METEOR: Evidence-weighted reaction selection reconciles enzyme prediction with growth feasibility

Kexin Niu[1] and Robert Hoehndorf[1,2,3,4,†]

[1]*Biological and Environmental Sciences & Engineering (BESE) Division,* [2]*KAUST Center of Excellence for Smart Health (KCSH),* [3]*KAUST Center of Excellence for Generative AI,* [4]*Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology, 4700 KAUST, Thuwal 23955, Kingdom of Saudi Arabia*

[†]*E-mail: robert.hoehndorf@kaust.edu.sa*

Enzyme function predictors score each protein in isolation and never ask whether the resulting assignments describe an organism that can grow. Metabolism supplies a different kind of constraint: steady-state fluxes must satisfy mass balance and can be required to support biomass production, regardless of whether a protein resembles anything already characterised. Here we present METEOR, which aggregates predicted per-protein Enzyme Commission (EC) scores into reaction-level scores and solves a mixed-integer linear program (MILP) over the SEED universal reaction database to return a growth-feasible reaction set. A penalty $\mu(1 - w_j)^p$ charges each reaction by the confidence behind it, a post-solve check repairs solutions that do not grow on their own, and a bounded per-EC update maps the selected context back to protein scores. On six curated models METEOR recovers 60.4% of curated ECs that DeepProZyme scores below 0.5, against 48.0% for thresholding followed by parsimonious repair and 48.3% for a size-matched top-scoring baseline, while EC precision falls from 45.5% to 42.2%. When predictor scores are deleted, the confidence-derived cost admits 4–58× fewer off-reference reactions than confidence-independent or shuffled costs. Across three predictors and 108 bacterial genomes, METEOR models grow, unlike thresholded networks; a candidate-mask ablation attributes growth to the shared growth-feasible scaffold and the choice of reactions that join it to evidence weighting. Network constraints operate at the organism level: METEOR returns a growth-feasible reaction set but improves neither size-matched pathway coverage nor individual per-protein accuracy.

Keywords: Enzyme function prediction; Genome-scale metabolic reconstruction; Constraint-based modelling; Mixed-integer linear programming; Parsimony.

1. Introduction

Enzyme function annotation has moved from homology transfer[1–3] to sequence-based deep learning, including contrastive,[4] CNN,[5,6] and transformer[7] architectures. Accuracy has improved substantially, but the unit of inference has not: these methods score each protein on its own sequence against what the training set contains, treating two proteins from the same genome independently and never asking whether the resulting assignments describe an organism that could exist.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Metabolism supplies a constraint that these predictors ignore. A feasible steady state must balance the production and consumption of every internal metabolite,[8] and EC assignments made one protein at a time need not satisfy it. A growth requirement can also favour a weakly scored reaction when that reaction completes a feasible route. Such a constraint informs the organism-level reaction set. Raising or lowering a confidence threshold does not impose it, and GLOBUS,[9] which adds genomic context, does not test whether the resulting assignment is metabolically consistent.

Two families of methods already couple reaction evidence with stoichiometric constraints. Context-specific extraction methods start from a generic model and reaction evidence that is already classified: FASTCORE extracts a flux-consistent subnetwork around a predefined core,[10] and CORDA retains reactions by confidence class together with the support they need.[11] Reconstruction tools such as CarveMe, gapseq and Reconstructor[12–14] commonly build a draft model and gap-fill it, adding reactions until the model grows;[15] ProbAnno ranks gap-fill candidates by homology-derived likelihood,[16] and other methods by subsystem, pathway or genomic context.[17–22] In these two-stage workflows the draft is fixed before repair: accepted reactions are never revisited, and rejected ones return only when growth demands them. Where likelihoods do become costs, as in ProbAnno, they apply only to reactions outside a fixed draft; none of these methods selects the whole reaction set under one confidence-derived cost or returns its result as protein scores.

METEOR makes the two decisions at once. It aggregates continuous per-protein EC scores into reaction-level costs and selects a growth-feasible reaction set from the SEED universal reaction database[23,24] in a single biomass-constrained optimisation, in which candidates compete under a cost that penalises weak support more than strong support (Section 2.2). The selected set is then written back to the protein–EC score matrix through a bounded update.

On 108 bacterial genomes, METEOR's selected sets grow where thresholded networks do not, a property the shared scaffold supplies (Section 3.1). METEOR recovers curated ECs that a fixed threshold discards, with higher recovery than a size-matched baseline and a measured precision cost (Section 3.2), and among the many growth-feasible sets the sequence-derived cost determines which is selected (Section 3.3). It improves neither size-matched pathway coverage nor per-protein accuracy (Sections 3.2 and 3.4). METEOR is predictor-agnostic and is freely available at `https://github.com/bio-ontology-research-group/Meteor/tree/v8.3-psb2027`.

2. Methods

2.1. *Datasets and benchmark construction*

Protein-level holdout. We built a temporal UniProtKB/Swiss-Prot[25] holdout of 227 bacterial proteins created between 2025-01-01 and 2026-01-28 (UniProt 2026_01), all postdating the evaluated predictors' training data. Among them, 125 have complete four-digit EC annotations across 92 classes. We linked proteins to genomes by strain-level `Proteomes` cross-reference, then manifest assembly, then exact sequence match within that proteome, preventing cross-organism matches from identical sequences. This yielded 66 proteins across 53 genomes for protein-level evaluation (metrics in Supplementary Table S2); we excluded 59 proteins for

missing strain proteomes, off-manifest assemblies, or absence from the genome panel.

Organism panel. Of the 227 holdout proteins, 208 map to an NCBI RefSeq assembly, covering 112 distinct assemblies. We excluded four assemblies with incomplete baseline coverage, yielding a panel of 108 bacterial genomes from 69 genera (63 Gram-negative, 45 Gram-positive). All organism-level analyses use this 108-genome panel; the 53 genomes containing a genome-resolved holdout protein also support the protein-level evaluation. Baseline predictions were matched to each proteome by exact sequence hash. Because this panel is derived from the holdout rather than sampled to represent bacterial diversity, organism-level claims are restricted to these 108 genomes.

