The Influence of DNA Sequence-Derived Features across the 'omics scales

Gregory Parkes, Mahesan Niranjan

Email: g.m.parkes@soton.ac.uk

UNIVERSITY OF Southampton

EPSRC

Overview

- What are sequence-derived features?
- What are ‘omics or multi-’omics?
- Exposure to examples and challenges with SDFs
- Use of sequence-derived features in yeast for protein prediction
- Use of sequence-derived features in human cell cycle study
- Expanding the number of SDFs, use as proxy
- SDFs contribute useful ‘static’ information into model

What are **sequence-derived features?**

DATABASE
(Uniprot, Refseq,
NCBI, EBI)

Extract sequence

Extract categorical metadata

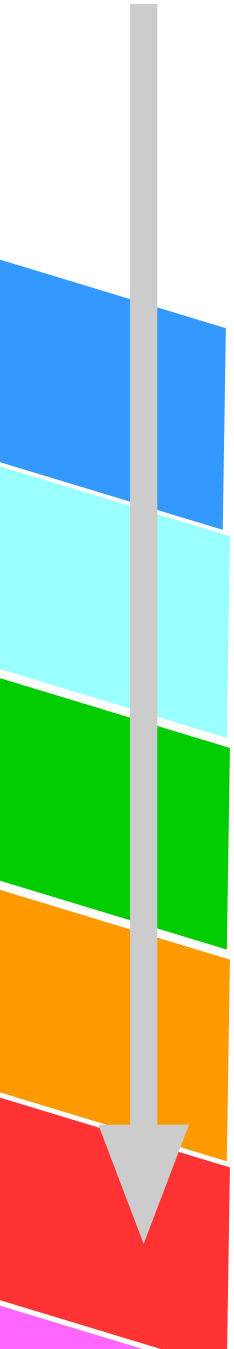
Feature engineering

P01920 Beta strand;Chain;Disul..
P18440 Active site;Beta strand;B.
Q15366 Glycosylation;Alternativ..
P31327 Transit peptide;Turn;Hel..
P00740 Active site;Mutagenesis;..
Q9BXJ4 Chain;Domain;Signal p..
Q9UBT6 Helix;Metal binding;Nat..
O75147 Beta strand;Region;Mod.

Preprocess

[1] Illumina website. https://www.illumina.com/techniques/sequencing/dna-sequencing.html. Taken with permission.

Analysis and **integration** of multiple levels of expression



Genomics

DNA base sequence

Illumina

DNA sequencing

Epigenomics

DNA methylation

3D chromatin structuring

Histone modification

ATAC-seq

Transcriptomics

mRNA expression

microarray

scRNA-seq

ncRNAs

microRNAs

RNA-seq

RNA interactions

Proteomics

peptides

Mass spectrometry

Protein structure

protein expression

Assays

Protein interactions

Metabolomics

metabolites

substrate

GC/LC-MS

Protein fragments

NMR spectroscopy

Phenomics

Phenotype

Application to disease/cancer

Disease stratification

Inherent challenges with using data from **different expression levels**

DNA

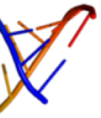
SEQUENCE
ATGTGGGCTTATAAATGTGCGGTACCAGCCCTGTGAATGAGTGGCCGTATACGCTCGATCGCCTGATC
STABLE

GC% Epigenetic markings, Chromosome, Locus

RNA

SEQUENCE
AUGACCUGACCUUCACUCUGGCGGCACAAAUGGUCCAUAUGCGCCAUCGUACGUGCGGCUAGCU
UNSTABLE (10h t1/2)

mRNA concentration, Codon bias, sequence length,*



Protein

SEQUENCE
MKKNRRSRWLVATAGKRSPSFLLIVAARERW**STOP**
UNSTABLE (20-40h t1/2)

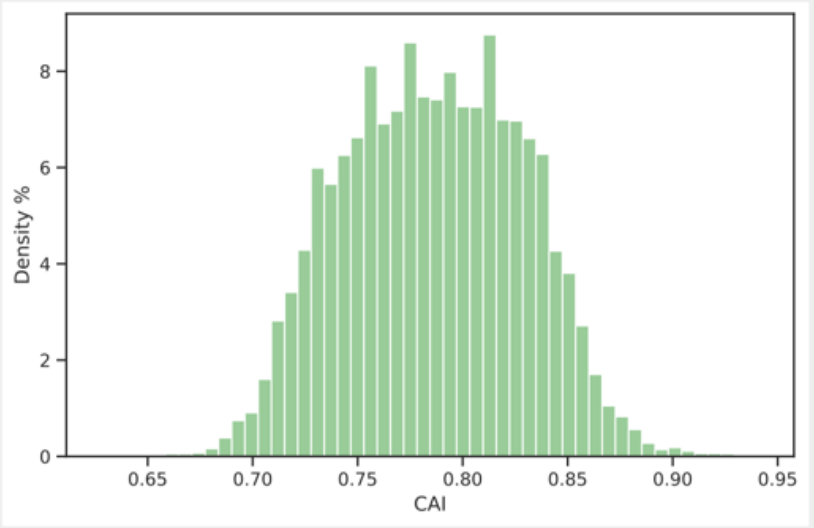
Translation concentration, Protein* concentration, a*

Example: Making a **sequence-derived feature**; RCB & MFE

Relative Codon Usage/Bias [1]

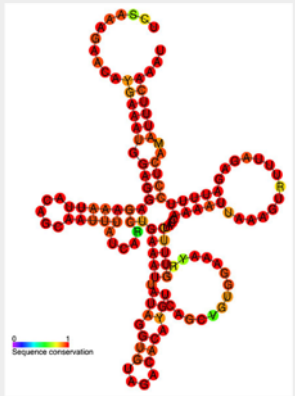
$$RCB_{xyz} = \frac{f(x,y,z)}{f_1(x)f_2(y)f_3(z)},$$

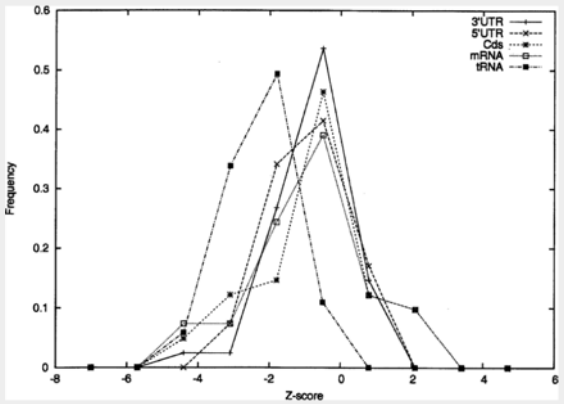
$$RCBS = \left(\prod_{l=1}^{L} RCB_{xyz}(l)\right)^{1/L} - 1$$

