Received: 10 July 2023 | Revised: 29 September 2023 | Accepted: 3 October 2023

DOI: 10.1002/jcb.30490

RESEARCH ARTICLE

Journal of Cellular Biochemistry WILEY

PortPred: Exploiting deep learning embeddings of amino acid sequences for the identification of transporter proteins and their substrates

Marco Anteghini[1,2,3] | **Vitor AP Martins dos Santos**[1,4] | **Edoardo Saccenti**[2]

[1]LifeGlimmer GmbH, Berlin, Germany

[2]Department of Systems and Synthetic Biology, Wageningen University & Research, Wageningen WE, The Netherlands

[3]Department of Visual and Data-Centric Computing, Zuse Institute Berlin, Berlin, Germany

[4]Department of Bioprocess Engineering, Wageningen University & Research, Wageningen WE, The Netherlands

Correspondence
Marco Anteghini and Edoardo Saccenti, Department of Systems and Synthetic Biology, Wageningen University & Research, Wageningen WE, The Netherlands.
Email: marco.anteghini@wur.nl and edoardo.saccenti@wur.nl

Funding information
HORIZON EUROPE Marie Sklodowska-Curie Actions; European Union's Horizon 2020 research and innovation program, Grant/Award Number: 812968

Abstract

The physiology of every living cell is regulated at some level by transporter proteins which constitute a relevant portion of membrane-bound proteins and are involved in the movement of ions, small and macromolecules across biomembranes. The importance of transporter proteins is unquestionable. The prediction and study of previously unknown transporters can lead to the discovery of new biological pathways, drugs and treatments. Here we present PortPred, a tool to accurately identify transporter proteins and their substrate starting from the protein amino acid sequence. PortPred successfully combines pre-trained deep learning-based protein embeddings and machine learning classification approaches and outperforms other state-of-the-art methods. In addition, we present a comparison of the most promising protein sequence embeddings (Unirep, SeqVec, ProteinBERT, ESM-1b) and their performances for this specific task.

KEYWORDS
membrane proteins, pre-trained embeddings, protein sequence embeddings, substrates prediction, transporter proteins

1 | INTRODUCTION

Since the first work by Rothman, Schekman and Sü dhof focused on unraveling the cell transport mechanisms[1–3] and the identification of the first water channel proteins, later called aquaporins, in 1985 by Benga,[4] the research on transporter proteins has continuously increased. Transporter proteins are now considered essential for the functioning of all living organisms; malfunctioning of transporters is often associated with diseases and they are frequently studied as drug targets.[5–7]

A transporter or membrane transport protein is a protein involved in the transport of ions, small molecules and macro-molecules across a biological membrane.[8,9] Transporter proteins are continuously identified and characterized. Nowadays, they are represented and

TABLE 1 Descriptions of substrate-specific transporters for all the classes considered in this study.

Substrate	Transporter description
Amino acid	Transporters for amino acid molecules that are mainly of the solute carrier family. Examples are transporters from the Amino Acid-Polyamine-Organocation (APC) Superfamily.[27]
Anion	The organic anion transporter (OAT) subfamily constitutes roughly half of the SLC22 (solute carrier 22) transporter family.[28]
Cation/hydrogen ion	Protein involved in the transport of hydrogen ions across a membrane. Used to power processes such as ATP synthesis and bacterial flagellar rotation.[29]
Electron	Electron transporter proteins form chains (ETCs) where each ETC is a series of protein complexes and other molecules that transfer electrons from electron donors to electron acceptors via redox reactions.[30]
Protein/mRNA	Membrane proteins involved in the movement of macromolecules, such as another protein or mRNA.[3,31]
Sugar	The sugar transporters are responsible for the binding and transport of carbohydrates, organic alcohols, and acids in a wide range of organisms.[32]
Lipid	ATP-dependent ABC and P4-ATPase lipid transporters are known to contribute to lipid translocation across the lipid bilayers on the cellular membranes.[33]
Other	This category includes all transporters that are not represented in the other classes. For example transporter proteins that move metal ions like iron, nickel, copper, and zinc.[34]

Abbreviation: mRNA, messenger RNA.

classified in the transporter classification system (http://www.tcdb.org/) that systematically classifies transport proteins according to their mode of transport, energy coupling mechanism, molecular phylogeny, and substrate specificity.[10,11]

The biological relevance of transporter proteins is reflected by the expanding body of research literature on the topic: between 2018 and 2022 18 295 papers appear on PUBMED (https://pubmed.ncbi.nlm.nih.gov/) containing the words "transporter protein" or "transporters" in their title or abstract, while 16 365 were found as published in the previous 4 years. The protein structure database PDB (https://www.rcsb.org/),[12] contains 5150 structures with resolution≤1.5 Å (14-12-2022), corresponding to 3424 proteins identified as transporters. UniProt reports 7391 reviewed sequences annotated as a transporter (14-12-2022); if the search is related to sequences automatically annotated, the number of transporters increases to a staggering 373 477. The huge amount of sequence data compared to structural data concerning transporter proteins indicates the necessity to rely on efficient sequence-based predictors to accurately identify transporter proteins.

Several tools predicting transporter proteins from amino acid sequences using machine learning approaches have emerged in recent years. At the time of this writing, the following tools were published: (1) Transporter Substrate Specificity Prediction (TrSSP);[13] (2) SCMMTP;[14] (3) Li et al.[15] approach; (4) FastTrans;[16] (5) TooT-T;[17] (6) TooT-BERT-T;[18] (7) TranCEP.[19]

All these tools exploit the amino acid composition of the protein sequences. However, the application of deep learning (DL) approaches to encode protein amino acid sequences has shown promising results for several tasks such as subcellular and sub-organelle classification, protein structure and function prediction, and protein-protein interactions (PPIs).[20–26]

Following our previous works on the use of sequence embeddings for the prediction of the subcellular localization of peroxisomal proteins,[22,26] we apply a similar framework to the development of PortPred, a prediction tool for the accurate identification of transporter proteins and multi-class classification of their transported substrates.

We reviewed and compared the most recent and frequently used DL-based protein embeddings (namely: UniRep,[20] SeqVec,[21] ProteinBERT,[25] and ESM-1b[23]) in predicting transporter proteins and their relative substrates, in combination with several machine learning approaches.

PortPred was developed by testing both the single embeddings and their combination thereof and finding

the best protein representation and machine learning classifier. PortPred was also tested against the state-of-the-art transporters predictors[13–17,19] for either binary classification (transporter vs. not-transporter) and multi-class classification related to the transporter substrates namely: cation, anion, electron, lipid, amino acid, protein/messenger RNA (mRNA), sugar, others. Details about the substrate are shown in Table 1.

We found PortPred to outperform the state-of-the-art in predicting transporter proteins and being accurate in predicting different substrates. Finally, after in-depth analysis of our results, we implemented PortPred which couples a combination of the four embeddings and Logistic Regression to perform the predictions.

2 | METHODS

2.1 | Database search queries

We obtained the number of transporter proteins available in biological databases, mentioned in the Introduction (Section 1), with the following queries.

PubMed query 1: `((transporter protein [Title/Abstract]) OR (transporters[Title/Abstract])) AND (("2014"[Date - Publication] : "2018"[Date - Publication])).`

PubMed query 2: `((transporter protein [Title/Abstract]) OR (transporters[Title/Abstract])) AND (("2018"[Date - Publication] : "3000"[Date - Publication])).`

PDB query: `(Structure Keywords HAS ANY OF WORDS "transport, transporter, transporters, transporter protein") AND (Refinement Resolution = [0 - 1.5]).`

UniProt query: `(cc_function:transporter)`

2.2 | Overview of existing methods for the prediction of transporter proteins and their substrate

Transporter Substrate Specificity Prediction[13]: It implements Support Vector Machines (SVM) classifier on six prediction modules considering the following features respectively: (1) amino acid composition;[35] (2) AAIndex that considers the biochemical composition of the amino acid residues.[36,37] In particular, a subset of the AAIndex database which has 49 selected physical, chemical, energetic, and conformational properties;[37] (3) The position-specific scoring matrix profile (PSSM)[38,39] run on the Swissprot data set.[29] PSSM captures the conservation pattern in the alignment and summarizes evolutionary information of the protein where the scoring matrix is at the basis of protein BLAST searches (BLAST and PSI-BLAST);[40] (4) combination of AAIndex/PSSM with the Swissport based PSSM; (5) PSSM run on UniRef90;[41] (6) a combination of AAIndex/PSSM (UniRef90).

Scoring card method (*SCM*) *for membrane transport proteins* (*SCMMTP*)[14]: It implements a SCM based on the dipeptide composition of the amino acid sequence to identify putative membrane transport proteins. In SCMMTP, the first step is the creation of a matrix of (20×20) 400 dipeptides which represents the normalized dipeptide propensity scores of the Membrane Transport Proteins (MTPs). This matrix is then optimized using the Improved Genetic Algorithm (IGA) for maximum satisfiability (MAX-SAT) problems,.[42] IGA optimizes the dipeptide propensity scores maximizing the prediction accuracy and conserving the original sequence information. The fitness function of IGA is concerned with the area under the receiver operating characteristic (ROC) curve (area under the curve [AUC])[43] and Pearson's correlation coefficient between the initial and optimized propensity scores of 20 amino acids.

Li et al.[15]: This approach first creates a hybrid feature representation of the amino acid sequence which integrates the PSSM,[38] the amino acid composition, biochemical properties from the PROFEAT (Protein Features),[44] and Gene Ontology (GO) terms. The hybrid feature is created by recursively selecting features using an SVM-based backward feature extraction model which is used to predict the substrate class of transmembrane transport proteins.

FastTrans[16]: It generates a word-embedding representation of the protein sequence implementing a natural language processing (NLP) approach. First, biological words are generated by splitting the amino acid sequence into overlapping fragments of the same length. Second, a word embedding vector for each biological word is generated using Skip gram[45] or Continuous Bags of Words (CBWO) models.[46] The classification is performed using SVM.[47]

TooT-T[17]: It is an ensemble classifier that combines the predictions from homology annotation transfer and machine-learning classifiers. The ensemble classifier uses six predictions (three from the homology annotation transfer and three from SVM classifiers) and outputs the final binary prediction (transporter vs non-transporter). It is implemented using the Gradient Boosting Machine, as available by caret package in R https://CRAN.R-project.org/package=caret. Given a query protein, this method starts with a homology search of the Transporter Classification Database[11] using BLAST.[48] The query

sequence is classified as a transporter if a hit is found using three predetermined sets of thresholds, thus generating the three homology modeling annotation transfer predictions. Second, three variations of newly generated features called psi-composition feature psiAAC, psiPAAC, and psiPseAAC are computed.[17] Psi-composition combines amino acid composition with alignment results from PSI-BLAST.[40] These psi-composition features are then used as input for three SVM classification models.

TooT-BERT-T[18]: This tool harnesses the power of BERT[49] representations to dissect and distinguish between transporters and non-transporters. The core of this approach lies in the utilization of a Logistic Regression classifier,[50] which leverages BERT's deep contextual understanding. The performances of both frozen and fine-tuned representations, originating from two distinct BERT models have been validated.

TranCEP[19]: It uses the pair amino acid composition (PAAC) encoding scheme, the TM-Coffee algorithm for generating multiple sequence alignments (MSAs),[51] and its relative transitive consistency score (TCS).[52] The predictor relies on eight SVM classifiers, one for distinguishing between each pair of classes of substrates. TranCEP has been further expanded into **TooT-Sc** which encompass a diverse data set of eleven substrate classes. The methodology underpinning TooT-SC combines the power of pairwise amino acid composition (PAAC), evolutionary insights from MSAs facilitated by TM-Coffee, and precise focus on critical alignment positions through TCS. Notably, experimental evaluations have showcased the remarkable performance of TooT-SC.[53,54]

2.2.1 | Software

We report the links to the tools mentioned and tested in this study (if available).

