Enrichment and Acceleration of Edge to Exascale Computational Steering STEM Workflow using Common Metadata Framework

Gayathri Saranathan
Hewlett Packard Labs
Hewlett Packard Enterprise
Singapore
gayathri.saranathan@hpe.com

Martin Foltin
Hewlett Packard Labs
Hewlett Packard Enterprise
Fort Collins, USA
martin.foltin@hpe.com

Aalap Tripathy
Hewlett Packard Labs
Hewlett Packard Enterprise
Austin, USA
aalap.tripathy@hpe.com

Annmary Justine
Hewlett Packard Labs
Hewlett Packard Enterprise
Fort Collins, USA
annmary.roy@hpe.com

Ayana Ghosh
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
ghosha@ornl.gov

Maxim Ziatdinov
Pacific Northwest National Lab,
Richland, WA 99352, USA
maxim.ziatdinov@pnnl.gov

Kevin Roccapriore
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
roccapriorkm@ornl.gov

Suparna Bhattacharya
Hewett Packard Labs
Hewlett Packard Enterprise
Banglore, India
suparna.bhattacharya@hpe.com

Paolo Faraboschi
Hewlett Packard Labs
Hewlett Packard Enterprise
San Jose, USA
paolo.faraboschi@hpe.com

Sreenivas Rangan Sukumaran
CTO, HPC and AI
Hewlett Packard Enterprise
Seattle, USA
sreenivas.sukumar@hpe.com

Computational steering of experiments in automated laboratories has potential to significantly increase scientific research productivity and accelerate discovery. New algorithms have been developed to better capture the structure-property relationships for driving experiments towards exploration of material sites of interest for specific function or that may exhibit new phenomena [1,2]. However, gaps still exist in effectively utilizing prior knowledge and integrating outputs from High Performance Computing (HPC) simulations. These require considerable computing resources whereas experiment control algorithms typically run at the instrument edge. To enable innovation in computational steering workflows spanning edge to HPC datacenter boundaries, new data management infrastructures are needed that facilitate capture and sharing of relevant data.

In this work we start with state-of-the-art Deep Learning – Gaussian Process Kernel (DKL) active learning algorithm running at the edge [1] and enhance it with meta-learning algorithm running at computing facility that helps to integrate prior knowledge from other experimental sites, reducing experiment time, cost, and sample degradation. First, we briefly recap the results published in [3] and focus on data management infrastructure details not published so far that enable algorithm optimization and data sharing between multiple edges and the datacenter. Then, we discuss work in progress on wrapping this inner loop of experiment control with outer loop that uses experimental results to calibrate Molecular Dynamic (MD) simulations in the HPC datacenter, where the simulations in turn help to reconfigure the DKL process for exploration of specific phenomena. We focus on the data management infrastructure that enables both loops in this complex workflow. While inner loop involves sharing of experimental data and AI models, the outer loop also involves sharing the details of how the DKL active learning arrived at structures of interest. This requires capture of the active learning history and sharing it between the edge and the datacenter. We take advantage of the Federated Common Metadata Framework (CMF) [4, 5] (see Figure 1) that captures the entire data lineage from the DKL with references to relevant data slices, enables data versioning for reproducibility, post-hoc analysis and explainability, and seamless sharing between edge-to-HPC sites employing Git-like paradigm. It decouples data and metadata to reduce data movement by limiting it to relevant data subsets.

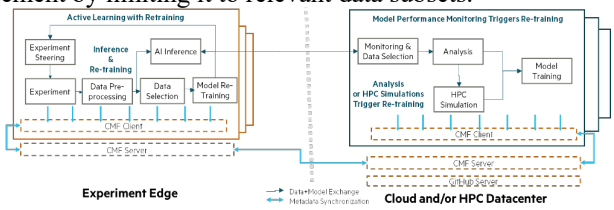


Figure 1: Federated CMF applied to Active-learning Workflow

The DKL, meta-learning, and data management described in this work is applied to Scanning Transmission Electron Microscopy (STEM) experiments at Oak Ridge National Laboratory (ORNL). However, we believe that it has more general applicability. The specific signals from the experiment to drive HPC simulations may be different in other domains, but the general need for providing context of experimental data for simulation by capturing the data lineage from experimental exploration by data infrastructures like CMF will be similar. CMF aids in HPE S/W stack like MLDE, Fed-SDK, etc. The overall workflow in context of a microscopy experiment is schematically depicted by Figure 2.

