MDPI

Article

Exploiting Protein Language Models for the Precise Classification of Ion Channels and Ion Transporters

Hamed Ghazikhani [1] 0000-0001-9587-8016, Gregory Butler [2] 0000-0002-6938-0879

[1] Department of Computer Science and Software Engineering, Concordia University; h_ghazik@encs.concordia.ca
[2] Centre for Structural and Functional Genomics, Concordia University; gregory.butler@concordia.ca
* Correspondence: hamed.ghazikhani@concordia.ca

Abstract: This study presents TooT-PLM-ionCT, a holistic framework that exploits the capabilities of six diverse Protein Language Models (PLMs) - ProtBERT, ProtBERT-BFD, ESM-1b, ESM-2 (650M parameters), and ESM-2 (15B parameters) - for precise classification of integral membrane proteins, specifically ion channels (ICs) and ion transporters (ITs). As these proteins play a pivotal role in the regulation of ion movement across cellular membranes, they are integral to numerous biological processes and overall cellular vitality. To circumvent the costly and time-consuming nature of wet lab experiments, we harness the predictive prowess of PLMs, drawing parallels with techniques in natural language processing. Our strategy engages six classifiers, embracing both conventional methodologies and a deep learning model, to segregate ICs and ITs from other membrane proteins, as well as differentiate ICs from ITs. Furthermore, we delve into critical factors influencing our tasks, including the implications of dataset balancing, the effect of frozen versus fine-tuned PLM representations, and the potential variance between half and full precision floating-point computations. Our empirical results showcase superior performance in distinguishing ITs from other membrane proteins and differentiating ICs from ITs, while the task of discriminating ICs from other membrane proteins exhibits results commensurate with the current state-of-the-art.

Keywords: Ion channels; Ion transporters; Membrane proteins; Drug discovery; Protein language models; Deep learning

Citation: Ghazikhani, H.; Butler, G. Exploiting Protein Language Models for the Precise Classification of Ion Channels and Ion Transporters. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

1.1. Background

Protein language models (PLMs) are a transformative development in the field of bioinformatics, leveraging the power of machine learning to predict protein structures and functions from their amino acid sequences [1–3]. These models, inspired by natural language processing (NLP) techniques [4–7], treat proteins as "sentences" composed of "words" (amino acids), enabling the prediction of protein properties based on sequence information alone [8]. The importance of PLMs lies in their potential to revolutionize our understanding of proteins, the building blocks of life, and to accelerate drug discovery and design processes [9]. They provide a powerful tool for predicting protein structures, which is crucial for understanding diseases and developing treatments [10]. Moreover, PLMs produce comprehensive representations of protein sequences that are useful for various applications in protein analysis, including predicting protein function, protein-protein interactions, and protein structure [1,3,11–17]. Unsal et al. [8] review the use of natural language models for protein representation from 2015 to the present.

The regulation of ion movement across cell membranes is a critical aspect of cellular function, with ion channels (ICs) and ion transporters (ITs) playing key roles [18]. These membrane proteins (MPs) are involved in maintaining ion homeostasis (the regulation and maintenance of a stable and balanced concentration of ions), regulating transmembrane potential, and facilitating electrical signaling, which are essential for various cellular processes such as proliferation, migration, apoptosis, and differentiation [19–21].

ITs, also known as ion pumps, actively transport ions against their concentration gradient, a process that requires potential energy [22]. On the other hand, ICs are transmembrane protein complexes located in the lipid bilayer membrane of all cells [23]. They facilitate the passive movement of ions across cell membranes, thereby helping cells maintain electrical properties and regulate functions [19,20].

Given their crucial role in cellular function, ICs have become a significant focus in membrane protein research and drug discovery [23]. They serve as promising therapeutic targets for various diseases, including neurological disorders, cardiovascular diseases, and cancer [24–27].

In an effort to expedite the drug discovery process and circumvent the high costs and time-consuming nature of wet lab experiments, computational methods have been developed. These innovative techniques efficiently predict the presence and function of ion channels, thereby accelerating the identification of potential drug targets [28,29].

Among these computational methods, PLMs have emerged as a particularly powerful tool [3]. By learning the sequence patterns of different protein families, PLMs can accurately classify proteins and predict their functions [17,23,30]. This capability not only streamlines the process of protein classification but also opens up new avenues for the discovery of therapeutic targets [31].

1.2. Review of Previous Work

There has been a significant amount of research on predicting ICs and ITs in the past, with an emphasis on developing computational methods that can accurately differentiate these proteins from other MPs [18,22,28,29,32–34]. These methods have often utilized traditional machine learning techniques, such as Support Vector Machines (SVM) and Random Forests (RF), which classify protein sequences based on features derived from their primary, secondary, and tertiary structures. These features can include information about the sequence itself, such as the presence of certain amino acid residues or motifs, as well as structural features, such as secondary structure elements or solvent accessibility [29,35]. The use of these features for ion channel prediction is thoroughly explained in Menke et al. [29] and Ashrafuzzaman [28].

The advent of deep learning has paved the way for novel opportunities in predicting ICs and ITs. Recent studies underscore the potential of these advanced techniques to generate intricate representations of protein sequences, thereby enhancing the efficiency of IC and IT prediction models [18,22]. In their respective methodologies, Taju and Ou [18], as well as Nguyen et al. [22], utilized position-specific scoring matrices (PSSM) for encoding proteins into feature vectors, while leveraging Convolutional Neural Networks (CNNs) for classifying ICs and ITs from other membrane proteins (MPs). These innovative models could discern complex patterns in protein sequences, employing this information to augment prediction performance, potentially surpassing the constraints of conventional machine learning approaches [18,22]. However, it is noteworthy that their work primarily focuses on distinguishing ion channels from other membrane proteins and ion transporters from other membrane proteins, rather than the task of differentiating ion channels from ion transporters.

Ghazikhani et al. pioneered the introduction of TooT-BERT-T [30] and TooT-BERT-C [23], sophisticated methods designed for distinguishing transmembrane transport proteins from non-transport proteins, as well as differentiating ICs from non-ICs. These methods incorporate a Logistic Regression (LR) classifier with fine-tuned representations derived from a PLM known as ProtBERT-BFD [3]. As the most advanced predictors for transporters and ICs, these approaches underscore the promising potential of employing protein language models for such tasks.

1.3. Research Overview and Objective

In this study, we conduct a comprehensive analysis of six PLMs with six different classifiers to differentiate ion channels, ion transporters, and other membrane proteins.

In pursuit of a deeper understanding of PLM performance in protein classification tasks, we scrutinize essential variables such as dataset balancing, representation tuning, and the precision of floating-point calculations.

The overarching goal of this paper is to present a pioneering, automated method for the precise categorization of ion transporters and ion channels within the expansive array of membrane proteins. By elucidating the complex nature of these vital biological components, we seek to facilitate their identification in bioinformatics research and potentially expedite the discovery of novel therapeutic targets for a variety of diseases.

1.4. Study of Impacts

In this study, we embarked on a meticulous investigation of three pivotal factors that could significantly influence the performance of PLMs in our tasks. These encompass:

- The choice between using frozen or fine-tuned PLM representations.
- The influence of balanced versus imbalanced datasets on model performance.
- The implications of half-precision versus full-precision floating-point computations.

Each of these elements represents a vital facet of the model's configuration and data management, thus underscoring the importance of their potential impacts on model performance. The forthcoming sections deliver a succinct synopsis of each factor, explicating the fundamental concept and the rationale for its incorporation in our study.

1.4.1. Frozen vs. Fine-tuned Representations

The concept of frozen and fine-tuned representations pertains to the degree of adaptation of pre-trained language models to a specific task. Frozen representations refer to the utilization of pre-trained models in their original state, without any further task-specific training. On the other hand, fine-tuned representations involve the additional step of task-specific training, where the pre-existing parameters of the pre-trained models are adjusted to enhance their performance on the given task.

The comparative study of frozen and fine-tuned versions of a PLM offers valuable insights into the performance dynamics of these models. It allows us to understand the inherent behavior of the original pre-trained models (as reflected in the frozen state) and to quantify the extent of improvement achievable through task-specific fine-tuning. This comparison can potentially expose the limitations of the pre-training process and highlight the areas where fine-tuning can yield significant benefits.

It is important to note that fine-tuning necessitates additional computational resources compared to the use of frozen models. Consequently, if the performance enhancement achieved through fine-tuning is marginal or negligible for a specific task, it might be more resource-efficient to employ the model in its frozen state. This aspect underscores the importance of our investigation into the relative merits of frozen and fine-tuned representations in the context of our tasks.

1.4.2. Balanced vs. Imbalanced Datasets

The terms "balanced" and "imbalanced" in machine learning refer to the distribution of classes within a dataset. A balanced dataset exhibits approximately equal representation of all classes, while an imbalanced dataset is characterized by unequal representation of classes. In the context of this study, these terms are used to describe the distribution of membrane protein sequences in the DS-C dataset (Table 2).

Imbalanced datasets, where certain classes are underrepresented, can significantly impact the performance of a machine learning model. The model may develop a bias towards the majority class, leading to suboptimal performance when predicting the minority class. In the realm of PLMs, this issue translates into a potential struggle for the model to accurately predict protein types that are underrepresented in the training data.

Furthermore, the bias introduced by an imbalanced dataset can result in a model that performs better for the class with greater representation in the data. For instance, if the dataset contains a significantly larger number of MPs compared to ICs or ITs, the

model may develop a bias towards MPs. This bias could compromise the model's ability to accurately predict ICs or ITs, underscoring the importance of considering the balance of classes in the dataset.

1.4.3. Half vs. Full Precision Floating Points Calculations

Half and full precision floating-point representations pertain to the level of numerical precision employed in model computations. Full precision, typically realized through 32-bit floats, provides superior numerical accuracy. Conversely, half precision, utilizing 16-bit floats, curtails memory usage and computational demands, albeit at the expense of a slight reduction in numerical accuracy.

The use of half-precision computations can expedite the training process, but it may also influence model performance due to the diminished numerical precision. It is crucial to evaluate whether this reduction in precision significantly affects the model's capacity to learn and generalize effectively.

Additionally, investigating the impact of half versus full precision provides valuable insights into the balance between computational efficiency and model performance. This understanding facilitates informed decision-making, taking into account the available computational resources and the precision requirements of the task at hand.

1.5. Paper Structure

This paper is organized as follows: Section 2 details our methodologies, including the datasets used and the process for balancing the membrane proteins dataset. It provides a brief overview of the employed PLMs and classifiers, elaborates on hyperparameter optimization, and discusses the evaluation metrics used to assess model performance. In Section 3, we present and dissect the results of our experimental analyses. This section evaluates the performance of different PLMs and classifiers for each task, sheds light on the impact of the three previously mentioned factors, and includes visualizations of protein representations. Additionally, it juxtaposes our findings with current state-of-the-art methodologies for each task. Finally, Section 4 encapsulates our contributions and the insights gleaned from our study. It also outlines potential future research avenues, emphasizing areas where additional exploration could enrich the understanding of protein classification using PLMs.

2. Materials and Methods

2.1. Methodology Overview

We have undertaken a comprehensive evaluation of representations derived from six distinct PLMs. These include ProtBERT, ProtBERT-BFD, and ProtT5 from ProtTrans project [3], as well as ESM-1b, ESM-2, and ESM-2_15B from ESM project [2,36].

To further our analysis, we have employed six classifiers with the aim of distinguishing ICs from other MPs, differentiating ITs from other MPs, and discriminating ICs from ITs. These classifiers encompass traditional methodologies such as LR, k-Nearest Neighbor (kNN), support vector machine (SVM), random forest (RF), and feed-forward neural network (FFNN). Additionally, we have incorporated a convolutional neural network (CNN), a deep learning model, for comparative analysis.

Our study also delves into the examination of several critical factors that could potentially influence the outcomes of our tasks. These include the impact of balancing the MP dataset on the results, the influence of frozen and fine-tuned representations from PLMs, and the potential differences between half and full precision floating-point calculations. By investigating these factors, we aim to provide a more nuanced understanding of the performance and applicability of PLMs in protein classification tasks. Refer to Table 1 for a comprehensive summary of the research methodology employed in this study.

Table 1. Comprehensive Overview of Research Methodology. This table encapsulates the various components of the research methodology employed in this study, providing a concise summary and brief description of each element.

Methodology Component	Details
Protein Language Models	ProtBERT, ProtBERT-BFD, ProtT5 (ProtTrans project), ESM-1b, ESM-2, ESM-2_15B (ESM project)
Tasks	Discrimination of ion channels vs other membrane proteins, ion transporters vs other membrane proteins, ion channels vs ion transporters
Classifiers	SVM, Logistic Regression (LR), Random Forest (RF), kNN, Feed-forward Neural Network (FFNN), CNN
Hyperparameter Optimization	Grid search using scikit-learn (for SVM, LR, RF, kNN, FFNN) and Optuna (for CNN)
Cross-Validation Technique	5-fold cross-validation
Evaluation Metrics	Accuracy, MCC, Sensitivity, Specificity
Statistical Significance Analysis	Paired Student t-test, ANOVA
Impacts Evaluated	1) Frozen vs. fine-tuned representations from PLMs, 2) Balanced vs. imbalanced datasets (Downsampling of MPs dataset), 3) Half vs. full precision floating point calculations
Presentation of Results	Tables and figures, grouped results by various aspects such as dataset balance, classifier type, PLM, representation type (frozen or fine-tuned), precision type (half or full), and UMAP projection figures for each PLM, task, and representation type
Optimal Configuration	Selected the best configuration for each task, ran on independent test set, and compared results with state-of-the-art
Limitations	Could not fine-tune large PLMs like ProtT5 and ESM-2_15B due to resource constraints (GPUs, memory), could not extract full precision floating point from these PLMs. This led to missing values in tables and figures.

2.2. Dataset

In our study, we employ the same dataset used in the DeepIon [18] and MFPS_CNN [22] projects, which was gathered from the UniProt database [37]. To ensure a diverse and representative collection, Taju and Ou [18] applied the BLAST algorithm [38] to remove protein sequences with more than 20% similarity. The resulting dataset comprises 4915 protein sequences, including 301 ion channels, 351 ion transporters, and 4263 (other) membrane proteins. The dataset was split into training and test sets for assessing model generalizability. The distribution of sequences in the dataset is presented in Table 2.

Table 2. DS-C, the ion channel and ion transporter dataset. This table displays the distribution of sequences in the dataset used in this study, separated into the training and test sets.

Class	Training	Test	Total
Ion channel (IC)	241	60	301
Ion transporter (IT)	281	70	351
Other membrane protein (MP)	3,413	850	4,263
Total	3,935	980	4,915

2.2.1. Balancing the Membrane Protein Dataset

As highlighted in Table 2, there exists a significant disparity in the number of membrane protein sequences in comparison to ion channel or ion transporter protein sequences.

For this study, our objective was to assess the performance of PLMs and classifiers employing both imbalanced and balanced datasets. To construct a balanced dataset (Figure 1), we implemented a random selection process to draw 280 sequences from the membrane protein training set. To enhance the accuracy of the results and mitigate potential variability, this process was reiterated ten times, each iteration using a distinct random state.

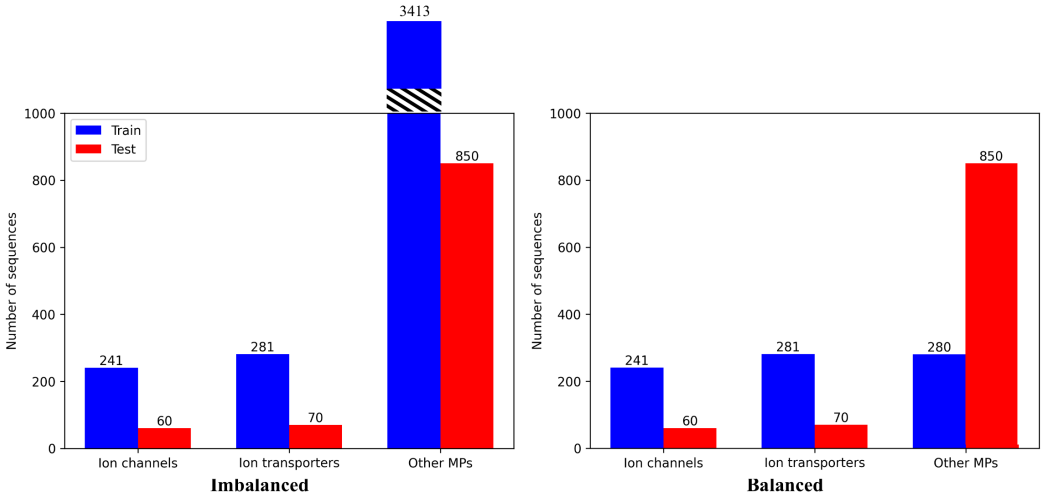

Figure 1. Visualization of Membrane Protein Dataset Balancing: This figure presents the distribution of sequences in each dataset, delineated as bar plots. The training set sequences are represented by the blue bars, whereas the red bars depict the sequences in the independent test set. The left-hand figure portrays the distribution within the imbalanced dataset of additional membrane proteins (MPs). Conversely, the right-hand figure exhibits the balanced dataset, which was achieved through undersampling of MPs in the training set.

2.3. Protein Language Models (PLMs)

This study leverages six distinct Protein Language Models (PLMs) for comparative analysis (Table 3): (1) *ProtBERT* [3] is an encoder-only model inspired by BERT [39], pre-trained on UniRef100 [40]. (2) *ProtBERT-BFD* [3], analogous to ProtBERT, is pre-trained on the BFD database [41] instead of UniRef100. (3) *ProtT5-XL* [3] (simplified to *ProtT5* for convenience), is an encoder-decoder model rooted in the T5 architecture [6]. It is initially trained on BFD and subsequently fine-tuned on Uniref50 [40]. (4) *ESM-1b* [2] is a Transformer model pre-trained on UniRef50. (5) *ESM-2* [36], while akin to ESM-1b, benefits from enhanced architecture, improved training parameters, and augmented computational resources and data. (6) *ESM-2_15B* [36], the largest PLM to date, is a more extensive version of ESM-2, incorporating 15 billion parameters.

Table 3. Implementation details for ProtBERT [3], ProtBERT-BFD [3], ProtT5 [3], ESM-1b [2], ESM-2 [36], ESM-2_15B [36].

	ProtBERT	ProtBERT-BFD	ProtT5	ESM-1b	ESM-2	ESM-2_15
Parameters	420M	420M	3B	650M	650M	15B
Dataset	UniRef100	BFD	BFD	UniRef50	UniRef50	UniRef50
Sequences	216M	2.1B	2.1B	27M	27M	27M
Embedding dim	1024	1024	1024	1280	1280	5120
Layers	30	30	24	33	33	48

To derive frozen representations, we harness feature vectors from the final layer of the PLMs, employing mean-pooling to generate a unique representation for each protein sequence. This process is consistent with the methodologies adopted in ProtTrans [3] and ESM [2,36].

For fine-tuning of the PLMs, we engage the Trainer API from the transformers library [42]. We primarily utilize the library's default hyperparameters but modify the number of epochs to 5, following the guidelines of the original BERT paper [39]. To mitigate memory constraints, we adopt a batch size of 1.

2.4. Classifiers

For our machine learning classifiers, we implement Support Vector Machine (SVM) [43], k-Nearest Neighbors (kNN) [44], Random Forest (RF) [45], Feed-Forward Neural Network (FFNN) [46], and Logistic Regression (LR) [47] using the scikit-learn library [48], whereas Convolutional Neural Network (CNN) [49] using PyTorch [50]. These classifiers are designed to provide a comprehensive comparison of various machine learning approaches in combination with the PLMs.

2.5. Hyperparameter Optimization

In this investigation, we incorporated an all-encompassing strategy for hyperparameter optimization, harnessing the prowess of scikit-learn grid search [48] and Optuna [51], an advanced Python library specifically designed for hyperparameter optimization. The primary objective was to discern the quintessential set of hyperparameters for each model to maximize the efficacy of our classification algorithms.

With respect to conventional classifiers such as SVM, RF, kNN, LR, and FFNN, we exploited grid search—an exhaustive technique that systematically scrutinizes a pre-defined subset of hyperparameters. This process was executed utilizing the scikit-learn library [48].

Each classifier was assigned a unique set of hyperparameters to investigate. The specific grids of hyperparameters tailored for each classifier were as follows:

- SVM: The investigation included cost parameters (C) of 0.1, 1, 10, and 100; kernel coefficients (gamma) of 0.1, 1, and 10; and kernel types (kernel) inclusive of linear, rbf, and sigmoid.
- RF: The search encompassed the number of trees in the forest (number of estimators) of 50, 100, and 200; the maximum tree depth (maximum depth) of 5, 10, and None; and the minimum samples required to split an internal node (minimum samples split) of 2, 5, and 10.
- kNN: The evaluation incorporated the number of considered neighbors (number of neighbors) of 3, 5, 7, and 9; the prediction weight function (weights) of uniform and distance; and the algorithm used for calculating the nearest neighbors (algorithm) of ball_tree, kd_tree, and brute.
- LR: The investigation comprised various penalty types (penalty) of l1 and l2; cost parameters (C) of 0.1, 1, 10, and 100; and optimization solvers (solver) of liblinear and saga.
- Feed-Forward Neural Network (FFNN): The search included the number of neurons in the hidden layer (hidden_layer_sizes) of (512, 256, 64), (512,), and (256,); the activation function for the hidden layer (activation) of relu and tanh; and the weight optimization solver (solver) of adam and sgd.

For the evaluation of model performance for each hyperparameter combination, we employed stratified 5-fold cross-validation. The optimization scoring metric was the Matthews Correlation Coefficient (MCC).

In the case of our Convolutional Neural Network (CNN) model, we utilized Optuna [51], a Python library adept at hyperparameter optimization. Optuna leverages a variety of optimization algorithms to traverse the hyperparameter space with the goal of identifying the optimal values that enhance the model's performance.

