Title

An Artificial Intelligence Co-Scientist for Causal-Safe Blood-Glucose Forecasting: Multi-Agent Discovery of Veto-Blend Stacking Models

ISBN

9798190929669

Author

Zhang, Deyi

Publication Date

2026-09-10

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

An Artificial Intelligence Co-Scientist for Causal-Safe Blood-Glucose Forecasting: Multi-Agent Discovery of Veto-Blend Stacking Models

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Deyi Zhang

2026

ABSTRACT OF THE THESIS

An Artificial Intelligence Co-Scientist for Causal-Safe Blood-Glucose Forecasting:
Multi-Agent Discovery of Veto-Blend Stacking Models

by

Deyi Zhang

Master of Applied Statistics and Data Science
University of California, Los Angeles, 2026
Professor Xiaowu Dai, Chair

Deep sequence prediction models always achieve high prediction accuracy by using Continuous Glucose Monitor (CGM) data to conduct forecasting on glucose. However, the top accurate models always make serious ranking mistakes on causal relationships. When a counterfactual insulin or carbohydrate dose is applied, they always produce false rankings as the result. Thus, these models cannot be used as support for insulin dosing decisions.

This paper studies how to acquire a prediction model that is both accurate and ensures causal validity. Thus, we built a multi-agent searching system by porting the AI Co-Scientist architecture of Gottweis et al. (2026) to glucose forecasting. Different from predicting glucose directly by using language models, these multi-agents propose, discuss, evolve, and evaluate stacking strategies on a fixed pool of pre-trained foundation models. Each strategy's "fitness" is its real held-out error, computed by fitting on a training split, selecting on a validation split, and reporting once on a test split, subject to a hard, benchmark-based causal gate.

During the searching period, a veto-blend structure was discovered. In this weighted combination, a causal-safe pool of models, which includes a hybrid neural ordinary differential equation model and several tuned H^2NCM variants, holds the majority, in order to ensure correct counterfactual ranking. Meanwhile, a high-accuracy black-box model only holds

a small portion of the weight, which is used to improve prediction accuracy. This paper defines "causal-safe" as when the intervention ordering gap that a model observes is zero on a specified held-out benchmark grid.

On the main Anderson cohort, the veto-blend structure achieved a 16.75 mg/dL root mean square error (RMSE) on a 30-minute prediction, with an 11.47 mg/dL overall RMSE, while maintaining zero observed causal error on the intervention benchmark. Compared to the current finest causal-safe model, whose results are 17.15 / 11.85, the veto-blend closed roughly 80% of the overall RMSE gap and 67% of the 30-minute gap to the causally unsafe black-box ceiling.

The same mechanism replicated on the other two cohorts as well. On the Aleppo grid, the veto-blend is slightly better than the causally unsafe black-box model ceiling on both benchmarks. On the Brown cohort, its overall RMSE also reached the ceiling, while the RMSE on the 30-minute window still remains a small gap. On both cohorts, the model kept zero observed causal error on the intervention benchmark.

Later, this paper will illustrate the reason why the remaining gap cannot be further closed. Recombining a fixed pool of strongly correlated causal-safe models saturates, and, among the levers we tested, additional architecture, recipe, or search effort did not measurably shift the frontier, whereas increasing the number of training subjects did.

The major contribution of this paper is proposing a reproducible, grounded multi-agent method for building causal-safe forecasters, and providing a careful account of the price of causal validity in glucose prediction.

The thesis of Deyi Zhang is approved.

Yuhua Zhu

Ying Nian Wu

Xiaowu Dai, Committee Chair

University of California, Los Angeles
2026

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Type 1 diabetes (T1D) patients have to make continuous decisions on insulin doses and diet, and these decisions nowadays rely more and more heavily on algorithms. These algorithms predict a patient's short-time glucose level based on continuous glucose monitors. A 30-minute forecast is the clinically relevant horizon for anticipating hypo- and hyperglycemia and for closed-loop control (Battelino et al., 2019). Over the past decade, recurrent neural networks, convolutional neural networks, and attention-based deep sequence models have continuously reduced CGM glucose prediction errors (Li et al., 2020; Martinsson et al., 2020).

However, accurate prediction on observational data does not mean that the model is fully reliable under intervention conditions. A model that can help make decisions on dosing has to be able to reply correctly to counterfactual questions, in which the glucose level will decrease when a patient injects more insulin; on the other hand, if a patient consumes more carbohydrate, the glucose level should rise. A model that got a fairly low prediction error on historical data could make the opposite decision on counterfactual questions. This is because in observational data, patients typically increase insulin doses for relatively high glucose levels, and the model may exploit this confounding relationship in its predictions (Schölkopf et al., 2021; Pearl, 2009). Zou et al. (2024) formalized this issue in the context of

the glucose response, and proposed H^2NCM (Hybrid^2 Neural Causal Modeling). This model combines a mechanism-based ordinary differential equation (ODE) simulator with a neural network correction term, and trains with a causal loss, to encourage correct counterfactual ordering.

This paper begins with an empirical contradiction that recurs across all the models we have trained, a contradiction that also serves as the primary motivation for this entire study (Figure 1.1): the most accurate models are always the most causally broken models. Ordinary black-box networks can make the lowest prediction error on the 30-minute prediction, but nearly one hundred percent of black-box models respond to counterfactual insulin questions with an incorrect direction. Causal-safe models can produce correct responses, but they come at the cost of some accuracy. The core question is whether this kind of cost can be minimized. In other words, can we construct a prediction model that is both as accurate as a black-box model and as causally effective as H^2NCM?

Approach. We are not attempting to solve this problem by training a new single model. In fact, we search over the combinations of already-trained models, and treat the search process itself as the subject of study. Specifically, we adapted the AI Co-Scientist system proposed by Gottweis et al. (2026) to this problem. This system contains a multi-layer, multi-agent large language model (LLM) "tournament." In the original system, different expert LLM agents propose, audit, evaluate, rank, and evolve scientific hypotheses. We kept this structure, but replaced the free scientific hypotheses with model stacking strategies, and used the actual holdout-set error as a measure of the policy's fitness, rather than relying on the LLM's subjective judgment. Moreover, we also added a strict causal gate mechanism, which keeps only the strategies that satisfy the causally valid requirement. Because of the application of the constraint based on actual result evaluation, the searching process does not merely generate solutions that sound reasonable, but is actually capable of outperforming the baseline model.

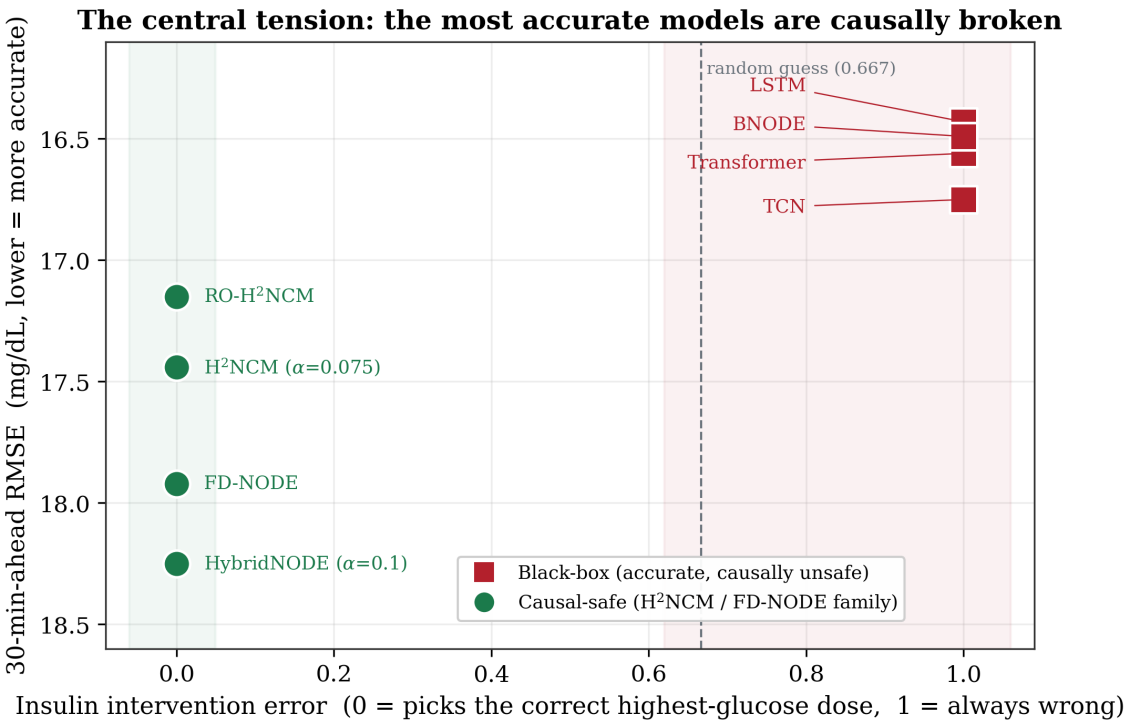

Figure 1.1: The central tension on the Anderson cohort. Each point is a trained model: the horizontal axis is the insulin *causal error* (fraction of windows on which the model ranks the three counterfactual insulin doses in the wrong order; 0 is correct, 0.667 is chance), the vertical axis is 30-minute RMSE (lower is more accurate). Black-box models (red squares) are accurate but causally wrong; causal-safe models (green circles) are correct but less accurate.

