Multi-Timescale State Space Models and Architectures for EEG-Like Signal Processing

Mamba variants, hierarchical SSMs, and multi-resolution architectures that handle multiple frequency bands, spatial relationships, and nonlocal dependencies have rapidly evolved from 2023-2025, offering powerful solutions for biomedical signal processing with multi-frequency characteristics.

Recent advances demonstrate three key approaches: explicit hierarchical SSM stacking (HiSS with 23% performance improvement), Hiss-csp +2 parallel multi-resolution processing (MS-SSM), and hybrid architectures combining SSMs with attention mechanisms. These architectures successfully address all four requirements—causal modeling, multi-frequency processing, spatial hierarchies, and nonlocal dependencies—while maintaining computational efficiency with linear time complexity.

Core State Space Model Architectures for Multi-Timescale Processing

Hierarchical State Space Models (HiSS) - The Breakthrough Architecture

HiSS (Bhirangi et al., CMU, February 2024) represents the most direct solution for multi-timescale SSM modeling. The architecture explicitly stacks SSMs vertically to create temporal hierarchy with **two levels capturing different timescales**. Hiss-csp arXiv The low-level SSM processes input sequences in chunks and outputs local chunk features, while the high-level SSM processes the sequence of chunk features to output global predictions. arXiv This hierarchical information flow from fast to slow timescales achieved **23% better MSE** than flat SSMs (S4, Mamba), Transformers, and LSTMs on six real-world sensor datasets. Hiss-csp +2

The architecture divides sequences into temporal chunks where the low-level SSM extracts fine-grained features within each chunk (fast timescale) and the high-level SSM captures coarse temporal dependencies across chunks (slow timescale). arXiv This design is particularly relevant for EEG applications where you need to capture both rapid event-related potentials (100-200ms) and slower oscillatory patterns (seconds to minutes). **Implementation**: https://hiss-csp.github.io/ with full code and CSP-Bench dataset.

Multi-Scale State Space Model (MS-SSM) - Parallel Multi-Resolution Processing

MS-SSM (2024, OpenReview) processes sequences at multiple resolutions simultaneously through multiple independent SSMs, one per resolution/scale. Each SSM is initialized for its target timescale with scale-dependent initialization, and an input-dependent scale-mixer dynamically fuses information across resolutions. The architecture captures both fine-grained, high-frequency patterns (fast timescale) and coarse, global trends (slow timescale) through parallel processing. OpenReview openreview

For EEG-like signals with distinct frequency bands (delta: 0.5-4 Hz, theta: 4-8 Hz, alpha: 8-13 Hz, beta: 13-30 Hz, gamma: 30-100 Hz), you could configure five parallel SSMs with timing parameters tuned to each band. PubMed Central The architecture consistently outperforms prior SSMs on Long Range Arena, hierarchical reasoning tasks, time series classification, and image recognition while maintaining enhanced memory efficiency. This addresses the limited effective memory of traditional SSMs while capturing multi-scale dependencies essential for complex structures. OpenReview openreview

Mamba and Mamba-2: Foundation Architectures with Learnable Timescales

Mamba (Gu & Dao, December 2023) introduced the Selective SSM (S6) with input-dependent parameters, including a **learnable timescale parameter (Δ)** that enables adaptive temporal resolution based on content. The discretization step Δ is learned from input, allowing selective temporal resolution adjustment—critical for signals where different segments require different processing speeds. With linear time and memory complexity $O(N)$ versus $O(N^2)$ for transformers, Mamba achieves 5× higher throughput while matching transformers twice its size. arXiv arxiv

Mamba-2 (Dao & Gu, May 2024) through State Space Duality (SSD) provides 2-8× faster training by using matrix multiplications instead of scan operations. The block decomposition algorithm combines linear recurrence with quadratic attention, enabling processing at different temporal resolutions through chunked computation. Goombalab Tri Dao The framework unifies SSMs and structured masked attention through semiseparable matrices, providing theoretical connections that enable hybrid algorithms combining benefits of both. arXiv Tri Dao **Code**: https://github.com/state-spaces/mamba with multiple pretrained models (130M to 2.8B parameters).

FlowState: IBM's Timescale-Invariant Architecture

FlowState (IBM Research, 2024) uses S5 SSM encoder with novel decoder and basis functions for continuous-time forecasting. The key innovation is **dynamic timescale adjustment**—analyze at one scale, predict at another—with timescale-invariant hidden state representation. IBM The architecture achieves seamless scale shifting that generalizes to unseen timescales through learnable timescale parameters adapting to different frequencies. IBM With only 9.1M parameters, FlowState ranks #2 on GIFT-Eval leaderboard for zero-shot forecasting, outperforming models 20× larger and representing the only SSM-based model in top rankings. IBM

Hybrid SSM Architectures Combining Multiple Mechanisms

Jamba: Production-Scale Hybrid Transformer-Mamba

Jamba (AI21 Labs, March 2024) interleaves Transformer and Mamba layers with Mixture-of-Experts (MoE) integration, using a 1:7 attention-to-Mamba layer ratio optimized for efficiency. Wikipedia The architecture handles 256K token context windows (largest among open models at release) with only 4GB KV cache versus 32GB for pure attention. arXiv OpenReview Transformer layers capture complex dependencies at multiple scales while Mamba layers provide efficient long-range context with constant memory, achieving 3× throughput versus Mixtral on long contexts. arXiv

For EEG applications, this translates to processing hours of continuous recording while maintaining fast oscillation detection capability. The hybrid design enables both local high-frequency feature extraction (Mamba) and global pattern recognition (Transformer). **Updated version**: Jamba-1.5 (August 2024) with 94B/12B active parameters and ExpertsInt8 quantization. arXiv **Models**: ai21labs/Jamba-v0.1 on HuggingFace.

