CellPress
OPEN ACCESS

Cell
Leading Edge

Perspective

Decoding the brain: From neural representations to mechanistic models

Mackenzie Weygandt Mathis,[1,2,*] Adriana Perez Rotondo,[1,2] Edward F. Chang,[3] Andreas S. Tolias,[4,5,6,7] and Alexander Mathis[1,2]

[1]Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland
[2]Neuro-X Institute, École Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland
[3]Department of Neurological Surgery, UCSF, San Francisco, CA, USA
[4]Department of Ophthalmology, Byers Eye Institute, Stanford University, Stanford, CA, USA
[5]Department of Electrical Engineering, Stanford University, Stanford, CA, USA
[6]Stanford BioX, Stanford University, Stanford, CA, USA
[7]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA
*Correspondence: mackenzie.mathis@epfl.ch
https://doi.org/10.1016/j.cell.2024.08.051

SUMMARY

A central principle in neuroscience is that neurons within the brain act in concert to produce perception, cognition, and adaptive behavior. Neurons are organized into specialized brain areas, dedicated to different functions to varying extents, and their function relies on distributed circuits to continuously encode relevant environmental and body-state features, enabling other areas to decode (interpret) these representations for computing meaningful decisions and executing precise movements. Thus, the distributed brain can be thought of as a series of computations that act to encode and decode information. In this perspective, we detail important concepts of neural encoding and decoding and highlight the mathematical tools used to measure them, including deep learning methods. We provide case studies where decoding concepts enable foundational and translational science in motor, visual, and language processing.

INTRODUCTION

Imagine sitting at a piano reading the sheet music, taking in the dark notes on the crisp white page of Rachmaninoff's Piano Concerto No. 2. You have a mental (internal) model of how the piano works; thus, the task for your brain is how to translate the notes into motor actions. The tactile sensation of the keys provides critical proprioceptive feedback, and your auditory senses are heightened as they immerse themselves in the melodic contours and harmonic progressions, listening intently to each note as you play and hum along. From this multi-sensory input, you constantly refine your next movements that ultimately create the rich tapestry of a motor skill in action. This scenario showcases the brain's ability to both encode sensory stimuli and decode this representation into meaningful actions while modulating your play with memories and emotions.

Within the brain, sensory areas must *encode* stimuli, such as the edges and dark contrast of the notes on the page, and downstream areas must *decode* these features to build an internal model of yourself and the environment,[1] transforming the statistical spiking properties of input neurons to construct new useful representations within other neurons (Figure 1A).[2] For example, internal models ultimately serve to select control policies that enable goal-directed actions,[1] having integrated multi-sensory information with the prior state of the body. Collectively, this process of neural encoding and decoding lies at the heart of one fundamental question in neuroscience: that is, how the brain computes to perceive, act, and learn.

"Decoding the brain" therefore has two meanings: one, as described above, is how neural dynamics *decode* and transform incoming information across distributed circuits to represent meaningful information about sensory and other task stimuli (Figure 1A). The other is how we can build "*decoder*" algorithms to measure information in the brain (representational level analysis) and use it for translational approaches like brain-computer interfaces (BCI)[8,9] (Figure 1A), but this does not necessarily link to neural mechanisms. Nonetheless, both avenues require recording from neurons and transforming action potentials (or other signals gleaned from fMRI, electroencephalogram [EEG], etc.) into lower-dimensional representations of the data or latent factors.

However, the inherent challenge lies in the brain's complexity, from its vast scale to the intricacies of its internal language, spanning from the 302 neurons in a worm to the 80 billion in the human brain.[10] Fortunately, technological advances allow us to record from large numbers[11,12] of neurons and build powerful machine learning models.[13,14] Deciphering the neural code involves grappling with non-linear, dynamic systems distributed across brain regions, functioning across temporal scales to integrate past experiences, the current state, and future predictions.

A significant challenge of decoding is determining the necessary dataset scale and even timescale to train and evaluate the performance and generalization of the learned model. Namely,

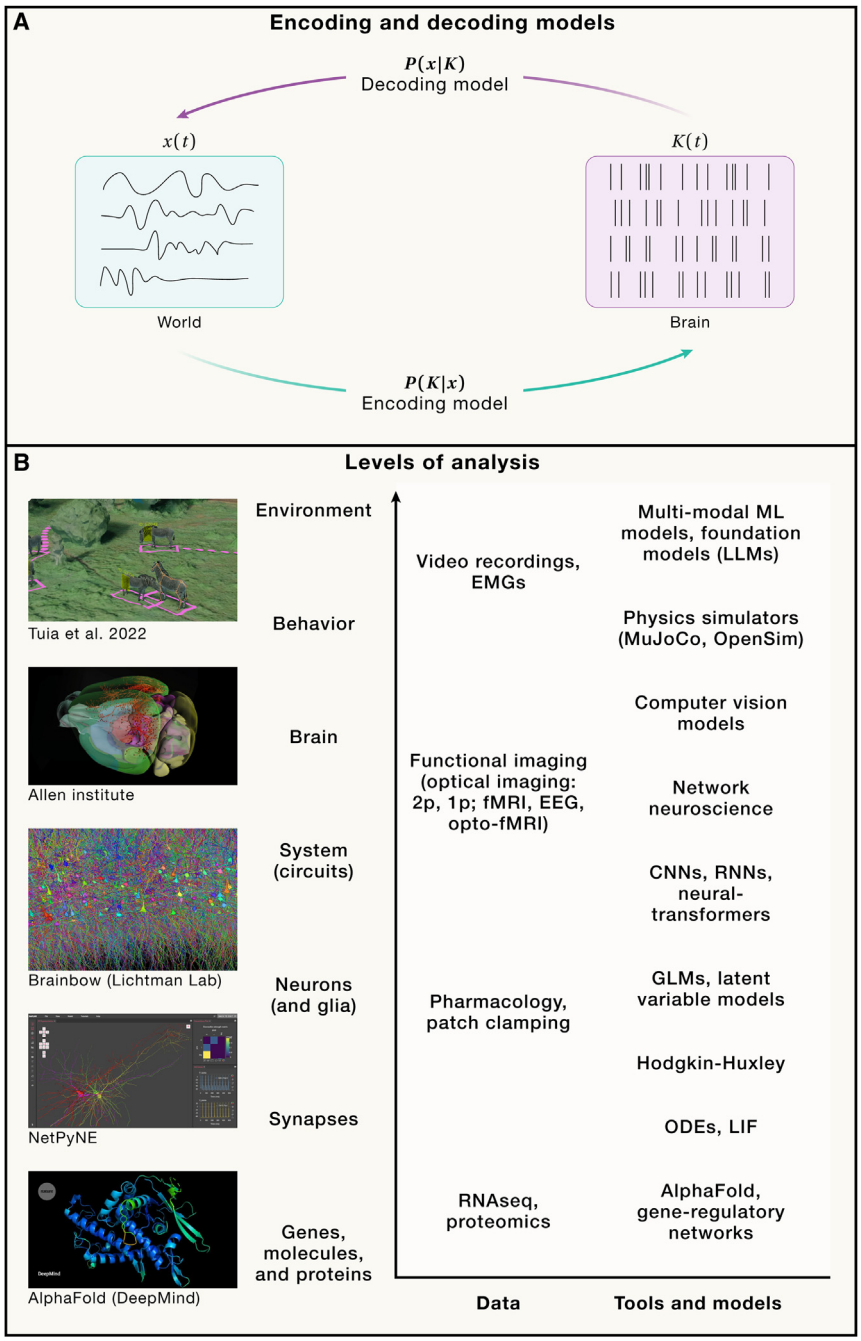

Figure 1. Encoding-decoding across scales

(A) An encoder represents the neural response of population $K(t)$ to stimulus $x(t)$ via $P(K|x)$, and a decoder aims to recover $x(t)$ given the neural activity $K(t)$ via $P(x|K)$.

(B) Systems neuroscience spans scales of descriptions and decoding algorithms can target any individual level and even span across scales. Here, we outline example scales (from genes to environment), the types of data we can collect (from genetic sequencing to whole-animal video analysis), and the classes of models the field has developed. On the far right is our mapping of scales, example data, and example tools to levels of understanding. Inset images adapted from: Tuia et al.,[3] Wang et al.,[4] Livet et al.,[5] Dura-Bernal et al.,[6] and Jumper et al.[7]

for constructing a decoder for BCI applications in motor control or language, for example, shorter time bins of spiking data from local neuronal populations are likely sufficient. However, learning hierarchical behavioral representations that span orders of magnitude in both time (from seconds to years) and space (from local environments to the entire world) in order to build foundational internal world models for downstream decoding tasks definitely requires richer, larger datasets. These datasets should ideally encompass recordings from individual neurons to the entire brain across multiple timescales (Figure 1B).

In this perspective, we begin by reviewing core principles of neural encoding-decoding and essential mathematical concepts and discuss work on using data-driven and normative machine learning models. Then, we provide case studies that highlight aspects of how neural decoding can allow for translational and foundational insights into the neural code. Lastly, we argue that ultimately the field should move toward causal modeling that allows us to infer and test causality in neural circuits.

NEURAL ENCODING

Decoding information from neural activity hinges on the assumption that the information is, in fact, encoded by the neural population in the first place (Figure 1). Thus, we begin by covering neural encoding principles, as it serves as a foundation for our understanding of how neurons encode information and how we interpret neural activity. Foundational research involving frogs, cats, mice, non-human primates, and humans has demonstrated that neurons convey information through their action potentials. As neuroscientists, we develop encoding models to quantify, from an information-theoretic perspective, the extent to which neural activity can be explained by externally observable or internally estimated variables.[15,16]

From a mathematical perspective, an encoder represents the neural response of population K to stimulus (or event) x:

$$P(K|x) \qquad \text{(Equation 1)}$$

Here K is a vector representing the activity of N neurons, and each entry represents, e.g., the number of spikes in some time bin or the rate response of that particular neuron. Fundamentally,

this statistical relationship summarizes how a group of neurons respond to an event x. As we discuss, there are different approaches for estimating these models.

Techniques such as linear regression, generalized linear models (GLMs), and artificial neural networks (ANNs) enable us to assess how individual neurons encode information.[16–24] This understanding is crucial because it forms the basis for constructing models of neural population dynamics, whether through population vectors, latent factor dynamical systems, or sophisticated hierarchical neural network models (which will be discussed in depth below).

In brief, linear regression models provide a basic framework by predicting neural responses through a linear relationship with stimulus features, while GLMs offer more flexibility by accommodating non-normal response distributions and non-linear link functions, making them well-suited for a broader range of neural data.[17,18] To quantify the amount of information neural responses convey about stimuli, information theory models such as mutual information are utilized, offering a measure of predictive accuracy without assuming a specific relationship form. ANNs consist of (multiple) layers of simplified (computational) neurons, whose connectivity patterns mimic the hierarchical, integrative properties of biological circuits. They are universal function approximators[25] and thus have emerged as powerful non-linear encoding models.[19–24]

DECODING WITHIN THE BRAIN

Once information is encoded, downstream areas must integrate information from upstream ensembles of neurons. For example, when reading the sheet music, the retina processes photons, and retinal ganglion cells transmit activity via the lateral geniculate nucleus to primary visual cortex. Even at the level of the retina, encoding and decoding approaches have been powerful to better estimate the variance in neural responses.[17]

The information encoded by specific groups of neurons, such as local contours in an early visual area of the visual hierarchy like V1 (although not all),[26] is processed or decoded by downstream neurons in higher visual areas like V4 to transform information and encode higher-order features like contours or textures.[23,27] Therefore, from the brain's point of view, neurons are encoding new information (i.e., representing specific latent world variables) by decoding and transforming information from upstream neurons. The encoding-decoding process must be thought of as two sides of the same coin, where neurons encode, transform, and process information from upstream neurons to more easily decode high-level relevant features of the world and drive behavior. Of course, sensory information cannot increase along a processing hierarchy, and thus all information about the visual world at a given instance is in the photon patterns impinging on the retina. However, the question lies in the simplicity of the decoder to extract relevant information, both from the experimenter's point of view and downstream neurons.[20,28]

For instance, the representation of a specific friend under all possible poses, lighting conditions, clutter, scales, etc., is encoded in the patterns of activity in the retina, forming a non-linear neural manifold embedded in a high-dimensional space. However, decoding or extracting the identity of that specific friend from all other possible images in a generalizable way requires a complex, non-linear decoder at the level of the retina.[28] In contrast, neurons in the inferotemporal (IT) cortex may allow for simpler decoding, potentially even with a linear decoder.[20,22,29] This progression through the visual processing hierarchy illustrates the shift from implicit to explicit encoding of information.[20,28] Early visual processing involves implicit encoding that does not directly convey object identity, whereas higher visual areas like the IT cortex provide more explicit representations tied to identifiable objects, making them easier to decode and more "human-interpretable." This view is, of course, oversimplified, as the visual cortex contains abundant redundant reciprocal connections, and this yin-yang of encoding and decoding is not simply done in a feed-forward manner.

One elegant line of work on how neurons might decode upstream spikes comes from studies on decision-making in rodents. Watabe-Uchida, Uchida, and colleagues have shown that dopaminergic neurons in the ventral tegmental area (VTA) encode reward prediction errors (RPEs), and GABAergic neurons encode a function akin to the estimated state value.[30] Notably, this maps exceedingly well to reinforcement learning algorithms.[30,31] Then, to address how dopamine neurons come to compute these RPEs, they measure their anatomical "inputome,"[32] and go on to record from these upstream neurons to find partially computed RPEs,[33] suggesting that dopamine neurons must decode these partially computed RPEs and inputs from GABAergic neurons in order to fully compute then broadcast RPEs.[34] This elegant example highlights how theoretical models and the neural decoding framework allow us to estimate what neurons decode in order to compute. Moreover, it offers us a mathematical framework to formalize mapping neural representations to neural computations.

NEURAL DECODING: MATHEMATICAL PRINCIPLES

As outlined, information processing in the brain can be conceptualized (in a simplified fashion) as a series of cascading encoding-decoding operations. Through these operations, the brain extracts relevant information from the environment, transforms it, and ultimately uses it to guide behavior. Decoding models serve as powerful tools in this context. This section goes into the mathematical principles of these decoders.

