A cautionary tale on organelle proteome prediction algorithms: limits and opportunities

Sven B. Gould, Jonas Magiera, Carolina García García, Parth K. Raval

Institute for Molecular Evolution, Heinrich–Heine–University Düsseldorf, 40225 Düsseldorf, Germany

Abstract

Mitochondria and plastids import thousands of proteins. Their experimental localisation remains a frequent task, but can be resource-intensive or even impossible especially for species that are genetically not accessible. Hence, hundreds of studies make use of (machine learning) algorithms that predict a sub-cellular localisation based on a protein's sequence. Their reliability across evolutionary diverse species is unknown. Here, we evaluate the performance of three commonly used algorithms (TargetP, Localizer and WoLFPSORT) for four photosynthetic eukaryotes, for which experimental plastid and mitochondrial proteome data is available. The match between algorithm-based predictions and experimental data ranges from 75% to as low as 2%, with up to thousands of false positives being predicted. Results depend on the algorithm used and the evolutionary distance between the training and query species. Specificity, sensitivity and precision analysis underscore severe limitations outside the training species and especially for plant mitochondria, for which the performance borders on random sampling. The results highlight current issues associated with prediction algorithms and present an opportunity for the next generation of protein localisation prediction tools that should train neural networks on an evolutionary more diverse set of organelle proteins for optimizing their performance and reliability.

Keywords: protein targeting, prediction algorithms, machine learning, mitochondria and plastid proteomes

Correspondence: gould@hhu.de; raval@hhu.de

Introduction

A plant encodes 20-30,000 proteins on average, of which many thousand are targeted to intracellular membrane bound compartments after or during translation[1–3]. The compartments owe their origins to bacterial ancestors directly or indirectly[4–12]. Mitochondria and plastids are of endosymbiotic origin and have transferred a majority of their coding capacity to the nuclear genome in the course of their transition from bacterium to organelle[13–15] . As a consequence, the vast majority of their proteins are translated in the cytosol and need to be imported. Protein translocation-related components of mitochondria such as TOM40, VDAC, TIM22, TIM23-PAM, OXA, SAM, HSP70, or the mitochondrial pre-sequence protease are likely of alphaproteobacterial origin[16–22], while many components of the plastid import machinery such as TOC75, OEP80, TIC20, the TAT pathway and also several signal processing peptidases are of cyanobacterial origin[23–34] . Despite their evolutionary independent roots, the import machineries of mitochondria and plastids are united by principles of how they recognize the vast majority of their cargo.

Cytosolically-translated proteins destined for the mitochondrial matrix or the plastid stroma, thousands in sum, carry N-terminal targeting sequences (pNTS for plastid; mNTS for mitochondria) with broad similarities and subtle differences. They concern the overall amino acid composition, processing peptidases and translocation motifs, and an overall charge difference among the more N-terminal region, in which mNTSs are enriched in arginine and pNTS are enriched in hydroxylated amino acids[35–39]. The subtle differences in NTS are still not fully understood, but determine whether a preprotein is targeted to mitochondria, plastids, or in the case of dual targeted proteins to both compartments simultaneously[40]. Considering the many remaining obstacles of *in vivo* protein localisation (time, resources, overexpression artefacts, impact of the tags on the cargo, or the simple unavailability of transfection methods for non-model systems)[41–47], hundreds of studies rely on algorithms that depend on the difference in NTS features for their localisation prediction. Furthermore, such prediction algorithms are integral parts of widely used databases such as Phytozome[48] or they are nested inside software packages such as InterProScan[49]. Hence, the algorithms are often used routinely, sometimes without a conscious decision to do so, and usually with a lack of knowledge on how reliable they work outside of the species on which they were trained.

In-silico localisation prediction from amino acid sequences were implemented concomitant with our understanding of cellular protein sorting[50–54]. Amino acid composition was used to differentiate between intracellular and secreted proteins[55–57], followed by the use of N-terminal features (e.g. charge and hydrophobicity) for signal sequence detection and cleavage site identification[52,58,59]. This channelled into early prediction algorithms such as PSORT[60] that relied on a relatively simple set of 'if and then' rules to predict signalling peptides and secreted proteins in Gram negative bacteria. PSORT II, an early formal expansion now including eukaryotic compartments[61], incorporated a more sophisticated technique of k-nearest neighbours (kNN), which searches the query against a database of proteins with known localisations and assigns localisation of the nearest neighbours to the query. PSORTb[62,63] introduced machine learning to the pipelines by including support vector machines for accumulating protein sequence features relevant to localisation. This culminated into WOLFPSORT (WPS from here), one of the first sophisticated machine learning algorithms[64,65]. The algorithm uses approximately 20 features of the query sequence to calculate feature vectors, closest neighbours of which from the database are used for assigning a localization prediction. More recently, supervised machine learning was included in a set of programs including Localizer and TargetP[66,67]. Localizer is a classifier algorithm trained to differentiate between N-terminal regions of known organellar and non-organellar proteins. Using the boundary conditions computed from the training dataset, it classifies query proteins. TargetP 2.0 (TargetP from here) is a more sophisticated algorithm that utilises bidirectional neural networks and multi-attention mechanisms on a network of interconnected, long short-term memory cells[67].

Apart from the training and sorting operations, the training datasets themselves also vary (Fig. 1a). WPS for example used a database of 2004 (Uniprot v45.0), a time at which no genomes for bryophytes, ferns, let alone streptophyte algae or multiple organelle proteomes were available. Its training dataset was almost exclusively based on eudicot (for plastid) and animal (for mitochondria) sequences and the proteins were selected based on their annotation from the gene ontology database (GO; evidence codes: TAS, IDA, IMP; cut-off 12.4.2004). Two of these evidence codes (TAS and IMP) are indirect[68] and when used as a starting point, prone to multiplying errors. Localizer was trained on several hundred Viridiplantae organelle proteins from Uniprot (database until March 2016) and validated on the cropPal dataset (barley, wheat, rice, maize) as well as Uniprot Viridiplantae organelle proteins that were added between March and September of 2016. Of these Viridiplantae proteins, a vast majority was of eudicot origin. TargetP used a relatively recent training data, including some green algal proteins, but still leaning heavily towards eudicots.

To date, TargetP, Localizer and WolFPSORT are among the algorithms with a superior reported accuracy and over the years, they have been used abundantly across disciplines (Fig. 1b) but are rarely benchmarked. Therefore, the impact of the skewed training on the performance and reliability of these algorithms outside angiosperms are unexplored. We made use of available, experimentally verified plant proteomes of mitochondria and plastids as well as protein clustering to investigate the reliability of these algorithms across species ranging from algae to angiosperms. Our analysis brings forth inadequacies of these algorithms, caused by a combination of their inherent *modus operandi*, a lack of training on a diverse dataset, the complex biology of the plant cell and the evolutionary dynamic nature of the plant organelles [69]. Tracing the error sources allows to sketch an approach towards developing better algorithms that are capable of serving the diversity of the plant kingdom.

