# Integration of metabolomics with genomics

## Metabolic gene prioritization using metabolomics data and genomic variant (CADD) scores

Bongaerts, Michiel ; Bonte, Ramon ; Demirdas, Serwet; Huidekoper, Hidde H. ; Langendonk, Janneke ; Wilke, Martina ; de Valk, Walter; Blom, Henk J. ; Reinders, Marcel J.T.; Ruijter, George J.G.

**DOI**
10.1016/j.ymgme.2022.05.002

**Publication date**
2022

**Document Version**
Final published version

**Published in**
Molecular Genetics and Metabolism

Molecular Genetics and Metabolism 136 (2022) 199–218

Contents lists available at ScienceDirect

Molecular Genetics and Metabolism

journal homepage: www.elsevier.com/locate/ymgme

# Integration of metabolomics with genomics: Metabolic gene prioritization using metabolomics data and genomic variant (CADD) scores

Michiel Bongaerts [a,*], Ramon Bonte [a], Serwet Demirdas [a], Hidde H. Huidekoper [b], Janneke Langendonk [c], Martina Wilke [a], Walter de Valk [a], Henk J. Blom [a], Marcel J.T. Reinders [d], George J.G. Ruijter [a,*]

[a] Department of Clinical Genetics, University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015, GD, Rotterdam, the Netherlands
[b] Department of Pediatrics, Center for Lysosomal and Metabolic Diseases, University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015, GD, Rotterdam, the Netherlands
[c] Department of Internal Medicine, Center for Lysosomal and Metabolic Diseases, University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015, GD, Rotterdam, the Netherlands
[d] Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft, Van Mourik Broekmanweg 6, 2628, XE, Delft, the Netherlands

ARTICLE INFO

*Article history:*
Received 16 August 2021
Received in revised form 6 April 2022
Accepted 17 May 2022
Available online 25 May 2022

*Keywords:*
Inborn errors of metabolism
ES
Untargeted metabolomics
Data integration
Metabolic pathways
CADD scores

ABSTRACT

The integration of metabolomics data with sequencing data is a key step towards improving the diagnostic process for finding the disease-causing genetic variant(s) in patients suspected of having an inborn error of metabolism (IEM). The measured metabolite levels could provide additional phenotypical evidence to elucidate the degree of pathogenicity for variants found in genes associated with metabolic processes. We present a computational approach, called *Reafect*, that calculates for each reaction in a metabolic pathway a score indicating whether that reaction is deficient or not. When calculating this score, *Reafect* takes multiple factors into account: the magnitude and sign of alterations in the metabolite levels, the reaction distances between metabolites and reactions in the pathway, and the biochemical directionality of the reactions. We applied *Reafect* to untargeted metabolomics data of 72 patient samples with a known IEM and found that in 81% of the cases the correct deficient enzyme was ranked within the top 5% of all considered enzyme deficiencies. Next, we integrated *Reafect* with Combined Annotation Dependent Depletion (CADD) scores (a measure for gene variant deleteriousness) and ranked the metabolic genes of 27 IEM patients. We observed that this integrated approach significantly improved the prioritization of the genes containing the disease-causing variant when compared with the two approaches individually. For 15/27 IEM patients the correct affected gene was ranked within the top 0.25% of the set of potentially affected genes. Together, our findings suggest that metabolomics data improves the identification of affected genes in patients suffering from IEM.



## 1. Introduction

DNA sequencing methods such as exome sequencing (ES) and whole genome sequencing (WGS) are powerful techniques to identify the pathogenic genetic variant(s) in patients suspected of a genetic disease [1–3]. Nevertheless, ES performed on a single person typically generates tens of thousands of variants [2]. With the reduced costs for sequencing, WGS becomes increasingly popular, generating even a few million of variants per patient [2]. Numerous filtering strategies have been developed to reduce the number of variants which needs human inspection. The Combined Annotation Dependent Depletion (CADD) score is widely explored as one of these filtering strategies [4]; prioritizing variants such as single nucleotide variants (SNV), deletions and insertions (InDels). CADD scores employ a machine learning based approach where 63 conservational - and functional genomic metrics are combined into a single metric. After various filtering steps, the investigator still needs to evaluate a substantial number of variants manually. The pathogenicity of these rare or novel variants is often unknown, leading to a clinically dissatisfactory classification.

Functional studies may provide evidence whether a variant of unknown significance should be considered pathogenic or not. For this purpose, metabolomics is catching more and more interest since it has the potential to resolve the degree of pathogenicity for genetic variants which are expected to have a deleterious effect on the patient's metabolism, i.e. inborn errors of metabolism (IEM) [5–7]. To integrate metabolomics and genomics, metabolomics results need to be interpreted to link metabolite abnormalities to potentially deficient reactions/enzymes and their corresponding genes. Some strategies have already

* Corresponding authors.
*E-mail addresses:* m.bongaerts@erasmusmc.nl (M. Bongaerts), g.ruijter@erasmusmc.nl (G.J.G. Ruijter).

https://doi.org/10.1016/j.ymgme.2022.05.002

been developed for this purpose. Haijes et al. applied expert knowledge to develop an algorithm that matches metabolic signatures obtained from metabolomics with expected metabolic signatures caused by each IEM, thereby ranking potential enzymatic deficiencies [8]. This approach, however, requires each IEM to be added manually and optimized individually, involving many different parameters. Alternatively, Baumgartner et al. explored the use of classification algorithms to distinguish multiple IEM based on differences in metabolite levels [9], which does not require manual parameter optimization. However, training such a classifier requires data from multiple patients having the same IEM. Since more than a 1000 different IEM exist with an overall birth prevalence of 51 per 100.000 [10] the creation of a proper training set is not likely to succeed. To overcome this limitation, Messa et al. explored the use of metabolic networks to simulate IEM specific metabolic profiles, which they then compared with real IEM profiles using a Siamese neural network to rank the most probable matching simulated IEM [11]. However, the successful detection of a certain IEM depends on the resemblance of the simulated IEM profiles with real IEM profiles. Furthermore, we argue that pathway information combined with (real) metabolomics profiles is sufficient to rank IEM, thereby removing the need for simulated IEM profiles. Another strategy involves the use of gene-metabolite sets for which an enrichment score can be calculated to rank potential affected genes [5]. Similarly, *metPropagate* uses gene-metabolite set enrichment scores, but additionally propagates these scores through a protein-protein network to rank potentially affected genes [7]. The main concern with these approaches is that enrichment scores require (*Z*-score) cutoffs for metabolite levels, potentially excluding subtle aberrations that do not exceed the thresholds. Furthermore, gene-metabolite set approaches neglect the signature where substrates normally catalyzed by the deficient enzyme increase in concentration, whereas related products decrease in concentration.

Two other methods that uses metabolomics data to detect potentially affected pathways/module were developed by Li et al. (*Mummichog*) [12] and Pirhaji et al. (*PIUMet*) [13]. *Mummichog* automatically annotates mass spectrometry (MS) features while inferring molecular pathways and modules that have increased 'activity' related to a set of significantly altered MS features. Similarly, *PIUMet* also automatically annotates MS features, infers dysregulated molecular pathways, but additionally scores metabolites and proteins to reflect their importance. Both *Mummichog* and *PIUMet* were developed as general tools to infer dysregulated molecular pathways/modules, but were not designed to rank/detect IEM. Moreover, the use of low confident feature annotations is doubtful in a clinical setting.

To integrate metabolomics data with genomic variant scores obtained from ES, or potentially WGS, we developed an algorithm, called *Reafect* (**Rea**ction de**fect**). *Reafect* combines information of metabolic pathways from KEGG [14] and the metabolite *Z*-scores obtained from annotated metabolomics data to calculate a 'deficient reaction score' for each reaction. Higher scores imply that there is more evidence of that reaction being deficient and vice versa. Our algorithm differs fundamentally from the approaches mentioned earlier, since 1) *Reafect* uses solely existing pathway information, thereby removing the need for manual addition of each IEM, 2) does not rely on the availability of a large metabolomics dataset containing multiple IEM patients, 3) contains only 3 parameters that need to be optimized, 4) uses the *Z*-scores in a continuous fashion without using cutoff values, and 5) explicitly takes the general metabolic signature of an IEM into account when searching for the most probable IEM. We evaluated *Reafect*'s performance on 36 distinct IEM using 72 plasma samples from patients diagnosed with an IEM.

Since each reaction is associated with genes coding for the enzyme catalyzing that reaction, we used *Reafect*'s deficient reaction scores in combination with CADD scores as an integrated model for prioritizing potentially affected genes. To evaluate this approach, we studied 27 IEM patients for which the pathogenic variant was identified and untargeted metabolomics data was obtained. This integrated model showed a significant improvement on ranking the correct affected genes when compared with using solely *Reafect* or CADD scores.

## 2. Material and methods

### 2.1. Reafect

An enzymatic deficiency generally leads to a build-up of the reaction substrate(s) and shortage of the product(s) formed by that reaction. *Z*-scores obtained from annotated metabolomics can be used to detect the accumulation of these substrates (i.e. positive Z-scores) as well as shortages of the products (i.e. negative Z-scores). Although the accumulation and shortage of metabolites occur for the metabolites directly involved in the deficient reaction, aberrant metabolite levels will also propagate through a biochemical pathway, leading to changes in metabolite levels that are multiple reaction steps away from the deficient reaction. We used this dogma to develop an algorithm, called *Reafect*, that calculates for each reaction in a pathway a score that reflects how deficient that reaction is. We called this score the 'deficient reaction score' or $S_R$ score. To calculate this score, *Reafect* weighs metabolite levels (*Z*-scores) which are further away from the considered reaction to a lesser extent than metabolite levels closer to the putative reaction, since we assume that more distant metabolites give less information about the reaction deficiency. For this purpose, *Reafect* uses a weighted version of the observed *Z*-scores, called 'effective Z-scores', and which are always relative to the considered reaction for which the deficient reaction score is calculated (see Fig. 1 and Eq. 1). The effective *Z*-score is determined by calculating a total decay factor over a reaction path when going from the metabolite (with *Z*-score) to that reaction. The more steps away from the considered reaction, the more the observed Z-score is decayed, thereby resulting in a lower absolute effective *Z*-score. To constrain the number of model parameters, we used three different decay factors ($a$, $b$, $c$) and distinguished five different decay types (see Eq. 2): 1) a decay factor $a$ for a metabolite with a positive *Z*-score taking a step downstream towards the considered reaction, 2) a decay factor $b$ for a metabolite with a positive Z-score taking a step upstream towards the considered reaction, 3) a decay factor $a$ for a metabolite with a negative Z-score taking a step upstream, 4) a decay factor $b$ for a metabolite with a negative *Z*-score taking a step downstream and 5) a decay factor $c$ for reversible reactions (independent of the *Z*-score sign) taking one step in the direction of the considered reaction. We want to emphasize that *Reafect* describes reaction paths as a chain of metabolite and reaction nodes (in a graph) to track all pathway information (see Fig. 1). Consequently, a reaction step is either a step from metabolite to reaction, or from reaction to metabolite. For example, consider a metabolite with a positive *Z*-score which takes three downstream steps to get to the considered reaction (Fig. 1a, $m_2$ to $R_3$). The effective Z-score for this metabolite would then be given by the *Z*-score multiplied by $a^3$, thus having a total decay factor of $a^3$. Similarly, if this metabolite had a negative *Z*-score the total decay factor for this reaction path would have been $b^3$. Obviously, a reaction path could also be more complex, resulting for example in a total decay factor of $c^2ba^2$. We justify the introduction of $a$ and $b$, by realizing that when $a > b$ the effective *Z*-scores remain relatively high for positive Z-scores located upstream of a deficiency, and the same holds for negative Z-scores downstream of the deficiency. The values of these decay factors ($a$, $b$ and $c$) are selected using the metabolomics data from 72 IEM patient samples (see Section *Tuning the model parameters*). Subsequently, *Reafect* aggregates all effective *Z*-scores resulting in the deficient reaction score (or $S_R$ score) where it takes into account whether a certain effective *Z*-score was located downstream or upstream of the considered reaction (Eq. 5).

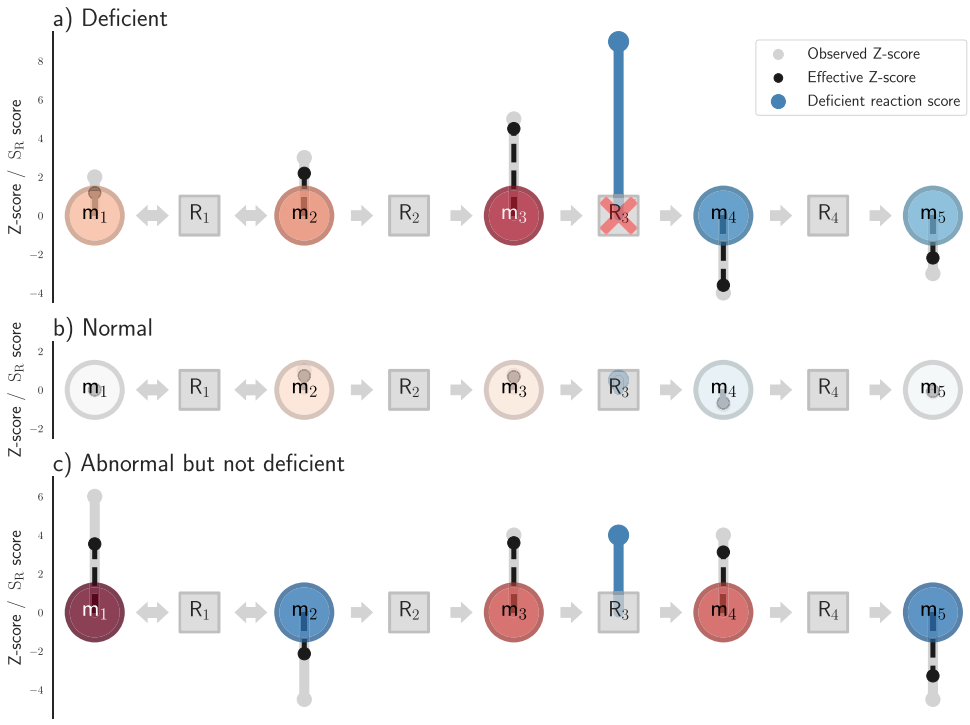


**Fig. 1.** Illustration of *Reafect*. A circle indicates a metabolite and a square a reaction (node), with the horizontal arrows indicating the directionality of the reaction. The vertical grey bars (with dot) indicate the observed *Z*-scores; pointing upwards indicating a positive Z-score and vice versa. The black dotted bars indicate the effective *Z*-score from the perspective of reaction $R_3$. Note that *Reafect* determines for each reaction a deficient reaction score but in these figures only the results are shown for $R_3$. **A)** Reaction $R_3$ is deficient. The effective *Z*-scores decay when going away from $R_3$ as visualized by the reduced magnitude of the black bars. The deficient reaction score, illustrated by the blue bar on $R_3$, is high since we observe net positive effective *Z*-scores upstream of $R_3$ and net negative effective *Z*-scores downstream of $R_3$. **B)**$R_3$ is not deficient and metabolite Z-scores around the reaction are normal, thereby resulting in a low deficient reaction score. Note that the blue bar at $R_3$ is small. **C)**$R_3$ is not deficient, but has still a relatively high deficient reaction score. Note that although the observed *Z*-scores for $m_3$ and $m_4$ are equal, the resulting effective *Z*-scores are different since the decay of the Z-scores also depends on the biochemical directionality (and also applies to $m_2$ and $m_5$). Metabolite $m_1$ has a relatively high observed *Z*-score, but its effective Z-scores is reduced since it is 5 reaction steps away from $R_3$. *Reafect* calculates per side of the reaction the net effective Z-scores. For example, the effective Z-scores for $m_4$ and $m_5$ roughly counter balance each other when looking at the downstream side of $R_3$. The upstream side has net positive effective *Z*-scores, therefore resulting in a positive deficient reaction score.

