Identification of Meiosis Proteins in Protists via Machine Learning

Qikai Jiang

Introduction

The Significance of Meiosis and Its Study in Protists

According to Page & Hawley (2003), "the separation of homologous chromosomes during meiosis in eukaryotes is the physical basis of Mendelian inheritance," highlighting the fundamental role of meiosis is sexual reproduction. Meiosis is a specialized cellular process reducing the chromosome number by half, transforming a diploid parental cell into haploid gametes. The core of meiosis is the first nuclear division (meiosis I), during which homologous chromosome pair, undergo recombination, and segregate from each other. Chromosome alignment and pairing facilitate homolog recognition, while reciprocal recombination which mediated by sister chromatid cohesion, ensures that homologs are physically connected until their accurate separation. A key feature of this process is the orientation of sister kinetochores to the same pole, which drives accurate chromosome segregation to offspring cells. Meiosis II resembles a mitotic division, in which the sister chromatids of each chromosome are separated into different daughter cells without an intervening DNA replication (Page & Hawley, 2003). Meiosis must be executed with remarkable precision which failures in any of the meiotic chromosome pairing, recombination, or segregation can lead to aneuploidy which is an abnormal number of chromosomes.

As noted in Schurko and Logsdon Jr. (2008), meiosis is not only crucial for reducing ploidy but also serve as a prerequisite of syngamy which by definition is the fusion of gametes. Syngamy restores diploidy by combining gametes to form a zygote, initiating embryogenesis and leading to the next diploid generation with a unique genetic composition. This fashion may occur between gametes from different individual (amphimixis), such as chordates like mammals and birds, or from the same individual (autogamy) in monoecious plants and heterothallic fungi. Without the accurate execution of meiosis, syngamy cannot proceed properly, as the process relies on the formation of balanced haploid gametes.

This demonstrates that meiosis is indispensable not only for maintaining genomic stability across generations but also enabling the fundamental cycle of sexual reproduction.

Building on this centrality of meiosis and syngamy in eukaryotic sexual reproduction, the absence of these processes has been used as defining characteristics of asexuality. As Judson and Normark (1996) explained, asexual are those that lack of meiosis or syngamy, instead of those species producing entirely without sex. Lineages described as "anciently asexual" are those hypothesized to have persisted for millions of years without any observable evidence of sexual reproduction. In classical evolutionary theory which dates to Weismann (1891) and echoed in more recent summary such as Cavalier-Smith (2002), many unicellular eukaryotes, including protists, were assumed to be asexual for a long period. These claims were rooted in the absence of visible sexual cycles or cytological evidence of recombination. However, these views predated the application of molecular phylogenetics and comparative genomics, which have since transformed the understanding of protist biology and challenged the notion that visible sex is required to infer the presence of meiosis.

One of the pivotal studies that marked a turning point in evolutionary biology's understanding of meiosis genes in protists is the work by Ramesh et al. (2005). This study applied comparative genomic approach to *Giardia intestinalis*, a diplomonad protist long believed to be an "ancient asexual." Despite the absence of observable sexual cycles, *Giardia* found to possess conserved meiosis-specific genes, including SPO11, DMC1 and MND1, genes known in homologous recombination and chromosome segregation during meiosis in well-studied sexual eukaryotes such as animals, plants, and fungi. This discovery challenged prior assumptions about asexuality in microbial eukaryotes.

Ramesh et al. (2005) suggests that the presence of meiosis genes should be using a newly adapted approach of phylogenetic analyses and comparative genomics not solely through life-cycle observation. This implies that many so called "asexual" protists may harbor cryptic or rare sexual processes that have remained undetected. Studying meiosis genes in protists uncovers hidden reproductive strategies and helps provide a more complete understanding of sexual evolution across all eukaryotes, not just traditional model organisms like animals, fungi and plants. Expanding the genomic survey of meiotic genes across diverse protists lineages not only enables biologists to identify lineage-specific adaptations and gain a deeper understanding of sexual diversity across eukaryotes but also contributes to resolving unclear branches in the eukaryotic tree of life through phylogenetic reconstructions.

Additionally, Ramesh et al. (2005) noted if *Giardia* represents a basal eukaryotic lineage, studying meiosis genes in protists could help trace the origin and diversification of meiotic machinery back to the last eukaryotic common ancestor (LECA), Koonin (2010). As the article suggests, comparing meiosis gene history across major eukaryotic groups (animals, fungi, plants, and protists), allows identifying which genes are conserved across lineages and which have been lost or modified over time. Such protist analyses help reconstruct how processes like homologous recombination and chromosome segregation evolved, which refines the understanding of eukaryotic evolution (Malik et al., 2007).

Previous Work

Ramesh et al. (2005) conducted a landmark investigation on evolutionary origins of meiosis by focusing on protists. They studied the evolutionary origins of meiosis by identifying core meiotic genes in several protist lineage. They selected well-characterized meiosis-specific genes (e.g., SPO11, MND1, DMC1, HOP1) and meiosis-related genes (e.g., MRE11, RAD52, MSH2, MSH6), as described in Table 1 of the article. To detect these genes in protists (*Giardia*, *Trichomonas*, *Entamoeba*, *Plasmodium*, and *Trypanosoma*), they performed Basic Local Alignment Search Tool for Proteins (BLASTP, Altschul et al., 1990) and Basic Local Alignment Search Tool for Nucleotides (BLASTN) to compare a protein (amino acid) query sequence or a nucleotide (DNA/RNA) query sequence against NCBI or TIGR databases. To identify potential homologs in unannotated genome regions, they used tBLASTn, which aligns protein queries to translated nucleotide databases in all six reading frames (Gertz et al., 2006). In *Giardia*, gene presence was confirmed by cloning and sequencing, with experimental validation.

The results from Table 2 are evident, *Giardia* contains 13 of 17 meiotic genes despite long thought to be an asexual linage. *Trypanosoma* and *Trichomonas* contained substantial subsets of meiotic genes, while *Entamoeba* and *Plasmodium* showed possible absences that the authors attributed in part to incomplete genome assemblies or sequencing limitations. Importantly, gene presence was interpreted conservatively where sequence similarity alone was not sufficient for confirmation. The authors used Bayesian phylogenetic analysis to rigorously assess gene orthology and rule out confusion with paralogous mitotic genes

To investigate the evolutionary conservation of meiosis-related proteins in protists, Ramesh et al. (2005) constructed Bayesian phylogenetic trees for key meiotic genes using the MrBayes software (Huelsenbeck and Ronquist, 2001). Multiple sequence alignments of inferred protein sequences from diverse eukaryotes, including protists, animals, fungi, and plants, as well as archaeal homologs were generated to examine deep evolutionary relationships. For each gene, Markov Chain Monte Carlo (MCMC) sampling was used to explore tree space, and 900–980 post–burn-in trees were retained to compute a consensus phylogeny with posterior probabilities assigned to branches. E.g., the unrooted phylogeny of HOP1, a meiosis-specific protein, revealed its broad conservation across eukaryotes, including protists. RecA family members RAD51 and DMC1, rooted with archaeal RadA, showed early divergence from a shared archaeal ancestor, reflecting their specialization in eukaryotic meiotic recombination, including in protist genomes.

In Schurko and Logsdon (2008), meiosis detection toolkit was introduced formally to build around a curated set of eight conserved, meiosis-specific genes including SPO11, HOP1, HOP2, MND1, REC8, DMC1, MSH4 and MSH5 based on table 1 and table 2 of the article. These genes were selected based on three key criteria relevant to protists: 1) highly specific towards meiosis; 2) span towards diverse functional roles of meiosis I and II through recombination, cohesion and synapsis; 3) evolutionarily conserved across all major eukaryotic groups, including deeply branching protist lineages such as *Entamoeba, Plasmodium, Trypanosoma* and *Giardia* making them ideal markers for probing meiosis.

Protist genomes are often incomplete, unannotated, or poorly assembled. To resolve this, Schurko and Logsdon (2008) make significant improvements from the previous version by using degenerate PCR premiers targeting each of the eight toolkit genes, designed from short and conserved amino acids motifs region (4-6 residues) found in the protein sequences of each meiosis gene (McPherson and Møller, 2006). These conserved motifs enable cross-species amplification even with the absence of full genomes. After amplifying candidate meiosis genes from protist DNA, sequences are translated and screened through BLAST, only hits with high similarity (E-value of smaller than $1e - 5$ or smaller to reduce false positives, percent identity of greater or equal to 30% over aligned regions, 60% of the reference meiotic gene should be covered) appropriate domain architecture (by InterProScan, Jones et al., 2014), and alignment over conserved regions through residue check is retained.

Identical towards Ramesh et al., 2005, due to gene duplications and paralogy, many meiosis genes (e.g., DMC1 and RAD51, REC8 and RAD21) have meiotic homologs. According to figure 2, to confidently identify meiosis-specific orthologs in protists, phylogenetic trees that include known meiosis and mitosis genes from diverse taxa needs to be built. Protist sequences clustered closely with meiosis specific clades should be interpreted as true orthologs. While ambiguous linage may reflect insufficient sequence data, gene divergence, or lineage-specific evolution which needs further study.

The experiment and investigation setting in Schurko and Logsdon (2008) are standardized and polished from Ramesh et al. (2005) with the introduction of degenerate PCR and the consolidate of meiosis-specific gene set. For interpretation, the presence of multiple toolkit genes in a protist lineage supports that it retains the capacity for meiosis. If genes are missing or with ambiguous identity, it does not indicate sexuality. This more standardized approach has been instrumental in re-evaluating supposedly asexual protists (e.g., *Giardia, Entamoeba*) and uncovering once hidden sexual cycles in protists.

Based on Table 3 from Schurko and Logsdon (2008), the meiosis detection toolkit is particularly valuable for studying protists, many lacks complete genome sequences. By targeting conserved amino acid motifs and degenerating PCR, the algorithm can amplify meiosis-specific genes directly from genomic DNA without the access of full proteome. Unlike population genetics that rely on large sample sizes to infer recombination, the toolkit allows researchers to access the potential for meiosis from single-cell which makes it especially useful for rare protists. From an epidemiological perspective, the algorithm helps trace the ancestry of major eukaryotic lineages. In parasitic protists such as *Plasmodium* and *Trypanosoma*, uncovering evidence of meiosis can have implications for understanding transmission dynamics and the evolution of drug resistance.

For disadvantages, PCR failure does not mean a meiosis gene is absent. Proteome might be divergent, or DNA quality or premier match could cause amplification failure, leading to false negatives. The toolkit only detects the presence while not how the gene being transcribed, translated, or functional in meiosis. With obscure annotation of protist sequences, these meiosis genes might remain inactive and no longer used. Many meiosis proteins like DMC1 vs. RAD51 has mitotic paralogs. In protists with high sequence divergence, it's hard to resolve orthology without intense use of phylogenetic analysis. Degenerating PCR on divergent protist sequences may yield very ambiguous results while typical short sequences requirement may be insufficient for confident identification. Finally, protist samples can be contaminated with DNA from fungi, algae or other eukaryotes. Amplification of foreign meiosis genes could produce misleading results unless proper negative controls were used.

Advantage of Machine Learning for tackling this project

Machine learning (ML)-based approaches to meiosis gene detection offer quite a few advantages, particularly when applied towards understudied and divergent eukaryotic lineages such as protists. One major advantage of ML models is their ability to leverage large-scale, high-dimensional feature sets including protein language model embeddings or evolutionary scale embeddings (ESM-2, Lin et al., 2023), amino acid composition (Chou, 2001), and physiochemical properties (Dubchak et al., 1995). These features enable the model to learn subtle structural patterns that might be missed by classical methods like only using BLAST. In contrast, toolkit-based methods typically rely on conserved motifs such as RecA-like fold in DMC1, which will fail to detect divergent homologs.

ML-based methods could identify functionally conserved but sequence-divergent orthologs. While degenerating PCR and homology searches often miss non-canonical meiosis proteins, especially in those deeply branching protists, ML classification trained on a broad range of positives and negatives can learn to generalize across evolutionary distance, recognizing meiosis-specific features even when primary sequence identities are very low (Bepler & Berger, 2021). Typically, toolkit-based methods also only recognize those of meiosis genes based on the algorithm setting, further limit the generalizability. This ability of ML-based models is important in protists, where gene losses and divergence can obscure standard homology signals.

