BIO Web of Conferences **163**, 01001 (2025) https://doi.org/10.1051/bioconf/202516301001
ICBBB 2025

TooT-SS: Transfer Learning using ProtBERT-BFD Language Model for Predicting Specific Substrates of Transport Proteins

Sima Ataei[1]e-mail: sima.ataei@concordia.ca and *Gregory* Butler[1]e-mail: gregory.butler@concordia.ca

[1]Concordia University, Montreal, Canada

Abstract. Transmembrane transport proteins are essential in cell life for the passage of substrates across cell membranes. Metabolic network reconstruction requires transport reactions that describe the specific substrate transported as well as the metabolic reactions of enzyme catalysis. We utilize a protein language model called ProtBERT (Protein Bidirectional Encoder Representations from Transformers) and transfer learning with a one-layer Feed-Forward Neural Network (FFNN) to predict 96 specific substrates. We automatically construct a dataset UniProt-SPEC-100 using the ChEBI and GO ontologies with 4,455 sequences from 96 specific substrates. This dataset is extremely imbalanced with a ratio of 1:408 between the smallest class and the largest. Our model TooT-SS predicts 83 classes out of 96 with an F1-score of 0.92 and Matthews Correlation Coefficient (MCC) of 0.91 on a hold-out test set. The results of 3-fold cross-validation experiments, particularly, on small classes show the potential of transfer learning from the ProtBERT language model for handling imbalanced datasets.

1 Introduction

Transmembrane transport proteins are an essential group of proteins located integrally in the membrane that transport chemical substrates intra- and inter-cellular; that is, across the cell membrane that separates the cell interior from its environment and across the membranes of the organelles, such as the nucleus and mitochondria, within the cell. They facilitate the selective passage of vital compounds ranging from small ions to macromolecules, for example, other proteins, through active transport or facilitated diffusion. Transport is a part of the overall biological processes of cell metabolism, regulation, and signaling. Membrane proteins are estimated to form one-third of the proteome, and transport proteins are a major proportion of membrane proteins. Despite their important role and number, transport proteins are not sufficiently characterized. Experimental characterization and determination of their 3D structure are difficult. This emphasizes the need for computational methods to study this group of proteins.

The process of genome-scale reconstruction of a metabolic network (GENRE) builds a map to represent the gene-protein-reaction (GPR) association between the gene, the protein as a gene product, to the reaction carried out by the protein [1]. Ideally, a GENRE covers cell metabolism, transport, regulation, and signaling. However,

to date they mainly cover metabolism as enzymes are well understood. A GENRE can assign gene-protein-reaction associations based on Enzyme Commission (EC) classification or molecular function terms of the genes on Gene Ontology. The transport reactions model transport of substrates across membranes. The prediction of GPR association for transport proteins, in particular, the specific substrate, or substrates, is not well-developed.

The main focus of this research is to predict the specific substrate transported by a specific transmembrane transport protein. In other words, given transmembrane transport protein, we need a model to predict the associated carried substrate. The uneven interest of studies in different substrates resulted in imbalanced known transport proteins associated with them. We exclude proteins carrying more than one substrate, giving us a multi-class classification problem on an imbalanced dataset.

Although imbalanced datasets are one of the most prevalent types of data in real-world problems, there are not many studies that are conducted without data manipulation. Facing this problem, researchers usually tackle the imbalance using the under-sampling and over-sampling techniques to balance the dataset. However, a new era of machine learning, the emergence of transformers and protein language models (PLM) provides transfer learning that can utilize a small labelled task-specific dataset and leverage the pre-trained PLM which is task-agnostic through self-supervised learning. Thus protein language models and transfer learning encourage us to reconsider the classification of proteins with dataset imbalance.

Inspired by other fields of machine learning transfer learning from pre-trained models is applied to protein characterization datasets. Here, we focus on utilizing the ProtBERT protein language model [2] is one of the state-of-the-art protein language models. We evaluate transfer learning using the ProtBERT-BFD language model and with a one-layer Feed-Forward Neural Network (FFNN) for the task of predicting the specific substrate of a transport protein. Furthermore, we study whether the model is able to be trained on imbalanced datasets by focusing on the performance of the model on small classes.

Our contributions in this paper are first to build new datasets of transmembrane transport proteins with their specific associated substrate based on `ChEBI` and Gene Ontology. Second, we introduce the *TooT-SS* model using transfer learning using ProtBERT-BFD, a one-layer Feed-Forward Neural Network (FFNN) and fine-tuning to predict the specific substrate transported by a transmembrane transport protein. Third, we study the impact of the imbalanced dataset.

2 Related Works

Related works of the research in substrate prediction are divided to two groups of studies. The first group focuses on the standard applications of similarity search. These researches find the homologs of known transporters using Hidden Markov Models (HMM) of orthologous protein families. The second group focuses on the *de novo* sequences prediction with none or remote homolog known protein sequences. However, these methods work on either predicting the protein sequences based on their substrate class, family or subfamily or they focus on a specific group e.g ion channels [3]. To our knowledge there are no previous studies conducted on prediction of the exhaustive specific substrate regardless of their category.

Standard similarity search studies such as BioV [4], TransATH [5], merlin [6–8], and Pantograph [9] usually rely on HMM models of orthologous protein families and have been effective for annotation of transporters. There are also studies, such

as TransportTP [10], implementing traditional homology methods combined with machine learning to improve performance.

