Universidade do Minho
Escola de Engenharia

Maria Adília Balacó Chócha Pessoa Monteiro

A model validation pipeline for healthy tissue genome-scale metabolic models

novembro de 2021

UMinho | 2021

Maria Adília Pessoa Monteiro

A model validation pipeline for healthy tissue genome-scale metabolic models

Universidade do Minho
Escola de Engenharia

Maria Adília Balacó Chócha Pessoa Monteiro

A model validation pipeline for healthy tissue genome-scale metabolic models

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho efetuado sob a orientação do
Doutor Miguel Francisco de Almeida Pereira da Rocha
e do
Doutor Pedro Gabriel Dias Ferreira

novembro de 2021

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Agradecimentos

Em primeiro lugar, agradeço aos meus orientadores, Doutor Miguel Rocha e Doutor Pedro Ferreira, por toda a paciência e orientação ao longo deste trabalho.

Em segundo lugar, agradeço aos meus “tutores”, Vítor Vieira, Jorge Ferreira e Tânia Barata, que me acompanham desde o projeto de 1° ano e cujo feedback e apoio também foram fundamentais.

Em último lugar, agradeço à minha família e amigos, especialmente à Natacha, Sónia e à Inês.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Resumo

Nos últimos anos, os métodos de alto rendimento disponibilizaram dados ómicos referentes a várias camadas da organização biológica, permitindo a integração do conhecimento de componentes individuais em modelos complexos, como modelos metabólicos à escala genómica (GSMMs). Estes podem ser analisados por métodos de modelação baseada em restrições (CBM), que facilitam abordagens preditivas *in silico*.

Os modelos metabólicos humanos têm sido usados para estudar tecidos saudáveis e as suas doenças metabólicas associadas, como obesidade, diabetes e cancro. Modelos humanos genéricos podem ser integrados com dados contextuais por meio de algoritmos de reconstrução, com vista a produzir modelos metabólicos contextualizados (CSMs), que são normalmente melhores a capturar a variação entre diferentes tecidos e tipos de células. Como o corpo humano contém uma grande variedade de tecidos e tipos de células, os CSMs são frequentemente adotados como um meio de obter modelos metabólicos mais precisos de tecido humano saudável.

No entanto, ao contrário de modelos de microrganismos e cancro, que acomodam vários métodos de validação, como a comparação de fluxos *in silico* ou de previsões de genes essenciais com dados experimentais, os métodos de validação facilmente aplicáveis a CSMs de tecido humano saudável podem ser mais limitados. Consequentemente, apesar de esforços continuados para atualizar os modelos humanos genéricos e algoritmos de reconstrução para extrair CSMs de alta qualidade, a sua validação continua a ser uma preocupação.

Este trabalho apresenta uma pipeline para a extração e validação básica de CSMs de tecidos humanos normais derivados da integração de dados transcriptómicos com um modelo humano genérico. Todos os CSMs foram extraídos do modelo genérico Human-GEM publicado recentemente por Robinson *et al.* (2020), usando o package *Troppo* em Python e nos algoritmos de reconstrução fastCORE e tINIT nele implementados. Os CSMs extraídos correspondem a 11 tecidos saudáveis disponíveis no conjunto de dados GTEx v8.

Antes da extração, métodos de aprendizagem máquina foram aplicados à seleção de um limiar para conversão em *gene scores*. Os modelos de maior qualidade foram obtidos com um limite mínimo global aplicado diretamente aos dados ómicos. A estratégia de validação focou-se no número de tarefas metabólicas passadas como um indicador de desempenho. Por último, este trabalho é acompanhado por Jupyter Notebooks, que incluem um guia de extração de modelos para novos utilizadores.

Palavras-chave: Modelação baseada em restrições; Modelos metabólicos contextualizados; Package *Troppo;* Tarefas metabólicas; Tecido humano saudável.

Abstract

In the past few years, high-throughput experimental methods have made omics data available for several layers of biological organization, enabling the integration of knowledge from individual components into complex models such as genome-scale metabolic models (GSMMs). These can be analysed by constraint-based modelling (CBM) methods, which facilitate *in silico* predictive approaches.

Human metabolic models have been used to study healthy human tissues and their associated metabolic diseases, such as obesity, diabetes, and cancer. Generic human models can be integrated with contextual data through reconstruction algorithms to produce context-specific models (CSMs), which are typically better at capturing the variation between different tissues and cell types. As the human body contains a multitude of tissues and cell types, CSMs are frequently adopted as a means to obtain accurate metabolic models of healthy human tissues.

However, unlike microorganisms' or cancer models, which allow several methods of validation such as the comparison of *in silico* fluxes or gene essentiality predictions to experimental data, the validation methods easily applicable to CSMs of healthy human tissue are more limited. Consequently, despite continued efforts to update generic human models and reconstruction algorithms to extract high quality CSMs, their validation remains a concern.

This work presents a pipeline for the extraction and basic validation of CSMs of normal human tissues derived from the integration of transcriptomics data with a generic human model. All CSMs were extracted from the Human-GEM generic model recently published by Robinson *et al.* (2020), relied on the open-source *Troppo* Python package and in the fastCORE and tINIT reconstruction algorithms implemented therein. CSMs were extracted for 11 healthy tissues available in the GTEx v8 dataset.

Prior to extraction, machine learning methods were applied to threshold selection for gene scores conversion. The highest quality models were obtained with a global threshold applied to the omics data directly. The CSM validation strategy focused on the total number of metabolic tasks passed as a performance indicator. Lastly, this work is accompanied by Jupyter Notebooks, which include a beginner friendly model extraction guide.

Keywords: Constraint-based modelling; Context-specific models; Healthy human tissue; Metabolic tasks; *Troppo* package.

Table of contents

List of abbreviations / Acronyms

CBM: Constraint-based modelling

cDNA: Complementary DNA

COBRApy: COnstraint-Based Reconstruction and Analysis for Python

CORDA: Cost Optimization Reaction Dependency Assessment

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

CSM: Context-specific model

EHMN: Edinburgh human metabolic network

fastCORE: Fast Consistent Reconstruction

FBA: Flux Balance Analysis

FDR: False discovery rate

FPKM: Fragments Per Kilobase per Million

FSGS: Focal Segmental Glomerulosclerosis

FVA: Flux variability analysis

GENRE: Genome-scale network reconstruction

GIMME: Gene Inactivation Moderated by Metabolism and Expression

GPRs: Gene-protein-reaction rules

GSMM: Genome-scale metabolic model

GTEx: Genotype-Tissue Expression project

HK: Housekeeping (tasks)

HMR: Human Model Reaction

iMAT: Integrative Metabolic Analysis Tool

INIT: Integrative Network Inference for Tissues

LP: Linear programming

MBA: Model-Building Algorithm

mCADRE: Metabolic Context Specificity Assessed by Deterministic Reaction Evaluation

MCC: Matthews correlation coefficient

Med: Per-sample Median

Med-pgQ2: Per-gene normalization after per-sample median

MFA: Metabolic flux analysis

MILP: Mixed Integer Linear Programming

ML: Machine learning

mRNA: Messenger RNA

NAFLD: Non-alcoholic fatty liver disease

NASH: Non-alcoholic steatohepatitis

NCI: National Cancer Institute

NGS: Next-generation sequencing

NHGRI: National Human Genome Research Institute

ODE: Ordinary differential equation

PCA: Principal component analysis

pFBA: Parsimonious Flux Balance Analysis

RIN: RNA Integrity Number

RMF: Required Metabolic Functionality

RNA-Seq: RNA Sequencing

RPKM: Reads Per Kilobase per Million mapped reads

S matrix: Stoichiometric matrix

TC: Total Counts

TCGA: The Cancer Genome Atlas

TMM: Trimmed Mean of M values

TPM: Transcript per Million

tSNE: t-Distributed Stochastic Neighbour Embedding

UQ: Per-sample Upper Quartile

UQ-pgQ2: Upper-quartile global scaling

Figure Index

Table Index

1. Introduction

1.1. Context and Motivation

Systems biology seeks to understand physiology and disease at the system-level, integrating knowledge from individual components into complex models. High-throughput methods have made available copious amounts of omics data from different layers of biological organization. These large-scale measurements can drive *in silico* predictive approaches, such as genome-scale metabolic models (GSMMs) [1, 2].

GSMMs are mathematically consistent representations of metabolic networks, composed of reactions and metabolites inferred through functional genomics. Constraint-based modelling (CBM) methods can be used to analyse GSMMs, ignoring unknown kinetic parameters by assuming that intracellular metabolite abundances are constant over time [1]. Furthermore, CBM methods usually require an objective function that is typically fulfilled by an artificial biomass reaction representing the demand for certain metabolites required for cell growth [1, 3].

Human metabolic models aim to be comprehensive, up-to-date collections of the components of human metabolism [3]. The first human generic model, Recon1, was followed by the Edinburgh human metabolic network (EHMN), the Human Model Reaction (HMR) database and updated model versions such as HMR2 and Recon3D [1,3]. One of the most recent models, Human1, was released in 2020 aiming to unify the two major human GSMM lineages, HMR and Recon [4]. Generic models can serve as a scaffold for integration with contextual data, or omics data, to better capture the variation between different tissues, cell types and environmental variability [1].

Context-specific models (CSMs) are typically extracted from generic models by removing inactive reactions based on omics data, and are, therefore, subsets of a template general model. Tailoring the model's reactions to capture the enzymatic profile of a certain tissue, cell type or condition, often results in greater predictive ability relative to this context [1].

Reconstruction algorithms can be used to integrate omics data in a generic model and extract a CSM. These approaches can be classified into 3 families: Gene Inactivation Moderated by Metabolism and Expression (GIMME)-like, integrative Metabolic Analysis Tool (iMAT)-like and Model-Building Algorithm (MBA)-like algorithms [2]. As there are multiple algorithm alternatives to choose from, no standard algorithm for model building exists.

CSMs based on these generic models have been built to explore the metabolism of various healthy human cell types and tissues, including adipocytes, as well as liver and kidney cells [1,3]. Healthy

CSMs have been used to compare wild-type and mutant cells and predict gene knockout phenotypes. CSMs also proved useful for studying diseases like non-alcoholic fat liver disease, type two diabetes, and cancer [1, 5]. CSMs for the latter have helped to identify biomarkers and therapeutic targets [6] and successfully simulate the Warburg effect [7].

Despite efforts to update generic human models over the years, coupled with algorithms developed to extract high quality CSMs, biological validation remains a concern. This work will address the validation of the predictive power of healthy human tissue models, which often face more challenges than their disease counterparts [1]. For instance, the rapid growth of microorganisms and cancer cells can be translated into biomass maximization as an objective function for CBM. In contrast, metabolic objective functions for human healthy cells and tissues are harder to define [1,3]. Given the difficulty in translating omics data into metabolic models, there is a need for a more unified pipeline for healthy CSM extraction and validation.

1.2. Research objectives

The aim of this work is to generate context-specific genome-scale models of healthy human tissues (for example, of liver, breast and renal tissue), establishing a pipeline for their reconstruction and validation. This optimization process aims to cover transcriptomics data pre-processing for integration with a generic model, extraction using reconstruction algorithms with separate approaches and common CSM validation methods. This will be achieved addressing the following scientific/technological objectives:

- Exploratory analysis and pre-processing of the transcriptomics obtained from the GTEx project [39] and other relevant data for their integration with metabolic models;
- Reconstruction and curation of tissue-specific models of healthy human tissues selected as case studies using the fastCORE and tINIT reconstruction algorithms implemented in the *Troppo* Python package [8], and the generic model Human1 as the template model [4];
- Analysis of the generated models based on metabolic tasks and other methods applicable to healthy human tissues, and their comparison to models of the relevant tissue;
- Development of a pipeline for reconstruction and validation of healthy tissue-specific models.

1.3. Report Outline

The main content of this thesis is divided into 4 sections, with Section 2 covering the state-of-the-art review. Sections 2.1.1 and 2.1.2 introduce the field of systems biology, define genome-scale metabolic models, distinguish between model types, and summarize the recent history of human models. Sections

2.1.3 and 2.1.4 explore the key principles of constraint-based models and how phenotype predictions may be applied to context-specific models of healthy human tissue. Section 2.2 addresses omics data types, sources, and pre-processing methods for integration with metabolic models, with a focus on transcriptomics. Section 2.3 includes a detailed review of context-specific metabolic model extraction and validation methods, including reconstruction algorithms, and lists some examples of applications of these models.

Section 3 (Methods) covers technical methods involved in the three stages of the pipeline, namely data pre-processing, model extraction and model validation, in detail. Section 4 (Results) presents the developed pipeline, all models extracted, the methods of validation used and compares the effects of several different extraction conditions in model quality. Firstly, Section 4.1 explores the pre-extraction phase of the pipeline, such as the gene scores threshold selection process. Section 4.2 focuses on reaction content and how the models separate by tissue. Section 4.3 and 4.4 focus on essential and full (more specific) metabolic tasks as model quality markers. Section 4.5 mainly contrasts the pre-processing methods employed by Robinson *et al.* and those explored in the previous sections. Finally, Section 5 (Discussion) describes the general conclusions obtained in this work, whereas Section 6 summarizes the main conclusions as well as examples of possible related future work.

2. State-of-the-art review

2.1. Systems biology

The study of the genotype–phenotype relationship is essential to the life sciences [1, 9]. Although single-omic layers have been the focus of the 20th century, complex biological systems cannot be understood with just the knowledge of its components [9]. Since then, technological advancement of high-throughput methods has gained renewed interest in the systems-level approach promoted by systems biology [10].

2.1.1. Genome-scale metabolic models

In systems biology, experimental data can be integrated into mathematical models to perform predictive simulations. Metabolic models are multi-omic approaches, capable of addressing biological and environmental systemic interactions that underlie phenotypes [9]. The kinetic and constraint-based approaches are the most popular methods to model a metabolic system [11], with key differences in their representation of enzyme kinetics [1].

Kinetic models employ ordinary differential equations (ODEs) and kinetic parameter values to model highly dynamic mechanisms of enzyme dependencies [1, 12]. However, they are often limited by experimental data availability for model calibration and high computational costs of solving many concurrent ODEs [12, 13]. On the other hand, constraint-based models focus on the global redistribution of metabolic fluxes to reach an objective function and are computationally cheap [12, 13]. However, neither method traditionally includes transcriptional or translational feedback, which also operate on a much larger timescale than reaction kinetics [13].

A genome-scale network reconstruction (GENRE) is an organism-specific collection of biochemical transformations, based on curated literature such as genome annotation. GENREs are converted into mathematical form as *in silico* GSMMs by assessing phenotypic properties [14]. Most GSMMs also employ gene-protein-reaction rules (GPRs), which represent the link between genes and proteins with Boolean logic (for example, AND for enzyme complexes and OR for isoenzymes) [1]. GPRs can serve as a scaffold for overlaying quantitative omics data and enable *in silico* gene perturbation experiments and exploration of their effect on metabolism [1, 4].

2.1.2. Human metabolic models

The past 15 years brought forth a community effort to develop and improve GSMMs for human metabolism, focused on updating and expanding metabolic reaction coverage [15, 4]. The first human generic model, Recon1 [16], was released in 2007, followed by the Edinburgh human metabolic network (EHMN) [17]. These first models formed the basis of the two major human GSMM lineages, Human Model Reaction (HMR) database and Recon. The Recon model series was succeeded by several updated versions [18, 19, 20], and the Human Model Reaction (HMR) database model [21] was also followed by HMR 2.0 [22], with revised gene-protein-reaction rules based on new human genome insight. The generic human model iHsa (2017) [23], built in parallel with the mouse model iRno, expanded upon the HMR2 model. Another entry in the Recon series, Recon3D [15], was presented in 2018, and Thiele *et al.* (2018) built upon it to generate gender-specific whole-body metabolism reconstructions, Harvey and Harvetta [24].

Robinson *et al.* (2020) integrated and reconciled information from HMR2, iHsa, and Recon3D to develop a consensus generic human model, Human1. This process included the removal of duplicated or inconsistent reactions and metabolites, revision of metabolite formulas, rebalancing of reaction equations and correction of reversibility for reactions and creating a new generic human biomass reaction. In addition, the authors compared genes predicted to be essential by HMR2, Recon3D and Human1 with Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) screen results of their respective cell lines. The gene essentiality predictions of Human1 derived CSMs exhibited the highest Matthews correlation coefficient (MCC) values, outperforming the other template models [4].

2.1.3. Principles of constraint-based modelling

CBM is primarily based on a numeric matrix, containing the stoichiometric coefficients of all reactions in a metabolic network, and the mass balance constraints they impose on how reactions operate, and how metabolites are produced and consumed [25]. The simplicity of constraint-based models, coupled with the smaller computational burden and their suitability for the integration of omics data layers, enables model reconstruction and analysis at the genome-scale [1].