- TrSSP—https://www.zhaolab.org/TrSSP/
- SCMMTP—http://iclab.life.nctu.edu.tw/iclab_webtools/SCMMTP/
- FastTrans—http://bio216.bioinfo.yzu.edu.tw/fasttrans/
- TranCEP—https://github.com/bioinformatics-group/TranCEP

2.3 | DL based protein sequence embeddings

We considered four recently proposed methods for the embedding of protein sequences based on deep-learning approaches and protein sequences:

The unified representation (*UniRep*)[20]: Is based on a 1900-hidden unit recurrent neural network architecture, able to capture evolutionary, chemical and biological information encoded in the protein sequence starting from 24 million UniRef50 sequences[41] where UniRef50 is a non-redundant sub-cluster of Uniprot.[29] In UniRep, the protein sequence is modeled by using a hidden state vector, which is recursively updated based on the previously hidden state vector. That means the method learns by scanning a sequence of amino acids, predicting the next one based on the sequence it has seen before. Using UniRep, a protein sequence can be represented by an embedding with a length of 64, 256, or 1900 units, depending on the neural network architecture. In this study, we used the 1900 units length (average final hidden array). For a detailed explanation of how to retrieve the UniRep embedding, we refer the reader to the specific GitHub repository https://github.com/churchlab/UniRep (11.2021) or the bio-embeddings GitHub repository https://github.com/sacdallago/bio_embeddings.

The sequence-to-vector embedding (*SeqVec*)[21]: Is based on a NLP approach. It embeds biophysical information of a protein sequence where amino acids are words and proteins are sentences. SeqVec is obtained by training ELMo,[55] on UniRef50.[41] ELMo is a deep contextualized word representation that models both complex characteristics of word use (e.g., syntax and semantics) and how these vary across linguistic contexts. It consists of a two-layer bidirectional LSTM[56] backbone pre-trained on a large text corpus. The SeqVec embedding can be obtained based on either a per-residue level (word level) or a per-protein level (sentence level). The per-residue level protein sequence embedding is informative in predicting the secondary structure or intrinsically disordered region; The per-protein level is useful to predict subcellular localization and to distinguish membrane-bound versus water-soluble proteins.[21] Here we use the per-protein level representation, where the protein sequence is represented by an embedding of length 1024. For a detailed explanation of how to retrieve the SeqVec embedding, we refer the reader to the specific GitHub repository https://github.com/mheinzinger/SeqVec or the bio-embeddings repository https://github.com/sacdallago/bio_embeddings.

ProteinBert[57]: Is inspired by the Bidirectional Encoder Representations from Transformers (BERT) which is a DL model that utilizes a transformer architecture to pretrain on large amounts of unlabeled text data, enabling it to generate high-quality contextualized word representations for various NLP tasks.[49] ProteinBERT was instead pretrained on the raw protein sequences available in Uniref100 (~106 million proteins).[41,49] The

original BERT model is trained on two tasks: (1) language modeling where 15% of tokens are masked and the model predicts the masked tokens from context; (2) next sentence prediction where BERT is trained to predict the probability of a chosen next sentence given the first sentence. BERT learns contextual embeddings for words and can be finetuned on small data sets for optimized predictions on specific tasks.[49] In ProteinBERT sequences are treated as separate documents, where the "next" sentence prediction is not used. The masking procedure works by training randomly masked protein sequences, similar to the original BERT model. In particular, the model takes a sequence (sentence) as input, masks 15% of the amino acids (words) from it and is asked to output the complete sequence. ProteinBert was pretrained on two simultaneous tasks. (1) Bidirectional language modeling of protein sequences (2) GO annotation prediction, which captures diverse protein functions.[58] The final embedding has a length of 1024. For a detailed explanation of how to retrieve the ProteinBert embedding, we refer the reader to the specific GitHub https://github.com/nadavbra/protein_bert.

The evolutionary scale modeling-1b (*ESM-1b*): Was trained on 250 million sequences of the UniParc database[29] and relies on a deep transformer architecture,[59,60] a powerful model architecture for representation learning and generative modeling in NLP. The peculiarity of the transformer architecture is that it is able to return for each amino acid (word) of the sequence (sentence), an embedding with contextual information. In other terms, it compares every amino acid (word) in the sequence (sentence) to every other amino acid (word) in the sequence (sentence), including itself, and reweighs the embeddings of each word. The modules responsible for this process are called self-attention blocks and consist of three main steps: (1) Dot product similarity and alignment scores; (2) Scores normalization and embedding weight; (3) Reweighing of the original embeddings. In ESM-1b, the transformer processes inputs through a series of blocks that alternate self-attention with feed-forward connections. In this case, since it has been trained on proteins, the self-attention blocks construct pairwise interactions between all positions in the sequence, so that the transformer architecture represents residue-residue interactions. In addition, ESM-1b was trained using the masked language modeling objective[60] which forces the model to identify dependencies between the masked site and the unmasked parts of the sequence to make the prediction of the masked parts. Finally, the model was optimized scaling the identified hyperparameters to train a model with~650 M parameters (33 layers) on the UR50/S data set, resulting in the ESM-1b Transformer.[23] The final length of the ESM-1b vector is 1280.

2.4 | Overview of PortPred development and benchmarking

The overall strategy for the development of the PortPred tool for the prediction of transporter proteins and their substrates is schematized in Figure 1. It consists of 4 main steps: (1) Curation of protein sequence data; (2) Generation of the embeddings (ESM-b1, UniRep, SeqVec, ProteinBERT) of the amino acid sequence; (3) Evaluation of different ML approaches; (4) Benchmarking with available tools.

2.5 | Data sets

Our ML architecture was trained on three different training sets. Training set 1 is a newly generated data set (the PortPred data set); Training set 2 is the TrSSP training set and Training set 3 is the FastTrans training set.[13,16] It was then tested against three different validation sets. Validation set 1 is an independent data set containing Peroxisomal proteins, Validation set 2 is an independent data set from the TrSSP predictor,[13] and Validation set 3 is an independent data set from the FastTrans predictor.[16] Each Validation set is independent from each Training set.

Training set 1, Training set 2 as well as Validation set 1 and Validation set 2 were used as benchmarks. See Sections 2.5.3, 2.5.4, and 2.5.1 for details. The newly generated data set, that contains peroxisomal proteins, was used as a specific real-world use case (see Section 2.5.2).

Some of the used data sets have differences among the classes of transporter proteins, in particular in Training set 2 the class "lipid" is not present while it is the only one containing the class "anion." Please note that the PortPred final tool does not have the "anion" label. A complete summary of the used data sets is available in Table 2.

Moreover, we invite researchers to consider the approach introduced by Alballa and Butler,[61] which aims to streamline the process, reduce subjectivity in data set curation, and eliminate external data set curator judgment with an automated tool.[61]

2.5.1 | Training set 1, the PortPred data set

Given the high percentage of sequence similarity (70%) present in the data set available in the literature (see 2.5.3), that could be considered redundant, we defined a novel data set that we consider more reliable for the final model training. The proteins were retrieved

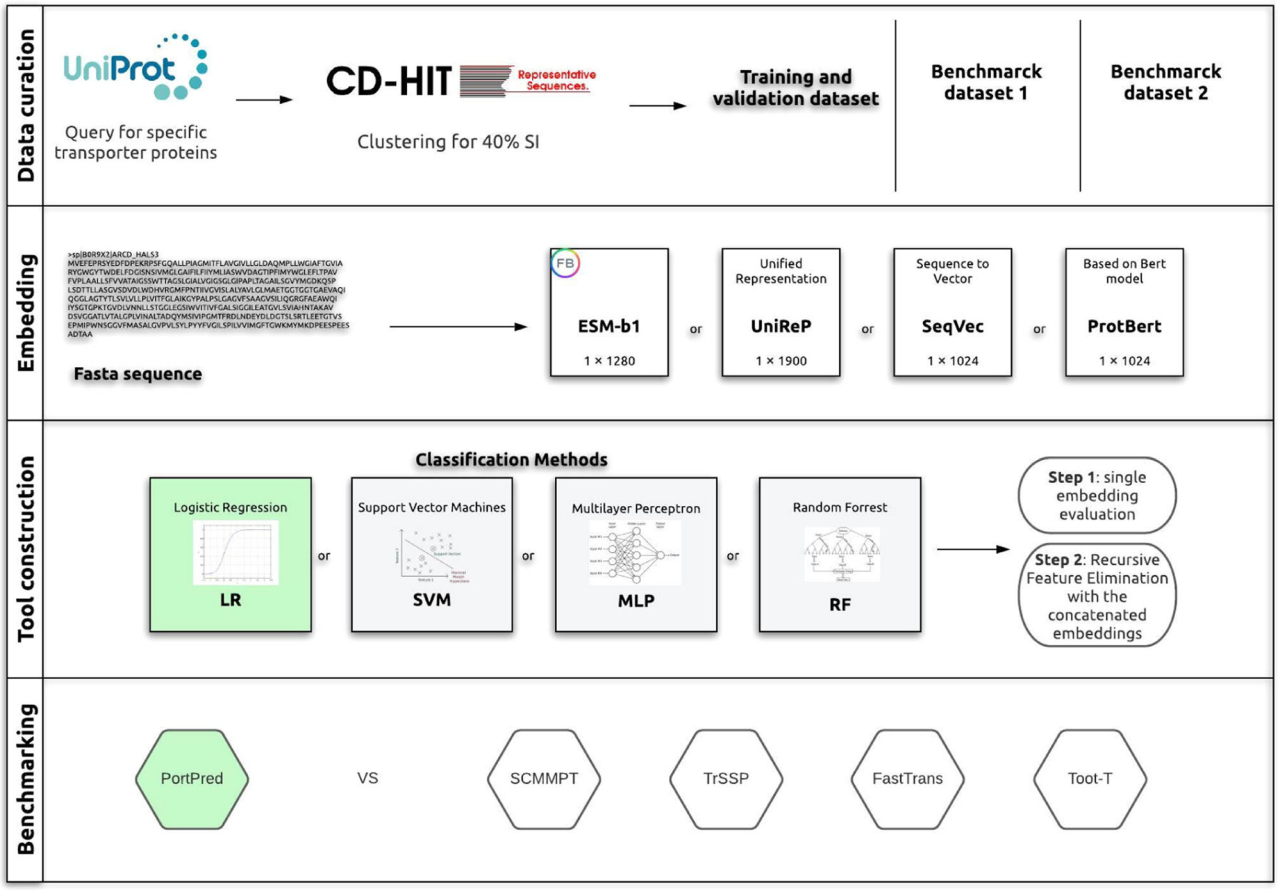

FIGURE 1 Overview of the PortPred development. Data curation: retrieval and selection of protein sequences (see Section 2.5). Embedding: conversion of protein sequences to standard encodings, namely: ESM-b1, unified representation (UniRep), sequence-to-vector (SeqVec), and ProteinBERT (ProtBERT). PortPred construction: application of different classification algorithms (Section 2.6), evaluation and selection of the best single embedding (step 1), evaluation and selection of the best combination of sequence embeddings using recursive feature elimination (RFE) (see Section 2.7.3) (step 2). Benchmarking: comparison of PortPred tool and or data set with the transporter classifiers available in literature: Scoring Card Method for Membrane Transport Proteins (SCMMPT); Transporter Substrate Specificity Prediction (TrSSP); FastTrans; TooT-T.

