bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Predicting proprioceptive cortical anatomy and neural coding with topographic autoencoders

Kyle P. Blum[1*], Max Grogan[2*], Yufei Wu[2*], J. Alex Harston[2], Lee E. Miller[1] & A. Aldo Faisal[2,3]
[*]contributed equally to the work
[1] Northwestern University
[2] Imperial College London
[3] University of Bayreuth

Proprioception is one of the least understood senses, yet fundamental for the control of movement. Even basic questions of how limb pose is represented in the somatosensory cortex are unclear. We developed a topographic variational autoencoder with lateral connectivity (topo-VAE) to compute a putative cortical map from a large set of natural movement data. Although not fitted to neural data, our model reproduces two sets of observations from monkey centre-out reaching: 1. The shape and velocity dependence of proprioceptive receptive fields in hand-centered coordinates despite the model having no knowledge of arm kinematics or hand coordinate systems. 2. The distribution of neuronal preferred directions (PDs) recorded from multi-electrode arrays. The model makes several testable predictions: 1. Encoding across the cortex has a blob-and-pinwheel-type geometry PDs. 2. Few neurons will encode just a single joint. Topo-VAE provides a principled basis for understanding of sensorimotor representations, and the theoretical basis of neural manifolds, with applications to the restoration of sensory feedback in brain-computer interfaces and the control of humanoid robots.

Keywords:

Proprioception, Cortical Maps, Topographic Mapping, Deep Learning, Natural Sensory Statistics, Sensory Ecology, Variational Autoencoder, Computational Neuroscience, Movement kinematics, Neural Activity, Primary Somatosensory Cortex, Natural Behaviour, Neuromechanics

Introduction

Somatosensation includes the familiar sense of touch, provided by receptors in the skin, and proprioception, the much less consciously perceived sense that informs us about the pose, motion and associated forces acting on our limbs. While the former has received much scientific attention, proprioception is often overlooked, yet this modality of sensory feedback is essential for our ability to plan, control and adapt movements. In engineering, the control of robotic movement would be impossible if the controller did not know the location of its actuators; correspondingly in human motor control (proprioceptive) feedback control theory is the preeminent explanation for the computations underlying limb control (Todorov and Jordan 2002; Scott 2004). Moreover, individuals with proprioceptive neurological deficits, such as patient IW, have profound motor deficits even in the presence of vision and an intact motor system (Tuthill and Azim 2018;

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sainburg, Poizner, and Ghez 1993). Similarly, recent major developments in neuroprosthetics are centred on restoring sensory feedback as well as limb motion through bi-directional interfaces (Flesher et al. 2021) and will likely require not only touch but also proprioceptive feedback to restore functional capability (Faisal 2021). Similarly, understanding proprioceptive encoding is essential for restoration of both motor action and sensory function in clinical rehabilitation (Formento et al. 2018).

Unlike hearing or vision, proprioceptive afferent pathways originate not from a single organ but from diverse families of mechanoreceptors within muscles, tendons, joints, and the skin itself (Blum et al. 2021). Proprioceptive information ascends within the dorsal column pathway through the dorsal root ganglia, the dorsal column nuclei, and thalamus before arriving in the primary somatosensory cortex (S1). Crucially, neurons are somatotopically organised throughout this pathway, meaning that neighbouring neurons encode stimuli from closely related parts of the body. This gives rise to the sensory homunculus, which results from the ordered mapping of tactile representations of the body's surface across the cortical surface (Penfield and Boldrey 1937).

Although proprioception is generally acknowledged to be critical to motor behaviour, the corresponding proprioceptive maps - particularly that of primate area 3a, but also the mixed modality area 2 - are much less distinct and well understood than those of the tactile submodality (areas 1 and 3b). We know that the proprioceptive and tactile systems often encode overlapping information, e.g. mechanoreceptors in our skin and interosseous membranes respond to deformation and vibration, contributing to the sense of body position and movement (Tuthill and Azim 2018). Indeed, the firing rates of somatosensory neurons with cutaneous receptive fields in area 2 can be used to decode limb movement as accurately as those with muscle fields (Weber et al. 2011). While proprioceptive cortical areas are critical for our ability to generate goal-directed complex behaviour it is unclear what properties of proprioceptive cortical coding facilitate this capability. Crucially, we lack an accepted hypothesis about the computational principles that drive the mapping of proprioceptive arm representations onto the cortex.

The limited nature of our understanding of proprioceptive neural representations in the brain has two major causes. First is the difficulty recording with many electrodes from sulcal proprioceptive areas in primates (Huffman and Krubitzer 2001), and second is the difficulty delivering independent proprioceptive stimuli in comparison to other senses, such as vision (Rossi, Harris, and Carandini 2020) and touch (Killebrew et al. 2007). To avoid these limitations, here we are combining computational modelling and natural movement kinematics data to test hypotheses of organisational mechanisms that are currently beyond the capability of experimental recording techniques. The combination of experiment and theory has been useful for explaining population-level coding in other sensory modalities, such as vision, olfaction, and touch (Stringer et al. 2019; Pehlevan, Genkin, and Chklovskii 2017), but has only begun to be applied for proprioception (Sandbrink et al. 2020). Much of the computational theoretical work on other senses has either focused on predicting specific neural coding features from natural sensory statistics (Olshausen and Field 1997; Stringer et al. 2019), or has ignored the details of neural coding and focused on the spatial distribution of stimulus representation across the cortex (Obermayer and Blasdel 1993).

Our system is the proprioception of the right arm (Fig. 1.A) for which we have collected natural behavioural data from daily life in humans (food preparation, eating, etc; Fig. 1.B and

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

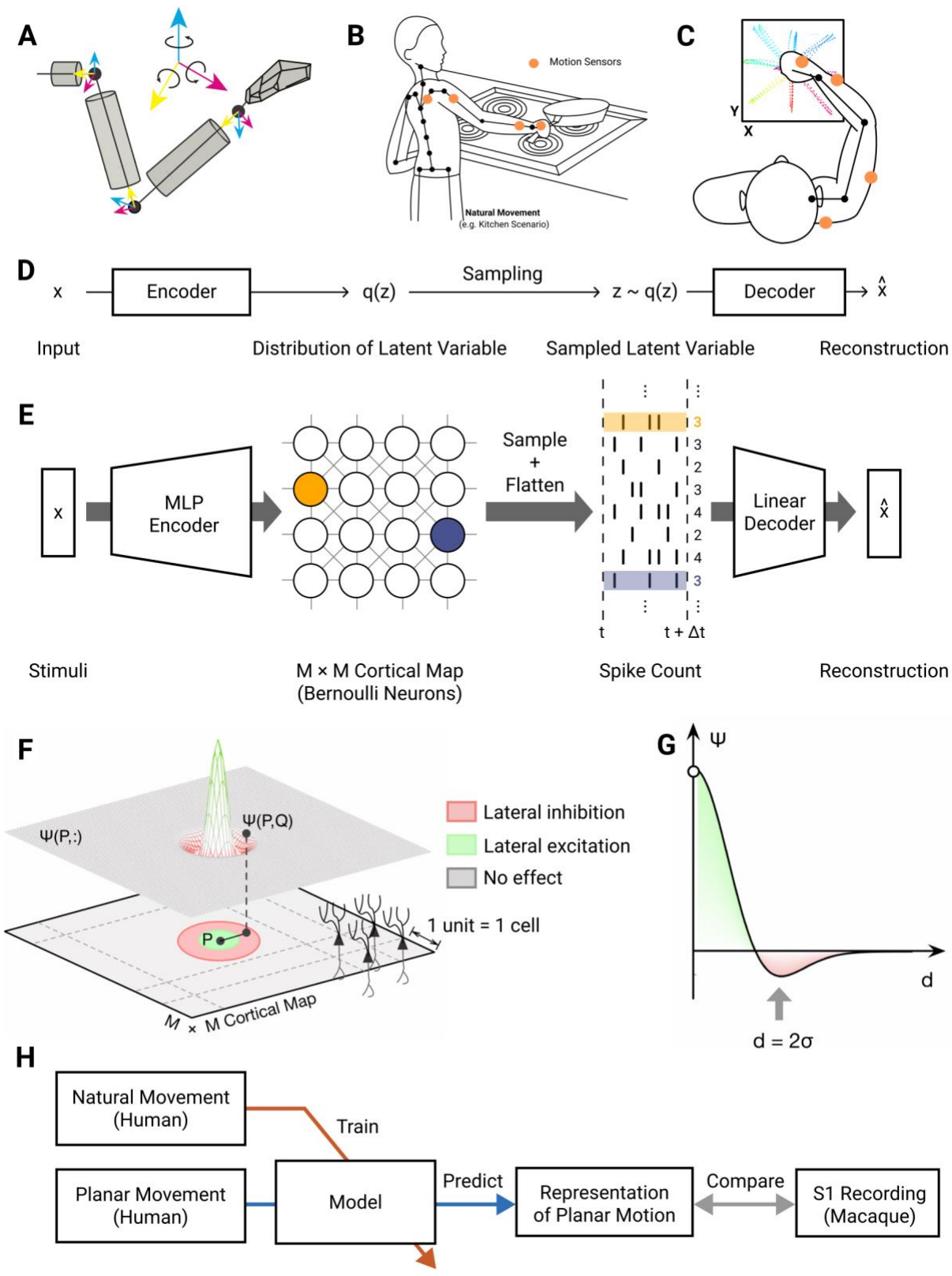
Supplemental Fig. S1) as well as constrained, planar centre-out reaching in both human and monkeys (Fig. 1.C). This kinematic data describes the dynamics of the pose of the body and thus stand in for the proprioceptive sensory state. We want to relate these kinematics to proprioceptive encoding using modelling and existing single-unit recordings from monkeys. We formulated a novel computational model that predicts both neural coding of single neurons and spatial organisation of these neurons across the cortex (Fig. 1.D-G). We used a small set of general computational elements reflecting principles and mechanisms found in sensory systems but not been previously unified. Models should: 1) use the information maximisation principle, which postulates that efficient sensory representations in the brain reflect natural sensory statistics, and implies that they are essential in shaping neural representations, 2) be stochastic, generative, and decodable, to reflect the natural variability in the data, produce neural activity to mimic that of the biological system, and offer the means to reconstruct the relevant original sensory input from its output, 3) implement (in Marr's sense; (Marr 1982)) neural computations that are performed by locally interacting neurons through synaptic interactions, rather than by an abstract computational machine.

In the context of proprioception, these principles require that our model should learn how to translate proprioceptive inputs from movements of the body into a latent representation of proprioception (which is read out as neural firing). We strove to remove human induction bias in the model building, except for specific characteristics relating to the principles laid out above. Therefore, we begin by using an unsupervised deep neural network model, a variational autoencoder (VAE, (Kingma and Welling 2014)) (Fig. 1.D) to perform efficient feature learning of proprioceptive stimuli. The primary training objective of an autoencoder is simply to reconstruct the input stimuli – thus implementing the infomax principle (Linsker 1988; Barlow 1961), our first outlined principle. Furthermore, the bottleneck of a VAE models a latent distribution of spiking neurons, which can be sampled from and decoded, satisfying our second principle.

In following with our third outlined principle, and the observation that spatial organization is relevant to cortical coding, we introduce a 2-dimensional structure to the latent ("bottleneck") layer of the VAE, representing a simplified proprioceptive cortex (Fig. 1.E). A conventional "vanilla" VAE model will capture the properties of encoded sensory features but would be devoid of any of the spatial properties of cortical neurons that are critical for understanding how sensory stimuli are represented across the cortical surface. Incorporating spatial relationships in neural coding models has not been well investigated, yet anatomical structure and function in neural computation are fundamentally linked by biophysical constraints (Sterling and Laughlin 2015). We therefore also implemented a topographical mechanism, using a lateral interaction term between the units in the VAE's latent layer (see Methods for detailed motivation and mathematical description; Fig. 1.F). This term corresponds biologically to a spatial distribution of short-range excitatory and longer-range inhibitory synaptic interactions between neurons (Fig. 1.G).

