**Supplementary Table**

**Supplementary Table 1.** Identifying and extracting keywords corresponding to transporters and enzymes from functional descriptions.

|  |  |
| --- | --- |
| Label | Keywords |
| Transporter | transporter, hypothetical transporter, putative transporter, symporter, antiporter, permease, export, import, efflux, influx, excrete, uptake |
| Enzyme | enzyme, synthase, transferase, ligase, nuclease, decarboxylase, helicase, epimerase, kinase, phosphatase, reductase |

**Supplementary Table 2.** Example of extracting transport directions from functional descriptions.

|  |  |
| --- | --- |
| Description | Direction label |
| Leucine efflux protein LeuE | Outward transport |
| L-cystine uptake protein TcyP | Inward transport |
| Threonine export carrier | Outward transport |
| Tyrosine transporter TyrP | None |
| ... | ... |

**Supplementary Table 3.** The number of strain proteomes downloaded from the GenBank database for 6 genera.

|  |  |
| --- | --- |
| Genera | The number of extracted strain proteomes |
| *Escherichia* | 30 |
| *Bacillus* | 10 |
| *Corynebacterium* | 10 |
| *Lactobacillus* | 7 |
| *Saccharomyces* | 5 |
| *Parageobacillus* | 3 |

**Supplementary Table 4.** Identifying and extracting keywords corresponding to transport directions from functional descriptions

|  |  |
| --- | --- |
| Label | Keywords |
| Inward transport | import, influx, uptake, intake |
| Outward transport | export, efflux, excrete, extrude, eject |

**Supplementary Table 5.** Example of extracting substrate names from functional descriptions.

|  |  |
| --- | --- |
| Description | Substrate label |
| Tyrosine transporter TyrP | Tyrosine |
| Sodium/proline symporter PutP | Proline |
| Glutamate/aspartate:proton symporter GltP | Glutamate |
| Glutamate/aspartate:proton symporter GltP | Aspartate |
| MFS transporter | None |
| ... | ... |

**Supplementary Table 6.** The distribution of transporters in the dataset used for constructing the phylogenetic tree.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sugar tps | Amino acid or Peptide tps | Lipid tps | Cation tps | Anion tps | Carboxylic acid tps | Nucleotide tps | Vitamin tps | Other | Total |
| Bacteria | 403 | 467 | 71 | 459 | 529 | 67 | 27 | 68 | 266 | 2357 |
| Fungi | 39 | 47 | 3 | 44 | 28 | 16 | 27 | 10 | 47 | 261 |

**Supplementary Table 7.** The performance metrics of XGB, RF, and SVM models trained on the UTP train set, along with their predictive Accuracy, Precision, Recall, ROC-AUC, and MCC on Test set 1 and Test set 2.

|  |  |  |  |
| --- | --- | --- | --- |
|  | XGB | RF | SVM |
| Accuracy % | 98.47 | 98.12 | 95.40 |
| Precision % | 98.41 | 99.42 | 99.69 |
| Recall % | 98.69 | 97.00 | 91.56 |
| ROC-AUC | 0.9980 | 0.9971 | 0.9974 |
| MCC | 0.9692 | 0.9627 | 0.9115 |
| Test set 1 Accuracy % | 96.86 | 90.81 | 93.72 |
| Test set 1 Precision % | 96.66 | 98.95 | 96.10 |
| Test set 1 Recall % | 95.59 | 78.24 | 88.15 |
| Test set 1 ROC-AUC | 0.9957 | 0.9914 | 0.9900 |
| Test set 1 MCC | 0.9349 | 0.8168 | 0.8704 |
| Test set 2 Accuracy % | 91.70 | 90.31 | 90.44 |
| Test set 2 Precision % | 81.29 | 86.50 | 80.53 |
| Test set 2 Recall % | 94.56 | 80.84 | 90.54 |
| Test set 2 ROC-AUC | 0.9759 | 0.9669 | 0.9658 |
| Test set 2 MCC | 0.8180 | 0.7680 | 0.7850 |

**Supplementary Table 8.** The predictive Accuracy of the XGB model in each strain subset of Test set 2.

|  |  |  |  |
| --- | --- | --- | --- |
| Subset in Test set 2 | Number of transporters | Number of enzymes | Accuracy % |
| *B.subtilis* str.168 | 434 | 101 | 90.09 |
| *B.subtilis* 6051-HGW | 22 | 650 | 90.77 |
| *B.subtilis* NCIB-3610 | 69 | 207 | 93.48 |
| *C.glutamicum* ATCC 13032 | 229 | 447 | 97.34 |
| *C.glutamicum* ATCC 21799 | 256 | 436 | 94.80 |
| *C.glutamicum* SCgG2 | 65 | 296 | 96.40 |
| *C.glutamicum* TQ2223 | 157 | 283 | 94.55 |
| *E.coli* BL21 | 391 | 695 | 94.38 |
| *E.coli* DSM 3003 | 419 | 779 | 93.82 |
| *E.coli* ETEC H10407 | 201 | 415 | 87.50 |
| *E.coli* MG1655 | 117 | 286 | 91.56 |
| *E.coli* O157:H7 str. Sakai | 291 | 669 | 92.92 |
| *E.coli* SE11 | 121 | 389 | 91.37 |
| *E.coli* SMS-3-5 | 269 | 561 | 93.25 |
| *E.coli* UT189 | 66 | 280 | 86.99 |
| *L.acidophilus* La-14 | 135 | 393 | 92.61 |
| *L.delbrueckii jakobsenii* ZN7a-9 | 181 | 395 | 91.67 |
| *L.gasseri* 2016 | 197 | 384 | 93.80 |
| *P.thermoglucosidasius* DSM2542 | 327 | 642 | 90.71 |
| *S.cerevisiae* S288C | 40 | 787 | 76.30 |

