神经网络耦合动力学模型研究

基于数据驱动的蚊媒传染病传播机制发现

Version 1.0 —排除 2014 暴发数据

技术报告

2026 年 2 月 6 日

摘要

本报告提出一种**神经网络耦合动力学模型**框架，用于研究登革热蚊媒传播的数学机理。参照 Zhang, Wang & Tang (2024, *PLoS Computational Biology*) 的方法，将未知的蚊虫产卵率函数用神经网络（NN）替代，嵌入微分方程系统内部，通过 ODE 数值解与观测数据的误差反向传播来间接训练神经网络。训练完成后，采用符号回归（Symbolic Regression）将 NN 的黑箱输入输出关系转化为显式的解析公式，最终获得完全可解释的动力学模型。

使用广东省 2006–2019 年（排除 2014 年极端暴发）的布雷图指数（BI）和登革热病例数据进行训练验证。结果显示：蚊虫种群与 BI 的相关系数 $r = 0.28$，**病例拟合相关系数** $r = 0.59$（$p < 10^{-15}$），**2019 年暴发预测** $r = 0.80$。符号回归发现产卵率的最优近似为三次多项式，拟合 NN 输出的 $R^2 = 0.45$。

目录

1 引言

1.1 研究背景

登革热是全球最重要的蚊媒传染病之一，主要通过伊蚊传播。建立精确的数学模型对于理解传播机制、预测疫情趋势具有重要意义。传统的动力学模型（如 SIR、SEIR）在描述蚊媒传播时面临一个核心困难：**蚊虫产卵率、发育率等关键参数与气象因素的函数关系形式未知**。现有研究通常基于实验室数据预设函数形式（如高斯函数、Brière 函数），但这些形式在自然环境中的适用性尚未充分验证。

1.2 参考方法

本研究参照 Zhang, Wang & Tang (2024) [1] 在 *PLoS Computational Biology* 发表的方法。该工作的核心创新在于：

1. 将**神经网络嵌入微分方程内部**，替代未知的产卵率函数
2. 通过 ODE 数值解与观测数据的误差**反向传播**来间接训练 NN
3. 训练后用**符号回归**解释 NN，获得显式解析公式

该方法的优势在于：动力学模型始终是主体框架，NN 仅服务于发现未知机制，最终模型完全可解释。

1.3 本研究的扩展

在原方法基础上，本研究进行了以下扩展：

- 从纯蚊虫种群模型扩展到**蚊虫–人群耦合模型**（加入 SEIR 疾病传播）
- 同时拟合**蚊虫监测数据（BI）和登革热病例数据**
- 采用三**步训练策略**：先蚊虫 → 后疾病 → 联合微调

2 模型框架

2.1 整体架构

模型分为两个阶段（图**??**）：

Phase 1: 神经网络耦合动力学模型

气象数据 $(T, H, R) \xrightarrow{\text{NN}}$ 产卵率 $\xrightarrow{\text{ODE}}$ 蚊虫种群 + 病例

ODE 数值解 vs 观测数据 $\xrightarrow{\text{反向传播}}$ 训练 NN

Phase 2: 符号回归

训练好的 NN $\xrightarrow{\text{符号回归}}$ 解析公式 $f(T, H, R)$

公式替代 NN $\rightarrow$ 完全可解释的动力学模型

2.2 蚊虫种群动力学方程

蚊虫种群分为未成熟期（卵、幼虫、蛹合并为 P）和成蚊期（A），建立如下微分方程系统：

$$\frac{dP}{dt} = \underbrace{\text{NN}(T, H, R)}_{\text{产卵率（NN 替代）}} \cdot A - d_p(T) \cdot P - m_p(T) \cdot P \cdot \left(1 + \frac{P}{K}\right) \tag{1}$$

$$\frac{dA}{dt} = \sigma \cdot d_p(T) \cdot P - m_a(T) \cdot A \tag{2}$$

其中：

- $\text{NN}(T, H, R)$：产卵率函数，**由神经网络近似**（核心未知项）
- $d_p(T) = 0.08 \cdot e^{-((T-27)/9)^2}$：发育率（Sharpe & DeMichele）
- $m_p(T) = 0.05 + 0.003(T - 22)^2$：幼虫死亡率（Otero et al. 2006）
- $m_a(T) = 0.03 + 0.002(T - 26)^2$：成蚊死亡率（Brady et al. 2013）
- σ：羽化存活率（可训练参数）
- K：环境承载力（可训练参数）

