Text S1

1. Algorithm of our proposed pipeline

The instance-feature file (see main article and Table S1), containing 384 essential and 3120 non-essential reaction-gene pairs (R_a_G_b) was given to the following algorithm. It is noteworthy to mention that, our methodology is not specific to only this problem of classification and hence, can be applied for classifying any other kinds of datasets. Further, 1000 randomized balanced datasets (equal number of positive and negative classes) were generated and given to the integrated pipeline.

Algorithm for choosing best feature combination (Part 1)

The following algorithm was used to choose best features from a total of 64 features:

// **BD:** Set of 1000 Balanced training Datasets
// **TRF:** Set of Top Ranking Features
// **PM:** Set of Performance Metrics (auROC)
// **perf:** Set of best performing metrics
// **BD_{temp}:** Set of Balanced Datasets giving high performance

for i=1: length(BD)
TRF[ ] = Features ranked (descending) using SVM-RFE
for j=1: length(TRF[ ])
PM[i][j] = auROC measured using SMO
end loop $_j$
BFC[i] = best features combination set from PM[i][j] which gives best auROC
for k=1: length(BD)
BFC_BD[i][k] = performance metric with BFC[i] measured using SMO
end loop $_k$
end Loop $_i$
for m=1: length(BD)
auROC[m] = **sum**(BFC_BD[m]) / **length**(BFC_BD[m]);
end loop $_m$
Sort (descending) auROC[ ]
Select Best feature combination set (BFC_{best}) which gives highest average performance (auROC[1])

Algorithm for parameter optimization (Part 2)

Training with BFC_{best} (obtained from Part 1) and tuning complexity parameter C

```
C[ ] = {0.01, 0.1, 1, 10, 100};
for i = 1 : length(C[ ])
        for j = 1: length(BD)
            PM[i][j] = performance metrics measured using SMO;
        end loop j
Sort (descending) PM[i] and choose corresponding BD[j]
perf[i] =  PM[1];                    //  best auROC for C[i]
BD_temp[i] = BD[1];                          // dataset giving best auROC for C[i]
avePERF_C[i]=sum(PM[i])/length(PM[i]);
end loop i
Sort (descending) perf[ ] and choose corresponding BD_temp[ ]
Sort (descending) avePERF_C[ ] and choose corresponding C[ ]
```

Output = Best feature combination, best penalty parameter, best dataset (**BFC_{best}, C_{best}, BD_{best}**) which gives highest performance

Algorithm for Model testing

The unbalanced instance-feature file again was given as the total master test set. The testing algorithm returns two results -

1) Predictions from the best model (C_{best}, BFC_{best}, BD_{best})

2) Predictions of the model (C_{best}, BFC_{best}) with respect to the 1000 random datasets -

Given - C_{best}, BFC_{best}

```
for i=1: length(Ra_Gb)
count = 0;
        for j=1: length(BD)
            if Ra_Gb[i] == "E"          // "E" means essential
              count=count+1;
            end if
        end loop j
PercentagePrediction = [count / length(BD)]*100;
End
```

2. Curated features for *E. coli* K-12 MG1655

Sequence-based, gene expression-based, metabolic network and flux coupled subnetwork based features was assembled for each reaction-gene combination within *E. coli* K-12 MG1655 metabolism. Thus, a total of 64 features were obtained for each combination (Table A, Table S1).

Nucleotide content and coding sequence length **-** Previous studies showed that GC content of bacterial genomes were either correlated with environmental niche in which the bacterium survives or as a proxy for horizontally transferred genes.[53] Thus, the underlying GC content within a considered genome can be an appropriate representative of gene essentiality. Frequency of A, T, G, and C nucleotides at the 3rd synonymous position of codons and percentage GC content at all the 3 codon positions in a gene and length of each coding sequence (A3, T3, G3, C3, GC1, GC2, GC3, CDSlen) were calculated using an in-house code.

Codon usage **-** Codon usage was previously used as a predictor of protein abundance.[54–56] Abundant highly expressing proteins might have a functional importance in metabolism and hence, can be essential. Codon usage is also strongly associated to GC content within *E. coli* and is a distinctive signature between genomes.[56–58] Codon usage features like Codon Adaptation Index (CAI),[55] and Effective Number of Codons (ENC)[57] were extracted using EMBOSS package version 6.6.0-1.[59] Total Number of codons (Num_codons) in a coding sequence was calculated using an in-house code.

