MSc. Thesis, 2025

A protein structure-informed approach to eukaryotic genome annotation

Gijs Berkenbosch[1]

[1]Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, Wageningen 6708 PB, the Netherlands

Abstract

Genome annotation is a crucial step to understand the functions of genes. Despite advances in the performance of predictive annotation tools and the assembly of genomes, challenges remain in structural genome annotation, particularly for less-studied and more complex eukaryotic genomes. Recent improvements in protein structure prediction have greatly increased the availability of predicted protein structure data. This offers new opportunities to enhance genome annotation by providing additional structural information. This study aims to improve annotation sensitivities by integrating protein structure data. We employ two approaches: structure-based filtering (SBF) to refine existing annotation predictions, and direct protein structure alignment (DPSA) to identify previously unannotated gene structures. In both approaches structural protein alignment is performed using FoldSeek's 3Di structural alphabet. The results indicated no clear added value of structure data over sequence data. When comparing the performance of sequence- and structure-based approaches, inherently difficult to annotate regions were not predicted better using structure compared to sequence approaches. These findings suggest that, while protein structure data may appear promising for improving genome annotation, its practical application is limited using our approach. More research is needed to explore how different types of structural data might improve existing genome annotation tools and help to address the complexities of gene prediction in eukaryotes. This work provides a basis for future approaches that integrate structural information into annotation pipelines.

Introduction

Next-Generation Sequencing has facilitated the generation of vast amounts of genome-wide sequence data. Furthermore, increasingly large and more complex genomes have been assembled due to the development of long-read sequencing technologies (Pucker et al., 2022). Unfortunately, structural and functional annotation of these assembled genomes remains challenging. Structural annotation involves the identification and marking of genomic components, such as gene boundaries, exon and intron regions, promoters and regulatory elements. This can be followed by functional annotation, which assigns biological functions to these identified components (Vuruputoor et al., 2023). For clarity, this work will refer to structural annotation as gene model annotation. The annotation of bacterial genomes is relatively simple, given roughly 90% of the genes are protein-coding, with relatively short intergenic distances (Salzberg, 2019). Eukaryotic genomes, on the other hand, are much more complex. These are often larger, contain long intergenic sections, have repetitive DNA sequences, and genes can contain introns (Cooper, 2000). The complex nature of eukaryotic genomes results in annotation errors that are propagated in subsequent analyses, as annotation is often based on homology. This problem is even more prevalent in organisms that do not have, or are not closely related to, an organism with a well-established genome annotation (Gabriel et al., 2024).

Annotation predictions are based on intrinsic and in some cases also extrinsic information. Intrinsic information is obtained from *ab initio* predictions based on DNA sequences, while extrinsic information is obtained from transcripts and homologous proteins (Ejigu and Jung, 2020). The recently developed genome annotation pipelines BRAKER3 by Gabriel et al. (2024) and Helixer by Triesch et al. (2023) are currently seen as state-of-the-art for the annotation of protein-coding genes in eukaryotes. However, they differ in their approach and incorporation of intrinsic and extrinsic information. BRAKER3 uses a machine learning approach that combines intrinsic and extrinsic information, integrating short-read RNA-seq and protein data using tools GeneMark-ETP and AUGUSTUS. GeneMark-ETP is a gene finder that generates so-called "high confidence genes", which are subsequently used for gene prediction using AUGUSTUS. Both tools use hidden Markov models (HMMs) for their predictions. Helixer, on the other hand, uses a deep learning approach that combines a deep neural network with an HMM for its predictions. It operates on only intrinsic information and is designed as a quick and user-friendly tool. Unfortunately,

the obtained gene annotations from BRAKER3 and Helixer are still quite far from perfect. For example, BRAKER3 achieves a sensitivity of just 77.9%, and a precision of 86.5% for *Arabidopsis thaliana*, a well-studied model organism.

In addition to the already abundant genomics, transcriptomics, and proteomics data, recent developments, like AlphaFold, have leveraged the vast availability of *in silico* translated sequence data and made (predicted) protein structure data widely available (Jumper et al., 2021). Ruperti et al. (2023) demonstrated the potential of protein structure data for the identification of homologous proteins over longer evolutionary distances. Given the higher evolutionary conservation of structure, compared to sequence, this contributed to reliable annotation for less-studied species, particularly those that have a large evolutionary distance from well-studied species (Ardell et al., 2009). For example, the sparsely annotated *Spongilla* proteome was aligned against structural databases to identify structurally similar proteins and assign the associated functional annotations. This approach increased the functional annotation coverage in the Spongilla proteome by 50%.

Based on the potential of structural data to complement sequence-based annotation, the main research question posed in this project is: to what extent can the use of protein structure data enhance eukaryotic gene model annotation? We aim to explore the potential of protein structure information to improve annotation sensitivity in gene model annotation, specifically for challenging annotations like over long evolutionary distances (Ardell et al., 2009). Many genes, even in very well-studied organisms, remain unannotated as a result from the low sensitivity of current annotation tools (Gabriel et al., 2024). To isolate the contribution of structure data, our approach will be compared to sequence-based approaches using Blastp. In addition to gene model predictions based on structure, we will investigate to what extent functional characteristics of proteins influence annotation performance in existing annotation practices. Genome annotations are at the core of a wide range of applications, from evolutionary research to drug design, as well as diagnostics of genetic disorders (Ejigu and Jung, 2020). Increasing the overall accuracy of these annotations can potentially avoid errors and aid the identification of previously missed features.

Methods

The potential of protein structure integration in genome annotation was explored using two approaches (Fig. 1). The first approach combines existing annotation pipelines (Fig. 1 B) and protein structure alignment (Fig. 1 D), which is referred to here as "Structure-based filtering" (SBF). SBF targets incorrectly predicted gene models, i.e. false positives. This approach attempts to improve gene model predictions from existing tools, BRAKER3 and Helixer, under the assumption that their prediction sensitivity can be improved by modifying input parameters. However, this comes at the cost of an inflated number of false positives. We reasoned that these can be removed using structure-based alignment, thereby improving precision. The second approach, called "Direct protein structure alignment" (DPSA), integrates a custom pipeline we developed, "Candidate protein fragment extraction" (CPFE) (Fig. 1 C), with protein structure alignment (Fig. 1 D). DPSA specifically targets gene models missed in existing gene model predictions, i.e. false negatives. Unlike traditional gene model prediction methods, CPFE directly translates the genome assembly into protein sequences without relying on gene predictions. This study assumes that this method captures at least parts of all possible protein sequences, which can then be used for structure-based alignment.

Pipelines

Annotation pipelines

The sensitivity of gene model predictions from existing tools can be increased by modifying the input parameters to adjust the precision-sensitivity trade-off. To assess this trade-off, BRAKER3 and Helixer were evaluated for their ability to improve sensitivity.

The most relevant parameters to improve sensitivity in BRAKER3 are temperature (t) and the reporting of alternative transcripts, either from evidence or sampling. Temperature scales the probabilities of predicted gene models by raising them to the power of $\frac{8-t}{8}$. Low values of t yield a more selective model, prioritizing the highest-likelihood gene models, resulting in a small selection of predictions. In contrast, high values provide greater variability and the reporting of less likely gene models. In our study t was set to 7, the maximum value, while the default value is 0. Alternative transcripts are not reported by default, but can be found when suggested by extrinsic information on gene models, i.e. hints, or when produced by probabilistic sampling (Gabriel et al., 2024). The parameters for the reporting of alternatives transcripts from evidence and sampling were tested both enabled and disabled. Each parameter was tested individually and in combination. Precomputed AUGUSTUS weights were used for all species.

Helixer relies on deep neural networks and a hidden Markov model for the prediction of gene models. The precision-sensitivity trade-off in Helixer is controlled by the peak-threshold parameter, which defines the minimum peak genic F1 score required for the acceptance of a candidate region. In our investigation, it was set to 0 instead of the default value of 0.8 (Triesch et al., 2023).

Candidate protein fragment extraction

In this pipeline, a sliding window strategy was used to identify open reading frames (ORFs). ORFs were defined as continuous sequences starting at the beginning of the window or directly after a stop codon, and ending with a stop codon, or in the absence of a valid stop codon within the window, until the window's boundary. All six reading frames were translated to amino acid sequences using the Biopython package (Cock et al., 2009). The amino acid sequences obtained were split at their stop codons, with the stop codons themselves excluded, and filtered on a minimal length to prioritize biologically relevant amino acid sequences and improve efficiency by decreasing runtime. The processed sequences are considered "Candidate protein fragments".

Using this approach, protein sequences were predicted on genomic regions that remained unannotated after processing with existing annotation pipelines to identify previously missed gene models. Regions with previously predicted gene models were also included in a subsequent experiment to evaluate the pipeline's performance on annotated and unannotated regions. The pipeline was optimized by adjusting the minimal protein fragment length threshold. A range of threshold values was tested to determine the optimal balance between efficiency while minimizing sensitivity loss.

