities.

Author contributions

Y.G.L. conceived the study, designed the experiments, analyzed the data, and wrote and modified the manuscript. X.Y.Z. designed correlation analysis experiments and modified the manuscript. Y.M.W. and Y.Q. supervised this study.

Competing interests

The authors declare no competing interests.

Supplementary Information

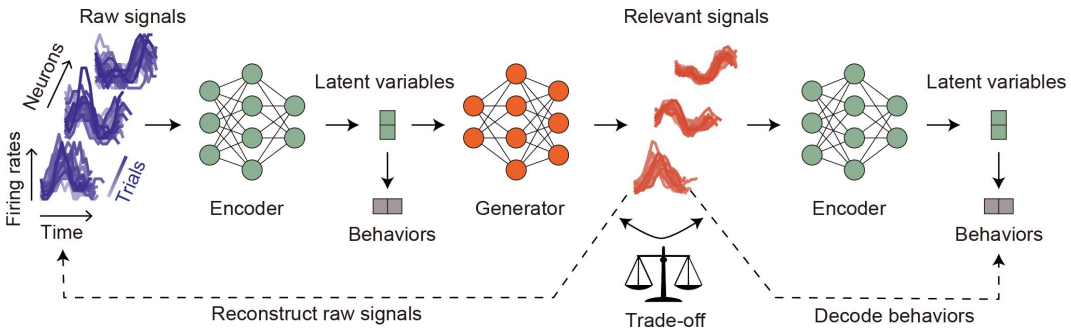


Figure S1.

Semantic overview of distill-VAE (d-VAE).

On the left, we present a set of raw signal examples (depicted by purple lines). The input neural signals fed into the d-VAE consist of single time step samples. Initially, the encoder compresses these input signals into latent variables. We constrain latent variables to decode behaviors to preserve behavioral information. Subsequently, these latent variables are transmitted to the generator, which produces behaviorally-relevant signals (depicted by red lines). To maintain the underlying neuronal properties, we constrain these generated signals to resemble the raw signals closely. At this juncture, relying solely on the constraint for signal resemblance to raw signals makes it challenging todetermine the extent of irrelevant signals present within the generated relevant signals. To tackle this hurdle, we introduce the generated signals back into the encoder. We then impose constraints on the resultant latent variables to decode behaviors. This approach is rooted in the assumption that irrelevant signals function as noise relative to relevant signals. Consequently, an excessive presence of irrelevant signals within the generated signals would lead to a degradation in their decoding performance. In essence, there exists a trade-off relationship between the decoding performance and the reconstruction performance of the generated signals. By striking a balance between these two constraints, we can effectively extract behaviorally-relevant signals.

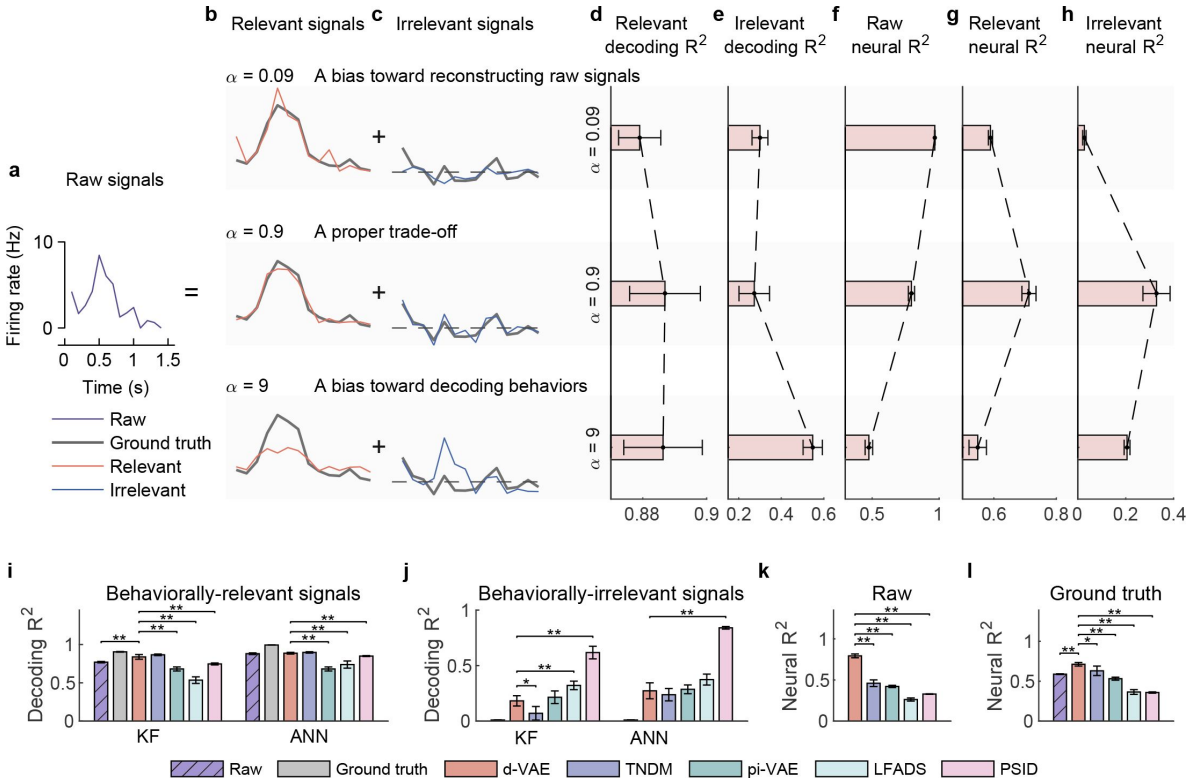


Figure S2.

Evaluation of separated signals on the synthetic dataset.

a, The temporal neuronal activity of raw signals (the purple line) of an example test trial, which is decomposed into relevant (**b**) and irrelevant (**c**) signals. **b**, Relevant signals (red lines) extracted by d-VAE under three distillation cases, where bold gray lines represent ground truth relevant signals. The hyperparameter α is very important to extracting behaviorally-relevant signals, which balances the trade-off between reconstruction loss and decoding loss. Results show that when α = 0.09, the relevant signals are too similar to raw signals but not similar to ground truth; when α = 0.9, the relevant signals are well similar to the ground truth; when α = 9, the relevant signals are not similar to the ground truth. **c**, Same as **b**, but for irrelevant signals (blue lines). Notably, when α = 9, some useful signals are left in irrelevant signals. **d**, The decoding $\mathbf{R}^2$ of distilled relevant signals of three cases. Error bars indicate mean ± s.d. across five cross-validation folds. Results demonstrate that decoding $\mathbf{R}^2$ increases as α increases. **e**, Same as **d**, but for irrelevant signals. Notably, when α = 9, irrelevant signals will contain large behavioral information. **f**, The neural similarity between relevant and raw signals. Results show that the neural $\mathbf{R}^2$ decreases as α increases. **g**, The neural $\mathbf{R}^2$ between relevant signals and the ground truth of relevant signals. Results show that d-VAE can utilize a proper trade-off to extract effective relevant signals that are similar to the ground truth. **h**, The neural $\mathbf{R}^2$ between irrelevant signals and the ground truth of irrelevant signals. Results show that d-VAE can utilize a proper trade-off to remove effective irrelevant signals that are similar to the ground truth. **i**, The decoding $\mathbf{R}^2$ between true velocity and predicted velocity of raw signals (purple bars with slash lines), the ground truth signals (gray) and behaviorally-relevant signals obtained by d-VAE (red), PSID (pink), pi-VAE (green), TNDM (blue), and LFADS (light green). Error bars denote mean ± standard deviation (s.d.) across five cross-validation folds. Asterisks represent significance of Wilcoxon rank-sum test with $^{*}P < 0.05$, $^{**}P < 0.01$. **j**, Same as **i**, but for irrelevant signals. **k**, The neural $\mathbf{R}^2$ between generated relevant signals and raw signals. **l**, Same as **k**, but for the ground truth of relevant signals.

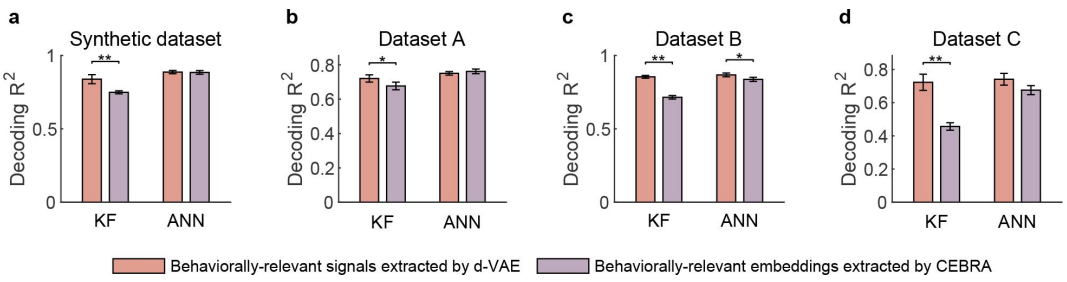


Figure S3.

Decoding performance comparison with CEBRA.

a, The decoding $\mathbf{R}^2$ comparison between d-VAE and CEBRA on synthetic dataset. The red bar represents the behaviorally-relevant signals extracted by d-VAE, and the light purple bar represents the behaviorally-relevant embeddings extracted by CEBRA. Error bars indicate mean ± s.d. across five cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{*}P < 0.05$, $^{**}P < 0.01$. **b-d**, Same as **a**, but for datasets A, B, and C, respectively.

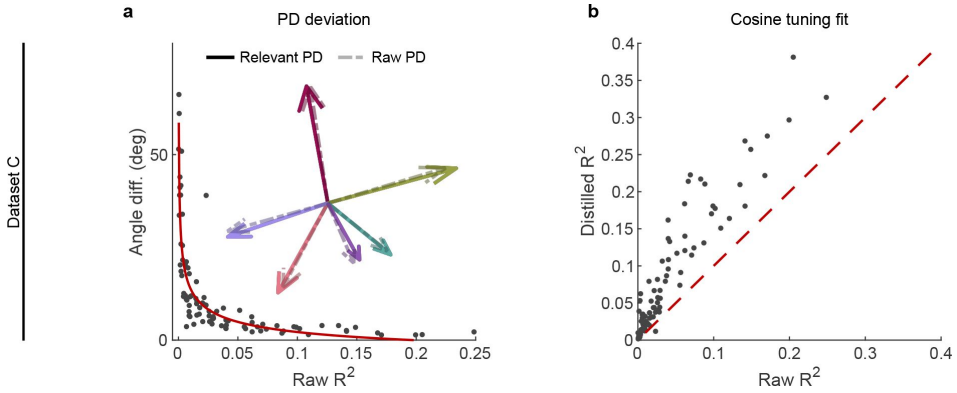


Figure S4.

The effect of irrelevant signals on relevant signals at the single-neuron level.

a,**b** Same as **Fig. 3**, but for dataset C. **a**, The angle difference (AD) of preferred direction (PD) between raw and distilled signals as a function of the $\mathbf{R}^2$ of raw signals. Each black point represents a neuron (n=91). The red curve is the fitting curve between $\mathbf{R}^2$ and AD. Five example larger $\mathbf{R}^2$ neurons' PDs are shown in the inset plot, where the solid line arrows represent the PD of relevant signals, and the dotted line arrows represent the PDs of raw signals. **b**, Comparison of the cosine tuning fit ($\mathbf{R}^2$) before and after distillation of single neurons (black points), where the x-axis and y-axis represent neurons' $\mathbf{R}^2$ of raw and distilled signals.

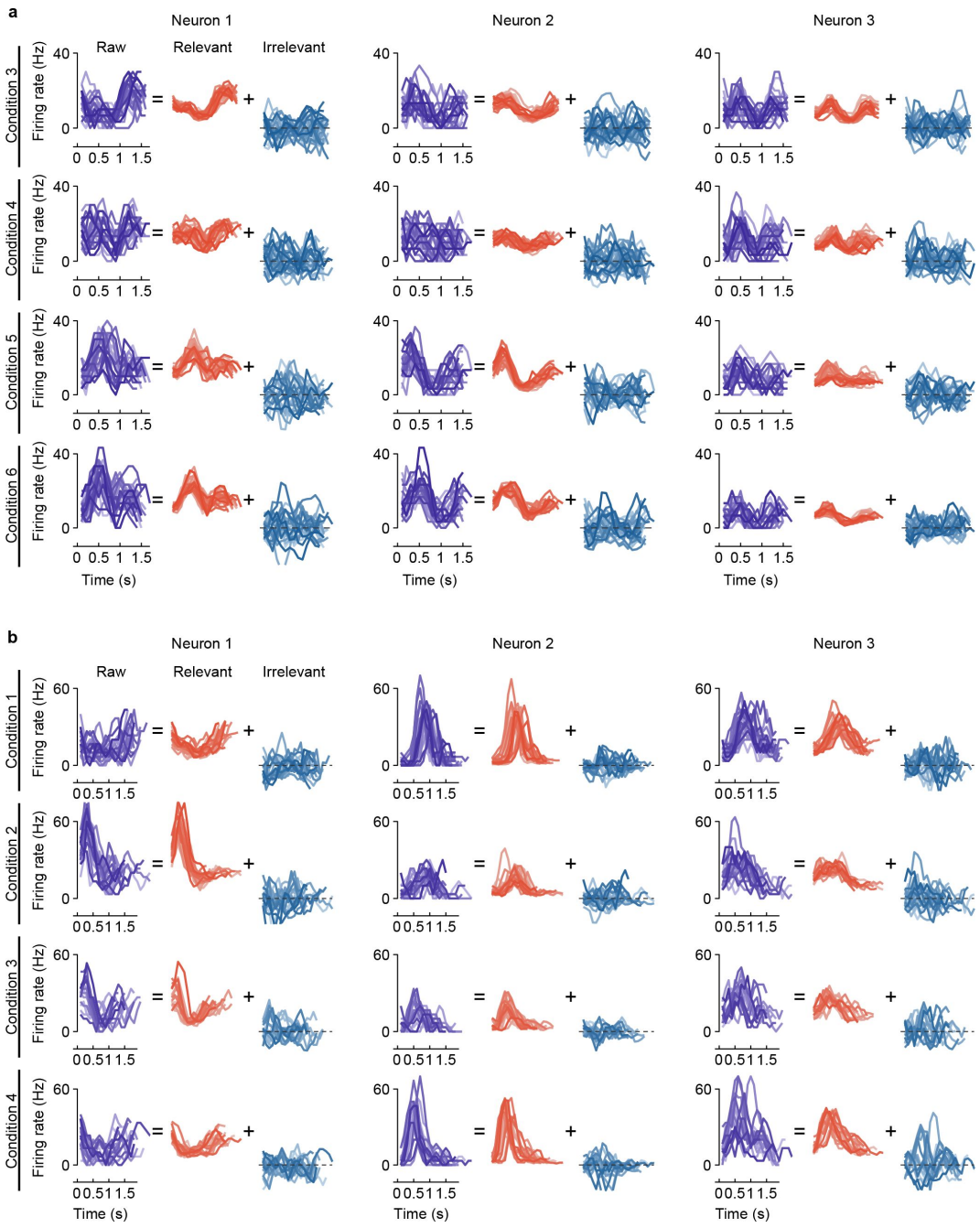


Figure S5.

