Single-Cell RNA-Sequencing Data Clustering via Locality Preserving Kernel Matrix Alignment

XIAO ZHENG,1†, JIAJIA CHEN,2†, CHANG TANG,3† (Member, IEEE), AND SUQIN ZHOU4,*

1School of Computer, National University of Defense Technology, Changsha 410073, China
2Department of Pharmacy, The Affiliated Huai’an Hospital, Xuzhou Medical University, Huai’an 223100, China
3School of Computer Science, China University of Geosciences, Wuhan 430074, China
4Department of Pharmacy, Lianshui People’s Hospital Affiliated to Kangda College, Nanjing Medical University, Huai’an 223300, China

*Corresponding author: Suqin Zhou (sqinzhou@163.com)
†Xiao Zheng and Jiajia Chen contributed equally as first author to this work.

This work was supported in part by the National Natural Science Foundation of China under Grant 61701451 and Grant 62076228, in part by the Fundamental Research Funds for the Central Universities, and in part by the China University of Geosciences, Wuhan, under Grant CUG170654.

ABSTRACT Single-cell RNA-sequencing (scRNA-seq) data provide opportunities to reveal new insights into many biological problems such as elucidating cell types. An effective approach to elucidate cell types in complex tissues is to partition the cells into several separated subgroups via clustering techniques, where the cells in a specific cluster belong to the same cell type based on gene expression patterns. In this work, we present a novel multiple kernel clustering framework for scRNA-seq data clustering via locality preserving kernel alignment. Specifically, we first generate a series of similarity kernel matrices by using different kernel functions. Then we transfer the clustering task to a multiple kernel k-means problem with the kernels aligned in a local manner, i.e., the similarity of a sample to its k-nearest neighbours are aligned with the ideal similarity matrix. In our method, the clustering process focuses on closer sample pairs that shall stay together, and avoids involving unreliable similarity evaluation for farther sample pairs. In addition, we construct a local Laplacian matrix for each sample to constrain that closer samples should be allocated similar labels. In such a manner, the local structure of the data can be well preserved and utilized to produce better alignment for clustering. An alternate updating algorithm with theoretical analysis is developed to solve the proposed problem. We evaluate the performance of the proposed method on various real scRNA-seq data, and the results show that our method can obtain superior results when compared with other state-of-the-art approaches.

INDEX TERMS scRNA-seq clustering, elucidating cell types, tissues, similarity kernel matrices, locality preservation.

I. INTRODUCTION Recent literature indicate that single-cell measurements play an important role in understanding cellular heterogeneity [1]–[5] and cell differentiation [6], [7]. Thanks to the rapid development of Single-cell RNA-sequencing (scRNA-seq) techniques, a tremendous amount of transcriptome datasets have been generated at single-cell resolution [8], [9]. On the one hand, these datasets provide opportunities to reveal new insights into many biological problems, e.g., elucidating cell types, on the other hand, there are also computational challenges due to the amount of data. A straightforward approach to elucidate cell types in complex tissues is to partition the cells into some separated subgroups via clustering techniques [10]–[13], which can be regarded as an unsupervised classification problem [14]–[16]. Many previous clustering techniques can be used for this task, such as principal component analysis (PCA) [17], spectral clustering [18], and k-means [19]. However, different to bulk RNA-seq or gene expression microarrays, there are high level of noise and many missing values in scRNA-seq data due to technical and sampling issues [20]–[22]. In addition, the high variability exists in gene expression levels even between cells of the same type, and this could degenerate the performance of those existing clustering approaches [23]–[27]. In order to address the issues in scRNA-seq data analysis, a various of novel clustering methods have been proposed in recent years. For subtype classification and detection of
relationships between the subtypes, some iterative clustering methods have been proposed [28]–[31]. Haghtorabi et al. [32] used the diffusion maps to perform dimension reduction of the data, which stresses continuity of cell states along putative developmental pathways. Nonnegative Matrix Factorization (NMF) technique has also been used to decompose the high-dimensional scRNA-seq data into biologically interpretable compositions [33], and the functional cell subgroups and biologically relevant features can be simultaneously obtained with NMF. Several graph theory-based algorithms have also been applied to scRNA-seq data clustering problems. Xu and Su [34] developed a quasi-clique-based clustering algorithm named SNN-Cliq to identify tight groups of highly similar nodes that are likely to belong to the same genuine clusters. In SNN-Cliq, the clusters are identified by using the proposed SNN graph. Spectral clustering, as a typical graph based clustering method, has also been successfully deployed for this task. Park and Zhao [35] first constructed a series of symmetric doubly stochastic similarity matrices by using the Gaussian kernel function with varying parameters, then they learned a target similarity matrix from the previous constructed matrices for spectral clustering. In [36], Wu et al. integrated dimension reduction and clustering of single-cell RNA-sequencing data into a unified framework.

Multiple kernel clustering is also a kind of popular modern clustering method which aims to optimally integrate a group of pre-specified kernels to improve clustering performance. A critical issue in multiple kernel clustering is to learn the kernel combination weights. Margolin [37] made the combination weights adaptively change with respect to samples to better capture their individual characteristics. Du et al. [38] presented a robust multiple kernel k-means algorithm that simultaneously finds the best clustering labels and the optimal combination of multiple kernels by replacing the squared error in k-means with an $l_2,1$-norm based one. Lu et al. [39] employed kernel alignment maximization to jointly perform the k-means clustering and multiple kernel learning. Wang et al. [40] presented an analytic framework (SIMLR) via multi-kernel learning which learns a similarity measure from scRNA-seq data in order to perform dimension reduction, clustering and visualization. Compared to other previous methods, SIMLR learns a distance metric that best fits the structure of the data by combining multiple kernels. Standard dimension reduction or clustering algorithms often work under certain statistical assumptions which the diverse statistical characteristics of scRNA-seq data could not easily fit well. Qi et al. [41] proposed to automatically learn similarity information from data and introduced a new clustering method in the form of a multiple kernel combination, which can directly discover groupings in scRNA-seq data. In this article, we propose a new scRNA-seq data clustering method via locality preserving multiple kernel alignment (referred to as LPKA briefly). Considering that previous kernel alignment methods often rigidly constrain closer and farther sample pairs to be equally aligned to the same ideal similarity, and the intra-cluster variation of samples is inappropriately neglected, we propose to align the kernels in a local manner, i.e., the similarity of a sample to only its k-nearest neighbours are aligned with the ideal similarity matrix. In our method, the clustering process focuses on closer sample pairs that shall stay together, and avoids involving unreliable similarity evaluation for farther sample pairs. In addition, we construct a local Laplacian matrix for each sample to constrain that closer samples should be allocated similar labels. In such a manner, the local structure of the data can be well preserved and utilized to produce better alignment for clustering. Experiments on 9 scRNA-seq datasets are conducted to demonstrate the superiority of our proposed method.

