MTL-FECAM: Bridging the stability-plasticity tradeoff in Exemplar-free Continual Learning

Sakshi Ranjan[1][0000−0002−1740−8366], Niraj Kumar[1][0009−0008−6722−323X], Jatin Kumar[2][0009−0004−3359−0948], and Sanjay Kumar Singh[1][0000−0002−9061−6313]

[1] Indian Institute of Technology (BHU), Varanasi, India-221005
[2] Vellore Institute of Technology, Vellore, India-632014
sakshiranjan.rs.cse21@itbhu.ac.in, nirajkumar.rs.cse22@itbhu.ac.in, jatin.kumar2022@vitstudent.ac.in, sks.cse@iitbhu.ac.in

Abstract. Exemplar-Free Class Incremental Learning (EFCIL) is a specialized form of Class Incremental Learning (CIL), where a model sequentially learns new classes without storing past data. As a subset of Continual Learning (CL), EFCIL presents greater challenges due to its heightened susceptibility to Catastrophic Forgetting (CF), stability-plasticity tradeoff, and feature drift. A recent trend in CIL for pre-trained models involves freezing the feature extractor after the first task and incrementally learning the classifier, attracting considerable attention. To address the research gap of the prototypical network for CIL leveraging new molecular class prototypes that can be generated using a frozen feature extractor, we propose a novel Multi-Task Learner (MTL) using Elastic Weight Consolidation (EWC) and Feature Covariance Aware Metric (FECAM) in ChemBERTa model named MTL-FECAM. The proposed framework uses the Mahalanobis metric for CIL setting for three molecular datasets BBBP, bitter, and sweet. The empirical analysis showcases that modeling feature covariance relationships outperform previous methods that sample features from normal distributions train the MTL and balance stability-plasticity tradeoff and minimum CF (Accuracy - 90.89%, 92.62%, 91.95% and Forgetting Measure (FM) - 0.27, 0.0, 0.0 on BBBP, Bitter, and Sweet datasets respectively). The proposed framework is adaptable to various molecular datasets for both Few-Shot CIL (FSCIL) and Many-Shot CIL (MSCIL), setting it apart from existing approaches. Remarkably, MTL-FECAM achieves state-of-the-art results on multiple standard CL benchmarks. Code link : https://github.com/sran07/MTLFECAM.

Keywords: Continual Learning · Machine Learning · Forgetting

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

1 Introduction

In CL, a model must continuously acquire knowledge from an evolving stream of tasks while only accessing data from the current task. This constraint makes it highly susceptible to CF, where previously learned information is lost. This challenge is particularly significant in CIL, where the goal is to incrementally learn new classes while maintaining high accuracy for all encountered classes without task labels indicating the origin of the evaluated samples [1]. A straightforward approach to mitigate forgetting involves storing exemplars of each class. However, this is impractical in scenarios where storage is limited, or data privacy is a concern, such as in medical imaging. Consequently, research has shifted towards EFCIL methods, which aim to distinguish between old and new classes without retaining past data. Existing EFCIL strategies handle this challenge in two primary ways. Some methods prioritize plasticity by training in new classes while preserving past knowledge via knowledge distillation. Others focus on stability, freezing the feature extractor after the first task and incrementally updating only the classifier [2]. However, a major drawback of freezing the feature extractor is its inability to learn new representations (Fig.1).

To address this, CIL methods, inspired by transfer learning, aim to maximize the utility of pre-trained representations while continuously adapting the classifier to new tasks. Feature-based approaches (e.g., Prototypes, Mahalanobis Classifier [6], Nearest-Class Mean (NCM) [4]) that store statistical summaries instead of data points. Architectural modifications (e.g., Expansion-based models) that add new neurons or layers for each task. These approaches leverage transformer models pre-trained on large-scale molecular datasets such as ZINC, HIV, etc., and primarily focus on refining the classifier while keeping the feature extractor frozen (Fig. 2). For task-wise incremental learning in Molecular property Prediction (MP): Train on BBBP (Task 1) implies learning features and storing feature representations; train on Bitter (Task 2) implies avoiding forgetting BBBP knowledge; train on Sweet (Task 3) implies avoiding forgetting both BBBP and Bitter while learning sweet [3].

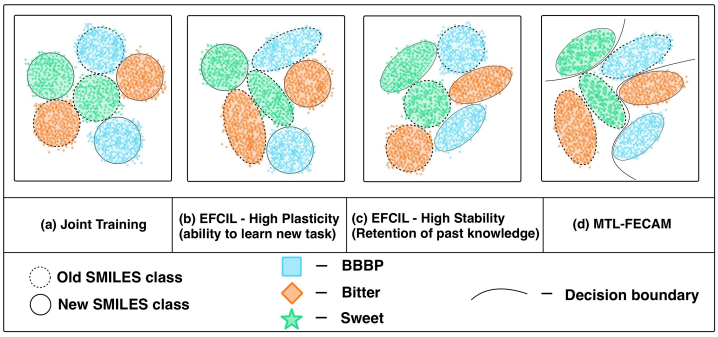

Fig. 1: Feature representations in EFCIL settings.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