Furthermore, ML models can be applied across entire proteomes, enable comprehensive scanning of all proteins in protists like *Plasmodium*, *Trypanosoma*, or *Entamoeba*. This contrasts with the gene specific detection for toolkit-based methods, where applying PCR amplification and phylogenetic tree are time-costly, labor-intensive and limited in throughput. Because of the scalability of those ML-based methods, they are suited for genome-wide studies that seek to uncover unannotated meiosis-related loci.

Another advantage of ML is its ability to incorporate multiple data modalities, especially for predicted structures (e.g., alpha-fold predictions, Jumper et al., 2021). These flexible structures allow ML-based classifiers to form advanced representations of protein function, way beyond domain-based annotations toolkit-based methods can be applied to.

Lastly, ML-based classifiers produce probabilistic outputs or confidence scores for predictions, providing quantified estimates of how likely a protein is to be meiosis-related. This is critical with ambiguous classification. Toolkit-based methods only yield binary results for detected which makes it much harder to reflect proteome's biological complexity.

Methods

Data

To build effective machine learning models for detecting meiosis-specific proteins, we identified meiosis-specific loci with well-conserved presence across eukaryotes. These genes are known for their roles as fundamental markers for toolkit-based models (Schurko and Logsdon, 2008).

SPO11 is a universally conserved gene that encodes a topoisomerase-like enzyme responsible for initiating meiotic recombination by generating double-strand breaks (DSBs) in DNA which are essential for homologous chromosome interactions, facilitating alignment during prophase I (Keeney et al., 1997).

DMC1 (Disrupted Meiotic cDNA 1) which is a recombinase to bacterial RecA, crucial for homolog pairing and strand invasion during meiotic recombination (Bishop et al., 1992). Its paralog RAD51, active in mitosis majorly, shares similarity but functions primarily in general DNA repair.

MND1 as nuclear division 1 is required for efficient homologous pairing and recombination during prophase I. The protein functions in complex with HOP2 while enhancing DMC1 and RAD51 by promoting strand invasion and stabilizing recombination intermediates (Pezza et al., 2007).

MSH4 and MSH5 are meiosis-specific members of the MutS homolog (MSH) family which are known for DNA mismatch repair. They form a unison that stabilizes recombination intermediates while promoting crossover between homologous chromosomes (Hollingsworth & Brill, 2004).

REC8 is a kleisin subunit of the cohesin complex, critical for maintaining sister chromatid cohesion and proper homolog separation during meiosis I. Its paralog, RAD21, serves similarly for cohesion during mitosis and meiosis II. REC8's specialized function in meiosis cannot be compensated by RAD21 (Parisi et al., 1999).

These six loci represent key functional categories in meiosis including recombination initiation (SPO11), strand invasion and pairing (DMC1, MND1), recombination resolution (MSH4, MSH5), and sister chromatids cohesion (REC8). We collected all known meiosis-specific sequences for these genes across eukaryotic lineages. 100 arthropods and 100 chordates are included, along with 80 from other animal groups such as mollusks, cnidarians, echinoderms, and nematodes, reflects Table 1 for distribution of genes in the meiosis detection toolkit among representative eukaryotes of Schurko and Logsdon Jr. (2008) for *Homo sapiens*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. For chordates, we collected approximately 35–40 mammalian species, with a balanced representation of fish, lizards, snakes, amphibians, sharks, and other groups.

We also included 100 fungal species, ensuring representation of diverse taxa such as *Saccharomyces*, *Schizosaccharomyces*, *Aspergillus*, and *Cryptococcus*. The 120 plant species sampled include eudicots (e.g., *Arabidopsis thaliana*), monocots (e.g., *Oryza sativa, Zea mays*), and some representatives from green or red algae to represent *Cyanidioschyzon*, though algae sequences remain very sparse from the database with 10-15 per loci. All sequences were retrieved through the National Center for Biotechnology Information (NCBI) database.

A significant proportion (>50%) of the collected meiosis-related proteins in arthropods, other animals, and fungi are homologs of the six core loci. Among chordates and plants, the sequences are more likely to represent orthologs, also in some cases, gene paralogs were also collected. To ensure sufficient sample size of true loci orthologs, we collected meiosis-related sequences from a total of 600 species.

Table 1 summarizes the number of training sequences for each core meiosis gene across the sampled eukaryotic taxa. We observe that many plant species contribute multiple sequences per species. Notably, MSH5 exhibits a striking example of sequence redundancy: although only 120 plant species were sampled, 631 sequences were retrieved, indicating an observable imbalance in sequence count across eukaryotic lineages and loci.

Table 1. Training Sequence Counts of Core Meiotic Proteins Across Eukaryotic Lineages

Locus	Arthropods	Chordates	Other animals	Fungi	Plant
SPO11	140	239	116	102	336
DMC1	121	184	86	100	179
MND1	120	248	102	101	178
MSH4	152	190	111	100	368
MSH5	156	257	129	101	631
REC8	168	198	104	101	271

For non-meiosis proteins, we collected ribosomal proteins which are core of ribosome, comprise both small (RPS3) and large (RPL10, RPL23A) subunits which are essential for translation in all cells (Wilson and Doudna Cate, 2012). These proteins serve as a baseline for non-meiotic genes.

Heat shock proteins, including HSP70 for protein folding, HSP90 for stabilization, and small HSPs for stress response, are molecular chaperones that might be strongly upregulated during stress conditions (Lindquist and Craig, 1988).

Cytoskeletal proteins, which are structural and motor proteins necessary for cell shape, transport, and division, include actin (filament-forming protein), tubulin (microtubule subunit), and motor proteins such as kinesin and myosin (Lodish et al., 2016). While these proteins play roles in mitosis and cytokinesis, they are not uniquely required for meiosis.

Metabolic enzymes are ubiquitous in energy metabolism and are unrelated to reproductive processes (Berg et al., 2002). We included GAPDH, enolase (glycolytic enzymes), hexokinase (initiates glycolysis), ICDH (a TCA cycle enzyme).

Translation factors are key regulators of mRNA translation and ribosome function (Jackson et al., 2010). They are essential for protein synthesis. EEF1A delivers aminoacyl-tRNAs to the ribosome, EEF2 raises ribosomal translocation, EIF4E initiates translation by binding, and EIF3 is a large multi-subunit complex that helps assemble the initiation.

These protein families are chosen as negative controls in the analysis because they mostly are not upregulated during the meiotic process while evolutionary conserved across all eukaryotes. Like meiosis genes, these non-meiosis genes are broadly expressed, making them help reduce false positives.

Table 2 summarizes the non-meiosis control proteins included in the training set. We selected 40 species from animals, 25 from fungi and 25 from plants for protein families mentioned. This distribution mirrors the balanced representation used for different species in Table 1 of Schurko and Logsdon Jr.

(2008). All proteins of category and family are listed with corresponding counts. Details and exceptions of these proteins are listed in the "notes" column.

Control protein families include multiple paralogs and isoforms, especially in animals, contributing to variable sequence counts. Kinesin and myosin show significant expansion in chordates due to multiple paralogs, whereas other proteins in animals are much more conserved which are mostly presented in single copy. As denoted *, for protein families with the most significant sequence count imbalance, top 3 of each species were included in the downstream analyses to ensure a more balanced non-meiosis class distribution.

Table 2. Sequence Counts of Non-Meiosis Control Protein Families Across Animals, Fungi, and Plants

Category	Protein Family	Animal (Count)	Fungi (Count)	Plants (Count)	Notes
Ribosomal Proteins	**RPS3**	45	25	26	Core small subunit protein
	RPL23A	44	25	27	Core large subunit protein
	RPL10	49	26	25	Core large subunit protein
Heat Shock Proteins	**HSP70**	54	26	27	Canonical heat shock family
	HSP90	44	25	37	Canonical heat shock family
	Small HSP	62	25	28	Includes diverse small HSPs (e.g., HSP20 in plants)
Cytoskeletal Proteins	**Actin**	46	25	25	Multiple paralogs in all taxa; highly variable
	Tubulin	52	25	28	Includes α-, β-tubulins; variable expansion
	Kinesin	106 (78*)	25	55 (43*)	KIF5B collected for chordates while kinesin heavy chain for other animals
	Myosin	364 (82*)	25	82 (43*)	Includes only heavy chains for animals; varying paralogs
Metabolic Enzymes	**GAPDH**	49	25	29	Generally single copy, conserved
	Hexokinase	88	28	30	Multiple isoforms per genome in animals
	ICDH	60	27	40	Includes NAD- or NADP-dependent forms with varying species
	Enolase	79	26	29	Multiple paralogs in animals and fungi
Translation Factors	**EEF1A**	50	27	35	Core elongation factor
	EEF2	44	25	27	Core elongation factor
	EIF4E	74	26	36	Cap-binding protein
	EIF3	53	27	30	Multi-subunit included

* Only the top three protein families with significant class imbalance (based on sequence counts) were selected for downstream analyses.

Table 3 presents the proteome data for selected protist species used in the post-training validation of the machine learning classification model. These species were chosen based on their inclusion in Table 1 of Schurko and Logsdon Jr. (2008), and their complete proteomes were obtained from the UniProt database. Duplicate protein entries were removed before analysis.

The inclusion of *Entamoeba histolytica*, *Plasmodium falciparum*, and *Trypanosoma brucei* as validation species enables a robust evaluation of the generalizability of the machine learning model across divergent eukaryotic lineages. These protists represent supergroups of Amoebozoa, Alveolata, and Excavata, each with reduced or divergent genomes that challenge homology-based approaches for detecting meiosis genes. All three are well-studied pathogens with well-annotated proteomes. Also, prior research on meiosis-related genes in these organisms provides a valuable reference for comparing predicted candidates. Lastly, the three's meiotic potential is uncertain, e.g. *Entamoeba*, where canonical meiosis genes are either missing or highly diverged, which is ideal test cases for assessing ML-based prediction methods.

"Unqualified" in Table 3 refers to duplicate sequences by species, sequences containing more than 5% ambiguous amino acid letters (X, B, or Z), or sequences significantly shorter than 50 amino acids.

Entamoeba (Entamoeba histolytica) is an anaerobic, intestinal protozoan parasite responsible for amoebiasis, a disease affecting 40–50 million people worldwide and a major cause of morbidity. It was the first human-infecting amoeba to have its genome fully sequenced (2005). After unqualified removal, 7,883 protein sequences remain for analysis.

Plasmodium (Plasmodium falciparum) is the most lethal malarial parasite in humans, responsible for over 90% of global malaria-related deaths. Transmitted by female Anopheles mosquitoes, its notable for its extremely low GC content (<20%). After filtering, 5,355 protein sequences are available for analysis.

Trypanosoma (Trypanosoma brucei brucei) is a protozoan parasite that causes African trypanosomiasis (sleeping sickness in humans and nagana in animals), endemic to 37 Sub-Saharan African countries. It is transmitted by the tsetse fly. We' one of the subspecies of Trypanosoma brucei. After unqualified duplicates, 8,478 protein sequences are retained.

Table 3: Proteome Sequence Counts and References for Selected Protist Species

Protist Name	Total Sequences	After Unqualified Removal	Reference
Entamoeba	7966	7883	UniProt Proteome: UP000001926
Plasmodium	5361	5355	UniProt Proteome: UP000001450
Trypanosoma	8587	8478	UniProt Proteome: UP000008524

Data Pre-processing

We collected individual species .faa files from NCBI and combined them into aggregated .faa files for each meiosis locus and each non-meiosis protein family.

We chose to disregard the accompanying .jsnl annotation files due to the complexity involved in parsing and matching them back to .faa sequences. Other potential concerns include the lack of standardized annotation formats and the frequent absence of annotations across rare species (Radivojac et al., 2013). Moreover, recent machine learning practices have shown a relatively strong performance with the sequence-only approach (Rives et al., 2021).

Next, we removed duplicate sequences from the combined .faa files for both meiosis and non-meiosis proteins, as well as the three protist proteomes.

To address extreme taxonomic imbalance among non-meiosis protein families, only the top three sequences per species in the highly imbalanced myosin and kinesin families from animal and plant taxa.

Evolutionary Scale Modeling

To represent protein sequences numerically without annotation, we used Evolutionary Scale Modeling (ESM-2) to extract informative representations for protein sequences for classifying meiosis and non-meiosis proteins. ESM-2 is a transformer-based protein language model developed by Lin et al., 2023. Using deep learning, the model understands and predicts protein structure and functions directly from primary amino acid sequences.

