Nature Communications

https://doi.org/10.1038/s41467-026-69696-3

Article in Press

Multi-species integration, alignment and annotation of single-cell RNA-seq data with CAMEX

Received: 6 March 2025

Accepted: 30 January 2026

Cite this article as: Guo, Z.-H., Huang, D.-S., Zhang, S. Multi-species integration, alignment and annotation of single-cell RNA-seq data with CAMEX. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-69696-3

Zhen-Hao Guo, De-Shuang Huang 黄德双 & Shihua Zhang 世

Multi-species integration, alignment and annotation of single-cell RNA-seq data with CAMEX

Zhen-Hao Guo[1,2,3], De-Shuang Huang (黄德双)[2,3*], Shihua Zhang (张世华)[4,5,6*]

[1]College of Electronics and Information Engineering, Tongji University, Shanghai 200123, China.
[2]Ningbo Institute of Digital Twin, Ningbo Key Laboratory of Multi-Omics & Multimodal Biomedical Data Mining and Computing, Eastern Institute of Technology, Ningbo 315201, Zhejiang, China.
[3]Institute for Regenerative Medicine, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai 200123, China.
[4]State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
[5]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.
[6]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China.

*****Co-corresponding authors. De-Shuang Huang, Tel: 86-18701810288; Email: dshuang@eitech.edu.cn. Shihua Zhang, Tel: 86-10-82541360; Email: zsh@amss.ac.cn.

Abstract

Single-cell RNA-seq (scRNA-seq) data from multiple species present remarkable opportunities to explore cellular origins and evolution. However, integrating and annotating scRNA-seq data across different species remains challenging due to the variations in sequencing techniques, ambiguity of homologous relationships, and limited biological knowledge. To tackle the above challenges, we introduce CAMEX, a heterogeneous Graph Neural Network (GNN) tool that leverages many-to-many homologous relationships for multi-species integration, alignment, and annotation of scRNA-seq data from multiple species. Notably, CAMEX outperforms state-of-the-art methods integration on various cross-species benchmarking datasets (ranging from one to eleven species). Besides, CAMEX facilitates the alignment of diverse species across different developmental stages, significantly enhancing our understanding of organ and organism origins. Furthermore, CAMEX enables the detection of species-specific cell types and marker genes through cell and gene embedding. In short, CAMEX holds the potential to provide invaluable insights into how evolutionary forces operate across different species at single-cell resolution.

Keywords:

Single-cell RNA-seq (scRNA-seq) data, multi-species, data integration, development alignment, cell annotation, heterogeneous graph neural network

Introduction

Single-cell omics techniques enable the profiling of thousands of cells in one experiment, revealing a comprehensive landscape of heterogeneous cells [1]. These techniques facilitate the discovery of rare cell states, infer gene regulatory networks, and enhance our understanding of development and differentiation processes at a finer resolution [2]. With advancements in sequencing methods, scRNA-seq data from different species, including humans [3], monkeys [4], mice [5], zebrafish [6], and more, have been generated by various studies. These datasets provide clues on the origin of cellular diversity and evolutionary drivers of cellular morphology [7]. Diverse studies have revealed many cryptic cell types across multiple species [8]. Comparing gene expression of individual cells across species uncovers species-specific transcriptomic regulation [9], provides novel insights into the phylogeny of cells and organs, and elucidates how cells function in different developmental stages [10]. Integrating scRNA-seq data across species advances our understanding of molecular and cellular biology and becomes an essential first step toward comparative biology [11, 12]. These scRNA-seq data enable cross-species comparisons and analyses, shedding light on cellular functions and evolutionary dynamics. Only through these comparisons can we eventually decipher and reconstruct species evolution [13].

Single-cell RNA sequencing experiments can only process a limited number of cells every time. Data from various batches exhibit significant variations due to differences in capture time, operating personnel, and sequencing platforms [14]. For instance, batch effects make it challenging to align the developmental stages of different species, hinder cross-species cell annotation and the identification of novel cell types and subtypes. Overall, these batch effects can obscure biological signals and interfere with our ability to analyze specific biological changes of interest [15]. Researchers have developed many methods, such as Harmony [16], scVI [17], and SingleCellNet (SCN) [18], for integrating and annotating scRNA-seq data within a single species. Furthermore, inspired by the rapid advancement of large models in natural language processing, analogous foundation models have emerged in the single-cell field, such as scGPT [19], which facilitates downstream tasks like batch effect removal. However, existing methods align features in the expression profile matrix across different species solely through one-to-one homologous genes, resulting in the loss of much biological information [20].

Several methods have made initial explorations into the integration and annotation of cross-species scRNA-seq data, such as BPA (Biological Process Activity) [21], SAMap [22], SATURN [23] and CAME [24]. However, the limited biological knowledge available for non-model organisms poses challenges for implementing BPA. SAMAP relies on sequence alignment, a computationally intensive process, and is primarily used for visualization purposes rather than generating embeddings for downstream tasks [19]. SATURN measures inter-species gene similarity using protein-language models but faces challenges from proteome-transcriptome mismatches (notably from alternative splicing events) and requires cell annotation inputs that might limit its application. While our

previous annotation method, CAME, utilizes many-to-many homologous gene relationships, it is restricted to pairwise species alignment (two species) within a shared embedding space and cannot perform integration without cell labels across multiple species datasets.

In a word, integrating scRNA-seq data obtained from diverse sources and non-model organisms faces challenges due to technical and biological factors. First, the data from different sequencing platforms vary greatly, including the difference of detected UMIs and genes. Second, aligning the features of expression profile matrices from other species through the relationships between orthologous genes poses a significant difficulty. Last, processing scRNA-seq data from non-model organisms with limited biological knowledge, e.g., specific regulatory networks, distinctive communication mechanisms, and individual diversity, is a complex challenge.

To this end, we introduce CAMEX, a heterogeneous graph neural network (GNN) tool that leverages many-to-many homologous relationships for integration, alignment, and annotation of scRNA-seq data from multiple species. Specifically, we construct a heterogeneous graph consisting of cells and genes based on expression matrices and many-to-many homology relationships. We design a heterogeneous graph neural network as an encoder to encode cells and genes from different species into the common embedding space. For the batch correction task, we design two decoders including a dot-product decoder to reconstruct the edges in the heterogeneous graph and a full-connected decoder to reconstruct the cell expression. For the cell annotation task, we feed the cell embedding into the graph attention network (GAT) for classification. CAMEX achieved competitive performance compared to state of the art (SOTA) methods in integrating scRNA-seq data in different scenarios including the liver (involving four species with two technologies), ovary (encompassing three species with rare populations) and pancreas (comprising three species with multiple batches). Furthermore, CAMEX successfully integrated a testis dataset across 11 different species, preserving the sperm differentiation trajectory. Additionally, CAMEX correctly integrated RNA-seq data from seven species, spanning multiple organs and various developmental stages. Notably, CAMEX could align and predict developmental time beyond the Carnegie stages across different species. Finally, we demonstrated that CAMEX effectively annotates multi-species cell labels, even for distantly related species, and can discover new subgroups and marker genes.

Results

Overview of CAMEX

CAMEX takes diverse single-cell datasets from various species generated using the same or different technologies as input (**Fig. 1**). The expression matrix can be transformed into a bipartite graph A, with cells and genes as vertices. An edge A_{nm} is established between a cell n and a gene m if the cell n expresses the gene m. Then, we construct a *k*-nearest neighbor (kNN) graph to connect cells according to their

expression similarity. In addition, self-loops are added for each cell and gene to prevent excessive smoothing in the GNN's message passing. Finally, we connect the genes of different species based on many-to-many homologous relationships. The initial feature of each cell node corresponds to its preprocessed expression, while the initial feature of each gene node is set to empty. CAMEX employs a heterogeneous GNN encoder to embed each node into a low-dimensional common space by nonlinearly propagating features from neighboring nodes to center nodes.

As to the integration task, CAMEX utilizes an encoder-decoder architecture with no cell type labels in any dataset. Specifically, it employs a heterogeneous GNN as an encoder, a dot product decoder to reconstruct the edges in the cell-gene heterogeneous graph, and a fully connected decoder to reconstruct the cell expression. The loss function comprises both cross-entropy loss between the predicted and raw labels of edges and mean square error (MSE) loss between the predicted and input expression of cells. CAMEX minimizes them to optimize the model parameters by the back-propagation algorithm in an unsupervised manner.

As to the annotation task, CAMEX adopts an encoder-classifier architecture with at least one given reference dataset containing cell labels. It employs a heterogeneous GNN as an encoder and a GAT to predict cell labels. The loss function computes the cross-entropy loss between the predicted and ground truth labels of cell annotations in the reference dataset. CAMEX minimizes it to optimize the model parameters by the back-propagation algorithm in a semi-supervised way.

We collected various cross-species scRNA-seq datasets to benchmark CAMEX and other baseline methods. In the integration task, we applied CAMEX to the datasets and compared its performance against other SOTA methods, i.e., Harmony [16], PyLiger [25, 26], scVI [17], Scanorama [27], scGEN [28], scGPT [19], Seurat [29], SAMap [22], and SATURN [30]. For SAMap and SATURN, we conducted our experiments following their respective tutorials (**Methods**). For CAMEX, we leveraged many-to-many homologous gene relationships to construct heterogeneous networks. For the other methods, we employed one-to-one homologous genes to align expression matrices for the baseline methods. We conducted a quantitative evaluation of batch correction and biological variation conservation. Specifically, the batch correction score using metrics including kBET (k-nearest-neighbor Batch Effect Test) [31], PCR batch (Principal Component Regression of the batch covariate) [31], ASW batch (Average Silhouette Width across batch) [31], and graph connectivity [14]. The biological variation conservation score using metrics including ARI (Adjusted Rand Index) [32], NMI (Normalized Mutual Information) [33], Cell-type ASW [14], ISO label F1 (Isolated label F1) [14], and ISO label silhouette (Isolated label silhouette) [14]. Higher values for each metric indicate better performance (**Methods**). Finally, we computed an overall score as a 1:1 weighted average of the batch correction score and the biological variation conservation score to indicate the overall model performance. Notably, some metrics for SAMap were not calculable. Although we computed its Batch Correction, Biological Variation Conservation, and Overall Average

Score, these metrics failed to reflect its ranking precisely. This finding is confirmed by the prior research [19], which has similarly pointed out the inapplicability of these metrics to SAMap. Furthermore, we evaluated the model's performance in retaining the developmental trajectory by calculating pseudo-time both before and after integration (**Methods**).

Similar to the integration task, here we applied CAMEX to the datasets and compared its performance against other baseline methods, i.e., SCN [18], scANVI [34], and Cell BLAST [35]. We conducted ACC (accuracy), Prec (precision), and F1 (F1 score) for the quantitative evaluation of each baseline method.

CAMEX achieves superior integration performance in different multi-species scenarios

We curated three multi-species scRNA-seq datasets, comprising the liver data (involving four species with two technologies), ovary data (encompassing three species with rare populations), and pancreas data (comprising three species with seven batches by six different technologies), to simulate a diverse range of real-world scenarios for benchmarking CAMEX against other baseline methods. In this unsupervised integration setting, CAMEX operates without the need for manually annotated cell type labels.

In the liver dataset, the scRNA-seq data for humans, macaques, and mice originating from the same study were generated using the same sequencing platform [36]. However, the zebrafish scRNA-seq data were obtained from the zebrafish atlas [13], where the detected number of Unique Molecular Identifiers (UMIs) and genes were relatively low (**Supplementary Fig. S1a, b**). Each of the datasets contains distinct cell types unique to that respective species. While the human, monkey, and mouse datasets exhibit balanced cell types, the zebrafish dataset predominantly consists of hepatocytes (**Supplementary Fig. S1c**). These two factors significantly increased the complexity of data integration. In this context, CAMEX was the only method capable of accurately integrating scRNA-seq data across different species and achieving the highest overall score (**Fig. 2a-c, Supplementary Fig. S1d, Supplementary Data1**). It effectively clustered hepatocytes, fibroblasts, cholangio, endothelial cells, B plasma cells, and Kupffer cells (macrophages). Compared to other methods, CAMEX achieved higher Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) scores, reflecting its superior ability to cluster cells of the same type. Although scVI also aggregated similar cell types across different species, its large intra-class distances led to lower kBET and ASW batch scores, positioning it below CAMEX for the overall ranking. PyLiger performed well in integrating hepatocytes but it cannot address other cell types well. Harmony excelled in integrating non-hepatocytes, but failed when dealing with hepatocytes. The integration result of Scanorama was similar to Harmony, but its overall performance lagged behind Harmony. Notably, scGen could integrate hepatocytes cells from four species while failing to integrate other cell types. In contrast, SATURN, Seurat v4, Harmony, scGPT, Scanorama, and SAMap achieved integration for all cell types except hepatocytes. In addition, all methods

successfully integrated liver-specific macrophages (macrophages and Kupffer cells). We examined the expression of typical marker genes in each cell type and verified the identities of macrophages and Kupffer cells (**Supplementary Fig. S1e**). The genes from each cell type vary considerably across species (**Supplementary Fig. S1f**). We also performed computational efficiency, hyperparameter comparison, and ablation experiments (**Supplementary Figs. S2-S4, Supplementary Data2-4**). Despite variations in evaluation metrics, most settings successfully isolated distinct cell types in the UMAP. An interesting observation was the clear separation of Kupffer cells and Macrophages when the hidden layer dimension was increased to 512 (**Supplementary Fig. S2c**).