Drug discovery integrates genomic, proteomic, chemical, clinical trials, and experimental data, forming a framework for CL in Artificial Intelligence (AI) research. CL enhances model accuracy by seamlessly incorporating diverse data over time, supporting personalized medicine with continuous updates for individualized treatment recommendations [38]. Integrating diverse data types can be time-consuming and error-prone. Cost and time constraints exist when validating new drugs. ML methods leverage complex molecular interactions, build predictive, efficient, and scalable models, integrate multi-modal data with feature selection and optimization, and generalize to unseen molecules to address the issue [36,37]. Furthermore, MSCIL occurs when an ML model receives a substantial number of labeled samples per class while learning new classes, allowing it to establish well-defined decision boundaries. For instance, in case of having thousands of SMILES molecules labeled as BBBP+ (1) or BBBP- (0), it is a MSCIL scenario. Each new molecular property (e.g., Bitter, Sweet) is learned with many examples per class in the incremental phase [39, 40]. In FSCIL, the Machine Learning (ML) model receives very few labeled samples per new class in the incremental phase, and it must generalize from limited examples, often using meta-learning, prototypes, or feature adaptation techniques. For instance, to introduce a new property like "Toxicity" with only 5–10 molecules labeled for each class (Toxic/Non-Toxic), the model must learn to classify Toxicity with minimal data while retaining knowledge of previously learned properties [5].

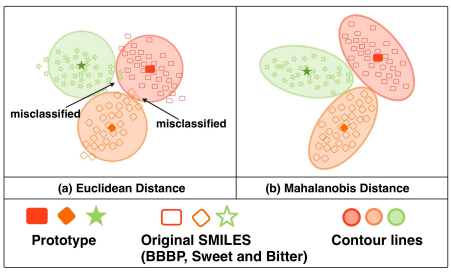

Fig. 2: Illustration of distances for SMILES input

MSCIL faces semantic drift as new tasks arrive, often mitigated by knowledge distillation or regularization, though these require storing past data—impractical for privacy-sensitive applications. We adopt an alternative that freezes the feature extractor after the first task, training only the classifier for new classes [8]. FSCIL, in contrast, deals with limited samples per class, often using meta-learning or variational inference for adaptation. Most methods average feature embeddings to form class prototypes, classifying via Euclidean distance. We explore prototype-based learning in both MSCIL and FSCIL for molecular property prediction, leveraging Mahalanobis distance for improved feature distribution modeling in cross-domain CL [7, 35]. We propose MTL-FECAM, a novel

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

MTL model that integrates FECAM (for stability-plasticity tradeoff and semantic drift mitigation), Mahalanobis metric, and a regularization approach - EWC (mitigating CF) in ChemBERTa model. FECAM leverages feature covariance normalization and a Bayes classifier for optimal decision boundaries learning by considering covariance-adjusted feature distributions, while EWC preserves key parameters using Fisher information. This approach enables scalable, privacy-preserving incremental learning for MP. The contributions of this work are summarized as follows :

- This study depicts a quantifiable association between stability and plasticity, a significant obstacle in CL for MP, which has not been tackled in bioinformatics. To overcome CF, we propose a novel model, MTL-FECAM, which uses an optimal Bayes classifier in the ChemBERTa model by modeling the covariance relations and molecular class prototypes to boost the model's performance.
- This framework models the feature covariance connections using a Mahalanobis metric to learn better non-linear decision boundaries for new classes of three molecular datasets - BBBP, bitter, and sweet.
- This framework is easily implementable and is used for both MSCIL and FSCIL approaches and depicts better learning of optimal decision boundaries. We calculate the covariance matrix for molecular classes using feature embeddings from the training samples and apply correlation normalization to standardize variances across class representations, ensuring reliable distance comparisons.
- Furthermore, the performance of the proposed model outperforms the baseline models upon comparison by conducting extensive empirical analysis and visualization through continuous retraining on the learned tasks.

The arrangement of this study is split into four sections. Section 2 explains the methodology used. Section 3 introduces the experimental analysis. Section 4 highlights the conclusion of the work.

2 Methodology

In this section, Fig. 1 shows the proposed methodology of the MTL-FECAM model, which is described below :

1. Dataset and Featurization The datasets used in this study are gathered from the MolecularNet [26] for the classification tasks. The pre-processing techniques used are - canonicalization, padding and truncation, and encodings. Molecules can have multiple valid SMILES notations (e.g., C1=CC=CC =C1 vs. c1ccccc1 for benzene). BBBP dataset (DATASET-I) is designed for modeling whether the compound will penetrate the blood-brain barrier. It comprised 2049 compounds with four attributes: Name, SMILES, Label, and Reference. For the bitter (DATASET-II) and sweet (DATASET-III) datasets, after pre-processing, i.e., eliminating the molecules that didn't contain the exact details about the bitter and sweet tastes or presenting some conflicting, unavailable facts, datasets