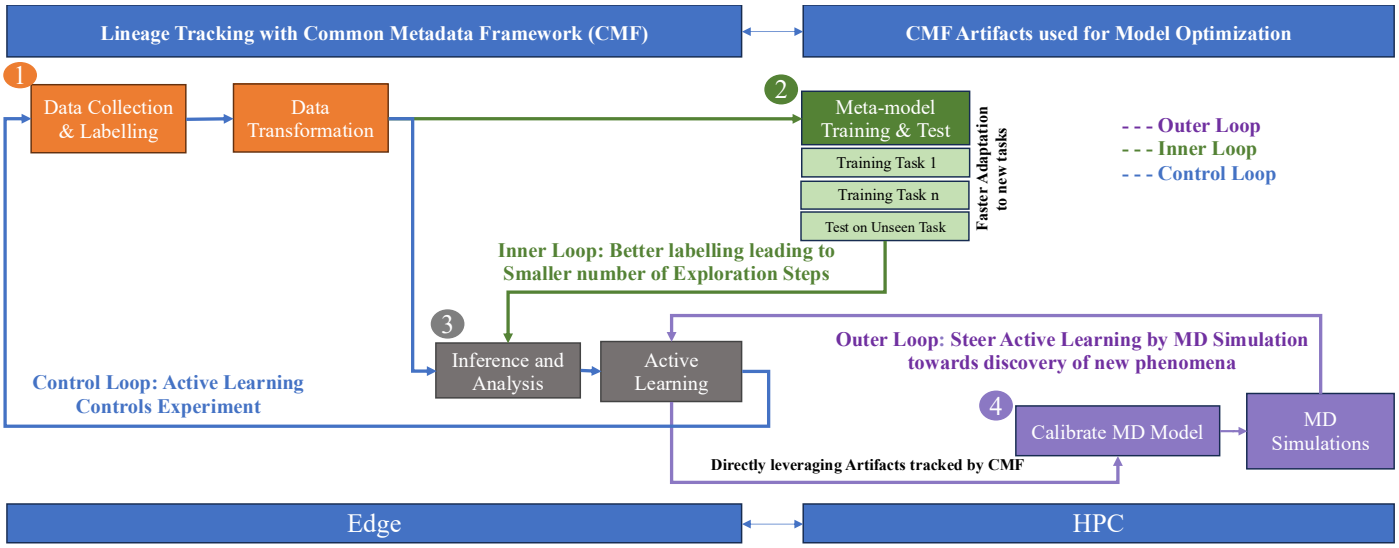


Figure 2: Optimizing Scientific Workflows: Active-Meta Learning with CMF

A. *Autonomous Experimentation Active-Meta Learning using Reptile Algorithm*

Autonomous Experimentation in Microscopy adopts Deep Kernel Learning (AE-DKL) [1] to predict energy spectra from structures, by employing a Gaussian Process Regressor combined with a 4-layer Multi-Layer Perceptron (MLP) feature extractor, enabling autonomous microscopy. It enables AI-driven discoveries of new phenomena by employing active learning that drives the spectroscopy- probe to structural sites with high probability of interesting spectra. It incurs challenges in cost, efficiency, generalization, reproducibility, and real-time analysis. We improve the AE- DKL in our Reptile-DKL workflow (Fig. 3) [3] by training a meta-model in HPC datacenter on prior experiments using Reptile algorithm [6] to provide seed for few-shot model adaptation at the instrumental edge, reducing active learning and experiment times by 30-40% [3].

