CSLAN: Cross-Species Latent Alignment Network for Trauma-Related Cell-Type Classification

Anonymous submission

Abstract

Transfer learning across domains with mismatched and non-mappable feature spaces is a fundamental challenge in machine learning. Existing methods often rely on brittle feature-mapping or risk catastrophic forgetting during fine-tuning. To address this, we introduce the Cross-Species Latent Alignment Network (CSLAN), a novel framework for robust knowledge transfer. CSLAN pioneers a three-pronged approach: (1) we employ a sparse regression model for principled selection of informative features in each domain, reducing noise and dimensionality. (2) We pre-train an encoder-decoder on a comprehensive source domain (mouse). (3) We introduce a biologically-inspired asymmetric fine-tuning strategy, where the pre-trained decoder and latent processor—encapsulating conserved class definitions—are frozen. A new target-specific encoder (for human) is then trained from scratch to project its distinct feature space into this preserved, semantically structured latent space. On a challenging cross-species trauma-related single-cell classification task, CSLAN achieved 95.83% accuracy using only a few hundred labeled human cells, significantly outperforming standard baselines. Our work establishes a powerful paradigm for aligning mismatched domains, demonstrating that decoupling feature projection from a conserved decision manifold is key to effective transfer.

1 Introduction

The ability to transfer learned knowledge across different domains is a cornerstone of modern artificial intelligence, enabling models to generalize from data-rich source domains to data-scarce target domains. While transfer learning has seen tremendous success (Pan and Yang 2009; Zhuang et al. 2020; Zhu et al. 2023), a significant and under-explored challenge arises when the source and target domains do not share an identical feature space (Li et al. 2023). In this scenario, where the source and target domains have different dimensions and semantics, standard fine-tuning approaches are often inapplicable or ineffective (Saito et al. 2020). This problem of transfer learning with mismatched feature spaces requires novel strategies that can align representations at a higher semantic level, rather than relying on direct feature-level correspondence.

This challenge is acutely manifested in translational bioinformatics, where a primary goal is to leverage vast datasets from model organisms (e.g., mice) to understand human biology and disease (Consortium* et al. 2022; Han et al. 2020). The evolutionary divergence between species means that their genetic feature spaces (i.e., gene expression profiles) are not directly comparable; there is no perfect one-to-one mapping for all genes. This biological reality presents a perfect, high-impact instantiation of the mismatched feature space problem, hindering the direct application of models trained on comprehensive mouse atlases to limited human data, especially given the high dimensionality and noise inherent in genomic dataset.

Current approaches to this cross-species problem fall into two main categories, each with significant limitations from a machine learning perspective. The first involves heuristic feature mapping, where researchers attempt to find corresponding genes (orthologs) between species to create a shared, but often incomplete, feature space (Stumpf et al. 2020; Theodoris et al. 2023). This approach is brittle, as it discards potentially crucial information from non-orthologous genes and is sensitive to the quality of the mapping. The second approach is full-model fine-tuning on the target data, which, given the typical scarcity of human samples, is highly susceptible to overfitting and catastrophic forgetting—erasing the robust features learned from the large source dataset (Kirkpatrick et al. 2017; Wang et al. 2023).

To overcome these limitations, we introduce the Cross-Species Latent Alignment Network (CSLAN), a transfer learning framework that combines principled feature selection with a structural fine-tuning strategy designed specifically for domains with mismatched features. Our approach begins by first tackling the high dimensionality of the data: we employ L1-regularized regression independently on each domain to identify a compact, highly informative subset of features (genes) critical for classification. Then, CSLAN's core idea comes into play: to decouple the learning of the input projection from that of the classification manifold. Operating on these curated feature sets, CSLAN pre-trains an encoder-decoder model on the source data (mouse cells), where the decoder learns to map latent representations to class labels. For transfer, we freeze the pre-trained decoder and latent processor, preserving the learned semantic structure of the classification space. We then instantiate a new, randomly initialized encoder to map target data (human cells) into this fixed latent space, using the frozen decoder's output as a supervised signal.

Our central hypothesis is that this combination of strategic feature reduction and asymmetric fine-tuning transforms the difficult task of learning a classifier from scratch into the much simpler and better-constrained task of learning a projection, by preserving the robustly learned decision boundaries. This forces the new encoder to align the target domain's representations with the source domain's latent structure, effectively distilling knowledge without direct feature mapping. We demonstrate that CSLAN achieves state-of-the-art performance on a challenging trauma-response dataset, providing a powerful, generalizable, and computationally efficient solution for a critical class of transfer learning problems.

The main contributions of this paper are summarized as follows:

- We propose a novel transfer learning framework, CSLAN, featuring an asymmetric fine-tuning strategy. This method preserves a pre-trained decision manifold by freezing the decoder and trains only a new, domain-specific encoder to align domains with mismatched feature spaces.
- We provide a comprehensive empirical validation of CSLAN on a challenging, real-world cross-species bioinformatics task. Our results demonstrate state-of-the-art performance in a few-shot setting, outperforming standard fine-tuning and from-scratch baselines.
- We demonstrate, both quantitatively and qualitatively through latent space visualization, that decoupling the input projection from a conserved decision manifold is a powerful, efficient, and generalizable paradigm for knowledge transfer across disparate domains.

