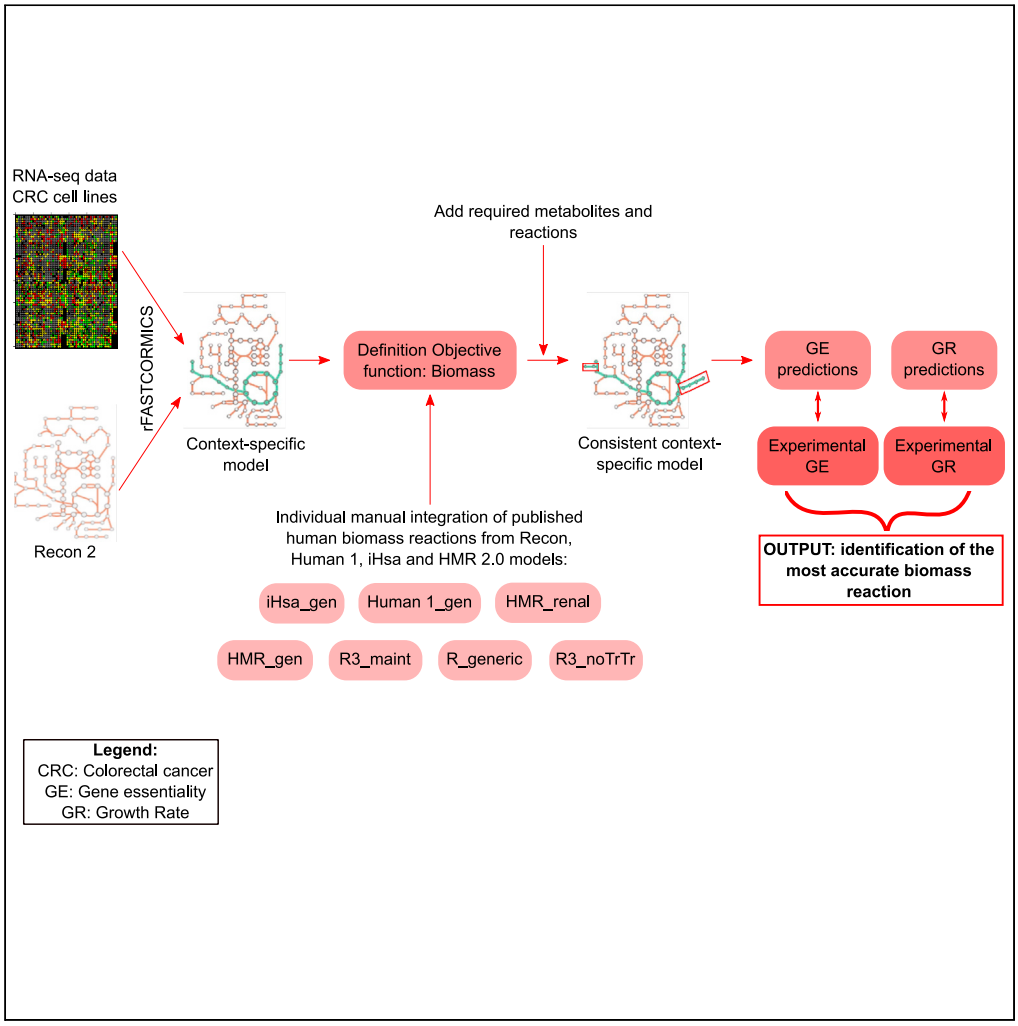
iScience

CellPress
OPEN ACCESS

Article

# Importance of the biomass formulation for cancer metabolic modeling and drug prediction




María Moscardó García, Maria Pacheco, Tamara Bintener, Luana Presta, Thomas Sauter

thomas.sauter@uni.lu

## Highlights

The definition of the biomass reaction is of utmost importance for model predictions

Growth rate predictions are affected by metabolite composition and their coefficients

Gene essentiality predictions are mainly affected by the metabolite composition

Need to find a standard biomass reaction for reproducibility and consistency purposes

Moscardó García et al., iScience 24, 103110
October 22, 2021 
https://doi.org/10.1016/j.isci.2021.103110

iScience

CellPress
OPEN ACCESS

Article

# Importance of the biomass formulation for cancer metabolic modeling and drug prediction

María Moscardó García,[1] Maria Pacheco,[1] Tamara Bintener,[1] Luana Presta,[1] and Thomas Sauter[1,2,*]

## SUMMARY

**Genome-scale metabolic reconstructions include all known biochemical reactions occurring in a cell. A typical application is the prediction of potential drug targets for cancer treatment. The precision of these predictions relies on the definition of the objective function. Generally, the biomass reaction is used to illustrate the growth capacity of a cancer cell. Today, seven human biomass reactions can be identified in published metabolic models. The impact of these differences on the metabolic model predictions has not been explored in detail. We explored this impact on cancer metabolic model predictions and showed that the metabolite composition and the associated coefficients had a large impact on the growth rate prediction accuracy, whereas gene essentiality predictions were mainly affected by the metabolite composition. Our results demonstrate the importance of defining a consensus biomass reaction compatible with most human models, which would contribute to ensuring the reproducibility and consistency of the results.**

## INTRODUCTION

The development of novel cancer drugs is costly and time-consuming. It takes on average 9 years and several hundred million dollars from compound discovery to clinical trials. Unfortunately, less than 10% of the drugs succeed (Ashburn and Thor, 2004). Most of them fail clinical trial phases II and III (Fogel, 2018), owing to differences between *in vitro* (animal and cell line models) and *in vivo* models or because of severe adverse effects for a subset of the population. Hence, drug repurposing, a more cost, and time-efficient approach has become popular in the last decade. It aims to find new indications for already approved drugs. With known pharmacokinetic profiles, toxicity, and side effects, the financial risks are highly reduced. Indeed, there are several examples of successfully repurposed drugs including Sildenafil, originally developed for the treatment of angina and further repurposed for erectile dysfunction (DeBusk et al., 2004).

Among others, high throughput *in vitro* screens such as drug, CRISPR-Cas9, and siRNA screens are used to identify drug candidates for repurposing. Although drug screens are relatively easy to perform, depending on the number of drugs and cell lines, they can be time-consuming as they require the determination of the IC-50 values for each drug and each cell line used. Further, essential genes determined by gene screenings, which can be used as a surrogate for drugs, vary across cell lines of the same cancer type, hence the efficiency of a drug is likely cell line-type dependent and the choice of the cell line is critical for the identification of drug targets. Furthermore, screens often fail to identify drug combinations promptly due to the high number of potential combinations. Consequently, computational approaches such as context-specific metabolic models reconstructed from patients or cell line RNA-seq data can be used to perform *in silico* knockouts to simulate the effects of drugs and their combinations across different cell lines, tissues, and patients in a fraction of time and money. Metabolic modeling allows simulating drugs on more tissues and cell lines, and thus to capture the heterogeneity of drug response or toxic effects that might affect only a subgroup of patients. Although simulations do not replace *in vitro* testing, they allow filtering out inefficient or toxic drugs for a certain subgroup. Hence, only the most promising candidates would be tested *in vitro,* and *in vivo* (Pacheco et al., 2019).

Previous studies have used gene expression data from The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE) to build metabolic models to predict drugs for lung, prostate, and colon cancer (Pacheco et al., 2019). Despite the success of these metabolic modeling studies, several problems

[1]Department of Life Sciences and Medicine, University of Luxembourg, 4367 Esch-sur-Alzette, Luxembourg

[2]Lead contact

*Correspondence: thomas.sauter@uni.lu

https://doi.org/10.1016/j.isci.2021.103110

need to be solved to apply *in silico* drug screenings for personalized medicine. These problems include the quality of the input reconstruction used as a scaffold for the building of the context-specific models and data integration, the accuracy of the Gene protein rules (GPR), and the algorithms themselves, as intensively discussed in (Pfau et al., 2016). Since then, efforts have been made to standardize the reconstruction such as Merlin (Dias et al., 2015), ModelSEED (Henry et al., 2010), and MetExplore (Cottret et al., 2018).

Nevertheless, one main issue, which directly impacts the quality of the context-specific models and thus *in silico* screenings, remains: the formulation of the objective function. The objective function represents the metabolic goal of the organism (Montezano et al., 2015). Though growth, reproduction, and maintenance may be the overall goal of a multicellular organism, it is not the aim of each of its cells e.g., neurons do not proliferate. Thus, other objective functions have been defined to describe the metabolism of non-proliferative cells such as the ATP production (Knorr et al., 2007), glucose transport (Maldonado et al., 2018), and combinations of multiple objective functions (Vo et al., 2004). There have also been formulations of objective functions that are tailored to the metabolism of entire organs such as the secretion and absorption of different metabolites for the human kidney (Chang et al., 2010) and very-low-density lipoprotein synthesis, bile formation, and heme biosynthesis for the liver (Gille et al., 2010). Regardless of the objective, every cell has to fulfill a multitude of functions (Schuetz et al., 2012); therefore, it is unlikely that a single objective reaction can capture every combination of functions. Nevertheless, in cancer for example, because of the high proliferation rate, biomass production is mainly used as the objective function (Wagner et al., 2013), whereas ATP maintenance is used for some non-proliferative cells (Schuetz et al., 2012).

Beyond the question of what function to use for which cell type or organism, there is no agreement on the formulation of the biomass reaction. In theory, the biomass reaction should encompass every cellular component required to sustain cellular proliferation, including the major macromolecules (DNA, RNA, proteins, and lipids), key coenzymes, inorganic ions, growth and non-growth associated maintenance costs, and in some cases, species-specific metabolites. However, several variants with slightly different formulations have been published for human cells (Table 1). Most currently used biomass reactions stem from standard metabolites related to the biomass generation, which has been indistinctly applied to *Escherichia coli*, *Saccharomyces cerevisiae*, and mammals (Feist and Palsson, 2010); thus, they might not be representative of human cells. The inconsistencies and the impact of the objective function on model prediction accuracy have led to the development of new algorithms to generate a data-driven biomass function (Lachance et al., 2019) and to standardize (Chan et al., 2017) condition and species-specific biomass objective functions.

BOFdat (Lachance et al., 2019) is one of the latest tools able to generate species-specific and condition-specific biomass reactions based on experimental data. BOFdat is a Python software package, which divides the biomass definition process into three steps:

- Step 1: calculation of the stoichiometric coefficients of the major cellular macromolecules (DNA, RNA, proteins, and lipids) using experimentally obtained omics data.
- Step 2: the addition of coenzymes and inorganic ions to the biomass objective function, based on available knowledge.
- Step 3: identification of the condition and species-specific biomass precursors, by using organism-specific data.

As the last step has been designed to optimize the phenotypic prediction and to improve the essential gene prediction, BOFdat could have applications in drug repurposing for cancer by generating a data-driven human cancer-specific objective function.

In this study, we tested seven different published biomass reaction formulas as well as a BOFdat-generated biomass reaction using Recon 2 or the corresponding home model as input model to build context-specific models based on transcriptomics data from colorectal cancer (CRC) cell lines. In addition, home models i.e., Recon 3, Human 1, HMR and iHsa were considered for validation purposes. *In silico* essential gene and growth rate predictions were performed to determine if its coefficients and metabolites have an impact on the prediction accuracy and to identify the best performing biomass formulation.