The optimization procedure was encapsulated in an objective function, which incorporated the hyperparameters to be optimized. The specific hyperparameters and their respective ranges or sets of values were as follows:

- Kernel Sizes: The possibilities included combinations of [3, 5, 7], [3, 7, 9], [5, 7], and [7, 7, 7].
- Output Channels: The combinations were [128, 64, 32].
- Dropout Probability: The range was set from 0.2 to 0.5.
- Optimizer: The options included Adam, RMSprop, and SGD.
- Learning Rate: The range extended from 1e-6 to 1e-2 on a logarithmic scale.

The model underwent training for 10 epochs, with the performance being assessed on the validation set using MCC as the performance metric. The pruning feature of Optuna was harnessed to curtail trials early if they lacked promise, thereby conserving computational resources.

Owing to the intensive computational requirements of this procedure in terms of time and memory, the optimization was carried out singularly for each task and dataset, thereby resulting in five distinct hyperparameter settings (IC-MP balanced, IC-MP imbalanced, IT-MP balanced, IT-MP imbalanced, and IC-IT). For balanced datasets, one dataset was randomly selected from a pool of 10 for consideration.

The optimization procedure was executed for 100 trials, with each trial embodying a complete execution of the objective function with a distinct set of hyperparameters. The Optuna study was configured to maximize the MCC, and the optimization procedure was expedited by using a GPU for increased efficiency.

2.6. Cross-Validation and Evaluation Metrics

The technique of k-fold cross-validation, ubiquitously utilized in model evaluation, necessitates partitioning the original dataset into k subsets or folds of equivalent size. During each iteration, a single fold is reserved for validation, while the remaining k-1 folds serve as the training set. This cycle is repeated k times, ensuring each fold is used precisely once as the validation set. The model's performance is then evaluated as the mean over the k iterations, delivering a more robust and accurate assessment of its capability to generalize. k-fold cross-validation plays a pivotal role in mitigating overfitting risk and curtailing bias in model evaluation. Our experimentation was conducted using 5-fold cross-validation, signifying the partitioning of the dataset into five subsets and repeated model training and validation over five iterations, with each fold serving as the validation set once.

In the context of this paper, we utilized four performance metrics to assess the efficacy of our approach for the tripartite tasks of IC-MP, IT-MP, and IC-IT. These metrics encompassed MCC, Accuracy, Sensitivity, and Specificity.

Accuracy represents an overall measure of correct classification rate, computed as the fraction of correct predictions relative to the total number of predictions. It is expressed as a percentage and can be determined using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sensitivity, also referred to as the true positive rate, measures the proportion of actual positive instances that are correctly identified. Its calculation is as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

Specificity, alternatively known as the true negative rate, quantifies the proportion of actual negative instances that are correctly identified. Its calculation is as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

MCC is esteemed as a reliable and stable evaluation metric when handling imbalanced data [52]. The MCC values span from -1 to 1, where 1 signifies perfect prediction, 0 denotes performance equivalent to random chance, and -1 represents a total misalignment between predictions and observations. A high MCC value suggests a predictor demonstrating high

accuracy for both positive and negative classes while maintaining a low misprediction rate for each class. In our research, we accord greater emphasis to the MCC metric due to its comprehensive nature and reliability.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

Here, TP (True Positive) denotes an instance where the classifier accurately predicts the positive class, TN (True Negative) signifies an instance where the classifier accurately predicts the negative class, FP (False Positive) represents an instance where the classifier erroneously predicts the positive class, and FN (False Negative) refers to an instance where the classifier inaccurately predicts the negative class.

2.7. Statistical Significance Analysis

The statistical significance of observed differences was tested using two methods: the paired Student's t-test [53] and Analysis of Variance (ANOVA) [54]. The paired Student's t-test, ideal for comparing means of two related groups, was employed for two sets of related observations. Conversely, ANOVA, which assesses means across three or more unrelated groups, was applied when more than two independent groups were to be compared. The outcomes were expressed as a p-value, a statistical measure estimating the probability of random chance producing the observed results. Conventionally, a p-value below 0.05 signifies statistical significance, indicating a minimal probability that the observed difference occurred due to random chance. In our analysis, the p-value was computed from MCC metric, which is deemed comprehensive and reliable.

2.8. Limitation

Our study was not without its limitations, primarily due to the constraints imposed by the available computational resources. The fine-tuning of large-scale PLMs such as ProtT5 (with 3 billions of parameters) and ESM-2_15B (with 15 billions of parameters) necessitates substantial computational resources and significant GPU memory, particularly for the extraction of full-precision floating-point representations [55**?**]. Given our limited resources, which included a single GPU V100, we were unable to perform these tasks, resulting in some missing results in Section 3 in our tables and figures.

Furthermore, the absence of results in Table 10 (ion channels vs. ion transporters) is attributed to the fact that the corresponding studies [18,22] do not report these specific results, and there are no readily available tools that can generate them. The primary focus of these papers is to classify ion channels and ion transporters against other membrane proteins, rather than against each other. However, in light of the data available to us, we chose to conduct this experiment and compare our models in this context as well.

3. Results and Discussion

This section presents a comprehensive exploration of the findings derived from our research, articulated through a combination of tables and figures to demonstrate and contrast varying facets of the study. We elucidate the performance of six distinct Protein Language Models (PLMs) as they engage with three specific tasks: differentiating ion channels (IC) from membrane proteins (MP), distinguishing ion transporters (IT) from MPs, and discerning IC from IT. We delve into the performance of six classifiers within these tasks, shedding light on three pivotal factors under investigation: the influence of frozen versus fine-tuned representations, the effect of balanced versus imbalanced datasets, and the impact of half versus full precision floating-point calculations.

Our findings are quantified using four evaluative metrics: Matthews Correlation Coefficient (MCC), Accuracy, Sensitivity, and Specificity. We present these results as mean ± standard deviation, obtained from a 5-fold cross-validation (CV). In our attempt to provide an overarching view, we compute averages over tasks, PLMs and classifiers, yielding a high-level depiction of our results. It should be noted, however, that results compared

against the state-of-the-art are derived from an independent test set, employed solely for this purpose, with all other evaluations conducted on the training set.

In our tables, the highest values for each column and category are highlighted in bold, facilitating immediate recognition. Where there are more than two comparable values, the second highest are underlined to illustrate the proximity between the best and second-best results. In the corresponding figures, we prioritize the MCC metric, owing to its reliability and comprehensive nature. Each bar in these figures denotes the mean MCC, with the error bar atop indicating the standard deviation from the 5-fold CV. A Δ symbol highlights the difference between pairs of bars.

To ascertain the statistical significance of our findings, we employ ANOVA [54], a method for comparing the means of three or more groups, and the paired t-test [53], used to compare the means of two related groups. A p-value of 0.05 or smaller indicates a significant difference. It is important to note that this section primarily discusses general findings; more detailed results can be found in the appendix of this paper.

3.1. Performance of PLMs for Classification Tasks

Table 4 presents a detailed evaluation of the six PLMs engaged in three distinct classification tasks: differentiating IC from MP, distinguishing IT from MP, and discerning IC from IT.

3.1.1. Performance of PLMs

Our findings underscore the superior performance of the ESM-1b PLM, as it eclipses other PLMs across all evaluation metrics and tasks. The lone exception is observed in the task of distinguishing IC from IT, where ESM-1b shares the lead position with ESM-2_15B. This indicates that ESM-1b consistently delivers high accuracy in predicting ICs and ITs from MPs.

However, the second-best performing model varies according to the task. ESM-2 exhibits commendable performance for differentiating IC from MP and distinguishing IT from MP, whereas ProtT5 excels in the IC-IT classification task. The significant variations in p-values across all PLMs further accentuate the formidable performance of ESM-1b.

In tasks pertaining to the differentiation of IC from MP and the distinction between IC and IT, the performance variance between the highest-ranking and the runner-up PLMs is minimally noticeable across all evaluation metrics. However, when tasked with discerning IT from MP, a notable performance discrepancy becomes apparent, particularly evident in the Matthews correlation coefficient (MCC) metric. This highlights a more substantial divergence in the proficiency of the two leading models, specifically ESM-2 and its predecessor, ESM-1b, within this particular task.

Factors Contributing to ESM-1b's Superior Performance Outcomes

Our study posits that the unique architectural design of ESM-1b substantially contributes to its superior performance. This hypothesis is supported by our observation that identical pretraining dataset sizes, as employed in ESM-1b, ESM-2, ESM-2_15B, and more data in ProtBERT, ProtBERT-BFD, and ProtT5, in conjunction with model dimensions varying from 650M (for ESM-1b) to 15B (for ESM-2_15B) parameters, does not affect the performance of the corresponding PLMs significantly. We attribute the performance differences primarily to two factors: positional encoding and dropout strategies.

Positional Encoding and Its Impact

ESM-1b [2] exhibits a unique approach to positional encoding. Diverging from the original Transformer architecture [4], it replaces the conventional static sinusoidal encoding with a learned encoding approach. This is markedly different from the approaches observed in the ESM-2 [36] and ProtTrans PLM [3] families.

Dropout Strategies and Their Influence

Dropout [56], a prominent regularization technique in deep learning, randomly disables certain neural network units during training. This strategy enforces the network to develop more robust and generalizable features by reducing overfitting.

Distinct dropout strategies underscore a significant differentiation between ESM-1b and other PLMs. For instance, ESM-2 chooses to completely forgo dropout within hidden layers and attention. This pattern is also discernible in ProtBERT and ProtBERT-BFD, where dropout appears to be absent from their architectures. Conversely, ESM-1b not only incorporates dropout in its architectural framework but also applies it across various tasks. Considering the potential benefits of overfitting prevention measures, especially pertinent to the tasks investigated in our study, this difference assumes substantial significance.

Thus, in light of these findings, we suggest that the distinctive architectural design of ESM-1b plays a crucial role in facilitating its superior performance outcomes.

Table 4. Performance overview of protein language models for protein classification tasks. This figure provides a comprehensive performance evaluation of various protein language models (PLMs), organized in the order of their parameter count, across three distinctive protein classification tasks: differentiating ion channels (IC) from membrane proteins (MP), distinguishing ion transporters (IT) from MPs, and discerning IC from IT. The evaluation metrics, captured through a 5-fold cross-validation approach, are presented as mean±standard deviation. The p-value accompanying each result measures the statistical significance of observed differences among the PLMs. The highest value achieved for each task and column is highlighted in bold, whereas the second highest value is underlined to allow for comparative analysis between top-performing models.

Task	PLM	MCC	Accuracy	Sensitivity	Specificity	P-value
IC-MP	ProtBERT	0.73±0.05	90.99±1.76	76.88±4.89	91.69±2.83	1.25e-06
	ProtBERT-BFD	0.74±0.05	91.46±1.63	76.18±4.82	92.27±2.60	
	ESM-1b	**0.84±0.03**	**94.15±1.17**	**88.44±3.39**	**94.33±1.91**	
	ESM-2	0.83±0.04	93.89±1.27	85.66±4.43	**94.39±1.94**	
	ProtT5	0.79±0.05	93.12±1.38	79.68±4.98	**94.35±1.81**	
	ESM-2_15B	0.78±0.04	93.16±1.23	81.52±4.38	93.13±1.71	
IT-MP	ProtBERT	0.71±0.05	90.75±1.41	75.66±4.69	91.58±2.34	2.49e-03
	ProtBERT-BFD	0.74±0.05	91.10±1.64	78.91±4.79	92.30±2.33	
	ESM-1b	**0.82±0.04**	**93.47±1.31**	**85.09±3.46**	**94.53±2.09**	
	ESM-2	0.78±0.04	92.64±1.36	82.06±4.20	93.41±2.26	
	ProtT5	0.75±0.04	92.78±1.13	77.55±4.42	93.58±1.94	
	ESM-2_15B	0.72±0.04	91.58±1.46	76.12±4.26	91.90±2.32	
IC-IT	ProtBERT	0.79±0.03	89.33±1.67	88.92±4.38	89.62±4.46	2.14e-06
	ProtBERT-BFD	0.78±0.05	88.71±2.46	88.29±5.12	89.29±4.67	
	ESM-1b	**0.85±0.04**	**92.46±2.25**	**92.83±3.42**	**92.12±4.21**	
	ESM-2	0.83±0.04	91.42±2.17	91.21±3.62	91.83±4.21	
	ProtT5	0.84±0.04	91.83±1.83	91.00±2.67	**92.50±3.83**	
	ESM-2_15B	**0.85±0.03**	**92.33±1.67**	91.50±2.67	**92.83±3.83**	

3.1.2. Impact of Dataset Balance and Fine-Tuning

This study observes that larger models, namely ProtT5 and ESM-2_15B, despite being precluded from fine-tuning due to resource constraints, managed to equal the performance of the smaller model, ESM-1b, on the balanced IC-IT dataset. Intriguingly, even with the application of fine-tuning to ESM-1b, the frozen representations demonstrated their efficacy when the dataset is balanced, as evidenced in the IC-IT case.

This finding is substantiated by Table A5 and Figure A5, which depict superior performance with frozen representation on the balanced dataset. However, the difference was

not statistically significant (with a p-value > 0.05) across most of the PLMs, rendering this observation as noteworthy, though not decisive.

The observed phenomenon intriguingly suggests a potential connection between dataset balance and the concepts of frozen and fine-tuned representations. Rather than treating these concepts as mutually exclusive, our study proposes that different tasks may warrant exploration of varying combinations of these methodologies, indicating the necessity for a more nuanced approach in their application.

3.1.3. Size of PLMs and Performance

Our findings challenge the prevailing notion that the performance of PLMs invariably scales in direct proportion to their size. Interestingly, we did not identify a clear linear correlation between the dimensionality of a PLM and its ensuing performance. As a case in point, ESM-1b, with its 650 million parameters, consistently outperformed ESM-2_15B, which boasts 15 billion parameters, even when dealing with frozen representations (refer to Table A1. This observation underscores the conclusion that the performance efficacy of a PLM does not hinge exclusively on its size. Instead, it is shaped by a more intricate interplay of factors, with architectural design playing a significant role.

3.2. Comparative Performance Analysis of Classifiers

Table 5 presents performance results grouped by various classifiers utilized for three distinct protein classification tasks: distinguishing IC from MP, differentiating IT from MP, and discerning IC from IT.

Table 5. Performance overview of classifiers across protein classification tasks. This table offers a comprehensive performance evaluation of each classifier across three distinct protein classification tasks: differentiating ion channels (IC) from membrane proteins (MP), distinguishing ion transporters (IT) from MPs, and discerning IC from IT. The results, captured via a 5-fold cross-validation approach, are represented as mean±standard deviation. An accompanying p-value quantifies the statistical significance of observed differences among the classifiers. The highest value achieved for each task and column is marked in bold, while the second highest value is underlined to facilitate a comparison between the top-performing models.

Task	Classifier	MCC	Accuracy	Sensitivity	Specificity	P-value
IC-MP	LR	0.82±0.04	93.99±1.30	85.53±4.03	94.69±1.97	2.29e-14
	kNN	0.68±0.05	87.52±1.71	82.96±4.62	82.13±2.68	
	SVM	**0.84±0.04**	**94.51±1.13**	85.76±3.69	95.66±1.71	
	RF	0.69±0.05	92.00±1.38	63.96±4.59	**96.86±1.52**	
	FFNN	0.83±0.04	**94.10±1.19**	**86.66±3.93**	94.66±1.82	
	CNN	0.83±0.05	93.96±1.93	85.07±5.63	95.40±3.84	
IT-MP	LR	0.80±0.04	**93.12±1.34**	83.71±3.74	94.19±2.21	4.77e-11
	kNN	0.69±0.05	88.54±1.76	80.58±4.21	85.93±2.56	
	SVM	**0.81±0.04**	**93.17±1.21**	**84.28±4.47**	94.62±1.96	
	RF	0.65±0.05	90.33±1.62	64.35±4.47	93.57±2.14	
	FFNN	**0.81±0.04**	**93.19±1.41**	**84.61±4.04**	94.03±2.43	
	CNN	**0.81±0.04**	**93.70±1.15**	82.66±4.80	**95.23±2.14**	
IC-IT	LR	0.82±0.03	91.22±1.61	91.00±3.11	91.44±3.44	1.38e-17
	kNN	0.74±0.06	86.44±3.22	89.83±4.33	83.56±5.56	
	SVM	0.85±0.04	**92.28±1.67**	91.67±3.61	**93.00±3.56**	
	RF	0.79±0.04	89.28±2.22	86.28±6.06	91.72±6.06	
	FFNN	0.84±0.04	**92.06±2.17**	**92.11±3.56**	92.11±3.94	
	CNN	**0.86±0.03**	**92.67±1.67**	91.61±3.17	**93.78±3.39**	

Our comprehensive investigation across distinct protein classification tasks, employing various classifiers, revealed a number of compelling insights.

3.2.1. Prominence of SVM and CNN Classifiers

Both the Support Vector Machine (SVM) and Convolutional Neural Network (CNN) classifiers consistently delivered superior performance across all tasks. These classifiers effectively navigate high-dimensional data and unravel complex patterns, contributing to their consistent performance. The CNN employs convolutional layers to identify local patterns in the representations and nonlinear relationships inherent in neural network layers, while the SVM excels at linear classification by distinguishing between classes efficiently by maximizing margins.

3.2.2. Comparison of Simple and Complex Models

Interestingly, a comparison of simple models, such as Logistic Regression (LR), and complex ones, like CNNs, indicated comparable performance levels. This observation counters the prevalent assumption that increasing model complexity necessarily results in superior performance. The consistent trend across all tasks and evaluation metrics suggests that in predicting IC and IT from MP, simpler models may deliver effectiveness on par with their more complex counterparts.

3.2.3. Less Effective Classifiers

However, not all classifiers showcased this level of effectiveness. Classifiers such as the k-Nearest Neighbors (kNN) and Random Forest (RF) were identified as the least effective across these tasks and representations derived from PLMs. This finding suggests that these classifiers may not align well with the specific nature of these tasks or the representations provided by the PLMs.

3.2.4. Performance Parallels Among Classifiers

Furthermore, our analysis disclosed an intriguing parallel in the performance metrics of LR and Feed-Forward Neural Networks (FFNN), and those of SVM and CNN. This pattern suggests that, despite inherent differences in their complexity and structure, these models can achieve similar results in these specific tasks.

3.2.5. Significance of Classifier Selection

Finally, the p-value analysis highlighted significant performance differences across the classifiers for all three tasks, emphasizing the crucial role of classifier selection in the outcomes of these prediction tasks. The observed variation implies that the effectiveness of a specific classifier may vary based on the unique characteristics of the task, underscoring the importance of thoughtful classifier selection.

3.3. Effects of Various Experimental Conditions

In this section, we delve deeper into our findings and their implications. We have conducted three distinct assessments to elucidate their impacts on the results and overall performance. The following subsections offer a comprehensive discussion on these critical areas of impact, namely, the implications of frozen vs. fine-tuned representations, the influence of balanced vs. imbalanced datasets, and the effects of half vs. full precision floating-point computations.

3.3.1. Frozen vs. Fine-tuned PLM Representations

Table 6 presents the impact of frozen and fine-tuned representations across the three tasks under consideration - IC-MP, IT-MP, and IC-IT. Additionally, Figure 2 underscores the performance, specifically focusing on the MCC metric across the three tasks. Note that a comprehensive analysis concerning the influence of frozen and fine-tuned representations is available in Section A.

Table 6. Comparison and evaluation of frozen and fine-tuned representations across diverse protein language models (PLMs). This table delineates the impact of utilizing both frozen and fine-tuned representations on three distinct tasks: differentiating Ion Channels (IC) from Membrane Proteins (MP), segregating Ion Transporters (IT) from MPs, and discriminating IC from IT, utilizing a range of PLMs. Four evaluation metrics have been computed using 5-fold cross-validation, presented as mean±standard deviation. The p-value is provided as a metric of the statistical significance of observed discrepancies among the models. Notably, the highest performance value for each task and each column is highlighted in boldface.

Task	Representation	MCC	Accuracy	Sensitivity	Specificity	P-value
IC-MP	frozen	0.75±0.05	**90.54±2.10**	**90.52±4.04**	90.65±4.33	1.57e-08
	finetuned	**0.83±0.04**	**90.75±2.08**	**90.33±3.92**	**91.17±4.32**	
IT-MP	frozen	0.70±0.05	**93.11±1.41**	**86.71±3.93**	**93.44±2.25**	2.33e-12
	finetuned	**0.83±0.04**	92.33±1.47	77.61±4.80	**93.06±2.26**	
IC-IT	frozen	**0.82±0.04**	**92.81±1.37**	**88.22±3.56**	**93.24±2.16**	7.15e-01
	finetuned	0.81±0.04	91.37±1.45	73.48±4.87	92.68±2.31	

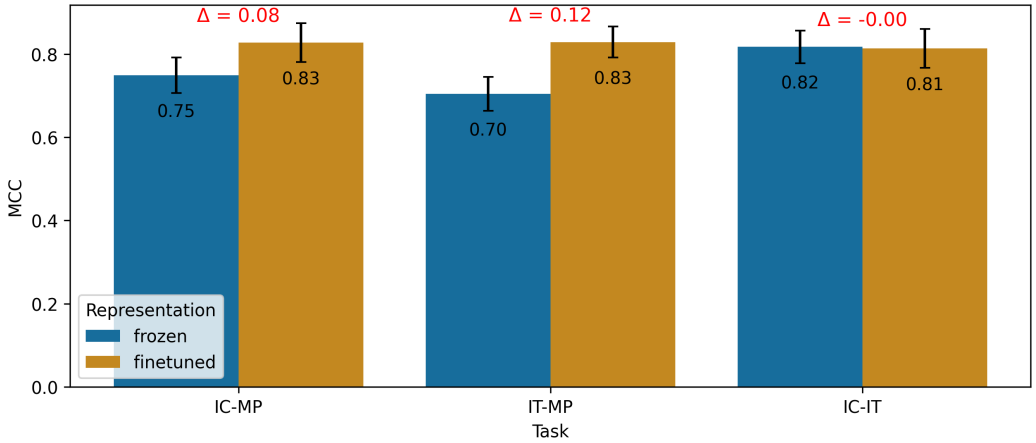

Figure 2. Graphical representation of the impact of frozen vs. fine-tuned representations on various tasks across different Protein Language Models (PLMs). This figure elucidates the impact of employing frozen and fine-tuned representations across a range of Protein Language Models (PLMs) for three distinct tasks: differentiating Ion Channels (IC) from Membrane Proteins (MP), distinguishing Ion Transporters (IT) from MPs, and discriminating IC from IT. The results are portrayed using the mean Matthew's Correlation Coefficient (MCC) values derived from 5-fold cross-validation. Each bar represents the mean MCC calculated across five cross-validation runs, while the error bars indicate the associated standard deviation. The symbol Δ is employed to denote the disparity between the corresponding pair of bars.