### II. MATERIALS AND METHODS

#### A. DATASETS COLLECTION AND KERNEL MATRICES GENERATION

1) DATASETS

In order to evaluate the efficacy of our proposed LPKA, we use some real-world scRNA-seq datasets to test the clustering performance. Similar to [35], we test the performance of LPKA on 9 scRNA-seq datasets which represent several types of dynamic processes such as cell cycle, cell differentiation, and response upon external stimulus. For each dataset, the types of cells are known as a priori. The number of cells, number of cell types, number of genes for each dataset are summarized in Table 1.

For the readability and integrity of this article, we also give the detailed information of the datasets as follows:

- Treutlein [42]. This dataset is composed of single cell RNA-seq expression data for 80 lung epithelial cells at E18.5 together with five putative cell types including AT1, AT2, Clara, BP, and ciliated. Similar to [42], we considered data with selected genes with 959 highest loadings in the first four PCA coefficients. The dataset is downloaded from: https://www.nature.com/articles/nature13173.

- Ting. This dataset contains 5 subtypes from Single-cell transcriptomes from MEFs, the NB508 pancreatic cancer cell line, normal WBCs, bulk primary tumors diluted to 10 or 100 pg of RNA, and classical CTC. We downloaded the data from GEO (GSE51372).

- Deng [43]. This dataset consists of transcriptomes for individual cells isolated from mouse embryos at different preimplantation stages. There are 135 cells and 19,703 genes, where cells belong to zygote, early

| Table 1: Brief Information of the 9 Used scRNA-seq Datasets in our Experiments. |
|-----------------|-----------|----------|----------|-----------|
| **Datasets**    | **No. of cells** | **No. of genes** | **No. of cell types** |
| Treutlein       | 80        | 9352     | 5        |
| Ting            | 114       | 14405    | 5        |
| Deng            | 135       | 12548    | 7        |
| Ginevoux        | 251       | 11834    | 3        |
| Buttner         | 182       | 8989     | 3        |
| Pollen          | 249       | 14805    | 11       |
| Tasic           | 1727      | 5832     | 49       |
| Sewel           | 3005      | 4412     | 47       |
| Mascosko        | 6418      | 12822    | 39       |
FIGURE 1. The objective value of our algorithm at each iteration on each dataset.

2-cell-stage, mid 2-cell-stage, late 2-cell-stage, 4-cell-stage, 8-cell-stage, and 16-cell-stage. The processed data is downloaded from GEO (GSE45719).

- Ginhoux. This dataset consists of the expression values of 15,752 genes for 251 dendritic cell progenitors in one of following three cellular states: Monocyte and Dendritic cell Progenitors (MDPs), Common Dendritic cell Progenitors (CDPs), and Pre-Dendritic Cells (PreDCs). The dataset contains 59 MDPs, 96 CDPs, and 96 PreDCs. We downloaded the processed data from GEO (GSE60783).

- Buettnert [44]. This dataset contains the transcriptional profile of 182 ESCs that has been staged for cell-cycle phase (G1, S, and G2M) based on sorting of the Hoechst 33342-stained cell area of a flow cytometry (FACS) distribution. The cells were sorted for three stages of the cell cycle, and they were validated using gold-standard Hoechst staining. The data have been deposited at ArrayExpress: E-MTAB-2805.

- Pollen. There are 249 single cells from 11 populations using microfluidics, including neural cells and blood cells. The 11 clusters in the dataset were from different sources (CRL-2338, CRL-2339, K562, BJ, HL60, hiPSC, Keratinocyte, Fetal cortex (GW21+3), Fetal cortex (GW21), Fetal cortex (GW16), and NPC) that are expected to show robust differences in gene expression. Data were pre-filtered to exclude genes where more than 90% of cells had zero measurements and include only single cells with greater than 500000 reads (n = 249).

- Tasic [30]. There are 49 transcriptomic cell types in this dataset, including 23 GABAergic, 19 glutamatergic and 7 non-neuronal types. To identify cell types, Tasic et al. [30] applied two parallel and iterative approaches for dimensionality reduction and clustering, iterative principal component analysis (PCA) and iterative weighted gene coexpression network analysis (WGCNA), and validated the cluster membership from each approach using a non-deterministic machine
learning method (random forest). We downloaded the processed data from GEO (GSE71585).

- Zeisel [28]. In this dataset, Zeisel et al. [28] used large-scale single-cell RNA sequencing to classify cells in the mouse somatosensory cortex and hippocampal CA1 region. 3005 Cells from the mouse cortex and hippocampus collected. Zeisel et al. (2015) found 47 molecularly distinct subclasses identified by hierarchical biclustering and validated by gene markers.

- Macosko. This dataset contains mouse retina cells with 39 subtypes, it is obtained by droplet-based high-throughput technique. The dataset consists of 44808 cells. The 39 cell types were identified via PCA and density-based clustering, and they were validated by differential gene expression. We filtered out cells with less than 1200 genes (yielding 6418 cells) for clustering analysis. We downloaded the data from GEO (GSE63473).

