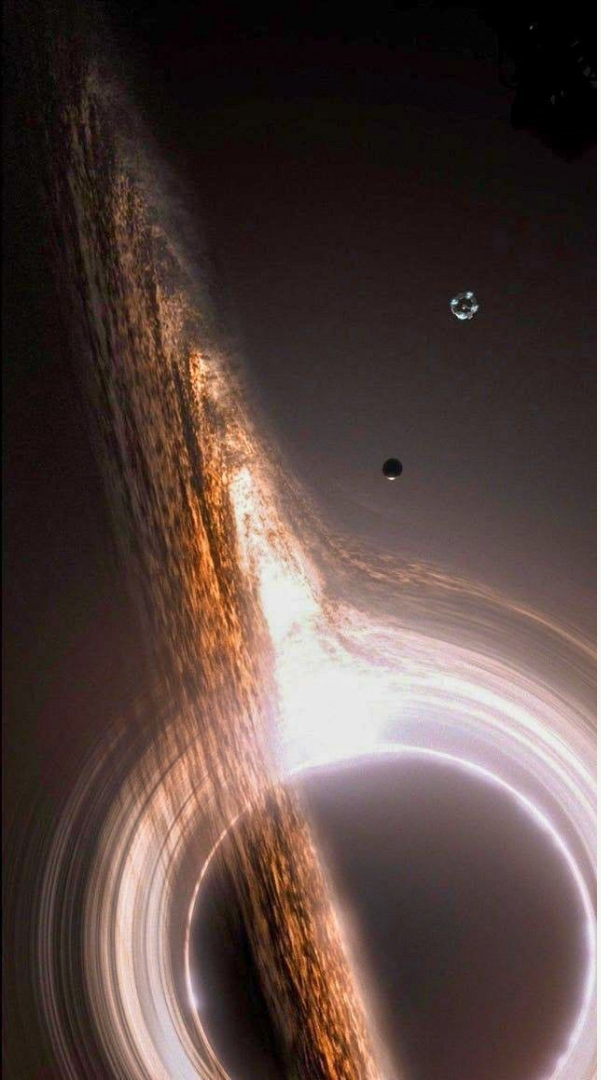


Protein Abundance Prediction

Roche PMDA Summer School 2025

Alessio D`Addio, Intern

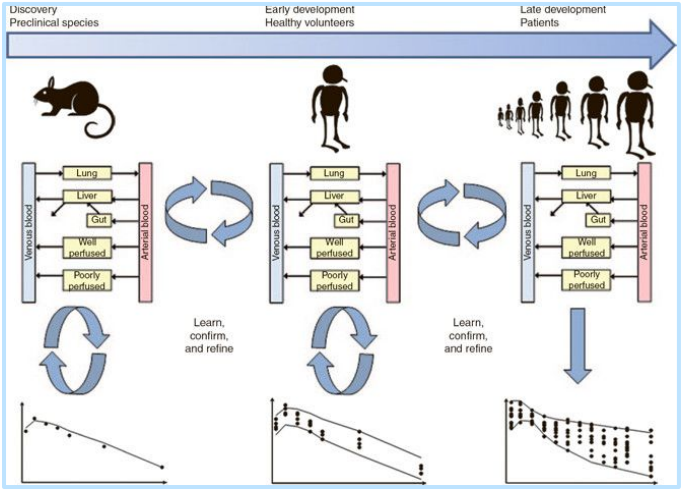
Table of contents

Why Protein Abundance Matters in Drug Discovery

Often missing Puzzle Piece

- **Pharmacokinetics (PK)**: Drug-metabolizing enzymes & transporters – their expression levels shape ADME (Absorption, Distribution, Metabolism, Excretion)
- **Pharmacodynamics (PD)**: Target protein expression – determines drug efficacy (a drug can only act where its target is sufficiently expressed)
- **Safety Profile**: Off-target proteins in critical tissues – high abundance can lead to side effects if a drug unintentionally binds them.
- **Translation to Humans**: Preclinical species (mouse, rat, etc.) protein expression informs human dosing and toxicity predictions

Physiologically Based Pharmacokinetic (PBPK)

