`slapnap`: Super LeArner Prediction of NAb Panels

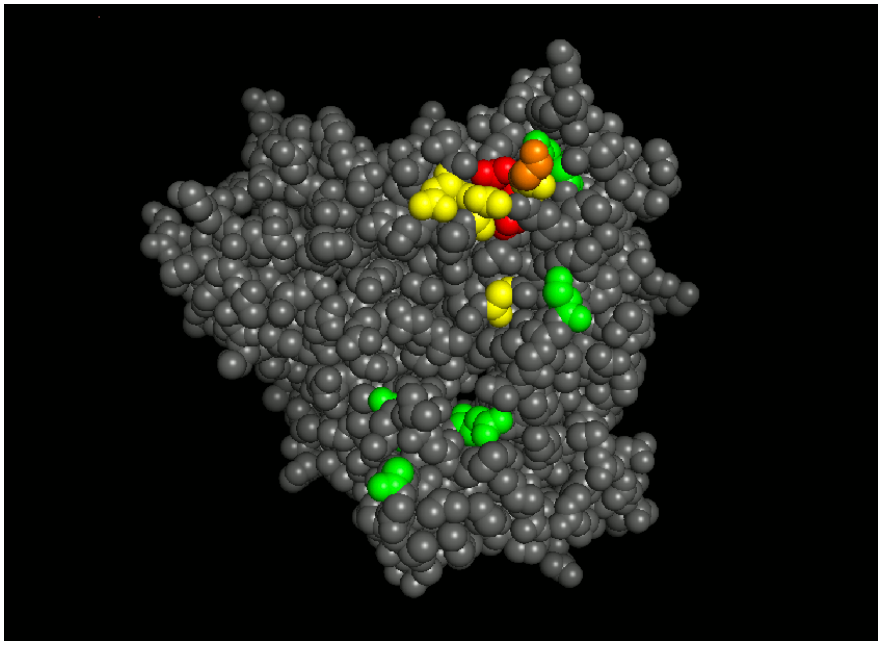
David Benkeser, Brian D. Williamson, Craig A. Magaret, Sohail Nizam, Peter B. Gilbert

June 08, 2020

Contents

Welcome

The `slapnap` container is a tool for using the Compile, Analyze and Tally NAb Panels (CATNAP) database to develop predictive models of HIV-1 neutralization sensitivity to one or several broadly neutralizing antibodies (bNAbs).



Crystal structure of HIV-1 gp120 glycoprotein. Highlighted residues indicating sites most-predictive of VRC01 neutralization resistance. [@magaret2019prediction]

In its simplest form, `slapnap` can be used simply to access and format data from CATNAP in a way that is usable for machine learning analysis. However, the tool also offers fully automated and customizable machine learning analyses based on up to five different neutralization endpoints, complete with automated report generation to summarize results and identify the most predictive features.

This document serves as the user manual for the `slapnap` container. Here, we describe everything needed to utilize `slapnap` and understand its output. The documentation is organized into the following sections:

- Section 1 provides a brief overview of Docker, including information on installing Docker and downloading the `slapnap` container.
- Section 2 provides a brief overview of the CATNAP database and the specifics of how and when these data were accessed to build the `slapnap` container.
- Section 3 provides a detailed description of how to make calls to `slapnap` and all options that are available at run time to customize its behavior.
- Section 4 includes example calls to `slapnap` for accomplishing different tasks.
- Section 5 describes the methodology used by `slapnap` to generate and analyze data.
- Section 6 describes the contents of the automated report generated by `slapnap`.
- Section 7 provides a description of the analysis data set created by `slapnap`.

If you have any issues or questions about using `slapnap`, please file an issue on GitHub.