2 Related Work

Heterogeneous Domain Adaptation (HDA)

The challenge of transferring knowledge across domains with mismatched feature spaces is a key problem in HDA (Li et al. 2023). While early methods focused on adversarial alignment (Tzeng et al. 2017), recent advancements have shifted towards more sophisticated strategies, particularly within the practical Source-Free Domain Adaptation (SFDA) setting where adaptation relies only on a pre-trained source model (Li et al. 2024).

CSLAN contributes a distinct and more direct strategy for the supervised, few-shot setting. It operates source-free during fine-tuning, but uses a standard supervised loss and a structurally-modified architecture—replacing the encoder—to achieve a stable and powerful alignment.

Parameter-Efficient Fine-Tuning (PEFT)

PEFT has become the dominant paradigm for adapting large pre-trained models. The state-of-the-art is dominated by methods like Low-Rank Adaptation (LoRA) (Hu et al. 2022) and its highly efficient successors (Dettmers et al. 2023). The field continues to evolve rapidly, new explorations to make these methods even more efficient, for example, by using shared random matrices (VeRA, (Kopiczko, Blankevoort, and Asano 2023)), or by developing new ways to compose PEFT modules across tasks (Wu et al. 2024).

CSLAN introduces a structural PEFT perspective to this landscape for the cross-domain challenge with feature space mismatch. It performs a substitutive adaptation—replacing the entire input-facing encoder. This structural change differs our approach from the additive modification methods, offering a targeted PEFT strategy specifically for heterogeneous domains.

Cross-Species Transfer Learning in Genomics

Within our application domain, prior work has established the feasibility of cross-species knowledge transfer and can be broadly categorized as two main streams.

Ortholog-Based Transfer: The dominant paradigm relies on creating a shared feature space by mapping one-to-one orthologous genes. This has been shown to be effective with models ranging from MLPs (Stumpf et al. 2020) to large-scale transformers like Geneformer, which learns gene embeddings from massive corpora of sequencing data (Theodoris et al. 2023; Ito et al. 2025). The primary limitation of this approach is its reliance on incomplete and potentially biased ortholog maps, forcing models to discard potentially critical non-orthologous genes.

Ortholog-Free and Integration Methods: To overcome this, other methods aim for alignment without direct gene mapping. Semi-supervised approaches like ItClust transfer a pre-trained encoder and use unsupervised clustering to adapt to the target domain (Hu et al. 2020). Other frameworks focus on data integration for building unified atlases, often using variational autoencoders or graph-based manifold alignment to create a shared latent space (Lotfollahi, Wolf, and Theis 2019; Tarashansky et al. 2021). While powerful, these integration-focused methods are often not optimized for a specific supervised task like high-fidelity classification. CSLAN is designed to resolve this trade-off: it is ortholog-free but uses a direct supervised signal to optimize for a specific classification task, outperforming standard fine-tuning strategies.

3 The CSLAN Framework

Datasets and Few-Shot Setup

We validate our framework using challenging mouse (source) and human (target) scRNA-seq datasets from a study of systemic immune response to severe trauma (Chen et al. 2021). The complexity of these data, which capture dynamic cellular states post-injury, provides a rigorous test for our method.

Source (Mouse): The source data was generated from mice subjected to a validated model of severe traumatic injury (polytrauma with hemorrhagic shock). Peripheral blood mononuclear cells (PBMCs) were collected at multiple time points post-injury alongside uninjured controls. The final dataset comprises 3,597 cells across 8 PBMC types (B, Cd4+T, Cd8+T, Mono, NK, NK-T, Neutrophils, RBC), with expression quantified for 12,398 genes. This dataset was split into training (80%), validation (10%), and test (10%) sets for pre-training.

Target (Human): The target data was collected from human trauma patients at three post-injury time points (4h, 24h, 72h) and from healthy volunteers, creating four distinct conditions. The dataset captures 6 corresponding PBMC types (B, Cd4+T, Cd8+T, Mono, NK, NK-T) with expression quantified for 17,038 genes. For the transfer learning phase, we construct a challenging few-shot learning scenario. The fine-tuning dataset consists of only 240 cells (10 cells per type per condition), and the validation and final test set each consists of 120 distinct cells (5 cells per type per condition). This setup rigorously tests the model's ability to generalize from extremely limited target data.

During every training stage, the validation set is used to select the model checkpoint with the best performance. This single, chosen model is then evaluated once on the distinct held-out test set for final, unbiased reporting of accuracy and F1-score. This strict separation ensures that our reported results are a true measure of the model's generalization ability.

