

# Package ‘ProteinGymR’

April 29, 2025

**Title** Programmatic access to ProteinGym datasets in R/Bioconductor

**Version** 1.3.0

**Description** The ProteinGymR package provides analysis-ready data resources from ProteinGym, generated by Notin et al., 2023, as well as built-in functionality to visualize the data. ProteinGym comprises a collection of benchmarks for evaluating the performance of models predicting the effect of point mutations. This package provides access to 1. deep mutational scanning (DMS) scores from 217 assays measuring the impact of all possible amino acid substitutions across 186 proteins, 2. model performance metrics and prediction scores from 79 variant prediction models in the zero-shot setting and 12 models in the semi-supervised setting.

**License** Artistic-2.0

**URL** <https://github.com/ccb-hms/ProteinGymR>

**BugReports** <https://github.com/ccb-hms/ProteinGymR/issues>

**Depends** R (>= 4.4.0)

**Imports** ExperimentHub, AnnotationHub, bio3d, r3dmol, ggExtra, dplyr,forcats, ggdist, gghalves, ggplot2, purrr, queryup, spdl, tidyverse, tidyselect, ComplexHeatmap, circlize, stringr

**Suggests** tibble, BiocStyle, knitr, testthat (>= 3.0.0)

**VignetteBuilder** knitr

**Encoding** UTF-8

**biocViews** ExperimentData, ExperimentHub, PackageTypeData, Homo\_sapiens\_Data, ReproducibleResearch, CellCulture, SequencingData, Proteome

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.3.2

**Config/testthat.edition** 3

**git\_url** <https://git.bioconductor.org/packages/ProteinGymR>

**git\_branch** devel

**git\_last\_commit** c788e3e

**git\_last\_commit\_date** 2025-04-25

**Repository** Bioconductor 3.22

**Date/Publication** 2025-04-29

**Author** Tram Nguyen [aut, cre] (ORCID: <<https://orcid.org/0000-0003-4809-6227>>),  
 Pascal Notin [aut],  
 Aaron Kollasch [aut],  
 Debora Marks [aut],  
 Ludwig Geistlinger [aut]

**Maintainer** Tram Nguyen <Tram\_Nguyen@hms.harvard.edu>

## Contents

am_scores . . . . .	2
available_models . . . . .	3
dms_corr_plot . . . . .	4
dms_substitutions . . . . .	5
plot_dms_heatmap . . . . .	6
plot_structure . . . . .	8
plot_zeroshot_heatmap . . . . .	10
supervised_substitutions . . . . .	11
zeroshot_DMS_metrics . . . . .	13

<b>Index</b>	<b>15</b>
--------------	-----------

---

am_scores	<i>AlphaMissense scores for ProteinGym variants</i>
-----------	---

---

### Description

AlphaMissense scores for ProteinGym variants

### Usage

```
am_scores(metadata = FALSE)
```

### Arguments

metadata	Logical, whether only experiment metadata should be returned. Default behavior is to return processed data with metadata included.
----------	--

### Details

am\_scores() loads in the AlphaMissense pathogenicity scores for substitutions matching those in the ProteinGym DMS assays. The table is taken from the AlphaMissense Supplementary Data by Cheng et al. 2023. See reference for details.

The columns contain:

DMS\_id: Character, ProteinGym assay identifier.

Uniprot\_ID: Character, UniProt accession identifier.

variant\_id: Character, variant identifier string matching ProteinGym. Protein position in the middle, and the reference and mutant amino acid residues to the left and right of the position, respectively.

AlphaMissense: Numeric, AlphaMissense pathogenicity score.

**Value**

Returns a `data.frame()`.

**References**

Cheng et al. (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 391, eadg7492. DOI:10.1126/science.adg7492.

**Examples**

```
data <- am_scores()  
data_meta <- am_scores(metadata = TRUE)
```

---

available\_models

*Benchmark Variant Effect Prediction Models*

---

**Description**

`benchmark_models()` plots one of the five model performance metrics ("AUC", "MCC", "NDCG", "Spearman", "Top\_recall") for up to 5 user-specified variant effect prediction tools listed in `available_models()`. See reference for more details about the metrics and models.