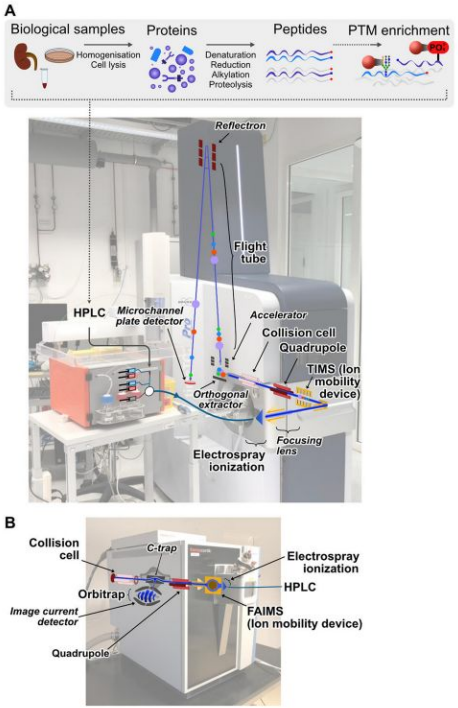


PBPK models use known physiological parameters and incorporate numerous compartments representing various organs and tissues in the body.

Basic Concepts in Physiologically Based Pharmacokinetic Modeling in Drug Discovery and Development

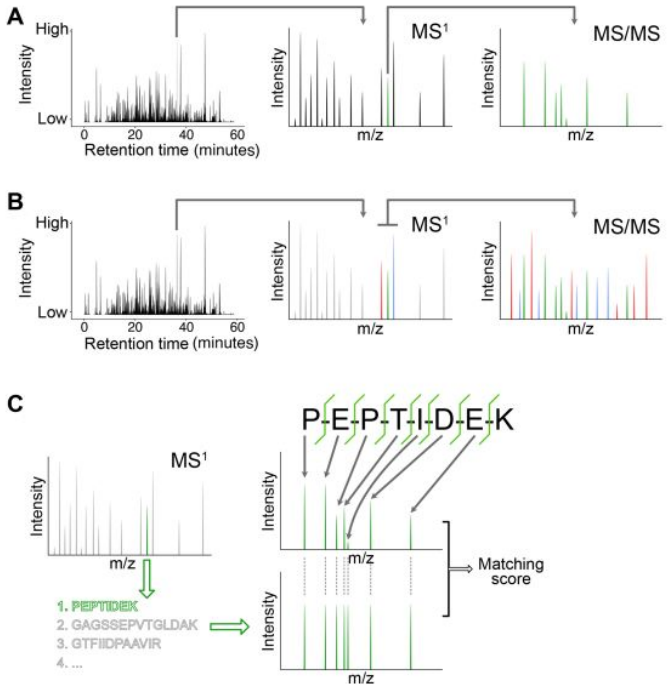
Mass spectrometry-based proteomics

Workflow



A Time-of-Flight (TOF) mass spectrometer
B Orbitrap mass spectrometer

Data-(In)Dependent Acquisition



For label-free quantification, we look at the initial MS1 survey scan

Isobaric labeling techniques like TMT (Tandem Mass Tag), the quantitative information comes from the MS/MS scan.

Challenges in Measuring the Complete Proteome

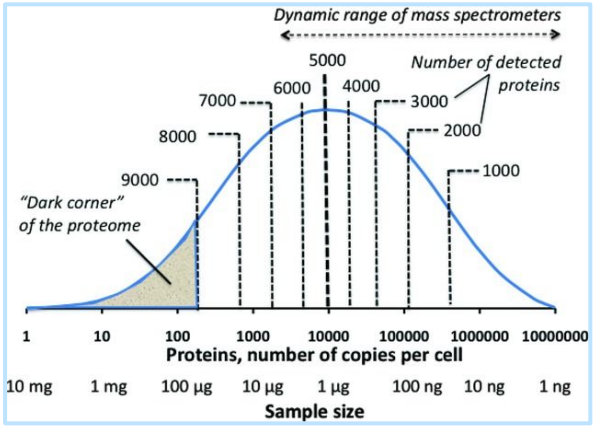
Missing proteome problem

Enormous Dynamic Range: Cellular proteins span ~7 orders of magnitude in concentration (from ~1 copy to 10 million copies per cell) – mass spectrometry instruments typically capture ~4–5 orders, leaving the extremes under-detected.

Low-Abundance Proteins Elude Detection: Transcription factors and kinases are often present in very low copy numbers and thus underrepresented in proteomic profiles – these include many key drug targets that are missed by standard assays.

Instrument & Method Biases: In mass spectrometry, highly abundant peptides dominate spectra, preventing selection of low-intensity signals. Some proteins yield peptides that ionize poorly, creating systematic blind spots.

