Meta-learning for Bridging Labeled and Unlabeled Data in Biomedicine

Maria Brbić, Chelsea Finn and Jure Leskovec
Stanford University
Open-World Learning
Where Are We Now?

- Optimization-based methods
- Non-parametric methods
- Prior knowledge enhanced methods
- Task-level
- Feature-level
- Novel class discovery
- Open-world semi-supervised learning

Missing data problems

Few-Shot Learning

Open-World Learning
What Will We Cover?

- How to discover novel, never-before-seen classes
- How to simultaneously discover novel classes and recognize previously seen classes

Learning goals:

- Open-world learning setting
- Basics of open-world learning techniques (& how to implement)
- Applications to cell type annotation task
Traditional machine learning assumes test data contains only classes encountered during training.

**Closed-World Setting**

**Training data**
- Classes: elephant, cheetah, zebra, rhino

**Test data**
- Classes: elephant, cheetah, zebra, rhino
Novel, never-before seen classes can appear during the model deployment.
Example: Disease Classification

- **Task:** Associate patients to diseases
- **Challenge:** Patients with rare diseases that had no labels in the train data can appear
Novel Class Discovery

Based on:
Problem Setting

Goal:
- Label every sample with its class in unannotated dataset
- Many classes are unique and can not be found in previously annotated datasets
Example: Single-Cell

Annotated experiments

Classes??

Unannotated experiment

Goal:
- Label every cell with its cell type in unannotated experiment
- Many cell types are unique and can not be found in any other experiment

Brbic et al., Nat Methods '20

Brbic et al., Nat Methods '20

ISMB 2021 Meta-Learning for Bridging Labeled and Unlabeled Data in Biomedicine
MARS: Reuse Data

Key idea: discover novel classes by using old, already annotated data
MARS Idea: Learn Embedding

Learn embedding function $f_\theta$ such that:
- Samples from same classes are close
- Samples from different classes are embedded far

**Note:** MARS considers single-cell experiments and cell type annotation task

Brbic et al., *Nat Methods* ‘20
Part 1: Samples Are Close to Their Ground-Truth Landmarks

For annotated datasets:

1. Minimize distance to class landmarks
Part 1: Samples Are Close to Their Ground-Truth Landmarks

For annotated datasets:

1. Minimize distance to class landmarks

Intuition: Samples with same class annotation have similar embedding vector
Part 1: Samples Are Close to Their Chosen Landmarks

For unannotated dataset:

1. Minimize distance to the closest landmark
Part 1: Samples Are Close to Their Chosen Landmarks

For unannotated dataset:

1. Minimize distance to the closest landmark

Both landmarks and embeddings are learned
Part 2: Put Landmarks Far Away

Within each dataset:

2. Maximize distance between landmarks
Part 2: Put Landmarks Far Away

Within each dataset:

2. Maximize distance between landmarks

Intuition: Landmarks of different classes have different embedding vectors
Annotated datasets: $\mathcal{D}_i = (X^{(i)}, y^{(i)})$ \hspace{1cm} $\mathcal{D}_{meta} = \bigcup_{i=1}^{T} \mathcal{D}_i$

$X^{(i)} = \{x_j^{(i)} \in \mathbb{R}^M \}_{j=1}^{N_i}$, $y^{(i)} = \{y_j \in \{1, \ldots, K_i\} \}_{j=1}^{N_i}$

class annotations

Unannotated dataset: $\mathcal{U} = X^{(u)}$

$X^{(u)} = \{x_j^{(u)} \in \mathbb{R}^G \}_{j=1}^{N}$

Goal: predict $y^{(u)}$

Embedding function: $f_\theta: \mathbb{R}^M \rightarrow \mathbb{R}^Z$

Class landmarks: $\left\{ \{p_k^{(i)} \in \mathbb{R}^Z \}_{j=1}^{K_i} \}_{i=1}^{M} \right\}$

Landmarks for annotated datasets

Landmarks for unannotated dataset
Math: Objective Function

- For each annotated dataset $\mathcal{D}_i$:

$$f_\theta: \mathbb{R}^M \to \mathbb{R}^Z$$

**Force samples close to their ground-truth landmarks**

$$\mathcal{L}_i = \frac{1}{N_i} \sum_{k=1}^{K_i} \sum_{j=1}^{N_i} \mathbb{I}_{\{y_j^{(i)} = k\}} \cdot d \left( f_\theta \left( x_j^{(i)} \right), p_k^{(i)} \right) - \lambda \frac{1}{K_i(K_i - 1)} \sum_{k_1=1}^{K_i} \sum_{k_2=1}^{K_i} \cdot d \left( p_{k_1}^{(i)}, p_{k_2}^{(i)} \right)$$