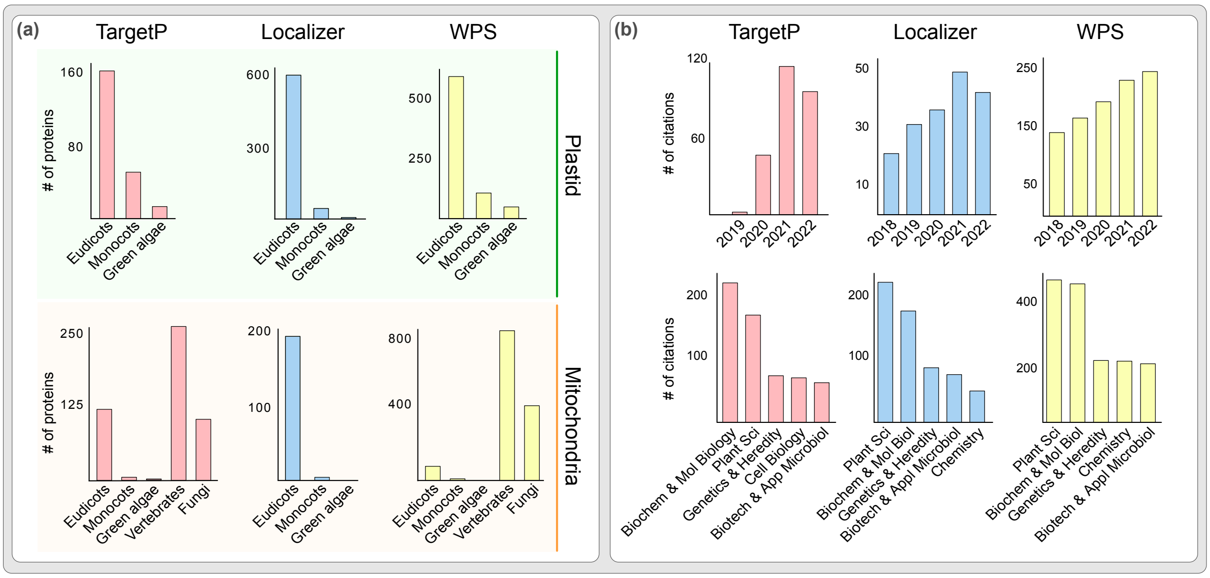
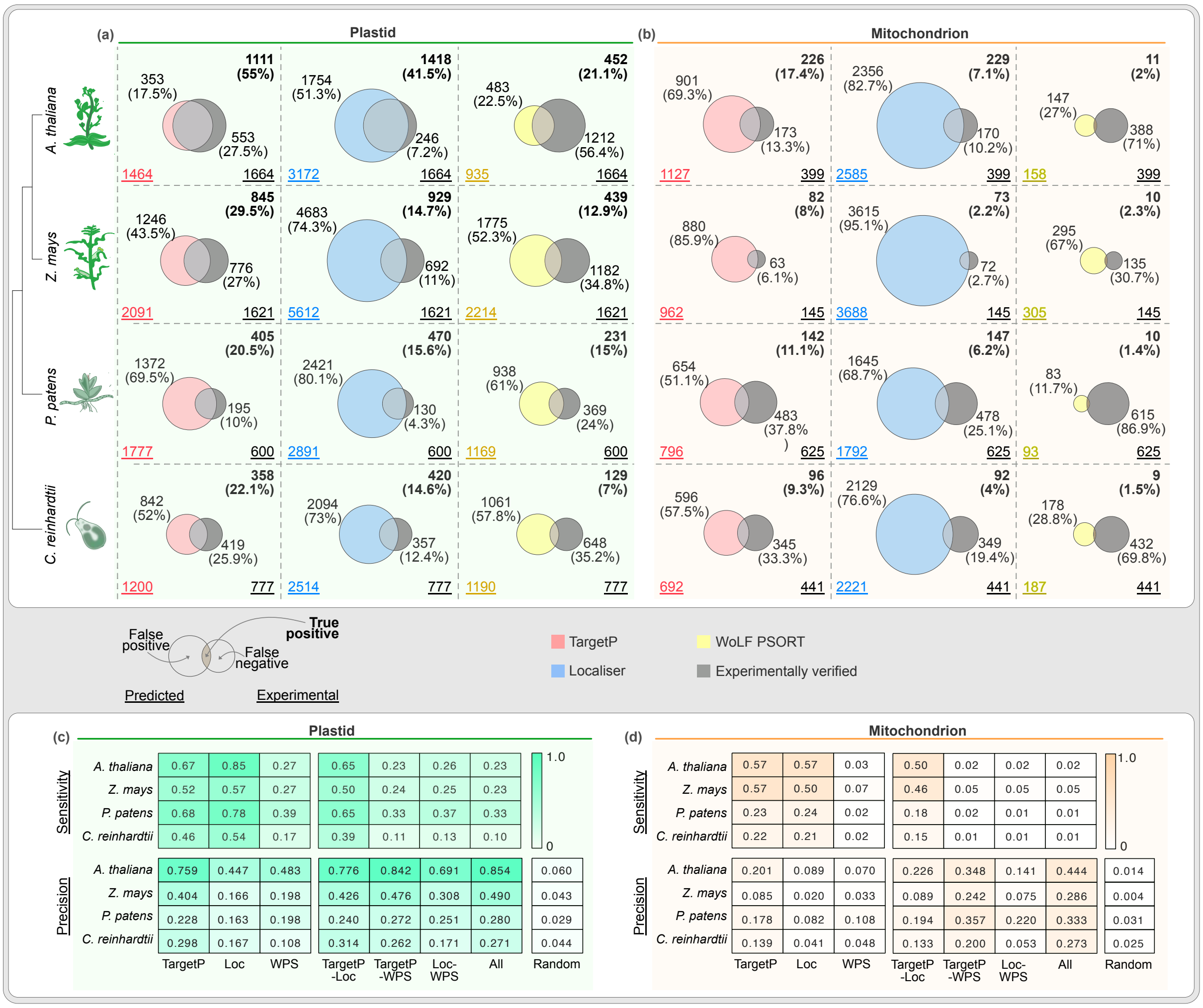


Fig. 1: Targeting prediction algorithms are frequently cited across disciplines and rely on a limited training set. (a) Taxonomic distribution of plastid and mitochondrial training datasets used for the three commonly used predictions tools TargetP, Localizer and WoLF PSORT (WPS). **(b)** Distribution of citations across different disciplines for the three commonly used predictions tools TargetP, Localizer and WoLF PSORT (WPS) and for a time period ranging from 2018 until 2022. Numbers according to the Web of Science.

Results

Algorithm performance declines with evolutionary distance from their training species

At first, we compared the organelle proteomes predicted by the algorithms (the *in-silico* proteomes) with those of experimentally verified organelle proteomes (the *in-vivo* proteomes). Across species, *in-silico* proteomes comprise 3-15% of the proteins encoded within the genome of a given species, in contrast to the *in-vivo* numbers that are below 5% (Fig. S1). Overlaps between *in-silico* and *in-vivo* proteomes show a substantial false positive rate for all, except for the *in-silico* plastid proteome predicted for *Arabidopsis* by TargetP (Fig. 2a,b). Localizer and WPS show larger fractions of false positives than TargetP, especially for mitochondria (Fig. 2b). The smallest overlap between *in-silico* and *in-vivo* proteomes are found for WPS. False negatives are generally predicted fewer on average than false positives, but still to a substantial number (Fig. 2a,b). The sensitivity of TargetP and Localizer are similar, above 0.5 for plastid (i.e., correctly identifying more than half of the plastid proteins) and below 0.5 for mitochondria, whereas that of WPS is 0.3 or lower (Fig. 2c,d). Since 2–5% of the proteins encoded in a nuclear genome have been localised to mitochondria or plastids (Fig. S1) *in-vivo* through proteomics or tagging, a random sampling has a precision of 0.02-0.05; a perfect algorithm should have a precision of or close to 1. Between these two theoretical extremes, established algorithms currently perform closer to random sampling than to the best-case scenario, especially for mitochondria. The best improvement over a random prediction is observed for TargetP on *Arabidopsis* data, which however shifts ever closer to random the greater the evolutionary distance from *Arabidopsis* gets. Combinations of algorithms also reflect similar trends, where TargetP and Localizer together perform marginally better than the two individually, as previously reported[70], albeit confined to the angiosperm plastid (Fig. 2c). For mitochondria, the same combination captured less than 50% of verified proteins across species, and any other combination captured less than 5% (Fig. 2d) due the poor performance of WPS. The precision too, of all combinations, was high in *Arabidopsis*, but declined moving towards *Chlamydomonas* and regardless of combination (Fig. 2c,d). To summarize, the predictions (for any individual algorithm or any combination) are more reliable for angiosperms and with a rapidly declining reliability with respect to algae and bryophytes (Fig. 2).