**Usage**

```
available_models()  
  
supervised_available_models()  
  
benchmark_models(  
  metric = c("AUC", "MCC", "NDCG", "Spearman", "Top_recall"),  
  models = available_models()  
)
```

**Arguments**

<code>metric</code>	character() a model performance metric to benchmark ("AUC", "MCC", "NDCG", "Spearman", "Top_recall").
<code>models</code>	character() a character vector of up to five variant effect prediction models to compare. Valid models can be seen with <code>available_models()</code> .

**Value**

`benchmark_models()` returns a `ggplot` object visualizing a chosen model performance metric across several variant effect prediction models, ordered by highest to lowest mean performance score.

## References

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

## Examples

```
# Currently support models
available_models()

benchmark_models(metric = "Spearman", models = c("Site_Independent",
"DeepSequence_single", "ESM2_15B", "GEMME", "CARP_640M"))

benchmark_models(models = "GEMME")
```

dms\_corr\_plot

*Integrate ProteinGym DMS and AlphaMissense Pathogenicity Scores*

## Description

`dms_corr_plot()` runs a Spearman correlation between ProteinGym deep mutational scanning (DMS) assay scores and AlphaMissense predicted pathogenicity scores. Returns a `ggplot` object for visualization.

## Usage

```
dms_corr_plot(uniprotId, alphamissense_table, dms_table)
```

## Arguments

<code>uniprotId</code>	character() a valid UniProt accession identifier.
<code>alphamissense_table</code>	a table containing AlphaMissense predictions for variants matching ProteinGym substitution mutants. The default is the supplemental table from the AlphaMissense paper. Alternatively, a user-defined <code>tibble::tbl_df</code> or <code>data.frame</code> can be supplied.
<code>dms_table</code>	a table containing deep mutational scanning (DMS) assay scores for mutations. The default table loads substitutions from <code>ProteinGym</code> . Alternatively, a user-defined <code>tibble::tbl_df</code> or <code>data.frame</code> can be supplied.

## Details

For `dms_corr_plot()`, `alphamissense_table` columns must include:

- `UniProt_id`: UniProt accession identifier.
- `mutant`: Mutant identifier string matching the `dms_table` format. Protein position in the middle, and the reference and mutant amino acid residues to the left and right of the position, respectively.

- AlphaMissense: AlphaMissense pathogenicity score.

dms\_table columns must include:

- UniProt\_id: UniProt accession identifier.
- mutant: Mutant identifier string matching AlphaMissense variants. Specifically, the set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').
- DMS\_score: Experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

## Value

dms\_corr\_plot() returns a ggplot object visualizing the Spearman correlation between experimental DMS scores and AlphaMissense predicted scores and prints the r and p-value of the analysis to console. Generally, a stronger negative correlation corresponds to a tighter

## References

Cheng et al., Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381, eadg7492. DOI:10.1126/science.adg7492.

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

## Examples

```
dms_corr_plot(uniprotId = "Q9NV35")
```

---

dms\_substitutions

*ProteinGym Deep Mutational Scanning (DMS) Scores for Substitutions*

---

## Description

ProteinGym Deep Mutational Scanning (DMS) Scores for Substitutions

## Usage

```
dms_substitutions(metadata = FALSE)
```

## Arguments

metadata	Logical, whether only experiment metadata should be returned. Default behavior is to return processed data with metadata included.
----------	--

## Details

`dms_substitutions()` loads in ProteinGym deep mutational scanning assays (DMS) scores for substitutions in 217 studies. The data is provided by Notin et. al 2023. See reference for details.

Each assay includes 6 columns:

`UniProt_id`: Character, UniProt accession identifier.

`DMS_id`: Character, ProteinGym assay identifier.

`mutant`: Character, set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').

`mutated_sequence`: Character, full amino acid sequence for the mutated protein.

`DMS_score`: Numeric, experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

`DMS_score_bin`: Factor, indicates whether the `DMS_score` is above the fitness cutoff (1 is fit, 0 is not fit).

## Value

Returns a [list\(\)](#) object of 217 individual assays.