from Uniprot (02-10-2021)[29] obtaining 6631 transporter protein sequences and 19 139 non-transporter sequences. The data set was clustered using cd-hit[62] for 40% of sequence identity and filtered to not overlap with the Training set 2 (see Section 2.5.3). We obtained a data set containing 1781 transporter proteins divided into 7 classes, namely: hydrogen ion transporters (116), electron transporters (262), lipid transporters, amino acid transporters (92), protein/mRNA transporters (656), sugar transporters (125), others transporters (465) which include calcium, cobalt, copper, porin, iron, potassium, sodium, zinc, nickel, neurotransmitter, oxygen, phosphate, sulfate and ammonia. An example query that can be applied to each class changing the specific keyword is: '(locations:(location:membrane) AND (reviewed:true) AND (keyword:KW-0762) AND (fragment:no) AND (existence: "evidence at protein levels")'. That means our data set only contains manually curated proteins. We also retrieved 1781 non-transporter proteins as negatives among our random sample of non-transporter proteins. The negatives were obtained searching for reviewed membrane proteins not associated to the keyword transport. The query is the following: '(reviewed:true) NOT (keyword:KW-0813) AND (fragment:no) AND (existence: "evidence at protein levels") AND (locations:(location:membrane)'. Moreover, we checked that 100% of the proteins are associated to a publication that confirms the class label for all classes including the negatives. Given some limitations in the generation of the embeddings for very long protein sequences (e.g., ESM-b1 does not embed proteins longer than 1024 residues), we removed them from the data set, which finally consists of a balanced and non-redundant data set of 1580 positives entries and 1621 negatives entries. The data set is available at https://github.com/MarcoAnteghini. An overview can be seen in Table 2.

TABLE 2 The six data sets used in this study.

	Amino acid	Anion	Cation/ hydrogen ion	Electron	Protein/ mRNA	Sugar	Lipid	Other	Positives	Negatives	Total
Training set 2	70	60	260	60	70	60	N.A.	200	780	600	1380
Validation set 2	15	12	36	10	15	12	N.A.	20	120	60	180
Training set 3	61	N.A.	73	184	380	71	66	165	1000	875	1875
Validation set 3	12	N.A.	15	37	75	13	12	33	197	167	372
Training set 1	92	N.A.	116	262	656	125	65	465	1781	1781	3562
Validation set 1	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	173

Note: In bold are Training set 1 and Validation set 1 which are newly generated data sets from this work. Validation set 1 is an independent data set which contains peroxisomal proteins. Note that the Validation set 1 does not contain information about the specific transporter substrates and it is used as a real-world case scenario. Training set 2 is the training set used in the Transporter Substrate Specificity Prediction (TrSSP) paper. The Validation set 2 is an independent data set also from the TrSSP paper. Training set 3 is the training set used in the FastTrans paper. The Validation set 3 is an independent data set also from the FastTrans paper.

Abbreviation: mRNA, messenger RNA.

2.5.2 | Validation set 1, a peroxisomal proteins data set

A data set for a specific use case scenario was created using peroxisomal protein only. We searched on Uniprot (20/03/2022) for reviewed peroxisomal proteins correlated to the GO term "transport," obtaining 173 entries using the query 'locations:(location:"Peroxisome membrane [SL-0203]") goa:("peroxisomal membrane [5778]") goa:("transport [6810]") AND reviewed:yes)'. In Table 2 it is shown that the transporter protein class and true transporter function of the peroxisomal protein in unknown when building the data set. This was decided to create a real-world use case scenario.

2.5.3 | Training set 2 and validation set 2 from TrSSP data set

The data set (composed of training and independent validation sets) for the model benchmarking with other available predictors was the same used in all of them.[14,16,63] This benchmarking data set (Training set 2 and Validation set 2) provided by Mishra et al.,[13] is collected from the Swiss-Prot database[29] release 2013_03 and has been filtered considering the 70% of sequence similarity using CD-HIT.[62] The TrSSP data set[13] contains a total of 1560 sequences, divided into Training set 2 and Validation set 2 as shown in Table 2. The categories related to the 900 transporters present in the data set are 85 amino acid/oligopeptide transporters, 72 anion transporters, 296 cation transporters, 70 electron transporters, 85 protein/mRNA transporters, 72 sugar transporters, 220 other transporters. Also, 660 non-transporters were included as negatives. The data set can either be found at https://www.zhaolab.org/TrSSP/?dowhat=datasets or on GitHub at https://github.com/MarcoAnteghini.

2.5.4 | Training set 3 and validation set 3 from FastTrans data set

As an additional data set for benchmarking our approach with the multiclass prediction, we used the same data set used for FastTrans by Nguyen et al.[16] This data set is divided into Training set 3 and Validation set 3 (see Table 2). The protein sequence in this data set was retrieved from UniProt[29] (release 2018_10) and contained proteins involved in the biological process of transporting ions/molecules. The data set does not contain fragmented sequences and sequences annotated with more than two substrate specificities. In addition, sequences with more than 20% similarity were removed using PSI Blast.[40] The data set consists of 1050 membrane proteins (negatives) and 1197 transporters (positives). Note that the hydrogen ion substrate category from Nguyen et al.[16] is either called hydrogen ion or cation and represents the same set of proteins.

2.6 | Classification algorithms

The determination of transporter and non-transporter proteins is easily translated into a binary classification

problem, while to distinguish among substrate categories we used a multi-class classification approach. For both tasks, we considered four widely used classification algorithms.

Support Vector Machines: Is a supervised learning algorithm for two-group classification which aims to find the maximal margin hyperplane separating the points in the feature space.[47,64] SVMs also perform non-linear classifications applying the kernel trick, thus implicitly mapping their inputs into high-dimensional feature spaces. In the case of multiple classes, multiple binary classification problems are performed. It can be done in two ways[65]: (1) *One-vs-One*, a binary classifier per each pair of classes; (2) *One-vs-Rest*, a binary classifier per class. In this study, we used the One-vs-Rest approach.

Random Forest (*RF*): Is an ensemble learning method that, in the case of a classification task constructs a multitude of decision trees and outputs the mode of the classes of the individual trees.[66,67]

Multilayer Perceptron (*MLP*): Is a class of feed-forward artificial neural networks that can distinguish among non-linearly separable data and uses backpropagation for training.[68,69] Each node in an MLP, with the exception of the input node, uses a nonlinear activation function. In this study, we used the ReLu activation function.[70]

Logistic Regression (*LR*): Estimates the parameters of a logistic model.[71] In binary classifications, the corresponding probability of the values associated with two different labels can vary between 0 and 1. The multinomial LR model, for K possible outcomes, runs K-1 independent binary logistic regression models, in which one outcome is chosen as a "pivot" and then the other K-1 outcomes are separately regressed against the pivot outcome. We used a penalized implementation of multivariable logistic regression.[50]

2.7 | PortPred implementation

2.7.1 | Model training and validation

In this study, we used three different training sets with no overlap between them and three independent validation sets.

Training

To be consistent with the other methods, each model was evaluated on the training data sets, respectively, using 10-fold cross-validation (10-CV).[72] In every iteration, a single fold was kept as the testing set, and the remaining nine sets were used to train the respective model. The trained model was then tested using the test set. The procedure stops when all 10 subsets are used as a test once. The average performance for each model was considered as a single estimation. To obtain a stable error estimation, we repeated the 10-CV 10 times with different random splits. The variations between runs were highlighted by the standard deviation (*SD*). The cross-validation performances are reported as mean± *SD* of the 10 different runs of the 10-CV.

The cross-validation procedures include a (hyper)grid search: for each set of hyperparameters, the average classification score is computed across the folds. The hyperparameters corresponding to the best classification score are then used to fit a classification model whose quality is assessed on the validation set. The reference metric is the F1 score. The Hyperparameters optimization details are shown in Table 3).

Validation

The independent data sets were used to perform additional validations. The data in the independent validation sets were not used during the cross-validation processes and are completely unknown to the models.

2.7.2 | Concatenation of embeddings

To obtain a comprehensive overview of the single embeddings capabilities, we first evaluated each model using a single embedding and finally, we run a training and test procedure where every protein was represented with a concatenation of all the available embeddings.

2.7.3 | Recursive feature elimination

Recursive Features Elimination (RFE) defines an optimal subset of informative features with respect to a given task. It starts considering all features in the training data set (the 4 concatenated embeddings in our case) and successfully removes one or more of them until the performance worsens or an arbitrary number of features remains. The performance is evaluated through a CV (10-CV here) classification. The approach creates a model where the desired input is a hybrid version of all the analysed embeddings. In particular, just the relevant features (values) of the concatenated embedding are kept (e.g., 2328 out of 5228). We used the RFECV function, available on scikit-learn that automatically selects the number of features chosen by RFE.[74] We adopted Logistic

TABLE 3 Hyperparameters for the grid searches.

Method	Hyperparameters	Description	Search space
SVM	C	Inverse of regularization strength	logspace(−2,10,13)
	gamma	Kernel coefficient	logspace(−9,3,13)
	kernel	Specifies the kernel type to be used in the algorithm	'linear','poly','rbf','sigmoid'
RF	n_estimators	The number of trees in the forest	15,25,50,75,100,200,300
	criterion	The function to measure the quality of a split	'gini','entropy'
	max_depth	The maximum depth of the tree	2,5,10,None
	min_samples_split	The minimum number of samples required to split an internal node	2,4,8,10
	max_features	The number of features to consider when looking for the best split	'sqtr','auto','log2'
MLP	hidden_layer_sizes	The number of neurons in each hidden layer	(200,),(100,),(50,),(200,100,6,1)
	activation	Activation function for the hidden layer	'relu'
	solver	The solver for weight optimization	'lbfgs'
	alpha	Strength of the L2 regularization term	1.0
	learning_rate	Learning rate schedule for weight updates	'constant'
LR	penalty	Specify the norm of the penalty	'l1','l2'
	solver	Algorithm to use in the optimization problem	'liblinear','saga'
	C	Inverse of regularization strength	logspace(−3,9,13)

Note: The logspace function, as available on NumPy[73] returns numbers spaced evenly on a log scale. In logscale(start, stop, numbers), the sequence starts at base start (base to the power of start), ends with base stop and numbers is the number of samples to generate. The listed methods are: Support Vector Machines (SVM); Random Forest (RF); Multilayer Perceptron (MLP); Logistic Regression (LR).

