

Structural Identifiability and Gradient Bounds for Hybrid-Activation Learning of Fractional HIV Dynamics

Okeke Ikenna Stephen¹

¹Industrial Mathematics and Health Statistics,, David Umahi Federal University of Health Sciences, Uburu, Nigeria, okekesi@dufuhs.edu.ng

Abstract.

The recovery of biological parameters from fractional HIV trajectories is examined alongside the mathematical properties of hybrid neural activations. Using the model of Okeke et al. [1] and the five-component activation mixture of Essang et al. [2], an exact parameter symmetry is derived: proliferation rate, healthy-cell death rate and carrying capacity cannot all be recovered separately from the model trajectories without additional information. A reduced parametrisation eliminates this redundancy. Conditional identifiability of the reduced coefficients is established through full-rank trajectory regressors for a known fractional order and fully observed states. The hybrid activation derivative is corrected, and global scalar slope bounds are proved. These bounds do not guarantee preservation of gradients through arbitrary network depth. A fractional residual learning framework is specified, and completed synthetic experiments examine parameter symmetry, regressor singular values, observation-window conditioning and slope products. The study establishes mathematical and computational limitations rather than empirical superiority of a trained learning algorithm.

Keywords: structural identifiability; fractional HIV model; hybrid activation; inverse problem; slope bounds; physics-informed neural networks.

1. Introduction

Fractional HIV modelling [1] and hybrid activation design [2] address different parts of a possible data-driven modelling pipeline. The former specifies biological state dynamics; the latter modifies the nonlinearities used by a neural estimator. Combining them raises two questions: which model parameters can the observations determine, and what does a positive activation slope actually guarantee?

Fractional physics-informed neural networks already address forward and inverse differential problems [5], while adaptive activations have been developed for physics-informed learning [4]. Consequently, neither the broad combination of fractional equations and networks nor the idea of learning activation behaviour is new. The present study focuses on an explicit symmetry in the specified HIV model, a reduced inverse formulation, and transparent slope bounds for the particular five-function mixture.

2. Related work and the scope of the inverse problem

The distinction between fitting a trajectory and determining its generating parameters is central to this study. Bellman and Åström [19] introduced structural identifiability as a property of the model and observation map. Raue et al. [18] develop profile-likelihood methods for examining structural and practical limitations in partially observed systems. Eisenberg and Hayashi [20] explain how

identifiable combinations can support reparameterisation when individual coefficients cannot be determined. I apply this perspective to the explicit proliferation term in the supplied fractional HIV model [1].

Within-host HIV modelling has a substantial biological and mathematical history. Ho et al. [12] and Perelson et al. [11] examined rapid turnover and viral dynamics, and the review by Perelson and Nelson [13] places such measurements in a modelling framework. These sources motivate state variables and time scales, but they do not justify recovering every coefficient of an arbitrarily enlarged model from sparse measurements. Fractional HIV models were studied by Arafa et al. [23], and the supplied model [1] is one particular logistic formulation. I make no claim that fractional HIV estimation is a new research area.

For fractional initial-value problems, Diethelm [6] gives the relevant mathematical framework. Numerical approaches include the predictor-corrector construction of Diethelm et al. [7], the methods reviewed by Garrappa [9], and L1-type approximations whose behaviour with nonsmooth data is analysed by Jin et al. [8]. Cong et al. [3] provide a local stability result for nonlinear fractional systems. Stability is relevant to long observation records because near-equilibrium trajectories can have weak excitation even when the underlying equilibrium is mathematically well behaved. The threshold theory in [10] supplies related context, but a stability theorem does not imply an identifiability theorem.

Neural inverse methods form a second line of literature. Raissi et al. [21] developed physics-informed neural networks for differential problems; Pang et al. [5] incorporated fractional operators. Adaptive activations in physics-informed problems were studied by Jagtap et al. [4]. The hybrid mixture used here is attributed to Essang et al. [2] and is interpreted within the wider trainable-activation taxonomy of Apicella et al. [22]. Rectifier and exponential components follow established activation families [15,16]. Glorot and Bengio [14] emphasise the relation between weight scaling, saturation and training difficulty. Adam [17] is one possible optimiser for a future learned implementation, but optimiser selection cannot remove an exact symmetry of the equations.

The contribution of this manuscript is consequently limited and testable. I derive the particular parameter equivalence class of the supplied model, state sufficient regressor conditions for identifying its reduced coefficients, and relate scalar activation bounds to the conditioning of a complete network. The numerical experiments evaluate symmetry, slope behaviour and trajectory excitation. They do not constitute a trained fractional physics-informed network benchmark. Keeping that distinction explicit makes the evidence reproducible and prevents a theoretical possibility from being reported as measured predictive performance.

3. Fractional model and observation assumptions

Work with nondimensional variables, a common known order $0 < q \leq 1$ and positive parameters:

$${}^C D_t^q T = s + (\alpha - d)T - \frac{\alpha}{K}T^2 - \beta TV, \quad {}^C D_t^q I = \beta TV - \delta I, \quad {}^C D_t^q V = \rho I - cV. \quad (1)$$

The Caputo derivative is used. Structural results concern ideal continuous trajectories and exact derivatives, not noisy finite samples. Fully observed T, I and V are assumed unless otherwise specified. Identifiability under these favourable assumptions is a prerequisite for meaningful unconstrained parameter estimation.