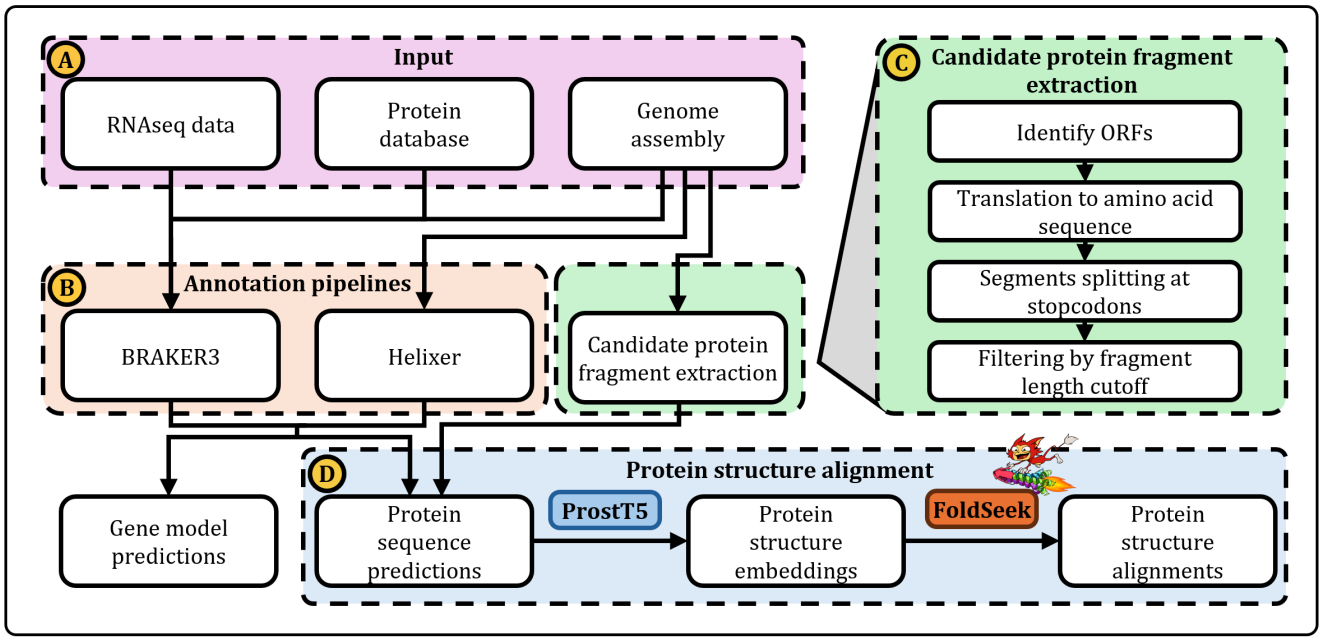


Fig. 1: Schematic overview of the project workflow for integrating protein structure data in genome annotation. (A) Input data (B) Genome annotation pipelines BRAKER3 and Helixer (C) Candidate protein fragment extraction (CPFE) (D) Protein structure alignment using ProstT5 and FoldSeek. (Kempen et al., 2023; Weisseno et al., 2024; Gabriel et al., 2024; Triesch et al., 2023)

Protein structure alignment

Protein structure alignment was used in both approaches to refine gene model predictions by filtering false positives and identifying missed gene models. Annotation tools provide both gene model and protein sequence predictions, while CPFE only yields the latter. 3Di structure embeddings were used to align predicted protein sequences to a protein structure database. These embeddings describe the amino acid backbone of proteins in a structural alphabet, enabling fast structural comparisons. ProstT5 was used to compute the embeddings, and FoldSeek to perform the alignment (Kempen et al., 2023; Weisseno et al., 2024). An embedded version of the SwissProt AlphaFold database was used for structural alignment to leverage its high-quality detailed annotations in subsequent analysis (Bairoch and Apweiler, 2000). These alignments can provide structural validation and can be used to assess the reliability of the predicted gene models.

FoldSeek embedding alignment, like protein sequence alignment, uses bit-scores as a statistical metric for alignment similarity. However, FoldSeek also accounts for structural similarity, resulting in the "Structural bit-score". This score is obtained by multiplying the the Smith-Waterman alignment bit-score with the geometric mean of the TM-score and the average alignment LDDT. This integrates both local (LDDT) and global (TM-score) structural similarity into the alignment statistic. A range of bit-score values was tested in order to find a suitable threshold value for alignment acceptance. The optimal threshold was selected based on its ability to balance precision and sensitivity in the alignments.

Performance evaluation

To assess the impact of SBF and DPSA on gene model predictions, we evaluated their performance using reference annotations and statistical metrics.

Structure-based filtering

The effect of parameter modification to adjust the precision-sensitivity trade-off in BRAKER3 and Helixer was evaluated using reference annotations and the tools' respective performance evaluation method, yielding precision and sensitivity metrics on the full gene model. However, the evaluation methods differ slightly between the tools. BRAKER3 computes performance metrics based on exact interval matches to the reference annotation, while Helixer reports base pair level accuracy, where a true positive (TP) is defined as a match between the predicted class and the annotated class (Gabriel et al., 2024; Stiehler et al., 2020). SBF was evaluated using a selection of proteins that contain folds with high structure and low sequence conservation. This was defined as having RMSD < 3.5 and amino acid identity $< 25\%$ (Sousounis et al., 2012). The total number of proteins present after filtering is compared with the total number of proteins present in the reference annotation. Only the folds present in relevant numbers were considered, meaning the organism has > 50 proteins in SwissProt that contain the fold. To isolate the effect of the use of protein structures in SBF, the approach was compared to sequence-based filtering using Blastp.

Direct protein structure alignment

The interval locations of structural alignments after DPSA were compared to the reference annotations to assess the predictive performance of the pipeline. Regions were extracted using the annotation pipeline that achieved the highest precision and sensitivity on *Arabidopsis thaliana*, despite the subtle differences

in performance evaluation between the tools. Furthermore, regions were extracted from the organism that achieved the lowest performance within the same pipeline. In addition to the extraction of candidate protein fragments from unannotated regions, annotated regions were also assessed for comparison. An important aspect in this evaluation is the definition of true positive alignments. Different definitions were explored; an exact match in intervals, overlap between intervals, and predictions falling within a certain margin from the reference interval. For a visual representation, see Supplementary Fig. 11. The runtimes of the DPSA pipelines were recorded using the *time* command for a single execution using 1 thread. Subsequent analysis of the structural alignments was performed based on protein length and functional annotation. The latter was performed based on a hyper-geometric distribution probability analysis of the Biological Process aspects in the plant Gene Ontology (GO) subset, using Benjamini–Hochberg procedure to correct for multiple hypothesis testing (Gene Ontology Consortium, 2024b). GO annotations of the organism's proteome and the complete SwissProt database were tested as background, and true positive protein alignments were used as sample (UniProt, 2024). The parent terms were included based on the GO-basic ontology for both the background and the sample (Gene Ontology Consortium, 2024a).

Materials

Materials for structure-based filtering

The species selected for BRAKER3 (v3.0.8) were *Arabidopsis thaliana* and *Physcomitrium patens*. For each species we selected, following Gabriel et al. (2024), a genome assembly and reference annotation were retrieved from Ensembl, 6 randomly selected short-read RNA-seq libraries were retrieved from NCBI, and a protein database of proteomes of closely related species was retrieved from NCBI, see Supplementary tables 4, 5, and 6 for more detail. The species selected for Helixer (v0.3.4) were *Arabidopsis thaliana*, *Selaginella moellendorffii*, and *Juglans regia*. *Arabidopsis thaliana* and *Selaginella moellendorffii* were included in the training dataset of Helixer, *Juglans regia* was not. The genome assemblies and reference annotations used for these species were retrieved from Ensembl and are outlined in Supplementary table 7. Protein folds with high structure and low sequence conservation were selected based on work from Sousounis et al. (2012). These were the histone fold, the helix-turn-helix motif, the zinc finger, and the globin-like fold; proteins containing these folds were extracted from the SwissProt database. Structure embedding alignment was performed using FoldSeek (Release 10), the ProstT5 (v0.0.01) model weights, and FoldSeek's prebuilt SwissProt AlphaFold V4 database (Kempen et al., 2023).

Materials for direct protein structure alignment

In DPSA the gene model predictions from Helixer on the GL377565 scaffold of *Selaginella moellendorffii* were used. Biopython (v1.85) was used for CPFE. Similar to SBF, structure embedding alignment was performed using FoldSeek (Release 10), the ProstT5 (v0.0.01) model weights, and FoldSeek's prebuilt SwissProt AlphaFold V4 database (Kempen et al., 2023).

Hardware

The SBF and DPSA pipelines were executed on AMD EPYC 7532 processors with 32 cores.

Results

As indicated previously, current sequence-based annotation pipelines still fail to annotate a significant part of the genome. In order to investigate this annotation gap and the potential use of protein structure information in genome annotation, we created two distinct workflows. The first approach, referred to as "Structure-based filtering" (SBF) focuses on incorrectly predicted gene models by combining existing annotation tools (Fig. 1 B) and protein structure alignment (Fig. 1 D). In this approach, we modified the parameters of existing annotation tools BRAKER3 and Helixer to increase gene model prediction sensitivity. We evaluated SBF using proteins with high structure but low sequence conservation folds, comparing the number of proteins remaining after filtering to those present in the reference annotation. The second approach, called "Direct protein structure alignment" (DPSA), targets gene models missed by existing annotation tools. It combines a custom pipeline referred to as "Candidate protein fragment extraction" (CPFE) (Fig. 1 C) with protein structure alignment (Fig. 1 D). In this approach, we compared the interval locations of structural alignments obtained from DPSA to reference annotations based on varying definitions of true positive matches, including exact matches, overlaps, and predictions within a specified margin. For a visual representation, see Supplementary Fig. 11. Further evaluation of DPSA was conducted by analyzing protein length and functional annotation using Gene Ontology (GO) terms.

Structure-based filtering

Limited gene model sensitivity improvement in BRAKER3 and Helixer

As a first step in assessing the potential of SBF, we examined whether the sensitivity of gene model predictions from BRAKER3 and Helixer could be increased through parameter adjustments. The aim was to determine whether sensitivity could be increased to serve as a baseline for subsequent false positive removal using protein structure alignment. However, as displayed in Fig. 2, in BRAKER3 only a marginal increase in sensitivity of less than 1% across all parameters settings in *Arabidopsis thaliana* and *Physcomitrella patens* was observed, resulting in overlapping markers. Predictions for *Arabidopsis thaliana* performed well, achieving high precision and sensitivity scores, whereas *Physcomitrella patens* showed low sensitivity scores, just below 40%. This pattern was consistent with a group of less extensively studied genomes (*Medicago truncatula, Parasteatoda tepidariorum, Solanum lycopersicum*) in the original BRAKER3 publication that showed similar precision (61.51–75.47%) and sensitivity (44.46–47.16%) (Gabriel et al., 2024). Using the number of NCBI Sequence Read Archieve (SRA) entries as a proxy for how well an organism has been studied, these genomes (331–35,048 entries) align with *Physcomitrella patens* (1986 entries)(National Center for Biotechnology Information, 2025). This indicates a relationship between gene model prediction performance and sequencing data availability, aligning with Gabriel et al. (2024).

Precision and sensitivity performance for Helixer also showed minimal improvements of less than 1% upon parameter adjustment. The predictive performance at gene level for *Arabidopsis thaliana*, *Selaginella moellendorffii* and *Juglas Regia* is visualized in Fig. 3. *Arabidopsis thaliana* had the highest precision and sensitivity of the studied species, followed by *Juglas Regia*, and *Selaginella moellendorffii*, which scored significantly

lower. Notably, Helixer's deep learning approach outperformed BRAKER3 in *Arabidopsis thaliana*, reaching higher precision and sensitivity scores.