The firing activity of example neurons.

a, Example of three neurons' raw firing activity decomposed into behaviorally-relevant and irrelevant parts using all trials in held-out test sets for four conditions (4 of 8 directions) of center-out reaching task. **b**, Example of three neurons' raw firing activity decomposed into behaviorally-relevant and irrelevant parts using all trials in held-out test sets for four conditions (4 of 12 conditions) of obstacle avoidance task.

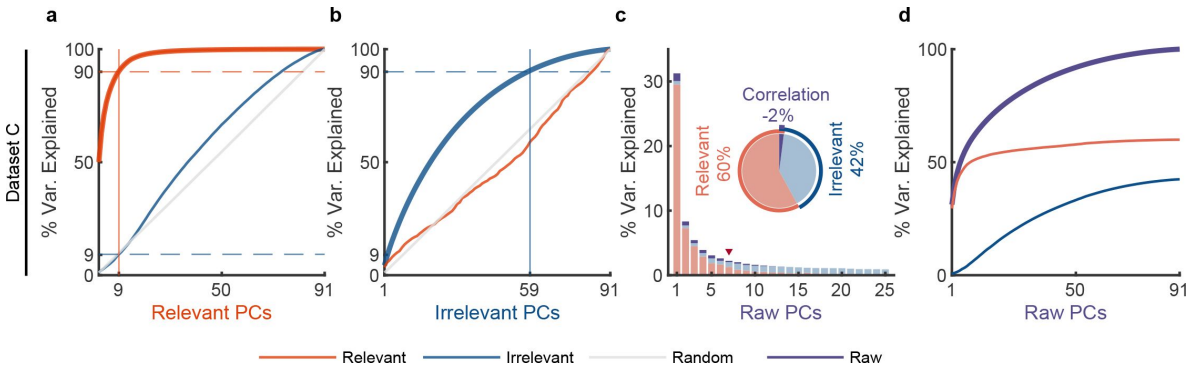


Figure S6.

The effect of irrelevant signals on analyzing neural activity at the population level.

a-d, Same as **Fig. 4**, but for dataset C. **a**,**b**, PCA is separately applied on relevant and irrelevant signals to get relevant PCs and irrelevant PCs. The thick lines represent the cumulative variance explained for the signals on which PCA has been performed, while the thin lines represent the variance explained by those PCs for other signals. Red, blue, and gray colors indicate relevant signals, irrelevant signals, and random Gaussian noise (for chance level). The cumulative variance explained for behaviorally-relevant (**a**) and irrelevant (**b**) signals got by d-VAE. **c**,**d**, PCA is applied on raw signals to get raw PCs. **c**, The bar plot represents the composition of each raw PC. The inset pie plot shows the overall proportion of raw signals, where red, blue, and purple colors indicate relevant signals, irrelevant signals, and the correlation between relevant and relevant signals. The PC marked with a red triangle indicates the last PC where the variance of relevant signals is greater than or equal to that of irrelevant signals. **d**, The cumulative variance explained by raw PCs for different signals, where the thick lines represent the cumulative variance explained for raw signals(purple), while the thin lines represent the variance explained for relevant (red) and irrelevant (blue) signals.

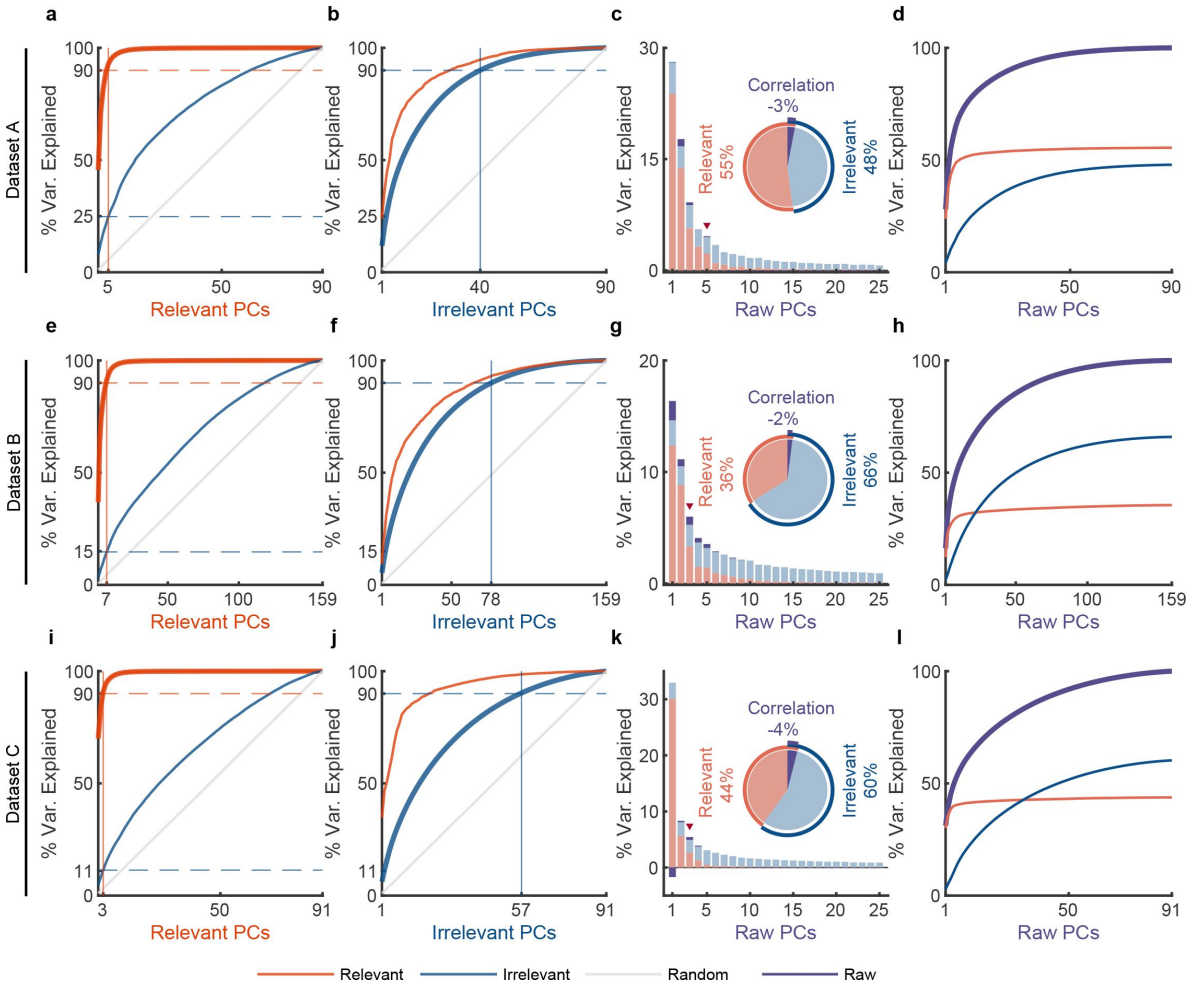


Figure S7.

The effect of irrelevant signals obtained by pi-VAE on analyzing neural activity at the population level.

a-l, Same as **Fig. 4** and **Fig. S6**, but for pi-VAE. **a**,**b**, PCA is separately applied on relevant and irrelevant signals to get relevant PCs and irrelevant PCs. The thick lines represent the cumulative variance explained for the signals on which PCA has been performed, while the thin lines represent the variance explained by those PCs for other signals. Red, blue, and gray colors indicate relevant signals, irrelevant signals, and random Gaussian noise (for chance level). The cumulative variance explained for behaviorally-relevant (**a**) and irrelevant (**b**) signals on dataset A. **c**,**d**, PCA is applied on raw signals to get raw PCs. **c**, The bar plot represents the composition of each raw PC. The inset pie plot shows the overall proportion of raw signals, where red, blue, and purple colors indicate relevant signals, irrelevant signals, and the correlation between relevant and relevant signals. The PC marked with a red triangle indicates the last PC where the variance of relevant signals is greater than or equal to that of irrelevant signals. **d**, The cumulative variance explained by raw PCs for different signals, where the thick line represents the cumulative variance explained for raw signals(purple), while the thin line represents the variance explained for relevant (red) and irrelevant (blue) signals. **e-h, i-l**, Same as **a-d**, but for datasets B and C.

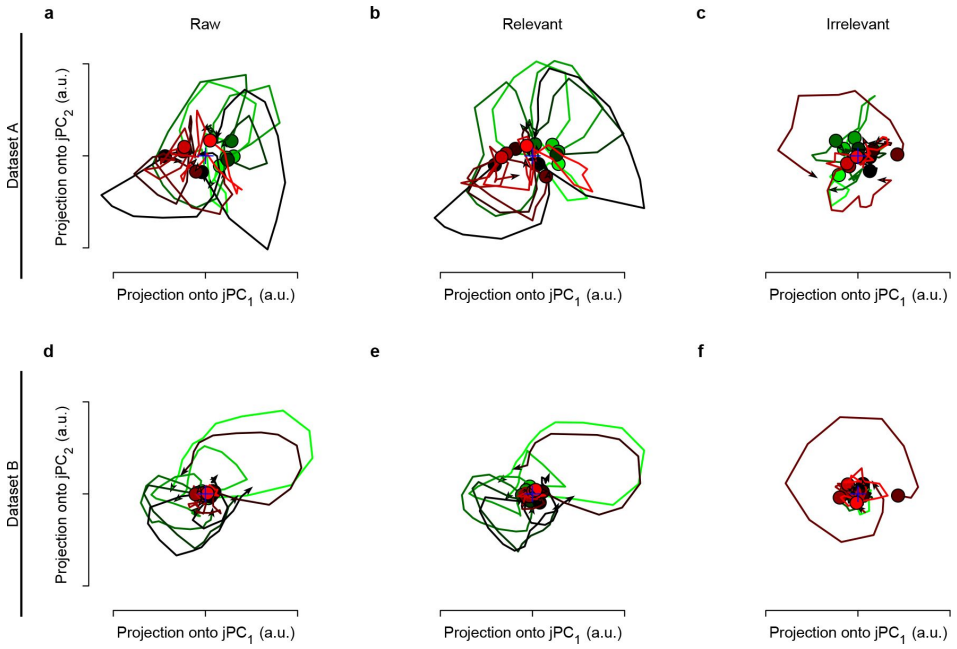


Figure S8.

The rotational dynamics of raw, relevant, and irrelevant signals.

datasets A and B have twelve and eight conditions, respectively. We get the trial-averaged neural responses for each condition, then apply jPCA to raw, relevant, and irrelevant signals to get the top two jPC, respectively. **a**, The rotational dynamics of raw neural signals. **b**, The rotational dynamics of relevant signals obtained by d-VAE. **c**, The rotational dynamics of irrelevant signals obtained by d-VAE. We can see that the rotational dynamics of behaviorally-relevant signals are similar to that of raw signals, but the rotational dynamics of behaviorally-irrelevant signals are irregular. **d-f**, Same as **a-c**, but for dataset B.

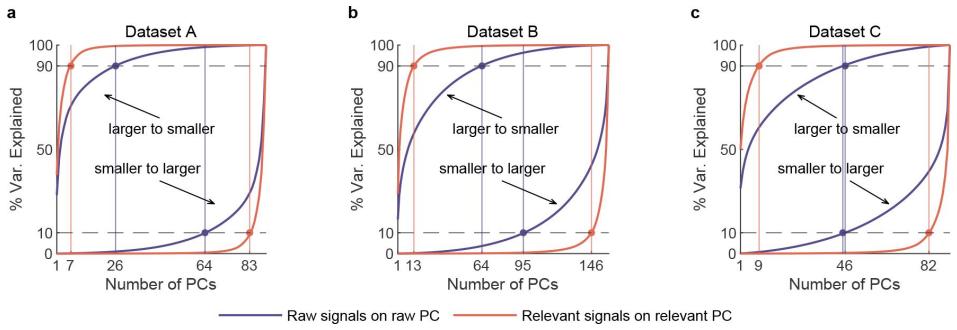


Figure S9.

The cumulative variance of raw and behaviorally-relevant signals.

a, PCA is applied separately on raw and distilled behaviorally-relevant signals to get raw PCs and relevant PCs. The cumulative variance of raw (purple) and behaviorally-relevant signals (red) on dataset A (n=90). Two upper left corner curves denote the variance accumulation from larger to smaller variance PCs. Two lower right corner curves indicate accumulation from smaller to larger variance PCs. The horizontal lines represent the 10%, and 90% variance explained. The vertical lines indicate the number of dimensions accounted for the last 10% and top 90% of the variance of behaviorally-relevant (red) and raw (purple) signals. Here we call the subspace composed by PCs of capturing top 90% variance the primary subspace, and the subspace composed by PCs of capturing last 10% variance the secondary subspace. We can see that the dimensionality of the primary subspace of raw signals is significantly higher than that of relevant signals, indicating that irrelevant signals make us overestimate the dimensionality of specific behaviors. **b**,**c**, Same as **a**, but for datasets B (n=159) and C (n=91).

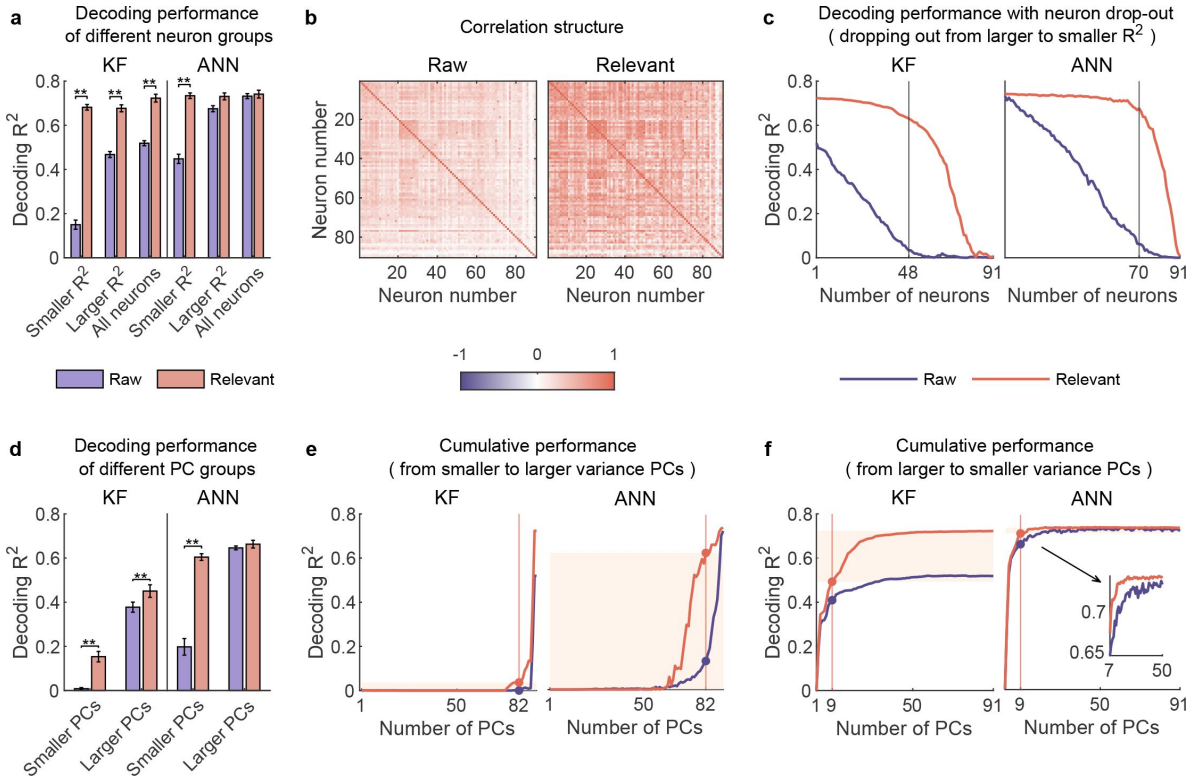


Figure S10.

neural responses usually considered useless encode rich behavioral information in complex nonlinear ways.

a-c, Same as **Fig. 5**, but for dataset C (n=91). **d-f**, Same as **Fig. 6**, but for dataset C. **a**, The comparison of decoding performance between raw (purple) and distilled signals (red) with different neuron groups, including smaller $\mathbf{R}^2$ neuron ($\mathbf{R}^2$ <= 0.03), larger $\mathbf{R}^2$ neuron ($\mathbf{R}^2$ > 0.03), and all neurons. Error bars indicate mean ± SD across cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{*}P < 0.05$, $^{**}P < 0.01$. **b**, The correlation matrix of all neurons of raw and behaviorally-relevant signals. **c**, The decoding performance of KF (left) and ANN (right) with neurons dropped out from larger to smaller $\mathbf{R}^2$. The vertical gray lines indicate the number of dropped neurons at which raw and behaviorally-relevant signals have the greatest performance difference. **d**, The comparison of decoding performance between raw (purple) and distilled signals (red) composed of different raw-PC groups, including smaller variance PCs (the proportion of irrelevant signals that make up raw PCs is higher than that of relevant signals), larger variance PCs (the proportion of irrelevant signals is lower than that of relevant ones). Error bars indicate mean ± s.d. across five cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{*}P < 0.05$, $^{**}P < 0.01$. **e**, The cumulative decoding performance of signals composed of cumulative PCs that are ordered from smaller to larger variance using KF (left) and ANN (right). The red patches indicate the decoding ability of the last 10% variance of relevant signals. **f**, Same as **e**, but PCs are ordered from larger to smaller variance. The red patches indicate the decoding gain of the last 10% variance signals of relevant signals superimposing on their top 90% variance signals.