4. Exact nonidentifiability of the original parametrisation

Theorem 1. The triple (α, d, K) is structurally nonidentifiable from model trajectories, even when all three state variables are observed exactly and every other parameter and q are known.

Proof. Choose any h with $\alpha + h > 0$ and $d + h > 0$, and set

$$\tilde{\alpha} = \alpha + h, \quad \tilde{d} = d + h, \quad \tilde{K} = K \frac{\alpha + h}{\alpha}. \quad (2)$$

Then $\alpha \sim d \sim \alpha - d$ and $\alpha \sim K \sim \alpha/K$. Equation (1) is unchanged at every state. The initial value problem is identical, so the trajectories are identical wherever its solution exists. An open interval of h around zero yields infinitely many admissible triples. $\square$

This obstruction is independent of neural architecture, sample density and optimisation accuracy. A small residual cannot resolve parameters that produce exactly the same differential equations.

Exact example. The triples (6.8,0.01,1300) and (7.8,1.01,1491.1764705882354) share $\alpha - d = 6.79$ and $\alpha/K = 6.8/1300$. With all remaining coefficients unchanged, they generate the same trajectory for any common q and the same initial conditions. No simulation is needed to establish this equality.

5. Reduced parametrisation and conditional recovery

Define $a_1 = \alpha - d$ and $a_2 = \alpha/K$. Estimate

$$\vartheta = (s, a_1, a_2, \beta, \delta, \rho, c), \quad a_2 > 0, \quad (3)$$

using

$${}^C D_t^q T = s + a_1 T - a_2 T^2 - \beta TV. \quad (4)$$

A positive a_1 is not required mathematically. The other rate coefficients remain positive.

Theorem 2. For known q and exact fully observed states on $[0, \tau]$, the reduced coefficients are uniquely recoverable if the following three Gram matrices are nonsingular:

$$G_T = \int_0^\tau r_T r_T^\top dt, \quad r_T = (1, T, -T^2, -TV)^\top, \\ G_I = \int_0^\tau r_I r_I^\top dt, \quad r_I = (TV, -I)^\top, \quad G_V = \int_0^\tau r_V r_V^\top dt, \quad r_V = (I, -V)^\top. \quad (5)$$

Proof. Two sets of coefficients giving the same observed states have differences satisfying $r_{-T}^\top \Delta \theta_{-T} = 0$, $r_{-I}^\top \Delta \theta_{-I} = 0$ and $r_{-V}^\top \Delta \theta_{-V} = 0$ almost everywhere. Squaring and integrating gives $\Delta \theta^\top G \Delta \theta = 0$ in each block. Nonsingularity of each positive semidefinite Gram matrix forces the coefficient differences to vanish. The infection coefficient β recovered in the first two blocks must agree. $\square$

These sufficient conditions are not asserted to hold for every trajectory. Constant equilibrium observations fail them, and nearly singular matrices imply poor practical conditioning. Positivity constraints and a shared β may help estimation but do not make these rank assumptions automatic.

Corollary 1. If d is known independently and the reduced coefficients are recovered, then $\alpha = a_1 + d$ and $K = (a_1 + d)/a_2$, provided $\alpha > 0$. If K is known, then $\alpha = a_2 K$ and $d = a_2 K - a_1$, provided $d > 0$.

Estimating q jointly requires an additional analysis; Theorem 2 does not establish joint identifiability of q and all rates. Viral-load-only data likewise require a separate observability and identifiability investigation.

6. Correct hybrid derivatives and bounds

Following [2], let

$$H(z) = w_1 \sigma(z) + w_2 \text{ReLU}(z) + w_3 \tanh(z) + w_4 \text{LReLU}_a(z) + w_5 \text{ELU}_\kappa(z), \quad (6)$$

with nonnegative weights summing to one, 00. LReLU equals z for $z > 0$ and az for $z \leq 0$. ELU equals z for $z > 0$ and $\kappa(e^z - 1)$ for $z \leq 0$.

For $z \neq 0$,

$$H'(z) = w_1 \sigma(z)(1 - \sigma(z)) + w_3 \operatorname{sech}^2(z) + \begin{cases} w_2 + w_4 + w_5, & z > 0, \\ aw_4 + \kappa w_5 e^z, & z < 0. \end{cases} \quad (7)$$

In particular, $\sigma'(z) = e^z / (1 + e^z)^2$. The e^{2z} numerator printed in the derivative calculations of [2] is not the sigmoid derivative. At zero the mixture is continuous, but differentiability requires equality of its one-sided derivatives; this cannot generally be assumed.