Fluxes represent the generation and/or depletion of metabolites by transport and enzymatic reactions, and the rates at which they occur. The difference between the rate of production and consumption of a metabolite equates to its change of concentration over time. As growth has a much larger time scale than reaction kinetics, studying metabolism in steady state is a reasonable assumption [26]. At steady state, the net concentration of metabolites (vector x of size m) is constant (Equation 1) and the law of conservation of mass is interpreted as equal generation and depletion fluxes for a given metabolite.

$$\frac{dx}{dt} = 0 \qquad (2.1)$$

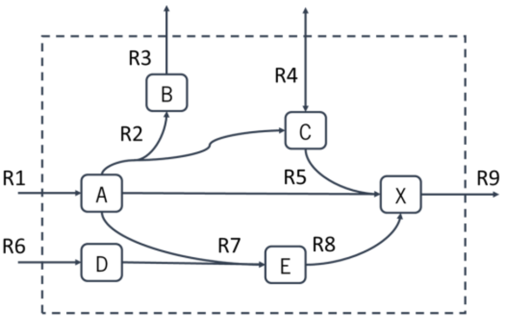
The stoichiometric (S) matrix (*m* rows of metabolites x *n* columns of reactions) represents all coefficients in metabolic reactions, connecting metabolites to their corresponding reactions, with positive and negative coefficients depicting production and consumption, respectively (**Figure 1**). Assuming a flux value vector *v* with *n* reactions, the system of mass balance equations at steady state can be defined by Equation 2:

$$S \cdot v = 0 \qquad (2.2)$$

Any flux vector *v* satisfying this equation makes up the constrained solution space. Fluxes in *v* are also constrained by upper and lower bounds which can be based on experimental data or network topology alone. The reaction constraints and the biomass reaction objective defined in Equation 2 describe a system of linear equations that can be solved by linear programming (LP), since these systems are typically underdetermined (*n* > *m*) [25]. A biological objective (Equation 3) is required for such approaches, formally represented as an objective function *Z* to maximize or minimize, with *c* representing a vector of weights of the reaction's contributions to the objective function:

$$\begin{aligned} &\text{maximize:} && Z(v) = c^T v \\ &\text{subject to:} && S \cdot v = 0 \qquad (2.3) \\ & && l_i \leq v_i \leq u_i, \forall_i \in \{1, \dots, n\} \\ & && c, v, l, u \in \Re^n, S \in \Re^{m,n} \end{aligned}$$

The system of equations and objective function definitions set the principles for Flux Balance Analysis (FBA), a method attempting to find the flux distribution(s) within the solution space that optimizes the defined objective [25].



$$\begin{array}{c|ccccccccc|}
 & R1 & R2 & R3 & R4 & R5 & R6 & R7 & R8 & R9 \\
A & 1 & -1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\
B & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
C & 0 & 1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\
D & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\
E & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\
X & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1
\end{array}$$

Figure 1 - Toy metabolic network adapted from Klamt & Gilles [27], with the corresponding S matrix. Boxes and arrows represent metabolites and reactions, respectively.

2.1.4. Phenotype Prediction

GSMM simulation has commonly relied on FBA to estimate cellular fluxes by assuming a cellular objective function to maximize or minimize [13]. Biomass maximization as an objective function for CBM is widely accepted in microbial and cancer models. However, defining growth as a main objective for human cells in typical physiological conditions may not lead to accurate intracellular flux predictions [1]. Growth rate varies between specialized cells, some of which generally do not divide after differentiation (skeletal muscle cells, neurons) or divide more frequently (fibroblasts, smooth muscle cells, liver cells) [28]. Cells may also have multiple objectives, competing or simultaneous, or settle on an evolutionary optimal compromise between objectives [1]. Several extensions of FBA have been developed to address and mitigate these issues.

Parsimonious Flux Balance Analysis (pFBA) adds a second optimization problem based on the primal objective value obtained with FBA, which minimizes the absolute sum of fluxes, yielding distributions that are optimal with the least amount of flux across all reactions [29]. This subset generally includes more efficient reactions or pathways, as pFBA favours activated genes that translate into fewer enzymatic steps [1, 30]. As there are many possible flux distributions resulting in the same objective value, especially with more complex models, pFBA is used to reduce the space of solutions [29].

Minimization of internal fluxes has also been used to predict flux distributions of healthy human cells instead of relying on a single biomass function, to validate if over 400 metabolic objectives could achieve non-zero stationary flux distributions in the liver specific HepatoNet1 model [3, 31].

Flux variability analysis (FVA) is used to establish the minimum and maximum fluxes for reactions after setting a minimum objective value, usually a fraction of the maximum theoretical biomass flux. FVA can be used to explore alternate optima, study flux distributions under suboptimal growth, investigate a model's flexibility and robustness and to identify essential reactions [32]. For instance, Robinson *et al.*

(2020) used FVA to compare the solution space between enzyme-constrained and regular CSMs derived from Human1 [4].

2.2. Omics data

A cell's phenotype is influenced by environmental conditions and the complex connections between several biomolecules [1]. The identification and quantification of these molecules can be achieved using the various omics techniques covering important biological layers, such as genomics, transcriptomics, proteomics, metabolomics and fluxomics.

2.2.1. Transcriptomics data sources

Genomics characterize the genetic code of a cell in the form of genomes. Transcriptomics, or the messenger RNA (mRNA) levels coded by the genome, provide a snapshot of dynamic gene expression activity considering development stage, environmental condition, and tissue type [33]. Proteomics represent the set of proteins produced by expressed genes [1], which undergo post-translational modifications that cannot be predicted solely with mRNA levels. Metabolomics represent the set of metabolites present in a cell, providing more information regarding enzyme activity and metabolic regulation. Finally, fluxomics involve quantifying fluxes, or the rates of metabolic reactions. These are traditionally measured using isotopic markers and may change without affecting the levels of metabolite intermediates [33].

DNA Microarrays and RNA Sequencing, or RNA-Seq, are two of the most popular methods for transcriptome profiling. Microarray technology can generally be divided into two phases, namely probe, and target complementary DNA (cDNA) production. After converting the mRNA to the more stable cDNA, the sequences are labelled with fluorochrome dyes and bound to a surface. As specific probes hybridize with the labelled targets, the signal identifies which mRNA sequences are present in the sample [35]. RNA-Seq technology quantifies expression by sequencing a cDNA library of all the RNA molecules transcribed by a certain tissue or cell type. These transcripts are then mapped to a reference genome [36].

One of the most common applications of RNA-Seq is the identification of differentially expressed genes between at least two conditions. The total read count, or sequencing depth, is fixed before sequencing. As such, the expression level of mRNA transcripts is measured by the proportion of total number of reads, or its abundance level. In addition to highly expressed transcripts having a correspondingly higher number of mapped reads, longer transcripts also have more mapped reads than shorter transcripts of comparable expression levels. Consequently, several normalization methods for RNA-Seq data exist to correct for library size, transcript length and GC-content bias [37].

Examples of read count normalization methods include per-sample Total Counts (TC), per-sample Upper Quartile (UQ), per-sample Median (Med), DESeq normalization (median-of-ratios), Trimmed Mean of M values (TMM), Reads Per Kilobase per Million mapped reads (RPKM) and Fragments Per Kilobase per Million (FPKM). Li *et al.* (2017) evaluated the performance of these methods and proposed two more, per-gene normalization after per-sample median (Med-pgQ2) and upper-quartile global scaling (UQ-pgQ2) [37].

As high-throughput methods such as next-generation sequencing (NGS) grew more affordable, more databases containing publicly available genomes and transcriptome datasets arose, such as GenBank [38], the Genotype-Tissue Expression (GTEx) project [39] and The Cancer Genome Atlas (TCGA) [40] databases, respectively. In addition, RNA-Seq is often favoured over microarray techniques, as RNA-Seq can identify low abundancy RNA and splice variants without requiring prior knowledge of the organism's genome [33, 36].

The TCGA project was launched in 2006 by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), and now provides data across 33 tumour types retrieved from over 11000 patients. This extensive dataset facilitates the study of specific genomic and molecular changes in cancer, the definition of a relevant taxonomy of cancer types and subtypes and the identification of potential targets for treatment [40].

Since its launch in 2010, the GTEx project has offered a catalogue of gene expression and its effects across many human tissues, with the goal of enlightening regulatory genetic variation and genetic associations with complex diseases. After quality control, the GTEx v8 dataset has a total of 838 donors and 17382 samples derived from 52 tissues and 2 cell lines [39]. Robinson *et al.* (2020) used GTEx RNA-Seq data alongside cancer RNA-Seq data from the TCGA project to extract CSMs and study human physiology and disease [4].

Particularly, the GTEx project uses Transcript per Million (TPM) [39], a type of within-sample normalization method meant to improve upon RPKM, which may not remove all length bias in gene expression or be overly influenced by relatively few transcripts. In contrast, TPM normalization adjusts for library size and gene length, in that order, and scales all samples to a common total sum of TPM values, making gene expression across samples more comparable than with RPKM normalization [41].

2.2.2. Omics data pre-processing

As with reconstruction algorithms, no universal pre-processing method for integrating transcriptomic data with metabolic networks exists. Richelle *et al*. (2019b) defines three categories of pre-processing methods. Firstly, how to approach genes and reactions without a one-to-one relationship (such as isozymes), or gene mapping. Secondly, how to define a gene as expressed, or a gene expression threshold. Lastly, the order of gene mapping and thresholding integration, in other words, if the activity cut-off is defined at the reaction or gene level [42].

Omics data pre-processing is necessary because algorithms operate at the reaction level. For this reason, gene mapping methods employ a combination of GPRs and gene expression data, to establish a reaction's enzyme activity. In the case of RNA-Seq data, absolute mRNA abundance is often used to define a gene expression threshold above which a gene is assumed to be active. This concept can be applied using single or multiple thresholds. For example, global thresholding applies a single unique threshold to all genes, while the contrasting local thresholding approach applies a different threshold per gene. A common local threshold is the mean expression value of a gene across samples [42].

Richelle *et al*. (2019b) compared the ability to capture metabolic pathways' observed ubiquity (ubiquitous, tissue-specific, or organ-specific) of various percentile-based thresholding methods, using principal component analysis (PCA). In addition to the contrasting global and local threshold definitions, the authors analysed the influence of the number of reactions states (ON/OFF, OFF/MAYBE ON or ON/MAYBE ON/OFF). The 3 reaction states are defined by a combination of global and local thresholds, where the gene expression is defined by the mean across all samples (local) but must be in between certain lower and higher percentiles (global). In summary, traditional (ON/OFF) global thresholding resulted in fewer differences between tissues and a higher false-negative rate, while the local25-75 thresholding method, which used the 25th and 75th percentiles of gene expression as lower and upper bounds, appeared to perform best [42].

2.3. Context-specific metabolic model reconstruction

Lower resolution models, like generic human models, aim to characterize a broader system but are less adequate for more specific variation. In contrast, with higher resolutions comes greater confidence in the model's biological accuracy [13]. A common application of CSMs of human tissues is the comparison of healthy and diseased models, such as cancer, to better understand both phenotypes.

Unlike GENREs, for which biological validation (or accuracy) is part of a long process of manual curation [14, 13], reconstruction algorithms and large omics data sets have allowed the batch generation of many draft CSMs at a time [3]. Methods used for biological validation of CSMs include metabolic tasks [4, 2, 31], gene essentiality screens [4, 2, 33] and fluxomics data [11].

2.3.1. Reconstruction algorithms

Reconstruction algorithms use omics data to extract context-specific models from template GSMMs in an automated way, through reaction removal. In general, the major decisions to consider when extracting context-specific models are how to constrain uptake and secretion fluxes and which template GSMM, gene expression threshold and reconstruction algorithm to use [33].

According to Robaina *et al*. (2014), reconstruction algorithms can be classified into 3 major groups (**Figure 2**): Gene Inactivation Moderated by Metabolism and Expression (GIMME)-like, integrative Metabolic Analysis Tool (iMAT)-like and Model-Building Algorithm (MBA)-like algorithms [43].

The GIMME-like family (GIMME, GIMMEp and GIM^{3}E) minimizes fluxes associated with low gene expression, while guaranteeing the Required Metabolic Functionality (RMF) objective, like growth or ATP production [33, 43]. As RMF evaluation typically uses FBA, selecting a RMF objective for multicellular eukaryotes is complicated and biomass maximization may not be adequate. GIMME is an LP approach that minimizes an inconsistency score function that penalizes reactions with expression levels (obtained through GPR values) beneath a certain user-defined threshold [44]. The original GIMME algorithm focuses on transcript profiles, but variants allow for the integration of proteomic (GIMMEp) [45] and metabolomic (GIM^{3}E) [46] data, respectively.

The iMAT-like family (iMAT, INIT and tINIT) reconstructs models based on experimental data without depending on RMF, instead matching reaction states to related data expression using Mixed Integer Linear Programming (MILP) [33, 43]. The Integrative Network Inference for Tissues (INIT) [47] algorithm integrates data directly in the objective function, including metabolomics data, while iMAT [48] does so in the constraints.

The tINIT extension to INIT [49] adds metabolic tasks representing production or consumption of metabolites, or the activation of certain pathways depending on model context. These tasks ensure the inclusion of their required reactions but not necessarily the smallest reaction set. Some tasks may also be redundant, to allow for a finer step-by-step analysis and reduce computational costs [49]. Despite their independence of RMF (with the exception of tINIT, where it is an optional parameter), the iMAT-like family uses MILP problems that are more computationally taxing than LP, specially iMAT as it solves two MILP problems per reaction [43, 49].

The MBA-like family (MBA, mCADRE and fastCORE) takes sets of reactions categorized as core (higher likelihood of being active) and non-core, after which the methods attempt to keep the model's consistency while removing or including non-core reactions [33, 50, 51, 52]. The Metabolic Context Specificity Assessed by Deterministic Reaction Evaluation (mCADRE) [50] algorithm scores reactions according to expression, connectivity, and confidence-level, defining them as core or non-core.

The MBA [51] algorithm subdivides the core set of reactions into 2 sets, based on the likelihood of being active. As MBA is affected by the order in which non-core reaction are removed, the algorithm is repeated, and the reactions ranked to form a consensus model. Unlike the stochastic MBA algorithm, mCADRE does not demand all core reactions to be included in the model. Instead of iteratively removing non-core reactions coupled with a consistency assessment, the Fast Consistent Reconstruction (fastCORE) [52] algorithm solves two LPs. After it maximizes the number of core reactions, comparing their values to a constant, the second LP minimizes the number of non-core reactions until core coherency is achieved. The fastCORMICS extension adds microarray data processing to the original algorithm [53].

The main advantages of the MBA-like family are the ability to integrate various types of data without needing to explicitly define a set of gene scores (instead relying on core reactions) compounded by the RMF independence of the iMAT-like family. In particular, fastCORE outperforms other MBA-like methods in terms of computation time [43]. The Cost Optimization Reaction Dependency Assessment (CORDA) algorithm is similar to the MBA-like family in its use of highly expressed core reactions but uses an artificial metabolite cost function [54].

Opdam *et al*. (2017) reported the choice of algorithm has the largest impact on model accuracy in gene essentiality predictions. The authors report greater accuracy in gene essentiality predictions using the INIT, MBA, and mCADRE algorithms for model extraction, particularly when considering stringent gene expression thresholds (top 10% and mean) [33]. Nonetheless, no single algorithm outperforms the others in all cases and no standard algorithm exists [33, 43, 11, 30], reinforcing the importance of appropriate gene expression thresholds [30].

On the other hand, progress has been made towards bridging the gap between different extraction methods. For example, Richelle *et al*. (2019a) proposed a framework of metabolic tasks inferred from omics data, prior to model reconstruction. The authors reported that the protection of reactions required by tasks reduces variability in the resulting models extracted using different algorithms. For example, if the model extraction relied on the fastCORE algorithm, the reactions required for task success would be included in the core set [2].

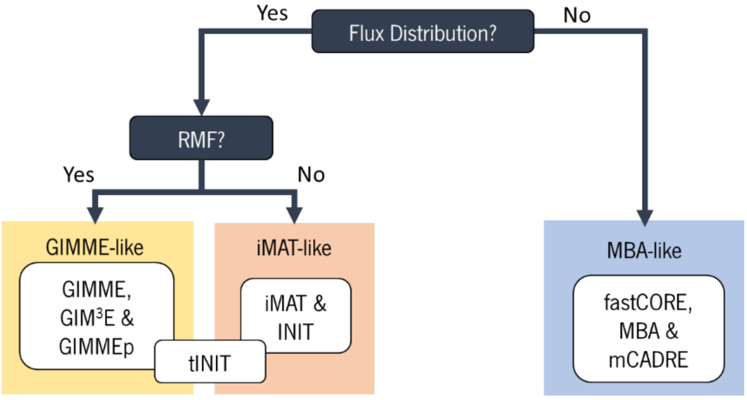


Figure 2 – The 3 major families of reconstruction algorithms, adapted from Robaina *et al.* (2014) [43]. The GIMME-like and iMAT-like families return a flux distribution as well as an extracted model. The GIMME-like family depends upon RMF, whereas the iMAT-like family does not. For the tINIT extension, it is optional.

2.3.2. Model Validation

GPRs enable *in silico* gene perturbation experiments, such as simulating phenotypes for essential gene deletions, that may be used to evaluate GSMM prediction accuracy when compared with experimental gene essentiality data. Essential genes predictions are commonly achieved by testing if a given gene, when deleted *in silico* sufficiently reduces the chosen biomass objective function in a simulation using the GSMM [4].

As was previously mentioned, objective functions defined as cell growth may be less appropriate for healthy human cells, which vary in growth rate [4, 28]. A less restrictive approach defines essential genes as those required for basic metabolic tasks necessary for cell viability. This broader definition is also estimated to increase prediction sensitivity, as it takes more metabolic functions into account [4].

According to Thiele *et al.* (2013), such metabolic tasks can be defined as the nonzero flux through a reaction or pathway, leading to the production of a target metabolite [18]. Metabolic tasks have since been used as tools for model benchmarking and comparison between different extraction methods, with several lists of tasks published [18, 49, 2].

For example, the set of liver-specific metabolic objectives used by Gille *et al.* (2010) was divided into network and physiological tasks. Each task consisted of two sets of metabolites, input and output, and the target metabolite to be produced [31]. Robinson *et al.* (2020) presented their own set of tasks adapted from Agren *et al.* (2014) [49], including a set of 57 essential tasks common to all human cell types that models based on Human1 are expected to pass (available in the Github repository[1]). These

[1] https://github.com/SysBioChalmers/Human-GEM

essential tasks were divided into general categories, namely re-phosphorylation of nucleoside triphosphates, *de novo* synthesis of nucleotides, key intermediates phospholipids, vitamins and co-factors and other compounds, uptake of essential amino acids, protein turnover, electron transport chain and tricarboxylic acid cycle, beta oxidation of fatty acids and growth (or feasible biomass production) [4].

Typically, metabolic fluxes that cannot be measured directly may be estimated *in silico* based on omics data. Popular computational algorithms used for this purpose include FBA, metabolic flux analysis (MFA) and ^{13}C MFA. In MFA, measured extracellular fluxes over time are used as input to calculate intracellular fluxes reactions, applying the stoichiometric model representation and steady state assumption likewise used by FBA. Experimental flux measurement commonly uses higher atomic mass isotopes, such as ^{13}C instead of ^{12}C, to label carbon and infer flux patterns. However, ensuring their accuracy is often difficult and usually limited to constraining smaller-scale metabolic models [55]. Fluxomic data are also rarer for human cells, for which data are often only available for cancer [1], such as the NCI-60 cell lines [56].

