AI, Robot Neuroscientist: Reimagining Hypothesis Generation

Jiaqi Shang[‡]
Program in Neuroscience
Harvard Medical School
Boston, MA 02115
jiaqishang@g.harvard.edu

Will Xiao[‡]
Department of Neurobiology
Harvard Medical School
Boston, MA 02115
xiaow@g.harvard.edu

Abstract

Neuroscience has long relied on human-conceived hypotheses, yet the brain's complexity fundamentally challenges this epistemology. Modern technologies and the large-scale data collection they enable throw this challenge into sharp relief. We champion the potential of AI for neuroscience exploration. We highlight both implicit, 'uninterpretable' models as aids in hypothesis formulation and symbolic regression for explicit hypothesis generation. For researchers from non-neuroscience backgrounds, we discuss domain-specific considerations in integrating AI into neuroscience research. By spotlighting the underexplored avenues for AI to accelerate neuroscience, we aim to induce both communities toward these exciting research opportunities.

1 Introduction

The brain is a complex matter, about which it is challenging to formulate hypotheses. Hypothesis generation in neuroscience has depended on heuristics that range from adapting psychology concepts [1], to simplified neuron [2, 3] and network models [4], anatomical localization, and serendipity [5].

Meanwhile, neuroscience is data-rich—current technology allows for simultaneously recording tens of thousands of neurons, a number that continues to grow in a Moore-like law [6]. Current analyses have just begun to make sense of these data, often with linear methods [7, 8, 9]. While linear decomposition can already detect on the order of 10^2 dimensions in the activity of 10^4 neurons, the interpretation of these dimensions is often further restricted to only a handful of task variables [10, 9]. Finally, much of high-dimensional neural activity is ascribed to 'mixed selectivity' that resists unraveling into simpler principles [11, 12].

The combination of complexity and data abundance makes neuroscience especially fertile ground for AI-driven hypothesis discovery. AI methods are uniquely suited to detect intricate patterns hidden in neural data. AI research has developed a plethora of tools suited for specific neuroscience questions. Much of this potential remains untapped, as neuroscience today still applies AI tools sparsely. Here, we review current statistical and AI methods for analyzing neural data and identify promising directions for future work.

1.1 Related perspectives

Many recent perspectives highlighted the potential for AI to inform neuroscience [13, 14, 15]. These perspectives propose AI models as embodiments of pre-existing hypotheses to be verified on neuroscience data. Examples of such hypotheses include temporal-difference learning, external

[‡]Equal contribution.

NeurIPS 2023 AI for Science Workshop.

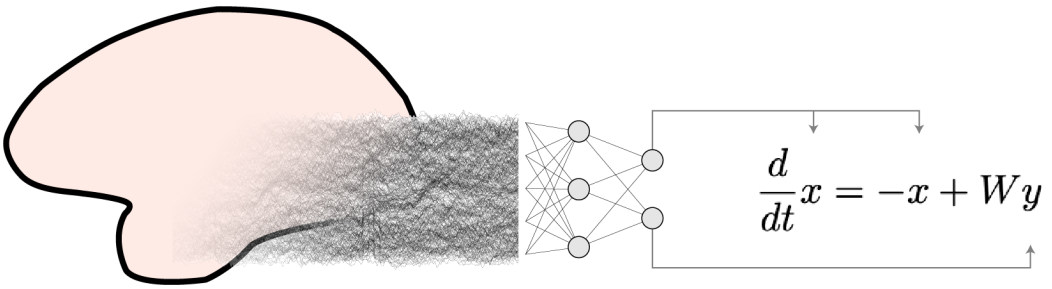

Figure 1: AI can help model and discover hypotheses from raw neural data.

memory, meta-reinforcement learning [13], task-optimized convolutional neural networks (CNNs) [16], and the learning algorithm, which is further analyzed into three components: the learning rule, objective function, and architecture [14]. In other words, these perspectives advocate for a systems identification approach, exemplified in work such as [17]. Comparing models and the brain to infer shared design principles has yielded some insights and also met challenges [17, 18].

We put a different emphasis. Instead of building hypotheses into AI and interpreting the model-brain match in those terms, we advocate for AI as a tool for hypothesis discovery from neural data (Figure 1). Our perspective is closer in spirit to [19], while we focus on the potential for AI to deliver conceptual neuroscience insights. Although we briefly discuss the potential of AI methods for neural data preprocessing, we prioritize the analysis of functional neural dynamics during perception, cognition, and behavior. Finally, we focus on large-scale electrical and optical physiology data because they best represent the rich, complex patterns that AI thrives on while promising further scaling.

2 Current methods for analyzing large neural data

2.1 Dimensionality reduction

Dimensionality reduction techniques are commonly used to analyze high-dimensional neural data. These methods transform the data into a lower-dimensional space while preserving the inherent variance. For instance, Principal Component Analysis (PCA), when applied to neural data from the mouse visual cortex [7], reveals a significant correlation between the first principal component and animal arousal indicators such as running, whisking, and pupil dilation. Other techniques, including Factor Analysis (FA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Independent Component Analysis (ICA) are also frequently employed [20, 21, 22]. After the reduced dimensions are extracted, they can be regressed against observable experimental variables to infer underlying neural mechanisms. A strong correlation between an activity dimension and its predictions from certain variables indicates the dimension contains information about those variables.