Theorem 3. Put

$$m = \min\{w_2 + w_4 + w_5, aw_4\}, \quad L = w_1/4 + w_3 + w_2 + w_4 \max(1, a) + w_5 \max(1, \kappa). \quad (8)$$

Then $m \leq H'(z) \leq L$ for $z \neq 0$. Furthermore, for all $x < y$,

$$m(y - x) \leq H(y) - H(x) \leq L(y - x). \quad (9)$$

Proof. The sigmoid derivative lies in $(0, 1/4]$, the tanh derivative lies in $(0, 1]$, and $0 < e^z \leq 1$ on $z < 0$. All weighted terms in (7) are nonnegative. The constant positive and negative branch contributions give the lower bounds, while the displayed component maxima give the upper bound. Integrate over $[x, y]$, splitting at zero if necessary; H is continuous and piecewise smooth. $\square$

If $w_4 > 0$, $m > 0$ and H is globally strictly increasing with a quantitative inverse Lipschitz bound. For $w = (0.1, 0.2, 0.25, 0.2, 0.25)$, $a = 0.01$ and $\kappa = 1$, the bounds are $m = 0.002$ and $L = 0.925$. The negative-tail derivative tends to 0.002, while the positive-tail derivative tends to 0.65.

A convex combination of functions does not imply a convex resulting function unless additional conditions hold. On $z > 0$, the sigmoid and tanh components have negative second derivatives, while the remaining components are linear. With $w_1 > 0$ or $w_3 > 0$, H is strictly concave on that branch. Thus “convex” here describes the weights, not global convexity of H or of the learning objective.

7. Why scalar slope bounds do not ensure deep gradient retention

For square layers $h_\ell = H(W_\ell h_{\ell-1} + b_\ell)$, away from activation kinks, write $D_\ell = \operatorname{diag}(H'(z_\ell))$. If each W_ℓ is nonsingular and $\sigma \min(W_\ell) \geq s_\ell$, then

$$\sigma_{\min}(J_{\text{net}}) \geq \prod_{\ell=1}^n m s_\ell, \quad \|J_{\text{net}}\|_2 \leq \prod_{\ell=1}^n L \|W_\ell\|_2. \quad (10)$$

Proof. Each D_ℓ has singular values between m and L . Apply the minimum singular value product inequality and the operator norm product inequality to the layer Jacobians $D_\ell W_\ell$. $\square$

These statements require compatible square nonsingular layers for the lower bound. Even with $s_\ell = 1$, m^n can decay rapidly. A positive scalar derivative therefore prevents an exactly flat scalar branch but does not eliminate depth-dependent vanishing gradients. Output transformations, the loss derivative and weight conditioning add further effects.

8. Fractional residual learning framework

Use a network $x\theta(t) = (T\theta, I\theta, V\theta)$. Positive outputs can be imposed using softplus; exact initial values can instead be imposed through a suitable trial construction, with any resulting sign constraints checked explicitly. Mixture coefficients may be fixed or parametrised on a constrained simplex. If a positive lower bound is desired, enforce $w_4 \geq \varepsilon > 0$ rather than merely using an unconstrained softmax whose weight can approach zero.

At $t_n = n\Delta t$ approximate the Caputo derivative, for $0 < q < 1$, by

$$\mathcal{D}_{\Delta t}^q x_n = \frac{1}{\Gamma(2-q)\Delta t^q} \sum_{j=0}^{n-1} [(j+1)^{1-q} - j^{1-q}] (x_{n-j} - x_{n-j-1}). \quad (11)$$

This history-dependent L1 formula is applied to sampled network outputs; an ordinary automatic first derivative is not a substitute for a fractional derivative. At $q=1$ use the ordinary derivative separately. Fractional initial-time singularities can reduce accuracy, so independent grid refinement is necessary [8].

Define residuals by subtracting the right-hand sides of (1), using the reduced parametrisation (4). Train with

$$\mathcal{L} = \lambda_d \mathcal{L}_{\text{data}} + \lambda_r \mathcal{L}_{\text{residual}} + \lambda_0 \mathcal{L}_{\text{initial}} + \lambda_w \mathcal{L}_{\text{regularisation}}. \quad (12)$$

State scales must be normalised in both the data and residual terms. The loss concerns state observations and the specified equations; a discrepancy between H and H' is not a model-data fitting loss for this problem.

9. Infinitesimal sensitivity and the singular information direction

The finite transformation in Theorem 1 also yields a local sensitivity statement. Assume that the solution depends differentiably on α , d and K on the interval under study. This assumption is natural for a smooth locally Lipschitz vector field away from a breakdown of the solution, but the result can also be read directly as a derivative of the exact invariance identity.

Proposition 1. For any differentiable state observation map $y(\theta)$, the sensitivities of the original parametrisation satisfy

$$\frac{\partial y}{\partial \alpha} + \frac{\partial y}{\partial d} + \frac{K}{\alpha} \frac{\partial y}{\partial K} = 0. \quad (13)$$

Consequently, an information matrix formed from these sensitivities is singular when all three coefficients are freely estimated.