2.3.3. Applications

This work will focus on metabolism and models of healthy human tissues. This section will cover 3 candidate tissues, taken as putative case studies, specifically the liver, kidney, and breast. Human tissues and their associated metabolic diseases make good modelling targets. After the first generic human models were published, several hepatocyte, adipocyte, and kidney CSMs followed.

Gille *et al.* (2010) [31] focused on nutrient and oxygen availability and their effect on ammonia detoxification in the liver, constructing the liver model HepatoNet1. HepatoNet1 was based on Recon1, manually curated, and comprised of 777 metabolites and 2539 reactions. This tissue-specific model marked a milestone in liver modelling efforts and was subsequently incorporated into later generic models [18]. Another example of a liver CSM is iHepatocyte2322 [57], extracted from HMR2 as a template model using INIT and with a total of 5686 metabolites and 7930 reactions. iHepatocyte2322 integrated lipid metabolism, merged previous hepatocyte models and provided insight into non-alcoholic fatty liver disease (NAFLD) and steatohepatitis (NASH) [57].

Mardinoglu *et al.* (2013) [58] published a manually curated CSM for adipocytes followed by an updated version in 2014 [59], named iAdipocytes1809 and iAdipocytes1850, respectively. The adipocyte CSMs were used to study differential metabolic activity in lean and obese subjects and to identify potential therapeutic targets for obesity. iAdipocytes1809 was built upon adipocyte-specific proteome and subcutaneous adipose tissue microarray data, with a total of 6160 reactions and 4550 metabolites.

iAdipocytes1850 updated model content using RNA-Seq data, to a total of 6230 reactions and 4577 metabolites.

Zhang *et al.* (2013) [60] published a kidney CSM extracted from Recon1 as a template model using MBA, with a total of 2904 reactions and 1898 metabolites. The CSM was used to study kidney-related disease metabolism and prediction of biomarkers for early diagnosis, focusing on diabetic kidney disease. Sohrabi-Jahromi *et al.* (2016) [61] also applied a kidney CSM to study disease and predicted possible drug targets, specifically for Focal Segmental Glomerulosclerosis (FSGS). The authors merged two previous kidney models, to a total of 3034 reactions and 1996 metabolites.

Finally, publications presenting reconstruction algorithms are also a source of CSMs. For example, Agren *et al.* (2014) published the tINIT algorithm alongside 83 healthy cell type-specific GEMs, extracted from HMR2 as the template model [49]. These models and Human1 are available in the Metabolic Atlas repository, which includes CSMs of liver hepatocyte and bile duct cells, breast adipocyte and glandular cells, kidney glomeruli and tubule, and soft tissue adipocyte models [4].

3. Methods

The Methods section covers methodological and technical details pertaining to the three stages of the pipeline developed in this work, specifically data pre-processing, model extraction and validation, with the latter focusing on metabolic task evaluation.

As most pre-processing is specific to the case study data used during pipeline development, it is detailed in Section 4 (Results) instead. In particular, the pre-extraction phase of the general pipeline includes blocked reaction removal from the generic model, genes scores threshold selection (for multi-tissue datasets), 3 types of RNA-Seq data aggregation methods, gene expression data conversion to gene scores of the chosen threshold and determination of reactions required by essential tasks.

As of November 2020, the *Troppo* Python package has the fastCORE, CORDA, GIMME, (t)INIT and iMAT algorithms implemented [8]. The extraction step utilizes the fastCORE and tINIT reconstruction algorithms. Additionally, the validation step employs metabolic tasks and basic analyses of model attributes, such as reaction content. A diagram of the general pipeline is shown in **Figure 3.**

Lastly, the developed software includes 3 Jupyter Notebooks available on Github[2], which cover part of the initial data pre-processing ("omics_to_genescores.ipynb"), all other statistical analyses and figures presented in this work ("results_graphs_clean.ipynb") and a quick tutorial on how to extract CSMs with the *Troppo* package ("extraction_example_guide.ipynb").

[2] https://github.com/MariaPessoa/thesis_annexes

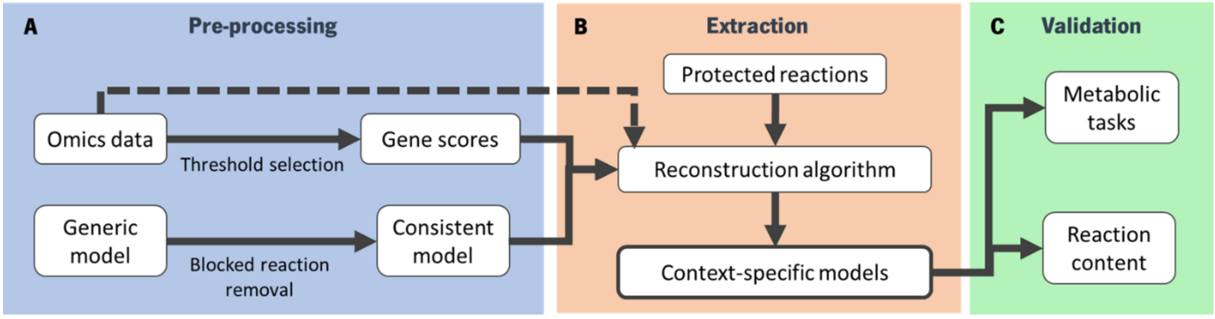


Figure 3 – Overview of the pipeline presented in this work. The pipeline can be divided into 3 phases, pre-processing, extraction, and validation. In the pre-processing phase (**A**), blocked reactions were removed from the generic model to obtain a consistent (template) model. The omics data was converted into gene scores after threshold selection or passed directly to the algorithm without conversion (dashed arrow). The extraction phase (**B**) considers whether the reconstruction algorithms chosen are supplied protected reactions required by essential tasks or not. Finally, the validation phase (**C**) is based on metabolic tasks and reaction content analysis.

3.1. Model extraction

Model extraction refers to the process of the extraction of a CSM from a generic template model by a reconstruction algorithm. The *Troppo* package's model extraction pipeline requires a template model, gene scores or gene expression data and an optional protected reaction set as direct input data. Furthermore, *Troppo* requires the selection of the appropriate activity threshold, of the algorithm(s) with which to extract the CSMs and the gene mapping method.

In particular, the *Troppo* package's implementation of (t)INIT does not accept a task file to determine which reactions to protect, but instead accepts a protected reaction set, similarly to fastCORE's core reactions. After extraction, the pipeline outputs the reaction content of all CSMs.

Specifically, gene mapping methods address genes and reactions without a one-to-one relationship, such as enzyme complexes (AND rule) and isoenzymes (OR rule). Two popular gene mapping methods incorporate a minimum expression value of all genes associated to an enzyme complex and either a maximum expression value (MINMAX) or sum of the expression values (MINSUM) of all genes related to an isoenzyme [42]. The MINMAX gene mapping method was used for all CSMs extracted in this work.

Drug-enzyme interactions mapped in GSMMs allow for the simulation of drug-related processes [18]. Extraction algorithms may keep such reactions in CSMs because they produce necessary intermediary metabolites, and, as such, may need to be removed from the template model prior to extraction. Consequently, a set of exchange reactions from the Drug metabolism subsystem and the template model's blocked reactions were removed. In addition, the blocked reactions were identified with the COnstraint-Based Reconstruction and Analysis for Python (COBRApy) package's [62] *find_blocked_reactions* function. The consistent version of the model without the Drug metabolism exchange reactions (n=11386 reactions, SBML format) was used as template for all extracted CSMs unless stated otherwise.

The *checkTasks* function of the MATLAB RAVEN 2.0 toolbox [63] was used to determine the reactions required by the essential tasks. Firstly, the (original) SBML version of Human-GEM was loaded using the MATLAB COBRA toolbox [64] function *importModel*, followed by conversion to RAVEN format using the *ravenCobraWrapper* function. Secondly, boundary metabolites were added using the *addBoundaryMets* function (available in the Human-GEM repository) and the essential metabolic tasks file was loaded with RAVEN's *parseTaskList* function. The original model and the essential tasks file were passed as arguments to the *checkTasks* function, which returns a sparse matrix (reactions x tasks) with non-zero values representing the required reactions.

Additionally, a reaction content matrix (detailing whether the template model's reactions are active or inactive in the CSM) was generated for the models of comparable tissue extracted by Robinson *et al.*, which were in MATLAB format (available on Zenodo[3]). The *compareModelField* function, contained within RAVEN's *compareMultipleModels* function, was extracted into a separate file, and used for this purpose[4].

3.2. Model validation with task evaluation

Task evaluation relies on the *Troppo* package's task parsing module and separate functions for metabolite nomenclature pre-processing (*get_essential_tasks* and *get_full_tasks*). The task evaluation of the SBML Human-GEM model used to determine the protected reaction set, for both sets of tasks, essential and the larger set of tasks (n= 256) available on Human-GEM's Github, was also performed using *checkTasks*. The process was repeated using the *Troppo* package to compare the two methods.

As part of the developed pipeline, machine learning (ML) methods were employed for gene scores threshold selection and are detailed in Section 4.1. The ML pipeline was additionally employed to validate CSM performance, using the reaction content as input. The objective, to identify the source tissue, was unchanged. The reaction content of the CSMs was further utilized to calculate a Hamming distance matrix, which was plotted using a t-Distributed Stochastic Neighbour Embedding (tSNE) projection (based on the

[3] https://doi.org/10.5281/zenodo.3577466

[4] https://github.com/SysBioChalmers/RAVEN/blob/master/core/compareMultipleModels.m

scipy package[5]). The learning rate was set to 5000 and the perplexity parameter to roughly 5% of the number of samples plotted.

Lastly, all statistical tests were performed using the *scipy* package unless stated otherwise and are available in the accompanying Jupyter Notebook[2] (“results_graphs_clean.ipynb”).

[5] https://www.scipy.org/index.html

4. Results

In addition to the 3 candidate tissues explored in Section 2.3.3, the liver, kidney, and breast, 7 other tissues were subsequently included to further test the pipeline. Specifically, the corresponding tissues in the GTEx v8 dataset were the adipose (subcutaneous), breast (mammary tissue), kidney (cortex), and liver samples. Likewise, the additional 7 tissues selected were the brain (cortex), colon (transverse), lung, skeletal muscle, pancreas, stomach, and whole blood.

The gene TPM and median gene-level TPM by tissue data, alongside sample attribute and phenotype subject metadata were downloaded from the GTEx Data Portal. Firstly, the samples corresponding to the four selected tissues (adipose, breast, kidney, and liver) were filtered by RNA Integrity Number (RIN, from sample attributes). Of the 1228 samples with RIN qualified for RNA sequence analysis[6] (RIN ≥ 6), the adipose tissue samples were the most numerous (592 samples), followed by the breast (390), liver (193) and kidney, which only had 53 samples.

The same RIN filter was applied to a second set of 7 tissues later added, of the (brain (cortex), colon (transverse), lung, muscle (skeletal), pancreas, stomach, and whole blood), with 3243 samples qualified for analysis in total. Of those, the skeletal muscle and brain were the most and least numerous, with 796 and 217 samples, respectively (all values provided in **Table 1**).

Table 1 – RIN filtered TPM (RIN ≥ 6) sample counts of all 11 tissues analysed, namely adipose (subcutaneous), brain (cortex), breast (mammary tissue), colon (transverse), kidney (cortex), liver, lung, muscle (skeletal), pancreas, stomach, and whole blood.

Tissue	Filtered sample count
Adipose (Subcutaneous)	592
Brain (Cortex)	217
Breast (Mammary Tissue)	390
Colon (Transverse)	350
Kidney (Cortex)	53
Liver	193
Lung	528
Muscle (Skeletal)	796
Pancreas	299
Stomach	307
Whole blood	746
Total	**4471**

[6] https://gtexportal.org/home/documentationPage

Additionally, the median values of the individual TPM sample dataset, without the RIN filter, were grouped by tissue, gender, and age group (20-year intervals), generating a set of 6 sample combinations per tissue, to test the influence of the input RNA-Seq data aggregation method used. The RIN filter was not applied because it removed all samples of certain age groups. This generated dataset is henceforth referred to as grouped samples. The median gene-level TPM by tissue data is assumed to have used all TPM samples[7] without any similar filter applied.

In total, 1615 CSMs were extracted with Human-GEM as the template model (**Table 2**). The models varied in GTEx RNA-Seq data aggregation method (TPM, grouped or median), input data pre-processing (gene expression data or gene scores, without null variance genes or with all genes), protected reactions required by the essential tasks (minimal or non-minimal) and reconstruction algorithm (fastCORE or tINIT) used, compounded by tissue-specific differences across the 11 tissues.

Table 2 – CSMs extracted per tissue and in total, by algorithm and extraction condition, namely data aggregation method (TPM, grouped or median), data pre-processing (gene expression data or gene scores, with null variance filter or with all genes) and protected reaction set (minimal or non-minimal) used. The absolute model count is in between parentheses.

	Data aggregation method						
	TPM					**Median**	
Algorithm	Minimal	Non-minimal	Non-minimal (all genes)	Gene expression data (all genes)	**Grouped**	Non-minimal	Gene expression data (all genes)
fastCORE	10 (40)	53 (583)	53 (212)	53 (212)	6 (66)	1 (11)	1 (11)
tINIT	10 (40)	53** (352)	–	–	6 (66)	1 (11)	1 (11)
Total	20* (80)	106** (935)	53* (212)	53* (212)	12 (132)	2 (22)	2 (22)

*Models were extracted for 4 of the 11 tissues, namely the adipose, breast, kidney, and liver tissues.
**As with fastCORE, 53 models per tissue (212) were extracted for the adipose, breast, kidney, and liver tissues, but only 20 models per tissue (140) were extracted for the remaining 7 tissues using tINIT.

The minimal TPM models were extracted with 3 protected reactions related to biomass (the biomass objective function, biomass_human, and 2 exchange reactions, HMR_10023 and HMR_10024). All gene expression data models rely on the protected reaction set obtained from RAVEN and are therefore non-minimal.

Likewise, the variance filter that discards genes with approximately null variance was not applied to all models prior to extraction. Gene scores TPM models were also extracted without the filter (with all genes), as were all gene expression data models. The task evaluation (for both sets of tasks) of the models

[7] https://gtexportal.org/home/documentationPage#staticTextDataProduction

of corresponding healthy tissue extracted by Robinson *et al.*, henceforth referred to as H1-CSMs, was also performed.

Lastly, the names of the tissues of median TPM GTEx dataset used by Robinson *et al.* (also available on Zenodo[3]) do not match exactly those present in the most recent GTEx TPM dataset available (v8). For example, the GTEx v8 dataset has two colon samples, transverse and sigmoid. In the authors' dataset, the relevant tissue is simply named colon. Consequently, the H1-CSMs are assumed to be of analogous (but not necessarily identical) tissue to the models presented in this work.

4.1. Pre-extraction pipeline

Prior to gene scores thresholding strategy selection, gene expression data conversion to gene scores, or CSM extraction, the RIN filtered (RIN ≥ 6) TPM gene expression data was visualized with PCA. **Figure 4** shows that the samples mainly differ by tissue of origin, as expected, with overlap between the adipose and breast and heavy overlap between the colon, pancreas, and stomach samples.

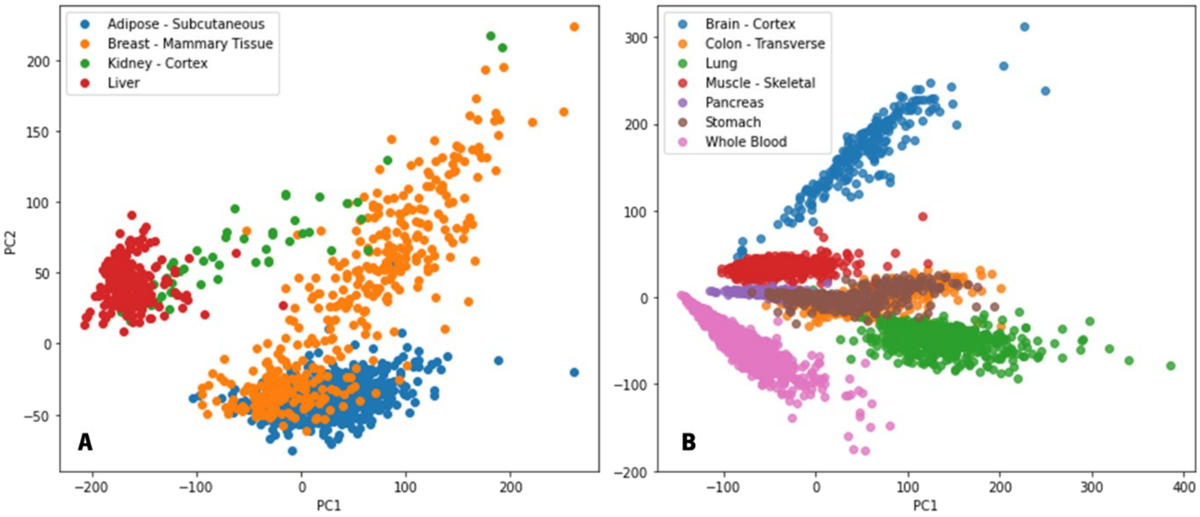


Figure 4 - Principal component analysis of the RIN filtered (RIN ≥ 6) GTEx v8 TPM gene expression dataset. (**A**) 1228 samples corresponding to 4 tissues: breast (mammary), liver, adipose (subcutaneous) and kidney (cortex). (**B**) 3243 samples corresponding to 7 tissues: brain (cortex), colon (transverse), lung, skeletal muscle, pancreas, stomach, and whole blood.