2.2 Dynamical models

A notable limitation of methods like PCA is that they treat each time point as independent. Temporally patterned behaviors, such as reaching, likely involve neural mechanisms with a temporal structure. Several methods have been proposed to explicitly address the temporal structure in neural mechanisms. They model the temporal evolution of the neural dynamics using tools such as Gaussian processes [23, 24], dynamical systems [25, 26, 27, 28, 29] and Recurrent Neural Networks (RNNs) [30, 31]. For example, Latent Factor Analysis via Dynamical Systems (LFADS) [30] uses an RNN to reconstruct recorded data across trials. After optimizing, the RNN can accurately predict behavioral variables, such as reaching directions for previously unseen neural data.

2.3 Latent variable models

Neural mechanisms often encompass factors not directly observable. Many complex cognitive processes involve multiple stages of input processing before behavioral outputs. To uncover these hidden factors, researchers have turned to Autoencoders, a type of artificial neural network (ANN). Autoencoders compress neural data through an encoder to produce a latent representation; this

representation is then used by a decoder to reconstruct the original data. This latent representation thus captures hidden features reflective of the intrinsic neural mechanisms. To illustrate, pi-VAE [32] employs the variational autoencoder to analyze hippocampal recordings from mice during a spatial navigation task. The model effectively extracts latent factors that separate the spatial and temporal information related to navigation. While Autoencoders can capture intricate mappings from neural data to latent representations, challenges persist in interpreting the extracted latent representation. Notably, interpretation is often confined to preconceived hypotheses, such as correlating with sensory inputs or motor outputs. Such a constraint can limit the breadth of hypothesis exploration, potentially missing out on novel neural mechanisms that operate beyond these predefined parameters.

2.4 Encoding models for visual neurons

Encoding models that allow image-computable predictions of visual neuron responses are among the first applications of AI in neuroscience [33, 34] and remain an active research direction. These models comprise a nonlinear core that is often trained without neural data and a linear mapping that is fitted to transform the model-core output to neural activity. Model cores trained without neural data thus do not extract any latent neural activity structure, although models that learn end-to-end from neural data pooled across many animals may represent implicit structures. Encoding models have been used, as advocated in related perspectives (Section 1.1), to instantiate preconceived conceptual hypotheses such as the convolution motif, learning objective and rule [14], and topological cortical organization [35, 36, 37].

2.5 Needs unmet by current methods

Current methods focus on discovering latent factors that have predictive power within a given dataset. The latent factors are sometimes given interpretations by correlation to external observables like stimuli and behaviors. However, these methods do not introduce novel concepts or rules that can extrapolate across datasets. In contrast, any scientific theories a 'robot neuroscientist' can produce should ultimately describe properties of the system (not merely of the data) that extrapolate to a wide range of situations unseen during model fitting [19].

3 'Uninterpretable' models as aids in hypothesis formulation

Connectionist AI models are opaque by default: An active research field of explainable AI is devoted to interpreting models. However, even 'uninterpretable' models, when coupled with an appropriate scientific framework, can help scientists form hypotheses. Here we use 'uninterpretable' to refer to models that are not expressly designed to be interpretable. Even though AI models are mathematically well-defined and explainable in this sense, such models do not explicitly represent conceptual insights. Below, we discuss two ways in which uninterpretable models can aid scientific discovery: feature attribution and factorizing complex interdependencies in data.

3.1 Feature attribution

Feature attribution on performance-optimized models can reveal relations between features in the data, thereby helping human scientists form hypotheses (Figure 2, left). Davies et al. (2021) used this approach to help mathematicians formulate two conjectures, one in knot theory and the other in representation theory, which were subsequently proven manually. Take the knot theory case for example. Davies and colleagues searched for unknown relationships between the *geometric invariants* and *algebraic invariants* of a knot by training supervised ANNs to predict the latter from the former using a synthetic dataset. They found that a knot K's signature $\sigma(K)$, an algebraic invariant, can be predicted above-chance from the geometric invariants of K and used feature attribution to identify the top three contributors: the meridional translation μ (real and imaginary parts) and the longitudinal translation λ. Human mathematicians subsequently conjectured and proved a novel theorem involving these quantities. (Notably, other geometric invariants entered the final theorem but had much lower attribution scores than the top three.) In the representation theory example, neural networks helped identify features of the Bruhat interval, a large directed graph too cumbersome for easy intuition, that ultimately led human mathematicians to a theorem and its (conjectured) generalization.

