Performance of localization prediction algorithms decreases rapidly with the evolutionary distance to the training set increasing

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 and sometimes impossible. Hence, hundreds of studies make use of algorithms that predict a localisation based on a protein's sequence. Their reliability across evolutionary diverse species is unknown. Here, we evaluate the performance of common algorithms (TargetP, Localizer and WoLFPSORT) for four photosynthetic eukaryotes for which experimental plastid and mitochondrial proteome data is available, and 171 eukaryotes using orthology inferences. The match between predictions and experimental data ranges from 75% to as low as 2%. Results worsen as the evolutionary distance between training and query species increases, especially for plant mitochondria for which performance borders on random sampling. Specificity, sensitivity and precision analyses highlight cross-organelle errors and uncover the evolutionary divergence of organelles as the main driver of current performance issues. The results encourage to train the next generation of neural networks on an evolutionary more diverse set of organelle proteins for optimizing performance and reliability.

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

Correspondence: gould@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 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 many 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 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 predictions 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 by including support vector machines for accumulating protein sequence features relevant to localisation. This culminated into WOLFPSORT (WPS from here on), 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 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 uses 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 deficiencies 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.

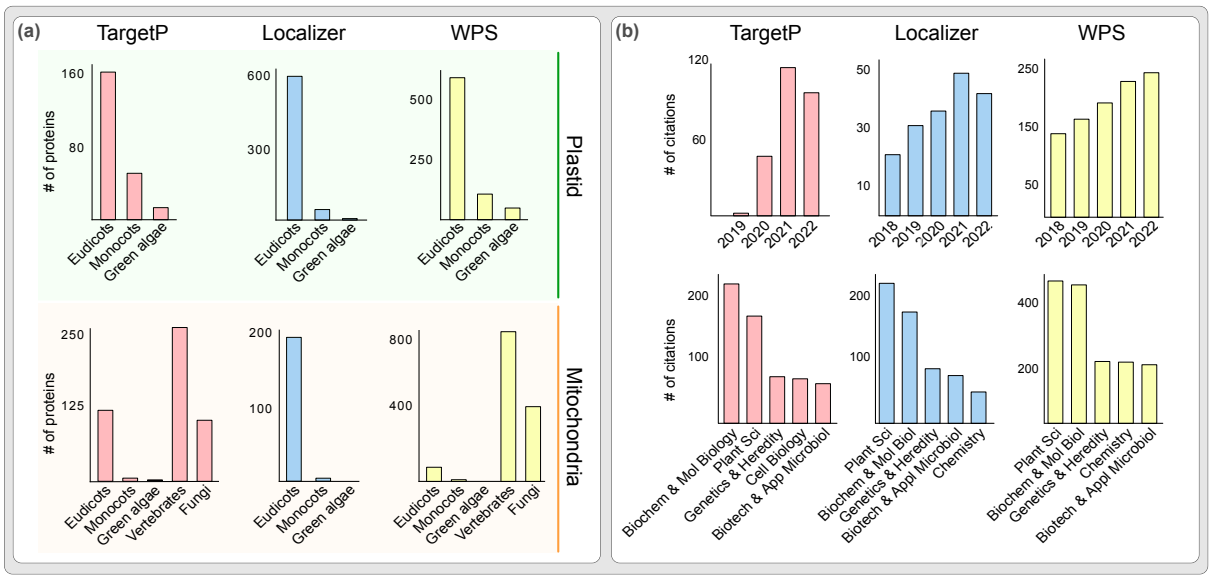

Fig. 1: Targeting prediction algorithms are frequently cited across disciplines and rely on a limited training set. (a) Taxonomic distribution of plastid and mitochondrial training datasets used for the three commonly used predictions tools TargetP, Localizer and WoLF PSORT (WPS). **(b)** Distribution 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

Algorithms perform poorly outside of their training species

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 by the genome of a given species, in contrast to the *in-vivo* numbers that are always below 5% (Fig. S1). Overlaps between *in-silico* and *in-vivo* proteomes show a substantial false positive rate 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 *in-vivo* through proteomics or tagging (Fig. S1), 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 reflect similar trends, where TargetP and Localizer together perform marginally better than each 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% due the poor performance of WPS (Fig. 2d). The precision was high in *Arabidopsis* for all combinations, too, 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).