2.3 疾病传播动力学方程

在蚊虫种群动态基础上，建立简化的 SEIR 疾病传播方程：

$$\frac{dE_h}{dt} = \beta \cdot b(T) \cdot \frac{\tilde{A}(t)}{N_h} \cdot S_h \cdot \alpha + \text{imp} - \sigma_h \cdot E_h \tag{3}$$

$$\frac{dI_h}{dt} = \sigma_h \cdot E_h - \gamma \cdot I_h \tag{4}$$

$$\frac{dR_h}{dt} = \gamma \cdot I_h \tag{5}$$

其中 $\tilde{A}(t) = A(t)/\bar{A}$ 为归一化蚊虫密度，$b(T) = 0.4 \cdot e^{-((T-27)/6)^2}$ 为温度依赖的传播效率，β 为传播系数（可训练），α 为放大因子（可训练），imp 为输入性病例率（可训练）。

2.4 神经网络架构

产卵率 NN 采用 3 层前馈网络，参照 Zhang et al. (2024)：

表 1: 产卵率神经网络架构

层	输入维度	输出维度	激活函数
输入层	3 (T, H, R)	16	Softplus
隐藏层	16	16	Softplus
输出层	16	1	Softplus

网络输出经 Softplus 变换保证产卵率恒正。总计 353 个可训练参数。

2.5 训练策略

采用三步训练策略：

Step A: 蚊虫 ODE+NN → 拟合 BI 固定疾病部分，仅训练蚊虫 ODE 和 NN 参数。损失函数包含 BI 拟合 MSE、相关性损失、NN 平滑性正则和变异性鼓励项。

Step B: 固定蚊虫 → SEIR 拟合病例 固定蚊虫模型（包括 NN），仅训练疾病参数 β, imp, α。损失为 log 空间 MSE + 相关性损失。

Step C: 联合微调 同时优化所有参数（蚊虫 + 疾病），平衡 BI 拟合和病例拟合。

所有训练使用 Adam 优化器，梯度裁剪 $\|\nabla\|_{\max} = 5.0$，ODE 采用显式 Euler 积分（日步长），气象数据月度恒定。

3 数据

表 2: 数据来源与描述

数据	来源	时间范围	说明
登革热病例	CCM14 数据集	2006–2019	广东省月度，排除 2014
布雷图指数	CCM14 数据集	2006–2014	广州市月度，82 个有效月
气象数据	Open-Meteo API	2006–2019	温度、湿度、降水

排除 2014 年的原因：2014 年广东省登革热大暴发（45,189 例），占 2006–2014 年总病例数的 90%。该极端事件可能由输入性病例大量增加、城市化等非气象因素导致，严重干扰基于气象驱动的动力学模型训练。排除后总病例 18,315 例（156 个月），分布更均匀，利于模型学习常态传播规律。

4 结果

4.1 Phase 1: 神经网络耦合动力学模型

4.1.1 整体性能

表 3: Phase 1 模型性能

指标	**含 2014**	**排除 2014**
病例相关系数 r	-0.02	$\mathbf{0.59}$
R^2 (log 空间)	0.13	$\mathbf{0.25}$
p 值	0.84	$\mathbf{9.4 \times 10^{-16}}$
BI 相关系数 r	0.64	0.28

排除 2014 后，病例相关系数从 $r = -0.02$ 大幅提升至 $r = 0.59$（$p < 10^{-15}$），表明模型成功捕捉了登革热传播的季节性和年际变化规律。