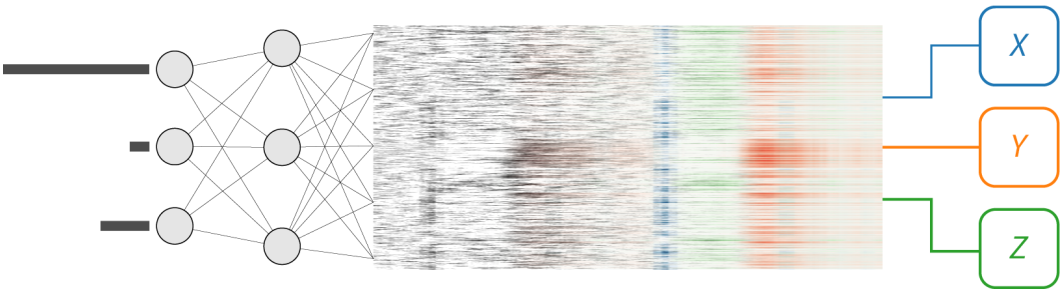

Figure 2: AI models can help evaluate feature contributions (left) to neural activity (center) or represent factors in the data (right).

In the feature attribution approach we propose, the AI model serves an analogous function to a generalized linear model (GLM). Another analog, symbolic regression [38, 39], promises to directly infer mathematical formulae that describe the laws governing a system; we discuss it in Section 4. However, deep learning with feature attribution enjoys some advantages over GLMs and symbolic regression. First, deep learning can learn complex, non-linear relations that are hard to capture in a GLM or a prespecified functional basis. Second, the same core architecture, such as the Transformer, can perform well across data domains [40, 41, 42]. Thus, an AI-based approach may be easier to apply across scientific fields.

In neuroscience, feature attribution can help identify the task and stimulus variables that most contribute to the ubiquitous and diverse mixed selectivity in neurons [11, 43]. Investigators can further use 'ablations'—withholding certain variables—to narrow the sufficient set of variables.

3.2 Factorizing complex dependencies in data

The ability of AI models to capture subtle interdependencies in the data can be leveraged to factor out, or control for, correlated variables in imbalanced data and allow scientists to ask otherwise intractable questions (Figure 2, right). For example, an intriguing study [44] examined whether non-robust image features (those alterable with small perturbations to the image) contain useful information for classification. It is not trivial to define robust and non-robust features in closed form, let alone manipulate them in high-dimensional images. However, non-robust CNNs implicitly represent non-robust features, and vice versa for adversarially trained CNNs. Thus, using the two types of CNNs and gradient descent, Ilyas and colleagues could create images that mostly contained robust or non-robust features for training classifier networks. They also trained networks to classify adversarial images according to *target* labels, because these images can now be interpreted as containing non-robust features that conflict with the remaining features. (E.g., the images look 'wrongly' labeled to humans.) Classifiers trained on either the non-robust or the conflicting features have good accuracy on unmodified validation images, indicating that non-robust features indeed contain useful information for categorization. This study exemplifies how AI models allow scientists to investigate the concepts models implicitly represent.

There is increasing interest in neuroscience to explain naturalistic data. However, drawing reliable conclusions from natural data is hindered by potential spurious correlations among variables [45]. AI methods can better 'regress out' uninteresting variables than linear regression and decomposition techniques. For example, the mouse visual cortex is strongly modulated by dimensions that are linearly independent of the visual stimulus but correlated with non-visual variables like pupil size, face movements, and running speed [7], as discussed in Section 2.1. However, some of the non-visual activity dimensions may be nonlinearly explainable by the visual activity, AI models can more fully remove the visual contributions to examine non-visual responses.

3.3 Limitations

Despite ways to derive insights from uninterpreted AI models, this process still requires a scientist in the loop to prespecify variables of interest. Ultimately, we desire an AI neuroscientist to extract 'concepts'—e.g., sparse features, relations between features, and connections to existing scientific and mathematical concepts. The next section reviews AI methods that hold such promises.

4 Methods that generate interpretable hypotheses

4.1 Symbolic regression

Recent advances in AI, especially in symbolic regression, offer promising avenues for discovering interpretable neural mechanisms from data. Symbolic regression aims to capture structure in the data with mathematical equations, bypassing the constraints of traditional methods where the structure, such as the linear subspaces in PCA, has to be presumed. The mathematical equations are constructed from a set of basic operators (such as addition) and basis functions (such as polynomials and trigonometric functions), and an optimization procedure selects the expression that both is succinct and fits the data well. As an example, Schmidt and Lipson [38] demonstrated that given recorded time-series data of position, velocity, and acceleration of a two-spring single mass system, symbolic regression can reproduce Newton's second law.

Though the optimization challenges in symbolic regression are substantial, recent advancements in heuristic methods using techniques like genetic programming [38, 46, 47, 48], Bayesian methods [49], sparse regression [39, 50] and neural networks [51, 52, 53] provide effective solutions. Furthermore, the development of specialized toolkits [54] has bolstered the applicability of these methods. Given these advancements, symbolic regression is poised to offer groundbreaking insights into neuroscience.