Informative Gene Selection

To reduce dimensionality and focus the model on the most predictive signals, we apply L1-regularized Multinomial Logistic Regression (L1-MLR) independently to the source and target training sets. Genes are ranked by the magnitude of their learned coefficients, and the top-k genes are selected. Based on preliminary experiments on the source domain (see Sec 4), we set k=100. This step provides a compact, high-signal feature set for each domain. All selected gene expression data is Z-score normalized using statistics computed only from the respective training sets.

L1-MLR Objective: For each domain $z \in \{m, h\}$ (mouse/human), let n_z be the number of labeled training samples (cells) and G_z the number of available genes (features) in that domain. Let $X_z \in \mathbb{R}^{n_z \times G_z}$ be the design matrix whose i-th row is $x_z^{(i)\top} \in \mathbb{R}^{G_z}$, and let $y \in \{1, \dots, C\}^{n_z}$ be class labels. Multinomial logistic regression with an ℓ_1 penalty solves

$$\min_{W_z, b_z} \frac{1}{n_z} \sum_{i=1}^{n_z} \left[-\log \frac{\exp\left((W_z x_z^{(i)} + b_z)_{y^{(i)}} \right)}{\sum_{c=1}^{C} \exp\left((W_z x_z^{(i)} + b_z)_c \right)} \right] + \lambda \|W_z\|_1, \tag{1}$$

where $W_z \in \mathbb{R}^{C \times G_z}$ are class-by-gene weights, $b_z \in \mathbb{R}^C$ are class biases, $\lambda > 0$.
We ranks genes by ℓ_2 norm of class coefficients. We keep the top-k genes (same k across domains).

Normalization. For each kept gene j, $x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$ with (μ_j, σ_j) computed on the training split only.

Model Architecture and Formalism

CSLAN employs an encoder-decoder architecture designed to explicitly decouple input projection from classification. We formalize the key components as follows:

- Source and Target Encoders (E_m, E_h): For each domain (mouse m, human h), a separate multi-layer perceptron (MLP) encoder, $E_z : \mathbb{R}^k \to \mathbb{R}^d$, maps a k-dimensional input gene vector to a d-dimensional latent embedding. Both encoders share the same architecture but have independent, domain-specific weights.
- Latent Processor (P): A shared module, $P : \mathbb{R}^d \to \mathbb{R}^d$, composed of a series of residual blocks. It processes the initial latent embeddings to model more complex feature interactions in a shared space.
- Decoder (Classifier) (D): A shared MLP, $D : \mathbb{R}^d \to \mathbb{R}^C$, that maps the final d-dimensional representation to a C-dimensional logit vector, where C is the number of cell types.

Our implementation uses a latent dimension of $d = 64$. All MLP modules consist of fully-connected layers followed by Batch Normalization, ReLU activation, and Dropout for regularization.

The CSLAN Two-Stage Training Strategy

Stage 1: Pre-training on Source Domain. In the first stage, we train the mouse-specific components end-to-end on the labeled mouse dataset. The objective is to minimize the standard cross-entropy loss:

$$\mathcal{L}_{\text{pre-train}} = \mathbb{E}_{(x_m, y_m) \sim \text{Data}_m} [\mathcal{L}_{CE}(f_m(x_m), y_m)], \tag{2}$$

where the composite function $f_m(x_m)$ is the full forward pass of the network for the source domain, defined as

$$f_m(x_m) = D(P(E_m(x_m; \theta_{E_m}); \theta_P); \theta_D). \tag{3}$$

And to mitigate class imbalance, we use Class-weighted loss:

$$\mathcal{L}_{\text{CE}}(x, y) = -\sum_{c=1}^{C} w_c \, \mathbf{1}[y = c] \log \pi_c, \quad w_c \propto \frac{1}{\text{freq}(c)}. \tag{4}$$

This stage learns a robust latent representation for cell types and a corresponding classification manifold within the frozen components P and D.

Stage 2: Asymmetric Fine-tuning on Target Domain. This stage contains the core novelty of CSLAN. We freeze the parameters θ_P and θ_D learned in Stage 1. We then instantiate a new, randomly initialized target encoder $E_h(\theta_{E_h})$ and train only its parameters θ_{E_h} on the few-shot human dataset. The objective is to minimize the transfer loss:

$$\mathcal{L}_{\text{transfer}} = \mathbb{E}_{(x_h, y_h) \sim \text{Data}_h} [\mathcal{L}_{CE}(f_h(x_h), y_h)], \tag{5}$$

where the human-specific forward pass

$$f_h(x) = D(P(E_h(x_h; \theta_{E_h}); \theta_P^{\text{frozen}}); \theta_D^{\text{frozen}}). \tag{6}$$

By freezing the decoder and latent processor, we force the new human encoder E_h to learn a projection that maps human cells into the pre-existing latent manifold in a way that is "understood" by the frozen classifier. This constrains the learning problem, prevents catastrophic forgetting, and enables effective knowledge transfer from minimal data.

The overall architecture is illustrated in Figure 1.

The detailed inference pipeline of CSLAN is demonstrated in Appendix A.4.