In the ovary dataset, the scRNA-seq data for humans [37-42], macaques [43], and mice [44] originating from different sequencing platforms resulted in distinct data types (**Supplementary Fig. S5a-c**). The macaque dataset contains a rare, unique oocyte cell type. In this scenario, CAMEX, scGen, SATURN, Seurat v4, PyLiger, scVI, and scGPT can integrate the scRNA-seq data from three species while preserving this rare cell population, but CAMEX achieved a higher overall score (**Supplementary Fig. S5d-e, Supplementary Fig. S6, Supplementary Data5, 6**).

In the pancreas dataset, the scRNA-seq data came from humans [39], macaques [45], and mice [39] with humans containing five batches (by five different technologies) of data (**Supplementary Fig. S7a**). We used this to simulate a multi-species, multi-batch scenario. The human batches in this dataset contain roughly the same cell types**.** In addition, mice and macaques contain cell types that are subsets of humans (**Supplementary Fig. S7b, c)**. In this scenario, Seurat v4, SATURN and CAMEX can preserve the strong biological signal and effectively separate different cell types (**Supplementary Fig. S7d, e, Supplementary Fig. S8, Supplementary Data7**, **8)**. Almost all methods, excluding scGPT, Scanorama, and SAMap exhibited relatively decent integration performance and can clearly distinguish various cell types. However, Harmony tended to confuse the acinar and ductal cell types. PyLiger confused the acinar and ductal cells and the delta and gamma cells but failed to integrate the alpha and beta cell types. Conversely, scVI cannot distinguish acinar and ductal cells well and faces challenges when integrating the gamma cells.

Overall, CAMEX has demonstrated competitive and robust performance across various multi-species scenarios. It can effectively handle substantial differences and multi-batch data and identify and preserve rare cell populations. Nevertheless, other baseline methods exhibit higher variability in performance, resulting in less stable outcomes.

CAMEX uncovers the conserved differentiation process in the testis across 11 species

The testis is a unique organ with the highest number of tissue-specific genes in its transcriptome, and sperm development represents a continuous and conserved process [46]. Consequently, comparing these continuous and critical biological processes across

different species is essential for evolutionary biology. Here, we collected and utilized the scRNA-seq testis data from 11 species, encompassing five distinct cell types, i.e., somatic, Sertoli, spermatogonia, spermatocytes, and spermatids [47]. Although this dataset originates from the same source and exhibits consistent cell types (**Supplementary Fig. S9a)**, integrating it across 11 species remains a significant challenge due to the scarcity of one-to-one homologous genes shared among them (**Supplementary Fig. S9b)**. Worth noting is that the original research only merges scRNA-seq data from seven primate datasets via LIGER (v.0.5.0) [26] based on primate one-to-one orthologues from Ensembl release 87 [47].

Firstly, we integrated the testis scRNA-seq data across 11 species using CAMEX and other baseline methods. In this unsupervised integration setting, both CAMEX and the baseline methods are precluded from using any manually annotated cell type labels. We found that SATURN and SAMap could not run normally due to prohibitive memory requirements (GPU memory for SATURN, CPU memory for SAMap), which led us to exclude them from our analysis. From the UMAP, we can see that CAMEX successfully preserved the continuous trajectory of sperm differentiation and separated distinct clusters corresponding to different cell types (**Fig. 3a, b, Supplementary Data9, 10**). Furthermore, we calculated the pseudo-times both before and after integration, using their correlation as the trajectory reservation metric. CAMEX achieved the highest value in this regard (**Supplementary Fig. S10)**. Analyzing the results of CAMEX, we observed that the Platypus dataset almost entirely lacked Sertoli cells, consistent with the original findings (**Supplementary Fig. S11)**. PyLiger partially retained the developmental trajectory but failed to integrate spermatids and spermatocytes, erroneously merging somatic cells with spermatids. scVI struggled to effectively integrate Sertoli and spermatocyte cells from chickens with those from other species. Harmony failed to integrate Sertoli cells from opossums and spermatocytes from chickens. Scanorama achieved the worst integration results.

Inspired by previous literature [48], to comprehensively evaluate CAMEX's performance across varying evolutionary distances and increasing numbers of species, we designed two integration scenarios. First, we conducted pairwise integration analyses between humans and each target species, covering a broad phylogenetic spectrum from closely related primates to distant vertebrates. The results clearly show CAMEX excelled in batch correction, while SATURN performed strongly in biological variation conservation. Overall, CAMEX demonstrated an advantage over other methods in both quantitative metrics, rankings, and qualitative UMAP visualization. Second, we created multi-species integration scenarios with progressively expanded species sets, systematically increasing from two to eleven species. CAMEX also consistently maintained high performance across all scenarios (**Supplementary Figs. S12-S24, Supplementary Data11-21)**. Overall, most integration algorithms, particularly those relying solely on one-to-one homologous genes, show diminished performance as the evolutionary distance between species increases. Additionally, we conducted an ablation experiment in which human

data were integrated separately with each of the other ten species using only one-to-one homologous genes. As shown in **Supplementary Figure S24b**, integration performance remained comparable to that achieved with many-to-many homologous genes when closely related primates were involved. However, as evolutionary distance increased—particularly in the human–chicken comparison—a marked decline in performance was observed. These findings suggest that the use of many-to-many homologous genes is a critical factor enabling CAMEX to maintain robust integration performance across large evolutionary spans. Overall, methods like CAMEX, which leverage many-to-many homologous genes, and STAURN, which utilizes sequence similarity, are less impacted by this increasing species divergence.

Secondly, given that the cell subtypes from seven primate species have been manually annotated, we leveraged the cell embeddings obtained from CAMEX to assess the quality of cell embedding in the CAMEX integration task. Using the human dataset as a reference, we applied a simple k-Nearest Neighbors (kNN) classifier to annotate the subtypes for other primate cells. The predicted results are almost consistent with the manual annotations (normalized confusion matrix) (**Supplementary Fig. S25**). Additionally, we annotated the cell subtypes of four other non-primate species. Cell subtypes exhibit similarities with the annotated major population (**Supplementary Fig. S26a**). CAMEX identified some cells in the platypus dataset as Sertoli cells. Through further differential analysis, we identified differentially expressed genes (DEGs) (**Supplementary Fig. S26b**). These genes are highly relevant to spermatogenesis in the testis. For example, the *CYP17A1* gene codes for a protein involved in testosterone synthesis and plays a vital role in cholesterol transport [49]. Similarly, the *STAR* gene encodes a protein critical for cholesterol transport across mitochondrial membranes, thereby regulating testosterone synthesis [50]. We speculate that the labeled cells may represent a rare population of potential platypus-specific Sertoli cells.

Thirdly, we obtained the mean embedding for each cell type across all species and conducted hierarchical clustering (**Supplementary Fig. S27**). The clustering outcomes indirectly confirm the accuracy of cell integration. Notably, opossum and mouse spermatogonia were mistakenly grouped with spermatocytes. We attribute this misclassification to the relatively young age of samples in the original dataset [47]. Interestingly, spermatogonia appears more similar to somatic cells than spermatocytes and spermatids. This finding suggests that the transcriptional differences during spermatogenesis are crucial [47, 51]. At the cell type level, we established a species expression hierarchy through clustering, although it remains incomplete. This exploration lays the groundwork for further research in this direction.

Finally, we acquired cell and gene embeddings via CAMEX. By clustering the gene embeddings, we identified specific functional modules within clusters 3 and 5 (**Fig. 3c, d**). Genes in cluster 3 are predominantly expressed in somatic cells. Enrichment analysis revealed associations with the cell surface receptor signaling pathway and extracellular matrix. Somatic cells are all the cells that make up the body of an organism, except for

reproductive cells. In the male reproductive organs, somatic cells play multiple roles such as supporting, protecting, secreting and regulating spermatogenesis [52]. Conversely, genes in cluster 5 are primarily expressed in Spermatocytes, and enrichment analysis linked them to sexual reproduction, meiotic cell cycle process, meiotic cell cycle, meiotic nuclear division, and reproductive process. Meiosis is a well-known process occurring in spermatocytes [53] and is consistent with gene function obtained by CAMEX. In summary, CAMEX can integrate conserved developmental processes across multiple and distant species and provide gene embeddings that shed light on cell functions.

CAMEX aligns various development stages of seven organs across seven different species

The evolution of gene expression during organ development of various species remains largely unknown, and different species exhibit distinct patterns. We applied CAMEX and competing methods to integrate and align the RNA-seq data from seven organs across seven species at various developmental stages [54]. In this unsupervised integration setting, CAMEX does not require any manually annotated cell type labels. We put the bulk data for all developmental stages of all species into CAMEX and other methods to perform the integration task and obtain the embedding together (**Fig. 4a**). Then, we split and visualize the samples of different organs respectively. Remarkably, CAMEX achieved the highest overall score, and scGen, Harmony, and CAMEX can preserve the continuous development trajectory and separate distinct clusters corresponding to different organs (**Fig. 4a**, **Supplementary Figs. S28, S29**, **Supplementary Data22**). We noted that the cerebellum and brain were well integrated due to their abundance of neurons and glial cells [55], while the testis and ovary, being gonadal cells in early development [56], exhibited higher similarity.

While most model organisms display distinct developmental timelines, these are typically mapped to Carnegie stages, spanning from gametogenesis through birth. The Carnegie stages themselves are fundamentally based on observable morphological development during embryogenesis. Crucially, they've only been defined by researchers for a handful of model organisms (e.g., human and mouse) and are limited to the period immediately following pregnancy and childbirth. An exciting challenge lies in exploring whether we can expand the Carnegie developmental staging concept beyond this limited set of organisms to encompass any species, or even to predict any point in a species' lifespan through integrated approaches.

After obtaining organ embeddings from the integration results of CAMEX, we calculated the similarity of these organ embeddings, allowing us to establish correspondence across diverse species. Initially, we selected rats and mice as our subjects and sampled from embryonic day 11 (e11) to postnatal day 112 (p112) for rats and from e10 to p63 for mice. CAMEX aligned the developmental stages of mouse and rat (9/10.5 - 16/17.5) within the Carnegie stages, consistent with existing literature, and even predicted their stages beyond the specified ranges (**Fig. 4b, Supplementary**

Data23). Additionally, we adopted CAMEX to analyze the brain transcriptome data from mouse (Carnegie stage 9–16 days) and opossum (Carnegie stage unknown) at multiple developmental stages. While the mouse brain data spanned from e10 to p63, the opossum brain data covered e13.5 to p180. Notably, the opossum central nervous system, including the brain, is markedly premature at birth (p0), resembling an embryonic e11 mouse, as confirmed by the similarity of development time calculated by CAMEX (**Supplementary Fig. S30**) and a previous study [57]. However, developmental stages across different species in the results are not perfectly aligned. We believe there are several contributing factors. Biological Considerations: From a biological perspective, there are two main reasons. First, the developmental speeds across different species might inherently vary. Second, while we align species based on transcriptome expression, the Carnegie developmental stages are primarily defined by morphological development during embryogenesis, which can lead to discrepancies. Model and Computational Considerations: From a modeling and computational standpoint, the observed differences can also stem from the randomness in the neural network optimization process and the various similarity metrics used for embedding calculations. Despite these points, CAMEX still generally achieves roughly alignment of developmental processes across species. For time points that are not well aligned, this can serve as a potential objective for studying the developmental differences across species.

CAMEX could achieve more accurate integration and annotation performance in both relatives and distant species

Integrating and annotating distant species presents a challenging task due to the tendency to share fewer orthologous genes, resulting in more specific cell-type markers. Here, we collected four scRNA-seq data from four species, i.e., adult human visual cortex, frontal cortex, cerebellum [55], mouse neocortex [58], and lizard and turtle pallium [59], which were profiled from three technology platforms. Each dataset has a specific cell type that is absent in the datasets of other species (**Supplementary Fig. S31a-c**).

First, we applied CAMEX and baseline methods to integrate the scRNA-seq data in the absence of cell type annotations. In this setting, all the methods except Scanorama achieved favorable integration performance, with CAMEX receiving the highest overall score (**Fig. 5a, Supplementary Fig. S31d-e, Supplementary Data24, 25**). From the UMAP plot (**Supplementary Fig. S32**), we can see that all the methods confused the excitatory and inhibitory neurons. CAMEX failed to integrate inhibitory neurons, Harmony and scVI complicated specifically with microglia cells, and scGen failed to mix astrocytes adequately. PyLiger did not properly integrate astrocytes and confused NPCs with excitatory neurons. SATURN misclassified CGCs as NPCs, and scGPT confused ExN with NPCs, in addition to failing to integrate oligodendrocytes.

Traditional cell-type annotation methods heavily depend on clustering and marker genes, posing a significant challenge in non-model species where such prior knowledge is insufficient. To this end, we introduce a GAT classifier after the encoder (Methods). It

alleviates the issue by leveraging many-to-many homology relationships, thereby improving annotation accuracy. In this semi-supervised scenario, the human dataset served as the labeled reference. We then applied CAMEX and other methods to annotate the other three unlabeled target datasets, with a focus on the lizard and turtle species. CAMEX and scANVI demonstrated better performance compared to the other two baseline methods, as evaluated by metrics including ACC, Prec, and F1 score (**Fig. 5b-c, Supplementary Fig. S33a-f**). CAMEX's annotations were accurate for most cell types, including smaller cell subgroups like brain pericytes in turtles (**Fig. 5c-d**). Unknown is the cell type that does not appear in the reference but appears in the query dataset. NPC does not appear in the reference human dataset but does appear in the query turtle and lizard dataset, so we have defined them as unknown. Due to the absence of neural progenitor cells (NPCs) in the human reference dataset, both CAMEX and scANVI cannot identify them. CAMEX misclassified them as excitatory neurons, whereas scANVI identified them as a mixture of CGCs and other cell types. Interestingly, both predictions offer a degree of reasonableness (**Supplementary Fig. S34**).

