Briefings in Bioinformatics, 2022, **23(2)**, 1–17

https://doi.org/10.1093/bib/bbab584

Case Study

Vec2image: an explainable artificial intelligence model for the feature representation and classification of high-dimensional biological data by vector-to-image conversion

Hui Tang, Xiangtian Yu, Rui Liu and Tao Zeng

Corresponding authors: Tao Zeng, Tel.: +86-18721274525; E-mail: zengtao@sibs.ac.cn or zeng_tao@gzlab.ac.cn; Rui Liu, Tel.: +86-13016098877; Fax: 0086-20-87110448; Email: scliurui@scut.edu.cn

Abstract

Feature representation and discriminative learning are proven models and technologies in artificial intelligence fields; however, major challenges for machine learning on large biological datasets are learning an effective model with mechanistical explanation on the model determination and prediction. To satisfy such demands, we developed Vec2image, an explainable convolutional neural network framework for characterizing the feature engineering, feature selection and classifier training that is mainly based on the collaboration of principal component coordinate conversion, deep residual neural networks and embedded k-nearest neighbor representation on pseudo images of high-dimensional biological data, where the pseudo images represent feature measurements and feature associations simultaneously. Vec2image has achieved better performance compared with other popular methods and illustrated its efficiency on feature selection in cell marker identification from tissue-specific single-cell datasets. In particular, in a case study on type 2 diabetes (T2D) by multiple human islet scRNA-seq datasets, Vec2image first displayed robust performance on T2D classification model building across different datasets, then a specific Vec2image model was trained to accurately recognize the cell state and efficiently rank feature genes relevant to T2D which uncovered potential T2D cellular pathogenesis; and next the cell activity changes, cell composition imbalances and cell–cell communication dysfunctions were associated to our finding T2D feature genes from both population-shared and individual-specific perspectives. Collectively, Vec2image is a new and efficient explainable artificial intelligence methodology that can be widely applied in human-readable classification and prediction on the basis of pseudo image representation of biological deep sequencing data.

Keywords: explainable artificial intelligence, deep residual neural network, classification, feature selection, single-cell sequencing, type 2 diabetes

Introduction

With the development of high-throughput sequencing technology, an abundance of biological datasets is accessible for exploration. Currently, there are many essential studies and application tasks involved in the convergence of biotechnology (BT) and information technology, which concerns the classification and prediction of biological samples on the basis of high-dimensional omics data [1]. For example, to recognize single cell types, many approaches involve unsupervised methods [2–4], which are usually followed by manual annotation based on previously identified marker genes [5]. However, the performance of these methods is unsatisfactory, with low sensitivity in sample/cell classification when integrating multiple datasets for follow-up analysis, and such annotation of samples/cells is also relatively subjective, which may introduce classification bias in the downstream analysis. Thus, an explainable computational model (e.g. artificial intelligence [AI]) is needed to help nonexperts correctly understand and obtain constructive inspiration from the data-intensive biological or biomedical studies on the basis of high-dimensional biological data (e.g. high-throughput deep-sequencing data) [6].

On one hand, in recent biomedical fields, the well-annotated biological sequencing datasets have become increasingly available, and machine learning (ML) approaches have shown great capability in prediction and classification [7, 8], such as xGBoost [9], Adaboost

Hui Tang is a postdoctoral fellow at the School of Mathematics, South China University of Technology. She is interested in bioinformatics, machine learning and data analysis.

Xiangtian Yu has a PhD in mathematics. She is an assistant professor in Clinical Research Center, Shanghai Jiao Tong University Affiliated Sixth People's Hospital. Her research interest is clinical medical data analysis and biostatistic.

Rui Liu is a full professor at the School of Mathematics, South China University of Technology. He received the BS and PhD degrees in applied mathematics from Peking University. His research interest includes nonlinear dynamics, modeling and computational methods.

Tao Zeng received BS, MS and PhD degrees from Wuhan University, Wuhan, China, in 2003, 2006 and 2010, respectively. Since 2013, he has been an associate professor in Key Laboratory of Systems Biology, SIBS, CAS and now in Guangzhou Laboratory, Guangzhou, China. His research interests include machine learning, bioinformatics, network biology, computational biology and graph theory.

Received: September 14, 2021. **Revised:** December 6, 2021. **Accepted:** December 20, 2021

[10] and random forest [11]. These algorithms and well-trained models have been widely applied [8, 12–14] and can automate the identification of sample annotations/types and reduce the bias existing in prior-given information [15]. In fact, most current methods are developed for a biological data sample in the form of a feature vector and then annotate/classify the vector into one of the defined types/classes. However, the features in the vector are generally selected and considered mutually independent by both the algorithm developer and user, so that they usually tend to ignore any (explainable) relationships between features [14, 16, 17] which leads to the contradiction as a 'black box' problem in computation. In addition, these models would require manual parameter adjustment, which actually aggravates the burden of the inexperienced user in understanding the model execution and outcome.

On the other hand, convolutional neural networks (CNNs) have recently gained wide attention due to their high performance in the recognition/classification of image data in the form of pixel matrices [18–21]. In brief, CNN has shown several computational advantages, including the feature extraction and classification via hidden layers; the detection of nonlinear correlations and high-order statistics of image data and the deep layer and fewer parameters of the network structure with weight sharing. However, the application of CNNs in the clinical scenarios is still limited in terms of medical images [20, 21]. This method is also difficult to directly apply to diverse biological sequencing data, such as genomic, transcriptomic and metabolomic data [22]. The predictions by using CNNs or similar neural network models are hard to interpret while they can provide excellent performance in practice, which are thought to lack explicit representations and direct consequent interpretation due to the lower meaning of a single pixel/feature, thus inevitably facing the 'black box' problem. Although a few studies has recently realized the transformation of omics data into a well-organized image form [12, 23–25], they focused more on the classification accuracy for algorithm developers rather than the outcome interpretation for inexperienced users. Thus, a model and algorithm for applying explainable artificial intelligence (xAI) [6, 26] in biological and biomedical studies are still lacking and need fundamental developments.

To address these practical challenges during AI application in biological big data, we proposed Vec2image, a xAI model for analyzing high-throughput (vector) data (Figure 1), whose framework is mainly based on principal component coordinate conversion, deep residual neural network learning and embedded k-nearest neighbor (KNN) representation. Vec2image is able to integratively execute feature engineering, feature selection, classifier learning and topic recommendation in the biological context. Technically, the workflow of Vec2image includes three steps: (i) adopting synthetic minority oversampling technique (SMOTE) to address class imbalances that occur in most biological sample imbalance or small sample sequencing datasets (Figure 1**A**); (ii) constructing a feature correlation/co-ordinate structure in a latent feature space using alternative feature extraction methods to transform the original vector data into the pseudo image/matrix, which contains both feature abundance and correlation information (Figure 1**B**) and (iii) implementing an interpretable CNN classifier for sample classification, feature/topic recommendation and insightful sample/feature association visualization simultaneously (Figure 1**C**), which is on the basis of paralleled ResNet architecture (Figure 1**D**) and KNN algorithm (Figure 1**E**), respectively. In these pseudo image of biological omics data, the single pixel/feature in image should represent the individual gene with strong expression signal and the pixel shape in image could indicate the grouped genes with tight expression correlation. This information can be simultaneously extracted by CNN or similar technology so as to provide abundant gene expression and co-expression pattern supporting classification model with biological interpretation. These capabilities of Vec2image have been validated on multiple benchmark datasets, which illustrates the accurate prediction in many biological learning tasks.

In particular, Vec2image has been applied to make a deep case study on type 2 diabetes (T2D) by multiple human islet Single-cell RNA sequencing (scRNA-seq) datasets at cross-domain level e.g. using single-cell sequencing data from separate cells (individual domain) and additional bulk sequencing data from whole tissues (population domain). In this case study, Vec2image achieved satisfactory performances in binary classification, multiclass classification, imbalance classification and cross-platform (transfer) classification on different kinds of high-throughput bulk and single-cell T2D sequencing datasets. As a candidate xAI model, Vec2image helped to identify cell-type-enriched and disease-related genes, which identified previously unknown diabetes-related genes enriched for alpha or beta cells. These findings of Vec2image provided deeper insights into the cell activity changes, cell composition imbalances and cell–cell communication dysfunctions that occur in pancreatic cells involved in diabetes. In summary, Vec2image indeed provides a comprehensive and widely applicable xAI model for analyzing diverse high-dimensional biological data.

Material and methods

Sample balance of Vec2image by oversampling

Vec2image combines SMOTE sampling, providing an efficient method for improving classification performance when trained on imbalanced data [27]. New sample data *D* are sampled from the given training data (Figure 1**A**), whose processing algorithm is as follows:

$$\begin{cases} \left(x_{new}^{1}, y^{1}\right) = \left(x_{raw}^{1} + \mathbf{rand}\left(0,1\right)\left(x_{rand}^{1} - x_{raw}^{1}\right), y^{1}\right) \\ \left(x_{new}^{2}, y^{2}\right) = \left(x_{raw}^{2} + \mathbf{rand}\left(0,1\right)\left(x_{rand}^{2} - x_{raw}^{2}\right), y^{2}\right) \\ \qquad\qquad\qquad\qquad \cdots \\ \left(x_{new}^{n}, y^{m}\right) = \left(x_{raw}^{m} + \mathbf{rand}\left(0,1\right)\left(x_{rand}^{m} - x_{raw}^{m}\right), y^{m}\right) \end{cases}$$