For *de novo* protein sequences, with at best only remote homologs, prediction is more challenging. Research in the labs of Gromiha [11–13], Helms [14–16], Zhao [10, 17], and Butler [5, 18–20] and sporadic contributions from teams working in machine learning [21–24] have been applying machine learning techniques to overcome the challenges. The best tools for classifying substrate classes, based on the performance measures in their papers, are TrSSP [17] and FastTrans method [24], both on seven substrate classes, and *TooT-SC* [19, 20] on eleven substrate classes.

Our first work on predicting specific substrates was *TooT-ICAT* [25]. A dataset of twelve specific inorganic cations and anions that had at least ten transmembrane transport proteins in UniProt was constructed as in Section 4.1. The dataset had 4112 proteins. Transfer learning was applied with ProtBERT-BFD and three downstream methods: logistic regression (LR), one-layer FFNN, and fine-tuning with the FFNN. On the hold-out test set they achieved MCC of 0.81, 0.88, and 0.95 respectively.

3 BERT Language Model for Proteins

Bidirectional Encoder Representations from Transformers (BERT) is a language model learning sequences representation using a self-attention mechanism [26, 27]. This multi-layer bidirectional transformer encoder is implemented in two phases: *pre-training* and *fine-tuning*. Pre-training of BERT includes two tasks: Masked Language Modeling (MLM), which masks a specific percentage of the input tokens at random to predict, and Next Sentence Prediction (NSP), which given two sentences predicts if they are in the correct order.

The BERT model has achieved state-of-the-art results for Natural Language Processing (NLP) tasks by applying transfer learning [28]. Transfer learning is a technique where a model learns general features of the specific type of data during the training process of one task, to be re-purposed on a second related task. This method is practical in the cases where data availability in one domain recompenses the lack of sufficient training data for another task. In such cases, the performance of learning is improved by avoiding expensive annotation efforts for data labeling. The BERT model profits transfer learning with two phases of self-supervised pre-training on a corpus of data and fine-tuning the pre-trained model on a specific task of Natural Language Processing using labeled data for supervised learning.

Protein primary structure is a composition of 20 essential amino-acids residues as a chain. The residues building the protein sequence have been interpreted as analogous to words building a sentence. This analogy has enabled researchers to utilize the advances in NLP transformer-based language models in the protein domain.

The ProtTrans project [2] trained and evaluated six transformer-based models on tasks associated with the protein domain. They developed ProtBERT, a BERT model with an architecture of 30 layers with a hidden size of 1024 and 16 attention heads. The models were pre-trained on the UniRef100 and BFD databases separately using MLM only. The BFD dataset [29] merges UniProt with proteins translated from multiple metagenomic sequencing projects. It has 2,122 million proteins and 393 billion amino acids. Here we focus on the ProtBERT-BFD PLM.

4 Dataset

4.1 Dataset Construction

Our dataset was constructed from UniProt release 2023_03 [30] using the methodology of *TooT-ICAT* [25]. This ontology-based method utilizes the Gene Ontology

(GO) and ChEBI ontology, and the connection for chemical compounds referred to in GO Molecular Function (MF) terms to ChEBI . From this, transport proteins (as UniProt IDs) can be linked to specific substrates (as ChEBI IDs), if that information is available in their annotations.

Since this study intends to predict all of the specific substrates carried by any transmembrane protein, we selected the root node *"CHEBI:24431"* (chemical entity) in ChEBI and GO term *"GO:0022857"* (transmembrane transporter activity–Molecular Function) as the start nodes. Extracting the Directed Acyclic Graph (DAG) of each node in their separate ontologies, we collect the leaves of each DAG. This process results in two separate leaf sets of ChEBI substrates and Gene Ontology molecular function terms.

According to Gene Ontology, catalytic activity has been classified as a transmembrane transporter activity. However since this type of activity is not a matter of interest in this study, their related GO terms are removed. To exclude the catalytic activity molecular function terms from the DAG, the GO terms derived from the node *GO:0003824 (catalytic activity)* have been eliminated from the GO leaf set.

The leaf sets are related by ontology mappings between GO and ChEBI provided by Gene Ontology. After map refinement, we use the extracted CHEBI2GO map to build our dataset. The protein sequences with a GO term from the CHEBI2GO map, are labeled as transporters of the mapping ChEBI substrate in the dataset.

The transmembrane transport proteins are collected from UniProt. The query to collect the proteins requires GO MF term

`GO22857:transmembrane transporter activity`

and existence evidence at the `protein level`. For querying SwissProt, we further add `reviewed:yes`. The query for UniProt is:

⟨`goa:(“transmembrane transporter activity [22857]”)`
`existence:“Evidence at protein level [1]”`⟩

The collected data includes the GO MF terms. Those associated with transport activity are labeled with their ChEBI substrates using the CHEBI2GO map. Sequences with multiple labels are eliminated from the dataset.