Consider a population of neurons encoding a stimulus x as described by Equation 1: $P(K|x)$. A natural question to ask is can we predict x from a spike count vector K? The aim of a decoder is to predict x from the neural response K. Mathematically, a decoder is a function that maps K to some estimate $\hat{x}(K)$. Naturally, many different decoders are possible, and we first describe one of the simplest—the linear decoder—in more detail. A linear decoder combines the activities of the different neurons in a linear fashion, i.e.,

$$\hat{x}(K) = w_1 \cdot K_1 + w_2 \cdot K_2 + \ldots + w_N \cdot K_N, \qquad \text{(Equation 2)}$$

where the different w_i are weight vectors that indicate how much the activity of neuron K_i contributes to the estimate. This decoder is biologically plausible, as it is rather natural to think of neurons to linearly combine their inputs.

A particularly instructive and simple example is given by the cercal system of the cricket.[35] The cricket has four neurons that integrate information of hair cells that are responsive to planar wind (Figure 2A). By combining their activity in a linear fashion, one can estimate the wind direction in a simple and effective way (Figure 2B). This linear combination is commonly called population vector, i.e., the weighted sum of the activity of all neurons. We will later see that the population vector enables strong movement decoding in primates,[36] and that linear decoders are essential for assessing learned representations of models (linear probing). Furthermore, linking to the earlier example, object identity can be accurately decoded from IT but not from upstream brain areas with linear decoders.[20,29]

Another simple yet powerful decoder is given by the k-nearest neighbors (k-NN) algorithm (illustrated in Figure 2C). Consider L recorded neural responses K^l each associated with a stimulus x^l: $(x^1, K^1), (x^2, K^2), \ldots (x^L, K^L)$. To decode the stimulus from a new trial with neural response K, the 1-NN algorithm finds the neural pattern K^j, which is closest to K, and assigns the new trial to the corresponding stimulus x^j (Figure 2C). In the more general case, one considers the k-NN and can also average their events.

Bayesian decoders directly use the probabilistic encoding model.[16,37] Concretely, a Bayesian decoder uses Bayes' theorem to compute the probability that, given a response K, the stimulus x was presented. Mathematically, let $P(x)$ denote the probability of a stimulus x and $P(K|x)$ the conditional probability of obtaining the population response K given the stimulus x (as in Equation 1; illustrated in Figure 2D). Bayes theorem states that:

$$P(x|K) = \frac{P(K|x) \cdot P(x)}{P(K)}, \qquad \text{(Equation 3)}$$

with $P(K) = \sum P(K|x) \cdot P(x)$. Using this expression of the posterior probability $P(x|K)$, we can predict the most likely stimulus: the x that maximizes $P(x|K)$. Bayesian decoders such as the naive Bayes decoder have been popular for position decoding from hippocampal activity, and here one can utilize priors representing the *a priori* expected location of the animal.[38] Another classic decoder builds on the Kalman filter[39] and allows levering the dynamics of the system.

How can we assess the quality of a decoder? For a continuous variable, like wind direction, we can just check how well it reconstructs the original stimulus x:

$$\mathbb{E}_{K \sim P(K|x)} \| \hat{x}(K) - x \|^2, \qquad \text{(Equation 4)}$$

where we average over samples K.[16] This is the variance of the decoder. We think that one particular decoder is better than another decoder if it has a lower variance or, in other words, if it is more accurate at estimating x from the neural response of the population.

Given the large number of potential decoders, establishing lower bounds to the variance is crucial to studying theoretical components of representations. The Cramèr-Rao inequality serves exactly this purpose. It states that the variance of any unbiased estimator can be bound from below by the inverse Fisher information.[16] Notably, the Fisher information can be calculated directly from the encoding model, $P(K|x)$, enabling the extraction of valuable insight for simple tuning curves (such as parameterized place and grid cells) within this framework.[40,41] For an excellent review on this topic, see Kriegeskorte and Wei.[42]

However, in many instances where the encoding model $P(K|x)$ is complex or not even explicitly known, calculating the posterior probability $P(x|K)$ is computationally challenging. As we discuss next, this is where data-driven and normative models, taking advantage of machine learning, can learn powerful statistical models to facilitate decoding.

DATA-DRIVEN MACHINE LEARNING

As delineated above, the key to a good decoder algorithm is a solid encoding model. There are several approaches to do so, namely data-driven or task-driven models. Data-driven decoding approaches, which we cover in this section, build powerful *statistical models* to assess how stimuli or behavior are encoded in neural activity, not necessarily focusing on mechanistic realism (Figure 3). Concretely, by *mechanistic models*, we mean they aim to model the biological basis of a given neural function, such as the Hodgkin-Huxley model.[43]

Historically, data-driven models were hindered by, one, the lack of large-enough datasets, but modern methods allow for single-neuron resolution of nearly 1 M neurons at a time,[11] and two, modeling approaches that could combine data across sessions and animals without averaging data.[13,14] Combined, this paradigm shift now greatly enables data-driven modeling approaches capable of capturing the complexity of high-dimensional population activity while remaining computationally feasible.

Two main categories of data-driven statistical models have emerged in neuroscience: fully observed models and latent variable models.[14,48–53] Fully observed models, such as GLMs and vine copula models, strive to explicitly delineate the interactions among neurons by directly modeling the joint activity of the population, operating under the assumption that the recorded population encompasses all relevant neuronal activity without the need to account for unrecorded neurons contributing to the neural manifold. On other hand, latent variable models, which we will mostly focus on, infer hidden (i.e., latent) variables that capture the underlying structure of the observed neural data through a joint probability distribution, acknowledging the possibility that unrecorded neurons or other unseen factors may contribute to the observed data. The goal of both approaches is to measure how much information, often measured through decoding or other information-theoretic approaches, about a given stimulus or behavior is captured by a model.

Given the intractability of modeling the entire space of neural activity, fully observed models rely on efficient descriptions of neuronal dynamics, often *a priori* ascribing the dynamics for simplicity.[51,54] On the other hand, latent variable models operate under the assumption that population activity is typically constrained on a low-dimensional manifold and can be summarized by a compact set of variables known as latent factors (or variables).[55]

How can we extract neural latents? Classic dimensionality reduction techniques such as principal-component analysis

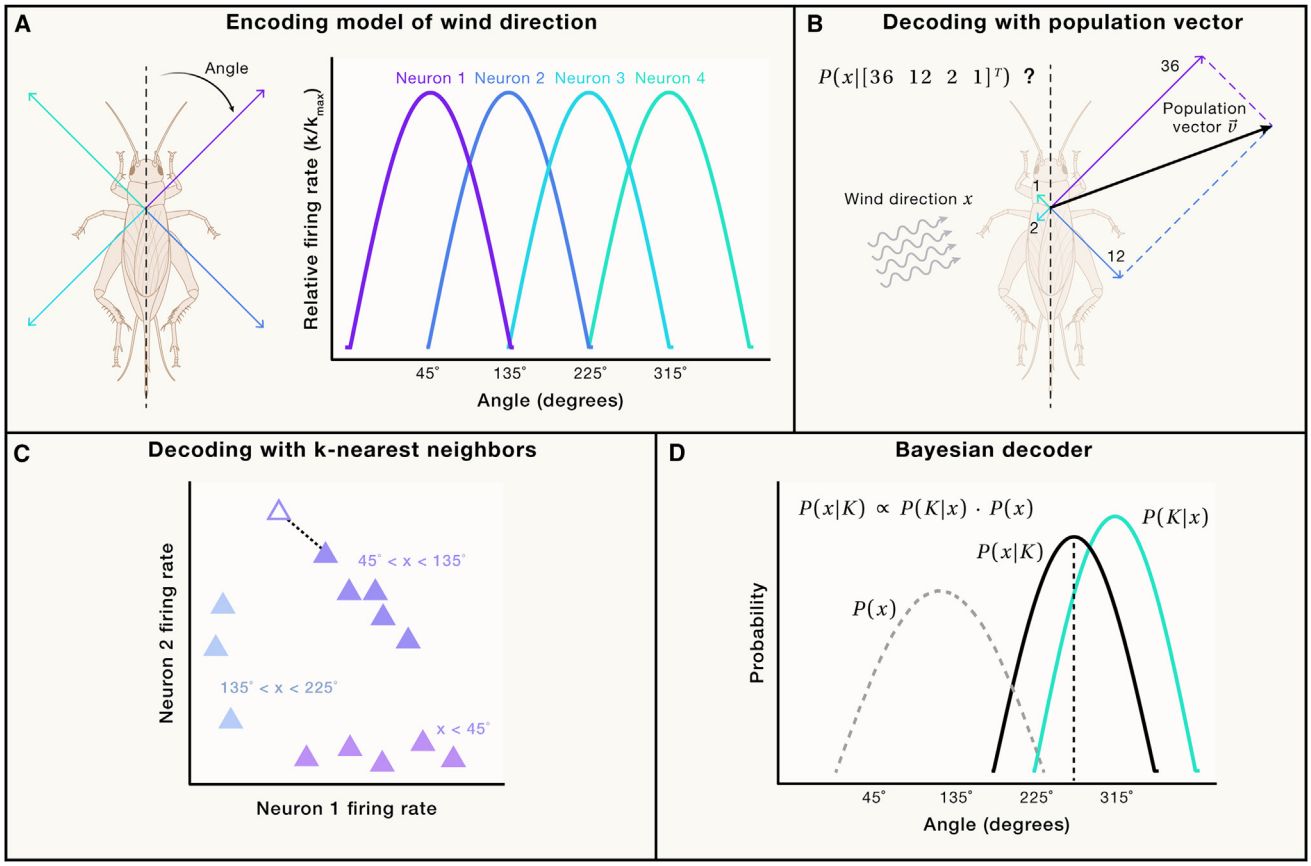

Figure 2. Encoding model and decoding methods

(A) The cercal system of the cricket has four interneurons that represent the wind direction. The preferred wind directions of the neurons are pointing in four cardinal directions and can be represented by orthogonal vectors (on the left). Each neuron responds with a firing rate approximated by a half-wave rectified cosine function. The maximum firing rate is elicited when the wind is blowing in the preferred direction.

(B) The wind direction x can be decoded as the direction of the population vector $\vec{x}$. This vector is the sum of the four preferred orientations scaled by their firing rate. An example is shown for neurons responding with activity $[36, 12, 2, 1]^T$. Note how the population vector closely matches the wind direction.

(C) In the k-nearest neighbors (k-NN) decoding method, neural activity K is represented within a neural activity space, which is illustrated here in 2D for two neurons for clarity (neuron 1 and neuron 2 from A). With these two neurons, angles between 0° and 225° can be represented. For simplicity, we focus on an NN variant with $k = 1$. As 1-NN is only able to decode discrete variables, we classify the angles in three ranges: 0°–45°, 45°–135°, and 135°–225°. Previously observed trials are color-coded by their associated wind direction ranges ($L = 13$). To decode the wind direction for a new trial (unfilled triangle), the k-NN (here, $k = 1$) in the activity space are identified. The decoded wind direction corresponds to the wind associated with the NN, highlighted by the sample connected to the observed sample via a dashed line.

(D) Bayesian decoders incorporate a prior $P(x)$ (dashed line) that reflects the probability of different wind directions before taking neural evidence into account and influences the decoded angle. For instance, if mainly wind directions around 125° have been experienced (mean of the prior $P(x)$), the decoded angle will be shifted toward this direction. The likelihood $P(K|x)$ (green-blue line) describes the probability of observing a particular neural response K given a specific wind direction. Following Bayes' theorem,[9] the prior $P(x)$ and the likelihood $P(K|x)$ are multiplied to obtain the posterior distribution $P(x|K)$ (solid black line). The posterior can be used to decode the wind direction, here 270°, based on the highest value for the observed neural activity K.

(PCA)[56] and independent component analysis (ICA)[57] simplify neural data by revealing the underlying (linear) structure essential for understanding the stimuli that are encoded. Of note, while such linear methods enhance interpretability, they often sacrifice performance and may over-estimate the true dimensionality.[58,59] A correct estimation of the intrinsic dimensionality is arguably critical for scientific interpretability. Namely, our aim is to reduce the dimensionality of the high-dimensional neuronal space into a latent space of reduced dimensions where the geometry of the latent space is interpretable with respect to behaviorally meaningful features like objects, actions, decisions, cognitive states, etc. Emerging frameworks have successfully illustrated this in many different domains.[14,59–63]

These latent factors are the mathematical representation underlying dynamics that give rise to the observable data (spikes) that we can directly record (Figure 4A). Thus, rather than explicitly modeling neuron correlations, these models capture neuronal relationships through the activation of these latent factors, often through non-linear machine learning approaches such as variational auto-encoders (VAEs),[64] or contrastive learning.[14,65,66] The aim of these methods is to generate a latent space, or so-called embedding, that captures the variance of the observed neural data within a smaller number of factors. This is akin to a dimensionality reduction procedure.

With those frameworks, complex, non-linear relationships between stimuli and neural responses can be assumed.