BlackMamba and MoE-Mamba: Efficient Conditional Processing

BlackMamba (Zyphra, February 2024) replaces both MLPs with MoE layers and attention with Mamba layers, offering **linear time and memory for generation**. The architecture combines Mamba's linear complexity for long sequences with MoE's conditional computation based on token content, enabling fast inference with cheap computation across scales. Wikipedia With 340M/1.5B and 630M/2.8B parameter configurations using 8 experts with top-2 routing, the model achieves competitive performance with better inference and training FLOPs. arXiv arXiv **Code**: https://github.com/Zyphra/BlackMamba

MoE-Mamba (January 2024) alternates Mamba layers with MoE feed-forward layers, where Mamba integrates entire sequence context (global temporal view) and MoE applies relevant expert per token (local specialized processing). This achieves **2.2× fewer training steps** than Mamba for same performance while preserving Mamba's inference gains versus Transformers. arXiv Emergent Mind

Multi-Timescale Architectures from Video and Audio Processing

SlowFast Networks: Dual-Pathway Multi-Rate Processing

SlowFast Networks (Feichtenhofer et al., Facebook AI, ICCV 2019) established the dual-pathway paradigm with biological inspiration from retinal parvocellular (P-cells: fine spatial detail, slow) and magnocellular cells (M-cells: high temporal frequency, less spatial). ScienceDirect +3 The **Slow Pathway** operates at low frame rate ($\tau=16$) capturing spatial semantics with full channel capacity, while the **Fast Pathway** operates at high frame rate ($\alpha=8$ more frames) capturing motion with lightweight $\beta=1/8$ channel capacity. Medium +2

Lateral connections fuse information between pathways through: (1) time-to-channel reshaping, (2) time-strided sampling, and (3) time-strided convolution with 5×1^2 kernel. Medium +2 For EEG applications, configure the Slow Pathway to process low-frequency bands (delta, theta, alpha: 0.5-13 Hz) with sparse temporal sampling but full spatial coverage across all electrodes, while the Fast Pathway processes high-frequency bands (beta, gamma: 13-100 Hz) with dense temporal sampling but reduced channel capacity focusing on key electrodes. PubMed Central The Fast Pathway represents only ~20% of total computation despite 8× sampling rate. Medium +2 **Code**: https://github.com/facebookresearch/SlowFast

Adaptations include **SF-TMN** (Zhang et al., 2024) for surgical phase recognition with frame-level and segment-level temporal modeling, PubMed +2 and **SlowFast-LLaVA** (2025) enabling 128 frame processing versus 16 frame baseline with only 3% computation increase. arXiv

Temporal Dendritic Heterogeneity in Spiking Neural Networks

DH-LIF (Dendritic Heterogeneity Leaky Integrate-and-Fire, Zheng et al., Nature Communications 2024) provides the most biologically-inspired multi-timescale architecture. The multi-compartment model includes 1 soma plus multiple dendrite compartments, where each dendritic branch has temporal memory with timing factor α_d and the soma has timing factor β for membrane potential. The architecture learns heterogeneous timing factors on different branches, enabling **intra-neuron multi-timescale integration**. nature +2

Mathematically, dendritic current dynamics follow: `i_d^(t+1) = α_d * i_d^t + I_d^t` and soma membrane potential: `u^(t+1) = β * u^t + Σ_d i_d^(t+1) - u_th * o^t` where timing factors are constrained α_d, β ∈ [0,1] via sigmoid. **Long-term memory** arises because dendritic currents never reset (unlike membrane potential), enabling branches with large α_d (0.8-0.95) to memorize low-frequency signals while branches with small α_d (0.2-0.5) track high-frequency signals. [nature] [Nature]

Performance on temporal tasks: **92.1% on Spiking Heidelberg Digits** (previous best: 90.4%), **82.46% on Spiking Speech Commands** (previous best: 74.2%), with 36-45% fewer parameters than baselines and superior robustness to noise. [Nature] The neuromorphic hardware implementation enables efficient real-time processing. [nature] For EEG, you would configure dendritic branches matching each frequency band—Branch 1 (α_d ≈ 0.9) captures delta/theta, Branch 2 (α_d ≈ 0.7) captures alpha/beta, Branch 3 (α_d ≈ 0.3) captures gamma/ripple. [PubMed Central] **Code**: https://github.com/eva1801/DH-SNN

Wavelet-Based Neural Networks for Frequency Decomposition

Wav-KAN (Wavelet Kolmogorov-Arnold Networks, 2024) replaces traditional activations with wavelet functions, using both Continuous Wavelet Transform (CWT) for potential and Discrete Wavelet Transform (DWT) for efficiency. The architecture provides **multiresolution analysis** capturing both high-frequency and low-frequency components with orthogonal/semi-orthogonal basis maintaining balance between data structure representation and noise overfitting. [arXiv +2]

DOG and Mexican Hat wavelets outperform standard Spl-KAN, with **no grid dependency** unlike B-splines. The architecture achieves better performance with fewer parameters, enhanced accuracy, faster training, and increased robustness to noise. For EEG, wavelets directly decompose into frequency bands (delta, theta, alpha, beta, gamma) with interpretable coefficients. [PubMed Central] **Code**: https://github.com/zavareh1/Wav-KAN with PyTorch/TensorFlow implementations.

