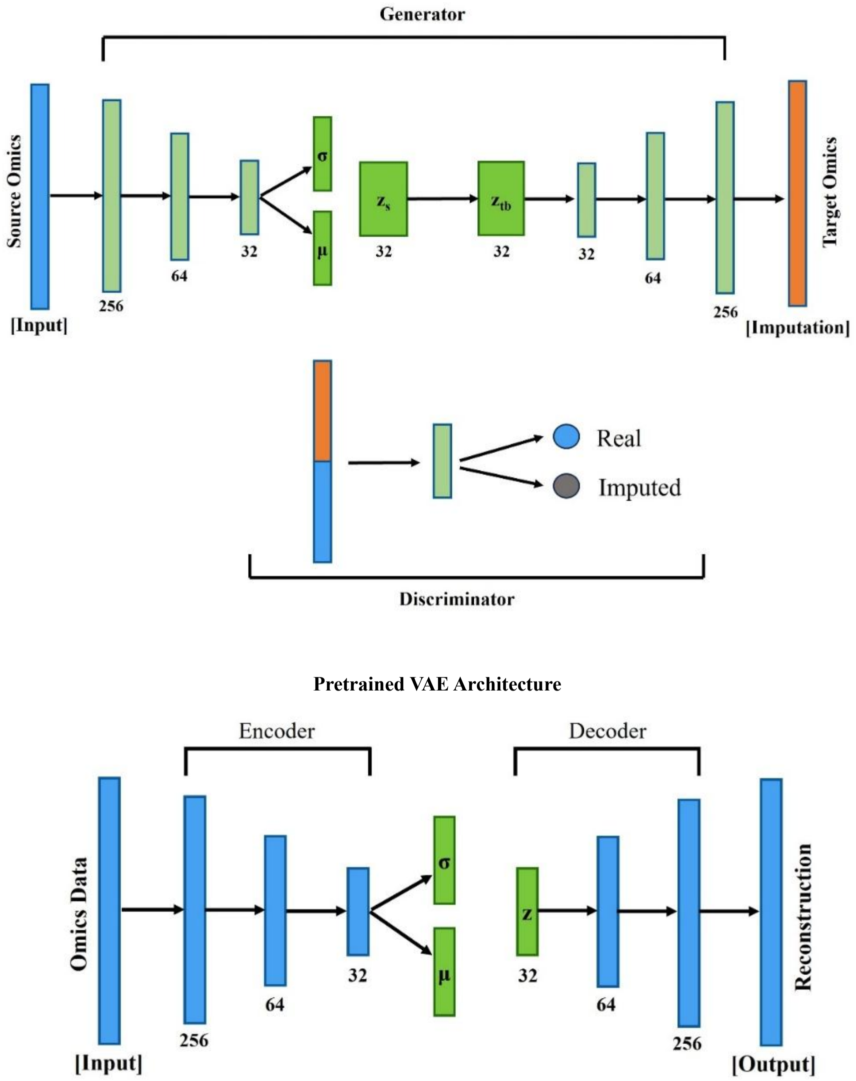
OmiImp: A Cross-Omics Imputation Framework Based on Improved Generative Adversarial Network