METEOR uses the ModelSEED `Prokaryote_Universal` reaction database distributed with Reconstructor v1.1.0 (47,880 reactions, 23,013 metabolites).[14,23,24] We evaluated canonical pathway completeness with KEGG pathway-to-EC mappings[26] accessed on 2026-06-30. Of the 172 KEGG Metabolism pathways we retained the 151 with at least five SEED-representable four-digit ECs, the smallest size at which our three stringencies (≥1 EC, ≥25%, ≥50%) require distinct EC counts.

Baseline predictions enter METEOR in each method's native score space, except for DeepProZyme,[6] whose EC-specific decision boundaries are mapped to τ=0.5 by per-EC piecewise-linear calibration. CLEAN[4] and EnzBERT[7] use a single native threshold. We use these bounded scores as confidence inputs, although they are not calibrated as comparable probabilities; reaction confidence therefore inherits each predictor's score scale. Supplementary Section S2 describes sequence-identity-filtered retraining.

2.2. *METEOR framework*

METEOR (Figure 1b) takes the score matrix $E \in [0,1]^{N\times C}$ of one organism's N proteins over C ECs and returns a selected reaction set and a refined matrix E^{new} of the same shape. It also exports the EC annotations of selected reactions as a separate organism-level set, which may extend beyond the input vocabulary. *Evidence* means model-derived predictor confidence: a protein–EC score or its reaction-level aggregate w_j.

(1) *Noisy-OR aggregation.* For each of the J = 47,880 SEED reactions j, let $M(j)$ denote its possibly empty set of EC annotations. Protein-level evidence is aggregated as

$$w_j = 1 - \prod_{p\in\mathcal{P},\, e\in M(j)} (1 - E_{p,e}). \tag{1}$$

We keep only each protein's top $k = 5$ EC scores before aggregation and set the rest to zero, limiting inflation by small score tails (sensitivity in Supplementary Section S5).

(2) *Evidence-weighted cost mapping.* Each reaction confidence is converted to a linear cost

$$c_j = -\log \frac{w_j + \epsilon}{1 - w_j + \epsilon} + \mu(1 - w_j)^p, \tag{2}$$

a log-odds evidence term plus an evidence-weighted parsimony penalty; at the defaults the cost is negative above $w_j \approx 0.61$ (Figure 1b, step 2). The exponent p controls how sharply the penalty concentrates on weak evidence; $p = 0$ charges every reaction the same μ, the

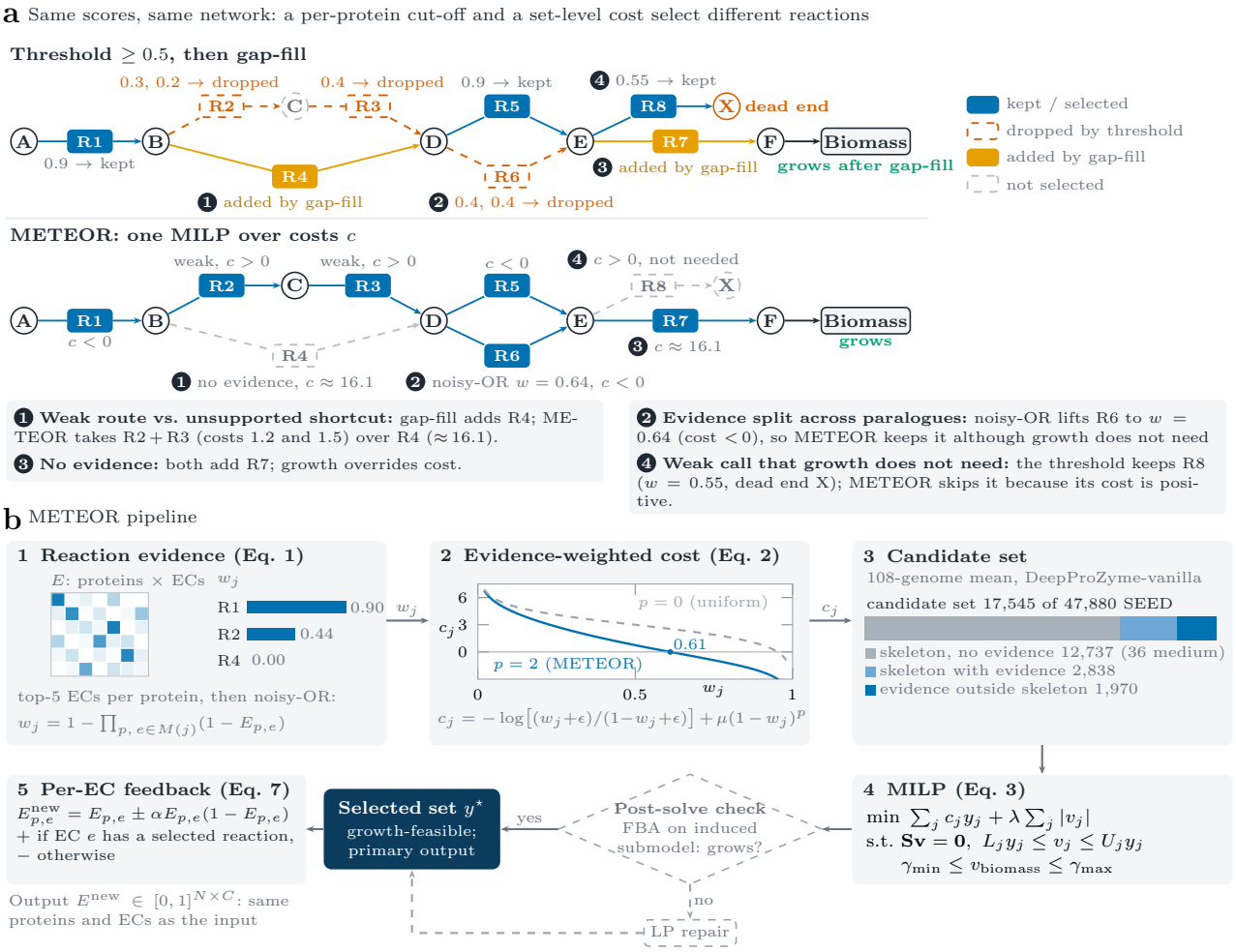


Fig. 1. METEOR vs. threshold-then-gap-fill. (a) Toy network with invented scores: from the same evidence, a per-protein cut-off and a set-level cost select different reactions (cases 1–4). (b) The pipeline; step 3 shows per-genome means on the DeepProZyme-vanilla panel. The selected set $y^\star$ is the primary output and is written back to E^{new}.

confidence-independent control of Section 3.3. The cost can retain evidence-supported alternative routes while discouraging unsupported ones (Figure 1a; Section 3.1). It cannot preserve or distinguish isozymes because the optimisation contains no protein-level variables. A reaction without evidence costs about 16.1 at the defaults (implementation details in Supplementary Section S1.1).