Traditionally, a gene activity score, or gene score, is assigned to all genes in a gene expression dataset to define which genes are active in each sample. After gene score conversion, a given gene is considered active if its higher or equal to a fixed activity threshold [2]. In this work, the gene score thresholding strategies used are the traditional global threshold (or global T1), which applies a single threshold to all genes in a dataset, and the local T2 threshold, which employs a combination of a local threshold and 2 global thresholds, which act as lower and upper bounds [42].

A ML pipeline was employed to determine the best gene score thresholding strategy (**Figure 5**). In total, 5 global and 50 local quantile-based thresholds were tested, based on combinations of the 10th, 25th, 50th, 75th and 90th quantiles, specifically. Random forest classification models (implemented based on the *scikit-learn* Python package[8]) were trained with 5-fold cross-validation and the objective of identifying the source tissue, and the process was repeated 20 times.

Prior to gene score conversion, genes with null variance across all samples were excluded (*VarianceThreshold* function of the *scikit-learn* package). After gene score conversion, the dataset underwent univariate K-Best feature selection (n=500 genes), which employs an ANOVA F-test to remove all but the *k* highest scoring features.

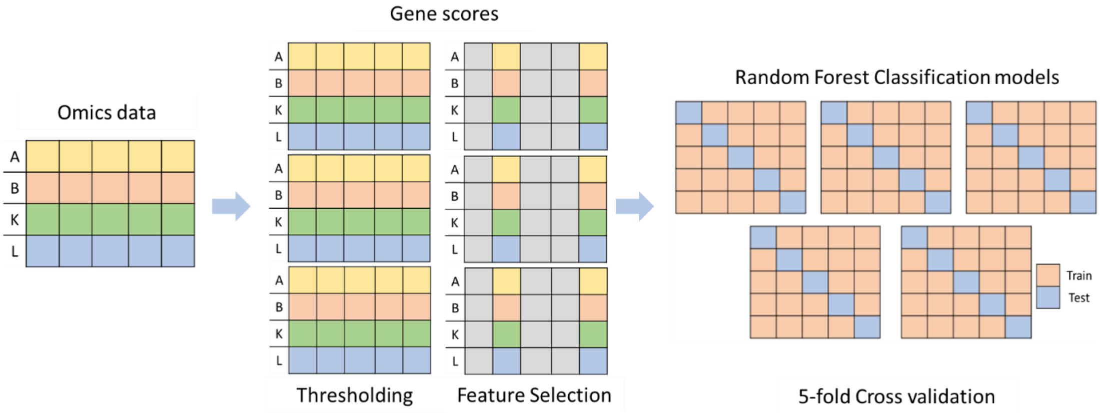


Figure 5 – Machine learning pipeline for thresholding strategy selection. Null variance genes were removed from the RIN filtered (RIN ≥ 6) TPM omics dataset, followed by conversion to gene scores and univariate K-Best feature selection (n=500 genes) to generate the input data. For every threshold, Random Forest classification models were trained and evaluated with 5-fold cross-validation and the MCC metric, with the objective of identifying the source tissue. This process was repeated 20 times for each of the 5 global and 50 local quantile-based thresholds tested.

For each threshold, the input dataset consisted of the gene scores calculated from the individual (TPM) gene expression data samples with the *global_thresholding* and *local2_thresholding* functions (available on Github[9], based on [42]). Both thresholding functions yield gene scores with an activity threshold of 0. The MCC was used as the performance metric. For reference, the same pipeline was applied to the gene expression data itself, without the thresholding step.

[8] https://scikit-learn.org/stable/

[9] https://github.com/BioSystemsUM/human_ts_models/blob/mcf7_devel/shared/src/thresholding.py

Figure 6 presents the mean cross-validation MCC values of the ML models trained with gene scores of the 4-tissue dataset (adipose, breast, kidney, and liver). All global and local thresholding methods outperformed the reference ML models trained with gene expression data, with very high (>0.90) MCC values overall. The global thresholds achieved MCC values between 0.93 and 0.95, whereas the local thresholds achieved values between 0.94 and 0.99. As the local strategy employing global50-90 and local50 thresholds had the highest mean MCC values (>0.98), all gene scores were calculated using this method unless stated otherwise.

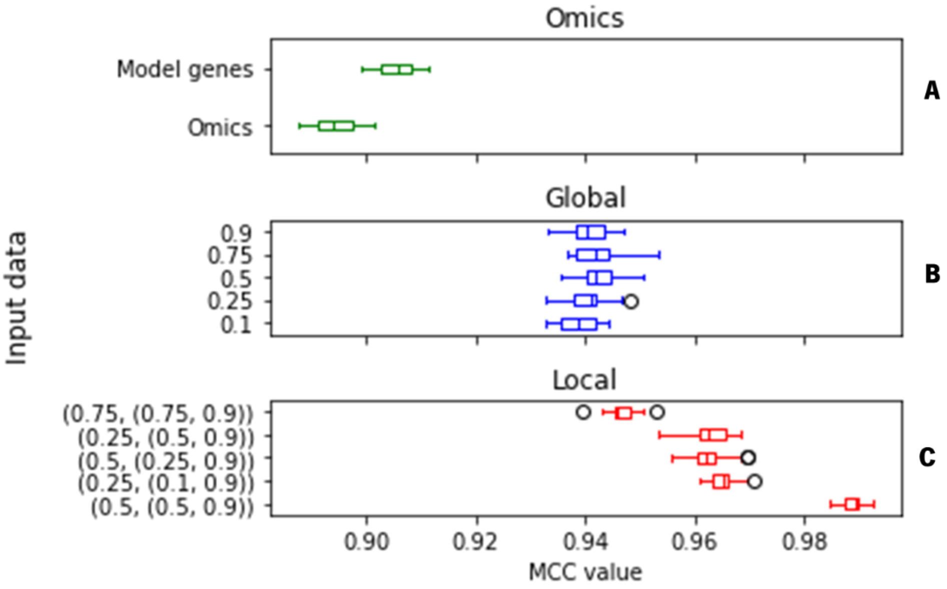


Figure 6 – Mean cross-validation MCC values from the ML pipeline for thresholding strategy selection (tissue identification) of the adipose, breast, kidney, and liver samples. (**A**) ML models trained with gene expression data. Model genes refers to the genes present in the SBML Human-GEM model. (**B**) ML models trained with global gene scores. (**C**) ML models trained with local gene scores. Only the top 5 local strategies are shown, based on maximum MCC value. The local thresholds correspond to "(local, (lower global, upper global))".

The same pipeline was later repeated for the remaining 7 tissues (figure available in the Jupyter Notebook[2] "omics_to_genescores.ipynb"), yielding extremely high (>0.99) MCC values regardless of input data or threshold type. Additionally, the ML models trained with gene scores of the 7 tissues did not outperform the corresponding gene expression data trained models. Consequently, the same local threshold chosen for the 4 tissues was used.

Lastly, the task evaluation of the *Troppo* and RAVEN packages were compared. Both approaches determined that the (original) SBML Human-GEM could achieve all essential tasks, but the task evaluation of the full set of tasks differed. According to the RAVEN package's *checkTasks* function, the model can fulfil all tasks in the larger set, but in *Troppo's* evaluation only 241 of the 256 tasks were passed. The 15 failed tasks all involved complete oxidation except for Tryptophan uptake and Aspartate degradation (also detailed in an accompanying Jupyter Notebook, “results_graphs_clean.ipynb”).

The task evaluation of the H1-CSMs, for both sets of tasks, was also performed using RAVEN and *Troppo*. The *checkTasks* essential task evaluation of the models established that all 11 models passed all tasks. The *Troppo* essential task evaluation determined that all models failed the GTP *de novo* synthesis task, and that the blood model also failed the Glucose 6-phosphate *de novo* synthesis task. Likewise, the *checkTasks* task evaluation of the full set of tasks consistently yielded slightly more passed tasks than *Troppo*'s, except for the breast model. A nonparametric one-tailed (“greater”) Mann-Whitney test for independent samples supported the tendency (U=95.5, p≈0.01). In total, 95 of the 256 tasks (37%) varied with the approach used in at least one model.

4.2. Reaction content

The reaction content of a CSM can be represented by a binary matrix representing whether a given reaction from the template model is active or inactive. Additionally, reactions are categorized by their respective metabolic subsystem and may be used to detect biological differences between model types.

Firstly, a Hamming distance matrix was calculated with the reaction content of the TPM and grouped models and plotted using a tSNE projection, by algorithm. Although it separated the two tissue sets, the overlapping tissues of the fastCORE TPM models (**Figure 7A**) differ only slightly from those observed when using the original gene expression data to conduct PCA (**Figure 4**), which showed overlap between the adipose and breast samples, as well as the stomach, colon, and the pancreas instead of the lung samples. Correspondingly, the tINIT TPM models (**Figure 7B**) reproduce the overlap between the adipose, breast, lung and a few of the colon and stomach samples.

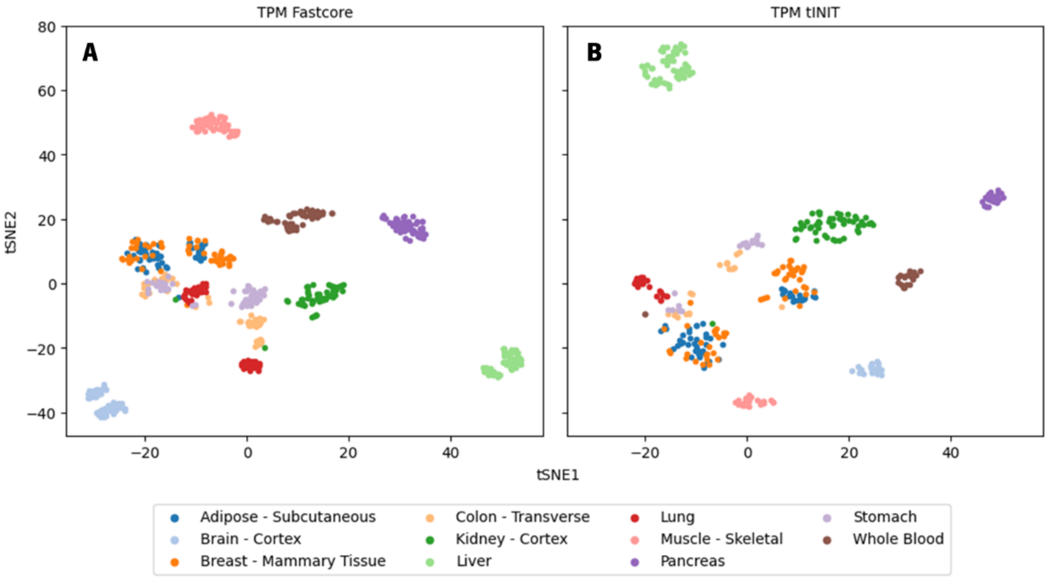


Figure 7 — Hamming distance of the non-minimal (gene scores) TPM models' reaction content visualized using tSNE projection. (**A**) Non-minimal fastCORE TPM models. (**B**) Non-minimal tINIT TPM models.

In contrast, the tSNE projection of the grouped models (**Figure 8**) only reproduces the similarity between the adipose, breast and lung models, regardless of algorithm, with otherwise good separation between tissues.

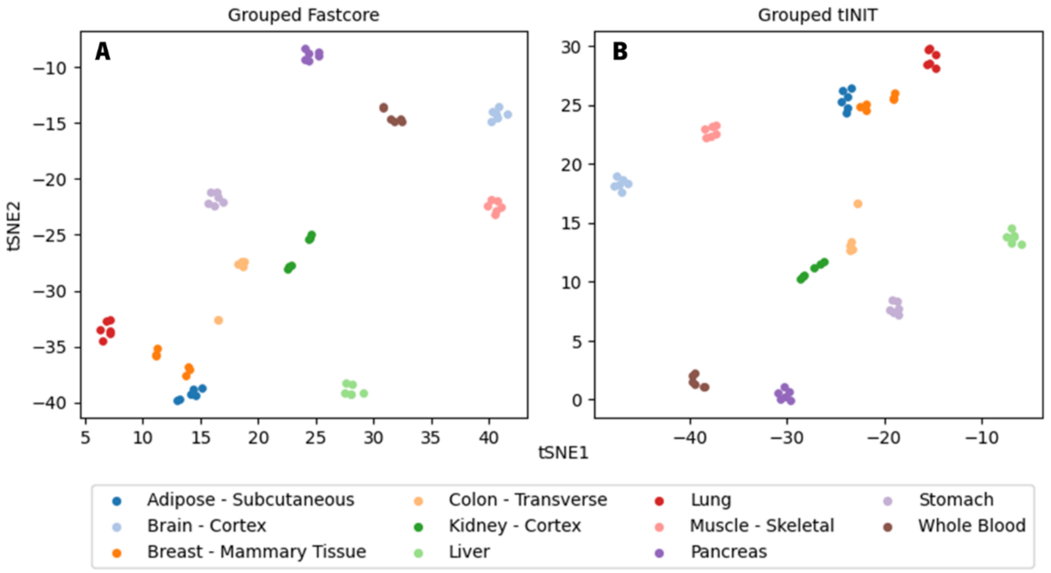


Figure 8 — Hamming distance of the grouped models' reaction content visualized using tSNE projection. (**A**) fastCORE grouped models. (**B**) tINIT grouped models.

In addition to a tSNE projection based on reaction content, ML models were trained using non-minimal TPM metabolic models' reaction content as input, with the objective of identifying the source tissue (**Figure 9**). ML models were trained for each tissue group individually and with models of all tissues, by algorithm, with all 11 tissues and the two tissue datasets separately. The 4-tissue group is composed of the adipose, breast, liver, and kidney tissues.

The best performing ML models were trained with the 7-tissue metabolic models' reaction content (**Figure 9C**), with both algorithms' ML models achieving a maximum cross-validation mean of 0.97. On the other hand, the 4-tissue tINIT models (**Figure 9B**) outperform their fastCORE counterparts, with the latter having the worst performance of all ML models trained based on reaction content, with a maximum cross-validation mean of 0.82. Finally, the ML models trained with the non-minimal TPM metabolic models of all tissues (**Figure 9A**) exhibit MCC values in between the previous two, with fastCORE outperforming tINIT instead.

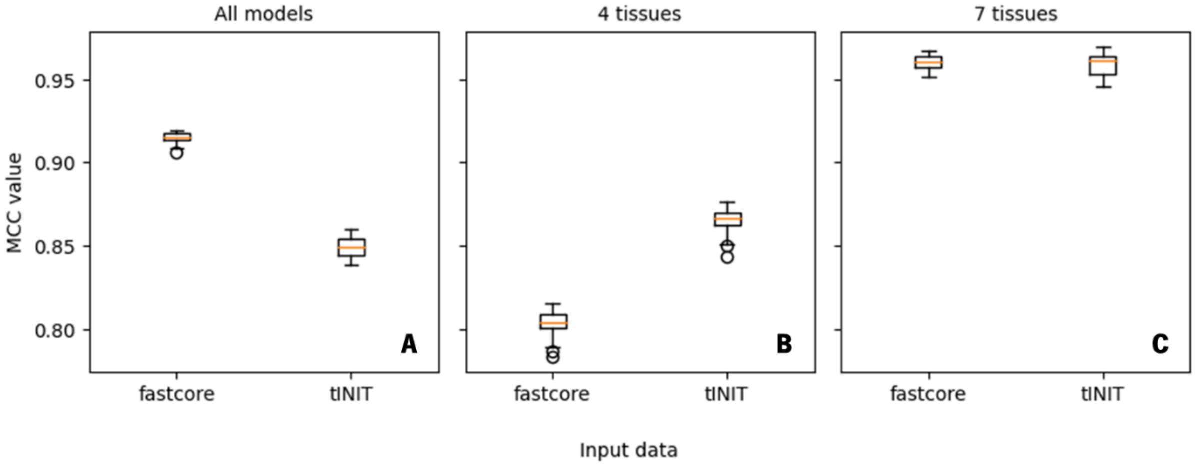


Figure 9 – Mean cross-validation results from the ML pipeline using reaction content of the non-minimal TPM models, by algorithm. (**A**) ML models trained with the non-minimal TPM models of all tissues. (**B**) ML models trained with the non-minimal TPM models of the adipose, breast, liver, and kidney tissues. (**C**) ML models trained with the non-minimal TPM models of the remaining 7 tissues (brain, colon, lung, skeletal muscle, pancreas, stomach, and whole blood).

Lastly, the absolute reaction count of the CSMs was investigated. **Figure 10** plots the mean reaction count of the (non-minimal) TPM models, by tissue and algorithm. Out of all non-minimal TPM tissue and algorithm combinations, the fastCORE whole blood and pancreas models had the lowest mean reaction counts (4891 and 4919) and the tINIT lung and colon models had the highest (8688 and 8227, respectively). Similarly, the grouped and median models had reaction counts ranging from 4101 (fastCORE grouped whole blood) to 8655 (tINIT grouped lung). Regardless of data aggregation method or tissue, all models extracted with tINIT had higher mean reaction counts than their fastCORE counterparts.

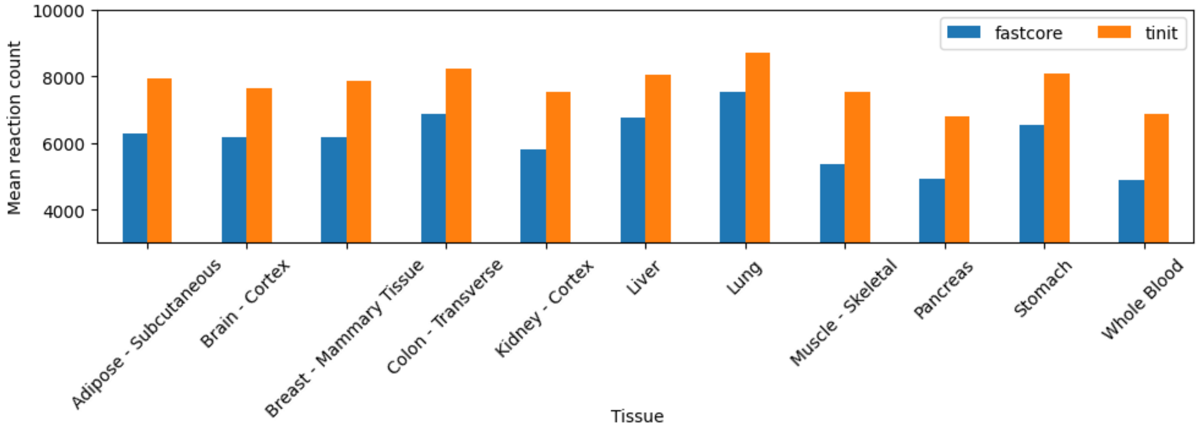


Figure 10 – Mean reaction count of the non-minimal (gene scores) TPM models of all tissues, by tissue and extraction algorithm.