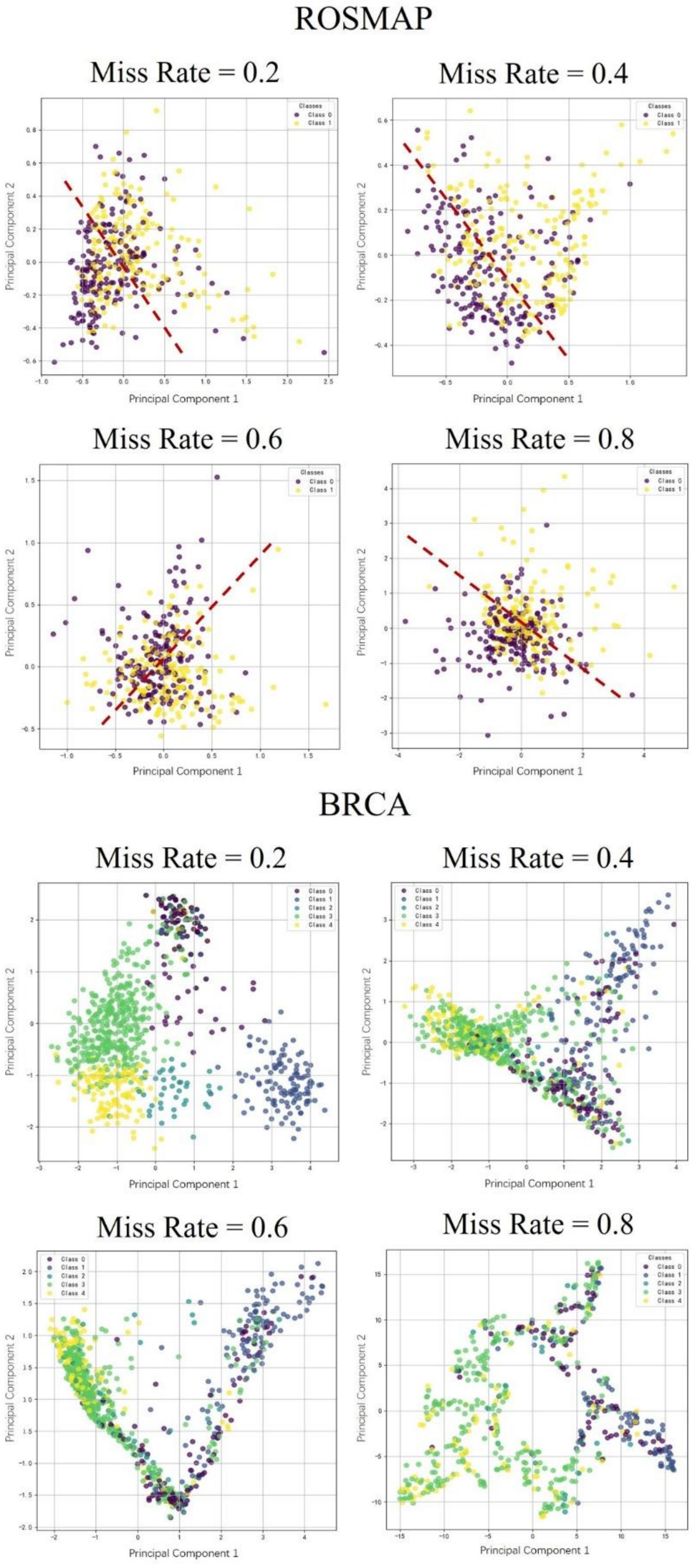
This document provides the supplementary figures and tables for the manuscript.

Supplementary Table 1: Dataset Description

Datasets	Num of Sample	Dimension of Source Omics	Dimension of target Omics
ROSMAP	**398**	**448**	**383**
BRCA	**875**	**1000**	**1000**



Supplementary Figure 1: The implementation of the OmiImp Architecture. The dimensionalities of the model are marked in the diagram.



Supplementary Figure 2. Evaluation of Imputation Quality Using PCA on RNA-seq Data. This figure illustrates the performance of OmiImp for cross-omics imputation across two datasets: ROSMAP and BRCA. For ROSMAP, imputation is performed from SNP to RNA-seq, while for BRCA, imputation is from DNA methylation to RNA-seq. The plots show the results of PCA applied to the imputed RNA-seq data to assess the quality of imputation. Each subplot represents different missing rates (0.2, 0.4, 0.6, 0.8), indicated at the top of each plot. The distinct colors represent different classes.