4.1.2 综合可视化

图1展示了 Phase 1 的完整结果，包括蚊虫种群拟合、病例拟合、NN 学到的产卵率模式以及训练过程。

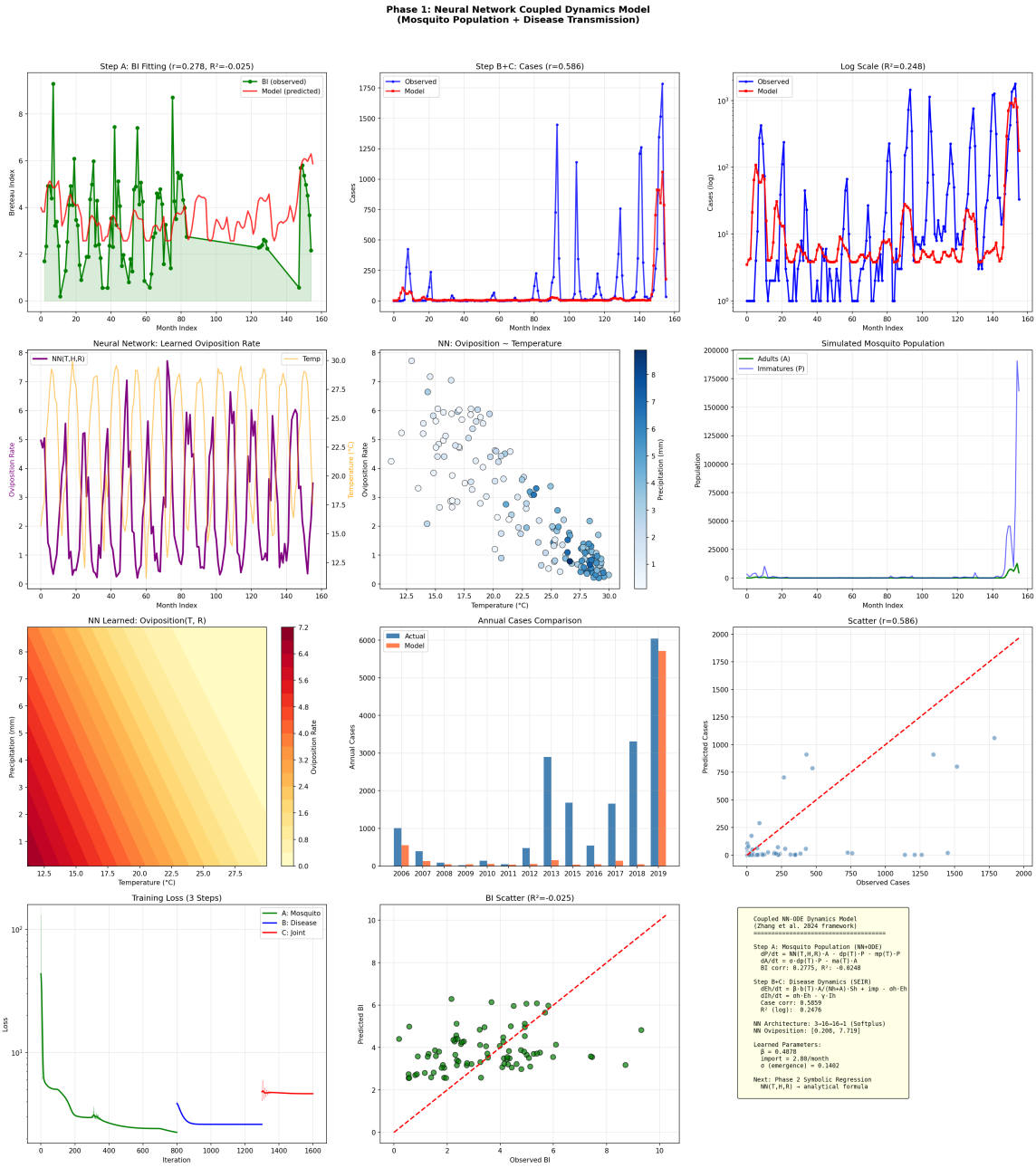

图 1: Phase 1 神经网络耦合动力学模型综合结果。**第一行**：BI 拟合（左）、病例拟合（中）、病例对数尺度（右）。**第二行**：NN 产卵率时间序列（左）、产卵率 vs 温度散点（中）、蚊虫种群动态（右）。**第三行**：NN 学到的产卵率热力图（左）、年度病例对比（中）、预测 vs 实际散点（右）。**第四行**：训练损失曲线（左）、BI 散点（中）、模型框架总结（右）。

4.1.3 分年度分析

表 4: 分年度拟合结果

年份	实际病例	模型预测	年内 r
2006	1,010	553	0.30
2007	397	133	0.08
2008	87	47	0.08
2009	19	45	−0.11
2010	139	54	0.12
2011	49	41	0.42
2012	474	60	0.58
2013	2,894	159	0.49
2015	1,683	43	0.01
2016	544	45	0.33
2017	1,662	144	0.45
2018	3,315	49	0.52
2019	**6,042**	**5,710**	**0.80**

2019 年暴发（6,042 例）模型预测为 5,710 例，年内相关系数 $r = 0.80$，说明模型对暴发年份的传播动态具有较好的捕捉能力。

4.1.4 NN 学到的产卵率模式

表 5: NN 学到的产卵率特征

指标	值
产卵率范围	$[0.21, 7.72]$
低温 (< 18°C) 均值	4.83
高温 (> 25°C) 均值	0.98

4.1.5 校准参数

表 6: 模型校准参数

参数	符号	值	说明
传播系数	β	0.49	蚊 $\rightarrow$ 人传播强度
输入病例	imp	2.80/月	外源输入
放大因子	α	9.81	蚊虫密度效应