DeSpaWN (Deep Sparse Wavelet Network, PNAS 2022) mimics Fast Discrete Wavelet Transform cascade using neural networks with **learnable filters** at each decomposition level. The architecture learns high-pass and low-pass filters with learnable hard-thresholding (continuous, differentiable approximation) while maintaining Conjugate Quadrature Filter (CQF) property ensuring perfect reconstruction. [PNAS] The tree-structured encoder-decoder with skip connections uses only a few hundred parameters, making it extremely lightweight while handling variable-length sequences. [PNAS] The unsupervised learning approach suits unlabeled data with sparse representation reducing noise.

Multi-Resolution Time-Series Transformer (MTST)

MTST (2024, AISTATS) employs multi-branch architecture for simultaneous modeling at different resolutions through **patch-based tokenization** with different patch sizes for different temporal scales. Large patches capture low-resolution coarse patterns and long-term trends, while small patches capture high-resolution fine-grained local patterns. arXiv arXiv **Relative positional encoding** proves better suited for extracting periodic components at different scales compared to absolute encoding. arXiv

The multi-resolution construction via patch-size adjustment (not pooling) avoids information loss from subsampling while maintaining chronological order. arxiv arXiv For EEG, configure short patches (50-100ms) for fast event-related potentials and longer patches (1-2 seconds) for slow oscillatory patterns. State-of-the-art performance on ETTh1, ETTh2, Weather, and Traffic benchmarks demonstrates superiority on complex temporal patterns. arxiv The architecture handles both amplitude and trend prediction better than PatchTST. **Code**: https://github.com/networkslab/MTST

Multi-Scale Neural Audio Codec (SNAC)

SNAC (NeurIPS 2024 Workshop) extends Residual Vector Quantization (RVQ) where quantizers operate at different temporal resolutions, creating **hierarchical token structure** with coarse tokens sampled at ~10 Hz and fine tokens at higher rates. arXiv arXiv The encoder-decoder architecture based on RVQGAN includes noise blocks for input-dependent Gaussian noise, depthwise convolutions for efficiency, and local windowed attention at lowest temporal resolution. MarkTechPost arXiv

Variable frame rates follow hierarchies: 44kHz model [16, 32, 64, 128] tokens per sequence, 24kHz model [12, 24, 48] tokens. Hugging Face Hugging Face Temporal resolution adaptation means coarse tokens cover broader time spans while fine tokens capture local details through average pooling downsampling at each iteration and nearest-neighbor upsampling for reconstruction. arxiv GitHub Performance achieves near-reference quality below 1 kbit/s for speech, outperforming EnCodec and DAC at comparable bitrates, with 3-minute context modeling possible with coarse tokens. arxiv +2

For EEG, configure hierarchical quantization matching frequency bands—coarse tokens (1-2 Hz) for delta/theta, medium tokens (8-16 Hz) for alpha/beta, fine tokens (30-100 Hz) for gamma. PubMed Central The discrete representation enables sequence modeling and generative approaches. **Code**: https://github.com/hubertsiuzdak/snac

Spatial-Temporal Graph Neural Networks for Sensor Arrays

Graph WaveNet: Adaptive Spatial Dependencies with Multi-Scale Temporal Processing

Graph WaveNet (Wu et al., IJCAI 2019) learns hidden spatial dependencies through node embeddings **without requiring pre-defined graph structure**. PubMed The adaptive dependency matrix computed as $\tilde{A}\_adp = SoftMax(ReLU(E1 E2^T))$ where E1, E2 are learnable node embedding matrices enables discovery of functional connectivity patterns in EEG electrode arrays even when anatomical distances don't capture functional relationships. IJCAI +2

Dilated causal convolution with stacked dilated 1D convolutions provides temporal modeling where receptive field grows exponentially (2^L for L layers), enabling very long sequence modeling. Multi-scale spatial-temporal processing through multiple graph convolution operations at different temporal scales creates short-term features at lower layers and long-term features at higher layers. Medium IJCAI Skip connections aggregate multi-scale features with dilation factors following [1, 2, 1, 2, 1, 2, ...] pattern for progressive temporal expansion. Medium

The architecture handles missing or incomplete graph information while capturing both local and global spatial patterns efficiently. For EEG, the adaptive matrix automatically discovers functional connectivity between electrodes based on signal correlations rather than physical proximity. **Code**: https://github.com/nnzhan/Graph-WaveNet (PyTorch)

Attention-Based Spatial-Temporal Graph Neural Networks (ASTGCN)

ASTGCN (Guo et al., AAAI 2019) implements **three independent components** modeling different temporal scales: Recent Component (last hour short-term), Daily-Periodic Component (same time yesterday), and Weekly-Periodic Component (same time last week). University of Northampt... This multi-component hierarchical structure directly addresses EEG periodicities like circadian rhythms, sleep-wake cycles, and ultradian rhythms.

Spatial attention dynamically assigns weights to different sensor nodes with attention score: $S = V_s \cdot \sigma(((X^h)^T U_s U_s^T X^h + b_s) \odot E)$, capturing dynamic spatial correlations that adapt to changing network importance. **Temporal attention** assigns importance to different time steps: $E = V_e \cdot \sigma((X^h)^T U_e + U_e^T X^h + b_e)$, modeling evolving temporal dependencies focused on relevant historical periods.