1 Docker

Docker is a free platform for building containers. Containers are standard units of software that package code and all its dependencies, so that the code can be executed reliably irrespective of computing environment. `slapnap` relies on machine learning implemented in the `R` language and relies on several packages. Achieving full reproducibility for such analyses is challenging in that it requires synchronization across the specific version of `R` and dependent packages. In other words, two users running two versions of `R` or two versions of the same `R` package may arrive at different output when running the same code. Containerization ensures that this does not happen. Any two runs of `slapnap` with the same input options will yield the same output every time.

Installing Docker is necessary for running `slapnap`. While it is not necessary for execution of the `slapnap` container, readers interested in learning more about Docker should consult the Docker documentation for information about getting started using Docker.

Once Docker has been installed on your local computer, you can download `slapnap` using the following command.

```
docker pull slapnap/slapnap
```

This command pulls the image from DockerHub. Once the image has been downloaded, we are ready to learn about how to execute `slapnap` jobs. The next section contains information on the source data used by `slapnap`. Users familiar with the CATNAP data may wish to skip directly to Section 3.1.

2 CATNAP

The CATNAP database is a web server hosted by Los Alamos National Laboratory [Yoon et al., 2015]. The database integrates antibody neutralization and HIV-1 sequence data from published studies. Neutralization is measured in terms of half maximal inhibitory concentration (IC_{50}) and 80% inhibitory concentration (IC_{80}). These measures of neutralization against HIV envelope pseudoviruses are available for many broadly neutralizing antibodies (bNAbs) and for some combination bNAbs. Also available on each pseudovirus are amino acid (AA) sequence features for the gp160 protein. These are detailed in Section 7.

During each build of the `slapnap` container, all raw data are downloaded from CATNAP. At run time, pseudovirus features are derived and measured sensitivity outcomes are derived from the raw CATNAP database files and merged into a `.csv` file that is used in subsequent predictive analyses.

The CATNAP data are updated periodically. The data are downloaded into the `slapnap` container at every build. The most recent build occurred on June 08, 2020.

3 Running `slapnap`

To run the `slapnap` container, we make use of the `docker run` command. Note that administrator (`sudo`) privileges are needed to execute this command.

There are several options that are necessary to include in this command to control the behavior of `slapnap`. These are discussed in separate subsections below.

3.1 `slapnap` run options

The user has control over many aspects of `slapnap`'s behavior. These options are passed in using the `-e` option[1]. Semi-colon separated strings are used to set options. For example, to provide input for the option `option_name`, we would used `-e option_name="a;semi-colon;separated;string"`. Note that there are no spaces between the option name and its value and no spaces after semi-colons in the separated list. See Section 4 for full syntax.

Each description below lists the default value that is assumed if the option is not specified. Note that many of the default values are chosen simply so that naive calls to `slapnap` compile quickly. Proper values should be determined based on scientific context.

-e options for `slapnap`

[1]This sets an environment variable in the container environment. These variables are accessed by the various `R` and `bash` scripts in the container to dictate how the container executes code.

