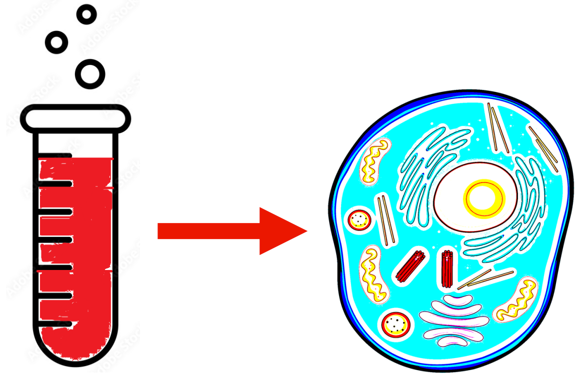
Cross Modality Representation Learning of Cell Painting and Transcriptomics data

Son Ha
PhD Student
Beerse, Belgium

Johnson&Johnson
Innovative Medicine

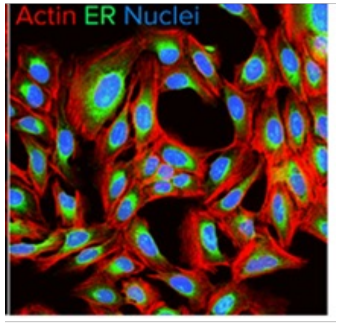
Introduction

Cell Painting (CP) and Transcriptomics (TX) data



Small Molecule

Cell



Cell Painting

→ Phenotypic Profile

Diam Cell	...	Area Nucleus
+0.54	...	-0.98

RNA-Seq Transcriptomics

→ Gene Expression Profile

Gene 1	...	Gene N
+0.41	...	-0.68

Cross Modality Representation Learning Motivation

- Real world problem: TX data is costly to generate

  → New compounds will only have CP data, TX missing

- Can we learn better single modality representations given unlabeled data from multiple modalities?
- Cross modality representation learning (Ngiam 2011):

Feature Learning	Downstream Tasks
CP + TX	CP

- Other multimodal representation learning benefits:
  - Integration of different data types for downstream tasks
  - Improve modelling capability of Modes of Actions/Bioassays.

Ngiam. Multimodal Deep Learning. 2011

Summary of this work

- Benchmark two cross modality representation learning methods for CP and TX data:
  - Contrastive Learning
  - Bimodal autoencoder

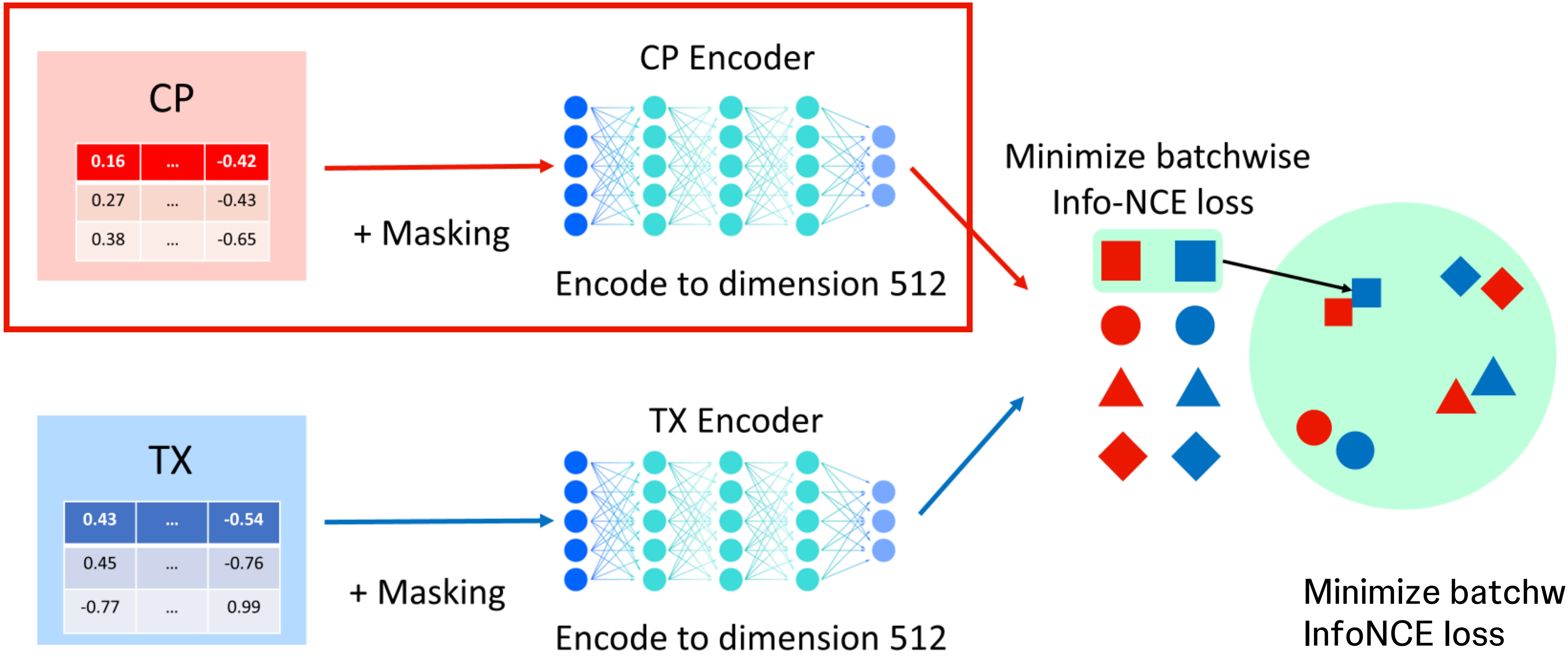
- Evaluate them on a variety of downstream tasks