Our investigation has uncovered noteworthy disparities in the performance of fine-tuned and frozen representations across various tasks, underscored by their responses to task-specific conditions, dataset sizes, classifier choices, and the underlying PLM's architecture.

Task-specific Performance Variations and the Impact of Dataset Imbalances

On differentiating IC from MP and IT from MP, fine-tuned representations have consistently outperformed frozen ones. This pattern, however, becomes less clear-cut in the IC-IT task. Statistical analysis further supports this pattern, revealing substantial performance discrepancies between frozen and fine-tuned representations in the IC-MP and IT-MP tasks. However, the IC-IT task showed no significant difference.

This relative performance convergence in the IC-IT task can be attributed to the balanced nature of its dataset, contrasting with potential imbalances in the MP dataset. This highlights the role of dataset balance in performance trends and suggests that evaluation metrics may capture varying aspects of model performance, particularly under conditions of dataset imbalance.

A case in point is the sensitivity metric for the IT-MP task. Here, frozen representations notably outshine their fine-tuned counterparts, contrasting with the general trend of fine-tuned superiority. This demonstrates the sensitivity metric's specific susceptibility to the effects of dataset imbalance. Whereas MCC metric, which accounts for all types of prediction errors, demonstrated equivalent performance for both representation types.

Influence of Dataset Size on Performance

Our analysis points towards a significant influence of dataset size on the performance of fine-tuned representations. The larger, albeit imbalanced, MP dataset, comprising 3,413 training sequences, rendered richer fine-tuned representations compared to the balanced dataset of 280 sequences (see Section 3.3.2). Consequently, the benefits of fine-tuning appear more distinct with larger datasets, underscoring the potential of using extensive data resources to enhance fine-tuned PLM representation performance.

The observed pattern suggests that larger models, such as ProtT5 and ESM-2_15B—currently unexplored due to computational limitations—could potentially exhibit improved performance given the feasibility of fine-tuning.

Performance Across Different Classifiers

A further probe into performance across all classifiers, as represented in Table A2 and Figure A2, demonstrated the consistent outperformance of fine-tuned over frozen representations. This observation reinforces the role of fine-tuning as a potent strategy to optimize PLM effectiveness across varied classifier architectures.

Performance across Diverse PLMs

Our findings, as showcased in Table A1 and Figure A1, reveal that performance remains relatively stable between diverse PLM sizes when using frozen representations. However, ESM-1b, a larger model with 650M parameters, outperformed smaller-sized PLMs like ProtBERT with 420M parameters. This observation suggests that the size of the underlying PLM can exert influence on the performance of frozen representations.

3.3.2. Balanced vs. Imbalanced Datasets

Table 7 and Figure 3 present the performance of the six PLMs when applied to either a balanced or imbalanced MP dataset. Our analysis suggests a profound effect of dataset balance on the performance of different representations across PLMs, classifiers, and tasks.

Performance Across PLMs

Our results, as presented in Table 7 and Figure 3, indicate that representations from imbalanced datasets outperform those from balanced datasets across six PLMs, with the exception of ProtT5 and ESM-2_15B. This inconsistency may arise from the lack of fine-tuned representations for these specific PLMs. Given the feasibility of fine-tuning, we expect that these PLMs would align with the overall trend, affirming the performance advantage of imbalanced datasets.

However, the reported p-value in Table 7 suggests no significant difference between balanced and imbalanced datasets for ProtBERT, ProtT5, and ESM-2_15B PLMs. As ProtT5 and ESM-2_15B were not fine-tuned, the observed p-value primarily reflects the impact of dataset balance on the performance of frozen representations for these PLMs.

Table 7. Performance of Protein Language Models (PLMs) on Balanced vs. Imbalanced Membrane Protein Datasets. This comprehensive evaluation examines the performance of various Protein Language Models (PLMs) on both balanced and imbalanced datasets of membrane proteins. The results, computed using 5-fold cross-validation, are represented as mean±standard deviation for the evaluation metrics. The p-value quantifies the statistical significance of observed differences amongst the classifiers. The highest values for each task and column are highlighted in bold. The PLMs are sorted based on their number of parameters.

PLM	Dataset	MCC	Accuracy	Sensitivity	Specificity	P-value
ProtBERT	balanced	0.70±0.06	89.14±2.42	**89.00±3.33**	89.27±3.89	2.52e-01
	imbalanced	**0.74±0.04**	**98.48±0.06**	84.52±3.52	**99.58±0.10**	
ProtBERT-BFD	balanced	0.71±0.06	88.55±2.45	**88.94±3.61**	88.24±4.12	1.57e-02
	imbalanced	**0.77±0.03**	**97.98±0.19**	78.79±5.02	**99.56±0.08**	
ESM-1b	balanced	0.79±0.05	87.83±2.52	**89.81±3.38**	85.87±3.78	1.38e-04
	imbalanced	**0.87±0.02**	**96.92±0.17**	67.83±5.25	**99.17±0.25**	
ESM-2	balanced	0.78±0.05	84.82±2.94	**85.83±4.14**	83.87±5.04	9.25e-03
	imbalanced	**0.83±0.03**	**96.92±0.23**	66.71±5.44	**99.40±0.12**	
ProtT5	balanced	**0.79±0.05**	85.31±3.19	**85.59±4.54**	85.07±4.77	4.33e-01
	imbalanced	0.75±0.04	**97.25±0.08**	69.50±5.06	**99.50±0.17**	
ESM-2_15B	balanced	**0.77±0.05**	89.08±2.35	**89.32±3.48**	88.77±3.67	6.05e-01
	imbalanced	0.73±0.03	**96.83±0.17**	67.92±5.92	**99.17±0.08**	

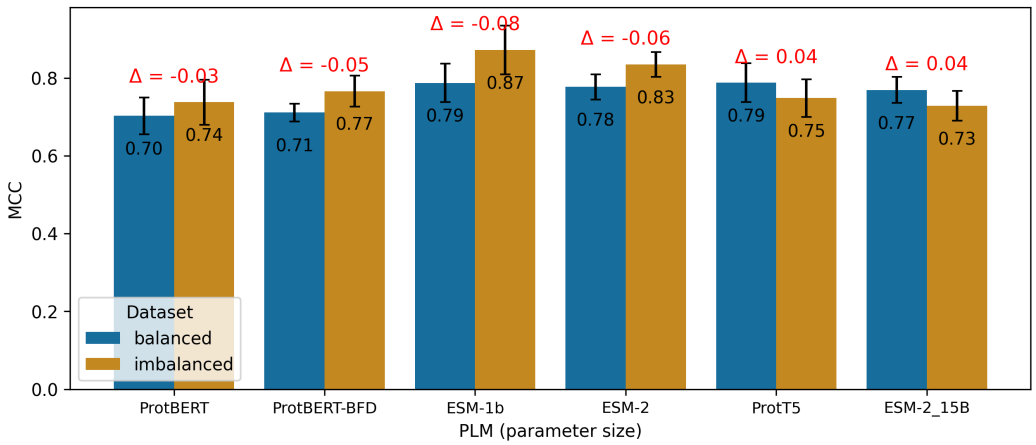

Figure 3. Evaluation of PLMs on balanced and imbalanced datasets of membrane proteins. This figure showcases a comprehensive evaluation of various protein language models (PLMs) on both balanced and imbalanced datasets of membrane proteins. The evaluation results are depicted as the mean Matthews Correlation Coefficient (MCC) calculated over five cross-validation runs, with error bars denoting the standard deviation. The symbol Δ indicates the difference between the corresponding pair of bars, providing insights into the performance disparities across the evaluated PLMs.

Task-specific Performance Variations

Evidence from Table A3 and Figure A3 indicates a superior performance of imbalanced datasets in the IC-MP and IT-MP tasks. These findings underscore the impact of dataset balance on model performance across these specific tasks.

Performance Across Different Classifiers

The comparison of classifier performances presented in Table A6 and Figure A6 suggests that imbalanced datasets outshine balanced datasets across all classifiers, except

for the RF classifier. This exception implies a particular sensitivity of the RF classifier to dataset balance, potentially explaining its performance divergence from the other classifiers.

Fine-Tuned vs. Frozen Representations

The performance patterns as seen in Table A4 and Figure A4 demonstrate that imbalanced datasets exhibit superior performance when employing fine-tuned representations across all fine-tuned PLMs. In contrast, balanced datasets perform better when using frozen representations, except for ProtBERT, where the p-value of 8.66e-02 indicates a statistically significant difference. These findings emphasize the significant impact of dataset balance on model performance, dependent on the choice of representation type (fine-tuned or frozen).

3.3.3. Half vs. Full Precision Floating Point Calculations

Table 8 and Figure 4 present the outcomes obtained from employing half and full precision floating-point calculations across the classifiers. Our analysis explores the influence of numerical precision—specifically half versus full precision floating-point calculations—on the performance of different tasks, classifiers, and PLMs.

Performance Across Different Classifiers

As evidenced by the results presented in Table 8 and Figure 4, the performance remains consistent across all classifiers, irrespective of whether half or full precision floating-point calculations are employed. This suggests that the level of numerical precision does not significantly affect classifier performance in the evaluated tasks.

Task-specific Performance Variations

Performance consistency extends to specific tasks as well. As shown in Table A7 and Figure A7, the IC-MP, IT-MP, and IC-IT tasks exhibit comparable performance levels regardless of the employed floating-point precision. These findings reinforce the notion that the numerical precision choice for the floating-point calculations does not materially affect model performance across these tasks.

Table 8. Performance of half vs. full precision floating-point across six classifiers. This table provides an overview of the performance of each classifier using half and full precision floating-point calculations. The results are presented using evaluation metrics on the 5-fold cross-validation, with the mean and standard deviation shown. The p-value indicates the statistical significance of the observed differences among the classifiers. The highest value for each task and each column is highlighted in bold.

Classifier	Precision	MCC	Accuracy	Sensitivity	Specificity	P-value
LR	half	**0.82±0.04**	**93.56±1.62**	**85.54±5.08**	94.99±3.04	9.69e-01
	full	0.81±0.04	**93.62±1.53**	**85.32±4.58**	**95.02±3.10**	
kNN	half	0.69±0.05	**93.20±1.50**	**86.95±3.90**	**93.78±2.56**	9.01e-01
	full	**0.70±0.05**	**93.43±1.45**	**86.92±3.90**	**93.99±2.44**	
SVM	half	**0.83±0.04**	92.93±1.42	**85.91±3.80**	93.65±2.44	9.22e-01
	full	**0.83±0.04**	**93.22±1.35**	**85.88±3.68**	**94.00±2.29**	
RF	half	0.69±0.05	**90.81±1.73**	**70.20±5.09**	**94.31±2.76**	9.64e-01
	full	**0.70±0.05**	**90.77±1.58**	67.29±4.63	**94.68±2.61**	
FFNN	half	**0.82±0.04**	**93.40±1.28**	**86.41±3.98**	**94.59±2.24**	9.27e-01
	full	**0.82±0.04**	**93.63±1.26**	**86.30±3.99**	**94.81±2.13**	
CNN	half	**0.83±0.04**	**87.85±2.04**	**83.29±4.37**	**84.35±3.19**	8.09e-01
	full	**0.83±0.04**	**87.60±2.03**	**83.46±4.42**	**83.60±3.22**	

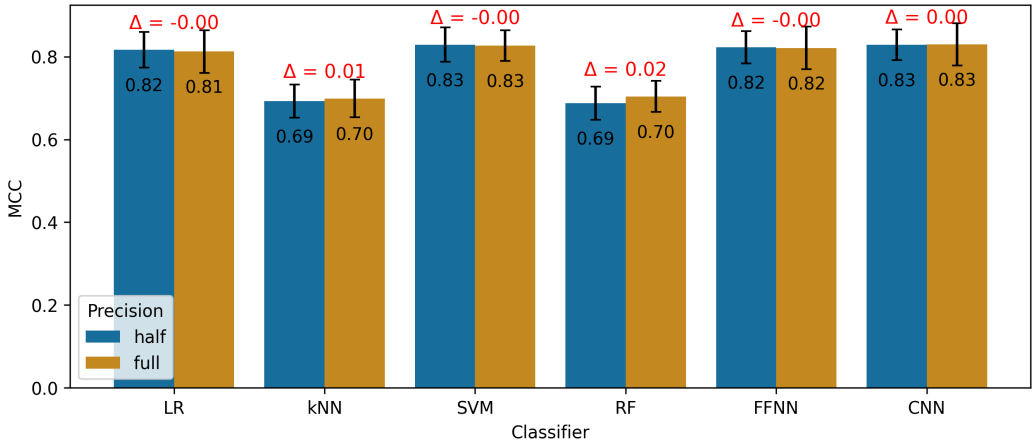

Figure 4. Half vs. full precision evaluation across classifiers. This evaluation compares the performance of different protein language models (PLMs) using both half and full precision floating-point calculations. The results are presented as the mean Matthews Correlation Coefficient (MCC) calculated across five cross-validation runs, with error bars indicating the standard deviation. The symbol Δ represents the difference between the corresponding pair of bars, providing insights into the impact of numerical precision on classifier performance.

Performance Across PLMs

The performance comparison among the six PLMs, as displayed in Table A8 and Figure A8, reveals minor performance variations when using both half and full precision floating-point calculations. This observation implies that the selection of floating-point precision has minimal impact on the performance of the evaluated PLMs.

Influence on Evaluation Metrics and Statistical Significance

An overarching analysis of evaluation metrics and p-values reveals no statistically significant differences between the usage of half and full precision floating-point calculations across varied tasks, classifiers, and PLMs. These findings underscore that the choice of floating-point precision does not exert a considerable influence on the outcomes of the prediction tasks assessed in this study.

3.4. Visualization of Representations: Insights and Implications

The UMAP projection matrix of representations derived from the ESM-1b PLM, presented in Figure 5, provides a compelling visualization of both frozen and fine-tuned representations for balanced and imbalanced datasets within the context of the IC-MP task on the training set. It is crucial to note that the representation shown for the balanced dataset is randomly selected from one of the ten available balanced datasets.

3.4.1. Fine-tuned Representations in Imbalanced Dataset

The Figure 5 visualization underscores the distinct clusters and patterns within the fine-tuned representations for the imbalanced dataset. The evident separation between ion channels and membrane proteins signifies the highly discriminative capability of fine-tuned representations, demonstrating their efficacy in this task. This insight underscores the prowess of fine-tuned representations in capturing the unique and distinguishable characteristics of ion channels, fostering precise classification and analysis.

3.4.2. Frozen Representations in Imbalanced Dataset

Notably, the visualization also indicates that the next best level of clarity is achieved using frozen representations with the imbalanced dataset. This suggests that the imbalanced dataset, enriched with a broader spectrum of other membrane proteins, enhances the performance of the frozen representations. This may be due to the diversity and complexity

of the other membrane proteins, requiring a larger dataset for effective representation and discrimination. Hence, this highlights the advantage of employing imbalanced datasets with frozen representations for capturing the intricacies of diverse membrane protein structures.

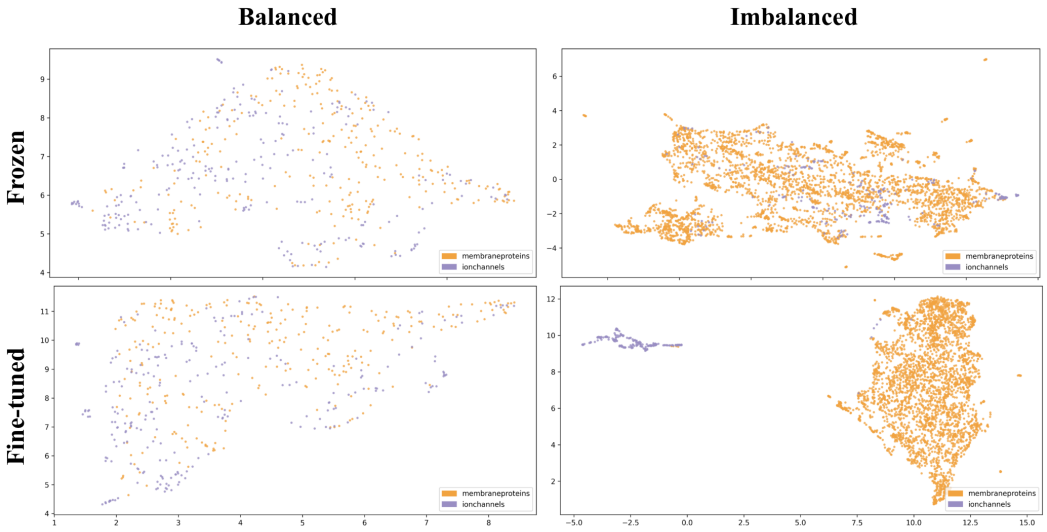

Figure 5. UMAP projection of representations from top PLM for ion channel discrimination. The figure showcases a UMAP projection of representations derived from ESM-1b, the highest-performing Protein Language Model (PLM) in the task of discriminating ion channels (IC) from membrane proteins (MP). The representations are visualized in four variations: frozen and fine-tuned representation types, along with balanced and imbalanced datasets. In the visualization, membrane protein representations are depicted in yellow, while ion channel protein representations are depicted in blue.

3.4.3. Impact of Undersampling on Classification Task

Our results accentuate the potential adverse consequences of undersampling the dataset on the classification task performance. Undersampling, which reduces the dataset size, can impair the model's ability to classify proteins accurately, underscoring the need for a sufficiently large dataset to ensure effective protein classification. A substantial dataset ensures the model's exposure to diverse and representative examples, facilitating the learning of robust, discriminative patterns that generalize well to unseen data. Consequently, securing a substantial dataset is of paramount importance for achieving optimal performance in protein classification tasks.

3.4.4. Implications for Balanced Dataset Representations

Examining the visualization of frozen and fine-tuned representations with balanced datasets, we find a lack of clear patterns. This signifies a less distinct characterization of ion channels compared to other membrane proteins, suggesting these representations may not effectively differentiate ion channels from other membrane proteins. This lack of clear patterns implies that the representations derived from balanced datasets may fail to capture unique features or discriminative information vital for robust ion channel classification. Hence, alternative representation strategies or dataset balancing techniques may warrant consideration to enhance model effectiveness.

3.4.5. Comprehensive Visualization of PLMs

The representation visualizations for all six PLMs, including both frozen and fine-tuned representations for the IC-MP, IT-MP, and IC-IT tasks, are provided in Section D. As shown in Figure A9, Figure A10, and Figure A11, these visualizations offer a holistic view of the performance and discriminative abilities of various PLMs and representations for these tasks. These comprehensive visualizations allow for an in-depth understanding of

how different PLMs capture the characteristics and separability of ion channels and other membrane proteins, illuminating their respective strengths and weaknesses.

3.5. Overview of Top Cross-Validation Results

Table 9. Top 5-fold CV results for each task and classifier, along with independent test set results. This table presents the best 5-fold cross-validation (CV) results for each task and classifier, as well as the corresponding results on the independent test set for comparison purposes. The tasks include discriminating ion channels (IC) from other membrane proteins (MP), ion transporters (IT) from MP, and IC against IT. The table displays the mean and standard deviation of the 5-fold CV results for each metric. The results for the IC-MP and IT-MP tasks are obtained from imbalanced datasets, while the dataset for the IC-IT task remains balanced. The best values for each task are shown in bold, and the second-best values are underlined. It is important to note that the independent test set results are provided solely for evaluating the models based on the CV results and not for selecting the best model, as the best models are chosen based on the CV results.

Task	Representation	Representer	Dataset	Classifier	MCC	
					CV	Independent
IC-MP	finetuned	ESM-1b	Imbalanced	SVM	0.99±0.01	**0.85**
				RF	0.98±0.01	0.84
				kNN	0.99±0.01	0.83
				LR	**1.00±0.00**	**0.85**
				FFNN	**1.00±0.01**	**0.85**
				CNN	0.99±0.01	**0.85**
IT-MP	finetuned	ESM-1b	Imbalanced	SVM	**1.00±0.00**	0.68
				RF	0.99±0.01	0.67
				kNN	0.99±0.01	**0.70**
				LR	**1.00±0.00**	0.69
				FFNN	**1.00±0.01**	0.67
				CNN	0.99±0.01	0.69
IC-IT	frozen	ESM-2_15B	Balanced	SVM	0.88±0.03	**0.88**
	finetuned	ESM-1b		RF	0.84±0.03	0.79
	frozen	ProtT5		kNN	0.81±0.03	0.75
	finetuned	ESM-1b		LR	0.88±0.05	0.79
	frozen	ESM-2		FFNN	0.88±0.05	0.74
	finetuned	ESM-2		CNN	**0.89±0.03**	0.87

The top results obtained from the 5-fold cross-validation (CV) for each task are detailed in Table 9. Results are stratified by classifier and presented in the CV column, showing the mean and standard deviation over the five folds. While independent test set results are provided for comparative purposes, they do not contribute to the selection of the best model, ensuring a robust and unbiased evaluation of classifier performance.

3.5.1. Superior Performance of ESM-1b PLM in IC-MP and IT-MP Tasks

As outlined in Table 9, the ESM-1b PLM, combined with fine-tuned representations and an imbalanced dataset, exhibits superior performance in the IC-MP and IT-MP tasks. The LR and FFNN classifiers, in particular, achieve a perfect MCC of 1.00, indicating flawless prediction on 5-fold CV. Other classifiers also present highly competitive results, with MCC values reaching 0.99, thereby emphasizing the exceptional efficacy of the ESM-1b PLM with fine-tuning and an imbalanced dataset.