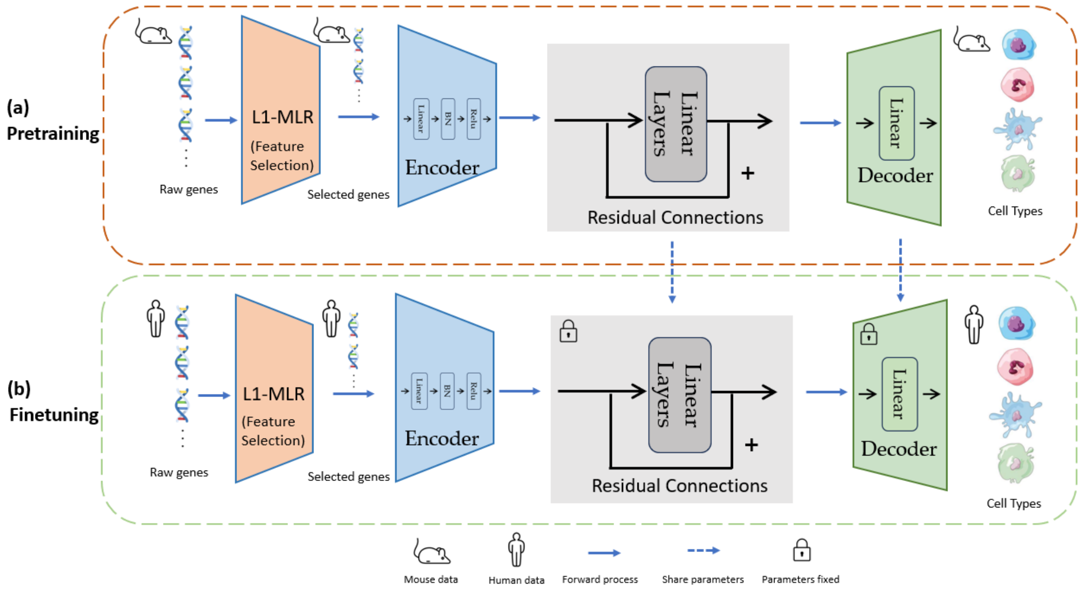


Figure 1: **The CSLAN Architecture and Transfer Learning Strategy.** The model consists of two species-specific encoders (Encoder-m for mouse, Encoder-h for human), a shared latent-space processor with residual connections, and a common decoder for cell type classification. The framework is utilized in a two-phase process: (a) **Pre-training:** The source encoder (mouse), latent processor, and decoder are trained end-to-end on mouse scRNA-seq data to learn robust cell type definitions and latent space representations. (b) **Fine-tuning:** The latent processor and decoder are frozen. A new target encoder (human) is then exclusively trained to project its inputs into the fixed, pre-trained latent space, enabling efficient knowledge transfer.

4 Experiments and Results

Evaluation Metrics

We evaluate model performance on the multi-class cell type classification task using several standard metrics to provide a comprehensive assessment.

- Overall Accuracy: The proportion of correctly classified cells out of the total number of cells in the test set.
- Macro F1-Score: To account for potential class imbalances in the test set, we report the macro-averaged F1-score. This is the unweighted mean of the F1-scores calculated for each cell type individually, providing a balanced measure of performance across all classes. The F1-score for each class c is the harmonic mean of its precision P_c and recall R_c.
- Confusion Matrix: To visualize per-class performance and identify specific error patterns (e.g., confusion between biologically similar cell types), we generate confusion matrices for all test set predictions. The y-axis corresponds to the ground truth label while the x-axis is the predicted label.

For qualitative assessment of the model's learned representations, we use Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018). UMAP plots are used to project the high-dimensional latent embeddings of cells into a two-dimensional space, allowing us to visually inspect the separation of cell type clusters in the latent space. All experiments were run with a fixed random seed of 42 for reproducibility.

Optimal Feature Selection and Source Model Performance

A critical preliminary step in our framework is to establish a strong source model by selecting an optimal, compact feature set. We used L1-regularized MLR to rank all 12,398 mouse genes by their predictive importance and selected the most informative k genes. We then trained and evaluated a mouse classifier for various feature subset sizes (k). A weighted random sampler was employed during the training process to handle the inherent class imbalance and promote a more robust model (Appendix A.1).

As shown in Table 1, performance peaked at k=100 genes, providing the optimal balance of accuracy and dimensionality and achieving 95.82% accuracy with a 0.9577 macro F1-score. The confusion matrix for this optimal k=100 model (Fig. 2(a)) shows high per-class accuracy, with minor confusion only between biologically related T-cell and NK-cell subtypes.

Effect of Feature Set Size (k):

Table 1's result confirms that a curated feature set is superior to using either too few genes (e.g. k=50), which leads to information loss, or too many (e.g. k=1000), which introduces noise and degraded performance. The detailed confusion matrices for each k is shown in Appendix B Fig. **??**.