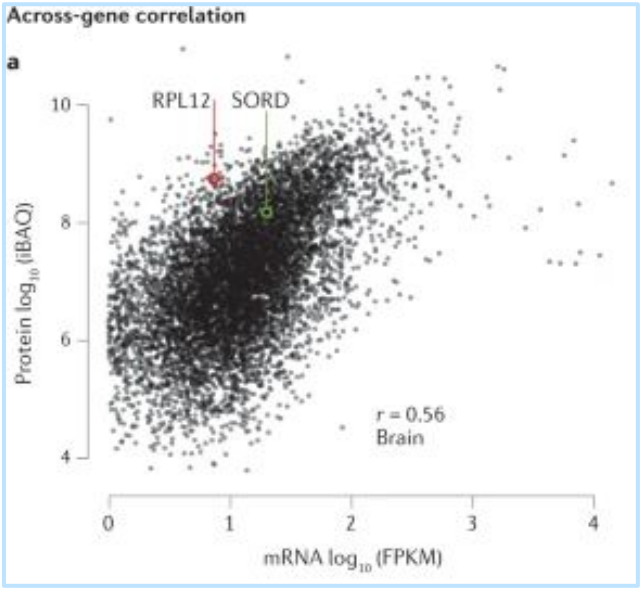
Incomplete Coverage: Even with deep sequencing runs, detecting beyond ~5,000 proteins requires exponentially more effort.



[The challenge of the proteome dynamic range and its implications for in-depth proteomics]

Why Prediction is Needed

mRNAs vs Proteins - a distinct challenge



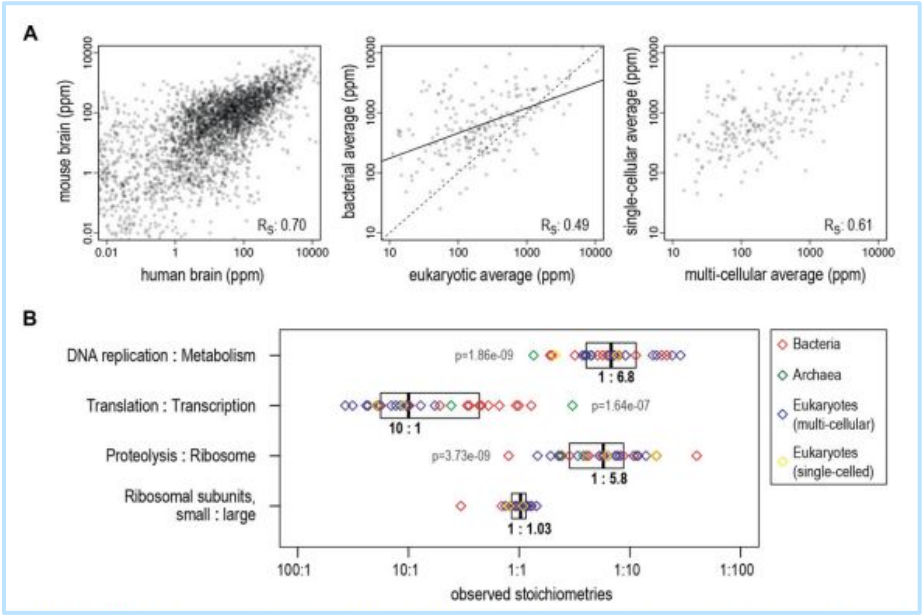
- Central Dogma ≠ 1:1 Output: mRNA level is an imperfect predictor of protein abundance. Genome-wide studies show only a moderate correlation between mRNA and protein levels (**often Pearson R ~0.4–0.6**) .
- **Post-Transcriptional Control**: Translation rates and protein degradation rates vary widely between genes.
- Cells can buffer protein levels – for example, long protein half-lives mean protein may remain abundant even if mRNA drops, and vice versa.
- **Non-Redundant Readouts**: Transcriptomics and proteomics each capture unique biology.

Figure: A deep proteome and transcriptome abundance atlas of 29 healthy human tissues, mRNAs, proteins and the emerging principles of gene expression control

Why Cross-Species Prediction?

Leveraging Prior Information

- **Conserved Tissue Signatures**: Many tissue types have similar proteomic profiles across mammals.
- **Leverage Housekeeping proteins:** Core cellular machinery and key organ-specific proteins often show predictable trends across species. These can anchor our predictions for a missing proteome
- **Species Differences Exist**:There are species-specific idiosyncrasies (a protein highly expressed in mouse liver might be lower in rat liver).