## References

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

## Examples

```
data <- dms_substitutions()
data_meta <- dms_substitutions(metadata = TRUE)
```

`plot_dms_heatmap`      *Visualize DMS Scores Along a Protein*

## Description

`plot_dms_heatmap()` plots DMS scores for amino acid substitutions along a protein in a defined DMS assay.

## Usage

```
plot_dms_heatmap(
  assay_name,
  dms_data,
  start_pos = NULL,
  end_pos = NULL,
  exact_coord = FALSE,
```

```

  cluster_rows = FALSE,
  cluster_columns = FALSE,
  color_scheme,
  ...
)

```

## Arguments

assay_name	character() a valid DMS assay name. For the full list of available assays, run names() on the list object loaded with ProteinGymR::dms_substitutions(). Alternatively, the name of a user-defined DMS assay.
dms_data	list() object of DMS assays loaded with ProteinGymR::dms_substitutions() if not specified by user. Alternatively, a user-defined list of DMS assays with names corresponding to assay_name param.
start_pos	integer() first amino acid position to plot. If missing, default start is at the first position along the protein where DMS scores are available.
end_pos	integer() last amino acid position to plot. If missing, default end is at the last position along the protein where DMS scores are available.
exact_coord	logical() TRUE will plot the precise start_pos and end_pos coordinates defined. By default, exact_coord is set to FALSE, plotting only amino acid positions with available data in the chosen assay.
cluster_rows	logical() defaults to FALSE. See argument details in <a href="#">ComplexHeatmap::Heatmap</a> .
cluster_columns	logical() defaults to FALSE. See argument details in <a href="#">ComplexHeatmap::Heatmap</a> .
color_scheme	character() defaults to blue, white, and red to represent positive, neutral, negative scores. Set argument equal to "EVE" to use the color scheme consistent with the popEVE portal.
...	additional arguments passed to internal plotting functions.

## Details

For `plot_dms_heatmap()`, `dms_data` must be a `list()` object with set names for each assay element matching `assay_name` parameter.

Each assay in the `dms_data()` must include the following columns:

- `mutant`: Mutant identifier string matching. Specifically, the set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').
- `DMS_score`: Experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

## Value

Returns a [ComplexHeatmap::Heatmap](#) plot of DMS scores for each position along a protein in a chosen DMS assay. The x-axis shows amino acid positions where a DMS mutation exist, and the y-axis represents possible amino acid residues, ordered by default based on the physiochemical groupings. Higher and lower DMS scores indicate a more positive or negative fitness effect after the mutation, respectively.

## Examples

```
dms_data <- dms_substitutions()

plot_dms_heatmap(assay_name = "A0A192B1T2_9HIV1_Haddox_2018",
                  start_pos = 10,
                  end_pos = 80)

plot_dms_heatmap(assay_name = "A0A192B1T2_9HIV1_Haddox_2018",
                  start_pos = 10,
                  end_pos = 80,
                  exact_coord = TRUE,
                  color_scheme = "EVE")

plot_dms_heatmap(assay_name = "A0A192B1T2_9HIV1_Haddox_2018",
                  start_pos = 50,
                  end_pos = 100,
                  cluster_rows = TRUE)
```

**plot\_structure**

*Visualize DMS and Model Scores on 3D Protein Structures*

## Description

`plot_structure()` plots DMS or model scores for amino acid substitutions on a 3D protein structure for a chosen assay.

## Usage

```
plot_structure(
  assay_name,
  pdb_file,
  dms_data,
  start_pos = NULL,
  end_pos = NULL,
  full_structure = FALSE,
  aggregate_fun = mean,
  color_scheme
)
```

## Arguments

<code>assay_name</code>	<code>character()</code> a valid DMS assay name. For the full list of available assays, run <code>names()</code> on the list object loaded with <code>ProteinGymR::dms_substitutions()</code> . Alternatively, the name of a
<code>pdb_file</code>	<code>string()</code> defaults to corresponding PDB FilePath on ExperimentHub. Alternatively, a file path to a user-defined PDB file.
<code>dms_data</code>	<code>list()</code> object of DMS assays loaded with <code>ProteinGymR::dms_substitutions()</code> . Alternatively, a user-defined list of DMS assays with names corresponding to <code>assay_name</code> param.
<code>start_pos</code>	<code>integer()</code> first amino acid position to plot. If missing, default start is the first position along the protein in the PDB file.

