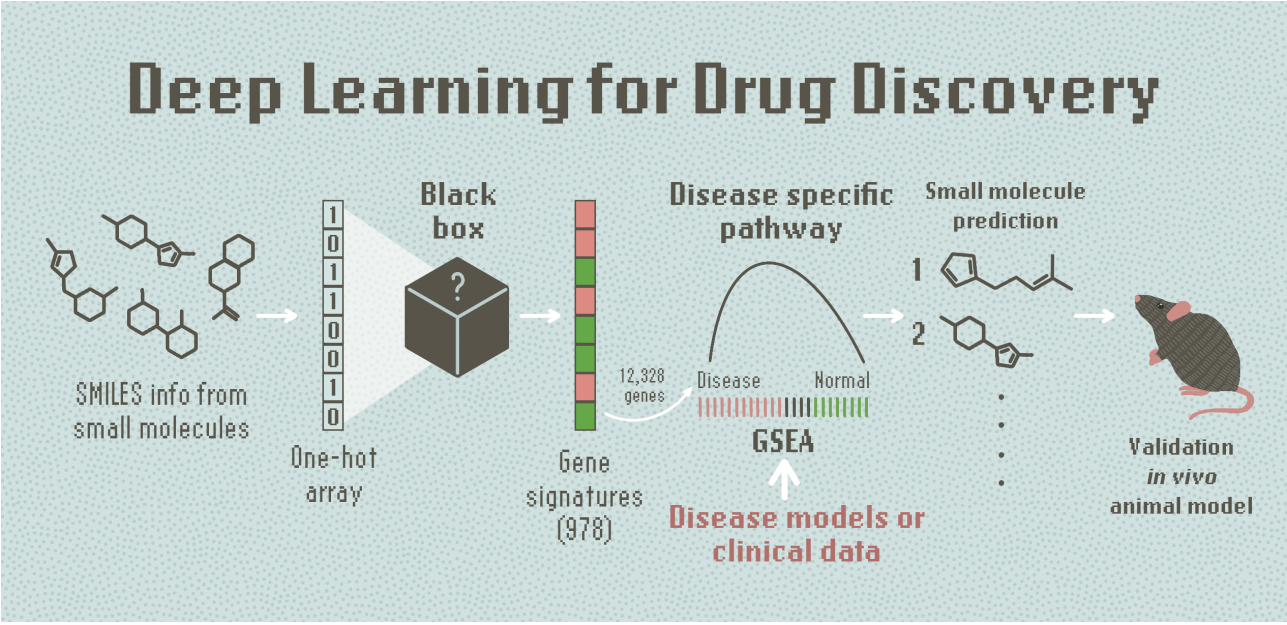
Molecules and Cells

Journal Club

Deep Learning Approach Based on Transcriptome Profile for Data Driven Drug Discovery

Eun-Ji Kwon and Hyuk-Jin Cha*

College of Pharmacy, Seoul National University, Seoul 08826, Korea
*Correspondence: hjcha93@snu.ac.kr
https://doi.org/10.14348/molcells.2023.2167
www.molcells.org



SMILES (simplified molecular-input line-entry system) information of small molecules parsed by one-hot array is passed to a convolutional neural network called black box. Outputs data representing a gene signature is then matched to the genetic signature of a disease to predict the appropriate small molecule. Efficacy of the predicted small molecules is examined by *in vivo* animal models. GSEA, gene set enrichment analysis.

Received November 2, 2022; revised November 19, 2022; accepted November 22, 2022; published online January 20, 2023