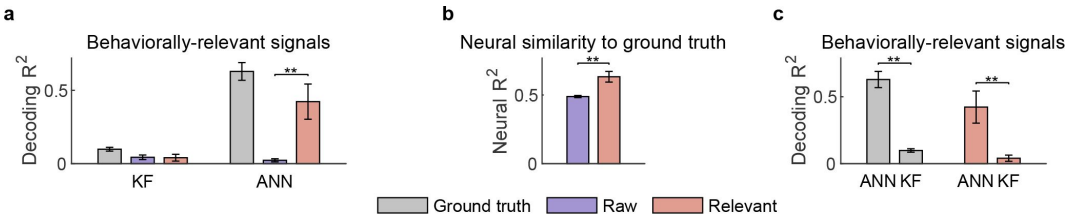


Figure S11.

Using synthetic data to demonstrate that conclusions are not a by-product of d-VAE.

a, b These results are used to demonstrate that d-VAE can utilize the larger R2 neurons to help the smaller R2 neurons restore their original face. **a**, The decoding $\mathbf{R}^2$ of the ground truth (gray), raw signals (purple), and distilled relevant signals (red) of smaller $\mathbf{R}^2$ neurons of synthetic data. Error bars indicate mean ± s.d. (n=5 folds). Asterisks denote the significance of Wilcoxon rank-sum test with $^{**}P < 0.01$. We can see that the ground truth of smaller $\mathbf{R}^2$ neurons contain a certain amount of behavioral information, but the behavioral information cannot be decoded from raw signals due to being covered by noise; d-VAE can indeed utilize the larger $\mathbf{R}^2$ neurons to help the smaller $\mathbf{R}^2$ neurons restore their damaged information. **b**, The neural similarity of raw signals and relevant signals to ground truth of smaller $\mathbf{R}^2$ neurons. We can see that d-VAE can obtain effective relevant signals that are more similar to the ground truth compared to raw signals. **c**, The decoding $\mathbf{R}^2$ of the ground truth (gray) and distilled relevant signals (red) of smaller $\mathbf{R}^2$ neurons of synthetic data. These results are used to demonstrate that d-VAE can not make the linear decoder achieve similar performance as the nonlinear decoder. We can see that KF is significantly inferior to ANN on ground truth signals. The KF decoding performance of the ground truth signals is notably low, leaving significant room for compensation by d-VAE. However, after processing with d-VAE, the KF decoding performance of distilled signals does not surpass its ground truth performance. The disparity between KF and ANN remains substantial. These results demonstrate that d-VAE can not make signals that originally require nonlinear decoding linearly decodable.

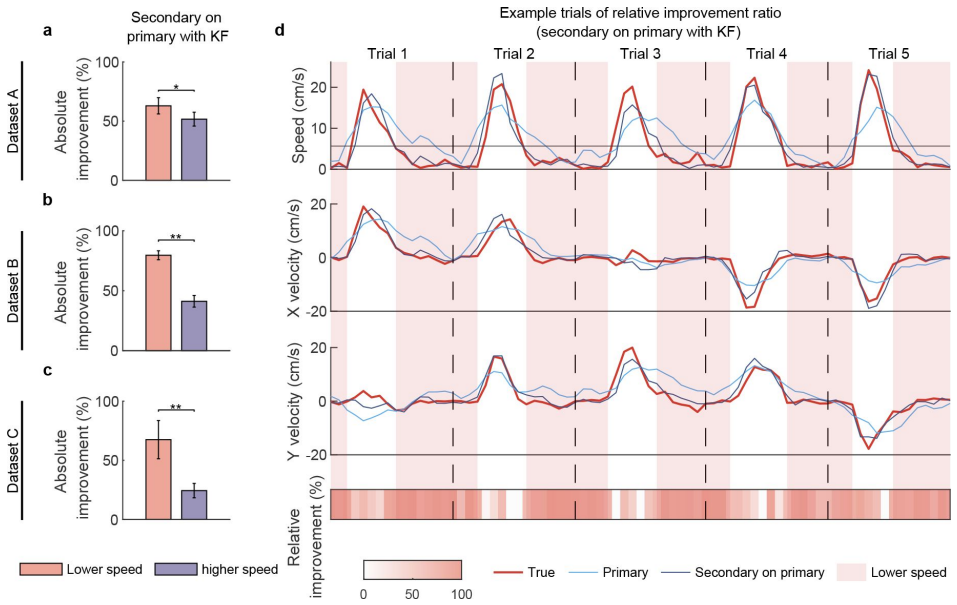


Figure S12.

Smaller variance PC signals preferentially improve lower-speed velocity.

a, The comparison of absolute improvement ratio between lower-speed (red) and higher-speed (purple) velocity when superimposing secondary signals on primary signals with KF on dataset A. Error bars indicate mean ± s.d. across five cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{*}P < 0.05$, $^{**}P < 0.01$. **b, c**, Same as **a**, but for datasets B and C. **d**, The comparison of relative improvement ratio between lower-speed (red patch) and higher-speed (no patch) velocity when superimposing secondary signals on primary signals with KF on dataset B. The first-row plot shows five example trials' speed profile of the decoded velocity using primary signals (light blue line) and full signals (dark blue line; superimposing secondary signals on primary signals) and the true velocity (red line). The black horizontal line denotes the speed threshold. The second and third-row plots are the same as the first-row plot, but for X and Y velocity. The fourth-row plot shows the relative improvement ratio for each point in trials.

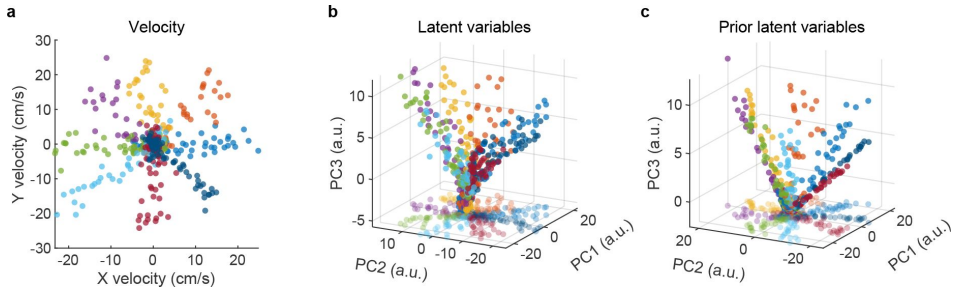


Figure S13.

Visualization of latent variables.

a, The velocity samples of one-fold test data. Different colors denote different directions of the eight-direction center-out. **b**, The distribution plot of the top three principal components (PCs) of latent variables. The points on the bottom plane represent the two-dimensional projections of the three-dimensional data. **c**, The distribution plot of the top three PCs of learned prior latent variables. We can see that the distribution of prior latent variables closely resembles that of latent variables, thus illustrating the effectiveness of the KL divergence.

References

Alstott J, Breakspear M, Hagmann P, Cammoun L, Sporns O. (2009) **Modeling the impact of lesions in the human brain** *PLoS computational biology* **5**

Altan E, Solla SA, Miller LE, Perreault EJ (2021) **Estimating the dimensionality of the manifold underlying multielectrode neural recordings** *PLoS computational biology* **17**

Azouz R, Gray CM (1999) **Cellular mechanisms contributing to response variability of cortical neurons in vivo** *Journal of Neuroscience* **19**:2209–2223

Carmena JM, Lebedev MA, Henriquez CS, Nicolelis MA (2005) **Stable ensemble performance with single-neuron variability during reaching movements in primates** *Journal of Neuroscience* **25**:10712–10716

Churchland MM, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, Ryu SI, Shenoy KV (2012) **Neural population dynamics during reaching** *Nature* **487**:51–56

Churchland MM *et al.* (2010) **Stimulus onset quenches neural variability: a widespread cortical phenomenon** *Nature neuroscience* **13**:369–378

Collinger JL, Wodlinger B, Downey JE, Wang W, Tyler-Kabara EC, Weber DJ, McMorland AJ, Velliste M, Boninger ML, Schwartz AB (2013) **High-performance neuroprosthetic control by an individual with tetraplegia** *The Lancet* **381**:557–564

Cunningham JP, Yu BM (2014) **Dimensionality reduction for large-scale neural recordings** *Nature neuroscience* **17**:1500–1509

Dhawale AK, Smith MA, Ölveczky BP (2017) **The role of variability in motor learning** *Annual review of neuroscience* **40**

Dyer EL, Gheshlaghi Azar M, Perich MG, Fernandes HL, Naufel S, Miller LE, Körding KP (2017) **A cryptography-based approach for movement decoding** *Nature biomedical engineering* **1**:967–976

Elsayed GF, Lara AH, Kaufman MT, Churchland MM, Cunningham JP (2016) **Reorganization between preparatory and movement population responses in motor cortex** *Nature communications* **7**:1–15

Faisal AA, Selen LP, Wolpert DM (2008) **Noise in the nervous system** *Nature reviews neuroscience* **9**:292–303

Fusi S, Miller EK, Rigotti M. (2016) **Why neurons mix: high dimensionality for higher cognition** *Current opinion in neurobiology* **37**:66–74

Gallego JA, Perich MG, Chowdhury RH, Solla SA, Miller LE (2020) **Long-term stability of cortical population dynamics underlying consistent behavior** *Nature neuroscience* **23**:260–270

Gallego JA, Perich MG, Miller LE, Solla SA (2017) **Neural manifolds for the control of movement** *Neuron* **94**:978–984

Gallego JA, Perich MG, Naufel SN, Ethier C, Solla SA, Miller LE (2018) **Cortical population activity within a preserved neural manifold underlies multiple motor behaviors** *Nature communications* **9**:1–13

Georgopoulos AP, Schwartz AB, Kettner RE (1986) **Neuronal population coding of movement direction** *Science* **233**:1416–1419

Glaser JI, Benjamin AS, Chowdhury RH, Perich MG, Miller LE, Kording KP (2020) **Machine learning for neural decoding** *Eneuro*

Hennig JA *et al.* (2021) **Learning is shaped by abrupt changes in neural engagement** *Nature Neuroscience* **24**:727–736

Hochberg LR *et al.* (2012) **Reach and grasp by people with tetraplegia using a neurally controlled robotic arm** *Nature* **485**:372–375

Hurwitz C, Srivastava A, Xu K, Jude J, Perich M, Miller L, Hennig M. (2021) **Targeted neural dynamical modeling** *Advances in Neural Information Processing Systems* **34**:29379–29392

Inoue Y, Mao H, Suway SB, Orellana J, Schwartz AB (2018) **Decoding arm speed during reaching** *Nature communications* **9**:1–14

Jiang X, Saggar H, Ryu SI, Shenoy KV, Kao JC (2020) **Structure in neural activity during observed and executed movements is shared at the neural population level, not in single neurons** *Cell reports* **32**

Keshtkaran MR, Sedler AR, Chowdhury RH, Tandon R, Basrai D, Nguyen SL, Sohn H, Jazayeri M, Miller LE, Pandarinath C. (2022) **A large-scale neural network training framework for generalized estimation of single-trial population dynamics** *Nature Methods*

Khemakhem I, Kingma D, Monti R, Hyvarinen A. (2020) **Variational autoencoders and nonlinear ica: A unifying framework** *International Conference on Artiicial Intelligence and Statistics PMLR* :2207–2217

Kingma DP, Welling M. (2013) **Auto-encoding variational bayes** *arXiv preprint*

Kobak D, Brendel W, Constantinidis C, Feierstein CE, Kepecs A, Mainen ZF, Qi XL, Romo R, Uchida N, Machens CK (2016) **Demixed principal component analysis of neural population data** *Elife* **5**

Kriegeskorte N, Douglas PK (2019) **Interpreting encoding and decoding models** *Current opinion in neurobiology* **55**:167–179

Li Y, Qi Y, Wang Y, Wang Y, Xu K, Pan G. (2021) **Robust neural decoding by kernel regression with Siamese representation learning** *Journal of Neural Engineering* **18**

Majaj NJ, Hong H, Solomon EA, DiCarlo JJ (2015) **Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance** *Journal of Neuroscience* **35**:13402–13418

Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A. (2014) **Information-limiting correlations** *Nature neuroscience* **17**:1410–1417

Musall S, Kaufman MT, Juavinett AL, Gluf S, Churchland AK (2019) **Single-trial neural dynamics are dominated by richly varied movements** *Nature neuroscience* **22**:1677–1686

Narayanan NS, Kimchi EY, Laubach M. (2005) **Redundancy and synergy of neuronal ensembles in motor cortex** *Journal of Neuroscience* **25**:4207–4216

Naufel S, Glaser JI, Kording KP, Perreault EJ, Miller LE (2019) **A muscle-activity-dependent gain between motor cortex and EMG** *Journal of neurophysiology* **121**:61–73

Nogueira R, Rodgers CC, Bruno RM, Fusi S. (2023) **The geometry of cortical representations of touch in rodents** *Nature Neuroscience* :1–12

O'Doherty JE, Cardoso M, Makin J, Sabes P. (2017) **Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology** *Zenodo*

