OmiTrans: Generative Adversarial Networks Based Omics-to-omics Translation Framework

Xiaoyu Zhang1,*, Yike Guo2,3
1 Academy of Military Sciences, Beijing, 100091, China
2 The Hong Kong University of Science and Technology, Hong Kong, China
3 Data Science Institute, Imperial College London, London, SW7 2AZ, UK

*Corresponding author: x.zhang18@imperial.ac.uk

Abstract—With the significant development of high-throughput experimental technologies, different types of omics (e.g., genomics, epigenomics, transcriptomics, proteomics, and metabolomics) data are produced from clinical samples at an unprecedented speed. The correlations between different types of omics data attract enormous research interest. In contrast, the study on genome-wide omics-to-omics translation (i.e., generation and prediction of one type of omics data from another type of omics data) is almost blank. Generative adversarial networks and variants are among the most state-of-the-art deep learning technologies, which have shown great success in image-to-image translation, text-to-image translation, etc. Here we propose OmiTrans, a deep learning framework that adopted the idea of generative adversarial networks to achieve omics-to-omics translation with promising results. OmiTrans can faithfully reconstruct genome-wide gene expression profiles from DNA methylation data with high accuracy and excellent model generalisation, as demonstrated in the experiments.

Index Terms—generative adversarial networks, deep learning, omics data, DNA methylation, gene expression

I. INTRODUCTION

The research on multi-omics data is an emerging topic that has attracted a great deal of interest recently [1]–[4], and the correlation between different omics types, especially the regulation of gene expression, is a classic research topic with long-term development [5], [6]. Nevertheless, the genome-wide generation and prediction of one type of omics data (e.g., gene expression profiles) from another type of omics data (e.g., DNA methylation profiles), a.k.a., omics-to-omics translation, has not been well studied. Zhong et al. [7] was one of the few works that discussed the possibility of predicting gene expression using DNA methylation data, which adopted locally-connected LASSO regression to the task. However, they came up with the conclusion that DNA methylation data had limited prediction power for gene expression, which is mainly because of the rather low capacity of their network, according to our experiments. TDimpute [8] is a more recent work in this field that applied a relatively straightforward neural network to impute missing RNA-Seq gene expression data from DNA methylation data. They claimed that TDimpute significantly outperformed other state-of-the-art methods, including singular value decomposition imputation (SVD), trans-omics block missing data imputation (TOBMI) [9], and LASSO [10]. TDimpute [8] is currently the most state-of-the-art method that can impute missing gene expression data from DNA methylation data, which is our primary comparing method.

Generative adversarial networks (GANs) [11] are one of the most emerging deep learning methodologies which have seen significant breakthroughs these years [12]–[14]. GANs and the variants [15]–[17], especially the idea of conditional GAN, have been adapted to solve tasks like image-to-image translation [17]–[19] and text-to-image translation [20]–[23]. Recently, there are also some attempts to adopt GANs to omics data, especially the omics data imputation task [24], [25]. However, there is currently no generative adversarial networks framework designed for the omics-to-omics translation task.

In this work, we proposed OmiTrans, a deep learning framework that adopted the idea of generative adversarial networks (GANs) to achieve omics-to-omics translation. With OmiTrans, we were able to faithfully reconstruct gene expression profiles from the corresponding DNA methylation data with high accuracy, which outperformed other methods, including the state-of-the-art TDimpute [8]. We also applied the pre-trained OmiTrans model to individual and previously unseen datasets, indicating the great applicability, practicability, and model generalisation of OmiTrans.

II. MATERIALS AND METHODS

A. Datasets

Multi-omics datasets from the Genomic Data Commons (GDC) data portal [26] were selected for the following experiments of OmiTrans. GDC is one of the most prestigious unified multi-omics data repositories, which supports several programmes at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) [27] and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [28], [29]. Although OmiTrans can facilitate data conversion between any two types of omics data, we chose DNA methylation and RNA-Seq gene expression profiling to demonstrate the omics-to-omics translation ability of OmiTrans due to the sufficient data amount for the training of GANs-based models. Therefore, we selected 9,081 samples from 33 different projects in GDC with both DNA methylation and gene expression profiles for further experiments. The multi-omics data of GDC were downloaded from the UCSC

Index Terms—generative adversarial networks, deep learning, omics data, DNA methylation, gene expression

OmiTrans: Generative Adversarial Networks Based Omics-to-omics Translation Framework