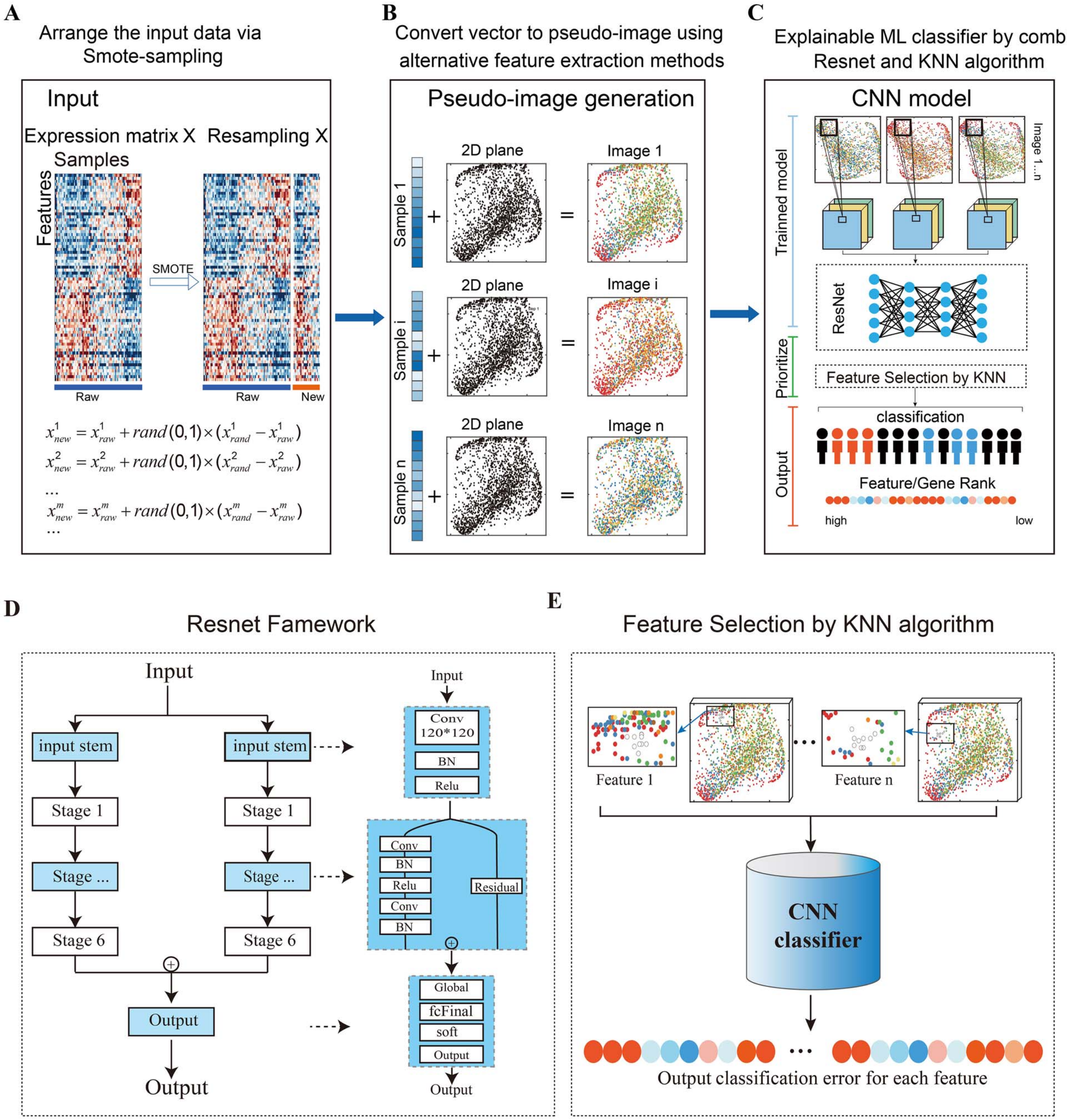

Figure 1. Overview of the Vec2image framework. (**A**) Preprocessing of input sample data with balance sampling in the first stage of Vec2image. (**B**) Transformation from original vector sample data to pseudo image sample data in the second stage of Vec2image, where the color of each dot represents the normalized expression value of each feature of one sample. (**C**) Schematic of the internal neural network structure of the classification model in the third stage of Vec2image, where sample classification and features/genes prioritization can be carried out via paralleled ResNet architecture combining with embedded KNN algorithm. (**D**) Parallel Resnet architecture, which consists of two parallel CNN architectures where each consists of six Resnet blocks. (**E**) Feature prioritization based on KNN.

SMOTE generates synthetic minority samples to oversample the minority class. For every x^i_{raw} in the minority class, we calculate its K (which is set to 5 by default) nearest neighbors from the whole training set, and then x^i_{rand} is randomly selected from them. After that, we can calculate a new synthetic sample x^i_{new} along the line between the minority example and its selected nearest neighbors. Therefore, the new balanced training data $D = (x_1, y_1), \ldots, (x_n, y_n)$ are generated, where $(x_i,\ y_i)$ represents a sample and its feature vector is x_i and its class variable is y_i.

Pseudo image generation of Vec2image by dimensionality reduction

Sequencing data of biological samples, which are inherently in numerical form, were transformed to an image form for them to be compatible with CNN architectures. First, the nonlinear dimensionality reduction methods (t-distributed stochastic neighbor embedding [t-SNE] [28] as the default algorithm) are used to transform the feature points from the observed data space to the latent data space, and meanwhile obtain the visualizable feature points in latent space with dimension as 2. Considering a training set D, the general sample points of D consist of n samples and m features, which can be defined as $X = (x_1, x_2, \ldots, x_n)$ and $\mathrm{x_i}$ $(i = 1, 2, \ldots, n)$ is a vector with values of m entries/features; in contrast, the feature points of D can be represented as $G = \{g_1, g_2, \ldots, g_m\}$, where g_j $(j = 1, 2, \ldots, m)$ is a vector with values of n entries/samples. Obviously, in the matrix form, G is just a transpose matrix of X. And here, the feature point matrix G is processed by t-SNE to get a 2D coordinates $\{(a_1, b_1), (a_2, b_2), \ldots, (a_\mathrm{m}, b_\mathrm{m})\}$ of features and the coordinates (a_j, b_j) in the 2D plane define the location of features/genes g_j in latent space. These nonlinear dimensionality reduction algorithms keep the high-dimensional topology of features to a 2D plane and preserve the local relationships among features in accordance with the similarity among features in high-dimensional space; therefore, the association structure of features in 2D plane can be determined. To avoid potential overfitting, the test of new datasets will directly use the feature locations (i.e. association structure of features) obtained from the training set in this work. Of note, the other data projections (e.g. Kernel Principal Component Analysis (KPCA), Uniform Manifold Approximation and Projection (UMAP)) [29, 30] were also implemented as additional pseudo image generators. The different dimensionality reduction techniques would affect downstream analysis although they might have comparable performances (Supplementary Figure S1), thus these projections (i.e. tSNE, KPCA, UMAP) are currently all available for custom selection during model training according to corresponding applications.

Next, a rotation is performed by a convex hull algorithm to frame the pseudo image (i.e. 2D plane structure) horizontally or vertically for input into the following CNN architecture [12, 31], and the Cartesian coordinates are converted to pixels, which determine the location of each feature in pixel frame.

Finally, the expression values of each sample within vector $\mathrm{x_i}$ are mapped to these pixel locations as detailed pixel values. The color shade of a single layer of an image ranges from 0 to 1. In this work, the pixel frame size is set to 120 × 120 by default, and the default value of each pixel is set to 1. Each $\mathrm{x_i}$ is normalized by the following manner:

$$step1 : \mathrm{x_i} = \log\left(\mathrm{x_i} + \left|\min_{samples} \mathrm{x_i}\right| + 1\right)$$

$$step\ 2 : \mathrm{x_i} = \frac{\mathrm{x_i}}{M}$$

The minimum value $\min\limits_{samples} \mathrm{x_i}$ is adjusted for each feature across samples, and the global maximum value $M = \max D$ of whole training data D is used to further scale the feature values between 0 and 1. Obviously, this normalization will retain the association topology of all features. Of note, if more than one feature acquires the same location in the pixel frame, then these features will be placed in the same location, and their average expression values will be given as the final value of this pixel. Therefore, unique pseudo image data $\mathrm{z_i}$ for each vector $\mathrm{x_i}$ of one sample i can be generated.

Through this pseudo image generation, n samples can be measured by the transformed n images, which would be directly plugged into following CNN architecture.

CNN model configuration of Vec2image by residual network

The above transformed pseudo image data $\mathrm{z_i}$ instead of original vector data $\mathrm{x_i}$ can be processed by CNN architecture (Figure 1**C**). The CNN models were constructed using a parallel ResNet architecture (Figure 1**D**). In brief, this architecture contains six residual building blocks in parallel. A convolution layers, a batch normalization layer and a ReLU activation layer in turn followed the input layer, which have the following implementations.

(1) The convolution layer is applied in the follow-up operation, where W is the learnable weight and b is a shift variable.

$$f_C(\mathrm{z_i}) = \mathrm{W}\mathrm{z_i} + b$$

(2) The batch normalization layer is calculated as below, where μ_D and σ_D represent the mean and variance of each channel data over all observation dimensions; ε is a constant improving numerical stability when facing a very small variance and α and β are learnable transform parameters and are updated in the network training process. Of note, after training, the training data will be passed through once more, and the fixed mean and variance will be determined (i.e. μ_D and σ_D) on whole data for prediction.

$$f_N(\mathrm{z_i}) = \alpha \frac{\mathrm{z_i} - \mu_D}{\sqrt{\sigma_D^2 + \varepsilon}} + \beta$$

(3) The ReLU activation layer is computed as below.

$$f_A(\mathrm{z_i}) = \begin{cases} \mathrm{z_i}, \mathrm{z_i} \geq 0 \\ 0, \mathrm{z_i} < 0 \end{cases}$$

Six building blocks in parallel followed the above ReLU activation layer, where each block consisted of two convolution layers, a batch normalization layer and two ReLU activation layers, and the outputs of the batch normalization layer of the sixth residual blocks were combined and fed to a global pooling layer. Finally, the SoftMax layer following the fully connected layer gives the final classification output. The architecture does not have much hyperparameters; in addition, the

'MaxEpochs' is fixed at 100 by default, and the adaptive moment estimation optimizer (Adam) is used in the current Vec2image implementation.

Interpretable informative feature ranking of Vec2image by embedded KNN representation

When classifier training is completed, Vec2image further uses the KNN algorithm embedding latent space to optimize feature selection. Using KNN, Vec2image resamples the input pseudo images, choosing the K-nearest features in the latent space for each feature (K = 10 by default), setting 1 of these features' shades for each sample and recomputing the classification error, which can indicate the impact of each feature on classification model (Figure 1**E**). By such a feature mask strategy, the rank of the importance/impact of each feature can be inferred dependent on their influence on classification errors, holding the assumption that the mask of the best features will cause the highest increase in classification error. The classification error for each feature reflects its impact on the overall model predictions, thus the features would be informative features when its shade causes the increase of classification error.

We also tested the threshold K of KNN with a range from 5 to 20 in multiple single-cell datasets and found that selection of informative features (e.g. informative feature genes [IFGs]) was consistent for different K values (Supplementary Figure S2).

Model learning configuration of Vec2image

For each dataset used for method evaluation and comparison, we drew a training sample *Z* from dataset *D*, uniformly at random without replacement of each category. By default, the size of *Z* was set to 80% that of *D*, and the remaining 20% of the dataset was used as the test set. And again, the 80% samples of *Z* were used for model learning, and 20% samples of *Z* were used for model validation. In another word, as a typical strategy for overcoming the potential overfitting issues in ML model building, there are training dataset (64% samples), validation dataset (16% samples) and test datasets (20%) randomly used in this study.

Comparison and evaluation of Vec2image

To comprehensively evaluate the Vec2image model, we have compared Vec2image with seven kinds of models, whose detail model characteristics and parameter settings can be seen in Supplementary Table S1. (i) xGBoost: the xgboost R package was used with the default parameters. (ii) DeepInsight: the public MATLAB source code was used, which is available from http://www.riken.jp/en/research/labs/ims/med_sci_math/ or http://www.alok-ai-lab.com. (iii, iv) RUSboost and Adaboost: The fitcensemble() MATLAB function was used with 50 learners. (v) Random forest: The treebagger() MATLAB function was used with 20 classification trees. (vi) SVM (support vector machines): The fitcsvm() MATLAB function was applied with default parameters. (vii) scClassify: The scClassify R package was carried out, which is available from http://github.com/SydneyBioX/scClassify.