Pagan M, Urban LS, Wohl MP, Rust NC (2013) **Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information** *Nature neuroscience* **16**:1132–1139

Pandarinath C *et al.* (2018) **Inferring single-trial neural population dynamics using sequential auto-encoders** *Nature methods* **15**:805–815

Pitkow X, Liu S, Angelaki DE, DeAngelis GC, Pouget A. (2015) **How can single sensory neurons predict behavior?** *Neuron* **87**:411–423

Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, Fusi S. (2013) **The importance of mixed selectivity in complex cognitive tasks** *Nature* **497**:585–590

Rouse AG, Schieber MH (2018) **Condition-dependent neural dimensions progressively shift during reach to grasp** *Cell reports* **25**:3158–3168

Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-Kabara EC, Yu BM, Batista AP (2014) **Neural constraints on learning** *Nature* **512**:423–426

Sani OG, Abbaspourazad H, Wong YT, Pesaran B, Shanechi MM (2021) **Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification** *Nature Neuroscience* **24**:140–149

Saxena S, Cunningham JP (2019) **Towards the neural population doctrine** *Current opinion in neurobiology* **55**:103–111

Schneider DM, Sundararajan J, Mooney R. (2018) **A cortical filter that learns to suppress the acoustic consequences of movement** *Nature* **561**:391–395

Schneider S, Lee JH, Mathis MW (2023) **Learnable latent embeddings for joint behavioural and neural analysis** *Nature* :1–9

Sun G, Zeng F, McCartin M, Zhang Q, Xu H, Liu Y, Chen ZS, Wang J. (2022) **Closed-loop stimulation using a multiregion brain-machine interface has analgesic effects in rodents** *Science Translational Medicine* **14**

Walker EY, Cotton RJ, Ma WJ, Tolias AS (2020) **A neural basis of probabilistic computation in visual cortex** *Nature Neuroscience* **23**:122–129

Wang F, Wang Y, Xu K, Li H, Liao Y, Zhang Q, Zhang S, Zheng X, Principe JC (2015) **Quantized attention-gated kernel reinforcement learning for brain–machine interface decoding** *IEEE transactions on neural networks and learning systems* **28**:873–886

Willsey MS, Nason-Tomaszewski SR, Ensel SR, Temmar H, Mender MJ, Costello JT, Patil PG, Chestek CA (2022) **Realtime brain-machine interface in non-human primates achieves high-velocity prosthetic finger movements using a shallow feedforward neural network decoder** *Nature Communications* **13**:1–14

Wodlinger B, Downey J, Tyler-Kabara E, Schwartz A, Boninger M, Collinger J. (2014) **Ten-dimensional anthropomorphic arm control in a human brainmachine interface: diffculties, solutions, and limitations** *Journal of neural engineering* **12**

Yan Y, Goodman JM, Moore DD, Solla SA, Bensmaia SJ (2020) **Unexpected complexity of everyday manual behaviors** *Nature communications* **11**:1–8

Yang Q, Walker E, Cotton RJ, Tolias AS, Pitkow X. (2021) **Revealing nonlinear neural decoding by analyzing choices** *Nature communications* **12**:1–13

Yatsenko D, Josić K, Ecker AS, Froudarakis E, Cotton RJ, Tolias AS (2015) **Improved estimation and interpretation of correlations in neural circuits** *PLoS computational biology* **11**

Yu BM, Cunningham JP, Santhanam G, Ryu S, Shenoy KV, Sahani M. (2008) **Gaussian-process factor analysis for lowdimensional single-trial analysis of neural population activity** *Advances in neural information processing systems*

Zhang J, Hughes RN, Kim N, Fallon IP, Bakhurin K, Kim J, Severino FPU, Yin HH (2022) **A one-photon endoscope for simultaneous patterned optogenetic stimulation and calcium imaging in freely behaving mice** *Nature Biomedical Engineering* :1–12

Zhou D, Wei XX (2020) **Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE** *Advances in Neural Information Processing Systems* **33**:7234–7247

Article and author information

Yangang Li

Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China, College of Computer Science and Technology, Zhejiang University, Hangzhou, China, The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China
ORCID iD: 0000-0002-7271-2993

Xinyun Zhu

Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China, College of Computer Science and Technology, Zhejiang University, Hangzhou, China, The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China

Yu Qi
Affliated Mental Health Center & Hangzhou Seventh People's Hospital and the MOE Frontier Science Center for Brain Science and Brain-machine Integration, Zhejiang University School of Medicine, Hangzhou, China, The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China
For correspondence: qiyu@zju.edu.cn

Yueming Wang
Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China, Zhejiang Brain-Computer Interface Institute, Hangzhou, China
For correspondence: ymingwang@zju.edu.cn

Copyright

Editors

Reviewing Editor
Juan Alvaro Gallego
Imperial College London, London, United Kingdom

Senior Editor
Tamar Makin
University of Cambridge, Cambridge, United Kingdom

Reviewer #1 (Public Review):

This work seeks to understand how behaviour-related information is represented in the neural activity of the primate motor cortex. To this end, a statistical model of neural activity is presented that enables a non-linear separation of behaviour-related from unrelated activity. As a generative model, it enables the separate analysis of these two activity modes, here primarily done by assessing the decoding performance of hand movements the monkeys perform in the experiments. Several lines of analysis are presented to show that while the neurons with significant tuning to movements strongly contribute to the behaviourally-relevant activity subspace, less or un-tuned neurons also carry decodable information. It is further shown that the discovered subspaces enable linear decoding, leading the authors to conclude that motor cortex read-out can be linear.

Strengths:

In my opinion, using an expressive generative model to analyse neural state spaces is an interesting approach to understand neural population coding. While potentially sacrificing interpretability, this approach allows capturing both redundancies and synergies in the code as done in this paper. The model presented here is a natural non-linear extension of a previous linear model PSID) and uses weak supervision in a manner similar to a previous non-linear model (TNDM).

Weaknesses:

This revised version provides additional evidence to support the author's claims regarding model performance and interpretation of the structure of the resulting latent spaces, in particular the distributed neural code over the whole recorded population, not just the well-tuned neurons. The improved ability to linearly decode behaviour from the relevant subspace and the analysis of the linear subspace projections in my opinion convincingly demonstrates that the model picks up behaviour-relevant dynamics, and that these are distributed widely across the population. As reviewer 3 also points out, I would, however, caution to interpret this as evidence for linear read-out of the motor system - your model performs a non-linear transformation, and while this is indeed linearly decodable, the motor system would need to do something similar first to achieve the same. In fact to me it seems to show the opposite, that behaviour-related information may not be generally accessible to linear decoders (including to down-stream brain areas).

As in my initial review, I would also caution against making strong claims about identifiability although this work and TNDM seem to show that in practise such methods work quite well. CEBRA, in contrast, offers some theoretical guarantees, but it is not a generative model, so would not allow the type of analysis done in this paper. In your model there is a para,eter \alpha to balance between neural and behaviour reconstruction. This seems very similar to TNDM and has to be optimised - if this is correct, then there is manual intervention required to identify a good model.

Somewhat related, I also found that the now comprehensive comparison with related models shows that the using decoding performance (R2) as a metric for model comparison may be problematic: the R2 values reported in Figure 2 (e.g. the MC_RTT dataset) should be compared to the values reported in the neural latent benchmark, which represent well-tuned models (e.g. AutoLFADS). The numbers (difficult to see, a table with numbers in the appendix would be useful, see: https://eval.ai/web/challenges/challenge-page/1256/leaderboard) seem lower than what can be obtained with models without latent space disentanglement. While this does not necessarily invalidate the conclusions drawn here, it shows that decoding performance can depend on a variety of model choices, and may not be ideal to discriminate between models. I'm also surprised by the low neural R2 for LFADS I assume this is condition-averaged) - LFADS tends to perform very well on this metric.

One statement I still cannot follow is how the prior of the variational distribution is modelled. You say you depart from the usual Gaussian prior, but equation 7 seems to suggest there is a normal prior. Are the parameters of this distribution learned? As I pointed out earlier, I however suspect this may not matter much as you give the prior a very low weight. I also still am not sure how you generate a sample from the variational distribution, do you just draw one for each pass?

Summary:

This paper presents a very interesting analysis, but some concerns remain that mainly stem from the complexity of deep learning models. It would be good to acknowledge these as readers without relevant background need to understand where the possible caveats are.

https://doi.org/10.7554/eLife.87881.2.sa2

Reviewer #2 (Public Review):

Li et al present a method to extract "behaviorally relevant" signals from neural activity. The method is meant to solve a problem which likely has high utility for neuroscience researchers. There are numerous existing methods to achieve this goal some of which the authors compare their method to-thankfully, the revised version includes one of the major

previous omissions (TNDM). However, I still believe that d-VAE is a promising approach that has its own advantages. Still, I have issues with the paper as-is. The authors have made relatively few modifications to the text based on my previous comments, and the responses have largely just dismissed my feedback and restated claims from the paper. Nearly all of my previous comments remain relevant for this revised manuscript. As such, they have done little to assuage my concerns, the most important of which I will restate here using the labels/notation (Q1, Q2, etc) from the reviewer response.

Q1) I still remain unconvinced that the core findings of the paper are "unexpected". In the response to my previous Specific Comment #1, they say "We use the term 'unexpected' due to the disparity between our findings and the prior understanding concerning neural encoding and decoding." However, they provide no citations or grounding for why they make those claims. What prior understanding makes it unexpected that encoding is more complex than decoding given the entropy, sparseness, and high dimensionality of neural signals (the "encoding") compared to the smoothness and low dimensionality of typical behavioural signals (the "decoding")?

Q2) I still take issue with the premise that signals in the brain are "irrelevant" simply because they do not correlate with a fixed temporal lag with a particular behavioural feature hand-chosen by the experimenter. In the response to my previous review, the authors say "we employ terms like 'behaviorally-relevant' and 'behaviorally-irrelevant' only regarding behavioral variables of interest measured within a given task, such as arm kinematics during a motor control task.". This is just a restatement of their definition, not a response to my concern, and does not address my concern that the method requires a fixed temporal lag and continual decoding/encoding. My example of reward signals remains. There is a huge body of literature dating back to the 70s on the linear relationships between neural and activity and arm kinematics; in a sense, the authors have chosen the "variable of interest" that proves their point. This all ties back to the previous comment: this is mostly expected, not unexpected, when relating apparently-stochastic, discrete action potential events to smoothly varying limb kinematics.

Q5) The authors seem to have missed the spirit of my critique: to say "linear readout is performed in motor cortex" is an over-interpretation of what their model can show.

Q7) Agreeing with my critique is not sufficient; please provide the data or simulations that provides the context for the reference in the fano factor. I believe my critique is still valid.

Q8) Thank you for comparing to TNDM, it's a useful benchmark.

https://doi.org/10.7554/eLife.87881.2.sa1

Reviewer #4 (Public Review):

I am a new reviewer for this manuscript, which has been reviewed before. The authors provide a variational autoencoder that has three objectives in the loss: linear reconstruction of behavior from embeddings, reconstruction of neural data, and KL divergence term related to the variational model elements. They take the output of the VAE as the "behaviorally relevant" part of neural data and call the residual "behaviorally irrelevant". Results aim to inspect the linear versus nonlinear behavior decoding using the original raw neural data versus the inferred behaviorally relevant and irrelevant parts of the signal.

Overall, studying neural computations that are behaviorally relevant or not is an important problem, which several previous studies have explored (for example PSID in (Sani et al. 2021), TNDM in (Hurwitz et al. 2021), TAME-GP in (Balzani et al. 2023), pi-VAE in (Zhou and Wei 2020), and dPCA in (Kobak et al. 2016), etc). However, this manuscript does not properly

put their work in the context of such prior works. For example, the abstract states "One solution is to accurately separate behaviorally-relevant and irrelevant signals, but this approach remains elusive", which is not the case given that these prior works have done that. The same is true for various claims in the main text, for example "Furthermore, we found that the dimensionality of primary subspace of raw signals (26, 64, and 45 for datasets A, B, and C) is significantly higher than that of behaviorally-relevant signals (7, 13, and 9), indicating that using raw signals to estimate the neural dimensionality of behaviors leads to an overestimation" (line 321). This finding was presented in (Sani et al. 2021) and (Hurwitz et al. 2021), which is not clarified here. This issue of putting the work in context has been brought up by other reviewers previously but seems to remain largely unaddressed. The introduction is inaccurate also in that it mixes up methods that were designed for separation of behaviorally relevant information with those that are unsupervised and do not aim to do so (e.g., LFADS). The introduction should be significantly revised to explicitly discuss prior models/works that specifically formulated this behavior separation and what these prior studies found, and how this study differs.

Beyond the above, some of the main claims/conclusions made by the manuscript are not properly supported by the analyses and results, which has also been brought up by other reviewers but not fully addressed. First, the analyses here do not support the linear readout from the motor cortex because i) by construction, the VAE here is trained to have a linear readout from its embedding in its loss, which can bias its outputs toward doing well with a linear decoder/readout, and ii) the overall mapping from neural data to behavior includes both the VAE and the linear readout and thus is always nonlinear (even when a linear Kalman filter is used for decoding). This claim is also vague as there is no definition of readout from "motor cortex" or what it means. Why is the readout from the bottleneck of this particular VAE the readout of motor cortex? Second, other claims about properties of individual neurons are also confounded because the VAE is a population-level model that extracts the bottleneck from all neurons. Thus, information can leak from any set of neurons to other sets of neurons during the inference of behaviorally relevant parts of signals. Overall, the results do not convincingly support the claims, and thus the claims should be carefully revised and significantly tempered to avoid misinterpretation by readers.

Below I briefly expand on these as well as other issues, and provide suggestions:

1. Claims about linearity of "motor cortex" readout are not supported by results yet stated even in the abstract. Instead, what the results support is that for decoding behavior from the output of the dVAE model -- that is trained specifically to have a linear behavior readout from its embedding -- a nonlinear readout does not help. This result can be biased by the very construction of the dVAE's loss that encourages a linear readout/decoding from embeddings, and thus does not imply a finding about motor cortex.

2. Related to the above, it is unclear what the manuscript means by readout from motor cortex. A clearer definition of "readout" (a mapping from what to what?) in general is needed. The mapping that the linearity/nonlinearity claims refer to is from the *inferred* behaviorally relevant neural signals, which themselves are inferred nonlinearly using the VAE. This should be explicitly clarified in all claims, i.e., that only the mapping from distilled signals to behavior is linear, not the whole mapping from neural data to behavior. Again, to say the readout from motor cortex is linear is not supported, including in the abstract.

3. Claims about individual neurons are also confounded. The d-VAE distilling processing is a population level embedding so the individual distilled neurons are not obtainable on their own without using the population data. This population level approach also raises the possibility that information can leak from one neuron to another during distillation, which is indeed what the authors hope would recover true information about individual neurons that wasn't there in the recording (the pixel denoising example). The authors acknowledge the possibility that information could leak to a neuron that didn't truly have that information and try to rule it out to some extent with some simulations and by comparing the distilled behaviorally relevant signals to the original neural signals. But ultimately, the distilled signals are different enough from the original signals to substantially improve decoding of low information neurons, and one cannot be sure if all of the information in distilled signals from any individual neuron truly belongs to that neuron. It is still quite likely that some of the improved behavior prediction of the distilled version of low-information neurons is due to leakage of behaviorally relevant information from other neurons, not the former's inherent behavioral information. This should be explicitly acknowledged in the manuscript.

