Adaptive Sampling for Chemical Kinetics: Zone-Based Strategy and Performance Evaluation

10/09/2025

1. Introduction

Problem Definition

Chemical kinetic modeling requires accurate determination of reaction rate coefficients (K) from experimental or simulated data. The traditional approach involves learning an inverse mapping from chemical composition data (C) to rate coefficients (K) using regression models. However, this process faces significant computational challenges:

- **Forward simulations** (K → C) using tools like LoKI are computationally expensive
- **Inverse problem** (C → K) requires sufficient training data to achieve accurate predictions
- **Sampling efficiency** is critical due to computational cost constraints
- Each rate coefficient has a single true value, creating opportunities for targeted sampling strategies

Previous Sampling Approaches

Prior work in this area typically employed standard sampling strategies for generating training datasets:

1. **Uniform Sampling**: Random uniform distribution across the parameter space
2. **Log-Uniform Sampling**: Uniform sampling in logarithmic space, suitable for parameters spanning multiple orders of magnitude
3. **Latin Hypercube Sampling (LHS)**:
    - **Uniform Latin Hypercube**: Ensures better space-filling properties than pure random sampling by dividing each dimension into equal intervals
    - **Log-Uniform Latin Hypercube**: Combines Latin hypercube structure with logarithmic scaling for wide-range parameters
4. **Morris Method**: A screening method that efficiently explores parameter sensitivity through structured sampling

While these methods provide reasonable coverage of the parameter space, they do not adaptively focus computational resources on regions where the model needs improvement.

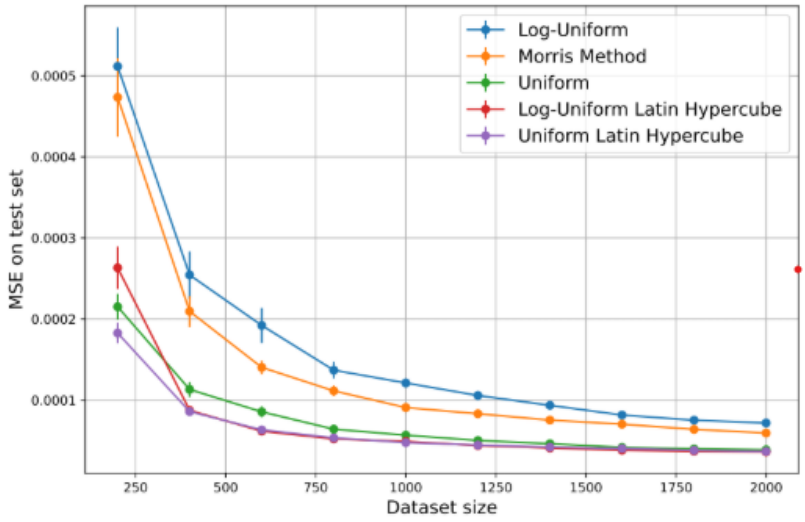


Figure 4.1: Sample efficiency of different methods. The model's performance improves as the number of training points increases until it reaches a plateau. The training dataset was divided into subsets using 20 random seeds to mitigate bias. Error bars denote the standard deviation across these seeds for each number of training points.

Motivation for Adaptive Sampling

To understand whether adaptive sampling could improve efficiency, we conducted a zone-based analysis to investigate how model performance varies across different regions of the parameter space.

Zone-Based Analysis Setup: We divided the parameter space into 6 zones based on **distance to the true rate coefficients (K_true)**, with uniform distribution of parameters across zones. Each zone represents a different "neighborhood" relative to the true solution.

Key Findings That Motivated Adaptive Sampling:

1. **Results are NOT uniform across the 6 zones**
    - Zone 1 (closest to K_true): Consistently lowest MSE (~10^{-5})
    - Zone 6 (farthest from K_true): Highest MSE (~10^{-4}), orders of magnitude worse
    - Clear performance degradation with increasing distance from true values
2. **MSE contribution across zones is NOT uniform**
    - **Zone 6 contributes 50-70% of total error** despite being only 1/6 of the data
    - Zones 1-3 contribute minimal error (~5-15% combined)
    - Error distribution remains highly skewed even with increased training data
3. **Most error reduction comes from Zone 6 decreasing**
    - During model training with both uniform and Latin hypercube sampling
    - Other zones show minimal improvement or plateau quickly
    - Zone 6 remains the primary bottleneck limiting overall model performance

Implication: Since Zone 6 dominates the error and drives most improvements, an adaptive sampling strategy that **targets problematic zones** could potentially achieve better performance with fewer total samples than

traditional uniform approaches.