Xiaoyu Zhang1,*, Yike Guo2,3
1 Academy of Military Sciences, Beijing, 100091, China
2 The Hong Kong University of Science and Technology, Hong Kong, China
3 Data Science Institute, Imperial College London, London, SW7 2AZ, UK

*Corresponding author: x.zhang18@imperial.ac.uk

Abstract—With the significant development of high-throughput experimental technologies, different types of omics (e.g., genomics, epigenomics, transcriptomics, proteomics, and metabolomics) data are produced from clinical samples at an unprecedented speed. The correlations between different types of omics data attract enormous research interest. In contrast, the study on genome-wide omics-to-omics translation (i.e., generation and prediction of one type of omics data from another type of omics data) is almost blank. Generative adversarial networks and variants are among the most state-of-the-art deep learning technologies, which have shown great success in image-to-image translation, text-to-image translation, etc. Here we propose OmiTrans, a deep learning framework that adopted the idea of generative adversarial networks to achieve omics-to-omics translation with promising results. OmiTrans can faithfully reconstruct genome-wide gene expression profiles from DNA methylation data with high accuracy and excellent model generalisation, as demonstrated in the experiments.

Index Terms—generative adversarial networks, deep learning, omics data, DNA methylation, gene expression

I. INTRODUCTION

The research on multi-omics data is an emerging topic that has attracted a great deal of interest recently [1]–[4], and the correlation between different omics types, especially the regulation of gene expression, is a classic research topic with long-term development [5], [6]. Nevertheless, the genome-wide generation and prediction of one type of omics data (e.g., gene expression profiles) from another type of omics data (e.g., DNA methylation profiles), a.k.a., omics-to-omics translation, has not been well studied. Zhong et al. [7] was one of the few works that discussed the possibility of predicting gene expression using DNA methylation data, which adopted locally-connected LASSO regression to the task. However, they came up with the conclusion that DNA methylation data had limited prediction power for gene expression, which is mainly because of the rather low capacity of their network, according to our experiments. TDimpute [8] is a more recent work in this field that applied a relatively straightforward neural network to impute missing RNA-Seq gene expression data from DNA methylation data. They claimed that TDimpute significantly outperformed other state-of-the-art methods, including singular value decomposition imputation (SVD), trans-omics block missing data imputation (TOBMI) [9], and LASSO [10]. TDimpute [8] is currently the most state-of-the-art method that can impute missing gene expression data from DNA methylation data, which is our primary comparing method.

Generative adversarial networks (GANs) [11] are one of the most emerging deep learning methodologies which have seen significant breakthroughs these years [12]–[14]. GANs and the variants [15]–[17], especially the idea of conditional GAN, have been adapted to solve tasks like image-to-image translation [17]–[19] and text-to-image translation [20]–[23]. Recently, there are also some attempts to adopt GANs to omics data, especially the omics data imputation task [24], [25]. However, there is currently no generative adversarial networks framework designed for the omics-to-omics translation task.

In this work, we proposed OmiTrans, a deep learning framework that adopted the idea of generative adversarial networks (GANs) to achieve omics-to-omics translation. With OmiTrans, we were able to faithfully reconstruct gene expression profiles from the corresponding DNA methylation data with high accuracy, which outperformed other methods, including the state-of-the-art TDimpute [8]. We also applied the pre-trained OmiTrans model to individual and previously unseen datasets, indicating the great applicability, practicability, and model generalisation of OmiTrans.

II. MATERIALS AND METHODS

A. Datasets

Multi-omics datasets from the Genomic Data Commons (GDC) data portal [26] were selected for the following experiments of OmiTrans. GDC is one of the most prestigious unified multi-omics data repositories, which supports several programmes at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) [27] and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [28], [29]. Although OmiTrans can facilitate data conversion between any two types of omics data, we chose DNA methylation and RNA-Seq gene expression profiling to demonstrate the omics-to-omics translation ability of OmiTrans due to the sufficient data amount for the training of GANs-based models. Therefore, we selected 9,081 samples from 33 different projects in GDC with both DNA methylation and gene expression profiles for further experiments. The multi-omics data of GDC were downloaded from the UCSC
The detailed information of the multi-omics datasets from GDC is shown in Supplementary Table 2.