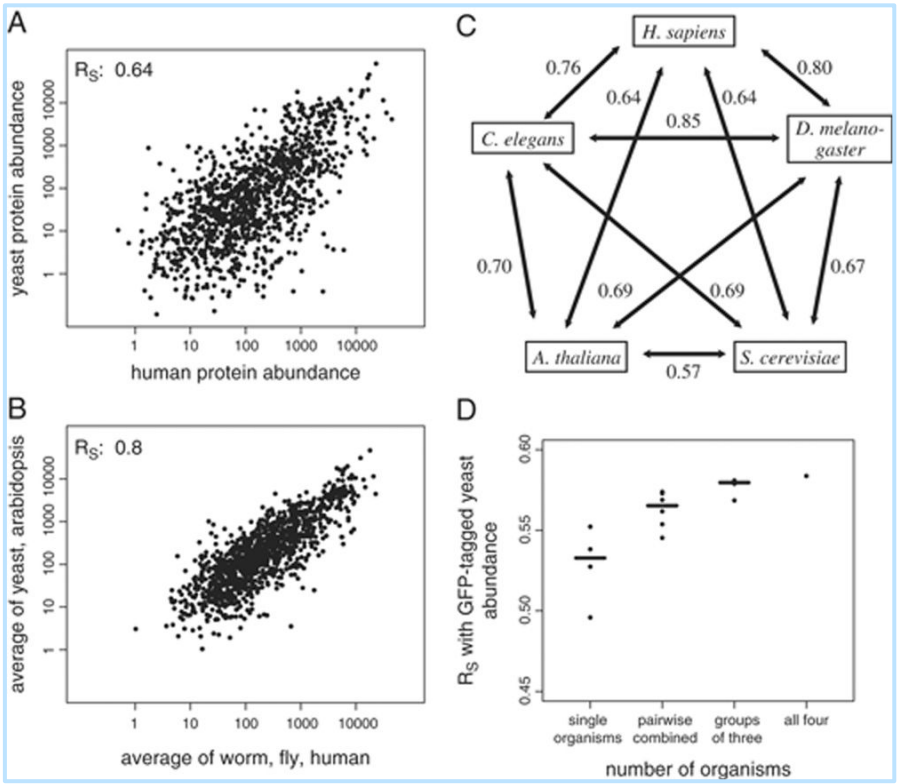
[Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines]

Using Ortholog Mapping

Orthologs = Backbone of Cross-Species Learning

The dataset provides [eggNOG ortholog] mappings. Please read the section about *Hierarchical Orthology** if you're not familiar with the annotation.

Missing Orthologs: Not all proteins have one-to-one counterparts. (Figure Notes C: In the case of organisms having more than one protein in an orthologous group, the abundances of these inparalogs in the organism were added up.)



Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome

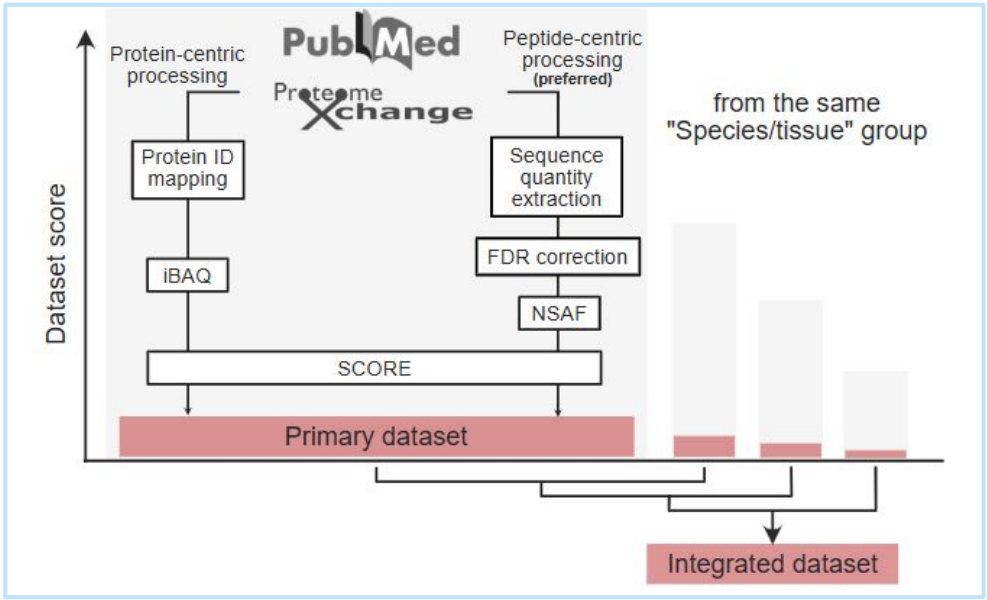
*PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life