2) KERNEL MATRICES GENERATION

In our experiments, three kinds of kernels are used to generate kernel matrices, including Gaussian kernel, linear kernel and polynomial kernel. For Gaussian kernel, it is one of the most widely used kernel function and it can obtain steadily performance whether the data size is large or small. We follow [40] and consider multiple kernel functions to construct kernel matrices as follows:

$$\text{GK}_{\theta,k}(i,j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2_{ij}}\right),$$

where $\sigma_{ij} = \frac{\theta(\pi_i + \pi_j)}{2}$, and $\pi_i = \frac{\sum_{l\in KNN(i)} ||x_l - x_i||}{k}$. $KNN(i)$ represents a set of sample indices that are the top $k$ nearest neighbors of the sample $x_i$. As can be seen, parameters $\theta$ and $k$ control the width of the neighborhoods. For generality, we also vary $\theta$ from $\{1, 1.25, \ldots, 2\}$ and $k$ from $\{10, 12, \ldots, 30\}$. Thus, a total number of 55 Gaussian kernel matrices can be obtained.

For linear kernel, it is suitable for the data samples which are linear separable. In addition, it has no parameter. We generate a linear kernel matrix as follows:

$$\text{LK}(i,j) = x_i \cdot x_j.$$  

For polynomial kernel, it projects low dimensional feature space to a higher dimensional feature space. It is defined as
follows:

$$PK(i,j) = ((x_i \cdot x_j) + 1)^d.$$  \hspace{1cm} (3)

In this work, we vary $d$ from \{0.1, 0.2, \ldots, 1\} and obtain 10 polynomial kernel matrices.

Finally, for each dataset, we combine different kinds of kernel matrices to obtain a kernel matrix with size $n \times n \times 66$ for further computation, where $n$ is the number of samples in the dataset.

**B. KERNEL K-MEANS CLUSTERING**

Given a set of $n$ data samples from $k$ clusters $\{x_i\}_{i=1}^n \subseteq \mathcal{X}$, let $\phi(\cdot) : x \in \mathcal{X} \mapsto \mathcal{H}$ be a feature mapping which maps $x$ from original space onto a reproducing kernel Hilbert space $\mathcal{H}$. Kernel k-means aims to minimize the sum-of-square loss over the cluster assignment matrix $Z \in \{0, 1\}^{n \times k}$, and the problem can be solved by minimizing following objective function:

$$\min_{Z \in \{0, 1\}^{n \times k}} \sum_{i=1}^n \sum_{c=1}^k Z_{ic}^2 ||\phi(x_i) - \xi_c||_2^2, \quad s.t. \quad \sum_{i=1}^n Z_{ic} = 1,$$  \hspace{1cm} (4)

where $n_c = \sum_{i=1}^n Z_{ic}$ and $\xi_c = \frac{1}{n_c} \sum_{i=1}^n \phi(x_i)$ represent the sample number and centroid of the $c$-th cluster in $\mathcal{H}$.

By some algebra, (4) can be transferred to following matrix form:

$$\min_{H \in \mathbb{R}^{n \times k}} \text{Tr}(K) - \text{Tr}(L^\frac{1}{2} Z^T K Z L^\frac{1}{2}), \quad s.t. \quad Z 1_k = 1_n.$$

where $\text{Tr}(\cdot)$ denotes the trace of a matrix and $K$ is a kernel matrix with $K_{ij} = \phi(x_i)^T \phi(x_i)$, $L = \text{diag}([n_1^{-1}, n_2^{-1}, \ldots, n_k^{-1}])$, and $1_l \in \mathbb{R}^l$ is a column vector with all elements 1.

The problem (5) is hard to solve since $Z$ is discrete. Fortunately, this problem can be usually approximated through relaxing $Z$ to take arbitrary real values. By defining $H = Z L^\frac{1}{2}$ and letting $H$ take real values, the relaxed version of (5) can be obtained as:

$$\min_{H \in \mathbb{R}^{n \times k}} \text{Tr}(K(I_h - HH^T)), \quad s.t. \quad H^TH = I_k,$$  \hspace{1cm} (6)

where $I_k$ is a $k \times k$ identity matrix. Since $Z^T Z = L^{-1}$, we have $L^\frac{1}{2} Z^T Z L^\frac{1}{2} = I_k$, then it is easy to get the orthogonality.

**FIGURE 3.** Purity evaluation of different clustering methods on different datasets.
constraint on $H$. Finally, (6) can be solved by taking the $k$ eigenvectors that correspond to the $k$ largest eigenvalues of $K$.

**C. MULTIPLE KERNEL K-MEANS CLUSTERING**

With a multiple kernel setting, each sample is represented via a group of feature mappings $\{\phi_j(x)\}_{j=1}^m$ [45]. In detail, each sample can be represented as $\phi_\xi(x) = [\xi_1 \phi_1(x)^T, \xi_2 \phi_2(x)^T, \ldots, \xi_m \phi_m(x)^T]^T$, where $\xi = [\xi_1, \xi_2, \ldots, \xi_m]^T$ represents the coefficients of each base kernel that we need to learn. As a consequence, the corresponding kernel function over the above mapping function can be written as:

$$k_\xi(x_i, x_j) = \phi_\xi(x_i)^T \phi_\xi(x_j) = \sum_{p=1}^m \xi_p^2 K_p(x_i, x_j).$$

(7)

By replacing the kernel matrix $K$ in (6) with $K_\xi$ calculated via (7), the multiple kernel k-means clustering problem can be re-rewritten as following form [37]:

$$\min_{H \in \mathbb{R}^{n \times k}, \xi \in \mathbb{R}_+^m} \text{Tr}(K_\xi(I_n - HH^T)), \text{ s.t. } H^T H = I_k, \quad \xi^T 1_m = 1.$$  

(8)

