WeightedLinearEnsemble evaluation

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
library(comprehenr)
```

```
## Warning: package 'comprehenr' was built under R version 4.0.5
```

```
library(stringr)
library(ungeviz)
library(relayer)
```

```
## Note: The package "relayer" is highly experimental. Use at your own risk.
```

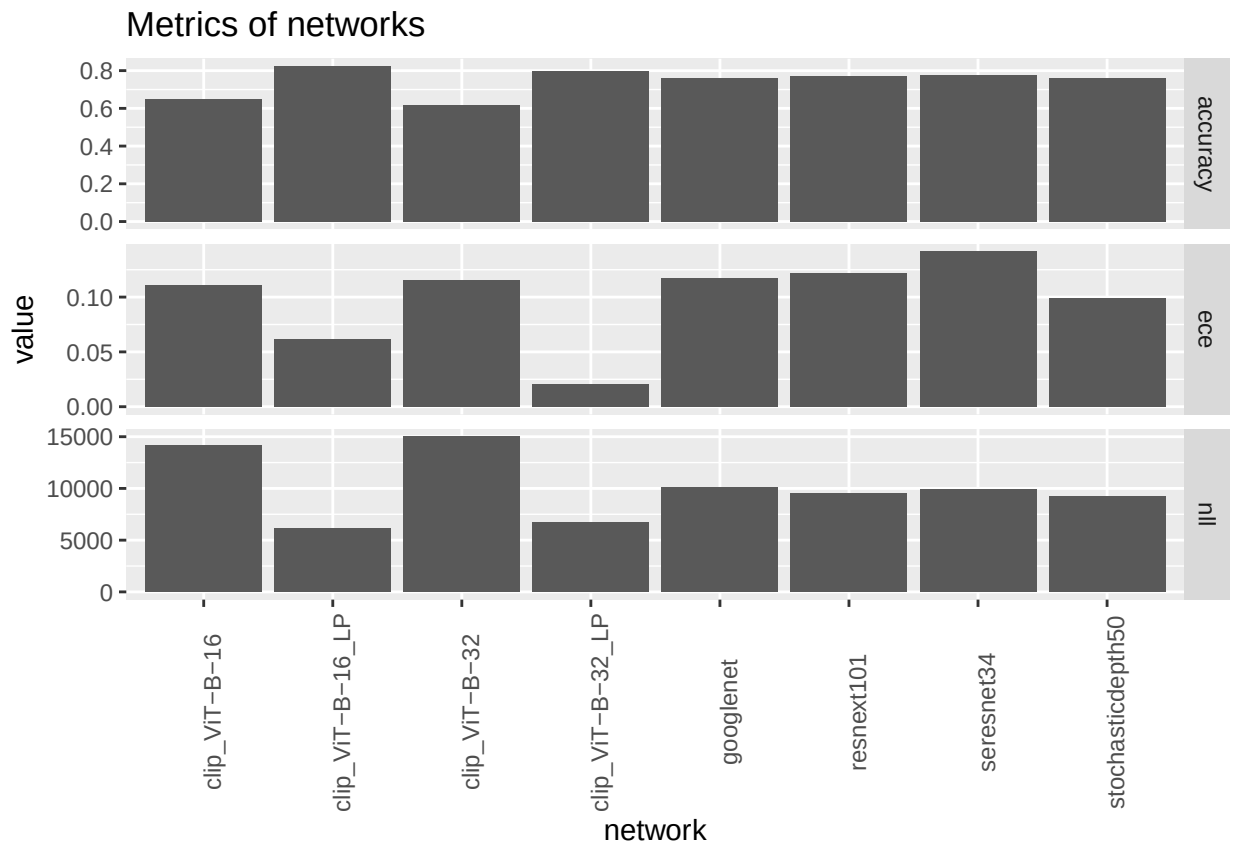
```
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.0.3
```

```
base_dir <- "D:/skola/1/weighted_ensembles/tests/test_cifar_2021/data/data_tv_5000_c100/0/"
net_df <- read.csv(file.path(base_dir, "net_metrics.csv"))
ens_df_cal <- read.csv(file.path(base_dir, "ens_cal_metrics_temp.csv"))
ens_df_pwc <- read.csv(file.path(base_dir, "ens_pwc_metrics_temp.csv"))

net_long <- pivot_longer(net_df, cols=c("accuracy", "nll", "ece"),
  names_to="metric", values_to="value")
ens_cal_long <- pivot_longer(ens_df_cal, cols=c("accuracy", "nll", "ece"),
  names_to="metric", values_to="value")
ens_pwc_long <- pivot_longer(ens_df_pwc, cols=c("accuracy", "nll", "ece"),
  names_to="metric", values_to="value")

networks <- net_df$network

comb_stats_df = data.frame(matrix(ncol=14, nrow=0,
  dimnames=list(NULL, c(
    "combination_size", "combination_id",
    "acc_min", "acc_max", "acc_avg", "acc_var",
    "nll_min", "nll_max", "nll_avg", "nll_var",
    "ece_min", "ece_max", "ece_avg", "ece_var"))))

for (sss in unique(ens_df_cal$combination_size))
{
  for (ssi in unique(ens_df_cal %>%
                     filter(combination_size == sss) %>%
                     pull(combination_id)))
  {
    cur_nets_vec <- to_vec(
      for (net in networks)
      if (str_replace_all(net, "-", ".") %in% colnames(ens_cal_long) &&
          (ens_cal_long %>%
            filter(combination_size == sss & combination_id == ssi) %>%
            pull(str_replace_all(net, "-", ".")))[1] == "True")
      net)
    cur_nets <- net_df %>% filter(network %in% cur_nets_vec)
    comb_stats_df[nrow(comb_stats_df) + 1, ] = c(sss, ssi,
      min(cur_nets$accuracy), max(cur_nets$accuracy), mean(cur_nets$accuracy), var(cur_nets$accuracy),
      min(cur_nets$nll), max(cur_nets$nll), mean(cur_nets$nll), var(cur_nets$nll),
      min(cur_nets$ece), max(cur_nets$ece), mean(cur_nets$ece), var(cur_nets$ece))

  }
}

ens_df_cal <- merge(ens_df_cal, comb_stats_df)
ens_df_cal$acc_imp_avg = ens_df_cal$accuracy - ens_df_cal$acc_avg
ens_df_cal$acc_imp_max = ens_df_cal$accuracy - ens_df_cal$acc_max

ens_df_pwc <- merge(ens_df_pwc, comb_stats_df)
ens_df_pwc$acc_imp_avg = ens_df_pwc$accuracy - ens_df_pwc$acc_avg
ens_df_pwc$acc_imp_max = ens_df_pwc$accuracy - ens_df_pwc$acc_max
```

```
nets_plot <- ggplot(data=net_long) +
  geom_col(mapping=aes(x=network, y=value)) +
```

```
  facet_grid(rows = vars(metric), scales="free") +
  theme(axis.text.x = element_text(angle=90)) +
  ggtitle("Metrics of networks")

nets_plot
```



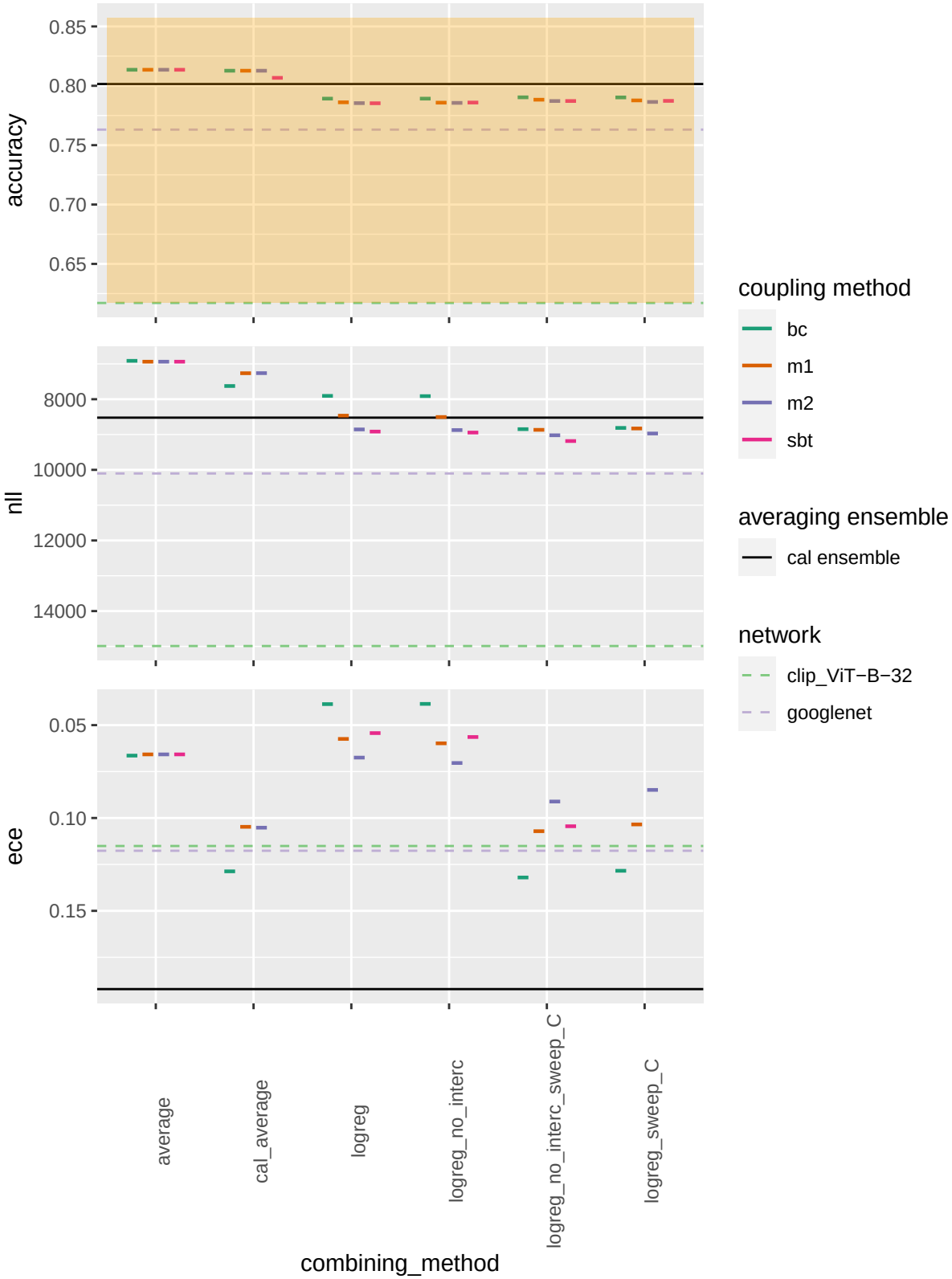

```
ens_pwc_plt_df <- ens_df_pwc %>% filter(combining_method != "lda")
ens_cal_plt_df <- ens_df_cal

for (sss in unique(ens_cal_plt_df$combination_size))
{
  for (ssi in unique(ens_cal_plt_df %>%
                       filter(combination_size == sss) %>%
                       pull(combination_id)))
  {
    cur_ens_cal <- ens_cal_plt_df %>% filter(combination_size == sss &
                                             combination_id == ssi)
    cur_ens_pwc <- ens_pwc_plt_df %>% filter(combination_size == sss &
                                             combination_id == ssi)
    cur_nets_vec <- to_vec(
      for (net in networks)
      if (str_replace_all(net, "-", ".") %in% colnames(cur_ens_cal) &&
            cur_ens_cal[[str_replace_all(net, "-", ".")]][1] == "True")
      net)
    cur_nets <- net_df %>% filter(network %in% cur_nets_vec)
```

```
acc_plot <- ggplot() +
  (
    geom_hline(
      data=cur_nets,
      mapping=aes(yintercept=accuracy, colour1=network),
      linetype="dashed") %>%
    rename_geom_aes(new_aes = c("colour" = "colour1"))
  ) +
  geom_hline(
    data=cur_ens_cal,
    mapping=aes(yintercept=accuracy, color="cal ensemble")) +
  (
    geom_hpline(
      data=cur_ens_pwc,
      mapping=aes(x=combining_method, y=accuracy,
                  colour2=coupling_method),
      size=0.8, width=0.11,
      position=position_dodge(width=0.65)) %>%
    rename_geom_aes(new_aes = c("colour" = "colour2"))
  ) +
  scale_colour_brewer(aesthetics="colour1", palette=1,
    name="network", type="qual") +
  scale_colour_brewer(aesthetics="colour2", palette=2,
    name="coupling method", type="qual") +
  scale_color_manual(values=c("black"), name="averaging ensemble") +
  theme(
    axis.text.x = element_blank(),
    axis.title.x = element_blank())

y_limits <- layer_scales(acc_plot)$y$get_limits()
x_limits <- layer_scales(acc_plot)$x$get_limits()
all_y_lim <- c(y_limits[1], cur_ens_cal$all_cor)

acc_plot <- acc_plot +
  geom_rect(
    data=cur_ens_cal,
    mapping=aes(
      xmin=0.5,
      xmax=length(x_limits) + 0.5,
      ymin=max(all_cor, y_limits[1]), ymax=all_cor + err_incons),
      fill="orange",
      alpha=0.3, color=NA)

if (all_y_lim[1] < all_y_lim[2])
{
  acc_plot <- acc_plot +
    geom_rect(
      data=cur_ens_cal,
      mapping=aes(
        xmin=0.5,
        xmax=length(x_limits) + 0.5,
        ymin=all_y_lim[1], ymax=all_y_lim[2]), fill="green",
        alpha=0.3, color=NA)
```

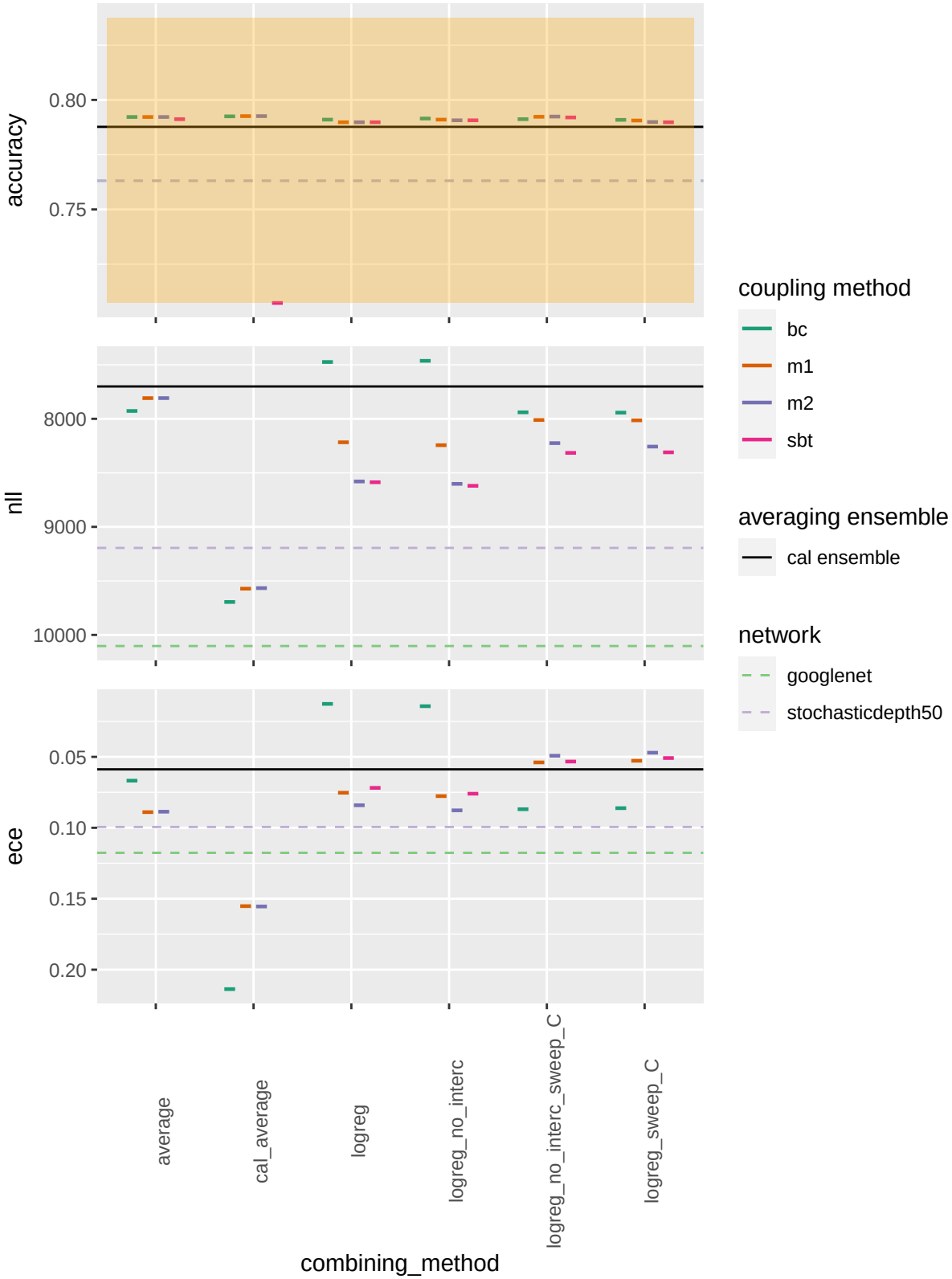
```
}

nll_plot <- ggplot() +
  (
    geom_hline(
      data=cur_nets,
      mapping=aes(yintercept=nll, colour1=network),
      linetype="dashed") %>%
    rename_geom_aes(new_aes = c("colour" = "colour1"))
  ) +
  geom_hline(
    data=cur_ens_cal,
    mapping=aes(yintercept=nll, color="cal ensemble")) +
  (
    geom_hpline(
      data=cur_ens_pwc,
      mapping=aes(x=combining_method, y=nll,
                  colour2=coupling_method),
      size=0.8, width=0.11,
      position=position_dodge(width=0.65)) %>%
    rename_geom_aes(new_aes = c("colour" = "colour2"))
  ) +
  scale_colour_brewer(aesthetics="colour1", palette=1,
    name="network", type="qual") +
  scale_colour_brewer(aesthetics="colour2", palette=2,
    name="coupling method", type="qual") +
  scale_color_manual(values=c("black"), name="averaging ensemble") +
  scale_y_reverse() +
  theme(
    axis.text.x = element_blank(),
    axis.title.x = element_blank())