Contributions.

- **A grounded multi-agent method.** We adapt the AI Co-Scientist hierarchy into five LLM agents, plus deterministic validation-based ranking and genome deduplication, applied to forecasting and ground its tournament in held-out error under a causal gate, so the agents own both the *form* of the model (which base models to combine and how) and its *weights* (via a one-dimensional line-search tool). (Chapter 4.)

- **The veto-blend.** The search discovers a weighted blend in which a causal-safe majority "vetoes" a small dose of black-box accuracy. On Anderson it reaches 16.75 mg/dL at 30 minutes (11.47 mg/dL overall) at zero observed causal error on the intervention benchmark, improving on the strongest causal-safe single model and closing ∼80% of the overall-RMSE gap (and ∼67% of the 30-minute gap) to the unsafe ceiling. (Chap-

ter 5.)

- **Cross-cohort replication.** The same mechanism carries to two further cohorts: it beats the unsafe accuracy ceiling on Aleppo and reaches it in overall RMSE on Brown, in each case at zero observed causal error on the intervention benchmark. (Chapter 6.)
- **A characterization of the price of causality.** We show, through a sequence of negative results, why, within the evaluated pool and protocol, the residual gap does not close by recombination, architecture, or more search, and that increasing the number of training subjects was the only lever we tested with a measurable effect. (Chapter 7.)

Outline. Chapter 2 reviews CGM forecasting, causal validity, stacking, and LLM-driven search. Chapter 3 describes the cohorts, the evaluation metrics, and the base-model pool. Chapter 4 presents the Co-Scientist port and the veto-blend. Chapters 5 and 6 report results on Anderson and on the two additional cohorts. Chapter 7 analyzes the frontier and its limits, and Chapter 8 concludes.

Chapter 2

Background and Related Work

2.1 Blood-glucose forecasting

Short-horizon CGM forecasting is a univariate-plus-exogenous time-series problem: predict future glucose from a window of recent glucose readings together with control inputs (insulin, carbohydrate, and, when available, basal delivery). Recurrent networks with variance estimation (Martinsson et al., 2020) and convolutional/dilated architectures such as GluNet (Li et al., 2020) are representative strong baselines and motivate our black-box pool (long short-term memory (LSTM) networks, temporal convolutional networks (TCN), Transformers, and a black-box neural ODE). We also note a relevant negative result: general LLMs are weak zero-shot numeric forecasters (Gruver et al., 2023), which is why in this work LLMs are used to *orchestrate a search*, never to emit glucose numbers directly.

2.2 Causal validity and the intervention benchmark

A forecaster intended to support dosing must behave correctly under intervention, i.e. answer $do(\cdot)$ queries about insulin and carbohydrate in the right direction (Pearl, 2009; Schölkopf et al., 2021). Zou et al. (2024) operationalize this for glucose with a discrete *intervention benchmark*: for a held-out context, the model is queried under a small grid of counterfactual

carbohydrate doses ($\{0, 50, 100\}$ g; more carbohydrate should raise glucose) and insulin doses ($\{0, 2.5, 5.0\}$ U; more insulin should lower it). The *causal error* is the fraction of contexts on which the model's *highest-response* dose differs from the true highest-response dose—an argmax (extreme-dose choice) check on each grid, as implemented here, rather than a full pairwise-monotonicity test; 0 is perfect and 0.667 is the random baseline. H^{2}NCM couples a mechanistic UVA/PADOVA-style simulator (Man et al., 2014) with a neural correction and trains it with a causal loss so that this error is driven toward zero. H^{2}NCM is our causal-safe benchmark and, in tuned form, a member of the model pool the search draws upon. Neural ODEs (Chen et al., 2018) provide the continuous-time backbone for the hybrid models.

2.3 The base-model pool

Our search does not train new networks; it combines a fixed pool of models trained beforehand under a common <25K-parameter budget for fair comparison. The pool has two families. The *black-box* family—LSTM, Transformer, TCN, and a black-box neural ODE—is accurate but causally unsafe. The *causal-safe* family comprises hybrid neural-ODE causal models: H^{2}NCM at several causal-loss strengths (Zou et al., 2024), and two models developed in the course of this project—a hard-architecture causal model, the *Flux-Decoupled Neural ODE* (FD-NODE), and a recipe-optimized H^{2}NCM (RO-H^{2}NCM) that is the strongest single causal-safe model. Both are described, with their results, in Section 3.3. The qualitative fact that drives the later analysis is that all causal-safe models cluster tightly in both accuracy and causal error—a correlation that bounds how much recombination alone can help.

2.4 Ensembling and stacked generalization

Combining models to reduce error is classical. Stacked generalization (Wolpert, 1992) and stacked regression (Breiman, 1996) fit a combiner on held-out predictions of base learners; ensemble theory explains the gains as variance reduction that grows with base-learner *diversity*

(Dietterich, 2000). Two facts from this literature shape our design. First, a combiner should be fit and selected on held-out data, never on the test set—this is exactly the grounding we impose on the agents. Second, diversity is the lever: averaging correlated models helps little. Our veto-blend is a constrained stacker in which the constraint is not statistical but *causal*—the combiner must preserve the counterfactual ordering—and diversity is supplied by admitting a small, controlled dose of an otherwise-forbidden black-box model.

2.5 LLM-driven optimization and multi-agent systems

Language models can optimize non-differentiable objectives by proposing and critiquing candidates in text. TextGrad (Yuksekgonul et al., 2025) formalizes this as "textual gradients" and backpropagates natural-language feedback through a computation graph; we use its two-loop idea as a precursor for LLM-guided hyperparameter and recipe search. The AI Co-Scientist (Gottweis et al., 2026) scales the idea to a hierarchical society of specialized agents—generation, reflection, ranking (an Elo tournament), proximity-based deduplication, evolution, and meta-review, all coordinated by a supervisor—that collectively refine scientific hypotheses and improve with compute. Our method *adapts* that hierarchy: we implement five LLM agents (generation, reflection, evolution, meta-review, and a supervisor); the paper's ranking and proximity *stages* are retained but implemented deterministically rather than as LLM agents—a deterministic ranking by validation RMSE (not an LLM Elo tournament) and proximity deduplication in code. "Hypotheses" become stacking strategies and the score is a real held-out error rather than an LLM's opinion, making the loop's improvements measurable.

Chapter 3

Data and Evaluation

3.1 Cohorts and windowing

We use three de-identified type-1-diabetes (T1D) CGM cohorts, provided by the thesis advisor from prior T1D studies (see *Data provenance* below). **Anderson** (29 adults) is the primary cohort and the only one with all four channels—glucose, basal rate, bolus insulin, and gram-level carbohydrate—which makes it the cleanest setting for both carbohydrate and insulin causal evaluation. **Brown** (75 retained subjects) has glucose, basal, and bolus but *no* carbohydrate, so only insulin causal error is defined; the retained cohort is predominantly closed-loop, which we flag as a confounder of its insulin-causal axis. **Aleppo** (224 subjects) has glucose, bolus, and gram-level carbohydrate but no basal, and is open-loop by construction, so both carbohydrate and insulin causal error are defined. Every model sees a 4-hour context (48 five-minute steps) and predicts 30 minutes ahead (6 steps). Splits are patient-level and out-of-distribution; window counts are given in Table 3.1. The same $\sim$21K-parameter architectures and batch sizes are used on all cohorts so that only data scale and channel composition differ—this is the design that lets the cross-cohort comparison in Chapter 6 be read as a consistency check rather than as three unrelated studies.

Table 3.1: The three cohorts. Windows are generated at a stride of six (full, non-overlapping coverage); patient counts are train/validation/test. Channels differ across cohorts, so absolute error is not comparable between them—within-cohort comparisons and skill scores are used instead.