We call our model the "topo-VAE" and compare it with existing recordings from area 2 of somatosensory cortex. It could equally well be applied to other proprioceptive areas such as 3a or even 5. Models with its basic form could be applied to other sensory modalities, including touch, vision, and hearing. Here, this approach has enabled us to uncover novel insights into proprioceptive coding by linking anatomical structure and function so that we can test our model's predictions using the sparse set of obtainable data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1: Task Context and model architecture. (A) *9-dimensional arm movement data acquired from motion capture suit is represented by Euler angles between body segments, following the ISB Euler angle extractions (G. Wu et al. 2005) in a ZXY coordinate.* ***(B)*** *Illustrative example of movements carried out in the natural scenario (cf. Fig. S1 for further examples).* ***(C)*** *Illustrative example of planar centre-out reaching movements.* ***(D)*** *General architecture of a VAE: Inputs are encoded as a latent distribution, from which samples are decoded to reconstruct the original input.* ***(E)*** *Topo-VAE architecture. Our cortical neural representation is modelled by an 80x80 cortical grid of artificial neurons in the latent layer q(*z*). In contrast to conventional VAEs which model these latent neurons as multi-dimensional Gaussian random variables, we used Poisson random variables. Input stimuli drive these model neurons via a multi-layer perception (MLP) encoder network (2 layers of 500 neurons each). A linear decoder is applied to the spike counts emitted from the cortical map within a time interval Δt, to reconstruct the sensory input stimuli.* ***(F)*** *To embed the model neurons in a cortex-like topographic context we use a Mexican-hat lateral effect between the latent layer neurons.* ***(G)*** *This interaction* $\Psi(P,Q)$ *is a function of the Euclidean distance* $d=||P-Q||$ *between a pair of neurons P,Q and is characterised by a length scale* σ*. Nearby neurons are excited, intermediate-range neurons are inhibited and there is no effect on distant neurons.* ***(H)*** *Flow of natural and planar movement data (joint angle velocities) in this work.*

Results

We have developed a novel model of cortical sensory representation (see Methods for details) that predicts both function and structure – the topo-VAE model. We trained, test and validate our model as follows (Fig. 1.H): We trained our topo-VAE model on natural daily human arm movement kinematics and explored the emergent proprioceptive representations in its cortical layer (i.e. the latent layer of the topo-VAE). After training, we used the model to generate neural responses by providing it kinematic data from a centre-out reaching task performed by human subjects. We then compared the properties of these simulated neural activities (i.e. the generated spike trains) to those of S1 neurons which were previously recorded from monkeys performing a planar centre-out reaching task. The human and monkey reaching data were rescaled so that the biomechanical differences between human and monkey arm movements are kept small. We characterised the proprioceptive neural representations at two levels of description: first, we considered the spatial tuning curves of individual neurons and second, we considered tuning preference maps across our model's cortical surface.

We used three classical measurements to summarise neural coding properties of S1 neurons for this centre-out task: 1) the preferred direction (PD) of single-neuron firing rates, 2) the full tuning surface defined by firing rate modulation as a function of both direction and speed of movement, and 3) the overall distribution of firing rates throughout the behaviour. First, we analysed PDs of neurons in the recordings and in our modelling (Fig. 2). The PD refers to the movement direction of the hand (when performing the centre-out reaching task) in which a neuron is most activated. Here, we find that overall, modelled neurons exhibit PD tuning that is similar to that of recorded neurons. We find many pairs of modelled and recorded neurons (Fig. 2.A,B respectively) with very similar tuning. Moreover, the distributions of PDs were quite similar for the topo-VAE modelled neurons (Fig. 2.C) and our recorded neurons (Fig. 2.D). To further quantify the similarity

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between PD distributions we compared their entropies; high entropy values imply that the PD distribution is more uniform, low ones the opposite. The entropy of the PD distributions of our topo-VAE model was similar to that of the recorded data (4.79 vs 5.10 bits, respectively: out of maximum entropy of 5.17 bits). Moreover, the orientations of the bimodal PD distributions (cf. Fig. 2.C,D) are well aligned between modelled and recorded neurons.

Second, we compared the full tuning surfaces of the modelled and recorded neurons (Fig. 3.A), a much more rigorous test than simply a comparison of PDs. We found that the cumulative distribution of half-peak widths was qualitatively similar for modelled and recorded neurons (Fig. 3.B). Since the tuning depth increases as a function of velocity, we also quantified the slope of this velocity gradient (Fig. 3.A), and again found them to be qualitatively similar (Fig 3.C). Note that the velocity gradient measurements were adjusted to control for differences in mean firing rate between modelled and recorded neurons. We also introduced a sanity check to ensure that the tuning structure we observe is not just the trivial result of the Poisson generalised linear model (GLM), we used to relate movement velocity to firing rate. To do so, we shuffled the firing rate data so that the distributions of input and output data were each the same, but the functional relationship between them was destroyed. We fit a GLM to this shuffled data as to the original. The shuffling induced a very large shift in the results, with most tuning surfaces having large half-peak widths, indicating weak tuning to direction. Likewise, the velocity gradients of modelled neurons were a much poorer match to recorded neurons after shuffling. This would suggest that both model and neural data results are not trivially the result of using a GLM (Fig. 3.B). In summary, the topo-VAE model and recordings exhibit similar structure in their tuning surfaces with respect to reach direction (Fig. 3.B-C).

Third, we looked at the overall firing rate distributions, which were well fit by Gamma curves for both modelled (Fig. 3.D) and recorded neurons (Fig. 3.E; $R^2 > 0.95$ for equal numbers of neurons). While topo-VAE generates spike count distributions over an arbitrary time window (cf. Fig. 3.D, x-axis), the functional form and relative change of the distributions with altered velocity are not arbitrary. Therefore, we compared the distributions at different endpoint velocities and found that they differ significantly in both modelled and recorded neurons ($p<0.005$; two-sided Wilcoxon signed-rank test), indicating velocity dependence in both the modelled and recorded neurons.

Thus, in summary, topo-VAE captures the variety of tuning properties and firing rate distributions of the recorded neurons without having been fitted to neural data. Instead, the correspondence is solely driven by the movement statistics. Moreover, this was the case even though the recordings were based on a planar centre-out reaching out task, while the model used unconstrained natural movements.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

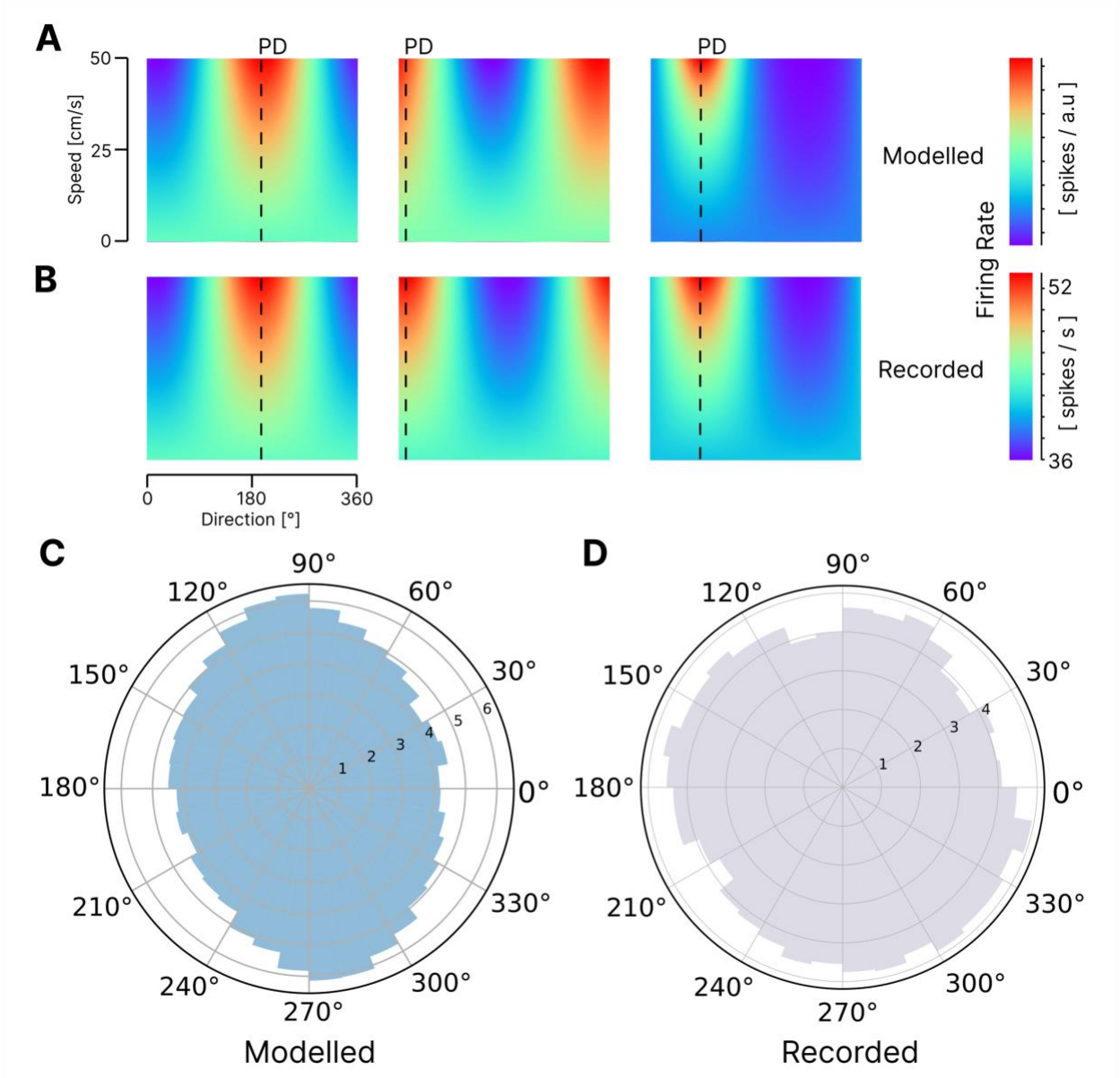


Figure 2. Comparing tuning of modelled and recorded neurons. (A) *Tuning maps of 3 example neurons observed in the latent layer of the topo-VAE under planar movements. Note that the time constant for rates is arbitrary, as it can be adjusted by changing the sampling rate of the input data.* ***(B)*** *Tuning maps of three example area 2 neurons during planar movements, chosen to be similar to those in (A).* ***(C)*** *Log10-frequency circular histograms of preferred directions in our topo-VAE model with lateral effects (n=6400) and* ***(D)*** *in neurons recorded from area 2 of 3 monkeys (n=383). Note, that the neural models were not fitted to monkey neural data, but directly predicted from the statistics of natural human body kinematics.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