**Table 1. List of biomass formulations that were benchmarked in the present study**

| GEM | Biomass abbreviation | References | Summary of the biomass reaction |
|---|---|---|---|
| Recon family (1, 2, and 3) | R_generic | (Brunk et al., 2018; Duarte et al., 2007; Thiele et al., 2013) | dNTPs + NTPs + amino acids + lipid precursors + carbohydrates ➡ H+ + ADP + Biomass |
| HMR 2.0 (biomass components) | HMR_gen | (Mardinoglu et al., 2014) | dNTPs + NTPs + amino acids + lipid precursors + cofactors and vitamins + glycogen ➡ H+ + ADP + Biomass |
| Human 1 | Human1_gen | (Robinson et al., 2020) | dNTPs + NTPs + amino acids + lipid precursors + glycogen storage pool + metabolite pool + protein pool + cofactor pool ➡ H+ + ADP + Biomass |
| iHsa | iHsa_gen | (Blais et al., 2017) | dNTPs + NTPs + amino acids + proteins + bile acid + glycogen storage pool + lipid precursors + misc metabolites ➡ H+ + ADP + Biomass |
| Recon 3 (maintenance biomass)[a] | R3_main | (Brunk et al., 2018) | NTPs + amino acids + lipid precursor s + carbohydrates ➡ H+ + ADP + Biomass |
| Recon 3 (noTrTr biomass)[b] | R3_noTrTr | (Brunk et al., 2018) | NTPs + lipid precursors + carbohydrates ➡ H+ + ADP + Biomass |
| HMR 2.0 (cancer renal biomass) | HMR_renal | (Mardinoglu et al., 2014) | dNTPs + NTPs + amino acids + lipid precursors + carbohydrates + vitamin pool + glycerides + cholesterol esters ➡ H+ + ADP + Biomass |
| BOFdat | BOFdat_Biomass | (Lachance et al., 2019) | dNTPs + NTPs + amino acids + lipid precursors + coenzymes and inorganic ions + specific metabolites ➡ H+ + ADP + Biomass |

[a]Recon 3 maintenance biomass represents the Recon generic biomass reaction without replication precursors.
[b]Recon 3 noTrTr biomass reaction represents the Recon generic biomass reaction without replication, transcription, and translation.

## RESULTS

The biomass reaction should include every metabolite required to sustain cell viability and proliferation such as the major macromolecules (DNA, RNA, proteins, and lipids), key coenzymes, inorganic ions, growth and non-growth associated maintenance costs, and species-specific metabolites for some. However, the lack of a global agreement on the methodology to define the biomass reaction has led to the development of GEMs with different biomass reaction compositions (Table 1). Thus, to benchmark the current human biomass objective functions and their impact on the metabolic model predictions, the biomass reactions were tested on CRC cell lines. CRC was selected due to the fact that it is one of the leading forms of cancer deaths across most high and middle-income countries and its trend of incidence has remained unchanged over the last 40 years (Global Burden of Disease Cancer Collaboration et al., 2018). Nevertheless, these analyses could be extended to any cancer type. In this study, sample-specific models were reconstructed via the rFASTCORMICS workflow using Recon 2 or the corresponding home model as input model and integrating CRC transcriptomic data. In addition, the model was constrained according to the media composition used in the experimental CRISPR (DepMap Achilles 19Q1, 2019) and growth rate analysis (O'Connor et al., 1997), respectively, and the biomass reaction was set as an objective function. These context-specific models were further assessed in gene essentiality (18 models considered) and growth rate analysis (5 models considered), to determine the impact of the biomass formulation.

### Different biomass reactions predict different essential genes

Although most key metabolites are shared between the biomasses, significant differences were observed between their compositions. To test the predictive power of the various formulations, the biomass reactions were individually integrated into Recon 2. Further, *in silico* knockouts with each biomass reaction as the objective function were performed. Later, precision, sensitivity, and specificity analyses were conducted using CRISPR data (DepMap Achilles 19Q1, 2019) to compare the predicted essential genes with experimentally identified essential genes. These results revealed that the use of different biomass reactions leads to different essential gene predictions.

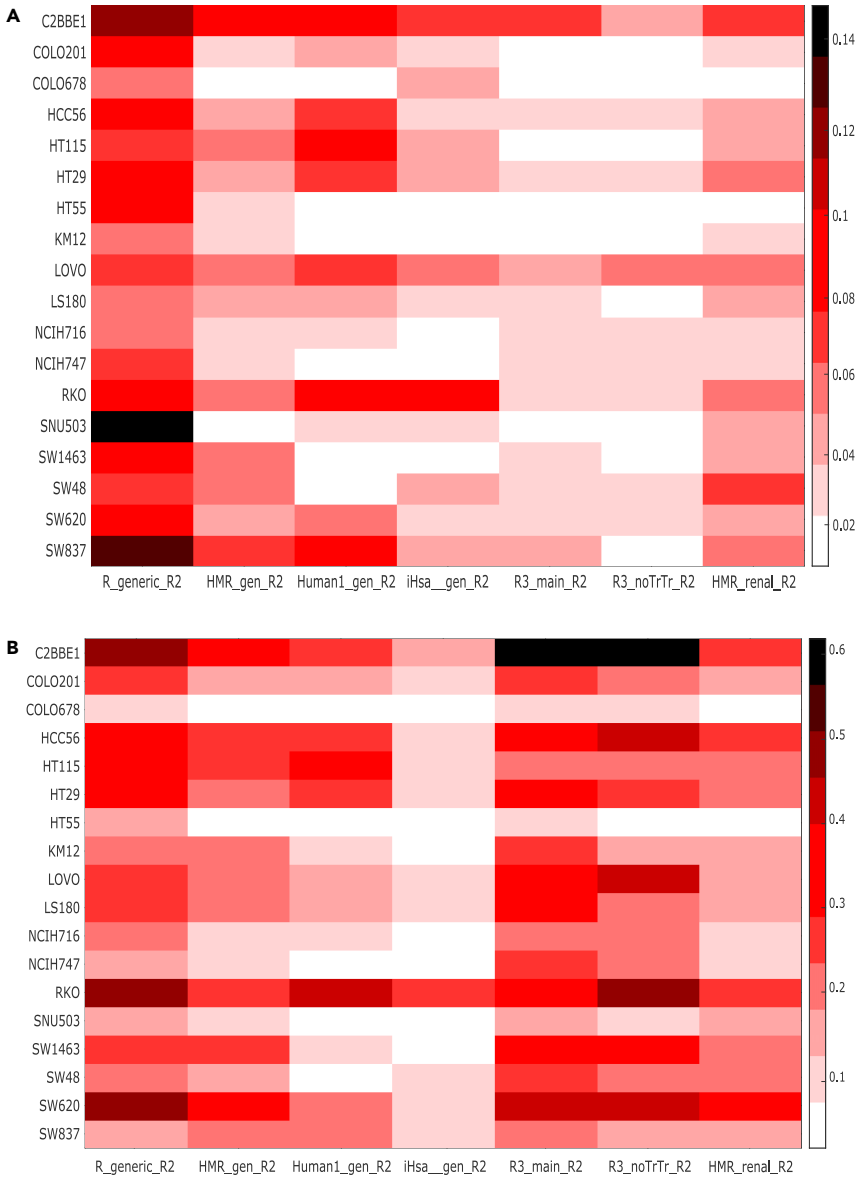



**Figure 1. Sensitivity and precision analysis for different biomass reactions using Recon 2 as input model**

(A) Sensitivity values were calculated based on TP/P.

(B) Precision values were calculated based on TP/(TP + FP). In both cases, a CRISPR-Cas 9 dataset containing experimentally identified essential genes was used (DepMap Achilles 19Q1, 2019). The y axis represents different CRC cell lines shared between CCLE and CRISPR datasets, whereas the x axis describes the biomass reaction set as an objective function. In all cases, Recon 2 was used as the input model (_R2). See also Figures S1, S2, and S3.

Overall, the R_generic_R2 biomass reaction led to the most sensitive ("_R2" indicates that Recon 2 was used as input model) and precise results in a majority of the cell lines, although significant differences could be observed across cell lines (Figure 1A). Nevertheless, the sensitivity values for all biomass reactions were relatively low (up to 0.15), meaning that most of the experimentally identified essential genes are not predicted as essential. Despite this, the precision results (Figure 1B) revealed that just up to 60% of the predicted essential genes were correctly predicted, representing a relatively high number of false positives in all biomass reactions, although it was higher in some objective functions such as Human1_gen_R2 and HMR_renal_R2. Alternatively, R_generic_R2 was giving the highest precision values for most of the cell lines. Last, specificity results show differences in the predicted non-essential genes based on the biomass reaction used as objective function (Figure S1). However, in this case, the values for all predictions

were close to 1, meaning that most of the non-essential genes were correctly identified as non-essential. Hence, although small differences were found between biomass reactions they were not significant.

Further, the number of predicted essential genes varies across the different biomass formulations, ranging from 83 genes for the R_generic_R2 prediction in the SW837 cell line to only 9 genes for the R3_noTrTr_R2 in the HT115 and LS180 cell lines (see Figure S2A). The highest number of essential genes was predicted by the HMR_renal_R2, Human1_gen_R2, and R_generic_R2 biomass reactions, although the number of genes common to the predictions of the three biomasses was low (Figure S3). Overall, these results hinted at the metabolites used to describe the biomass reaction as key players for the essential gene predictions.

Nevertheless, by including only a minimal set of reactions to guarantee flux through the biomass reaction (as done to test the biomass reactions using Recon 2 as input model) additional alternative pathways present in the input might have been missed, overestimating the essential genes. Thus, control tests were performed using each biomass reaction within the home model, e.g., iHsa_gen was also tested using iHsa as input model. Then, *in silico* single gene deletions were performed and their predictive capacity was assessed through precision, sensitivity, and specificity analysis (Figure 2). These results revealed significant differences in terms of the number of predicted essential genes (Figure S2B). In addition, sensitivity values increased for all biomass reactions, except for Recon 3 model, compared to the results obtained with Recon 2 as input model (Figure 2A), meaning that more known essential genes were being predicted as essential. Similarly, the precision values were increased for some of the biomass reactions such as Human1_gen_H1 and R_generic_R3, although the values were more similar to those obtained with Recon 2 as input model (Figure 2B). As previously mentioned, specificity values were calculated based on the true negative rate (Figure S4), and no significant differences were identified. In general, the generic Recon biomass reaction used in Recon 2 remains one of the most accurate essential gene predictions considering the three parameters, although the use of Recon 3 and the recon family biomass reaction led to higher precision values whereas the sensitivity values remained low, because the number of predicted essential genes was quite low as compared to Recon 2. On the other hand, Human1_gen_H1 within the home model showed a good performance, leading to higher sensitivity and precision values in most of the cell lines.

