# Computational Analysis of Physiological Systems at Multiple Time and Length Scales

by

Hongyang Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2019

Doctoral Committee:

Associate Professor Yuanfang Guan, Chair
Professor Daniel Burns, Jr
Professor Heather Carlson
Professor Kayvan Najarian
Professor Gilbert S. Omenn
Professor David Sept

Hongyang Li

hyangl@umich.edu

ORCID: 0000-0002-7902-4157

*To my love Shuai*

## ACKNOWLEDGMENTS

I would like to express my deepest gratitude and appreciation to Dr. Yuanfang Guan for her essential role and guidance in my PhD study. Without her, I might not be able to finish my PhD and the future track of my life would be completely different. Everytime I had difficulties in my research projects, Yuanfang was always there standing by me and promptly replied to my questions. Since she loves research and programmings by herself, her suggestions were practically useful and helped me overcome obstacles in the shortest path. Yuanfang is incomparably fast in learning and problem solving. I learn a lot from her and will continue benefit from these characteristics. Discussing projects with Yuanfang is a joyful journey and often enlightens me, because she is extremely sharp, full of vim and vigor, and always has marvelous ideas. I am so grateful to her for her trust in me and offering me the opportunities to participate data challenges with her, from which I "evolved" into a much better me. She is a frank person of integrity, which is rare and invaluable. As a group leader, Yuanfang is a true philosopher-queen who is exceptionally intelligent, considerate, reliable, and responsible for her students and projects. For me, it is the best of times - having such a great mentor like her; it is the worst of times - meeting her only the second half of my PhD since 2017. Beyond science and research, the interaction with Yuanfang often inspired me to meditate on life and universe. "Art itself is the relief of a artist; Science itself is the relief of a scientist", said by her, will accompany and encourage me to fearlessly pursue Veritas in the future.

I would like to thank Dr. Daniel Burns, Dr. Heather Carlson, Dr. Kayvan Najarian, Dr. Gilbert Omenn, and Dr. David Sept. They served on my dissertation committee and guided me all the way through my PhD study. They were always willing to support me and help me towards graduation. I would like to thank Dr. Yang Zhang, the rotation in his lab was the initial step for me to transit from Chemical Biology to Bioinformatics, which was more suitable for me. I also would like to thank two previous committee members, Dr. Barry Grant and Dr. John Tesmer, who helped me develop the structural dynamic approach for investigating evolutionarily-related protein families.

I would like to thank all my friends at University of Michigan. Together with them, I had so many enjoyable experiences and irreplaceable memories. I would like to thank my parents for their unconditional love and support for me since I was born. Finally, I would like to thank my wife, Shuai Hu, for her love, encouragement and strong support for me to overcome hard times during my PhD study.

## TABLE OF CONTENTS

## List of Figures

## List of Tables

## List of Abbreviations

MD: Molecular Dynamics

GTPase: Guanosine Triphosphate Phosphohydrolase

GTP: Guanosine Triphosphate

GDP: Guanosine Diphosphate

PCA: Principal Component Analysis

ANN: Artificial Neural Network

GPU: Graphics Processing Unit

RF: Random Forest

GAP: GTPase-activating protein

GEF: Guanine nucleotide Exchange factors

GDI: GDP Dissociation Inhibitor

Gα: heterotrimeric G protein α subunit

Gβγ: heterotrimeric G protein βγ subunits

H-Ras: H isoform of Ras

EF-Tu: Elongation Factor Thermo unstable

RMSF: Root Mean Square Fluctuation

NCI: National Cancer Institute

CPTAC: Clinical Proteomic Tumor Analysis Consortium

TCGA: The Cancer Genome Atlas

DREAM: Dialogue on Reverse Engineering Assessment and Method

NRMSE: Normalized Root Mean Square Error

GO: Gene Ontology

CNV: Copy Number Variation

KEGG: Kyoto Encyclopedia of Genes and Genomes

RERA: respiratory effort-related arousals

AASM: American Academy of Sleep Medicine

CNN: Convolutional Neural Network

EEG: ElectroEncephaloGraphy

EOG: ElectroOculoGraphy

EMG: ElectroMyoGraphy

ECG: ElectroCardioGram

AUROC: Area Under Receiver Operating Characteristic curve

AUPRC: Area Under Precision-Recall Curve

REM: Rapid Eye Movement

ICC: Intra-class Correlation Coefficient

ReLU: Rectified Linear Unit

LSTM: Long Short-Term Memory

GRU: Gated Recurrent Unit

STFT: Short-Time Fourier Transform

## Abstract

The fast advancement of computers in the past decade has revolutionized the way we explore and understand the world. Meanwhile, the development of algorithms and methods enables us to efficiently analyze data, build models, and even perform *in silico* experiments through simulations. This is especially true in the era of big data - how to leverage large-scale multi-source information to model physiological phenomena and interpret observations from an unprecedented computational perspective. In this dissertation, I focus on multiple physiological and biological systems at different scales, ranging from biological molecules at nanosecond time scale to human physiological signals spanning hours.

First, I present a network analysis approach for comparing the structural dynamics of three major GTPase superfamilies based on extensive molecular dynamics simulations. GTPases are essential biological macromolecules that regulate a variety of cellular processes. They share a common core structure supporting nucleotide binding and hydrolysis, yet their biological functions diverge dramatically. Many efforts have been made to compare their sequences and 3D structures, however, the similarity and differences of their physical movements remain unclear. I investigated the structural dynamic characteristics of three typical GTPases, and identified common and family-specific residues mediating the coupling of functional sites. I further performed mutational simulations and demonstrated the dynamic effects of disrupting key couplings.

Second, I describe a first-place algorithm in the 2017 NCI-CPTAC DREAM Proteogenomics Challenge, which unbiasedly evaluated computational methods for predicting the proteomics profiles in breast and ovarian cancer patients. Decoding the determinants controlling protein levels is crucial for understanding the regulatory mechanisms underlying cancers. Predicting the protein abundance from mRNA levels is challenging, due to the large variations across cancer patients and weak correlations between mRNA and protein levels. I investigated several critical determinants of protein abundance, including the rule of multi-omics data, the interdependencies among various genes, and how to harness information from two different cancer tissues to extend our understanding of protein abundance regulation. While for the first two we gave an improved modelling method over previous studies, the last aspect is unexplored in literature for proteomic expression level modelling. In addition, the prediction correlation of our method approaches the theoretical upper limit calculated from experimental replicates. Key functional pathways and gene-gene interaction network modules associated with cancer proteome regulation were further revealed.

Third, I present a first-place algorithm in the 2018 PhysioNet/Computing in Cardiology Challenge. I developed a deep learning approach, DeepSleep, to automatically segment sleep arousals from polysomnographic recordings, including physiological signals from brain activity, heart, breath, and body movement during sleep. DeepSleep enables fast segmentation of sleep records at millisecond resolution. Compared with the theoretical upper limit based on annotation replicates by different sleep experts, our method approximates human performance in detecting sleep arousals. Moreover, the pattern of our predictions differs from human annotations, especially at the low-confident boundary regions. This indicates that *in silico* annotations is a complement to

human annotations and potentially advances the current binary label system and scoring criteria for sleep arousals.

# CHAPTER I

# Introduction

## Molecular dynamics

Molecular dynamics (MD) is a technique for studying the physical movements of atoms and molecules through computer simulations (Frenkel and Smit 2002). Given a system of interest, we first compute the forces on all particles in the system based on molecular mechanics force fields, which can be derived from physicochemical experiments and/or calculations in quantum mechanics. Then we numerically integrate Newton's equations of motion and obtain the new positions of all particles. Repeating the force and position calculations result in the dynamic evolution or trajectory of the system. The first development of MD simulation dates back to early 1950s (Alder and Wainwright 1959). At that time, MD was mainly used to simulate atoms and small molecules in the fields of physics and chemistry . The first MD simulation of a biomolecule was published in 1977 (McCammon et al. 1977). As the improvement of computational powers, nowadays MD has been widely used to study biological macromolecules (Karplus 2002). In 2013, the Nobel chemistry prize was awarded to Martin Karplus, Michael Levitt, and Arieh Warshel, for their significant contribution to the development of multiscale models for complex chemical systems (https://www.nobelprize.org/prizes/chemistry/2013/summary/). Meanwhile, massively parallel supercomputers specially designed for MD simulations such as Anton, can simulate processes on long above microsecond time scales. These advancements open a new avenue for us to investigate and understand details of particle motions in a variety of biophysical problems,

including protein-ligand interactions, protein folding, conformational dynamics of protein and protein complexes (Guo et al. 2016; Chung et al. 2015; Rosenbaum et al. 2011).

**Molecular switches**

Molecular switches are molecules that can reversibly transit between two or more states (Vale 1996). Guanosine triphosphate phosphohydrolases (GTPases) are ubiquitous molecular switches that regulate a multitude of essential cellular processes ranging from cell division and differentiation to protein synthesis and translocation (Scheffzek and Ahmadian 2005; Vetter and Wittinghofer 2001). They operate through hydrolyzing guanosine triphosphate (GTP) into guanosine diphosphate (GDP) with associated conformational changes that modulate affinity for specific binding partners. There are three major GTPase superfamilies: Ras-like GTPases (Wennerberg 2005), heterotrimeric G proteins (Milligan and Kostenis 2009) and protein-synthesizing GTPases (Maracci and Rodnina 2016). As the primary coupling molecule to membrane receptors, Gα together with its partner βγ subunits (Gβγ) mediate the very early stage signal transduction initiated by extracellular stimuli. In contrast, small GTPase does not interact with receptors directly and regulates more downstream events in the cascade. The protein-synthesizing GTPases participate in initiation, elongation and termination of mRNA translation. Although they contain a similar nucleotide-binding architecture, the detailed mechanisms by which these structurally and functionally diverse superfamilies operate remain unclear (Vetter and Wittinghofer 2001). MD provides us the opportunity to study their nucleotide-associated dynamics *in silico*.

**Machine learning**

Machine learning is a category of algorithm that computer systems learn patterns from data and perform specific tasks without being explicitly programmed (Ziegel 2003; Bishop 2016). Two major types of machine learning algorithms are unsupervised learning and supervised learning. In a unsupervised learning task, the data only contain the inputs. In a supervised learning task, the data contain both the inputs/features and the desired supervisory outputs/labels. The aim is to model the relationship between features and labels, and ultimately make predictions based on new inputs from held-out test datasets.

**Artificial neural network**

Artificial neural network (ANN) is a machine learning algorithm that is inspired by the structure of biological neural networks in animal brains (Hopfield 1982; LeCun et al. 2015). Multiple layers of artificial neurons are used to learn the representation and abstraction of the intricate structures in data at multiple levels (Hopfield 1982; LeCun et al. 2015). ANN was created back to the 1970s and 1980s (Werbos 1974; Rumelhart et al. 1986), yet the lacks of large datasets and computational powers limited the application of ANN. In recent years, with the big data explosion and the advancement of computer hardwares such as graphics processing unit (GPU), ANN has demonstrated extraordinary breakthroughs in image recognition and speech recognition, and dramatically outperformed conventional machine learning models. Unlike traditional machine learning models, deep ANN approaches do not depend on manually crafted features and can automatically extract information from large datasets in an implicit way (LeCun et al. 2015). Without stringent assumptions and restrictions, deep ANN can approximate complex mathematical functions and models to address those problems. Currently, these powerful tools

have also been successfully applied to biomedical image analysis and signal processing (Litjens et al. 2017; Shen et al. 2017; Faust et al. 2018).

**Data challenge**

A typical problem of machine learning algorithms is overfitting, which means the models excellently fit the training data but fail to make predictions on new data. This usually occurs when the dataset is small and the machine learning model is complex. A common technique to address this is cross-validation (Devyver and Kittler 1982), in which the dataset is randomly partitioned into two subsets for model training and testing, respectively. However, each time we evaluate a model using the internal test set, we probe the dataset and fit our model to it, ultimately leading to overfitting. Data challenge provide the opportunity for researchers to develop methods that are less likely to overfit, since the test dataset is stringently held-out. Meanwhile, data challenge is a unique platform to unbiasedly evaluate methods of a field in the same format. Similar to the Olympics, all data science "athletes" are evaluated according to the same standard during a data challenge. In contrast, studies reported in literatures may use different datasets or formats, making it hard to truly compare the performances of different models. Furthermore, data challenge encourages scientists to efficiently advance the state-of-the-art (Guan 2019), during which new findings may occur.

**Genomics, transcriptomics, and proteomics**

The central dogma of information flow from DNA to mRNA to protein has been applied for nearly six decades (Crick 1958). Yet, the cell functions as a whole: besides the translation from mRNA to protein, many other features are important to the complex protein expression process, including

microRNA (Lovett and Rogers 1996a), upstream open reading frame (Lovett and Rogers 1996b), cap-binding proteins (Raczynska et al. 2010), poly(A) tails (Guhaniyogi and Brewer 2001), nonsense-mediated decay (Chang et al. 2007) or alternative splicing (Black 2003). In addition, the mRNA and protein abundances are dynamic, due to ubiquitination and other degradation mechanisms to fulfill diverse condition-dependent functional requirements (Liu et al. 2016a). These complicated regulatory mechanisms underlying protein translation lead to the weak correlations between mRNA and protein abundances, when evaluating the same gene across multiple samples (Liu et al. 2016a; Vogel and Marcotte 2012; Ning et al. 2012; Zhang et al. 2014, 2016; Mertins et al. 2016). Identifying the missing factors affecting transcriptomic and proteomic correlation is important to understanding the biological mechanisms behind phenotypic variances and diseases. This is particularly true in cancers. Transcriptomic and proteomic variations across individuals are expected in diverse cancers, such as colorectal, breast, and ovarian cancers (Mertins et al. 2016; Zhang et al. 2016, 2014). These variations have important clinical consequences and implications, due to activation of different functional pathways, leading to different subtypes in the same organ, and biomarkers indicative of high- and low-risk patients in survival analysis (Zhang et al. 2014; Mertins et al. 2016; Zhang et al. 2016). These transcriptional and proteomic expression profiles provide invaluable information to studying cancer mechanisms. However, compared with the fast, inexpensive RNA sequencing profiles, large-scale high-quality proteomic data are costlier to obtain, despite remarkable progress. Therefore, a computational model to predict protein abundance from mRNA data could not only help to quickly obtain an estimation of proteomic data, but also, to understand what are the important players in cancers.

**Sleep arousal**

Sleep plays an important role in our health and wellbeing. Inadequate sleep results in many negative outcomes, including obesity, cardiovascular dysfunction, hypotension, irritability, impaired memory, and depression. About one third of the general population in United States are affected by insufficient sleep (Liu et al. 2016b). The prevalence of inadequate sleep results in large economic costs (Hillman et al. 2018). Sleep arousals are transient intrusions of wakefulness into sleep. Excessive arousals due to disturbances are harmful resulting in fragmented sleep, daytime sleepiness and sleep disorders (Bonnet 1985, 1986; Ting and Malhotra 2005). Unlike common sleep stages (wakefulness, stage1, stage2, stage3, and rapid eye movement), sleep arousals are very brief and sparsely distributed during sleep, which makes the detection difficult. Typically, each sleep stage lasts more than ten minutes and transition between sleep stages forms a unique architecture, the sleep circle. In contrast, sleep arousals are extremely short, being less than one minute, and sparsely distributed during sleep. The accumulated length of sleep arousals is usually less than 10 percent of the total sleep time. Therefore the prediction of sleep arousals is a highly imbalanced classification problem. In addition, the arousal patterns vary dramatically across individuals (e.g. some individuals do not have any arousal while others may have hundreds of arousals per night), further complexing the situation and rendering it a much more difficult task than sleep staging. Currently, polysomnographic recordings are manually examined by human experts to annotate sleep arousal events. This requires significant time and effort, due to the fact that one sleep record may contain millions of data points to be analyzed. Although pioneering progress has been made (Olsen et al. 2018; Basner et al. 2007; Behera et al. 2014; Fernández-Varela et al. 2017; Alvarez-Estevez and Fernández-Varela 2019), there is a great demand for an accurate, robust, generalizable, and fast computational tool to automatically detect sleep arousals.

**Thesis outline**

In this dissertation, I mainly focus on three projects related to different physiological systems at multiple time and length scales. In Chapter II, I describe a novel method to compare and contrast structural dynamics of evolutionarily-related proteins through PCA and MD simulations at the atom level. In Chapter III, I describe a novel trans-tissue approach for predicting proteomics from transcriptomics in cancer patient at the tissue level, using a classical machine learning model random forest (RF). In Chapter IV, I describe a novel deep neural network method for automatic segmentation of sleep arousals based on polysomnograms at the organism level. Finally in Chapter V, I summarize my work and propose future directions for these studies.

## References


Alder BJ, Wainwright TE. 1959. Studies in Molecular Dynamics. I. General Method. *J Chem Phys* **31**: 459–466.

Alvarez-Estevez D, Fernández-Varela I. 2019. Large-scale validation of an automatic EEG arousal detection algorithm using different heterogeneous databases. *Sleep Med*. https://linkinghub.elsevier.com/retrieve/pii/S1389945718303198.

Basner M, Griefahn B, Müller U, Plath G, Samel A. 2007. An ECG-based Algorithm for the Automatic Identification of Autonomic Activations Associated with Cortical Arousal. *Sleep* **30**: 1349–1361.

Behera CK, Reddy TK, Behera L, Bhattacarya B. 2014. Artificial neural network based arousal detection from sleep electroencephalogram data. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pp. 458–462, IEEE.

Bishop CM. 2016. *Pattern Recognition and Machine Learning*. Springer.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291–336.

Bonnet MH. 1985. Effect of Sleep Disruption on Sleep, Performance, and Mood. *Sleep* **8**: 11–19.

Bonnet MH. 1986. Performance and Sleepiness as a Function of Frequency and Placement of Sleep Disruption. *Psychophysiology* **23**: 263–271.

Chang Y-F, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**: 51–74.

Chung HS, Piana-Agostinetti S, Shaw DE, Eaton WA. 2015. Structural origin of slow diffusion in protein folding. *Science* **349**: 1504–1510.

Crick FH. 1958. On protein synthesis. *Symp Soc Exp Biol* **12**: 138–163.

Devyver PA, Kittler J (1946-). 1982. *Pattern Recognition: A Statistical Approach*. Prentice Hall.

Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. 2018. Deep learning for healthcare applications based on physiological signals: A review. *Comput Methods Programs Biomed* **161**: 1–13.

Fernández-Varela I, Hernández-Pereira E, Álvarez-Estévez D, Moret-Bonillo V. 2017.

Combining machine learning models for the automatic detection of EEG arousals. *Neurocomputing* **268**: 100–108.

Frenkel D, Smit B. 2002. Molecular Dynamics Simulations. *Understanding Molecular Simulation* 63–107. http://dx.doi.org/10.1016/b978-012267351-1/50006-7.

Guan Y. 2019. Waking up to data challenges. *Nature Machine Intelligence* **1**: 67–67. http://dx.doi.org/10.1038/s42256-018-0011-2.

Guhaniyogi J, Brewer G. 2001. Regulation of mRNA stability in mammalian cells. *Gene* **265**: 11–23.

Guo D, Pan AC, Dror RO, Mocking T, Liu R, Heitman LH, Shaw DE, IJzerman AP. 2016. Molecular Basis of Ligand Dissociation from the Adenosine A2A Receptor. *Mol Pharmacol* **89**: 485–491.

Hillman D, Mitchell S, Streatfeild J, Burns C, Bruck D, Pezzullo L. 2018. The economic cost of inadequate sleep. *Sleep* **41**. http://dx.doi.org/10.1093/sleep/zsy083.

Hopfield JJ. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A* **79**: 2554–2558.

Karplus M. 2002. Molecular Dynamics Simulations of Biomolecules. *Accounts of Chemical Research* **35**: 321–323. http://dx.doi.org/10.1021/ar020082r.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521**: 436–444.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. 2017. A survey on deep learning in medical image analysis. *Med Image Anal* **42**: 60–88.

Liu Y, Beyer A, Aebersold R. 2016a. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**: 535–550.

Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB. 2016b. Prevalence of Healthy Sleep Duration among Adults — United States, 2014. *MMWR Morb Mortal Wkly Rep* **65**: 137–141.

Lovett PS, Rogers EJ. 1996a. Ribosome regulation by the nascent peptide. *Microbiol Rev* **60**: 366–385.

Lovett PS, Rogers EJ. 1996b. Ribosome regulation by the nascent peptide. *Microbiol Rev* **60**: 366–385.

Maracci C, Rodnina MV. 2016. Review: Translational GTPases. *Biopolymers* **105**: 463–475. http://dx.doi.org/10.1002/bip.22832.

McCammon JA, Andrew McCammon J, Gelin BR, Karplus M. 1977. Dynamics of folded proteins. *Nature* **267**: 585–590. http://dx.doi.org/10.1038/267585a0.

Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, et al. 2016. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**: 55–62.

Milligan G, Kostenis E. 2009. Heterotrimeric G-proteins: a short history. *British Journal of Pharmacology* **147**: S46–S55. http://dx.doi.org/10.1038/sj.bjp.0706405.

Ning K, Fermin D, Nesvizhskii AI. 2012. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J Proteome Res* **11**: 2261–2271.

Olsen M, Schneider LD, Cheung J, Peppard PE, Jennum PJ, Mignot E, Sorensen HBD. 2018. Automatic, electrocardiographic-based detection of autonomic arousals and their association with cortical arousals, leg movements, and respiratory events in sleep. *Sleep* **41**. http://dx.doi.org/10.1093/sleep/zsy006.

Raczynska KD, Simpson CG, Ciesiolka A, Szewc L, Lewandowska D, McNicol J, Szweykowska-Kulinska Z, Brown JWS, Jarmolowski A. 2010. Involvement of the nuclear cap-binding protein complex in alternative splicing in Arabidopsis thaliana. *Nucleic Acids Res* **38**: 265–278.

Rosenbaum DM, Zhang C, Lyons JA, Holl R, Aragao D, Arlow DH, Rasmussen SGF, Choi H-J, Devree BT, Sunahara RK, et al. 2011. Structure and function of an irreversible agonist-β(2) adrenoceptor complex. *Nature* **469**: 236–240.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. *Nature* **323**: 533–536. http://dx.doi.org/10.1038/323533a0.

Scheffzek K, Ahmadian MR. 2005. GTPase activating proteins: structural and functional insights 18 years after discovery. *Cell Mol Life Sci* **62**: 3014–3038.

Shen D, Wu G, Suk H-I. 2017. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng* **19**: 221–248.

Ting L, Malhotra A. 2005. Disorders of sleep: an overview. *Prim Care* **32**: 305–18, v.

Vale RD. 1996. Switches, latches, and amplifiers: common themes of G proteins and molecular motors. *J Cell Biol* **135**: 291–302.

Vetter IR, Wittinghofer A. 2001. The guanine nucleotide-binding switch in three dimensions. *Science* **294**: 1299–1304.

Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**: 227–232.

Wennerberg K. 2005. The Ras superfamily at a glance. *Journal of Cell Science* **118**: 843–846. http://dx.doi.org/10.1242/jcs.01660.

Werbos P. 1974. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*.

Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, et al. 2014. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**: 382–387.

Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou J-Y, Petyuk VA, Chen L, Ray D, et al. 2016. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166**: 755–765.

Ziegel ER. 2003. The Elements of Statistical Learning. *Technometrics* **45**: 267–268. http://dx.doi.org/10.1198/tech.2003.s770.

# CHAPTER II

## Comparative Structural Dynamic Analysis of GTPases


### Abstract

GTPases regulate a multitude of essential cellular processes ranging from movement and division to differentiation and neuronal activity. These ubiquitous enzymes operate by hydrolyzing GTP to GDP with associated conformational changes that modulate affinity for family-specific binding partners. There are three major GTPase superfamilies: Ras-like GTPases, heterotrimeric G proteins and protein-synthesizing GTPases. Although they contain similar nucleotide-binding sites, the detailed mechanisms by which these structurally and functionally diverse superfamilies operate remain unclear. Here we compare and contrast the structural dynamic mechanisms of each superfamily using extensive molecular dynamics (MD) simulations and subsequent network analysis approaches. In particular, dissection of the cross-correlations of atomic displacements in both the GTP and GDP-bound states of Ras, transducin and elongation factor EF-Tu reveals analogous dynamic features. This includes similar dynamic communities and subdomain structures (termed lobes). For all three proteins the GTP-bound state has stronger couplings between equivalent lobes. Network analysis further identifies common and family-specific residues mediating the state-specific coupling of distal functional sites. Mutational simulations demonstrate how disrupting these couplings leads to distal dynamic effects at the nucleotide-binding site of each family. Collectively our studies extend current understanding of GTPase



The work presented in Chapter II are adapted from: Li H, et.al. "Comparative structural dynamic analysis of GTPases" *PLoS Comput Biol.* 2018 Nov 9;14(11):e1006364

allosteric mechanisms and highlight previously unappreciated similarities across functionally diverse families.

**Author Summary**

GTPases are a large superfamily of essential enzymes that regulate a variety of cellular processes. They share a common core structure supporting nucleotide binding and hydrolysis, and are potentially descended from the same ancestor. Yet their biological functions diverge dramatically, ranging from cell division and movement to signal transduction and translation. It has been shown that conformational changes through binding to different substrates underlie the regulation of their activities. Here we investigate the conformational dynamics of three typical GTPases by *in silico* simulation. We find that these three GTPases possess overall similar substrate-associated dynamic features, beyond their distinct functions. Further identification of key common and family-specific elements in these three families helps us understand how enzymes are adapted to acquire distinct functions from a common core structure. Our results provide unprecedented insights into the functional mechanism of GTPases in general, which potentially facilitates novel protein design in the future.

**Introduction**

Guanosine Triphosphate Phosphohydrolases (GTPases) are ubiquitous molecular machines mediating a variety of essential cellular processes (Scheffzek and Ahmadian 2005). Harnessing the GTP hydrolysis to modulate the affinity of partner molecule binding, GTPases transduce intracellular signals, control cell division and differentiation, and direct protein synthesis and translocation (Bourne et al. 1991; Simon et al. 1991; Takai et al. 2001; Jackson et al. 2010). In

general, GTP-bound GTPases in the active state are able to interact with partner effectors and regulate effector-mediated processes. GTP hydrolysis leads to the dissociation of GTPases from effectors, whereas exchange of GDP for GTP activates GTPases and restarts the signaling or protein synthesis cycle (Sprang 1997; Vetter and Wittinghofer 2001). Two classes of accessory proteins are involved in regulating this reaction cycle. GTPase-activating proteins (GAPs) accelerate the GTPase activity and the inactivation of GTPases, whereas guanine nucleotide exchange factors (GEFs) promote GDP dissociation and subsequent GTP binding, activating GTPases (Cherfils and Zeghouf 2013; Ross and Wilkie 2000; Hollinger and Hepler 2002).

There are three major GTPase superfamilies: small Ras-like GTPase, heterotrimeric G protein α subunit (Gα) and protein-synthesizing GTPase. Both small and heterotrimeric G proteins participate in signal transduction. As the primary coupling molecule to membrane receptors, Gα together with its partner βγ subunits (Gβγ) mediate the very early stage signal transduction initiated by extracellular stimuli. In contrast, small GTPase does not interact with receptors directly and regulates more downstream events in the cascade. Finally, the protein-synthesizing proteins participate in initiation, elongation and termination of mRNA translation. Underlying this functional difference are the low sequence identity (<20%) and overall different molecular shapes among these three types of GTPases. In particular, whereas small G protein consists of a single canonical Ras-like catalytic domain (RasD), Gα has an extra α-helical domain (HD) inserted and elongation factor EF-Tu has two extra β-barrel domains (D2 and D3) subsequent to the C-terminus (**Figure 2.1**). In addition, Gα can form a complex with Gβγ and undergoes a cycle of altered oligomeric states during function.

In contrast to the functional and structural diversity, GTPases display significant conservation in the core structure of the catalytic domain. Small GTPase, Gα and EF-Tu contain a RasD consisting of six β strands (β1-β6) and five α helices (α1-α5) flanking on both sides of the β sheet (**Figure 2.1**). Three highly conserved loops named P-loop (PL), switch I (SI), and switch II (SII) constitute the primary sites coordinating the nucleotide phosphates. This structural similarity suggests that at a fundamental level small GTPase, Gα and EF-Tu may utilize the same mode of structural dynamics for their allosteric regulation, which is likely inherited from their common evolutionary ancestor (Vale 1996; Leipe et al. 2002). However, it is currently unclear what are the general atomistic mechanisms underlying GTPase allostery and how these common mechanisms can be adapted to have specific function.

Recent computational and experimental studies have gained much insight into the allosteric mechanisms of individual small and heterotrimeric G protein systems. Principal component analysis (PCA) of crystallographic structures and molecular dynamics (MD) simulations characterized the structural dynamics of small GTPase Ras and revealed an intriguing dynamical partitioning of Ras structure into two lobes: the N-terminal nucleotide binding lobe (lobe1) and the C-terminal membrane anchoring lobe (lobe2) (Gorfe et al. 2008; Grant et al. 2009). Several allosteric sites were identified in lobe 2 or between lobes, including L3 (the loop between β2 and β3), L7 (the loop between α3 and β5), and α5. Importantly, α5 is the major membrane-binding site and has been related to the nucleotide modulated Ras/membrane association (Abankwa et al. 2008). In addition, binding of small molecules at L7 has been reported to affect the ordering of SI and SII (Buhrman et al. 2010). Intriguingly, recent studies of Gα have revealed nucleotide associated conformational change and bilobal substructures in the catalytic domain largely resembling those

in Ras (Yao and Grant 2013; Yao et al. 2016). The allosteric role of lobe 2, which contains the major binding interface to receptors, has also been well established for Gα (Yao et al. 2016; Marin et al. 2001; Oldham et al. 2006; Chung et al. 2011; Rasmussen et al. 2011; Kaya et al. 2014; Alexander et al. 2014; Dror et al. 2015; Sun et al. 2015; Flock et al. 2015). Furthermore, the comparison between G proteins and translational factors via sequence and structural analysis indicates a conserved molecular mechanism of GTP hydrolysis and nucleotide exchange, and cognate mutations of key residues in the nucleotide-binding regions showed similar functional effects among these systems (Bourne et al. 1991; Sprang 1997; Vetter and Wittinghofer 2001; Leipe et al. 2002). Collectively, these consistent findings from separate studies support the common allosteric mechanism hypothesis of GTPases and underscore a currently missing detailed residue-wise comparison of the structural dynamics among different GTPase superfamilies.

In this study, we compare and contrast the nucleotide-associated conformational dynamics between H-Ras (H isoform of Ras), Gαt (transducin α subunit) and EF-Tu (elongation factor thermo unstable), and describe how this dynamics can be altered by single point mutations in both common and family-specific ways. This entails the application of an updated PCA of crystallographic structures, multiple long time (80-ns) MD simulations, and recently developed network analysis approach of residue cross-correlations (Yao et al. 2016). In particular, we identify highly conserved nucleotide dependent correlation patterns across GTPase families: the active GTP-bound state displays stronger correlations both within lobe1 and between lobes, exhibiting an overall "dynamical tightening" consistent with the previous study in Gα alone (Yao et al. 2016). Detailed inspection of the residue level correlation networks along with mutational MD simulations reveal several common key residues that are potentially important for mediating the

inter-lobe communications. Point mutations of these residues substantially disrupt the couplings around the nucleotide binding regions in Ras, Gαt and EF-Tu. In addition, with the same network comparison analysis, we identify Gαt and EF-Tu specific key residues. Mutations of these residues significantly disrupt the couplings in Gαt and EF-Tu but have no or little effect in Ras. Our results are largely consistent with findings from experimental mutagenesis, with a number of dynamical disrupting mutants have been shown to have altered activities in either Ras or Gα. Our new predictions can be promising targets for future experimental testing.

## Results

### Principal component analysis (PCA) of Ras, Gαt/i and EF-Tu crystallographic structures reveals functionally distinct conformations.

Previous PCA of 41 Ras crystallographic structures revealed distinct GDP, GTP and intermediate mutant conformations (Gorfe et al. 2008). Updating this analysis to include the 121 currently available crystallographic structures reveals consistent results but with two additional conformations now evident (**Figure 2.2A**). In addition to GDP (green in **Figure 2.2A**), GTP (red), and mutant forms, GEF-bound nucleotide free (purple) and so-called 'state 1' forms (orange) are now also apparent. In the GEF-bound form, the SI region is displaced in a distinct manner – 12Å away from the nucleotide-binding site coincident with the insertion of a helix of GEF into the PL-SI cleft. The state 1 GTP-bound form was first observed via NMR and later high-resolution crystal structures were solved (Geyer et al. 1996; Araki et al. 2011; Muraoka et al. 2012). In contrast to the canonical GTP-bound conformation (red), the state 1 form (orange) lacks interaction between the two switches and the γ-phosphate of GTP, resulting in a moderate 7Å displacement of SI away from its more closed GTP conformation.

The first two PCs capture more than 75% of the total mean-square displacement of all 121 Ras structures. Residue contributions from SI and SII dominate PC1 and PC2 (**Figure 2.2D**). The height of each bar in **Figure 2.2D** displays the relative contribution of each residue to a given PC. PC1 mainly describes the opening and closing of SI – more open in GEF-bound and state 1 forms, and more closed in nucleotide bound structures. PC1 also captures smaller scale displacement of L8 (the loop between β5 and α4), which resides 5Å closer to the nucleotide-binding pocket in the GEF-bound structures than the GTP-bound structure set. PC2 depicts SII displacements and clearly separates GTP from GDP bound forms (red and green, respectively). As we expect, the lack of γ-phosphate in the GDP releases SII from the nucleotide, whereas in the GTP form SII is fixed by the hydrogen bond of the backbone amide of G60 with the γ-phosphate oxygen atom. This is also shown in the state 1 form where the hydrogen bond is disrupted with SII moderately displaced from the nucleotide (4Å on average from the canonical GTP group structures).

PCA of 53 available Gαt/i structures described recently revealed three major conformational groups: GTP (red in **Figure 2.2B**), GDP (green) and GDI (GDP dissociation inhibitor; blue) bound forms (Yao et al. 2016). The first two PCs capture over 65% of the total variance of Cα atom positions in all structures. The dominant motions along PC1 and PC2 are the concerted displacements of SI, SII and SIII in the nucleotide-binding region as well as a relatively small-scale rotation of the helical domain with respect to RasD (**Figure 2.2E**).

PC1 separates GDI-bound from non-GDI bound forms. In GDI-bound structures the GDI interacts with both the HD and the cleft between SII and SIII of the Ras-like domain, increasing the distance

between SII and SIII. Similar to Ras, PC2 of Gαt/i clearly distinguishes the GTP and GDP-bound forms, where again the unique γ-phosphate (or equivalent atom in GTP analogs) coordinates SI and SII. In addition, the SIII is displaced closer to the nucleotide, effectively closing the nucleotide-binding pocket.

PCA of 23 available full-length EF-Tu structures reveals distinct GTP and GDP conformations. PC1 dominantly captures nearly 95% of the total structural variance of Cα atom positions (**Figure 2.2C**). It mainly describes the dramatic conformational transition in SI as well as the large rotation of two β-barrel domains D2 and D3 (**Figure 2.2F**). In the GTP-bound form, the C-terminal SI is coordinated to the γ-phosphate and $Mg^{2+}$ ion, forming a small helix near SII. Meanwhile, D2 and D3 are close to RasD and create a narrow cleft with SI, serving as the binding site for tRNA (Nissen et al. 1995). In the GDP-bound form, the C-terminal helix in SI unwinds and forms a β-hairpin, protruding towards D2 and D3 (Polekhina et al. 1996). The highly conserved residue T62 (T35 in Ras) of EF-Tu moves more than 10Å away from its position in the GTP form and loses interaction with the $Mg^{2+}$ ion. In addition, D3 rotates towards SI and D2 moves far away from the Ras-like domain. In contrast to PC1, PC2 only captures a very small portion (3.59%) of the structural variance in EF-Tu (**Figure 2.2F**). The major conformational change along PC2 is a small-scale rotation of D2 and D3 with respect to RasD in the GTP form.

PCA of Ras, Gαt/i and EF-Tu demonstrates that the binding of different nucleotides and protein partners can lead to a rearrangement of global conformations in a consistent manner. In particular, within RasD, these three families display conserved nucleotide-dependent conformational distributions with major contributions from the switch regions. In the GTP-bound form of these

proteins, SI and SII are associated with the nucleotide through interacting with γ-phosphate. Despite these similarities, critical questions about their functional dynamics remain unanswered: How does nucleotide turnover lead to allosteric regulation of distinct partner protein-binding events? To what extent are the structural dynamics of these proteins similar beyond the switch region displacements evident in accumulated crystal structures? How do distal disease-associated mutations affect the functional dynamics for each family and are there commonalities across families? In the next section, we report MD simulations that address these questions, which are not answered by accumulated static experimental structures.