According to Lin et al., 2023, ESM models work directly on raw amino acid sequences and do not need multiple sequence alignments making it extremely suitable for our only protist protein sequences on hand. ESM learns patterns across evolutionary scale data, capturing features relevant to biological function, especially useful for poorly annotated taxa like rare AFP and protists. While the whole model is based on transfer learning which is already trained on diverse proteins from UniProt and UniRef which is significantly helpful to generalize towards all eukaryotic lineages including protists. As also mentioned in the article, ESM-2 can predict accurately on protein structures with no detectable homologs through structure similarity not sequence similarity. This fit perfectly with under-annotated protist proteins which are highly divergent. The scalability of transfer learning is another knock-off, this leads to a much faster structural prediction than alpha-fold, enable it to perform fast on whole proteome. The transfer learning model was used in structure prediction of more than 617 million proteins which shows impressive generalizability and its good function in exploring unknown proteome space.

ESM-2 is comparatively faster than ESM-1b (Rives et al., 2021), for its improvements in deep learning architecture, optimized training parameters, and enhanced parallelization. In addition, ESM-2 captures richer biological and structural features which include atomic-level structural information, result in more informative representations.

We used the esm2_t30_150M_UR50D for ESM model building for all meiosis, non-meiosis and protists, which contain 30 transformer layers and 150 million parameters. The output of this is a fixed-size 640-dimensional embedding for each protein, extracted from the classification (CLS) token of the final layer. The CLS embedding serves as a global summary of sequences with contextual and structural information being incorporated for later classification. Compared to larger models, this offers a balance between computational cost and richness in representation, suitable for a few thousand samples of meiosis and non-meiosis sequences need to train and the environment of Google Colab.

To accelerate computation, we implemented a parallelized batch structure using Python's multiprocessing. Batch size is flexible based on sample size of meiosis, non-meiosis and protists. This stage of the pipeline avoids the use of BLAST or alpha-fold, as only sequence-based training is currently implemented.

Amino-acid features

To build an interpretable baseline for classifying meiosis versus non-meiosis and later predicting meiosis genes in protists, we selected a comprehensive set of amino-acid sequence derived features based on established bioinformatics literatures which are listed in the "Reference" column in Table 4. These 588 features altogether capture multiple levels of information about protein sequences including local composition and order (Dipeptide Composition, PseAAC), global structural tendency (θ correlation, Shannon entropy, CTD), physicochemical properties (10 properties, Z-scale values). Importantly, all features are computable directly from sequences and all are sequence length invariant, making them ideal for machine learning and a direct comparison to 640 ESM's CLS embedding computed previously. We do not desire to use BLAST or alpha-fold for model training which is another reason is sequence-derived features are critical in this pipeline.

We chose these features as they provide a comprehensive representation of protein sequences especially which may detect the difference between meiosis and non-meiosis:

Dipeptide Composition captures the frequencies of all 400 dipeptide pairs, enabling detection of conserved local motifs that may underline functional domains of meiosis-related proteins even in divergent sequences.

Pseudo Amino Acid Composition extends basic amino acid composition towards sequence-order effect using correlation to memorize both composition and residue order information, which are essential for proteins involved in complex interactions like homologous recombination (Bhasin and Raghava, 2004). Given the redundancy between basic AAC and the former two, we excluded basic AAC to reduce feature overlap.

Sequence-order-correlated factors (θ) are to capture long-range correlations between residues that are separated by a lag of lambda to reflect folding tendencies and structural constraints. This is relevant for meiosis-specific proteins as many functions within large multi-protein assemblies or are in dynamic interactions during homologous recombination (Handel and Schimenti, 2010).

10 average physicochemical properties capture overall chemical profile and biophysical tendencies of proteins that in these meiosis proteins may exhibit conserved functional chemistry (Ramesh et al., 2005).

Shannon Entropy, which is the degree of sequence variability which can distinguish ordered or disordered regions. Meiosis protein may contain intrinsic disordered regions (Brown et al., 2002). This is a must kept feature after feature selection since the entropy does not heavily correlate with other features here.

Composition, Transition, and Distribution (CTD) features capture how amino acid residues are distributed across seven key physicochemical properties. These features provide view for sequence include overall composition, transitions between residue types, spatial distribution patterns along the sequence. CTD features are sensitive to whether a protein adopts a globular structure (enzymatic proteins of meiosis) or a more extended conformation (not specific to meiosis) (Dubchak et al., 1995). Additionally, CTD features can help detect localized sequence motifs involved in meiosis-specific processes such as DNA binding, homologous recombination, and chromosome cohesion (Malik et al., 2007).

Z-scale value which is the PCA descriptor which is valuable in distinguishing functionally specific meiosis-related proteins from general non-meiosis housekeeping proteins.

Table 4. Summary of Amino Acid Sequence-Based Features Used for Protein Characterization

Features	Counts	Description	Reference
Dipeptide Composition (DPC)	**400**	Frequency of all possible contiguous amino acid pairs (e.g., AA, AC, ..., YY), capturing local sequence-order information.	Bhasin, M., & Raghava, G. P. S. (2004)
Pseudo Amino Acid Composition (PseAAC) – Type I	**20**	Extension of traditional amino acid composition incorporating sequence-order information via physicochemical correlation factors.	Chou, K.C. (2001)
Sequence-order-correlated factors (θ)	**5**	Global descriptors measure correlations of amino acid properties between residues at defined sequence lags, based on PseAAC λ-correlation.	Chou, K.C. (2005)
Average Physicochemical Properties of Amino Acids	**10**	Average values of key physicochemical properties (Hydrophobicity, Hydrophilicity, Side Chain Mass, Van der Waals Volume, Polarity, Polarizability, Relative Solvent Accessibility, Flexibility, Isoelectric Point, Bulkiness) calculated across the entire amino acid sequence.	Kawashima & Kanehisa (2000)
Shannon Entropy	**1**	Entropy of amino acid distribution within the sequence, reflecting sequence variability.	Sander & Schneider (1991)
Composition, Transition, Distribution (CTD) Features	**147**	Describe residue distribution patterns across 7 physicochemical properties (Hydrophobicity, Polarity, Charge, Solvent Accessibility, Polarizability, Van der Waals Volume, Secondary structure tendency) using:	Dubchak et al. (1995)

		Tiered Composition (C1–C3): Frequency of residues in three property-based classes (3×7 = 21). Pairwise Transitions (T12, T13, T23): Class-to-class transitions (3×7 = 21). Distance-based Distributions (Dxxxx): Positional spread at 5 percentiles across 3 classes (5×3×7 = 105).	
Z-scale Values	**5**	Standardized quantitative descriptors of amino acids derived from principal component analysis (PCA) of multiple physicochemical properties, captures hydrophobicity, size, polarity, and electronic effects.	Sandberg et al. (1998)

Training and Test Split, Feature Selection

We constructed six binary classification datasets, each corresponding to one meiosis-specific protein: SPO11, DMC1, MND1, MSH4, MSH5, REC8. For each dataset, known sequences of the meiosis protein were labeled as 1, and non-meiosis control proteins were labeled as 0. These datasets are composed of only annotated eukaryotic proteins and do not include protist sequences at this stage.

For each of the six meiosis loci, we prepared two feature sets: 1): ESM CLS token embeddings, 2): Manually engineered amino acid-based features. This resulted in 12 datasets (6 loci × 2 feature types).

To avoid data leakage, we split each dataset into training and test sets in an approximately 70:30 ratio, ensuring that all sequences from the same species are assigned entirely to either training or test set to avoid data leakage (James et al., 2013).

We standardized the features in the training set to have zero mean and unit variance. The test set and protist data were standardized using the mean and standard deviation of the training set, maintaining consistency in feature scaling for fair model evaluation and later prediction.

We then perform feature selection on the training set using approximate Minimum Redundancy Maximum Relevance (approximate mRMR), as proposed by Peng et al. (2005). The mRMR algorithm selects features that are highly relevant for target class (meiosis vs. non-meiosis) and minimally redundant with respect to one another, based on mutual information (MI).

This approach is very satisfying for biological classification problems where feature spaces are high-dimensional, and features are highly correlated. Both the amino acid derived features and ESM's CLS embeddings contain hundreds of potentially overlapping features with both feature spaces being greater than 500.

Using mRMR helps to identify a subset of features that are strongly informative for meiosis-related proteins while reducing redundancy, which is crucial to avoid overfitting, improve generalization, especially targeting the divergent protist proteins later. This results in selected features of biologically meaningful and machine learning effective.

To make the algorithm scales better and cost efficient for our dataset of larger than 500 features. The algorithm for ESM's CLS embedding is the following:

Correlation-based Clustering: We compute the pairwise correlation matrix among features and derive a distance matrix (1 - |correlation|) to measure redundancy. Features are then grouped into clusters using agglomerative clustering based on this distance.

Mutual Information Scoring: For each feature, we calculate mutual information (MI) with the class labels (meiosis vs. non-meiosis), which measures individual predictive power.

Redundancy Reduction: From each cluster, we select features with the highest MI score, ensuring selected features are both informative and not redundant.

Top-k Feature Selection: Lastly, we rank the selected features by MI and retain the top k features for downstream.

For the amino acid derived-features, we incorporate biological structure by performing clustering within predefined groups to minimize within-group redundancy before global selection: Dipeptide Composition (DPC), Pseudo-Amino Acid Composition (PseAAC Type I), Sequence-Order-Correlated Factors (θ), Average Physicochemical Properties, Shannon Entropy, CTD Features, and Z-scale values. Importantly, Shannon Entropy will be included, even if it is not among the top k features due to its low correlation with other amino acid features and its biological relevance.

While we keep cluster number equals feature number, we select top 50 and top 100 for the ESM's CLS embedding while top 50 + Shannon entropy and top 100 + Shannon entropy for amino acid-based features based on approximate mRMR.

For the test and protist datasets, we retain the features selected from the training set to ensure consistency and applicability of the machine learning model. This results in training and test sets with either 50 or 100 features for each of the six loci, yielding a total of 24 training sets and 24 corresponding test sets (6 loci × 2 feature counts × 2 feature types (CLS embedding vs. amino acid features)).

Support Vector Machine with Radial Basis Function Kernel

We adopt Support Vector Machines (SVM) as the primary classification model for distinguishing between meiosis and non-meiosis, based on both the nature of protein generated features dataset and the demonstrated effectiveness of SVMs in a similar biological classification task.

A previous demonstration comes from Bhasin and Raghava (2004), who developed ESLpred, an SVM-based method for predicting the subcellular localization of eukaryotic proteins. This work showed that SVMs achieve high performance when using high-dimensional sequence-derived features like dipeptide composition and PSI-BLAST profiles. This is highly related to our study, where we also use high-dimensional input vectors including amino acid derived features and ESM's CLS embeddings. Though we performed feature selection before training, both settings are still similar to ESLpred.

As described in *An Introduction to Statistical Learning* (James et al., 2021), our dataset includes around 500–1000 meiosis protein sequences and ~2000 non-meiosis sequences per locus. While this creates a modest class imbalance, SVMs are less sensitive to such imbalance compared to models that rely on minimizing total classification error. SVMs focus only on the most informative samples near the decision boundary which is the support vectors. With boundary points hardest to classify, maximizing the margin between class is the best way to model imbalance.

Even after feature selection using approximate mRMR, the retained features (50 vs 100 per set) can still have correlations, which is common for the relevant task. As explained in ISLR, SVMs work well with less assumption on feature independence and are optimized to find a decision boundary that minimizes overfitting. The margin-maximization principle stated in the book ensures that noise has limited impact on the model, an advantage in protein classification where subtle signal must be extracted from complex feature spaces.

The main reason we use RBF kernel is for amino acid features and CLS embeddings often encode non-linear relationships. A linear classifier of other kernels would fail to capture this complexity. According to ISLR, the RBF kernel allows the SVM to create flexible decision boundaries in a transformed space. This eases to separate meiosis and non-meiosis proteins even when they're not linearly separable in the original feature space.

The ultimate goal is to apply this classifier to protist genomes which are highly divergent from the animals, fungi, and plants used in training. The flexibility of the RBF kernel becomes particularly valuable. RBF SVMs can learn smooth, adaptable boundaries that are more likely to generalize across evolutionary distances than other models which means our model may detect meiosis proteins in protists, even if their sequences differ substantially from those in the training set.

Finally, SVMs with RBF kernel are much easier to tune. Unlike deep learning models that require extensive chance on training procedures, SVMs only require tuning two hyperparameters: the regularization parameter C and the kernel width γ. According to ISLR, these can be optimized efficiently using grid search and cross-validation.

