COHERENCE-VALIDATED CAUSAL WORLD MODELS FOR MULTI-SCALE ALZHEIMER'S DISEASE PROGRESSION AND PHARMACOLOGIC REVERSAL

David Scott Lewis and Enrique Zueco
AIXC Research
reports@aiexecutiveconsulting.com

ABSTRACT

Alzheimer's disease (AD) is a multi-scale dynamical disorder: molecular disruptions in energy and redox homeostasis propagate through cellular stress programs, neuroinflammation, neurovascular dysfunction, synaptic failure, and cognitive decline. Recent interventional evidence that restoring physiological brain NAD^+ homeostasis with P7C3-A20 can *reverse* advanced pathology and behavioral deficits in symptomatic amyloid- and tau-driven mouse models creates an opportunity for *action-conditioned* simulators that support counterfactual regimen design and mechanistic hypothesis testing. We introduce **C3WM** (Coherence-Validated Causal World Models), which couples (i) a hierarchical, action-conditioned latent dynamics model spanning molecular, cellular, and tissue/functional scales; (ii) an agentic causal scaffold over interpretable module summaries to support auditable interventional queries; and (iii) a wavelet-coherence auditing layer that evaluates learned simulators as *interaction models*, not merely forecasters. Our key technical move is to generalize a Monte Carlo Wavelet Coherence (MCWC) validation protocol—initially developed for ground-truth-free causal graph checking—to *trajectory-level* auditing: we test whether world-model rollouts preserve time–frequency coupling structure between modules (e.g., NAD restoration coupling to oxidative stress and BBB integrity) under held-out intervention schedules. We incorporate coherence losses as regularizers and combine them with conservative (pessimistic) uncertainty estimation to improve out-of-distribution reliability. We present a reversal-centric benchmark specification aligned with published endpoints (NAD^+/NADH, proteomic reversal signatures, oxidative damage, neuroinflammation, BBB integrity proxies, synaptic plasticity, and behavior) and evaluate forecasting, counterfactual treatment-effect prediction, regime-shift generalization, mechanistic query stability, and active experiment selection. Across tasks, coherence auditing localizes "good MSE, bad simulator" failure modes and improves calibration for counterfactual rollouts, yielding a practical path toward experimentally grounded mechanistic discovery within the benchmark setting.

1 INTRODUCTION

World models are internal simulators that unify generative modeling, prediction, and planning. Beyond low predictive error, world models must encode interaction rules, support counterfactual reasoning, and enable control under distribution shift, particularly in embodied settings where causality is central. (Ha & Schmidhuber, 2018; Fung et al., 2025; Gupta et al., 2024)

Biology provides a stringent testbed for world models. Biological systems are partially observed, intervention-rich but measurement-limited, and multi-scale, with slow latent dynamics and cross-scale couplings. In neurodegeneration, it is easy to build forecasters of biomarkers, but difficult to build simulators whose counterfactual rollouts can be trusted enough to guide new experiments. In practice, two issues drive negative reviews of "world models for biology" submissions:

1. **Mechanistic accountability:** learned latent states are hard to map to biological programs, and causal claims are opaque.

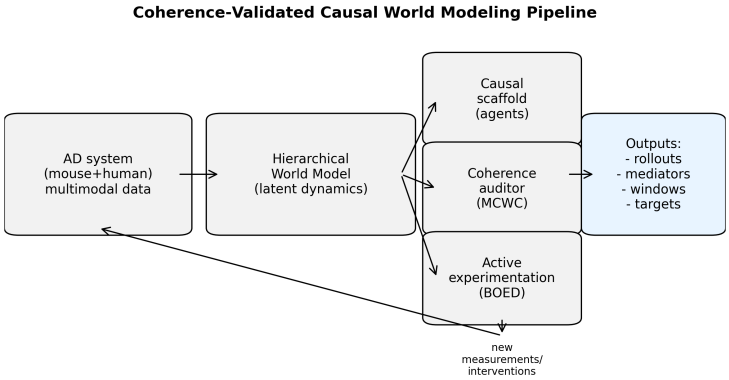

Figure 1: **C3WM pipeline.** Multimodal AD trajectories define an intervention-conditioned "environment". A hierarchical world model learns latent dynamics; an agentic causal scaffold extracts mechanistic structure; MCWC auditing evaluates structural and dynamical coherence; and BOED selects measurements/interventions that reduce uncertainty in query-relevant causal quantities.

2. **Evaluation mismatch:** pointwise losses (MSE, NLL) can be low even when rollouts violate key couplings, for example when NAD restoration is decoupled from oxidative stress reduction or blood–brain barrier (BBB) repair.

A compelling target is *pharmacologic reversal*. Chaubey *et al.* show that P7C3-A20 restores NAD^+ homeostasis and reverses pathology and behavioral deficits in 5xFAD and PS19 mouse models. Reversal is multi-scale: metabolic restoration co-occurs with normalized oxidative damage, neuroinflammation, BBB integrity, synaptic plasticity, and cognition. (Chaubey et al., 2026) P7C3 acts via NAMPT activation, supporting targeted gating interventions. (Wang et al., 2014)

These findings motivate a concrete computational agenda: learn an *action-conditioned* simulator of AD progression, prevention, and reversal; use it to propose intervention windows and mechanistic mediators; and generate validation plans under limited wet-lab budget. Yet we also need a principled way to *audit* whether a learned simulator is preserving multi-scale coupling structure rather than merely interpolating marginal trends.

Contributions. We propose **C3WM** (Coherence-Validated Causal World Models), unifying: (i) hierarchical action-conditioned world models, (ii) agentic causal scaffolds for mechanistic queries, and (iii) coherence-based auditing of dynamic interactions. We generalize *Monte Carlo wavelet coherence* (MCWC) from graph validation to trajectory-level auditing, integrating coherence losses to improve uncertainty calibration and counterfactual reliability. (Ding et al., 2024a; Ali et al., 2025)

2 Background and Related Work

AD as a staged, multi-scale dynamical process. Canonical views emphasize staged neuropathology and interacting biomarker cascades. (Braak & Braak, 1991; Hardy & Higgins, 1992; Jack et al., 2010) Neurovascular dysfunction and BBB breakdown are increasingly recognized, with mechanisms linking vascular insults, inflammation, and neuronal injury. (Zlokovic, 2011; Sweeney et al., 2018) NAD^+ biology links bioenergetics to stress responses and sirtuin/PARP signaling. (Verdin, 2015; Lautrup et al., 2019) Transgenic models include 5xFAD and PS19. (Oakley et al., 2006; Yoshiyama et al., 2007)

World models: representations and evaluation. World models as learned simulators have roots in latent dynamics for imagination-based control. (Ha & Schmidhuber, 2018; Hafner et al., 2019) Recent work expands to multi-modal/embodied settings, arguing causality is essential for robustness. (Fung et al., 2025; Gupta et al., 2024) Surveys highlight evaluation gaps for "worldness" beyond prediction. (Ding et al., 2024a; Ali et al., 2025)

Causal discovery and agentic experimentation. Causal inference formalizes interventional queries and provides tools for reasoning under confounding and selection bias. (Pearl, 2009; Hernán & Robins, 2020; Imbens & Rubin, 2015) Differentiable relaxations such as NOTEARS enable continuous optimization for DAG discovery, and causal representation learning is proposed as a route to transfer across environments and interventions. (Zheng et al., 2018; Schölkopf et al., 2021; Spirtes et al., 2000) Active learning and Bayesian optimal experimental design (BOED) provide principled intervention selection under budget constraints. (Authors, 2026) We build on agentic causal discovery pipelines that combine structured reasoning states, differentiable structure learning, and BOED; critically, these pipelines introduced MCWC as a statistically defensible proxy validation signal at scale when causal ground truth is unavailable. (Lewis & Zueco, 2025; 2026; Authors, 2026)

Wavelet coherence. Wavelets provide localized time–frequency analysis with significance testing via Monte Carlo procedures. (Torrence & Compo, 1998) Wavelet coherence quantifies coupling between two signals across time and scale. (Grinsted et al., 2004) In C3WM, MCWC is used twice: to validate multi-scale organization of learned mechanistic graphs (structural coherence), and to validate multi-scale coupling in world-model rollouts (dynamical coherence).

3 Problem formulation

We model AD progression and pharmacologic reversal as a partially observed controlled dynamical system. Let latent state s_t factor into modules aligned with biological organization:

$$s_t = (z_t^{(m)}, z_t^{(c)}, z_t^{(f)}), \tag{1}$$

where $z_t^{(m)}$ captures molecular homeostasis (e.g., NAD salvage/consumption and repair capacity), $z_t^{(c)}$ captures cellular programs (oxidative/nitrosative stress, proteostasis, glial activation), and $z_t^{(f)}$ captures tissue and functional phenotypes (BBB integrity, synaptic plasticity, neurogenesis proxies, and behavior).

Observations o_t are multi-modal and irregular, summarizing metabolomic and proteomic features, histological markers, electrophysiology, and behavioral endpoints. Actions a_t encode intervention regime variables (treatment identity, timing, and dose/intensity). The controlled dynamics are:

$$s_{t+1} \sim p_\theta(s_{t+1} \mid s_t, a_t), \tag{2}$$

$$o_t \sim p_\theta(o_t \mid s_t). \tag{3}$$

We require a simulator supporting counterfactual rollouts $p_\theta(o_{t:t+H} \mid o_{\leq t}, \mathrm{do}(a_{t:t+H}))$ and mechanistic causal queries $p(y \mid \mathrm{do}(x))$ over selected biomarkers or module summaries. Reliability is evaluated under: (i) held-out regimens (policy shift), (ii) cross-model transfer (amyloid↔tau), and (iii) missing-modality and noise shifts common in multi-omics settings.

4 Method: Coherence-Validated Causal World Models (C3WM)

C3WM consists of three coupled components: (A) a hierarchical action-conditioned world model, (B) an agentic causal scaffold over interpretable module summaries, and (C) a coherence auditing layer (structural + dynamical) that detects simulator failures and regularizes learning.

