Revealing unexpected mechanisms of complex encoding but simple decoding in motor cortex via task-relevant neural signal distillation

Yangang Li[1,2,4], **Yu Qi**[3,4] †, **Yueming Wang**[1,4] †

[1] Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[3] Affiliated 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
[4] The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China

Abstract. Understanding how motor cortex encodes and decodes behaviors is a fundamental goal of neuroscience, which faces significant challenges because task-relevant neural responses are entangled with considerable task-irrelevant nuisance signals. Due to the entanglement, it is difficult to determine to extract which part and how much neural activity to study neural mechanisms. To cope with this puzzle, current studies usually extract a subset of neural responses that satisfy empirical hypotheses from noisy signals for analysis, which may lead to a one-sided understanding of neural mechanisms. Here we proposed a new perspective — extracting complete task-relevant signals — to explore neural mechanisms and offered a distillation framework to distill signals. Analyzing data in three monkeys performing different reaching tasks, distilled task-relevant signals reveal unexpected complex encoding but simple decoding mechanisms in motor cortex.

† Corresponding authors: Yu Qi (qiyu@zju.edu.cn) and Yueming Wang (ymingwang@zju.edu.cn).

1. Introduction

Understanding how the motor cortex encodes and decodes movement behavior is a fundamental goal of neuroscience [1, 2]. This goal faces significant challenges because task-relevant neural responses are entangled with some task-irrelevant nuisance factors such as mixed selective responses for various other task variables [3, 4] and considerable noise [5, 6]. Due to the entanglement of relevant and irrelevant signals, it is difficult to distinguish which are true changes in underlying task-relevant neural activity and which are changes caused by irrelevant signals, which significantly complicates the study of the encoding and decoding mechanism. Thus, eliminating the influence of irrelevant signals is a necessary pathway to approaching the truth of neural mechanisms for motor control.

Over the past few decades, researchers directly used noisy signals to explore neural mechanisms. The research approaches are broadly divided into two doctrines, neuronal doctrine [7–10] (assumes single neurons directly tune external stimuli and relates neuronal activity to movement behavior by building the encoding model) and neural population doctrine [11–16] (speculates that the coordinated activation of the neural population serves as the basis for generating movement commands and relates identified population structures to movement behavior), both of which have advanced the understanding of neural mechanisms. However, to be faced with the interference of irrelevant signals, approaches from both doctrines encounter the same puzzle about which part and how much neural activity needs to be characterized. To cope with the puzzle, researchers chose strong neural responses that satisfy empirical hypotheses for analysis, such as simple encoding models [7,8] and low-dimensional population structures [11–14], which can reduce the influence of irrelevant signals to some extent. However, these strong neural patterns only explain a small subset of neural responses, while large amounts of signals that do not meet the hypothesized structures are discarded. The arising concerns are whether the extracted patterns are accurate under the interference of irrelevant signals, even if they are strong patterns. Whether the discarded signals appearing useless are critical to a comprehensive understanding of neural mechanisms? All these issues remain unclear yet.

Generally, using incomplete noisy signals for analysis may yield a one-sided understanding of neural mechanisms. Given the disadvantages of using incomplete noisy signals, we propose a new perspective — extracting complete task-relevant neural signals — to explore neural mechanisms. However, since the ground truth of task-relevant signals is unknown, methods of extracting complete task-relevant signals are elusive to date. To this end, we proposed a distillation framework that exploits the trade-off between encoding and decoding capabilities of task-relevant signals to distill task-relevant signals maximally. By dissociating relevant and irrelevant signals, analyzing noise-free neural responses benefits the study for both single-neuron and population levels and enables us to examine whether the extracted patterns are accurate, explore the function of discarded trivial neural signals, and look for discoveries that are difficult to be spotted in noisy signals.

Experiments were carried out with three datasets of different reaching tasks in monkeys. We found that trial-to-trial neuronal variability is mainly caused by task-irrelevant signals, and irrelevant signals are almost evenly distributed across neural dimensions of relevant signals and vice versa. Furthermore, irrelevant signals have a small effect on strong neural responses (signals composed of larger variance principal components (PCs) or larger R^2 neurons) but a large effect on trivial neural responses (signals composed of smaller variance PCs or smaller R^2 neurons) which are previously considered useless. Yet, these ignored activities encode unexpected rich task information in a complex nonlinear way, suggesting that task information is widely distributed on PCs and neurons, thus reflecting the complexity of the encoding mechanism of the motor cortex. By eliminating the interference of irrelevant signals, we surprisingly found that linear decoders achieve comparable performance to nonlinear decoders, which provides strong evidence for the linear readout of the neural population in the motor cortex and demonstrates the unexpected simplicity of decoding mechanisms. Moreover, our work brings a new perspective to developing robust and accurate BMIs and provides a powerful distillation tool, which can also be applied to other cortical data to discover new findings that are difficult to be spotted in noisy signals.

2. Results

2.1. Distillation framework for extracting task-relevant neural signals

2.1.1. What are task-relevant neural signals. Since the ground truth of task-relevant signals is unknown, the task-relevant signals are not well defined yet. Here we consider the task-relevant signals capable of thoroughly analyzing neural mechanisms should satisfy two requirements: the task-relevant signals should be similar to raw signals to maintain neuronal properties (encoding requirement); the task-relevant signals should contain task information maximally (decoding requirement). If satisfying the two requirements, we call the signals complete task-relevant signals.

2.1.2. How to distill task-relevant signals. Since raw signals contain numerous task-irrelevant signals, the distilling model is faced with the dilemma of how closely generated task-relevant signals should resemble raw signals. Specifically, if generated signals are too similar to raw signals, these signals will contain a large amount of task-irrelevant signals, which hinders exploring neural mechanisms; if generated signals are not similar to raw signals, generated signals may lose task information, affecting the comprehensive understanding of neural mechanisms. To address the dilemma, in the present study, we exploited the trade-off between the encoding and decoding capabilities of task-relevant signals to distill task-relevant signals (for details, see Methods). The core assumption of our model is that task-irrelevant signals are noise relative to task-relevant signals, which would harm the decoding generalization of generated task-relevant signals. Generally, the distilling model is faced with three cases: a bias toward generating (Fig. 1f), a bias

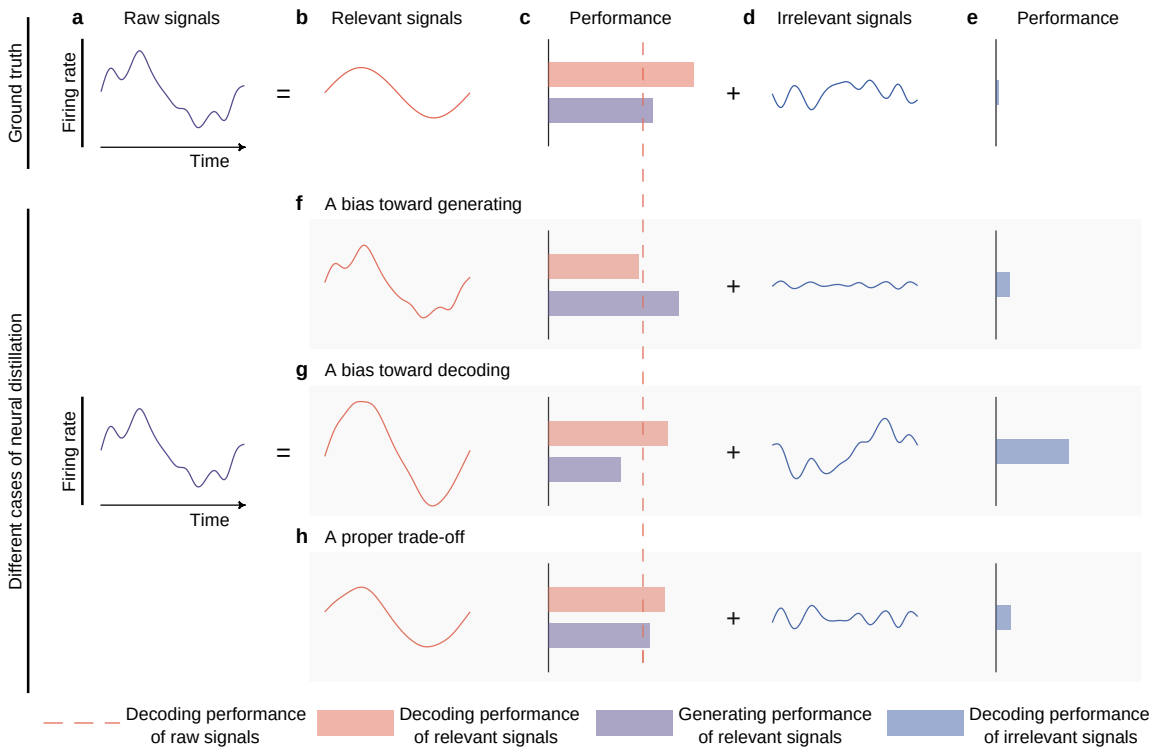

Figure 1: **Semantic illustration of distilling and evaluating task-relevant signals. a-e** The ideal decomposition of raw signals. **a**, The temporal firing activity of raw signals, where x-axis denotes time, and y-axis represents firing rate. Raw signals are decomposed to relevant (**b**) and irrelevant (**d**) signals. The ground truth of relevant signals decode task information perfectly (**c**, orange bar) and be similar to raw signals to some extent (**c**, purple bar), and the ground truth of irrelevant signals contain little task information (**e**). **f-h**, Three different cases of task-relevant signals distillation. **f**, When the distilling model is biased toward generating task-relevant signals that are similar to raw signals, the generating performance will be high while the decoding performance will be low because generating too many irrelevant signals will harm decoding performance. Since it is difficult for models to extract complete task-relevant signals, the residuals will also contain a small amount of task information. **g**, When the model is biased toward generating signals that contain decoding information while ignoring being similar to raw signals, the decoding performance is high and the generating performance is low; meanwhile, the residuals will have much task information. **g**, When the model balances the trade-off of decoding and generating capabilities of relevant signals, both decoding and generating performance are good, and the residuals contain a small amount of task information.

toward decoding (Fig. 1g), and a proper trade-off (Fig. 1h). If the distilling model is biased toward extracting signals similar to raw signals, distilled task-relevant signals will contain too much task-irrelevant information, which would affect the decoding

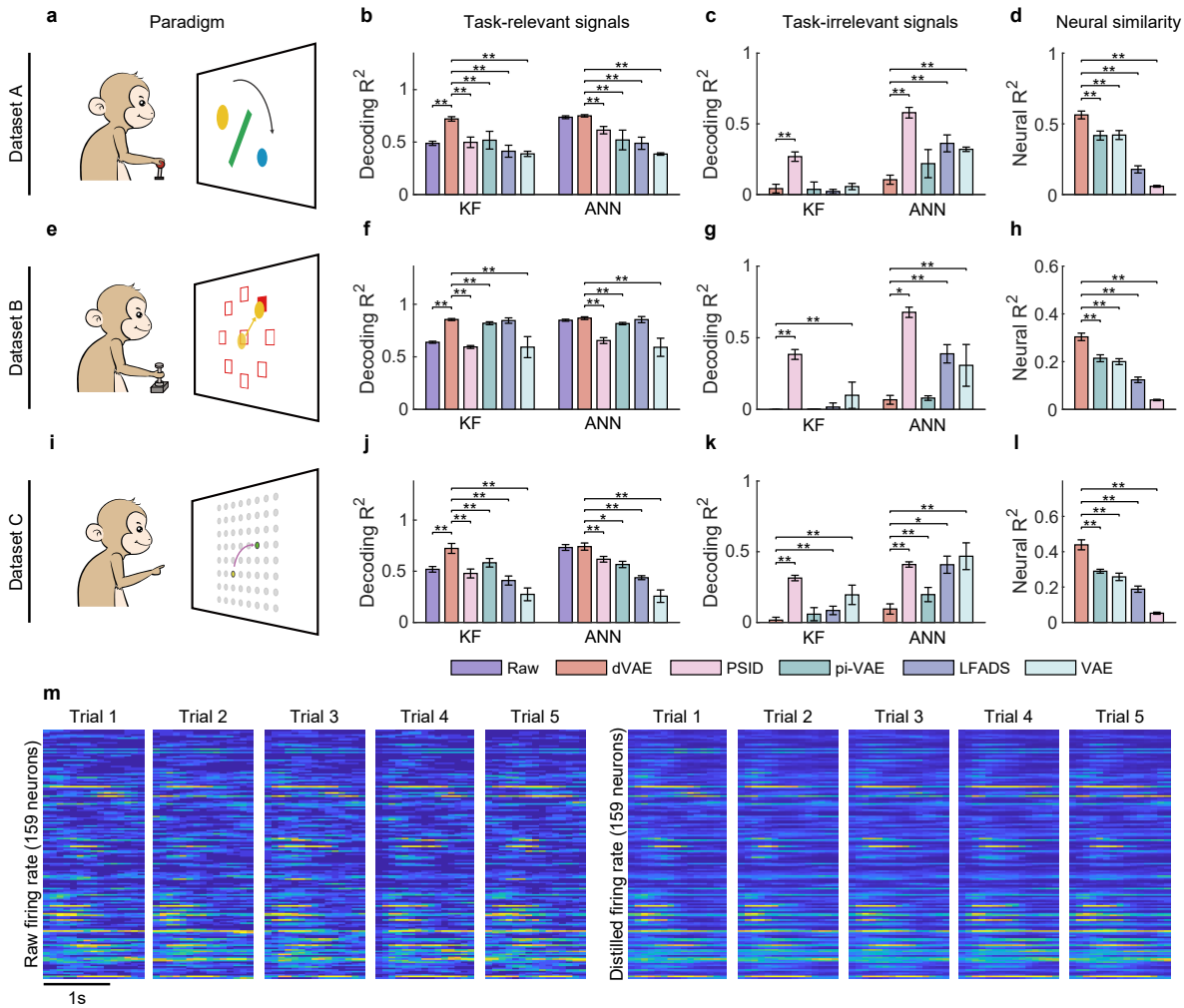

Figure 2: **Evaluation of task-relevant signals distillation.** The plots in the top three rows represent the results for three datasets recorded from the motor cortex with different paradigms, where the paradigms of datasets A (first row), B (second row), and C (third row) are obstacle avoidance (**a**), center-out (**e**), and self-paced reaching (**i**), respectively. **b**, The decoding R^2 between true velocity and predicted velocity of raw signals and task-relevant signals obtained by dVAE, PSID, pi-VAE, LFADS, and VAE 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 task-irrelevant signals obtained by five different methods. **d**, The neural similarity (R^2) between raw signals and task-relevant signals extracted by dVAE, PSID, pi-VAE, LFADS, and VAE. Error bars represent mean ± s.d. across five cross-validation folds. Asterisks indicate significance of Wilcoxon rank-sum test with $^{**}P < 0.01$. **f-h** and **j-l**, Same as **b-d**, but for datasets B and C. **m**, Example trials of firing rates of raw and distilled signals by dVAE on dataset B.

generalization of distilled 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 useful but redundant signals will be left in the residuals (task-irrelevant signals) so that the residuals contain a lot of task information (Fig. 1g). Neither of these two cases is what we desire because the former loses decoding performance, whereas the latter loses some useful neural signals, which are not conducive to our subsequent analysis of the relationship between task-relevant signals and tasks. The task-relevant signals we want should be similar to raw signals and preserve the task information maximally, which can be obtained by balancing the generating and decoding capabilities of generated task-relevant signals (Fig. 1h).