Graph convolution layers apply to spatially-attended features using Chebyshev polynomial approximation for multi-hop spatial relationships. **Hierarchical feature fusion** combines components: $\hat{Y} = W_h * Y_h + W_d * Y_d + W_w * Y_w$ with learned optimal weights for different temporal scales in end-to-end trainable fashion. The attention mechanism provides nonlocal interactions while multi-scale temporal hierarchy captures different periodicities. **Code**: https://github.com/guoshnBJTU/ASTGCN-2019-pytorch (PyTorch)

Diffusion Convolutional Recurrent Neural Network (DCRNN)

DCRNN (Li et al., ICLR 2018) models spatial dependencies as **diffusion processes on directed graphs** through bidirectional random walks. The diffusion convolution formula $X :G f = \Sigma(k=0 \text{ to } K-1) (\theta_{k,1}(D_O^{-1} W)^k + \theta_{k,2}(D_I^{-1} W^T)^k) X$ captures both upstream and downstream influences through K-step diffusion capturing K-hop neighborhoods with weighted adjacency. IJCAI +2

The encoder-decoder architecture uses GRU cells with diffusion convolution where matrix multiplication is replaced with diffusion convolution in GRU operations: $r_t = \sigma(\Theta_r :G [X_t, H_{t-1}] + b_r)$ and $u_t = \sigma(\Theta_u :G [X_t, H_{t-1}] + b_u)$, enabling spatially-aware recurrent modeling. Wikipedia Scheduled sampling gradually transitions from teacher forcing during training, capturing long-term temporal dependencies while maintaining spatial awareness. ACM Digital Library

For large-scale networks, the **graph-partitioning approach** (Mallick et al., 2020) decomposes networks with 11,160 sensors into smaller sub-graphs with independent training and overlapping nodes, enabling scalability. Nature For EEG, bidirectional diffusion captures both afferent (sensory input) and efferent (motor output) signal propagation patterns across cortical networks. **Code**: https://github.com/liyaguang/DCRNN (TensorFlow), https://github.com/chnsh/DCRNN_PyTorch (PyTorch)

Non-Local Neural Networks for Global Context

Non-local Neural Networks (Wang et al., Facebook AI, CVPR 2018) compute response at each position as weighted sum of features at **ALL positions**, providing direct nonlocal dependency modeling. The nonlocal operation $y_i = 1/C(x) \Sigma_j f(x_i, x_j) g(x_j)$ includes pairwise function $f(x_i, x_j)$ computing affinity between positions i and j, and unary function $g(x_j)$ for feature representation. arXiv

The **embedded Gaussian** variant $f(x_i, x_j) = e^{(\theta(x_i)^T \varphi(x_j))}$ is equivalent to self-attention, learning which positions to attend while capturing long-range dependencies directly. Spacetime nonlocal operations extend to 3D (space + time) for video/sensor data, capturing dependencies across both spatial locations AND time in single operations for long-range spacetime correlations. Medium

Nonlocal blocks can be inserted into existing architectures with residual connections $z_i = W_z y_i + x_i$, using bottleneck design that reduces channels by half for efficiency. Medium For EEG sensor arrays, this enables modeling of distant electrode interactions (e.g., frontal-occipital coupling) without requiring explicit graph paths. The plug-and-play module adds computational cost $O(T \times H \times W)^2$ manageable with bottleneck, typically added in middle layers. **Code**: https://github.com/facebookresearch/video-nonlocal-net

Hierarchical Graph Neural Networks (HGNet)

HGNet (Rampasek & Wolf, 2021) guarantees message-passing paths of **logarithmic length**—for any two connected nodes, O(log N) path length versus standard GNN K-hop limitation. The multi-level hierarchical structure includes original input network (Level 0) and auxiliary network layers (Levels 1, 2, ..., L) where each level represents coarser graph abstraction.

Node features update through horizontal connections (within level) and vertical connections (between levels), simultaneously learning individual node features and aggregated network features with variable resolution representation. Community-based hierarchy using Louvain community detection creates super-nodes from communities at each level, providing multi-grained semantic encoding capturing meso and macro-level patterns. Springer

For EEG, hierarchical structure could represent: Level 0 (individual electrodes), Level 1 (local clusters/brain regions), Level 2 (hemispheric networks), Level 3 (global brain states). Hierarchical shortcuts reduce path lengths enabling efficient nonlocal aggregation through hierarchy with improved convergence and stability. **Code**: https://github.com/rampasek/HGNet

EEG-Specific Architectures with Multi-Frequency Modeling

Graph Neural Networks for EEG: DAMGCN and GGN

DAMGCN (Dual Attention Mechanism GCN, Chen et al., 2024) decomposes EEG into **5 frequency bands** using Short-Time Fourier Transform: δ (1-4 Hz), θ (4-8 Hz), α (8-13 Hz), β (13-30 Hz), γ (>30 Hz), extracting Differential Entropy features: $h(x) = (1/2)\log_2(2\pi e\sigma^2)$. The 3D spatial adjacency matrix constructed from electrode coordinates feeds into two-layer graph convolutional network using self-connected adjacency matrix $\tilde{A} = A + I$ with graph convolution $H^{(l+1)} = \sigma(\tilde{D}^{(-1/2)}\tilde{A}\tilde{D}^{(-1/2)}H^{(l)}W^{(l)})$ and residual connections preventing gradient vanishing. frontiersin

Dual attention mechanism includes electrode channel attention (weights for spatial locations) and frequency band attention (weights for different bands) using Transformer-based multi-head attention: $\text{Attention}(Q,K,V) = \text{softmax}(QK^T/\sqrt{d_k})V$. Learned weight coefficients for each band (initially 0.2, adapted during training) revealed δ band showed lowest weight while γ band was most prominent for emotion recognition. frontiersin Performance: **99.42% accuracy on SEED dataset, 97.50% on DEAP**, modeling functional connectivity between brain regions across frontal, temporal, parietal, and occipital lobes. frontiersin

GGN (Graph-Generative Neural Network, Li et al., 2022) dynamically discovers brain functional connectivity rather than using fixed topology. Nature nature The connectivity graph generator includes Para-Learner (three independent GNNs learning mixture Gaussian distribution parameters), Mix-Gaussian Module (generates probabilistic distribution of connections p_{ij}), and Gumbel-Sampler (makes sampling differentiable). nature Connection strength: $S_{ij} = \exp((\exp(p_{ij} \times \varepsilon) - 1)/\sigma)$.