4. Given the nuances involved in appropriate comparisons across methods and since two of the datasets are public, the authors should provide their complete code (not just the dVAE method code), including the code for data loading, data preprocessing, model fitting and model evaluation for all methods and public datasets. This will alleviate concerns and allow readers to confirm conclusions (e.g., figure 2) for themselves down the line.

5. Related to 1) above, the authors should explore the results if the affine network h(.) (from embedding to behavior) was replaced with a nonlinear ANN. Perhaps linear decoders would no longer be as close to nonlinear decoders. Regardless, the claim of linearity should be revised as described in 1) and 2) above, and all caveats should be discussed.

6. The beginning of the section on the "smaller R2 neurons" should clearly define what R2 is being discussed. Based on the response to previous reviewers, this R2 "signifies the proportion of neuronal activity variance explained by the linear encoding model, calculated using raw signals". This should be mentioned and made clear in the main text whenever this R2 is referred to.

7. Various terms require clear definitions. The authors sometimes use vague terminology (e.g., "useless") without a clear definition. Similarly, discussions regarding dimensionality could benefit from more precise definitions. How is neural dimensionality defined? For example, how is "neural dimensionality of specific behaviors" (line 590) defined? Related to this, I agree with Reviewer 2 that a clear definition of irrelevant should be mentioned that clarifies that relevance is roughly taken as "correlated or predictive with a fixed time lag". The analyses do not explore relevance with arbitrary time lags between neural and behavior data.

8. CEBRA itself doesn't provide a neural reconstruction from its embeddings, but one could obtain one via a regression from extracted CEBRA embeddings to neural data. In addition to decoding results of CEBRA (figure S3), the neural reconstruction of CEBRA should be computed and CEBRA should be added to Figure 2 to see how the behaviorally relevant and irrelevant signals from CEBRA compare to other methods.

References:

Kobak, Dmitry, Wieland Brendel, Christos Constantinidis, Claudia E Feierstein, Adam Kepecs, Zachary F Mainen, Xue-Lian Qi, Ranulfo Romo, Naoshige Uchida, and Christian K Machens.

2016. "Demixed Principal Component Analysis of Neural Population Data." Edited by Mark CW van Rossum. eLife 5 (April): e10989. https://doi.org/10.7554/eLife.10989.

Sani, Omid G., Hamidreza Abbaspourazad, Yan T. Wong, Bijan Pesaran, and Maryam M. Shanechi. 2021. "Modeling Behaviorally Relevant Neural Dynamics Enabled by Preferential Subspace Identification." Nature Neuroscience 24 (1): 140-49. https://doi.org/10.1038/s41593-020-00733-0.

Zhou, Ding, and Xue-Xin Wei. 2020. "Learning Identifiable and Interpretable Latent Models of High-Dimensional Neural Activity Using Pi-VAE." In Advances in Neural Information Processing Systems, 33:7234-47. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/510f2318f324cf07fce24c3a4b89c771-Abstract.html.

Hurwitz, Cole, Akash Srivastava, Kai Xu, Justin Jude, Matthew Perich, Lee Miller, and Matthias Hennig. 2021. "Targeted Neural Dynamical Modeling." In Advances in Neural Information Processing Systems. Vol. 34. https://proceedings.neurips.cc/paper/2021/hash/f5cfbc876972bd0d031c8abc37344c28-Abstract.html.

Balzani, Edoardo, Jean-Paul G. Noel, Pedro Herrero-Vidal, Dora E. Angelaki, and Cristina Savin. 2023. "A Probabilistic Framework for Task-Aligned Intra- and Inter-Area Neural Manifold Estimation." In . https://openreview.net/forum?id=kt-dcBQcSA.

https://doi.org/10.7554/eLife.87881.2.sa0

Author Response

The following is the authors' response to the previous reviews.

To the Senior Editor and the Reviewing Editor:

We sincerely appreciate the valuable comments provided by the reviewers, the reviewing editor, and the senior editor. After carefully reviewing and considering the comments, we have addressed the key concerns raised by the reviewers and made appropriate modifications to the article in the revised manuscript.

The main revisions made to the manuscript are as follows:

1. We have added comparison experiments with TNDM (see Fig. 2 and Fig. S2).
2. We conducted new synthetic experiments to demonstrate that our conclusions are not a by-product of d-VAE (see Fig. S2 and Fig. S11).
3. We have provided a detailed explanation of how our proposed criteria, especially the second criterion, can effectively exclude the selection of unsuitable signals.
4. We have included a semantic overview figure of d-VAE (Fig. S1) and a visualization plot of latent variables (Fig. S13).
5. We have elaborated on the model details of d-VAE, as well as the hyperparameter selection and experimental settings of other comparison models.

We believe these revisions have significantly improved the clarity and comprehensibility of the manuscript. Thank you for the opportunity to address these important points.

> ***Reviewer #1***
>
> *Q1: "First, the model in the paper is almost identical to an existing VAE model (TNDM) that makes use of weak supervision with behaviour in the same way [1]. This paper should at least be referenced. If the authors wish they could compare their model to TNDM, which combines a state space model with smoothing similar to LFADS. Given that TNDM achieves very good behaviour reconstructions, it may be on par with this model without the need for a Kalman filter (and hence may achieve better separation of behaviour-related and unrelated dynamics)."*

Our model significantly differs from TNDM in several aspects. While TNDM also constrains latent variables to decode behavioral information, it does not impose constraints to maximize behavioral information in the generated relevant signals. The trade-off between the decoding and reconstruction capabilities of generated relevant signals is the most significant contribution of our approach, which is not reflected in TNDM. In addition, the backbone network of signal extraction and the prior distribution of the two models are also different.

It's worth noting that our method does not require a Kalman filter. Kalman filter is used for post hoc assessment of the linear decoding ability of the generated signals. Please note that extracting and evaluating relevant signals are two distinct stages.

Heeding your suggestion, we have incorporated comparison experiments involving TNDM into the revised manuscript. Detailed information on model hyperparameters and training settings can be found in the Methods section in the revised manuscripts.

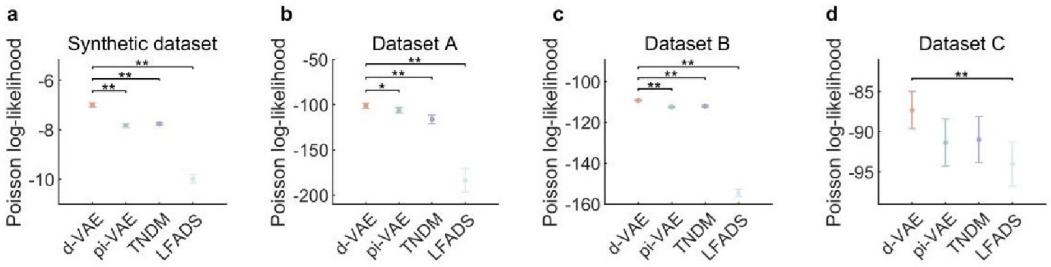
Thank you for your valuable feedback.

> *Q2: "Second, in my opinion, the claims regarding identifiability are overstated - this matters as the results depend on this to some extent. Recent work shows that VAEs generally suffer from identifiability problems due to the Gaussian latent space [2]. This paper also hints that weak supervision may help to resolve such issues, so this model as well as TNDM and CEBRA may indeed benefit from this. In addition however, it appears that the relative weight of the KL Divergence in the VAE objective is chosen very small compared to the likelihood (0.1%), so the influence of the prior is weak and the model may essentially learn the average neural trajectories while underestimating the noise in the latent variables. This, in turn, could mean that the model will not autoencode neural activity as well as it should, note that an average R2 in this case will still be high (I could not see how this is actually computed). At the same time, the behaviour R2 will be large simply because the different movement trajectories are very distinct. Since the paper makes claims about the roles of different neurons, it would be important to understand how well their single trial activities are reconstructed, which can perhaps best be investigated by comparing the Poisson likelihood (LFADS is a good baseline model). Taken together, while it certainly makes sense that well-tuned neurons contribute more to behaviour decoding, I worry that the very interesting claim that neurons with weak tuning contain behavioural signals is not well supported."*

We don't think our distilled signals are average neural trajectories without variability. The quality of reconstructing single trial activities can be observed in Figure 3i and Figure S4. Neural trajectories in Fig. 3i and Fig. S4 show that our distilled signals are not average neural trajectories. Furthermore, if each trial activity closely matched the average neural trajectory, the Fano Factor (FF) should theoretically approach 0. However, our distilled signals exhibit a notable departure from this expectation, as evident in Figure 3c, d, g, and f. Regarding the diminished influence of the KL Divergence: Given that the ground truth of latent variable distribution is unknown, even a learned prior distribution might not accurately reflect the true distribution. We found the pronounced impact of the KL divergence would prove

detrimental to the decoding and reconstruction performance. As a result, we opt to reduce the weight of the KL divergence term. Even so, KL divergence can still effectively align the distribution of latent variables with the distribution of prior latent variables, as illustrated in Fig. S13. Notably, our goal is extracting behaviorally-relevant signals from given raw signals rather than generating diverse samples from the prior distribution. When aim to separating relevant signals, we recommend reducing the influence of KL divergence. Regarding comparing the Poisson likelihood: We compared Poisson log-likelihood among different methods (except PSID since their obtained signals have negative values), and the results show that d-VAE outperforms other methods.

Author response image 1.



Regarding how R2 is computed: $R^2 = 1 - \sum_i (x_i - \hat{x}_i)^2 / \sum_i (x_i - \bar{x})^2$, where x_i, $\hat{x}_i$, and $\bar{x}$ denote ith sample of raw signals, ith sample of distilled relevant signals, and the mean of raw signals. If the distilled signals exactly match the raw signals, the sum of squared error is zero, thus R2=1. If the distilled signals $\hat{x}_i$ always are equal to $\bar{x}$, R2=0. If the distilled signals are worse than the mean estimation, R2 is negative, negative R2 is set to zero.

Thank you for your valuable feedback.

> *Q3: "Third, and relating to this issue, I could not entirely follow the reasoning in the section arguing that behavioural information can be inferred from neurons with weak selectivity, but that it is not linearly decodable. It is right to test if weak supervision signals bleed into the irrelevant subspace, but I could not follow the explanations. Why, for instance, is the ANN decoder on raw data (I assume this is a decoder trained fully supervised) not equal in performance to the revenant distilled signals? Should a well-trained non-linear decoder not simply yield a performance ceiling? Next, if I understand correctly, distilled signals were obtained from the full model. How does a model perform trained only on the weakly tuned neurons? Is it possible that the subspaces obtained with the model are just not optimally aligned for decoding? This could be a result of limited identifiability or model specifics that bias reconstruction to averages (a well-known problem of VAEs). I, therefore, think this analysis should be complemented with tests that do not depend on the model."*

Regarding "Why, for instance, is the ANN decoder on raw data (I assume this is a decoder trained fully supervised) not equal in performance to the relevant distilled signals? Should a well-trained non-linear decoder not simply yield a performance ceiling?": In fact, the decoding performance of raw signals with ANN is quite close to the ceiling. However, due to the presence of significant irrelevant signals in raw signals, decoding models like deep neural networks are more prone to overfitting when trained on noisy raw signals compared to behaviorally-relevant signals. Consequently, we anticipate that the distilled signals will

demonstrate superior decoding generalization. This phenomenon is evident in Fig. 2 and Fig. S1, where the decoding performance of the distilled signals surpasses that of the raw signals, albeit not by a substantial margin.

Regarding “Next, if I understand correctly, distilled signals were obtained from the full model. How does a model perform trained only on the weakly tuned neurons? Is it possible that the subspaces obtained with the model are just not optimally aligned for decoding?”：Distilled signals (involving all neurons) are obtained by d-VAE. Subsequently, we use ANN to evaluate the performance of smaller and larger R2 neurons. Please note that separating and evaluating relevant signals are two distinct stages.

Regarding the reasoning in the section arguing that smaller R2 neurons encode rich information, we would like to provide a detailed explanation:

1. After extracting relevant signals through d-VAE, we specifically selected neurons characterized by smaller R2 values (Here, R2 signifies the proportion of neuronal activity variance explained by the linear encoding model, calculated using raw signals). Subsequently, we employed both KF and ANN to assess the decoding performance of these neurons. Remarkably, our findings revealed that smaller R2 neurons, previously believed to carry limited behavioral information, indeed encode rich information.

2. In a subsequent step, we employed d-VAE to exclusively distill the raw signals of these smaller R2 neurons (distinct from the earlier experiment where d-VAE processed signals from all neurons). We then employed KF and ANN to evaluate the distilled smaller R2 neurons. Interestingly, we observed that we could not attain the same richness of information solely through the use of these smaller R2 neurons.

3. Consequently, we put forth and tested two hypotheses: First, that larger R2 neurons introduce additional signals into the smaller R2 neurons that do not exist in the real smaller R2 neurons. Second, that larger R2 neurons aid in restoring the original appearance of impaired smaller R2 neurons. Our proposed criteria and synthetic experiments substantiate the latter scenario.

Thank you for your valuable feedback.

> *Q4: “Finally, a more technical issue to note is related to the choice to learn a non-parametric prior instead of using a conventional Gaussian prior. How is this implemented? Is just a single sample taken during a forward pass? I worry this may be insufficient as this would not sample the prior well, and some other strategy such as importance sampling may be required (unless the prior is not relevant as it weakly contributed to the ELBO, in which case this choice seems not very relevant). Generally, it would be useful to see visualisations of the latent variables to see how information about behaviour is represented by the model.”*

Regarding "how to implement the prior?": Please refer to Equation 7 in the revised manuscript; we have added detailed descriptions in the revised manuscript.

Regarding "Generally, it would be useful to see visualizations of the latent variables to see how information about behavior is represented by the model.": Note that our focus is not on latent variables but on distilled relevant signals. Nonetheless, at your request, we have added the visualization of latent variables in the revised manuscript. Please see Fig. S13 for details.

Thank you for your valuable feedback.

> *Recommendations: "A minor point: the word 'distill' in the name of the model may be a little misleading - in machine learning the term refers to the construction of smaller models with the same capabilities.*
>
> *It should be useful to add a schematic picture of the model to ease comparison with related approaches."*

In the context of our model's functions, it operates as a distillation process, eliminating irrelevant signals and retaining the relevant ones. Although the name of our model may be a little misleading, it faithfully reflects what our model does.

I have added a schematic picture of d-VAE in the revised manuscript. Please see Fig. S1 for details.

Thank you for your valuable feedback.

> ***Reviewer #2***
>
> *Q1: "Is the apparently increased complexity of encoding vs decoding so unexpected given the entropy, sparseness, and high dimensionality of neural signals (the "encoding") compared to the smoothness and low dimensionality of typical behavioural signals (the "decoding") recorded in neuroscience experiments? This is the title of the paper so it seems to be the main result on which the authors expect readers to focus. "*

We use the term "unexpected" due to the disparity between our findings and the prior understanding concerning neural encoding and decoding. For neural encoding, as we said in the Introduction, in previous studies, weakly-tuned neurons are considered useless, and smaller variance PCs are considered noise, but we found they encode rich behavioral information. For neural decoding, the nonlinear decoding performance of raw signals is significantly superior to linear decoding. However, after eliminating the interference of irrelevant signals, we found the linear decoding performance is comparable to nonlinear decoding. Rooted in these findings, which counter previous thought, we employ the term "unexpected" to characterize our observations.