A Shapiro-Wilk normality test ascertained that only the median models' reaction counts (fastCORE: W≈0.89, p≈0.14; tINIT: W≈0.93, p≈0.43) were of approximately normal distribution. A nonparametric Kruskal-Wallis H-test confirmed that the reaction counts of the 6 combinations of data aggregation method and algorithm model types were also significantly different (H≈590.90, p≈1.88e-125). A Dunn's test with multiple test p-value adjustment established that the non-minimal gene scores models' total number of reactions followed a distinct pattern, where all models differed significantly by algorithm, regardless of RNA-Seq data aggregation method (**Table 3**). Therefore, the test supports that tINIT consistently produces CSMs with higher reaction counts than fastCORE.

Table 3 – Dunn's test (*scikit_posthocs* package) with Benjamini/Hochberg (non-negative) multiple test p-value adjustment for the differences in **reaction count** of the 6 data aggregation method (TPM, grouped/GRP and median/MED) and algorithm (fastCORE/FT and tINIT) combinations of non-minimal models. P-values in cells coloured green are significant (p < 0.05) and rounded to 2 decimal places.

	tINIT TPM	**FT MED**	**tINIT MED**	**FT GRP**	**tINIT GRP**
FT TPM	1,71E-108	2,86E-01	1,34E-04	1,37E-01	1,16E-24
tINIT TPM	–	2,29E-09	3,90E-01	8,08E-37	3,28E-01
FT MED	–	–	3,20E-04	6,60E-01	2,56E-07
tINIT MED	–	–	–	2,49E-05	6,60E-01
FT GRP	–	–	–	–	6,76E-19

4.3. Essential tasks

Robinson *et al.* provided a set of 57 essential tasks that models of healthy tissue based on Human1 are expected to pass, and their evaluation returns a pass/fail binary dataset. Consequently, the essential task evaluation of the CSMs enables a basic comparison of the algorithms' performance and to assess the effect of the protected reactions and of the RNA-Seq data used.

The minimal CSMs (**Figure 11A**) extracted using the tINIT algorithm had a higher mean number of essential tasks passed (23) in comparison to fastCORE (12), which vary in reaction count but not in tasks passed. The corresponding fastCORE models extracted with protected reactions (**Figure 11B, full opacity**) show an improvement in model quality, with a higher mean of essential tasks passed (29). By comparison, the non-minimal tINIT models' quality improvement is much less pronounced (25).

However, when considering all non-minimal TPM models (**Figure 11B**), the mean number of passed tasks is only 26 and 22 out of 57 for fastCORE and tINIT models, respectively. After excluding the fastCORE and tINIT models with less than 6000 and 7000 reactions, the means improve slightly, to 30 and 23 tasks passed. The median (**Figure 11C**) and grouped (**Figure 11D**) models follow the same trend, with fastCORE outperforming tINIT. On average, the fastCORE and tINIT median models passed 27 and 22 tasks, while the grouped models passed 28 and 23 tasks.

Out of all non-minimal TPM tissue and algorithm combinations (**Figure 12**) the whole blood and pancreas models tINIT models had the lowest mean number of essential tasks passed (12 and 11) and the fastCORE adipose and breast models had the highest (34 and 32, respectively). The whole blood grouped models passed the least essential tasks (10), including models of all algorithms, genders, and age groups, whereas the adipose, kidney and liver grouped models of female samples aged between 20 and 39 had the highest, at 42 passed tasks.

Firstly, the *Troppo* package's implementation of the fastCORE and tINIT algorithms were compared, as fastCORE appears to produce models more capable of fulfilling essential tasks when supplied with protected reactions, with fewer reactions, regardless of data aggregation method.

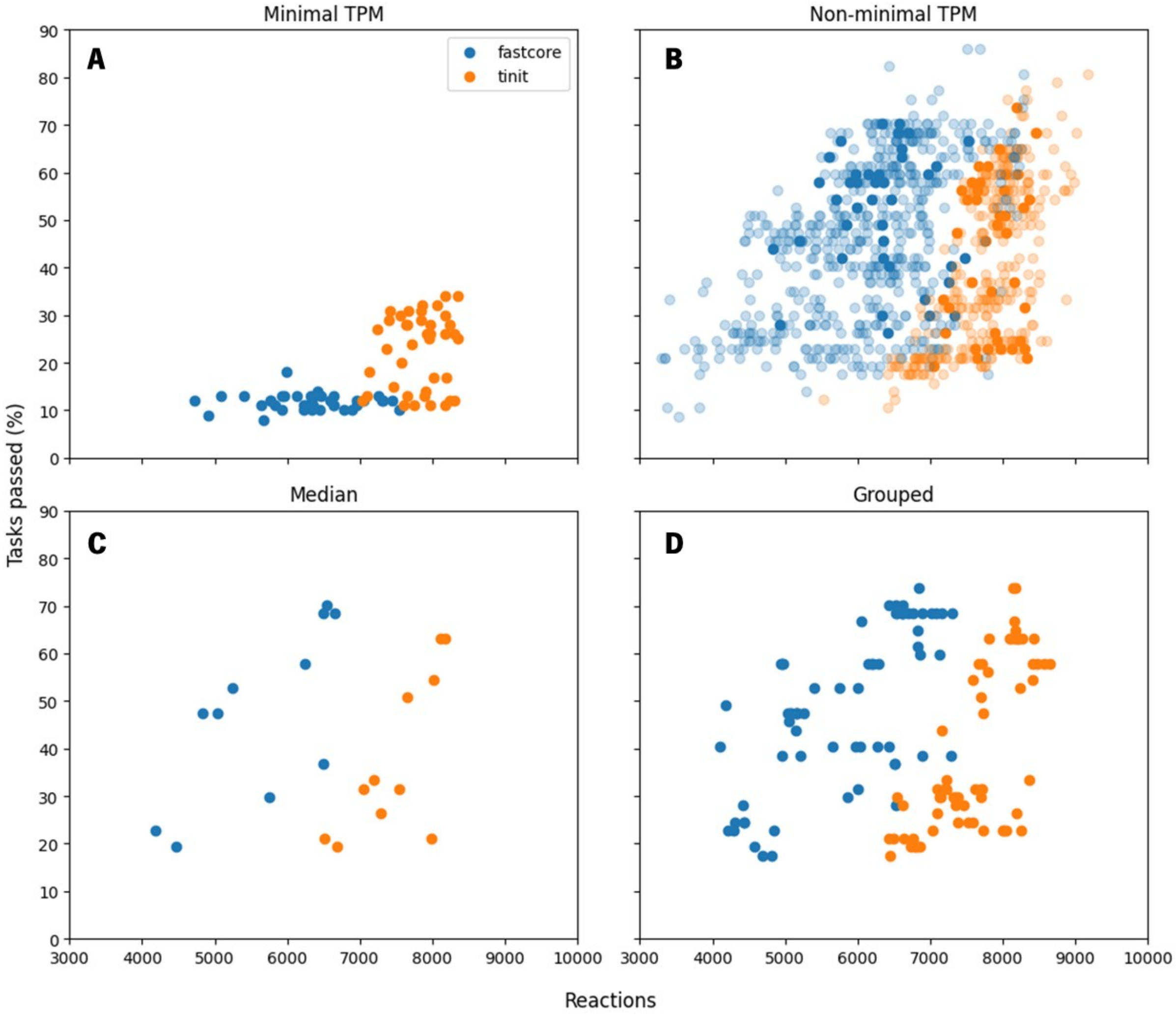


Figure 11 – Percentage of essential metabolic tasks passed (out of 57) by the minimal TPM, non-minimal TPM, grouped, and median (gene scores) CSMs extracted using the tINIT and fastCORE algorithms, compared to their respective reaction counts. (**A**) Minimal TPM models (n=80) of the adipose, breast, kidney, and liver tissues, extracted without the RAVEN protected reactions. (**B**) Non-minimal TPM models of all tissues (n=935). The points marked with full opacity correspond to the same samples used for the minimal models. (**C**) Median models of all tissues (n=22). (**D**) Grouped models of all tissues (n=132).

A Shapiro-Wilk normality test ascertained that only the median models' essential tasks (fastCORE: W≈0.93, p≈0.37; tINIT: W≈0.86, p≈0.06) were of approximately normal distribution. Consequently, a nonparametric Kruskal-Wallis H-test confirmed that the essential tasks of the 6 combinations of (non-minimal) model data aggregation method and algorithm were significantly different (H≈34.48, p≈1.91e-6). A *post hoc* Dunn's test (*scikit_posthocs* package[10]) with Benjamini/Hochberg (non-negative) multiple

[10] https://scikit-posthocs.readthedocs.io/en/latest/installation/

test p-value adjustment determined that the TPM and grouped non-minimal models differed by algorithm, whereas the median models did not. Additionally, the models did not differ solely by data aggregation method (**Table 4**).

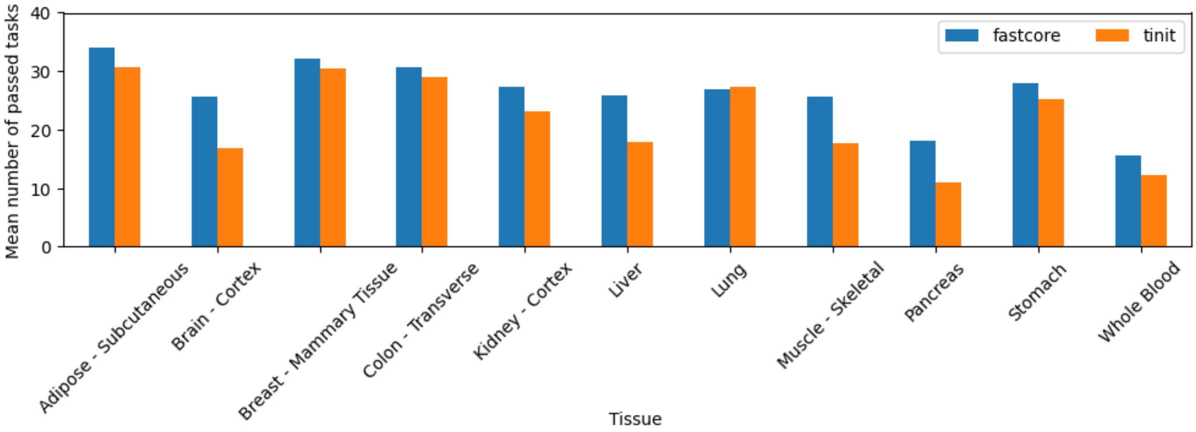


Figure 12 – Mean number of essential metabolic tasks passed by the non-minimal (gene scores) TPM models of all tissues, by tissue and extraction algorithm.

Thirdly, the effect of the protected reactions was investigated by comparing minimal and non-minimal models extracted from the same samples. A nonparametric one-tailed (“greater”) Mann-Whitney U rank test established that the non-minimal fastCORE (U=1595, p≈8.57e-15) models passed significantly more essential tasks than their minimal model counterparts, but the non-minimal tINIT models did not (U=949.5, p≈0.08).

Table 4 – Dunn’s test (*scikit_posthocs* package) with Benjamini/Hochberg (non-negative) multiple test p-value adjustment for the differences in **essential metabolic tasks passed** by the 6 data aggregation method (TPM, grouped/GRP and median/MED) and algorithm (fastCORE/FT and tINIT) combinations of non-minimal models. P-values in cells coloured green are significant ($p < 0.05$) and rounded to 2 decimal places.

	tINIT TPM	**FT MED**	**tINIT MED**	**FT GRP**	**tINIT GRP**
FT TPM	2,24E-05	8,37E-01	2,62E-01	3,09E-01	1,73E-02
tINIT TPM	–	3,09E-01	7,93E-01	1,95E-03	8,07E-01
FT MED	–	–	3,09E-01	8,07E-01	3,09E-01
tINIT MED	–	–	–	1,29E-01	8,07E-01
FT GRP	–	–	–	–	1,10E-02

In contrast to the gene scores CSMs, the H1-CSMs passed all essential tasks according to the MATLAB RAVEN evaluation, and up to 56 tasks according to *Troppo*. Furthermore, the H1-CSMs had comparable reaction counts, between 6584 (muscle) and 8164 (kidney). A (two-sided) nonparametric

Mann-Whitney U rank test determined that the H1-CSMs only had significantly more reactions than the non-minimal fastCORE TPM models (fastCORE: U=5861, p≈2.52e-6; tINIT: U=1421, p≈0.13).

Lastly, the number of essential tasks passed by the grouped and non-minimal TPM models, by age and gender, was plotted (**Figure 13**). A possible trend, where tasks passed decrease as the age group increases, is more apparent in the TPM models.

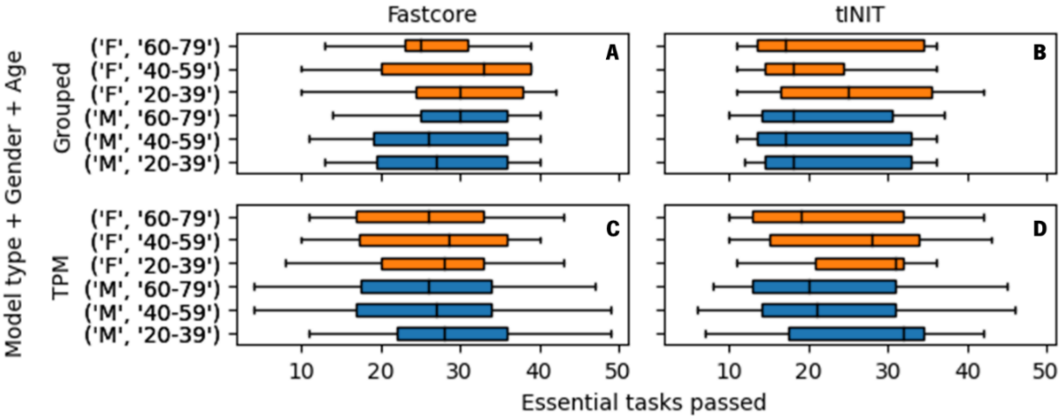


Figure 13 – Essential metabolic tasks passed by the (gene scores) TPM and grouped RNA-Seq data aggregation model types, by age, gender, and algorithm. Boxplots pertaining to female (F) and male (M) models are coloured orange and blue, respectively. Age groups denote years. (**A**) Grouped fastCORE models. (**B**) Grouped tINIT models. (**C**) Non-minimal fastCORE TPM models. (**D**) Non-minimal tINIT TPM models.

As the samples within each combined category of gender and age are independent, with a minimum of 10 samples, a nonparametric Kruskal-Wallis H-test was employed for each of the 4 model groups. Of the 4, only the tINIT TPM models' essential tasks were found to be significantly different (H≈12.7, p≈0.03). Afterwards, a Dunn's test with multiple test p-value adjustment determined that the only significantly different pair (p≈0.04) was the male tINIT models aged between 20 and 39 and the male tINIT models aged between 60 and 79.

4.4. Full set of tasks

Unlike the essential tasks, which should be attainable by all healthy human tissues, the full set of metabolic tasks can better characterize differences between model types. According to Uhlén *et al.* (2015) [65], from which this set of tasks was adapted for Human-GEM, 192 of the 256 tasks were classified as housekeeping (HK), as models of all tissues extracted by the authors could achieve them. As the *Troppo* and RAVEN task evaluations can differ greatly, the *Troppo* task evaluation of the H1-CSMs was used for comparison.

Firstly, the mean number of tasks from this set passed by the three RNA-Seq data model types, (non-minimal) TPM, grouped and median, was plotted by tissue and algorithm. The tINIT TPM pancreas and whole blood models (**Figure 14A**) had the lowest mean number of tasks passed (51 and 54, respectively), whereas the fastCORE TPM adipose models had the highest (154). The grouped (**Figure 14B**) and median (**Figure 14C**) models both had the same tissues on the lower end of the spectrum, but the median fastCORE lung model (183) outperformed the corresponding adipose tissue model (178) as best performing tissue instead. In contrast, the H1-CSMs surpass the average performance of the *Troppo* CSMs, ranging from 207 tasks passed by the H1-blood model to 219 tasks passed by the H1-liver model (according to the *Troppo* task evaluation).

As with the essential tasks, models extracted with fastCORE appear to pass more tasks overall, for all tissues apart from the skeletal muscle. A Shapiro-Wilk normality test ascertained that only the median models' tasks from the full set were of approximately normal distribution (fastCORE: W≈0.90, p≈0.17; tINIT: W≈0.89, p≈0.15). Firstly, a nonparametric Kruskal-Wallis H-test determined that the number of tasks passed by the 6 combinations of model data aggregation method and algorithm were also significantly different (H≈11.9429, p≈0.04). However, the *post hoc* Dunn's test with Benjamini/Hochberg (non-negative) multiple test p-value adjustment established that there were no actual significant differences between the groups.

Secondly, the effect of the protected reactions was also investigated in terms of number of tasks passed, specifically by comparing minimal and non-minimal TPM models extracted from the same samples. On average, the minimal fastCORE and tINIT TPM models passed 54 and 102 tasks, respectively. Surprisingly, a nonparametric one-tailed ("greater") Mann-Whitney U rank test established that neither the non-minimal fastCORE (U=1571.5, p≈0.53) nor tINIT models (U=949.5, p≈0.07) passed significantly more tasks than their minimal model counterparts. However, when the tasks passed by the minimal TPM models were compared to all non-minimal TPM models, the difference became significant for the fastCORE (U=21742.5, p≈0.03) but not for the tINIT (U=7891.5, p≈0.11) models.