Meanwhile, there were four categories of computational tasks and benchmark datasets for method comparison and evaluation. (a) For the binary classification task, there were five datasets, including one single-cell dataset (bladder), one vowel dataset (PD), one mass-spectrometry dataset (Arcene) and two artificial datasets (Ringnorm-DELVE and Medelon). Here, the raw data matrices were used as the same inputs to Vec2image and all other methods, and different methods would have their own preprocession on the input data (e.g. pseudo image transformation by Vec2image and deepInsight, or log normalization by scClassify, as listed in Supplementary Table S1). The AUC measurements with 10 times running on resampling data were used to measure the performance of each binary classification model and also other kind of models in the following discussions. (b) For the imbalance classification task, the proportion of two sample categories in the above binary class datasets was set to 1/50, 1/20, 1/10, 1/5 and 9/10 to generate multiple imbalanced datasets by random sampling without replacement. (c) For the multiclass classification task, 20 organ-specific single-cell datasets from the Tabula Muris were used, and the accuracy and kappa measurements were applied to evaluate all methods. For the Tabula Muris data, we have used the ROGUE [32] to identify genes that are most relevant for biological heterogeneity before model training, and such data after feature selection were used as the same inputs to Vec2image and all other methods. Noted that, scClassify is originally a multiscale classification method for single-cell data, and it is especially used in the comparison of this multiclass classification of single-cell data as a key baseline method, where the weighted KNN + Pearson +DE was set when using scClassify. (d) For the cross-platform classification task, four human pancreas datasets (GSE86469, E-MTAB-5061, GSE84133 and GSE83139; Supplementary Table S4) generated from three platforms (Illumina NextSeq 500, Illumina HiSeq 2000 and Illumina HiSeq 2500) by independent studies were combined together to assess the Vec2image performance and further applied in a deep case study, where all overlapping genes in four datasets were used to training models.

Assessment of IFGs identified by Vec2image

As a baseline, the Seurat R package [33] and SCCAF (Single Cell Clustering Assessment Framework) package [34] were used to identify the candidate marker genes of cell clusters/types in each single-cell dataset from the Tabula Muris. The 1000 top-ranked IFGs selected by Vec2image were used for consistency evaluation with these candidates. The functional enrichment analysis of candidate marker genes and IFGs was carried out by hypergeometric test (adjusted *P*-value: **0.05/25). To evaluate the impact of IFGs on CNN performance, several subgroup datasets of each single-cell dataset were constructed: namely, a group including IFGs (IFG+), a group including non-IFGs (IFG−) and a group including randomly selected genes (shuffle), with gene proportions of 5%, 10%, 15% and 20%, respectively. The accuracy was calculated to compare the performance in the classification of IFG+, IFG− and shuffle for each single-cell dataset, where the classifier used was the independent xGBoost approach.

In addition, some popular feature selection methods, such as the S-E model [32], were also used to compare and evaluate IFGs. S-E model is an entropy-based feature selection method that can detect variable genes by hypothesis testing. On the integrated human pancreas single-cell datasets, the 1000 selected top-ranked IFGs by Vec2image and the S-E model, respectively, were used to test their contribution to the following cell clustering performance, where the t-SNE and adjust rand index (ARI) were used to qualitatively visualize and quantitatively evaluate the clustering accuracy based on the selected feature genes by Vec2image and S-E model, respectively.

Case study of T2D by Vec2image

Data preprocessing

The three integrated pancreas single-cell datasets (E-MTAB-5061, GSE84133, GSE83139) were used to train classifier by Vec2image. Meanwhile, another pancreas dataset (GEO (Gene Expression Omnibus) accession GSE86469) was used for independent validation and investigation, which includes eight human cadaveric organ donors, and each islet sample from one donor was processed to generate scRNA-seq libraries and paired bulk RNA sequencing (RNA-seq) libraries at three different stages of islet processing (baseline, intact and dissociated). Considering the requirement of sample size and statistical power, three bulk datasets from different stages were merged for further analysis.

Execution of Vec2image

We trained a Vec2image classifier based on the integrated single-cell dataset to classify cells from the nondiabetic (ND) and T2D states, respectively. The proportion of training and validation was set to 80:20. The dimensional reduction technique t-SNE was used to obtain 2D plane, the pixel size was set 120×120 and 'MaxEpochs' was fixed in 300.

Filtering of IFGs

To increase the statistical power, we used paired three groups from merged samples with similar age (ND 30 – T2D 42; ND 53 – T2D 51; ND 56 – T2D 55) to prioritize IFGs, respectively. The overlapping 373 genes of top 1000 IFGs from three groups are considered as important IFGs in distinguishing T2D and ND cells/states. The threshold K of KNN was set 10.

Deconvolution of bulk gene expression data

To evaluate cell-type-specific differences between ND and T2D individuals, we used the merged bulk RNA-seq dataset (GEO accession GSE86469) which contains gene expression profiles of 24 samples, 9 of which are identified as diabetic and 15 of which are healthy controls. This resulted in a gene expression matrix of 25,169 genes and 24 samples. Based on single-cell dataset and bulk dataset of GSE86469, the Bseq-SC [35] was used to estimate the proportions of cell types for each bulk sample. And the proportions of alpha and beta cells in each bulk sample were further used for the cell-type-specific analysis, which were two main cell types involved in T2D and their proportion distribution would be most variable between ND and T2D subjects.

Estimation of cell activity

AUCell [36] was used to estimate the cell activity based on IFGs in the ND and T2D cells. The input to AUCell is the single cell expression matrix of gene set with 373 IFGs, and the output is the cell activity matrix (i.e. gene set 'activity') of all ND and T2D cells.

Reconstruction of intercellular communication networks

Cellchat [37] was used to further investigate the potential role of IFGs in cell–cell communication. The IFGs' expression matrix and the whole gene expression matrix from single-cell dataset were used as the input of CellChat, respectively, which models the probability of cell–cell communication by integrating gene expression with prior knowledge of receptors, their cofactors and the interaction between signaling ligands.

Ligand-target analysis

NicheNet [38] was applied to ligand-target prediction between alpha and beta cells. As potential active ligands, we consider all ligands that were expressed in alpha cells. As the target gene set, we consider the list of IFGs significantly differentially expressed in beta cells between nondiabetic sample (NDS) and type 2 diabetes sample (T2DS), and we consider all genes expressed in beta cells as background. NicheNet firstly prioritized the most probable ligands regulating IFGs by ligand activity prediction, and then assessed the ligands in alpha cell regulating the IFGs in beta cells.

Statistics analysis

The ROC (receiver operating characteristic) curve was used to illustrate the overall performances of different modeling methods. The decision curves analysis (DCA) was used to test the clinical usefulness of different models in prediction. About 10-fold cross-validation was performed to calculate the Cohen's kappa coefficient using the K-fold cross-validation routine of the R and MATLAB. The functional enrichment analysis between IFGs and markers was carried out by hypergeometric test (adjusted *P*-value: ** 0.05/25). Generally, statistical significance was assessed by unpaired *t*-test. Significance levels were set to $P = 0.05$. Significance for comparisons: *$P < 0.05$; **$P < 0.01$.

Enrichment analysis

Gene Ontology functional analysis was performed by using g:Profiler [39] and Pathway Commons [40].

Availability of data

The Ringnorm dataset [41] is available from University of Toronto at https://www.cs.toronto.edu/&#x007E;delve/data/ringnorm/desc.html. The scRNA-seq data of bladder are available from the Tabula Muris [42]. Madelon dataset [43] is available at UCI repository http://archive.ics.uci.edu/ml/datasets/madelon. The Arcene dataset [43] is available at http://archive.ics.uci.edu/ml/datasets/Arcene. PD disease [44] is available at http://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification. The Tabula Muris mouse data were downloaded from https://tabula-muris.ds.czbiohub.org/. The accession numbers for the pancreas datasets in NCBI GEO are GSE86469 [45], GSE84133 [35], GSE83139 [46], and that of a dataset in ArrayExpress (EBI) is E-MTAB-5061 [47].

Results

Accurate and robust performance of Vec2image in the classification of heterogeneous biological samples

Evaluation on binary classification

First of all, Vec2image was evaluated on a binary classification task with a collection of five different kinds of datasets (Supplementary Table S2), including a single-cell dataset (bladder), a vowel dataset (Parkinson's disease), a mass-spectrometry dataset (Arcene) and two artificial datasets (Ringnorm-DELVE and Medelon), which have been widely used in previous studies. As evaluation controls, six popular supervised classifiers in binary classification, including Adaboost, DeepInsight, Random Forest, RUSboost, SVM and xGBoost [9, 12, 48–51], were used to provide comparable baselines. The accuracy of each model on each dataset is summarized in Figure 2**A**, and the ROC curves and DCA curves were also compared on the five datasets in Figure 2**C** and **D**. In addition, we also evaluated the efficiency of CNN models training on image matrix data, and original data vector, and latent data vector after dimension reduction for samples, respectively, where the image-based CNN model still showed better performance than nonimage-based CNN models (Supplementary Figure S3), indicating the contribution of feature structures in addition to feature values in our produced pseudo images. These results demonstrate that Vec2image robustly led to higher accuracy and would bring significant benefit to clinical application than other models on the binary classification task in most datasets.

Evaluation on unbalanced classification

Next, Vec2image was evaluated in the above datasets with different sample imbalance settings, where the same five benchmark datasets were randomly split to form unbalanced training datasets, and the balance proportions of the two sample classes were 2%, 5%,10% 20% and 45%. As shown in Figure 2**B**, Vec2image still significantly outperformed the other classifiers. In particular, approaches such as boosting alone, boosting with undersampling and even extreme gradient boosting did not achieve the highest performance, and the most pronounced performance differences among the different methods were indeed observed when the training set had a balance proportion of 2% (Figure 2**B**). Therefore, Vec2image is able to achieve satisfactory performance even with imbalanced class labels or rare biological sample types.

Evaluation on multiclass classification

Then, we evaluated the accuracy and robustness of Vec2image for cell type classification [5, 52], which is usually a multiclass classification task, using the Tabula Muris datasets including 20 well-organized organ-specific scRNA-seq datasets (Supplementary Table S3). We have compared the effectiveness of Vec2image with the six abovementioned methods and an advanced single-cell-specific supervised learning method (i.e. scClassify) [13]. According to performance measurements (i.e. accuracy and kappa coefficient), Vec2image outperformed many existing methods again during dissimilar learning tasks with significant performance promotion (Figure 3**A** and **B**). Of note, Vec2image still has larger average performance than xGBboost in this condition although it is not significant enough, and thus the model evaluation on more complicated datasets would uncover more pros and cons of different methods in future.

Evaluation on selected features

In addition, we carried out an assessment to compare the IFGs selected by Vec2image and the marker genes selected by two popular cell marker identification methods Seurat (Figure 3**C** and Supplementary Figure S4) and SCCAF (Supplementary Figure S5**A** and **B**) [33, 34, 53] (see the 'Materials and Methods' section), on above single-cell sequencing samples. Indeed, the IFGs selected by Vec2image have satisfactory efficiency and high consistency with those candidate cell marker genes (Figure 3**C** and Supplementary Figures S5 and S6). The significant test (Supplementary Figure S5**C**) and hypergeometric test (Supplementary Table S3) are also performed to verify the effectiveness of feature selection of Vec2image. These results indicated that IFGs have significant enrichments in candidate cell markers. Of note, according to this comparison, different methods indeed can find a large number of the same cell markers; in contrast, each method can also detect some specific makers dependent on their model hypothesis. For Vec2image, it would identify markers considering both feature values and associations, and thus it might select the important features/genes on networks (e.g.