- number of examples in dataset $i$
- landmark of class $k$ in dataset $i$
- number of landmarks in dataset $i$

**Force landmarks far away**
Math: Objective Function

For unannotated dataset $\mathcal{U}$:

$$f_\theta : \mathbb{R}^M \rightarrow \mathbb{R}^Z$$

**Force samples close to their closest landmark**

$$\mathcal{L}_\mathcal{U} = \frac{1}{N} \sum_{k=1}^{K} \sum_{j=1}^{N} \min_{k=1,\ldots,K} d(f_\theta(x_j), p_k) - \frac{\lambda}{K(K-1)} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} d(p_{k_1}, p_{k_2})$$

- $N$: number of examples in unannotated dataset
- $K$: number of landmarks in unannotated dataset
- $p_k$: landmark of class $k$

**Force landmarks far away**

Brbic et al., *Nat Methods* ‘20

ISMB 2021 Meta-Learning for Bridging Labeled and Unlabeled Data in Biomedicine
Math: Objective Function

- **MARS objective function:**

\[ \mathcal{L}_{MARS} = \min_{\theta, \{p_k^{(i)}\}_i, \{p_k\}_k} \sum_{i=1}^{T} \mathcal{L}_i + \tau \mathcal{L}_U \]

- Optimize over landmarks and embedding parameters \( \theta \)
- Annotated datasets
- Unannotated dataset

Brbic et al., *Nat Methods* ‘20

ISMB 2021 Meta-Learning for Bridging Labeled and Unlabeled Data in Biomedicine
Learning Function $f_\theta$

- Fully connected neural network

$$h_i^{(k)} = \phi(W^{(k-1)} h_i^{(k-1)})$$

\(\phi\) : non-linearity

Again, you can replace backbone depending on your input type

Brbic et al., *Nat Methods* ‘20
Pretraining

**Approach:** Pretrain network before fine-tuning on the designed objective function

Autoencoder pretraining:

Pretraining strategy depends on the application

$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \hat{X}_i)^2$

Brbic et al., *Nat Methods* ‘20
After Pretraining

- Remove the decoder part and use encoder weights to initialize neural network
- Initialize landmarks with K-means clustering in the embedding space
- Fine-tune network with the designed objective function

Brbic et al., *Nat Methods* ‘20
Inference

- Embed data points from the unannotated dataset in the learned shared embedding space.
- Assign data points to the class of the closest landmark from the unannotated dataset.
Application: Cell type discovery in single-cell genomics

Based on:
How to discover novel cell types across tissues, species and sequencing technologies without any annotations?
Experimental Setup

Datasets:

- **Tabula Muris** and **Tabula Muris Senis**
- >100,000 cells across 23 tissues
- Each cell described by its gene expression profile (23k genes)
Experimental Setup

Can we generalize to unseen cell types across tissues?

Training and prediction: Embed into 100 dim.

Given source tissues with annotations and a target unannotated tissue:

1) Train the model on source tissues
   - Use Tabula Muris cell annotations for source tissues
   - No annotations of the target tissue are used!
2) Predict cell annotations for the target tissue
Baseline Methods

Louvain:
- Methodology used to annotate Tabula Muris [Nature’18]:
  1. Perform PCA
  2. Construct k-NN graph in low-dimensional PC space
  3. Run Louvain community detection method

SIMLR\(^1\):
- Unsupervised similarity-learning approach for clustering and dimensionality reduction
- Learns cell-to-cell similarity metric by multiple kernel-learning

ScVi\(^2\):
- Hierarchical Bayesian model with conditional distributions learned by deep neural networks
- We pretrain deep neural network with the same data used for pretraining our approach

\(^1\)Wang et al. Nat Methods ’17, \(^2\)Lopez et al., Nat Methods ‘20
Results: Performance