Homology based features **-** A gene might be more important if it is conserved across evolutionarily related organisms residing in different environments. In bacteria, essential genes were observed to be more evolutionarily conserved as compared to non-essential genes irrespective of the environment.[60,61] Phyletic Retention (PR) is defined as the number of organisms in which ortholog of a given gene is present.[8] For computing PR, protein orthologs amongst 710 bacterial genomes (leaving *E. coli* K-12 substr MG1655) available from the 2014 update of the COG database[36] was searched. An ortholog was defined such that, it is the only bi-directional best hit of the query gene in an organism and possessed atleast 40% identity with the query gene along with an E-value cut-off 10^{-7}. Bi-directional best hits for each gene were identified using BLAST version 2.2.26 along with the above parameters. Further, the number of

homologs in the 710 genomes with respect to the hits obtained with different E-value cut-offs ranging from 10^{-3} to 10^{-30} (H3, H5, H7, H10, H20, H30) was also calculated [15].

Peptide sequence features - Biased amino acid usage is a property of essential genes in bacterial genomes. The amino acid usage within a protein sequence is largely dependent on the physicochemical properties that an amino acid provides. In *E. coli*, proteins from essential genes might have specific structural properties important for its function.[27] A total of 20 peptide sequence-based features with respect to the frequencies of the 20 amino acids for each protein related to a particular gene (Gly_f, Met_f, Ala_f, Val_f, Leu_f, Ile_f, Pro_f, Phe_f, Trp_f, Tyr_f, Asn_f, Arg_f, His_f, Glu_f, Gln_f, Ser_f, Thr_f, Asp_f, Cys_f, Lys_f) was calculated. Similarly, protein length is also an important factor for determining function of a gene in *E. coli.* Protein length and amino acid usage was calculated using EMBOSS package version 6.6.0-1.[59]

Gene expression features - Essential genes tend to express at higher rates as compared to non-essential genes across bacteria.[45] To calculate gene expression based features, 101 microarray experimental samples (from 16 different studies given in Table B) for *E. coli* K-12 substr. MG1655 that were performed under different environmental stress conditions was collected. The microarray studies were curated from *E. coli* gene expression database.[62] Microarray studies carried out on mutant strains was not considered, as our aim was to predict essential genes in a wild type strain, subject to an array of environmental conditions. Also, the microarray experiments related to gene expression based features were chosen such that it covers the expression profiles of genes under various environmental stress conditions to get a universal definition of gene irrespective of the environment. Average mRNA Expression of a gene (aveEXP) and mRNA Expression Fluctuation (mEF) which is the standard deviation of log2 normalized gene expression values of the Cy3/Cy5 intensity ratio of each gene from the above samples were calculated. From previous studies, it was reported that a gene might be important if it co-regulated with many other genes.[63] Hence, the Number of Genes with Similar Expression (NGSE), (number of gene pairs having a Pearson correlation coefficient: $r < -0.8$ and $r > 0.8$)[15] was also calculated.

Reaction and Flux-coupled sub-network features - The *E. coli* iJO1366 metabolic network was transformed as a static undirected reaction network representation, in which each node represents

an enzyme (reaction) and the edge signifying the connection between two reaction such that product of previous reaction is the substrate of the next reaction. Different commonly used network features that emphasize biological importance of an enzyme with respect to its position and connectivity to other enzymes in a network were calculated.[64] A highly connected and central enzyme in biological networks is necessarily essential as it represents either a hub or bottleneck within the network.[65] Each network feature would represent the importance of a reaction to interconvert substrates/products given by/to a subsequent reaction. Features like Degree Centrality (DC),[14,15] Closeness Centrality (CC),[14,15] Betweenness Centrality (BC),[14,15] Eccentricity Centrality (EC),[65] Page Rank (PageRank),[66] Eigenvector Centrality (EvC),[15] Modularity (M),[67] Clustering Coefficient (ClustCoef),[14,15] Number of triangles (Num_triangles),[68] Hub (HS)[69] and Authority (AS)[69] Scores were calculated for the reaction network using Gephi version 0.8.6[70]. The aforementioned network topological features were calculated for the flux-coupled sub-graph as well.