For RNA-Seq gene expression data, all of the 60,483 identifiers obtained from the Illumina RNA-Seq array were kept for the experiments without any feature filtering. The gene expression level of each identifier was quantified by the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) measurement. The original profiles of DNA methylation data comprise 485,577 probes from the Infinium HumanMethylation450 BeadChip (450K) array. The feature filtering process was applied to the original DNA methylation profiles to balance the dimensionality between the two types of omics data and reduce the computational requirements. Feature filtering is a two-step process. First, specific probes in the 450K array were removed according to the similar criteria mentioned in [3]: probes excluded in the current Infinium MethylationEPIC BeadChip (EPIC) array (n = 32,260), probes containing the single-nucleotide polymorphism (SNP) dbSNP132Common within five base pairs of the targeted CpG site (n = 7,998), probes not uniquely mapping to the human reference genome (hg19) with one mismatch allowed (n = 3,965), the non-CpG loci probes (n = 3,091), the SNP assay probes (n = 65), and probes with missing values (N/A) in more than 10% of samples (n = 2). Unlike the filtering criteria in [3], we kept the CpG sites in sex chromosomes because they were indispensable for the reconstruction of corresponding gene expression levels in the sex chromosomes. 46,330 probes in total were filtered in the first step. Then in the next step, the variance of each remaining CpG site was calculated and sorted in descending order. Probes with variance lower than 0.05 were removed for further experiments. Eventually, 39,464 CpG sites in the 450K array remained after the two-step feature filtering process. The average feature value replaced any remaining N/A values in DNA methylation and gene expression datasets.

Location alignment is the final preprocessing step, where the feature orders of both the DNA methylation and gene expression profiles were rearranged according to their locations in the chromosomes, and the chromosomes were sorted by the order from 1 to 22 and X Y.

Furthermore, we selected the brain tumour DNA methylation dataset (BTM) GSE109381 [30] as an individual dataset to further evaluate the model generalisation of OmiTrans. The GSE109381 BTM dataset is one of the most extensive DNA methylation datasets of central nervous system tumours and the chromosomes were sorted by the order from 1 to 22 and X Y.

B. Overview of OmiTrans

The OmiTrans deep generative framework was proposed to implement the data translation between any two omics types (e.g., gene expression, DNA methylation, and miRNA expression) based on the idea of generative adversarial networks (GANs) and its variants [11], [15]–[17]. In a nutshell, OmiTrans learns a mapping from the profile of one omics type x and a random noise vector z, to the profile of another omics type y, \( G : \{x, z\} \rightarrow y \). The discriminator \( D \) is trained to classify between the synthetic (fake) \( \{x, y\} \) omics pairs and the original (real) \( \{x, y\} \) omics pairs. The overall diagram of the OmiTrans framework is illustrated in Figure 1.

In the original version of GANs [11], the generator \( G \) produces the fake data \( y' \) only from the noise \( z \) and the discriminator \( D \) classifies between the real data and fake data without the observation of \( x \). The value function \( V(G, D) \) of a vanilla GAN can be expressed as

\[
V_{GAN}(G, D) = \mathbb{E}_{y}[\log D(y)] + \mathbb{E}_{z}[\log(1 - D(G(z))].
\] (1)

Moving one step forward, we can provide \( x \) to generator \( G \) to facilitate the generation of \( y' \) based on corresponding \( x \). The value function \( V(G, D) \) can be therefore written as

\[
V_{GAN}(G, D) = \mathbb{E}_{y}[\log D(y)] + \mathbb{E}_{x,z}[\log(1 - D(G(x, z))].
\] (2)

By also conditioning \( x \) in the discriminator \( D \), we can get the value function in its complete expression:

\[
V_{GAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))].
\] (3)

where the generator \( G \) tries to minimise the function to fool the discriminator \( D \), whereas the discriminator \( D \) tries to...
maximise the function to spot the fake data generated by $G$, i.e.,

$$\arg\min_G \max_D V_{GAN}(G, D).$$ (4)

To make the synthetic omics data $y^t$ more correlated to the corresponding $x$, we further added a distance constrain to the objective of OmiTrans. The generator $G$ tries to produce the synthetic omics data $y^t$ that not only fool the discriminator $D$ but be as similar as possible to the original omics data $y$, and the value function of the distance constrain is

$$V_{dist}(G) = \mathbb{E}_{x,y,z}[\text{dist}(y - G(x, z))].$$ (5)

In the OmiTrans framework, we provide L1, mean square error (MSE) and cross-entropy (CE) as the distance function $\text{dist}()$. With the distance constrain, the final objective of OmiTrans can be formalised as

$$\arg\min_G \max_D V_{GAN}(G, D) + \lambda V_{dist}(G)$$ (6)

where $\lambda$ balances the GAN loss and the distance loss.