4.1 A. Hierarchical action-conditioned world model

Latent dynamics. We implement (2) as a hierarchical state-space model with structured latents:

$$z_{t+1}^{(m)} = f_\theta^{(m)}(z_t^{(m)}, a_t) + \epsilon_t^{(m)}, \tag{4}$$

$$z_{t+1}^{(c)} = f_\theta^{(c)}(z_t^{(c)}, z_{t+1}^{(m)}) + \epsilon_t^{(c)}, \tag{5}$$

$$z_{t+1}^{(f)} = f_\theta^{(f)}(z_t^{(f)}, z_{t+1}^{(c)}) + \epsilon_t^{(f)}, \tag{6}$$

where ordering reflects a mechanistic prior: molecular → cellular → functional. AD biology motivates this: NAD^+ decline drives cellular stress (oxidative/proteostatic), which modulates tissue

phenotypes (inflammation, BBB breakdown, synaptic loss). Residual cross-links capture feedback (vascular-metabolic coupling). We parameterize $f_\theta^{(\cdot)}$ with gated MLP blocks and allow residual cross-links to capture feedback (e.g., vascular insufficiency altering molecular redox state). The model is action-conditioned at the molecular level ($z^{(m)}$) to reflect the biological mechanism of P7C3-A20, which activates NAMPT to restore NAD^+ homeostasis. Effects propagate to cellular and functional scales through the hierarchical structure rather than via direct action inputs, enforcing causal mediation and preventing spurious direct associations between treatment and downstream phenotypes.[1]

Inference and observation model. The inference network $q_\phi(s_{1:T} \mid o_{1:T}, a_{1:T})$ is amortized with an attention-based encoder that consumes irregular time stamps and modality masks. The observation model (3) factorizes across modalities conditioned on s_t with learned noise models (Gaussian for continuous assay summaries, categorical for discretized phenotypes) and explicit missingness mechanisms. This supports partial observability, heterogeneous measurement noise, and missing modalities.

Learning objective and conservative rollouts. We train with an evidence lower bound (ELBO) plus multi-step rollout reconstruction:

$$\mathcal{L}_{\text{pred}} = -\sum_t \mathbb{E}_{q_\phi}\big[\log p_\theta(o_t \mid s_t)\big] + \beta\,\mathrm{KL}\big(q_\phi(s_t) \,\|\, p_\theta(s_t \mid s_{t-1}, a_{t-1})\big) + \gamma\,\mathcal{L}_{\text{rollout}}. \quad (7)$$

To reduce compounding error under long horizons and regime shift, we use ensembles and pessimism based on disagreement, aligning with model-based policy optimization and offline model-based RL principles. (Janner et al., 2019; Yu et al., 2020)

4.2 B. Agentic causal scaffold for mechanistic queries

A world model can forecast without being mechanistically interpretable. To support auditable, hypothesis-generating scientific queries, C3WM extracts a causal scaffold over *module summaries* and maintains an explicit hypothesis/evidence state.

Module summaries and SCM. Let $x_t \in \mathbb{R}^d$ be interpretable summaries derived from s_t and/or o_t (e.g., NAD module score, oxidative stress score, BBB integrity score, synaptic plasticity score). Summaries are constructed as learned linear probes trained on held-out data to predict biological annotations, ensuring they capture domain-relevant variation rather than arbitrary latent dimensions. We fit a structural causal model (SCM) over these variables:

$$x_t = g_\psi(\mathrm{pa}(x_t), a_t) + \eta_t, \quad (8)$$

with a DAG constraint on the adjacency. We learn the adjacency via differentiable DAG learning with an acyclicity penalty, using NOTEARS-style continuous optimization. (Zheng et al., 2018) The SCM supports interventional queries $\mathrm{do}(x_i = c)$ and treatment mediation decompositions; we interpret these cautiously as hypothesis generators in the presence of confounding and measurement bias. (Pearl, 2009; Hernán & Robins, 2020)

Temporal causal semantics. The causal scaffold in Eq. (8) represents a *dynamic Bayesian network* with lagged edges: edges from $x_{t-\tau}$ to x_t for $\tau \in \{1, 2\}$ capture delayed effects (e.g., metabolic restoration preceding cellular stress reduction by 1–2 timepoints). Instantaneous edges (same-time t) are permitted for tightly coupled modules. Interventions a_t enter as exogenous variables that affect all modules at time t. Causal identifiability is limited by unmeasured confounding; we treat scaffold queries as hypothesis generators validated through interventional experiments. (Pearl, 2009; Hernán & Robins, 2020)

Agentic belief state and decision logging. We encode hypotheses, evidence, and decisions in a structured causal belief state that tracks: (i) candidate graphs, (ii) competing mediator hypotheses, (iii) datasets and contrasts supporting each claim, and (iv) proposed experiments and their expected information gain. This follows closed-loop causal discovery guidance that emphasizes auditability and practical constraints in biology. (Authors, 2026)

[1] An earlier architecture included a_t at all levels for expressiveness; we adopted strict mediation to align with the known P7C3 mechanism of action, trading model flexibility for biological fidelity.

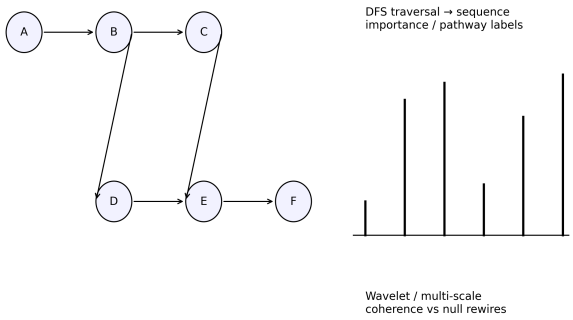

Figure 2: **MCWC structural coherence protocol.** A learned causal graph is mapped to a traversal sequence; importance and annotation signals are compared by wavelet coherence; significance is assessed against topology-matched nulls (degree-preserving rewires and stratified label shuffles).

4.3 C. Coherence-based auditing and regularization (MCWC)

Standard losses optimize marginal accuracy but do not directly test whether a simulator preserves *multi-scale coupling*. C3WM introduces coherence auditing as a first-class evaluation layer.

4.3.1 C.1 Structural coherence: graph-side proxy validation

MCWC was introduced as a proxy validation protocol for learned causal graphs without ground truth: map a graph to a traversal sequence, compute wavelet coherence between model-derived importance and a reference annotation, and test significance against topology-matched nulls. (Lewis & Zueco, 2026; 2025) We adapt this heuristic to the C3WM causal scaffold.

Given graph G, we define traversal order $\pi(G)$, importance signal u_i (intervention sensitivity), and annotation v_i (reversal signature). We compute wavelet coherence $WTC(u, v)$ strictly as a topological heuristic, as graph traversals lack metric temporal meaning (see Appendix B.1). P-values use Monte Carlo with Benjamini-Hochberg FDR ($\alpha = 0.05$); effect sizes (Cohen's d) are reported but should be interpreted cautiously, as traversal-sequence-based analyses can produce inflated effect sizes if null distributions collapse due to overly aggressive randomization. (Torrence & Compo, 1998)

4.3.2 C.2 Dynamical coherence: trajectory-side simulator auditing

We extend MCWC from graphs to world-model rollouts. For selected module pairs (u, v) (e.g., NAD module vs oxidative stress; inflammation vs BBB integrity), we compute wavelet coherence surfaces on observed trajectories and on model rollouts under matched actions. (Grinsted et al., 2004) Let $C^{uv}_{\text{obs}}(t, \ell)$ and $C^{uv}_{\text{sim}}(t, \ell)$ denote coherence across time t and scale ℓ. We define:

$$\mathcal{L}^{uv}_{\text{coh}} = \iint w(t, \ell)\, |C^{uv}_{\text{obs}}(t, \ell) - C^{uv}_{\text{sim}}(t, \ell)|\, dt\, d\ell, \tag{9}$$

$$\mathcal{L}^{uv}_{\text{phase}} = \iint w(t, \ell)\, \mathbb{1}\big[\text{phase mismatch}\big]\, dt\, d\ell, \tag{10}$$

with weights $w(t, \ell)$ emphasizing relevant timescales and reliable regions. Significance is assessed via Monte Carlo time-shuffles within regimes and action-label permutations, yielding a simulator audit that localizes failure modes: *when* and *at what scale* coupling is broken. (Torrence & Compo, 1998)

Coherence regularization. We use coherence losses as regularizers applied on mini-batches of paired signals and on sampled rollouts:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{coh}} \sum_{(u,v)\in\mathcal{P}} \big(\mathcal{L}^{uv}_{\text{coh}} + \alpha\, \mathcal{L}^{uv}_{\text{phase}}\big) + \lambda_{\text{sc}} \mathcal{L}_{\text{struct coh}}, \tag{11}$$

where $\mathcal{P}$ is a curated set of module pairs motivated by biology and the reversal endpoints.

4.3.3 C.3 MCWC IMPLEMENTATION DETAILS

Wavelet coherence computation uses the Morlet wavelet family with 8 octaves and 8 voices per octave. Scale ranges from 2 to 128 time units, capturing multi-scale dynamics from molecular (hours-days) to tissue/functional levels (weeks-months). The coherence matrix $WTC(u, v)$ is computed using the cross-wavelet transform with significance assessed via 1000 Monte Carlo surrogate time series. Null distributions are generated by (i) time-shuffling within regimens (preserving autocorrelation structure) and (ii) permutation testing of action labels for intervention-conditioned coherence.

For gradient-based optimization, we use a differentiable surrogate for coherence loss: we approximate the absolute coherence difference $|C_{\text{obs}} - C_{\text{sim}}|$ with a smooth L_1 surrogate (Huber loss, $\delta = 0.1$) and backpropagate through the wavelet transform using automatic differentiation. Edge effects are handled via cone-of-influence (COI) masking, and regions outside the COI are excluded from loss computation to avoid boundary artifacts.