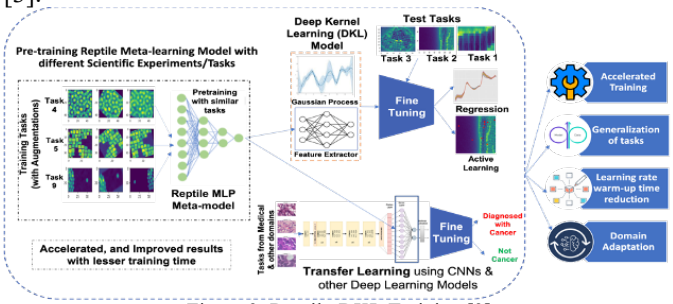


Figure 3: Reptile-DKL Training [3]

The meta-model training involves exploration of optimum sub-sets of prior experiments (the "task" sets) for the best generalization, and optimization of various hyper-parameters, including the number of Reptile iterations. Results and lineages from these explorations (see examples from Table 1 and Figure 4) are captured in CMF to help accelerate meta-model retraining after addition of new experiments, seeding "task" sets and hyper-parameters from historical experience.

Our meta-model trained on plasmonic images also shows good adaptation to other domain: acceleration of tumor classification in Breast Cancer images [7-9] through cross-domain Transfer Learning [3].

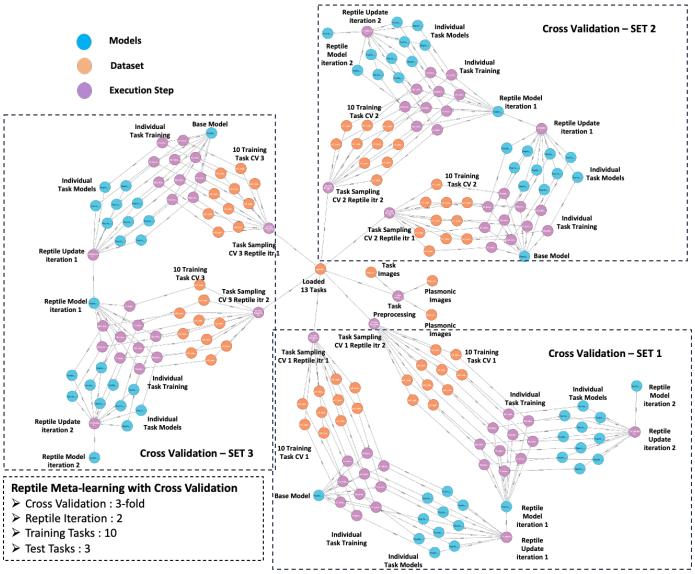


Figure 4: Reptile Reptile-DKL CMF Lineages - training with different task sets

Table 1: Reptile-DKL Accuracies for different task sets and hyperparameter explorations

Reptile Epsilon Parameter: 1e-2			Number of Reptile Iterations			
			30		50	
	Train Tasks	Test Tasks	MLL	RMSE	MLL	RMSE
Cross Validation	[6, 1, 8, 10, 11, 12, 5, 7, 3, 9]	[2, 4, 13]	**22.66**	**0.0109**	**22.81**	**0.0130**
	[10, 13, 6, 8, 5, 7, 2, 4, 12, 3]	[1, 9, 11]	**22.85**	**0.0120**	**23.10**	**0.0178**
	[8, 4, 1, 2, 10, 12, 13, 9, 11, 3]	[5, 6, 7]	**22.36**	**0.0086**	**22.215**	**0.0092**

B. *Bridging Simulation and Experiment: CMF-Enhanced Microscopy Workflow*

Molecular Dynamics (MD) simulations can explore significantly broader set of structures than laboratory experiments. The Reptile DKL method involves 3 stages or execution steps: Task Preprocessing, Full train-dataset training with meta learning, and test-dataset used for inference with Active Learning. The Meta-model training involves Task-wise Batch Sampling stage, Individual Task Training stage producing Individual Task Models, which is concatenated parallelly and used to update the base model parameters. The model has learnt on multiple similar yet different tasks to generalize much faster on the new/unseen tasks.

Therefore, MD has the potential to steer the active learning in Reptile-DKL by informing it what spectral features to prioritize when looking for new phenomena or specific functional behaviors. (The meta learning lineage of the Reptile DKL for 3 meta-model training iterations are shown in Figure 5) However, MD simulation models need to be calibrated to reflect correctly the most interesting features (e.g., defects). In our presentation, we will show initial results from an iterative workflow that employs experimental data to refine simulation model, and simulation results to inform scalarization of experimental spectra to drive active learning.