To calculate the deficient reaction score for a certain reaction, we first consider the decay of the *Z*-score over a path *p* leading for metabolite *m* to reaction *R*:

$$E_{m,R,p}(Z_m) = \left(\prod_{s}^{\text{steps}\in p} \gamma_s(Z_m, D_s)\right)|Z_m| \tag{1}$$

Here, $E_{m,R,p}(Z_m)$ is the effective Z-score for metabolite *m* from the perspective of reaction *R* along reaction path *p*. $\gamma_s(Z_m,D_s)$ is the decay factor for step *s* and depends on the biochemical directionality of the step ($D_s$) (upstream, downstream, reversible) and the sign of the *Z*-score ($Z_m$):

$$\gamma_s(Z_m, D_s) = \begin{cases} a & \text{if } \operatorname{sgn}(Z_m) = 1 \text{and } D_s = \text{downstream} \\ a & \text{if } \operatorname{sgn}(Z_m) = -1 \text{and } D_s = \text{upstream} \\ b & \text{if } \operatorname{sgn}(Z_m) = 1 \text{and } D_s = \text{upstream} \\ b & \text{if } \operatorname{sgn}(Z_m) = -1 \text{and } D_s = \text{downstream} \\ c & \text{if } \operatorname{sgn}(Z_m) = 1 \text{and } D_s = \text{reversible} \\ c & \text{if } \operatorname{sgn}(Z_m) = -1 \text{and } D_s = \text{reversible} \end{cases} \tag{2}$$

Since more paths (*p*'s) could be possible between metabolite *m* and reaction *R*, and these could have different lengths, we calculated a normalized effective *Z*-score for every path:

$$\tilde{E}_{m,R,p} = \frac{\left[E_{m,R,p}(Z_m)\right]^2}{\sum_{p'} |E_{m,R,p'}(Z_m)|} \tag{3}$$

where $\tilde{E}_{m,R,p}$ is the normalized effective Z-score for path *p*. The summation over $p'$ indicates all paths leading from *m* to *R*. In this way, paths originating from *m* with (relatively) low effective *Z*-score strengths (such as longer paths) are weighted less in the normalized effective Z-score whereas short paths get more weight since their effective *Z*-score is relatively large (when compared to the other paths). All paths (*p*'s) were determined by constructing an 'ego graph' around each metabolite, selecting a subset of neighbouring metabolites and reactions around this central metabolite. To reduce computational cost, we set a limit of 15 reaction steps (metabolite-reaction or reaction-metabolite) around this ego graph, and a maximum of 10 paths for travelling from *m* to *R*.

Next, we summed all normalized effective *Z*-scores but we made a distinction between normalized effective Z-scores where its path is connected to the upstream or the downstream side of reaction *R*. For clarity, let us consider a direct substrate *m* of reaction *R*, which has a direct connection at the upstream side of the reaction. Let us also assume that there is a path going from *m*, via other reactions, which ends at the downstream side of the reaction. Since we have two paths, we have two normalized effective Z-scores; one belonging to the direct connection, the other belonging to the longer path. Since the latter path is longer, its normalized effective Z-score will be less than the normalized effective Z-score of the direct connection (Eq. 3). We aggregated all metabolite normalized effective Z-scores based on the Z-score sign and connection to the reaction (downstream or upstream):

$$E^R_{x,y} = \sum_{m \in \Omega^R_x} \sum_{p \in \Omega^R_y} \widetilde{E}_{m,R,p} \tag{4}$$

with x ∈ {positive Z − score, negative Z − score} indicates the set of metabolites having a Z-score sign equal to x and y ∈ {downstream, upstream} indicates the set of paths from *m* to *R* which are connected to the y-side of reaction *R* (i.e. downstream or upstream side). Since reversible reactions lack a clear defined up – and downstream side, we assigned one of each side to the up – or downstream side while making sure that product/substrate information was conserved.

At last, we defined the deficient reaction score for reaction *R* as:

$$S_R = \begin{cases} \left(E^R_{+,up} - E^R_{-,up}\right) + \left(E^R_{-,down} - E^R_{+,down}\right) & \text{if } R \text{ irrev} \\ \left|\left(E^R_{+,up} - E^R_{-,up}\right) + \left(E^R_{-,down} - E^R_{+,down}\right)\right| & \text{if } R \text{ rev} \end{cases} \tag{5}$$

where we replaced 'positive Z-score' and 'negative Z-score' for the symbol '+' and '-', respectively. We replaced 'downstream' and 'upstream' for 'down' and 'up', respectively. We observe that $S_R$ increases for net positive normalized effective Z-scores located at the upstream side of the reaction and for net negative normalized effective Z-scores located at the downstream side of the reaction, while $S_R$ decreases for the opposite cases. When a reaction is reversible we decided to take the absolute value, arguing that we are interested in an imbalance of the net positive and negative normalized effective Z-scores across the reaction regardless of which side of the reaction these normalized effective Z-scores were positioned.

Finally, *Reafect* prioritizes all reactions by sorting the $S_R$ scores on their magnitude, with higher scores indicating that a reaction is more likely to be deficient. Next to prioritizing the reactions, *Reafect* can prioritize enzymes and corresponding genes on their potential of being deficient. As enzymes can be involved in multiple reactions the final $S_R$ score for an enzyme is taken to be the maximum $S_R$ score of the set of reactions the enzyme may catalyze (Method). Furthermore, we need to realize that some enzymes can catalyze the same reaction (s). In this study we dealt with this issue by taking the maximum occurring $S_R$ score for each enzyme, even if that same $S_R$ score was already assigned to another enzyme.

### 2.2. Metabolomics data

#### 2.2.1. In-house dataset

Metabolomics data was obtained as described by Bonte et al. [15]. Samples obtained from IEM patients were measured in 20 separate batches. Features were annotated using an in-house database with retention times of each metabolite and by matching data dependent MS/MS spectra with an in-house MS/MS database [15]. For the 72 patient samples, a median of 119 annotated metabolites was obtained (when combining positive – and negative ion mode), and a minimum of 95 annotated metabolites was available for each sample. The complete list of annotated metabolites can be found at the Github repository of *Reafect*: https://github.com/mbongaerts/Reafect/. Note, that some rare metabolites were also measured and annotated. Two samples were obtained from the same individual; a patient with isovaleric aciduria (S21, S71) and a patient with citrullinemia (S04, S62). Thus, 70 unique IEM patients were included in this study. From the 72 IEM samples investigated, 51 were from patients on specific treatment, while 19 were not treated at the time of sample collection (see Appendix A1). For only 28/70 IEM patients the pathogenic variant was identified using either Sanger sequencing, a SNParray or ES (see Appendix A2 for more details). In agreement with national legislation and institutional guidelines, all patients or their guardians approved the possible de-identified use of the remainder of their samples for method validation and research purposes. The study was conducted in accordance with the Declaration of Helsinki.

#### 2.2.2. Z-score calculation

Z-scores were calculated using two different approaches. Metabolites that were annotated in at least 7 batches were merged, a Box-Cox transform was applied and normalized using *Metchalizer* [16]. Z-scores were determined using a regression model with age and sex as covariates [16]. For metabolites that were annotated in less than 7 batches, the Z-scores were determined from 15 within-batch samples, where abundancies were first normalized using Probabilistic Quotient Normalization [17] and Box-Cox transformed. These 15 samples could originate from controls or (un)diagnosed patients. We used to following procedure to prevent that outlier samples were used as reference for the Z-score calculation:

1. Calculated Z-scores for the selected samples.
2. Keep the samples with a |Z-score| < 3.
3. Repeat step 1. and 2. five times.
4. Determine Z-scores based on the mean and standard deviation of 15 (random) samples from the remaining samples (step 3).

When a metabolite was annotated in both positive- and negative ion mode, the Z-score of the ion mode with the largest median abundancy (over all samples) was taken. Since three technical replicates were measured for all patient samples, we used the average of these three Z-scores as the final Z-score.

Since the deficient reaction scores (Eq. 5) aggregate multiple metabolite Z-scores, we do not wish a single extreme Z-score to dominate this measure. To prevent this, we used the following Z-score transformation to down scale extreme Z-scores:

$$\widetilde{Z} = \text{sign}(Z)\left(\alpha(Z)\,|Z|^{0.75} + [1 - \alpha(Z)]|Z|\right) \tag{6}$$

with

$$\alpha(Z) = \frac{1}{1 + \exp\,(2 - |Z|)} \tag{7}$$

This transform behaves approximately linear for the region $0 < |Z| < 2$, but scales down Z-scores when $|Z| >> 2$ (Fig. 2).

#### 2.2.3. Miller dataset

To evaluate our approach we used a second metabolomics dataset that was published by Miller et al. [18]. This dataset is available via https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4626538/. Z-score transform Eq. 6 was not applied to this dataset since all metabolite Z-scores were considered to be within an acceptable range.

### 2.3. Retrieving human metabolic reactions

We used the KGML parser from https://github.com/biopython/biopython (20−03−2020) to process KEGG [14] pathways and modules, where we filtered on reactions involved in humans (using the hsa pre-fix). When retrieving the KEGG networks, some reactions were associated with more than one enzyme, for which KEGG returns the same unique reaction as many times as it is associated with the different enzymes, leading to a multiplicity for these reactions. We removed this multiplicity but we remained all the associated enzymes with this reaction. In other words, in these cases the same $S_R$ score for that reaction was assigned to all associated enzymes.

To increase the overlap between the metabolites measured in plasma and metabolites in the pathways/modules (from KEGG), we manually added some reactions. These can be found in Appendix A3. Most of these reactions were obtained from Recon / Virtual Metabolic Human [19].

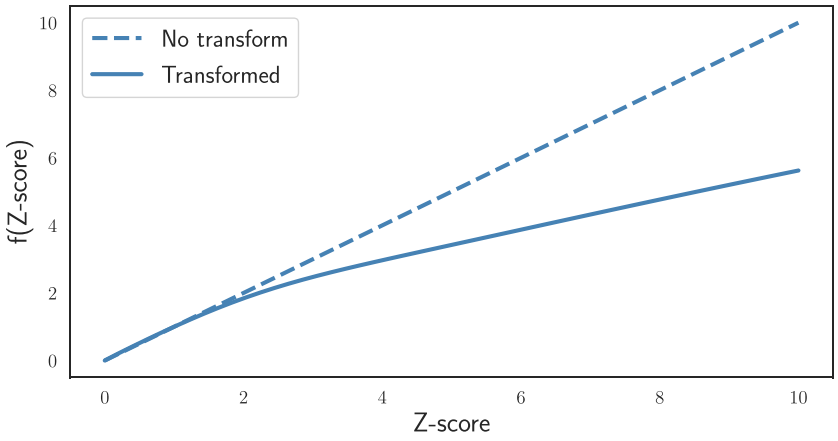


**Fig. 2.** The effect of the transformation given by Eq. 6 on the Z-scores.

### 2.4. Overall performance of Reafect using bootstrapped AUC

Annotation of metabolites in the metabolomics data was performed per batch, which resulted in an unequal number of annotations per batch. This difference also affected the number of unique enzymes on which ranking was based per patient (see Fig. 5, *Total number of enzymes*). To correct for this, we expressed the (absolute) rank in as a percentile by dividing by the total number of enzymes multiplied by 100%. The overall performance of *Reafect* for a certain choice of ($a$, $b$, $c$) was measured by displaying the percentage (vertical axis) of the IEM patients having the percentile rank of the correct IEM lower than a predefined value (horizontal axis). While increasing this value we obtained a curve, where the area under the curve (AUC) gives a measure for the overall performance. Higher AUC indicates that a larger percentage of the IEM patients have a lower rank (steeper increase of the curve). We used a bootstrap procedure where we selected 1000 times a random 75% of the total IEM patients for which we calculated the AUC. By taking the 50th percentile of these 1000 AUCs we obtained a more robust overall performance for each ($a$, $b$, $c$).

### 2.5. MetPropagate and comparison with Reafect

We downloaded the weighted STRING network (v11) from https://github.com/emmagraham/metPropagate (07-08-2020). ME scores were calculated in the exact same manner as described by Linck et al. Using the same terminology, metabolites having $|Z - \text{score}| > 1.5$ where considered as 'differentially abundant metabolites'. ME scores were propagated using the Local and Global Consistency (LGC) algorithm with settings max_iter = 30 and alpha = 0.99.

To objectively compare *Reafect* with *metPropagate* we took several factors into account:

1. Only metabolites were included with (HMDB) identifiers in the pathways/modules used by *Reafect* and which were also present in the gene-metabolite sets used by *metPropagate*.
2. Before determining ranks, the propagated ME scores for every gene were assigned to the associated enzyme(s). We removed genes (and thus enzymes) which did not overlap in the output of both algorithms. Thus, both outputs contained the exact same number of unique enzymes on which ranking was performed.
3. The ranks for *metPropagate* were calculated using the propagated ME scores on the enzyme level. Note that we took the maximum propagated ME score for an enzyme when more genes were associated with that enzyme. Similarly, the ranks for *Reafect* were determined from the $S_R$ scores (as described above).

### 2.6. ES data

Exome Sequencing (ES) data was acquired over a longer time period (2013—2021), and was performed either using the Agilent Clinical Research Exome V1 (sureselect SSCRE V1) or Agilent Clinical Research Exome V2 (sureselect SSCRE V2) on a Illumina NovaSeq sequencer using paired-end reads with a read-length of 150 bp. Reads were aligned to human reference genome build GRCh37 hg19 (ucsc.hg19.nohap.fasta) using the BWA alignment algorithm [20]. The VCF-files were obtained using GATK3 [21] and ANNOVAR was used to annotate gene names and variants [22]. All patients included in this study from which ES data was used gave written consent for de-identified use of their data for research purposes.

### 2.7. CADD scores and gene prioritization

Variants called by GATK3 were annotated with CADD scores from Genome build GRCh37/ hg19 v1.6 (https://cadd.gs.washington.edu/download) for both SNVs and InDels. In this study we used the CADD (Phred) scores in two manners: 1) ranking genes based solely on the maximum CADD score occurring in each gene and 2) ranking genes using the deficient reaction score ($S_R$ score) from *Reafect* combined with the CADD scores. Note, that only genes were included in this ranking for which a $S_R$ score was determined and which were present in the ES data.