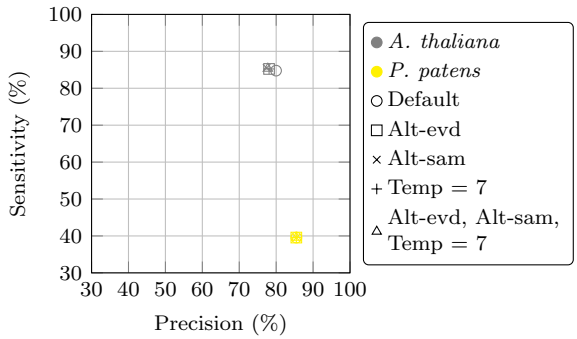


Fig. 2: Gene-level precision and sensitivity for BRAKER3 in *Arabidopsis thaliana* and *Physcomitrella patens* with various parameter configurations: using default settings (Default), alternatives-from-evidence (Alt-evd) = True, alternatives-from-sampling (Alt-sam) = True, Temperature (Temp) = 7 and the combination of Alt-evd = True, Alt-sam = True ,and Temp = 7.

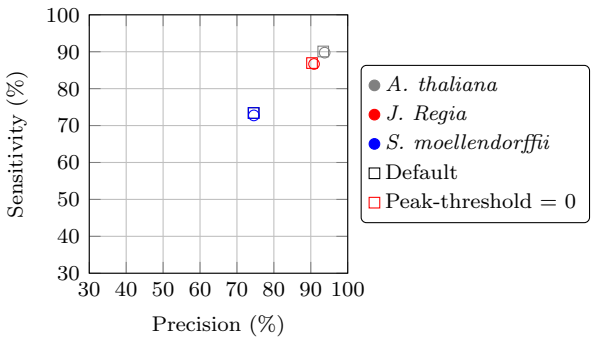


Fig. 3: Gene-Level precision and sensitivity for Helixer across species *Arabidopsis thaliana*, *Selaginella moellendorffii*, and *Juglans regia* under default settings and with peak-threshold = 0.

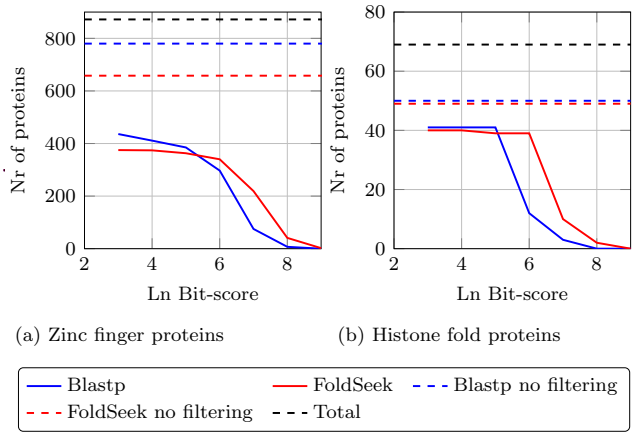


Fig. 4: Number of proteins retained after filtering across different bit-scores with Blastp and different structural bit-scores with FoldSeek in BRAKER3. *Arabidopsis thaliana* proteins containing zinc finger and histone fold motifs selected from SwissProt AlphaFold. The total number of proteins is the number of proteins, with the respective motif, that were identified using Blastp or FoldSeek without bit-score filtering. The x-axis shows the natural log of the bit-scores.

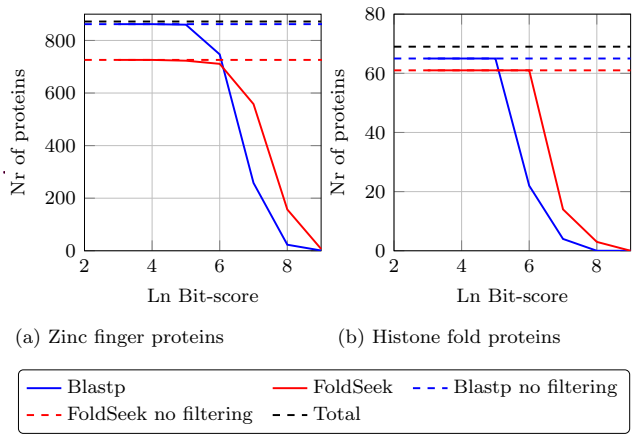


Fig. 5: Number of proteins retained after filtering across different bit-scores with Blastp and different structural bit-scores with FoldSeek in Helixer. *Arabidopsis thaliana* proteins containing zinc finger and histone fold motifs selected from SwissProt AlphaFold. The total number of proteins is the number of proteins, with the respective motif, that were identified using Blastp or FoldSeek without bit-score filtering. The x-axis shows the natural log of the bit-scores.

Structure-based filtering of zinc finger and histone fold proteins

A key assumption of using structure to infer homology is that structural features are likely more conserved throughout evolution than sequence. To assess whether using structure in genome annotation indeed provides advantages over longer evolutionary distances, we investigated a small set of proteins containing folds with high structure but low sequence conservation from *Arabidopsis thaliana*. Given that *Arabidopsis thaliana* contains more than 50 proteins with these folds, the zinc finger and histone fold were chosen as motifs of interest. Proteins containing these folds were identified through structure alignments using FoldSeek based on BRAKER3 and Helixer sequence predictions using default settings. Subsequently, the alignments were filtered at bit-score cutoffs from e^3 to e^9, and compared them to sequence-based filtering with Blastp.

Sensitivity trade-offs in sequence and structure alignment

The total number of proteins retained after filtering using Blastp and FoldSeek at various bit-score cutoff values on BRAKER3 output is plotted in Fig. 4, and on Helixer output in Fig. 5. The number of proteins identified by each approach is indicated by the "no-filtering" line. Blastp consistently identifies more proteins than FoldSeek, with this difference being larger for proteins containing zinc finger folds compared to histone fold. This indicates a higher sensitivity of sequence-based alignments. However, it should be noted that these alignments are based on protein sequence predictions from default BRAKER3 and Helixer runs, given the sensitivity could not be improved. This introduces a bias, as these sequences used for alignment are derived from

sequence-based annotations, which may inflate the sensitivity of sequence-based alignment at low bit-scores.

Furthermore, the number of proteins retained at lower bit-score thresholds is higher for sequence-based alignment (Blastp) compared to structure-based alignment (FoldSeek) for both motifs in the BRAKER3 and Helixer outputs. However, around e^5 an intersection point is observed after which FoldSeek retains more proteins than Blastp. This suggests a trade-off where sequence-based alignment initially identifies more proteins, while structure-based alignment is more sensitive at higher bit-score thresholds. However, the direct comparison of these approaches is influenced by differences in bit-score calibration.

Statistical calibration affecting bit-score comparability

The bit-scores of FoldSeek and Blastp are not directly comparable. FoldSeek's structural bit-score integrates structural similarity scores TM-score and LDDT, but since these values scale between 0-1, they only act as a correction factor. However, the difference in retained proteins may be influenced by the calibration of the statistical significance in each tool, rather than reflecting true differences in alignment sensitivity. Blastp calibrates its bit-scores based on the distribution of random sequence alignments, here significance is measured relative to alignments expected by chance under an extreme value distribution (Pearson, 2014). In contrast, FoldSeek calibrates its bit-scores using the SCOPe40 structural database based on the distributions of TP bit-scores (hits within the same SCOPe superfamily) and false positive bit-scores (hits to another fold) (Kempen et al., 2023). These calibrations are used to estimate the statistical significance of a match, so similar bit-scores in both tools do not correspond to the same level of significance. This discrepancy suggests that the observed difference in retained proteins might be influenced by bit-score calibration rather than actual differences in alignment sensitivity.

Higher bit-scores on Helixer output

It is also noticed that the number of proteins retained in BRAKER3 outputs decreases more after filtering at low bit-scores compared to Helixer, indicating that Helixer predictions achieve higher bit-scores. In addition, the difference between the number of proteins identified and the total number of proteins present in the reference is also smaller in Helixer compared to BRAKER3.

In conclusion, these results indicate that parameter adjustments in BRAKER3 and Helixer provide marginal improvements in gene model prediction sensitivity, leaving little room for further optimization by SBF. In addition, our findings indicate that Helixer outperforms BRAKER3 in predictive performance in *Arabidopsis thaliana*, as well as achieving protein alignments with higher bit-scores. The filtering of zinc finger and histone fold containing proteins indicated that sequence-based alignment is more sensitive at low bit scores, whereas structure-based alignment is more sensitive at high bit-scores. However, this trend may be affected by bit-score calibration differences between Blastp and FoldSeek rather than reflecting true differences in alignment sensitivity.

Direct protein structure alignment

Given the limited potential for annotation improvement using SBF, we explored the possible detection of previously missed gene models. To do so, we applied DPSA to unannotated regions extracted from Helixer predictions on the GL377565 scaffold of *Selaginella moellendorffii*. Moreover, Helixer outperformed BRAKER3 in previous analysis, in both statistical performance and achieved bit-scores, and within Helixer *Selaginella moellendorffii* achieved the lowest precision and sensitivity scores, thereby providing many unannotated regions. The total length of the GL377565 scaffold was 6.95 million bp, of which 3.21 million bp remained unannotated. DPSA was applied to both unannotated and annotated regions, and its performance was compared to using sequence-based alignment instead of structure-based alignment as baseline.

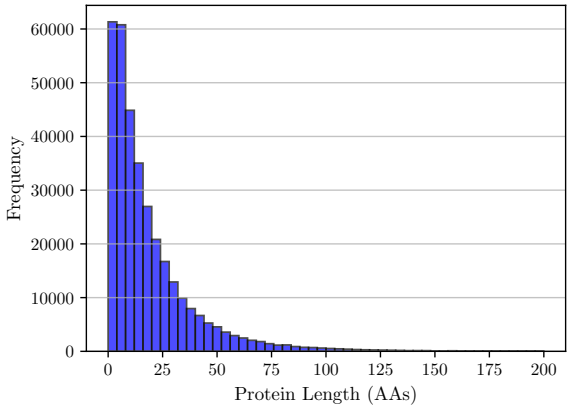


Fig. 6: Size distribution of candidate protein fragments obtained by "Candidate protein fragment extraction" (CPFE) from regions unannotated by Helixer on the GL377565 scaffold of *Selaginella moellendorffii*.

