ChatGPT

# Evaluating NeuroFM-X and Enhancing Its Mechanistic Insights

## Current Capabilities of NeuroFM-X

NeuroFM-X is already an ambitious *neural foundation model* integrating many state-of-the-art components for neural data analysis. The codebase (in `neuros-neurofm`) outlines a pipeline that can ingest various neural modalities (spike times, calcium imaging, LFP/iEEG signals, behavioral data) and process them through a multi-module architecture [1] [2]. Key implemented features include:

- **Selective State-Space Backbone (Mamba):** A linear-time sequence model that handles very long neural time series more efficiently than standard Transformers [3] [4]. The Mamba SSM uses multi-rate processing (multiple time-scale streams at 5ms, 20ms, 80ms) with cross-scale gating, enabling it to capture neural dynamics across fast and slow timescales [5] [6]. This addresses the challenge of long recordings (up to millions of timesteps) without the $O(T^2)$ cost of attention.

- **Perceiver-IO Multi-Modal Fusion:** A latent-space attention module that merges inputs from different modalities (spikes, LFP, behavior, etc.) efficiently [7] [8]. A set of learnable latent vectors attend to each data modality via cross-attention, achieving integration of neural and behavioral streams in a shared latent space. This design (inspired by Perceiver-IO) allows flexible input sizes and focuses computational effort on a fixed latent bottleneck.

- **Population Transformer (PopT) Aggregator:** A module to achieve *population-level* representations that are permutation-invariant to neuron identities [9] [10]. PopT treats each neuron's features as a token and uses a small Transformer or pooling mechanism to aggregate across neurons. This helps the model generalize across sessions or subjects where the recorded neurons differ (a form of **cross-session neural alignment**). It implements the concept of **population-level coding**, ensuring the model's latent state isn't tied to any one neuron's identity – an important step toward biological plausibility.

- **Latent Diffusion Generative Model:** A two-stage generative module that learns to *simulate neural activity* [11] [12]. Neural data is encoded into a latent distribution, and a conditional diffusion (e.g. DDPM/UNet) model generates or forecasts plausible neural activity sequences [13] [14]. Applications include predicting future neural states (short-term forecasting of 1–2s), imputing missing neurons or data, and augmenting datasets with synthetic but realistic neural activity [15]. This component endows NeuroFM-X with a **"digital twin"** capability – the model can produce synthetic neural signals that mimic real ones, useful for testing hypotheses in silico.

- **Multi-Task Heads (Decoding, Encoding, Contrastive):** On top of the core, NeuroFM-X has separate output heads for various objectives [16] [17]. For example, a *Decoding head* predicts behavioral variables from neural data (e.g. kinematics, EMG, phonemes) [18]; an *Encoding head* predicts held-out neural activity (like a generative encoding model for missing neurons) [19]; and a *Contrastive head*

implements **CEBRA**-style representation learning [17]. The contrastive objective aligns latent neural representations with behavior or task context, encouraging the model's latent space to capture meaningful, behaviorally-relevant neural dynamics. This is in line with recent advances like **CEBRA** (Contrastive Embedding via Behavioral Retrieval) which learn latent factors explaining neural activity and behavior jointly [20] [21].

- **Adapter-Based Transfer Learning:** The codebase supports adding *adapters* to handle new neurons, sessions, or species without retraining the entire model [22] [23]. For example, Unit-ID adapters learn embedding vectors for new recorded neurons [22], and LoRA (Low-Rank Adaptation) layers allow tuning a few low-rank weight matrices instead of the full network [23]. This design lets NeuroFM-X adapt to new data with minimal updates – crucial for real-world use where each experiment may record a different neural population.

Overall, NeuroFM-X's design is **comprehensive and cutting-edge**. It already fuses many modern machine learning architectures (SSMs, Transformers, diffusion models, contrastive learners) with neuroscience-informed ideas (spike tokenization, population-level invariances, session transfer). The *production-ready* aspects (Lightning training pipeline, NWB data interfaces, Docker deployment) further strengthen it as a practical platform [24] [25]. These foundations mean NeuroFM-X can handle *cross-task learning* (multiple behaviors), *cross-species generalization*, and efficient training on large neural datasets [26].

