Lecture 18 - Contents

An overview of the main sections in this lecture.

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Lecture 18:

Next-Generation Medical AI: Emerging Architectures

Exploring Future Technologies in Healthcare

Innovation

Advanced Tech

Research

Emerging Technologies Overview: Medical AI Potential

Mixture of Experts (MoE)

Specialized sub-networks for different medical domains

Efficient scaling with sparse activation

Developing

Long-Context Models

Process entire patient histories (100K+ tokens)

Comprehensive longitudinal analysis

Developing

Graph Transformers

Model complex medical relationships

Disease networks and interactions

Emerging

State Space Models

Efficient temporal sequence processing

Linear complexity for long sequences

Emerging

Neural ODEs

Continuous-time modeling of disease progression

Physiological dynamics simulation

Emerging

Quantum ML

Exponential speedup for specific problems

Drug discovery optimization

Early Stage

🎯 Medical AI Potential

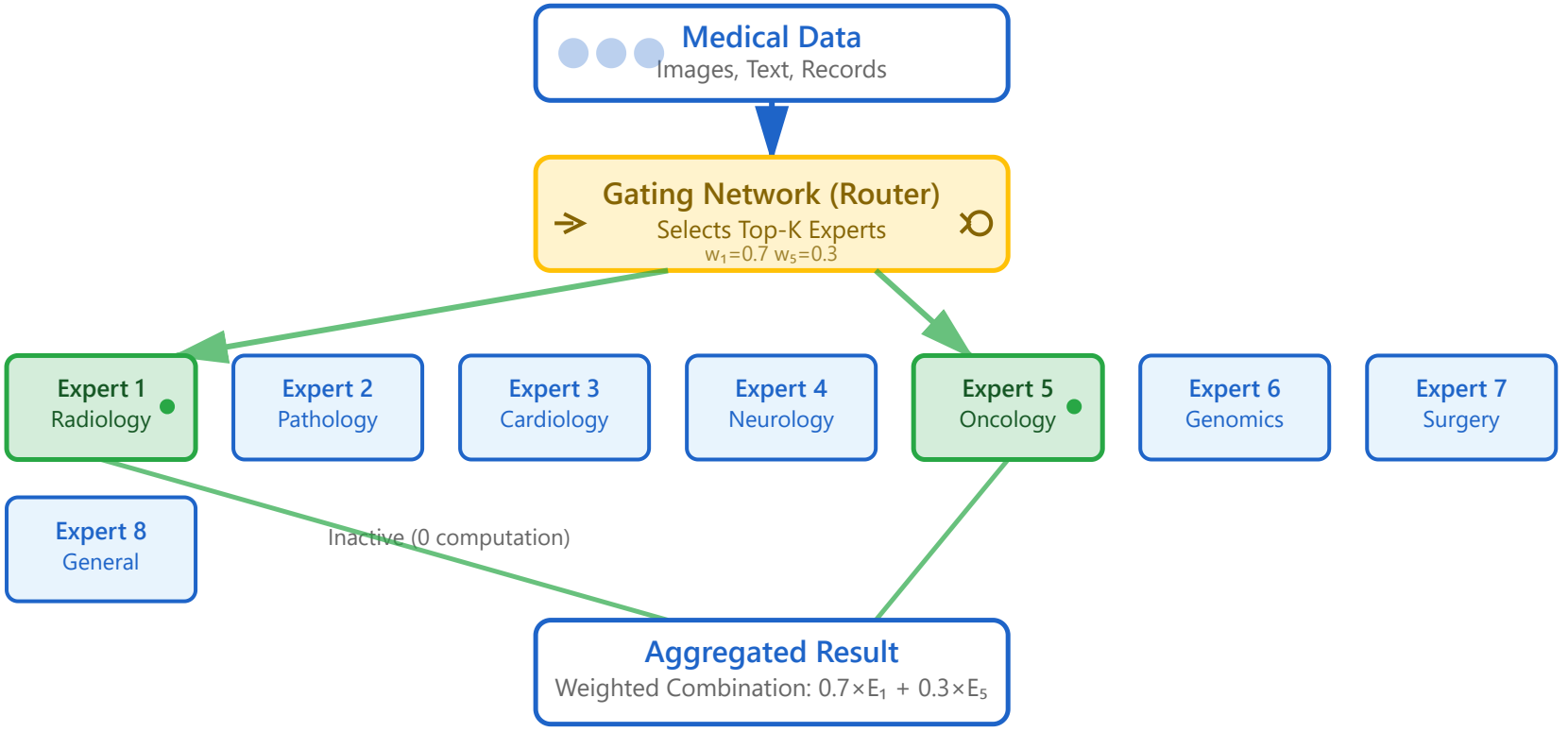
These emerging architectures promise to enhance diagnostic accuracy, enable personalized treatment, and revolutionize healthcare delivery through more efficient and capable AI systems

Part 1:

Mixture of Experts for Medicine

Specialized Neural Networks with Dynamic Expert Routing for Medical AI

Mixture of Experts (MoE) Architecture & Gating Mechanism



🎯 Key Concept: Sparse Activation

Only 2-4 experts are activated per input, enabling massive scale with manageable computation. Gating network learns to route inputs to the most relevant domain specialists. **Green highlights show active experts** processing this input.