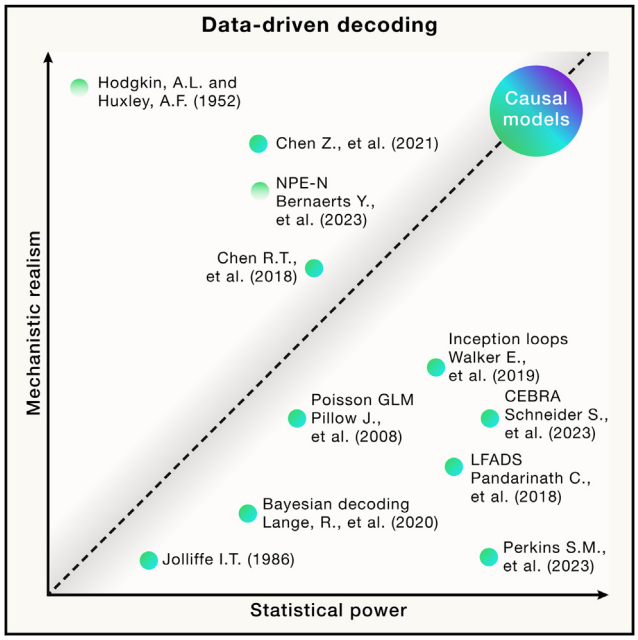

Figure 3. Data-driven models: Statistical power vs. mechanistic realism

Ultimately, as a field, we want to map mechanism to computation in order to have causal, testable models. On one side we have mechanistic models, and on the other statistical models that aim to best encode neural dynamics, but there is a large gap between them. We provide a non-exhaustive selection of contributions: Schneider et al., [14] Pillow et al.,[17] Lange et al.,[37] Hodgkin and Huxley,[43] Jones et al.,[44] Bernaerts et al.,[45] Chen et al.,[46] and Chen et al.[47]

ANNs are particularly useful for quantifying non-linear relationships in neural data.[13,19,20,26,29] Recent strides in non-linear disentangled representation learning and self-supervised learning have paved the way for new methods that can be jointly applied to behavioral and neural recordings,[65,67] unveiling meaningful lower-dimensional neural population dynamics. Recently, a new dimensionality reduction method called CEBRA introduced a new paradigm for joint modeling of time-series data with a generalized contrastive learning algorithm[14] (Figure 4A). Critically, the data-sampling scheme (the selection of "positive" samples of paired (x, K) data vs. the "negative" samples that are ultimately contrasted against, ϕ) and the model optimization (minimally denoted by $\phi(\cdot, \cdot)$) are directly linked such that the resulting latent factors can be interpreted based on the input auxiliary variables (such as behaviors [kinematics], animal identification, rewards, estimated internal states, etc.).

Given an auxiliary variable x, such as the continuous position of an animal, and neural data K, one can select the positive distribution (samples) of paired data, $p(x|K)$, to explicitly test the relationship of the auxiliary variable to the neural data. This is then contrasted with a negative distribution $q(x|K)$ to optimize an ANN. It was empirically shown that if x does not influence K, the model cannot falsely fit the data (in fact, it collapses on the manifold used for training[14]). Thus, this method can be used to construct data-driven models that allow for hypothesis testing, and this model can be simply linked to a decoder algorithm of choice for downstream use, whether for BCIs or for interpretation of the latent variables.[14,68,69]

Note, the goal of this contrastive approach, in comparison to auto-encoders (like VAEs), is not to reconstruct the input data (i.e., spikes) but rather invert the data generating process to extract latent variables that give rise to the recorded data (whether from spikes, fMRI, ECoG, or calcium imaging[14,68]). A promising direction of this work is to interrogate how neurons across time contribute to latent variables.[69] Here one can build on advances in interpretable machine learning. Moreover, we can begin to derive the underpinning ordinary differential equations (ODEs) that govern the latent representation.[46,47,70] Critically, this approach could be additionally merged with normative approaches (see below) in order to ultimately link mechanistic to statistical approaches to build causal models (Figure 3).

NORMATIVE, TASK-DRIVEN MACHINE LEARNING

Normative models address the question of why a system exhibits particular features,[71,72] aiming to better link mechanism to function. For example, Horace Barlow proposed the efficient coding hypothesis, which postulates that a sensory system minimizes the number of action potentials (energy constraints) to efficiently represent sensory information.[73] One can, for instance, deduce that simple cells emerge as a consequence of sparsely representing natural scenes.[74] Sparse coding also explained coding properties in many other systems.

Machine learning has elevated normative modeling to new heights in the name of task-driven modeling, where the normative principle is the goals (or computational task, i.e., Figure 1B) that the system is trained to achieve.[20,72] It is important to keep in mind that just 15 years ago, it was considered challenging to train machines to recognize objects in natural scenes.[28] Yet, while robustness issues persist,[75] in the meantime, these models are the best models at predicting the neural representations of the ventral pathway in non-human primates.[20,22] Importantly, for decoding the brain, this suggests that we can gain insights into neural coding with a complementary approach to data-driven modeling. This approach is particularly successful in the (neural) data-poor regime.

The normative approach leverages the computational power of ANNs to explore how neural circuits implement complex tasks. The underlying assumption is that neural responses emerge as an interplay of task objectives, neural circuit architecture, and learning constraints. This approach involves training models on specific tasks to test hypotheses regarding the computations carried out by neural circuits. By doing so, they provide insights into the functional significance of neural activity patterns and their role in mediating behavior. The emerging representations are then compared neural data via linear probing, i.e., one checks how well the activity on test stimuli can be decoded from the learned representations via a linear readout.[20,29]

This normative framework is general and has been applied beyond vision,[19–21,23,24] in studies on audition,[76] proprioception,[77,78] heat perception,[79] and path integration.[80,81] In contrast to neural-data-heavy (data-driven) approaches, here one needs large-scale stimulus datasets, tasks, and architectures. Thus, for instance, in the context of path integration,

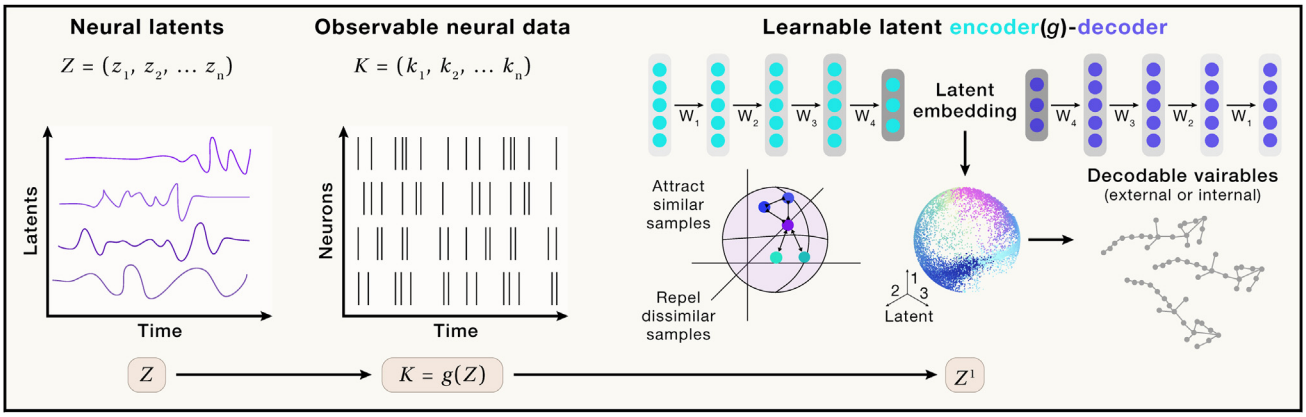

Figure 4. Learnable latent variable models

On the path to building more causal models are new frameworks, such as CEBRA,[14] that allow for learning a mapping from the observable data *K* to the latent dynamics *Z*. Here, the aim is to use identifiable models with contrastive learning (the encoder), then invert this model or use another decoder framework to probe the relationship between the estimated latents, Z^1, and a variable such as an externally observable state (behavior), internal, or sensory (i.e., recover some stimulus space (*x*)).

one simply needs movement trajectories.[81] In the context of object recognition, one needs large-scale image datasets with annotations of objects.[19,20] If primary data for a particular stimulus is not easily available, physics engines can help to generate the relevant large-scale datasets.[77,78,82] For instance, Marin Vargas et al.[78] used musculoskeletal simulators to generated synthetic proprioceptive inputs resulting from passive, naturalistic movement at scale, which they then used to train neural network models on computational tasks reflecting hypotheses about proprioception. Subsequently, they tested if the network's learned internal representations resembled those of proprioceptive brain areas and found that task-driven models could more accurately predict single-trial neural dynamics than classical encoding models. For the majority of computational tasks, architectures that performed better at solving the computational tasks based on biomechanical data were better at explaining the neural data.[78] This work highlights that we can gain insight into neural processing via constraints from the body, including muscles.

However, as our capabilities to scale up recording techniques improve and we enter the regime of very large-scale data in neuroscience,[11] including in behaving non-human primates,[83] we are beginning to see data-driven models outperform task-driven ones.[84] This shift opens exciting research directions to develop new normative models to bridge these gaps. One such intriguing principle is the idea that biological neural systems have evolved to reflect the symmetries of the natural world. For example, in object recognition, translation, scale, and 3D pose represent group symmetries or transformations that do not alter the identity of the object. In the context of group theory, a symmetry is characterized as a transformation that maintains the identify of an object through the relevant operations. Leveraging the principle that neurons learn representations invariant to 2D translation, recent works have derived filters similar to those of V1 neurons, proposing an alternative normative model to sparse coding that predicts the characteristics of neurons in V1.[85,86]

VISUAL FEATURE DECODING

Understanding the brain's algorithms of visual perception requires comprehending how natural visual scenes translate into neural activity. This subject has been explored using two complementary methods: encoding methods, which describe neural responses based on the stimuli, and decoding methods, which aim to reconstruct the visual scene or specific attributes from neural data without learning an encoding model first.

Historically, predicting how higher-order visual neurons respond to visual stimuli has long been an open challenge.[28] Yamins et al.[29] proposed the task-driven modeling approach and could substantially improve our ability to predict the neural activity of V4 and IT neurons. Subsequently, progress was fast, and more powerful encoding models were created gaining insights into core-object recognition in primates. This approach was also successful for modeling neural responses to static images in mice[24,26,87] and studying how brain states modulate neuronal tuning.[88]

These models have mostly focused on static images, and learning encoding models for dynamic scenes has been a challenge. By following the data-driven foundation model paradigm,[89] recent works have built a video encoding model that was trained on large-scale data (> 70, 000 neurons) combing data from many mice and visual areas.[90] Foundation models, by definition, are built in order to provide a strong encoder model for many downstream decoding tasks. Indeed, the model improved the predictive power for natural videos and many other stimulus domains that the model was not trained on.[90] It could also predict synaptic connectivity in the MICrONS data,[91] which combined functional imaging with synaptic connectivity measured through electron microscopy.[92] Conversely, anatomy can also be used to build better encoding-decoding models. The recent emergence of a partial connectome in *Drosophila* has also already been leveraged in models of vision using a connectome-constrained deep mechanistic network that was able to predict neural responses across the fly visual system at single-neuron resolution.[93]

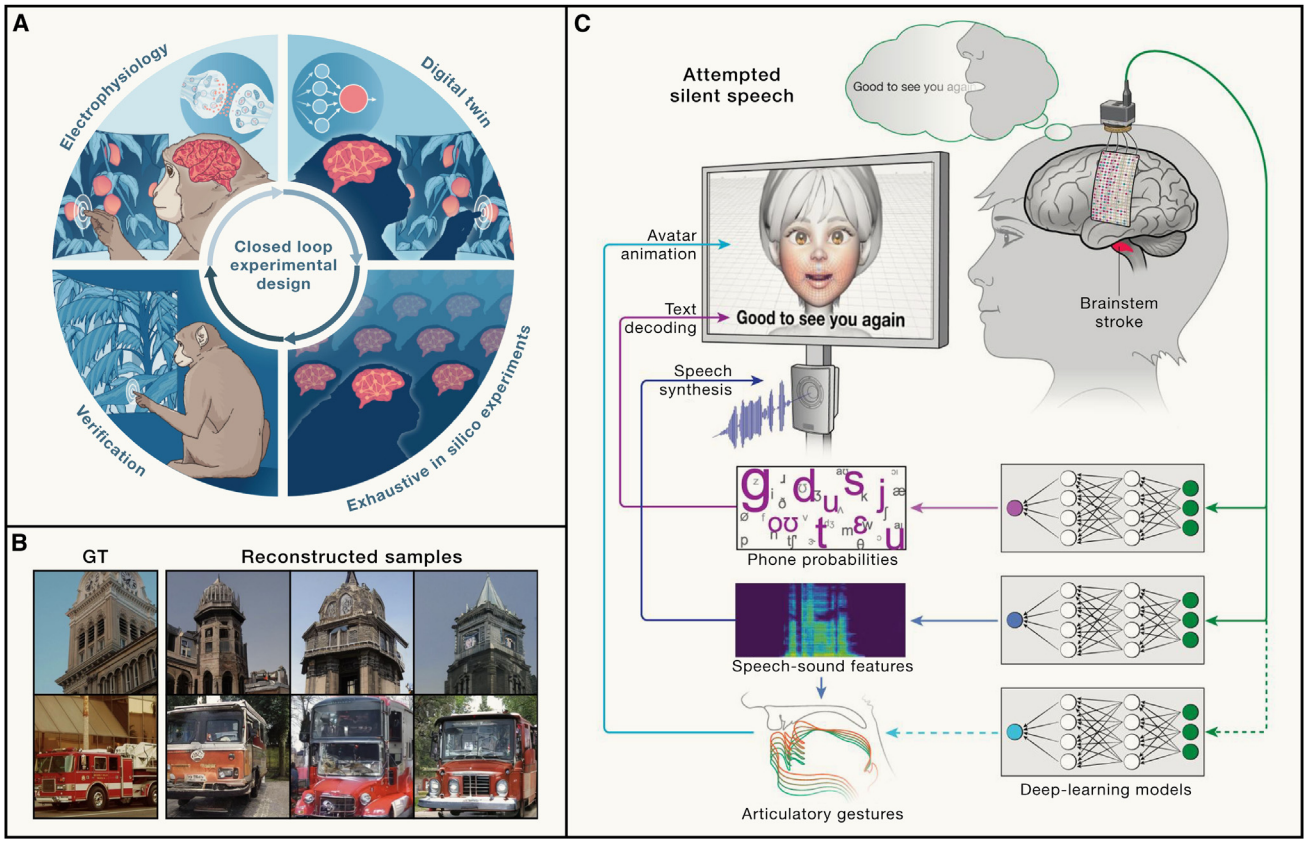

Figure 5. Examples of decoding from motor, vision, and language areas

(A) Close loop experiments using digital twins: schematic of an inception loop, depicted clockwise from the upper left: (1) presentation of large entropy natural stimuli and tasks while recording large-scale neural activity; (2) deep learning models accurately predict neural activity, creating a functional digital twin of the recorded neurons; (3) the *in silico* model facilitates unlimited experiments and employs mechanistic interpretability tools to characterize neural tuning; and (4) images and hypotheses synthesized *in silico* are validated back *in vivo*.