CAMEX facilitates the discovery of new populations and markers in the primate granular dorsolateral prefrontal cortex (DLPFC) data

The discovery of new cell populations and gene markers holds immense importance in understanding evolutionary organisms and identifying novel targets. The granular dorsolateral prefrontal cortex, unique to primates in evolution, plays a pivotal role in cognitive processes. We investigated specific cell types and gene markers within the collected DLPFC dataset across four species, i.e., humans, rhesuses, marmosets, and chimpanzees [60]. The cell class, subclass, and subtype are manually annotated by clustering and marker genes in the original study [60].

Firstly, we integrated all the cells in the DLPFC dataset among the four species by CAMEX in an unsupervised way. Almost all the cell types match accurately, and the major classes (e.g., excitatory neurons, inhibitory neurons, non-neural cells, and glial cells) of cell types are relatively conserved, indicating the capability of CAMEX for cross-species integration (**Fig. 6a, Supplementary Fig. S35a**).

Then, we simulated the following scenario, assuming we were familiar with one species or model organism. We could adopt its comprehensive reference scRNA-seq atlas to transfer its annotations to other species and reveal several species-specific cell populations relating to species divergence. Considering that during the process of nervous system evolution from a diffuse to a centralized form, glial cells and neurons co-evolved, glial cells, in particular, diversified into a more varied family of cells than the latter [61]. Here, we use human cells as the reference (training set) to annotate cells in other species (test set) in a semi-supervised way by using a GAT classifier (**Methods**). Glial cells have four subclasses: astrocytes, oligodendrocytes, oligodendrocyte precursor cells (OPC), and microglia cells. Early studies indicated that OPC and oligodendrocytes among four species shared all subtypes, and the annotation embedding also confirmed that

(**Supplementary Fig. S35b**). As to the astrocytes, CAMEX did not isolate the *AQP4 OSMR* subtype that exists in humans and chimpanzees (**Supplementary Fig. S35b**). This might be due to the low proportion of this subtype in the training set (<0.25%).

In addition, microglia are more diverse, and those microglial cells are divided into three distinct cell subtypes in the original study, i.e., *APBB1IP*, *CCL3*, and *GLDN* clusters. Specifically, *CCL3* is exclusive to humans, *GLDN* is found in humans and chimpanzees, and *APBB1IP* appeared across four species (**Fig. 6a left**). Leveraging CAMEX with a GAT classifier, we mapped the other three species query into the human reference dataset, resulting in the distinct isolation of three subtypes (**Fig. 6b, Supplementary Fig. S36a-b**). At the same time, we investigated the expression of *GLDN*, *CCL3*, and *APBB1IP* in the original manual annotation and the CAMEX prediction (**Supplementary Fig. S36c**). We found that the prediction of CAMEX has certain rationality, for instance, the expression of *CCL3* in the subgroups of chimpanzees. However, there are also false positives, such as *CCL3* and *GLDN* in marmosets. Overall, we therefore wish to clarify that CAMEX should be considered a bioinformatics tool for generating hypotheses about potential cell subpopulations. These computational predictions need further validation through subsequent wet-lab experiments.

Moreover, we conducted differential analysis for each subtype, comparing them with other clusters to identify subtype-specific genes (**Supplementary Fig. S37**). The differential genes associated with the *CCL3* subtype primarily function in immunity. For instance, *CD83* encodes a protein known as CD83, which stabilizes MHC II, costimulatory molecules, and CD28 in the cell membrane by counteracting E3 ubiquitin ligases [62, 63]. On the other hand, the differential genes within the *GLDN* subtype are primarily involved in signal transduction. SPP1 (secreted phosphoprotein 1) is an extracellular protein closely linked to tumor biology, including processes like proliferation, migration, and invasion [64, 65].

In addition, we analyzed species differences. Even within the *APBB1IP* subtype, we observed abundant differential genes between species (**Fig. 6c**). *NAV2* and *NAV3* belong to the neuronal navigator family and are predominantly expressed in the nervous system. They likely play roles in axon guidance, cell migration, and neurodevelopment [66]. PAM is involved in the amidation process of peptide hormones and neurotransmitters and is crucial for regulating their activity and stability [67]. Overall, the human and chimpanzee transcriptomes exhibit a high consistency at the single-cell resolution and can serve as model organisms for exploring disease mechanisms or drug development. However, focusing on more species-specific genes is essential when studying marmoset and rhesus.

Discussion

The scRNA-seq data are becoming increasingly popular, encompassing a wider range of species. Comprehensive cross-species comparisons and analyses at the single-cell resolution can significantly enhance our understanding of cellular origins and evolutionary

mechanisms. By exploring conserved and specific features of cell states, one can gain insights into fundamental evolutionary processes and how cellular diversity correlates with physiological connections across species. For instance, after removing the batch effect, we can align the Carnegie stages of different species and discover new cell types and subtypes. Consequently, there is a pressing demand for robust tools to integrate multi-species datasets.

While previous approaches have made significant efforts toward cross-species integration and annotation, their widespread applicability to current multi-species data remains limited due to various constraints. It's worth mentioning that cross-species single-cell integration remains a relatively nascent field with few established methods. For instance, integration algorithms designed for same-species data often focus solely on one-to-one homologous genes when applied to multi-species datasets. Unfortunately, these algorithms falter when dealing with remote species due to the distinct number of gene copies involved in evolutionary processes [68]. Another type of algorithm constructs correlations between diverse species based on gene sets (pathway and module) [21]. However, the availability of prior knowledge regarding gene sets is often restricted to model organisms, rendering this algorithm ineffective when applied to non-model organisms. To our knowledge, SAMap pioneered this area, but its reliance on sequence alignment leads to significant computational time consumption. Furthermore, SAMap's CPU memory usage increases sharply with a large number of species, posing a practical limitation.

SATURN, while an advanced algorithm that combines protein language models with transcriptional expression, requires cell type annotation as input. This prerequisite significantly limits its broad application. Our previous method, CAME, is a semi-supervised model that relies on reference cell labels for annotation and cannot perform integration tasks.

Here, we introduce CAMEX for integrating, aligning, and annotating multi-species scRNA-seq data using heterogeneous graph neural networks. We treat the dataset for each species as a bipartite graph and connect these bipartite graphs through many-to-many homology relationships. This design allows us to leverage homologous relationships as much as possible. For integration tasks, CAMEX employs an unsupervised encoder-decoder architecture. For annotation tasks, it utilizes a semi-supervised encoder-classifier architecture. We demonstrate the effectiveness of CAMEX on multiple single-cell datasets from different organs across various species.

Additionally, we integrate sperm development data from up to 11 species and validate somatic and spermatocyte functions through gene embedding. We align multi-species data across different developmental stages, which extends beyond the Carnegie Stage. The cell embeddings obtained by CAMEX facilitate the discovery of species-specific cell subtypes and novel marker genes. CAMEX also serves as a robust cell annotation tool, showcasing competitive performance compared to baseline methods across multiple datasets. In short, we expect CAMEX to become a significant tool for understanding the

conservation and diversification of cell types across species, revealing fundamental evolutionary processes at the single-cell transcriptomic resolution.

While CAMEX demonstrates significant utility, it also exhibits some limitations that warrant improvement. First, the similarity of homologous genes can be obtained through sequence alignment. However, other methods for calculating similarity have not been considered in CAMEX. Fine-tuning this aspect could enhance its performance. Second, despite incorporating many-to-many homologous relationships into the graph structure, CAMEX still relies on one-to-one homologous genes during initializing cell node features. This limitation impacts its implementation and performance when integrating data across highly diverse species. Third, CAMEX leverages cell and gene embeddings for interpretability. However, the interpretability analysis based on graph neural networks remains an ongoing challenge. Exploring computational methods to elucidate biological phenomena, such as inferring gene regulation networks and considering subgraph structures as potential pathways, could yield unexpected insights. Fourthly, Single-cell RNA sequencing studies in plants face greater challenges than those in animals due to gene family expansions from whole-genome duplications, which complicate cross-species integration [69]. In addition, recent studies have extended cross-species integration models to encompass the construction of developmental evolutionary trees and perturbation predictions [70], representing highly promising applications in both biological and pharmaceutical research. It is anticipated that CAMEX will incorporate additional components in the future to better adapt to emerging tasks. Lastly, as technology evolves, we envision CAMEX broadening its scope to cover more cross-species data [71] and to encompass multi-omics and spatial transcriptomics. This approach would provide a more comprehensive map for species origin and evolution by leveraging multi-dimensional information.

Methods

CAMEX model

Heterogeneous graph of cells and genes. A scRNA-seq dataset comprising N cells and M genes is represented as a matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)^T \in R^{N \times M}$. Following a standard preprocessing pipeline (normalization, log-transformation, and selection of highly variable genes, clustering, and obtaining differential genes), we construct a cell-gene bipartite graph $\boldsymbol{A}$. In this graph, an edge connects cell n and gene m if and only if cell n expresses gene m. Subsequently, cell-cell connections are established by applying the kNN algorithm within the PCA space derived from this processed matrix. In addition, we added a self-loop to each cell and gene, respectively. Then, we connected genes in each $\boldsymbol{A}$ from different species through the many-to-many-homologous gene relationships. In short, there are two types of nodes including cells and genes, and six types of edges (relations) including ‘a cell expresses a gene’, ‘a gene is expressed by a cell’, ‘two cells that are kNN neighbors’, ‘self-loop of a cell’, ‘self-loop of a gene’ and ‘two

genes that are orthologous ones' in the heterogeneous graph. We initialized cell node features with preprocessed expression data (normalized, log-transformed values from the top 2,000 highly variable genes (HVGs)), while gene nodes were initialized as zero vectors. The resulting heterogeneous graph incorporated both cell and gene nodes with their respective features.

Integration task:

Heterogeneous graph neural network encoder. CAMEX adopts a heterogeneous graph neural network as the encoder inspired by the relational graph convolutional network [72]. We design six independent parameter matrices for each of the above relationships as follows. $\boldsymbol{W}_{cg}$ (a cell expresses a gene), $\boldsymbol{W}_{gc}$ (a gene is expressed by a cell), $\boldsymbol{W}_{cc}$ (two cells that are kNN neighbors), $\boldsymbol{W}_{gg}$ (two genes that are orthologous ones), $\boldsymbol{W}_{c}$ (self-loop of a cell), and $\boldsymbol{W}_{g}$ (self-loop of a gene). The encoder transforms the features of the neighbors in a non-linear way and converges to the center node according to the message-passing paradigm. The $(l+1)$-th layer representation of cell $\boldsymbol{h}_{ni}^{(l+1)}$ and gene $\boldsymbol{h}_{mi}^{(l+1)}$ can be defined as follows:

$$\boldsymbol{h}_{ni}^{(l+1)} = \sigma\left(\sum_{r\in \mathrm{R}}\left(\sum_{j\in N_i^r}\frac{1}{c_{i,r}}\boldsymbol{W}_r^{(l)}\boldsymbol{h}_j^{(l)} + \boldsymbol{W}_c^{(l)}\boldsymbol{h}_{ni}^{(l)}\right)\right) \quad (1)$$

$$\boldsymbol{h}_{mi}^{(l+1)} = \sigma\left(\sum_{r\in \mathrm{R}}\left(\sum_{j\in \mathrm{N}_i^r}\frac{1}{c_{i,r}}\boldsymbol{W}_r^{(l)}\boldsymbol{h}_j^{(l)} + \boldsymbol{W}_g^{(l)}\boldsymbol{h}_{mi}^{(l)}\right)\right) \quad (2)$$

where $\boldsymbol{h}_{ni}^{(l)}$ and $\boldsymbol{h}_{mi}^{(l)}$ represent the l-th layer representation of cells and genes, respectively, $\boldsymbol{W}_c^{(l)}$ and $\boldsymbol{W}_g^{(l)}$ represent the l-th layer weight matrices of 'self-loop of a cell' and 'self-loop of a gene', respectively, N_i^r denotes the set of neighbor indices of the node (cell or gene) i under relations $r \in \mathrm{R}$, $c_{i,r}$ is a normalization factors, $\boldsymbol{W}_r^{(l)}$ is the weight matrix of the l-th layer transformation of the relationship r, $\boldsymbol{h}_j^{(l)}$ is the l-th layer representation of neighbor *j*, and σ is the activation function (i.e., LeakyReLU here).

To reduce the number of parameters to prevent overfitting, we add a hyperparameter 'share' that can share weight matrices of the same relationship in different layers such as $\boldsymbol{W}_{cc}^{(l)} = \boldsymbol{W}_{cc}^{(l+1)}$. After the message passing through several layers of neural networks, the cell and gene representation can be mapped into the same embedding space.