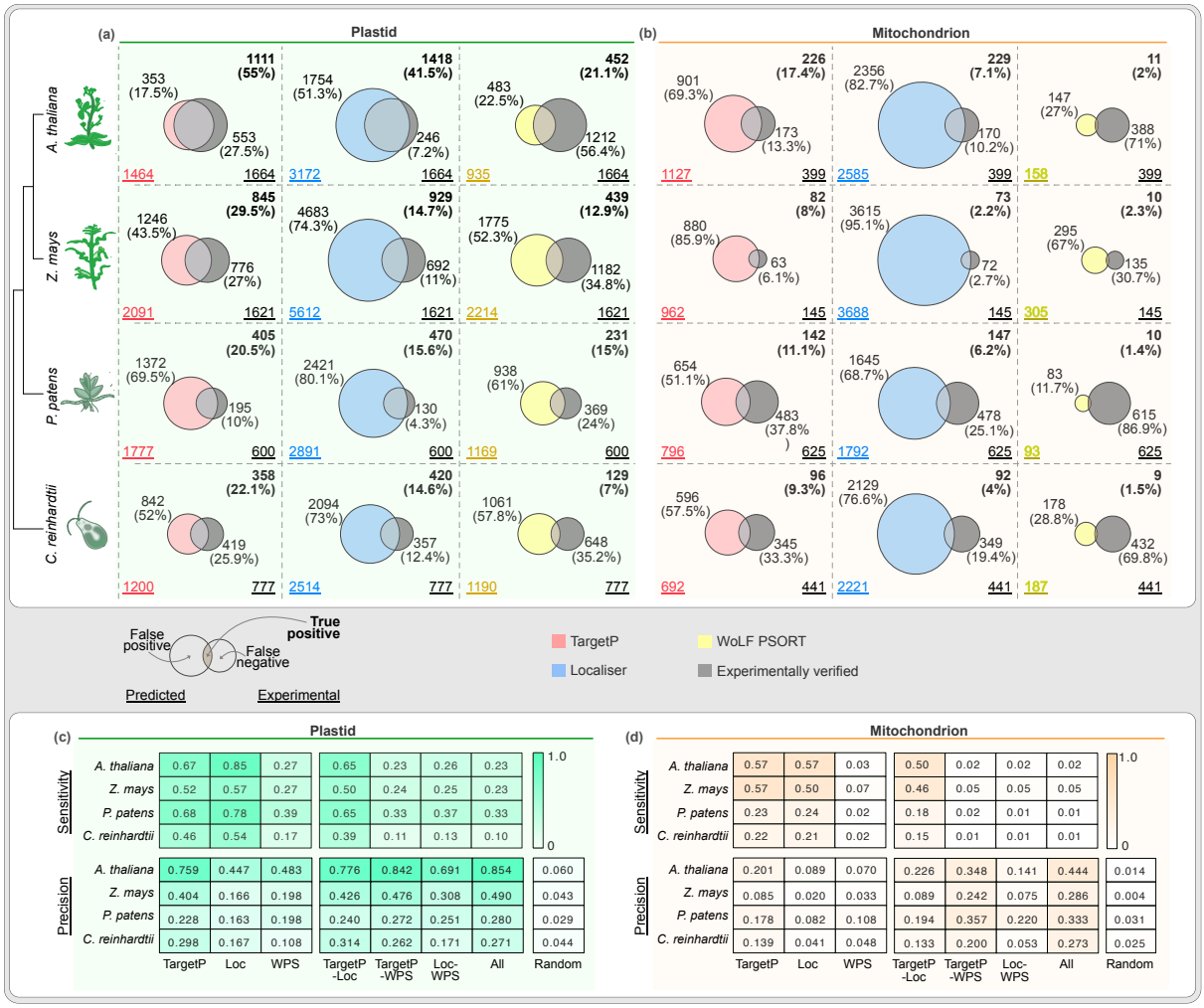

Fig. 2: Performance of algorithms outside the training species. Comparison of predicted versus experimentally localised plastid **(a)** and mitochondrial **(b)** proteome numbers. Each Venn diagram of the top panel shows an overlap between predicted (left circles) and experimentally verified organelle 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)**.