(c) Plastid

		TargetP	Loc	WPS	TargetP-Loc	TargetP-WPS	Loc-WPS	All	Random
Sensitivity	*A. thaliana*	0.67	0.85	0.27	0.65	0.23	0.26	0.23	
	Z. mays	0.52	0.57	0.27	0.50	0.24	0.25	0.23	
	P. patens	0.68	0.78	0.39	0.65	0.33	0.37	0.33	
	C. reinhardtii	0.46	0.54	0.17	0.39	0.11	0.13	0.10	
Precision	*A. thaliana*	0.759	0.447	0.483	0.776	0.842	0.691	0.854	0.060
	Z. mays	0.404	0.166	0.198	0.426	0.476	0.308	0.490	0.043
	P. patens	0.228	0.163	0.198	0.240	0.272	0.251	0.280	0.029
	C. reinhardtii	0.298	0.167	0.108	0.314	0.262	0.171	0.271	0.044

(d) Mitochondrion

		TargetP	Loc	WPS	TargetP-Loc	TargetP-WPS	Loc-WPS	All	Random
Sensitivity	*A. thaliana*	0.57	0.57	0.03	0.50	0.02	0.02	0.02	
	Z. mays	0.57	0.50	0.07	0.46	0.05	0.05	0.05	
	P. patens	0.23	0.24	0.02	0.18	0.02	0.01	0.01	
	C. reinhardtii	0.22	0.21	0.02	0.15	0.01	0.01	0.01	
Precision	*A. thaliana*	0.201	0.089	0.070	0.226	0.348	0.141	0.444	0.014
	Z. mays	0.085	0.020	0.033	0.089	0.242	0.075	0.286	0.004
	P. patens	0.178	0.082	0.108	0.194	0.357	0.220	0.333	0.031
	C. reinhardtii	0.139	0.041	0.048	0.133	0.200	0.053	0.273	0.025

Fig. 2: Performance of algorithms decline with increasing evolutionary distance from *Arabidopsis*. Comparison of predicted versus experimentally localised plastid **(a)** and mitochondrial **(b)** proteome numbers. Each Venn diagram of the top panel shows an overlap between predicted (left circles) and experimentally verified organelle proteomes (right circles, grey). The underscored numbers in the bottom corners show the total number of predicted (bottom left) and experimentally confirmed proteins (bottom right). The numbers of proteins that overlap (true positives) are provided in the top right corner in bold, while the numbers of non-overlapping ones (false positives) are shown next to each circle. See also the key for the Venn diagrams on the bottom left. Sensitivity, specificity and precision of individual algorithms and their combinations for plastid **(c)** and mitochondria **(d)**.

The training bias of algorithms causes *in-silico* cross-organelle contamination

One likely source of false positives is the errors between the two organelles, especially considering the evolutionary similarities between them and their protein import machineries. For example, a plastid protein can contaminate an *in-silico* mitochondrial proteome (Fig. 3a) or vice versa (Fig. 3b). Such errors can be quantified by overlapping the *in-vivo* proteome of one organelle with the *in-silico* proteome of the another: an overlap between the *in-vivo* plastid proteome and the *in-silico* mitochondrial proteome, highlights those plastid proteins that “contaminated” the *in-silico* mitochondrial proteome (Fig. 3a). We observed that on average about a hundred or more plastid proteins were found across the four species in the *in-silico* mitochondrial proteomes (more frequently so with Localizer, in particular for the bryophyte and alga, Fig. S2) and a smaller number of mitochondrial proteins were identified in the *in-silico* plastid proteomes.

While N–terminal targeting sequences of plastid and mitochondrial proteins (pNTS and mNTS, respectively) share similarities, an mNTS contains a statistically significant higher net positive charge, while pNTSs contain a high number of serine and threonine residues among their first 20 amino acids[36]. It seems these differences became more pronounced later in plant evolution, since they are most striking in the angiosperms (Fig. 3c,d, vertical green and orange lines) – this is a good time to remember that more than 95% of discussed training datasets come from angiosperms (Fig. 1a). Algorithms are inclined to sort NTSs based on these features and any NTS that deviates would be prone to an erroneous cross-organelle prediction, declining the performance of the algorithm. Indeed, NTSs of plastid proteins that showed a higher charge and/or a lower number of phosphorylatable amino acids than the average *Arabidopsis* pNTS, were predicted to be mitochondrial (Fig. 3c) and NTSs of mitochondrial proteins that showed a lower charge and/or higher number of phosphorylatable amino acids than the average *Arabidopsis* mNTS were predicted to be plastid proteins (Fig. 3d). These differences underscored that algorithms are trained to recognise and sort evolutionary late angiosperm NTSs, a bias that causes error when they are faced with NTSs of algae and more ancient plant species.

The substantial number of cross-organelle prediction errors motivated us to investigate the predictability of proteins that are *in vivo* targeted to both, plastid and mitochondria. More than hundred such dually targeted proteins are identified in Arabidopsis[40], the plant proteomes of plastids and mitochondria corroborate such numbers and that is how we treated all proteins that overlapped in the proteome analyses. Algorithms can also predict the same protein to be plastid and mitochondria localised, either explicitly (by listing both these compartments) or implicitly (by providing similar probability scores for these two compartments). We considered such cases as predicted dual targeted proteins. *In-vivo* and *in-silico* dual targeted proteins hardly overlap, with hundreds of false positive and false negatives (Fig. 3e). Except for maize, TargetP predicted most of the experimentally dual localized proteins (i.e. plastid and mitochondrion) to be only plastid localized or not to be organellar at all (Fig. 3e,f). Localizer performed better than the other two with respect to quantity, but at the substantial cost of hundreds of false positives, and WPS failed to predict dual targeted proteins altogether. On the whole, all algorithms perform poorly on this task, sorting experimentally dual targeted proteins to only the plastid or no organelle at all, while also labelling non-organellar or plastid proteins falsely as being dual targeted likely as a result of cross-organelle errors (Fig. 3a-d, Fig. S2).

In summary, a combination of training bias and the evolution of targeting sequences ever since the origin of eukaryotes with mitochondria, culminates into cross-organelle errors which also affect the predictability of the dual targeted proteins.

