Contribution to

EBRAINS 10-Year Roadmap 2026 – 2036

MechAIBrain: mechanistic AI brain models for behavioral, neural, and biophysical predictions

Acronym: **MechAIBrain**
Co-lead proposer: Martin **Schrimpf** (EPFL, CH), Walter **Senn** (UniBE, CH)
Contributors: Alexander **Mathis** (EPFL, CH), Mihai **Petrovici** (UniBE, CH), Rainer **Goebel** (U Maastricht, NL), Ausra **Saudargienė** (LSMU, LT); Philippe **Ryvlin** (CHUV, CH), Marmaduke **Woodman** (U Aix-Marseille), Viktor **Jirsa** (U Aix-Marseille, F)

Institutions: Ecole Polytechnique Fédérale de Lausanne (EPFL, CH), University of Bern (CH), Centre Hospitalier Universitaire Vaudois (CHUV, CH), Aix-Marseille University (F), University of Maastricht (NL), Lithuanian University of Health Sciences (LT)

SUMMARY
This proposal introduces a paradigm shift for EBRAINS through a Hybrid Modeling Strategy designed to create a Mechanistic Foundation Model of the Human Brain and Mind. For enabling new medical applications, we need to merge mechanistic models of the brain with the computational power of AI, and to optimize these hybrid brain models for behavioral performance and alignment with neural data. Ultimately, this will transform neuroscience towards a predictive, engineering-grade science, providing powerful in-silico tools for basic research and translations to clinical interventions.

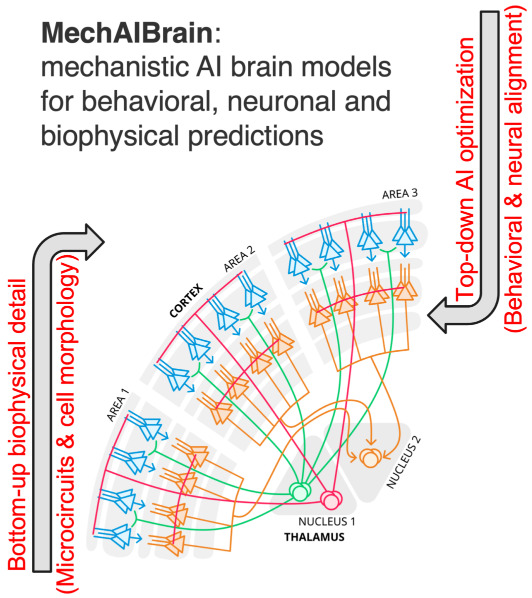

Table of Contents

1. A mechanistic foundation model of the human brain and mind

Historically, brain modeling has been divided into data-driven "bottom-up" approaches maximizing biological realism, and task-driven "top-down" approaches maximizing functional performance. This proposal introduces a paradigm shift for EBRAINS through a hybrid modeling strategy designed to create a Mechanistic Human Brain Foundation Model. By merging biophysical models of the brain with computational paradigms of AI, and optimizing these hybrid brain models for behavioral performance and alignment with neural data, we aim to bridge the gap between biophysics and behavior. This framework leverages EBRAINS' infrastructures with brain simulation environments and anatomy atlases to constrain predictive AI models to the real biology. It also links to the US BRAIN Initiative aiming to advance AI-based neurotechnologies to capture brain activity at the level of interacting cells and circuits. While unprecedented progress has been made with modern AI in predicting and influencing brain activity and behavioural outcomes, the alignment to brain data seems to have saturated and these models lack biophysical detail. On the other hand, mechanistic brain models lack the predictive power of AI models in reproducing large-scale brain activity, cognition and behaviour. Casting the computational principles of predictive AI in the form of mechanistic brain models is a promising avenue for further progress in understanding and treating the brain. The suggested roadmap lays out how to construct such mechanistic foundation models of the brain and mind, and proposes to apply these models to study the brain and mind in function and dysfunction.

2. Grand challenge: merging predictive AI with mechanistic brain models

The grand challenge addressed is the historical dichotomy in brain modeling, which has prevented a holistic understanding of human intelligence.