2.1.3. How to evaluate task-relevant signals. To evaluate whether the distilled signals are complete, we proposed three criteria. The first criterion is that the decoding ability of the task-relevant signals should be higher than that of raw signals. Because strong decoding models, such as deep neural networks, are more prone to overfit to noisy raw signals than task-relevant signals, distilled signals should achieve better decoding generalization than raw signals. The second criterion is that the task-irrelevant signals contain little task, which is very important and, from the opposite side, measures whether distilled signals preserve task information maximally. If the distilled task-relevant signals are incomplete or compensated for additional task information which does not exist in real task-relevant signals, the uncharacterized or over-characterized task information would be reflected in the residuals (task-irrelevant signals), and thus the residuals can be decoded a lot of task information. In this case, the distilled signals are not suitable for analysis. The third criterion is that the task-relevant signals should be similar to raw signals to maintain neuronal properties. If the distilled signals do not resemble raw signals, these signals do not retain the essential characteristics of raw signals, which are not qualified for subsequent analysis. Overall, if satisfying the above three criteria, we consider the distilled signals are complete.

2.2. dVAE distills complete task-relevant signals

To assess whether the distilled task-relevant signals are complete, we evaluated distilled signals' velocity decoding ability (determination of coefficients, R^2) using two decoders, a linear Kalman filter (KF) and a nonlinear artificial neural network (ANN), and evaluated the similarity (R^2) between distilled task-relevant and raw signals on three benchmark datasets with different task paradigms (Fig. 2a,e,i; see Methods for details). To demonstrate the effectiveness of distilled signals, we compared dVAE with representative approaches, including pi-VAE [17], PSID [12] and LFADS [18], and a vanilla VAE [19]. The details of model settings are presented in Methods.

Overall, dVAE extracts task-relevant signals that satisfy the three criteria abovementioned. Specifically, for the decoding performance of task-relevant signals (Fig. 2b,f,j), the signals distilled by dVAE achieve higher decoding R^2 than raw signals and signals extracted by competitors with both KF and ANN decoders. The

results demonstrate that dVAE effectively extracts task-relevant signals that preserve task information maximally. For the decoding performance of task-irrelevant signals (Fig. 2c,g,k), task-irrelevant signals filtered by dVAE hardly contain task information, whereas the irrelevant signals got by other models possess more task information. The results reflect the superiority of our method in extracting task-relevant signals from another angle and demonstrate that most existing approaches only distill part of task-relevant signals, leaking more task information in the residuals. For the neural similarity between relevant and raw signals (Fig. 2d,h,i), distilled signals by dVAE achieve the highest performance among competitors ($P < 0.01$, Wilcoxon rank-sum test). Besides, to visually compare the similarity, we showed the firing rate of neurons for five held-out test trials in the same condition of dataset B (Fig. 2m). We can see that the firing pattern of distilled signals is similar to the corresponding raw signals, thus indicating that task-relevant signals distilled by dVAE are similar to raw signals. Furthermore, we found that the pattern of distilled signals is more consistent among different trials than that of raw signals. The results suggest that eliminating the interference of irrelevant signals makes us easier to spot consistent patterns across trials.

In summary, dVAE distills effective task-relevant signals that preserve task information maximally and are similar to raw signals. Meanwhile, the task-irrelevant signals discarded by dVAE contain little task information. Therefore, these signals are reliable for exploring the relationship between task-relevant and irrelevant signals, as well as the encoding and decoding mechanisms of relevant signals.

2.3. How do task-irrelevant signals affect task-relevant signals at the single-neuron level?

Given that we partitioned raw signals into relevant and irrelevant signals, we explored how task-irrelevant signals affect task-relevant signals. Here we first investigated how irrelevant signals affect relevant signals at the single-neuron level. It is well known that the preferred direction (PD) is an essential property of neuronal activity. The first question we want to know is how irrelevant signals affect neurons' PD. Besides, a significant challenge in analyzing neuronal activity is that individual neurons exhibit considerable trial-to-trial variability. The second question we are concerned about is whether irrelevant signals mainly cause variability.

To explore how irrelevant signals affect neurons' PD, we quantified the PD deviation by the angle difference between raw and distilled signals. Results show that the PD of neurons with higher R^2 are mostly stable with signal distillation (see example PDs in the inset), while neurons with lower R^2s show larger PD deviation(Fig. 3a,e and Fig. S1a). The possible reason is smaller R^2s neurons have a lower degree of linear encoding but a higher degree of nonlinear encoding, and highly nonlinear encoding structures are more susceptible to interference from irrelevant signals. The results indicate that irrelevant signals have a small effect on strongly tuned neurons but a large effect on weakly tuned neurons. Besides, after filtering out the task-irrelevant signals, the cosine tuning fit

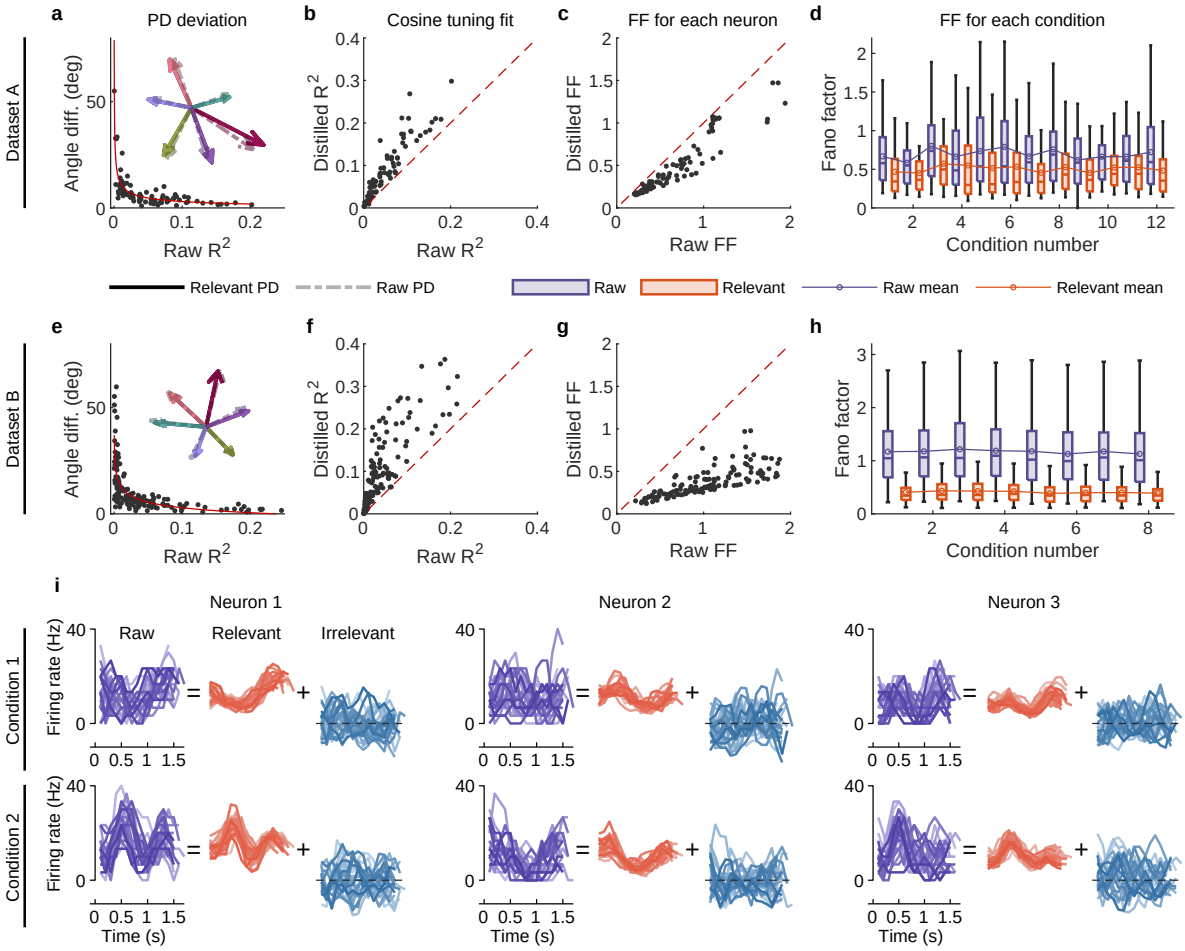

Figure 3: **The effect of irrelevant signals on relevant signals at the single-neuron level.** The plots in the first and second rows represent the results for datasets A and B, respectively. **a**, The change of angle difference (AD) of PD deviation between raw and distilled signals with R^2 of raw signals on datasets A and B. Each black point represents a neuron (n = 90). The red curve is the fitting curve between R^2 and AD. Five example neurons' PDs before and after distillation are shown in the inset plot, where the solid line arrows represent the PDs of relevant signals, and the dotted line arrows represent the PDs of raw signals. **b**, Comparison of the cosine tuning fit (R^2) before and after distillation of single neurons (black points), where the x-axis and y-axis represent neurons' R^2 of raw and distilled signals. **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 variability of different trials in the same condition. **d**, Boxplots of raw and distilled signals under different conditions for all neurons (12 and 8 conditions for datasets A and B). The broken lines represent the mean FF across all neurons, where purple and orange colors denote raw and distilled signals. **e-h**, Same as **a-d**, but for dataset B (n=159). **i**, Example of three neurons' raw firing activity decomposed into task-relevant and irrelevant parts using all trials under two conditions (2 of 8 directions) in held-out test sets on dataset B.

(R^2) of neurons increases ($P < 10^{-20}$, Wilcoxon signed-rank test; Fig. 3b,f and Fig. S1b), indicating that irrelevant signals reduce the neurons' tuning expression. Notably, although removing the irrelevant signals, the linear encoding model only explains a small fraction of neural responses, indicating that neuronal activity encodes task information in a complex nonlinear way.

To investigate whether irrelevant signals mainly cause variability, we compared the neuronal variability (measured with the Fano factor [20], 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 and median FFs of all neurons with different conditions are also significantly lower than those of raw signals ($P < 0.01$, Wilcoxon signed-rank test; Fig. 3d,h), indicating that the variability of single neurons is mainly due to irrelevant signals. We then visualized the single-trial neuronal activity of exemplar neurons under different conditions (Fig. 3i and Fig. S2). Results demonstrate that the variability of raw signals is mainly explained by irrelevant signals, and task-relevant signals also have some variability; the patterns of relevant signals are more consistent and stable among trials under different conditions than raw signals.

2.4. How do task-irrelevant signals affect task-relevant signals at the population level?

The population structure of neurons is also an essential characteristic of neural activity. Here, we explored the population properties and relationship of relevant and irrelevant signals and investigated how the two parts make up raw signals and the difference between raw and distilled signals at the population level.

To explore the population properties of relevant and irrelevant signals, we applied principal component analysis (PCA) separately on relevant and irrelevant partitions to get the corresponding cumulative variance curve in a descending variance order. As shown in Fig.4a,e and Fig. S3a, the primary subspace (capturing the top 90% variance) of relevant signals is only explained by a few dimensions (7, 13, and 9 for each dataset), indicating that task-relevant signals exist in a lower-dimensional subspace. In contrast, the primary subspace of irrelevant signals needs more dimensions (46, 81, and 59). The variance distribution of task-irrelevant signals across dimensions (thick blue line, Fig. 4b,f and Fig. S3b) is more even than task-relevant signals (thick orange line, Fig. 4a,e and Fig. S3a) but not as uniform as Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (thin gray line, Fig.4b,f and Fig. S3b), indicating that irrelevant signals are not pure noise but also bear some significant structure.

