

# T-TRP0DS Supporting the Foundation of a Data Intensive Study Center

## **OVERVIEW**

Tufts University T-TRIPODS Phase I Institute supports a culture of interdisciplinary research and learning in data sciences across multiple departments, fostering collaboration among mathematicians, computer scientists, and electrical engineers, as well as scientists and scholars in a wide range of application domains.

#### **INSTITUTE STRUCTURE**

Many TRIPODS support the Tufts Data Intensive Studies Center (DISC): The structure of T-TRIPODS fosters specific and deep connections with applications domain experts in several areas, leading to translational research. Teaching and curriculum development efforts at the undergraduate, graduate, and professional levels complement the research effort.



# ACKNOWLEDGEMENTS

vww.PosterPresentations.cor

This project is part of the NSF's Harnessing the Data Revolution (HDR) Big Idea Activity under award 1934553.

# **FOCUSED RESEARCH TOPICS**

The model is built around overlapping, three-year research topics with an offset timeline. Each year, the oldest research topic sunsets while a new research topic is added. For each focused research topic, T-TRIPODS convenes interdisciplinary teams of computer scientists, electrical engineers, mathematicians, and statisticians to address timely questions and solve important problems at the frontiers of data science. 21 Tufts faculty working in data science foundations spread over the three core departments are T-TRIPODS affiliates.



#### R1/R2 so far: 30+ publications

ALTU	Department of Computer Science, Tufts University, Medford, MA	Computer Science	
MOTIVATION	METHODOLOGY	CONCLUSION	
ixisting PPI networks are often noisy and noomplete. In many datasets, there may be monotate interactions that are undiscovered, but	0.7 0.3 0.7 0.3	Our tests have shown that GLIDER-based network augmentation results in slightly better	
ay have an impact on downstream tasks like	0.4	function prediction on most of the DREAM network.	
IDER, our GLIDE-based link prediction and	Augmentation	FUTURE WORK	
diction method <sup>[3]</sup> to predict the most likely		We would like to further investigate how to	
ten applies a distance based reweighting	0.1 Predictor	classify and remove noisy edges in the PPI network, along with network augmentation.	
network.	Graph	We would like to improve the efficiency of     Our algorithm by reducing the time	
tion prediction results between the new		complexity.	
ph obtained after network augmentation using DER and the original graph.	Function 0.7	• We would like to investigate other reweighting procedures, like effective	
plying it on several human PPI networks from EAM disease module identification challenge <sup>[5]</sup> ,		<ul> <li>We would like to investigate other function</li> </ul>	
th GLIDER, we obtained some improvement in curacy for the classical function prediction	g = GO:0004416 KNN-DSD	prediction approaches that could be applied to the GLIDER-outputs, like cascading <sup>[7]</sup> .	
ethods.	a = GO:0006357, GO:0003345 b = GO:0003345, GO:0004416		
<b>OUR CONTRIBUTIONS</b>	c = GO:0006357, GO:0002435 d = GO:0002435, GO:0004416	ACKNOWLEDGEMENTS	
rincipled way of predicting and adding the	Figure 2: Description of GLIDER based network augmentation, reweighting and Function Prediction.	Thanks to the Tufts T-Tripods institute (NSF HDR grant 1934553). We also thank the Tufts	
ly missing links in a PPI network by using a bined link prediction and reweighting	Augmentation Reweighting F. Prediction	BCB group for helpful discussions.	
nework. nparing the GLIDER output on different	1. Uses GLIDE to find the top 1. Normalize the network weight to lie 1. To predict the function label of a	<u>REFERENCES</u>	
tion prediction methods.	scoring links with default glide in [0,1]. parameters. 2. Let (m, n) be the newly predicted a node m:	<sup>11]</sup> Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather	
DATASETS	2. Local metric used: Common Weighted Normalized 3. Look for the shortest path from <b>m</b> to between <b>m</b> and all other nodes in the using GLIDE.	Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. <i>Nature</i> genetics 25,1(2000), 25.	
rk compendium Network type #Genes #Edges Degree distribution	3. Global metric used: cDSD <sup>[2][4]</sup> n in the original graph. Let the 4. A certain percentage (p) of these     n in the original graph. Let the shortest path be: m->r1->>rN->n     b. Select the 10 nearest node based on the cDSD distance	<sup>[2]</sup> Mengfei Cao, Christopher M Pietras, Xian Feng, Kathryn J Doroschak, Thomas Schaffner, Jisoo Park, Hao Zhang, Lenore J Cowen, and Benjamin J Hescott. 2014. New directions for diffusion-based network prediction of protein function:	
9 2 Protein Interaction 12,420 397,309 9 3 3 Signaling 5,254 21,826 4 Co-expression 12,588 1,000,000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	predicted links augmented to 4. The the new distance d(m, n) is: c. Use them to perform majority original PPI graph.	incorporating pathways with confidence. Bioinformatics 30, 12 (2014), 1219-1227. <sup>Dip</sup> Devkota, K., Murphy, J. M., & Cowen, L. J. (2020). GLIDE: combining local	
6 Homology 10,405 4,223,606	$\frac{1}{d(m,n)} = \frac{1}{d(m,r_1)} + \frac{1}{d(n,r_N)} + \sum_{k=2}^{N-1} \frac{1}{d(r_{k-1},r_k)}$ the label of m. 2. We call this method KNN-DSD	methods and diffusion state embeddings to predict missing interactions in biological networks. Bioinformatics, 36/Supplement, 1), 1464-1473. 14 <sup>(Convent</sup> , L. Dewichat, K. Mu, X. Murthu, J. M., & Wu, K. (2020). Diffusion	
Anonymize networks: IL17 JAK 2 3	<u>RESULTS</u>	State Distances: Multitemporal Analysis, Fast Algorithms, and Applications to Biological Networks. arXiv preprint arXiv:2003.03616.	
Figure 1: Description of the DREAM networks es: <i>H. Sapiens</i> (human)	network $p = 0$ $p = 0.1$ $p = 0.2$ $p = 0.4$ $p = 0.6$ $p = 0.8$ $p = 1$	<ul> <li>Biologiani P. March, The Collegation of the Month Soft Networks (Network) (Network)</li></ul>	
etwork: DREAM1-4 (human). These are the mark networks for DREAM disease module	DREAM1         0.3091         0.3053         0.3004         0.3120         0.3119         0.2974         0.3100           DREAM2         0.2122         0.2155         0.2151         0.2152         0.2152         0.2152	2491-2306, Polizianzi et al. 2010 and a second and complexitions, 452(10), Plazarsfeld, J., Rodriguez, J., Erden, M., Liu, Y., & Cowen, L. J. (2019, Santamber, Mariotry Units Crassifiers, X-basis Supported Texamends for	
ication challenge. <sup>[5]</sup> onal Categories:	DREAM3 0.2439 0.2439 0.2422 0.2403 0.2145 0.2145 0.2418	Improving Protein Function Prediction. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (an. 51-64)	
Gene Ontology <sup>[1]</sup> : only labels from the molecular function and biological process hierarchies are	DREAM4 0.0833 0.0847 0.0845 0.0824 0.0831 0.0819 0.0835		
used; anything labeled by less than 50 or more	Table 1: Function Prediction Scores from 5 fold crossvalidation, using KNN-DSD on different DREAM networks.		

