these immense advantages, accurately understanding the relationship between chemical structure and molecular activity has proven to be a challenging problem in the general sense. Finlayson et al. employ an approach to train a set of neural networks to learn how to associate the structure of a given small molecule with its effect on changes in gene expression. This method attempts to jointly optimize representations of chemical structure and the transcriptional changes resulting from exposure to these chemicals. Despite observing mixed performance when attempting to generalize to new tissues, this method shows great potential to make progress on a longstanding, challenging problem and may lead to the ability to more effectively perform *in silico* prioritization of molecules to elicit specific transcriptional responses.

Digital pathology has seen an immense amount of activity where deep learning and convolutional neural networks have been applied to analyze pathology images. One unique challenge in digital pathology has been that whole slide images are too large for many of these neural network approaches to process. Levy et al. propose methods that use a combination of topological domain analysis and graph neural networks to reduce the need to break whole slide images into smaller patches of images that are computationally tractable; this latter approach is lossy yet common. Importantly, their topological analysis allows Levy et al. to quantify a graph neural network's quality of fit and help determine regions of interest. The combination of topological domain analysis and graph neural networks showed significant improvement over traditional convolutional neural networks applied to the task.

3 Data Integration with Applications to Cancer

One of the most genetically, functionally, and medically heterogeneous diseases afflicting humans in modern times is cancer. This disease, typified by one's own cells growing and dividing uncontrollably while evading the immune system to form tumors, is still challenging to detect, diagnose, and treat due to the various molecular mechanisms involved and diverse medical presentations. The papers we highlight here have employed biomedical data integration specifically to address cancer-specific challenges. In general, multiple distinct data modalities can be integrated via novel techniques to more effectively use poorly labeled or unlabeled data. Data sources that can be examined together include: molecular 'omics data (genomics, transcriptomics, proteomics, metabolomics etc.), medical imaging data, free text, and longitudinal outcomes data. The inclusion and integration of new and novel data sources can help examine and understand various biological processes, many of which have been implicated in cancer progression.

Scott et al. highlight the lack of heterogeneity in discovery populations and subsequent inability to accurately translate biomarkers for general use in the clinic. To address this challenge, the authors attempted to leverage heterogeneity, both biological and technical, across independent cohorts to find biomarkers more likely to generalize. By utilizing a primary dataset that includes 23 different cancers and combining it with 57 independent microarray datasets, they found the gene KRT8 to be significantly hypomethylated in the 57 independent datasets and overexpressed in 22 out of 23 cancers. Scott et al. then performed additional validation steps, including single-cell analyses, immunohistochemistry of tumor biopsies, and finally, detecting levels of KRT8 in the serum of patients with pancreatic cancer vs. healthy controls. While they have not yet shown its ability as a predictive marker for cases who have not yet been identified by other means, these validation steps show great potential for translational applicability.

Durmaz et al.'s approach to subgraph analyses allowed for the examination of single cancers using The Cancer Genome Atlas (TCGA) pan-cancer data. Their approach enabled the data-driven identification of patient clusters across different TCGA disease codes based on related dysregulation patterns and led to elucidating significant differences between survival for various disease codes, including lower grade gliomas and uterine cancer. The survival differences illustrate the potential of applying pathway-based functional networks to stratify cancer as compared to traditional gene-centric models. Additionally,

considering cancer-relevant dysregulation at the pathway level versus the gene level provides additional insight into disease etiology.

Similarly, Levy et al.'s combination of topological data analysis with a graph neural network allows for identification of regions of interest in whole slide pathology images. The authors were subsequently able to measure the degree of overlap between regions of interest in a tumor and in the adjacent normal tissue and then associate these regions of interest with clinical outcomes by means of cancer staging. Their approach allows for human-readable highlighted regions of interest as well as a prediction of cancer stage where they found they were able to predict advanced colon cancer staging and positive lymph nodes at >0.9 AUC.