4.2 Identity Reduction

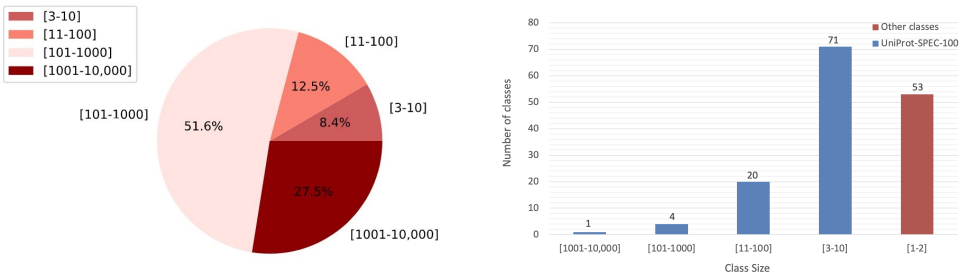
To reduce the pairwise sequence identity, we use CD-HIT [31] to remove sequences with more than 100% and 60% identity. Applying two pairwise identity thresholds on SwissProt and UniProt datasets separately, results in four different datasets. Table 1 shows the size of each dataset.. The detailed information about each dataset is presented in Table 5. As their name presents these data sets contain transmembrane transport proteins from SwissProt or UniProt with 2 different thresholds. For example, dataset UniProt-SPEC-100 represents the transporter sequences extracted from UniProt with less than 100% pairwise identity labeled with their specific substrate.

4.3 Class Size Threshold

The full collection of proteins and labels from UniProt contains 149 substrates. There are 53 classes of size 1 or 2 (shown in Figure 1). Our study is to investigate how well transfer learning deals with imbalanced datasets, and how it deals with small classes, We set a threshold of three as the minimum size of a class. However, no maximum threshold is applied. The threshold of thrree allows us to split the dataset with a test-train ratio of on 34% to 66%. This ratio secures one sequence from the smallest class for the stratified test-train split. Table 1 shows the number of sequences and classes in each of the four datasets.

Table 1: Four extracted datasets

Dataset	Size	Train	Test	# Classes
UniProt-SPEC-100 (UP-100)	4,455	2,970	1,485	96
UniProt-SPEC-60 (UP-60)	1,628	1,085	543	69
SwissProt-SPEC-100 (SP-100)	2,340	1,560	780	86
SwissProt-SPEC-60 (SP-60)	1,262	841	421	62



(a) Pie Chart (b) Histogram

Figure 1: Frequency of Class Sizes in UniProt-SPEC-100 Dataset

4.4 Dataset Analysis

Figure 1 shows that the dataset UniProt-SPEC-100 has four classes with size in [101..1000] and they have 51.6% of the sequences. These classes are CHEBI:29108 CALCIUM(2+), CHEBI:29103 POTASSIUM(1+), CHEBI:17996 CHLORIDE, and CHEBI:29101 SODIUM(1+). Furthermore, 27.5% of the proteins are in one majority class: CHEBI:24636 PROTON with 1,226 samples.

On the other hand, the dataset UniProt-SPEC-100 includes 71 classes with 3 to 10 sequences. These classes include 8.4% of the sequences. This group of data includes CHEBI:29036 COPPER(2+), CHEBI:29034 IRON(3+) and CHEBI:17992 SUCROSE which are considered important substrates for the cell metabolism. Figure 1b also presents that there are 20 classes with sizes between 11 to 1000 which include 12.5% percent of the dataset. The 53 substrate classes with less than 3 sequences are not included the UniProt-SPEC-100 dataset.

In summary, 79.1% of the data samples are distributed in the 5 largest classes and the remaining 20.9% are scattered in 91 small classes. The ratio of class size in the UniProt-SPEC-100 dataset is 1:408 which shows not only that the data is imbalanced but exhibits *extreme class imbalance* as it exceeds the 1:100 ratio set in [32].

5 Methods

5.1 Fine-tuning ProtBERT-BFD model

A pre-trained ProtBERT-BFD model has been fine-tuned on each dataset separately to update the model weights for the specified classification task. Our model includes a single-layer Feed-Forward Neural Network (FFNN) followed by a soft-max function to predict the probabilities of each class in the dataset. We used a cross entropy loss function and Adam optimizer [33] with learning rate $= 5 \times 10^{-5}$ for batch size 1. The

implementation used the function `torch.optim.Adam` of the PyTorch package [34] and the pre-trained ProtBERT-BFD model from the HuggingFace website [35]. The fine-tuning process achieves acceptable results in 15 epochs with a maximum MCC at 13 epochs (Figure 2). Hence our model *TooT-SS* was built using 13 epochs of fine-tuning

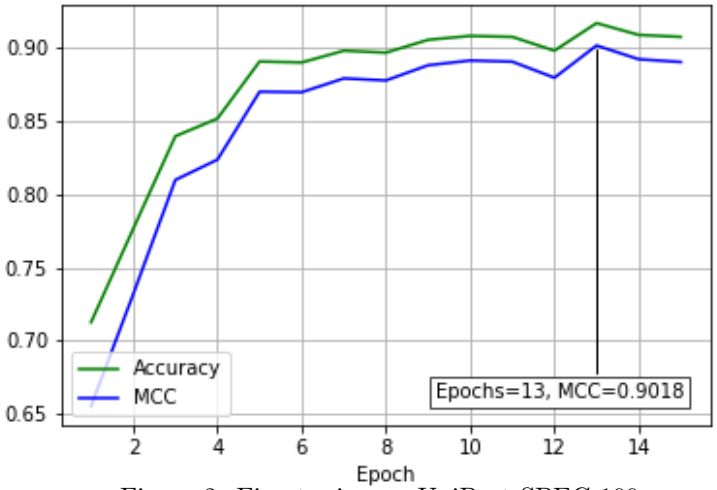


Figure 2: Fine-tuning on UniProt-SPEC-100

5.2 Cross-validation

Due to the existence of small classes in the datasets, results are expected to be highly affected by the randomness of the train-test split. For the classes of size three there are two sequences in the training set and only one sequence in the hold-out test set. Accordingly, a 3-fold cross-validation experiment was performed to examine the consistency of the results. The UniProt-SPEC-100 dataset was divided into 3-folds with stratification to include each class in both test and training sets. Then, the fine-tuning process has been implemented on each fold separately. The performance variation across the 3-fold cross-validation was reported as the standard deviation.