### The biomass coefficients

Because the different biomass reactions do not only differ in their metabolites but also the coefficients assigned to each metabolite, the impact of the coefficients' values was assessed in Recon 2 using the R_generic as objective function. In this case, the coefficient of each metabolite included in the R_generic was individually altered, considering values from 0.1 times to 10 times the original value. Then, *in silico* knockouts were performed for each coefficient value to further determine if these variations affect the prediction of essential genes. Out of the 38 metabolites included in the R_generic reaction, only coefficient variations of 7 metabolites (arginine, asparagine, ATP, cholesterol, cardiolipin, phosphatidylethanolamine, and sphingomyelin betaine) led to a change in the number of predicted essential genes (Figure 3). The enrichment results for these 7 relevant metabolites are represented in the C2BBE1 cell line, although similar results were obtained for the remaining 17 CRC cell lines (data not shown). Although 20 values were tested for each coefficient, only the lower and/or higher values led to different predictions. Hence, to ease the representation and understanding of the results, just 5 values covering the whole range were selected. Figure 3 represents the percentage of predicted essential genes that have been experimentally identified as essential (y axis) for each metabolite (color code) in the biomass reaction affected by the change of the coefficient values (x axis). Additionally, metabolic genes, referring to the percentage of genes within the metabolic model that have been identified as essential, representing the metabolic essential genes, showed a lower enrichment as compared to the predicted essential genes. The results showed that the discrepancies were small and were found just on a few values, especially on lowest and highest values, suggesting that variations in the coefficients do not significantly affect the essential gene prediction. In addition, by individually setting all coefficients to 0, we could also identify those metabolites to which the essential gene prediction is more sensitive, including the amino acid asparagine, the main source of energy ATP, and some lipids such as cardiolipin, cholesterol, phosphatidylethanolamine, and sphingomyelin betaine. The removal of these lipids decreased the number of predicted essential genes while keeping more or less the number of true positives, thus the sensitivity and specificity values increased, whereas the removal of ATP and arginine led to a higher number of predicted essential genes without impacting the number of true positives, leading to an increase in the number of false positives. For validation purposes, the analysis was repeated on the Recon 3 model using the same biomass objective function on the C2BBE1

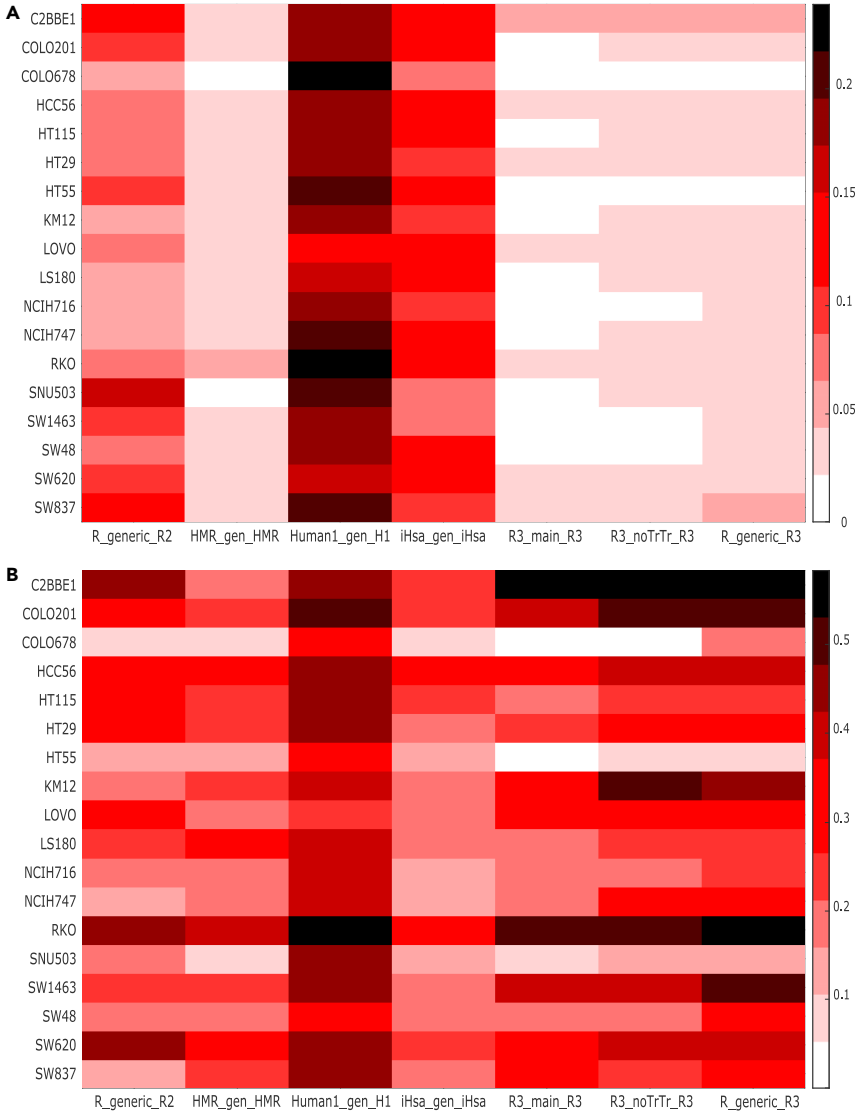



**Figure 2. Sensitivity and precision analysis for different biomass reactions using home models**

(A) Sensitivity values were calculated based on TP/P.

(B) Precision values were calculated based on TP/(TP + FP). In both cases, a CRISPR-Cas 9 dataset containing experimentally identified essential genes was used (DepMap Achilles 19Q1, 2019). The y axis represents different CRC cell lines shared between CCLE and CRISPR datasets, whereas the x axis describes the biomass reaction set as an objective function. Each biomass reaction was used within its own home model, represented by _InputModel (e.g.,_HMR). See also Figure S4.

cell line (Figure S5). In this case, the rate of true positives was increased as compared to Recon 2 predictions. However, this increase could be related to the lower number of predicted essential genes in Recon 3 and not to a better prediction capacity. Similar to Recon 2 results, the variation of the coefficients was significant just for highest and lowest values, whereas intermediate coefficient values did not impact the identification of essential genes. On the other hand, most of the affected metabolites were shared between both models although dttp was just identified using Recon 3.

## Different biomass reactions predict different drugs for repurposing

Finally, to identify currently approved drugs that could be repurposed for cancer treatment, DrugBank was data-mined to retrieve approved drugs that have an inhibitory effect on the predicted essential genes. In this case, predicted essential genes by each biomass reaction were pooled together as a unique list of

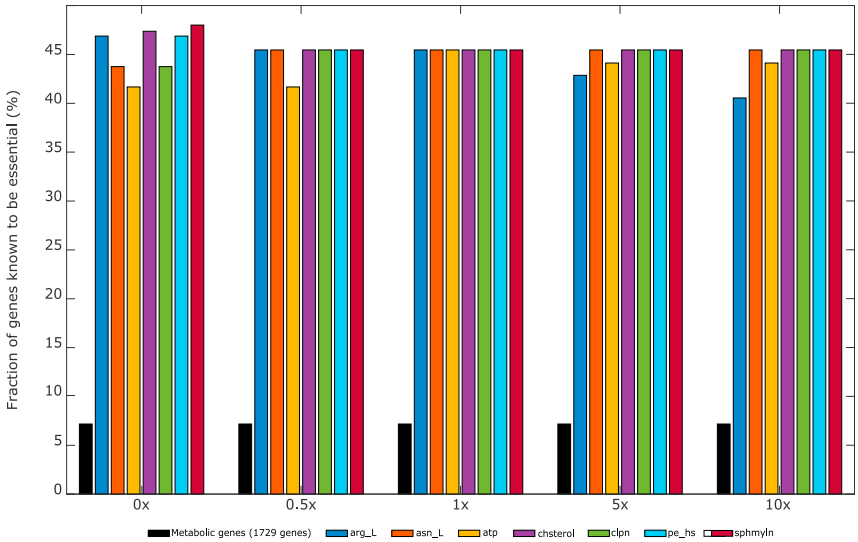



**Figure 3. Coefficients impact on essential gene prediction C2BBE1 cell line using Recon 2 as input model.**
The y axis represents the percentage of predicted genes matching the known essential genes from the CRISPR-Cas 9 dataset (DepMap Achilles 19Q1, 2019). On the other hand, the x axis represents the different values tested as coefficients, values multiplied by the original coefficient (0.5x would be a value 5 times less than the original). Essential genes show a higher enrichment as compared to metabolic genes, those genes within the metabolic reconstruction known as essential genes. In addition, the enrichment results are more or less constant despite the coefficient alteration. For representation purposes, only 5 out of the 20 tested coefficient values are represented on the plot. In the same manner, only metabolites that induced a change are depicted, 7 out of the 38 metabolites within the biomass reaction. See also Figure S5.

candidate genes. Then the effects of each drug were simulated by simultaneously knocking out all targeted genes. The use of different biomass reactions led to a heterogeneous list of drugs for repurposing (Figure 4A) ranging from 12 drugs predicted by R3_noTrTr_R2 to 97 drugs predicted by Human1_gen_R2. Thus, a comparison matrix has been used to display the differences in the lists of predicted drugs given by each biomass reaction. HMR_renal_R2, Human1_gen_R2, and R_generic_R2 identified the largest number of drugs, although the overlap with R_generic_R2 prediction was close to 50%. Simultaneously, this assay was done using home models as input models instead of Recon 2 (Figure 4B). This test revealed that the use of a different input model is also leading to different drug predictions, giving really high numbers of predicted drugs as compared to Recon 2 analysis. Despite the increase in the number of predicted drugs, the number of drugs approved for antineoplastic therapy in the SEER*RX database (Table 2) did not increase much in most of the cases, being the R_generic_R2 predicting the highest percentage of antineoplastic drugs, representing a higher level of confidence, while being able to identify novel drugs for repurposing in CRC.

## Different biomass reactions predict different growth rates

The versatility and predictive accuracy of the biomass reactions was also assessed on the growth rate prediction. The growth rates were predicted for five CRC cell lines from the NCI-60, namely HCT116, HCT15, HT29, SW620, and KM12. The selection of these cell lines was done based on exometabolomic data availability. Hence, just those CRC cell lines present in the CCLE and Zielinski et al. (2017) datasets were selected. Context-specific models using transcriptomic data from the cell lines mentioned above, the RPMI composition according to the experimental conditions used in (O'Connor et al., 1997) and the experimentally measured fluxes of carbon sources such as glucose, lactate, and several amino acids were obtained. In addition, to improve the predictions, the remaining carbon sources were constrained to 1% of the sum of experimental constraints, which represent 99% of the system's carbon source. Then, predicted growth rates were simulated via FBA and they were compared to the corresponding experimentally obtained growth rates (O'Connor et al., 1997) to evaluate the model prediction (Figure 5).

The growth rate analysis shows that the metabolic model predictions are still far from the experimentally measured data. However, those values closer to the experimental data were obtained with R_generic_R2, R3_main_R2, and R3_noTrTr_R2. Moreover, the model predictions also capture the general trend observed

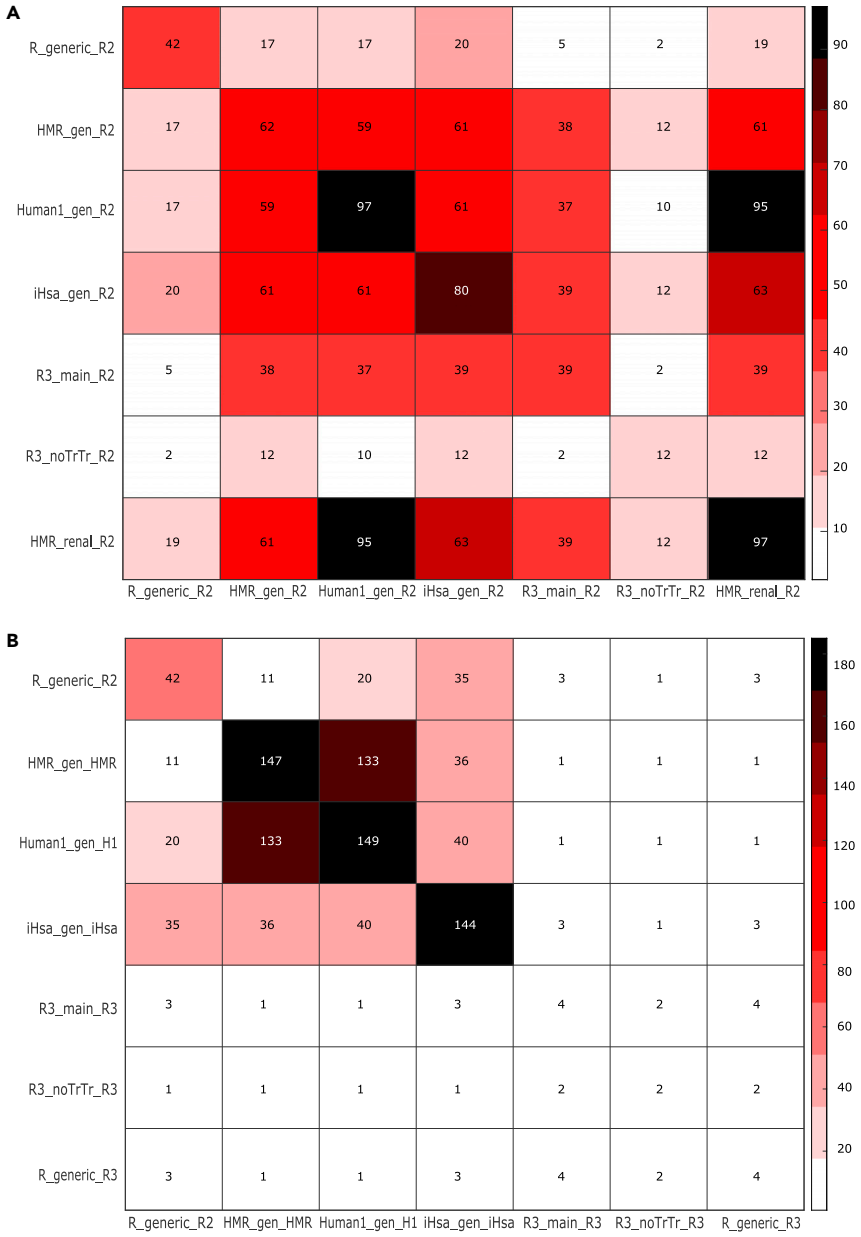



**Figure 4. Comparative analysis of drugs proposed for repurposing depending on the biomass reaction used as the objective function**

(A) Drugs predicted by each biomass reaction using Recon 2 as input model.