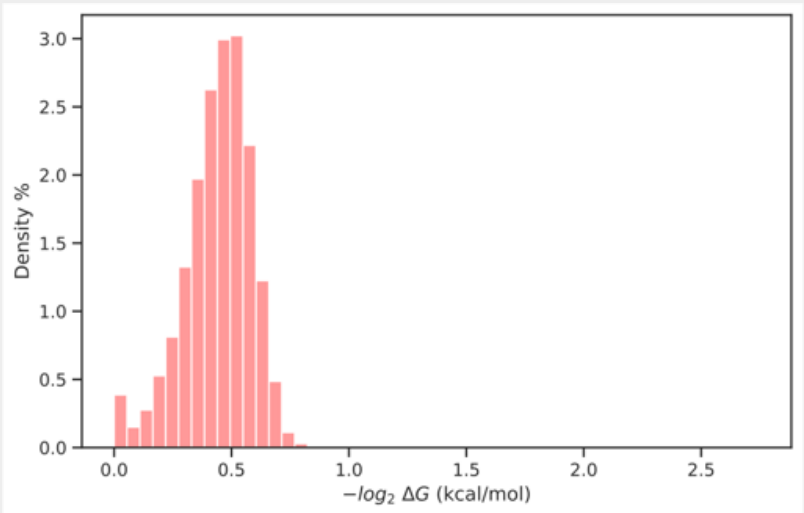


[1] Fox, J. M., & Erill, I. (2010). Relative codon adaptation: a generic codon bias index for prediction of gene expression. DNA research : an international journal for rapid publi

Minimum Free Energy (MFE) [2]