(3) *Metabolic network optimisation.* The MILP runs over a per-genome candidate set rather than all 47,880 SEED reactions. It contains the reactions with evidence $w_j \geq 0.01$ in that genome, the *medium reactions* (exchange reactions that supply the defined growth medium across the system boundary), and a biomass-feasible skeleton. The skeleton is computed once per Gram group by maximising biomass flux over the universal model and is shared by every genome in the group, so its reactions are selectable without evidence in a given organism; about 12,700 of the 17,545 candidates per genome are of this kind (Figure 1b, step 3; Supplementary Section S1.1).

Let $y_j \in \{0,1\}$ indicate reaction inclusion, $v_j \in \mathbb{R}$ its steady-state flux, $\mathbf{v}$ the flux vector over all candidate reactions, and $\mathbf{S}$ the stoichiometric matrix (metabolites by reactions). METEOR selects the reaction set by solving

$$\min_{y,v} \quad \sum_j c_j y_j + \lambda \sum_j |v_j| \tag{3}$$

$$\text{s.t.} \quad \mathbf{Sv} = \mathbf{0}, \tag{4}$$

$$L_j y_j \le v_j \le U_j y_j \quad \forall j, \tag{5}$$

$$\gamma_{\min} \le v_{\text{biomass}} \le \gamma_{\max}. \tag{6}$$

The objective minimises evidence-derived reaction cost together with a flux parsimony term, subject to mass balance (each internal metabolite is produced as fast as it is consumed), inclusion-gated flux bounds and a biomass-flux interval ($\gamma_{\max} = 2.5$). Medium reactions have $y_j = 1$; non-candidate reactions have $y_j = 0$. With CBC[27] at a 5% relative gap and a 600 s limit, 97 of 108 DeepProZyme-vanilla solve calls took less than 600 s and 11 took at least 600 s, consistent with time-limited runs. The solve call (including solver input and output) took a median of 137 s per genome (range 72–738 s) on four cores. Across all twelve configurations 624 of 1,296 solve calls took at least 600 s, most for CLEAN (Supplementary Table S1); settings, media and fixed reactions are in Supplementary Section S1.1.

(4) *Verify-and-repair.* The flux-selection coupling constraints $L_j y_j \le v_j \le U_j y_j$ operate as standard big-M bounds.[28] Because parsimony penalises skeleton reactions without evidence, the solver can leave them at $y_j \approx 0$ within its integrality tolerance while $U_j \approx 1000$ still passes the small flux the biomass floor needs. Consequently, the induced submodel (only the selected reactions and exchange reactions) may fail independent growth verification despite formal MILP feasibility. After each solve we run flux balance analysis (FBA) on the induced submodel and require $\max v_{\text{biomass}} \ge 0.05$. Where this fails, we add the support (the reactions carrying nonzero flux) of a linear program (LP) that minimises flux through reactions not already selected while satisfying the biomass floor. Every growth count in this paper refers to the repaired set (repair frequency in Section 3.1).

(5) *Constraint feedback propagation.* For each EC e, let $A_e = \{j : e \in M(j) \text{ and } y_j^\star = 1\}$ be the selected reactions associated with e. We update per-protein scores

$$E_{p,e}^{\text{new}} = \begin{cases} E_{p,e} + \alpha\, E_{p,e}(1 - E_{p,e}) & \text{if } A_e \neq \emptyset, \\ E_{p,e} - \alpha\, E_{p,e}(1 - E_{p,e}) & \text{if } A_e = \emptyset, \end{cases} \tag{7}$$

with step size α; the $E_{p,e}(1-E_{p,e})$ factor keeps updates within $[0,1]$. All proteins sharing an EC receive the same monotone transformation, preserving their ordering within that EC column but not necessarily the ordering of ECs within a protein. A reaction with several ECs (8.2% of EC-annotated SEED reactions) boosts all of them when selected; the method attributes nothing to a specific protein or paralogue.

We fixed ϵ=10^{-6}, $\gamma_{\min}$=0.1, λ=10^{-4} and α=0.4 before holdout evaluation. We chose p=2 and μ=3 on panel and holdout metrics, so the holdout is not an independent test of these two (sensitivity analyses in Supplementary Section S5).

Structural metrics. Dead-end metabolites (produced but never consumed, or the reverse) and mass-imbalanced reactions are counted with MEMOTE[29] on each selected submodel, identically for both methods. The dead-end count also includes metabolites an organism legitimately only secretes or takes up, so it is a structural diagnostic whose biological interpretation depends on exchange and secretion.

3. Results

3.1. *Growth feasibility comes from the shared scaffold*

For each predictor we compared a baseline network, containing every reaction with at least one protein score above $\tau = 0.5$, against METEOR's selected set across the 108-genome panel (Figure 2a; per-predictor selected-reaction counts and mass-imbalance fraction in Supplementary Section S7.7; Figure 2 also previews additional recovery and stability results from Sections 3.2–3.3).