Proof. Differentiate $y(\alpha+h, d+h, K(\alpha+h)/\alpha) = y(\alpha, d, K)$ at $h=0$. This gives (13). If S is the sensitivity matrix, it has the nonzero null vector $v = (1, 1, K/\alpha)$ in the relevant three coordinates. For any positive definite observation weighting matrix W , $v^T S^T W S v$ equals zero. Therefore $S^T W S$ is singular. The same conclusion holds if further identifiable coefficients are appended to the parameter vector, because v can be extended by zero entries. $\square$

This is a structural rank defect rather than merely a small numerical singular value. Measurement noise may perturb an empirical information estimate, and regularisation can make a penalised Hessian invertible, but neither change makes the data identify the original triple. If a prior is used, the resulting estimate must be described as conditional on that prior. In particular, a regularisation penalty that favours one carrying capacity selects a representative of the equivalence class; it does not demonstrate that the selected capacity is uniquely supported by the trajectory.

10. A precise inverse formulation

For a known order and fully observed states, the reduced system is linear in its coefficients once the trajectories and their fractional derivatives are available. If b_T , b_I and b_V denote the derivative observations, the coefficients can be obtained from least-squares problems with the design matrices in (5). In an ideal continuous experiment these correspond to the Gram matrices of Theorem 2. In a finite experiment, matrix rank must be checked at the actual sample points rather than assumed from the number of samples alone.

The infection coefficient appears in both the T and I equations. A joint formulation should enforce the same value instead of fitting unrelated infection rates in the two blocks. Positivity constraints on s , a_2 , β , δ , ρ and c are appropriate; the net growth a_1 may have either sign. If noisy derivatives are estimated by smoothing the data, the smoothing bias and the history dependence of the Caputo operator must be included in an uncertainty analysis. Treating fitted derivative values as noiseless would overstate parameter precision.

A neural formulation can replace direct derivative estimation by evaluating a history operator on a smooth state approximation. It must still estimate the reduced coefficients or impose an independently justified parameter constraint. If q is learned simultaneously, several orders can produce similar finite records, and a separate profile or sensitivity calculation is needed. The present reduced-rank result does not imply that learning q is impossible; it only prevents the known-order recovery theorem from being extended without proof.

11. Constraining the hybrid mixture

A convenient coefficient parametrisation uses nonnegative weights on the simplex. To enforce a leaky-component floor ϵ , let π_j be a softmax vector and define

$$w_4 = \epsilon + (1 - \epsilon)\pi_4, \quad w_j = (1 - \epsilon)\pi_j \quad (j \neq 4), \quad 0 < \epsilon < 1. \quad (14)$$

The weights remain nonnegative and sum to one. Since a is in $(0,1]$, the positive-branch constant contribution is at least w_4 and the negative-branch contribution is at least $a w_4$.

Proposition 2. Under (14), H satisfies

$$H(y) - H(x) \geq a\epsilon(y - x), \quad x < y. \quad (15)$$

Moreover, if a , κ and ϵ are fixed, its upper derivative bound is uniformly bounded by the largest component derivative bound.

Proof. The lower bound follows from Theorem 3 and w_4 at least ϵ . For the upper bound, H prime is a convex combination of component derivatives almost everywhere; hence it cannot exceed their largest individual supremum. The continuity of H allows integration across zero. $\square$

A useful consequence is that H is injective and its scalar inverse on its range is Lipschitz with constant no greater than $1/(a\epsilon)$. This conclusion is about a scalar activation, not an entire network. A rectangular weight matrix can erase directions even if every activation is injective. A small lower slope can also give a large inverse Lipschitz constant and poor practical conditioning.

For the coefficients inherited from [2], I use $a=0.01$ and $\kappa=1$. These component settings must be stated because the mixture weights alone do not determine the activation. The analytical lower bound is 0.002 and the stated component-wise upper bound is 0.925. I do not select these values by optimising a dataset. They provide a reproducible example for checking the derivative formulas.

12. Completed computational experiments

I generate a synthetic trajectory for the dimensionless parameter vector (0.1, 0.8, 0.1, 2, 0.5, 0.4, 1, 0.8), with initial state (1.5,0.05,0.1), common order $q=0.8$, step $h=0.025$ and horizon 40. The implicit L1 solver and nonlinear residual check are specified in the accompanying script. All coefficient-order conventions are recorded in the machine-readable results. The trajectory is used as a design example, not as a substitute for observed patient data.

The activation is evaluated on 2,401 evenly spaced points in $[-12,12]$. The derivative at zero is not included in a claim of classical differentiability. The symmetry calculation evaluates 201 displacements h between -0.09 and 1.5. Each transformed triple is positive. I compare vector fields along the same synthetic trajectory rather than comparing two separately integrated trajectories,

so the numerical comparison isolates the exact algebraic invariance without an integration-error confound.