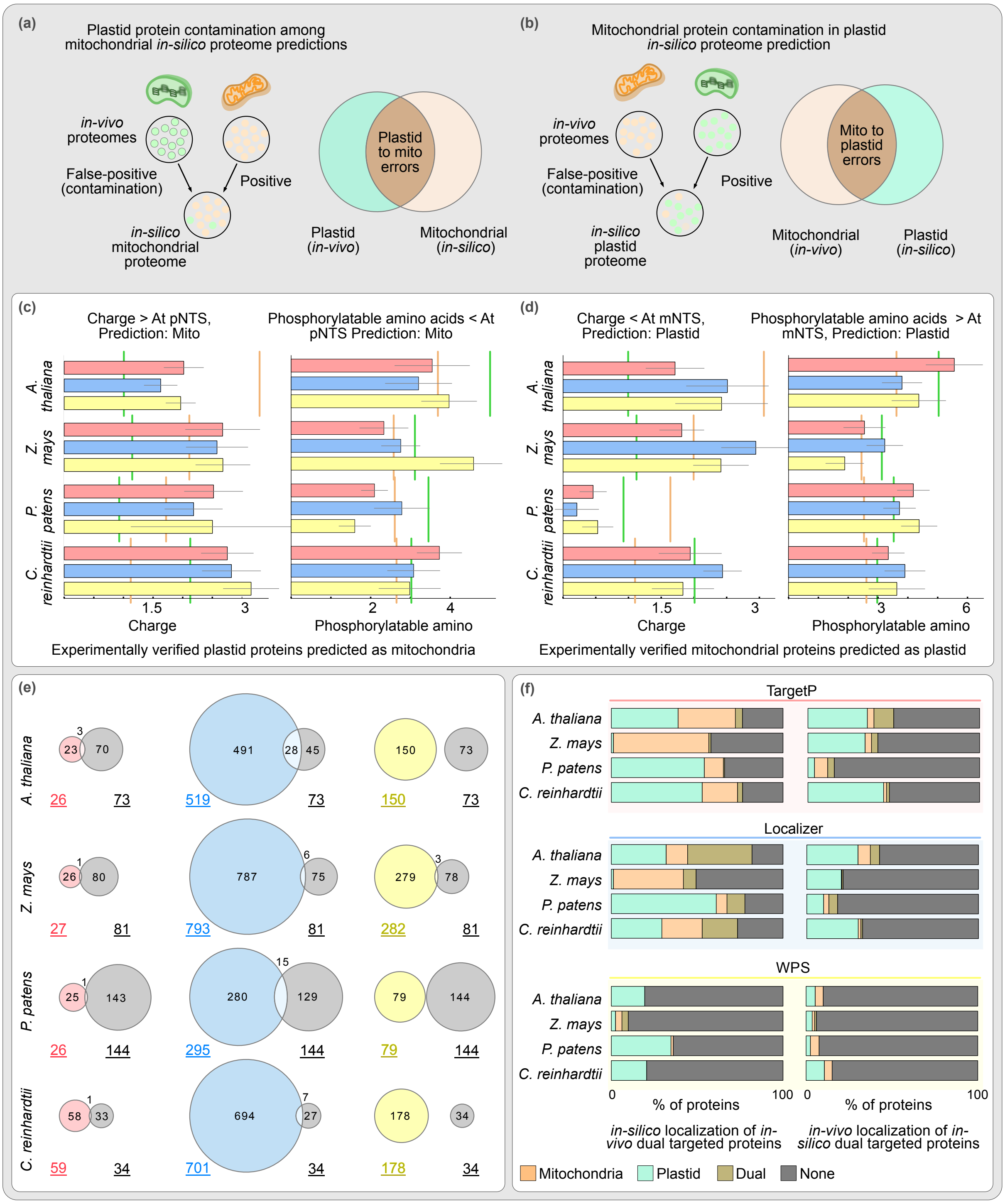


Fig. 3: Cross-organelle errors in proteome prediction due to physio-chemical properties of the NTS. Cross organelle prediction errors could be either because an *in-vivo* plastid protein is *in-silico* mitochondria localised **(a)** or *vice versa* **(b)**. The overlaps between cross-organelle *in-vivo* and *in-silico* proteomes identifies these predictions errors. Analysis of the first 20 amino acids of pNTSs incorrectly predicted to be mitochondrial **(c)** and *vice versa* **(d)**. Average charge and phosphorylatable amino acids for NTS from all verified organelle proteins of each species are indicated by vertical green (pNTS) and orange (mNTS) lines. Error bars indicate standard error of mean (N=4-331, Figure S2) **(e)** Overlap between predicted (left) and experimentally localised (right, in grey) dual targeted proteins. **(f)** Predicted (*in-silico)* intracellular locations of experimentally verified (*in-vivo*) dual targeted proteins (left column) and experimentally verified (*in-vivo*) intracellular locations of proteins that are predicted (*in-silico*) to be dual targeted (right column).

Evolutionary dynamics and the diversity of organelles contribute to prediction inaccuracy

The endosymbiotic organelles of algae and plants have been co-evolving for over a billion-years and their proteomes continue to change and adapt[69,71,72]. During plant terrestrialization for instance, the plastid proteome of the algal ancestor expanded from a few hundred to that of the angiosperm plastid housing about 1500 proteins[69]. The algorithms predict there to be 1000 to 2000 plastid (and mitochondrial) organellar proteins even outside of angiosperms, 25% or less of which appear to be true positives (Fig. 2). Together with the general pattern of the prediction performance worsening with the evolutionary distance to model angiosperms increasing, it prompted us to consider evolutionary dynamics of organelle proteomes as another error source.

We clustered all proteins from the four species into protein families[73], filtered the experimentally verified organelle protein families, and sorted them to be conserved (present in all four species) or to be unique (present in only one species) (Fig. S3, Table S1). Around 150 protein families were found to be conserved across all proteomes, whereas a few hundred were unique. TargetP and Localizer missed around 30% of the conserved proteins, and WPS missed more (Fig. 4a). For the unique plastid proteins, TargetP and Localizer performed well for *Arabidopsis* with declining success for the other species. WPS missed more than 75% of the unique proteins across the species (Fig. 4a). For the conserved mitochondrial protein families, Localizer and TargetP predicted 50-70% correctly, whereas WPS missed more than 90% (Fig. 4c). For mitochondria-unique proteins, the success rate ranged from 20-50% for Localizer and TargetP in *Arabidopsis* and other species, while WPS missed more than 90% across the species (Fig. 4c). More than half of all protein missed out across the algorithms (i.e. false negatives of Fig. 2), were present in only one of a given species (Fig. 4b,d) and likely missed because of a lack of diverse training datasets. With the growing notion of organelle 'pan-proteomes', i.e. organelle proteins present in selected species or organelle sub-types[69,70,72,74–78] , our analysis shows that algorithms are inadequate at capturing this pan-proteome or even the distant homologues of conserved proteins. Training the future algorithms on these missed proteins from across (Fig. 2a-b) and within species [79] would be the first step towards developing algorithms that can cover a larger span of phylogenetic and intracellular (eg. organelle types) diversity.

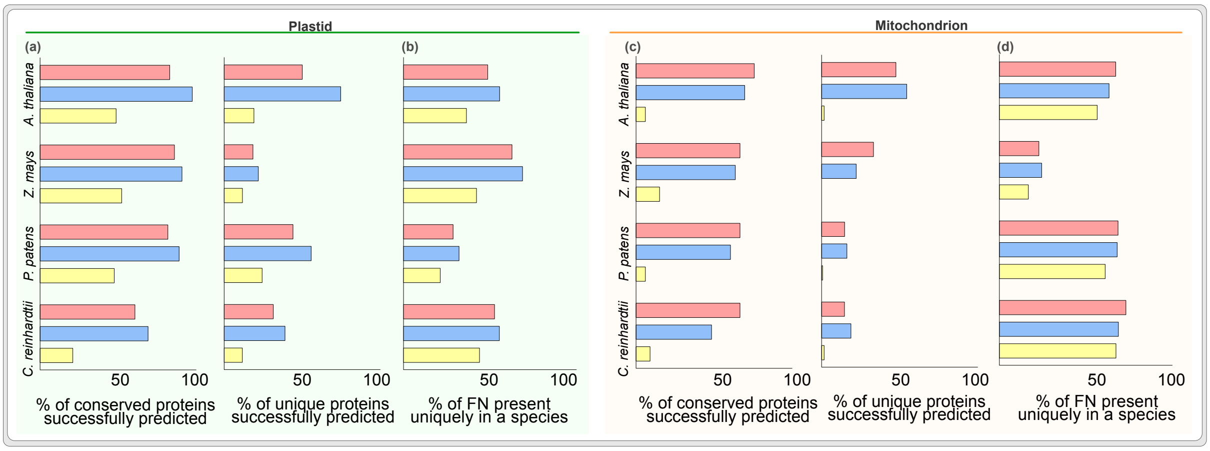


Fig. 4: Success rate of predicting unique versus conserved organelle proteins. Success rate (sensitivity) of predicting experimentally verified conserved and unique proteins for **(a)** plastids and **(c)** mitochondria. Percentage of total plastid **(b)** and mitochondrial **(d)** false negatives explained by their unique presence in a given species.

Discussion

After its cytosolic translation, a plant protein needs to be targeted to the correct compartment if it is not to remain in the cytosol. Machine learning algorithms are used abundantly to determine where proteins are targeted, but they are trained on phylogenetically constrictive datasets (Fig. 1a). As a consequence, the three algorithms evaluated here perform poorly outside of (model) angiosperms, especially for mitochondrial cargo for which the targeting prediction is only slightly better than random sampling. TargetP, the best performing among the three, has a fifty-fifty chance of sorting an algal plastid protein correctly and twice the chance of predicting a false positive. For mitochondria, the error margins are worse. For WPS, the most cited of the three analysed (Fig. 1b), the chances of a wrong prediction are several times higher for plastid- and tens of times higher for mitochondrial proteins. Thus far, such systematic error margins were unavailable and the outputs of these algorithms have been widely accepted directly, across individual studies (Fig. 1b), as well as indirectly, as integral component of widely used software packages and databases. These databases contain genomes from hundreds of diverse species, including taxa with ecological and academic relevance[80–86] such as on streptophyte algae and bryophytes. Proteins from these newly accumulating genomes, however, continue to receive their intracellular localisation annotation from the same set of algorithms and, on an average, 70-80% of these predicted annotations might be unreliable. Algorithms trained on phylogenetically diverse datasets would thus improve reliability of large datasets, while being equally useful to diverse areas of fundamental and applied life sciences (Fig. 1b).