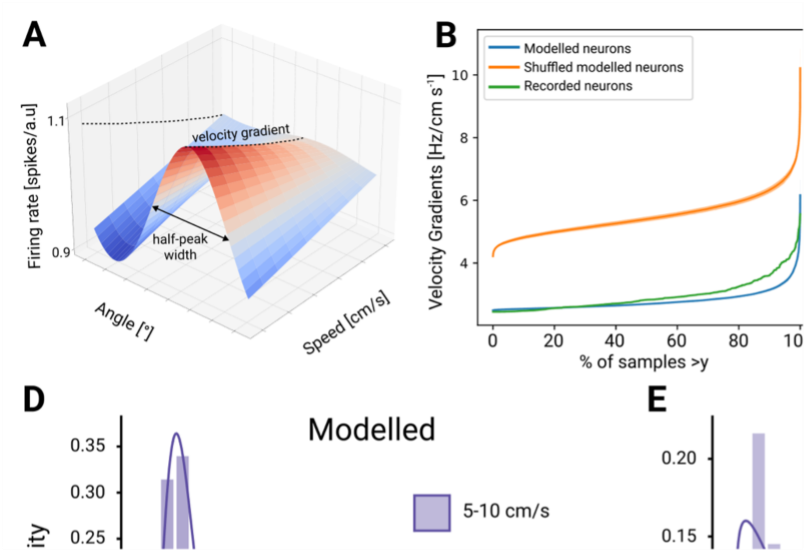


Figure 3. Comparing velocity tuning surface features and firing rate distributions of modelled and recorded neurons. *(**A**) Tuning surface from a single modelled neuron, illustrating two measures we computed to summarise neuronal modulation with velocity: half-peak width of spatial tuning (at maximum end-point speed) and velocity gradient. (**B**) Cumulative distributions of half peak-widths for modelled (n=6400) and recorded neurons (n=383) from 10 bootstrapped subsets of the data. Dashed lines separate results before and after shuffling. Solid lines denote mean, and the shaded areas (visible only for the shuffled velocity gradient curves) in the same colour denote the standard error bounds. (**C**) Cumulative distributions of log-gradients for modelled and recorded neurons, calculated from 10 subsets each of original and shuffled data. (**D**) Firing rate depth of modulation (peak-mean) for two hand-velocity ranges across all reach directions in topo-VAE (6400 neurons). (**E**) Same as D for data from all three monkeys (383 neurons). Histograms represent actual distributions, whereas lines represent fitted gamma distributions.*

We next compared the spatial relationship of PDs in our model neurons to those in the monkey brain within the constraints of the 400 μm spacing of the recording electrodes (Fig. 4). While neurons recorded at two adjacent electrodes correspond to a distance that is substantially beyond the effective range of the Mexican hat distance function, many electrodes record data from more than one distinguishable neuron. The distance between these neurons is within the local neighbourhood of neurons in our model. We can thus compare the similarity of neural coding properties (specifically, the PD) for neurons recorded on the same electrode to that of neurons on different electrodes (cf Fig. 4.A-C). In recorded neurons, same-electrode PD differences followed an exponential distribution, such that the probability that two neurons had similar (within 30 degrees) PDs was much higher than chance (p=0.77; two-sided Wilcoxon signed-rank test; Fig.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.C., blue histogram). Conversely, the probability that two neurons on the same electrode had nearly opposite PDs (between 150º and 180º) was very low (p=0.03; two-sided Wilcoxon signed-rank test; Fig. 4.C orange histogram). We find the same properties reflected in the PD difference distributions of the model (Fig. 4.B).

The topo-VAE model allows us to directly predict the spatial organization of PD tuning across the proprioceptive cortex: PDs are clustered in blob-like structures of similar PDs with boundary regions where the PDs rotate smoothly toward neighbouring directions (Fig. 4.D). This arrangement is interrupted by small, sparsely spaced "pinwheel" regions, representing all PDs in a small neighbourhood. The same-electrode similarity of Fig. 4.B is consistent with this structure, but our 400um electrode spacing does not provide adequate spatial resolution to review the full detail of the actual cortical map at this scale.

In the case of the model, different choices of the neighbourhood hyperparameter, σ, led to differences in the exponential distributions of PD difference (Supplemental Fig. S2.B,C). Small neighbourhood values (σ=1, Supplemental Fig. S2.B,C, left) led to more uniform distributions, whereas larger neighbourhood values (σ={2,3}, Supplemental Fig. 2.B,C, middle and right) led to exponential relationships like those we found in the recorded data, with the closest match to the model being $\sigma = 2$. Note, that for our simple cortical grid model we only considered integer values of σ.

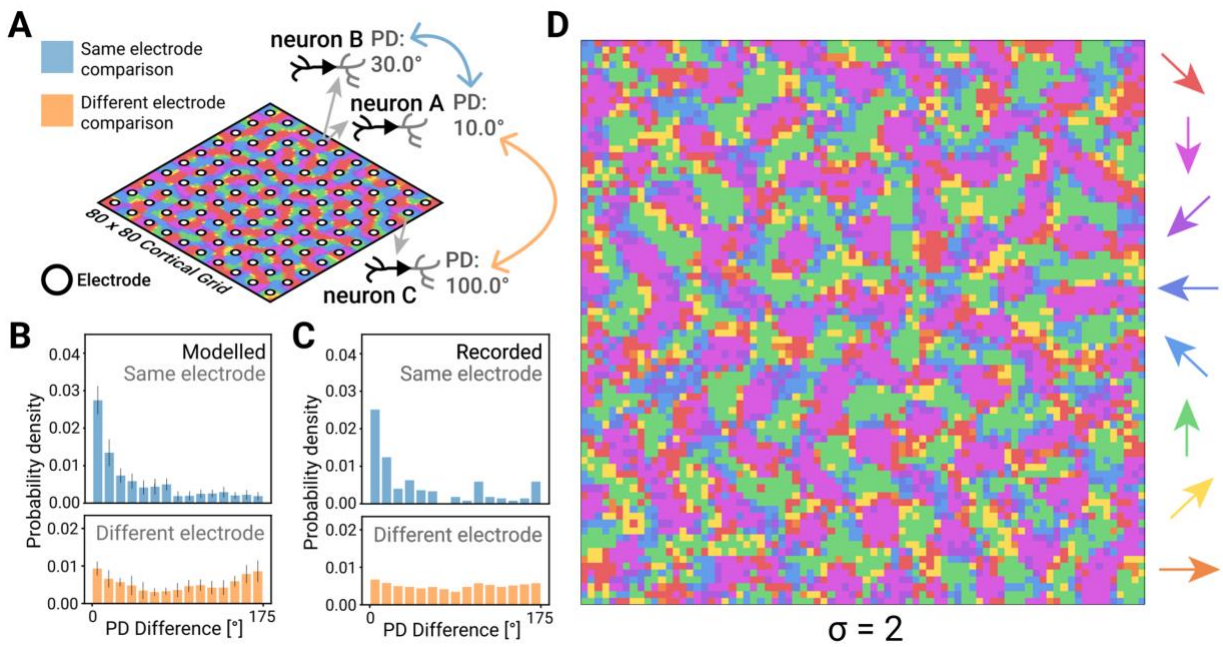


Figure 4. Predicting the topography of proprioceptive cortex. *(**A**) Illustration of pairwise comparisons of preferred directions recorded on same (blue) and different (orange) electrodes, performed in recorded and modelled neurons. Neuron A and B are recorded from the same electrode, Neuron C is recorded from a distant electrode. (**B**) Distribution of PD difference for same electrode (blue histograms) and different electrode (orange histograms) comparisons in modelled neurons and (**C**) recorded neurons (length scale $\sigma = 2$); error bars are* standard deviation. ***(D)*** *The PD map in a topo-VAE with $\sigma = 2$, the hyperparameter value which well*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

approximates the recorded neuron topography (cf. Supplemental Fig. 2 for other values of length scale σ).

We also found that only when trained on the natural 3D human movement data (cf. Fig. 2.C,D), did the orientation of the topo-VAE PD distribution consistently match the PD distribution of recorded neurons (Fig. 5.A-B). Training on the planar centre-out reaching task data led to inconsistent orientations (Fig. 5.C-D). Thus, training the topo-VAE with the full distribution of natural movement data better reproduces the neural coding of the planar centre-out task than does training with the more limited kinematic data alone. These findings at the PD distribution level suggest that the topo-VAE model matches the PD distribution empirical data well, without ever having been fitted to it.

Next, we examined which specific elements of the lateral interaction were important for reproducing the recorded data. We verified through model ablation the importance of the shape of the Mexican hat function (black curve in Fig. 5.F), which determines the strength of excitatory and inhibitory connections as a function of distance between two neurons. Flipping the function upside down (i.e. short range inhibition and intermediate range excitation, orange curve in Fig. 5.F) resulted in no obvious topographic structure (Fig. 5.G, bottom left quadrant). Similarly, removing either the excitatory (red curve in Fig. 5.F) or inhibitory component (green curve in Fig 5.F) produced results unlike the localised topographic structure of our primary model (Fig. 5.G, bottom right and top right quadrants, respectively), suggesting that the specific Mexican hat shape was important for reproducing biological results.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