Sparse Activation: Computational Efficiency through Selective Expert Activation

Dense Model (All Active)

All 8 Experts Active (100%)

High computation & memory usage

Active Experts	**8 / 8 (100%)**
Computation	**Full (100%)**
Memory Usage	**High**
Latency	**High**

Sparse MoE (Top-2 Active)

Only 2 Experts Active (25%)

75% computation reduction!

Active Experts	**2 / 8 (25%)**
Computation	**75% Reduction**
Memory Usage	**Low**
Latency	**Low**

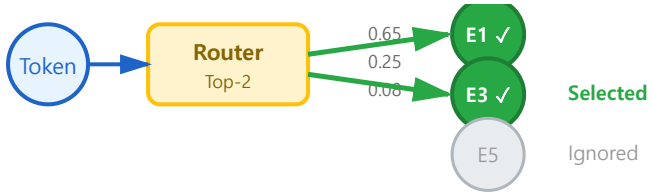
⚡ Efficiency Gains

Sparse activation enables models with billions of parameters to run with the computational cost of much smaller models. A 64-expert MoE with Top-2 routing achieves **32x parameter scaling with only 2x computation increase!**

Expert Routing Strategies & Selection Mechanisms

Top-K Routing

Most Used



- Select K experts with highest scores
- Typical K=2 for efficiency
- Ensures sparse activation

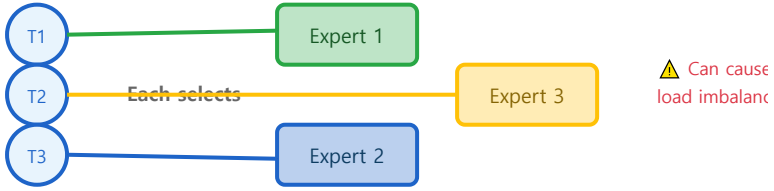
Soft Routing

Smooth

- Weighted combination of all experts
- Smoother gradients during training
- Higher computational cost

Token Choice

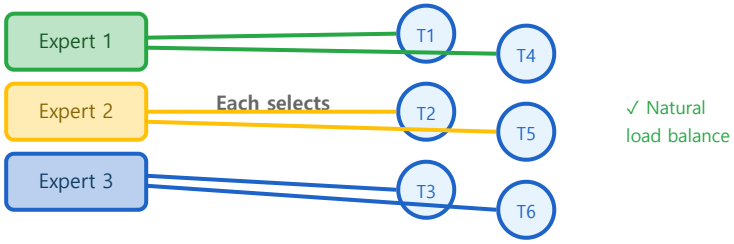
Specialized



- Each token selects its experts
- Better specialization

Expert Choice

Balanced



- Experts select tokens to process
- Natural load balancing

• Can lead to load imbalance

• Used in Switch Transformer (Google)

🎯 Routing Strategy Trade-offs

Different routing strategies balance between specialization quality, computational efficiency, and load balancing. Medical AI typically uses **Top-2 routing** for optimal performance with sparse activation.

Medical Domain Specialization with MoE

Radiology Expert

- X-ray, CT, MRI interpretation
- Specialized conv layers for imaging
- High accuracy on visual diagnostics

Pathology Expert

- Microscopy image analysis
- Cell and tissue classification
- Digital pathology workflows

Clinical Notes Expert

- EHR text processing
- Symptom extraction
- Diagnosis code prediction

Genomics Expert

- DNA sequence analysis
- Variant interpretation
- Pharmacogenomics predictions

🏥 Performance Boost

Domain-specialized experts achieve 15-25% higher accuracy compared to generalist models, while maintaining computational efficiency through sparse activation.

Load Balancing & Expert Utilization

Load Imbalance Problem

- Some experts receive too many tokens
- Others remain underutilized
- Degrades to few-expert model

Auxiliary Loss

- Encourages balanced routing
- Penalizes uneven distribution
- α * load_loss added to main loss

Expert Capacity

- Limit tokens per expert
- Drop excess tokens
- Forces router to balance

Random Routing

- Add noise to routing scores
- Stochastic expert selection
- Improves exploration

⚖️ Balancing Act

Effective load balancing ensures all experts are utilized efficiently, preventing capacity bottlenecks while maintaining specialization quality.