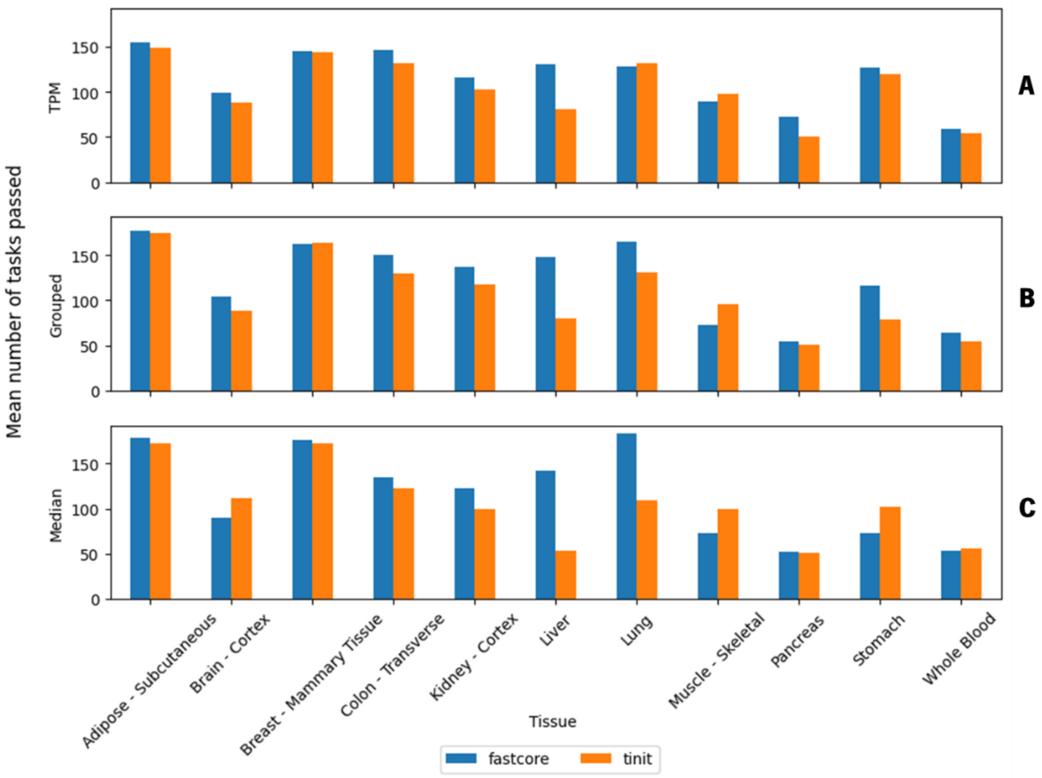


Figure 14 – Mean number of metabolic tasks passed from the full set of tasks, by all three non-minimal (gene scores) RNA-Seq data aggregation model types, TPM, grouped and median, by tissue and algorithm. (**A**) TPM models. (**B**) Grouped models. (**C**) Median models.

Thirdly, the absolute tasks passed by the grouped and non-minimal TPM models, by age and gender, was also plotted (**Figure 15**). Nonparametric Kruskal-Wallis tests were employed to discern if any significant differences in number of tasks passed exist within each of the 4 model groups. Once again, of the 4 groups tested, only the differences between the tINIT TPM models were considered significant (H≈12.97, p≈0.02). A Dunn's test with multiple test p-value adjustment determined that only the tasks passed by models of male samples differed by age group (20-39 & 40-59: p≈0.03; 20-39 & 60-79: p≈0.03; 40-59 & 60-79: p≈0.97), with an additional significant pair (p≈0.03) between the male models aged between 20 and 39 and the female models aged between 60 and 79.

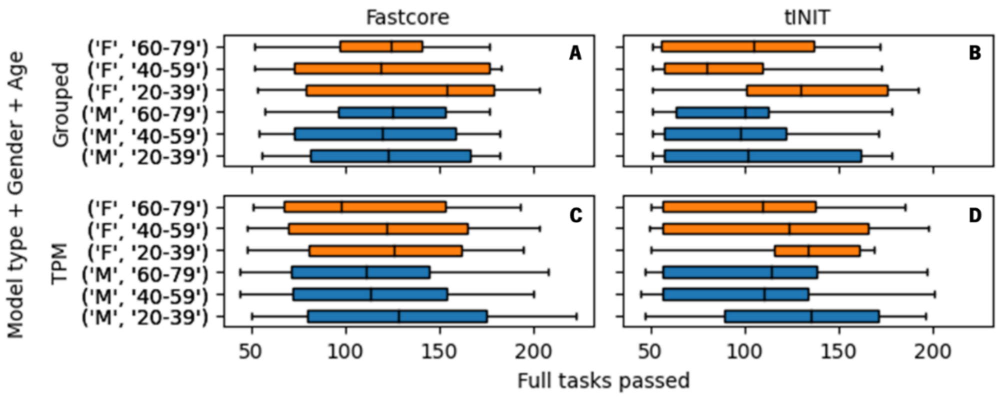


Figure 15 – Total number of metabolic tasks passed from the full set of tasks, by the (gene scores) TPM and grouped RNA-Seq data aggregation model types, by age, gender, and algorithm. Boxplots pertaining to female (F) and male (M) models are coloured orange and blue, respectively. Age groups denote years. (**A**) Grouped fastCORE models. (**B**) Grouped tINIT models. (**C**) Non-minimal fastCORE TPM models. (**D**) Non-minimal tINIT TPM models.

Subsequently, the type and frequency of tasks passed were explored. The tasks were categorized as universally passed, universally failed, or neither. HK tasks were distinguished from non-HK (other) tasks. According to **Table 5**, the median and TPM models were generally the most and least homogeneous, respectively, and non-HK tasks appear to be (universally) passed more often than HK tasks. Of all tasks considered, 28 HK and 12 non-HK tasks were universally passed, whereas 7 HK and 3 non-HK were universally failed, regardless of model type or algorithm.

Table 5 – Metabolic task evaluation of the 3 RNA-Seq data aggregation methods of (gene scores) CSMs, (non-minimal) TPM, grouped and median, by algorithm and task type (Housekeeping, HK), of the full set of tasks. The percentage of tasks passed, per task type, is in between parentheses.

		Task type					
		Universally passed		**Universally failed**		**Neither**	
Model type	**Algorithm**	HK	Other	HK	Other	HK	Other
TPM	fastCORE	28 (14.58%)	12 (18.75%)	8 (4.17%)	3 (4.69%)	156 (81.25%)	49 (76.56%)
	tINIT	30 (15.63%)	13 (20.31%)	25 (13.02%)	7 (10.94%)	137 (71.35%)	44 (68.75%)
Grouped	fastCORE	33 (17.19%)	12 (18.75%)	34 (17.71%)	9 (14.06%)	125 (65.10%)	43 (67.19%)
	tINIT	35 (18.23%)	13 (20.31%)	41 (21.35%)	9 (14.06%)	116 (60.42%)	42 (65.63%)
Median	fastCORE	35 (18.23%)	12 (18.75%)	56 (29.17%)	9 (14.06%)	101 (52.60%)	43 (67.19%)
	tINIT	36 (18.75%)	14 (21.88%)	63 (31.81%)	13 (20.31%)	93 (48.44%)	37 (57.81%)

Lastly, as the models of healthy tissue extracted by Robinson *et al.* relied upon GTEx median TPM data, it was expected that the most similar *Troppo* CSMs would be the tINIT median models. For that reason, the percentage of models that pass the 123 tasks neither passed nor failed by all H1-CSMs, tINIT median and tINIT TPM models (93 HK and 30 non-HK tasks) was plotted, by tissue. According to **Figure**

16, the median tINIT models are more similar to the TPM tINIT models than to the H1-CSMs (full size figure available in the accompanying Jupyter Notebook[2], “results_graphs_clean.ipynb”). Furthermore, the heatmap somewhat reflects the tendency presented in **Table 5**, where HK tasks appear to fail more often than non-HK tasks, which are concentrated on the right side of the heatmap.

Unsurprisingly, the H1-CSMs were noticeably more homogeneous. According to the *Troppo* task evaluation, the H1-CSMs universally passed 77.6% and 85.9% of HK and non-HK tasks, respectively, with tasks neither failed nor passed by the 11 models between 3.13% (non-HK) and 9.38% (HK). However, according to the MATLAB RAVEN task evaluation, the H1-CSMs universally pass nearly all HK tasks (189, 98.44%) and only have non-HK universally failed tasks (29, 45.31%).

Consequently, the percentage of tasks passed, by task type, of the non-minimal TPM models based on samples extracted with both algorithms was used to ascertain whether non-HK tasks truly passed in greater proportion than HK tasks. Nonparametric paired one-tailed (“greater”) Wilcoxon Signed-rank tests supported the trend in the TPM (fastCORE: W=147419.5, p≈3.04e-60; tINIT: W=61030, p≈7.52e-57), grouped (fastCORE: W=2043.5, p≈1.02e-9; tINIT: W=2211, p≈8.08e-13) and median (Both: W=66, p≈5e-4) model sets. In addition, the trend was also significant in the *Troppo* task evaluation (W=66, p≈5e-4) of the H1-CSMs.

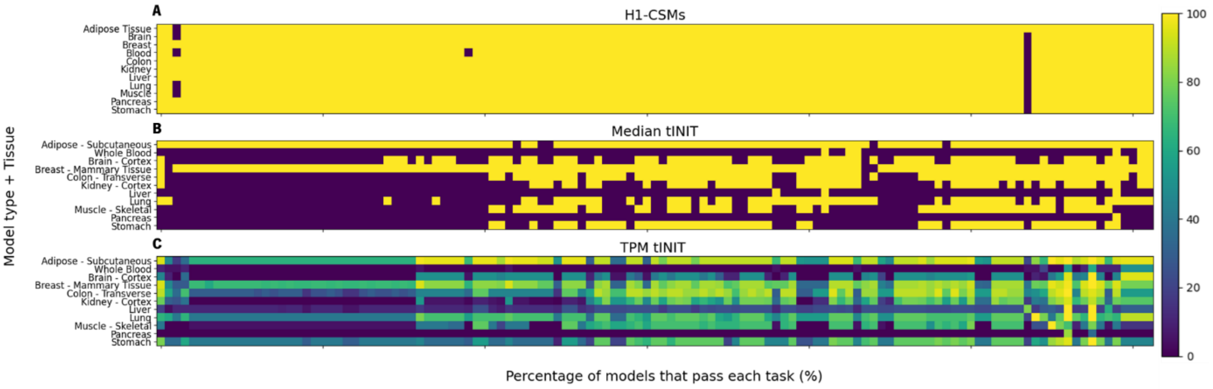


Figure 16 – Heatmap of the 123 metabolic tasks from the full set, which were not all passed or failed by the models extracted by Robinson *et al.* (H1-CSMs), the median and TPM models extracted with the tINIT algorithm, by tissue. The last 30 tasks (from the left) are classified as non-housekeeping (other). (**A**) Task evaluation of the corresponding H1-CSMs, performed using *Troppo*. (**B**) Task evaluation of the tINIT median CSMs. (**C**) Percentage of tINIT TPM models that pass a given task.

4.5. *Troppo* versus RAVEN

The null variance filter used during the threshold selection and gene score conversion process could have removed active housekeeping genes as well as unexpressed genes and affect the quality of the models. To control for this oversight, a second batch of non-minimal TPM models of the adipose, breast, kidney, and liver tissues were extracted with fastCORE (n=212), differing only in the absence of the variance filter. In other words, no genes were excluded from the gene scores dataset prior to extraction, which was otherwise obtained with the same local threshold. Unsurprisingly, the default non-minimal fastCORE TPM models had a slightly lower mean number of reactions (6246) than the corresponding revised models (6416). However, both model sets passed almost the same mean number of essential tasks (default: 29.87, revised: 29.91).

On the other hand, the CSMs extracted by Robinson *et al.* followed a distinct pre-processing method, where instead of converting the gene expression data into gene scores prior to extraction, the gene expression data was passed to the reconstruction algorithm directly, with a threshold of 1 TPM[11]. In an effort to obtain models more analogous to those produced by the authors, additional (non-minimal) TPM CSMs were extracted with the gene expression data (1 TPM threshold, without the null variance filter) and the fastCORE algorithm. Likewise, additional median models were extracted with the gene expression data directly, with both algorithms.

Figure 17 presents the percentage of essential tasks passed (out of 57) of all three sets of additional non-minimal models (gene scores without the variance filter, TPM gene expression data and median gene expression data models), alongside the H1-CSMs *Troppo* task evaluation for reference, with their corresponding reaction counts, with and without blocked reactions. As expected, the gene scores models (**Figure 17A**) overlap heavily regardless of the variance filter. In contrast, the TPM gene expression data models with all reactions (**Figure 17B**) had a higher mean number of reactions (8867) and essential tasks passed (41), while appearing more condensed.

[11] https://sysbiochalmers.github.io/Human-GEM-guide/gem_extraction/

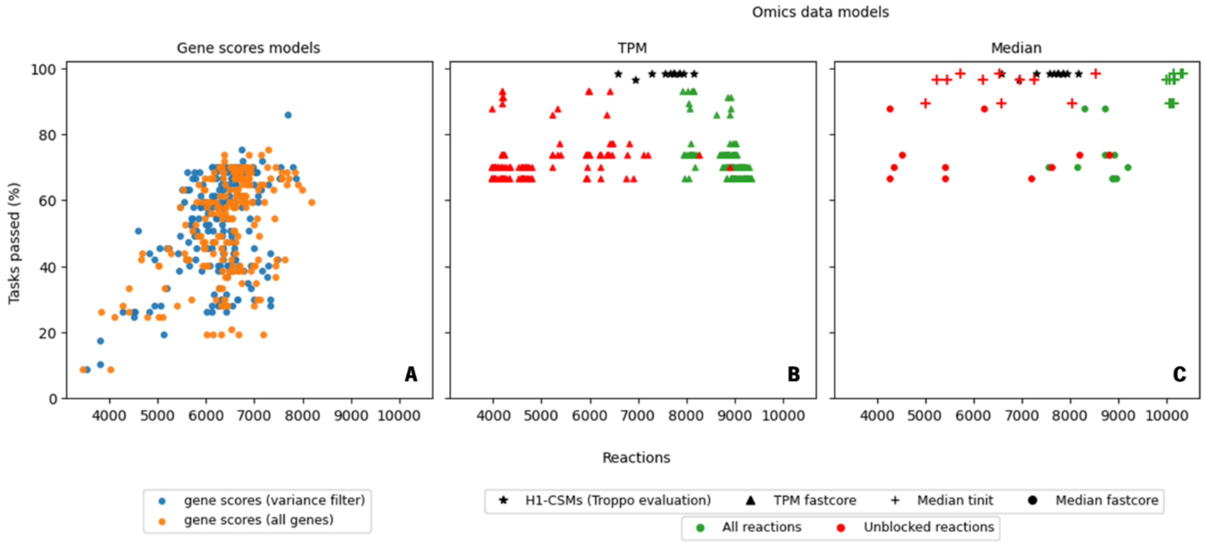


Figure 17 — Essential metabolic task evaluation of the additional non-minimal gene scores and gene expression data models and their respective reaction counts. The *Troppo* task evaluation of the models extracted by Robinson *et al.* (H1-CSMs) is shown for reference. (**A**) Gene scores fastCORE TPM models with and without the variance filter (all genes) of the same samples (n=212) belonging to the adipose, breast, liver, and kidney tissues. (**B**) Gene expression data fastCORE TPM models (all genes) corresponding to the 4 previously mentioned tissues, with and without blocked reactions. (**C**) Gene expression data median models (all genes) of all 11 tissues, with and without blocked reactions, extracted with the fastCORE and tINIT algorithms.

Particularly, according to the COBRApy package's *find_blocked_reactions* function, the gene expression data CSMs had a sizable proportion of blocked reactions, which were identified and removed. After the blocked reactions were removed, the TPM gene expression data models' reaction counts lower remarkedly (4746), with a similar spread to the gene scores models. The median gene expression data models (**Figure 17C**) appear to repeat the pattern, with the consistent reaction versions of the models having a larger spread.

However, unlike most previous comparisons between algorithms, the median fastCORE models also appear to be outperformed by tINIT. The median fastCORE and tINIT models had a mean of 42 and 54 essential tasks passed and 8640 and 10146 reactions, respectively. After blocked reaction(s) removal, the means become 6023 and 6503, with the median gene expression data tINIT models overlapping the H1-CSMs. For the gene expression TPM models, the percentage of reactions identified as blocked ranged between 0% and 57%, with a mean of 46% blocked reactions. For the median gene expression models, the percentage ranged between 0 and 46% for fastCORE and between 20 and 50% for tINIT, with a mean of 31% and 36% blocked reactions, respectively. A one-tailed Mann-Whitney test (U=52, p≈0.30) determined that the difference in blocked reaction proportion by algorithm was not truly significant in the median gene expression models.

A Kruskal-Wallis test confirmed that the three sets of *Troppo* models (gene scores, TPM gene expression data and median gene expression data models) had significantly different reaction counts (H≈661.41, p≈1.44e-139). Another test determined that the three sets of models and the H1-CSMs also differed in absolute essential tasks passed (H≈ 350.88, p≈1.13e-73). A subsequent Dunn's test with multiple test p-value adjustment established that the null variance filter did not actually affect the gene scores TPM models in terms of reaction count (**Table 6**). Furthermore, the gene scores models' reaction count differed from all other gene expression data models', except for the consistent version of the median gene expression data models. Additionally, the difference in number of reactions between all TPM and median gene expression data models, and their corresponding consistent model versions, was significant.

Table 6 – Dunn's test (*scikit_posthocs* package) with Benjamini/Hochberg (non-negative) multiple test p-value adjustment for the differences in **total number of reactions** of the 8 additional data aggregation method (TPM or median/MED), pre-processing (gene scores with variance threshold/GS, gene scores with all genes/GS-NV or gene expression data with all genes/OM), algorithm (fastCORE/FT and tINIT) and post-processing (with and without blocked reactions/NB) combinations of non-minimal models. P-values in cells coloured green are significant ($p < 0.05$) and rounded to 2 decimal places.