Reflecting on the sources of the prediction errors and in light of the evolutionary cell biology of plants, allows us to sketch improvement strategies for the future algorithms. More than a billion years of co-evolution has resulted in plastid and mitochondrial proteomes and their import machinery, nuances of which affect the predictability of protein sorting. For instance, likely due to a selection pressure against plastid mistargeting, mitochondrial protein import evolved specific receptors such as TOM20 and TOM70[87–91] that are unique to plant mitochondria and have binding sites for cargo that are different from that of animal mitochondria[92–94]. Such changes are likely to be reflected in plant mNTS as well, but not accounted for by the algorithms that are hitherto trained almost exclusively on animal mNTSs (Fig. 1a). Consequently, algorithms require a major upgrade to be able to predict plant mitochondrial proteomes and training them on plant mitochondrial proteins is essential. The impact of organelle co-evolution appears to be more pronounced in the angiosperms NTS (the training dataset), which evolved features different from other clades, such as longer pNTSs and different physicochemical properties of NTSs in general[35,95–97]. However, the details of NTSs are mostly studied in a few angiosperms[98–102] and in league with the skewed training (Fig. 1a) compromises the performance of algorithms outside of angiosperms. A better understanding of NTSs outside of angiosperm remains a bottleneck for developing better algorithms, as much as it remains an unchartered territory in the field of protein import evolution.

Some NTSs are ambiguous and identified equally well by the import machineries of mitochondria and plastids. Although these dual targeted proteins are small in number, they play a key role in information processing[103,104] and have been theorized to reroute whole metabolic pathways[105]. The process of dual targeting appears to be conserved [106,107], rarely lost [106] and can arise by small changes in the NTS[108]. Therefore, it is likely to be common across species, although outside of the model systems, the identification of dually targeted proteins is limited. Algorithms are unlikely to help for now, as they sort dual targeted proteins usually only to plastids, sort plastid proteins to mitochondria as reported previously[109], and falsely predict many plastid proteins to be dually localised. *In vitro* protein import assays with purified organelles also localise many plastid proteins to plastid and mitochondria both, which complicates the matter[110–113]. Such *in vitro* and *in silico* errors on dual targeted proteins limit our understanding of *in vivo* dual targeting mechanisms. Studying protein dual targeting outside of angiosperms would elucidate general strategies of dual targeting. In the interim, explicitly training algorithms on verified dual targeted proteins could help to identify targets for experimental investigation.

Our understanding of cellular protein sorting is far from complete and even experimental approaches do not escape contradictions[45–47]. Localisation prediction reliability, too, varies based on the subcellular compartment in question, consider for instance the varying reliability for nucleus vs. endoplasmic reticulum localized proteins[114]. Our analysis shows that reliability significantly declines when phylogenetically diverse species come into play. This should motivate to plot common benchmarks for other eukaryotes, while underscoring a need for major updates in prediction algorithms for plants. Inclusion of data from diverse techniques into the training of algorithms has been attempted[115] – the inclusion of phylogenetically diverse species should be next. When doing so, and in the absence of proteome data, one could commence with canonical and universally accepted organellar marker proteins. Moreover, not all proteins are equally abundant in organelles, but they often contribute equally to the training process of algorithms. It is conceivable that NTSs have evolved differences based on protein abundance. Inclusion of relative abundance of proteins in the training process might improve the predictions and reveal novel strategies of protein sorting. As advances in in proteomics[116,117] , genomics[84,86,118–122], and machine learning[123,124] set a stage for future prediction algorithms, our analysis serves as a reminder that considering evolutionary diversity is key to a better modelling of protein sorting. One can tailor an algorithm for a given species[125] or a clade[109,126] , but computational power and AI-guided tools likely now make it possible to design a comprehensive prediction algorithm that can serve evolutionarily diverse species and in addition help to better understand the mechanisms of protein sorting in eukaryotic cells in more general.

Methods

Algorithms

All algorithms were installed on a local server supported by the ZIM at the HHU Düsseldorf. Full proteomes were analyzed using TargetP 2.0 (https://services.healthtech.dtu.dk/services/TargetP-2.0/) with the setting 'pl' (plant derived); with Localizer 1.0.4 (https://localizer.csiro.au/software.html) with Python 2.7 and setting '-p'; WPS 0.2 (https://github.com/fmaguire/WoLFPSort) with setting 'plant'. The number of citations for each algorithm were retrieved from the web of science.

Source genomes and organelle proteomes

Genomes of all chloroplastida species were downloaded from Kyoto Encyclopedia of Genes and Genomes (KEGG) [127]. Experimental organelle proteomes were retrieved from published literature and databse as follows: *Chlamydomonas reinhardtii* (chlorophyte algae)[128,129], *Physcomitrium patens* (bryophyte) [130], *Zea mays* (monocot) [42], *Arabidopsis thaliana* (eudicot) [42]

Evaluation of algorithms

We evaluated the performance in species from four diverse chloroplastida species. A protein present in verified proteome and absent in prediction was categorised as false negative. A protein absent in verified proteome and present in prediction was categorised as false positive. A protein present in both, verified and experimental, proteome was categorised as true positive. Sensitivity (ie true positive rate) was calculated as a ratio of true positive and true positive + false negative. Specificity (true negative rate) was calculated as a ratio of true negative and true negative + false positive. Precision was calculated as a ratio of true positive and all predictions. For a combinatorial approach, organelle proteomes were predicted individual by each algorithm and proteins present in the prediction of both or all three algorithms were filtered for further evaluation against experimental proteome. TargetP2.0 predicted 'thylakoid' proteins as a category distinct from 'chloroplast' and therefore around 100 thylakoid proteins were not included under ‘chloroplast predicted’ category. Inclusion of these proteins do not change broad patterns by more than a few percentage (Fig S4, as compared to Fig 1a).

Protein family clustering and annotation

Whole proteomes of all species were clustered into protein families using Orthofinder version 2.5.4 [73]. Source genomes of all species was taken from KEGG [127]. Functional annotations were retrieved using KOID annotated to each of the gene IDs.

Analysis of N-terminal targeting sequences and prediction of the dual targeted proteins

The first 20 amino acids of each protein were retrieved from the whole genome assemblies. Charge was determined by assigning -1 to D,E; +1 to K,R; +0.5 to H and 0 to the rest of the amino acids. The total number of serine and threonine were counted as phosphorylatable amino acids. The verified dual targeted proteins were inferred from overlapping the experimental proteomes of mitochondria and plastid for each species. TargetP sorts proteins to only one intracellular locations that gets the highest probability. However, if probability of mitochondria and plastid both were above 0.35, we considered that protein to be dually targeted. WPS and Localizer predicted more than one locations explicitly, and hence proteins predicted as plastid and mitochondria, were labelled dually targeted.

Author contributions

SBG: Conceptualization, Project administration, Funding acquisition, Writing - original draft, review and editing. JM: Software, Investigation, Data curation, Formal analysis CGG: Writing - original draft, review and editing PKR: Conceptualization, Supervision, Experimental design, Methodology, Investigation, Formal analysis, Visualization, Writing - original draft, review and editing

Funding

We thank the Deutsche Forschungsgemeinschaft (SFB 1208-2672 05415 and SPP2237–440043394) for funding.

Acknowledgments

We acknowledge support from the high-performance computing cluster (HILBERT; ZIM at the HHU Düsseldorf) and Michael R. Knopp from the Heinrich–Heine–University Düsseldorf. We also thank William Martin for a discussion on including protein quantity as a parameter.