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Leiden</th>
<th>Louvain</th>
<th>ScVi²</th>
<th>SIMLR¹</th>
<th>MARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aorta</td>
<td>0.377</td>
<td>0.386</td>
<td>0.490</td>
<td>0.511</td>
<td>0.544</td>
</tr>
<tr>
<td>BAT</td>
<td>0.705</td>
<td>0.694</td>
<td>0.783</td>
<td>0.192</td>
<td>0.854</td>
</tr>
<tr>
<td>Bladder</td>
<td>0.168</td>
<td>0.262</td>
<td>0.997</td>
<td>0.992</td>
<td>0.997</td>
</tr>
<tr>
<td>Brain_Non-Myeloid</td>
<td>0.581</td>
<td>0.563</td>
<td>0.557</td>
<td>0.731</td>
<td>0.754</td>
</tr>
<tr>
<td>Diaphragm</td>
<td>0.500</td>
<td>0.605</td>
<td>0.674</td>
<td>0.897</td>
<td>0.916</td>
</tr>
<tr>
<td>GAT</td>
<td>0.247</td>
<td>0.375</td>
<td>0.438</td>
<td>0.587</td>
<td>0.872</td>
</tr>
<tr>
<td>Heart</td>
<td>0.287</td>
<td>0.427</td>
<td>0.488</td>
<td>0.724</td>
<td>0.910</td>
</tr>
<tr>
<td>Kidney</td>
<td>0.667</td>
<td>0.702</td>
<td>0.639</td>
<td>0.658</td>
<td>0.776</td>
</tr>
<tr>
<td>Large_Intestine</td>
<td>0.332</td>
<td>0.391</td>
<td>0.490</td>
<td>0.389</td>
<td>0.473</td>
</tr>
<tr>
<td>Limb_Muscle</td>
<td>0.404</td>
<td>0.549</td>
<td>0.674</td>
<td>0.788</td>
<td>0.958</td>
</tr>
<tr>
<td>Liver</td>
<td>0.357</td>
<td>0.375</td>
<td>0.378</td>
<td>0.486</td>
<td>0.828</td>
</tr>
<tr>
<td>Lung</td>
<td>0.509</td>
<td>0.507</td>
<td>0.382</td>
<td>0.689</td>
<td>0.585</td>
</tr>
<tr>
<td>MAT</td>
<td>0.366</td>
<td>0.441</td>
<td>0.483</td>
<td>0.734</td>
<td>0.940</td>
</tr>
<tr>
<td>Mammary_Gland</td>
<td>0.300</td>
<td>0.341</td>
<td>0.926</td>
<td>0.560</td>
<td>0.893</td>
</tr>
<tr>
<td>Pancreas</td>
<td>0.399</td>
<td>0.437</td>
<td>0.495</td>
<td>0.588</td>
<td>0.723</td>
</tr>
<tr>
<td>SCAT</td>
<td>0.353</td>
<td>0.477</td>
<td>0.590</td>
<td>0.668</td>
<td>0.931</td>
</tr>
<tr>
<td>Skin</td>
<td>0.178</td>
<td>0.231</td>
<td>0.366</td>
<td>0.413</td>
<td>0.530</td>
</tr>
<tr>
<td>Spleen</td>
<td>0.160</td>
<td>0.206</td>
<td>0.213</td>
<td>0.209</td>
<td>0.875</td>
</tr>
<tr>
<td>Thymus</td>
<td>0.301</td>
<td>0.268</td>
<td>0.378</td>
<td>0.360</td>
<td>0.565</td>
</tr>
<tr>
<td>Tongue</td>
<td>0.119</td>
<td>0.131</td>
<td>0.463</td>
<td>0.265</td>
<td>0.588</td>
</tr>
<tr>
<td>Trachea</td>
<td>0.378</td>
<td>0.406</td>
<td>0.502</td>
<td>0.689</td>
<td>0.861</td>
</tr>
<tr>
<td>Median</td>
<td>0.357</td>
<td>0.406</td>
<td>0.490</td>
<td>0.588</td>
<td>0.854</td>
</tr>
</tbody>
</table>

Performance is measured as adjusted Rand index.