As shown in Table 5, hyperparameter tuning was performed using a grid search over a broad range of values for the regularization parameter C of [0.01,0.1,1,10,100,1000] and the RBF kernel coefficient γ of [0.0001,0.001,0.01,0.1,1,10,scale,auto]. This choice of expansive search space enables control between margin maximization and misclassification tolerance, crucial for the classification task involving high-dimensional, noisy, and imbalanced protein features. Specifically, C allows the model to navigate between underfitting and overfitting of controlling margin and noise acceptance, while tuning γ targeting how far the influence of a single training point extends, effectively modifying the decision boundary. Given that meiosis proteins are functionally coherent but evolutionary diverse, especially when generalizing from other proteins to those from protists, this flexibility of tuning is crucial. The grid search ensures that the SVM identifies a robust boundary that accommodates feature noise and evolutionary divergence, making it well-suited for detecting meiosis-related proteins in novel and phylogenetically distant organisms.

Table 5. Tuned Hyperparameters for SVM (RBF Kernel) and Multi-Layer Perceptron Models

Models	Hyperparameters Tuned
Support Vector Machine (SVM) with Radial Basis Function (RBF) Kernel	Regularization Parameter (C): [0.01, 0.1, 1, 10, 100, 1000] Kernel Coefficient (gamma): 0.0001, 0.001, 0.01, 0.1, 1, 10, 'scale', 'auto'
Multi-layer perceptron	Hidden Layer Dimension (hidden_dim): [64, 128] Dropout Rate (dropout): [0.2, 0.3] Learning Rate (lr): [0.001, 0.0005] Batch Size (batch_size): 32 Number of Hidden Layers (num_layers): [1,2] Activation Function (activation): ReLU Weight Decay (L2 Regularization) (weight_decay): [0, 0.0001] Optimization Algorithm (optimizer): AdamW

Multi-layer Perceptron

In this study, we employ a multi-layer perceptron (MLP), a feedforward neural network architecture, to classify meiosis versus non-meiosis proteins. The MLP enables a practical balance between expressive power and simplicity, making it well-suited to this task. While more sophisticated convolutional networks are capable of learning complex sequence motifs, they often require extensive labeled training data and tend to overfitting when the sample size is limited (Zou et al., 2019). For our case, the number of labeled meiosis proteins is relatively small which on the order of 500 to 1000 sequences and the feature space, though refined to the top 50–100 dimensions, still carries noise. Also, the dataset is imbalanced, with meiosis proteins underrepresented relative to those of non-meiosis.

A further challenge arises from the later application of predicting meiosis proteins in distant protist genomes. These genomes may contain functionally conserved proteins like SPO11, yet exhibit substantial sequence divergence from the animal, fungal, and plant proteins used during training (Schurko & Logsdon, 2008; Tekle et al., 2017). Given this, a simple MLP might offer a computationally efficient solution. It is capable of modeling non-linear interactions in protein feature space such as ESM embeddings or amino acid features, while being less susceptible to overfitting than complex architectures (Goodfellow et al., 2016; Shrikumar et al., 2017).

To further improve generalization to unseen taxa, we employ regularization techniques such as dropout and weight decay. **Early stopping** is also introduced, using a patience of 5 epochs with the validation F1 score as the monitored metric (Prechelt, 2002). This mechanism halts training if no improvement in F1 is observed over five consecutive epochs to help the model avoid overfitting to training data and encourages it to capture generalizable biological signals. This is especially critical when applying the model to protists, whose proteins might be underrepresented or absent from training yet may share core functional roles in meiosis.

The hyperparameters used for tuning the MLP are by prior work in protein classification and deep learning (Min et al., 2017). Specifically:

- The **hidden layer dimension** (64 or 128) controls model capacity; 128 units may better capture nuanced signals in complex and noisy datasets.
- **Dropout rates** of 0.2 or 0.3 are applied during training to prevent reliance on specific neurons overly, addressing potential overfitting caused by feature redundancy or class imbalance (Srivastava et al., 2014).
- The **learning rate** (0.0005 or 0.001) affects how fast the model updates weights; the lower rate helps stabilize training with noisy protein features, while the higher rate promotes faster convergence.
- A **batch size** of 32 provides a balanced approach between computational efficiency and gradient stability, well-suited to the moderate dataset size.

- The number of **hidden layers** (1 or 2) allows modeling shallow hierarchical patterns, which may support generalization to taxonomically distant species.
- The **ReLU activation function** enables the model to learn non-linear relationships between features and function.
- **Weight decay** (0 or 0.0001) introduces L2 regularization, discouraging large weights. A small positive value promotes smoother generalization without overly constraining the model.
- Finally, the **AdamW optimizer** is used for training, which improves convergence stability by decoupling weight decay from the learning rate schedule (Loshchilov & Hutter, 2019).

Altogether, this MLP architecture is purposefully designed to balance simplicity, flexibility, and generalization. These choices reflect both the biological complexity of meiosis proteins and the computational challenges posed by sparse and heterogeneous training data across the eukaryotic.

Training Models

We employ two classification strategies to distinguish meiosis from non-meiosis proteins: a Support Vector Machine with Radial Basis Function (RBF) kernel and a multi-layer perception (MLP). The SVM serves as a simpler, well-established baseline, with previous work demonstrating its success through Bhasin and Raghava, 2004. The MLP introduces greater capacity to model non-linear relationships. Using both allows us to directly compare performance and to assess the added performance of deep learning in this specific classification setting.

Before training either model, we apply a sample weighting strategy to address two sources of bias: class imbalance of meiosis vs. non-meiosis (Table 2 vs. Table 1) and the significant imbalance of taxa (Table 1) for each of the 6 meiosis loci (Japkowicz and Stephen, 2002). To this, we assign higher weights to meiosis proteins from targeted eukaryotic lineages of arthropods, chordates, other animals, fungi and plants. Specifically, we upweight those of fungi arthropods and other animals which are much less representative. While globally, we also apply weight that upweight all meiosis protein to match with a larger amount of non-meiosis. This ensures rare but biologically important meiosis protein taxa exert greater influence during model training, thereby improving the ability to generalize to a very distant and diverse taxa of protist. These weights are inputted into both SVM and MLP classifiers to emphasize the functional signal present in conserved but underrepresented lineages like fungi, arthropods and other animals.

We trained both weighed versions of the SVM and MLP models to address class imbalance and improve classification performance across the 6 selected loci. Each model is trained using 70% of the standardized training set, with feature selection applied to both amino acid-derived features and ESM's CLS embeddings. For every locus, we tested two feature representations (amino acid-derived vs. ESM embeddings) and two feature set sizes (top 50 and top 100), applying both SVM and MLP classifiers to each configuration, which yields a total of 48 model training instances (6 loci × 2 feature types × 2 feature sizes × 2 models).

Hyperparameter tuning was conducted using 10-fold cross-validation, a robust model evaluation technique described in *An Introduction to Statistical Learning* (James et al., 2021), which reduces variance in performance estimation by ensuring every observation is used for both training and validation. To select the best hyperparameter configuration for each model, F1 score which is a harmonic mean of precision and recall over accuracy. Since F1 provides a more informative metric in class imbalance and uncertainty of labeling according to the nature of those datasets and settings. As noted in ISLR, cross-validation is very effective in moderate sample size which help access a model's ability to generalize towards the test set and the protist prediction.

Test Models across All Configurations

After identifying the optimal hyperparameter across all 48 configurations through cross-validation, we refit each model on the full 70% training set using those best parameters. Predictions were then made on the remaining 30% standardized test set across all six loci, using both feature types (amino acid-derived features and ESM CLS embeddings), two feature set sizes (top 50 and top 100), and two model architectures (SVM and MLP), totaling 48 test evaluations. These test sets span meiosis proteins from animals, fungi, and plants, along with matched non-meiosis controls, enabling us to assess generalization across eukaryotes groups. To evaluate model performance, we recorded four classification metrics: accuracy, precision, recall, and F1-score, as summarized in Table 6 with interpretation being provided in the Result section.

Following *An Introduction to Statistical Learning* (James et al., 2021), accuracy measures the overall proportion of correctly classified samples but can be misleading with imbalance presence. Precision measures the proportion of true positive predictions among all predicted positives, reflecting a model's ability to mitigate false positives. Recall measures the proportion of actual positives that are correctly identified which indicates how the model avoids false negatives. F1-score provides a single metric that balances the trade off between precision and recall and is especially informative when here meiosis seems to be imbalanced though we have applied weights. F1-score might serve as the most reliable performance metric and should be a guide of how the model performance is.

Prediction on Protist

For prediction, we applied the trained classifiers to protist genomes using standardized features for each locus–species pair (e.g., SPO11 & Entamoeba). Predictions were made separately for each of the six loci, using both amino acid-derived features and ESM CLS embeddings, with two feature sizes (top 50 and top 100 selected features), and two classification models (SVM and MLP), each with optimal hyperparameters. This resulted in 2 feature sizes × 2 feature types × 2 models × 6 loci = 48 configurations. Since each locus includes three protist species (*Entamoeba, Plasmodium, and Trypanosoma*), the total number of prediction settings was 6 loci × 3 species × 2 feature types × 2 feature sizes × 2 models = 144.

For each setting, we applied the trained model to all candidate sequences from the given protist and recorded the predicted class probabilities. A threshold of 0.90 on the predicted probability was used to mark on high-confidence meiosis candidates, and we recorded the count of predictions for each configuration (Table 7). In addition to binary classification, we also identified the top 3 protein sequences with the highest predicted probability (above the 90% threshold) as prioritized candidates (Table 9).

After filtering predicted probabilities above 0.90, we retained the corresponding protist sequences and recorded these in filtered .faa files. We then searched for potential meiosis proteins by checking for relevant keywords, based on Table 2 of Schurko and Logsdon Jr. (2008) or the locus descriptions in the *Data* section above, for each of the six loci (SPO11, DMC1, MND1, MSH4, MSH5, REC8). We will record the results in Table 9.

Results

Performance on Training

Table 6 presents the classification performance based on Accuracy, Precision, Recall, and F1 score of models distinguishing meiosis-related proteins across six loci of SPO11, DMC1, MND1, MSH4, MSH5, REC8 using two feature types (ESM's CLS embeddings and amino acid- derived features), two feature dimensions (Top 50, Top 100) and two classifiers (SVM with RBF kernel and MLP).

We observe overall high performance across all configurations. F1 score ranges from 0.947 to 0.997, indicating a very strong ability across loci. DMC1and MSH5 perform extremely well, which consistently are above 0.990 across both classifiers and feature types. While REC8 and MND1 generally have lower F1 values, REC8 with amino acid-derived features using SVM are the lowest amongst all metrics.

In terms of feature type comparison, amino acid-derived features perform only slightly worse, or nearly identically, to ESM CLS embeddings. For loci such as MSH5 and DMC1, both feature types yield comparable performance across all configurations. Regarding classifier performance, the SVM more often performs slightly better than the MLP, but this is not the case in general. For feature dimension effect, more often, top 100 features perform better than top 50 features. However, since all F-1 values are very high and in a very close range on metrics, we cannot easily determine which of the feature counts though there might be some exception.

For what we observed, we have shown for all configurations strong signals in known meiosis proteins. This means these meiosis proteins are highly conserved and functionally distinct, enabling very robust separation from non-meiosis. Both ESM embedding and amino acid-derived features likely capture biologically meaningful patterns specific to meiosis-related sequences. Additionally, these models were trained on a well-annotated set of meiosis and non-meiosis proteins across eukaryotes we collected which minimize noise while maximizing training ability. Non-meiosis though is majority class are functionally distinct which contributes to much clearer classification boundaries than expected. Lastly, both SVM with RBF kernel and MLP are suitable for such non-linear, imbalanced classifiers capable of capturing complex relationships in high-dimensional data when ESM's CLS embedding and amino acid-derived features are very informative.

These results are too high and might be very impractical. Overfitting is very possibly a cause which models are very possibly memorizing training examples rather than generalizing. This risk is higher if cross-validation isn't done properly based on ISLR. Also, since test proteins are extremely similar in distribution to training proteins which are all mixing up for taxa with significant overlaps which may result in a subtle data leakage, this result in unintentionally shared evolutionary patterns or convergent features across both training and test sets, leading to the model make easy predictions based on taxonomic signals rather than true functional properties.