To study the subspace overlap between relevant and irrelevant signals, we calculated how many variances of irrelevant signals can be captured by projecting irrelevant signals onto relevant PCs and vice versa [21–23](see Methods). We found that the variance of irrelevant signals increases relatively uniformly over relevant PCs (blue thin line, Fig. 4a,e and Fig. S3a), like random noise's variance accumulation explained by relevant PCs (gray thin line, Fig. 4a,e and Fig. S3a); similar results are observed for relevant signals

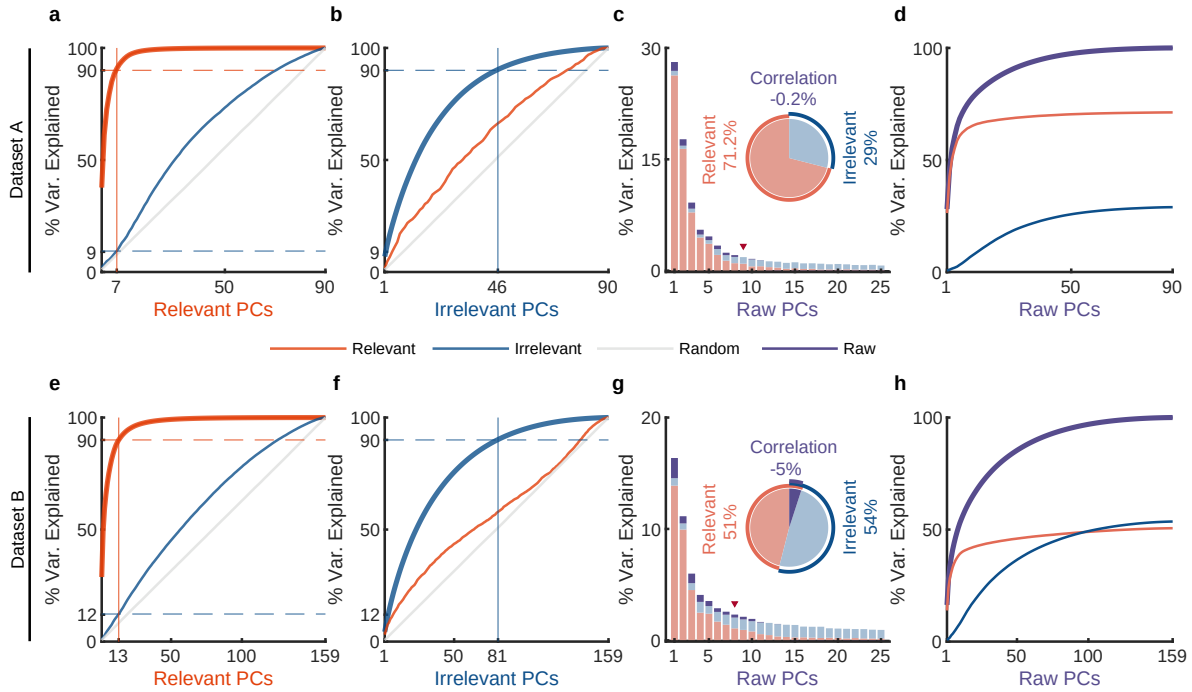

Figure 4: **The effect of irrelevant signals on relevant signals at the population level. a,b**, PCA is separately applied on relevant and irrelevant signals to get relevant PCs and irrelevant PCs. The thick lines represent the cumulative variance explained for the signals on which PCA has been performed, while the thin lines represent the variance explained by those PCs for other signals. Orange, blue, and gray colors indicate relevant signals, irrelevant signals, and random Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (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 task-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 task-relevant (**a**) and irrelevant (**b**) signals got by dVAE 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 orange, 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 (orange) and irrelevant (blue) signals. **e-h**, Same as **a-d**, but for dataset B.

explained by irrelevant PCs (orange thin line, Fig. 4b,f and Fig. S3b). These results indicate that relevant PCs can not match the informative dimensions of irrelevant signals and vice versa, thus the dimensions of task-relevant and irrelevant signals are mostly unrelated. It is worth mentioning that the signals obtained by pi-VAE are contrary to our findings. Its results show that a few relevant PCs can explain a considerable variance of irrelevant signals (thin orange line, Fig. S4b,f,j), which indicates the relevant PCs are close to irrelevant PCs. The possible reason is that the pi-VAE leaks a lot of information into the irrelevant signals. Notably, Fig. 4a,e and Fig. S3a show that irrelevant signals got by dVAE projecting on task-relevant primary subspace only capture a little variance information (9%, 12%, and 9%), indicating that the primary subspace of task-relevant signals is nearly orthogonal to irrelevant space.

To investigate how the two partitions make up raw signals, we performed PCA on raw neural signals to obtain raw PCs and then projected relevant and irrelevant signals onto raw PCs to assess how many variances of raw signals can be explained by relevant and irrelevant signals, respectively. Firstly, we analyzed the overall proportion of relevant and irrelevant signals that make up the raw signals (the inset pie plot, Fig. 3c,g and Fig. S3c). The variance of the raw signals consists of three parts: the variance of relevant signals, the variance of irrelevant signals, and the correlation between relevant and irrelevant signals (see Methods). Results demonstrate that the irrelevant signals account for a large proportion of raw signals, suggesting the motor cortex encodes considerable information unrelated to the measured behaviors. In addition, there is only a slight correlation between relevant and irrelevant signals, implying that the 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 occupy the dominant proportions in the larger variance PCs (before the PC marked with a red triangle), while irrelevant signals occupy the dominant proportions in the smaller variance PCs (after the PC marked with a red triangle) (Fig. 3c,g and Fig. S3c). The same results are obtained in the accumulation of each raw PC (Fig. 3d,h and Fig. S3d); results show that the variance accumulation of raw signals in larger variance PCs is mainly contributed by relevant signals, while irrelevant signals primarily contribute to the lower variance PCs. These results demonstrate that irrelevant signals have a small effect on larger variance PCs but a large effect on smaller variance PCs, clearing the doubts that irrelevant signals may also significantly affect the top few PCs of raw signals to draw inaccurate conclusions. To further test these findings, we used the top six PCs as jPCA [11] did to examine the rotational dynamics of distilled and raw signals (Fig. S5). Results show that the rotational dynamics of distilled signals are similar to raw signals.

Finally, to directly compare the population properties between raw and relevant signals, we showed the cumulative variance curve of raw and relevant signals (Fig. S7). Results (upper left corner curves, Fig. S7) show that the cumulative variance curve of relevant signals accumulates faster than that of raw signals in the preceeding larger variance PCs, indicating that the variance of the relevant signal is more concentrated in

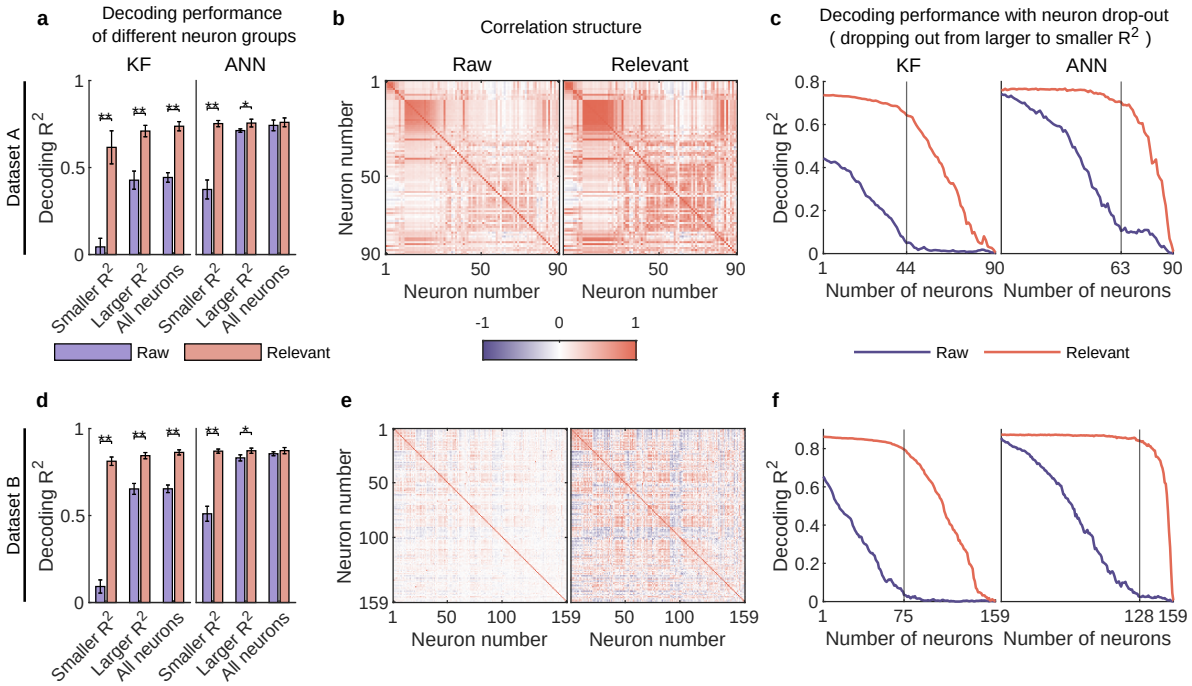

Figure 5: **Smaller R^2 neurons encode rich task information.** **a**, The comparison of decoding performance between raw (purple color) and distilled signals (orange color) on dataset A with different neuron groups, including smaller R^2 neuron ($R^2 <= 0.03$), larger R^2 neuron ($R^2 > 0.03$), and all neurons. Error bars indicate mean ± s.d. across five cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{\star}P < 0.05$, $^{\star\star}P < 0.01$. **b**, The correlation matrix of all neurons of raw and task-relevant 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 R^2 on dataset A. The vertical gray line indicates the number of dropped neurons at which raw and task-relevant signals have the greatest performance difference. **d-f**, Same as **a-c**, for dataset B.

the few 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 task-relevant signals (7, 13, and 9), indicating that task-irrelevant signals make us overestimate the neural dimensionality of task-relevant signals.

2.5. Distilled task-relevant signals uncover that smaller R^2 neurons encode rich task information in a complex nonlinear way.

The above results show that irrelevant signals significantly impact smaller R^2 neurons. Under the interference of irrelevant signals, it is difficult to explore the amount of task information in neuronal activity. After removing the influence of irrelevant signals, we asked whether these neurons contained sufficient task information.

To answer the question, we divided the neurons into two groups of smaller R^2 ($R^2 <= 0.03$) and larger R^2 ($R^2 > 0.03$) and compared the decoding performance of raw and distilled signals. As shown in Fig. 5a, for the smaller R^2 neuron group, both KF and ANN decode task information poorly on raw signals while they can achieve high decoding performance using relevant signals. Specifically, with the KF decoder, the decoding R^2 improves significantly from 0.044 to 0.616 (improves by about 1300%; $P < 0.01$, Wilcoxon rank-sum test) after signal distillation; and with the ANN decoder, the improvement is from 0.374 to 0.753 (improves by about 100%; $P < 0.01$, Wilcoxon rank-sum test). Similar results are obtained with datasets B (improves by 410% and 70% for KF and ANN; Fig. 5d) and C (improves by 780% and 70% for KF and ANN; Fig. S6d). These results indicate that smaller R^2 neurons indeed encode rich task information, which has not been found in the past only because task information is covered by irrelevant signals.

The above results about the smaller R^2 neurons seem unexpected. One might doubt that the larger R^2 neurons leak their information to the smaller R^2 neurons, making them contain too much task information. However, it is worth noting that the residuals contain almost no information (Fig. 2c,g,k); thus, this doubt does not hold. Another reasonable explanation is that the larger R^2 neurons help the smaller R^2 neurons to restore their original appearance. To verify this case, we simulated a situation where the smaller R^2 neurons contain a certain amount of task information, but the task information cannot be decoded from these neurons due to being covered by noise to see if dVAE can restore these neurons' original appearance. Results (Fig. S9a-c) show that for smaller R^2 neurons, the decoding R^2 of distilled signals (0.362, 0.401 for KF and ANN) is significantly higher than that of raw signals (0.019, 0.063 for KF and ANN; $P < 0.01$, Wilcoxon rank-sum test), and is more closer to that of ground truth (0.454, 0.467 for KF and ANN), demonstrating that dVAE indeed utilizes the larger R^2 neurons to help the smaller R^2 neurons restore their original appearance. These simulation results further confirm that the smaller R^2 neurons contain rich task information, which is just covered by irrelevant signals.

Furthermore, it is worth noting that the sum of the decoding performance of smaller R^2 neurons and larger R^2 neurons is significantly greater than that of all neurons for relevant signals (orange bar, Fig. 5a,d and Fig. S6a), indicating movement parameters are encoded redundantly in neuronal population. In addition, we can see that distilled signals have more redundant information than raw signals (purple bar, Fig. 5a,d and Fig. S6a), indicating that irrelevant signals cover the task information. Generally, the cortical redundancy can arise from neuronal correlation. Thus we visualized the neurons' ordered correlation matrix (see Methods) of raw and relevant signals (Fig. 5b,e and Fig. S6b). We can find that the neuronal correlation of relevant signals is stronger than that of raw signals, demonstrating that irrelevant signals weaken the neuronal correlation, affecting the accurate estimation of the neuronal correlation.