Slide legal disclaimer goes here if required

Methods

Johnson&Johnson
Innovative Medicine

Contrastive Learning Pretraining



CP

0.16	...	-0.42
0.27	...	-0.43
0.38	...	-0.65

TX

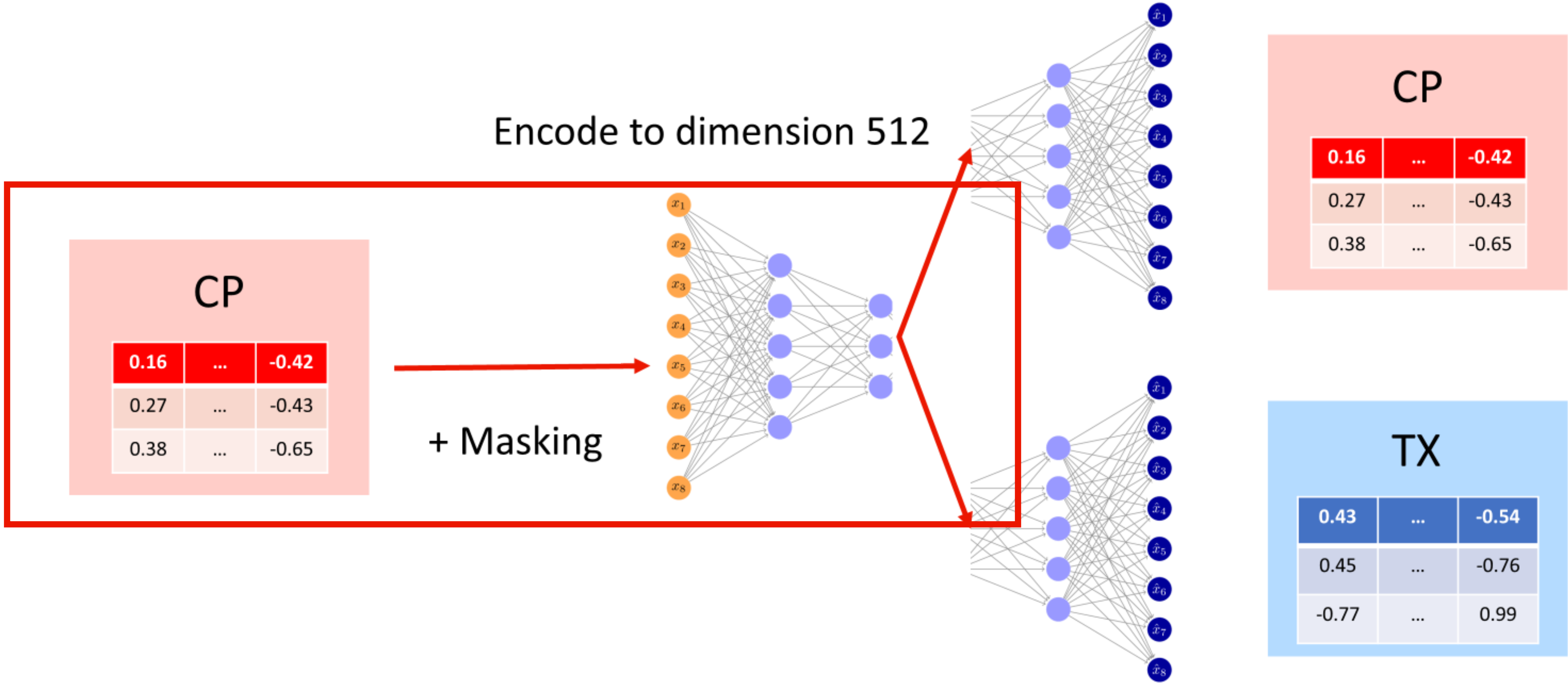
0.43	...	-0.54
0.45	...	-0.76
-0.77	...	0.99