## **Ethics Case Studies Project**

T-TRIPODS ethics subcommittee is working on a project to design ethics case studies for use in undergraduate CS and data science classes. Short case studies are paired with expert opinions. Initial set of 8-10 case studies will be published in May 2023. More info/how to contribute linked off the "Ethics" tab of the T-Tripods website.

# https://tripods.tufts.edu

R3: Data Guarantees: data quality, fairness, transparency privacy, and trust (2021 - 2023)

# R3 Programming in 2022:

Colloquium series: Making eal-world data science responsible data science;

2-day virtual workshop on **Jnderstanding Misinformation and** Disininformation

New seminar series and working group in 2022-2023 focused on FAIR data and metadata: best practices.



Lenore Cowen T-TRIPODS Director (CS & MATH)



T-Tripods founded DIAMONDS with the goal of Broadening participation in data science.

**DIAMONDS** stands for Directed, Intensive and **Measured Opportunities** In Data Science and is aimed at broadening participation in data science, especially for groups historically under-represented in STEM. This is a program of The Center for STEM Diversity at Tufts University.





#### **LEADERSHIP TEAM**

















Abani Patra Huang DISC (CS)Director (CS & MATH)

# **TRIPODS GRADUATE FELLOWS**



The centerpiece of our Ph.D. advising model is the Advising Trio: T-TRIPODS Graduate Fellows from CS/ECE/MATH have a primary research advisor in their own department. As T-TRIPODS Fellows, they are also assigned co-advisors in the other two departments.

# **DATA SCIENCE FOR ALL**

**Undergraduate Summer Research Opportunities in Data Science** 



**DIAMONDS** is now a stand-alone REU site in Data Science (NSF2149871: PI Ellise LaMotte) sites.tufts.edu/diamondsreu