Given that neuronal activities have rich redundancy and strong correlation, we asked whether decoders can utilize redundant information from other neurons to cope

with strongly tuned neurons drop-out for robust decoding. We investigated this question by evaluating the decoding performance of dropping out neurons from larger R^2 to smaller R^2 on raw and relevant signals (Fig. 5c,f and Fig. S6c). Results show that the decoding performance of the KF and ANN on raw signals decreases steadily and progressively before the number of neurons marked, and the remaining smaller R^2 neurons decode task information poorly. In contrast, even if many neurons are lost, the decoding performance of KF and ANN on relevant signals is stable and maintains high accuracy. These findings demonstrate that decoders can effectively deal with the problem of neuron drop-out on task-relevant signals, and task-irrelevant signals significantly affect the decoders to utilize other neurons' information for robust decoding. Notably, ANN performs much better than KF on smaller R^2 neurons (Fig. 5a,d and Fig. S6a) and is more robust than KF in the case of neuron drop-out especially when only smaller R^2 neurons are left (Fig. 5c,f and Fig. S6c), indicating that smaller R^2 neurons encode task information in a strongly nonlinear way.

2.6. Distilled task-relevant signals uncover that signals composed of smaller variance PCs encode rich task information in a complex nonlinear way.

The above results show that smaller variance PCs of raw signals are dominated by irrelevant signals. Typically, irrelevant signals can drown out useful information and make it appear useless. Here, we want to know whether the signals composed of smaller variance PCs (after the PC marked with a red triangle in Fig. 4c,g and Fig. S3c) contain sufficient task information.

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, according to the ratio of relevant to irrelevant signals in the raw-PC (see Fig. 4c,g and Fig. S3c). 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 Fig. S6a). These results demonstrate that the signals composed of smaller variance PCs encode much task information, and it has not been found before only because irrelevant signals cover their information. In contrast, irrelevant signals have relatively a small effect on larger variance PCs.

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. Since we got the task-relevant signals, we want to explore how task information of distilled signals is distributed on relevant PCs. To address the question, we used decoding models to assess how much task information is contained in cumulative PCs of relevant

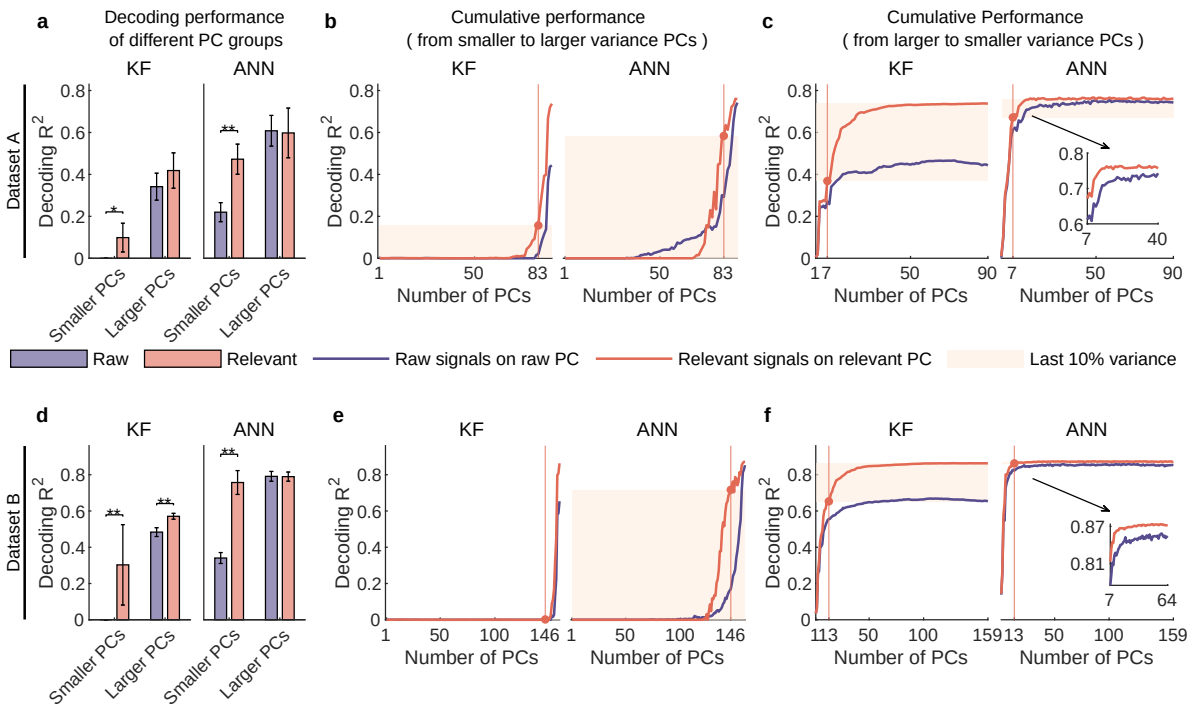

Figure 6: **Signals composed of small variance PCs encode rich task information in a complex nonlinear way. a**, The comparison of decoding performance between raw (purple) and distilled signals (orange) composed of different raw-PC groups, including smaller variance PCs (the proportion of irrelevant 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 $^{\star}P < 0.05$, $^{\star\star}P < 0.01$. **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 orange 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 orange 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.

signals (using raw signals for 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 Fig. S7.

Here we first investigated secondary signals' decoding ability solely by accumulating PCs from smaller to larger variance. Results show that, for relevant signals, KF hardly decodes task information solely using secondary signals (orange line, Fig. 6b,e and Fig.

S6e left), but ANN can decode rich information (orange line, Fig. 6b,e and Fig. S6e right). These results demonstrate that the relevant signals composed of smaller variance PCs encode rich task information in a complex nonlinear way. In contrast, this finding can not be discovered from raw signals (purple line, Fig. 6b,e and Fig. S6e) because irrelevant signals cover the truth.

Then we accumulated PCs from larger to lower variance to explore the effect of superimposing secondary signals on primary signals. Results (Fig. 6c,f and Fig. S6f) show that both KF and ANN's decoding performance of relevant signals (orange line) is always higher than that of raw signals (purple lines) during accumulating PCs. In particular, around the top 5 to 20 PCs, the decoding performance of relevant signals increases significantly over that of raw ones, suggesting these PCs are critical for improving decoding performance. Besides, it is interesting to find that although secondary signals nonlinearly encode task information and are decoded by linear decoders poorly, they can improve KF performance considerably by superimposing on primary signals (Fig. 6c,f and Fig. S6f left), and the sum of the sole decoding performance of primary and secondary primary signals is lower than the decoding performance of full signals, indicating that the combination of smaller and larger variance PCs will produce a synergy effect so that signals which cannot be linearly decoded can be linearly decoded, implying that smaller variance PC signals encode task information in a subtly nonlinear way.

Finally, we explored which aspect of movement parameters was improved by superimposing the secondary signals on the primary ones with KF. In BMIs, directional control has achieved great success [7, 9, 10], but precise speed control has always been a challenging problem [8, 24], especially for lower-speed control, such as hold or stop. Thus we speculated that these signals might improve the lower-speed velocity. To verify this conjecture, we divided samples into lower-speed and higher-speed regions and saw which part improved the most by superimposing the secondary signals (see details in Methods). After superimposing the secondary signals, the absolute improvement ratio of the low-speed part is significantly higher than that of the high-speed part ($P < 0.05$, Wilcoxon rank-sum test; Fig.S8a,b,c). Furthermore, we visualized five example trials of relative improvement ratio for the two parts. Results (Fig. S8d) also show that secondary signals significantly improve the estimation of lower speed. These results indicate that the secondary signals can improve the lower-speed control and inspire us to focus on the smaller variance PC signals.

2.7. Distilled task-relevant signals reveal that movement behaviors are generated from neural population in motor cortex by linear readout.

The readout mechanism of the motor cortex remains unclear yet, which is an essential problem for neuroscience and neural engineering. By filtering out the interference of task-irrelevant signals, we found a stunning result that the linear decoder KF achieves comparable performance to 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 the possibility of linear readout in the motor cortex. However, one might suspect that our distillation model makes nonlinearly decodable signals linearly decodable. To clear the doubt, we conducted a simulation experiment where linear decoding is significantly inferior to nonlinear decoding to see if signals after processing by our model can make linear decoders achieve similar performance as nonlinear decoders. Results (Fig. S9d-f) show that the difference of decoding R^2 between KF and ANN on distilled signals is 0.146, which is comparable to the difference (0.162) on raw signals ($P = 0.15$, Wilcoxon rank-sum test), demonstrating that our model can not make linear decoders achieve similar performance as nonlinear decoders, thus clearing the doubt, further suggesting that task information is linearly decoded from neural responses in the motor cortex. In summary, these results demonstrate that the linear readout is sufficient to achieve high decoding performance, providing strong evidence for the linear readout of neural responses to generate movement behavior and demonstrating the unexpected simplicity of decoding mechanisms.

3. Discussion

In this study, we proposed a new perspective — distilling complete task-relevant signals — to explore the relationship between neural activity and task information, and offered a distilling model which can effectively distill task-relevant neural signals. By disassociating task-relevant and irrelevant signals, we found that trivial neural responses encode rich task information in a complex nonlinear way. Furthermore, we found that linear decoders can achieve comparable performance to nonlinear decoders, providing strong evidence for the linear readout of neural population in motor cortex. Taken together, these results reveal unexpected complex encoding but simple decoding mechanisms in the motor cortex.

3.1. Pursuing complete task-relevant neural signals by dVAE

To reduce irrelevant signals' interference, researchers extract obvious strong neural patterns, including the signal level (single-neuron level) and latent neural mode level (neural population level), for analysis. Here our model targets the signal level, although we use the latent variable model to achieve this. Extracting task-relevant information at the signal level can preserve the properties of single neurons and is helpful for electrophysiological [25] or optogenetic [26] interventions in closed-loop stimulation experiments.

For the signal level, researchers usually get the trial-averaged neural responses of the same task parameters for analysis [22,27]; however, trial-averaged responses sacrifice single-trial information, thus providing an incomplete characterization of neural activity.

In contrast, our model can preserve single-trial information. For the latent neural mode level, existing latent variable models [12, 17, 18, 28] focus on modeling the properties of latent variables, such as linear or nonlinear dynamics [11, 12, 18], temporal smoothness [28], and interpretability [17]; however, these restrictive assumptions inevitably lead to the loss of task information, yielding no guarantee that the generated signals preserve task information maximally. In contrast, our model aims to extract complete task information at the signal level. To make sure of achieving this, we explicitly constrained the model to generate signals that maintain the neuronal properties while preserving the task information maximally; besides, we deliberately avoided imposing additional constraints on the model, thus allowing the model to generate neural signals faithfully.

Although we have taken corresponding measures to achieve the goal, the model still cannot perfectly extract task-relevant neural signals, leaking a small amount of task information in the residuals. In the future, better methods can be proposed to extract task-relevant signals. Nevertheless, these distilled signals are reliable, and the minor inadequacies do not affect the conclusions.

3.2. The benefits of disassociating relevant and irrelevant signals

Studying neural mechanisms through noisy signals is like looking at flowers in a fog, which is difficult to know the truth. Thus, removing irrelevant signals' interference is necessary and beneficial for studying neural mechanisms no matter at the single neuron level or population level.

At the single neuron level, trial-to-trial neuronal variability is a significant obstacle to studying the actual neuronal pattern changes. The interpretation of variability ranges from meaningless noise [6] to meaningful but task-irrelevant neural processes [29, 30] to intrinsic components of neural encoding [31,32]. To what extent each source contributes to the variability remains unclear [6]. Previous studies [20, 29–32] used raw signals to study neural variability, thus it was difficult to distinguish the contribution of different sources of variability. Contrary to these studies, we separated relevant and irrelevant signals to study variability. Therefore we could determine to what extent these two parts contribute to the variability and explore which type of variability these two parts may contain. Our results demonstrated that task-irrelevant signals mainly contribute to variability, which may include both meaningless noise and meaningful but task-irrelevant signals because task-irrelevant signals are not pure noise and carry some structures (thick blue line, Fig. 4b,f and Fig. S3b); task-relevant signals also bear some variability, which may come from intrinsic components of neural encoding, providing the neural basis for motor learning [33]. In particular, eliminating the variability caused by irrelevant signals allows us to better observe and compare actual neuronal pattern changes and hopefully facilitates the study of learning mechanisms [34–37].

At the population level, it is challenging to determine the neural dimensionality of specific tasks from raw signals because we don't know how many variances stand for irrelevant signals, which usually severely depend on signal quality. A previous

study [38], through simulation experiments of adding different level noises, demonstrated that noises make methods overestimate the neural dimension of the task. Consistent with their results, we found that task-irrelevant signals make us overestimate the neural dimensionality of task-relevant signals (Fig. S7). These results recommend that it is necessary to filter out irrelevant signals to estimate the neural dimensionality accurately. Moreover, this perspective of signal disassociation also has implications for other studies. For example, researchers can distill neural signals corresponding to different tasks and then explore their shared and exclusive patterns, thus uncovering underlying common and unique mechanisms of different behaviors [13].

3.3. The complex encoding but simple decoding mechanisms in motor cortex

At the single-neuron level, our results reveal that smaller R^2 neurons previously considered useless encode unexpected rich task information in a complex nonlinear way, which demonstrates the complexity of neuronal encoding, also implying that it is hard to find a simple and unified formula to characterize the complex and heterogeneous properties of single neurons. Furthermore, previous literature [39, 40] showed the redundant encoding of neuronal ensembles in the motor cortex. Our results add support to the redundant encoding and found more redundant task information in distilled signals than in raw signals. Specifically, their results show 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. These findings support 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 [41].

