

Hierarchical Gradient-Based Genetic Sampling for Accurate Prediction of Biological Oscillations

Heng Rao¹, Yu Gu^{1*}, Jason Zipeng Zhang², Ge Yu¹, Yang Cao³, Minghan Chen^{2*}

¹College of Computer Science and Engineering, Northeastern University, Shenyang, China

²Department of Computer Science, Wake Forest University, Winston-Salem, NC, USA

³Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
guyu@mail.neu.edu.cn, chenm@wfu.edu

Abstract

Biological oscillations are periodic changes in various signaling processes crucial for the proper functioning of living organisms. These oscillations are modeled by ordinary differential equations, with coefficient variations leading to diverse periodic behaviors, typically measured by oscillatory frequencies. This paper explores sampling techniques for neural networks to model the relationship between system coefficients and oscillatory frequency. However, the scarcity of oscillations in the vast coefficient space results in many samples exhibiting non-periodic behaviors, and small coefficient changes near oscillation boundaries can significantly alter oscillatory properties. This leads to non-oscillatory bias and boundary sensitivity, making accurate predictions difficult. While existing importance and uncertainty sampling approaches partially mitigate these challenges, they either fail to resolve the sensitivity problem or result in redundant sampling. To address these limitations, we propose the Hierarchical Gradient-based Genetic Sampling (HGGs) framework, which improves the accuracy of neural network predictions for biological oscillations. The first layer, Gradient-based Filtering, extracts sensitive oscillation boundaries and removes redundant non-oscillatory samples, creating a balanced coarse dataset. The second layer, Multigrid Genetic Sampling, utilizes residual information to refine these boundaries and explore new high-residual regions, increasing data diversity for model training. Experimental results demonstrate that HGGs outperforms seven comparative sampling methods across four biological systems, highlighting its effectiveness in enhancing sampling and prediction accuracy.

1 Introduction

In biological systems, oscillation refers to the repetitive, cyclical behaviors of cell signaling and biological processes over time, such as fluctuations in concentrations of biochemical substances (Tamate et al. 2017), rhythmic activities in cellular functions (Goldbeter 2002), or periodic changes in physiological states (Kurosawa, Mochizuki, and Iwasa 2002). Understanding these oscillations is crucial for comprehending how biological systems maintain stability, respond to external stimuli, and regulate complex processes. To study oscillations, researchers often use ordinary differential equations (ODEs) to model the temporal dynamics

that characterize oscillatory patterns in biological systems. Recently, the application of machine learning to biological systems has gained significant attention. In many studies (Daneker et al. 2023; Yazdani et al. 2020; Szep, Dalchau, and Csikász-Nagy 2021), system-level mathematical models of biological reactions are integrated with machine learning models to establish relationships between raw data and system coefficients. Of these, Neural Networks (NNs) show promising results, offering an effective tool for predicting biological oscillations based on system coefficients.

However, the relationship between system coefficients and oscillatory behaviors is highly complex (Stark, Chan, and George 2007). Variations in these coefficients can disrupt oscillations, causing non-oscillatory states. In biological systems, coefficients that can lead to oscillatory states are relatively rare; most coefficient combinations result in non-oscillatory behavior, especially in high-dimensional domains. The prevalence of non-oscillatory states creates a significant data imbalance, challenging the prediction of oscillatory frequency—a key characteristic of oscillation, defined as the inverse of the oscillatory period ($\frac{1}{\text{period}}$). When oscillation is absent, the frequency is set to zero. Fig. 1(a) illustrates the mapping between system coefficients and oscillatory frequencies of a cell cycle system (Liu et al. 2012). The domain shows data imbalance with 70% non-oscillation (blue). Moreover, minor changes in these coefficients can lead to substantial changes in oscillatory frequencies, particularly in boundary areas where oscillatory states transition to non-oscillatory states. This heightened sensitivity further complicates accurate predictions and is reflected in the poor performance of NNs in these regions, as evidenced by high residuals (absolute error) marked as black shades in Fig. 1. Overall, we face two challenges:

- **Non-Oscillatory Bias.** The majority of data samples exhibit non-oscillatory behaviors, resulting in an abundance of redundant information that does not improve model accuracy and may even reduce efficiency.
- **Oscillatory Boundary Sensitivity.** Sharp transitions between non-oscillatory and oscillatory states occur in narrow boundaries of the coefficient space, which are easily overlooked by random sampling and lead to high errors.

Therefore, a more effective sampling strategy is necessary to obtain more balanced and representative samples for

*Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

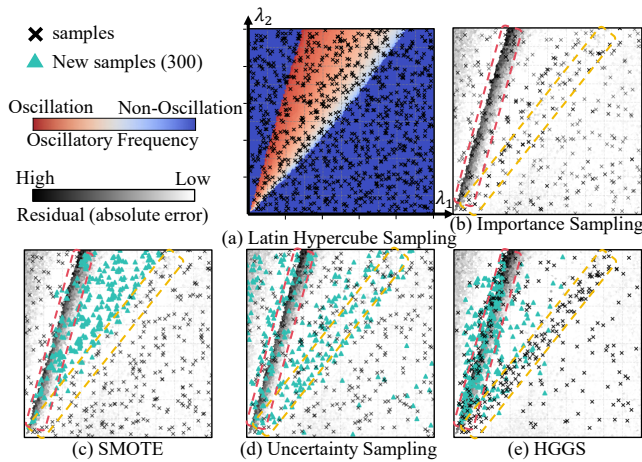


Figure 1: Examples of different sampling methods applied to the system coefficients of a cell cycle model (Liu et al. 2012). The orange-blue and black-white color bars indicate trends in oscillation and residual changes, respectively. (a) LHS performs a random sampling of 1000 points in the coefficient domain. (b) IS fails to introduce new samples; (c-d) SMOTE and US lack sufficient coverage in high residual regions (red box). US also produces redundant samples spreading across the domain; (e) Our proposed HGGGS method extracts the boundary information (red, yellow boxes) and effectively generates new samples (tear triangle) concentrating on high-residual regions (red box).

model training. Importance Sampling (IS) (Liu et al. 2021; Lu et al. 2023) can be used to resample more minority class instances, balancing the majority-minority ratio. As shown in Fig. 1(b), IS resamples the oscillation class, with darker samples indicating higher resampling, but it cannot generate new samples to alleviate boundary sensitivity issues. Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002; Torgo et al. 2013) offers another way to balancing the dataset by generating synthetic minority samples. However, it insufficiently covers high-residual areas (red box), see Fig. 1(c). Targeting data sensitivity, Uncertainty Sampling (US) (Liu and Li 2023) selects informative samples based on predicted importance. However, due to limited diversity in the random selection process, US often produces redundant samples, reducing efficiency. This is evident in Fig. 1(d), where new samples are scattered across the domain but fail to adequately cover high-residual regions (red box) compared to our method in Fig. 1(e). Thus, these sampling strategies fail to effectively address the challenges.