# Opportunities for Improvement and Expansion

To **significantly expand NeuroFM-X's capabilities**, we can integrate advanced mathematical and conceptual methods that focus on *mechanistic interpretability* – in other words, extracting human-understandable explanations of how neural computations are organized. The ultimate goal is to have NeuroFM-X not just *predict* neural and behavioral data well, but also to provide insights into **how the brain carries out its functions**. Below we outline several directions, grounded in recent research, that could enhance NeuroFM-X:

## 1. Mechanistic Interpretability Techniques

**Extract Low-Dimensional Circuit Mechanisms:** One way to shed light on *how* the model (and by analogy, the brain) computes is to distill the complex high-dimensional network dynamics into a **simplified circuit** of interacting latent variables. Recent work by Langdon & Engel (2025) introduced a *"latent circuit model"* that infers a low-dimensional recurrent network explaining heterogeneous neural responses [27]. Instead of just correlating neural activity with task variables, their approach discovers how different latent factors (e.g. "context" vs "stimulus" signals) **interact via recurrent connections** to produce behavior [28]. Applying such *latent circuit inference* to NeuroFM-X's learned representations could identify interpretable modules or feedback loops within the model. For example, after training, one might find a small circuit where a "decision" unit inhibits a "sensory" unit to implement context-dependent suppression of irrelevant inputs – exactly the mechanism that Langdon & Engel uncovered in monkeys and in trained RNNs [29]. By fitting a simplified dynamical system to NeuroFM-X's internal activations, we could map its internals to a **causal circuit diagram** (with excitatory/inhibitory connections among a few latent nodes) that mirrors a hypothesis of brain circuitry. This provides *mechanistic insight*: the model's behavior can be explained by a known motif (e.g. inhibitory gating of inputs), which can then be compared to biological neural circuit motifs in cortex.

**Algorithmic and Circuit Analysis of the Model's Neurons:** In the spirit of **mechanistic interpretability in AI**, we can adopt techniques that have been used to **reverse-engineer artificial neural networks**. In AI research, mechanistic interpretability seeks to break down a trained network into human-understandable *pieces (features and circuits)* that explain its behavior [30]. For NeuroFM-X, this could mean analyzing the learned weight matrices and activations to identify *sub-networks* that perform distinct computations. For example, in a Transformer or the PopT module, one can inspect **attention patterns** or weight matrices to see if certain neurons (or channels) correspond to specific concepts (like tuning to a particular stimulus feature or behavioral variable). Tools like **causal intervention** or ablation can be used: systematically lesion or silence subsets of units in the model *in silico* to see how it affects output. This is analogous to lesion studies in neuroscience and can pinpoint which components of NeuroFM-X are responsible for which function (e.g., a set of latent units might be crucial for decoding movement intention). Recent *circuit tracing* methods in deep networks use techniques such as **activation patching** (replace parts of the network's activations with alternatives to test their influence) and **feature visualization** (optimize an input that maximally excites a neuron) to understand neural representations. Bringing these into NeuroFM-X would let us characterize, for instance, what input pattern a particular latent unit is most responsive to – effectively probing the *"receptive field"* of units in latent space. Incorporating such analysis tools (perhaps as a post-training interpretability suite) would directly align with the vision of using *"mechanistic interpretability tools to characterize neural tuning"* in silico [31]. This was highlighted in a recent Cell perspective, which described how deep models can serve as **"digital twins"** of neural systems: once trained to emulate real neurons, the model can be probed to generate hypotheses and stimuli, which are then validated back in vivo [31]. Enabling NeuroFM-X with these capabilities transforms it from a predictive model into a *hypothesis-generating scientific tool*.