Gene ranking using *Reafect* in combination with CADD scores was done as follow:

1. Per enzyme the maximum $S_R$ score was determined for all associated reactions (KEGG). For each enzyme, all associated genes (KEGG) were determined and the same maximum $S_R$ score was assigned to these genes.
2. The maximum CADD (Phred) score per gene was determined.
3. The $S_R$ score was multiplied with the CADD score (Phred) for each gene.
4. Genes were ranked on their integrated score.

For a subset of the IEM patients included in this study the disease-causing variant was identified either using exome sequencing (ES), Sanger sequencing or using an SNP array. Since ES data was not available for most IEM patients where the disease-causing variant(s) is identified, we assumed that we could include these patients using 15 random ES backgrounds while inserting the known disease-causing variant in each background. Consequently, we obtained 15 different rankings for each affected gene. We assumed that the average of these

15 rankings is a good estimate of the rank when a real ES background was used (Discussion).

### 2.8. Excluded IEM patients

Although some IEM patients were initially measured they were not included in this study, which had two main reasons. First, in some cases there was no reaction in a known (KEGG) pathway relating (measured) metabolites to the IEM. Because of this, we left out a patient with a mutation in the *MMACHC* gene, one with a mutation in the *MOCS3* gene and two patients with glutaric acidemia type 2 (*ETFDH*, *ETFA*, *ETFB* genes). Secondly, since *Reafect* does not make a distinction between different compartments within the body or cell, the inclusion of enzymatic deficiencies related to transport proteins is complicated. In these transport reactions the metabolite itself does not change, only its location changes, and therefore build-up of these metabolites are expected only in certain parts of the body or cell. For this reason, we were not able to include a few patients with lysinuric protein intolerance (*SLC7A7* gene), and a patient with organic cation transporter 2 deficiency (*SLC22A5* gene).

From the Miller dataset we excluded all patients having lysinuric protein intolerance, cobalamin biosynthesis deficiency, glutaric aciduria type 2, 3, 4, 5 for the same reasons mentioned earlier.

## 3. Results

### 3.1. Tuning the model parameters

Per IEM patient, potential deficient enzymes were ranked by their maximum associated $S_R$ score (Methods) and the rank of the true deficient enzyme in that patient was reported. Since the total number of enzymes on which the ranking was based varied among the patients, we determined the percentile rank (PR) by dividing by the total number of enzymes multiplied by 100% (Methods). A lower PR indicates an improved ranking performance and vice versa. The overall performance of *Reafect* was measured by calculating how often a PR was smaller or equal than a predefined value across the 72 IEM patient samples. When increasing this predefined value a curve is generated as displayed in Fig. 3. We used the area under this curve (AUC) to indicate the overall performance of *Reafect*, where higher AUCs imply better performances.

Since *Reafect* uses three model parameters ($a$, $b$, $c$), we used a parameter sweep over these parameters to explore how the performance (AUC) was affected. We performed a bootstrap procedure to obtain a robust performance AUC (Methods). Fig. 4 shows these bootstrapped AUCs for each combination of ($a$, $b$, $c$). For region $b > a$, *Reafect* performs less than for region $b < a$. This can be understood by realizing that when $a > b$ the effective $Z$-scores for metabolites having positive Z-scores decay faster for downstream steps than for upstream steps (and the opposite for negative $Z$-scores), resulting in reduced evidence for the deficient reaction. Furthermore, for region $c < 0.5$, *Reafect*'s overall performance is poor. The highest performance was reached for $a = 0.85$, $b = 0.35$, and $c = 0.75$ (see Fig. 4B). In further evaluations of *Reafect*, we set the parameters $a$, $b$, $c$ to these values.

### 3.2. Enzyme ranking for IEM patients

We applied *Reafect* to 72 IEM patient samples and determined the percentile rank (PR) of the true enzyme deficiency. For 62% of these samples the PR was within the top 2.5% of all considered enzyme deficiencies, and for 81% of the samples the PR was within the top 5% (Fig. 3B). Evaluating *Reafect* on 106 IEM patients from the Miller dataset resulted in similar performance, where about 59% and 74% of the patients was ranked within the top 2.5% and 5% respectively (Fig. 3D). The ranks per IEM patient can be found in Appendix A5.

Additionally, we compared *Reafect* with *metPropagate* [7], while taking several factors into account such as overlapping metabolites and genes between the two approaches to objectively compare the performances (Methods). We found that *Reafect* has a 21% increase in the AUC when compared to *metPropagate*. Considering that lower percentile ranks (<10%) are more interesting (Fig. 3B), we observe that for this region the partial AUC of *Reafect* is 71% higher than the partial AUC of *metPropagate*. A detailed overview of the PRs per IEM patient for both approaches can be found in Appendix A4. A comparison using the Miller dataset resulted in similar differences, with *Reafect*'s AUC having an 18% increase over *metPropagate*'s AUC. For the partial AUC this increase is 60%.

Fig. 5 shows a detailed overview of the results per IEM patient sample in our dataset. From this figure it is clear that for the same IEM but different patients, *Reafect* can return different PRs. For example, one patient with maple syrup urine disease (*BCKDH(A)(B)/DBT* genes) has a PR of 0.55%, whereas for the other patient this is 4.29%. This can be explained by the difference in the magnitude of the $Z$-scores for the disease-related metabolites leucine and isoleucine, namely for the patient with the low rank $Z = 6.7$ and $Z = 5.4$, respectively, and for the patient with the higher rank these Z-scores were less extreme, $Z = 2.49$ and $Z = 2.65$, respectively. Similarly, for two patients having long-chain-3-hydroxyacyl-CoA dehydrogenase deficiency (*HADHA* gene), one has a PR of 0.55% and for the other this is 1.66%. Again, this difference in PRs can be understood by differences in for example 3-hydroxyhexadecanoylcarnitine, which had a Z-score of $Z = 13.3$ for the patient with the lower PR, while the other was more subtle with $Z = 6.4$. Also, one patient with carbamoyl phosphate synthetase I deficiency (*CPS1* gene) ranked at 1.23%, had $Z = 1.8$ for L-glutamine, while the other patient (ranked at 10.89%) seemed to have a normal L-glutamine level ($Z = 0.2$), thereby explaining also the difference between these ranks.

Some IEM were poorly ranked due to the absence of clear aberrations in the metabolomics data. For both patients with alkaptonuria (homogentisate 1,2-dioxygenase deficiency, *HGD* gene), homogentisic acid was not increased in our analysis ($Z = 0.4$ and $Z = 0.5$), which clarifies why *Reafect* poorly ranked these patients. The patient with mevalonate kinase deficiency (*MVK* gene) was also ranked poorly, which was a consequence of two reasons: 1) only one metabolite involved in calculating the $S_R$ score i.e. mevalonic acid, was annotated in the metabolomics data and 2) the Z-score of this metabolite was $Z = 0.7$.

*Reafect* ranked the patient with arginase I deficiency (*ARG1* gene) at 0.36%. This was considered to be a relatively good ranking, since 14 metabolites were found to have a Z-score above 2.1, while the disease related metabolites arginine and ornithine had $Z = 2.1$ and $Z = -2.4$ respectively. From a naive perspective we would expect about 14 other enzyme deficiencies to have lower (better) ranks than arginase I. However, this relatively good performance can be explained by the fact that arginase I catalyzes the conversion of arginine into ornithine (plus urea), and the substrate (arginine) is increased while the product (ornithine) is reduced. Consequently, *Reafect* assigned a relatively high $S_R$ score to this reaction (see Eq. 5).

Another interesting observation is the poor rank obtained for the patient having guanidinoacetate *N*-methyltransferase deficiency (*GAMT* gene). This patient was under treatment with creatine supplementation, which explains the poor rank. Although guanidinoacetate ($Z = 3.1$) was high in this patient, the presence of the high creatine level ($Z = 6.7$) led to high Z-scores on both sides of the guanidinoacetate *N*-methyltransferase reaction R01883 which reduces the $S_R$ score, as can be observed in Eq. 5 (Methods). Similarly, we observed that the patients with guanidinoacetate *N*-methyltransferase deficiency in the Miller dataset also have high ranks (Appendix A5). This can be explained by the relatively high concentrations of creatine (likely due to treatment) in these patients and the fact that guanidinoacetate was not measured in this dataset.

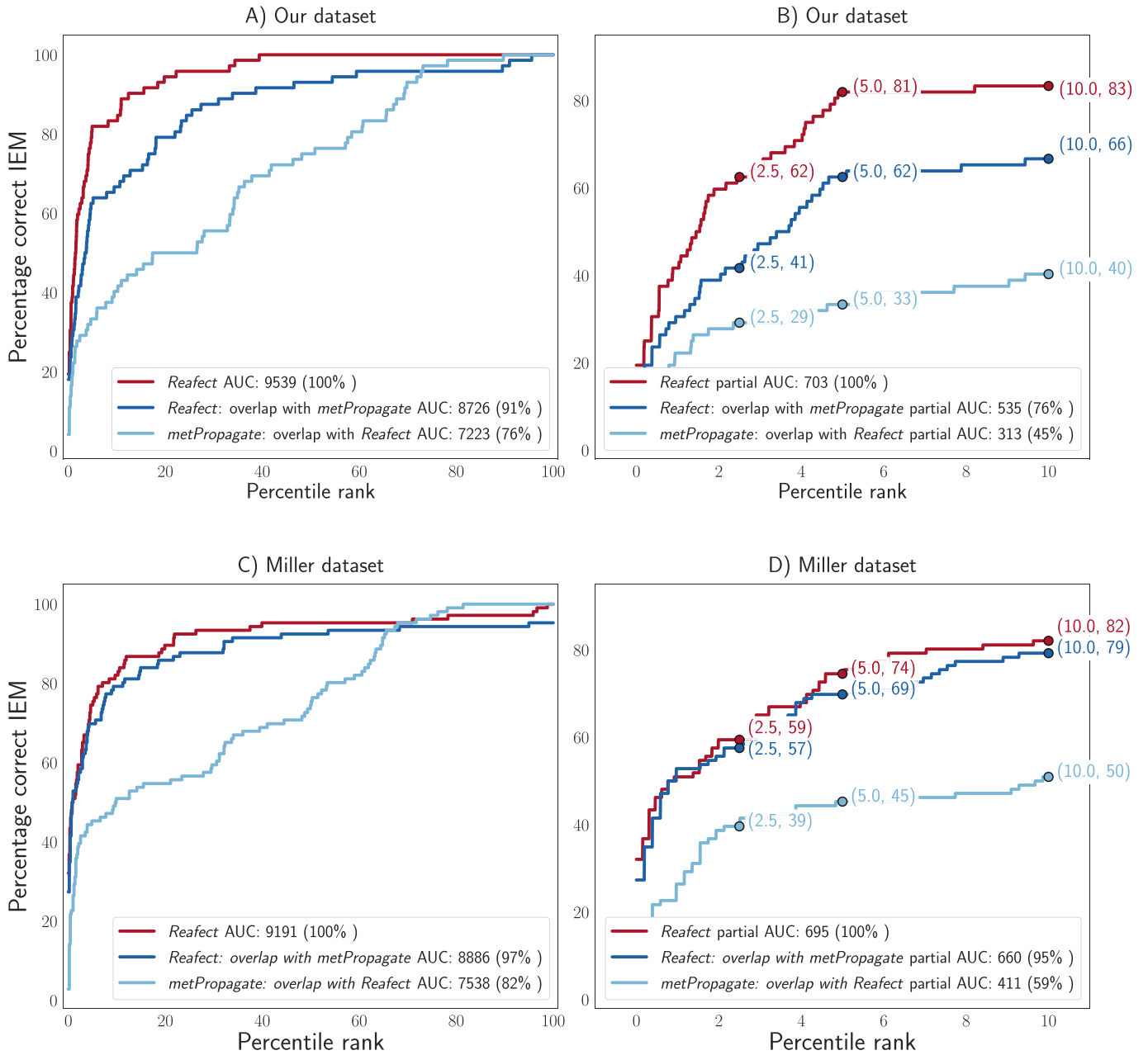


**Fig. 3.** IEM ranking performances for different approaches as indicated by the legend. Each curve shows the percentage of IEM patient samples for which the percentile rank (PR) of the true enzyme deficiency is within the top x% (horizontal axis) of all considered enzyme deficiencies. Model settings for *Reafect*: $a = 0.85$, $b = 0.35$ and $c = 0.75$. **A)** Full performance curves using our datatset. **B)** Performance curves with PR $<= 10\%$ using our dataset. **C)** Full performance curves using the Miller datatset. **D)** Performance curves with PR $<= 10\%$ using the Miller dataset. To perform a meaningful comparison between *Reafect* and *metPropagate* a subset of the data was analyzed that contained only metabolites and genes that were included in both approaches (Methods). Note that this selection reduced the performance of *Reafect* to 91% and 97% of its original performance on our dataset and the Miller dataset respectively.

### *3.3. Gene prioritization for IEM patients using CADD scores and Reafect*

We hypothesized that potentially affected (metabolic) genes could be better prioritized when we combine the CADD (Phred) scores obtained from gene variants in ES data with the deficient reaction scores obtained from *Reafect*. Since an increase in both scores is expected to be associated with increased pathogenicity we chose to multiply the deficient reaction score with the maximum CADD score observed in the variants of the gene corresponding to that enzyme. Next, we used this combined score to rank the genes (Methods).

Since ES data was only available for two IEM patients, we evaluated this gene ranking based on two approaches: 1) using the ES background belonging to that patient if the ES data was available (see asterisks in Table 1) and 2) using 15 random ES backgrounds while inserting the (known) disease-causing variant of the patient (Methods). Table 1 shows the PRs for 28 IEM patients for which the pathogenic variant was identified, using solely *Reafect*, solely *CADD* as well as the integrated approach. For 12/28 patients *Reafect* scored better than *CADD* (marked blue). For 21/28 and 20/28 patients, the integrated approach led to improved ranking when compared only to *Reafect* or *CADD*, respectively. Especially the gain in ranking performance for patients S43 (*ACADVL* gene), S56 (*ACAT1* gene), S18 and S67 (*GLDC* gene), and S70 (*OGDH* gene) is noteworthy (marked orange).

