MyESL Directives

No	Directives	Descriptions
Model Options		
1.	`Position 1 <text file>`	A text file containing full paths for all gene sequence alignments in FASTA format
2.	`Position 2`	A phylogenetic tree in newick format with clade ID
3.	`--clade_list <text_file>`	A text file containing the list of all clades for which clade models are built. Each line in the file contains a clade ID.
4.	`--gen_clade_list <lower, upper>`	This directive is used when the input phylogeny does not contain predefined clades. It automatically generates clade IDs for all internal nodes whose descendant species count falls within the specified **lower** and **upper** bounds. By default, both values are set to 0, which results in assigning clade IDs to **all internal nodes** in the phylogeny.
5.	`--auto_name_length <int>`	Specifies the number of characters to extract from species names when generating internal node labels for clade IDs. The default is 3, meaning the first three letters of species names will be used to construct clade identifiers.
6.	`--clade_size_cutoff_lower <num>`	This directive is used when no predefined clades are present in the input phylogeny. It enables users to automatically generate candidate clades for model building by specifying a lower and upper bound on clade size (both numeric). Clades containing a number of species within the specified size range will be selected, and ESL models will be built for each of them.
7.	`--clade_size_cutoff_upper`	
8.	`--class_bal <method>`	This directive addresses **class imbalance** in clade modeling—i.e., when the number of species inside the focal clade differs significantly from the number outside. Without

No	Directives	Descriptions
		correction, such an imbalance can bias the model toward the larger class, reducing predictive accuracy. MyESL provides four strategies for class balancing: ● **up**: Upsampling the smaller class to match the size of the larger class. ● **down**: Downsampling the larger class to match the smaller class ● **weighted**: Applying inverse class weights during model training ● **phylo**: Phylogeny-aware balancing that selects training samples based on evolutionary relationships Use this option to ensure balanced and unbiased model training, especially when clade sizes are uneven.
9.	`--classes <text file>`	A tab-separated text file listing the **response values for all species** included in the clade model. Each line should contain a **species name** followed by its corresponding **binary response label**: ● +1 for species **inside** the focal clade ● -1 for species **outside** the clade This file is required when building clade-specific models. This file serves as the basis for building clade-specific models in the absence of a Newick-format phylogeny.
10.	`--lambda1 <float [0-1]>`	Specifies the site sparsity penalty parameter that controls the inclusion of features (e.g., sites) in the clade model. The value must be a float between 0 and 1, where higher values apply stronger penalization, resulting in a sparser model that selects fewer sites.

No	Directives	Descriptions
11.	`--lambda2 <float [0-1]>`	Specifies the group sparsity penalty parameter that controls the inclusion of feature groups (e.g., genes) in the clade model. The value must be a float between 0 and 1, where higher values apply stronger penalization, resulting in a sparser model that selects fewer groups.
12.	`--no_group_penalty`	Enables monolevel Lasso analysis by disabling the group-level sparsity penalty. This directive must be used in combination with `--lambda1`, and ensures that sparsity is applied only at the individual feature level, not at the group (e.g., gene) level.
13.	`--lambda1_grid <float, float, float [0-1]>`	Specifies a range of feature-level sparsity parameters (`lambda1`) for building multiple clade models, one model for each parameter. This directive takes three float values: the starting value, the maximum value, and the step size. All values should be between 0 and 1. The parameter values increase linearly from the start to the maximum using the specified step, allowing exploration of different levels of feature sparsity.
14.	`--lambda1_grid <float, float, float [0-1]>`	Specifies a range of feature-level sparsity parameters (`lambda1`) for building multiple clade models, one model for each parameter. This directive takes three float values: the starting value, the maximum value, and the step size. All values should be between 0 and 1. The parameter values increase linearly from the start to the maximum using the specified step, allowing exploration of different levels of feature sparsity.