Minimize batchwise InfoNCE loss

$$L_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \ln \frac{\exp\left(\text{sim}\left(\mathbf{x}_i, \mathbf{z}_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)} - \frac{1}{N}\sum_{i=1}^{N} \ln \frac{\exp\left(\text{sim}\left(\mathbf{x}_i, \mathbf{z}_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(\mathbf{z}_j, \mathbf{z}_i\right)/\tau\right)},$$

Alec Radford, et al.. (2021). Learning Transferable Visual Models From Natural Language Supervision.

Bimodal Autoencoder Pretraining



Encode to dimension 512

CP

0.16	...	-0.42
0.27	...	-0.43
0.38	...	-0.65

+ Masking

CP

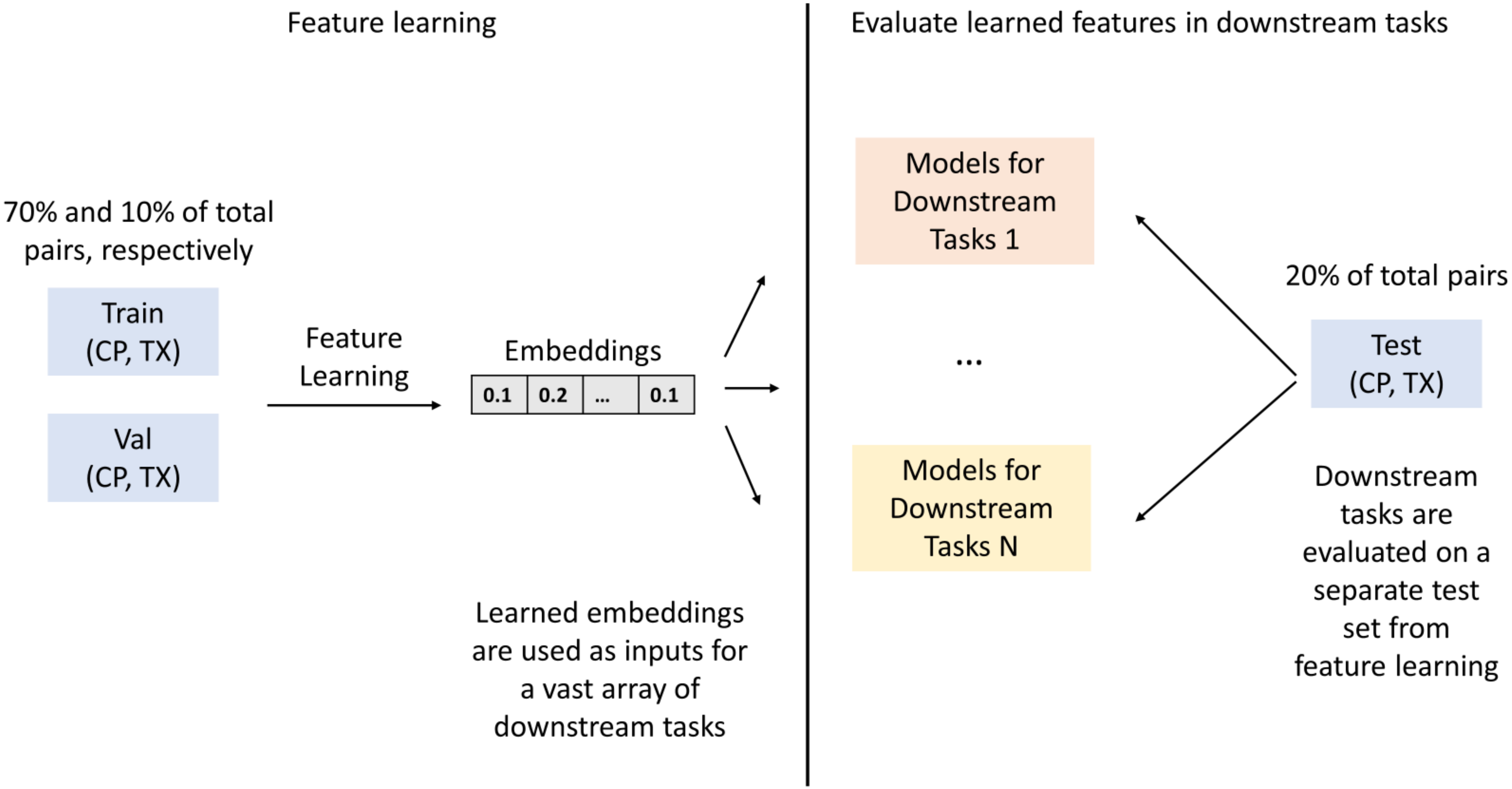
0.16	...	-0.42
0.27	...	-0.43
0.38	...	-0.65