Optimization for single exon-based structure alignment

An initial assessment of DPSA revealed that single exons were already sufficient to obtain protein structure alignments with high bit-scores, ranging from e^5 to e^8 (Supplementary table 8). This indicated that a single exon contains enough structural information for reliable protein identification. Consequently, we implemented a minimal fragment length threshold to filter candidate protein fragments extracted by CPFE, optimizing efficiency while limiting sensitivity loss. The protein fragment length threshold was optimized to prioritize the retention of the largest exons of the genes. The size distribution of candidate protein fragments extracted from regions unannotated by Helixer on the GL377565 scaffold of *Selaginella moellendorffii* without applying minimal length filtering is visualized in Fig. 6.

The size distribution of protein fragment candidates reveals the presence of many short fragments (< 25 amino acids) after CPFE, suggesting the extraction of many non-coding sequences. Given the average regular exon length in *Selaginella moellendorffii* is 207 bases (69 amino acids) sequence length cutoffs between 60 and 100 amino acids were tested (Zhu et al., 2017). Additionally, bit-score thresholds ranging from e^3 to e^9 and different TP definitions (exact match, margin match and overlap match) were evaluated for both sequence, and structure alignment. Complete overview of the performance statistics is provided in Supplementary table 10 to table 39.

Applying DPSA to unannotated regions of *Selaginella moellendorffii* revealed its potential in identifying previously missed gene models. Using structure-based alignment, F1 scores

up to 0.51 were achieved, while sequence-based yielded even higher F1 scores up to 0.56. These scores show the potential of DPSA to recover gene models that were missed by existing annotation tools.

Overlap-based true positive definition

To asses the reliability of these predictions, our different TP definitions need to be evaluated (exact match, margin match and overlap match). The margin definition, in which the interval prediction had to fall within a defined window around the reference interval, yielded the highest F1 scores, 0.51 (structure alignment) and 0.56 (sequence alignment). However, this margin also facilitated the inclusion of false positive observations. For instance, protein predictions for the SELMODRAFT 402209 gene (encoding for the D8QPX7 protein), when applying a margin of 5k bp to the predictions (table 1). It can be seen that half of the predictions do not overlap with the reference annotation. These protein predictions are obtained from amino acid sequences of neighboring non-coding regions and, therefore, do not provide predictive value for the presence of the D8QPX7 protein. This highlights the risk of too lenient TP definitions.

Table 1. True Positive protein structure alignments of *Selaginella moellendorffii* with a margin of 5k bp for TPs, a protein length threshold of 60 amino acids, and a bit score threshold of e^4.

Protein ID	Start	End	Bit Score
***Selaginella moellendorffii* v1.0 INSDC Assembly**			
D8QPX7	1,489,407	1,491,467	-
True positives from predictions			
P96638	1,490,662	1,490,880	86.0
P0DUH1	1,491,248	1,491,541	97.0
Q9V9Y9	1,491,752	1,491,931	67.0
C3N5R8	1,491,925	1,492,251	58.0

On the other hand, the strict exact match definition, similar to what is used in genome annotation tools, may exclude TP predictions that extent slightly beyond the reference interval. For example, the predicted match (P0CT39) for the SELMODRAFT 29411 gene (encoding for the D8QQT1 protein) spans a longer region than the reference (table 2), for a schematic visualization of such an instance see Supplementary Fig. 11 (a). Despite this, both proteins share key GO-annotations on metal ion binding activities, suggesting the prediction is biologically relevant and should be considered as TP (UniProt Consortium, 2025a,b).

Table 2. False Positive protein structure alignment predictions of *Selaginella moellendorffii* with an exact interval match for TPs, a protein length threshold of 60 amino acids, and a bit score threshold of e^4.

Protein ID	Start	End	Bit Score
***Selaginella moellendorffii* v1.0 INSDC Assembly**			
D8QQT1	3,213,312	3,213,551	-
True Positives (TP) from Predictions			
P0CT39	3,213,204	3,213,638	489.0

Given these trade-offs, the overlap TP definition was chosen for further analysis, ensuring both biological relevancy and flexibility of interval matches. A minimal overlap length of 1 bp was required in the experiments, but most reference-prediction overlaps ranged between 200-400 bp, see Supplementary Fig. 9 and Supplementary Fig. 10 for the respective reference-prediction overlap size distribution in Blastp and Foldseek. This suggests that most relevant predictions have an overlap substantially larger than 1 bp, and further optimization of this threshold can improve the biological relevance of predicted intervals.

Limited difference between sequence and structure alignment at high sensitivity

Under the overlap TP definition, the highest F1 score achieved by Blastp was 0.34, obtained using a bit-score threshold of e^4 and protein sequence length cutoffs of 60, 70 and 80 amino acids (AA)(Fig. 7). Among these cutoffs, the highest sensitivity of 55% was achieved at 60 AA. The highest F1 score achieved by FoldSeek was 0.27, obtained using a bit-score threshold of e^5 and a protein fragment length cutoff of 90 AA. However, the primary aim of this approach is to uncover previously missed genes, making sensitivity a critical factor in our evaluation. When applying a minimum 90% sensitivity threshold, the highest F1 score obtained by both tools is 0.26, with Blastp achieving 96% sensitivity, and FoldSeek reaching 98% sensitivity, both at 15% precision and e^4 bit-score threshold. This indicates that, despite minor differences in performance variation across bit-scores, Blastp and FoldSeek perform similarly overall. This, similar to SBF, suggests a difference in the scaling of their precision-sensitivity trade-offs.

The impact of the protein fragment length threshold on the predictive performance of DPSA appears to be limited, as indicated by the highly similar F1 scores across thresholds (Fig. 7). On the other hand, the runtime is largely influenced by the chosen cutoff, with runtimes increasing from 3 to 10 hours in Blastp, and from 40 to 110 in FoldSeek for cutoffs of 100 and 60 amino acids, respectively. The difference in runtimes between Blastp and FoldSeek highlights the computational costs associated with structural embedding computation. A complete overview of all runtimes can be found in Supplementary table 9.

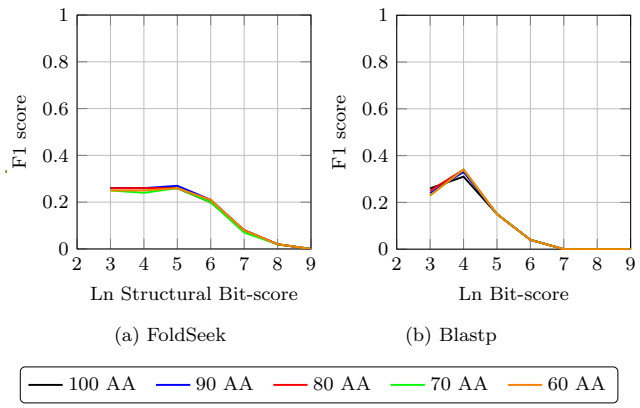


Fig. 7: F1 scores for overlap matches for Blastp and FoldSeek using direct protein structure alignment on regions of the GL377565 scaffold from *Selaginella moellendorffii* which remained unannotated by Helixer, at (structural) bit-score thresholds of e^3 to e^9 and protein fragment length cutoff 100 to 60 amino acids.

Improved performance of Direct protein structure alignment on annotated regions

To gain a comprehensive understanding of DPSA, it is essential not only to compare sequence with structure-based annotations but also to compare the performance on annotated and unannotated regions. By applying CPFE to previously annotated regions, we aim to assess whether DPSA shows a similar prediction performance pattern as existing annotation tools. For this analysis, we used a protein length threshold of 60 AA with a bit-score threshold of e^4, given this setting yielded the highest F1 score in Blastp while maintaining consistent performance in FoldSeek across low bit-score thresholds. The performance metrics of Blastp and FoldSeek on annotated and unannotated regions are summarized in table 3. The results indicate that both methods perform much better on annotated compared to unannotated regions, while still maintaining similar performance, achieving F1 scores of 0.85 (sequence alignment) and 0.82 (structure alignment). This suggests that DPSA experiences similar prediction difficulties in identifying these specific gene models as existing annotation pipelines

Table 3. Performance metrics for overlap matches for Blastp and FoldSeek using direct protein structure alignment on regions of the GL377565 scaffold from *Selaginella moellendorffii* which were annotated and unannotated by Helixer, at a bit-score threshold of e^4 and 60 amino acids protein fragment length cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
Unannotated regions						
Foldseek e^4	1068	6615	15	99.0	14.0	0.25
Blastp e^4	155	466	127	55.0	25.0	0.34
Annotated regions						
Foldseek e^4	11318	4828	4	100.0	70.1	0.82
Blastp e^4	2355	488	361	86.7	82.8	0.85

Difficulties in functional analysis

To better understand the challenges in gene annotation using existing tools and to explore the possible differences between sequence- and structure-based approaches, we analyzed TP protein predictions from DPSA (60 AA protein sequence length cutoff, e^4 bit-score threshold) based on protein length and associated GO terms. Protein length is expected to be correlated with complexity, possibly causing them to be more difficult to annotate (Tan et al., 2005). The size distributions of proteins identified by DPSA employing Blastp (sequence alignment) and FoldSeek (structure alignment) are categorized by identification method and plotted in Fig. 8. The results indicate a slight difference in size distribution, with Blastp identifying relatively longer proteins compared to FoldSeek. However, this trend is likely influenced by the different bit-score calibration in both tools, as noted earlier, where this influences their sensitivity thresholds. At the tested bit-score value, FoldSeek exhibits higher sensitivity (99%), identifying a broader range of proteins fragments. In contrast, Blastp applied a stricter threshold resulting in a lower sensitivity (55%), leading to the identification of relatively longer proteins.