The architecture uses 5-second sliding windows for dynamic connectivity detection, capturing transitions across resting state, onset, and recovery phases with non-linear functional connectivity through multi-layer generative networks. nature Band-filtered frequencies emphasize gamma band (31-51 Hz) for seizure detection. Performance: **91% accuracy classifying 7 seizure types** (outperforms CNN: 65%, GNN: 74%, Transformer: 82%). nature **Code**: https://github.com/ICLab4DL/GGN

Multi-Branch Parallel Processing: MFBPST-3D-DRLF and

MFBPST-3D-DRLF (Multiple Frequency Bands Parallel Spatial-Temporal 3D Deep Residual Learning, 2022) uses **parallel 3D deep residual CNNs** for each frequency band with separate branches processing delta, theta, alpha, beta, gamma independently. ScienceDirect Optimal frequency band selection via group sparse regression adapts per subject. 3D feature representation preserves spatial-temporal-spectral features reflecting neural signatures of different emotions. ScienceDirect Two-dimensional spatial attention automatically learns importance of spatial brain regions with adaptive weighting across electrode locations. ScienceDirect Band-specific feature extraction before fusion enables learning optimal band combinations per subject, achieving state-of-the-art on SEED and SEED-IV datasets.

MSDCGTNet (Multi-Scale Dynamic CNN and Gated Transformer Network, Cheng et al., 2024) performs end-to-end emotion recognition without time-frequency conversion. Multi-Scale Dynamic 1D CNN uses temporal convolution with multiple kernel sizes determined by frequency relationships: $K_i = \lfloor f \times \alpha_i \rfloor$ where for i=1,2,3: $\alpha = [0.125, 0.0625, 0.03125]$, capturing 8 Hz, 16 Hz, 32 Hz features covering alpha, beta, gamma bands. Spatial convolution (C×1) models electrode relationships with average pooling (1×4) for dimensionality reduction. nature Nature

Gated Transformer Encoder provides improved multi-head self-attention with linear complexity $O(n \cdot k)$ versus $O(n^2 \cdot d)$, projecting keys/values to lower dimension $k < n$. GLU (Gated Linear Unit) layers $h(x) = (W * x + b) \otimes \sigma(U * x + c)$ control information flow. Temporal Convolutional Network (TCN) with stacked residual blocks using dilated causal convolution extracts long-term temporal dependencies. Nature +2 Performance: **99.66% DEAP, 98.85% SEED, 99.67% SEED-IV**. Nature Implicit frequency band capture through multi-scale kernels avoids information loss from time-frequency conversion. nature

Transformer-Based Architectures: ETST and CTNet

EEG Temporal-Spatial Transformer (ETST) (Du et al., 2022) uses dual-branch architecture with Temporal Transformer Encoder (attention over sampling points/ time domain calculating correlation among temporal features with sine-cosine positional encoding) and Spatial Transformer Encoder (attention over electrode channels capturing coupling relationships with channel-specific positional encoding). Sequential processing follows: raw EEG → TTE → STE → classifier, preserving both temporal dynamics and spatial topology through multi-head attention: $\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O$ with feed-forward networks and residual connections plus layer normalization.

CTNet (Convolutional Transformer Network, 2024) combines convolutional module inspired by EEGNet (temporal convolution (1, 64) capturing 4+ Hz frequencies, depthwise convolution (22, 1) for spatial filtering, one-dimensional convolution) with Transformer encoder applying multi-head attention on convolutional features for global dependencies and critical feature emphasis. Two fully-connected layers with ReLU map to motor imagery classes, achieving state-of-the-art on BCI Competition IV-2a dataset. Nature

Phase Relationship Modeling: Cross-Frequency Coupling

Phase-Amplitude Coupling (PAC) using Reduced Interference Distribution (RID)-Rihaczek method couples low-frequency phase with high-frequency amplitude without requiring bandpass filtering while maintaining high frequency resolution through time-frequency distribution. Nature PubMed Central Measures modulation of high-frequency amplitude by low-frequency phase with applications in sensory detection, attention, visual perception, and memory processing. Nature PubMed Central Slow oscillations (delta, theta, alpha, beta) provide top-down control while fast oscillations (gamma, high-gamma) carry bottom-up information, with phase synchronization coordinating distributed neural processing. Cell Press

Cross-Frequency Phase Synchrony (CFS) implements n:m phase synchronization between frequency bands with stable phase difference between coupled oscillations. PLOS +2 Small frequency ratios (1:2, 1:3) prove most reliable with binary hierarchy showing 1:2 frequency relationships including delta-alpha coupling, theta-gamma coupling, and alpha phase coupling with beta/gamma amplitudes (1:2 to 1:4 ratios). PubMed Central PLOS Implementation requires filtering artifact control with validation through multiple methods to avoid false positives from non-sinusoidal signals.