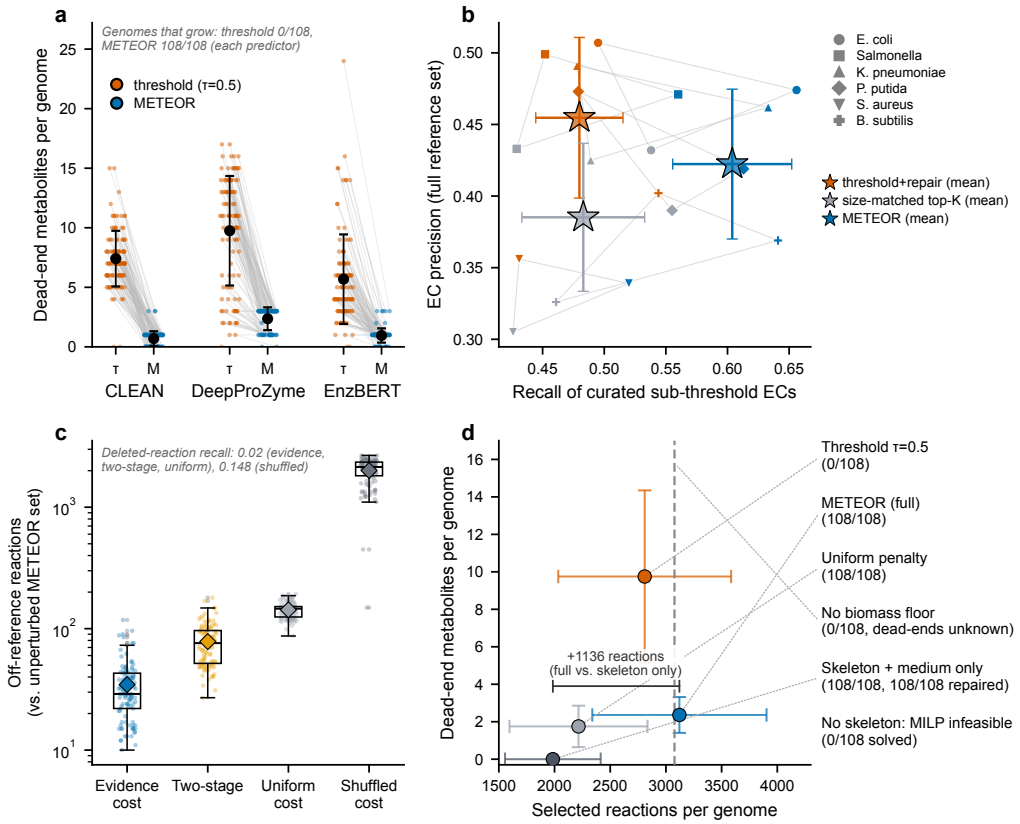


Fig. 2. Properties of METEOR-selected networks (mean ± s.d.; small points are genomes or organisms). **(a)** Dead ends per genome, threshold ($\tau = 0.5$, does not grow) vs. METEOR; a threshold network repaired to grow has a similar count (Section 3.1). **(b)** Recall of curated sub-threshold ECs vs. EC precision, six curated models; stars are method means. **(c)** Off-reference reactions under four deleted-evidence cost regimes (log scale; boxes: IQR; diamonds: mean). **(d)** Selected reactions vs. dead ends for five DeepProZyme ablations; open markers do not grow.

No thresholded network grows even with every exchange open: FBA on the submodel returns zero biomass flux for every predictor, and about half of the reactions in each submodel fail elemental-balance checks (Figure 2a; Supplementary Section S7.7).

Relative to the raw thresholded networks, mean dead-end counts fall by 76–91% (paired Wilcoxon $p < 10^{-18}$), with fewer dead ends in 104–106 of 108 genomes and ties in the rest. The candidate-mask ablation attributes this reduction to growth-feasible construction: the skeleton alone yields growing sets with no dead ends, and METEOR's 2.4 dead ends per genome (DeepProZyme) arrive with the evidence-supported reactions added to the skeleton (Section 3.3). A threshold network repaired to grow with the same LP (Section 3.2) has slightly fewer dead ends (2.23 against 2.36 per genome; equal in 95 of 108, METEOR higher in 11); METEOR's lower dead-end fraction (0.84 against 0.90 per 1,000 metabolites) reflects only its larger network (2,723 against 2,381 metabolites). The parsimony penalty controls size: without the $\mu(1 - w_j)^p$ term, log-odds evidence alone selects a mean of 5,966 reactions per genome against 3,121. About half of the reactions in every submodel fail elemental balance whichever method selects them, and the imbalanced fraction moves only from 0.49–0.51 to 0.46–0.50 (0.478 for the repaired threshold network against METEOR's 0.464). METEOR retains the SEED stoichiometry of each selected reaction; the small shift reflects which reactions are selected. Reaction selection ($y_j = 1$) permits both zero and nonzero flux. Flux variability analysis shows that 54.5% of METEOR's selected reactions can carry nonzero flux on average, compared with 40.5% in thresholded networks. METEOR improves both the flux-consistent fraction and absolute count in all 108 genomes (paired Wilcoxon $p = 1.9 \times 10^{-19}$ for both comparisons). In a single biomass-maximising flux solution only 13.3% of selected reactions carry flux, so the selected set describes candidate reactions across possible flux states.

Growth feasibility is imposed by the MILP but counted only after post-solve verify-and-repair (Section 2.2), which 77 of 1,296 raw solutions (5.9%) across the twelve configurations require and which adds a median of 98 reactions (range 1–230; Supplementary Table S1). After repair, all 108 METEOR submodels grow under the defined medium. Exported in the Systems Biology Markup Language (SBML) and re-tested with MEMOTE under a medium it infers per model, the repaired models grow in 105–108 of 108 genomes in the vanilla configurations and 83–108 across all twelve. MEMOTE's unbounded-flux fraction (lower is better) is 0.73–0.94 for METEOR against 0.545 for CarveMe on the same panel, a consequence of the universal template that METEOR does not correct (Supplementary Sections S7.1 and S7.2).

Thresholded networks are repairable by the same flux-parsimony LP, which restores growth in every genome (Section 3.3). The sequence-derived cost determines which feasible selection is returned. We next assess those selections against curated reference models.