3.5.2. Results from Multiple PLMs in IC-IT Task

The IC-IT task, employing a balanced dataset, sees a range of PLMs delivering notable results. The top-performing classifier, CNN, leverages the ESM-2 PLM with fine-tuned representations, yielding an impressive MCC of 0.89. Notably, larger PLMs like ProtT5 and

ESM-2_15B produce comparable results to their smaller counterparts such as ESM-1b and ESM-2. This suggests that the size of the PLM does not necessarily influence performance enhancement for the IC-IT task.

3.5.3. Comparative Performance of Classifiers for IC-IT Task

While the CNN classifier utilizing the ESM-2 PLM's fine-tuned representations achieves the top result for the IC-IT task, other classifiers also demonstrate comparable performances. The SVM classifier with frozen representations from ESM-2_15B, the LR classifier with fine-tuned representations from ESM-1b, and the FFNN classifier with frozen representations from ESM-2 deliver similar results to the CNN classifier. This suggests that a diverse set of classifiers can deliver equivalent performance levels, depending on the selected PLM and representation type.

3.5.4. Comprehensive Analysis of Results

A detailed examination of the results for each task - IC-MP, IT-MP, and IC-IT - is provided in Section E. Here, the evaluation metrics are delineated in detail across various tables for each task. This thorough breakdown offers an exhaustive and nuanced understanding of the performance of the employed models, classifiers, and representations. Delving into the evaluation metrics' specifics enables readers to gain deeper insights into the results, providing valuable information for future research in the prediction of ion channels and ion transporters from other membrane proteins.

3.6. Performance Comparison with State-of-the-Art Projects

A detailed comparison of TooT-PLM-ionCT's performance against state-of-the-art projects is provided in Table 10 and Figure 6 for the IC-MP, IT-MP, and IC-IT tasks. This analysis includes established methodologies such as DeepIon [18], MFPS_CNN [22], and TooT-BERT-C [23], providing a comprehensive assessment of TooT-PLM-ionCT's relative performance.

As shown in Table 10 and Figure 6, TooT-PLM-ionCT outperforms its counterparts in the IT-MP and IC-IT tasks. However, in the IC-MP task, its performance aligns closely with TooT-BERT-C. These results underscore the capability of TooT-PLM-ionCT to accurately predict ion channels and ion transporters from other membrane proteins, demonstrating its superiority or competitive performance.

It's worth noting that DeepIon [18] and MFPS_CNN [22] do not report specific results for the IC-IT task, as they focus predominantly on differentiating ion channels and ion transporters from other membrane proteins. This further underscores the unique contribution of our study in exploring the IC-IT task and offering crucial insights into the categorization of ion channels and ion transporters from other membrane proteins.

Model Selection Process

The model selection was driven by the top-performing models in our experiments, as detailed in Table 9. In instances where multiple classifiers achieved the same MCC, we favored the simpler and more straightforward classifier for the IC-MP and IT-MP tasks. However, for the IC-IT task, despite the SVM classifier's marginally better performance on the independent test set, the CNN classifier was selected based on superior CV results. This decision balanced the need for performance with model simplicity, while considering the unique demands and constraints of each task.

Table 10. Comparative performance of TooT-PLM-ionCT with state-of-the-art. This table provides a comparative analysis of the performance of TooT-PLM-ionCT with the state-of-the-art methods on the independent test set. The performance is evaluated for classifying membrane proteins (MP), ion channels (IC), and ion transporters (IT). The absence of results is denoted by a "-" when corresponding studies and tools do not report ion channel and ion transporter classification against each other. The boldface highlights the highest values in the accuracy and Matthews Correlation Coefficient (MCC) columns, while the underline indicates the second-highest values.

Task	Project	Encoder	Classifier	Accuracy	MCC
IC-MP	DeepIon [18]	PSSM	CNN	86.53	0.37
	MFPS_CNN [22]	PSSM	CNN	<u>94.60</u>	<u>0.62</u>
	TooT-BERT-C [23]	ProtBERT-BFD	LR	**98.24**	**0.85**
	TooT-PLM-ionCT	ESM-1b	LR	**98.24**	**0.85**
IT-MP	DeepIon [18]	PSSM	CNN	83.78	0.37
	MFPS_CNN [22]	PSSM	CNN	93.30	0.59
	TooT-BERT-C [23]	ProtBERT-BFD	LR	<u>95.43</u>	<u>0.64</u>
	TooT-PLM-ionCT	ESM-1b	LR	**95.98**	**0.69**
IC-IT	DeepIon [18]	-	-	-	-
	MFPS_CNN [22]	-	-	-	-
	TooT-BERT-C [23]	ProtBERT-BFD	LR	<u>85.38</u>	<u>0.71</u>
	TooT-PLM-ionCT	ESM-2	CNN	**93.07**	**0.87**

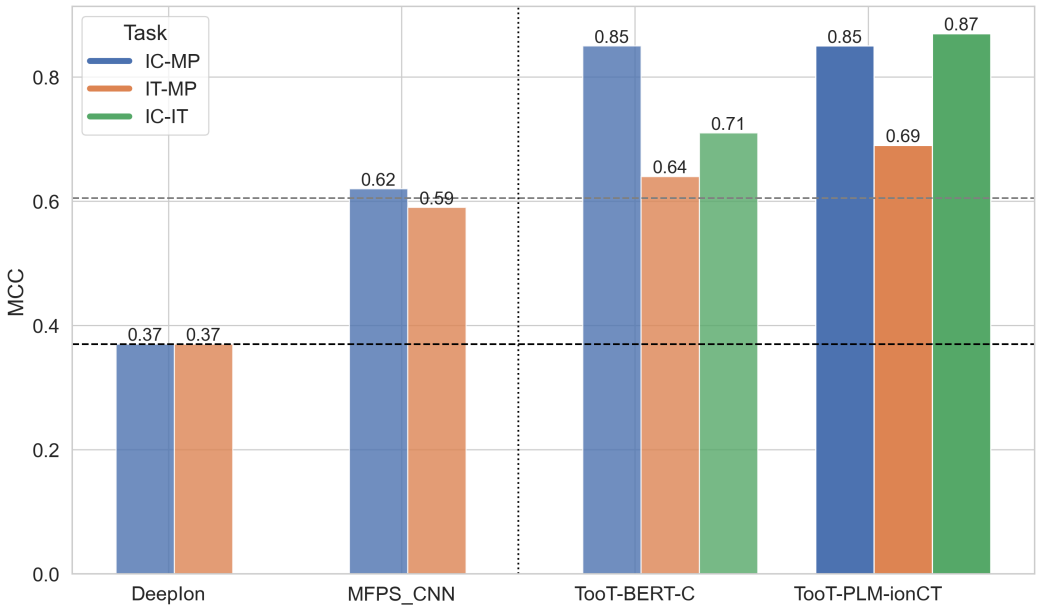

Figure 6. Comparative performance with state-of-the-art. This figure presents the comparative performance of TooT-PLM-ionCT on the independent test set, showcasing the classification results for membrane proteins (MP), ion channels (IC), and ion transporters (IT). The absence of bars indicates studies that focused on classifying ion channels and ion transporters against membrane proteins, rather than against each other, resulting in no available results from either publications or tools. The horizontal dashed lines represent two baselines, while the vertical dashed line distinguishes between traditional and PLM-based representations.

4. Conclusions

This study presented TooT-PLM-ionCT, a framework developed for distinguishing ion channels (IC) from other membrane proteins (MP), ion transporters (IT) from MP, and IC from IT. Six Protein Language Models (PLMs) were utilized: ProtBERT, ProtBERT-BFD,

and ProtT5 from the ProtTrans project, along with ESM-1b, ESM-2 (650M parameters), and ESM-2 (15B parameters) from the ESM project. These were employed alongside a range of traditional (Logistic Regression, kNN, Random Forest (RF), SVM, and Feed-Forward Neural Network) and deep learning (Convolutional Neural Network) classifiers.

The study scrutinized the effects of dataset balance, the comparison of frozen and fine-tuned representations, and the performance difference between half-precision and full-precision floating-point calculations. The significant findings from our analysis are discussed below:

- **PLM Performance:** ESM-1b PLM outshone its peers in most metrics and tasks, with the exception of distinguishing IC from IT, where it shared the top spot with ESM-2_15B. The second-best performing model, however, varied with the task at hand. ESM-2 proved effective in differentiating IC from MP and IT from MP, while ProtT5 excelled in IC-IT classification. The substantial variation in p-values of statistical analysis across all PLMs further emphasized ESM-1b's formidable performance.
- **Dataset Balance:** Our study found that imbalanced datasets outperformed balanced datasets across most PLMs, except for ProtT5 and ESM-2_15B, where we saw inconsistency due to the absence of fine-tuned representations. Additionally, a comparison of classifier performance revealed that imbalanced datasets outperformed balanced datasets across all classifiers. The sole exception was the RF classifier, which exhibited a heightened sensitivity to balanced datasets and therefore yielded superior results with them.
- **Fine-Tuned Representations:** Fine-tuned representations consistently performed better than frozen ones for differentiating IC from MP and IT from MP, while for the IC-IT task, the performance was equivocal. The size of the dataset appeared to significantly influence the performance of fine-tuned representations. Thus, larger datasets, despite their imbalanced nature, seemed to benefit more from fine-tuning.
- **Floating-Point Precision:** Our study found negligible performance variations between half and full precision floating-point calculations across tasks, classifiers, and PLMs. This suggests that the numerical precision choice does not considerably impact the performance in the prediction tasks examined in this study.
- **Impact of Undersampling:** Results highlighted the potential detrimental effects of undersampling, emphasizing the need for larger, more representative datasets for accurate protein classification.
- **Comparison of PLM Sizes:** Our analysis showed an intriguing pattern where a 650M-parameter PLM exhibited comparable performance to a 15B-parameter PLM and surpassed a 450M-parameter model in terms of frozen representation.
- **Computational Cost vs. Improvement:** The improvement in performance for the IC-IT task justified the associated computational cost, a contrast to the IC-MP and IT-MP tasks where the benefit did not outweigh the cost.

In our future endeavors, we aspire to probe the feasibility of augmenting the representations produced by PLMs with additional sources of knowledge. Concurrently, we aim to assess more sophisticated techniques for sequence representation, pushing the boundaries of current methodologies to further enhance the depth and breadth of our protein analysis.

We are also committed to expanding the scope of our approach, testing its efficacy on larger and more diverse protein datasets. By doing so, we aim not only to validate our methodology's robustness but also to potentially broaden its range of applicability.

Author Contributions: Conceptualization, HG and GB; methodology, HG and GB; software, HG; validation, HG; formal analysis, HG and GB; investigation, HG; resources, HG and GB; data curation, HG; writing—original draft preparation, HG and GB; writing—review and editing, HG and GB; visualization, HG; supervision, GB; project administration, GB; All authors have read and agreed to the published version of the manuscript.

Funding: Both authors are supported by Natural Sciences and Engineering Research Council of Canada (NSERC), Genome Québec, and Genome Canada and Concordia University.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data used in this study is available at https://tootsuite.encs.concordia.ca/datasets/TooT-BERT-C, and the code for the proposed method, TooT-PLM-ionCT, is available at https://github.com/bioinformatics-group/tootplmionct.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IC	Ion channel
IT	Ion transporter
MP	Membrane protein
PLM	Protein Language Model
kNN	k-Nearest Neighbors
RF	Random Forest
SVM	Support Vector Machine
LR	Logistic Regression
FFNN	Feed-Forward Neural Network
CNN	Convolutional Neural Network
MCC	Matthews Correlation Coefficient
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

Appendix A. Frozen vs. Fine-tuned Representations

Table A1. Frozen vs. fine-tuned representations across protein language models. This table presents a comparison and evaluation of frozen versus fine-tuned representations across a range of protein language models (PLMs). The assessment is based on four evaluation metrics computed using a 5-fold cross-validation procedure and is presented as the mean ± standard deviation. Statistical significance of observed discrepancies among the models is denoted by the provided p-value. Please note, instances of 'None' indicate that due to resource constraints, we were unable to fine-tune larger PLMs such as ProtT5 and ESM-2 with 15 billion parameters.

PLM	Representation	MCC	Accuracy	Sensitivity	Specificity	P-value
ProtBERT	frozen	0.69±0.05	94.14±1.48	93.80±2.65	94.17±2.42	2.78e-06
	finetuned	0.78±0.04	92.94±1.41	82.16±4.20	93.76±2.46	
ProtBERT-BFD	frozen	0.70±0.05	93.35±1.40	90.32±3.55	93.62±2.46	1.40e-06
	finetuned	0.79±0.05	92.44±1.57	80.35±4.81	93.36±2.58	
ESM-1b	frozen	0.79±0.04	92.36±1.41	81.36±3.99	92.58±2.38	5.31e-07
	finetuned	0.88±0.03	91.09±1.51	83.56±4.47	91.45±2.84	
ESM-2	frozen	0.77±0.05	90.04±1.69	74.04±4.94	91.01±3.08	5.40e-07
	finetuned	0.85±0.04	91.34±1.74	84.63±4.56	91.97±2.80	
ProtT5	frozen	0.78±0.04	90.19±1.86	74.76±5.17	91.41±3.02	None
ESM-2_15B	frozen	0.77±0.04	92.73±1.37	81.09±4.29	93.67±2.27	None

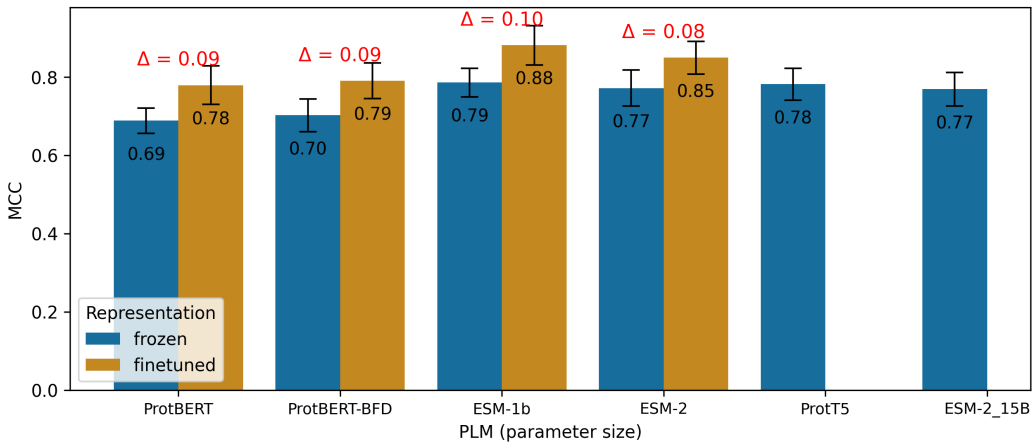

Figure A1. This figure provides a graphical display of the differential impact of employing frozen and fine-tuned representations across various Protein Language Models (PLMs). The comparison is made using the mean Matthew's Correlation Coefficient (MCC) values, as determined from 5-fold cross-validation. Each bar signifies the mean MCC obtained across the cross-validation sets, with error bars representing the standard deviation. The delta symbol (Δ) illustrates the difference between the associated pair of bars. Absent bars denote the inability to fine-tune large PLMs such as ProtT5 and ESM-2, each containing 15 billion parameters, due to resource limitations.

Table A2. Frozen vs. fine-tuned representations across classifiers. This table presents a comparison and evaluation of frozen versus fine-tuned representations across a range of classifiers. The assessment is based on four evaluation metrics computed using a 5-fold cross-validation procedure and is presented as the mean ± standard deviation. Statistical significance of observed discrepancies among the models is denoted by the provided p-value.

Classifier	Representation	MCC	Accuracy	Sensitivity	Specificity	P-value
LR	frozen	0.79±0.04	93.94±1.53	89.42±3.84	95.08±2.97	2.56e-05
	finetuned	0.84±0.04	93.32±1.60	82.21±5.58	94.95±3.16	
kNN	frozen	0.65±0.05	93.62±1.47	89.93±3.49	94.01±2.49	1.13e-05
	finetuned	0.75±0.05	93.09±1.48	84.53±4.24	93.81±2.49	
SVM	frozen	0.81±0.04	93.35±1.36	89.28±3.35	93.94±2.34	2.67e-05
	finetuned	0.85±0.03	92.88±1.39	83.18±4.03	93.77±2.38	
RF	frozen	0.61±0.05	91.74±1.61	82.22±4.70	94.43±2.66	3.02e-06
	finetuned	0.80±0.04	90.03±1.67	57.67±4.94	94.58±2.69	
FFNN	frozen	0.80±0.04	93.73±1.25	89.67±3.48	94.69±2.21	3.97e-05
	finetuned	0.85±0.04	93.36±1.28	83.69±4.39	94.74±2.15	
CNN	frozen	0.81±0.04	88.47±1.97	87.94±3.98	84.67±3.11	4.68e-06
	finetuned	0.86±0.04	87.10±2.09	79.74±4.74	83.35±3.28	

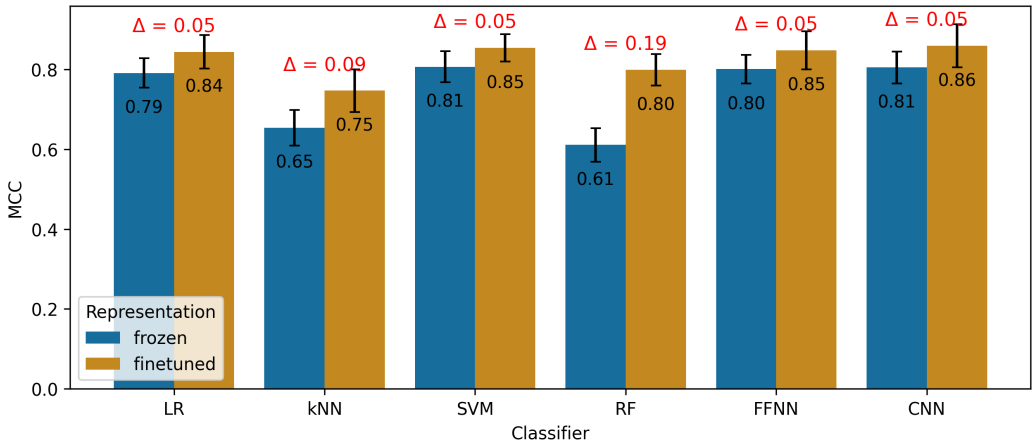

Figure A2. This figure provides a graphical display of the differential impact of employing frozen and fine-tuned representations across various classifiers. The comparison is made using the mean Matthew's Correlation Coefficient (MCC) values, as determined from 5-fold cross-validation. Each bar signifies the mean MCC obtained across the cross-validation sets, with error bars representing the standard deviation. The delta symbol (Δ) illustrates the difference between the associated pair of bars.

Appendix B. Balanced vs. Imbalanced Datasets

Table A3. Balanced vs. imbalanced dataset performance across tasks. This table presents a comparison and evaluation of balanced versus imbalanced dataset peformance across the tasks of ion channels vs. other membrane proteins (MP) and ion transporters vs. MP. The assessment is based on four evaluation metrics computed using a 5-fold cross-validation procedure and is presented as the mean ± standard deviation. Statistical significance of observed discrepancies among the models is denoted by the provided p-value.

Task	Dataset	MCC	Accuracy	Sensitivity	Specificity	P-value
IC-MP	balanced	0.76±0.05	87.47±2.73	88.10±4.03	86.88±4.42	5.50e-04
	imbalanced	0.81±0.03	97.89±0.16	75.21±4.80	99.58±0.09	
IT-MP	balanced	0.74±0.05	86.77±2.70	87.07±3.67	86.50±4.32	1.44e-02
	imbalanced	0.78±0.03	97.25±0.13	72.99±4.91	99.36±0.16	

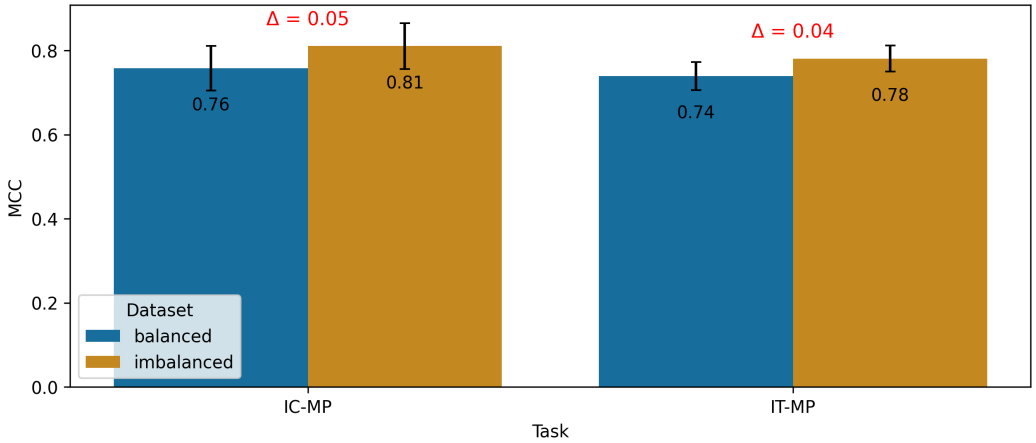

Figure A3. This figure provides a graphical display of the differential impact of employing balanced and imbalanced dataset across various tasks of ion channels (IC) vs. other membrane proteins (MP) and ion transporters (IT) vs. MP. The comparison is made using the mean Matthew's Correlation Coefficient (MCC) values, as determined from 5-fold cross-validation. Each bar signifies the mean MCC obtained across the cross-validation sets, with error bars representing the standard deviation. The delta symbol (Δ) illustrates the difference between the associated pair of bars.

Table A4. Balanced vs. imbalanced dataset performance across fine-tuned protein language models. This table presents a comparison and evaluation of balanced versus imbalanced dataset peformance across the fine-tuned protein language models. The assessment is based on four evaluation metrics computed using a 5-fold cross-validation procedure and is presented as the mean ± standard deviation. Statistical significance of observed discrepancies among the models is denoted by the provided p-value.