4.2 Applying symbolic regression to neuroscience

How can symbolic regression aid in deciphering neural mechanisms from data? We explore this with a specific example. Neural responses to an identical stimulus have been empirically observed to differ across trials [55, 56, 57] and this variability is correlated across neurons [58, 59]. However, the mechanisms driving this neural variability remain undefined.

Traditional neural analysis methods, reviewed in Section 2, have yet to fully explain this variability. Neuroscientists have approached this problem by positing models based on various hypotheses. These hypotheses explored how a single factor of population activity could modulate the neural stimulus-response, either through additive offset [60, 61], multiplicative gain [62], or a hybrid (termed 'multi-gain') [63, 64, 65]. Comparing the fit of each model to experimental data indicated a predilection for the multi-gain model. This model-fitting approach has limitations: it rests on predefined assumptions about the underlying mechanisms. It is unclear whether the assumptions hold and conceivable that the neural sensory response is modulated by population activity through alternative nonlinear interactions.

Symbolic regression emerges as a potent method capable of unveiling mathematical relationships in data without relying on pre-defined models (Figure 3). To understand the link between population activity and neural sensory responses, we can create an input-output pairing dataset. Inputs include sensory stimuli (e.g., orientation for drifting gratings) and neural population activity, while outputs capture neuronal activity in the visual cortex. Following Udrescy and Tegmark [52], we first fit the dataset to a black-box model, such as a multi-layer perceptron, to capture the input-output relation. This model is geared to accurately predict unseen neural activity and offers a more generalizable data

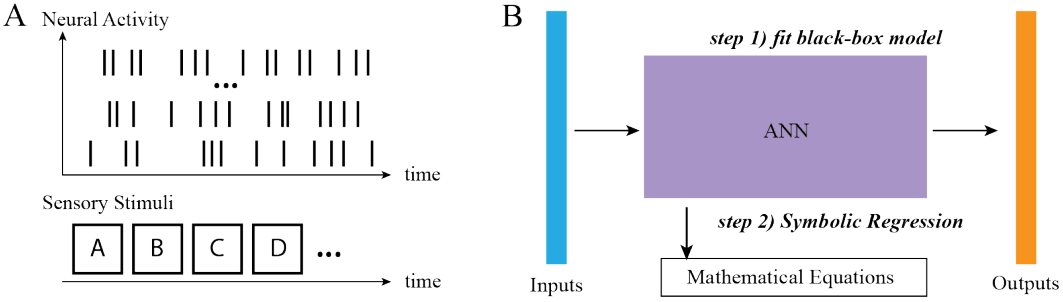

Figure 3: Applying symbolic regression to neuroscience. A) Raw data, including neural activity and sensory stimuli. B) The workflow involves structuring the raw data into input-output pairs, fitting with a black-box model, and applying symbolic regression to extract mathematical equations.

representation, facilitating the exploration of data properties, especially at input values absent from the dataset. Using symbolic regression, we then distill mathematical equations from this model, revealing an experimentally testable formula for how population activity affects neural sensory responses.

If the resultant equation features combined multiplicative and additive terms, it suggests that the population activity has a multi-gain effect, aligning with conclusions from previous model-based approaches. Nevertheless, the modulatory effect could potentially be more intricate. Symbolic regression enables an exploration of the equation landscape more expansive than the previous manual model formulation approach. Consequently, symbolic regression holds the promise of unveiling novel neural mechanisms in a less biased and more efficient manner.

4.3 Limitations: identifying variables

While symbolic regression offers a promising approach for discovering mathematical relations within data, its success heavily relies on the appropriate choice of input variables. As an example, Schmidt and Lipson [38] showed that given only the position and velocity data for a two-spring system, symbolic regression converged on the energy laws instead of Newton's second law. The need to select the right variables poses a challenge: the brain is vastly complicated, spanning multiple levels of complexity and making it difficult to preemptively identify the most pertinent variables, which are liable to be complex functions of the raw neural activities. Future work should ideally explore how to extract both the relevant variables and their mathematical relations from neural data.

4.4 Other methods

Discovering both variables and their relations within data is at the frontier of AI research. This problem is challenging due to the expansive search space and likely requires proper inductive biases [66]. Nevertheless, recent advances offer promise. For example, Chen et al. [67] introduced a method to identify state variables in high-dimensional data. Given an observed system, the algorithm first models the system's dynamics using a black-box neural network with bottleneck latent embeddings. It then estimates the number of state variables from these latent embeddings using geometric manifold learning techniques. Lastly, the algorithm trains another latent reconstruction neural network with the exact identified number of latent variables to identify the system's governing mechanisms. The algorithm was demonstrated to successfully extract the angle and angular velocity as key variables from a video of a pendulum in motion. Such techniques hold exciting potential for unveiling complex patterns and structures within neural data.

Casual discovery [68, 69, 70] is another promising avenue for progress. Here, the aim is to learn the ground-truth causal generative process of the data. For instance, Shen et al., 2020 [71] explored an enhancement to the VAE (Section 2), where the latent factors can be causally related. One advantage of integrating causality is its capacity to support intervention on causal variables. Such intervention predictions can then be validated experimentally through neural perturbation studies [72].