C. Network Architecture

The OmiTrans framework supports various implementations of generator $G$ and discriminator $D$. For the similar image translation task, most solutions [17], [32]–[35] adopted deep convolutional neural networks (CNNs) to implement $G$ and $D$ because the characteristics of images fit the local connectivity and parameter sharing properties of convolutional layers, and the underlying structure of the input image is aligned with that of the output image. However, it is not applicable to omics data because the format of omics data is not 2D/3D grid where each pixel/voxel is connected to its neighbours normally with similar values, and the input omics data and the output data with different dimensionality do not share the same underlying structure.

Therefore, the most straightforward implementation of the OmiTrans generator $G$ is an encoder-decoder network using multiple fully-connected (FC) blocks, which is comprised of a fully-connected layer, a normalisation layer, a dropout layer and an activation layer, as illustrated in Supplementary Figure 1. Unlike the translation of image data, the correlation between the input features and output features is not just one-to-one mapping but a combination of one-to-one, one-to-many, many-to-one, and many-to-many mappings. Some of the correlations between molecular features (e.g., gene, CpG site, miRNA, and protein) of two omics types have been discovered by biologists, whereas others are still unknown. Fully-connected layers can capture any type of the correlations mentioned above, which justifies the application of FC blocks in the generator. Similar to the FC generator, the OmiTrans discriminator $D$ can also be implemented by fully-connected blocks as shown in Supplementary Figure 2. The input vectors of omics A and omics B are first concatenated together and then fed to a multi-layer fully-connected network to produce an output vector discriminating whether the input vector of omics A is real or fake.

Although convolutional neural networks are naturally not suitable for omics data, we can still make 1D convolutional layers workable by applying some additional preprocessing steps to the omics data, including balancing the dimension- alities of the two omics types and rearranging the orders of the molecular features (e.g., gene, CpG sites, miRNA) based on their genomic location (a.k.a., location alignment). Thus, CNN-based networks can also be adapted to the generator and discriminator of OmiTrans. We modified the U-Net architecture [36] to make it compatible with omics data, as illustrated in Supplementary Figure 3. The discriminator of OmiTrans can also be implemented by a CNN as shown in Supplementary Figure 4. The performance of both the FC-based architecture and the CNN-based architecture was evaluated in our experiments, and any state-of-the-art or future network can also be added to the OmiTrans framework as the generator or discriminator using source code from our open source repository with minimal modification.

D. Comparing Methods

The OmiTrans framework was built on PyTorch [37], and has been made open source through GitHub. OmiTrans is compatible with the OmiEmbed [3] multi-omics multi-task framework and well-organised with modular code structures, predefined packages and easy-to-follow tutorials. The model of OmiTrans was trained on our experiment platform with two NVIDIA Titan X GPUs, one 6-core 3.40GHz Intel Core i7-6800K CPU, and 96 GB of memory, which is a standard configuration that is easy to reproduce.

We have also compared OmiTrans with other methods, including the latest TDimpute [8] that utilised transfer learning-based neural networks to impute missing RNA-Seq data from DNA methylation, the traditional linear regression (LR), and LASSO regression [10]. Unlike the implementation of [7] that only connected each gene with CpG sites mapped to it, we fully connected each gene to each CpG site for LR and LASSO to capture all of the potential correlations between genome-wide DNA methylation profile and gene expression profile. TDimpute was run on the same platform mentioned above for OmiTrans. In contrast, LR and LASSO could not perform on our experiment platform due to the extremely high memory usage and the extremely long running time. Therefore, we ran the experiments of LR and LASSO on Alibaba Cloud Elastic Container Service (ECS) with 80 vCPU of Intel Xeon (Cascade Lake) Platinum 8269CY and 192 GB of memory using multiprocessing and celer [38], [39]. In order to minimise the running time, we applied process-based parallelism to solve LASSO, and estimated the expression values of each gene separately in an individual process instead of calculating the output matrix together. Even with the top configuration and the parallelism strategy, the training time of LASSO is around 154 hours, which is unacceptable for everyday use compared to the 12-hour training time of OmiTrans. As for the implementation of linear regression, we built a one-

3https://github.com/zhangxiaoyu11/OmiTrans/
A. Reconstruction Performance

Since both the DNA methylation profiles and the corresponding gene expression profiles are available in the GDC datasets, we are able to compare the synthetic data with the original data to evaluate the reconstruction performance of each model. 9,081 samples in the GDC datasets were randomly separated into training, validation and testing sets. Results on the testing set using the trained model were used to evaluate the performance of each method.

