# Train-Test Splitting Strategy for CHLA Prediction

In this project, we deviated from the splitting approach used in some previous literature, where a fixed number of lakes (e.g., 49 for training and 50 for testing) were selected to construct train and test datasets. Instead, we opted for a more scientifically robust and reproducible strategy.

## Why Fixed-Lake Splitting Is Problematic

1. Statistical Imbalance: Lakes vary greatly in observation count. Fixed-lake splitting can create uneven distributions of temporal data.  
2. Reduced Generalization: Models trained on a few lakes may overfit specific lake dynamics, performing poorly on unseen lakes.  
3. Loss of Intra-Lake Patterns: Splitting entire lakes removes valuable temporal trends from training data.

## Our Approach: Sample-Wise Stratified Splitting

Instead of assigning entire lakes to train or test sets, we shuffled and split individual weekly sequences across all lakes (e.g., 80% train, 20% test). Lake\_ID was retained and encoded using embeddings.

This guarantees that each lake contributes proportionally to both training and testing, without leaking sequence-level information.

## Benefits of Our Method

- Ensures balanced representation across lakes.  
- Enables generalization across similar lake behaviors.  
- Avoids overfitting to specific lakes.  
- Fully utilizes the dataset's statistical richness.

## Conclusion

This method allows a more data-efficient, scientifically valid approach to CHLA prediction with LSTMs. It also aligns well with modern deep learning practices by respecting the temporal sequence structure while avoiding biased evaluation.

# References

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[2] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.