5 Neuroscience-specific considerations

5.1 Finding the right level of abstraction for hypotheses

Unlike the data AI typically handles, such as images or text, the brain is a complicated system spanning multiple levels of complexity: molecules, including neurotransmitters and receptors. regulate processes like synaptic transmission; individual neurons integrate inputs and relay signals to their neighbors; assemblies of neurons form microcircuits dedicated to specific computations; finally, groups of microcircuits converge to execute cognitive functions. Crucially, these levels are closely linked: Changes at one level can propagate and cause changes at another level. For example, molecular changes can alter neuron functions and even affect behavior [73, 74].

The brain's deep hierarchical structure presents distinct challenges when employing AI to elucidate neural mechanisms. When AI proposes a neural 'law,' it is crucial to interpret it in the appropriate biological context. For example, mechanisms identified by AI at network levels may not decompose neatly into cellular or molecular mechanisms, thereby missing a mechanistic level of explanation. Conversely, some key mechanisms may only emerge when examining data across multiple levels. One way to navigate these complexities is to integrate existing knowledge in neurobiology. For

instance, preconfiguring AI models with a set of biologically plausible canonical computations can enhance interpretability and relevance [75].

5.2 Non-stationarity

Biological systems are adaptive and multi-scaled, properties that pose additional challenges to AI-for-science techniques developed for physical and engineering sciences. For example, symbolic regression to discover physical laws assumes that the same equation describes system evolution in a time-invariant way. However, the brain is plastic at multiple time scales. We provide a representative but non-exhaustive list to illustrate the relevance of non-stationarities. At the 10^{-2}–10^{0} s scale, neural activities evolve due to circuit properties like adaptation and recurrent processing even without changes in the external drive (e.g., viewing a static image). At the 10^{0}–10^{2} s scales, sensory responses continue to manifest adaptation, while internal brain states including arousal, attention, motivation, and neuromodulation also fluctuate. At yet longer time scales (hours to days), the brain manifests the effects of learning. At time scales of a species' life history (birth to adulthood to senescence), the brain develops, remodels, ages, and evolves. Theseus' paradox involves all of these scales.

5.3 Data modalities

Neuron-level brain recording today remains in the regime of extremely sparse sampling. Leading edge techniques can record about 10^{4} neurons, but a tiny fraction in moderately complex brains. For comparison, the mouse brain contains on the order of 10^{8} neurons, while the human brain has almost 10^{11}. Although recording from most neurons in a mammalian brain is ultimately possible physically [76], sampling by current recording techniques—primarily calcium imaging and high-density electrophysiology like Neuropixels—is sparse and biased, with the two techniques differently deviating from i.i.d. The neocortex is locally organized like a flat sheet with layers (imagine puff pastry but with six layers of varying depth). Neuropixel recordings are confined to a narrow sheath around a linear track (the voltage signal amplitude decreases ten-fold $\tilde{5}0$ um away from the neuron [77]; this slender track may be normal or tangential to the cortical surface. Meanwhile, calcium imaging more evenly samples the tangent plane but penetrates to a limited depth (typically layers 2/3) within accessible brain areas (e.g., at the top of the head and outside cortical convolutions).

Calcium imaging uses genetically encoded fluorescence reporters that restrict the signal to a defined cell type and, usually, a sub-cellular compartment (e.g., cell bodies vs. dendrites). In contrast, electrophysiology provides essentially no cell type or compartment information and is probably dominated by large neurons that generate high-amplitude spikes.

The time resolution also differs between the two modalities. Ephys has sub-spike resolution (below 1 ms), whereas calcium imaging is currently limited to about 10^{-1} s by the indicator dynamics [78].

Ephys comes with non-single-neuron signal types—multiunit activities (MUAs) and local field potentials (LFPs)—that have no close analogs in calcium imaging. MUAs are functionally and conceptually (under a linear readout assumption) similar to single-unit activity, whereas LFPs have distinct interpretations.

6 Other opportunities

6.1 Data preprocessing

Data preprocessing uses statistical methods that AI can supply. For clearly defined end goals (e.g., to localize cells, align imaging sessions, and sort spikes), current methods serve well and leave less margin for improvement by AI. However, AI may reveal additional information in raw data in hard-to-expect ways. Just as images contain category-relevant 'adversarial features' unrecognizable to humans, so there may be informative patterns amidst the diffuse fluorescence in calcium images after cell extraction or within the voltage trace fluctuations after spike sorting, although we stress that the existence of any residual information and its form are unexpectable by definition. Thus, data preprocessing may yet hold surprising rewards for applying AI.