**MD simulations reveal distinct nucleotide-associated flexibility and cross-correlation near functional regions.**

MD simulations reveal distinct nucleotide-associated flexibility at known functional regions. Representatives of the distinct GTP and GDP-bound conformations of Ras, Gαt and EF-Tu were selected as starting points for MD simulation. Five replicated 80-ns MD simulations of these three proteins for each state (GTP and GDP totaling 2.4us; see **Materials and Methods**) exhibit high flexibility in the SI, SII, SIII/α3 and loop L3, L7, L8 and L9 regions (**Figure 2.3A-C**). The Cα atom root-mean-square fluctuation (RMSF) in Gαt shows that SI is significantly more flexible in the GDP-bound state (**Figure 2.3B**). The C-terminal SI of Ras and EF-Tu, corresponding to the shorter SI in Gαt, is also more flexible with GDP bound (**Figure 2.3A & C**). Interestingly, the middle part of SI in Ras and EF-Tu show higher fluctuations in the GTP-bound state. Moreover, SII is more flexible in the GTP-bound state in Ras. Detailed inspection reveals that SII always stays away from the nucleotide during the GDP-bound state MD simulations, whereas SII sometimes moves close to and interacts with the unique γ-phosphate of GTP, leading to higher

flexibility in the GTP-bound state. In contrast, the flexibility of SII in Gαt has no significant difference between states, whereas SII in EF-Tu is less flexible with GTP bound. This is due to the relatively compact interactions between SII and the unique D2 and D3 in the GTP-bound EF-Tu. In fact, D2 and D3 show extremely higher flexibility in the GDP state (**Figure 2.3C**). Overall, the nucleotide-dependent flexibility of RasD in Ras, Gαt and EF-Tu are quite similar except for SII.

The cross-correlations of atomic displacements derived from MD simulations also manifest conserved nucleotide-associated coupling in these three systems (**Figure 2.3D-F**). In both Ras and Gαt, significantly stronger couplings within the catalytic lobe 1 between PL, SI and SII can be found only in the GTP-bound state (red rectangles in **Figure 2.3D & E**). Interestingly, a unique inter-lobe coupling between SII and SIII/α3 also characterizes the GTP-bound state in both systems (blue rectangles in **Figure 2.3D & E**). In EF-Tu, the intra-lobe 1 and inter-lobe couplings are similar between states (red and blue rectangles in **Figure 2.3F**). Intriguingly, a lot of negative correlations between D2 and RasD of EF-Tu are found in the GDP-bound state, indicating the swing motion of D2 with respect to RasD during MD simulations (lower triangle in **Figure 2.3F**).

**Correlation network analysis displays similar nucleotide-associated correlation in Ras, Gαt and ET-Tu**

Consensus correlation networks for each nucleotide state were constructed from the corresponding replicate MD simulations. In these initial networks, each node is a residue linked by edges whose weights represent their respective correlation values averaged across simulations (see **Materials and Methods**). These residue level correlation networks underwent hierarchical clustering to

identify groups of residues (termed communities) that are highly coupled to each other but loosely coupled to other residue groups. Nine communities were identified for Ras and eleven for Gαt and EF-Tu (**Figure 2.4**). The two additional family specific communities not present in Ras correspond to two regions of HD in Gαt and D2 and D3 in EF-Tu.

In the resulting community networks the width of an edge connecting two communities is the sum of all the underlying residue correlation values between them. Interestingly, Ras, Gαt and EF-Tu community networks can be partitioned into two major groups (dashed lines in **Figure 2.4**) corresponding to the previously identified lobes for Ras and the RasD in Gαt (Gorfe et al. 2008; Yao et al. 2016). The boundary between lobes is located at the loop between α2 and β4. In these proteins, lobe1 includes the nucleotide-binding communities (PL, SI and SII) as well as the N-terminal β1-β3 and α1 structural elements. Lobe2 includes α3-α5, L8 and the C-terminal β4-β6 strands.

Comparing the GTP and GDP community networks of these three proteins reveals common nucleotide-dependent coupling features. In particular, for Ras and Gαt, comparing the relative strength of inter-community couplings in GTP and GDP networks using a nonparametric Wilcoxon test across simulation replicates reveals common significantly distinct coupling patterns (colored edges in **Figure 2.4A & B**). Within lobe1 stronger couplings between PL, SI and SII are observed for the GTP state of both families. This indicates that the γ-phosphate of GTP leads to enhanced coupling of these proximal regions. This is consistent with our PCA results above, where PC2 clearly depicts the more closed conformation of SI and SII in the GTP bound structures (**Figure 2.2D & E**). In addition, a significantly stronger inter-lobe correlation between SII and α3

is evident for the GTP state of both families, which is not available from analysis of the static experimental ensemble alone. This indicates that nucleotide turnover can lead to distinct structural dynamics not only at the immediate nucleotide-binding site in lobe 1 but also at the distal lobe 2 region.

Intriguingly, similar patterns of intra and inter-lobe dynamic correlations are observed in EF-Tu (**Figure 2.4C**). Within lobe1, significantly stronger correlations between PL-SI and PL-SII are evident in the GTP state, although SI-SII coupling becomes weaker in this state. In fact, the C-terminal β-hairpin of SI moves towards and interacts extensively with SII and D3 in the GDP bound state, leaving the nucleotide-binding site widely open. Moreover, our results reveal that SII and SIII/α3 of EF-Tu are more tightly coupled in the GTP state, resembling the strong inter-lobe couplings in the GTP bound Ras and Gαt. It is worth noting that this conserved structural dynamic coupling is evident only from the comparative network analysis and is not accessible from PCA of crystal structures.

**The common residue-wise determinants of structural dynamics in Ras, Gαt and EF-Tu.**

Comparative network analysis highlights the common residue-wise determinants of nucleotide-dependent structural dynamics. Besides correlations within lobe1, inter-lobe couplings are also significantly stronger in the GTP state networks of Ras, Gαt and EF-Tu. Inspection of the residue-wise correlations between communities reveals common major contributors to the SII – α3 couplings in the three proteins (red residues in **Table S2.1**). In particular, $M72^{Ras}$ in SII and $V103^{Ras}$ in α3 act as primary contributors to inter-lobe correlations in Ras. Interestingly, the equivalent residues in the other two systems, $F211^{G\alpha t}$ or $I93^{EF\text{-}Tu}$ in SII and $F255^{G\alpha t}$ or $V126^{EF\text{-}Tu}$ in

α3/SIII also contribute to the inter-lobe couplings. We further examined the importance of these residues by MD simulations of mutant GTP-bound systems. Results indicate that each single mutation $M72A^{Ras}$ and $V103A^{Ras}$ can significantly reduce the couplings between SI and PL, indicating that these mutations disturb couplings at distal sites of known functional relevance (**Figure 2.5A & D**). Moreover, the cognate mutations $F211A^{G\alpha t}$ and $F255^{G\alpha t}$ in Gαt not only decouple SI and PL but also SI and SII (**Figure 2.5B & E**). Similarly, the analogous mutation $I93A^{EF\text{-}Tu}$ decreases the correlations between PL and SI, whereas $V126A^{EF\text{-}Tu}$ decouples PL and SII (**Figure 2.5C & F**). The simulation results indicate that single alanine mutation of residues contributing to SII-α3 couplings diminishes the couplings of the nucleotide binding regions, and this allosteric effect is common in all the three proteins.

Inter-lobe couplings that are distal from the nucleotide binding regions are also shown to be critical for the nucleotide dependent dynamics in Ras, Gαt and EF-Tu. By inspecting the residue level couplings between L3 and α5, we identified common distal inter-lobe couplings in the three proteins. Mutational simulations indicate that the substitutions $K188A^{G\alpha t}$ and $D337A^{G\alpha t}$ significantly decouple SI from the PL and SII regions (**Figure 2.6B & E**). Interestingly, the mutations $K188A^{G\alpha t}$ and $D337A^{G\alpha t}$ have been reported to cause a 6-fold and 2-fold increase in nucleotide exchange, respectively, but no direct structural dynamic mechanism was established (Marin et al. 2001). We further tested mutations of analogous residues in Ras. We considered both $D47^{Ras}$ and $E49^{Ras}$ as the equivalent residues to $K188^{G\alpha t}$ (due to the longer L3 region of Ras), and $R164^{Ras}$ as the equivalent residue to $D337^{G\alpha t}$. Both double mutation $D47A/E49A^{Ras}$ and single mutation $R164A^{Ras}$ significantly reduce the correlations between PL and SI (**Figure 2.6A & D**). We note that the functional consequences of mutating these residues in Ras has been highlighted

in a previous study, in which the salt bridges between D47/E49$^{Ras}$ in L3 and R161/R164$^{Ras}$ in α5 were shown to be involved in the reorientation of Ras with respect to the plasma membrane, and enhanced activation of MAPK pathway (Abankwa et al. 2008). Moreover, substitutions of analogous residues R75A$^{EF\text{-}Tu}$ (L3) and D207A$^{EF\text{-}Tu}$ (α5) also significantly reduce the couplings between PL and SI (**Figure 2.6C & F**). Our results indicate that the conserved interactions between L3 and α5 are important for maintaining the close coordination of the distal SI, SII and PL around the nucleotide, and this is common to these three proteins.

**Network analysis identifies family-specific residue substitutions that can also perturb structural dynamics.**

Comparison of the GTP-bound residue-wise networks of Ras, Gαt and EF-Tu reveals that the N-terminus of α3 strongly couples SII only in Gαt and EF-Tu. In particular, we identified residues R201$^{Gαt}$ or A86$^{EF\text{-}Tu}$ (SII) and E241$^{Gαt}$ or Q115$^{EF\text{-}Tu}$ (α3) as underlying these strong couplings (blue residues in **Table S2.1**). These residues are specific to Gαt and EF-Tu because the corresponding residues E62$^{Ras}$ in SII and K88$^{Ras}$ in α3 have no contribution in Ras (green residues in **Table S2.1**). Mutational MD simulations indicate that substitutions E241A$^{Gαt}$ and Q115A$^{EF\text{-}Tu}$ have a similar drastic effect on the coupling of nucleotide binding regions (**Figure S2.1**). In particular, the couplings between PL, SII and PL are all significantly reduced (**Figure S2.1B & C**). We note that E241A$^{Gαt}$ in Gαs (the α subunit of the stimulatory G protein for adenylyl cyclase) was previously reported to impair GTP binding but the structural basis for this allosteric effect has been unknown (Iiri et al. 1997, 1999). Our results indicate that weakened correlations of the nucleotide-binding regions in E241A$^{Gαt}$ as a consequence of allosteric mutations in SIII/α3 and SII likely underlie the reported impaired GTP binding. Moreover, we identified residue E232$^{Gαt}$ as a Gαt-specific primary

contributor to the inter-lobe couplings in SIII, which has no direct counterparts in Ras or EF-Tu due to the absence of SIII (purple residues in **Table S2.1**). The simulation of mutation E232A$^{G\alpha t}$ shows diminished couplings between PL, SI and SII, as well (**Figure S2.2A**). Similar effects of mutations R201A$^{G\alpha t}$ and D234A$^{G\alpha t}$ are also observed (**Figure S2.2B & C**).

Mutations of the counterpart residues E62A$^{Ras}$ and K88A$^{Ras}$ result in no significant change in the coupling of nucleotide binding loops in Ras (**Figure S2.1A**). Collectively these findings indicate that in Gαt and EF-Tu both N- and C-terminal α3 positions dynamically couple with SII, whereas in Ras the communication between α3 and SII is mainly through the C-terminus of α3. In addition, our results suggest that SIII plays a unique role in Gαt not only mediating the couplings between the two lobes but also allosterically maintaining the tight correlations between SI, SII and PL.

**Discussion**

In this work, our updated PCA of Ras structures captures two new conformational clusters representing the GEF-bound state and "state 1", respectively, in addition to the canonical GTP and GDP forms. By comparing the Ras PCA to PCA of Gαt/i and EF-Tu, we reveal common nucleotide dependent collective deformations of SI and SII across G protein families. Our extensive MD simulations and network analyses reveal common nucleotide-associated conformational dynamics in Ras, Gαt and EF-Tu. Specifically, these three systems have stronger intra-lobe1 (PL – SI and PL – SII) and inter-lobe (SII – SIII/α3) couplings in the GTP-bound state. Meanwhile, with the network comparison approach we further identify residue-wise determinants of commonalities and specificities across families. Residues M72$^{Ras}$ (SII), V103$^{Ras}$ (α3), D47/E49$^{Ras}$ (L3) and R164$^{Ras}$ (α5) are predicted to be crucial for inter-lobe communications in Ras. Mutations of these distal

residues display decreased coupling strength in SI – PL. Interestingly, the analogous residues in the other two proteins, F211$^{G\alpha t}$/I93$^{EF\text{-}Tu}$ (SII), F255$^{G\alpha t}$/V126$^{EF\text{-}Tu}$ (α3), K188$^{G\alpha t}$/R75$^{EF\text{-}Tu}$ (L3) and D337$^{G\alpha t}$/D207$^{EF\text{-}Tu}$ (α5) also have important inter-lobe couplings and show similar decoupling effects upon alanine mutations. Besides the key residues that are common in the three systems, residues mediating inter-lobe couplings only in Gαt and EF-Tu are identified. These include R201$^{G\alpha t}$/A86$^{EF\text{-}Tu}$ and E241$^{G\alpha t}$/Q115$^{EF\text{-}Tu}$, whose cognates in Ras do not have significant effect on the nucleotide-binding regions upon mutation. In addition, Gαt specific residue E232$^{G\alpha t}$ in SIII (which is missing in Ras and EF-Tu) is identified to be important to the couplings of the nucleotide-binding regions. Importantly, some of our highlighted mutants (D47A/E49A$^{Ras}$, K188A$^{G\alpha t}$, D207A$^{G\alpha t}$ and R241A$^{G\alpha t}$) have been reported to have functional effects by *in vitro* experiments. Our analysis provides insights into the atomistic mechanisms of these altered protein functions.

Using differential contact map analysis of crystallographic structures, Babu and colleagues recently suggested a universal activation mechanism of Gα (Flock et al. 2015). In their model, structural contacts between α1 and α5 act as a 'hub' mediating the communications between α5 and the nucleotide. These contacts are broken upon the binding of receptor at α5, leading to a more flexible α1 and the destabilization of nucleotide binding. According to their studies, however, these critical α1/α5 contacts do not exist in Ras structures. Thus, they concluded that, unlike Gα, α5 in Ras does not have allosteric regulation of the nucleotide. It is worth noting that Babu's work is purely based on the comparison of structures without considering protein dynamics. In fact, our study indicates that functionally important communications may not be directly observed from static structures. For example, the inter-lobe couplings between SII and SIII/α3 are not captured by PCA of structure ensemble, but they are clearly shown in our network analysis of structural

dynamics. By inspecting structural dynamics, we find that α5 in Ras actually plays an allosteric role, in which point mutation (R164A) substantially disrupts the couplings in the nucleotide binding regions. The potential salt bridges between D47/E49 in L3 and R161/R164 in α5 are shown in **Figure S2.3**.

A previous study of Ras GTPases via an elastic network model – normal mode analysis (ENM-NMA) revealed similar bilobal substructures and found that functionally conserved modes are localized in the catalytic lobe1, whereas family-specific deformations are mainly found in the allosteric lobe2 (Raimondi et al. 2010). The subsequent study via MD, in contrast, indicated that the conformational dynamics of Ras and Gαt are distinct, especially in the GDP state (Raimondi et al. 2011). We note that in that study only a single MD simulation trajectory was analyzed, which is insufficient to assess the significance of the observed difference. Moreover, few atomistic details were given in that work. In our study, we make improvements by building ensemble-averaged networks based on multiple MD simulations instead of a single trajectory. This increases the robustness of the networks and largely reduces statistical errors. In addition, our correlation analysis provides residue wise predictions of potential important positions that mediate communications between functional regions. Overall, separation of functionally conserved and specific residues in conformational dynamics provides us unprecedented insights into protein evolution and engineering.

## Materials and Methods

### Crystallographic structures preparation

Atomic coordinates for all available Ras, Gαt/i and EF-Tu crystal structures were obtained from the RCSB Protein Data Bank (Rose et al. 2017) via sequence search utilities in the Bio3D package version 2.2 (Grant et al. 2006; Skjærven et al. 2014). Structures with missing residues in the switch regions were not considered in this study, resulting in a total of 143 chains extracted from 121 unique structures for Ras, 53 chains from 36 unique structures for Gαt/i and 34 chains from 23 unique structures for EF-Tu. Prior to analyzing the variability of the conformational ensemble, all structures were superposed iteratively to identify the most structurally invariable region. This procedure excludes residues with the largest positional differences (measured as an ellipsoid of variance determined from the Cartesian coordinate for equivalent Cα atoms) before each round of superposition, until only invariant "core" residues remained (Gerstein and Altman 1995). The identified "core" residues were used as the reference frame for the superposition of both crystal structures and subsequent MD trajectories.

**Principal component analysis**

PCA was employed to characterize inter-conformer relationships of both Ras and Gαt/i. PCA is based on the diagonalization of the variance-covariance matrix, Σ, with element $\Sigma_{ij}$ built from the Cartesian coordinates of Cα atoms, r, of the superposed structures:

$$\Sigma_{ij} = \langle (r_i - \langle r_i \rangle) \rangle \cdot \langle (r_j - \langle r_j \rangle) \rangle,$$

where i and j enumerate all 3N Cartesian coordinates (N is the number of atoms being considered), and $\langle \cdot \rangle$ denotes the average value. The eigenvectors, or principal components, of Σ correspond to a linear basis set of the distribution of structures, whereas each eigenvalue describes the variance of the distribution along the corresponding eigenvector. Projection of the conformational ensemble

onto the subspace defined by the top two largest PCs provides a low-dimensional display of structures, highlighting the major differences between conformers.

**Molecular dynamics simulations**

Similar MD simulation protocols as those used in (Yao et al. 2016) were employed. Briefly, the AMBER12 (http://ambermd.org/) and corresponding force field ff99SB (Hornak et al. 2006) were exploited in all simulations. Additional parameters for guanine nucleotides were taken from Meagher *et al.* (Meagher et al. 2003). The $Mg^{2+}$·GDP-bound Ras crystal structure (PDB ID: 4Q21), Gαt structure (PDB ID: 1TAG) and EF-Tu structure (PDB ID: 1TUI) were used as the starting point for GDP-bound simulations. The $Mg^{2+}$·GNP (PDB ID: 5P21), the $Mg^{2+}$·GSP (PDB ID: 1TND) and the $Mg^{2+}$·GNP (PDB ID: 1TTT) bound structures were used as the starting point for GTP-bound simulations of Ras, Gαt and EF-Tu, respectively. These structures were identified as cluster representatives from PCA of the crystallographic structures. Prior to MD simulations, the sulfur (S1γ)/nitrogen (N3β) atom in the GTP-analogue was replaced with the corresponding oxygen (O1γ) / oxygen (O3β) of GTP. All Asp and Glu were deprotonated whereas Arg and Lys were protonated. The protonation state of each His was determined by its local environment via the PROPKA method (Olsson et al. 2011). Each protein system was solvated in a cubic pre-equilibrated TIP3P water box, where the distance was at least 12Å from the surface of the protein to any side of the box. Then sodium ions ($Na^+$) were added to neutralize the system. Each MD simulation started with a four-stage energy minimization, and each stage employed 500 steps of steepest descent followed by 1500 steps of conjugate gradient. First, the atomic positions of ligands and protein were fixed and only solvent was relaxed. Second, ligands and protein side chains were relaxed with fixed protein backbone. Third, the full atoms of ligands and protein were relaxed with fixed solvent. Fourth, all atoms were free to relax with no constraint. Subsequent to energy

minimization, 1ps of MD simulation was performed to increase the temperature of the system from 0K to 300K. Then 1ns of simulations at constant temperature (T=300K) and pressure (P=1bar) was further performed to equilibrate the system. Finally, 80ns of production MD was performed under the same condition as the equilibration. For long-range electrostatic interactions, particle mesh Ewald summation method was used, while for short-range non-bonded Van der Waals' interactions, an 8Å cutoff was used. In addition, a 2-fs time step was use. The center-of-mass motion was removed every 1000 steps and the non-bonded neighbor list was updated every 25 steps.

We performed a total of 1,920 ns MD simulations and analyzed results from multiple production phase 80ns simulations for each of our 3 systems, including the wild type in two nucleotide states and (5 x ras / 8 x Gat / 5 x EF-Tu) mutant systems in the GTP-bound states (i.e. 80ns x (7 + 10 +7) = 1,920 ns; see details in **Tables S2.2**). The RMSD time courses for the above systems are shown in **Figure S2.4**.

**Correlation network construction**

Consensus correlation networks were built from MD simulations to depict dynamic couplings among functional protein segments. A weighted network graph was constructed where each node represents an individual residue and the weight of edge between nodes, i and j, represents their Pearson's inner product cross-correlation value cij (Ichiye and Karplus 1991) during MD trajectories. The approach is similar to the dynamical network analysis method introduced by Luthey-Schulten and colleagues (Sethi et al. 2009). However, instead of using a 4.5Å contact map of non-neighboring residues to define network edges, which were further weighted by a single

correlation matrix, we constructed consensus networks based on five replicate simulations in the same way as described before (Yao et al. 2016).

**Network community**

Hierarchical clustering was employed to identify residue groups, or communities, that are highly coupled to each other but loosely coupled to other residue groups. We used a betweenness clustering algorithm similar to that introduced by Girvan and Newman (Girvan and Newman 2002). However, instead of partitioning according to the maximum modularity score, which is usually used in unweighted networks, we selected the partition closest to the maximum score but with the smallest number of communities (i.e. the earliest high scoring partition). This approach avoided the common cases that many small communities were generated with equally high partition scores. The resulting networks under different nucleotide-bound states showed largely consistent community partition in Ras, Gαt and EF-Tu, with differences mainly localized at the nucleotide binding PL, SI, SII and α1 regions. To facilitate comparison between states and families, the boundary of these regions was re-defined based on known conserved functional motifs. Re-analysis of the original residue cross-correlation matrices with the definition of communities was then performed. Only inter-community correlations were of interest, which were calculated as the sum of all underlying residue correlation values between two given communities satisfying that the smallest atom-atom distance between corresponding residue pairs was less than 4.5Å (for Gαt and EF-Tu) or 6 Å (for Ras) for more than 75% of total simulation frames. A larger cutoff was selected for Ras because the overall residue level correlations are weaker in Ras. A standard nonparametric Wilocox test was performed to evaluate the significance of the differences of inter-community correlations between distinct states.

## Figures

**Figure 2.1 Structural comparison of Ras, Gαt and EF-Tu reveals common canonical Ras-like domain.** The Ras-like domains of Ras (**A**), Gαt (**B**) and EF-Tu (**C**) are shown in cartoon and the extra domains in Gαt and EF-Tu are shown as gray tubes. Highly conserved regions (PL, SI, and SII) and helices (α1, α3, α4, and α5) are labeled. The PDB IDs of these three structures are 5P21 (Ras), 1TND (Gαt) and 1TTT (EF-Tu).

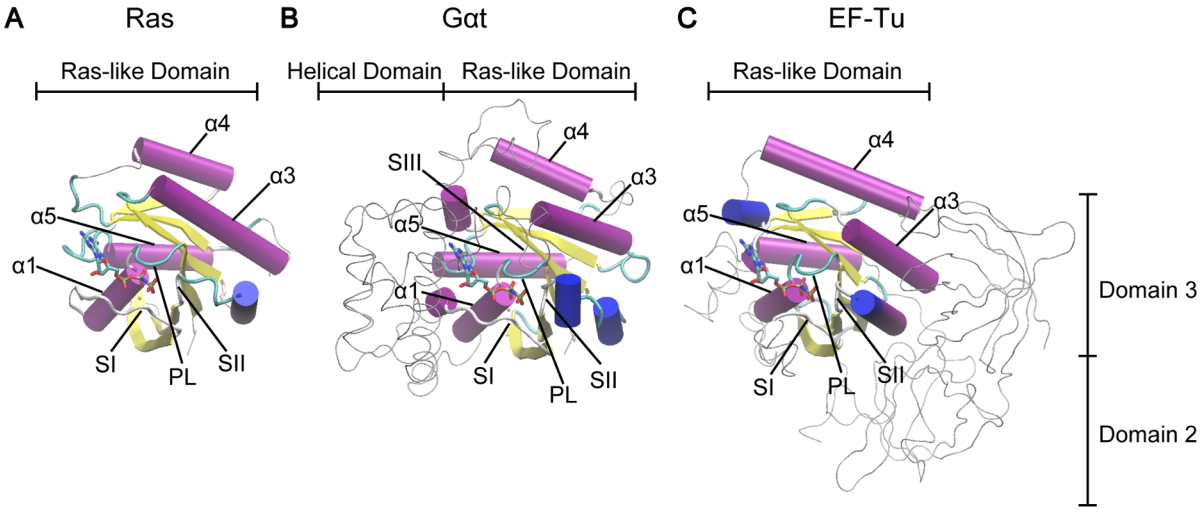


**Figure 2.2 Principal component analysis of Ras, Gαt/i and EF-Tu crystallographic structures reveals distinct nucleotide-associated conformations.** (**A-C**) Projection of 121 Ras (**A**), 53 Gαt/i (**B**) and 23 EF-Tu (**C**) PDB structures (represented as squares) onto the first two PCs reveals different conformational clusters corresponding to GTP (red), GDP (green), GEF (purple) and GDI (blue) bound states. A distinct cluster of GTP-bound structures in Ras corresponds to the "State 1" state (orange). The inserted figures show that the first two PCs capture 76.1%, 65.4% and 97.7% of the total structural variances in Ras, Gαt/i and EF-Tu, respectively. (**D-F**) The contributions of each residue to PC1 (brown) and PC2 (grey) show that the switch regions mainly correspond to the accumulated structural differences in Ras (**D**) and Gαt/i (**E**). In addition to switch regions, Domain 2 and Domain 3 also contribute to the structure differences in EF-Tu (**F**). The marginal black and grey rectangles with labels on top of them represent the location of alpha-helix and beta-strand secondary structures.

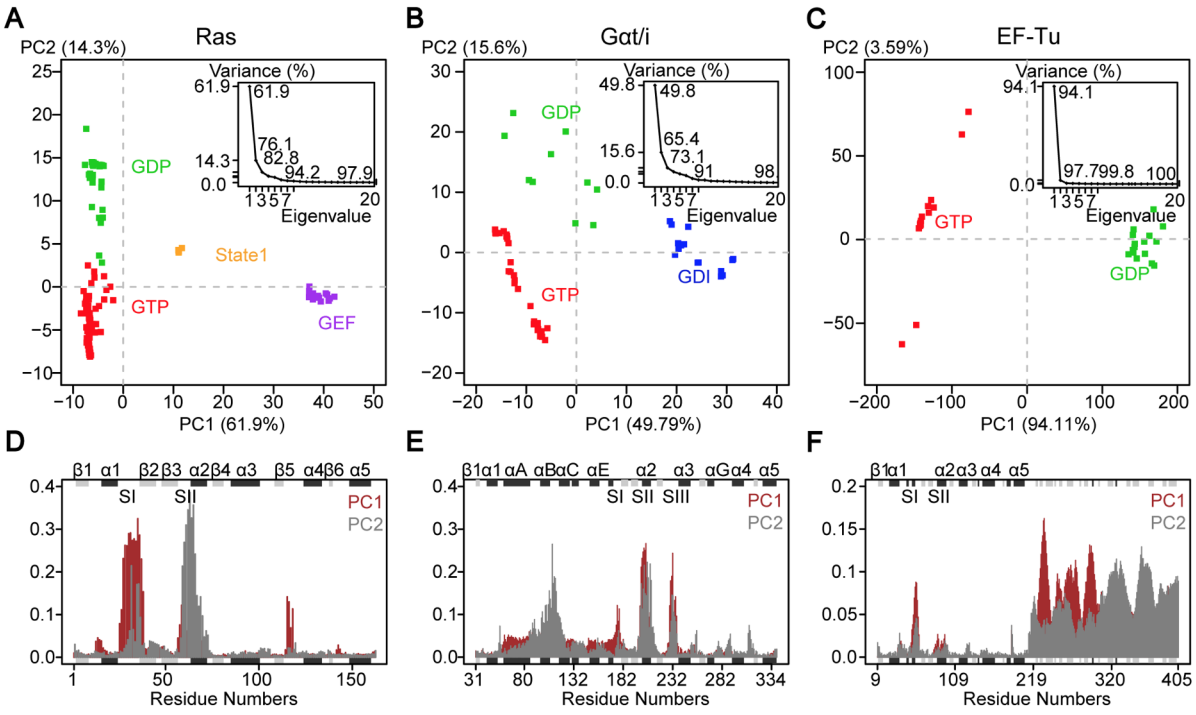


**Figure 2.3 Nucleotide specific residue fluctuations and cross-correlations of atomic displacements from molecular dynamics simulations.** (**A-C**) The ensemble averaged root-mean-square fluctuation (RMSF) reveals nucleotide dependent flexibilities that are consistent in the Ras-like domain of Ras (**A**), Gαt (**B**) and EF-Tu (**C**). Residues with significant differences (*p*-value < 0.01) between GTP and GDP bound states are highlighted with dashed lines. (**D-F**) The cross-correlations reveal stronger intra-lobe1 couplings between PL, SI and SII (red rectangles) and inter-lobe couplings between SII and SIII/α3 (blue rectangles) in the GTP-bound state (upper triangle) for both Ras (**D**) and Gαt (**E**).

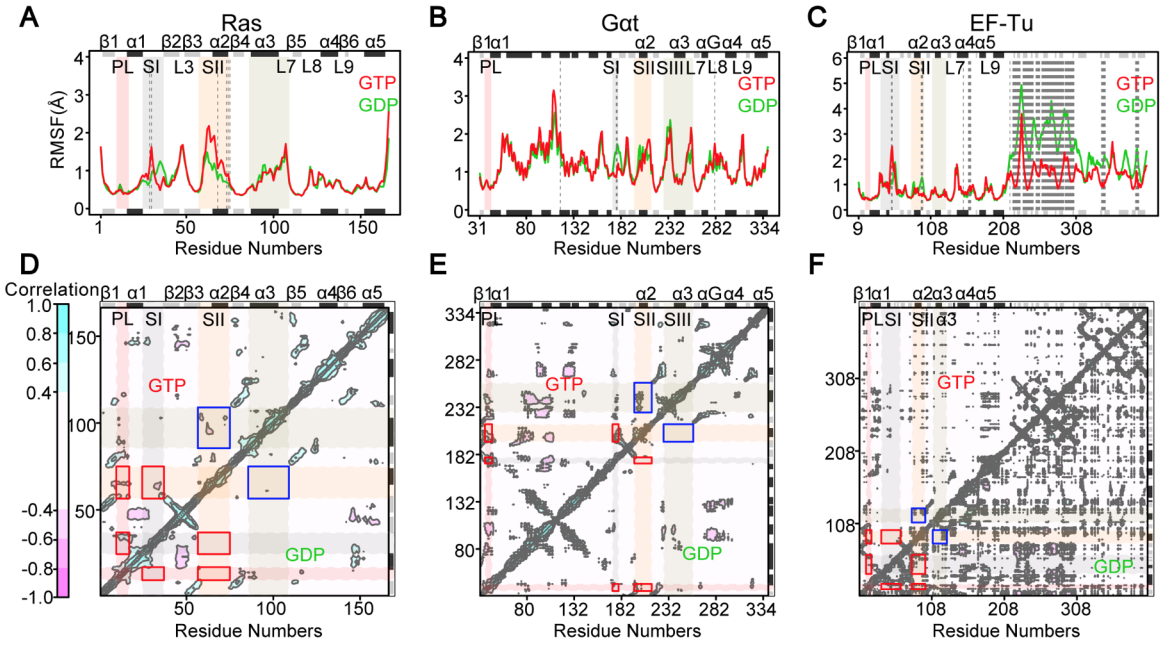


**Figure 2.4 Correlation network analysis reveals similar patterns of nucleotide-dependent couplings in Ras, Gαt and EF-Tu.** (**A**) Network communities are represented as colored circles with different radius indicating the number of residues within the community. The width of an edge is determined by the summation of all residue level correlation values between two connected communities. Red and green edges indicate enhanced GTP or GDP couplings that are significantly ($p$-value < 0.05) or more than two-fold stronger in one state than the other. All other lines are colored gray. Dashed lines with a light gray background represent the two-lobe substructures. (**B & C**) Similar nucleotide-associated network patterns are evident in the GTP (top) and GDP (bottom) bound state of Gαt (**B**) and EF-Tu (**C**), except for the SI and SII coupling.

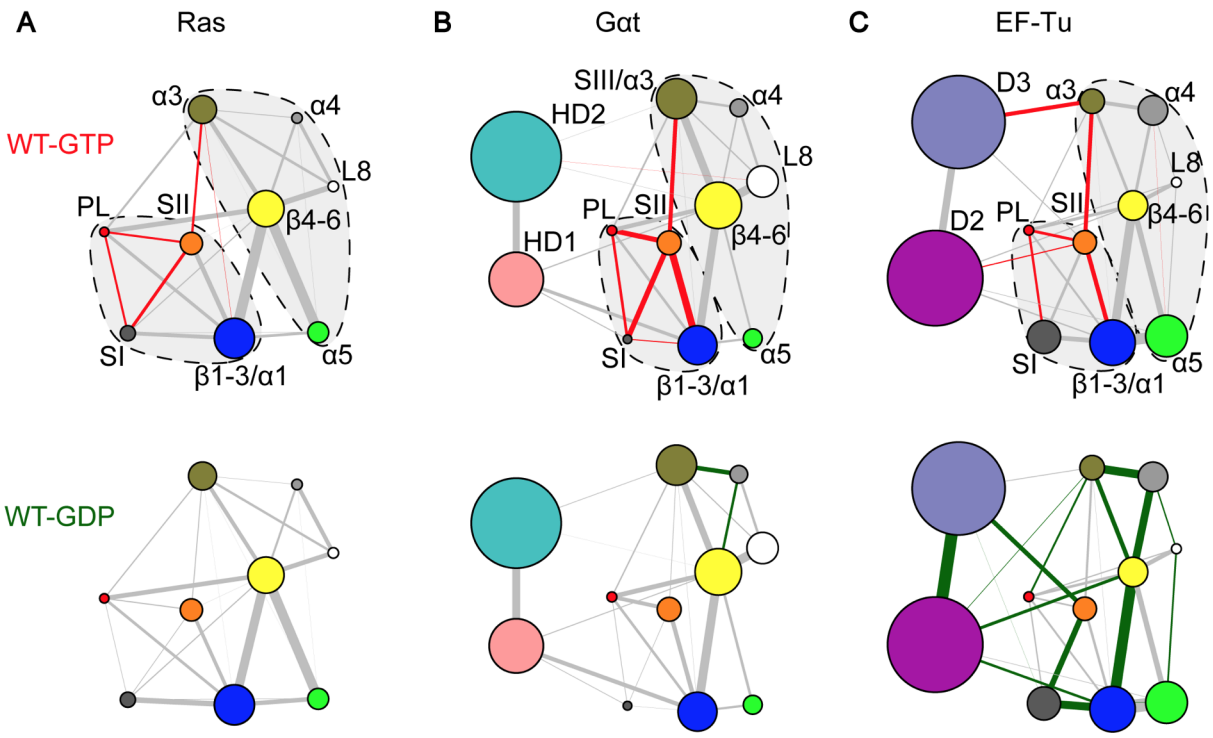


**Figure 2.5 Mutations of common residue-wise determinants of structural dynamics between SII and α3 have similar effects in Ras, Gαt and EF-Tu.** Mutations M72A$^{Ras}$ in SII (**A**) and V103A$^{Ras}$ in α3 (**D**) significantly reduce the couplings between PL and SI. The counterpart mutations in Gαt and EF-Tu, F211A$^{Gαt}$ in SII (**B**), F255A$^{Gαt}$ in α3 (**E**), I93A$^{EF-Tu}$ in SII (**C**) and V126A$^{EF-Tu}$ in α3 (**F**) have similar effects in the nucleotide-binding region – significantly reducing the coupling between PL, SI and SII.

