

Dynamic Mathematical PDE Modeling and Machine Learning for TANGO2 Identification and Crisis Prediction

Abhay Tippimath
McNeil High School
Austin, Texas, USA
abhay.tippi03@gmail.com

Abstract

TANGO2 (Transport And Golgi Organization 2) deficiency disorder (TDD) is a rare autosomal recessive condition marked by unpredictable, life-threatening metabolic crises, including rhabdomyolysis, cardiac arrhythmias, and metabolic encephalopathy. Supportive treatment, particularly B-complex vitamin supplementation and fasting avoidance, substantially reduces the frequency and severity of crises, so undiagnosed patients remain at the greatest risk. This work presents a two-stage framework that couples supervised machine learning classification with numerical simulation of a stochastic partial differential equation (PDE). The first stage identifies TDD patients from their stable baseline presentation, and the second models future crisis behavior as a threshold-crossing dynamical system. Using deidentified Human Phenotype Ontology (HPO) profiles from 90 molecularly confirmed TDD patients and 141 diagnosed Undiagnosed Diseases Network (UDN) controls, we exclude all 23 crisis-defining phenotypes and restrict the primary analysis to the 30 HPO terms recorded by both annotating teams, so that classification depends only on information available before a crisis. A Random Forest classifier, evaluated by repeated stratified tenfold cross-validation, attains an AUC of 0.993 (variance-corrected 95% interval 0.980 to 1.000), with 95.6% sensitivity and 100% specificity. The result withstands a label permutation test ($p = 0.001$), stricter feature sets, and ablation of the most important features, indicating a signal distributed across many correlated phenotypes. Two phenotypes, paroxysmal lethargy and compensatory head posture, were recorded only in TDD patients; because the control annotating team never used these terms, they are reported as hypotheses that cannot yet be separated from annotation style. Separately, a stochastic reaction-diffusion PDE with a patient-specific source term reproduces the episodic nature of metabolic crises and exhibits a critical source amplitude that partitions patients into stable and recurrent regimes. Scalar models without diffusion cannot reproduce this behavior without recalibration, and the patient-specific parameter is recoverable from a single simulated timeline ($r = 0.992$). Together, these results indicate that phenotypes present before any crisis can identify TDD before life-threatening events occur, while the PDE offers a testable hypothesis for how continuous metabolic risk produces discrete, episodic crises.

Keywords

TANGO2, rare disease, random forest, reaction diffusion equation, partial differential equations, Human Phenotype Ontology, crisis prediction, bioinformatics, machine learning

ACM Reference Format:

Abhay Tippimath. 2026. Dynamic Mathematical PDE Modeling and Machine Learning for TANGO2 Identification and Crisis Prediction. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 4)*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.67149/yhjs2024.5/1542temi>

1 Introduction

1.1 Clinical and Molecular Background

TANGO2 (Transport And Golgi Organization 2) deficiency disorder is caused by biallelic loss of function variants in *TANGO2* on chromosome 22q11.21, most often a recurrent deletion of exons 3 to 9 [8, 9]. The protein localizes to the endoplasmic reticulum, the Golgi apparatus, and mitochondria, and its loss impairs lipid homeostasis, membrane trafficking, and mitochondrial fatty acid oxidation. Its precise molecular function is still being characterized, so no targeted therapy exists and care depends on early recognition and supportive treatment.

The natural history has two states. Between crises, patients show a relatively stable neurodevelopmental phenotype that includes developmental delay, intellectual disability, dysarthria, and gait disturbance. During a crisis, usually triggered by fasting or febrile illness, patients decompensate with rhabdomyolysis, metabolic acidosis, encephalopathy, and dangerous ventricular arrhythmias [4]. Fourteen of the 90 patients analyzed here are deceased. Figure 1 summarizes the mechanism: loss of TANGO2 disrupts three converging processes, and under metabolic stress, cellular energy production fails in the tissues with the highest demand. B complex supplementation, particularly pantothenate (B5) and folate (B9), is thought to act upstream by supporting coenzyme A availability, which is why it reduces crisis frequency and severity.

1.2 Two Problems with Complementary Structure

This history creates two clinical problems. The *diagnostic* problem is that the stable phenotype overlaps with many neurogenetic conditions, so TDD is often not suspected until a first crisis has already occurred, even though prophylaxis exists. Computationally, this is classification of a sparse binary phenotype vector. The *prognostic* problem is that clinicians cannot anticipate the next crisis, so monitoring is uniform rather than stratified. Computationally, this is a dynamical question about a latent risk that accumulates until it

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

IJSCAR VOL. 3, ISSUE 4,
© 2026 The Author(s). Published by IJSCAR.
<https://doi.org/10.67149/yhjs2024.5/1542temi>

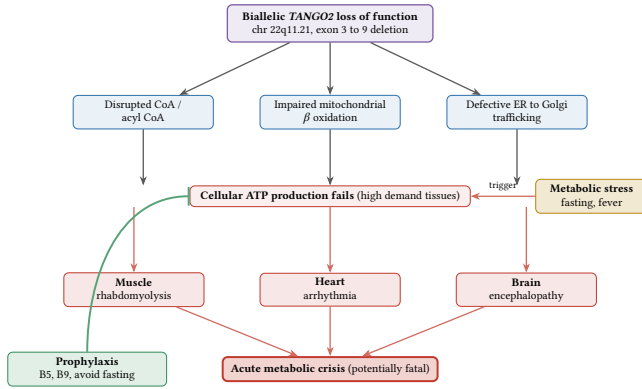


Figure 1: Pathophysiology of TANGO2 deficiency disorder. Loss of function disrupts three converging molecular processes. Under metabolic stress this causes energy failure and injury to muscle, heart, and brain, which together produce the acute crisis. B complex prophylaxis (green) acts upstream.