ece_plot <- ggplot() +
  (
    geom_hline(
      data=cur_nets,
      mapping=aes(yintercept=ece, colour1=network),
      linetype="dashed") %>%
    rename_geom_aes(new_aes = c("colour" = "colour1"))
  ) +
  geom_hline(
    data=cur_ens_cal,
    mapping=aes(yintercept=ece, color="cal ensemble")) +
  (
    geom_hpline(
      data=cur_ens_pwc,
      mapping=aes(x=combining_method, y=ece,
                  colour2=coupling_method),
      size=0.8, width=0.11,
      position=position_dodge(width=0.65)) %>%
    rename_geom_aes(new_aes = c("colour" = "colour2"))
  ) +
  scale_colour_brewer(aesthetics="colour1", palette=1,
```

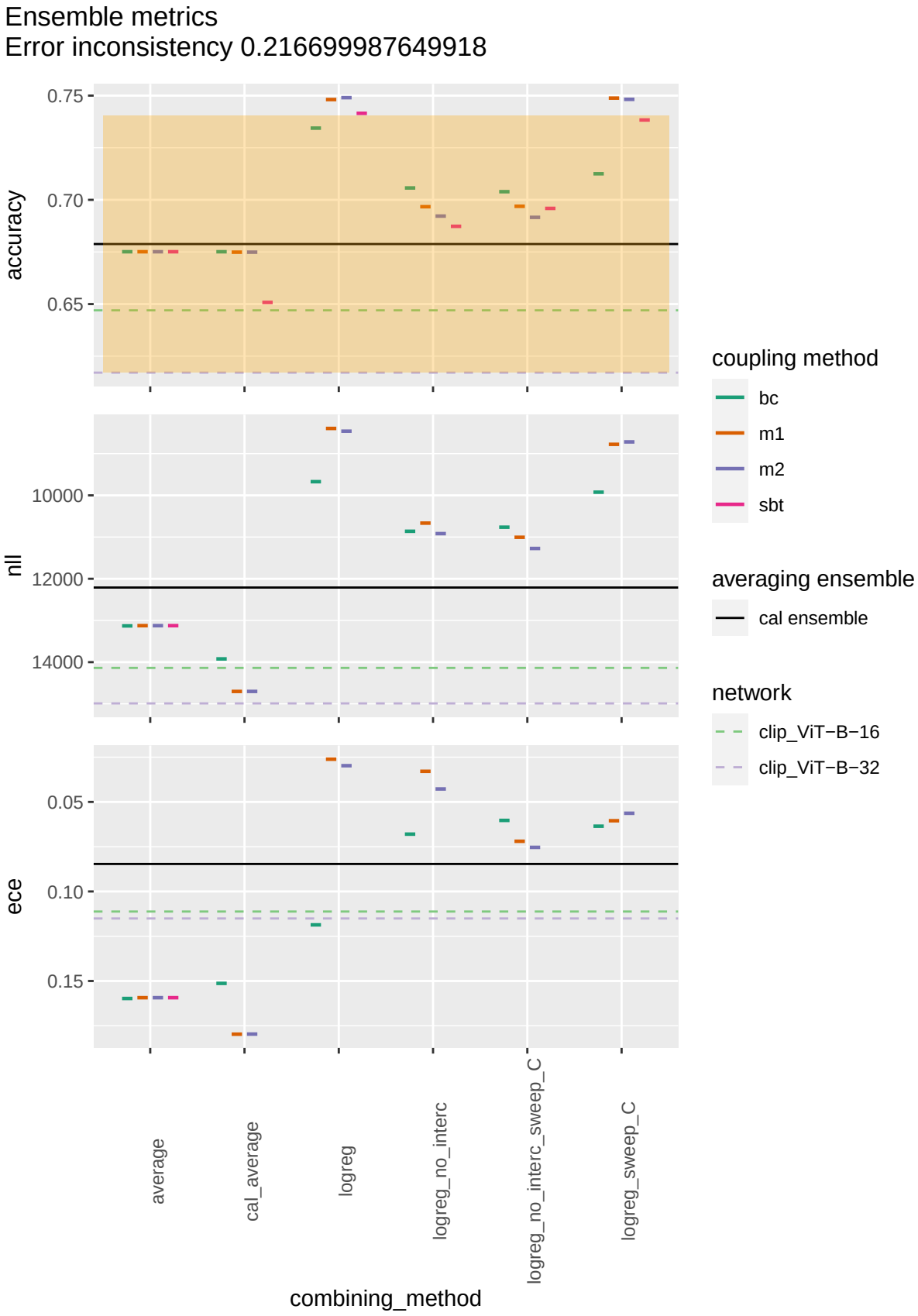
```
        name="network", type="qual") +
      scale_colour_brewer(aesthetics="colour2", palette=2,
        name="coupling method", type="qual") +
      scale_color_manual(values=c("black"), name="averaging ensemble") +
      scale_y_reverse() +
      theme(axis.text.x = element_text(angle=90))