PACNet (Phase-Amplitude Coupling Network, 2023) leverages PAC for adaptive filter bank design with end-to-end decoder providing task-specific frequency band location. Frontiers Extracts fine frequency bands from gamma range dynamically, improving ECoG neural decoding performance through PAC-guided band selection.

Classical Foundation: EEGNet and Variants

EEGNet (Lawhern et al., Army Research Laboratory, 2018) remains fundamental with temporal convolution (kernel size (1, K) where K ≈ sampling_rate/2) learning frequency filters similar to bandpass filters with F1 filters for initial feature maps. Depthwise spatial convolution (kernel (C, 1) where C = channels) learns spatial filters similar to CSP with depth multiplier D for feature expansion. Separable convolution (temporal (1, 16) followed by 1×1 pointwise) reduces parameters while maintaining expressiveness. GitHub +2 Max-norm constraints on weights, average pooling for temporal downsampling, and dropout layers provide regularization. Towards Data Science

Temporal filters implicitly learn frequency bands capturing alpha (8-13 Hz), beta (14-30 Hz), gamma (>30 Hz) without explicit frequency decomposition. Variants include **EEGNeX** (deeper with dilated convolutions), **3D-EEGNet** (3D spatial representation), and **ATCNet** (attention-augmented temporal convolution). Braindecode **Code**: https://github.com/vlawhern/arl-eegmodels (Keras/TensorFlow)

Architecture Selection Framework and Implementation Guide

Matching Architectures to EEG Requirements

For Causal Relationships in Markovian Processes: Choose architectures with explicit state evolution like **Mamba/Mamba-2** (selective state space with input-dependent dynamics preserving causal temporal flow), **DCRNN** (diffusion on directed graphs capturing directional influences), or **dilated causal convolutions** in Graph WaveNet and TCN variants (maintain temporal causality through masked future information).

For Multiple Frequency Band Modeling: Select **DAMGCN** (explicit 5-band decomposition with attention weighting), PubMed Central Frontiers **Wav-KAN** (wavelet activation functions providing interpretable multi-frequency decomposition), PubMed Central **MS-SSM** (parallel SSMs tuned to different bands), **DH-LIF** (dendritic branches with heterogeneous timing factors), or **MSDCGTNet** (multi-scale kernels implicitly capturing frequency bands). Nature Phase relationship modeling requires **PACNet** for explicit PAC Frontiers or attention mechanisms in transformers for implicit cross-frequency coupling.

For Spatial Relationships and Hierarchical Aggregation: Implement **Graph WaveNet** (adaptive adjacency learning functional connectivity), PubMed **ASTGCN** (spatial attention with multi-component temporal hierarchy), University of Northampton **HGNet** (logarithmic message-passing paths through graph hierarchy), **GGN** (dynamic connectivity graph generation), or **Non-local Networks** (all-to-all spatial interactions). Hierarchical aggregation achieved through **multi-level graph structures** (HGNet), **community detection** for coarse-graining, or **spatial attention mechanisms** (ASTGCN, DAMGCN).

For Nonlocality: Use **Non-local Neural Networks** (explicit global context computation with $O(N^2)$ all-to-all connections), **Transformer self-attention** in ETST/CTNet (query-key-value attention over all positions), **Graph WaveNet** (learned adaptive dependencies beyond fixed graph), or **DCRNN** with high K-step diffusion (propagates information across many hops capturing distant dependencies). Attention mechanisms provide dynamic nonlocal weighting while graph-based approaches enable structured nonlocal propagation.

Recommended Architecture Combinations

High-Performance Research System: Combine **MS-SSM** (parallel multi-frequency SSMs) for temporal processing with **ASTGCN** (multi-component spatial-temporal graph attention) for spatial modeling and hierarchical temporal scales. Add **Non-local blocks** at middle layers for explicit long-range dependencies. Expected benefits: comprehensive multi-scale coverage, explicit frequency band modeling, dynamic spatial attention, multiple temporal periodicities, and state-of-the-art accuracy with interpretability through attention weights and frequency decomposition.

Production/Real-Time System: Use **HiSS** (hierarchical SSM) as backbone for efficient two-level temporal hierarchy with **Graph WaveNet** for spatial modeling (adaptive adjacency without preprocessing). Implement in **Mamba-2** framework for 2-8× faster training and inference. Expected benefits: linear time complexity O(N), constant memory for generation, learned spatial dependencies, causal processing, and neuromorphic-compatible design enabling edge deployment.

Maximum Interpretability System: Deploy **Wav-KAN** (wavelet neural network) with explicit frequency decomposition into physiological bands, **DAMGCN** (graph attention with learnable spatial and frequency weights), and **dual attention mechanisms** showing spatial and spectral importance. Add **DH-LIF** neurons if spiking implementation desired for biological plausibility. Expected benefits: interpretable wavelet coefficients, attention weight visualization, frequency band contributions, learned electrode importance, and biologically-grounded architecture.

Large-Scale/Long-Duration System: Implement **MTST** (multi-resolution transformer) with different patch sizes for fast events and slow trends, **DCRNN with graph partitioning** for spatial scalability to thousands of electrodes, and **Temporal Gaussian Hierarchy** for efficient long-sequence processing with constant memory. Add **SNAC-style hierarchical tokenization** for compression. Expected benefits: O(log N) scaling through hierarchy, handles hours of continuous recording, efficient compression, multi-resolution temporal modeling, and parallel processing capability.