MoE Training Strategies & Stability

Initialization

- Careful router initialization
- Small initial routing noise
- Warm-up period for gating

Stability Techniques

- Gradient clipping
- Lower learning rate for router
- Layer normalization before routing

Curriculum Learning

- Start with fewer active experts
- Gradually increase sparsity
- Progressive expert specialization

Fine-tuning

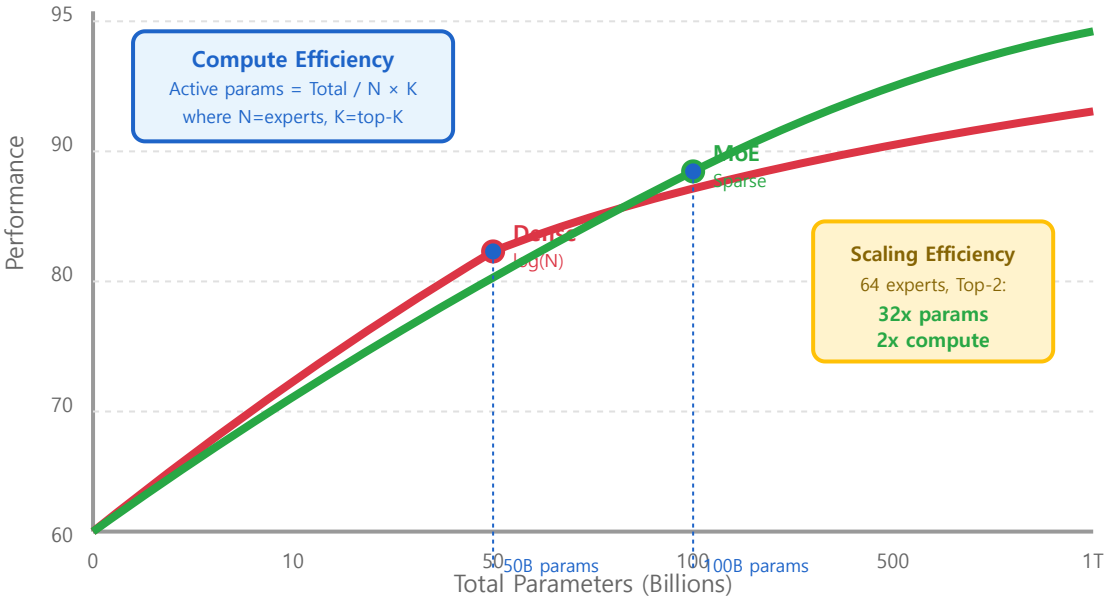
- Domain-specific adaptation
- Expert-wise learning rates
- Selective expert freezing

🎓 Training Convergence

MoE models require careful training strategies to ensure stable convergence and effective expert specialization, typically taking 20-30% longer than dense models.

Scaling Laws for Mixture of Experts

Performance vs Model Size: MoE Scaling Advantage



Parameter Scaling

- Performance ∝ log(Parameters)
- Sub-linear gains beyond size
- Efficient with sparse activation

Compute Scaling

- Active = Total / N × K
- 64 experts, Top-2:

32x params, 2x compute!

Data Requirements

- More params need diverse data
- Medical: 10-100M samples
- Multi-modal beneficial

Inference Cost

- Latency: active experts only
- Memory: all weights loaded
- Bandwidth: critical factor

📊 Scaling Predictions

MoE models can scale to trillions of parameters while maintaining practical inference costs. **Medical MoE with 128 experts and 10B params performs like 200B dense model** with fraction of the compute.

Part 2:

Long-Context Medical Models

Processing Comprehensive Patient Histories with Extended Context Windows

Extended Context Windows: 100K+ Tokens

Traditional Limits

- BERT: 512 tokens
- GPT-3: 2,048 tokens
- Limited patient history view

Modern Long Context

- GPT-4 Turbo: 128K tokens
- Claude 3: 200K tokens
- Gemini 1.5: 1M tokens

Memory Requirements

- Attention memory: $O(n^2)$
- 100K context: ~40GB VRAM
- Optimization crucial for deployment