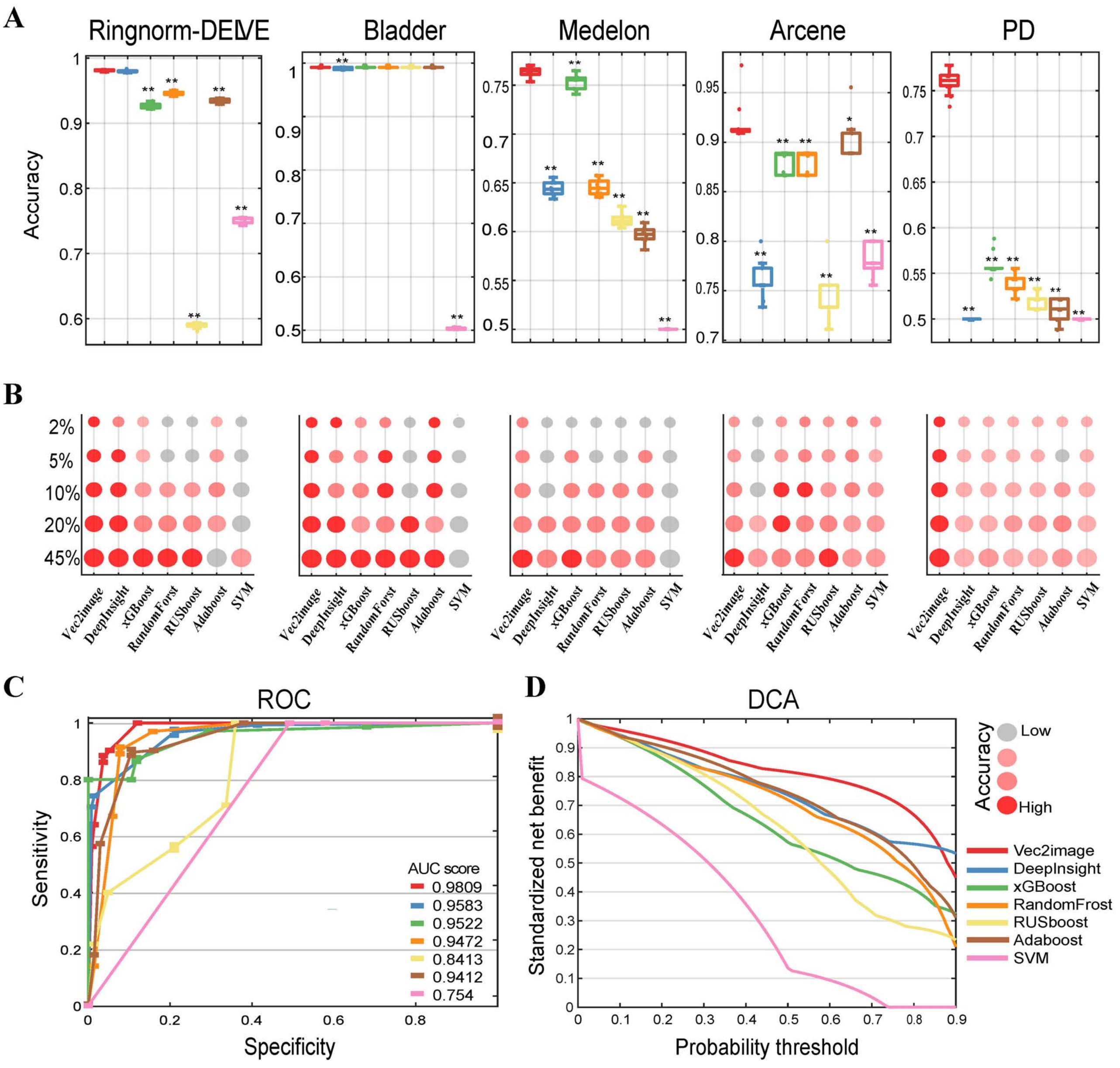

Figure 2. Performance and robustness evaluations of Vec2image on binary classification tasks. (**A**) Performance evaluation for Vec2image and six compared methods on five datasets, according to classification model accuracy. The star indicates the significance of performance promotion of Vec2image compared with other methods (*P < 0.05, **P < 0.01). (**B**) Performance evaluation for Vec2image and six compared methods on five groups of datasets with different imbalance ratio, where the dot color indicates the degree of accuracy, and the dot size represents the proportion of the sample size of two imbalanced class categories. (**C**) ROC curves of each method on 10 times running on random shuffling data, which indicate the prediction power of classification models. (**D**) DCA of each method on 10 times running on random shuffling data, which indicate the clinical usefulness of corresponding models.

network hubs), which should be deeply explored in future work in single-cell fields.

Finally, as an independent evaluation to the selected features/genes on original expressions, we ranked the impact of IFGs on the corresponding classification model by comparing the performance of xGBoost classifiers learned on different groups/alternatives of candidate feature genes. This comparison included the Vec2image-selected IFGs+, the randomly selected genes except IFGs (IFG−) and the same number of randomly selected genes; each of these three gene groups had proportion 5%, 10%, 15% and 20%, respectively. The results in Figure 3**D** show that IFGs are indeed essential for achieving high performance for cell classification e.g. the detachment of IFGs from model learning could cause a remarkable performance decrease. In particular, the distribution of these IFGs in the feature correlation structure (i.e. the pseudo image of Vec2image) displays an insightful visualization that key feature genes are grouped or clustered in the hidden feature space (Figure 3**E**); they are actually automatedly recognized and weighted by principal component coordinate conversion and the CNN model integrated in Vec2image. Besides, the confusion matrix demonstrated (Supplementary Figure S6) again

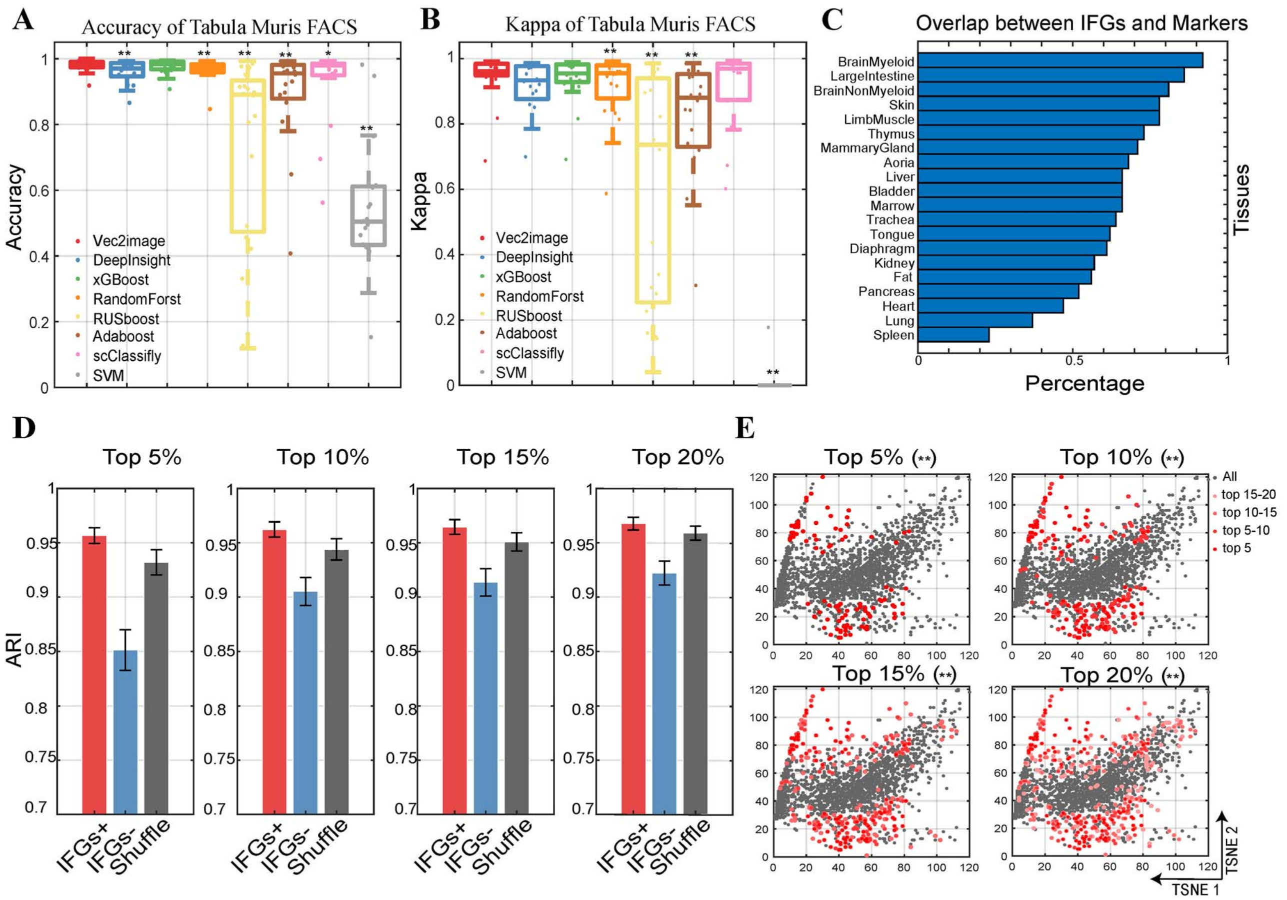

Figure 3. Performance and feature selection evaluations of Vec2image on multiclassification tasks. (**A**) Classification accuracy was summarized for Vec2image and 7 compared methods on 20 single-cell datasets from the Tabula Muris Atlas. The star indicates the significance of performance promotion of Vec2image compared with other methods (*$P < 0.05$, **$P < 0.01$). (**B**) Kappa coefficient evaluation was summarized for Vec2image and 7 compared methods on the same 20 datasets, using 10 times running on random shuffling data. The star indicates the significance of performance promotion of Vec2image compared with other methods (*$P < 0.05$, **$P < 0.01$). (**C**) Overlapping percentage of feature genes identified by Vec2image and cell makers recognized by Seurat in each dataset. (**D**) Performance of cell classification on the basis of different feature genes, including different proportions of IFGs+, IFGs− and shuffle genes. The bar indicates the mean ARI of 20 datasets and the error bar indicates the variance confidence interval (adjust by std/5), where ARI is applied to evaluate the match of prior-known types and predicted clusters of cells. (**E**) Distribution of different proportions of top-ranked IFGs in the latent space as feature association structure displayed in tSNE plot. Each dot represents a gene and the star indicates the P-value of hypergeometric test between IFGs and cell markers (**$P < 0.01$).

that Vec2image can acquire high accuracy in identifying rare cell types (e.g. small cell groups/categories) compared with other methods like DeepInsight and xGBoost. Of note, we expect here to independently test expressions of IFGs on distinguishing classes/cell types. And in above evaluation experimental results, xGBoost was found to achieve a best performance using original expression values, thus it was applied as an independent classifier to validate the IFGs identified by Vec2image. In fact, any other compared models can also be applied, and they should obtain the similar conclusion on the same IFGs.