Implementation Roadmap

Phase 1 - Foundation (Weeks 1-2): Implement **EEGNet** baseline for comparison, set up **Graph WaveNet** for spatial processing with learned adjacency, and integrate **Mamba** for temporal modeling with selective SSM. Establish data pipeline with frequency band decomposition (STFT/wavelet), spatial adjacency matrix construction (Euclidean distance + learned adaptive), and preprocessing (bandpass filtering, artifact removal, normalization).

Phase 2 - Multi-Scale Temporal (Weeks 3-4): Add **HiSS-style hierarchical stacking** with 2-3 temporal levels, implement **MS-SSM parallel branches** (5 branches for 5 EEG frequency bands), integrate **dilated temporal convolutions** with exponentially growing receptive fields, and add **temporal attention** for adaptive temporal weighting. Configure timing parameters per branch (α_d values: 0.9 for delta/theta, 0.7 for alpha/beta, 0.4 for gamma).

Phase 3 - Spatial Hierarchy (Weeks 5-6): Implement **multi-level graph structure** using community detection (Louvain) for brain region hierarchy, add **spatial attention mechanisms** (ASTGCN-style with learnable electrode weights), integrate **Non-local blocks** at middle layers for global context, and implement **graph pooling** for coarse-to-fine spatial processing. Create 3-level hierarchy: electrodes → regions → hemispheres → global.

Phase 4 - Multi-Component Integration (Weeks 7-8): Add **multi-component architecture** (Recent/Daily/Weekly as in ASTGCN or Fast/Slow as in SlowFast), implement **lateral connections** for information flow between components, add **hierarchical feature fusion** with learned weights, and integrate **phase-amplitude coupling** layers for cross-frequency interactions. Weight initialization: equal for all components, learned during training.

Phase 5 - Optimization (Weeks 9-10): Integrate **Mamba-2 optimizations** for 2-8× training speedup, implement **mixed precision training** (FP16/BF16), add **gradient checkpointing** for memory efficiency, optimize **graph operations** using PyTorch Geometric sparse tensors, and implement **efficient attention** (FlashAttention-2 for transformer components). Profile and optimize bottlenecks.

Phase 6 - Validation (Weeks 11-12): Test on **standard benchmarks** (BCI Competition IV-2a, DEAP, SEED, Temple University Hospital), perform **ablation studies** (remove each component to measure contribution), validate **frequency band contributions** (analyze learned attention weights), assess **spatial pattern interpretability** (visualize electrode importance), and benchmark **computational efficiency** (FLOPs, memory, latency).

Code Resources and Starting Points

Core SSM Implementations: Mamba/Mamba-2 at https://github.com/state-spaces/mamba (official implementation with CUDA kernels, pretrained models 130M-2.8B parameters), HiSS at https://hiss-csp.github.io/ (hierarchical SSM with CSP-Bench dataset), BlackMamba at https://github.com/Zyphra/BlackMamba (MoE + Mamba hybrid).

Graph Neural Network Libraries: PyTorch Geometric for graph operations (GCN, GAT, graph pooling layers), Graph WaveNet at https://github.com/nnzhan/Graph-WaveNet (adaptive adjacency + dilated convolutions), ASTGCN at https://github.com/guoshnBJTU/ASTGCN-2019-pytorch (multi-component attention-based STGCN), DCRNN at https://github.com/chnsh/DCRNN_PyTorch (diffusion convolution + encoder-decoder), HGNet at https://github.com/rampasek/HGNet (hierarchical GNN with O(log N) paths), GGN at https://github.com/ICLab4DL/GGN (dynamic graph generation for EEG).

Multi-Scale Architectures: MTST at https://github.com/networkslab/MTST (multi-resolution transformer), TimesNet at https://github.com/thuml/TimesNet (temporal 2D-variation with multi-periodicity), Wav-KAN at https://github.com/zavareh1/Wav-KAN (wavelet neural networks), SNAC at https://github.com/hubertsiuzdak/snac (multi-scale neural audio codec), SlowFast at https://github.com/facebookresearch/SlowFast (dual-pathway video networks), DH-SNN at https://github.com/eva1801/DH-SNN (dendritic heterogeneity spiking networks).

EEG-Specific Tools: EEGModels at https://github.com/vlawhern/arl-eegmodels (EEGNet, ShallowConvNet, DeepConvNet in Keras/TensorFlow), EEG-DL at https://github.com/SuperBruceJia/EEG-DL (comprehensive EEG deep learning library with CNN, RNN, LSTM, GCN, Transformer), Braindecode at https://braindecode.org (PyTorch-based EEG analysis with preprocessing pipelines), Non-local Networks at https://github.com/facebookresearch/video-nonlocal-net (spacetime nonlocal operations).

Technical Configuration Guidelines

Frequency Band Configuration (5 parallel branches for EEG): Delta branch (0.5-4 Hz, α_d=0.95, slow SSM, 4-8 second receptive field), Theta branch (4-8 Hz, α_d=0.90, 2-4 second receptive field), Alpha branch (8-13 Hz, α_d=0.80, 1-2 second receptive field), Beta branch (13-30 Hz, α_d=0.65, 0.5-1 second receptive field), Gamma branch (30-100 Hz, α_d=0.40, 100-500ms receptive field). Use learnable parameters initialized at these values, adapted during training.

Spatial Graph Configuration: Construct multi-level hierarchy with Level 0 (64 individual electrodes, 10-20 system), Level 1 (8-12 brain regions via Louvain clustering based on functional connectivity), Level 2 (2 hemispheres + subcortical), Level 3 (global brain state). Initialize adjacency with Gaussian kernel on Euclidean distance: $W_{ij} = \exp(-d_{ij}^2/\sigma^2)$ where σ determined by average inter-electrode distance. Add learnable adaptive adjacency matrix as in Graph WaveNet. Use message-passing across levels: bottom-up aggregation and top-down refinement.