- **nab**: A semicolon-separated list of bNAbs (default = "VRC01"). A list of possible bNAbs can be found here. If multiple bNAbs are listed, it is assumed that the analysis should be of estimated outcomes for a combination of bNAbs (see Section 5.1 for details on how estimated outcomes for multiple bNAbs are computed).
- **outcomes**: A semicolon-separated string of outcomes to include in the analysis. Possible values are "ic50" (included in default), "ic80", "iip", "sens" (included in default), "estsens", "multsens". If only a single nab is specified, use sens to include a dichotomous endpoint. If multiple nabs are specified, use estsens and/or multsens. For detailed definitions of outcomes see Section 5.1.
- **sens_thresh** A numeric value defining the IC_{50} threshold for defining a sensitive versus resistant pseudovirus (default = 1). The dichotomous sensitivity/resistant outcomes are defined as the indicator that (estimated) IC_{50} is greater than or equal to sens_thresh.
- **multsens_nab** A numeric value used for defining whether a pseudovirus is resistant to a multi-nAb cocktail. Only used if multsens is included in outcome and more than one nab is requested. The dichotomous outcome multsens is defined as the indicator that a virus has IC_{50} greater than sens_thresh for at least multsens_nab nAbs.
- **learners**: A semicolon-separated string of machine learning algorithms to include in the analysis. Possible values include "rf" (random forest, default), "xgboost" (eXtreme gradient boosting), and "lasso" (elastic net). See Section 5.2 for details on how tuning parameters are chosen. If more than one algorithm is included, then it is assumed that a cross-validated-based ensemble (i.e., a super learner) is desired (see Section 5.3).
- **cvtune**: A boolean string (i.e., either "TRUE" or "FALSE" [default]) indicating whether the learners should be tuned using cross validation and a small grid search. Defaults to "FALSE". If multiple learners are specified, then the super learner ensemble includes three versions of each of the requested learners with different tuning parameters.
- **cvperf**: A boolean string (i.e., either "TRUE" or "FALSE" [default]) indicating whether the learners performance should be evaluated using cross validation. If cvtune="TRUE" or learners includes multiple algorithms, then nested cross validation is used to evaluate the performance of the cross validation-selected best value of tuning parameters for the specified algorithm or the super learner, respectively.
- **nfolds**: A numeric string indicating the number of folds to use in cross validation procedures (default = "2").
- **importance_grp**: A semicolon-separated string indicating which group-level variable importance measures should be computed. Possible values are none "" (default), marginal "marg", conditional "cond". See Section 5.4.1 for details on these measures.
- **importance_ind**: A semicolon-separated string indicating which individual-level variable importance measures should be computed. Possible values are none "" (default), learner-level "pred", marginal "marg" and conditional "cond". The latter two take significant computation time to compute.
- **same_subset** If "FALSE" (default) all data available for each outcome will be used in the analysis. If "TRUE", when multiple outcomes are requested, the data will be subset to just those sequences that have all measured outcome, and, if iip is requested, for which iip can be computed (i.e., measured IC_{50} and IC_{80} values are different). Thus, if "TRUE" all requested outcomes will be evaluated using the same_subset of the CATNAP data.
- **report_name**: A string indicating the desired name of the output report (default = report_[_-separated list of nabs]_[date].html).
- **return**: A semicolon-separated string of the desired output. Possible values are "report" (default), "learner" for a .rds object that contains the algorithm for each endpoint trained using the full analysis data, "data" for the analysis dataset, "figures" for all figures from the report, and "vimp" for variable importance objects.
- **view_port**: A boolean string indicating whether the compiled report should be made viewable on localhost (default "FALSE"). If "TRUE" then -p option should be used in the docker run command to identify the port. See example in Section 4.2 for details.

3.2 Returning output

At the end of a `slapnap` run, user-specified output will be saved (see option `return` in Section 3.1). To retrieve these files from the container, there are two options: mounting a local directory (Section 3.2.1) or, if the report is the **only** desired output, viewing and saving the report in a web browser (Section 3.2.2).

3.2.1 Mounting a local directory

To mount a local directory to the output directory in the container (`/home/output/`), use the `-v` option. Any items saved to the output directory in the container (file path in the container `/home/output/`) will be available in the mounted directory. Conversely, all files in the mounted local directory will be visible to programs running inside the container.

Suppose `/path/to/local/dir` is the file path on a local computer in which we wish to save the output files from a `slapnap` run. A `docker run` of `slapnap` would include the option `-v /path/to/local/dir:/home/output`. After a run completes, the requested output should be viewable in `/path/to/local/dir`. See Section 4 for full syntax.

To avoid possible naming conflicts and file overwrites in the mounted directory, **we recommend mounting an empty directory** to store the output.

3.2.2 Viewing report in browser

An alternative option to mounting local directories for viewing and downloading the report is to set the `view_port` option to `"TRUE"` and open a port to the container via the `-p` option in the `docker run` statement. In this case, rather than exiting upon completion of the analysis, the container will continuing to run and broadcast the compiled report to `localhost` at the specified port (see examples below). The report can be downloaded from the web browser directly in this way.