(B) Illustration of decoded images from fMRI using diffusion models: ground truth (GT) vs. decoded images generated by Chen et al.[96] from human fMRI with a diffusion model. Note that decoded images share similar color, shape, and semantics.

(C) Multi-modal speech decoding: adapted from Metzger et al.,[97] this panel shows the decoding pipeline, where neural activity was used to train an ANN to predict phone probabilities, speech-sound features, and articulatory gestures. A decoder was then constructed to produce text, generate audible speech, and animate an avatar, respectively.

When modeling more complex neural circuits, such as those in the neocortex, deep learning models that lack implementation-level details remain the gold standard. Introducing mechanistic-level details of biological circuits to build more accurate predictive models is challenging. This is formidable due to the inherent complexity of cortical circuits and the difficulties associated with global connectivity mapping. Moreover, even though the architecture and elements of networks might not limit the range of functions they can realize, they significantly influence which function is selected during the learning process when data is limited. This means that the inductive bias of architectural elements is intricately linked to the learning rule, underscoring that having the correct circuit structure may not be enough without the appropriate learning rules to guide the model toward better predictions.

A common criticism of using ANNs to model the brain is that we are simply swapping one complex system (the brain) for another (ANNs). However, unlike the brain, we have a degree of control over how ANNs are constructed.[72] While we may not fully grasp the nuances of their capabilities, ANNs are to some extent explainable through their architecture, task performance, and training data. Furthermore, encoding-decoding models can also be leveraged for closed-loop physiology experiments to gain insights into neural coding (Figure 5A). A few years ago, several groups made exciting breakthroughs in this domain.[26,94,95]

The development of accurate predictive models of neural activity that function as digital twins of brain circuits has opened the door to conducting *in silico* experiments. These models, combined with emerging tools from the field of *interpretability* research (or *mechanistic interpretability*),[98] enable the generation of hypotheses that can be tested *in vivo* through closed-loop experiments (inception loops, see Figure 5A). These models have proven effective in predicting how the brain responds to novel images[26,94,95] and have paved the way for creating synthetic images that maximally activate specific neurons or selectively drive a particular neuron while inactivating another group.[26,94] This approach has yielded significant insights: for

example, evidence for a columnar organization of tuning to spatial patterns in visual area V4,[23] the characterization of single-neurons invariances,[99] and the exploration of contextual modulations at the single-neuron level.[100] Recent advancements, such as the use of diffusion mechanisms, have further improved the process of predicting the *most exciting stimuli*,[84] offering faster and images that can generalize better across varying model architectures compared to earlier gradient-based.[26,94]

In parallel, different groups also succeeded in reconstructing visual input from neural activity. Using CEBRA with a *k*-NN decoder to predict the best matching video frame from neural activity in the visual cortex of mice.[14] The predicted frame was close to the frame that the mice actually saw (with ≈ 95% accuracy). Here, the authors did not decode images at the pixel level. Yet several other teams, based on paired data of images and fMRI activity, trained data-driven diffusion models to tackle this question.[96,101,102] While these demonstrated that a diffusion model could decode realistic and semantically correct images (Figure 5B), others have observed a significant decrease in performance when applying these methods to datasets that were specifically designed to prevent category overlaps between training and test sets, underscoring the limitations of these fMRI approaches.[103]

Taken together, the fusion of large-scale neural activity recording with the latest advancements in machine learning has significantly enhanced the precision of encoding-decoding models, even in the context of higher visual processing. Although encoding models are evaluated based on how accurately they predict neural responses, interpreting these models and what specific groups of neurons represent becomes complex when dealing with natural images due to the highly non-linear nature of neuronal tuning.

Decoding models are instrumental in translating neuronal group activities into understandable stimulus features, such as stimulus orientation, motion direction, or the detection of specific objects. The effectiveness of these decoding approaches is heavily influenced by the selection of a quality metric or training loss function, which guides the estimation of these decoded features. For tasks like identifying an object's class or the motion direction of a stimulus, the choice of loss function is straightforward. However, when tackling more complex challenges, such as the reconstruction of whole visual scenes, defining an appropriate loss function becomes more daunting.

Metrics based on image properties, like the widely utilized mean-squared-error of pixel intensities, often fail to align with how humans perceive similarity. This issue becomes particularly pronounced when attempting to develop an effective loss function for reconstructing intricate natural scenes in higher visual areas. These areas are proficient at extracting specific latent features from visual scenes, including textures, shapes, colors, and faces. However, the specific visual features encoded are not known in advance. The key measure of success lies in the ability of the decoded stimuli to accurately reflect the brain activity that initiated their reconstruction, essentially creating visual equivalents, or metamers, that are indistinguishable to a given neuronal population when compared to the original image.

Lastly, in visual decision-making tasks, decoding can be very powerful during closed-loop experiments. An elegant example is work from Peixoto et al.[104] where the authors record from the motor cortex during a visual discrimination (motion dot) task. They constructed a "decision-variable" (DV) decoder such that at each time step, in real time, they have a continuous readout of the decoder's prediction (concretely, the logistic model's log odds ratio) if the macaque will choose left or right. They found that the within-trial DV fluctuations could predict behavioral choices (starting within only 250 ms) substantially better than the condition-averaged DV or the visual stimulus alone (and it correlated with the strength of the motion coherence). Moreover, analogous to the inception-loop paradigm, they used the DV to terminate the trial at a particular DV threshold to test how accurate the model was: if they terminated when the model strongly predicted the macaque would "choice left," for example, it was over 90% accurate, suggesting that this readout from motor cortex was tightly linked to the perceptual decision-making of the animal.

DECODING THE MOTOR CORTEX

The study of the motor system testifies to the diversity and evolution of approaches to decode the brain: from deciphering the role of individual neurons to decoding the computations underlying motor control. Pioneering work in the early 20th century established the fundamental link between individual neuron activity and motor function. Singular motor neurons firing in the spinal cord activate specific muscles directly. Upstream, in the primary motor cortex (M1), the activity of single neurons was found to be correlated with a range of movement-related variables,[105] such as force,[106] muscle activity,[107] and joint kinematics.[107] Given that layer 5 M1 neurons project directly onto spinal alpha motor neurons (in many primates), it is expected to find the representation of low-level movement variables.

If the activity of some neurons in M1 relates to lower-level kinetic and kinematic features, what about others? In the 1980s, Georgopolous and colleagues showed that the direction of whole-arm movements in monkeys could be predicted by simultaneously recording from multiple M1 neurons.[108,109] They found many neurons that are broadly tuned to a preferred direction of hand movement. By constructing a population vector—a weighted sum of the firing rates of multiple neurons—the direction of the hand movement can be accurately decoded (akin to the linear decoder in Figure 2). This framework highlighted a simple mathematical relationship between the reaching direction and population neural activity (cf. Sussillo et al.[8]). The population vector approach could successfully find representations of higher-level kinematic variables in the motor cortex.[109–111]

BCIs seek to build a link between neural activity and various tools, e.g., in order to control computer cursors,[112–116] robotic arms,[117–119] or prosthetic devices for paralyzed patients.[120,121] Another consideration, especially for BCIs, is the impact of volitional control on neural activity. Subjects have been shown to quickly learn to optimize control by modulating neural activity, allowing for effective operation of devices like robotic arms or functional electrical stimulation onto muscles,[116,122] irrespective of the neurons' original encoding.

BCIs thus rely on brain decoders and are ideal for testing the performance of different algorithms. Early BCIs primarily used

linear decoders for extracting movement kinematics from neural population activity.[123] However, incorporating Kalman filters has demonstrably enhanced accuracy.[119,124–126] As indicated above, the Kalman filter[39] helps estimate the evolution of the state (e.g., cursor velocity) over time and updates its estimates according to the observations (neural recordings) and evolving predictions. The success of Kalman filter decoders highlights a key limitation of the classic population vector approach. While population vectors offer a static snapshot of neural activity, movements are inherently dynamic. Furthermore, identifying representations of movement-related variables within a population vector does not illuminate how either the representation or the movement itself is generated.

The dynamical systems perspective focuses on how the state of a neural population evolves over time. Here, the representation of a neural population is viewed as a dynamical system that performs computations, such as generating movements, through its temporal evolution. External inputs and intrinsic neural dynamics, which dictate how the current neural state influences the next, govern the evolution of neural activity.[127,128] Let the vector $K(t)$ describe the firing rates of N neurons at time t. We can express the evolution of this vector with the following equation

$$\frac{dK(t)}{dt} = f(K(t), u(t)) \qquad \text{(Equation 5)}$$

where the vector $\frac{dK}{dt}$ is the temporal derivative of K, u is a vector describing the external inputs to the neural population, and f is a function that defines the dynamics of the neural population.

This perspective shifted the focus from deciphering the information encoded by the motor cortex to understanding how it generates movement. Within this framework, how does Equation 5 aid in decoding the computations underlying movement control? Let's consider the following: to execute a specific movement at time t_m, the neural activity, represented by $K(t_m)$, needs to reside within a certain subset of states—a specific configuration of activity across individual neurons. Prior to target presentation, nothing prevents the neural activity K to vary across trials. However, as the brain prepares for movement, Equation 5 governs the evolution of $K(t)$, effectively constraining it toward the required subspace for the intended movement. This convergence toward a specific subspace manifests itself as the reduction in variability across trials. Importantly, this analysis offers valuable insight into the computational goal of motor preparation. From a dynamical systems perspective, the goal becomes driving the neural activity toward the subspace before movement onset to produce the necessary motor commands for the desired movement.

The dynamical systems perspective has been further leveraged for studying movement preparation.[129–132] Kaufman and colleagues identified distinct subspaces within the dynamics of the overall neural population during movement preparation and execution.[133,134] These subspaces were termed the "null space" and the "potent space." During preparation, neural activity can evolve within the null space without triggering unwanted movement. In contrast, movement execution is characterized by neural activity primarily in the potent space, which controls muscle activity. However, a part of the neural activity does not drive the movement directly. Instead, it serves a supportive role in "untangling" the neural dynamics—separating the neural states that can result in different future behaviors.[135] Further analysis of movement-related neural activity has revealed that the dynamics have a strong rotational component and that their phase and amplitude are determined by the neural state reached during movement preparation.[9,54,134,136]

So far, we have reviewed how the dynamical systems perspective sheds light on experimental recordings of neural populations. By analyzing these recordings, researchers have been able to formulate principles governing the computations that take place during motor preparation and generation. In parallel, computational models have provided valuable insights.

Data-driven models based on recurrent neural networks (RNNs) have been used to model the system's dynamics Equation 5. These RNNs embody abstract representations of the underlying neural circuit while driving movement. For instance, RNNs have been trained to generate dynamics of muscle activity patterns.[137,138] Remarkably, even without being trained to fit neural data, these models can reproduce responses from individual cortical neurons and capture features of the observed population dynamics.[137–139] Furthermore, training RNNs on different motor tasks can test hypotheses about the computations performed by distinct motor areas.[140,141] Additionally, RNNs have shown promise in increasing decoder robustness to temporal variations.[8,142,143] More recently, latent factors have been leveraged to train spiking neural networks to perform two distinct motor tasks,[144] and data-driven latent variable models that use time contrastive learning have been shown to have excellent performance for decoding movement in sensorimotor areas.[14]

In the past decade, the convergence of advancements in understanding cortical dynamics and powerful machine learning tools has opened exciting avenues for BCIs.[121,143] Incorporating latent factors and their dynamics into decoding algorithms has improved performance.[143,145,146] For instance, data-driven machine learning has played a critical role in achieving long-term control of a four-limb exoskeleton by a tetraplegic individual using a BCI.[147] Transformers, with their ability to be pre-trained on a wide range or motor BCI datasets, offer the potential to enhance BCI adaptability across experimental contexts.[148]

Beyond therapeutic applications, BCIs have emerged as powerful tools for studying motor learning and adaptation.[122,149–151] In a typical setup, awake monkeys control a computer cursor using their neural activity while receiving visual feedback of the cursor's position.[152] Unlike traditional motor tasks, where complex and largely unknown transformations convert cortical activity to muscle activity and movement, in a BCI, the decoder characterizes this transformation. By manipulating this mapping, experimenters can study the adaptation required to compensate for the perturbation.[152] For instance, it was found that monkeys could learn to compensate for perturbations within the original intrinsic manifold within a session, but it took many sessions to learn perturbations that required control in directions outside the manifold (off-manifold).[150] This finding reinforces the idea that although neural state space is high-dimensional, neural population dynamics lives in a low-dimensional manifold reflecting intrinsic constraints.

