Automatic and Structure-Aware Sparsification of Hybrid Neural ODEs with Application to Glucose Prediction

ICLR 2026 Poster Presentation

Bob Junyi Zou[1] Lu Tian[2]

[1]Institute for Computational and Mathematical Engineering, Stanford University
[2]Department of Biomedical Data Science, Stanford University

May 2026

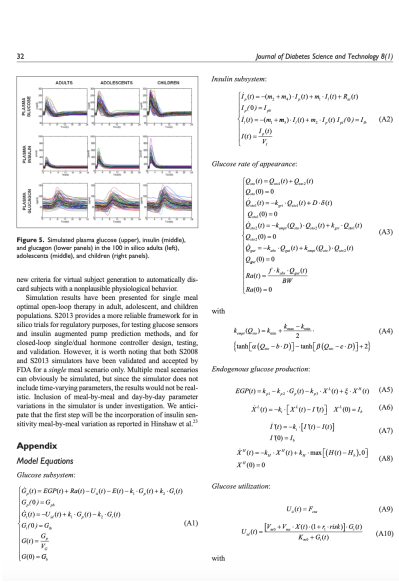
The Challenge: Hybrid Models are Powerful but "Bloated"

The Promise of Hybrid Modeling

- Combines mechanistic priors (ODEs) with neural flexibility.
- Essential for data-scarce healthcare settings (e.g., T1D glucose forecasting).

The Reality Gap

- **Over-parameterization:** Mechanistic models are designed for biological completeness, not predictive efficiency.
- **Redundancy:** A glucose model might have 20+ latent states but only 4 observables.
- **Instability:** Cycles and feedback loops lead to stiffness and training failures.



32 Journal of Diabetes Science and Technology 8(1)

Figure 5. Simulated plasma glucose (upper), insulin (middle), and glucagon (lower panels) in the 100 in silico adults (left), adolescents (middle), and children (right panels).

new criteria for virtual subject generation to automatically discard subjects with a nonplausible physiological behavior.

Simulation results have been presented for single meal optimal open-loop therapy in adult, adolescent, and children populations. S2013 provides a more reliable framework for in silico trials for regulatory purposes, for testing glucose sensors and insulin augmented pump prediction methods, and for closed-loop single/dual hormone controller design, testing, and validation. However, it is worth noting that both S2008 and S2013 simulators have been validated and accepted by FDA for a *single* meal scenario only. Multiple meal scenarios can obviously be simulated, but since the simulator does not include time-varying parameters, the results would not be realistic. Inclusion of meal-by-meal and day-by-day parameter variations in the simulator is under investigation. We anticipate that the first step will be the incorporation of insulin sensitivity meal-by-meal variation as reported in Hinshaw et al.[23]

Appendix

Model Equations

Glucose subsystem:

$$\begin{cases} \dot{G}_p(t) = EGP(t) + Ra(t) - U_{ii}(t) - E(t) - k_1 \cdot G_p(t) + k_2 \cdot G_t(t) \\ G_p(0) = G_{p,b} \\ \dot{G}_t(t) = -U_{id}(t) + k_1 \cdot G_p(t) - k_2 \cdot G_t(t) \\ G_t(0) = G_{t,b} \\ G(t) = \dfrac{G_p}{V_G} \\ G(0) = G_b \end{cases} \quad \text{(A1)}$$

Insulin subsystem:

$$\begin{cases} \dot{I}_p(t) = -(\mathbf{m}_1 + \mathbf{m}_3) \cdot I_p(t) + \mathbf{m}_1 \cdot I_l(t) + R_a(t) \\ I_p(0) = I_{p,b} \\ \dot{I}_l(t) = -(\mathbf{m}_1 + \mathbf{m}_2) \cdot I_l(t) + \mathbf{m}_1 \cdot I_p(t) \; I_{p,l}(0) = I_{l,b} \\ I(t) = \dfrac{I_p(t)}{V_I} \end{cases} \quad \text{(A2)}$$

Glucose rate of appearance:

$$\begin{cases} \dot{Q}_{sto}(t) = Q_{sto1}(t) + Q_{sto2}(t) \\ Q_{sto}(0) = 0 \\ \dot{Q}_{sto1}(t) = -k_{gri} \cdot Q_{sto1}(t) + D \cdot \delta(t) \\ Q_{sto1}(0) = 0 \\ \dot{Q}_{sto1}(t) = -k_{empt}(Q_{sto}) \cdot Q_{sto1}(t) + k_{gri} \cdot Q_{sto1}(t) \\ Q_{sto2}(0) = 0 \\ \dot{Q}_{gut} = -k_{abs} \cdot Q_{gut}(t) + k_{empt}(Q_{sto}) \cdot Q_{sto1}(t) \\ Q_{gut}(0) = 0 \\ Ra(t) = \dfrac{f \cdot k_{abs} \cdot Q_{gut}(t)}{BW} \\ Ra(0) = 0 \end{cases} \quad \text{(A3)}$$

with

$$k_{empt}(Q_{sto}) = k_{min} + \frac{k_{max} - k_{min}}{2} \cdot \left[\tanh\left\{\alpha\left(Q_{sto} - b \cdot D\right)\right\} - \tanh\left\{\beta\left(Q_{sto} - c \cdot D\right)\right\} + 2\right] \quad \text{(A4)}$$