This work introduces the Hierarchical Gradient-based Genetic Sampling (HGGGS) method, a novel two-layer framework designed to enhance the effectiveness of selecting representative samples and improve the accuracy of predictions for biological oscillations. Specifically, starting with an initial candidate set obtained by Latin Hypercube Sampling (LHS), we first apply Gradient-based Filtering (GF) to select samples near boundary regions and balance the proportions of non-oscillation (majority) and oscillation (minority)

samples. This GF layer yields a balanced coarse dataset that guides subsequent refinement. In the second layer, Multigrid Genetic Sampling (MGS) dynamically constructs grids at multiple levels based on residual (absolute error) information from training data. By sampling within these grids, MGS not only enriches existing boundaries but also explores new high-residual areas. Through continual learning, HGGGS iteratively refines the training dataset with more representative samples, consistently improving model performance. The main contributions of this paper are summarized below:

- We propose a simple and efficient Gradient-based Filtering technique that can extract oscillation boundaries, which often entail high residuals, and removes redundant non-oscillatory samples. The GF layer generates balanced coarse data, enabling efficient MGS refinement.
- Our Multigrid Genetic Sampling strategy leverages residual information to refine existing boundaries and explore new high-residual regions. The MGS layer systematically samples across grids at different levels, reducing sensitivity and enhancing oscillation diversity.
- The proposed HGGGS method ensures a representative dataset through continuous adaption to the evolving error landscape during training, showing superior accuracy over seven baselines across four biological systems. HGGGS is versatile and can be applied to predict various systematic features beyond biological oscillations.

2 Related Work

Sampling for Data Imbalance. Data imbalance often results in biased model performance, favoring majority samples while underperforming on minority ones. To address this issue, several sampling techniques have been proposed, including undersampling, oversampling, and importance sampling techniques. Undersampling randomly removes majority samples (Wilson 1972; Tomek 1976; Guo et al. 2008) to balance the imbalance ratio, but it risks discarding informative data. On the other hand, oversampling mitigates data imbalance by augmenting minority class samples, as in SMOTE (Chawla et al. 2002; Torgo et al. 2013). While SMOTE generates new minority samples to improve balance, it can introduce noise due to interpolated labels and may struggle to create truly representative minority samples. Importance Sampling (IS) (Liu et al. 2021; Lu et al. 2023) combines elements of both undersampling and oversampling by resampling more informative minority samples and excluding well-performing majority samples. However, it is prone to overfitting due to resampling and does not incorporate new data. Although these techniques offer various ways to alleviate data imbalance, they all struggle with the data sensitivity challenge.

Dynamic Sampling. Dynamic sampling offers a more adaptive approach to selecting informative samples, enhancing model training by adjusting sample selection based on the model’s evolving state. This idea is widely utilized in Active Learning (AL), where various dynamic sampling methods leverage direct or indirect information from the model to select the most representative samples from unlabeled

data. One prominent method in AL is Uncertainty Sampling (US) (Lewis and Catlett 1994; Zhu et al. 2010; Liu and Li 2023), which selects samples based on their uncertainty scores. In classification, these scores can be calculated using techniques such as entropy uncertainty (Shannon 1948) or confidence margin uncertainty (Sharma and Bilgic 2017). Diversity-based strategies aim to select a broad range of samples based on data distribution, employing methods like gradient representation (Saran et al. 2023) and switch events (Benkert et al. 2023). Query by committee methods (Burbidge, Rowland, and King 2007; Kee, Del Castillo, and Runger 2018; Hino and Eguchi 2023) aggregate outputs from multiple models to form new discriminative criteria, identifying the most representative samples for labeling by considering the underlying data distribution. However, uncertainty sampling and diversity-based sampling often introduce significant redundancy, reducing sampling efficiency. Additionally, most of these methods are designed for nominal target variables and are rarely applicable to regression problems with continuous targets (Liu and Li 2023). In contrast, our HGGs method leverages residual information to ensure targeted sampling in high-residual areas and avoid redundancy, boosting both effectiveness and efficiency.

3 Preliminary

In this paper, we utilize a neural network model $\hat{y}(\boldsymbol{\lambda}) = f_{nn}(\boldsymbol{\lambda}; \Theta)$ to approximate the oscillatory frequency $y(\boldsymbol{\lambda}) = f_P(\mathbf{u}(t, \boldsymbol{\lambda})) \in \mathbb{R}^{D'}$ of a system under initial state \mathbf{u}_0 , given any set of coefficients $\boldsymbol{\lambda} \in \mathbb{R}^D$ in the ODEs presented below.

$$\begin{aligned} \frac{d\mathbf{u}}{dt} &= \mathcal{N}(\mathbf{u}), t \in [0, T] \\ \mathbf{u}(t, \boldsymbol{\lambda})|_{t=0} &= \mathbf{u}_0(\boldsymbol{\lambda}), \boldsymbol{\lambda} \in \Omega, \end{aligned} \quad (1)$$

where f_P is the oscillatory operator used to calculate the oscillatory frequency $f_P(\mathbf{u}(t, \boldsymbol{\lambda}))$ of \mathbf{u} over the time span $t = [0, t_1, t_2, \dots, T]^\top$ using (Apicella et al. 2013). $\mathcal{N}(\mathbf{u})$ denotes a nonlinear operator consisting of variables vector \mathbf{u} , $\boldsymbol{\lambda} \in \Omega$ is a D -dimensional system coefficient vector, Ω is a subset of \mathbb{R}^D , \mathbf{u}_0 is the initial condition vector in D' dimensions, and Θ is the parameter of the neural network. Our goal is to minimize the error in approximating the original function $f_P(\mathbf{u}(t, \boldsymbol{\lambda}))$ using a neural network $f_{nn} \in F : \mathbb{R}^D \rightarrow \mathbb{R}^{D'}$, where F represents an appropriate function mapping space. We achieve this by employing supervised learning with the following loss function.

$$L(f_{nn}(\boldsymbol{\lambda}; \Theta), y) = \|l(f_{nn}(\boldsymbol{\lambda}; \Theta), y)\|_{2, \Omega}^2, \quad (2)$$

where $l(f_{nn}(\boldsymbol{\lambda}; \Theta), y) = |f_{nn}(\boldsymbol{\lambda}; \Theta) - y|$ and $\|\cdot\|_2$ denotes Euclidean normalization.