**Supplementary Table 9.** The performance metrics of XGB, RF, and SVM models trained on the DirectIO train set, along with their predictive Accuracy, Balanced Accuracy, ROC-AUC, and MCC on Test set 3, as well as Accuracy on Test set NMN.

|  |  |  |  |
| --- | --- | --- | --- |
|  | XGB | RF | SVM |
| Accuracy % | 90.87 | 93.68 | 92.51 |
| Balanced Accuracy % | 81.17 | 88.16 | 85.92 |
| ROC-AUC | 0.9693 | 0.9781 | 0.9823 |
| MCC | 0.7244 | 0.8126 | 0.7762 |
| Test set 3 Accuracy % | 78.87 | 82.73 | 67.68 |
| Test set 3 Balanced Accuracy % | 80.87 | 81.04 | 77.86 |
| Test set 3 ROC-AUC | 0.8771 | 0.9028 | 0.9061 |
| Test set 3 MCC | 0.5535 | 0.5850 | 0.4936 |
| Test set NMN Accuracy % | 66.67 | 100.00 | 66.67 |

**Supplementary Table 10.** The performance metrics of SPOTESM1b\_ECFP, SPOTESM1b\_ChemBERTa and SPOTIC models, along with their predictive Accuracy on Test set 4.

|  |  |  |  |
| --- | --- | --- | --- |
|  | SPOTESM1b\_ECFP | SPOTESM1b\_ChemBERTa | SPOTIC |
| SPOT Test set Accuracy % | 91.50 | 92.40 | 91.17 |
| SPOT Test set Precision % | 84.00 | 88.00 | 83.60 |
| SPOT Test set Recall % | 80.21 | 81.19 | 82.22 |
| SPOT Test set ROC-AUC | 0.9560 | 0.9610 | 0.9510 |
| SPOT Test set MCC | 0.7800 | 0.8000 | 0.7696 |
| SPOTIC Test set Accuracy % | 91.07 | 90.49 | 91.04 |
| SPOTIC Test set Precision % | 86.37 | 84.14 | 84.03 |
| SPOTIC Test set Recall % | 79.05 | 79.34 | 82.05 |
| SPOTIC Test set ROC-AUC | 0.9530 | 0.9455 | 0.9509 |
| SPOTIC Test set MCC | 0.7670 | 0.7532 | 0.7695 |
| Test set 4 Accuracy % | 42.34 | 5.24 | 78.63 |

**Supplementary Table 11.** The proportions of data points from different transporter categories that formed clusters in the phylogenetic tree of bacterial-derived transporters (TB) and the phylogenetic tree of fungal-derived transporters (TF) were calculated relative to the total number of data points in each category.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sugar tps** | **Amino acid or Peptide tps** | **Lipid tps** | **Cation tps** | **Anion tps** | **Carboxylic acid tps** | **Nucleotide tps** | **Vitamin tps** |
| Bacteria | 314 (77.92%) | 386 (82.66%) | 59 (83.10%) | 345 (85.61%) | 488 (92.25%) | 15 (23.39%) | 12 (44.44%) | 51 (75.00%) |
| Fungi | 29 (74.36%) | 17 (36.17%) | 0 | 0 | 0 | 0 | 7 (25.93%) | 0 |

**Supplementary Table 12.** A functional comparison was conducted for pairs of members with similarity scores of 60 or higher.

|  |  |  |
| --- | --- | --- |
| Similarity score | Member1 and Function | Member2 and Function |
| 63 | P15993 (Active transport of all three aromatic amino acids, phenylalanine, tyrosine and tryptophan) | P24207(Active transport of phenylalanine) |
| 64 | P0AA47 (Putrescine importer) | P76037 (Uptake of putrescine) |
| 91 | P40812 (L-asparagine permease) | P77610 (L-asparagine permease) |
| 65 | K7WL41 (Transporter of lysine, histidine and arginine) | K7VV21(The lysine specific transporter) |
| 62 | F2HQ24 (Transports alanine, serine and glycine) | F2HQ25(Transports L-serine, L-threonine and L-cysteine) |
| 68 | F2HL52 (Arginine/Ornithine antiporter) | F2HL56 (Arginine/Ornithine antiporter) |
| 96 | P37460 (Proline-specific permease) | P0AAE2 (Proline-specific permease) |
| 60 | O31199 (Histidine permease) | P0AAE2 (Proline-specific permease) |
| 60 | O31199 (Histidine permease) | P37460 (Proline-specific permease) |