3.2. *METEOR recovers curated sub-threshold ECs at a measured precision cost*

A hard threshold at τ=0.5 discards every EC scored below it, including ECs the organism carries. We measure that loss against curated reference models. For each of six curated bacterial genome-scale metabolic models (GEMs),[30–35] we collect the ECs annotated by the reference model that DeepProZyme scores between 0 and 0.5, then measure the recovery rate for each

Table 1. Recovery of sub-threshold ECs across six curated bacterial GEMs (*E. coli*, *Salmonella*, *K. pneumoniae*, *P. putida*, *S. aureus*, *B. subtilis*); denominator: reference ECs that DeepProZyme scores in $(0, 0.5)$, 102 ± 16 per organism, 612 in total. Mean ± s.d.; precision against the full reference EC set.

	Threshold + parsimonious repair	METEOR
Recall, curated sub-threshold ECs (n=612)	0.480 ± 0.039	$\mathbf{0.604 \pm 0.053}$
Precision, full EC set	0.455 ± 0.061	0.422 ± 0.057

method. This is the set that standard thresholding drops despite nonzero predictive evidence. The baseline thresholds the same predictor at 0.5 and repairs the draft model by minimising flux through external reactions, using the identical reaction database, medium, and biomass constraints; it differs from METEOR's binary selection in its two-stage objective and candidate pool (Supplementary Section S7.4).

METEOR recovers more of these ECs at a moderate precision cost (Table 1). The recall gain holds across all six organisms ($+0.124 \pm 0.030$), while the precision loss is smaller and equally consistent (-0.032 ± 0.012). Pooling across organisms rather than averaging yields the same trend, 368/612 against 293/612. Measured on DeepProZyme alone and on six genomes, these data apply only to this comparison.

The additional output contains a small fraction of reference-supported ECs. Across the six models METEOR emits 773 more ECs than the baseline (1,402 against 1,274 per model). Of these, 77 match curated annotation, 75 of those being curated sub-threshold ECs, yielding a marginal precision of 0.10 for the extra calls. Many remaining predictions fall outside the reference EC vocabulary (Section 3.4), where the comparison can neither confirm nor refute them. In a size-matched comparison, a baseline of the top-K ECs by score, with K equal to METEOR's EC count, recovers 0.483 of the sub-threshold set at precision 0.385, compared with METEOR's 0.604 at 0.422, and is lower in all six organisms (Supplementary Section S7.5).

The recovered ECs are specific capabilities such as dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133, O-antigen assembly; checked against ENZYME[36] and KEGG[26]); every recovered EC is listed in Supplementary Section S7.5.

Pathway coverage depends on the number of ECs selected. Against the raw threshold METEOR gains 4.7–9.1 percentage points of mean KEGG pathway coverage, but against the size-matched top-K baseline it loses 2.7–2.8 points, in 107 of 108 genomes for CLEAN and in all 108 for DeepProZyme and EnzBERT (Supplementary Section S3.1). This comparison attributes the threshold-relative gain to output size.

3.3. *The sequence-derived cost chooses among feasible sets*

Many reaction sets satisfy the growth constraint equally well; when several reactions could supply the same biomass precursor, only the evidence separates them. To quantify this reliance on evidence, we remove predictive evidence and measure the resulting shift in the selected reaction set. For each genome we zero the EC scores behind 60 randomly selected, evidence-bearing, non-skeleton reactions (fixed seed; this may also strip evidence from co-annotated reactions) and re-solve over the same candidate set under four regimes: the evidence-derived

cost, a uniform cost ($p = 0$), shuffled evidence-derived costs, and a two-stage threshold-then-weighted-gap-fill of the kind ProbAnno[16] uses, run on METEOR's own costs and candidates. We count *off-reference* reactions, those absent from the unperturbed METEOR set. The reference is METEOR's own output, so this experiment measures selected-set stability under evidence deletion.

All four regimes grow, but they select very different sets (Figure 2c). The evidence-derived cost admits 34.6 off-reference reactions per genome, against 78.2 for the two-stage arm (2.3×), 142.9 for uniform cost (4.1×) and 2,006.6 for shuffled costs (58×), with the same ordering in both Gram groups (Supplementary Section S9). The uniform and shuffled arms keep METEOR's binary objective and change only the costs, so their gap isolates the evidence-to-cost mapping. The two-stage arm also changes the objective (costs per unit of flux outside a fixed draft), so its count is descriptive. Recovery of deleted reactions is low: shuffled costs reach 0.148 while adding 2,015 reactions. The comparison quantifies changes in the selected set.

Ablations on DeepProZyme (108 genomes; Figure 2d) show which component supplies what. Feasibility comes from the biomass floor and the skeleton: without the floor the set barely changes in size (3,077 vs. 3,121 reactions) but no submodel grows (0/108), and without the skeleton the MILP is infeasible in every genome. The skeleton with the medium alone already yields growing sets of 1,985 reactions with no dead ends, all needing repair (Supplementary Table S7). Evidence weighting decides what is added on top: the full cost selects 1,136 more reactions (3,121 against 1,985) at 2.4 dead ends, while a uniform penalty in place of $\mu(1 - w_j)^2$ keeps growth but shrinks the network to 2,215 reactions and retains less of the nonzero sub-threshold evidence (10.8% against 19.1%).

3.4. *Protein-level interface and evaluation on six curated organisms*

The confidence update step projects reaction-level selections back onto a protein–EC score matrix sharing the original vocabulary. Because its monotone per-EC transformation can still reorder ECs within a protein, we test protein-level predictions, mapping baseline and METEOR through the identical EC vocabulary (*net-only*). On the 66 temporal-holdout proteins, METEOR alters no top-1 call for any predictor and $F_{\max}$ shifts by at most 0.012; the same holds in all twelve baseline×cut-off configurations (Supplementary Table S2). For DeepProZyme on 69,666 NCBI-annotated proteins, top-1 shifts from 90.2% to 90.1%. On the temporal holdout, micro-F_1@0.5 changes by +0.015 for CLEAN, +0.008 for DeepProZyme and 0.000 for EnzBERT. Full-proteome results are stratified by training-set identity in Supplementary Section S8.