end_pos	integer() last amino acid position to plot. If missing, default end is the last position along the protein in the PDB file.
full_structure	logical() defaults to FALSE and will only plot protein regions where there is DMS data available in the assay. If start_pos and end_pos coordinates are specified, plotting is restricted to this defined region. Setting full_structure() to TRUE will display full protein structure in the PBD file, and grey out regions where no DMS data is available.
aggregate_fun	method for aggregating DMS scores for each residue. For example, give <code>min</code> , <code>max</code> , or <code>var</code> to return the minimum, maximum, or variance of scores for each position, respectively. aggregate_fun can also take in a user-defined function with a numeric vector as input. By default, the mean DMS score across mutations at each position is calculated.
color_scheme	character() defaults to blue, white, and red to represent positive, neutral, negative scores. Set argument equal to "EVE" to use the color scheme consistent with the popEVE portal.

## Details

For `plot_structure()`, `dms_data` must be a `list()` object with set names for each assay element matching `assay_name` parameter.

Each assay in the `dms_data()` must include the following columns:

- `mutant`: Mutant identifier string matching. Specifically, the set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').
- `DMS_score`: Experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

Each PBD table in `pdb_file` must include the following columns:

## Value

`plot_structure()` returns a `r3dmol::r3dmol` object of DMS scores for each position along a protein in a chosen DMS assay. The x-axis shows amino acid positions where a DMS mutation exist, and the y-axis represents possible amino acid residues, ordered by default based on the physiochemical groupings. Higher and lower DMS scores indicate a more positive or negative fitness effect after the mutation, respectively.

## Examples

```
# Using default dms_data
plot_structure(assay_name = "ACE2_HUMAN_Ch_2020")

# Use EVE model color scheme
plot_structure(assay_name = "ACE2_HUMAN_Ch_2020",
               color_scheme = "EVE")

plot_structure(assay_name = "C6KNH7_9INFA_Lee_2018",
               start_pos = 20,
               end_pos = 50,
               full_structure = FALSE,
               aggregate_fun = min)
```

`plot_zeroshot_heatmap` *Visualize Zero Shot Scores Along a Protein*

## Description

`plot_zeroshot_heatmap()` plots predicted model scores under the zero-shot model for amino acid substitutions along a protein in a defined DMS assay.

## Usage

```
plot_zeroshot_heatmap(
  assay_name,
  model_data,
  model = NULL,
  start_pos = NULL,
  end_pos = NULL,
  exact_coord = FALSE,
  cluster_rows = FALSE,
  cluster_columns = FALSE,
  color_scheme,
  ...
)
```

## Arguments

<code>assay_name</code>	<code>character()</code> a valid assay name. For the full list of available assays, run <code>names()</code> on the list object loaded with <code>ProteinGymR::zeroshot_substitutions()</code> . Alternatively, the name of a user-defined DMS assay.
<code>model_data</code>	<code>list()</code> object of zero-shot assays loaded with <code>ProteinGymR::zeroshot_substitutions()</code> . Alternatively, a user-defined list of assays with names corresponding to <code>assay_name</code> param.
<code>model</code>	<code>character()</code> one of the 79 zero-shot models to plot. To view the list of models, runs <code>ProteinGymR::available_models()</code> .
<code>start_pos</code>	<code>integer()</code> first amino acid position to plot. If missing, default start is at the first position along the protein where zero shot scores are available.
<code>end_pos</code>	<code>integer()</code> last amino acid position to plot. If missing, default end is at the last position along the protein where zero shot scores are available.
<code>exact_coord</code>	<code>logical()</code> TRUE will plot the precise <code>start_pos</code> and <code>end_pos</code> coordinates defined. By default, <code>exact_coord</code> is set to FALSE, plotting only amino acid positions with available data in the chosen assay.
<code>cluster_rows</code>	<code>logical()</code> defaults to FALSE. See argument details in <a href="#">ComplexHeatmap::Heatmap</a> .
<code>cluster_columns</code>	<code>logical()</code> defaults to FALSE. See argument details in <a href="#">ComplexHeatmap::Heatmap</a> .
<code>color_scheme</code>	<code>character()</code> defaults to blue, white, and red to represent positive, neutral, negative scores. Set argument equal to "EVE" to use the color scheme consistent with the popEVE portal.
<code>...</code>	additional arguments passed to internal plotting functions.