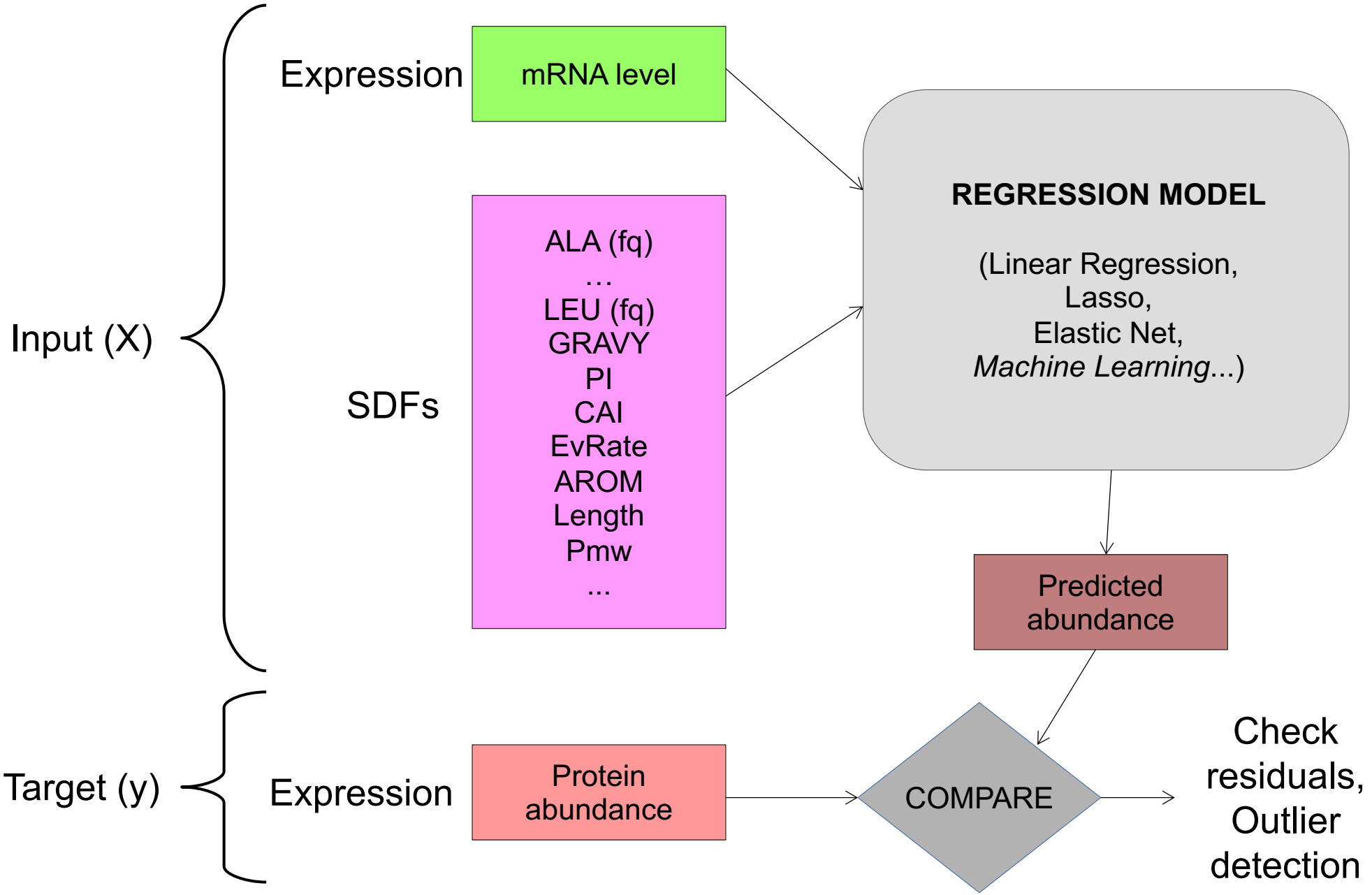




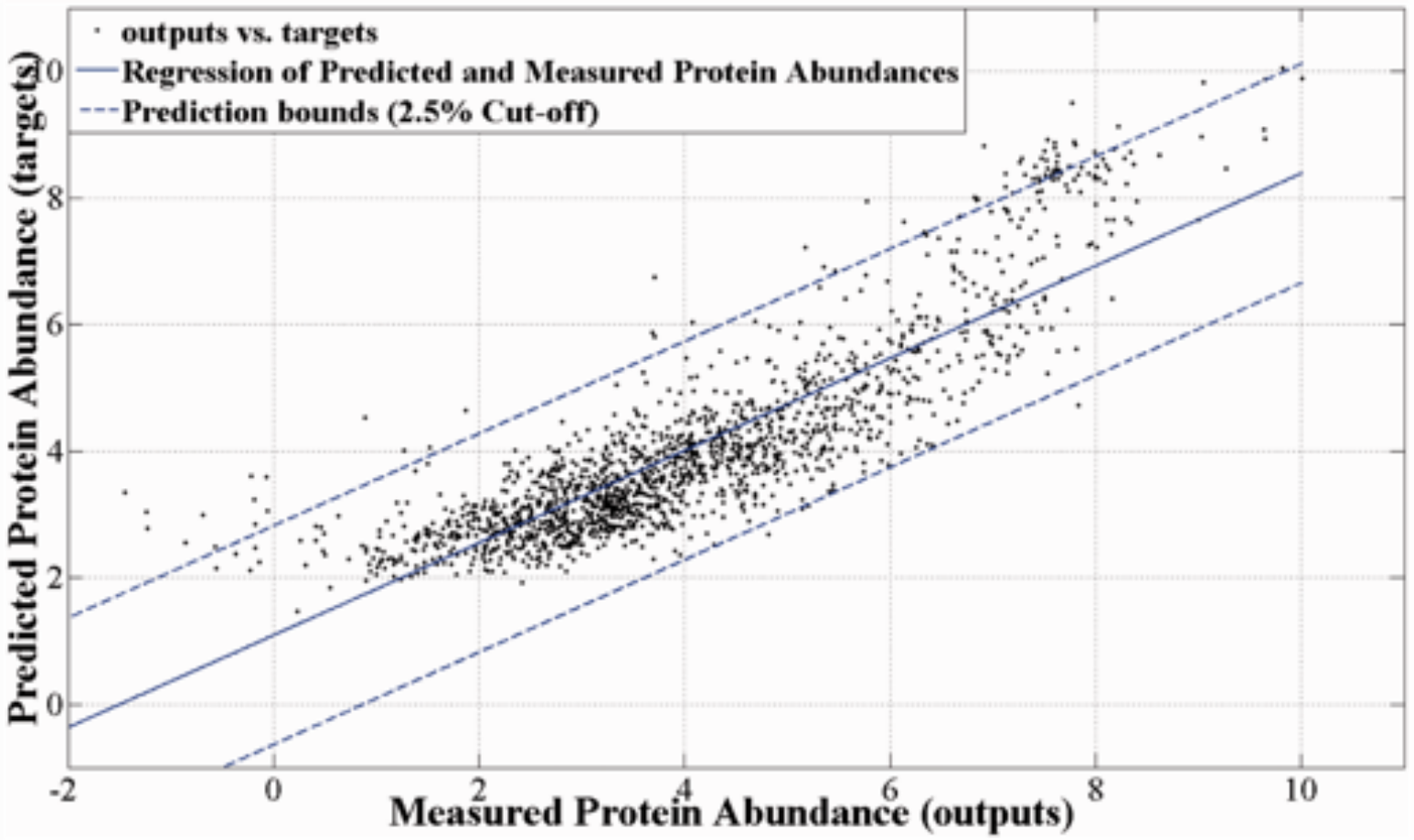


[2] Ringnér M, Krogh M (2005) Folding Free Energies of 5′-UTRs Impact Post-Transcriptional Regulation on a Genomic Scale in Yeast. PLOS Computational Biology 1(7): e72.