**D. PROPOSED MULTIPLE KERNEL K-MEANS CLUSTERING VIA LOCALITY PRESERVING MULTIPLE KERNEL ALIGNMENT**

Kernel alignment maximization has been widely used to learn kernel parameters in supervised learning. However, it is not suitable to unsupervised learning that the true labels are absent [46]. An effective solution is to update kernel coefficients by maximizing the alignment between the combined kernel $K_\xi$ and $HH^T$, where $H$ can be regarded as pseudo-labels in the last iteration [47]. In specific, the kernel alignment maximization for multiple kernel clustering can be formulated as following:

$$\max_{H \in \mathbb{R}^{n \times k}, \xi \in \mathbb{R}_+^m} \frac{\langle K_\xi, HH^T \rangle}{\sqrt{\langle K_\xi, K_\xi \rangle}} \text{ s.t. } H^T H = I_k, \quad \xi^T 1_m = 1,$$

(9)

where $\langle K_\xi, HH^T \rangle = \text{Tr}(K_\xi HH^T)$, $\{K_\xi, K_\xi\} = \xi^T M \xi$ with $\xi = [\xi_1, \xi_2, \ldots, \xi_m]^T$ and $M$ is a positive semi-definite matrix with $M_{pq} = \text{Tr}(K_p^T K_q)$ [46]. Directly optimizing (9) is difficult since it is a fourth-order fractional optimization.
problem. Thus, we need to derive a new and approximated optimization problem.

Following we use Theorem 2 to get a second-order upper bound for the denominator in (9).

**Lemma 1:** $\xi^T M \xi$ is an upper bound of $\tilde{\xi}^T M \tilde{\xi}$.

**Proof:** For a pair of semi-definite matrices $K_p$ and $K_q$, there exists matrices $U_p$ and $U_q$ such that $K_p = U_p U_p^T$ and $K_q = U_q U_q^T$. As a results, we have $M_{pq} = \text{Tr}(K_p K_q) = \text{Tr}(U_p U_p^T U_q U_q^T) = \text{Tr}((U_p^T U_q)(U_p^T U_q)^T) = ||U_p^T U_q||^2_F \geq 0$, where $|| \cdot ||_F$ denotes the Frobenius norm of a matrix. We also have $\xi^T M \xi = \sum_{p,q=1}^m M_{pq} \tilde{\xi}_p \tilde{\xi}_q \geq \sum_{p,q=1}^m M_{pq} \xi_p^2 \xi_q^2 = \tilde{\xi}^T M \tilde{\xi}$. This completes the proof.

$\xi^T M \xi$ is much easier to handle than $\tilde{\xi}^T M \tilde{\xi}$ since it leads to a well studied quadratic programming. In addition, this term also works as a regularization on the kernel coefficients to prevent $\xi_p$ and $\xi_q$ from being jointly assigned to a large weight if $M_{pq}$ is relatively high.

In addition, we find that minimizing the negative of numerator, i.e., $-\text{Tr}(K_q H H^T)$, together with $\xi^T M \xi$ simultaneously cannot guarantee that the whole objective is convex w.r.t $\xi$.
with fixed $H$. This would degenerate the quality of solution at each iteration, leading to sub-optimal solution. Here we use the following theorem to give a good substitute of $-\text{Tr}(K_{\xi}H H^T)$ while with convexity.

**Lemma 2:** $\text{Tr}(K_{\xi}(I_n - H H^T))$ is convex w.r.t $\xi$ with fixed $H$.

**Proof:** Since $H^T H = I_k$, we have $H H^T H = H$. By denoting $H = [h_1, h_2, \cdots, h_k]$, we can obtain that $H H^T h_c = h_c$, $\forall 1 \leq c \leq k$. This means $H H^T$ has $k$ eigenvalues with 1 and its rank is no more than $k$, which implies that it has $n - k$ eigenvalues with 0. As a consequence, $I_n - H H^T$ has $n - k$ and $k$ eigenvalues with 1 and 0. This induces $\text{Tr}(K_{\xi}(I_n - H H^T)) \geq 0$. With $\text{Tr}(K_{\xi}(I_n - H H^T)) = \sum_{p=1}^{m} \xi_p^2 \text{Tr}(K_p(I_n - H H^T))$, we can conclude that $\text{Tr}(K_{\xi}(I_n - H H^T))$ is convex w.r.t $\xi$ with fixed $H$.

According to the above-mentioned observations, the maximization problem described by (9) can be turned to following minimization problem:

$$
\min_{H \in \mathbb{R}^{n \times k}, \xi \in \mathbb{R}^m} \text{Tr}(K_{\xi}(I_n - H H^T)) + \frac{\lambda}{2} \xi^T M \xi
$$

s.t. $H^T H = I_k$, $\xi^T 1_m = 1$,  \hspace{1cm} (10)

where $\lambda$ is a parameter used to balance the two terms.
As can be seen from (9) and (10), they maximize the kernel alignment between the combined kernel matrices $K_\xi$ and the ideal kernel matrix $H H^T$ globally. In such a way, closer and farther sample pairs will be equally aligned to the same ideal similarity, intra-cluster variation of samples will be neglected. In other words, the discrimination between samples are not fully exploited. In addition, the closer samples are not constrained to share similar label vectors. Therefore, the locality of samples are not well preserved, which is a critical priori in unsupervised learning. In this article, we propose to locally align the similarity of each sample to its $k$-nearest neighbours with corresponding ideal kernel matrix rather than enforce the global alignment of all the samples, which is flexible and able to well handle the intra-cluster variations. In addition, for each sample, we construct a local Laplacian matrix to regularize that closer samples being allocated similar labels. Thus the locality of data samples can be well preserved.