At the population level, previous studies suggest that motor control occupies low-dimensional neural manifolds [15, 42]. However, our results reveal that signals composed of smaller variance PCs nonlinearly encode large amounts of task information, thus indicating neural activity exists in a higher dimensional space than expected. Interestingly, although smaller variance PC signals nonlinearly encode task information, the task information of these signals can also be linearly decoded by superimposing these signals to larger variance PC signals. This finding is consistent with previous findings that nonlinear mixed selectivity yields high dimensional neural responses, thus allowing linear readout of task information by downstream neurons [3, 4]. Besides, we found that smaller variance PCs of neural signals can improve precise motor control, such as lower speed control. In analogous, recent studies found smaller variance PCs of hand postures are task-dependent and relate to the precise and complex postures [43]. These results suggest that lower variance PC signals are related to precise movement control and motivate us to focus on these neglected but valuable smaller variance PCs.

In other cortices, such as inferotemporal cortex (IT) [44], and perirhinal cortex (PRH) [45], studies suggested that the downstream neurons may linearly read out task

information. In the motor cortex, in what form neurons decode task information remains unclear. By filtering out task-irrelevant signals, we found a stunning result that the linear decoder KF achieves comparable performance to the nonlinear decoder ANN, suggesting the linear readout of neural responses in motor cortex to generate movement behaviors. Notably, linear decoding needs a certain number of smaller R^2 neurons or smaller variance PCs that nonlinearly encode task information, suggesting that the complexity of encoding mechanisms underlies the simplicity of decoding mechanisms. Furthermore, these results show that the most critical factor affecting the decoding performance is the interference of irrelevant signals. These findings provide a new avenue for developing robust and accurate BMIs: develop preprocessing distilling algorithms or develop decoding algorithms that contain the function of filtering out task-irrelevant signals.

4. Reference

[1] Shreya Saxena and John P Cunningham. Towards the neural population doctrine. *Current opinion in neurobiology*, 55:103–111, 2019.

[2] Nikolaus Kriegeskorte and Pamela K Douglas. Interpreting encoding and decoding models. *Current opinion in neurobiology*, 55:167–179, 2019.

[3] Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, 37:66–74, 2016.

[4] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[5] Rony Azouz and Charles M Gray. Cellular mechanisms contributing to response variability of cortical neurons in vivo. *Journal of Neuroscience*, 19(6):2209–2223, 1999.

[6] A Aldo Faisal, Luc PJ Selen, and Daniel M Wolpert. Noise in the nervous system. *Nature reviews neuroscience*, 9(4):292–303, 2008.

[7] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

[8] Yoh Inoue, Hongwei Mao, Steven B Suway, Josue Orellana, and Andrew B Schwartz. Decoding arm speed during reaching. *Nature communications*, 9(1):1–14, 2018.

[9] Meel Velliste, Sagi Perel, M Chance Spalding, Andrew S Whitford, and Andrew B Schwartz. Cortical control of a prosthetic arm for self-feeding. *Nature*, 453(7198):1098–1101, 2008.

[10] Leigh R Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y Masse, John D Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S Cash, Patrick Van Der Smagt, et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–375, 2012.

[11] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

[12] Omid G Sani, Hamidreza Abbaspourazad, Yan T Wong, Bijan Pesaran, and Maryam M Shanechi. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience*, 24(1):140–149, 2021.

[13] Juan A Gallego, Matthew G Perich, Stephanie N Naufel, Christian Ethier, Sara A Solla, and Lee E Miller. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature communications*, 9(1):1–13, 2018.

[14] Juan A Gallego, Matthew G Perich, Raeed H Chowdhury, Sara A Solla, and Lee E Miller.

Long-term stability of cortical population dynamics underlying consistent behavior. *Nature neuroscience*, 23(2):260–270, 2020.

[15] John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[16] Jonathan C Kao, Paul Nuyujukian, Stephen I Ryu, Mark M Churchland, John P Cunningham, and Krishna V Shenoy. Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. *Nature communications*, 6(1):1–12, 2015.

[17] Ding Zhou and Xue-Xin Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. *Advances in Neural Information Processing Systems*, 33:7234–7247, 2020.

[18] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10):805–815, 2018.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] Mark M Churchland, Byron M Yu, John P Cunningham, Leo P Sugrue, Marlene R Cohen, Greg S Corrado, William T Newsome, Andrew M Clark, Paymon Hosseini, Benjamin B Scott, et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3):369–378, 2010.

[21] Gamaleldin F Elsayed, Antonio H Lara, Matthew T Kaufman, Mark M Churchland, and John P Cunningham. Reorganization between preparatory and movement population responses in motor cortex. *Nature communications*, 7(1):1–15, 2016.

[22] Adam G Rouse and Marc H Schieber. Condition-dependent neural dimensions progressively shift during reach to grasp. *Cell reports*, 25(11):3158–3168, 2018.

[23] Xiyuan Jiang, Hemant Saggar, Stephen I Ryu, Krishna V Shenoy, and Jonathan C Kao. Structure in neural activity during observed and executed movements is shared at the neural population level, not in single neurons. *Cell reports*, 32(6):108006, 2020.

[24] B Wodlinger, JE Downey, EC Tyler-Kabara, AB Schwartz, ML Boninger, and JL Collinger. Ten-dimensional anthropomorphic arm control in a human brain- machine interface: difficulties, solutions, and limitations. *Journal of neural engineering*, 12(1):016011, 2014.

[25] Guanghao Sun, Fei Zeng, Michael McCartin, Qiaosheng Zhang, Helen Xu, Yaling Liu, Zhe Sage Chen, and Jing Wang. Closed-loop stimulation using a multiregion brain-machine interface has analgesic effects in rodents. *Science Translational Medicine*, 14(651):eabm5868, 2022.

[26] Jinyong Zhang, Ryan N Hughes, Namsoo Kim, Isabella P Fallon, Konstantin Bakhurin, Jiwon Kim, Francesco Paolo Ulloa Severino, and Henry H Yin. A one-photon endoscope for simultaneous patterned optogenetic stimulation and calcium imaging in freely behaving mice. *Nature Biomedical Engineering*, pages 1–12, 2022.

[27] Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E Feierstein, Adam Kepecs, Zachary F Mainen, Xue-Lian Qi, Ranulfo Romo, Naoshige Uchida, and Christian K Machens. Demixed principal component analysis of neural population data. *Elife*, 5:e10989, 2016.

[28] Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in neural information processing systems*, 21, 2008.

[29] Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22(10):1677–1686, 2019.

[30] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019.

[31] Edgar Y Walker, R James Cotton, Wei Ji Ma, and Andreas S Tolias. A neural basis of probabilistic

computation in visual cortex. *Nature Neuroscience*, 23(1):122–129, 2020.

[32] Martina Valente, Giuseppe Pica, Giulio Bondanelli, Monica Moroni, Caroline A Runyan, Ari S Morcos, Christopher D Harvey, and Stefano Panzeri. Correlations enhance the behavioral readout of neural population activity in association cortex. *Nature neuroscience*, 24(7):975–986, 2021.

[33] Ashesh K Dhawale, Maurice A Smith, and Bence P Ölveczky. The role of variability in motor learning. *Annual review of neuroscience*, 40:479, 2017.

[34] Patrick T Sadtler, Kristin M Quick, Matthew D Golub, Steven M Chase, Stephen I Ryu, Elizabeth C Tyler-Kabara, Byron M Yu, and Aaron P Batista. Neural constraints on learning. *Nature*, 512(7515):423–426, 2014.

[35] Emily R Oby, Matthew D Golub, Jay A Hennig, Alan D Degenhart, Elizabeth C Tyler-Kabara, Byron M Yu, Steven M Chase, and Aaron P Batista. New neural activity patterns emerge with long-term learning. *Proceedings of the National Academy of Sciences*, 116(30):15210–15215, 2019.

[36] Jay A Hennig, Emily R Oby, Matthew D Golub, Lindsay A Bahureksa, Patrick T Sadtler, Kristin M Quick, Stephen I Ryu, Elizabeth C Tyler-Kabara, Aaron P Batista, Steven M Chase, et al. Learning is shaped by abrupt changes in neural engagement. *Nature Neuroscience*, 24(5):727–736, 2021.

[37] Matthew D Golub, Patrick T Sadtler, Emily R Oby, Kristin M Quick, Stephen I Ryu, Elizabeth C Tyler-Kabara, Aaron P Batista, Steven M Chase, and Byron M Yu. Learning by neural reassociation. *Nature neuroscience*, 21(4):607–616, 2018.

[38] Ege Altan, Sara A Solla, Lee E Miller, and Eric J Perreault. Estimating the dimensionality of the manifold underlying multi-electrode neural recordings. *PLoS computational biology*, 17(11):e1008591, 2021.

[39] Jose M Carmena, Mikhail A Lebedev, Craig S Henriquez, and Miguel AL Nicolelis. Stable ensemble performance with single-neuron variability during reaching movements in primates. *Journal of Neuroscience*, 25(46):10712–10716, 2005.

[40] Nandakumar S Narayanan, Eyal Y Kimchi, and Mark Laubach. Redundancy and synergy of neuronal ensembles in motor cortex. *Journal of Neuroscience*, 25(17):4207–4216, 2005.

[41] Jeffrey Alstott, Michael Breakspear, Patric Hagmann, Leila Cammoun, and Olaf Sporns. Modeling the impact of lesions in the human brain. *PLoS computational biology*, 5(6):e1000408, 2009.

[42] Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

[43] Yuke Yan, James M Goodman, Dalton D Moore, Sara A Solla, and Sliman J Bensmaia. Unexpected complexity of everyday manual behaviors. *Nature communications*, 11(1):1–8, 2020.

[44] Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015.

[45] Marino Pagan, Luke S Urban, Margot P Wohl, and Nicole C Rust. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nature neuroscience*, 16(8):1132–1139, 2013.

[46] Fang Wang, Yiwen Wang, Kai Xu, Hongbao Li, Yuxi Liao, Qiaosheng Zhang, Shaomin Zhang, Xiaoxiang Zheng, and José C Principe. Quantized attention-gated kernel reinforcement learning for brain–machine interface decoding. *IEEE transactions on neural networks and learning systems*, 28(4):873–886, 2015.

[47] Eva L Dyer, Mohammad Gheshlaghi Azar, Matthew G Perich, Hugo L Fernandes, Stephanie Naufel, Lee E Miller, and Konrad P Körding. A cryptography-based approach for movement decoding. *Nature biomedical engineering*, 1(12):967–976, 2017.

[48] Joseph E O'Doherty, MMB Cardoso, JG Makin, and PN Sabes. Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology. *Zenodo http://doi. org/10.5281/zenodo*, 583331, 2017.

5. Methods

5.1. Dataset and preprocessing

Three datasets with different paradigms are employed, including obstacle avoidance task dataset [46], center-out reaching task dataset [47], and self-paced reaching task dataset. [48].

The first dataset 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 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 is publicly available and provided by Kording Lab [47]. 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 is publicly available and provided by Sabes Lab (Zenodo dataset) [48]. 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 dataset A and B, respectively; trials with a length less than 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).

5.2. Simulation experiments

In this experiment, we simulated four neurons, including two larger R^2 neurons (greater than 0.4) and two smaller R^2 neurons (lower than 0.03). The larger R^2 neurons are strongly tuned neurons and are added with small Gaussian noise. In contrast, the smaller R^2 neurons strongly nonlinearly encode task information and are added with large Gaussian noise. We regarded the signals without adding Gaussian noise as ground-truth signals and those with Gaussian signals as raw signals. We used the velocity of dataset B (20161011) as the velocity for simulation. The simulation dataset A includes the four neurons, which simulates the situation where the smaller R^2 neurons contain a lot of task information, but the task information cannot be decoded from these neurons due to being covered by noise. The simulation dataset B includes the two larger R^2 neurons, which simulates the situation where linear decoding is significantly inferior to nonlinear decoding.

The encoding model of larger R^2 neurons is as follows:

$$x_i(t) = b + m \mid \mathbf{v}(t) \mid \cos(\theta(t) - \theta_{PDi}) + b_s \mid \mathbf{v}(t) \mid + \epsilon_i(t), \tag{1}$$

where the subscript i refers to the ith unit of N neurons, $N = 2$ in this experiment, $b = 30$ Hz represents the baseline firing rate of neurons, $m = 0.25$ Hz per cm s^{-1} indicates the modulation depth, the PDs of neurons θ_{PDi} (i=1, 2) are set to 30° and 60°, $\epsilon_i(t) = \mathcal{N}(0, 0.3)$ denotes random Gaussian noise where the mean is zero and the standard deviation is 0.3, $\mathbf{v}$ represents velocity, $\mid \mathbf{v} \mid$ represents the norm (speed) of velocity, $b_s = 0.25$ Hz per cm s^{-1} denotes speed bias.

The encoding model of smaller R^2 neurons is as follows:

$$x_i(t) = b + m \mid \mathbf{v}(t) \mid \cos(\theta(t) - \theta_{PDi}) + b_s \mid \mathbf{v}(t) \mid + b_{ss} \mid \mathbf{v}(t) \mid^2 + \epsilon_i(t), \tag{2}$$

where $b = 30$ Hz, $m = 0.25$ Hz per cm s^{-1}, the PDs of neurons θ_{PDi} (i=1, 2) are set to 40° and 50°, $\epsilon_i(t) = \mathcal{N}(0, 1)$ denotes random Gaussian noise where the mean is zero and the standard deviation is one, $b_s = 0.25$ Hz per cm s^{-1}, $b_{ss} = 0.1$ Hz per cm^2 s^{-1} denotes the bias of speed square for increasing the encoding nonlinearty of neurons.