5.3 Evaluation Metrics

To evaluate the performance of the method, five different metrics were considered:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Table 2: 3-fold cross-validation on UniProt-SPEC-100

Fold	**# Classes**	**# Effective Classes**	**F1-score**	**MCC**
1	96	83	0.923	0.908
2	96	84	0.939	0.927
3	96	83	0.902	0.884
Average	96	83.33 ± 0.58	0.921 ± 0.02	0.906 ± 0.02

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

The Matthews Correlation Coefficient (MCC) is less influenced by imbalanced data and is arguably the best single assessment metric in this case [36–38]. The overall performance across all classes is the micro-average of the individual results [39] and we used the multi-class version of MCC [40].

6 Results and Discussion

Table 3 shows the overall performance evaluation results of fine-tuned ProtBERT-BFD with FFNN model on the hold-out test set of the four different datasets. The two columns "Classes" and "Effective Classes" indicate the total number of classes in the dataset (afer applying the threshold of three) and the number of classes the classifier was able to identify in the hold-out test set. For example, UniProt-SPEC-100 was trained on 96 substrate classes; 83 classes out of 96 get a prediction in the hold-out test set and 13 substrate classes do not have either a True Positive (TP) or a False Positive (FP) prediction associated with them, even though they are involved in both the training and hold-out test set. Note that the reported metric values are calculated for effective classes.

The results for UniProt-SPEC-100 outperforms other datasets in all of the metrics. The model fine-tuned on this dataset is selected as *TooT-SS* for predicting the specific substrate prediction of a transporter.

Table 3: Results on hold-out test sets

Dataset	**Classes**	**Effective Classes**	**F1-score**	**MCC**
UniProt-SPEC-100	96	83	0.923	0.908
UniProt-SPEC-60	69	52	0.849	0.818
SwissProt-SPEC-100	86	73	0.872	0.848
SwissProt-SPEC-60	62	44	0.786	0.743

Table 2 shows the results for each fold of the 3-fold cross-validation for the UniProt-SPEC-100 dataset. There is only minor variation in the number of effective classes and in the two metrics F1-score and MCC.

For *TooT-SS* details of performance on the hold-out test set for each class a shown in Table 6. The classes which could not be predicted, that is, non-effective, have metric values reported with NaN for Precision and MCC in Table 6. Each of these 13 classes has five or fewer sequences in the dataset.

On a positive note, Table 6 shows that 34 out of 71 classes with less than 10 sequences in the dataset have a MCC of 1.0 so they are predicted completely correct.

The non-effective classes were foreseeable due to the dataset imbalance and the low number of training samples for these classes. The performance of *TooT-SS* on effective small classes supports the ProtBERT-BFD language model's potential for transfer learning in the classification of small classes in imbalanced datasets.

However, it is important to mention the impact of randomness in the test-train ratio on the final results. Due to the minimum of 3 sequences set on the number of samples in each class, there are 62 classes with only one sequence in their test set. Prediction of one test sequence based on two sequences in the training set is dependent on the randomness of the test-train split.

Therefore, the experiment is repeated two more times to include each of the samples of the smallest dataset separately in the test set. Table 2 shows the results of this 3-fold cross-validation experiment. They indicate that the performance is consistent on the three folds. The model predicted one extra class reaching 84 effective classes for the second fold. Furthermore, the classes that are non-effective vary from fold to fold though the total number is consistent between 12 to 13. Table 4 gives more detail on the full set of 24 classes that are non-effective in at least one of the folds. It is not surprising that the randomness in the test-train choice or the fold split has such an impact.

Table 4: MCC of unclassified classes in each fold

Substrate	Fold1	Fold2	Fold3
copper(1+)	0.925	0.913	NaN
2-oxoglutarate(2-)	0.816	0.816	NaN
maltose	0.707	NaN	0.499
iron(3+)	1.000	NaN	0.407
hydrogencarbonate	0.515	NaN	0.499
N-ribosylnicotinamide	NaN	0.707	1.000
gamma-aminobutyric acid zwitterion	0.707	NaN	NaN
L-arabinose	NaN	0.499	NaN
bacteriocin	0.707	1.000	NaN
L-histidine zwitterion	NaN	-0.001	NaN
L-methionine zwitterion	NaN	NaN	NaN
boric acid	NaN	1.000	0.707
maltodextrin	NaN	1.000	1.000
folate(2-)	NaN	NaN	NaN
molybdate	NaN	NaN	NaN
N-retinylidenephosphatidylethanolamine	1.000	NaN	NaN
sn-glycerol 3-phosphate(2-)	NaN	NaN	NaN
flavin adenine dinucleotide	NaN	NaN	0.707
CMP-N-acetyl-beta-neuraminate(2-)	0.707	1.000	NaN
L-tryptophan zwitterion	-0.001	NaN	0.577
phosphonatoenolpyruvate	1.000	NaN	1.000
guanine	NaN	1.000	1.000
xanthine	NaN	1.000	1.000
glycine betaine	NaN	NaN	NaN

7 Conclusion

In the domain of transport proteins, the expensive process of experimental characterization of the data results in a lack of labeled proteins. Transfer learning has been purposed to solve the problem. Applying pre-trained protein language models, such as ProtBERT-BFD, for capturing information is a feasible solution.