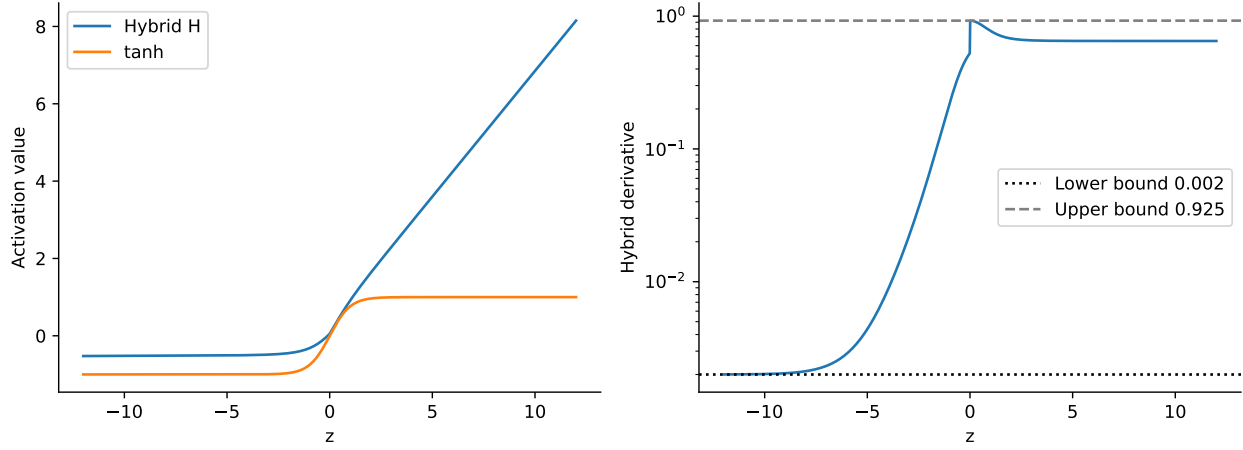


Figure 1: Hybrid activation and its analytical derivative for the stated weights, $a=0.01$ and $\kappa=1$. The horizontal lines indicate the proven slope bounds. The kink at zero is not treated as a smooth derivative point.

The negative tail approaches the leaky contribution 0.002, while the positive tail approaches 0.65. The sigmoid and tanh terms affect the transition near the origin. These limits are consistent with the analytical formula and do not constitute evidence that one activation trains faster than another.

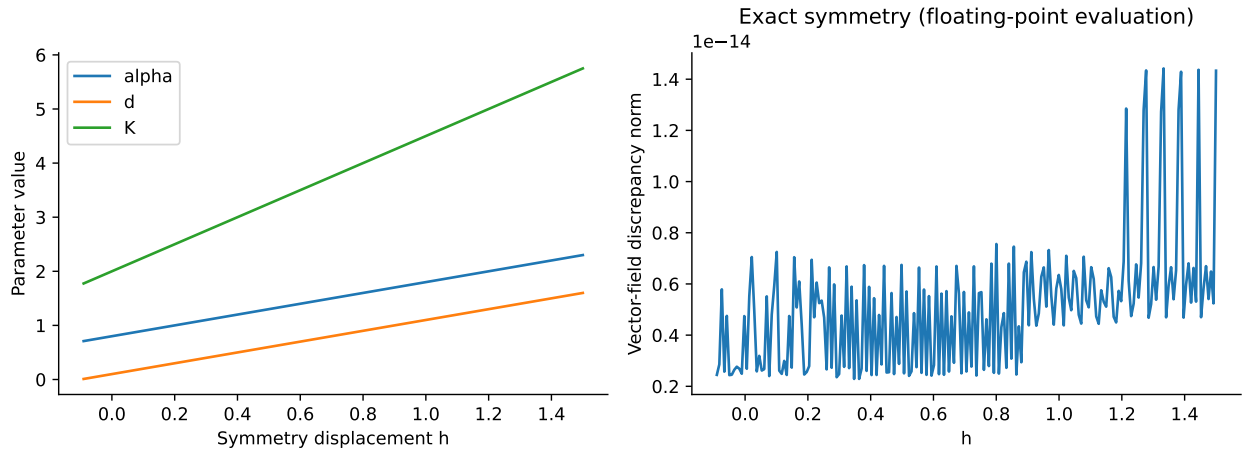


Figure 2: Parameter displacement along the exact equivalence class and the corresponding vector-field discrepancy norm. The discrepancy is at floating-point scale despite substantial changes in individual parameters.

The largest vector-field discrepancy norm is approximately 1.44 times 10 to the power minus fourteen. This nonzero computational value is consistent with floating-point arithmetic; the analytical discrepancy is exactly zero. The graph visually distinguishes parameter change from dynamical change. In particular, the simultaneous increase of α and d leaves net growth unchanged, while the compensating change in K preserves the quadratic coefficient.