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

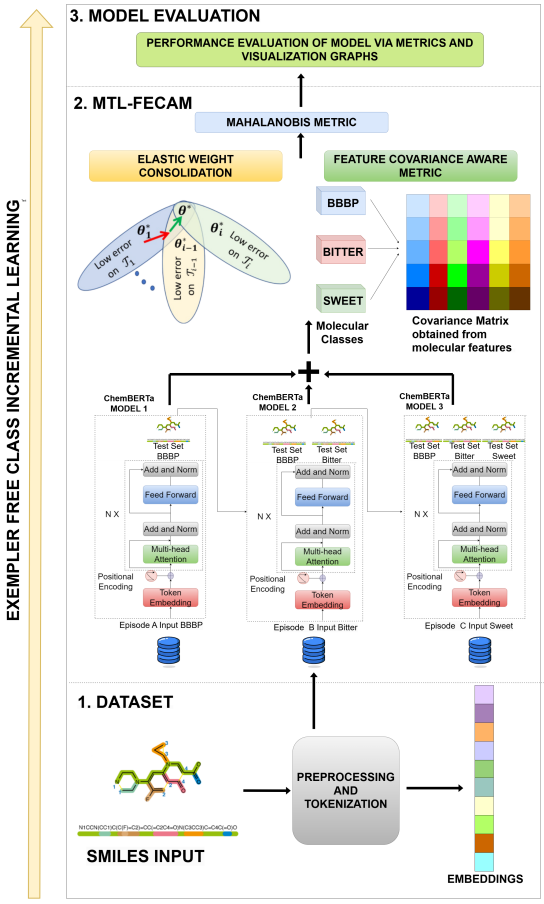

Fig. 3: Proposed Methodology

comprised 1698 bitter and 2860 non-bitter (sweet/tasteless) compounds and 2411 sweet and 2147 non-sweet (bitter/tasteless) compounds.

2. MTL-FECAM Model - In MP, feature representations derived from Deep Learning (DL) models play a crucial role in classification tasks. For CIL on molecular datasets - BBBP, Bitter, and Sweet, it is essential to determine how to assign molecules, represented as SMILES, to their respective classes. Traditional classification methods, such as the NCM classifier, employ the squared Mahalanobis distance D_M instead of the Euclidean distance to assign a given sample x to the closest class mean μ_y:

$$y^* = \arg\min_{y=1,\ldots,Y} D_M(x, \mu_y), \quad D_M(x, \mu_y) = (x - \mu_y)^T M (x - \mu_y) \tag{1}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

where Y is the number of molecular property classes (e.g., BBBP, Bitter, Sweet), and M is a positive definite matrix. The class mean μ_y is computed as the average of feature representations of all molecules belonging to class y. With the shift toward deep molecular representations, where a neural network $\phi : X \to \mathbb{R}^D$ extracts feature vectors from molecular SMILES, learning the Mahalanobis metric M may no longer be necessary. Instead, the isotropic Euclidean distance is often used for classification:

$$y^* = \arg\min_{y=1,\ldots,Y} D_E(\phi(x), \mu_y), \quad D_E(\phi(x), \mu_y) = (\phi(x) - \mu_y)^T(\phi(x) - \mu_y) \quad (2)$$

where $\phi(x)$ represents the learned molecular feature vector, and μ_y is the class prototype (average feature vector of class y). In Euclidean space, $M = I$, where I is the identity matrix. The effectiveness of the NCM classifier with Euclidean distance has been widely adopted in CIL settings. In EFCIL with molecular data, feature representations evolve incrementally, making Euclidean distance less effective. We compare Euclidean and Mahalanobis distances in a CL model trained on Bitter and BBBP molecules and later extended to Sweet. While Euclidean distance performs well on old classes, it struggles with new ones. Mahalanobis distance, however, consistently improves classification by better handling heterogeneous molecular feature distributions

Given that feature distributions follow a multivariate normal distribution $N(\mu_y, \Sigma_y)$, the probability of a molecular feature x belonging to class y for ChemBERTa model is given by:

$$P(x|C = y) \approx \exp\left(-\frac{1}{2}(x - \mu_y)^T \Sigma_y^{-1}(x - \mu_y)\right) \quad (3)$$

This corresponds to the optimal Bayesian classifier, where classification is performed based on the posterior probability:

$$\arg\max_y P(Y|X) = \arg\max_y P(X|Y)P(Y) = \arg\max_y P(X|Y) \quad (4)$$

Since applying the logarithm preserves the ordering, the decision boundary can be rewritten in terms of the squared Mahalanobis distance:

$$\arg\max_y \log P(X|Y) = \arg\min_y D_M(x, \mu_y) \quad (5)$$

$$D_M(x, \mu_y) = (x - \mu_y)^T \Sigma_y^{-1}(x - \mu_y). \quad (6)$$

Feature Representation with ChemBERTa - We extract molecular representations from SMILES using ChemBERTa, denoted as $\phi(x)$, where $\phi : X \to \mathbb{R}^D$ maps input SMILES sequences to a D-dimensional feature space. These features serve as the basis for classifying molecular properties in a MTL framework.