Thank you for your valuable feedback.

> *Q2: "I take issue with the premise that signals in the brain are "irrelevant" simply because they do not correlate with a fixed temporal lag with a particular behavioural feature hand-chosen by the experimenter. As an example, the presence of a reward signal in motor cortex [1] after the movement is likely to be of little use from the perspective of predicting kinematics from time-bin to time-bin using a fixed model across trials (the apparent definition of "relevant" for behaviour here), but an entire sub-field of neuroscience is dedicated to understanding the impact of these reward-related signals on future behaviour. Is there method sophisticated enough to see the behavioural "relevance" of this brief, transient, post-movement signal? This may just be an issue of semantics, and perhaps I read too much into the choice of words here. Perhaps the authors truly treat "irrelevant" and "without a fixed temporal correlation" as synonymous phrases and the issue is easily resolved with a clarifying parenthetical the first time the word "irrelevant" is used. But I remain troubled by some claims in the paper which lead me to believe that they read more deeply into the "irrelevancy" of these components."*

In this paper, we employ terms like 'behaviorally-relevant' and 'behaviorally-irrelevant' only regarding behavioral variables of interest measured within a given task, such as arm kinematics during a motor control task. A similar definition can be found in the PSID[1].

Thank you for your valuable feedback.

[1] Sani, Omid G., et al. "Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification." Nature Neuroscience 24.1 (2021): 140-149.

> *Q3: "The authors claim the "irrelevant" responses underpin an unprecedented neuronal redundancy and reveal that movement behaviors are distributed in a higher-dimensional neural space than previously thought." Perhaps I just missed the logic, but I fail to see the evidence for this. The neural space is a fixed dimensionality based on the number of neurons. A more sparse and nonlinear distribution across this set of neurons may mean that linear methods such as PCA are not effective ways to approximate the dimensionality. But ultimately the behaviourally relevant signals seem quite low-dimensional in this paper even if they show some nonlinearity may help."*

The evidence for the "useless" responses underpin an unprecedented neuronal redundancy is shown in Fig. 5a, d and Fig. S9a. Specifically, the sum of the decoding performance of smaller R2 neurons and larger R2 neurons is significantly greater than that of all neurons for relevant signals (red bar), demonstrating that movement parameters are encoded very redundantly in neuronal population. In contrast, we can not find this degree of neural redundancy in raw signals (purple bar).

The evidence for the "useless" responses reveal that movement behaviors are distributed in a higher-dimensional neural space than previously thought is shown in the left plot (involving KF decoding) of Fig. 6c, f and Fig. S9f. Specifically, the improvement of KF using secondary signals is significantly higher than using raw signals composed of the same number of dimensions as the secondary signals. These results demonstrate that these dimensions, spanning roughly from ten to thirty, encode much information, suggesting that behavioral information exists in a higher-dimensional subspace than anticipated from raw signals.

Thank you for your valuable feedback.

> *Q5: "there is an apparent logical fallacy that begins in the abstract and persists in the paper: "Surprisingly, when incorporating often-ignored neural dimensions, behavioral information can be decoded linearly as accurately as nonlinear decoding, suggesting linear readout is performed in motor cortex." Don't get me wrong: the equivalency of linear and nonlinear decoding approaches on this dataset is interesting, and useful for neuroscientists in a practical sense. However, the paper expends much effort trying to make fundamental scientific claims that do not feel very strongly supported. This reviewer fails to see what we can learn about a set of neurons in the brain which are presumed to "read out" from motor cortex. These neurons will not have access to the data analyzed here. That a linear model can be conceived by an experimenter does not imply that the brain must use a linear model. The claim may be true, and it may well be that a linear readout is implemented in the brain. Other work [2,3] has shown that linear readouts of nonlinear neural activity patterns can explain some behavioural features. The claim in this paper, however, is not given enough"*

Due to the limitations of current observational methods and our incomplete understanding of brain mechanisms, it is indeed challenging to ascertain the specific data the brain acquires to generate behavior and whether it employs a linear readout. Conventionally, the neural data recorded in the motor cortex do encode movement behaviors and can be used to analyze neural encoding and decoding. Based on these data, we found that the linear decoder KF achieves comparable performance to that of the nonlinear decoder ANN on distilled relevant signals. This finding has undergone validation across three widely used datasets, providing substantial evidence. Furthermore, we conducted experiments on synthetic data to show that

this conclusion is not a by-product of our model. In the revised manuscript, we added a more detailed description of this conclusion.

Thank you for your valuable feedback.

> *Q6: "Relatedly, I would like to note that the exercise of arbitrarily dividing a continuous distribution of a statistic (the "R2") based on an arbitrary threshold is a conceptually flawed exercise. The authors read too much into the fact that neurons which have a low R2 w.r.t. PDs have behavioural information w.r.t. other methods. To this reviewer, it speaks more about the irrelevance, so to speak, of the preferred direction metric than anything fundamental about the brain."*

We chose the R2 threshold in accordance with the guidelines provided in reference [1]. It's worth mentioning that this threshold does not exert any significant influence on the overall conclusions.

Thank you for your valuable feedback.

[1] Inoue, Y., Mao, H., Suway, S.B., Orellana, J. and Schwartz, A.B., 2018. Decoding arm speed during reaching. Nature communications, 9(1), p.5243.

> *Q7: "I am afraid I may be missing something, as I did not understand the fano factor analysis of Figure 3. In a sense the behaviourally relevant signals must have lower FF given they are in effect tied to the temporally smooth (and consistent on average across trials) behavioural covariates. The point of the original Churchland paper was to show that producing a behaviour squelches the variance; naturally these must appear in the behaviourally relevant components. A control distribution or reference of some type would possibly help here."*

We agree that including reference signals could provide more context. The Churchland paper said stimulus onset can lead to a reduction in neural variability. However, our experiment focuses specifically on the reaching process, and thus, we don't have comparative experiments involving different types of signals.

Thank you for your valuable feedback.

> *Q8: "The authors compare the method to LFADS. While this is a reasonable benchmark as a prominent method in the field, LFADS does not attempt to solve the same problem as d-VAE. A better and much more fair comparison would be TNDM [4], an extension of LFADS which is designed to identify behaviourally relevant dimensions."*

We have added the comparison experiments with TNDM in the revised manuscript (see Fig. 2 and Fig. S2). The details of model hyperparameters and training settings can be found in the Methods section in the revised manuscripts.

Thank you for your valuable feedback.

> ***Reviewer #3***
>
> *Q1.1: "TNDM: LFADS is not the best baseline for comparison. The authors should have compared with TNDM (Hurwitz et al. 2021), which is an extension of LFADS that (unlike LFADS) actually attempts to extract behaviorally relevant factors by adding a behavior term to the loss. The code for TNDM is also available on Github. LFADS is not even supervised by behavior and does not aim to address the problem that d-VAE aims to address, so it is not the most appropriate comparison. "*

We have added the comparison experiments with TNDM in the revised manuscript (see Fig. 2 and Fig. S2). The details of model hyperparameters and training settings can be found in the Methods section in the revised manuscripts.

Thank you for your valuable feedback.

> *Q1.2: "LFADS: LFADS is a sequential autoencoder that processes sections of data (e.g. trials). No explanation is given in Methods for how the data was passed to LFADS. Was the moving averaged smoothed data passed to LFADS or the raw spiking data (at what bin size)? Was a gaussian loss used or a poisson loss? What are the trial lengths used in each dataset, from which part of trials? For dataset C that has back-to-back reaches, was data chopped into segments? How long were these segments? Were the edges of segments overlapped and averaged as in (Keshtkaran et al. 2022) to avoid noisy segment edges or not? These are all critical details that are not explained. The same details would also be needed for a TNDM comparison (comment 1.1) since it has largely the same architecture as LFADS.*
>
> *It is also critical to briefly discuss these fundamental differences between the inputs of methods in the main text. LFADS uses a segment of data whereas VAE methods just use one sample at a time. What does this imply in the results? I guess as long as VAEs outperform LFADS it is ok, but if LFADS outperforms VAEs in a given metric, could it be because it received more data as input (a whole segment)? Why was the factor dimension set to 50? I presume it was to match the latent dimension of the VAE methods, but is the LFADS factor dimension the correct match for that to make things comparable?*
>
> *I am also surprised by the results. How do the authors justify LFADS having lower neural similarity (fig 2d) than VAE methods that operate on single time steps? LFADS is not supervised by behavior, so of course I don't expect it to necessarily outperform methods on behavior decoding. But all LFADS aims to do is to reconstruct the neural data so at least in this metric it should be able to outperform VAEs that just operate on single time steps? Is it because LFADS smooths the data too much? This is important to discuss and show examples of. These are all critical nuances that need to be discussed to validate the results and interpret them."*

Regarding "Was the moving averaged smoothed data passed to LFADS or the raw spiking data (at what bin size)? Was a gaussian loss used or a poisson loss?": The data used by all models was applied to the same preprocessing procedure. That is, using moving averaged smoothed data with three bins, where the bin size is 100ms. For all models except PSID, we used a Poisson loss.

Regrading "What are the trial lengths used in each dataset, from which part of trials? For dataset C that has back-to-back reaches, was data chopped into segments? How long were these segments? Were the edges of segments overlapped and averaged as in (Keshtkaran et al. 2022) to avoid noisy segment edges or not?":

For datasets A and B, a trial length of eighteen is set. Trials with lengths below the threshold are zero-padded, while trials exceeding the threshold are truncated to the threshold length from their starting point. In dataset A, there are several trials with lengths considerably longer than that of most trials. We found that padding all trials with zeros to reach the maximum length (32) led to poor performance. Consequently, we chose a trial length of eighteen, effectively encompassing the durations of most trials and leading to the removal of approximately 9% of samples. For dataset B (center-out), the trial lengths are relatively consistent with small variation, and the maximum length across all trials is eighteen. For dataset C, we set the trial length as ten because we observed the video of this paradigm and

found that the time for completing a single trial was approximately one second. The segments are not overlapped.

Regarding “Why was the factor dimension set to 50? I presume it was to match the latent dimension of the VAE methods, but is the LFADS factor dimension the correct match for that to make things comparable?”: We performed a grid search for latent dimensions in {10,20,50} and found 50 is the best.

Regarding “I am also surprised by the results. How do the authors justify LFADS having lower neural similarity (fig 2d) than VAE methods that operate on single time steps? LFADS is not supervised by behavior, so of course I don't expect it to necessarily outperform methods on behavior decoding. But all LFADS aims to do is to reconstruct the neural data so at least in this metric it should be able to outperform VAEs that just operate on single time steps? Is it because LFADS smooths the data too much?”: As you pointed out, we found that LFADS tends to produce excessively smooth and consistent data, which can lead to a reduction in neural similarity.

Thank you for your valuable feedback.

> *Q1.3: “PSID: PSID is linear and uses past input samples to predict the next sample in the output. Again, some setup choices are not well justified, and some details are left out in the 1-line explanation given in Methods.*

Why was a latent dimension of 6 chosen? Is this the behaviorally relevant latent dimension or the total latent dimension (for the use case here it would make sense to set all latent states to be behaviorally relevant)? Why was a horizon hyperparameter of 3 chosen? First, it is important to mention fundamental parameters such as latent dimension for each method in the main text (not just in methods) to make the results interpretable. Second, these hyperparameters should be chosen with a grid search in each dataset (within the training data, based on performance on the validation part of the training data), just as the authors do for their method (line 779). Given that PSID isn't a deep learning method, doing a thorough grid search in each fold should be quite feasible. It is important that high values for latent dimension and a wider range of other hyperparmeters are included in the search, because based on how well the residuals (x_i) for this method are shown predict behavior in Fig 2, the method seems to not have been used appropriately. I would expect ANN to improve decoding for PSID versus its KF decoding since PSID is fully linear, but I don't expect KF to be able to decode so well using the residuals of PSID if the method is used correctly to extract all behaviorally relevant information from neural data. The low neural reconstruction in Fid 2d could also partly be due to using too small of a latent dimension.

Again, another import nuance is the input to this method and how differs with the input to VAE methods. The learned PSID model is a filter that operates on all past samples of input to predict the output in the "next" time step. To enable a fair comparison with VAE methods, the authors should make sure that the last sample "seen" by PSID is the same as then input sample seen by VAE methods. This is absolutely critical given how large the time steps are, otherwise PSID might underperform simply because it stopped receiving input 300ms earlier than the input received by VAE methods. To fix this, I think the authors can just shift the training and testing neural time series of PSID by 1 sample into the past (relative to the behavior), so that PSID's input would include the input of VAE methods. Otherwise, VAEs outperforming PSID is confounded by PSID's input not including the time step that was provided to VAE.”

Thanks for your suggestions for letting PSID see the current neural observations. We did it per your suggestions and then performed a grid search for the hyperparameters for PSID. Specifically, we performed a grid search for the horizon hyperparameter in {2,3,4,5,6,7}. Since the relevant latent dimension should be lower than the horizon times the dimension of

behavior variables (two-dimensional velocity in this paper) and increasing the dimension will reach performance saturation, we directly set the relevant latent dimensions as the maximum. The horizon number of datasets A, B, C, and synthetic datasets is 7, 6, 6 and 5, respectively.

And thus the latent dimension of datasets A, B, and C and the synthetic dataset is 14, 12, 12 and 10, respectively.

Our experiments show that KF can decode information from irrelevant signals obtained by PSID. Although PSID extracts the linear part of raw signals, KF can still use the linear part of the residuals for decoding. The low reconstruction performance of PSID may be because the relationship between latent variables and neural signals is linear, and the relationship between latent variables and behaviors is also linear; this is equivalent to the linear relationship between behaviors and neural signals, and linear models can only explain a small fraction of neural signals.

Thank you for your valuable feedback.

> *Q1.4: "CEBRA: results for CEBRA are incomplete. Similarity to raw signals is not shown. Decoding of behaviorally irrelevant residuals for CEBRA is not shown. Per Fig. S2, CEBRA does better or similar ANN decoding in datasets A and C, is only slightly worse in Dataset B, so it is important to show the other key metrics otherwise it is unclear whether d-VAE has some tangible advantage over CEBRA in those 2 datasets or if they are similar in every metric. Finally, it would be better if the authors show the results for CEBRA on Fig. 2, just as is done for other methods because otherwise it is hard to compare all methods."*

CEBRA is a non-generative model, this model cannot generate behaviorally-relevant signals. Therefore, we only compared the decoding performance of latent embeddings of CEBRA and signals of d-VAE.

Thank you for your valuable feedback.