5.3. Variational autoencoders

Variational autoencoders (VAEs) [26] are famous latent variable generative models and assume latent representations $\mathbf{z} \sim p(\mathbf{z})$ can generate data $\mathbf{x}$ by conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{z})$, that is, $p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{z}) p_\theta(\mathbf{x} \mid \mathbf{z}) \mathrm{d}\mathbf{z}$. The usual training objective is maximum marginal likelihood $\log p_\theta(\mathbf{x})$. Since directly computing the integration is intractable, VAE propose an amortized inference distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$ and jointly optimize the evidence lower bound (ELBO) to the log-likelihood:

$$\mathcal{L}_{ELBO}(x) = \mathbb{E}_{z \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[p_\theta(\mathbf{x} \mid \mathbf{z}) \right] - D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})) \leq \log p_\theta(\mathbf{x}), \tag{3}$$

where the first term of right hand side of Eq. 3 is called reconstruction term, the second term is called regularization term, $q_\phi(\mathbf{z} \mid \mathbf{x})$ is usually parameterized with a neural

network called encoder or inference model, and $p_\theta(\mathbf{x} \mid \mathbf{z})$ is also parameterized with a neural network called decoder or generative model. $p(\mathbf{z})$ is the prior over the latent representations and is Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in vallina VAE, D_{KL} denotes the Kullback-Leibler divergence.

5.4. Distill-VAE (dVAE)

Notations. $\mathbf{x} \in \mathbb{R}^n$ denotes raw neural signals. $\mathbf{x}_r \in \mathbb{R}^n$ represents task-relevant signals. $\mathbf{x}_i = \mathbf{x} - \mathbf{x}_r \in \mathbb{R}^n$ represents task-irrelevant signals. $\mathbf{z} \in \mathbb{R}^d$ denotes the latent neural representations. $\mathbf{y} \in \mathbb{R}^k$ represents kinematics. $f : \mathbb{R}^n \to \mathbb{R}^d$ represents the inference model (encoder) of VAE. $g : \mathbb{R}^d \to \mathbb{R}^n$ represents the generative model (decoder) of VAE.

Our approach utilizes the trade-off between the decoding and generation ability of $\mathbf{x}_r$ to distill task-relevant neural signals. The basic assumption is generating task-relevant signals that contain task-irrelevant signals harms decoding ability. To distill task-relevant signals, we have three steps, including identifying latent representations $\mathbf{z}$, generating task-relevant signals $\mathbf{x}_r$, and decoding task-relevant signals $\mathbf{x}_r$.

5.4.1. Identfiying latent representations $\mathbf{z}$ Identifying latent representations containing task-relevant information is the crucial part because latent representations influences the subsequent generation. Good representations are more likely to generate good task-relevant neural signals $\mathbf{x}_r$. dVAE identifies latent representations with inference model f, that is, $\mathbf{z} = f(\mathbf{x})$, and guides latent representations containing task-information through an affine map $h : \mathbb{R}^d \to \mathbb{R}^k$ under the loss $\mathcal{L}_{dec1}$,

$$\mathcal{L}_{dec1} = \mathrm{MSE}(h(\mathbf{z}), \mathbf{y}), \tag{4}$$

where MSE denotes mean squared loss. In other words, we constrain latent representations to decode kinematics to distill task-relevant information.

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

$$\mathcal{L}_{rec} = \mathrm{Poisson}(\mathbf{x}_r, \mathbf{x}), \tag{5}$$

where Poisson denotes Poisson negative log likelihood loss. It is worth noting that making generated task-relevant signals to reconstruct noisy raw signals as closely as possible may cause generated task-relevant signals to contain more task-irrelevant neural signals, which deviates from our original intention of extracting task-relevant signals. In the following subsection, we will introduce how to avoid generating task-irrelevant signals.

5.4.3. Decoding task-relevant signals $\mathbf{x}_r$ As mentioned above, if generated task-relevant signals $\mathbf{x}_r$ are only constrained by $\mathcal{L}_{rec}$, generated task-relevant signals may contain more task-irrelevant signals. To avoid $\mathbf{x}_r$ generating task-irrelevant signals, we introduce decoding loss $\mathcal{L}_{dec2}$ to constrain $\mathbf{x}_r$ to decode task information. The basic assumption is that task-irrelevant signals act like noise for decoding task information and are detrimental to decoding. Thus, there is a trade-off between generating and decoding ability of $\mathbf{x}_r$: the more task-irrelevant signals $\mathbf{x}_r$ contains, the more decoding performance $\mathbf{x}_r$ loses. The decoding loss $\mathcal{L}_{dec2}$ is as follows:

$$\mathcal{L}_{dec2} = \mathrm{MSE}(h(f(\mathbf{x}_r)), \mathbf{y}). \tag{6}$$

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:

$$\mathcal{L}_{dec} = \frac{1}{2}(\mathcal{L}_{dec1} + \mathcal{L}_{dec2}). \tag{7}$$

5.4.4. Learning the prior distribution with task information The prior distribution of latent representation is crucial because inappropriate prior assumptions can affect 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 be more complex than Gaussian. Therefore we adopt neural networks $m : \mathbb{R}^k \to \mathbb{R}^d$ to learn the prior representation $\mathbf{z}_{prior}$ with kinematics $\mathbf{y}$, that is, $\mathbf{z}_{prior} = m(\mathbf{y})$. The prior representation $\mathbf{z}_{prior}$ is aligned to $\mathbf{z}$ guided by the Kullback–Leibler divergence:

$$\mathcal{L}_{KL} = D_{KL}(\mathbf{z} \| \mathbf{z}_{prior}).$$

Since $\mathbf{z}_{prior}$ is aligned to $\mathbf{z}$ 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.

5.4.5. End-to-end optimization dVAE 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 that determine the trade-off between reconstruction and distilling task-relevant signals.

In the training stage, we feed raw neural signals into the inference network f to get latent representation $\mathbf{z}$, then use $\mathbf{z}$ to decode kinematics $\mathbf{y}$ by linear readout layer h and send $\mathbf{z}$ to the generative networks g to generate neural signals $\mathbf{x}_r$. To ensure $\mathbf{x}_r$ contains decoding ability, we send $\mathbf{x}_r$ to f and h to decode $\mathbf{y}$. In addition, kinematics $\mathbf{y}$ pass the network m to get $\mathbf{z}_{prior}$. The whole model is trained in an end-to-end manner under the guidance of total loss. After training, we can feed raw neural signals to the model to distill task-relevant neural signals and generate task-relevant neural signals from $\mathbf{z}_{prior}$.

5.5. 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 folds, taking three folds as the training set, one as the validation set, and one as the test set. The validation sets are used to choose hyperparameters. The reported performance is averaged over test sets of five folds. To avoid overfitting, we apply the early stopping technology. Specifically, we assess the criteria (loss for training distilled method, 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 specified epochs, we stop the training process.

5.6. Implementation details for methods

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

- dVAE. The encoder of dVAE uses two hidden layer multilayer perceptron (MLP) with 300 and 100 hidden units. The decoder of dVAE is symmetric with the encoder. The dimension of the latent variable is set to 50. The prior networks use one hidden layer MLP with 300 units. The β is set to 0.001. The α is set to 0.3, 0.4, 0.7 for dataset A, B, and C, respectively. The choice of α and β is based on the decoding R^2 of task-relevant (D_{re}) and irrelevant (D_{ir}) signals using ANN with the formula $0.75 \times D_{re} + 0.25 \times (1 - D_{ir})$ on validation sets.
- pi-VAE. The original paper obtains the approximate posterior using the label prior(Eq.6 in the original paper), so the label information is used in the test stage. However, we can not use label information (kinematics) to extract task-relevant signals in the test stage, so we infer the posterior without using the label prior. The parameters of the encoder, decoder and prior networks of pi-VAE are the same as dVAE.
- VAE. The parameters of the encoder and decoder of VAE are the same as dVAE.
- LFADS. The hidden units of the encoder for the generator's initial conditions, the controller, the generator are set to 200. The dimensionality of latent factor is set to 50. The dimensionality of inferred inputs is set to 1. The training strategy follows the practice of the original paper.
- PSID. The horizon parameter of PSID is set to 3, and the dimensionality of latent variable is set to 6.
- ANN. ANN has two hidden layers with 300 and 100 hidden units.
- KF. The matrix parameters of observation and state transition process are optimized with the least square algorithm.

5.7. Percentage of explained variance captured in a subspace.

We applied PCA to task-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 projecting irrelevant signals onto a subspace composed of some relevant PCs and vice versa. The percentage of variance captured is:

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

where $\mathbf{D} \in \mathbb{R}^{N\times d}$ is a d-dimensional population basis over N neurons, $\mathbf{C} \in \mathbb{R}^{N\times N}$ is the covariance matrix of the neural signals, and $\mathrm{Tr}\left(\mathbf{C}\right) = \sum_{i=1}^{N} \lambda_i$, where λ_i is the ith largest eigenvalue for $\mathbf{C}$. The matrices $\mathbf{D}$ and $\mathbf{C}$ can be computed for both task-irrelevant and relevant signals. The percentage variance is a quantity between 0% and 100%.

5.8. The composition of raw signals' variance.

Suppose $\mathbf{X} \in \mathbb{R}^{1\times T}$, $\mathbf{Y} \in \mathbb{R}^{1\times T}$, and $\mathbf{Z} \in \mathbb{R}^{1\times T}$ are the random variable for a single neuron of task-relevant signals, task-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:

$$\begin{aligned}
\mathrm{Var}(\mathbf{Z}) &= \mathrm{Var}(\mathbf{X}+\mathbf{Y}) = \mathbb{E}\left\{[\mathbf{X}-\mathbb{E}(\mathbf{X})] + [\mathbf{Y}-\mathbb{E}(\mathbf{Y})]\right\}^2 \\
&= \mathbb{E}[\mathbf{X}-\mathbb{E}(\mathbf{X})]^2 + \mathbb{E}[\mathbf{Y}-\mathbb{E}(\mathbf{Y})]^2 + 2\mathbb{E}[\mathbf{X}-\mathbb{E}(\mathbf{X})][\mathbf{Y}-\mathbb{E}(\mathbf{Y})] \\
&= \mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{Y}) + 2\mathrm{Cov}(\mathbf{X},\mathbf{Y}),
\end{aligned}$$

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 $\mathrm{Var}(\mathbf{X})$, the variance of irrelevant signals $\mathrm{Var}(\mathbf{Y})$, and the correlation between relevant and irrelevant signals $2\mathrm{Cov}(\mathbf{X},\mathbf{Y})$. If there are N neurons, calculate the composition of each neuron separately and then add it up to get the total composition of raw signals.

5.9. Reordered correlation matrix of neurons.

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

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.

5.10. The improvement ratio of lower and higher speed parts

5.10.1. Split lower and higher speed parts Since the speed ranges of different datasets are different, it is hard to determine a common speed threshold to split lower and higher speed parts. Here we used the squared error to split the two speed parts. Specifically, first, we can calculate small 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 are lower speed parts, and those whose speed is greater than the speed threshold are higher speed parts. 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).

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

$$AIR_{low} = -\frac{E_{p+s}^{low} - E_p^{low}}{E_p^{low}} \approx -\frac{E_{p+s}^{low} - 0.5E_p}{0.5E_p} \tag{9}$$

$$AIR_{high} = -\frac{E_{p+s}^{high} - E_p^{high}}{E_p^{high}} \approx -\frac{E_{p+s}^{high} - 0.5E_p}{0.5E_p} \tag{10}$$

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

5.10.3. 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:

$$RIR^i = -\frac{E_{p+s}^i - E_p^i}{E_p^i}, \tag{11}$$

where i denotes the ith sample.

6. Supplementary Materials

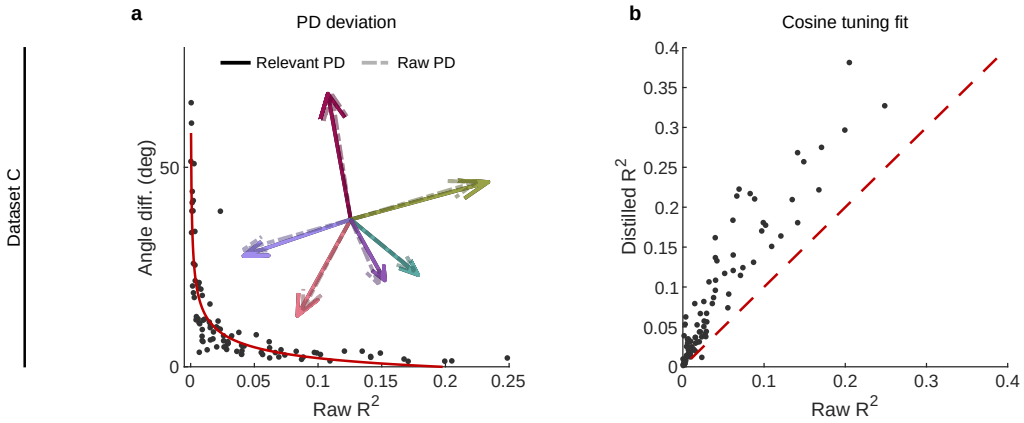

Figure S1: **The effect of irrelevant signals on relevant signals at the single-neuron level. a,b** Same as Fig. 3, but for dataset C. **a**, The change of angle difference (AD) of PD deviation between raw and distilled signals with R^2 of raw signals. Each black point represents a neuron (n=91). The red curve is the fitting curve between R^2 and AD. Five exemplar neurons' PDs before and after distillation 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 (R^2) before and after distillation of single neurons (black points), where the x-axis and y-axis represent neurons' R^2 of raw and distilled signals.