## Details

For `plot_zeroshot_heatmap()`, `model_data` must be a `list()` object with set names for each assay element matching `assay_name` parameter.

Each assay in the `model_data()` must include the following columns:

- `mutant`: Mutant identifier string matching. Specifically, the set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').
- A column with predicted model scores for each substitution. Column name should match `model` argument string.

## Value

Returns a [ComplexHeatmap::Heatmap](#) plot of zero-shot model scores for each position along a protein in a chosen DMS substitution assay. The x-axis shows amino acid positions where a DMS mutation exist, and the y-axis represents possible amino acid residues, ordered by default based on the physiochemical groupings. Higher and lower DMS scores indicate a more positive or negative fitness effect after the mutation, respectively.

## Examples

```
available_models()

model_data <- zeroshot_substitutions()

plot_zeroshot_heatmap(assay_name = "A0A192B1T2_9HIV1_Haddox_2018",
                      model_data = model_data,
                      model = "GEMME",
                      start_pos = 600,
                      end_pos = 700,
                      color_scheme = "EVE")

plot_zeroshot_heatmap(assay_name = "SRC_HUMAN_Nguyen_2022",
                      model_data = model_data,
                      model = "CARP_38M")

plot_zeroshot_heatmap(assay_name = "A0A192B1T2_9HIV1_Haddox_2018",
                      model_data = model_data,
                      model = "EVE_ensemble",
                      start_pos = 10,
                      end_pos = 80,
                      exact_coord = TRUE)
```

## supervised\_substitutions

*Load Semi-Supervised Model Predictions for Substitutions in 217 Assays*

## Description

Load Semi-Supervised Model Predictions for Substitutions in 217 Assays

Load Semi-Supervised Model Summary Metrics

## Usage

```
supervised_substitutions(
  metadata = FALSE,
  fold_scheme = c("contiguous", "modulo", "random")
)
supervised_metrics(metadata = FALSE)
```

## Arguments

<b>metadata</b>	Logical, whether only experiment metadata should be returned. Default behavior is to return processed data with metadata included.
<b>fold_scheme</b>	Character, which validation folding scheme to load. Options include: "contiguous", "modulo", or "random". Default behavior loads "contiguous". For more information about the different folding schemes, refer to the original publication.

## Details

`supervised_substitutions()` loads prediction scores outputted by semi-supervised models run on the 217 DMS substitution assays.

For raw model predictions, each assay includes 18 columns:

**UniProt\_id:** Character, UniProt accession identifier.

**DMS\_id:** Character, ProteinGym assay identifier.

**mutant:** Character, set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').

**mutated\_sequence:** Character, full amino acid sequence for the mutated protein.

**DMS\_score:** Numeric, experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

**DMS\_score\_bin:** Factor, indicates whether the DMS\_score is above the fitness cutoff (1 is fit, 0 is not fit).

**Columns 7:18:** Respective semi-supervised model name.

`supervised_metrics()` loads in model performance summary metrics ("Spearman" and "MSE") from semi-supervised models in ProteinGymR run on 217 DMS substitution assays.

A metric summary table with 7 columns:

**UniProt\_id:** Character, UniProt accession identifier.

**DMS\_id:** Character, ProteinGym assay identifier.

**mutant:** Character, set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').

**model\_name:** Character, semi-supervised model used.

**fold\_variable\_name:** Character, the folding scheme used.

**Spearman:** Numeric, Spearman performance metric.

**MSE:** Numeric, MSE the Spearman performance metric.

**Value**

Returns a [list\(\)](#) object of 217 individual assays.

Returns a [data.frame\(\)](#) with 7 columns.