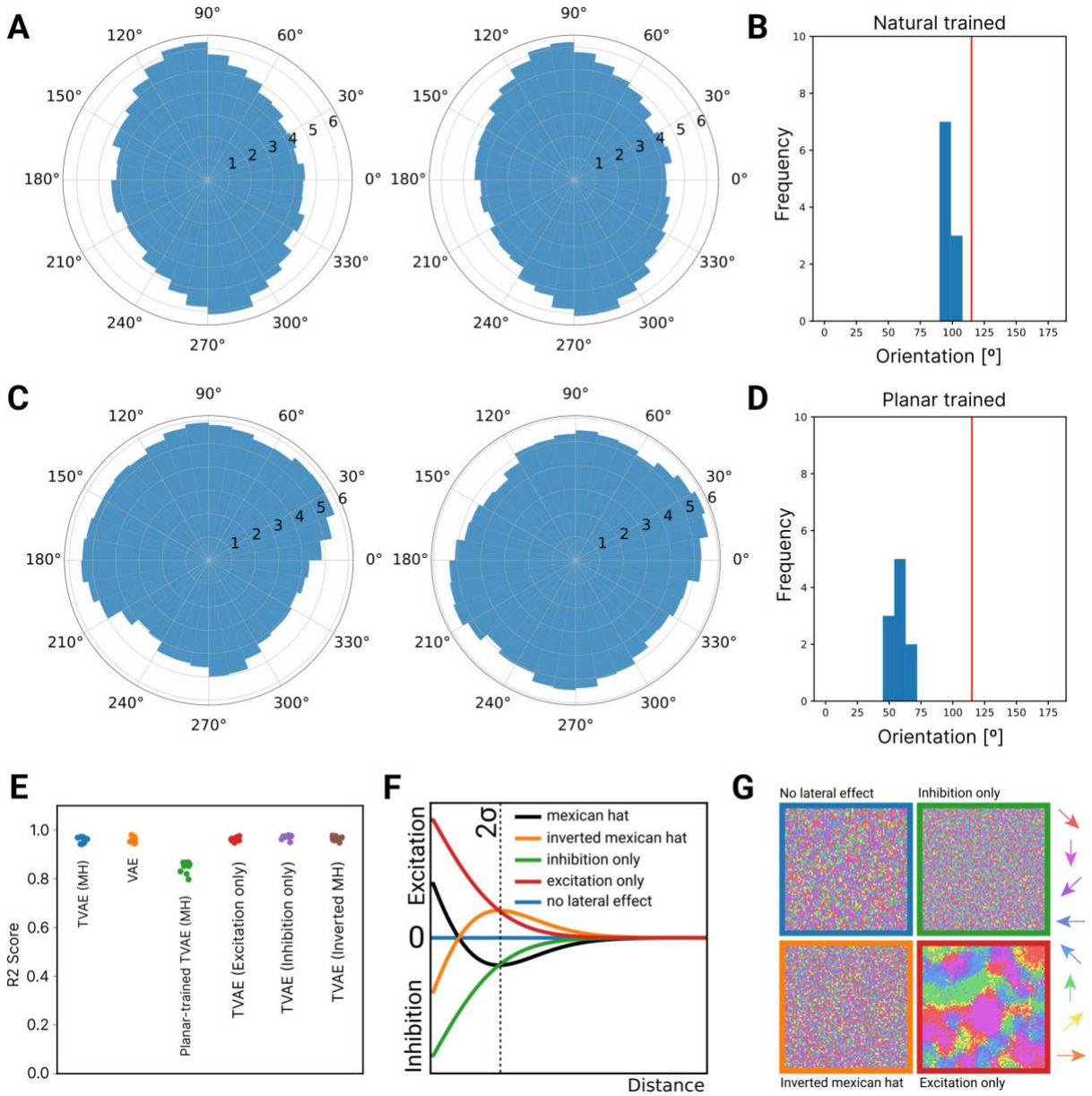


Figure 5. Lateral effects and natural behavioural data are essential for reproducing recorded neuron properties. ***(A)*** *Example PD distributions from neurons in topo-VAE models with different weight initialisations trained on natural movement data only (n=6400).* ***(B)*** *Orientations of the PD distribution modes for all 10 models.* ***(C)*** *Example PD distributions from neurons in topo-VAE models with different weight initialisations trained on planar movement data only (n=6400).* ***(D)*** *Orientations of the PD distribution modes across all models (n=10). Red lines indicate orientation of PD distribution in recorded neurons.* ***(E)*** *Reconstruction performance of different models (n=10), as measured by R^2 score.* ***(F)*** *Alternative lateral effect functions tested in control experiments, compared to the Mexican hat (black line), and* ***(G)*** *the resulting PD maps for each function.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We were able to make testable predictions about the representation of movements involving multiple joints (wrist, elbow, shoulder) across the neural population. To obtain a measure of how the 3 joints (Fig 6.A) are encoded across the cortical surface relative to each other, we found the average correlation with firing rate for all the degrees of freedom of each joint. All joints were evenly represented across the whole population (Fig 6.B), but no neurons correlated with only one joint. Many modelled neurons were correlated even with non-adjacent joints, e.g. elbow-wrist or even all three joints, possibly a result of the natural covariance patterns that exist between joints. For pairwise joint sensitivity comparisons see scatter plots (Fig. S3).

Fig. 6.C shows in more detail, the response strength of our cortical model neurons for individual degrees of freedom. These maps have a spatial pattern of ripple-like transitions between positive (white) and negative (black) correlations (Fig. 6.C), producing topographic clusters of neurons that are sensitive to particular joints (Fig. 6.D). The scale of these clusters approximately matches the that of the Mexican-hat lateral connection range. Regions of pure red, green or blue in Fig. 6D would correspond to neurons responsive to only shoulder, elbow or wrist respectively. Instead, colours that are the result of blending two or more joints predominate.

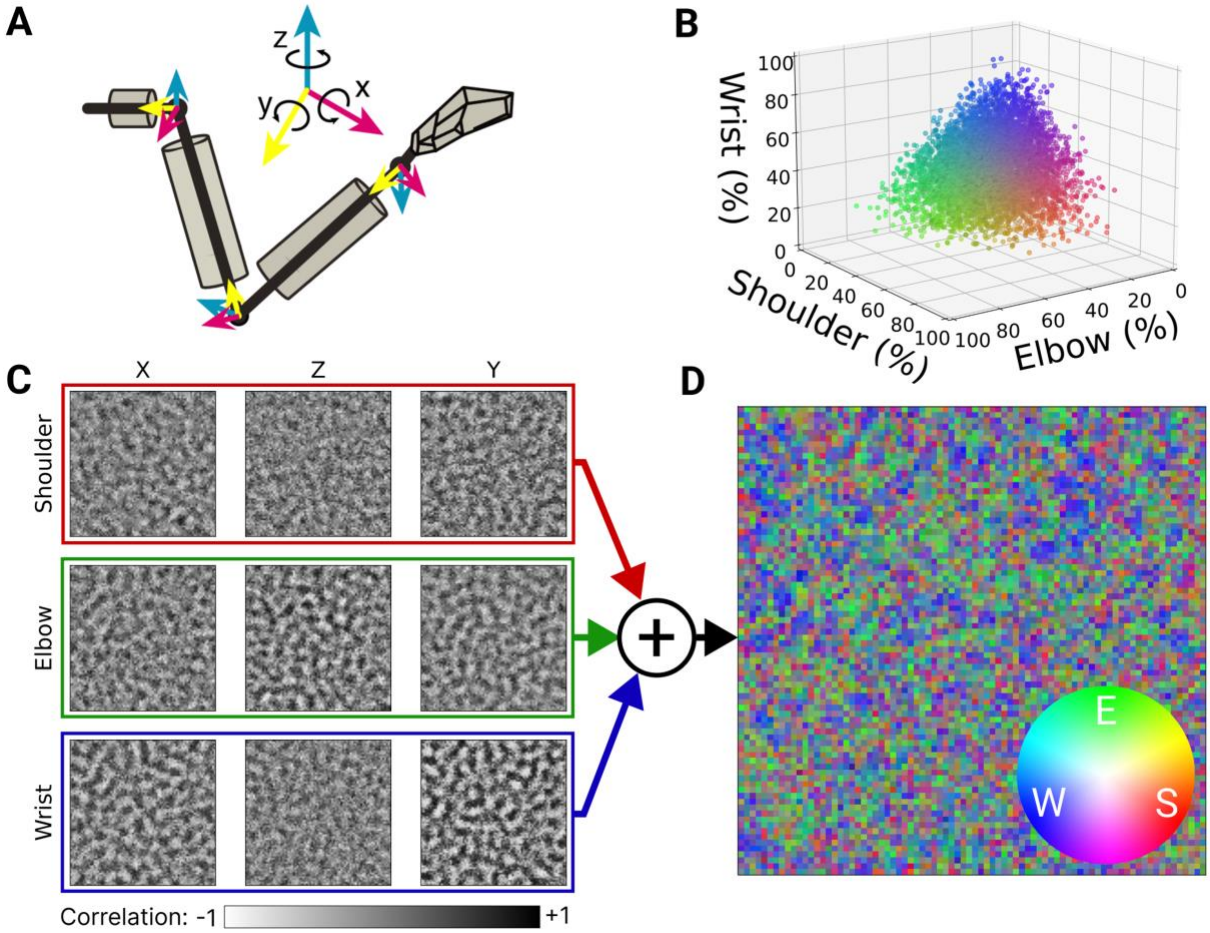


Figure 6. Sensitivity of latent layer neurons in topo-VAE to different input dimensions. (A) *Joint angle axes, 3 per joint.* ***(B)*** *3D plot showing the relative sensitivity of each neuron to joint inputs. Sensitivity is defined as the sum of correlations across X/Z/Y axes of all joints for a given neuron*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(colour normalised by maximum values per joint) (n=6400). ***(C)*** *Correlations between each dimension of input joint motion and neural response.* ***(D)*** *Joint sensitivity map where each pixel represents a neuron and the pixels RGB values (colour wheel inset) are reflecting the correlation with shoulder ("S"), elbow ("E"), and wrist ("W") angular velocity, respectively (colours normalised by maximum value across joints).*

Discussion

We built a model of neural coding and topographical organization of neurons in somatosensory cortex to help develop an understanding of proprioceptive coding, allowing us to paint a picture beyond that we can see through the peephole of neural recordings. Core to our approach was allowing the model to learn to represent the statistics of natural human movements, which are key to understanding sensory representations in the nervous system (Ejaz, Hamada, and Diedrichsen 2015; Laughlin 1981; Simoncelli and Olshausen 2001; Ganguli and Simoncelli 2016; 2014). We combine these natural proprioceptive statistics with a novel model that learns to generate movement-related neural activity.

Our topo-VAE is designed to reflect information processing in cortex: we hypothesise that the cortex learns, through experience, an efficient representation of the somatosensory world which we mimic here as a deep autoencoder network. The network learns a nonlinear mapping of its kinematic inputs into a "cortical" latent space, which embodies a stochastic, generative model. It performs a transformation from the high-dimensional proprioceptive inputs onto the two-dimensional cortical surface, thereby creating a population of neurons on a square grid. The topo-VAE includes a novel lateral interaction term between neighbouring neurons that shapes inputs into local neighbourhoods, allowing it to learn cortical-like spatial structure as well as single-neuron activity temporal patterns, which we compare to recorded neural data.

Machine learning methods with minimal assumptions built into their architecture can be used to reduce human inductive bias. For example, the topo-VAE knows nothing of the biomechanics of an arm or of scientific theories of the coding of planar hand-movements. Yet, it reproduces a range of experimental results, emergent properties of an information processing infrastructure trained on natural movement data and guided by three computational principles of information representation. By mimicking successive feed-forward stages or recurrent processing between regions, deep learning methods have been used to develop or confirm theories of the brain (Nayebi et al. 2018; Richards et al. 2019). As with all VAEs, the encoder in the topo-VAE learns latent stochastic variables that form statistically efficient representations of the input (Kingma and Welling 2014; Higgins et al. 2017; Eslami et al. 2016). The decoder portion of the VAE then attempts to linearly reconstruct the input variables from samples of the latent variables. VAEs have been used to learn both single neuron and population level features from spike train data (A. Wu et al. 2017; Speiser et al. 2017).

However, a vanilla VAE (with a Euclidean latent space) cannot induce a topographic mapping on the data, causing nearby input variables (stimuli) to be represented in arbitrary locations across the latent space. This phenomenon is known as manifold mismatch and a number of solutions have been sought mathematically (Davidson et al. 2018; Falorsi et al. 2018), yet none offers a biologically plausible neuroanatomical implementation, as does our topo-VAE. Our

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distance-dependent excitation and inhibition extension allowed us to link proprioceptive predictions not only to neural coding, but also to the distribution of neural activity across the cortex. A recent machine learning model with a similar name (the "topographic VAE"; (Keller and Welling 2021) has a more complex mathematical structure, but is focused on modelling sequential data, not lateral excitation and inhibition.

Our topo-VAE forms a spatially organised representation of proprioceptive inputs from which it generates spiking neuron outputs that are amenable to direct comparison with neural data. Previous self-organising maps (e.g. (Obermayer and Blasdel 1993; Aflalo and Graziano 2006) operate on less biologically plausible grounds. For example, Kohonen-type maps operate on winner-takes-all, non-spiking activation (and consequently, their training updates synaptic connections in a winner-takes-all form as well). Therefore, only one neuron can ever be active for given a sensory stimulus in a Kohonen map, a property very unlike actual proprioceptive cortex. In contrast, any number of neurons in our topo-VAE can be active simultaneously (and consequently, all synaptic connections are updated as sensory information is processed). Likewise, Poisson GLMs consider only the input statistics when producing simulated neural responses. While it is possible to reproduce the coding properties of individual neurons from convergence of peripheral inputs to a GLM, those models can tell us little about the role of neighbouring cells in shaping receptive fields or neuroanatomy-neural function relationships in general.