In practice, Monte Carlo (MC) approximation is used to estimate the overall error L using a finite set of sampled points from the domain, denoted as L_N . Thus, the loss function can be approximated as:

$$\begin{aligned} L(\Theta) \approx L_N(\Theta) &= \|l(f_{nn}(\boldsymbol{\lambda}; \Theta), y)\|_{2, S_\Omega}^2 \\ &= \frac{1}{N} \sum_{i=1}^N \|l(f_{nn}(\boldsymbol{\lambda}^{(i)}; \Theta), y^{(i)})\|_2^2, \end{aligned} \quad (3)$$

where N is the number of samples and $S_\Omega = \{\boldsymbol{\lambda}^{(i)}\}_{i=1}^N$ represents the set of sampled coefficients from Ω . To theoretically prove that sampling methods can further reduce the error between f_{nn} and f_P , we build on the framework from (Tang et al. 2023). Through analyzing the MC approximate loss function L_N , we derive the theoretical upper bound of the optimal model $f_{nn}(\cdot)_N^*$ approximating to oscillatory frequency $y(\cdot)$. Additionally, we introduce the theoretically optimal model $f_{nn}(\cdot)^*$ from overall error L , categorizing the error into two parts:

$$\begin{aligned} E(\|f_{nn}(\cdot)_N^* - y(\cdot)\|_\Omega) &\leq E(\|f_{nn}(\cdot)_N^* - f_{nn}(\cdot)^*\|_\Omega) \\ &\quad + \|f_{nn}(\cdot)^* - y(\cdot)\|_\Omega. \end{aligned}$$

Here, E is the expectation operator, and $\|\cdot\|_\Omega$ is the norm operator in the function space F . The first term represents the statistical error, while the second term reflects the model approximation error determined by model structure. Sampling new data from the residual distribution over Ω can reduce statistical error. Proof details are provided in the Appendix.

4 Method

This section describes the Hierarchical Gradient-based Genetic Sampling (HGGs) framework (Fig. 2), which divides the sampling process into two layers: Gradient-based Filtering and Multigrid Genetic Sampling. HGGs yields a diverse, balanced dataset, drastically reducing model statistical error.

Gradient-based Filtering

As demonstrated in Fig. 1, biological systems often contain a significant amount of redundant data in non-oscillatory regions. Removing these redundant non-oscillatory samples not only preserves the model's overall performance but also increases the proportion of oscillatory samples in the dataset, thereby mitigating the effects of non-oscillatory bias.

To achieve this, we design an effective Gradient-based Filtering (GF) technique that identifies sensitive regions, particularly boundary areas where small changes can lead to significant shifts in oscillatory frequency (see red, yellow boxes in Fig. 1(e)), which are also characterized by higher residuals. In this context, we first rank all samples from the initial dataset S_Ω using Eq. 4.

$$gd(\boldsymbol{\lambda}^{(i)})_{S_\Omega} = \frac{1}{K} \sum_{j=1}^K \frac{\|y^{(j)} - y^{(i)}\|_2^2}{\|\boldsymbol{\lambda}^{(j)} - \boldsymbol{\lambda}^{(i)}\|_2^2}, \boldsymbol{\lambda}^i, \boldsymbol{\lambda}^j \in S_\Omega, i \neq j, \quad (4)$$

where $gd(\boldsymbol{\lambda}^{(i)})_{S_\Omega}$ represents the gradient degree of collocation sample i from the training dataset S_Ω and K is a hyperparameter denoting the top- K nearest collocation samples to i in the training dataset S_Ω . This function assesses the importance of collocation sample i ; a higher $gd(\boldsymbol{\lambda}^{(i)})_{S_\Omega}$ value indicates that the sample is closer to boundary areas.

According to this gradient ranking, we filter the top- $r\%$ samples $S_{\Omega_{gf1}}$ from S_Ω , where r is the filtering ratio to extract the boundary information. Next, to enhance the model's adaptability to non-oscillatory conditions, we retain an essential global dataset, $S_{\Omega_{gf2}}$, from the remaining samples in $S_\Omega - S_{\Omega_{gf1}}$ using uncertainty sampling (Liu and

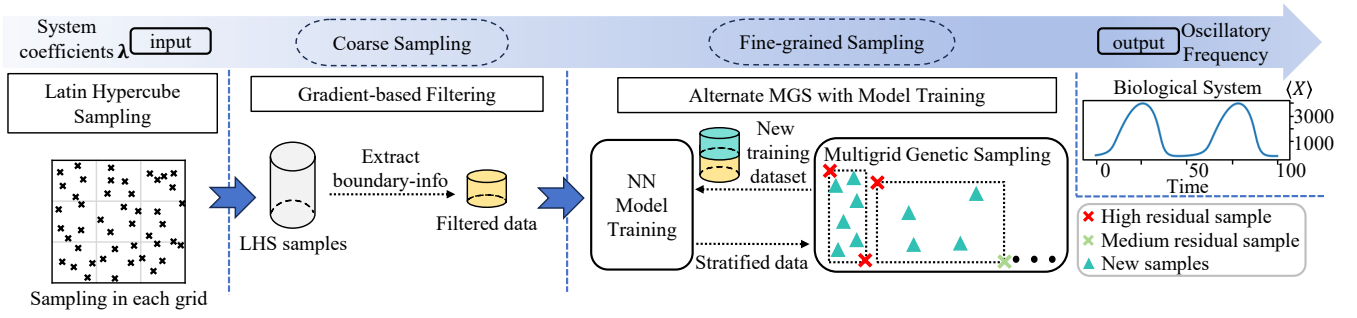


Figure 2: Overview of Hierarchical Gradient-based Genetic Sampling framework: The input consists of initial samples obtained via LHS. Gradient-based Filtering is then employed to extract sensitive boundary information and eliminate redundant samples, generating balanced coarse data. Next, Multigrid Genetic Sampling constructs candidate sampling from stratified data to sharpen boundary precision. During training, new instances in high-residual areas are explored to continuously improve model performance. Finally, given any set of coefficients λ , the model predicts the oscillatory frequency of the biological system.