Figure 1: Zone-based MSE evolution for Uniform sampling

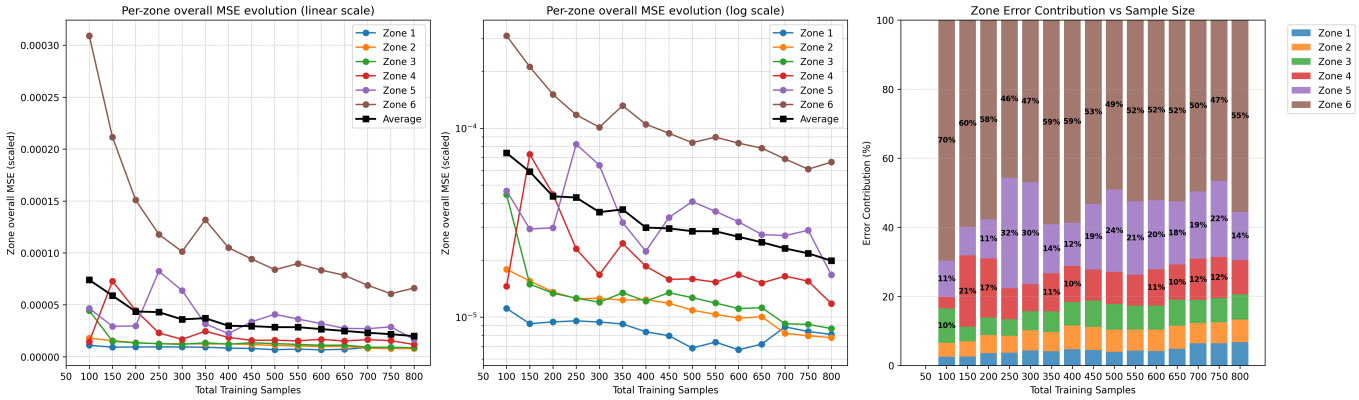
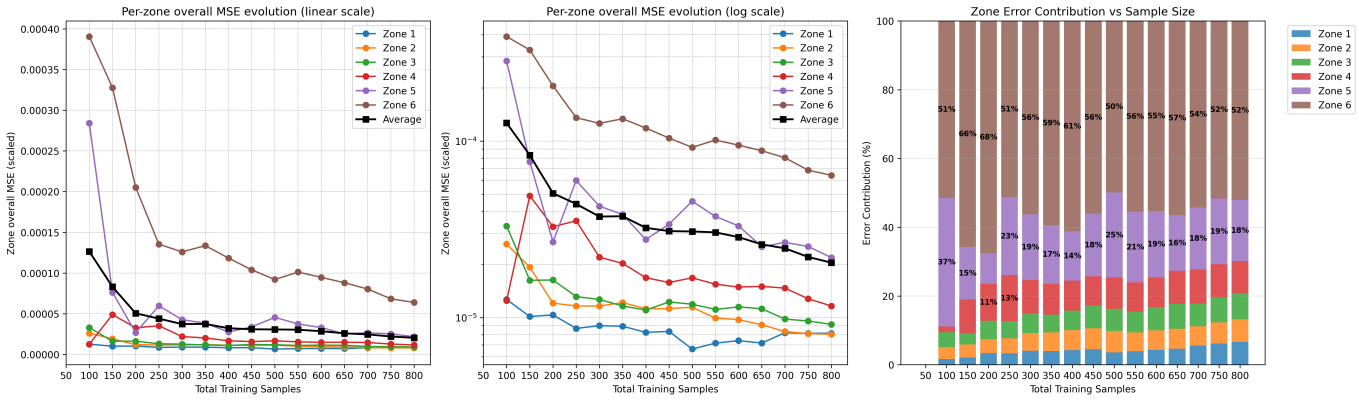


Figure 2: Zone-based MSE evolution for Latin Hypercube sampling



The plots clearly demonstrate the non-uniform error distribution across zones and Zone 6's dominant contribution to both total error and error reduction dynamics.