In addition to protein complexity, we explored whether there are functional differences between genes that are better predicted by either sequence- or structure-based approaches. This is analyzed using GO-terms, and may provide insights into

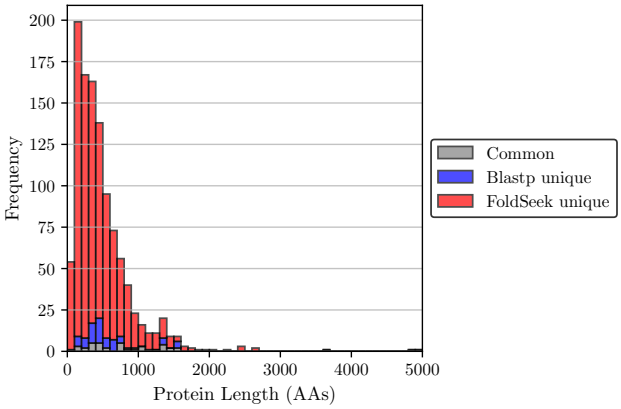


Fig. 8: Size distribution of true positive protein matches identified by direct protein structure alignment (DPSA) on regions of the GL377565 scaffold in *Selaginella moellendorffii* that remained unannotated by Helixer. Predictions are categorized based on their detection method: proteins identified by both Blastp (sequence alignment) and FoldSeek (structure alignment), labeled as "Common", or uniquely by either method.

functional characteristics that influence annotation performance. This was performed using both *Selaginella moellendorffii* proteins on the GL377565 scaffold from UniProt as background, as well as the complete SwissProt database, with the plant-specific GO subset as the selected category. UniProt entries from *Selaginella moellendorffii* were used given UniProt contained 33,386 associated entries, of which only 39 are reviewed and contained in SwissProt. The predicted proteins used in the analysis were obtained from alignments to the SwissProt (AlphaFold) database and categorized based on their identification by both methods or uniquely by either Blastp or FoldSeek. Both backgrounds introduced significant limitations. The UniProt-based *Selaginella moellendorffii* background ensured greater protein-level similarity between the sample and background but was weakly annotated compared to SwissProt, resulting in an uneven enrichment of GO-terms. The predicted proteins had an average of 51.2 GO-terms per protein, compared to 19.8 in the background, resulting in an inflated number of overrepresented specific terms and underrepresented high-level terms. On the other hand, using SwissProt provided a more in depth annotation but introduced a bias due to a lack of protein similarity between the sample and background. This resulted in an excessive number of enriched terms, with 261 enriched significant GO-terms (($p \leq 0.05$ and log_2 fold enrichment greater than 0.5) for just 37 proteins commonly identified by both methods, further complicating interpretation.

Due to these limitations resulting from the sparsely annotated *Selaginella moellendorffii* proteome, the GO-enrichment results are not included in the main report, but the results using the UniProt-based *Selaginella moellendorffii* background can be found in the supplementary materials Fig. 12, 13, 14.

Discussion

We aimed to improve genome annotation by removing incorrectly annotated intervals from predictions using structure-based filtering (SBF) and to improve the annotation of unannotated intervals

by direct protein structure alignment (DPSA). The application of SBF proved to be challenging given limited improvements in annotation sensitivity in BRAKER3 and Helixer. This resulted in a minimal increase in false positive annotations, providing little margin for structure-based filtering. The filtering of proteins containing high structure low sequence folds revealed different sensitivity trade-offs between Blastp and FoldSeek, likely influenced by the different bit-score calibrations of both tools which limits the comparability. DPSA achieved moderate performance in identifying proteins in unannotated regions, with an F1 score of 0.34 for Blastp and 0.27 for FoldSeek. However, it performed notably better on previously annotated regions, achieving F1 scores of 0.85 and 0.82 for Blastp and FoldSeek, respectively. Despite this, little difference between sequence- and structure-based approaches was observed in predictive performance and protein characteristics. Observed differences were likely driven by the tools' distinct bit-score calibrations, which influence sensitivity and comparability. Additionally, functional analysis of the proteins was troubled by limitations in background composition and annotation depth, resulting from the sparsely annotated *Selaginella moellendorffii* proteome.

Restricted precision-sensitivity trade-off BRAKER3 and Helixer

The precision-sensitivity trade-off reflects the balance between the fraction of correctly predicted annotations from the reference annotation, i.e. sensitivity ($\frac{TP}{TP+FN}$), and the fraction of correctly predicted annotations from the predicted annotations, i.e. precision ($\frac{TP}{TP+FP}$). Depending on the application of the predictive model focus can be on the restriction of either false positive (FP) or false negative (FN) observations. Parameter optimization to improve the sensitivity of existing annotation tools BRAKER3 and Helixer, had a negligible effect. Even for organisms with less studied genomes and lower baseline performance, little to no changes in the precision-sensitivity trade-off were observed. The difference in performance evaluation between BRAKER3 and Helixer may have influenced the observed variation. However, as the comparisons were performed within tools, this is expected to have a limited effect. The observations suggest that both models are inherently restricted in their precision-sensitivity trade-off and that this is independent of the applied prediction technique. One possible explanation for this could be that the models are designed to prioritize precision over sensitivity to minimize the risk of incorrect annotations which can be propagated in downstream analyses. In the end, this choice is a design preference. For instance, Gabriel et al. (2024) explains their choice when discussing the BUSCO completeness score of their model by stating, "BRAKER3 weights the trade-off between fewer missing genes and fewer false positive genes in favor of the latter.". Unfortunately, this preference is fixed in both models, limiting the possibility to explore alternative trade-offs. Allowing users to modify this trade-off can be valuable for future exploratory research in genome annotation. For example, a more sensitive setting could be applied for novel gene exploration. On the other hand, a stricter setting could be applied when targeting high-confidence genes. However, implementing such flexibility is likely challenging due to the complexity of these tools.

Sequence versus structure-based protein prediction

We evaluated the performance of sequence- and structure-based alignment in SBF and DPSA using various methods. Analysis of SBF of proteins with high structure low sequence conservation folds indicated that sequence-based alignments were more sensitive at lower bit-score thresholds, whereas structure-based alignments showed greater sensitivity at higher bit-score thresholds. Although FoldSeek appears to be more sensitive at higher bit-scores, this effect is likely influenced by differences in bit-score calibration between the tools, which can cause equal bit-scores to correspond to different levels of significance. DPSA allowed for a more thorough evaluation of the two alignment methods, covering predictive performance and protein size distribution. The direct comparison of BRAKER3 and Helixer resulted in Helixer being used in DPSA. However, BRAKER3's performance metrics are defined stricter, based on exact matches, which likely influenced the comparison, but Helixer also achieved higher bit-scores in SBF. Nonetheless, this choice is expected to have limited impact on the performance and evaluation of DPSA, as Helixer was purely used for the classification of genomic regions.

The predictive performance of DPSA indicated that despite minor differences in performance across bit-scores, aligning with the observations mentioned earlier, Blastp and FoldSeek perform very similarly. In line with comparable prediction performance of both methods, limited difference was observed in the size distribution of the proteins predicted by the methods. However, this observation was likely caused by the equal bit-score thresholds being used in Blastp and FoldSeek. This limitation, introduced by the calibration differences between Blastp and FoldSeek, largely impacted the evaluation of SBF, as no direct comparisons could be made. DPSA overcame this limitation by expressing performance in precision and sensitivity of protein alignments, rather than the number of proteins retained at a specific bit-score. This allowed for a more equal comparison across bit-scores thresholds.

The similarity between sequence and structure-based protein predictions could have various causes. Firstly, it could result from the use of 3Di embeddings to capture protein structure information. Given that the structural alphabet represents only 20 states of geometric confirmations, the representation of more complex protein structures is limited (Kempen et al., 2023). In this way, it may primarily capture more common structural configurations which can also be achieved using sequence alignment. The experimental design also has its limitations. Structural alignment in FoldSeek was performed against the embedded SwissProt AlphaFold database, which only contains well-characterized proteins. As a result, the diversity of structures is limited, and the sequences of these proteins are likely well-studied and also accessible through sequence alignment. This results in overlap between the data sources and could explain the limited observed difference between sequence and structure. Lastly, the translated amino acid sequences in DPSA were split at their stop codons. This practice could make it harder to detect truncated proteins or psuedogenes. The split at stop codons results in incomplete alignments that fail to capture these variations. Partial alignment of these structures could be useful for successful prediction of these gene models to gain a better understanding of them.

True positive definition in direct protein structure alignment

True positive (TP) predictions in genome annotation pipelines are defined in various ways. BRAKER3 computes performance metrics at the gene, transcript, and exon level. At the transcript and exon level, TP annotation predictions are defined as exact

interval matches to the reference annotation, and at gene level as at least 1 match between the predicted alternative transcripts and the reference transcript (Gabriel et al., 2024). Helixer reports base pair level accuracy, where a TP is defined as a match between the predicted class, e.g. CDS, and the annotated class of a specif base (Stiehler et al., 2020). In case of DPSA, protein predictions could be used as hints to guide existing tools rather than to treat them as final predictions. This approach would allow for greater flexibility in defining TP predictions as it enables a more versatile integration of these protein predictions. While this application of hints could, in theory, help refine genome annotation in challenging regions, the results presented here do not indicate a substantial difference between sequence- and structure-based predictions, therefore the practical benefit of this TP flexibility remains uncertain.

Challenges in predicting proteins from unannotated regions

DPSA revealed to experience similar difficulties to existing annotation pipelines when predicting proteins on fragments of unannotated regions compared to those of previously annotated regions. It performed much better on the annotated regions, achieving F1 scores of 0.85 and 0.82 for Blastp and FoldSeek, respectively, compared to 0.34 and 0.27 for the unannotated regions. This indicates the inherent difficulties associated with the identification of the specific gene models in these regions. Unfortunately, a functional analysis using GO-terms was not feasible due to data limitations. In addition, these annotation difficulties may also be caused by other factors, such as alternative splicing variants which can further complicate gene model predictions. In the end, these observations highlight the need for improved annotation strategies to address these difficult gene models.

Future outlook

The results of this study highlight the challenges in genome annotation and the limited potential of protein structure data integration. Current genome annotation tools were shown to be restricted in their precision-sensitivity trade-offs. For future work, it would be beneficial to modify these trade-offs to enable users to adjust parameters based on their specific goals. In addition, further exploration of alternative splicing variants and other complex gene models may be needed to improve genome annotation of currently missed gene models. Regarding the integration of structural data, the results of this study did not indicate a clear added value of structure data over sequence data. However, exploring the use of structure alignments without relying on embeddings, and using more diverse structural databases could improve prediction sensitivity.