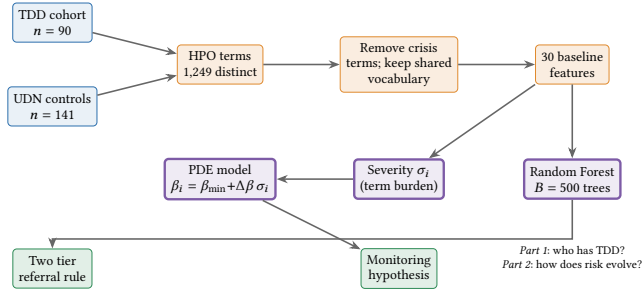


Figure 2: Overview of the framework. Shared baseline phenotypes drive a Random Forest classifier (Part 1). A severity score computed from the same phenotypes sets the patient specific source amplitude of the PDE (Part 2).

crosses a threshold. Figure 2 shows how the framework connects the two.

1.3 Contributions

- (1) We show that phenotypes observable strictly before any crisis identify TDD accurately, after removing every crisis defining term and restricting to vocabulary shared by both annotating teams, and we stress test the result with a permutation test, a variance corrected interval, stricter feature sets, and alternative models.
- (2) We report two phenotypes recorded only in TDD patients in this cohort, explain why their apparent specificity cannot yet be separated from annotation style, and contrast them with a more sensitive neurological triad recorded by both annotating teams.
- (3) We formulate a stochastic reaction diffusion model of crisis risk, derive a critical source amplitude β^* separating quiescent and recurrent regimes, and test empirically what the spatial term adds over scalar models without diffusion.

- (4) We verify the solver numerically, quantify parameter sensitivity, and show through a synthetic recovery test that the patient specific parameter is identifiable from crisis timelines, which prepares the model for future fitting to longitudinal data.

2 Related Work

Phenotype driven rare disease analysis. The Human Phenotype Ontology (HPO) provides a standardized vocabulary of phenotypic abnormalities and underpins computational rare disease analysis [6]. The Undiagnosed Diseases Network (UDN), which supplied our controls, applies systematic phenotyping to patients whose presentations defeated standard workup [15]. Whereas phenotype driven tools usually rank candidate genes for one patient, we learn a classifier that separates one confirmed disorder from its realistic differential using only features available before a crisis.

Classification on sparse binary features. Random Forests average many decorrelated trees, need no feature scaling, capture interactions, and expose feature importances [2]. These properties suit small, sparse, correlated clinical data where an interpretable rule is the goal. Because leakage and dataset artifacts frequently inflate clinical machine learning results, we pair the model with explicit safeguards and falsification tests, including label permutation [12].

Reaction diffusion models and stiff solvers. Reaction diffusion equations descend from the Fisher KPP equation [5, 7] and are widely used for pattern formation and physiology [10]. We apply this machinery to an abstract coordinate of coupled physiological subsystems. Diffusion terms are stiff, so we use an implicit explicit (IMEX) time integrator [1] verified against standard numerical criteria [14].

3 Data

3.1 Cohorts

Two deidentified cohorts were provided by the Department of Molecular and Human Genetics at Baylor College of Medicine. The TDD cohort contains 90 individuals with molecularly confirmed TANGO2 deficiency (49 male, 41 female; 76 living, 14 deceased; median age at evaluation 8.8 years across the 86 with recorded age). The control cohort contains 141 UDN participants with a confirmed diagnosis other than TANGO2 deficiency, most with neurological ($n = 67$) or multiple congenital anomaly ($n = 39$) diagnoses. The task is therefore TDD versus a phenotypically overlapping differential rather than versus healthy children. The cohorts contain 1,249 distinct HPO terms and no missing phenotype records.

3.2 Feature Construction and Analysis Timeline

Two safeguards shape the features. First, every phenotype describing a crisis or its aftermath was removed, because a classifier that sees *rhabdomyolysis* would simply detect a crisis that has already occurred. The 23 excluded categories (35 HPO terms) include *rhabdomyolysis*, *myoglobinuria*, *elevated creatine kinase*, *cardiac arrest*, *cardiomyopathy*, *arrhythmia*, *prolonged QT*, *metabolic and lactic acidosis*, *hyperammonemia*, *hypoglycemia*, *encephalopathy*, and *coma*. Keeping terms recorded in at least three patients leaves 261 candidate features.

Table 1: Random Forest hyperparameters, fixed before evaluation.

Setting	Value	Setting	Value
Trees	500	Max depth	none
Split criterion	Gini	Features per split	$\lfloor \sqrt{30} \rfloor = 5$
Bootstrap	yes	Class weight	balanced
Min samples split	2	Min samples leaf	1
Random seed	fixed	Threshold	0.5

Second, the cohorts were phenotyped by different teams with nearly disjoint vocabularies. A term used by only one team separates the cohorts regardless of biology, so we retained only the 31 candidate terms used by both teams. We then excluded the single electrocardiographic term, because an ECG abnormality is a test result rather than a bedside observation, leaving a primary set of 30 features. The two candidate markers of Section 5.4 were never recorded in controls and are therefore outside this set; the headline results are achieved without them. By the same reasoning, their absence among controls may reflect annotation vocabulary rather than biology.

Reviewers rightly ask when such choices were made. The crisis exclusion list and the three patient threshold were fixed before any model was trained. The shared vocabulary restriction was added after an audit of term overlap, and the ECG exclusion was made later on the advice of a reviewing clinician, after initial results had been seen. Both later restrictions made the task harder rather than easier (AUC 0.996 for the 261 candidate features, 0.994 for 31 shared terms, 0.993 for the final 30), so neither was chosen to improve performance, and results at every stage are reported. Section 5.2 further shows that the result does not depend on the frequency threshold.

Each patient is represented as a binary vector $\mathbf{x} \in \{0, 1\}^{30}$ whose entry j is 1 when the patient carries term j . Binary indicators need no scaling, and the absence of a term is treated as a true negative. The 90 to 141 class imbalance is handled by inverse frequency class weights rather than resampling.

4 Methods

4.1 Random Forest Classifier

A Random Forest of $B = 500$ trees estimates $\hat{P}(y = 1 \mid \mathbf{x}) = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x})$, and a patient is classified as TDD when $\hat{P} \geq 0.5$. Feature importance is the mean decrease in Gini impurity $G(m) = 2p_m(1 - p_m)$, which rewards features that are both common in cases and rare in controls. Because correlated phenotypes share credit, importances identify informative clusters rather than independent causes. All hyperparameters (Table 1) were fixed in advance and held constant across folds, and the implementation uses the RandomForestClassifier of Pedregosa et al. [13].

