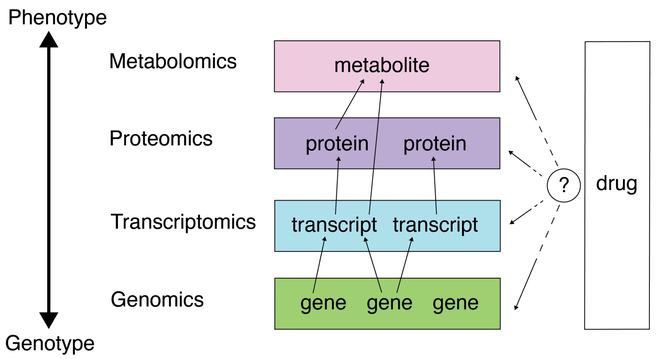
Learning molecular cross-talk: deep learning for multi-omic translation

Jeannie She[1], Nikolaos Meimetis[2], Douglas Lauffenburger[2]
[1]MIT Department of Electrical Engineering and Computer Science, [2]MIT Department of Biological Engineering

Multi-omic data captures a comprehensive view of biological systems and disease states

- Complex biological systems are better understood with multiple **omics assays** which characterize genotype and phenotype at varying granularity.
- Experimentally perturbing biological systems with drugs results in changes that are **comprehensively captured** through multi-omic analysis.

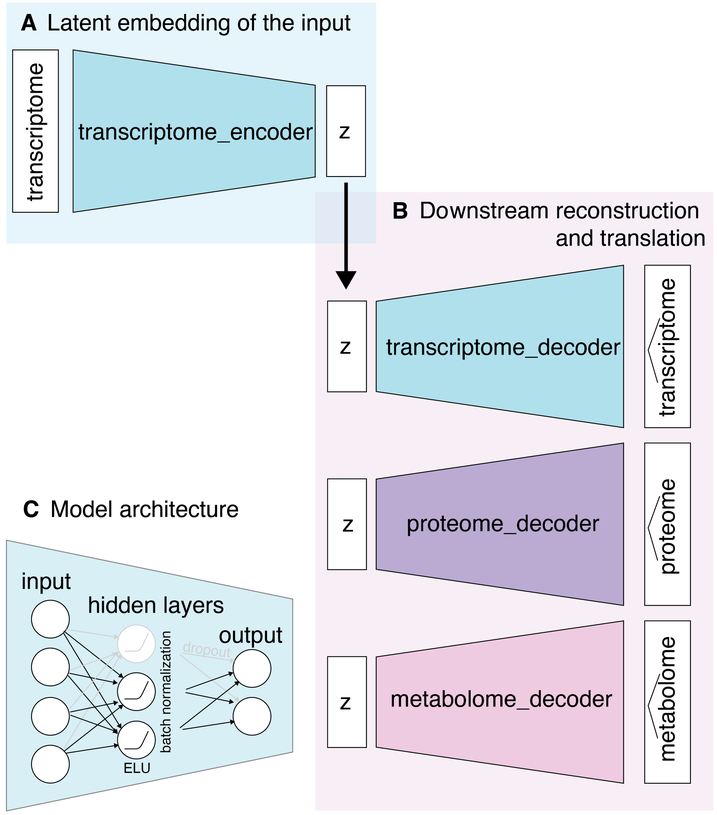


- **Problem**: Performing assays for each drug perturbation is experimentally expensive and combinatorically infeasible. Most often, transcriptomic data is the most abundant.
- **Goal:** To bridge experimental gaps by utilizing deep learning to **translate the transcriptome to the proteome and metabolome**, learning cell-line specific cross-talk interactions.

Autoencoder learns a shared latent representation that captures key biological features across omic layers

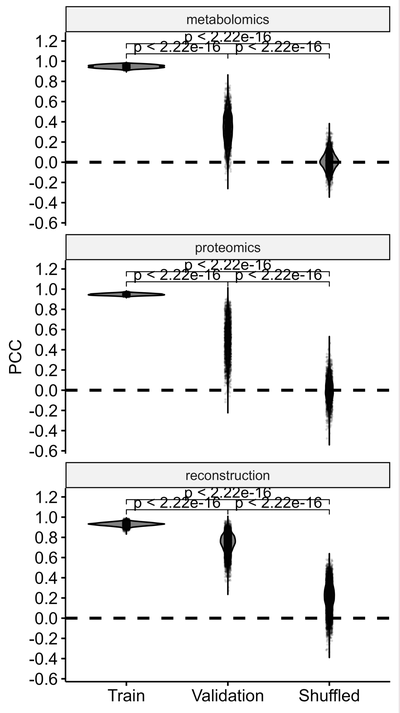
A multi-decoder autoencoder, largely inspired by PRNet[1], is trained on multi-omic data from 881 cell lines at the Cancer Cell Line Encyclopedia[2,3].

- Encoder maps high-dimensional transcriptomes to a **shared latent embedding (z)** capturing underlying biological insight.
- Three parallel decoders generate:
  - Reconstructed transcriptome
  - Translated proteome
  - Translated metabolome



Model performance

- Loss function combines **mean squared error across all three outputs,** weighted equally.
- The model performs significantly better on the validation dataset than on a shuffled dataset, indicating that the **learned representation captures true, meaningful relationships** between the omic layers.
- The reconstruction of the transcriptome and translation into the proteome performs exceptionally, while translation into the metabolome performs slightly lower on average.



Significance and next steps

By looking closer at the autoencoder's latent space, we can uncover information about which features are most representative for reconstruction and translation into other omic layers.

This model will be utilized to generate simulated data for the purpose of *in silico* drug discovery, where we will train a deep learning model to learn and predict the impact of drug perturbations on cell lines.

Thank you to Nikolaos Meimetis and the Lauffenburger Lab for the support on this project.

[1]Qi, X., Zhao, L., Tian, C. *et al.* Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nat Commun* **15**, 9256 (2024). https://doi.org/10.1038/s41467-024-53457-1

[2]Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, ... Todd R. Golub, Levi A. Garraway & William R. Sellers. 2019. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 569, 503–508 (2019). https://doi.org/10.1038/s41586-019-1186-3

[3]Haoxin Li, Shaoyang Ning, Mahmoud Ghandi, Gregory V. Kryukov, Shuba Gopal, ... Levi A. Garraway & William R. Sellers. The landscape of cancer cell line metabolism. *Nature Medicine* 25, 850-860 (2019). https://doi.org/10.1038/s41591-019-0404-8