To qualitatively compare the effect of k values on the inherent structure of the input data, Figure 3 shows the UMAP projections of the mouse test set using the raw expression data. While k=50 shows some clustering, the cell type separation becomes significantly clearer and more distinct at

Table 1: Mouse cell type classification accuracy on test set with varying numbers of top k genes selected.

No. Genes (k)	Accuracy (%)	Macro F1-Score
10	86.94	0.8706
50	92.50	0.9268
100	**95.82**	**0.9577**
500	89.72	0.8971
1000	82.78	0.8163

k=100. Increasing the feature set to k=1000 introduces noise and leads to more diffuse, less-separated clusters. This visual evidence aligns with the quantitative findings that k=100 provides the best balance of signal and noise.

Comparison of Feature Selection Methods:

To validate our choice of Multinomial Logistic Regression, we benchmarked it against several common feature selection strategies (Table 2). We used each method to select the top 100 genes from the mouse training dataset and then trained our classifier architecture on that subset.

The L1-MLR outperforms all other methods. As expected, Random Selection performs poorly, confirming that an intelligent, data-driven selection strategy is essential. The heuristic of selecting genes with High Variance also performs poorly, indicating that expression variance alone is not a reliable proxy for predictive importance in this context. While the Random Forest and Mutual Information methods are strong performers, they do not reach the level of L1-MLR. This is likely due to L1-MLR's inherent ability to handle sparse, high-dimensional data by shrinking the coefficients of redundant or irrelevant genes, resulting in a more robust and potent feature set for classification.

Model Performance:

The value of our deep learning approach is visually demonstrated in Figure 4. While the raw top-100 input genes provide a reasonable initial separation of cell types (Fig. 3(b)), the 64-dimensional latent space learned by the pre-trained CSLAN model exhibits a qualitatively superior representation, with significantly tighter and more distinct clusters, as shown in Fig. 4(a). This comparison highlights the model's ability to learn a powerful non-linear transformation that enhances class separability. The UMAP visualization is calculated and plotted based on the intermediate results of the model, which effectively projects the high-dimensional geometric structure of the latent space, providing direct visual evidence of this improved representation.

Overall, the high quantitative accuracy and the clear qualitative improvement in representation provide compelling proof of effective source model training and ability of representative cell identity learning, establishing a robust foundation for our transfer learning experiments.

CSLAN Outperforms Baselines in Few-Shot Transfer

We now turn to the primary evaluation: the few-shot, cross-species transfer task. CSLAN with the proposed asymmetric fine-tuning strategy was compared against several key baselines on the human test set. The results are shown in Table 3.

Table 2: Comparison of feature selection methods, with performance evaluated on the mouse test set using the top 100 genes selected by each method.

Method	Accuracy (%)	Macro F1-Score
L1-MLR	**95.82**	**0.9577**
Random Selection	44.51	0.4282
High Variance	33.24	0.2946
Random Forest	93.32	0.9303
Mutual Information	92.63	0.9209

CSLAN achieves a state-of-the-art accuracy of 95.83% and a macro F1-score of 0.9591. This performance surpasses all key baselines:

- **Human-Only (all-genes):** A naive baseline trained from scratch on all 17,038 human genes performs poorly (80.83%), highlighting the challenge of learning from high-dimensional, low-sample data.
- **Human-Only (k=100):** Adding our feature selection step substantially improves the "training-from-scratch" performance to 92.50%, yet this still falls short of CSLAN, demonstrating that feature selection alone is insufficient.
- **Zero-Shot Transfer:** Directly applying the pre-trained mouse model to the 100 selected human genes yields an accuracy of 17.51%, which is near random chance. This is expected, as the mouse encoder's learned feature mapping is meaningless for the distinct human gene space, confirming that adaptation is essential to bridge the species gap.
- **Full Fine-tuning (Full-FT):** The most direct transfer learning competitor, fine-tuning the entire model, is overfitting on the few-shot data and only reaches a accuracy of 92.12% (compared with 95.83%). This accuracy is close to that of the "training from scratch" results, as the dataset and model size is limited in this scenario. This highlights that CSLAN's constrained tuning is a superior regularization strategy, preventing catastrophic forgetting and common overfitting for small dataset.

The validation curves in Figure 2(c) show that CSLAN converges effectively on the small fine-tuning set with a small batch size (see Appendix A.3 for training dynamics). The resulting classifier demonstrates high and balanced performance, as shown by the confusion matrix in Figure 2(b). A detailed comparison with the primary baselines, Full Fine-tuning and Human-Only (k=100), reveals the superiority of our approach. As shown in Figure 5, both baseline models struggle to learn the fine-grained distinctions between lymphocyte subtypes from the limited few-shot data, exhibiting confusion between the Cd4+ T, Cd8+ T, and NK-T populations. In contrast, CSLAN achieves a much cleaner separation of these challenging classes.