> *Q2: "Given the fact that d-VAE infers the latent (z) based on the population activity (x), claims about properties of the inferred behaviorally relevant signals (x_r) that attribute properties to individual neurons are confounded.*

The authors contrast their approach to population level approaches in that it infers behaviorally relevant signals for individual neurons. However, d-VAE is also a population method as it aggregates population information to infer the latent (z), from which behaviorally relevant part of the activity of each neuron (x_r) is inferred. The authors note this population level aggregation of information as a benefit of d-VAE, but only acknowledge it as a confound briefly in the context of one of their analyses (line 340): "The first is that the larger R2 neurons leak their information to the smaller R2 neurons, causing them contain too much behavioral information". They go on to dismiss this confounding possibility by showing that the inferred behaviorally relevant signal of each neuron is often most similar to its own raw signals (line 348-352) compared with all other neurons. They also provide another argument specific to that result section (i.e., residuals are not very behavior predictive), which is not general so I won't discuss it in depth here. These arguments however do not change the basic fact that d-VAE aggregates information from other neurons when extracting the behaviorally relevant activity of any given neuron, something that the authors note as a benefit of d-VAE in many instances. The fact that d-VAE aggregates population level info to give the inferred behaviorally relevant signal for each neuron confounds several key conclusions. For example, because information is aggregated across neurons, when trial to trial variability looks smoother after applying d-VAE (Fig 3i), or reveals better cosine tuning

(Fig 3b), or when neurons that were not very predictive of behavior become more predictive of behavior (Fig 5), one cannot really attribute the new smoother single trial activity or the improved decoding to the same single neurons; rather these new signals/performances include information from other neurons. Unless the connections of the encoder network (z=f(x)) is zero for all other neurons, one cannot claim that the inferred rates for the neuron are truly solely associated with that neuron. I believe this a fundamental property of a population level VAE, and simply makes the architecture unsuitable for claims regarding inherent properties of single neurons. This confound is partly why the first claim in the abstract are not supported by data: observing that neurons that don't predict behavior very well would predict it much better after applying d-VAE does not prove that these neurons themselves "encode rich[er] behavioral information in complex nonlinear ways" (i.e., the first conclusion highlighted in the abstract) because information was also aggregated from other neurons. The other reason why this claim is not supported by data is the characterization of the encoding for smaller R2 neurons as "complex nonlinear", which the method is not well equipped to tease apart from linear mappings as I explain in my comment 3."

We acknowledge that we cannot obtain the exact single neuronal activity that does not contain any information from other neurons. However, we believe our model can extract accurate approximation signals of the ground truth relevant signals. These signals preserve the inherent properties of single neuronal activity to some extent and can be used for analysis at the single-neuron level.

We believe d-VAE is a reasonable approach to extract effective relevant signals that preserve inherent properties of single neuronal activity for four key reasons:

1. d-VAE is a latent variable model that adheres to the neural population doctrine. The neural population doctrine posits that information is encoded within interconnected groups of neurons, with the existence of latent variables (neural modes) responsible for generating observable neuronal activity [1, 2]. If we can perfectly obtain the true generative model from latent variables to neuronal activity, then we can generate the activity of each neuron from hidden variables without containing any information from other neurons. However, without a complete understanding of the brain's encoding strategies (or generative model), we can only get the approximation signals of the ground truth signals.

2. After the generative model is established, we need to infer the parameters of the generative model and the distribution of latent variables. During the inference process, inference algorithms such as variational inference or EM algorithms will be used. Generally, the obtained latent variables are also approximations of the real latent variables. When inferring the latent variables, it is inevitable to aggregation the information of the neural population, and latent variables are derived through weighted combinations of neuronal populations [3].

This inference process is consistent with that of d-VAE (or VAE-based models).

1. Latent variables are derived from raw neural signals and used to explain raw neural signals. Considering the unknown ground truth of latent variables and behaviorally-relevant signals, it becomes evident that the only reliable reference at the signal level is the raw signals. A crucial criterion for evaluating the reliability of latent variable models (including latent variables and generated relevant signals) is their capability to effectively explain the raw signals [3]. Consequently, we firmly maintain the belief that if the generated signals closely resemble the raw signals to the greatest extent possible, in accordance with an equivalence principle, we can claim that these obtained signals faithfully retain the inherent properties of single neurons. d-VAE explicitly constrains the generated signal to closely resemble the raw signals. These results demonstrate that d-VAE can extract effective relevant signals that preserve inherent properties of single neuronal activity.

Based on the above reasons, we hold that generating single neuronal activities with the VAE framework is a reasonable approach. The remaining question is whether our model can obtain accurate relevant signals in the absence of ground truth. To our knowledge, in cases where the ground truth of relevant signals is unknown, there are typically two approaches to verifying the reliability of extracted signals:

1. Conducting synthetic experiments where the ground truth is known.
2. Validation based on expert knowledge (Three criteria were proposed in this paper). Both our extracted signals and key conclusions have been validated using these two approaches.

Next, we will provide a detailed response to the concerns regarding our first key conclusion that smaller R2 neurons encode rich information.

We acknowledge that larger R2 neurons play a role in aiding the reconstruction of signals in smaller R2 neurons through their neural activity. However, considering that neurons are correlated rather than independent entities, we maintain the belief that larger R2 neurons assist damaged smaller R2 neurons in restoring their original appearance. Taking image denoising as an example, when restoring noisy pixels to their original appearance, relying solely on the noisy pixels themselves is often impractical. Assistance from their correlated, clean neighboring pixels becomes necessary.

The case we need to be cautious of is that the larger R2 neurons introduce additional signals (m) that contain substantial information to smaller R2 neurons, which they do not inherently possess. We believe this case does not hold for two reasons. Firstly, logically, adding extra signals decreases the reconstruction performance, and the information carried by these additional signals is redundant for larger R2 neurons, thus they do not introduce new information that can enhance the decoding performance of the neural population. Therefore, it seems unlikely and unnecessary for neural networks to engage in such counterproductive actions. Secondly, even if this occurs, our second criterion can effectively exclude the selection of these signals. To clarify, if we assume that x, y, and z denote the raw, relevant, and irrelevant signals of smaller R2 neurons, with x=y+z, and the extracted relevant signals become y+m, the irrelevant signals become z-m in this case. Consequently, the irrelevant signals contain a significant amount of information. It's essential to emphasize that this criterion holds significant importance in excluding undesirable signals.

Furthermore, we conducted a synthetic experiment to show that d-VAE can indeed restore the damaged information of smaller R2 neurons with the help of larger R2 neurons, and the restored neuronal activities are more similar to ground truth compared to damaged raw signals. Please see Fig. S11a,b for details.

Thank you for your valuable feedback.

[1] Saxena, S. and Cunningham, J.P., 2019. Towards the neural population doctrine. Current opinion in neurobiology, 55, pp.103-111.

[2] Gallego, J.A., Perich, M.G., Miller, L.E. and Solla, S.A., 2017. Neural manifolds for the control of movement. Neuron, 94(5), pp.978-984.

[3] Cunningham, J.P. and Yu, B.M., 2014. Dimensionality reduction for large-scale neural recordings. Nature neuroscience, 17(11), pp.1500-1509.

> *Q3: "Given the nonlinear architecture of the VAE, claims about the linearity or nonlinearity of cortical readout are confounded and not supported by the results.*
>
> *The inference of behaviorally relevant signals from raw signals is a nonlinear operation, that is x_r=g(f(x)) is nonlinear function of x. So even when a linear KF is used to decode behavior from the inferred behaviorally relevant signals, the overall decoding from raw signals to predicted behavior (i.e., KF applied to g(f(x))) is nonlinear. Thus, the result that decoding of behavior from inferred behaviorally relevant signals (x_r) using a linear KF and a nonlinear ANN reaches similar accuracy (Fig 2), does not suggest that a "linear readout is performed in the motor cortex", as the authors claim (line 471). The authors acknowledge this confound (line 472) but fail to address it adequately. They perform a simulation analysis where the decoding gap between KF and ANN remains unchanged even when d-VAE is used to infer behaviorally relevant signals in the simulation. However, this analysis is not enough for "eliminating the doubt" regarding the confound. I'm sure the authors can also design simulations where the opposite happens and just like in the data, d-VAE can improve linear decoding to match ANN decoding. An adequate way to address this concern would be to use a fully linear version of the autoencoder where the f(.) and g(.) mappings are fully linear. They can simply replace these two networks in their model with affine mappings, redo the modeling and see if the model still helps the KF decoding accuracy reach that of the ANN decoding. In such a scenario, because the overall KF decoding from original raw signals to predicted behavior (linear d-VAE + KF) is linear, then they could move toward the claim that the readout is linear. Even though such a conclusion would still be impaired by the nonlinear reference (d-VAE + ANN decoding) because the achieved nonlinear decoding performance could always be limited by network design and fitting issues. Overall, the third conclusion highlighted in the abstract is a very difficult claim to prove and is unfortunately not supported by the results."*

We aim to explore the readout mechanism of behaviorally-relevant signals, rather than raw signals. Theoretically, the process of removing irrelevant signals should not be considered part of the inherent decoding mechanisms of the relevant signals. Assuming that the relevant signals we extracted are accurate, the conclusion of linear readout is established. On the synthetic data where the ground truth is known, our distilled signals show a significant improvement in neural similarity to the ground truth when compared to raw signals (refer to Fig. S2l). This observation demonstrates that our distilled signals are accurate approximations of the ground truth. Furthermore, on the three widely-used real datasets, our distilled signals meet the stringent criteria we have proposed (see Fig. 2), also providing strong evidence for their accuracy.

Regarding the assertion that we could create simulations in which d-VAE can make signals that are inherently nonlinearly decodable into linearly decodable ones: In reality, we cannot achieve this, as the second criterion can rule out the selection of such signals. Specifically,z=x+y=n^2+y, where z, x, y, and n denote raw signals, relevant signals, irrelevant signals and latent variables. If the relevant signals obtained by d-VAE are n, then these signals

can be linear decoded accurately. However, the corresponding irrelevant signals are n^2-n+z; thus, irrelevant signals will have much information, and these extracted relevant signals will not be selected. Furthermore, our synthetic experiments offer additional evidence supporting the conclusion that d-VAE does not make inherently nonlinearly decodable signals become linearly decodable ones. As depicted in Fig. S11c, there exists a significant performance gap between KF and ANN when decoding the ground truth signals of smaller R2 neurons. KF exhibits notably low performance, leaving substantial room for compensation by d-VAE. However, following processing by d-VAE, KF's performance of distilled signals fails to surpass its already low ground truth performance and remains significantly inferior to ANN's performance. These results collectively confirm that our approach does not convert signals that are inherently nonlinearly decodable into linearly decodable ones, and the conclusion of linear readout is not a by-product by d-VAE.

Regarding the suggestion of using linear d-VAE + KF, as discussed in the Discussion section, removing the irrelevant signals requires a nonlinear operation, and linear d-VAE can not effectively separate relevant and irrelevant signals.

Thank you for your valuable feedback.

> *Q4: "The authors interpret several results as indications that "behavioral information is distributed in a higher-dimensional subspace than expected from raw signals", which is the second main conclusion highlighted in the abstract. However, several of these arguments do not convincingly support that conclusion.*
>
> *4.1) The authors observe that behaviorally relevant signals for neurons with small principal components (referred to as secondary) have worse decoding with KF but better decoding with ANN (Fig. 6b,e), which also outperforms ANN decoding from raw signals. This observation is taken to suggest that these secondary behaviorally relevant signals encode behavior information in highly nonlinear ways and in a higher dimensions neural space than expected (lines 424 and 428). These conclusions however are confounded by the fact that A) d-VAE uses nonlinear encoding, so one cannot conclude from ANN outperforming KF that behavior is encoded nonlinearly in the motor cortex (see comment 3 above), and B) d-VAE aggregates information across the population so one cannot conclude that these secondary neurons themselves had as much behavior information (see comment 2 above).*
>
> *4.2) The authors observe that the addition of the inferred behaviorally relevant signals for neurons with small principal components (referred to as secondary) improves the decoding of KF more than it improves the decoding of ANN (red curves in Fig 6c,f). This again is interpreted similarly as in 4.1, and is confounded for similar reasons (line 439): "These results demonstrate that irrelevant signals conceal the smaller variance PC signals, making their encoded information difficult to be linearly decoded, suggesting that behavioral information exists in a higher-dimensional subspace than anticipated from raw signals". This is confounded by because of the two reasons explained in 4.1. To conclude nonlinear encoding based on the difference in KF and ANN decoding, the authors would need to make the encoding/decoding in their VAE linear to have a fully linear decoder on one hand (with linear d-VAE + KF) and a nonlinear decoder on the other hand (with linear d-VAE + ANN), as explained in comment 3.*
>
> *4.3) From S Fig 8, where the authors compare cumulative variance of PCs for raw and inferred behaviorally relevant signals, the authors conclude that (line 554): "behaviorally-irrelevant signals can cause an overestimation of the* neural dimensionality of behaviorally-relevant responses *(Supplementary Fig. S8)." However, this analysis does not really say anything about overestimation of "behaviorally relevant" neural dimensionality since the comparison is done with the dimensionality of "raw" signals. The next sentence is ok though: "These findings highlight the need to filter out relevant*

> *signals when estimating the* neural dimensionality.*", because they use the phrase "neural dimensionality" not "neural dimensionality of behaviorally-relevant responses".”*

Questions 4.1 and 4.2 are a combination of Q2 and Q3. Please refer to our responses to Q2 and Q3.

Regarding question 4.3 about “behaviorally-irrelevant signals can cause an overestimation of the neural dimensionality of behaviorally-relevant responses”: Previous studies usually used raw signals to estimate the neural dimensionality of specific behaviors. We mean that using raw signals, which include many irrelevant signals, will cause an overestimation of the neural dimensionality. We have modified this sentence in the revised manuscripts.

Thank you for your valuable feedback.