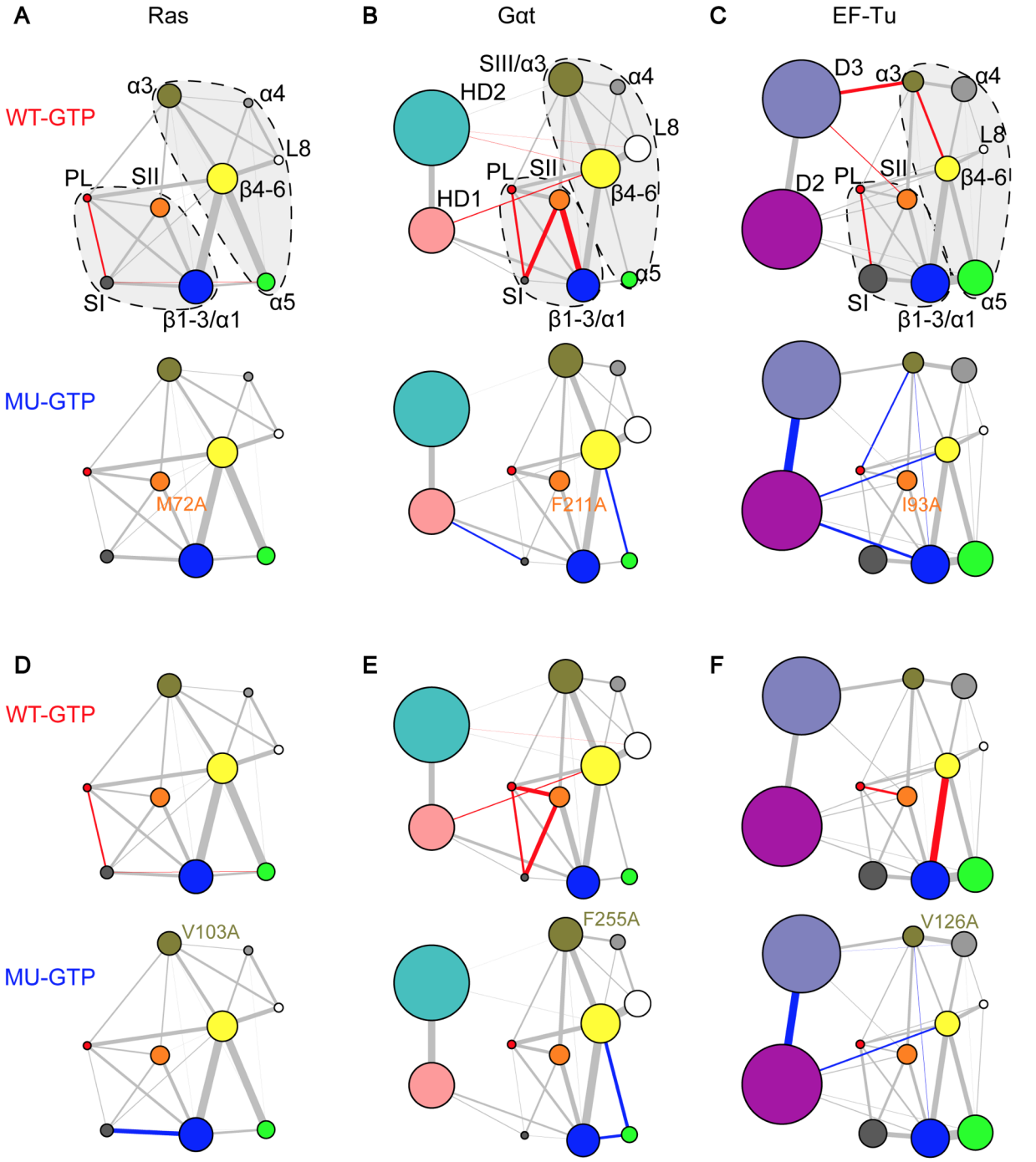
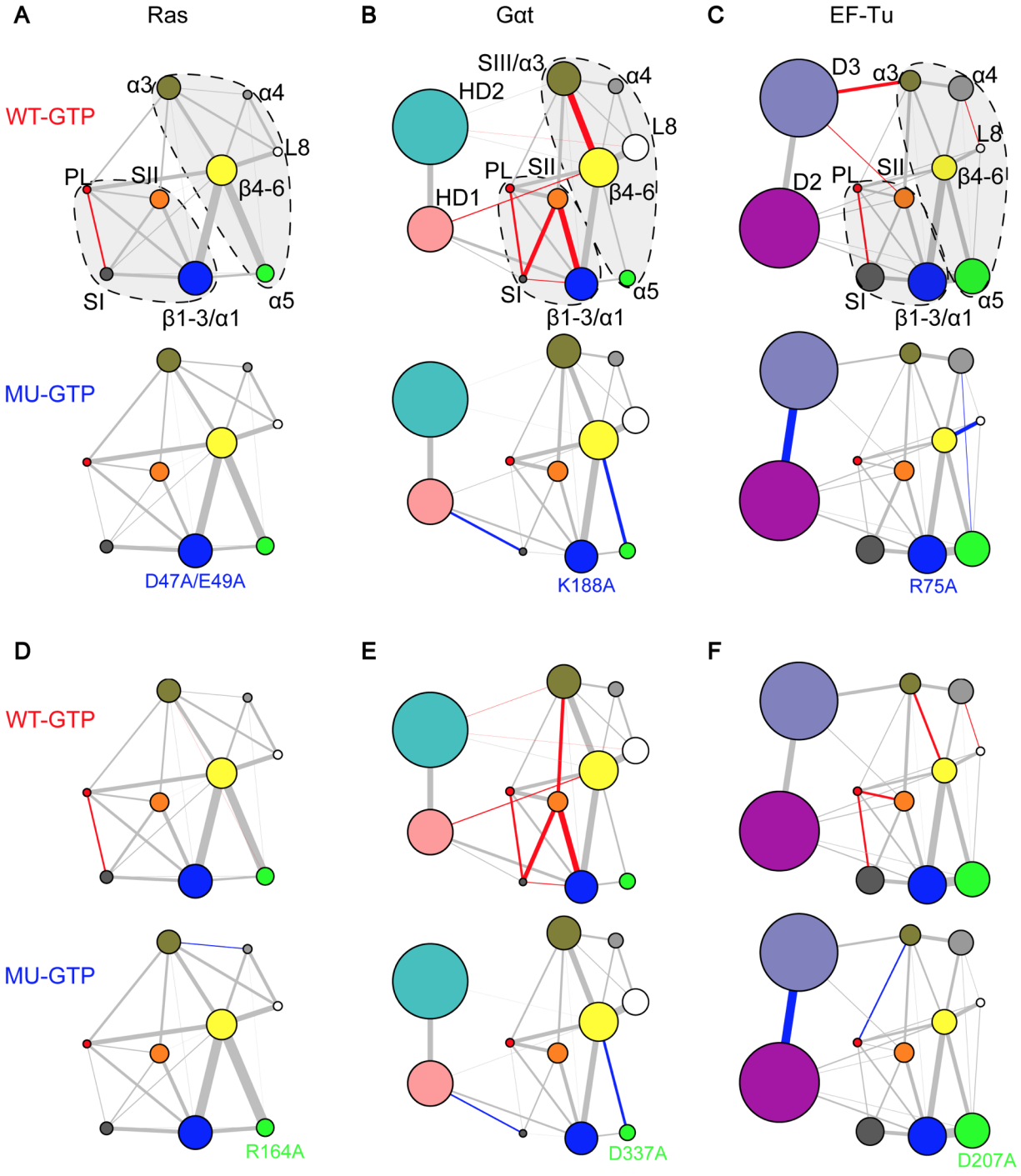


**Figure 6. Mutations of common residue-wise determinants of structural dynamics between L3 and α5 have similar effects in Ras, Gαt and EF-Tu.** Mutations D47A/E49A$^{Ras}$ in L3 (**A**) and R164A$^{Ras}$ in α5 (**D**) significantly reduce the couplings between PL and SI. The counterpart mutations in Gαt and EF-Tu, K188A$^{Gαt}$ in L3 (**B**), D337A$^{Gαt}$ in α5 (**E**), R75A$^{EF-Tu}$ in L3 (**C**) and D207A$^{EF-Tu}$ in α5 (**F**) have similar effects in the nucleotide-binding region – significantly reducing the coupling between PL, SI and SII.



**Supplementary Figures and Tables**

**Figure S2.1 Mutations of distal Gαt and EF-Tu specific residues perturb structural dynamics at nucleotide binding regions.**
In each panel, networks of wild type GTP-bound (WT-GTP, top) and mutant GTP-bound (MU-GTP, bottom) are compared. Red and blue edges indicate enhanced WT or MU couplings that are significantly (*p*-value <0.05). All other lines are colored gray. Specific mutations E241A$^{Gαt}$ (**B**) and Q115A$^{EF\text{-}Tu}$ (**C**) in α3 dramatically reduce the couplings between the functional regions PL, SI and SII, whereas the counterpart mutation K88A$^{Ras}$ (**A**) has minor effects.

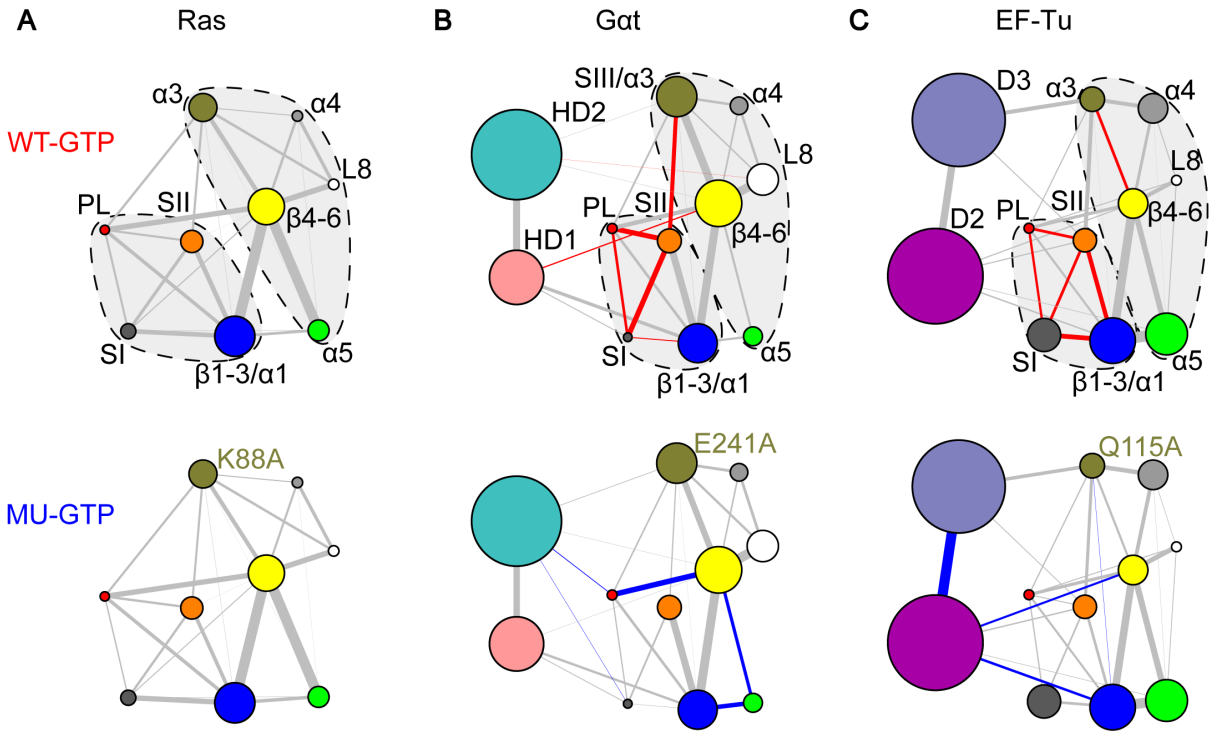


**Figure S2.2 Mutations of distal Gαt specific residues perturb structural dynamics at nucleotide binding regions.**
In each panel, networks of wild type GTP-bound (WT-GTP, top) and mutant GTP-bound (MU-GTP, bottom) are compared. Red and blue edges indicate enhanced WT or MU couplings that are significantly ($p$-value <0.05). All other lines are colored gray. Gαt specific mutations E232A$^{Gαt}$ (**A**) in SIII dramatically reduce the couplings between the functional regions PL, SI and SII. Similar effects of mutations R201A$^{Gαt}$ (**B**) and D234A$^{Gαt}$ (**C**) are also observed in Gαt.

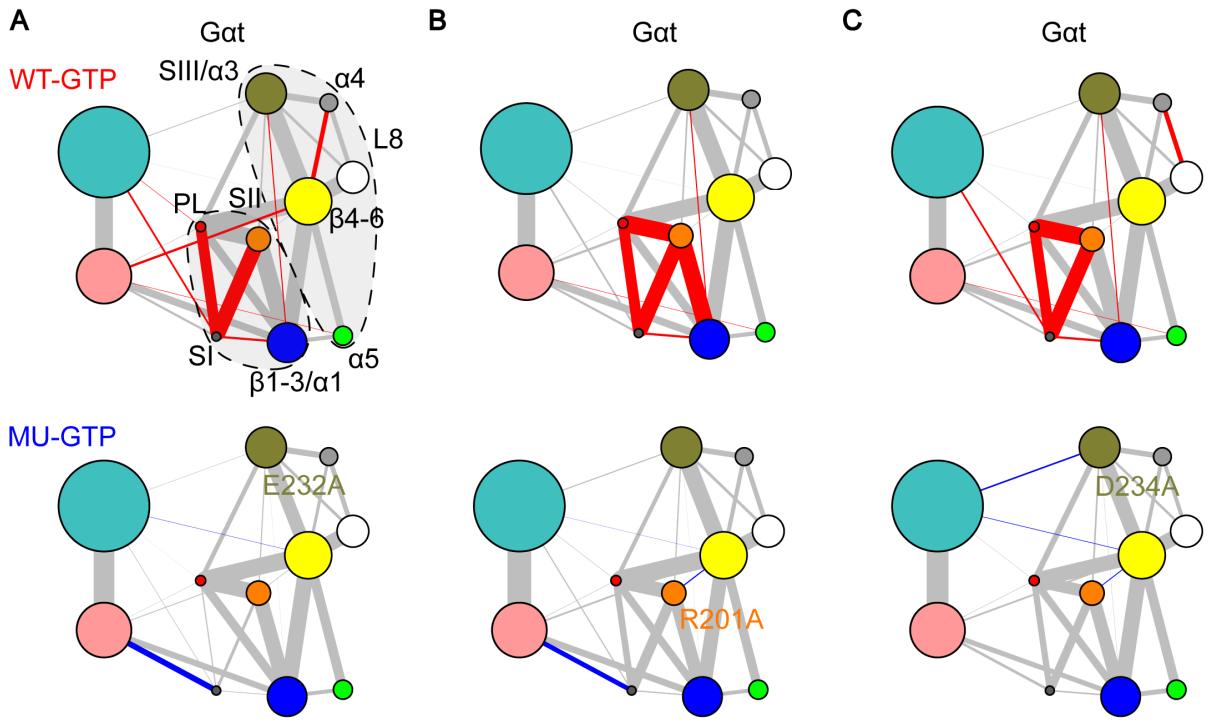


**Figure S2.3 The potential salt bridges between D47/E49 in L3 and R161/R164 in α5 in Ras-GTP wild type.**

The L3 loop and helix α5 are shown as secondary structure cartoons in blue and green respectively. The side chains of the noted residues are highlighted, with oxygen atoms in red and nitrogen atoms in blue. Labeled distances are in the unit of Angstrom (Å).

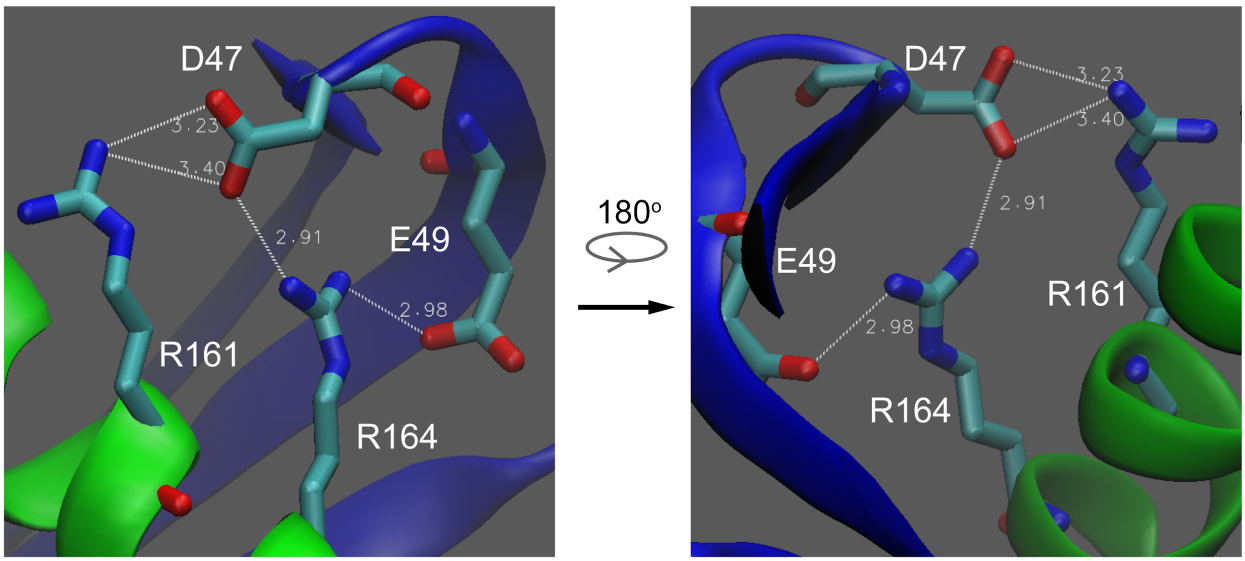


**Figure S2.4 The RMSD time-course plots of all 24 MD simulation systems.**
In each system, the five simulation replicates are shown in five different colors.

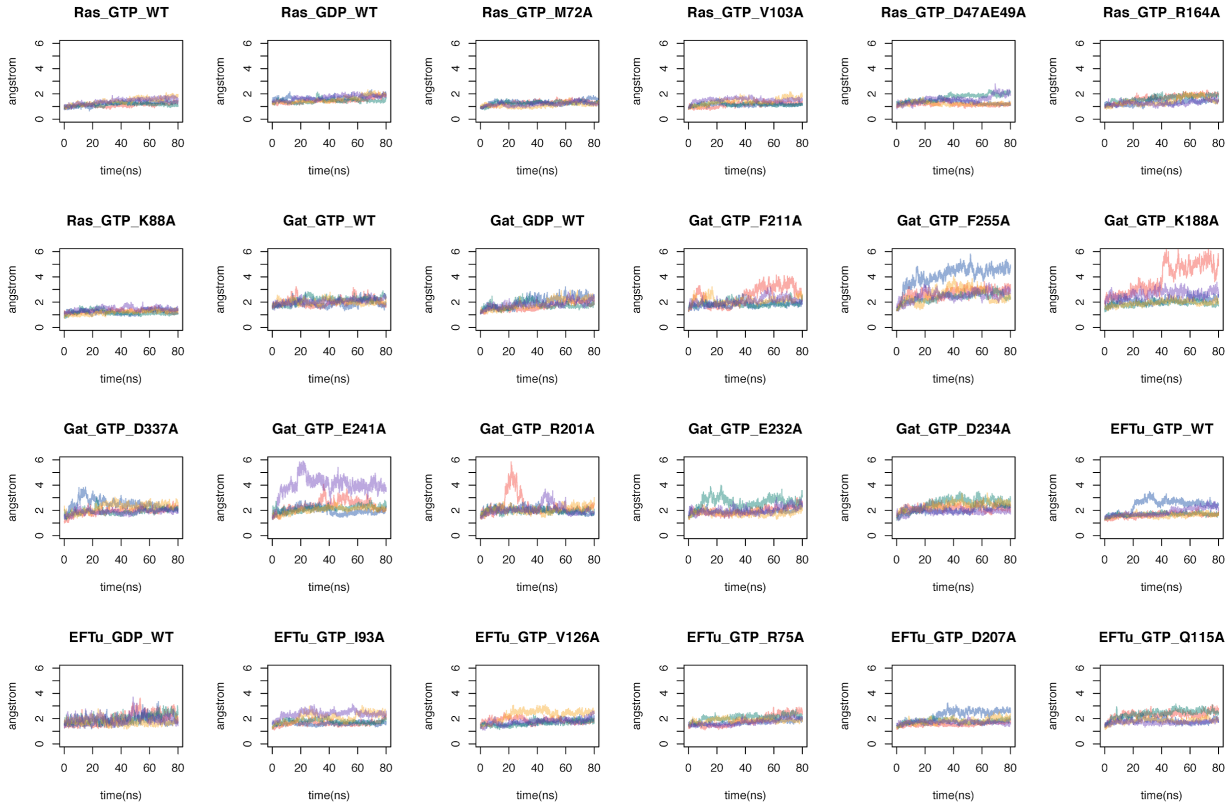


**Table S2.1 Residue-wise contributions to inter-community couplings.**
The numbers represent the residue-wise contributions to inter-community couplings. For example, the sum of correlations between residue M72 in SII and all residues in SIII/ α3 is 1.19 (after filtering by contact map). The first row contains common counterpart residues (red) connecting SII and SIII/α3 in three proteins. The second row contains family-specific functional residues: residues in Gαt and EF-Tu (blue) contribute to the dynamic correlations between SII and SIII/α3, whereas their counterparts in Ras (green) have no contributions. The third row contains Gαt specific residue in SIII, which has no counterparts in the other two proteins.

| SII | | | SIII/α3 | | |
|---|---|---|---|---|---|
| Ras | Gαt | EF-Tu | Ras | Gαt | EF-Tu |
| M72 | F211 | I93 | V103 | F255 | V126 |
| 1.19 | 0.5 | 0.88 | 0.96 | 0.26 | 1.71 |
| E62 | R201 | A86 | K88 | E241 | Q115 |
| 0 | 1.63 | 0.23 | 0 | 1.14 | 0.23 |
| | | | NA | E232(SIII) | NA |
| | | | | 1.03 | |

**Table S2.2 Summary of systems simulated.**

| Protein | Nucleotide | Mutation | Simulation Length |
|---|---|---|---|
| Ras | GTP | WT | 80 ns |
| Ras | GDP | WT | 80 ns |
| Ras | GTP | M72A | 80 ns |
| Ras | GTP | V103A | 80 ns |
| Ras | GTP | D47A; E49A | 80 ns |
| Ras | GTP | R164A | 80 ns |
| Ras | GTP | K88A | 80 ns |
| Gαt | GTP | WT | 80 ns |
| Gαt | GDP | WT | 80 ns |
| Gαt | GTP | F211A | 80 ns |
| Gαt | GTP | F255A | 80 ns |
| Gαt | GTP | K188A | 80 ns |
| Gαt | GTP | D337A | 80 ns |
| Gαt | GTP | E241A | 80 ns |
| Gαt | GTP | R201A | 80 ns |
| Gαt | GTP | E232A | 80 ns |
| Gαt | GTP | D234A | 80 ns |
| EF-Tu | GTP | WT | 80 ns |
| EF-Tu | GDP | WT | 80 ns |
| EF-Tu | GTP | I93A | 80 ns |
| EF-Tu | GTP | V126A | 80 ns |
| EF-Tu | GTP | R75A | 80 ns |
| EF-Tu | GTP | D207A | 80 ns |
| EF-Tu | GTP | Q115A | 80 ns |
| EF-Tu | GTP | D234A | 80 ns |
| | | | total: 1920 ns |

## References


Abankwa D, Hanzal-Bayer M, Ariotti N, Plowman SJ, Gorfe AA, Parton RG, McCammon JA, Hancock JF. 2008. A novel switch region regulates H-ras membrane orientation and signal output. *EMBO J* **27**: 727–735.

Alexander NS, Preininger AM, Kaya AI, Stein RA, Hamm HE, Meiler J. 2014. Energetic analysis of the rhodopsin-G-protein complex links the α5 helix to GDP release. *Nat Struct Mol Biol* **21**: 56–63.

Araki M, Shima F, Yoshikawa Y, Muraoka S, Ijiri Y, Nagahara Y, Shirono T, Kataoka T, Tamura A. 2011. Solution structure of the state 1 conformer of GTP-bound H-Ras protein and distinct dynamic properties between the state 1 and state 2 conformers. *J Biol Chem* **286**: 39644–39653.

Bourne HR, Sanders DA, McCormick F. 1991. The GTPase superfamily: conserved structure and molecular mechanism. *Nature* **349**: 117–127.

Buhrman G, Holzapfel G, Fetics S, Mattos C. 2010. Allosteric modulation of Ras positions Q61 for a direct role in catalysis. *Proc Natl Acad Sci U S A* **107**: 4931–4936.

Cherfils J, Zeghouf M. 2013. Regulation of small GTPases by GEFs, GAPs, and GDIs. *Physiol Rev* **93**: 269–309.

Chung KY, Rasmussen SGF, Liu T, Li S, DeVree BT, Chae PS, Calinski D, Kobilka BK, Woods VL Jr, Sunahara RK. 2011. Conformational changes in the G protein Gs induced by the β2 adrenergic receptor. *Nature* **477**: 611–615.

Dror RO, Mildorf TJ, Hilger D, Manglik A, Borhani DW, Arlow DH, Philippsen A, Villanueva N, Yang Z, Lerch MT, et al. 2015. SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* **348**: 1361–1365.

Flock T, Ravarani CNJ, Sun D, Venkatakrishnan AJ, Kayikci M, Tate CG, Veprintsev DB, Babu MM. 2015. Universal allosteric mechanism for Gα activation by GPCRs. *Nature* **524**: 173–179.

Gerstein M, Altman RB. 1995. Average core structures and variability measures for protein families: application to the immunoglobulins. *J Mol Biol* **251**: 161–175.

Geyer M, Schweins T, Herrmann C, Prisner T, Wittinghofer A, Kalbitzer HR. 1996. Conformational transitions in p21ras and in its complexes with the effector protein Raf-RBD and the GTPase activating protein GAP. *Biochemistry* **35**: 10308–10320.

Girvan M, Newman MEJ. 2002. Community structure in social and biological networks. *Proc Natl Acad Sci U S A* **99**: 7821–7826.

Gorfe AA, Grant BJ, McCammon JA. 2008. Mapping the nucleotide and isoform-dependent

structural and dynamical features of Ras proteins. *Structure* **16**: 885–896.

Grant BJ, Gorfe AA, McCammon JA. 2009. Ras conformational switching: simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. *PLoS Comput Biol* **5**: e1000325.

Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. 2006. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22**: 2695–2696.

Hollinger S, Hepler JR. 2002. Cellular regulation of RGS proteins: modulators and integrators of G protein signaling. *Pharmacol Rev* **54**: 527–559.

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**: 712–725.

Ichiye T, Karplus M. 1991. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* **11**: 205–217.

Iiri T, Bell SM, Baranski TJ, Fujita T, Bourne HR. 1999. A Gsalpha mutant designed to inhibit receptor signaling through Gs. *Proceedings of the National Academy of Sciences* **96**: 499–504.

Iiri T, Farfel Z, Bourne HR. 1997. Conditional activation defect of a human Gsalpha mutant. *Proc Natl Acad Sci U S A* **94**: 5656–5661.

Jackson RJ, Hellen CUT, Pestova TV. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11**: 113–127.

Kaya AI, Lokits AD, Gilbert JA, Iverson TM, Meiler J, Hamm HE. 2014. A conserved phenylalanine as a relay between the α5 helix and the GDP binding region of heterotrimeric Gi protein α subunit. *J Biol Chem* **289**: 24475–24487.

Leipe DD, Wolf YI, Koonin EV, Aravind L. 2002. Classification and evolution of P-loop GTPases and related ATPases. *J Mol Biol* **317**: 41–72.

Marin EP, Krishna AG, Sakmar TP. 2001. Rapid activation of transducin by mutations distant from the nucleotide-binding site: evidence for a mechanistic model of receptor-catalyzed nucleotide exchange by G proteins. *J Biol Chem* **276**: 27400–27405.

Meagher KL, Redman LT, Carlson HA. 2003. Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem* **24**: 1016–1025.

Muraoka S, Shima F, Araki M, Inoue T, Yoshimoto A, Ijiri Y, Seki N, Tamura A, Kumasaka T, Yamamoto M, et al. 2012. Crystal structures of the state 1 conformations of the GTP-bound H-Ras protein and its oncogenic G12V and Q61L mutants. *FEBS Lett* **586**: 1715–1718.

Nissen P, Kjeldgaard M, r. Thirup S, Polekhina G, Reshetnikova L, Clark BFC, Nyborg J. 1995. Crystal Structure of the Ternary Complex of Phe-tRNAPhe, EF-Tu, and a GTP Analog. *Science* **270**: 1464–1472. http://dx.doi.org/10.1126/science.270.5241.1464.

Oldham WM, Van Eps N, Preininger AM, Hubbell WL, Hamm HE. 2006. Mechanism of the receptor-catalyzed activation of heterotrimeric G proteins. *Nat Struct Mol Biol* **13**: 772–777.

Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. 2011. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J Chem Theory Comput* **7**: 525–537.

Polekhina G, Thirup S, Kjeldgaard M, Nissen P, Lippmann C, Nyborg J. 1996. Helix unwinding in the effector region of elongation factor EF-Tu-GDP. *Structure* **4**: 1141–1151.

Raimondi F, Orozco M, Fanelli F. 2010. Deciphering the deformation modes associated with function retention and specialization in members of the Ras superfamily. *Structure* **18**: 402–414.

Raimondi F, Portella G, Orozco M, Fanelli F. 2011. Nucleotide binding switches the information flow in ras GTPases. *PLoS Comput Biol* **7**: e1001098.

Rasmussen SGF, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, et al. 2011. Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **477**: 549–555.

Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, et al. 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* **45**: D271–D281.

Ross EM, Wilkie TM. 2000. GTPase-activating proteins for heterotrimeric G proteins: regulators of G protein signaling (RGS) and RGS-like proteins. *Annu Rev Biochem* **69**: 795–827.

Scheffzek K, Ahmadian MR. 2005. GTPase activating proteins: structural and functional insights 18 years after discovery. *Cell Mol Life Sci* **62**: 3014–3038.

Sethi A, Eargle J, Black AA, Luthey-Schulten Z. 2009. Dynamical networks in tRNA:protein complexes. *Proc Natl Acad Sci U S A* **106**: 6620–6625.

Simon M, Strathmann M, Gautam N. 1991. Diversity of G proteins in signal transduction. *Science* **252**: 802–808. http://dx.doi.org/10.1126/science.1902986.

Skjærven L, Yao X-Q, Scarabelli G, Grant BJ. 2014. Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics* **15**: 399.

Sprang SR. 1997. G protein mechanisms: insights from structural analysis. *Annu Rev Biochem* **66**: 639–678.

Sun D, Flock T, Deupi X, Maeda S, Matkovic M, Mendieta S, Mayer D, Dawson R, Schertler GFX, Madan Babu M, et al. 2015. Probing Gαi1 protein activation at single-amino acid resolution. *Nat Struct Mol Biol* **22**: 686–694.

Takai Y, Sasaki T, Matozaki T. 2001. Small GTP-binding proteins. *Physiol Rev* **81**: 153–208.

Vale RD. 1996. Switches, latches, and amplifiers: common themes of G proteins and molecular motors. *J Cell Biol* **135**: 291–302.

Vetter IR, Wittinghofer A. 2001. The guanine nucleotide-binding switch in three dimensions. *Science* **294**: 1299–1304.

Yao X-Q, Grant BJ. 2013. Domain-opening and dynamic coupling in the α-subunit of heterotrimeric G proteins. *Biophys J* **105**: L08–10.

Yao X-Q, Malik RU, Griggs NW, Skjærven L, Traynor JR, Sivaramakrishnan S, Grant BJ. 2016. Dynamic Coupling and Allosteric Networks in the α Subunit of Heterotrimeric G Proteins. *J Biol Chem* **291**: 4742–4753.

# CHAPTER III

## Transfer Learning Improves The State of The Art for Protein Abundance Prediction in Cancers


### Abstract

The mechanism by which information is translated from transcriptome to proteome and ultimately to phenotype has long been an intriguing problem. The observed baseline Pearson's correlation between mRNA and protein levels across cancer samples is low (corr=0.40). Here we report a method for predicting proteome from transcriptome. First, we establish a generic model capturing the correlation between mRNA and protein abundance of a single gene. Second, we build a gene-specific model capturing the inter-dependencies among multiple genes in a regulatory network. Third, we create a cross-tissue model by transfer learning the information of shared regulatory networks and pathways across cancer tissues. This method ranked first in the 2017 NCI-CPTAC DREAM Proteogenomics Challenge, which is a benchmark platform to unbiasedly evaluate prediction accuracy of proteome based on genomic and transcriptomic data in breast and ovarian cancer patients. The performance of our method (corr=0.53) on the held-out test dataset is approaching the accuracy of experimental replicates (corr=0.59). Key functional pathways and network modules controlling the proteomic abundance in cancers were revealed.


---

The work presented in Chapter III will be submitted as Li H, et.al. "Transfer Learning Improves The State of The Art for Protein Abundance Prediction in Cancers".

## Significance Statement

Understanding the controllers of protein abundance is important for understanding the mechanisms driving the phenotypic differences across individuals. The proteomic data are invaluable sources of information to understanding the regulation of gene expression. However, it requires considerable time and effort to measure proteome experimentally. Here we present a novel method for predicting protein abundance from mRNA levels by transfer learning the shared gene regulatory network between breast and ovarian cancers. The performance of our method approaches the accuracy of experimental replicates.

## Introduction

The central dogma of information flow from DNA to mRNA to protein has been applied for nearly six decades (Crick 1958). Yet, the cell functions as a whole: besides the translation from mRNA to protein, many other features are important to the complex protein expression process, including microRNA (Lovett and Rogers 1996a), upstream open reading frame (Lovett and Rogers 1996b), cap-binding proteins (Raczynska et al. 2010), poly(A) tails (Guhaniyogi and Brewer 2001), nonsense-mediated decay (Chang et al. 2007) or alternative splicing (Black 2003). In addition, the mRNA and protein abundances are dynamic, due to ubiquitination and other degradation mechanisms to fulfill diverse condition-dependent functional requirements (Liu et al. 2016). These complicated regulatory mechanisms underlying protein translation lead to the weak correlations between mRNA and protein abundances, when evaluating the same gene across multiple samples (Liu et al. 2016; Vogel and Marcotte 2012; Ning et al. 2012; Zhang et al. 2014, 2016; Mertins et al. 2016). Identifying the missing factors affecting transcriptomic and proteomic correlation is important to understanding the biological mechanisms behind phenotypic variances and diseases.

This is particularly true in cancers. Transcriptomic and proteomic variations across individuals are expected in diverse cancers, such as colorectal, breast, and ovarian cancers (Mertins et al. 2016; Zhang et al. 2016, 2014). These variations have important clinical consequences and implications, due to activation of different functional pathways, leading to different subtypes in the same organ, and biomarkers indicative of high- and low-risk patients in survival analysis (Zhang et al. 2014; Mertins et al. 2016; Zhang et al. 2016). These transcriptional and proteomic expression profiles provide invaluable information to studying cancer mechanisms. However, compared with the fast, inexpensive RNA sequencing profiles, large-scale high-quality proteomic data are costlier to obtain, despite remarkable progress. Therefore, a computational model to predict protein abundance from mRNA data could not only help to quickly obtain an estimation of proteomic data, but also, to understand what are the important players in cancers.

The National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Ellis et al. 2013) and The Cancer Genome Atlas (TCGA) provide large datasets of proteomic and transcriptomic data in many cancers, which is an unprecedented source for exploring the regulatory process of protein expression. In 2017, the Dialogue on Reverse Engineering Assessment and Method (DREAM) (Stolovitzky et al. 2007) organized the NCI-CPTAC Proteogenomics Challenge. This challenge provides a systematic benchmark to evaluate computational methods for predicting proteomic profiles in breast and ovarian cancers. Here, we describe the best-performing algorithm in this challenge, and reveal the insights derived. Our approach pinpoints the relative importance of the innate correlations between mRNA and protein levels, and the global direct and indirect interactions across all genes in controlling the expression level of a protein.

Based on the intuition that the regulatory mechanism may be shared across different cancer types, we built a new model that shares parameters across two cancers, and improved prediction performance in both cancers. This reveals a new, unexplored aspect of the regulatory mechanism that is previously not captured in single tissue modelling approaches. Pathway analysis and gene-gene interaction network indicate that functionally different gene sets have different predictability profiles and regulatory powers. In sum, our approach offers a new field standard for protein abundance prediction across cancer patients, and the key features used in our model and the innovation of transfer learning across two cancer types will be instructive for future method development and protein expression regulatory mechanism exploration.

## Results

### Overview of the experimental design for protein abundance prediction

In this study, we use a training dataset provided by NCI-CPTAC, which consists of the transcriptome and proteome data from 77 breast and 105 ovarian cancer samples. To unbiasedly evaluate prediction methods, a docker image system was used in the NCI-CPTAC DREAM challenge for participants to submit their code and score on a held-out testing dataset of the proteomic data from 108 breast and 82 ovarian cancer samples (**Figure 3.1** top-center). For each protein, the primary evaluation metric was the Pearson's correlation between predictions and observations across samples. The final score was calculated by averaging the prediction correlations of all proteins under consideration. In addition, the Normalized Root Mean Square Error (NRMSE) between predictions and observations was used as the secondary scoring metrics to evaluate models.

We developed three major components in order to extract informative features and exploit the training data. First, the intrinsic correlation between mRNA and protein levels was considered in the generic model (**Figure 3.1** top-left). Second, for each protein under investigation, we utilized the nonlinear interdependencies among all genes in the gene-specific model (**Figure 3.1** bottom). Third, the model weights were interchangeable between cancer tissues, capturing the shared regulatory mechanism in the trans-tissue model (**Figure 3.1** top-right). By integrating these components, we enhanced the prediction of protein abundance in both breast and ovarian cancers.

**Dissection of critical components in determining protein abundance**

To quantify the relative contributions of features that determine protein abundance, we investigated the performance gain of each component. The average Pearson's correlations of the generic model were 0.37 and 0.40 in breast and ovarian cancer, respectively (**Figure 3.2A** left; **Table S3.1**). By combining the predictions from the gene-specific model, we significantly improved the correlations to 0.40 (breast) and 0.46 (ovary) (**Figure 3.2A** middle; $p<2.2e-16$; see **Materials and Methods**). To consider the similarity across cancer tissues, we further integrated the trans-tissue model and achieved the highest correlations of 0.41 (breast) and 0.47 (ovary) (**Figure 3.2A** right; $p<2.2e-16$; see **Materials and Methods**). In addition, the RMSEs of these components were also calculated (**Figure S3.1A**).