As a result, our topo-VAE model can make several predictions about the anatomical organisation of neural coding across the cortical surface that can be tested in a straightforward manner. First, our model neurons encoded combinations of joints, both adjacent (shoulder-elbow) and distant (shoulder-wrist). Such convergence in monkey somatosensory cortex has been observed for the hand (Costanzo and Gardner 1981; Warren, Santello, and Tillery 2011) and must be present to some extent in the proximal arm, given its multi-articular muscles. Our model used only joint-based inputs (i.e. the kinematic state or pose of the arm) but knows nothing about the musculoskeletal mechanics of the limb (e.g. the fact that bi-articulate muscles span multiple joints). Nonetheless, multi-joint coding emerged for most neurons in our model. These neurons predicted coding properties of neurons recorded from centre-out-reaching tasks data well. This is non-trivial, as the centre-out-reaching task is highly stereotyped, with heavily skewed joint correlation statistics. Crucially, when trained only on the stereotyped centre-out-data, the topo-VAE predicted that same data less well than when it was trained on the richer natural movement data.

The joint correlations we observe in the natural data are substantial (4 principal components explain 80% of the variance of the seven degrees of freedom of the arm) and are the result of three main factors, 1. the biomechanics of the body, including the way many muscles span multiple joints, 2. the way the brain controls movements and 3. the tasks performed. Arguably, task requirements in particular drive a substantial amount of the joint correlations. The statistics of the highly varied, robust set of natural movements we used to train the topo-VAE allowed it to generalise from these natural tasks, to humans and even monkeys doing planar centre-out movements. This same dataset used in an fMRI study of the representation of finger movements also found stable representation across subjects (Ejaz et al, 2015). They showed that the pairwise similarity of finger-specific activity patterns in the human sensorimotor

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cortex was well preserved across individuals, and this invariant organisation of movement activity was better explained by the correlation structure of everyday hand movements than it was by correlations in muscle activity. Therefore, the emergence of multi-joint proprioceptive receptive fields may represent yet to be investigated higher-order features of movements (Thomik, Fenske, and Faisal 2015) analogous to higher-order features of visual receptive fields, such as edges in V1 (Olshausen and Field 1997), that we will not understand without studying them in the context of natural stimuli, which for proprioception implies natural behaviour.

The joints of the body support a high-dimensional sensory space, representing multiple movement directions and sensory modalities, which must be compressed down onto a two-dimensional cortical surface. This problem has been famously resolved in the visual system by the discovery of pinwheels based on orientation selectivity of neighbouring neurons (Obermayer and Blasdel 1993). These may arise from the interaction of organised excitatory input from the periphery and isotropic Mexican hat interactions within cortex (Kang, Shelley, and Sompolinsky 2003). The breaks in the 2D data manifold of the cortical surface, necessitated by the dimensionality reduction occasionally cause very different stimuli to be encoded by nearby neurons. In our model, we find structure akin to the pinwheels of the visual cortex, here representing hand movement direction instead of orientation selectivity. While the relatively large spacing of electrodes used to record neural data from area 2 does not allow us to confirm the pinwheel anatomical structure directly, there are signatures of it in the recordings, such as the fact that neurons recorded on a given electrode tend to have more similar PDs than those recorded on separate electrodes – a necessary, but not sufficient property for proprioceptive pinwheels.

The well-known homunculus represents only the tactile component of somatosensation and its well-ordered map of the skin receptors. Since the human tactile homunculus has driven much of neuroscience's intuition about somatosensory representations, it is tempting to hypothesise that this is how proprioceptive representations might also be structured. However, proprioception is driven not by receptors embedded in a "simple" two-dimensional sheet, but rather by a set of dynamically quite different receptors in muscles that span one, two, and even three joints. The expectation that proprioception and touch might share a similar homunculus may not be reasonable.

So how should body and limb pose be represented in cortex, if not analogously to the tactile representation? When faced with the same problem, robotics engineers consider the configuration of each joint as a node in a graph, with limb segments between joints representing the edges (Teh et al. 2018; Farber 2008). In the kinematic hierarchy of this "proprioceptive representation", joints have neighbourhood relationships that form a tree-like graph with branches formed by the limbs.

The neurodevelopmental process, in which genes determine the differentiation of motor neuron pools, offers an alternate perspective. It appears that the topographic organisation of neurons in the spinal cord follow not the kinematic or muscular relationship of neighbouring joints but instead a grouping according to flexor and extensor muscles across the entire limb (Tsuchida et al. 1994), which one might expect to be transferred to the cortical level.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Emergent from our model were higher-level neural features similar to those observed in both S1 and M1, such as cosine directional tuning and movement speed modulation (Fig. 2&3). This coding was present in all neurons in the cortical layer and resulted in PD distributions with entropy values close to those of a uniform distribution. This result is consistent with the need for the brain to control reaching movements in all directions, but is a bit unexpected, given that musculoskeletal mechanics cause a strongly bimodal distribution of muscle PDs due to nonuniform distribution of muscle stretch (Versteeg, Chowdhury, and Miller 2021).

Because of its accessibility with Utah multielectrode arrays, we compared our modelled neurons to those recorded in area 2 of the somatosensory cortex, which combine cutaneous and muscle information. Area 3a, on the other hand, has only muscle-receptor inputs, and shares many features with the adjacent motor cortex. It is certainly intimately involved in movement execution. One might expect an even closer correspondence between our modelled cortex and area 3. Recent theory has pointed to the existence of a low-dimensional neural manifold in which motor cortices may operate to simplify the neural computations involved in controlling movement (Gallego et al. 2017; Churchland and Shenoy 2007; Churchland et al. 2012). These experiments and theory have focused primarily on the motor systems (although see (Stringer et al. 2019) for visual system examples) but may also be quite relevant for the proprioceptive system.

Perhaps one role of a topographic proprioceptive map (e.g. in area 3a) is to help translate peripheral feedback into the language of a low-dimensional neural motor manifold (Gallego et al. 2018). In principle, the projection of sensory information onto a 2-dimensional cortical map supports efficient wiring and short range access to multi-dimensional information (Chklovskii and Koulakov 2004), an architecture that we speculate may facilitate efficient learning and control of motor systems. Our topographic map is also compatible with the theory of optimal feedback control of movement (Scott 2004), with proprioceptive cortices being optimised to spatially transform their inputs into feedback control signals for motor cortices. This is because a 2D sensory state representation can be mapped efficiently (in terms of wiring geometry and length) onto a 2D motor representation. This could be achieved by linear operations using the synaptic connections between the sensory and motor cortices. Moreover, modulation of these sensorimotor connections by higher order areas could switch between and blend different feedback controllers, allow the efficient implementation and learning of complex control strategies.

A key assumption in the confirmation of our modelling results is that the proprioceptive organizational principles captured by our topo-VAE model from the kinematics of human movement would extend to monkeys, given their musculoskeletal similarity. Crucially, we are able to reproduce the coding properties of the neural recordings without a detailed biomechanical model, and with only a few computational principles, but only when training with natural behaviour joint kinematics in three dimensions (Fig 5.B,I). Planar centre-out kinematics were not adequate, even to represent those same centre-out movements. Our results are thus in line with other work showing that characterising neural activity in the context of natural behaviour may produce rich and unexpected results (Haar, van Assel, and Faisal 2020). This observation may be of great importance for the majority of proprioceptive and motor

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

neurophysiology experiments that have been conducted in highly constrained lab settings, settings that may not contain adequate ethologically relevant kinematic statistics to uncover the true coding of cortical neurons. Undertaking electrophysiological experiments with a broader repertoire of movements may affect proprioceptive neuroscience as much as the adoption of natural images did for understanding vision.

Methods

Human Behaviour Scenarios and Natural Movement Data

We recorded full-body movements from 18 healthy right-handed participants in two experimental scenarios. In the natural behaviour scenario (see Fig. 1.B & Supplemental Fig. 1), subjects performed unconstrained daily tasks in a working kitchen environment. As food preparation and feeding are universal behaviours, the only direction given to subjects was to prepare and eat an omelette. For this modelling work we used only the arm movement data (including wrist but not digits). The average recording time across subjects was 22 minutes. In the second movement scenario, a subject performed planar centre-out reaches in a 20x20 cm horizontal task space (see Fig. 1.C) to mimic the movement data for the monkey task. The horizontal task space was aligned 20 cm below the subject's shoulder and centred on the mid-line at 30 cm forward of the chest.

Arm movements in both scenarios were recorded at 60 Hz by an XSENS 3D motion tracking suit, a full-body sensor network based on inertial sensors. We used biomechanical models and fusion algorithms (including calibration and validation) to estimate joint angles. Fig. 1.A shows the biomechanical structure and coordinate system used in this paper. Arm movement datasets were formatted as time series of angles between segments following the International Society of Biomechanics (ISB) Euler angle extractions (G. Wu et al. 2005) in a ZXY coordinate. For the elbow and the wrist, angular rotations of Z, X and Y represent flexion/extension, abduction/adduction and internal/external rotation, respectively, since the biomechanics of the human body cannot be fully described by a rotation around a single axis and must instead be described with respect to 3 rotational axes. During planar movements, we used optical tracking as well as the motion tracking suit for capturing the end-point (hand) position on the task square. Data from the inertial and optical motion tracking systems were synchronised manually via cue-based movements before, during and after the recording period.

Variational Autoencoder with Topographic Latent Space

In the following we lay out the rationale for building our model and the model itself, the various forms of data we collected for model training and validation, and the validation methodology.

The topo-VAE model (cf. Fig. 1.D-G) uses the autoencoder framework to model sensory representations. Autoencoders are a type of artificial neural network used to learn efficient encodings of unlabelled data (Hinton and Salakhutdinov 2006). The neuroscientific mechanism is referred to as the Infomax principle, i.e. unsupervised learning to discover structure in the sensory data by maximising the match between the inputs (in our case, the somatosensory world) and their neural representation (Linsker 1988; Barlow 1961). We choose specifically a variational

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

autoencoder (Kingma and Welling 2014) because we want the latent cortical layer to be able to capture the stochasticity and variability inherent to neural representations (Orbán et al. 2016). We modelled the latent neurons as Bernoulli processes using the Gumbel-max trick (Gumbel 1954) where neuron firing is sampled from a categorial distribution of $\{0,1\}$. This forces the model to learn an encoding based on spike counts, which for a fixed-value input will approximate the Poisson distributed statistics of biological neurons. This allowed us to use the same data analysis pipeline as for our recorded neural data. However, this simple VAE model would be devoid of any spatial relationships between the neurons. Therefore, we added a simple organisational mechanism that would link learning between neighbouring latent neurons, effectively implementing cortical lateral connectivity with short range excitation and longer-range inhibition. Thus, our enhanced VAE learns both receptive field tuning properties (in an unsupervised manner through deep feature learning) and establishes a topographic relationship between neurons.

Let $[X] = \{\overrightarrow{x_n}\}_{n=1,\dots,N}, x_n \in \mathbb{R}^D$ be the sensory stimuli, following the natural behaviour distribution $p(\vec{x})$. A group of cortical neurons, arranged with a topographic structure, are activated by the sensory stimuli $[X]$ and generate firing patterns $[Z] = \{\overrightarrow{z_n}\}_{n=1,\dots,N}, z_n \in \mathbb{R}^{M\times M}, (M^2 \gg D)$. We aim to find a decoder $p_\Phi([X] \mid [Z])$ and corresponding neural responses $[Z]$ to optimally represent the sensory stimuli $[X]$, namely maximise the marginal likelihood of $p([X])$. The variational lower bound of the log likelihood $\log p(X)$ is derived as:

$$\begin{aligned} \log p([X]) &= log \int p([X]|[Z])p([Z])\, d\,[Z] \\ &\geq \mathbb{E}_{Z\sim q([Z])}[logp_\phi([X]|[Z]) - KL(q([Z])||p([Z])) \end{aligned} \tag{1}$$

where $q(Z)$ is the variational parameter approximating the intractable true posterior $p(Z|X)$.