- **Data-driven, bottom-up approaches** (such as The Virtual Brain) offer high biological fidelity but **face significant scaling challenges** and the risk of overfitting to microscopic details. In fact, beside the cognitive computation our brains implement, major parts of the biophysical processes are in place to sustain the functionality of the tissue and keep it alive. These biophysical processes introduce correlations across various spatial and temporal scales that makes the bottom-up extraction of the underlying computational principle a nearly unsolvable endeavor.

- Conversely, **task-driven, top-down approaches** (such as Artificial Neural Networks of predictive AI) excel at complex behavioral tasks and neural population activity (see e.g. Brain-Score), but they often **lack biological realism**. In fact, current AI models are structurally under-determined to explain known computational principles in the brain, and hence might miss crucial mechanisms necessary to explain brain diseases.

- This proposal seeks to resolve this grand challenge by creating human brain models that are both biophysically grounded and functionally competent, starting from a pragmatic **middle ground between biophysical realism and functional abstraction**.

- The key new ingredient is that "middle ground hybrid models", capturing both cortical structures and AI computational principles, are **for the first time able to solve cognitive tasks** (such as human language processing), and **can be extended in both directions, towards more biophysical details and other behavioural tasks** (e.g. Granier & Senn, 2025).

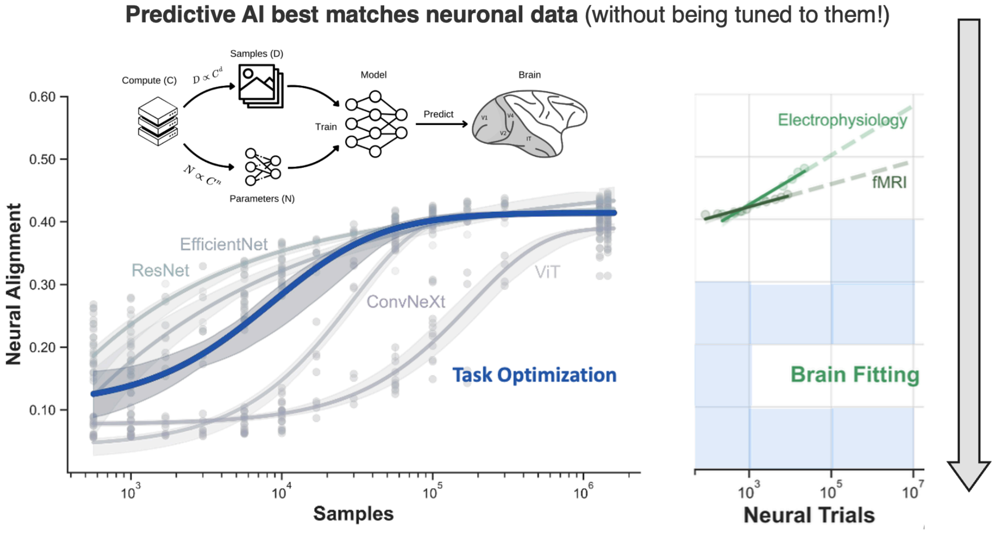

Figure 2: Top-down models of predictive AI do not further improve the neuronal alignment when providing more data samples (left; adapted from [Gokce & Schrimpf, ICML 2025](); brain data from [Brain-Score]()). By training these AI models on specific neuronal data the alignment further improves (right).

3. Why EBRAINS? – EBRAINS offers a unique infrastructure of mechanistic brain models

EBRAINS' infrastructure and computational capabilities and the substantial prior EU investments in data acquisition are essential components of this hybrid modeling strategy.

- The existing "bottom-up" data repositories on **EBRAINS**—such as detailed atlases, connectomes, and electrophysiological recordings—can serve as the **vital ground-truth constraints for these mechanistic foundation models.**

- This approach aims to provide functional and behavioral context to EBRAINS' rich existing data repositories, bringing them "to life."

- High-performance infrastructure, such as **EBRAINS' JUPITER, are essential for scaling the envisioned hybrid models**.

- Furthermore, validating these computational models requires **rigorous benchmarking to evaluate representational similarity and neural alignment**, which will build directly on the infrastructure and data archives already **established by EBRAINS**.