This confirms our central hypothesis: preserving a learned decision manifold while adapting only the input projection

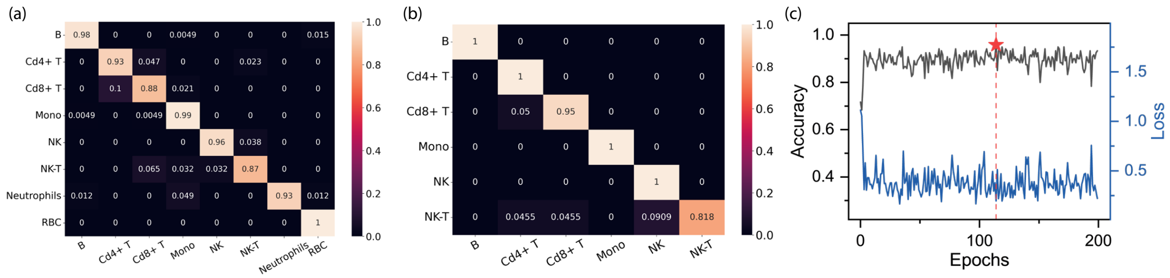


Figure 2: **CSLAN Performance on Source and Target Domains.** (a) Confusion matrix for the best-performing pre-trained model on the mouse test set, achieving 95.82% accuracy. (b) Confusion matrix for the asymmetrically fine-tuned CSLAN model on the 6-class few-shot human test set, achieving 95.83% accuracy. (c) Validation loss and accuracy curves during the human fine-tuning phase. The final model was selected from the epoch with the highest validation accuracy (dashed line).

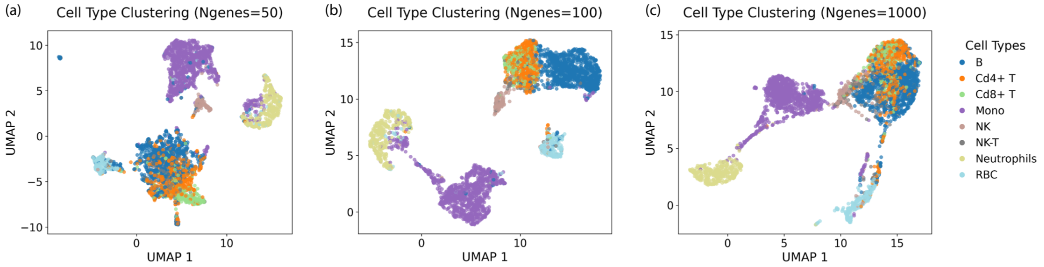


Figure 3: **UMAP visualization** of the mouse input data for different numbers of top-k genes (k = (a)50, (b)100, (c)1000). The clearest separation of cell type clusters is observed at k=100, validating our choice of feature subset size.

Table 3: CSLAN performance on the human few-shot test set. Our method (CSLAN) with encoder-only tuning outperforms all baselines. Baselines include training from scratch on human data (Human-Only), full-model fine-tuning (Full-FT), and zero-shot transfer from the mouse model.

Model	**Accuracy (%)**	**Macro F1**
CSLAN (Ours)	**95.83**	**0.9591**
Baselines		
Human-Only (k=100)	92.50	0.9232
Human-Only (All Genes)	80.83	0.7989
Full-FT (k=100)	92.12	0.9181
Zero-Shot (k=100)	17.51	0.1793

is a superior regularization strategy for few-shot transfer. It effectively prevents catastrophic forgetting and preserves the crucial feature distinctions necessary for high-fidelity classification across mismatched domains.

Visualization of Latent Space Alignment

To qualitatively validate that CSLAN achieves a meaningful alignment, we visualized the topology of the learned latent spaces for both the mouse and human domains (Figure 4). While the global orientation of a UMAP projection is arbitrary, the relative topology of the cell type clusters provides powerful insights into the success of the transfer. This visualization serves as the ultimate test of our framework's ability to align the target domain with the structured manifold learned from the source.

The analysis reveals that CSLAN successfully transfers core topological features. For instance, the B-cell lineage is clearly isolated from other lymphocytes and the Cd8+T is at the top-right corner in both the mouse and human manifolds. Similarly, the close proximity of NK and NK-T cells is recapitulated in the human space, correctly mirroring their biological similarity. The preservation of these key relative geometries demonstrates that CSLAN has learned a inherited semantic map of cell identity. This confirms that our asymmetric fine-tuning strategy effectively bridges the species gap by aligning the human data to a biologically meaningful manifold. An interesting divergence feature is also captured. While Cd4+ and Cd8+ T-cells are adjacent in the mouse space, they are more distinctly separated in the human manifold. This may reflect a genuine species-specific transcriptional response to trauma that our model has successfully learned to represent.

In summary, the preservation of these essential topological features and the clear clustering for both species provide strong visual evidence that our asymmetric fine-tuning strategy works. CSLAN does not merely learn a new classifier; it effectively bridges the species gap by aligning the human data to a shared, biologically meaningful semantic map.