Brbic et al., Nat Methods ‘20
Positive Knowledge Transfer

Higher score indicates better performance

17% improvement

Takeaway:
- Incorporating more tasks improves performance

Brbic et al., *Nat Methods* ‘20
Example Embeddings: Diaphragm and Liver

Takeaway:
- Cell types naturally form clusters

Colors indicate ground-truth annotations

Brbic et al., Nat Methods ‘20
Example Embeddings: Diaphragm and Liver

Can we go beyond cell-type discovery task and provide human understandable names to discovered cell types?

Brbic et al., Nat Methods ‘20
Key idea: Use landmarks from annotated experiments to generate interpretable names for discovered cell types

Approach:
- Probabilistically define cell type based on distances from landmarks of annotated tasks.

How to estimate probabilities?

Brbic et al., Nat Methods ‘20
Naming Cell Types

- Estimate probabilities in proportion to the probability density under Gaussian distribution centered at the mean of the discovered cluster:

\[
p_{k|j} = \frac{\exp \left(-\|\mathbf{p}_k - \mu_j\|^2 / 2\sigma_j^2 \right)}{\sum_{i=1}^{M} \sum_{k'}^{K_i} \exp \left(-\|\mathbf{p}_k^{(i)} - \mu_j\|^2 / 2\sigma_j^2 \right)}
\]

\(\mu_j\): centroid of cluster \(j\)
\(\sigma_j\): standard deviation estimated based on pairwise distances of cells assigned to cluster \(j\)

Probability that \(j\)th cluster has the same name as \(k\)th landmark

Brbic et al., *Nat Methods* ‘20
Naming Cell Types: Example

Example: Limb muscle used as unannotated experiment
Naming Cell Types: Example

Example: Limb muscle used as unannotated experiment

MARS:
46% B-cell
18% T-cell
36% Other

Brbic et al., Nat Methods ‘20

ISMB 2021 Meta-Learning for Bridging Labeled and Unlabeled Data in Biomedicine
Naming Cell Types: Example

Example: Limb muscle used as unannotated experiment

MARS:
88% Macrophage
4% Myeloid cell
8% Other

Brbic et al., Nat Methods ‘20
Naming Cell Types: Example

Example: Limb muscle used as unannotated experiment

What if we want to recognize previously seen and discover novel cell types simultaneously?
Open-World Semi-Supervised Learning

Based on:
Open-World Semi-Supervised Learning

- **So far:** Discovering novel classes

  Pretrain, then learn to cluster by reusing previously annotated tasks.

- **Challenge:**

  What if we want to simultaneously discover new classes and assign examples to previously seen classes?
Supervised Learning

Setting:

Task: Assign examples to previously seen classes

Training data

Test data

“elephant”

“octopus”

“cheetah”

Seen classes
Novel Class Discovery

Setting:

Training data

Test data

Novel classes

Task: Discover novel classes
Task: Simultaneously discover novel and recognize previously seen classes
Open-World Semi-Supervised Learning

- Why is it hard?
  - Novel classes data can negatively affect the classification performance on seen classes
  - Novel class discovery task is hard by the nature of the problem

- Why is it important?
  - Very realistic setting: we can not expect to identify and label all classes in advance

Cao*, Brbic*, Leskovec. arXiv ‘21
Example: Disease Classification

- **Task:** Train a classifier that associates patients to previously seen diseases, but also discovers unseen disease states.

![Train data and Test data diagram](image)
**Task:** Train a classifier that associates patients to previously seen diseases, but also discovers unseen disease states.

---

**Train data**

**Test data**

unseen disease
Example: Disease Classification

**Task:** Train a classifier that associates patients to previously seen diseases, but also discovers unseen disease states.
**Example: Disease Classification**

**Task:** Train a classifier that associates patients to previously seen diseases, but also discovers unseen disease states.
Open-World SSL Objective

- **ORCA Approach**: Introduce additional classification heads for each newly discovered class

\[ \mathcal{L} = \mathcal{L}_{CE} + \eta_1 \mathcal{L}_{BCE} + \eta_2 \mathcal{L}_R \]

\[ p(\text{"octopus"}) \]

\[ p(\text{"class 1"}) \]

\[ \mathcal{L}_{CE} \quad \text{supervised objective} \]

\[ \mathcal{L}_{BCE} \quad \text{unsupervised objective} \]

\[ \mathcal{L}_R \quad \text{regularization towards uniform distribution} \]

Cao*, Brbic*, Leskovec. *arXiv* '21
Unsupervised Objective $\mathcal{L}_{BCE}$

- **Unsupervised objective** $\mathcal{L}_{BCE}$: Predict whether two examples belong to the same class

  - For labeled data, we can use existing labels.
  - How to get labels for unlabeled data?