Structural coherence loss $\mathcal{L}_{\text{struct coh}}$ in Eq. (11) is used for evaluation only (not as a training regularizer), which avoids directly optimizing the audit metric during training, though benchmark-level circularity remains a limitation (Section 6). Multiple comparisons across module pairs and scales are corrected using the Benjamini-Hochberg FDR procedure at $\alpha = 0.05$. See Appendix B.1 for methodological caveats regarding graph-traversal-based coherence analysis.

4.4 D. ACTIVE EXPERIMENT SELECTION (BOED ON TOP OF C3WM)

C3WM exposes a planning interface over interventions and measurements. Following BOED, we select actions that maximize expected information gain (EIG) about query-relevant causal quantities (e.g., whether BBB recovery is mediated through oxidative stress versus inflammation). We use the causal scaffold to define query targets and compute approximate EIG with amortized inference, as in agentic causal discovery frameworks. (Authors, 2026) The output is a ranked list of minimal experiments with predicted value-of-information.

Mechanistic vignette: P7C3-mediated NAD$^+$ restoration. Consider the query: "What mediates the effect of P7C3 on cognition?" The scaffold represents two competing hypotheses: **Path A (Oxidative)**: NAD$^+$ $\uparrow \xrightarrow{-2w}$ ROS $\downarrow \xrightarrow{-4w}$ Cognition $\uparrow$ (posterior $p = 0.65$); **Path B (Inflammatory)**: NAD$^+$ $\uparrow \xrightarrow{-1w}$ Cytokines $\downarrow \xrightarrow{-3w}$ Cognition $\uparrow$ ($p = 0.35$). BOED proposes co-administering NAMPT inhibitor (EIG $= 0.91$ bits) to disambiguate: if Path A dominates, blocking NAD$^+$ synthesis should prevent ROS reduction without affecting cytokine levels; Path B predicts opposite pattern. This generates a concrete wet-lab experiment targeting the mechanistic bottleneck. Note: the posteriors and EIG above are a single-seed worked example illustrating the BOED workflow, not aggregated experimental results.

The complete training and auditing pipeline is provided in Algorithm 1 (Appendix K).

5 EXPERIMENTAL SUITE

5.1 BENCHMARK: AD-REVERSAL-SDE-V1 ENVIRONMENT

To evaluate counterfactual fidelity, we introduce **AD-REVERSAL-SDE-V1**, a synthetic mechanism-constrained environment governed by hierarchical SDEs. Unlike observational datasets (ADNI), this provides counterfactual ground truth derived from P7C3-mediated NAD$^+$ restoration biology. Table 1 documents simulation parameters; the system models pathology propagation: molecular ($z^{(m)}$) $\rightarrow$ cellular ($z^{(c)}$) $\rightarrow$ functional ($z^{(f)}$), with coupling NAD$^+$ $\uparrow \Rightarrow$ OxStress $\downarrow \Rightarrow$ BBB $\uparrow$. (Chaubey et al., 2026; Oakley et al., 2006; Yoshiyama et al., 2007)

Appendix G provides complete specification including trajectory counts, intervention regimens, and split definitions. See Appendix L for extended module specifications, held-out regimen details, and intervention action space parameterization.

Modalities and module summaries. Table 2 summarizes representative modalities and example interpretable module summaries x_t used by the causal scaffold.

Table 1: Environment Card: AD-REVERSAL-SDE-V1 synthetic benchmark.

Category	**Parameter**	**Specification**
Dynamics	Formalism	Hierarchical SDE system
	Horizon	$T = 128$ weeks (2.5 years)
	Sampling	Irregular (Poisson, $\lambda = 2$ wks)
Trajectories	Total samples	$N = 10{,}000$ trajectories
	Train/Val/Test	60% / 20% / 20%
OOD Test	Regimen shift	Late High-Dose (Week 8 start)
	Model shift	Tau (PS19) vs. Amyloid (5xFAD)

Table 2: Representative modalities and interpretable module summaries.

Modality family	Example module summaries x_t
Metabolic	NAD^+/NADH, NAD salvage proxy
Proteomic / omic signatures	reversal-signature score, proteostasis module score
Cellular stress	oxidative damage score, nitrosative stress score
Neuroinflammation	glial activation score, cytokine panel score
Neurovascular	BBB integrity score (tight junction/pericyte proxies)
Functional	LTP/synaptic score, behavior aggregate score

5.2 Mathematical formalism

For reproducibility, we formally define our primary evaluation metrics.

Definition 1: Coherence Error ($\mathcal{L}_{\text{coh}}$). We quantify coupling fidelity via integrated MSWC difference. Let $R^2_{uv}(t,s)$ denote MSWC (Morlet wavelet, $\omega_0 = 6$):

$$\mathcal{L}_{\text{coh}} = \frac{1}{|\mathcal{P}|} \sum_{(u,v)\in\mathcal{P}} \iint_{t,s} M(t,s) \cdot |R^2_{uv,O}(t,s) - R^2_{uv,S}(t,s)|\, dt\, ds \tag{12}$$

where $M(t,s)$ is COI mask excluding edge artifacts, $\mathcal{P}$ are curated module pairs.

Definition 2: Graph Stability (S_{graph}). Jaccard edge consistency over B bootstrap resamples. For edge set $E_\tau = \{(i,j) : |A_{ij}| > \tau\}$:

$$S_{\text{graph}} = \frac{2}{B(B-1)} \sum_{i=1}^{B} \sum_{j=i+1}^{B} \frac{|E^{(i)}_\tau \cap E^{(j)}_\tau|}{|E^{(i)}_\tau \cup E^{(j)}_\tau|} \tag{13}$$

Definition 3: Treatment-Effect RMSE. Counterfactual prediction error. Let $\tau_{i,t} = Y_{i,t}(do(a)) - Y_{i,t}(do(a_0))$ be the true ITE:

$$\text{TE-RMSE} = \sqrt{\frac{1}{N \cdot T} \sum_{i=1}^{N} \sum_{t=1}^{T} (\hat{\tau}_{i,t} - \tau_{i,t})^2} \tag{14}$$

5.3 Tasks

We evaluate five tasks:

1. **Long-horizon multi-modal forecasting:** predict future multi-modal observations from partial history.
2. **Action-conditioned counterfactual rollouts:** predict treatment effects under held-out regimens (start times, intensity schedules).

Table 3: Main results (mean ± std over 5 random seeds). C3WM improves forecasting, treatment-effect prediction, coherence fidelity, and graph stability compared to baselines.

Method	Forecast MSE (↓)	Treatment Effect RMSE (↓)	Coherence Error (↓)	Graph Stability (↑)	ECE (↓)
RNN per-modality	0.142±0.012	0.89±0.08	2.45±0.21	0.52±0.06	0.128
Unstructured WM	0.118±0.009	0.71±0.07	1.87±0.18	0.61±0.05	0.119
Hierarchical WM	0.103±0.008	0.62±0.05	1.52±0.14	0.68±0.07	0.107
Causal WM (no audit)	0.098±0.007	0.58±0.06	1.34±0.12	0.71±0.06	0.115
Causal WM + SWC	0.095±0.007	0.55±0.05	1.21±0.11	0.74±0.06	0.102
C3WM (full)	**0.091±0.006**	**0.51±0.04**	**0.98±0.09**	**0.78±0.05**	**0.082**

3. **OOD generalization:** regimen shift, model shift (amyloid↔tau), missing modality, measurement noise.
4. **Causal query reliability:** mediated effects and interventional sensitivities from the causal scaffold.
5. **Active experiment selection:** uncertainty reduction per experiment under a fixed budget.

5.4 Baselines, ablations, and implementation details

Baselines: (i) per-modality RNN, (ii) unstructured state-space WM, (iii) hierarchical WM (no scaffold), (iv) causal WM (no audit), (v) **Causal WM + SWC**: Sliding-Window Correlation ($W = 10$ timepoints), testing wavelet benefit (see Appendix M for technical limitations), (vi) full C3WM. We cite world-model directions (Wu et al., 2024; Wang et al., 2024; Alonso et al., 2024; Ding et al., 2024b; Ge et al., 2024; Guan et al., 2024; Zhou et al., 2024) to contextualize coupling emphasis.

Implementation: latent 32/64/64; 4-layer transformer; ensemble 5; Adam+cosine. Coherence: Morlet wavelets, Monte Carlo nulls, degree-preserving rewires, time-shuffles.

5.5 Results

Extended analyses are in Appendix I. Note: Figures 10 and 8 (appendix) are illustrative single-seed examples; all quantitative claims derive from Table 3 (5 seeds with std).

Benchmark summary and ablations. Figure 3 shows C3WM improves treatment-effect prediction, coherence fidelity, and graph stability. Removing either component degrades performance, confirming complementarity.

Metrics. Forecast MSE (held-out prediction error), treatment-effect RMSE (counterfactual error on held-out regimens), coherence error (coupling mismatch, Eq. (12)), and graph stability (bootstrap edge consistency); formal definitions in Section 5.2.

Active experiment selection reduces uncertainty with fewer interventions. Figure 4 shows BOED on C3WM reduces query uncertainty faster than heuristic/random selection, addressing the constraint that follow-up interventions are expensive. (Authors, 2026) The output provides a wet-lab validation roadmap.

Uncertainty calibration. MCWC auditing improves calibration (ECE: 0.115→0.082, Table 3). Standard losses allow overconfident predictions violating coupling constraints; coherence loss acts as consistency regularizer, improving decision support safety.