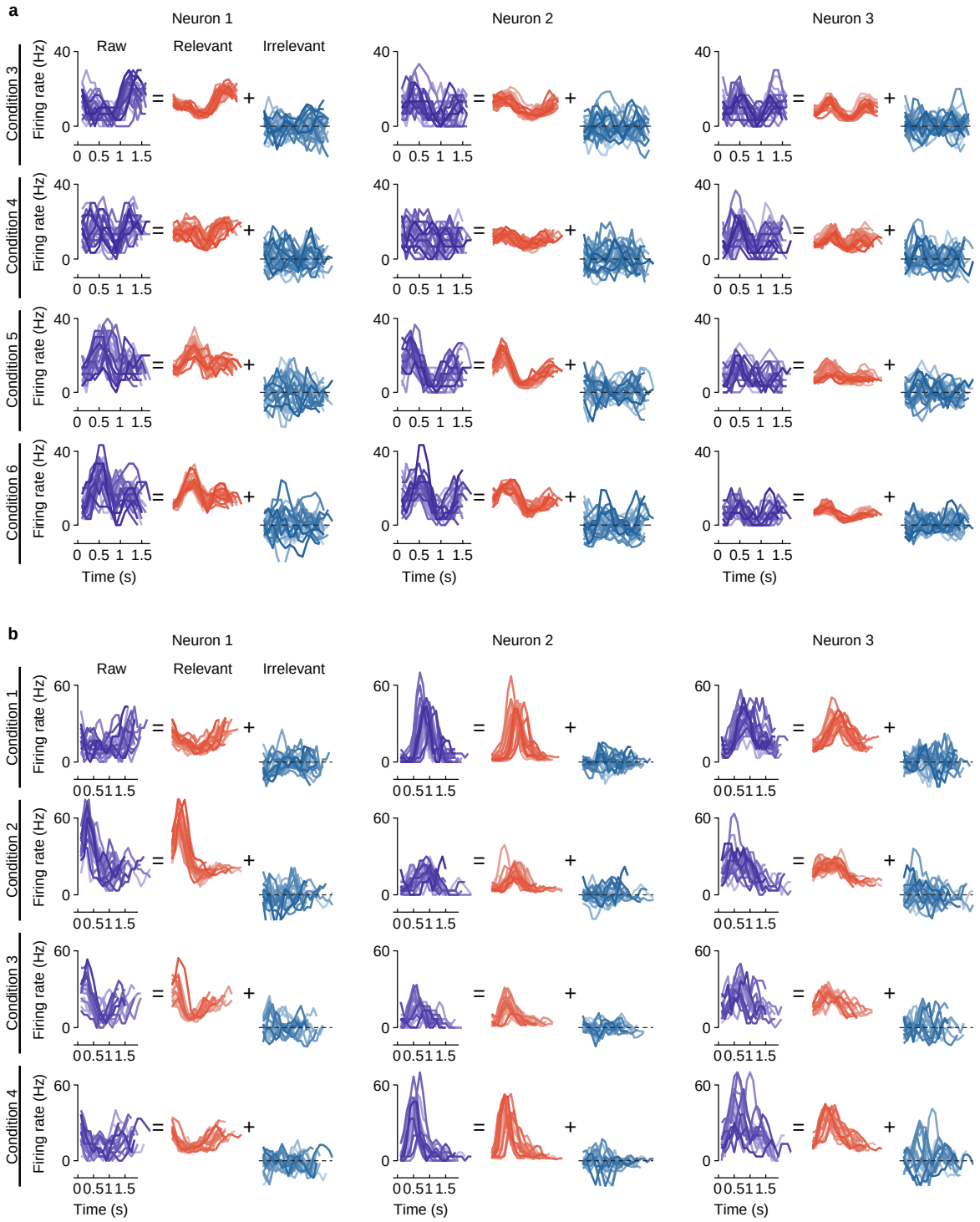

Figure S2: **The firing activity of exemplar neurons a**, Example of three neurons' raw firing activity decomposed into task-relevant and irrelevant parts using all trials in held-out test sets for four conditions (4 of 8 directions) of center-out reaching task. **b**, Example of three neurons' raw firing activity decomposed into task-relevant and irrelevant parts using all trials in held-out test sets for four conditions (4 of 12 conditions) of obstacle avoidance task.

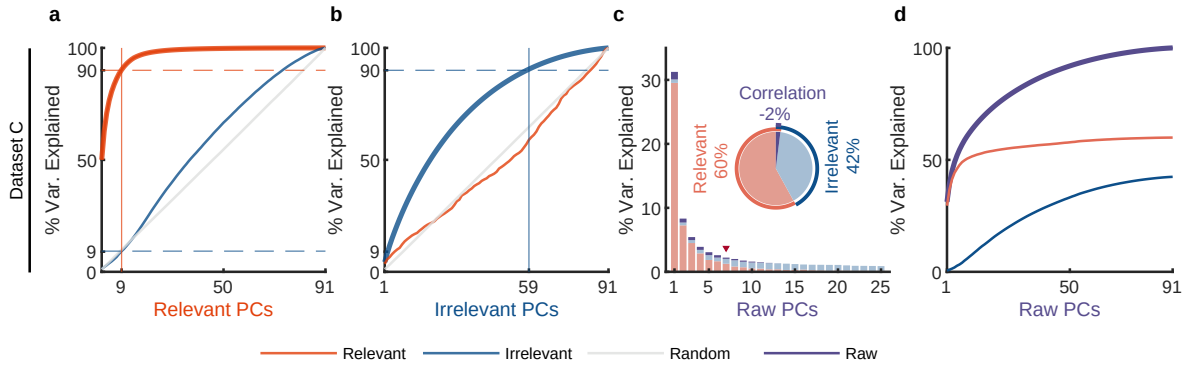

Figure S3: **The effect of irrelevant signals on relevant signals at the population level. a-d**, Same as Fig. 4, but for dataset C. **a,b**, PCA is separately applied on relevant and irrelevant signals to get relevant PCs and irrelevant PCs. The thick lines represent the cumulative variance explained for the signals on which PCA has been performed, while the thin lines represent the variance explained by those PCs for other signals. Orange, blue, and gray colors indicate relevant signals, irrelevant signals, and random Gaussian noise (for chance level). The cumulative variance explained for task-relevant (**a**) and irrelevant (**b**) signals got by dVAE 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 orange, 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 and 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 (orange) and irrelevant (blue) signals.

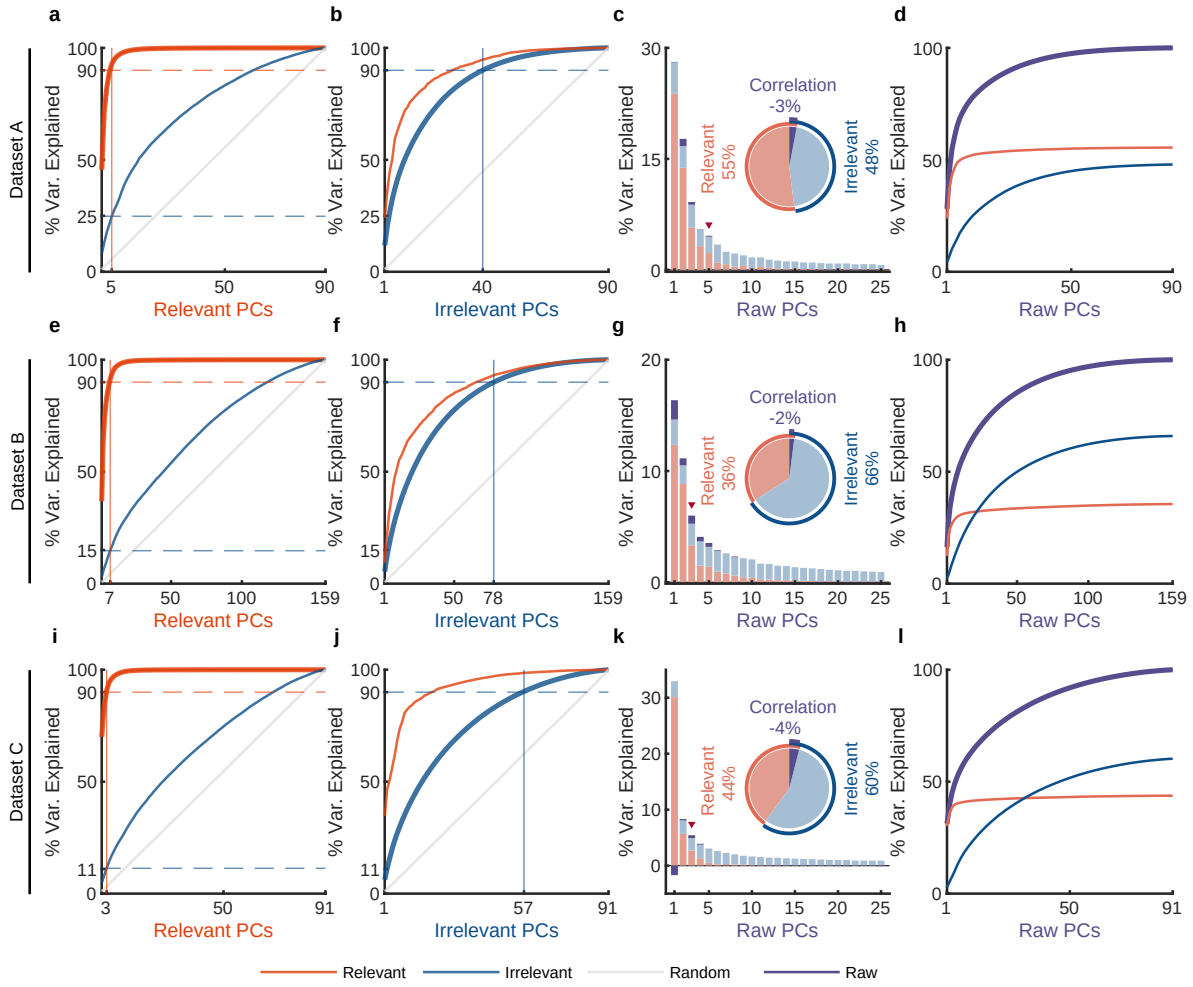

Figure S4: **The effect of irrelevant signals on relevant signals at the population level. a-l**, Same as Fig. 4 and Fig. S3, but for pi-VAE. **a,b**, PCA is separately applied on relevant and irrelevant signals to get relevant PCs and irrelevant PCs. The thick lines represent the cumulative variance explained for the signals on which PCA has been performed, while the thin lines represent the variance explained by those PCs for other signals. Orange, blue, and gray colors indicate relevant signals, irrelevant signals, and random Gaussian noise (for chance level). The cumulative variance explained for task-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 orange, 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 and 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 (orange) and irrelevant (blue) signals. **e-h, i-l**, Same as **a-d**, but for datasets B and C.

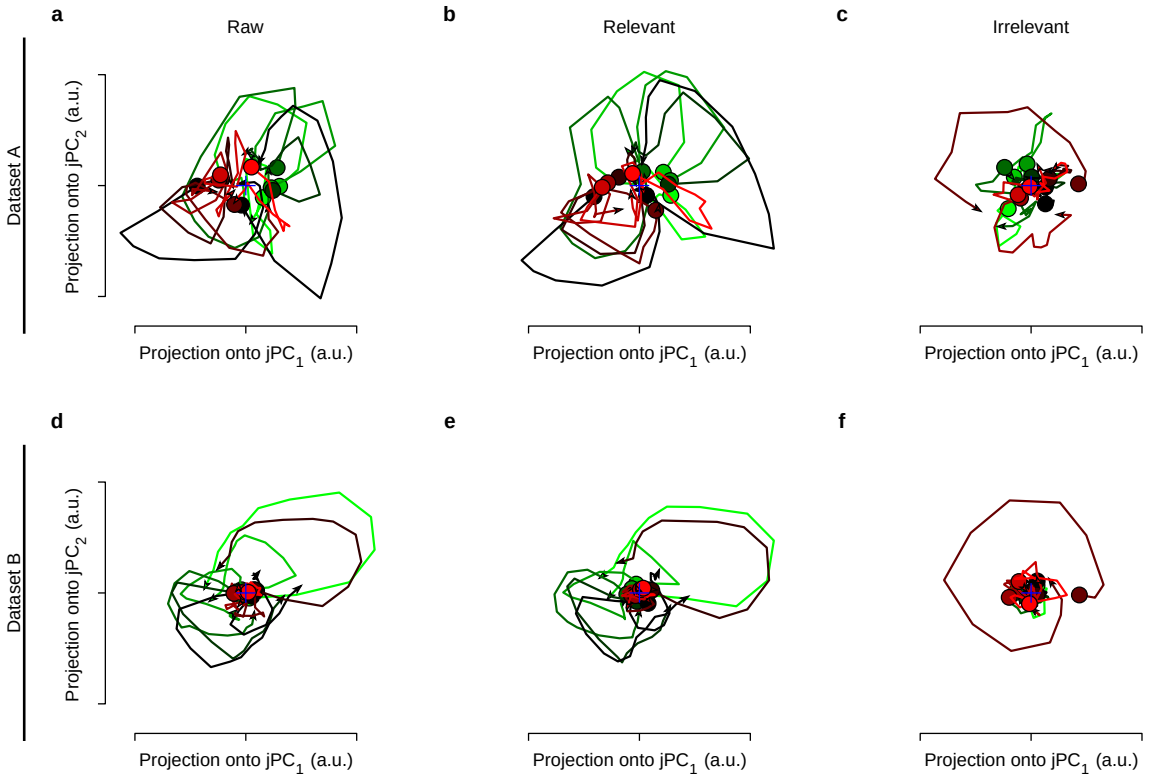

Figure S5: **The rotational dynamics of raw, relevant, and irrelevant signals.** datasets A and B have twelve and eight conditions, respectively. We get the trial-averaged neural responses for each condition, then apply jPCA to raw, relevant, and irrelevant signals to get the top two jPC, respectively. **a**, The rotational dynamics of raw neural signals. **b**, The rotational dynamics of relevant signals obtained by dVAE. **c**, The rotational dynamics of irrelevant signals obtained by dVAE. **d-f**, Same as **a-c**, but for dataset B.