A more serious and obvious consideration is that meiosis and non-meiosis we collected might differ too drastically, making the task way easier than Bhasin & Raghava (2004) which resolve on functionally similar protein classification (e.g., apoptosis-related proteins). All the non-meiosis proteins we include are housekeeping proteins like ribosomal, metabolic, cytoskeletal, heat shock, and translation-related proteins which are with core functions unrelated to meiosis at all. These proteins are functionally distinct from meiosis-related proteins, which may allow models to achieve artificially high-performance metrics by only learning broad functional differences.

Moreover, we did not include proteins functionally similar but non-meiosis, such as core mitosis, DNA repair, or recombination not of meiosis. Another 1000 of those way more similar proteins might be able to reveal the actual classification power for SVM with RBF kernel vs. MLP, ESM's CLS embedding vs. amino acid derived features and feature count differences. To properly perform this classification task, we should be able to distinguish subtle differences in proteins like Bhasin & Raghava (2004).

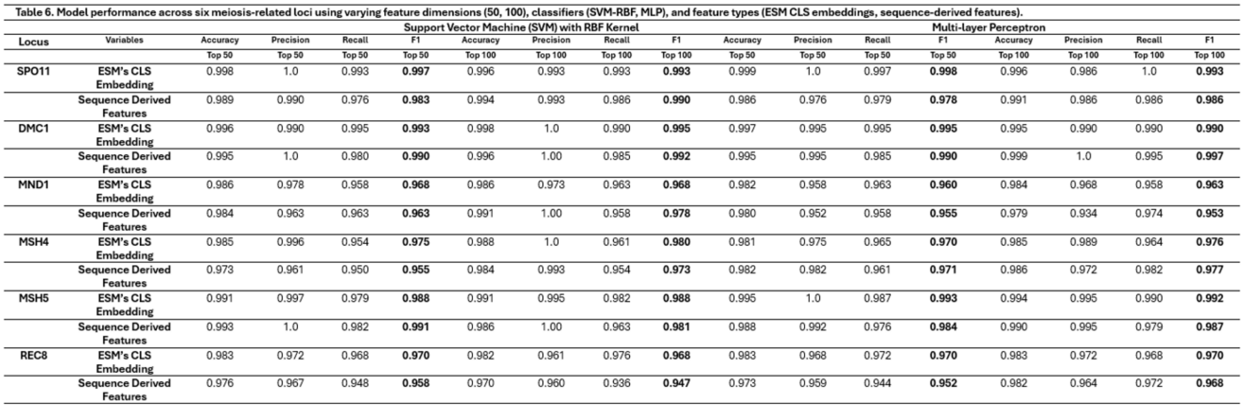

Table 6. Model performance across six meiosis-related loci using varying feature dimensions (50, 100), classifiers (SVM-RBF, MLP), and feature types (ESM CLS embeddings, sequence-derived features).

		Support Vector Machine (SVM) with RBF Kernel								Multi-layer Perceptron							
Locus	Variables	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
		Top 50	Top 50	Top 50	Top 50	Top 100	Top 100	Top 100	Top 100	Top 50	Top 50	Top 50	Top 50	Top 100	Top 100	Top 100	Top 100
SPO11	ESM's CLS Embedding	0.998	1.0	0.993	**0.997**	0.996	0.993	0.993	**0.993**	0.999	1.0	0.997	**0.998**	0.996	0.986	1.0	**0.993**
	Sequence Derived Features	0.989	0.990	0.976	**0.983**	0.994	0.993	0.986	**0.990**	0.986	0.976	0.979	**0.978**	0.991	0.986	0.986	**0.986**
DMC1	ESM's CLS Embedding	0.996	0.990	0.995	**0.993**	0.998	1.0	0.990	**0.995**	0.997	0.995	0.995	**0.995**	0.995	0.990	0.990	**0.990**
	Sequence Derived Features	0.995	1.0	0.980	**0.990**	0.996	1.00	0.985	**0.992**	0.995	0.995	0.985	**0.990**	0.999	1.0	0.995	**0.997**
MND1	ESM's CLS Embedding	0.986	0.978	0.958	**0.968**	0.986	0.973	0.963	**0.968**	0.982	0.958	0.963	**0.960**	0.984	0.968	0.958	**0.963**
	Sequence Derived Features	0.984	0.963	0.963	**0.963**	0.991	1.00	0.958	**0.978**	0.980	0.952	0.958	**0.955**	0.979	0.934	0.974	**0.953**
MSH4	ESM's CLS Embedding	0.985	0.996	0.954	**0.975**	0.988	1.0	0.961	**0.980**	0.981	0.975	0.965	**0.970**	0.985	0.989	0.964	**0.976**
	Sequence Derived Features	0.973	0.961	0.950	**0.955**	0.984	0.993	0.954	**0.973**	0.982	0.982	0.961	**0.971**	0.986	0.972	0.982	**0.977**
MSH5	ESM's CLS Embedding	0.991	0.997	0.979	**0.988**	0.991	0.995	0.982	**0.988**	0.995	1.0	0.987	**0.993**	0.994	0.995	0.990	**0.992**
	Sequence Derived Features	0.993	1.0	0.982	**0.991**	0.986	1.00	0.963	**0.981**	0.988	0.992	0.976	**0.984**	0.990	0.995	0.979	**0.987**
REC8	ESM's CLS Embedding	0.983	0.972	0.968	**0.970**	0.982	0.961	0.976	**0.968**	0.983	0.968	0.972	**0.970**	0.983	0.972	0.968	**0.970**
	Sequence Derived Features	0.976	0.967	0.948	**0.958**	0.970	0.960	0.936	**0.947**	0.973	0.959	0.944	**0.952**	0.982	0.964	0.972	**0.968**

Performance on Validation

The first critical issue is that generating CLS embeddings using Evolutionary Scale Modeling (ESM) proved infeasible due to time and memory constraints in the Google Colab environment. Despite implementing a parallelized solution via esm_parallel.ipynb, processing a single protist proteome took over 60 hours. Given these limitations, I decided to exclude ESM embeddings from the predictions summarized in Tables 7–9 and instead focused on amino acid-derived features. I acknowledge this constraint and appreciate your understanding.

Table 7 shows the number of candidate meiosis proteins predicted with high probability (greater than 95% or greater than 99%) across six loci of SPO11, DMC1, MND1, MSH4, MSH5, REC8) and three protists (*Entamoeba, Plasmodium, Trypanosoma*)., using two classifiers (SVM and RBF kernel and MLP) and two feature sets (Top 50 or Top 100 amino-acid-derived features + Shannon entropy).

From the observation, MLP consistently predicts a larger number of high-probability candidates across all loci and protists across all loci and protists compared to SVM, at least 3 times more. Also, MLP is observed to have higher sensitivity, flagging more potential meiosis, while SVM appears to be very conservative may with less false positive. While from the table, SVM's output fluctuates much more than MLP with a increase of threshold from 95% to 99%. For low-signal loci like DMC1 in Plasmodium, SVM failed to return any candidate above 0.95 which we denoted in footnote, whereas MLP still found some, possibly risking overfitting. This demonstrates that SVM is much better if aiming for interpretability which may result in a more optimal prediction.

We observe that MLP at 95% threshold, MLP returned more than 1000 proteins in multiple loci (e.g., MSH5: 1416 in Entamoeba, 1157 in Trypanosoma) which is highly implausible even if you count. Even with Entamoeba's canonical meiosis genes are either missing or highly diverged, the 1157 in Trypanosoma is unrealistic with at least 10% being meiosis genes. Based on Min et al., 2017, MLP might be easily overfit with a training dataset of around 2000-3000 with 50 or 100 features, this leads to the model being memorizing patterns correlate with training set biases instead of generalizable features of meiosis proteins. More importantly, for MLP, our parameter tuning in Multi-layer Perceptron section focused on avoiding overfitting given its known sensitivity to high-dimensional data and small training sets. We applied some amount of regularization during tuning, but the result shown here with significant amount of positive still indicates overfitting which more regularization should be applied.

In this case, our final model chosen should be SVM with RBF kernel. SVM is with much higher precision which returns much fewer positives with 95% threshold or 99%. SVM reduces false positives by a significant amount. Contrast to MLP, SVM avoids overfitting with much less of hyperparameters need to tune with only C and gamma. Prediction scales very realistically with less than 100 positives for 99% threshold. Smooth drop off across threshold is also more realistic compared to much more stable calibration in MLP. SVM with RBF kernel is certainly more realistic while MLP is a better candidate for exploratory analysis here.

Table 7. Number of Predicted Meiosis Protein Candidates in Protist Genomes by SVM and MLP Classifiers Across Feature Sets and Thresholds.

Locus	Features	Threshold (%)	SVM with RBF Kernel			Multi-layer Perceptron		
			Entamoeba	Plasmodium	Trypanosoma	Entamoeba	Plasmodium	Trypanosoma
SPO11	Top 50 + Shannon Entropy	95	227	82	69	1276	391	527
	Top 50 + Shannon Entropy	99	42	15	15	909	241	373
	Top 100 + Shannon Entropy	95	138	32	40	813	377	654
	Top 100 + Shannon Entropy	99	17	5	8	478	247	499
DMC1	Top 50 + Shannon Entropy	95	3	0*	2	52	11	70
	Top 50 + Shannon Entropy	99	1	0	2	25	7	35
	Top 100 + Shannon Entropy	95	10	5	7	79	16	72
	Top 100 + Shannon Entropy	99	4	1	2	39	11	32
MND1	Top 50 + Shannon Entropy	95	71	29	44	548	456	302
	Top 50 + Shannon Entropy	99	8	3	9	435	385	233
	Top 100 + Shannon Entropy	95	41	14	26	351	359	189
	Top 100 + Shannon Entropy	99	6	2	4	210	228	107
MSH4	Top 50 + Shannon Entropy	95	202	69	64	837	234	492
	Top 50 + Shannon Entropy	99	51	12	8	548	191	414
	Top 100 + Shannon Entropy	95	230	12	194	900	160	461
	Top 100 + Shannon Entropy	99	77	2	29	706	97	307
MSH5	Top 50 + Shannon Entropy	95	176	30	109	1416	536	1157
	Top 50 + Shannon Entropy	99	33	2	17	1311	483	994
	Top 100 + Shannon Entropy	95	87	3	144	1028	261	1311
	Top 100 + Shannon Entropy	99	12	2	23	902	219	1161
REC8	Top 50 + Shannon Entropy	95	16	17	152	454	370	681
	Top 50 + Shannon Entropy	99	1	6	49	370	306	571
	Top 100 + Shannon Entropy	95	10	9	49	636	268	646
	Top 100 + Shannon Entropy	99	2	4	26	558	229	535

* DMC1's highest prediction probability being 0.941 for top 50 features + Shannon entropy

Table 8 presents predicted meiosis-related proteins in protist genomes with >90% probability, using selected feature sets (Top 50 + Shannon entropy vs. Top 100 + Shannon entropy). Several predicted hits are topoisomerases associated with SPO11, which align with the explanation in the *Data* section suggesting these may represent true SPO11 orthologs or related proteins. Notably, while screening protist .faa files, we observed that at least three topoisomerases' sequences were missing from 90% threshold for all three protists, and some showed very low scores when cross-checked against the full .faa. Among the top hits, *tr|C4M624|C4M624_ENTH1* appears with high confidence as a candidate SPO11, ranking 48th in the Top 50 + entropy set and 13th in the Top 100 + entropy set. In *Plasmodium*, three topoisomerases were identified with high confidence in both feature sets, while in *Trypanosoma*, a DNA topoisomerase is ranked 27th only in Top 100 + entropy.

For DMC1, predictions are highly confident across all three protist genomes, with *Meiotic recombination protein DMC1* and *DNA repair protein RAD51 homolog* consistently ranked among the top three. This demonstrates a strong and accurate prediction for DMC1 using the current model.

For MND1, we observe that the *Meiosis-specific nuclear structural protein* ranks relatively low in *Entamoeba* (rank 63 in Top 50 + entropy). However, the related *Meiotic coiled-coil protein* ranks much higher in the same dataset. In the Top 100 + entropy set, only the coiled-coil protein surpasses the 90% threshold. In *Plasmodium*, a *Meiotic nuclear division protein 1, putative* appears only in the Top 100 + entropy set, highlighting it's clearly a borderline cases of very close to threshold. In *Trypanosoma*, the *Meiosis-specific nuclear structural protein* ranks 5th in Top 100 + entropy but is not detected in the Top 50 + entropy set which emphasizes the importance of the broader feature set is necessary.