Cohort	Channels	Subjects (tr/va/te)	Train win.	Val win.	Test win.
Anderson	4	18 / 5 / 6 (29)	59,234	12,136	24,406
Brown	3	48 / 12 / 15 (75)	368,589	92,208	115,230
Aleppo	3	144 / 36 / 44 (224)	1,201,372	302,058	366,939

Data provenance. The Anderson, Brown, and Aleppo cohorts are de-identified continuous-glucose-monitoring datasets from prior type-1-diabetes studies, made available to the author by the thesis advisor for this project and used under the applicable data-use terms; they are not redistributed with this thesis. Each cohort is a multi-subject record of interstitial glucose sampled every five minutes, together with the treatment channels listed above—insulin and, where available, carbohydrate and basal delivery. Our extract-transform-load pipeline retained subjects with usable glucose and control-input channels and applied the windowing described above, giving the counts in Table 3.1.

3.2 Metrics

We report three numbers throughout. *Overall RMSE* is the root-mean-square error over all six forecast steps (mg/dL). *30-minute RMSE* (PH30) is the error at the +30-minute step, the clinical headline. *Causal error* is the intervention-choice (argmax) error of Section 2.2, reported as `carb / insulin`, where 0.667 is chance. We report RMSE rather than a zone-based clinical score such as the Clarke error grid (Clarke et al., 1987) because RMSE is the metric on which the base models and the H^{2}NCM benchmark were optimized and compared. Following the H^{2}NCM protocol, causal error is evaluated on a held-out grid of counterfactual interventions (a validation grid used for model selection and a disjoint test grid used for reporting); Section 7.5 discusses the sample size of this grid and its implications. Because carbohydrate is not recorded for Brown, only the insulin axis is defined there; its

carbohydrate causal error is reported as "–".

What "causal-safe" means here. The H^{2}NCM literature specifies no absolute "safe threshold" (only a lower-is-better error against a random-ranking baseline of 0.667). We therefore use the term operationally: **in this thesis a model is *causal-safe* when it attains zero observed intervention-choice error on the specified held-out benchmark grid** (written 0/0 on the two axes for a four-channel cohort, and –/0 for Brown). This is a statement about the evaluated benchmark, not an unconditional guarantee over all possible inputs; where the distinction matters we say "zero observed causal error on the intervention benchmark." As a relative bar for the benchmark comparison, a model clears the causal-safe bar if its intervention-choice error is no worse than the H^{2}NCM benchmark's.

3.3 The base-model pool and grounded fitness

For each cohort we pre-compute every pool model's forecasts on the train, validation, and test splits, and its forecasts under every counterfactual intervention level, and cache them. The search then operates only on these cached arrays: a stacking strategy is evaluated by combining cached forecasts, an operation that is deterministic, fast, and—crucially—LLM-free, so that the tournament's "score" is an exact held-out error.

Two causal-safe base models developed in this work. Two members of the causal-safe pool were built as part of this project and merit description, both because they are the strongest causal-safe constituents the search draws on and because their results motivate the search itself.

The Flux-Decoupled Neural ODE (FD-NODE) pursues causal validity through architecture rather than through a soft penalty. Writing G for glucose, $c \geq 0$ for the carbohydrate

input and $u \geq 0$ for the insulin input, its state evolves as

$$\frac{dG}{dt} \;=\; f_{\text{mech}}(G, z_{\text{basal}}) \;+\; \phi_c(c) \;-\; \phi_u(u) \;+\; g_\theta(\mathbf{x}_{\text{ctx}}), \tag{3.1}$$

where f_{mech} is a UVA/PADOVA-style mechanistic term (Man et al., 2014) driven only by basal state, ϕ_c and ϕ_u are the decoupled treatment "flux" channels, and g_θ is a neural correction. Two design choices give the counterfactual guarantee. First, ϕ_c and ϕ_u are constrained to be monotonically non-decreasing (implemented with non-negative-weight monotone dense layers), so raising c can only raise dG/dt and raising u can only lower it. Second—an *intervention takeover*—the treatment inputs enter *only* through ϕ_c, ϕ_u: the mechanistic term sees basal state alone and the correction g_θ sees only the pre-intervention context $\mathbf{x}_{\text{ctx}}$, never c, u, or any treatment-derived state. Consequently the sign of the model's response to a carbohydrate or insulin intervention is fixed by the monotone channels regardless of the fitted weights, so FD-NODE is designed to enforce the required monotonic response *by construction* and is observed to achieve zero error on the intervention benchmark (no causal-loss weight to tune). Empirically it attains a 30-minute RMSE of 17.92 mg/dL and overall RMSE 12.19 mg/dL at 0/0 on the benchmark (20,225 parameters). Measuring the accuracy *cost of causal safety* against the causal-unconstrained hybrid baseline (30-minute RMSE 17.18), FD-NODE's cost is $17.92 - 17.18 = 0.74$ mg/dL, versus $18.25 - 17.18 = 1.07$ mg/dL for H^{2}NCM's soft causal loss—about a third lower—evidence that a hard architectural guarantee can be cheaper than a soft one. Architecture and training configuration are summarized in Appendix A.

Recipe-optimized H^{2}NCM (RO-H^{2}NCM) leaves the H^{2}NCM architecture unchanged but improves how it is trained. An LLM code-editing agent (Claude Opus), constrained to edit only the training loop, searched the training recipe and found a configuration—smaller batches, mild weight decay, a denser sampling stride, and a longer schedule—that lowers 30-minute RMSE to 17.15 mg/dL (overall RMSE 11.85) at 0/0 causal error, making RO-

Table 3.2: Representative members of the Anderson base-model pool (test set). Black-box models are accurate but causally unsafe; causal-safe models are correct but cluster tightly in accuracy. RO-H^{2}NCM is a recipe-optimized H^{2}NCM and the strongest single causal-safe model; FD-NODE is the author's hard-architecture causal model (Section 3.3).

Family	Model	Overall RMSE	PH30 RMSE	Causal (carb/ins)
Black-box	LSTM	11.54	16.43	0.59 / 1.00
	Transformer	11.38	16.56	0.68 / 1.00
	TCN	11.48	16.75	0.91 / 1.00
	Black-box NODE	11.63	16.49	0.99 / 1.00
Causal-safe	HybridNODE (α=0.1)	12.39	18.25	0.00 / 0.00
	FD-NODE	12.19	17.92	0.00 / 0.00
	H^{2}NCM (α=0.075)	11.96	17.44	0.00 / 0.00
	RO-H^{2}NCM	11.85	17.15	0.00 / 0.00

H^{2}NCM the strongest single causal-safe model in the pool. The same agentic search run with a weaker language model did not find the improvement, foreshadowing the value of capable agents in the multi-agent search of Chapter 4.

Table 3.2 lists the Anderson pool. Two facts drive everything downstream: black-box models own the low-error region but have insulin causal error near 1.0, whereas the causal-safe models—including FD-NODE and RO-H^{2}NCM—are both less accurate and tightly clustered (their 30-minute errors span barely one mg/dL and their causal errors are all ≈ 0).

From strong single models to a structured search. FD-NODE and RO-H^{2}NCM advance the causal-safe frontier about as far as careful single-model design reasonably can in this setting, yet a visible accuracy gap to the black-box ceiling remains, and Chapter 7 shows that pushing a single architecture further yields diminishing returns. This motivates a change of strategy: rather than engineer yet another individual model, we treat the pool as fixed raw material and ask whether a structured, multi-agent search over *combinations* of these models can raise causal-safe accuracy—and its ceiling—beyond what any single member attains. The remainder of the thesis develops and evaluates that search.

Chapter 4

Method: An Artificial Intelligence Co-Scientist for Grounded Stacking Search

4.1 Overview

The method is a hierarchical multi-agent tournament that searches the space of stacking strategies over the fixed model pool (Figure 4.1). It *adapts* the AI Co-Scientist (Gottweis et al., 2026) with five LLM agents plus two deterministic steps: a *supervisor* parses the goal and allocates compute; a *generation* agent proposes candidate strategies; a *reflection* agent peer-reviews them for soundness and for over-fitting or leakage risk; an *evolution* agent recombines and mutates the strongest candidates into new ones; and a *meta-review* agent synthesizes the round's reviews into feedback appended to every agent's context in the next round—feedback propagation without gradient descent. The original system's ranking and proximity stages are retained but implemented deterministically rather than as LLM agents: candidates are *ranked* by validation RMSE (a deterministic ordering, not an LLM Elo tournament), and near-duplicate genomes are removed by a code-based *proximity* check.