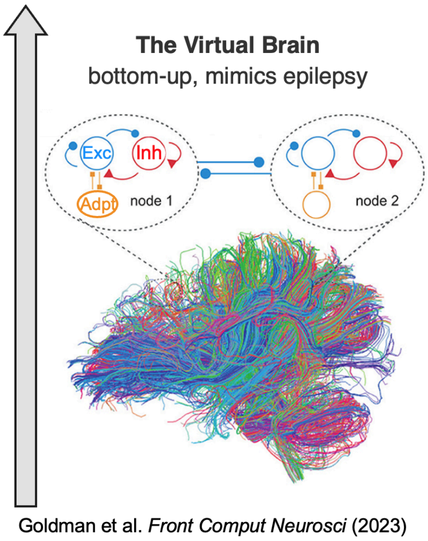

Figure 3: Bottom-up models of EBRAINS infrastructure, like [The Virtual Brain] (TVB) based on whole-brain neural mass models (image adapted from Goldman et al, 2023), outperform classical models to estimate the epileptogenic zone ([Wang et al., 2025]). Yet, as compared to predictive AI models, they align less well to neuronal and behavioral data involving cognitive tasks (such as language production, Fig. 2).

4. Key methods: morphing predictive AI to a mechanistic brain model

The core approach is a **hybrid modeling strategy** that integrates both bottom-up and top-down philosophies.

- Models will be simultaneously optimized for behavioral performance and alignment with neural data at multiple scales.
- Recent work demonstrates that models jointly trained on ecological tasks and neural recordings generalize better (Fig. 2).
- The top-down objective will act as a functional regularizer, restricting the solution space to mechanisms that are behaviorally effective.
- The bottom-up objective will act as a biological regularizer, restricting the solution space to mechanisms that are biologically feasible.
- A major activity involves enabling high-throughput "in-silico" experimentation to perform neural hypothesis searches over thousands of possibilities.

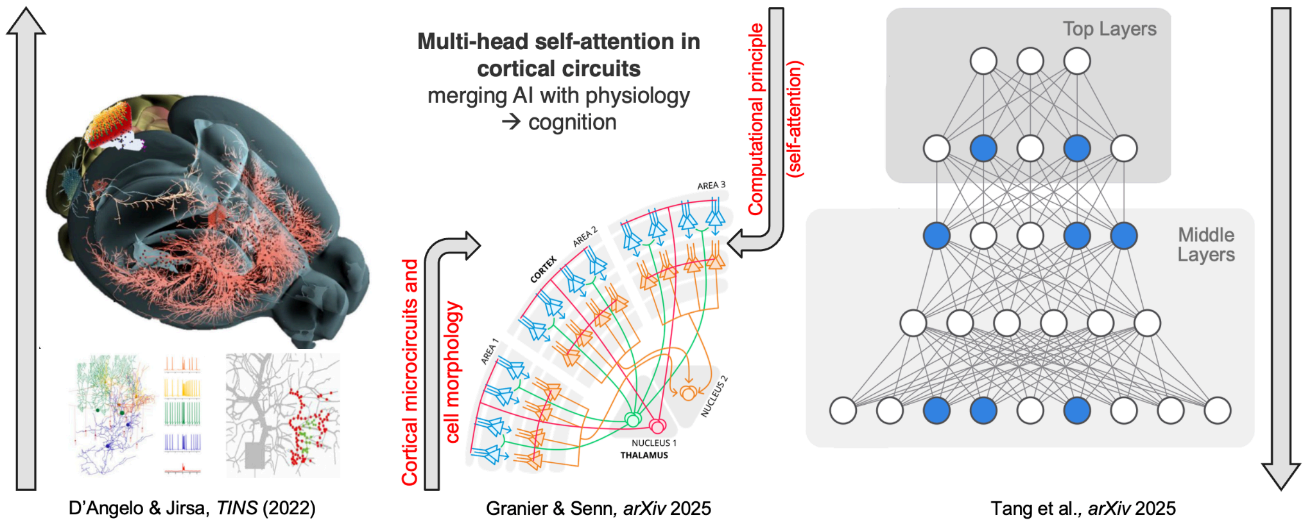

Figure 4: Merging bottom-up mechanistic models from EBRAINS (left) with top-down predictive AI models (right) to form a mechanistic foundational model of the human brain and mind (middle, example from [Granier & Senn, 2025]). Such hybrid models promise highest alignment to brain data and best predictive power to cure brain diseases.