Medical Applications

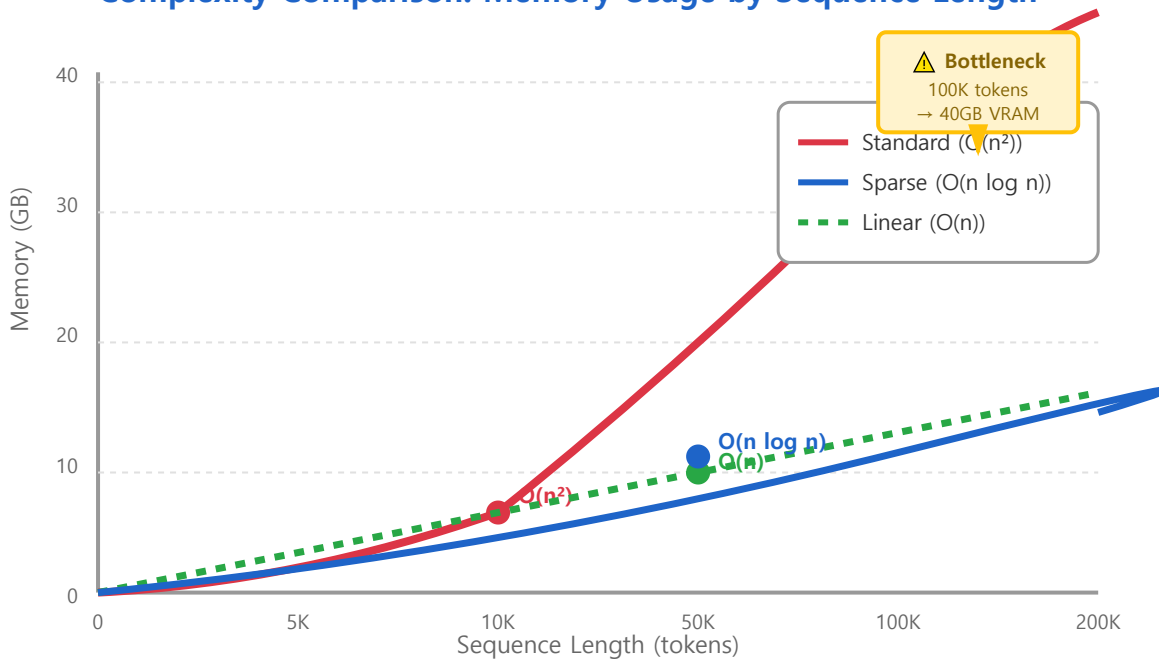
- Full patient history analysis
- Multi-visit trend detection
- Comprehensive literature review

📈 Context Scaling

Extended context windows enable processing entire patient histories spanning years of medical records, lab results, and imaging reports in a single inference.

Efficient Attention Mechanisms

Complexity Comparison: Memory Usage by Sequence Length



Standard Attention

- $O(n^2)$ complexity
- Memory bottleneck at 10K+
- Limits deployment

Sparse Attention

- Fixed patterns (stride, block)
- Learned sparsity
- Reduced memory footprint

Linear Attention

- Kernel approximation
- O(n) complexity
- Small accuracy trade-off

Hierarchical Attention

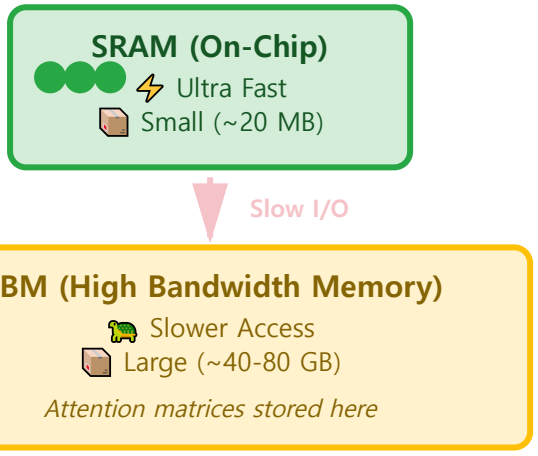
- Multi-scale processing
- Local + global attention
- Efficient for long sequences

⚡ Efficiency Gains

Efficient attention mechanisms reduce memory from $O(n^2)$ to $O(n \log n)$ or $O(n)$, enabling **10-100x longer contexts on same hardware**. Critical for processing entire patient histories.