**Disentangling Mixed Selectivity:** Biological neurons are notoriously **polysemantic**, often responding to multiple overlapping features or task variables [32]. This mixed selectivity makes it hard to interpret single-neuron responses. NeuroFM-X's units (being artificial neurons) likely also learn mixed representations (e.g., a latent dimension might encode a combination of speed, position, and reward signals). Advanced methods can help *disentangle* these. One intriguing approach is **algorithm unrolling for interpretability**. Demba Ba and colleagues (2025) developed an interpretable deep model called **DUNL (Deconvolutional Unrolled Neural Learning)** to decompose neural activity into distinct components [33] [34]. The idea is to start from a *mechanistic generative model* (e.g., that a neuron's activity is a sum of latent causes convolved with distinct temporal kernels) and unroll the inference algorithm for that model into a neural network architecture. By training this unrolled network on neural data, DUNL was able to recover human-understandable components of neural activity without averaging across trials [35]. For example, a dopamine neuron's firing might be explained as the sum of a "value" signal and a "salience" signal – DUNL aims to tease those apart in single trials [36]. **Integrating an "unrolled" sparse factor model** into NeuroFM-X could allow it to explicitly represent multiple causes for each neuron's activity. Concretely, one could add a module (or regularize an existing layer) that forces the representation to be a superposition of a few *interpretable factors* (perhaps by designing it as a sparse autoencoder with convolutional kernels, per DUNL). This would encourage NeuroFM-X to represent neural data in terms of latent **sources** (like movement vs. sensory drive, or different stimulus features), each of which could be traced and understood. Not only does this improve interpretability, it aligns with the neuroscience goal of explaining neural responses in terms of *meaningful latent causes* rather than opaque high-dimensional vectors. As the Kempner Institute blog notes, aligning latent dimensions with human-understandable variables (space, time, task factors) and disentangling overlapping components are key challenges for interpretable brain models [37] [32].

**Causal/Intervention-Based Analysis:** To truly claim a **mechanistic understanding**, one must test causal hypotheses. NeuroFM-X can be augmented with tools to perform *in silico experiments*. For instance, using the trained generative model, we could simulate the effect of activating or silencing certain neurons: e.g., set a subset of spike-token inputs to a high rate and see how the model's prediction of behavior changes. Because NeuroFM-X integrates with real data streams, one could imagine a *closed-loop* setup: the model suggests a perturbation (like "if neurons in this latent cluster were inhibited, the model predicts movement X would fail"), and then that perturbation could be approximated in real experiments (via optogenetics or stimulation) to see if the behavior actually changes. This loop of **predictive causal analysis** is forward-looking, but even without real experiments, *internally* NeuroFM-X's structure allows virtual perturbation tests. For example, the **Population Transformer** aggregator could be analyzed by zeroing out certain neuron embeddings to see if some subset of neurons is critical for a given task output. Likewise, the **cross-attention weights** in the Perceiver module can be examined: if the model attends strongly to LFP vs spikes for a certain prediction, that suggests the relevant pathway for that task. By analyzing these attention weight patterns, we might infer which modality or brain region is driving a given cognitive function – a form of interpretability that parallels neuroscientists asking "is this behavior more encoded in spikes or in LFP rhythms?"

**Representational Similarity and Identifiability:** Another mechanistic angle is comparing *the model's internal representations with the brain's representations*. Since NeuroFM-X is trained on neural data, one can compute **representational similarity** between the model's latent features and recorded neural activity or known theoretical features. If, say, the model has a 512-dim latent in the Perceiver, we could ask: do linear combinations of these dimensions correlate with known neural population dynamics (like principal components of the neural data) or known task variables? Recent methods like CEBRA already push the model to align latent space with behavioral variables [20]. We can go further by evaluating whether the latent space is **identifiable** and interpretable: for instance, does one latent dimension consistently encode "running speed" across animals, and another "heading direction"? If not, we might impose additional structure or priors during training (e.g. an orthogonal constraint or supervised factors) so that these factors *emerge as separate latent axes*. The Mathis & Mathis (2024) Cell perspective emphasizes that ultimately we want models that bridge the gap between high predictive power and **mechanistic realism**, yielding "causal, testable models" rather than just black-box function approximators [38]. To that end, enforcing that certain latent units correspond to known mechanistic quantities (like tuning curves, oscillation phase, etc.) could be a productive improvement.