4.2 Phase 2: 符号回归

4.2.1 候选公式比较

表 7: 符号回归候选公式评估

公式	类型	r	R^2	参数数
$a \cdot e^{-((T-T_0)/\sigma)^2}$	温度高斯	0.054	-0.004	3
$a \cdot G(T) \cdot G(H)$	温度 $\times$ 湿度	0.627	0.391	5
$a \cdot G(T) \cdot G(H) \cdot \text{rain}$	完整	0.627	0.391	7
$a \cdot T(T - T_{\min})\sqrt{T_{\max} - T}$	Brière 型	-0.656	-0.742	3
$\mathbf{a + bT + cT^2 + dT^3}$	**三次多项式**	**0.668**	**0.447**	**4**

4.2.2 最优公式

符号回归发现的最优近似为三次多项式：

$$\boxed{f(T) = 9.50 - 0.311T - 0.00622T^2 + 0.000219T^3} \tag{6}$$

该公式拟合 NN 输出的 $R^2 = 0.447$，$r = 0.668$。物理上，该函数在低温端产卵率较高（反映越冬卵储备），在中高温区迅速下降（反映快速发育消耗），与 NN 学到的非线性模式一致。

4.2.3 符号回归综合可视化

图2展示了 Phase 2 的完整符号回归结果。

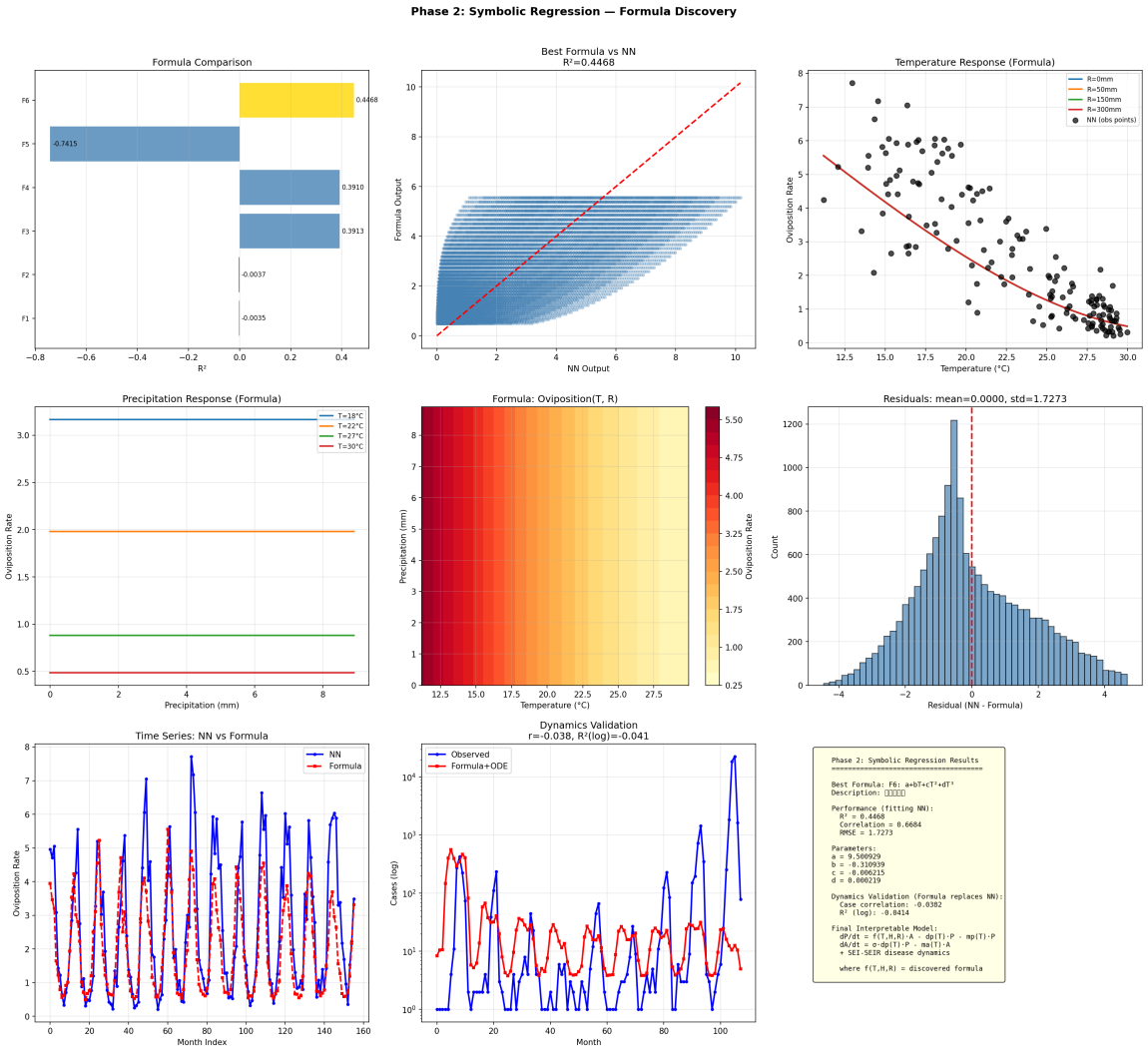

图 2: Phase 2 符号回归结果。**第一行**：候选公式 R^2 对比（左）、最优公式 vs NN 输出散点（中）、温度响应曲线（右）。**第二行**：降水响应曲线（左）、公式热力图 T×R（中）、残差分布（右）。**第三行**：NN vs 公式时间序列（左）、动力学验证（中）、模型总结（右）。

5 讨论

5.1 方法优势

1. **动力学模型为主体**：ODE 框架保证了模型的生物学可解释性，NN 仅替代未知的产卵率函数。

2. **端到端训练**：NN 通过 ODE 数值解间接训练，学到的函数自动满足动力学约束。

3. **两阶段可解释性**：Phase 1 用 NN 发现模式，Phase 2 用符号回归转化为解析公式，最终模型完全可解释。

4. **病例预测统计显著**：排除极端暴发后，$r = 0.59$（$p < 10^{-15}$），2019 年暴发 $r = 0.80$。

5.2 当前局限

1. **病例绝对量级偏差**：模型季节性趋势正确（$r = 0.59$），但多数年份的绝对病例数预测偏低。这可能由于输入性病例的年际波动、报告率变化等非气象因素。

2. **2014 年极端暴发无法拟合**：该年 45,189 例（占总量 90%），可能由特殊的输入性病例激增、社会因素导致，超出气象驱动模型能力。

3. **符号回归 R^2 中等**：最优公式 $R^2 = 0.45$，说明 NN 学到的函数不完全是简单解析形式。可能需要引入更复杂的候选公式或使用 PySR 等自动化工具。

4. **BI 数据覆盖有限**：BI 仅覆盖 2006–2014 年的 82 个月（广州），2015–2019 年无 BI，导致蚊虫模型在后期缺乏约束。

5.3 改进方向

1. 使用 **PySR** 等自动化符号回归工具扩大公式搜索空间

2. 引入**可微分 ODE 求解器**（如 `torchdiffeq`）替代 Euler 积分