    print(acc_plot / nll_plot / ece_plot + plot_layout(guides="collect") +
      plot_annotation(title=paste(
        "Ensemble metrics",
        paste(
          c("Error inconsistency", cur_ens_cal$err_incons[[1]]),
          collapse=" "),
        sep="\n")))
  }
}
```

Ensemble metrics
Error inconsistency 0.334399998188019

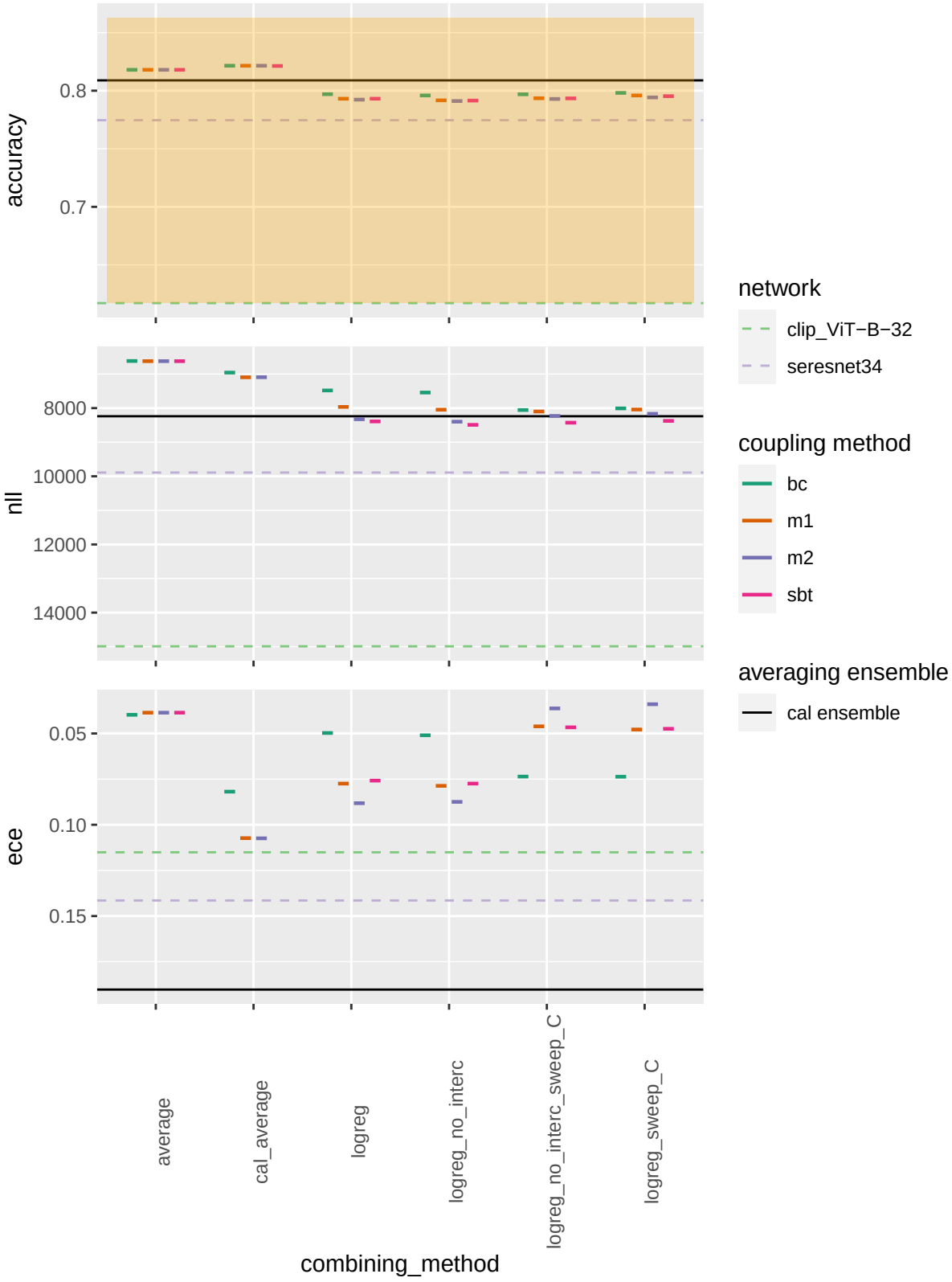


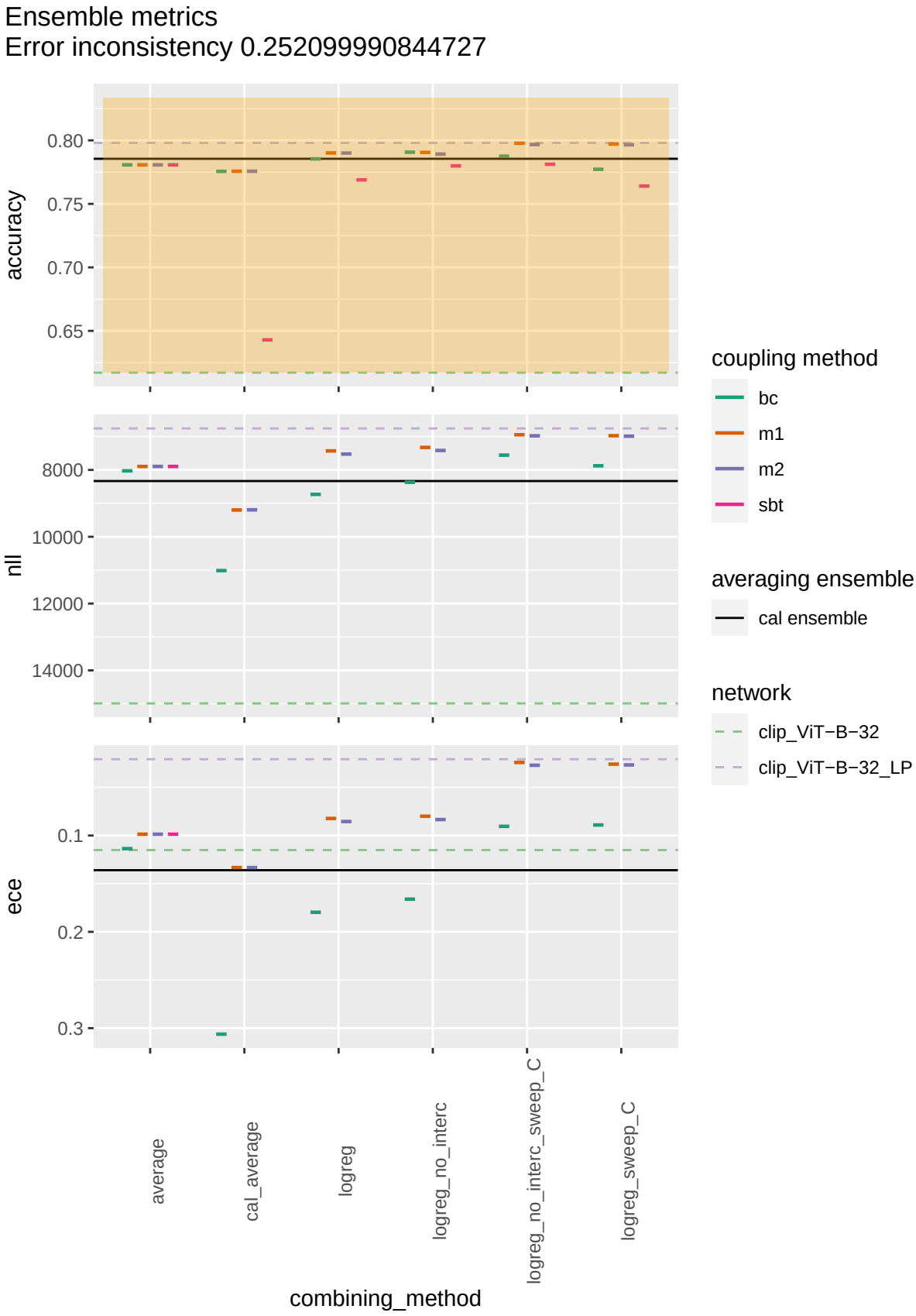
Ensemble metrics
Error inconsistency 0.148800000548363

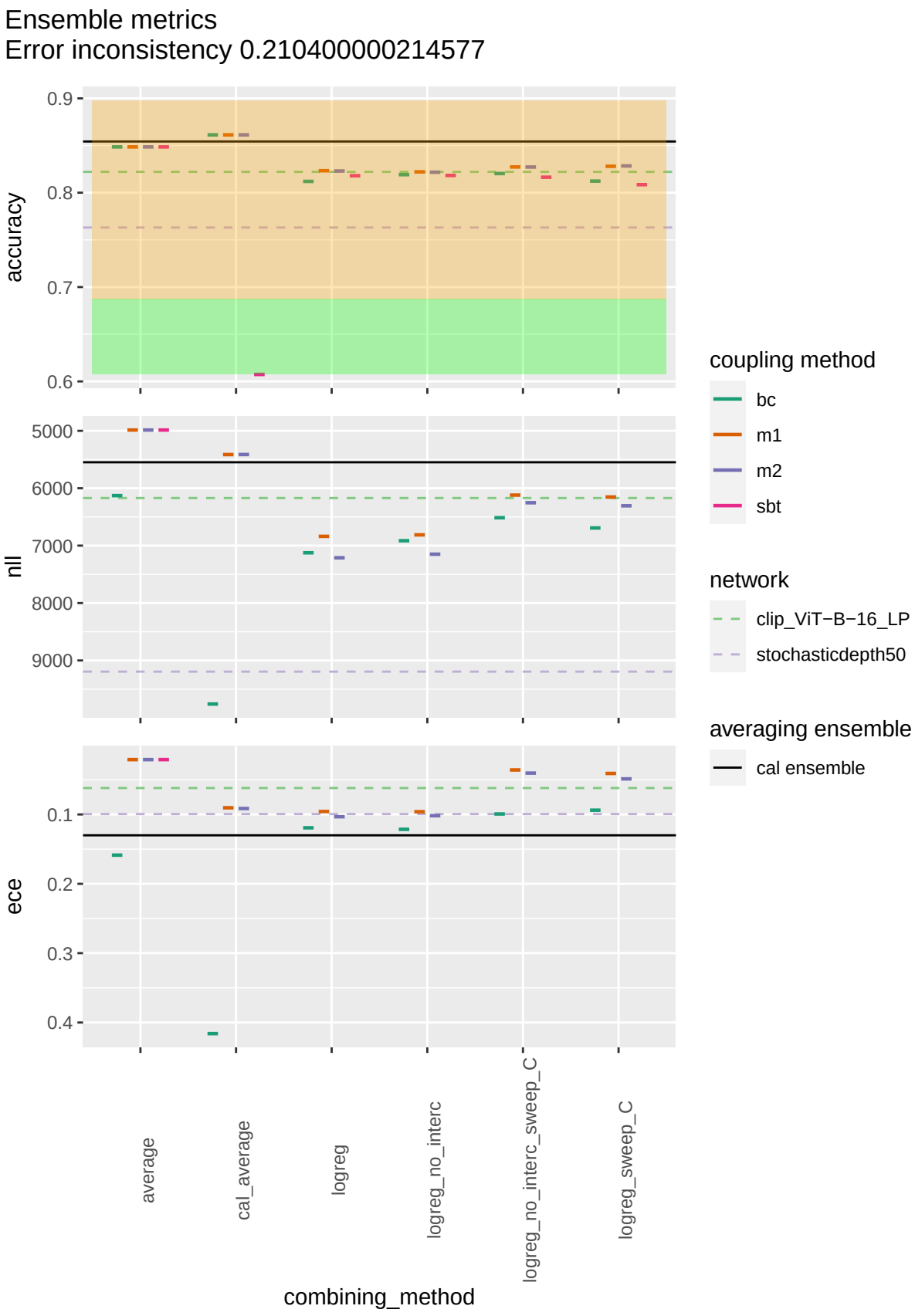


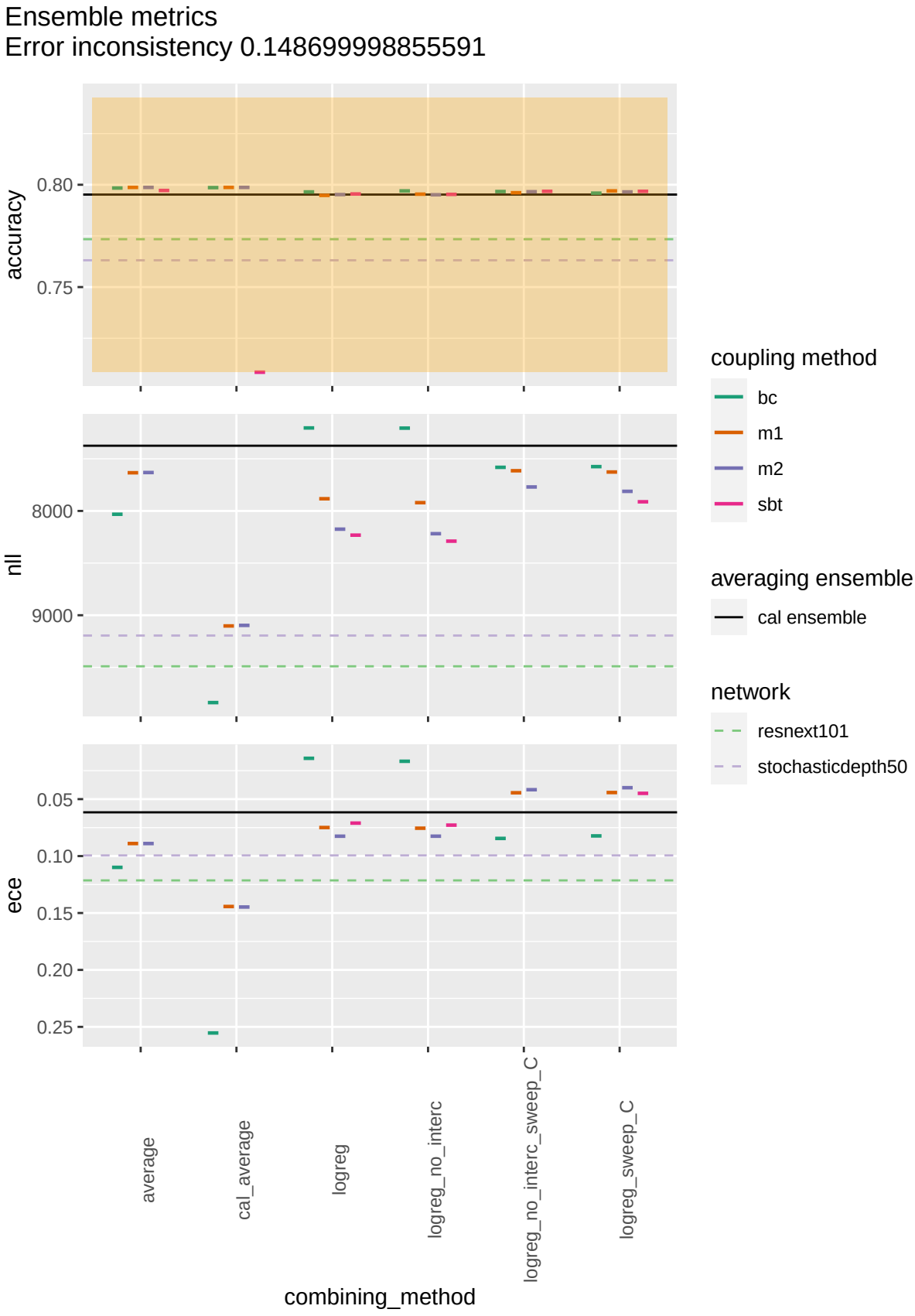

Ensemble metrics
Error inconsistency 0.216699987649918
accuracy
0.75
0.70
0.65
nll
10000
12000
14000
ece
0.05
0.10
0.15
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−32

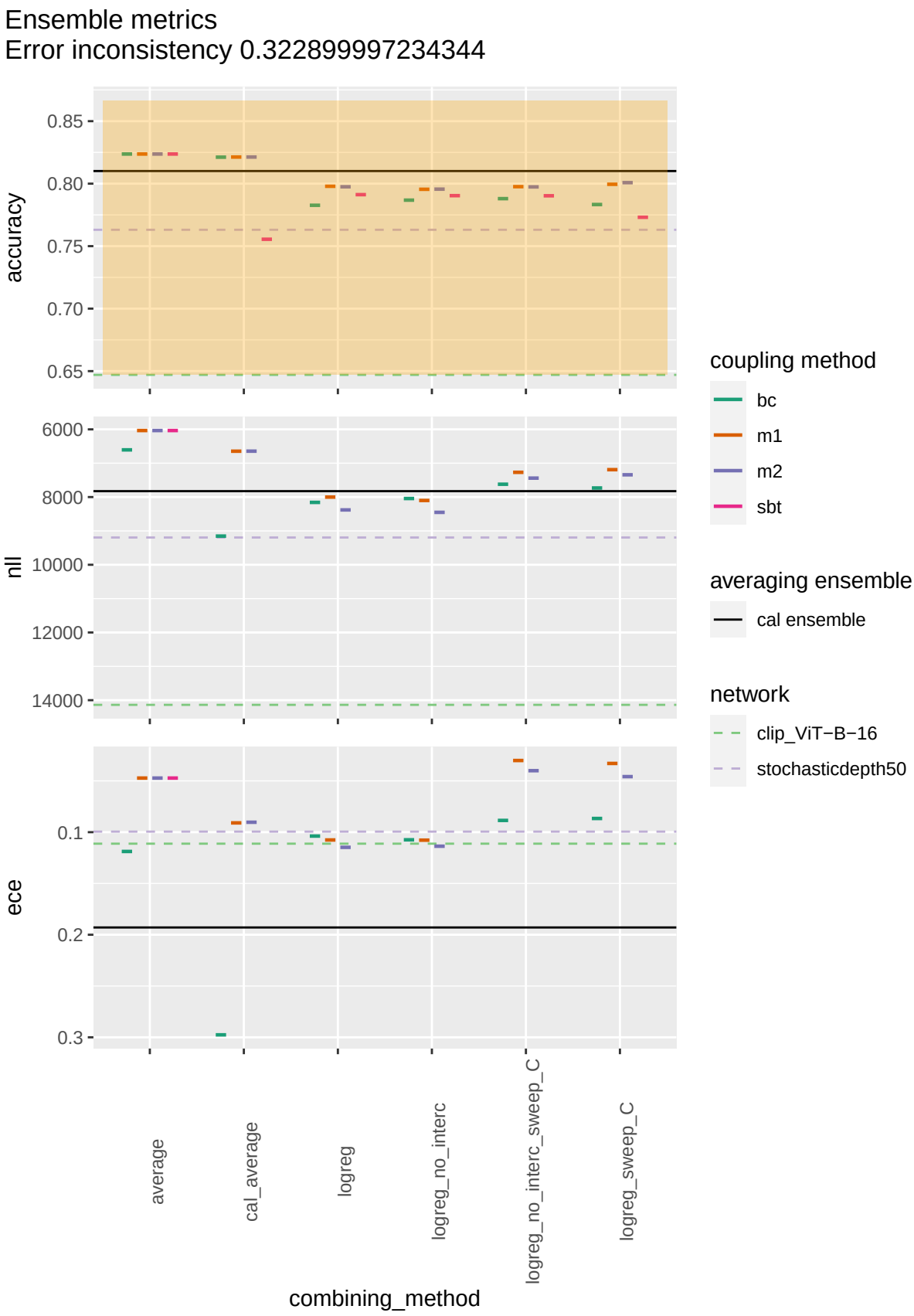
Ensemble metrics
Error inconsistency 0.33390000462532

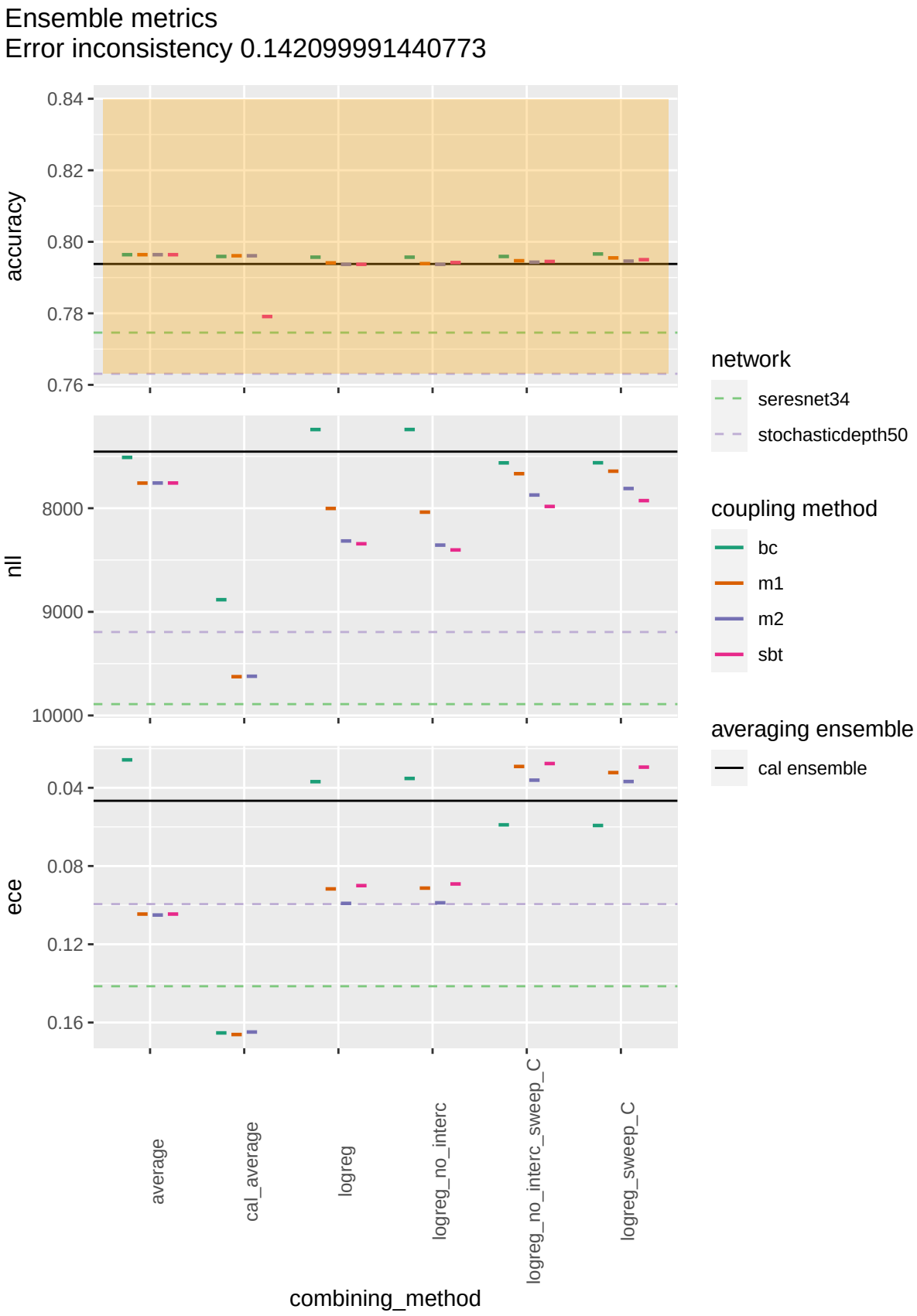


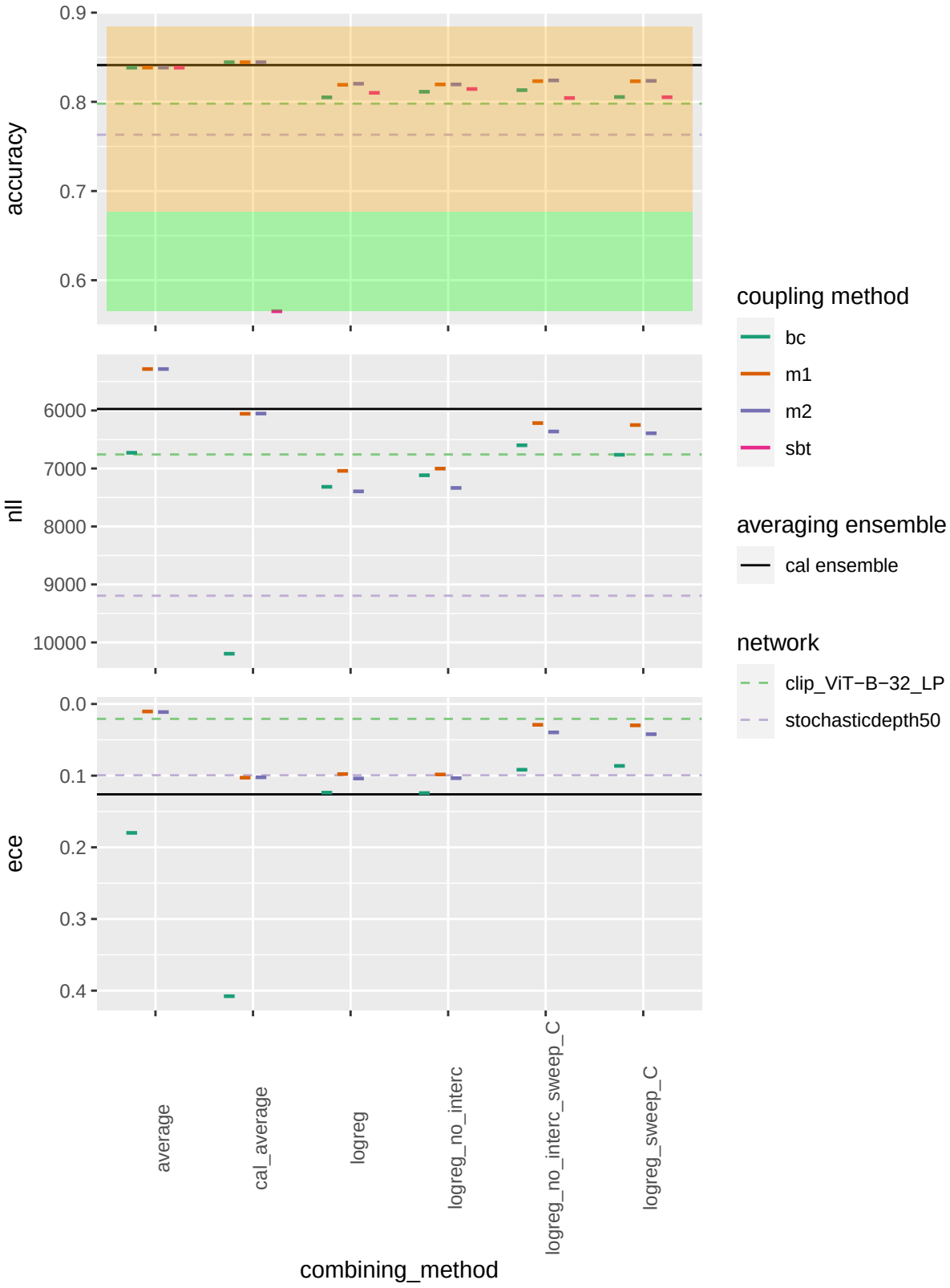

Ensemble metrics
Error inconsistency 0.252099990844727
accuracy
0.80
0.75
0.70
0.65
nll
8000
10000
12000
14000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