PaxDB

Objective & Strategy

$$a = \frac{\sum_i number(p_i) \cdot length(p_i)}{\sum_j length(q_j) \cdot f(q_j)}$$

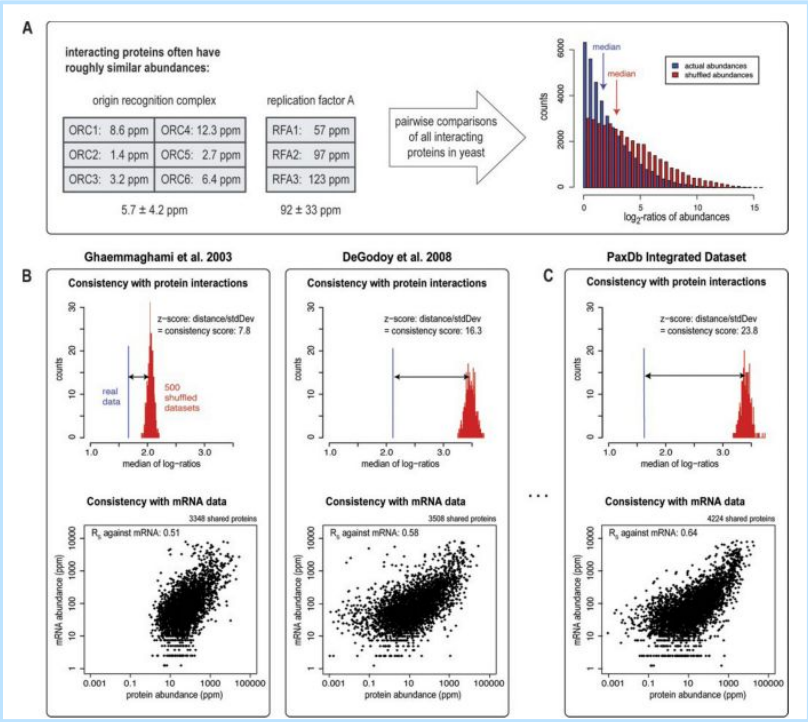
a = protein abundance
p = identified peptides
q = tryptic peptides (in silico digest)
f(q) = peptide length correction factor



- PaxDb does not use the original relative values (like log2 ratios) directly but reprocesses the data into a common, standardized unit: "parts per million" (ppm).
- **Abundance in "ppm" is essentially describing each protein with reference to the entire expressed proteome.This means each protein entity is enumerated relative to all other protein molecules in the sample.**
- The procedure depends on the input data type:
- Spectral Counting: Processed using an in-house pipeline that accounts for protein size and estimated peptide detectability.

PaxDB

Quality Score



- Datasets in PaxDb are quality scored using the "**interaction consistency score**" which assesses if interacting proteins (from STRING) show similar abundance levels.
- A higher Z-score, derived from comparing observed to randomized medians, indicates a better dataset quality in PaxDb.
- For organisms or tissues with multiple datasets, PaxDb creates a single **integrated** dataset by computing a weighted average of the individual, standardized datasets.

Provided Dataset Overview

- **Format**: We provide the PaxDB-derived data as a Parquet file in my github repo. You can load it with common data analysis libraries (Pandas, etc.)

- **Data Fields**
  - Organism (Name and taxonomy ID, e.g. Rattus norvegicus, taxid 10116)
  - Sample Cell/Tissue/Organ (e.g. LIVER, BRAIN, KIDNEY – standardized names often manually mapped and *not UBERON annotation)
  - Protein identifiers: UniProt accession & gene name; Ensembl protein/transcript IDs; a PaxDB/STRING ID linking to ortholog groups
  - Abundance value (in ppm) – the normalized protein quantity
  - Dataset details: dataset ID, description (e.g. method like “spectral counting”), a quality score (e.g. 4.2), and coverage (% of proteome identified in that experiment)