References

1. Wiedemann, N. & Pfanner, N. Mitochondrial Machineries for Protein Import and Assembly. (2017) doi:10.1146/annurev.

2. Rochaix J. D. (2022). Chloroplast protein import machinery and quality control. *The FEBS journal*, **289**(22), 6908–6918. https://doi.org/10.1111/febs.16464

3. Gamerdinger, M., & Deuerling, E. (2023). Cotranslational sorting and processing of newly synthesized proteins in eukaryotes. *Trends in biochemical sciences*, S0968-0004(23)00258-X. Advance online publication. https://doi.org/10.1016/j.tibs.2023.10.003

4. Gould, S. B., Garg, S. G. & Martin, W. F. Bacterial Vesicle Secretion and the Evolutionary Origin of the Eukaryotic Endomembrane System. *Trends Microbiol* **24**, 525–534 (2016).

5. Raval, P. K., Garg, S. G., & Gould, S. B. (2022). Endosymbiotic selective pressure at the origin of eukaryotic cell biology. *eLife*, **11**, e81033. https://doi.org/10.7554/eLife.81033

6. Archibald, J. M. Endosymbiosis and Eukaryotic Cell Evolution. Current biology : CB, **25**(19), R911–R921. *Current Biology* Preprint at https://doi.org/10.1016/j.cub.2015.07.055 (2015).

7. Keeling, P. J. The endosymbiotic origin, diversification and fate of plastids. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, **365**(1541), 729–748. https://doi.org/10.1098/rstb.2009.0103 (2010).

8. Martin, W. F., Garg, S. & Zimorski, V. Endosymbiotic theories for eukaryote origin. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, (2015).

9. Dacks, J. B. & Field, M. C. Evolution of the eukaryotic membrane-trafficking system: Origins, tempo and mode. *J Cell Sci* **120**, 2977–2985 (2007).

10. Eliáš, M. Patterns and processes in the evolution of the eukaryotic endomembrane system. *Molecular membrane biology*, **27**(8), 469–489. https://doi.org/10.3109/09687688.2010.521201 (2010).

11. Elliott, L., Moore, I. & Kirchhelle, C. Spatio-temporal control of post-Golgi exocytic trafficking in plants. *J Cell Sci* **133**, (2020).

12. Gould, S. B. Membranes and evolution. *Current Biology* **28**, R381–R385 (2018).

13. Kelly, S. The economics of organellar gene loss and endosymbiotic gene transfer. *Genome Biol* **22**, (2021).

14. Timmis, J. N., Ayliff, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* **5**, 123–135 (2004).

15. Green, B. R. Chloroplast genomes of photosynthetic eukaryotes. *Plant Journal* **66**, 34–44 (2011).

16. Hewitt, V., Alcock, F. & Lithgow, T. Minor modifications and major adaptations: the evolution of molecular machines driving mitochondrial protein import. *Biochimica et biophysica acta*, **1808**(3), 947–954. https://doi.org/10.1016/j.bbamem.2010.07.019 (2011).

17. Hewitt, V., Lithgow, T. & Waller, R. F. Modifications and innovations in the evolution of mitochondrial protein import pathways. in *Endosymbiosis* vol. **9783709113035** 19–35 (Springer-Verlag Wien, 2014).

18. Scotti, P. A. *et al.* YidC, the Escherichia coli homologue of mitochondrial Oxa1p, is a component of the Sec translocase. *EMBO Journal* **19**, 542–549 (2000).

19. Hennon, S. W., Soman, R., Zhu, L. & Dalbey, R. E. YidC/Alb3/Oxa1 family of insertases. *Journal of Biological Chemistry* vol. **290** 14866–14874 Preprint at https://doi.org/10.1074/jbc.R115.638171 (2015).

20. Diederichs, K. A., Buchanan, S. K. & Botos, I. Building Better Barrels – β-barrel Biogenesis and Insertion in Bacteria and Mitochondria. *Journal of Molecular Biology* vol. **433** Preprint at https://doi.org/10.1016/j.jmb.2021.166894 (2021).

21. Jiang, J. H., Tong, J., Tan, K. S. & Gabriel, K. From evolution to Pathogenesis: The link between β-barrel assembly machineries in the outer membrane of mitochondria and Gram-

negative bacteria. *International Journal of Molecular Sciences* vol. **13** 8038–8050 Preprint at https://doi.org/10.3390/ijms13078038 (2012).

22. Moro, F., Fernández-Sáiz, V., Slutsky, O., Azem, A. & Muga, A. Conformational properties of bacterial DnaK and yeast mitochondrial Hsp70: Role of the divergent C-terminal α-helical subdomain. *FEBS Journal* **272**, 3184–3196 (2005).

23. Endow, J. K., Singhal, R., Fernandez, D. E. & Inoue, K. Chaperone-assisted post-translational transport of plastidic type i signal peptidase 1. *Journal of Biological Chemistry* **290**, 28778–28791 (2015).

24. Teixeira, P. F. & Glaser, E. Processing peptidases in mitochondria and chloroplasts. *Biochim Biophys Acta Mol Cell Res* **1833**, 360–370 (2013).

25. Ziehe, D., Dünschede, B. & Schünemann, D. From bacteria to chloroplasts: Evolution of the chloroplast SRP system. *Biological Chemistry* vol. **398** 653–661 Preprint at https://doi.org/10.1515/hsz-2016-0292 (2017).

26. Schein, A. I., Kissinger, J. C. & Ungar, L. H. *Chloroplast transit peptide prediction: a peek inside the black box*. *Nucleic Acids Research* vol. **29** (2001).

27. Chen, Y., Soman, R., Shanmugam, S. K., Kuhn, A. & Dalbey, R. E. The role of the strictly conserved positively charged residue differs among the gram-positive, gram-negative, and chloroplast YidC homologs. *Journal of Biological Chemistry* **289**, 35656–35667 (2014).

28. Day, P. M., Potter, D. & Inoue, K. Evolution and targeting of omp85 homologs in the chloroplast outer envelope membrane. *Front Plant Sci* **5**, (2014).

29. Knopp, M., Garg, S. G., Handrich, M. & Gould, S. B. Major Changes in Plastid Protein Import and the Origin of the Chloroplastida. *iScience* **23**, 100896 (2020).

30. Paila, Y. D. *et al.* Multi-functional roles for the polypeptide transport associated domains of Toc75 in chloroplast protein import. *Elife*. Mar 21;5:e12631 (2016)

31. Richardson, L. G. L. & Schnell, D. J. Origins, function, and regulation of the TOC-TIC general protein import machinery of plastids. *Journal of Experimental Botany* vol. **71** 1226–1238 Preprint at https://doi.org/10.1093/jxb/erz517 (2020).

32. Berks, B. C. The twin-arginine protein translocation pathway. *Annual Review of Biochemistry* vol. **84** 843–864 Preprint at https://doi.org/10.1146/annurev-biochem-060614-034251 (2015).

33. New, C. P., Ma, Q. & Dabney-Smith, C. Routing of thylakoid lumen proteins by the chloroplast twin arginine transport pathway. *Photosynthesis Research* vol. **138** 289–301 Preprint at https://doi.org/10.1007/s11120-018-0567-z (2018).

34. Robinson, C. & Bolhuis, A. Tat-dependent protein targeting in prokaryotes and chloroplasts. *Biochimica et Biophysica Acta - Molecular Cell Research* vol. **1694** 135–147 Preprint at https://doi.org/10.1016/j.bbamcr.2004.03.010 (2004).

35. Ge, C., Spånning, E., Glaser, E. & Wieslander, Å. Import determinants of organelle-specific and dual targeting peptides of mitochondria and chloroplasts in arabidopsis thaliana. *Mol Plant* **7**, 121–136 (2014).

36. Garg, S. G. & Gould, S. B. The Role of Charge in Protein Targeting Evolution. *Trends Cell Biol* **26**, 894–905 (2016).

37. Bhushan, S., Kuhn, C., Berglund, A. K., Roth, C. & Glaser, E. The role of the N-terminal domain of chloroplast targeting peptides in organellar protein import and miss-sorting. *FEBS Lett* **580**, 3966–3972 (2006).

38. Lee, D. W. *et al.* Molecular Mechanism of the Specificity of Protein Import into Chloroplasts and Mitochondria in Plant Cells. *Mol Plant* **12**, 951–966 (2019).

39. Schleiff, E. & Becker, T. Common ground for protein translocation: Access control for mitochondria and chloroplasts. *Nat Rev Mol Cell Biol* **12**, 48–59 (2011).