PLM	Dataset	MCC	Accuracy	Sensitivity	Specificity	P-value
ProtBERT	balanced	0.71±0.06	89.21±2.40	89.00±3.33	89.39±3.89	1.55e-08
	imbalanced	0.85±0.03	100.00±0.00	99.12±1.29	100.00±0.00	1.55e-08
ProtBERT-BFD	balanced	0.71±0.06	88.57±2.42	88.96±3.66	88.26±4.03	3.27e-10
	imbalanced	0.87±0.03	99.04±0.04	91.17±3.50	99.88±0.04	3.27e-10
ESM-1b	balanced	0.79±0.05	84.96±2.86	85.95±4.00	84.05±4.93	1.43e-14
	imbalanced	0.99±0.01	98.00±0.21	78.29±5.17	99.83±0.04	1.43e-14
ESM-2	balanced	0.78±0.05	85.48±3.18	85.66±4.57	85.41±4.66	6.96e-10
	imbalanced	0.93±0.02	98.42±0.00	81.58±4.46	99.92±0.04	6.96e-10

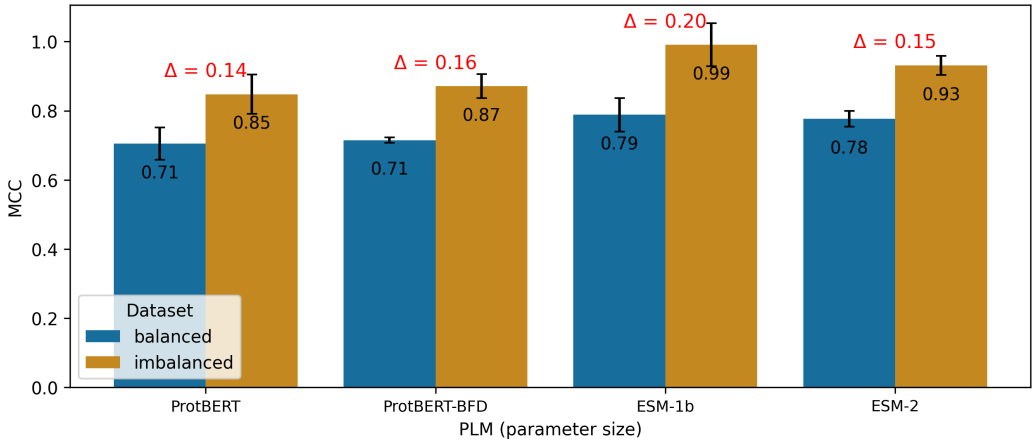

Figure A4. Balanced vs. imbalanced dataset performance across fine-tuned PLMs. This figure provides a graphical display of the differential impact of employing balanced and imbalanced datasets across various fine-tuned Protein Language Models (PLMs). The comparison is made using the mean Matthew's Correlation Coefficient (MCC) values, as determined from 5-fold cross-validation. Each bar signifies the mean MCC obtained across the cross-validation sets, with error bars representing the standard deviation. The delta symbol (Δ) illustrates the difference between the associated pair of bars.

Table A5. Balanced vs. imbalanced dataset performance across frozen protein language models. This table presents a comparison and evaluation of balanced versus imbalanced dataset peformance across the frozen protein language models. The assessment is based on four evaluation metrics computed using a 5-fold cross-validation procedure and is presented as the mean ± standard deviation. Statistical significance of observed discrepancies among the models is denoted by the provided p-value.

PLM	Dataset	MCC	Accuracy	Sensitivity	Specificity	P-value
ProtBERT	balanced	0.70±0.06	89.07±2.45	89.01±3.34	89.15±3.89	8.66e-02
	imbalanced	0.63±0.04	96.96±0.12	69.92±5.75	99.17±0.21	
ProtBERT-BFD	balanced	0.71±0.06	88.52±2.47	88.91±3.56	88.22±4.20	1.34e-01
	imbalanced	0.66±0.04	96.92±0.33	66.42±6.54	99.25±0.12	
ESM-1b	balanced	0.79±0.05	87.83±2.52	89.81±3.38	85.87±3.78	2.34e-01
	imbalanced	0.75±0.04	96.92±0.17	67.83±5.25	99.17±0.25	
ESM-2	balanced	0.78±0.05	84.67±3.02	85.71±4.28	83.70±5.16	2.46e-01
	imbalanced	0.74±0.04	95.83±0.25	55.12±5.71	98.96±0.21	
ProtT5	balanced	0.79±0.05	85.14±3.20	85.52±4.52	84.73±4.87	4.33e-01
	imbalanced	0.75±0.04	96.08±0.17	57.42±5.67	99.08±0.29	
ESM-2_15B	balanced	0.77±0.05	89.08±2.35	89.32±3.48	88.77±3.67	6.05e-01
	imbalanced	0.73±0.03	96.83±0.17	67.92±5.92	99.17±0.08	

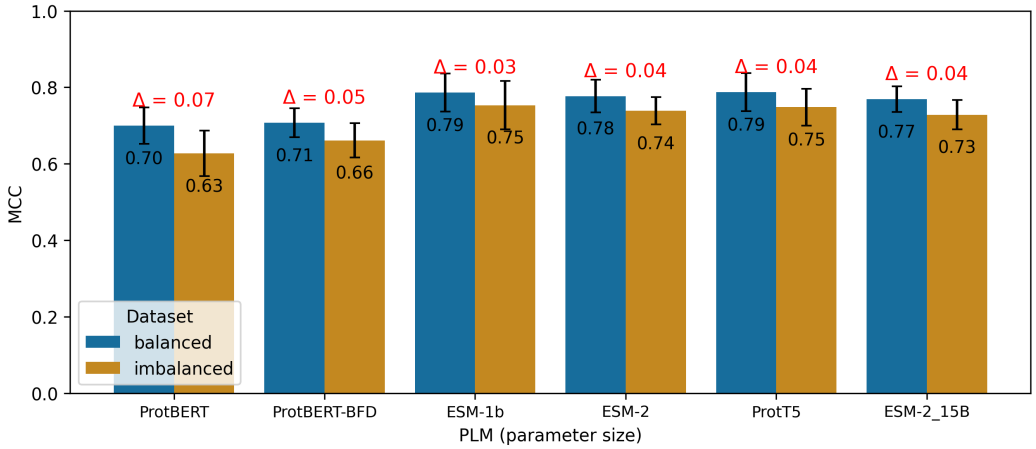

Figure A5. Balanced vs. imbalanced dataset performance across frozen PLMs. This figure provides a graphical display of the differential impact of employing balanced and imbalanced dataset across various frozen Protein Language Models (PLMs). The comparison is made using the mean Matthew's Correlation Coefficient (MCC) values, as determined from 5-fold cross-validation. Each bar signifies the mean MCC obtained across the cross-validation sets, with error bars representing the standard deviation. The delta symbol (Δ) illustrates the difference between the associated pair of bars.

Table A6. Balanced vs. imbalanced dataset performance across classifiers. This table presents a comparison and evaluation of balanced versus imbalanced dataset peformance across classifiers. The assessment is based on four evaluation metrics computed using a 5-fold cross-validation procedure and is presented as the mean ± standard deviation. Statistical significance of observed discrepancies among the models is denoted by the provided p-value.

Classifier	Dataset	MCC	Accuracy	Sensitivity	Specificity	P-value
LR	balanced	0.78±0.05	89.49±2.81	87.85±4.74	91.07±5.68	5.91e-04
	imbalanced	0.84±0.03	98.17±0.28	79.89±5.69	99.56±0.31	
kNN	balanced	0.60±0.06	89.32±2.55	89.38±3.31	89.27±4.17	1.99e-07
	imbalanced	0.77±0.03	97.97±0.06	81.89±4.67	99.42±0.08	
SVM	balanced	0.79±0.05	89.14±2.53	88.71±3.35	89.47±4.18	1.83e-04
	imbalanced	0.85±0.03	97.97±0.11	80.53±4.42	99.42±0.00	
RF	balanced	0.73±0.06	86.14±2.92	81.34±4.81	90.43±3.66	1.13e-02
	imbalanced	0.62±0.04	96.19±0.08	46.97±4.25	100.00±0.00	
FFNN	balanced	0.79±0.05	89.60±2.31	88.34±3.21	90.76±3.56	1.28e-04
	imbalanced	0.85±0.03	98.08±0.03	81.69±4.94	99.53±0.11	
CNN	balanced	0.80±0.05	79.03±3.17	89.90±3.69	69.14±4.99	8.20e-04
	imbalanced	0.85±0.03	97.03±0.31	73.64±5.14	98.92±0.25	

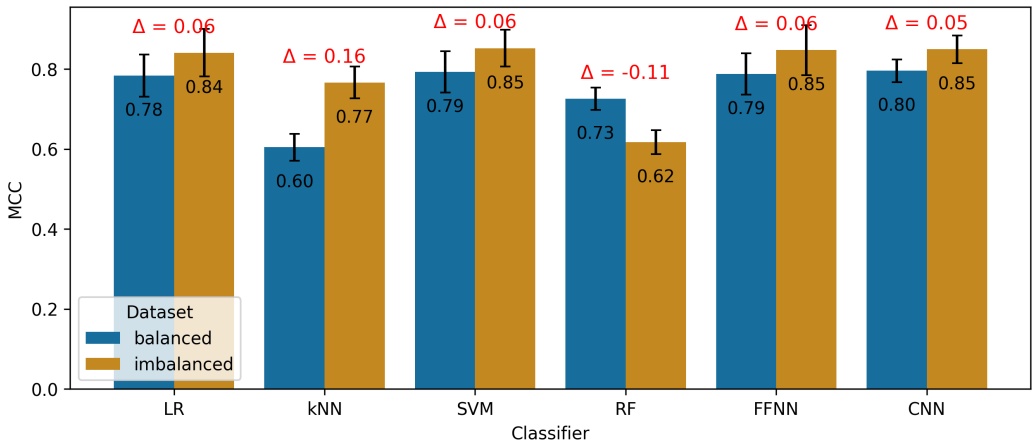

Figure A6. Balanced vs. imbalanced dataset performance across classifiers. This figure provides a graphical display of the differential impact of employing balanced and imbalanced dataset across various classifiers. The comparison is made using the mean Matthew's Correlation Coefficient (MCC) values, as determined from 5-fold cross-validation. Each bar signifies the mean MCC obtained across the cross-validation sets, with error bars representing the standard deviation. The delta symbol (Δ) illustrates the difference between the associated pair of bars.

Appendix C. Half vs. Full Precision Floating Point Calculations

Table A7. Half vs. full precision floating point calculations across tasks. This table presents a comparison and evaluation of half versus full precision floating-point across the tasks of ion channels (IC) vs. other membrane proteins (MP), ion transporter (IT) vs. MP, and IC vs. IT. The assessment is based on four evaluation metrics computed using a 5-fold cross-validation procedure and is presented as the mean ± standard deviation. Statistical significance of observed discrepancies among the models is denoted by the provided p-value.

Task	Precision	MCC	Accuracy	Sensitivity	Specificity	P-value
IC-MP	half	0.78±0.04	90.46±2.17	90.21±4.19	90.73±4.44	9.75e-01
	full	0.78±0.04	90.82±2.03	90.58±3.80	91.10±4.23	
IT-MP	half	0.76±0.04	92.65±1.46	81.89±4.43	93.18±2.29	7.48e-01
	full	0.76±0.04	92.71±1.42	81.47±4.40	93.27±2.23	
IC-IT	half	0.82±0.04	92.03±1.45	80.62±4.40	92.99±2.25	9.34e-01
	full	0.81±0.04	92.00±1.39	79.56±4.20	92.88±2.23	

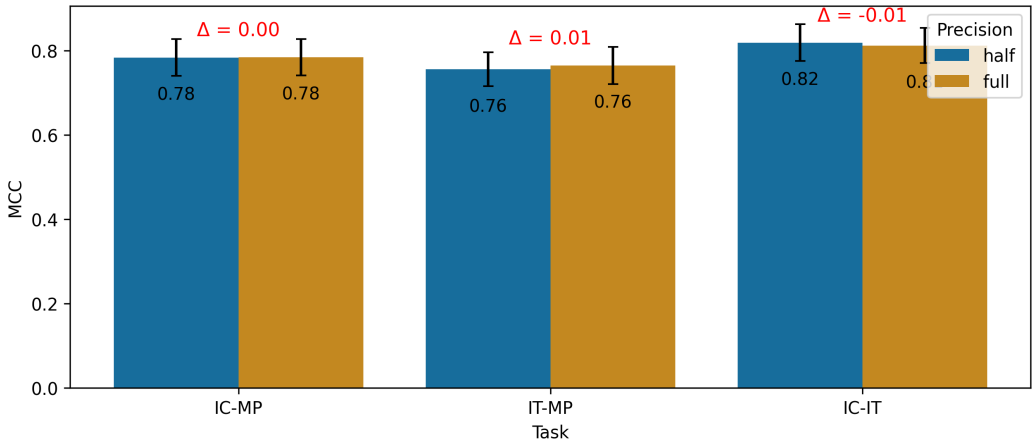

Figure A7. Half vs. full precision floating point calculations across tasks. This figure provides a graphical display of the differential impact of employing half and full precision floating-point calculation across various tasks of ion channels (IC) vs. other membrane proteins (MP), ion transporters (IT) vs. MP and IC vs. IT. The comparison is made using the mean Matthew's Correlation Coefficient (MCC) values, as determined from 5-fold cross-validation. Each bar signifies the mean MCC obtained across the cross-validation sets, with error bars representing the standard deviation. The delta symbol (Δ) illustrates the difference between the associated pair of bars.

Table A8. Half vs. full precision floating point calculations across protein language models. This table presents a comparison and evaluation of half versus full precision floating-point across protein language models (PLMs). The assessment is based on four evaluation metrics computed using a 5-fold cross-validation procedure and is presented as the mean ± standard deviation. Statistical significance of observed discrepancies among the models is denoted by the provided p-value. Please note, instances of 'None' indicate that due to resource constraints, we were unable to fine-tune larger PLMs such as ProtT5 and ESM-2 with 15 billion parameters.

PLM	Precision	MCC	Accuracy	Sensitivity	Specificity	P-value
ProtBERT	half	0.73±0.04	93.52±1.45	87.91±3.57	93.98±2.46	7.41e-01
	full	0.74±0.05	93.56±1.44	88.05±3.28	93.96±2.42	
ProtBERT-BFD	half	0.75±0.05	92.94±1.48	85.44±4.09	93.55±2.46	9.59e-01
	full	0.75±0.05	92.85±1.49	85.23±4.26	93.43±2.58	
ESM-1b	half	0.83±0.04	92.36±1.41	81.36±3.99	92.58±2.38	9.13e-01
	full	0.83±0.04	90.60±1.65	79.20±4.87	91.23±2.98	
ESM-2	half	0.81±0.04	90.52±1.55	78.39±4.54	91.24±2.94	8.09e-01
	full	0.81±0.04	90.78±1.80	79.65±4.95	91.71±2.92	
ProtT5	half	0.78±0.04	90.75±1.80	79.74±4.78	91.67±2.89	None
ESM-2_15B	half	0.77±0.04	92.73±1.37	81.09±4.29	93.67±2.27	None

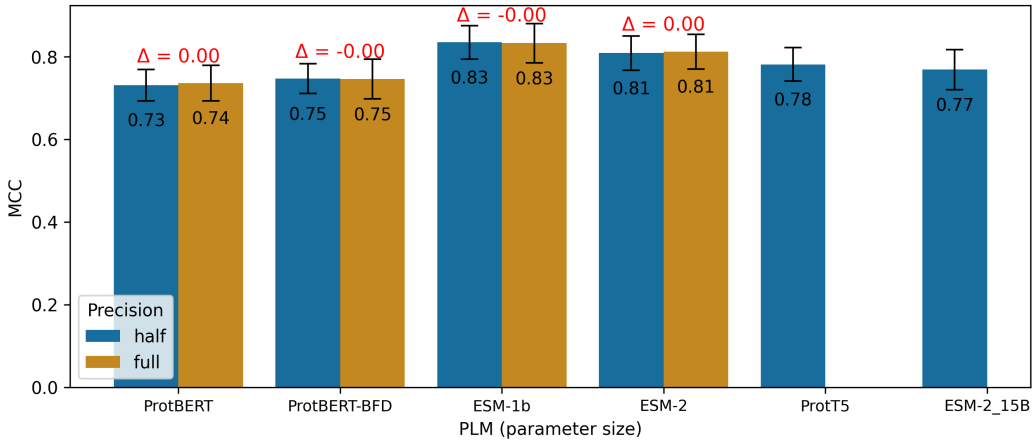

Figure A8. Half vs. full precision floating point calculations across PLMs. This figure provides a graphical display of the differential impact of employing half and full precision floating-point calculation across various Protein Language Models (PLMs). The comparison is made using the mean Matthew's Correlation Coefficient (MCC) values, as determined from 5-fold cross-validation. Each bar signifies the mean MCC obtained across the cross-validation sets, with error bars representing the standard deviation. The delta symbol (Δ) illustrates the difference between the associated pair of bars. Absent bars denote the inability to fine-tune large PLMs such as ProtT5 and ESM-2, each containing 15 billion parameters, due to resource limitations.

Appendix D. Protein visualization

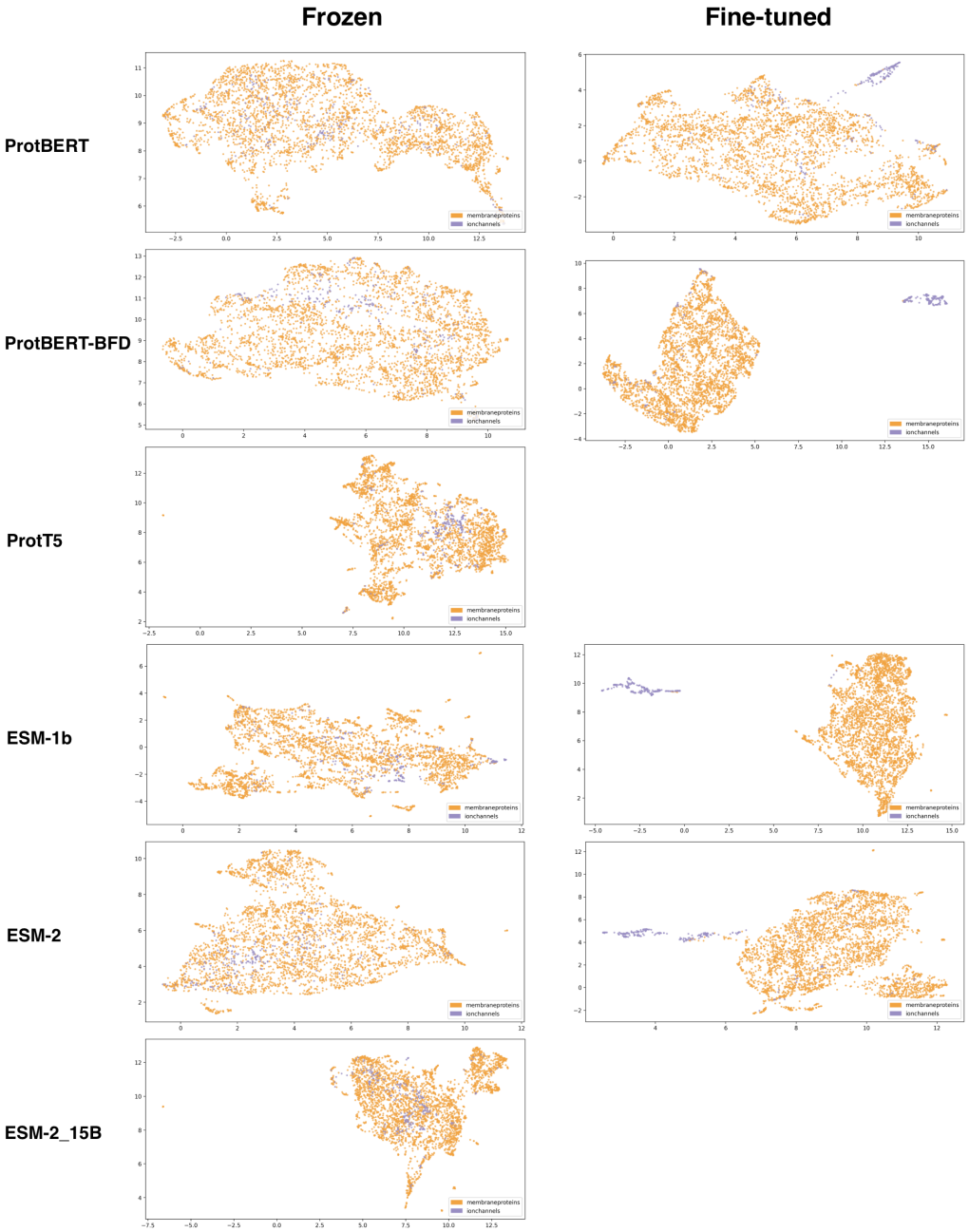

Figure A9. This figure illustrates a UMAP projection visualizing the separation of ion channels and an imbalanced dataset of other membrane proteins. The visualization encompasses all six protein language models and includes both frozen and fine-tuned representation types. Membrane proteins are represented by yellow points, while ion channels are depicted in blue.