To explore the overall differences in ranking performances between the three methods, we plotted the PRs in a boxplot (Fig. 6). We removed the patient with guanidinoacetate *N*-methyltransferase deficiency from

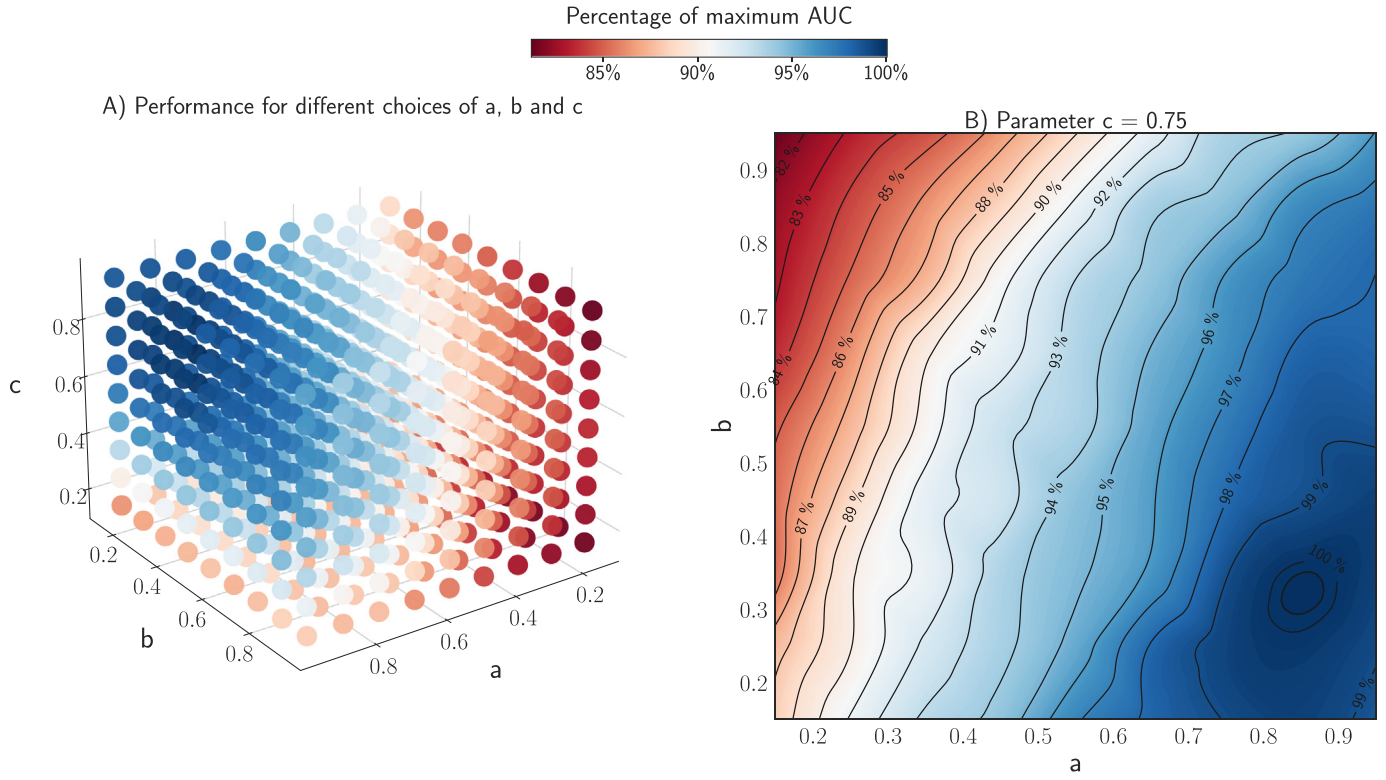


**Fig. 4. A)** Bootstrapped AUCs (Methods) for different combinations of *Reafect*'s hyper parameters ($a$, $b$, $c$). The colors indicate the percentage of the maximum obtained AUC. **B)** Contour plot of the (cubic interpolated) bootstrapped AUCs while fixing $c = 0.75$ and varying $a$ and $b$. The contour levels indicate the percentage of the maximum AUC reached at $a = 0.85$, $b = 0.35$, $c = 0.75$.

this analysis, arguing that the metabolic profile of this patient was not representative for this IEM because of the treatment. Using the Mann-Whitney *U* test, we observe that the performance between *Reafect* and *CADD* did not significantly differ ($p$-value $> 0.05$). However, the integrated approach significantly (Wilcoxon signed-rank test, p-value $<0.05$) improved the ranking performance when compared to using solely *Reafect* or *CADD* scores. In other words, by combining the two scores we gained improved IEM ranking/ gene prioritization.

## 4. Discussion

Our aim was to use metabolomics data as additional evidence for filtering genetic variants found in ES data. For this purpose, we developed *Reafect*, an algorithm that scores the efficacy of each reaction in a pathway. To calculate these scores, *Reafect* combines four types of information: 1) the magnitude and 2) sign of the metabolite *Z*-scores, 3) the biochemical directionality of reactions, and 4) the reaction distances between the metabolites and reactions in a pathway. We observed that *Reafect* ranked the true deficient enzyme for 81% of the 72 IEM patient samples within the top 5% of all considered enzyme deficiencies. On the independent Miller dataset we found similar performance, where 74% of the 106 patients were ranked within the top 5%. *Reafect* showed improved ranking performance when compared to *metPropagate*. We anticipate that this improvement may at least partially be explained by four differences between *Reafect* and *metPropagate*. First, since *metPropagate* uses cutoff values for the metabolite *Z*-scores when calculating the enrichment scores, we expect relevant but subtle aberrant metabolites to be neglected. *Reafect* uses the Z-scores in a continuous fashion, therefore even subtle aberrations contribute to the deficient reaction scores and positively impact IEM ranking (see Appendix A8). Secondly, metabolite-gene set enrichment approaches only consider metabolites which have a direct relationship with a gene, such as well-known biomarkers. Metabolite levels which are multiple reaction steps away from the deficiency may still be informative but will not contribute to the enrichment score when these metabolites are not included in the metabolite-gene set. Thirdly, *metPropagate*, and approaches like the ones suggested by Pirhaji et al. [13] and Kerkhofs et al. [5], do not explicitly take the directionality of reactions and the sign of metabolite levels (decreased/increased) into account. We observed that *Reafect*'s IEM ranking performance is reduced when all reactions in a pathway are considered to be reversible (Appendix A6), emphasizing that including reaction directionality contributes to IEM ranking. Lastly, *Reafect* explicitly searches for reactions that have (net) positive *Z*-scores at the upstream side of the reaction and (net) negative Z-scores at the downstream side of the reaction; a signature that is expected in presence of an enzymatic deficiency. In case a clear up- and downstream side of the reaction is absent (i.e. reversible reaction), *Reafect* is still capable of finding such signature by assigning one of the reaction sides to the 'upstream' - or 'downstream' side. This property also explains why *Reafect*'s IEM ranking performance remains relatively stable when all reactions are considered reversible (Appendix A6).

Integration of metabolomics with ES was achieved by multiplying the maximum deficient reaction scores with the maximum CADD score found for each enzyme and corresponding gene respectively (Methods). This integrated approach resulted in a significant improvement of ranking the true affected genes (see Fig. 6), where the median percentile rank (PR) was 1.43% lower than the median PR obtained from *Reafect*, and was 0.64% lower than the median PR obtained from using solely CADD scores.

In reality the human metabolome is one interconnected network of metabolites and reactions. In this study we have chosen to use isolated metabolic modules/pathways for two reasons. First, the (KEGG) pathways are clusters of highly interdependent reactions, for which we expect multiple metabolite levels to be affected if a pathway contains an enzymatic deficiency. Secondly, the direct use of a complete metabolic network would introduce metabolic 'hubs' that would connect more

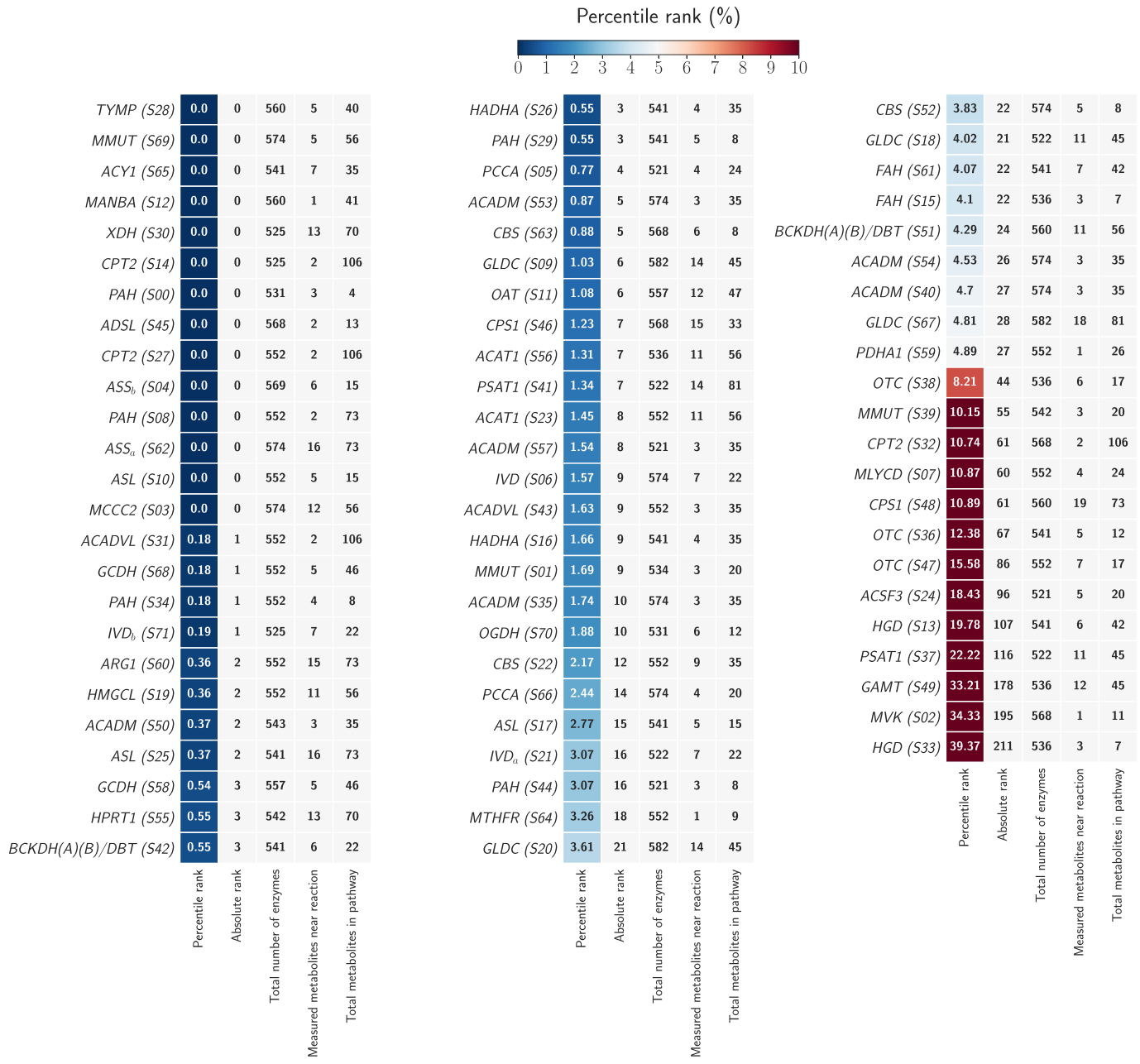


**Fig. 5.** Detailed overview of the results obtained per patient using *Reafect*. The first column indicates the PR (for the known deficient enzyme) for a given patient. Blue colors indicate PRs lower than 5%, orange/red colors indicate PRs above 5% (see colour bar). The second column shows the absolute rank of the deficient enzyme. The third column indicates the total number of the ranks/ unique enzymes on which the ranking was based (this number varies across patients due to differences in metabolite annotations). The fourth column indicates the number of annotated metabolites in the pathway on which the deficient reaction score was based. The fifth column shows the total number of metabolites present in that pathway. For the *HADHA* gene, which encodes two enzymatic functions, we selected enzyme EC 1.1.1.211. The patient samples $ASS_a$ (S62) and $ASS_b$ (S04) originate from the same patient, but were acquired on different dates. The same holds for the samples $IVD_a$ (S21) and $IVD_b$ (S71).

distinct parts of the metabolism. This entanglement of pathways/reactions may have unwanted consequences for the deficient reaction scores since also less relevant metabolite *Z*-scores would be involved in the calculation of these scores. A negative consequence of using isolated modules/pathways might be that some important reactions are not included. Although the goal was to develop an algorithm with minimum manual adjustments, we needed to add several reactions, such as glycine conjugation and carnitine esterification, to increase the overlap between metabolites measured in plasma and the metabolites included in the pathways (Appendix A3).

*Reafect* also has some limitations. First, if not all metabolites in the KEGG pathway are measured and annotated, this may lead to wrong conclusions. A single metabolite with a relatively high Z-score will cause all (downstream) reactions to have high deficient reaction scores. The inclusion of more measured metabolites could prevent this behavior, since metabolite Z-scores with the same sign on both sides of the reaction reduce the deficient reaction score (Methods, Eq. 5). The IEM ranking performance of *Reafect* is therefore affected by the number of metabolites being measured within each pathway. Similarly, the inclusion of more rare metabolites in both the pathways and obtained

**Table 1**
Overview of the IEM and affected gene ranks for 28 IEM patients using *Reafect*, *CADD* and the integrated approach. The first column indicates the patient, second columns the deficient enzyme with EC identifier. The third columns refers to the affected gene. Next columns contain the PRs for each method as indicated by the column name; *Reafect* (only), *CADD* (only), and the integrated approach. The approaches using the 15 random ES backgrounds report the mean, minimum and maximum obtained PR across the 15 backgrounds. Blue marked results indicate that the PR of *Reafect* is lower than the PR of *CADD*. Orange marked results indicate a clear improvement of the integrated approach over the individual approaches.