clip_ViT−B−32_LP

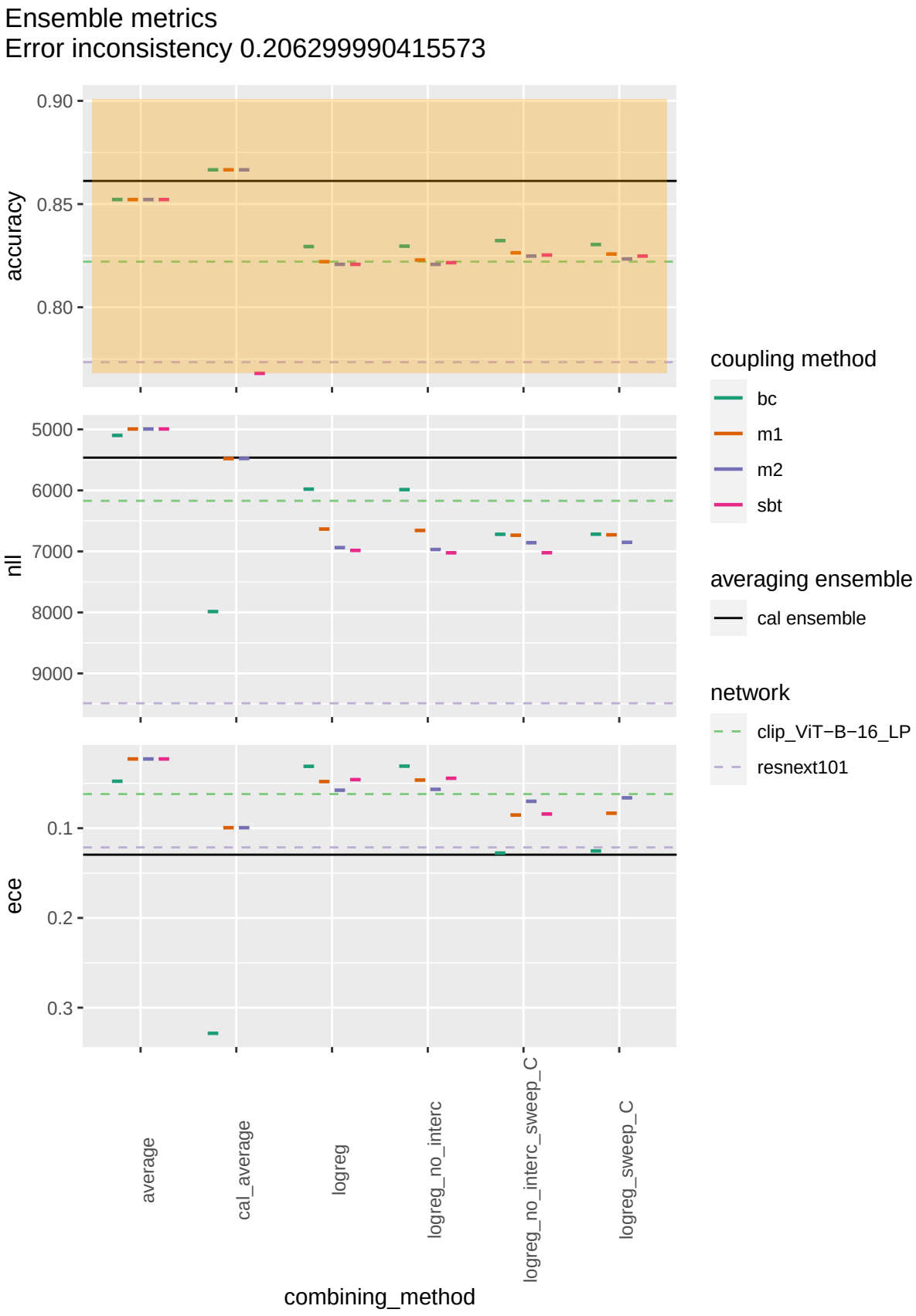

Ensemble metrics
Error inconsistency 0.210400000214577
accuracy
nll
ece
0.9
0.8
0.7
0.6
5000
6000
7000
8000
9000
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
network
clip_ViT−B−16_LP
stochasticdepth50
averaging ensemble
cal ensemble

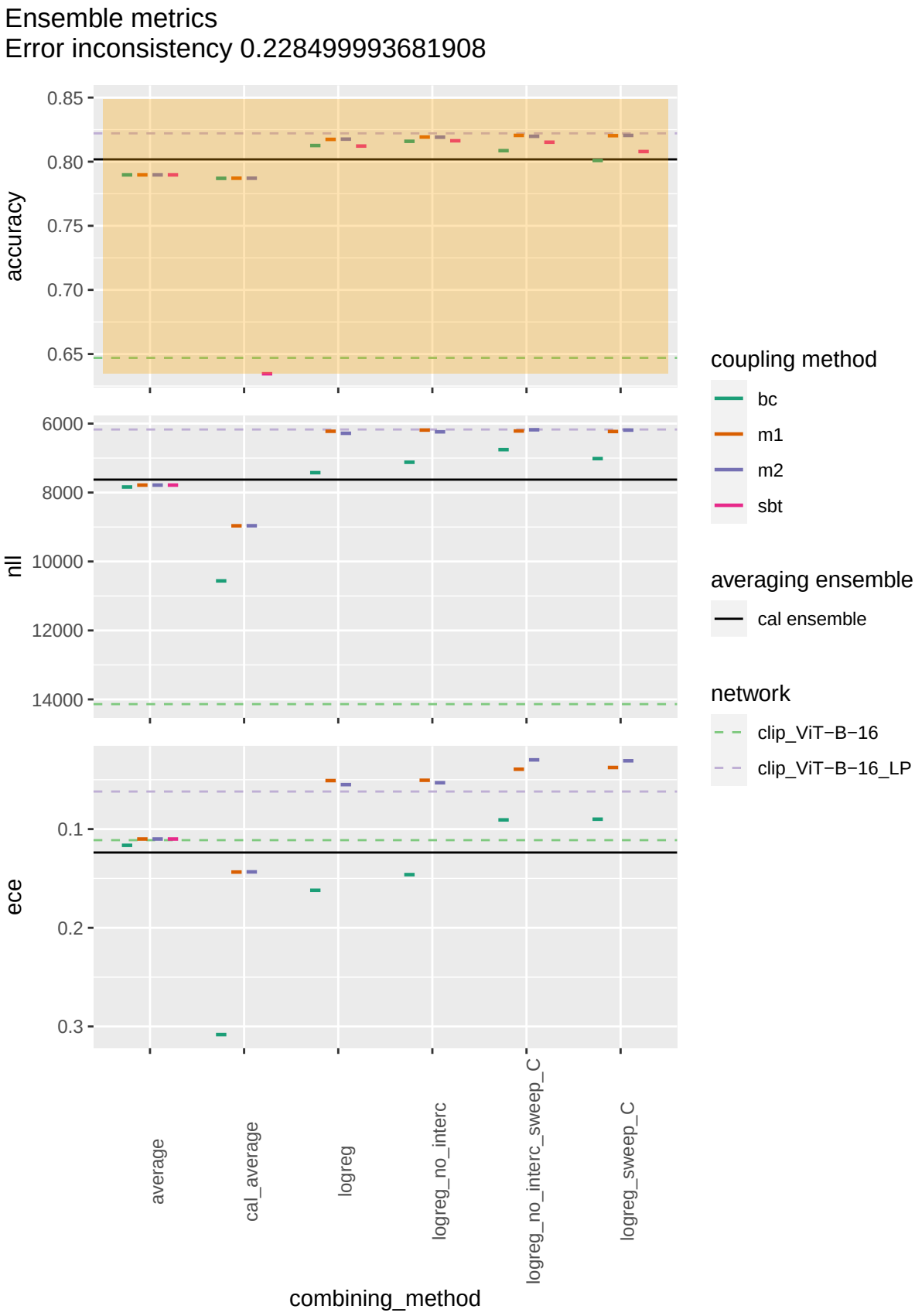

Ensemble metrics
Error inconsistency 0.148699998855591
accuracy
0.80
0.75
nll
8000
9000
ece
0.05
0.10
0.15
0.20
0.25
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
resnext101
stochasticdepth50

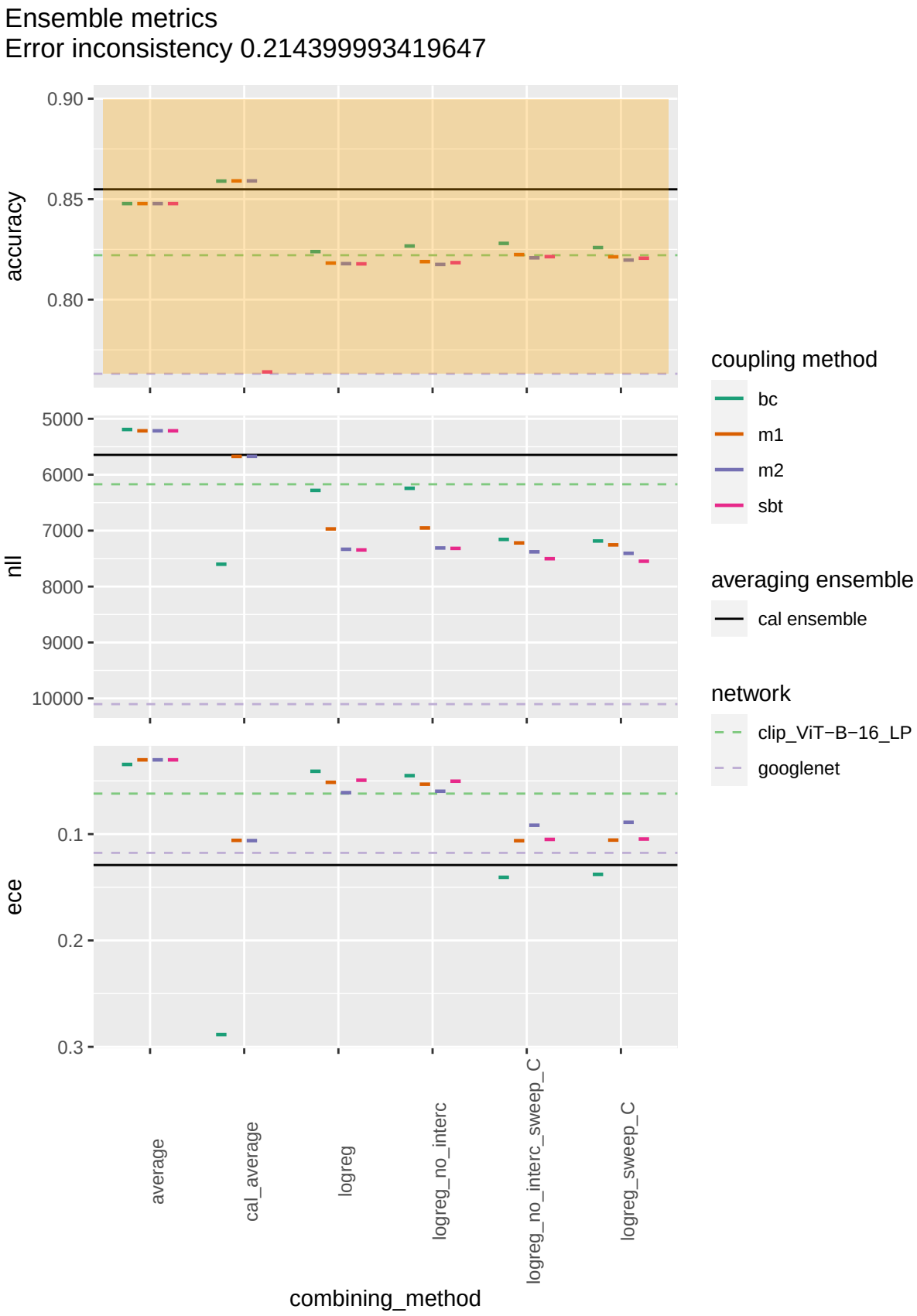

Ensemble metrics
Error inconsistency 0.322899997234344
accuracy
0.85
0.80
0.75
0.70
0.65
nll
6000
8000
10000
12000
14000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
stochasticdepth50


Ensemble metrics
Error inconsistency 0.142099991440773
accuracy
0.84
0.82
0.80
0.78
0.76
nll
8000
9000
10000
ece
0.04
0.08
0.12
0.16
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
network
seresnet34
stochasticdepth50
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble

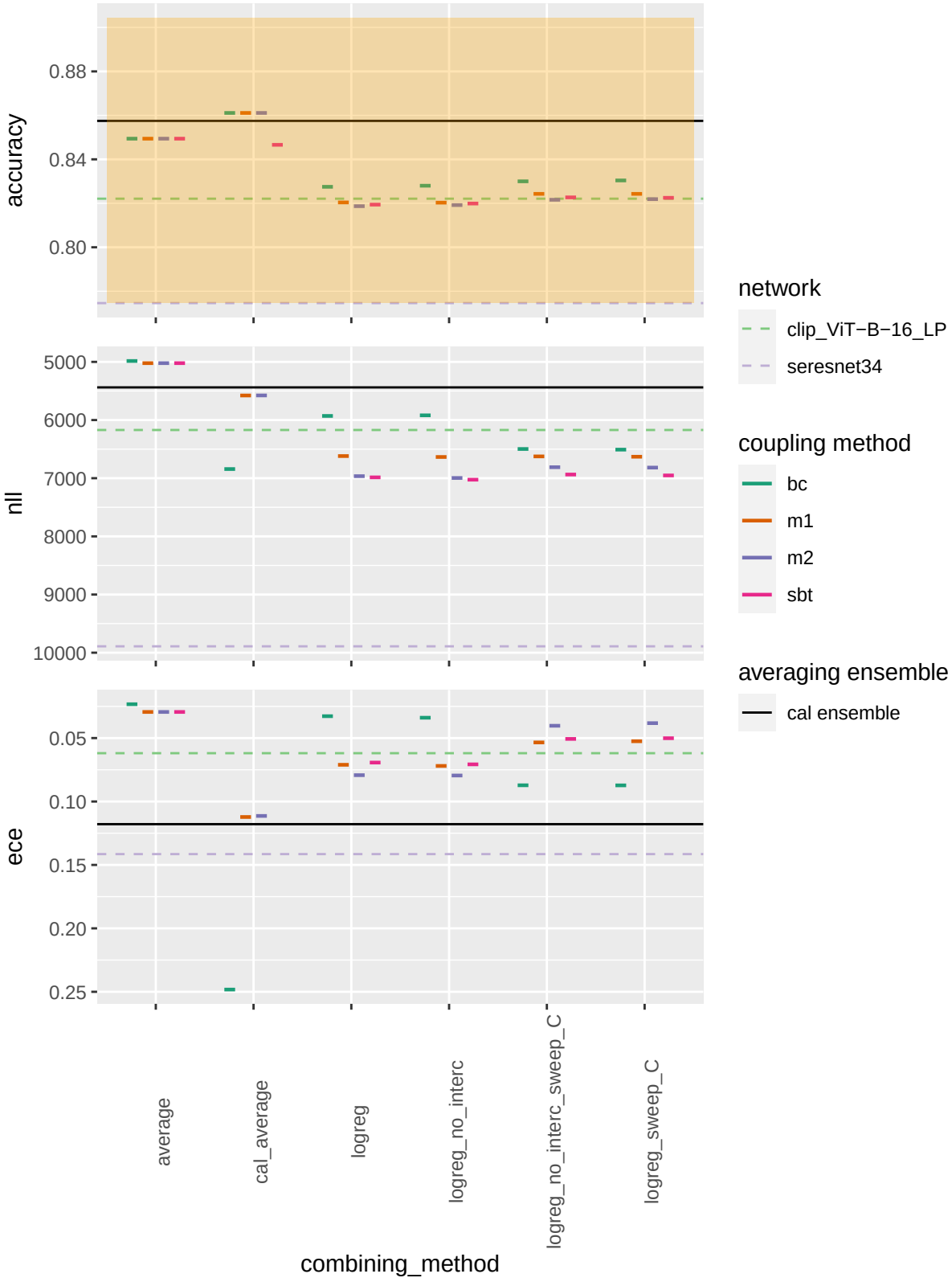

Ensemble metrics
Error inconsistency 0.207699999213219
accuracy
0.9
0.8
0.7
0.6
nll
6000
7000
8000
9000
10000
ece
0.0
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32_LP
stochasticdepth50

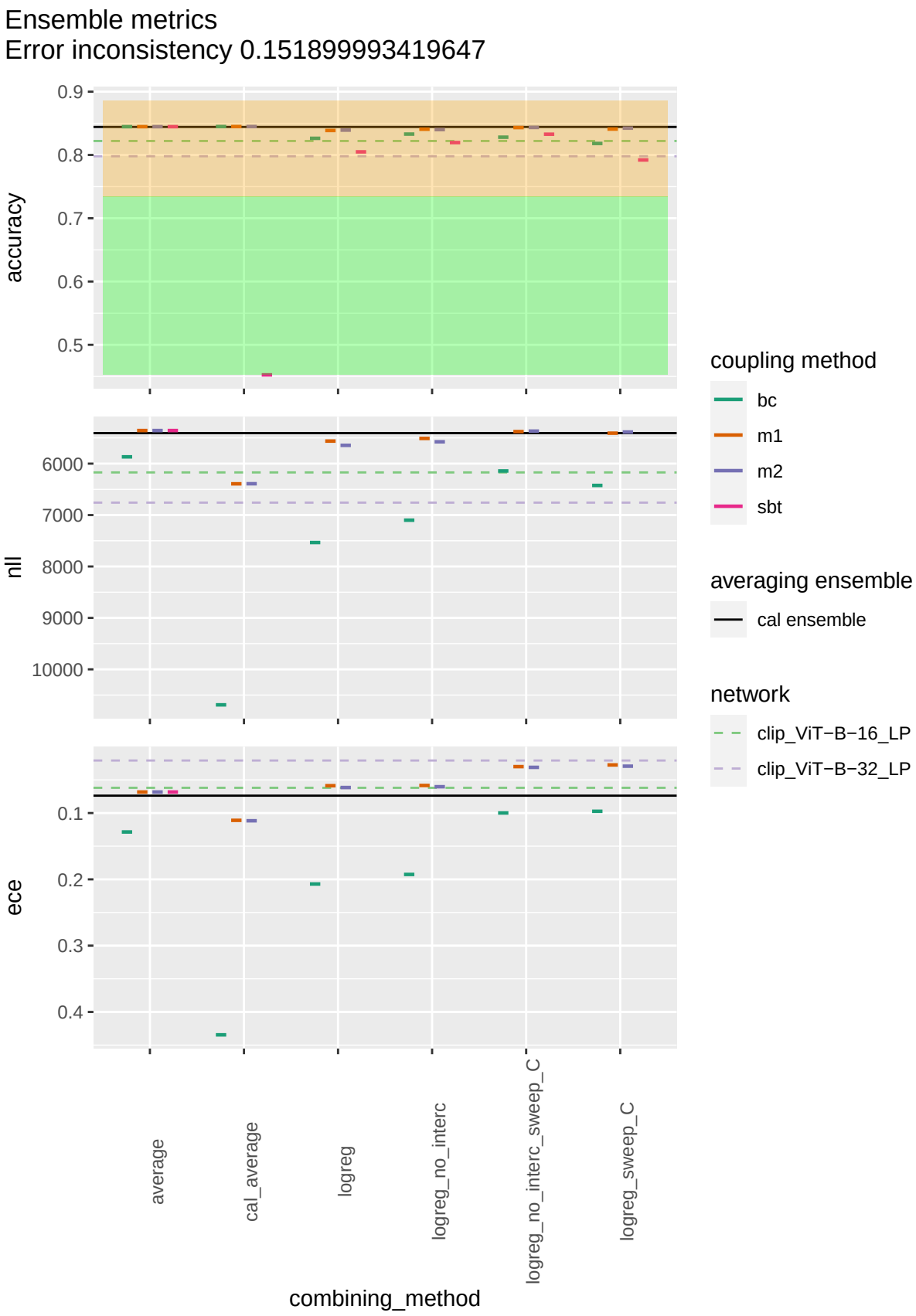

Ensemble metrics
Error inconsistency 0.206299990415573
accuracy
0.90
0.85
0.80
nll
5000
6000
7000
8000
9000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16_LP
resnext101

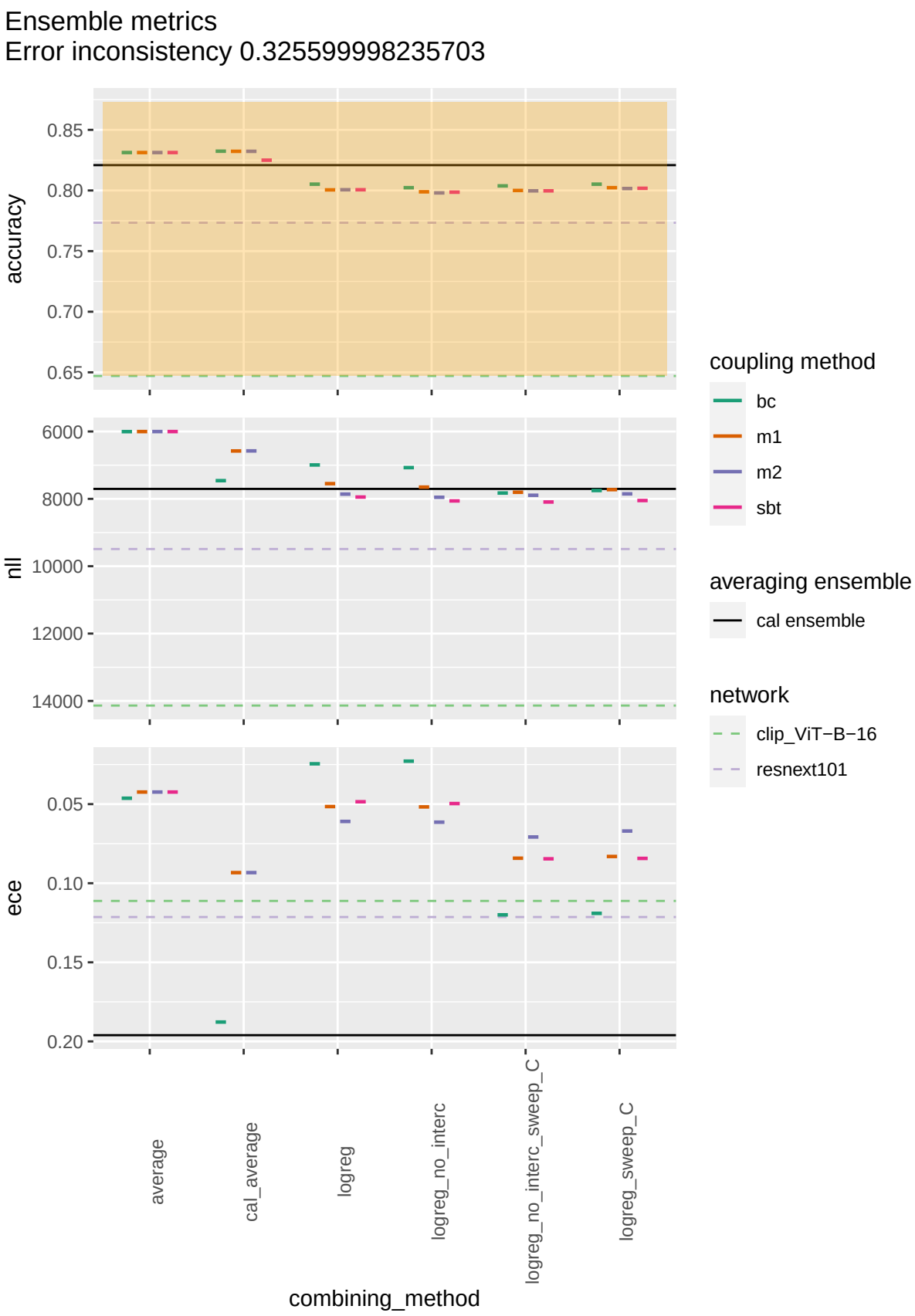

Ensemble metrics
Error inconsistency 0.228499993681908
accuracy
0.85
0.80
0.75
0.70
0.65
nll
6000
8000
10000
12000
14000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−16_LP

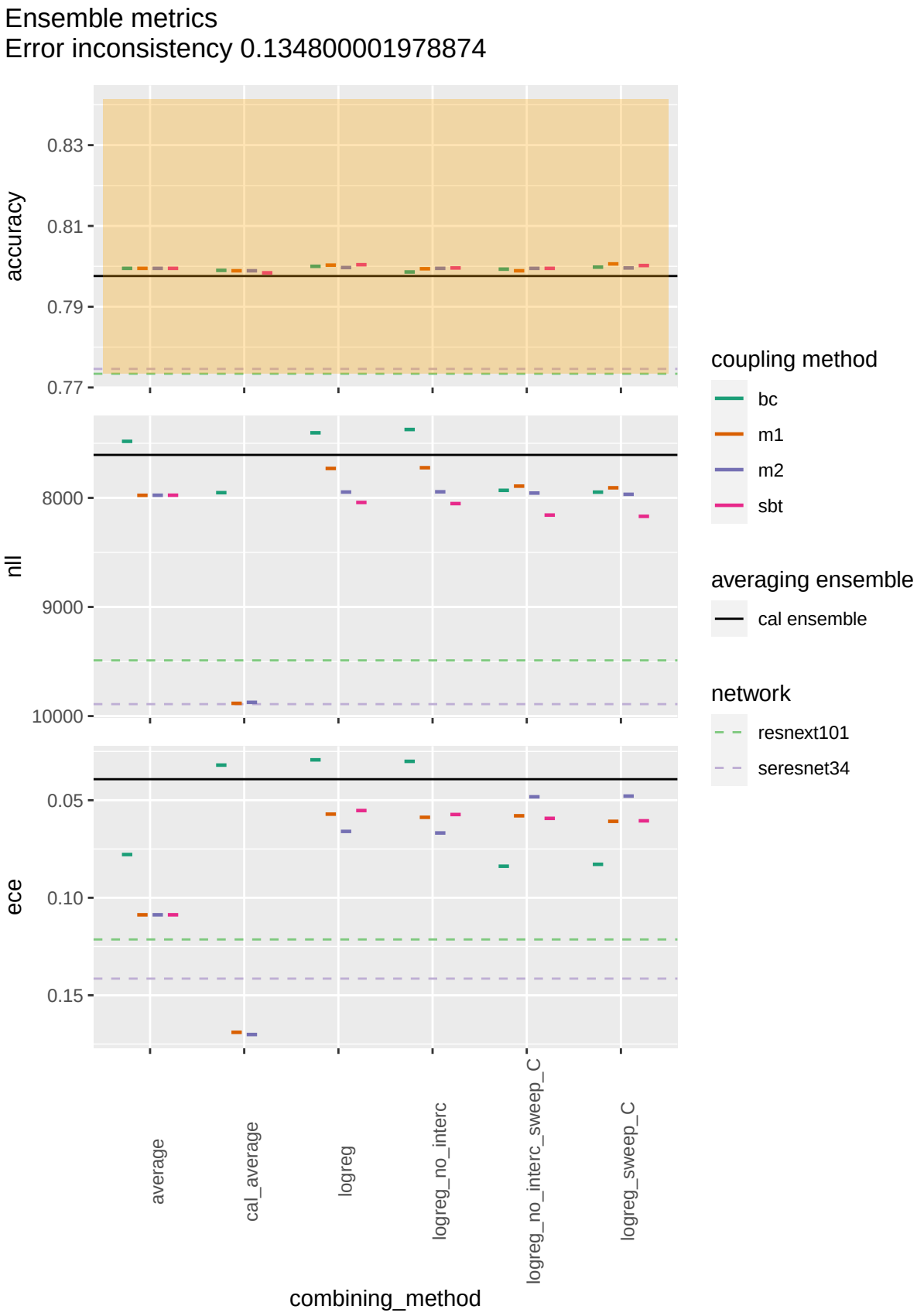

Ensemble metrics
Error inconsistency 0.214399993419647
accuracy
0.90
0.85
0.80
nll
5000
6000
7000
8000
9000
10000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16_LP
googlenet

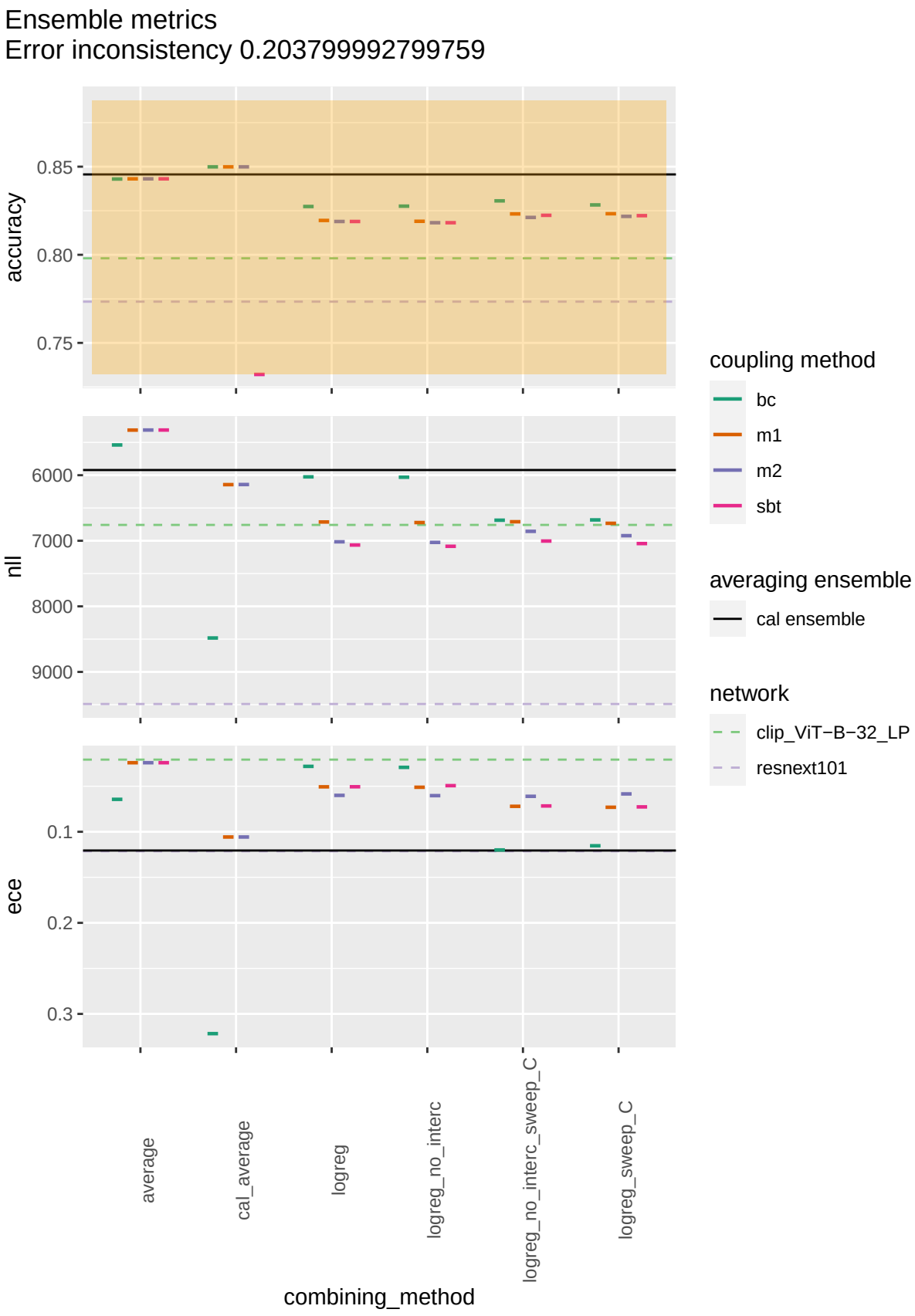
Ensemble metrics
Error inconsistency 0.212099999189377



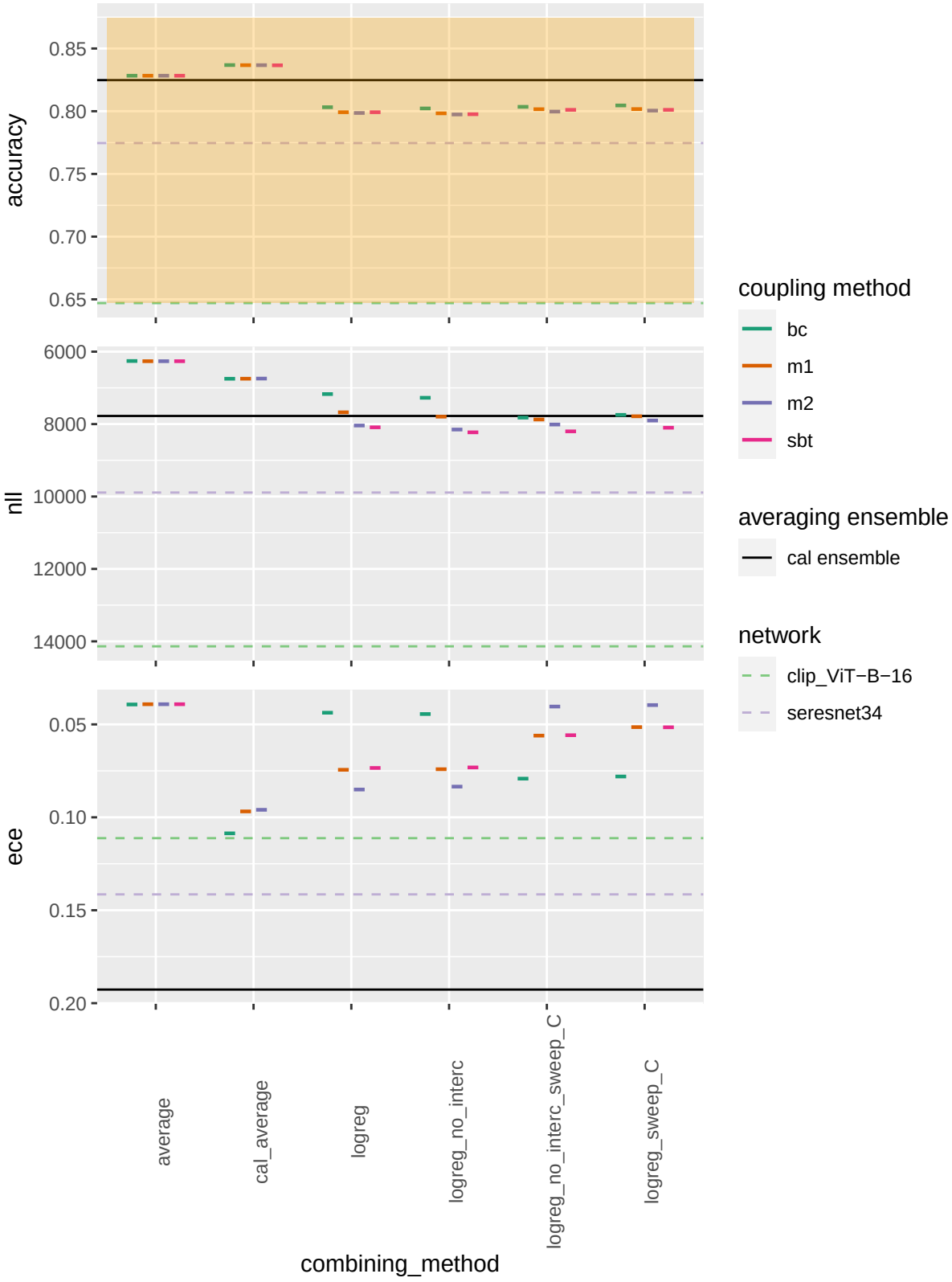

Ensemble metrics
Error inconsistency 0.151899993419647