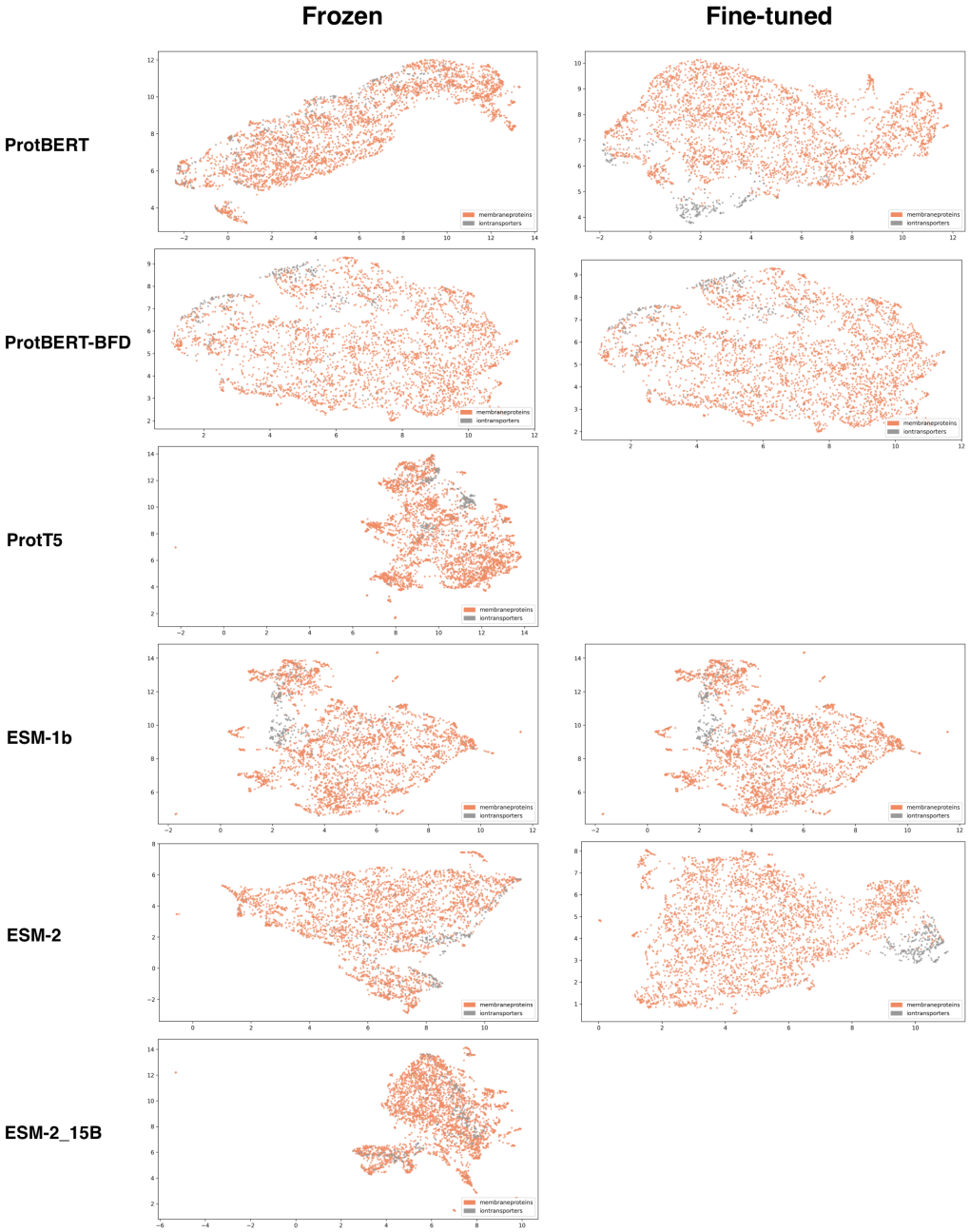

Figure A10. This figure illustrates a UMAP projection visualizing the separation of ion transporters and an imbalanced dataset of other membrane proteins. The visualization encompasses all six protein language models and includes both frozen and fine-tuned representation types. Membrane proteins are represented by red points, while ion transporters are depicted in grey.

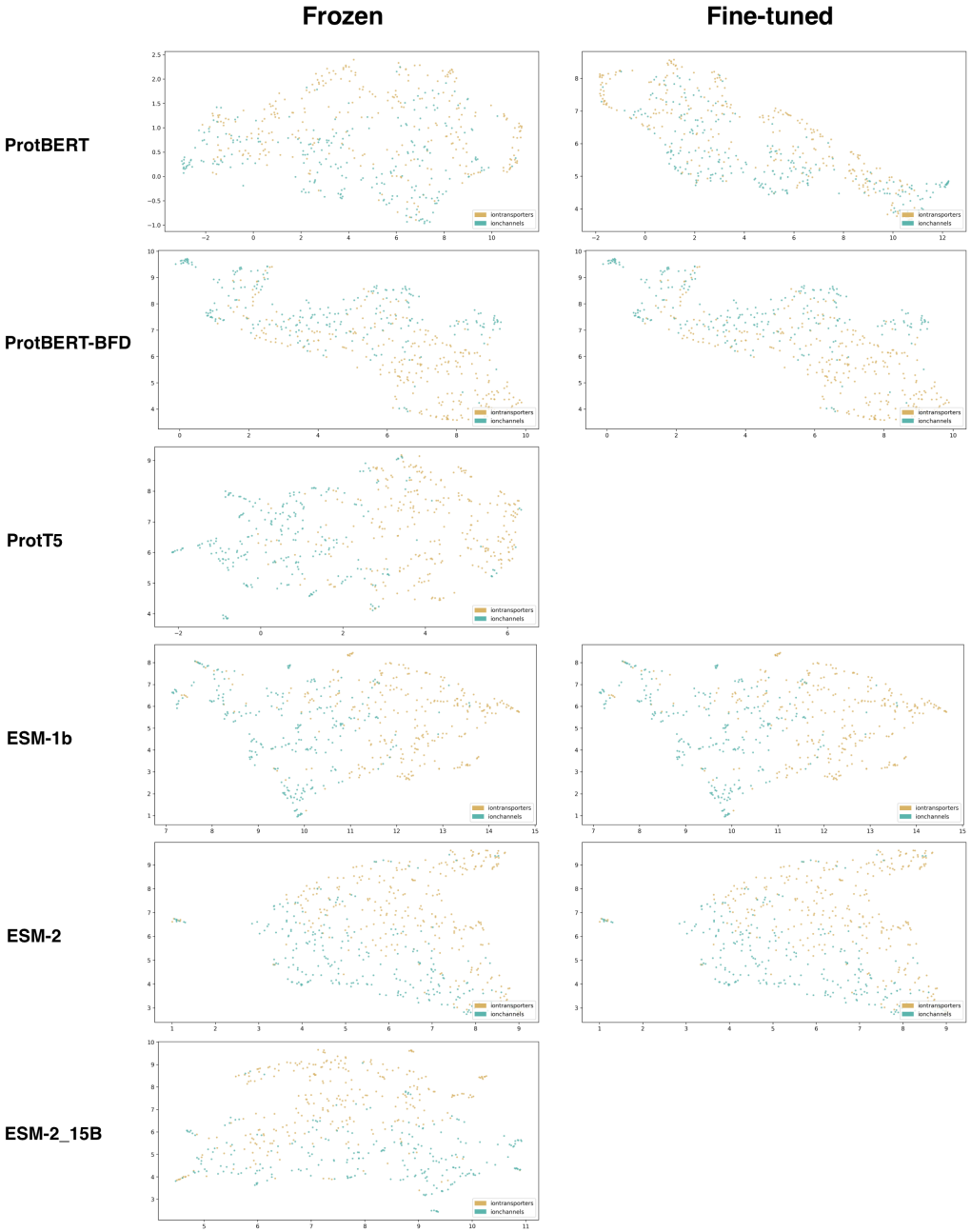

Figure A11. This figure illustrates a UMAP projection visualizing the separation of ion channels and ion transporters. The visualization encompasses all six protein language models and includes both frozen and fine-tuned representation types. Ion channels are represented by yellow points, while ion transporters are depicted in green.

Appendix E. Detailed five-fold cross-validation results

Appendix E.1. Ion channels vs. other membrane proteins

Table A9. Comparison of representations and classifiers performance for discriminating **ion channels from membrane proteins** on Accuracy metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Dataset	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-2_15B	frozen	imbalanced	half	99.00±0.00	99.00±1.00	95.00±0.00	96.00±0.00	99.00±0.00	99.00±0.00
		balanced	half	92.20±2.20	93.10±1.90	89.80±2.30	68.90±3.20	93.50±2.00	93.40±2.20
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ProtBERT	frozen	imbalanced	half	98.00±0.00	97.00±0.00	94.00±0.00	95.00±1.00	98.00±0.00	97.00±0.00
		balanced	half	86.60±4.60	87.80±3.10	86.20±2.40	72.50±4.10	86.70±3.40	87.60±2.70
		imbalanced	full	98.00±1.00	97.00±0.00	94.00±0.00	95.00±1.00	98.00±0.00	97.00±0.00
		balanced	full	86.78±3.56	87.70±3.10	86.30±2.70	72.50±4.20	86.70±3.40	87.50±2.60
	finetuned	imbalanced	half	98.00±0.00	98.00±0.00	97.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00
		balanced	half	86.90±3.60	87.70±2.90	86.50±2.60	73.20±3.90	87.30±2.80	87.70±2.50
		imbalanced	full	99.00±1.00	99.00±0.00	98.00±1.00	98.00±0.00	98.00±1.00	98.00±1.00
		balanced	full	86.50±3.80	87.80±3.20	86.10±2.50	72.90±4.30	87.30±2.80	87.70±2.70
ESM-2	frozen	imbalanced	half	99.00±1.00	99.00±0.00	95.00±0.00	97.00±1.00	98.00±0.00	98.00±0.00
		balanced	half	91.00±3.40	92.40±1.70	88.10±3.00	80.50±2.90	91.80±2.00	91.90±2.00
		imbalanced	full	99.00±1.00	99.00±0.00	95.00±0.00	97.00±1.00	98.00±0.00	98.00±0.00
		balanced	full	91.90±3.00	92.40±1.70	88.00±2.90	80.30±2.80	91.90±2.00	92.00±1.90
	finetuned	imbalanced	half	100.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	half	91.60±2.80	92.30±2.00	88.00±3.00	80.40±2.40	91.80±2.00	91.70±2.00
		imbalanced	full	100.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	full	91.90±2.50	92.30±1.80	88.20±2.90	80.50±2.60	91.90±2.00	92.00±1.80
ESM-1b	frozen	imbalanced	half	98.00±1.00	99.00±0.00	96.00±0.00	97.00±0.00	98.00±0.00	98.00±0.00
		balanced	half	90.40±4.00	92.80±1.80	88.50±2.70	80.70±2.70	92.00±1.70	91.90±1.80
		imbalanced	full	98.00±0.00	99.00±0.00	96.00±0.00	97.00±0.00	98.00±0.00	98.00±0.00
		balanced	full	91.00±2.30	92.80±1.80	88.70±2.60	80.70±2.70	91.90±1.80	91.90±1.80
	finetuned	imbalanced	half	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	half	90.70±3.90	92.80±1.70	88.30±2.70	81.20±2.80	91.80±1.70	91.70±1.80
		imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	91.20±2.50	92.80±1.50	88.50±2.50	81.40±2.70	91.80±1.80	91.70±1.90
ProtT5	frozen	imbalanced	half	98.00±1.00	98.00±0.00	95.00±0.00	97.00±1.00	98.00±0.00	98.00±0.00
		balanced	half	91.00±2.90	92.00±2.20	88.80±2.30	80.10±3.10	90.70±1.80	90.90±2.30
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ProtBERT-BFD	frozen	imbalanced	half	97.00±1.00	97.00±0.00	94.00±0.00	96.00±0.00	97.00±0.00	97.00±0.00
		balanced	half	87.50±3.80	88.30±2.20	86.30±2.90	77.60±3.20	86.40±3.60	87.40±2.90
		imbalanced	full	97.00±1.00	97.00±0.00	94.00±0.00	96.00±0.00	97.00±0.00	97.00±0.00
		balanced	full	86.20±4.30	88.30±2.20	86.30±3.30	77.60±3.10	86.70±3.50	87.50±3.10
	finetuned	imbalanced	half	98.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00
		balanced	half	87.40±4.40	88.60±2.50	86.20±2.50	78.30±2.90	87.20±3.70	88.10±2.80
		imbalanced	full	98.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00
		balanced	full	87.67±4.00	88.60±2.40	86.20±2.90	78.30±3.10	87.30±3.70	88.00±3.20

Table A10. Comparison of representations and classifiers performance for discriminating **ion channels from membrane proteins** on MCC metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Dataset	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-1b	finetuned	imbalanced	half	0.99±0.01	0.99±0.01	0.98±0.01	0.99±0.01	1.00±0.00	1.00±0.01
		balanced	half	0.82±0.07	0.85±0.04	0.77±0.05	0.66±0.05	0.84±0.04	0.83±0.04
		imbalanced	full	0.99±0.01	0.99±0.01	0.97±0.01	0.98±0.01	0.99±0.01	0.99±0.01
		balanced	full	0.83±0.04	0.85±0.04	0.77±0.05	0.66±0.05	0.84±0.04	0.83±0.04
	frozen	imbalanced	half	0.83±0.07	0.88±0.03	0.58±0.03	0.78±0.03	0.83±0.04	0.85±0.04
		balanced	half	0.81±0.07	0.85±0.04	0.78±0.05	0.65±0.05	0.84±0.04	0.84±0.04
		imbalanced	full	0.87±0.04	0.88±0.03	0.59±0.04	0.78±0.03	0.83±0.04	0.85±0.04
		balanced	full	0.82±0.04	0.85±0.04	0.78±0.05	0.65±0.05	0.84±0.04	0.84±0.04
ESM-2	finetuned	imbalanced	half	0.97±0.02	0.95±0.03	0.90±0.01	0.90±0.03	0.95±0.02	0.95±0.02
		balanced	half	0.84±0.05	0.85±0.04	0.76±0.05	0.64±0.05	0.83±0.04	0.84±0.04
		imbalanced	full	0.97±0.01	0.95±0.01	0.91±0.02	0.90±0.02	0.95±0.02	0.95±0.03
		balanced	full	0.84±0.05	0.84±0.03	0.77±0.06	0.64±0.05	0.83±0.04	0.84±0.04
	frozen	imbalanced	half	0.88±0.05	0.88±0.03	0.51±0.05	0.75±0.05	0.87±0.04	0.86±0.04
		balanced	half	0.83±0.06	0.85±0.04	0.76±0.06	0.64±0.06	0.84±0.04	0.84±0.04
		imbalanced	full	0.87±0.05	0.88±0.03	0.52±0.06	0.75±0.05	0.87±0.04	0.86±0.04
		balanced	full	0.84±0.05	0.85±0.04	0.77±0.06	0.63±0.06	0.84±0.04	0.84±0.04
ESM-2_15B	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	frozen	imbalanced	half	0.88±0.03	0.88±0.05	0.40±0.03	0.72±0.03	0.89±0.03	0.88±0.03
		balanced	half	0.85±0.04	0.86±0.04	0.80±0.05	0.47±0.06	0.87±0.04	0.87±0.04
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ProtBERT	finetuned	imbalanced	half	0.87±0.02	0.86±0.03	0.73±0.02	0.79±0.03	0.83±0.03	0.84±0.03
		balanced	half	0.75±0.06	0.76±0.06	0.73±0.06	0.51±0.07	0.75±0.06	0.75±0.05
		imbalanced	full	0.88±0.05	0.88±0.04	0.80±0.06	0.84±0.04	0.85±0.05	0.87±0.05
		balanced	full	0.74±0.07	0.76±0.06	0.72±0.05	0.50±0.08	0.74±0.06	0.75±0.05
	frozen	imbalanced	half	0.81±0.02	0.78±0.03	0.31±0.05	0.54±0.05	0.79±0.03	0.79±0.03
		balanced	half	0.75±0.08	0.75±0.06	0.72±0.05	0.49±0.08	0.73±0.07	0.75±0.05
		imbalanced	full	0.81±0.05	0.78±0.03	0.30±0.04	0.54±0.06	0.79±0.03	0.78±0.04
		balanced	full	0.75±0.06	0.76±0.06	0.73±0.06	0.49±0.08	0.74±0.07	0.75±0.05
ProtT5	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	frozen	imbalanced	half	0.87±0.05	0.86±0.03	0.54±0.07	0.76±0.06	0.82±0.04	0.84±0.03
		balanced	half	0.83±0.06	0.84±0.04	0.78±0.05	0.64±0.06	0.82±0.04	0.82±0.05
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ProtBERT-BFD	finetuned	imbalanced	half	0.82±0.02	0.87±0.04	0.86±0.03	0.84±0.03	0.86±0.04	0.85±0.04
		balanced	half	0.76±0.08	0.77±0.05	0.72±0.05	0.60±0.05	0.75±0.08	0.76±0.05
		imbalanced	full	0.82±0.03	0.83±0.03	0.82±0.05	0.81±0.04	0.83±0.04	0.82±0.04
		balanced	full	0.77±0.07	0.77±0.04	0.73±0.06	0.59±0.06	0.75±0.07	0.76±0.06
	frozen	imbalanced	half	0.78±0.05	0.75±0.04	0.34±0.04	0.63±0.03	0.72±0.03	0.74±0.03
		balanced	half	0.76±0.07	0.77±0.04	0.73±0.06	0.58±0.07	0.73±0.07	0.75±0.06
		imbalanced	full	0.80±0.04	0.75±0.04	0.33±0.07	0.63±0.03	0.72±0.03	0.74±0.01
		balanced	full	0.74±0.07	0.77±0.04	0.72±0.06	0.58±0.06	0.73±0.07	0.75±0.06

Table A11. Comparison of representations and classifiers performance for discriminating **ion channels from membrane proteins** on Sensitivity metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Dataset	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-1b	finetuned	imbalanced	half	100.00±1.00	99.00±2.00	98.00±2.00	98.00±2.00	100.00±1.00	100.00±1.00
		balanced	half	89.50±3.60	90.50±2.70	80.70±4.80	95.00±3.20	91.20±2.40	92.00±2.60
		imbalanced	full	99.00±2.00	99.00±2.00	98.00±2.00	98.00±2.00	100.00±1.00	100.00±1.00
		balanced	full	89.30±4.20	90.50±2.70	80.90±4.40	94.80±3.10	91.20±2.80	92.10±2.50
	frozen	imbalanced	half	82.00±7.00	81.00±6.00	36.00±4.00	82.00±6.00	84.00±5.00	84.00±5.00
		balanced	half	89.30±5.10	90.70±2.70	80.20±4.50	95.20±2.80	91.90±2.20	92.10±2.60
		imbalanced	full	82.00±7.00	81.00±6.00	36.00±5.00	82.00±6.00	84.00±5.00	85.00±5.00
		balanced	full	88.80±3.90	90.70±2.70	81.10±4.00	95.30±2.70	91.80±2.20	92.20±2.40
ESM-2	finetuned	imbalanced	half	97.00±3.00	93.00±4.00	83.00±2.00	85.00±6.00	93.00±3.00	93.00±3.00
		balanced	half	90.20±4.50	91.40±2.60	81.10±6.20	92.20±4.20	91.60±2.60	92.00±3.00
		imbalanced	full	96.00±3.00	94.00±3.00	85.00±3.00	86.00±4.00	94.00±4.00	94.00±4.00
		balanced	full	89.90±4.90	91.30±2.80	81.60±6.20	92.00±4.10	91.30±2.70	92.40±2.60
	frozen	imbalanced	half	81.00±8.00	83.00±6.00	28.00±6.00	71.00±6.00	84.00±6.00	85.00±6.00
		balanced	half	89.00±6.50	91.40±2.40	81.60±5.90	92.40±4.40	91.40±2.70	92.20±2.80
		imbalanced	full	81.00±8.00	83.00±6.00	29.00±6.00	71.00±6.00	84.00±6.00	85.00±7.00
		balanced	full	90.10±4.70	91.40±2.40	81.60±5.80	92.30±4.30	91.40±2.60	92.10±2.80
ESM-2_15B	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	frozen	imbalanced	half	85.00±6.00	87.00±6.00	17.00±3.00	78.00±7.00	85.00±5.00	84.00±5.00
		balanced	half	88.10±4.90	92.00±2.60	80.80±4.40	95.90±2.80	92.80±2.90	92.70±2.90
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ProtT5	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	frozen	imbalanced	half	84.00±6.00	79.00±6.00	31.00±7.00	73.00±9.00	79.00±5.00	80.00±4.00
		balanced	half	87.90±6.80	88.30±4.00	79.40±3.90	93.70±2.80	90.00±2.40	90.90±2.80
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ProtBERT-BFD	finetuned	imbalanced	half	76.00±2.00	80.00±3.00	76.00±4.00	75.00±5.00	78.00±6.00	80.00±7.00
		balanced	half	85.20±9.00	87.40±3.20	79.80±4.60	90.90±3.20	87.20±4.80	88.40±3.30
		imbalanced	full	70.00±3.00	74.00±5.00	72.00±5.00	70.00±7.00	74.00±5.00	73.00±5.00
		balanced	full	87.67±7.67	87.90±3.20	80.10±5.00	90.40±3.20	87.00±4.80	88.50±3.30
	frozen	imbalanced	half	71.00±9.00	67.00±3.00	13.00±2.00	53.00±6.00	64.00±5.00	70.00±6.00
		balanced	half	86.00±8.90	87.40±2.70	80.60±5.90	90.70±3.60	86.70±4.30	88.40±3.40
		imbalanced	full	75.00±8.00	67.00±3.00	13.00±5.00	53.00±6.00	64.00±5.00	69.00±4.00
		balanced	full	86.00±7.90	87.30±2.70	80.50±6.30	90.70±3.20	86.70±4.50	88.00±3.60
ProtBERT	finetuned	imbalanced	half	81.00±2.00	81.00±7.00	56.00±4.00	68.00±5.00	77.00±5.00	80.00±5.00
		balanced	half	85.80±7.00	87.20±3.40	78.00±3.80	89.50±3.80	87.00±3.80	87.50±3.70
		imbalanced	full	85.00±6.00	84.00±4.00	68.00±9.00	80.00±5.00	80.00±8.00	83.00±6.00
		balanced	full	82.90±8.60	87.20±3.40	77.50±4.40	89.10±3.70	87.00±3.50	87.70±3.70
	frozen	imbalanced	half	68.00±4.00	74.00±4.00	11.00±4.00	52.00±8.00	73.00±5.00	77.00±7.00
		balanced	half	83.90±8.70	87.40±3.10	78.40±4.20	89.10±3.70	86.40±4.30	87.40±3.70
		imbalanced	full	76.00±2.00	74.00±4.00	10.00±3.00	53.00±8.00	73.00±5.00	74.00±5.00
		balanced	full	84.11±8.89	87.50±3.50	78.70±5.00	89.30±3.60	86.50±4.40	87.30±3.90