	FT GS-NV	**FT OM**	**FT OM-NB**	**FT MED-OM**	**tINIT MED-OM**	**FT MED-OM-NB**	**tINIT MED-OM-NB**
FT GS	1,09E-01	9,25E-51	3,10E-20	5,64E-05	1,75E-09	6,93E-01	6,93E-01
FT GS-NV	–	5,93E-40	4,43E-28	5,12E-04	4,56E-08	3,97E-01	9,13E-01
FT OM	–	–	1,01E-131	6,37E-01	1,84E-01	5,87E-07	3,70E-05
FT OM-NB	–	–	–	3,62E-12	2,18E-19	1,72E-02	1,19E-03
FT MED-OM	–	–	–	–	1,84E-01	1,32E-03	1,02E-02
tINIT MED-OM	–	–	–	–	–	3,63E-06	5,64E-05
FT MED-OM-NB	–	–	–	–	–	–	6,14E-01

Another Dunn's test with multiple test p-value adjustment between the three sets of *Troppo* models and the H1-CSMs determined that the variance filter also did not affect the gene scores models in terms of essential tasks passed (**Table 7**). In addition, all gene expression data models passed significantly more tasks than the gene scores models. The only *Troppo* gene expression data models that did not have absolute essential tasks passed comparable to the H1-CSMs were the fastCORE gene expression data models.

Table 7 – Dunn's test (*scikit_posthocs* package) with Benjamini/Hochberg (non-negative) multiple test p-value adjustment for the differences in **essential metabolic tasks passed** by the H1-CSMs and 5 additional data aggregation method (TPM or median/MED), pre-processing (gene scores with variance threshold/GS, gene scores with all genes/GS-NV or gene expression data with all genes/OM) and algorithm (fastCORE/FT and tINIT) combinations of non-minimal models. P-values in cells coloured green are significant (p < 0.05) and rounded to 2 decimal places.

	FT TPM-GS-NV	**FT TPM-OM**	**FT MED-OM**	**tINIT MED-OM**	**H1-CSMs**
FT TPM-GS	9,10E-01	2,62E-46	5,17E-06	3,56E-12	1,88E-12
FT TPM-GS-NV	–	1,09E-45	5,67E-06	4,13E-12	1,98E-12
FT TPM-OM	–	–	9,10E-01	1,57E-02	1,10E-02
FT MED-OM	–	–	–	1,04E-01	8,81E-02
tINIT MED-OM	–	–	–	–	9,10E-01

The differences in task behaviour of the full set of tasks between the three sets of models were also analysed. As with the essential tasks, the gene expression data models passed more tasks on average than the gene scores models. The default and revised gene scores models had a mean of 136.2 and 136.6, and the TPM and median fastCORE gene expression data models had a mean of 206.8 and 206.6 tasks passed, respectively. In contrast, the median tINIT gene expression data models had a mean of 245.7 tasks passed, seemingly outperforming all other model types, including the H1-CSMs (212.5).

A Kruskal-Wallis test confirmed that the three sets of models also differed in number of tasks passed from the full set (H≈466.03, p≈1.71e-98). Another Dunn's test with multiple test p-value adjustment specified which groups differed (**Table 8**). Once again, the gene scores models did not differ from each other, whereas all gene expression data models were significantly different from the gene scores models in terms of number of tasks passed, but not from each other or from the H1-CSMs.

In addition to passing more metabolic tasks, the gene expression data models were also more homogeneous than their gene scores counterparts. For example, 75% of housekeeping tasks were neither passed nor failed by all revised gene scores models, whereas the TPM and median gene expression data models extracted with fastCORE only reached 25% and 32%, respectively. However, not all the gene expression data models appear to maintain the trend of non-HK tasks being passed more often than HK tasks seen in the gene scores models. A nonparametric paired one-tailed ("greater") Wilcoxon Signed-rank test indicated the trend to be statistically significant in the fastCORE TPM and median gene expression data models (TPM: W=22578, p≈5.49e-37; Median: W=66, p≈5e-14), but not in the median tINIT gene expression data models (W=10, p≈0.98).

Table 8 – Dunn's test (*scikit_posthocs* package) with Benjamini/Hochberg (non-negative) multiple test p-value adjustment for the differences in **metabolic tasks passed from the full set of tasks** by the H1-CSMs and 5 additional data aggregation method (TPM or median/MED), pre-processing (gene scores with variance threshold/GS, gene scores with all genes/GS-NV or gene expression data with all genes/OM) and algorithm (fastCORE/FT and tINIT) combinations of non-minimal models. P-values in cells coloured green are significant ($p < 0.05$) and rounded up to 2 decimal places.

	FT TPM-GS-NV	FT TPM-OM	FT MED-OM	tINIT MED-OM	H1-CSMs
FT TPM-GS	7,78E-01	2,11E-66	6,70E-07	2,31E-13	5,98E-12
FT TPM-GS-NV	–	1,37E-64	9,34E-07	3,39E-13	9,35E-12
FT TPM-OM	–	–	7,78E-01	6,15E-02	1,56E-01
FT MED-OM	–	–	–	1,18E-01	2,05E-01
tINIT MED-OM	–	–	–	–	7,78E-01

Lastly, the number of tasks passed from the essential and full sets by the fastCORE TPM gene expression models was plotted by age and gender (**Figure 18**). Unlike the identical analyses of the gene scores variance filter models (of all 11 tissues), the TPM gene expression models are more condensed but have several outliers. Likewise, the gene expression TPM models are also affected by unbalanced sample sizes, as all metadata categories had at least 25 samples apart from the models of female patients aged between 20 and 39, with only 5 models (the minimum value according to the *scipy* documentation[12]). Subsequently, a Kruskal-Wallis test established that the TPM gene expression models vary by age and gender in terms of number of tasks passed from the full task set (W≈14.0, p≈0.02), but not from the essential set (W≈8.33, p≈0.14). Nevertheless, the follow-up Dunn's test with multiple value adjustment determined that there were no actual differences in number of tasks passed, by age and gender.

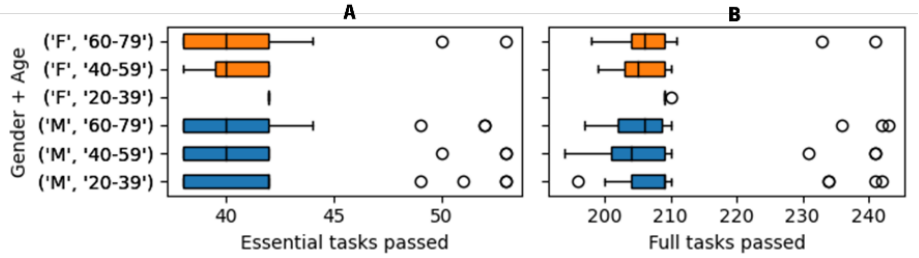


Figure 18 – Total number of metabolic tasks passed from the essential and full sets of tasks, by the TPM gene expression data models extracted with fastCORE, by age and gender. Boxplots pertaining to female (F) and male (M) models are coloured orange and blue, respectively. Age groups denote years. (**A**) Essential tasks passed by the gene expression TPM models. (**B**) Tasks passed from the full set by the gene expression TPM models.

[12] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html

5. Discussion

CSMs were extracted for 11 tissues available in the GTEx v8 dataset, specifically the adipose (subcutaneous), brain (cortex), breast (mammary tissue), colon (transverse), kidney (cortex), liver, lung, muscle (skeletal), pancreas, stomach, and whole blood. It was hypothesized that the three types of RNA-Seq data aggregation methods used, TPM, grouped and median, may vary in how well their respective models differ by tissue. For example, the median samples may have resulted in models which were very similar to each other, regardless of tissue. Likewise, the (TPM) CSMs extracted with non-minimal protected reactions were expected to perform much better than their minimal counterparts, in other words, that the protected reaction set would lead to a significant increase in number of essential tasks passed, regardless of reconstruction algorithm used. Finally, of the two reconstruction algorithms considered, fastCORE was expected to vastly outperform the tINIT algorithm in terms of computation time, but also to achieve similar performance in terms of metabolic tasks passed when compared to CSMs extracted with tINIT.

5.1. Applied machine learning

Of all the pre-processing steps evaluated by Richelle *et al.* (2019b) [42] for data integration with a metabolic model, only thresholding, or how to define a gene (or reaction) as expressed, was considered. Specifically, only the global and local (T2 state definition) approaches were assessed. Subsequently, machine learning (classification) methods were employed to automate the threshold selection process, enabling the analysis of a greater number of quantile-based global and local thresholds, 55 in total. Although the chosen classification objective, identification of the sample's tissue, renders the process dependent on a multiple tissue dataset, it can also be easily adjusted towards more complex objectives.

In spite of the application of the null variance filter, the selection pipeline (**Figure 6**) yielded high MCC values overall for ML models trained with gene scores of the adipose, breast, kidney, and liver tissues, with a local threshold visibly outperforming all others, including the ML models trained with gene expression data directly. However, the pipeline did not prove as effective for the other 7 tissues, where no threshold clearly outperformed the others, whether global or local, with extremely high MCC values (>0.99) overall. As a greater amount of data was passed as input for the 7-tissue threshold selection, it is unlikely that the high overall MCC values are merely a case of overfitting. Instead, it is more likely that the objective itself, tissue identification, may not be the most adequate. For example, a slightly more complex objective could be a combination of the sample metadata categories, such as the identification of the tissue, gender, and age group of the sample.

Post extraction, a reaction content binary matrix of CSMs enables a quantitative comparison of how the models differ structurally [4]. Accordingly, the Hamming similarity of the reaction content of the non-minimal gene scores models was plotted with a tSNE projection. The tSNE projection indicated that the overlap between tissues, seen in the gene expression data PCA prior to any data pre-processing other than filtering, persisted in the derived TPM models (**Figure 7**). Additionally, as the tSNE projection of the TPM models did not separate the two tissue groups, it also revealed previously unrecognized overlaps. Nevertheless, this type of general analysis lacks the biological context provided by other validation methods. For example, the subsystem coverage, or the number of reactions from a given reaction subsystem present in each CSM, could offer a more meaningful exploration of the CSMs structure [4].

Despite the shortcomings of the currently implemented ML pipeline, its flexibility was also utilized to train ML models with the reaction content of the non-minimal gene scores models, by algorithm. Unlike the ML pipeline employed for threshold selection, ML models were also trained without separation by tissue group (**Figure 9**). Once again, ML methods were less successful when applied to data derived from the 7-tissue dataset, which attained extremely high MCC values (up to 0.97) but showed no apparent differences between algorithms. The reaction content ML models were also inconsistent in their evaluation of algorithm performance, as the ML models trained with the reaction content of all tissues attained MCC values in between those of the separate tissue sets. Therefore, the importance of an appropriate classification objective is further emphasized.

5.2. Gene scores CSM validation

Unlike other methods of validation, such as the comparison of *in silico* metabolic model fluxes or gene essentiality predictions to experimental data, metabolic tasks are not biomass centric or as constrained by data availability. Furthermore, they should be applicable to most models of a given organism, whether the phenotypes in question involve healthy tissue or a disease state and have been used to compare between the two [4]. Therefore, the task evaluation of the CSMs was central to the appraisal of the various extraction conditions tested.

5.2.1. Essential task evaluation of the gene scores CSMs

Ideally, the generic essential tasks should be universally passed by all models, regardless of tissue. Instead, as none of the gene scores models passed all 57 essential tasks, and the models extracted under different conditions appeared to differ in that regard (**Figure 11**), the total number of metabolic tasks passed by each model was used as a basic performance indicator.

Firstly, the TPM models extracted with fastCORE appeared to pass more essential tasks than their tINIT counterparts while simultaneously having fewer reactions. In addition to confirming this trend, the Kruskal-Wallis and *post hoc* Dunn's tests also determined that the number of tasks passed was seemingly unaffected by the RNA-Seq data aggregation method in isolation.

Secondly, Mann-Whitney tests indicated that the non-minimal protected reaction set only led to a significant increase in absolute essential tasks passed for the fastCORE models. As a member of the MBA-like family of algorithms, fastCORE relies primarily on core reactions rather than a set of supplied gene scores [43]. Since the protected reactions were explicitly defined as core, this may explain why it affected the fastCORE algorithm more strongly.

However, there was a stark difference in quality between the H1-CSMs extracted by Robinson *et al.*, which passed up to 56 of the 57 essential tasks according to *Troppo,* and the *Troppo* gene scores models. As similar data (median) and template model were used for CSM extraction, and the differences in absolute essential tasks passed could not be explained solely by a difference in total number of reactions or between the *Troppo* package's and MATLAB RAVEN's task evaluation, the different pre-processing method was identified as the most likely cause.

5.2.2. Full task evaluation of the gene scores CSMs

In contrast to the generic essential tasks, the full set of metabolic tasks includes some that may be tissue-dependent and, therefore, enable a functional analysis of a CSM. Furthermore, Uhlén *et al.* (2015) [65] subdivided the tasks into 2 categories, housekeeping (HK) and non-housekeeping.

Firstly, Kruskal-Wallis and *post hoc* Dunn's tests indicated that there were few significant differences in the number of tasks passed by the gene scores models, by RNA-Seq data aggregation method or algorithm. Secondly, Mann-Whitney tests established that the non-minimal TPM models based on the same samples did not outperform their minimal counterparts. As the non-minimal protected reaction set is derived from the essential set instead of the full set of tasks, it is not entirely unexpected that their absence has a greater effect in the number of essential tasks passed. Regardless, this analysis only further emphasizes the need to explore alternate methods for CSM extraction.

Thirdly, the type and frequency of tasks was explored by calculating the percentage of tasks universally passed, universally failed or neither, by data aggregation method and algorithm. According to **Table 5**, HK tasks appear to be passed in greater proportion than non-HK tasks. Subsequently, paired Wilcoxon Signed-rank tests confirmed the trend in all non-minimal gene scores model types.

In contrast to the *Troppo* gene scores models, the H1-CSMs passed more tasks in total and were also more homogeneous. However, whereas the same trend in task type proportion is present in the *Troppo* task evaluation of the H1-CSMs, it was absent in the MATLAB RAVEN evaluation of the same models. In other words, the trend of HK tasks passing in greater proportion than the corresponding non-HK tasks appears to be an occurrence specific to the *Troppo* task evaluation.

5.3. Gene expression data CSM validation

As differences in data pre-processing were identified as the most likely cause behind the disparity in task behaviour between the *Troppo* models and H1-CSMs, more models were extracted to test alternative methods. Furthermore, the oversight in applying the null variance filter, which may have removed housekeeping genes from the gene scores dataset and, consequently, from their derived models, was addressed. Due to the algorithm's vastly better performance in terms of computation time, the models based on gene scores without the variance filter applied (all genes) were only extracted with fastCORE. Likewise, the extraction of CSMs following the distinct pre-processing method used by Robinson *et al.* was attempted. Instead of converting the RNA-Seq data into gene scores, the authors passed the gene expression data directly to their implementation of tINIT, with a global threshold of 1 TPM[11].

According to **Figure 17**, the gene expression models appeared to have many more reactions than the revised gene scores models based on the same samples. Unlike the default and revised gene scores models, which differed in the number of genes (and respective expressions) available, the revised gene scores and gene expression datasets differ solely by data pre-processing. Consequently, the apparent difference in number of reactions between the revised gene scores and gene expression models may signify that the 1 TPM global threshold is more permissive than the local gene scores threshold, as the number of genes determined to be active decreases with the global threshold value [42].

Particularly, a more in-depth analysis showed that the gene expression data models had a sizable proportion of blocked reactions. As was discussed previously, translating gene expression data into accurate CSMs is a difficult process. Despite their continued development, reconstruction algorithms are not infallible, and each implementation has its own disadvantages. The presence of blocked reactions in extracted CSMs represents another symptom of this issue, and is, therefore, not unexpected.

However, a Mann-Whitney test determined there were no significant differences in the blocked reaction proportion of the median gene expression models, by algorithm. In addition, the type of reactions removed was not explored. For example, the procedure may be removing reactions that are active under certain conditions, instead of merely inactive when the objective is biomass production. Furthermore, by including seemingly blocked reactions the algorithms may be excluding relevant reactions and, consequently, lose information during the reconstruction process. Nevertheless, as the proportion of blocked reactions was only identified for the gene expression models, of which only the median gene expression models were extracted with both algorithms, the analysis is somewhat limited.

Once again, Kruskal-Wallis and Dunn's tests were employed to investigate differences in total number of reactions and tasks passed from the essential and full tasks sets. Firstly, the tests established that the null variance filter oversight did not significantly affect the quality of the CSMs in the measurements considered, namely the total number of reactions and metabolic tasks passed. In other words, the exclusion of the filter after threshold selection did not lead to an increase in model quality. This may suggest that the null variance filter instead affected the threshold selection process itself, and that the chosen local threshold for gene scores conversion might not have been optimal. Additionally, the tests supported the hypothesis that a global threshold is more permissive, as all gene expression models except the median models' consistent versions had more reactions in total than the gene scores (local threshold) CSMs.

Secondly, the tests confirmed that all gene expression data models passed significantly more essential tasks than either type of gene scores CSM. Moreover, the median gene expression CSMs passed a comparable number of essential tasks to the H1-CSMs, regardless of algorithm, outperforming the gene expression TPM CSMs. However, none of the *Troppo* gene expression CSMs significantly differed from the H1-CSMs, or from each other, with a comparable number of tasks passed from the full set.

In addition to passing more tasks than the gene scores models, the gene expression CSMs were also more homogeneous, with much lower percentages of tasks neither passed nor failed by all models. Likewise, a Wilcoxon Signed-rank test confirmed the trend where non-HK tasks were passed more often HK tasks in the TPM and median gene expression models extracted with fastCORE, but not in the tINIT median gene expression CSMs. As the tINIT median gene expression CSMs passed the most tasks, this does not support the hypothesis that there is a task type bias in the *Troppo* task evaluation. Instead, the presence of the tendency may simply be another side-effect of the differences in task evaluation between the *Troppo* package and RAVEN.