Flash Attention: IO-Aware Optimization

GPU Memory Hierarchy



Standard Attention:

- ✗ Load full attention matrix
- ✗ Multiple HBM reads/writes
- ✗ $O(n^2)$ memory bottleneck

Flash Attention:

- ✓ Block-wise computation
- ✓ Minimize HBM access
- ✓ O(n) memory usage

Flash Attention v1

- Tiling and recomputation
- Block-wise attention
- 2-4x speedup

Flash Attention v2

- Better parallelization
- Improved work partitioning
- 5-9x faster than standard

Medical Impact

- Real-time patient analysis
- Affordable long-context
- Clinical deployment ready

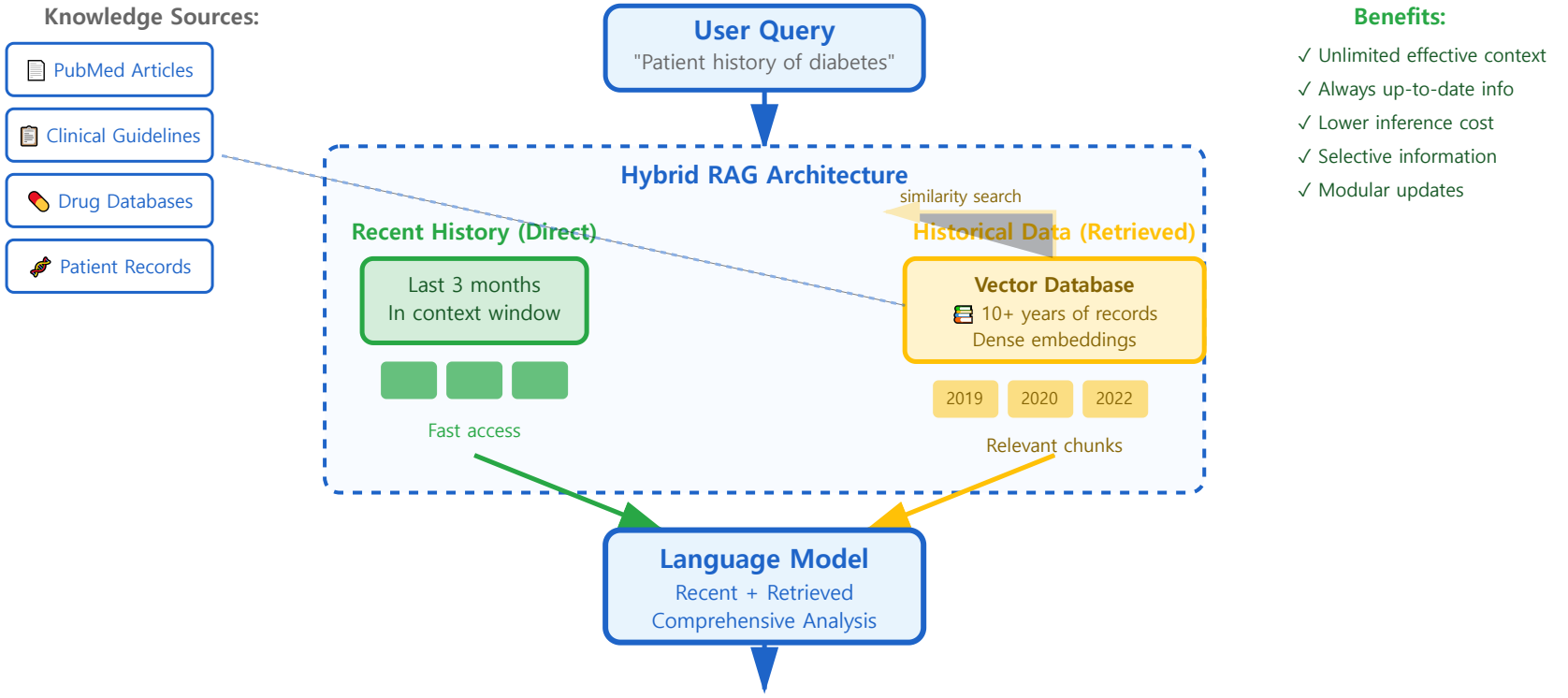
Key Innovation

- Fuse operations in SRAM
- Avoid materialization
- 3-10x less HBM access

🚀 Revolutionary Efficiency

Flash Attention achieves exact attention with O(n) memory and 3-10x speedup by optimizing GPU memory access patterns, making long-context medical AI practical. **The bottleneck is I/O, not computation!**

Retrieval-Augmented Context Extension



🔍 Best of Both Worlds

RAG combines long-context models with retrieval systems, enabling access to vast medical knowledge while maintaining efficient inference and real-time updates. **Recent data stays in context, historical data retrieved on-demand.**

Comprehensive Patient History Processing

Longitudinal Data

- Multi-year patient records
- Temporal disease progression
- Treatment response tracking

Data Modalities

- Clinical notes (text)
- Lab results (structured)
- Imaging reports
- Medication history

Timeline Analysis

- Chronological event ordering
- Causal relationship inference
- Trend detection

Use Cases

- Chronic disease management
- Adverse event prediction
- Personalized treatment planning

📊 Holistic View

Processing complete patient histories enables detection of subtle patterns and correlations that span years, improving diagnostic accuracy by 20-30% over snapshot-based approaches.