Our topo-VAE encoder can be considered as a multi-layer neuronal structure delivering sensory stimuli to the cortex from sensory afferents, through brainstem nuclei, to the thalamus and cortex. This implies that at the encoder level we do not attribute or consider specific representations at these intermediate stages of proprioceptive processing (including how different sensory systems are integrated. From the perspective of computational modelling, it is also the amortised inference for inferring the optimal representation $q([Z])$, helping to avoid smoothness problems in over-complete representations (`back-constraint'). Fig. 1b illustrates the detailed structure of our topo-VAE, with a multi-layer perceptron (MLP) encoder and a linear decoder. The latent layer contains spiking neurons whose spike counts z_n within a given time interval Δt (i.e. over multiple input steps) follow a Poisson distribution. The encoder infers the distribution $q([Z])$ of responses for a group of cortical neurons, and z_n is sampled from this distribution using the Gumbel-max trick:

$$q(\vec{z}_n) = \underset{i}{\operatorname{argmax}}[g_i + \log \pi_i] \tag{2}$$

Where $\pi_i = [p \quad 1-p]$ is a vector defining the probability of a neuron spiking vs. not spiking, and p is set by the output of the encoder network for each neuron.

The decoder is a linear mapping from neural activities $[Z]$ to the reconstructions $[\hat{Z}]$ of the sensory stimuli. To summarise, the standard VAE model (which typically uses normally

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distributed random variables in the latent layer) is replaced by a latent layer of Bernoulli-type spike count distributions that are immediately applicable to neural signal analysis.

In addition to the log-likelihood function (Eq. 1) of the standard variational encoder, we include lateral effects in the latent layer. This was done by defining a distance-dependent function acting on the neurons, which are arranged in an $M \times M$ topographic map. A natural choice for cortical neurons is the Mexican-hat function, which transitions with distance from excitation to inhibition before vanishing (Amari 1977) (Fig. 1.F,G). As we are modelling an entire population, we can represent the interaction between neurons as a matrix $[\Psi]$. Each element $[\Psi_{p,q}]$ represents the lateral effect between neurons p and q, calculated as:

$$\Psi_{p,q} = (1 - \frac{d_{p,q}^2}{2\sigma^2})e^{-\frac{d_{p,q}^2}{2\sigma^2}} \tag{3}$$

where $d_{i,j}$ represents the Euclidean distance and σ is a hyperparameter defining the common length scale of local excitation and intermediate-range inhibition. As shown in Fig. 1.G, the transition from maximum excitation to maximum inhibition spans a distance of 2σ and the lateral effect vanishes at about 4σ.

The total loss function governing our topo-VAE model is given by:

$$\begin{aligned} L(\Phi,\Theta) = &-\mathbb{E}_{[Z]\sim q_\Phi([Z])}[log\, p_\Theta([X]|[Z])] + KL(q([Z]) \,||\, p([Z])) \\ &- \gamma \mathbb{E}_{Z\sim q_\Phi([Z])}[Tr([Z']^T[\Psi][Z'])] \end{aligned} \tag{4}$$

where $p([Z])$ is the prior distribution, set to be an independent Poisson distribution with rate r_p, $Z' = Z - z_b$. z_b is the base firing rate and γ is a constant controlling the impact of the lateral effects. Φ and Θ represent the trainable parameters in the encoder and the decoder, respectively. The KL divergence term in the loss function performs as a constraint on temporal sparsity, penalizing firing rates far from the expected rate r_p in the prior distribution $p([Z])$. r_p is usually set to be small (due to constraints such as metabolic cost (Attwell and Laughlin 2001) and allows us to naturally control the temporal sparsity of neural activity. In addition to the temporal sparsity, our topo-VAE also involves structured spatial sparsity. The lateral effect, as represented by the topographic term $\gamma \mathbb{E}_{[Z]\sim q_\Phi([Z])}[Tr([Z']^T[\Psi][Z'])]$ in Eq. 4 introduces topographic structure on the latent space and specifies the firing dependencies between neurons. Lateral excitation dominates the formation of pattern patterns in a structured space while lateral inhibition encourages spatial sparsity by penalising co-activation of non-nearby neurons. From the perspective of probabilistic inference, this regularisation item is equivalent to amending the prior distribution $p([Z])$ and modifies the target function as:

$$\begin{aligned} &L(\Phi,\Theta) = -\mathbb{E}_{[Z]\sim q_\Phi([Z])}[log\, p_\Theta([Z]|[Z])] + KL\big(q([Z]) \,||\, p^*([Z])\big) + const\,, \\ &p^*([Z]) = \frac{p([Z])e^{\gamma Tr([Z']^T[\Psi][Z'])}}{B}, \end{aligned} \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where the normalisation factor $B = \int p([Z]) e^{Tr\left(\gamma [Z']^T [\Psi][Z']\right)} dZ$ is a constant. In this form, both the firing sparsity and the lateral effect are expressed within the amended prior distribution $q^*([Z])$ and the target function is maintained in a standard variational autoencoder framework.

We demonstrate through ablation and parameter variation experiments the need for the specific design elements of our model to explain the biological data.

<u>Model parameter and design choices inspired by biological cortex</u>
The encoder component of the model contains two fully connected feed-forward layers of size 50 and 100 neurons, respectively, and a final linear readout layer. Tanh activation functions were used for all neurons in the first two layers. The reconstruction error of the model is measured using mean squared error.

To compare our modelling and recording results, we chose hyperparameters of the topographic map with consideration of 1) spatial densities of S1 neurons and 2) characteristics of the neural recording devices. The density of neurons in S1 is about 8M – 17M per cm^3 (Turner et al. 2016; Collins et al. 2016) of which 70% to 90% are pyramidal cells (Kaas 2006). The thickness of cortex varies from 1 mm to 4.5 mm (Fischl and Dale 2000; Wagstyl et al. 2015). The Utah electrode array has 100 microelectrodes arranged in a 10 x 10 configuration with 400 µm separation along each axis, thus spanning 3.6 mm x 3.6 mm of the cortex. We model neuronal anatomy as voxels or cubes, where the number of pyramidal cells contained within 1mm x 1mm x 1mm of cortex varies from 12,000 to 1,000,0000, which means every 1 mm along the cortical surface crosses about 20 - 50 pyramidal cells (see Fig. 1c). A surface-parallel slice captures a grid of 80 x 80 neurons. This conceptually simplified arrangement allows us to formulate a computationally tractable design of the latent layer neurons in topo-VAE. We can give a bit of intuition of our topographic model's parameters to neuroanatomy: the range of the lateral effect parameter σ translates to about 1-2 neuron spacing on our cortical model grid (assuming a radius of dendritic input to these cortical neurons of around 200 µm (Braitenberg and Schüz 2013)). Its precise value was determined pot-hoc using model selection by numerically sweeping for a range of σ and selecting the best fitting value.

To observe the effect of the neighbourhood range parameter σ on topography in the latent representation, we test values σ={1,2,3}. For the expected firing rate *p(x)*, we find stable topography around *p(x)*=0.01. The training loss function contains multiple components which can be given individual weightings. Here we use weightings of 10, 0.4, and 0.005 for the reconstruction, lateral effect, and firing sparsity components, respectively. In addition, we include an L2 norm on the model weights β=0.1. When sampling from the latent space during decoding, we parameterise the scaling of the rate parameter, to simulate sampling across different time windows. Varying this parameter in the range [0,1000] yielded optimal reconstruction at a scaling factor of 40.

Model Training & Validation overview

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The topo-VAE model was implemented in Python (Van Rossum et al. 1995) using PyTorch (Paszke et al. 2019) and run on a GPU workstation. All models were trained for 4000 epochs using the Adam optimiser with a learning rate of 10^{-5} and a batch size of 400.

To train our topo-VAE we needed to pre-process the input data. Empirical distribution of joint angular velocities during movements in our tasks is symmetric, unimodal, with sharp peaks at zero and heavy tails towards large speeds. To ensure equal contributions to model optimisation, these features are standardised without mean centring (since they are already approximately zero-centred, and to preserve the correspondence of a zero-valued input to zero joint angle velocity). This pre-processing also improves the training efficiency and convergence of our model without breaking the spatial structure of natural movements. In addition, since the original kinematic data were sampled at a frequency of 60Hz, we subsampled the training data at a depth of 1%, to remove redundancy and improve training times, without significant effects on the outcome of the model.

Optimisation of VAEs by backpropagation requires estimation of parameter gradients with respect to loss. Since gradients cannot be estimated for the Gumbel-max function, we implemented the Gumbel-softmax reparameterisation trick (Jang et al. 2015) during training, where *z* is sampled from *q(z)* using the Gumbel-softmax function:

$$q(\vec{z}_n) = \frac{\exp(\frac{\log(\pi_i) + g_i}{\tau})}{\sum_{j=1}^{2} \exp(\frac{\log(\pi_j) + g_j}{\tau})} \tag{2}$$

And g_i is a random sample from the Gumbel distribution $G(0,1)$:

$$P(x) = e^{-(x+e^{-x})} \tag{3}$$

We wanted to perform feature learning of proprioceptive representations with our topo-VAE and this required large amounts of natural movement data. Therefore, we performed human experiments (see above) and measured the full-body kinematics of human subjects. We use human full-body behaviour as a functional proxy for non-human primate movements of the arm (Young, Wagner, and Hallgrímsson 2010), as human data is much easier and more precisely obtainable. We used the human data to drive the training of our topo-VAE, then froze our model parameters to evaluate it. Our topo-VAE model is generative, so by playing back any limb movement data (time series of body poses) we obtain spike trains for each neuron in our cortical grid.

To compare our model's predictions to those of recorded neurons, we use the kinematic data from humans performing the same centre-out task as the monkeys (see above) to drive the frozen topo-VAE model and compare its output to the actual recorded neural data.

Model Robustness and Parameter Variation

The topographic property of the VAE arises from the lateral effect component of the loss function (Eq. 4). As a control, we tested our model with no lateral effects (Fig. 5.A-D), and with several

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

different distance functions (Mexican hat, inverted Mexican hat, excitation only, inhibition only; Fig. 5.F,G). The inverted Mexican hat lateral effect is the additive inverse of the Mexican hat function (Eq. 3). Excitation-only lateral effect is defined as:

$$\Psi_{p,q} = e^{-\frac{d_{p,q}^2}{2\sigma^2}} \tag{6}$$

and the inhibition-only lateral effect is the additive inverse of the excitation-only lateral effect (Eq. 6).

Model Analysis

To quantify the sensitivity of neurons to specific joints (Fig. 6), we individually perturb each input feature of the model and measure the Pearson correlation between individual neurons in the latent space and that feature. Since each joint is represented by three input features (angular velocity in the Z, X, and Y axes), we use the mean of the absolute correlation across all three axes to quantify the sensitivity of a given neuron to a particular joint.

We computed angular velocity profiles from the recorded data and analysed the natural movement dataset X_{nat} and the planar movement dataset X_{pl}. Supplemental Fig. 3.A illustrates the principal component analysis (PCA) of angular velocities in the recorded planar movement and natural movement datasets. The first two PCs of the planar movements explained over 95% of the total variance, but only half of the variance of the natural movements. This reveals that joint velocities of planar movements are highly constrained, as expected, restricted largely to a 2-dimensional subspace. We applied the manipulative complexity metric (Belić and Faisal 2015) to quantify the complexity of the movements, which is defined as:

$$C = 1 - \frac{2}{D-1}\sum_{j=1}^{D}\sum_{i=1}^{j}\left(VAF_i - \frac{1}{D}\right)$$

Where VAF_i is the variance captured by the ith PC. Larger values of C indicate higher complexity and a value of $C = 1$ means that all PCs contribute equally to the total variance. The complexity of planar movements was 0.06, much lower than the natural movement complexity (C = 0.5). We then show a series of planar movements (colour-coded with respect to movement direction) in the end-point velocity space (Supplemental Fig. 3.B) and the angular velocity subspace spanned by the first 2 PCs (Supplemental Fig. 3.C). We defined the direction θ and speed v of planar movements in world coordinates: 90°/270° are respectively away from and towards the chest, 180°/0° are to the left and right.