Lastly, the possible influence of age and gender on the number of tasks passed from the essential and full tasks sets, of the TPM gene expression CSMs extracted with fastCORE, was investigated. Once again, no significant differences between any of the groups were established. As previous identical analyses did not produce any noteworthy patterns, it indicates that the age and gender of the samples do not influence the number of metabolic tasks passed by the respective models.

6. Conclusions and Future work

The main goal of this work was to develop a pipeline for the extraction and validation of CSMs of normal human tissue. Prior to the extraction step, this pipeline covers transcriptomics data pre-processing for integration through gene scores threshold selection and conversion, generic human model pre-processing and determination of protected reactions required by essential tasks. The extraction step itself relied upon two reconstruction algorithms with different approaches, namely fastCORE and tINIT. Post extraction, the task evaluation framework of the *Troppo* and MATLAB RAVEN packages was also contrasted. Finally, metabolic task evaluation and machine learning methods were employed to validate the CSMs extracted and compare them to the state-of-the-art models available.

The development of the pipeline enabled the analysis of the extraction and validation processes themselves. In particular, the type of input data aggregation method, respective pre-processing and protected reaction set used to extract metabolic models were compared. Furthermore, all statistical analyses presented are transparently available in the accompanying Jupyter Notebook ("results_graphs_clean.ipynb"). Jupyter Notebooks concerning part of the initial data pre-processing and pre-extraction pipeline ("omics_to_genescores.ipynb") and a simple tutorial of CSM extraction with the *Troppo* package ("extraction_example_guide.ipynb") are also available[2].

Alongside the development of the pipeline itself, another main conclusion of this work was that extracted CSMs passed more metabolic tasks when the reconstruction algorithms were supplied with gene expression data directly, with a global threshold of 1 TPM, rather than with the chosen local threshold (global50-90 and local50). However, as the chosen local threshold may not have been optimal, it does not necessarily signify that a global threshold may outperform a local one. Regardless, the gene expression data CSMs extracted in this work achieved comparable performance to models available in the state-of-the-art extracted by Robinson *et al.* [4] in terms of total number of metabolic tasks passed.

Likewise, the CSMs extracted directly with gene expression data were also more homogeneous. As stated previously, according to the *Troppo* evaluation, the H1-CSMs of the 11 different tissues considered differed in only 12 tasks between the models that passed the least and most tasks from the full set, respectively. According to the RAVEN evaluation, the difference is of 13 tasks, between 209 and 222 tasks passed in total. Subsequently, the increased homogeneity cannot be explained by differences in the task evaluations of the two packages. Instead, it is more likely to be a side-effect of the use of a global threshold, as global thresholding has been observed to mark a higher number of genes as active, in all tissues, than local approaches [42].

The results presented in this work were made possible by focusing on general model attributes, such as the number of metabolic tasks passed, rather than more in-depth analyses. Naturally, this approach had its own drawbacks. For example, the total number of tasks passed did not differ by the CSMs' sample metadata, specifically by age and gender. Although the analyses were in part hindered by small or unbalanced sample sizes, a more specific approach would most likely be better at detecting those differences in model behaviour.

In summary, the pipeline presented in this work achieved its main objective, enabling extraction and basic CSM validation applicable to models of healthy human tissue. Nonetheless, the current pipeline could also be expanded upon, and several unexplored questions remain.

Firstly, the threshold selection pipeline's effectiveness could be more appropriately tested without the null variance filter oversight, enabling the confirmation of whether the gene expression data 1 TPM global threshold can truly outperform local threshold approaches based on gene scores. Secondly, an alternative objective for machine learning analyses, possibly applicable to single-tissue datasets, could be developed. The flexibility of machine learning methods could also be further applied post extraction. For example, the reaction content of extracted CSMs could be supplied as input data with a more complex metadata-based classification objective, such as age and gender. Thirdly, the pipeline could be further improved by the implementation of an equivalent function to RAVEN's *checkTasks* [63] in *Troppo*, to determine reactions required by essential tasks, as at present the pipeline still partly depends on MATLAB, a commercial platform.

Although fastCORE vastly outperformed the *Troppo* implementation of tINIT in the variance filter gene scores models, the best models obtained were extracted with tINIT, specifically with median gene expression data and a global threshold (of 1 TPM). As such, the two algorithms could be compared further. Likewise, no minimal models were extracted with gene expression data as input and the effect of the protected reactions alongside the alternative pre-processing method was not explored. Moreover, the specific source behind the differences in task evaluation between the *Troppo* package and RAVEN was not investigated outside of the context of the comparisons between CSMs.

Lastly, by narrowing the scope to a single tissue, CSMs could more easily be validated on an individual task level. In other words, effort could be devoted towards the automation of a typically tedious, manual CSM curation process. A good candidate tissue would be the liver, which has been extensively modelled in the past. Particularly, a manually curated liver model, HepatoNet1, is available for comparison [31].

References

1. Angione C. (2019). Human Systems Biology and Metabolic Modelling: A Review-From Disease Metabolism to Precision Medicine. BioMed research international, 2019, 8304260. https://doi.org/10.1155/2019/8304260
2. Richelle, A., Chiang, A. W., Kuo, C. C., & Lewis, N. E. (2019). Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions. PLoS computational biology, 15(4), e1006867. https://doi.org/10.1371/journal.pcbi.1006867
3. Ryu, J. Y., Kim, H. U., & Lee, S. Y. (2015). Reconstruction of genome-scale human metabolic models using omics data. Integrative Biology, 7(8), 859-868. https://doi.org/10.1039/c5ib00002e
4. Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P. E., Cook, D., Nilsson, A., Anton, M., Ferreira, R., Domenzain, I., Billa, V., Limeta, A., Hedin, A., Gustafsson, J., Kerkhoven, E. J., Svensson, L. T., Palsson, B. O., Mardinoglu, A., Hansson, L., Uhlén, M., & Nielsen, J. (2020). An atlas of human metabolism. Science signaling, 13(624), eaaz1482. https://doi.org/10.1126/scisignal.aaz1482
5. Zhang, C., & Hua, Q. (2016). Applications of genome-scale metabolic models in biotechnology and systems medicine. Frontiers in physiology, 6, 413. https://doi.org/10.3389/fphys.2015.00413
6. Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., & Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. Molecular systems biology, 7(1), 501. https://doi.org/10.1038/msb.2011.35
7. Shlomi, T., Benyamini, T., Gottlieb, E., Sharan, R., & Ruppin, E. (2011). Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. PLoS Comput Biol, 7(3), e1002018. https://doi.org/10.1371/journal.pcbi.1002018
8. Ferreira, J., Vieira, V., Gomes, J., Correia, S., & Rocha, M.: Troppo-A Python Framework for the Reconstruction of Context-Specific Metabolic Models. In International Conference on Practical Applications of Computational Biology & Bioinformatics, pp. 146-153. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23873-5_18
9. Bordbar, A., Monk, J. M., King, Z. A., & Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. Nature Reviews Genetics, 15(2), 107-120. https://doi.org/10.1038/nrg3643
10. Kitano, H. (2002). Systems biology: a brief overview. science, 295(5560), 1662-1664. https://doi.org/ 10.1126/science.1069492
11. Machado, D., & Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol, 10(4), e1003580. https://doi.org/10.1371/journal.pcbi.1003580
12. Tummler, K., & Klipp, E. (2018). The discrepancy between data for and expectations on metabolic models: How to match experiments and computational efforts to arrive at quantitative predictions?. Current Opinion in Systems Biology, 8, 1-6. https://doi.org/10.1016/j.coisb.2017.11.003
13. Cook, D. J., & Nielsen, J. (2017). Genome-scale metabolic models applied to human health and disease. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 9(6), e1393. https://doi.org/10.1002/wsbm.1393
14. Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols, 5(1), 93. https://doi.org/10.1038/nprot.2009.203
15. Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., ... & Prlić, A. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. Nature biotechnology, 36(3), 272. https://doi.org/10.1038/nbt.4072

16. Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., ... & Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences, 104(6), 1777-1782. https://doi.org/10.1073/pnas.0610772104
17. Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., & Goryanin, I. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. Molecular systems biology, 3(1), 135.https://doi.org/10.1038/msb4100177
18. Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., ... & Thorleifsson, S. G. (2013). A community-driven global reconstruction of human metabolism. Nature biotechnology, 31(5), 419-425. https://doi.org/10.1038/nbt.2488
19. Smallbone, K. (2013). Striking a balance with Recon 2.1. arXiv:1311.5696. https://arxiv.org/abs/1311.5696
20. Swainston, N., Smallbone, K., Hefzi, H., Dobson, P. D., Brewer, J., Hanscho, M., ... & Mendes, P. (2016). Recon 2.2: from reconstruction to model of human metabolism. Metabolomics, 12(7), 1-7. https://doi.org/10.1007/s11306-016-1051-4
21. Mardinoglu, A., & Nielsen, J. (2012). Systems medicine and metabolic modelling. Journal of internal medicine, 271(2), 142-154. https://doi.org/10.1111/j.1365-2796.2011.02493.x
22. Mardinoglu, A., Gatto, F., & Nielsen, J. (2013). Genome-scale modeling of human metabolism–a systems biology approach. Biotechnology journal, 8(9), 985-996. https://doi.org/10.1002/biot.201200275
23. Blais, E. M., Rawls, K. D., Dougherty, B. V., Li, Z. I., Kolling, G. L., Ye, P., ... & Papin, J. A. (2017). Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. Nature communications, 8(1), 1-15. https://doi.org/10.1038/ncomms14250
24. Thiele, I., Sahoo, S., Heinken, A., Heirendt, L., Aurich, M. K., Noronha, A., & Fleming, R. M. (2018). When metabolism meets physiology: Harvey and Harvetta. BioRxiv, 255885. https://doi.org/10.1101/255885
25. Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis?. Nature biotechnology, 28(3), 245-248. https://doi.org/10.1038/nbt.1614.
26. Schilling, C. H., Schuster, S., Palsson, B. O., & Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnology progress, 15(3), 296-303. https://doi.org/10.1021/bp990048k
27. Klamt, S., & Gilles, E. D. (2004). Minimal cut sets in biochemical reaction networks. Bioinformatics, 20(2), 226-234. https://doi.org/10.1093/bioinformatics/btg395
28. Alberts, B., Bray, D., Hopkin, K., Johnson, A. D., Lewis, J., Raff, M., ... & Walter, P. (2013). Essential cell biology. Garland Science.
29. Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., ... & Weitz, K. K. (2010). Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular systems biology, 6(1), 390. https://doi.org/10.1038/msb.2010.47
30. Cyrielle, C., Joshi, C., Lewis, N. E., Laetitia, M., & Andersen, M. R. (2019). Adaption of Generic Metabolic Models to Specific Cell Lines for Improved Modeling of Biopharmaceutical Production and Prediction of Processes. Cell Culture Engineering: Recombinant Protein Production, 127-162. https://doi.org/10.1002/9783527811410.ch6
31. Gille, C., Bölling, C., Hoppe, A., Bulik, S., Hoffmann, S., Hübner, K., ... & Weidlich, M. (2010). HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Molecular systems biology, 6(1), 411. https://doi.org/10.1038/msb.2010.62

32. Gudmundsson, S., & Thiele, I. (2010). Computationally efficient flux variability analysis. BMC bioinformatics, 11(1), 489. https://doi.org/10.1038/msb.2010.47
33. Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D. C., & Lewis, N. E. (2017). A systematic evaluation of methods for tailoring genome-scale metabolic models. Cell systems, 4(3), 318-329. https://doi.org/10.1016/j.cels.2017.01.010
34. Rai, A., & Saito, K. (2016). Omics data input for metabolic modeling. Current opinion in biotechnology, 37, 127-134. https://doi.org/10.1016/j.copbio.2015.10.010
35. Govindarajan, R., Duraiyan, J., Kaliyappan, K., & Palanisamy, M. (2012). Microarray and its applications. Journal of pharmacy & bioallied sciences, 4(Suppl 2), S310. https://dx.doi.org/10.4103%2F0975-7406.100283
36. Chen, L., Sun, F., Yang, X., Jin, Y., Shi, M., Wang, L., ... & Wang, Q. (2017). Correlation between RNA-Seq and microarrays results using TCGA data. Gene, 628, 200-204. https://doi.org/10.1016/j.gene.2017.07.056
37. Li, X., Brock, G. N., Rouchka, E. C., Cooper, N. G., Wu, D., O'Toole, T. E., ... & Rai, S. N. (2017). A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data. PloS one, 12(5), e0176185. https://doi.org/10.1371/journal.pone.0176185
38. NCBI Resource Coordinators (2018). Database resources of the National Center for Biotechnology Information. Nucleic acids research, 46(D1), D8–D13. https://doi.org/10.1093/nar/gkx1095
39. GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318-1330. https://doi.org/10.1126/science.aaz1776
40. Hutter, C., & Zenklusen, J. C. (2018). The cancer genome atlas: creating lasting value beyond its data. Cell, 173(2), 283-285. https://doi.org/10.1016/j.cell.2018.03.042
41. Johnson, K. A., & Krishnan, A. (2020). Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. bioRxiv. https://doi.org/10.1101/2020.09.22.308577
42. Richelle, A., Joshi, C., & Lewis, N. E. (2019). Assessing key decisions for transcriptomic data integration in biochemical networks. PLoS computational biology, 15(7), e1007185. https://doi.org/10.1371/journal.pcbi.1007185
43. Robaina Estévez, S., & Nikoloski, Z. (2014). Generalized framework for context-specific metabolic model extraction methods. Frontiers in plant science, 5, 491. https://doi.org/10.3389/fpls.2014.00491
44. Becker, S. A., & Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol, 4(5), e1000082. https://doi.org/10.1371/journal.pcbi.1000082
45. Bordbar, A., & Palsson, B. O. (2012). Using the reconstructed genome-scale human metabolic network to study physiology and pathology. Journal of internal medicine, 271(2), 131-141. https://doi.org/10.1111/j.1365-2796.2011.02494.x
46. Schmidt, B. J., Ebrahim, A., Metz, T. O., Adkins, J. N., Palsson, B. Ø., & Hyduke, D. R. (2013). GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics, 29(22), 2900-2908. https://doi.org/10.1093/bioinformatics/btt493
47. Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., & Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput Biol, 8(5), e1002518. https://doi.org/10.1371/journal.pcbi.1002518
48. Zur, H., Ruppin, E., & Shlomi, T. (2010). iMAT: an integrative metabolic analysis tool. Bioinformatics, 26(24), 3140-3142. https://doi.org/10.1093/bioinformatics/btq602

49. Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., & Nielsen, J. (2014). Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Molecular systems biology, 10(3), 721. https://doi.org/10.1002/msb.145122
50. Wang, Y., Eddy, J. A., & Price, N. D. (2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC systems biology, 6(1), 153. https://doi.org/10.1186/1752-0509-6-153
51. Jerby, L., Shlomi, T., & Ruppin, E. (2010). Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Molecular systems biology, 6(1), 401. https://doi.org/10.1038/msb.2010.56
52. Vlassis, N., Pacheco, M. P., & Sauter, T. (2014). Fast reconstruction of compact context-specific metabolic network models. PLoS Comput Biol, 10(1), e1003424. https://doi.org/10.1371/journal.pcbi.1003424
53. Pacheco, M. P., John, E., Kaoma, T., Heinäniemi, M., Nicot, N., Vallar, L., ... & Sauter, T. (2015). Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network. BMC genomics, 16(1), 809. https://doi.org/10.1186/s12864-015-1984-4
54. Schultz, A., & Qutub, A. A. (2016). Reconstruction of tissue-specific metabolic networks using CORDA. PLoS computational biology, 12(3), e1004808. https://doi.org/10.1371/journal.pcbi.1004808
55. Martín, H. G., Kumar, V. S., Weaver, D., Ghosh, A., Chubukov, V., Mukhopadhyay, A., ... & Keasling, J. D. (2015). A method to constrain genome-scale models with 13 C labeling data. PLoS Comput Biol, 11(9), e1004363. https://doi.org/10.1371/journal.pcbi.1004363
56. Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A. L., ... & Mootha, V. K. (2012). Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. Science, 336(6084), 1040-1044. https://doi.org/10.1126/science.1218595
57. Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., & Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature communications, 5(1), 1-11. https://doi.org/10.1038/ncomms4083
58. Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Nookaew, I., Jacobson, P., ... & Nielsen, J. (2013). Integration of clinical data with a genome-scale metabolic model of the human adipocyte. Molecular systems biology, 9(1), 649. https://doi.org/10.1038/msb.2013.5
59. Mardinoglu, A., Kampf, C., Asplund, A., Fagerberg, L., Hallstrom, B. M., Edlund, K., ... & Nielsen, J. (2014). Defining the human adipose tissue proteome to reveal metabolic alterations in obesity. Journal of proteome research, 13(11), 5106-5119. https://doi.org/10.1021/pr500586e
60. Zhang, A. D., Dai, S. X., & Huang, J. F. (2013). Reconstruction and analysis of human kidney-specific metabolic network based on omics data. BioMed research international, 2013. https://doi.org/10.1155/2013/187509
61. Sohrabi-Jahromi, S., Marashi, S. A., & Kalantari, S. (2016). A kidney-specific genome-scale metabolic network model for analyzing focal segmental glomerulosclerosis. Mammalian Genome, 27(3), 158-167. https://doi.org/10.1007/s00335-016-9622-2
62. Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: constraints-based reconstruction and analysis for python. BMC systems biology, 7(1), 1-6. https://doi.org/10.1186/1752-0509-7-74
63. Wang, H., Marcišauskas, S., Sánchez, B. J., Domenzain, I., Hermansson, D., Agren, R., ... & Kerkhoven, E. J. (2018). RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLoS computational biology, 14(10), e1006541. https://doi.org/10.1371/journal.pcbi.1006541

64. Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., ... & Fleming, R. M. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. Nature protocols, 14(3), 639-702. https://doi.org/10.1038/s41596-018-0098-2
65. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., ... & Pontén, F. (2015). Tissue-based map of the human proteome. Science, 347(6220). https://science.sciencemag.org/lookup/doi/10.1126/science.1260419