Endogenous glucose production:

$$EGP(t) = k_{p1} - k_{p2} \cdot G_p(t) - k_{p3} \cdot X^*(t) + \xi \cdot X^*(t) \quad \text{(A5)}$$

$$\dot{X}^*(t) = -k_i \cdot \left[X^*(t) - I'(t)\right] \quad X^*(0) = I_b \quad \text{(A6)}$$

$$I'(t) = k_i \cdot \left[I'(t) - I(t)\right] \quad I'(0) = I_b \quad \text{(A7)}$$

$$\dot{X}^{**}(t) = -k_d \cdot X^{**}(t) + k_a \cdot \max\left[\left(H(t) - H_b\right), 0\right] \quad X^{**}(0) = 0 \quad \text{(A8)}$$

Glucose utilization:

$$U_{ii}(t) = F_{cns} \quad \text{(A9)}$$

$$U_{id}(t) = \frac{\left[V_{m0} + V_{mx} \cdot X(t) \cdot (1 + r_1 \cdot risk)\right] \cdot G_t(t)}{K_{m0} + G_t(t)} \quad \text{(A10)}$$

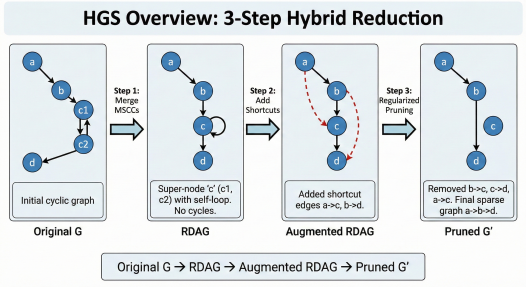
with

Why not use standard pruning?

- **Stability:** Feedback loops can cause training instabilities.
- **Timescale Separation:** Standard pruning doesn't account for "fast" variables that can be effectively skipped.
- **Gradient-Free Methods:** Greedy search or genetic algorithms are too slow for training complex ODEs.
- **Pure Data-Driven GNN Pruning:** Ignores physical laws; risks breaking reachability or creating nonsensical pathways.

Our Solution: Hybrid Graph Sparsification (HGS)
A gradient-based, structure-aware pipeline.

Methodology: The HGS Pipeline



1. **Merge Cycles (Stability):** Collapse Maximal Strongly Connected Components (MSCCs) into super-nodes.
2. **Add Shortcuts (Expressivity):** Augment pathways via partial transitive closure to model timescale separation.
3. **Regularized Pruning (Sparsity):** Train with L_1 penalty on edge weights to select the best subgraph.

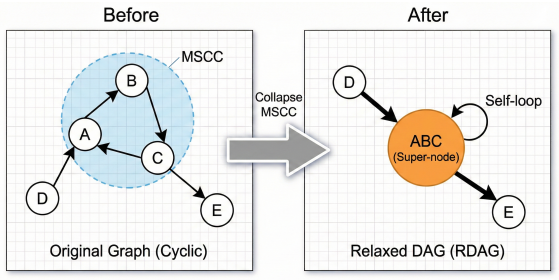
Step 1: Merging MSCCs for Stability

Problem:

- Physiological models are full of feedback loops.
- Cycles causes stiffness and exploding gradients.

HGS Action:

- Collapse cycles into “super-nodes”.
- **Result:** A Relaxed DAG (RDAG) that preserves the high-level mechanistic causal structure. System Jacobian becomes block upper-triangular.



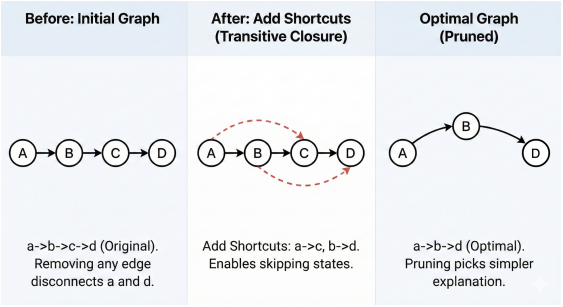
Step 2: Mechanistic Shortcuts

Intuition:

- In biology, A $\rightarrow$ B $\rightarrow$ C might happen so fast that A $\rightarrow$ C is a better predictive model (Timescale Separation).
- Standard pruning can only *remove* edges, not *skip* steps.