When we built the gene-specific model, a key question was how many genes should be used as features for predicting protein abundance. As we expected, as the number of features increased (the top 10, 100, or 1,000 expressed genes), the predictive performances consistently improved in terms of both correlation (**Figure S3.2**) and RMSE (**Figure S3.3**). Interestingly, filtering feature

genes based on prior knowledge of Gene Ontology (GO) (Ashburner et al. 2000; The Gene Ontology Consortium 2017) related to 'translation' and 'gene expression' did not improve the performance, whereas using all genes as features achieved the highest correlations ("GO-features" and "All-features" in **Figure 3.2B**) and lowest RMSEs (**Figure S3.1B**). These results indicate that the abundance of a single protein is regulated by the commonly existing gene-gene associations; the regulatory contributions are not from a small set of genes but universally distributed among all genes.

To further investigate contributions of these three models, we performed the grid-search of various weighting ratios of them. We observed similar "dark" right arms of the ternary plots in both breast and ovarian cancers (**Figure 3.2C**,**D**), where the correlations were relatively low. This is because the gene-specific and trans-tissue models captured non-redundant regulatory information, compared with the generic model. When integrating different types of models, we significantly improved the correlations, leading to the sudden color change moving from the right arms towards the left-bottom generic model. Furthermore, when moving along the right arms towards the trans-tissue model, the correlation gradually increased (the color becomes brighter), since the trans-tissue model contributed more to the final prediction. The best combination ratios of the generic, gene-specific and trans-tissue models were 2:3:5 in breast and 1:4:5 in ovary, where the trans-tissue model had the largest weights in both cancers (golden stars in **Figure 3.2C,D**).

**Regulatory information of protein abundance is transferable between breast and ovarian cancers**

Regulatory pathways are expected to be shared to certain extent across different tissues, which motivates us to develop a model that shares the weights between tissues. To investigate the effect of transferring information between cancer tissues, we trained a "Combined-samples" model by combining samples from these two cancers, and directly compared it with the model training on one cancer only. The "Combined-samples" model largely increased the prediction correlation from 0.27 to 0.32 in breast and from 0.36 to 0.49 in ovary ("All-features" and "Combined-samples" in **Figure 3.2B**). In fact, the performance was highly dependent on the number of training samples. When we used 40%, 60%, 80% or 100% of the samples to train the model, the performances gradually increased in terms of both correlation (**Figure S3.4**) and RMSE (**Figure S3.5**). These results demonstrate that current prediction performance is limited by the relatively small sample size. Therefore, we combined samples from the two types of cancers and trained the trans-tissue model, assuming that the same protein is regulated in a similar fashion in these two cancers. As we expected, the trans-tissue model achieved higher correlations since it was trained on more samples.

In addition to the transcriptomic data, we also investigated other types of data that could potentially contribute to the prediction of protein abundance (**Figure 3.3**). We first considered DNA copy number variation (CNV) as the approximation for proteome. Compared with RNA, CNV provided much less information and the prediction correlation of CNV itself was only 0.2 in both breast and ovarian cancers (RNA and CNV in **Figure 3.3A,B**). We next used the RNA and CNV values of a gene as features and trained a random forest model on all available proteins, yet the performance was worse than RNA itself. Nevertheless, the cross-tissue models either trained on separated or

combined data improves the correlation ("RF", "RF+cross1", and "RF+cross2" in **Figure 3.3A,B**). These results indicate that the RNA level itself is already a good approximation for the protein abundance, better than CNV or the simple model trained on RNA and CNV. Therefore, the CNV data was not used in our final model. To reduce the potential batch effects across individuals, different normalization methods were also tested (**Figure S3.6**).

We further explored the effects of adding features of protein sequence and class. For each amino acid, we counted the number of occurrences in a protein sequence as an extra feature, improving the correlations in both cancers ("RF+aa" and "RF+aaKR" in **Figure 3.3C,D**). Similarly, we considered the protein classes defined by CATH protein structure classification database as extra features, improving the performance ("RF+class" and "RF+aaKR+class" in **Figure 3.3C,D**). However, when assembling models using these features into the final model, we didn't observe any improvement. Therefore, these features were not used in our final model.

**Transfer learning approaches experimental replicate level accuracy**

Since proteomics data have intrinsic noises due to batch effects and fluctuations, we further estimated the theoretical best performance based on the experimental replicates for the overlapping samples measured at two different cohorts. To be specific, there are 32 ovarian cancer samples measured at both JHU and PNNL. For these samples, we calculated the Pearson's correlation (0.59) and RMSE (0.179) between the experimental replicates at two cohorts. Meanwhile, the prediction correlation and RMSE of our method on the held-out testing dataset during the NCI-CPTAC DREAM challenge were 0.53 (**Figure S3.7**) and 0.186 (**Figure S3.8**), respectively. These results indicate that the protein abundance prediction is a relatively hard task, due to the intrinsic noises

of the measurements across cancer samples. Although our method only achieved a medium prediction correlation of 0.53, it is in fact close to the correlation of 0.59 between experimental replicates. In terms of RMSE, our method is even closer to the accuracy of experimental replicates and the error is only 3.9% higher, which is calculated from (0.186 - 0.179) / 0.179 = 3.9%. Currently, our method was built on 77 breast and 105 ovarian cancer samples by transfer learning. We foresee that this method would become even closer to the performance of experimental replicates with more training samples, since we have observed the gradually increased performance as the training set becomes larger (**Figure S3.4**).

**Functionally diverse gene sets display different predictability spectrums**

To investigate the relationship between protein functions and ease of predictability, we performed functional enrichment analysis of all considered proteins. We found that gene sets of different predictability were functionally enriched in different Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto 2000). The overall distributions of correlations between our predictions and observations for breast and ovarian cancers are shown in **Figure 3.4 A** and **B**, respectively. Based on the predictability, we partitioned the proteins into four groups: the top 0%-25% proteins that are easy to predict, the 25%-50% and 50%-75% proteins that are medium to predict, and the bottom 75%-100% proteins that are hard to predict. For each group, the functional enrichment analysis was performed against KEGG pathways, and significantly enriched functional pathways were shown in **Figure 3.4**. In the breast cancer, the gene group easy to predict was highly associated with the "Metabolism" category, including pathways of amino acids and other biomolecules metabolism (red genes in **Figure 3.4C**). In contrast, the genes hard to predict were usually associated with the "Genetic Information Processing" and "Human Disease" categories,

including pathways of ribosome, spliceosome, proteasome and three neurodegenerative diseases (blue and purple genes in **Figure 3.4C**, respectively). Interestingly, it has been reported that cancers and neurodegenerative disease share common mechanisms of molecular abnormalities (Du and Pertsemlidis 2011; Spencer et al. 2012). In particular, microRNA (miRNA)-based regulation of mRNA translation is a potential common regulator of both cancer and neurodegenerative disease (Cooper et al. 2009). Mutations in genes associated with cell cycle regulation, protein turnover and DNA repair have been implicated in these two type of diseases (Morris et al. 2010). We observed similar distribution of functionally different gene sets in ovarian cancers (**Figure 3.4D**). These results are consistent with the previous observations that stable and housekeeping proteins usually have weak mRNA-protein correlations, whereas dynamic proteins tend to have strong correlations (Zhang et al. 2014; Mertins et al. 2016; Zhang et al. 2016).

To further understand the regulatory patterns of different genes, we performed similar functional enrichment analysis on genes ranked by the prediction improvement after integrating the gene-gene interdependencies of the gene-specific model. We found that in general the housekeeping proteins, associated with RNA transport, ribosome, spliceosome and proteasome, benefited more than the metabolism-related genes in both cancers (**Figure S3.9**). In addition, several disease-related gene sets gained relatively large improvements in the ovarian cancer, including Parkinson's, Alzheimer's and Huntington's diseases. In sum, we find similar mapping landscapes between protein abundance prediction improvement and functional pathways in breast and ovarian cancers.

**Metabolism-related genes are essential in regulating the protein abundance**

Metabolism-related gene sets make major contributions to predicting protein abundance. To evaluate the feature importance of a gene, the mRNA values of each gene across samples were permuted and the prediction performance was re-evaluated. Permutation of more important genes resulted in larger drops in performances, which were considered as the feature importance. Based on the importance, we ranked all genes and performed the functional enrichment analysis on the important "driver" genes. We found that genes of the KEGG "Metabolism" category played an essential role (**Figure 3.5**). As we expected, among pathways of carbon metabolism, biosynthesis of amino acids was more critical in determining the protein abundance.

To further investigate these "driver" genes, we mapped them to a gene functional network (Li et al. 2015, 2016; Guan et al. 2008). This network was constructed based on a Bayesian integration of diverse genetic and functional genomic data. We hypothesize that these "driver" genes form a nexus module which dictates certain core functions in the cell. We extracted a subnetwork that contained only the driver genes as well as edges that had high estimated probability of the co-functioning relationship (**Figure 3.6**). The high-confidence connections encompassed 674 "driver" and "target" genes in ovarian cancer, and 568 in breast cancer. Then, we applied the Girvan-Newman community clustering algorithm to the subnetwork. The algorithm iteratively identifies and cuts the sparse connections that connect different modules to maximize a modularity score (Newman and Girvan 2004; Newman 2006).

The resulting clusters are a collection of gene modules that are highly connected within the cluster but loosely connected to other genes. The GO term enrichment analysis was further performed on

the resulting modules. The important enriched pathways fell into a number of naturally forming groups. Specifically, the processes of gene expression, protein metabolics, transcription initiation and regulation were enriched. The initiation of protein translation is known to be the bottleneck step of the protein synthesis (Guimaraes et al. 2014). The pathways of cell cycle regulation and DNA/RNA modification were also prominently featured. Additionally, the immune response, signal transduction, response to wounds, and morphological development were all enriched. Interestingly, it has been reported that cellular stress responses and the wound healing are related to cancer treatment resistance and metastasis (Chircop and Speidel 2014; Arnold et al. 2015; Sundaram et al. 2017). The results confirmed our expectation that the nexus modules formed by these genes are loosely but confidently associated with other genes. The translation level of a protein is controlled by a complex network consisting of diverse regulatory elements in the cells.

**Discussion**

From the central dogma to the complex protein functional networks and pathways, our understanding of protein expression regulation has been revolutionized over the past 60 years. Although macromolecular interactions require specific physicochemical interfaces (Liddington 2004), indirect interactions and high-level associations exist in cellular environment. In terms of predicting protein abundance from transcriptomic data, these ubiquitous associations among all genes play an indispensable role. This indicates that in addition to the idea of functional pathways and protein-protein interaction networks, considering the general direct and indirect interactions among all genes is a complement towards understanding the underlying mechanisms.

Many pioneering efforts have been made to characterize the proteogenomic features of various cancers (Zhang et al. 2014; Mertins et al. 2016; Zhang et al. 2016; Robertson et al. 2017). However, how to integrate information from multiple cancers to foster cancer research remains unclear. In this study, we propose a simple yet effective attempt to address this problem, facilitating the prediction of protein abundance. It would be interesting to see where the information is shareable among diverse cancers or other diseases, beyond breast and ovarian cancers. Intriguingly, we observe that protein subsets that are hard to predict are enriched in several neural degenerative diseases.

## Materials and Methods

### Data collection

For both breast and ovarian cancers, the proteome data were acquired using the isobaric Tags for Relative and Absolute Quantification protein quantification method. The proteomics data were downloaded from CPTAC data portal. For breast proteome, 77 samples were analyzed at the Broad Institute (BI). For ovarian proteome, 84 and 122 samples were analyzed at Pacific Northwest National Laboratory (PNNL) and Johns Hopkins University (JHU), respectively. The protein log ratios of the protein abundance were calculated including only peptides that map unambiguously to the protein. Among the 84+122=206 samples, only 105 samples had the corresponding TCGA RNA-seq data. Since we needed both the RNA-seq (as features) and the proteomic data (as labels) to build machine learning models, only these 105 samples were used in this work. The transcriptomics data for the corresponding breast and ovarian cancer samples were downloaded from TCGA firehose.

**Generic model**

For each gene i, the mRNA levels across patients were used as the baseline predictions for the corresponding protein abundance across the same patients (top-left in **Figure 3.1**). If the mRNA values were missing, we used the average of all non-missing RNA observations of the same gene as the imputation:

$$x_{missing} = (\sum_{i=0}^{n_{non-missing}} x_i)/n_{non-missing}$$

where $x_i$ represents the mRNA level of a non-missing sample and $n_{non\text{-}missing}$ represents the number of non-missing samples.

**Gene-specific model**

The entire RNA-seq data is represented by a m-by-n matrix X,

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & \cdots & x_{mn} \end{bmatrix}$$

where rows represent genes and columns represent samples. An element $x_{ij}$ denotes the mRNA level of gene i from sample j. Similar to mRNA, the proteomic data is represented by a s-by-n matrix Y,

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{s1} & \cdots & \cdots & y_{sn} \end{bmatrix}$$

where rows represent proteins and columns represent samples. For each gene, we created a gene-specific random forest (RF) model (Breiman 2001), with maximum depth of 3 and 100 trees

(bottom in **Figure 3.1**). As one of the tree-based models, RF has been reported to avoid overfitting and capture nonlinear interactions between features (Li et al. 2018c, 2018b, 2018a, 2018d). For example, for gene i, we treated the protein levels of this gene across n samples ($y_{i1}$, $y_{i2}$, … , $y_{in}$) as n targets. For each sample $y_{ik}$, we use its corresponding mRNA levels of all m genes ($x_{1k}$, $x_{2k}$, … , $x_{mk}$) as a vector of m features . In this way, we trained a model using n samples. And for a different gene j, we created a different model since the target values across n samples ($y_{j1}$, $y_{j2}$, … , $y_{jn}$) are different. Thus, we call this a gene-specific model. After excluding genes with missing mRNA values, the total numbers of feature genes are 8,738 and 5,837 in breast and ovarian cancers, respectively. These models were implemented using the function called ensemble.RandomForestRegressor of python module scikit learn.

**Trans-tissue model**

The numbers of proteins to be predicted are 10,006 and 7,061 in breast and ovarian cancers, respectively. Among them, 6934 proteins are common in the two cancers. To pool regulatory information between two cancers, we combined the patient samples for each common protein and trained the trans-tissue random forest model in the same way as the gene-specific model (top-right in **Figure 1**). The total number of training samples is 182 (77 breast and 105 ovarian).

**Statistical analysis**

To compare the prediction correlations among different models, the bootstrap sampling with replacement was performed. Specifically, 5,000 genes were randomly selected to calculate the overall prediction correlation of a model in each bootstrap sample. The sampling was performed 1,000 times for each model, followed by the Wilcoxon signed-rank test to compare two models.

The differences between all pairs of models in **Figure 3.2A-B** were statistically significant ($p <$ 2.2e-16). The *p*-values were calculated using the default function wilcox.test in R version 3.4.4.

**Five-fold cross validation**

To systematically compare the performance of different models and features, five-fold cross validation was performed on the training data of 77 breast and 105 ovarian cancer samples. For each cancer, the entire training samples were randomly partitioned into 5 non-overlapping subsets. In each validation, 4 subsets were used to train a model and 1 subset was used to validate the performance of this model. This resulted in 5 scores, reflecting the overall performance of a model on the entire dataset.

**Comparing models using different numbers of features**

To evaluate the effects of using different number of features, the top 10, 100, and 1000 highly expressed genes, and all genes (8738 breast genes and 5837 ovarian genes) were used to train the gene-specific models. We further evaluated the filtered gene subset based on GO terminology (GO 0010467: gene expression and GO 0010468: regulation of gene expression), resulting in 4472 and 4473 feature genes in the GO breast and ovarian cancer models.

**Comparing models trained on different numbers of samples**

To evaluate the effects of training different numbers of samples, 20%, 40%, 80% and 100% of training samples were randomly selected to train the gene-specific model. Then the samples from the breast and ovarian cancers were combined and trained the trans-tissue model.

**Model ensemble**

For each protein, the weighted average predictions from the generic and the gene-specific models were calculated, with the weighting ratio of 1:3. For the 6934 common proteins, the predictions from the trans-tissue model were added, with the weighting ratio of 1:1. It should be noted that, for non-common proteins, the trans-tissue model is not applicable. These weights were used to generate predictions. To evaluate the effect of different weighting ratios, we performed a grid search of all possible weights from 0 to 10 among the generic, gene-specific and trans-tissue models.

**Evaluation metrics**

To evaluate the performance of different models, the Pearson's correlation between observed and predicted abundances across all samples was calculated for each protein. We then took the mean correlations of all proteins as the primary evaluation score. In addition, the Normalized Root Mean Square Error (NRMSE) was used as the secondary metric to compare models.

The formula for computing the Pearson correlation r, is as follow:

$$r = \frac{1}{n_{obs} - 1} \sum_{i=1}^{n_{obs}} \frac{(x_i - \underline{x})(y_i - \underline{y})}{S_x * S_y}$$

The formula for computing NRMSE is as follows:

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n_{obs}} (y_i - x_i)^2 / n_{obs}}}{y_{max} - y_{min}}$$

The observed and predicted values are denoted by y and x, respectively. $S_y$ and $S_x$ are their standard deviations. For each protein, nobs is the number of observed samples. And $y_{max}$ and $y_{min}$ are the respective maximal and minimal value across all observed samples.

**Correlations and RMSEs between experimental replicates**

There were 32 overlapping ovarian cancer samples measured at both JHU and PNNL. These overlapping samples were used to estimate the theoretical best performance that could be achieved by a computational prediction method. The Pearson's correlations and RMSEs for all 5,218 proteins under consideration were calculated across the 32 ovarian cancer samples.

**Feature importance**

Random forest enables us to estimate the importance of each chemical feature by permuting the values of a feature across samples and computing the increase in prediction error, delta-error. More important feature genes have larger delta-error. Based on the delta-error, we evaluate the importance of all feature genes.

**Functional enrichment analysis**

All the evaluated proteins were quantile partitioned into four subsets based on the prediction performance. For each subset, functional annotation was performed using DAVID. We further analyzed the functional enrichment of proteins ranked by the improvement compared with the baseline mRNA and protein levels, and proteins playing important roles in regulating the protein abundance of all genes.

**Functional network analysis**

The top 500 genes with the highest feature importance ("driver" genes) were mapped to a gene functional network. A subset of highly connected genes were selected for the clustering analysis

(674 genes in breast and 568 genes in ovary). These genes, together with edges among these genes, were extracted to a subnetwork. The network was then fed into GLay community clustering method. The clustering method is based on the Girvan-Newman algorithm [23] and implemented in ClusterMaker2, a Cytoscape plugin. The method dissects the original subnetwork into multiple modules. Each of the modules was then fed into BINGO, a Cytoscape plugin, for GO term enrichment analysis.

**Figure preparation**

The figures were prepared using R package ggplot2, ggtern and GGally. The protein structures shown as 3D illustration in Figure 1 were downloaded from Protein Data Bank. Their IDs are 1cr5, 1ctq, 1grn, 1jbb, 1kpc, 1tnd, 1yfp and 1zho. These images were generated by VMD 1.9.3.

**Availability of data and material**

Source code: https://github.com/GuanLab/CPTAC_sub2

Challenge dataset repository: https://www.synapse.org/#!Synapse:syn8228304/wiki/448379

Breast cancer proteomic data: https://cptac-data-portal.georgetown.edu/cptac/s/S029

Ovarian cancer proteomic data: https://cptac-data-portal.georgetown.edu/cptac/s/S026

Transcriptomic data: https://portal.gdc.cancer.gov/

## Figures

**Figure 3.1 Overview of the algorithm design for predicting proteomic expression from transcriptomic data.**
The overall task of this study is to transform the red matrix, representing the transcriptomic level expression across different individuals, to the blue-grey matrix, representing the proteomic level expression (top-center). Three models are created to address this problem: 1. Generic model, which captures the innate correlation between mRNA and protein level (top-left); 2. Gene-specific model, which captures how multiple genes work in a network to control the protein level under investigation through random forest aggregation of multiple base learners (bottom); 3. Trans-tissue model, which captures the shared functional networks across cancer types (top-right).

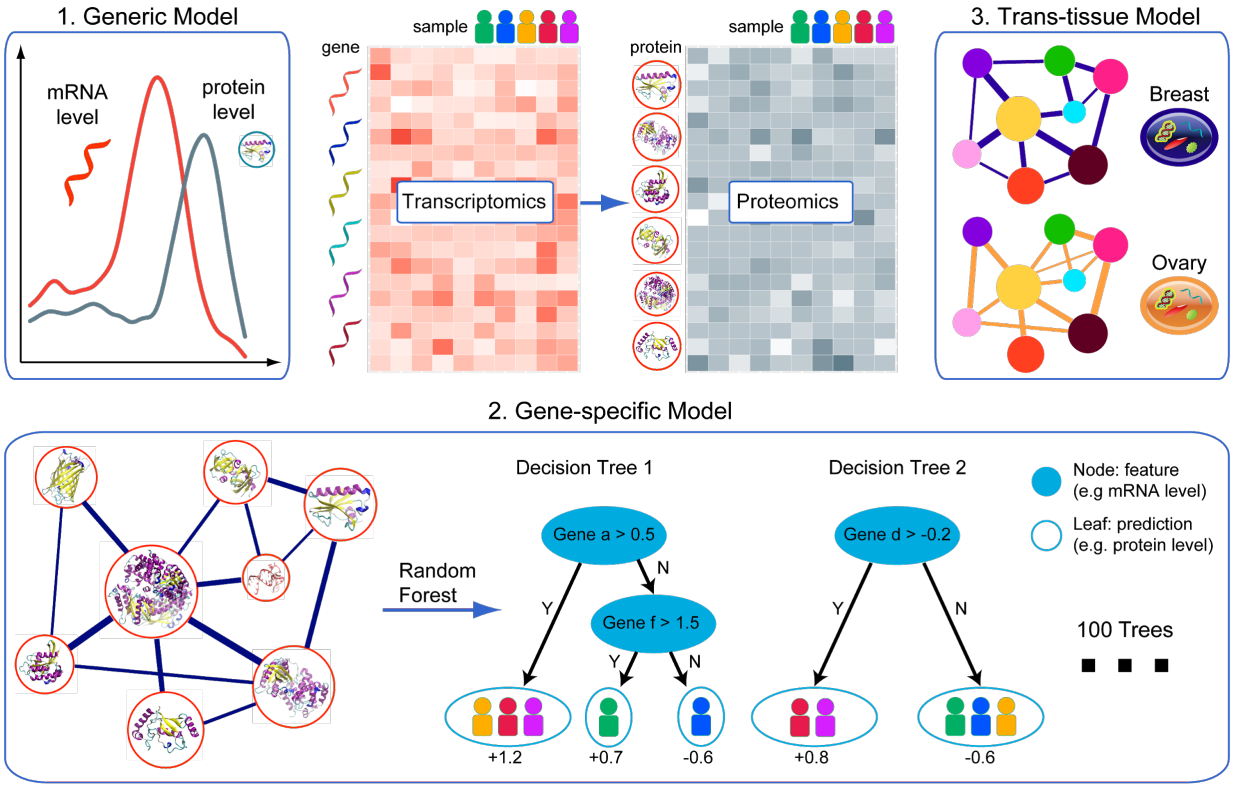


**Figure 3.2 The contributions of different models to predicting proteome in breast and ovarian cancers.**

**A.** From left to right, the correlations were calculated by assembling the following three models step by step (blue: breast; red: ovary): 1) The generic model, which only uses the transcript-level expression of a target protein as the only feature; 2) The gene-specific model, which uses the transcript-level expressions of all genes as features for predicting a target protein; 3) The trans-tissue model, which is similar to the gene-specific model yet combines both breast and ovarian cancer samples. **B.** Dissection of the gene-specific model by using different sets of features and samples. 1） Sub-selecting all genes related to 'gene expression' as features. 2） Using all transcripts as features to predict the target protein. 3) Combining samples from two tissues to train. The correlations between all pairs of models are significantly different ($p<2.2e-16$) using Wilcoxon signed-rank test, after bootstrap sampling for 1,000 times. **C-D.** The contributions of the generic, gene-specific, and trans-tissue models to the final predictions in **(C)** breast and **(D)** ovary. The combination that achieves the highest correlation is labeled by the golden star, where the best combination ratios of the generic, gene-specific and trans-tissue models are 2:3:5 in breast and 1:4:5 in ovary. Notably, the right arms of both triangle are in "darker" color (lower correlations), representing large correlation increases when new models are integrated.

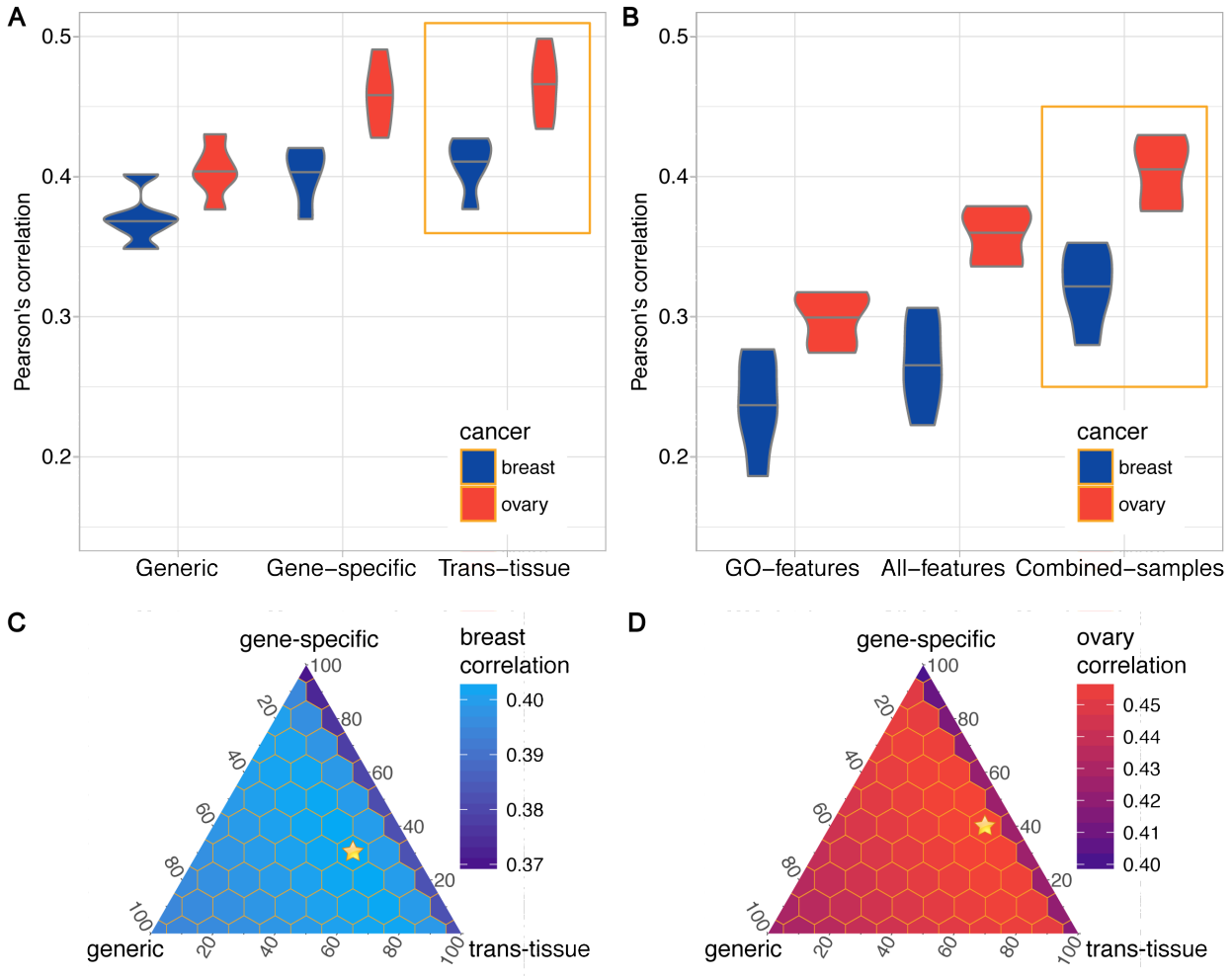


**Figure 3.3 Prediction performance using different input features.**
**A-B.** The Pearson's correlation between predictions and observations across patients in (**A**) breast and (**B**) ovary. The x-axis represents different methods. Specifically, RNA and CNV simply use the mRNA and DNA copy number variation values as approximations for the proteomic values, respectively. RF is the random forest model trained across all available proteins using two features, the corresponding RNA and CNV values of a protein. RF+cross1 and RF+cross2 are the random forest models transferring information cross breast and ovarian cancers. In RF+cross1, we trained two RF models on breast or ovary data separately and assembled the predictions of them, while in RF+cross2, we only trained one RF model on the combined breast and ovary data. **C-D**. The prediction performance using protein sequence and class information in (**C**) breast and (**D**) ovary. In addition to RNA and CNV, in RF+aa we add twenty features, each representing the number of an amino acid in a protein. In RF+aaKR, we add only the numbers of two amino acids, lysine (K) and arginine (R), which are the cleavage targets of trypsin in proteomics mass spectrometry. In RF+class, we add four binary features, representing the four protein classes defined by the CATH protein structure classification database. In RF+aaKR+class, we add features of both the number of amino acids and protein classes.

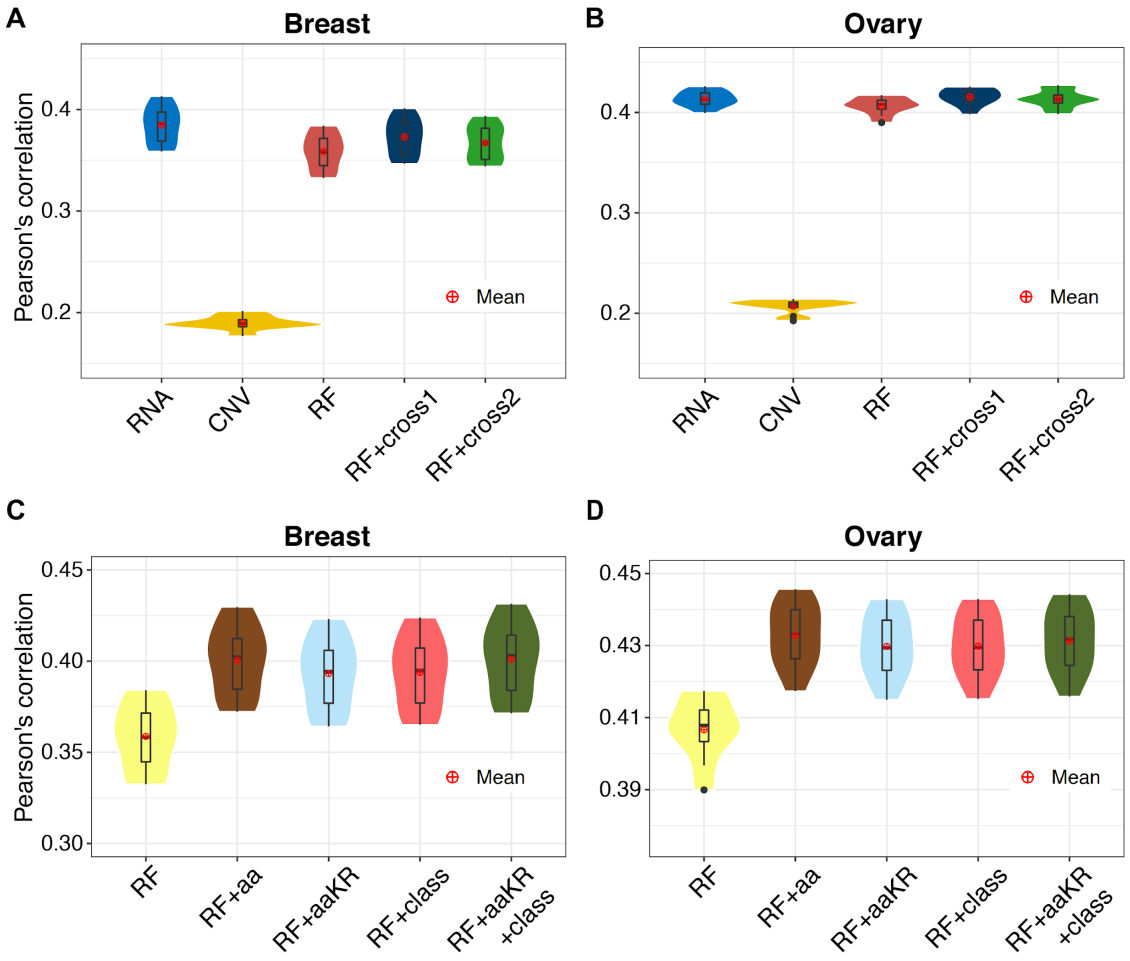


**Figure 3.4 The functional enrichment analysis of gene sets with different predictability spectrums.**

**A-B**. The overall distribution of the Pearson's correlations between observations and our predictions in (**A**) breast and (**B**) ovarian cancers. **C-D**. Based on the predictability, we partitioned the proteins into four groups: the top 0%-25% easiest proteins to predict, the median 25%-50% and 50%-75% predictable group and the bottom 75%-100% hardest proteins to predict. For each group, the functional enrichment analysis was performed against KEGG pathways. The colors represent the major KEGG categories. Genes with high prediction correlations are mainly associated with "Metabolism", whereas genes with low prediction correlations are mainly associated with "Genetic Information Processing" and "Human Diseases".

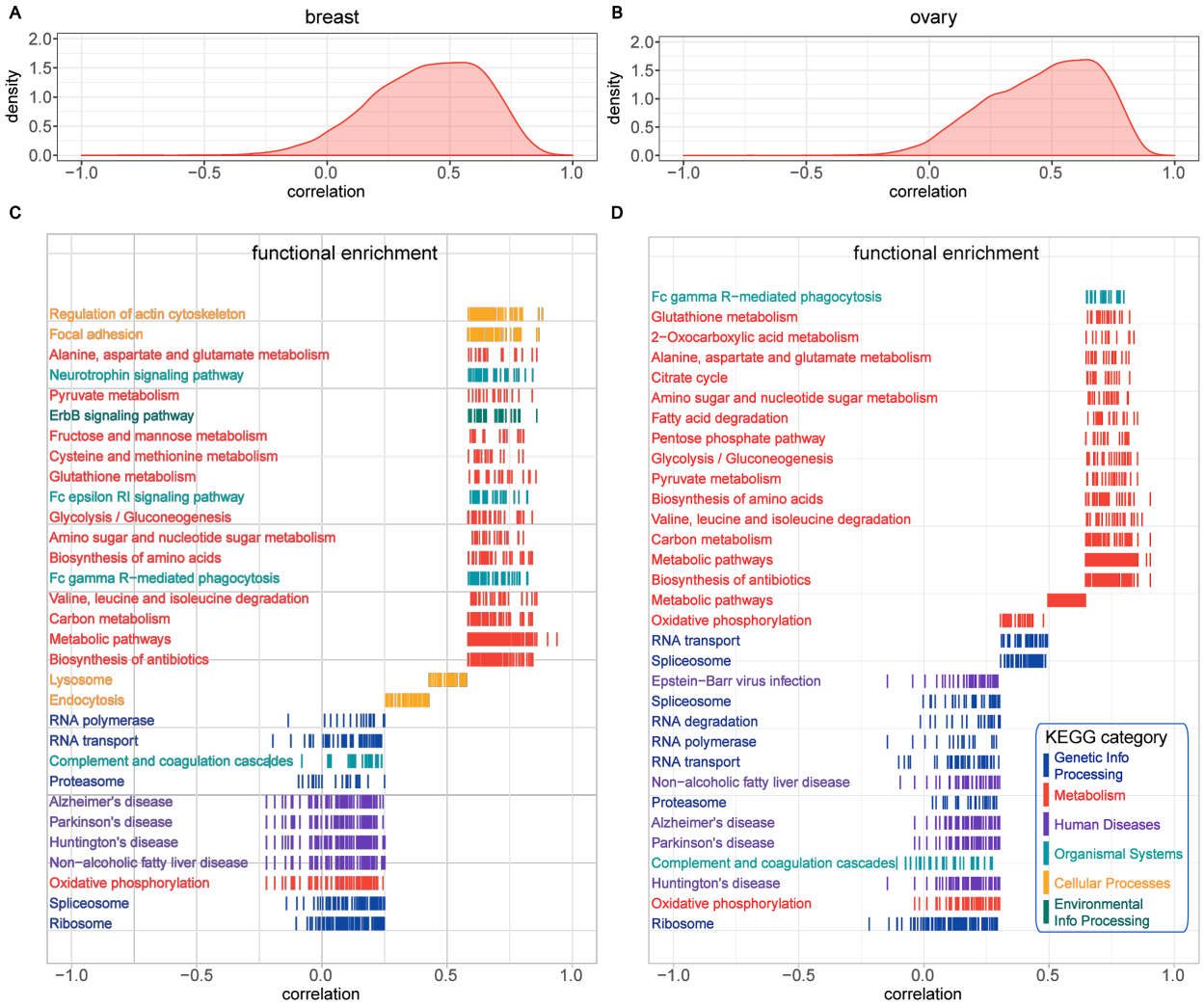


**Figure 3.5 The functional enrichment analysis of gene sets that drive the regulation of protein abundance.**

**A-B**. The overall distribution of the gene importance in predicting protein abundance are shown in (**A**) breast and (**B**) ovarian cancers. **C-D**. Functional enrichment analysis was performed on gene subsets based on the feature importance. The colors represent the major KEGG categories and the x-axis is the feature importance. The genes and pathways on the right have higher feature importance and contribute more to regulating protein abundance. In both breast and ovarian cancers, genes and pathways associated "Metabolism" are the most informative for predicting protein abundance and cross-sample correlations.