Supplementary Table 2: Comparison of OmiImp with Baselines on BRCA (Missing Rate = 0.2)

	MAE ↓	**MSE ↓**	**RMSE ↓**	**Mean R² ↑**	**% R²_f > 0.3 ↑**
OmiImp	0.0539	0.0064	0.0800	0.8110	99.3%
Lasso	**0.0414**	**0.0038**	**0.0616**	**0.8895**	**100.0%**
TOBMI	0.1419	0.0508	0.2253	0.1164	77.1%
OmiTrans	0.0621	0.0103	0.1015	0.7121	97.9%
TDimpute	0.0435	0.0092	0.0959	0.8790	**100.0%**
OmicsNMF	0.0591	0.0094	0.0970	0.7792	99.1%
DeepGAMI	0.0615	0.0077	0.0877	0.7933	92.1%
TMO-NET	0.0915	0.0100	0.1000	0.5919	74.6%

Supplementary Table 3: Comparison of OmiImp with Baselines on BRCA (Missing Rate = 0.4)

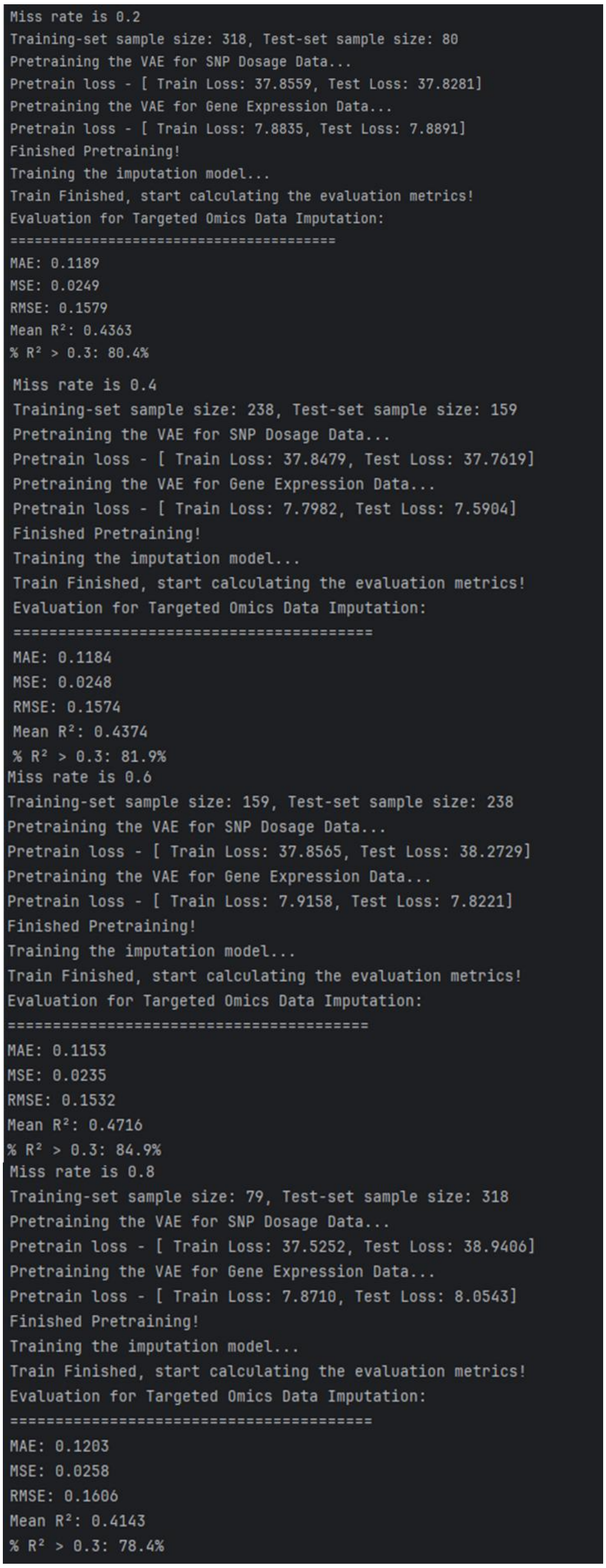
	MAE ↓	**MSE ↓**	**RMSE ↓**	**Mean R² ↑**	**% R²_f > 0.3 ↑**
OmiImp	0.0560	0.0068	0.0825	0.8094	**99.2%**
Lasso	**0.0441**	**0.0044**	**0.0663**	**0.8334**	97.8%
TOBMI	0.1148	0.0368	0.1919	0.1871	82.6%
OmiTrans	0.0711	0.0094	0.0970	0.7421	87.3%
TDimpute	0.0655	0.0104	0.1020	0.7704	98.2%
OmicsNMF	0.0796	0.0101	0.1005	0.7113	88.1%
DeepGAMI	0.0613	0.0057	0.0755	0.7745	96.0%
TMO-NET	0.0841	0.0176	0.1326	0.6782	79.6%