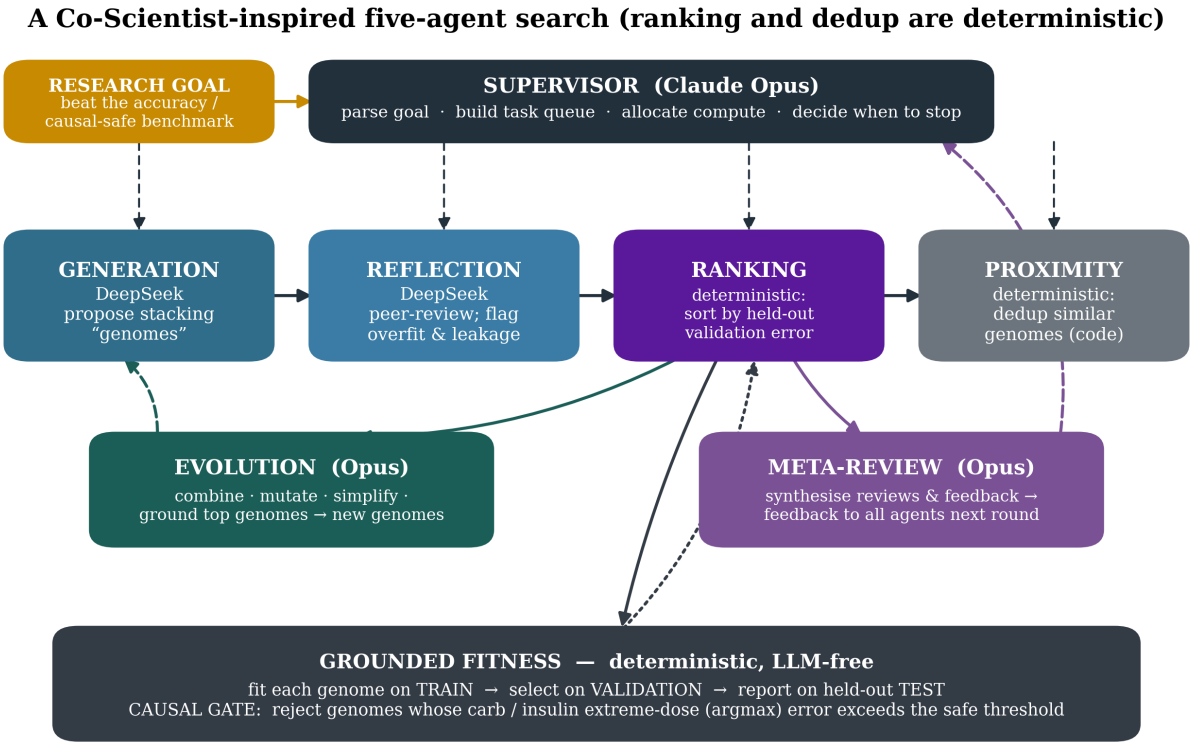

Figure 4.1: A Co-Scientist-inspired search (Gottweis et al., 2026) with *five* LLM agents (generation, reflection, evolution, meta-review, supervisor); the original system's Elo tournament and proximity agent are replaced by deterministic ranking-by-validation-error and code-based deduplication. The modification that makes the search deployable is the *grounded fitness* block: every strategy is fit on the training split, selected on validation, and evaluated on held-out test, and a *causal gate* rejects any strategy whose extreme-dose (argmax) intervention error exceeds the safe threshold.

Bulk roles (generation, reflection) run on a cost-efficient model (DeepSeek) and the high-value seats (supervisor, evolution, meta-review) on Claude Opus.

4.2 Strategies as safe, parameterized genomes

A candidate strategy—a "genome"—is a small JSON specification, not arbitrary LLM-written code. This choice is deliberate: it removes the security and robustness hazards of executing model-authored programs while still giving the agents a rich design space. The

grammar includes: `single` (use one base model), `mean` (unweighted average of a subset), `linear` (a ridge/non-negative least-squares combiner fit per horizon), `blend` (a weighted mixture of sub-genomes), and the causal-aware `vetoblend` described below. Genomes that fit parameters do so on the training split only; any residual-correction term is restricted to context-only features so that it *cannot* alter the counterfactual ordering, preserving causal safety by construction.

4.3 Grounded fitness and the causal gate

The single most important departure from the original system is that a genome's fitness is not an LLM judgment but a measured quantity. Each genome is fit on train and evaluated on validation; candidates are ranked deterministically by their validation error. Orthogonally, a *causal gate* recomputes each genome's extreme-dose (argmax) intervention error on the causal grid and rejects any genome whose error exceeds the safe threshold. Grounding plus the gate is what allows the loop to make measurable, monotone progress on the held-out validation objective rather than to converge on confident-but-wrong ideas.

Selection protocol, and a leakage caveat. Because grounding is central to the method's claim, we state the protocol—and its limitation—plainly. The *selection rule* uses validation only: the winning genome reported in Chapters 5–6 is the one with the best held-out validation error under the causal gate, and its weight is set by a line search on the validation intervention grid. This follows the stacked-generalization discipline of fitting and selecting a combiner on held-out data (Wolpert, 1992). However, we do *not* claim a fully test-blinded run: in the implementation, each round's leaderboard—which is passed into the generation, evolution, and meta-review agents—also contained the corresponding test metrics, so the search *path* (which genomes were proposed and evolved) had visibility into test, even though the final *selection* was made on validation. On Anderson, test metrics never entered the causal gate's threshold or the weight line search—the causal gate was evaluated on the

validation intervention grid only. The cross-cohort runs (Brown, Aleppo) are an exception we disclose: there the weight tool's causal gate also consulted the causal-*test* grid, so those replications are *exploratory* (Chapter 6). A fully test-blinded re-run—with all test arrays and metrics withheld from every agent until the winner is frozen—is the cleanest way to remove this residual leakage and is left as future work (Section 7.5).

4.4 The veto-blend and its weight-optimization tool

The causal gate makes a specific composition attractive, which the search discovers on its own. Consider a blend

$$\hat{y} \;=\; w_{\text{safe}} \cdot \underbrace{\text{mean}(\text{causal-safe models})}_{\text{preserves counterfactual order}} \;+\; (1 - w_{\text{safe}}) \cdot \underbrace{(\text{black-box model})}_{\text{sharpens the point forecast}} . \tag{4.1}$$

When w_{safe} is large enough, the causal-safe majority *dominates the counterfactual ranking*, so the blend inherits the majority's zero causal error; meanwhile the small black-box weight lowers RMSE. We call this a *veto-blend*: the safe majority "vetoes" the causal misbehavior of the black-box term while keeping its accuracy contribution (Figure 4.2). The corresponding genome, `vetoblend`, names the safe base and the black-box set; a specialized *tool*—in the sense of tool-using agents—then performs a one-dimensional line search for the largest black-box weight that the causal gate still admits, using only the validation forecasts and the validation intervention grid. Because it optimizes a single scalar under a hard constraint, it has substantially lower capacity—and therefore lower validation-overfitting risk—than an unconstrained high-dimensional weight vector, and it hands the tournament a concrete, safe blend to score.

Why this is more than reweighting. The veto-blend is not merely a tuned ensemble. It encodes a causal invariant—keep the safe majority in control of the ordering—and it uses black-box *diversity*, at individually small weights, precisely so that the black boxes' causal

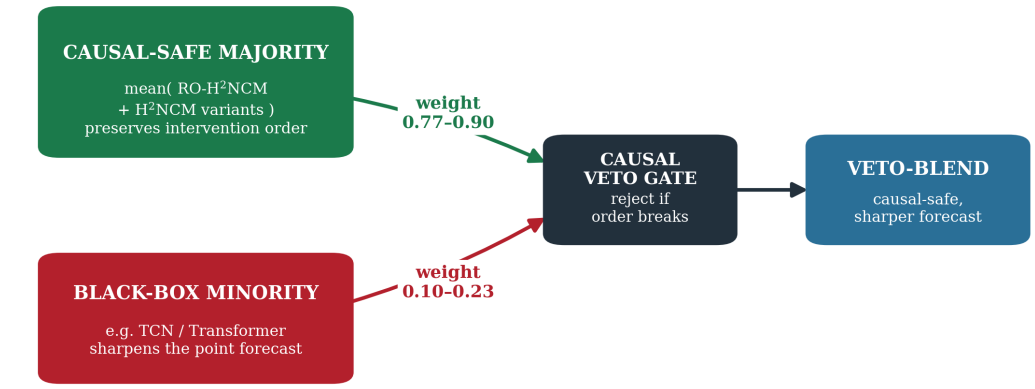

Figure 4.2: The veto-blend of Equation (4.1). A causal-safe majority carries a large weight and preserves the intervention ordering; a small weight on a black-box model (e.g. TCN or Transformer) sharpens the forecast. A causal gate rejects any weight at which the ordering breaks.

disturbances partially cancel while their accuracy adds. That is why, in Chapter 5, adding a second and third diverse black box under the veto keeps lowering error at a fixed zero-error benchmark operating point.