5. Expected Impact – Personalized brain model to cure diseases

Crucially, the integration of AI in this roadmap extends far beyond using artificial neural networks as mere classifiers. Instead, we aim to build **stimulus-computable digital twins**. While current NeuroAI models may not yet capture the macroscopic biophysical detail of platforms like The Virtual Brain, they uniquely process the exact same sensory inputs as the biological brain to predict the responses of individual neurons across large populations. This stimulus-computability bridges the gap between sensory input, neural representation, and behavioral output, enabling unprecedented interventional applications.

Model Building and Validation
To construct these brain models, our approach relies on several proven scaling and validation principles:

- **Task-Optimization & Alignment:** Task-optimization inherently drives brain alignment. By training models on ecological tasks across multiple modalities—including vision (images and videos), language, and audio—internal representations naturally converge with biological brain processing (Yamins et al. 2014; Khaligh-Razavi et al. 2014; Kell et al. 2018; Schrimpf et al., 2018, 2021; Caucheteux et al. 2022; Vargas et al. 2024; Tang et al., 2025).

- **Large-Scale Benchmarking:** Validating these computational proxies requires rigorous, large-scale benchmarking to advance NeuroAI models, leveraging established community platforms like Brain-Score, Algonauts, or Sensorium, to systematically test neural and behavioral alignment against ground-truth data (Schrimpf et al., 2020; Cichy et al., 2019; Willeke et al. 2022).

- **Scaling Laws and Neural Data:** While task optimization alone eventually plateaus in its ability to align with brain data, recent scaling laws for human fMRI and primate data demonstrate that directly training on more neural data log-linearly improves model predictions (Antonello et al. 2023; Gokce & Schrimpf, 2025).

- **Generalization:** We build upon strong foundational evidence that large-scale foundation models trained on extensive animal data (e.g., mouse visual cortex and behavioral data) generalize remarkably well to new subjects and task domains, providing a technical blueprint for scaling up to human brain foundation models (Wang et al., 2025; Ye et al., 2024).

- **Gradient-based biological learning rules:** To connect to cortical circuits with realistic plasticity rules, that at the same time ensure gradient-descent learning on behavioural loss functions, the mechanistic AI brain models will explore biologically plausible learning rules (e.g., Haider et al., 2021, Senn et al., 2024, Ellenberger et al., 2025, Granier & Senn, 2025).

Towards Clinical Applications and Behavioral Impact

The expected impact spans basic science discovery and direct clinical translation, leveraging the predictive power of digital twins across distinct input-output pathways:

- **Image ⟶ Neurons (Sensory Stimulation):** Digital twins enable the neural control of electrodes in primate inferotemporal (IT) cortex via model-synthesized images, demonstrating that computationally generated visual stimuli can predictably push targeted neural populations into desired states (Bashivan et al., 2019; Walker et al. 2019; Guo et al. 2022; Wang et al., 2026).

- **Text ⟶ Neurons (Sensory Modulation of Cognitive Areas):** In higher-level cognition, large language models can identify the specific sentences that drive or suppress the human language system, offering a powerful, non-invasive tool to map and modulate functional networks in individuals (Tuckute et al., 2024).

- **Neurons ⟶Behavior (Restorative Prosthetics):** By mapping how neural activity translates to perception, model-guided microstimulation can induce complex visual percepts and steer primate behavior (Mehrer et al., 2025; Moure et al., 2025). This establishes the critical "software" for next-generation cortical prosthetics to bypass damaged sensory organs.

- **Image ⟶ Behavior (Accelerating Learning):** Foundation models can also optimize behavioral training directly; utilizing a model-informed curriculum improves human visual category learning and could accelerate cognitive rehabilitation (Talbot et al., 2025).