Prediction performance declines as a function of evolutionary distance from the training data

To expand benchmarking across a larger evolutionary scale, we used TargetP (the best performing among the three) to predict organelle proteomes for 171 photosynthetic eukaryotes, for which genomes are available. Since organelle proteomes are scarce, we utilised orthology inferred organelle proteomes based on their sequence similarity with experimentally validated organelle proteomes [69]. Around 70-80% of predicted plastid proteins in eudicots could be validated by orthology based predictions (Fig. 3). Precision is lower for eudicot land plant sister lineages and algae, however, in accordance with patterns observed across the four species for which there is proteome data available (Fig. 2). The precision for mitochondrial protein is lower, including for *Arabidopsis* (and related eudicots) and below 10% for algae. This large and diverse sample size allowed us to systematically and quantitatively test the impact of evolutionary distance between training and test species on the performance of the algorithms. TargetP is trained on 227 plastid and 499 mitochondrial proteins almost exclusively of

eudicot or metazoan origin (Fig. S5). We calculated cophenetic (evolutionary) distances for each of the 171 test species from the most prominent training species, which reveals a significant negative corelation between the precision of algorithms across test species and the evolutionary distance of the test species from the training species. It underscores the need for an algorithm, whose performance is optimized with respect to evolutionary diversity. To this end, we next investigated sources of the prediction errors.