Fig. 1.H illustrates our use of the two human datasets with our computational model. The natural movement dataset X_{nat} can be viewed as a group of samples generated from the natural movement statistic. Following the idea of natural sensory coding (Olshausen and Field 1997), we used this natural movement dataset as training data to learn an optimal neural coding scheme. After training, we tested the converged model with data from the planar reaching task. To compare with hand-based coding properties of area 2 neurons (Prud'homme and Kalaska 1994; Chowdhury, Glaser, and Miller 2020), we found a linear mapping between joint angular velocities and planar hand

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

velocity. This allowed us to assess the relationship between hand movement direction and topo-VAE firing rates.

Nonhuman Primate Behaviour and Data Collection

We used a combination of previously recorded data in which three rhesus macaques performed a planar, centre-out reaching task while seated, using a two-link planar manipulandum. A cursor displayed on a monitor tracked the position of the manipulandum and provided visual feedback for the monkey as he reached for a target on-screen. The monkey moved the cursor to a central target in the workspace. After a random delay period, 1 of 8 targets spaced evenly in a circle around the central target appeared on the screen and the monkey moved the cursor toward it upon an audible 'go' cue. After placing the cursor in the target for a random hold time of 0-500 ms, the monkey received a liquid reward and returned the cursor to the central target. We used 6 experimental sessions across three monkeys; two contain data that has been previously published (Chowdhury, Glaser, and Miller 2020), and the rest are unpublished. All procedures were in accordance with the Guide for the Care and Use of Laboratory Animals, and were approved by the institutional animal care and use committee of Northwestern University under protocol #IS00000367.

Once a monkey was trained on the experimental apparatus, a 96-channel microelectrode array with 1 mm iridium-oxide coated electrodes (Blackrock Microsystems, Inc.) was pneumatically inserted in Brodmann's area 2, near the intraparietal sulcus (Chowdhury, Glaser, and Miller 2020). The implantation site was chosen to avoid cerebral vasculature and maximise proximal arm representation. All surgery was performed under isoflurane gas anaesthesia (1-2 percent) except during intraoperative recording to identify the arm representations, when the monkey was transitioned to a mixture of <0.5% isoflurane and remifentanil (0.4 ug/kg/min).

The data were recorded from the microelectrode array using the Cerebus multichannel data recording system (Blackrock Microsystems, Inc.). Thresholded waveforms and timing of behavioural task events were synchronised and recorded for offline analyses. The position of the handle was recorded at 1kHz. We discriminated single neurons using Offline Sorter (Plexon, Inc., Dallas TX).

To calculate the preferred direction of a neuron, we used a simple bootstrapping procedure. For each iteration, we drew random points from the dataset conforming to a uniform distribution of movement directions. We then fit Poisson generalised linear models (GLM) with angular velocity inputs to the firing rate for the sampled timepoints. The GLM models are defined by:

$$[\hat{f}] = Poisson(\lambda), \qquad \lambda = e^{[X][\beta]}$$

where $[\hat{f}]$ is a T (number of time points) by N (number of neurons) matrix of firing rate estimates of the recorded rates $[f]$, X is a T by 2 (number of velocity inputs) matrix, and β is a P by N matrix of encoding parameters. β was found using maximum likelihood estimation. The preferred direction was then calculated from the encoding vector of the GLM as:

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$PD_i = \tan^{-1}(\beta_y, \beta_x), \qquad r_i = \sqrt{\beta_y^2, \beta_x^2}$$

In these equations, a bootstrap PD estimate of neuron i was defined by β_y and β_x, the bootstrap encoding parameters for hand velocity in the y and x directions, respectively. We took the circular mean of the PD estimates over all bootstrap iterations to find the PD for each neuron. This method of PD calculation was used for both recorded and modelled neurons.

Author contributions

AAF conceived the study. LEM & AAF supervised the study. YW, AAF developed the theory and model. JAH, AAF collected the behavioural data. KPB collected the unpublished neural data. YW, MG implemented the model code. YW, MG implemented the behavioural data analysis code. KPB implemented the neural data analysis code. KPB, MG, YW, AAF, LEM analysed the data. KPB, LEM, AAF drafted the manuscript. All authors discussed the results and reviewed the manuscript.

Funders

We acknowledge: KPB was supported by NIH BRAIN NRSA Postdoctoral Fellowship F32MH120893 and [S1R01]. YW was supported by an NIHR Imperial College BRC Deep Phenotyping Grant and Project eNHANCE (http://www.enhance-motion.eu) under the European Union's Horizon2020 research and innovation programme (Grant No. 644000). MG was supported by the Wellcome Trust PhD Program "Bioinformatics & Theoretical Systems Biology" (222888/Z/21/Z). JAH was supported by an EPSRC Doctoral Training Award. LEM was supported by NINDS grant #R01NS095251. AAF acknowledges his UKRI Turing AI Fellowship (EP/V025449/1).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Financial Disclosures

KB, YW, JAH, MG, LEM, AAF have no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Aflalo, T. N., and M. S. A. Graziano. 2006. "Possible Origins of the Complex Topographic Organization of Motor Cortex: Reduction of a Multidimensional Space onto a Two-Dimensional Array." *Journal of Neuroscience* 26 (23): 6288–97. https://doi.org/10.1523/JNEUROSCI.0768-06.2006.

Amari, Shun-ichi. 1977. "Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields." *Biological Cybernetics* 27 (2): 77–87. https://doi.org/10.1007/BF00337259.

Attwell, David, and Simon B. Laughlin. 2001. "An Energy Budget for Signaling in the Grey Matter of the Brain." *Journal of Cerebral Blood Flow & Metabolism* 21 (10): 1133–45. https://doi.org/10.1097/00004647-200110000-00001.

Barlow, H. B. 1961. "Possible Principles Underlying the Transformations of Sensory Messages." In *Sensory Communication*, edited by Walter A. Rosenblith, 216–34. The MIT Press. https://doi.org/10.7551/mitpress/9780262518420.003.0013.

Belić, Jovana J., and A. Aldo Faisal. 2015. "Decoding of Human Hand Actions to Handle Missing Limbs in Neuroprosthetics." *Frontiers in Computational Neuroscience* 9: 27. https://doi.org/10.3389/fncom.2015.00027.

Blum, Kyle P., Christopher Versteeg, Joseph Sombeck, Raeed H. Chowdhury, and Lee E. Miller. 2021. "Proprioception: A Sense to Facilitate Action." In *Somatosensory Feedback for Neuroprosthetics*. Elsevier. https://doi.org/10.1016/B978-0-12-822828-9.00017-4.

Braitenberg, Valentino, and Almut Schüz. 2013. *Cortex: Statistics and Geometry of Neuronal Connectivity*. Springer Science & Business Media.

Chaudhuri, Rishidev, Berk Gerçek, Biraj Pandey, Adrien Peyrache, and Ila Fiete. 2019. "The Intrinsic Attractor Manifold and Population Dynamics of a Canonical Cognitive Circuit across Waking and Sleep." *Nature Neuroscience* 22 (9): 1512–20. https://doi.org/10.1038/s41593-019-0460-x.

Chklovskii, Dmitri B., and Alexei A. Koulakov. 2004. "MAPS IN THE BRAIN: What Can We Learn from Them?" *Annual Review of Neuroscience* 27 (1): 369–92. https://doi.org/10.1146/annurev.neuro.27.070203.144226.

Chowdhury, Raeed H, Joshua I Glaser, and Lee E Miller. 2020. "Area 2 of Primary Somatosensory Cortex Encodes Kinematics of the Whole Arm." Edited by Tamar R Makin, Joshua I Gold, and Tamar R Makin. *ELife* 9 (January): e48198. https://doi.org/10.7554/eLife.48198.

Churchland, Mark M., John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. 2012. "Neural Population Dynamics during Reaching." *Nature* 487 (7405): 51–56. https://doi.org/10.1038/nature11129.

Churchland, Mark M., and Krishna V. Shenoy. 2007. "Temporal Complexity and Heterogeneity of Single-Neuron Activity in Premotor and Motor Cortex." *Journal of Neurophysiology* 97 (6): 4235–57. https://doi.org/10.1152/jn.00095.2007.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Collins, Christine E., Emily C. Turner, Eva Kille Sawyer, Jamie L. Reed, Nicole A. Young, David K. Flaherty, and Jon H. Kaas. 2016. "Cortical Cell and Neuron Density Estimates in One Chimpanzee Hemisphere." *Proceedings of the National Academy of Sciences* 113 (3): 740–45. https://doi.org/10.1073/pnas.1524208113.

Costanzo, Richard M, and Esther P Gardner. 1981. "Multiple-Joint Neurons in Somatosensory Cortex of Awake Monkeys." *Brain Research*.

Davidson, Tim R., Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. 2018. "Hyperspherical Variational Auto-Encoders." *ArXiv:1804.00891 [Cs, Stat]*, September. http://arxiv.org/abs/1804.00891.

Ejaz, Naveed, Masashi Hamada, and Jörn Diedrichsen. 2015. "Hand Use Predicts the Structure of Representations in Sensorimotor Cortex." *Nature Neuroscience* 18 (7): 1034–40. https://doi.org/10.1038/nn.4038.

Eslami, S. M. Ali, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. 2016. "Attend, Infer, Repeat: Fast Scene Understanding with Generative Models." *Advances in Neural Information Processing Systems* 29. https://proceedings.neurips.cc/paper/2016/hash/52947e0ade57a09e4a1386d08f17b656-Abstract.html.

Faisal, A. Aldo. 2021. "Putting Touch into Action." *Science* 372 (6544): 791–92. https://doi.org/10.1126/science.abi7262.

Falorsi, Luca, Pim de Haan, Tim R. Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S. Cohen. 2018. "Explorations in Homeomorphic Variational Auto-Encoding." *ArXiv:1807.04689 [Cs, Stat]*, July. http://arxiv.org/abs/1807.04689.

Farber, Michael. 2008. *Invitation to Topological Robotics*. European Mathematical Society.

Fischl, Bruce, and Anders M. Dale. 2000. "Measuring the Thickness of the Human Cerebral Cortex from Magnetic Resonance Images." *Proceedings of the National Academy of Sciences* 97 (20): 11050–55.

Flesher, Sharlene N., John E. Downey, Jeffrey M. Weiss, Christopher L. Hughes, Angelica J. Herrera, Elizabeth C. Tyler-Kabara, Michael L. Boninger, Jennifer L. Collinger, and Robert A. Gaunt. 2021. "A Brain-Computer Interface That Evokes Tactile Sensations Improves Robotic Arm Control." *Science* 372 (6544): 831–36. https://doi.org/10.1126/science.abd0380.

Formento, Emanuele, Karen Minassian, Fabien Wagner, Jean-Baptiste Mignardot, Camille G. Le Goff-Mignardot, Andreas Rowald, Jocelyne Bloch, Silvestro Micera, Marco Capogrosso, and Grégoire Courtine. 2018. "Electrical Spinal Cord Stimulation Must Preserve Proprioception to Enable Locomotion in Humans with Spinal Cord Injury." *Nature Neuroscience*, November, 1–21. https://doi.org/10.1038/s41593-018-0262-6.