Edge and feature reconstruction decoder. When integrating different scRNA-seq data in an unsupervised manner, we employ two decoders to reconstruct edges and features, respectively. Specifically, the dot-product decoder aims to reconstruct the edges of the input heterogeneous graph. Edge predictions in the heterogeneous graph are generated

using the dot product of cell and gene embeddings:

$$\boldsymbol{A}' = \boldsymbol{H}_N \cdot \boldsymbol{H}_M \quad (3)$$

where $\boldsymbol{H}_N$ and $\boldsymbol{H}_M$ are the cell and gene embeddings, respectively, and $\boldsymbol{A}'$ is the reconstructed adjacency matrix.

We measure the quality of each edge in the adjacency matrix using cross-entropy loss:

$$L_A = -\Sigma_{i=1}^{N}\Sigma_{j=1}^{M}[\boldsymbol{A}_{ij} log\widehat{\boldsymbol{A}}_{ij} + (1-\boldsymbol{A}_{ij}) log(1-\widehat{\boldsymbol{A}}_{ij})] \quad (4)$$

where $\boldsymbol{A}_{ij}$ is the edge between the *i*-th cell and the *j*-th gene in the original graph.

The feature reconstruction decoder aims to reconstruct cell features through a fully connected layer:

$$\boldsymbol{h}_0' = \boldsymbol{W}_{fc}(\boldsymbol{h}_{ni}^{(l)}) \quad (5)$$

where $\boldsymbol{W}_{fc}$ is the weight of a fully connected neural network and $\boldsymbol{h}_0'$ is the reconstructed cell feature.

We utilize the mean squared error to measure the distance between the original $\boldsymbol{h}_0$ and the predicted $\boldsymbol{h}_0'$:

$$L_F = \frac{1}{N}\Sigma_{i=1}^{N}(\boldsymbol{h}_0 - \boldsymbol{h}_0') \quad (6)$$

Loss function and optimizer in the integration task. The unsupervised integration task loss function can be defined as follows:

$$L_{intergration} = L_A + L_F \quad (7)$$

where L_A and L_F are optimized in parallel.

The integration task utilizes a graph attention network autoencoder with an encoder-decoder architecture, which is trained end-to-end through backpropagation [73]. The model optimization employs the Adam optimizer [74], a variant of the Stochastic Gradient Descent algorithm, for parameter optimization and is implemented in PyTorch [75] using a mini-batch training strategy. We use an NVIDIA GPU laptop RTX 3070 to accelerate the training process. By default, both the encoder and decoder architectures contain two bottleneck layers. We applied a hyperparameter 'share' to allow these layers to share parameters, respectively, thereby reducing GPU memory consumption.

Annotation task:

Graph attention classifier. Cell labels are not mandatory during the integration task. However, during annotation, it is essential to have one or more reference datasets with pre-defined manual annotations. Consequently, the model could be trained in a semi-supervised manner. The cell annotations provided by the reference dataset can be denoted as: $\boldsymbol{Y} = (y_1, y_2, \ldots, y_{Nr}) \in R^C$, where Nr represents the number of cells in the reference dataset and C represents the number of cell types. When performing cell annotation tasks, cell embeddings can be input into a GAT classifier. Specifically, the attention coefficients are first computed as follows:

$$e_{ij} = a\left([\boldsymbol{W}\boldsymbol{h}_i^{(l)} || \boldsymbol{W}\boldsymbol{h}_j^{(l)}]\right) \tag{8}$$

where $\boldsymbol{h}_i^{(l)}$ and $\boldsymbol{h}_j^{(l)}$ is the l-th layer representations of cells *i* and *j*.

Shared parameter $\boldsymbol{W}$ maps cell embedding to a common space. The concatenation operation (denoted by ||) is applied. Finally, the function a maps the concatenated result to a scalar representing the relevance between nodes *i* and *j*. Subsequently, the relevance scores are normalized to obtain attention coefficients:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}(e_{ij})\right)}{\Sigma_{k \in N_i} \exp\left(\text{LeakyReLU}(e_{ij})\right)} \tag{9}$$

We use the multi-head self-attention mechanism to predict the final results:

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} \hat{y}^k = \frac{1}{K}\sum_{k=1}^{K} \sum_{j \in N_i^r} \alpha_{ij}\, \boldsymbol{W}_r^{(l)} \boldsymbol{h}_{ni}^{(l)} \tag{10}$$

where K is the total number of attention-heads, set as eight by default.

Loss function and optimizer in the annotation task. The semi-supervised annotation task loss function is a cross-entropy loss function that can be defined as follows:

$$L_{annotation} = -\frac{1}{N}\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c}\, log\left(\hat{y}_{i,c}\right) \tag{11}$$

where N represents sample number, C represents class number, $y_{i,c}$ represents the true label of sample i on category c, $\hat{y}_{i,c}$ represents the model predicts the probability that sample i belongs to category c. The optimizer, and implementation details are the same as the integration task in 'Loss function and optimizer in the integration task'.

In a word, integration and annotation tasks share the same encoder architecture. During integration, unsupervised training is achieved without requiring cell type labels, while annotation necessitates a reference dataset with cell type labels for semi-supervised training.

Training details.

(1) Graph sampling: For large-scale data with a large number of cells or involving more species, we partition the heterogeneous graph into multiple subgraphs and train the graph neural network using a mini-batch approach. We implement this using the Deep Graph Library (DGL) [76].

(2) Checkpoint selection: In deep learning model training, checkpoints are commonly employed to prevent overfitting. However, in the context of unsupervised integration and semi-supervised annotation tasks, these checkpoints can sometimes lead to over-correction or over-fitting. To address this, we introduce a novel metric for evaluating checkpoints. Specifically, we cluster cells based on their embeddings and assess whether to halt training or revert to a checkpoint using the clustering metric ARI.

(3) Graph storage: To efficiently store large-scale graphs, we implemented directed

heterogeneous attribute graphs and graph neural networks using the framework DGL library [76]. The DGL utilizes the coordinate list matrix (COO) matrix for sparse data representation. Its storage logic follows a directed graph structure, mapping relationships from source nodes to target nodes. To maintain consistency with the COO matrix format, we save six kinds of edges independently and opted not to merge the W_{gc} and W_{cg}.

(4) Hyperparameters:

CAMEX employs a set of default hyperparameters, including: train_mode: 'mini_batch', dim_hidden: 128, batch_size: 1024, gnn_layer_num: 2, epoch_integration: 10, epoch_annotation: 10, and encoder: 'GCN'. All experiments except for bulk RNA-seq integration used these defaults. For bulk RNA-seq data, which has a small sample size, we recommend setting the train_mode to ‘full_batch’ and setting epoch_integration to 200. Further details are provided in the GitHub tutorials.

Baseline methods

Integration methods:

Harmony: Harmony [19] has demonstrated superior performance in various benchmarking studies for scRNA-seq data integration. We applied the Python package harmony (v0.0.9) in this study.

scVI: scVI [77] is an extensible Python toolkit with diverse applications. We applied the Python package scvi-tools (v0.14.6) in this study.

pyLiger: LIGER [26] is a robust online integration model. It has recently been re-implemented as PyLiger [25] in Python with improved computational efficiency. We applied the Python package PyLiger (v0.1.3) in this study.

Scanorama: Scanorama [27] is a sensitive algorithm that identifies and merges shared cell types between all datasets and accurately integrates heterogeneous scRNA-seq datasets. We applied the Python package Scanorama (v1.7.3) in this study.

Seurat V4: Seurat V4 [29] is one of the most popular single-cell analysis tools based on the R language. We used Seurat V4.2.3 in this study.

scGen: scGen [28] is a VAE-based method that can predict the gene expression in different perturbations of single-cell data. We use version (2.0.0) in this study.

SAMap: SAMap [22] is an algorithm that integrates cross-species single-cell transcriptomic data by utilizing the similarity of gene or protein sequences. We use version (1.0.15) in this study.

SATURN: SATURN [30] is a deep learning method for obtaining universal cell embeddings from different species. We download it from GitHub (https://github.com/snap-stanford/saturn) and use it following the tutorials (https://github.com/snap-stanford/SATURN/tree/main/Vignettes/frog_zebrafish_embryogenesis).

scGPT: scGPT [19] is a single-cell foundation model trained based on human and mouse scRNA-seq data. We use the latest version (0.2.4) in this study.

Annotation methods:

scANVI: scANVI [34] is a probabilistic approach to annotate cells in a semi-supervised

way. We applied the Python package scvi-tools (v0.14.6).

CellBlast: CellBlast [78] is an easy-to-use Python package for cell-type annotation. We applied the Python package CellBlast (v0.5.0) in this study.

SingleCellNet: SingleCellNet [18] is a tree model-based R package that can annotate single-cell datasets including cross-species RNA-seq data, and the authors recently migrated this approach into Python (https://github.com/pcahan1/PySingleCellNet/).

It is crucial to note that the benchmarking strategy is task-dependent. For the unsupervised integration and alignment task, CAMEX is compared against label-free methods such as Harmony. For the semi-supervised cell-type annotation task, however, the comparison is made against reference-based methods like SingleCellNet (SCN), which require at least one labeled reference dataset. In addition, we use default hyperparameters in all experiments for all methods.

Evaluation Metrics

Integration metrics. We group the batch correction metrics into two categories: 1) batch correction and 2) conservation of biological variance. Specifically, the batch correction score using metrics including kBET (k-nearest-neighbor Batch Effect Test) [31], PCR batch (Principal Component Regression of the batch covariate) [31], ASW batch (Average Silhouette Width across batch) [31], and Graph connectivity [14]. The conservation of biological variation score using metrics including ARI (Adjusted Rand Index) [32], NMI (Normalized Mutual Information) [33], Cell type ASW [14], ISO label F1 (Isolated label F1) [14], and ISO label silhouette (Isolated label silhouette) [14]. We implemented all the above integration metrics by scIB [14].

kBET: The kBET algorithm assesses whether the label composition of a cell's k nearest neighborhood is consistent with the expected label distribution. This evaluation is performed by repeating the test on a random subset of cells, and the outcomes are aggregated as a rejection rate across all tested neighborhoods.

PCR batch: PCR batch quantifies the batch effect removal by comparing the variance contributions of the batch effects of datasets before integration (VC_{before}) and after integration (VC_{after}), respectively.

ASW Batch: ASW Batch calculates the silhouette width of cells with respect to batch labels.

Graph connectivity: Graph connectivity assesses whether the graph correctly connects cells of the same cell-type labels among batches.

ARI: The adjusted Rand index (ARI) is a common external evaluation metric for clustering. It assesses the effectiveness of clustering by calculating the number of sample pairs assigned to the same or different clusters in both the true labels and the clustering results.

NMI: NMI computes normalized mutual information between two clustering results.

Cell type ASW: Cell-type ASW evaluates ASW with respect to cell-type labels, where a higher score means that cells are closer to cells of the same cell type.

Isolated label F1: Isolated label F1 is developed to measure the ability of integration

methods to preserve dataset-specific cell types. We adopted the Scanpy pipeline to cluster cells and evaluated the cluster assignment of dataset-specific cell types based on the F1 score.

Isolated label silhouette: Isolated label silhouette measures the conservation of dataset-specific cell types. It evaluates the ASW of dataset-specific cell types.

Annotation metrics: To comprehensively compare the performance of different annotation methods, we employed three kinds of evaluation metrics, including accuracy, precision, and F1 score. Cell annotation is a multi-classification task. Considering the difference in the number of different types of cells in a dataset, all evaluation metrics are weighted. We implemented all the above annotation metrics by Scikit-learn [33].

Accuracy: The accuracy can be defined as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN},$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Precision: The precision can be defined as follows:

$$precision = \frac{TP}{TP+FP},$$

where TP is the number of true positives and FP is the number of false positives.

F1 score: The F1 score can be interpreted as a harmonic mean of precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The F1 can be defined as follows:

$$F1 = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})},$$

Batch mean = (kBET + PCR batch + ASW Batch + Graph connectivity) / 4), Biological score (biological mean = (ARI + NMI + Cell type ASW + Isolated label F1 + Isloated label silhouette) / 5),

Overall score = (batch mean + biological mean) / 2.

Trajectory Conservation Analysis:

Inspired by previous literature [20, 48], we applied and adjusted a quantitative metric to evaluate trajectory preservation during integration as Luecken et al. [14]. This metric is calculated in the following steps:

Pre-integration Pseudotime Calculation: For each unintegrated (raw) dataset, we first exclude cell types outside our focus. For example, in the testis dataset across 11 species, we excluded somatic and Sertoli cells. We then designate a spermatogonia cell as the trajectory origin ("iroot") and compute pseudotime values (t_unintegrate) using the scanpy.tl.dpt() function.

For bulk RNA-seq datasets with known developmental timepoints, we normalize the ground truth temporal values by linearly scaling them to the continuous interval [0,1],

where 0 and 1 represent the start and end points of the developmental process, respectively. This normalization allows for direct comparison with pseudotime estimates while preserving relative temporal relationships.

Post-integration Pseudotime Analysis: In the integrated dataset (e.g., from CAMEX or Seurat), we select a human spermatogonia cell as the trajectory origin ("iroot") within the testis dataset. We then compute pseudotime values (t_integrate) using the scanpy.tl.dpt() function.

Trajectory Conservation Metric: We quantify trajectory preservation using correlation analysis between the pre- and post-integration pseudotime values. Spearman's rank correlation is used for this quantification.

One-to-one and many-to-many homologous relationships

For the baseline methods, we map the gene symbols to the first or reference species dataset using one-to-one homologous relationships. The many-to-many homology relationships were obtained from the Ensembl database [79]. The details of how to use it can be found in the GitHub tutorials.