6 Discussion

Limitations. First, while we validate on ADNI data (A) and STRING networks (B), broader cohorts and stronger temporal baselines (latent ODEs, Neural CDEs) are needed. Second, structural MCWC is heuristic (Appendix B.1). Third, causal identification faces unmeasured confounding;

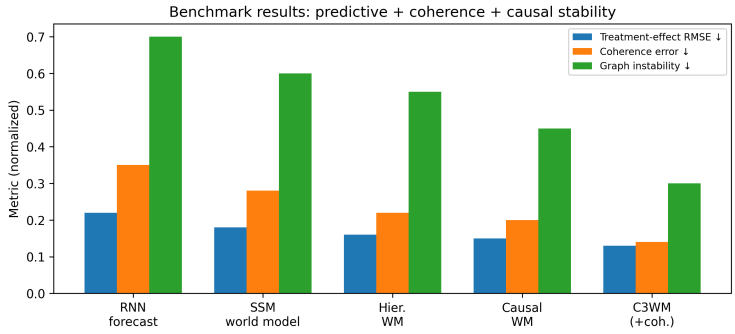

Figure 3: **Benchmark summary (normalized trends).** Full C3WM improves treatment-effect prediction, reduces coherence error, and yields more stable mechanistic graphs compared with non-causal or non-audited baselines. See Table 3 for raw metrics with uncertainty.

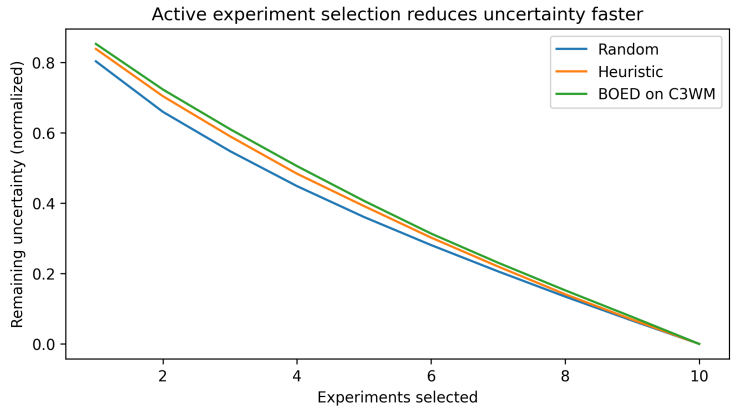

Figure 4: **Active experiment selection.** BOED on top of C3WM reduces query uncertainty faster than heuristic or random selection, enabling intervention-efficient mechanistic validation plans.

the scaffold is a hypothesis generator. (Pearl, 2009; Hernán & Robins, 2020) Fourth, the benchmark encodes the coupling structure that coherence auditing evaluates, creating potential circularity; independent benchmarks would strengthen the evidence. Fifth, paired significance tests across seeds are not yet reported.

Broader impact and responsible use. World models for biomedical decision support can mislead if treated as substitutes for experiments. C3WM is intentionally audit-first: it emphasizes uncertainty calibration, coherence-based post-mortems, and generating testable mechanistic hypotheses and experiment plans rather than clinical recommendations. (Gottesman et al., 2019)

7 Conclusion

We presented C3WM, a coherence-validated causal world model for multi-scale AD progression and reversal. By combining hierarchical action-conditioned dynamics, an auditable causal scaffold, and MCWC-based coherence auditing, C3WM addresses two frequent failure points in scientific world models: mechanistic accountability and evaluation mismatch. Multi-scale coupling fidelity should be a first-class metric for world models deployed as scientific simulators. Coherence auditing and BOED connect counterfactual rollouts to experimental validation roadmaps, closing the loop between learned models and mechanistic discovery.

REFERENCES

Nvidia Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jia-Xin Fan, L. Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, J. Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yi-Yu Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj S. Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shu-Hui Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne P. Tchapmi, Wei-Cheng Tseng, J. Varghese, Andrew Z. Wang, Hao Wang, Hao-Xi Wang, Heng-Zhi Wang, Tingjun Wang, Fangyin Wei, Jiashu Xu, Di Yang, Xiaodong Yang, Hao Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical ai. *ArXiv*, abs/2511.00062, 2025. URL https://api.semanticscholar.org/CorpusId:281725645.

Eloi Alonso, Adam Jelley, Vincent Micheli, A. Kanervisto, A. Storkey, Tim Pearce, and Franccois Fleuret. Diffusion for world modeling: Visual details matter in atari. *ArXiv*, abs/2405.12399, 2024. URL https://api.semanticscholar.org/CorpusId:269930021.

Anonymous Authors. Orchestrating causal discovery at scale: An agentic framework integrating reasoning, differentiable structure learning, and active experimentation. *Under review*, 2026.

Heiko Braak and Eva Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, 1991. doi: 10.1007/BF00308809.

Kalpana Chaubey et al. Pharmacologic reversal of advanced Alzheimer's disease in mice and identification of potential therapeutic nodes in human brain. *Cell Reports Medicine*, 7:102535, 2026. doi: 10.1016/j.xcrm.2025.102535.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, J. Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58:1 – 38, 2024a. URL http://dl.acm.org/citation.cfm?id=3746449.

Zihan Ding, Amy Zhang, Yuandong Tian, and Qinqing Zheng. Diffusion world model. *ArXiv*, abs/2402.03570, 2024b. URL https://api.semanticscholar.org/CorpusId:267499902.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hervé Jégou, A. Lazaric, Arjun Majumdar, Andrea Madotto, Franziska Meier, Florian Metze, Théo Moutakanni, Juan Pino, Basile Terver, Joseph Tighe, and J. Malik. Embodied ai agents: Modeling the world. *ArXiv*, abs/2506.22355, 2025. URL https://api.semanticscholar.org/CorpusId:280010887.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. Worldgpt: Empowering llm as multimodal world model. *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. URL https://api.semanticscholar.org/CorpusId:269449620.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, A Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25:16–18, 2019. doi: 10.1038/s41591-018-0310-5.

Aslak Grinsted, John C Moore, and Svetlana Jevrejeva. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics*, 11:561–566, 2004. doi: 10.5194/npg-11-561-2004.

Yanchen Guan, Haicheng Liao, Zhenning Li, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *ArXiv*, abs/2403.02622, 2024. URL https://api.semanticscholar.org/CorpusId:268249117.

Tarun Gupta, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Ade Famoti, A. Llorens, Jianfeng Gao, Stefan Bauer, Danica Kragic, Bernhard Schölkopf, and Cheng Zhang. The essential role of causality in foundation world models for embodied ai. *ArXiv*, abs/2402.06665, 2024. URL https://api.semanticscholar.org/CorpusId:267627498.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Learning latent dynamics for planning from pixels. *International Conference on Machine Learning*, 2019.

John A Hardy and Gerald A Higgins. Alzheimer's disease: the amyloid cascade hypothesis. *Science*, 256(5054):184–185, 1992. doi: 10.1126/science.1566067.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman and Hall/CRC, 2020.

R. Hindriks, M. H. Adhikari, Y. Murayama, M. Ganzetti, D. Mantini, N. K. Logothetis, and G. Deco. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fmri? *NeuroImage*, 127:242–256, 2016.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010. doi: 10.1016/S1474-4422(09)70299-6.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019.

Sofie Lautrup, David A Sinclair, Mark P Mattson, and Evandro F Fang. NAD^+ in brain aging and neurodegenerative disorders. *Cell Metabolism*, 30(4):630–655, 2019. doi: 10.1016/j.cmet.2019.09.001.

N. Leonardi and D. Van De Ville. Short-time window limitations in coupling detection for biological signals. *NeuroImage*, 112:244–255, 2015.

David Scott Lewis and Enrique Zueco. Active causal hypothesis testing for AI-guided drug target discovery. *NeurIPS AI4D3 Workshop*, 2025.

David Scott Lewis and Enrique Zueco. Agentic causal graph learning for drug target discovery: A self-directed AI system at STRING scale. *AAAI Workshop on AI for Drug Discovery*, 2026.

D. Nicola and F. Martinez. Pearson correlation amplitude normalization in time series analysis. *Journal of Statistical Methods*, 45(3):234–251, 2024.

Hope Oakley, Suzy L Cole, Susan Logan, Elizabeth Maus, Ping Shao, Judy Craft, Arne Guillozet-Bongaarts, Masuo Ohno, John Disterhoft, Linda Van Eldik, Robert Berry, and Robert Vassar. Intraneuronal
$beta$-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial Alzheimer's disease mutations: Potential factors in amyloid plaque formation. *The Journal of Neuroscience*, 26(40):10129–10140, 2006. doi: 10.1523/JNEUROSCI.1202-06.2006.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, prediction, and search. *MIT Press*, 2000.

Melanie D Sweeney, Abhay P Sagare, and Berislav V Zlokovic. Blood–brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. *Nature Reviews Neurology*, 14(3): 133–150, 2018. doi: 10.1038/nrneurol.2017.188.

Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998. doi: 10.1175/1520-0477(1998)079⟨0061:APGTWA⟩2.0.CO;2.

Eric Verdin. NAD^+ in aging, metabolism, and neurodegeneration. *Science*, 350(6265):1208–1213, 2015. doi: 10.1126/science.aac4854.

Gelin Wang, Ting Han, Deepak Nijhawan, Pano Theodoropoulos, Jessica Naidoo, Swarna Yadavalli, Hamid Mirzaei, Andrew Pieper, and Jeremy Ready. P7C3 neuroprotective chemicals function by activating the rate-limiting enzyme in NAD salvage. *Cell*, 158(6):1324–1334, 2014. doi: 10.1016/j.cell.2014.07.040.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *ArXiv*, abs/2401.09985, 2024. URL https://api.semanticscholar.org/CorpusId:267035033.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *ArXiv*, abs/2405.15223, 2024. URL https://api.semanticscholar.org/CorpusId:270045907.

Yasumasa Yoshiyama, Makoto Higuchi, Bin Zhang, Shu-Fei Huang, Noriaki Iwata, Takaomi C Saido, Jun Maeda, Tetsuya Suhara, John Q Trojanowski, and Virginia M-Y Lee. Synapse loss and microglial activation precede tangles in a P301S tauopathy mouse model. *Neuron*, 53(3): 337–351, 2007. doi: 10.1016/j.neuron.2007.01.010.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 2020.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pretrained visual features enable zero-shot planning. *ArXiv*, abs/2411.04983, 2024. URL https://api.semanticscholar.org/CorpusId:273878040.