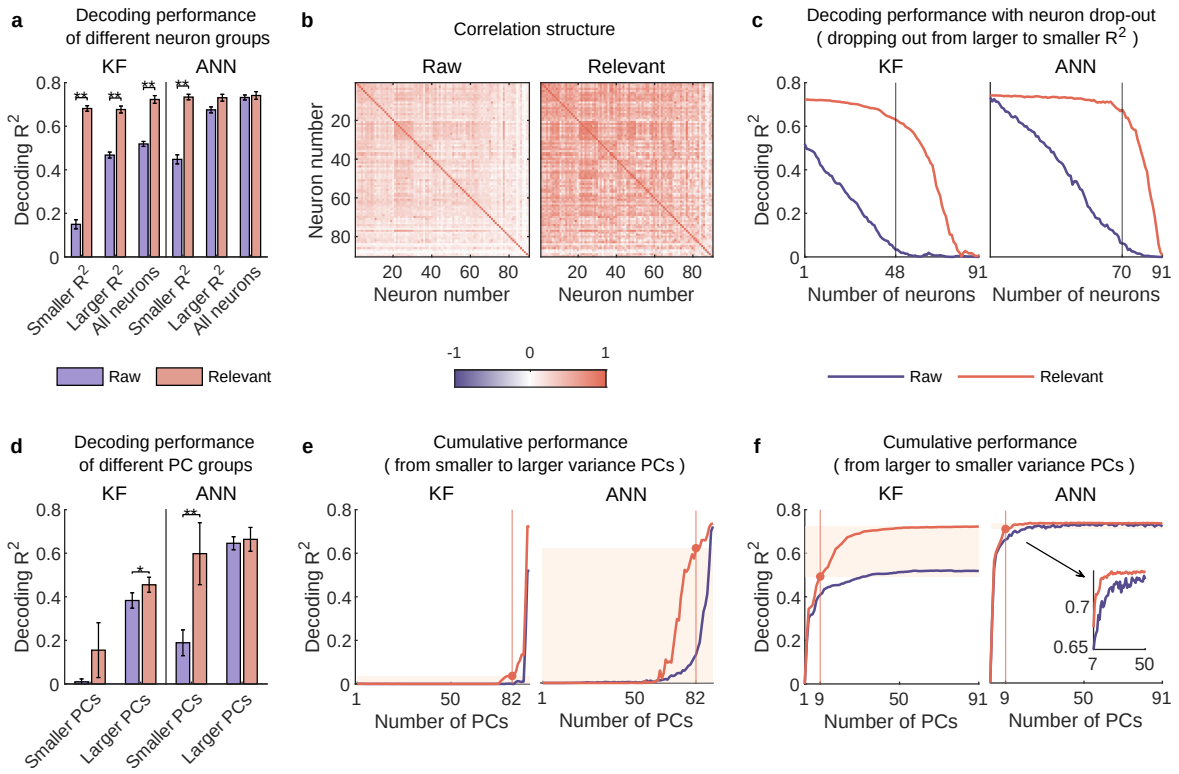

Figure S6: **trivial neural activities encode rich task information. a-c**, Same as Fig. 5, but for dataset C. **d-f**, Same as Fig. 6, but for dataset C. **a**, The comparison of decoding performance between raw (purple) and distilled signals (orange) with different neuron groups, including smaller R^2 neuron ($R^2 <= 0.03$), larger R^2 neuron ($R^2 > 0.03$), and all neurons. Error bars indicate mean $\pm$ SD across cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{\star}P < 0.05$, $^{\star\star}P < 0.01$. **b**, The correlation matrix of all neurons of raw and task-relevant signals. **c**, The decoding performance of KF (left) and ANN (right) with neurons dropped out from larger to smaller R^2. The vertical gray lines indicate the number of dropped neurons at which raw and task-relevant signals have the greatest performance difference. **d**, The comparison of decoding performance between raw (purple) and distilled signals (orange) 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 $\pm$ s.d. across five cross-validation folds. Asterisks denote significance of Wilcoxon rank-sum test with $^{\star}P < 0.05$, $^{\star\star}P < 0.01$. **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 orange 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 orange patches indicate the decoding gain of the last 10% variance signals of relevant signals superimposing on their top 90% variance signals.

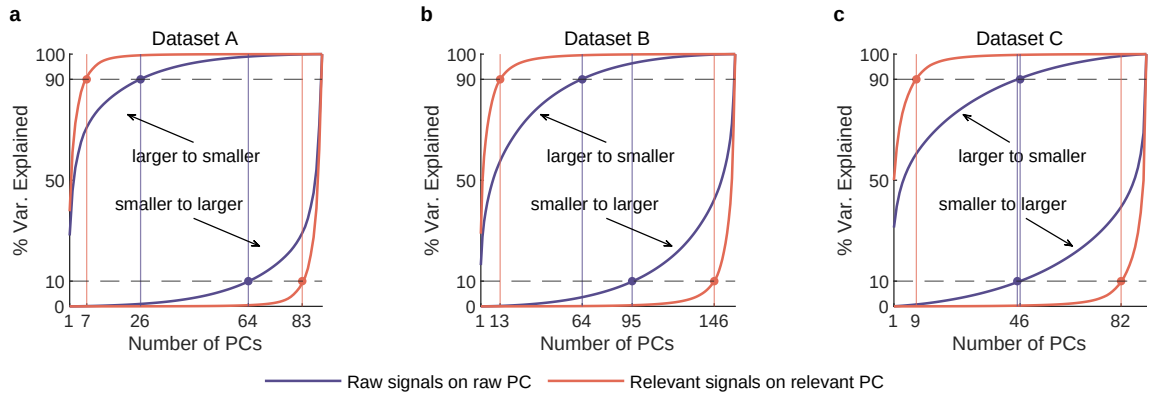

Figure S7: **The cumulative variance of raw and distilled signals.** **a**, PCA is applied separately on raw and distilled task-relevant signals to get raw PCs and relevant PCs. The cumulative variance of raw (purple) and task-relevant signals (orange) on dataset A. Two upper left corner curves denote 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 in task-relevant (orange) 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 tasks. **b,c**, Same as **a**, but for datasets B and C.

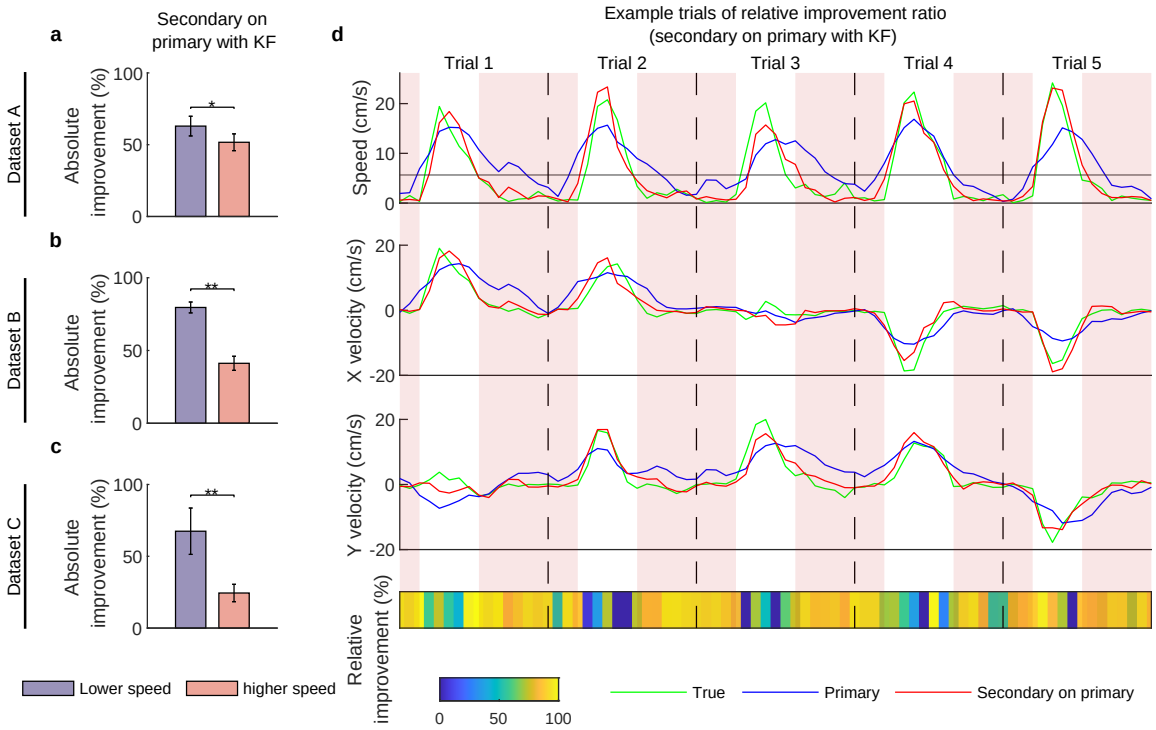

Figure S8: **Smaller variance PC signals preferentially improve lower-speed velocity.** **a**, The comparison of absolute improvement ratio between lower-speed (purple) and higher-speed (orange) velocity when superimposing secondary signals on primary signals with KF on dataset A. Error bars indicate 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 (blue line) and full signals (red line; superimposing secondary signals on primary signals) and the true velocity (green line). 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.

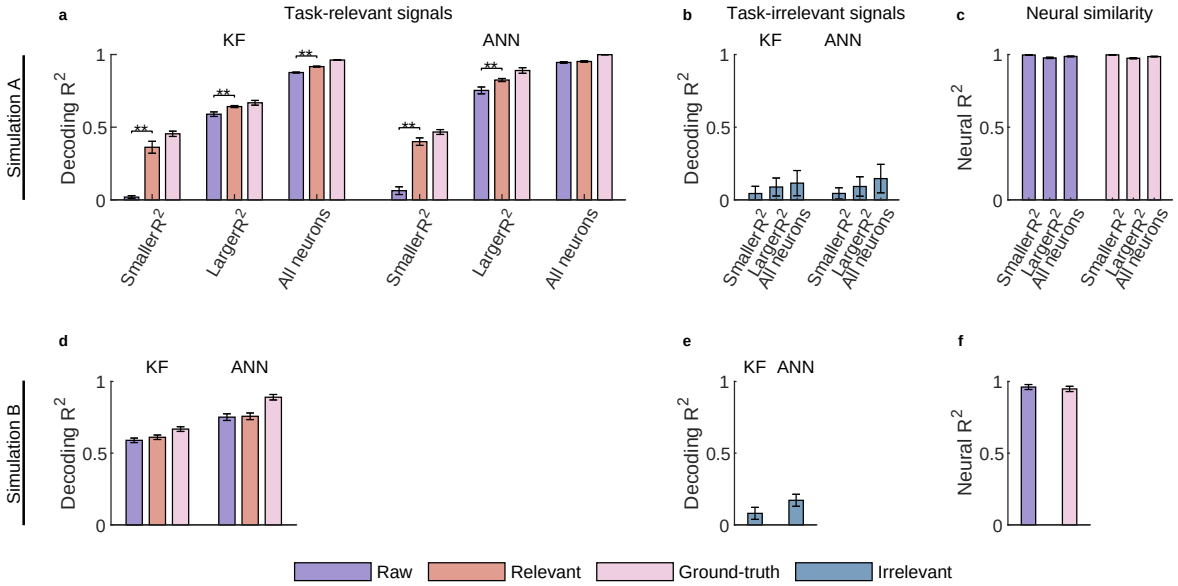

Figure S9: **Simulation results. a-c**, The simulation dataset A simulated a situation where the smaller R^2 neurons contain a certain amount of task information, but the task information cannot be decoded from these neurons due to being covered by noise to demonstrate that dVAE can utilize the larger R^2 neurons to help the smaller R^2 neurons restore their original face. **a**, The comparison of decoding performance between raw (purple), distilled (orange) and ground-truth (pink) signals with different neuron groups, including smaller R^2 neuron ($R^2 <= 0.03$), larger R^2 neuron ($R^2 > 0.03$), and all neurons. Error bars indicate mean ± s.d. (n=5 folds). Asterisks denote significance of Wilcoxon rank-sum test with $^{**}P < 0.01$. **b**, The decoding performance of task-irrelevant signals got by dVAE with different neuron groups. Error bars indicate mean ± s.d. (n=5 folds). **c**, The neural similarity between distilled task-relevant and raw signals (purple) and between relevant and ground-truth signals (pink) with different neuron groups. Error bars indicate mean ± s.d. (n=5 folds). **d-f**, The simulation dataset B simulated the situation where linear decoding is significantly inferior to nonlinear decoding to demonstrate that dVAE can not make the linear decoder achieve similar performance as the nonlinear decoder. **d**, The comparison of decoding performance between raw (purple), distilled (orange), and ground-truth (pink) signals. Error bars indicate mean ± s.d. (n=5 folds). **e**, The decoding performance of task-irrelevant signals got by dVAE. Error bars indicate mean ± s.d. (n=5 folds). **f**, The neural similarity between distilled task-relevant and raw signals (purple) and between relevant and ground-truth signals (pink). Error bars indicate mean ± s.d. (n=5 folds).