- ***Preprocessing***: *The data is already integrated and normalized by PaxDB. Nonetheless, consider filtering or transforming as needed. No additional normalization across species is required since ppm is comparable by design.*

Challenge Stage 1 – Cross-Species Tissue Imputation

Mus musculus KIDNEY Proteome Prediction

- **Goal:** Predict an entire tissue proteome for a species using all other available data.
- **Setup:** We withhold one organ's proteomic data for a particular organism during training. The model is built on proteomes of other tissues and other organisms.
- **Example Scenario (A)**: *Mus musculus KIDNEY Proteome Prediction* – Train on all other species' data (human, mouse, etc.) and on mouse data from other tissues (**training data)**, but with no mouse kidney data. The task is to predict the protein abundance in mouse kidney for all proteins that are known to be expressed there (**test set**).
- **Testing Generalization:** This Scenario tests how well your models can extrapolate to a new combination (a tissue–species pair it hasn't seen) by leveraging cross-organism patterns or whatever features you want to add. **You can use and compute any features you want and find except mouse kidney data form PaxDB**.

Challenge Stage 2 – Within-Tissue Imputation

Mus musculus KIDNEY Proteome Prediction

- **Goal:** Predict missing fraction of proteins for a partially observed tissue proteome
- **Setup:** Partway through the project (*Wednesday*), we will reveal most (**80–90%**) of the previously withheld tissue data, leaving a portion (10–20% of proteins) still held-out.
- **Refinement Scenario (B)**: Now the model has some data for the target tissue (e.g., most of mouse kidney proteome). Use this to refine predictions and impute the remaining sparse proteins that are still unmeasured.
- **Tests Interpolation**: Can your models accurately predict a set of proteins that were deliberately left out, using both cross-species info and the partial in-species-tissue data?
- **Practical Insight**: Mirrors real scenarios where an experiment detects many proteins, but one wants to estimate the ones that were missed due to detection limits.

Test Data Evaluation

Metrics

- **RMSE (Root Mean Square Error):** Average magnitude of prediction error (in ppm).
- **R^2 (Coefficient of Determination)**
- **Spearman Rank Correlation**: A model that correctly ranks most proteins (high vs low abundance) is biologically useful.
- ***Goal is to achieve low RMSE, high R^2, and high Spearman ρ.***
- **Evaluation Procedure:** We will evaluate metrics per missing proteins in tissue (mouse kidney) .

  In Stage 1, you submit the filled out test_csv file with the abundance predictions (in ppm) for each protein in the mouse kidney.
- **In Stage 2,** on the smaller held-out set. For RMSE/R^2 calculated on raw ppm values (not log transformed!).
- Might add more metrics, happy for suggestions.

Train and Test Data

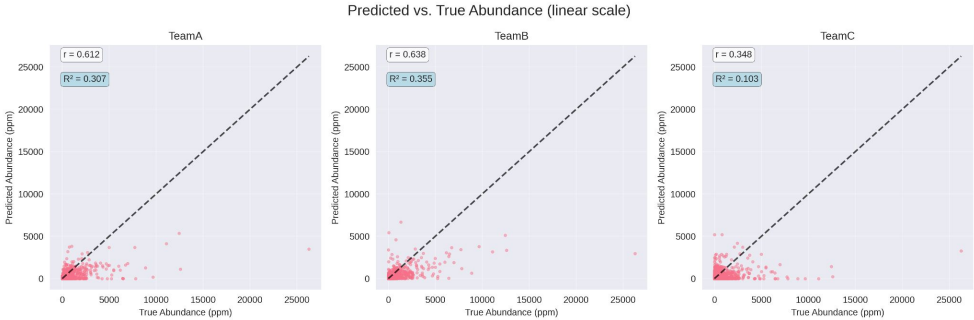
Paxdb Protein Abundance Data

- Original dataset has 4,812,859 rows.
- Removed 13,408 rows where *organism_name* was 'M.musculus' and *sample_organ* was 'KIDNEY'.
- The main filtered **training** dataset now has 4,799,451 rows (<1.57GB parquet file gzip compressed). (please verify after downloading from repo).
  - I advise to focus on the sets of organism with high proteome coverage first. (but you can use every species)
  - I suggest you split the dataset by organism into multiple files (parquet > jsonl > csv)
  - Every row is one datapoint. (is_Integrated= True, has only one abundance value per tissue, organism and protein)
- Masked abundance in 5,654 rows where *is_integrated* was True for the target organism/organ.
- => 5,654 unique proteins (*EnsemblProteinID*) within the masked data
- => Removed 269 rows from the test file due to missing *Sequence* information in uniprot.
- **There are 5,385 unique proteins left to predict** (test_masked_integrated_rows_M.musculus_KIDNEY.csv)