Table A. Curated features for *E. coli* K-12 MG1655

Feature Name	**Description**	**# Features**
Sequence Features	A3, T3, G3, C3, GC1, GC2, GC3, CDSlen, CAI, ENC, Num_Codon, Gly_f, Met_f, Ala_f, Val_f, Leu_f, Ile_f, Pro_f, Phe_f, Trp_f, Tyr_f, Asn_f, Arg_f, His_f, Glu_f, Gln_f, Ser_f, Thr_f, Asp_f, Cys_f, Lys_f, PL, PR, H3, H5, H7, H10, H20, H30	39
Expression features	NGSE, aveEXP, mEF	3
Network features		
Reaction Network (RN)	RN_DC, RN_EC, RN_CC, RN_ BC, RN_EvC, RN_ HS, RN_AS, RN_PageRank, RN_ClustCoeff, RN_Num_triangles, RN_M	11
Flux Coupled Sub Network (FCA)	FCA_DC, FCA_ EC, FCA_CC, FCA_BC, FCA_EvC, FCA_HS, FCA_AS, FCA_PageRank, FCA_ClustCoeff, FCA_Num_triangles, FCA_M	11

Experiments curated for calculating gene expression-based features **-** Gene expression based features like average mRNA Expression (aveEXP), mRNA Expression Fluctuation (mEF) and the gene co-expression feature NGSE were calculated from 16 microarray experiments.

Table B. Microarray experiments curated for *E. coli* K-12 substr. MG1655 from *E. coli* gene expression database

Accession	Conditions	# Samples	References
GSE1730	mRNA abundance from wild type grown in LB or M9	12	[71]
GSE1981	Gene expression profiles of *E.coli* grown in LB and minimal medium (DM) at OD 600 =1	3	[72]
GSE4344	Transcript abundance in LB and L9 at 30^0 C and OD 600=0.8	4	[73]
GSE4359	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	6	[74]
GSE4363	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	11	[74]
GSE4364	Responses of genes over time as they recover from one stationary phase In rich media at OD 0.4	7	[74]
GSE4370	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	7	[74]
GSE4371	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	7	[74]
GSE4373	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	6	[74]
GSE4374	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	6	[74]
GSE4375	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	6	[74]
GSE4376	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	6	[74]
GSE4380	Responses of genes over time as they recover from one stationary phase in rich media at OD 0.4	6	[74]
GSE4706	Response and adaptation to growth with low glucose	6	[75]

	concentration		
GSE4735	Transcriptome dynamic analysis during transition from aerobic and anaerobic conditions	4	[76]
GSE6644	Transcriptome dynamic analysis during transition from aerobic and anaerobic conditions	4	[76]

3. Flux coupling analysis

Flux coupling analysis (FCA) is a flux-based optimization procedure that calculates reaction subsets that are either coupled with each other via flux or represent a set of block reactions, given specific environmental exchange constraints.[31,32] Let v_1 and v_2 be fluxes through reactions R_1 and R_2. Keeping either v_1 or v_2 as objective functions to be optimized, if a non-zero flux in v_1 imposes a non-zero flux in v_2 or vice versa, the two reaction fluxes are termed to be coupled with each other. If zeroing the flux of one reaction does not produce any effect on any other reaction within the metabolic network, then the reaction is termed to be uncoupled. If maximum or minimum of a particular reaction flux objective equals zero, then the reaction is termed to be blocked. Considering v_1 or v_2 to be objective functions, the coupled reactions can be classified into:

1) **Fully coupled:** If $v_1 = 0$ implies $v_2 = 0$ and if $v_2 = 0$ implies $v_1 = 0$, and $v_1 = v_2$, then the reaction pair is fully coupled.
2) **Directionally coupled**: If $v_1 = 0$ implies $v_2 = 0$ but if $v_2 = 0$ does not imply $v_1 = 0$, then the reaction pair is directionally coupled.
3) **Partially coupled:** If $v_1 = 0$ implies $v_2 = 0$ and if $v_2 = 0$ implies $v_1 = 0$, and $v_1 \neq v_2$, then the reaction pair is partially coupled.

Performing FCA on the iJO1366 metabolic network, 1527 fully, 41049 directionally, and 7438 partially coupled reaction pairs and 865 blocked reactions were obtained. As our aim was to find a flux-coupled subnetwork, the nature/property of each reaction pair can be represented within an adjacency matrix (1718 x 1718) where each reaction pair can be given a value of 1 or 0 corresponding to whether they are either coupled or not. Here, we give a value of 0 to both uncoupled and blocked reaction pairs. The adjacency matrix represents a flux-coupled subgraph, which can be used to extract biologically relevant topological features dependent on predicted physiological flux relationships.