(B) Drugs predicted by each biomass reaction using home models. The numbers in the squares represent the shared number of drugs proposed for repurposing by each biomass reaction for each combination of couples of models. In the diagonal, since it is a comparison with itself, the numbers represent the number of drugs predicted by each model and biomass reaction.

in the measurements i.e., cell lines having a higher experimental growth rate also had a higher predicted growth rate. Alternatively, the remaining biomass reactions did not predict the experimentally measured growth rates as proficiently. Although HMR_renal_R2 was leading to predictions higher than the experimental values, predictions by the Human1_gen_R2 and iHsa_gen_R2 dropped to almost 0, being between 0.0001 and 0.001. In these cases, the applied constraints could have been too stringent, making it more difficult to produce biomass. On the other hand, as previously done for the essential gene prediction,

**Table 2. Number of predicted drugs for each biomass reaction corresponding to approved antineoplastic drugs according to SEER*RX database**

| Input model | R_generic biomass | Human1_ genbiomass | iHsa_gen biomass | R3_main biomass | R3_noTrTr biomass | R3_renal biomass | HMR_gen biomass |
|---|---|---|---|---|---|---|---|
| Recon2 | 19 (45.24%) | 22 (22.68%) | 14 (17.5%) | 2 (5.13%) | 1 (8.33%) | 22 (22.68%) | 12 (19.35%) |
| Home models | 1 (25%) | 38 (26.39%) | 49 (34.03%) | 1 (25%) | 1 (50%) | 1 (20%) | 31 (21.09%) |

the control tests were performed using each biomass reaction with the home models to predict the growth rate (Figure S6). Nevertheless, although a small improvement was observed on the R3_noTrTr_R3 and R3_main_R3 predictions, it did not lead to better growth rate predictions, whereas slightly better prediction values were obtained for some cell lines with R_generic_R3. Indeed, as previously observed in Figure 5, Human 1 and iHsa models led to a predicted growth rate close to 0. Thus, the most accurate growth rate predictions were obtained with the R_generic_R2 and the R_generic_R3, which led to almost the same results.

In an effort to explain the discrepancies between the measured and predicted fluxes, we stepwise modified each coefficient as explained previously. In total, 23 metabolites (ATP, cardiolipin pool, CTP, cysteine, dNTPs, glucose-6-phosphate, GTP, histidine, isoleucine, leucine, lysine, methionine, phosphatidylinositol, phosphatidylcholine pool, phosphatidylethanolamine pool, glycerophospholipid pool, phenylalanine, sphingomyelin, threonine, tyrosine, UTP, and valine) were identified that significantly change the growth rate prediction when their coefficients were modified. The impact was observed across the tested values (from 0.1 to 10 times more), although only some of them were included in the representations to facilitate the understanding. For instance, focusing on the HCT116 cell line results (Figure 6), changes in the coefficients of some metabolites led to different growth rate predictions, even dropping the growth rate close to 0 in the case of ATP. Similar results were obtained for the other CRC cell lines (Figure S7) as well as for the Recon 3 model with the generic biomass reaction (Figure S8) suggesting that, unlike the essential gene prediction, modifications of the coefficients may lead to significant changes in the predicted growth rates. Hence, we could conclude that the growth rate predictions are more sensitive to the stoichiometric coefficients of ATP, GTP, threonine, lysine, methionine, CTP, phosphatidylcholine, UTP, glucose 6 phosphate, histidine, leucine, phenylalanine, and valine, although they are less sensitive to the stoichiometric coefficients of cysteine, phosphatidylinositol, phosphatidylethanolamine, cardiolipin, dGTP, dCTP, dATP, dTTP, isoleucine, and sphingomyelin (Table S1). Alternatively, there are some metabolites whose coefficients are not impacting the growth rate prediction at all, such as water, glutamate, aspartic acid, asparagine, and alanine, among others. In addition, by individually setting all coefficients to 0, we could identify those metabolites to which the growth prediction is more sensitive, being ATP the one having the highest impact, leading to a value 2 times higher than the predicted value. Other metabolites relevant to the growth rate prediction are GTP, methionine, CTP, phosphatidylcholine, UTP, glucose-6-phosphate, and histidine. Nevertheless, a more profound analysis would be needed to determine why some metabolites of the biomass reactions are limiting the predictions. In conclusion, the disparities between the compositions of the biomass reactions directly affect the growth rate prediction, underlining the importance of describing the most representative and realistic biomass reaction using an appropriate and reproducible method.

## Cell-specific or tailored biomass reactions

To assess how data-driven algorithms for the creation of tailored biomass reactions perform in comparison to already published formulations, we have repeated the same analysis with a biomass reaction formula generated by the BOFdat algorithm (Lachance et al., 2019). Genomics, transcriptomics, metabolomics, and lipidomics data were required to run BOFdat. However, because of ethical concerns, human data is restricted and access to this data could not be obtained in a reasonable time, consequently data from the HeLa cell line was used instead for the biomass formulation. The formulation of the resulting BOFdat biomass reaction is shown in the Data S1.

Once the biomass reaction was obtained, in silico KOs were performed and the predictions were compared to the R_generic_R2 predictions using precision, sensitivity, and specificity analyses (Figure 7). Overall, the number of predicted essential genes was completely different between BOFdat_Biomass and R_generic_R2 predictions, ranging from 35 genes for the R_generic_R2 to 116 genes for the

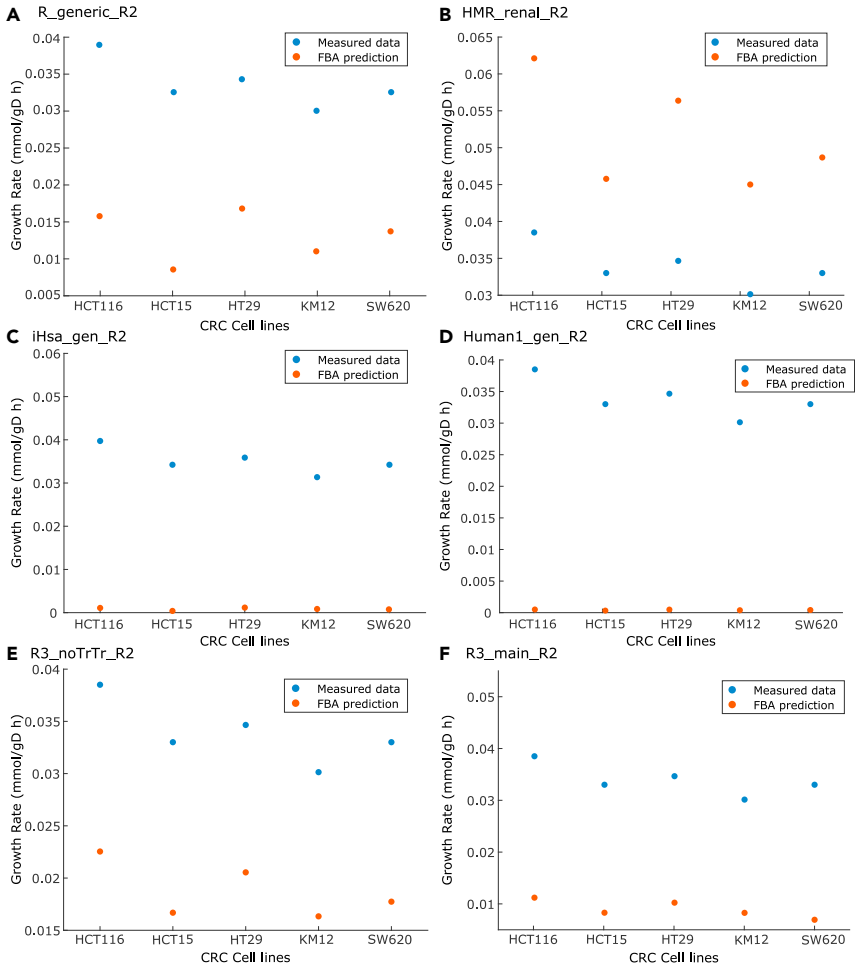



**Figure 5. Comparison of predicted growth rates and experimentally measured growth rates for different biomass reactions using Recon 2 as input model.**

The x axis indicates CRC cell lines present in the CCLE and the exometabolomic datasets whereas the y axis represents the growth rate (1/h).

(A–E) The predicted (orange dots) and experimental (blue dots) growth rates are plotted for each biomass reaction: (A) R_generic_R2, (B) HMR_renal_R2, (C) iHsa_gen_R2, (D) Human1_gen_R2, (E) R3_noTrTr_R2, and (F) R3_main_R2. See also Figure S6.

BOFdat_Biomass. Consequently, sensitivity, and specificity values were different. Sensitivity values revealed that the number of correctly predicted essential genes as compared to the total number of experimentally identified essential genes was higher than in the R_generic_R2 predictions (Figure 7A). Alternatively, precision values were higher for R_generic_R2 as compared to BOFdat_Biomass, although the values were relatively close (Figure 7B). Hence, the power of predicting essential genes increased with the BOFdat biomass reaction, leading to higher sensitivity values for all datasets, although the precision of such prediction was lower. Finally, the specificity values were higher for the R_generic_R2; however, the difference was not significant. Thus, we could conclude that the BOFdat biomass reaction led to more sensitive essential gene predictions.

## DISCUSSION

In this paper, we have benchmarked different biomass formulations from different metabolic reconstructions by integrating them into Recon 2. Cancer-specific metabolic models were extracted by the rFASTCORMICS workflow (Pacheco et al., 2019) and each biomass was added and used to predict essential genes and growth rates to minimize the impact of the input model structure on the prediction.