Temporal Configuration: Implement three-component system following ASTGCN: Recent Component (processes last 30-60 seconds for immediate context, high temporal resolution, captures transients and events), Periodic Component (samples same time windows from previous cycles, daily rhythm for experiments spanning days, ultradian rhythm 90-120 min for sleep studies), Slow Component (downsampled 4-8× for long-term trends, hours-long context, captures state changes). Fusion weights learned through training with balanced initialization (0.33 each).

Attention Mechanisms: Multi-head attention with 8-16 heads for diversity, key/query/value dimensions d_k = 64-128 (trade-off between capacity and efficiency), dropout 0.1-0.2 on attention weights for regularization, learned positional encodings (sinusoidal for long sequences). Apply spatial attention over electrode dimension, temporal attention over time dimension, and cross-attention between frequency bands for phase coupling.

Training Hyperparameters: Learning rate 1e-3 to 1e-4 with cosine annealing schedule, batch size 16-64 depending on sequence length and memory, gradient clipping at norm 1.0 to prevent instability, weight decay 1e-4 for regularization, mixed precision training (FP16/BF16) for 2× speedup. Use warmup for first 5-10% of training, curriculum learning starting with shorter sequences then increasing length, scheduled sampling for sequence-to-sequence tasks transitioning from teacher forcing to autoregressive.

Computational Considerations: For **64 electrodes, 1000 Hz sampling, 60-second windows**: Mamba backbone processes 64,000 timesteps in O(64,000) = O(N) time with constant memory per token, Graph convolution on 64-node graph with K=3 hops requires O(E·K) ≈ O(64·10·3) operations using sparse matrices, Multi-head attention over electrodes: O(64^2·d_k) = O(4096·64) acceptable for spatial dimension, Temporal attention avoided or use efficient variants (Linformer, Performer) for O(N) complexity. Use gradient checkpointing to trade computation for memory, enabling 4-8× longer sequences. Expected memory: 4-8 GB for training, 1-2 GB for inference with mixed precision.

Performance Benchmarks and Validation

Language Modeling Comparisons: Mamba-3B matches Transformer-7B in performance, Mamba-2 achieves 2-8× faster training than Mamba-1, Jamba handles 256K context on single 80GB GPU (4GB KV cache vs 32GB for pure attention), BlackMamba shows 2.2× fewer training steps than standard Mamba for equivalent performance.

Vision and Sensor Tasks: Vision Mamba (Vim) provides 2.8× faster processing with 86.8% less memory than DeiT at 1248^2 resolution, HiSS achieves 23% better MSE than Transformers/LSTMs/S4/Mamba on sensor datasets, FlowState ranks #2 on GIFT-Eval leaderboard with only 9.1M parameters (smallest in top 10).

EEG-Specific Benchmarks: DAMGCN: 99.42% accuracy SEED, 97.50% DEAP (emotion recognition), MSDCGTNet: 99.66% DEAP, 98.85% SEED, 99.67% SEED-IV (emotion recognition), GGN: 91% accuracy classifying 7 seizure types (vs 82% Transformer, 74% GNN, 65% CNN), CTNet: State-of-the-art on BCI Competition IV-2a motor imagery (~89% with attention mechanisms), DH-LIF SNNs: 92.1% Spiking Heidelberg Digits (previous best 90.4%), 82.46% Spiking Speech Commands (previous best 74.2%), with 36-45% fewer parameters and superior noise robustness.

Spatial-Temporal Graph Networks: Graph WaveNet: Superior performance on METR-LA and PEMS-BAY traffic forecasting without pre-defined graphs, ASTGCN: Outperforms baselines on PeMSD4 (307 sensors) and PeMSD8 (170 sensors), DCRNN: Effective on sensor networks up to 11,160 sensors with graph partitioning, STGCN: 14× faster training than RNN-based methods with fewer parameters (455K vs 9.4M for FC-LSTM).

Conclusion: Integrated Architecture for EEG Multi-Timescale Processing

The optimal architecture for EEG-like signals combines: **Multi-Scale SSMs** (MS-SSM or HiSS) providing parallel frequency band processing with learned timescale parameters (α_d = 0.4-0.95 range), **Graph Neural Networks** (Graph WaveNet or ASTGCN) with adaptive spatial dependencies, attention-weighted electrode importance, and multi-level hierarchy (electrodes → regions → hemispheres), **Temporal Components** (Recent/Periodic/Slow) capturing multiple temporal scales from milliseconds to hours, and **Nonlocal Mechanisms** (Non-local blocks or Transformer attention) at middle layers for global context and cross-frequency coupling.

This integrated design preserves causal relationships through selective SSMs and dilated causal convolutions, models multiple frequency bands simultaneously via parallel SSM branches with band-specific timing factors and wavelet decomposition, accounts for spatial relationships through graph convolution with learned adjacency and hierarchical aggregation, and handles nonlocality via attention mechanisms and adaptive graph dependencies enabling distant electrode interactions.

Implementation path: Start with Mamba-2 backbone (efficient SSM with linear complexity), add Graph WaveNet spatial processing (learned adjacency), integrate HiSS hierarchical structure (2-3 temporal levels), incorporate ASTGCN multi-component system (Recent/Periodic/Slow), and enhance with attention mechanisms for interpretability and nonlocal modeling. Full implementations available at referenced GitHub repositories with production-ready code enabling rapid prototyping and deployment on neuromorphic hardware for real-time EEG analysis.