For the T design, the singular values are approximately 1.878884, 0.675883, 0.113411 and 0.010776. For the I design they are 1.410188 and 0.106631; for the V design they are 1.413244 and 0.052368. These finite-sample values support the sufficient rank condition in this example. They do not prove rank for all parameter sets or all trajectories. Column normalisation is essential for a meaningful

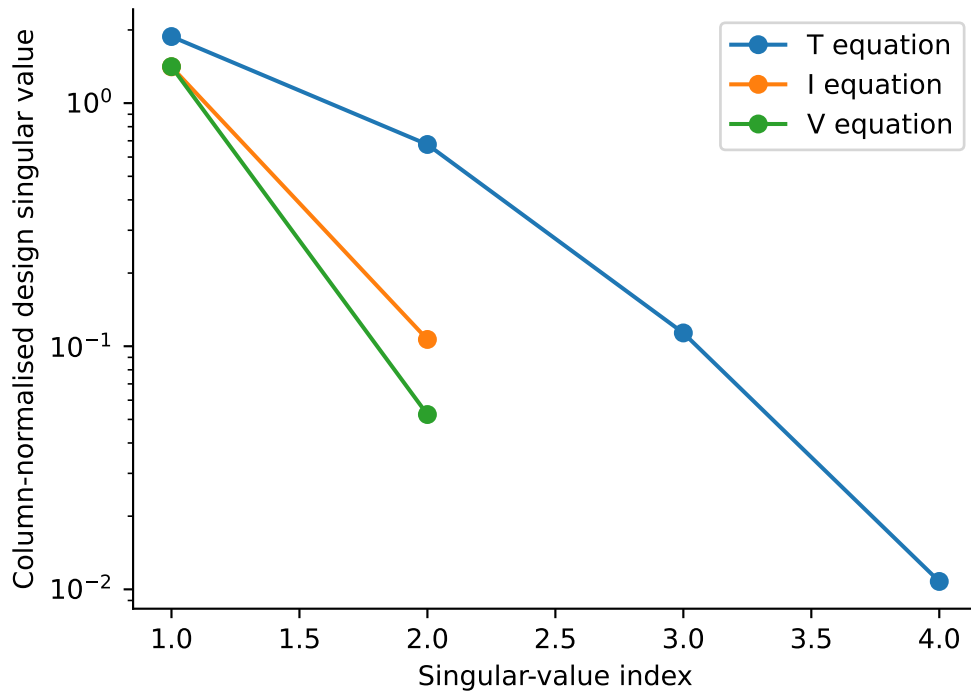


Figure 3: Singular values of the column-normalised reduced design blocks over $[0, 40]$. Positive values support full discrete rank for this example; unequal magnitudes show unequal conditioning.

comparison because raw polynomial and viral regressors can have different scales.

The T design condition number decreases from approximately 16,421.8 at horizon 2 to 163.2 at horizon 20, before increasing to 174.4 at horizon 40. The I and V blocks also vary nonmonotonically. Early observations may not sufficiently separate the polynomial regressors; later observations can add information but may also accumulate nearly equilibrium states. This illustrates why observation design concerns excitation and conditioning, not simply record length.

The log-scale product graph makes the distinction between a strictly positive local slope and a usable deep gradient explicit. For example, the product of the negative-tail slope over n layers is approximately (0.002) to the power n in the corresponding tail regime. Even an upper bound below one decays exponentially when multiplied repeatedly. Weight singular values can compensate, suppress or amplify these effects; an activation-only plot cannot determine a complete training outcome.

13. Discussion, limitations and recommendations for model fitting

The analytical symmetry persists for every common fractional order because it leaves the entire right-hand side unchanged. The result is therefore not an artefact of the selected numerical grid or initial condition. On the other hand, the discrete conditioning results depend on the synthetic parameter set, sampling interval and observation horizon. Their practical role is to demonstrate an assessment procedure, not to specify an optimal observation schedule for clinical use.

The computational study does not train a state network, recover coefficients from noisy clinical records, or compare neural generalisation errors. The neural residual objective is a fully specified methodological extension that respects the established identifiability restrictions. A later benchmark should compare fixed tanh, the supplied hybrid mixture and a floor-constrained mixture using equal capacity, repeated seeds and an independent reference solver. Any accuracy comparison must identify