Covariance Matrix Approximation - estimating the covariance matrices from ChemBERTa's feature space to model feature distributions include:

1. Common Covariance Matrix ($\Sigma_{1:t}$) - It is incrementally updated as the mean covariance across all MP tasks seen up to task t:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

$$\Sigma_{1:t} = \Sigma_{1:t-1} \cdot \frac{|Y_{1:t-1}|}{|Y_{1:t}|} + \Sigma_t \cdot \frac{|Y_{1:t}| - |Y_{1:t-1}|}{|Y_{1:t}|} \tag{7}$$

where Σ_t represents the covariance matrix of new molecular classes introduced in task t.

2. Class-Specific Covariance Matrices (Σ_y) - maintains separate covariance matrices for each molecular class y by storing one covariance matrix per class, providing a more flexible representation of feature distributions:

$$\Sigma_y = \frac{1}{|X_y|} \sum_{x \in X_y} (\phi(x) - \mu_y)(\phi(x) - \mu_y)^T. \tag{8}$$

Normalization of Covariance Matrices - Covariance matrices are normalized using the correlation matrix to ensure comparability:

$$\hat{\Sigma}_y(i,j) = \frac{\Sigma_y(i,j)}{\sigma_y(i)\sigma_y(j)}, \tag{9}$$

where $\sigma_y(i) = \sqrt{\Sigma_y(i,i)}$ and $\sigma_y(j) = \sqrt{\Sigma_y(j,j)}$ are the standard deviations along feature dimensions.

Covariance Shrinkage - is used where the feature dimensionality is large, the covariance matrix may be non-invertible due to a limited number of training samples per molecular class:

$$\Sigma_s = \Sigma + \gamma_1 V_1 I + \gamma_2 V_2 (1 - I), \tag{10}$$

where V_1 is average diagonal variance, V_2 is average off-diagonal covariance, and I is identity matrix.

Tukey's Ladder of Powers Transformation - It reduces feature skewness and enforces a more Gaussian-like distribution:

$$\tilde{\phi}(x) = \begin{cases} \phi(x)^\lambda, & \text{if } \lambda \neq 0, \\ \log(\phi(x)), & \text{if } \lambda = 0. \end{cases} \tag{11}$$

where λ is a transformation parameter (set to 0.5 in our experiments). Normalized features $\tilde{\phi}(x)$ are then used for covariance matrix estimation.

Final Prediction - Using transformed features and class-specific covariance matrices, the classification decision is made as:

$$y^* = \arg\min_{y=1,\ldots,Y} D_M(\tilde{\phi}(x), \mu_{\tilde{y}}), \tag{12}$$

$$D_M(\tilde{\phi}(x), \mu_{\tilde{y}}) = (\tilde{\phi}(x) - \mu_{\tilde{y}})^T \hat{\Sigma}_y^{-1} (\tilde{\phi}(x) - \mu_{\tilde{y}}) \tag{13}$$

and $\hat{\Sigma}_y^{-1}$ denotes the inverse of the covariance matrix after shrinkage and normalization.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

3 Experimental Analysis and Discussions

The empirical analysis outcomes are described in four aspects. Models are trained and tested on a T4 GPU with 52GB RAM and coded in Python; training for 3000 epochs with a patience value of 1000 using performance metrics are Test Accuracy (TA), Forgetting Measure (FM), and Anytime Average Accuracy (AAA).

3.1 Comparison of the Model's Performance

MTL-FECAM achieves the highest accuracy across all three datasets at all incremental tasks, FM = 0.0, indicating no CF. This suggests that MTL-FECAM effectively preserves past knowledge while learning new tasks. This is because MTL learns a shared representation across multiple MP tasks and prevents task interference by optimizing task-invariant and task-specific features together. The covariance matrix models the covariance between different feature dimensions. It helps in the consolidation of important features by preserving not only individual weight importance but also feature relationships. Additionally, the Mahalanobis metric considers feature distributions and class prototypes, ensuring better separation between classes. In contrast, oEWC and LwF-MC show strong accuracy but still suffer some forgetting (FM $>$ 0.0). DeeSIL, MUC, SDC, FeTrIL, and SSRE exhibit low accuracy and high FM, implying they struggle with knowledge retention. PASS and IL2A perform well but still exhibit minor forgetting. These methods fail because they train models sequentially without retaining prior task knowledge and constrain weight updates but do not actively promote knowledge transfer (Table 1 and Table 2).

Table 1: Incremental test accuracy and forgetting measure in exemplar-free MSCIL with different incremental tasks.