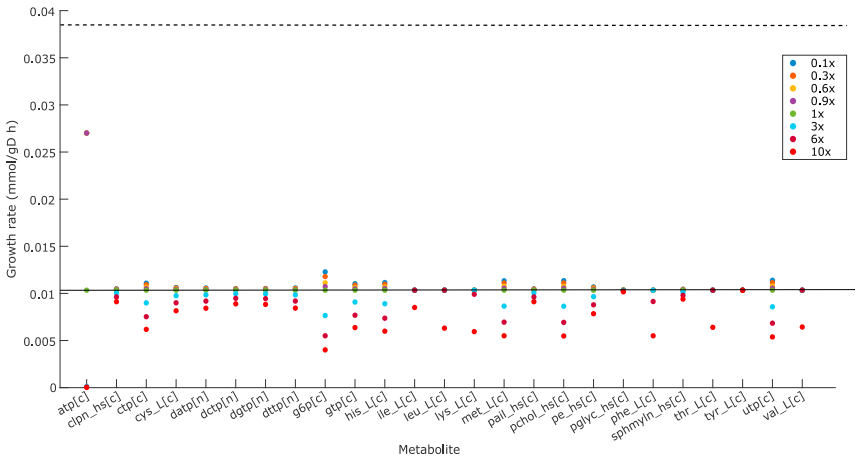



**Figure 6. Evaluation of the coefficient impact on the growth rate prediction, HCT116 cell line**

The x axis encompasses those metabolites whose coefficients were modified whereas the y axis represents the values for the growth rate prediction. Although 20 values were tested for every metabolite, from ten times less to ten times more, just nine values covering the whole range were included in the representation to ease the understanding. The legend stands for the value multiplied by the original coefficient, for instance 0.1x represents 0.1 times the original coefficient, meaning a coefficient 10 times lower. The line represents the growth rate value predicted by the R_generic_R2 and it is used as a guide to easily observe the impact of the coefficients, whereas the dashed line stands for the experimental value. See also Figures S7, S8, and Table S1.

## Models

Recon 2, although less recent than Recon 3, was used because it was shown in previous results to be an adequate input reconstruction for rFASTCORMICS (Pacheco et al., 2019). Furthermore, Recon 3 has a high percentage of artificial reactions, lacking evidence to occur *in vivo*, which could negatively affect the prediction power of extracted context-specific models. Indeed, as the validation results shown, the use of Recon 3 with its own biomass reaction does not exceed the Recon 2 results. Although the precision values were slightly higher in Recon 3, the sensitivity values dropped significantly due to the low number of predicted essential genes, which could be related to the high number of reactions as compared to Recon 2 while the number of genes is not drastically increased; hence the gene essentiality analyses are not significantly affected. In addition, the preliminary observations of the Recon 3 model shown that the number of alternative pathways could be too high, leading to a significant reduction in the identification of essential genes. On the other hand, Human 1 could have been considered as an input model, but the reconstruction was published after the start of the project, so the lack of time and experience was considered too low to use it as input for a benchmarking workflow. Furthermore, the exometabolomic data adjusted for GEM integration was just found for Recon 2 model, leading to more accurate and realistic results. Considering these facts, Recon 2 was seen as the most appropriate model for a benchmarking workflow.

Nevertheless, Recon 2 harbors many dead ends and gaps, causing a large number of reactions to be blocked (2123 out of 7440) possibly causing an overestimation of essential genes and false positives, because alternative pathways that are able to rescue a phenotype *in vivo* might be shut down. Ideally, dead-ends and gaps should be filled and further extended to accurately represent human metabolism. As dead-ends often result from a lack of knowledge or missing data on a specific reaction in the metabolism, it is difficult to fill these gaps without introducing reactions with poor evidence that might jeopardize the model's prediction power. Further, the existence of loops (a chain of reactions that form an internal cycle violating the thermodynamics laws) and stoichiometric inconsistencies can be another source of blocked reactions. Some algorithms (Schellenberger et al., 2011) have been published to break these thermodynamically infeasible loops by blocking a reaction and forcing the flux to exit the circle. However, the lack of availability of this type of data restricts their applicability to constraint-based modeling. Hence, to avoid bias and discussion that might arise from manual curation, we decided to limit the curation to the addition or removal of reactions that are necessary to allow the biomass functions to carry flux.

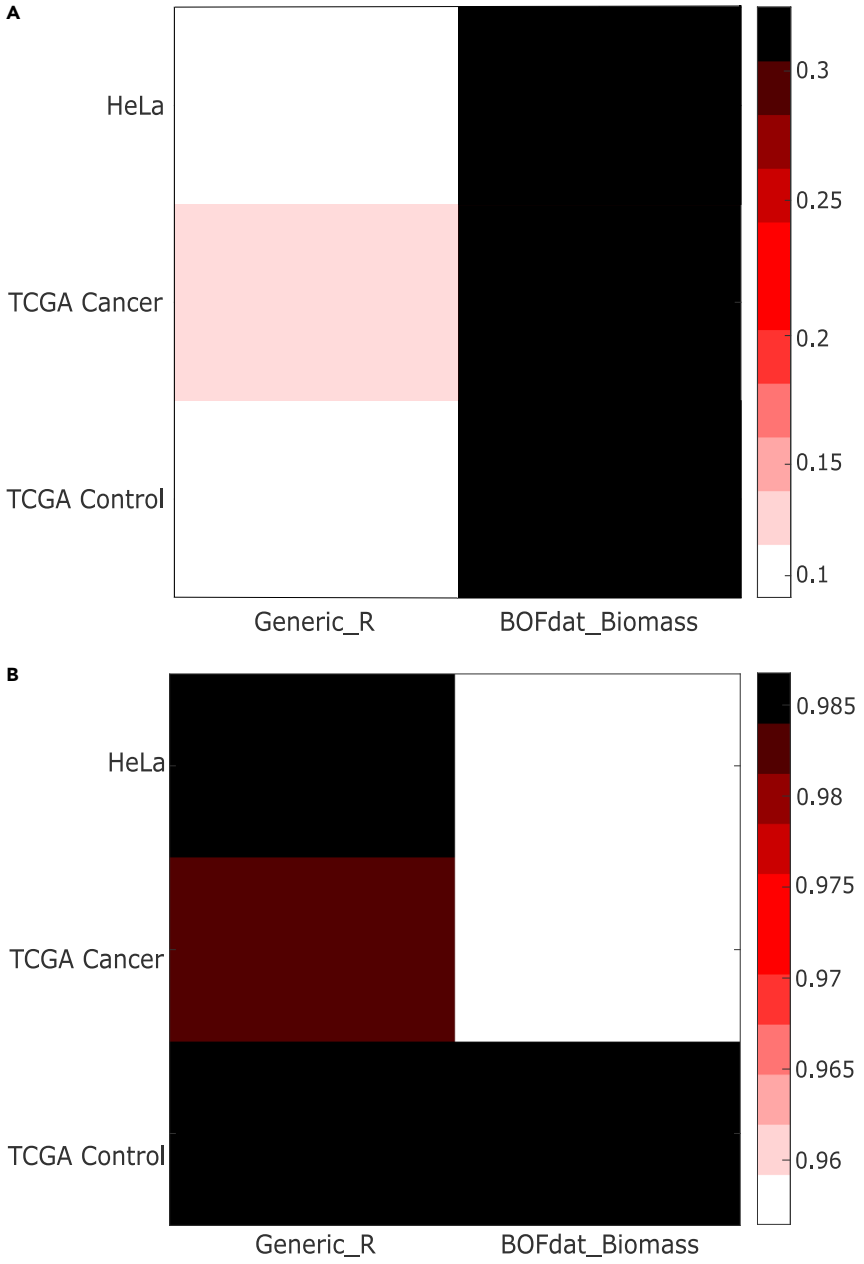



**Figure 7. Sensitivity and precision analyses for the HeLa cell line and TCGA data using two different biomass reactions**

(A) Sensitivity values were calculated based on TP/P, where TP is the number of predicted essential genes present in the CRISPR dataset (Hart et al., 2015) whereas P is the number of metabolic genes present in the CRISPR dataset.

(B) Precision values were calculated based on TP/(TP + FP), where FP are the number of genes predicted as essential which are non-essential according to the CRISPR-Cas 9 dataset. Recon 2 was used as an input model in both cases, although R_generic and BOFdat_biomass reactions were individually set as objective functions.

The input model also plays a role in the prediction of essential genes and growth rate as the number of reactions, metabolites, and especially genes vary from one reconstruction to another. Since the GPR rules define which genes need to be activated for a reaction to being able to carry a flux, they are important in the prediction of essential genes. Therefore, an improvement in the definition of these rules could lead to better enrichment results for most biomass reactions (Robinson et al., 2020). Despite this, the control tests revealed that although the precision of the essential gene predictions was slightly improved by the use of

Human 1 and Recon 3 models with its own biomass reaction, sensitivity values dropped. This could be explained by the fact that a drastic increase in the number of reactions was observed in Human 1 and Recon 3 followed by an increase in the number of genes and metabolites, although not to the same extent. This could come from an increase in the number of reactions involving the same main metabolites, or more reactions with no associated genes. Hence, the number of genes has increased, but maybe not sufficient to counterbalance the explosion of reactions.

### Gene screenings

Another limitation comes with the CRISPR-Cas 9 knockout screens that were used to validate gene essentiality. A comparison between the used CRISPR-Cas 9 KO from Project Score (Behan et al., 2019) and the DepMap dataset from the Broad Institute (DepMap Achilles 19Q1, 2019) revealed significant differences between both of them, which could be explained by the different experimental protocols and reagents. Furthermore, the data is provided as a binary matrix for each gene and cell line (1 standing for essential gene and 0 for non-essential); hence, a thresholding issue might arise from the conversion of continuous to binary data, leading to differential identification of essential genes. Thus, a potential bias coming from the validation data should be considered.

### Effect of the biomass reaction on gene predictions

The biomass reaction is a valid optimization function for highly proliferative cells such as cancer; however, the prediction power is dependent on the formulation. Even though we showed that the impact of the coefficients on the essential gene predictions is non-significant, it is for the growth rate prediction. Alternatively, the metabolite composition is crucial for both of them. For example, the R3_noTrTr_R2 and R3_main_R2 predicted fewer essential genes than the R_generic_R2. Both reactions lack replication precursors (and transcription and translation precursors for the R3_noTrTr), hence, the removal of these components leads to overlooking potential essential genes and a reduction in the predicted essential genes while still keeping a high percentage of correctly predicted genes. Following this hypothesis, the addition of metabolites into the biomass reaction formula slightly increased the number of predicted essential genes when using the HMR_renal_R2 or Human1_gen_R2 while decreasing the prediction accuracy. Because additional reactions had to be included to allow for metabolite production and consumption, a disruption in the added pathway by a gene deletion will lead to an overestimation of essential genes. The test controls revealed that the use of the original input model improves the sensitivity values for most of the reactions although it does not increase the precision and specificity values, except for the Human 1 model which showed higher values than those obtained with Recon 2 as input model. Last, Recon 3 with the generic biomass reaction showed a lower number of predicted essential genes, leading to lower sensitivity values, although the precision values increased. This reduction could be linked to the fact that many new reactions were integrated in Recon 3 as compared to Recon 2, although the increase of genes was not as significant. Thus, the higher number of alternative pathways keeping a relatively low number of genes could lead to the underestimation of the number of essential genes. Hence, these results demonstrate the importance of the biomass formulation to correctly and accurately predict essential genes, suggesting that the R_generic_R2 and Human1_gen_Input can be considered as the most appropriate biomass reaction for essentiality analysis.

The heterogeneity in terms of predicted essential genes for different cell lines could be related to the different medium compositions under study. Context-dependent essential genes are considered as essential in some specific circumstances or growth conditions but not in others. Thus, the different compositions could have led to the identification of some genes as essential which are not essential in a different medium. Similarly, the context-specific reconstruction, meaning the reactions included in the context-specific model, could be affecting the essential gene identification. If the alternative pathways that should be present in the context-specific model are associated with lowly expressed genes, rFASTCORMICS might not include this alternative pathway, as it is not supported by the data, leading to the wrong identification of a gene as essential and increasing the number of false positives. The same is true if the GPR is incomplete or wrong. If an isozyme that is controlling this pathway and expressed in the context of interest is missing in the rules, the target reaction might not be included in the output model, contributing to increase false positive rate. In addition, genetic differences between cell lines could also contribute to diverse essential gene identification. Nevertheless, most of the cell lines predictions are far from the experimental values, which could be explained by the fact that our predictions allow the identification of growth rate-related essential genes, owing to the set of the biomass objective function. Alternatively, the experimental lists of essential

genes contain known essential genes related to several metabolic functions. Thus, other necessary metabolic tasks besides growth such as *de novo* synthesis of nucleotides, uptake of essential amino acids, and beta-oxidation of fatty acids have previously been defined (Robinson et al., 2020). Future analyses are required to determine which essential genes can be linked to known metabolic tasks and which ones are resulting from limitations of the reconstructions. However, because cancer cells inside a tumor are highly heterogeneous, the identification of these tasks for cancer is not evident because of the existence of possible sub-clones that each have altered metabolisms and requirements.