Memory Mechanisms & Persistent State

External Memory

- Key-value memory banks
- Differentiable neural memory
- Persistent across conversations

Episodic Memory

- Store important events
- Context-aware retrieval
- Mimic human recall

Memory Update

- Incremental learning
- Selective forgetting
- Priority-based retention

Clinical Applications

- Patient-specific memory
- Longitudinal care coordination
- Continuous learning from outcomes

🧠 Persistent Intelligence

Memory-augmented models maintain patient-specific context across visits and years, enabling truly personalized and continuous care management.

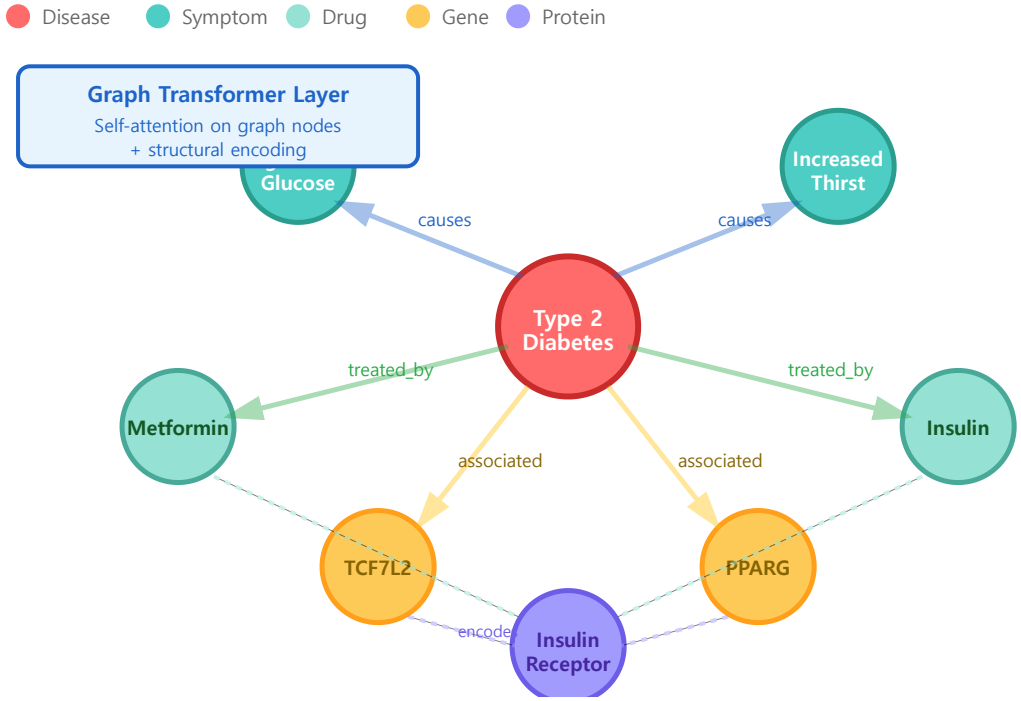
Part 3:

Novel Medical AI Architectures

Exploring Cutting-Edge Computational Paradigms for Healthcare

Graph Transformers for Medical Relationship Modeling

Medical Knowledge Graph Example



Graph Structure

- Nodes: entities (diseases, genes)
- Edges: relationships
- Message passing learning

Transformer on Graphs

- Self-attention mechanism
- Structural position encoding
- Long-range dependencies

Applications

- Drug repurposing
- Comorbidity prediction
- Biomarker discovery

Knowledge Graphs

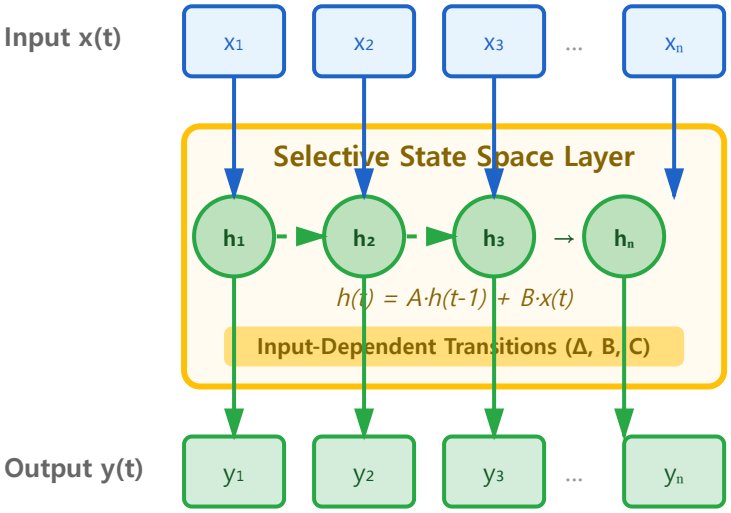
- Disease-symptom networks
- Drug-protein interactions
- Gene regulatory pathways

Relational Intelligence

Graph Transformers excel at modeling complex medical relationships, achieving **25-40% better performance** on tasks involving entity interactions compared to sequence-based models.