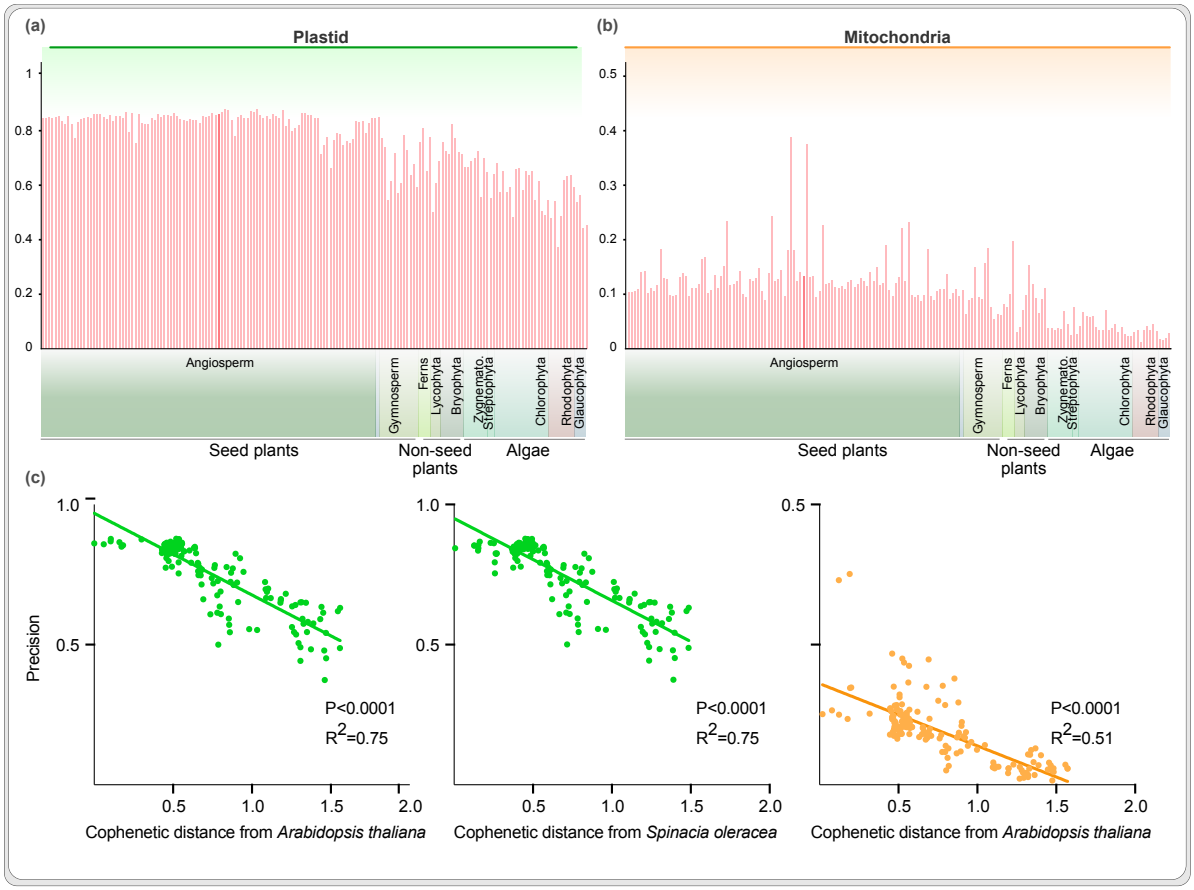

Fig. 3: Strong negative correlation between the precision of algorithms and the evolutionary distance from the training data. Precision of TargetP across eukaryotes for plastid **(a)** and mitochondria **(b)**; *A. thaliana* is shown in darker shade. **(c)** Precision of TargetP as a function of evolutionary distance between the training species and 171 test genomes (plastid in green and mitochondria in orange).

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

One likely source of false positives is the errors between the two organelles, caused also by the similarities in how their protein import machineries evolved. For example, a plastid protein can contaminate an *in-silico* mitochondrial proteome (Fig. 4a) or vice versa (Fig. 4b). 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. 4a). 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 NTSs of plastid and mitochondrial proteins 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. 4c,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. 4c) 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. 4d). These differences underscore that algorithms are trained to recognise and sort evolutionary late angiosperm targeting sequences, a bias that increases the error rate when facing proteins of algae and early branching plant species such as bryophytes.

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. 4e). 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. 4e,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. 4a-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.