Chapter 5

Results on Anderson

We report two experiments on Anderson: an accuracy-only search that establishes the search can beat strong baselines, and the causal-safe search that is the thesis's main result.

5.1 Accuracy-only: the search beats best-single and equal-mean

When the causal gate is disabled and the objective is accuracy alone, the tournament (six rounds, LLM cost $\approx$ \$0.92) selects an unweighted mean of a small, diverse four-model subset (a Transformer, a black-box NODE, a TCN, and one H^{2}NCM variant). Table 5.1 and Figure 5.1 show that it beats both the best single model and the equal-weight mean of the full pool, and that validation error falls monotonically over rounds. The robust finding here is methodological: averaging a *small, diverse* subset beats both the best single model and the naive full-pool average, and the tournament's meta-review correctly articulated the rule ("lock the diverse core, prune the rest"). This winner is, however, *not* causal-safe—its black-box-heavy composition has insulin causal error ≈ 1.0—which is exactly why the causal gate of the next section is necessary.

Table 5.1: Accuracy-only search (Anderson test). The Co-Scientist selects a diverse four-model mean that improves on the best single model by 0.24 overall RMSE and 0.34 at 30 minutes. This operating point is not causal-safe and serves only to validate the search.

Strategy	Overall RMSE	PH30 RMSE
Equal-weight mean (all 22 pool models)	11.673	17.003
Best single model (Transformer)	11.381	16.559
Co-Scientist (diverse-4 mean)	**11.139**	**16.221**

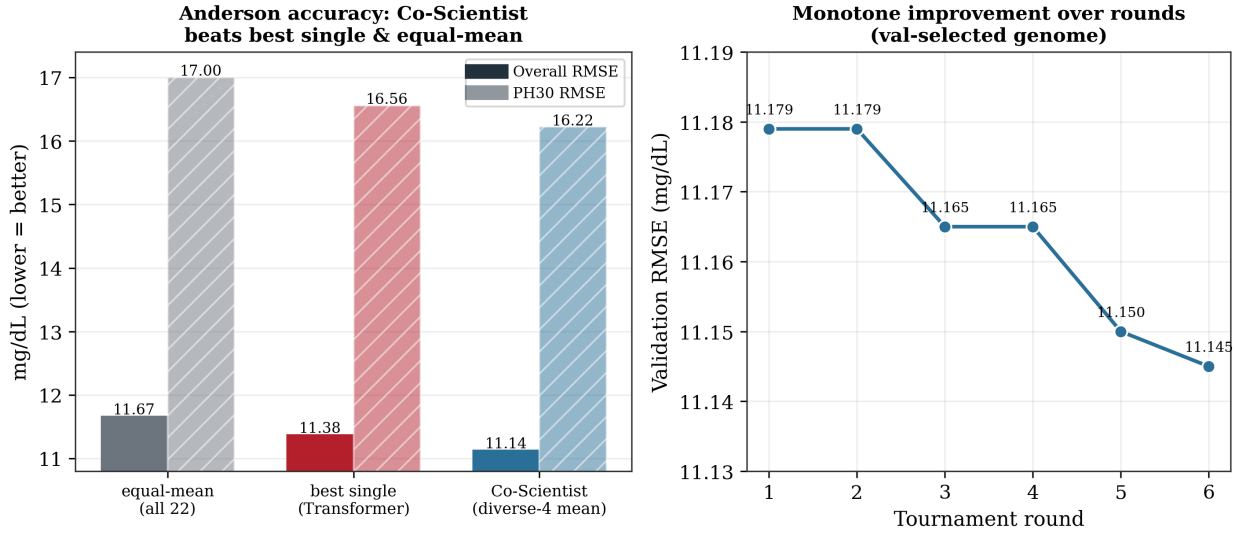

Figure 5.1: Left: the accuracy-only Co-Scientist beats the best single model and the equal-weight mean. Right: validation RMSE of the selected genome falls monotonically over the six tournament rounds.

5.2 The causal-safe frontier and the veto-blend

We now enable the causal gate and require zero observed causal error on the intervention benchmark. A pure causal-safe stack—an average of only causal-safe models—*saturates*: it ties the strongest single causal-safe model (RO-H^{2}NCM) at 17.27 vs. 17.15 PH30, because the causal-safe pool is collinear and averaging correlated models adds nothing. The veto-blend breaks the saturation. Adding one, then two, then three diverse black-box models under the veto, each at a small weight chosen by the line-search tool, lowers error monotonically while the causal gate holds the operating point at 0/0 (Figure 5.3). Allowed to choose the

Table 5.2: The causal-safe frontier on the Anderson intervention benchmark. The pure safe stack saturates; adding diverse black-box models under the veto closes most of the gap to the unsafe black-box ceiling while the causal gate holds the operating point at zero observed 80-context benchmark error. "bb" = black-box. Causal columns are argmax (extreme-dose) error on the 80-context benchmark.

Strategy	Overall RMSE	PH30 RMSE	Causal (carb/ins)
Safe-only stack (saturates)	11.858	17.267	0.00 / 0.00
RO-H^{2}NCM (strongest safe single)	11.853	17.146	0.00 / 0.00
+ 1 bb (w=0.10)	11.647	16.957	0.00 / 0.00
+ 2 bb (w=0.14)	11.589	16.829	0.00 / 0.00
+ 3 bb (w=0.17)	11.535	16.775	0.00 / 0.00
Veto-blend (search winner)	**11.473**	**16.751**	**0.00 / 0.00**
Transformer (unsafe accuracy ceiling)	11.381	16.559	0.68 / 1.00

composition itself, the tournament settles on

$$0.77 \cdot \text{mean}\Big(\text{RO-H}^2\text{NCM},\ \text{H}^2\text{NCM}(\alpha{=}0.075),\ \text{H}^2\text{NCM}(\alpha{=}0.060)\Big)\ +\ 0.23 \cdot \text{TCN},$$

a three-model causal-safe base plus one black-box term, with the black-box weight pinned at 0.23 by the line-search tool. Table 5.2 and Figure 5.2 report the frontier. The validation-selected veto-blend reaches PH30 = 16.75 and overall RMSE = 11.47 at zero observed causal error on the 80-context benchmark, improving on the strongest causal-safe single model by 0.38 overall RMSE and 0.40 at 30 minutes, and closing roughly **80% of the overall-RMSE gap and 67% of the 30-minute gap** to the causally unsafe black-box ceiling (the Transformer). One caveat attaches here: the strict 0/0 result holds on the 80-context intervention benchmark; a subsequent full-test audit over all 24,406 windows shows this veto-blend has 0.033% carbohydrate and 0.795% insulin error (the causal-safe base alone is near-clean at 0.016% / 0.012%), so the small black-box weight introduces a low but non-zero full-test causal error that the 80-context subset does not capture (Section 7.5).