Case study 1: Use of SDFs in modelling protein abundance in yeast

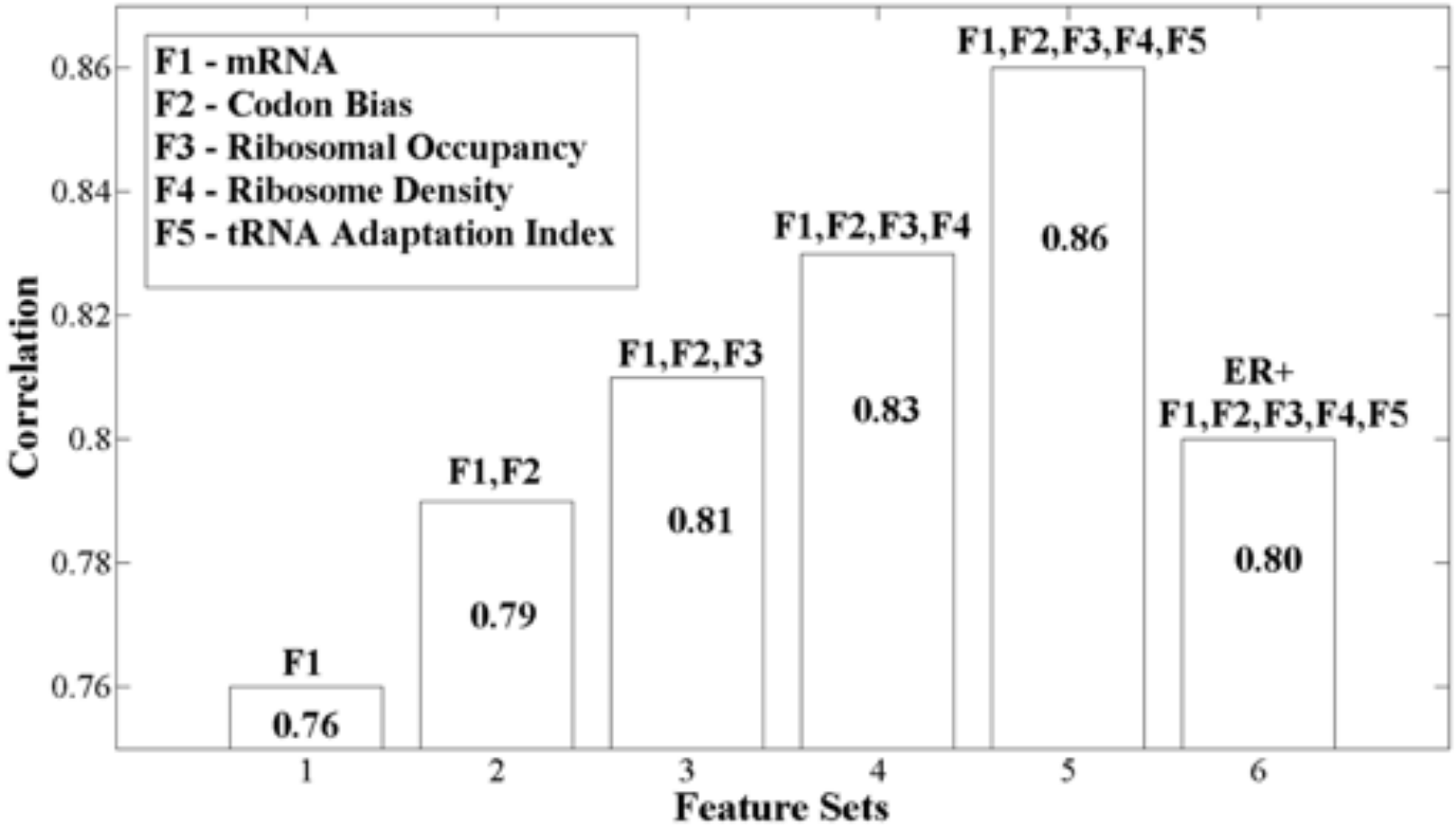


Linear models of mRNA proxy for protein level



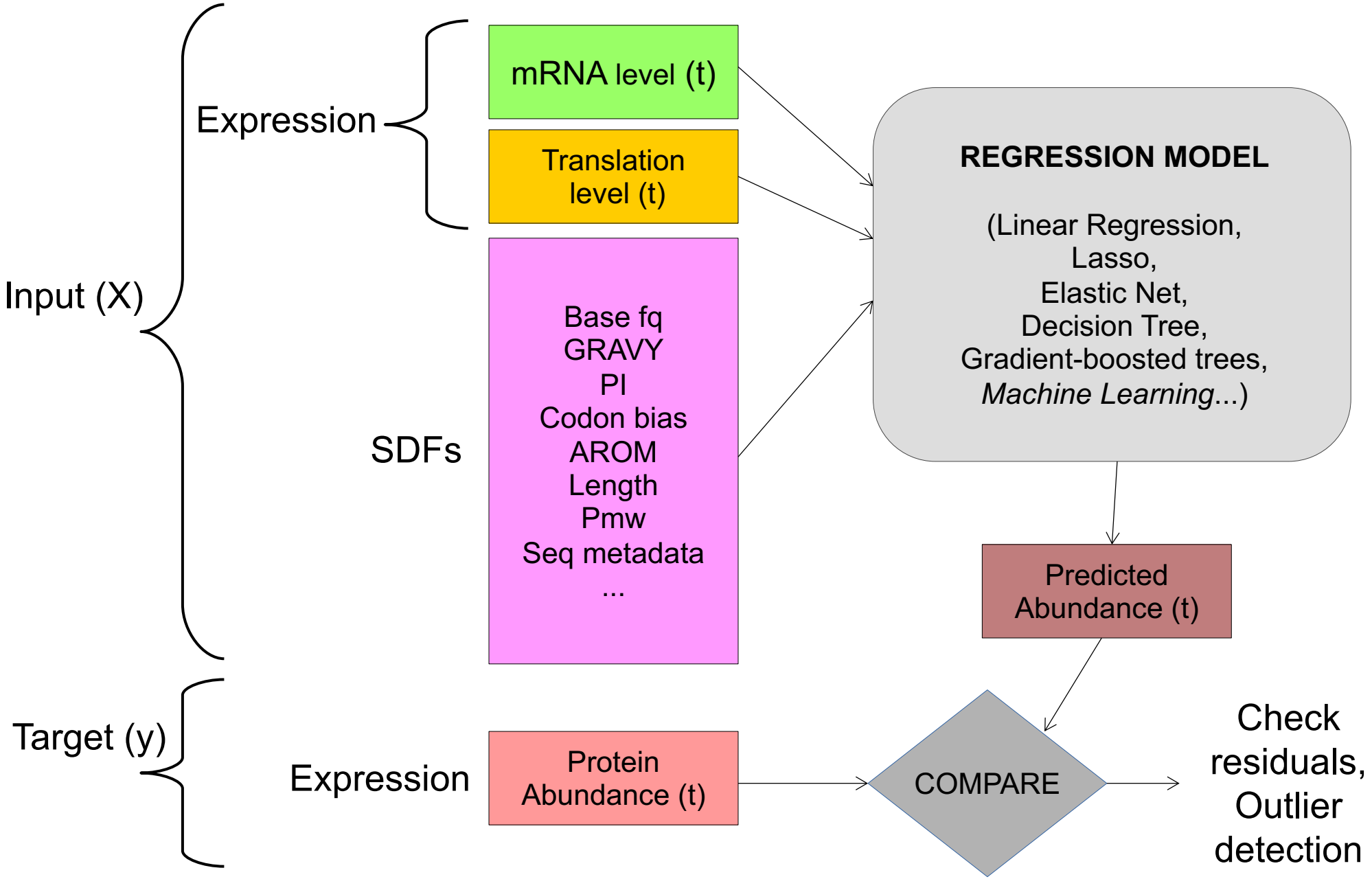
Gunawardana, Yawwani & Niranjan, Mahesan. (2013). Bridging the Gap Between Transcriptome and Proteome Measurements Identifi