Carnegie stages

Carnegie stages represent a standardized system used to define the 23 stages of embryonic development in vertebrates. The data on ‘the number of days’ in Carnegie stages refers to the days following embryo formation and should be considered an approximate value. These stages track development from the formation of a fertilized egg to an embryo's progress over approximately 60 days (8 weeks). For instance, a human embryo takes around 60 days to reach Carnegie Stage 23, whereas a chicken embryo completes the same stage in just 10 days. This system enables scientists to compare developmental characteristics across species at equivalent stages, thereby enhancing our understanding of the universal principles governing embryonic development.

Data Availability

The details of all datasets can be found in **Supplementary Data26**. We preprocessed all raw data following the pipeline by Scanpy [80] and upload the processed data in h5ad format to Google Driver. This dataset is freely accessible without requiring a password:
https://drive.google.com/drive/folders/1rwdjEvWFEFw82a0x2JzMi2jXICbUc5eb?usp=sharing
and the dataset can also be available at:
https://figshare.com/articles/dataset/Dataset_for_CAMEX/31131808.
Source data are provided with this paper.

Code Availability

The source codes of CAMEX package, along with code and detailed tutorials for reproducibility, is available at https://github.com/zhanglabtools/CAMEX/ under MIT license, and in Zenodo (https://zenodo.org/records/17991379) [81].

Reference

[1] M. D. Luecken and F. J. Theis, "Current best practices in single‐cell RNA‐seq analysis: a tutorial," Molecular systems biology, vol. 15, no. 6, p. e8746, 2019.

[2] W. Chen et al., "A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples," Nature Biotechnology, vol. 39, no. 9, pp. 1103-1114, 2021.

[3] O. Rozenblatt-Rosen, J. W. Shin, J. E. Rood, A. Hupalowska, A. Regev, and H. Heyn, "Building a high-quality human cell atlas," Nature Biotechnology, vol. 39, no. 2, pp. 149-153, 2021.

[4] L. Han et al., "Cell transcriptomic atlas of the non-human primate Macaca fascicularis," Nature, vol. 604, no. 7907, pp. 723-731, 2022.

[5] T. Iram and T. M. Consortium, "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris," Nature, vol. 562, no. 7727, pp. 367-372, 2018.

[6] J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier, "Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis," Science, vol. 360, no. 6392, p. eaar3131, 2018.

[7] A. Tanay and A. Sebé-Pedrós, "Evolutionary cell type mapping with single-cell genomics," Trends in genetics, vol. 37, no. 10, pp. 919-932, 2021.

[8] X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, and C. Trapnell, "Single-cell mRNA quantification and differential analysis with Census," Nature methods, vol. 14, no. 3, pp. 309-315, 2017.

[9] S. S. Potter, "Single-cell RNA sequencing for the study of development, physiology and disease," Nature Reviews Nephrology, vol. 14, no. 8, pp. 479-492, 2018.

[10] E. V. Koonin, M. Y. Galperin, E. V. Koonin, and M. Y. Galperin, "Comparative genomics and new evolutionary biology," Sequence—Evolution—Function: Computational Approaches in Comparative Genomics, pp. 227-294, 2003.

[11] M. E. Shafer, "Cross-species analysis of single-cell transcriptomic data," Frontiers in cell and developmental biology, vol. 7, p. 175, 2019.

[12] J. Wang et al., "Tracing cell-type evolution by cross-species comparison of cell atlases," Cell Reports, vol. 34, no. 9, 2021.

[13] R. Wang et al., "Construction of a cross-species cell landscape at single-cell level," Nucleic Acids Research, vol. 51, no. 2, pp. 501-516, 2023.

[14] M. D. Luecken et al., "Benchmarking atlas-level data integration in single-cell genomics," Nature methods, vol. 19, no. 1, pp. 41-50, 2022.

[15] H. T. N. Tran et al., "A benchmark of batch-effect correction methods for single-cell RNA sequencing data," Genome biology, vol. 21, pp. 1-32, 2020.

[16] I. Korsunsky et al., "Fast, sensitive and accurate integration of single-cell data with Harmony," Nature methods, vol. 16, no. 12, pp. 1289-1296, 2019.

[17] G. P. Way and C. S. Greene, "Bayesian deep learning for single-cell analysis," Nature methods, vol. 15, no. 12, pp. 1009-1010, 2018.

[18] Y. Tan and P. Cahan, "SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species," Cell systems, vol. 9, no. 2, pp. 207-213. e2,

2019.

[19] H. Cui et al., "scGPT: toward building a foundation model for single-cell multi-omics using generative AI," Nature Methods, vol. 21, no. 8, pp. 1470-1480, 2024.

[20] Y. Song, Z. Miao, A. Brazma, and I. Papatheodorou, "Benchmarking strategies for cross-species integration of single-cell RNA sequencing data," Nature Communications, vol. 14, no. 1, p. 6495, 2023.

[21] H. Ding, A. Blair, Y. Yang, and J. M. Stuart, "Biological process activity transformation of single cell gene expression for cross-species alignment," Nature communications, vol. 10, no. 1, p. 4899, 2019.

[22] A. J. Tarashansky et al., "Mapping single-cell atlases throughout Metazoa unravels cell type evolution," Elife, vol. 10, p. e66747, 2021.

[23] Y. Rosen, M. Brbić, Y. Roohani, K. Swanson, Z. Li, and J. Leskovec, "Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN," Nature Methods, pp. 1-9, 2024.

[24] X. Liu, Q. Shen, and S. Zhang, "Cross-species cell-type assignment from single-cell RNA-seq data by a heterogeneous graph neural network," Genome Research, vol. 33, no. 1, pp. 96-111, 2023.

[25] L. Lu and J. D. Welch, "PyLiger: scalable single-cell multi-omic data integration in Python," Bioinformatics, vol. 38, no. 10, pp. 2946-2948, 2022.

[26] J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko, "Single-cell multi-omic integration compares and contrasts features of brain cell identity," Cell, vol. 177, no. 7, pp. 1873-1887. e17, 2019.

[27] B. Hie, B. Bryson, and B. Berger, "Efficient integration of heterogeneous single-cell transcriptomes using Scanorama," Nature biotechnology, vol. 37, no. 6, pp. 685-691, 2019.

[28] M. Lotfollahi, F. A. Wolf, and F. J. Theis, "scGen predicts single-cell perturbation responses," Nature methods, vol. 16, no. 8, pp. 715-721, 2019.

[29] Y. Hao et al., "Integrated analysis of multimodal single-cell data," Cell, vol. 184, no. 13, pp. 3573-3587. e29, 2021.

[30] Y. Rosen, M. Brbić, Y. Roohani, K. Swanson, Z. Li, and J. Leskovec, "Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN," Nature Methods, vol. 21, no. 8, pp. 1492-1500, 2024.

[31] M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis, "A test metric for assessing single-cell RNA-seq batch correction," Nature methods, vol. 16, no. 1, pp. 43-49, 2019.

[32] L. Hubert and P. Arabie, "Comparing partitions," Journal of classification, vol. 2, pp. 193-218, 1985.

[33] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol. 12, pp. 2825-2830, 2011.

[34] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, "Probabilistic harmonization and annotation of single‐cell transcriptomics data with deep generative models," Molecular systems biology, vol. 17, no. 1, p. e9620, 2021.

[35] Z.-J. Cao, L. Wei, S. Lu, D.-C. Yang, and G. Gao, "Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST," Nature communications, vol. 11, no. 1, p. 3458, 2020.

[36] M. Guilliams et al., "Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches," Cell, vol. 185, no. 2, pp. 379-396. e38, 2022.

[37] X. Fan et al., "Single-cell reconstruction of follicular remodeling in the human adult ovary," Nature communications, vol. 10, no. 1, p. 3164, 2019.

[38] M. J. Muraro et al., "A single-cell transcriptome atlas of the human pancreas," Cell systems, vol. 3, no. 4, pp. 385-394. e3, 2016.

[39] M. Baron et al., "A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure," Cell systems, vol. 3, no. 4, pp. 346-360. e4, 2016.

[40] Y. Xin et al., "RNA sequencing of single human islet cells reveals type 2 diabetes genes," Cell metabolism, vol. 24, no. 4, pp. 608-615, 2016.

[41] Å. Segerstolpe et al., "Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes," Cell metabolism, vol. 24, no. 4, pp. 593-607, 2016.

[42] N. Lawlor et al., "Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes," Genome research, vol. 27, no. 2, pp. 208-222, 2017.

[43] S. Wang et al., "Single-cell transcriptomic atlas of primate ovarian aging," Cell, vol. 180, no. 3, pp. 585-600. e19, 2020.

[44] X. Han et al., "Mapping the mouse cell atlas by microwell-seq," Cell, vol. 172, no. 5, pp. 1091-1107. e17, 2018.

[45] J. Li et al., "A single-cell transcriptomic atlas of primate pancreatic islet aging," National science review, vol. 8, no. 2, p. nwaa127, 2021.

[46] M. Soumillon et al., "Cellular source and mechanisms of high transcriptome complexity in the mammalian testis," Cell reports, vol. 3, no. 6, pp. 2179-2190, 2013.

[47] F. Murat et al., "The molecular evolution of spermatogenesis across mammals," Nature, vol. 613, no. 7943, pp. 308-316, 2023.

[48] H. Zhong et al., "Benchmarking cross-species single-cell RNA-seq data integration methods: towards a cell type tree of life," Nucleic Acids Research, vol. 53, no. 1, p. gkae1316, 2025.

[49] P. Acién and M. Acién, "Disorders of sex development: classification, review, and impact on fertility," Journal of clinical medicine, vol. 9, no. 11, p. 3555, 2020.

[50] L. Fagerberg et al., "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics," Molecular & cellular proteomics, vol. 13, no. 2, pp. 397-406, 2014.

[51] J. A. MacLean II and M. F. Wilkinson, "Gene regulation in spermatogenesis," Current topics in developmental biology, vol. 71, pp. 131-197, 2005.

[52] L. O'Donnell, P. Stanton, and D. M. de Kretser, "Endocrinology of the male reproductive system and spermatogenesis," 2015.

[53] M. D. Griswold, "Spermatogenesis: the commitment to meiosis," Physiological reviews, vol. 96, no. 1, pp. 1-17, 2016.
[54] M. Cardoso-Moreira et al., "Gene expression across mammalian organ development," Nature, vol. 571, no. 7766, pp. 505-509, 2019.
[55] B. B. Lake et al., "Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain," Nature biotechnology, vol. 36, no. 1, pp. 70-80, 2018.
[56] L. Garcia-Alonso et al., "Single-cell roadmap of human gonadal development," Nature, vol. 607, no. 7919, pp. 540-547, 2022.
[57] K. K. Smith, "Early development of the neural plate, neural crest and facial region of marsupials," The Journal of Anatomy, vol. 199, no. 1-2, pp. 121-131, 2001.
[58] B. Tasic et al., "Shared and distinct transcriptomic cell types across neocortical areas," Nature, vol. 563, no. 7729, pp. 72-78, 2018.
[59] M. A. Tosches, T. M. Yamawaki, R. K. Naumann, A. A. Jacobi, G. Tushev, and G. Laurent, "Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles," Science, vol. 360, no. 6391, pp. 881-888, 2018.
[60] S. Ma et al., "Molecular and cellular evolution of the primate dorsolateral prefrontal cortex," Science, vol. 377, no. 6614, p. eabo7257, 2022.
[61] A. Verkhratsky, M. S. Ho, and V. Parpura, "Evolution of neuroglia," Neuroglia in Neurodegenerative Diseases, pp. 15-44, 2019.
[62] Z. Li et al., "CD83: activation marker for antigen presenting cells and its therapeutic potential," Frontiers in immunology, vol. 10, p. 460131, 2019.
[63] L. Grosche et al., "The CD83 molecule–an important immune checkpoint," Frontiers in Immunology, vol. 11, p. 721, 2020.
[64] K. Zhao and Z. Ma, "Comprehensive analysis to identify SPP1 as a prognostic biomarker in cervical cancer," Frontiers in genetics, vol. 12, p. 732822, 2022.
[65] X. Yi et al., "SPP1 facilitates cell migration and invasion by targeting COL11A1 in lung adenocarcinoma," Cancer Cell International, vol. 22, no. 1, p. 324, 2022.
[66] C. Klein et al., "Neuron navigator 3a regulates liver organogenesis during zebrafish embryogenesis," Development, vol. 138, no. 10, pp. 1935-1945, 2011.
[67] M. Satani, K. Takahashi, H. Sakamoto, S. Harada, Y. Kaida, and M. Noguchi, "Expression and characterization of human bifunctional peptidylglycine α-amidating monooxygenase," Protein expression and purification, vol. 28, no. 2, pp. 293-302, 2003.
[68] H. Innan and F. Kondrashov, "The evolution of gene duplications: classifying and distinguishing between models," Nature Reviews Genetics, vol. 11, no. 2, pp. 97-108, 2010.
[69] M. J. Passalacqua and J. Gillis, "Coexpression enhances cross-species integration of single-cell RNA sequencing across diverse plant species," Nature Plants, vol. 10, no. 7, pp. 1075-1080, 2024.
[70] H. Zhong et al., "Unify: Learning Cellular Evolution with Universal Multimodal Embeddings," bioRxiv, p. 2025.09. 07.674681, 2025.
[71] P. Wang et al., "scCompass: An integrated cross-species scRNA-seq database for AI-

ready," bioRxiv, p. 2024.11. 12.623138, 2024.

[72] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15, 2018: Springer, pp. 593-607.