As tabulated in Table I, we used nine different metrics to evaluate the reconstruction performance of six methods: FC-based OmiTrans, CNN-based OmiTrans, TDimpute [8], LASSO regression [10], conditional GAN, and traditional linear regression (LR). The first four metrics were used to measure the distance between the original omics data and the synthetic omics data, including mean square error (MSE), root mean square error (RMSE), mean absolute error (Mean AE) and median absolute error (Median AE). The remaining five metrics are variants of the coefficient of determination ($R^2$) which is a more intuitively informative metric that measures how well observed data are replicated by a model [40]. Zhong et al. [7] used featurewise $R^2$ ($R^2_f$), while Zhou et al. [8] used samplewise $R^2$ ($R^2_s$). We included both variants in the evaluation to represent the results more comprehensively. Since featurewise (i.e., calculate the coefficient of determination for each gene) $R^2_f$ more faithfully, we further analysed $R^2_f$ by calculating both the mean $R^2_f$ and median $R^2_f$, counted the number of genes with $R^2_f$ larger than the threshold, and computed the percentage of genes with $R^2_f$ larger than the threshold. We set the threshold to 0.3 in our analysis.

OmiTrans with the FC-based architecture got the best performance in all nine metrics, outperforming the CNN-based OmiTrans with a U-Net generator, because the fundamental properties of convolutional neural networks like local connectivity and parameter sharing do not fit the format of omics data even with the location alignment. Figure 2 shows scatter graphs of the real gene expression levels and the synthetic expression levels for six given samples with highest reconstruction performance using FC-based OmiTrans. Moreover, the scatter graphs for the top six genes with highest $R^2$ values were illustrated in Supplementary Figure 5.

The OmiEmbed [3] multi-omics multi-task framework can be used to compare the performance of each method further. First, we fully trained an OmiEmbed [3] model on the pan-
Fig. 3. The scatter graphs of the original training set of GDC, the fake testing set of GDC synthesised by OmiTrans, and both sets together. Samples with different tumour types were marked with different colours, original data and synthetic data were marked with different symbols, as shown in the legend. The full name of each tumour type is listed in Supplementary Table 2.

TABLE I
Reconstruction performance of six methods using nine different metrics. For the first four metrics, lower value means better performance. For the remaining metrics, higher value means better performance.

| Method           | MSE    | RMSE   | Mean AE  | Median AE | Mean $R^2_s$ | Mean $R^2_f$ | Median $R^2_f$ | # $R^2_f > 0.3$ | % $R^2_f > 0.3$ |
|------------------|--------|--------|----------|-----------|--------------|--------------|----------------|----------------|----------------|
| OmiTrans-FC      | 0.1097 | 0.3204 | 0.1538   | 0.0396    | 0.9453       | 0.3556       | 0.4167         | 37,717         | 62.36%         |
| OmiTrans-CNN     | 0.1747 | 0.4062 | 0.1948   | 0.0453    | 0.9311       | -3.4496      | 0.0506         | 15,145         | 25.04%         |
| TDimpute         | 0.1141 | 0.3279 | 0.1809   | 0.0872    | 0.9432       | -27.8159     | 0.3210         | 31,097         | 51.41%         |
| LASSO            | 0.1162 | 0.3295 | 0.1645   | 0.0513    | 0.9422       | 0.3491       | 0.3440         | 32,798         | 54.23%         |
| Conditional GAN  | 0.1988 | 0.4345 | 0.2035   | 0.0413    | 0.9010       | -0.3598      | 0.0950         | 14,023         | 23.19%         |
| Linear Regression| 0.4602 | 0.6666 | 0.5128   | 0.4178    | 0.7701       | -2078.2833   | -3.7674        | 4,922          | 8.14%          |

cancer multi-class classification benchmark task mentioned in [41] using our GDC training set. The pre-trained model was then fed with the original testing data and the six fake testing data synthesised by the six methods mentioned above. The classification performance of each method is shown in Table II. Better classification performance means the synthetic
data is more distinguishable, considering the cancer label of each sample, and more similar to the original data. FC-based OmiTrans still got the best performance in all the five classification metrics, including accuracy, precision, recall, F1 score (F1) and area under the receiver operating characteristic curve (AUC).