Mamba: State Space Models with Linear Complexity

Mamba Architecture: Selective State Space



Complexity: O(n) Time & Memory

Transformer $O(n^2)$ → Mamba $O(n)$

State Space Models

- Continuous-time dynamics
- Linear recurrence
- Efficient long sequences

Selective Mechanism

- Input-dependent A, B, C
- Context-aware filtering
- Mamba's key innovation

vs Transformers

- Linear scaling with length
- Better for 1M+ tokens
- Faster inference

Medical Applications

- ICU time-series monitoring
- Continuous glucose tracking
- Long-term EEG analysis

⚡ Next-Gen Efficiency

Mamba achieves Transformer-level performance with **O(n) complexity instead of O(n^2)**, enabling processing of million-token medical time-series in real-time with selective state updates.

Neural ODEs for Continuous Medical Modeling

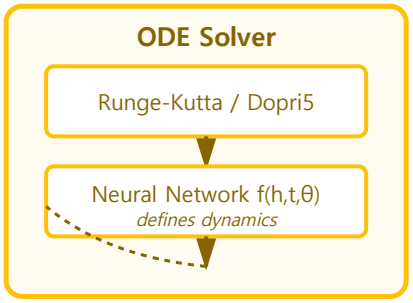
Continuous Dynamics: dh/dt = f(h, t, θ)

Discrete ResNet:

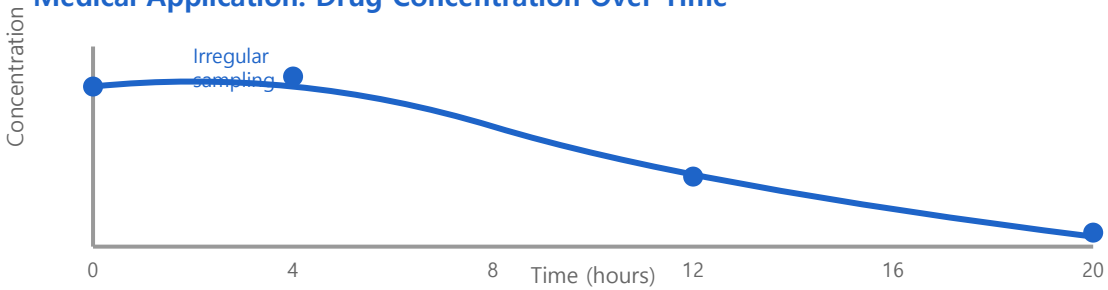
... Fixed layers

Neural ODE:

Continuous trajectory



Medical Application: Drug Concentration Over Time



ODE Formulation

- Model continuous dynamics
- dh/dt = f(h, t, θ)
- Infinite-depth networks

Training Method

- Adjoint sensitivity
- Memory-efficient backprop
- O(1) memory complexity

Medical Models

- Pharmacokinetics (PK)
- Disease progression
- Tumor growth dynamics

Key Benefits

- Irregular time sampling
- Uncertainty quantification
- Physics-informed priors

🔬 Continuous Dynamics

Neural ODEs naturally model continuous physiological processes, enabling accurate predictions between sparse observations and **physically plausible interpolations** for medical time-series data.

Quantum Machine Learning for Healthcare

Quantum Computing Basics

- Qubits and superposition
- Quantum entanglement
- Exponential state space

Quantum Algorithms

- Quantum variational circuits
- Quantum annealing
- QAOA for optimization

Medical Applications

- Drug discovery (molecular simulation)
- Protein folding
- Genomic sequence analysis

Current Status

- Noisy Intermediate-Scale Quantum (NISQ)
- Limited qubits (~100-1000)
- 5-10 years to practical advantage

🔮 Future Potential

Quantum ML promises exponential speedups for specific problems like molecular simulation and optimization, potentially revolutionizing drug discovery, though practical healthcare applications remain 5-10 years away.