[73] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," nature, vol. 323, no. 6088, pp. 533-536, 1986.

[74] K. Diederik, "Adam: A method for stochastic optimization," (No Title), 2014.

[75] A. Paszke et al., "Automatic differentiation in pytorch," 2017.

[76] M. Y. Wang, "Deep graph library: Towards efficient and scalable deep learning on graphs," in ICLR workshop on representation learning on graphs and manifolds, 2019.

[77] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," Nature methods, vol. 15, no. 12, pp. 1053-1058, 2018.

[78] Z.-J. Cao, L. Wei, S. Lu, D.-C. Yang, and G. Gao, "Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST," Nature communications, vol. 11, no. 1, pp. 1-13, 2020.

[79] K. L. Howe et al., "Ensembl 2021," Nucleic acids research, vol. 49, no. D1, pp. D884-D891, 2021.

[80] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," Genome biology, vol. 19, pp. 1-5, 2018.

[81] Zhen-Hao Guo. (2025). zhanglabtools/CAMEX: Multi-species integration, alignment and annotation of single-cell RNA-seq data with CAMEX (v0.0.1). Zenodo. https://doi.org/10.5281/zenodo.17991379

Acknowledgements

This work has been supported by the National Key Research and Development Program of China [No. 2021YFA1302500 to S.Z.], the National Natural Science Foundation of China [Nos. 32341013, 12326614 to S.Z., Nos. 62333018, 62372255, 62432013, W2412087, 62402250, 62433001, U22A2039 to D.S.H.], the CAS Project for Young Scientists in Basic Research [No. YSBR-034 to S.Z.], the Natural Science Foundation of Zhejiang Province [No. LMS25F020001 to D.S.H.], the Key Research and Development Program of Ningbo City [Nos. 2024Z112, 2023Z219, 2023Z226 to D.S.H.], the Yongjiang Talent Project of Ningbo, Yongrencaifa [No. 2024-4 to D.S.H.], and the Basic Research Program Project of Department of Science and Technology of Guizhou Province [No. ZK2024ZD035 to D.S.H.].

Author Contributions

S.Z. conceived and supervised the project. Z. G. collected the datasets and developed the algorithm. Z. G., D. H. and S.Z. performed the analyses and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Figure captions

Figure 1. Overview of CAMEX. a, Construction of a heterogeneous graph representing the relationships between cells and genes across different species. This multi-species cell-gene heterogeneous graph consists of six types of edges, i.e., ‘cell express gene’, ‘gene expressed by cell’, ‘cell kNN’, ‘gene orthologous relationship’, ‘cell self-loop’, and ‘gene self-loop’. **b**, CAMEX leverages a heterogeneous GNN encoder to embed cells and genes from various species into a shared space. The integration task employs the reconstruction decoders for both edges and features. The cell annotation task utilizes a graph attention network classifier. Training details outline the message propagation paradigm for six types of edges and graph sampling techniques. **c**, CAMEX serves several applications, including multi-species integration, alignment of developmental stages, and cell annotation.

Figure 2. Benchmarking CAMEX and SOTA methods on the liver dataset across four species. a, Overall score evaluation on the liver dataset. The overall score considers both batch correction and biological variation conservation performance. Rankings are depicted using a color gradient, where lighter colors indicate better performance. **b**, UMAP visualization of joint embedding space on the liver dataset. Colors represent species annotations. Circles represent the hepatocytes cell population from different

species. **c**, UMAP visualization of joint embedding space on the liver dataset. Colors represent cell type annotations. Squares represent the Kupffer cells and Macrophages as well as cholangio cell population from different species.

Figure 3. CAMEX achieved superior integration performance and preserved the differentiation process in the testis dataset across 11 species. a, Overall score evaluated on testis and phylogenetic tree generated by TimeTree (v.5) (http://www.timetree.org/). **b**, UMAP visualization of a joint embedding space on the testis dataset. Colors denote species annotations (first row) and cell type annotations (second row). **c**, UMAP visualization of CAMEX integration result where color denotes the mean embedding of gene clusters 3 and 5 generated by CAMEX. **d**, Gene set enrichment analysis of clusters 3 and 5 was performed using the g:Profiler package, applying a one-sided hypergeometric test with g:SCS (Set Counts and Sizes correction) multiple testing correction.

Figure 4. CAMEX aligns various development stages of seven organs across seven different species. a, Phylogenetic tree created using TimeTree (v.5) (http://www.timetree.org/). UMAP visualization of joint embedding space on this multi-species dataset, encompassing various organs and developmental stages. Cells are colored by the species (left), organs (middle), and stages (right). **b**, Similarities between all organs in mice and rats with the mouse as a reference, ensuring their developmental times were well aligned. The Carnegie stages in mouse and rat correspond to gestational days 9-16 and 10.5-17.5, respectively. Here, "ek" and "pk" denote the k-th embryonic day and postnatal day, respectively.

Figure 5. CAMEX integrates and annotates relatives and distant species cortex scRNA-seq dataset. a, UMAP visualization of joint embedding space. Colors denote species annotations (first row) and cell type annotations (second row). Circles represent the cell population that can not be integrated very well. The cell types are abbreviated as follows: Astrocyte (Ast), Brain Pericyte (BP), Endothelial (End), Excitatory Neuron (ExN), Inhibitory Neuron (InN), Microglial cell (Micro), Neural Progenitor Cell (NPC), Oligodendrocyte (Oligo), Oligodendrocyte Precursor Cell (OPC), Purkinje cell (PC), Cerebellar Granule Cell (CGC), Ependymal cell (EC), Macrophage (mø), UnKnown (UK). **b**, ACC, PRC, and F1 scores achieved by CAMEX and baseline methods on the turtle dataset. Unknown is the cell type dost not appear in reference but appear in query dataset. NPC dost not appear in reference human dataset but appear in query turtle and lizard dataset, so we defined them as unknown. **c**, Row-normalized confusion matrix of CAMEX and baseline methods on turtle dataset. Row-normalized confusion matrix is provided where the rows represent the true labels and the columns correspond to the predicted labels. Row normalization scales values within the matrix to a range of [0, 1], where a darker color indicates a value closer to 1, signifying higher accuracy or agreement. d,

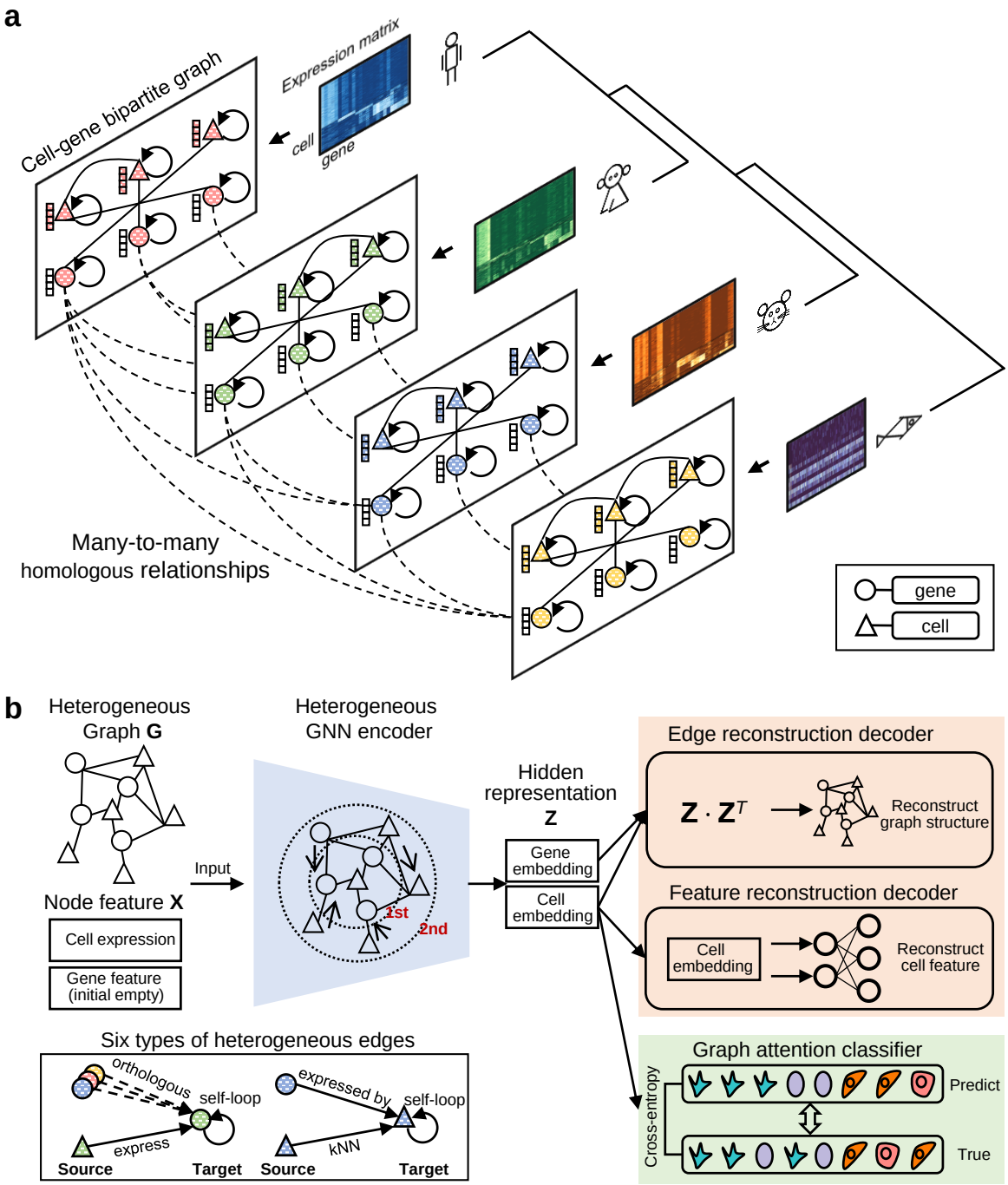
UMAP visualization of turtle dataset. Colors denote manual and predicted annotations. Circles represent the cell population that cannot be integrated very well.

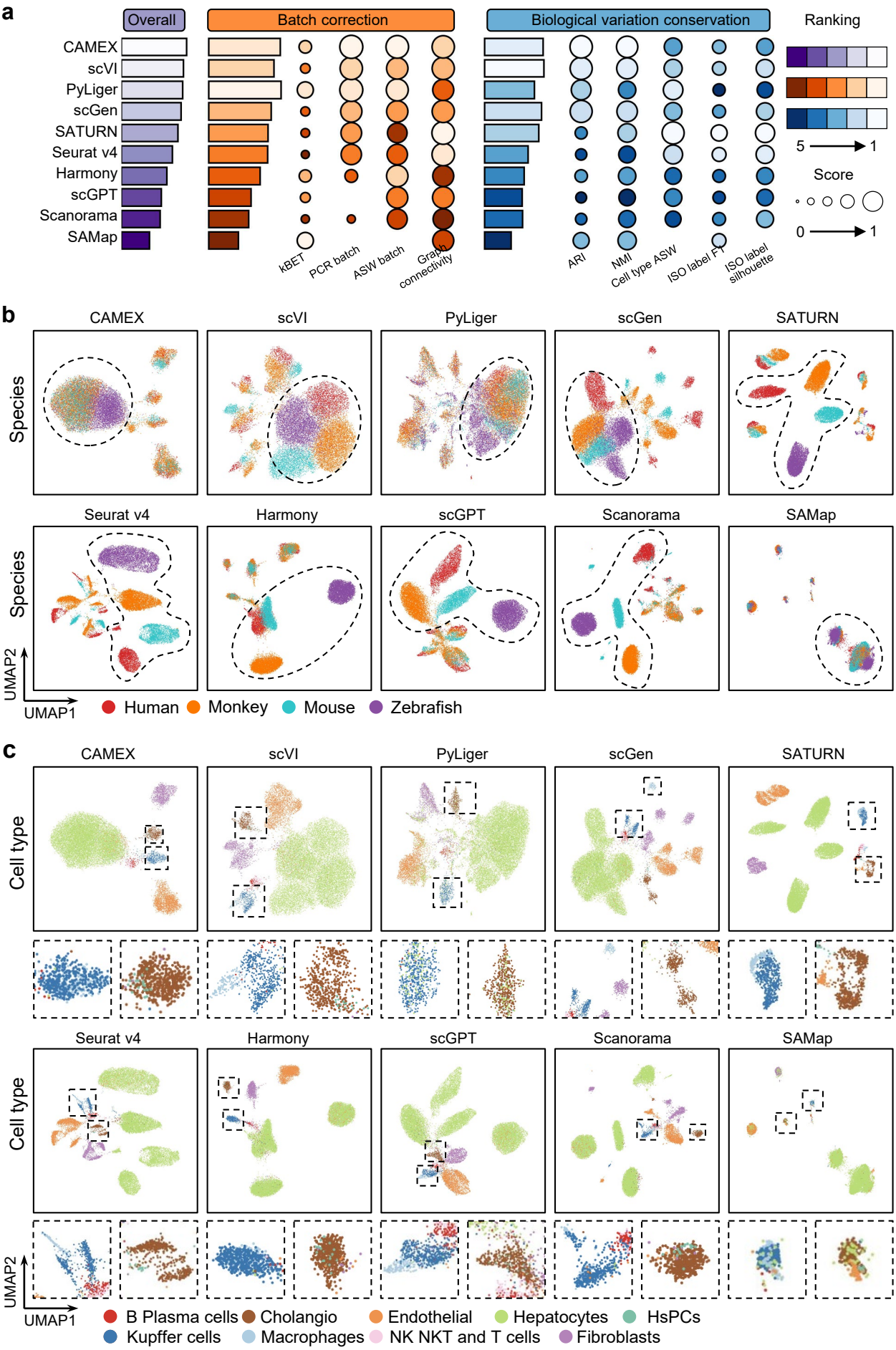
Figure 6. Integrating the dorsolateral prefrontal cortex (DLPFC) dataset of four species, i.e., chimpanzees, marmosets, rhesus, and humans by CAMEX. a, The class and subclass of cell types in DLPFC of four species, i.e., human, chimpanzee, marmoset, and rhesus. Note that two kinds of subtypes in the microglial are species specifically. UMAP visualization of a joint embedding space on the DLPFC dataset across species. Colors denote cell type annotations. Specifically, the subtypes and abbreviations of excitatory and inhibitory neurons are defined by their marker genes and locations: intratelencephalic (IT) extratelencephalic (ET) near-projecting (NP); corticothalamic (CT), abbreviation for glial cells, oligodendrocytes (Oligo), microglia (Micro), OPC (Oligodendrocyte Precursor Cell), astrocyte (Astro), abbreviation for non-neuronal cells: vascular leptomeningeal cells (VLMC), red blood lineage cells (RB), endothelial cells (Endo), immune cells (Immune), pericyte (PC), smooth muscle cells (SMC). **b**, UMAP visualization of a joint embedding space on microglial. **c**, Dot plots showing the differential expression genes in different species in the APBB1IP subtype.