eISSN: 0219-1032

Once a target protein has been associated with the onset and progression of a certain disease, the initial 'hit' compounds are largely discovered by either high-throughput screening (HTS) or chemical design based on the target protein structure (Schenone et al., 2013). This process, which is commonly referred to as target-based drug discovery (TDD), has been quite successful. However, the TDD approach requires 'target identification,' which demands tremendous efforts during the pre-discovery process. However, once a cellular platform to explore pathogenic phenotypes is established (e.g., disease-in-a-dish from patient-derived stem cells), HTS (or content screening) can be performed to identify candidate compounds to alleviate or reverse the pathogenic phenotypes (Schenone et al., 2013). The recent advances in phenotype-based drug discovery and gene editing technology have enabled the establishment of isogenic pairs of patient-derived stem cells (Park et al., 2022). Disease-induced transcriptome profiles are not only the consequences of disease but can also provide clues regarding its potential causes (i.e., the mechanism of disease onset). Therefore, in-depth analyses of transcriptome profiles from disease models often provide crucial insights into the mechanisms of pathogenesis and enable the discovery of valuable target proteins (Casamassimi et al., 2017). Furthermore, a growing number of studies have incorporated disease-associated transcriptome signatures into the drug discovery process (Kwon et al., 2019). For example, the connectivity Map (CMap) approach can be used to inversely match drug- or compound-induced transcriptome profiles with disease signatures (Kwon et al., 2020).

Zhu et al. (2021) developed a deep learning-based efficacy prediction system (DLEPS) model to facilitate drug discovery. Unlike CMAP, where chemically inducible changes in transcriptional profiles (CTPs) are experimentally assessed, the DLEPS model is a deep neural network that is trained using input from simplified molecular-input line-entry system (SMILES) data (i.e., a line-based specification for the description of chemical structures) and CTPs from the L1000 project (Subramanian et al., 2017), an extended transcriptome profile of chemical or genetic perturbations. First, each small molecule, which is first represented by SMILES, is vectorized into a single array and projected into a two-dimensional latent space. This latent space is a virtual space that is further decoded into 978 landmark gene signatures. These landmark genes, which are converted into 12,328 genes via linear transformation, are used as input to perform gene set enrichment analysis based on the ranks of their expression levels. Through DLEPS, the authors virtually produce a correlation network between the chemical structure of small molecules and putative gene expression signatures. Therefore, this DLEPS algorithm allows for the establishment of links between gene signatures, pathogenic responses, and specific small molecules.

To validate this system, the authors first attempted to predict putative anti-obesity molecules based on 150 unbiased up- and down-regulated genes identified by comparing brown adipose tissue (BAT) to white adipose tissue (WAT) using a library containing a total of 3680 small molecules (D3680 library: including U.S. Food and Drug Administration approved drug or natural compounds). The four predicted molecules, including isoginkgetin, chelidonine, loureirin B, and chikusetsusaponin IV, were tested in a high-fat diet mouse model to confirm their weight reduction effect. Interestingly, such weight loss is achieved through the activation of genes associated with BAT and adaptive thermogenesis without significant alterations in food intake or daily exercise activity, suggesting the potential of these compounds as novel anti-obesity drug candidates. Therefore, the authors concluded that treating the mice with the natural compounds predicted by the DLEPS model promotes the browning of WAT. Similarly, this model was applied to identify candidate drugs for the treatment of hyperuricemia (HUA), a chronic metabolic disorder characterized by the occurrence of inflammation and renal fibrosis. Based on the corresponding gene signature, the authors calculated an HUA score and inflammation/fibrosis score as the inputs for the DLEPS model and identified four candidate compounds out of the D3680 library. Experimental validation with an HUA mouse model demonstrated that perillene, a natural compound derived from '*Perilla frutescens*,' could effectively lower blood uric acid levels and this therapeutic effect was similar to that of commercial anti-HUA drugs such as allopurinol, benzbromarone, topiroxostat, and febuxostat. Other HUA symptoms such as blood urea nitrogen, serum creatinine, alanine aminotransferase (ALT), aspartate aminotransferase (AST), kidney index (ratio of kidney and body weight), and the level of fibrosis of renal tubules were also ameliorated by perillene treatment. The authors also reported that perillene inhibits xanthine oxidase by direct interaction, which was likely the mode of action of this compound. A similar approach was adopted to identify drug candidates for nonalcoholic steatohepatitis (NASH) from a clinical trial database including 11,293 compounds (D11294 library). The two predicted compounds trametinib and GI02002 effectively reduced the level of ALT, AST, and triglyceride levels in a NASH mouse model, which was established by feeding the mice with a methionine-and choline-deficient diet. Trametinib is an inhibitor of MEK1 and impedes ERK1/2 activation, which was previously demonstrated to inhibit glucose and lipid metabolism. The authors also examined the effectiveness of other ERK inhibitors such as ravoxertinib and FR180204.