This research is a study on the prediction of the specific substrate transported by a transmembrane transport protein using a pre-trained ProtBERT-BFD language model as a foundation for transfer learning. We investigated the ability of transfer learning with ProtBERT-BFD model to predict small classes. Transfer learning using ProtBERT-BFD is a promising approach for further studies *TooT-SS* is the ProtBERT-BFD model with a one-layer FFNN ine-tuned for 13 epochs with the UniProt-SPEC-100 dataset. *TooT-SS* can predict 83 specific substrates (out of 96) and achieves an MCC of 0.91 on the hold-out test set. where there are imbalanced datasets and small classes. Software and datasets will be available at https://github.com/simaataei upon publication.

In our future work, we will continue a focus on small classes and investigate one-shot and zero-shot classification for protein classification tasks.

Acknowledgements This work was supported by Genome Canada, Genome Québec, Natural Sciences & Engineering Research Council of Canada (NSERC), and Concordia University.

References

[1] I. Thiele, B.Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction, Nature Protocols **5**, 93 (2010).

[2] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger et al., ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, IEEE Trans. Pattern Analysis & Machine Intelligence pp. 7112–7127 (2022).

[3] K. Han, M. Wang, L. Zhang, Y. Wang, M. Guo, M. Zhao, Q. Zhao, Y. Zhang, N. Zeng, C. Wang, Predicting ion channels genes and their types with machine learning techniques, Frontiers in Genetics **10**, 399 (2019).

[4] V.S. Reddy, M.H. Saier Jr, BioV Suite–a collection of programs for the study of transport protein evolution, FEBS Journal **279**, 2036 (2012).

[5] F. Aplop, G. Butler, TransATH: Transporter prediction via annotation transfer by homology, ARPN J. of Engineering & Applied Sciences **12** (2017).

[6] O. Dias, M. Rocha, E.C. Ferreira, I. Rocha, Reconstructing genome-scale metabolic models with merlin, Nucleic Acids Research **43**, 3899 (2015).

[7] J. Capela, D. Lagoa, R. Rodrigues, E. Cunha, F. Cruz, A. Barbosa, J. Bastos, D. Lima, E.C. Ferreira, M. Rocha et al., Merlin, an improved framework for the reconstruction of high-quality genome-scale metabolic models, Nucleic Acids Research **50**, 6052 (2022).

[8] E. Cunha, D. Lagoa, J.P. Faria, F. Liu, C.S. Henry, O. Dias, TranSyT, an innovative framework for identifying transport systems, Bioinformatics **39**, btad466 (2023).

[9] N. Loira, A. Zhukova, D.J. Sherman, Pantograph: A template-based method for genome-scale metabolic model reconstruction, J. of Bioinformatics & Computational Biology **13**, 1550006 (2015).

[10] H. Li, V.A. Benedito, M.K. Udvardi, P.X. Zhao, TransportTP: a two-phase classification approach for membrane transporter prediction and characterization, BMC Bioinformatics **10**, 418 (2009).

[11] M.M. Gromiha, Y. Yabuki, Functional discrimination of membrane proteins using machine learning techniques, BMC Bioinformatics **9**, 135 (2008).

[12] Y.Y. Ou, S.A. Chen, M.M. Gromiha, Classification of transporters using efficient radial basis function networks with position-specific scoring matrices & biochemical properties, Proteins: Structure, Function & Bioinformatics **78**, 1789 (2010).

[13] S.A. Chen, Y.Y. Ou, T.Y. Lee, M.M. Gromiha, Prediction of transporter targets using efficient RBF networks with PSSM profiles & biochemical properties, Bioinformatics **27**, 2062 (2011).

[14] N.S. Schaadt, J. Christoph, V. Helms, Classifying substrate specificities of membrane transporters from *Arabidopsis thaliana*, J. Chemical Information & Modeling **50**, 1899 (2010).

[15] N. Schaadt, V. Helms, Functional classification of membrane transporters and channels based on filtered TM/non-TM amino acid composition, Biopolymers **97**, 558 (2012).

[16] A. Barghash, V. Helms, Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs, BMC Bioinformatics **14**, 343 (2013).

[17] N.K. Mishra, J. Chang, P.X. Zhao, Prediction of membrane transport proteins and their substrate specificities using primary sequence information, PLoS ONE **9**, e100278 (2014).

[18] M. Alballa, F. Aplop, G. Butler, TranCEP: Predicting the substrate class of transmembrane transport proteins using compositional, evolutionary, and positional information, PLoS ONE **15**, e0227683 (2020).

[19] M. Alballa, Ph.D. thesis, Concordia University (2020)

[20] M. Alballa, G. Butler, TooT-SC: Predicting eleven substrate classes of transmembrane transport proteins, bioRxiv (2022). 10.1101/2022.01.25.477715

[21] H. Lin, L. Han, C. Cai, Z. Ji, Y. Chen, Prediction of transporter family from protein sequence by support vector machine approach, Proteins: Structure, Function & Bioinformatics **62**, 218 (2006).

[22] Y.F. Liou, T. Vasylenko, C.L. Yeh, W.C. Lin, S.H. Chiu, P. Charoenkwan, L.S. Shu, S.Y. Ho, H.L. Huang, SCMMTP: identifying and characterizing membrane transport proteins using propensity scores of dipeptides, BMC Genomics **16**, S6 (2015).

[23] L. Li, J. Li, W. Xiao, Y. Li, Y. Qin, S. Zhou, H. Yang, Prediction the substrate specificities of membrane transport proteins based on support vector machine and hybrid features, IEEE Trans. Computational Biology & Bioinformatics **13**, 947 (2016).