Li 2023). This subset is specifically selected to capture the overall characteristics and maintain performance for non-oscillatory regions. We then formulate the coarse training set $S_{\Omega}^{(0)} = S_{\Omega_{gf}} = S_{\Omega_{gf1}} \cup S_{\Omega_{gf2}}$, where $S_{\Omega_{gf}} \subseteq S_{\Omega}$ and $|S_{\Omega_{gf}}| < |S_{\Omega}|$. In $S_{\Omega_{gf}}$, most redundant non-oscillation samples from the original dataset are removed, while a diverse set of oscillatory data is retained.

Multigrid Genetic Sampling

From the prior GF layer, we obtained a balanced coarse set of informative samples $S_{\Omega}^{(0)}$. Our next goal is to sharpen boundary precision and explore high-residual areas through effective sampling. This requires constructing an accurate residual distribution, as non-oscillatory regions have low probabilities and contribute little to model performance. We focus on constructing high-residual distributions.

However, the small sample set S_{Ω} obtained through random sampling does not sufficiently represent the high-residual distribution p_g over the domain Ω . To address this limitation, we construct multiple sampling grids of varying sizes to capture the high-residual distribution p_g .

We first use a Gaussian Mixture model (Reynolds et al. 2009) to characterize three potential residual distributions within S_{Ω} : low-, medium-, and high-residual distributions, respectively, denoted as $S_{\Omega_{lr}}$, $S_{\Omega_{mr}}$, and $S_{\Omega_{hr}}$.

Next, each grid is defined as a hypercube bounded by points $\lambda^{(i)}$ and $\lambda^{(j)}$, both originating from $S_{\Omega}^{(0)} - S_{\Omega_{lr}}^{(0)}$. To ensure comprehensive coverage of the high-residual distribution, it is necessary to construct a large number of grids at different sizes, each composed of diverse samples.

Building on the idea of multigrid, we integrate genetic sampling into a method called Multigrid Genetic Sampling (MGS) to approximate the high-residual distribution over Ω and further refine these areas. Specifically, we sample new points within each hypercubic grid according to Eq. 5:

$$\lambda_{new} = \alpha \odot (\lambda^{(i)} - \lambda^{(j)}) + \lambda^{(j)}; \alpha \in [0, 1]^D, i \neq j, \quad (5)$$

where α is a D-dimensional weight vector with each value in the range $[0,1]$, and \odot denotes component-wise multipli-

cation. λ_{new} is a randomly sampled point within the hypercube enclosed by $\lambda^{(i)}$ and $\lambda^{(j)}$.

In order to obtain a fine-grained dataset from $S_{\Omega}^{(0)}$, we alternate MGS with model training to continuously refine the high-residual areas. We define m_c as the number of sampling cycles and m_e as the number of training epochs. The k' -th sampled dataset by MGS is denoted as $S_{\Omega_{gs}}^{(k')}$, and the k -th training dataset is $S_{\Omega}^{(k)} = \bigcup_{k'=1}^k S_{\Omega_{gs}}^{(k')} \cup S_{\Omega_{gf}}$.

For the $(k+1)$ -th sampling, we utilize the current k -th residual samples, categorized in low-, medium-, and high-residual distributions, denoted as $S_{\Omega_{lr}}^{(k)}$, $S_{\Omega_{mr}}^{(k)}$, and $S_{\Omega_{hr}}^{(k)}$. Next, to exploit and explore the high-residual domain, we defined crossover and mutation operations over medium- and high-residual sets as follows:

- **Multigrid Crossover:** Randomly select two distinct samples $\lambda_{hr}^{(i)}$ and $\lambda_{hr}^{(j)}$ from the high-residual subset $S_{\Omega_{hr}}^{(k)}$, and randomly sample a new point λ_{hh} by Eq. 5 and add it to the sample set $S_{\Omega_{gs}}^{(k+1)}$. This operation serves to exploit and refine existing high-residual areas.
- **Multigrid Mutation:** Randomly select $\lambda_{hr}^{(i)}$ from $S_{\Omega_{hr}}^{(k)}$ and $\lambda_{mr}^{(j)}$ from $S_{\Omega_{mr}}^{(k)}$. Form a hypercubic grid using the two samples and sample a new point λ_{hm} by Eq. 5. Add this new sample to $S_{\Omega_{gs}}^{(k+1)}$. The mutation operation serves to explore global boundaries and new high-residual areas.

Before the $(k+1)$ -th training round, our method samples n_{v1} and n_{v2} points using the aforementioned operations, such that $|S_{\Omega_{gs}}^{(k+1)}| = n_{v1} + n_{v2}$ and $S_{\Omega_{gs}}^{(k+1)} = \bigcup_{i=1}^{n_{v1}} \lambda_{hh}^{(i)} \cup \bigcup_{i=1}^{n_{v2}} \lambda_{hm}^{(i)}$. Thus, the sample set for the $(k+1)$ -th training round is $S_{\Omega}^{(k+1)} = S_{\Omega_{gs}}^{(k+1)} \cup S_{\Omega}^{(k)}$. At the end of sampling, $n_s = m_c \times (n_{v1} + n_{v2})$ and $S_{\Omega_{gs}} = \bigcup_{k=1}^{m_c} S_{\Omega_{gs}}^{(k)}$.

The GF followed by the MGS layer, allows for effective and efficient sampling within high residuals. The pseudocode is presented in Algorithm 1.