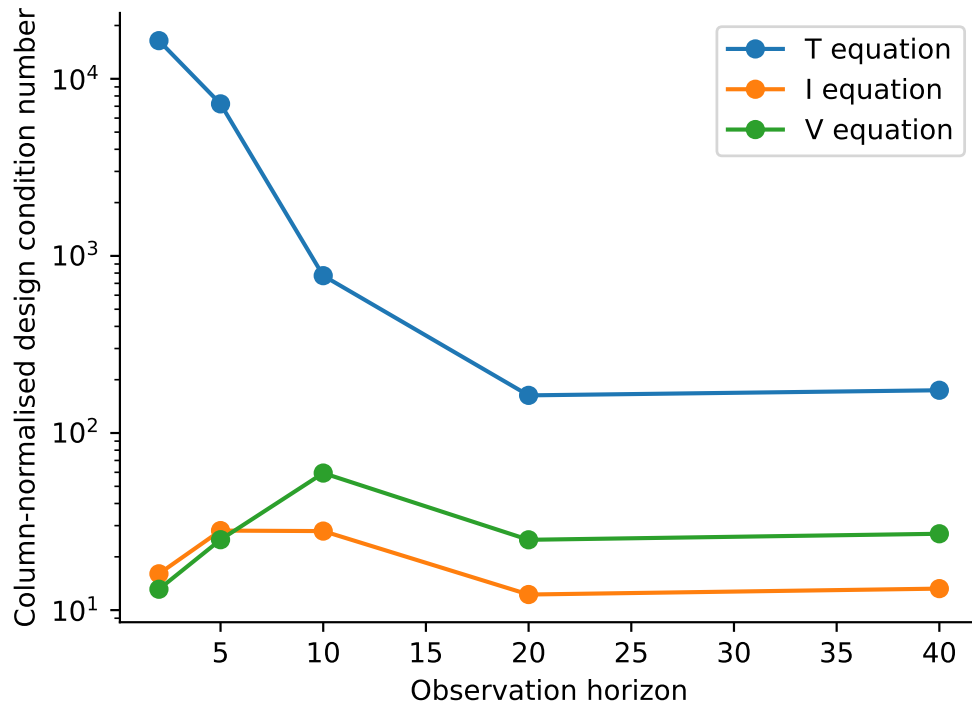


Figure 4: Condition numbers of column-normalised design blocks for observation horizons 2, 5, 10, 20 and 40. A longer horizon does not monotonically improve every design block.

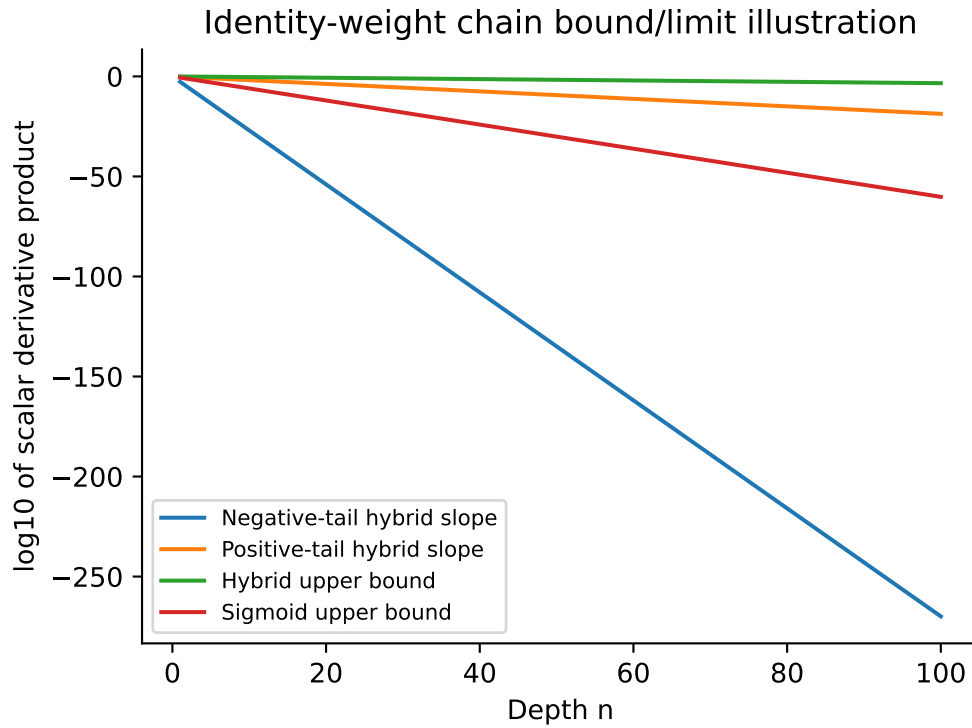


Figure 5: Logarithms of scalar slope products through identity-weight chains. The curves are bounds or tail-limit illustrations, not measured backpropagation gradients in trained networks.

whether the estimator predicts states, estimates reduced combinations, or selects original parameters using priors.

The single-author perspective of this manuscript is reflected in its analytical scope: I derive the symmetry and bounds, implement the illustrative checks, and distinguish completed results from proposed learning experiments. The supplied papers [1,2] are the starting references, not substitutes for independent derivations. A publication claim of a wholly new identifiability phenomenon would require comparison with specialised HIV inverse-model literature beyond the fundamental sources cited here.

14. Data availability and computational details

The supplementary package includes the source code, figure files in PDF and PNG form, the numerical trajectory CSV files and a JSON record of the computed quantities. The implementation uses NumPy, SciPy and Matplotlib, with exact versions recorded in that JSON file. No patient data were used. The calculations require no trained weights or external dataset. Re-running the script regenerates the numerical illustrations; graph titles and captions explicitly state the synthetic or analytical nature of each result.

15. Conclusion

The model identifies combinations α -d and α/K rather than α, d, K separately. Hybrid activations cannot overcome this exact symmetry. A reduced parametrisation, independently known coefficients and informative trajectories are necessary ingredients of a defensible inverse study. Correct derivative bounds clarify what the mixture in [2] can guarantee, while the network product bounds show why empirical claims about deep gradient retention still require controlled experiments. This inverse-learning focus is distinct from an equilibrium-stability study and should be assessed on its own literature and computational evidence.

16. Declarations

Funding: No funding information was supplied for this manuscript.

Competing interests: A competing-interest statement must be supplied by the author before journal submission.

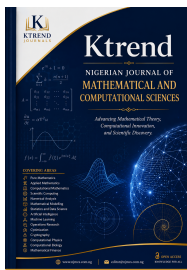
Ethics and consent: The reported computations use synthetic model states and no human participant data.