Regarding MSH4, based on Table 1 of Schurko & Logsdon (2008), there are likely no true MSH4 orthologs present in *Plasmodium.* In *Entamoeba*, *DNA mismatch repair protein mutS* ranks 7th in the Top 50 + entropy set but drops significantly to 167th in the Top 100 + entropy set. A large number of predicted positives (230 proteins above the 90% threshold) suggest the possibility of false positives or functional analogs and warrant further analysis. In *Trypanosoma*, only one predicted protein (ranked 68th in Top 50 + entropy) is recovered, but the same ID is ranked 15th in Top 100 + entropy, closely aligning with an MSH5 hit at rank 14.

For MSH5, the same ID as in MSH4 appears in *Entamoeba*, ranking 250th. Further annotation checking is required to determine whether this hit truly corresponds to MSH4 or MSH5. In *Trypanosoma*, MSH5 is clearly predicted, ranking within the top 5 in both feature sets.

Finally, for REC8, only *DNA repair protein Rad21*, a known paralog of REC8, was recovered across the datasets. In *Trypanosoma*, Rad21 is confidently detected, ranking 13th and 16th in the Top 50 and 100 + entropy sets, respectively. In *Plasmodium*, *Rad21/Rec8-like* is detected at rank 12 in Top 50 + entropy but is not detected in the Top 100 + entropy set.

In summary, DMC1 is reliably detected across all protists, while MND1 is confidently predicted in *Entamoeba* and *Trypanosoma*. Other loci, including SPO11, MSH4, MSH5, and REC8, show promising hits but require additional downstream validation using methods such as phylogenetic analysis, domain annotation, or structural modeling.

Table 8. Predicted Meiosis-Related Proteins in Protist Genomes with >90% Probability Using Selected Feature Sets (Top 50 + Shannon Entropy vs. Top 100 + Shannon Entropy)

<table>
<tr><th>Loci</th><th>Protists</th><th>Top 50 + Shannon Entropy (ID, related proteins, ranking, prediction probability)</th><th>Top 100 + Shannon Entropy (ID, related proteins, ranking, prediction probability)</th></tr>
<tr><td><b>SPO11</b></td><td><b>Entamoeba</b></td><td>tr|C4M624|C4M624_ENTH1, DNA topoisomerase (ATP-hydrolyzing), 48, 0.98934
tr|C4MBH4|C4MBH4_ENTH1, Topoisomerase, putative, 267, 0.943338</td><td>tr|C4M624|C4M624_ENTH1, DNA topoisomerase (ATP-hydrolyzing), 13, 0.993477</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|Q8I5N7|Q8I5N7_PLAF7, DNA topoisomerase (ATP-hydrolyzing), 21, 0.988011
tr|A0A143ZZR0|A0A143ZZR0_PLAF7, DNA topoisomerase (ATP-hydrolyzing), 22, 0.987053
tr|C6S3D8|C6S3D8_PLAF7, DNA topoisomerase (ATP-hydrolyzing), 41, 0.977784</td><td>tr|A0A143ZZR0|A0A143ZZR0_PLAF7, DNA topoisomerase (ATP-hydrolyzing), 44, 0.94122
tr|C6S3D8|C6S3D8_PLAF7, DNA topoisomerase (ATP-hydrolyzing), 20, 0.964667
tr|Q8I5N7|Q8I5N7_PLAF7, DNA topoisomerase (ATP-hydrolyzing), 32, 0.950047</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>Not hitting</td><td>tr|Q57U90|Q57U90_TRYB2 DNA topoisomerase (ATP-hydrolyzing), 27, 0.967623</td></tr>
<tr><td><b>DMC1</b></td><td><b>Entamoeba</b></td><td>tr|C4LTR6|C4LTR6_ENTH1, Meiotic recombination protein DMC1, 1, 0.999996
tr|C4M4K4|C4M4K4_ENTH1, DNA repairprotein RAD51 homolog, 2, 0.985599</td><td>tr|C4LTR6|C4LTR6_ENTH1, Meiotic recombination protein DMC1, 1, 0.997214
tr|C4M4K4|C4M4K4_ENTH1, DNA repair protein RAD51 homolog, 1, 0.992874</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|Q8IB05|Q8IB05_PLAF7, Meiotic recombination protein DMC1, putative, 1, 0.940955</td><td>tr|Q8IB05|Q8IB05_PLAF7, Meiotic recombination protein DMC1, putative, 3, 0.972753</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q38E34|Q38E34_TRYB2, RAD51/dmc1 protein, 1, 1
DNA repair protein RAD51 homolog, DNA repair protein RAD51 homolog, 2, 0.997418</td><td>tr|Q38E34|Q38E34_TRYB2, RAD51/dmc1 protein, 3, 0.982291</td></tr>
<tr><td><b>MND1</b></td><td><b>Entamoeba</b></td><td>tr|C4LZN7|C4LZN7_ENTH1, Meiosis-specific nuclear structural protein, 63, 0.958734
tr|C4M8I3|C4M8I3_ENTH1, Meiotic coiled-coil protein, 3, 0.997074
tr|C4LYK6|C4LYK6_ENTH1, Chromosome segregation in meiosis protein 3 domain-containing protein, 107, 0.924571</td><td>tr|C4M8I3|C4M8I3_ENTH1, Meiotic coiled-coil protein, 3, 0.993527</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>Not hitting</td><td>tr|C6S3J7|C6S3J7_PLAF7, Meiotic nuclear division protein 1, putative, 41, 0.901728</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>Not hitting</td><td>tr|Q587B8|Q587B8_TRYB2, Meiosis-specific nuclear structural protein 1, 5, 0.989237</td></tr>
<tr><td><b>MSH4</b></td><td><b>Entamoeba</b></td><td>tr|C4M7H8|C4M7H8_ENTH1, DNA mismatch repair protein mutS, 7, 0.999987
tr|C4M0Z9|C4M0Z9_ENTH1, DNA mismatch repair proteins mutS family domain-containing protein, 89, 0.981594</td><td>tr|C4M7H8|C4M7H8_ENTH1, DNA mismatch repair protein mutS, putative,167, 0.9708
tr|C4M0Z9|C4M0Z9_ENTH1, DNA mismatch repair proteins mutS family domain-containing protein, 180, 0.967428</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|C0H4L8|C0H4L8_PLAF7, DNA mismatch repair protein MSH2, putative, 5, 0.99576
tr|Q8ILI9|Q8ILI9_PLAF7, DNA mismatch repair protein MSH2, putative, 18, 0.985615</td><td>tr|Q8ILI9|Q8ILI9_PLAF7, DNA mismatch repair protein MSH2, putative, 9, 0.961088
tr|C0H4L8|C0H4L8_PLAF7, DNA mismatch repair protein MSH2, putative, 11, 0.952933</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q38CB3|Q38CB3_TRYB2, Mismatch repair protein, putative, 68, 0.947553</td><td>tr|Q580L4|Q580L4_TRYB2, Mismatch repair protein MSH5, putative, 14, 0.994934
tr|Q38CB3|Q38CB3_TRYB2, Mismatch repair protein, putative, 15, 0.994598
tr|Q38AW5|Q38AW5_TRYB2, Mismatch repair protein MSH8, putative, 185, 0.903082
tr|Q38F29|Q38F29_TRYB2, Mismatch repair protein MSH3, putative, 95, 0.959754</td></tr>
<tr><td><b>MSH5</b></td><td><b>Entamoeba</b></td><td>tr|C4M7H8|C4M7H8_ENTH1, DNA mismatch repair protein mutS, putative, 250, 0.925776</td><td>Not hitting</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|C0H4L8|C0H4L8_PLAF7, DNA mismatch repair protein MSH2, putative, 58, 0.922412</td><td>Not hitting</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q580L4|Q580L4_TRYB2, Mismatch repair protein MSH5, putative, 5</td><td>tr|Q580L4|Q580L4_TRYB2, Mismatch repair protein MSH5, putative, 1, 0.999997
tr|Q38F29|Q38F29_TRYB2, Mismatch repair protein MSH3, putative, 189, 0.931659</td></tr>
<tr><td><b>REC8</b></td><td><b>Entamoeba</b></td><td>tr|C4M385|C4M385_ENTH1, UV excision repair protein RAD23, putative, 1, 1
tr|C4LYH9|C4LYH9_ENTH1, DNA repair protein Rad21, putative, 12, 0.964607</td><td>tr|C4M385|C4M385_ENTH1, UV excision repair protein RAD23, putative, 1, 0.999998
tr|C4M4K4|C4M4K4_ENTH1, DNA repair protein RAD51 homolog, 15, 0.937256
tr|C4LYH9|C4LYH9_ENTH1, DNA repair protein Rad21, putative, 25, 0.918044</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|Q8IJS8|Q8IJS8_PLAF7, UV excision repair protein RAD23, 15, 0.960119
tr|Q8IL69|Q8IL69_PLAF7, Rad21/Rec8-like protein N-terminal domain-containing protein, 12, 0.970143</td><td>tr|Q8IJS8|Q8IJS8_PLAF7 UV, excision repair protein RAD23, 8, 0.962463</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q57XR0|Q57XR0_TRYB2, Double-strand-break repair protein rad21 homolog, putative, 13, 0.999984</td><td>tr|Q57XR0|Q57XR0_TRYB2, Double-strand-break repair protein rad21 homolog, putative, 16, 0.993865
tr|Q38DL8|Q38DL8_TRYB2, DNA repair protein RAD2, putative, 114, 0.921253</td></tr>
</table>

Table 9 shows that for most protist genomes, a lot of the top three predicted meiosis-associated proteins aren't actually annotated as meiosis-related, or they're not annotated at all. The only consistent hit across all species is DMC1, which seems to confirm that the SVM model, especially using an RBF kernel and amino acid-derived features, is capable of pulling out at least some highly conserved meiosis genes, even in very divergent genomes like those of protists.

That said, there's a clear pattern in which lots of non-meiosis proteins are scoring high. There are a few likely reasons. One is that certain non-meiotic proteins such as helicases, topoisomerases, and general DNA-binding proteins, have overlapping features with some actual meiosis genes. They're involved in recombination, genome stability, and chromatin dynamics, so it's not surprising they might mimic meiosis proteins to the model. If you're feeding its features from proteins like SPO11, MND1, or REC8, which all play roles in those same processes, it makes sense that the model can get a bit confused. That overlap is real and very hard to avoid (Marcotte et al., 1999).

Another issue is the quality of the annotations. Many protist proteins are just labeled as "hypothetical" or "uncharacterized," which doesn't tell us much. Some of them might actually be meiosis-related, but without functional studies or better annotation databases, we're barely guessing. This was already pointed out by Schurko and Logsdon (2008), and others like Malik et al. (2008) have echoed it too just because something isn't annotated doesn't mean it's irrelevant. We might be missing real meiosis genes simply because no one's done the work to verify what they are.

On top of that, there's the issue of overfitting. SVMs with RBF kernels are powerful but also risky in high-dimensional space, especially when the features are things like amino acid compositions or other sequence-derived features. It's easy for the model to latch onto superficial patterns that don't actually reflect function. So yeah, it can predict proteins that *look* like meiosis proteins in terms of their biochemical profiles, but that doesn't mean they actually are. Some false positives are inevitable in ML-based setup (Cai et al., 2020).

The negative training set also plays a role. If the "non-meiosis" proteins used for training are mostly housekeeping types, ribosomal, metabolic, cytoskeletal, and so on, then the model might learn to distinguish meiosis proteins from those. But that's not the whole picture. What about other DNA-binding proteins that aren't meiosis-specific? If they weren't included in the negative set, the model has no reason to learn to separate them from meiosis proteins, and they'll likely show up among the high-scoring false positives.

Finally, there's the biological context. Meiosis doesn't happen in isolation—it's connected to the cell cycle, DNA repair, chromosome segregation, and more. Many proteins involved in those adjacent processes might show up in the predictions, not because they *are* meiosis proteins, but because they're nearby in function or regulation. So, in a way, the model isn't entirely wrong—but it's also not as precise as we'd like. That ambiguity makes it tricky to interpret predictions without follow-up analysis.