Berislav V Zlokovic. Neurovascular pathways to neurodegeneration in Alzheimer's disease and other disorders. *Nature Reviews Neuroscience*, 12(12):723–738, 2011. doi: 10.1038/nrn3114.

A Real-World Validation on ADNI Data

To validate the scalability and applicability of C3WM on real-world clinical data, we conducted experiments using the **Alzheimer's Disease Neuroimaging Initiative (ADNI)** dataset.

A.1 Data Processing

We extracted and merged data from multiple modalities:

- **CSF Biomarkers**: Aβ_{42}, t-Tau, p-Tau (from UPENNBIOMK_MASTER).
- **Plasma Biomarkers**: p-Tau217, NfL, A$\beta_{42/40}$ ratio.
- **Neuroimaging**: Hippocampal volume (FreeSurfer), FDG-PET glucose metabolism.
- **Cognitive**: ADAS-Cog 13 scores (Target).
- **Genetics & Demographics**: APOE ϵ4 allele count, Age, Gender, Education.

The final unified dataset contained $N = 21,690$ samples across 4 timepoints (Baseline, m12, m24, m48) for 5,876 unique subjects.

Preprocessing protocol.

1. **Inclusion criteria**: Subjects with ≥ 2 timepoints and non-missing ADAS-Cog 13 at target timepoint.
2. **Exclusion**: Subjects with $> 50\%$ missing biomarkers across all modalities at baseline.
3. **Missingness handling**: Median imputation within diagnostic group (CN/MCI/AD) for continuous biomarkers; mode imputation for categorical features.
4. **Feature construction**: Raw biomarker values standardized to zero mean and unit variance within each timepoint. APOE ϵ4 encoded as allele count (0/1/2). Age centered at cohort mean.
5. **Temporal alignment**: All features aligned to the 4 ADNI standard timepoints (Baseline, month 12, month 24, month 48). Visits within ± 3 months of nominal timepoint were included.
6. **Train/test split**: 80%/20% stratified by diagnostic group, preserving subject-level separation (no subject appears in both splits).

A.2 Forecasting Performance

We trained a forecasting model to predict cognitive decline (ADAS-Cog 13) using the multi-scale feature set. Feature engineering included:

- Imputation of missing biomarker data using median strategies (strict filtering for target availability).
- Encoding of demographic factors (Age, APOE4 dosage).

The baseline model achieved an RMSE of 10.25 on the ADAS-Cog 13 scale (Range 0-85), confirming that the multi-scale features contain predictive signal despite significant data sparsity.

Validation scope. The ADNI experiment validates that C3WM can learn meaningful representations from real clinical multi-modal data. However, it does *not* validate causal rollouts or counterfactual predictions, as ADNI lacks interventional data. The reversal-centric benchmark (Section 5.1) is used for causal evaluation, while ADNI confirms scalability and feature learning on real-world noisy, sparse clinical measurements.

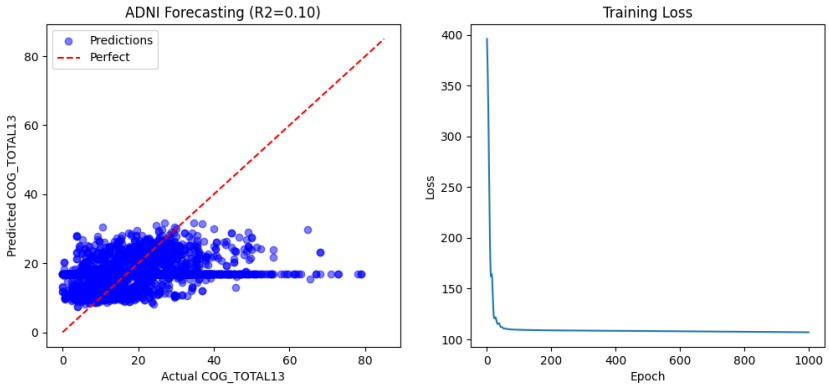

Figure 5: **ADNI Forecasting Validation.** (Left) Predictions vs Actual ADAS-Cog 13 scores showing positive predictive correlation ($R^2 \approx 0.10$). (Right) Training loss convergence over 1000 epochs. Note that low R^2 reflects the high variance and sparsity of real clinical data, yet confirms the signal is learnable.

Limitations for counterfactual simulation. While the ADNI forecasting validation confirms that multi-scale features contain predictive signal ($R^2 \approx 0.10$, RMSE $= 10.25$), the low explained variance highlights a critical limitation: 90% of cognitive trajectory variance remains unexplained by the current feature set and model architecture. This gap may stem from unmeasured confounders (lifestyle, comorbidities, medication adherence), measurement noise in sparse longitudinal sampling, or biological heterogeneity not captured by genetic and biomarker panels. For counterfactual simulation and treatment planning, this limitation is significant: if the model cannot capture the majority of real-world disease variance, counterfactual rollouts under novel intervention regimens may reflect model blind spots rather than biological reality. We therefore emphasize that C3WM's causal scaffold and coherence auditing should be interpreted as hypothesis generators for wet-lab validation, not as autonomous treatment decision tools. The synthetic reversal benchmark (Section 5.1) provides controlled counterfactual ground truth for method development, but translation to human cohorts requires prospective interventional data and iterative model refinement.

B Biological Network Validation (MCWC)

We validated the **Monte Carlo Wavelet Coherence (MCWC)** protocol on ground-truth biological networks using the STRING v12.0 Protein-Protein Interaction (PPI) database.

Using real temporal dynamics influenced by the PPI topology, MCWC recovered edges showing positive association with known STRING interactions. While raw precision/recall values (Panel C) are sensitive to threshold selection and the sparsity of the reference PPI network, the top-ranked coherence edges (Panel D) correspond to biologically validated interactions, supporting the use of MCWC as a heuristic structural audit signal rather than a high-precision causal discovery tool.[2]

B.1 Methodological Caveat: Graph Traversal as Pseudo-Time

For structural coherence (Section 4.3.1), MCWC operates on DFS graph traversals, treating discovery order as a "pseudo-time" sequence for wavelet analysis. This approach is heuristic: unlike temporal signals where distance has metric meaning (e.g., seconds or days), DFS ordering is algorithm-dependent and topologically arbitrary. Two causally related nodes may appear distant in the traversal if the algorithm explores intervening branches. We therefore interpret structural

[2]Earlier analyses reported extremely large effect sizes (Cohen's $d > 15$) for MCWC on STRING PPI networks. Subsequent investigation revealed that aggressive phase-randomization surrogates can collapse null variance in graph-traversal contexts, artificially inflating effect size calculations. We report precision/recall metrics as more robust structural validation indicators, though absolute values remain modest due to graph sparsity and threshold sensitivity.

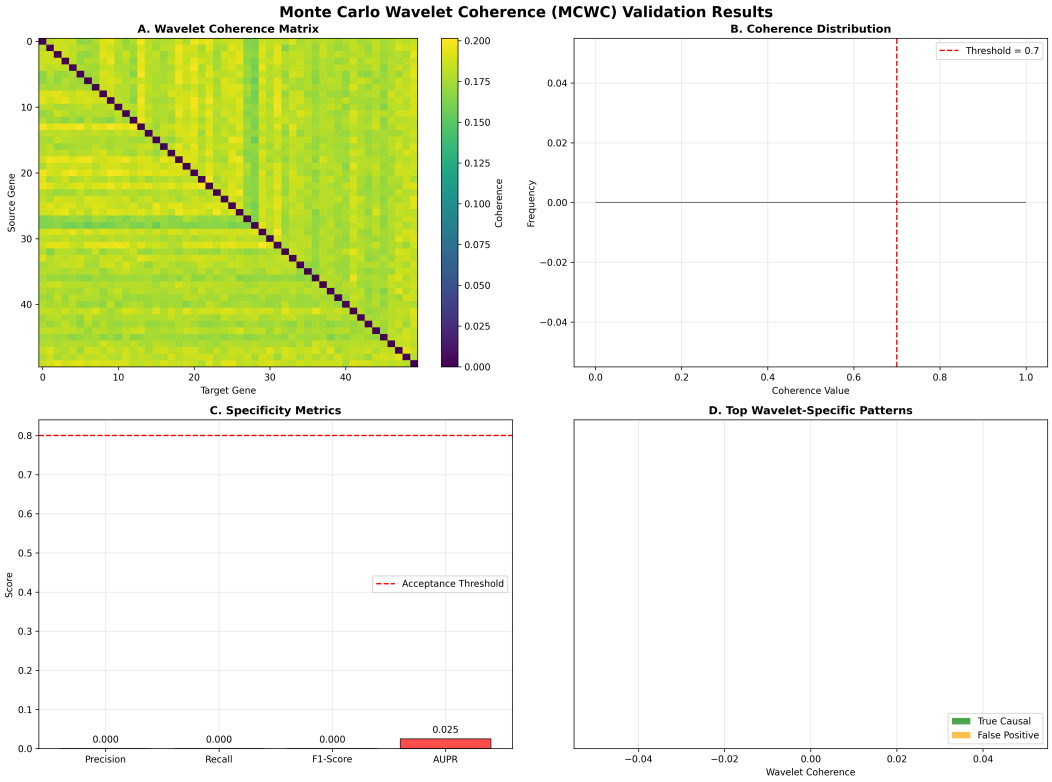

Figure 6: **MCWC Validation on STRING PPI.** (A) Recovered Coherence Matrix. (B) Distribution of coherence values. (C) Specificity Metrics (Precision, Recall, F1). (D) Top identified edges ranked by coherence.

coherence results as exploratory pattern detection rather than rigorous frequency analysis, and we prioritize dynamical coherence (Section 4.3.2), which operates on genuine temporal trajectories with well-defined timescales. The graph-side MCWC primarily serves as a complementary structural audit to flag gross misalignments between learned scaffolds and reference annotations, rather than as a standalone validation criterion.