**References**

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

**Examples**

```
data <- supervised_substitutions()
data_random <- supervised_substitutions(fold_scheme = "random")
meta <- supervised_substitutions(metadata = TRUE)

data <- supervised_metrics()
meta <- supervised_metrics(metadata = TRUE)
```

**zeroshot\_DMS\_metrics** *Load Zero-shot Model Predictions and Metrics for Substitutions in 217 DMS Assays*

**Description**

Load Zero-shot Model Predictions and Metrics for Substitutions in 217 DMS Assays

**Usage**

```
zeroshot_DMS_metrics(metadata = FALSE)

zeroshot_substitutions(metadata = FALSE)
```

**Arguments**

metadata	Logical, whether only experiment metadata should be returned. Default behavior is to return processed data with metadata included.
----------	--

**Details**

`zeroshot_DMS_metrics()` loads in the model performance metrics ("AUC", "MCC", "NDCG", "Spearman", "Top\_recall") calculated on the DMS substitutions in the zero-shot setting for 79 models updated in ProteinGym v1.2.

Each data.frame contains the following columns:

**DMS\_ID:** Character, Assay name for the DMS study.

**Columns 2:80:** Numeric, Corresponding to the average performance score of each of the 79 models tested.

**Number\_of\_Mutants:** Numeric, Number of protein mutants evaluated.

**Selection\_Type:** Character, Protein function grouping.

**UniProt\_ID:** Character, UniProt protein entry name identifier.

**DMS\_score:** Numeric, experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

**DMS\_score\_bin:** Factor, indicates whether the DMS\_score is above the fitness cutoff (1 is fit, 0 is not fit).

**MSA\_Neff\_L\_category:** Multiple sequence alignment category.

**Taxon:** Taxon group.

`zeroshot_substitutions()` loads prediction scores outputted by models in the zero-shot setting evaluated on the 217 DMS substitution assays. To examine all model options, run `available_models()`.

For raw model predictions, each data.frame includes 85 columns:

**UniProt\_id:** Character, UniProt accession identifier.

**DMS\_id:** Character, ProteinGym assay identifier.

**mutant:** Character, set of substitutions to apply on the reference sequence to obtain the mutated sequence (e.g., A1P:D2N implies the amino acid 'A' at position 1 should be replaced by 'P', and 'D' at position 2 should be replaced by 'N').

**mutated\_sequence:** Character, full amino acid sequence for the mutated protein.

**DMS\_score:** Numeric, experimental measurement in the DMS assay. Higher values indicate higher fitness of the mutated protein.

**DMS\_score\_bin:** Factor, indicates whether the DMS\_score is above the fitness cutoff (1 is fit, 0 is not fit).

**Columns 7:85:** Respective zero-shot model name.

## Value

Returns a `list()` object with 5 `data.frame()` corresponding to a model metric table.

Returns a `list()` object of 217 individual assays.

## References

Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., & Marks, D. (2023). ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36, pp. 64331-64379). Curran Associates, Inc.

## Examples

```
data <- zeroshot_DMS_metrics()
data_meta <- zeroshot_DMS_metrics(metadata = TRUE)

data <- zeroshot_substitutions()
data_meta <- zeroshot_substitutions(metadata = TRUE)
```

# Index

am\_scores, 2  
available\_models, 3  
  
benchmark\_models (available\_models), 3  
  
ComplexHeatmap::Heatmap, 7, 10, 11  
  
data.frame, 4  
data.frame(), 3, 13, 14  
dms\_corr\_plot, 4  
dms\_substitutions, 5  
  
list(), 6, 13, 14  
  
max, 9  
min, 9  
  
plot\_dms\_heatmap, 6  
plot\_structure, 8  
plot\_zeroshot\_heatmap, 10  
  
r3dmol::r3dmol, 9  
  
supervised\_available\_models  
    (available\_models), 3  
supervised\_metrics  
    (supervised\_substitutions), 11  
supervised\_substitutions, 11  
  
tibble::tbl\_df, 4  
  
var, 9  
  
zeroshot\_DMS\_metrics, 13  
zeroshot\_substitutions  
    (zeroshot\_DMS\_metrics), 13