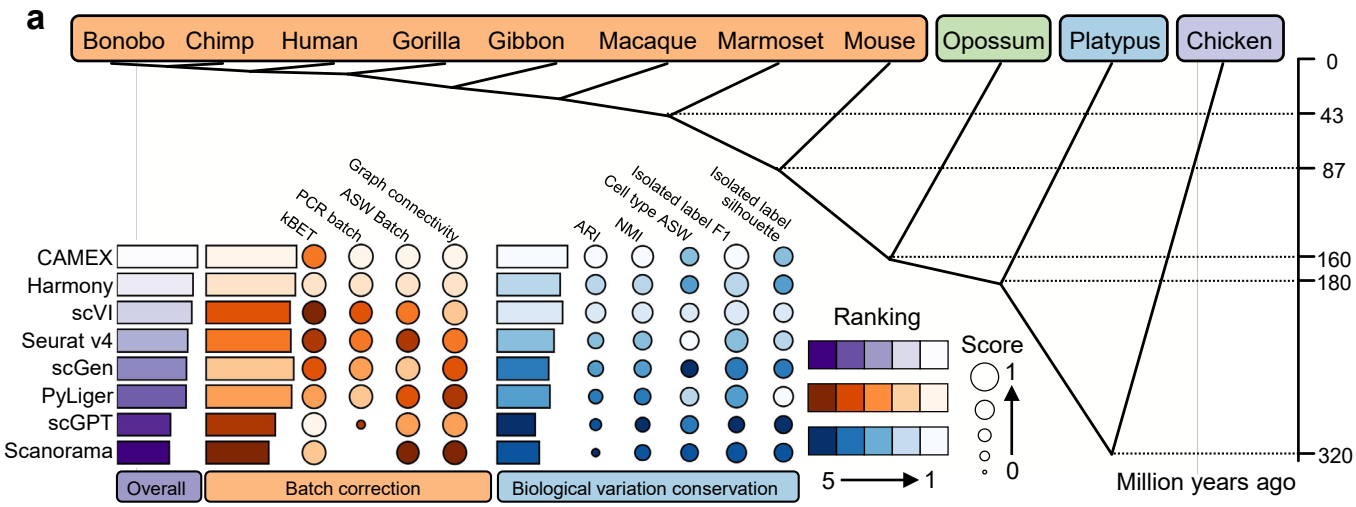
Editor's Summary

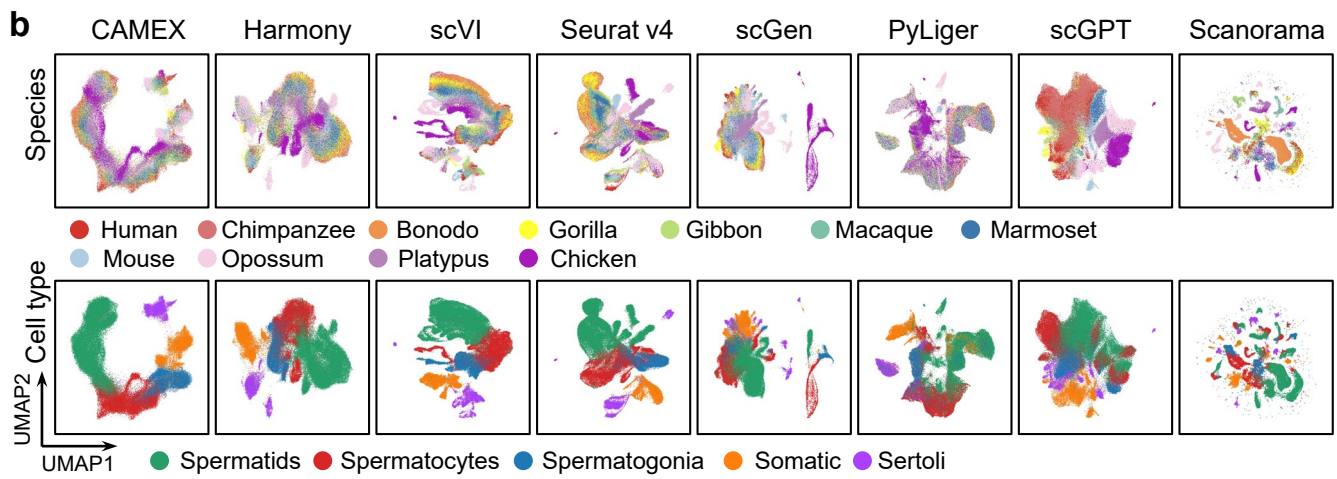
Single-cell RNA-sequencing data enable cross-species comparisons, but integrating them remains hard. Here, authors present CAMEX, which leverages many-to-many gene homology to integrate, align, and annotate multi-species single-cell data, revealing both conserved and species-specific cellular characteristics.

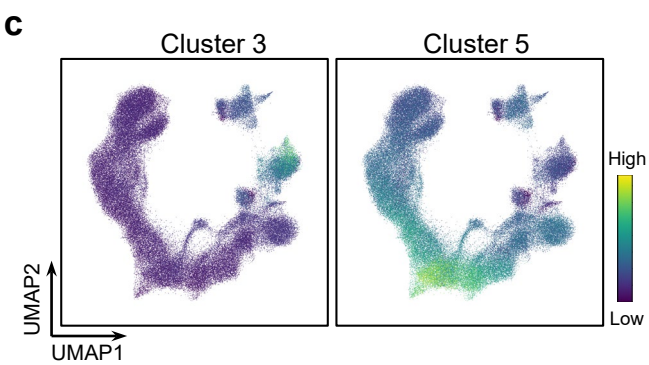
Peer Review Information: *Nature Communications* thanks Doron Betel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

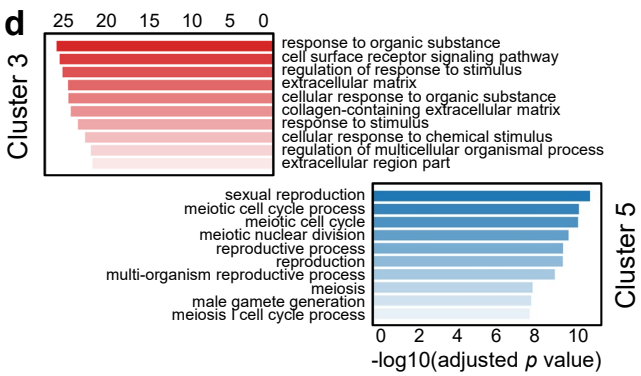

a
Cell-gene bipartite graph
Expression matrix
cell
gene
Many-to-many
homologous relationships
gene
cell
b
Heterogeneous
Graph G
Heterogeneous
GNN encoder
Input
Node feature X
Cell expression
Gene feature
(initial empty)
1st
2nd
Hidden
representation
Z
Gene
embedding
Cell
embedding
Edge reconstruction decoder
Z · Z^T
Reconstruct
graph structure
Feature reconstruction decoder
Cell
embedding
Reconstruct
cell feature
Six types of heterogeneous edges
orthologous
self-loop
express
Source
Target
expressed by
self-loop
kNN
Source
Target
Graph attention classifier
Cross-entropy
Predict
True
c
Multi-species
integration
Multi-species
development stage alignment
Cross-species
annotation

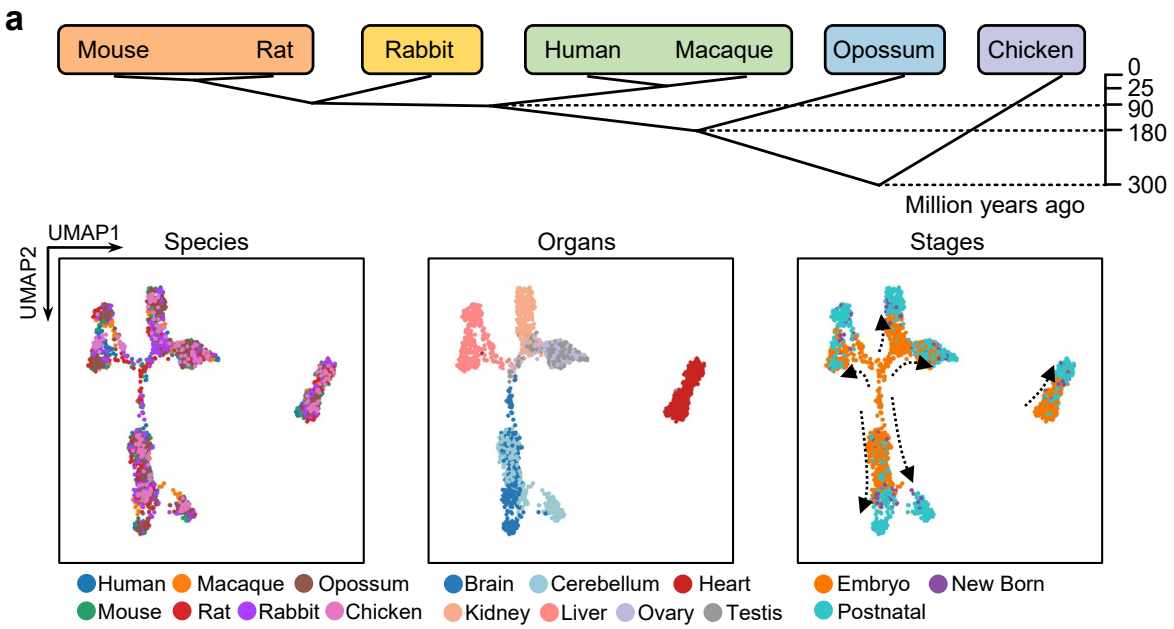

a
Overall
Batch correction
Biological variation conservation
Ranking
CAMEX
scVI
PyLiger
scGen
SATURN
Seurat v4
Harmony
scGPT
Scanorama
SAMap
kBET
PCR batch
ASW batch
Graph connectivity
ARI
NMI
Cell type ASW
ISO label F1
ISO label silhouette
5
1
Score
0
1
b
CAMEX
scVI
PyLiger
scGen
SATURN
Species
Seurat v4
Harmony
scGPT
Scanorama
SAMap
UMAP2
UMAP1
Human
Monkey
Mouse
Zebrafish
c
CAMEX
scVI
PyLiger
scGen
SATURN
Cell type
Seurat v4
Harmony
scGPT
Scanorama
SAMap
UMAP2
UMAP1
B Plasma cells
Cholangio
Endothelial
Hepatocytes
HsPCs
Kupffer cells
Macrophages
NK NKT and T cells
Fibroblasts