Table A12. Comparison of representations and classifiers performance for discriminating **ion channels from membrane proteins** on Specificity metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Dataset	Precision	CNN	SVM	RF	kNN	LR	FFNN
ProtBERT-BFD	frozen	imbalanced	half	99.00±1.00	99.00±0.00	100.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	half	88.80±8.70	89.00±3.80	90.90±3.80	66.20±4.90	86.20±4.60	86.70±4.40
		imbalanced	full	99.00±1.00	99.00±0.00	100.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	full	86.50±10.30	89.10±3.90	90.70±3.90	66.40±5.10	86.30±4.70	87.10±4.30
	finetuned	imbalanced	half	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	half	89.20±8.70	89.70±4.10	91.50±3.70	67.70±4.70	87.30±4.90	87.80±4.20
		imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	88.11±8.56	89.30±4.30	91.50±3.60	67.80±4.70	87.60±4.60	87.40±4.40
ESM-1b	frozen	imbalanced	half	99.00±1.00	100.00±0.00	100.00±0.00	98.00±0.00	99.00±0.00	99.00±0.00
		balanced	half	91.10±6.60	94.40±2.30	95.60±2.70	68.40±4.60	91.80±3.50	91.40±3.20
		imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	98.00±0.00	99.00±0.00	99.00±0.00
		balanced	full	93.00±5.20	94.40±2.30	95.20±2.70	68.30±4.60	91.80±3.60	91.60±3.20
	finetuned	imbalanced	half	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	half	91.80±7.50	94.60±2.50	94.90±2.70	69.10±4.90	92.30±3.40	91.40±3.30
		imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	92.50±5.10	94.50±2.50	95.40±2.70	69.80±4.60	92.20±3.40	91.20±3.60
ProtBERT	frozen	imbalanced	half	100.00±0.00	99.00±1.00	100.00±0.00	98.00±1.00	99.00±0.00	99.00±0.00
		balanced	half	89.20±10.00	87.80±4.50	92.70±3.10	58.20±6.50	87.00±5.20	87.80±4.10
		imbalanced	full	99.00±1.00	99.00±1.00	100.00±0.00	98.00±1.00	99.00±0.00	99.00±0.00
		balanced	full	89.22±9.56	87.80±4.50	92.80±3.30	57.80±6.40	87.10±5.30	87.70±4.20
	finetuned	imbalanced	half	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	99.00±0.00
		balanced	half	88.00±9.50	88.20±5.00	93.50±3.20	59.10±6.10	87.50±4.30	87.50±4.10
		imbalanced	full	100.00±1.00	100.00±0.00	100.00±0.00	99.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	89.70±8.00	88.20±4.90	93.10±2.90	58.90±6.90	87.60±4.50	87.80±4.00
ProtT5	frozen	imbalanced	half	99.00±0.00	100.00±0.00	100.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	half	93.80±4.90	95.20±2.40	96.70±2.50	68.30±5.70	91.40±3.10	90.80±3.10
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ESM-2	frozen	imbalanced	half	100.00±0.00	100.00±0.00	100.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	half	92.50±7.80	93.30±2.60	93.70±2.90	70.20±4.90	91.90±3.40	91.80±3.00
		imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	full	93.50±6.50	93.30±2.50	93.70±3.30	70.00±4.60	92.00±3.40	92.00±3.00
	finetuned	imbalanced	half	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	half	93.10±6.30	93.10±2.60	93.90±3.50	70.50±4.60	91.90±3.10	91.80±3.40
		imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	93.70±5.40	93.10±2.60	93.80±2.70	70.40±4.50	92.00±3.20	91.70±3.20
ESM-2_15B	frozen	imbalanced	half	100.00±0.00	99.00±0.00	100.00±0.00	98.00±1.00	100.00±0.00	100.00±0.00
		balanced	half	95.50±4.80	93.90±2.30	97.40±1.60	45.50±5.10	94.10±2.80	94.20±2.90
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-

Appendix E.2. Ion transporters vs. other membrane proteins

Table A13. Comparison of representations and classifiers performance for discriminating **ion transporters from membrane proteins** on Accuracy metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Dataset	Precision	CNN	SVM	RF	kNN	LR	FFNN
ProtBERT-BFD	finetuned	imbalanced	full	99.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	full	87.70±2.90	86.70±3.00	82.80±4.00	80.60±3.30	86.40±3.10	86.30±3.00
		imbalanced	half	98.00±0.00	99.00±0.00	99.00±0.00	98.00±0.00	99.00±0.00	99.00±0.00
		balanced	half	87.60±2.80	86.40±2.90	82.60±3.90	80.10±3.20	86.20±3.20	86.30±2.90
	frozen	imbalanced	full	97.00±0.00	96.00±0.00	94.00±0.00	95.00±1.00	96.00±0.00	97.00±0.00
		balanced	full	87.40±3.10	86.30±2.70	82.50±3.90	79.90±3.40	86.30±3.20	86.40±2.90
		imbalanced	half	96.00±0.00	97.00±0.00	94.00±0.00	95.00±1.00	96.00±0.00	97.00±0.00
		balanced	half	87.50±2.90	86.40±2.60	82.10±4.40	79.70±3.30	86.40±3.30	86.30±2.90
ESM-2	finetuned	imbalanced	full	100.00±0.00	99.00±0.00	98.00±0.00	99.00±0.00	99.00±0.00	99.00±0.00
		balanced	full	91.11±1.78	89.60±2.20	85.80±2.70	80.60±3.70	89.80±1.90	89.90±2.80
		imbalanced	half	100.00±0.00	99.00±0.00	98.00±0.00	98.00±1.00	99.00±0.00	99.00±0.00
		balanced	half	91.30±2.00	89.40±2.20	85.60±2.90	80.60±3.50	89.70±1.90	89.70±2.80
	frozen	imbalanced	full	97.00±0.00	97.00±0.00	94.00±1.00	95.00±0.00	97.00±1.00	97.00±1.00
		balanced	full	91.20±1.70	89.30±2.40	85.70±2.50	80.80±3.40	89.60±2.00	89.70±3.00
		imbalanced	half	97.00±0.00	97.00±0.00	94.00±0.00	95.00±0.00	97.00±1.00	97.00±0.00
		balanced	half	91.10±1.90	89.40±2.30	85.70±2.60	80.70±3.40	89.50±2.00	89.70±2.70
ProtBERT	finetuned	imbalanced	full	99.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00
		balanced	full	88.20±1.90	86.90±2.30	82.80±3.10	77.90±3.30	87.50±2.70	87.50±2.30
		imbalanced	half	98.00±0.00	98.00±0.00	97.00±0.00	97.00±0.00	98.00±1.00	98.00±0.00
		balanced	half	88.20±2.10	86.90±2.20	82.90±3.40	78.00±3.30	87.10±2.10	87.60±2.30
	frozen	imbalanced	full	96.00±1.00	96.00±0.00	93.00±0.00	94.00±1.00	96.00±0.00	96.00±0.00
		balanced	full	88.10±2.00	86.50±2.60	82.30±3.50	77.50±3.20	87.10±2.40	87.20±2.80
		imbalanced	half	96.00±0.00	96.00±0.00	93.00±0.00	94.00±1.00	96.00±0.00	96.00±0.00
		balanced	half	88.20±1.70	86.50±2.50	82.30±3.80	77.30±2.90	87.10±2.30	87.20±2.80
ESM-1b	finetuned	imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	90.80±2.20	90.70±2.10	87.40±2.50	84.50±2.90	90.00±2.60	89.90±2.90
		imbalanced	half	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	half	90.80±1.90	90.80±2.10	87.60±2.40	84.50±2.80	90.10±2.60	90.10±3.00
	frozen	imbalanced	full	96.00±1.00	97.00±0.00	94.00±1.00	96.00±0.00	97.00±0.00	97.00±0.00
		balanced	full	90.90±2.10	90.40±2.10	87.00±2.70	84.20±2.90	89.90±2.70	90.00±2.40
		imbalanced	half	97.00±0.00	97.00±0.00	94.00±0.00	96.00±0.00	97.00±0.00	97.00±0.00
		balanced	half	90.56±2.67	90.40±2.20	86.90±3.10	84.20±2.80	89.90±2.60	90.00±2.80
ESM-2_15B	finetuned	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
	frozen	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	97.00±0.00	97.00±0.00	93.00±0.00	95.00±1.00	97.00±0.00	97.00±0.00
		balanced	half	90.80±2.10	90.70±2.70	86.00±2.70	74.00±3.30	91.00±2.80	90.50±2.90
ProtT5	finetuned	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
	frozen	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	97.00±0.00	97.00±0.00	94.00±0.00	96.00±0.00	97.00±0.00	97.00±0.00
		balanced	half	91.80±1.70	91.70±2.40	88.00±2.30	82.40±2.70	90.80±2.00	90.70±2.50

Table A14. Comparison of representations and classifiers performance for discriminating **ion transporters from membrane proteins** on MCC metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Dataset	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-1b	finetuned	imbalanced	full	0.99±0.01	1.00±0.00	0.99±0.01	0.99±0.01	1.00±0.00	1.00±0.01
		balanced	full	0.82±0.04	0.82±0.04	0.75±0.05	0.70±0.05	0.80±0.05	0.80±0.06
		imbalanced	half	0.99±0.01	0.99±0.01	0.98±0.01	0.99±0.01	1.00±0.00	1.00±0.00
		balanced	half	0.82±0.04	0.82±0.04	0.75±0.05	0.69±0.05	0.80±0.05	0.80±0.06
	frozen	imbalanced	full	0.74±0.04	0.77±0.04	0.42±0.10	0.74±0.02	0.77±0.03	0.79±0.03
		balanced	full	0.82±0.04	0.81±0.04	0.74±0.06	0.69±0.06	0.80±0.05	0.80±0.05
		imbalanced	half	0.77±0.03	0.77±0.04	0.45±0.03	0.74±0.02	0.78±0.04	0.79±0.03
		balanced	half	0.82±0.05	0.81±0.04	0.74±0.06	0.69±0.05	0.80±0.05	0.80±0.05
ESM-2	finetuned	imbalanced	full	0.98±0.01	0.95±0.01	0.86±0.04	0.90±0.03	0.95±0.02	0.94±0.02
		balanced	full	0.83±0.04	0.80±0.04	0.72±0.06	0.62±0.07	0.80±0.04	0.80±0.06
		imbalanced	half	0.97±0.01	0.94±0.03	0.88±0.04	0.87±0.03	0.93±0.03	0.93±0.03
		balanced	half	0.83±0.04	0.79±0.04	0.72±0.06	0.62±0.07	0.80±0.04	0.79±0.06
	frozen	imbalanced	full	0.77±0.02	0.77±0.04	0.44±0.07	0.65±0.03	0.75±0.04	0.76±0.04
		balanced	full	0.83±0.04	0.79±0.05	0.72±0.05	0.63±0.07	0.80±0.04	0.80±0.06
		imbalanced	half	0.74±0.04	0.77±0.04	0.43±0.07	0.65±0.03	0.74±0.05	0.76±0.03
		balanced	half	0.82±0.04	0.79±0.04	0.72±0.05	0.63±0.07	0.79±0.04	0.80±0.06
ProtBERT	finetuned	imbalanced	full	0.92±0.03	0.89±0.04	0.81±0.03	0.86±0.04	0.89±0.03	0.88±0.03
		balanced	full	0.76±0.04	0.74±0.04	0.66±0.06	0.57±0.07	0.75±0.05	0.75±0.05
		imbalanced	half	0.88±0.02	0.87±0.03	0.76±0.02	0.81±0.04	0.87±0.04	0.87±0.03
		balanced	half	0.77±0.04	0.74±0.05	0.66±0.07	0.57±0.07	0.74±0.04	0.75±0.05
	frozen	imbalanced	full	0.64±0.14	0.72±0.04	0.22±0.07	0.51±0.06	0.68±0.03	0.71±0.04
		balanced	full	0.76±0.04	0.73±0.05	0.65±0.07	0.56±0.06	0.74±0.05	0.75±0.06
		imbalanced	half	0.69±0.03	0.72±0.04	0.23±0.05	0.52±0.05	0.68±0.03	0.71±0.03
		balanced	half	0.77±0.04	0.73±0.05	0.65±0.07	0.56±0.06	0.74±0.05	0.75±0.06
ProtBERT-BFD	finetuned	imbalanced	full	0.90±0.03	0.92±0.03	0.92±0.03	0.92±0.02	0.92±0.02	0.92±0.02
		balanced	full	0.76±0.06	0.73±0.06	0.66±0.08	0.61±0.07	0.73±0.06	0.73±0.06
		imbalanced	half	0.89±0.03	0.90±0.02	0.90±0.01	0.88±0.01	0.90±0.01	0.90±0.01
		balanced	half	0.75±0.06	0.73±0.05	0.66±0.08	0.61±0.07	0.73±0.06	0.73±0.06
	frozen	imbalanced	full	0.74±0.03	0.74±0.04	0.41±0.03	0.62±0.06	0.71±0.02	0.75±0.01
		balanced	full	0.75±0.06	0.73±0.06	0.65±0.08	0.60±0.07	0.72±0.06	0.73±0.06
		imbalanced	half	0.71±0.05	0.74±0.04	0.43±0.05	0.62±0.06	0.72±0.01	0.75±0.02
		balanced	half	0.75±0.06	0.73±0.05	0.65±0.09	0.60±0.07	0.73±0.07	0.73±0.06
ProtT5	finetuned	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
	frozen	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	0.76±0.03	0.80±0.01	0.42±0.05	0.73±0.03	0.79±0.02	0.79±0.04
		balanced	half	0.83±0.03	0.83±0.05	0.76±0.05	0.66±0.06	0.82±0.04	0.82±0.05
ESM-2_15B	finetuned	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
	frozen	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	0.79±0.03	0.78±0.03	0.27±0.03	0.66±0.06	0.80±0.03	0.79±0.02
		balanced	half	0.82±0.04	0.82±0.05	0.72±0.06	0.53±0.06	0.81±0.06	0.81±0.06

Table A15. Comparison of representations and classifiers performance for discriminating **ion transporters from membrane proteins** on Sensitivity metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Dataset	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-1b	finetuned	imbalanced	half	99.00±1.00	99.00±1.00	98.00±2.00	99.00±1.00	99.00±1.00	100.00±1.00
		balanced	half	88.50±3.70	88.20±3.50	85.50±3.10	87.60±4.00	89.30±2.90	90.00±2.90
		imbalanced	full	99.00±1.00	100.00±1.00	99.00±1.00	99.00±1.00	100.00±0.00	99.00±1.00
		balanced	full	88.50±2.90	88.00±3.50	85.70±4.00	87.60±4.20	89.40±3.10	89.90±3.10
	frozen	imbalanced	half	72.00±6.00	76.00±7.00	23.00±4.00	69.00±4.00	74.00±5.00	77.00±5.00
		balanced	half	89.22±3.67	88.20±3.60	84.40±4.40	87.40±3.90	89.20±2.80	90.10±3.00
		imbalanced	full	72.00±11.00	76.00±7.00	21.00±8.00	69.00±4.00	74.00±5.00	76.00±5.00
		balanced	full	89.00±3.40	88.20±3.50	84.70±4.30	87.40±3.90	89.10±2.80	90.10±3.00
ESM-2	finetuned	imbalanced	half	98.00±2.00	95.00±3.00	81.00±6.00	89.00±3.00	93.00±4.00	93.00±3.00
		balanced	half	89.00±3.20	89.10±2.90	82.50±4.30	88.60±3.90	89.30±3.20	90.20±3.40
		imbalanced	full	98.00±2.00	95.00±3.00	77.00±7.00	89.00±4.00	94.00±2.00	93.00±3.00
		balanced	full	89.00±2.67	89.20±2.70	82.60±4.40	88.80±4.30	89.30±3.10	90.40±3.30
	frozen	imbalanced	half	64.00±7.00	72.00±7.00	22.00±6.00	61.00±7.00	69.00±6.00	72.00±7.00
		balanced	half	89.30±2.90	88.90±2.80	82.70±3.70	88.80±3.90	88.90±2.90	89.90±3.60
		imbalanced	full	71.00±5.00	72.00±7.00	24.00±7.00	61.00±7.00	69.00±6.00	72.00±8.00
		balanced	full	89.80±2.50	88.80±2.90	82.50±2.80	88.80±3.80	89.00±2.90	89.60±3.40
ESM-2_15B	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	frozen	imbalanced	half	71.00±8.00	77.00±5.00	8.00±2.00	70.00±7.00	77.00±4.00	75.00±5.00
		balanced	half	87.80±2.80	90.00±3.40	83.40±4.10	94.80±2.60	89.50±3.60	89.90±3.60
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ProtT5	finetuned	imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
	frozen	imbalanced	half	69.00±8.00	74.00±4.00	19.00±5.00	75.00±6.00	76.00±5.00	76.00±6.00
		balanced	half	91.60±2.30	90.30±3.40	85.40±3.50	92.30±3.60	91.00±2.80	91.00±3.50
		imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
ProtBERT-BFD	finetuned	imbalanced	half	87.00±4.00	87.00±5.00	86.00±2.00	86.00±4.00	88.00±3.00	87.00±2.00
		balanced	half	86.10±3.80	84.70±4.30	78.70±6.80	85.10±4.40	85.90±4.10	85.70±3.80
		imbalanced	full	85.00±7.00	91.00±5.00	91.00±5.00	91.00±4.00	90.00±5.00	91.00±4.00
		balanced	full	87.00±3.50	85.40±4.50	79.40±6.80	85.20±4.20	86.20±3.90	86.00±4.30
	frozen	imbalanced	half	60.00±11.00	71.00±10.00	21.00±5.00	61.00±5.00	67.00±3.00	69.00±4.00
		balanced	half	86.00±3.20	85.30±3.80	78.60±6.40	84.70±4.40	86.00±4.00	85.90±4.10
		imbalanced	full	64.00±8.00	70.00±11.00	20.00±3.00	61.00±5.00	66.00±4.00	69.00±5.00
		balanced	full	85.90±3.30	85.40±4.00	78.90±5.60	84.70±4.70	86.10±3.80	86.00±4.10
ProtBERT	finetuned	imbalanced	half	83.00±5.00	85.00±5.00	61.00±3.00	75.00±6.00	82.00±5.00	82.00±6.00
		balanced	half	88.50±2.90	85.80±3.50	83.10±4.90	86.90±4.30	86.60±3.50	87.70±3.60
		imbalanced	full	88.00±5.00	85.00±5.00	68.00±5.00	78.00±4.00	85.00±4.00	84.00±5.00
		balanced	full	87.90±2.50	86.20±3.10	82.40±4.30	86.70±3.40	86.90±3.90	87.70±3.30
	frozen	imbalanced	half	59.00±8.00	68.00±6.00	6.00±2.00	45.00±4.00	63.00±6.00	68.00±6.00
		balanced	half	88.00±2.70	85.60±3.80	82.40±5.20	86.20±3.70	86.30±3.70	87.00±3.70
		imbalanced	full	48.00±19.00	68.00±6.00	6.00±4.00	45.00±5.00	63.00±6.00	69.00±6.00
		balanced	full	87.80±2.80	85.60±3.70	82.60±5.30	86.40±3.40	86.40±3.80	86.80±3.90

Table A16. Comparison of representations and classifiers performance for discriminating **ion transporters from membrane proteins** on Specificity metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Dataset	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-2	frozen	imbalanced	full	99.00±1.00	99.00±0.00	100.00±0.00	98.00±0.00	99.00±0.00	99.00±1.00
		balanced	full	92.90±3.50	90.20±4.20	89.40±4.30	72.80±4.60	90.30±4.30	89.60±5.40
		imbalanced	half	99.00±0.00	99.00±0.00	100.00±0.00	98.00±0.00	99.00±0.00	99.00±1.00
		balanced	half	92.80±4.50	90.10±4.10	88.90±4.10	72.80±4.70	90.10±4.20	89.50±5.00
	finetuned	imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	99.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	93.33±3.33	90.20±4.20	88.90±3.90	72.40±4.80	90.10±4.30	89.40±5.10
		imbalanced	half	100.00±0.00	99.00±0.00	100.00±0.00	99.00±1.00	100.00±0.00	100.00±0.00
		balanced	half	93.80±3.60	90.30±4.20	88.80±4.10	72.50±5.00	90.10±4.00	89.30±5.10
ProtBERT-BFD	frozen	imbalanced	full	99.00±0.00	99.00±1.00	100.00±0.00	98.00±1.00	99.00±0.00	99.00±0.00
		balanced	full	89.10±4.60	87.20±4.00	85.80±4.90	74.80±4.30	86.30±4.90	86.80±4.70
		imbalanced	half	99.00±1.00	99.00±1.00	100.00±0.00	98.00±1.00	99.00±0.00	99.00±0.00
		balanced	half	88.80±4.90	87.30±3.80	85.80±4.20	74.70±4.20	86.50±5.10	86.50±4.90
	finetuned	imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	89.00±4.50	88.00±3.80	86.20±4.40	75.70±4.30	86.70±4.60	86.90±4.80
		imbalanced	half	99.00±1.00	100.00±0.00	100.00±0.00	99.00±0.00	100.00±0.00	100.00±0.00
		balanced	half	89.20±4.40	87.90±3.70	86.80±3.90	75.30±4.00	86.30±4.60	87.00±4.40
ESM-1b	frozen	imbalanced	full	98.00±1.00	99.00±0.00	100.00±0.00	99.00±1.00	99.00±0.00	99.00±0.00
		balanced	full	93.00±3.80	92.90±2.80	89.20±4.60	81.00±4.70	90.70±4.20	89.80±3.80
		imbalanced	half	99.00±1.00	99.00±0.00	100.00±0.00	99.00±1.00	99.00±0.00	99.00±0.00
		balanced	half	92.22±4.67	92.90±2.60	89.30±4.50	81.10±4.70	90.70±4.20	89.70±4.40
	finetuned	imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	93.40±3.40	93.40±2.70	89.20±3.90	81.60±4.70	90.50±4.30	90.00±5.00
		imbalanced	half	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	half	92.90±3.50	93.30±2.40	89.50±3.80	81.30±4.40	90.60±4.30	90.00±4.80
ESM-2_15B	frozen	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	99.00±0.00	99.00±0.00	100.00±0.00	97.00±1.00	99.00±0.00	99.00±1.00
		balanced	half	93.90±3.10	91.30±4.10	88.80±4.10	53.00±6.30	91.70±3.90	91.10±4.40
	finetuned	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
ProtT5	frozen	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	99.00±1.00	99.00±0.00	100.00±0.00	98.00±0.00	99.00±0.00	99.00±0.00
		balanced	half	91.90±3.40	92.90±3.60	90.60±3.60	72.50±4.40	90.80±3.40	90.30±3.90
	finetuned	imbalanced	full	-	-	-	-	-	-
		balanced	full	-	-	-	-	-	-
		imbalanced	half	-	-	-	-	-	-
		balanced	half	-	-	-	-	-	-
ProtBERT	frozen	imbalanced	full	100.00±0.00	99.00±0.00	100.00±0.00	98.00±0.00	98.00±0.00	98.00±0.00
		balanced	full	88.40±4.00	87.40±4.70	82.40±4.40	68.10±5.40	87.80±4.90	87.60±4.80
		imbalanced	half	99.00±0.00	99.00±0.00	100.00±0.00	98.00±0.00	98.00±0.00	99.00±0.00
		balanced	half	88.30±4.00	87.50±4.60	82.60±5.20	68.20±5.30	87.90±5.10	87.40±4.80
	finetuned	imbalanced	full	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
		balanced	full	88.40±3.90	87.70±4.70	83.30±4.80	69.10±5.40	88.10±4.60	87.60±4.70
		imbalanced	half	100.00±0.00	99.00±0.00	100.00±0.00	99.00±0.00	100.00±0.00	100.00±0.00
		balanced	half	88.00±4.10	88.00±4.30	83.00±4.40	69.50±4.90	87.70±4.60	87.60±4.50