Algorithm 1: HGGS for predicting biological oscillations

Input: NN model $f_{nn}(\lambda; \Theta)$, neighbors for gradient estimation K , initial sample size N , filtering ratio r , sampling cycle m_c , Multigrid Genetic Sampling budget $\{n_{v1}, n_{v2}\}$

Initialization: LHS $S_\Omega = \{(\lambda^{(i)}, y^{(i)})\}_{i=1}^N$

Output: Target model $f_{nn}(\lambda; \Theta^*)$

- 1: Apply Gradient-based Filtering to S_Ω to generate $S_\Omega^{(0)}$:
 $S_\Omega^{(0)} \leftarrow GF(S_\Omega, r, K)$ where $S_\Omega^{(0)} \subseteq S_\Omega$
 - 2: Update $f_{nn}(\lambda; \Theta)$ by minimizing $L_{|S_\Omega^{(0)}|}$ in Eq. 3
 - 3: **for** $k = 0, 1, \dots, m_c - 1$ **do**
 - 4: Compute residual $l = \{|f_{nn}(\lambda^{(i)}; \Theta^{(k)}) - y^{(i)}|\}_{i=1}^{|S_\Omega^{(k)}|}$
 - 5: Stratify $S_\Omega^{(k)}$ into 3 subdomains based on residual using Gaussian Mixture: $\{S_{\Omega_{lr}}^{(k)}, S_{\Omega_{mr}}^{(k)}, S_{\Omega_{hr}}^{(k)}\}$
 - 6: $S_{\Omega_{gs}}^{(k+1)} \leftarrow MGS(n_{v1}, n_{v2}, S_{\Omega_{lr}}^{(k)}, S_{\Omega_{mr}}^{(k)}, S_{\Omega_{hr}}^{(k)})$
 - 7: $S_\Omega^{(k+1)} \leftarrow S_\Omega^{(k)} \cup S_{\Omega_{gs}}^{(k+1)}$ // Update datasets
 - 8: Update $f_{nn}(\lambda; \Theta)$ by minimizing $L_{|S_\Omega^{(k+1)}|}$ in Eq. 3
 - 9: **end for**
 - 10: **return** $f_{nn}(\lambda; \Theta^*)$
-

5 Experiments

To validate the proposed HGGS, we conducted experiments on four biological systems known for their oscillatory behaviors. We also performed an ablation study to assess Gradient-based Filtering and Multigrid Genetic Sampling, and examined the sensitivity of model hyperparameters.

Biological System Dataset

Benchmark datasets from four biological systems were used for method evaluation: the Brusselator system (Prigogine 1978), the Cell Cycle system (Liu et al. 2012), the Mitotic Promoting Factor (MPF) system (Novak and Tyson 1993), and the Activator Inhibitor system (Murray 2002). For each system, we generated 20k–70k sets of system coefficients using LHS, ran simulations to produce system dynamics, and determined the oscillatory frequency using (Apicella et al. 2013). This data was then used for training and testing of our method. Descriptions of the four biological systems, their corresponding ODEs, and detailed simulation settings are provided in the Appendix.

Baselines

The proposed HGGS was compared with seven baselines:

- Latin Hypercube Sampling (LHS) (Stein 1987): Divides the coefficient domain into equal grids and randomly samples from each grid to ensure full coverage.
- Importance Sampling (IS) (Lu et al. 2023): Resamples based on residual information, updating the training set after each epoch. IS[†] is an optimization that updates only when the model shows no improvement.
- Uncertainty Sampling (US) (Liu and Li 2023): Selects the top- o new samples from a candidate set based on residual distribution, and adds them to the training set.

We implemented the candidate set in two ways: pool-based (US-P) and streaming-based (US-S).

- Weight Reservoir Sampling (WRS) (Efrimidis and Spirakis 2006): Selects the top- o new samples from the candidate set based on residuals to replace part of the current data, keeping a constant training sample size.
- Volume Sampling for Streaming Active Learning (VeSS) (Saran et al. 2023): Selects samples based on their gradient distribution relative to the model, which is a type of diversity-based sampling in AL.

Implementation

Our algorithm was implemented using the PyTorch framework on a single NVIDIA A6000 GPU. We utilized $N = 10k$ samples for initial training and 5k samples for validation for each experiment. For a thorough evaluation, our testing data consists of four subsets, characterizing different types of the coefficient domain: overall (entire testing data), majority (non-oscillatory samples only), minority (oscillatory samples only), and boundary (top 20% samples ranked by gradient using Eq. 4). The total size of the testing data varies between 7k–60k, depending on the oscillatory systems.

The neural network (Multi-Layer Perceptron), consisting of 3 or 4 hidden layers, was trained for 3k epochs per sampling cycle using the Adam optimizer with a learning rate of $2\text{--}2.5 \times 10^{-3}$, employing full batch training and early stopping. For key hyperparameters, the GF filtering ratio was set to $r = 20\%$, with $K = 5$ nearest neighbors and a GF sample size of $n_f = N/2$. During the sampling cycles, the MGS ratio was set to $n_{v1} : n_{v2} = 6 : 4$, with an MGS sample size of $n_s = N/2$. Implementation details and other parameters for each experiment can be found in the Appendix. All reported results below were based on five independent experiments.

Results

Metrics. Root Mean Square Error (RMSE) is our primary metric. Imbalance Ratio (IR) and Gini Index (GI) quantify the proportion of non-oscillatory to oscillatory samples and the diversity of oscillatory frequency labels, respectively. Lower IR and GI indicate more effective handling of non-oscillatory bias and oscillatory boundary sensitivity by the sampling method. RMSE, IR, and GI are defined below.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N \|f_{nn}(\lambda^{(i)}; \Theta) - y^{(i)}\|_2^2}$$

$$\text{IR} = \frac{\sum_{i=1}^N I(y^{(i)} = 0)}{\sum_{i=1}^N I(y^{(i)} \neq 0)}; \quad \text{GI} = \frac{\sum_{i=1}^N \sum_{j=1}^N |y^{(i)} - y^{(j)}|}{2N \sum_{i=1}^N y^{(i)}}$$