Active learning lineages captured in the CMF help inform initial conditions for MD simulations. The uncertainties are indicative of structural features that the AI model is least certain about (i.e., not represented in prior knowledge) and should be included in MD calibration. The order in which the locations have been explored is important because it captures points of probe-induced sample heating that affect correct matching between experiment and simulation. Figure 6 shows the active learning lineage captured in CMF along with example locations (shown in Fig. 7) and uncertainties (shown in Table 2). Experimental results with spectral intensities from active learning step are also captured in CMF and used to help parametrize MD simulations. More detail will be given in the presentation.

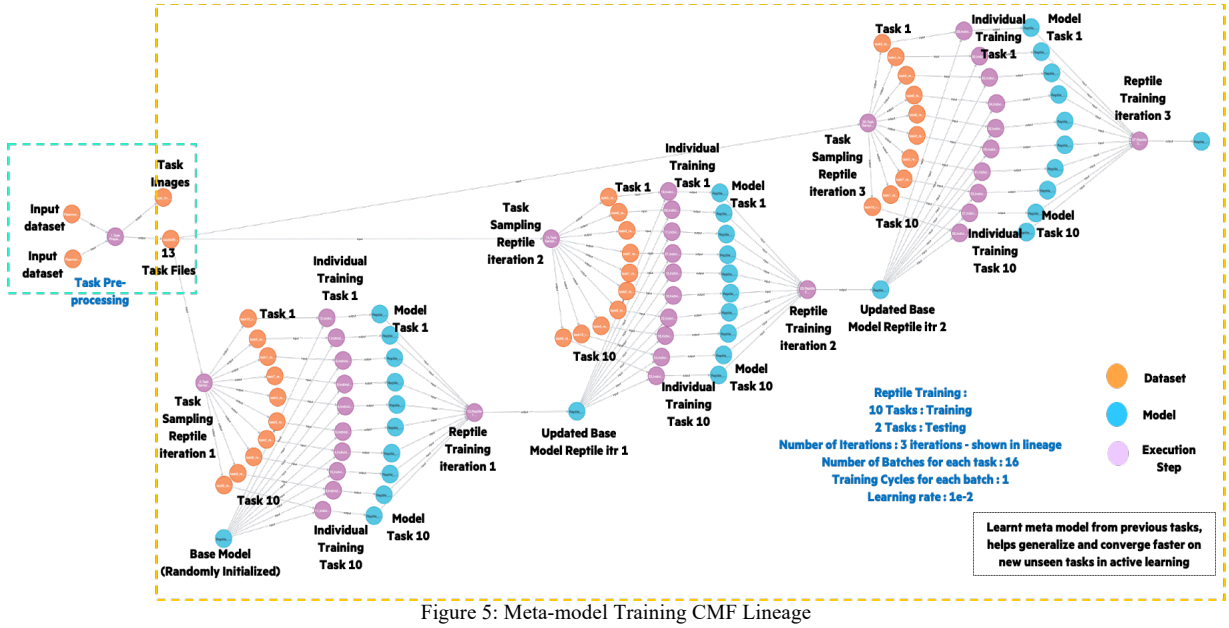


Figure 5: Meta-model Training CMF Lineage

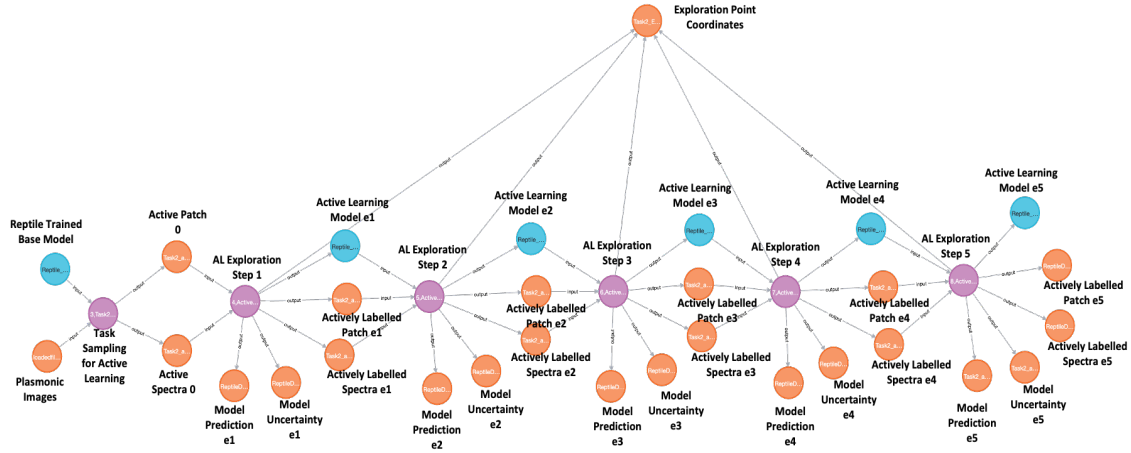
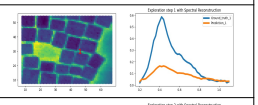
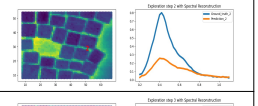
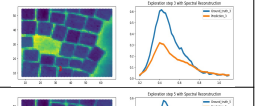
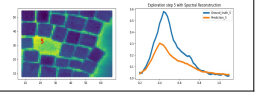