2. Adaptive Zone-Based Sampling Strategy

Core Algorithm

Based on the zone-based analysis findings, we developed an adaptive sampling strategy that targets high-error zones while maintaining fair comparison with traditional methods:

Initialization:

- **Start with 100 uniform samples** (same as baseline methods)
- Same SVR models and hyperparameters as traditional approaches
- No model architecture changes to ensure fair comparison

Zone-Based Targeting Process:

1. **Zone Definition:** Divide parameter space into **6 zones** based on distance to K_true
2. **Validation-Based Evaluation:**
    - Use dedicated **validation dataset (500 datapoints)** for zone performance assessment
    - Calculate **average MSE per zone** on validation set only
    - **Prevent test set contamination** by never using test data for sampling decisions
3. **Zone Selection:** Model samples from **3 highest-error zones** (top 50% of problematic zones)
4. **Sample Addition:** Add 100 samples per iteration, distributed across the 3 target zones

5. **Model Retraining:** Retrain SVR models with augmented dataset using same hyperparameters

Key Innovation - Clean Validation:

- **Validation set is completely separate** from both training and test data
- **Zone evaluation performed only on validation samples** to guide sampling decisions
- **Test set remains untouched** until final performance evaluation
- This prevents any form of test set leakage or contamination in the sampling process

Experimental Controls:

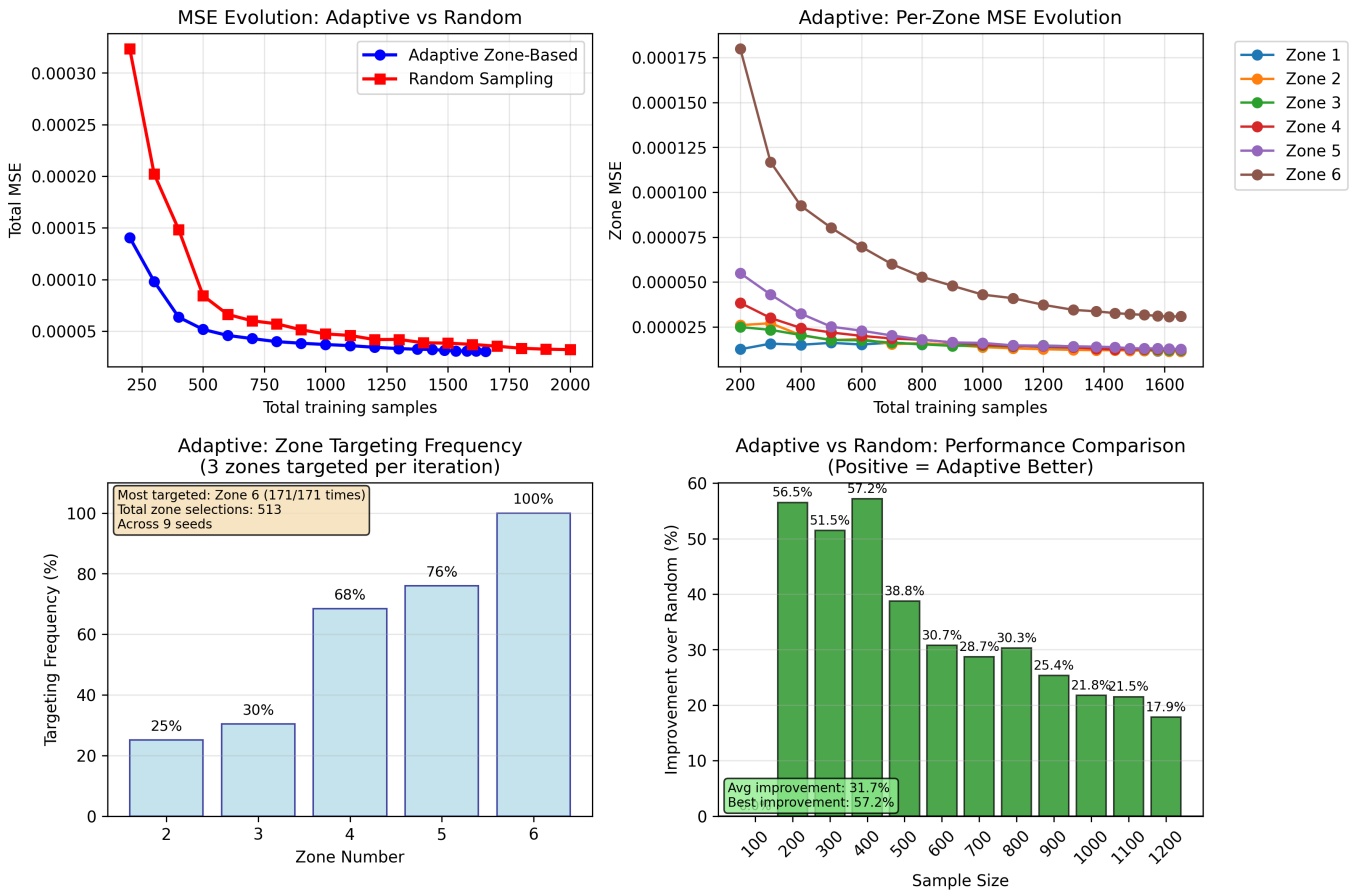
- **Same models:** Identical SVR architecture and hyperparameters as baseline methods
- **Same evaluation:** All methods evaluated on identical test sets
- **Seed strategy:** Multiple random seeds used to shuffle both uniform and batch data
- **Fair comparison:** Only difference is the sampling strategy, not the learning algorithm