Use of sequence-derived features to additive model

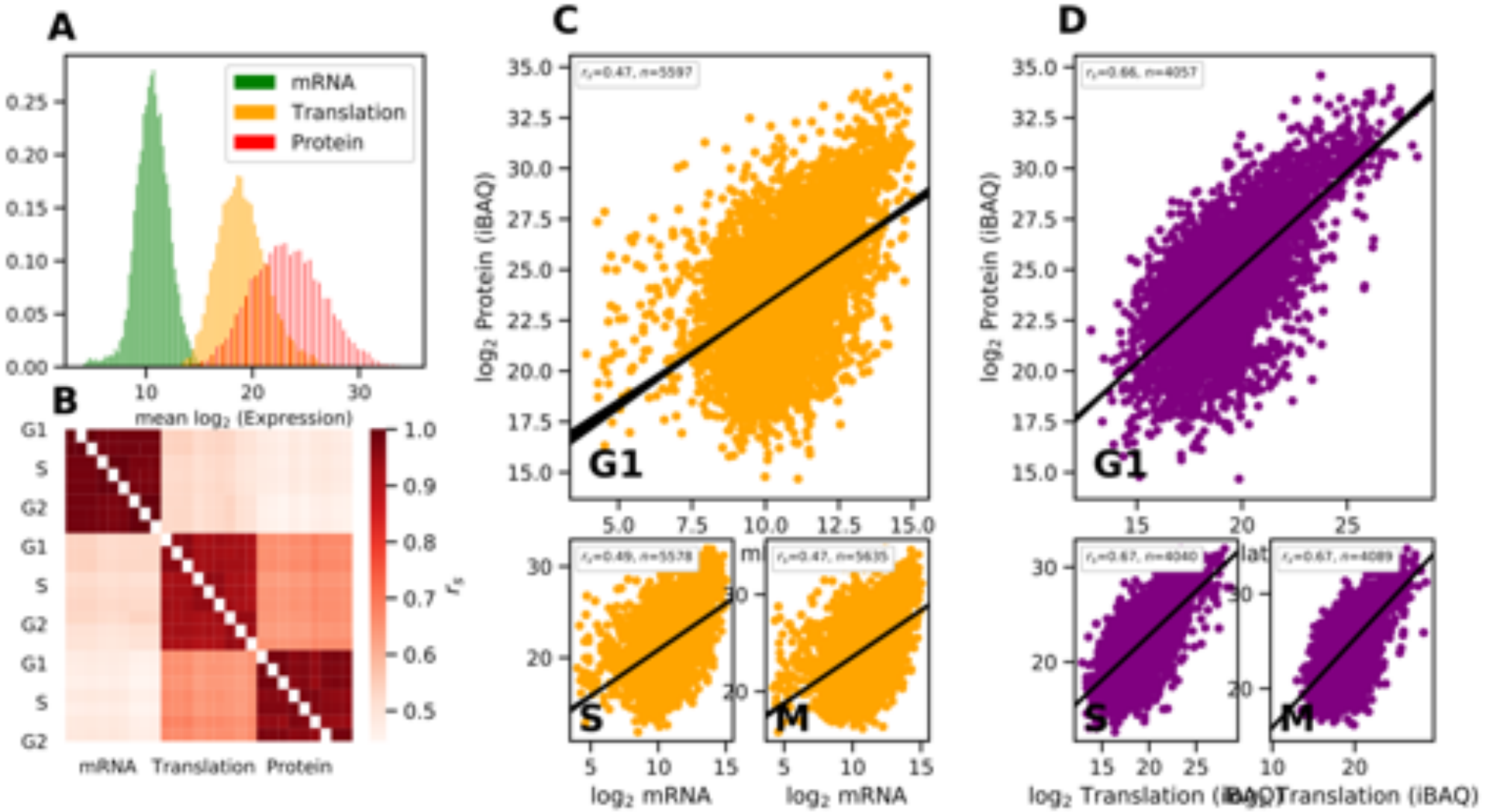


Gunawardana, Yawwani & Niranjan, Mahesan. (2013). Bridging the Gap Between Transcriptome and Proteome Measurements Identifi