Doing now what patients need next

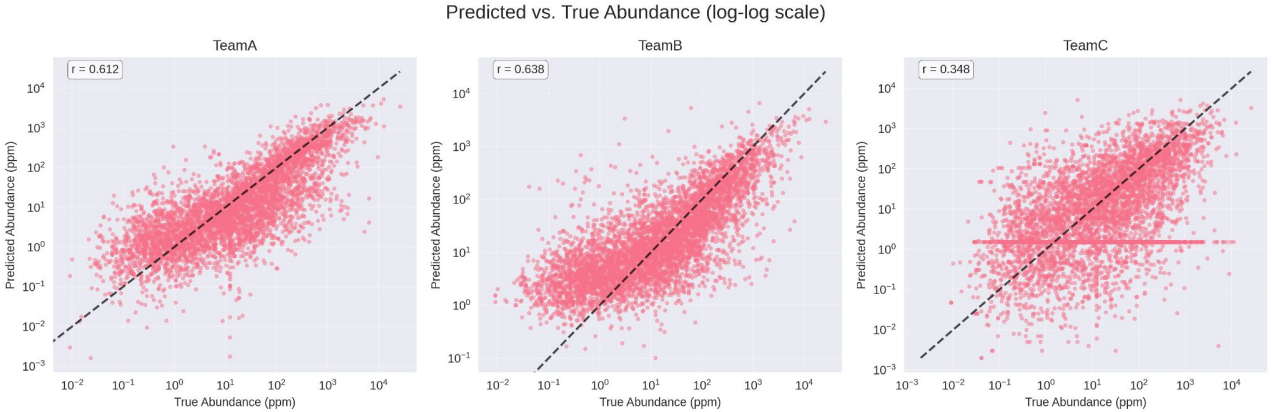
Stage 1

Tissue Imputation



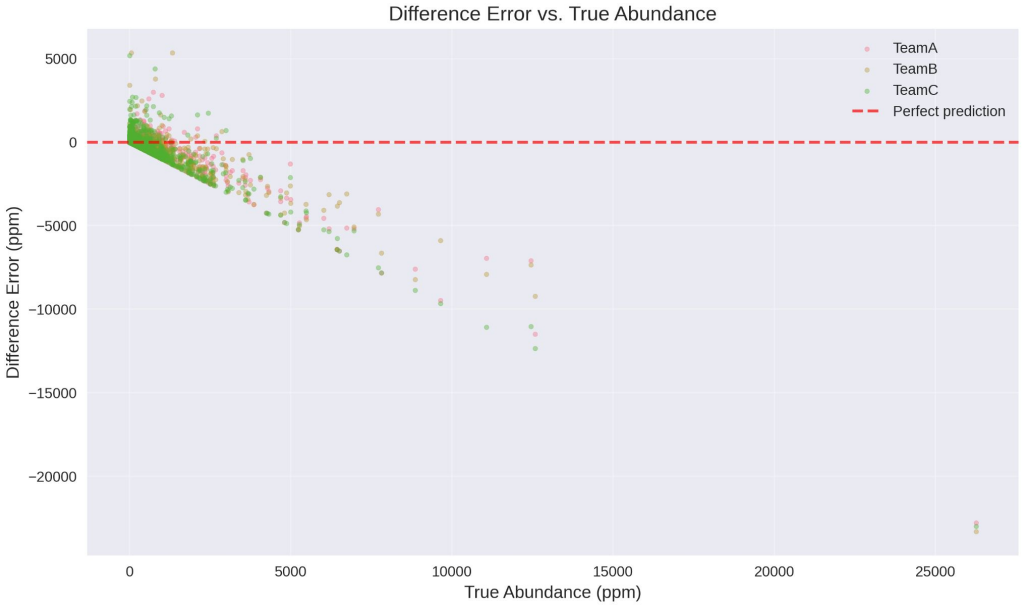
--- Performance Leaderboard (sorted by RMSE) ---