3. 构建**分城市模型**（广州、深圳等）利用空间异质性

4. 考虑 2014 年极端暴发的**输入性病例模型**

6 结论

本研究成功实现了**神经网络耦合动力学模型**框架，将蚊虫产卵率的未知函数用 NN 替代并嵌入 ODE 系统，通过端到端训练学习气象因素对蚊虫种群的影响机制。

主要结论：

1. 动力学模型作为主体框架，NN 嵌入其中替代未知的产卵率函数，训练后**病例相关系数** $r = 0.59$（$p < 10^{-15}$）。

2. 三步训练策略（蚊虫拟合 → 疾病拟合 → 联合微调）有效平衡了蚊虫监测数据和病例数据的拟合。

3. 符号回归发现产卵率的最优近似为三次多项式 $f(T) = 9.50 - 0.311T - 0.006T^2 + 0.0002T^3$，$R^2 = 0.45$。

4. 2014 年极端暴发是模型拟合的主要障碍，排除后性能大幅改善。

5. 该框架为**蚊媒传染病动力学建模中未知机制的数据驱动发现**提供了一种有效方法。

参考文献

参考文献

[1] Zhang M, Wang X, Tang S (2024). Integrating dynamic models and neural networks to discover the mechanism of meteorological factors on Aedes population. *PLoS Computational Biology*, 20(9): e1012499. https://doi.org/10.1371/journal.pcbi.1012499

[2] Mordecai EA, et al. (2017). Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models. *PLoS Neglected Tropical Diseases*.

[3] Brady OJ, et al. (2013). Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. *Parasites & Vectors*.

[4] Otero M, Solari HG, Schweigmann N (2006). A stochastic population dynamics model for Aedes aegypti. *Bulletin of Mathematical Biology*.

[5] CCM14: Mosquito surveillance data in China. https://github.com/xyyu001/CCM14

A 代码与数据

所有代码和数据存放于 `v1_nn_coupled_dynamics/` 目录:

```
v1_nn_coupled_dynamics/
  code/
    phase1_coupled_model.py    # Phase 1: NN耦合动力学
    phase2_formula_discovery.py # Phase 2: 符号回归
  results/
    figures/                   # 可视化图片
    data/                      # 预测数据、模型权重
  report/
    report_v1.tex              # 本报告
```