Table 9. Top Three High-Confidence Predicted Meiosis-Associated Proteins per Protist for Each Locus, Based on Different Feature Sets (Top 50 + Shannon Entropy vs. Top 100 + Shannon Entropy)

<table>
<tr><th>Loci</th><th>Protists</th><th>Top 50 + Shannon Entropy (ID, protein names)</th><th>Top 100 + Shannon Entropy (ID, protein names)</th></tr>
<tr><td><b>SPO11</b></td><td><b>Entamoeba</b></td><td>tr|C4M7J7|C4M7J7_ENTH1 Poly(A) RNA polymerase mitochondrial-like central palm domain-containing protein
tr|C4LV10|C4LV10_ENTH1 Poly(A) polymerase, putative
tr|C4LZP7|C4LZP7_ENTH1 Rab-GAP TBC domain-containing protein</td><td>tr|C4LZP7|C4LZP7_ENTH1 Rab-GAP TBC domain-containing protein
tr|C4LY77|C4LY77_ENTH1 Uncharacterized protein
tr|C4LXJ9|C4LXJ9_ENTH1 Zinc finger domain containing protein</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|Q8ILY9|Q8ILY9_PLAF7 protein-synthesizing
sp|Q8I615|ORC1_PLAF7 Origin recognition complex subunit 1
tr|Q8IL65|Q8IL65_PLAF7 Allantoicase, putative</td><td>tr|Q8I5S7|Q8I5S7_PLAF7 Glycerol-3-phosphate 1-O-acyltransferase
tr|Q8I3G9|Q8I3G9_PLAF7 RING zinc finger protein, putative
tr|C6KSW1|C6KSW1_PLAF7 Basal complex transmembrane protein 1</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q383Y0|Q383Y0_TRYB2 CobW/HypB/UreG nucleotide-binding domain-containing protein
tr|Q582E7|Q582E7_TRYB2 Transmembrane protein
tr|Q582E1|Q582E1_TRYB2 Transmembrane protein</td><td>tr|Q380Z0|Q380Z0_TRYB2 Serine/threonine-protein phosphatase
tr|Q387C8|Q387C8_TRYB2 Uncharacterized protein
tr|Q38B82|Q38B82_TRYB2 General transcription factor IIH subunit 4</td></tr>
<tr><td><b>DMC1</b></td><td><b>Entamoeba</b></td><td>tr|C4LTR6|C4LTR6_ENTH1 Meiotic recombination protein DMC1
tr|C4M4K4|C4M4K4_ENTH1 DNA repair protein RAD51 homolog
tr|C4M8L1|C4M8L1_ENTH1 Uncharacterized protein</td><td>tr|C4LTR6|C4LTR6_ENTH1 Meiotic recombination protein DMC1
tr|B1N2M7|B1N2M7_ENTH1 rRNA processing protein, putative
tr|B1N4G0|B1N4G0_ENTH1 Serine/threonine kinase, putative</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|Q8IB05|Q8IB05_PLAF7 Meiotic recombination protein DMC1, only one above 90% cutoff</td><td>tr|Q8IEN2|Q8IEN2_PLAF7 40S ribosomal protein S27
tr|Q8I2R8|Q8I2R8_PLAF7 Polyadenylate-binding protein 2, putative
tr|Q8IB05|Q8IB05_PLAF7 Meiotic recombination protein DMC1, putative</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q38E34|Q38E34_TRYB2 RAD51/dmc1 protein, putative
tr|Q384K0|Q384K0_TRYB2 DNA repair protein RAD51 homolog
tr|Q384B8|Q384B8_TRYB2 Uncharacterized protein</td><td>tr|Q383S4|Q383S4_TRYB2 Uncharacterized protein
tr|Q384N0|Q384N0_TRYB2 Uncharacterized protein
tr|Q38E34|Q38E34_TRYB2 RAD51/dmc1 protein, putative</td></tr>
<tr><td><b>MND1</b></td><td><b>Entamoeba</b></td><td>tr|C4LUZ8|C4LUZ8_ENTH1 Uncharacterized protein
tr|C4LZS7|C4LZS7_ENTH1 High mobility group (HMG) box domain containing protein
tr|C4M8I3|C4M8I3_ENTH1 Meiotic coiled-coil protein</td><td>tr|C4LWC7|C4LWC7_ENTH1 Dephospho-CoA kinase, putative
tr|C4M6W3|C4M6W3_ENTH1 High mobility group (HMG) box domain containing protein
tr|C4M8I3|C4M8I3_ENTH1 Meiotic coiled-coil protein</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|Q8IBA6|Q8IBA6_PLAF7 Uncharacterized protein
tr|C0H4D5|C0H4D5_PLAF7 Apicomplexan specific coiled coil protein
tr|C0H560|C0H560_PLAF7 Pre-mRNA-splicing factor SYF2</td><td>tr|Q8IBA6|Q8IBA6_PLAF7 Uncharacterized protein
tr|Q8I3L9|Q8I3L9_PLAF7 50S ribosomal protein L12, apicoplast, putative
tr|C0H4D5|C0H4D5_PLAF7 Apicomplexan specific coiled coil protein</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q585W4|Q585W4_TRYB2 Uncharacterized protein
tr|Q38BT0|Q38BT0_TRYB2 Uncharacterized protein
tr|Q388K8|Q388K8_TRYB2 Uncharacterized protein</td><td>tr|Q585W4|Q585W4_TRYB2 Uncharacterized protein
tr|Q584F1|Q584F1_TRYB2 Uncharacterized protein
tr|Q384U1|Q384U1_TRYB2 Dynein regulatory complex subunit 2</td></tr>
<tr><td><b>MSH4</b></td><td><b>Entamoeba</b></td><td>tr|B1N482|B1N482_ENTH1 Uncharacterized protein
tr|C4LWT5|C4LWT5_ENTH1 HEAT repeat-containing protein 1
tr|C4LW42|C4LW42_ENTH1 Uncharacterized protein</td><td>tr|C4LVB5|C4LVB5_ENTH1 Importin beta-3 subunit family protein, putative
tr|C4M0B4|C4M0B4_ENTH1 HEAT repeat domain containing protein
tr|C4M5I7|C4M5I7_ENTH1 Importin, putative</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|Q8IHR6|Q8IHR6_PLAF7 Coatomer subunit gamma
tr|C6KT96|C6KT96_PLAF7 Separase
tr|A0A5K1K8Q1|A0A5K1K8Q1_PLAF7 Exportin-T</td><td>tr|A0A5K1K8Q1|A0A5K1K8Q1_PLAF7 Exportin-T
tr|C0H530|C0H530_PLAF7 Exportin-7, putative
tr|O97283|O97283_PLAF7 Oocyst capsule protein Cap380</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q38A51|Q38A51_TRYB2 AP complex subunit beta
tr|Q57Y15|Q57Y15_TRYB2 Uncharacterized protein
tr|Q4GYW2|Q4GYW2_TRYB2 Pumillio RNA binding protein, putative</td><td>tr|Q583W5|Q583W5_TRYB2 Uncharacterized protein
tr|Q585X9|Q585X9_TRYB2 PAP-associated domain-containing protein
tr|Q38A51|Q38A51_TRYB2 AP complex subunit beta</td></tr>
<tr><td><b>MSH5</b></td><td><b>Entamoeba</b></td><td>tr|C4LWT5|C4LWT5_ENTH1 HEAT repeat-containing protein 1
tr|C4M2H2|C4M2H2_ENTH1 Importin alpha
tr|C4M748|C4M748_ENTH1 CNH domain-containing protein</td><td>tr|C4M3U1|C4M3U1_ENTH1 VPS9 domain-containing protein
tr|C4LUA7|C4LUA7_ENTH1 HEAT repeat domain containing protein
tr|C4LWT5|C4LWT5_ENTH1 HEAT repeat-containing protein 1</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|C0H530|C0H530_PLAF7 Exportin-7, putative
tr|Q8I5S7|Q8I5S7_PLAF7 Glycerol-3-phosphate 1-O-acyltransferase
tr|A0A5K1K8Q1|A0A5K1K8Q1_PLAF7 Exportin-T</td><td>tr|Q8I5S7|Q8I5S7_PLAF7 Glycerol-3-phosphate 1-O-acyltransferase
tr|C0H530|C0H530_PLAF7 Exportin-7, putative
tr|O96153|O96153_PLAF7 26S proteasome regulatory subunit RPN1, putative</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q38AQ5|Q38AQ5_TRYB2 ATP-dependent RNA helicase
tr|Q38D32|Q38D32_TRYB2 Nucleoporin
tr|Q586T7|Q586T7_TRYB2 Uncharacterized protein</td><td>tr|Q580L4|Q580L4_TRYB2 Mismatch repair protein MSH5, putative
tr|Q382T4|Q382T4_TRYB2 Mon2 C-terminal domain-containing protein
tr|Q586T7|Q586T7_TRYB2 Uncharacterized protein</td></tr>
<tr><td><b>REC8</b></td><td><b>Entamoeba</b></td><td>tr|C4M385|C4M385_ENTH1 UV excision repair protein RAD23, putative
tr|C4LVW0|C4LVW0_ENTH1 Uncharacterized protein
tr|C4M3G5|C4M3G5_ENTH1 Serine/threonine-protein phosphatase</td><td>tr|C4LZP7|C4LZP7_ENTH1 Rab-GAP TBC domain-containing protein
tr|C4LY77|C4LY77_ENTH1 Uncharacterized protein
tr|C4LXJ9|C4LXJ9_ENTH1 Zinc finger domain containing protein</td></tr>
<tr><td></td><td><b>Plasmodium</b></td><td>tr|Q8IIM8|Q8IIM8_PLAF7 Ubiquitin domain-containing protein DSK2, putative
tr|Q8IHY0|Q8IHY0_PLAF7 protein-serine/threonine phosphatase
tr|Q76NM2|Q76NM2_PLAF7 Thrombospondin-related anonymous protein</td><td>tr|Q8IK70|Q8IK70_PLAF7 Plasmodium RESA N-terminal domain-containing protein
tr|Q8IDX9|Q8IDX9_PLAF7 MSP7-like protein
tr|Q76NM2|Q76NM2_PLAF7 Thrombospondin-related anonymous protein</td></tr>
<tr><td></td><td><b>Trypanosoma</b></td><td>tr|Q389F4|Q389F4_TRYB2 Uncharacterized protein
tr|Q582K0|Q582K0_TRYB2 Uncharacterized protein
tr|Q38EY2|Q38EY2_TRYB2 Nucleosome assembly protein-like protein</td><td>tr|Q38FX0|Q38FX0_TRYB2 T. brucei spp.-specific protein
tr|Q585N8|Q585N8_TRYB2 Uncharacterized protein
tr|Q587I3|Q587I3_TRYB2 Uncharacterized protein</td></tr>
</table>

Discussion

Summarize Results

In this study, we first surveyed why meiosis is essential to sexual reproduction and genomic stability, and why studying meiosis genes in protists, once thought to be asexual, is crucial for uncovering hidden sexual processes and understanding eukaryotic evolution. We then reviewed prior work on meiosis proteins in protists, focusing on Ramesh et al. (2005), who first used comparative genomics and phylogenetics to identify conserved meiotic genes in lineages like Giardia and Trypanosoma, challenging the notion of protist asexuality. Schurko and Logsdon (2008) refined this approach by introducing the "meiosis detection toolkit," a curated set of conserved, meiosis-specific genes coupled with degenerate PCR to detect these markers even in incomplete genomes. Together, these studies provided a standardized framework for uncovering cryptic sex in protists and highlighted the evolutionary conservation of meiotic machinery across eukaryotes. Later, we discussed how machine learning offers key advantages for studying meiosis genes in protists. Unlike motif-based or PCR methods, ML can detect divergent homologs using features like ESM embeddings and amino acid properties. It scales to whole proteomes, integrates structural data, and provides confidence scores, making it ideal for uncovering meiosis genes in understudied, divergent protists.

We then outlined the data and methodology used in our analysis. The dataset included six known meiosis-specific loci, and a control set composed primarily of housekeeping proteins. Our prediction targets were the full proteomes of three protist species. To classify meiosis vs. non-meiosis proteins, we implemented a supervised learning pipeline using a weighted Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel and a Multi-layer Perceptron (MLP). We justified the model choice and corresponding hyperparameters, and applied standardization and approximate mRMR feature selection prior to training. Input features included ESM-2 protein embeddings and amino acid-derived descriptors; two feature sets, containing the top 50 and top 100 ranked features were set up for 48 training configurations. Models were tuned using 10-fold cross-validation and evaluated on a 70:30 train-test split, with F1 score used as the primary performance metric for the 30% test set. Finally, we applied the trained models to the protist proteomes, recording prediction probabilities for each sequence and examining known meiosis loci to assess performance and identify limitations of the pipeline. Preliminary conclusions were drawn about the prediction of selected protists.