## 2. Biologically-Plausible Modeling Enhancements

**Incorporate Biophysical Neuron Models (Differentiable Spiking Networks):** NeuroFM-X currently uses rate-based or token-based neural representations. An improvement is to make parts of the model more *biophysically grounded*. For example, integrating **spiking neural network units** (with membrane potential dynamics and spiking nonlinearity) could allow the model to capture neuronal properties like refractory periods or precise spike timing, which are lost in coarse time-bin tokens. Recent advances in **surrogate gradient training** make it feasible to include spiking neuron models in deep networks. By replacing or augmenting certain layers with spiking neuron simulations, NeuroFM-X could simulate *actual neural circuit dynamics* more faithfully. This is not just for realism – it can aid interpretability, as spiking models have explicit parameters (membrane time constants, thresholds, synaptic weights) that relate to physical properties. If NeuroFM-X had, say, a spiking recurrent layer that learns a certain oscillatory pattern, we could interpret that in familiar neuroscientific terms (e.g. a gamma oscillation emerging from excitatory-inhibitory interactions). The challenge is ensuring differentiability; however, tools now exist for differentiable

spiking circuits [39]. In fact, researchers have demonstrated *differentiable simulators* of biophysical neural models that can be trained with gradient descent, bridging deep learning with mechanistic brain simulation [39]. Incorporating a **differentiable simulator of a neural circuit** (for instance, a small network of Hodgkin-Huxley or Izhikevich model neurons) inside NeuroFM-X could let the system tune actual biophysical parameters to fit data. This would yield a **hybrid model**: part deep learning, part mechanistic simulation. The benefit is twofold – improved biological fidelity and the ability to directly interpret those mechanistic parameters (e.g. "synaptic weight X increased, representing potentiation between these neurons").

**Enforce Neurobiological Constraints:** Another improvement is adding architectural or regularization constraints derived from neuroscience knowledge. For example, **Dale's law** (neurons are either excitatory or inhibitory, not both) could be imposed on the model's interneuron connections. In practice, this means splitting latent units into excitatory (weights constrained to $\ge 0$ for certain connections) and inhibitory (weights $\le 0$) sets. Enforcing such structure might make the learned model more interpretable, as you could then identify "this latent unit acts like an inhibitory interneuron suppressing a population." Similarly, we might structure layers to mimic known circuit motifs – e.g., an E-I feedback loop, or a feedforward hierarchy reflecting sensory → association → motor areas. In fact, a *2024 study* showed that reorganizing a neural network's layers to **mirror the structure of a known mechanistic model** (in their case, a pharmacological pathway) led to better training efficiency and maintained interpretability of the system [40]. In the context of NeuroFM-X, this could mean if we have prior knowledge of the brain system (say we know area A projects to B which projects to C in the task), we could constrain the model to have modules corresponding to A, B, C with unidirectional connections. The model then wouldn't be a free-form black box but rather a *semi-mechanistic model* reflecting the actual information flow. This approach retains some flexibility of deep learning but grounds it in plausible anatomy, making results easier to map to brain circuitry [41] [42]. Another example of biological constraint is **population coupling structure** – e.g., ensuring that some "principal components" of population activity (like a low-rank factor) drive the output, analogous to neural modes. Regularizing the model to have a low effective rank in certain layers might encourage it to learn a few key dynamical patterns rather than many inscrutable ones.

**Multi-Scale and Multi-Level Modeling:** The brain operates across multiple scales (from cellular electrophysiology up to whole-brain networks). NeuroFM-X focuses on the neural population level, but integrating multi-scale aspects could enhance interpretability. For instance, adding modules that simulate **mesoscopic dynamics** (like local field potential generation from aggregate synaptic activity) would allow the model to connect single-unit activity with population-level rhythms. If the model can predict LFP or EEG from its internal state, we can validate whether it's capturing known oscillatory mechanisms. Likewise, considering *inter-area interactions*: one could incorporate a high-level model of brain area connectivity (e.g., a matrix describing how signals flow between cortical regions). Even if data from multiple areas is not directly available, hypothesizing and fitting an inter-area coupling can be done (similar to The Virtual Brain project's large-scale brain network models). The recent **Virtual Brain Inference (VBI)** toolkit, for example, allows probabilistic fitting of global brain network models to data [43]. While full brain simulation may be beyond NeuroFM-X's scope, a simplified version – e.g., a two-region model (motor and sensory cortex interacting) – could be embedded. This again would give interpretable parameters (like coupling strengths) that connect to neuroscience theories of brain oscillations or communication. The overarching idea is to **tie NeuroFM-X's components to real physiological or anatomical counterparts** wherever possible: each time we do so, we gain a foothold for understanding. Instead of just saying "layer 5 did X," we could say "the model's 'hippocampal formation module' did X," which is more meaningful if that module was designed to mimic hippocampal circuit function.