Conclusion

This study aimed to improve genome annotation by integrating protein structure information. The refinement of existing annotations using structure-based filtering (SBF) and identification of previously missed gene models using direct protein structure alignment (DPSA) did not indicate an advantage of structure-based compared to sequence-based approaches. However, DPSA showed good performance on annotated regions, and achieved a similar performance pattern to existing annotation tools. Comparison of FoldSeek and Blasp revealed the impact of statistical calibration on comparability and sensitivity thresholds. Functional analysis was limited by the sparse annotation of the *Selaginella moellendorffii* proteome, and restricted this aspect of the project. The restricted precision-sensitivity trade-off in both Helixer and BRAKER3 across tested species highlights the need for more flexibility in annotation tools. Genome annotation remains a challenging practice troubled by inherently difficult to annotate gene models, and our findings suggest that structure information alone may not be able to overcome this. Nonetheless, further refinement of the integration of protein structure data is needed to conclude this, employing structure alignment without the use of embeddings and utilizing more diverse structural databases.

References

Ardell, D. H., Elofsson, A., and Illerga, K. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins*, 77(3):499–508.

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. 28(1):45–48.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Cooper, G. M. (2000). *The Cell: A Molecular Approach*. Sinauer Associates, Sunderland, MA, 2nd edition.

Ejigu, G. F. and Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing.

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., and Stanke, M. (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research*, 34(5):769–777.

Gene Ontology Consortium (2024a). Go basic ontology. Format version: 1.2. This ontology includes various subsets such as ChEBI terms, GO slim terms for various organisms, and terms not to be used for direct annotation.

Gene Ontology Consortium (2024b). Go slim plant ontology. Format version: 1.2. This ontology includes various subsets such as ChEBI terms, GO slim terms for various organisms, and terms not to be used for direct annotation.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Kempen, M. V., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. pages 1–10.

National Center for Biotechnology Information (2025). Sequence Read Archive (SRA) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for

Biotechnology Information; [cited 2025 Feb 23]. Available from: https://www.ncbi.nlm.nih.gov/sra.

Pearson, W. R. (2014). An Introduction to Sequence Similarity (“ Homology ”) Searching. pages 1–9.

Pucker, B., Irisarri, I., De Vries, J., and Xu, B. (2022). Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology*, 3.

Ruperti, F., Papadopoulos, N., Musser, J. M., Mirdita, M., Steinegger, M., and Arendt, D. (2023). Cross-phyla protein annotation by structural prediction and alignment. *Genome Biology*, 24(1):1–21.

Salzberg, S. L. (2019). Next-generation genome annotation: We still struggle to get it right. *Genome Biology*, 20(1):19–21.

Sousounis, K., Haney, C. E., Cao, J., Sunchu, B., and Tsonis, P. A. (2012). Conservation of the three-dimensional structure in non-homologous or unrelated proteins. pages 1–10.

Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M., and Denton, A. K. (2020). Genome analysis Helixer : cross-species gene annotation of large eukaryotic genomes using deep learning. 36(December):5291–5298.

Tan, T., Frenkel, D., Gupta, V., and Deem, M. W. (2005). Length , protein – protein interactions , and complexity. 350:52–62.

Triesch, S., Kindel, F., Kiel, N., Saadat, N., and Ebenhöh, O. (2023). HELIXER – de novo PREDICTION OF PRIMARY EUKARYOTIC GENE MODELS COMBINING DEEP LEARNING AND A HIDDEN MARKOV MODEL.

UniProt (2024). Go annotations and gene product information. Generated by UniProt. This file contains all GO annotations and gene product information for proteins in the UniProt KnowledgeBase (UniProtKB), ComplexPortal protein complexes, and RNAcentral identifiers.

UniProt Consortium (2025a). D8QQT1 - Uncharacterized protein - Homo sapiens (Human). Accessed: 2025-02-10.

UniProt Consortium (2025b). P0CT39 - Beta-defensin 103 - Schizosaccharomyces pombe (Fission yeast). Accessed: 2025-02-10.

Vuruputoor, V. S., Monyak, D., Fetter, K. C., Webster, C., Bhattarai, A., Shrestha, B., Zaman, S., Bennett, J., McEvoy, S. L., Caballero, M., and Wegrzyn, J. L. (2023). Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. *Applications in Plant Sciences*, 11(4):1–17.

Weisseno, K., Gomez, J., Henkel, A., Mir, M., and Rost, B. (2024). Bilingual language model for protein sequence and structure. (November).

Zhu, Y., Chen, L., Zhang, C., Hao, P., Jing, X., and Li, X. (2017). Global transcriptome analysis reveals extensive gene remodeling , alternative splicing and differential transcription profiles in non-seed vascular plant Selaginella moellendorffii. *BMC Genomics*, 18(Suppl 1):1–15.

float

Supplementary materials

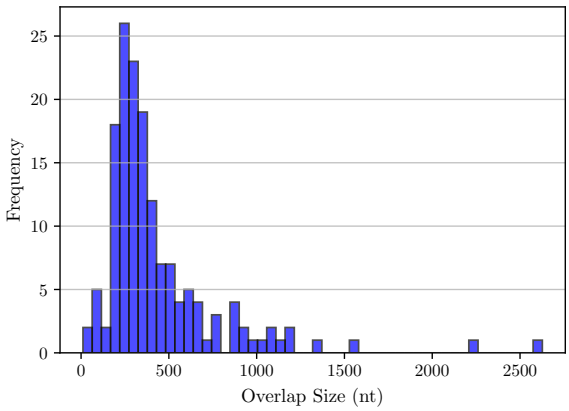


Fig. 9: Overlap size distribution between reference and predicted intervals of TP alignments using Blastp and 60 amino acids as minimal protein candidate fragment length.

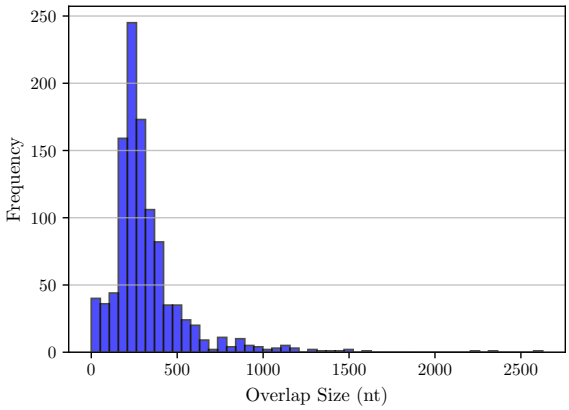


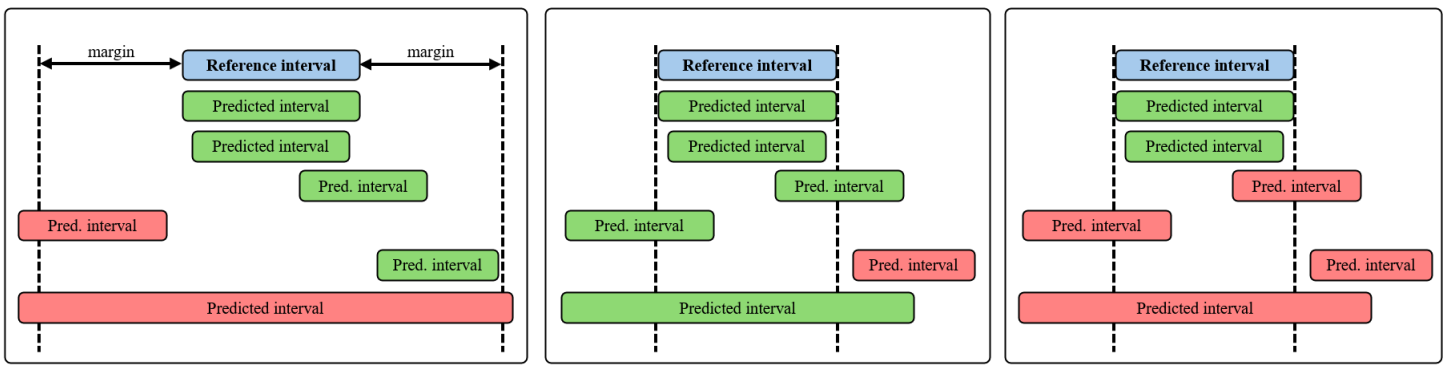
Fig. 10: Overlap size distribution between reference and predicted intervals of TP alignments using FoldSeek and 60 amino acids as minimal protein candidate fragment length.

Table 4. Genome assemblies and reference annotations of species used in BRAKER3 retrieved from EnsemblPlants. Repeat masking and annotation processing according to https://github.com/gatech-genemark/EukSpecies-BRAKER2 and at https://github.com/gatech-genemark/GeneMark-ETP-exp.

Species	Genome assembly file	Reference annotation
Arabidopsis thaliana	GCF_000001735.4_TAIR10.1_genomic.fna.gz	Araport11; Oct, 2024
Physcomitrium patens	Physcomitrium_patens.Phypa_V3.dna_sm.toplevel.fa	Phypa V3; Jan, 2018

Table 5. Species and reference protein files for close relative protein database of species used in BRAKER3 retrieved from EnsemblPlants.

Species	Reference protein file
Arabidopsis thaliana	
Arabidopsis lyrata subsp. lyrata	GCF_000004255.2_v.1.0_protein.faa.gz
Arabidopsis thaliana x Arabidopsis arenosa	GCA_019202795.1_ASM1920279v1_protein.faa.gz
Camelina sativa	GCF_000633955.1_Cs_protein.faa.gz
Arabidopsis suecica	GCA_019202805.1_ASM1920280v1_protein.faa.gz
Capsella rubella	GCF_000375325.1_Caprub1_0_protein.faa.gz
Physcomitrium patens	
Ceratodon purpureus	GCA_014871385.1_CpurpureusR40_1_0_protein.faa
Sphagnum fallax	GCA_021442195.1_S.fallax_v1.1_protein.faa
Marchantia polymorpha subsp. ruderalis	GCA_037833965.1_MpTak2_v7.1_protein.faa
Selaginella moellendorffii	GCF_000143415.4_v1.0_protein.faa
Ziziphus jujuba	GCF_031755915.1_ASM3175591v1_protein.faa



a) Margin match b) Overlap match c) Exact match

Fig. 11: Different definitions of true positive matches between reference and prediction intervals.

Table 6. Short-read RNA-seq libraries of species used in BRAKER3 retrieved from NCBI's Sequence Read Archive.