CIL METHODS	**BBBP**				**BITTER**				**SWEET**			
	ACCURACY			**FM**	**ACCURACY**			**FM**	**ACCURACY**			**FM**
	T=5	**T=10**	**T=20**		**T=5**	**T=10**	**T=20**		**T=5**	**T=10**	**T=20**	
oEWC [12]	85.80	83.71	90.80	0.0	86.52	85.26	89.97	0.0	84.20	87.42	88.74	0.0
LwF-MC [18]	86.20	86.44	86.54	0.0	86.55	89.05	92.14	0.0	82.54	88.24	91.09	0.0
DeeSIL [14]	47.71	65.61	58.15	0.125	42.26	40.65	41.11	0.011	49.46	47.23	49.92	0.499
MUC [15]	74.73	72.29	76.10	0.125	61.79	59.57	64.50	0.002	58.59	64.87	57.45	0.307
SDC [16]	68.63	58.20	65.61	0.064	49.98	61.89	63.14	0.208	64.69	56.40	61.16	0.223
PASS [10]	84.97	86.63	84.76	0.005	84.19	86.60	86.11	0.010	82.29	84.90	85.59	0.019
IL2A [11]	76.10	23.56	39.66	0.627	43.93	45.36	41.84	0.140	47.29	49.71	57.36	0.284
SSRE [17]	54.83	51.07	47.32	0.199	63.52	54.59	53.30	0.252	47.46	58.49	58.30	0.346
FeTrIL [9]	23.56	23.56	45.07	0.444	34.31	38.15	38.15	0.064	52.73	48.46	48.33	0.470
Eucl-NCM [4]	77.12	53.02	60.68	0.167	63.73	56.28	62.77	0.186	63.83	61.51	62.89	0.148
MTL-FECAM	**89.32**	**86.24**	**90.89**	**0.272**	**85.63**	**89.78**	**92.62**	**0.0**	**84.95**	**86.75**	**91.95**	**0.0**

The key observations from Table 3 are as follows: MTL-FECAM outperforms other methods by effectively balancing learning stability, task adaptability, and

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

Table 2: Anytime average accuracy in exemplar-free MSCIL with different incremental tasks.

CIL METHODS	BBBP			BITTER			SWEET		
	T=5	T=10	T=20	T=5	T=10	T=20	T=5	T=10	T=20
oEWC [12]	86.88	85.85	87.67	85.17	86.81	88.13	84.86	88.57	90.08
LwF-MC [18]	89.80	90.34	90.88	85.34	85.79	87.21	84.84	86.34	87.91
DeeSIL [14]	74.05	74.49	74.73	49.35	51.06	50.52	40.22	40.39	42.60
MUC [15]	70.93	66.12	57.46	61.60	61.74	62.32	58.01	60.45	59.21
SDC [16]	76.29	69.71	66.62	55.05	55.11	55.35	63.83	60.43	60.68
PASS [10]	89.51	89.08	89.12	84.19	85.24	85.53	89.29	83.59	84.26
IL2A [11]	89.00	88.63	88.36	87.30	87..62	88.18	84.00	86.07	86.41
SSRE [17]	63.27	58.68	56.47	62.45	62.20	62.44	66.88	61.84	63.65
FeTrIL [9]	55.26	54.71	45.20	61.14	64.85	61.12	54.26	58.66	56.14
Eucl-NCM [4]	64.88	64.88	64.88	72.24	72.24	72.24	73.68	73.68	73.68
MTL-FECAM	**90.17**	**90.67**	**91.11**	**89.41**	**90.22**	**90.93**	**88.61**	**89.57**	**90.41**

memory efficiency while preventing catastrophic forgetting. It leverages a multi-task framework and memory consolidation to enhance feature discrimination across tasks. In contrast, methods like CoIL, WA, BiC, FOSTER, DER, and MEMO exhibit negative FM values, indicating a decline in accuracy due to inadequate knowledge retention. PODNet and iCaRL struggle with task adaptation as they rely on fixed feature extractors or rigid distillation constraints, limiting flexibility. WA and BiC underperform due to simple feature alignment, which fails to generalize to evolving molecular representations. DER and MEMO suffer from overfitting and inefficient memory utilization, leading to poor generalization.

Table 3: Comparison of the proposed method with exemplar-based methods storing 2000 exemplars for #P parameters.

CIL METHODS	#P	BBBP			BITTER			SWEET		
		TA	AAA	FM	TA	AAA	FM	TA	AAA	FM
iCaRL [18]	11.17	64.88	64.88	0.00	72.24	72.24	0.00	73.68	73.68	0.00
PODNet [19]	11.17	87.41	86.72	0.00	80.83	80.09	0.00	79.89	79.66	0.00
CoiL [20]	11.17	90.00	89.20	-0.0141	83.77	82.50	-0.0127	84.27	82.87	-0.0163
WA [21]	11.17	89.46	86.08	-0.0449	84.84	83.51	-0.0196	83.52	82.75	-0.0194
BiC [22]	11.17	89.80	88.46	-0.0312	84.67	83.64	-0.0211	83.06	81.99	-0.0250
FOSTER [23]	11.17	89.07	88.20	-0.0146	85.34	82.37	-0.0578	84.69	82.79	-0.0346
DER [24]	67.02	90.00	88.28	-0.0293	85.92	84.15	-0.0363	84.54	82.62	-0.0338
MEMO [25]	53.14	88.05	88.05	0.0015	86.38	84.80	-0.0386	84.54	82.33	-0.0434
MTL-FECAM	**11.17**	**92.44**	**91.11**	**0.00**	**94.49**	**90.93**	**0.00**	**94.64**	**90.41**	**0.00**

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

3.2 Comparison with Baselines

Table 4 shows MTL-FECAM achieves the best accuracy across all datasets, outperforming baseline models. The MTL enables shared feature representations across tasks, leading to better generalization, while the FECAM stabilizes feature importance over different molecular representations, reducing the risk of feature drift. Additionally, the Mahalanobis metric improves class separation, leading to better discrimination of molecular properties in all datasets. Graph-based SSL models struggle with forgetting, leading to lower performance. Transformer-based models are strong but lack explicit continual learning mechanisms. Weave and MoMu perform well on specific datasets but fail to generalize. Traditional graph-based SSL methods struggle with CF and generalization, leading to significantly lower performance, while transformer-based models are strong, they lack effective feature stability mechanisms, making them less robust in incremental molecular learning

Table 4: Comparative analysis of test accuracy for baselines.

Model	Sweet	ToxCast	SIDER	Bitter	COCONUT
MoMu [27]	77.60±0.56	66.23±0.66	56.50±0.88	73.90±0.36	85.00±0.25
Mole-BERT [28]	81.80±0.25	67.30±0.47	56.80±0.85	76.90±0.40	84.00±0.56
InfoGraph [29]	51.00±0.15	64.52±0.45	55.59±1.25	67.00±0.57	86.00±0.15
GPTGNN [30]	54.00±0.22	63.80±1.25	56.63±0.66	63.45±0.59	87.00±0.41
DGI [31]	54.80±0.62	64.20±1.36	56.12±0.55	64.00±0.68	87.49±2.10
MGSSL [32]	55.50±0.34	65.56±0.37	55.91±0.49	65.00±0.22	89.00±0.68
Transformer [33]	81.00±0.50	66.00±0.14	56.00±0.34	82.00±0.49	87.00±0.30
Weave [34]	80.50±0.42	64.00±0.42	54.00±0.34	81.00±0.29	84.00±0.47
MTL-FECAM	**91.95±0.63**	**68.55±0.16**	**66.71±0.29**	**92.62±0.36**	**91.07±0.15**

3.3 Ablation Study

Table 5 shows that the best performance is achieved with Mahalanobis distance when Turkey's equation, shrinkage estimation, and normalization are applied together, reinforcing the importance of feature correlation and distribution adjustments in MP tasks. In this study, Mahalanobis considers feature correlations, making it better suited for high-dimensional molecular data. Furthermore, the covariance matrix allows Mahalanobis distance to account for feature correlations, improving discriminability. Additionally, adding Turkey's and Shrinkage's equation further boosts accuracy by refining feature distribution adjustment, leading to better class separation, and helps stabilize covariance estimation, especially when sample sizes are small. The aggregation of these equations depicts a synergistic effect, reducing distortions caused by correlated features, where each component refines Mahalanobis distance, making it more robust.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

3.4 Graphical Analysis

In Fig. 4, MTL-FECAM effectively balances stability and plasticity in CL, making it a strong candidate for MP in exemplar-free, privacy-preserving settings. It maintains higher accuracy as the number of incremental sessions increases, demonstrating superior resistance to CF and better learning of new tasks. The performance gap increases over time, highlighting its better long-term stability and ensures optimal decision boundaries by adjusting class feature distributions. BBBP shows the steepest decline in accuracy, suggesting that retaining prior knowledge for this dataset is more challenging, while in Bitter and Sweet datasets performance degrades more gradually, indicating better feature separability and incremental adaptation. In Fig. 5, MTL-FECAM has the slowest decline in accuracy compared to other methods, suggesting superior knowledge retention and plasticity-stability balance. IL2A and FeTrIL show competitive results, indicating that their strategies for feature adaptation and incremental learning are effective. PASS and SSRE rely on distillation, and fail to maintain knowledge over time. DeeDIL and MUC perform slightly better but still experience significant CF. In Fig. 6 ensures better feature alignment and better feature space organization, preserving key parameters using Turkey's transformation.

Table 5: Ablation study using test accuracy.

DISTANCE	Covariance Matrix	Turkey Eqn.	Shrinkage Eqn.	Normalization Eqn.	BBBP	BITTER	SWEET
Eucledian	-	×	-	-	55.77	58.18	59.47
Eucledian	-	✓	-	-	62.00	59.09	50.27
Mahalanobis	Full	×	×	×	60.00	57.99	65.75
Mahalanobis	Full	✓	×	×	63.12	64.31	50.69
Mahalanobis	Full	×	×	✓	50.61	61.93	52.67
Mahalanobis	Full	×	✓	×	62.54	62.77	52.48
Mahalanobis	Full	✓	✓	×	66.44	63.75	64.27
Mahalanobis	Full	×	✓	✓	71.80	62.64	64.52
Mahalanobis	Full	✓	✓	✓	**73.85**	**68.04**	**67.00**