Beyond intrinsic constraints, the physical properties of muscles, tendons, and bones as well as the sensory feedback streams, play a significant role in determining how neural signals create movement.[153–156] Recently, a study in fruit flies has highlighted how the brain controls head movements by adding a bias to proprioceptive feedback loops.[157] In a BCI setting in humans, stimulating the somatosensory cortex to mimic tactile feedback has been shown to improve robotic arm control.[158] Going forward, we need better models that incorporate biomechanics and biophysics. Advances in biomechanics simulators[159–161] and continuous control learning algorithms[162–164] are being actively combined to further address these challenges.[164–168]

LANGUAGE DECODING

The power of machine-learning-fueled decoding algorithms has perhaps most clearly been illustrated for speech decoding, where the action spaces are naturally very high dimensional. Speech is a fundamental mode of human communication, intricate in its orchestration of neural signals and motor actions. The advent of speech neuroprosthetics promises to restore communication capabilities by bridging the gap between neural activity and speech production. Central to understanding speech neuroprosthetics is deciphering the complex neural mechanisms orchestrating speech production. From conceptualization to articulation, speech involves a finely choreographed interplay of neural circuits spanning cortical and subcortical regions. Broca's area, Wernicke's area, primary motor cortex, and supplementary motor area are among the key brain regions implicated in speech and language generation. Perhaps the best characterized so far is the motor and premotor cortices in the precentral gyrus.[169,170]

The lateral aspect of the precentral gyrus features a somatotopic organization of orofacial and unique dual laryngeal vocal-tract articulator representations.[171,172] Accumulating evidence suggests that the precentral gyrus is not only critical for executing but also for planning speech movements, a function that is commonly mis-attributed to Broca's area.[173] Kinematic analyses have revealed the encoding of dynamical, low-dimensional patterns of speech movements called "gestures."[174] One example is how the specific movements of forward tongue raise and jaw closure are coordinated to create a "d" sound at one local site in the precentral gyrus. The motor cortex contains a complete inventory of speech-related movements to create all of the sounds of a given language.[174]

The core objective of speech neuroprosthetics is to decode neural signals associated with speech production and translate them into intelligible output. Traditional approaches have focused on text output and synthetic voice. This entails harnessing BCI technologies coupled with machine learning algorithms to interpret neural activity patterns and reconstruct spoken words or phrases. The first successful demonstration of speech decoding of full words and sentences was carried out in 2021 in a man with severe paralysis after a brainstem stroke.[175] The efficacy of speech neuroprosthetics critically hinges on the development of robust decoding algorithms capable of discerning nuanced patterns of neural activity corresponding to different phonetic units. Recent strides in machine learning, particularly deep learning architectures such as RNNs, have propelled the field forward by enabling more precise and efficient decoding of speech-related brain signals.

More recent BCI approaches decode subword linguistic units, such as phonemes or characters, rather than individual words or sentences. This is a common technique in automatic speech recognition where language models—trained to capture the statistical patterns of subword units and words—are used to convert decoded phoneme or character sequences into sentences. Progress has been rapid with recent demonstrations of fast and large vocabulary decoding. The approaches leveraged RNN models trained to map an input sequence of neural activity to an output sequence of phonemes without the need for any alignment.[97,176] Language models then mapped decoded phoneme sequences into text words and sentences.

Metzger et al.[97] also used a similar approach to decode neural activity into synthesized speech; however, rather than regressing the acoustic mel-spectrogram, they decoded input sequences of neural activity into output sequences of discrete acoustic-speech units. During training, a large self-supervised audio model (HuBert[177]) was adapted to convert target waveforms (generated from a text-to-speech model) into sequences of discrete acoustic-speech units (Figure 5C). During online inference, the decoded discrete acoustic-speech unit sequences were decoded into intelligible sentence-level speech. They first used this voice-conversion approach to personalize synthesized speech for an individual with vocal-tract paralysis using speech samples recorded before the injury.[97]

While still in its nascent stages, speech neuroprosthetics hold immense potential for revolutionizing the rehabilitation and communication landscape for individuals with speech impairments.[178] Beyond restoring speech functionality, these neurotechnological innovations may pave the way for novel therapeutic interventions targeting a range of neurological disorders affecting speech and language. However, several challenges, including achieving high decoding accuracy, minimizing invasiveness, and ensuring long-term reliability, warrant further investigation to realize the full clinical potential of speech neuroprosthetics.

TOWARD CAUSAL MODELS

Overall, encoding-decoding has been a powerful mathematical framework that enables understanding perception, action, and cognition, as illustrated by several examples. But there are still challenges and opportunities ahead for better understanding the brain and building better BCI decoders.

One major thrust toward this challenge will be coming from foundation models.[89] Training large models on diverse data gives rise to stupendous capabilities, and those models can be adapted for many tasks relevant for neuroscience, such as behavioral analysis via in-context learning[179] and speech recognition.[180] As we outlined above, within neuroscience, training large-scale models on data from many studies will also open up many possibilities.

Furthermore, we are advocating for advancing models that are mechanistic (biologically plausible) and statistical (can account for neural data) at the same time (Figure 3). Encoding-decoding models lie at the heart of this pursuit. As we have summarized,

neuroscience has created many powerful statistical models of brain function for vision, motor control, and language. Beyond neuroscience, advances in machine learning have led to models with high explanatory power grounded in statistical learning theory.[181] Next, we illustrate this envisoned future via examples. What all examples have in common is that these models have good (statistical) explanatory power while being grounded via a mechanism from the lower level(s) of analysis (Figure 1B).

One excellent avenue is exploring the relationship between neural physiology and gene expression, which naturally describe a neuron at different levels, but are linked via biophysical mechanisms such as ion-channel models.[43,45] Indeed, recent work created mechanical ion-channel models that can be constrained statistically from transcriptomics data (called NPE-N), thus also linking across scales (Figure 1B). Given that (classically) neuronal cell types are characterized via their anatomical and physiological characteristics, single-cell transcriptomics continues to refine our taxonomies of cell types.[182]

Another interesting example is studying the conditions under which hexagonal firing fields emerge in RNNs that were trained to path integrate.[81] Classical mechanistic models for path integration are built on attractor models.[183] The connectivity of these models is designed to allow path integration. Remarkably, the normative models learned similar mechanisms to represent the location of animals (via center-surround connectivity), and to integrate movement updates (via shift circuits). They also provide a mathematical mechanism, grounded in pattern formation theory, why hexagonal tuning curves emerge in the abstract ANN models.[81]

How might we continue to push further? Combining statistical models and mechanistic models may require new mathematical tools. Relevant approaches are being developed in other fields such as in causal machine learning. Schölkopf and von Kügelgen[184] propose using *causality* to build systems that do not only rely on statistical dependencies of data and that generalize better to out-of-distribution examples. In particular, causal representation learning aims to extract the relevant causal variables from data and learn representations that contain not just statistical information but support interventions, reasoning, and planning. We can think of this approach as an extension of the classic latent variable analysis. Indeed, just as with latent variables, we can apply a decoder to the causal variables to extract the relevant behavioral variables. However, causal variables also come with the corresponding mechanisms and a causal graph that captures the relationships between each other. Concretely, let $X_1, \ldots, X_d$ be some experimental variables (with some of them corresponding to neural data K for different brain areas, others to sensory inputs, and motor outputs)—they possess a causal factorization:

$$p(X_1, \ldots, X_d) = \prod_{i=1}^{d} p(X_i|\mathbf{PA}_i), \qquad \text{(Equation 6)}$$

where $\mathbf{PA}_i$ denotes the direct causes of X_i and $p(X_i|\mathbf{PA}_i)$ represents the mechanisms of the model. Although learning this causal factorization from the recorded data is a challenging task, if we have some knowledge of the causal structure (which variable is caused by which) then only the mechanisms need to be inferred.[185] As the relevant behavioral variable x can be decoded from the causal variables $x = g(X_1, \ldots, X_d)$, the causal factorization can ultimately link the mechanisms with behavior. Although this approach is promising and has already generated concrete tools for machine learning (causal auto-encoders, self-supervised causal representation learning, or independent mechanism analysis), it might take further refinements to apply to the field of neuroscience.[186,187]

Another avenue to explore is merging learning dynamical systems (such as ODEs or physics-informed neural networks [PINNs], see below) with latent variable models. Historically, in latent variable models, the dynamical system has been *a priori* ascribed to the generative model.[51] One notable exception is CEBRA,[14] where the authors explore the implicit dynamics of the learned latent variables but do not explore the explicit equations that best model the latents. A potential challenge could be having several sets of differential equations (solutions), even for these theoretically identifiable models.

Here, we can turn to a recent success within the machine learning field regarding modeling dynamical systems via neural ODEs (N-ODE).[46] Yet, despite N-ODE's success in fitting temporal data for various applications like forecasting, control, and system identification, the approach lacks interpretability, as one simply gets a neural network that fits the data.

Another important direction is building knowledge into neural network models, e.g., via PINNs, which offer a different approach to bridge the gap between data-driven decoding of brain activity and understanding the underlying mechanisms.[188,189] Unlike traditional black-box models (in machine learning), PINNs can incorporate physics knowledge, such as biomechanics or neural dynamics, as constraints during training. By combining the data-fitting power of ANNs with physical constraints, PINNs are guided toward solutions that align with real-world principles. This not only enhances model interpretability but also allows PINNs to achieve good performance even with limited data. Furthermore, these physical constraints promote better generalization to unseen scenarios and offer mechanistic insights into the decoded information by revealing the potential processes at play. Demonstrating remarkable success in modeling physical systems,[190–192] and, more recently, biological processes,[193,194] PINNs hold promise for understanding neural mechanisms. For instance, recent work has developed models of muscle spindles that can accurately predict neural dynamics while also including biophysical mechanisms.[195] By bridging the gap between biomechanics and neural dynamics, this model offers a comprehensive understanding of muscle spindle function as adaptable signal processors in sensorimotor control.

Historically, differential equations provided the gold standard for understanding a system. Thus, an alternative method is symbolic regression (SR), which provides human-readable mathematical expressions directly from data and has recently shown some success in (re-)discovering natural laws.[196,197] SR has mostly been used for discovering $f(x)$ from paired observations $(x, f(x))$ but has been extended for dynamics (ODE $\dot{x} = f(x)$), even with multi-modal and noisy data.[70] Here, transformers can be trained to predict the differential equation from raw data alone.[70] This approach, called ODEFormer, reaches

state-of-the-art performance on a number of classic problems, such as Lotka-Volterra, a binocular rivalry model, a cell cycle model, or a Van der Pol oscillator.[70]

Ultimately, we would like to build meaningful "standard models" for causal understanding that link statistical firing properties of neural populations to their underlying mechanisms. A formidable goal would be to develop the neural dynamics equivalent to gold standards in physics, such as Newton's laws of motion.

CONCLUSIONS

Where will we be in 50 years?

The progress of machine learning in the last year alone makes it a daunting task to look 50 years ahead. Yet, the future certainly will be a place where we can build powerful causal models that predict across scales. Thus, we must consider how the framing of causal encoding-decoding can be natively built into foundation models for (Neuro)Science. Current efforts to build foundation models revolve around tokenizing a multitude of datatypes, leveraging GPT-4-style (inferred) training with human alignment and reinforcement learning. The goal is to build good "next prediction" models for all these datatypes in relevant downstream tasks. However, the future might dig deeper: can we build fully mechanistic models that predict, behave, or even show signs of cognitive, compositional thinking? Perhaps this hinges on what the ultimate tasks will be for GPT-42 (or such). This answer will lie in the values of society 50 years from now and our understanding of the neural mechanisms that give rise to action, perception, and cognition.

ACKNOWLEDGMENTS

The authors thank E. Fetz and the anonymous reviewers for helpful comments on the manuscript. We also acknowledge S. Sanborn, K. Franke, and X. Pitkow for helping with the design of Figure 5A and Joana Carvalho for creating the illustration. M.W.M. acknowledges Swiss NSF grant 320030_227871 and NIH BRAIN 1UF1NS126566-01, A.S.T. acknowledges NIH BRAIN 1UF1NS126566-01, E.F.C. acknowledges NIH U01DC018671, and A.M. and A.P.R. acknowledge Swiss NSF grant 310030_212516.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Wolpert, D.M., Miall, R.C., and Kawato, M. (1998). Internal models in the cerebellum. Trends Cogn. Sci. *2*, 338–347. https://doi.org/10.1016/s1364-6613(98)01221-2.
2. In this work we define "neural representations" to mean how a given neuron, or population of neurons, encodes an internal state or external stimulus.
3. Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., van Langevelde, F., Burghardt, T., et al. (2022). Perspectives in machine learning for wildlife conservation. Nat. Commun. *13*, 792. https://doi.org/10.1038/s41467-022-27980-y.
4. Wang, Q., Ding, S.L., Li, Y., Royall, J., Feng, D., Lesnar, P., Graddis, N., Naeemi, M., Facer, B., Ho, A., et al. (2020). The allen mouse brain common coordinate framework: A 3d reference atlas. Cell *181*, 936–953.e20. https://doi.org/10.1016/j.cell.2020.04.007.
5. Livet, J., Weissman, T.A., Kang, H., Draft, R.W., Lu, J., Bennis, R.A., Sanes, J.R., and Lichtman, J.W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature *450*, 56–62. https://doi.org/10.1038/nature06293.
6. Dura-Bernal, S., Suter, B.A., Gleeson, P., Cantarelli, M., Quintana, A., Rodriguez, F., Kedziora, D.J., Chadderdon, G.L., Kerr, C.C., Neymotin, S.A., et al. (2019). Netpyne, a tool for data-driven multiscale modeling of brain circuits. eLife *8*, e44494. https://doi.org/10.7554/eLife.44494.
7. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
8. Sussillo, D., Nuyujukian, P., Fan, J.M., Kao, J.C., Stavisky, S.D., Ryu, S., and Shenoy, K. (2012). A recurrent neural network for closed-loop intracortical brain–machine interface decoders. J. Neural Eng. *9*, 026027. https://doi.org/10.1088/1741-2560/9/2/026027.
9. Pandarinath, C., Gilja, V., Blabe, C.H., Nuyujukian, P., Sarma, A.A., Sorice, B.L., Eskandar, E.N., Hochberg, L.R., Henderson, J.M., and Shenoy, K.V. (2015). Neural population dynamics in human motor cortex during movements in people with als. eLife *4*, e07436. https://doi.org/10.7554/eLife.07436.
10. White, J.G., Southgate, E.L., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode caenorhabditis elegans. Philos. Trans. R. Soc. Lond. B Biol. Sci. *314*, 1–340. https://doi.org/10.1098/rstb.1986.0056.
11. Manley, J., Demas, J., Kim, H., Traub, F.M., and Vaziri, A. (2024). Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. Neuron. https://doi.org/10.1016/j.neuron.2024.02.011.
12. Stevenson, I.H., and Kording, K.P. (2011). How advances in neural recording affect data analysis. Nat. Neurosci. *14*, 139–142. https://doi.org/10.1038/nn.2731.
13. Keshtkaran, M.R., Sedler, A.R., Chowdhury, R.H., Tandon, R., Basrai, D., Nguyen, S.L., Sohn, H., Jazayeri, M., Miller, L.E., and Pandarinath, C. (2022). A large-scale neural network training framework for generalized estimation of single-trial population dynamics. Nat. Methods *19*, 1572–1577. https://doi.org/10.1038/s41592-022-01675-0.
14. Schneider, S., Lee, J.H., and Mathis, M.W. (2023a). Learnable latent embeddings for joint behavioural and neural analysis. Nature *617*, 360–368. https://doi.org/10.1038/s41586-023-06031-6.
15. Borst, A., and Theunissen, F.E. (1999). Information theory and neural coding. Nat. Neurosci. *2*, 947–957. https://doi.org/10.1038/14731.
16. Dayan, P., and Abbott, L.F. (2001). Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems (The MIT Press).
17. Pillow, J.W., Shlens, J., Paninski, L., Sher, A., Litke, A.M., Chichilnisky, E.J., and Simoncelli, E.P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature *454*, 995–999. https://doi.org/10.1038/nature07140.
18. van Gerven, M.A.J. (2017). A primer on encoding models in sensory neuroscience. J. Math. Psychol. *76*, 172–183. https://doi.org/10.1016/j.jmp.2016.06.009.
19. Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447.
20. Yamins, D.L.K., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365. https://doi.org/10.1038/nn.4244.
21. Sinz, F., Ecker, A.S., Fahey, P., Walker, E., Cobos, E., Froudarakis, E., Yatsenko, D., Pitkow, Z., Reimer, J., and Tolias, A. (2018). Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. Advances in Neural Information Processing Systems *31*.

22. Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M.C., DiCarlo, J.J., and Yamins, D.L.K. (2021). Unsupervised neural network models of the ventral visual stream. Proc. Natl. Acad. Sci. USA *118*, e2014196118. https://doi.org/10.1073/pnas.2014196118.

23. Willeke, K.F., Restivo, K., Franke, K., Nix, A.F., Cadena, S.A., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A.S., et al. (2023). Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. Preprint at bioRxiv.

24. Nayebi, A., Kong, N.C.L., Zhuang, C., Gardner, J.L., Norcia, A.M., and Yamins, D.L.K. (2023). Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. PLoS Comput. Biol. *19*, e1011506. https://doi.org/10.1371/journal.pcbi.1011506.

25. Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: A view from the width. Advances in neural information processing systems *30*.

26. Walker, E.Y., Sinz, F.H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P.G., Ecker, A.S., Reimer, J., Pitkow, X., and Tolias, A.S. (2019). Inception loops discover what excites neurons most using deep predictive models. Nat. Neurosci. *22*, 2060–2065. https://doi.org/10.1038/s41593-019-0517-x.

27. van Essen, D.C., Newsome, W.T., Maunsell, J.H.R., and Bixby, J.L. (1986). The projections from striate cortex (v1) to areas v2 and v3 in the macaque monkey: Asymmetries, areal boundaries, and patchy connections. J. Comp. Neurol. *244*, 451–480.

28. DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434. https://doi.org/10.1016/j.neuron.2012.01.010.

29. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111.

30. Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85–88. https://doi.org/10.1038/nature10754.

31. Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., and Botvinick, M.M. (2020). A distributional code for value in dopamine-based reinforcement learning. Nature *577*, 671–675. https://doi.org/10.1038/s41586-019-1924-6.

32. Watabe-Uchida, M., Zhu, L., Ogawa, S.K., Vamanrao, A., and Uchida, N. (2012). Whole-brain mapping of direct inputs to midbrain dopamine neurons. Neuron *74*, 858–873. https://doi.org/10.1016/j.neuron.2012.03.017.

33. Tian, J., Huang, R., Cohen, J.Y., Osakada, F., Kobak, D., Machens, C.K., Callaway, E.M., Uchida, N., and Watabe-Uchida, M. (2016). Distributed and mixed information in monosynaptic inputs to dopamine neurons. Neuron *91*, 1374–1389. https://doi.org/10.1016/j.neuron.2016.08.018.

34. Watabe-Uchida, M., Eshel, N., and Uchida, N. (2017). Neural circuitry of reward prediction error. Annu. Rev. Neurosci. *40*, 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109.

35. Theunissen, F.E., and Miller, J.P. (1991). Representation of sensory information in the cricket cercal sensory system. ii. information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. J. Neurophysiol. *66*, 1690–1703. https://doi.org/10.1152/jn.1991.66.5.1690.

36. Schwartz, A.B. (2016). Movement: how the brain communicates with the world. Cell *164*, 1122–1135. https://doi.org/10.1016/j.cell.2016.02.038.

37. Lange, R.D., Shivkumar, S., Chattoraj, A., and Haefner, R.M. (2023). Bayesian encoding and decoding as distinct perspectives on neural coding. Nat. Neurosci. *26*, 2063–2072. https://doi.org/10.1038/s41593-023-01458-6.

38. Zhang, K., Ginzburg, I., McNaughton, B.L., and Sejnowski, T.J. (1998). Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. J. Neurophysiol. *79*, 1017–1044. https://doi.org/10.1152/jn.1998.79.2.1017.

39. Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. ASME J. Basic Eng. *82*, 35–45. https://doi.org/10.1115/1.3662552.

40. Zhang, K., and Sejnowski, T.J. (1999). Neuronal tuning: To sharpen or broaden? Neural Comput. *11*, 75–84. https://doi.org/10.1162/089976699300016809.

41. Mathis, A., Stemmler, M.B., and Herz, A.V.M. (2015). Probable nature of higher-dimensional symmetries underlying mammalian grid-cell activity patterns. eLife *4*, e05979. https://doi.org/10.7554/eLife.05979.

42. Kriegeskorte, N., and Wei, X.X. (2021). Neural tuning and representational geometry. Nat. Rev. Neurosci. *22*, 703–718. https://doi.org/10.1038/s41583-021-00502-3.

43. Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. Bull. Math. Biol. *52*, 25–71.

44. Jones, S.A.H., Cressman, E.K., and Henriques, D.Y.P. (2010). Proprioceptive localization of the left and right hands. Exp. Brain Res. *204*, 373–383. https://doi.org/10.1007/s00221-009-2079-8.

45. Bernaerts, Y., Deistler, M., Gonçalves, P.J., Beck, J., Stimberg, M., Scala, F., Tolias, A.S., Macke, J., Kobak, D., and Berens, P. (2023). Combined statistical-mechanistic modeling links ion channel genes to physiology of cortical neuron types. Preprint at bioRxiv.

46. Chen, T.Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D.K. (2018). Neural ordinary differential equations. In Neural Information Processing Systems.

47. Chen, Z., Liu, Y., and Sun, H. (2021). Physics-informed learning of governing equations from scarce data. Nat. Commun. *12*, 6136. https://doi.org/10.1038/s41467-021-26434-1.

48. Louizos, C., Shalit, U., Mooij, J.M., Sontag, D.A., Zemel, R.S., and Welling, M. (2017). Causal effect inference with deep latent-variable models. Neural Information Processing Systems.

49. Anandkumar, A., Ge, R., Hsu, D.J., Kakade, S.M., and Telgarsky, M. (2012). Tensor decompositions for learning latent variable models. J. Mach. Learn. Res. *15*, 2773–2832.

50. Rainer, G., and Miller, E.K. (2000). Neural ensemble states in prefrontal cortex identified using a hidden markov model with a modified em algorithm. Neurocomputing *32–33*, 961–966. https://doi.org/10.1016/S0925-2312(00)00266-6.

51. Hurwitz, C.L., Kudryashova, N.N., Onken, A., and Hennig, M.H. (2021). Building population models for large-scale neural recordings: Opportunities and pitfalls. Curr. Opin. Neurobiol. *70*, 64–73. https://doi.org/10.1016/j.conb.2021.07.003.

52. Pei, F., Joel, Y., Zoltowski, D.M., Wu, A., Chowdhury, R.H., Sohn, H., O'Doherty, J.E., Shenoy, K.V., Kaufman, M.T., Churchland, M.M., et al. (2021). Neural latents benchmark '21: Evaluating latent variable models of neural population activity. Advances in Neural Information Processing Systems (NeurIPS) *34*.

53. Jha, A., Ashwood, Z.C., and Pillow, J.W. (2024). Active learning for discrete latent variable models. Neural Comput. *36*, 437–474. https://doi.org/10.1162/neco_a_01646.

54. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56. https://doi.org/10.1038/nature11129.

55. While there is some debate in neuroscience if "factors" and "variables" are semantically describing the same features, for simplicity here we mean them to be the same thing.

56. Jolliffe, I.T. (1986). Principal Component Analysis. Springer Series in Statistics, Second Edition (Springer) https://doi.org/10.1007/b98835.

57. Bell, A.J., and Sejnowski, T.J. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural Comput. *7*, 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129.

58. Altan, E., Solla, S.A., Miller, L.E., and Perreault, E.J. (2021). Estimating the dimensionality of the manifold underlying multi-electrode neural recordings. PLoS Comput. Biol. *17*, e1008591. https://doi.org/10.1371/journal.pcbi.1008591.

59. Jazayeri, M., and Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Curr. Opin. Neurobiol. *70*, 113–120. https://doi.org/10.1016/j.conb.2021.08.002.

60. Curto, C. (2017). What can topology tell us about the neural code? Bull. Amer. Math. Soc. *54*, 63–78. https://doi.org/10.1090/bull/1554.

61. Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., and Fiete, I.R. (2019). The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. Nat. Neurosci. *22*, 1512–1520. https://doi.org/10.1038/s41593-019-0460-x.

62. Gardner, R.J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N.A., Dunn, B.A., Moser, M.B., and Moser, E.I. (2022). Toroidal topology of population activity in grid cells. Nature *602*, 123–128. https://doi.org/10.1038/s41586-021-04268-7.

63. Safaie, M., Chang, J.C., Park, J., Miller, L.E., Dudman, J.T., Perich, M.G., and Gallego, J.A. (2023). Preserved neural dynamics across animals performing similar behaviour. Nature *623*, 765–771. https://doi.org/10.1038/s41586-023-06714-0.

64. Kingma, D.P., and Welling, M. (2019). An introduction to variational autoencoders. Found. Trends Mach. Learn. *12*, 307–392. https://doi.org/10.1561/2200000056.

65. Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In The 22nd International Conference on Artificial Intelligence and Statistics (PMLR), pp. 859–868.

66. Sun, J., Li, M., Chen, Z., Zhang, Y., Wang, S., and Moens, M.F. (2024). Contrast, attend and diffuse to decode high-resolution images from brain activities. Advances in Neural Information Processing Systems *36*.

67. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. Preprint at arXiv.

68. Merk, T., Köhler, R.M., Peterson, V., Lyra, L., Vanhoecke, J., Chikermane, M., Binns, T.S., Li, N., Walton, A., Bush, A., et al. (2023). Invasive neurophysiology and whole brain connectomics for neural decoding in patients with brain implants. Preprint at Research Square. https://doi.org/10.21203/rs.3.rs-3212709/v1.

69. Schneider, S., Laiz, R.G., Frey, M., and Mathis, M.W. (2023). Identifiable attribution maps using regularized contrastive learning. NeurIPS 2023 Workshop: Self-Supervised Learning–Theory and Practice.

70. d'Ascoli, S., Becker, S., Mathis, A., Schwaller, P., and Kilbertus, N. (2024). Odeformer: Symbolic regression of dynamical systems with transformers. International Conference on Learning Representations (ICLR).

71. Geisler, W.S. (2011). Contributions of ideal observer theory to vision research. Vision Res. *51*, 771–781. https://doi.org/10.1016/j.visres.2010.09.027.

72. Cao, R., and Yamins, D. (2024). Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. Cogn. Syst. Res. *85*, 101200. https://doi.org/10.1016/j.cogsys.2023.101200.

73. Barlow, H.B., and Rosenblith, W.A. (1961). Possible principles underlying the transformations of sensory messages. In Sensory Communication, *1* (MIT Press), pp. 217–234.

74. Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607–609. https://doi.org/10.1038/381607a0.

75. Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M.H., and Sabokrou, M. (2022). A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. Trans. Mach. Learn. Res. *2021*, 234.

76. Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron *98*, 630–644.e16. https://doi.org/10.1016/j.neuron.2018.03.044.

77. Sandbrink, K.J., Mamidanna, P., Michaelis, C., Bethge, M., Mathis, M.W., and Mathis, A. (2023). Contrasting action and posture coding with hierarchical deep neural network models of proprioception. eLife *12*, e81499. https://doi.org/10.7554/eLife.81499.

78. Marin Vargas, A., Bisi, A., Chiappa, A.S., Versteeg, C., Miller, L.E., and Mathis, A. (2024). Task-driven neural network models predict neural dynamics of proprioception. Cell *187*, 1745–1761.e19. https://doi.org/10.1016/j.cell.2024.02.036.

79. Haesemeyer, M., Schier, A.F., and Engert, F. (2019). Convergent temperature representations in artificial and biological neural networks. Neuron *103*, 1123–1134.e6. https://doi.org/10.1016/j.neuron.2019.07.003.

80. Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M.J., Degris, T., Modayil, J., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. Nature *557*, 429–433. https://doi.org/10.1038/s41586-018-0102-6.

81. Sorscher, B., Mel, G.C., Ocko, S.A., Giocomo, L.M., and Ganguli, S. (2023). A unified theory for the computational and mechanistic origins of grid cells. Neuron *111*, 121–137.e13. https://doi.org/10.1016/j.neuron.2022.10.003.

82. Zhuang, C., Kubilius, J., Hartmann, M.J., and Yamins, D.L. (2017). Toward goal-driven neural network models for the rodent whisker-trigeminal system. Advances in Neural Information Processing Systems *30*.

83. Trautmann, E.M., Hesse, J.K., Stine, G.M., Xia, R., Zhu, S., O'Shea, D.J., Karsh, B., Colonell, J., Lanfranchi, F.F., Vyas, S., et al. (2023). Large-scale high-density brain-wide neural recording in nonhuman primates. Preprint at bioRxiv. https://doi.org/10.1101/2023.02.01.526664.

84. Pierzchlewicz, P., Willeke, K., Nix, A., Elumalai, P., Restivo, K., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Franke, K., et al. (2024). Energy guided diffusion for generating neurally exciting images. Advances in Neural Information Processing Systems *36*.