4 Examples

4.1 Basic call to `slapnap`

A call to `slapnap` with all default options can be run using the following command.

```
docker run -v /path/to/local/dir:/home/output slapnap/slapnap
```

Note that this call mounts the local directory `path/to/local/dir` to receive output from the container (see Section 3.2.1).

When this command is executed, messages will print to indicate the progress of the container. The first message will report the name of the log file, which will appear in `/path/to/local/dir`. The container will then compile an analytic data set from the CATNAP database for the default bNAb (VRC01), train the default learner (random forest [Breiman, 2001]) for the default outcomes (`ic50` and `sens`), evaluate its performance using two-fold (default for `nfolds`) cross validation, and compile a report detailing the results, and place the compiled report in `path/to/local/dir`.

4.2 Viewing report in browser

To have the results viewable in a web browser execute the following command[2].

[2]In this command, we use the escape character `\` to break the command over multiple lines, which will work on Linux and Mac OS. In Windows Command Prompt, the equivalent escape character is `^`. In both cases, take care not to include a space

```
docker run -v /path/to/local/dir:/home/output \
           -e view_port="TRUE" -p 80:80 \
           slapnap/slapnap
```

This command opens port 80 on the container. Once the report has been compiled, the container will not close down automatically. Instead it will continue to run, broadcasting the report to port 80. Open a web browser on your computer and navigate to `localhost:80` and you should see the compiled report. Many web browsers should allow downloading of the report (e.g., by right-clicking and selecting save `Save As...`).

The container will continue to run until you tell it to `stop`. To do that, retrieve the container ID by executing `docker container ps`[3]. Copy the ID of the running container, which will be a string of numbers and letters (say `a1b2c3d4`) and then execute `docker stop a1b2c3d4` to shut down the container.

Note that in the above command, we have still mounted a local directory, which may be considered best practice in case other output besides the report is desired to be returned.

4.3 Super learning

If multiple `learner`s are specified, then a super learner ensemble [van der Laan et al., 2007] is constructed based on the requested `learner`s and a predictor that simply returns the empirical average of the outcome (i.e., ignores all features entirely). In the following command, we construct an ensemble based on a random forest [Breiman, 2001] and elastic net [Zou and Hastie, 2005]. Note that the execution time for super learners can be considerably greater than for single `learner`s because of the extra layer of cross validation needed to construct the ensemble.

```
docker run -v /path/to/local/dir:/home/output \
           -e learners="rf;lasso" \
           slapnap/slapnap
```

For specific details on the super learner algorithms implemented in `slapnap`, see Section 5.3.

4.4 Train an algorithm

The previous commands train learners and evaluate their performance using cross validation. However, at times we may wish only to use `slapnap` to train a particular algorithm, while avoiding the additional computational time associated with evaluating its cross-validated performance and compiling a report. We show an example of this below using sensitivity as the outcome.

```
docker run -v /path/to/local/dir:/home/output \
           -e learners="rf" \
           -e return="learner" \
           -e cvperf="FALSE" \
           -e outcomes="sens" \
           slapnap/slapnap
```

After completion of this run, `learner_sens.rds` will appear in `/path/to/local/dir` that contains an `R` object of class `ranger` (the `R` package used by `slapnap` to fit random forests). You can confirm this from the command line by executing

```
Rscript -e "learner <- readRDS('/path/to/local/dir/learner_sens.rds'); class(learner)"
```

after the escape character.

[3] To execute this command, you will need to hit `control + c` to return to the command prompt in the current shell or open a new shell. Alternatively, you could add the `-d` option to the `docker run` command, which will run the container in detached mode.