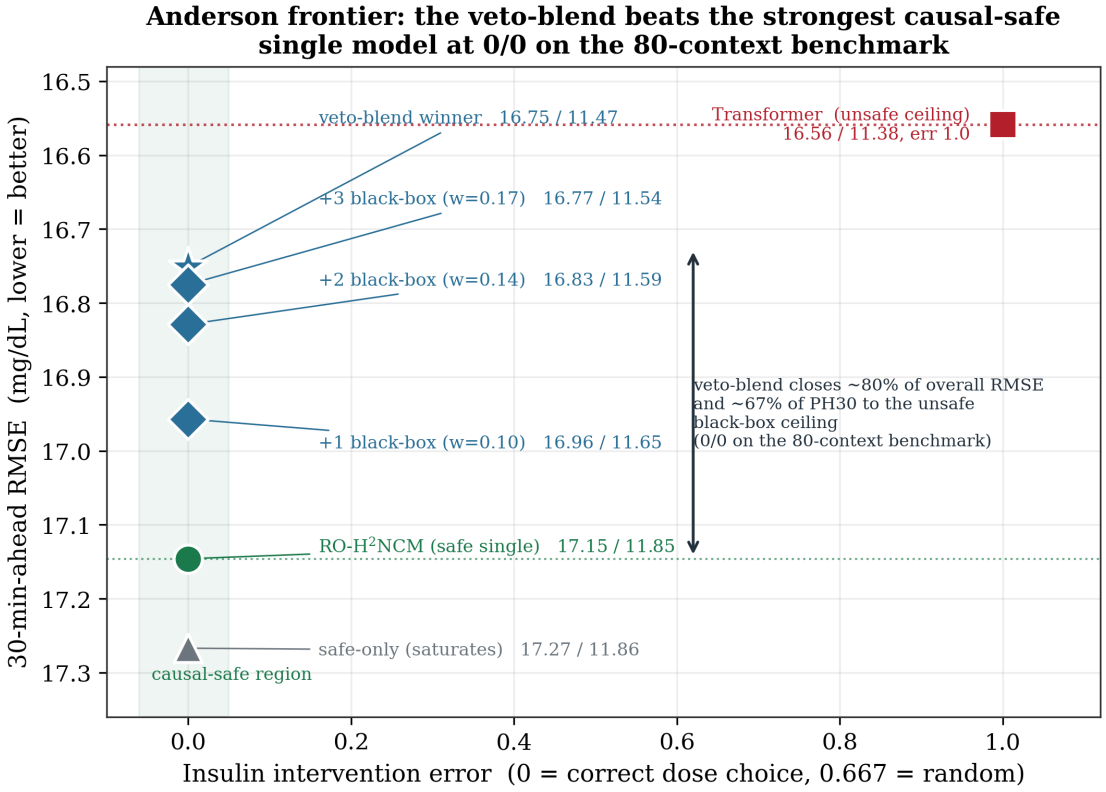

Figure 5.2: The veto-blend beats the strongest causal-safe single model at zero observed causal error on the 80-context benchmark. The veto-blend winner (blue star), the intermediate blends (blue diamonds), and the strongest safe single model (green circle) all sit at zero insulin argmax error; the unsafe black-box ceiling (red square) sits at error 1.0. The vertical arrow marks the gap the veto-blend closes.

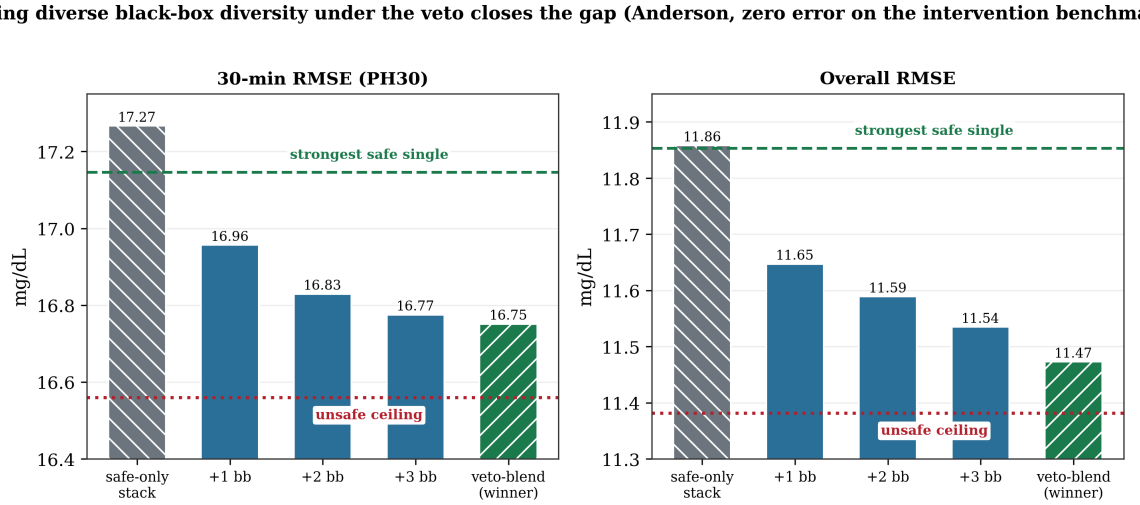

Figure 5.3: Adding diverse black-box models under the veto lowers both overall and 30-minute error toward the unsafe black-box ceiling (red dotted) while the operating point stays at zero observed causal error on the 80-context benchmark. The pure safe stack (grey) merely ties the strongest causal-safe single model (green dashed).

5.3 The structure owns both method and weights

Two features distinguish this result from a purely hand-tuned ensemble. First, the composition—which causal-safe models form the base and which black-box term to add—was chosen by the tournament under the validation-error ranking rather than hand-picked by us. Second, the blend weight was set not by us but by the line-search tool the agents invoke. In other words, the Co-Scientist-inspired structure owns both the *form* of the model and its *weight*, at an LLM cost under one dollar per run. This is the sense in which the search procedure, and not merely the veto-blend idea, is the contribution. (As noted in Section 5.2, the reported winner is the validation-selected genome; the search path's visibility into test is discussed in Sections 4.4 and 7.5.)

Chapter 6

Cross-Cohort Replication

A single-cohort result invites the worry that the mechanism is an artifact of Anderson. We therefore repeated the causal-safe search on Brown and Aleppo, holding architectures, parameter budgets, and batch sizes fixed and changing only the data. Because the cohorts have different channels, absolute error is not comparable across them; the question is purely whether the *mechanism* replicates—whether a veto-blend again reaches or beats the unsafe accuracy ceiling at zero observed causal error on each cohort's intervention benchmark. As on Anderson, these are single-seed runs and the selection caveat of Section 7.5 applies. One difference is important to disclose: unlike the Anderson run, whose causal gate used the validation grid only, the cross-cohort weight tool evaluated its causal gate on *both* the validation and the test intervention grids. The Brown and Aleppo zero-error results are therefore *exploratory* replications—the causal-test labels informed weight selection—rather than fully held-out tests, and should be read as consistency evidence pending a validation-only re-run.

It does (Table 6.1, Figure 6.1). On **Brown**, where the causal-safe RO-H^{2}NCM model trailed the black-box ceiling by a real margin (0.43 overall RMSE), the veto-blend $0.65 \cdot$ RO-H^{2}NCM $+ 0.35 \cdot$ TCN matches it in overall RMSE (11.01 vs. 11.02) at zero observed insulin argmax error (carbohydrate is undefined for Brown); at the 30-minute horizon it

Table 6.1: The veto-blend on Brown and Aleppo (test sets); "RO-H^{2}NCM $\rightarrow$ veto" shows the safe single model improving to the veto-blend, and "unsafe" is the strongest causally unsafe single model on that cohort. On Aleppo the veto-blend is marginally lower on both metrics (within single-seed noise); on Brown it matches the reference in overall RMSE but remains above it at 30 minutes (bold marks the better of RO-H^{2}NCM/veto-blend). Causal columns are argmax error on each cohort's benchmark and are zero on every reported axis; carbohydrate is undefined for Brown, shown "–". Values are single-seed and not comparable across cohorts (different channels).

Cohort	Overall RMSE		PH30 RMSE		Causal
	RO-H^{2}NCM $\rightarrow$ veto	unsafe	RO-H^{2}NCM $\rightarrow$ veto	unsafe	(veto)
Brown	11.44 $\rightarrow$ **11.01**	11.02	17.16 $\rightarrow$ **16.90**	16.67	– / 0
Aleppo	12.56 $\rightarrow$ **12.49**	12.55	18.72 $\rightarrow$ **18.61**	18.68	0 / 0

closes most of the gap but remains 0.23 mg/dL above the ceiling (16.90 vs. 16.67). This is the strongest "tradeoff-narrowed" case. On **Aleppo**, causal safety was already nearly free (RO-H^{2}NCM trailed the unsafe ceiling by only 0.01–0.04), and the veto-blend $0.65 \cdot \text{mean}(\text{RO-H}^2\text{NCM}, \text{HybridNODE}) + 0.35 \cdot \text{LSTM}$ is marginally lower than the unsafe LSTM ceiling on both metrics—by about 0.06–0.07 mg/dL, which is within single-seed variation and not established with confidence intervals—at zero observed argmax error (both carbohydrate and insulin defined here). The two cohorts thus bracket the two regimes the mechanism is designed for: when there is a real accuracy price for causal safety, the veto-blend pays most of it back; when there is little price to begin with, it holds the line without harm. Each run cost roughly \$0.6 in LLM calls.