> *Q5: “Imprecise use of language in many places leads to inaccurate statements. I will list some of these statements”*
>
> *5.1) In the abstract: "One solution is to accurately separate behaviorally-relevant and irrelevant signals, but this approach remains elusive due to the unknown ground truth of behaviorally-relevant signals". This statement is not accurate because it implies no prior work does this. The authors should make their statement more specific and also refer to some goal that existing linear (e.g., PSID) and nonlinear (e.g., TNDM) methods for extracting behaviorally relevant signals fail to achieve.*
>
> *5.2) In the abstract: "we found neural responses previously considered useless encode rich behavioral information" => what does "useless" mean operationally? Low behavior tuning? More precise use of language would be better.*
>
> *5.3) "... recent studies (Glaser 58 et al., 2020; Willsey et al., 2022) demonstrate nonlinear readout outperforms linear readout." => do these studies show that nonlinear "readout" outperforms linear "readout", or just that nonlinear models outperform linear models?*
>
> *5.4) Line 144: "The first criterion is that the decoding performance of the behaviorally-relevant signals (red bar, Fig.1) should surpass that of raw signals (the red dotted line, Fig.1).". Do the authors mean* linear *decoding here or decoding in general? If the latter, how can something extracted from neural* surpass *decoding of neural data, when the extraction itself can be thought of as part of decoding? The operational definition for this "decoding performance" should be clarified.*
>
> *5.5) Line 311: "we found that the dimensionality of primary subspace of raw signals (26, 64, and 45 for datasets A, B, and C) is significantly higher than that of behaviorally-relevant signals (7, 13, and 9), indicating that behaviorally-irrelevant signals lead to an overestimation of the neural dimensionality of behaviorally-relevant signals." => here the dimensionality of the total PC space (i.e., primary subspace of raw signals) is being compared with that of inferred behaviorally-relevant signals, so the former being higher does* not *indicate that neural dimensionality of* behaviorally-relevant *signals was overestimated. The former is simply not behavioral so this conclusion is not accurate.*
>
> *5.6) Section "Distilled behaviorally-relevant signals uncover that smaller R2 neurons encode rich behavioral information in complex nonlinear ways". Based on what kind of R2 are the neurons grouped? Behavior decoding R2 from raw signals? Using what mapping? Using KF? If KF is used, the result that small R2 neurons benefit a lot from d-VAE could be somewhat expected, given the nonlinearity of d-VAE: because only ANN would have the capacity to unwrap the nonlinear encoding of d-VAE as needed. If decoding performance that is used to group neurons is based on data, regression to the mean could also partially explain the result: the neurons with worst raw decoding are most likely to benefit from a change in decoder, than neurons that already had good*

> *decoding. In any case, the R2 used to partition and sort neurons should be more clearly stated and reminded throughout the text and I Fig 3.*
>
> *5.7) Line 346 "...it is impossible for our model to add the activity of larger R2 neurons to that of smaller R2 neurons" => Is it really* impossible*? The optimization can definitely add small-scale copies of behaviorally relevant information to all neurons with minimal increase in the overall optimization loss, so this statement seems inaccurate.*
>
> *5.8) Line 490: "we found that linear decoders can achieve comparable performance to that of nonlinear decoders, providing compelling evidence for the presence of linear readout in the motor cortex." => inaccurate because no d-VAE decoding is really linear, as explained in comment 3 above.*
>
> *5.9) Line 578: ". However, our results challenge this idea by showing that signals composed of smaller variance PCs nonlinearly encode a significant amount of behavioral information." => inaccurate as results are confounded by nonlinearity of d-VAE as explained in comment 3 above.*
>
> *5.10) Line 592: "By filtering out behaviorally-irrelevant signals, our study found that accurate decoding performance can be achieved through linear readout, suggesting that the motor cortex may perform linear readout to generate movement behaviors." => inaccurate because it us confounded by the nonlinearity of d-VAE as explained in comment 3 above."*

Regarding "5.1) In the abstract: "One solution is to accurately separate behaviorally-relevant and irrelevant signals, but this approach remains elusive due to the unknown ground truth of behaviorally-relevant signals". This statement is not accurate because it implies no prior work does this. The authors should make their statement more specific and also refer to some goal that existing linear (e.g., PSID) and nonlinear (e.g., TNDM) methods for extracting behaviorally relevant signals fail to achieve":

We believe our statement is accurate. Our primary objective is to extract accurate behaviorally-relevant signals that closely approximate the ground truth relevant signals. To achieve this, we strike a balance between the reconstruction and decoding performance of the generated signals, aiming to effectively capture the relevant signals. This crucial aspect of our approach sets it apart from other methods. In contrast, other methods tend to emphasize the extraction of valuable latent neural dynamics. We have provided elaboration on the distinctions between d-VAE and other approaches in the Introduction and Discussion sections.

Thank you for your valuable feedback.

Regarding "5.2) In the abstract: "we found neural responses previously considered useless encode rich behavioral information" => what does "useless" mean operationally? Low behavior tuning? More precise use of language would be better.":

In the analysis of neural signals, smaller variance PC signals are typically seen as noise and are often discarded. Similarly, smaller R2 neurons are commonly thought to be dominated by noise and are not further analyzed. Given these considerations, we believe that the term "considered useless" is appropriate in this context. Thank you for your valuable feedback.

Regarding "5.3) "... recent studies (Glaser 58 et al., 2020; Willsey et al., 2022) demonstrate nonlinear readout outperforms linear readout." => do these studies show that nonlinear "readout" outperforms linear "readout", or just that nonlinear models outperform linear models?":

In this paper, we consider the two statements to be equivalent. Thank you for your valuable feedback.

Regarding “5.4) Line 144: "The first criterion is that the decoding performance of the behaviorally-relevant signals (red bar, Fig.1) should surpass that of raw signals (the red dotted line, Fig.1).". Do the authors mean *linear* decoding here or decoding in general? If the latter, how can something extracted from neural *surpass* decoding of neural data, when the extraction itself can be thought of as part of decoding? The operational definition for this "decoding performance" should be clarified.”:

We mean the latter, as we said in the section “Framework for defining, extracting, and separating behaviorally-relevant signals”, since raw signals contain too many behaviorally-irrelevant signals, deep neural networks are more prone to overfit raw signals than relevant signals. Therefore the decoding performance of relevant signals should surpass that of raw signals. Thank you for your valuable feedback.

Regarding “5.5) Line 311: "we found that the dimensionality of primary subspace of raw signals (26, 64, and 45 for datasets A, B, and C) is significantly higher than that of behaviorally-relevant signals (7, 13, and 9), indicating that behaviorally-irrelevant signals lead to an overestimation of the neural dimensionality of behaviorally-relevant signals." => here the dimensionality of the total PC space (i.e., primary subspace of raw signals) is being compared with that of inferred behaviorally-relevant signals, so the former being higher does *not* indicate that neural dimensionality of *behaviorally-relevant* signals was overestimated. The former is simply not behavioral so this conclusion is not accurate.”: In practice, researchers usually used raw signals to estimate the neural dimensionality. We mean that using raw signals to do this would overestimate the neural dimensionality. Thank you for your valuable feedback.

Regarding “5.6) Section "Distilled behaviorally-relevant signals uncover that smaller R2 neurons encode rich behavioral information in complex nonlinear ways". Based on what kind of R2 are the neurons grouped? Behavior decoding R2 from raw signals? Using what mapping? Using KF? If KF is used, the result that small R2 neurons benefit a lot from d-VAE could be somewhat expected, given the nonlinearity of d-VAE: because only ANN would have the capacity to unwrap the nonlinear encoding of d-VAE as needed. If decoding performance that is used to group neurons is based on data, regression to the mean could also partially explain the result: the neurons with worst raw decoding are most likely to benefit from a change in decoder, than neurons that already had good decoding. In any case, the R2 used to partition and sort neurons should be more clearly stated and reminded throughout the text and I Fig 3.”:

When employing R2 to characterize neurons, it indicates the extent to which neuronal activity is explained by the linear encoding model [1-3]. Smaller R2 neurons have a lower capacity for linearly tuning (encoding) behaviors, while larger R2 neurons have a higher capacity for linearly tuning (encoding) behaviors. Specifically, the approach involves first establishing an encoding relationship from velocity to neural signal using a linear model, i.e., y=f(x), where f represents a linear regression model, x denotes velocity, and y denotes the neural signal. Subsequently, R2 is utilized to quantify the effectiveness of the linear encoding model in explaining neural activity. We have provided a comprehensive explanation in the revised manuscript. Thank you for your valuable feedback.

[1] Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J., Velliste, M., Boninger, M.L. and Schwartz, A.B., 2013. High-performance neuroprosthetic control by an individual with tetraplegia. The Lancet, 381(9866), pp.557-564.

[2] Wodlinger, B., et al. "Ten-dimensional anthropomorphic arm control in a human brain–machine interface: difficulties, solutions, and limitations." Journal of neural engineering 12.1

(2014): 016011.

[3] Inoue, Y., Mao, H., Suway, S.B., Orellana, J. and Schwartz, A.B., 2018. Decoding arm speed during reaching. Nature communications, 9(1), p.5243.

Regarding Questions 5.7, 5.8, 5.9, and 5.10:

We believe our conclusions are solid. The reasons can be found in our replies in Q2 and Q3. Thank you for your valuable feedback.

> *Q6: "Imprecise use of language also sometimes is not inaccurate but just makes the text hard to follow.*
>
> *6.1) Line 41: "about neural encoding and decoding mechanisms" => what is the definition of encoding/decoding and how do these differ? The definitions given much later in line 77-79 is also not clear.*
>
> *6.2) Line 323: remind the reader about what R2 is being discussed, e.g., R2 of decoding behavior using KF. It is critical to know if linear or nonlinear decoding is being discussed.*
>
> *6.3) Line 488: "we found that neural responses previously considered trivial encode rich behavioral information in complex nonlinear ways" => "trivial" in what sense? These phrases would benefit from more precision, for example: "neurons that may seem to have little or no behavior information encoded". The same imprecise word ("trivial") is also used in many other places, for example in the caption of Fig S9.*
>
> *6.4) Line 611: "The same should be true for the brain." => Too strong of a statement for an unsupported claim suggesting the brain does something along the lines of nonlin VAE + linear readout.*
>
> *6.5) In Fig 1, legend: what is the operational definition of "generating performance"? Generating what? Neural reconstruction?"*

Regarding "6.1) Line 41: "about neural encoding and decoding mechanisms" => what is the definition of encoding/decoding and how do these differ? The definitions given much later in line 77-79 is also not clear.":

We would like to provide a detailed explanation of neural encoding and decoding. Neural encoding means how neuronal activity encodes the behaviors, that is, y=f(x), where y denotes neural activity and, x denotes behaviors, f is the encoding model. Neural decoding means how the brain decodes behaviors from neural activity, that is, x=g(y), where g is the decoding model. For further elaboration, please refer to [1]. We have included references that discuss the concepts of encoding and decoding in the revised manuscript. Thank you for your valuable feedback.

[1] Kriegeskorte, Nikolaus, and Pamela K. Douglas. "Interpreting encoding and decoding models." Current opinion in neurobiology 55 (2019): 167-179.

Regarding "6.2) Line 323: remind the reader about what R2 is being discussed, e.g., R2 of decoding behavior using KF. It is critical to know if linear or nonlinear decoding is being discussed.":

This question is the same as Q5.6. Please refer to the response to Q5.6. Thank you for your valuable feedback.

Regarding "6.3) Line 488: "we found that neural responses previously considered trivial encode rich behavioral information in complex nonlinear ways" => "trivial" in what sense? These phrases would benefit from more precision, for example: "neurons that may seem to

have little or no behavior information encoded". The same imprecise word ("trivial") is also used in many other places, for example in the caption of Fig S9.":

We have revised this statement in the revised manuscript. Thanks for your recommendation.

Regarding "6.4) Line 611: "The same should be true for the brain." => Too strong of a statement for an unsupported claim suggesting the brain does something along the lines of nonlin VAE + linear readout."

We mean that removing the interference of irrelevant signals and decoding the relevant signals should logically be two stages. We have revised this statement in the revised manuscript. Thank you for your valuable feedback.

Regarding "6.5) In Fig 1, legend: what is the operational definition of "generating performance"? Generating what? Neural reconstruction?"":

We have replaced "generating performance" with "reconstruction performance" in the revised manuscript. Thanks for your recommendation.

> *Q7: "In the analysis presented starting in line 449, the authors compare improvement gained for decoding various speed ranges by adding secondary (small PC) neurons to the KF decoder (Fig S11). Why is this done using the KF decoder, when earlier results suggest an ANN decoder is needed for accurate decoding from these small PC neurons? It makes sense to use the more accurate nonlinear ANN decoder to support the fundamental claim made here, that smaller variance PCs are involved in regulating precise control"*

Because when the secondary signal is superimposed on the primary signal, the enhancement in KF performance is substantial. We wanted to explore in which aspect of the behavior the KF performance improvement is mainly reflected. In comparison, the improvement of ANN by the secondary signal is very small, rendering the exploration of the aforementioned questions inconsequential. Thank you for your valuable feedback.

> *Q8: "A key limitation of the VAE architecture is that it doesn't aggregate information over multiple time samples. This may be why the authors decided to use a very large bin size of 100ms and beyond that smooth the data with a moving average. This limitation should be clearly stated somewhere in contrast with methods that can aggregate information over time (e.g., TNDM, LFADS, PSID) "*

We have added this limitation in the Discussion in the revised manuscript. Thanks for your recommendation.

> *Q9: "Fig 5c and parts of the text explore the decoding when some neurons are dropped. These results should come with a reminder that dropping neurons from behaviorally relevant signals is not technically possible since the extraction of behaviorally relevant signals with d-VAE is a population level aggregation that requires the raw signal from all neurons as an input. This is also important to remind in some places in the text for example:*
>
> - *Line 498: "...when one of the neurons is destroyed."*
> - *Line 572: "In contrast, our results show that decoders maintain high performance on distilled signals even when many neurons drop out.""*

We want to explore the robustness of real relevant signals in the face of neuron drop-out. The signals our model extracted are an approximation of the ground truth relevant signals and

thus serve as a substitute for ground truth to study this problem. Thank you for your valuable feedback.

> *Q10: "Besides the confounded conclusions regarding the readout being linear (see comment 3 and items related to it in comment 5), the authors also don't adequately discuss prior works that suggest nonlinearity helps decoding of behavior from the motor cortex. Around line 594, a few works are discussed as support for the idea of a linear readout. This should be accompanied by a discussion of works that support a nonlinear encoding of behavior in the motor cortex, for example (Naufel et al. 2019; Glaser et al. 2020), some of which the authors cite elsewhere but don't discuss here."*

We have added this discussion in the revised manuscript. Thanks for your recommendation.

> *Q11: "Selection of hyperparameters is not clearly explained. Starting line 791, the authors give some explanation for one hyperparameter, but not others. How are the other hyperparameters determined? What is the search space for the grid search of each hyperparameter? Importantly, if hyperparameters are determined only based on the training data of each fold, why is only one value given for the hyperparameter selected in each dataset (line 814)? Did all 5 folds for each dataset happen to select exactly the same hyperparameter based on their 5 different training/validation data splits? That seems unlikely."*

We perform a grid search in {0.001, 0.01,0.1,1} for hyperparameter beta. And we found that 0.001 is the best for all datasets. As for the model parameters, such as hidden neuron numbers, this model capacity has reached saturation decoding performance and does not influence the results.

Regarding "Importantly, if hyperparameters are determined only based on the training data of each fold, why is only one value given for the hyperparameter selected in each dataset (line 814)? Did all 5 folds for each dataset happen to select exactly the same hyperparameter based on their 5 different training/validation data splits": We selected the hyperparameter based on the average performance of 5 folds data on validation sets. The selected value denotes the one that yields the highest average performance across the 5 folds data.

Thank you for your valuable feedback.

> *Q12: "d-VAE itself should also be explained more clearly in the main text. Currently, only the high-level idea of the objective is explained. The explanation should be more precise and include the idea of encoding to latent state, explain the relation to pip-VAE, explain inputs and outputs, linearity/nonlinearity of various mappings, etc. Also see comment 1 above, where I suggest adding more details about other methods in the main text."*

Our primary objective is to delve into the encoding and decoding mechanisms using the separated relevant signals. Therefore, providing an excessive amount of model details could potentially distract from the main focus of the paper. In response to your suggestion, we have included a visual representation of d-VAE's structure, input, and output (see Fig. S1) in the revised manuscript, which offers a comprehensive and intuitive overview. Additionally, we have expanded on the details of d-VAE and other methods in the Methods section.

Thank you for your valuable feedback.

> *Q13: "In Fig 1f and g, shouldn't the performance plots be swapped? The current plots seem counterintuitive. If there is bias toward decoding (panel g), why is the irrelevant residual so good at decoding?"*

The placement of the performance plots in Fig. 1f and 1g is accurate. When the model exhibits a bias toward decoding, it prioritizes extracting the most relevant features (latent variables) for decoding purposes. As a consequence, the model predominantly generates signals that are closely associated with these extracted features. This selective signal extraction and generation process may result in the exclusion of other potentially useful information, which will be left in the residuals. To illustrate this concept, consider the example of face recognition: if a model can accurately identify an individual using only the person's eyes (assuming these are the most useful features), other valuable information, such as details of the nose or mouth, will be left in the residuals, which could also be used to identify the individual.

Thank you for your valuable feedback.