Against published BiGG GEMs for the six curated organisms of Section 3.2, METEOR matches CarveMe[12] on recall of the ECs annotated to selected reactions (0.884 vs. 0.893) while returning nearly twice as many ECs (1,521 vs. 818), and trails on every set-sensitive metric: precision 0.438 vs. 0.822, F1 0.585 vs. 0.854 (Supplementary Section S7.1). Much of the gap is vocabulary mismatch. METEOR emits ECs from the 6,598-EC ModelSEED mapping; the curated models and CarveMe use a 1,234-EC BiGG-derived[37] subset of it, and 76% of METEOR's apparent false positives fall outside that subset; their correctness requires a broader reference or targeted validation. Restricting both methods to the reference vocabulary raises

METEOR's precision to 0.761 against CarveMe's 0.822, narrowing the precision gap from 38 points to 6 (F1 0.817 against 0.854). CarveMe's BiGG template shares a lineage with this restricted evaluation standard; the residual gap is largest for the two Gram-positive species. Both vocabularies and every false positive are deposited.

Against a *same-predictor* baseline, constructed by thresholding at 0.5 and then minimising flux through reactions outside the draft (Supplementary Section S7.4), METEOR recovers marginally more (0.784 vs. 0.767, +0.017 mean; 6/6, p=0.031); both models are growth-feasible. This recall is lower than the 0.884 above because it scores only ECs in the predictor's vocabulary (1,402 per model), whereas the full reaction-associated EC set (1,521) also carries superseded EC numbers that SEED and the curated models still use (108–124 per organism). The reference sets are identical; the score matrix retains its input columns. The corresponding precision and sub-threshold results are reported in Section 3.2, with controls in Supplementary Section S3.

Gene-essentiality prediction shows no general advantage. Against experimentally determined essential-gene sets for the same six organisms, in eight organism-medium-reference combinations (Supplementary Section S7.6), METEOR beats the threshold-and-repair baseline only on *B. subtilis* (both media tested; mean Matthews correlation coefficient, MCC, 0.323 vs. 0.247 in rich medium), ties it on *E. coli* in minimal medium, is inconclusive on *S. aureus*, and loses on *E. coli* in rich medium and on *Salmonella*, *K. pneumoniae* and *P. putida*. The two *E. coli* cases are scored against different reference tables, and a 2×2 decomposition shows that the table, not the medium, drives the reversal: against the Keio reference[38] METEOR trails in both media (MCC 0.052 vs. 0.105 in minimal medium), and switching the medium alone moves scores by only 0.01–0.05. At the gene level, METEOR's wrong-direction calls on the medium switch under DeepProZyme are all five de novo purine genes, consistent with parsimony removing salvage routes; the effect is small. For the other three organisms, tested only under a rich medium, most of the gap has a different cause: the shared candidate pool contains reactions annotated as eukaryotic (e.g. mitochondrial nucleotide carriers) that the baseline's gap-fill admits regardless of evidence, and removing them roughly halves the MCC gap, reversing it for one organism-predictor pair. Of the reactions METEOR still declines with no alternative route in its own network, only 19% are in that organism's curated model, against 74% for reactions it keeps, which descriptively points to candidate-pool quality rather than a parsimony failure.

4. Discussion

4.1. *Reaction-level scope and protein attribution*

Two limits belong to the formulation. Evidence cannot reach ECs outside the predictor's vocabulary or lift a zero score, and unannotated proteins contribute none; only the candidate set can supply reactions for such ECs. Stratifying curated-EC recovery by vocabulary on the six curated organisms measures how much the candidate set supplies: of the 7–12% of curated ECs outside a predictor's vocabulary, METEOR's reaction-associated EC set recovers 55–71% against 31–40% for the skeleton alone (6/6 organisms for each predictor), and 2.5–4.4% of curated ECs have no candidate reaction at all (Supplementary Section S7.8). Selection also

runs post hoc as a discrete MILP, so training is separate from selection. A differentiable relaxation[39,40] with protein–reaction variables could couple these steps, at up to $O(NJ)$ variables; testing protein attribution would also require suitable labels and a defined vocabulary.

METEOR's evidence-weighted parsimony, which favours the fewest supported reactions, controls network size (Section 3.1) but has a cost for essentiality prediction: by removing redundant routes it can make a retained gene look essential when it is not, as for five purine genes in *E. coli* (Section 3.4).

4.2. *Limitations*

Several limitations bound what this paper claims. *Lack of protein attribution.* METEOR evaluates reactions at the organism level without resolving specific protein-to-reaction assignments. When a selected reaction maps to multiple ECs, all associated proteins receive uniform feedback updates, leaving paralogues indistinguishable and unannotated zero-score proteins unrefined.

Protein attribution and validation. Top-1 calls remain unchanged in every holdout configuration (Section 3.4). The evaluations use curated models, KEGG definitions, published experimental essential-gene sets, and the method's own solutions. Prospective enzyme assays and growth experiments would directly test the recovered functions.

Evidence aggregation and truncation. The noisy-OR formulation assumes independent, probability-like confidence inputs, whereas raw predictor scores are uncalibrated and biological functional annotations often exhibit inter-protein correlations. Reaction confidence can consequently increase with proteome or paralogue count. Although top-k ($k = 5$) truncation limits this effect, the available k sweep predates the final evidence-weighted formulation and shows an approximately twofold change in selected network size (Supplementary Section S5).

Reaction mappings and skeleton. EC-to-reaction mappings in the SEED universal contain errors that propagate into the MILP, and SEED sustains thermodynamically infeasible cycles relative to curated references.[41] The biomass-feasible skeleton is the support of one biomass-maximising LP solution per Gram group, not per genome, with alternative optima unsampled. Every genome in a Gram group shares this scaffold, which supplies most of the candidate set without organism evidence and, by the candidate-mask ablation (Section 3.3), both growth and the absence of dead ends.