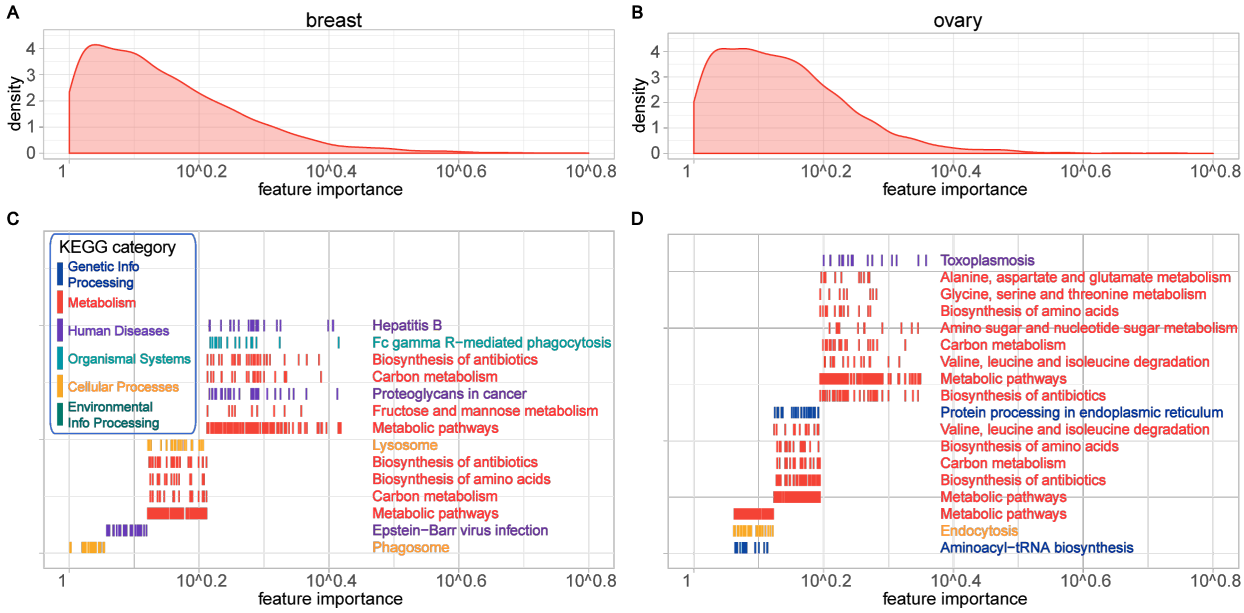
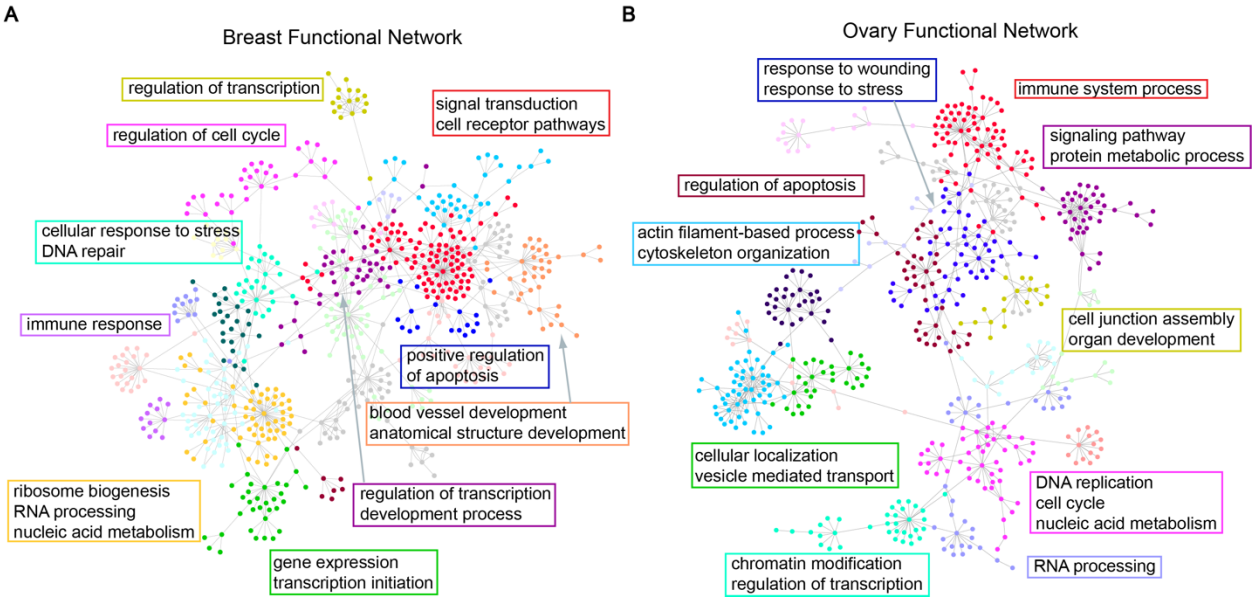


**Figure 3.6. Functional clusters in the gene-gene interaction network that drive the regulation of protein abundance.**

**A-B**. Decomposition of gene functional network among "driver" genes in breast (**A**) and ovarian (**B**) cancers reported important metabolism pathways. The gene clusters were shown in different colors and visualized using a gene-gene interaction network. The shared biological processes of selected clusters were labeled in rectangles.



## Supplementary Figures and Tables

**Figure S3.1 The RMSEs of different models in predicting proteome in breast and ovarian cancers.**
**A.** From left to right, the RMSEs were calculated by assembling the following three models step by step (blue: breast; red: ovary): 1) The generic model, which only uses the transcript-level expression of a target protein as the only feature; 2) The gene-specific model, which uses the transcript-level expressions of all genes as features for predicting a target protein; 3) The trans-tissue model, which is similar to the gene-specific model yet combines both breast and ovarian cancer samples. **B.** Dissection of the gene-specific model by using different sets of features and samples. 1） Sub-selecting all genes related to 'gene expression' as features. 2） Using all transcripts as features to predict the target protein. 3) Combining samples from two tissues to train.

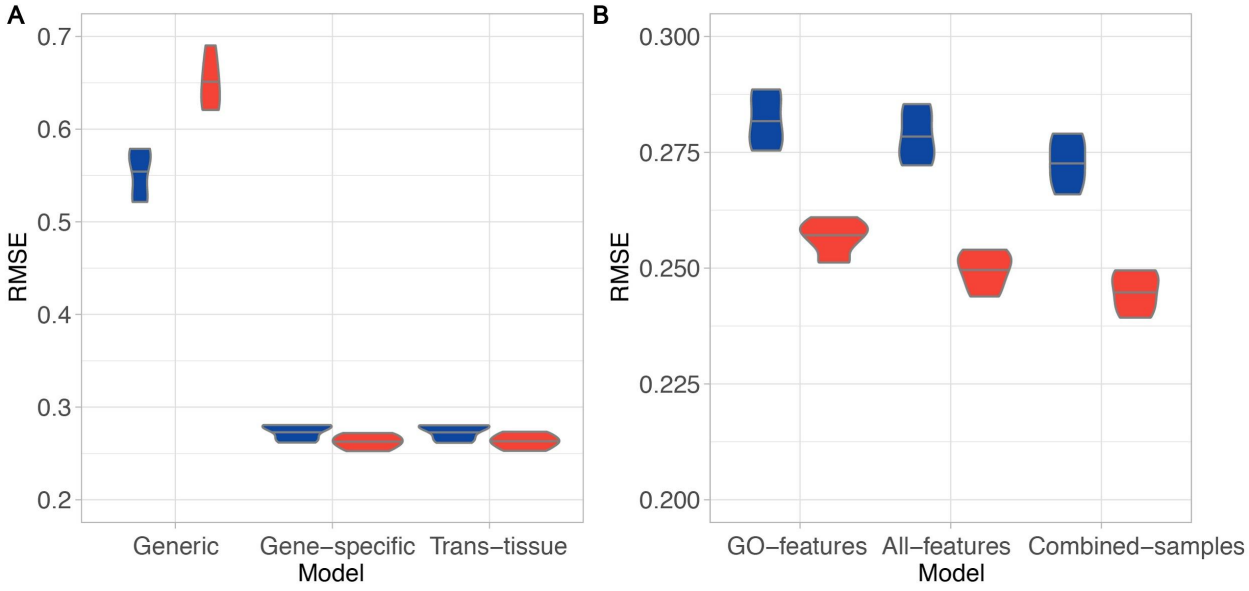


**Figure S3.2 The correlation comparison of models using different number of genes as features.**

The Pearson's correlations of five-fold cross validation results are shown in blue (breast) and red (ovary). From left to right, the number of feature genes used in the gene-specific model increases. The first three numbers represent models using the top 10, 100, and 1,000 expressed genes as features. In addition, the gene subsets associated with GO terms (0010467: gene expression and 0010468: regulation of gene expression) are also evaluated, which contain 4,472 and 4,473 genes in breast and ovary, respectively. Our final gene-specific model uses all genes (8,738 genes in breast and 5,837 genes in ovary) as features and achieves highest correlations (the orange box).

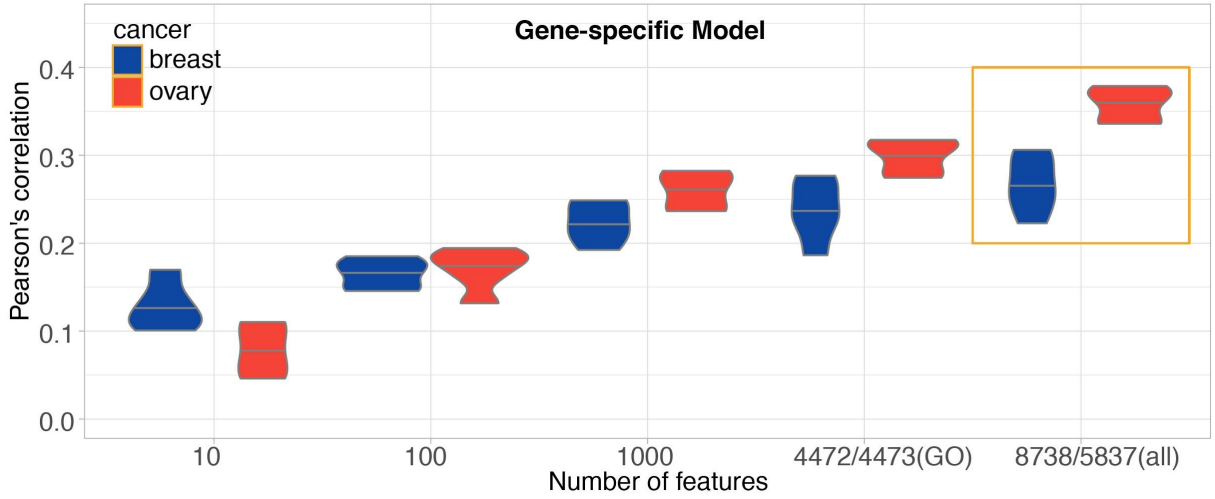


**Figure S3.3 The RMSE comparison of models using different number of genes as features.** The RMSEs of five-fold cross validation results are shown in blue (breast) and red (ovary). From left to right, the number of feature genes used in the gene-specific model increases. The first three numbers represent models using the top 10, 100, and 1,000 expressed genes as features. In addition, the gene subsets associated with GO terms (0010467: gene expression and 0010468: regulation of gene expression) are also evaluated, which contain 4,472 and 4,473 genes in breast and ovary, respectively. Our final gene-specific model uses all genes (8,738 genes in breast and 5,837 genes in ovary) as features and achieves lowest RMSEs (the orange box). With greater number of features, the random forest model can learn the nonlinear interdependencies and the regulatory relationship between more genes. Therefore, models with more genes as features can estimate protein abundance more accurately, increasing the prediction correlation and reducing the prediction RMSE. Of note, the RMSEs in the ovarian cancer are overall lower than those in the breast cancer mainly due to the cohort effect. These two types of cancer samples were collected from two different cohorts (see details in the section "**Materials and Methods – Data collection**").

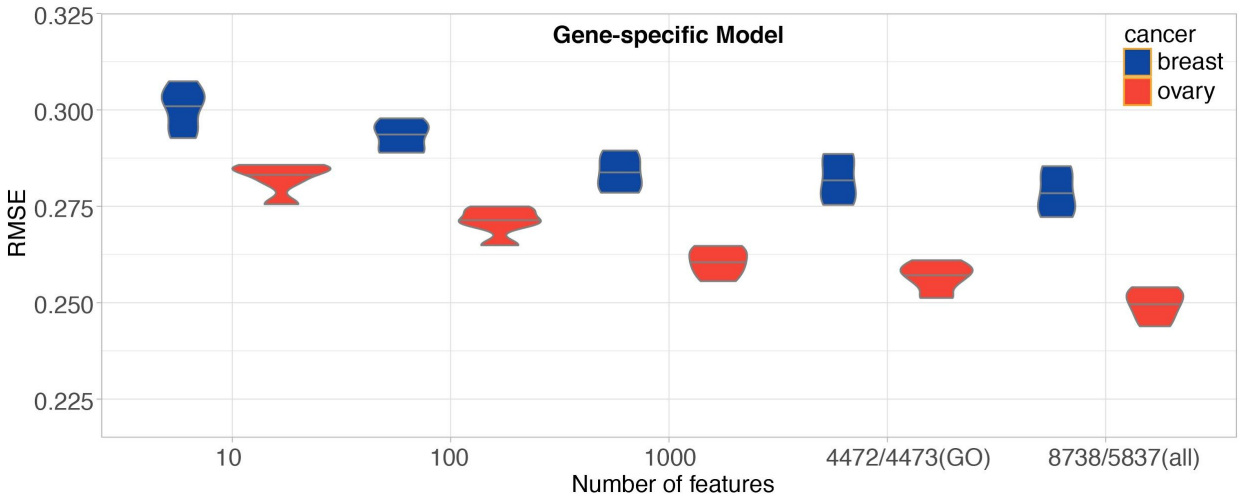


**Figure S3.4 The correlation comparison of models trained on different number of samples.** From left to right, the Pearson's correlations were calculated for models using (1) 40% of the training samples (2) 60% of the training samples (3) 80% of the training samples (4) 100% of the training samples (5) 100% of the combined training samples from two cancer tissues. Of note, the exact number of training samples in (4) is listed in the parentheses (b for breast and o for ovary). Our final trans-tissue model combines samples from two cancers and achieves highest correlations (the orange box).

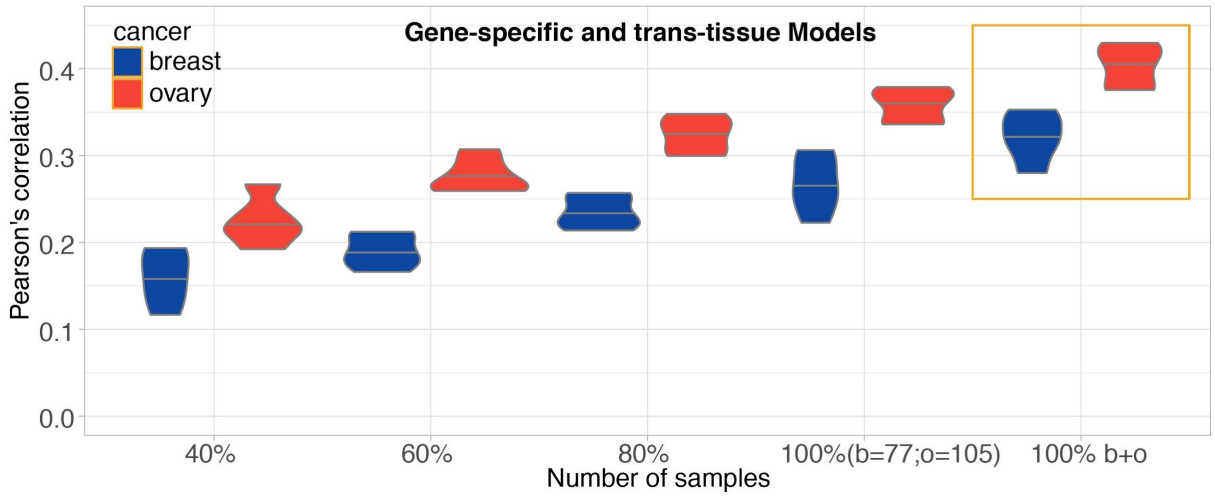


**Figure S3.5 The RMSE comparison of models trained on different number of samples.** From left to right, the RMSEs were calculated for models using (1) 40% of the training samples (2) 60% of the training samples (3) 80% of the training samples (4) 100% of the training samples (5) 100% of the combined training samples from two cancer tissues. Of note, the exact number of training samples in (4) is listed in the parentheses (b for breast and o for ovary). Our final trans-tissue model combines samples from two cancers and achieves lowest RMSEs.

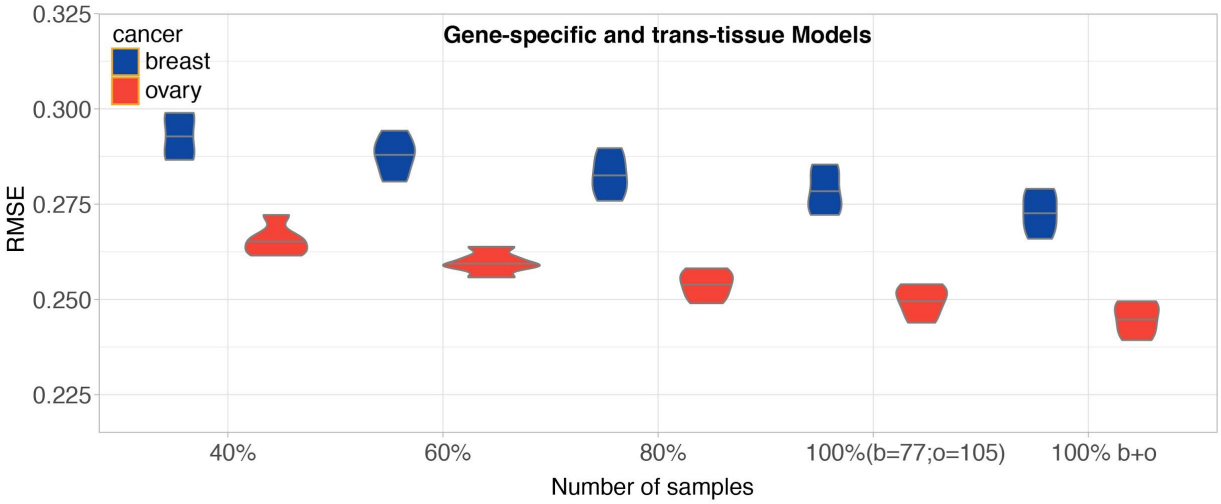


**Figure S3.6 The effects of different training scenarios and normalization strategies.**
**A-B**. The random forest model was trained on the two features, RNA and CNV values of a gene in (**A**) breast and (**B**) ovary. Two training scenarios were applied: 1. in the "local" model, only the samples of the same gene were used to train and 2. in the "global" model, the samples of all genes were used to train. To reduce the potential batch effects across individuals, in the "global+quantile" model, we quantile mapped the RNA or CNV profile of each individual to the corresponding reference profile, which was the average expression level across all individuals. We further tested the effects of adjusting the overall expression profile of an individual by multiplying an individual-specific ratio.

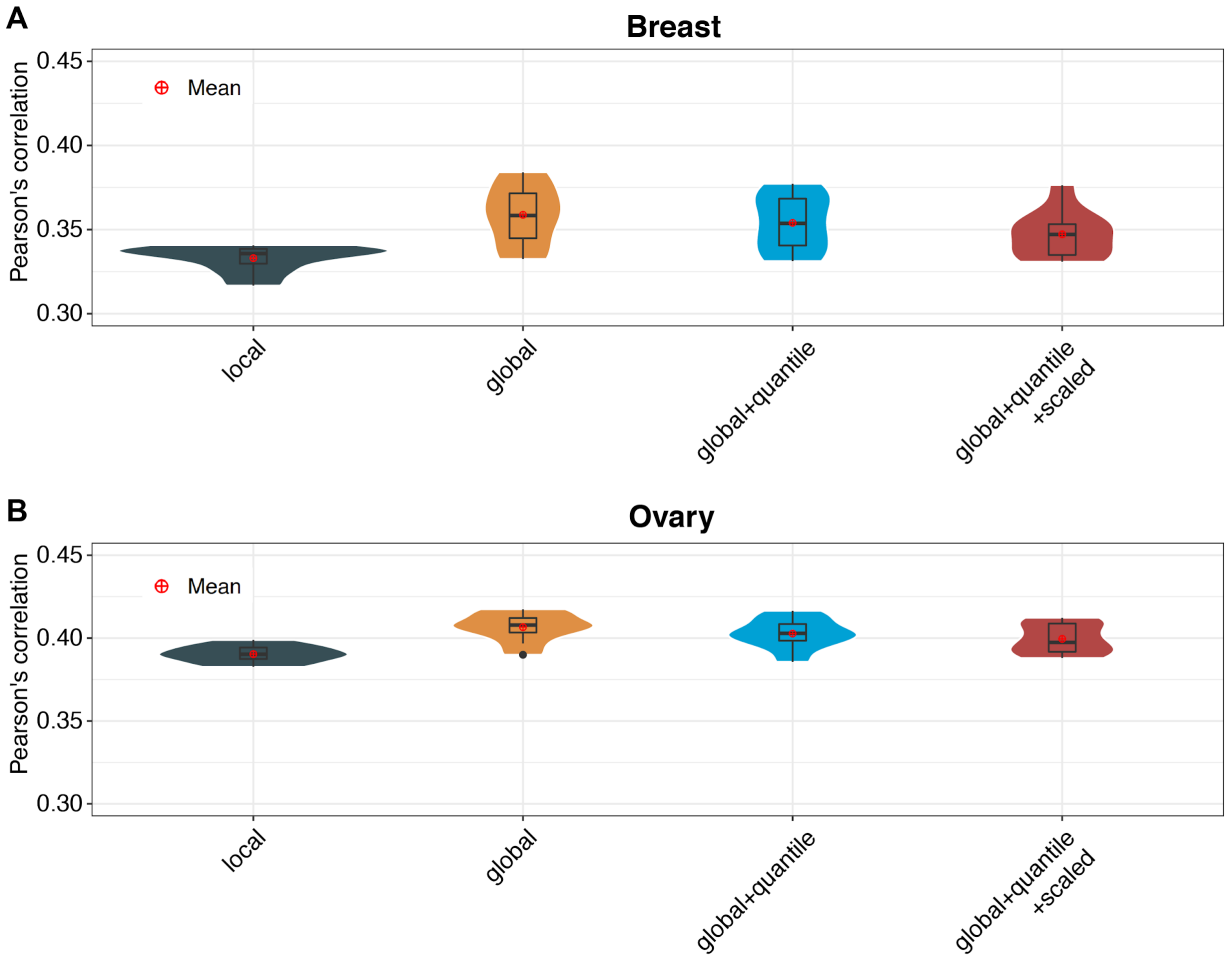


**Figure S3.7 The correlation comparison of predictions by our method and experimental replicates.**
We calculated the Pearson's correlations across 32 overlapping ovarian cancer samples measured at both JHU and PNNL for all proteins (blue). Meanwhile, the prediction correlation of our method on the held-out testing dataset during the NCI-CPTAC DREAM challenge were shown in red. The dashed line represents the average correlation.

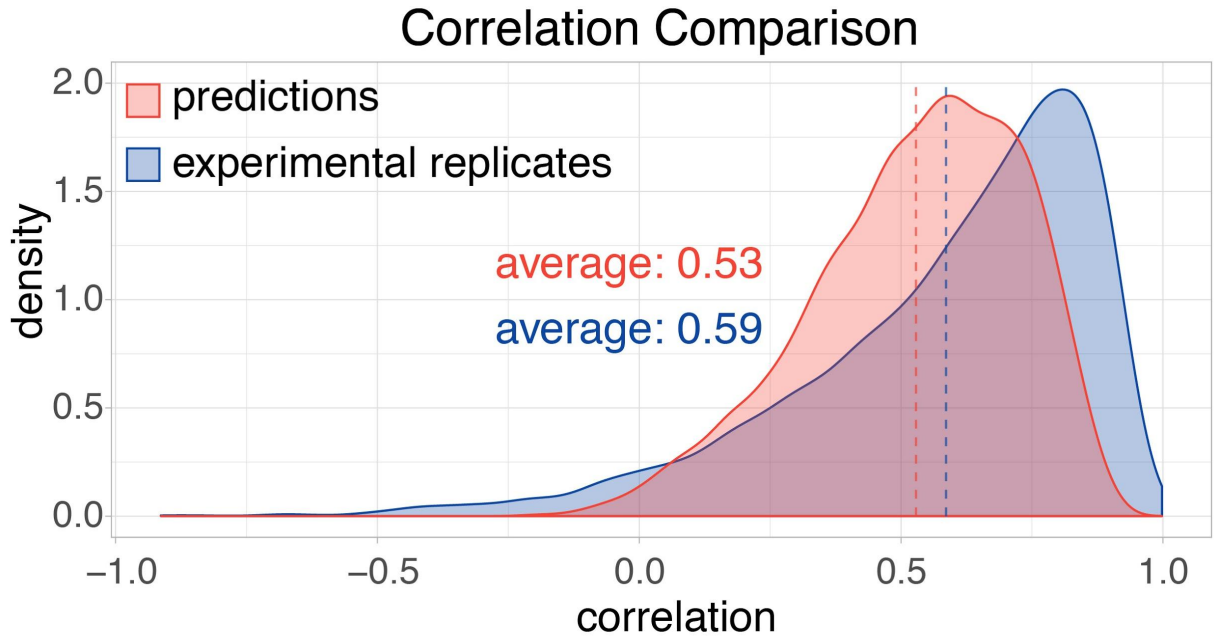


**Figure S3.8 The RMSE comparison of predictions by our method and experimental replicates.**
We calculated the RMSEs across 32 overlapping ovarian cancer samples measured at both JHU and PNNL for all proteins (blue). Meanwhile, the prediction RMSEs of our method on the held-out testing dataset during the NCI-CPTAC DREAM challenge were shown in red. The two dashed line represents the average RMSE.

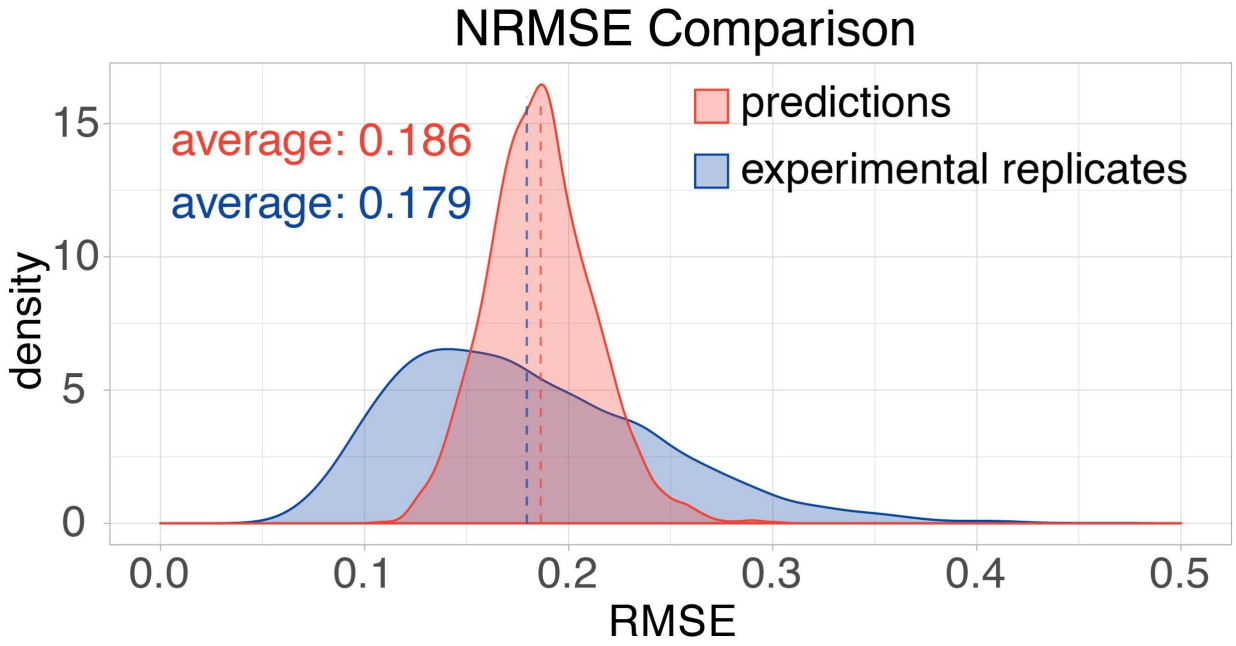


**Figure S3.9 The functional enrichment analysis of gene sets with different correlation increases.**

The overall distribution of the Pearson's correlation increases using our method, compared with the baseline mRNA-protein correlation in **A**. breast and **B**. ovarian cancers. **C-D**. Functional enrichment analysis was performed on gene subsets based on the improvement.

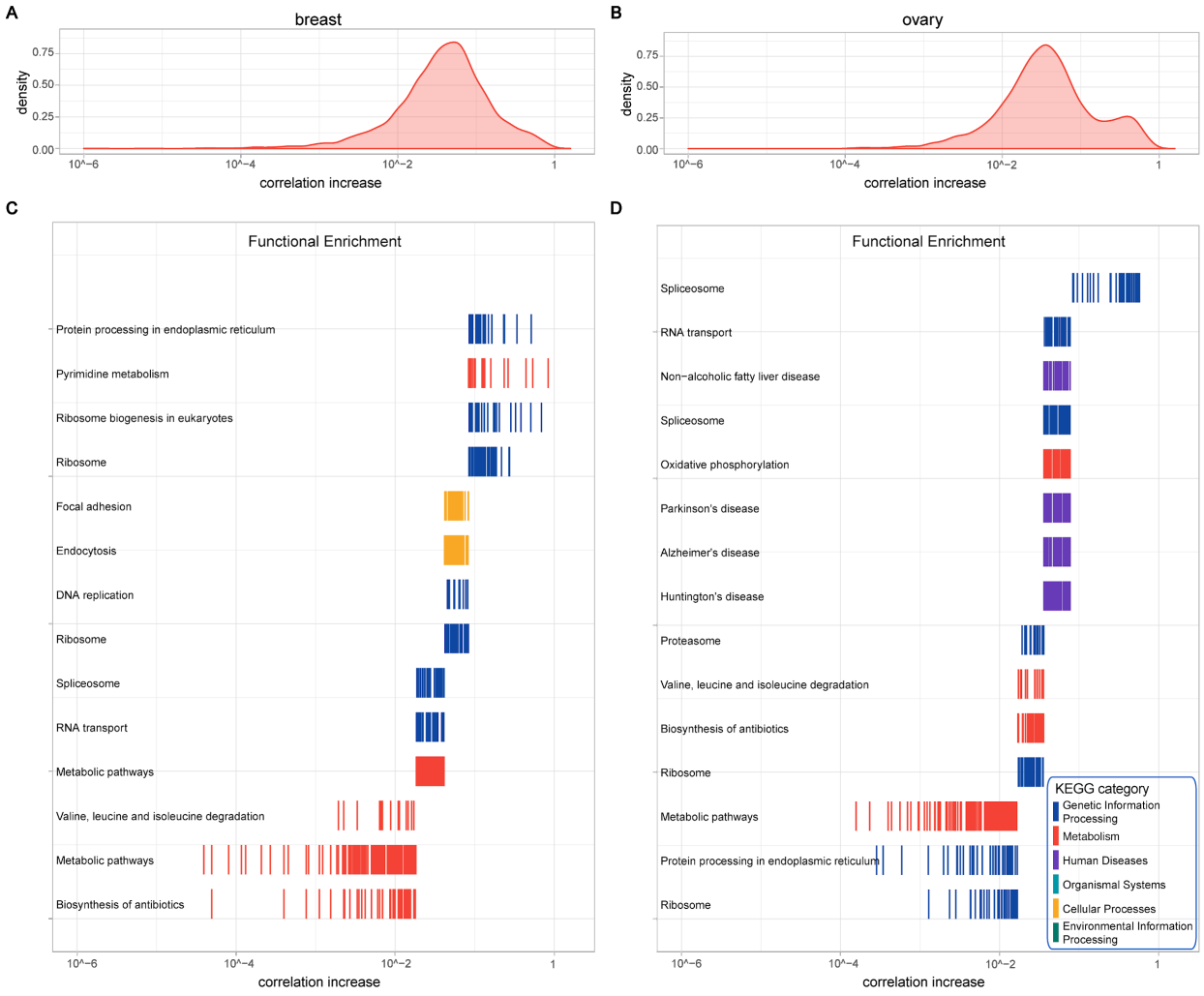


**Table S3.1 The five-fold Pearson's correlations of the generic, gene-specific and trans-tissue models.**

| breast | generic | gene-specific | trans-tissue |
|---|---|---|---|
| cross-validation1 | 0.362 | 0.408 | 0.413 |
| cross-validation2 | 0.348 | 0.370 | 0.377 |
| cross-validation3 | 0.371 | 0.396 | 0.404 |
| cross-validation4 | 0.370 | 0.420 | 0.423 |
| cross-validation5 | 0.402 | 0.420 | 0.427 |
| ovary | generic | gene-specific | trans-tissue |
| cross-validation1 | 0.377 | 0.428 | 0.434 |
| cross-validation2 | 0.413 | 0.470 | 0.478 |
| cross-validation3 | 0.402 | 0.459 | 0.468 |
| cross-validation4 | 0.430 | 0.491 | 0.498 |
| cross-validation5 | 0.398 | 0.444 | 0.449 |

## References


Arnold KM, Opdenaker LM, Flynn D, Sims-Mourtada J. 2015. Wound healing and cancer stem cells: inflammation as a driver of treatment resistance in breast cancer. *Cancer Growth Metastasis* **8**: 1–13.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291–336.

Breiman L. 2001. 10.1023/A:1010933404324. *Machine Learning* **45**: 5–32. http://link.springer.com/10.1023/A:1010933404324.

Chang Y-F, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**: 51–74.

Chircop M, Speidel D. 2014. Cellular stress responses in cancer and cancer therapy. *Front Oncol* **4**: 304.

Cooper TA, Wan L, Dreyfuss G. 2009. RNA and disease. *Cell* **136**: 777–793.

Crick FH. 1958. On protein synthesis. *Symp Soc Exp Biol* **12**: 138–163.

Du L, Pertsemlidis A. 2011. Cancer and neurodegenerative disorders: pathogenic convergence through microRNA regulation. *J Mol Cell Biol* **3**: 176–180.

Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, Rodland KK, Townsend RR, Kinsinger C, Mesri M, Rodriguez H, et al. 2013. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* **3**: 1108–1112.

Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, Troyanskaya OG. 2008. A genomewide functional network for the laboratory mouse. *PLoS Comput Biol* **4**: e1000165.

Guhaniyogi J, Brewer G. 2001. Regulation of mRNA stability in mammalian cells. *Gene* **265**: 11–23.

Guimaraes JC, Rocha M, Arkin AP. 2014. Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic Acids Res* **42**: 4791–4799.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27–30.

Liddington RC. 2004. Structural basis of protein–protein interactions. In *Protein-Protein*

*Interactions*, pp. 003–014, Humana Press.

Li H-D, Menon R, Eksi R, Guerler A, Zhang Y, Omenn GS, Guan Y. 2016. A network of splice isoforms for the mouse. *Sci Rep* **6**: 24507.

Li H-D, Menon R, Govindarajoo B, Panwar B, Zhang Y, Omenn GS, Guan Y. 2015. Functional networks of highest-connected splice isoforms: from the chromosome 17 human proteome project. *J Proteome Res* **14**: 3484–3491.

Li H, Hu S, Neamati N, Guan Y. 2018a. TAIJI: Approaching Experimental Replicates-Level Accuracy for Drug Synergy Prediction. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/bty955.

Li H, Li T, Quang D, Guan Y. 2018b. Network Propagation Predicts Drug Synergy in Cancers. *Cancer Res* **78**: 5446–5457.

Li H, Panwar B, Omenn GS, Guan Y. 2018c. Accurate prediction of personalized olfactory perception from large-scale chemoinformatic features. *Gigascience* **7**. http://dx.doi.org/10.1093/gigascience/gix127.

Li H, Quang D, Guan Y. 2018d. Anchor: Trans-cell type prediction of transcription factor binding sites. *Genome Res*. http://dx.doi.org/10.1101/gr.237156.118.

Liu Y, Beyer A, Aebersold R. 2016. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**: 535–550.

Lovett PS, Rogers EJ. 1996a. Ribosome regulation by the nascent peptide. *Microbiol Rev* **60**: 366–385.

Lovett PS, Rogers EJ. 1996b. Ribosome regulation by the nascent peptide. *Microbiol Rev* **60**: 366–385.

Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, et al. 2016. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**: 55–62.

Morris LGT, Veeriah S, Chan TA. 2010. Genetic determinants at the interface of cancer and neurodegenerative disease. *Oncogene* **29**: 3453–3464.

Newman MEJ. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**: 8577–8582.

Newman MEJ, Girvan M. 2004. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **69**: 026113.

Ning K, Fermin D, Nesvizhskii AI. 2012. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J Proteome Res* **11**: 2261–2271.