- **Computational Psychiatry & Aging:** Because they are stimulus-computable, these models allow researchers to simulate circuit disruptions *in-silico*. This provides mechanistic targets for conditions like Schizophrenia, personalized targeting for Deep Brain Stimulation in depression (e.g. Scangos et al., 2021), and the ability to predict cognitive decline trajectories by artificially "aging" network parameters (Honarmand et al., 2026).

- **Resting state activity as early diagnostic marker**. Capturing the computational properties of generative AI in a mechanistic brain model might also allow us to model mind wandering, and relate this to resting state or dream activity in the absence of external stimuli (e.g. Deperrois et al., 2022). Recreating resting state activity with a direct semantic interpretation likely improves diagnostic markers, e.g. for schizophrenia and psychosis (Wolff & Northoff, 2024).

6. Synergies – Links to (i)EEG, fMRI, BCI and NeuroAI initiatives

The MechAIBrain proposal establishes direct synergies with various other proposals and initiatives:

- It supports **The Virtual Brain** (TVB) in approaching e.g. schizophrenia (Viktor Jirsa) and connects to the EBRAINS **Virtual Brain Twin** for Personalised Treatment of Psychiatric Disorders. It supports the interpretation of **intracranial EEG** data from **EBRAINS Switzerland** (Philippe Ryvlin, the **Human Intracerebral EEG Platform**, HIP), and from

EBRAINS Italy (Maurizio Corbetta, the **Platform for Human Imaging**, PHI).

- The mechanistic AI-Brain models will help to interpret the next-generation (14 Tesla) **human fMRI** data at (Nijmegen / Maastricht, Rainer Goebel), as well as large scale intracranial EEG (>100 participants, Maastricht, Christian Herff). MechAIBrains will be fed by the clinical data provided from **EBRAINS Lithuania** (Ausra Saudargiene), and will contribute to the **Functional Whole-Brain Models** (Mario Senden) and **AI benchmarking** initiatives (LITMUS, Fabian Sinz).

- The MechAIBrain models also integrate with predictive AI validation platforms in our community, such as **Brain-Score** (Martin Schrimpf, EBRAINS Switzerland**)** with its tight links to AI-boosted neurotechnology and **Brain Computer Interfaces** (BCIs) developed by the **US BRAIN Initiative**, the BIM4BMI consortium, and **EBRAINS-Neurotech**.

7. Implementation Outlook – Milestones, timeframe, resources

To realize the Mechanistic Foundation Model within the 2026–2036 timeframe, we propose the following phased approach and resource allocation:

Milestones & Timeframe:

- **Phase 1: Integration & Benchmarking (Years 1–3):** Merge EBRAINS' bottom-up data repositories (e.g., The Virtual Brain, Julich Brain Atlas) with top-down predictive AI frameworks.
    - *Deliverable:* A unified, EBRAINS-hosted benchmarking infrastructure (e.g., building on Brain-Score) to rigorously evaluate both biophysical feasibility, and neural & behavioral alignment.

- **Phase 2: Scaling & High-Throughput In-Silico Trials (Years 4–6):** Scale hybrid models to perform rapid, high-throughput "in-silico" hypothesis testing, establishing the virtuous cycle of computational prediction and in-vivo verification.
    - *Deliverable:* Validated mechanistic targets for function and dysfunction (e.g., circuit disruption in Schizophrenia) and early-stage identification of personalized neuromodulation protocols.

- **Phase 3: Clinical Translation & Personalization (Years 7–10):** Deploy patient-specific Foundation Models into clinical environments.
    - *Deliverable:* Direct clinical application of "Virtual Brain Twins" for neuroprosthetics (universal decoders) and personalized Deep Brain Stimulation (DBS) targeting conditions like treatment-resistant depression.

We believe that such a MechAIBrain roadmap would deserve a **specific ERC call**, comparable to the second round of the US BRAIN Initiative.