85. Sanborn, S., Shewmake, C., Olshausen, B., and Hillar, C. (2022). Bispectral neural networks. Preprint at arXiv.

86. Marchetti, G.L., Hillar, C., Kragic, D., and Sanborn, S. (2023). Harmonics of learning: Universal fourier features emerge in invariant networks. Preprint at arXiv.

87. Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., and Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. Advances in Neural Information Processing Systems *34*, 25164–25178.

88. Franke, K., Willeke, K.F., Ponder, K., Galdamez, M., Zhou, N., Muhammad, T., Patel, S., Froudarakis, E., Reimer, J., Sinz, F.H., et al. (2022). State-dependent pupil dilation rapidly shifts visual feature selectivity. Nature *610*, 128–134. https://doi.org/10.1038/s41586-022-05270-3.

89. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. Preprint at arXiv.

90. Wang, E.Y., Fahey, P.G., Ponder, K., Ding, Z., Chang, A., Muhammad, T., Patel, S., Ding, Z., Tran, D., Fu, J., et al. (2023). Towards a foundation model of the mouse visual cortex. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.21.533548.

91. Ding, Z., Fahey, P.G., Papadopoulos, S., Wang, E.Y., Celii, B., Papadopoulos, C., Kunin, A.B., Chang, A., Fu, J., Ding, Z., et al. (2023a). Functional connectomics reveals general wiring rule in mouse visual cortex. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.13.531369.

92. MICrONS Consortium, Bae, J.A., Baptiste, M., Bishop, C.A., Bodor, A.L., Brittain, D., Buchanan, J., Bumbarger, D.J., Castro, M.A., Celii, B., et al. (2021). Functional connectomics spanning multiple areas of mouse visual cortex. Preprint at bioRxiv.

93. Lappalainen, J.K., Tschopp, F.D., Prakhya, S., McGill, M., Nern, A., Shinomiya, K., Takemura, S.-Y., Gruntman, E., Macke, J.H., and Turaga, S.C. (2024). Connectome-constrained deep mechanistic networks predict neural responses across the fly visual system at single-neuron resolution. Nature. https://doi.org/10.1038/s41586-024-07939-3.

94. Bashivan, P., Kar, K., and DiCarlo, J.J. (2019). Neural population control via deep image synthesis. Science *364*, eaav9436. https://doi.org/10.1126/science.aav9436.

95. Ponce, C.R., Xiao, W., Schade, P.F., Hartmann, T.S., Kreiman, G., and Livingstone, M.S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell *177*, 999–1009.e10. https://doi.org/10.1016/j.cell.2019.04.005.

96. Chen, Z., Qing, J., Xiang, T., Yue, W.L., and Zhou, J.H. (2023). Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22710–22720. https://doi.org/10.1109/CVPR52729.2023.02175.

97. Metzger, S.L., Littlejohn, K.T., Silva, A.B., Moses, D.A., Seaton, M.P., Wang, R., Dougherty, M.E., Liu, J.R., Wu, P., Berger, M.A., et al. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature *620*, 1037–1046. https://doi.org/10.1038/s41586-023-06443-4.

98. Zhang, Y., Tiňo, P., Leonardis, A., and Tang, K. (2021). A survey on neural network interpretability. IEEE Trans. Emerg. Top. Comput. Intell. *5*, 726–742. https://doi.org/10.1109/TETCI.2021.3100641.

99. Ding, Z., Tran, D.T., Ponder, K., Cobos, E., Ding, Z., Fahey, P.G., Wang, E., Muhammad, T., Fu, J., Cadena, S.A., et al. (2023). Bipartite invariance in mouse primary visual cortex. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.15.532836.

100. Fu, J., Shrinivasan, S., Ponder, K., Muhammad, T., Ding, Z., Wang, E., Ding, Z., Tran, D.T., Fahey, P.G., Papadopoulos, S., et al. (2024). Pattern completion and disruption characterize contextual modulation in mouse visual cortex. Preprint at bioRxiv.

101. Xia, W., de Charette, R., Oztireli, C., and Xue, J.H. (2024). Dream: Visual decoding from reversing human visual system. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 8211–8220. https://doi.org/10.1109/WACV57701.2024.00804.

102. Luo, A., Henderson, M., Wehbe, L., and Tarr, M. (2024). Brain diffusion for visual exploration: Cortical discovery using large scale generative models. Advances in Neural Information Processing Systems *36*.

103. Shirakawa, K., Nagano, Y., Tanaka, M., Aoki, S.C., Majima, K., Muraki, Y., and Kamitani, Y. (2024). Spurious reconstruction from brain activity. Preprint at arXiv.

104. Peixoto, D., Verhein, J.R., Kiani, R., Kao, J.C., Nuyujukian, P., Chandrasekaran, C., Brown, J.R., Fong, S., Ryu, S.I., Shenoy, K.V., et al. (2021). Decoding and perturbing decision states in real time. Nature *591*, 604–609. https://doi.org/10.1038/s41586-020-03181-9.

105. Fetz, E.E. (1992). Are movement parameters recognizably coded in the activity of single neurons? Behav. Brain Sci. *15*, 679–690.

106. Evarts, E.V. (1968). Relation of pyramidal tract activity to force exerted during voluntary movement. J. Neurophysiol. *31*, 14–27. https://doi.org/10.1152/jn.1968.31.1.14.

107. Thach, W.T. (1978). Correlation of neural discharge with pattern and force of muscular activity, joint position, and direction of intended next movement in motor cortex and cerebellum. J. Neurophysiol. *41*, 654–676. https://doi.org/10.1152/jn.1978.41.3.654.

108. Georgopoulos, A.P., Schwartz, A.B., and Kettner, R.E. (1986). Neuronal population coding of movement direction. Science *233*, 1416–1419. https://doi.org/10.1126/science.3749885.

109. Georgopoulos, A.P., Kettner, R.E., and Schwartz, A.B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. ii. coding of the direction of movement by a neuronal population. J. Neurosci. *8*, 2928–2937. https://doi.org/10.1523/JNEUROSCI.08-08-02928.1988.

110. Kalaska, J.F., Cohen, D.A., Hyde, M.L., and Prud'homme, M. (1989). A comparison of movement direction-related versus load direction-related activity in primate motor cortex, using a two-dimensional reaching task. J. Neurosci. *9*, 2080–2102. https://doi.org/10.1523/JNEUROSCI.09-06-02080.1989.

111. Moran, D.W., and Schwartz, A.B. (1999). Motor cortical representation of speed and direction during reaching. J. Neurophysiol. *82*, 2676–2692. https://doi.org/10.1152/jn.1999.82.5.2676.

112. Wessberg, J., Stambaugh, C.R., Kralik, J.D., Beck, P.D., Laubach, M., Chapin, J.K., Kim, J., Biggs, S.J., Srinivasan, M.A., and Nicolelis, M.A.L. (2000). Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. Nature *408*, 361–365. https://doi.org/10.1038/35042582.

113. Serruya, M.D., Hatsopoulos, N.G., Paninski, L., Fellows, M.R., and Donoghue, J.P. (2002). Instant neural control of a movement signal. Nature *416*, 141–142. https://doi.org/10.1038/416141a.

114. Taylor, D.M., Tillery, S.I., and Schwartz, A.B. (2002). Direct cortical control of 3d neuroprosthetic devices. Science *296*, 1829–1832. https://doi.org/10.1126/science.1070291.

115. Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., and Donoghue, J.P. (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature *442*, 164–171. https://doi.org/10.1038/nature04970.

116. Moritz, C.T., Perlmutter, S.I., and Fetz, E.E. (2008). Direct control of paralysed muscles by cortical neurons. Nature *456*, 639–642. https://doi.org/10.1038/nature07418.

117. Carmena, J.M., Lebedev, M.A., Crist, R.E., O'Doherty, J.E., Santucci, D.M., Dimitrov, D.F., Patil, P.G., Henriquez, C.S., and Nicolelis, M.A.L. (2003). Learning to control a brain–machine interface for reaching and grasping by primates. PLoS Biol. *1*, E42. https://doi.org/10.1371/journal.pbio.0000042.

118. Velliste, M., Perel, S., Spalding, M.C., Whitford, A.S., and Schwartz, A.B. (2008). Cortical control of a prosthetic arm for self-feeding. Nature *453*, 1098–1101. https://doi.org/10.1038/nature06996.

119. Hochberg, L.R., Bacher, D., Jarosiewicz, B., Masse, N.Y., Simeral, J.D., Vogel, J., Haddadin, S., Liu, J., Cash, S.S., van der Smagt, P., et al. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature *485*, 372–375. https://doi.org/10.1038/nature11076.

120. Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J.C., Velliste, M., Boninger, M.L., and Schwartz, A.B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. Lancet *381*, 557–564. https://doi.org/10.1016/S0140-6736(12)61816-9.

121. Shokur, S., Mazzoni, A., Schiavone, G., Weber, D.J., and Micera, S. (2021). A modular strategy for next-generation upper-limb sensory-motor neuroprostheses. Med. *2*, 912–937. https://doi.org/10.1016/j.medj.2021.05.002.

122. Fetz, E.E. (2007). Volitional control of neural activity: implications for brain–computer interfaces. J. Physiol. *579*, 571–579. https://doi.org/10.1113/jphysiol.2006.127142.

123. Schwartz, A.B., Cui, X.T., Weber, D.J., and Moran, D.W. (2006). Brain-controlled interfaces: movement restoration with neural prosthetics. Neuron *52*, 205–220. https://doi.org/10.1016/j.neuron.2006.09.019.

124. Wu, W., Gao, Y., Bienenstock, E., Donoghue, J.P., and Black, M.J. (2006). Bayesian population decoding of motor cortical activity using a kalman filter. Neural Comput. *18*, 80–118. https://doi.org/10.1162/089976606774841585.

125. Mulliken, G.H., Musallam, S., and Andersen, R.A. (2008). Decoding trajectories from posterior parietal cortex ensembles. J. Neurosci. *28*, 12913–12926. https://doi.org/10.1523/JNEUROSCI.1463-08.2008.

126. Kim, S.P., Simeral, J.D., Hochberg, L.R., Donoghue, J.P., and Black, M.J. (2008). Neural control of computer cursor velocity by decoding motor cortical spiking activity in humans with tetraplegia. J. Neural Eng. *5*, 455–476. https://doi.org/10.1088/1741-2560/5/4/010.

127. Shenoy, K.V., Sahani, M., and Churchland, M.M. (2013). Cortical control of arm movements: a dynamical systems perspective. Annu. Rev. Neurosci. *36*, 337–359. https://doi.org/10.1146/annurev-neuro-062111-150509.

128. Vyas, S., Golub, M.D., Sussillo, D., and Shenoy, K.V. (2020). Computation through neural population dynamics. Annu. Rev. Neurosci. *43*, 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115.

129. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Ryu, S.I., and Shenoy, K.V. (2010). Cortical preparatory activity: representation of movement or first cog in a dynamical machine? Neuron *68*, 387–400. https://doi.org/10.1016/j.neuron.2010.09.015.

130. Churchland, M.M., and Shenoy, K.V. (2007). Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. J. Neurophysiol. *97*, 4235–4257. https://doi.org/10.1152/jn.00095.2007.

131. Rickert, J., Riehle, A., Aertsen, A., Rotter, S., and Nawrot, M.P. (2009). Dynamic encoding of movement direction in motor cortical neurons. J. Neurosci. *29*, 13870–13882. https://doi.org/10.1523/JNEUROSCI.5441-08.2009.

132. Churchland, M.M., and Shenoy, K.V. (2024). Preparatory activity and the expansive null-space. Nat. Rev. Neurosci. *25*, 213–236. https://doi.org/10.1038/s41583-024-00796-z.

133. Kaufman, M.T., Churchland, M.M., Ryu, S.I., and Shenoy, K.V. (2014). Cortical activity in the null space: permitting preparation without movement. Nat. Neurosci. *17*, 440–448. https://doi.org/10.1038/nn.3643.

134. Elsayed, G.F., Lara, A.H., Kaufman, M.T., Churchland, M.M., and Cunningham, J.P. (2016). Reorganization between preparatory and movement population responses in motor cortex. Nat. Commun. *7*, 13239. https://doi.org/10.1038/ncomms13239.

135. Russo, A.A., Bittner, S.R., Perkins, S.M., Seely, J.S., London, B.M., Lara, A.H., Miri, A., Marshall, N.J., Kohn, A., Jessell, T.M., et al. (2018). Motor cortex embeds muscle-like commands in an untangled population response. Neuron *97*, 953–966.e8. https://doi.org/10.1016/j.neuron.2018.01.004.

136. Michaels, J.A., Dann, B., and Scherberger, H. (2016). Neural population dynamics during reaching are better explained by a dynamical system than representational tuning. PLoS Comput. Biol. *12*, e1005175. https://doi.org/10.1371/journal.pcbi.1005175.

137. Hennequin, G., Vogels, T.P., and Gerstner, W. (2014). Optimal control of transient dynamics in balanced networks supports generation of complex movements. Neuron *82*, 1394–1406. https://doi.org/10.1016/j.neuron.2014.04.045.

138. Sussillo, D., Churchland, M.M., Kaufman, M.T., and Shenoy, K.V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. Nat. Neurosci. *18*, 1025–1033. https://doi.org/10.1038/nn.4042.

139. Michaels, J.A., Schaffelhofer, S., Agudelo-Toro, A., and Scherberger, H. (2020). A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. Proc. Natl. Acad. Sci. USA *117*, 32124–32135. https://doi.org/10.1073/pnas.2005087117.

140. Russo, A.A., Khajeh, R., Bittner, S.R., Perkins, S.M., Cunningham, J.P., Abbott, L.F., and Churchland, M.M. (2020). Neural trajectories in the supplementary motor area and motor cortex exhibit distinct geometries, compatible with different classes of computation. Neuron *107*, 745–758.e6. https://doi.org/10.1016/j.neuron.2020.05.020.

141. Saxena, S., Russo, A.A., Cunningham, J., and Churchland, M.M. (2022). Motor cortex activity across movement speeds is predicted by network-level strategies for generating muscle activity. eLife *11*, e67620. https://doi.org/10.7554/eLife.67620.