TX

0.43	...	-0.54
0.45	...	-0.76
-0.77	...	0.99

Minimize average of the 2
MSE reconstruction losses

Training and Evaluating Learned Embeddings



Feature learning

Evaluate learned features in downstream tasks

70% and 10% of total pairs, respectively

Train (CP, TX)

Val (CP, TX)

Feature Learning

Embeddings

0.1	0.2	...	0.1

Learned embeddings are used as inputs for a vast array of downstream tasks

Models for Downstream Tasks 1

...

Models for Downstream Tasks N

20% of total pairs

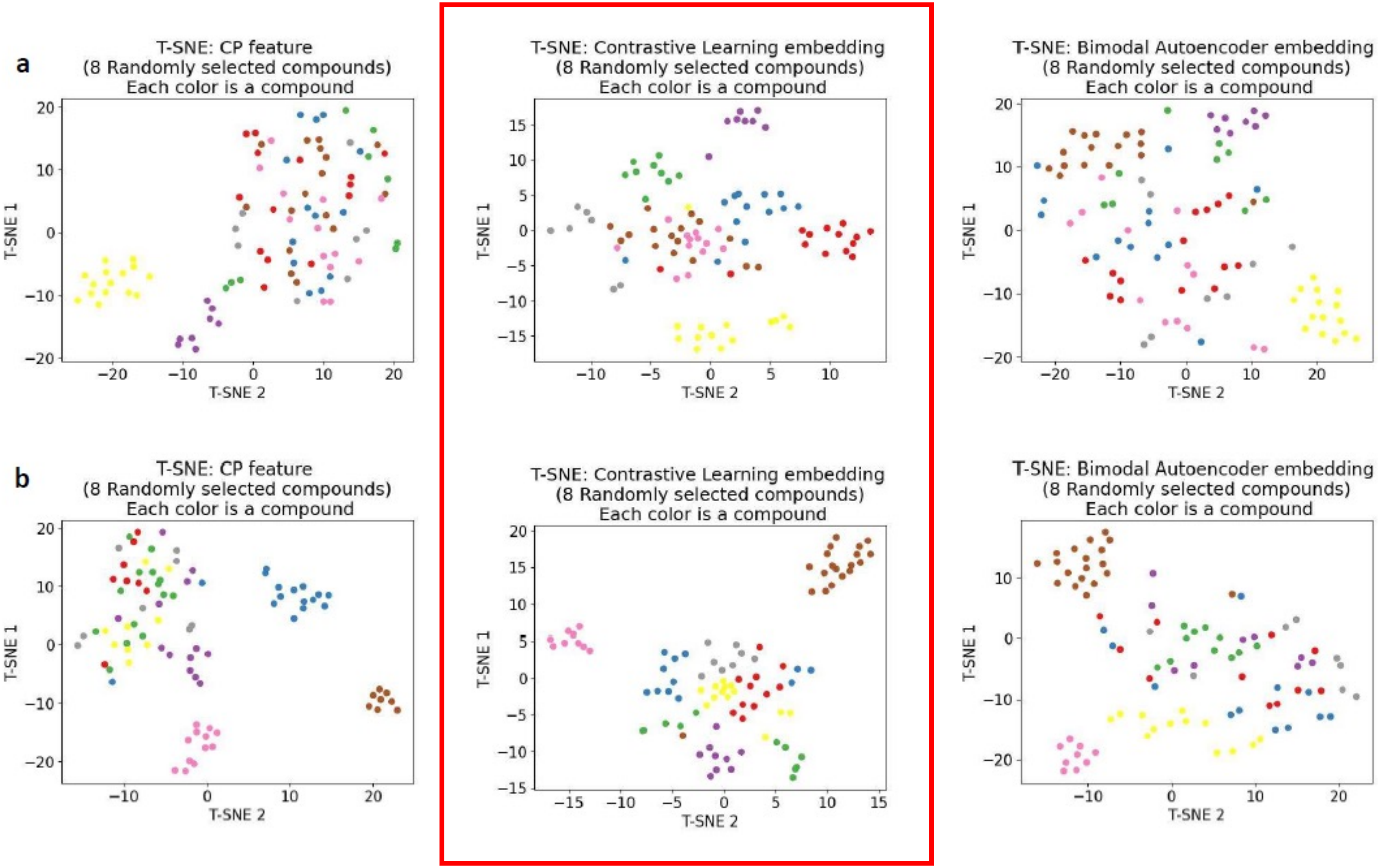
Test (CP, TX)