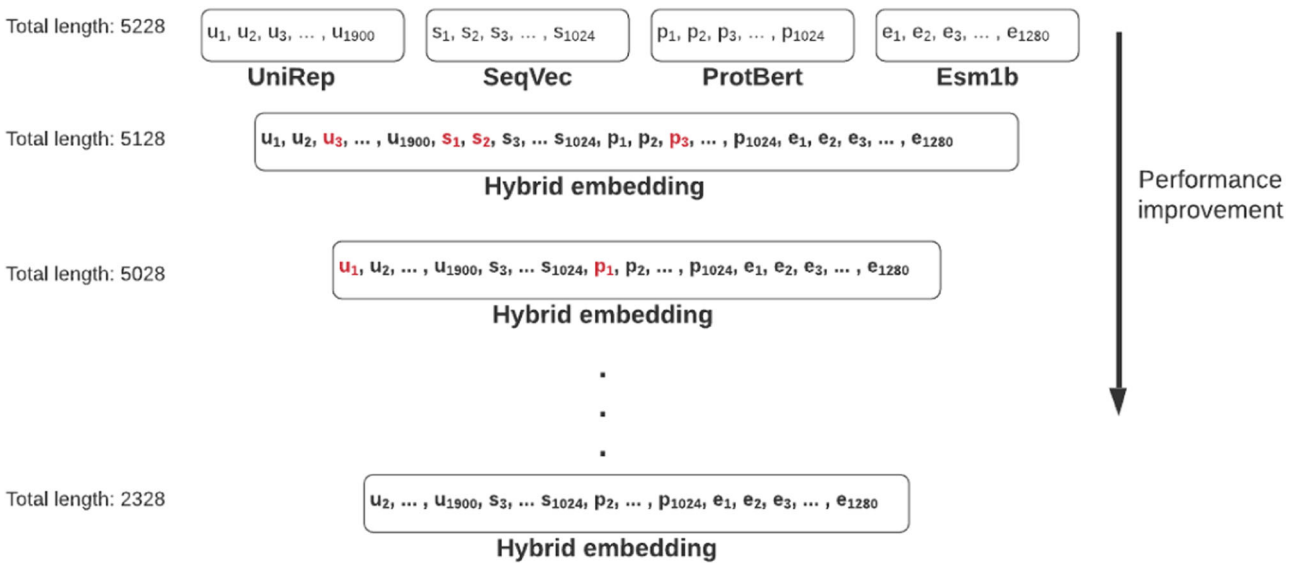

FIGURE 2 Schematic representation of the recursive feature elimination process. The initial data set contains a vector of length 5228. Each iteration remove a fixed number of random features, in this case, 100. The performance is then evaluated with the reduced embedding and the process continues until it worsens or the minimum number of features to consider has been reached. The final vector (length 2328) representation is saved.

Regression as an estimator within the RFECV function, given its consistency during our initial estimations and its capability of working with both binary and multiclass classification tasks. In the RFECV function, the number of features to remove at each iteration must be specified, we used 100 to have a granular but fast process. The chosen metric for the performance optimization was the F1 score. A detailed explanation of the metric can be seen in Section 2.8. An overview of the process is shown in Figure 2.

2.8 | Metrics

We used several metrics to quantify the quality of the classification models, namely: sensitivity (SEN), specificity (SPE), accuracy (ACC), F_1 score,[75] Matthews correlation coefficient (MCC)[76] and the AUC of the ROC. Given that *TP* is the number of true positives, FP is the number of false positives; TN and FN are the numbers of true and false negatives, respectively, the following formulas are defined as:

Sensitivity (SEN) or True positive rate (TPR) is defined as

$$\mathrm{SEN/TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

Specificity (SPE) is defined as

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2}$$

Accuracy (ACC) is defined as

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

F_1 score[75] is defined as

$$F_1 = 2 \times \frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} \tag{4}$$

where PPV is the positive predicted value (or precision)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

TheF_1 score is the harmonic mean of recall and precision and varies between 0, if the precision or the recall is 0, and 1 indicating perfect precision and recall.

Matthews correlation coefficient (MCC)[76] is defined as

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{6}$$

MCC is the correlation coefficient between the true ad predicted class: it is bound between¯−1 (total disagreement between prediction and observation) and +1 (perfect prediction); 0 indicates no better than random prediction. The MCC is appropriate also in presence of class unbalance.[77]

The AUC of the ROC curve which plots the true positive proportion or the Sensitivity against the Specificity, is defined as

$$\begin{aligned} \mathrm{AUC} &= \int_{x=0}^{1} \mathrm{TPR}(\mathrm{FPR}^{-1}(x))dx \\ &= \int_{\infty}^{-\infty} \mathrm{TPR}(T)\mathrm{FPR}'(T)dT \end{aligned} \tag{7}$$

The AUC analysis enables the evaluation of the performance of a binary classifier system according to the variation of the discrimination threshold. A perfect prediction has an AUC score of 1.0 while an AUC of 0.5 indicates randomness.[78]

2.9 | Data availability and software

The data that support the findings of this study are openly available in PortPred at https://github.com/MarcoAnteghini/PortPred. A stand-alone version of the tool is available at https://github.com/MarcoAnteghini/PortPred. Moreover, the data sets, together with an explanatory Jupyter notebook, are available at https://drive.google.com/drive/folders/1L_zdaDa2EoPTWQzOdNqSHCweQFixcsHe.

The final version of PortPred generates an hybrid representation of the four concatenated embeddings and perform its classification with a Logistic Regression.

3 | RESULTS

3.1 | Embeddings correlation

Different embeddings store different information and in some cases, concatenating two or more embeddings can improve the performances.[22] In this study we first checked for a possible correlation between the embeddings using Pearson's correlation coefficient.[79] We observed that combining four different encodings

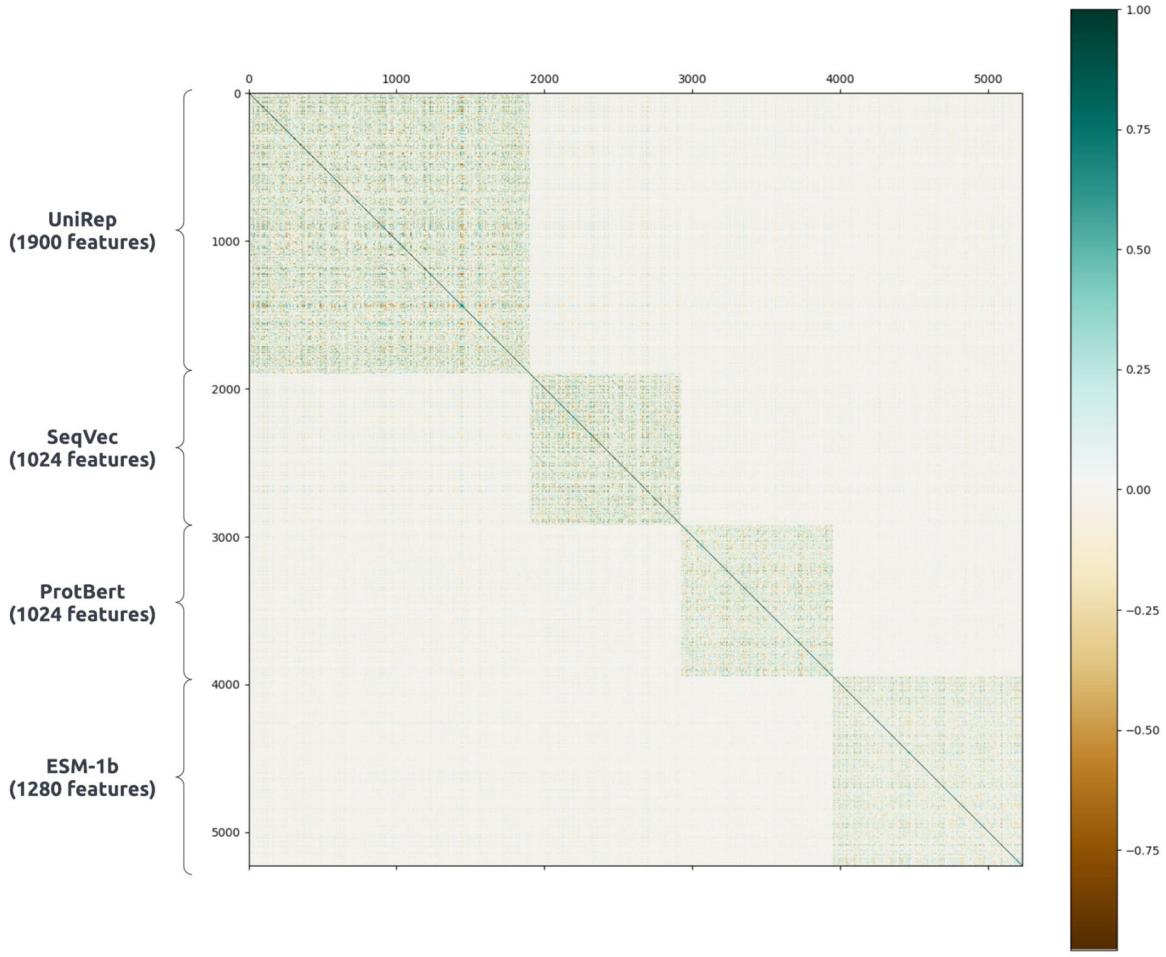

FIGURE 3 Correlation among UniRep (1900 features), SeqVec (1024 features), PortPred (1024 features), and ESM-1b (1280 features) protein sequence embeddings. Pearson's linear correlation is used and are calculated over 1580 transporter protein sequences of the PorPred training set. The four embeddings are uncorrelated.

and/or embeddings gives a better prediction of transporter proteins and their substrates. In particular, concatenating UniRep, SeqVec, ProteinBERT and ESM-1b showed a noticeable improvement in the performances. That indicates that the four embeddings carry different and complementary information about the properties of the protein sequence, as given in Figure 3, which shows how the four embeddings are not correlated.

3.2 | PortPred development

We selected Logistic Regression for the final models to predict transporter proteins and their substrates, using the PortPred data set. These models were subsequently employed in the final tool. For the binary classification task, we obtained a model with the following hyperparameters: C=10.0, class_weight={0:1, 1:1}, solver='liblinear'. For the multiclass classification task, we derived a model with the following hyperparameters: C=1000000.0, class_weight={0:1, 1:1}, multi_class='multinomial', random_state=10, solver='newton-cg'.

The final tool consists of two classification steps. First, the classifier distinguishes between transporter and non-transporter proteins. Second, a multiclass classification is performed to assign proteins to specific transporter categories, which include lipid, sugar, protein/mRNA, electron, hydrogen ion, amino acid, and other. A schematic representation of the tool's functionalities is provided in Figure 4.

3.2.1 | Transporter versus non-transporter prediction on Training set 1

We analysed the performances of each embedding for a binary classification task ("transporter" vs. "non-transporter") on the Training set 1 and the Validation

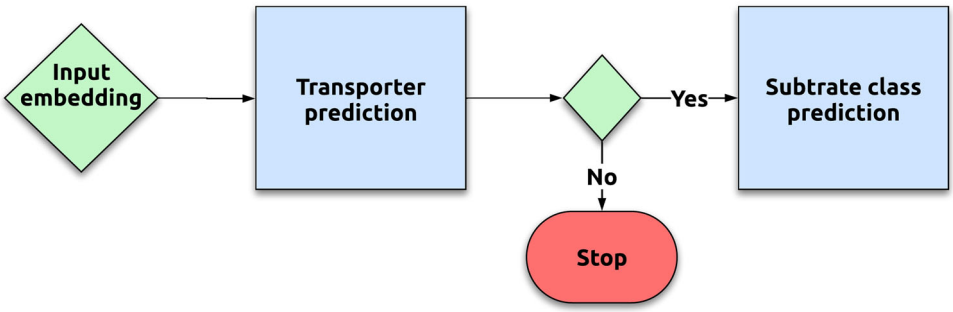

FIGURE 4 A schematic representation of the PortPred tool functionalities in three steps. (1) The algorithm takes as input the embedding of a protein sequence; (2) IF condition. If the protein is predicted as transporter the algorithm proceed, otherwise it stops; (3) The algorithm predicts which substrate the transporter carries. The two prediction steps (1) transporter vs. non-transporter; (2) substrate class prediction relies on two separate models, independently trained.

set 2 (reported in table as Ind.), explained in Section 2.5. As classifiers, we tested LR, RF, SVM, and MLP (see Section 2.6). The results in terms of sensitivity, specificity, accuracy, AUC and f1 score are shown in Table 4 (see Section 2.8 for details about the metrics). The ESM-1b embedding coupled with an SVM classifier reaches the best performances with an F1 score of 84.65% and ACC of 84.67% during cross-validation.