6.2 Benchmark datasets

Large-scale, high-quality datasets are pivotal for the advancement and evaluation of AI methods. In neuroscience, there have been concerted efforts to offer standardized and accessible large-scale datasets. For instance, Brain-Score [18] assesses the feature alignment between task-pretrained deep neural networks and the primate ventral visual stream. The Allen Institute for Brain Science [79] and the sensorium competition [80] provide expansive data from mouse visual areas in response to natural images. The Neural Latent Benchmark [81] provides monkey responses across sensory and motor areas for a diverse set of cognitive behaviors.

The majority of these datasets are designed for predicting neural activity from sensory stimuli. While building precise prediction models is useful, interpreting those models is also vital. A comparable initiative in the realm of physics is the Feynman Symbolic Regression Database [52]. This database is organized into input-output pairs: the inputs are observational data, and the outputs correspond to equations of physical law. By this token, an aspirational dataset for neuroscience would have neural data as inputs and equations that describe those data as outputs. Yet, neuroscience poses a distinct challenge: The ground-truth mechanisms are often unknown. A pragmatic intermediary approach might entail representing outputs as AI-generated testable predictions. These predictions could then be either cross-referenced with existing literature or tested experimentally.

6.3 Using prior knowledge

Scientific discoveries do not occur in a vacuum [82]; they heavily depend on and build upon previous knowledge. In formulating hypotheses, prior knowledge not only constrains the hypotheses to ensure alignment with established knowledge but also sparks inspiration by providing frameworks for novel ideas. When generating neuroscience hypotheses using AI methods, it is crucial to incorporate prior knowledge. For example, in feature learning 3, established neuroscientific principles, such as the firing patterns of specific neural types, can be integrated using regularization terms to guide the learning process. Neuroscience knowledge presented as concepts or equation-based models can be directly incorporated into symbolic regression 4 as candidate variables or starting points, thereby guiding the symbolic search.

7 Conclusion

AI has begun to transform all fields of science [83, 84]. If any field should be exceptional, it is neuroscience: The study of natural information processing systems has always been tightly entwined with AI, in history and inherently. We reviewed current applications of AI in neuroscience, which emphasize their parallels. Further from the streetlight, we spotlight opportunities for using AI as analysis tools to unlock insights into neuroscience data.

References

[1] MD György Buzsáki. *The brain from inside out*. Oxford University Press, 2019.

[2] Alan L Hodgkin and Andrew F Huxley. "A quantitative description of membrane current and its application to conduction and excitation in nerve". In: *The Journal of physiology* 117.4 (1952), p. 500.

[3] LM Lapicque. "Recherches quantitatives sur l'excitation electrique des nerfs". In: *J Physiol Paris* 9 (1907), pp. 620–635.

[4] John J Hopfield. "Neural networks and physical systems with emergent collective computational abilities." In: *Proceedings of the national academy of sciences* 79.8 (1982), pp. 2554–2558.

[5] David H Hubel and Torsten N Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". In: *The Journal of physiology* 160.1 (1962), p. 106.

[6] Ian H Stevenson and Konrad P Kording. "How advances in neural recording affect data analysis". In: *Nature neuroscience* 14.2 (2011), pp. 139–142.

[7] Carsen Stringer et al. "Spontaneous behaviors drive multidimensional, brainwide activity". In: *Science* 364.6437 (2019), eaav7893.

[8] Carsen Stringer et al. “High-dimensional geometry of population responses in visual cortex”. In: *Nature* 571.7765 (2019), pp. 361–365.

[9] Sadegh Ebrahimi et al. “Emergent reliability in sensory cortical coding and inter-area communication”. In: *Nature* 605.7911 (2022), pp. 713–721.

[10] Peiran Gao et al. “A theory of multineuronal dimensionality, dynamics and measurement”. In: *BioRxiv* (2017), p. 214262.

[11] Mattia Rigotti et al. “The importance of mixed selectivity in complex cognitive tasks”. In: *Nature* 497.7451 (2013), pp. 585–590.

[12] Shih-Yi Tseng et al. “Shared and specialized coding across posterior cortical areas for dynamic navigation decisions”. In: *Neuron* 110.15 (2022), pp. 2484–2502.

[13] Demis Hassabis et al. “Neuroscience-inspired artificial intelligence”. In: *Neuron* 95.2 (2017), pp. 245–258.

[14] Blake A Richards et al. “A deep learning framework for neuroscience”. In: *Nature neuroscience* 22.11 (2019), pp. 1761–1770.

[15] Adrien Doerig et al. “The neuroconnectionist research programme”. In: *Nature Reviews Neuroscience* (2023), pp. 1–20.

[16] Daniel LK Yamins and James J DiCarlo. “Using goal-driven deep learning models to understand sensory cortex”. In: *Nature neuroscience* 19.3 (2016), pp. 356–365.

[17] Aran Nayebi et al. “Identifying learning rules from neural network observables”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2639–2650.

[18] Martin Schrimpf et al. “Brain-score: Which artificial neural network for object recognition is most brain-like?” In: *BioRxiv* (2018), p. 407007.

[19] Luca Pion-Tonachini et al. “Learning from learning machines: a new generation of AI technology to meet the needs of science”. In: *arXiv preprint arXiv:2111.13786* (2021).