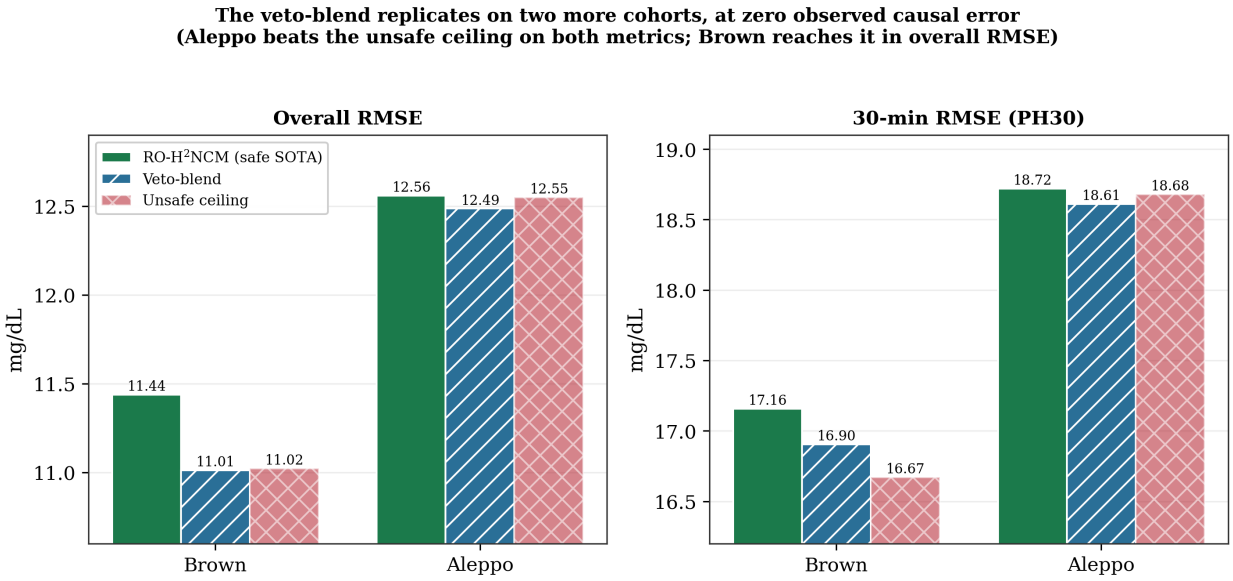

Figure 6.1: The veto-blend on two additional cohorts. On Aleppo the veto-blend (blue) is marginally lower than the unsafe ceiling (red) on both metrics—within single-seed variation; on Brown it matches that model in overall RMSE while remaining slightly above it at 30 minutes. In every case the veto-blend attains zero observed argmax causal error on the benchmark (carbohydrate is undefined for Brown). Bars are also distinguished by hatch pattern for grayscale readability.

Chapter 7

Discussion: The Price of Causality

The veto-blend closes most, but not all, of the gap to the unsafe ceiling. This chapter asks why the remainder does not close, and argues that the answer is informative rather than a failure of effort.

7.1 Why recombination saturates

The causal-safe members of the pool are strongly correlated: forcing several architectures to obey the same counterfactual constraints collapses them onto a common ∼17 mg/dL PH30 frontier with highly correlated errors. Ensemble theory says variance reduction requires diversity (Dietterich, 2000); correlated models supply none, so a pure safe stack cannot beat its best member. The veto-blend's gains come precisely from importing *outside* diversity (black-box models) in a controlled, causally vetoed dose. But that dose is bounded by the gate, which is why the frontier stops short of the black box: the black box's residual accuracy *is* its causal violation, and the gate forbids exactly the part that would close the last sliver.

7.2 Levers that did not move the frontier

We verified, one lever at a time, that the residual gap is not an artifact of insufficient search:

- **More recombination.** Input-conditioned dynamic gating, multi-LLM committees, and longer tournaments all tie the static veto-blend—the pool is collinear, so more search finds the same frontier.

- **More architecture / recipe tuning.** An LLM code-editing agent tuning the causal-safe base did not beat the hand-tuned recipe; single-model architecture and loss variants moved error by less than the seed-to-seed noise ($\sim$0.08 overall RMSE).

- **Post-hoc repair.** Forcing a correct counterfactual response onto a black box after training only degrades its accuracy, because that model's accuracy does not *rely* on the insulin signal it gets wrong.

Read together, these are single-configuration observations rather than a controlled ablation with confidence intervals; we present them as such. Within the evaluated pool and search budget, near-black-box accuracy appears difficult to reconcile with a zero-argmax-error operating point, and the residual increment appears to be the price of that operating point, concentrated at the longest horizon where the insulin effect is smallest and the constraint binds hardest. We state this as an empirical observation over the models, data, and budget studied here, not as a general impossibility; a formal ablation with per-lever configurations and patient-level uncertainty is future work.

7.3 The lever with evidence: data

Among the levers we evaluated, the one associated with a measurable frontier shift is data. A learning-curve study on Aleppo (fixing the test set and varying the number of training subjects from 25 to 144) shows 30-minute error still descending at the largest size tried, without a visible plateau. Anderson has only 29 subjects; whether it sits on the steep part of *its own* curve is suggestive rather than established, since the cohorts differ in channels, scale, and treatment regime and the learning curve was measured on Aleppo. Subject to

that caveat, the lever most associated with a measurable frontier shift in our experiments is more compatible four-channel T1D data plus a retraining of the causal-safe components, after which the veto-blend can be re-run—rather than more recombination, architecture, or search.

7.4 Two coexisting ceilings

It helps to separate two limits. *Subject count* still has value—error falls with more subjects—but it will not, by itself, let a causal-safe model beat the black box, because the black box benefits from data too; the best it buys is a tie at 0/0. *Information-per-window* (which channels are observed) plus aleatoric noise sets each cohort's asymptote; lowering that asymptote needs new input modalities (heart rate, activity, meal composition) that these cohorts lack. Within the current pool and information budget, the veto-blend is the best causal-safe operating point we found.

7.5 Limitations

Several limitations bound the claims, and we state them plainly.

The causal metric is an argmax check, not a full ranking. As implemented, "causal error" is the fraction of contexts on which the model's *highest-response* dose is not the true highest-response dose (Section 2.2). It therefore verifies the extreme-dose choice, not the full pairwise monotonicity of the three-dose response; a model that ranks the middle dose incorrectly can still score zero. A strict pairwise-monotonicity metric would be a stronger test and is the appropriate next refinement.

Zero causal error is a benchmark-subset result; the full-test audit is not zero. The strict 0/0 operating points are on the sampled intervention benchmark (80 contexts on Anderson, 120 on each cohort). On a subsequent audit over all 24,406 Anderson windows, the reported veto-blend has 0.033% carbohydrate and 0.795% insulin argmax error; the

causal-safe base alone is near-clean (0.016% / 0.012%), so the small black-box weight is what introduces the non-zero full-test error. Claims of "causal-safe" and "zero causal error" should be read as "zero on the evaluated benchmark subset," not window-for-window on the full test set. Bringing the full-test error to zero (for example with a per-window constraint on the black-box weight) is a natural extension we do not pursue here.

Selection used validation, but the search path saw test. As disclosed in Section 4.4, each round's leaderboard passed to the agents also contained test metrics, so although the reported winner is the validation-selected genome, the sequence of proposed genomes was not fully blinded to test. On Anderson the causal gate itself used the validation grid only; the cross-cohort weight tool, however, also consulted the causal-test grid, so the Brown and Aleppo zero-error results are exploratory replications (Chapter 6). A clean, fully test-blinded re-run is the way to remove this residual leakage and to let the reported test numbers stand as a single independent evaluation.

"Causal-safe" is not clinical safety. The property verified here is intervention-direction consistency on a benchmark, not a guarantee of safe dosing: effect *magnitude*, hypoglycemia risk, and individual (as opposed to population-average) treatment effects are not evaluated.

Statistical inference and single seeds. Test cohorts contain only 6/15/44 subjects and the windows within a subject are highly correlated, so the appropriate inference is a patient-level bootstrap rather than treating windows as independent; the point estimates here are single-seed and are not accompanied by confidence intervals. The small cross-cohort margins (e.g. Aleppo's $\sim$0.06–0.07 mg/dL) should be read with this in mind. Multi-seed runs with patient bootstrap confidence intervals would put the accuracy and margin claims on firmer footing.

Cohort confounds and comparisons. Brown's insulin axis is largely closed-loop-confounded, so its causal numbers are consistency evidence rather than an independent open-loop test.

Chapter 8

Conclusion

This paper aims to resolve a contradiction that is easy to describe but difficult to resolve: in glucose prediction, the models with the highest accuracy are not trustworthy at the same time under intervention conditions. The method to solve this problem is not to propose a single new model or structure; rather, we build a multi-agent searching system based on real evaluation results.