[24] Q.T. Ho, D.V. Phan, Y.Y. Ou et al., Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters, Analytical Biochemistry **577**, 73 (2019).

[25] S. Ataei, G. Butler, Predicting the specific substrate for transmembrane transport proteins using BERT language model, in *2022 IEEE Conf. on Computational Intelligence in Bioinformatics & Computational Biology* (IEEE, 2022), pp. 1–8

[26] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805 (2018).

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems **30** (2017).

[28] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge & Data Engineering **22**, 1345 (2009).

[29] M. Steinegger, J. Söding, Clustering huge protein sequence sets in linear time, Nature Communications **9**, 1 (2018).
[30] The UniProt Consortium, UniProt: the universal protein knowledgebase in 2023, Nucleic Acids Research **51**, D523 (2022). 10.1093/nar/gkac1052
[31] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering & comparing biological sequences, Bioinformatics **26**, 680 (2010).
[32] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence **5**, 221 (2016).
[33] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014).
[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., PyTorch: An imperative style, high-performance deep learning library, Advances in Neural Information Processing Systems **32** (2019).
[35] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv:1910.03771 (2019).
[36] Z. Ding, Ph.D. thesis, Georgia State University (2011)
[37] G.M. Weiss, F. Provost, Learning when training data are costly: The effect of class distribution on tree induction, Journal of Artificial Intelligence Research **19**, 315 (2003).
[38] M. Bekkar, H.K. Djemaa, T.A. Alitouche, Evaluation measures for models assessment over imbalanced data sets, J. Information Engineering & Applications **3**, 27 (2013).
[39] C.D. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval (Cambridge University Press, 2008)
[40] J. Gorodkin, Comparing two K-category assignments by a K-category correlation coefficient, Computational Biology & Chemistry **28**, 367 (2004).

Table 5: Substrates in Four Datasets (after threshold)

	CHEBI	Substrate	SP-60	UP-60	SP-100	UP-100
0	CHEBI:24636	proton	443	563	813	1,226
1	CHEBI:29108	calcium(2+)	153	209	347	865
2	CHEBI:29103	potassium(1+)	146	193	299	726
3	CHEBI:17996	chloride	70	124	165	442
4	CHEBI:29101	sodium(1+)	69	83	148	266
5	CHEBI:29105	zinc(2+)	47	51	67	80
6	CHEBI:16189	sulfate	17	33	20	67
7	CHEBI:28938	ammonium	21	26	32	52
8	CHEBI:15377	water	8	9	38	43
9	CHEBI:57986	riboflavin(1-)	22	23	29	37
10	CHEBI:36080	protein	19	18	28	28
11	CHEBI:24337	glutathione derivative	5	7	13	27
12	CHEBI:50211	retinol	4	8	15	27
13	CHEBI:17632	nitrate	17	17	25	26
14	CHEBI:29033	iron(2+)	13	16	16	24
15	CHEBI:15361	pyruvate	8	9	13	20
16	CHEBI:30616	ATP(4-)	2	5	9	18
17	CHEBI:49552	copper(1+)	6	9	7	17
18	CHEBI:16199	urea	4	4	9	16
19	CHEBI:58070	dehydroascorbide(1-)	4	4	13	14
20	CHEBI:15354	choline	6	6	9	13