Case study of T2D discrimination on bulk and single-cell islet transcriptome by Vec2image

Consensus classification of islet cell types

Many deep-sequencing techniques have been developed [54, 55] in the BT field. Ideally, the classifiers should be able to accurately and robustly identify sample/cell types in datasets generated from different platforms i.e. there should be consensus analysis results, although some systematic differences exist in sequencing data. In our collected pancreas datasets related to T2D (GSE86469 [45], E-MTAB-5061 [47], GSE84133 [35] and GSE83139 [46]), they have a similar biological context generated from three sequencing platforms (Methods and Supplementary Table S4), thus we first evaluated the classification consensus of Vec2image on cell type discrimination.

Here, two evaluation strategies were carried out for Vec2image in a dataset-based cross-validation manner. The first strategy is to train the classifier on any three datasets and perform validation on the remaining dataset. The second strategy is to train the classifier on any two datasets and perform validation on the two remaining datasets. The kappa coefficient was used to qualify the performance accuracy and robustness of Vec2image and comparable models (Supplementary Figure S7). Noted that, the four datasets indeed came from different sequencing platforms and laboratories, so that, they should have independent

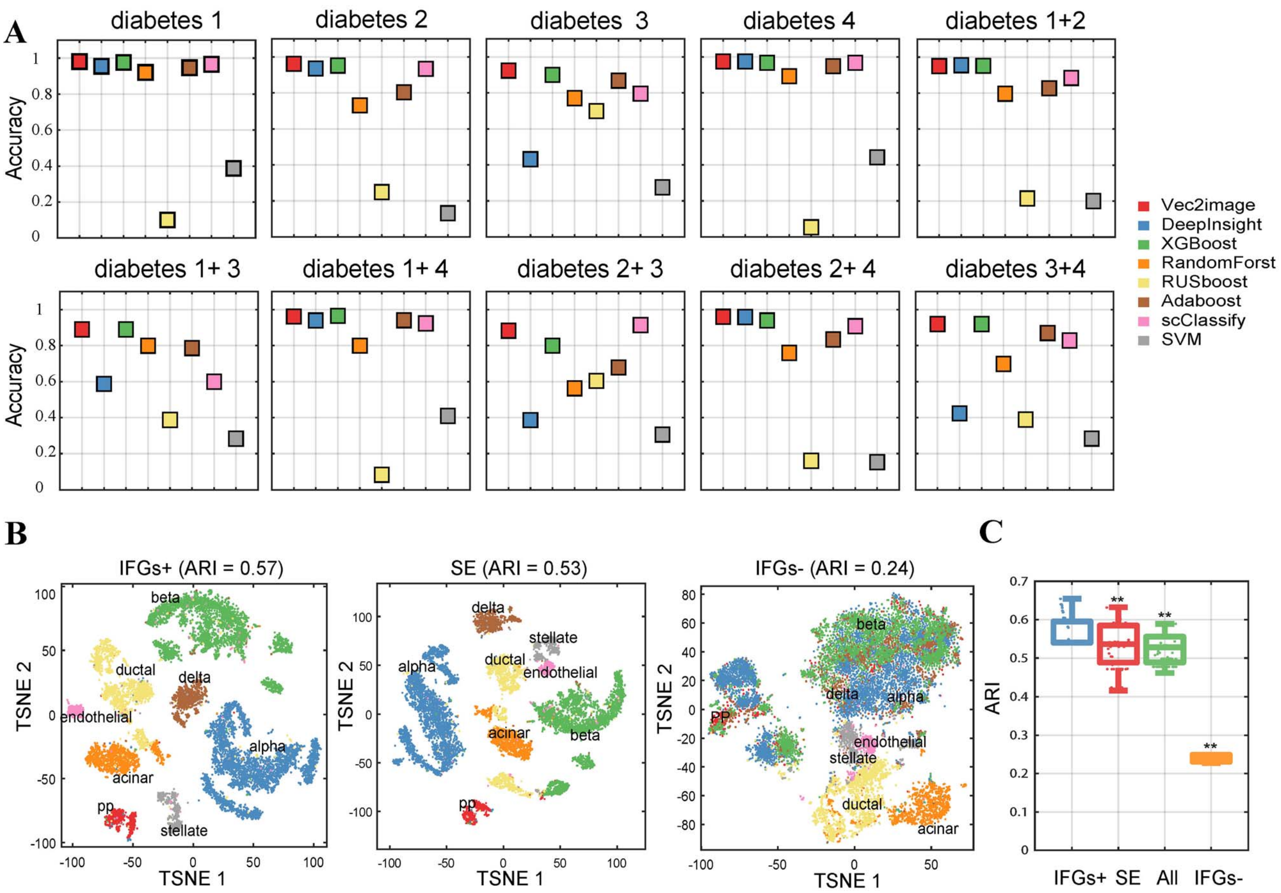

Figure 4. Accurate and consensus analysis of cell type classification by Vec2image on T2D-related scRNA-seq datasets generated from different platforms. (**A**) Classification accuracies of cross-dataset validation measured for Vec2image and seven compared methods on four pancreas scRNA-seq datasets (diabetes 1: GSE86469; diabetes 2: E-MTAB-5061; diabetes 3: GSE84133; diabetes 4: GSE83139). In each loop, the training datasets consist of three or two pancreas datasets of four pancreas scRNA-seq datasets, and the test dataset consists of the remaining datasets. For example, the first experiment is to train model on three datasets (i.e. diabetes 2, 3 and 4), and test model on remaining dataset (i.e. diabetes 1). (**B**) Unsupervised cell clustering within dimensional reduction by tSNE on the basis of informative genes selected by Vec2image (IFGs+), IFG-shaded genes (IFGs−) and genes selected by S-E model (SE), respectively. (**C**) Comparison of clustering result and prior-known clusters based on IFGs+, IFGs−, SE, and all genes. Here, the combination of all four pancreas datasets is used, and the cell clustering is based on tSNE, and the ARI are calculated by running k-means 50 times separately in each clustering. The significance test of performance comparison is carried out between IFGs+ and other groups of genes (SE, All, IFGs−), ** $P < 0.01$.

sample pattern/distribution in data. Such dataset-based cross-validation might also be an effective form of external validation to overcome potential overfitting issue in performance evaluation. These results clearly illustrated again that Vec2image can provide more robust performance of cell type identification than the other methods (Figure 4**A**), and especially Vec2image outperforms others in the classification task on such cross-platform data.

Consensus identification of islet cell-type-specific genes

As an important component of xAI, the identified IFGs of Vec2image should be constructive for downstream analyses and improve final outcomes, such as (unsupervised) cell clustering. Thus, the above four scRNA-seq datasets were combined together, 80% of the samples were used for selecting IFGs by Vec2image, and the remaining 20% of the samples were used for cell clustering, whose efficiency was evaluated by the ARI [56] using the previously known cell types as the gold-standard. As qualitatively shown in the Sankey plot (Supplementary Figure S8), Vec2image achieved high accuracy on cell type recovery. As quantitative measurement and for comparison, the ARI performance (Figure 4**C**) strongly indicated the ability of Vec2image to identify and select consensus IFGs. Here, the cell clustering was carried out by using different candidate genes: IFGs selected by Vec2image; feature genes identified by other state-of-the-art feature selection methods (S-E) [32] and randomly selected genes without IFGs (IFG−) as controls (Figure 4**B** and **C**).

Therefore, Vec2image achieves consensus performance on sample/cell classification and key feature/-gene selection, which also shows higher accuracy and sensitivity than the conventional CNN model (Supplementary Figure S9).

Accurate T2D state prediction of islet cell

Using Vec2image, we constructed a classifier of T2D to predict ND and T2D cells (states), by learning on integrative scRNA-seq data from three pancreas datasets (i.e. E-MTAB-5061, GSE84133 and GSE83139; including 9533 ND cells from 15 ND individuals and 1433 T2D cells from 6 diabetic individuals). As a supervised method, Vec2image acquire high accuracy in identifying cell states compared with SCCAF (Supplementary Figure S10). Next, we applied this classifier to classify the states of cells from an independent pancreas scRNA-seq dataset (GSE86469, including 380 ND cells and 258 T2D cells) and achieved significantly high accuracy (Accuracy score (ACC) = 0.9034) (Figure 5**A**). On the basis of this classifier, 373 IFGs related to T2D were selected (see the 'Materials and Methods' section). According to the accuracy of the baseline model rebuilt on shuffle datasets, these IFGs indeed played key roles in cell type classification. Of note, two strategies were used to produce those shuffle datasets: one is by removing the expression information of all IFGs, and the other one is by removing the expression information of the same number of randomly selected genes (Figure 5**A**).

Cell-type-specific IFGs relevant to T2D state

Among the above IFGs, there were four differentially expressed genes (DEGs, $P < 0.01$) between the ND and T2D states detected in acinar cells, 12 DEGs in alpha cells and 63 DEGs in beta cells (Supplementary Figures S11–S13), suggesting a stronger signal of dysfunction of T2D-related beta cells than other cell types. Moreover, we estimated the average total counts of these IFGs in all single-cell samples respectively belonging to five cell types and we found the significant decreases in alpha and beta cells from T2D ($P < 0.05$) (Figure 5**B**) and no significant proportion differences were detected in acinar, delta and ductal cells. These results were consistent with those from the analysis of all genes and all DEGs (Supplementary Figure S14). In addition, we applied AUCell [36] to score the activity of IFGs on ND and T2D cells (Figure 5**C**–**D**) and observed that the activity of IFGs in T2D cells was lower than that in ND cells (Figure 5**C**), and the activity of IFGs was significantly different between the ND and T2D states for alpha or beta cells (Figure 5**D**). Overall, the loss of expression and activity of IFGs could mainly result in the biological dysfunctions of alpha and beta cells in T2D.

As key feature genes selected by Vec2image, IFGs should discriminate cell types and also reveal insightful biological functions. By functional enrichment analysis of IFGs (see the 'Materials and Methods' section), IFGs are significantly correlated with metabolic processes (primary metabolic process: $P = 4.35\text{e-}5$, cellular metabolic process: $P = 8.24\text{e-}5$, organic substance metabolic process: $P = 0.008$) (Supplementary Table S5) relevant to abnormal metabolism in T2D. The four DEGs/IFGs in acinar cells were highly expressed in T2D, among which the high expression of *PTPRF* can contribute to the pathogenesis of insulin resistance [57–59] and *PDCD4* is known to be involved in pancreatic cancer [60] (Figure 5**E**, Supplementary Table S6). In contrast, the most IFGs/DEGs in alpha and beta cells had low expressions in the T2D state (Figure 5**E** and Supplementary Figure S13), and the top-ranked IFGs/DEGs, such as *REEP3* and *NACA*, were highly correlated with endoplasmic reticulum (ER) relevant to functional links between ER stress and T2D [61–63] (Supplementary Table S6, Supplementary Figure S15). Some other IFGs have not been reported to be associated with T2D currently, and they would be underestimated due to their cell type-specific dysfunction under T2D states, so they warrant further research.