Figure 6: Reptile-DKL guided Active Learning

Table 2: Active Learning Results (Only first 5 steps shown for simplicity)

Exploration Step	Maximum Uncertainty Objective Score	Reconstruction Accuracy	Active Learning Exploration Step
1	0.1865	0.0090	
2	0.1791	0.0035	
3	0.1341	0.0015	
5	0.1622	0.0008	

In summary, our data management infrastructure built on CMF enhances computational steering of experiments:

i) Enables seamless sharing of relevant data subsets between edge and HPC datacenter,

ii) Helps to accelerate AI model training by learning from historical experience,

iii) Validates meta- model adaptability by capturing its few-shot learning trajectory for further examination, and

iv) Captures the evolution of experiments to help calibrate the simulations.

CMF also helps to address time disparity between experiments (fast) and simulations (slow) by aiding workflow manager to select relevant data subsets for staging. These workflows efficiently save training time and reduce compute requirements through reproducibility. Implementation can be found in [10].

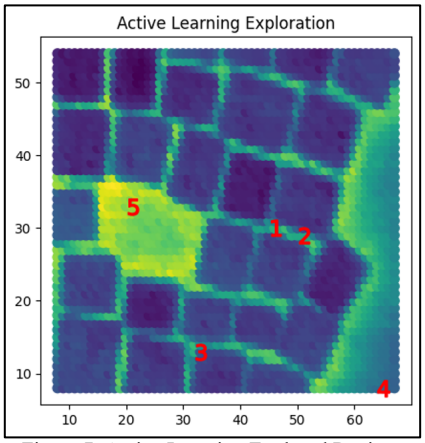


Figure 7: Active Learning Explored Regions

C. *Identifying Stable Trajectories*

This section delves into the application of Ab Initio Molecular Dynamics (AIMD) simulations for investigating the dynamics and characteristics of materials such as MoS2 (molybdenum disulfide) with sulfur (S) defects. AIMD simulations involve modeling atomic and molecular behavior based on fundamental physical principles, like quantum mechanics, without relying on empirical parameters. Here, AIMD simulations are utilized to explore the dynamics and properties of MoS2 with S defects. The dataset comprises image representations of the MoS2 structure at various time points during the AIMD simulations as given in Figure 8, while the ground truths represent corresponding data indicating defect locations or states within these images. These masks serve as reference data for validating or training models to automatically detect defects.

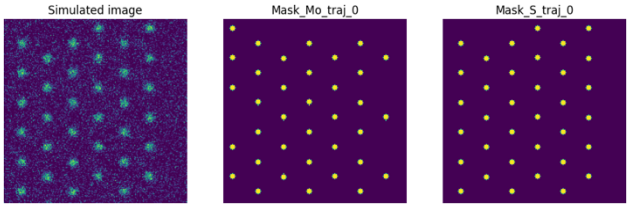


Figure 8: AIMD MoS2 Simulated Structure and Defect locations

By extracting energies linked to different configurations from AIMD trajectories, our aim is to pinpoint the most energetically stable structure among them. In materials science, comprehending the energetic stability of different configurations is vital for predicting their stability under diverse conditions and designing materials with desired properties.

Figure 9 illustrates the exploration of stable energy regions in AIMD simulation: red points are selected through active learning, and blue points are ground truth energies.