Fixed medium and biomass. All genomes solve under a single minimal medium with oxygen, so aerobic pathways are available to every organism (including obligate anaerobes) and growth-coupled pathways are detected more reliably than non-coupled ones (secondary metabolism, denitrification). METEOR also presupposes a defined medium and a biomass reaction in the universal, of which SEED's prokaryote universal carries two Gram-specific variants. Eukaryotes would need a compartmentalised universal and their own biomass composition, and unculturable taxa an inferred medium that is itself a source of error, so we claim nothing beyond culturable bacteria.

Optimisation and evaluation. Four-threaded CBC with a 5% gap and no fixed seed can return different near-optimal sets between runs (re-solving the DeepProZyme-vanilla panel changed a mean of 1.6 reactions per genome, Supplementary Section S1.1). Nearly half of

the solve calls exceed the configured time limit; these wall-time counts are proxies for time-limited runs. Certification of termination reasons and achieved gaps requires solver logs and bounds. Solve time does not rise with candidate-set size (Spearman $\rho = -0.34$ over a narrow 16,000–17,900 range; Supplementary Section S7.3), so scaling to larger universals is untested; construction time scales roughly linearly with the universal. The 108 genomes include related taxa, so genome-wise tests are not 108 independent samples, and the full-proteome comparison is descriptive rather than a holdout test.

Scope. The sub-threshold comparison covers six curated models and DeepProZyme alone; the deletion experiment uses one perturbation per genome against METEOR's own output. Broader claims require additional predictors, organisms and perturbations.

5. Conclusion

METEOR converts per-protein EC scores into reaction costs, selects a growth-feasible reaction set in one optimisation, and writes the result back through a bounded per-EC update at the organism level. The shared scaffold makes the selected sets grow; evidence weighting decides which reactions join it, recovering 60.4% of curated sub-threshold ECs against 48.0% for threshold-and-repair and 48.3% for a size-matched baseline, at a marginal precision of 0.10 on the extra calls. METEOR shows no pathway-inference or per-protein gain, and no general gene-essentiality advantage across six organisms; the losses follow the reference table and candidate-pool quality more than parsimony (Section 3.4).

Availability. Code, results and supplement:
`https://github.com/bio-ontology-research-group/Meteor` (tag `v8.3-psb2027`; Zenodo `https://doi.org/10.5281/zenodo.21716682`). Baseline prediction matrices must be regenerated. The repository lists NCBI RefSeq accessions; Supplementary Section S1 gives solver settings and parameters.

Acknowledgments

This work was supported by KAUST baseline funding and by grants FCC/1/5932-11-01 and FCC/1/5940-07-02. We thank the KAUST Ibex HPC team for computational support. Large language models were used for language and structural editing, formatting and consistency checks; the authors developed all scientific content and remain responsible for the text.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, Basic local alignment search tool, *Journal of molecular biology* **215**, 403 (1990).
2. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer *et al.*, The pfam protein families database, *Nucleic acids research* **32**, D138 (2004).
3. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork and J. Huerta-Cepas, eggnog-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale, *Molecular biology and evolution* **38**, 5825 (2021).

4. T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang and H. Zhao, Enzyme function prediction using contrastive learning, *Science* **379**, 1358 (2023).
5. J. Y. Ryu, H. U. Kim and S. Y. Lee, Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers, *Proceedings of the National Academy of Sciences* **116**, 13996 (2019).
6. G. B. Kim, J. Y. Kim, J. A. Lee, C. J. Norsigian, B. O. Palsson and S. Y. Lee, Functional annotation of enzyme-encoding genes using deep learning with transformer layers, *Nature Communications* **14**, p. 7370 (2023).
7. N. Buton, F. Coste and Y. Le Cunff, Predicting enzymatic function of protein sequences with attention, *Bioinformatics* **39**, p. btad620 (2023).
8. J. D. Orth, I. Thiele and B. Ø. Palsson, What is flux balance analysis?, *Nature biotechnology* **28**, 245 (2010).
9. G. Plata, T. Fuhrer, T.-L. Hsiao, U. Sauer and D. Vitkup, Global probabilistic annotation of metabolic networks enables enzyme discovery, *Nature Chemical Biology* **8**, 848 (2012).
10. N. Vlassis, M. P. Pacheco and T. Sauter, Fast reconstruction of compact context-specific metabolic network models, *PLoS Computational Biology* **10**, p. e1003424 (2014).
11. A. Schultz and A. A. Qutub, Reconstruction of tissue-specific metabolic networks using CORDA, *PLoS Computational Biology* **12**, p. e1004808 (2016).
12. D. Machado, S. Andrejev, M. Tramontano and K. R. Patil, Fast automated reconstruction of genome-scale metabolic models for microbial species and communities, *Nucleic Acids Research* **46**, 7542 (September 2018).
13. J. Zimmermann, C. Kaleta and S. Waschina, gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models, *Genome Biology* **22**, p. 81 (December 2021).
14. M. L. Jenior, E. M. Glass and J. A. Papin, Reconstructor: a COBRApy compatible tool for automated genome-scale metabolic network reconstruction with parsimonious flux-based gap-filling, *Bioinformatics* **39**, p. btad367 (June 2023).
15. I. Thiele, N. Vlassis and R. M. Fleming, fastgapfill: efficient gap filling in metabolic networks, *Bioinformatics* **30**, 2529 (2014).
16. M. N. Benedict, M. B. Mundy, C. S. Henry, N. Chia and N. D. Price, Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models, *PLoS Computational Biology* **10**, p. e1003882 (2014).
17. A. Osterman and R. Overbeek, Missing genes in metabolic pathways: a comparative genomics approach, *Current Opinion in Chemical Biology* **7**, 238 (2003).
18. R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko and V. Vonstein, The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, *Nucleic Acids Research* **33**, 5691 (2005).
19. M. L. Green and P. D. Karp, A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases, *BMC Bioinformatics* **5**, p. 76 (2004).
20. M. N. Price, A. M. Deutschbauer and A. P. Arkin, GapMind: Automated annotation of amino acid biosynthesis, *mSystems* **5**, e00291 (2020).
21. C. Chen, C. Liao and Y.-Y. Liu, Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning, *Nature Communications* **14**, p. 2375 (2023).
22. M. D. Boer, C. Melkonian, H. Zafeiropoulos, A. F. Haas, D. R. Garza and B. E. Dutilh, Improving genome-scale metabolic models of incomplete genomes with deep learning, *iScience* **27**, p. 111349