Cao*, Brbic*, Leskovec. arXiv ‘21
Unsupervised Objective $\mathcal{L}_{BCE}$

Key idea: Generate pseudo-labels

Find closest neighbors in the embedding space and regard them as true pairs.

$$\mathcal{L}_{BCE} = \frac{1}{N} \sum_{z_i, z'_i \in \mathcal{E}(Z, Z')} -\log(\sigma(W^T z_i), \sigma(W^T z'_i))$$

Cao*, Brbic*, Leskovec. arXiv '21
**Supervised Objective $\mathcal{L}_{CE}$**

- **Challenge:**
  Using standard cross-entropy loss on seen classes results in learning seen classes faster than novel classes (gradient is updated for seen but not for novel classes)

$$\mathcal{L}_{CE} = \frac{1}{N_l} \sum_{z_i \in Z_l} - \log \frac{\exp(W_{y_i}^T z_i)}{\sum_j \exp(W_{y_j}^T z_i)}$$

How to mitigate the bias towards seen classes?

Cao*, Brbic*, Leskovec. *arXiv '21*
Supervised Objective

Key idea: Uncertainty based adaptive margin to avoid learning seen classes too fast

\[ \mathcal{L}_\text{CE} = \frac{1}{m} \sum_{z_i \in Z_l} - \log \frac{\exp(W_{y_i}^T z_i + \lambda \hat{u})}{\exp(W_{y_i}^T z_i + \lambda \hat{u}) + \sum_{j \neq i} \exp(W_{y_j}^T z_i)} \]

\( \hat{u} \): estimated uncertainty  
\( \lambda \): regularizer defining strength of uncertainty

How to estimate uncertainty \( \hat{u} \)?

From the confidence of unlabeled samples computed from the output of the classifier.

Cao*, Brbic*, Leskovec. *arXiv* '21
Supervised Objective

- **Intuition:**
  - **Beginning of the training:** Larger uncertainty $\rightarrow$ larger margin $\rightarrow$ larger intra class variance of seen classes
  - **Close to the end:** Smaller uncertainty $\rightarrow$ margin nearly zero (standard CE) $\rightarrow$ small intra class variance
Application: Cell type annotation of CODEX data

Based on:
- Ongoing work: Brbic*, Cao*, Hickey*, Tan, Snyder, Nolan, Leskovec
CODEX Technology

- Multiplexed imaging data technology
- Provides full spatial context
- High throughput single-cell information
- ~50 protein targets within a single tissue section

How to accurately identify cell types?
How to generalize across tissues and patients?
How to accurately identify cell types?
How to generalize across tissues and patients?

New challenges compared to scRNA-seq data:

- Additional sources of noise from imperfect segmentation of cells, imaging artifacts, and issues with image processing
- Only ~50 proteins per cell
- Includes spatial context
Experimental Setup

- **Dataset:**
  - Human small intestine and colon generated within Human BioMolecular Atlas Program (HuBMAP)

- **Validation:**
  - Quantitative validation (using the ground truth):
    - 4 regions of small intestine with ground truth annotations
  - Expert evaluation: across donors

- **Setting:**
  - Leave-one-region-out
  - Leave-one-donor-out
Approach

- **ORCA** for assigning cells to previously labeled cell types and discovering new cell types

- **Backbone**: 2-layer fully connected neural network
Results: Accuracy

- **Leave-one-region-out**
  - ORCA outperforms other methods and agrees with human annotators

- **Leave-one-donor-out**
  - Identification is accurate across cell types by expert validation

Validation based on the marker genes
Open-World Learning: Summary

- We covered:
  - How to generalize to unseen classes and transfer knowledge across different biological contexts
  - Novel-class discovery methods: MARS
  - Open-world semi-supervised learning methods: ORCA
  - Applications: single-cell annotation, CODEX annotation
If You Want To Know More

Papers:

Tools:

Website [data, code & more]:
- [MARS] [http://snap.stanford.edu/mars/](http://snap.stanford.edu/mars/)
- [ORCA] [http://snap.stanford.edu/orca/](http://snap.stanford.edu/orca/)