8. References

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Deperrois, N., Petrovici, M. A., Senn, W., & Jordan, J. (2022). Learning cortical representations through perturbed and adversarial dreaming. *ELife*, 1–34. https://doi.org/10.7554/elife.76384

Ellenberger, B., Haider, P., Benitez, F., Jordan, J., Max, K., Jaras, I., Kriener, L., & Petrovici, M. A. (2025). Backpropagation through space , time and the brain. *Nature Communications*, *December*(July 2024), 1–23. https://doi.org/10.1038/s41467-025-66666-z

Gokce, A., & Schrimpf, M. (2025). Scaling Laws for Task-Optimized Models of the Primate Visual Ventral Stream. *Proceedings of the 42 Nd International Conference on Machine Learning*, vol. 267, 1–22. https://arxiv.org/pdf/2411.05712

Goldman, J. S., Kusch, L., Aquilue, D., Yalçınkaya, B. H., Depannemaecker, D., Ancourt, K., Nghiem, T. E., Jirsa, V., & Destexhe, A. (2023). A comprehensive neural simulation of slow-wave sleep and highly responsive wakefulness dynamics. *Frontiers in Computational Neuroscience*, *January*, 1–17. https://doi.org/10.3389/fncom.2022.1058957

Granier, A., & Senn, W. (2025). Multihead self-attention in cortico-thalamic circuits. *Arxiv*, *v2*, 1–15. https://arxiv.org/pdf/2504.06354

Haider, P., Kriener, L., Ellenberger, B., Jordan, J., Senn, W., & Petrovici, M. (2021). Latent Equilibrium: A unified theory for arbitrarily fast computation with arbitrarily slow neurons. *NeurIPS*, 1–20.

Mehrer, J., Lonnqvist, B., Mitola, A., Gokce, A., Papale, P., & Schrimpf, M. (2025). Model-Guided Microstimulation Steers Primate Visual Behavior. OpenReview / ICLR. https://arxiv.org/abs/2510.03684

Scangos, K. W., Khambhati, A. N., Daly, P. M., Makhoul, G. S., Sugrue, L. P., Zamanian, H., Liu, T. X., Rao, V. R., Sellers, K. K., Dawes, H. E., Starr, P. A., Krystal, A. D., & Chang, E. F. (2021). Closed-loop neuromodulation in an individual with treatment-resistant depression. *Nature Medicine*, *27*(October), 1696--1700. https://doi.org/10.1038/s41591-021-01480-w

Schrimpf, M., et al. (2020). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? bioRxiv. https://doi.org/10.1101/407007

Schrimpf, M., Asher, I., Tuckute, G., Kauf, C., & Hosseini, E. A. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *PNAS*, *4*. https://www.pnas.org/doi/10.1073/pnas.2105646118

Talbot, M. B., Kreiman, G., DiCarlo, J. J., & Gaziv, G. (2025). L-WISE: Boosting Human Visual Category Learning Through Model-Based Image Selection and Enhancement. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2412.09765

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. Nature Human Behaviour, 8(3), 544-561. https://doi.org/10.1038/s41562-023-01783-7

Wang, E. Y., Fahey, P. G., Ding, Z., Papadopoulos, S., Patel, S., Ding, Z., Tran, D., Fu, J., & Schneider-mizell, C. M. (2025). Foundation model of neural activity predicts response to new stimulus types. *Nature*, *640*(April 2023), 470–483. https://doi.org/10.1038/s41586-025-08829-y

Wang, H. E., Dollomaja, B., Triebkorn, P., Duma, G. M., Williamson, A., Makhalova, J., Lemarechal, J., Bartolomei, F., & Jirsa, V. (2025). Virtual brain twins for stimulation in epilepsy. *Nature Computational Science*, *5* (September 2025), 754-768. https://doi.org/10.1038/s43588-025-00841-6

Wolff, A., & Northoff, G. (2024). Temporal imprecision of phase coherence in schizophrenia and psychosis — dynamic mechanisms and diagnostic marker. *Molecular Psychiatry*, *29*(April 2023), 425–438. https://doi.org/10.1038/s41380-023-02337-z

Ye, S., Filippova, A., Lauer, J., Schneider, S., Vidal, M., Qiu, T., Mathis, A., & Mathis, M. W. (2024). SuperAnimal pretrained pose estimation models for behavioral analysis. *Nature Communications*, *March 2023*. https://doi.org/10.1038/s41467-024-48792-2