Raczynska KD, Simpson CG, Ciesiolka A, Szewc L, Lewandowska D, McNicol J, Szweykowska-Kulinska Z, Brown JWS, Jarmolowski A. 2010. Involvement of the nuclear cap-binding protein complex in alternative splicing in Arabidopsis thaliana. *Nucleic Acids Res* **38**: 265–278.

Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, Hinoue T, Laird PW, Hoadley KA, Akbani R, et al. 2017. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* **171**: 540–556.e25.

Spencer P, Fry RC, Kisby GE. 2012. Unraveling 50-year-old clues linking neurodegeneration and cancer to cycad toxins: are microRNAs common mediators? *Front Genet* **3**: 192.

Stolovitzky G, Monroe D, Califano A. 2007. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann N Y Acad Sci* **1115**: 1–22.

Sundaram GM, Ismail HM, Bashir M, Muhuri M, Vaz C, Nama S, Ow GS, Vladimirovna IA, Ramalingam R, Burke B, et al. 2017. EGF hijacks miR-198/FSTL1 wound-healing switch and steers a two-pronged pathway toward metastasis. *J Exp Med* **214**: 2889–2900.

The Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* **45**: D331–D338.

Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**: 227–232.

Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, et al. 2014. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**: 382–387.

Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou J-Y, Petyuk VA, Chen L, Ray D, et al. 2016. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166**: 755–765.

Broad GDAC Firehose. https://gdac.broadinstitute.org/ (Accessed April 22, 2018a).

CPTAC Data Portal. https://cptac-data-portal.georgetown.edu/cptacPublic/ (Accessed April 21, 2018b).

TCGA. *The Cancer Genome Atlas - National Cancer Institute*. https://cancergenome.nih.gov/ (Accessed March 29, 2018c).

# CHAPTER IV

## DeepSleep: Near-perfect Detection of Sleep Arousals at Millisecond Resolution by Deep Learning


### Abstract

Sleep has an essential impact on our health and wellbeing. Sleep arousals are transient periods of wakefulness punctuated into sleep. Excessive arousals lead to fragmented sleep and have various negative effects. Accurate diagnosis of sleep arousal disorders requires high-quality annotations of sleep records. Currently, sleep arousal annotations are performed by human experts manually looking at millions of data points, which requires considerable time and effort. There exist automatic sleep arousal detection tools, however, their performance is unsatisfactory. Here we present a deep learning approach, DeepSleep, which empowers fast and automatic detection of sleep arousals within 10 seconds per sleep record. This method ranked first in the 2018 PhysioNet Challenge for segmenting sleep arousal regions based on polysomnographic recordings. Compared with the reported theoretical upper limit, DeepSleep approximates human performance in detecting sleep arousals. Moreover, the pattern of DeepSleep segmentations differs from human annotations for sleep arousals, especially at the low-confident boundary regions. These results indicate that computer-assisted segmentations can serve as an alternative to human annotations, and potentially allow for improvement of the current scoring criteria and binary-label system.


---

The work presented in Chapter IV will be submitted as Li H, et.al. "DeepSleep: Near-perfect Detection of Sleep Arousals at Millisecond Resolution by Deep Learning".

## Introduction

Sleep is important for our overall health and quality of life (Mukherjee et al. 2015; Buysse 2014; Takahashi 2012). Inadequate sleep is often associated with many negative outcomes, including obesity (Gangwisch et al. 2005; St-Onge 2017; Miller et al. 2018b), irritability (Gangwisch et al. 2005; St-Onge 2017; Miller et al. 2018b; Paiva et al. 2015), cardiovascular dysfunction (Tobaldini et al. 2017; Bauters et al. 2016), hypotension (Lewis et al. 2015), impaired memory (Banks and Dinges 2007) and depression (Vitiello 2018; Okun et al. 2018). About one third of the general population in United States are affected by insufficient sleep (Liu et al. 2016). The prevalence of inadequate sleep results in large economic costs (Hillman et al. 2018) and continues to increase in various nations (Ford et al. 2015; St-Onge et al. 2016; Kronholm et al. 2016). Spontaneous sleep arousals, defined as brief intrusions of wakefulness into sleep (1992; Halasz et al. 2004), are a common characteristic of brain activity during sleep. Excessive arousals due to disturbances can be harmful, resulting in fragmented sleep, daytime sleepiness and sleep disorders (Bonnet 1985, 1986; Ting and Malhotra 2005). There are different types of arousing stimulus, including obstructive sleep apneas or hypopneas, respiratory effort-related arousals (RERA), hyperventilations, bruxisms (teeth grinding), snoring, vocalizations, and leg movements. Together with sleep stages (wakefulness, stage1, stage2, stage3, and rapid eye movement), sleep arousals are labeled through visual inspections of polysomnographic recordings according to the American Academy of Sleep Medicine (AASM) scoring manual (Berry et al. 2017). Of note, an 8-hour sleep record sampled at 200Hz with 13 different physiological measurements contains a total of 75 million data points. It takes hours to manually annotate such a large-scale sleep record.

Extensive research efforts have been made in developing computational methods for automatic sleep stage scoring (Hsu et al. 2013; Sharma et al. 2017; Suzuki et al. 2017; Alickovic and Subasi 2018; ’t Wallant et al. 2016; Sousa et al. 2015; de Carli et al. 1999; Sugi et al. 2009; Shahrbabaki et al. 2015; Cho et al. 2005; Shmiel et al. 2009; Fernández-Varela et al. 2017a; Phan et al. 2019; Biswal et al. 2018; Patanaik et al. 2018; Anderer et al. 2005; Sun et al. 2017; Ebrahimi et al. 2008; Malafeev et al. 2018; Tsinalis et al. 2016; Supratak et al. 2017; Ronzhina et al. 2012; Huang et al. 2018; Andreotti et al. 2018; Zhang et al. 2016; Sors et al. 2018; Sun et al. 2018; Chambon et al. 2018) and arousal detection (Olsen et al. 2018; Basner et al. 2007; Behera et al. 2014; Fernández-Varela et al. 2017b; Alvarez-Estevez and Fernández-Varela 2019) based on polysomnographic recordings. These methods mainly focus on 30-second epochs, and extract statistical features in the time and frequency domains through Fourier transform or in-house feature engineering. These features and/or raw signals are subsequently fed into classical machine learning or neural network models to classify different sleep stages and events. Typically, each sleep stage lasts more than ten minutes and transition between sleep stages forms a unique architecture, the sleep circle. In contrast, sleep arousals are extremely short, being less than one minute, and sparsely distributed during sleep. The accumulated length of sleep arousals is usually less than 10 percent of the total sleep time. Therefore the prediction of sleep arousals is a highly imbalanced classification problem. In addition, the arousal patterns vary dramatically across individuals (e.g. some individuals do not have any arousal while others may have hundreds of arousals per night), further complexing the situation and rendering it a much more difficult task than sleep staging. A key question is how to build an accurate, generalizable, and robust model to quickly detect sleep arousals. In particular, how to preprocess the raw data or extract features before training models? Which types of machine

learning models are well suited? What is the optimal input length (e.g. 30-second epochs or full-length records)? Which types of physiological signals should be used?

Here we investigate these questions and benchmark state-of-the-art methods in sleep staging and arousal detection. We describe a novel deep learning approach, DeepSleep, for automatic detection of sleep arousals. This approach ranked first in the 2018 "You Snooze, You Win" PhysioNet/Computing in Cardiology Challenge (Ghassemi et al. 2018), in which computational methods were systematically evaluated for predicting non-apnea arousal regions based on polysomnographic recordings (Guan 2019). The workflow of DeepSleep is schematically illustrated in **Figure 4.1**. We built a deep convolutional neural network (CNN) to capture long-range and short-range interdependencies between timepoints across an entire sleep record. Information at different resolutions and scales was integrated to improve the performance. Intriguingly, we found that similar EEG and EMG channels were interchangeable, which was used as a special augmentation in our approach. Compared with the theoretical upper limit calculated from annotation replicates by different sleep experts, DeepSleep achieved near-perfect detection of sleep arousals at millisecond resolution, approximating human performance.

## Results

### Overview of the experimental design for predicting sleep arousals from polysomnogram

In this work, we used the 994 polysomnographic records provided in the 2018 PhysioNet challenge, which were collected at the Massachusetts General Hospital. In each record, 13 physiological measurements were sampled at 200Hz (Location and Data in **Figure 4.1**), including six electroencephalography (EEG) signals at F3-M2, F4-M1, C3-M2, C4-M1, O1-M2 and O2-M1;

one electrooculography (EOG) signal at E1-M2; three electromyography (EMG) signals of chin, abdominal and chest movements; one measure of respiratory airflow; one measure of oxygen saturation ($SaO_2$); one electrocardiogram (ECG). Each time point in the polysomnographic record was labeled as "Arousal" or "Sleep" by sleep experts, excluding some non-scoring regions such as apnea or hypopnea arousals. To exploit the information of the training records, we employed a nested train-validate-test framework, in which 60% of the data was used to train the neural network, 15% of the data was used to validate for parameter selection and 25% of the data was used to evaluate the performance of the model (Cross-validation in **Figure 4.1**). To capture the long-range and short-range information at different scales, we adapted a classic neural network (Model in **Figure 4.1**), U-Net, which was originally designed for image segmentation. Multiple data augmentation strategies, including swapping similar polysomnographic channels, were used to expand the training data space and enable the generalizability of the model. Finally, the prediction performance was evaluated by the area under receiver operating characteristic curve (AUROC) and the area under precision-recall curve (AUPRC) on the held-out test dataset of 989 records (Evaluation in **Figure 4.1**) during the challenge. Since sleep arousal events are extremely rare (<10% in terms of length), the performances of different methods are not apparent in the Receiver Operating Characteristic (ROC) curve (Davis and Goadrich 2006; Li et al. 2019), where the y-axis is the True Positive Rate (TPR) and the x-axis is the False Positive Rate (FPR). This is because when the number of negative events ("Sleep"; 92.8%), is much larger than the positive ones ("Arousal"; 7.2%), the FPR is always very small and will barely change even if a poor model makes many FP predictions. Therefore, in addition to the commonly used AUROC, we evaluated our model and various strategies using ARPRC (Li et al. 2018). In the Precision-Recall space, the Precision and Recall are defined as The Precision is very sensitive to FP when the number of TP

is relatively small. Therefore, the AUPRC metric is able to distinguish the performances in highly unbalanced data such as the annotations of sleep arousals.

**Highly heterogeneous sleep records among individuals**

By investigating the annotations of these sleep records, we found high levels of heterogeneity among individuals. In **Figure 4.2a**, we randomly selected sleep records of 20 individuals and presented the annotations in different colors. There are 8 major annotation categories: "Arousal", "Undefined", "REM" (Rapid Eye Movement), "N1" (Non-REM stage 1), "N2" (Non-REM stage 2), "N3" (Non-REM stage 3), "Wake" and "Apnea". The distribution of these categories differs dramatically among individuals (different colors in **Figure 4.2a**). Clearly, different individuals display distinct patterns of sleep, including the length of total sleep time and multiple sleep stages. Notably, the sleep arousal regions are relatively short and sparsely distributed along the entire record for most individuals (yellow regions in **Figure 4.2a**).

We further investigated the occurrence of arousals and found that the median number of arousals during sleep was 29, indicating the prevalence of sleep arousals. A total of 43 individuals (4.33%) had solid sleep without any arousal, whereas 82 individuals (8.25%) had more than 100 arousals during their sleep (y-axis in **Figure 4.2b)**, lasting around 10% of the total sleep duration (x-axis in **Figure 4.2b**). In addition, there was no significant correlation between the total sleep time and the total length of sleep arousals (**Figure 4.2c**), which was expected since quality of sleep is not determined by sleep length. In sum, the intrinsically high heterogeneity of sleep records across individuals rendered the segmentation of sleep arousals a very difficult problem.

**Deep U-Net captures the long-range and short-range information at different scales and resolutions**

Current manual annotation of sleep arousals is defined by the AASM scoring manual (Berry et al. 2017), in which sleep experts focus on a short period (less than a minute) and make decisions about sleep arousal events. However, it remains unclear whether the determinants of sleep arousals reside only within a short range, or long-range information across minutes and even hours plays an indispensable role in detecting sleep arousals. Although sleep arousal is in nature a transient event, it may be associated with the overall sleep pattern through the night. Intriguingly, when we trained the convolutional neural networks on longer sleep records, we consistently achieved better performances (**Supplementary Figure 4.1**). Therefore, we used the entire sleep record as input to make predictions, instead of small segments of a sleep record.

To learn the long-range association between data points across different time scales (second, minute, and hours), we develop an extremely deep convolutional neural network, which contains a total of 35 convolutional layers (**Figure 4.3a**). This network architecture has two major components, the encoder and the decoder. The encoder takes a full-length sleep record of $2^{23}$ = 8,388,608 time points and gradually encrypts the information into a latent space (the red trapezoid in **Figure 4.3a**). Sleep records with different lengths are made uniform to the same 8-million length by padding zeros at both the beginning and the end. To be specific, the convolution-convolution-pooling (hereafter referred to as "ccp") block is used to gradually reduce the size from $2^{23}$ = 8,388,608 to $2^8$ = 256 (**Figure 4.3b** top; see details in **Methods**). Meanwhile, the number of channels gradually increases from 13 to 480 to encode more information, compensating the loss

of resolution in the time domain. In each convolutional layer, the convolution operation is applied on the data along the time axis to aggregate the neighborhood information. Since the sizes of data in these convolutional layers are different, the encoded information is unique within each layer. For example, in the input layer, 10 successive time points (sampled at 200Hz) correspond to a short time interval of 0.05 seconds, whereas in the center layer (size = $2^8$), 10 time points correspond to a much longer time interval of $0.05 * 2^{23-8}$ = 1,638 seconds, nearly 30 minutes. Therefore, this deep encoder architecture allows us to capture and learn about the interactions across data points at multiple time scales.

Similar to the encoder, the second component of our network architecture is a decoder to decrypt the compressed information from the center latent space. In contrast to the "ccp" block, the convolution-convolution-upscaling (hereafter referred to as "ccu") block is used (**Figure 4.3b** bottom; see details in **Methods**), which gradually increases the size and decreases the number of channels of the data (the purple trapezoid in **Figure 4.3a**). In addition, the concatenation is used to integrate the information from both the encoder and the decoder at each time scale (green horizontal arrows in **Figure 4.3**). The concatenation is a unique feature of U-Net, without which the performance decreases (**Supplementary Figure 4.2-4.3**). Finally, the output is the segmentation of the entire sleep record, where high prediction values indicate sleep arousal events and low values indicate sleep.

**Deep learning enables accurate predictions of sleep arousals**

By capturing the information at multiple resolutions, DeepSleep achieves high performance in automatic segmentation of sleep arousals. Since deep neural networks are iteration-based machine

learning approaches, a validation subset is used for monitoring the underfitting or overfitting status of a model and approximating the generalization ability on unseen datasets. A subset of 15% randomly selected records was used as the validation set during the training process (Cross-validation in **Figure 4.1**) and the cross entropy was used to measure the training and validation losses (see details in **Methods**). Since the 13 polysomnographic channels complemented each other, using all of them instead of one type of these signals enabled the neural network to capture interactions between channels and achieved the highest performance (**Supplementary Figure 4.4**). We developed three basic models called "1/8", "1/2" and "full", according to the resolution of the neural network input. The "full" resolution means that the original 8-million ($2^{23}$ = 8,388,608) length data were used as input. The "1/2" or "1/8" resolution means that the original input data were first shrunk to the length of 4-million ($2^{22}$) or 1-million ($2^{20}$) by averaging every 2 or 8 successive time points, respectively. We observed similar validation losses of the "full", "1/2" and "1/8" models (solid lines in **Figure 4.4a**). The final evaluation was based on the AUROC and AUPRC scores of predicting 25% of the data. In **Figure 4.4b**, each blue dot represented one sleep record and we observed a significant yet weak correlation = 0.308 between the AUROCs and AUPRCs. The baselines of random predictions were shown as red dashed lines. Notably, the AUPRC baseline of 0.072 corresponded to the ratio of the average total sleep arousal length over the total sleep time, which was considerably low and made it a hard task due to the intrinsic sparsity of sleep arousal events.

To build a robust and generalizable model, multiple data augmentation strategies were used in DeepSleep. After carefully examining the data, we found that signals belonging to the same physiological categories were very similar and synchronized, including two EMG channels and

six EEG channels (see Data in **Figure 4.1b**). We applied a novel augmentation strategy by randomly swapping these similar channels during the model training process, assuming that these signals were interchangeable in determining sleeping arousals. This channel swapping strategy was bold but effective, adapting which largely improved the prediction performance ("1/8_no_swap" versus "1/8" in **Figure 4.4c-d**). In addition, we multiplied the polysomnographic signals by a random number between 0.90 and 1.15 to simulate the inherent fluctuation and noise of the data. Furthermore, to address the heterogeneity and batch effects among individuals, we quantile normalized each sleep record to a reference, which was generated by averaging all the records. This step effectively removed the biases introduced by the differences of individuals and instruments. Finally, we assembled the predictions from the "1/8" , "1/2" and "full" resolution models as the final prediction in DeepSleep (red violin plots in **Figure 4.4c-d**).

We further compared different machine learning models and strategies, and benchmarked current methods in segmenting sleep arousals and stages. We first tested a classical model, logistic regression, and found that our deep learning approach had a much higher performance (**Supplementary Figure 4.5**). It has also been reported that neural network approaches significantly outperformed classical machine learning methods, including random forest (Biswal et al. 2018), logistic regression (Biswal et al. 2018), support vector machine (Alvarez-Estévez and Moret-Bonillo 2011), linear and quadratic models (Alvarez-Estévez and Moret-Bonillo 2011; Becq et al. 2005). In fact, 8 out of the top 10 teams (Howe-Patterson et al. 2018; Már Þráinsson et al. 2018; He et al. 2018; Varga et al. 2018; Patane et al. 2018; Miller et al. 2018a; Warrick and Nabhan Homsi 2018; Bhattacharjee et al. 2018; Szalma et al. 2018) used neural network models in the 2018 PhysioNet Challenge (red blocks in **Supplementary Figure 4.6**) (Ghassemi et al.

2018). Two types of network structures (convolutional and recurrent) were mainly used, and integrating Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) or Gated Recurrent Unit (GRU) (Cho et al. 2014) into DeepSleep did not improve the performance (**Supplementary Figure 4.7-4.8**). In terms of input length, most previous studies focused on short epochs (equal or less than 30 seconds) (Phan et al. 2019, 2018b, 2018a; Biswal et al. 2018; Sun et al. 2017). We found that increasing input length significantly improved the performance (**Supplementary Figure 4.1**), and full-length records were used by three teams (blue blocks in **Supplementary Figure 4.6**). We also compared DeepSleep with recent state-of-the-art methods in sleep stage scoring. These methods extracted features from 30-second epochs through short-time Fourier transform (STFT) (Phan et al. 2019, 2018b, 2018a) or Thomson's multitaper (Biswal et al. 2018; Sun et al. 2017), but they were not transferred very well to the task of sleep arousal detection (**Supplementary Figure 4.9**). In contrast, deep learning approaches can model informative features in an implicit way without tedious feature crafting (Sors et al. 2018), and neural networks using raw data as input were frequently used by half of the top 10 teams (orange blocks in **Supplementary Figure 4.6**).

To comprehensively investigate the effects of various network structures and parameters on predictions, we further performed experiments with different modifications, including shallow neural network, large convolution kernel size, average pooling, and loss functions. These modifications had either similar or lower prediction performances. We concluded that the neural network architecture and augmentation strategies in DeepSleep were optimized for the current task of segmenting sleep arousals. Subsequent analysis of the relationships between the prediction performance and multiple statistics were investigated. As we expected, the prediction AUPRC was

correlated with the number of arousals in a sleep record (**Supplementary Figure 4.10**). The individuals who had more sleep arousals during sleep were relatively easier to predict. Moreover, we tested the runtime of DeepSleep with GPU acceleration and segmenting sleep arousals of a full sleep record can be finished within 10 seconds on average (**Supplementary Figure 4.11**). The time cost of DeepSleep is much lower than that of manual annotations, which requires hours for one sleep record.

Unlike the sleep stage annotation, the sleep arousal annotation is a hard task. The relatively weak agreement across human experts has been reported with the Intraclass Correlation Coefficients (ICCs) of 0.520 and 0.575 for scoring sleep arousals in non-REM and REM, respectively (Kuna et al. 2013). Similar to Pearson's correlation, ICC is a statistic for quantifying the degree to which data from the same group resemble each other. The medium ICCs indicate that the gold standard itself is not perfect and the annotations for the same sleep record vary across different human experts. The theoretical upper bound for the task of computation prediction is thus around 0.520 to 0.575. To estimate the performance of DeepSleep, we calculated the ICC between our predictions and the gold standard annotations. DeepSleep achieved the ICC of 0.497 (**Supplementary Figure 4.12**), which is close to the theoretical upper limit in literature.

**Visualization of DeepSleep predictions**

In addition to the abstract AUROC and AUPRC scores, we directly visualized the prediction performance of DeepSeep at 5-millisecond resolution (corresponding to the 200Hz sample rate). An example 7.5-hour sleep record with the prediction AUROC of 0.960 and AUPRC of 0.761 is shown in **Figure 4.5**. From top to bottom, we plotted the multi-stage annotations, sleep arousal

labels, predictions and cross entropy losses long the time x-axis. By comparing the prediction and gold standard, we can see the general prediction pattern of DeepSleep correlates well with the gold standard across the entire record (the second and third rows in **Figure 4.5a**). We further zoom into a short interval of 12.5 minutes and DeepSleep successfully identifies and segments seven sleep arousal events out of eight (yellow in **Figure 4.5b**), although one arousal around 25,600 is missed. Intriguingly, DeepSleep predictions display a different pattern from the gold standard annotated by sleep experts: DeepSleep assigns continuous prediction values with lower probabilities near the arousal-sleep boundaries, whereas the gold standard is strictly binary either arousal = 1 or sleep = 0 based on the AASM scoring manual (Berry et al. 2017). This becomes clearer when examining the cross entropy loss at each time point and the boundary region has higher losses shown in red (the bottom row in **Figure 4.5b**). This is expected because in general we will have a higher confidence of annotation in the central region of sleep arousal or other sleep events. Yet due to the limit of time and effort, it is practically infeasible to introduce rules for manually annotating each time point via a probability scenario. Additionally, binary annotation of sleep records containing millions of data points has already required significant effort. DeepSleep opens a new avenue to reconsider the way of defining sleep arousals or other sleep stage annotations by introducing the probability system.

**Discussion**

In this study, we created a deep learning approach, DeepSleep, to automatically segment sleep arousal regions in a sleep record based on the corresponding polysomnographic signals. A deep convolutional neural network architecture was designed to capture the long-range and short-range interactions between data points at different time scales and resolutions. Unlike traditional machine

learning models, deep learning approaches do not depend on manually crafted features and can automatically extract information from large datasets in an implicit way (LeCun et al. 2015). Using traditional approaches to define rules and craft features for modelling sleep problems in real life would become much too tedious. In contrast, without assumptions and restrictions, deep neural networks can approximate complex mathematical functions and models to address those problems. Currently, these powerful tools have also been successfully applied to biomedical image analysis and signal processing (Litjens et al. 2017; Shen et al. 2017; Faust et al. 2018).

Overfitting is a common issue in deep learning models, especially when the training dataset is small and the model is complex. Many previous studies only trained and evaluated models on less than 100 polysomnographic recordings (Ebrahimi et al. 2008; Malafeev et al. 2018; Huang et al. 2018; Chambon et al. 2018; Supratak et al. 2017; Sun et al. 2018; Zhang et al. 2016), which may not generalize well. Moreover, even if we use a large dataset and perform cross-validation, we will gradually and eventually overfit to the data. This is because each time we evaluate a model using the internal test set, we probe the dataset and fit our model to it. In contrast to previous studies, the 2018 PhysioNet Challenge offered us a unique opportunity to truly evaluate the performances and compare cutting-edge methods on a large external hidden test set of 989 samples (Guan 2019). In addition, we demonstrate that deep convolutional neural networks trained on full-length records and multiple physiological channels have the best performance in detecting sleep arousals, which are quite different from current approaches extracting features from short 30-second epochs (Phan et al. 2019; Biswal et al. 2018; Huang et al. 2018; Patanaik et al. 2018; Chambon et al. 2018; Supratak et al. 2017; Sors et al. 2018; Andreotti et al. 2018). The ideas embedded in our approach will provide unprecedented insights into future method development for automatic scoring of sleep.

An interesting observation is that when we used records of different lengths as input to train deep learning models, the model using full-length records largely outperformed models using short periods of records. This observation brings about the question of how to accurately detect sleep arousals based on polysomnography. Current standards mainly focus on short time intervals of less than one minute (Berry et al. 2017), yet the segmentations among different sleep experts are not very consistent in determining sleep arousals. One reason is that it is hard for human to directly read and process millions of data points at once. In contrast, computer is good at processing large-scale data and discover the intricate interactions and structures between data points across seconds, minutes and even hours. Our results indicate that sleep arousal events are not be solely determined by the local physiological signals but associated with much longer time intervals even spanning hours. It would be interesting to foresee the integration of computer-assisted annotations to improve definitions of sleep arousals or other sleep stages.

In addition to the unique long-range information captured by DeepSleep, a clear advantage of computational approaches lies in the annotations for the boundary regions between arousal and sleep. Since current sleep annotations are binary only, it would be a more accurate and appropriate approach to introduce the probability of the annotation confidence, especially at the boundary regions. Machine learning approaches such as DeepSleep naturally provide the continuous predictions for each time point. It would be interesting to see improved annotation systems using continuous values instead of binary labels. A simple approach could be directly integrating the computer predictions with annotations by human sleep experts. The proposed annotation systems

would provide more accurate information for the diagnosis of sleep disorders and the evaluation of sleep quality in the future.

## Methods

### Polysomnographic recordings

The dataset used in this study contains a total of 994 polysomnographic sleep records from different individuals and their corresponding labels at each time point. Specifically, the arousal region is labeled by "1" and other sleep regions are labeled by "0", except for the wakefulness regions, apnea arousal regions and hypopnea arousal regions labeled by "-1". These "-1" regions will not be scored in the challenge, and we mainly focused on non-apnea arousals that interrupted the sleep of an individual, including spontaneous arousals, respiratory effort related arousals (RERA), bruxisms, hypoventilations, hypopneas, apneas (central, obstructive and mixed), vocalizations, snores, periodic leg movements, Cheyne-Stokes breathing or partial airway obstructions (https://physionet.org/challenge/2018/). The final test dataset consists of 989 unseen polysomnographic recordings from different individuals. For each time point sampled at 200Hz in each test sleep record, the participants needed to provide a prediction value between 0 and 1. A 8-hour sleep record contained nearly 75 million data points (8*60*60*200*13=74,880,000). Our model made predictions for all the time points, at the resolution of 5 milliseconds (1/200Hz = 5 milliseconds).

### Partition of the training, validation and testing sleep records

The 994 sleep records were randomly partitioned into three sets: 60% of them as the training set, 15% of them as the validation set and 25% of them as the testing set. The validation set was used

for monitoring the training-validation losses and avoiding the problems of overfitting or underfitting.

**AUROC and AUPRC**

Since sleep arousal events are extremely rare (<10% in terms of length), the performances of different methods are not apparent in the Receiver Operating Characteristic (ROC) curve (Davis and Goadrich 2006; Li et al. 2019), where the y-axis is the True Positive Rate (TPR) and the x-axis is the False Positive Rate (FPR). The TPR and FPR are defined as

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

where TP is True Positive, FN is False Negative, FP is False Positive, and TN is True Negative. This is because when the number of negative events ("Sleep"; 92.8%), or TN, is much larger than the positive ones ("Arousal"; 7.2%), the FPR is always very small and will barely change even if a poor model makes many FP predictions. Therefore, in addition to the commonly used AUROC, we evaluated our model and various strategies using ARPRC (Li et al. 2018). In the Precision-Recall space, the Precision and Recall are defined as

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = TPR = \frac{TP}{TP + FN}$$

The Precision is very sensitive to FP when the number of TP is relatively small. Therefore, the AUPRC metric is able to distinguish the performances in highly unbalanced data such as the annotations of sleep arousals.

**Convolutional neural network architectures**

The classic U-Net architecture was adapted in DeepSleep. The original U-Net is a 2D convolutional neural network designed for 2D image segmentation (Ronneberger et al. 2015). We transformed the structure into 1D for the time-series sleep records and largely increased the number of convolutional layers from the original 18 to 35 for extracting the information at different scales. Similar to U-Net, we had convolution, max pooling and concatenation layers. The kernel size of 7 was used in the convolution operation and increasing the kernel size didn't significantly change the performance. The nonlinear activation after each convolution operation is a Rectified Linear Unit (ReLU) (Nair and Hinton 2010) defined as

$$f(x) = max(0, x)$$

where x is the input to a neuron and f(x) is the output. Only positive values active a neuron and ReLU allows for fast and effective training of neural networks compared to other complex activation functions. In addition, batch normalization was used after each convolutional layer. In the final output layer, we used the sigmoid activation unit defined as

$$f(x) = \frac{1}{1 + e^{-x}}$$

where x is the input to a neuron and f(x) is the output. During the training process, the Adam optimizer (Kingma and Ba 2014) was used with the learning rate of 1e-4 and the decay rate of 1e-5.

Other network structures were also tested, including AlexNet (Krizhevsky et al. 2017) (**Supplementary Fig. 2**), Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) and Gated Recurrent Unit (GRU) (Cho et al. 2014) (**Supplementary Fig. 9**). They have worse

(AlexNet) or similar (LSTM and GRU) performances. Therefore, we kept the U-Net based structure.

**Training Losses**

The cross entropy loss, or log loss, was used for model training in DeepSleep. The cross entropy loss is defined as

$$H(y,\hat{y}) = \sum_{i=1}^{N} [-y_i \cdot log\hat{y_i} - (1-y_i) \cdot log(1-\hat{y_i})]$$

where $y_i$ is the gold standard label of sleep=0 or arousal=1 at time point i, $\hat{y}_i$ is the prediction value at time point i, N is the total number of time points, $y$ is the vector of the gold standard labels and $\hat{y}$ is the vector of predictions. Ideally, an "AUPRC loss" should be used for optimizing the prediction AUPRC. However, the "AUPRC loss" doesn't exist because the AUPRC function is not mathematically differentiable, which is required in the neural network model training through the back propagation algorithm (Rumelhart et al. 1986). Therefore, we need to use cross entropy loss to approximate the "AUPRC loss". Another option is using the Sorensen-dice coefficient defined as

$$S(y,\hat{y}) = \frac{\sum_{i=1}^{N}(y_i \cdot \hat{y_i})}{\sum_{i=1}^{N}(y_i) + \sum_{i=1}^{N}(\hat{y_i})}$$

where $y_i$ is the gold standard label of sleep=0 or arousal=1 at time point i, $\hat{y}_i$ is the prediction value at time point i, N is the total number of time points, $y$ is the vector of the gold standard labels and $\hat{y}$ is the vector of predictions. We have tested the cross entropy loss, the Sorensen dice loss and combining these two losses. Using the cross entropy loss achieved the best performance in DeepSleep.

**Intraclass correlation coefficient**

Intraclass correlation coefficient is a statistic for quantifying the degree to which data from the same group resemble each other. When the paired units are under consideration, the definition of ICC is similar to Pearson's correlation, except for using the pull the average and variance. In this study, we paired the prediction with the annotation by sleep expert for each time point in each sleep record. The ICC is defined as (Bartko 1966)

$$ICC(y,\hat{y}) = \frac{1}{Ns^2}\sum_{i=1}^{N}(y_i - \underline{y})(\hat{y_i} - \underline{y})$$

$$\underline{y} = \frac{1}{2N}\sum_{i=1}^{N}(y_i + \hat{y_i})$$

$$s^2 = \frac{1}{2N}\left\{\sum_{i=1}^{N}(y_i - \underline{y})^2 + \sum_{i=1}^{N}(\hat{y_i} - \underline{y})^2\right\}$$

where $y_i$ is the gold standard label of sleep=0 or arousal=1 at time point i, $\hat{y}_i$ is the prediction value at time point i, N is the total number of time points, $y$ is the vector of the gold standard labels and $\hat{y}$ is the vector of predictions. In contrast to Pearson's correlation, the $\underline{y}$ and $s^2$ are the pulled average and variance, respectively.

Since the annotations for sleep arousals are binary, we first transformed the predictions into binary values using a cutoff before calculating the ICC. The 7% percentile value among all prediction values was selected as the cutoff, based on the positive ("Arousal") ratio of 0.07 observed in the training data. All the prediction values larger than the cutoff were set to "1" and others were set to "0", resulting in 7% of the processed predictions were "1" and 93% were "0".

**Overall AUPRC**

The overall AUPRC, or the gross AUPRC, is defined as

$$AUPRC = \sum_j P_j(R_j - R_{j+1})$$

$$P_j = \frac{number\ of\ arousal\ data\ points\ with\ predicted\ probability\ (j/1000)\ or\ greater}{total\ number\ of\ data\ points\ with\ predicted\ probability\ (j/1000)\ or\ greater}$$

$$R_j = \frac{number\ of\ arousal\ data\ points\ with\ predicted\ probability\ (j/1000)\ or\ greater}{total\ number\ of\ arousal\ data\ points}$$

where the Precision ($P_j$) and Recall ($R_j$) were calculated at each cutoff j and j = 0, 0.001, 0.002, …, 0.998, 0.999, 1. For a test dataset of multiple sleep records, this overall AUPRC is similar to the "weighted AUPRC", which is different from simply averaging the AUPRC values of all test records. This is because the overall AUPRC considers the length of each record and longer records contributing more to the overall AUPRC, resulting in a more accurate performance description of a model. The overall AUPRC was also used as the primary scoring metric in the 2018 PhysioNet Challenge.

**Data availability**

The dataset used in this study is publicly available at the 2018 PhysioNet Challenge website: https://physionet.org/physiobank/database/challenge/2018/

**Code availability**

The code of DeepSleep is available at:

https://github.com/GuanLab/DeepSleep

## Figures

**Figure 4.1 Schematic Illustration of DeepSleep workflow.**
DeepSleep is a deep neural network model for automatic detecting sleep arousals based on polysomnograms, which contain multiple human physiological signals during sleep. **Location.** The 13-channel polysomnogram monitored multiple body functions, including brain activity (six EEG channels of F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1), eye movement (one EOG channel of E1-M2), muscle activity (three EMG channels of Chin, ABD, and Chest), heartbeat (one channel of ECG), airflow, and saturation of oxygen (SaO2). **Data.** A 50-second sleep record with the gold standard label of arousal/sleep on the top and the corresponding 13 physiological features. **Cross-validation.** In the nested train-validate-test framework, 60%, 15%, and 25% of the data were used to train, validate, and evaluate the model. **Model.** The classic U-Net architecture was adapted to capture the information at different scales and allowed for detecting sleep arousals at millisecond resolution. **Evaluation.** DeepSleep achieved high prediction AUROC of 0.927 and AUPRC of 0.550.

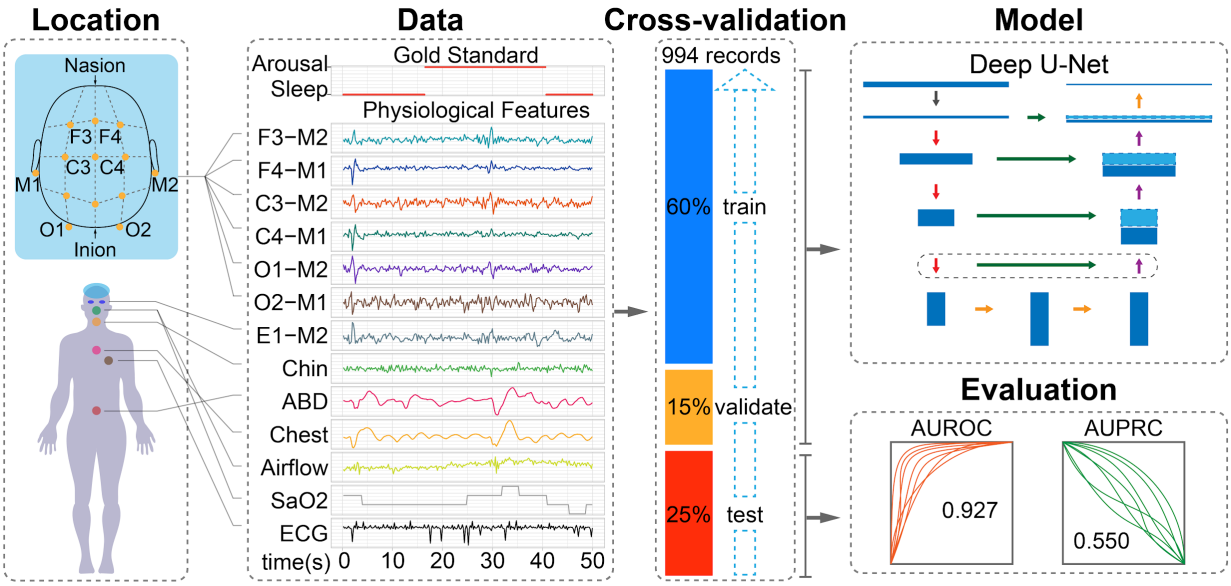


**Figure 4.2 Sleep arousals sparsely distributed in the heterogenous sleep records among individuals.**
**a.** The 8 major annotation categories are shown in different colors for 20 randomly selected sleep records. **b.** The relationship between the number of sleep arousals (y-axis) and the percentage of total sleep arousal time over total sleep time (x-axis) in the 994 sleep records. In general, more arousal events lead to longer accumulated arousal time and the correlation is significantly strong. **c.** The length of sleep (x-axis) has no significant correlation with the accumulated length of sleep arousals (y-axis).