4 Discussion

Pattern recognition has already had and will continue to have a large role in understanding both biology and medicine. Technological developments are leading to larger and more varied biomedical datasets. Both novel and repurposed methodologies must be developed and applied to these data in order to derive insights that can drive more precise patient care, yield novel therapeutics, guide earlier interventions and in general provide greater understanding of biomedicine.

The work highlighted here targets these developments. Finlayson et al. aim to make the identification of therapeutically-relevant small molecules possible at faster speeds, and Durmaz et al. aspire to characterize the molecular mechanisms of cancer development and progress through computation that considers graph structure in protein interaction networks. Levy et al. propose novel methods to precisely extract regions of interest from histopathology images and to identify prognostic predictors to enable more precise patient care. Finally, Scott et al. identified a biomarker that may help lead to earlier and more accurate diagnoses of cancer. Each of these works is guided by the common theme of using pattern recognition to go beyond computational performance and to drive biomedical discovery and understanding.

References

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., … Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface / the Royal Society*, *15*(141). https://doi.org/10.1098/rsif.2017.0387

Durmaz, A., Henderson, T. A. D., Bebek, G. (2020). "Frequent Subgraph Mining of Functional Interaction Patterns Across Multiple Cancers." *Pac Symp Biocomput.* To appear.

Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. *Science*, *328*(5981), 1029–1031.

Finlayson, S. G., McDermott, M. B. A., Pickering, A. V., Lipnick, S. L., Kohane, I. S. (2020). "Cross-modal representation alignment of molecular structure and perturbation-induced transcriptional profiles." *Pac Symp Biocomput.* To appear.

Lakhani, C. M., Tierney, B. T., Manrai, A. K., Yang, J., Visscher, P. M., & Patel, C. J. (2019). Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nature Genetics*, *51*(2), 327–334.

Lee, H., Huang, A. Y., Wang, L.-K., Yoon, A. J., Renteria, G., Eskin, A., Signer, R. H., Dorrani, N., Nieves-Rodriguez, S., Wan, J., Douine, E. D., Woods, J. D., Dell'Angelica, E. C., Fogel, B. L., Martin, M. G., Butte, M. J., Parker, N. H., Wang, R. T., Shieh, P. B., … Nelson, S. F. (2020). Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *22*(3), 490–499.

Levy, J., Haudenschild, C., Barwick, C., Christensen, B., Vaickus, L. (2020). "Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks." *Pac Symp Biocomput.* To appear.

Le Goallec, A., Tierney, B. T., Luber, J. M., Cofer, E. M., Kostic, A. D., & Patel, C. J. (2020). A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. *PLoS Computational Biology*, *16*(5), e1007895.

Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, *23*(5), 899–908.

Pasco, P. M. D., Ison, C. V., Muňoz, E. L., Magpusao, N. S., Cheng, A. E., Tan, K. T., Lo, R. W., Teleg, R. A., Dantes, M. B., Borres, R., Maranon, E., Demaisip, C., Reyes, M. V. T., & Lee, L. V. (2011). Understanding XDP through imaging, pathology, and genetics. *The International Journal of Neuroscience*, *121 Suppl 1*, 12–17.

Scott, M. K. D., Ozawa, M. G., Chu, P., Limaye, M., Nair, V. S., Schaffert, S., Koong, A. C., West, R., Khatri, P. (2020). "A multi-scale integrated analysis identifies KRT8 as a pan-cancer early biomarker." *Pac Symp Biocomput.* To appear.

Turro, E., Astle, W. J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H. L., Sanchis-Juan, A., Frontini, M., Thys, C., Stephens, J., Mapeta, R., Burren, O. S., Downes, K., Haimel, M., Tuna, S., Deevi, S. V. V., Aitman, T. J., Bennett, D. L., … Ouwehand, W. H. (2020). Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, *583*(7814), 96–102.

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., & Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, *7*, 12474.