## 3. Advanced Transformer and Diffusion Approaches for Insight

**Analyzing Attention and Representations:** NeuroFM-X's use of Transformer-derived ideas (Perceiver IO and PopT) can itself be harnessed for interpretability. **Transformer attention weights** are often interpretable as a sort of *soft connectivity* or influence matrix – they tell us which inputs or tokens the model is focusing on at each layer. For example, the Perceiver's cross-attention might reveal *which modality provides the most information* for a given task epoch. If we inspect the attention scores, we might discover that during movement, the model attends mostly to spikes from motor cortex, whereas during rest it attends more to LFP theta oscillations (just as a hypothetical). This aligns with neuroscience intuitions about task-dependent information routing. By visualizing these patterns (perhaps averaging attention weights for trials of a certain type), we effectively get a *data-driven circuit diagram* of "who listens to whom" in the model. There is also the concept of **token embeddings**: in spike tokenization, each neuron or time bin has an embedding vector. We could perform clustering or dimensionality reduction on these embeddings to see if the model groups neurons by functional similarity (e.g., all neurons with similar tuning end up with similar embedding vectors). This could reveal *emergent functional categories* learned by the model, analogous to how neuroscientists classify cells into functional types.

**Diffusion Models for Decoding and Hypothesis Generation:** The inclusion of a latent diffusion model in NeuroFM-X opens a novel avenue: using generative models to **visualize and test what the brain is encoding**. A striking recent example is work that used **latent diffusion models (like Stable Diffusion)** to decode images directly from brain fMRI activity [31]. In that work, the researchers could reconstruct an image a person was viewing by guiding a diffusion model with the person's brain signals (reference "Chen et al." in the Cell paper) [31]. We can apply a similar concept to neural population data: since NeuroFM-X already has a diffusion prior that can generate neural activity, we can condition it on *desired outputs* and see what neural patterns it produces. For instance, we might ask: *what neural activity pattern would cause the model to decode a "leftward reach" at time $t$?* By sampling from the diffusion model conditioned on the decoder predicting "left reach," we obtain synthetic neural trajectories that the model believes correspond to that movement. Analyzing these synthetic trajectories could reveal what features are key (perhaps an increase in firing of a certain neuron group, or a particular oscillatory burst). Those, in turn, are predictions about the real brain: e.g., *the model predicts that a burst of beta oscillation in these channels causes a leftward reach*. This approach leverages the generative model as an **interpretability tool** – much like how one would use a well-calibrated *digital twin* to run virtual experiments. The Cell perspective specifically illustrates this concept: a loop where **"in silico models employ mechanistic interpretability tools to characterize neural tuning, and then images/hypotheses synthesized in silico are validated back in vivo"** [44]. Using diffusion-driven synthesis of *stimuli* is another exciting path: one could train a diffusion model that generates, say, visual stimuli (images) that would maximally drive a given neural population state. This is akin to *dreaming up an image that a brain cell "wants to see."* Such images can offer intuitive clues (e.g., a cell's "dream image" might have vertical stripes, suggesting it's tuned to vertical orientations). In summary, **diffusion models and other deep generative models** integrated with NeuroFM-X can be turned toward interpretability: decoding brain states into images or sounds, generating new data to test model predictions, and creating rich visual explanations for neural activity patterns.

**Differentiable Environment Simulators:** Beyond neural data alone, one can consider NeuroFM-X operating within a *differentiable simulated environment* (e.g., a virtual avatar or a physics simulator). This would allow *end-to-end learning and interpretation of sensorimotor loops*. For example, imagine a differentiable simulator of an arm that takes muscle commands and outputs movements. If NeuroFM-X's output (decoded behavior) is fed into such a simulator, we could directly map neural patterns to predicted

movements and even optimize those. More interestingly, because the simulator is differentiable, one could propagate error back into the neural latent space. This could facilitate discovering *mechanistic strategies*: if the task is to reach a target, the combined model+simulator might learn interpretable intermediate variables (like desired trajectory, error corrections) in its latent space because those are useful for the physical task. This is somewhat speculative, but it aligns with the idea of *"closing the loop"* for a more biologically grounded evaluation of the model. It also relates to **normative models**: many normative theories (optimal control, Bayesian inference models, etc.) exist for sensorimotor tasks. By embedding the neural model in an environment loop, we can compare its performance and internal solutions to those normative standards. Any deviation is an insight – perhaps the model (and thus the brain) uses a strategy that is robust to noise at the expense of absolute optimality, for instance.