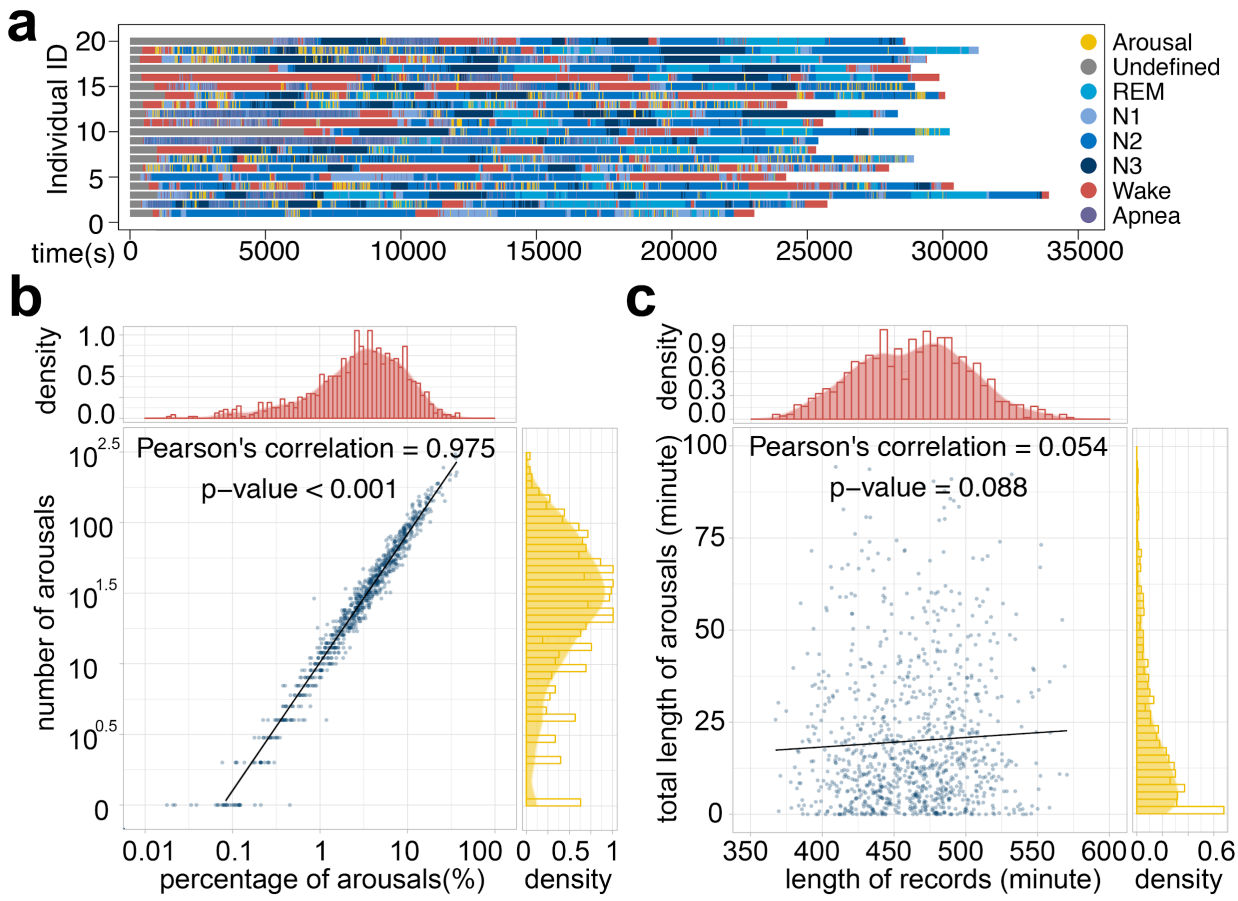


**Figure 4.3 The deep convolutional neural network architecture in DeepSleep.**
**a.** The classic U-Net structure was adapted in DeepSleep, which has two major components of the encoder (the red trapezoid on the left) and the decoder (the purple trapezoid on the right). **b.** The building blocks of DeepSleep are the convolution-convolution-pooling block (red), the concatenation (green) and the convolution-convolution-upscaling block (purple).

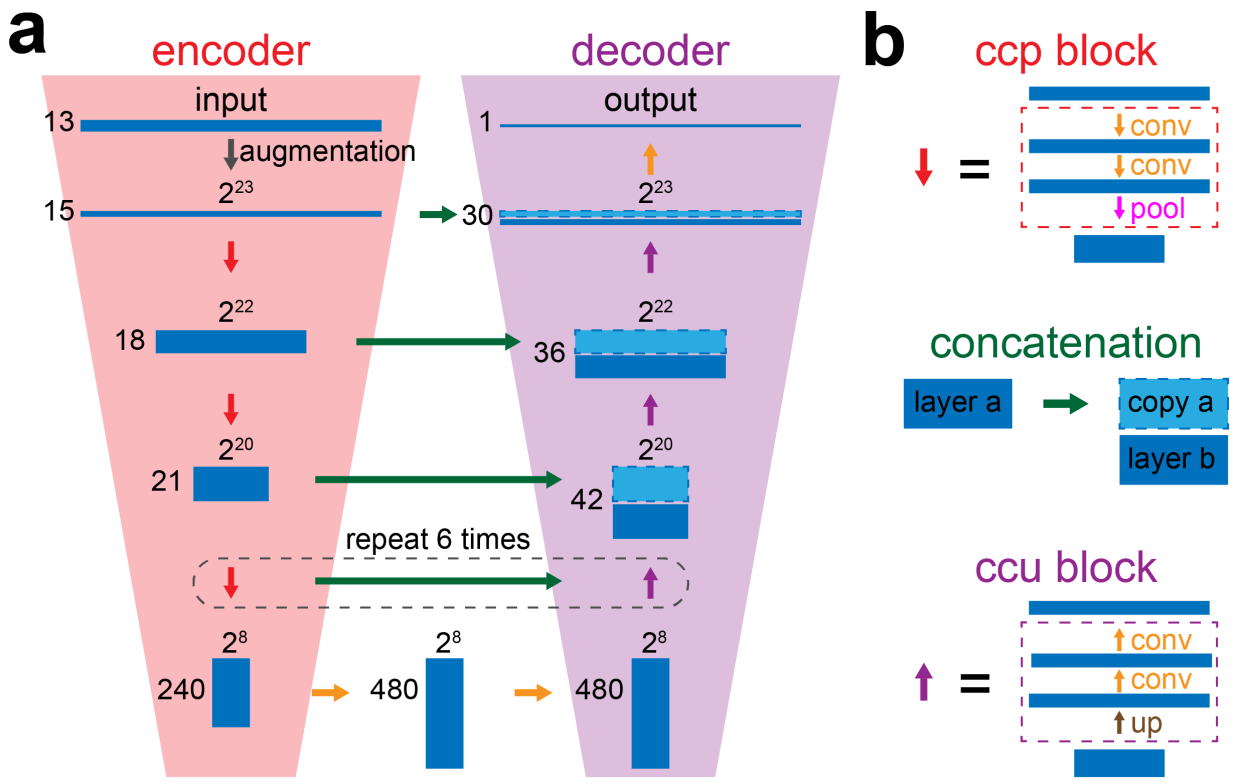


**Figure 4.4 The performance comparison of DeepSleep using different model training strategies.**
**a.** The training and validation cross entropy losses are shown in the dashed and solid lines, respectively. The models using sleep records at different resolutions are shown in different colors. **b.** The prediction of each sleep record in the test set is shown as a blue dot in the AUROC-AUPRC space. A weak correlation is observed between AUROCs and AUPRCs with a significant p-value < 0.001. The 95% percent confidence interval is shown as the yellow bend. The baselines of random predictions are shown as red dashed lines. The prediction **c.** AUPRCs and **d.** AUROCs of models using different resolution or strategies were calculated. The "1/8_no_swap" model corresponds to the model using the "1/8" resolution records as input without any channel swapping, whereas the "1/8", "1/2" and "full" models use the strategy of swapping similar polysomnographic channels. The final "1/8+1/2+full" model of DeepSleep is the ensemble of models at 3 different resolutions, achieving the highest AUPRC of 0.550 and AUROC of 0.927.

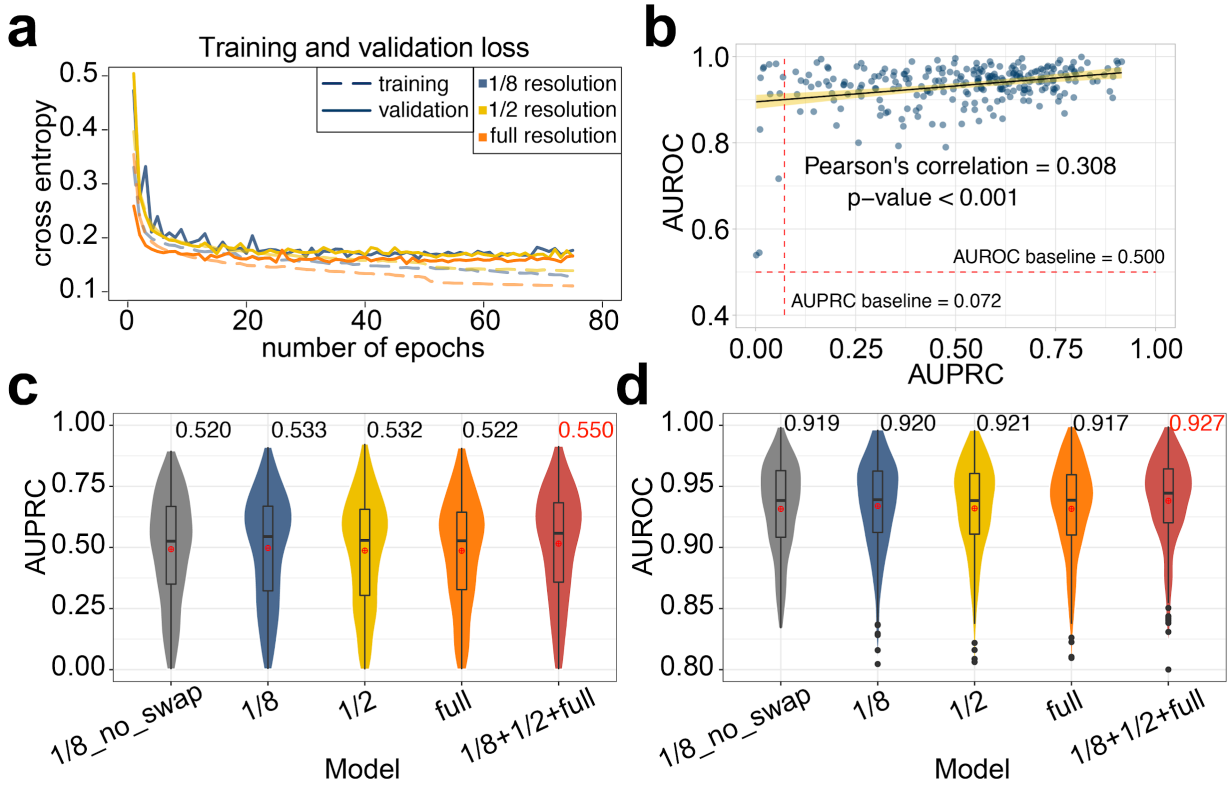
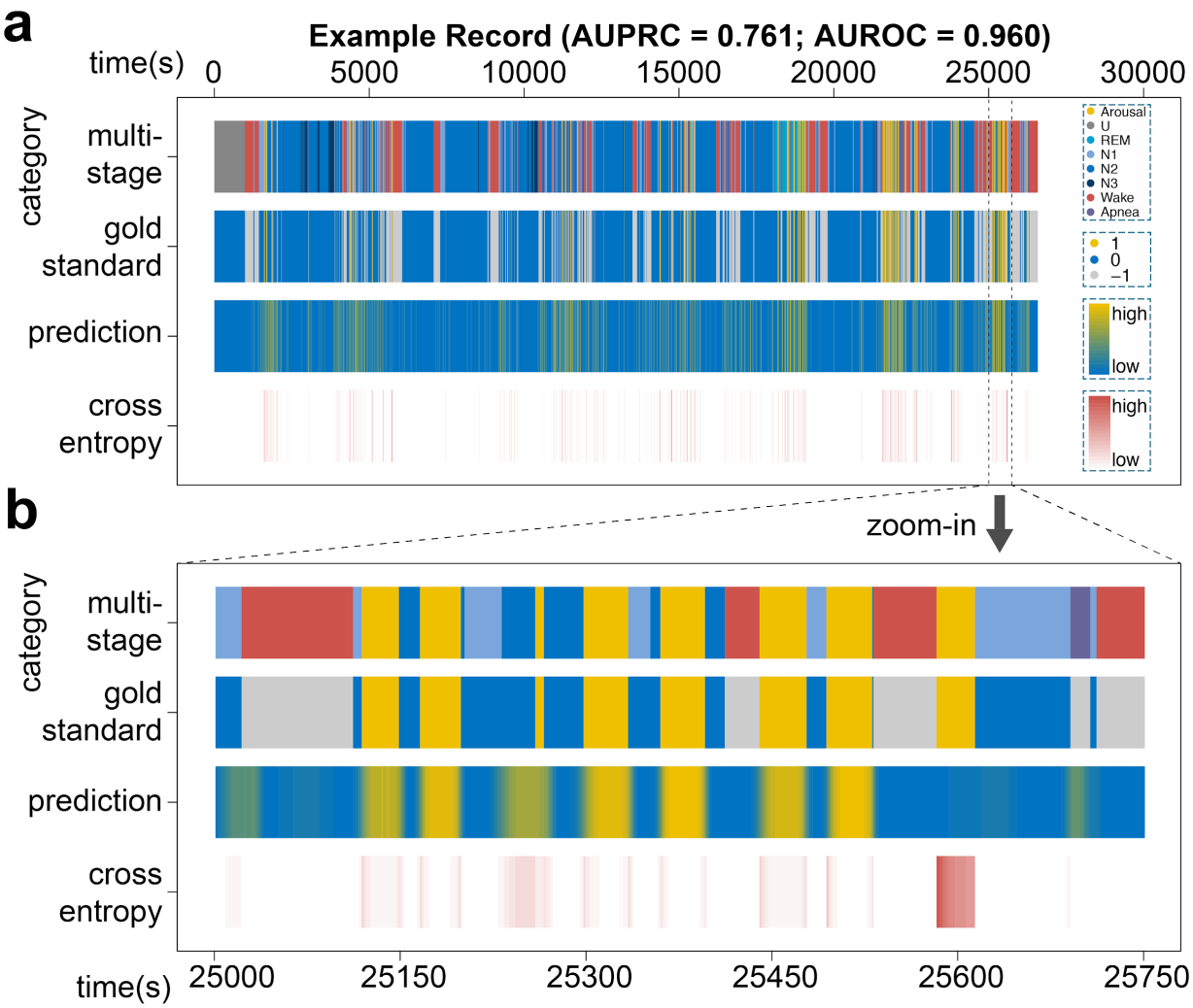
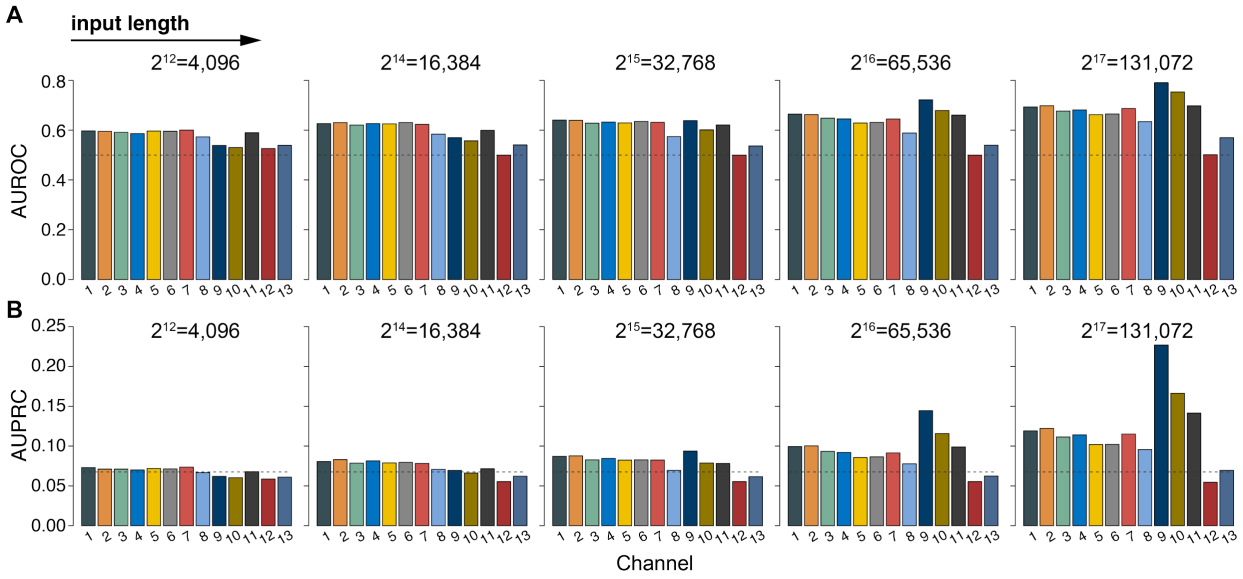


**Figure 4.5 Visualization of DeepSleep predictions and the gold standard annotations.** **a.** A 7.5-hour sleep record (id=tr05-1034) with the prediction AUROC of 0.960 and AUPRC of 0.761 is used as an example. From top to bottom along the y-axis, the four rows correspond to the 8 annotation categories, the binary label of arousal (yellow), sleep (blue) and the non-scoring regions (gray), the continuous prediction, and the cross entropy loss at each time point along the x-axis. **b.** The zoomed in comparison of a 12.5-minute period of this sleep record.



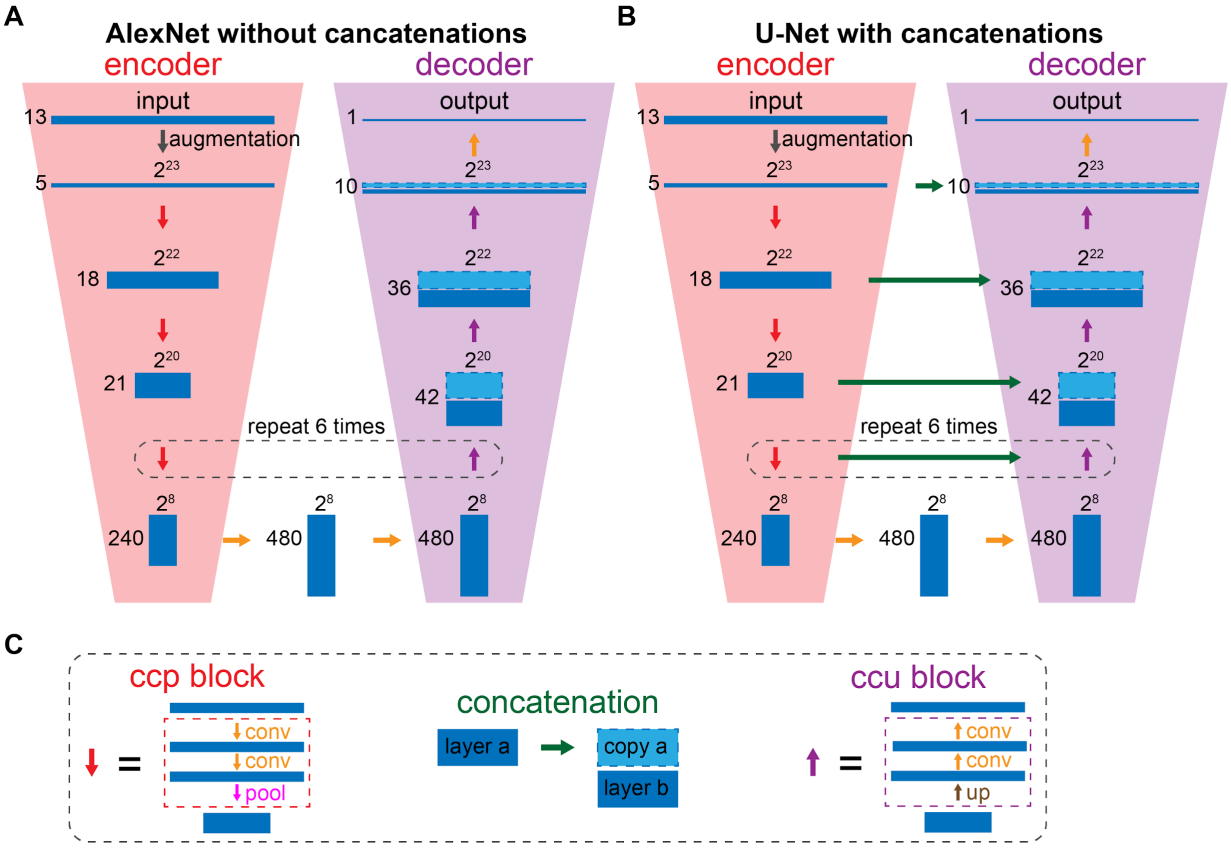
## Supplementary Figures

**Supplementary Figure 4.1 The prediction performances of models using various lengths of polysomnographic recordings as input.**
The **A**. AUROCs and **B**. AUPRCs of models using different lengths of polysomnographic recordings as input. From left to right, the length of input gradually increases from 4,096 (about 20 seconds) to 131,072 (about 11 minutes). Each color represents a model using one of the 13 polysomnographic signals. These signals correspond to the 13 channels from top to bottom in **Figure 4.1 - "Data"**: 1. F3-M2; 2. F4-M1; 3. C3-M2; 4. C4-M1; 5. O1-M2; 6. O2-M1; 7. E1-M2; 8. Chin; 9. ABD; 10. Chest; 11. Airflow; 12. $SaO_2$; 13. ECG. The dashed lines represent the baseline of random predictions in the AUROC space (baseline=0.500) and the AUPRC space (baseline=0.072).

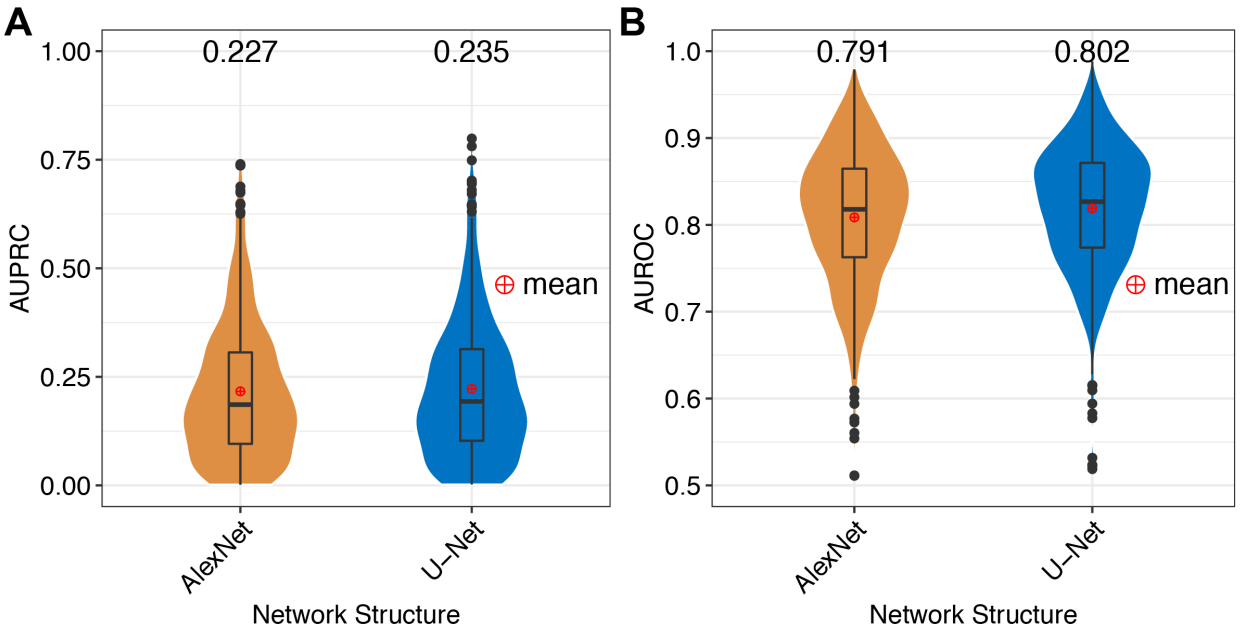


**Supplementary Figure 4.2 The comparison of network structures between AlexNet and U-Net.**

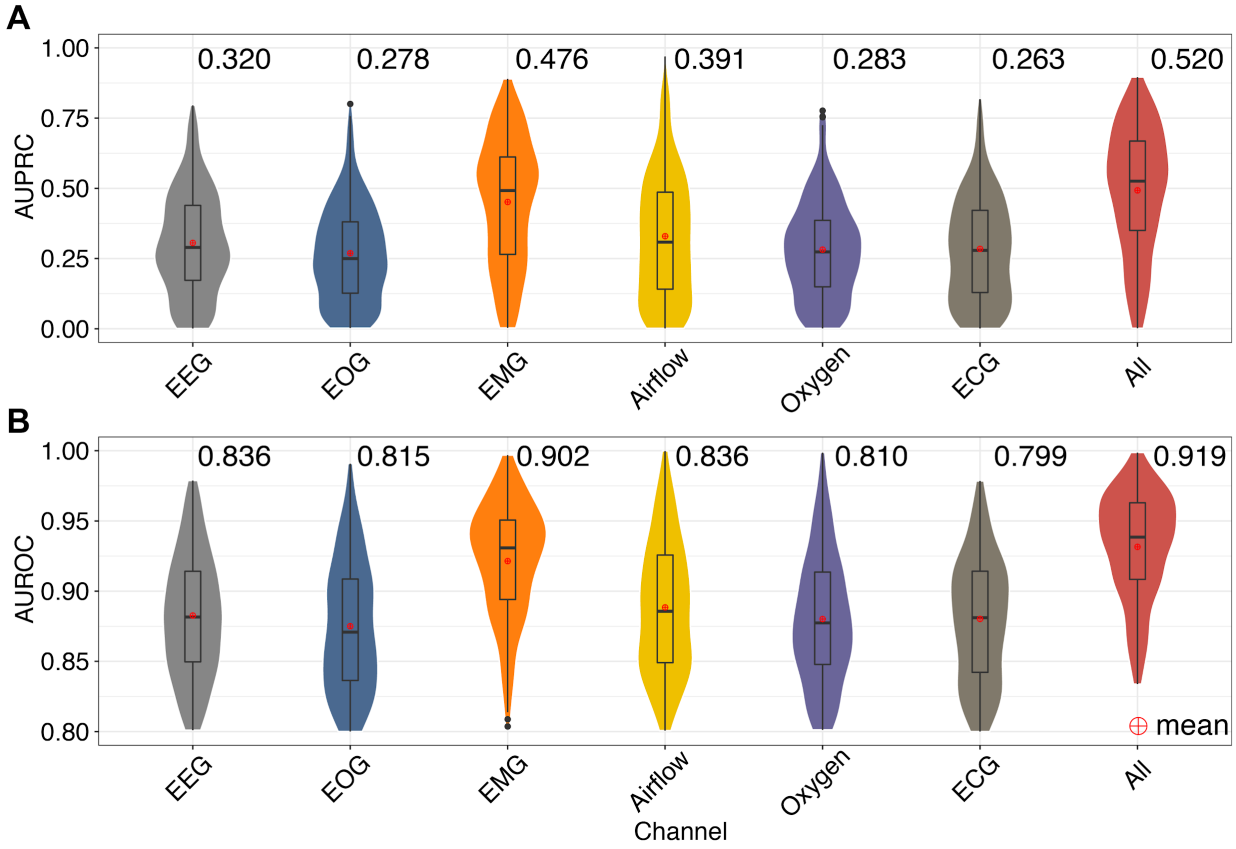
Both **A**. AlexNet and **B**. U-Net contain the encoder and the decoder, whereas AlexNet does not have the concatenation operations (horizontal green arrows) to directly transfer the information from the encoder to the decoder in each feature map. The feature map is the output of the ccp or ccu block. The convolutional/pooling/upscaling layers within the ccp and ccu blocks, and the concatenation operation are shown in **C**.



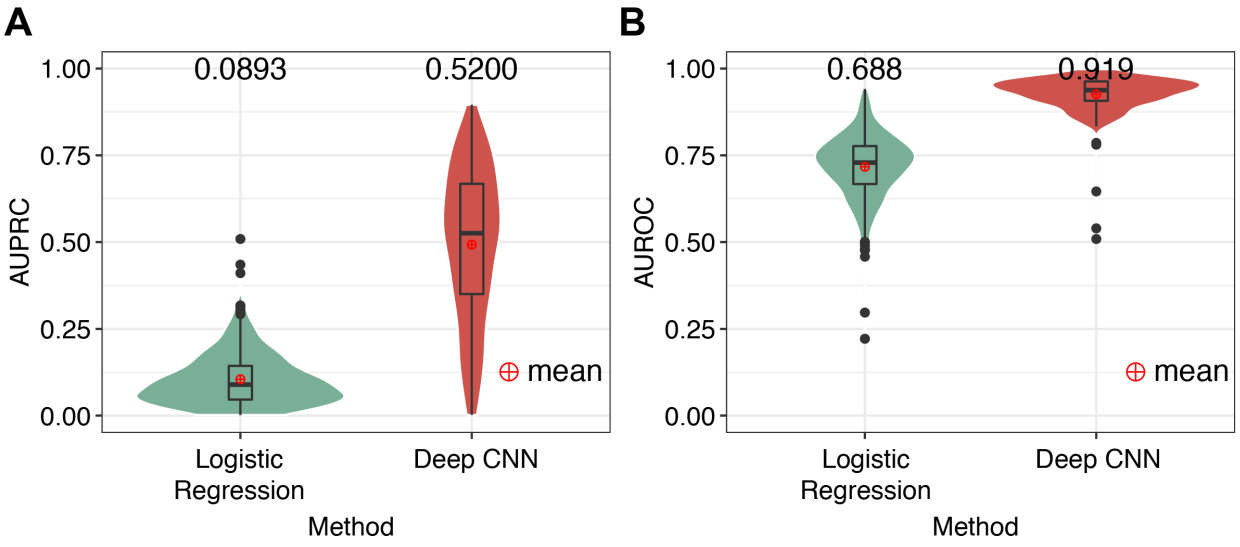
**Supplementary Figure 4.3 The performance comparison between AlexNet and U-Net.** The prediction **A**. AUPRCs and **B**. AUROCs of AlexNet and U-Net in segmenting sleep arousal regions. The only difference is that AlexNet does not have the concatenation operations between the encoder and the decoder (**Supplementary Figure 4.2**). U-Net outperformed AlexNet, in terms of both AUPRC and AUROC.



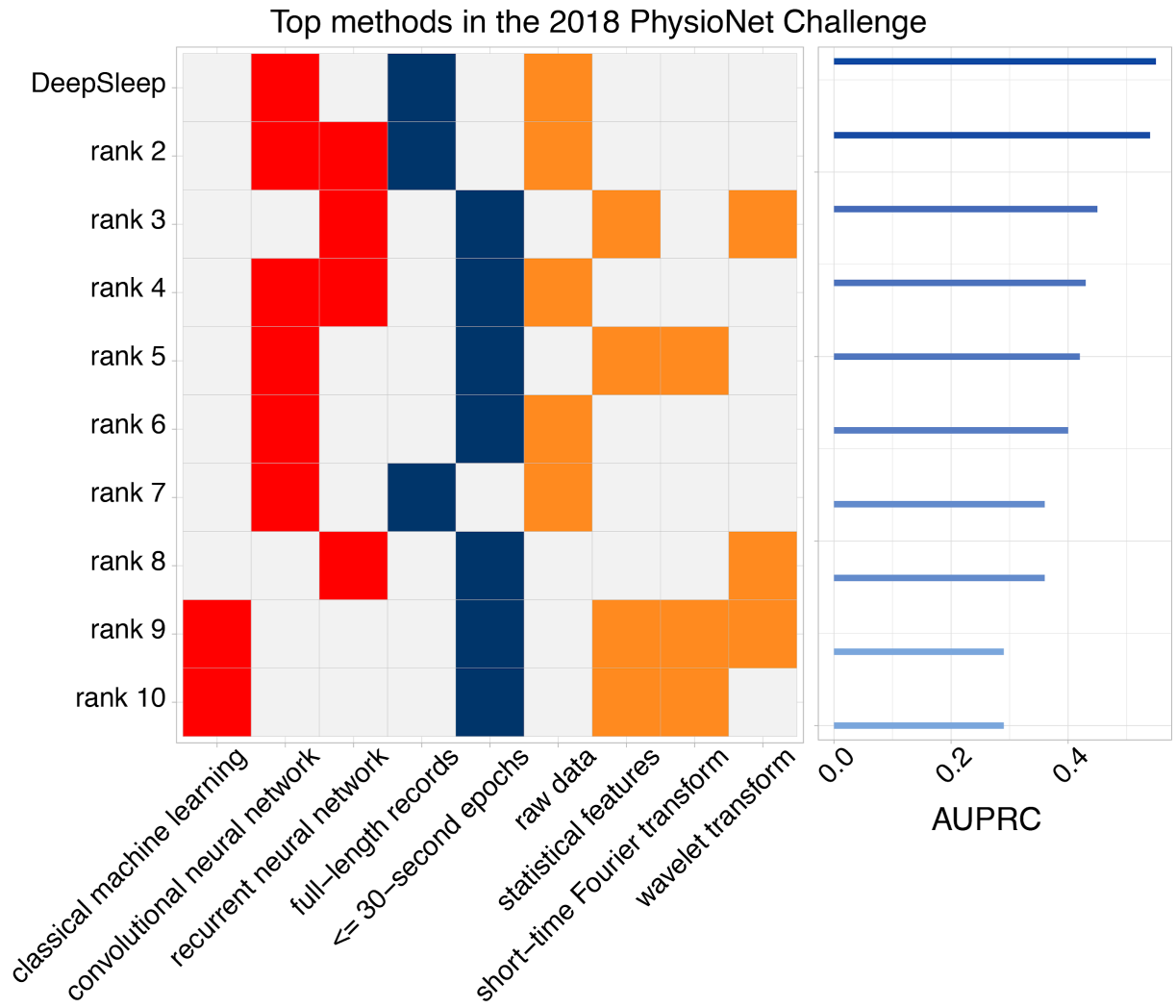
**Supplementary Figure 4.4 The performance comparison of models using different types of polysomnographic signals.**
From left to right, the first six categories are EEG (channel 1-6), EOG (channel 7), EMG (channel 8-10), Airflow (channel 11), saturation of Oxygen (channel 12) and ECG (channel 13). The last one, "All", represents the model using all these 13 channels as input. The prediction **A**. AUPRCs and **B**. AUROCs of models using different types of signals are shown in different colors. Of note, the model "All" using all 13 polysomnographic signals achieved the best performance.



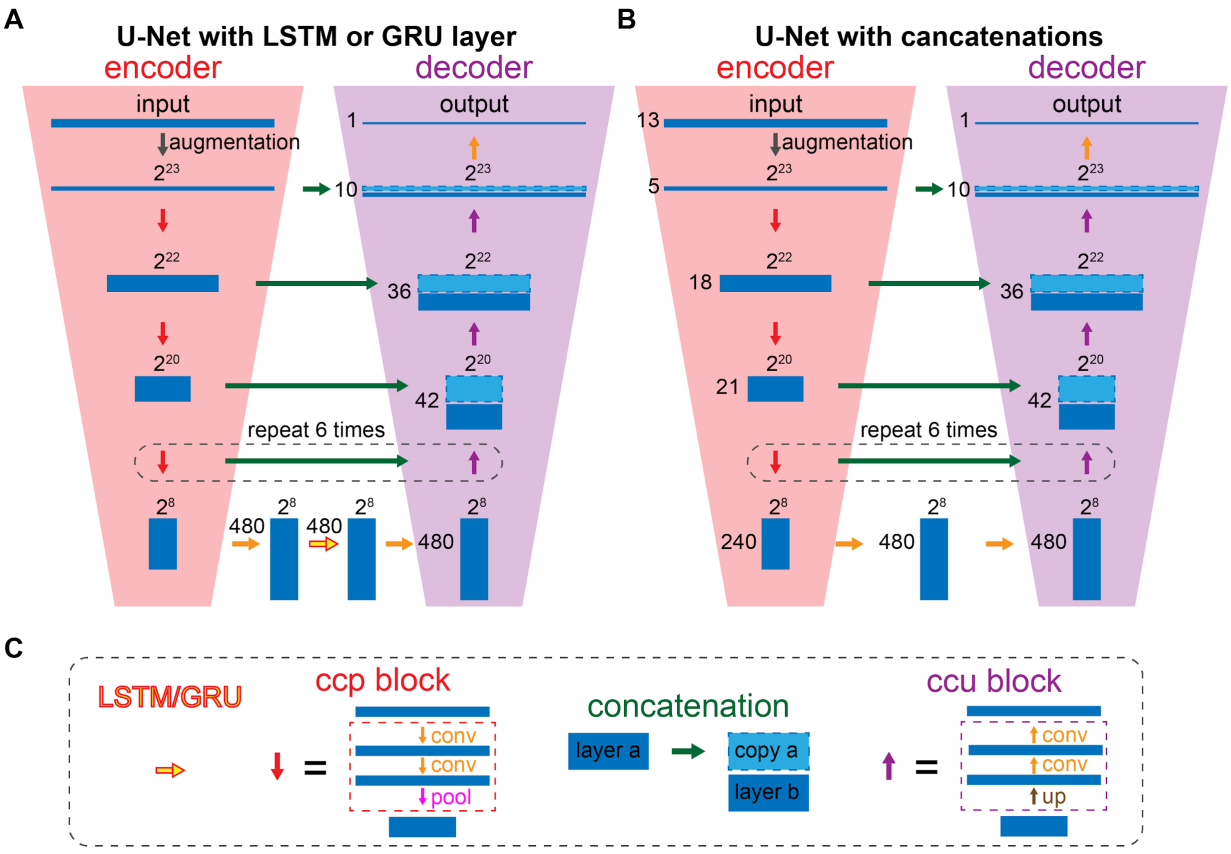
**Supplementary Figure 4.5 The performance comparison of deep CNN and the traditional approach of logistic regression.**
The prediction **A**. AUPRCs and **B**. AUROCs of deep convolutional neural network and logistic regression are shown in different colors. Clearly, the deep CNN had much higher performance in terms of both AUPRC and AUROC.