Gallego, Juan A., Matthew G. Perich, Lee E. Miller, and Sara A. Solla. 2017. "Neural Manifolds for the Control of Movement." *Neuron* 94 (5): 978–84. https://doi.org/10.1016/j.neuron.2017.05.025.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gallego, Juan A., Matthew G. Perich, Stephanie N. Naufel, Christian Ethier, Sara A. Solla, and Lee E. Miller. 2018. "Cortical Population Activity within a Preserved Neural Manifold Underlies Multiple Motor Behaviors." *Nature Communications* 9 (1): 4233. https://doi.org/10.1038/s41467-018-06560-z.

Ganguli, Deep, and Eero P. Simoncelli. 2014. "Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations." *Neural Computation* 26 (10): 2103–34. https://doi.org/10.1162/NECO_a_00638.

Ganguli, Deep, and Eero P. Simoncelli. 2016. "Neural and Perceptual Signatures of Efficient Sensory Coding." *ArXiv:1603.00058 [q-Bio]*, February. http://arxiv.org/abs/1603.00058.

Gumbel, E.J., 1954. *Statistical theory of extreme values and some practical applications: a series of lectures* (Vol. 33). US Government Printing Office.

Haar, Shlomi, Camille M. van Assel, and A. Aldo Faisal. 2020. "Motor Learning in Real-World Pool Billiards." *Scientific Reports* 10 (1): 20046. https://doi.org/10.1038/s41598-020-76805-9.

Higgins, Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. "β-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK," 22.

Hinton, G. E., and R. R. Salakhutdinov. 2006. "Reducing the Dimensionality of Data with Neural Networks." *Science* 313 (5786): 504–7. https://doi.org/10.1126/science.1127647.

Huffman, K. J., and L. Krubitzer. 2001. "Area 3a: Topographic Organization and Cortical Connections in Marmoset Monkeys." *Cerebral Cortex (New York, N.Y. : 1991)* 11 (9): 849–67.

Jang, E., Gu, S. and Poole, B., 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kaas, Jon H. 2006. "Evolution of the Neocortex." *Current Biology* 16 (21): R910–14. https://doi.org/10.1016/j.cub.2006.09.057.

Kang, Kukjin, Michael Shelley, and Haim Sompolinsky. 2003. "Mexican Hats and Pinwheels in Visual Cortex." *Proceedings of the National Academy of Sciences* 100 (5): 2848–53.

Keller, T Anderson, and Max Welling. 2021. "Topographic VAEs Learn Equivariant Capsules." *ArXiv:2109.01394*, 27.

Killebrew, Justin H., Sliman J. Bensmaïa, John F. Dammann, Peter Denchev, Steven S. Hsiao, James C. Craig, and Kenneth O. Johnson. 2007. "A Dense Array Stimulator to Generate Arbitrary Spatio-Temporal Tactile Stimuli." *Journal of Neuroscience Methods* 161 (1): 62–74. https://doi.org/10.1016/j.jneumeth.2006.10.012.

Kingma, Diederik P., and Max Welling. 2014. "Auto-Encoding Variational Bayes." *ArXiv:1312.6114 [Cs, Stat]*, May. http://arxiv.org/abs/1312.6114.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Laughlin, Simon. 1981. "A Simple Coding Procedure Enhances a Neuron's Information Capacity." *Zeitschrift Für Naturforschung C* 36 (9–10): 910–12. https://doi.org/10.1515/znc-1981-9-1040.

Linsker, R. 1988. "Self-Organization in a Perceptual Network." *Computer* 21 (3): 105–17. https://doi.org/10.1109/2.36.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, Mass: MIT Press.

Nayebi, Aran, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J. DiCarlo, and Daniel L. K. Yamins. 2018. "Task-Driven Convolutional Recurrent Models of the Visual System." *ArXiv:1807.00053 [Cs, q-Bio]*, October. http://arxiv.org/abs/1807.00053.

Obermayer, K., and G. G. Blasdel. 1993. "Geometry of Orientation and Ocular Dominance Columns in Monkey Striate Cortex." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 13 (10): 4114–29.

Olshausen, Bruno A., and David J. Field. 1997. "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?" *Vision Research* 37 (23): 3311–25. https://doi.org/10.1016/S0042-6989(97)00169-7.

Orbán, Gergő, Pietro Berkes, József Fiser, and Máté Lengyel. 2016. "Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex." *Neuron* 92 (2): 530–43. https://doi.org/10.1016/j.neuron.2016.09.038.

Pehlevan, Cengiz, Alexander Genkin, and Dmitri B. Chklovskii. 2017. "A Clustering Neural Network Model of Insect Olfaction." In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, 593–600. https://doi.org/10.1109/ACSSC.2017.8335410.

Penfield, Wilder, and Edwin Boldrey. 1937. "Somatic Motor and Sensory Representation in the Cerebral Cortex of Man as Studied by Electrical Stimulation." *Brain* 60 (4): 389–443. https://doi.org/10.1093/brain/60.4.389.

Prud'homme, M. J., and J. F. Kalaska. 1994. "Proprioceptive Activity in Primate Primary Somatosensory Cortex during Active Arm Reaching Movements." *Journal of Neurophysiology* 72 (5): 2280–2301. https://doi.org/10.1152/jn.1994.72.5.2280.

Richards, Blake A., Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, et al. 2019. "A Deep Learning Framework for Neuroscience." *Nature Neuroscience* 22 (11): 1761–70. https://doi.org/10.1038/s41593-019-0520-2.

Rossi, L. Federico, Kenneth D. Harris, and Matteo Carandini. 2020. "Spatial Connectivity Matches Direction Selectivity in Visual Cortex." *Nature* 588 (7839): 648–52. https://doi.org/10.1038/s41586-020-2894-4.

Sainburg, R. L., H. Poizner, and C. Ghez. 1993. "Loss of Proprioception Produces Deficits in Interjoint Coordination." *Journal of Neurophysiology* 70 (5): 2136–47. https://doi.org/10.1152/jn.1993.70.5.2136.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sandbrink, Kai J., Pranav Mamidanna, Claudio Michaelis, Mackenzie Weygandt Mathis, Matthias Bethge, and Alexander Mathis. 2020. "Task-Driven Hierarchical Deep Neural Network Models of the Proprioceptive Pathway." *BioRxiv*, May, 2020.05.06.081372. https://doi.org/10.1101/2020.05.06.081372.

Scott, Stephen H. 2004. "Optimal Feedback Control and the Neural Basis of Volitional Motor Control." *Nature Reviews Neuroscience* 5 (7): 532–45. https://doi.org/10.1038/nrn1427.

Simoncelli, Eero P, and Bruno A Olshausen. 2001. "Natural Image Statistics and Neural Representation." *Annual Review of Neuroscience* 24 (1): 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193.

Speiser, Artur, Jinyao Yan, Evan Archer, Lars Buesing, Srinivas C. Turaga, and Jakob H. Macke. 2017. "Fast Amortized Inference of Neural Activity from Calcium Imaging Data with Variational Autoencoders." *ArXiv:1711.01846 [Cs, q-Bio, Stat]*, November. http://arxiv.org/abs/1711.01846.

Sterling, Peter, and Simon Laughlin. 2015. *Principles of Neural Design*. MIT Press.

Stringer, Carsen, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. 2019. "High-Dimensional Geometry of Population Responses in Visual Cortex." *Nature* 571 (7765): 361–65. https://doi.org/10.1038/s41586-019-1346-5.

Teh, Thomas, Chaiyawan Auepanwiriyakul, John Alexander Harston, and A. Aldo Faisal. 2018. "Generalised Structural CNNs (SCNNs) for Time Series Data with Arbitrary Graph Topology." *ArXiv:1803.05419 [Cs, Stat]*, May. http://arxiv.org/abs/1803.05419.

Thomik, Andreas A. C., Sonja Fenske, and A. Aldo Faisal. 2015. "Towards Sparse Coding of Natural Movements for Neuroprosthetics and Brain-Machine Interfaces." In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, 938–41. https://doi.org/10.1109/NER.2015.7146780.

Todorov, Emanuel, and Michael I. Jordan. 2002. "Optimal Feedback Control as a Theory of Motor Coordination." *Nature Neuroscience* 5 (11): 1226–35. https://doi.org/10.1038/nn963.

Tsuchida, T., M. Ensini, S. B. Morton, M. Baldassare, T. Edlund, T. M. Jessell, and S. L. Pfaff. 1994. "Topographic Organization of Embryonic Motor Neurons Defined by Expression of LIM Homeobox Genes." *Cell* 79 (6): 957–70. https://doi.org/10.1016/0092-8674(94)90027-2.

Turner, Emily C., Nicole A. Young, Jamie L. Reed, Christine E. Collins, David K. Flaherty, Mariana Gabi, and Jon H. Kaas. 2016. "Distributions of Cells and Neurons across the Cortical Sheet in Old World Macaques." *Brain, Behavior and Evolution* 88 (1): 1–13. https://doi.org/10.1159/000446762.

Tuthill, John C., and Eiman Azim. 2018. "Proprioception." *Current Biology* 28 (5): R194–203. https://doi.org/10.1016/j.cub.2018.01.064.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.10.472161; this version posted August 27, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Versteeg, Christopher, Raeed H Chowdhury, and Lee E Miller. 2021. "Cuneate Nucleus: The Somatosensory Gateway to the Brain." *Current Opinion in Physiology* 20 (April): 206–15. https://doi.org/10.1016/j.cophys.2021.02.004.

Wagstyl, Konrad, Lisa Ronan, Ian M. Goodyer, and Paul C. Fletcher. 2015. "Cortical Thickness Gradients in Structural Hierarchies." *NeuroImage* 111 (May): 241–50. https://doi.org/10.1016/j.neuroimage.2015.02.036.

Warren, Jay P., Marco Santello, and Stephen I. Helms Tillery. 2011. "Effects of Fusion between Tactile and Proprioceptive Inputs on Tactile Perception." *PLOS ONE* 6 (3): e18073. https://doi.org/10.1371/journal.pone.0018073.

Weber, Douglas J., Brian M. London, James A. Hokanson, Christopher A. Ayers, Robert A. Gaunt, Ricardo R. Torres, Boubker Zaaimi, and Lee E. Miller. 2011. "Limb-State Information Encoded by Peripheral and Central Somatosensory Neurons: Implications for an Afferent Interface." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 19 (5): 501–13. https://doi.org/10.1109/TNSRE.2011.2163145.

Willmore, B., and D. J. Tolhurst. 2001. "Characterizing the Sparseness of Neural Codes." *Network: Computation in Neural Systems* 12 (3): 255–70. https://doi.org/10.1088/0954-898X/12/3/302.

Wu, Anqi, Nicholas G Roy, Stephen Keeley, and Jonathan W Pillow. 2017. "Gaussian Process Based Nonlinear Latent Structure Discovery in Multivariate Spike Train Data," 10.

Wu, Ge, Frans C. T. van der Helm, H. E. J. (DirkJan) Veeger, Mohsen Makhsous, Peter Van Roy, Carolyn Anglin, Jochem Nagels, et al. 2005. "ISB Recommendation on Definitions of Joint Coordinate Systems of Various Joints for the Reporting of Human Joint Motion—Part II: Shoulder, Elbow, Wrist and Hand." *Journal of Biomechanics* 38 (5): 981–92. https://doi.org/10.1016/j.jbiomech.2004.05.042.

Young, Nathan M., Günter P. Wagner, and Benedikt Hallgrímsson. 2010. "Development and the Evolvability of Human Limbs." *Proceedings of the National Academy of Sciences* 107 (8): 3400–3405. https://doi.org/10.1073/pnas.0911856107.