T2D-related cell composition involved with IFGs

Based on aforementioned recognized cell type and marker gene information from single-cell sequencing data, the cell type proportions and cell type-specific differential expression can be efficiently estimated from bulk sequencing data by statistical deconvolution methods [35, 64], which would aid in the investigation of individual-specific disease/cellular heterogeneity.

The islet bulk RNA-seq dataset of 24 samples (including 15 NDS/individual) and 9 T2DS /individual were analyzed. We first used a set of discriminative marker genes (Supplementary Table S7) to estimate the proportion of each cell type (acinar, alpha, beta, delta and ductal) in the bulk samples by Bseq-SC (Figure 6**A**). Significant proportion increases ($P < 0.05$) in alpha and ductal cells were detected in T2DS compared with NDS, whereas the proportion of beta, acinar and delta cells decreased. Considering that the proportion variability of different cells may affect much of the reported differences in bulk gene expression, we fit two models of gene expression differences between ND and T2D by using *EdgeR* to detect a set of genes that are differentially regulated at the cell type level. The base model included only the group variable (sex, age) and then the estimated cell type proportions (alpha, beta) as additional covariates formed an extended model. Following the adjustment of cell type proportion, we identified a set of 252 genes with differential expression between ND and T2D samples, among which 69 genes showed significant differences between NDS and T2DS (67 genes had significant upregulation in T2DS, $P < 0.05$) (Supplementary Figure S16). Next, we used a regression-based method *csSAM* [65] on the above subset of adjusted genes and an additional set of five IFGs related to ER stress (*ALDH1A1, NACA, REEP3, OSTC* and *REEP5*) that were identified in the previous single-cell data analysis [34], aiming to identify cell type-specific differential expression and estimate interaction terms between a group of variables/genes and the cell proportions of alpha and beta cells. We observed a strong signal for gene upregulation in alpha cells and downregulation in beta cells of T2D (false discovery rate (FDR) < 0.05) (Figure 6**B**). All five of IFGs showed cell type-specific differences, upregulated in alpha cells and downregulated in beta cells. This result suggests that the beta cell heterogeneity identified in the single-cell dataset may be smaller in T2D than in the normal state; in contrast, the alpha cells showed the opposite effect. These findings indicate that there

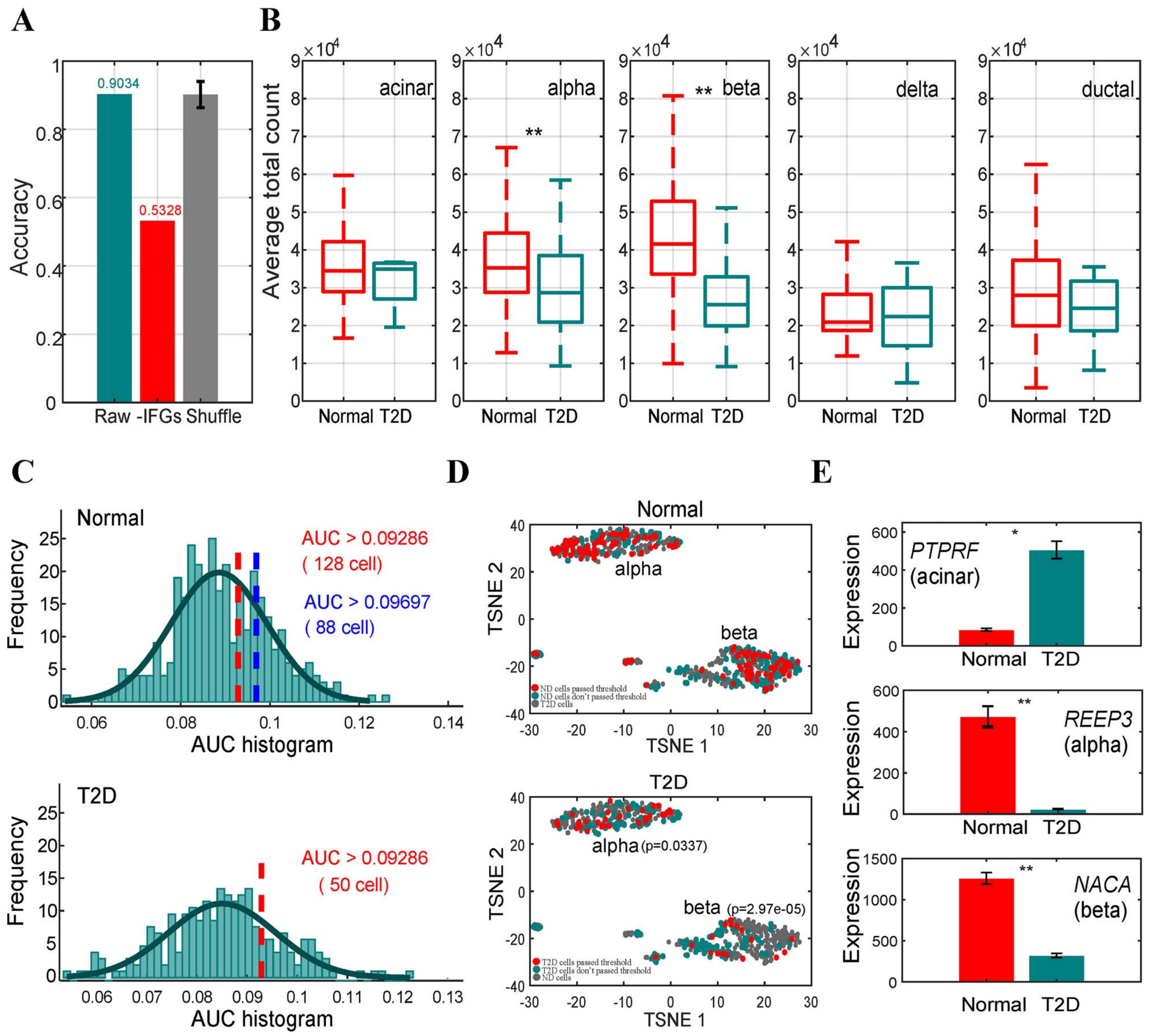

Figure 5. Efficient classifier for T2D single cells learnt by Vec2image with cell-type-specific diabetes-related genes identification. (**A**) Classification accuracy of Vec2image on different feature space e.g. the raw test data, the IFGs− test data (whose IFGs were set 1 for shades) and shuffle data (whose randomly selected genes were set 1 for shades). Of note, the shuffle groups were randomly selected 10 times, whose average and variance of accuracies were summarized and illustrated. (**B**) Summary of average total count of IFGs in ND and T2D cells for five cell types (** $P < 0.01$ for significance test). (**C**) Quantification of cell activity (AUC score) based on IFGs in ND and T2D cells. The vertical dotted line indicates the threshold selected by default. (**D**) Distribution of cell activity of each cell in latent space (e.g. tSNE plot) based on IFGs. The red dots indicate the case cells pass the threshold (AUC > 0.09286), and the blue dots represent the case cells do not pass the threshold, where the gray dots represent the control cells (e.g. T2D cells as cases in upper subfigure or ND cells as cases in bottom subfigure). Significant differences of AUC score are detected in alpha and beta cells between ND and T2D states by unpaired *t*-test. (**E**) Detailed expression changes between ND and T2D cells for the most differentially expressed genes of IFGs in acinar (upper), alpha (median) and beta (under) cells, respectively (*$P < 0.05$, **$P < 0.01$).

are actually necessary requirements for investigating the cellular heterogeneity underlying the individual heterogeneity of T2D for the precise diagnosis and prognosis of T2D.

T2D-related cell–cell communication involved with IFGs

In addition, the intercellular communication was investigated by CellChat [37], which showed that five cell types in ND and T2D were enriched in totally different signaling pathways (Figure 6**C**). We found that ductal cells dominate the outgoing and incoming signaling in both the ND and T2D states. Beta cells reduce communication patterns in the T2D state e.g. turning off signaling such as *WNT* (incoming) and *FGF*. In contrast, alpha cells did the opposite (Figure 6**C**) e.g. turning on/increasing signaling such as *SEMA3*, *SEMATOSTATIN*, *MK* and *ANGPTL*. In addition, we also studied the detailed changes in the signaling of IFGs (Supplementary Figure S17). Interestingly, we found that alpha and beta cells changed their major outgoing and incoming signaling between the ND state and the T2D state. In the T2D state, beta cells turn on *NEGR* signaling, and alpha cells turn off signaling, such as *CDH* and *CDH5*, and turn on the *L1CAM* signaling. The results suggested that the IFGs prominently redesigned their outgoing and incoming signaling in different cell types between the ND state and the T2D state.

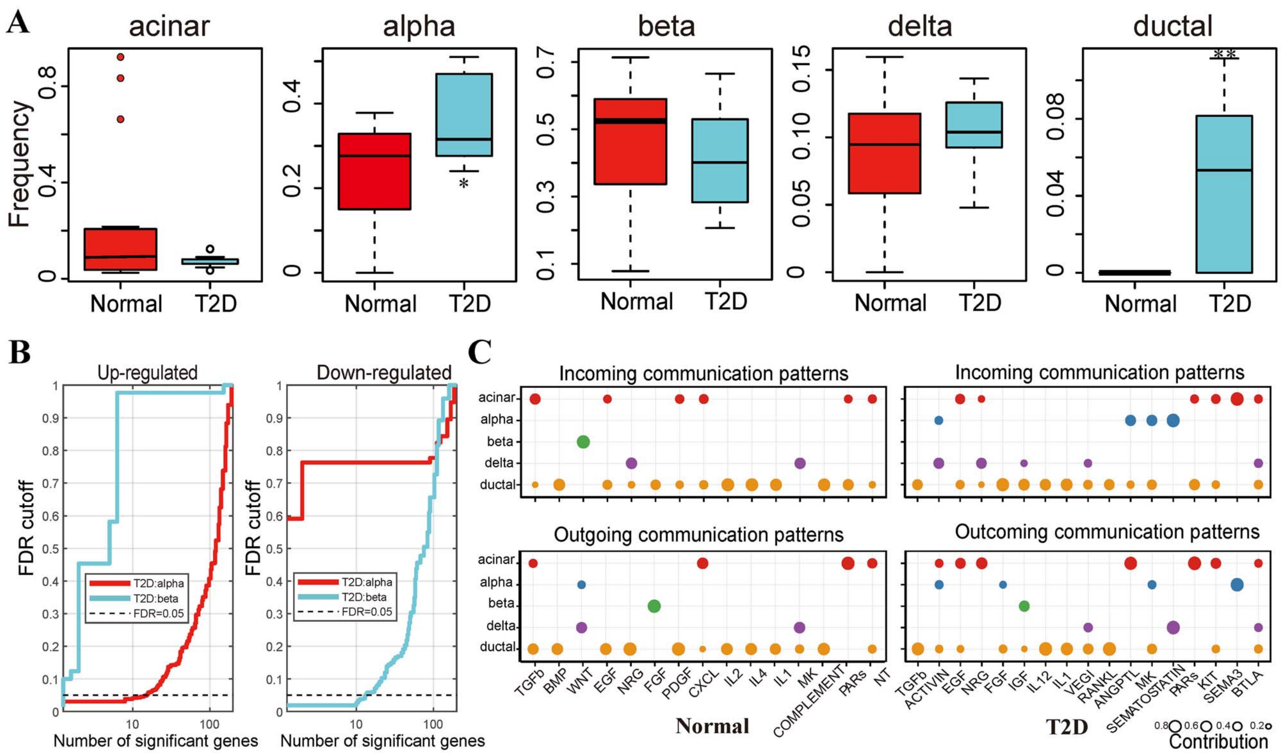

Figure 6. Functional role of T2D relevant IFGs identified by Vec2image in cell-type-specific expression/composition and cell–cell communications. (**A**) Proportion differences of pancreas cell types between ND and T2D (*$P < 0.05$, **$P < 0.01$). (**B**) Cell-type specificity analysis based on FDR plot from Bseq-SC. The *x*-axis shows the number of upregulated genes and downregulated genes at a given FDR cutoff (*y*-axis). (**C**) Comparison of incoming and outgoing signaling patterns of different cell types between ND and T2D sates. The dot size is proportional to the contribution score computed from pattern recognition analysis, where a higher contribution score implies the signaling pathway is more enriched in the corresponding type of cells.