4. Definitions of Performance metrics

In our data set, two classes essential (positive class) and non-essential (negative class) genes are considered. The classifier has four outcomes:

True positive (TP): Number of essential instances correctly predicted by the classifier.

*False positive (FP)***:** Number of non-essential instances wrongly predicted as essential.

True negative (TN): Number of non-essential instances correctly predicted by the classifier.

False negative (FN): Number of essential instances wrongly predicted as non-essential.

With respect to the above model outcomes, a set of model performance metrics can be calculated (See Table 1 of Main Article for formulae). The metrics are defined as,

True Positive Rate (TPR) or Sensitivity: It is defined as the proportion of positive (essential) instances predicted correctly by the model.

False Positive Rate (FPR): It is defined as the proportion of negative (non-essential) instances predicted as positive by the model.

*Precision***:** It determines the measure of correctness (i.e., how many instances predicted as essential class really belongs to positive class. .

*Recall***:** It measures the proportion of essential instances correctly predicted by the model.

*F-measure***:** This performance metric is defined as the harmonic mean between precision and recall. A high value of F-measure suggests that the predictive performance better on essential class.

*Matthews Correlation Coefficient (MCC)***:** This performance metric was proposed by biochemist Brain W. Matthews in 1975.[77] This measure is less influenced by imbalanced data sets.

*Area under Receiver Operating Characteristic curve (auROC)***:** Area which is calculated from ROC curve. Wilcoxon-Mann-Whitney test statistic is used to calculate auROC.

5. Comparison with other available methods – Proof of training set independence

Apart from Hwang et al., 2009,[14] our strategy was also compared with other recent supervised classification studies on essential gene identification.[16,33] To compare the performance of our strategy with these classification methods, training (*Escherichia coli* genes) and test dataset (*Bacillus subtilis* genes) considered in these studies were provided to our methodology for generating a best SVM model and for further testing, respectively. The best model generated from our methodology using the previously available training dataset (consisting of sequence features of *E. coli* genes) was further tested with test dataset (sequence-based features of *B. subtilis* genes) of the available methods.

Testing results in the form of auROC and precision, indicate that the best model generated through our strategy outperforms both the available supervised classification methods (Table C). The achieved sensitivity and specificity from our methodology is the highest suggesting an enhanced model performance, irrespective of the given input training dataset.

Table C. Comparison of our proposed strategy with methods proposed by Song et al. 2014[33] and Deng et al. 2011[16]

Performance metric	Our Method	Song et al. 2014[33]	Deng et al. 2011[16]
auROC	0.966	0.930	0.800
Precision	0.970	0.730	0.540

6. List of selected features

After applying SVM-RFE algorithm for feature selection, combinations of 26 features that give highest model performance were shortlisted. The ranks (contributions to classification) of these features based on SVM-RFE are given in Table D. Also, after performing Wilcoxon rank-sum test, 21 features out of the total 26 have significantly different median values between essential and non-essential genes. *P*-values calculated by comparing the distributions of features between essential and non-essential classes (Fig. 3 of Main Article) are also indicated in Table D.

Table D: The 26 selected best features

Feature Name	Rank (SVM-RFE)	*P*-value
PR	1	2.2E-16
H3	2	2.2E-16
RN_CC	3	2.2E-16
H5	4	2.2E-16
Ala_f	5	0.000052
Arg_f	6	0.00996
H7	7	2.2E-16
CAI	8	0.2447*
FCA_AS	9	2.2E-16
FCA_Num_triangles	10	2.2E-16
FCA_Hub	11	2.2E-16
Leu_f	12	0.2669*
RN_BC	13	0.000309
GC1	14	2.2E-16
T3	15	1.35E-10
NGSE	16	0.2399*
Gly_f	17	2.2E-16
Val_f	18	0.5293*
Ser_f	19	0.009535
FCA_ClusteringCoef	20	3.08E-12
Phe_f	21	2.21E-15
C3	22	1.65E-05
Glu_f	23	0.001545
GC2	24	0.2302*
Asn_f	25	2.2E-16
FCA_BC	26	2.93E-15

* *P*-values insignificant at a threshold of $P < 0.05$

References cited in this ESI Text S1 are provided in the Reference list of the Main Article.