Author contributions: The manuscript is presented under the sole authorship of Okeke Ikenna Stephen. Final responsibility for the scientific content and submission rests with the named author.

References

- [1] Okeke, S. I., Nwagor, P., Yakubu, H., & Ogoegbulem, O. (2019). Modelling HIV infection of CD4+ T cells using fractional order derivatives. *Asian Journal of Mathematics and Applications*, 2019, Article ama0519, 1–6.
- [2] Essang, S. O., Ante, J. E., Otobi, A. O., Okeke, S. I., Akpan, U. D., Francis, R. E., Auta, J. T., Essien, D. E., Fadugba, S. E., Kolawole, O. M., Asanga, E. E., & Ita, B. I. (2025). Optimizing neural networks with convex hybrid activations for improved gradient flow. *Advanced Journal of Science, Technology and Engineering*, 5(1), 10–26. <https://doi.org/10.52589/AJSTE-UOBYFV1B>
- [3] Cong, N. D., Doan, T. S., Siegmund, S., & Tuan, H. T. (2015). *Linearized asymptotic stability for fractional differential equations* [Preprint]. arXiv. <https://arxiv.org/abs/1512.04989>
- [4] Jagtap, A. D., Kawaguchi, K., & Karniadakis, G. E. (2020). Adaptive activation functions accelerate

- convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404, Article 109136. <https://doi.org/10.1016/j.jcp.2019.109136>
- [5] Pang, G., Lu, L., & Karniadakis, G. E. (2019). fPINNs: Fractional physics-informed neural networks. *SIAM Journal on Scientific Computing*, 41(4), A2603–A2626. <https://doi.org/10.1137/18M1229845>
 - [6] Diethelm, K. (2010). *The analysis of fractional differential equations: An application-oriented exposition using differential operators of Caputo type*. Springer. <https://doi.org/10.1007/978-3-642-14574-2>
 - [7] Diethelm, K., Ford, N. J., & Freed, A. D. (2002). A predictor-corrector approach for the numerical solution of fractional differential equations. *Nonlinear Dynamics*, 29, 3–22. <https://doi.org/10.1023/A:1016592219341>
 - [8] Jin, B., Lazarov, R., & Zhou, Z. (2016). An analysis of the L1 scheme for the subdiffusion equation with nonsmooth data. *IMA Journal of Numerical Analysis*, 36(1), 197–221. <https://doi.org/10.1093/imanum/dru063>
 - [9] Garrappa, R. (2018). Numerical solution of fractional differential equations: A survey and a software tutorial. *Mathematics*, 6(2), Article 16. <https://doi.org/10.3390/math6020016>
 - [10] van den Driessche, P., & Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180(1–2), 29–48. [https://doi.org/10.1016/S0025-5564\(02\)00108-6](https://doi.org/10.1016/S0025-5564(02)00108-6)
 - [11] Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., & Ho, D. D. (1996). HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255), 1582–1586. <https://doi.org/10.1126/science.271.5255.1582>
 - [12] Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., & Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373, 123–126. <https://doi.org/10.1038/373123a0>
 - [13] Perelson, A. S., & Nelson, P. W. (1999). Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Review*, 41(1), 3–44. <https://doi.org/10.1137/S0036144598335107>
 - [14] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Vol. 9, pp. 249–256). PMLR. <https://proceedings.mlr.press/v9/glorot10a.html>
 - [15] Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Vol. 15, pp. 315–323). PMLR. <https://proceedings.mlr.press/v15/glorot11a.html>
 - [16] Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). *Fast and accurate deep network learning by exponential linear units (ELUs)* [Preprint]. arXiv. <https://arxiv.org/abs/1511.07289>
 - [17] Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization* [Preprint]. arXiv. <https://arxiv.org/abs/1412.6980>
 - [18] Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15), 1923–1929. <https://doi.org/10.1093/bioinformatics/btp358>
 - [19] Bellman, R., & Åström, K. J. (1970). On structural identifiability. *Mathematical Biosciences*, 7(3–4), 329–339. [https://doi.org/10.1016/0025-5564\(70\)90132-X](https://doi.org/10.1016/0025-5564(70)90132-X)
 - [20] Eisenberg, M. C., & Hayashi, M. A. L. (2014). Determining identifiable parameter combinations using subset profiling. *Mathematical Biosciences*, 256, 116–126. <https://doi.org/10.1016/j.mbs.2014.08.008>



- [21] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707. <https://doi.org/10.1016/j.jcp.2018.10.045>
- [22] Apicella, A., Donnarumma, F., Isgrò, F., & Prevete, R. (2021). A survey on modern trainable activation functions. *Neural Networks*, 138, 14–32. <https://doi.org/10.1016/j.neunet.2021.01.012>
- [23] Arafa, A. A. M., Rida, S. Z., & Khalil, M. (2012). Fractional modeling dynamics of HIV and CD4+ T-cells during primary infection. *Nonlinear Biomedical Physics*, 6, Article 1. <https://doi.org/10.1186/1753-4631-6-1>