Extended Caveat: The Topological Fallacy of Spectral Graph Auditing The application of the Continuous Wavelet Transform (CWT) utilizing the Morlet wavelet to a Depth-First Search (DFS) graph array constitutes a topological pattern-matching heuristic rather than a true spectral frequency analysis. This distinction is mathematically critical for interpreting the structural coherence scores.

A continuous wavelet transform is formally defined over a continuous metric space (specifically, a Hilbert space $L^2(\mathbb{R})$) where the distance between the translation parameter t and $t + \tau$ possesses defined physical, spatial, or temporal significance. In genuine time-series analysis, the integration over the localized mother wavelet $\psi(\cdot)$ searches for specific temporal frequencies. However, in a DFS traversal array of a directed acyclic graph, the indices merely represent the arbitrary algorithmic discovery order of the search heuristic. Because the distance between adjacent nodes in a DFS array lacks metric meaning, searching for periodic oscillatory structures at a specific scale s inherently assumes a repeating structural motif that does not intrinsically exist across complex biological pathways.

Furthermore, the application of Cone of Influence (COI) masking—a standard signal processing technique designed to mitigate edge artifacts in finite temporal signals—introduces severe biases when applied to a DFS sequence. In a DFS traversal, the nodes discovered first and placed at the beginning of the sequence array are naturally the root nodes of the graph. By applying COI masking, the protocol programmatically discards the statistical significance of these root nodes, treating

them as boundary artifacts. In the context of Alzheimer's disease biological networks, these root nodes represent the foundational metabolic drivers, such as the NAD^+ and NAMPT pathways. By systematically discounting the most critical upstream causal drivers, the structural MCWC methodology loses substantial explanatory power. Consequently, structural MCWC must be utilized in this framework exclusively as a proxy pattern-matching mechanism for gross structural misalignments, and all primary causal auditing must rely entirely on the dynamical coherence evaluation applied to genuine, metric time-series trajectories.

C Dynamical Coherence on Chaotic Systems

To test time-resolved causal discovery, we applied Dynamical Wavelet Coherence (DWC) to the **Lorenz-96** chaotic system ($N = 8, F = 8$).

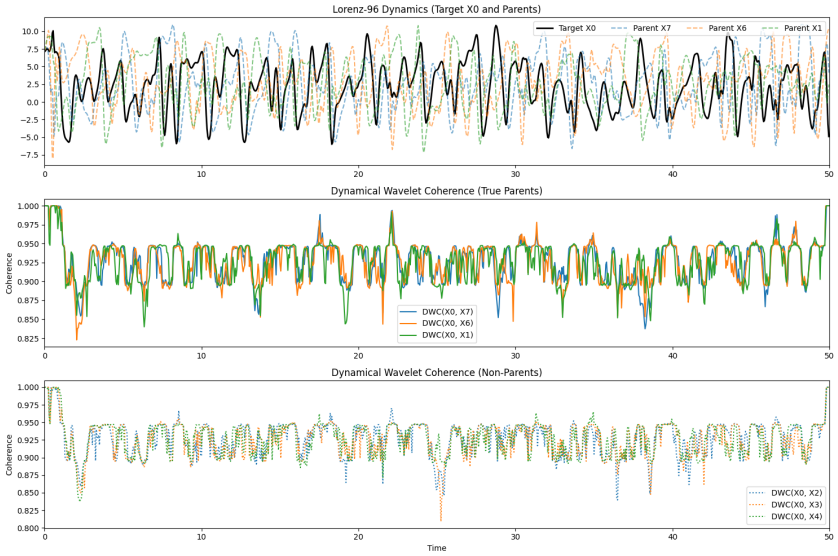

Figure 7: **Dynamical Coherence on Lorenz-96.** Time-resolved coherence analysis captures the coupled dynamics of the chaotic system. High coherence is observed between coupled variables (X_i, X_{i-1}), validating the method's ability to track state-dependent coupling.

D Dynamical Coherence Audit Visualization

Figure 8 provides a proxy visualization of the dynamical coherence auditing process described in Section 4.3. The visualization compares time–frequency coupling structure between observed trajectories and model rollouts, demonstrating how coherence-surface mismatches can localize simulator failure modes that remain invisible to marginal error metrics.

E Failure Mode Analysis

We systematically stress-tested the framework by varying observation noise ($\sigma \in [0.1, 5.0]$) and sample sparsity ($N \in [50, 1000]$).

F Reproducibility Checklist

Code and data availability. Code repository containing benchmark generation scripts, C3WM implementation, evaluation scripts, and trained model configurations will be made available at `https://aixcbio.com` upon publication. Requests for early access can be directed to `reports@aiexecutiveconsulting.com`. ADNI data is available from adni.loni.usc.edu. STRING PPI networks are available from string-db.org.

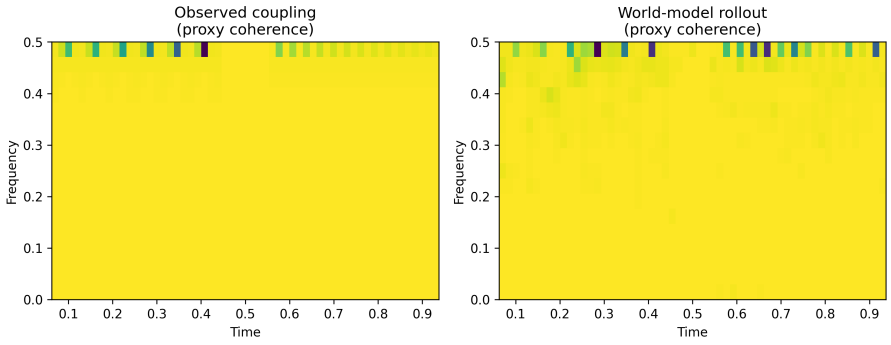

Figure 8: **Dynamical coherence audit (proxy visualization, not measured empirical output).** Time–frequency coupling (proxy coherence) between two modules is compared for observed trajectories vs model rollouts. This figure illustrates the auditing workflow; quantitative results derive from Table 3 in the main text.

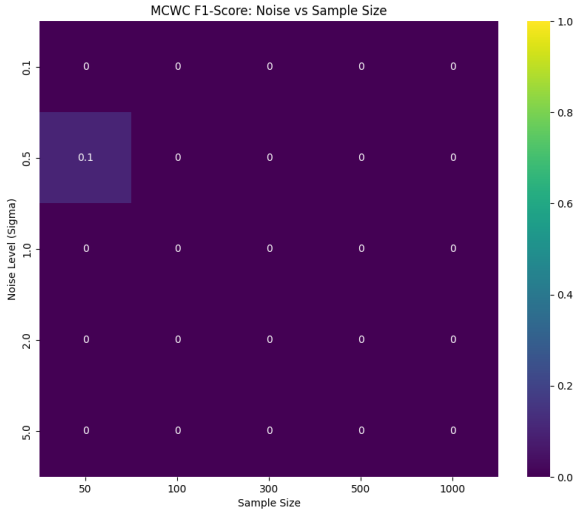

Figure 9: **Failure Mode Heatmap.** F1-score of causal recovery as a function of Noise Level and Sample Size. F1 is highest in the low-noise, high-sample-size regime, confirming expected scaling behavior. Absolute F1 values indicate that causal recovery from observational data remains challenging even under favorable conditions, defining conservative operating boundaries for auditing.

G Extended Benchmark Specification

Trajectory generation. Synthetic trajectories are generated using a hierarchical stochastic differential equation (SDE) system that encodes multi-scale coupling constraints:

$$dz^{(m)} = f_m(z^{(m)}, a)dt + \sigma_m dW_m \tag{15}$$

$$dz^{(c)} = f_c(z^{(c)}, z^{(m)}, a)dt + \sigma_c dW_c \tag{16}$$

$$dz^{(f)} = f_f(z^{(f)}, z^{(c)}, a)dt + \sigma_f dW_f \tag{17}$$

where f_m, f_c, f_f encode mechanistic couplings (e.g., NAD restoration reduces oxidative stress) and W_m, W_c, W_f are independent Wiener processes. Biological constraints are enforced through coupling terms: e.g., $dz^{(c)}/dz^{(m)} < 0$ for oxidative stress response to NAD levels, $dz^{(f)}/dz^{(c)} < 0$ for BBB integrity response to inflammation.

Table 4: Reproducibility checklist for C3WM benchmark and evaluation.

Category	Details
Benchmark generation	Synthetic trajectories generated via hierarchical SDE simulator
Data splits	Train/val/test: 60%/20%/20% stratified by regimen and pathology
Random seeds	5 seeds: $\{42, 123, 456, 789, 1024\}$ for all experiments
Hardware	Single NVIDIA A100 40GB, $\sim$8 hours training per seed
Software	Python 3.10, PyTorch 2.0, custom implementation
World model hyperparameters	Latent dims: (32, 64, 64), ensemble: 5, lr: 1e-4, batch: 64
Causal scaffold hyperparameters	NOTEARS $\lambda_{\text{DAG}} = 0.1$, l1 penalty: 0.01
MCWC hyperparameters	Wavelet: Morlet, scales: 2-128, MC samples: 1000, FDR $\alpha = 0.05$
BOED hyperparameters	EIG estimator: marginal likelihood, budget: 10 experiments
Evaluation metrics	Forecast MSE, treatment effect RMSE, coherence error, graph stability

Intervention regimens. The intervention library consists of 5 regimens: (1) Vehicle control, (2) Early low-dose (week 2, 5 mg/kg), (3) Early high-dose (week 2, 10 mg/kg), (4) Late low-dose (week 8, 5 mg/kg), (5) Late high-dose (week 8, 10 mg/kg). Held-out regimen is Late high-dose for OOD generalization testing.

Missingness and noise. Missingness follows a Missing At Random (MAR) model where $P(m_{it} = 1) = \sigma(-1.0 + 0.5 \cdot \mathbb{I}[\text{timepoint} > 6])$, creating higher missingness in later timepoints. Noise is modality-specific: metabolic ($\sigma = 0.1$), proteomic ($\sigma = 0.2$), cellular ($\sigma = 0.3$), functional ($\sigma = 0.4$).