40. Carrie, C. & Small, I. A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts. *Biochim Biophys Acta Mol Cell Res* **1833**, 253–259 (2013).

41. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* vol. **73** 2092–2123 Preprint at https://doi.org/10.1016/j.jprot.2010.08.009 (2010).

42. Sun, Q. *et al.* PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res* **37**, (2009).

43. Hooper, C. M., Castleden, I. R., Tanz, S. K., Aryamanesh, N. & Millar, A. H. SUBA4: The interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Res* **45**, D1064–D1074 (2017).

44. Hooper, C. M., Castleden, I. R., Aryamanesh, N., Jacoby, R. P. & Millar, A. H. Finding the Subcellular Location of Barley, Wheat, Rice and Maize Proteins: The Compendium of Crop Proteins with Annotated Locations (cropPAL). *Plant Cell Physiology* (2015) doi:10.4225/23/556.

45. Lisenbee, C. S., Karnik, S. K. & Trelease, R. N. Overexpression and mislocalization of a tail-anchored GFP redefines the identity of peroxisomal ER. *Traffic* **4**, 491–501 (2003).

46. Jeong, K., Kim, S. & Bandeira, N. False discovery rates in spectral identification. *BMC Bioinformatics* **13 Suppl 16**, (2012).

47. van Wijk, K. J. & Baginsky, S. Plastid proteomics in higher plants: Current state and future goals. *Plant Physiol* **155**, 1578–1588 (2011).

48. Goodstein, D. M. *et al.* Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* **40**, (2012).

49. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res* **51**, 418–427 (2022).

50. Nakai, K. & Kanehisa, M. *A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells*. *GENOMICS* vol. **14** (1992).

51. Reczko, M. & Hatzigeorgiou, A. Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. *Proteomics* **4**, 1591–1596 (2004).

52. Von Heijne, G. *A new method for predicting signal sequence cleavage sites*. *Nucleic Acids Research* vol. **14** (1986).

53. Bedwell, D. M. *et al. Sequence and Structural Requirements of a Mitochondrial Protein Import Signal Defined by Saturation Cassette Mutagenesis The Saccharomyces cerevisiae Fl-ATPase , subunit precursor contains redundant mitochondrial protein import information at its NH2 terminus (D. MOLECULAR AND CELLULAR BIOLOGY* vol. 9 (1989).

54. Nielsen, H., Tsirigos, K. D., Brunak, S. & von Heijne, G. A Brief History of Protein Sorting Prediction. *Protein Journal* vol. 38 200–216 Preprint at https://doi.org/10.1007/s10930-019-09838-3 (2019).

55. Nishikawa, K. *Correlation of the Amino Acid Composition of a Protein to Its Structural and Biological Characters1. COMMUNICATION J. Biochem* vol. **91** (1982).

56. Nishikawa, K., Kubota, Y. & Ooi, T. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *J Biochem* **94**, 981–995 (1983).

57. Nishikawa, K., Kubota, Y. & Ooi, T. Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. *J Biochem* **94**, 997–1007 (1983).

58. Mcgeoch, D. J. *On the predictive recognition of signal peptide sequences. Virus Research* vol. **3** (1985).

59. Nakai, K. & Kanehisa, M. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function, and Bioinformatics* **11**, 95–110 (1991).

60. Walker, J. M. *PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization*. (Humana Press, 1998).

61. Nakai, K. & Horton, P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24**, 34–36 (1999).

62. Gardy, J. L. *et al.* PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res* **31**, 3613–3617 (2003).

63. Yu, N. Y. *et al.* PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608–1615 (2010).

64. Horton, P. *et al.* WoLF PSORT: Protein localization predictor. *Nucleic Acids Res* **35**, (2007).

65. Horton, P. A., Park, K. A., Obayashi, T. B. & Nakai, K. C. *Protein subcellular localisation prediction using WOLF PSORT*. (2005) doi:10.1142/9781860947292_0007.

66. Sperschneider, J. *et al.* LOCALIZER: Subcellular localization prediction of both plant and effector proteins in the plant cell. *Sci Rep* **7**, (2017).

67. Armenteros, J. J. A. *et al.* Detecting sequence signals in targeting peptides using deep learning. *Life Sci Alliance* **2**, 1–14 (2019).

68. Blake, J. A. *et al.* Gene ontology annotations and resources. *Nucleic Acids Res* **41**, (2013).

69. Raval, P. K., Macleod, A. I. & Gould, S. B. A molecular atlas of plastid and mitochondrial adaptations across the evolution from chlorophyte algae to angiosperms. *bioRxiv* (2023) doi:10.1101/2023.09.01.555919.

70. Christian, R. W., Hewitt, S. L., Roalson, E. H. & Dhingra, A. Genome-Scale Characterization of Predicted Plastid-Targeted Proteomes in Higher Plants. *Sci Rep* **10**, 1–22 (2020).

71. de Vries, J., Stanton, A., Archibald, J. M. & Gould, S. B. Streptophyte Terrestrialization in Light of Plastid Evolution. *Trends Plant Sci* **21**, 467–476 (2016).

72. Schreiber, M., Rensing, S. A. & Gould, S. B. The greening ashore. *Trends in Plant Science* vol. **27** 847–857 Preprint at https://doi.org/10.1016/j.tplants.2022.05.005 (2022).

73. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 1–14 (2019).

74. Heinnickel, M. L. & Grossman, A. R. The GreenCut: Re-evaluation of physiological role of previously studied proteins and potential novel protein functions. *Photosynth Res* **116**, 427–436 (2013).

75. Schaeffer, S. M. *et al.* Comparative ultrastructure of fruit plastids in three genetically diverse genotypes of apple (Malus × domestica Borkh.) during development. *Plant Cell Rep* **36**, 1627–1640 (2017).

76. Richly, E. & Leister, D. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of Arabidopsis and rice. *Gene* **329**, 11–16 (2004).

77. Li, L. & Yuan, H. Chromoplast biogenesis and carotenoid accumulation. *Archives of Biochemistry and Biophysics* vol. **539** 102–109 Preprint at https://doi.org/10.1016/j.abb.2013.07.002 (2013).

78. Choi, H., Yi, T. & Ha, S. H. Diversity of Plastid Types and Their Interconversions. *Frontiers in Plant Science* vol. **12** Preprint at https://doi.org/10.3389/fpls.2021.692024 (2021).

79. Kleffmann, T., Hirsch-Hoffmann, M., Gruissem, W. & Baginsky, S. plprot: A comprehensive proteome database for different plastid types. *Plant Cell Physiol* **47**, 432–436 (2006).

80. Boval, M. & Dixon, R. M. The importance of grasslands for animal production and other functions: A review on management and methodological progress in the tropics. in *Animal* vol. **6** 748–762 (2012).

81. José, J., Karlusich, P., Ibarbalz, F. M. & Bowler, C. Phytoplankton in the Tara Ocean. (2019) doi:10.1146/annurev-marine-010419.

82. Linder, H. P., Lehmann, C. E. R., Archibald, S., Osborne, C. P. & Richardson, D. M. Global grass (Poaceae) success underpinned by traits facilitating colonization, persistence and habitat transformation. *Biological Reviews* **93**, 1125–1144 (2018).

83. Frangedakis, E. *et al.* What can hornworts teach us? *Frontiers in Plant Science* vol. **14** Preprint at https://doi.org/10.3389/fpls.2023.1108027 (2023).

84. Li, F. W. *et al.* Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. *Nat Plants* **6**, 259–272 (2020).

85. Rensing, S. A., Goffinet, B., Meyberg, R., Wu, S. Z. & Bezanilla, M. The moss physcomitrium (Physcomitrella) patens: A model organism for non-seed plants. *Plant Cell* **32**, 1361–1376 (2020).

86. Lang, D. *et al.* The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant Journal* **93**, 515–533 (2018).

87. Murcha, M. W. *et al.* Protein import into plant mitochondria: signals, machinery, processing, and regulation. *J Exp Bot* **65**, 6301–6335 (2014).