Essential genes can be used as a surrogate for drug targets; in this case we considered as essential those genes whose deletion was leading to a growth decrease larger or equal to 50% as compared to the wild type. Indeed, targets allowing for a partial decrease of the growth rate might be relevant for drug targeting in situations where full inhibitions might e.g., lead to strong side effects. Therefore, differentially essential gene predictions also led to a distinct identification of drugs for repurposing. Nevertheless, a high overlap was found across predictions from different biomasses, indicating that the majority of the predicted inhibiting drugs were targeting essential genes. In the case of predictions made by the Human1_gen_R2 and HMR_renal_R2, more essential genes and, therefore, additional drugs were predicted that often had intermediate effects on the growth rate. Despite the increase in the number of predicted drugs in these reactions, R_generic_R2 presented the highest percentage of known antineoplastic drugs, increasing the level of confidence, while being able to predict new drugs that could be repurposed for the treatment of CRC.

### Effect of the biomass reaction on growth rate predictions

Another important feature of GEM-based analyses is the ability to predict growth rates and intracellular reaction fluxes. Hence, we have further assessed the effects of each biomass formulation through the growth rate prediction accuracy. Because metabolite uptake and secretion rates were not available for all cell lines to further constrain the models, exometabolomic data coupled to cell size and growth (Zielinski et al., 2017) were used for the context-specific growth rate prediction for CRC.

The applied constraints led to relatively accurate growth rate predictions for the R_generic_R2, R3_main_R2, and R3_noTrTr_R2; however, lowly predicted growth rates could represent a limitation in glucose uptake by the model and highly predicted growth rates could be related to the added reaction to allow for metabolite exchange. Similar results were observed for the control tests, using each biomass reaction with its respective model. Nevertheless, the used data has been tailored for Recon 2 and its use with other input models could be biasing the results. The recalculation of the upper and lower bounds for each input model used in this paper seemed outside of the scope of this study, with the core being the validation of the performance of different biomass reactions within the same input model, in this case Recon 2. Thus, future studies focused on the precise calculation of the growth rate should include a step adjusting the current values to the input model under study. Despite that, further exploration of individual metabolic systems needs to be performed to identify reactions that affect the predicted growth rate while also taking into account the appropriate upper and lower bound of the exchange reactions. Overall, and unlike the gene essentiality results, Human1_gen_R2 was not able to predict an accurate growth rate, whereas the R_generic_R2 and R_generic_R3 predictions were closer to the experimental measurements. Thus, so far, the generic biomass reaction from the Recon family seems to be the most appropriate considering its predictive capacity in terms of gene essentiality and growth rate.

### BOFdat

It is commonly accepted that the formulation of the biomass objective function depends on the composition of the cell, energetic requirements to generate biomass from metabolite precursors, and on the different species and cell types. Thus, the possibility of obtaining a human-specific biomass reaction based on experimental data using BOFdat was seen as an opportunity to improve model predictions. However, the required input data limits its applicability in drug target identification because of the availability of human omics data. To circumvent this issue, we have used the publicly accessible HeLa cell line data instead of CRC data. Using BOFdat, we were able to obtain a HeLa-specific biomass reaction, although the ideal scenario would have been to obtain a CRC tissue-specific biomass reaction. The resulting biomass reaction includes metabolites traditionally related to biomass generation such as amino acids, dNTPs, NTPs, and pools of several lipids. Interestingly, some of the predicted metabolites have never been observed in other

biomass reactions, although they are directly related with the biomass generation or the generation of metabolites directly involved with the biomass generation, including gamma-carboxyethyl-hydroxychroman metabolite (a metabolite of Vitamin E), thiosulfate and carbamoyl phosphate (which play a role in the biosynthesis of amino acids) among others. Other metabolites are related to the keratan sulfate I degradation products and biosynthesis precursors (ksi_pre29, ksi_deg16, and ksi_deg29, etc). Keratan sulfate is a glycoprotein that plays an important role in corneal transparency, developmental biology, cell signaling, adhesion, and migration (Habuchi et al., 2002). Although some studies reported the potential role of keratan sulfate on several cancers such as liver, lung, and pancreas (Miyamoto et al., 2011), there is no clear evidence of the role of such metabolites on tumor growth or biomass generation. Similarly, l2fn2m2masn metabolite is involved in the N-glycan biosynthesis, but there is no link between this pathway and the biomass generation. On the other hand, some metabolites suggested by BOFdat have not been experimentally identified in humans, such as 25aics, although they are present in human reconstructions. Last, the information of some metabolites included in the biomass formulation, such as m2gacpail_hs and g2m8masn, is scarce and no evidence about their role in the metabolism and their potential link with biomass generation were found. Hence, these metabolites might actually be stemming from missing pathways in the host model. These metabolites belong to a part of the metabolism which is not well described in the models and the inclusion of these metabolites in the biomass reaction could lead to wrong predictions. Consequently, an improvement in the algorithm and/or a process of manual curation after the generation of the biomass reaction could be performed to ensure that all metabolites included in the species-specific biomass reaction are related to the biomass generation.

The implementation of the HeLa biomass into the CRC model yielded poor results, demonstrating the HeLa specificity of the biomass reaction and the need for cell type-specific biomass formulations. To avoid mixing cell types, the HeLa biomass was implemented into a HeLa-specific metabolic model and compared with the R_generic_R2. The HeLa-specific biomass increases the gene enrichment by 15% and the true positive rate was also significantly higher although the false positive rate remains high. However, the performance of the BOFdat biomass could not be fully assessed because of the lack of exometabolomic data for the HeLa cell line. Thus, future experiments should include this analysis to conclude the predictive capacity and accuracy of species-specific biomass reactions generated using the BOFdat algorithm. Though, to correctly assess the BOFdat_biomass and compare the predictions with our previous results, the BOFdat_biomass should be based on CRC-specific data and integrated into a CRC-specific model. Moreover, ideally a manual curation of the data-driven biomass reaction would be performed to exclusively include genes related to biomass generation.

### Summary predictions

In conclusion, the integration of transcriptomic data into a metabolic model allowed the identification of potential targets for cancer treatment based on currently approved inhibiting drugs. However, the precision of these predictions lies in the definition of the biomass objective function. The role of the coefficients was remarkably more important on the growth rate predictions than on the essential gene prediction. Modifications of some coefficients reduced the predicted growth rate to almost zero, suggesting interplay between the biomass coefficients and media compositions, unlike the effect on the essential gene predictions. However, the formulation of the biomass in terms of the metabolites used to define it had a high impact on both predictions. The R_generic formulation from the Recon family models leads to the most accurate results in terms of essential gene and growth rate predictions in both Recon 2 and Recon 3 models, although the use of Recon 3 could be misleading to a low number of essential genes due to the increase in the number of reactions. Although Human1_gen led to more sensitive results for the essential gene predictions, its capacity to predict the growth rate was very low as compared to the R_generic. In addition, a data-driven algorithm was used and, although the essentiality analysis revealed an improvement in the model predictive capacity as compared to Recon 2, some of the metabolites included in the biomass formulation lack strong evidence of their role on the biomass generation. However, the biomass formulation should be tailored to its host model. It is not advisable to include more complex metabolites with various types of lipids if the related pathway in the host model is far from being complete or is only present in a simplified form as this would lead to an overestimation of the essential genes. As a general rule of thumb, the biomass formulation should ideally encompass metabolites that are connected to pathways that are well described in literature and in the model. Simpler formulation allows capturing a large fraction of known targets. Nevertheless, more complex or cancer-specific formulations could be helpful to propose more tailored drug treatments. But again, investing in developing more advanced formulation

would only make sense if it goes hand-in-hand with the curation of the related pathway and its GPR rules in the host model.

Overall, the results presented in this study showed that the definition of the biomass reaction is of utmost importance in prediction studies, such as growth rate and essentiality analysis, showing the need to find a standard definition of human cell-type-specific biomass reactions within the systems biology community.

### Limitations of study

There were several limitations to the present study. The benchmark of the biomass has been performed using Recon 2 as input model. As shown by the analysis performed using home models, various input models could lead to different predictions, although for most of them, including the most recently published reconstructions Recon 3 and Human 1, the overall predictive capacity is not improved. Hence, a benchmark analysis using another input model, could have led to slightly different results. Additionally, the CRISPR-Cas 9 data used for validation could be constraining the results. Two different datasets under similar conditions were found, although they showed different results. Thus, the use of one or another, as well as the discretization of the continuous values outputting the CRISPR-Cas 9 analysis could be limiting the validation of the predictions. Last, the exometabolomic data used for constraint purposes in the growth rate predictions, was fitted by the authors to the Recon 2 model, meaning that the use of this dataset using a different home model could be misleading.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data availability
- METHOD DETAILS
  - Input model
  - Data
  - Quality control
  - Model reconstruction
  - Medium composition
  - In silico screens
  - Precision, sensitivity, and specificity analyses
  - Drug target prediction
  - Biomass reaction integration
  - Growth rate prediction
  - BOFdat

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103110.

## ACKNOWLEDGMENTS

This study was supported by the University of Luxembourg and the Luxembourg National Research Fund (FNR PRIDE PRIDE15/10675146/CANBIO).

## AUTHOR CONTRIBUTIONS

M.M.G carried out the experiments. M.P.P and T.S conceived and planned the experiment. M.M.G, M.P.P and T.S analyzed the data. M.M.G and M.P.P wrote the manuscript. T.B and L.P assisted with the coding. All the authors read and revised the manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 25, 2021
Revised: July 27, 2021
Accepted: September 8, 2021
Published: October 22, 2021

## REFERENCES

Ashburn, T.T., and Thor, K.B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. Nat. Rev. Drug Discov. *3*, 673–683. https://doi.org/10.1038/nrd1468.

Behan, F.M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C.M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., et al. (2019). Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature *568*, 511–516. https://doi.org/10.1038/s41586-019-1103-9.

Bekker-Jensen, D.B., Kelstrup, C.D., Batth, T.S., Larsen, S.C., Haldrup, C., Bramsen, J.B., Sørensen, K.D., Høyer, S., Ørntoft, T.F., Andersen, C.L., et al. (2017). An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. Cell Syst. *4*, 587–599.e4. https://doi.org/10.1016/j.cels.2017.05.009.

Blais, E.M., Rawls, K.D., Dougherty, B.v., Li, Z.I., Kolling, G.L., Ye, P., Wallqvist, A., and Papin, J.A. (2017). Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. Nat. Commun. *8*, 14250. https://doi.org/10.1038/ncomms14250.

Blighe, K., and Lewis, M. (2019). PCAtools: Everything Principal Components Analysis, R. Package Version 1.0.0.

Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., et al. (2018). Recon3 enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. *36*, 272–281. https://doi.org/10.1038/nbt.4072.

Chan, S., Cai, J., and Wang, L. (2017). Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. Bioinformatics *33*, 3603–3609. https://doi.org/10.1093/bioinformatics/btx453.

Chang, R.L., Xie, L., Xie, L., Bourne, P.E., and Palsson, B.Ø. (2010). Drug off-target effects predicted using structural analysis in the context of a metabolic network model. PLoS Comput. Biol. *6*, e1000938. https://doi.org/10.1371/journal.pcbi.1000938.