Species	Short-read RNA-seq libraries
Arabidopsis thaliana	SRR8714016,SRR8759751,SRR4010853,SRR7289569,SRR12547664,SRR12076896
Physcomitrium patens	SRR14784393, SRR14784396, DRR121336, DRR121278, DRR002562, DRR002558

Table 7. Genome assemblies and reference annotations of species used in Helixer retrieved from EnsemblPlants.

Species	Genome assembly file	Reference annotation
Arabidopsis thaliana	GCF_000001735.4_TAIR10.1_genomic.fna.gz	Araport11; Oct, 2024
Selaginella moellendorffii	Selaginella_moellendorffii.v1.0.dna.toplevel.fa	v1.0, INSDC Assembly; Jul, 2010
Juglans regia	Juglans_regia.Walnut_2.0.dna.toplevel.fa	Walnut_2.0, INSDC Assembly; Jul, 2020

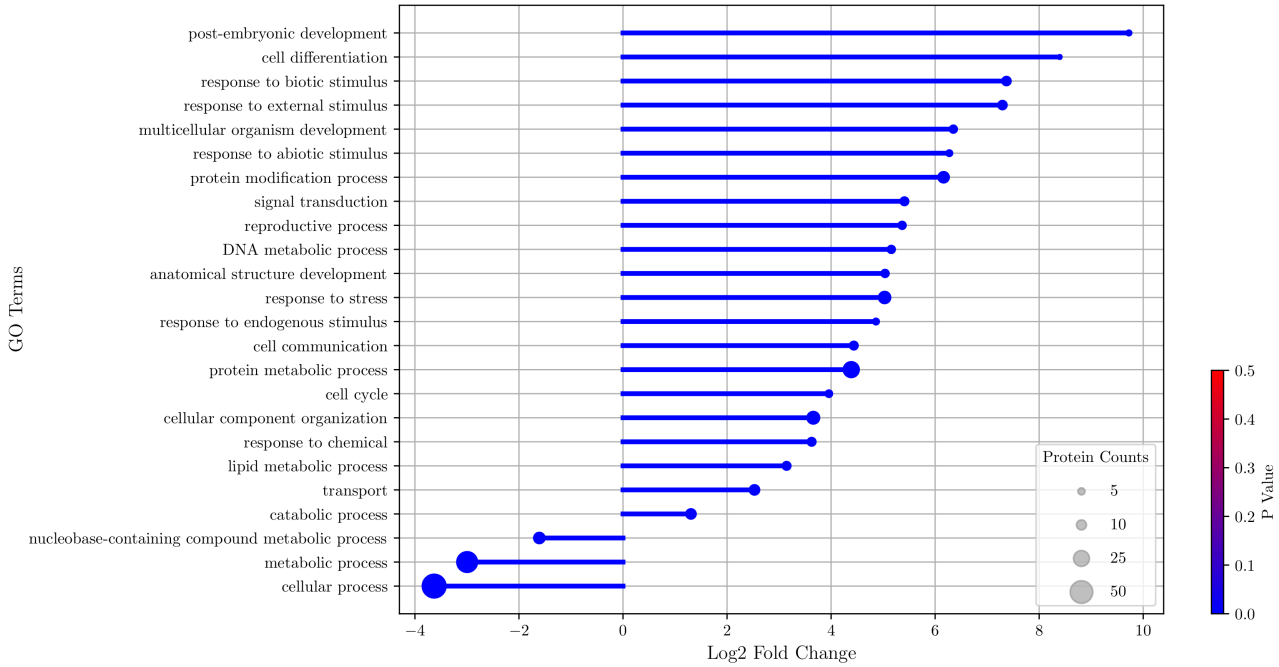


Fig. 12: GO enrichment analysis of proteins uniquely predicted by Blastp in by Helixer unannotated regions of *Selaginella moellendorffii* GL377565 scaffold, *Selaginella moellendorffii* proteome as background including all terms with counts > 1, log_2 fold enrichment greater than 0.5, and $p \leq 0.05$.

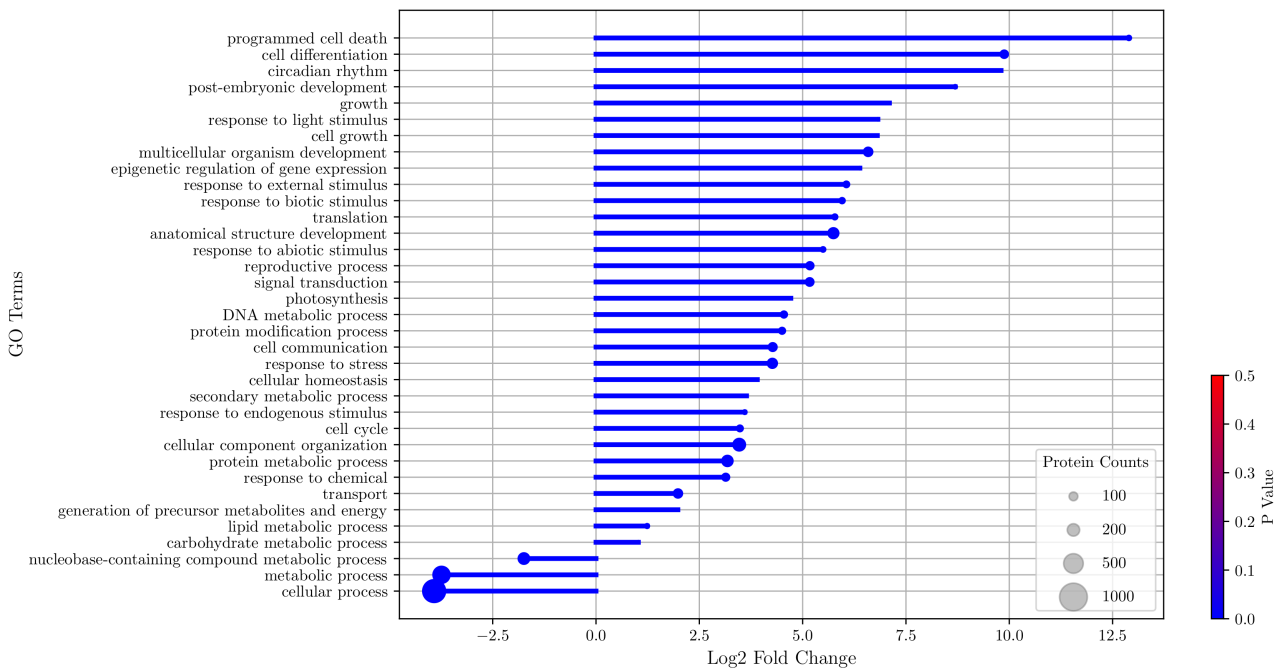


Fig. 13: GO enrichment analysis of TP proteins uniquely predicted by FoldSeek in by Helixer unannotated regions of *Selaginella moellendorffii* GL377565 scaffold, *Selaginella moellendorffii* proteome as background including all terms with counts > 1 fold enrichment greater than 0.5, and $p \leq 0.05$.

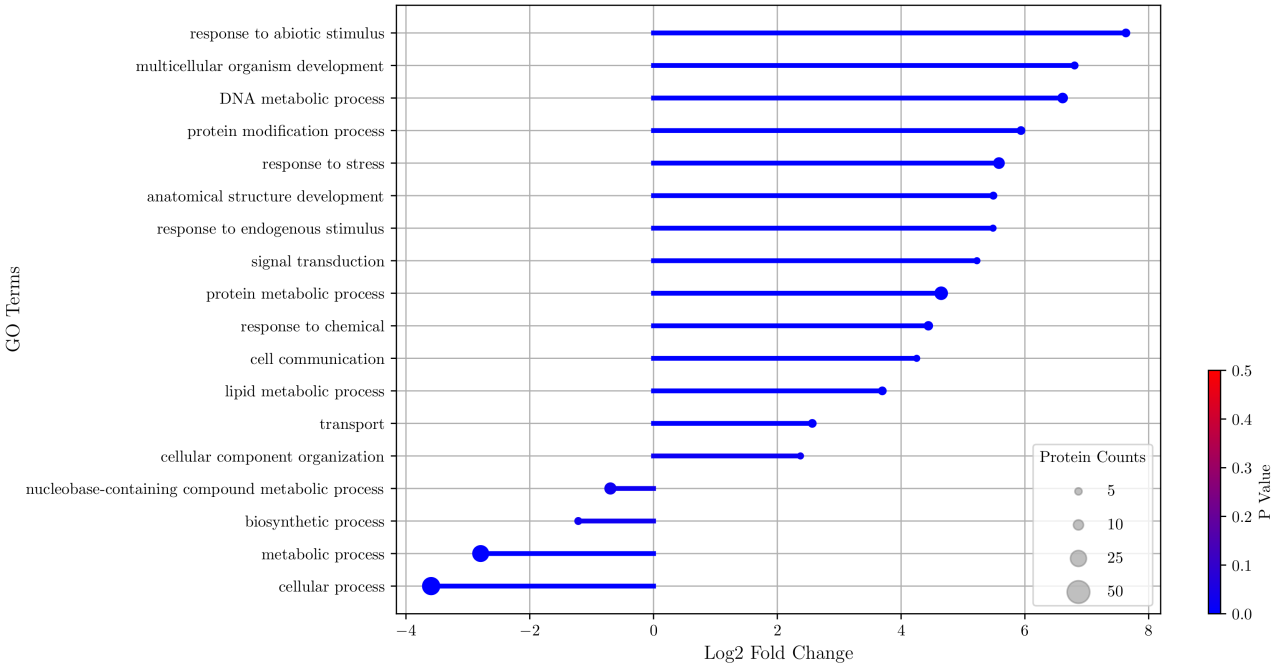


Fig. 14: GO enrichment analysis of proteins commonly annotated by both Blastp and FoldSeek in by Helixer unannotated regions of *Selaginella moellendorffii* GL377565 scaffold, *Selaginella moellendorffii* proteome as background including all terms with counts > 1 fold enrichment greater than 0.5, and $p \leq 0.05$.

Table 8. True positive exon predictions compared to Araport 11 annotation for Arabidopsis (Chr1:1-36000). Predictions were made using FoldSeek (fragment size cutoff: 100, bit score cutoff: e5) with only forward strand reads.