By applying AI Co-Scientist (Gottweis et al., 2026) to this mission, replacing free-form scientific hypotheses with stacked strategies, and scoring these strategies using real holdout-set errors under causal gating constraints, the search process discovered a "veto-blend" architecture. This architecture is formed mostly with causal-safe models, allowing the incorrect causal decisions made by a small number of various black-box models to be "vetoed," while retaining the latter's accuracy advantages.

In the Anderson cohort, the veto-blend surpassed the strongest causal-safe single model in our pool, shrank the gap to the causal safety performance ceiling, while ensuring zero observed causal error on the intervention benchmark. This mechanism has also been successfully applied to the Brown and Aleppo cohorts. Subsequently, through a series of experiments that failed to further improve the results, cutting-edge analysis indicates that, given the currently evaluated model pool and experimental protocols, the remaining gap appears

to be the cost required to achieve causal safety. The analysis also pointed out that better-matched four-channel data is the factor most closely associated with measurable shifts in the performance frontier.

There are two threads most valuable for moving forward. The first direction is the data. Integrate additional Type 1 diabetes cohorts containing four channels, retrain the causal-safety model, and run the veto-blend again. The evidence of the learning curve shows that, compared with any recombination of current models, it is most likely to push the performance frontier by adding more T1D data with four channels.

The second direction is the evaluation. Conduct a comprehensive audit of the model's causal safety for each test window and report confidence intervals under multiple random seeds, thereby replacing the current evidence based on a sampled intervention grid with evidence covering the entire test set.

Besides glucose prediction, this real-evaluation-results-based Co-Scientist method has wide application potential even in other fields. In this study, multiple agents discuss and propose different model pairs, and the adaptability of each pair is determined by the set-aside metric measured under domain constraints. Therefore, it is expected to excel on other similar missions that ensure continuous improvement in prediction accuracy and maintain certain kinds of strict correctness properties.

Appendix A

Reproducibility Details

This appendix records the settings needed to reproduce the results. The base-model checkpoints, the cached forecast arrays, the agent prompt templates, and the figure-generation code are retained by the author and are available on request.

A.1 Data and splits

All three cohorts use a patient-level, out-of-distribution split with fixed seed 42; per-cohort subject and window counts are in Table 3.1. Each window is a 4-hour context (48 five-minute steps) predicting 30 minutes ahead (6 steps). Windows are generated at stride 6 (full, non-overlapping coverage); the recipe-optimized H^2NCM uses training stride 4. Channels: Anderson = [glucose, basal, bolus, carbohydrate]; Brown = [glucose, basal, bolus]; Aleppo = [glucose, bolus, carbohydrate]. Inputs are normalized with per-cohort training statistics.

A.2 Base models and training

All base models respect the <25K-parameter budget. Training uses AdamW; for the H^2NCM family the mechanistic parameters use learning rate 10^{-1} and all other parameters 2×10^{-3}; benchmarks use batch size 128 and the H^2NCM family 64. The causal loss is $L = (1 -$

$\alpha)\,\mathrm{MSE}_z + \alpha\,\mathrm{CE}_{\mathrm{causal}}$ with softmax temperature $\varphi = 1.0$. Key per-model settings are in Table A.1.

Table A.1: Base-model configurations. All models are under 25K parameters. The same architectures and parameter counts are used for Brown and Aleppo; only the input-channel dimension changes with the cohort's available channels (Section 3).

Model	Params	Epochs	Key settings
LSTM (h=48)	20,593	25	batch 128
Transformer (d=32)	21,697	25	batch 128
TCN (c=24)	20,782	25	batch 128
Black-box NODE (z=24)	14,921	25	batch 128
HybridNODE (α=0.1)	11,201	20	causal loss, batch 64
H^{2}NCM (α=0.075)	11,201	20	causal loss, batch 64, stride 6
RO-H^{2}NCM	11,201	30	batch 64, weight decay 10^{-5}, stride 4
FD-NODE	20,225	30†	monotone flux channels + intervention takeover; synthetic UVA/PADOVA pretrain, then finetune; best-validation checkpoint retained

† Finetuning epochs, preceded by a synthetic-data (UVA/PADOVA) pretraining phase; the best-validation checkpoint is retained rather than training to a fixed epoch.

A.3 Causal intervention grid

Counterfactual doses follow the H^{2}NCM protocol: carbohydrate $\in \{0, 50, 100\}$ g (true order: more carbohydrate $\rightarrow$ higher glucose) and insulin $\in \{0, 2.5, 5.0\}$ U (true order: more insulin $\rightarrow$ lower glucose). Causal error is the fraction of contexts whose predicted ordering is incorrect. The benchmark grid is sampled into a validation grid (used for selection and the causal gate) and a disjoint test grid (used once for reporting); the Anderson grids contain 80 contexts each and the Brown/Aleppo grids 120 each. Section 7.5 notes the small-sample caveat this implies.

A.4 The Co-Scientist search

The generation and reflection agents run on DeepSeek models; the high-value seats (supervisor, evolution, meta-review) run on Claude Opus. Ranking (by validation RMSE) and proximity deduplication are deterministic code steps, not agents. The accuracy-only run used six rounds; the causal-safe run used four rounds (validation error 11.419 $\rightarrow$ 11.351, then plateau). The genome grammar is `single`, `mean`, `linear`, `blend`, and `vetoblend`; the `vetoblend` weight is set by a one-dimensional line search over the global black-box weight under the causal gate, evaluated on the validation splits only. Sampling temperature is left at each provider's default; because the winning genome is chosen by the deterministic grounded fitness and causal gate rather than by the agents' sampled text, the outcome does not hinge on the sampling temperature, though the search *path* does—each setting was run at a single tournament seed (Section 7.5). Reported LLM cost is provider token pricing times tokens consumed: about $0.92 for the Anderson accuracy-only run and $0.70 for the Anderson causal-safe run, and roughly $0.6 per additional cohort.

A.5 Software and hardware

Models are implemented in PyTorch; stacking genomes are fit deterministically with NumPy/scikit-learn. Anderson models were trained on CPU (with `OMP_NUM_THREADS=1` to avoid a platform-specific segfault) and the larger Brown/Aleppo campaigns on a single GPU. All figures are generated by a matplotlib script from the recorded result files.

A.6 Reproducibility caveats

Two aspects limit exact reproducibility and are noted for transparency. First, the LLM agents call external providers at each provider's default sampling temperature; identical re-runs are therefore not bit-for-bit reproducible, and the search *path* can vary run to run

even though the grounded fitness and causal gate are deterministic. Exact library versions, hardware, random seeds, and verbatim prompt templates are recorded in the author's project files rather than inline, and are available on request. Second, as noted in Sections 4.4 and 7.5, the round leaderboards shown to the agents included test metrics; a fully test-blinded re-run is the recommended way to reproduce the reported numbers as a single independent evaluation. For statistical inference we recommend a patient-level bootstrap (resampling test *subjects*, not windows) with multiple training seeds, given the small number of test subjects and the strong within-subject correlation of windows.

Bibliography

Tadej Battelino, Thomas Danne, Richard M. Bergenstal, Stephanie A. Amiel, Roy Beck, Torben Biester, et al. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, 42(8):1593–1603, 2019. doi: 10.2337/dci19-0028.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996. doi: 10.1007/BF00117832.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 6572–6583. Curran Associates, Inc., 2018.

William L. Clarke, Daniel Cox, Linda A. Gonder-Frederick, William Carter, and Stephen L. Pohl. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, 10(5):622–628, 1987. doi: 10.2337/diacare.10.5.622.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS)*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL `https://www.nature.com/articles/s41586-026-10644-y`.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Kezhi Li, Chengyuan Liu, Taiyu Zhu, Pau Herrero, and Pantelis Georgiou. GluNet: A deep learning framework for accurate glucose forecasting. *IEEE Journal of Biomedical and Health Informatics*, 24(2):414–423, 2020. doi: 10.1109/JBHI.2019.2931842.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal*

of Diabetes Science and Technology, 8(1):26–34, 2014. doi: 10.1177/1932296813514502.

John Martinsson, Alexander Schliep, Björn Eliasson, and Olof Mogren. Blood glucose prediction with variance estimation using recurrent neural networks. *Journal of Healthcare Informatics Research*, 4(1):1–18, 2020. doi: 10.1007/s41666-019-00059-y.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 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.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025. doi: 10.1038/s41586-025-08661-4.

Bob Junyi Zou, Matthew E. Levine, Dessi P. Zaharieva, Ramesh Johari, and Emily B. Fox. Hybrid^2 neural ODE causal modeling and an application to glycemic response. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 62934–62963. PMLR, 2024.