4.2 Evaluation and Stress Tests

With 231 patients, a single split is unreliable, so we use repeated stratified tenfold cross validation: 10 repeats, 100 model fits, about 208 training and 23 validation patients per fold, and every patient predicted once per repeat by a model that never saw it. Calibration

is measured with the Brier score $BS = \frac{1}{n} \sum_i (\hat{p}_i - y_i)^2$ and its skill relative to a model that always predicts the base rate.

Because the 100 fits reuse the same 231 patients, they are not independent, and a naive standard deviation understates uncertainty. We therefore also report the corrected resampled t interval of Nadeau and Bengio [11], which inflates the variance of fold estimates by the ratio of test to training size. To test whether the AUC could arise from a small, easily separable dataset, we ran a label permutation test [12]: the labels were shuffled 1,000 times and the full cross validation was repeated each time (100 trees per forest for tractability). Further checks were stricter feature sets, the annotation depth of each record, duplicate records, ablation of the most important features, and two alternative model families, logistic regression and a support vector machine with a radial basis kernel.

4.3 Reaction Diffusion Model of Crisis Risk

We model crisis risk as a field $r(x, t) \in [0, 1]$, where t is time in days and $x \in [0, 1]$ is a latent coordinate indexing coupled physiological subsystems:

$$\frac{\partial r}{\partial t} = D \frac{\partial^2 r}{\partial x^2} + \alpha r(1 - r) + \beta S(x) - \delta r, \quad (1)$$

with no flux boundaries and $r(x, 0) = 0.13$. Diffusion represents spread of instability between coupled subsystems; logistic growth represents accumulation that reinforces itself, capped at 1 [5, 7]; the Gaussian source $S(x) = \exp[-(x - x_0)^2 / 2\sigma^2]$ represents a constitutional vulnerability whose amplitude β is the only patient specific parameter; and $-\delta r$ represents homeostatic clearance. A crisis is declared when $\max_x r \geq r^*$, after which the reset $r \leftarrow r_{\text{rest}} + \kappa(r - r_{\text{rest}})$ with $\kappa = 0.52$ and $r_{\text{rest}} = 0$ removes 48 percent of the accumulated risk while leaving β unchanged, so crises recur from the dynamics rather than from an imposed period. Because real crises follow discrete triggers, we add multiplicative noise $\sigma_\eta \sqrt{r(1 - r)} S(x) \dot{W}(x, t)$, which vanishes at 0 and 1 and so keeps r in range.

The coordinate x is a modeling abstraction rather than a measured quantity. The reason for a spatial model is that risk in TANGO2 appears to be concentrated in a vulnerable subsystem before spreading to coupled ones, and a scalar model cannot distinguish concentrated risk near threshold from the same total risk spread harmlessly. Section 5.6 tests this claim empirically rather than asserting it.

4.4 Parameter Selection

Parameters appear in Table 2. Only β varies between patients. The shared parameters were chosen by a coordinate wise grid search to reproduce three qualitative features reported in the TANGO2 literature [4, 9]: recurrence on a timescale of months, more frequent crises with greater baseline burden, and, in the deterministic model without noise, no crises for the mildest patients, that is, those with β below the critical amplitude β^* of Section 4.6. Crises in these patients under noise are a model prediction rather than a calibration target (Section 5.5). The parameters were not fitted to patient crisis dates, which the dataset does not contain. The resting level is $r_{\text{rest}} = 0$; it does not affect β^* , which is defined with the reset disabled. The threshold r^* sets the scale of the dimensionless risk; α and δ set the rebuild time after a crisis; D keeps the risk profile spatially structured over the horizon; x_0 and σ place and size the vulnerable

Table 2: Model parameters and their basis. Only β varies between patients, and no parameter was fitted to observed crisis dates.

Symbol	Meaning	Value	Basis
r^*	crisis threshold	0.72	scale of r
κ	retention after crisis	0.52	partial recovery
r_{rest}	resting risk level	0	recovery floor
β	source amplitude	0.0121 to 0.0176	per patient
β^*	critical amplitude	0.01273	derived
α, δ	growth, clearance	0.011, 0.007	crisis timescale
D	diffusion coefficient	0.0012	spatial structure
x_0, σ	locus, width	0.58, 0.06	vulnerable band
σ_η	noise intensity	0.045	timing variability
$N, \Delta t, T$	grid, step, horizon	260, 0.5 d, 900 d	numerics

Algorithm 1 IMEX solver with crisis detection

```

1: Assemble tridiagonal  $A = I - \Delta t DL$  with no flux rows;  $r \leftarrow r_0 \mathbf{1}$ ;  $\mathcal{E} \leftarrow \emptyset$ 
2: for  $n = 0$  to  $T/\Delta t - 1$  do
3:    $g \leftarrow \alpha r \odot (1 - r) + \beta S - \delta r$  ▷ explicit reaction
4:    $\xi \leftarrow \sigma_\eta \sqrt{r} \odot (1 - r) \odot S \odot \Delta W^n$  ▷ noise
5:   Solve  $Ar^+ = r + \Delta t g + \xi$ ;  $r \leftarrow \text{clip}(r^+, 0, 1)$ 
6:   if  $\max_j r_j \geq r^*$  then  $\mathcal{E} \leftarrow \mathcal{E} \cup \{n\Delta t\}$ ;  $r \leftarrow r_{\text{rest}} \mathbf{1} + \kappa(r - r_{\text{rest}} \mathbf{1})$ 
7:   end if
8: end for
9: return crisis times  $\mathcal{E}$ 

```

band; and σ_η gives roughly a week of timing variability between realizations. Severity is the min to max normalized count of retained baseline terms, $\sigma_i = (n_i - n_{\min}) / (n_{\max} - n_{\min})$, mapped linearly to $\beta_i = \beta_{\min} + (\beta_{\max} - \beta_{\min})\sigma_i$. This linear mapping is a modeling assumption, so simulated patient trajectories illustrate the model rather than predict individual outcomes.