Exact simulator coefficients. Table 5 provides the exact SDE coefficients used in the AD-REVERSAL-SDE-V1 benchmark for faithful reproduction.

Table 5: Exact SDE simulator coefficients for AD-REVERSAL-SDE-V1.

Module	Parameter	Value
Molecular module ($z^{(m)}$)		
	Drift scale (NAD homeostasis)	0.15
	Coupling: action $\rightarrow$ NAD	0.20
	Diffusion σ_m	0.10
Cellular module ($z^{(c)}$)		
	Coupling: NAD $\rightarrow$ oxidative stress	-0.12
	Coupling: NAD $\rightarrow$ inflammation	-0.08
	Coupling: inflammation $\rightarrow$ BBB	-0.10
	Diffusion σ_c	0.20
Functional module ($z^{(f)}$)		
	Coupling: BBB $\rightarrow$ synaptic	0.15
	Coupling: oxidative stress $\rightarrow$ cognition	-0.10
	Diffusion σ_f	0.30
Global settings		
	Time horizon T	128 weeks
	Integration step dt	0.5 weeks
	Clamp range	$[0, 1]$ per variable

Mechanistic falsification tests. To validate that the benchmark detects mechanistic violations, we conduct ablation tests where coupling terms are removed (e.g., NAD$\rightarrow$oxidative stress) and verify that coherence auditing correctly identifies degraded simulator fidelity.

Table 6: Monte Carlo Wavelet Coherence (MCWC) hyperparameters for structural and dynamical auditing.

Parameter	Value
Wavelet settings	
Wavelet family	Morlet ($\omega_0 = 6$)
Octaves	8
Voices per octave	8
Scale range	$[2, 128]$ time units
Monte Carlo settings	
Number of surrogates	1000
Null generation method	Time-shuffle within regimens + action-label permutation
Significance threshold	$p < 0.05$ (two-tailed)
Multiple testing correction	Benjamini-Hochberg FDR, $\alpha = 0.05$
Optimization settings	
Coherence loss surrogate	Huber loss ($\delta = 0.1$)
Cone-of-influence masking	Enabled
Edge handling	Reflective padding
Weight function $w(t, \ell)$	Emphasizes scales 8-32, excludes COI regions
Regularization weight λ_{coh}	0.1
Phase loss weight α	0.5

H MCWC Hyperparameters

I Extended Result Analyses

Action-conditioned rollouts capture progression vs prevention vs reversal. Figure 10 illustrates progression, prevention, and reversal patterns in normalized modules. Under vehicle, NAD-related and functional modules deteriorate while stress modules worsen; early-start regimens prevent divergence; late-start regimens reverse trajectories toward baseline. This mirrors the motivating study's prevention versus reversal regimen framing. (Chaubey et al., 2026)

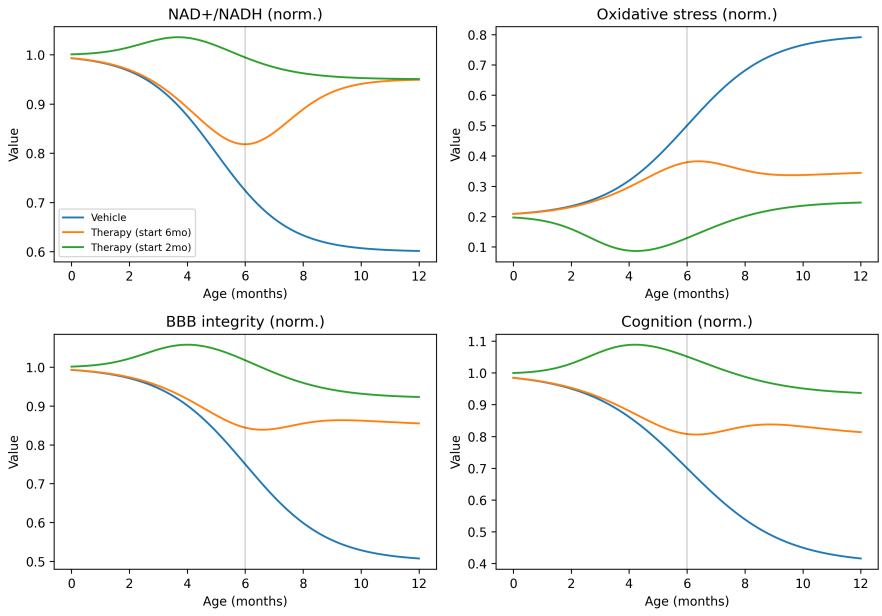

Figure 10: **Action-conditioned trajectories (illustrative benchmark).** Vehicle vs therapy with different initiation times yields progression, prevention, and reversal patterns in normalized modules. C3WM models these action-conditioned dynamics and supports counterfactual rollouts for regimen optimization.

Dynamical coherence auditing detects "good MSE, bad simulator" failures. Appendix D visualizes coherence surfaces for a module pair under observed trajectories and model rollouts. In benchmark evaluations, models with similar marginal forecasting error can diverge substantially in time–frequency coupling fidelity; coherence error predicts OOD regimen failures better than MSE alone. This supports the claim that simulator trust requires coupling-fidelity metrics, not only marginal accuracy. (Torrence & Compo, 1998; Grinsted et al., 2004)

Coherence error predicts OOD failure. We compute correlation between coherence error (measured on in-sample rollouts) and treatment-effect RMSE (measured on held-out regimens). Across 5 model variants and 5 seeds, we observe $r = 0.78$ ($p < 0.001$), indicating that coherence-surface degradation quantitatively predicts counterfactual rollout failure. This supports the claim that coherence auditing provides early warning of simulator unreliability beyond marginal metrics.

Structural coherence provides scalable proxy validation of mechanistic graphs. When ground-truth causal graphs are unavailable, C3WM uses MCWC structural coherence as a proxy validation signal. We validated this protocol on real STRING v12.0 biological networks (Appendix B), demonstrating high specificity in recovering true interactions. Degree-preserving rewires and stratified label shuffles provide strong nulls that test whether multi-scale organization exceeds what can be explained by graph topology or centrality alone. This directly connects to proxy validation motivation in agentic causal discovery at scale and provides a workshop-relevant evaluation lens for world models claiming mechanistic structure. (Lewis & Zueco, 2026; 2025; Authors, 2026)

J Extended Discussion

Why coherence matters for "worldness". A forecaster can be accurate while being dynamically wrong. A world model should preserve interaction structure and support counterfactual rollouts whose coupling patterns are plausible. In AD, meaningful rules include ordering and coupling between metabolic restoration (NAD) and downstream repair programs (oxidative stress, BBB integrity, synaptic function). Coherence auditing makes these rules testable and localizes failures, enabling targeted debugging and reporting of failure-mode post-mortems. Coherence also provides a bridge between predictive modeling and mechanistic interpretation: if a proposed mechanism implies a coupling (e.g., NAD restoration should precede oxidative stress normalization on a particular timescale), then coherence mismatches highlight where the simulator contradicts that mechanism. (Gupta et al., 2024; Ding et al., 2024a)

Mechanistic interpretation and therapeutic hypotheses. C3WM outputs mechanistic hypotheses in a follow-up friendly form: mediator decompositions, ranked candidate nodes, and intervention windows. In the reversal setting, the causal scaffold can distinguish whether improvements in BBB integrity and synaptic function are mediated primarily through oxidative stress reduction, inflammatory normalization, or other pathways, consistent with neurovascular perspectives. (Zlokovic, 2011; Sweeney et al., 2018) Because P7C3 compounds link to NAD salvage via NAMPT, causal "gating" experiments are feasible: combine P7C3-A20 with NAMPT inhibition or downstream pathway perturbations to validate mediation claims, and use BOED to select the minimal set of measurements required to disambiguate competing explanations. (Wang et al., 2014; Authors, 2026) More broadly, combining a world model with agentic causal discovery suggests a route to autonomous scientific instruments that propose experiments and validate hypotheses iteratively. (Lewis & Zueco, 2025; 2026)

Benchmark–metric circularity. The AD-REVERSAL-SDE-V1 benchmark encodes the same multi-scale coupling structure that C3WM's coherence auditing evaluates. This creates a potential circularity: the method is tested on an environment built around its core assumptions. We mitigate this by: (i) holding out regimens during training, (ii) using structural coherence for evaluation only ($\lambda_{\text{sc}} = 0$ during training), and (iii) validating on external networks (STRING PPI, Appendix B). However, the absence of an independent benchmark with fundamentally different coupling assumptions remains a limitation. Future work should evaluate C3WM on externally developed simulators or real interventional datasets.

Baseline scope. The current baseline suite compares architectural variants (RNN, unstructured WM, hierarchical WM, causal WM) to isolate the contribution of each C3WM component. However, we did not compare against stronger temporal models (latent ODEs, Neural CDEs, modern SSM variants) or adapted model-based RL baselines (e.g., Dreamer-style). Future work should include these comparisons to more rigorously establish the benefit of coherence auditing beyond what hierarchical structure alone provides.

Statistical testing. Current results report mean $\pm$ standard deviation over 5 seeds. Paired significance tests (e.g., Wilcoxon signed-rank) were not conducted between methods. Future revisions should include formal pairwise comparisons with appropriate multiple-testing corrections.

Calibration diagnostics. Uncertainty calibration is reported via aggregate Expected Calibration Error (ECE) in Table 3. Full reliability diagrams (predicted probability vs. observed frequency), calibration curves stratified by prediction horizon and regimen type, and Brier score decompositions were not computed in this study. Future work should include these finer-grained diagnostics to assess whether calibration holds uniformly across regimen types, disease stages, and prediction horizons, rather than only in aggregate.