Case study 2: Use of SDFs in modelling protein abundance in human cell cycle

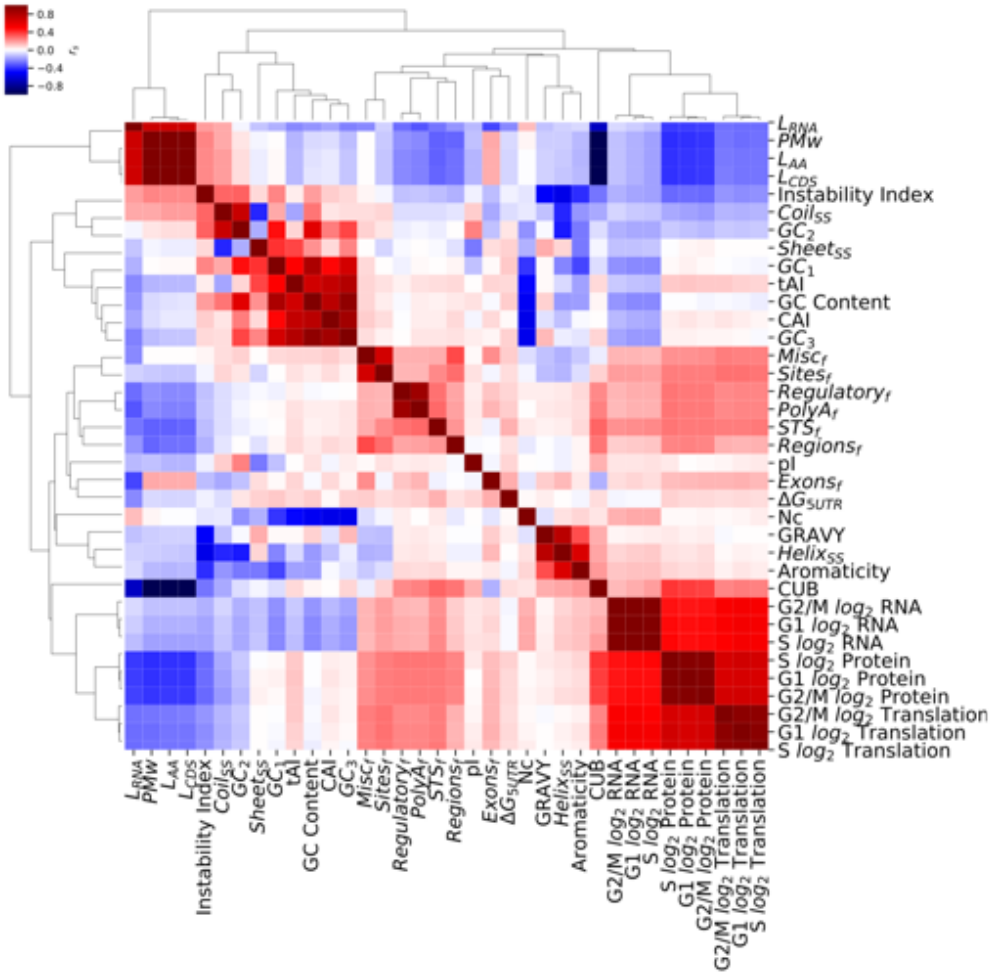


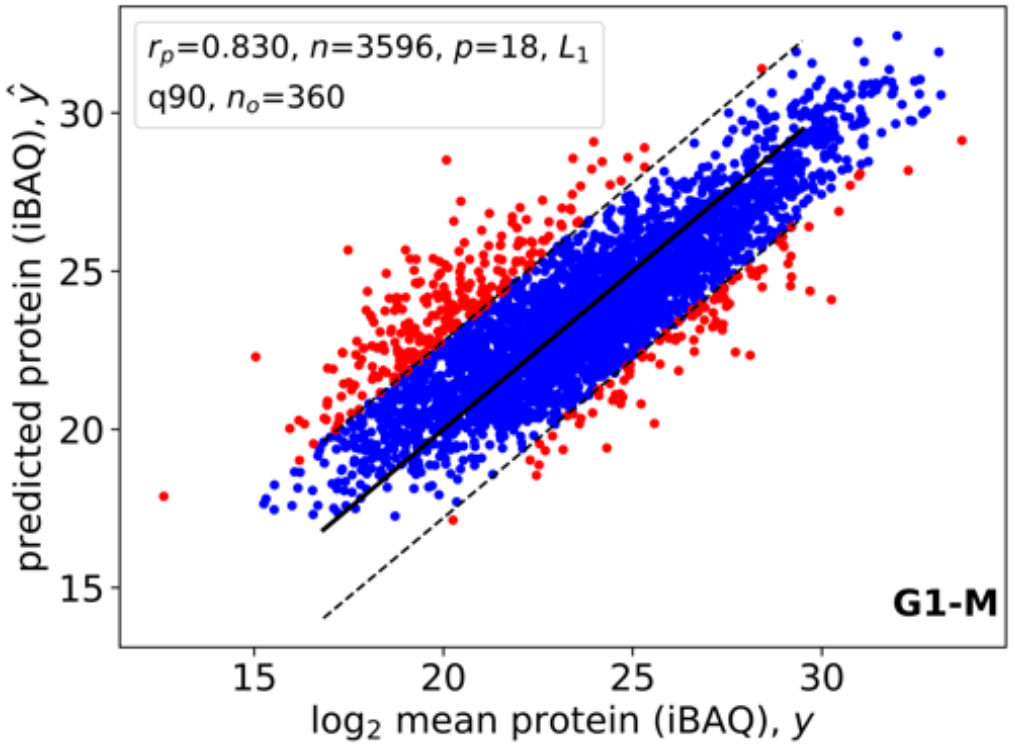
Linear models of relationship between mRNA, translation and protein



Parkes, Gregory & Niranjan, Mahesan. (2019). Uncovering Extensive Post-Translation Regulation During Human Cell Cycle Progress

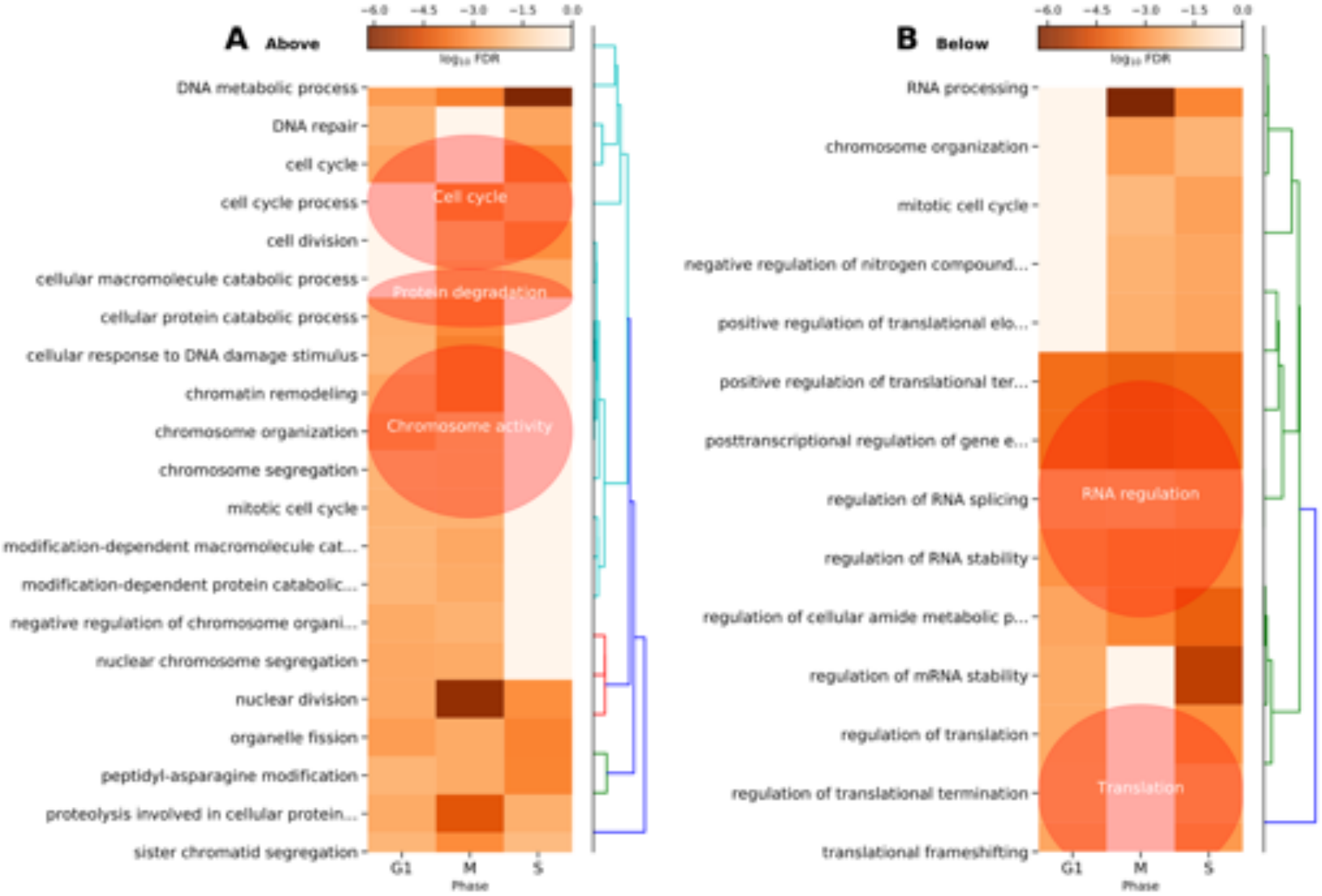
Inclusion of SDF demonstrates increase in model prediction