Supplementary Table 4: Comparison of OmiImp with Baselines on BRCA (Missing Rate = 0.6)

	MAE ↓	**MSE ↓**	**RMSE ↓**	**Mean R² ↑**	**% R²_f > 0.3 ↑**
OmiImp	**0.0581**	0.0072	0.0850	**0.7853**	**98.8%**
Lasso	0.0672	0.0090	0.0949	0.7535	96.0%
TOBMI	0.0814	0.0180	0.1342	0.4608	88.7%
OmiTrans	0.0891	0.0092	0.0950	0.6941	91.2%
TDimpute	**0.0651**	0.0076	0.0777	0.7243	98.1%
OmicsNMF	0.0836	0.0100	0.1000	0.7018	84.9%
DeepGAMI	0.0941	**0.0043**	**0.0655**	0.7749	97.4%
TMO-NET	0.0786	0.0139	0.1179	0.7447	89.3%

Supplementary Table 5: Comparison of OmiImp with Baselines on BRCA (Missing Rate = 0.8)

	MAE ↓	**MSE ↓**	**RMSE ↓**	**Mean R² ↑**	**% R²_f > 0.3 ↑**
OmiImp	0.0583	0.0063	0.0794	0.7819	99.0%
Lasso	0.0682	0.0081	0.0900	**0.8485**	96.6%
TOBMI	0.0966	0.0259	0.1609	0.3371	85.9%
OmiTrans	0.0725	0.0091	0.0954	0.7119	**99.4%**
TDimpute	**0.0498**	**0.0054**	**0.0734**	0.7471	**100.0%**
OmicsNMF	0.0741	0.0075	0.0866	0.7324	97.1%
DeepGAMI	0.0828	0.0090	0.0949	0.6809	98.6%
TMO-NET	0.0879	0.0077	0.0877	0.7771	94.8



```
Miss rate is 0.2
Training-set sample size: 318, Test-set sample size: 80
Pretraining the VAE for SNP Dosage Data...
Pretrain loss - [ Train Loss: 37.8559, Test Loss: 37.8281]
Pretraining the VAE for Gene Expression Data...
Pretrain loss - [ Train Loss: 7.8835, Test Loss: 7.8891]
Finished Pretraining!
Training the imputation model...
Train Finished, start calculating the evaluation metrics!
Evaluation for Targeted Omics Data Imputation:
========================================
MAE: 0.1189
MSE: 0.0249
RMSE: 0.1579
Mean R²: 0.4363
% R² > 0.3: 80.4%
Miss rate is 0.4
Training-set sample size: 238, Test-set sample size: 159
Pretraining the VAE for SNP Dosage Data...
Pretrain loss - [ Train Loss: 37.8479, Test Loss: 37.7619]
Pretraining the VAE for Gene Expression Data...
Pretrain loss - [ Train Loss: 7.7982, Test Loss: 7.5904]
Finished Pretraining!
Training the imputation model...
Train Finished, start calculating the evaluation metrics!
Evaluation for Targeted Omics Data Imputation:
========================================
MAE: 0.1184
MSE: 0.0248
RMSE: 0.1574
Mean R²: 0.4374
% R² > 0.3: 81.9%
Miss rate is 0.6
Training-set sample size: 159, Test-set sample size: 238
Pretraining the VAE for SNP Dosage Data...
Pretrain loss - [ Train Loss: 37.8565, Test Loss: 38.2729]
Pretraining the VAE for Gene Expression Data...
Pretrain loss - [ Train Loss: 7.9158, Test Loss: 7.8221]
Finished Pretraining!
Training the imputation model...
Train Finished, start calculating the evaluation metrics!
Evaluation for Targeted Omics Data Imputation:
========================================
MAE: 0.1153
MSE: 0.0235
RMSE: 0.1532
Mean R²: 0.4716
% R² > 0.3: 84.9%
Miss rate is 0.8
Training-set sample size: 79, Test-set sample size: 318
Pretraining the VAE for SNP Dosage Data...
Pretrain loss - [ Train Loss: 37.5252, Test Loss: 38.9406]
Pretraining the VAE for Gene Expression Data...
Pretrain loss - [ Train Loss: 7.8710, Test Loss: 8.0543]
Finished Pretraining!
Training the imputation model...
Train Finished, start calculating the evaluation metrics!
Evaluation for Targeted Omics Data Imputation:
========================================
MAE: 0.1203
MSE: 0.0258
RMSE: 0.1606
Mean R²: 0.4143
% R² > 0.3: 78.4%
```