21	CHEBI:16113	cholesterol	4	3	9	12
22	CHEBI:46911	L-ornithinium(1+)	4	7	7	12
23	CHEBI:35491	L-cystine zwitterion	8	7	11	11
24	CHEBI:61082	lipid-linked peptidoglycan	10	10	11	11
25	CHEBI:3098	bile acid	1	1	4	10
26	CHEBI:29985	L-glutamate(1-)	3	3	8	10
27	CHEBI:16947	citrate(3-)	4	3	8	10
28	CHEBI:30823	oleate	2	3	7	9
29	CHEBI:59789	S-adenosyl-L-methionine zwitterion	3	7	4	9
30	CHEBI:30411	cobalamin	8	8	9	9
31	CHEBI:30413	heme	3	4	6	9
32	CHEBI:58367	UDP-D-glucose(2-)	6	7	8	9
33	CHEBI:57925	glutathionate(1-)	3	3	8	8
34	CHEBI:58339	3'-phosphonato-5'-adenylyl sulfate(4-)	4	4	5	8
35	CHEBI:326268	1,4-butanediammonium	7	6	8	8
36	CHEBI:57705	UDP-N-acetyl-alpha-D-glucosamine(2-)	6	6	6	8
37	CHEBI:29036	copper(2+)	5	7	6	8
38	CHEBI:57586	biotinate	6	6	6	8
39	CHEBI:46819	urate salt	1	3	4	8
40	CHEBI:17992	sucrose	5	5	7	7
41	CHEBI:57288	acetyl-CoA(4-)	0	2	0	7
42	CHEBI:16810	2-oxoglutarate(2-)	3	4	4	7
43	CHEBI:29034	iron(3+)	5	5	7	7
44	CHEBI:17544	hydrogencarbonate	1	2	4	7
45	CHEBI:61304	phosphoglycerate	3	4	3	7
46	CHEBI:17306	maltose	6	5	7	7
47	CHEBI:32682	L-argininium(1+)	3	3	6	6
48	CHEBI:30408	iron-sulfur cluster	1	0	4	6
49	CHEBI:60039	L-proline zwitterion	3	2	6	6
50	CHEBI:17051	fluoride	3	5	4	6
51	CHEBI:13389	NAD	3	3	6	6
52	CHEBI:15740	formate	4	4	4	6
53	CHEBI:15355	acetylcholine	4	4	5	6
54	CHEBI:15927	N-ribosylnicotinamide	5	5	5	5
55	CHEBI:59888	gamma-aminobutyric acid zwitterion	4	4	5	5
56	CHEBI:25197	mercury cation	4	4	5	5
57	CHEBI:37554	fatty acyl-CoA	2	1	4	5
58	CHEBI:63063	cadmium cation	2	2	5	5
59	CHEBI:26441	pyrimidine nucleotide	2	4	3	5
60	CHEBI:64608	GDP-fucose	3	3	3	5
61	CHEBI:35780	phosphate ion	5	5	5	5
62	CHEBI:38290	L-ascorbate	1	3	1	4
63	CHEBI:57932	D-methionine zwitterion	4	4	4	4
64	CHEBI:15904	long-chain fatty acid	3	4	3	4
65	CHEBI:25140	maltodextrin	2	2	4	4
66	CHEBI:57844	L-methionine zwitterion	4	4	4	4
67	CHEBI:57595	L-histidine zwitterion	2	2	3	4
68	CHEBI:26333	prostaglandin	1	1	3	4
69	CHEBI:30849	L-arabinose	4	4	4	4
70	CHEBI:57305	glycine zwitterion	2	2	2	4
71	CHEBI:33118	boric acid	3	3	4	4
72	CHEBI:48081	bacteriocin	2	4	2	4
73	CHEBI:71501	c-GMP-AMP(2-)	1	1	3	3
74	CHEBI:24265	gluconate	2	2	3	3
75	CHEBI:17750	glycine betaine	2	2	3	3
76	CHEBI:60903	N-(4-aminobenzoyl)-L-glutamate	1	3	1	3
77	CHEBI:16412	lipopolysaccharide	1	2	1	3
78	CHEBI:57632	UDP-alpha-D-xylose(2-)	1	1	3	3
79	CHEBI:30049	teichoic acid	2	3	2	3
80	CHEBI:28793	beta-D-glucan	1	1	3	3
81	CHEBI:24040	flavin adenine dinucleotide	2	2	3	3
82	CHEBI:16235	guanine	2	2	3	3
83	CHEBI:62501	folate(2-)	0	1	2	3
84	CHEBI:22629	arsenate ion	3	3	3	3
85	CHEBI:36264	molybdate	3	3	3	3
86	CHEBI:58702	phosphonatoenolpyruvate	3	3	3	3
87	CHEBI:71063	N-retinylidenephosphatidylethanolamine	1	1	2	3
88	CHEBI:46502	tungstate	3	3	3	3
89	CHEBI:16704	uridine	2	2	3	3
90	CHEBI:39127	magnesium cation	2	2	3	3
91	CHEBI:57597	sn-glycerol 3-phosphate(2-)	2	1	3	3
92	CHEBI:57912	L-tryptophan zwitterion	2	2	3	3
93	CHEBI:15318	xanthine	3	3	3	3
94	CHEBI:57812	CMP-N-acetyl-beta-neuraminate(2-)	1	1	2	3
95	CHEBI:16347	(R)-carnitine	1	1	3	3