Second, we analysed the concatenated embeddings performances, and the hybrid embeddings performances (details in Section 2.7.3) on the same data set. Results are shown in Table 5. In this case, the hybrid embeddings obtained with an RFE procedure slightly outperformed the concatenated embeddings. The best classifier in handling the hybrid embedding is SVM, reaching ACC of 84.27% in the cross-validation.

We reported the results with and without the ESM-1b embedding to have an additional comparison since it cannot handle long protein sequences (longer than 1024 residues).

3.2.2 | Validation on the peroxisomal proteins data set

We tested the tool with a subset of peroxisomal proteins called Validation set 1 (Section 2.5.2). The predictor produced promising results, thus allowing us to suggest new transporter protein candidates in peroxisomes. In particular, we analyzed the predictor performances in highlighting transporter proteins in a generic subset of peroxisomal proteins that have been associated with transport functions in Uniprot. A total of 26 proteins out of 167 were identified as non-transporter proteins. Looking into this predicted negative data set, we realized that just 3 out of 26 had a clear transporter function (True Negatives) while the remaining are part of more complex machinery not directly connected with transporter function. For example PEX12 (UniprotID: Q8VC48), predicted as negative, is a peroxisome assembly protein. More precisely, it is a component of a retrotranslocation channel required for peroxisome organization. This proteins only forms a channel once assembled with PEX2 and PEX10. The complete list of predictions is available at https://drive.google.com/drive/folders/1XKnORs8uEb_T61Nhgi0aCx8Pzsc0rRNA.

We performed a manual curation for all the entries (available at https://github.com/MarcoAnteghini/PortPred/blob/main/peroxisomal_proteins_dataset/Validation_set1.csv). From the manual curation the metrics concerning the binary prediction task are ACC: 70.66%, ROC AUC: 82.62%, MCC: 0.4756, SEN: 65.24%, SPE: 1.0%, F1: 65.23%.

3.2.3 | Transporter versus non-transporter prediction on Training set 2 and benchmarking

We first analysed the performances of each embedding for a binary classification task ("transporter" vs. "non-transporter") on the Training set 2 and the Validation set 2 (TrSSP benchmark data set). As classifiers, we tested LR, RF, SVM, and MLP (see Section 2.6). The results in terms of sensitivity, specificity, accuracy, Mattew correlation coefficient, AUC and f1 score are shown in Table 6 (see Section 2.8 for details). The ESM-1b embedding coupled with an SVM classifier reaches the best performances with an F1 score of 85.54% and ACC of 88.70%.

Second, we analysed the concatenated embeddings performances and the hybrid embeddings performances (details in Section 2.7.3) on the same benchmark data set (TrSSP). Results are shown in Table 7. In this case, the

TABLE 4 Performances of each embeddings in predicting transporter proteins in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), Mattew correlation coefficient (MCC), area under the curve (ROC AUC) and F1 score (F1).

(a) UniRep												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**ROC AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	84.17	76.70 ± 0.47	70.00	76.61 ± 0.49	79.44	76.65 ± 0.03	0.5395	0.5333 ± 0.0061	77.08	76.65 ± 0.03	84.52	76.65 ± 0.03
RF	89.17	75.81 ± 0.28	78.33	78.63 ± 0.32	85.56	77.22 ± 0.21	0.6750	0.5451 ± 0.0042	83.75	77.22 ± 0.21	89.17	77.21 ± 0.21
SVM	83.33	77.58 ± 0.82	73.33	77.20 ± 0.66	80.0	77.39 ± 0.27	0.5581	0.5482 ± 0.0053	78.33	77.39 ± 0.27	84.75	77.53 ± 0.26
MLP	90.00	76.24 ± 0.51	75.00	75.08 ± 0.56	85.00	75.66 ± 0.19	0.6587	0.5136 ± 0.0038	82.5	75.66 ± 0.19	88.89	75.65 ± 0.19
(b) SeqVec												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**ROC AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	89.17	78.62 ± 0.38	73.33	79.65 ± 0.55	83.89	79.13 ± 0.35	0.6334	0.5830 ± 0.0070	81.25	79.13 ± 0.35	88.07	79.13 ± 0.35
RF	90.00	74.99 ± 0.50	78.33	81.20 ± 0.56	86.11	78.10 ± 0.33	0.6862	0.5633 ± 0.0067	84.17	78.10 ± 0.33	89.63	78.07 ± 0.33
SVM	91.67	80.65 ± 0.43	81.67	80.16 ± 0.31	88.33	80.40 ± 0.30	0.7365	0.6084 ± 0.0060	86.67	80.40 ± 0.30	91.29	80.40 ± 0.30
MLP	90.83	79.87 ± 0.49	73.33	78.61 ± 0.51	85.00	79.24 ± 0.36	0.6567	0.5850 ± 0.0071	82.08	79.24 ± 0.36	88.98	79.24 ± 0.36
(c) ProteinBERT												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**ROC AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	90.83	78.40 ± 0.41	78.33	79.02 ± 0.50	86.67	78.71 ± 0.29	0.6977	0.5745 ± 0.0059	84.58	78.71 ± 0.29	90.08	78.71 ± 0.29
RF	90.83	78.40 ± 0.39	70.00	76.73 ± 0.33	83.89	77.56 ± 0.22	0.6292	0.5515 ± 0.0044	80.42	77.56 ± 0.22	88.26	77.55 ± 0.21
SVM	93.33	78.70 ± 0.45	73.33	81.09 ± 0.44	86.67	79.89 ± 0.45	0.6934	0.5984 ± 0.0089	83.33	79.90 ± 0.45	90.32	79.89 ± 0.45
MLP	89.17	79.19 ± 0.32	76.67	78.62 ± 0.60	85.00	78.91 ± 0.43	0.6611	0.5784 ± 0.0088	82.92	78.91 ± 0.43	88.80	78.90 ± 0.43
(d) ESM-1b												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**ROC AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	89.72	82.42 ± 0.40	76.36	83.16 ± 0.54	85.19	82.80 ± 0.38	0.6670	0.6560 ± 0.0074	83.04	82.79 ± 0.37	88.89	82.78 ± 0.38
RF	92.52	80.23 ± 0.55	85.45	81.30 ± 0.39	90.12	80.78 ± 0.24	0.7798	0.6157 ± 0.0050	88.99	80.76 ± 0.24	92.52	80.75 ± 0.24
SVM	96.26	84.25 ± 0.58	80.00	85.06 ± 0.57	90.74	84.67 ± 0.41	0.7909	0.6935 ± 0.0080	88.13	84.66 ± 0.41	93.21	84.65 ± 0.41
MLP	90.65	81.44 ± 0.47	74.55	81.61 ± 0.66	85.19	81.53 ± 0.39	0.6648	0.6307 ± 0.0076	82.60	81.52 ± 0.38	88.99	81.51 ± 0.39

Note: Each classifier, namely: Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), Multilayer Perceptron (MLP) was evaluated with a 10-fold cross-validation (see columns CV). Each model was trained on the Training set 1 and then validated on an independent data set (see columns Ind.). The independent data set is the one described in Validation set 2 (Section 2.5). The subtables represent the single embedding performances: (a) UniRep; (b) SeqVec; (c) ProteinBERT; (d) ESM-1b. The results refer to the prediction task "transporter vs non-transporter."

TABLE 5 Performances of the concatenated embeddings based models, trained on the Training set 1 (PortPred data set).

(a) UniRep + SeqVec + ProteinBERT												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**ROC AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	90.08	79.35 ± 0.46	81.36	80.96 ± 0.44	87.22	80.15 ± 0.21	0.7114	0.6035 ± 0.0043	85.72	80.16 ± 0.21	90.46	80.15 ± 0.21
RF	90.32	78.16 ± 0.51	85.71	80.96 ± 0.41	88.89	79.56 ± 0.36	0.7467	0.5917 ± 0.0070	88.02	79.56 ± 0.36	91.8	79.55 ± 0.36
SVM	92.31	79.82 ± 0.41	80.95	81.97 ± 0.37	88.33	80.89 ± 0.26	0.7412	0.6183 ± 0.0052	86.63	80.90 ± 0.26	91.14	80.89 ± 0.26
MLP	89.34	80.26 ± 0.43	81.03	79.43 ± 0.39	86.67	79.85 ± 0.20	0.6977	0.5973 ± 0.0040	85.19	79.84 ± 0.19	90.08	79.84 ± 0.19
(b) RFE − UniRep + SeqVec + ProteinBERT												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**ROC AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	90.00	84.18 ± 0.38	80.00	84.89 ± 0.24	86.67	84.54 ± 0.24	0.7000	0.6909 ± 0.0047	85.00	84.54 ± 0.24	90.00	84.53 ± 0.24
RF	91.20	78.17 ± 0.42	89.09	81.54 ± 0.37	90.56	79.86 ± 0.30	0.7846	0.5978 ± 0.0060	90.15	79.86 ± 0.30	93.06	79.85 ± 0.30
SVM	92.31	84.60 ± 0.36	80.95	83.93 ± 0.61	88.33	84.27 ± 0.36	0.7412	0.6857 ± 0.0073	86.63	84.26 ± 0.36	91.14	84.26 ± 0.36
MLP	87.50	83.60 ± 0.46	84.62	82.98 ± 0.50	86.67	83.29 ± 0.38	0.6934	0.6661 ± 0.0078	86.06	83.29 ± 0.38	90.32	83.29 ± 0.38
(c) UniRep + SeqVec + ProteinBERT + ESM-1b												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**ROC AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	89.19	83.49 ± 0.30	84.31	82.65 ± 0.28	87.65	83.07 ± 0.26	0.7209	0.6617 ± 0.0052	86.75	83.07 ± 0.26	90.83	83.07 ± 0.26
RF	90.09	80.02 ± 0.50	86.27	80.72 ± 0.46	88.89	80.37 ± 0.28	0.7490	0.6078 ± 0.0056	88.18	80.37 ± 0.29	91.74	80.36 ± 0.28
SVM	90.00	84.47 ± 0.50	84.62	83.84 ± 0.41	88.27	84.16 ± 0.39	0.7356	0.6834 ± 0.0078	87.31	84.15 ± 0.39	91.24	84.15 ± 0.39
MLP	88.24	82.75 ± 0.58	71.67	80.82 ± 0.51	82.10	81.80 ± 0.38	0.6109	0.6363 ± 0.0076	79.95	81.78 ± 0.38	86.12	81.78 ± 0.38
(d) RFE − UniRep + SeqVec + ProteinBERT + ESM-1b												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**ROC AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	89.19	83.74 ± 0.42	84.31	83.12 ± 0.60	87.65	83.43 ± 0.43	0.7209	0.6690 ± 0.0088	86.65	83.43 ± 0.43	90.83	83.42 ± 0.43
RF	90.91	80.46 ± 0.52	86.54	81.10 ± 0.50	89.51	80.78 ± 0.38	0.7635	0.6159 ± 0.0077	88.72	80.78 ± 0.38	92.17	80.77 ± 0.38
SVM	90.00	84.18 ± 0.33	84.62	84.37 ± 0.47	88.27	84.27 ± 0.30	0.7356	0.6859 ± 0.0061	87.31	84.27 ± 0.31	91.24	84.27 ± 0.30
MLP	87.27	81.52 ± 0.43	78.85	80.56 ± 0.68	84.57	81.05 ± 0.37	0.6519	0.6212 ± 0.0075	83.06	81.05 ± 0.37	88.48	81.04 ± 0.37

Note: Performances reflect the models capability in predicting transporter proteins in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), Mattew correlation coefficient (MCC), area under the curve (ROC AUC) and F1 score (F1). Each classifier, namely: Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), Multilayer Perceptron (MLP) was evaluated with a 10-fold cross-validation (see columns CV). Each model was then validated on an independent data set (see columns Ind.). The independent data set is the one described in Validation set 2 (Section 2.5). The subtables represent the concatenated embedding performances. (a) Represent the concatenated embeddings performances with the exclusion of ESM-1b; (b) represent the concatenated embeddings performances of all the embeddings; (c) represent the concatenated embeddings performances after performing a Recursive Feature Elimination and with the exclusion of ESM-1b; (d) represent the concatenated embeddings performances after performing a Recursive Feature Elimination. The results refer to the prediction task "transporter vs non-transporter."