No	Directives	Descriptions
15.	`--grid_rmse_cutoff<float>`	This directive is to be used when lambda grids are used with multiple models. This helps to reduce the number of models to build and which models to use for aggregation.
16.	`--grid_acc_cutoff <float [0 1)>`	Sets the minimum training accuracy threshold for retaining models during grid search. Any model with accuracy below this cutoff will be discarded.
17.	`--grid_summary_only`	This directive retains only the summary of all models generated during the grid search and discards the results of individual models.
18.	`--min_group <num>`	Specifies the minimum number of groups that must be retained in each model during grid search. Models with fewer groups will be discarded.
19.	`--method <method>`	Specifies the type of loss function or regression to be used in model training. The default option is `logistic.` • `logistic:` which performs sparse group lasso or lasso regression with logistic loss and is suitable for binary response variables (e.g., +1/–1). • `leastr:` applies sparse group lasso or lasso regression using least squares loss, and should be used when the response variable is continuous, ranging from $-\infty$ to $+\infty$.
20.	`--slep_opts <text file>`	A text file containing different options used for model optimization.
Data Preprocessing		
21.	`--include_singletons`	By default, MyESL removes all conserved sites and singleton variants (sites with only one differing sequence). Use this option to retain singleton sites during clade model construction if they are relevant to your

No	Directives	Descriptions
		analysis.
22.	`--bit_ct <num>`	Excludes sites where mutations are observed fewer than `<num>` times across all sequences. This helps filter out extremely rare variants that may be uninformative or noisy.
23.	`--data_type <type>`	Specifies the type of input sequence data MyESL will process. Supported options are: ● `nucleotide`: For DNA sequences ● `protein`: For amino acid (AA) sequences ● `molecular`: For datasets containing both DNA and protein sequences ● `universal`: Allows any IUPAC characters, additional symbols, and numeric values (0–9), treating each unique character, including numbers, as distinct. This mode is case-sensitive, unlike the others, which are not. Choose the appropriate type based on the format and diversity of your input data.
Model Output		
24.	`--output <name>`	An output directory name can be defined by users. This directory will be created in the current working directory. If no output directory name is defined by users, a directory named “output” will be created by default.
25.	`--gene_display_limit <num>`	Specifies the maximum number of gene groups (columns) to display in the model grid generated by MyESL. The default is 20, or fewer if the number of selected groups is less than 20.
26.	`--species_display_limit <num>`	Specifies the maximum number of taxa (columns) from the clade of interest to display in the model grid generated by MyESL. The default is 20, or fewer if the number of selected taxa is less than 20.
27.	`--gene_display_cutoff`	Defines the number of genes or groups

No	Directives	Descriptions
	`<num>`	(columns) to display in the model grid. The default is 20, or fewer if fewer than 20 groups are selected.
28.	`--m_grid <row,col>`	Sets the number of rows and columns in the model grid displayed by MyESL. The default is 20 for both dimensions, or fewer if the available rows or columns are less than 20.
29.	`--stats_out <PGHS>`	Specifies the types of sparsity statistics to output from the model. Users can include any combination of the following: P: Site-level (position) sparsity scores G: Gene or group sparsity scores H: Hypothesis sparsity score S: Species prediction scores
Miscellaneous		
30.	`--DrPhylo`	Enables DrPhylo analysis for a specified clade, either defined by clade ID using a phylogeny in a NEWICK format or via a response file in text format. DrPhylo builds multiple sparse models across combinations of site and group sparsity parameters. By default, it explores values from 0.1 to 0.9 (in steps of 0.1), generating 81 models. Users can customize this range using the `--lambda_grid` directive. To reduce computation, DrPhylo skips models that include fewer than three genes. It outputs summary statistics, including PSS, GSS, and HSS scores, as well as a model grid. The default grid size is 20×20 (20 rows for species and 20 columns for genes), which can be adjusted using the `--m_grid <row, col>` directive.
31.	`--subsets <int>`	Enables subset-based analysis by dividing feature groups (e.g., genes) into multiple subsets. Models are trained on each subset, and informative groups are selected from each. A final model is then built using the combined set of selected groups. If

No	Directives	Descriptions
		`--subsets 0`, MyESL automatically determines the optimal number of subsets based on the system's computational capacity.