Performance Discussion

The classifiers (SVM and MLP) for meiosis vs. meiosis on gathered data achieved very high F1 scores (0.947–0.997) across all loci, features, and feature counts, indicating that meiosis proteins have distinct, learnable patterns. Both ESM and amino acid-derived features encode biologically relevant signals, and the models effectively distinguish between classes. However, there are serious concerns about overfitting due to the small dataset size, potential taxonomic overlap between training and test sets, and an overly simplistic classification task, since non-meiosis housekeeping proteins collected are functionally distant. As a result, the impressive metrics may overstate the true generalizability of the models.

Due to severe time and memory constraints in Google Colab, generating ESM embeddings for protist proteomes proved infeasible, so the final predictions in Tables 7–9 rely solely on amino acid-derived features.

MLP predicted significantly more high-probability meiosis candidates (often >3 times of SVM), showing higher sensitivity and signs of overfitting, especially at the 95% threshold where some predictions (e.g., >1000 proteins) are biologically implausible. SVM with RBF kernel was more conservative, with fewer predictions and more fluctuation across thresholds, but likely more realistic and interpretable, with lower false positives and smoother threshold calibration. The conclusion favors SVM with RBF kernel as the preferred model due to better precision, interpretability, and resistance to overfitting, making it a more believable choice for this task.

High-probability meiosis protein predictions (>90%) in three protist genomes using amino acid-derived features (Top 50 vs. Top 100 + Shannon entropy) delivered enthusiastic results:

- **SPO11**: Several predicted proteins are topoisomerases, aligning with expectations. Some key topoisomerases were missing or had low scores. One strong candidate, *tr|C4M624|*, consistently ranks highly, especially in the Top 100 set.
- **DMC1**: Predictions are highly accurate across all protists, with DMC1 and RAD51 homologs consistently top-ranked, showing strong model performance for this locus.
- **MND1**: Predictions vary a lot here. In Entamoeba, the true MND1 ortholog ranks low, while a related coiled-coil protein ranks higher. Broader feature sets of Top 100 help recover borderline cases, especially in Plasmodium and Trypanosoma.
- **MSH4**: Likely absent in Plasmodium. In Entamoeba and Trypanosoma, some predicted hits may be false positives or analogs; large numbers of high-scoring proteins suggest overprediction.
- **MSH5**: A shared candidate with MSH4 appears in Entamoeba but needs further annotation. In Trypanosoma, MSH5 is confidently predicted in both feature sets.
- **REC8**: Only paralog Rad21 is recovered. Detected in all three protists, but inconsistently between feature sets, supporting it as a borderline or paralogous case.

As shown in Table 9, DMC1 stands out as the only meiosis gene consistently predicted across the surveyed protist genomes. This points to the SVM classifier’s ability to recognize deeply conserved meiosis-related proteins, even when working with highly divergent sequences. That said, the broader list of top hits includes a significant number of proteins that are either unannotated or completely unrelated to meiosis, which isn’t surprising. A few factors likely contribute such as overlaps in features between meiosis proteins and general DNA-binding or repair proteins, incomplete or low-quality genome annotations in many protist species, and the relatively narrow taxonomic range of the negative training set.

These issues, especially when combined with the risk of overfitting in a high-dimensional feature space, introduce noise into the predictions and can lead to inflated confidence in false positives. In other words, while the model performs well on paper, some of its top predictions in real protist data aren't biologically meaningful without further validation.

Together, the results support the idea that machine learning, and SVMs in particular, can be valuable tools for detecting conserved meiotic machinery in lineages where traditional homology methods might fail. But they also highlight the limits of relying solely on ML for functional inference, especially in organisms where the underlying annotations are unreliable. The tendency for meiosis-related features to overlap with those of other chromatin-associated proteins can obscure true signals, and performance metrics like F1 score may overstate the model's effectiveness outside controlled benchmarks.

Ultimately, while ML provides a helpful first pass for prioritizing candidate proteins, its output needs to be interpreted in a very cautious way. Follow-up validation which through approaches like domain-specific annotation, phylogenetic placement, and structural prediction, remains essential for drawing reliable biological conclusions otherwise the generated results can hardly be conclusive.

Contrast to older work

Compared to the approach taken by Schurko and Logsdon (2008), which relied on a defined "meiosis detection toolkit" using BLAST searches and PCR analysis, the machine learning (ML) method used here represents a move toward more flexible, data-driven inference. Both methods aim to identify evidence of meiosis in eukaryotic genomes, particularly in understudied lineages, but they go about it in fundamentally different ways.

The original toolkit focused on detecting a core set of highly conserved meiosis genes, such as SPO11, DMC1, and REC8, by looking for close sequence homologs. This strategy is powerful when dealing with well-annotated proteins that haven't diverged much over time. But its reliance on sequence similarity thresholds means it can miss functionally conserved genes that have drifted too far in sequence to be recognized by standard BLAST cutoffs which is too conservative often. In such cases, even genuinely meiosis-related genes could go undetected. From toolkit-based methods, it is more direct to get the answers like Table 1 that definite about whether a meiosis loci is presented with an exact tabulation.

By contrast, the ML method doesn't depend on one-to-one homology in design. Instead, it learn representations and from known meiosis proteins, such as their amino acid composition, structural features, or embedding-based representations, and tries to generalize those patterns to unknowns. This makes it particularly useful in protists, where gene annotation is often incomplete, and protein sequences can evolve rapidly. However, the flexibility of ML comes at a cost. Models can overfit, and sometimes they mistake unrelated proteins—like DNA repair enzymes or general chromatin remodelers—for meiosis genes, simply because they share certain biochemical or structural features. While here under our investigation, we have a probabilistic number on whether these annotated/unannotated genes have a higher chance of being meiosis which is still necessary to use BLAST or toolkit based methods to scan for its homolog or paralog or ortholog later on.

Each method, then, has strengths and trade-offs. The toolkit approach is highly specific and interpretable but risks missing divergent cases. The ML approach is more sensitive to remote homologs but also more prone to false positives. In practice, the two methods may work best together. A promising direction for future research is to use ML models to cast a wide net and then refine those predictions using traditional homology tools, creating a complementary pipeline that balances sensitivity and specificity.

Future Work

Future work will focus on clustering high-confidence candidate proteins, specifically those scoring above the 90% prediction threshold, to investigate whether they group into families with shared structural or functional characteristics suggestive of meiosis-related roles. To accomplish this, we plan to use the Markov clustering algorithm (MCL; Enright et al., 2002), which has been widely used for detecting functionally coherent groups in protein similarity networks. Particular attention will be given to possible paralogs near canonical meiosis genes such as REC8, MSH4, MSH5, and SPO11, as well as to any high-scoring proteins lacking annotation. These will be examined using standard sequence similarity tools like BLASTP and domain prediction with InterProScan incorporating with toolkit-based methods, following approaches described by Schurko and Logsdon (2008). For selecting proteins that meet prediction confidence criteria but remain uncharacterized, we will also generate AlphaFold structural models to evaluate whether their predicted folds may possibly support a meiosis-related function. To improve scalability and avoid the memory limitations encountered in Google Colab, we will re-run the ESM-based feature extraction on a local high-performance computing cluster (Rives et al., 2021). Additionally, the training set will be broadened to include proteins involved in mitosis and other core cell cycle processes. These biologically relevant negatives should improve model generalizability and help reduce overly optimistic performance metrics. Finally, given the unexpectedly high F1 scores observed in test metrics, we will incorporate more rigorous evaluation strategies, including leave-one-group-out cross-validation and orthogonal methods such as phylogenetic profiling, as suggested by Ramesh et al. (2005), to better assess the robustness of the model and guard against overfitting.

References:

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002). *Biochemistry* (5th ed.). W. H. Freeman. [Chapters on Glycolysis and TCA cycle]

Bepler, T., & Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell systems*, *12*(6), 654-669.

Bishop, D. K., Park, D., Xu, L., & Kleckner, N. (1992). DMC1: a meiosis-specific yeast homolog of E. coli recA required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell*, *69*(3), 439-456.

Bhasin, M., & Raghava, G. P. S. (2004). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Bioinformatics*, 20(9), 1335–1341.

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., ... & Keith Dunker, A. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of molecular evolution*, *55*(1).

Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., ... & Pei, J. (2020). Transfer learning for drug discovery. *Journal of Medicinal Chemistry*, *63*(16), 8683-8694.

Cavalier-Smith T. (2002). Origins of the machinery of recombination and sex. *Heredity*, *88*(2), 125–141.

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo‐amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, *43*(3), 246-255.

Chou, K.C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21(1), 10–19.

Dubchak, I., Muchnik, I., Holbrook, S. R., & Kim, S. H. (1995). Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences*, 92(19), 8700–8704.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, *30*(7), 1575-1584.

Gertz, E. M., Yu, Y.-K., Agarwala, R., Schäffer, A. A., & Altschul, S. F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology, 4*(1), 41.

Goodfellow, I. (2016). Deep learning.

Handel, M. A., & Schimenti, J. C. (2010). Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nature Reviews Genetics*, *11*(2), 124-136.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754-755.

Jackson, R. J., Hellen, C. U., & Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature reviews Molecular cell biology*, *11*(2), 113-127.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, No. 1). New York: springer.

Japkowicz, N., & Stephen, S. (2002). *The class imbalance problem: A systematic study*. Intelligent data analysis, 6(5), 429–449.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236-1240.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.

Kawashima, S., & Kanehisa, M. (2000). AAindex: amino acid index database. *Nucleic acids research*, *28*(1), 374-374.

Keeney, S., Giroux, C. N., & Kleckner, N. (1997). Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*, *88*(3), 375–384

Koonin, E. V. (2010). The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biology*, 11, 209.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123-1130.

Lindquist, S., & Craig, E. A. (1988). The heat-shock proteins. *Annual review of genetics*, *22*(1), 631-677.

Lodish, H., Berk, A., Kaiser, C. A., et al. (2016). *Molecular Cell Biology* (8th ed.). W. H. Freeman. [Chapter: The Cytoskeleton]

Malik, S. B., Ramesh, M. A., Hulstrand, A. M., & Logsdon Jr, J. M. (2007). Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. *Molecular biology and evolution*, *24*(12), 2827-2841.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, *285*(5428), 751-753.

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in bioinformatics*, *18*(5), 851-869.

McPherson, M., & Møller, S. (2000). *PCR*. Taylor & Francis.

Page, S. L., & Hawley, R. S. (2003). Chromosome choreography: the meiotic ballet. *Science (New York, N.Y.)*, *301*(5634), 785–789.

Parisi, S., McKay, M. J., Molnar, M., Thompson, M. A., van der Spek, P. J., van Drunen-Schoenmaker, E., ... & Kohli, J. (1999). Rec8p, a meiotic recombination and sister chromatid cohesion phosphoprotein of the Rad21p family conserved from fission yeast to humans. *Molecular and cellular biology*, *19*(5), 3515-3528.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, *27*(8), 1226-1238.

Pezza, R. J., Voloshin, O. N., Vanevski, F., & Camerini-Otero, R. D. (2007). Hop2/Mnd1 acts on two critical steps in Dmc1-promoted homologous pairing. *Genes & development*, *21*(14), 1758-1766.

Prechelt, Lutz. "Early stopping-but when?." *Neural Networks: Tricks of the trade*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. 55-69.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., ... & Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature methods*, *10*(3), 221-227.

Ramesh, M. A., Malik, S. B., & Logsdon, J. M. (2005). A phylogenomic inventory of meiotic genes: evidence for sex in Giardia and an early eukaryotic origin of meiosis. *Current biology*, *15*(2), 185-191.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118.

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., & Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. *Peptide Research*, 11(3), 119–123.

Sander, C., & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, 9(1), 56–68.

Schurko, A. M., & Logsdon Jr, J. M. (2008). Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. *BioEssays*, *30*(6), 579-589.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Reverse-complement parameter sharing improves deep learning models for genomics. *BioRxiv*, 103663.

Snowden, T., Acharya, S., Butz, C., Berardini, M., & Fishel, R. (2004). hMSH4-hMSH5 recognizes Holliday Junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Molecular cell*, *15*(3), 437-451.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Weismann, A. (1891). *Essays upon heredity and kindred biological problems* (Vol. 1). Clarendon press.

Wilson, D. N., & Cate, J. H. D. (2012). The structure and function of the eukaryotic ribosome. *Cold Spring Harbor perspectives in biology*, *4*(5), a011536.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., ... & Hsieh, C. J. (2019). Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.

Zou, X., Xu, S., Li, S., Chen, J., & Zou, W. (2019). Optimization of the Brillouin instantaneous frequency measurement using convolutional neural networks. *Optics letters*, *44*(23), 5723–5726.