accuracy
nll
ece
0.9
0.8
0.7
0.6
0.5
6000
7000
8000
9000
10000
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16_LP
clip_ViT−B−32_LP

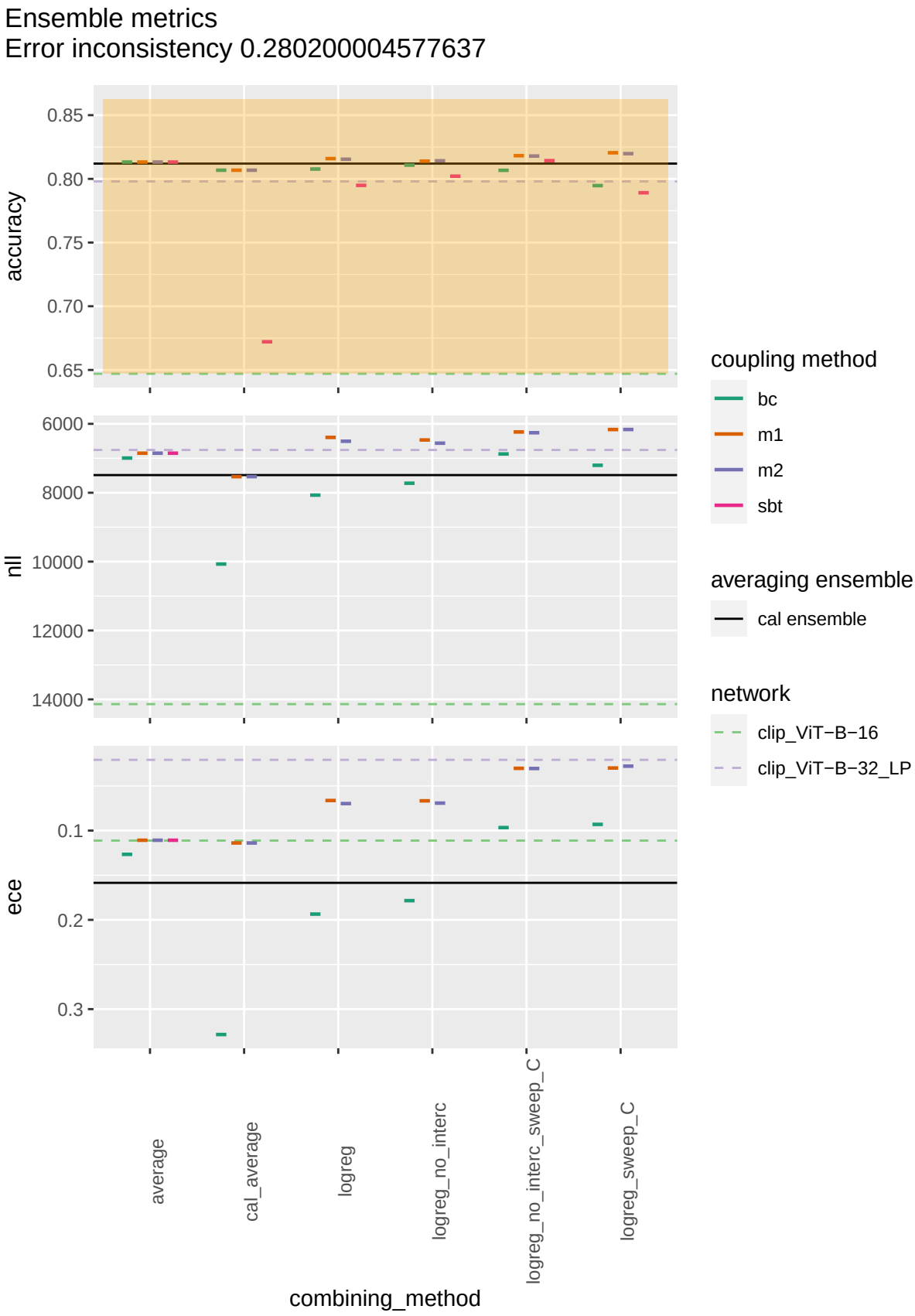

Ensemble metrics
Error inconsistency 0.325599998235703
accuracy
0.85
0.80
0.75
0.70
0.65
nll
6000
8000
10000
12000
14000
ece
0.05
0.10
0.15
0.20
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
resnext101


Ensemble metrics
Error inconsistency 0.134800001978874
accuracy
0.83
0.81
0.79
0.77
nll
8000
9000
10000
ece
0.05
0.10
0.15
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
resnext101
seresnet34

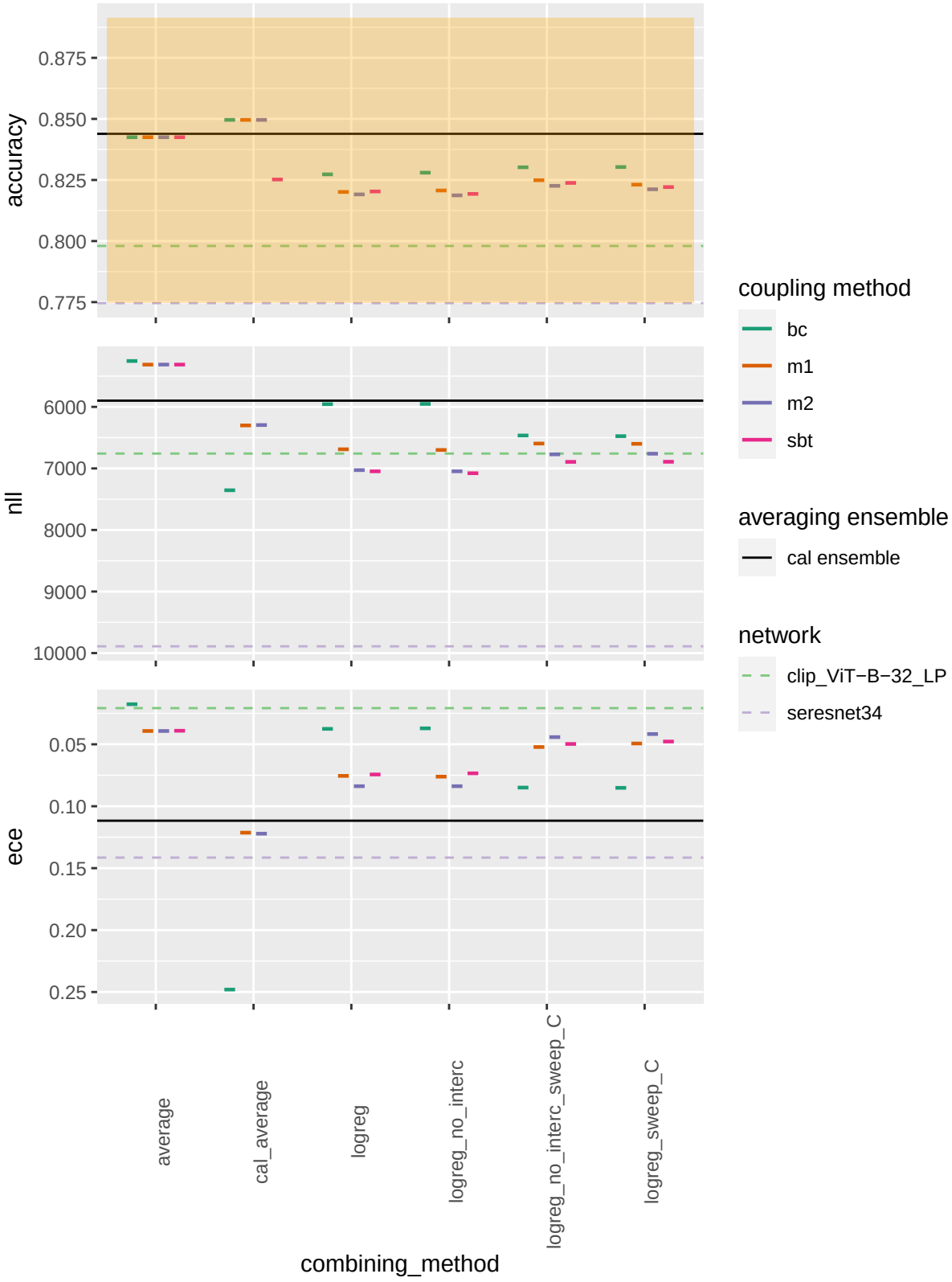

Ensemble metrics
Error inconsistency 0.203799992799759
accuracy
0.85
0.80
0.75
nll
6000
7000
8000
9000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32_LP
resnext101

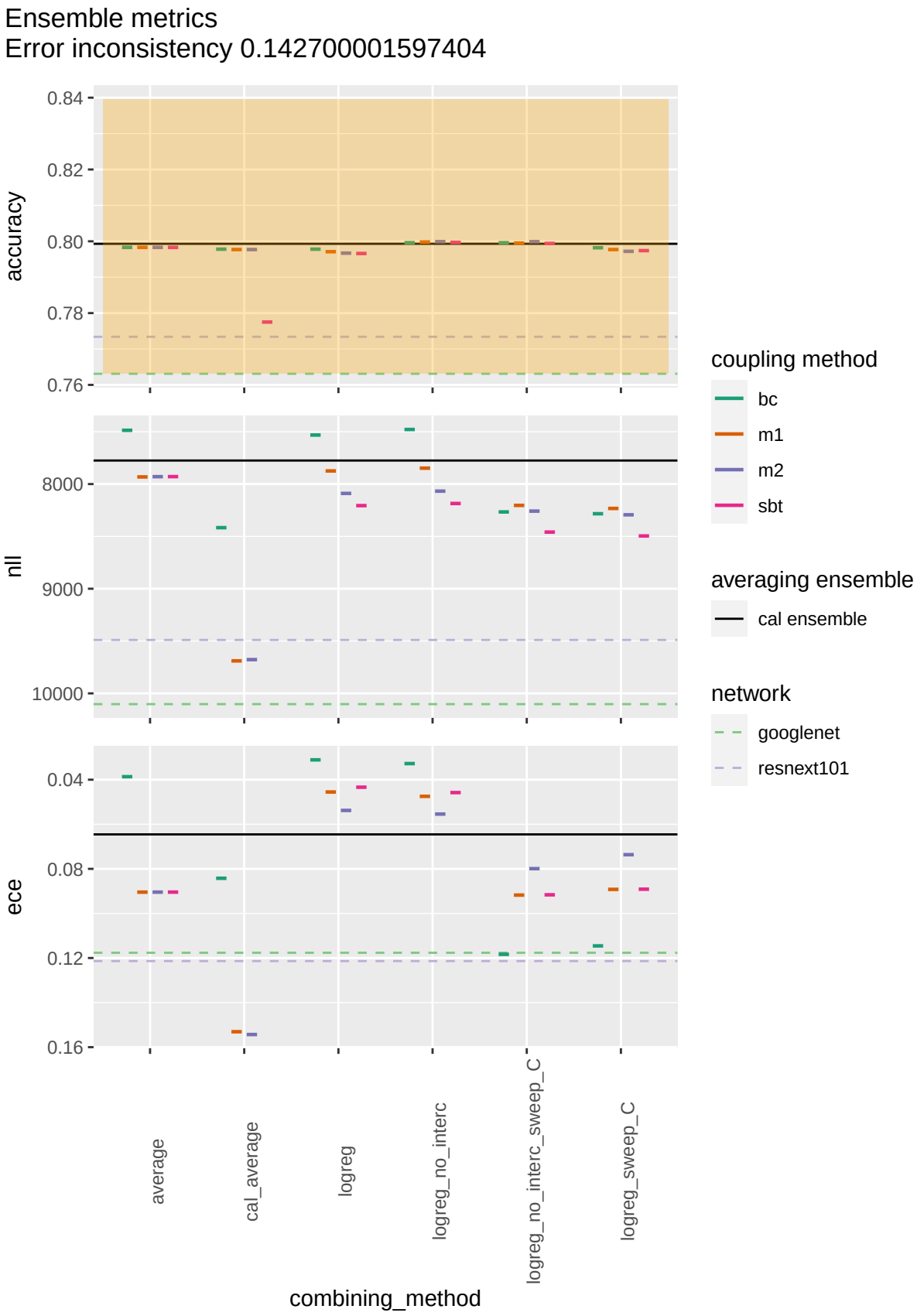
Ensemble metrics
Error inconsistency 0.327399998903275

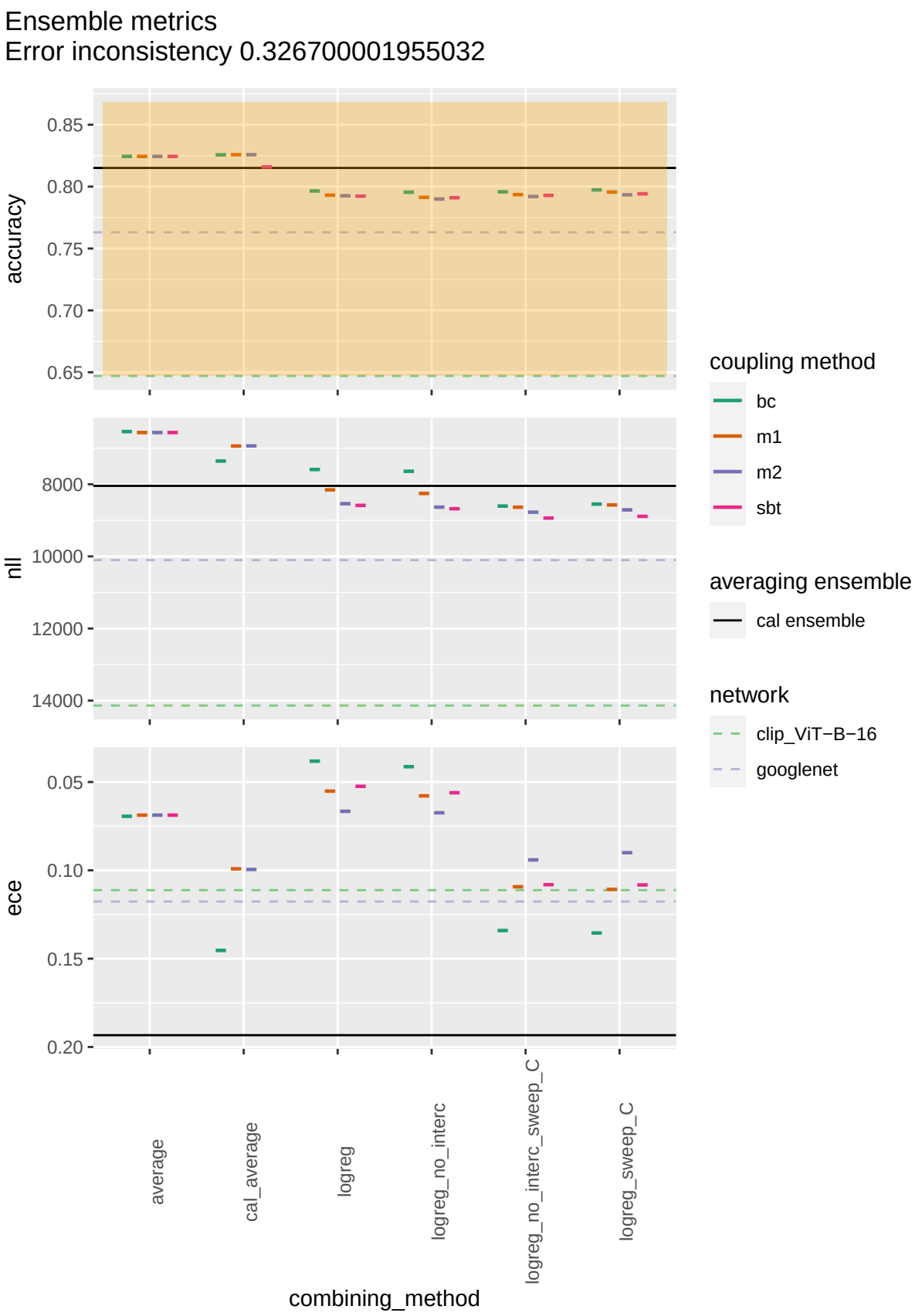


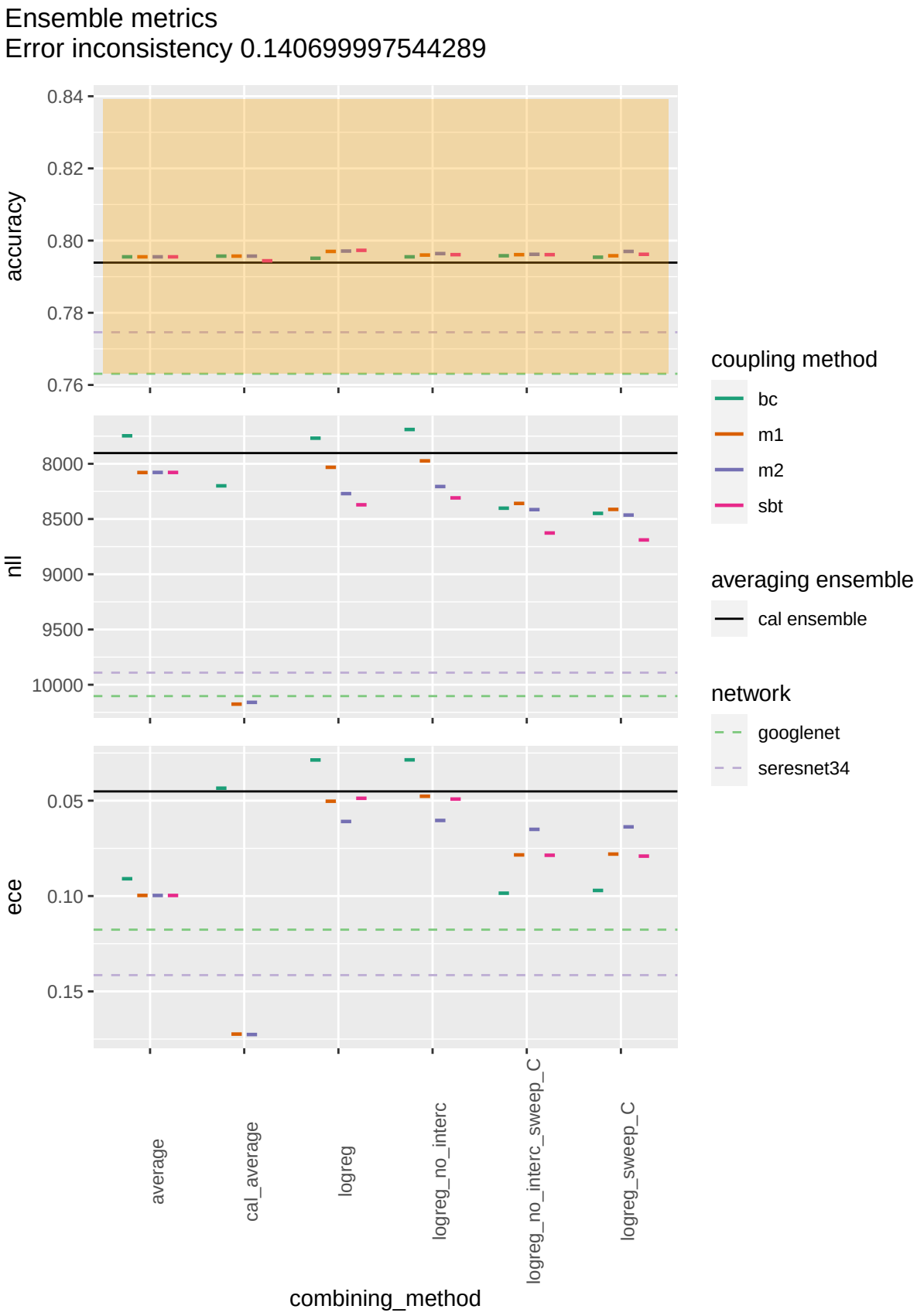

Ensemble metrics
Error inconsistency 0.280200004577637
accuracy
0.85
0.80