Brusselator. In Fig. 3(a), our method achieves the lowest RMSE for the overall testing data (0.0085). Notably, for crucial minority and boundary cases, we observe 24%–71% and 13%–66% improvement over the baseline approaches. In this low-dimensional, densely-packed space, IS[†] (minority: 0.0119; boundary: 0.0320) performs well but lacks sufficient boundary information. HGGS, however, effectively targets more boundary instances, particularly in high-residual boundary regions, as illustrated in Appendix. Due to

Model	Brusselator System		Cell Cycle System		MPF System		Activator Inhibitor System	
	IR	GI	IR	GI	IR	GI	IR	GI
LHS	1.47±0.00	0.76±0.00	4.43±0.00	0.85±0.00	5.27±0.00	0.89±0.00	11.84±0.00	0.94±0.00
WRS	1.37±0.05	0.75±0.00	4.41±0.12	0.85±0.00	5.16±0.18	0.89±0.00	11.32±0.48	0.94±0.00
VeSS	1.42±0.12	0.75±0.01	3.98±0.16	0.84±0.01	3.57±0.37	0.85±0.01	10.96±0.56	0.94±0.00
IS	1.35±0.23	0.77±0.03	1.86±0.36	0.70±0.04	3.07±0.44	0.82±0.02	6.04±1.15	0.88±0.02
IS [†]	1.20±0.15	0.75±0.02	2.00±0.44	0.71±0.04	3.11±0.78	0.82±0.03	6.09±2.07	0.88±0.04
US-S	1.18±0.11	0.72±0.01	3.93±1.53	0.83±0.05	3.75±0.37	0.85±0.01	11.58±2.90	0.94±0.02
US-P	1.08±0.17	0.70±0.02	2.3±0.22	0.76±0.02	2.75±0.41	0.81±0.02	7.12±0.24	0.91±0.00
Ours	1.18±0.12	0.71±0.03	1.10±0.02	0.58±0.01	1.11±0.07	0.65±0.01	2.88±0.13	0.77±0.01

Table 1: Imbalance Ratio (IR) and Gini Index (GI) for seven baseline methods and our HGGs across four biological systems. HGGs achieves the lowest IR and GI in most cases.

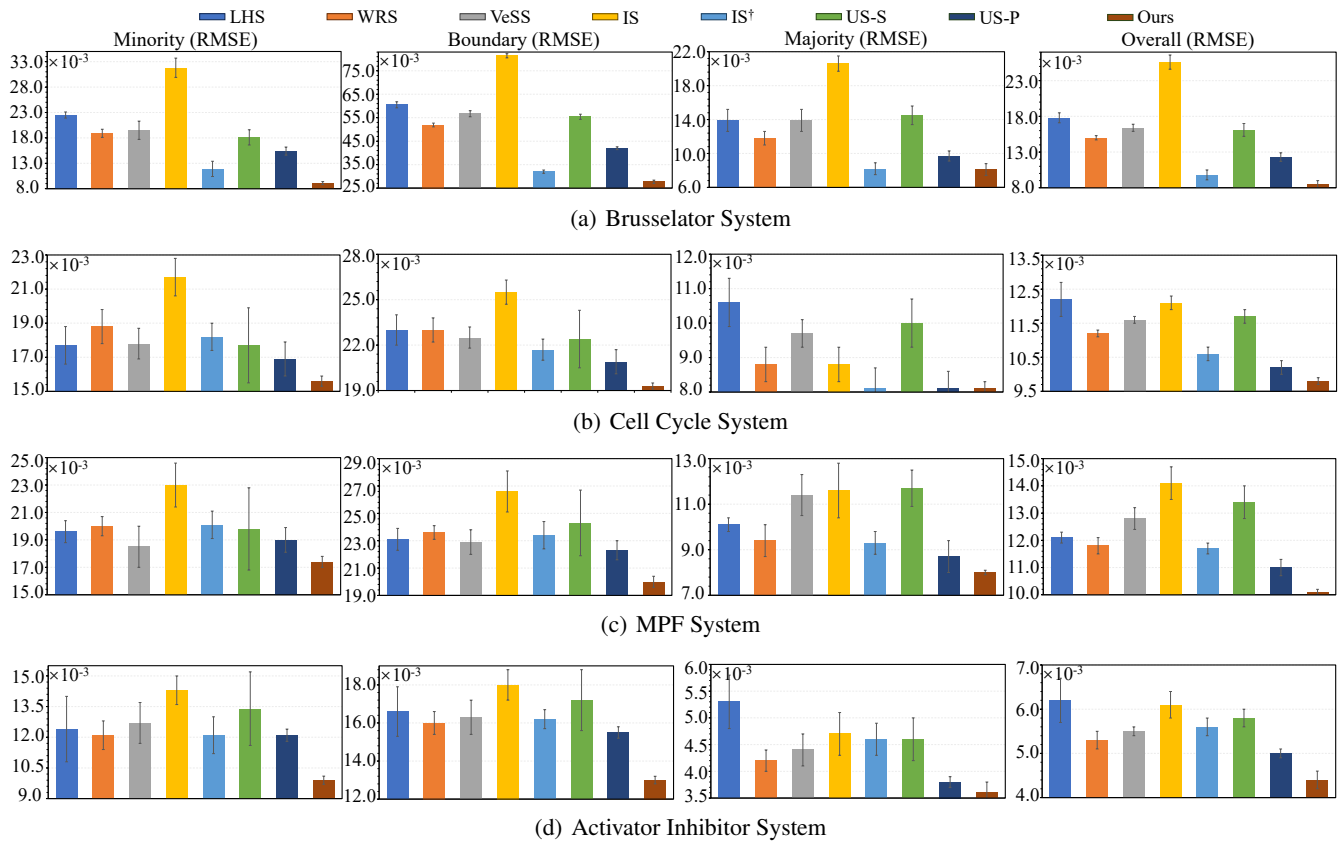


Figure 3: Accuracy comparison of seven baseline methods (LHS, WRS, VeSS, IS, IS[†], US-S, US-P) and our HGGs across four biological systems. HGGs obtained the lowest RMSEs across all testing subsets: minority, boundary, majority, and overall.

the relatively mild data imbalance in the Brusselator system, the differences in IR and GI between our method, US-S, and US-P are minimal. Nonetheless, HGGs efficiently samples at the boundaries and improves prediction accuracy compared to the others.