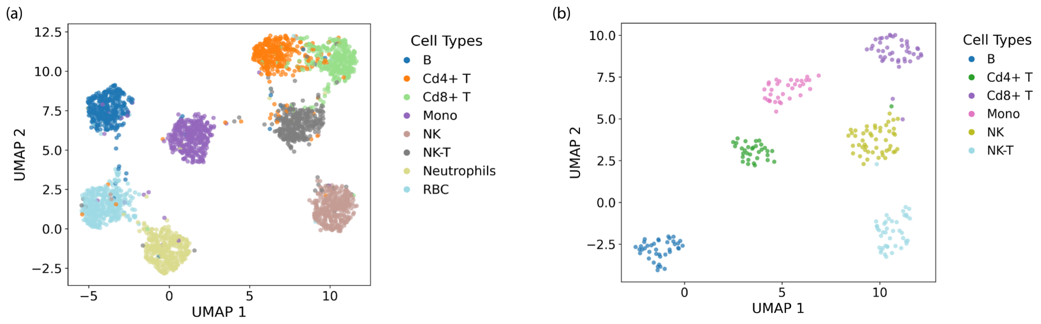
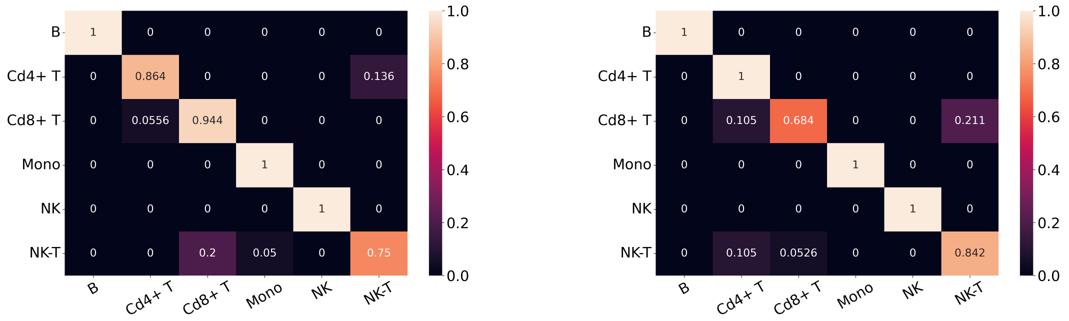


Figure 4: **Aligned Latent Spaces.** (a) The pre-trained CSLAN model transforms the raw top-100 selected mouse genetic data into a highly structured latent space with significantly clearer class separation than initial clustering. Significant improvement is shown for Cd4+/8+ T cells' separation. (b) After asymmetric fine-tuning, the CSLAN human encoder successfully projects human cells into a topologically similar latent space, preserving key inter-class relationships with effective alignment.



(a) Human-Only (From Scratch, k=100)

(b) Full Fine-tuning (Full-FT, k=100)

Figure 5: **Confusion matrices for baseline models** on the 6-class human test set. (a) The model trained from scratch achieves 92.50% accuracy but shows notable confusion between specific cell types. (b) The Full Fine-tuning model achieves 92.12% accuracy and exhibits similar error patterns, failing to significantly improve upon the from-scratch performance. Both are outperformed by CSLAN's targeted approach (Figure 2(b)).

Table 4: Knowledge ablation study. Performance of CSLAN on the 6-class human test set when the pre-trained decoder is missing specific biological priors from the source model. The baseline uses priors for all corresponding classes.

Missing Prior(s) in Source Model	**Accuracy (%)**
None (Baseline)	95.83
Single Priors Removed	
NK-T	95.00
Mono	95.00
Cd8+ T-cell	93.33
B-cell	90.83
Cd4+ T-cell	**98.35**
Multiple Priors Removed	
Cd4+ T & Cd8+ T	90.83

Dissecting the Learned Manifold

To investigate the compositionality of the knowledge within the frozen decoder and the impact of individual biological priors, we conducted a series of knowledge ablation experiments. We pre-trained source models on different subsets of the mouse cell types and then transferred them to the full 6-class human task. For each transfer, the human decoder was constructed by inheriting the weights for only those classes present in its specific pre-training run. Weights for any human class not seen during pre-training were randomly initialized and fine-tuned alongside the human encoder.

The results, summarized in Table 4, reveal a complex interplay between conserved knowledge and species-specific features. As expected, removing the pre-trained priors for B-cells or Cd8+ T-cells reduced performance to 90.83% and 93.33% respectively, demonstrating the high value of transferring these conserved cell type definitions.

Remarkably, not all priors were beneficial. The baseline model, using all 6 corresponding priors, achieved 95.83% accuracy. However, when the pre-trained prior for Cd4+ T-cells was removed, the model's performance increased to a new state-of-the-art of 98.35%. This fascinating result suggests a case of subtle negative transfer, where the pre-trained mouse definition of a trauma-response Cd4+ T-cell may be a suboptimal constraint for representing its human counterpart. By allowing the model to learn the human Cd4+ T-cell representation from scratch, unconstrained by the mouse prior, it discovers a more effective, human-specific solution.