Downstream tasks are evaluated on a separate test set from feature learning

Result

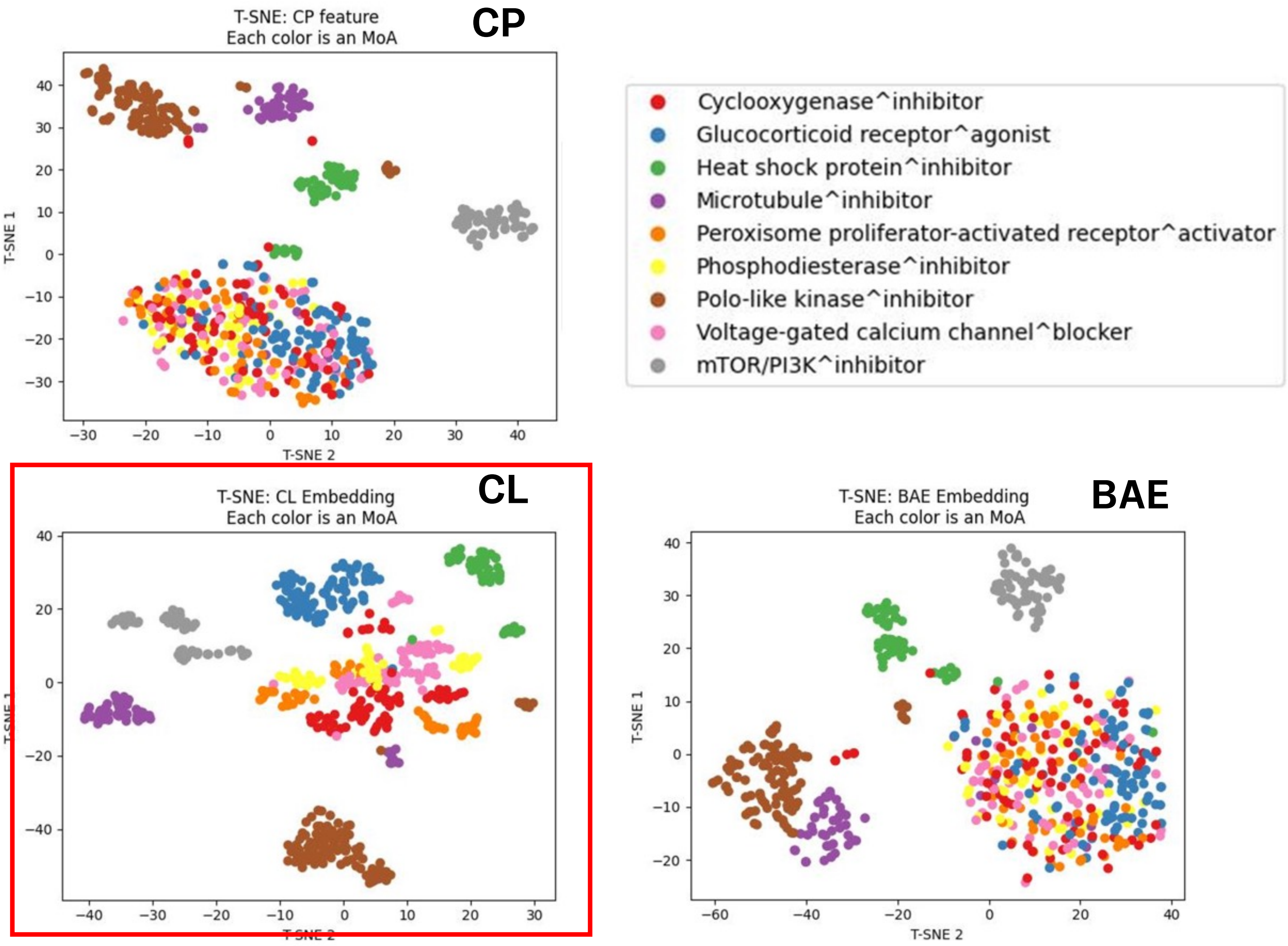
Johnson&Johnson
Innovative Medicine

Unsupervised Task: CP replicates Clustering



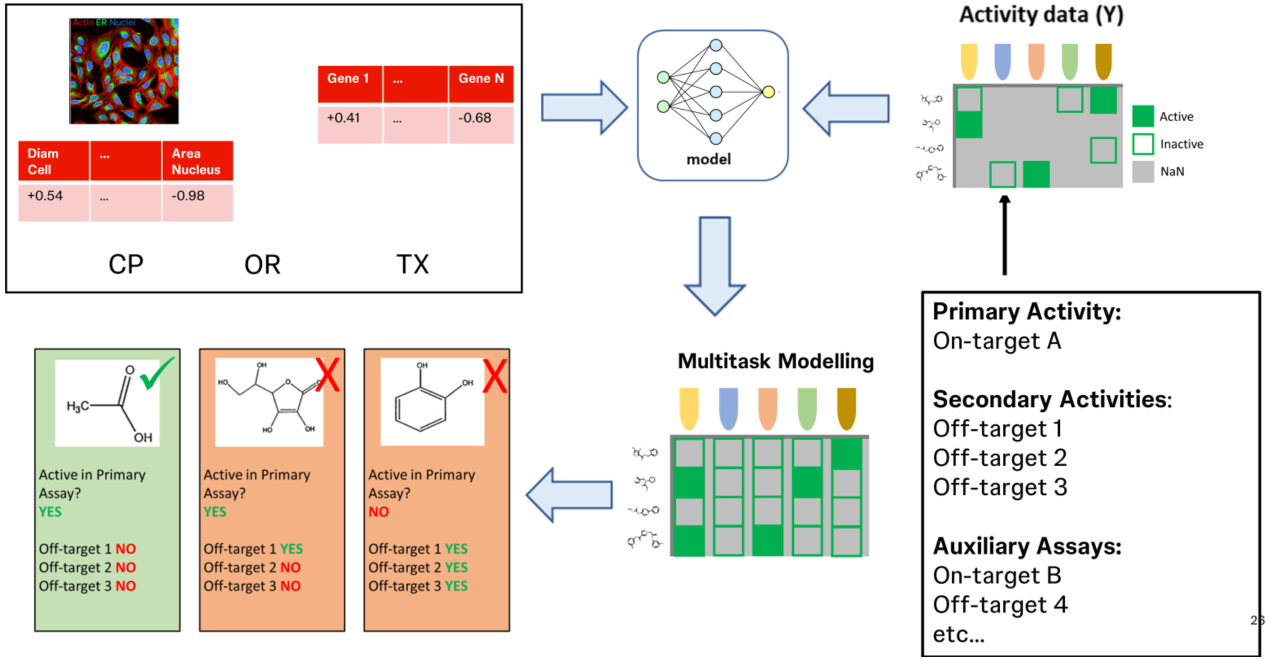
Feature Type	kNN Accuracy CP Replicates
CP	0.416
CL Embedding	**0.805**
BAE Embedding	0.428

Unsupervised Task: Mode of Action Clustering



Feature Type	kNN Accuracy MoA
CP	0.784
CL Embedding	**0.952**
BAE Embedding	0.784

Supervised Task – Multitask Bioactivity Classification



Primary Activity:
On-target A

Secondary Activities:
Off-target 1
Off-target 2
Off-target 3

Auxiliary Assays:
On-target B
Off-target 4
etc...

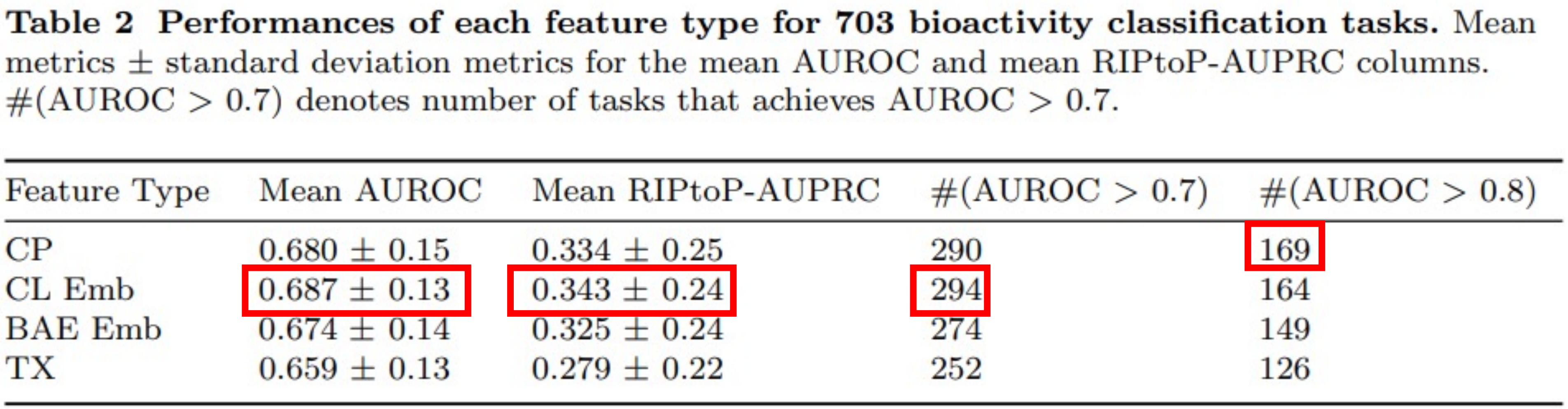
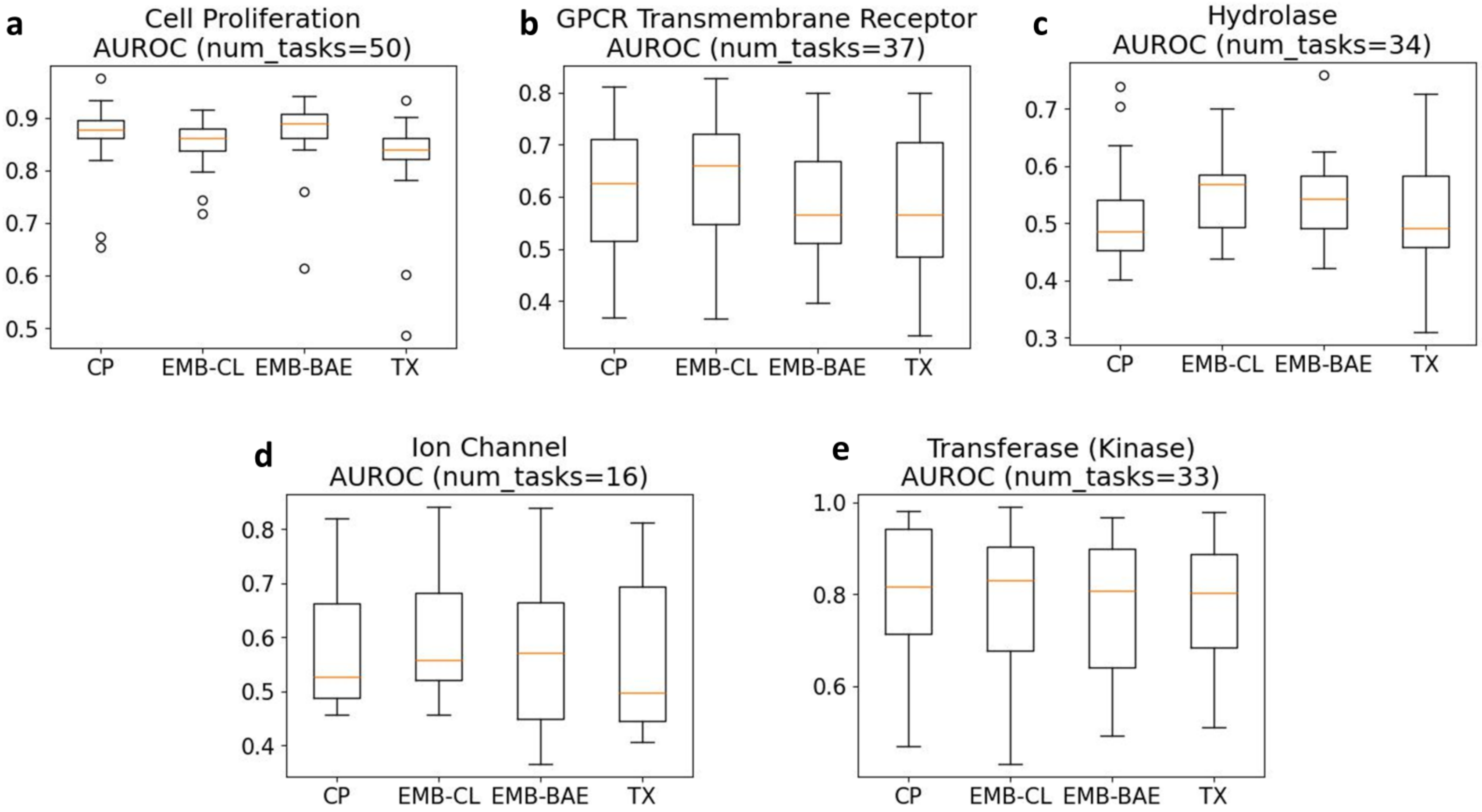


Table 2 Performances of each feature type for 703 bioactivity classification tasks. Mean metrics ± standard deviation metrics for the mean AUROC and mean RIPtoP-AUPRC columns. #(AUROC > 0.7) denotes number of tasks that achieves AUROC > 0.7.

Feature Type	Mean AUROC	Mean RIPtoP-AUPRC	#(AUROC > 0.7)	#(AUROC > 0.8)
CP	0.680 ± 0.15	0.334 ± 0.25	290	169
CL Emb	0.687 ± 0.13	0.343 ± 0.24	294	164
BAE Emb	0.674 ± 0.14	0.325 ± 0.24	274	149
TX	0.659 ± 0.13	0.279 ± 0.22	252	126

Bioactivity Classification Grouped by Protein Family

- CL embedding outperforms CP feature in GPCR, Hydrolase and Ion Channel tasks.
- BAE embedding, surprisingly, outperforms CP feature and CL embedding in Cell Proliferation.



Learned Embedding improves upon underperforming CP tasks that TX does well

- Motivation:
  - TX costly to generate → new compounds will only have CP but not TX
    → Lose out on ‘good TX models’
  - Can embedding improves underperforming CP models that TX does well?
  - Yes, we achieve improvement with statistical significance .

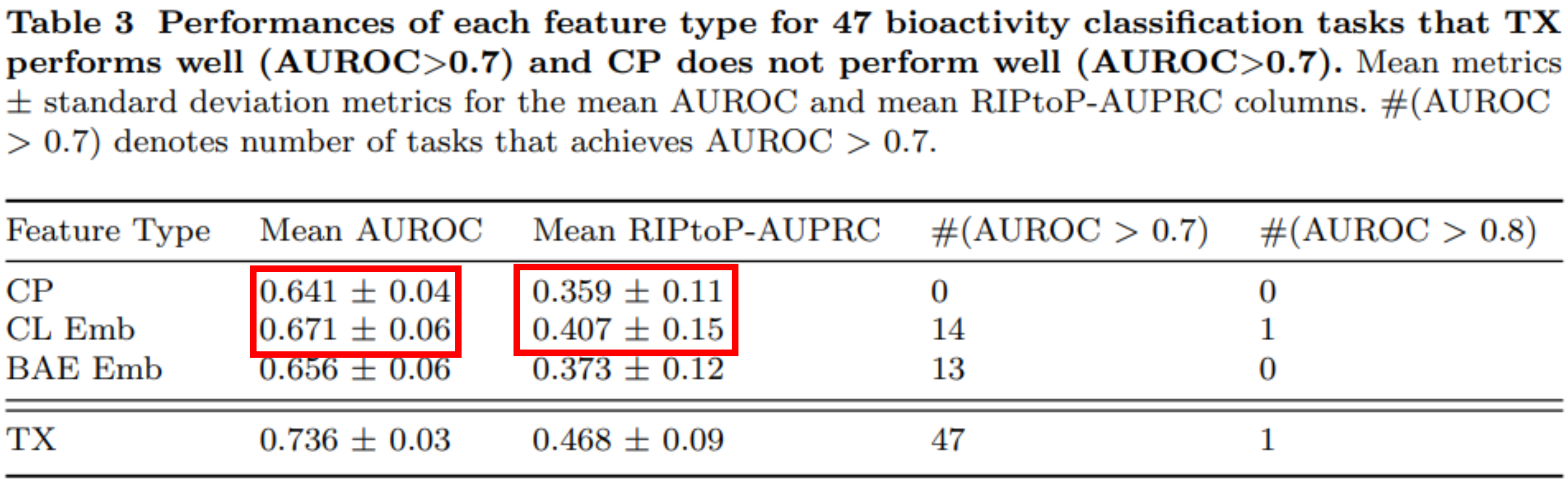


Table 3 Performances of each feature type for 47 bioactivity classification tasks that TX performs well (AUROC>0.7) and CP does not perform well (AUROC>0.7). Mean metrics ± standard deviation metrics for the mean AUROC and mean RIPtoP-AUPRC columns. #(AUROC > 0.7) denotes number of tasks that achieves AUROC > 0.7.

Feature Type	Mean AUROC	Mean RIPtoP-AUPRC	#(AUROC > 0.7)	#(AUROC > 0.8)
CP	0.641 ± 0.04	0.359 ± 0.11	0	0
CL Emb	0.671 ± 0.06	0.407 ± 0.15	14	1
BAE Emb	0.656 ± 0.06	0.373 ± 0.12	13	0
TX	0.736 ± 0.03	0.468 ± 0.09	47	1

Tasks criteria: (TX tasks >0.7 AUROC, CP tasks <0.7 AUROC, at least 20 positives and 20 negatives)

Discussion

Johnson&Johnson
Innovative Medicine

Discussion

- Supervised learning (bioactivity classification):
  - CL embedding achieves higher mean AUROC and RIPtoP-AUPRC over CP feature.
  - CL embedding outperforms CL feature in GPCR, Hydrolase and Ion Channel tasks, while BAE outperforms CL feature in Cell Proliferation tasks.
  - For tasks that TX performs well and CP performs badly, embeddings from CP improve performance over CP features.
- Unsupervised clustering:
  - CL embedding achieves highest kNN Accuracy, while BAE embedding achieves minimal improvement.
  - Visual inspection agrees with the above results.