88. Murcha, M. W., Wang, Y., Narsai, R. & Whelan, J. The plant mitochondrial protein import apparatus — The differences make it interesting. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1840**, 1233–1245 (2014).

89. Heidorn-Czarna, M., Maziak, A. & Janska, H. Protein Processing in Plant Mitochondria Compared to Yeast and Mammals. *Front Plant Sci* **13**, (2022).

90. Carrie, C., Murcha, M. W. & Whelan, J. An in silico analysis of the mitochondrial protein import apparatus of plants. *BMC Plant Biol* **10**, 249 (2010).

91. Lister, R. *et al.* Functional definition of outer membrane proteins involved in preprotein import into mitochondria. *Plant Cell* **19**, 3739–3759 (2007).

92. Perry, A. J., Hulett, J. M., Likić, V. A., Lithgow, T. & Gooley, P. R. Convergent Evolution of Receptors for Protein Import into Mitochondria. *Current Biology* **16**, 221–229 (2006).

93. Rimmer, K. A. *et al.* Recognition of mitochondrial targeting sequences by the import receptors Tom20 and Tom22. *J Mol Biol* **405**, 804–818 (2011).

94. Chew, O. *et al.* A plant outer mitochondrial membrane protein with high amino acid sequence identity to a chloroplast protein import receptor. *FEBS Lett* **557**, 109–114 (2004).

95. Huang, S., Taylor, N. L., Whelan, J. & Millar, A. H. Refining the definition of plant mitochondrial presequences through analysis of sorting signals, n-terminal modifications, and cleavage motifs. *Plant Physiol* **150**, 1272–1285 (2009).

96. Zhang, X.-P. & Glaser, E. Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci* **7**, 14–21 (2002).

97. Razzak, M. A., Lee, D. W., Yoo, Y. J. & Hwang, I. Evolution of rubisco complex small subunit transit peptides from algae to plants. *Sci Rep* **7**, (2017).

98. Sáiz-Bonilla, M., Martín Merchán, A., Pallás, V. & Navarro, J. A. Molecular characterization, targeting and expression analysis of chloroplast and mitochondrion protein import components in Nicotiana benthamiana. *Front Plant Sci* **13**, 1040688 (2022).

99. Schnell, D. J. The TOC GTPase Receptors: Regulators of the Fidelity, Specificity and Substrate Profiles of the General Protein Import Machinery of Chloroplasts. *Protein J* **38**, (2019).

100. Yan, J., Campbell, J. H., Glick, B. R., Smith, M. D. & Liang, Y. Molecular characterization and expression analysis of chloroplast protein import components in tomato (Solanum lycopersicum). *PLoS One* **9**, (2014).

101. Paul, P. *et al.* The protein translocation systems in plants - composition and variability on the example of Solanum lycopersicum. *BMC Genomics* **14**, 1–16 (2013).

102. Stengel, A., Benz, J. P., Buchanan, B. B., Soll, J. & Bölter, B. Preprotein import into chloroplasts via the Toc and Tic complexes is regulated by redox signals in Pisum sativum. *Mol Plant* **2**, 1181–1197 (2009).

103. Elo, A., Lyznik, A., Gonzalez, D. O., Kachman, S. D. & Mackenzie, S. A. Nuclear genes that encode mitochondrial proteins for DNA and RNA metabolism are clustered in the Arabidopsis genome. *Plant Cell* **15**, 1619–1631 (2003).

104. Carrie, C., Giraud, E. & Whelan, J. Protein transport in organelles: Dual targeting of proteins to mitochondria and chloroplasts. *FEBS J* **276**, 1187–1195 (2009).

105. Martin, W. Evolutionary origins of metabolic compartmentalization in eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 847–855 (2010).

106. Xu, L., Carrie, C., Law, S. R., Murcha, M. W. & Whelan, J. Acquisition, Conservation, and Loss of Dual-Targeted Proteins in Land Plants. *Plant Physiol* **161**, 644 (2013).

107. Morgante, C. V. *et al.* Conservation of dual-targeted proteins in Arabidopsis and rice points to a similar pattern of gene-family evolution. *Molecular Genetics and Genomics* **281**, 525–538 (2009).

108. Burak, E., Yogev, O., Sheffer, S., Schueler-Furman, O. & Pines, O. Evolving dual targeting of a prokaryotic protein in yeast. *Mol Biol Evol* **30**, 1563–1573 (2013).

109. Tardif, M. *et al.* Predalgo: A new subcellular localization prediction tool dedicated to green algae. in *Molecular Biology and Evolution* vol. **29** 3625–3639 (2012).

110. Cleary, S. P. *et al.* Isolated Plant Mitochondria Import Chloroplast Precursor Proteins in Vitro with the Same Efficiency as Chloroplasts. *Journal of Biological Chemistry* **277**, 5562–5569 (2002).

111. Chew, O., Rudhe, C., Glaser, E. & Whelan, J. Characterization of the targeting signal of dual-targeted pea glutathione reductase. *Plant Mol Biol* **53**, 341–356 (2003).

112. Lister, R., Chew, O., Rudhe, C., Lee, M. N. & Whelan, J. Arabidopsis thaliana ferrochelatase-I and -II are not imported into Arabidopsis mitochondria. *FEBS Lett* **506**, 291–295 (2001).

113. Hurt, E. C., Soltanifar, N., Goldschmidt-Clermont, M., Rochaix, J.-D. & Schatz, G. The cleavable pre-sequence of an imported chloroplast protein directs attached polypeptides into yeast mitochondria. *EMBO J* **5**, 1343–1350 (1986).

114. Xiong, E., Zheng, C., Wu, X. & Wang, W. Protein Subcellular Location: The Gap Between Prediction and Experimentation. *Plant Mol Biol Report* **34**, 52–61 (2016).

115. Breckels, L. M. *et al.* Learning from Heterogeneous Data Sources: An Application in Spatial Proteomics. *PLoS Comput Biol* **12**, (2016).

116. Mulvey, C. M. *et al.* Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nat Protoc* **12**, 1110–1135 (2017).

117. Geladaki, A. *et al.* Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat Commun* **10**, (2019).

118. Wang, S. *et al.* Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. *Nat Plants* **6**, 95–106 (2020).

119. Cheng, S. *et al.* Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* **179**, 1057-1067.e14 (2019).

120. Bowman, J. L. *et al.* Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell* **171**, 287-304.e15 (2017).

121. Nishiyama, T. *et al.* The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell* **174**, 448-464.e24 (2018).

122. Hori, K. *et al.* Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nat Commun* **5**, (2014).

123. Bordin, N. *et al.* Novel machine learning approaches revolutionize protein knowledge. *Trends in Biochemical Sciences* vol. **48**, 345–359 Preprint at https://doi.org/10.1016/j.tibs.2022.11.001 (2023).

124. Hesami, M., Alizadeh, M., Jones, A. M. P. & Torkamaneh, D. Machine learning: its challenges and opportunities in plant system biology. *Applied Microbiology and Biotechnology* **106**, 3507–3530 Preprint at https://doi.org/10.1007/s00253-022-11963-6 (2022).

125. Wang, L. *et al.* A chloroplast protein atlas reveals punctate structures and spatial organization of biosynthetic pathways. *Cell* **186**, 3499-3518.e14 (2023).

126. Gruber, A., Rocap, G., Kroth, P. G., Armbrust, E. V. & Mock, T. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant Journal* **81**, 519–528 (2015).

127. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587–D592 (2023).

128. Atteia, A. *et al.* A proteomic survey of chlamydomonas reinhardtii mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the α-proteobacterial mitochondrial ancestor. *Mol Biol Evol* **26**, 1533–1548 (2009).

129. Terashima, M., Specht, M. & Hippler, M. The chloroplast proteome: A survey from the Chlamydomonas reinhardtii perspective with a focus on distinctive features. *Current Genetics* vol. **57** 151–168 Preprint at https://doi.org/10.1007/s00294-011-0339-1 (2011).

130. Mueller, S. J. *et al.* Quantitative analysis of the mitochondrial and plastid proteomes of the moss Physcomitrella patens reveals protein macrocompartmentation and microcompartmentation. *Plant Physiol* **164**, 2081–2095 (2014).