TABLE 6 Performances of each embeddings in predicting transporter proteins in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), Mattew correlation coefficient (MCC), area under the curve (AUC) and F1 score (F1).

(a) UniRep												
	SEN %		**SPE %**		**ACC %**		**MCC**		**AUC %**		**F1 %**	
Classifier	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	87.50	83.73 ± 0.78	70.00	82.47 ± 0.60	81.67	83.18 ± 0.42	0.5827	0.6604 ± 0.0080	78.75	83.10 ± 0.39	86.42	82.95 ± 0.42
RF	89.17	81.47 ± 0.81	81.67	87.48 ± 0.62	86.67	84.09 ± 0.33	0.7027	0.6849 ± 0.0058	85.42	84.48 ± 0.29	89.92	83.98 ± 0.33
SVM	86.67	83.58 ± 1.520	76.67	83.80 ± 01.48	83.33	83.67 ± 0.31	0.6283	0.6717 ± 0.0048	81.67	83.69 ± 0.23	87.39	83.47 ± 0.29
MLP	88.33	84.67 ± 0.68	66.67	80.63 ± 0.88	81.11	82.91 ± 0.60	0.5658	0.6533 ± 0.0119	77.50	82.65 ± 0.61	86.18	82.62 ± 0.61
(b) SeqVec												
	SEN %		**SPE %**		**ACC %**		**MCC**		**AUC %**		**F1 %**	
Classifier	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	87.50	85.01 ± 0.63	88.33	85.02 ± 1.04	87.78	85.01 ± 0.55	0.7373	0.6979 ± 0.0114	87.92	85.01 ± 0.58	90.52	84.82 ± 0.56
RF	87.50	82.22 ± 0.42	83.33	88.18 ± 0.78	86.11	84.81 ± 0.37	0.6952	0.6990 ± 0.0078	85.42	85.20 ± 0.40	89.36	84.70 ± 0.38
SVM	86.67	82.35 ± 01.68	91.67	89.85 ± 1.39	88.33	85.61 ± 0.47	0.7556	0.7170 ± 0.0079	89.17	86.10 ± 0.36	90.83	85.52 ± 0.45
MLP	90.00	85.54 ± 0.64	90.00	83.33 ± 01.21	90.00	84.57 ± 0.71	0.7826	0.6874 ± 0.0146	90.00	84.42 ± 0.75	92.31	84.33 ± 0.73
(c) ProteinBERT												
	SEN %		**SPE %**		**ACC %**		**MCC**		**AUC %**		**F1 %**	
Classifier	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	93.33	85.77 ± 0.47	83.33	82.97 ± 0.71	90.0	84.55 ± 0.46	0.7734	0.6871 ± 0.0091	88.33	84.37 ± 0.47	92.56	84.30 ± 0.47
RF	91.67	84.63 ± 0.62	78.33	77.58 ± 0.86	87.22	81.57 ± 0.43	0.7094	0.6249 ± 0.0089	85.00	81.11 ± 0.45	90.53	81.17 ± 0.44
SVM	95.00	84.87 ± 0.13	81.67	84.90 ± 0.91	90.56	84.88 ± 0.57	0.7846	0.6955 ± 0.0106	88.33	84.89 ± 0.50	93.06	84.69 ± 0.55
MLP	95.83	86.00 ± 0.79	81.67	82.70 ± 0.76	91.11	84.57 ± 0.59	0.7972	0.6871 ± 0.0118	88.75	84.35 ± 0.59	93.50	84.30 ± 0.60
(d) ESM-1b												
	SEN %		**SPE %**		**ACC %**		**MCC**		**AUC %**		**F1 %**	
Classifier	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	94.39	88.70 ± 0.31	90.91	85.60 ± 0.37	93.21	87.34 ± 0.28	0.8493	0.7441 ± 0.0052	92.65	87.15 ± 0.28	94.84	87.14 ± 0.28
RF	95.33	88.53 ± 0.43	85.45	78.42 ± 0.88	91.98	84.08 ± 0.37	0.8194	0.6769 ± 0.0078	90.39	83.47 ± 0.41	94.01	83.70 ± 0.39
SVM	92.52	89.47 ± 0.81	90.91	87.71 ± 0.46	91.98	88.70 ± 0.35	0.8241	0.7718 ± 0.0070	91.72	88.59 ± 0.30	93.84	88.54 ± 0.34
MLP	92.52	88.34 ± 01.09	87.27	83.65 ± 01.31	90.74	86.28 ± 0.36	0.7945	0.7224 ± 0.0073	89.9	86.00 ± 0.38	92.96	86.05 ± 0.37

Note: Each classifier, namely: Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), Multilayer Perceptron (MLP) was evaluated with a 10-fold cross-validation (see columns CV) on the Training set 2. Each model was then validated on the Validation set 2 (see columns Ind.). The subtables represent the single embedding performances: (a) UniRep; (b) SeqVec; (c) ProteinBERT; (d) ESM-1b. The results refer to the prediction task "transporter vs non-transporter."

TABLE 7 Performances of the concatenated embeddings based models, trained on the Training set 2.

(a) UniRep + SeqVec + ProteinBERT												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	91.60	85.70 ± 0.70	81.97	85.50 ± 0.5	88.33	85.60 ± 0.40	0.7386	0.7090 ± 0.0007	86.78	85.6 ± 0.40	91.21	85.40 ± 0.40
RF	94.83	82.17 ± 0.57	84.38	88.28 ± 0.93	91.11	84.83 ± 0.53	0.8043	0.6996 ± 0.0112	89.6	85.23 ± 0.56	93.22	84.72 ± 0.54
SVM	93.28	85.6 ± 0.70	85.25	88.17 ± 0.42	90.56	86.72 ± 0.40	0.7884	0.7340 ± 0.0077	89.26	86.88 ± 0.37	92.89	86.58 ± 0.4
MLP	90.83	86.44 ± 0.61	81.67	84.10 ± 0.93	87.78	85.42 ± 0.42	0.7250	0.7049 ± 0.0082	86.25	85.27 ± 0.45	90.83	85.19 ± 0.44
(b) RFE − UniRep + SeqVec + ProteinBERT												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	90.16	85.60 ± 0.58	82.76	85.48 ± 0.61	87.78	85.55 ± 0.45	0.7229	0.7086 ± 0.0090	86.46	85.54 ± 0.45	90.91	85.36 ± 0.46
RF	95.65	82.95 ± 0.23	84.62	88.65 ± 0.69	91.67	85.43 ± 0.32	0.8179	0.7110 ± 0.0071	90.13	85.8 ± 0.36	93.62	85.32 ± 0.33
SVM	90.16	84.99 ± 0.93	82.76	88.72 ± 0.48	87.78	86.61 ± 0.62	0.7229	0.7328 ± 0.0119	86.46	86.85 ± 0.58	90.91	86.48 ± 0.62
MLP	94.07	86.23 ± 0.85	85.48	84.28 ± 1.23	91.11	85.38 ± 0.74	0.8019	0.7043 ± 0.0150	89.78	85.26 ± 0.76	93.28	85.16 ± 0.75
(c) UniRep + SeqVec + ProteinBERT + ESM-1b												
Classifier	**SEN %**		**SPE %**		**ACC %**		**MCC**		**AUC %**		**F1 %**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	94.34	88.46 ± 0.38	87.50	86.95 ± 0.70	91.98	87.79 ± 0.37	0.8219	0.7546 ± 0.0077	90.92	87.70 ± 0.37	93.90	87.62 ± 0.37
RF	95.33	84.42 ± 0.45	90.91	87.84 ± 0.70	93.83	85.92 ± 0.38	0.8624	0.7200 ± 0.0074	93.12	86.13 ± 0.39	95.33	85.80 ± 0.38
SVM	94.44	89.59 ± 0.51	90.74	88.04 ± 0.59	93.21	88.90 ± 0.30	0.8480	0.7764 ± 0.0064	92.59	88.75 ± 0.30	94.88	88.81 ± 0.31
MLP	95.15	88.84 ± 0.45	84.75	85.42 ± 0.73	91.36	87.34 ± 0.47	0.8118	0.7445 ± 0.0094	89.95	87.13 ± 0.49	93.33	87.13 ± 0.48
(d) RFE − UniRep + SeqVec + ProteinBERT + ESM-1b												
Classifier	**SEN**		**SPE**		**ACC**		**MCC**		**ROC AUC**		**F1**	
	Ind.	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
LR	95.24	95.06 ± 0.37	87.72	93.85 ± 0.31	92.59	94.53 ± 0.26	0.8366	0.8896 ± 0.0054	91.48	94.46 ± 0.26	94.34	94.45 ± 0.27
RF	94.39	85.61 ± 0.52	89.09	89.22 ± 0.84	92.59	87.19 ± 0.57	0.8348	0.7452 ± 0.0114	91.74	87.41 ± 0.59	94.39	87.09 ± 0.57
SVM	94.29	95.02 ± 0.48	85.96	93.84 ± 0.48	91.36	94.50 ± 0.33	0.8093	0.8890 ± 0.0067	90.13	94.42 ± 0.33	93.4	94.43 ± 0.32
MLP	95.24	95.01 ± 0.31	87.72	93.71 ± 0.36	92.59	94.44 ± 0.20	0.8366	0.8877 ± 0.0039	91.48	94.36 ± 0.20	94.34	94.35 ± 0.20

Note: Performances reflect the models capability in predicting transporter proteins in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), Mattew correlation coefficient (MCC), area under the curve (ROC AUC) and F1 score (F1). Each classifier, namely: Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), Multilayer Perceptron (MLP) was evaluated with a 10-fold cross-validation (see columns CV). Each model was then tested on the Validation set 2 (see columns Ind.). The subtables represent the concatenated embedding performances. (a) Represent the concatenated embeddings performances with the exclusion of ESM-1b; (b) represent the concatenated embeddings performances of all the embeddings; (c) represent the concatenated embeddings performances after performing a Recursive Feature Elimination and with the exclusion of ESM-1b; (d) represent the concatenated embeddings performances after performing a Recursive Feature Elimination. The results refer to the prediction task "transporter vs. non-transporter."

hybrid embeddings obtained with an RFE procedure outperformed the concatenated embeddings. Both, in general, outperform the single embeddings' performances. The best classifier in handling the hybrid embedding is LR, reaching an F1 score of 94.45% and ACC of 94.53% in the cross-validation.

We reported the results with and without the ESM-1b embedding to have an additional comparison since it cannot handle long protein sequences (longer than 1024 residues). Nevertheless, the average length of a transporter protein sequence found on Swissprot (20.01.2023) is 447 residues and the median is 347. We computed this value by averaging the length of 7420 proteins. These proteins were found with the query `'(cc_function:transporter) AND (length:[50 TO *]) AND (reviewed:true) AND (fragment:false)'`.