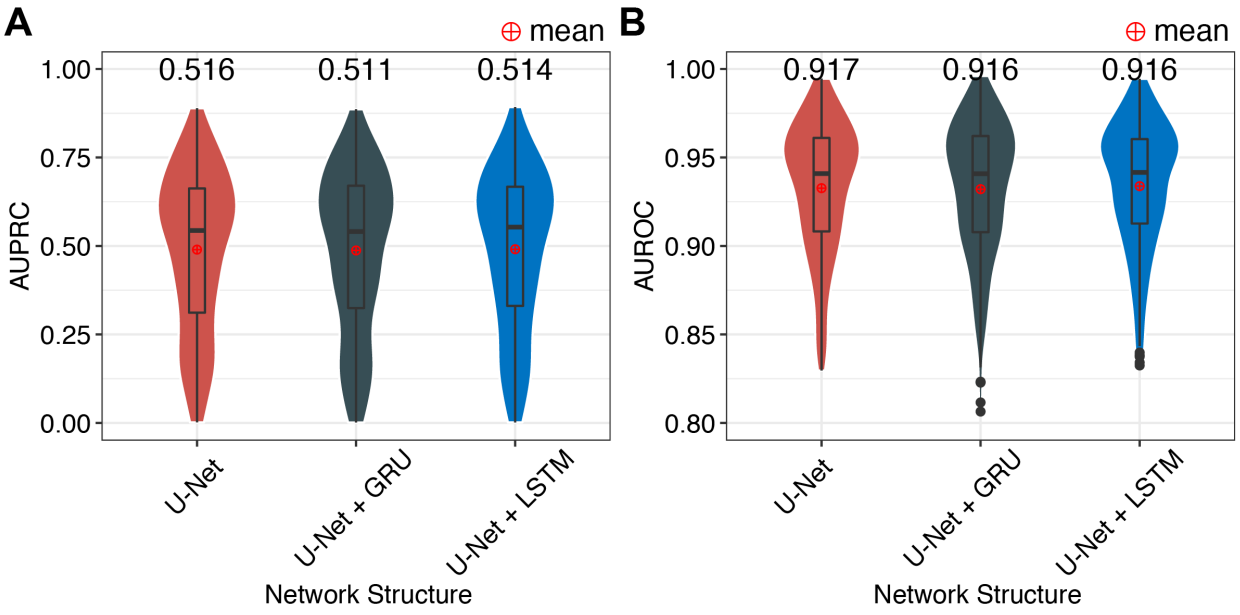
**Supplementary Figure 4.6 The comparison of the top 10 teams in the 2018 PhysioNet Challenge.**
In the left panel, top methods (Howe-Patterson et al. 2018; Már Þráinsson et al. 2018; He et al. 2018; Varga et al. 2018; Patane et al. 2018; Miller et al. 2018a; Warrick and Nabhan Homsi 2018; Bhattacharjee et al. 2018; Szalma et al. 2018) are compared in terms of machine learning models (red blocks), input length for models (blue blocks), and the types of input (orange blocks). In particular, the input are either raw polysomnogram data, or features extracted by statistical analysis, short-time Fourier transform, or wavelet transform. The corresponding prediction performances of these methods are shown in the right panel.



**Supplementary Figure 4.7 The comparison of U-Net structures with or without recurrent layers.**
Both **A**. U-Net with LSTM or GRU layer and **B**. U-Net has components of the encoder, the decoder and concatenation. The only difference lies at the bottom of U-Net, where a recurrent unit of LSTM or GRU layer is inserted. The the recurrent layer, the layers within the ccp and ccu blocks, and the concatenation operation are shown in **C**.

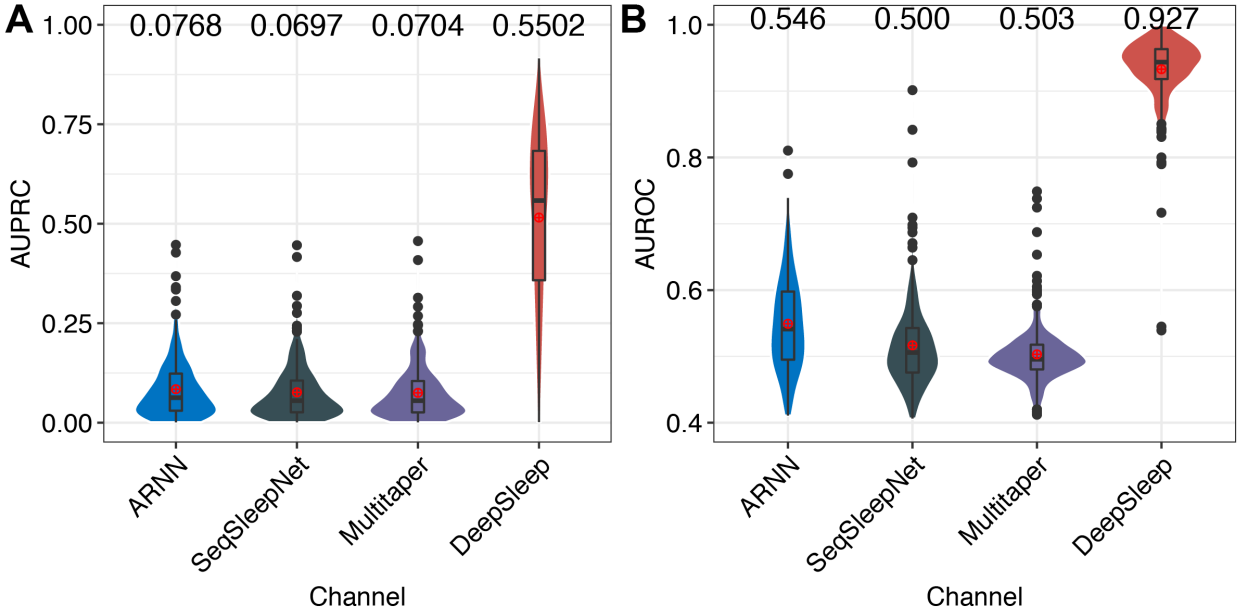


**Supplementary Figure 4.8 The performance comparison of different U-Net structures with or without recurrent units.**
The prediction **A**. AUPRCs and **B**. AUROCs of U-Net, U-Net with GRU and U-Net with LSTM are shown in different colors. The recurrent layer, GRU or LSTM, was implemented at the center of U-Net (**Supplementary Figure 4.5**). Adding the recurrent layer did not improve the performance. We used U-Net without recurrent layers as in our final model.

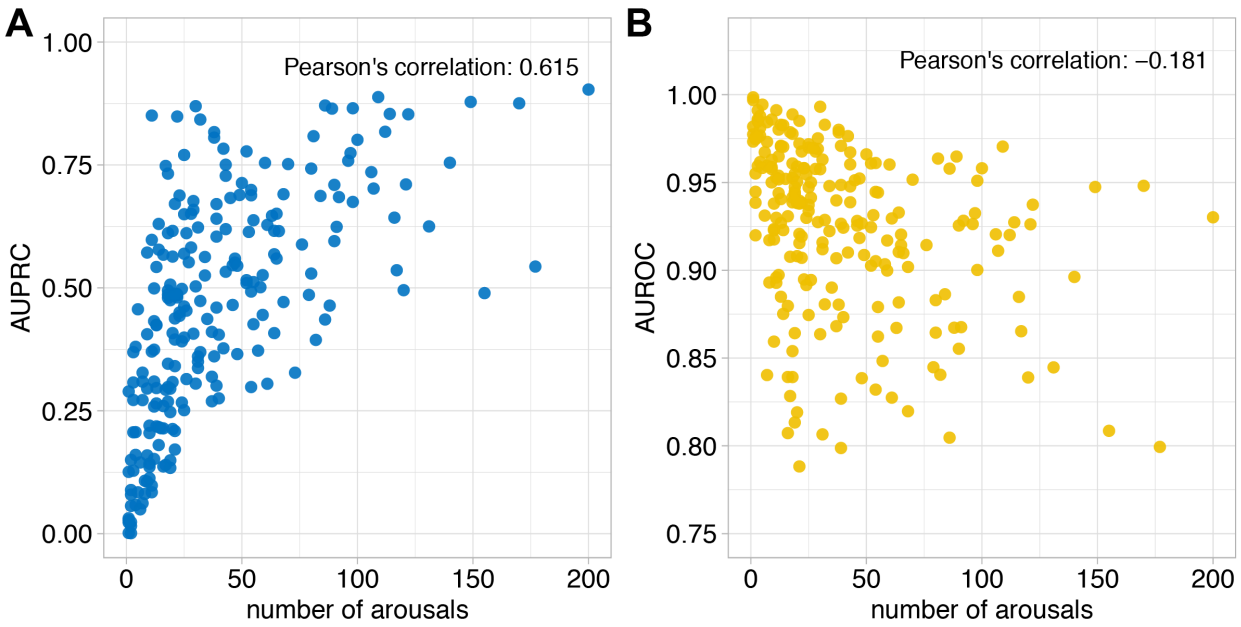


**Supplementary Figure 4.9 The comparison of DeepSleep with current methods for sleep staging.**

The prediction **A**. AUPRCs and **B**. AUROCs of (1) attention recurrent neural network (ARNN), (2) SeqSleepNet using features from short-time Fourier transform, (3) a method using features from Thomson's multitaper, and (4) our DeepSleep approach are shown in different colors.

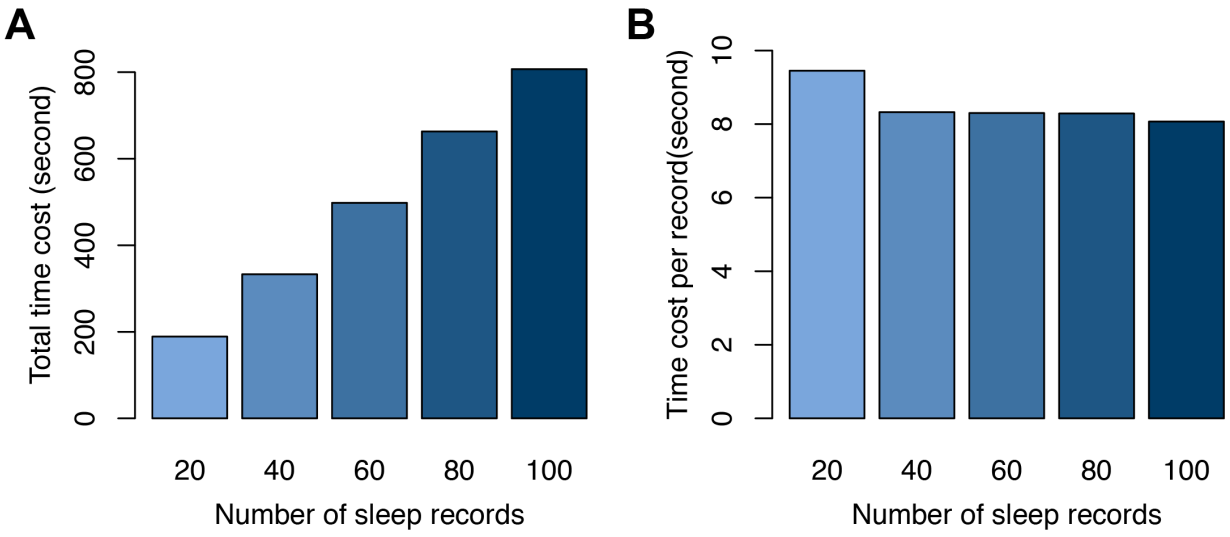


**Supplementary Figure 4.10 The relationship between prediction performance and the number of arousals.**
Each dot represents one sleep record. The prediction **A**. AUPRCs and **B**. AUROCs are shown by the y-axis. The AUPRC has a medium correlation with the number of sleep arousals.

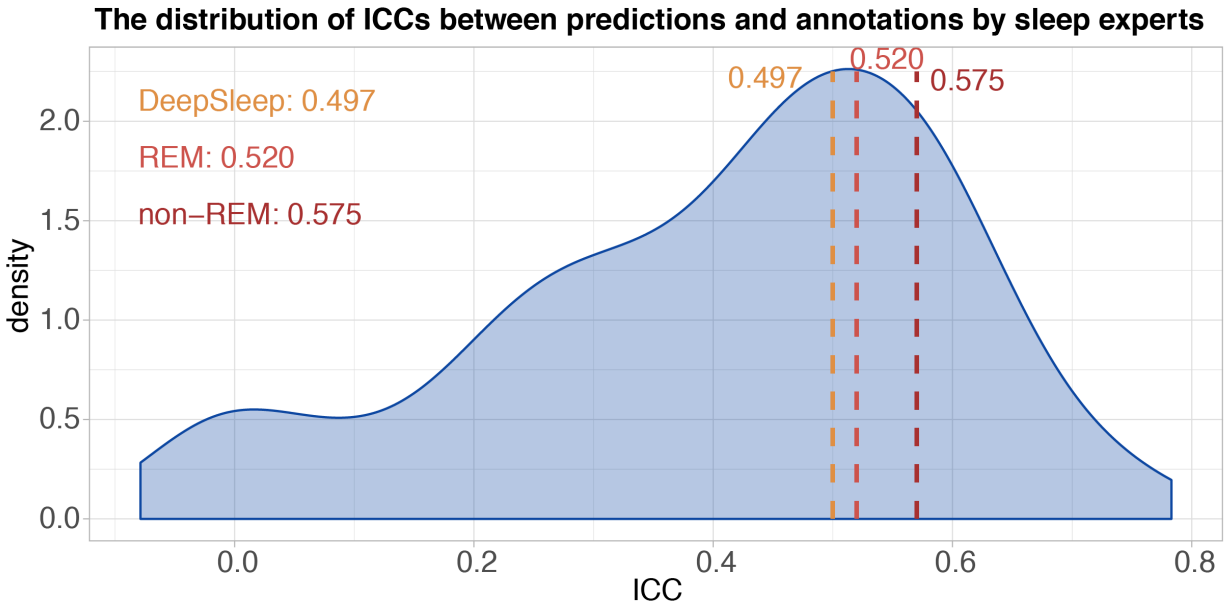


**Supplementary Figure 4.11 The runtimes for predicting sleep arousals at millisecond resolution.**

The **A**. total time cost and **B**. average time cost per sleep record are shown in bar plots. Notably, the average runtime per sleep record is less than 10 seconds and gradually decreases as the total number of records to be analyzed increases. This results from the overhead time of loading the large neural network models before the prediction step.



**Supplementary Figure 4.12 The distribution of Intraclass Correlation Coefficient values for all the test sleep records between our predictions and human labels.**
The overall ICC of DeepSleep is 0.497, which approaches the reported theoretical upper limit between 0.520 (arousals in REM regions) and 0.575 (arousal in non-REM regions).



## References


Alickovic E, Subasi A. 2018. Ensemble SVM Method for Automatic Sleep Stage Classification. *IEEE Trans Instrum Meas* **67**: 1258–1265.

Alvarez-Estevez D, Fernández-Varela I. 2019. Large-scale validation of an automatic EEG arousal detection algorithm using different heterogeneous databases. *Sleep Med*. https://linkinghub.elsevier.com/retrieve/pii/S1389945718303198.

Alvarez-Estévez D, Moret-Bonillo V. 2011. Identification of electroencephalographic arousals in multichannel sleep recordings. *IEEE Trans Biomed Eng* **58**: 54–63.

Anderer P, Gruber G, Parapatics S, Woertz M, Miazhynskaia T, Klosch G, Saletu B, Zeitlhofer J, Barbanoj MJ, Danker-Hopfe H, et al. 2005. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 x 7 utilizing the Siesta database. *Neuropsychobiology* **51**: 115–133.

Andreotti F, Phan H, Cooray N, Lo C, Hu MTM, De Vos M. 2018. Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks. *Conf Proc IEEE Eng Med Biol Soc* **2018**: 171–174.

Banks S, Dinges DF. 2007. Behavioral and physiological consequences of sleep restriction. *J Clin Sleep Med* **3**: 519–528.

Bartko JJ. 1966. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* **19**: 3–11.

Basner M, Griefahn B, Müller U, Plath G, Samel A. 2007. An ECG-based Algorithm for the Automatic Identification of Autonomic Activations Associated with Cortical Arousal. *Sleep* **30**: 1349–1361.

Bauters F, Rietzschel ER, Hertegonne KBC, Chirinos JA. 2016. The Link Between Obstructive Sleep Apnea and Cardiovascular Disease. *Curr Atheroscler Rep* **18**: 1.

Becq G, Charbonnier S, Chapotot F, Buguet A, Bourdon L, Baconnier P. 2005. Comparison Between Five Classifiers for Automatic Scoring of Human Sleep Recordings. In *Classification and Clustering for Knowledge Discovery* (eds. S. K. Halgamuge and L. Wang), Vol. 4 of *Studies in Computational Intelligence*, pp. 113–127, Springer Berlin Heidelberg, Berlin, Heidelberg.

Behera CK, Reddy TK, Behera L, Bhattacarya B. 2014. Artificial neural network based arousal detection from sleep electroencephalogram data. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pp. 458–462, IEEE.

Berry RB, Brooks R, Gamaldo C, Harding SM, Lloyd RM, Quan SF, Troester MT, Vaughn BV. 2017. AASM Scoring Manual Updates for 2017 (Version 2.4). *J Clin Sleep Med* **13**: 665–666.

Bhattacharjee T, Das D, Alam S, Rao M A V, Kumar Ghosh P, Ranjan Lohani A, Banerjee R, Dutta Choudhury A, Pal A. 2018. SleepTight: Identifying Sleep Arousals Using Inter and Intra-Relation of Multimodal Signals. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-245.pdf.

Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT. 2018. Expert-level sleep scoring with deep neural networks. *J Am Med Inform Assoc* **25**: 1643–1650.

Bonnet MH. 1985. Effect of Sleep Disruption on Sleep, Performance, and Mood. *Sleep* **8**: 11–19.

Bonnet MH. 1986. Performance and Sleepiness as a Function of Frequency and Placement of Sleep Disruption. *Psychophysiology* **23**: 263–271.

Buysse DJ. 2014. Sleep health: can we define it? Does it matter? *Sleep* **37**: 9–17.

Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. 2018. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Trans Neural Syst Rehabil Eng* **26**: 758–769.

Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. http://arxiv.org/abs/1406.1078 (Accessed December 2, 2018).

Cho S, Lee J, Park H, Lee K. 2005. Detection of arousals in patients with respiratory sleep disorders using a single channel EEG. *Conf Proc IEEE Eng Med Biol Soc* **3**: 2733–2735.

Davis J, Goadrich M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning - ICML '06* http://dx.doi.org/10.1145/1143844.1143874.

de Carli F, Nobili L, Gelcich P, Ferrillo F. 1999. A Method for the Automatic Detection of Arousals During Sleep. *Sleep* **22**: 561–572.

Ebrahimi F, Mikaeili M, Estrada E, Nazeran H. 2008. Automatic sleep stage classification based on EEG signals by using neural networks and wavelet packet coefficients. *Conf Proc IEEE Eng Med Biol Soc* **2008**: 1151–1154.

Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. 2018. Deep learning for healthcare applications based on physiological signals: A review. *Comput Methods Programs Biomed* **161**: 1–13.

Fernández-Varela I, Alvarez-Estevez D, Hernández-Pereira E, Moret-Bonillo V. 2017a. A simple and robust method for the automatic scoring of EEG arousals in polysomnographic recordings. *Comput Biol Med* **87**: 77–86.

Fernández-Varela I, Hernández-Pereira E, Álvarez-Estévez D, Moret-Bonillo V. 2017b.

Combining machine learning models for the automatic detection of EEG arousals. *Neurocomputing* **268**: 100–108.

Ford ES, Cunningham TJ, Giles WH, Croft JB. 2015. Trends in insomnia and excessive daytime sleepiness among U.S. adults from 2002 to 2012. *Sleep Med* **16**: 372–378.

Gangwisch JE, Malaspina D, Boden-Albala B, Heymsfield SB. 2005. Inadequate sleep as a risk factor for obesity: analyses of the NHANES I. *Sleep* **28**: 1289–1296.

Ghassemi M, Moody B, Lehman L-W, Song C, Li Q, Sun H, Westover B, Clifford G. 2018. You Snooze, You Win: The PhysioNet/Computing in Cardiology Challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-049.pdf.

Guan Y. 2019. Waking up to data challenges. *Nature Machine Intelligence* **1**: 67–67.

Halasz P, Terzano M, Parrino L, Bodizs R. 2004. The nature of arousal in sleep. *J Sleep Res* **13**: 1–23.

He R, Wang K, Liu Y, Zhao N, Yuan Y, Li Q, Zhang H. 2018. Identification of Arousals With Deep Neural Networks Using Different Physiological Signals. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-060.pdf.

Hillman D, Mitchell S, Streatfeild J, Burns C, Bruck D, Pezzullo L. 2018. The economic cost of inadequate sleep. *Sleep* **41**. http://dx.doi.org/10.1093/sleep/zsy083.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput* **9**: 1735–1780.

Howe-Patterson M, Pourbabaee B, Benard F. 2018. Automated Detection of Sleep Arousals From Polysomnography Data Using a Dense Convolutional Neural Network. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-232.pdf.

Hsu Y-L, Yang Y-T, Wang J-S, Hsu C-Y. 2013. Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing* **104**: 105–114.

Huang G, Chu C-H, Wu X. 2018. A Deep Learning-Based Method for Sleep Stage Classification Using Physiological Signal: International Conference, ICSH 2018, Wuhan, China, July 1–3, 2018, Proceedings. In *Smart Health* (eds. H. Chen, Q. Fang, D. Zeng, and J. Wu), Vol. 10983 of *Lecture Notes in Computer Science*, pp. 249–260, Springer International Publishing, Cham.

Kingma DP, Ba J. 2014. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980 (Accessed December 2, 2018).

Krizhevsky A, Sutskever I, Hinton GE. 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM* **60**: 84–90.

Kronholm E, Partonen T, Härmä M, Hublin C, Lallukka T, Peltonen M, Laatikainen T. 2016. Prevalence of insomnia-related symptoms continues to increase in the Finnish working-age population. *J Sleep Res* **25**: 454–457.

Kuna ST, Benca R, Kushida CA, Walsh J, Younes M, Staley B, Hanlon A, Pack AI, Pien GW, Malhotra A. 2013. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep* **36**: 583–589.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521**: 436–444.

Lewis NCS, Jones H, Ainslie PN, Thompson A, Marrin K, Atkinson G. 2015. Influence of nocturnal and daytime sleep on initial orthostatic hypotension. *Eur J Appl Physiol* **115**: 269–276.

Li H, Li T, Quang D, Guan Y. 2018. Network Propagation Predicts Drug Synergy in Cancers. *Cancer Res* **78**: 5446–5457.

Li H, Quang D, Guan Y. 2019. Anchor: trans-cell type prediction of transcription factor binding sites. *Genome Res* **29**: 281–292.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. 2017. A survey on deep learning in medical image analysis. *Med Image Anal* **42**: 60–88.

Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB. 2016. Prevalence of Healthy Sleep Duration among Adults — United States, 2014. *MMWR Morb Mortal Wkly Rep* **65**: 137–141.

Malafeev A, Laptev D, Bauer S, Omlin X, Wierzbicka A, Wichniak A, Jernajczyk W, Riener R, Buhmann J, Achermann P. 2018. Automatic Human Sleep Stage Scoring Using Deep Neural Networks. *Front Neurosci* **12**: 781.

Már Þráinsson H, Ragnarsdóttir H, Fannar Kristjansson G, Marinósson B, Finnsson E, Gunnlaugsson E, Ægir Jónsson S, Skírnir Ágústsson J, Helgadóttir H. 2018. Automatic Detection of Target Regions of Respiratory Effort-Related Arousals Using Recurrent Neural Networks. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-126.pdf.

Miller D, Ward A, Bambos N. 2018a. Automatic Sleep Arousal Identification From Physiological Waveforms Using Deep Learning. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-242.pdf.

Miller MA, Kruisbrink M, Wallace J, Ji C, Cappuccio FP. 2018b. Sleep duration and

incidence of obesity in infants, children, and adolescents: a systematic review and meta-analysis of prospective studies. *Sleep* **41**. http://dx.doi.org/10.1093/sleep/zsy018.

Mukherjee S, Patel SR, Kales SN, Ayas NT, Strohl KP, Gozal D, Malhotra A, American Thoracic Society ad hoc Committee on Healthy Sleep. 2015. An Official American Thoracic Society Statement: The Importance of Healthy Sleep. Recommendations and Future Priorities. *Am J Respir Crit Care Med* **191**: 1450–1458.

Nair V, Hinton GE. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 807–814, Omnipress.

Okun ML, Mancuso RA, Hobel CJ, Schetter CD, Coussons-Read M. 2018. Poor sleep quality increases symptoms of depression and anxiety in postpartum women. *J Behav Med* **41**: 703–710.

Olsen M, Schneider LD, Cheung J, Peppard PE, Jennum PJ, Mignot E, Sorensen HBD. 2018. Automatic, electrocardiographic-based detection of autonomic arousals and their association with cortical arousals, leg movements, and respiratory events in sleep. *Sleep* **41**. http://dx.doi.org/10.1093/sleep/zsy006.

Paiva T, Gaspar T, Matos MG. 2015. Sleep deprivation in adolescents: correlations with health complaints and health-related quality of life. *Sleep Med* **16**: 521–527.

Patanaik A, Ong JL, Gooley JJ, Ancoli-Israel S, Chee MWL. 2018. An end-to-end framework for real-time automatic sleep stage classification. *Sleep* **41**. http://dx.doi.org/10.1093/sleep/zsy041.

Patane A, Ghiasi S, Pasquale Scilingo E, Kwiatkowska M. 2018. Automated Recognition of Sleep Arousal Using Multimodal and Personalized Deep Ensembles of Neural Networks. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-332.pdf.

Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. 2018a. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. *IEEE Trans Biomed Eng*. http://dx.doi.org/10.1109/TBME.2018.2872652.

Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. 2019. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Trans Neural Syst Rehabil Eng*. http://dx.doi.org/10.1109/TNSRE.2019.2896659.

Phan H, Andreotti F, Cooray N, Chen OY, Vos MD. 2018b. Automatic Sleep Stage Classification Using Single-Channel EEG: Learning Sequential Features with Attention-Based Recurrent Neural Networks. *Conf Proc IEEE Eng Med Biol Soc* **2018**: 1452–1455.

Ronneberger O, Fischer P, Brox T. 2015. U-Net: Convolutional Networks for Biomedical

Image Segmentation. In *Lecture Notes in Computer Science*, pp. 234–241.

Ronzhina M, Janoušek O, Kolářová J, Nováková M, Honzík P, Provazník I. 2012. Sleep scoring using artificial neural networks. *Sleep Med Rev* **16**: 251–263.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. *Nature* **323**: 533–536.

Shahrbabaki SS, Dissanayaka C, Patti CR, Cvetkovic D. 2015. Automatic detection of sleep arousal events from polysomnographic biosignals. In *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4, IEEE.

Sharma R, Pachori RB, Upadhyay A. 2017. Automatic sleep stages classification based on iterative filtering of electroencephalogram signals. *Neural Comput Appl* **28**: 2959–2978.

Shen D, Wu G, Suk H-I. 2017. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng* **19**: 221–248.

Shmiel O, Shmiel T, Dagan Y, Teicher M. 2009. Data mining techniques for detection of sleep arousals. *J Neurosci Methods* **179**: 331–337.

Sors A, Bonnet S, Mirek S, Vercueil L, Payen J-F. 2018. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomed Signal Process Control* **42**: 107–114.

Sousa T, Cruz A, Khalighi S, Pires G, Nunes U. 2015. A two-step automatic sleep stage classification method with dubious range detection. *Comput Biol Med* **59**: 42–53.

St-Onge M-P. 2017. Sleep-obesity relation: underlying mechanisms and consequences for treatment. *Obes Rev* **18 Suppl 1**: 34–39.

St-Onge M-P, Grandner MA, Brown D, Conroy MB, Jean-Louis G, Coons M, Bhatt DL, American Heart Association Obesity, Behavior Change, Diabetes, and Nutrition Committees of the Council on Lifestyle and Cardiometabolic Health; Council on Cardiovascular Disease in the Young; Council on Clinical Cardiology; and Stroke Council. 2016. Sleep Duration and Quality: Impact on Lifestyle Behaviors and Cardiometabolic Health: A Scientific Statement From the American Heart Association. *Circulation* **134**: e367–e386.

Sugi T, Kawana F, Nakamura M. 2009. Automatic EEG arousal detection for sleep apnea syndrome. *Biomed Signal Process Control* **4**: 329–337.

Sun H, Jia J, Goparaju B, Huang G-B, Sourina O, Bianchi MT, Westover MB. 2017. Large-Scale Automated Sleep Staging. *Sleep* **40**. http://dx.doi.org/10.1093/sleep/zsx139.

Sun Y, Wang B, Jin J, Wang X. 2018. Deep Convolutional Network Method for Automatic Sleep Stage Classification Based on Neurophysiological Signals. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*

*(CISP-BMEI)*, pp. 1–5, IEEE.

Supratak A, Dong H, Wu C, Guo Y. 2017. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Trans Neural Syst Rehabil Eng* **25**: 1998–2008.

Suzuki Y, Sato M, Shiokawa H, Yanagisawa M, Kitagawa H. 2017. MASC: Automatic Sleep Stage Classification Based on Brain and Myoelectric Signals. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)* http://dx.doi.org/10.1109/icde.2017.218.

Szalma J, Bánhalmi A, Bilicki V. 2018. Detection of Respiratory Effort-Related Arousals Using a Hidden Markov Model and Random Decision Forest. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-089.pdf.

Takahashi M. 2012. Prioritizing sleep for healthy work schedules. *J Physiol Anthropol* **31**: 6.

Ting L, Malhotra A. 2005. Disorders of sleep: an overview. *Prim Care* **32**: 305–18, v.

Tobaldini E, Costantino G, Solbiati M, Cogliati C, Kara T, Nobili L, Montano N. 2017. Sleep, sleep deprivation, autonomic nervous system and cardiovascular diseases. *Neurosci Biobehav Rev* **74**: 321–329.

Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. 2016. Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks. http://arxiv.org/abs/1610.01683 (Accessed February 27, 2019).

't Wallant DC, Muto V, Gaggioni G, Jaspar M, Chellappa SL, Meyer C, Vandewalle G, Maquet P, Phillips C. 2016. Automatic artifacts and arousals detection in whole-night sleep EEG recordings. *J Neurosci Methods* **258**: 124–133.

Varga B, Görög M, Hajas P. 2018. Using Auxiliary Loss to Improve Sleep Arousal Detection With Neural Network. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-247.pdf.

Vitiello MV. 2018. The interrelationship of sleep and depression: new answers but many questions remain. *Sleep Med* **52**: 230–231.

Warrick P, Nabhan Homsi M. 2018. Sleep Arousal Detection From Polysomnography Using the Scattering Transform and Recurrent Neural Networks. In *2018 Computing in Cardiology Conference (CinC)*, Vol. 45 of *Computing in Cardiology Conference (CinC)*, Computing in Cardiology http://www.cinc.org/archives/2018/pdf/CinC2018-368.pdf.

Zhang J, Wu Y, Bai J, Chen F. 2016. Automatic sleep stage classification based on sparse deep belief net and combination of multiple classifiers. *Transactions of the Institute of*

*Measurement and Control* **38**: 435–451.

1992. EEG arousals: scoring rules and examples: a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association. *Sleep* **15**: 173–184.

# CHAPTER V

## Summary and Conclusion

### Summary and future directions

Advancements in unsupervised and supervised machine learning algorithms have provided new insights into data patterns and generated prediction models for practical usage. More importantly, these machine learning approaches have been evolving our understanding of various physiological systems. In this dissertation, I have contributed to developing cutting-edge computational models to address multiple problems, including the comparison of evolutionarily-related protein families, modelling the relationship of multi-omics in cancers, and automatic segmentation of sleep arousals.

In Chapter II, I performed PCA of existing crystallographic structures of three GTPase families. In addition to the canonical GTP and GDP forms, I found two new conformational clusters representing the GEF-bound state in Gαt/i and the "state 1" in Ras. By comparing the Ras PCA to PCA of Gαt/i and EF-Tu, I revealed common nucleotide dependent collective deformations of SI and SII across G protein families. I further performed extensive MD simulations and network analyses, which reveal common nucleotide-associated conformational dynamics in Ras, Gαt and EF-Tu. Specifically, these three systems have stronger intra-lobe1 (PL–SI and PL–SII) and inter-lobe (SII–SIII/α3) couplings in the GTP-bound state. Meanwhile, through the network comparison approach, I further identified residue-wise determinants of commonalities and specificities across families. Mutations of identified distal residues display decreased coupling strength in SI–PL.

Besides the key residues that are common in the three systems, residues mediating inter-lobe couplings only in Gαt and EF-Tu are identified. Importantly, some of the highlighted mutants have been reported to have functional effects by *in vitro* experiments. This study provides insights into the atomistic mechanisms of these altered protein functions. Overall, separation of functionally conserved and specific residues in conformational dynamics provides us unprecedented insights into protein evolution and engineering. In addition to molecular switches, this approach can be broadly used in the comparison of multiple protein families in the future. For example, molecular motors such as ATPases have a similar nucleotide-dependent functional circle. A structural dynamic comparison between GTPases and ATPases will potential reveal key determinants of these two evolutionarily-related enzyme superfamilies.

In Chapter III, I created a machine learning algorithm for predicting protein abundances from the mRNA levels. This approach pinpointed the relative importance of the innate correlations between mRNA and protein levels, and the global direct and indirect interactions across all genes in controlling the expression level of a protein. Based on the intuition that the regulatory mechanism may be shared across different cancer types, I built a new model that shares parameters across breast and ovarian cancers, and improved prediction performance in both cancers. This revealed a new, unexplored aspect of the regulatory mechanism that is previously not captured in single tissue modelling approaches. Pathway analysis and gene-gene interaction network indicated that functionally different gene sets had different predictability profiles and regulatory powers. In sum, this approach offers a new field standard for protein abundance prediction across cancer patients,

and the key features used in our model and the innovation of transfer learning across two cancer types will be instructive for future method development and protein expression regulatory mechanism exploration. In addition, decoding the determinants modulating protein phosphorylation is also crucial for understanding the regulatory mechanisms underlying cancers. Similar ideas can be potentially used in predicting the phosphoproteomic profiles from the corresponding genomic, transcriptomic, and proteomic data in the future.

In Chapter IV, I investigated a novel deep learning approach, DeepSleep, for automatic detection of sleep arousals. I built a deep convolutional neural network (CNN) to capture long-range and short-range interdependencies between timepoints across an entire sleep record. Information at different resolutions and scales was integrated to improve the performance. I found that similar EEG and EMG channels were interchangeable, which was used as a special augmentation in our approach. Compared with the theoretical upper limit calculated from annotation replicates by different sleep experts, DeepSleep achieved near-perfect detection of sleep arousals at millisecond resolution, approximating human performance. Furthermore, a clear advantage of computational approaches lies in the annotations for the boundary regions between arousal and sleep. Since current sleep annotations are binary only, it would be a more accurate and appropriate approach to introduce the probability of the annotation confidence, especially at the boundary regions. Machine learning approaches such as DeepSleep naturally provide the continuous predictions for each time point. It would be interesting to see improved annotation systems using continuous values instead of binary labels. A simple approach could be directly integrating the computer predictions with

annotations by human sleep experts. The proposed annotation systems would provide more accurate information for the diagnosis of sleep disorders and the evaluation of sleep quality in the future.

**Conclusion**

In the big data era, data explosion brings up critical problems - how to efficiently distinguish true signals from noises and artifacts, build high-performing prediction models for practical usage, and ultimately reveal new insights from the computational perspective. In this dissertation, I apply a variety of machine learning algorithms to multiple problems. I believe these improved approaches will facilitate data analysis in the fields of structural dynamic comparison of proteins, proteogenomics, and signal processing of sleep recordings.