4.5 Numerical Scheme

With $N = 260$ nodes ($\Delta x \approx 3.85 \times 10^{-3}$), an explicit scheme would need $\Delta t \leq \Delta x^2 / 2D \approx 6.2 \times 10^{-3}$ days, about 80 times smaller than our step of 0.5 days. We therefore treat diffusion implicitly and reaction explicitly [1]:

$$(I - \Delta t DL) r^{n+1} = r^n + \Delta t [\alpha r^n \odot (1 - r^n) + \beta S - \delta r^n] + \xi^n, \quad (2)$$

where L is the three point Laplacian and ξ^n is the Euler Maruyama noise increment. The system is tridiagonal, so each step costs $O(N)$ with the Thomas algorithm (Algorithm 1).

PROPOSITION 4.1. *The implicit diffusion update is unconditionally stable for every $\Delta t > 0$.*

PROOF. For a Fourier mode $r_j^n = g^n e^{i\theta j}$ the amplification factor is $g(\theta) = [1 + 4\lambda \sin^2(\theta/2)]^{-1}$ with $\lambda = D\Delta t / \Delta x^2 > 0$. The denominator is at least 1, so $0 < g(\theta) \leq 1$ for every θ . $\square$

4.6 Critical Source Amplitude

We define the critical amplitude from the steady state of the deterministic PDE with the reset disabled,

$$\beta^* = \inf\{\beta > 0 : \max_x r_\infty(x; \beta) \geq r^*\}, \quad (3)$$

and bisection gives $\beta^* = 0.01273$. Patients below β^* settle into a bounded state under the threshold, and patients above it cross repeatedly. The calibrated range 0.0121 to 0.0176 contains β^* , so the mildest patients sit just below it.

It is instructive to compare two scalar reductions without diffusion, each with fixed point $r_+ = [(\alpha - \delta) + \sqrt{(\alpha - \delta)^2 + 4\alpha b}] / 2\alpha$ for effective source b . A scalar model of the *peak*, which receives the full source $b = \beta$ with no leakage, has r_+ between 1.25 and 1.46 across the cohort. These fixed points lie above the admissible range $[0, 1]$ of r , which means risk in this model never settles below the threshold; in simulation r is clipped at 1, so every patient crosses. A scalar model of the *average*, with $b = \beta \bar{S}$ where $\bar{S} = 0.150$ is the mean of S , has r_+ between 0.63 and 0.70, so no patient would cross, and its own critical amplitude (0.0188) lies outside the cohort range. The spatial model sits between these extremes because diffusion leaks risk away from a localized peak.

4.7 Model Tests

Three tests address what the model can and cannot show. (i) *Scalar baselines*: both scalar models above were simulated with the same threshold, reset, and noise (noise scaled by $\bar{S}$ for the average model), 200 realizations per amplitude. (ii) *Equal total source*: the width σ was scaled by factors of 0.5 to 2 while β was adjusted to hold the total source $\int \beta S dx$ fixed. Any model of the average depends only on this total and therefore predicts no change, so a change isolates the contribution of spatial concentration. (iii) *Synthetic recovery*: to check that β is identifiable before fitting real data, we built a reference table of mean time to first crisis and mean crisis count on a grid of β values (0.0115 to 0.0185 in steps of 0.00025, 100 realizations each), simulated one independent 900 day timeline for each of 60 synthetic patients with known β drawn uniformly from the cohort range, and estimated β as the grid value minimizing the standardized distance between observed and reference features, with the shared parameters held at their calibrated values.

5 Results

5.1 Classification Performance

On the 30 shared features, the Random Forest reached AUC 0.993 ± 0.002 , accuracy 98.3 percent, sensitivity 95.6 percent, specificity 100 percent, and Brier score 0.017 (Table 3). Averaged across folds, it identified 86 of 90 TDD patients and all 141 controls. Figure 3 shows a representative fold. The patients it missed were the most sparsely annotated TDD records, carrying one to five shared terms and lacking the full neurological triad, so the misses reflect missing documentation rather than a rejected signature. Logistic regression and the support vector machine reached AUCs of 0.990 and 0.992 on the same features.

5.2 Stress Tests of the Classifier

An AUC near 0.99 with perfect specificity warrants deliberate attempts to break it (Figure 4, Table 4).

Permutation. Across 1,000 label shuffles, the cross validated AUC averaged 0.494 (SD 0.051) and never exceeded 0.689, against 0.993 with true labels ($p = 0.001$, the smallest value attainable with 1,000 permutations). A dataset that is merely small and easy to separate

Table 3: Repeated stratified tenfold cross validation (10 repeats) on the 30 shared features. The corrected interval accounts for overlap between folds.

Model	AUC	Sens.	Spec.	Brier
Random Forest	0.993	95.6%	100%	0.017
Logistic regression	0.990	96.2%	100%	0.013
SVM (radial kernel)	0.992	98.4%	100%	0.010
Random Forest AUC per repeat: 0.992 to 0.996				
Corrected 95% interval for AUC [11]: 0.980 to 1.000				

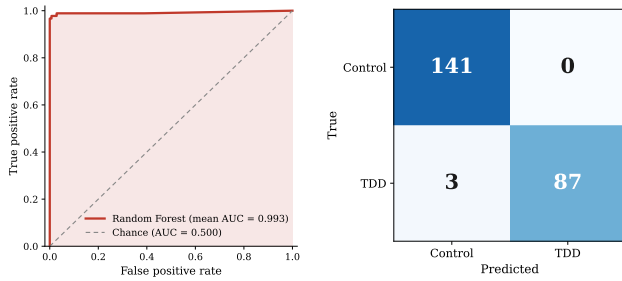


Figure 3: Left: ROC curve (legend reports the repeated cross validation mean AUC). Right: confusion matrix for a representative fold, with 87 of 90 TDD patients and all 141 controls classified correctly.

by chance would produce high AUCs under shuffled labels, and this one does not.

Uncertainty. The 10 repeats gave AUCs from 0.992 to 0.996, so the result is not an artifact of one favorable partition. Because folds overlap, the naive interval understates uncertainty; the corrected interval is wider, 0.980 to 1.000, and we treat it as the honest range.