Finally, we report the performances of our best classifier trained on the Training set 2 against the results of recently published works.[13,14,16,17] Results are visible in Table 8. Our model outperforms the state-of-the-art in predicting transporter proteins with an ACC of 94.53% on the independent validation set.

TABLE 8 Performance comparison between the proposed method (PortPred) and those of recently published works in predicting transporter proteins trained on the Training set 2 and validated on Validation set 2.

	SEN %		SPE %		ACC %		MCC	
Tool	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**	**Ind.**	**CV**
PortPred	95.24	92.59	87.72	93.85	92.59	94.53	0.84	0.89
SCMMTP	80.00	83.76	68.33	77.68	76.11	81.12	0.47	0.62
TrSSP	76.67	76.67	81.67	78.46	80.00	78.99	0.57	0.58
FastTrans	100.00	83.14	77.50	84.48	85.00	83.94	0.73	0.68
TooT-T	94.17	90.15	88.33	89.97	92.22	90.07	0.82	0.80

Note: Performances are measured in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), Mattew correlation coefficient (MCC), area under the curve (ROC AUC) and F1 score (F1). The evaluation was performed on 10-fold cross-validation data (see CV columns) and on an independent data set (see Ind. columns).

TABLE 9 Performances of the PortPred model, trained on the Training set 1 and the Training set 3 in terms of F1 score (macro average) and Mattew correlation coefficient (MCC).

Data set	Classifier	F1 %		MCC	
		CV	**Ind.**	**CV**	**Ind.**
1		88.24 ± 0.57	89.14	0.9184 ± 0.0031	0.9071
	LR				
3		96.65 ± 0.24	90.26	0.9706 ± 0.0024	0.9137
1		64.13 ± 0.29	66.63	0.7406 ± 0.0064	0.7512
	RF				
3		61.42 ± 0.37	65.04	0.7099 ± 0.0067	0.7619
1		88.15 ± 0.61	88.92	0.9183 ± 0.0044	0.8980
	SVM				
3		88.44 ± 0.67	89.89	0.8967 ± 0.0058	0.9135
1		86.55 ± 0.64	88.51	0.9057 ± 0.0032	0.9274
	MLP				
3		87.45 ±;0.66	91.56	0.8900 ± 0.0058	0.9283

Note: The first column indicates the data set; the second column indicates the classifier among Logistic Regression (LR), Random forest (RF), Support Vector Machines (SVM) and Multilayer perceptron (MLP); from the third column the performances are shown in terms of F1 and MCC, whit the CV indicating the cross-validation process on the specific Training set (1 or 3) and the column Ind. indicating the performances of the model trained on the specific Training set (1 or 3) but tested only against Validation set 3. The results refer to the prediction task "substrate category prediction."

TABLE 10 Performances of the PortPred model in a multiclass prediction task.

Training set	Substrate	SEN		SPE		ACC		AUC		MCC	
		CV	Ind.	CV	Ind.	CV	Ind.	CV	Ind.	CV	Ind.
1		0.90 ± 0.10	0.83	0.99 ± 0.01	0.99	0.99 ± 0.01	0.98	0.95 ± 0.05	0.91	0.91 ± 0.06	0.86
	Amino acid										
3		0.89 ± 0.12	0.75	0.99 ± 0.01	0.99	0.99 ± 0.01	0.98	0.94 ± 0.04	0.87	0.93 ± 0.07	0.81
1		0.94 ± 0.02	0.99	0.99 ± 0.01	0.99	0.99 ± 0.01	0.99	0.98 ± 0.02	0.99	0.95 ± 0.02	0.99
	Electron										
3		0.99 ± 0.02	0.99	0.99 ± 0.01	0.99	0.99 ± 0.01	0.98	0.99 ± 0.01	0.99	0.98 ± 0.02	0.95
1		0.90 ± 0.08	0.99	0.99 ± 0.01	0.99	0.99 ± 0.01	0.99	0.95 ± 0.04	0.99	0.93 ± 0.03	0.99
	Hydrogen ion										
3		0.91 ± 0.11	0.93	0.99 ± 0.01	0.98	0.99 ± 0.01	0.99	0.95 ± 0.06	0.96	0.93 ± 0.09	0.93
1		0.50 ± 0.16	0.99	0.99 ± 0.01	0.99	0.98 ± 0.01	0.99	0.75 ± 0.08	0.99	0.64 ± 0.12	0.99
	Lipid										
3		0.87 ± 0.11	0.78	0.99 ± 0.01	0.99	0.99 ± 0.01	0.99	0.93 ± 0.06	0.89	0.92 ± 0.08	0.88
1		0.98 ± 0.01	0.99	0.98 ± 0.01	0.99	0.98 ± 0.01	0.99	0.98 ± 0.01	0.99	0.95 ± 0.01	0.99
	Protein/mRNA										
3		0.99 ± 0.01	0.99	0.98 ± 0.01	0.98	0.99 ± 0.01	0.98	0.99 ± 0.01	0.98	0.98 ± 0.03	0.96
1		0.93 ± 0.09	0.99	0.99 ± 0.01	0.99	0.99 ± 0.01	0.99	0.96 ± 0.05	0.99	0.92 ± 0.09	0.96
	Sugar										
3		0.99 ± 0.01	0.99	0.99 ± 0.01	0.99	0.99 ± 0.01	0.99	0.99 ± 0.01	0.99	0.96 ± 0.03	0.93
1		0.94 ± 0.02	0.97	0.97 ± 0.19	0.99	0.96 ± 0.01	0.99	0.95 ± 0.01	0.98	0.90 ± 0.02	0.96
	Others										
3		0.98 ± 0.03	0.91	0.99 ± 0.01	0.99	0.99 ± 0.01	0.98	0.98 ± 0.02	0.95	0.95 ± 0.04	0.92

Note: The model was trained on Training set 1 and Training set 3. Results are shown in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), Mattew correlation coefficient (MCC) and area under the curve (ROC AUC) (F1). The first column indicates the data set; the second column indicates the substrates; from the third column the performances are shown, whit the CV indicating the cross-validation process on the specific Training set (1 or 3) and the column Ind. indicating the performances on the model trained on the specific Training set (1 or 3) but tested only against Validation set 3. The results refer to the prediction task "substrate category prediction."

Abbreviation: mRNA, messenger RNA.

3.2.4 | Transporter substrate categories prediction and benchmarking

Our method's capability in predicting substrate-specific transporter proteins is shown in Table 9. The results in terms of F1 score and MCC show the consistency of our method when trained on two different data sets (Training set 1 and Training set 3) and validated on the same independent data set (Validation set 3). The Training set 3 comes from the work of Nguyen et al.,[16] reported as FastTrans. Training set 1 is a newly generated PortPred data set. For details about the data set refer to Section 2.5.

As an additional comparison, Table 10 reports similar performances of PortPred trained with Training set 1 and Training set 3 in classifying specific kinds of transporter proteins. Moreover performances of our PortPred model validated on the Validation set 3, retrieved from the FastTrans paper, are visible as confusion matrix in Figure 5.

4 | DISCUSSION AND CONCLUSION

Transporter proteins play a crucial role in the transport of ions, small molecules, and macromolecules across biological membranes. They are essential for the functioning of all living organisms and are frequently studied as drug targets due to their association with various diseases. The research on transporter proteins has significantly increased since their first discovery and characterization. In this study, we focused on developing

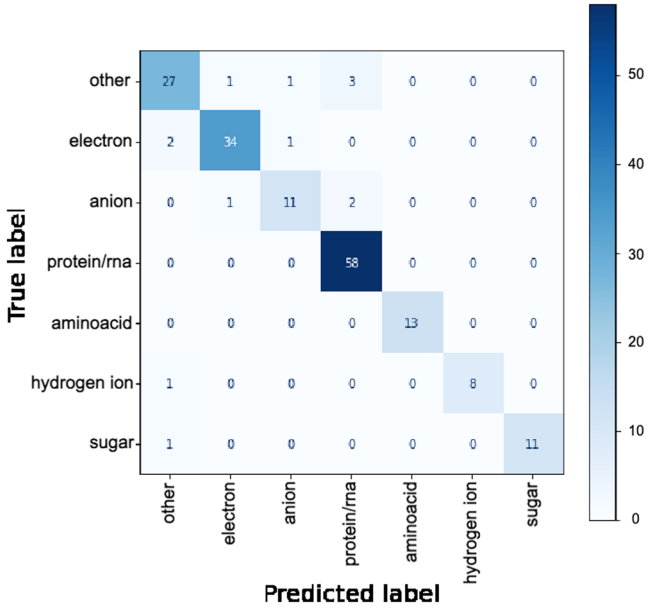

FIGURE 5 Confusion matrix of the Logistic Regression model for the different transporter categories. This matrix represent the results of the PortPred model trained on the Training set 3 and tested on Validation set 3.

PortPred, a prediction tool for transporter proteins, using DL approaches and protein sequence embeddings.

DL-based sequence embeddings have shown promising results in various bioinformatics tasks, including subcellular and sub-organelle classification, protein structure and function prediction, and PPIs. Inspired by these advancements, we explored the use of DL-based sequence embeddings for the accurate identification and classification of transporter proteins.

To develop PortPred, we reviewed and compared several DL-based protein embeddings, including UniRep, SeqVec, ProteinBERT, and ESM-1b. These embeddings capture the underlying features and representations of protein sequences and can be used to encode protein information for downstream prediction tasks. We evaluated the performance of each embedding and their combination in predicting transporter proteins and differentiating among various categories of transporters.

Our comprehensive analysis revealed that hybrid embeddings, which combine multiple embeddings with a feature selection procedure, generally outperformed single embeddings alone. The combination of embeddings provided a more informative representation of transporter proteins, leading to improved prediction performance. However, we observed that the ESM-1b embedding alone showed comparable or even higher performance than the hybrid embedding, demonstrating the effectiveness of this DL-based approach.

In particular, ESM-1b proved to be the most influential embedding even within the hybrid representation, while the others (UniRep, SeqVec, ProteinBERT) exhibited similar performance levels. Therefore, we emphasize the adaptability of the ESM-1b embedding for fine-tuned prediction tasks. Additionally, these results align with our prior studies on DL-based protein embeddings, as demonstrated in Anteghini et al.[22] and Anteghini et al.[26]

Our research initially focused on a binary prediction task, distinguishing transporter from non-transporter proteins. PortPred exceeded expectations with an accuracy of 94.53%. Encouraged by this success, we expanded our analysis to multiclass prediction, categorizing various transporter protein classes. PortPred maintained its excellence, achieving an average accuracy of 98.71%. These results underscore PortPred's robustness and reliability for transporter protein prediction, starting from a binary task.

Furthermore, we emphasize the importance of critically evaluating and validating DL methodologies in bioinformatics. It is essential to avoid the trend of simply improving the state-of-the-art without thoroughly assessing the reliability and interpretability of the results. Therefore, we performed consistent benchmark analyses to validate the performance of PortPred and ensure the reproducibility of our findings in a FAIR (Findable, Accessible, Interoperable, and Reusable) manner.

In practical applications, PortPred shows promising potential. When applied to a real-world case scenario involving peroxisomal proteins, PortPred achieved an accuracy of 82.62% and exhibited high specificity, making it a reliable tool for avoiding false positives. This highlights the practical utility of PortPred in identifying transporter proteins and can aid in the understanding of their biological functions and implications.