Cottret, L., Frainay, C., Chazalviel, M., Cabanettes, F., Gloaguen, Y., Camenen, E., Merlet, B., Heux, S., Portais, J.-C., Poupin, N., et al. (2018). MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res. *46*, W495–W502. https://doi.org/10.1093/nar/gky301.

DeBusk, R.F., Pepine, C.J., Glasser, D.B., Shpilsky, A., DeRiesthal, H., and Sweeney, M. (2004). Efficacy and safety of sildenafil citrate in men with erectile dysfunction and stable coronary artery disease. Am. J. Cardiol. *93*, 147–153. https://doi.org/10.1016/j.amjcard.2003.09.030.

DepMap Achilles 19Q1 (2019). Public. https://doi.org/10.6084/m9.figshare.7655150.v1.

Dias, O., Rocha, M., Ferreira, E.C., and Rocha, I. (2015). Reconstructing genome-scale metabolic models with merlin. Nucleic Acids Res. *43*, 3899–3910. https://doi.org/10.1093/nar/gkv294.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc. Natl. Acad. Sci. *104*, 1777–1782. https://doi.org/10.1073/pnas.0610772104.

Feist, A.M., and Palsson, B.O. (2010). The biomass objective function. Curr. Opin. Microbiol. *13*, 344–349. https://doi.org/10.1016/j.mib.2010.03.003.

Fogel, D.B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp. Clin. Trials Commun. *11*, 156–164. https://doi.org/10.1016/j.conctc.2018.08.001.

Gille, C., Bölling, C., Hoppe, A., Bulik, S., Hoffmann, S., Hübner, K., Karlstädt, A., Ganeshan, R., König, M., Rother, K., et al. (2010). HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Mol. Syst. Biol. *6*, 411. https://doi.org/10.1038/msb.2010.62.

Global Burden of Disease Cancer Collaboration, Fitzmaurice, C., Akinyemiju, T.F., Al Lami, F.H., Alam, T., Alizadeh-Navaei, R., , et al.Allen, C., Alsharif, U., Alvis-Guzman, N., Amini, E. (2018). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer Groups, 1990 to 2016: a systematic analysis for the Global Burden of Disease Study. JAMA Oncol. https://doi.org/10.1001/jamaoncol.2018.2706.

Habuchi, H., Nogami, K., and Kimata, K. (2002). Heparan sulfate: structure, biosynthesis, and functions. Connective Tissue *34*, 249–259.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell *163*, 1515–1526. https://doi.org/10.1016/j.cell.2015.11.015.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat. Protoc. *14*, 639–702. https://doi.org/10.1038/s41596-018-0098-2.

Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., and Stevens, R.L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat. Biotechnol. *28*, 977–982. https://doi.org/10.1038/nbt.1672.

Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A.L., Kafri, R., Kirschner, M.W., Clish, C.B., and Mootha, V.K. (2012). Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. Science *336*, 1040–1044. https://doi.org/10.1126/science.1218595.

Jeucken, A., and Brouwers, J. (2019). High-throughput screening of lipidomic adaptations in cultured cells. Biomolecules *9*, 42. https://doi.org/10.3390/biom9020042.

Knorr, A.L., Jain, R., and Srivastava, R. (2007). Bayesian-based selection of metabolic objective functions. Bioinformatics *23*, 351–357. https://doi.org/10.1093/bioinformatics/btl619.

Koch, F., and Koch, G. (1985). The Molecular Biology of Poliovirus (Springer). https://doi.org/10.1007/978-3-7091-7000-7.

Lachance, J.-C., Lloyd, C.J., Monk, J.M., Yang, L., Sastry, A.v., Seif, Y., Palsson, B.O., Rodrigue, S., Feist, A.M., King, Z.A., and Jacques, P.-É. (2019). BOFdat: generating biomass objective functions for genome-scale metabolic models from experimental data. PLoS Comput. Biol. *15*, e1006971. https://doi.org/10.1371/journal.pcbi.1006971.

Lee, S., Lee, K.H., Song, M., and Lee, D. (2011). Building the process-drug–side effect network to discover the relationship between biological Processes and side effects. BMC Bioinform. *12*, S2. https://doi.org/10.1186/1471-2105-12-S2-S2.

Maldonado, E.M., Fisher, C.P., Mazzatti, D.J., Barber, A.L., Tindall, M.J., Plant, N.J., Kierzek, A.M., and Moore, J.B. (2018). Multi-scale, whole-system models of liver metabolic adaptation to fat and sugar in non-alcoholic fatty liver disease. Npj Syst. Biol. Appl. *4*, 33. https://doi.org/10.1038/s41540-018-0070-3.

Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., and Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat. Commun. *5*, 3083. https://doi.org/10.1038/ncomms4083.

Miyamoto, T., Ishii, K., Asaka, R., Suzuki, A., Takatsu, A., Kashima, H., and Shiozawa, T. (2011). Immunohistochemical expression of keratan sulfate: a possible diagnostic marker for carcinomas of the female genital tract. J. Clin. Pathol. *64*, 1058–1063. https://doi.org/10.1136/jclinpath-2011-200231.

Montezano, D., Meek, L., Gupta, R., Bermudez, L.E., and Bermudez, J.C.M. (2015). Flux balance analysis with objective function defined by proteomics data—metabolism of *Mycobacterium tuberculosis* exposed to mefloquine. PLoS One *10*, e0134014. https://doi.org/10.1371/journal.pone.0134014.

O'Connor, R., Kauffmann-Zeh, A., Liu, Y., Lehar, S., Evan, G., Baserga, R., and Blättler, R. (1997). The IGF-I receptor domains for protection from apoptosis are distinct from those required for proliferation and transformation. Mol. Cell. Biol. *17*, 427–435.

Pacheco, M.P., Bintener, T., Ternes, D., Kulms, D., Haan, S., Letellier, E., and Sauter, T. (2019). Identifying and targeting cancer-specific metabolism with network-based drug target prediction. EBioMedicine *43*, 98–106. https://doi.org/10.1016/j.ebiom.2019.04.046.

Pfau, T., Pacheco, M.P., and Sauter, T. (2016). Towards improved genome-scale metabolic network reconstructions: Unification, transcript specificity and beyond. Briefings in Bioinformatics *17*, 1060–1069. https://doi.org/10.1093/bib/bbv100.

R Core Team (2019). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing), Retrieved from. https://www.r-project.org/.

Robinson, J.L., Kocabaş, P., Wang, H., Cholley, P.-E., Cook, D., Nilsson, A., Anton, M., Ferreira, R., Domenzain, I., Billa, V., et al. (2020). An atlas of human metabolism. Sci. Signal. *13*, eaaz1482. https://doi.org/10.1126/scisignal.aaz1482.

Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat. Protoc. *6*, 1290–1307. https://doi.org/10.1038/nprot.2011.308.

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional optimality of microbial metabolism. Science *336*, 601–604. https://doi.org/10.1126/science.1216882.

Thiele, I., Swainston, N., Fleming, R.M.T., Hoppe, A., Sahoo, S., Aurich, M.K., Haraldsdottir, H., Mo, M.L., Rolfsson, O., Stobbe, M.D., et al. (2013). A community-driven global reconstruction of human metabolism. Nat. Biotechnol. *31*, 419–425. https://doi.org/10.1038/nbt.2488.

Tu, Y., Wu, X., Yu, F., Dang, J., Wei, Y., Yu, H., Liao, W., Zhang, Y., and Wang, J. (2020). Tristetraprolin-RNA interaction map reveals a novel TTP-RelB regulatory network for innate immunity gene expression. Mol. Immunol. *121*, 59–71. https://doi.org/10.1016/j.molimm.2020.02.004.

Vlassis, N., Pacheco, M., and Sauter, T. (2014). Fast Reconstruction of Compact Context-Specific Metabolic Network Models. PLoS Computational Biology *10*. https://doi.org/10.1371/journal.pcbi.1003424.

Vo, T.D., Greenberg, H.J., and Palsson, B.O. (2004). Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. J. Biol. Chem. *279*, 39532–39540. https://doi.org/10.1074/jbc.M403782200.

Wagner, A., Zarecki, R., Reshef, L., Gochev, C., Sorek, R., Gophna, U., and Ruppin, E. (2013). Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious. Proc. Natl. Acad. Sci. *110*, 19166–19171. https://doi.org/10.1073/pnas.1312361110.

Zielinski, D.C., Jamshidi, N., Corbett, A.J., Bordbar, A., Thomas, A., and Palsson, B.O. (2017). Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism. Sci. Rep. *7*, 41241. https://doi.org/10.1038/srep41241.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| CCLE's dataset | Broad Institute | https://depmap.org/portal/download/ |
| CRISPR-Cas 9 dataset | Broad Institute | https://depmap.org/portal/download/ |
| Growth rate's dataset | O'Connor et al., 1997 | PMID: 9331090 |
| Tu's dataset | NCBI SRA | GSE114216 |
| Bekker-Jensen's dataset | ProteomeCentral | PXD004452 |
| Jeucken & Brouwers's dataset | Jeucken & Brouwers, 2019 | doi: https://doi.org/10.3390/biom9020042 |
| Hart's dataset | NCBI SRA | GSE75189 |
| Exometabolomic's dataset | Zielinski et al., 2017 | doi: https://doi.org/10.1038/srep41241 |
| Software and algorithms | | |
| rFASTCORMICS | Pacheco et al., 2019 | https://github.com/sysbiolux/rFASTCORMICS |
| BOFdat | Lachance et al., 2019 | https://github.com/jclachance/BOFdat |
| Cobra Toolbox | Heirendt et al., 2019 | https://github.com/opencobra/cobratoolbox/tree/master/src |
| Other | | |
| Recon 2 | VMH | https://www.vmh.life/#downloadview |
| Recon 3 | VMH | https://www.vmh.life/#downloadview |
| iHsa | GitHub | https://github.com/csbl/ratcon1 |
| HMR | BioModels | MODEL1402200003 |
| Human 1 | GitHub | https://github.com/SysBioChalmers/Human-GEM/tree/master/model |

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for code or datasets should be directed to and will be fulfilled by the lead contact, Dr. Thomas Sauter (Thomas.sauter@uni.lu).

#### Materials availability

This study did not generate new unique reagents.

#### Data availability

- This paper analyses existing, publicly available data. The source of the data is listed in the Key resources table.
- All original code has been deposited at GitHub (https://github.com/sysbiolux/Biomass_formulation/) and is publicly available as of the date of publication.
- Any additional information to reproduce the results reported in this paper is available from the lead contact upon request.

### METHOD DETAILS

#### Input model

The human genome-scale metabolic reconstruction (GEM) Recon 2 (Thiele et al., 2013) was used as an input model for the reconstruction of different cancer cell lines and patient models via the rFASTCORMICS workflow (Pacheco et al., 2019). Recon 2 (Thiele et al., 2013) is composed of 7440 reactions, 5063 metabolites, and 2140 genes whereas 1733 are unique genes. Since Recon 2 contains blocked reactions, a flux consistent version of the Recon 2 model (5317 reactions, 2960 metabolites, 1913 genes, and 1550 unique genes) was

obtained by running FASTCC (Vlassis et al., 2014). Recon 2 was used to obtain CRC context-specific models and to perform the assessment of the biomass reactions by individually integrating each of them and running gene essentiality and growth rate analyses. Additionally, for validation purposes HMR 2.0 (Mardinoglu et al., 2014) (7573 reactions, 4906 metabolites and 3765 genes), iHsa (Blais et al., 2017) (5838 reactions, 3591 metabolites and 2315 genes), Human 1 (Robinson et al., 2020) (11902 reactions, 7140 metabolites and 3628 genes) and Recon 3 (Brunk et al., 2018) (10546 reactions, 5816 metabolites, 2248 genes, and 1866 unique genes) models were used with its own biomass reaction to identify any bias related to the used input model. These models were obtained from different sources: BioModels for HMR2.0 (https://www.ebi.ac.uk/biomodels/MODEL1402200003#Files), VMH for the Recon family models (https://www.vmh.life/#downloadview), SysBioChalmers GitHub page for Human 1 model (https://github.com/SysBioChalmers/Human-GEM/tree/master/model) and Csbl GitHub page for iHsa model (https://github.com/csbl/ratcon1).