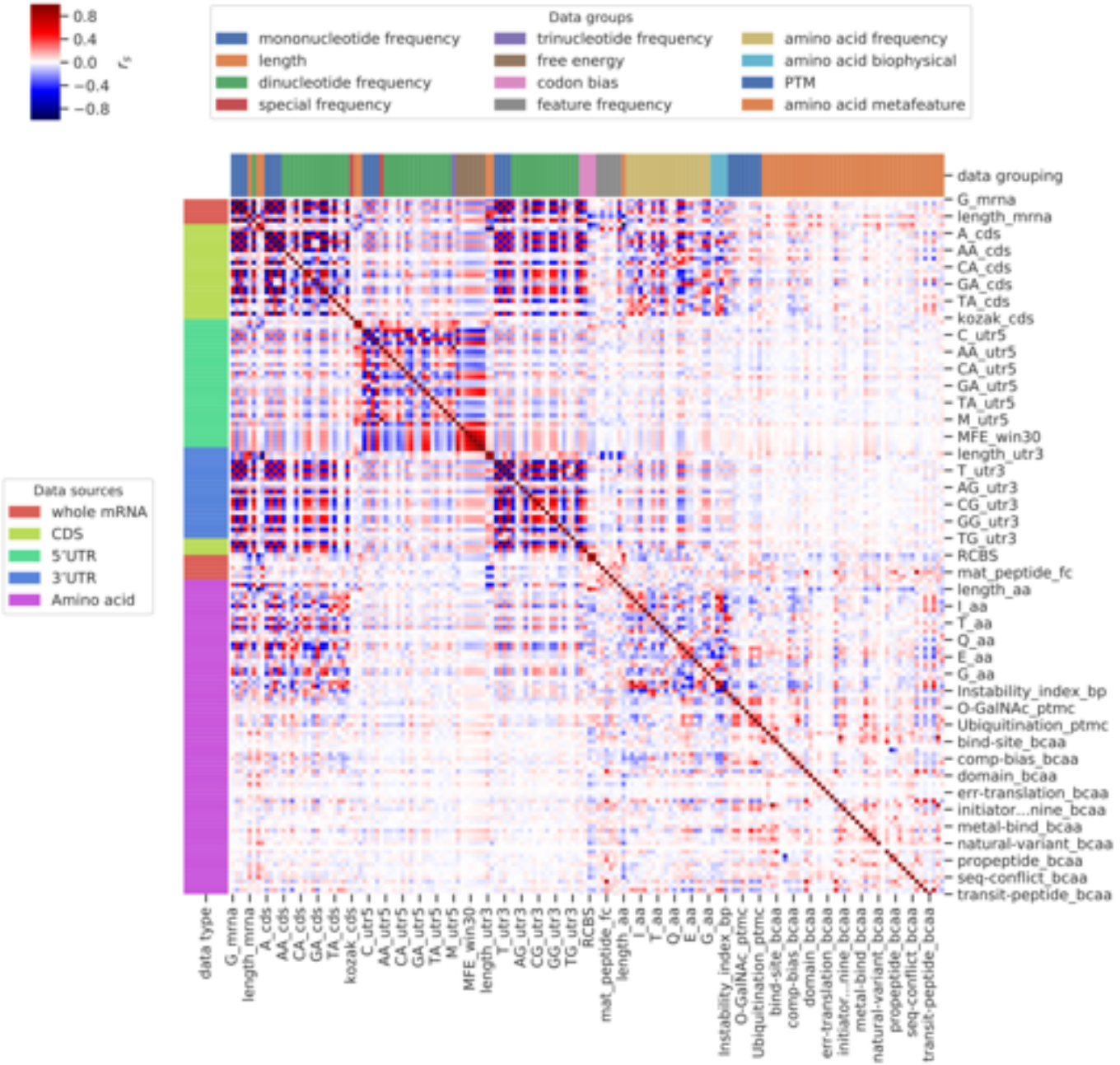
Parkes, Gregory & Niranjan, Mahesan. (2019). Uncovering Extensive Post-Translation Regulation During Human Cell Cycle Progress