The continuous development of large-scale datasets of drug (or small molecule)-induced transcriptomic patterns (i.e., drug-omics datasets) will thus enable the creation of learning models to link small molecules to gene signatures associated with specific diseases. In turn, this would greatly promote the development of data-driven drug discovery. The refinement of deep learning models also poses several important challenges (Zhao and So, 2019) and therefore additional efforts are needed to experimentally validate these models. The accuracy of a deep learning model greatly depends on the size of the training dataset. However, current datasets from compounds and transcriptomes of disease models are relatively insufficient. Therefore, alternative approaches such as 'transfer learning' could enable the development of deep learning-based drug discovery (Chiu et al., 2021) by minimizing overfitting and uncertainty.

The study by Zhu et al. (2021) successfully validated the efficacy of drugs and natural compounds predicted by

their deep learning model based on disease-specific gene signatures. Considering the complexity of the molecular mechanisms that underly disease onset and progression, disease-specific gene signatures must be precisely characterized and defined to avoid overlap with other similar diseases for further clinical applications. Various deep learning systems are also being extensively developed to interpret biological networks such as complex protein-protein interactions and signaling cascades (Muzio et al., 2021). Therefore, highly sophisticated deep learning systems for drug prediction could be developed by training the algorithms with disease-specific gene signatures and biological network datasets.

ACKNOWLEDGMENTS

This work was supported by the Seoul National University Research Grant in 2022.

AUTHOR CONTRIBUTIONS

H.J.C. participated in conception and reviewing manuscript. E.J.K. participated in manuscript writing.

CONFLICT OF INTEREST

The authors have no potential conflicts of interest to disclose.

ORCID

Eun-Ji Kwon https://orcid.org/0000-0001-7057-3953
Hyuk-Jin Cha https://orcid.org/0000-0001-9277-2662

REFERENCES

Casamassimi, A., Federico, A., Rienzo, M., Esposito, S., and Ciccodicola, A. (2017). Transcriptome profiling in human diseases: new advances and perspectives. Int. J. Mol. Sci. *18*, 1652.

Chiu, Y.C., Zheng, S., Wang, L.J., Iskra, B.S., Rao, M.K., Houghton, P.J., Huang, Y., and Chen, Y. (2021). Predicting and characterizing a cancer dependency map of tumors with deep learning. Sci. Adv. *7*, eabh1275.

Kwon, O.S., Kim, W., Cha, H.J., and Lee, H. (2019). In silico drug repositioning: from large-scale transcriptome data to therapeutics. Arch. Pharm. Res. *42*, 879-889.

Kwon, O.S., Lee, H., Kong, H.J., Kwon, E.J., Park, J.E., Lee, W., Kang, S., Kim, M., Kim, W., and Cha, H.J. (2020). Connectivity map-based drug repositioning of bortezomib to reverse the metastatic effect of GALNT14 in lung cancer. Oncogene *39*, 4567-4580.

Muzio, G., O'Bray, L., and Borgwardt, K. (2021). Biological network analysis with deep learning. Brief. Bioinform. *22*, 1515-1530.

Park, J.C., Kim, J., Jang, H.K., Lee, S.Y., Kim, K.T., Kwon, E.J., Park, S., Lee, H.S., Choi, H., Park, S.Y., et al. (2022). Multiple isogenic GNE-myopathy modeling with mutation specific phenotypes from human pluripotent stem cells by base editors. Biomaterials *282*, 121419.

Schenone, M., Dancik, V., Wagner, B.K., and Clemons, P.A. (2013). Target identification and mechanism of action in chemical biology and drug discovery. Nat. Chem. Biol. *9*, 232-240.

Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. Cell *171*, 1437-1452.e17.

Zhao, K. and So, H.C. (2019). Using drug expression profiles and machine learning approach for drug repurposing. Methods Mol. Biol. *1903*, 219-237.

Zhu, J., Wang, J., Wang, X., Gao, M., Guo, B., Gao, M., Liu, J., Yu, Y., Wang, L., Kong, W., et al. (2021). Prediction of drug efficacy from transcriptional profiles with deep learning. Nat. Biotechnol. *39*, 1444-1452.