0.75
0.70
0.65
nll
6000
8000
10000
12000
14000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−32_LP

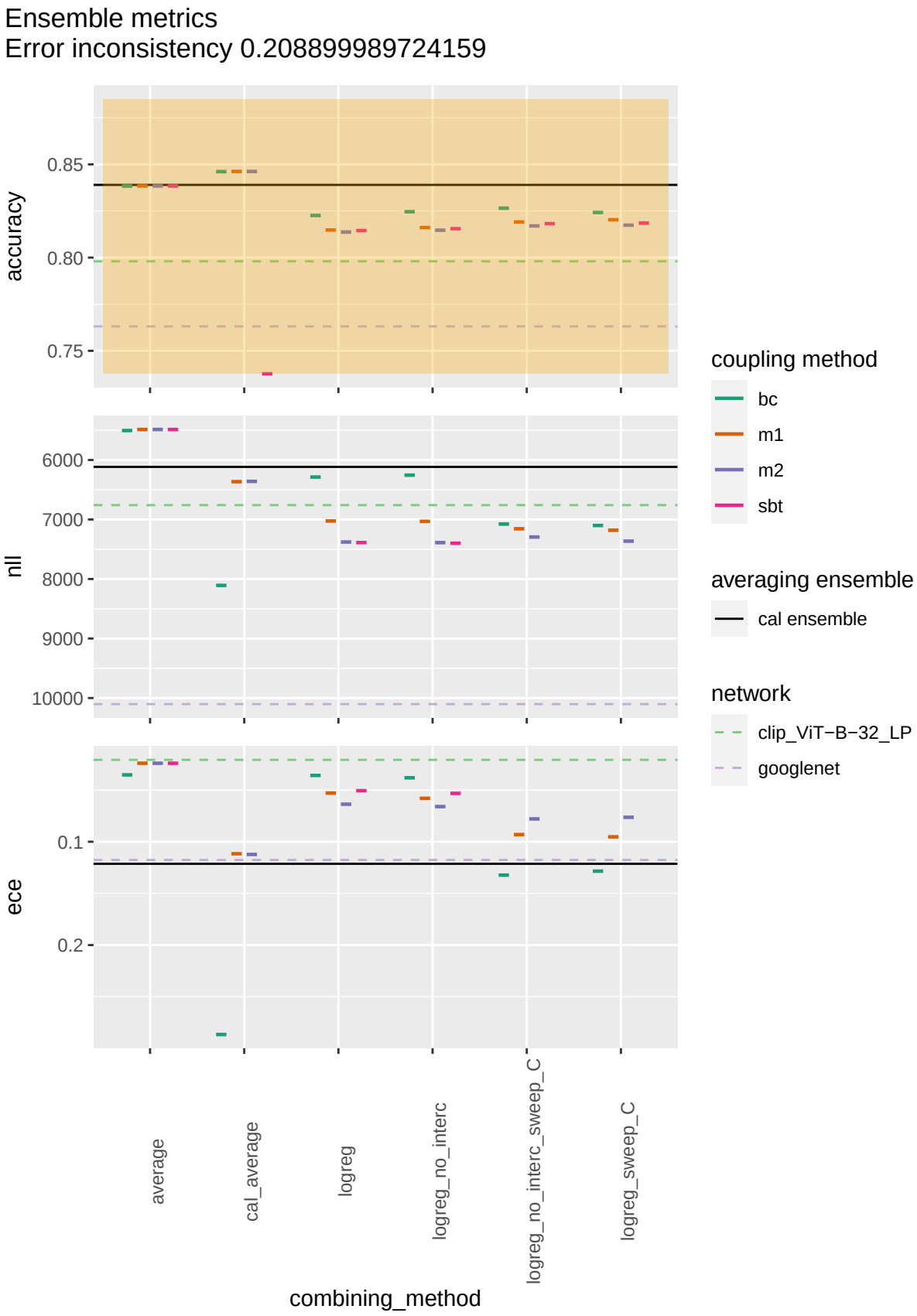
Ensemble metrics
Error inconsistency 0.210199996829033

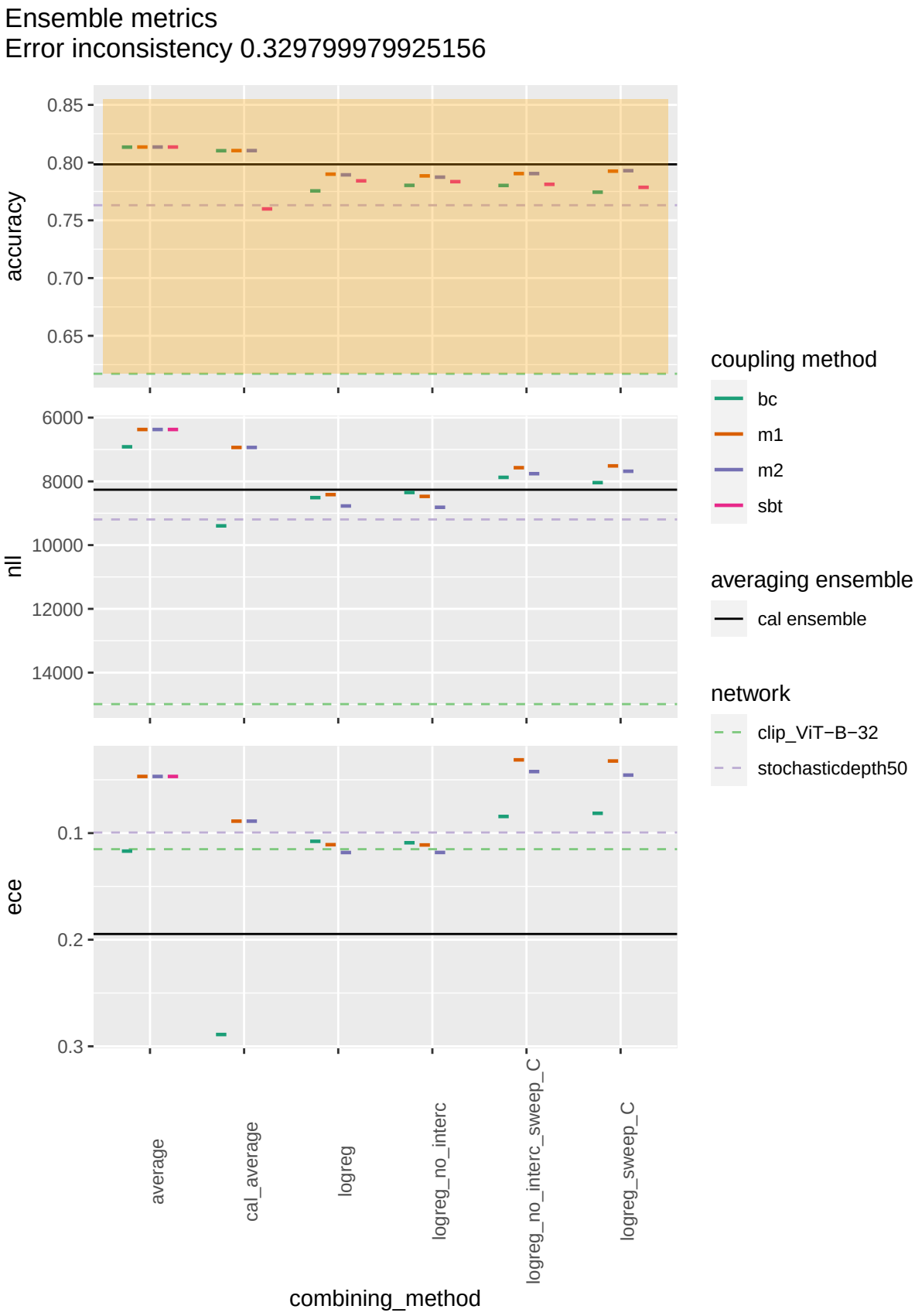


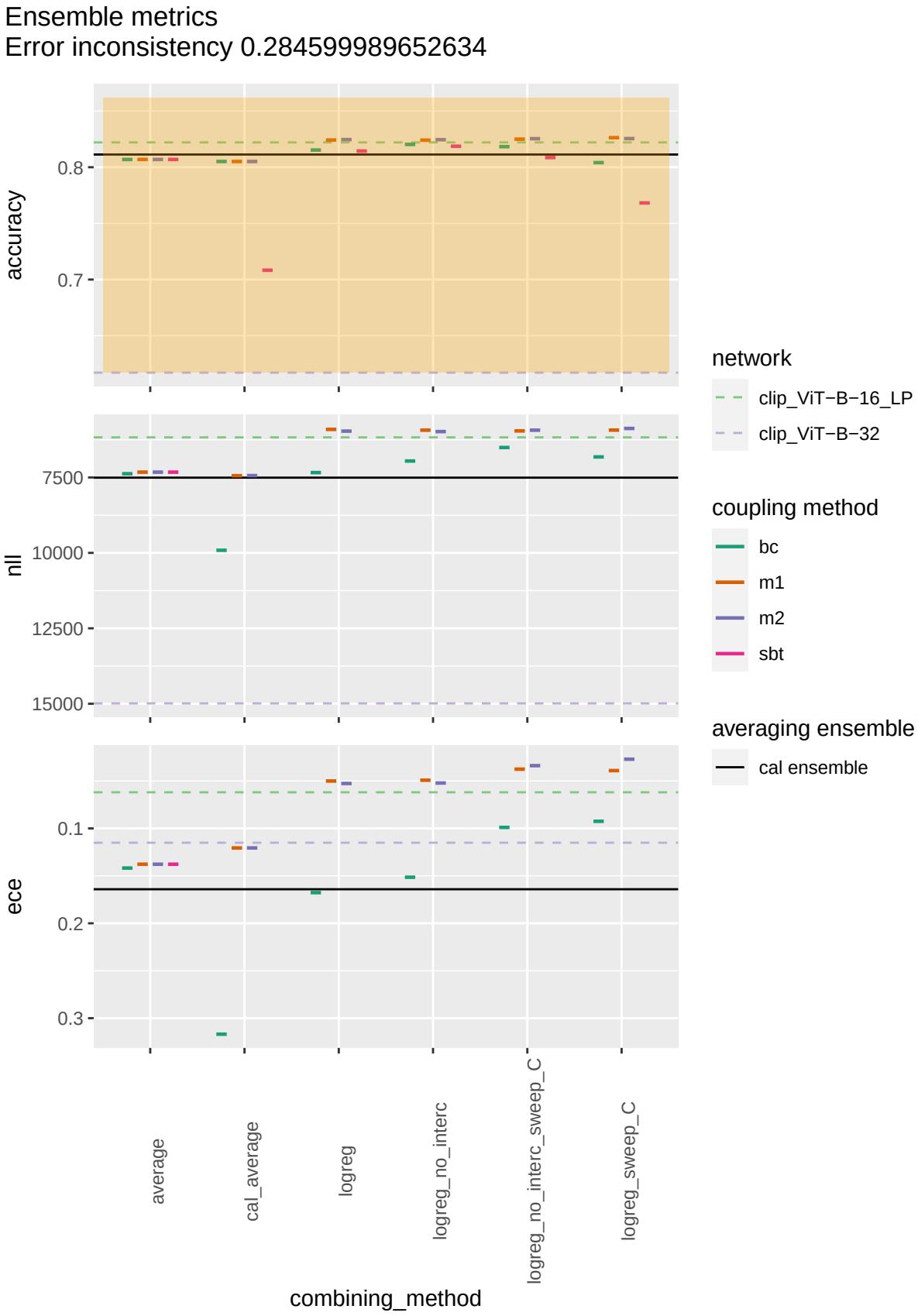

Ensemble metrics
Error inconsistency 0.142700001597404
accuracy
0.84
0.82
0.80
0.78
0.76
nll
8000
9000
10000
ece
0.04
0.08
0.12
0.16
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
googlenet
resnext101

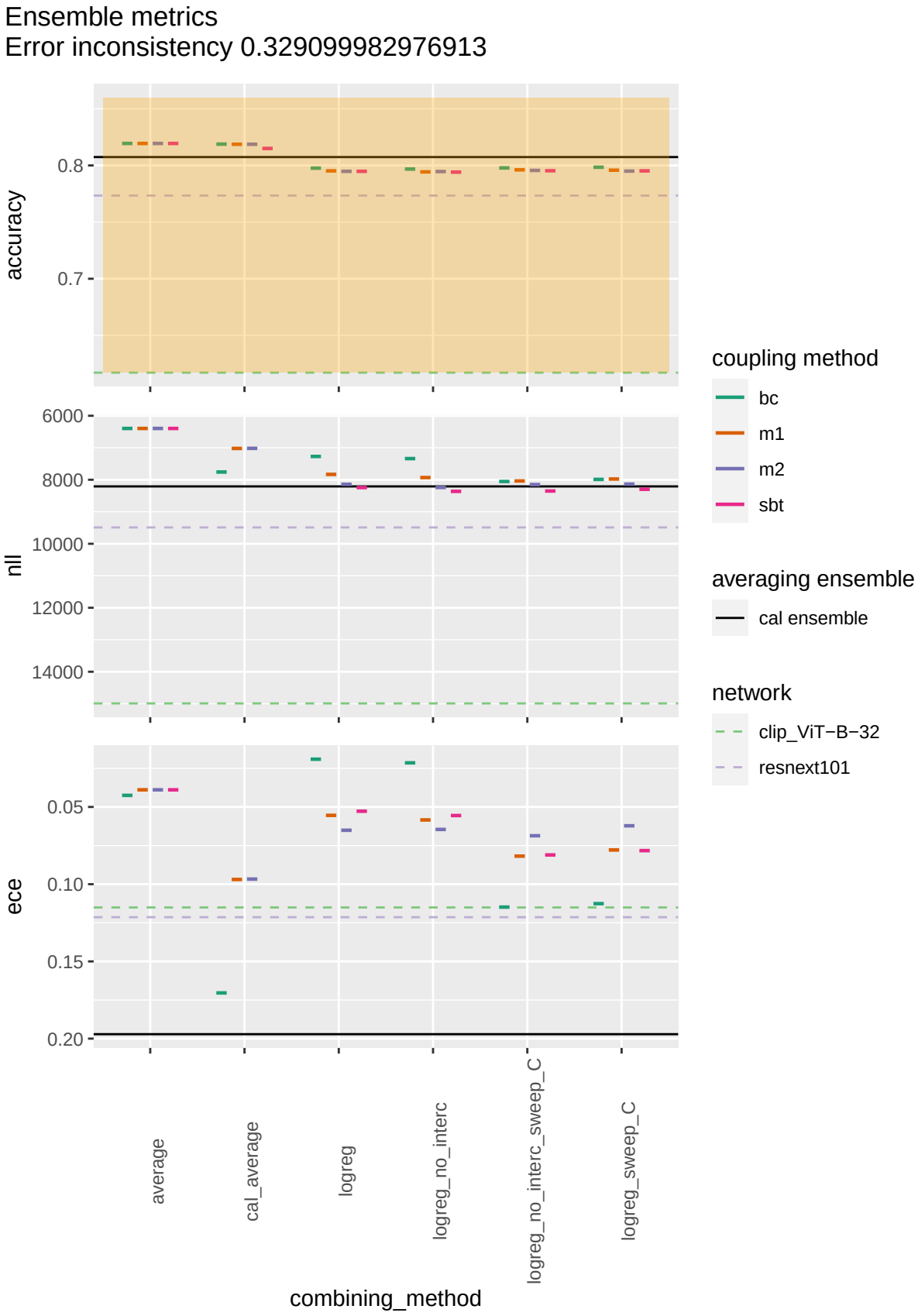

Ensemble metrics
Error inconsistency 0.326700001955032
accuracy
0.85
0.80
0.75
0.70
0.65
nll
8000
10000
12000
14000
ece
0.05
0.10
0.15
0.20
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
googlenet


Ensemble metrics
Error inconsistency 0.140699997544289
accuracy
0.84
0.82
0.80
0.78
0.76
nll
8000
8500
9000
9500
10000
ece
0.05
0.10
0.15
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
googlenet
seresnet34

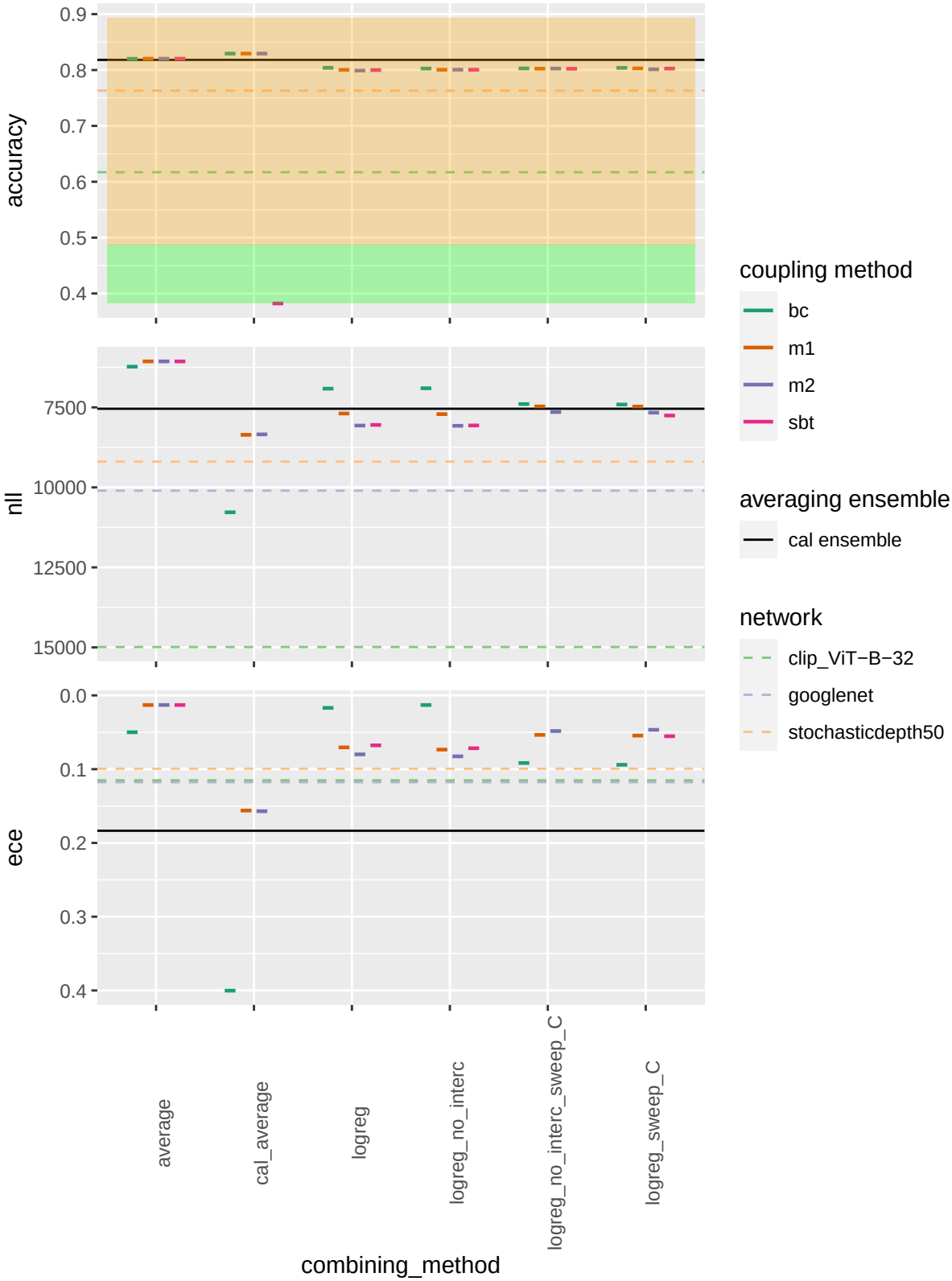

Ensemble metrics
Error inconsistency 0.208899989724159
accuracy
0.85
0.80