4.5 Pull and clean data

The `slapnap` container can also be used to return cleaned CATNAP data suitable for analyses not supported by the `slapnap` pipeline. In this case, the container avoids training machine learning algorithms and report generation, returning a data set and associated documentation. In the following call, `return` only includes `"data"`; thus, options pertaining to the machine learning portions of `slapnap` are essentially ignored by `slapnap`. The inputted `outcomes` are also irrelevant, as all `outcomes` are included in the resultant data set.

```
docker run -v /path/to/local/dir:/home/output \
           -e return="data" \
           slapnap/slapnap
```

4.6 Interactive sessions

To simply enter the container and poke around, use an interactive session by including `-it` and overriding the container's entry-point.

```
docker run -it slapnap/slapnap /bin/bash
```

This will enter you into the container in a bash terminal prior to any portions of the analysis being run. This may be useful for exploring the file structure, examining versions of `R` packages that are included in the container, etc.

To enter the container interactively *after* the analysis has run, you can execute the following commands. Here we add the `-d` option to start the container in detached mode.

```
docker run -d -p 80:80 -e view_port="TRUE" slapnap/slapnap

# ...wait for analysis to finish...

# use this command to enter the container
docker exec -it /bin/bash
```

To close the interactive session type `exit` at the container command prompt and hit `Return`. This will close the container and stop its running.

5 Methods

5.1 Outcomes

5.1.1 Single bNAb

If a single bNAb or combination of bNAbs that are measured directly in the CATNAP data is requested (i.e., the `nab` option is a single string of a single bNAb or measured combination of bNAbs from the CATNAP database), then the possible outcomes are:

- `ic50` = IC_{50}: the half maximal inhibitory concentration;
- `ic80` = IC_{80}: the 80% maximal inhibitory concentration;
- `iip` = IIP [Shen et al., 2008, Wagh et al. [2016]]: the instantaneous inhibitory potential, computed as

$$\frac{10^m}{\text{IC}_{50}{}^m + 10^m},$$

where $m = \log_{10}(4)/(\log_{10}(\text{IC}_{80}) - \log_{10}(\text{IC}_{50}))$; and

- `sens` = sensitivity: the binary indicator that $\text{IC}_{50} <$ `sens_thresh`, the user-specified sensitivity threshold.

5.1.2 Multiple bNAbs

If multiple bNAbs are requested (i.e., the `nab` option is a semi-colon separated string of more than one bNAb from CATNAP), then the possible `outcomes` that can be requested are:

- `ic50` = estimated IC_{50}: for J bNAbs,

$$\text{estimated IC}_{50} = \left(\sum_{j=1}^{J} \text{IC}_{50,j}^{-1} \right)^{-1},$$

  where $\text{IC}_{50,j}$ denotes the measured IC_{50} for antibody j [Wagh et al., 2016];
- `ic80` = estimated IC_{80}: for J bNAbs,

$$\text{estimated IC}_{80} = \left(\sum_{j=1}^{J} \text{IC}_{80,j}^{-1} \right)^{-1},$$

  where $\text{IC}_{80,j}$ denotes the measured IC_{80} for antibody j [Wagh et al., 2016];
- `iip` = IIP: computed as

$$\frac{10^m}{\text{estimated IC}_{50}{}^m + 10^m},$$

  where $m = \log_{10}(4)/(\log_{10}(\text{estimated IC}_{80}) - \log_{10}(\text{estimated IC}_{50}))$; and
- `estsens` = estimated sensitivity: the binary indicator that estimated IC_{50} (defined above) is less than `sens_thresh`; and
- `multsens` = multiple sensitivity: the binary indicator that measured IC_{50} is less than the sensitivity threshold (`sens_thresh`) for a number of bNAbs defined by `multsens_nab`.

5.2 Learners

There are three possible `learners` available in `slapnap`: random forests [Breiman, 2001], as implemented in the `R` package `ranger` [Wright and Ziegler, 2017]; elastic net [Zou and Hastie, 2005] as implemented in `glmnet` [Friedman et al., 2010]; and boosted trees [Friedman, 2001, Chen and Guestrin, 2016] as implemented in `xgboost` [Chen et al., 2019].