In particular, the ligands in alpha cells can regulate the expression of IFGs/DEGs in beta cells, thus we applied NicheNet [38] to assess the changes in the ligand-target association between the ND and T2D states (see the 'Materials and Methods' section). We selected the top 20-ranked ligands and found that the five top-ranked ligands already dominate the regulation of IFGs in both the ND and T2D states. On the one hand, the top-ranked ligands in the two states are different e.g. one of the top-ranked ligands, *SLIT2*, which encodes a member of the slit family of secreted glycoproteins and is related to pancreatic cancer [66, 67], exist in only the T2D state. On the other hand, the top-ranked ligands regulated more IFGs in beta cells in the T2D state than in those in the ND state (Supplementary Figure S18). These results suggest that many IFGs indeed play important roles in T2D by being involved in the dysfunctional communication from alpha cells to beta cells.

Discussion

In the era of biological big data, a critical and basic problem is to capture the small variations in genomic data that differentiates phenotypes. To benefit from deep learning technologies for analyzing high-dimensional biological data, a necessary way of xAI model would bridge nonimage information and image-based learning. With that in mind, we proposed Vec2image to realize classification of biological samples on transformed pseudo image data in a qualitative and quantitative manner. Vec2image addresses several limitations of conventional ML architecture by two techniques: one is to transform the vector information to pseudo images by similarity embedding with feature abundance, and the other one is to unlock the black box by adapting KNN search in latent space. By evaluation on many benchmark datasets, Vec2image showed higher accuracy and sensitivity than competing approaches, and it was also able to rank and select informative features related to sample discrimination and feature association simultaneously. In a deep case study of bulk and single-cell integrated pancreas RNA-seq data for T2D, we learned a disease classifier at both the cell and tissue levels, selected 373 IFGs to distinguishing ND and T2D states and detected significant cell proportion differences and IFGs activity changes in alpha and beta cells. Dissimilar to traditional analysis, Vec2image automatically represented, quantified and detected such informative features (e.g. IFGs including DEGs), which might regulate the cell state, affect cell activity related to ER stress, further have an influence on the cell–cell communication and finally lead to relevant T2D dysfunctions such as cell-type-specific insulin resistance and glucagon imbalance. Thus, these new findings from Vec2image analysis should provide an understanding of pancreatic dysfunction and potential therapeutics for

T2D from a joint viewpoint from individuals and their organism cells, and the IFGs identified by Vec2image are explainable and constructive candidates for follow-up wet-experiments at both the molecular and cellular levels.

In this work, the structure of feature associations was actually holding the same as a shape description in pseudo image, and the pixel values corresponding to feature values of each sample could be different. Indeed, the nonlinear dimensionality reduction algorithm we used (e.g. tSNE) has the ability to map high dimensional data to a low dimensional data (i.e. 2D plane) and can preserve the local relationships among features (genes). Thus, the shape in pseudo image contains both feature abundance and correlation information, which is benefit for feature extraction of feature values and associations simultaneously by using CNN or similar algorithm. And another advantage is to further rank features by embedding KNN algorithm, which can also consider the influence of feature values and associations. Indeed, the usage of different shapes (e.g. different feature structures generated by different dimensionality reduction algorithms) can also be considered (e.g. using multichannel image for CNN where one channel includes one structure), which would be deeply investigated in future work by diverse external validations.

To evaluate the performance of Vec2image, we have compared seven typical supervised learning methods. To the best of our knowledge, Adaboost, random forest and SVM are general ML methods in classification with different theoretical and computational technologies, which can be used as general baseline methods. RUSboost is able to combine data sampling and boosting, which can be used to further assess the ability of Vec2image in dealing with imbalanced datasets. xGBoost has shown better performance and computation cost in previous studies. And scClassify has been compared with many popular single-cell focused supervised methods (especially including scVI using deep neural networks [68], and ACTINN using a neural network with three hidden layers [69]) and shows high accuracy in single-cell classification. Thus, this work adopted the wide comparison with these well-known methods, and the combination or integration with other models in Vec2image would provide improved accurate and robust xAI model in diverse application scenarios. For example, the identification of unassigned cells and follow-up cell clustering is also a problem in the application of single-cell data analysis. Motivated by scClassify on the basis of tree structure, a potential expansion of Vec2image would be able to identify the weak-evidence supported samples/cells in CNN classification model (e.g. with small prediction weights for all classes) as candidate unassigned cells, and then using any CNN-based or alternative single-cell clustering methods to cluster these cell candidates for new cell types, which can be further understand and analyzed by biostatisticians or cytologists.

In technical terms, the unsupervised learning methods would be easily influenced by various confounding factors in the data e.g. it is difficult to differentiate disease and normal cells in single-cell datasets, which tend to give priority to clustering cell types (Supplementary Figure S19). Meanwhile, supervised learning methods help directly detect the discrimination of different target phenotypes; however, they usually ignore the relationships between features, so as to enlarge the side effects of black-box computing in biological and biomedical applications. To benefit the computational analysis of biological big data with advanced AI technologies, Vec2image was developed on a highly applicable and interpretable CNN-based framework. It realizes the construction and deployment of accurate and robust classification/prediction for nonimage biological omics datasets. In Vec2image, nonimage sample data can be transformed into a pseudo image sample data with feature structure and state simultaneous embedding, which should be conveniently processed by CNN or similar approaches. Vec2image also tries to interpret the results of the original CNN black box algorithms by supplying additional feature extraction and selection approaches e.g. ranking features based in their impact on CNN performance during network disturbance and screening features through its manifold structure distribution in latent space embedded in image structure. To the best of our knowledge, this is the first xAI model and algorithm for the deep mining of a variety of biological sequencing datasets that has great potential to be applied in biological and biomedical fields.

Of note, Vec2image uses CNN for supervised learning, considering that CNN or related image-intuitive methods can help automatic feature extraction and reduce the need for neurons and sharing weights in large-scale computation even with GPU (graphics processing unit) utilization. In particular, the ResNet framework enables a much deeper level of model training and promises high accuracy in classification; thus, it is utilized for pseudo image classification in Vec2image. The choice of evaluation metric is also important for comparing the performance of Vec2image and other state-of-the-art methods. This work has carried out a wide benchmark assessment on multiple computational tasks, including binary classification, multiclass classification, sample imbalanced classification, cross-platform classification and cross-sample classification. In particular, class imbalance occurs in many biological sequencing datasets, due to the easy collection of normal/control samples rather than disease/case samples. Deep neural network architectures encompass the advantages of upsampling, and SMOTE sampling can be easily incorporated in the CNN framework to correct the class imbalance, which is critical for Vec2image to achieve optimal performance in imbalanced datasets.

In addition, there are already some interpretable ML models have already been proposed recently. For example, CellBox is an interpretable ML tool for perturbation biology [70], and another work used an interpretable ML framework to identify T2D-related gut microbiome features [71]. These works indicate that deep learning is becoming an effective data-driven framework capable of generating predictions for complex biological systems. However, these models and applications are still limited to specific biological contexts, far short of the emerging demands of xAI [72, 73]. Our developed Vec2image aims to provide a general ML architecture for xAI applications in the biological and biomedicine fields. Vec2image enables the CNN framework to classify nonimage samples by converting the expressions data of each sample to pixel data, with low sensitivity on the format and distribution of input data; and it also enlarges features interpretability by quantifying its impact and tracing its relationship within feature communities in latent feature space. Considering that Vec2image can deal with pseudo image data and also native image data, it has immense potential to be applied to clinical diagnosis where conventional biomedical images and new omics data are provided adequately.

Vec2image is implemented using the CNN-based framework so that it can make full use of various advanced ML techniques, such as dropout, mini-batching and Bayesian optimization techniques, to further improve the learning efficiency and accuracy and reveal previously nonobserved informative features and their associations. The current framework of Vec2image employs grayscale or the 2D layer of pseudo images for classification, thus further extension can consider the more complicated structure of pseudo image data.

Vec2image would be scalable when its computational components can be further enhanced. For the imbalance sampling, the time complexity is already linear with data size (e.g. the calculation of SMOTE in Methods). For the nonlinear dimensionality reduction (e.g. tSNE), the time complexity would be $O(N^2)$; however, many approach can reduce it to $O(N(\log(N))$ [74], where N is the size of samples. For the CNN, its general computational complexity of each layer could be $O(M^2 * K^2 * C_{in} * C_{out})$, where M is the (edge) size of feature map, K is the (edge) size of Kernel, C_{in} and C_{out} are the number of input and output channels of Kernel. As well-known, the CNN with ResNet can be easily speed-up by GPU in actual applications, which have been implemented well in MATLAB and Python environments. And the implementation of Vec2image on web server or cloud computing will be an important future work for users in different fields.

With the rapid development of biological sequencing technology, the whole genomic sequence information of numerous organisms has become publicly available, and different kinds of omics data (e.g. gene expression, methylation, ATAC-seq (Assay for Transposase Accessible Chromatin using sequencing)) need to be jointly analyzed for downstream analysis and prediction. Vec2image has the potential to integrate and analyze bulk and single-cell sequencing data. However, it is still necessary to consider further algorithm extension to incorporate multiple data layers into pseudo images to realize the integration of multiomics datasets. One potential direction of future work could integrate multiomics datasets by reducing multiomics feature to medium dimensions with enough variations and extracting the regulatory network map of different molecules (e.g. RNA-MicroRNA) [75, 76], which combines with ML-related methods (e.g. graph neural network) can realize phenotype predictions, key feature selection and key sub-regulatory network reconstruction simultaneously.

Collectively, Vec2image can provide a comprehensive and widely applicable xAI model and approaches for analyzing high-dimensional biological data.

Key Points

- The Vec2image model is an explainable CNN framework for characterizing the feature engineering, feature selection, classifier learning and topic recommendation, by considering the feature values and feature correlation structure in an integrative image manner.
- Vec2image has shown better performance than other compared methods on different binary classification and multiclass classification tasks, especially dealing well with imbalance datasets and cross-platform datasets.
- Vec2image is also efficient on feature selection as cell marker identification from tissue-specific single-cell datasets.
- Vec2image provides accurate and robust T2D classifier to recognize cell state in disease or normal individuals, and further recommends feature genes relevant to T2D cellular pathogenesis including cell activity change, cell composition imbalance and cell–cell communication dysfunction.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

National Natural Science Foundation of China (Nos 11871456, 12026608, 62172164 and 61803360), Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB38050200), the Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01) and Guangdong Basic and Applied Basic Research Foundation, China (Grant Nos 2019B151502062, 2021A1515012317).