HGS Action:

- Transitive Closure: Add shortcut edges $(u \rightarrow v)$ if u is an ancestor of v.
- Allows the optimizer to "bypass" redundant latent states without violating mechanistic reachability.

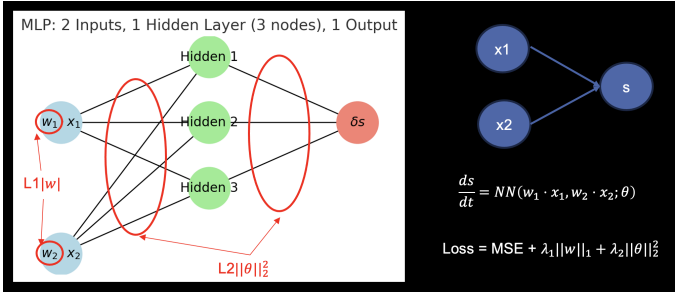


Step 3: Regularized Pruning

We learn a sparse subgraph using a composite loss function:

$$\mathcal{L} = \text{MSE} + \lambda_1 \sum_{(u,v)} |w_{u,v}| + \lambda_2 ||\Theta||_2^2$$

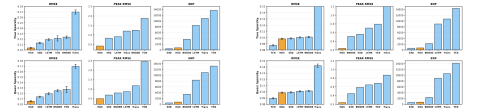
- L_1 **on Edge Weights (w):** Encourages sparsity.
- L_2 **on Parameters (Θ):** Ensures stability.
- **Equivalence:** Equivalent to a first-layer Group LASSO.



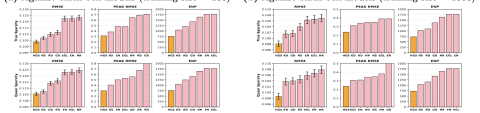
Results: Synthetic Experiments

Setup: Ground truth sparse ODE hidden inside a redundant comprehensive graph.

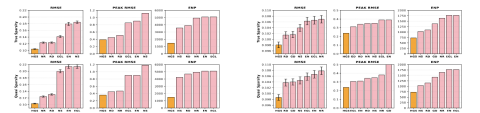
- **HGS vs. Black Box:** Outperforms LSTM/TCN in low-data regimes.
- **HGS vs. Other Pruning:** Achieves lower RMSE and better sparsity (ENP) than NeuralSparse or Group LASSO.



(a) Against black-box models (training size = 100) (b) Against black-box models (training size = 1000)



(c) Against other reduction methods (refined initial graph, training size = 100) (d) Against other reduction methods (refined initial graph, training size = 1000)



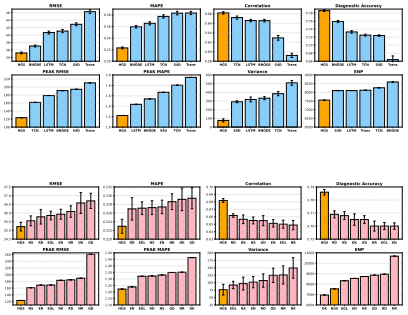
(e) Against other reduction methods (comprehensive initial graph, training size = 100) (f) Against other reduction methods (comprehensive initial graph, training size = 1000)

Application: T1D Glucose Forecasting

Data: T1D Exercise Initiative (T1DEXI), 342 time series from 105 patients.

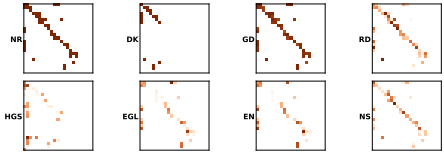
Performance:

- Surpasses Black-box Neural ODEs and standard mechanistic models.
- **Robustness:** Significantly lower Peak RMSE (worst-case error).



Interpretability:

- HGS pruned glucagon feedback loops.
- Suggests impaired glucagon response during exercise-induced hypoglycemia.



Conclusion

Summary

We propose **HGS**, a three-step pipeline for Hybrid Neural ODEs:

1. **Structure-Aware:** Respects mechanistic plausibility.
2. **Efficient:** Gradient-based (no expensive greedy search).
3. **Effective:** Improves accuracy and robustness in medical time-series.

Acknowledgment

This publication is based on research using data from Jaeb Center for Health Research Foundation that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication. We thank Alex Wang, Dessi Zeharieva, Emily Fox, Matthew Levine and Ramesh Johari for providing assistance with data acquisition and preliminary discussions on T1D and model reduction.

Paper available at the ICLR poster session.