Appendix E.3. Ion channels vs. ion transporters

Table A17. Comparison of representations and classifiers performance for discriminating **ion channels from ion transporters** on Accuracy metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-1b	finetuned	half	93.00±1.00	93.00±2.00	92.00±2.00	89.00±3.00	94.00±2.00	94.00±3.00
		full	91.00±5.00	93.00±2.00	91.00±3.00	89.00±3.00	94.00±2.00	94.00±3.00
	frozen	half	93.00±2.00	94.00±2.00	92.00±2.00	90.00±2.00	94.00±2.00	93.00±2.00
		full	93.00±1.00	94.00±2.00	92.00±2.00	90.00±2.00	94.00±2.00	93.00±2.00
ESM-2	finetuned	half	93.00±1.00	93.00±2.00	90.00±1.00	87.00±5.00	92.00±1.00	94.00±2.00
		full	94.00±1.00	93.00±2.00	90.00±2.00	87.00±4.00	92.00±1.00	93.00±3.00
	frozen	half	92.00±2.00	93.00±2.00	89.00±2.00	87.00±5.00	92.00±1.00	94.00±2.00
		full	94.00±1.00	93.00±2.00	89.00±2.00	87.00±5.00	92.00±1.00	94.00±2.00
ESM-2_15B	finetuned	half	-	-	-	-	-	-
		full	-	-	-	-	-	-
	frozen	half	94.00±1.00	94.00±1.00	90.00±1.00	89.00±4.00	94.00±1.00	93.00±2.00
		full	-	-	-	-	-	-
ProtT5	finetuned	half	-	-	-	-	-	-
		full	-	-	-	-	-	-
	frozen	half	93.00±1.00	93.00±2.00	89.00±2.00	90.00±2.00	93.00±2.00	93.00±2.00
		full	-	-	-	-	-	-
ProtBERT	finetuned	half	93.00±1.00	92.00±0.00	89.00±2.00	82.00±4.00	90.00±1.00	91.00±1.00
		full	93.00±0.00	92.00±0.00	89.00±2.00	82.00±4.00	90.00±1.00	91.00±1.00
	frozen	half	92.00±0.00	92.00±1.00	88.00±3.00	82.00±3.00	90.00±2.00	91.00±2.00
		full	92.00±1.00	92.00±1.00	88.00±3.00	82.00±3.00	90.00±2.00	91.00±2.00
ProtBERT-BFD	finetuned	half	92.00±3.00	90.00±2.00	87.00±3.00	86.00±2.00	89.00±2.00	90.00±2.00
		full	92.00±3.00	90.00±3.00	88.00±2.00	85.00±2.00	88.00±2.00	90.00±2.00
	frozen	half	92.00±3.00	90.00±2.00	87.00±3.00	86.00±2.00	87.00±2.00	89.00±4.00
		full	92.00±3.00	90.00±2.00	87.00±3.00	86.00±3.00	87.00±2.00	89.00±2.00

Table A18. Comparison of representations and classifiers performance for discriminating **ion channels from ion transporters** on MCC metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-2	finetuned	full	0.89±0.03	0.87±0.04	0.80±0.03	0.74±0.08	0.84±0.03	0.86±0.05
		half	0.86±0.03	0.87±0.04	0.80±0.01	0.75±0.09	0.84±0.01	0.87±0.05
	frozen	full	0.88±0.02	0.87±0.04	0.78±0.03	0.74±0.09	0.84±0.02	0.88±0.05
		half	0.85±0.04	0.87±0.04	0.77±0.03	0.74±0.09	0.85±0.02	0.87±0.04
ESM-2_15B	finetuned	full	-	-	-	-	-	-
		half	-	-	-	-	-	-
	frozen	full	-	-	-	-	-	-
		half	0.89±0.02	0.88±0.03	0.80±0.03	0.79±0.07	0.87±0.03	0.87±0.03
ESM-1b	finetuned	full	0.83±0.08	0.86±0.05	0.83±0.06	0.79±0.05	0.88±0.05	0.87±0.06
		half	0.87±0.02	0.87±0.05	0.84±0.03	0.80±0.05	0.88±0.04	0.87±0.06
	frozen	full	0.85±0.02	0.87±0.04	0.83±0.05	0.80±0.05	0.88±0.03	0.87±0.04
		half	0.87±0.03	0.87±0.04	0.84±0.03	0.80±0.05	0.88±0.03	0.87±0.04
ProtT5	finetuned	full	-	-	-	-	-	-
		half	-	-	-	-	-	-
	frozen	full	-	-	-	-	-	-
		half	0.85±0.03	0.86±0.04	0.79±0.05	0.81±0.03	0.86±0.03	0.85±0.04
ProtBERT	finetuned	full	0.86±0.01	0.84±0.01	0.78±0.04	0.66±0.08	0.80±0.02	0.81±0.02
		half	0.86±0.02	0.84±0.01	0.78±0.05	0.66±0.08	0.80±0.02	0.81±0.02
	frozen	full	0.85±0.02	0.84±0.02	0.77±0.06	0.65±0.06	0.81±0.03	0.82±0.04
		half	0.84±0.00	0.84±0.02	0.77±0.05	0.65±0.06	0.80±0.03	0.82±0.04
ProtBERT-BFD	finetuned	full	0.84±0.06	0.81±0.05	0.76±0.05	0.71±0.04	0.76±0.04	0.81±0.05
		half	0.84±0.06	0.81±0.04	0.75±0.06	0.71±0.05	0.77±0.03	0.81±0.04
	frozen	full	0.84±0.07	0.81±0.04	0.75±0.05	0.72±0.05	0.74±0.05	0.79±0.05
		half	0.84±0.06	0.81±0.04	0.75±0.06	0.72±0.05	0.75±0.04	0.78±0.08

Table A19. Comparison of representations and classifiers performance for discriminating **ion channels from ion transporters** on Sensitivity metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-1b	frozen	full	91.00±1.00	93.00±2.00	89.00±6.00	95.00±3.00	93.00±3.00	95.00±2.00
		half	93.00±1.00	93.00±2.00	90.00±6.00	95.00±3.00	93.00±3.00	95.00±2.00
	finetuned	full	88.00±13.00	94.00±3.00	88.00±5.00	95.00±3.00	95.00±2.00	94.00±3.00
		half	93.00±2.00	94.00±3.00	88.00±6.00	95.00±3.00	95.00±2.00	94.00±3.00
ESM-2	frozen	full	93.00±2.00	93.00±2.00	85.00±4.00	90.00±7.00	92.00±3.00	93.00±3.00
		half	93.00±3.00	93.00±2.00	85.00±7.00	90.00±7.00	93.00±3.00	93.00±3.00
	finetuned	full	92.00±2.00	93.00±2.00	87.00±4.00	91.00±6.00	92.00±2.00	93.00±3.00
		half	93.00±3.00	93.00±2.00	87.00±5.00	91.00±6.00	90.00±2.00	94.00±4.00
ESM-2_15B	frozen	full	-	-	-	-	-	-
		half	94.00±2.00	92.00±2.00	85.00±5.00	92.00±3.00	93.00±2.00	93.00±2.00
	finetuned	full	-	-	-	-	-	-
		half	-	-	-	-	-	-
ProtT5	frozen	full	-	-	-	-	-	-
		half	91.00±2.00	90.00±4.00	86.00±4.00	94.00±1.00	92.00±2.00	93.00±3.00
	finetuned	full	-	-	-	-	-	-
		half	-	-	-	-	-	-
ProtBERT-BFD	frozen	full	92.00±4.00	88.00±8.00	85.00±7.00	85.00±4.00	88.00±5.00	90.00±5.00
		half	91.00±3.00	88.00±8.00	86.00±8.00	85.00±3.00	88.00±5.00	89.00±6.00
	finetuned	full	91.00±3.00	88.00±7.00	87.00±7.00	87.00±3.00	88.00±4.00	90.00±5.00
		half	91.00±2.00	90.00±6.00	86.00±8.00	86.00±4.00	88.00±4.00	92.00±4.00
ProtBERT	frozen	full	91.00±5.00	92.00±3.00	85.00±7.00	85.00±6.00	89.00±4.00	90.00±4.00
		half	89.00±4.00	92.00±3.00	85.00±7.00	85.00±6.00	89.00±4.00	90.00±4.00
	finetuned	full	91.00±2.00	92.00±3.00	84.00±7.00	88.00±5.00	90.00±3.00	90.00±4.00
		half	92.00±3.00	92.00±3.00	85.00±6.00	88.00±5.00	90.00±3.00	90.00±4.00

Table A20. Comparison of representations and classifiers performance for discriminating **ion channels from ion transporters** on Specificity metric as m±d, where m is the mean and d is the standard deviation across the five runs of the cross-validation. The symbol "-" indicates that results are unavailable due to the extensive computational resources needed for fine-tuning large PLMs, which could not be accommodated by our limited resources.

Representer	Representation	Precision	CNN	SVM	RF	kNN	LR	FFNN
ESM-2	finetuned	full	96.00±2.00	94.00±4.00	92.00±5.00	83.00±7.00	92.00±4.00	94.00±4.00
		half	93.00±3.00	94.00±4.00	92.00±5.00	84.00±7.00	94.00±3.00	94.00±2.00
	frozen	full	96.00±2.00	94.00±3.00	91.00±5.00	84.00±7.00	93.00±3.00	95.00±3.00
		half	92.00±6.00	94.00±3.00	92.00±5.00	84.00±7.00	92.00±4.00	95.00±3.00
ESM-1b	finetuned	full	93.00±6.00	92.00±5.00	95.00±6.00	84.00±5.00	93.00±4.00	94.00±4.00
		half	94.00±3.00	93.00±4.00	95.00±3.00	85.00±5.00	93.00±3.00	94.00±4.00
	frozen	full	94.00±2.00	94.00±3.00	94.00±5.00	85.00±5.00	94.00±5.00	92.00±4.00
		half	94.00±3.00	94.00±3.00	94.00±5.00	85.00±5.00	94.00±5.00	92.00±4.00
ESM-2_15B	finetuned	full	-	-	-	-	-	-
		half	-	-	-	-	-	-
	frozen	full	-	-	-	-	-	-
		half	94.00±3.00	95.00±3.00	94.00±5.00	86.00±7.00	94.00±2.00	94.00±3.00
ProtT5	finetuned	full	-	-	-	-	-	-
		half	-	-	-	-	-	-
	frozen	full	-	-	-	-	-	-
		half	94.00±2.00	95.00±3.00	93.00±6.00	87.00±4.00	94.00±3.00	92.00±5.00
ProtBERT	finetuned	full	94.00±3.00	92.00±3.00	92.00±7.00	78.00±7.00	90.00±3.00	91.00±3.00
		half	94.00±2.00	92.00±3.00	92.00±7.00	78.00±7.00	90.00±3.00	91.00±3.00
	frozen	full	93.00±4.00	91.00±4.00	91.00±7.00	80.00±7.00	91.00±2.00	91.00±4.00
		half	95.00±3.00	91.00±4.00	91.00±8.00	80.00±7.00	91.00±3.00	92.00±3.00
ProtBERT-BFD	finetuned	full	93.00±4.00	93.00±3.00	88.00±7.00	84.00±4.00	88.00±4.00	90.00±6.00
		half	93.00±4.00	90.00±4.00	88.00±8.00	85.00±3.00	89.00±3.00	89.00±5.00
	frozen	full	93.00±4.00	93.00±4.00	89.00±8.00	86.00±3.00	87.00±4.00	89.00±5.00
		half	93.00±5.00	93.00±4.00	88.00±7.00	86.00±3.00	87.00±4.00	89.00±6.00

References

1. Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; Rost, B. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **2019**, *20*, 723.
2. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C.L.; Ma, J.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2016239118. Publisher: Proceedings of the National Academy of Sciences.
3. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, pp. 1–1.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, [1706.03762].
5. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach **2019**.
6. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text Transformer, 2020. arXiv:1910.10683 [cs, stat].
7. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); Association for Computational Linguistics: New Orleans, Louisiana, 2018; pp. 2227–2237.
8. Unsal, S.; Atas, H.; Albayrak, M.; Turhan, K.; Acar, A.C.; Doğan, T. Learning functional properties of proteins with language models. *Nature Machine Intelligence* **2022**, *4*, 227–245.
9. Askr, H.; Elgeldawi, E.; Aboul Ella, H.; Elshaier, Y.A.M.M.; Gomaa, M.M.; Hassanien, A.E. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review* **2022**.
10. Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **2020**, *117*, 1496–1503. Publisher: Proceedings of the National Academy of Sciences.
11. Asgari, E.; Mofrad, M.R.K. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE* **2015**, *10*, e0141287.
12. Rao, R.M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A. MSA transformer. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021, pp. 8844–8856. ISSN: 2640-3498.
13. Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; Song, Y. Evaluating protein transfer learning with TAPE. In Proceedings of the Advances in Neural Information Processing Systems; Wallach, H.; Larochelle, H.; Beygelzimer, A.; Alché-Buc, F.d.; Fox, E.; Garnett, R., Eds. Curran Associates, Inc., 2019, Vol. 32.
14. Unsal, S.; Ataş, H.; Albayrak, M.; Turhan, K.; Acar, A.C.; Doğan, T. Evaluation of methods for protein representation learning: A quantitative analysis. Technical report, bioRxiv, 2020.
15. Kotsiliti, E. De novo protein design with a language model. *Nature Biotechnology* **2022**, *40*, 1433–1433.
16. Ghazikhani, H.; Butler, G. A study on the application of protein language models in the analysis of membrane proteins. In Proceedings of the Distributed Computing and Artificial Intelligence, Special Sessions, 19th International Conference; Machado, J.M.; Chamoso, P.; Hernández, G.; Bocewicz, G.; Loukanova, R.; Jove, E.; del Rey, A.M.; Ricca, M., Eds.; Springer International Publishing: Cham, 2023; Lecture Notes in Networks and Systems, pp. 147–152.
17. Ghazikhani, H.; Butler, G. TooT-BERT-M: Discriminating membrane proteins from non-membrane proteins using a BERT representation of protein primary sequences. In Proceedings of the 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2022, pp. 1–8.
18. Taju, S.W.; Ou, Y.Y. DeepIon: Deep learning approach for classifying ion transporters and ion channels from membrane proteins. *Journal of Computational Chemistry* **2019**, *40*, 1521–1529.
19. Hille, B. *Ionic Channels of Excitable Membranes*, 3 ed.; Vol. 21, Springer, 2001.
20. Nogueira, J.J.; Corry, B. Ion Channel Permeation and Selectivity. In *The Oxford Handbook of Neuronal Ion Channels*; Bhattacharjee, A., Ed.; Oxford University Press, 2019.
21. Restrepo-Angulo, I.; De Vizcaya-Ruiz, A.; Camacho, J. Ion channels in toxicology. *Journal of Applied Toxicology* **2010**, *30*, 497–512.
22. Nguyen, T.T.D.; Ho, Q.T.; Tarn, Y.C.; Ou, Y.Y. MFPS_CNN: Multi-filter pattern scanning from position-specific scoring matrix with convolutional neural network for efficient prediction of ion transporters. *Molecular Informatics* **2022**, p. e2100271.
23. Ghazikhani, H.; Butler, G. TooT-BERT-C: A study on discriminating ion channels from membrane proteins based on the primary sequence's contextual representation from BERT models. In Proceedings of the Proceedings of the 9th International Conference on Bioinformatics Research and Applications; Association for Computing Machinery: Berlin, Germany, 2023; ICBRA '22, pp. 23–29.
24. Eisenberg, B. From structure to function in open ionic channels. *Journal of Membrane Biology* **1999**, *171*, 1–24.
25. Kulbacka, J.; Choromańska, A.; Rossowska, J.; Wężgowiec, J.; Saczko, J.; Rols, M.P. Cell Membrane Transport Mechanisms: Ion Channels and Electrical Properties of Cell Membranes. In *Transport Across Natural and Modified Biological Membranes and its*

Implications in Physiology and Therapy; Kulbacka, J.; Satkauskas, S., Eds.; Advances in Anatomy, Embryology and Cell Biology, Springer International Publishing: Cham, 2017; pp. 39–58.
26. Clare, J.J. Targeting ion channels for drug discovery. *Discovery Medicine* **2010**, *9*, 253–260.
27. Picci, G.; Marchesan, S.; Caltagirone, C. Ion channels and transporters as therapeutic agents: From biomolecules to supramolecular medicinal chemistry. *Biomedicines* **2022**, *10*, 885.
28. Ashrafuzzaman, M. Artificial intelligence, machine learning and deep learning in ion channel bioinformatics. *Membranes* **2021**, *11*, 672.
29. Menke, J.; Maskri, S.; Koch, O. Computational ion channel research: From the application of artificial intelligence to molecular dynamics simulations. *Cellular Physiology and Biochemistry: International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology* **2021**, *55*, 14–45.
30. Ghazikhani, H.; Butler, G. TooT-BERT-T: A BERT Approach on Discriminating Transport Proteins from Non-transport Proteins. In Proceedings of the Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022); Fdez-Riverola, F.; Rocha, M.; Mohamad, M.S.; Caraiman, S.; Gil-González, A.B., Eds.; Springer International Publishing: Cham, 2023; Lecture Notes in Networks and Systems, pp. 1–11.
31. Liu, J.; Jiang, T.; Lu, Y.; Wu, H. Drug-target interaction prediction based on transformer. In Proceedings of the Intelligent Computing Theories and Application; Huang, D.S.; Jo, K.H.; Jing, J.; Premaratne, P.; Bevilacqua, V.; Hussain, A., Eds.; Springer International Publishing: Cham, 2022; Lecture Notes in Computer Science, pp. 302–309.
32. Zhao, Y.W.; Su, Z.D.; Yang, W.; Lin, H.; Chen, W.; Tang, H. IonchanPred 2.0: A tool to predict ion channels and their types. *International Journal of Molecular Sciences* **2017**, *18*, 1838. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.
33. Gao, J.; Cui, W.; Sheng, Y.; Ruan, J.; Kurgan, L. PSIONplus: Accurate sequence-based predictor of ion channels and their types. *PLOS ONE* **2016**, *11*, e0152964. Publisher: Public Library of Science.
34. Gao, J.; Wei, H.; Cano, A.; Kurgan, L. PSIONplusm server for accurate multi-label prediction of ion channels and their types. *Biomolecules* **2020**, *10*, 876. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.
35. Lin, H.; Chen, W. Briefing in Application of Machine Learning Methods in Ion Channel Prediction. *The Scientific World Journal* **2015**, *2015*, e945927. Publisher: Hindawi.
36. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130. Publisher: American Association for the Advancement of Science.
37. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Research* **2004**, *32*, D115–D119.
38. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology* **1990**, *215*, 403–410.
39. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]* **2019**.
40. Suzek, B.E.; Wang, Y.; Huang, H.; McGarvey, P.B.; Wu, C.H.; the UniProt Consortium. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015**, *31*, 926–932.
41. Steinegger, M.; Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications* **2018**, *9*, 2542.
42. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2020**, [1910.03771].
43. Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.
44. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **1967**, *13*, 21–27.
45. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1998**, *20*, 832–844.
46. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, *61*, 85–117.
47. Tolles, J.; Meurer, W.J. Logistic regression: relating patient characteristics to outcomes. *JAMA* **2016**, *316*, 533–534.
48. Kramer, O. Scikit-Learn. In *Machine Learning for Evolution Strategies*; Kramer, O., Ed.; Springer International Publishing: Cham, 2016; pp. 45–53.
49. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Computation* **1989**, *1*, 541–551. Conference Name: Neural Computation.
50. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 2019, Vol. 32.
51. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Association for Computing Machinery: New York, NY, USA, 2019; KDD '19, pp. 2623–2631.
52. Chicco, D.; Jurman, G. The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, *21*, 6.
53. Mowery, B.D. The paired t-test. *Pediatric Nursing* **2011**, *37*, 320–322. Publisher: Jannetti Publications, Inc.
54. Sthle, L.; Wold, S. Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems* **1989**, *6*, 259–272.

55. Elnaggar, A.; Essam, H.; Salah-Eldin, W.; Moustafa, W.; Elkerdawy, M.; Rochereau, C.; Rost, B. Ankh: optimized protein language model unlocks general-purpose modelling, 2023.
56. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **2014**, *15*, 1929–1958.