For each `learner`, there is a `default` choice of tuning parameters that is implemented if `cvtune="FALSE"`. If instead `cvtune="TRUE"`, then there are several choices of tuning parameters that are evaluated using `nfold` cross validation, Table 1.

Tuning parameters not mentioned in the table are set as follows:

- `rf`: `num.trees = 500`, `min.node.size = 5` for continuous outcomes and `= 1` for binary outcomes;
- `xgboost`: `nrounds = 1000`, `eta = 0.1`, `min_child_weight = 10`, objective = `binary:logistic` for binary outcomes and `objective=reg:squarederror` for continuous outcomes.

5.3 Super learner

If multiple `learners` are specified, then a super learner ensemble [van der Laan et al., 2007] is constructed using `nfold` cross validation, as implemented in the `R` package `SuperLearner` [Polley et al., 2019]. Specifically, the data are randomly partitioned into `nfold` chunks of approximately equal size. For binary outcomes, this partitioning is done in such a way as to ensure an approximately even number of sensitive/resistant

Table 1: Labels for ‘learners‘ in report and description of their respective tuning parameters

‘learner‘	Tuning parameters
‘rf_default‘	‘mtry‘ equal to square root of number of predictors
‘rf_1‘	‘mtry‘ equal to one-half times square root of number of predictors
‘rf_2‘	‘mtry‘ equal to two times square root of number of predictors
‘xgboost_default‘	maximum tree depth equal to 4
‘xgboost_1‘	maximum tree depth equal to 2
‘xgboost_2‘	maximum tree depth equal to 6
‘xgboost_3‘	maximum tree depth equal to 8
‘lasso_default‘	λ selected by 5-fold CV and α equal to 0
‘lasso_1‘	λ selected by 5-fold CV and α equal to 0.25
‘lasso_2‘	λ selected by 5-fold CV and α equal to 0.5
‘lasso_3‘	λ selected by 5-fold CV and α equal to 0.75

pseudoviruses in each chunk. A so-called super learner *library* of candidate algorithms is constructed by including different `learners`:

- the algorithm `mean`, which reports back the sample mean as prediction for all observations is always included;
- if `cvtune="FALSE"` then the `default` version of each `learner` (Section 5.2) is included;
- if `cvtune="TRUE"` then each choice of tuning parameters for the selected `learners` in Table 1 is included.

The cross-validated risk of each algorithm in the library is computed. For binary outcomes, mean negative log-likelihood loss is used; for continuous outcomes, mean squared-error is used. The single algorithm with the smallest cross-validated risk is reported as the `cv selector` (also known as the *discrete* super learner). The super learner ensemble is constructed by selecting convex weights (i.e., each algorithm is assigned a non-negative weight and the weights sum to one) that minimize cross-validated risk.

When `cvperf="TRUE"` and a super learner is constructed, an additional layer of cross validation is used to evaluate the predictive performance of the super learner and of the `cv selector`.

5.4 Variable importance

If `importance_grp` or `importance_ind` is specified, variable importance estimates are computed based on the `learners`. Both biological and prediction importance can be obtained; we discuss each in the following two sections.

5.4.1 Biological importance

Biological importance may be obtained by specifying `importance_grp`, `importance_ind`, or both. We provide two types of biological importance: marginal and conditional, accessed by passing `"marg"` and `"cond"`, respectively, to one of the importance variables. Both types of biological importance are based on the population prediction potential of features [Williamson et al., 2020a], as implemented in the `R` package `vimp` [Williamson et al., 2020b]. We measure prediction potential using nonparametric R^2 for continuous outcomes (i.e., IC_{50}, IC_{80}, or IIP) and using the nonparametric area under the receiver operating characteristic curve (AUC) for binary outcomes (i.e., sensitivity, estimated sensitivity, or multiple sensitivity). In both marginal and conditional importance, we compare the population prediction potential including the feature(s) of interest to the population prediction potential excluding the feature(s) or interest; this provides a measure of the biological importance of the feature(s). The two types of biological importance differ only in the other adjustment variables that we consider: conditional importance compares the prediction potential of all