## 4. Guidance from Recent Research & Future Directions

The above ideas are heavily inspired by the latest trends at the intersection of neuroscience and AI. The overarching theme in literature is bridging the gap between high-performing *statistical models* and *true mechanistic understanding* [38]. Some key references and how they guide NeuroFM-X's evolution:

- **"Decoding the brain: from neural representations to mechanistic models" (Mathis & Mathis, 2024)** – This Cell perspective emphasizes that while we've gotten good at *encoding/decoding models* (i.e., mapping stimuli to neural responses and vice versa), we need to push toward models that also explain *why* and *how* those responses arise [20] [38]. It advocates using deep learning not just for prediction but as a means to build **"causal, testable models"** of brain function [45]. NeuroFM-X is a perfect platform to implement this vision. For instance, its generative module can serve in the *"digital twin"* role described: generating hypotheses about neural tuning that can then be tested [44]. Following this guidance, we should ensure NeuroFM-X isn't a black box – adding the interpretability and intervention abilities described above will answer the call put forth by this paper.

- **Mechanistic Interpretability in NeuroAI (e.g., DUNL by Ba et al. 2025)** – The Neuron paper and blog by Demba Ba's group introduce a concrete method (DUNL) to make deep models more interpretable by design [33] [46]. The success of DUNL in breaking neural activity into human-understandable components validates the approach of adding *sparsity and structure* to our models. NeuroFM-X could incorporate a similar *unrolled optimization* component (for example, unrolling a sparse coding algorithm that separates latent causes of neural firing). This would directly tackle the mixed selectivity problem by giving each latent factor a clear meaning. The key lesson is that **interpretable deep learning is feasible**: we don't have to accept a trade-off of accuracy vs. interpretability if we design the model cleverly [46]. By adopting these methods, NeuroFM-X can maintain its performance while becoming far more transparent about what each part of the network represents.

- **Latent Factor & Dynamical Models (e.g., LFADS, latent circuit models)** – Prior works like LFADS (Latent Factor Analysis via Dynamical Systems) and the more recent latent circuit model by Engel et al. show the power of assuming an explicit dynamical system underneath neural data. NeuroFM-X already has a notion of dynamics via the SSM and diffusion models, but we can strengthen the *interpretability of those dynamics*. The latent circuit approach, which discovered a specific *inhibitory mechanism* in a trained RNN and then found it in real brain data [27] [29], is a template for how we should use NeuroFM-X. After training NeuroFM-X on a task, we should perform a similar analysis: identify if there are latent inhibitory connections or other motifs in the model and then check if the

real neural data supports that. In essence, **NeuroFM-X can be used to generate mechanistic hypotheses which are then validated on neural recordings** – a powerful scientific workflow enabled by these advanced methods.

- **Differentiable Simulators and Hybrid Models:** Innovations in differentiable simulation (e.g., biomimetic spiking networks, Virtual Brain models) demonstrate that we can combine deep learning with actual biophysical modeling [39]. For NeuroFM-X, incorporating even a simplified differentiable neuron model or circuit can bring it closer to the biophysical realm, enabling direct mapping to things neuroscientists measure (like synaptic conductances or firing thresholds). The **Scientific Reports 2024 study by Mann et al.** further shows that structuring networks after mechanistic models not only improves interpretability but can improve training speed and accuracy [40]. This is an encouraging sign that *biologically-informed design* is not just philosophically satisfying, but practically beneficial. Embracing this, we can refactor parts of NeuroFM-X's architecture to more closely mimic known brain architecture (for instance, segregating sensory processing vs. motor command pathways within the model). Doing so might reduce the solution space the model has to search (making learning easier) and yield components that correspond to known brain subsystems (making interpretation easier).