Feature choices. Changing the frequency threshold to two or five patients gave AUCs of 0.994 and 0.994. A stricter set keeping only terms recorded in at least three patients *within each cohort*, which removes terms the controls’ team used only rarely, retained 18 features and gave 0.994. Removing the six most important features left 0.993; removing eight lowered it to 0.980.

Leakage checks. The number of terms per record classifies at AUC 0.569 on its own, near chance, so annotation depth is not the signal. All 231 phenotype sets are unique, with none shared across cohorts.

Specificity. Zero false positives among 141 controls does not imply a zero rate. By the rule of three, the one sided 95 percent upper bound on the false positive rate is about $3/141 \approx 2.1$ percent. A leave one team out validation, suggested by a reviewer, is not possible here because each cohort was annotated by exactly one team, so team and label are identical; the strict vocabulary set above is the closest available check, and Section 8 describes the external test that would settle it.

5.3 Calibration

The Brier score of 0.017 compares with 0.238 for a model that always predicts the base rate, a skill of 0.93. The calibration curve lies

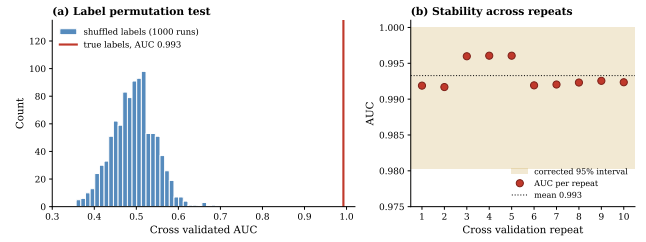


Figure 4: Stress tests. (a) Cross validated AUC under 1,000 label permutations (blue) versus true labels (red). (b) AUC for each of the 10 repeats, with the variance corrected 95 percent interval (shaded).

Table 4: Sensitivity of the classifier to feature construction (mean AUC over 5 repeats for alternative sets).

Feature set	Features	AUC
All 261 candidates (no vocabulary restriction)	261	0.996
Shared vocabulary including ECG term	31	0.994
Primary: shared vocabulary, ECG removed	30	0.993
Frequency threshold of 2 patients	30	0.994
Frequency threshold of 5 patients	28	0.994
At least 3 patients in <i>each</i> cohort	18	0.994
Primary minus 6 most important	24	0.993
Primary minus 8 most important	22	0.980
Record length only	1	0.569

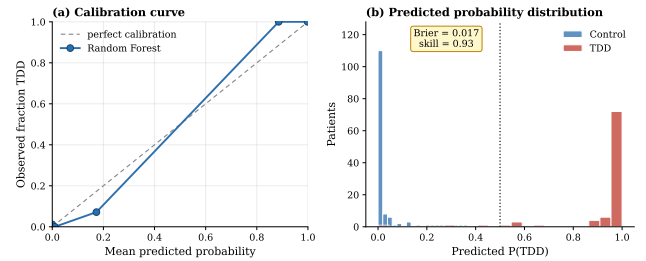


Figure 5: Calibration. (a) Reliability curve, quantile binned. (b) Predicted probabilities by true class.

near the diagonal and predicted probabilities are strongly bimodal (Figure 5), which supports the fixed 0.5 threshold. These probabilities reflect the 39 percent TDD share of this study, not the far lower prevalence of a general clinic, a point we return to in Section 6.

5.4 Candidate Markers and a Neurological Triad

All phenotypes are structured HPO annotations rather than free text. Paroxysmal lethargy (82.2 percent of TDD patients) and compensatory head posture (70.0 percent) were never recorded in the 141 controls. Because only the TDD annotating team used these terms, the reasoning of Section 3.2 applies to them directly: their absence among controls may reflect differences in annotation vocabulary rather than biology. We therefore treat them as hypotheses about specificity that this dataset cannot confirm, to be tested

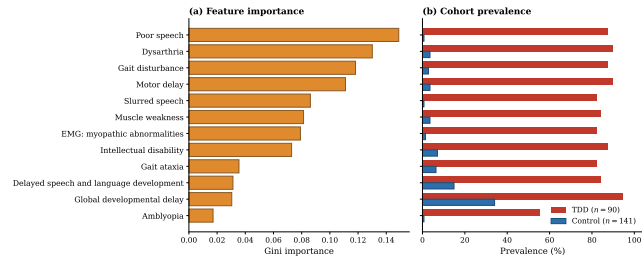


Figure 6: (a) Gini importance of the twelve most informative shared phenotypes. (b) Their prevalence in TDD ($n = 90$) and control ($n = 141$) patients.

Table 5: The ten most informative shared phenotypes.

Baseline phenotype	TDD (%)	Control (%)	Importance
Poor speech	87.8	0.7	0.149
Dysarthria	90.0	3.5	0.130
Gait disturbance	87.8	2.8	0.118
Motor delay	90.0	3.5	0.111
Slurred speech	82.2	0.7	0.086
Muscle weakness	84.4	3.5	0.081
EMG: myopathic abnormalities	82.2	1.4	0.079
Intellectual disability	87.8	7.1	0.073
Gait ataxia	82.2	6.4	0.036
Delayed speech and language	84.4	14.9	0.031

when TDD patients and controls are annotated by the same team (Section 8.2). They are also only moderately sensitive, since head posture is absent in about a third of TDD patients.

The neurological triad is the more sensitive signal. Dysarthria and motor delay each appear in 81 of 90 TDD patients (90.0 percent) versus 3.5 percent of controls, gait disturbance in 87.8 versus 2.8 percent, and intellectual disability in 87.8 versus 7.1 percent (Figure 6, Table 5). Gini importance rewards features that are common in cases and rare in controls, so these phenotypes drive the classifier’s decisions, and their combined presence is what the ensemble exploits. Unlike the candidate markers, every one of them was recorded by both annotating teams.