142. Sussillo, D., Stavisky, S.D., Kao, J.C., Ryu, S.I., and Shenoy, K.V. (2016). Making brain–machine interfaces robust to future neural variability. Nat. Commun. *7*, 13749. https://doi.org/10.1038/ncomms13749.

143. Pandarinath, C., O'Shea, D.J., Collins, J., Józefowicz, R., Stavisky, S.D., Kao, J.C., Trautmann, E.M., Kaufman, M.T., Ryu, S.I., Hochberg, L.R., et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods *15*, 805–815. https://doi.org/10.1038/s41592-018-0109-9.

144. DePasquale, B., Sussillo, D., Abbott, L.F., and Churchland, M.M. (2023). The centrality of population-level factors to network computation is demonstrated by a versatile approach for training spiking networks. Neuron *111*, 631–649.e10. https://doi.org/10.1016/j.neuron.2022.12.007.

145. Kao, J.C., Nuyujukian, P., Ryu, S.I., Churchland, M.M., Cunningham, J.P., and Shenoy, K.V. (2015). Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. Nat. Commun. *6*, 7759. https://doi.org/10.1038/ncomms8759.

146. Pandarinath, C., Nuyujukian, P., Blabe, C.H., Sorice, B.L., Saab, J., Willett, F.R., Hochberg, L.R., Shenoy, K.V., and Henderson, J.M. (2017). High performance communication by people with paralysis using an intracortical brain-computer interface. eLife *6*, e18554. https://doi.org/10.7554/eLife.18554.

147. Benabid, A.L., Costecalde, T., Eliseyev, A., Charvet, G., Verney, A., Karakas, S., Foerster, M., Lambert, A., Morinière, B., Abroug, N., et al. (2019). An exoskeleton controlled by an epidural wireless brain–machine interface in a tetraplegic patient: a proof-of-concept demonstration. Lancet Neurol. *18*, 1112–1122. https://doi.org/10.1016/S1474-4422(19)30321-7.

148. Joel, Y., Collinger, J., Wehbe, L., and Gaunt, R. (2024). Neural data transformer 2: multi-context pretraining for neural spiking activity. Advances in Neural Information Processing Systems *36*.

149. Jackson, A., Mavoori, J., and Fetz, E.E. (2006). Long-term motor cortex plasticity induced by an electronic neural implant. Nature *444*, 56–60. https://doi.org/10.1038/nature05226.

150. Sadtler, P.T., Quick, K.M., Golub, M.D., Chase, S.M., Ryu, S.I., Tyler-Kabara, E.C., Yu, B.M., and Batista, A.P. (2014). Neural constraints on learning. Nature *512*, 423–426. https://doi.org/10.1038/nature13665.

151. Orsborn, A.L., Moorman, H.G., Overduin, S.A., Shanechi, M.M., Dimitrov, D.F., and Carmena, J.M. (2014). Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control. Neuron *82*, 1380–1393. https://doi.org/10.1016/j.neuron.2014.04.048.

152. Jarosiewicz, B., Chase, S.M., Fraser, G.W., Velliste, M., Kass, R.E., and Schwartz, A.B. (2008). Functional network reorganization during learning in a brain-computer interface paradigm. Proc. Natl. Acad. Sci. USA *105*, 19486–19491. https://doi.org/10.1073/pnas.0808113105.

153. Todorov, E. (2000). Direct cortical control of muscle activation in voluntary arm movements: a model. Nat. Neurosci. *3*, 391–398. https://doi.org/10.1038/73964.

154. Ajemian, R., Green, A., Bullock, D., Sergio, L., Kalaska, J., and Grossberg, S. (2008). Assessing the function of motor cortex: single-neuron models of how neural response is modulated by limb biomechanics. Neuron *58*, 414–428. https://doi.org/10.1016/j.neuron.2008.02.033.

155. Lillicrap, T.P., and Scott, S.H. (2013). Preference distributions of primary motor cortex neurons reflect control solutions optimized for limb biomechanics. Neuron *77*, 168–179. https://doi.org/10.1016/j.neuron.2012.10.041.

156. Loeb, G.E. (2021). Learning to use muscles. J. Hum. Kinet. *76*, 9–33. https://doi.org/10.2478/hukin-2020-0084.

157. Gorko, B., Siwanowicz, I., Close, K., Christoforou, C., Hibbard, K.L., Kabra, M., Lee, A., Park, J.-Y., Li, S.Y., Chen, A.B., et al. (2024). Motor neurons generate pose-targeted movements via proprioceptive sculpting. Nature *628*, 596–603. https://doi.org/10.1038/s41586-024-07222-5.

158. Flesher, S.N., Downey, J.E., Weiss, J.M., Hughes, C.L., Herrera, A.J., Tyler-Kabara, E.C., Boninger, M.L., Collinger, J.L., and Gaunt, R.A. (2021). A brain-computer interface that evokes tactile sensations improves robotic arm control. Science *372*, 831–836. https://doi.org/10.1126/science.abd0380.

159. Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In 2012 IEEE/RSJ (IEEE Publications), pp. 5026–5033. https://doi.org/10.1109/IROS.2012.6386109.

160. Caggiano, V., Wang, H., Durandau, G., Sartori, M., and Kumar, V. (2022). MyoSuite – a contact-rich simulation suite for musculoskeletal motor control. In Proceedings of The 4th Annual Learning for Dynamics and Control Conference.

161. Wang-Chen, S., Stimpfling, V.A., Özdil, P.G., Genoud, L., Hurtak, F., and Ramdya, P. (2023). Neuromechfly 2.0, a framework for simulating embodied sensorimotor control in adult drosophila. Preprint at bioRxiv.

162. Peng, X.B., Abbeel, P., Levine, S., and Van de Panne, M. (2018). Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Trans. Graph. *37*, 1–14. https://doi.org/10.1145/3197517.3201311.

163. Schumacher, P., Haeufle, D., Büchler, D., Schmitt, S., and Martius, G. (2022). Dep-rl: Embodied exploration for reinforcement learning in overactuated and musculoskeletal systems. In The Eleventh International Conference on Learning Representations.

164. Chiappa, A.S., Marin Vargas, A., Huang, A.Z., and Mathis, A. (2023). Latent exploration for reinforcement learning. Advances in Neural Information Processing Systems.

165. Caggiano, V., Durandau, G., Wang, H., Chiappa, A., Mathis, A., Tano, P., Patel, N., Pouget, A., Schumacher, P., Martius, G., et al. (2022). Myochallenge 2022: Learning contact-rich manipulation using a musculoskeletal hand. In NeurIPS 2022 Competition Track. Proceedings of the Machine Learning Research, pp. 233–250.

166. Chiappa, A.S., Tano, P., Patel, N., Ingster, A., Pouget, A., and Mathis, A. (2024). Acquiring musculoskeletal skills with curriculum-based reinforcement learning. Preprint at bioRxiv.

167. Almani, M.N., Lazzari, J., Chacon, A., and Saxena, S. (2024). μsim: A goal-driven framework for elucidating the neural control of movement through musculoskeletal modeling. Preprint at bioRxiv.

168. Melis, J.M., Siwanowicz, I., and Dickinson, M.H. (2024). Machine learning reveals the control mechanics of an insect wing hinge. Nature *628*, 795–803. https://doi.org/10.1038/s41586-024-07293-4.

169. Hickok, G. (2012). Computational neuroanatomy of speech production. Nat. Rev. Neurosci. *13*, 135–145. https://doi.org/10.1038/nrn3158.

170. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402. https://doi.org/10.1038/nrn2113.

171. Bouchard, K.E., Mesgarani, N., Johnson, K., and Chang, E.F. (2013). Functional organization of human sensorimotor cortex for speech articulation. Nature *495*, 327–332. https://doi.org/10.1038/nature11911.

172. Dichter, B.K., Breshears, J.D., Leonard, M.K., and Chang, E.F. (2018). The control of vocal pitch in human laryngeal motor cortex. Cell *174*, 21–31.e9. https://doi.org/10.1016/j.cell.2018.05.016.

173. Silva, A.B., Liu, J.R., Zhao, L., Levy, D.F., Scott, T.L., and Chang, E.F. (2022). A neurosurgical functional dissection of the middle precentral gyrus during speech production. J. Neurosci. *42*, 8416–8426. https://doi.org/10.1523/JNEUROSCI.1614-22.2022.

174. Chartier, J., Anumanchipalli, G.K., Johnson, K., and Chang, E.F. (2018). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. Neuron *98*, 1042–1054.e4. https://doi.org/10.1016/j.neuron.2018.04.031.

175. Moses, D.A., Metzger, S.L., Liu, J.R., Anumanchipalli, G.K., Makin, J.G., Sun, P.F., Chartier, J., Dougherty, M.E., Liu, P.M., Abrams, G.M., et al. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. N. Engl. J. Med. *385*, 217–227. https://doi.org/10.1056/NEJMoa2027540.

176. Willett, F.R., Kunz, E.M., Fan, C., Avansino, D.T., Wilson, G.H., Choi, E.Y., Kamdar, F., Hochberg, L.R., Druckmann, S., Shenoy, K.V., et al. (2023). A high-performance speech neuroprosthesis. Nature *620*, 1031–1036. https://doi.org/10.1038/s41586-023-06377-x.

177. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. *29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291.

178. Card, N.S., Wairagkar, M., Iacobacci, C., Hou, X., Singer-Clark, T., Willett, F.R., Kunz, E.M., Fan, C., Vahdati Nia, M.V., Deo, D.R., et al. (2024). An accurate and rapidly calibrating speech neuroprosthesis. N. Engl. J. Med. *391*, 609–618. https://doi.org/10.1056/NEJMoa2314132.

179. Shaokai, Y., Lauer, J., Zhou, M., Mathis, A., and Mathis, M.W. (2023). Amadeusgpt: a natural language interface for interactive animal behavioral analysis. Thirty-seventh Conference on Neural Information Processing Systems.

180. Ling, S., Hu, Y., Qian, S., Ye, G., Qian, Y., Gong, Y., Lin, E., and Zeng, M. (2024). Adapting large language model with speech for fully formatted end-to-end speech recognition. In ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE Publications), pp. 11046–11050. https://doi.org/10.1109/ICASSP48485.2024.10448204.

181. Von Luxburg, U., and Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In Handbook of the History of Logic, *10* (Elsevier), pp. 651–706. https://doi.org/10.1016/B978-0-444-52936-7.50016-1.

182. Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R.S., Aldridge, A.I., Ament, S.A., Bartlett, A., Behrens, M.M., Van den Berge, K., et al. (2021). A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature *598*, 103–110. https://doi.org/10.1038/s41586-021-03500-8.

183. Khona, M., and Fiete, I.R. (2022). Attractor and integrator networks in the brain. Nat. Rev. Neurosci. *23*, 744–766. https://doi.org/10.1038/s41583-022-00642-0.

184. Schölkopf, B., and von Kügelgen, J. (2022). From statistical to causal learning. In Proceedings of the International Congress of Mathematicians, pp. 5540–5593. https://doi.org/10.4171/icm2022/173.

185. Wendong, L., Kekić, A., von Kügelgen, J., Buchholz, S., Besserve, M., Gresele, L., and Schölkopf, B. (2024). Causal component analysis. Advances in Neural Information Processing Systems *36*.

186. Siddiqi, S.H., Kording, K.P., Parvizi, J., and Fox, M.D. (2022). Causal mapping of human brain function. Nat. Rev. Neurosci. *23*, 361–375. https://doi.org/10.1038/s41583-022-00583-8.

187. Ross, L.N., and Bassett, D.S. (2024). Causation in neuroscience: keeping mechanism meaningful. Nat. Rev. Neurosci. *25*, 81–90. https://doi.org/10.1038/s41583-023-00778-7.

188. Raissi, M., Perdikaris, P., and Karniadakis, G.E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comp. Phys. *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

189. Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. Nat. Rev. Phys. *3*, 422–440. https://doi.org/10.1038/s42254-021-00314-5.

190. Cai, S., Mao, Z., Wang, Z., Yin, M., and Karniadakis, G.E. (2021a). Physics-informed neural networks (pinns) for fluid mechanics: A review. Acta Mech. Sin. *37*, 1727–1738. https://doi.org/10.1007/s10409-021-01148-1.

191. Cai, S., Wang, Z., Wang, S., Perdikaris, P., and Karniadakis, G.E. (2021b). Physics-informed neural networks for heat transfer problems. J. Heat Transf. *143*, 060801. https://doi.org/10.1115/1.4050542.

192. Calicchia, M.A., Mittal, R., Seo, J.H., and Ni, R. (2023). Reconstructing the pressure field around swimming fish using a physics-informed neural network. J. Exp. Biol. *226*, jeb244983. https://doi.org/10.1242/jeb.244983.

193. Lagergren, J.H., Nardini, J.T., Baker, R.E., Simpson, M.J., and Flores, K.B. (2020). Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. PLoS Comput. Biol. *16*, e1008462. https://doi.org/10.1371/journal.pcbi.1008462.

194. Sel, K., Mohammadi, A., Pettigrew, R.I., and Jafari, R. (2023). Physics-informed neural networks for modeling physiological time series for cuffless blood pressure estimation. NPJ Digit. Med. *6*, 110. https://doi.org/10.1038/s41746-023-00853-4.

195. Perez Rotondo, A., Marin Vargas, A., Dimitriou, M., and Mathis, A. (2024). Modeling Sensorimotor Processing with Physics-Informed Neural Networks. Preprint at bioRxiv. https://doi.org/10.1101/2024.09.14.613030.

196. Aréchiga, N., Chen, F., Chen, Y.Y., Zhang, Y., Iliev, R., Toyoda, H., and Lyons, K. (2021). Accelerating understanding of scientific experiments with end to end symbolic regression. Preprint at arXiv.

197. Udrescu, S.M., and Tegmark, M. (2020). Symbolic pregression: Discovering physical laws from raw distorted video. Phys. Rev. E *103*, 043307. https://doi.org/10.1103/PhysRevE.103.043307.