Outliers to SDF have functionally predictable GO terms

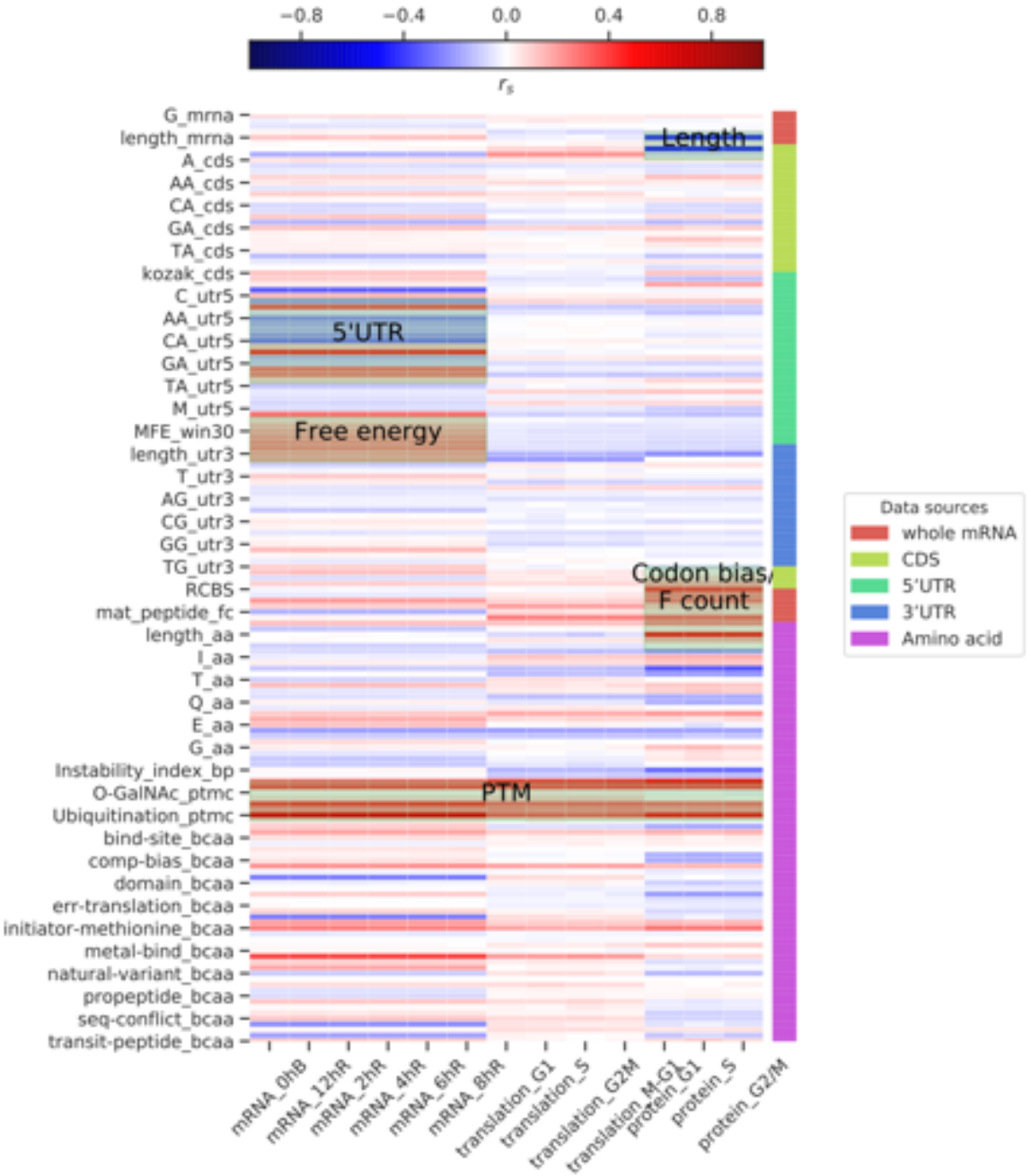


Parkes, Gregory & Niranjan, Mahesan. (2019). Uncovering Extensive Post-Translation Regulation During Human Cell Cycle Progress

Case study 3: Correlation matrix of SDFs reveals data-source subgroups

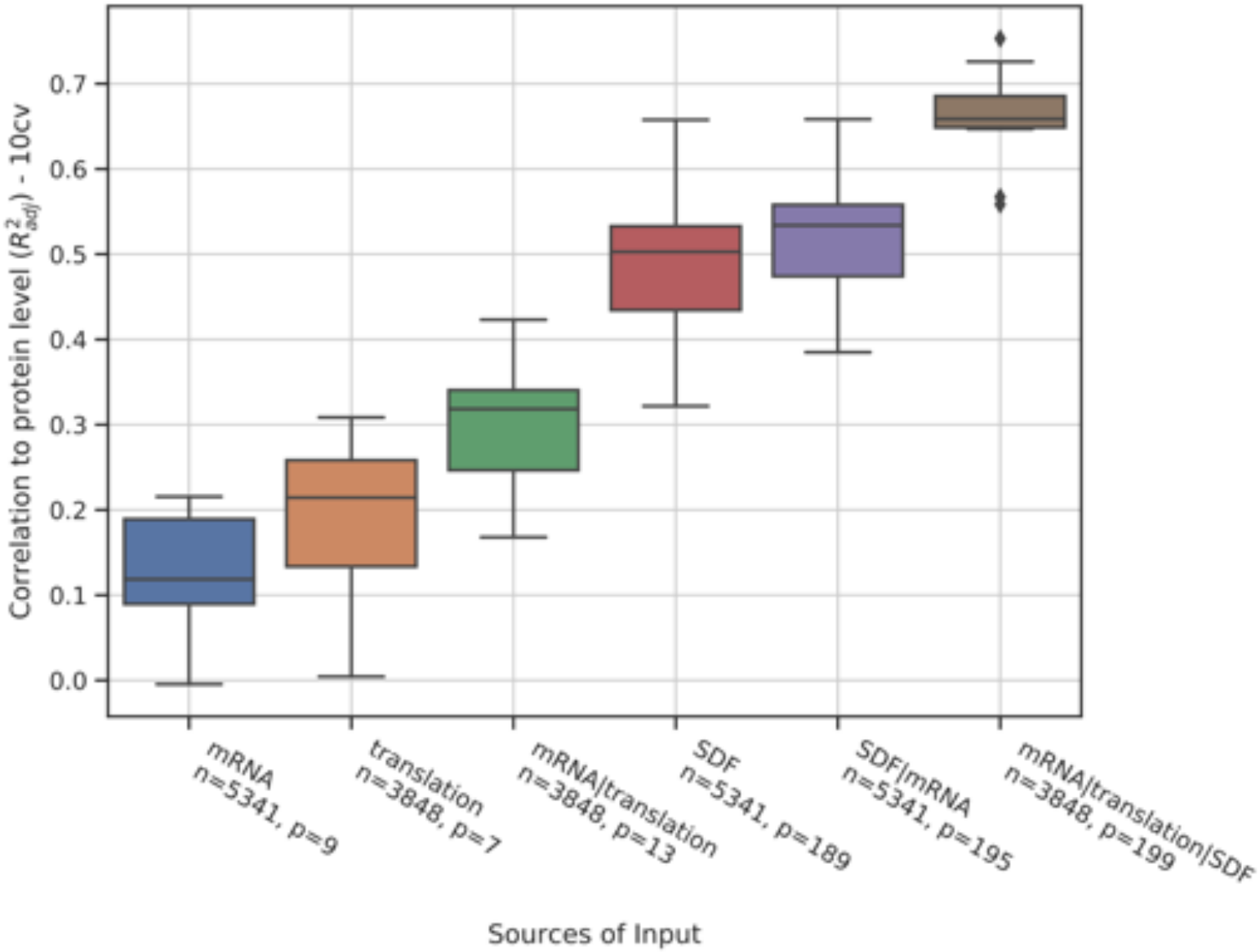


Correlation to multi-’omic features reveals SDF importance



Feature	Target / Correlation R-sq
Acetylation (PTM)	mRNA (0.06), Translation (~0.13) Protein (~0.19)
Ubiquitination (PTM)	mRNA (~0.3) , Translation (~0.09), Protein (~0.19)
Length (CDS, mRNA, AA)	Protein (~0.15-0.17)
Modified residue (bcAA)	mRNA (~0.13), Translation (~0.03)
Proline (AA)	Translation (0.01), Protein (0.04)
RCBS	Protein (~0.13)
Initiator-Methionine (bcAA)	mRNA (0.06), Translation (0.02), Protein (0.07)

Can SDFs act as a reliable proxy for mRNA abundance?



Summary

- What are sequence-derived features?
- What are ‘omics or multi-’omics?

- Exposure to examples and challenges with SDFs
- Use of sequence-derived features in yeast for protein prediction
- Use of sequence-derived features in human cell cycle study

- Expanding the number of SDFs, use as proxy
- SDFs contribute useful ‘static’ information into models