(2024).
23. C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay and R. L. Stevens, High-throughput generation, optimization and analysis of genome-scale metabolic models, *Nature Biotechnology* **28**, 977 (2010).
24. S. M. Seaver, F. Liu, Q. Zhang, J. Jeffryes, J. P. Faria, J. N. Edirisinghe, M. Mundy, N. Chia, E. Noor, M. E. Beber *et al.*, The modelseed biochemistry database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes, *Nucleic acids research* **49**, D575 (2021).
25. B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan *et al.*, The swiss-prot protein knowledgebase and its supplement trembl in 2003, *Nucleic acids research* **31**, 365 (2003).
26. M. Kanehisa and S. Goto, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic acids research* **28**, 27 (2000).
27. J. Forrest, T. Ralphs, S. Vigerske, H. G. Santos, L. Hafer, B. Kristjansson, M. Lubin, M. Saltzman *et al.*, coin-or/cbc: Release 2.10.12 (August 2024).
28. J. P. Vielma, Mixed integer linear programming formulation techniques, *SIAM Review* **57**, 3 (2015).
29. C. Lieven, M. E. Beber, B. G. Olivier, F. T. Bergmann, M. Ataman, P. Babaei, J. A. Bartell, L. M. Blank, S. Chauhan, K. Correia, C. Diener, A. Dräger, B. E. Ebert, J. N. Edirisinghe, J. P. Faria, A. M. Feist, G. Fengos, R. M. T. Fleming, B. García-Jiménez, V. Hatzimanikatis, W. Van Helvoirt, C. S. Henry, H. Hermjakob, M. J. Herrgård, A. Kaafarani, H. U. Kim, Z. King, S. Klamt, E. Klipp, J. J. Koehorst, M. König, M. Lakshmanan, D.-Y. Lee, S. Y. Lee, S. Lee, N. E. Lewis, F. Liu, H. Ma, D. Machado, R. Mahadevan, P. Maia, A. Mardinoglu, G. L. Medlock, J. M. Monk, J. Nielsen, L. K. Nielsen, J. Nogales, I. Nookaew, B. O. Palsson, J. A. Papin, K. R. Patil, M. Poolman, N. D. Price, O. Resendis-Antonio, A. Richelle, I. Rocha, B. J. Sánchez, P. J. Schaap, R. S. Malik Sheriff, S. Shoaie, N. Sonnenschein, B. Teusink, P. Vilaça, J. O. Vik, J. A. H. Wodke, J. C. Xavier, Q. Yuan, M. Zakhartsev and C. Zhang, MEMOTE for standardized genome-scale metabolic model testing, *Nature Biotechnology* **38**, 272 (March 2020).
30. J. M. Monk, C. J. Lloyd, E. Brunk, N. Mih, A. Sastry, Z. King, R. Takeuchi, W. Nomura, Z. Zhang, H. Mori, A. M. Feist and B. O. Palsson, iML1515, a knowledgebase that computes escherichia coli traits, *Nature Biotechnology* **35**, 904 (2017).
31. I. Thiele, D. R. Hyduke, B. Steeb, G. Fankam, D. K. Allen, S. Bazzani, P. Charusanti, F.-C. Chen, R. M. Fleming, C. A. Hsiung, S. C. De Keersmaecker, Y.-C. Liao, K. Marchal, M. L. Mo, E. Özdemir, A. Raghunathan, J. L. Reed, S.-I. Shin, S. Sigurbjörnsdóttir, J. Steinmann, S. Sudarsan, N. Swainston, I. M. Thijs, K. Zengler, B. O. Palsson, J. N. Adkins and D. Bumann, A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2, *BMC Systems Biology* **5**, p. 8 (2011).
32. Y.-C. Liao, T.-W. Huang, F.-C. Chen, P. Charusanti, J. S. J. Hong, H.-Y. Chang, S.-F. Tsai, B. O. Palsson and C. A. Hsiung, An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228, *Journal of Bacteriology* **193**, 1710 (2011).
33. J. Nogales, J. Mueller, S. Gudmundsson, F. J. Canalejo, E. Duque, J. Monk, A. M. Feist, J. L. Ramos, W. Niu and B. O. Palsson, High-quality genome-scale metabolic modelling of Pseudomonas putida highlights its broad metabolic capabilities, *Environmental Microbiology* **22**, 255 (2020).
34. Y. Seif, J. M. Monk, N. Mih, H. Tsunemoto, S. Poudel, C. Zuniga, J. Broddrick, K. Zengler and B. O. Palsson, A computational knowledge-base elucidates the response of Staphylococcus aureus to different media types, *PLOS Computational Biology* **15**, p. e1006644 (2019).
35. Y.-K. Oh, B. O. Palsson, S. M. Park, C. H. Schilling and R. Mahadevan, Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and

gene essentiality data, *Journal of Biological Chemistry* **282**, 28791 (2007).
36. A. Bairoch, The ENZYME database in 2000, *Nucleic Acids Research* **28**, 304 (2000).
37. J. Schellenberger, J. O. Park, T. M. Conrad and B. Ø. Palsson, Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions, *BMC bioinformatics* **11**, 1 (2010).
38. T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner and H. Mori, Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection, *Molecular Systems Biology* **2**, p. 2006.0008 (2006).
39. B. Amos and J. Z. Kolter, OptNet: Differentiable optimization as a layer in neural networks, in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, , Proceedings of Machine Learning Research Vol. 70 (PMLR, 2017).
40. A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond and J. Z. Kolter, Differentiable convex optimization layers, in *Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2019).
41. C. J. Fritzemeier, D. Hartleb, B. Szappanos, B. Papp and M. J. Lercher, Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal, *PLoS computational biology* **13**, p. e1005494 (2017).