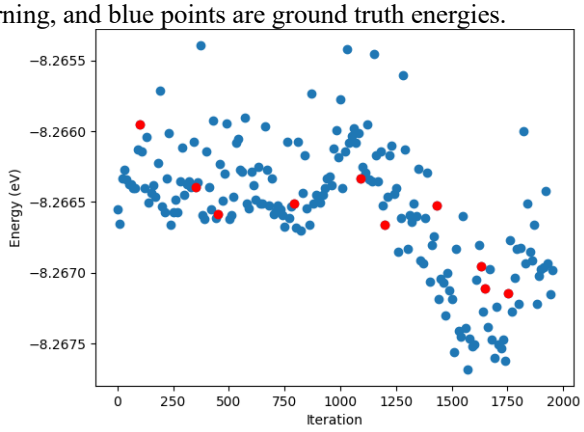


Figure 9: Active Learning Explored Regions in AIMD Simulation

Using an experimentally trained model for active learning on simulations involves leveraging insights gained from physical experiments to guide the selection of data points in simulated environments. This process typically entails employing a meta-model, often constructed using techniques like Bayesian Optimization and deep kernel learning, trained on data acquired from real-world experiments. The meta-model learns to approximate the underlying relationships between input parameters and desired outcomes, enabling efficient exploration of the simulated space. The pre-trained meta-model aids in efficiently identifying stable energy structures across different trajectories, as depicted in Figure 8, based on the maximum uncertainty acquisition function. By actively selecting data points based on the meta-model's predictions, particularly focusing on areas of high significance, the simulation process can be optimized to identify key insights or optimal solutions. The integration of experimentally derived knowledge with simulation-based exploration enhances the efficiency and effectiveness of the overall research or optimization process. This research is ongoing, with continuous experimentation.

D. Future Work

We will explore different ways to steer the active learning in ReptileDKL algorithm by MD simulations to uncover characteristics of various defects sites and other interesting new physical phenomena. Sample evolution study will also be performed based on the performance characteristics and the dynamic changes of the molecules that are logged in CMF. We also intend to deploy it in real-time microscopic sites of ORNL to simultaneously work on the current experimental setup.

Acknowledgements This research (A.G.) is sponsored by the INTERSECT Initiative as part of the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

REFERENCES

[1] Roccapriore, Kevin et. al., *Physics Discovery in Nanoplasmonic Systems via Autonomous Experiments in Scanning Transmission Electron Microscopy*. https://doi.org/10.1002/advs.202203422

[2] Bhowmik, Debsindhu, et. al., *Building an edge computing infrastructure for rapid multi-dimensional electron microscopy*. https://www.osti.gov/servlets/purl/1813209.

[3] Saranathan, Gayathri et. al., *Towards Rapid Autonomous Electron Microscopy with Active Meta-Learning*. https://doi.org/10.1145/3624062.3626085

[4] A. Justine, et. al., *Self-Learning Data Foundation for Scientific AI*, SMC 2022, CCIS volume 1690, Springer 2023, https://link.springer.com/chapter/10.1007/978-3-031-23606-8_2

[5] A. Justine, et. al., *The Case for Decentralized AI Metadata Tracking and Lineage in Science and Engineering Workflows.* 2nd Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE); https://ai-2- ase.github.io/papers/24%5CSubmission%5CAAAI-2023-camera- ready.pdf

[6] Nichol, et. al., *Reptile: a Scalable Metalearning Algorithm*. https://openai.com/research/reptile

[7] Pengyu Yuan et. al., Few Is Enough: Task-Augmented Active Meta-Learning for Brain Cell Classification, https://doi.org/10.48550/arXiv.2007.05009

[8] Abu Al-Haija et. al., Breast Cancer Diagnosis in Histopathological Images Using ResNet-50 Convolutional Neural Network. https://doi.org/10.1109/IEMTRONICS51293.2020.9216455.

[9] Gupta, et. al., Analysis of Histopathological Images for Prediction of Breast Cancer Using Traditional Classifiers with Pre-Trained CNN. https://doi.org/10.1016/j.procs.2020.03.427

[10] https://github.com/atripathy86/AE-DKL/blob/aldkl_reptile/Reptile_DKL_Batch_CMF.ipynb