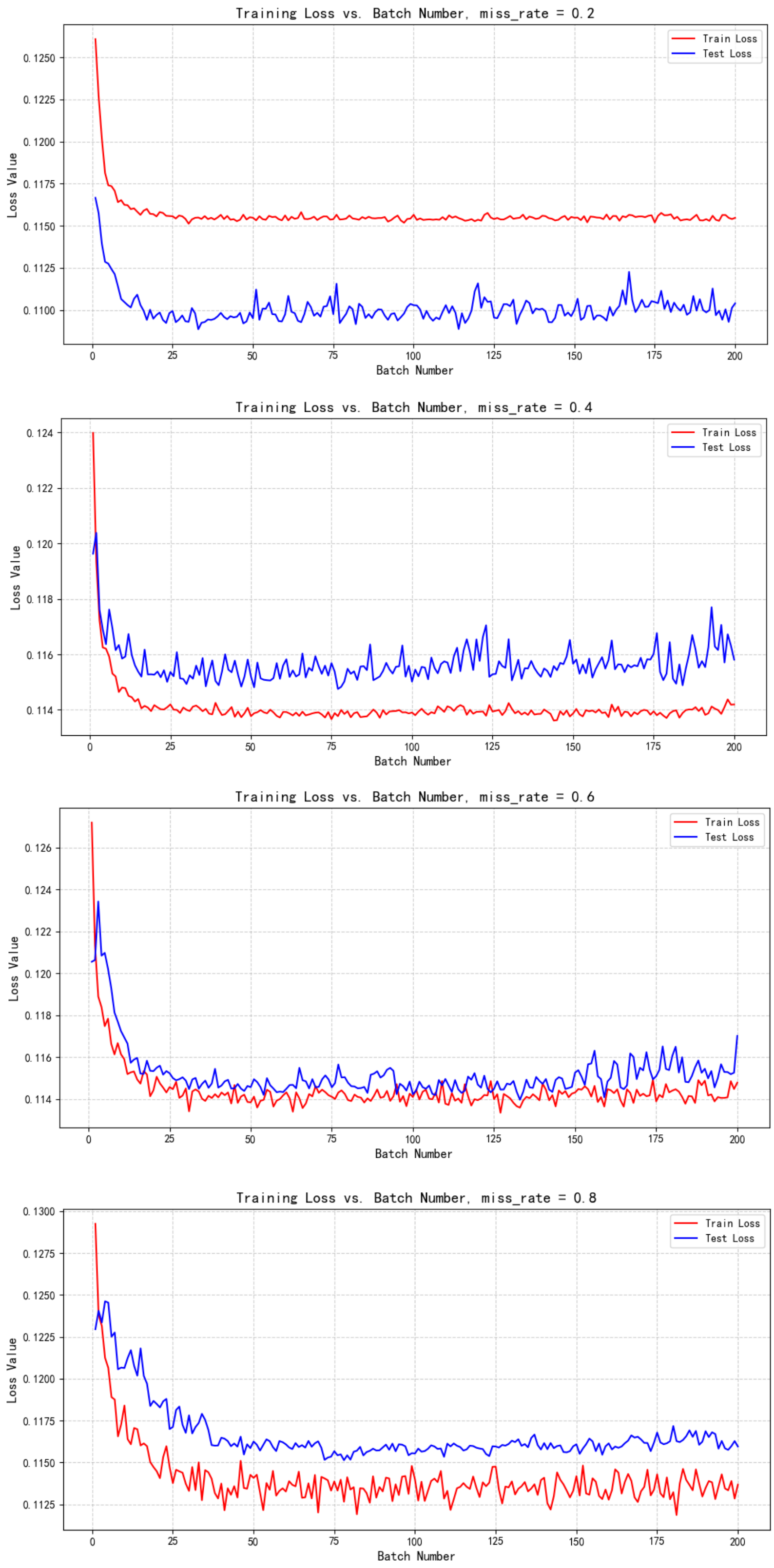
Supplementary Figure 3. Results from Running on the ROSMAP. Due to the limitation of the ROSMA, OmiImp cannot provide the raw data. Therefore, we present the terminal results of OmiImp's performance on the ROSMAP. Additionally, we have provided the specific location for obtaining the source data and have made the complete data processing procedures available on GitHub for method replication.

Supplementary Table 6: Genes with high imputation errors in the ROSMAP for enrichment analysis.

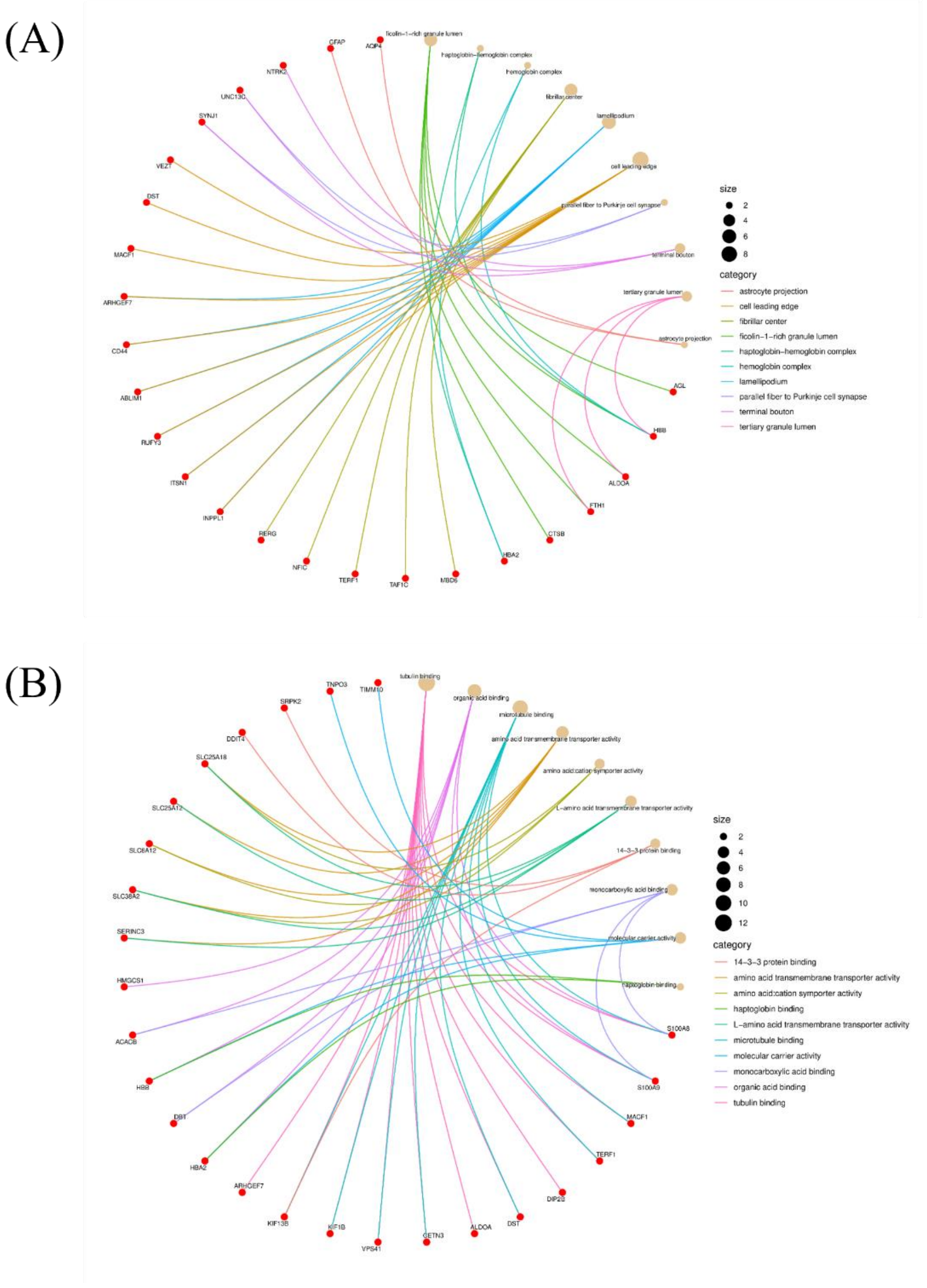
ROSMAP (gene, mean error)
('IL1RL1',0.2769794166088104), ('VGF',0.17709074914455414), ('RRAGD',0.17358125746250153), ('FKBP5', 0.1711604744195938), ('SST', 0.1705254316329956), ('RASL12', 0.16886985301971436), ('GFAP', 0.16580837965011597), ('CHCHD2', 0.1634678691625595), ('HBA2.1', 0.16123740375041962), ('PDK4', 0.15991394221782684), ('SERINC3', 0.15939074754714966), ('PPP2R2B', 0.1580149084329605), ('HBA2.3', 0.15725208818912506), ('PSMC6', 0.15625983476638794), ('INPPL1', 0.15590783953666687), ('DBT', 0.15479812026023865), ('SRRM2', 0.15351635217666626), ('STK36', 0.1533292979001999), ('TEAD2', 0.15329138934612274), ('ACADVL', 0.1523379534482956), ('ACADVL.1', 0.15134285390377045), ('ITGB5.1', 0.15053997933864594), ('CGNL1', 0.15043772757053375), ('ITGB5.5', 0.1503130942583084), ('PTBP1', 0.15029725432395935), ('PTBP1.1', 0.1500462293624878), ('SRPK2', 0.15000177919864655), ('HBA2.2', 0.14905591309070587), ('DDIT4', 0.14880827069282532), ('S100A9', 0.14866122603416443), ('S100A8', 0.14865577220916748), ('SYTL2.1', 0.1485312134027481), ('TERF1', 0.1483224481344223), ('ZNF160', 0.1481068730354309), ('ATPIF1', 0.1478842943906784), ('NRXN1.1', 0.14742854237556458), ('VEZT', 0.1471608281135559), ('AGL', 0.14673134684562683), ('AP1G2.1', 0.14670568704605103), ('PLXNB1', 0.14664065837860107), ('HOPX.2', 0.14630421996116638), ('ACACB', 0.14617487788200378), ('TAF1C', 0.14616002142429352), ('TMEM14B', 0.14604848623275757), ('TCF4', 0.14600972831249237), ('RGS5.1', 0.14577001333236694), ('DIP2B.1', 0.14568381011486053), ('RGS5', 0.145640030503273), ('CREG2', 0.14530889689922333), ('ITPKB', 0.14513573050498962)