Data availability

An open-source implementation of Vec2image in MATLAB and alternative python script for ResNet are available at GitHub: https://github.com/ztpub/Vec2image. This package has been tested on Linux and Windows platforms based on MATLAB2020 version and python with example datasets, which is default to use multi-GPU in Linux system.

References

1. Trapnell C. Defining cell types and states with single-cell genomics. *Genome Res* 2015;**25**:1491–8.
2. Kiselev VY, Kirschner K, Schaub MT, *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;**14**:483–6.
3. Wang B, Zhu J, Pierson E, *et al.* Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* 2017;**14**:414–6.
4. Tang H, Tang Y, Zeng T, *et al.* Gene expression analysis reveals the tipping points during infant brain development for human and chimpanzee. *BMC Genomics* 2020;**21**:74.
5. Kolodziejczyk AA, Kim JK, Svensson V, *et al.* The technology and biology of single-cell RNA sequencing. *Mol Cell* 2015;**58**:610–20.
6. Zeng T, Yu X, Chen Z. Applying artificial intelligence in the microbiome for gastrointestinal diseases: a review. *J Gastroenterol Hepatol* 2021;**36**:832–40.
7. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019;**20**:e262–73.
8. Abdelaal T, Michielsen L, Cats D, *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**:194.
9. Chen T, Guestrin C. XGBoost: a scalable tree boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA: Association for Computing Machinery), 2016;785–794.
10. Wang P, Shen CH, Barnes N, *et al.* Fast and robust object detection using asymmetric totally corrective boosting. *IEEE Trans Neural Netw Learn Syst* 2012;**23**:33–46.
11. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell* 1998;**20**:832–44.
12. Sharma A, Vans E, Shigemizu D, *et al.* DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture. *Sci Rep* 2019;**9**:–11399.
13. Lin YX, Cao Y, Kim HJ, *et al.* scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol Syst Biol* 2020;**16**:e9389.
14. Wang L, Catalan F, Shamardani K, *et al.* Ensemble learning for classifying single-cell data and projection across reference atlases. *Bioinformatics* 2020;**36**:3585–7.
15. Zhao X, Wu S, Fang N, *et al.* Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Brief Bioinform* 2020;**21**:1581–95.
16. Chen P, Liu R, Aihara K, *et al.* Autoreservoir computing for multi-step ahead prediction based on the spatiotemporal information transformation. *Nat Commun* 2020;**11**:4568.
17. Bakken T, Cowell L, Aevermann BD, *et al.* Cell type discovery and representation in the era of high-content single cell phenotyping. *BMC Bioinformatics* 2017;**18**:559.
18. Harada S, Akita H, Tsubaki M, *et al.* Dual graph convolutional neural network for predicting chemical networks. *BMC Bioinformatics* 2020;**21**:94.
19. Wong LM, Ai QYH, Mo FKF, *et al.* Convolutional neural network in nasopharyngeal carcinoma: how good is automatic delineation for primary tumor on a non-contrast-enhanced fat-suppressed T2-weighted MRI? *Jpn J Radiol* 2021;**39**:571–9.
20. Yamlome P, Akwaboah AD, Marz A, *et al.* Convolutional neural network based breast cancer histopathology image classification. *Annu Int Conf IEEE Eng Med Biol Soc* 2020;**2020**:1144–7.
21. Yasaka K, Akai H, Kunimatsu A, *et al.* Deep learning with convolutional neural network in radiology. *Jpn J Radiol* 2018;**36**:257–72.
22. Lyu B, Haque A. Deep learning based tumor type classification using gene expression data. *In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Washington, DC, USA: Association for Computing Machinery), 2018;89–96.
23. Kalakoti Y, Yadav S, Sundar D. SurvCNN: a discrete time-to-event cancer survival estimation framework using image representations of omics data. *Cancer* 2021;**13**:3106.
24. Zhu Y, Brettin T, Xia F, *et al.* Converting tabular data into images for deep learning with convolutional neural networks. *Sci Rep* 2021;**11**:11325.
25. Naz M, Shah JH, Khan MA, *et al.* From ECG signals to images: a transformation based approach for deep learning. *PeerJ Comput Sci* 2021;**7**:e386.
26. Lauritsen SM, Kristensen M, Olsen MV, *et al.* Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat Commun* 2020;**11**:3852.
27. Chawla NV, Bowyer KW, Hall LO, *et al.* SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321–57.
28. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
29. Scholkopf B, Smola A, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 1998;**10**:1299–319.
30. McInnes L, Healy J. UMAP: uniform manifold approximation and projection for dimension reduction. *ArXiv e-prints* 2018; 1802.03426.
31. Barber CB, Dobkin DP, Huhdanpaa HT. The Quickhull algorithm for convex hulls. *ACM Transac Math Software* 1996;**22**:469–83.
32. Liu B, Li C, Li Z, *et al.* An entropy-based metric for assessing the purity of single cell populations. *Nat Commun* 2020;**11**:3155.
33. Butler A, Hoffman P, Smibert P, *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.
34. Miao Z, Moreno P, Huang N, *et al.* Putative cell type discovery from single-cell gene expression data. *Nat Methods* 2020;**17**:621–8.
35. Baron M, Veres A, Wolock SL, *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst* 2016;**3**:346–60.e4.
36. Aibar S, Gonzalez-Blas CB, Moerman T, *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6.
37. Jin S, Guerrero-Juarez CF, Zhang L, *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* 2021;**12**:1088.
38. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* 2020;**17**:159–62.
39. Raudvere U, Kolberg L, Kuzmin I, *et al.* G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 2019;**47**:W191–8.
40. Rodchenkov I, Babur O, Luna A, *et al.* Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic Acids Res* 2020;**48**:D489–97.
41. Breiman L. Bias, variance, and arcing classifiers. *Addit Polym* 1996;**2002**:10.

42. Tabula Muris C, Overall C, Logistical C, *et al*. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 2018;**562**:367–72.
43. Guyon I, Gunn S, Hur AB, *et al*. Result analysis of the nips 2003 feature selection challenge. *Advances in Neural Information Processing Systems 17* [Vancouver, British Columbia, Canada], 2004;**4**:545–552.
44. Sakar CO, Serbes G, Gunduz A, *et al*. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Appl Soft Comput* 2018;**74**:255–63.
45. Lawlor N, George J, Bolisetty M, *et al*. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* 2017;**27**: 208–22.
46. Wang YJ, Schug J, Won KJ, *et al*. Single-cell Transcriptomics of the human endocrine pancreas. *Diabetes* 2016;**65**:3028–38.
47. Segerstolpe A, Palasantza A, Eliasson P, *et al*. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab* 2016;**24**:593–607.
48. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
49. Seiffert C, Khoshgoftaar T, Van Hulse J, Napolitano A RUSBoost: improving classification performance when training data is skewed. *In: 2008 19th International Conference on Pattern Recognition* (ICPR 2008)[IEEE, December 8-11, 2008, Tampa, Florida, USA]. 2008:1–4.
50. Freund Y, Schapire RE. A decision-theoretic generalization of online learning and an application to boosting. *J Comput Syst Sci* 1997;**55**:119–39.
51. Cortes C, Vapnik VN. Support-vector networks. *Mach Learn* 1995;**20**:273–97.
52. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* 2018;**18**:35–45.
53. Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–1902.e21.
54. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* 2018;**13**: 599–604.
55. Jiang D, Zhang X, Pang Y, *et al*. Terminal transfer amplification and sequencing for high-efficiency and low-bias copy number profiling of fragmented DNA samples. *Protein Cell* 2019;**10**: 229–33.
56. Hubert LAP. Comparing partitions. *J Classif* 1985;**14**: 193–218.
57. Miscio G, Tassi V, Coco A, *et al*. The allelic variant of LAR gene promoter −127 bp T–>a is associated with reduced risk of obesity and other features related to insulin resistance. *J Mol Med* 2004;**82**:459–66.
58. Mander A, Hodgkinson CP, Sale GJ. Knock-down of LAR protein tyrosine phosphatase induces insulin resistance. *FEBS Lett* 2005;**579**:3024–8.
59. Seki N, Yagi S, Suzuki Y, *et al*. Protein tyrosine phosphatase regulation in fibroblasts from patients with an insulin receptor gene mutation. *Horm Metab Res* 2008;**40**:833–7.
60. Zhang LL, Yao J, Li WY, *et al*. Micro-RNA-21 regulates cancer-associated fibroblast-mediated drug resistance in pancreatic cancer. *Oncol Res* 2018;**26**:827–35.
61. Ozcan U, Cao Q, Yilmaz E, *et al*. Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. *Science* 2004;**306**:457–61.
62. Kumar D, Golchoubian B, Belevich I, *et al*. REEP3 and REEP4 determine the tubular morphology of the endoplasmic reticulum during mitosis. *Mol Biol Cell* 2019;**30**:1377–89.
63. Hotokezaka Y, Van Leyen K, Lo EH, *et al*. alphaNAC depletion as an initiator of ER stress-induced apoptosis in hypoxia. *Cell Death Differ* 2009;**16**:1505–14.
64. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr Opin Immunol* 2013;**25**:571–8.
65. Shen-Orr SS, Tibshirani R, Khatri P, *et al*. Cell type-specific gene expression differences in complex tissues. *Nat Methods* 2010;**7**: 287–9.
66. Secq V, Leca J, Bressy C, *et al*. Stromal SLIT2 impacts on pancreatic cancer-associated neural remodeling. *Cell Death Dis* 2015;**6**:e1592.
67. Gohrig A, Detjen KM, Hilfenhaus G, *et al*. Axon guidance factor SLIT2 inhibits neural invasion and metastasis in pancreatic cancer. *Cancer Res* 2014;**74**:1529–40.
68. Lopez R, Regier J, Cole MB, *et al*. Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8.
69. Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics* 2020;**36**:533–8.
70. Yuan B, Shen C, Luna A, *et al*. CellBox: interpretable machine learning for perturbation biology with application to the Design of Cancer Combination Therapy. *Cell Syst* 2021;**12**:128–140 e124.
71. Gou W, Ling CW, He Y, *et al*. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. *Diabetes Care* 2021;**44**:358–66.
72. Thelisson, E. Towards trust, transparency and liability in AI/AS systems. In: *Twenty-Sixth International Joint Conference on Artificial Intelligence*, (Melbourne, Australia: AAAI Press) 2017;5215–5216.
73. Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 2019;**40**(2):44–58.
74. Pezzotti N, Lelieveldt BPF, Van Der Maaten L, *et al*. Approximated and user steerable tSNE for progressive visual analytics. *IEEE Trans Vis Comput Graph* 2017;**23**:1739–52.
75. Tang H, Zeng T, Chen L. High-order correlation integration for single-cell or bulk RNA-seq data analysis. *Front Genet* 2019;**10**:371.
76. Shi Q, Zhang C, Peng M, *et al*. Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics* 2017;**33**:2706–14.