features to the prediction potential of all features excluding the feature(s) of interest, and thus importance must be interpreted conditionally; whereas marginal importance compares the prediction potential of the feature(s) of interest plus geographic confounders to the prediction potential of the geographic confounders alone.

Both marginal and conditional biological importance can be computed for groups of features or individual features. The available feature groups are detailed in Section 7. Execution time may increase when biological importance is requested, depending upon the other options passed to `slapnap`: a separate `learner` (or super learner ensemble) must be trained for each feature group (or individual feature) of interest. Marginal importance tends to be computed more quickly than conditional importance, but both types of importance provide useful information about the population of interest and the underlying biology.

If biological importance is requested, then point estimates, confidence intervals, and p-values (for a test of the null hypothesis that the biological importance is equal to zero) will be computed and displayed for each feature or group of features of interest. All results are based on first creating two independent splits of the data: the population prediction potential including the feature(s) of interest is estimated on one half of the data, while the population prediction potential excluding the feature(s) of interest is estimated on the remaining half. This ensures that the procedure has the desired type I error rate.

In the following command, we request marginal biological importance for the feature groups defined in Section 7. We do not specify a super learner ensemble to reduce computation time; however, in most problems we recommend an ensemble to protect against model misspecification.

```
docker run -v /path/to/local/dir:/home/output \
           -e importance_grp="marg" \
           slapnap/slapnap
```

The raw `R` objects (saved as `.rds` files) containing the point estimates, confidence intervals, and p-values for biological importance can be saved by passing `"vimp"` to `return`.

5.4.2 Predictive importance

`learner`-level predictive importance may be obtained by including `"pred"` in the `importance_ind` option. If a single `learner` is fit, then the predictive importance is the `R` default for that type of learner:

- `rf`: the `impurity` importance from `ranger` [Wright and Ziegler, 2017] is returned. The impurity importance for a given feature is computed by taking a normalized sum of the decrease in impurity (i.e., Gini index for binary outcomes; mean squared-error for continuous outcomes) over all nodes in the forest at which a split on that feature has been conducted.
- `xgboost`: the `gain` importance from `xgboost` [Chen et al., 2019] is returned. Interpretation is essentially the same as for `rf`'s `impurity` importance.
- `lasso`: the absolute value of the estimated regression coefficient at the cross-validation-selected λ is returned.

Note that these importance measures each have important limitations: the `rf` and `xgboost` measures will tend to favor features with many levels, while the `lasso` variable importance will tend to favor features with few levels. Nevertheless, these commonly reported measures can provide some insight into how a given learner is making predictions.

If multiple `learners` are used, and thus a super learner is constructed, then the importance measures for the `learner` with the highest weight in the super learner are reported.

If a single `learner` is used, but `cvtune="TRUE"` then importance measures for the `cv selector` are reported.

In the following command, we request predictive importance for a simple scenario. Predictive importance is displayed for the top 15 features.

```
docker run -v /path/to/local/dir:/home/output \
           -e importance_ind="pred" \
           slapnap/slapnap
```

6 Report

The `slapnap` report consists of an executive summary followed by results for each requested outcome.

The executive summary contains:

- descriptions of `outcomes` (including how any derived outcomes are generated);
- descriptive statistics detailing the number of sequences extracted from CATNAP, the number of sequences with complete feature and outcome information, and the number of estimated sensitive and resistant sequences (defined based on sensitivity, estimated sensitivity, and/or multiple sensitivity);
- a table describing the `learners` used to predict each outcome;
- a table of cross-validated prediction performance for each outcome (if `cvperf = TRUE`);
- a table of ranked marginal biological prediction performance for each feature group and outcome (if `"marg"` is included in `importance_grp`); and
- a table of ranked conditional biological prediction performance for each feature group and outcome (if `"cond"` is included in `importance_grp`).