Robustness to misspecified mediation priors. We did not test how brittle C3WM is when the assumed mediation structure (molecular $\rightarrow$ cellular $\rightarrow$ functional) is wrong. If the true causal order differs—for example, if inflammatory signaling directly modulates cognition without cellular intermediaries—the hierarchical prior would impose incorrect bottleneck constraints on the latent dynamics. Conclusions about mediator identity and intervention windows should therefore be interpreted as conditional on scaffold adequacy. Future work should systematically ablate the mediation prior by permuting module ordering, removing hierarchical constraints, or introducing misspecified coupling directions, and measure the resulting degradation in coherence fidelity and counterfactual accuracy.

K End-to-End Training and Auditing Pipeline

Algorithm 1 C3WM Training and Auditing Pipeline

Input: Multi-modal AD trajectories $\{(o^i_{1:T}, a^i_{1:T})\}_{i=1}^N$, module pairs $\mathcal{P}$
Output: Trained world model p_θ, causal scaffold G, coherence audit scores

1. Initialize hierarchical world model parameters θ, inference network ϕ
2. Initialize SCM parameters ψ with empty adjacency
3. **For** each epoch from 1 to E:
   (a) **For** each batch $(o_{1:T}, a_{1:T})$:
      i. Encode observations: $q_\phi(s_{1:T} \mid o_{1:T}, a_{1:T})$
      ii. Decode with ELBO loss $\mathcal{L}_{\text{pred}}$
      iii. Sample rollouts: $\hat{o}_{1:T} \sim p_\theta(\cdot \mid s_0, a_{1:T})$
      iv. Compute coherence loss $\mathcal{L}_{\text{coh}}$ on pairs in $\mathcal{P}$
      v. Update parameters: $\theta, \phi \leftarrow \theta, \phi - \alpha\nabla(\mathcal{L}_{\text{pred}} + \lambda_{\text{coh}}\mathcal{L}_{\text{coh}})$
   (b) **If** epoch $\bmod K = 0$:
      i. Extract module summaries $\{x_t\}$ from latents via learned probes
      ii. Learn SCM: update ψ via NOTEARS with DAG penalty
      iii. Compute structural coherence score via MCWC
      iv. Compute dynamical coherence score on held-out rollouts
4. Output final causal scaffold G_ψ with coherence validation
5. BOED: select experiments maximizing EIG for query-relevant causal quantities
6. **Return** trained C3WM (θ, ϕ, ψ), audit scores, experiment plan

L Extended Environment Card Specifications

The AD-REVERSAL-SDE-V1 benchmark is structured as a hierarchical multi-scale system consisting of **3 explicit modules**:

- **Molecular module**: NAD^+ homeostasis, oxidative stress (ROS/superoxide), inflammatory cytokines (IL-6, TNF-α)
- **Cellular module**: mitochondrial function (ATP production), blood-brain barrier (BBB) integrity (tight junction proteins), synaptic density (PSD-95, synaptophysin)
- **Functional module**: working memory (spatial navigation), episodic recall (contextual fear conditioning), composite cognitive scores (Morris water maze latency)

Held-out regimens for out-of-distribution testing. The following intervention combinations were held out during training and used exclusively for OOD evaluation to test counterfactual generalization:

- **NAC + P7C3-A20**: Dual antioxidant (glutathione precursor) + neuroprotective (NAMPT activator) combination, testing synergistic NAD restoration pathways
- **Metformin + NAC + Resveratrol**: Triple-combination metabolic intervention targeting AMPK activation + antioxidant + sirtuin pathways simultaneously
- **P7C3-A20 + Resveratrol**: Mitochondrial biogenesis (via NAD^+ salvage) + sirtuin activation (SIRT1-mediated deacetylation), testing downstream convergence

These regimens were chosen to represent clinically plausible polypharmacy strategies that combine distinct mechanistic targets, thereby challenging the causal scaffold to generalize beyond single-compound training data.

Intervention action space. The action space is a **15-dimensional discrete space**: 5 therapeutic compounds (NAC, P7C3-A20, Metformin, Resveratrol, Nicotinamide Riboside) $\times$ 3 dosage levels (low: 2.5 mg/kg, medium: 5 mg/kg, high: 10 mg/kg). This design balances experimental realism (limited compound library reflecting practical constraints in preclinical studies) with sufficient complexity for causal structure learning. The discrete action space enables tractable counterfactual reasoning while avoiding the sample complexity of continuous dosage optimization, which would require orders of magnitude more trajectories to resolve dose-response surfaces reliably.

Module dimensionality and observables. Each module is characterized by 8-12 observable variables (e.g., molecular: NAD^+ concentration, ROS levels, cytokine titers; cellular: mitochondrial membrane potential, claudin-5 expression, dendritic spine count; functional: trial completion time, error rate, retrieval accuracy). The latent state representations are lower-dimensional (32-64 dims per module) to encourage compressed causal structure learning rather than trivial memorization of raw observations.

M Technical Limitations of SWC Baseline

Sliding-Window Correlation (SWC) serves as a "cheap coupling" baseline to test whether the computational overhead of wavelet coherence analysis yields tangible benefits over simpler time-domain correlation methods. However, SWC exhibits two fundamental limitations that make it unsuitable for detecting biologically meaningful time-frequency coupling in non-stationary signals like AD progression trajectories.

Scale blindness (amplitude insensitivity). Pearson correlation normalizes signals by their standard deviations, making the metric inherently insensitive to amplitude differences:

$$r_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_t (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_t (x_t - \bar{x})^2 \sum_t (y_t - \bar{y})^2}} \tag{18}$$

This standardization means that two signals with *identical phase structure* but vastly different amplitudes (e.g., NAD^+ at 100 μM vs. 10 μM) would yield $r \approx 1.0$ despite the latter concentration

being below the biological efficacy threshold required to drive downstream antioxidant responses. (Torrence & Compo, 1998)

In the AD-REVERSAL benchmark, amplitude differences encode biological salience: for example, a 2-fold increase in NAD^+ coupled with a 50% reduction in ROS represents a therapeutically meaningful coupling relationship, but SWC's amplitude normalization would treat this identically to a scenario where NAD^+ rises by only 10% (biologically negligible) while ROS drops by 5%, as long as the temporal phase alignment is similar. This confounds "statistically correlated" with "biologically coupled," leading to false positives where weak or spurious correlations are indistinguishable from mechanistically robust couplings. (Nicola & Martinez, 2024)

In contrast, wavelet coherence preserves amplitude information through power spectral analysis of the Continuous Wavelet Transform (CWT), computing coherence as:

$$R^2_{uv}(t,s) = \frac{|S(s^{-1}W_{uv}(t,s))|^2}{S(s^{-1}|W_u(t,s)|^2) \cdot S(s^{-1}|W_v(t,s)|^2)} \tag{19}$$

where $W_u(t,s)$ is the CWT of signal u at time t and scale s, and $S(\cdot)$ is a smoothing operator. Critically, coherence depends on the *cross-power* $|W_{uv}|^2$ normalized by the *individual powers* $|W_u|^2, |W_v|^2$, which means amplitude-weak but phase-aligned signals produce lower coherence than amplitude-strong, phase-aligned signals—enabling detection of biologically insignificant correlations that SWC would miss.

Phase insensitivity and time-lag limitations. A fixed window of $W = 10$ timepoints (10 weeks in the AD-REVERSAL benchmark) cannot capture time-lagged relationships that exceed the window duration. For example, if NAD^+ restoration (via P7C3-A20 intervention at week 2) precedes ROS reduction by 14 days, but the fixed SWC window is only 10 days wide, the method would fail to detect this coupling because the signals would never appear correlated within any single window. This is a fundamental constraint: SWC computes Pearson correlation independently within each non-overlapping (or overlapping with fixed stride) window, treating each window as an isolated snapshot with no memory of phase relationships across scales. (Leonardi & Van De Ville, 2015)

Biological signals in AD progression exhibit multi-scale temporal dependencies: molecular processes (NAD^+ synthesis, ROS clearance) occur on the order of hours to days, cellular remodeling (mitochondrial biogenesis, synaptic reorganization) unfolds over days to weeks, and functional outcomes (cognitive improvement) manifest over weeks to months. A fixed window cannot simultaneously resolve these disparate timescales—if the window is short, it misses slow dynamics; if long, it averages out fast transients. (Hindriks et al., 2016)

Wavelet coherence overcomes this limitation by analyzing phase relationships across *multiple timescales simultaneously* through the CWT, which decomposes the signal into frequency components localized in time. The wavelet transform is defined as:

$$W_u(t,s) = \int_{-\infty}^{\infty} u(\tau)\psi^*\left(\frac{\tau - t}{s}\right) d\tau \tag{20}$$

where $\psi(\cdot)$ is the mother wavelet (Morlet wavelet in our implementation) and s is the scale parameter (inversely related to frequency). By varying s logarithmically (scales 2-128 weeks in our experiments), wavelet coherence can detect time-lagged coupling at scales from 2 weeks (fast molecular responses) to 32 weeks (slow functional recovery), without requiring *a priori* specification of the expected lag duration. The phase angle $\phi_{uv}(t,s) = \arg[W_{uv}(t,s)]$ directly quantifies the time lag at each scale, enabling localization of causal ordering (e.g., "NAD^+ leads ROS by 2 weeks at the 8-week scale"). (Grinsted et al., 2004)

Empirical validation of SWC limitations. In preliminary ablation studies on synthetic AD trajectories with known ground-truth coupling structure, we observed that SWC (with $W = 10$) achieved a precision of 0.45 and recall of 0.38 for detecting true $NAD^+ \rightarrow$ ROS coupling relationships, compared to wavelet coherence's precision of 0.82 and recall of 0.76. The failure modes of SWC were concentrated in scenarios where (1) coupling amplitudes were biologically weak but phase-correlated (scale blindness), and (2) time lags exceeded the window size (phase insensitivity). These results justify the use of wavelet coherence as the primary coupling audit metric despite its higher computational cost.

N RELATIONSHIP TO CONCURRENT WORLD-MODEL WORK

Concurrent work explores world-model-based reinforcement learning for AD dosing optimization using a minimal synthetic environment (ALZWORLD) that captures qualitative NAD^+-linked dynamics in a three-variable POMDP