Since the FC-based OmiTrans got the best performance of reconstructing genome-wide expression profile from DNA methylation data, we further analysed the results from OmiTrans (henceforth, OmiTrans refers to the FC-based OmiTrans). The histogram of the samplewise $R^2$ and the featurewise $R^2$ were illustrated in Supplementary Figure 6 and Supplementary Figure 7. The mean and median $R^2$ values were also marked in the histogram. To observe the distribution of the gene expression data synthesised by OmiTrans visually, we used OmiEmbed [3] to reduce the dimensionalities of the synthetic testing set and the original training set of TCGA to 128. We then used t-SNE [42] to visualise both the synthetic and original data in scatter graphs as shown in Figure 3.

As can be seen in Figure 3, the synthetic data shared a similar distribution in the latent space, which indicated that OmiTrans faithfully reconstructed the gene expression profiles from the DNA methylation data.

B. Testing on Individual Dataset

All of the aforementioned experiments were tested on datasets from GDC. Thus, it remained unknown whether an OmiTrans model trained on datasets from GDC still worked on another individual dataset. To further evaluate the generalisation of OmiTrans, we applied the generator of an OmiTrans model that had been fully trained on the GDC training dataset to a new and previously unseen dataset: the brain tumour DNA methylation dataset (BTM) GSE109381 [30]. The OmiTrans generator synthesised the gene expression profile from each DNA methylation profile in the BTM dataset. We then used OmiEmbed [3] to reduce the dimensionalities of the original DNA methylation data and the synthetic gene expression data of BTM to 128 and then used t-SNE [42] to visualise them, as illustrated in Supplementary Figure 8.

Even though the OmiTrans model was trained on the GDC training set and had never seen any data from the BTM dataset, it is able to synthesise gene expression profiles that we did not have for every sample in BTM. As we can see in Supplementary Figure 8, the synthetic gene expression data kept the multi-level hierarchical clustering pattern of the original DNA methylation data, although the OmiTrans model had zero information about the BTM dataset and the brain tumour classification criteria, which means a pre-trained OmiTrans model can be easily transferred to another dataset without any a priori knowledge. Unlike BTM (3,905 samples) and the GDC datasets (9,081 samples), most omics datasets are relatively small-scale, which is unsuitable for deep learning. However, with the model generalisation of OmiTrans, we can pre-train an OmiTrans model on a large multi-omics dataset like GDC and adopt the model to the dataset-of-interest with a small sample size, which vastly increases the applicability and practicability of OmiTrans.

IV. Conclusion

Omic-to-omics translation is a brand new research area that has been hardly studied. Inspired by deep learning-based machine translation, we proposed the assumption that each type of omics profile of the same clinical sample shared the same latent representation, and each type of omics profile can be generated from this latent representation. Based on this idea, we introduce OmiTrans, which is the first GANs based omics-to-omics translation framework to the best of our knowledge. OmiTrans is able to faithfully generate synthetic omics data from corresponding original data of another omics type. With the model generalisation ability, an OmiTrans model can be trained on a large-scale multi-omics dataset like TCGA and adopted to small-scale dataset-of-interest, which vastly increase the applicability and practicability. In future, we will further upgrade OmiTrans using the cycle-consistent adversarial networks and take advantage of the enormous amount of unpaired omics data from public data repositories.

V. Availability

The source code have been made publicly available on GitHub. The TCGA dataset can be downloaded from the UCSC Xena data portal. The latest release of OmiTrans has also been stored online in Zenodo repository under the doi:10.5281/zenodo.5728496. The BTM dataset is available from GEO with the accession ID GSE109381.

VI. Supplementary

Supplementary file is available online at GitHub.

REFERENCES

[1] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, “Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification,” Nature Communications, vol. 12, no. 1, pp. 1–13, 2021.

[4]https://github.com/zhangxiaoyu11/OmiTrans/

[5]https://xenabrowser.net/datapages/

[6]https://zenodo.org/record/5728496

[7]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109381

[8]https://github.com/zhangxiaoyu11/OmiTrans/blob/main/documents/supplementary.pdf