The rest of the report is organized by outcome. Each of these sections contains descriptive statistics including summaries of the distribution of the outcome (raw and log-transformed) for each bNAb for continuous outcomes and number sensitive/resistant for binary outcomes. Based on the specific options passed to `slapnap`, the following subsections may also be present:

- a table of super learner weights (Section 5.3) if an ensemble is used;
- cross-validated prediction performance for the fitted learner (or super learner): figures showing cross-validated prediction performance (all outcomes), cross-validated receiver operating characteristic (ROC) curves (binary outcomes), and cross-validated predicted probabilities of resistance (binary outcomes); and
- variable importance: biological importance (group and individual) and predictive importance.

Finally, if group biological importance is requested, then the variable groups are displayed in a section immediately preceding the references.

7 Data

The analysis dataset includes neutralization outcomes for the requested bNAb(s) and AA sequence features for the gp160 protein. The possible outcomes are described in Section 5.1.

The additional groups of variables in the data include:

- **geographic information**: binary indicator variables describing the region of origin of each pseudovirus;
- **subtype**: binary indicator variables denoting the HIV-1 subtype for the given pseudovirus;
- **AA sequence features**: binary indicator variables denoting presence or absence of a residue containing a specific AA at each HXB2-referenced site in gp160;
- **viral geometry features**: length of Env, gp120, V2, V3, V5;
- **numbers of sequons**: number of sequons in Env, gp120, V2, V3, V5; and
- **numbers of cysteines**: number of cysteines in Env, gp120, V2, V3, V5.

8 References

References

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Tianqi Chen and Carlos Guestrin. xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. *xgboost: Extreme Gradient Boosting*, 2019. URL https://CRAN.R-project.org/package=xgboost. R package version 0.82.1.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001. doi: 10.1214/aos/1013203451.

Eric Polley, Erin LeDell, Chris Kennedy, and Mark van der Laan. *SuperLearner: Super Learner Prediction*, 2019. URL https://CRAN.R-project.org/package=SuperLearner. R package version 2.0-25.

Lin Shen, Susan Peterson, Ahmad R Sedaghat, Moira A McMahon, Marc Callender, Haili Zhang, Yan Zhou, Eleanor Pitt, Karen S Anderson, Edward P Acosta, et al. Dose-response curve slope sets class-specific limits on inhibitory potential of anti-HIV drugs. *Nature Medicine*, 14(7):762–766, 2008. doi: 10.1038/nm1777.

Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. doi: 10.2202/1544-6115.1309.

Kshitij Wagh, Tanmoy Bhattacharya, Carolyn Williamson, Alex Robles, Madeleine Bayne, Jetta Garrity, Michael Rist, Cecilia Rademeyer, Hyejin Yoon, Alan Lapedes, et al. Optimal combinations of broadly neutralizing antibodies for prevention and treatment of HIV-1 clade C infection. *PLoS Pathogens*, 12(3), 2016. doi: 10.1371/journal.ppat.1005520.

Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A unified approach for inference on algorithm-agnostic variable importance. *arXiv preprint*, 2020a. URL https://arxiv.org/abs/2004.03683.

Brian D. Williamson, Noah Simon, and Marco Carone. vimp: Perform inference on algorithm-agnostic variable importance. 2020b. URL https://CRAN.R-project.org/package=vimp. R package version 2.0.2.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

Hyejin Yoon, Jennifer Macke, Anthony P West Jr, Brian Foley, Pamela J Bjorkman, Bette Korber, and Karina Yusim. CATNAP: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Research*, 43(W1):W213–W219, 2015. doi: 10.1093/nar/gkv404.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.