Cell Cycle. In this complex cell cycle system, our method distinguishes itself with the lowest IR and GI in Table 1, highlighting its dual strength in balancing data distribution and enriching sample diversity. HGGs also obtains the lowest RMSE error (overall: 0.0098), as shown in Fig. 3(b).

Specifically for minority and boundary cases, our method shows improvements of 8%–28% and 8%–24% over the other seven baselines. US-P (minority: 0.0169; boundary: 0.0209) includes high-residual samples but suffers from redundancy due to random sampling. In contrast, our method directly targets high-risk areas, avoiding this redundancy and demonstrating superior efficiency and effectiveness.

MPF. Fig. 3(c) shows that our method excels in achieving minimal RMSE (overall: 0.0101), a decrease of 7%–27% compared to other methods. It also stands out for minority

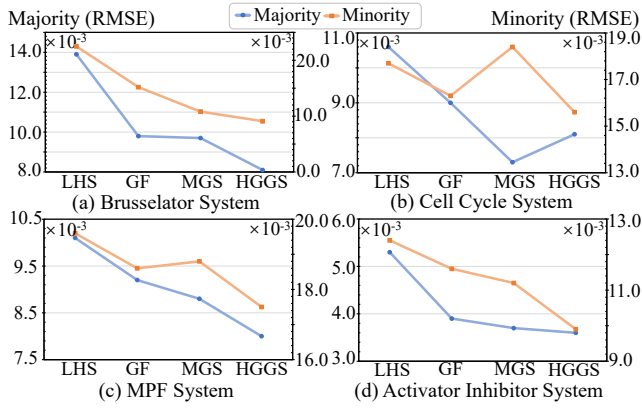


Figure 4: Ablation study for majority and minority classes across four oscillatory systems. Both Gradient-based Filtering (GF) and Multigrid Genetic Sampling (MGS) layers contribute to the improvement in model accuracy.

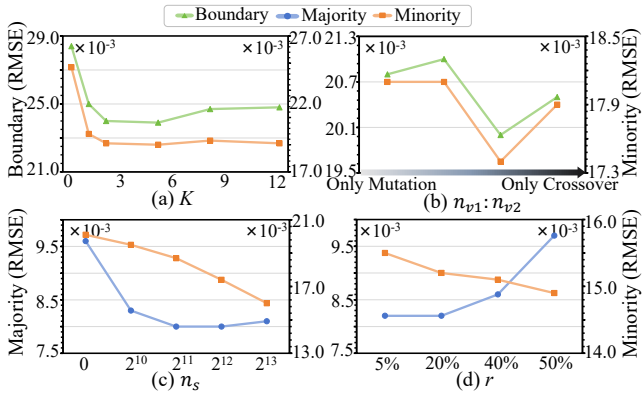


Figure 5: Sensitivity analysis on (a) number of neighbors K , (b) MGS ratio $n_{v1} : n_{v2}$, (c) MGS sampling size n_s , and (d) GF filtering ratio r . (a-c) are performed on the MPF system, while (d) is conducted on the Cell Cycle system.

and boundary cases (minority: 0.0174; boundary: 0.0200). VeSS performs adequately for minority and boundary cases, but its weaker performance in the majority case reduces its overall effectiveness. Table 1 further underscores the effectiveness of HGGS in mitigating non-oscillatory bias in the MPF system. Our method achieves the lowest GI among all baselines, reflecting its superior ability to uncover informative oscillatory patterns and enhance diversity.

Activator Inhibitor. The Activator Inhibitor system is characterized by a pronounced data imbalance, with a staggering IR of 11.84 under LHS. However, as highlighted in Table 1, our method exhibits remarkable resilience against such extreme scenarios, effectively reducing the IR to about 2.88 while simultaneously enhancing data diversity to a GI of 0.77. Moreover, HGGS has the lowest error across all categories (minority: 0.0099; boundary: 0.0130; majority: 0.0036; overall: 0.0044) (Fig. 3(d)). It particularly excels in minority and boundary cases, showing improvements of 18%–31% and 16%–28% over other baselines.

Ablation Study

We conducted an ablation study on four biological systems to evaluate the efficacy of the two layers: Gradient-based Filtering (GF) and Multigrid Genetic Sampling (MGS).

Effect of Gradient-based Filtering. Relying solely on GF followed by random sampling does not effectively address non-oscillatory bias and oscillatory boundary sensitivity. Although GF offers initial benefits, subsequent random sampling fails to sustain improvements. As shown in Fig. 4, GF yields higher RMSEs for majority and minority cases across four systems than HGGS, indicating that MGS produces more informative samples than random sampling.

Effect of Multigrid Genetic Sampling. When MGS is used without prior GF, it struggles to precisely identify informative minority samples, leading to higher RMSEs for minority class across all four systems compared to HGGS (Fig. 4). This suggests that GF provides critical guidance for MGS to effectively target and refine minority samples.

These findings underscore the synergy between GF and MGS within HGGS. GF enables an efficient extraction of sample domains, identifying coarse boundaries rich in critical information. Informed by GF, MGS operates with greater precision to refine minority samples and explore new high-risk areas. Together, they enhance data representativeness and diversity, improving model performance.

Sensitivity Analysis

Fig. 5(a-c) illustrates the effects of key model parameters, including the number of neighbors K , MGS ratio $n_{v1} : n_{v2}$, and MGS sampling size n_s on the MPF system. In Fig. 5(a), varying K produces an elbow curve, with $K = 5$ offering optimal performance. Increasing K beyond this point has little effect on gradient estimation, while decreasing K leads to significant errors in oscillatory frequency estimation. In Fig. 5(b), MGS cannot achieve optimal performance using only crossover or mutation operation. Our experiments found that an MGS ratio of 6:4 yielded the lowest RMSE. In Fig. 5(c), increasing the MGS sampling size n_s continues to reduce RMSE for the minority case, while the majority case shows no further improvement beyond 2k new samples. Fig. 5(d) shows the effect of GF filtering ratio on the Cell Cycle system. Increasing r reduces the sampling of majority (non-oscillation) class, leading to higher error, while the opposite is true for minority class. A ratio of $r = 20\%$ offers the best balance between majority and minority samples.