5.5 Episodic Crises and the Critical Amplitude

Simulated trajectories reproduce the qualitative course of the disease: risk accumulates near the vulnerable locus, fluctuates, crosses the threshold, is partially reset, and rebuilds (Figure 7). Over 200 realizations per stratum, a high burden patient ($\beta = 0.0167$) averages 4.79 ± 0.40 crises in 900 days with the first near day 229, a moderate patient ($\beta = 0.0158$) averages 4.00 with the first near day 256, and a mild patient ($\beta = 0.0139$) averages 2.76 ± 0.43 with the first near day 346. Across the 90 patients, severity correlates with simulated crisis count ($r = 0.97$) and inversely with time to first crisis ($r = -0.90$). Because severity sets β by assumption, these correlations describe how the model behaves rather than confirming a clinical relationship.

Figure 8 shows the steady state peak against β . Below β^* no crisis occurs without noise, which is the calibration target of Section 4.4.

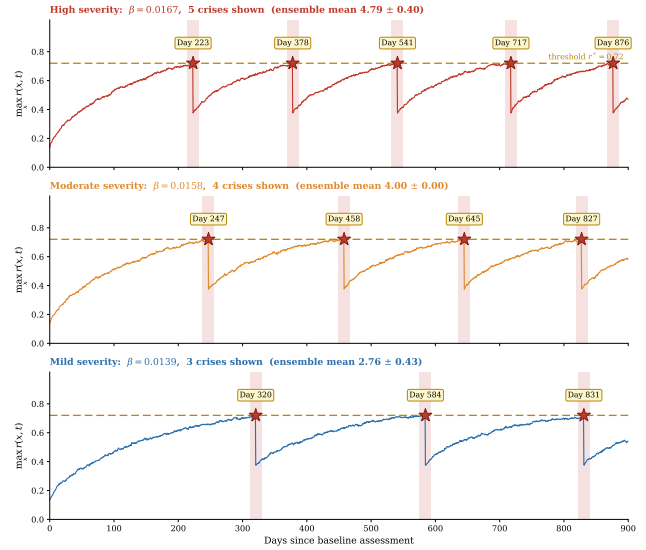


Figure 7: Single stochastic realizations of $\max_x r(x, t)$ for three severity strata, with the ensemble mean and standard deviation over 200 realizations in each title. Stars mark simulated crises and the dashed line is $r^* = 0.72$.



Figure 8: (a) Steady state peak risk versus β , with $\beta^* = 0.01273$ separating quiescent (green) and recurrent (red) regimes. (b) Mean simulated crises in 900 days versus β (bars: SD over 120 realizations).

With noise, a patient at $\beta = 0.0121$ still has at least one crisis in 83 percent of realizations, and each of those realizations has exactly one crisis, so the mean of 0.83 crises in 900 days equals that fraction, compared with 4.79 for a high burden patient. The calibration target is therefore met in the deterministic model, and the stochastic model adds a prediction beyond it: a low baseline burden makes crises rare rather than impossible. Above β^* , crisis count rises smoothly with no chaotic regime, so predictions degrade gradually under error in β .

5.6 What the Spatial Term Adds

Figure 9a compares the PDE with the two scalar baselines under identical parameters. The scalar model of the peak produced 26.9 to 40.9 crises in 900 days, roughly one every three to five weeks for every patient. The scalar model of the average produced 0.0 to 1.4, so even the most severe patients would average only about

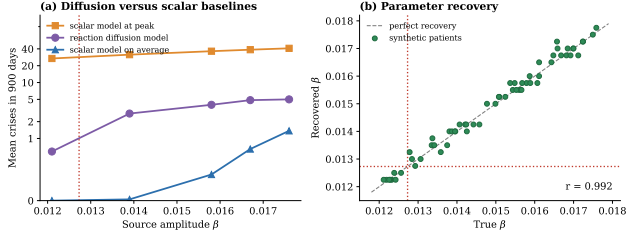


Figure 9: (a) Mean crises in 900 days for the PDE and two scalar models without diffusion, identical parameters, 200 realizations each (dotted line: β^*). (b) Synthetic recovery of β from one simulated 900 day timeline per patient.

one crisis every two years. The mildest patients having few or no crises is not the shortcoming here, since the PDE behaves the same way deterministically; the shortcoming is that crises are far too rare across the whole cohort. The PDE produced 0.8 to 5.0, on the timescale of months reported clinically. Neither scalar reduction reproduces that behavior without recalibration, and each gives an essentially all or none answer across the cohort.

The equal total source test isolates the specific contribution of diffusion. Holding $\int \beta S dx$ fixed while scaling the width by 0.5, 0.75, 1, 1.5, and 2 changed the PDE's mean crisis count to 4.5, 4.0, 4.0, 3.4, and 3.0. A scalar model of the average cannot register this change, because it sees only the total. The PDE therefore makes a distinct, testable claim: patients with equally large but more concentrated vulnerability should decompensate more often. We note that a scalar model could be recalibrated to place a threshold inside the cohort range; what it cannot represent is this dependence on concentration.

5.7 Identifiability of the Patient Parameter

From a single simulated 900 day timeline per synthetic patient, the recovered β correlated with the true value at $r = 0.992$ (Figure 9b), with mean absolute error 1.6×10^{-4} , median relative error 0.9 percent, and negligible bias. The regime, above or below β^* , was classified correctly for all 60 patients. This shows that β is recoverable from the kind of data a longitudinal study would collect, provided the shared parameters are known; jointly estimating the shared parameters would require data from many patients and is left to future work.

5.8 Numerical Verification and Sensitivity

Both resolutions were refined sixteenfold: the grid from $N = 65$ to 1,040 against a reference of $N = 2,080$, and the time step from 0.5 to 0.03125 days against a reference of 0.0039 days. Error fell monotonically at approximately first order in both, with observed temporal orders of 1.01 to 1.10, as expected for a first order time integrator and a first order boundary treatment. With reaction and source removed, total mass was conserved to a relative drift of 1.3×10^{-12} , and the measured amplification factor was 1.000000, matching the proposition.