These findings provide three key insights. First, the high performance across all experiments demonstrates the robustness of the CSLAN framework, showing it can successfully learn to classify target categories even when no corresponding prior exists. Second, the results suggest the learned latent space is highly modular, where the semantic definitions for individual cell types are largely independent. This compositionality, allowing the model to learn ”new” classes without catastrophic interference, is a key feature of a well-structured representation. Third, and most critically, our results highlight that while transferring conserved knowledge is powerful, the flexibility to discard suboptimal priors and learn species-specific features from scratch is equally vital for achieving optimal performance.

5 Conclusion

We introduced CSLAN, a novel transfer learning framework for the challenging setting of mismatched and non-mappable feature spaces. Our core innovation is an asymmetric fine-tuning strategy—preserving the learned biological priors while exclusively learning a new domain-specific projection to a fixed semantic space. CSLAN achieves state-of-the-art, few-shot classification performance in a critical cross-species bioinformatics task, bridging the translational gap between mouse and human.

While our validation is on a single mouse-to-human context, the core principle of CSLAN is generalizable. The decoupling of the domain-specific input projection from the decision manifold learning presents a powerful, parameter-efficient, and interpretable strategy for knowledge transfer. We believe this paradigm holds broad implications for other AI domains facing feature space mismatches, as well as for accelerating discovery in the sciences.

Data Availability

The datasets and code used in this study are available upon reasonable request. A public release is planned and will be shared via an online repository upon publication.

References

Chen, T.; Delano, M. J.; Chen, K.; Sperry, J. L.; Namas, R. A.; Lamparello, A. J.; Deng, M.; Conroy, J.; Moldawer, L. L.; Efron, P. A.; et al. 2021. A road map from single-cell transcriptome to patient classification for the immune response to trauma. *JCI insight*, 6(2): e145108.

Consortium*, T. T. S.; Jones, R. C.; Karkanias, J.; Krasnow, M. A.; Pisco, A. O.; Quake, S. R.; Salzman, J.; Yosef, N.; Bulthaup, B.; Brown, P.; et al. 2022. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594): eabl4896.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36: 10088–10115.

Han, X.; Zhou, Z.; Fei, L.; Sun, H.; Wang, R.; Chen, Y.; Gu, H.; Liu, W.; Ye, F.; Li, T.; et al. 2020. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808): 321–328.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, J.; Li, X.; Hu, G.; Lyu, Y.; Susztak, K.; and Li, M. 2020. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nature machine intelligence*, 2(10): 607–618.

Ito, K.; Hirakawa, T.; Shigenobu, S.; Fujiyoshi, H.; and Yamashita, T. 2025. Mouse-Geneformer: A deep learning model for mouse single-cell transcriptome and its cross-species utility. *Plos Genetics*, 21(3): e1011420.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Kopiczko, D. J.; Blankevoort, T.; and Asano, Y. M. 2023. Vera: Vector-based random matrix adaptation. *arXiv preprint arXiv:2310.11454*.

Li, H.; Li, F.; Liu, J.; and Gong, C. 2023. A survey on heterogeneous domain adaptation. *ACM Computing Surveys*, 55(9): 1–37.

Li, J.; Yu, Z.; Du, Z.; Zhu, L.; and Shen, H. T. 2024. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5743–5762.

Lotfollahi, M.; Wolf, F. A.; and Theis, F. J. 2019. scGen predicts single-cell perturbation responses. *Nature methods*, 16(8): 715–721.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Saito, K.; Kim, D.; Sclaroff, S.; and Saenko, K. 2020. Universal domain adaptation through self supervision. *Advances in neural information processing systems*, 33: 16282–16292.

Stumpf, M.; Parth, M.; Rüdisser, M.; Wagner, L.; Lotfollahi, M.; Hölzl-Gruber, J.; Nieto-Pelegrín, E.; Dander, M.; Löffler, J.; Trajanoska, K.; et al. 2020. Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell RNA sequencing. *Communications biology*, 3(1): 643.

Tarashansky, A. J.; Musser, J. M.; Khariton, M.; Li, P.; Arendt, D.; Quake, S. R.; and Wang, B. 2021. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife*, 10: e66747.

Theodoris, C. V.; Xiao, Y.; Zvyagin, M.; Dela-Merced, F.; Luo, Y.; Tran, V.; Dai, H.; Tan, M.; and Chang, H. Y. 2023. Transfer learning for gene expression analysis. *Nature Methods*, 20(4): 521–528.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

Wang, L.; Zhang, X.; Liu, K.; Hang, H.; and Liu, J. 2023. A comprehensive survey on catastrophic forgetting in deep learning. *arXiv preprint arXiv:2303.07223*.

Wu, C.; Zhang, Z.; Liu, Y.; Liu, Z.; Li, J.; Niu, S. Z.; Gonzalez, J. E.; and Stoica, I. 2024. LoRAHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. In *International Conference on Learning Representations (ICLR)*.

Zhu, Z.; Lin, K.; Jain, A. K.; and Zhou, J. 2023. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13344–13362.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) partial

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/-partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) no

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) no

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes