DIFFUSE: Predicting isoform functions from sequences and expression via deep learning (Supplementary Materials)

Hao Chen[1], Dipan Shaw[1], Jianyang Zeng[2], Dongbo Bu[3,4] and Tao Jiang[1,5]

[1]Department of Compute Science and Engineering, University of California, Riverside, CA 92521, USA
[2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China
[3]Key Lab of Intelligent Information Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[4]University of Chinese Academy of Sciences, Beijing 100049, China
[5]Bioinformatics Division, BNRIST/Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

1 Supplementary figures

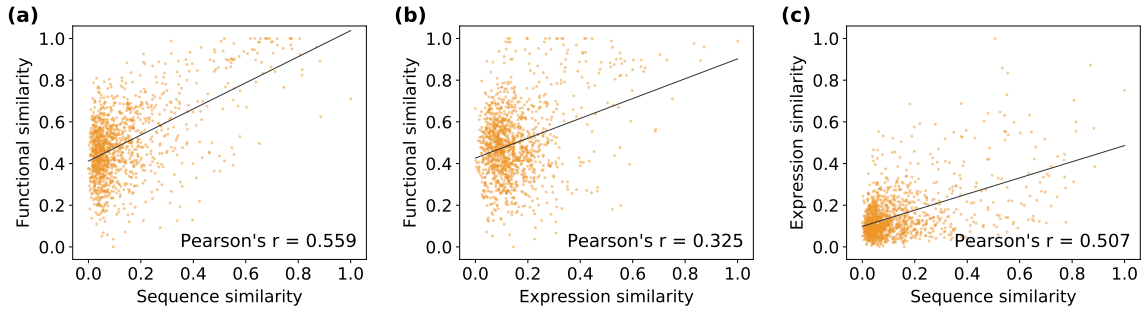

Figure S1: Correlations between functional similarity, sequence similarity and expression similarity at the gene level. The genes are grouped into 1254 clusters with sizes in the range of [10, 20] based on hierarchical clustering. The average pairwise functional similarity, sequence similarity and expression similarity are estimated for each cluster. The Pearson correlation coefficient (PCC) is used to measure the correlations.

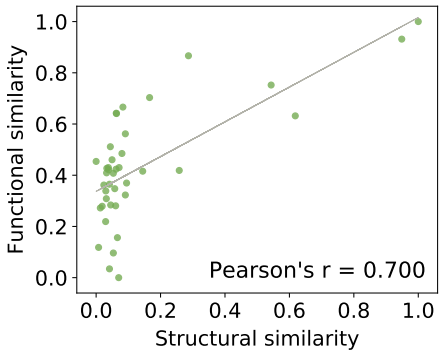

Figure S2: The correlation between functional similarity and structural similarity measured on 600 SIGs that are grouped into 40 clusters with sizes in the range of [10, 20]. The functional similarity is estimated based on the annotated functions of the SIGs. The result is consistent with the PCC value in Figure 7(a).

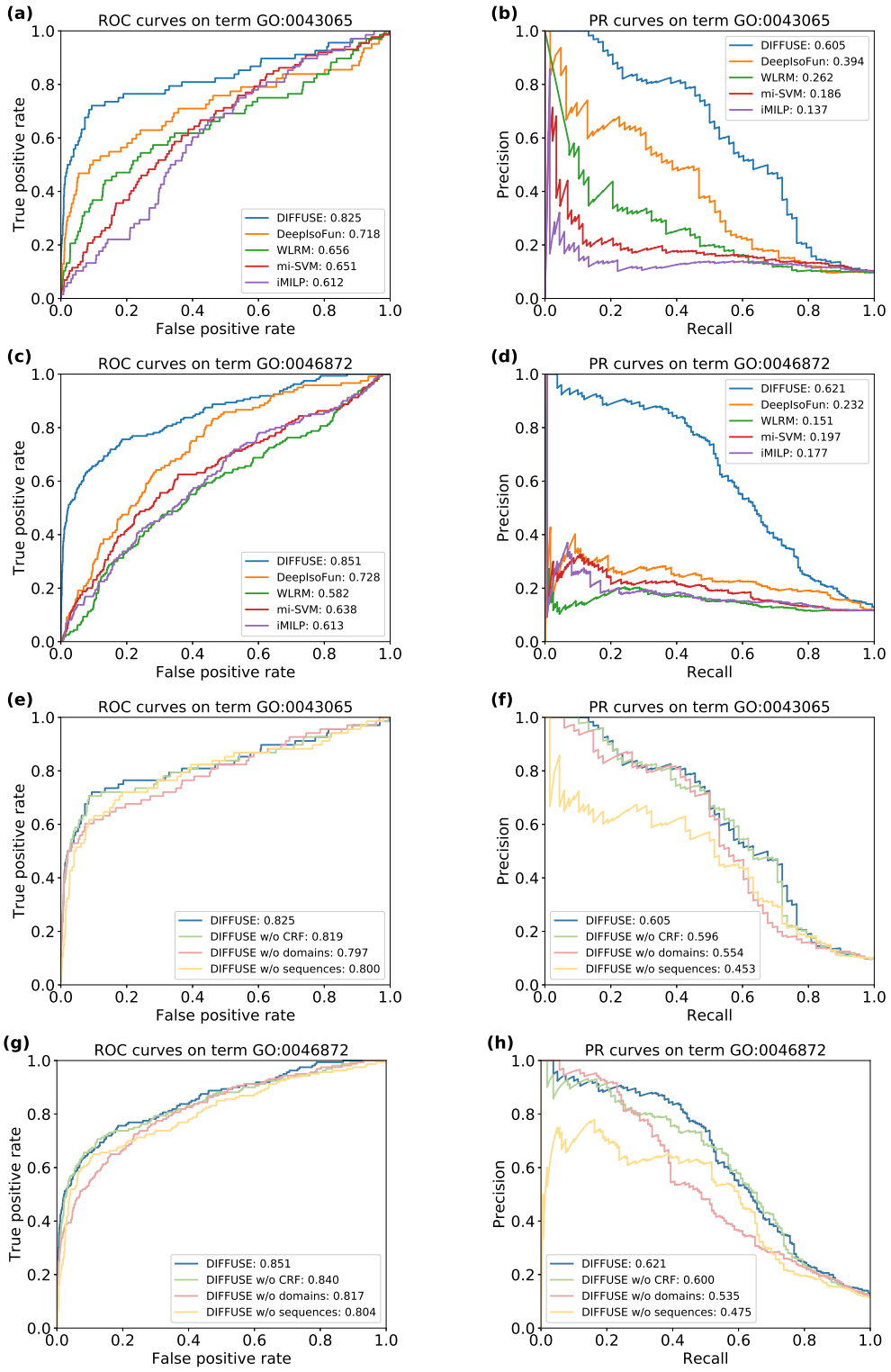

Figure S3: Example receiver operating characteristic (ROC) curves and precision-recall (PR) curves on terms GO:0043065 (positive regulation of apoptotic process) and GO:0046872 (metal ion binding). The term GO:0043065 has been studied extensively in the literature (Li et al., 2013; Shaw et al., 2018) and the term GO:0046872 is used to validate our prediction results in Table 2. The plots (a-d) illustrate the ROC and PR curves achieved by the methods compared in Table 1. The curves on both GO terms demonstrate that DIFFUSE performs better than the other methods across all thresholds on false positive rate or recall. The plots (e-h) illustrate the ROC and PR curves on the same GO terms achieved by the four variants of DIFFUSE discussed in section 3.1.3. The PR curves on both GO terms suggest that after removing sequence features from DIFFUSE, the model predicts more false positives with high scores, which may explain why the AUPRC of DIFFUSE drops so significantly without using sequences.

2 Supplementary tables

Table S1: Comparison of the performance of DIFFUSE in training and testing on all three datasets. The average performance gaps across the three datasets are 6.3% in terms of AUC and 8.3% in terms of AUPRC. These are well within acceptable ranges reported in the literature (Zhang et al., 2016; Keskar et al., 2016) and thus likely to indicate that our model was not grossly overtrained in the experiments.

	Dataset#1		Dataset#2		Dataset#3	
	AUC	AUPRC	AUC	AUPRC	AUC	AUPRC
DIFFUSE *training*	0.891	0.632	0.882	0.588	0.875	0.574
DIFFUSE *test*	0.835	0.585	0.828	0.537	0.817	0.524

Table S2: Consistency between the presence or absence of sequence feature 'Metal ion binding site' and the function predictions concerning GO term GO:0046872 (metal ion binding). Note that a metal ion may have several binding sites. We treat the binding sites that correspond to the same metal ion as a group. Each isoform sequence may contain multiple metal ion binding site groups. If all metal ion binding site groups of an isoform have binding sites affected by alternative splicing, we treat the sequence feature 'Metal ion binding site' as absent in this isoform, noted by a cross. Otherwise, we treat it as present in this isoform, noted by a circle. Positive and negative predictions are represented by circles and crosses as well.

Gene	Isoform	Sequence feature	Predictions				
			DIFFUSE	DeepIsoFun	WLRM	mi-SVM	iMILP
ACE	P12821-1	○	○	○	○	○	○
	P12821-3	○	○	○	×	○	○
	P12821-4	○	○	○	○	○	○
ACMSD	Q8TDX5-1	○	○	○	○	○	○
	Q8TDX5-2	×	○	○	○	○	○
ADAM33	Q9BZ11-1	○	○	○	○	○	×
	Q9BZ11-2	○	○	○	○	○	○
ADAMTS13	Q76LX8-1	○	×	○	○	○	○
	Q76LX8-2	○	○	○	○	○	○
	Q76LX8-3	○	×	○	○	○	○
ALKBH2	Q6NS38-1	○	○	○	○	○	○
	Q6NS38-2	×	×	×	○	○	○
ALKBH8	Q96BT7-1	○	○	○	○	○	○
	Q96BT7-4	○	○	○	○	○	○
ALOX15B	O15296-1	○	○	○	×	×	×
	O15296-2	○	○	○	○	○	○
	O15296-4	○	○	○	○	○	○
ALOX5	P09917-1	○	○	○	×	○	×
	P09917-2	○	○	○	○	×	×
	P09917-3	○	○	○	○	○	○
AOC3	Q16853-1	○	○	○	○	○	○
	Q16853-2	×	×	○	○	○	○
	Q16853-3	○	×	○	○	×	○
APOBEC3F	Q8IUX4-1	○	×	○	○	○	○
	Q8IUX4-3	×	○	○	○	○	○
APOBEC3G	Q9HC16-1	○	○	○	○	○	○
ARG1	P05089-1	○	○	○	○	○	○
	P05089-2	○	○	○	○	○	○
ARSA	P15289-2	×	×	○	○	○	×
ASPH	Q12797-1	○	○	○	○	○	○
	Q12797-10	○	○	○	○	×	○
	Q12797-2	×	×	○	○	×	○
	Q12797-3	×	×	○	○	○	○
	Q12797-4	×	×	○	○	○	○
	Q12797-5	×	×	○	○	×	○
	Q12797-6	×	×	○	○	×	○
	Q12797-7	×	×	×	○	○	○
	Q12797-8	×	×	○	○	×	○
	Q12797-9	×	×	○	○	○	○
ATP1A1	P05023-1	○	○	○	○	○	○
	P05023-3	○	○	○	○	○	○
	P05023-4	○	○	×	○	○	○
ATP7A	Q04656-1	○	○	○	○	○	○
	Q04656-5	○	○	○	○	○	○
ATP7B	P35670-1	○	○	○	○	○	○
	P35670-3	○	○	○	○	○	○
DIS3L2	Q8IYB7-1	○	○	×	○	○	○
	Q8IYB7-3	○	×	○	○	○	○
	Q8IYB7-4	×	×	×	○	○	○
DPP3	Q9NY33-1	○	○	○	○	×	○
	Q9NY33-4	○	○	○	○	○	○

ENDOV	Q8N8Q3-1	O	O	O	O	O	O
	Q8N8Q3-2	O	O	O	O	O	O
	Q8N8Q3-3	O	O	O	O	O	O
ENOSF1	Q7L5Y1-1	O	O	O	O	O	×
	Q7L5Y1-6	O	×	×	O	O	O
EPHX2	P34913-1	O	×	O	O	O	O
	P34913-2	×	×	O	×	O	O
	P34913-3	×	O	O	O	O	O
ERAP2	Q6P179-1	O	O	O	O	×	O
	Q6P179-3	O	O	O	O	O	O
	Q6P179-4	×	×	O	O	O	×
FAN1	Q9Y2M0-1	O	O	O	×	O	O
	Q9Y2M0-2	×	O	O	O	O	O
GALNT2	Q10471-1	O	O	O	×	O	O
GALT	P07902-1	O	O	×	×	O	O
	P07902-2	O	×	O	O	O	O
GCH1	P30793-1	O	O	O	×	×	O
	P30793-2	×	×	×	×	O	×
	P30793-4	×	×	O	O	O	O
HEPH	Q9BQS7-3	O	O	O	O	O	O
	Q9BQS7-4	O	O	O	O	×	O
HMGCL	P35914-1	O	O	O	×	×	O
	P35914-2	O	O	O	O	O	O
HMGCLL1	Q8TB92-1	O	O	O	O	O	O
	Q8TB92-2	O	O	O	O	O	×
	Q8TB92-5	O	O	O	O	O	O
IDE	P14735-1	O	O	×	O	O	O
	P14735-2	×	×	O	×	×	O
IMPA1	P29218-1	O	O	O	O	×	O
	P29218-2	O	O	O	O	O	O
	P29218-3	O	O	O	O	O	O
LHPP	Q9H008-1	O	O	O	O	O	O
	Q9H008-2	×	×	O	O	×	O
MASP1	P48740-1	O	O	O	O	O	O
	P48740-2	O	O	O	O	O	O
	P48740-3	O	O	O	O	×	O
MPPE1	Q53F39-1	O	O	O	O	O	O
	Q53F39-4	O	O	O	O	O	O
NOS1	P29475-1	O	O	O	O	O	O
	P29475-3	O	O	O	O	O	O
	P29475-5	O	O	O	O	O	O
PGM1	P36871-1	O	O	O	O	O	O
	P36871-2	O	O	O	O	O	O
	P36871-3	×	×	O	×	×	O
PPM1M	Q96MI6-4	O	O	×	O	×	O
PPP3CA	Q08209-1	O	O	O	O	O	×
	Q08209-2	O	O	O	×	O	O
	Q08209-3	O	O	O	O	O	O
QPCTL	Q9NXS2-1	O	O	O	O	O	O
	Q9NXS2-3	×	×	O	O	×	O
RGN	Q15493-1	O	O	O	O	O	O
	Q15493-2	×	×	O	O	O	O
RPE	Q96AT9-1	O	O	O	O	O	×
	Q96AT9-3	×	O	O	O	×	O
	Q96AT9-4	×	O	O	O	×	×
	Q96AT9-5	×	O	O	O	O	×
SOD2	P04179-1	O	O	O	×	×	O
	P04179-2	×	O	O	O	×	O
	P04179-3	O	O	O	O	O	O
	P04179-4	O	O	O	×	O	O
SUV39H2	Q9H5I1-1	O	O	O	O	O	O
	Q9H5I1-2	O	O	×	O	O	O
	Q9H5I1-3	O	×	O	O	O	O
TET2	Q6N021-1	O	O	×	×	O	O
	Q6N021-2	×	×	O	O	O	O
THTPA	Q9BU02-1	O	O	O	O	O	O
	Q9BU02-2	×	×	O	O	O	O
USP16	Q9Y5T5-1	O	O	×	O	O	O
	Q9Y5T5-2	O	O	O	O	O	O
XPNPEP1	Q9NQW7-1	O	O	O	O	O	O
	Q9NQW7-3	O	O	O	×	O	O
	Q9NQW7-4	O	O	O	O	O	O
Jaccard index			0.674	0.548	0.514	0.534	0.560

Table S3: Consistency between the presence or absence of sequence feature 'ATP binding site' and the function predictions concerning GO term GO:0005524 (ATP binding). Again, note that an ATP maye have several binding sites. We treat the binding sites that correspond to the same ATP as a group. Each isoform sequence may contain multiple ATP binding site groups. If all ATP binding site groups of an isoform have binding sites affected by alternative splicing, we treat the sequence feature 'ATP binding site' as absent in this isoform. Otherwise, we treat it as present in this isoform.

Gene	Isoform	Sequence feature	Predictions				
			DIFFUSE	DeepIsoFun	WLRM	mi-SVM	iMILP
ACLY	P53396-1	O	O	O	O	×	O
	P53396-2	O	×	O	O	×	O
BRSK2	Q8IWQ3-1	O	O	O	O	O	O
	Q8IWQ3-2	O	O	O	O	O	O
	Q8IWQ3-3	O	O	O	O	O	O
	Q8IWQ3-5	O	O	O	O	×	O
	Q8IWQ3-6	×	O	O	O	×	O
CDK11A	Q9UQ88-1	O	O	O	O	O	O
	Q9UQ88-2	O	O	O	O	O	O
	Q9UQ88-4	O	O	×	O	O	O
IDE	P14735-1	O	O	×	×	O	O
	P14735-2	×	×	O	×	×	O
LATS1	O95835-1	O	O	O	O	O	O
	O95835-2	×	×	×	O	×	O
MARK1	Q9P0L2-1	O	O	O	×	O	O
	Q9P0L2-3	O	O	O	O	O	×
MTHFS	P49914-1	O	O	O	O	×	O
OAS2	P29728-1	O	O	O	O	O	O
	P29728-2	O	O	O	O	O	O
	P29728-3	×	×	×	O	×	O
PFKP	[Q01813-2	×	O	O	×	×	×
	Q01813-1	O	O	O	O	O	O
SIK3	Q9Y2K2-5	O	O	O	O	O	O
	Q9Y2K2-8	O	O	O	O	O	O
STK26	Q9P289-1	O	O	O	O	O	O
	Q9P289-2	×	×	O	O	O	O
	Q9P289-3	O	×	O	O	O	O
TAOK1	Q7L7X3-1	O	×	O	O	O	O
	Q7L7X3-3	O	O	O	×	×	×
TSSK4	Q6SA08-1	O	O	O	O	O	O
	Q6SA08-2	O	O	O	×	O	O
	Q6SA08-3	×	×	O	O	O	O
UHMK1	Q8TAS1-1	O	O	O	O	O	O
	Q8TAS1-2	O	O	O	O	×	O
	Q8TAS1-3	×	×	O	×	O	O
WNK1	Q9H4A3-1	O	O	×	O	O	×
	Q9H4A3-5	O	O	O	O	O	O
	Q9H4A3-6	O	O	O	O	O	O
WNK4	Q96J92-1	O	O	O	O	×	O
Jaccard index			0.700	0.595	0.578	0.517	0.581

Table S4: Consistency between the presence or absence of sequence feature 'Nuclear localization signal' and the function predictions concerning GO term GO:0005634 (nucleus). Note that each isoform sequence may contain multiple nuclear localization signals. If all the nuclear localization signals of an isoform are affected by alternative splicing, we treat the sequence feature 'Nuclear localization signal' as absent in this isoform. Otherwise, we treat it as present in this isoform.

Gene	Isoform	Sequence feature	Predictions				
			DIFFUSE	DeepIsoFun	WLRM	mi-SVM	iMILP
ADK	P55263-1	O	O	O	O	O	×
	P55263-2	×	O	×	O	O	O
	P55263-3	O	O	O	×	×	×
	P55263-4	×	×	O	×	×	×
AIFM1	O95831-1	O	O	O	O	O	O
	O95831-3	O	O	O	O	×	×
	O95831-4	×	×	×	O	×	×
APTX	Q7Z2E3-1	O	O	O	O	O	O
	Q7Z2E3-10	O	O	O	O	×	O
	Q7Z2E3-11	O	O	O	O	O	O
	Q7Z2E3-3	×	×	O	O	×	O
	Q7Z2E3-5	O	O	O	O	O	O
	Q7Z2E3-7	O	O	O	O	O	O
	Q7Z2E3-9	O	O	O	O	×	O
DDX25	Q9UHL0-1	O	O	O	O	O	O
	Q9UHL0-2	×	O	O	O	×	×
DNMT1	P26358-1	O	O	O	O	O	O
	P26358-2	O	O	O	O	O	O
ERBB2	P04626-1	O	O	O	×	×	O
	P04626-4	O	O	O	O	O	O
	P04626-5	O	O	O	O	O	O
ERCC2	P18074-1	O	O	×	O	O	O
	P18074-2	×	×	O	×	×	×
HIPK2	Q9H2X6-1	O	O	O	O	O	O
	Q9H2X6-3	O	O	×	O	O	×
JMJD6	Q6NYC1-1	O	O	O	O	O	O
	Q6NYC1-3	O	O	O	O	O	O
MAPK7	Q13164-1	O	O	O	O	O	O
	Q13164-2	O	×	O	O	O	O
MDM2	Q00987-11	O	O	O	O	O	O
	Q00987-5	×	O	O	×	O	O
OGFOD1	Q8N543-2	×	×	O	O	O	×
PAPOLA	P51003-1	O	O	O	O	O	O
	P51003-2	×	×	×	×	O	×
PIAS1	O75925-1	O	O	×	O	O	O
	O75925-2	O	O	O	O	O	O
PIAS2	O75928-1	O	O	O	O	O	O
	O75928-2	O	O	O	O	O	O
	O75928-3	×	×	O	O	O	O
PIK3C2A	O00443-1	O	O	O	O	O	O
PPP1R8	Q12972-1	O	O	O	O	O	O
	Q12972-2	O	O	O	×	O	O
	Q12972-3	×	×	O	×	×	O
REV1	Q9UBZ9-2	O	O	O	O	O	×
RTEL1	Q9NZ71-1	O	O	O	O	O	O
	Q9NZ71-6	O	O	O	×	O	×
	Q9NZ71-7	O	O	O	O	×	O
	Q9NZ71-9	×	O	O	×	×	×
SPAST	Q9UBP0-1	O	O	O	O	O	×
	Q9UBP0-2	O	O	O	O	O	O
USP4	Q13107-1	O	O	×	×	O	O
	Q13107-2	O	×	O	O	O	O
	Q13107-3	×	×	×	O	O	O
WWOX	Q9NZC7-1	O	×	O	O	O	O
	Q9NZC7-3	O	O	O	O	O	×
Jaccard index			0.700	0.579	0.580	0.569	0.521

References

Keskar, N. S., Mudigere, D., Nocedal, J., et al. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Li, W., Kang, S., Liu, C.-C., et al. (2013). High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic acids research*, 42(6):e39–e39.

Shaw, D., Chen, H., and Jiang, T. (2018). DeepIsoFun: a deep domain adaptation approach to predict isoform functions. *Bioinformatics*, page bty1017.

Zhang, C., Bengio, S., Hardt, M., et al. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.