- **Generative AI for Brain-Computer Interfaces:** The Cell paper and others also highlight the exciting use of generative models (like diffusion models) in brain-computer interface contexts [31]. By extending NeuroFM-X's generative capabilities (perhaps linking its latent diffusion model with image or text generators), we could create a system that *decodes* brain activity into rich outputs (reconstructed images a subject saw, predicted speech they intend, etc.). For example, coupling NeuroFM-X's latent state with a **text-generating transformer** could allow decoding of internal speech or thoughts from neural data – a frontier being explored with large language models and brain recordings. While these applications are beyond the current scope, the modular design of NeuroFM-X (with Perceiver for multi-modal input and diffusion for generation) means it could serve as the core brain-state representation that feeds into such decoders. This underscores that improving NeuroFM-X's interpretability and fidelity to neural mechanisms will not only advance basic understanding, but also enhance its utility in practical BCIs (where trust and transparency of the model are crucial).

## Conclusion and Next Steps

In summary, **NeuroFM-X is a powerful foundation** that can be elevated further by infusing it with cutting-edge mechanistic interpretability methods. By implementing the improvements above, we transform NeuroFM-X from a sophisticated predictor into a *scientific discovery platform*. It would be capable of generating hypotheses about neural circuit function, testing them in silico, and guiding real experiments – essentially fulfilling the vision of a *"neuroscience GPT"* that helps us understand the brain.

Concretely, next steps would involve: (a) adding analysis modules to the code (for computing attention maps, conducting ablations, factorizing latent features, etc.), (b) exploring architectural adjustments (like enforcing E/I separation or adding spiking neuron dynamics layers), and (c) integrating additional models for decoding (e.g., image diffusion for fMRI or language models for speech decoding). Throughout, we should validate that these additions indeed capture known neurobiological phenomena on reference datasets (IBL, Allen Institute data, etc.) – for example, does a latent factor correspond to running speed as expected, or does a learned inhibitory connection correspond to an actual interneuron-mediated effect

seen in optogenetic studies? By iterating between model insights and experimental evidence, NeuroFM-X can continuously improve both as a brain emulator and as a brain **theory**.

The aspiration to *"uniquely analyze how the brain works"* with such a model is bold, but entirely plausible with these innovations. With NeuroFM-X serving as a unifying platform, and by incorporating biologically grounded, interpretable methods, we take one more step toward **changing the world** – where understanding the brain's algorithms leads to breakthroughs in medicine, AI, and beyond. Each enhancement rooted in advanced math or neuroscience brings us closer to a model that is not just powerful, but also **transparent, explanatory, and truly brain-like** in its operation.

---


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 22 23 NEUROFM_X_PLAN.md
https://github.com/sidhulyalkar/neurOS-v1/blob/6788263906f59e8d1a3fe1c84f25551b84fa7591/docs/NEUROFM_X_PLAN.md

20 31 38 44 45 Decoding the brain: From neural representations to mechanistic models - PubMed
https://pubmed.ncbi.nlm.nih.gov/39423801/?utm_source=FeedFetcher&utm_medium=rss&utm_campaign=None&utm_content=1FcSKUuQkQG3EOhBkj91uHouHARUJNZOFRrbjh1Bk9KB1gWgq8&fc=

24 25 26 README.md
https://github.com/sidhulyalkar/neurOS-v1/blob/6788263906f59e8d1a3fe1c84f25551b84fa7591/packages/neuros-neurofm/README.md

27 28 29 Latent circuit inference from heterogeneous neural responses during cognitive tasks | Nature Neuroscience
https://www.nature.com/articles/s41593-025-01869-7?error=cookies_not_supported&code=acfbaabc-dac4-4dad-9b11-1e856d55364b

30 Multilevel interpretability of artificial neural networks: leveraging framework and methods from neuroscience
https://arxiv.org/html/2408.12664v2

32 33 34 35 36 37 46 Mechanistic Interpretability: A Challenge Common to Both Artificial and Biological Intelligence - Kempner Institute
https://kempnerinstitute.harvard.edu/research/deeper-learning/mechanistic-interpretability-a-challenge/

39 Building mechanistic models of neural computations with simulation ...
https://www.world-wide.org/bernstein-24/building-mechanistic-models-neural-903ef33b

40 41 42 Mechanism-based organization of neural networks to emulate systems biology and pharmacology models | Scientific Reports
https://www.nature.com/articles/s41598-024-59378-9?error=cookies_not_supported&code=ac7ee91b-4c96-4520-bcb7-dc4a0a025fa7

43 Virtual Brain Inference (VBI): A flexible and integrative toolkit for ...
https://elifesciences.org/reviewed-preprints/106194v1