Supplementary Table 7: Genes with high imputation errors in the BRCA for enrichment analysis

BRCA (gene, mean error)
('TFF1', 0.17003846168518066), ('C20orf114', 0.14417332410812378), ('CYP2B7P1', 0.14252127707004547), ('SOX10', 0.14150941371917725), ('ANKRD30A', 0.13983199000358582), ('ABCC11', 0.1390864998102188), ('GP2', 0.13765811920166016), ('GRPR', 0.13590297102928162), ('KRT6B', 0.13482166826725006), ('C1orf64', 0.13093417882919312), ('TFF3', 0.13061878085136414), ('PGR', 0.1276731640100479), ('KLK7', 0.12760448455810547), ('CA9', 0.12704510986804962), ('GABRP', 0.12472674995660782), ('PGLYRP2', 0.12440887093544006), ('MSLN', 0.12377319484949112), ('WNK4', 0.12341295927762985), ('NEK10', 0.12325531989336014), ('ANKRD30B', 0.12306058406829834), ('KLK6', 0.12265763431787491), ('CAPN8', 0.12232110649347305), ('KLHDC7A', 0.1214643344283104), ('PTPRT', 0.11971765756607056), ('FABP7', 0.11964398622512817), ('PTPRZ1', 0.11960896104574203), ('SERPINA11', 0.11894254386425018), ('SLC6A14', 0.11867886781692505), ('SERPINA5', 0.11739855259656906), ('DSC3', 0.11620905250310898), ('CHAD', 0.1158611848950386), ('AGR3', 0.11563336849212646), ('CYP4Z2P', 0.1151723697185516), ('ZBTB16', 0.11483877152204514), ('SYTL5', 0.11371797323226929), ('SYT9', 0.11263220757246017), ('AGR2', 0.11231307685375214), ('C2orf40', 0.1120322048664093), ('KRT16', 0.11175376921892166), ('DSG1', 0.11148282885551453), ('ANXA8L2', 0.11117774993181229), ('KLK8', 0.11028177291154861), ('SOSTDC1', 0.10960229486227036), ('LY6D', 0.10957033187150955), ('A2ML1', 0.10879737883806229), ('NOVA1', 0.1086173802614212), ('MIA', 0.10805920511484146), ('GFRA1', 0.10800483077764511), ('CLDN19', 0.10653168708086014), ('SERPINB5', 0.10632789880037308)



Supplementary Figure 4. Loss Reduction of OmiImp on ROSMAP. The figure illustrates that the introduction of a pre-trained VAE enables the generator to produce good imputation performance from the initial stage, which also promotes rapid convergence of the model. The figure compares the training and testing loss of the model under missing rates of [0.2, 0.4, 0.6, 0.8], further validating the robustness of the model.



Supplementary Figure 5. Validation of No Significant Gene Enrichment for Genes with High Imputation Errors. Panels A and B present the enrichment analysis results for genes with high imputation errors, categorized into Cellular Component (CC) and Molecular Function (MF), respectively. In Panel A (CC), the analysis does not reveal any significant enrichment of genes within specific cellular components. Similarly, Panel B (MF) shows no significant enrichment in molecular functions associated with these genes, with the network displaying a dispersed pattern without clear functional groupings. This analysis helps to validate the assumption that the genes with larger imputation errors do not have a significant impact on the biological interpretations derived from enrichment analysis.