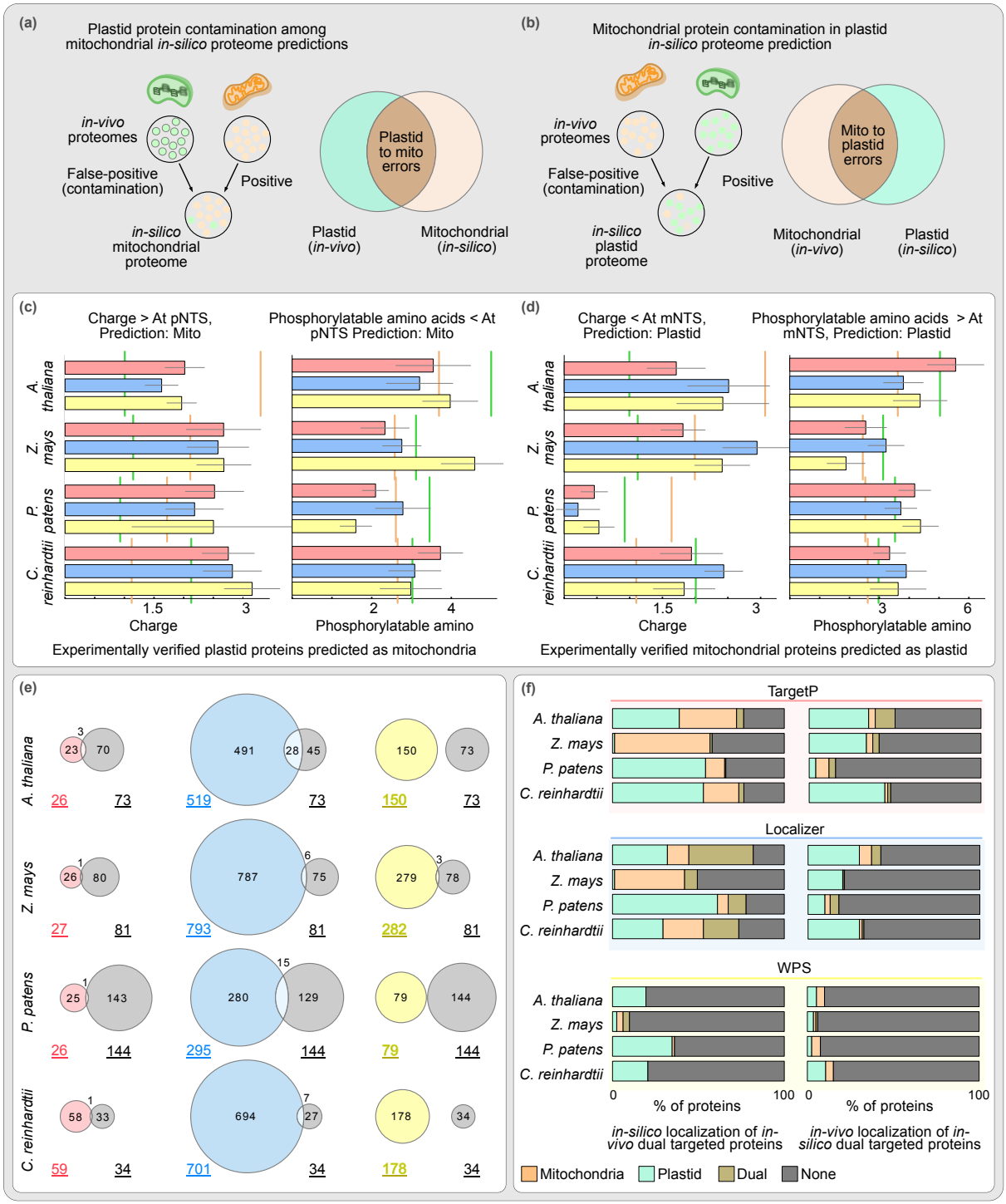

Fig. 4: Cross-organelle errors in proteome prediction due to physio-chemical properties of the NTS. Cross organelle prediction errors could be either because an *in-vivo* plastid protein is *in-silico* mitochondria localised **(a)** or *vice versa* **(b)**. The overlaps between cross-organelle *in-vivo* 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. 5a). 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. 5a). For the conserved mitochondrial protein families, Localizer and TargetP predicted 50-70% correctly, whereas WPS missed more than 90% (Fig. 5c). 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. 5c). 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. 5b,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.

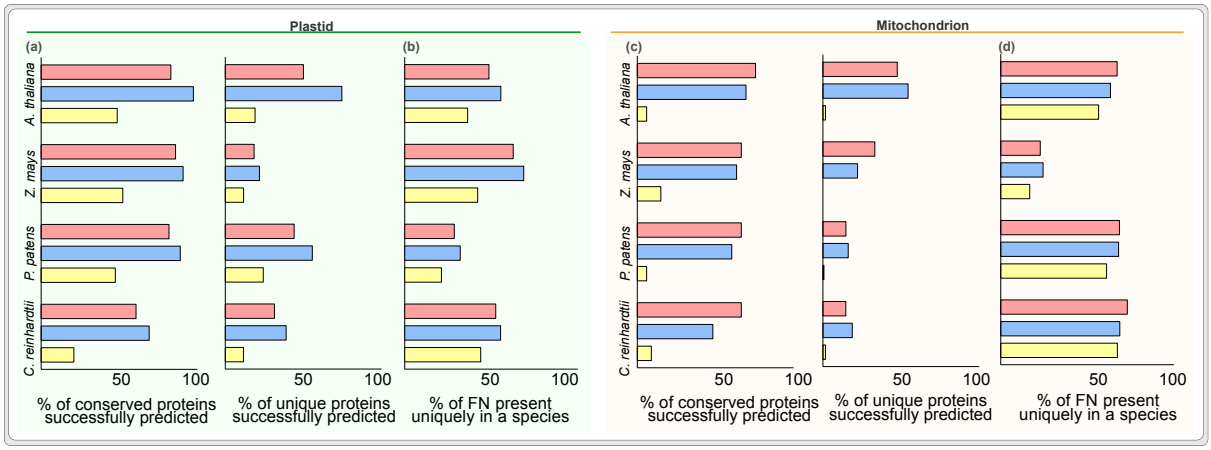

Fig. 5: Success rate of predicting unique versus conserved organelle proteins. Success rate (sensitivity) of predicting experimentally verified conserved and unique proteins for **(a)** plastids and **(c)** mitochondria. Percentage of total plastid **(b)** and mitochondrial **(d)** false negatives explained 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. Such systematic error margins are a real issue, yet the output is trusted across individual studies directly (Fig. 1b) or indirectly through being part of software packages and databases. Databases contain genomes from hundreds of diverse species, including for instance bryophyte and algae taxa of ecological and academic relevance[80–86]. Their genome annotations, however, receive localisation predictions from the same set of algorithms, but on an average 70-80% of them are likely wrong (Fig. 3). 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).