Protein ID	Exon(s)	Start	End	Bit Score
Araport 11 Annotation				
Q0WV96	(6 exons)	2760	5630	-
Q9SP32	(23 exons)	23121	31227	-
True Positives (TP) from Predictions				
Q0WV96	1	3610	3930	242
Q0WV96	2	3992	4339	559
Q9SP32	1	23504	24487	1283
Q9SP32	4	24668	24991	410
Q9SP32	21	30338	30817	777
Q9SP32	9+10	26292	26777	501
Q9SP32	19+20	29160	30317	2022

Table 9. Runtimes of Blastp and FoldSeek at various protein length thresholds.

Protein length threshold	Runtime
BlastP	
100 amino acids	186m9.805s
90 amino acids	279m16.794s
80 amino acids	358m52.455s
70 amino acids	462m40.769s
60 amino acids	618m55.409s
FoldSeek	
100 amino acids	2443m53.124s
90 amino acids	2916m53.650s
80 amino acids	3810m55.209s
70 amino acids	4849m45.994s
60 amino acids	6575m38.404s

Table 10. Performance metrics for exact matches from Blastp at 100 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	445	4245	76	85.0	9.0	0.16
e^4	68	386	164	29.0	15.0	0.20
e^5	12	85	196	6.0	12.0	0.08
e^6	2	11	203	1.0	15.0	0.02
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 11. Performance metrics for margin matches from Blastp at 100 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1526	3164	25	98.0	33.0	0.49
e^4	198	256	105	65.0	44.0	0.52
e^5	36	61	178	17.0	37.0	0.23
e^6	5	8	200	2.0	38.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 12. Performance metrics for overlap matches from Blastp at 100 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	689	4001	29	96.0	15.0	0.26
e^4	111	343	138	45.0	24.0	0.31
e^5	23	74	187	11.0	24.0	0.15
e^6	4	9	201	2.0	31.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 13. Performance metrics for exact matches from Blastp at 90 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	592	5651	61	91.0	9.0	0.16
e^4	79	413	162	33.0	16.0	0.22
e^5	12	85	196	6.0	12.0	0.08
e^6	2	11	203	1.0	15.0	0.02
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 14. Performance metrics for margin matches from Blastp at 90 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	2041	4202	21	99.0	33.0	0.49
e^4	219	273	102	68.0	45.0	0.54
e^5	36	61	178	17.0	37.0	0.23
e^6	5	8	200	2.0	38.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 15. Performance metrics for overlap matches from Blastp at 90 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	892	5351	18	98.0	14.0	0.24
e^4	123	369	135	48.0	25.0	0.33
e^5	23	74	187	11.0	24.0	0.15
e^6	4	9	201	2.0	31.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 16. Performance metrics for exact matches from Blastp at 80 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	820	7766	45	95.0	10.0	0.18
e^4	90	458	158	36.0	16.0	0.22
e^5	12	85	196	6.0	12.0	0.08
e^6	2	11	203	1.0	15.0	0.02
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 17. Performance metrics for margin matches from Blastp at 80 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	2802	5784	18	99.0	33.0	0.49
e^4	248	300	98	72.0	45.0	0.55
e^5	36	61	178	17.0	37.0	0.23
e^6	5	8	200	2.0	38.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 18. Performance metrics for overlap matches from Blastp at 80 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1179	7407	11	99.0	14.0	0.25
e^4	137	411	130	51.0	25.0	0.34
e^5	23	74	187	11.0	24.0	0.15
e^6	4	9	201	2.0	31.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 19. Performance metrics for exact matches from Blastp at 70 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1143	10723	34	97.0	10.0	0.18
e^4	100	487	155	39.0	17.0	0.24
e^5	12	85	196	6.0	12.0	0.08
e^6	2	11	203	1.0	15.0	0.02
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 20. Performance metrics for margin matches from Blastp at 70 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	3873	7993	17	100.0	33.0	0.50
e^4	267	320	98	73.0	45.0	0.56
e^5	36	61	178	17.0	37.0	0.23
e^6	5	8	200	2.0	38.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 21. Performance metrics for overlap matches from Blastp at 70 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1590	10276	5	100.0	13.0	0.23
e^4	149	438	128	54.0	25.0	0.34
e^5	23	74	187	11.0	24.0	0.15
e^6	4	9	201	2.0	31.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 22. Performance metrics for exact matches from Blastp at 60 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1599	15383	23	99.0	9.0	0.16
e^4	106	515	153	41.0	17.0	0.24
e^5	12	85	196	6.0	12.0	0.08
e^6	2	11	203	1.0	15.0	0.02
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 23. Performance metrics for margin matches from Blastp at 60 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	5420	11562	16	100.0	32.0	0.48
e^4	280	341	95	75.0	45.0	0.56
e^5	36	61	178	17.0	37.0	0.23
e^6	5	8	200	2.0	38.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 24. Performance metrics for overlap matches from Blastp at 60 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	2157	14825	4	100.0	13.0	0.23
e^4	155	466	127	55.0	25.0	0.34
e^5	23	74	187	11.0	24.0	0.15
e^6	4	9	201	2.0	31.0	0.04
e^7	0	1	205	0.0	0.0	0.00
e^8	0	0	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 25. Performance metrics for exact matches from Foldseek at 100 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	392	3670	85	82.0	10.0	0.18
e^4	344	3288	90	79.0	9.0	0.16
e^5	78	694	158	33.0	10.0	0.15
e^6	26	239	184	12.0	10.0	0.11
e^7	4	44	201	2.0	8.0	0.03
e^8	0	9	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 26. Performance metrics for margin matches from Foldseek at 100 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1349	2713	26	98.0	33.0	0.49
e^4	1201	2431	28	98.0	33.0	0.49
e^5	262	510	91	74.0	34.0	0.47
e^6	91	174	146	38.0	34.0	0.36
e^7	16	32	190	8.0	33.0	0.13
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 27. Performance metrics for overlap matches from Foldseek at 100 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	617	3445	34	95.0	15.0	0.26
e^4	552	3080	39	93.0	15.0	0.26
e^5	134	638	121	53.0	17.0	0.26
e^6	49	216	166	23.0	18.0	0.20
e^7	10	38	195	5.0	21.0	0.08
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 28. Performance metrics for exact matches from Foldseek at 90 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	510	4639	69	88.0	10.0	0.18
e^4	425	3990	78	84.0	10.0	0.18
e^5	88	738	156	36.0	11.0	0.17
e^6	27	239	184	13.0	10.0	0.11
e^7	4	44	201	2.0	8.0	0.03
e^8	0	9	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 29. Performance metrics for margin matches from Foldseek at 90 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1730	3419	25	99.0	34.0	0.51
e^4	1464	2951	27	98.0	33.0	0.49
e^5	287	539	85	77.0	35.0	0.48
e^6	92	174	146	39.0	35.0	0.37
e^7	16	32	190	8.0	33.0	0.13
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 30. Performance metrics for overlap matches from Foldseek at 90 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	773	4376	25	97.0	15.0	0.26
e^4	656	3759	33	95.0	15.0	0.26
e^5	146	680	118	55.0	18.0	0.27
e^6	50	216	166	23.0	19.0	0.21
e^7	10	38	195	5.0	21.0	0.08
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 31. Performance metrics for exact matches from Foldseek at 80 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	653	5902	54	92.0	10.0	0.18
e^4	528	4815	64	89.0	10.0	0.18
e^5	95	812	151	39.0	10.0	0.16
e^6	26	240	185	12.0	10.0	0.11
e^7	4	44	201	2.0	8.0	0.03
e^8	0	9	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 32. Performance metrics for margin matches from Foldseek at 80 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	2211	4344	21	99.0	34.0	0.51
e^4	1798	3545	24	99.0	34.0	0.51
e^5	318	589	80	80.0	35.0	0.49
e^6	91	175	146	38.0	34.0	0.36
e^7	16	32	190	8.0	33.0	0.13
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 33. Performance metrics for overlap matches from Foldseek at 80 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	952	5603	17	98.0	15.0	0.26
e^4	782	4561	24	97.0	15.0	0.26
e^5	155	752	114	58.0	17.0	0.26
e^6	49	217	167	23.0	18.0	0.20
e^7	10	38	195	5.0	21.0	0.08
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 34. Performance metrics for exact matches from Foldseek at 70 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	843	7600	42	95.0	10.0	0.18
e^4	654	5804	56	92.0	10.0	0.18
e^5	100	853	149	40.0	10.0	0.16
e^6	26	240	184	12.0	10.0	0.11
e^7	3	44	202	1.0	6.0	0.02
e^8	0	9	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 35. Performance metrics for margin matches from Foldseek at 70 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	2846	5597	19	99.0	34.0	0.51
e^4	2170	4288	21	99.0	34.0	0.51
e^5	337	616	78	81.0	35.0	0.49
e^6	91	175	146	38.0	34.0	0.36
e^7	15	32	191	7.0	32.0	0.11
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 36. Performance metrics for overlap matches from Foldseek at 70 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1194	7249	11	99.0	14.0	0.25
e^4	933	5525	20	98.0	14.0	0.24
e^5	162	791	111	59.0	17.0	0.26
e^6	49	217	166	23.0	18.0	0.20
e^7	9	38	196	4.0	19.0	0.07
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 37. Performance metrics for exact matches from Foldseek at 60 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1068	9866	31	97.0	10.0	0.18
e^4	762	6921	48	94.0	10.0	0.18
e^5	105	897	148	42.0	10.0	0.16
e^6	27	240	184	13.0	10.0	0.11
e^7	4	44	201	2.0	8.0	0.03
e^8	0	9	205	0.0	0.0	0.00
e^9	0	0	205	0.0	0.0	0.00

Table 38. Performance metrics for margin matches from Foldseek at 60 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	3618	7316	17	100.0	33.0	0.50
e^4	2547	5136	18	99.0	33.0	0.49
e^5	347	655	78	82.0	35.0	0.49
e^6	92	175	146	39.0	34.0	0.36
e^7	16	32	190	8.0	33.0	0.13
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00

Table 39. Performance metrics for overlap matches from Foldseek at 60 amino acids protein fragment cutoff.

Bit score	TP	FP	FN	Sensitivity(%)	Precision(%)	F1 Score
e^3	1482	9452	8	99.0	14.0	0.25
e^4	1068	6615	15	99.0	14.0	0.25
e^5	167	835	110	60.0	17.0	0.26
e^6	50	217	166	23.0	19.0	0.21
e^7	10	38	195	5.0	21.0	0.08
e^8	2	7	203	1.0	22.0	0.02
e^9	0	0	205	0.0	0.0	0.00