| Sample ID | Enzyme | Gene | *Reafect* Percentile rank (%) | *CADD* Percentile rank (%) mean [min, max] 15 random ES backgrounds | *Reafect with CADD* Percentile rank (%) mean [min, max] 15 random ES backgrounds |
|---|---|---|---|---|---|
| S35 | 1.3.8.7 | *ACADM* | 1.74 | 0.39 [0.12,0.59] | 0.0 [0.0,0.0] |
| S53 | 1.3.8.7 | *ACADM* | 0.87 | 6.08 [5.37,7.27] | 0.27 [0.11,0.37] |
| S54 | 1.3.8.7 | *ACADM* | 4.53 | 0.39 [0.12,0.59] | 0.14 [0.0,0.24] |
| S57 | 1.3.8.7 | *ACADM* | 1.54 | 0.43 [0.12,0.63] | 0.04 [0.0,0.12] |
| S43 | 1.3.8.9 | *ACADVL* | 1.63 | 2.06 [1.28,2.65] | 0.28 [0.12,0.5] |
| S31 | 1.3.8.9 | *ACADVL* | 0.18 | 1.58 [1.04,2.07] | 0.06 [0.0,0.36] |
| S56 | 2.3.1.9 | *ACAT1* | 1.31 | 2.45 [1.32,3.37] | 0.42 [0.24,0.74] |
| S23 | 2.3.1.9 | *ACAT1* | 1.45 | 0.05 [0.0,0.23] | 0.01 [0.0,0.12] |
| S17 | 4.3.2.1 | *ASL* | 2.77 | 1.87 [1.19,2.48] | 2.44 [2.06,2.68] |
| S32 | 2.3.1.21 | *CPT2* | 10.74 | 0.73 [0.36,1.31] | 4.5 [3.8,5.13] |
| S27 | 2.3.1.21 | *CPT2* | 0 | 0.23 [0.0,0.47] | 0.0 [0.0,0.0] |
| S15 | 3.7.1.2 | *FAH* | 4.1 | 4.04 [2.51,5.21] | 4.98 [4.55,5.31] |
| S49 | 2.1.1.2 | *GAMT* | 33.21 | 0.06 [0.0,0.13] | 5.61 [4.62,6.65] |
| S58 | 1.3.8.6 | *GCDH* | 0.54 | 0.36 [0.12,0.6] | 0.18 [0.0,0.46] |
| S18 | 1.4.4.2 | *GLDC* | 4.02 | 2.29 [1.3,3.16] | 0.79 [0.48,1.07] |
| S67 | 1.4.4.2 | *GLDC* | 4.81 | 2.15 [1.21,2.96] | 0.35 [0.11,0.55] |
| S26 | 1.1.1.211 | *HADHA* | 0.55 | 1.05 [0.64,1.48] | 0.11 [0.0,0.25] |
| S26 | 4.2.1.17 | *HADHA* | 0.18 | 1.05 [0.64,1.48] | 0.11 [0.0,0.25] |
| S12* | 3.2.1.25 | *MANBA* | 0 | 0.41 [0.12,0.61] | 0.0 [0.0,0.0] |
| S03 | 6.4.1.4 | *MCCC2* | 0 | 1.54 [1.01,1.97] | 0.0 [0.0,0.0] |
| S39 | 5.4.99.2 | *MMUT* | 10.15 | 0.84 [0.36,1.32] | 1.04 [0.47,1.37] |
| S64 | 1.5.1.20 | *MTHFR* | 3.26 | 0.53 [0.23,0.96] | 0.76 [0.47,1.05] |
| S02 | 2.7.1.36 | *MVK* | 34.33 | 0.68 [0.36,1.19] | 12.59 [11.15,13.7] |
| S70 | 1.2.4.2 | *OGDH* | 1.88 | 8.14 [2.86,10.15] | 0.2 [0.0,0.48] |
| S38 | 2.1.3.3 | *OTC* | 8.21 | 0.49 [0.23,0.74] | 1.55 [1.21,1.9] |
| S34 | 1.14.16.1 | *PAH* | 0.18 | 2.1 [1.28,2.77] | 0.0 [0.0,0.0] |
| S00 | 1.14.16.1 | *PAH* | 0 | 1.22 [0.76,1.69] | 0.0 [0.0,0.0] |
| S59** | 1.2.4.1 | *PDHA1* | 4.89 | 0.63 [0.35,1.08] | 0.21 [0.12,0.49] |
| S28 | 2.4.2.4 | *TYMP* | 0 | 0.41 [0.24,0.61] | 0.0 [0.0,0.0] |

* The PR for *CADD* was 0.25% using the real ES, and 0.0% for *Reafect with CADD* using the real ES.
** The PR for *CADD* was 0.58% using the real ES, and 0.35% for *Reafect with CADD* using the real ES.

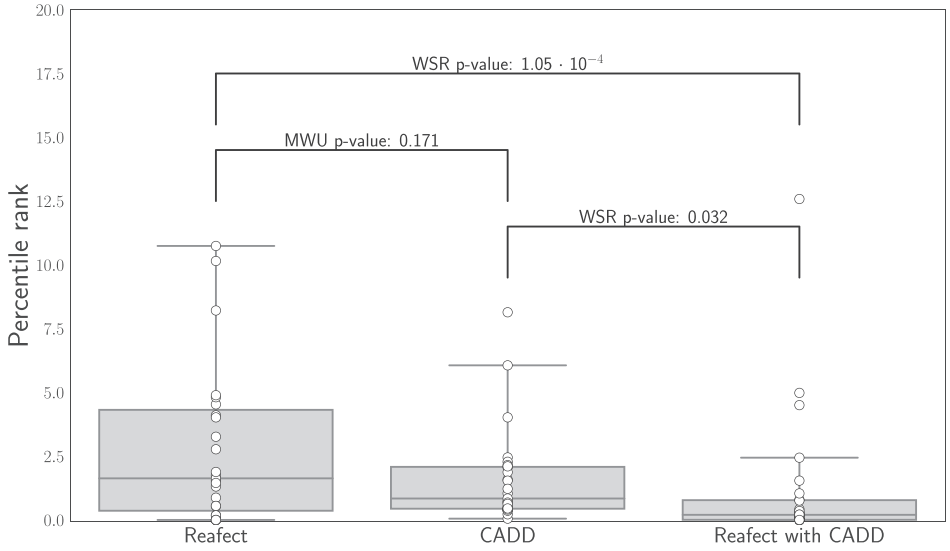


**Fig. 6.** Boxplots of the percentile ranks (PRs) obtained from the different approaches; *Reafect* (only), *CADD* (only), and the integrated approach. For *CADD* and the integrated approach we used the average PR obtained from the 15 random ES. Significance was determined using the Wilcoxon signed-rank test (WSR) when comparing *Reafect with CADD* with *CADD* or *Reafect*. We used the Mann-Whitney U test (MWU) for comparing *CADD* with *Reafect*, arguing that the PRs for *CADD* and *Reafect* are independent since they are obtained from two separate datasets and approaches.

metabolomics data may positively affect IEM ranking. Secondly, *Reafect* is based on the assumption that IEM have the signature where substrates of the deficient reaction become more abundant and the products decrease in abundancy. In case such a signature does not hold for a certain IEM, we expect *Reafect* to detect these kinds of IEM poorly. Finally, *Reafect* ignores compartmentalization of different metabolic processes. A substantial number of metabolic reactions occur within certain compartments of the cell such as the mitochondrion. Similarly, different organs contain different sets of metabolic reactions, therefore the concentration of the affected metabolites for an IEM may be very different from the concentrations measured in plasma on which our Z-scores are based. The inclusion of (extra) metabolite Z-scores obtained from metabolomics data from other body fluids (e.g. urine) could theoretically extend the analysis of *Reafect*.

For most IEM patients with an identified disease-causing gene variant in this study, the putative gene was directly sequenced, and therefore no ES data was obtained. We inserted the identified disease-causing variant in 15 random ES backgrounds, to enable the inclusion of these patients in our study. We assumed that the average ranking obtained from these 15 backgrounds was still a good estimate of the ranking which would have been obtained when the real ES data was used. Due to our limited number of patients with real ES data ($N = 2$), a reliable comparison between both rankings is not possible, and thus we cannot validate the accuracy of this assumption. We realize that the current assessment should be considered as a proof-of-concept, and that future studies should validate the accuracy of our findings. Still, note that the PRs obtained from the real ES fall within the minimum and maximum PR obtained from the 15 ES backgrounds (Table 1).

*Reafect* uses only three decay factors ($a$, $b$, $c$) which we optimized using an overall performance metric (see Results, Fig. 4). Ideally, these decay factors are optimized using a training set while using a separate validation set for evaluating the IEM ranking performances. Due to the low number of IEM patients included in our dataset we decided to use all samples for optimization and validation, arguing that splitting the dataset into a training - and validation set would lead to less accurate estimates of the decay factors and would give less insights into the overall performance of *Reafect* on distinct IEM. Consequently, the results based on our dataset might have been biased. However, we obtained similar ranking performances when *Reafect* was applied to the (independent) Miller dataset, which underlines the validity of our conclusions.

We realize that the use of three decay factors is a simplification, and that these factors should ideally be reaction specific. Kinetic parameters, such as the Michaelis–Menten constant, could be used to establish such reaction dependent decay factor. Currently accurate kinetic parameters are only available for a subset of reactions. Besides the additional complexity introduced by these reaction specific decay factors, the use of just three decay factors offered us the opportunity to demonstrate the overall importance of choosing different decay factors for reaction directionality and the sign of the Z-score, as we clearly observed in Fig. 4. Still, we anticipate that *Reafect* 's performance on ranking IEM/genes could improve when reaction specific decay factors are incorporated.

*Reafect* may not only be useful in the context of IEM but could be applicable in a wider context since the deficient reaction scores are a direct readout of potential accumulations and/or reductions of metabolites before/after a reaction. For example, *Reafect* is potentially useful in drug screening research for generating an overview of drug candidates which have the potential to inhibit metabolic enzymes. Namely, we expect that the inhibition of an enzyme by a drug will result in metabolic signatures similar to the ones caused by an IEM where the same enzyme is affected.

## 5. Conclusions

In conclusion, the integration of metabolomics data with ES data by using *Reafect*'s deficient reaction scores and CADD scores, significantly improved the prioritization of the affected genes in patients suffering from an IEM. A next step is to investigate the use of *Reafect* as part of a clinical screening procedure.

## Funding

This work was funded by the Erasmus Medical Centre, department of Clinical Genetics.

## Availability of data and materials

*Reafect* is available at https://github.com/mbongaerts/Reafect. All (KEGG) associations between reactions, enzymes and genes used in this study can be found here. The Z-score data for the 72 IEM patient samples is also available from this source.

## Authors' contributions

The development of the methods was done by MB, MR and GR. RB performed all the experimental work, among which compound identification of the metabolomics data. MB developed the software and performed the computational experiments. WdV contributed in methods to analyze and process ES data. The interpretation of the results was done by MB, MR, HB and GR. The manuscript was written by MB, MR, HB and GR. SD, JL, HH and MW provided data and resources. The research was under supervision of GR.

## Declaration of Competing Interest

All authors state that they have no conflict of interest to declare. None of the authors accepted any reimbursements, fees, or funds from any organization that may in any way gain or lose financially from the results of this study. The authors have not been employed by such an organization. The authors do not have any other conflict of interest.

## Acknowledgements

We dedicate this study to the memory of Professor Robert Hofstra for his continuous scientific support, open mind and care for all employees of his department. We furthermore want to thank and acknowledge Dr. Geert Geeven for his comments and feedback on the manuscript.

## Appendices

### A.1. Overview of which patients received treatment

**Table A1**
Overview of which IEM patient samples received treatment.

| Sample | Treatment | Comment |
|---|---|---|
| S00 | Protein restriction, amino acid supplement | |
| S01 | Liver transplantation, protein restriction, carnitine supplement | |
| S02 | None | At diagnosis |
| S03 | None | |
| S04 | Protein restriction, arginine supplement, carnitine supplement, benzoate, phenylbutyrate | |
| S05 | Protein restriction, carnitine supplement, Carbaglu | |
| S06 | Carnitine supplement | |
| S07 | Carnitine supplement | |
| S08 | Protein restriction, amino acid supplement | |
| S09 | None | At diagnosis |
| S10 | Protein restriction, arginine supplement, phenylbutyrate | |
| S11 | Protein restriction, carnitine supplement, Carbaglu | |
| S12 | None | At diagnosis |
| S13 | None | |
| S14 | None | |
| S15 | Liver transplantation, protein restriction, Nitisinone | |
| S16 | Carnitine supplement | |
| S17 | Protein restriction, arginine supplement, phenylbutyrate | |
| S18 | None | At diagnosis |
| S19 | Carnitine supplement | |
| S20 | None | At diagnosis |
| S21 | Carnitine supplement | |
| S22 | B12 supplement, folate supplement, betaine | |
| S23 | Carnitine supplement | |
| S24 | None | |
| S25 | Protein restriction | |
| S26 | Carnitine supplement | |
| S27 | None | |
| S28 | None | |
| S29 | Protein restriction, amino acid supplement | |
| S30 | None | At diagnosis |
| S31 | LCT restriction, MCT supplement | |
| S32 | None | |
| S33 | None | |
| S34 | Protein restriction, amino acid supplement | |
| S35 | None | |
| S36 | Protein restriction, citrulline supplement, carnitine supplement, benzoate, phenylbutyrate | |
| S37 | Serine supplement | |
| S38 | Protein restriction, citrulline supplement, carnitine supplement, benzoate, phenylbutyrate | |
| S39 | Liver transplantation, protein restriction, carntine supplement, B12 supplement | |
| S40 | None | |
| S41 | Serine supplement | |
| S42 | Protein restriction | |
| S43 | LCT restriction, MCT supplement | |
| S44 | Protein restriction, Kuvan | |
| S45 | None | At diagnosis |
| S46 | Protein restriction | |
| S47 | Protein restriction, carnitine supplement, benzoate, phenylbutyrate | |
| S48 | None | |
| S49 | Creatine supplement, ornithine supplement, benzoate | |
| S50 | Carnitine supplement | |
| S51 | Protein restriction | |
| S52 | B12 supplement, folate supplement, betaine | |
| S53 | Carnitine supplement | |
| S54 | Carnitine supplement | |
| S55 | Allopurinol | |
| S56 | None | |
| S57 | Carnitine supplement | |
| S58 | Carnitine supplement | |
| S59 | Thiamine supplement | |
| S60 | Protein restriction, carnitine supplement, benzoate | |
| S61 | Protein restriction, Nitisinone | |
| S62 | Protein restriction, arginine supplement, carnitine supplement, benzoate, phenylbutyrate | |
| S63 | B12 supplement, folate supplement, betaine | |
| S64 | B12 supplement, folate supplement, betaine | |
| S65 | None | |
| S66 | Protein restriction, carnitine supplement, Carbaglu | |
| S67 | None | At diagnosis |

**Table A1** (*continued*)

| Sample | Treatment | Comment |
|---|---|---|
| S68 | Carnitine supplement | |
| S69 | Protein restriction, carntine supplement, B12 supplement, Carbaglu | |
| S70 | Thiamine, carnitine supplement | |
| S71 | Carnitine supplement | |

## *A.2. Overview of variants*

**Table A2**
Information about the variants found in the 28 IEM patients. ACMG classification was provided for each variant [23].