Evaluating the error source in light of the cellular complexity and evolutionary cell biology of plants, allows to sketch improvement strategies for 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. They have binding sites for cargo that differs from that of animal mitochondria[92–94]. Such changes are likely to be mirrored by a plant's mNTS, but is not accounted for by the algorithms that are hitherto trained almost exclusively on animal sequences (Fig. 1a). Consequently, algorithms require an 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 angiosperms sequences (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]. The details of NTSs are mostly studied in a few angiosperms[98–102], however, and in league with the skewed training (Fig. 1a) compromises the performance of algorithms outside of angiosperms. A better understanding of NTSs outside of angiosperms 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, but outside of the model systems the identification of dually targeted proteins is limited. Algorithms are currently of little use in this respect, as they assign dual targeted proteins usually only to the plastid, sort plastid proteins to mitochondria as reported previously[109], or falsely predict many sole 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 us 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[125,126] or a clade[109,127] , 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) [128]. Experimental organelle proteomes were retrieved from published literature and databse as follows: *Chlamydomonas reinhardtii* (chlorophyte algae)[129,130], *Physcomitrium patens* (bryophyte) [131], *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 [128]. 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 location 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 location explicitly, and hence proteins predicted as plastid and mitochondria, were labelled dually targeted.

Corelation between evolutionary distance between training and test species and precision

The genomes of 202 eukaryotes and its phylogeny, along with inferred plastid and mitochondria localisation protein families were retrieved from previous study [69] . TargetP analysis and calculation of prevision was performed as described above. The protein identification numbers of the training data and their species details were retrieved from TaregtP 2.0 (https://services.healthtech.dtu.dk/services/TargetP-2.0/) and uniprot (http://uniprot.org/) websites respectively. The frequency at which different species were represented in the training data was calculated and top two plant species were chosen for evolutionary distance analysis. Their phylogenetic distance from each of the test species was calculated in rstudio using APE [132]. The precision v.s. evolutionary distance plots and linear regression were conducted in Graphpad prism.

Data availability

Supplemental data (figures and tables) are available with the bioRxiv preprint (https://doi.org/10.1101/2023.12.18.572050).

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. Chloroplast protein import machinery and quality control. *FEBS Journal* vol. 289 6908–6918 Preprint at https://doi.org/10.1111/febs.16464 (2022).

3. Gamerdinger, M. & Deuerling, E. Cotranslational sorting and processing of newly synthesized proteins in eukaryotes. *Trends Biochem Sci* (2023) doi: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. Endosymbiotic selective pressure at the origin of eukaryotic cell biology. *eLife* vol. 11 Preprint at https://doi.org/10.7554/ELIFE.81033 (2022).

6. Archibald, J. M. Endosymbiosis and Eukaryotic Cell Evolution. *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 B Biological Sciences* Preprint at 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* Preprint at 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 - Biomembranes vol. 1808 947–954 Preprint at 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* (2016) doi:10.7554/eLife.12631.001.

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).

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

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

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

482 35. Ge, C., Spånning, E., Glaser, E. & Wieslander, Å. Import determinants of organelle-specific
483 and dual targeting peptides of mitochondria and chloroplasts in arabidopsis thaliana. *Mol Plant*
484 **7**, 121–136 (2014).

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

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

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

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

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

496 41. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for
497 peptide and protein identification in shotgun proteomics. *Journal of Proteomics* vol. 73 2092–
498 2123 Preprint at https://doi.org/10.1016/j.jprot.2010.08.009 (2010).

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

500 43. Hooper, C. M., Castleden, I. R., Tanz, S. K., Aryamanesh, N. & Millar, A. H. SUBA4: The
501 interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Res*
502 **45**, D1064–D1074 (2017).

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

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

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

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

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

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

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

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

520 52. Von Heijne, G. *A New Method for Predicting Signal Sequence Cleavage Sites*. *Nucleic Acids*
521 *Research* vol. 14 (1986).

522 53. Bedwell, D. M. *et al. Sequence and Structural Requirements of a Mitochondrial Protein*
523 *Import Signal Defined by Saturation Cassette Mutagenesis The Saccharomyces Cerevisiae Fl-*
524 *ATPase , Subunit Precursor Contains Redundant Mitochondrial Protein Import Information at*
525 *Its NH2 Terminus (D. MOLECULAR AND CELLULAR BIOLOGY* vol. 9 (1989).