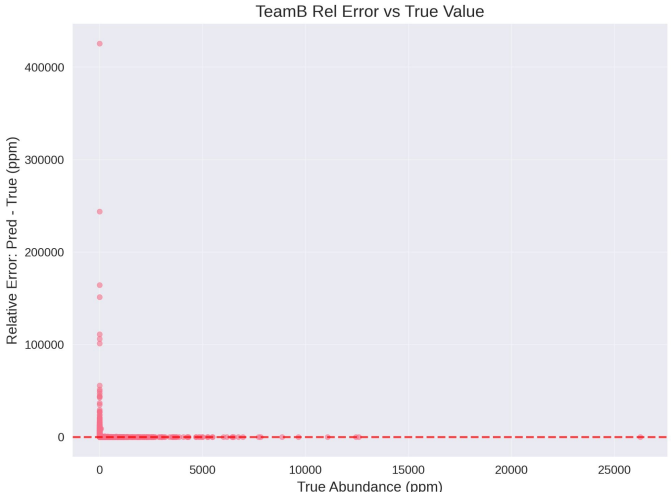
	team	RMSE	MSE	R2	Spearman	n
1	TeamB	568.9352	323687.2178	0.3552	0.7590	5387
0	TeamA	589.9136	347998.0644	0.3070	0.6056	5385
2	TeamC	670.9172	450129.8972	0.1033	0.4843	5387

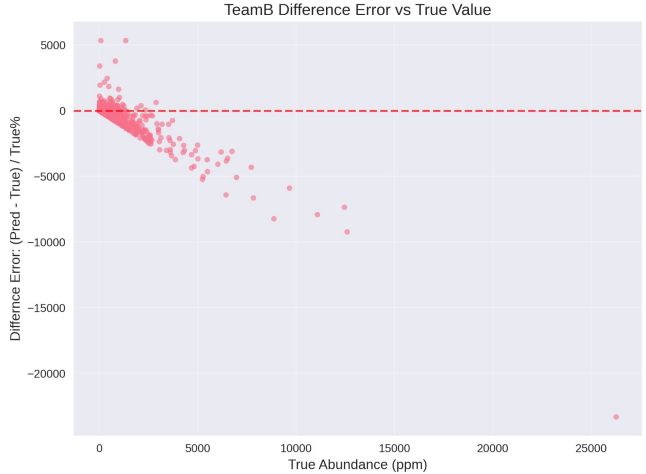


Bland-Altman Plot

Tims-plot







- Absolute scale: a systematic underestimation of high abundance proteins
- Relative scale: many-fold over estimation of low abundance proteins

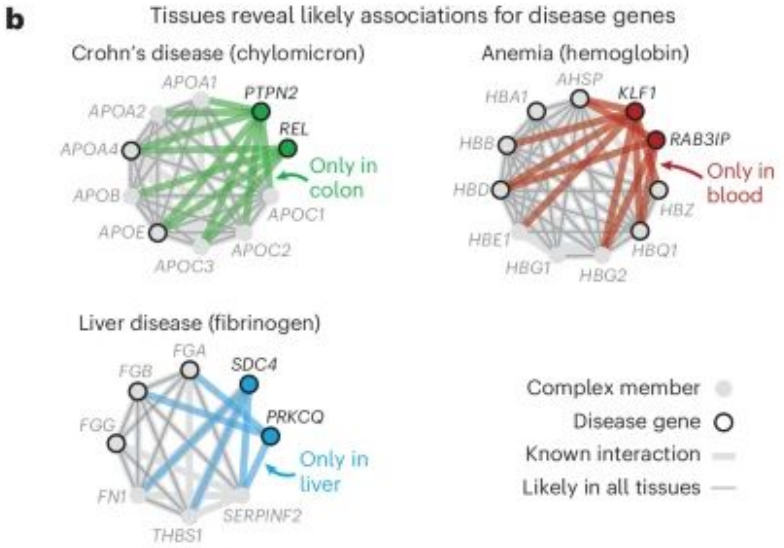
Protein Abundance Prediction

Stage2

- - 12129 rows -> train_added_M.musculus_KIDNEY_data.csv.gz
- - 540 rows -> test2_unmasked_dropped_M.musculus_KIDNEY_data.csv.gz (unmasked)

Stratified random sampling was performed on the integrated dataset. Proteins were divided into three abundance strata (low, mid, high) based on the terciles of their log10-transformed abundance. The final 10% test set was sampled proportionally from each of these strata.

Applications of Predicted Proteomes



A tissue-specific atlas of protein–protein associations enables prioritization of candidate disease genes