| Sample ID | Gene symbol | CADD (Phred) | Homozygous/Heterozygous | ACMG classification | Conclusion |
|---|---|---|---|---|---|
| S00 | *PAH* | 27.5 | Heterozygous | PM1, PP2, PM2, PM5, PP3, PP5 | Pathogenic |
| S00 | *PAH* | 28.5 | Heterozygous | PP2, PM2, PM5, PP3, PP5 | Pathogenic |
| S02 | *MVK* | 32 | Heterozygous | PVS1, PM2 | (Likely)pathogenic |
| S03 | *MCCC2* | 27 | Heterozygous | PM2, PP3, PP2, PP5 | Pathogenic |
| S03 | *MCCC2* | 23 | Heterozygous | PM2, PP3, PP2 | VUS |
| S12 | *MANBA* | 33 | Heterozygous | PM2 | VUS |
| S15 | *FAH* | 24 | Homozygous | PM2, PP3, PP2, PP5 | Pathogenic |
| S17 | *ASL* | 26.5 | Homozygous | PM1, PP2, PM2, PM5, PP3, PP5 | pathogenic |
| S18 | *GLDC* | 26 | Heterozygous | PM2, PM5, PP2, PP3 | VUS, almost likely pathogenic |
| S23 | *ACAT1* | 38 | Homozygous | PVS1, PM2 | Likely pathogenic |
| S26 | *HADHA* | 29 | Homozygous | PM2, PP3, PP5, | Pathogenic |
| S27 | *CPT2* | 34 | Heterozygous | PM2, PP3, PP5, | Pathogenic |
| S27 | *CPT2* | 24 | Heterozygous | PM2, PM5, PP3, PM1 | Likely pathogenic |
| S28 | *TYMP* | 33 | Homozygous | PM2, PM1, PM5, PP3, PP5 | Pathogenic |
| S31 | *ACADVL* | 27.5 | Homozygous | PM2, PM1, PP2, PP3 | VUS (strong) |
| S32 | *CPT2* | 31 | Homozygous | PM2, PP3, PP5 | Pathogenic |
| S34 | *PAH* | 26 | Heterozygous | PM1, PP2, PM2, PP3, PP5 | Pathogenic |
| S35 | *ACADM* | 33 | Homozygous | PM1, PP2, PM2, PP3, PP5 | Pathogenic |
| S38 | *OTC* | 33 | Heterozygous | PM2, PM1, PP2, PM5, PP3, PP5 | Pathogenic |
| S39 | *MMUT* | 30 | Homozygous | PM2, PM5, PM1, PP2, PP3, PP5 | Pathogenic |
| S43 | *ACADVL* | 26 | Heterozygous | PM2, PM1, PP2, PP3, PP5 | Pathogenic |
| S43 | *ACADVL* | 24.5 | Heterozygous | PM2, PM1, PP2, PP3, PP5 | Pathogenic |
| S49 | *GAMT* | 39 | Homozygous | PVS1, PM2, PP5 | Pathogenic |
| S53 | *ACADM* | 23 | Homozygous | PM1, PP2, PM2, PM5, PP5 | Pathogenic |
| S54 | *ACADM* | 33 | Homozygous | PP2, PM2, PP3, PP5 | Pathogenic |
| S56 | *ACAT1* | 25.5 | Homozygous | PM2, PP3, PP2, PP5 | Likely pathogenic |
| S57 | *ACADM* | 33 | Homozygous | PP2, PM2, PP3, PP5 | Pathogenic |
| S58 | *GCDH* | 21 | Heterozygous | PS1, PM5, PM1, PP2, PP3, PP5 | Pathogenic |
| S58 | *GCDH* | 34 | Heterozygous | PM1, PP2, PM2, PM5, PP3, PP5 | Pathogenic |
| S59 | *PDHA1* | 32 | Heterozygous | PM2, PM5, PP3, PP2, PP5, PS2 | (likely) Pathogenic |
| S64 | *MTHFR* | 0.15 | Heterozygous | PM2, PP5 | VUS |
| S64 | *MTHFR* | 33 | Heterozygous | PM2, PP3 | VUS |
| S67 | *GLDC* | 26 | Heterozygous | PM2, PP3, PP2, PP5 | Likely Pathogenic |
| S67 | *GLDC* | 25.5 | Heterozygous | PM2, PP3, PP2 | VUS (strong) |
| S70 | *OGDH* | 22.5 | Homozygous | PM2, PP2 | VUS |

### A.3. Manually added reactions

The KEGG pathways and modules were extended with some additional reactions (see Table A3) to increase the overlap between metabolites present in the pathways/modules and metabolites measured in plasma. Note, that a reaction is defined as a graph which also includes a reaction node.

**Table A3**
Manually added reactions. The second and fourth column indicate the directionality of the reaction.'< = >' indicates that the reaction is reversible whereas'=>' indicates the direction of an irreversible reaction. Note that these reactions passes through a reaction node (Reaction ID). Most reactions originate from Recon3D.

| Metabolite 1 | | Reaction ID | | Metabolite 2 | Source |
|---|---|---|---|---|---|
| 3-Hydroxypyruvic acid | => | HPYRR2x | => | Glyceric acid | https://vmh.life/#reaction/HPYRR2x |
| 2-Methylbutyrylglycine | <=> | RE2428M | <=> | 2-Methylbutanoyl-CoA | https://vmh.life/#reaction/RE2428M |
| 2-Methylbutyrylglycine | <=> | RE2428M | <=> | Glycine | https://vmh.life/#reaction/RE2428M |
| C16OH \| 3-Hydroxyhexadecanoylcarnitine | <=> | C16OHc | <=> | (*S*)-3-Hydroxyhexadecanoyl-CoA | https://vmh.life/#reaction/C16OHc |
| 3-Methylcrotonylglycine | <=> | RE2111M | <=> | 3-Methylcrotonyl-CoA | https://vmh.life/#reaction/RE2111M |
| 3-Methylcrotonylglycine | <=> | RE2111M | <=> | Glycine | https://vmh.life/#reaction/RE2111M |
| glcnac-man | => | B_MANNASEly | => | N-Acetyl-D-glucosamine | https://vmh.life/#reaction/B_MANNASEly |
| glcnac-man | => | B_MANNASEly | => | D-Mannose | https://vmh.life/#reaction/B_MANNASEly |
| C10 \| Decanoylcarnitine | <=> | C100CPT1 | <=> | Decanoyl-CoA | https://vmh.life/#reaction/C100CPT1 |
| Glycylproline | <=> | GLYPROPRO1c | <=> | Glycine | https://vmh.life/#reaction/GLYPROPRO1c |
| Glycylproline | <=> | GLYPROPRO1c | <=> | Proline | https://vmh.life/#reaction/GLYPROPRO1c |
| C6 \| Hexanoylcarnitine | <=> | C60CPT1 | <=> | Hexanoyl-CoA | https://vmh.life/#reaction/C60CPT1 |
| Homocysteine thiolactone | <=> | RE1933C | <=> | Homocysteine | https://vmh.life/#reaction/RE1933C |
| Isobutyrylglycine | <=> | RE2429M | <=> | Glycine | https://vmh.life/#reaction/RE2429M |
| Isobutyrylglycine | <=> | RE2429M | <=> | 2-Methylpropanoyl-CoA | https://vmh.life/#reaction/RE2429M |
| C5 \| Isovalerylcarnitine | <=> | C50CPT1 | <=> | 3-Methylbutanoyl-CoA | https://vmh.life/#reaction/C50CPT1 |
| Isovalerylglycine | <=> | RE2427M | <=> | Glycine | https://vmh.life/#reaction/RE2427M |
| Isovalerylglycine | <=> | RE2427M | <=> | 3-Methylbutanoyl-CoA | https://vmh.life/#reaction/RE2427M |
| Malonyl-CoA | => | r0430 | => | C3DC \| Malonylcarnitine | https://vmh.life/#reaction/r0430 |
| N-Acetylasparagine | <=> | RE2032M | <=> | Asparagine | https://vmh.life/#reaction/RE2032M |
| C14 \| Tetradecanoylcarnitine | <=> | C140CPT1 | <=> | Tetradecanoyl-CoA | https://vmh.life/#reaction/C140CPT1 |
| C5:1 \| Tiglylcarnitine | <=> | C51CPT1 | <=> | 2-Methylbut-2-enoyl-CoA | https://vmh.life/#reaction/C51CPT1 |
| Octanoyl-CoA | <=> | C80CPT1 | <=> | C8 \| Octanoylcarnitine | https://vmh.life/#reaction/C80CPT1 |
| Butanoyl-CoA | <=> | C40CPT1 | <=> | C4 \| Butyrylcarnitine | https://vmh.life/#reaction/C40CPT1 |
| Propanoyl-CoA | <=> | C30CPT1 | <=> | C3 \| Propionylcarnitine | https://vmh.life/#reaction/C30CPT1 |
| (2*S*,3*S*)-3-Hydroxy-2-methylbutanoyl-CoA | <=> | R_2M3HBUc | <=> | 2-Methyl-3-hydroxybutyric acid | https://vmh.life/#reaction/2M3HBUc |
| 2-Methylbut-2-enoyl-CoA | <=> | R_TIGGLYc | <=> | Tiglylglycine | https://vmh.life/#reaction/TIGGLYc |
| N-Acetylmethionine | <=> | RE2640C | <=> | Methionine | https://vmh.life/#reaction/RE2640C |
| N-Acetylalanine | <=> | RE2642C | <=> | L-Alanine | https://vmh.life/#reaction/RE2642C |
| Glutaryl-CoA | <=> | FAOXC5C5DCc | <=> | C5DC \| Glutarylcarnitine | https://vmh.life/#reaction/FAOXC5C5DCc |
| 3-Methylglutaconyl-CoA | <=> | 3mgcoac61dcmgccrn | <=> | C6:1DC \| 3-Methylglutaconylcarnitine | EC 2.3.1.7, EC 2.3.1.137 https://doi.org/10.1016/j.bbadis.2013.02.012 https://doi.org/10.1016/-S0163-7827(99)00002-8 |
| 3-Methylglutaconyl-CoA | <=> | MGCHrm | <=> | (*S*)-3-Hydroxy-3-methylglutaryl-CoA | https://vmh.life/#reaction/MGCHrm |
| (*S*)-3-Hydroxy-3-methylglutaryl-CoA | <=> | hmgcoac6dcmgcrn | <=> | C6DC \| 3-Methylglutarylcarnitine | EC 2.3.1.7, EC 2.3.1.137 |
| 3-Hydroxyisovaleryl-CoA | <=> | C059983ivcrn | <=> | C5OH \| 3-Hydroxyisovalerylcarnitine | https://vmh.life/#reaction/FAOXC5OHc |
| 3-Hydroxyisovaleryl-CoA | <=> | C059983CE2028 | <=> | 3-Hydroxyisovaleric acid | EC 3.1.2.20 https://doi.org/10.1111/j.1365-2362.2005.01447.x |
| Adenylosuccinate | => | C03794succinyladenosine | => | Succinyladenosine | EC 3.1.3.5 |
| L-Aspartate | <=> | ASPCTr | <=> | N-Carbamoyl-L-aspartate | https://vmh.life/#reaction/ASPCTr |
| Carbamoylphosphate | <=> | ASPCTr | <=> | N-Carbamoyl-L-aspartate | https://vmh.life/#reaction/ASPCTr |
| Dihydroorotic acid | => | DHORTS | => | N-Carbamoyl-L-aspartate | https://vmh.life/#reaction/DHORTS |
| Dihydroorotic acid | => | DHORD9 | => | Orotic acid | https://vmh.life/#reaction/DHORD9 |
| Malonic acid | => | C00383malcoa | => | Malonyl-CoA | EC 3.1.2.20 https://doi.org/10.1016/S0021-9258(17)44433-4 |
| Malonyl-CoA | => | MCDm | => | Acetyl-CoA | https://vmh.life/#reaction/MCDm |
| 7-Dehydrocholesterol | => | HMR_2114 | => | Vitamine D3 | https://vmh.life/#reaction/HMR_2114 |
| Cholesterol sulfate | <=> | RE1100L | <=> | Cholesterol | https://vmh.life/#reaction/RE1100L |

### A.4. Comparison of ranks IEM patients Reafect versus MetPropagate

We compared the IEM ranking performance of *Reafect* with *metPropagate* (Fig. 3). The percentile ranks obtained for each patient and both methods are displayed in Fig. A1.

| | Percentile rank for metPropagate | Percentile rank for Reafect |
|---|---|---|
| ACADM (S35) | 34.21 | 16.54 |
| ACADM (S40) | 5.83 | 23.31 |
| ACADM (S50) | 27.61 | 2.04 |
| ACADM (S53) | 26.5 | 18.05 |
| ACADM (S54) | 28.01 | 21.99 |
| ACADM (S57) | 9.43 | 18.07 |
| ACADVL (S31) | 10.94 | 0.57 |
| ACADVL (S43) | 60.59 | 23.14 |
| ACAT1 (S23) | 37.84 | 1.57 |
| ACAT1 (S56) | 32.79 | 1.42 |
| ACSF3 (S24) | 65.62 | 16.31 |
| ACY1 (S65) | 12.17 | 0.0 |
| ADSL (S45) | 0.38 | 0.0 |
| ARG1 (S60) | 2.35 | 0.0 |
| ASL (S10) | 0.59 | 0.0 |
| ASL (S17) | 1.33 | 2.85 |
| ASL (S25) | 1.32 | 0.38 |
| $ASS_a$ (S62) | 0.94 | 0.0 |
| $ASS_b$ (S04) | 0.72 | 0.72 |
| BCKDH(A)(B)/DBT (S42) | 50.85 | 0.56 |
| BCKDH(A)(B)/DBT (S51) | 3.67 | 25.48 |
| CBS (S22) | 10.2 | 2.16 |
| CBS (S52) | 7.71 | 4.14 |
| CBS (S63) | 17.3 | 0.95 |
| CPS1 (S46) | 3.99 | 1.52 |
| CPS1 (S48) | 34.17 | 10.62 |
| CPT2 (S14) | 0.2 | 0.0 |
| CPT2 (S27) | 0.0 | 0.0 |
| CPT2 (S32) | 5.89 | 11.41 |
| FAH (S15) | 71.86 | 4.45 |
| FAH (S61) | 46.33 | 3.95 |
| GAMT (S49) | 0.0 | 30.96 |
| GCDH (S58) | 89.71 | 89.51 |
| GCDH (S68) | 69.41 | 90.98 |
| GLDC (S09) | 41.85 | 1.3 |
| GLDC (S18) | 58.42 | 3.85 |
| GLDC (S20) | 4.63 | 3.7 |
| GLDC (S67) | 0.93 | 4.26 |
| HADHA (S16) | 41.43 | 54.43 |
| HADHA (S26) | 33.33 | 46.52 |
| HGD (S13) | 35.17 | 17.87 |
| HGD (S33) | 60.77 | 38.65 |
| HMGCL (S19) | 26.6 | 27.36 |
| HPRT1 (S55) | 17.36 | 0.38 |
| IVD (S06) | 57.14 | 3.76 |
| $IVD_a$ (S21) | 65.52 | 3.25 |
| $IVD_b$ (S71) | 57.59 | 2.56 |
| MANBA (S12) | 78.19 | 95.56 |
| MCCC2 (S03) | 72.74 | 0.0 |
| MLYCD (S07) | 73.14 | 59.41 |
| MMUT (S01) | 67.06 | 5.1 |
| MMUT (S39) | 69.81 | 9.43 |
| MMUT (S69) | 69.2 | 0.0 |
| MTHFR (S64) | 0.0 | 3.4 |
| MVK (S02) | 67.68 | 33.84 |
| OAT (S11) | 1.75 | 1.17 |
| OGDH (S70) | 14.07 | 1.54 |
| OTC (S36) | 33.27 | 12.74 |
| OTC (S38) | 9.04 | 7.88 |
| OTC (S47) | 33.92 | 15.29 |
| PAH (S00) | 0.19 | 0.0 |
| PAH (S08) | 0.2 | 0.0 |
| PAH (S29) | 0.56 | 0.38 |
| PAH (S34) | 0.2 | 0.2 |
| PAH (S44) | 0.79 | 2.95 |
| PCCA (S05) | 1.38 | 0.79 |
| PCCA (S66) | 48.12 | 2.63 |
| PDHA1 (S59) | 15.47 | 4.53 |
| PSAT1 (S37) | 34.89 | 24.34 |
| PSAT1 (S41) | 36.31 | 4.67 |
| TYMP (S28) | 0.19 | 0.0 |
| XDH (S30) | 0.39 | 0.0 |