If we use $K^{(i)}_\xi$ and $H^{(i)}$ to represent the sub-matrices of $K_\xi$ and $H$, and their indices are specified by the $\tau$-nearest neighbors of the $i$-th sample, then we have $K^{(i)}_\xi = S^{(i)^T} K_\xi S^{(i)}$ and $H^{(i)} = S^{(i)^T} H$, where $S^{(i)} \in \{0, 1\}^{n \times \tau}$ is a matrix indicating the $\tau$-nearest neighbors of the $i$-th sample. For each pair of sample, if they are close to each other, then their label vector should be also similar, this can be formulated as following:

$$
\min_{H \in \mathbb{R}^{n \times k}} \frac{1}{2} \sum_{i=1}^{n} \sum_{p=1}^{\tau} ||H^{(i)}_{p} - h^{(i)}_{p}||_2^2 [K^{(i)}_\xi]_{pq},
$$

FIGURE 7. ARI of our method w.r.t $\tau$ and $\lambda$ on different datasets ($\beta = 1$).
where \( h_p^{(i)} \) and \( h_q^{(i)} \) represent the \( p \)-th and \( q \)-th row of \( H^{(i)} \), respectively. (11) can be easily rewritten as the following trace form:

\[
\min_{H \in \mathbb{R}^{n \times k}} \sum_{i=1}^{n} \text{Tr}(H^{(i)T}L^{(i)}_{\tau}H^{(i)}),
\]

where \( L^{(i)}_{\tau} \) is the local Laplacian matrix of sample \( i \) with local similarity matrix \( K^{(i)}_{\xi} \).

Then we re-written (10) in a locally regularized form and combine it with (12) to induce our final LPKA model:

\[
\min_{H \in \mathbb{R}^{n \times k}, \xi \in \mathbb{R}^{m}} \sum_{i=1}^{n} \text{Tr}(K^{(i)}_{\xi}(I_{\tau} - H^{(i)T}H^{(i)})) + \frac{\lambda}{2} \xi^{T}M^{(i)}\xi + \beta \text{Tr}(H^{(i)T}L^{(i)}_{\tau}S^{(i)}H^{(i)})
\]

\[
s.t. \quad H^{T}H = I_{k}, \quad \xi^{T}1_{m} = 1,
\]

where \( I_{\tau} \) and is an identity matrix with size \( \tau \times \tau \). By defining \( A^{(i)} = S^{(i)}S^{(i)T} \), we obtain the objective function of our proposed LPKA:

\[
\min_{H \in \mathbb{R}^{n \times k}, \xi \in \mathbb{R}^{m}} \sum_{i=1}^{n} \text{Tr}(K_{\xi}(A^{(i)} - A^{(i)HHT}A^{(i)})) + \frac{\lambda}{2} \xi^{T}M^{(i)}\xi
\]

\[
+ \beta \text{Tr}(H^{(i)T}S^{(i)}L^{(i)}_{\tau}S^{(i)T}H)
\]

\[
s.t. \quad H^{T}H = I_{k}, \quad \xi^{T}1_{m} = 1
\]

1) OPTIMIZATION ALGORITHM

Note that (14) is not jointly convex with respect to \( H \) and \( \xi \), while it is convex to each variable if the other one is fixed. Thus, we design a two-step algorithm to solve this problem alternately.
Step 1: Updating $H$ with fixed $\xi$
When $\xi$ is fixed, $H$ can be obtained by solving the following optimization problem:

$$\max_{H \in \mathbb{R}^{n \times k}} \text{Tr}(H^T \sum_{i=1}^{n} (A^{(i)} K A^{(i)}) H)$$

$$- \beta \text{Tr}(H^T \sum_{i=1}^{n} S^{(i)T} L^{(i)} S^{(i)T} H)$$

$$s.t. \ H^T H = I_k, \ \ \ \ \xi^T 1_m = 1$$  \hspace{0.5cm} (15)

If we set $\Omega = \sum_{i=1}^{n} A^{(i)} K A^{(i)}$ and $\Theta = \sum_{i=1}^{n} S^{(i)T} L^{(i)} S^{(i)T}$, then (15) can be rewritten as:

$$\max_{H \in \mathbb{R}^{n \times k}} \text{Tr}(H^T (\Omega - \beta \Theta) H), \ \ s.t. \ H^T H = I_k,$$  \hspace{0.5cm} (16)

which is a standard kernel k-means clustering problem and can be efficiently solved.

Step 2: Updating $\xi$ with fixed $H$
Given fixed $H$, the optimization (14) w.r.t $\xi$ is a quadratic programming with linear constraints, which turns to the

FIGURE 9. Purity of our method w.r.t $\tau$ and $\beta$ on different datasets ($\lambda = 1$).
following problem:

$$\min_{\xi \in \mathbb{R}^m} \sum_{i=1}^{n} \frac{1}{2} \xi^T (2\Delta + \lambda \Psi) \xi \text{ s.t. } \xi^T 1_m = 1,$$

(17)

where $\Delta = \text{diag}(\text{Tr}(K_1 V), \cdots, \text{Tr}(K_m V))$, $V = \sum_{i=1}^{n} (A^{(i)} - A^{(i)} H H^T A^{(i)})$, and $\Psi_{pq} = \sum_{i=1}^{n} \text{Tr}(K_p A^{(i)} K_q A^{(i)})$.

In summary, our algorithm for solving (14) can be outlined in Algorithm 1. $\text{obj}^t$ denotes the objective value at the $t$-th iterations.

2) CONVERGENCE ANALYSIS

In Algorithm 1, the neighborhood of each sample is kept unchanged during the optimization. Specifically, the $\tau$-nearest neighbors of sample $i$ are measured by $K^\tau_i$. In such a manner, the objective value of Algorithm 1 is guaranteed to be monotonically decreased when updating one variable with the other fixed at each iteration. Meantime, the whole optimization problem is lower-bounded. As a sequence, the proposed algorithm can be guaranteed to be convergent. In order to empirically study the convergence of Algorithm 1, we show the variation of the objective values of Eq. (14) on different datasets.
datasets in Figure 1, which demonstrate that our proposed optimization algorithm is very efficient, i.e., the objective value is monotonically decreased and the algorithm quickly converges in less than five iterations.

III. RESULTS AND DISCUSSIONS

A. EXPERIMENT RESULTS

1) EXPERIMENTAL SETTINGS

To systematically evaluate the clustering performance of LPKA on the collected datasets, three metrics including Normalized Mutual Information (NMI) [48], Purity [49], and Adjusted Rand Index (ARI) [50] are used in our experiments. NMI and Purity take on values between 0 and 1, but ARI can be negative. The three metrics measure the concordance between the ground truth cell types and the cell types calculated by clustering algorithms, thus higher values indicate better performance for all metrics.