Multi-Seed Validation:

- **9 different random seeds** used to shuffle training data order
- Each seed produces different data presentation sequences
- Results averaged across seeds to account for data order effects
- Provides robust statistical comparison across sampling strategies

3. Results

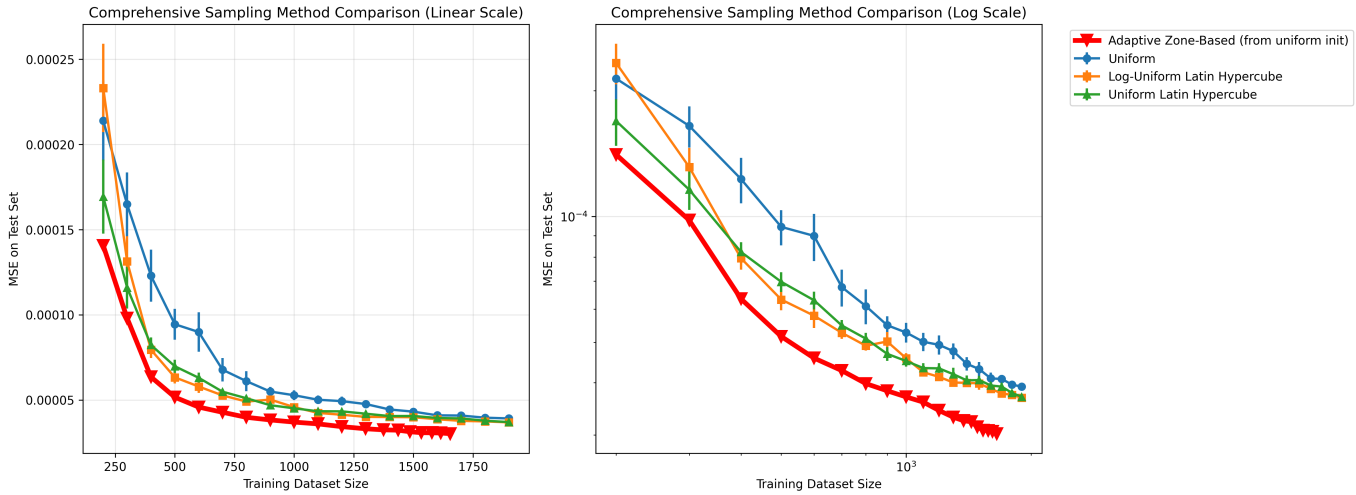
Adaptive vs Random Sampling Performance



Key Performance Findings:

1. **Non-Uniform Zone Targeting**
    - Target zones are **NOT uniformly distributed** across the 6 zones
    - **Zone 6 is always chosen** (100% targeting frequency across all iterations)
    - Zone selection reflects the underlying error distribution, with highest-error zones receiving priority
2. **Superior Performance at Small Sample Sizes**
    - **Adaptive sampling significantly outperforms random sampling**, especially for smaller datasets
    - **Average improvement: 31.7%** across all sample sizes
    - **Best improvement: 57.2%** achieved at smaller sample sizes
    - Performance advantage most pronounced in the critical 200-600 sample range
3. **Consistent Advantage Throughout Training**
    - Adaptive maintains better performance across the entire training curve
    - Improvement diminishes at very large sample sizes (>1000 samples) but remains positive
    - Demonstrates efficient use of computational resources by targeting high-impact regions

Comprehensive Sampling Method Comparison



Adaptive vs All Traditional Methods:

The comprehensive comparison reveals that adaptive zone-based sampling **outperforms all traditional sampling strategies**:

- **Best overall performer:** Adaptive Zone-Based achieves lowest final MSE (3.02×10^{-5})
- **Efficiency advantage:** Reaches best performance with fewer samples (~1654) compared to traditional methods (~1900)
- **Consistent superiority:** Outperforms Uniform, Log-Uniform Latin Hypercube, and Uniform Latin Hypercube across most sample sizes
- **Significant improvement margins:** 18-27% better than the best traditional methods

Strategic Validation: The results validate the core hypothesis that **targeting problematic zones** (particularly Zone 6) leads to more efficient learning compared to uniform space exploration strategies.