526 54. Nielsen, H., Tsirigos, K. D., Brunak, S. & von Heijne, G. A Brief History of Protein Sorting
527 Prediction. *Protein Journal* vol. 38 200–216 Preprint at https://doi.org/10.1007/s10930-019-
528 09838-3 (2019).

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

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

533 57. Nishikawa, K., Kubota, Y. & Ooi, T. Classification of proteins into groups based on amino
534 acid composition and other characters. II. Grouping into four types. *J Biochem* **94**, 997–1007
535 (1983).

536 58. Mcgeoch, D. J. *On the Predictive Recognition of Signal Peptide Sequences*. *Virus Research*
537 vol. 3 (1985).

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

540 60. Walker, J. M. *PSORT: A Program for Detecting Signals in Proteins and predicting Their*
541 *Subcellular*. (Humana Press, 1998).

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

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

546 63. Yu, N. Y. *et al.* PSORTb 3.0: Improved protein subcellular localization prediction with refined
547 localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**,
548 1608–1615 (2010).

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

550 65. Horton, P. A., Park, K. A., Obayashi, T. B. & Nakai, K. C. *Protein Subcellular Localisation*
551 *Prediction Using WOLF PSORT*. (2005) doi:10.1142/9781860947292_0007.

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

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

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

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

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

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

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

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

568 74. Heinnickel, M. L. & Grossman, A. R. The GreenCut: Re-evaluation of physiological role of
569 previously studied proteins and potential novel protein functions. *Photosynth Res* **116**, 427–
570 436 (2013).

571 75. Schaeffer, S. M. *et al.* Comparative ultrastructure of fruit plastids in three genetically diverse
572 genotypes of apple (Malus × domestica Borkh.) during development. *Plant Cell Rep* **36**, 1627–
573 1640 (2017).

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

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

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

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

583 80. Boval, M. & Dixon, R. M. The importance of grasslands for animal production and other
584 functions: A review on management and methodological progress in the tropics. in *Animal* vol.
585 6 748–762 (2012).

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

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

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

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

595 85. Rensing, S. A., Goffinet, B., Meyberg, R., Wu, S. Z. & Bezanilla, M. The moss physcomitrium
596 (Physcomitrella) patens: A model organism for non-seed plants. *Plant Cell* **32**, 1361–1376
597 (2020).

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

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

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

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

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

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

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

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

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

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

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

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

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

627 99. Schnell, D. J. The TOC GTPase Receptors: Regulators of the Fidelity, Specificity and
628 Substrate Profiles of the General Protein Import Machinery of Chloroplasts. *Protein J* **38**,
629 (2019).

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

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

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

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

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

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

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

647 107. Morgante, C. V. *et al.* Conservation of dual-targeted proteins in Arabidopsis and rice points to
648 a similar pattern of gene-family evolution. *Molecular Genetics and Genomics* **281**, 525–538
649 (2009).

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

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

654 110. Cleary, S. P. *et al.* Isolated Plant Mitochondria Import Chloroplast Precursor Proteinsin Vitro
655 with the Same Efficiency as Chloroplasts. *Journal of Biological Chemistry* **277**, 5562–5569
656 (2002).

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

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

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

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

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

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

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

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

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

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

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

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

682 123. Bordin, N. *et al.* Novel machine learning approaches revolutionize protein knowledge. *Trends*
683 *in Biochemical Sciences* vol. 48 345–359 Preprint at https://doi.org/10.1016/j.tibs.2022.11.001
684 (2023).

685 124. Hesami, M., Alizadeh, M., Jones, A. M. P. & Torkamaneh, D. Machine learning: its challenges
686 and opportunities in plant system biology. *Applied Microbiology and Biotechnology* vol. 106
687 3507–3530 Preprint at https://doi.org/10.1007/s00253-022-11963-6 (2022).

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

690 126. Burstein, D. *et al.* A machine learning approach to identify hydrogenosomal proteins in
691 trichomonas vaginalis. *Eukaryot Cell* **11**, 217–228 (2012).

692 127. Gruber, A., Rocap, G., Kroth, P. G., Armbrust, E. V. & Mock, T. Plastid proteome prediction
693 for diatoms and other algae with secondary plastids of the red lineage. *Plant Journal* **81**, 519–
694 528 (2015).

695 128. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for
696 taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587–D592
697 (2023).

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

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

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

707 132. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R
708 language. *Bioinformatics* **20**, 289–290 (2004).

709