[20] George Dimitriadis, Joana P Neto, and Adam R Kampff. “t-SNE visualization of large-scale neural recordings”. In: *Neural computation* 30.7 (2018), pp. 1750–1774.

[21] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. “Learnable latent embeddings for joint behavioural and neural analysis”. In: *Nature* (2023), pp. 1–9.

[22] John P Cunningham and Byron M Yu. “Dimensionality reduction for large-scale neural recordings”. In: *Nature neuroscience* 17.11 (2014), pp. 1500–1509.

[23] Byron M Yu et al. “Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity”. In: *Advances in neural information processing systems* 21 (2008).

[24] Yuan Zhao and Il Memming Park. “Variational latent gaussian process for recovering single-trial dynamics from population spike trains”. In: *Neural computation* 29.5 (2017), pp. 1293–1316.

[25] Jakob H Macke et al. “Empirical models of spiking in neural populations”. In: *Advances in neural information processing systems* 24 (2011).

[26] Lars Buesing, Jakob H Macke, and Maneesh Sahani. “Learning stable, regularised latent models of neural population dynamics”. In: *Network: Computation in Neural Systems* 23.1-2 (2012), pp. 24–47.

[27] Yuan Zhao and Il Memming Park. “Interpretable nonlinear dynamic modeling of neural trajectories”. In: *Advances in neural information processing systems* 29 (2016).

[28] Byron M Yu et al. “Extracting dynamical structure embedded in neural activity”. In: *Advances in neural information processing systems* 18 (2005).

[29] Yuanjun Gao et al. “Linear dynamical neural population models through nonlinear embeddings”. In: *Advances in neural information processing systems* 29 (2016).

[30] Chethan Pandarinath et al. “Inferring single-trial neural population dynamics using sequential auto-encoders”. In: *Nature methods* 15.10 (2018), pp. 805–815.

[31] David Sussillo. “Neural circuits as computational dynamical systems”. In: *Current opinion in neurobiology* 25 (2014), pp. 156–163.

[32] Ding Zhou and Xue-Xin Wei. “Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7234–7247.

[33] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. "Deep supervised, but not unsupervised, models may explain IT cortical representation". In: *PLoS computational biology* 10.11 (2014), e1003915.

[34] Daniel LK Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624.

[35] Hyodong Lee et al. "Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network". In: *bioRxiv* (2020), pp. 2020–07.

[36] Fenil R Doshi and Talia Konkle. "Cortical topographic motifs emerge in a self-organized map of object space". In: *Science Advances* 9.25 (2023), eade8187.

[37] Zejin Lu et al. "End-to-end topographic networks as models of cortical map formation and human visual behaviour: moving beyond convolutions". In: *arXiv preprint arXiv:2308.09431* (2023).

[38] Michael Schmidt and Hod Lipson. "Distilling free-form natural laws from experimental data". In: *science* 324.5923 (2009), pp. 81–85.

[39] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proceedings of the national academy of sciences* 113.15 (2016), pp. 3932–3937.

[40] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[41] Andrew Jaegle et al. "Perceiver io: A general architecture for structured inputs & outputs". In: *arXiv preprint arXiv:2107.14795* (2021).

[42] Kevin Lu et al. "Pretrained transformers as universal computation engines". In: *arXiv preprint arXiv:2103.05247* 1 (2021).

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

[44] Andrew Ilyas et al. "Adversarial examples are not bugs, they are features". In: *Advances in neural information processing systems* 32 (2019).

[45] Célian Bimbard et al. "Behavioral origin of sound-evoked activity in mouse visual cortex". In: *Nature neuroscience* 26.2 (2023), pp. 251–258.

[46] Douglas Adriano Augusto and Helio JC Barbosa. "Symbolic regression via genetic programming". In: *Proceedings. Vol. 1. Sixth Brazilian symposium on neural networks*. IEEE. 2000, pp. 173–178.

[47] Steven Gustafson, Edmund K Burke, and Natalio Krasnogor. "On improving genetic programming for symbolic regression". In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 1. IEEE. 2005, pp. 912–919.

[48] Qiang Lu, Jun Ren, and Zhiguang Wang. "Using genetic programming with prior formula knowledge to solve symbolic regression problem". In: *Computational intelligence and neuroscience* 2016 (2016), pp. 1–1.

[49] Ying Jin et al. "Bayesian symbolic regression". In: *arXiv preprint arXiv:1910.08892* (2019).

[50] Markus Quade et al. "Sparse identification of nonlinear dynamics for rapid model recovery". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.6 (2018).

[51] Brenden K Petersen et al. "Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients". In: *arXiv preprint arXiv:1912.04871* (2019).

[52] Silviu-Marian Udrescu and Max Tegmark. "AI Feynman: A physics-inspired method for symbolic regression". In: *Science Advances* 6.16 (2020), eaay2631.

[53] Miles Cranmer et al. "Discovering symbolic models from deep learning with inductive biases". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17429–17442.

