**Gerstein lab experience with Interpretable Models Integrating Genomics and Epigenomics Features which can Serve as Distinct Biomarkers of Key Molecular Mechanisms Demonstrating Changes with Aging**

We developed a deep generative single-cell Deep Mixed-Modal Clustering and Imputation method for uniformly fusing heterogeneous single-cell modalities from multiple sources to dissect a cell atlas across multiple brain regions and identify epigenome and transcriptome perturbations in a cell-type-specific manner. Distinct from existing methods, our noise-aware model evaluates each modality's sensitivity in separating various cell types and flexibly weight them for robust clustering and imputation.

To integrate epigenomics into our aging analysis, we have analyzed ATAC-seq data from PFC brain bulk tissues to extract features, for males and females separately, demonstrating more open chromatin with large variability with aging. For this, we measured the Gini score, representing inequality, across individuals for each peak split by age group. Changes in chromatin variability were demonstrated with aging using Gino score. More recently, our group used their developed epigenetic biomarker “DNAm PhenoAge” to predict the ages of different tissues from four individuals in the EN-TEx database1. We showed that (A) different tissues of the same individuals have quite different predicted ages, suggesting that different tissues age at different speeds, yet the predicted age is still highly correlated with the chronological age, and (B) the model is accurate for capturing changes in tissues with chronological age.

To enhance the understanding of deep neural networks (DNNs) and their learning process in relation to personalized medicine2, we developed an DNN architecture integrating importance score calculations for genomic and epigenomics features to accurately predict regulatory elements3, 4**.** Furthermore, we have successfully applied a DNN-encoder-decoder architecture on transcriptomic datasets to extract predictive features that serve as distinct biomarkers of key molecular mechanisms for asthma5. Moreover, we developed a rank-statistic-based interpretation scheme that allows us to functionally annotate hidden nodes and prioritize them relative to disorders6. Our model improved disease prediction by 6-fold compared to additive polygenic risk scores for schizophrenia, highlighted key genes for schizophrenia and other disorders, and allowed imputation of missing transcriptome information from genotype alone7.

We have much experience in scaling up our machine learning approaches8 to a large scale. Specifically, we were the lead analysis group for various NIH consortia, including ENCODE9,10 PsychENCODE11, modENCODE12, Pan-Cancer Analysis of Whole Genomes (International Cancer Genome Consortium/The Cancer Genome Atlas)13, and SCORCH. Our cell-type-specific gene expression signatures have been widely used by PsychENCODE to prioritize risk factors in multiple disorders.

1. Rozowsky J, Drenkow J, Yang YT, Gursoy G, Galeev T, Borsari B, Epstein CB, Xiong K, Xu J, Gao J, Yu K, Berthel A, Chen Z, Navarro F, Liu J, Sun MS, Wright J, Chang J, Cameron CJ, Shoresh N, Gaskell E, Adrian J, Aganezov S, Balderrama-Gutierrez G, Banskota S, Corona GB, Chee S, Chhetri SB, Cortez Martins GC, Danyko C, Davis CA, Farid D, Farrell NP, Gabdank I, Gofin Y, Gorkin DU, Gu M, Hecht V, Hitz BC, Issner R, Kirsche M, Kong X, Lam BR, Li S, Li B, Li T, Li X, Lin KZ, Luo R, Mackiewicz M, Moore JE, Mudge J, Nelson N, Nusbaum C, Popov I, Pratt HE, Qiu Y, Ramakrishnan S, Raymond J, Salichos L, Scavelli A, Schreiber JM, Sedlazeck FJ, See LH, Sherman RM, Shi X, Shi M, Sloan CA, Strattan JS, Tan Z, Tanaka FY, Vlasova A, Wang J, Werner J, Williams B, Xu M, Yan C, Yu L, Zaleski C, Zhang J, Cherry JM, Mendenhall EM, Noble WS, Weng Z, Levine ME, Dobin A, Wold B, Mortazavi A, Ren B, Gillis J, Myers RM, Snyder MP, Choudhary J, Milosavljevic A, Schatz MC, Guigó R, Bernstein BE, Gingeras TR, Gerstein M. Multi-tissue integrative analysis of personal epigenomes [Internet]. Genomics; 2021 Apr [cited 2022 Sep 21]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.04.26.441442

2. Hussein Mohsen MG. Weight-based Neural Network Interpretability using Activation Tuning and Personalized Products. In.

3. Chen Z, Zhang J, Liu J, Dai Y, Lee D, Min MR, Xu M, Gerstein M. DECODE: a Deep-learning framework for Condensing enhancers and refining boundaries with large-scale functional assays. Bioinformatics. 2021 Jul 12;37(Suppl\_1):i280–8.

4. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). 2017. p. 618–26.

5. Lou S, Li T, Spakowicz D, Yan X, Chupp GL, Gerstein M. Latent-space embedding of expression data identifies gene signatures from sputum samples of asthmatic patients. BMC Bioinformatics. 2020 Oct 15;21(1):457.

6. Warrell J, Mohsen H, Gerstein M. Rank Projection Trees for Multilevel Neural Network Interpretation [Internet]. arXiv; 2018 [cited 2022 Sep 19]. Available from: <http://arxiv.org/abs/1812.00172>.

7. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FC, Clarke D, Gu M, Emani P, Yang YT. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362(6420): eaat8464.

8. Kumar S, Harmanci A, Vytheeswaran J, Gerstein MB. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. Genome Biol. 2020 Nov 9;21(1):274.

9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

10. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91–100.

11. Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S, Geschwind DH, Mill J, Nairn AC, Abyzov A, Pochareddy S, Prabhakar S, Weissman S, Sullivan PF, State MW, Weng Z, Peters MA, White KP, Gerstein MB, Amiri A, Armoskus C, Ashley-Koch AE, Bae T, Beckel-Mitchener A, Berman BP, Coetzee GA, Coppola G, Francoeur N, Fromer M, Gao R, Grennan K, Herstein J, Kavanagh DH, Ivanov NA, Jiang Y, Kitchen RR, Kozlenkov A, Kundakovic M, Li M, Li Z, Liu S, Mangravite LM, Mattei E, Markenscoff-Papadimitriou E, Navarro FCP, North N, Omberg L, Panchision D, Parikshak N, Poschmann J, Price AJ, Purcaro M, Reddy TE, Roussos P, Schreiner S, Scuderi S, Sebra R, Shibata M, Shieh AW, Skarica M, Sun W, Swarup V, Thomas A, Tsuji J, van Bakel H, Wang D, Wang Y, Wang K, Werling DM, Willsey AJ, Witt H, Won H, Wong CCY, Wray GA, Wu EY, Xu X, Yao L, Senthil G, Lehner T, Sklar P, Sestan N. The PsychENCODE project. Nature Neuroscience. 2015 Dec 1;18(12):1707–12.

12. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ. Comparative analysis of the transcriptome across distant species. Nature. 2014;512(7515):445–8.

13. Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, Harmanci A, Martinez-Fundichely A, Chan CW, Nielsen MM. Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. Cell. 2020;180(5):915–27.