Finally, Recon 2 was used as input model for the HeLa context-specific model to test the BOFdat biomass reaction.

### Data

For this study, cancer cell line data from the Cancer Cell Line Encyclopedia (CCLE) was used. The data was retrieved from the Broad Institute CCLE project website (https://portals.broadinstitute.org/ccle), containing FPKM values for 56 CRC cell lines. However, not all of them were considered since the validation data used for the gene essentiality and growth rate predictions was not available for all of them. Thus, 18 CRC cell lines were finally considered for the gene essentiality analysis whereas just 5 cell lines were included in the growth rate predictions.

### Quality control

Both datasets were individually log2-transformed and then a principal component analysis (PCA) was performed in R (R Core Team, 2019) using the PCAtools package (Blighe and Lewis, 2019) to assess data segregation and quality. The PCA clusters showed that cell lines and patient data, clearly clustered in the function of the cancer and control condition.

### Model reconstruction

Context-specific models were reconstructed with rFASTCORMICS (Pacheco et al., 2019) using different consistent models and RNA-seq data as input. rFASTCORMICS allows building two types of models: (i) consensus models, corresponding to the different phenotypes (i.e., CRC, controls) for which all the samples of the same condition were pooled, thus only genes expressed or not expressed in 90% of the samples are tagged as active or inactive, respectively; (ii) sample-specific models, where individual models are built for each sample. In this case, sample-specific models were obtained by using CRC cell line transcriptomic data. Additionally, the models were individually constraint according to the media composition used in the experimental data (Data S2) considered for validation purposes, while the remaining metabolites were unconstraint. Last, the biomass reaction was set as objective function for all the models. Once the model reconstructions were obtained, gene essentiality and growth rate predictions were run using these models.

### Medium composition

Context-specific models were constraint according to the culture medium used in the experimental CRISPR-Cas 9 (DepMap Achilles 19Q1, 2019) and growth rate analysis (O'Connor et al., 1997), in order to obtain comparable results. The 18 cell lines considered in gene essentiality analysis were cultured in 7 different media, including RPMI, DMEM, EMEM, F12K, McCoys, Leibovitz and MEM media. Alternatively, the 5 cell lines included in the growth rate analysis were treated with RPMI medium. The composition of these media was obtained from the supplier website and each metabolite was written in a form compatible with the input model (i.e. glycine was included as gly[c]). However, in some cases, the addition of just those metabolites included in the technical sheet of the supplier were not sufficient for the biomass reaction to carry a flux. Hence, essential metabolites were identified and added to the medium composition. This process was repeated for every medium and biomass reaction, to obtain compatible media allowing the flux of the biomass reactions. A detailed description of the medium composition used for constraining the models can be found in the Data S2 as well as in the GitHub repository as MATLAB files to ease the reproducibility of the results.

### In silico screens

*In silico* knock-outs for context-specific models were simulated to identify growth-inhibiting drug targets. Thus, a modified version of the *singleGeneDeletion* function from the COBRA toolbox (Heirendt et al., 2019) was used in Matlab 2019a and each biomass reaction was individually set as the objective function. Genes were considered to be essential if the growth after the gene deletion was less or equal to 50% of the wild type growth.

### Precision, sensitivity, and specificity analyses

Three statistical measures were used to assess the performance of the gene essentiality predictions: precision (TP/TP + FN), sensitivity (TP/P), and specificity (TN/N). Thus, genome-scale CRISPR-Cas 9 screenings performed on hundreds of human cancer cell lines from the Broad Institute were used (DepMap Achilles 19Q1, 2019). The DepMap Achilles dataset contains the results of genome-scale CRISPR knockout screens for 17,634 genes in 563 cell lines, of which just 18 samples corresponding to CRC samples present in the CCLE dataset were selected.

### Drug target prediction

The predicted essential genes can be considered as potential targets for drugs. Hence, the predicted essential genes by different biomass reactions were pooled together and were used as input for the drug identification. Their Entrez Gene IDs were converted to the corresponding UniProtKB/Swiss-Prot Identifiers (UniProtID), which were then matched to the DrugBank database dictionary to find all the interacting drugs. Only approved drugs with an inhibiting effect were selected.

### Biomass reaction integration

The main goal of the study was to identify how the definition of the biomass objective function affects the model predictions. Thus, we identified all human-related biomass reactions that had been defined in the literature in order to evaluate all of them to determine their impact and the most accurate biomass objective function. The evaluation of the different biomass reactions required the individual integration of each of them into Recon 2 to further optimise and compare results. First, the formulation of the biomass reaction to be tested was obtained by using the model containing such reaction. For example, according to the Human 1 model its biomass reaction is:

45 m01371c + 0.0267 m01721n + 45 m02040c + 0.1124 m02847c + 0.4062 m03161c + 0.0012001c m10012c + 5.3375 m10013c + 0.2212 m10014c + 0.4835 m10015c -> 45 m02039c + 45 m02751c + temp

Then, the names of the components of each biomass reaction were converted into Recon 2 compatible names by using the Metabolic Atlas repository (Robinson et al., 2020):

45 atp[c] + 0.0267 dna[c] + 45 h2o[c] + 0.1124 rna[c] + 0.4062 gly[c] + 0.0012 cofactor_pool_biomass[c] + 5.3375 protein_pool_biomass[c] + 0.2212 lipid_pool_biomass[c] + 0.4835 metabolite_pool_biomass[c] -> 45 h+[c] + 45 pi[c] + biomass[c]

However, not all metabolites encompassed in the biomass reaction under study were present in Recon 2, thus the missing metabolites and their associated reactions were added to the metabolic reconstruction to obtain a consistent model that could be further assessed, according to the information in the Metabolic Atlas (Robinson et al., 2020). For example, dna[c] does not exist as such in the Recon 2 model, thus the reaction producing it was integrated:

0.3 datp[c] + 0.2 dctp[c] + 0.2 dgtp[c] + 0.3 dttp[c] -> dna[c] + ppi[c]

In some cases, the newly added reaction contained metabolites not present in the Recon 2 model, hence additional reactions were required. For instance, N-retinylidene-N-retinylethanolamine (m02624c), part of the cofactor pool of the Human 1 biomass reaction, is not present in Recon 2 and its integration requires several extra reactions to lead to a consistent model. Hence, since this could be an endless process a maximum number of four additional reactions per metabolite was considered. Otherwise the metabolite was excluded from the biomass formulation. Following this approach, just a few metabolites were excluded

in some biomass reactions: 4 metabolites from the Human 1 biomass (m00611c, m00611r, m02394c, m02624c), 1 metabolite from the iHsa biomass (m01580c), 3 metabolites from the Renal biomass (m00209c, m01631m, m02624c), and 4 metabolites from the HMR biomass (m00209c, m01631m, m02624c, m02629c). In this case, the use of Recon 3 would have not significantly improved the biomass addition step since most of the removed metabolites (m00209c, m01631m, m02624c, m02629c, m00611c, m00611r) and their associated reactions are also missing in the consistent Recon 3 model. Similarly, those metabolites whose addition was causing the formation of loops were not considered, such as the protein pools of Human 1 and iHsa biomass reactions. A detailed composition of each biomass reconstruction including the final considered metabolites and the additional reactions required for the model consistency can be found in the Data S1. Finally, the generic biomass reaction present by default in Recon 2 was replaced by the biomass to be tested. Then, the impact of each biomass reaction was assessed by analysing the essential gene and the growth rate prediction.

### Growth rate prediction

After obtaining the context-specific models using transcriptomic data and the RPMI medium composition as previously described in the Model reconstruction section, Flux Balance Analysis (FBA) was used to calculate the flux of the objective function, using the *optimizeCbModel* function from the COBRA toolbox (Heirendt et al., 2019) in Matlab 2019a. Additionally, metabolite uptake and secretion profiles for several human cancer cell lines were integrated as FBA constraints in each model, to improve the model predictions. The consumption and release of 219 metabolites measured via mass spectrometry from media for the NCI-60 cell lines (Jain et al., 2012) coupled to growth and cell size data (Zielinski et al., 2017) were used. In total, the experimental measurements of 23 metabolites were considered, including glucose, lactate, and amino acids, which represent 99% of the observed metabolic exchange fluxes in the NCI-60 cell line dataset (Lee et al., 2011). These values were used to set the upper and lower bounds of the production and uptake reactions, respectively. In addition, in order to have the maximum metabolic effort into the biomass production, the lower bounds of the remaining carbon exchange reactions included in each context-specific model were set to 1% of the sum of all the experimental values.

This analysis was performed on five CRC cell lines (HCT116, HCT15, SW620, KM12, and HT29) shared between the NCI-60 and the CCLE datasets. Furthermore, the biomass reactions were individually integrated into each cell line-specific model, to assess the performance of each reaction in each model. Finally, the predicted growth rates were compared to experimentally obtained doubling times (O'Connor et al., 1997). These doubling times were represented in hours, thus the corresponding growth rate was obtained by the following formula assuming exponential growth: growth rate = ln (2)/doubling time.

### BOFdat

BOFdat algorithm was used to obtain a data-driven human-specific biomass reaction (Lachance et al., 2019). Step 1 aims to generate the stoichiometric coefficients for major macromolecules (DNA, RNA, proteins, and lipids) using different -omics data. In this case, due to the limited accessibility to some -omics data on CRC samples, we focused on HeLa cell line data. Thus, the dry weight of the HeLa cell constituents was obtained (Koch and Koch, 1985) and the reference human genome GRCh38 (https://www.ncbi.nlm.nih.gov/genome/51) was used as genomic data since variations between human genomes are assumed to be very low. Transcriptomic, proteomic and lipidomic data were retrieved from GSE114216 (Tu et al., 2020) (Bekker-Jensen et al., 2017), and (Jeucken and Brouwers, 2019), respectively. Next, Step 2, which aims to find inorganic ions and coenzymes and to generate their stoichiometric coefficients, was based on currently available knowledge. The promiscuous nature of coenzymes was used to identify them by determining the number of reactions in which each metabolite participates and selecting the most promiscuous. Alternatively, as none of the biomass reactions included inorganic elements they were not considered for the biomass formulation. Lastly, step 3 aims to identify the remaining species-specific metabolic objective and generate their stoichiometric coefficients. In this case, experimentally obtained HeLa essential genes (Hart et al., 2015) were given as input and multiple iterations of a genetic algorithm (GA) were applied for the identification of such metabolites. A GA was applied to the initial populations by iteratively applying genetic operators (mutation, mating, and selection). It led to a clustered distance matrix containing the frequency of apparition of each metabolite across all individuals. Finally, the weight of each cluster determined the number of metabolites that should be added to the final biomass objective function suggested by BOFdat, leading to the final formulation of the biomass objective function (Table S1). Once the new biomass reaction was obtained from the BOFdat algorithm, it was integrated into Recon 2 to perform

*in silico* knockouts and to assess the predictive accuracy of the new objective function. First, HeLa context-specific models were obtained by using FPKM data from the HeLa cell line (Tu et al., 2020) and constraining the model with the DMEM media composition. Then, precision, sensitivity and specificity analysis were performed on CRISPR-Cas 9 data from the HeLa cell line (Hart et al., 2015). Unfortunately, the assessment of the new biomass reaction on the growth rate prediction could not be done due to the lack of exometabolomics data for the HeLa cell line on the Zielinski et al. (2017) dataset.