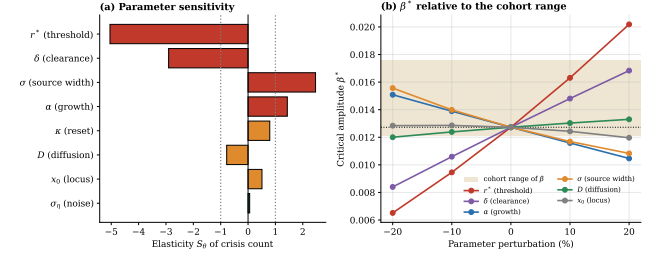


Figure 10: (a) Elasticity of crisis count for each parameter (red sensitive, orange moderate, green robust). (b) Critical amplitude β^* under 10 and 20 percent perturbations of each parameter that affects it; the shaded band is the cohort range of β , so points outside it place the entire cohort on one side of β^* . The reset κ and noise σ_η do not affect β^* .

Figure 10 reports, for each parameter θ , the elasticity of the crisis count n ,

$$S_\theta = \frac{\Delta n/n}{\Delta \theta/\theta}. \quad (4)$$

The model is most sensitive to the threshold r^* ($S = -5.0$) and clearance δ ($S = -2.9$), which encode the core assumptions, and nearly insensitive to the noise intensity σ_η ($S = 0.06$). The locus x_0 has a moderate effect on crisis count ($S = 0.51$), but moving it by 10 percent shifts β^* by at most 2.3 percent, so it barely changes which patients fall in each regime. The reset κ leaves β^* exactly unchanged, as it must, since a reset cannot alter a steady state. A critical amplitude exists under every perturbation tested, and crisis count still rises with β , but its position relative to the cohort is sensitive. Increasing r^* by 20 percent raises β^* to 0.0202, above the entire cohort range of 0.0121 to 0.0176, so no patient would cross the threshold; decreasing r^* by 20 percent lowers it to 0.0065, below the entire range, so every patient would. Some 10 percent changes also move β^* below the cohort range, namely lowering r^* or δ or raising α or σ , and lowering D or raising x_0 by 20 percent moves it just below (Figure 10b). Which patients the model classifies as quiescent therefore depends directly on the calibrated values of r^* , δ , α , and σ , and should not be interpreted for individual patients until these are estimated from data.

6 Clinical Application

The classifier suggests a two tier referral rule (Figure 11) that needs no new laboratory infrastructure. **Tier 2** is the operative rule: a child with the neurological triad, led by dysarthria, would receive a genetic panel that includes *TANGO2*. It rests on phenotypes recorded by both annotating teams and catches most cases, because nearly all TDD patients show the triad. **Tier 1** is conditional: a child with developmental delay who shows paroxysmal lethargy or compensatory head posture would be referred directly for *TANGO2* sequencing, but only if the external test of Section 8.2 confirms that these markers remain specific when TDD patients and controls are annotated by the same team. Until then, Tier 1 is a hypothesis and should not be used to route patients. The design reflects an asymmetry: a false positive costs a genetic test, while a missed diagnosis can cost a preventable crisis.

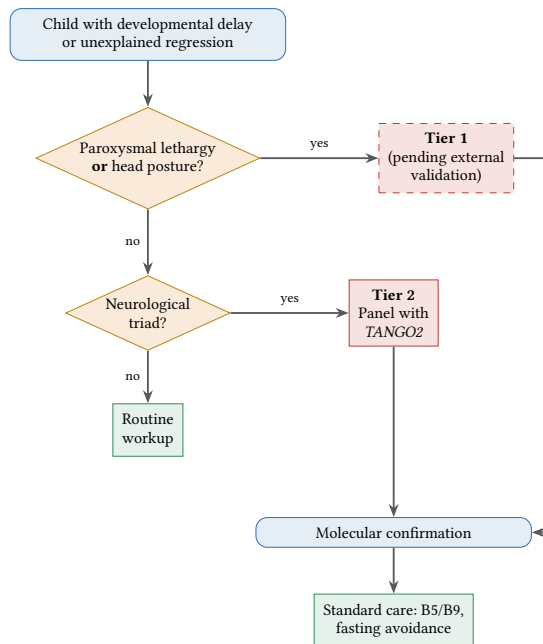


Figure 11: Proposed two tier referral rule for evaluation. Tier 2 uses the neurological triad, recorded by both annotating teams. Tier 1 (dashed) uses the candidate markers and applies only if the external test of Section 8.2 confirms their specificity. Diagnosis always rests on molecular confirmation.

Two cautions shape how the rule should be read. First, predicted probabilities reflect a study in which 39 percent of patients have TDD; in a general clinic the prevalence is far lower. Illustratively, if 1 in 1,000 children with developmental delay had TDD, then even at the 97.9 percent specificity lower bound most positive flags would be false positives. The output is therefore a trigger for confirmatory genetic testing, never a diagnosis, and probabilities would need recalibration to local prevalence. Second, the rule is a proposal derived from cohort statistics. It has been shaped by feedback from collaborating geneticists, but before any clinical use it requires formal review by clinical experts in TANGO2 and prospective evaluation. The PDE component is for research only and is not intended to guide individual care.

7 Ethics and Data Governance

This study is a secondary analysis of deidentified records. The data were provided by Baylor College of Medicine under a data use agreement covering their use for this analysis. No institutional review board approval or exemption determination was obtained for this specific analysis. The data were deidentified by the data custodian before being shared with the author, and the custodian approved publication of the results. The author received only HPO term lists with study identifiers, sex, vital status, and age at evaluation; no names, birth dates, medical record numbers, or contact details were accessed, the author had no access to any key linking records to individuals, and there was no contact with patients or families. Under the U.S. Common Rule, research using information

that the investigator cannot readily link to identifiable individuals does not meet the regulatory definition of human subjects research [3]. The original records were collected by the source programs at Baylor College of Medicine and the Undiagnosed Diseases Network under their own institutional review board oversight.

Because the TDD cohort is small and includes deceased individuals, reidentification risk is not zero even for deidentified data. We therefore report only aggregate statistics, describe individual patients only in aggregate terms, and do not release patient level records; the public code runs on user supplied data and contains no patient information.

8 Discussion

The classification result shows that the information needed to suspect TDD is present before the first crisis. This was not guaranteed: the stable phenotype could have been indistinguishable from other developmental disorders, with discriminating information arriving only with the crisis. Instead, a model restricted to shared, pre-crisis phenotypes separates TDD from a difficult control cohort, the result survives permutation, variance correction, stricter feature sets, ablation, and alternative models, and the signature is spread across many correlated findings rather than resting on a few. The most directly usable outcome is the neurological triad, recorded by both annotating teams, which could prompt earlier genetic testing; the two candidate markers are promising but require external confirmation first.