In conclusion, our study demonstrates the adaptability and effectiveness of DL-based sequence embeddings for transporter protein prediction. We encourage the scientific community to make informed choices when selecting and utilizing DL-based pre-trained representations, considering the specific requirements and characteristics of their prediction tasks. Furthermore, we advocate for rigorous validation, adaptation, and reporting of the limitations of these embeddings to ensure their reliability and usefulness in extracting meaningful biological insights.

AUTHOR CONTRIBUTIONS

Conceptualization: Marco Anteghini, Edoardo Saccenti and Vitor Martins dos Santos. *Methodology*: Marco Anteghini and Edoardo Saccenti; software, Marco Anteghini. *Validation*: Marco Anteghini, Edoardo

Saccenti; formal analysis, Marco Anteghini. *Investigation*: Marco Anteghini and Edoardo Saccenti; resources, Edoardo Saccenti and Vitor Martins dos Santos. *Data curation*: Marco Anteghini. *Writing—original draft preparation*: Marco Anteghini and Edoardo Saccenti. *Writing—review and editing*: Edoardo Saccenti and Vitor Martins dos Santos. *Visualization*: Marco Anteghini and Edoardo Saccenti. *Supervision*: Edoardo Saccenti and Vitor Martins dos Santos. *Project administration*: Edoardo Saccenti. *Funding acquisition*: Vitor Martins dos Santos. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

This project was developed in the context of the PerICo International Training Network and has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska Curie grant agreement No. 812968.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in PorPred at https://github.com/MarcoAnteghini/PortPred

ORCID

Marco Anteghini http://orcid.org/0000-0003-2794-3853

REFERENCES

1. Balch WE, Dunphy WG, Braell WA, Rothman JE. Reconstitution of the transport of protein between successive compartments of the Golgi measured by the coupled incorporation of N-acetylglucosamine. *Cell*. 1984; 39: 405-416.
2. Kaiser CA, Schekman R. Distinct sets of SEC genes govern transport vesicle formation and fusion early in the secretory pathway. *Cell*. 1990; 61: 723-733.
3. Hata Y, Slaughter CA, Sü dhof TC. Synaptic vesicle fusion complex contains unc-18 homologue bound to syntaxin. *Nature*. 1993; 366: 347-351.
4. Benga G. Water channel proteins: from their discovery in Cluj-Napoca, Romania in 1985, to the 2003 Nobel Prize in chemistry and their implications in molecular medicine. *Keio J Med*. 2006; 55: 64-69.
5. Hediger MA, Clémençon B, Burrier RE, Bruford EA. The ABCs of membrane transporters in health and disease (SLC series): introduction. *Mol Aspects Med*. 2013; 34: 95-107.
6. Sahoo S, Aurich M, Jonsson J, Thiele I. Membrane transporters in a human genome-scale metabolic knowledgebase and their implications for disease. *Front physiol*. 2014; 5: 91.
7. Robey RW, Pluchino KM, Hall MD, et al. Revisiting the role of ABC transporters in multidrug-resistant cancer. *Nat Rev Cancer*. 2018; 18: 452-464.
8. Dahl SG, Sylte I, Ravna AW. Structures and models of transporter proteins. *J Pharmacol Exp Ther*. 2004; 309: 853-860.
9. Saier MH. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev*. 2000; 64: 354-411.
10. Busch W, Saier MH. The transporter classification (TC) system, 2002. *Crit Rev Biochem Mol Bio*. 2002; 37: 287-337.
11. Saier Jr. MH, Reddy VS, Moreno-Hagelsieb G, et al. The transporter classification database (TCDB): 2021 update. *Nucleic Acids Res*. 2020; 49: D461-D467.
12. Berman HM. The protein data bank. *Nucleic Acids Res*. 2000; 28: 235-242.
13. Mishra NK, Chang J, Zhao PX. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PLoS ONE*. 2014; 9: e100278.
14. Liou YF, Vasylenko T, Yeh CL, et al. SCMMTP: identifying and characterizing membrane transport proteins using propensity scores of dipeptides. *BMC Genomics*. 2015; 16: S6.
15. Li L, Li J, Xiao W, et al. Prediction the substrate specificities of membrane transport proteins based on support vector machine and hybrid features. *IEEE/ACM Trans Comput Biol Bioinform*. 2016; 13: 947-953.
16. Nguyen TTD, Le NQK, Ho QT, et al. Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters. *Anal Biochem*. 2019; 577: 73-81.
17. Alballa M, Butler G. TooT-T: discrimination of transport proteins from non-transport proteins. *BMC Bioinformatics*. 2020; 21: 25.
18. Ghazikhani H, Butler G. TooT-BERT-T: A BERT approach on discriminating transport proteins from non-transport proteins. In Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022). Springer International Publishing; 2022: 1-11.
19. Alballa M, Aplop F, Butler G. TranCEP: predicting the substrate class of transmembrane transport proteins using compositional, evolutionary, and positional information. *PLOS ONE*. 2020; 15: e0227683.
20. Alley EC, Khimulya G, Biswas S, et al. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods*. 2019; 16: 1315-1322.
21. Heinzinger M, Elnaggar A, Wang Y, et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*. 2019; 20: 723.
22. Anteghini M, dos Santos VAM, Saccenti E. In-Pero: exploiting deep learning embeddings of protein sequences to predict the localisation of peroxisomal proteins. *Int. J. Mol*. 2021; 22: 6409.
23. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA*. 2021;118:e2016239118.
24. Nambiar A, Heflin M, Liu S, et al. Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '20. Association for Computing Machinery, New York, NY, USA. ISBN 9781450379649.
25. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: towards cracking the language of lifes code through

self-supervised deep learning and high performance computing. *IEEE Trans Pattern Anal Mach Intell*. 2021; 1.
26. Anteghini M, Haja A, dos Santos VAM, et al. OrganelX web server for sub-peroxisomal and sub-mitochondrial protein localization and peroxisomal target signal detection. *Comput Struct Biotechnol J*. 2022; 21: 128-133.
27. Vastermark A, Wollwage S, Houle ME, et al. Expansion of the APC superfamily of secondary carriers. *Proteins*. 2014; 82: 2797-2811.
28. Nigam SK, Bush KT, Martovetsky G, et al. The organic anion transporter (OAT) family: a systems biology perspective. *Physiol Rev*. 2015; 95: 83-123.
29. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2020;49:D480-D489.
30. Lyall F. Biochemistry. In *Basic Science in Obstetrics and Gynaecology*. Elsevier; 2010: 143-171.
31. Huang Y, Anderle P, Bussey KJ, et al. Membrane transporters and channels. *Cancer Res*. 2004; 64: 4294-4301.
32. Mueckler M, Caruso C, Baldwin SA, et al. Sequence and structure of a human glucose transporter. *Science*. 1985; 229: 941-945.
33. Ristovski M, Farhat D, Bancud SEM, Lee JY. Lipid transporters beam signals from cell membranes. *Membranes*. 2021; 11: 562.
34. Ma Z, Jacobsen FE, Giedroc DP. Coordination chemistry of bacterial metal transport and sensing. *Chem Rev*. 2009; 109: 4644-4681.
35. Agarwal S, Mishra NK, Singh H, Raghava GP. Identification of mannose interacting residues using local composition. *PloS one*. 2011; 6: e24039.
36. Chen SA, Ou YY, Lee TY, Gromiha MM. Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties. *Bioinformatics*. 2011; 27: 2062-2067.
37. Kawashima S, Pokarowski P, Pokarowska M, et al. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*. 2007; 36: D202-D205.
38. Attwood T. *Profile (Position-Specific Scoring Matrix, Position Weight Matrix, PSSM, Weight Matrix)*. American Cancer Society; 2004. ISBN 9780471650126.
39. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res*. 1982; 10: 2997-3011.
40. Altschul S, Madden T, Shaffer A, et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl Acids Res*. 1996; 25: 3389-3402.
41. Suzek BE, Wang Y, Huang H, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*. 2014; 31: 926-932.
42. Boughaci D, Benhamou B, Drias H. IGA: an improved genetic algorithm for MAX-SAT problems. In: Prasad B, ed. Proceedings of the 3rd Indian International Conference on Artificial Intelligence, Pune, India, December 17-19, 2007. IICAI; 2007: 132-150.
43. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*. 1997; 30: 1145-1159.
44. Li ZR, Lin HH, Han LY, et al. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res*. 2006; 34: W32-W37.
45. Guthrie D, Allison B, Liu W, et al. A closer look at skip-gram modelling. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association (ELRA), Genoa, Italy; 2006.
46. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space, 2013.
47. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. COLT '92. Association for Computing Machinery, New York, NY, USA. 1992: 144-152. ISBN 089791497X.
48. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol*. 1990; 215: 403-410.
49. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.
50. Cramer J. The origins of logistic regression. Tinbergen Institute, Tinbergen Institute Discussion Papers. 2002.
51. Tommaso PD, Moretti S, Xenarios I, et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res*. 2011; 39: W13-W17.
52. Chang JM, Tommaso PD, Notredame C. TCS: A new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol Biol Evol*. 2014; 31: 1625-1637.
53. Alballa M. Predicting Transporter Proteins and Their Substrate Specificity. Ph.D. thesis, Concordia University, 2020. Unpublished.
54. Alballa M, Butler G. TooT-SC: predicting eleven substrate classes of transmembrane transport proteins. *bioRxiv*. 2022.
55. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. In Proc. of NAACL. 2018.
56. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*. 1997; 9: 1735-1780.
57. Brandes N, Ofer D, Peleg Y, et al. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*. 2022;38:2102-2110.
58. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*. 2000; 25: 25-29.
59. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process*. 2017:5998-6008.
60. Harris ZS. Distributional structure. *Word*. 1954; 10: 146-162.
61. Alballa M, Butler G. Ontology-based transporter substrate annotation for benchmark datasets. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019: 2613-2619.
62. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006; 22: 1658-1659.
63. Li Y, Ilie L. SPRINT: ultrafast protein-protein interaction prediction of the entire human interactome. *BMC Bioinformatics*. 2017; 18: 485.
64. Cristianini N, Ricci E. *Support Vector Machines*. Springer US, 2008: 928-932. ISBN 978-0-387-30162-4.
65. Seliya N, Zadeh AA, Khoshgoftaar TM. A literature review on one-class classification and its potential applications in big data. *J Big Data*. 2021; 8: 122.

66. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. 1998; 20: 832-844.
67. Breiman L. Random forests. *Mach Learn*. 2001; 45: 5-32.
68. Murtagh F. Multilayer perceptrons for classification and regression. *Neurocomputing*. 1991; 2: 183-197.
69. Linnainmaa S. Taylor expansion of the accumulated rounding error. *BIT*. 1976; 16: 146-160.
70. Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*. 1980; 36: 193-202.
71. Tolles J, Meurer WJ. Logistic regression. *JAMA*. 2016; 316: 533.
72. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Series B Stat Methodol*. 1974; 36: 111-133.
73. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020; 585: 357-362.
74. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mac Learn Res*. 2011; 12: 2825-2830.
75. Rijsbergen CJV. *Information Retrieval*. 2nd ed. Butterworth-Heinemann; 1979.
76. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975; 405: 442-451.
77. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PloS one*. 2017; 12: e0177678.
78. Melo F. Area under the ROC curve. In *Encyclopedia of Systems Biology*. Springer 2013: 38-39.
79. Saccenti E, Hendriks MHWB, Smilde AK Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. *Sci Rep*. 2020; 10: 438.

How to cite this article: Anteghini M, Santos VAMd, Saccenti E. PortPred: exploiting deep learning embeddings of amino acid sequences for the identification of transporter proteins and their substrates. *J Cell Biochem*. 2023;124:1803-1824. doi:10.1002/jcb.30490