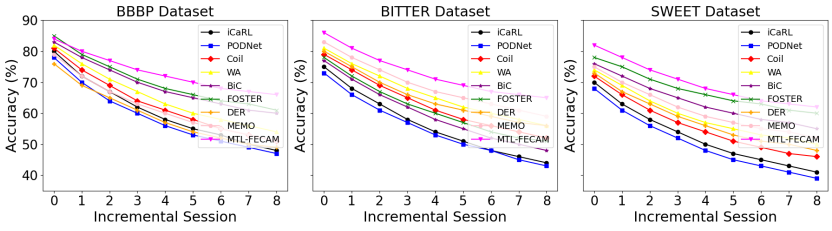

Fig. 4: FSCIL methods accuracy of each incremental task for molecular datasets

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

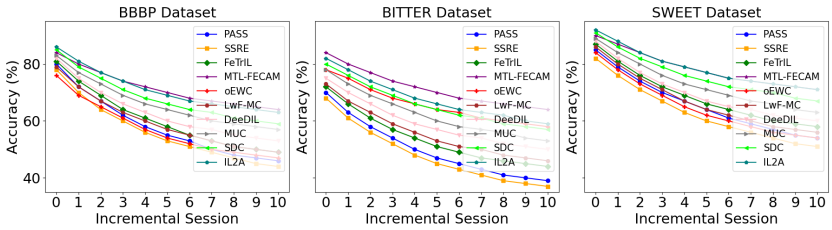

Fig. 5: Accuracy of each incremental task for molecular datasets and multiple MSCIL methods.

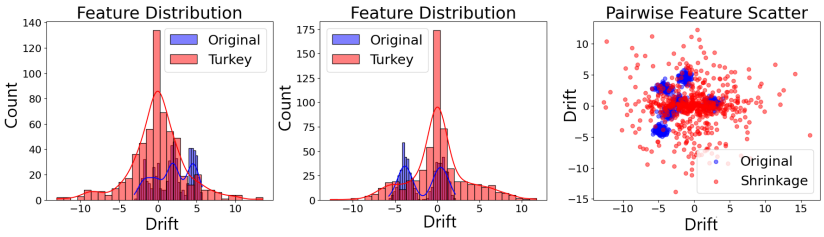

Fig. 6: Scatterplots for old & new molecular classes with Turkey's transformation.

3.5 Key Findings

The key takeaways of this study are described as follows:

1. Superior Incremental Test Accuracy & Forgetting Minimization - MTL-FECAM consistently outperforms all EFCIL methods in TA across all incremental tasks. FM is minimal or zero for MTL-FECAM, indicating effective knowledge retention. Competing methods like oEWC and LwF-MC show some knowledge retention but struggle with feature stability. Methods such as DeeSIL, MUC, SDC, FeTrIL, and SSRE exhibit higher CF.

2. Stability-Plasticity tradeoff in Incremental Learning - MTL-FECAM achieves the highest AAA across incremental tasks, confirming its adaptability in a CIL setting. Compared to oEWC, LwF-MC, and IL2A, MTL-FECAM better balances the stability-plasticity tradeoff, ensuring both knowledge retention and adaptability to new tasks.

3. Baselines on Molecular Property Prediction - MTL-FECAM achieves the highest accuracy compared to other state-of-the-art models like MoMu, InfoGraph, GPTGNN, and Transformer-based methods. This highlights its effectiveness in learning molecular features while preventing CF.

4. Comparison with Exemplar-Based CIL Methods - MTL-FECAM achieves comparable or superior performance against exemplar-based methods like iCaRL, PODNet, and DER, even though it operates without storing exemplars. The proposed method outperforms FOSTER, MEMO, and other replay-based methods in terms of AAA and TA while maintaining minimal forgetting.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

4 Conclusion

This study presents MTL-FECAM, a novel MTL framework that integrates FECAM, the Mahalanobis metric, and EWC within ChemBERTa to address CL challenges in MP tasks and bridge a critical gap in bioinformatics CL. The model effectively tackles CF and semantic drift while balancing stability and plasticity - a major limitation in bioinformatics overlooked by researchers. By leveraging covariance-adjusted feature normalization and an optimal Bayes classifier, MTL-FECAM improves CIL performance The key takeaway of this study includes - MTL-FECAM effectively quantifies and mitigates the stability-plasticity tradeoff, improving long-term model performance in bioinformatics CIL. Feature covariance modeling enables a more informative and adaptive representation. The Mahalanobis metric helps learn optimal non-linear boundaries for new molecular classes. MTL-FECAM does not rely on exemplars, making it suitable for privacy-based biomedical applications. Empirical results and continuous retraining highlight the model's robustness and adaptability over time, hence confirming that MTL-FECAM outperforms existing CL methods. This work lays the foundation for future privacy-preserving, exemplar-free CL models in bioinformatics and cheminformatics applications.

Acknowledgement

The authors thank the Department of Science and Technology (DST-INSPIRE) for the Senior Research Fellowship for carrying out the research work.