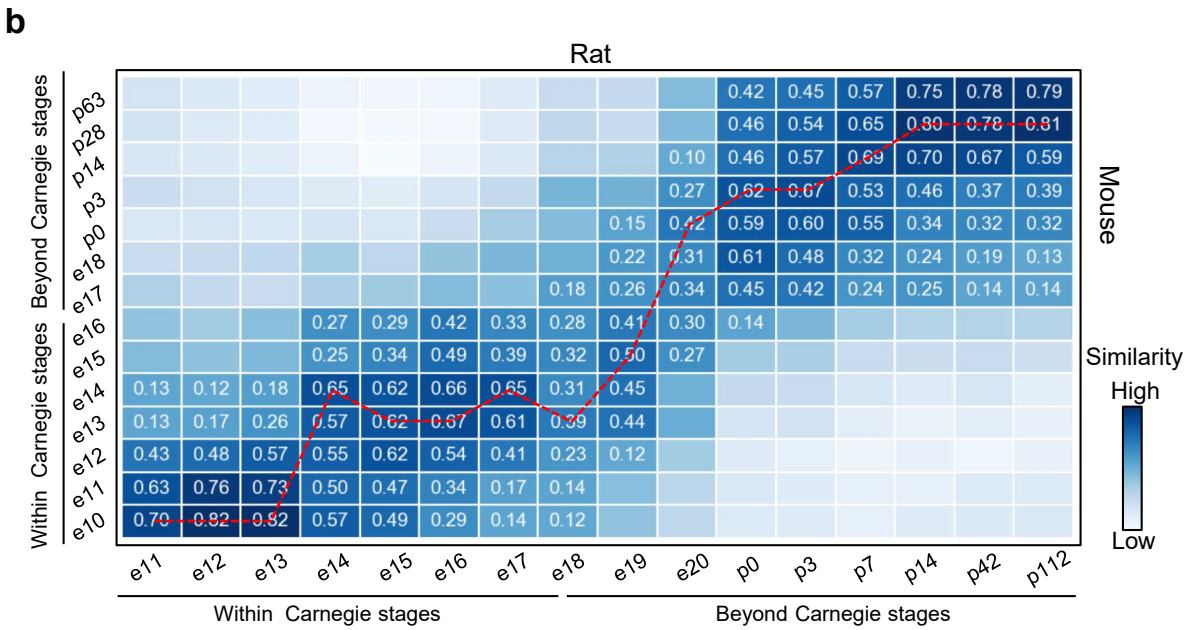

a
Bonobo
Chimp
Human
Gorilla
Gibbon
Macaque
Marmoset
Mouse
Opossum
Platypus
Chicken
0
43
87
160
180
320
Million years ago
kBET
PCR batch
ASW Batch
Graph connectivity
ARI
NMI
Cell type ASW
Isolated label F1
Isolated label silhouette
CAMEX
Harmony
scVI
Seurat v4
scGen
PyLiger
scGPT
Scanorama
Ranking
Score
1
0
5
1
Overall
Batch correction
Biological variation conservation

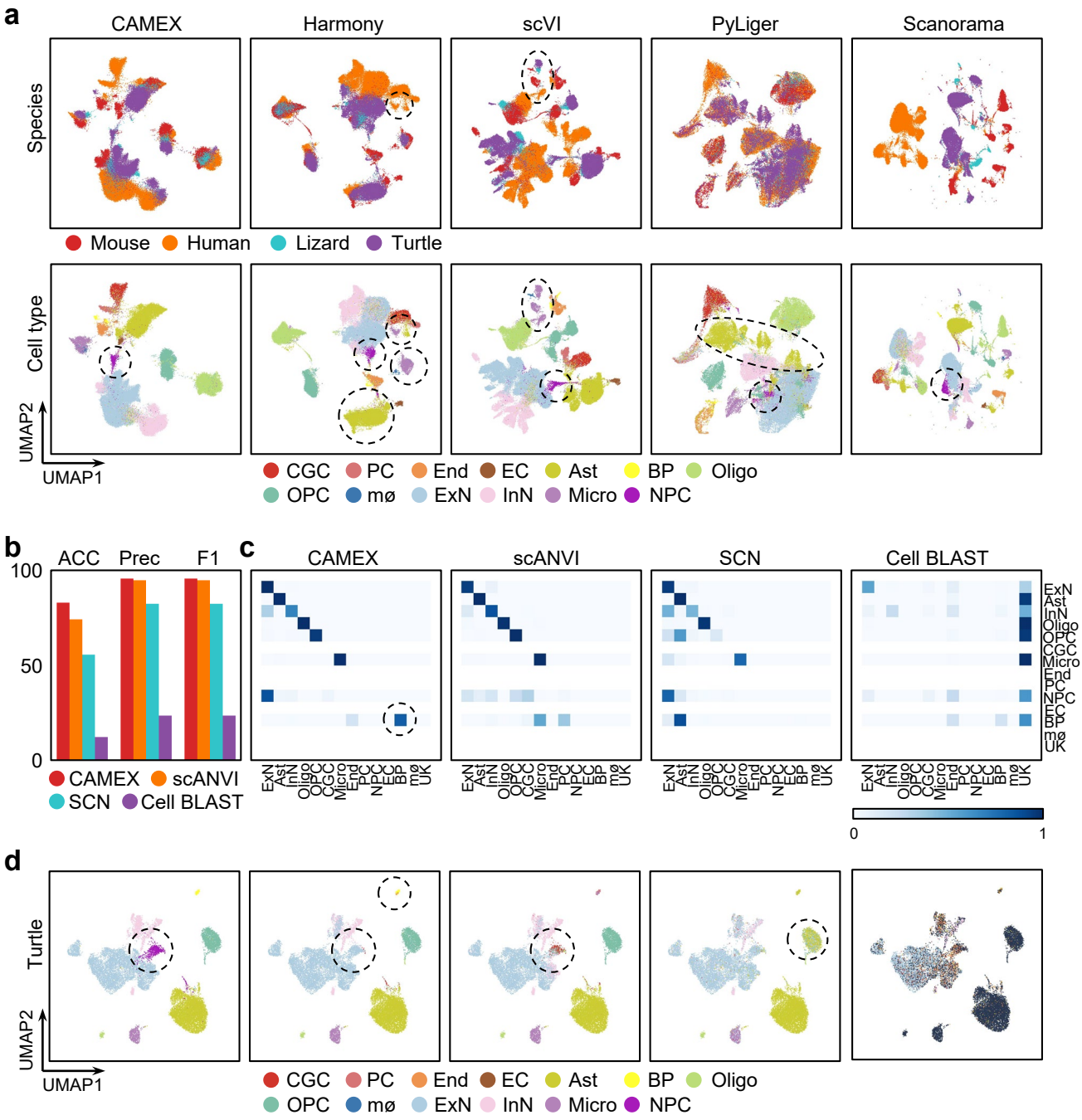

b
CAMEX
Harmony
scVI
Seurat v4
scGen
PyLiger
scGPT
Scanorama
Species
Human
Chimpanzee
Bonobo
Gorilla
Gibbon
Macaque
Marmoset
Mouse
Opossum
Platypus
Chicken
Cell type
UMAP2
UMAP1
Spermatids
Spermatocytes
Spermatogonia
Somatic
Sertoli

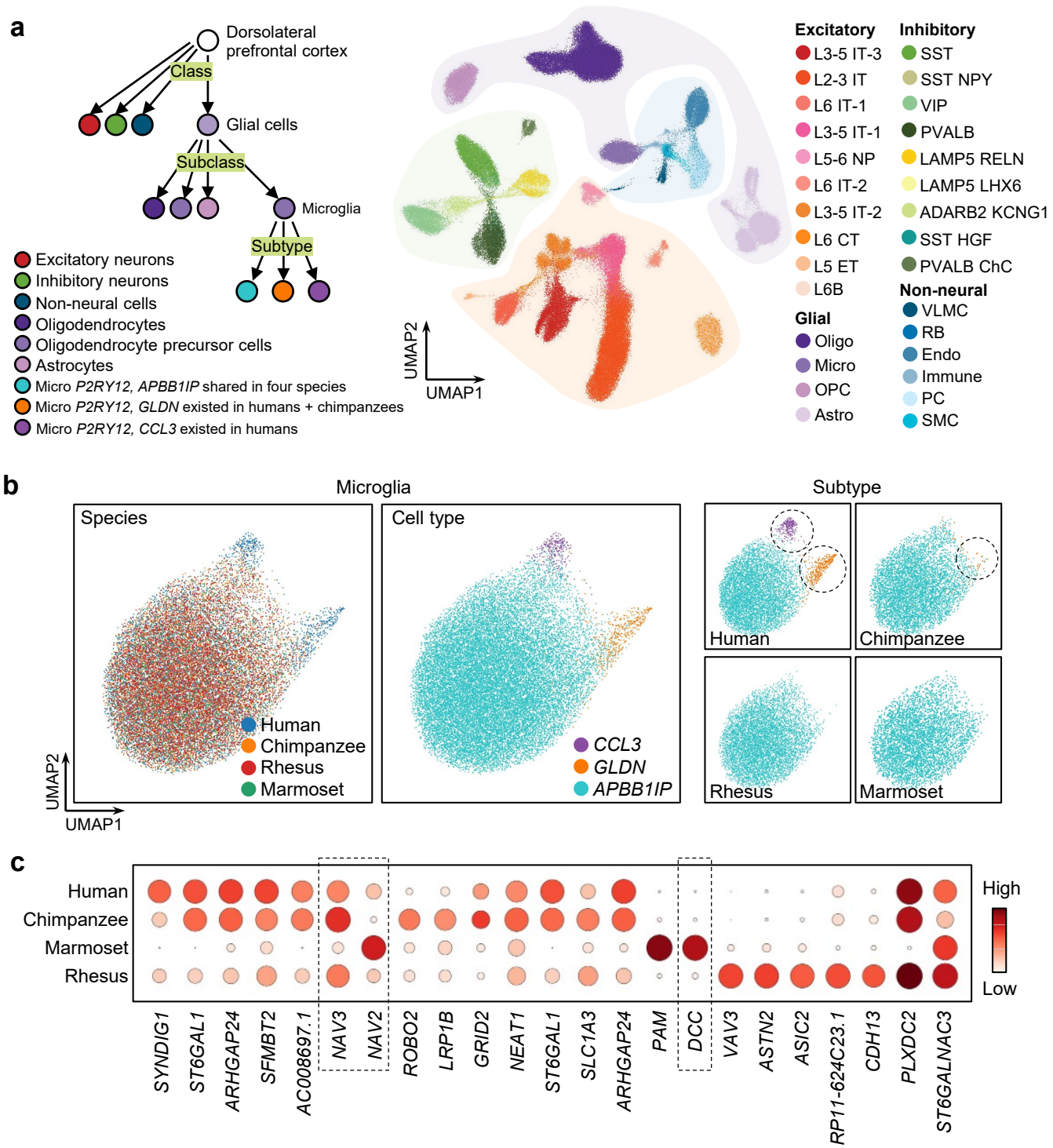

c
Cluster 3
Cluster 5
High
Low
UMAP2
UMAP1


d
Cluster 3
25
20
15
10
5
0
response to organic substance
cell surface receptor signaling pathway
regulation of response to stimulus
extracellular matrix
cellular response to organic substance
collagen-containing extracellular matrix
response to stimulus
cellular response to chemical stimulus
regulation of multicellular organismal process
extracellular region part
Cluster 5
sexual reproduction
meiotic cell cycle process
meiotic cell cycle
meiotic nuclear division
reproductive process
reproduction
multi-organism reproductive process
meiosis
male gamete generation
meiosis I cell cycle process
0
2
4
6
8
10
-log10(adjusted p value)


a
Mouse
Rat
Rabbit
Human
Macaque
Opossum
Chicken
0
25
90
180
300
Million years ago
UMAP1
UMAP2
Species
Organs
Stages
Human
Macaque
Opossum
Mouse
Rat
Rabbit
Chicken
Brain
Cerebellum
Heart
Kidney
Liver
Ovary
Testis
Embryo
New Born
Postnatal


b
Rat
Mouse
Beyond Carnegie stages
Within Carnegie stages
p63
p28
p14
p3
p0
e18
e17
e16
e15
e14
e13
e12
e11
e10
e11
e12
e13
e14
e15
e16
e17
e18
e19
e20
p0
p3
p7
p14
p42
p112
Within Carnegie stages
Beyond Carnegie stages
Similarity
High
Low


a
CAMEX
Harmony
scVI
PyLiger
Scanorama
Species
Mouse
Human
Lizard
Turtle
Cell type
UMAP2
UMAP1
CGC
PC
End
EC
Ast
BP
Oligo
OPC
mø
ExN
InN
Micro
NPC
b
ACC
Prec
F1
100
50
0
CAMEX
scANVI
SCN
Cell BLAST
c
CAMEX
scANVI
SCN
Cell BLAST
ExN
Ast
InN
Oligo
OPC
CGC
Micro
End
PC
NPC
EC
BP
mø
UK
0
1
d
Turtle
UMAP2
UMAP1
CGC
PC
End
EC
Ast
BP
Oligo
OPC
mø
ExN
InN
Micro
NPC


a
Dorsolateral prefrontal cortex
Class
Glial cells
Subclass
Microglia
Subtype
Excitatory neurons
Inhibitory neurons
Non-neural cells
Oligodendrocytes
Oligodendrocyte precursor cells
Astrocytes
Micro P2RY12, APBB1IP shared in four species
Micro P2RY12, GLDN existed in humans + chimpanzees
Micro P2RY12, CCL3 existed in humans
UMAP2
UMAP1
Excitatory
L3-5 IT-3
L2-3 IT
L6 IT-1
L3-5 IT-1
L5-6 NP
L6 IT-2
L3-5 IT-2
L6 CT
L5 ET
L6B
Glial
Oligo
Micro
OPC
Astro
Inhibitory
SST
SST NPY
VIP
PVALB
LAMP5 RELN
LAMP5 LHX6
ADARB2 KCNG1
SST HGF
PVALB ChC
Non-neural
VLMC
RB
Endo
Immune
PC
SMC
b
Microglia
Species
Human
Chimpanzee
Rhesus
Marmoset
Cell type
CCL3
GLDN
APBB1IP
Subtype
Human
Chimpanzee
Rhesus
Marmoset
UMAP2
UMAP1
c
Human
Chimpanzee
Marmoset
Rhesus
SYNDIG1
ST6GAL1
ARHGAP24
SFMBT2
AC008697.1
NAV3
NAV2
ROBO2
LRP1B
GRID2
NEAT1
ST6GAL1
SLC1A3
ARHGAP24
PAM
DCC
VAV3
ASTN2
ASIC2
RP11-624C23.1
CDH13
PLXDC2
ST6GALNAC3
High
Low