Given two clustering results $U$ and $V$ on a set of $N$ data points with $N_U$ and $N_V$ clusters, respectively, the mutual

FIGURE 11. NMI of our method w.r.t $\lambda$ and $\beta$ on different datasets ($\tau = 0.3$).
The information NMI is defined as

$$NMI(U,V) = \frac{\sum_{p=1}^{N_U} \sum_{q=1}^{N_V} |U_p \cap V_q| \log \frac{N|U_p \cap V_q|}{|U_p| \times |V_q|}}{\max \left( \sum_{p=1}^{N_U} U_p \log \frac{|U_p|}{N}, \sum_{q=1}^{N_V} V_q \log \frac{|V_q|}{N} \right)}$$

where the numerator is the mutual information between $U$ and $V$, and the denominator represents the entropy of the clustering $U$ and $V$.

For Purity, each identified cluster is assigned to the one which is most frequent in the cluster, and then the accuracy of this assignment is computed by counting the number of correctly assigned samples divided by the number $N$:

$$Purity(U,V) = \frac{\max_{p} |U_p \cap V_q|}{N}.$$  

The ARI depends on the following four quantities:

- $O_{uv}$, the number of objects in a pair that are placed in the same group in $U$ and $V$;

FIGURE 12. Purity of our method w.r.t $\lambda$ and $\beta$ on different datasets ($\tau = 0.3$).
• $O_u$, the number of objects in a pair that are placed in the same group in $U$ but in different groups in $V$;
• $O_v$, the number of objects in a pair that are placed in the same group in $V$ but in different groups in $U$;
• $O$, the number of objects in a pair that are placed in different groups in $U$ and $V$.

Then, ARI is defined as

$$ARI(U, V) = \frac{\binom{n}{2}(A) - [(B)(C) + (O_v + O)(O_u + O) - (O_{uv})]}{\binom{n}{2} - [(B)(C) + (O_v + O)(O_u + O)]}, \tag{20}$$

where $A = O_{uv} + O$, $B = O_{uv} + O_u$ and $C = O_{uv} + O_v$.

As can be seen from Eq. (14), there are three parameters in LPKA including $\lambda$, $\beta$ and $\tau$ need to be set. In our experiments, we tune all the parameters by a “grid-search” strategy. Specifically, $\lambda$ and $\beta$ are chosen from $\{0.01, 0.1, 1, 10, 100\}$, and $\tau$ is chosen from $\{0.3n, 0.4n, 0.5n\}$. Finally, the best clustering results are used for comparison.

2) COMPARED WITH OTHER METHODS
We compare the proposed LPKA with several existing methods, including PCA, traditional k-means, t-SNE [51], spectral clustering (“SC”), sparse spectral clustering (“SSC”) [52]
Algorithm 1 Optimization Algorithm for the Proposed LPKA

Input: Multiple kernel matrices \( \{K_p\}_{p=1}^{m} \), \( k \), \( \lambda \), \( \beta \) and a small positive constant \( \varepsilon \).

Initialization: \( \xi = 1/m \), and \( t = 1 \). Calculating \( S^{(i)} \) for the \( i \)-th sample \( (1 \leq i \leq n) \) by \( K_{\xi} \).

while not converged do
1. Update \( K_{\xi} \) by \( K_{\xi} = \sum_{p=1}^{m} (\xi_{p}^{2})K_{p} \);
2. Update \( H^{t} \) by solving Eq. (15);
3. Update \( \xi^{t} \) by solving Eq. (17);
4. \( t = t + 1 \)
5. Check convergence condition: \( (\text{obj}^{t-1} - \text{obj}^{t})/\text{obj}^{t} \leq \varepsilon \).
end while

Output: \( H \) and \( \xi \).

and SIMLR [40]. For all the compared methods, their parameters are turned carefully as suggested in their corresponding papers for fair comparison. We show the NMI, Purity and ARI of different clustering algorithms on different datasets in Figure 2, Figure 3 and Figure 4, respectively. As can be seen, the proposed LPKA has higher performance than all of other methods on the 9 datasets in terms of three different metrics, which demonstrates the superiority of LPKA. Therefore, our proposed LPKA can be used as a reliable pre-processing step to distinguish different cell types, which can reveal new insights into many other biological problems.

3) PARAMETER SENSITIVITY

In our experiment, we turn the parameters \( \lambda \), \( \beta \) and \( \tau \) to obtain the optimal results. To study the sensitivity of the LPKA with regard to the parameters in Eq. (14), we conduct experiments by fixing one of the three parameters and varying the other two. Firstly, we fix \( \beta = 1 \), and vary \( \lambda \) and \( \tau \). Figure 5-7 plot the values of the NMI, Purity and ARI, respectively, with fixed \( \beta \). Secondly, we fix \( \lambda = 1 \), and vary \( \tau \) and \( \beta \). Figure 8-10 plot the values of the NMI, Purity and ARI, respectively, with fixed \( \lambda \). Thirdly, we fix \( \tau = 0.4n \), and vary \( \lambda \) and \( \beta \). Figure 11-13 plot the values of the NMI, Purity and ARI, respectively, with fixed \( \tau \). As can be seen, the clustering results are robust with respect to the varying of \( \lambda \), while the results are a little sensitive to \( \beta \) and \( \tau \) to some extend, which demonstrate the significance of preserving the local structure of original data. We can obtain optimal clustering results with different combinations of the three parameters.

IV. CONCLUSION

In this work, we present a novel multiple kernel clustering framework for scRNA-seq data clustering via locality preserving kernel matrix alignment. A series of similarity kernel matrices are firstly generated by using different kernel functions. Then we transfer the clustering task to a multiple kernel k-means problem with the kernels aligned in a local manner. In order to preserve the local structure of the data for boosting final clustering performance, we construct a local Laplacian matrix for each sample to constrain that closer samples should be allocated similar labels. An alternate updating algorithm with theoretical analysis is developed to solve the proposed problem. Experiments with parameter sensitivity analysis on various real scRNA-seq data are conducted to demonstrate that our method can obtain superior results when compared with other state-of-the-art approaches.