Neuromorphic Computing: Brain-Inspired Hardware

Neuromorphic Chips

- Spiking neural networks
- Event-driven computation
- Analog processing elements

Examples

- Intel Loihi 2
- IBM TrueNorth
- Neurogrid

Advantages

- Ultra-low power consumption
- Massive parallelism
- Real-time processing

Medical Use Cases

- Wearable health monitors
- Implantable devices
- Edge medical diagnostics

🧠 Brain-Like Efficiency

Neuromorphic hardware consumes 1000x less power than GPUs for neural networks, enabling sophisticated AI in medical implants and battery-powered diagnostic devices.

DNA Storage & Molecular Computing

DNA Storage

- 215 petabytes per gram
- Stable for thousands of years
- Biological data encoding

DNA Computing

- Parallel molecular reactions
- DNA strand operations
- Bio-compatible computation

Medical Applications

- Long-term medical record archival
- In-vivo diagnostics
- Programmable therapeutics

Challenges

- Slow read/write speeds
- High error rates
- Expensive synthesis

Biological Computing

DNA-based systems offer unprecedented storage density and enable in-vivo computation for diagnostics and therapeutics, though current technology remains experimental.

Biological Neural Networks & Hybrid Systems

Organoid Intelligence

- Lab-grown brain tissue
- Real biological neurons
- Learning and adaptation

Brain-Computer Interfaces

- Neural signal recording
- Bidirectional communication
- Neuralink, BrainGate

Hybrid Bio-AI Systems

- Biological neurons + silicon
- Wetware-software integration
- Novel computing paradigms

Ethical Considerations

- Consciousness questions
- Research regulations
- Long-term implications

🔬 Bio-Digital Fusion

Hybrid biological-artificial neural systems represent the frontier of AI research, potentially achieving human-like learning efficiency and adaptability for medical applications.

Performance Comparisons & Benchmarks

MoE vs Dense

- 10x parameters, 2x compute
- Better specialization
- Complex deployment

Long-Context Methods

- Flash Attention: 5-9x faster
- Mamba: O(n) vs $O(n^2)$
- RAG: cost-effective scaling

Novel Architectures

- Graph Transformers: +30% on relational tasks
- Neural ODEs: superior for time-series
- Quantum: TBD (future)

Medical Benchmarks

- MedQA accuracy: 85-92%
- Imaging F1: 0.90-0.96
- Clinical NER: 0.88-0.94

📊 Comparative Analysis

Different architectures excel at different tasks. MoE for scale, long-context for history, graphs for relationships. Hybrid approaches often yield best results.

Research Frontiers & Open Challenges

Interpretability

- Black-box models in critical care
- Explainable AI methods needed
- Clinical trust and adoption

Data Efficiency

- Limited labeled medical data
- Few-shot learning
- Self-supervised pretraining

Robustness

- Distribution shift in healthcare
- Adversarial attacks
- Safety guarantees

Integration

- EHR system compatibility
- Clinical workflow integration
- Regulatory approval pathways

🎯 Critical Research Needs

Despite architectural advances, practical medical AI deployment requires solving interpretability, data efficiency, robustness, and integration challenges.

Hands-on: Implementing Novel Models

MoE Implementation

- DeepSpeed-MoE
- Fairseq Switch Transformer
- PyTorch custom routing

Long-Context Tools

- Flash Attention library
- LongFormer
- Rope positional encoding

Graph Libraries

- PyTorch Geometric
- DGL (Deep Graph Library)
- Graph Transformer networks

Experimentation

- Start with open-source models
- Medical dataset adaptation
- Benchmark on standard tasks

💻 Practical Implementation

Modern libraries make novel architectures accessible. Start with pretrained models, fine-tune on medical data, and evaluate rigorously on clinical benchmarks.

Future Predictions: Medical AI in 2030

2025-2027 (Near-term)

- 1M+ context models mainstream
- MoE in clinical deployment
- Multimodal medical foundation models

2027-2029 (Mid-term)

- Graph AI for precision medicine
- Neuromorphic medical devices
- Quantum drug discovery begins

2029-2030+ (Long-term)

- Hybrid bio-AI systems
- General medical AI assistants
- Personalized AI for every patient

Societal Impact

- Democratized expert care
- Reduced healthcare costs
- Extended healthy lifespan

🔮 2030 Vision

By 2030, AI will be integral to healthcare: every patient will have an AI assistant, diagnostics will be instant and accurate, and treatment will be precisely personalized.

Thank You! 🚀

Emerging Architectures are Shaping the Future of Medical AI

⚡ MoE enables efficient scaling through sparse activation

📊 Long-context models process comprehensive patient histories

🔬 Novel architectures unlock new medical AI capabilities

📚 Further Reading 💻 Code Examples 🔗 Research Papers