**Fig. A1.** Comparison of the percentile ranks between *Reafect* and *metPropagate* per IEM patient.

*A.5. Overview ranks determined by Reafect on the Miller dataset*

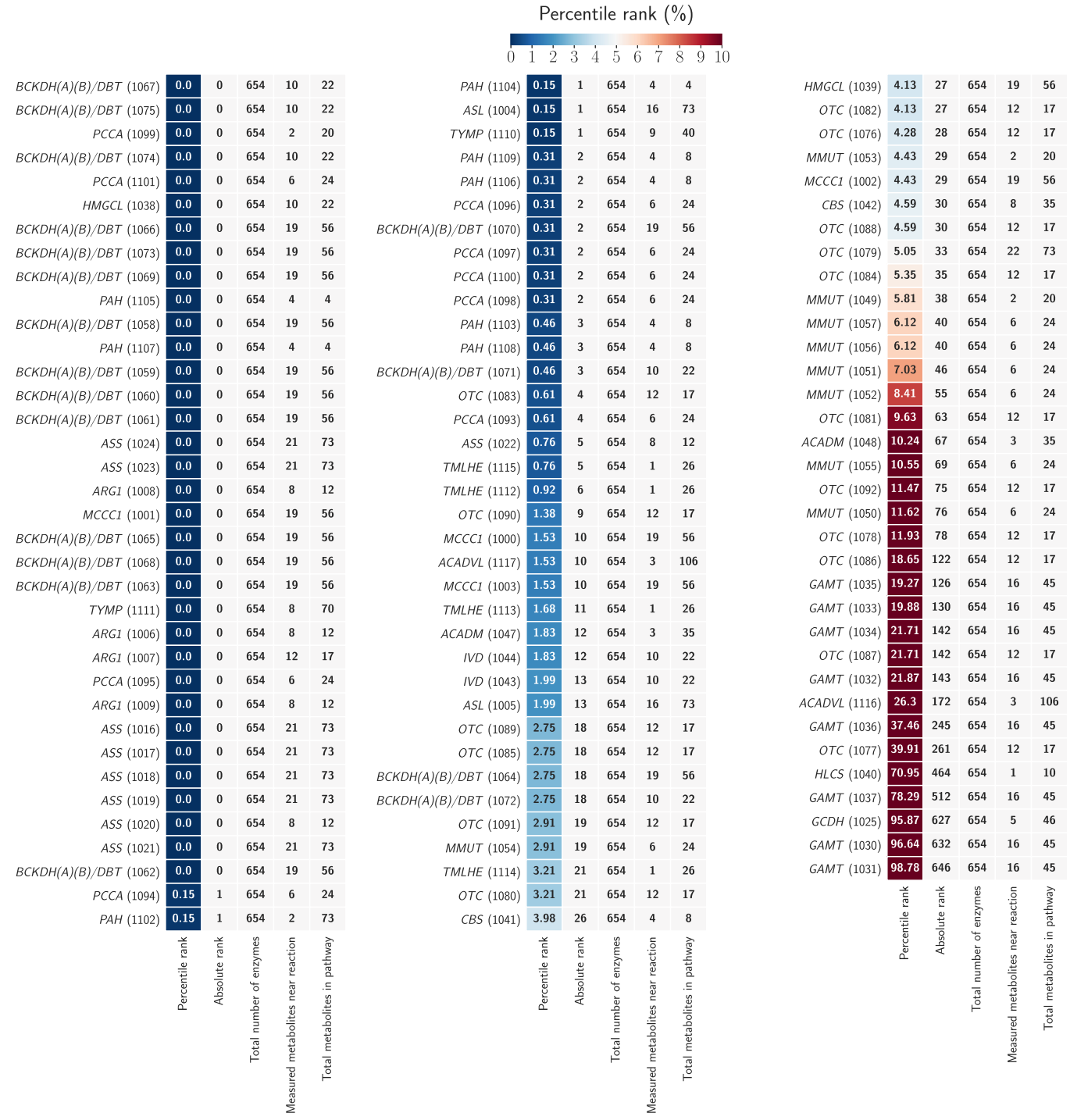


**Fig. A2.** Detailed overview of the results obtained per patient using *Reafect* applied on the Miller dataset. The first column indicates the PR (for the known deficient enzyme) for a given patient. Blue colors indicate PRs lower than 5%, orange/red colors indicate PRs above 5% (see colour bar). The second column shows the absolute rank of the deficient enzyme. The third column indicates the total number of the ranks/ unique enzymes on which the ranking was based (this number varies across patients due to differences in metabolite annotations). The fourth column indicates the number of annotated metabolites in the pathway on which the deficient reaction score was based. The fifth column shows the total number of metabolites present in that pathway.

### *A.6. The effect of removing reaction directionality on IEM ranking performance*

To explore the importance of taking the biochemical directionality into account, we compared the situation where *Reafect* was applied to pathways that include reaction directionality (default setting) with the situation where all reactions were considered reversible (bidirectional). We observe that the overall IEM ranking performance was dropped by 2% when all reactions were considered bidirectional (Fig. A3A). For the partial AUC of the ranking performance, we observe a 10% decrease in performance. A detailed comparison between the ranks obtained from both approaches can be found in Fig. A4. Note that for reversible reactions *Reafect* also searches for signatures where (net) positive *Z*-scores are found at one side of the reaction and (net) negative Z-scores are found at the other side of the reaction. This explains why the overall IEM ranking performance of the situation where all reactions are reversible remains relatively stable. Still, the inclusion of reaction directionality does contribute positively to IEM ranking.

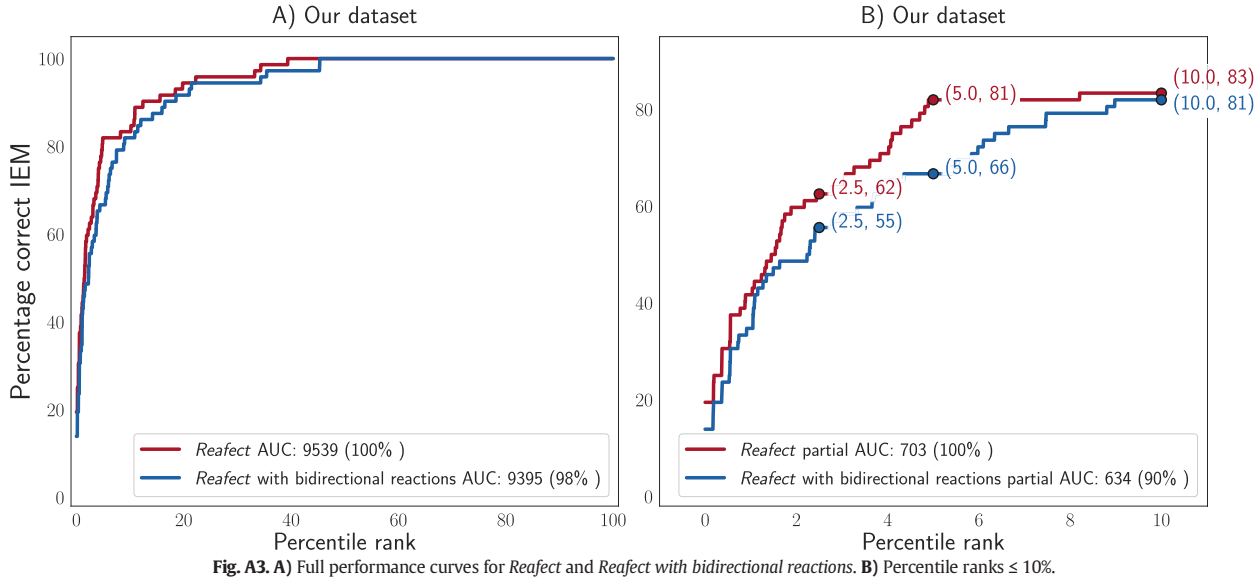


**Fig. A3. A)** Full performance curves for *Reafect* and *Reafect with bidirectional reactions*. **B)** Percentile ranks ≤ 10%.

| | Percentile rank for Reafect | Percentile rank for Reafect with bidirectional reactions |
|---|---|---|
| ACADM (S35) | 1.74 | 2.79 |
| ACADM (S40) | 4.7 | 5.4 |
| ACADM (S50) | 0.37 | 0.37 |
| ACADM (S53) | 0.87 | 1.05 |
| ACADM (S54) | 4.53 | 6.1 |
| ACADM (S57) | 1.54 | 2.3 |
| ACADVL (S31) | 0.18 | 0.36 |
| ACADVL (S43) | 1.63 | 1.27 |
| ACAT1 (S23) | 1.45 | 1.63 |
| ACAT1 (S56) | 1.31 | 1.49 |
| ACSF3 (S24) | 18.43 | 45.3 |
| ACY1 (S65) | 0.0 | 0.0 |
| ADSL (S45) | 0.0 | 0.18 |
| ARG1 (S60) | 0.36 | 1.09 |
| ASL (S10) | 0.0 | 0.0 |
| ASL (S17) | 2.77 | 3.33 |
| ASL (S25) | 0.37 | 0.0 |
| $ASS_a$ (S62) | 0.0 | 0.17 |
| $ASS_b$ (S04) | 0.0 | 0.53 |
| BCKDH(A)(B)/DBT (S42) | 0.55 | 0.74 |
| BCKDH(A)(B)/DBT (S51) | 4.29 | 11.43 |
| CBS (S22) | 2.17 | 5.98 |
| CBS (S52) | 3.83 | 4.36 |
| CBS (S63) | 0.88 | 1.06 |
| CPS1 (S46) | 1.23 | 3.7 |

| | Percentile rank for Reafect | Percentile rank for Reafect with bidirectional reactions |
|---|---|---|
| CPS1 (S48) | 10.89 | 5.54 |
| CPT2 (S14) | 0.0 | 0.0 |
| CPT2 (S27) | 0.0 | 0.0 |
| CPT2 (S32) | 10.74 | 8.98 |
| FAH (S15) | 4.1 | 7.46 |
| FAH (S61) | 4.07 | 6.65 |
| GAMT (S49) | 33.21 | 0.37 |
| GCDH (S58) | 0.54 | 1.08 |
| GCDH (S68) | 0.18 | 0.54 |
| GLDC (S09) | 1.03 | 2.23 |
| GLDC (S18) | 4.02 | 3.83 |
| GLDC (S20) | 3.61 | 2.41 |
| GLDC (S67) | 4.81 | 18.56 |
| HADHA (S16) | 1.66 | 2.4 |
| HADHA (S26) | 0.55 | 0.55 |
| HGD (S13) | 19.78 | 21.44 |
| HGD (S33) | 39.37 | 45.34 |
| HMGCL (S19) | 0.36 | 0.91 |
| HPRT1 (S55) | 0.55 | 11.99 |
| IVD (S06) | 1.57 | 3.66 |
| $IVD_a$ (S21) | 3.07 | 7.47 |
| $IVD_b$ (S71) | 0.19 | 2.29 |
| MANBA (S12) | 0.0 | 0.0 |
| MCCC2 (S03) | 0.0 | 0.17 |
| MLYCD (S07) | 10.87 | 0.54 |

| | Percentile rank for Reafect | Percentile rank for Reafect with bidirectional reactions |
|---|---|---|
| MMUT (S01) | 1.69 | 8.8 |
| MMUT (S39) | 10.15 | 21.03 |
| MMUT (S69) | 0.0 | 1.05 |
| MTHFR (S64) | 3.26 | 6.34 |
| MVK (S02) | 34.33 | 34.33 |
| OAT (S11) | 1.08 | 0.72 |
| OGDH (S70) | 1.88 | 5.84 |
| OTC (S36) | 12.38 | 3.88 |
| OTC (S38) | 8.21 | 14.18 |
| OTC (S47) | 15.58 | 15.94 |
| PAH (S00) | 0.0 | 0.0 |
| PAH (S08) | 0.0 | 0.0 |
| PAH (S29) | 0.55 | 0.55 |
| PAH (S34) | 0.18 | 0.18 |
| PAH (S44) | 3.07 | 1.15 |
| PCCA (S05) | 0.77 | 1.34 |
| PCCA (S66) | 2.44 | 2.96 |
| PDHA1 (S59) | 4.89 | 10.87 |
| PSAT1 (S37) | 22.22 | 35.44 |
| PSAT1 (S41) | 1.34 | 16.48 |
| TYMP (S28) | 0.0 | 0.0 |
| XDH (S30) | 0.0 | 0.0 |

**Fig. A4.** Comparison of the percentile ranks between *Reafect* and *Reafect with bidirectional reactions* per IEM patient.

### *A.7. CADD scores for ES and pathogenic variants*

The distribution of CADD scores for all variants (in metabolic genes) found in each ES background are displayed in Fig. A5A. Additionally, we show the CADD scores for the disease-causing variants for the 28 IEM patients included in Table 1. Fig. A5B shows the distribution of CADD scores (Phred) for the affected genes (i.e. the genes for which a pathogenic variant is found in one of the 28 IEM patients) in the 15 ES backgrounds. There was only one case where a CADD score from one of the ES backgrounds was higher than the variant of a IEM patient.