ACKNOWLEDGMENT

(Xiao Zheng and Jiajia Chen contributed equally to this work.)

V. AVAILABILITY OF DATA AND MATERIALS

The implementation code and supporting files are available from http://tangchang.net/codes/LPKACode.zip

REFERENCES

[1] J. M. Raser and E. K. O’Shea, “Control of stochasticity in eukaryotic gene expression,” Science, vol. 304, no. 5678, pp. 1811–1814, Jun. 2004.
[2] T. Kalisky and S. R. Quake, “Single-cell genomics,” Nature Methods, vol. 8, no. 4, pp. 311–314, 2011.
[3] L. Pelkmans, “Using cell-to-cell variability—A new era in molecular biology,” Science, vol. 336, no. 6080, pp. 425–426, 2012.
[4] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, “A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data,” Comput. Methods Programs Biomed., vol. 166, pp. 99–105, Nov. 2018.
[5] D. Gokuluk, G. Zararsiz, S. Korkmaz, V. Elden, G. E. Zararsiz, E. Orcetin, A. Ozurk, and A. E. Karagargil, “MLSeq: Machine learning interface for RNA-sequencing data,” Comput. Methods Programs Biomed., vol. 175, pp. 223–231, Jul. 2019.
[6] Z. Xue, K. Huang, C. Cai, L. Cai, C.-Y. Jiang, Y. Feng, Z. Liu, Q. Zeng, L. Cheng, Y. E. Sun, J.-Y. Liu, S. Horvath, and G. Fan, “Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing,” Nature, vol. 500, no. 7464, pp. 593–597, Aug. 2013.
[7] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang, “Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells,” Nature Struct. Mol. Biol., vol. 20, no. 9, pp. 1131–1139, Sep. 2013.
[8] A. E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, “Single-cell RNA-seq: Advances and future challenges,” Nucleic Acids Res., vol. 42, no. 14, pp. 8845–8860, 2014.
[9] R. Kabir and R. Islam, “Chemical reaction optimization for RNA structure prediction,” Appl. Intell., vol. 49, no. 2, pp. 352–375, 2019.
[10] Y.-S. Sun, L. Ou-Yang, and D.-Q. Dai, “LRSK: A low-rank self-representation K-means method for clustering single-cell RNA-sequencing data,” Mol. Omics, vol. 16, no. 5, pp. 465–473, 2020.
[11] J. Hua, H. Liu, B. Zhang, and S. Jin, “Lak: Lasso and K-means based single-cell RNA-seq data clustering analysis,” IEEE Access, vol. 8, pp. 129679–129688, 2020.
[12] Y. Wu, Y. Guo, X. Xiao, and S. Lao, “AAE-SC: A scRNA-seq clustering framework based on adversarial autoencoder,” IEEE Access, vol. 8, pp. 178962–178975, 2020.
[13] F. Zhou, F. Ye, and L. Du, “Unsupervised robust multiple kernel learning via extracting local and global noises,” IEEE Access, vol. 7, pp. 34451–34461, 2019.
[14] C. Tang, X. Zhu, X. Liu, M. Li, P. Wang, C. Zhang, and L. Wang, “Learning a joint affinity graph for multiview subspace clustering,” IEEE Trans. Multimedia, vol. 21, no. 7, pp. 1724–1736, Jul. 2019.
[15] Z. Li, C. Tang, J. Chen, C. Wan, W. Yan, and X. Liu, “Diversity and consistency learning guided spectral embedding for multi-view clustering,” Neurocomputing, vol. 370, pp. 128–139, Dec. 2019.
[16] C. Tang, X. Liu, X. Zhu, E. Zhu, Z. Luo, L. Wang, and W. Gao, “CGD: Multi-view clustering via cross-view graph diffusion,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 5924–5931.
[17] R. Vidal, Y. Ma, and S. Sastry, “Generalized principal component analysis (GPCA),” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 12, pp. 1945–1959, Dec. 2005.
W. Wu and X. Ma, “Joint learning dimension reduction and clustering,” *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

E. W. Forgy, “Cluster analysis of multivariate data: Efficiency versus interpretability of classifications,” *Biometrics*, vol. 21, no. 3, pp. 41–52, 1965.

P. Brennecke, S. Anders, J. K. Kim, A. A. Kocodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler, “Accounting for technical noise in single-cell RNA-seq experiments,” *Nature Methods*, vol. 10, no. 11, pp. 1093–1095, Nov. 2013.

D. Grün, L. Kester, and A. van Oudenaarden, “Validation of noise models for single-cell transcriptomics,” *Nature Methods*, vol. 11, no. 6, pp. 637–640, Jun. 2014.

R. Bacher and C. Kendziorski, “Design and computational analysis of single-cell RNA-seq experiments,” *Genome Biol.*, vol. 17, no. 1, p. 63, Dec. 2015.

Y. Buganim, D. A. Faddah, A. W. Cheng, E. Isakovitch, S. Markoulaki, K. Ganz, S. L. Klemm, A. van Oudenaarden, and R. Jaensch, “Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase,” *Cell.*, vol. 150, no. 6, pp. 1209–1222, Sep. 2012.

T. Hashimshony, F. Wagner, N. Sher, and I. Yanai, “CEL-seq: Single-cell RNA-seq by multiplexed linear amplification,” *Cell Rep.*, vol. 2, no. 3, pp. 666–673, Sep. 2012.

V. Y. Kiselev, T. S. Andrews, and M. Hemberg, “Challenges in unsupervised clustering of single-cell RNA-seq data,” *Nature Rev. Genet.*, vol. 20, no. 5, pp. 273–282, May 2019.

H. Peng, X. Zeng, Y. Zhou, D. Zhang, R. Nussinov, and F. Cheng, “A component overlapping attribute clustering (COAC) algorithm for single-cell RNA sequencing data analysis and potential pathobiological implications,” *PLOS Comput. Biol.*, vol. 15, no. 2, Feb. 2019, Art. no. e1006772.