The PDE result is different in kind. It is a mechanistic hypothesis: episodic crises whose frequency rises with severity can arise from slow, continuous accumulation of risk crossing a threshold, without assuming that crises are random events. The model makes testable predictions, namely a critical amplitude dividing quiescent from recurrent patients, noise induced crises just below that amplitude, and a dependence on how concentrated a patient's vulnerability is, which scalar models cannot express. Whether these predictions hold in patients remains to be determined.

8.1 Scope and Limitations

The classifier was evaluated on 231 specialist referred patients from two cohorts, so performance in an unselected population will be lower and requires prospective confirmation. Because each cohort was annotated by one team, annotation style and diagnosis cannot be fully separated in this data, even with the shared vocabulary safeguards. This applies most strongly to the two candidate markers, which only one team recorded, and in weaker form to shared terms with extreme prevalence contrasts such as poor speech (87.8 percent of TDD patients versus 0.7 percent of controls): both teams used the term, but a difference in how often each team recorded it could still widen the contrast. The position of β^* relative to the cohort also depends on calibrated parameters (Section 5.8). The PDE uses an abstract coordinate, a linear severity mapping, and parameters calibrated to qualitative patterns rather than fitted to crisis dates, so its simulated timelines illustrate mechanisms rather than forecast individual patients. The recovery test establishes identifiability of β only when the shared parameters are known.

8.2 Future Work

The most decisive next test is external: TDD patients diagnosed within the UDN are annotated by the same team as the controls, so classifying them would directly separate disease signal from annotation style. It would also test whether paroxysmal lethargy and compensatory head posture remain specific when recorded by the same team as the controls, which is the condition for adopting Tier 1. Prospective use of the referral rule at first developmental assessment would measure how many cases are found before a first crisis. For the PDE, longitudinal crisis dates would allow each β_i to be fitted with the recovery procedure validated here, the model to be scored against held out crises, and the prediction that B complex supplementation lowers effective β to be tested directly.

9 Conclusion

Restricting a Random Forest to phenotypes observable before any metabolic crisis, and to vocabulary shared by both annotating teams, separated TANGO2 deficiency from a phenotypically overlapping cohort with an AUC of 0.993 (corrected interval 0.980 to 1.000), a result that withstood a 1,000 shuffle permutation test, stricter feature sets, and alternative models. A highly prevalent neurological triad recorded by both annotating teams suggests a practical referral rule for prospective evaluation, and two candidate markers recorded only in TDD patients are hypotheses awaiting an external test. A stochastic reaction diffusion model, verified numerically, reproduces the episodic course of the disease, identifies a critical amplitude separating quiescent and recurrent regimes, captures an effect of vulnerability concentration that scalar models cannot represent, and has an identifiable patient parameter, preparing it for testing against longitudinal data.

Code Availability

Analysis code and a research prototype interface are available at <https://github.com/abhaytippi/TANGO2-Forecast>. The prototype is intended for research and education, has not been validated for clinical use, and must not be used to make decisions about individual patients.

Acknowledgments

The author thanks Dr. Seema Lalani and Dr. Zhandong Liu of Baylor College of Medicine for providing the deidentified cohorts, for clinical and scientific guidance, and for mentorship, and thanks the reviewers, whose suggestions led to the permutation, baseline, and recovery analyses reported here.

References

- [1] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton. 1995. Implicit explicit methods for time dependent partial differential equations. *SIAM Journal on Numerical Analysis* 32, 3 (1995), 797–823.
- [2] L. Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.
- [3] U.S. Department of Health and Human Services. 2018. Protection of Human Subjects, 45 CFR 46.102, Definitions. *Code of Federal Regulations*.
- [4] J. N. Dines, K. Golden-Grant, A. LaCroix, et al. 2019. TANGO2: expanding the clinical phenotype and spectrum of pathogenic variants. *Genetics in Medicine* 21, 3 (2019), 601–607.
- [5] R. A. Fisher. 1937. The wave of advance of advantageous genes. *Annals of Eugenics* 7, 4 (1937), 355–369.

- [6] S. Köhler, L. Carmody, N. Vasilevsky, et al. 2019. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research* 47, D1 (2019), D1018–D1027.
- [7] A. N. Kolmogorov, I. G. Petrovsky, and N. S. Piskunov. 1937. A study of the diffusion equation with increase in the amount of substance. *Bulletin of Moscow University, Mathematics Series A* 1 (1937), 1–25.
- [8] L. S. Kremer, F. Distelmaier, B. Alhaddad, et al. 2016. Biallelic truncating mutations in TANGO2 cause infancy onset recurrent metabolic crises with encephalocar-diomyopathy. *American Journal of Human Genetics* 98, 2 (2016), 358–362.
- [9] S. R. Lalani, P. Liu, J. A. Rosenfeld, et al. 2016. Recurrent muscle weakness with rhabdomyolysis, metabolic crises, and cardiac arrhythmia due to biallelic TANGO2 mutations. *American Journal of Human Genetics* 98, 2 (2016), 347–357.
- [10] J. D. Murray. 2002. *Mathematical Biology I: An Introduction* (3rd ed.). Springer.
- [11] C. Nadeau and Y. Bengio. 2003. Inference for the generalization error. *Machine Learning* 52, 3 (2003), 239–281.
- [12] M. Ojala and G. C. Garriga. 2010. Permutation tests for studying classifier performance. *Journal of Machine Learning Research* 11 (2010), 1833–1863.
- [13] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. 2011. Scikit learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [14] A. Quarteroni, R. Sacco, and F. Saleri. 2007. *Numerical Mathematics* (2nd ed.). Springer.
- [15] R. B. Ramoni, J. J. Mulvihill, D. R. Adams, et al. 2017. The Undiagnosed Diseases Network: accelerating discovery about health and disease. *American Journal of Human Genetics* 100, 2 (2017), 185–192.