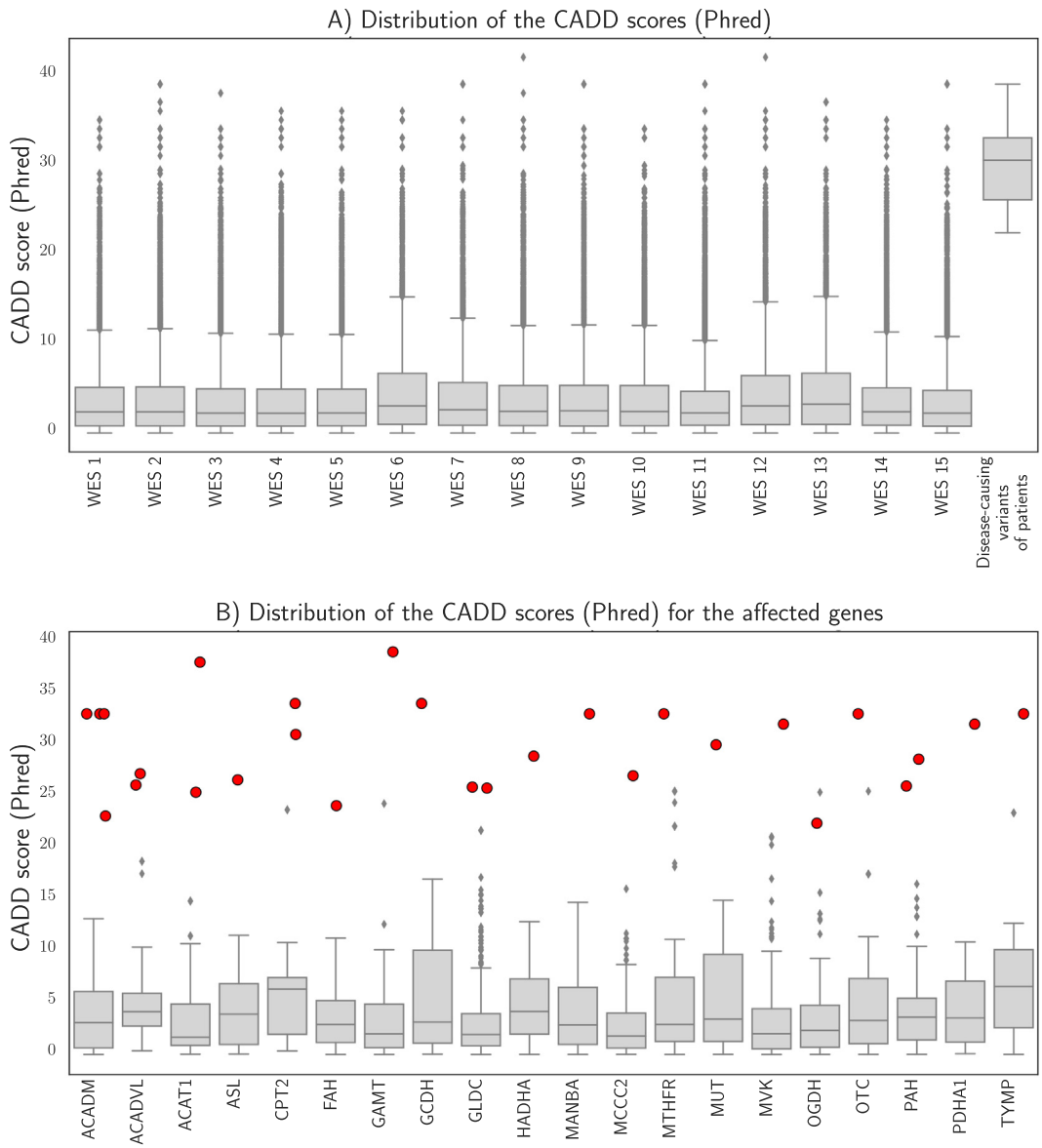


**Fig. A5. A)** Each boxplot indicates the distribution of the CADD (Phred) scores for variants in metabolic genes obtained in 15 random ES files. The last boxplot shows the CADD scores (Phred) for the disease-causing variants found in the IEM patients (Table 1). **B)** Each boxplot indicates the distribution of the CADD (Phred) scores found in the 15 ES backgrounds for the affected genes. The red dots indicate the CADD scores (Phred) for the variants in the 28 IEM patients.

### *A.8. Contribution of subtle metabolite Z-scores on IEM ranking*

We explored the contribution of more subtle metabolite Z-scores to the IEM ranking performance of *Reafect*. This was investigated by creating performance curves for various Z-score cutoffs, where we included only metabolite Z-scores for which |Z-score| < cutoff (Fig. A6A) or |Z-score| > cutoff (Fig. A6B). These results show that for decreasing cutoff values and |Z-score| < cutoff, the overall performance on IEM ranking also declines. This can be understood by realizing that for decreasing cutoff values, also more informative (disease-related) metabolites are excluded. More importantly, we observe that even for the lower cutoff values the overall performance is still positive (above the diagonal line), suggesting that more subtle metabolite Z-scores also contribute to IEM ranking. The same conclusion can be drawn from the experiment where we included only metabolites having a |Z-score| > cutoff. When increasing the cutoff values, we observe that the IEM ranking performance also decreases. Since in these cases only more extreme Z-scores are available for ranking, we conclude that more subtle metabolite Z-scores normally also contribute to IEM ranking.

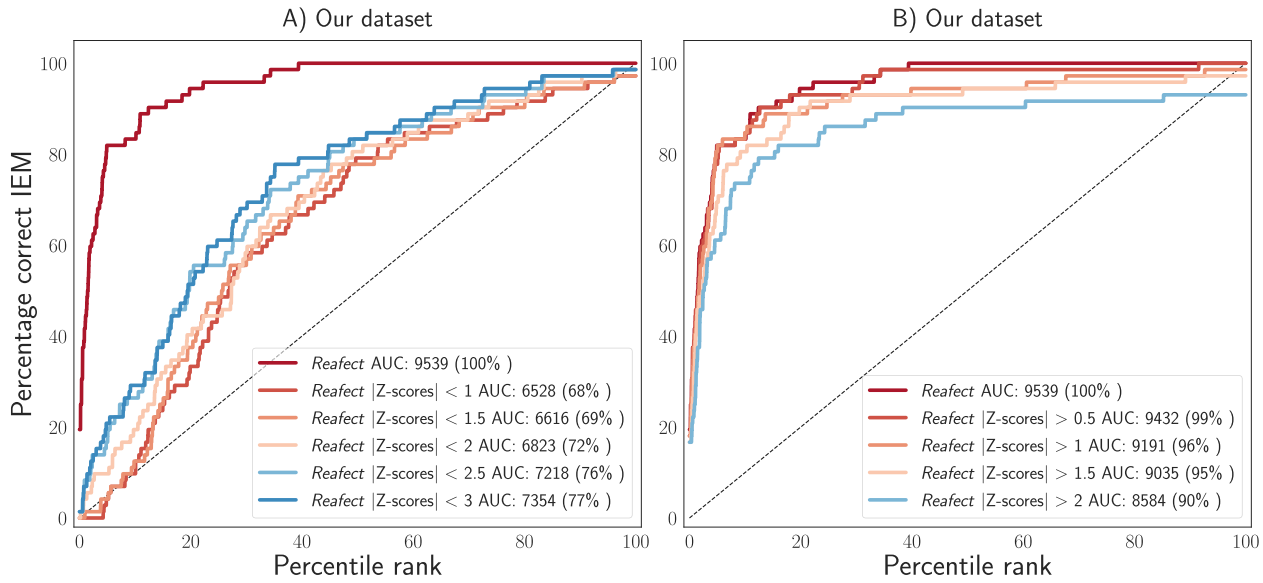


**Fig. A6. A)** Full performance curves for *Reafect* for various Z-score cutoff values, and |Z-score| < cutoff. Cutoff values are indicated by the legend. **B)** Full performance curves for Reafect for various Z-score cutoff values, and |Z-score| > cutoff.

## References


[1] E. Pronicka, D. Piekutowska-Abramczuk, E. Ciara, J. Trubicka, D. Rokicki, A. Karkucińska-Wieckowska, M. Pajdowska, E. Jurkiewicz, P. Halat, J. Kosińska, A. Pollak, M. Rydzanicz, P. Stawinski, M. Pronicki, M. Krajewska-Walasek, R. Płoski, New perspective in diagnostics of mitochondrial disorders: two years' experience with whole-exome sequencing at a national paediatric Centre, J. Transl. Med. 14 (1) (2016) https://doi.org/10.1186/s12967-016-0930-9.

[2] C.F. Wright, D.R. FitzPatrick, H.V. Firth, Paediatric genomics: diagnosing rare disease in children, Nat. Rev. Genet. 19 (5) (2018) 253–268, https://doi.org/10.1038/nrg.2017.116.

[3] D.J. Stavropoulos, D. Merico, R. Jobling, S. Bowdin, N. Monfared, B. Thiruvahindrapuram, T. Nalpathamkalam, G. Pellecchia, R.K.C. Yuen, M.J. Szego, R.Z. Hayeems, R.Z. Shaul, M. Brudno, M. Girdea, B. Frey, B. Alipanahi, S. Ahmed, R. Babul-Hirji, R.B. Porras, M.T. Carter, L. Chad, A. Chaudhry, D. Chitayat, S.J. Doust, C. Cytrynbaum, L. Dupuis, R. Ejaz, L. Fishman, A. Guerin, B. Hashemi, M. Helal, S. Hewson, M. Inbar-Feigenberg, P. Kannu, N. Karp, R.H. Kim, J. Kronick, E. Liston, H. MacDonald, S. Mercimek-Mahmutoglu, R. Mendoza-Londono, E. Nasr, G. Nimmo, N. Parkinson, N. Quercia, J. Raiman, M. Roifman, A. Schulze, A. Shugar, C. Shuman, P. Sinajon, K. Siriwardena, R. Weksberg, G. Yoon, C. Carew, R. Erickson, R.A. Leach, R. Klein, P.N. Ray, M.S. Meyn, S.W. Scherer, R.D. Cohn, C.R. Marshall, Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. Npj, Genom. Med. 1 (1) (2016) https://doi.org/10.1038/npjgenmed.2015.12.

[4] P. Rentzsch, D. Witten, G.M. Cooper, J. Shendure, M. Kircher, CADD: predicting the deleteriousness of variants throughout the human genome, Nucleic Acids Res. 47 (D1) (2018) 886–894, https://doi.org/10.1093/nar/gky1016.

[5] M.H.P.M. Kerkhofs, H.A. Haijes, A.M. Willemsen, K.L.I. van Gassen, M. van der Ham, J. Gerrits, M.G.M. de Sain-van der Velden, H.C.M.T. Prinsen, H.W.M. van Deutekom, P.M. van Hasselt, N.M. Verhoeven-Duif, J.J.M. Jans, Cross-omics: Integrating genomics with metabolomics in clinical diagnostics, Metabolites 10 (5) (2020) 206, https://doi.org/10.3390/metabo10050206.

[6] J.T. Alaimo, K.E. Glinton, N. Liu, J. Xiao, Y. Yang, V.R. Sutton, S.H. Elsea, Integrated analysis of metabolomic profiling and exome data supplements sequence variant interpretation, classification, and diagnosis, Genetics in Med. (2020) https://doi.org/10.1038/s41436-020-0827-0.

[7] E.J.G. Linck, P.A. Richmond, M. Tarailo-Graovac, U. Engelke, L.A.J. Kluijtmans, K.L.M. Coene, R.A. Wevers, W. Wasserman, C.D.M. van Karnebeek, S. Mostafavi, metPropagate: network-guided propagation of metabolomic information for prioritization of metabolic disease genes. Npj, Genom. Med. 5 (1) (2020) https://doi.org/10.1038/s41525-020-0132-5.

[8] H.A. Haijes, M. van der Ham, H.C.M.T. Prinsen, M.H. Broeks, P.M. van Hasselt, de Sain-van der Velden, M.G.M., Verhoeven-Duif, N.M., Jans, J.J.M., Untargeted metabolomics for metabolic diagnostic screening with automated data interpretation using a knowledge-based algorithm, Int. J. Mol. Sci. 21 (3) (2020) 979, https://doi.org/10.3390/ijms21030979.

[9] C. Baumgartner, C. Bohm, D. Baumgartner, G. Marini, K. Weinberger, B. Olgemoller, B. Liebl, A.A. Roscher, Supervised machine learning techniques for the classification of metabolic disorders in newborns, Bioinformatics 20 (17) (2004) 2985–2996, https://doi.org/10.1093/bioinformatics/bth343.

[10] D. Waters, D. Adeloye, D. Woolham, E. Wastnedge, S. Patel, I. Rudan, Global birth prevalence and mortality from inborn errors of metabolism: a systematic analysis of the evidence. Journal of, Glob. Health 8 (2) (2018) https://doi.org/10.7189/jogh.08.021102.

[11] G.M. Messa, F. Napolitano, S.H. Elsea, D. di Bernardo, X. Gao, A siamese neural network model for the prioritization of metabolic disorders by integrating real and simulated data, Bioinformatics 36 (Supplement_2) (2020) 787–794, https://doi.org/10.1093/bioinformatics/btaa841.

[12] S. Li, Y. Park, S. Duraisingham, F.H. Strobel, N. Khan, Q.A. Soltow, D.P. Jones, B. Pulendran, Predicting network activity from high throughput metabolomics, PLoS Comput. Biol. 9 (7) (2013) 1003123, https://doi.org/10.1371/journal.pcbi.1003123.

[13] L. Pirhaji, P. Milani, M. Leidl, T. Curran, J. Avila-Pacheco, C.B. Clish, F.M. White, A. Saghatelian, E. Fraenkel, Revealing disease-associated pathways by network integration of untargeted metabolomics, Nat. Methods 13 (9) (2016) 770–776, https://doi.org/10.1038/nmeth.3940.

[14] M. Kanehisa, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30, https://doi.org/10.1093/nar/28.1.27.

[15] R. Bonte, M. Bongaerts, S. Demirdas, J.G. Langendonk, H.H. Huidekoper, M. Williams, W. Onkenhout, E.H. Jacobs, H.J. Blom, G.J.G. Ruijter, Untargeted metabolomics-based screening method for inborn errors of metabolism using semi-automatic sample preparation with an UHPLC- orbitrap-MS platform, Metabolites 9 (12) (2019) 289, https://doi.org/10.3390/metabo9120289.

[16] M. Bongaerts, R. Bonte, S. Demirdas, E.H. Jacobs, E. Oussoren, A.T. van der Ploeg, M.A.E.M. Wagenmakers, R.M.W. Hofstra, H.J. Blom, M.J.T. Reinders, G.J.G. Ruijter, Using out-of-batch reference populations to improve untargeted metabolomics for screening inborn errors of metabolism, Metabolites 11 (1) (2020) 8, https://doi.org/10.3390/metabo11010008.

[17] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in1h NMR metabonomics, Anal. Chem. 78 (13) (2006) 4281–4290, https://doi.org/10.1021/ac051632c.

[18] M.J. Miller, A.D. Kennedy, A.D. Eckhart, L.C. Burrage, J.E. Wulff, L.A.D. Miller, M.V. Milburn, J.A. Ryals, A.L. Beaudet, Q. Sun, V.R. Sutton, S.H. Elsea, Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism, J. Inherit. Metab. Dis. 38 (6) (2015) 1029–1039, https://doi.org/10.1007/s10545-015-9843-7.

[19] A. Noronha, J. Modamio, Y. Jarosz, E. Guerard, N. Sompairac, G. Preciat, A.D. Danelsdóttir, M. Krecke, D. Merten, H.S. Haraldsdóttir, A. Heinken, L. Heirendt, S. Magnúsdóttir, D.A. Ravcheev, S. Sahoo, P. Gawron, L. Friscioni, B. Garcia, M. Prendergast, A. Puente, M. Rodrigues, A. Roy, M. Rouquaya, L. Wiltgen, A. Žagare, E. John, M. Krueger, I. Kuperstein, A. Zinovyev, R. Schneider, R.M.T. Fleming, I. Thiele, The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease, Nucleic Acids Res. 47 (D1) (2018) 614–624, https://doi.org/10.1093/nar/gky992.

[20] H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760, https://doi.org/10.1093/bioinformatics/btp324.

[21] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (9) (2010) 1297–1303, https://doi.org/10.1101/gr.107524.110.

[22] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (16) (2010) 164, https://doi.org/10.1093/nar/gkq603.

[23] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W.W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H.L. Rehm, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology, Genet. Med. 17 (5) (2015) 405–424, https://doi.org/10.1038/gim.2015.30.