L. Sun, X.-Y. Zhang, Y.-H. Qian, J.-C. Xu, S.-G. Zhang, and Y. Tian, “Joint neighborhood entropy-based gene selection method with Fisher score for tumor classification,” *Appl. Intell.*, vol. 49, no. 4, pp. 1245–1259, 2019.

A. Zeisel, A. B. Muñozmachado, S. Codeluppi, P. Lönnnerberg, M. G. La, A. Jurés, S. Marques, H. Munguba, L. H., and C. Betsholtz, “Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq,” *Science*, vol. 347, no. 6226, pp. 1138–1142, 2015.

E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, and E. M. Martersteck, “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets,” *Cell.*, vol. 161, no. 5, pp. 1202–1214, 2015.

B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarssky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, and T. Dolbeare, “Adult mouse cortical cell taxonomy revealed by single cell transcriptomics,” *Nature Neurosci.*, vol. 19, no. 2, pp. 335–346, Feb. 2016.

M. Mojarad, S. Nejatian, H. Parvin, and M. Mohammadpoor, “A fuzzy clustering ensemble based on cluster clustering and iterative fusion of base clusters,” *Appl. Intell.*, vol. 49, no. 7, pp. 2567–2581, 2019.

L. Haghverdi, F. Buettner, and F. J. Theis, “Diffusion maps for high-dimensional single-cell analysis of differentiation data,” *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, Sep. 2015.

C. Shao and T. Höfer, “Robust classification of single-cell transcriptome data by nonnegative matrix factorization,” *Bioinformatics*, vol. 33, no. 2, pp. 235, 2016.

C. Xu and Z. Su, “Identification of cell types from single-cell transcriptomes using a novel clustering method,” *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, 2015.

S. Park and H. Zhao, “Spectral clustering based on learning similarity matrix,” *Bioinformatics*, vol. 34, no. 12, pp. 2069–2076, Jun. 2018.

W. Wu and X. Ma, “Joint learning dimension reduction and clustering of single-cell RNA-sequencing data,” *Bioinformatics*, vol. 36, no. 12, pp. 3825–3832, Jun. 2020.

A. A. Margolin, “Localized data fusion for kernel K-means clustering with application to cancer biology,” in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 1305–1313.

L. Du, P. Zhou, L. Shi, H. Wang, M. Fan, W. Wang, and Y. D. Shen, “Robust multiple kernel K-means using 121-norm,” in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 3476–3482.

Y. Lu, L. Wang, J. Lu, J. Yang, and C. Shen, “Multiple kernel clustering based on centered kernel alignment,” *Pattern Recognit.*, vol. 47, no. 11, pp. 3656–3664, Nov. 2014.

B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, “Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning,” *Nature Methods*, vol. 14, no. 4, p. 414, 2017.

R. Qi, J. Wu, F. Guo, L. Xu, and Q. Zou, “A spectral clustering with self-weighted multiple kernel learning method for single-cell RNA-seq data,” *Briefings Bioinf.*, Oct. 2020, doi: 10.1093/bib/bbaa216.

B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake, “Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq,” *Nature*, vol. 509, no. 7500, pp. 371–375, May 2014.

Q. Deng, D. Ramskold, B. Reinius, and R. Sandberg, “Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells,” *Science*, vol. 343, no. 6167, pp. 193–196, Jan. 2014.

F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, “Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells,” *Nature Biotechnol.*, vol. 33, no. 2, pp. 60–155, 2015.

B. Zhao, J. T. Kwik, and C. Zhang, “Multiple kernel clustering,” in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2009, pp. 638–649.

C. Cortes, M. Mohri, and A. Rostamizadeh, “Algorithms for learning kernels based on centered alignment,” *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 795–828, Jan. 2012.

X. Liu, Y. Dou, J. Yin, L. Wang, and E. Zhu, “Multiple kernel K-means clustering with matrix-induced regularization,” in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1888–1894.

A. Strehl and J. Ghosh, “Cluster ensembles—A knowledge reuse framework for combining multiple partitions,” *J. Mach. Learn. Res.*, vol. 3, no. 12, pp. 583–617, 2003.

S. Wagner and D. Wagner, *Comparing Clusterings: An Overview*. Karlsruhe, Germany: Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

S. Wagner and D. Wagner, “Comparing clusterings—An overview,” in *Analysis*, vol. 4769, 2007, pp. 1–19.

L. van der Maaten, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

C. Lu, S. Yan, and Z. Liu, “Convex sparse spectral clustering: Single-view to multi-view,” *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2833–2843, Jun. 2016.

XIAO ZHENG received the M.Med. degree from Tianjin Medical University, Tianjin, China, in 2014. She is currently pursuing the Ph.D. degree with the Department of Pharmacy, National University of Defense Technology. Her current research interests include machine learning and bioinformatics.

JIAJIA CHEN received the M. Med. degree from the School of Pharmacy, Nanjing Medical University, in 2013. She is currently a Pharmacist with the Department of Pharmacy, Huai’an Second People’s Hospital Affiliated, Xuzhou Medical College. Her research interest includes medical data analysis.
CHANG TANG (Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2016. From September 2014 to September 2015, he was with the AMRL Lab, University of Wollongong. He is currently an Associate Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. He has published more than 50 peer-reviewed articles, including those in highly regarded journals and conferences such as IEEE T-PAMI, IEEE T-MM, IEEE T-KDE, IEEE T-HMS, IEEE SPL, ICCV, CVPR, IJCAI, AAAI, and ACMM. His current research interests include machine learning and data mining. He served on the Technical Program Committees for NIPS 2020, IJCAI 2018/2019/2020, ICME 2018/2019/2020, AAAI 2019/2020, IJCAI 2018/2019/2020, CVPR 2019/2020, and ICCV 2019. He serves as an Associate Editor for BMC Bioinformatics and BioMed Research International.

SUQIN ZHOU is currently a Pharmacist with the Department of Pharmacy, People’s Hospital of Lian’shui County, Nanjing Medical University. Her current research interests include medical data analysis for disease diagnosis and drug data analysis.

* * *