Conclusion

This paper introduces Hierarchical Gradient-based Genetic Sampling, a two-layer framework designed to address non-oscillatory bias and boundary sensitivity in predicting biological oscillations. The first layer, gradient-based filtering, selects a representative subset from initial random sampling, creating a balanced coarse dataset. The second layer, genetic sampling, refines minority instances and explores new high-residual information, enhancing data diversity. Experiments show that HGGS achieves the best accuracy across four biological systems, particularly for the oscillation class.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U23B2019) and the Fundamental Research Funds of the Central Universities (N2216017).

References

- Apicella, B.; Bruno, A.; Wang, X.; and Spinelli, N. 2013. Fast Fourier Transform and autocorrelation function for the analysis of complex mass spectra. *International Journal of Mass Spectrometry*, 338: 30–38.
- Benkert, R.; Prabhushankar, M.; AlRegib, G.; Pacharmi, A.; and Corona, E. 2023. Gaussian Switch Sampling: A Second-Order Approach to Active Learning. *IEEE Transactions on Artificial Intelligence*, 5(1): 38–50.
- Burbidge, R.; Rowland, J. J.; and King, R. D. 2007. Active learning for regression based on query by committee. In *Intelligent Data Engineering and Automated Learning-IDEAL 2007: 8th International Conference, Birmingham, UK, December 16-19, 2007. Proceedings 8*, 209–218. Springer.
- Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.
- Daneker, M.; Zhang, Z.; Karniadakis, G. E.; and Lu, L. 2023. Systems biology: Identifiability analysis and parameter identification via systems-biology-informed neural networks. In *Computational Modeling of Signaling Networks*, 87–105. Springer.
- Efraimidis, P. S.; and Spirakis, P. G. 2006. Weighted random sampling with a reservoir. *Information processing letters*, 97(5): 181–185.
- Goldbeter, A. 2002. Computational approaches to cellular rhythms. *Nature*, 420(6912): 238–245.
- Guo, X.; Yin, Y.; Dong, C.; Yang, G.; and Zhou, G. 2008. On the class imbalance problem. In *2008 Fourth international conference on natural computation*, volume 4, 192–201. IEEE.
- Hino, H.; and Eguchi, S. 2023. Active learning by query by committee with robust divergences. *Information Geometry*, 6(1): 81–106.
- Kee, S.; Del Castillo, E.; and Runger, G. 2018. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*, 454: 401–418.
- Kurosawa, G.; Mochizuki, A.; and Iwasa, Y. 2002. Comparative study of circadian clock models, in search of processes promoting oscillation. *Journal of theoretical biology*, 216(2): 193–208.
- Lewis, D. D.; and Catlett, J. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, 148–156. Elsevier.
- Liu, H.; Wang, X.; Li, J.; and So, A. M.-C. 2021. Low-Cost Lipschitz-Independent Adaptive Importance Sampling of Stochastic Gradients. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2150–2157. IEEE.
- Liu, S.; and Li, X. 2023. Understanding Uncertainty Sampling. *arXiv preprint arXiv:2307.02719*.
- Liu, Z.; Pu, Y.; Li, F.; Shaffer, C. A.; Hoops, S.; Tyson, J. J.; and Cao, Y. 2012. Hybrid modeling and simulation of stochastic effects on progression through the eukaryotic cell cycle. *The Journal of chemical physics*, 136(3).
- Lu, S.; Hu, Y.; Yang, L.; Sun, Z.; Mei, J.; Tan, J.; and Song, C. 2023. Pa&da: Jointly sampling path and data for consistent nas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11940–11949.
- Murray, J. D. 2002. Mathematical biology: I. An introduction. Interdisciplinary applied mathematics. *Mathematical Biology*, Springer, 17.
- Novak, B.; and Tyson, J. J. 1993. Modeling the cell division cycle: M-phase trigger, oscillations, and size control. *Journal of theoretical biology*, 165(1): 101–134.
- Prigogine, I. 1978. Time, structure, and fluctuations. *Science*, 201(4358): 777–785.
- Reynolds, D. A.; et al. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).
- Saran, A.; Yousefi, S.; Krishnamurthy, A.; Langford, J.; and Ash, J. T. 2023. Streaming active learning with deep neural networks. In *International Conference on Machine Learning*, 30005–30021. PMLR.
- Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.
- Sharma, M.; and Bilgic, M. 2017. Evidence-based uncertainty sampling for active learning. *Data Mining and Knowledge Discovery*, 31: 164–202.
- Stark, J.; Chan, C.; and George, A. J. 2007. Oscillations in the immune system. *Immunological reviews*, 216(1): 213–231.
- Stein, M. 1987. Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2): 143–151.
- Szep, G.; Dalchau, N.; and Csikász-Nagy, A. 2021. Parameter inference with bifurcation diagrams. *Advances in Neural Information Processing Systems*, 34: 5620–5630.
- Tamate, R.; Ueki, T.; Shibayama, M.; and Yoshida, R. 2017. Effect of substrate concentrations on the aggregation behavior and dynamic oscillatory properties of self-oscillating block copolymers. *Physical Chemistry Chemical Physics*, 19(31): 20627–20634.
- Tang, K.; Zhai, J.; Wan, X.; and Yang, C. 2023. Adversarial adaptive sampling: Unify PINN and optimal transport for the approximation of PDEs. *arXiv preprint arXiv:2305.18702*.
- Tomek, I. 1976. Two Modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics*, 6: 769–772.
- Torgo, L.; Ribeiro, R. P.; Pfahringer, B.; and Branco, P. 2013. Smote for regression. In *Portuguese conference on artificial intelligence*, 378–389. Springer.
- Wilson, D. 1972. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, 2(3): 408–421.
- Yazdani, A.; Lu, L.; Raissi, M.; and Karniadakis, G. E. 2020. Systems biology informed deep learning for inferring parameters and hidden dynamics. *PLoS computational biology*, 16(11): e1007575.

Zhu, J.; Wang, H.; Tsou, B. K.; and Ma, M. 2010. Active Learning With Sampling by Uncertainty and Density for Data Annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6): 1323–1331.