[54] Michael Schmidt and Hod Lipson. "Eureqa (version 0.98 beta)[software]". In: *Nutonian, Somerville, Mass, USA* (2013).

[55] P Heggelund and K Albus. "Response variability and orientation discrimination of single cells in striate cortex of cat". In: *Experimental Brain Research* 32 (1978), pp. 197–211.

[56] David J Tolhurst, J Anthony Movshon, and Andrew F Dean. "The statistical reliability of signals in single neurons in cat and monkey visual cortex". In: *Vision research* 23.8 (1983), pp. 775–785.

[57] Rufin Vogels, Werner Spileers, and Guy A Orban. "The response variability of striate cortical neurons in the behaving monkey". In: *Experimental brain research* 77 (1989), pp. 432–436.

[58] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. "Neural correlations, population coding and computation". In: *Nature reviews neuroscience* 7.5 (2006), pp. 358–366.

[59] Michael R Deweese and Anthony M Zador. "Shared and private variability in the auditory cortex". In: *Journal of neurophysiology* 92.3 (2004), pp. 1840–1855.

[60] Alexander S Ecker et al. "State dependence of noise correlations in macaque primary visual cortex". In: *Neuron* 82.1 (2014), pp. 235–248.

[61] Michael Okun et al. "Diverse coupling of neurons to populations in sensory cortex". In: *Nature* 521.7553 (2015), pp. 511–515.

[62] Robbe LT Goris, J Anthony Movshon, and Eero P Simoncelli. "Partitioning neuronal variability". In: *Nature neuroscience* 17.6 (2014), pp. 858–865.

[63] I-Chun Lin et al. "The nature of shared cortical variability". In: *Neuron* 87.3 (2015), pp. 644–656.

[64] Iñigo Arandia-Romero et al. "Multiplicative and additive modulation of neuronal tuning with population activity affects encoded information". In: *Neuron* 89.6 (2016), pp. 1305–1316.

[65] Neil C Rabinowitz et al. "Attention stabilizes the shared gain of V4 populations". In: *Elife* 4 (2015), e08998.

[66] Francesco Locatello et al. "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[67] Boyuan Chen et al. "Automated discovery of fundamental variables hidden in experimental data". In: *Nature Computational Science* 2.7 (2022), pp. 433–442.

[68] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. "D'ya like dags? a survey on structure learning and causal discovery". In: *ACM Computing Surveys* 55.4 (2022), pp. 1–36.

[69] Murat Kocaoglu et al. "Causalgan: Learning causal implicit generative models with adversarial training". In: *arXiv preprint arXiv:1709.02023* (2017).

[70] Mengyue Yang et al. "Causalvae: Disentangled representation learning via neural structural causal models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9593–9602.

[71] Xinwei Shen et al. "Weakly supervised disentangled generative causal representation learning". In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 10994–11048.

[72] Karl Deisseroth. "Optogenetics". In: *Nature methods* 8.1 (2011), pp. 26–29.

[73] Eric J Nestler. "Molecular basis of long-term plasticity underlying addiction". In: *Nature reviews neuroscience* 2.2 (2001), pp. 119–128.

[74] Philip Seeman. "Dopamine receptors and the dopamine hypothesis of schizophrenia". In: *Synapse* 1.2 (1987), pp. 133–152.

[75] Christof Koch. *Biophysics of computation: information processing in single neurons*. Oxford university press, 2004.

[76] David Kleinfeld et al. "Can one concurrently record electrical spikes from every neuron in a mammalian brain?" In: *Neuron* 103.6 (2019), pp. 1005–1015.

[77] György Buzsáki, Costas A Anastassiou, and Christof Koch. "The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes". In: *Nature reviews neuroscience* 13.6 (2012), pp. 407–420.

[78] Yan Zhang et al. "Fast and sensitive GCaMP calcium indicators for imaging neural populations". In: *Nature* 615.7954 (2023), pp. 884–891.

[79] Saskia EJ de Vries et al. "A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex". In: *Nature neuroscience* 23.1 (2020), pp. 138–151.

[80] Konstantin F Willeke et al. "The Sensorium competition on predicting large-scale mouse primary visual cortex activity". In: *arXiv preprint arXiv:2206.08666* (2022).

[81] Felix Pei et al. "Neural Latents Benchmark'21: evaluating latent variable models of neural population activity". In: *arXiv preprint arXiv:2109.04463* (2021).

[82] Wikipedia contributors. *Spherical cow — Wikipedia, The Free Encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Spherical_cow&oldid=1176405805`. [Online; accessed 3-October-2023]. 2023.

[83] Hanchen Wang et al. “Scientific discovery in the age of artificial intelligence”. In: *Nature* 620.7972 (2023), pp. 47–60.

[84] *How artificial intelligence can revolutionise science*. Sept. 14, 2023. URL: https://www.economist.com/leaders/2023/09/14/how-artificial-intelligence-can-revolutionise-science (visited on 09/28/2023).