References

1. Robins, Anthony. "Catastrophic forgetting, rehearsal and pseudorehearsal." Connection Science 7.2 (1995): 123-146.
2. Ma, Chunwe, et al. "Progressive voronoi diagram subdivision enables accurate data-free class-incremental learning." (2023).
3. Hinton, Geoffrey. "Distilling the Knowledge in a Neural Network." arXiv preprint arXiv:1503.02531 (2015).
4. Janson, Paul, et al. "A simple baseline that questions the use of pretrained-models in continual learning." arXiv preprint arXiv:2210.04428 (2022).
5. Zhang, Chi, et al. "Few-shot incremental learning with continually evolved classifiers." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
6. Mensink, Thomas, et al. "Distance-based image classification: Generalizing to new classes at near-zero cost." IEEE transactions on pattern analysis and machine intelligence 35.11 (2013): 2624-2637.
7. Tao, Xiaoyu, et al. "Few-shot class-incremental learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
8. Castro, Francisco M., et al. "End-to-end incremental learning." Proceedings of the European conference on computer vision (ECCV). 2018.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

9. Petit, Grégoire, et al. "Fetril: Feature translation for exemplar-free class-incremental learning." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023.
10. Zhu, Fei, et al. "Prototype augmentation and self-supervision for incremental learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
11. Zhu, Fei, et al. "Class-incremental learning via dual augmentation." Advances in Neural Information Processing Systems 34 (2021): 14306-14318.
12. Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." Proceedings of the national academy of sciences 114.13 (2017): 3521-3526.
13. Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.
14. Belouadah, Eden, and Adrian Popescu. "DeeSIL: Deep-Shallow Incremental Learning." Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018.
15. Liu, Yu, et al. "More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16. Springer International Publishing, 2020.
16. Yu, Lu, et al. "Semantic drift compensation for class-incremental learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
17. Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion
18. Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.
19. Douillard, Arthur, et al. "Podnet: Pooled outputs distillation for small-tasks incremental learning." Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16. Springer International Publishing, 2020.
20. Zhou, Da-Wei, Han-Jia Ye, and De-Chuan Zhan. "Co-transport for class-incremental learning." Proceedings of the 29th ACM International Conference on Multimedia. 2021.
21. Zhao, Bowen, et al. "Maintaining discrimination and fairness in class incremental learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
22. Wu, Yue, et al. "Large scale incremental learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
23. Wang, Fu-Yun, et al. "Foster: Feature boosting and compression for class-incremental learning." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.
24. Yan, Shipeng, Jiangwei Xie, and Xuming He. "Der: Dynamically expandable representation for class incremental learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
25. Zhou, Da-Wei, et al. "A model or 603 exemplars: Towards memory-efficient class-incremental learning." arXiv preprint arXiv:2205.13218 (2022).
26. Wu, Zhenqin, et al. "MoleculeNet: a benchmark for molecular machine learning." Chemical science 9.2 (2018): 513-530.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23

27. Edwards, Carl, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. "Translation between molecules and natural language." arXiv preprint arXiv:2204.11817 (2022).
28. Li, Juncai, and Xiaofei Jiang. "Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction." Wireless Communications and Mobile Computing 2021, no. 1 (2021): 7181815.
29. Sun, Fan-Yun, Jordan Hoffmann, Vikas Verma, and Jian Tang. "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization." arXiv preprint arXiv:1908.01000 (2019).
30. Hu, Ziniu, et al. "Gpt-gnn: Generative pre-training of graph neural networks." Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020.
31. Velickovic, Petar, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. "Deep graph infomax." ICLR (Poster) 2, no. 3 (2019): 4.
32. Zhang, Zaixi, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. "Motif-based graph self-supervised learning for molecular property prediction." Advances in Neural Information Processing Systems 34 (2021): 15870-15882.
33. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).
34. Kearnes, Steven, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. "Molecular graph convolutions: moving beyond fingerprints." Journal of computer-aided molecular design 30 (2016): 595-608.
35. Goswami, Dipam, et al. "Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning." Advances in Neural Information Processing Systems 36 (2024).
36. Shaheen, Hina, and Roderick Melnik. "Neural dynamics in parkinson's disease: Integrating machine learning and stochastic modelling with connectomic data." International Conference on Computational Science. Cham: Springer Nature Switzerland, 2024.
37. Abramova, Yulia, and Vasiliy Leonenko. "The Past Helps the Future: Coupling Differential Equations with Machine Learning Methods to Model Epidemic Outbreaks." International Conference on Computational Science. Cham: Springer Nature Switzerland, 2024.
38. Catalfamo, Alessio, et al. "Machine Learning Workflows in the Computing Continuum for Environmental Monitoring." International Conference on Computational Science. Cham: Springer Nature Switzerland, 2024.
39. Tanade, Cyrus, and Amanda Randles. "HarVI: Real-time intervention planning for coronary artery disease using machine learning." International Conference on Computational Science. Cham: Springer Nature Switzerland, 2024.
40. Leifsson, Leifur. "Cost-Efficient Multi-Objective Design of Miniaturized Microwave Circuits Using Machine Learning and Artificial Neural Networks."

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_23