Table 6: Results for hold-out test set with *TooT-SS* trained on UniProt-SPEC-100

No.	Substrate	Train	Test	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
0	proton	817	409	407	46	2	1030	0.9677	0.8985	0.9951	0.9443	0.9240
1	calcium(2+)	576	289	276	21	13	1175	0.9771	0.9293	0.9550	0.9420	0.9279
2	potassium(1+)	484	242	218	7	24	1236	0.9791	0.9689	0.9008	0.9336	0.9221
3	chloride	295	147	135	5	12	1333	0.9886	0.9643	0.9184	0.9408	0.9348
4	sodium(1+)	178	88	72	5	16	1392	0.9859	0.9351	0.8182	0.8727	0.8674
5	zinc(2+)	53	27	26	1	1	1457	0.9987	0.9630	0.9630	0.9630	0.9623
6	sulfate	44	23	19	0	4	1462	0.9973	**1.0000**	0.8261	0.9048	0.9077
7	ammonium	35	17	17	2	0	1466	0.9987	0.8947	**1.0000**	0.9444	0.9453
8	water	29	14	14	0	0	1471	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
9	riboflavin(1-)	24	13	11	0	2	1472	0.9987	**1.0000**	0.8462	0.9167	0.9192
10	protein	18	10	7	0	3	1475	0.9980	**1.0000**	0.7000	0.8235	0.8358
11	glutathione derivative	18	9	9	0	0	1476	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
12	retinol	18	9	8	0	1	1476	0.9993	**1.0000**	0.8889	0.9412	0.9425
13	nitrate	18	8	5	1	3	1476	0.9973	0.8333	0.6250	0.7143	0.7204
14	iron(2+)	16	8	8	1	0	1476	0.9993	0.8889	**1.0000**	0.9412	0.9425
15	pyruvate	14	6	6	0	0	1479	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
16	ATP(4-)	12	6	6	0	0	1479	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
17	copper(1+)	11	6	6	1	0	1478	0.9993	0.8571	**1.0000**	0.9231	0.9255
18	urea	10	6	6	0	0	1479	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
19	dehydroascorbide(1-)	9	5	5	0	0	1480	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
20	choline	9	4	4	0	0	1481	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
21	cholesterol	8	4	4	1	0	1480	0.9993	0.8000	**1.0000**	0.8889	0.8941
22	L-ornithinium(1+)	8	4	4	1	0	1480	0.9993	0.8000	**1.0000**	0.8889	0.8941
23	lipid-linked peptidoglycan	8	3	2	0	1	1482	0.9993	**1.0000**	0.6667	0.8000	0.8162
24	L-cystine zwitterion	8	3	2	1	1	1481	0.9987	0.6667	0.6667	0.6667	0.6660
25	bile acid	6	4	4	0	0	1481	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
26	L-glutamate(1-)	7	3	2	2	1	1480	0.9980	0.5000	0.6667	0.5714	0.5764
27	citrate(3-)	7	3	3	0	0	1482	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
28	cobalamin	6	3	2	2	1	1480	0.9980	0.5000	0.6667	0.5714	0.5764
29	oleate	6	3	3	2	0	1480	0.9987	0.6000	**1.0000**	0.7500	0.7741
30	heme	6	3	3	0	0	1482	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
31	S-adenosyl-L-methionine zwitterion	6	3	2	0	1	1482	0.9993	**1.0000**	0.6667	0.8000	0.8162
32	UDP-D-glucose(2-)	6	3	3	0	0	1482	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
33	UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	2	0	1	1482	0.9993	**1.0000**	0.6667	0.8000	0.8162
34	1-4-butanediammonium	6	2	0	2	2	1481	0.9973	0.0000	0.0000	0.0000	-0.0013
35	copper(2+)	5	3	2	0	1	1482	0.9993	**1.0000**	0.6667	0.8000	0.8162
36	3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
37	glutathionate(1-)	6	2	2	1	0	1482	0.9993	0.6667	**1.0000**	0.8000	0.8162
38	urate salt	5	3	2	0	1	1482	0.9993	**1.0000**	0.6667	0.8000	0.8162
39	biotinate	5	3	1	0	2	1482	0.9987	**1.0000**	0.3333	0.5000	0.5770
40	sucrose	4	3	3	0	0	1482	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
41	acetyl-CoA(4-)	5	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
42	2-oxoglutarate(2-)	5	2	2	1	0	1482	0.9993	0.6667	**1.0000**	0.8000	0.8162
43	maltose	5	2	2	2	0	1481	0.9987	0.5000	**1.0000**	0.6667	0.7066
44	phosphoglycerate	4	3	3	0	0	1482	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
45	iron(3+)	5	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
46	hydrogencarbonate	4	3	2	3	1	1479	0.9973	0.4000	0.6667	0.5000	0.5152
47	iron-sulfur cluster	4	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
48	L-proline zwitterion	4	2	0	1	2	1482	0.9980	0.0000	0.0000	0.0000	-0.0010
49	L-argininium(1+)	4	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
50	NAD	4	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
51	acetylcholine	4	2	1	0	1	1483	0.9993	**1.0000**	0.5000	0.6667	0.7069
52	fluoride	4	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
53	formate	4	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
54	pyrimidine nucleotide	3	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
55	cadmium cation	3	2	2	1	0	1482	0.9993	0.6667	**1.0000**	0.8000	0.8162
56	GDP-fucose	3	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
57	fatty acyl-CoA	3	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
58	mercury cation	4	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
59	phosphate ion	4	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
60	N-ribosylnicotinamide	4	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
61	gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	0.9993	**1.0000**	0.5000	0.6667	0.7069
62	long-chain fatty acid	3	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
63	L-arabinose	2	2	0	0	2	1483	0.9987	NaN	0.0000	0.0000	NaN
64	bacteriocin	3	1	1	1	0	1483	0.9993	0.5000	**1.0000**	0.6667	0.7069
65	glycine zwitterion	3	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
66	L-histidine zwitterion	3	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
67	L-ascorbate	3	1	1	1	0	1483	0.9993	0.5000	**1.0000**	0.6667	0.7069
68	L-methionine zwitterion	3	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
69	boric acid	2	2	0	0	2	1483	0.9987	NaN	0.0000	0.0000	NaN
70	D-methionine zwitterion	2	2	2	0	0	1483	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
71	prostaglandin	3	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
72	maltodextrin	3	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
73	folate(2-)	2	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
74	molybdate	2	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
75	N-retinylidenephosphatidylethanolamine	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
76	lipopolysaccharide	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
77	beta-D-glucan	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
78	gluconate	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
79	tungstate	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
80	sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
81	UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
82	magnesium cation	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
83	flavin adenine dinucleotide	2	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
84	(R)-carnitine	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
85	CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	0.9993	0.5000	**1.0000**	0.6667	0.7069
86	c-GMP-AMP(2-)	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
87	L-tryptophan zwitterion	2	1	0	1	1	1483	0.9987	0.0000	0.0000	0.0000	-0.0007
88	N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
89	phosphonatoenolpyruvate	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
90	arsenate ion	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
91	teichoic acid	2	1	1	0	0	1484	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
92	guanine	2	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
93	xanthine	2	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
94	glycine betaine	2	1	0	0	1	1484	0.9993	NaN	0.0000	0.0000	NaN
95	uridine	2	1	0	1	1	1483	0.9987	0.0000	0.0000	0.0000	-0.0007
Macro Average	-	-	-	-	-	-	-	0.998	None	0.758	0.744	-
Micro Average	-	-	-	-	-	-	-	0.923	0.923	0.923	0.923	0.908