0.75
nll
6000
7000
8000
9000
10000
ece
0.1
0.2
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32_LP
googlenet


Ensemble metrics
Error inconsistency 0.329799979925156
accuracy
0.85
0.80
0.75
0.70
0.65
nll
6000
8000
10000
12000
14000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
stochasticdepth50

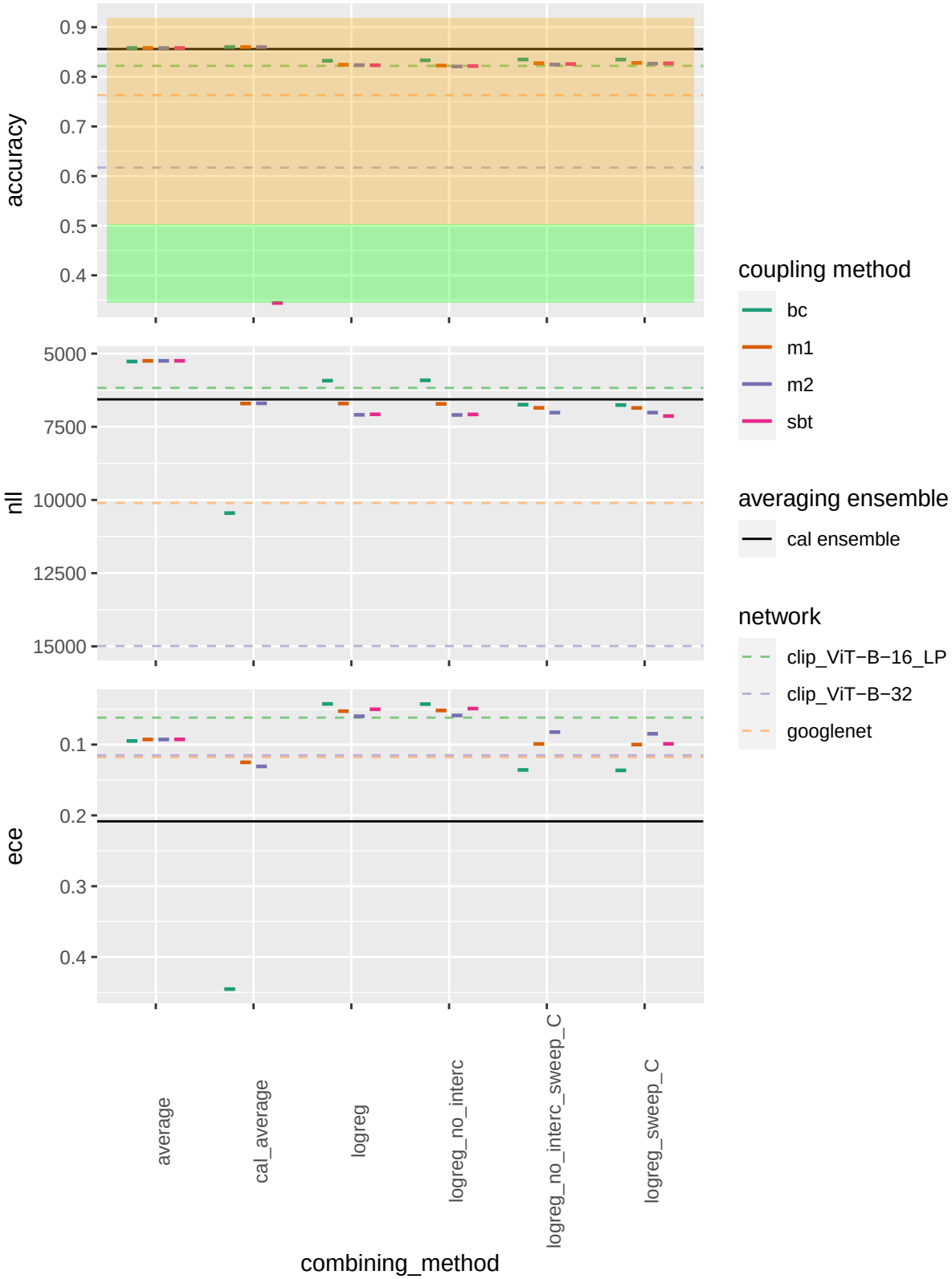

Ensemble metrics
Error inconsistency 0.284599989652634
accuracy
0.8
0.7
nll
7500
10000
12500
15000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
network
clip_ViT−B−16_LP
clip_ViT−B−32
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble

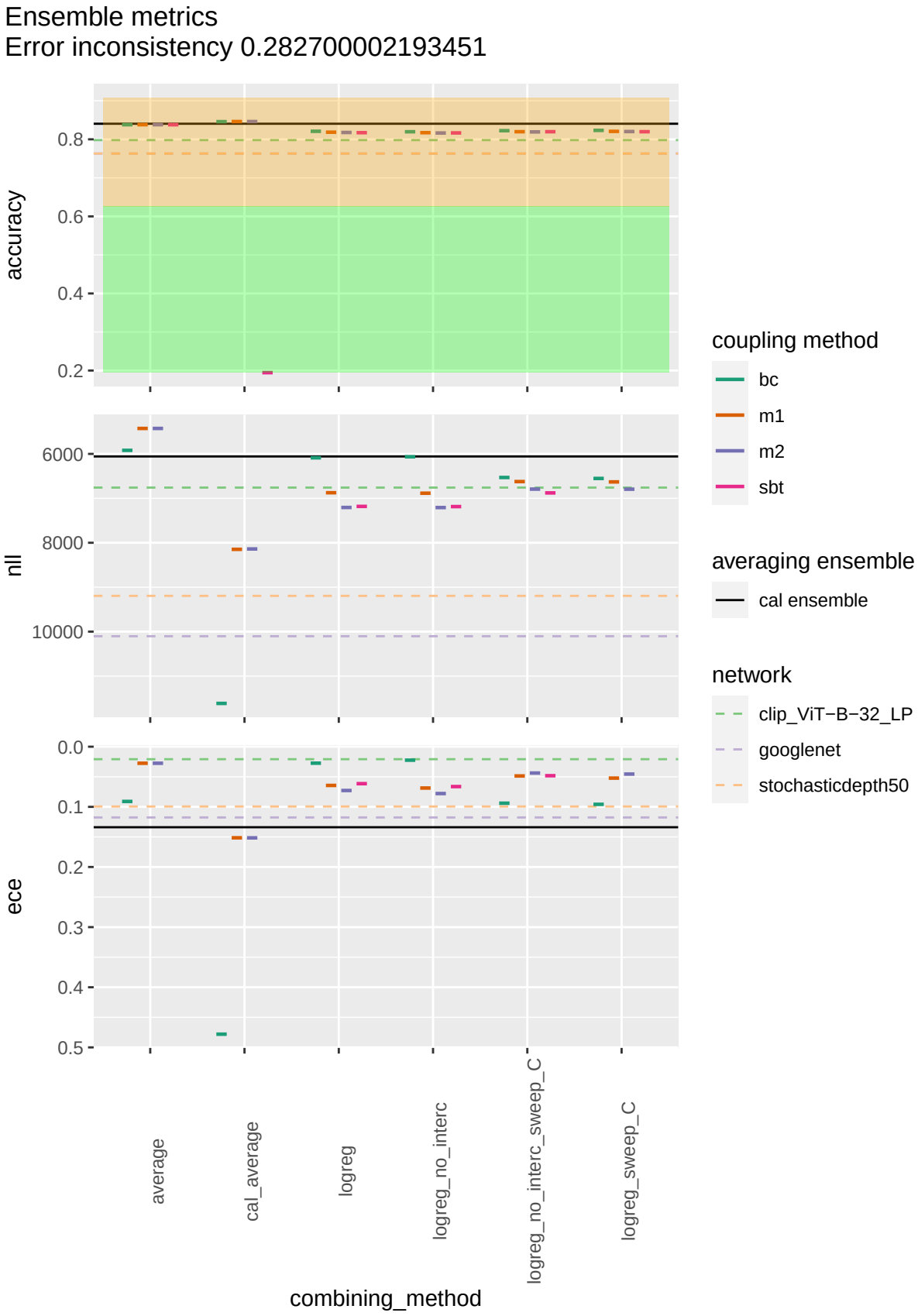

Ensemble metrics
Error inconsistency 0.329099982976913
accuracy
0.8
0.7
nll
6000
8000
10000
12000
14000
ece
0.05
0.10
0.15
0.20
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
resnext101

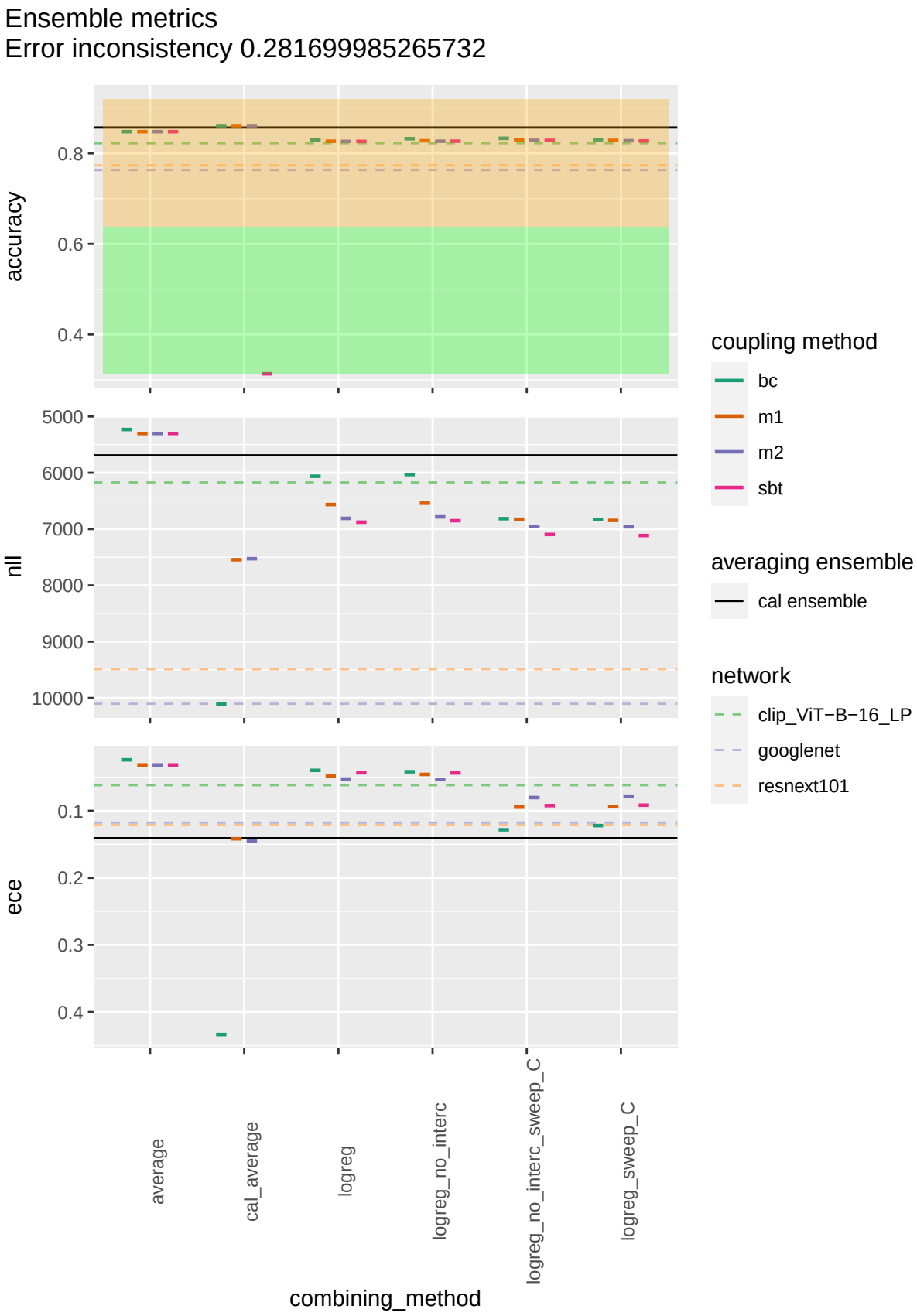
Ensemble metrics
Error inconsistency 0.406499981880188

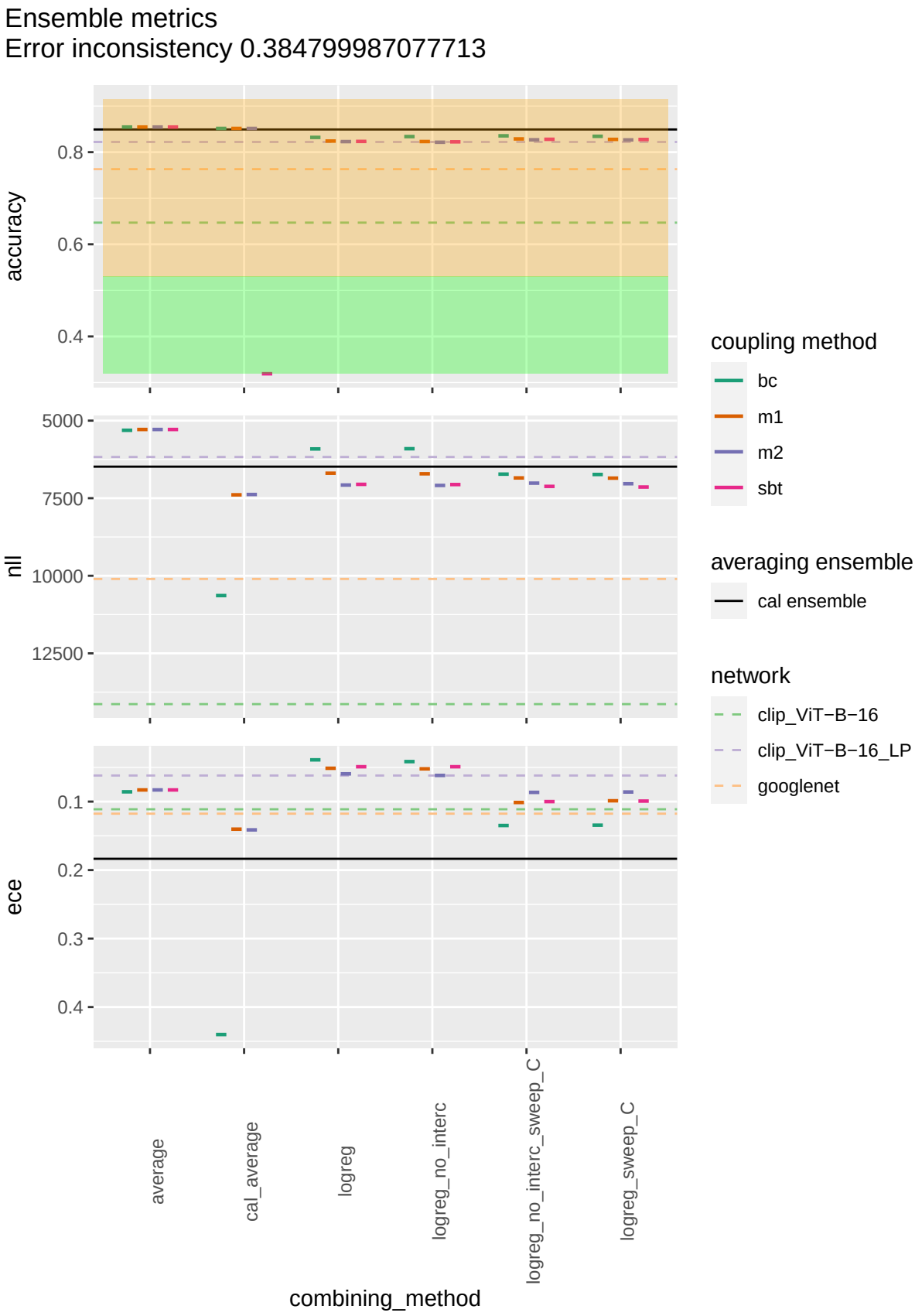


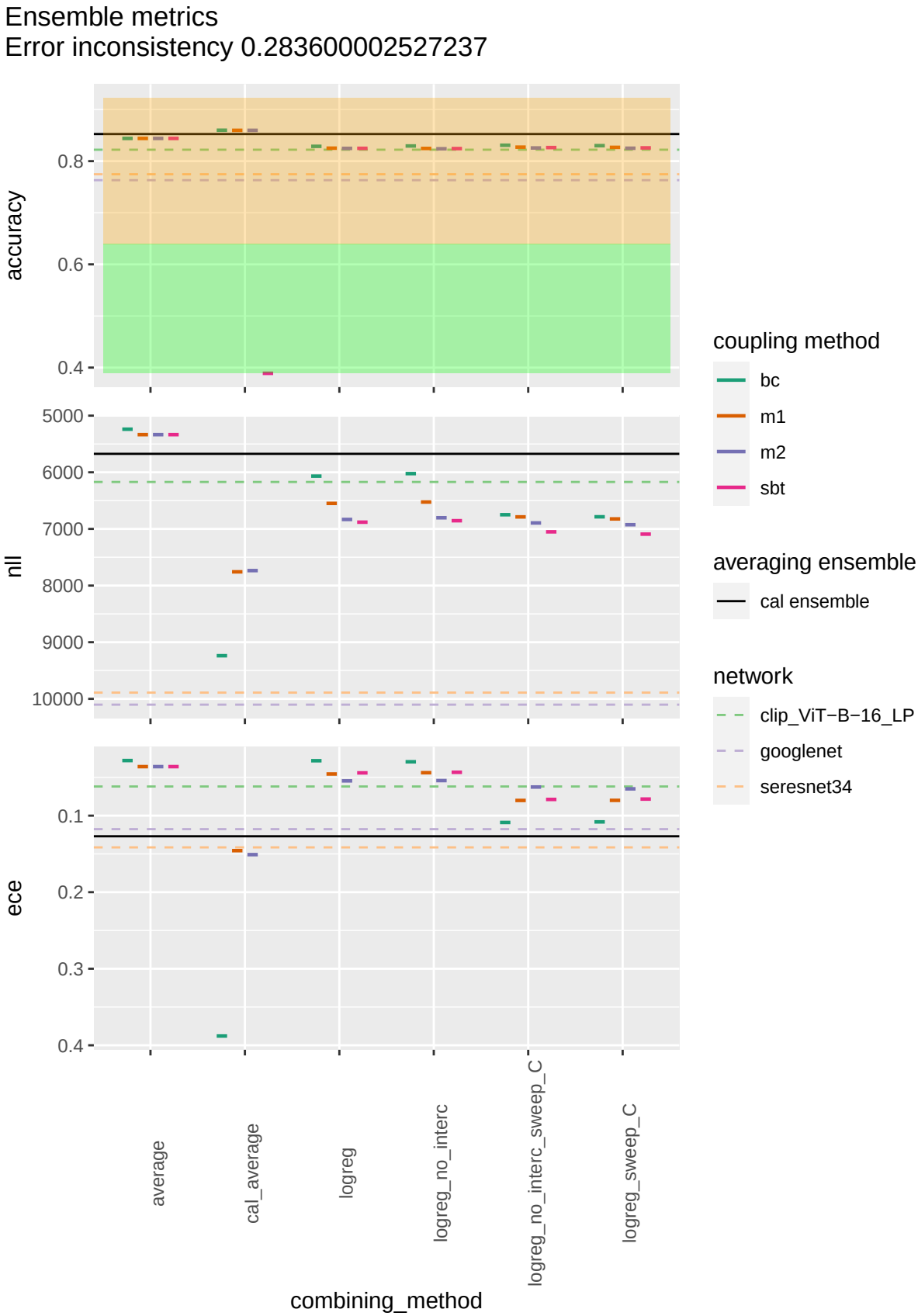
Ensemble metrics
Error inconsistency 0.416699975728989

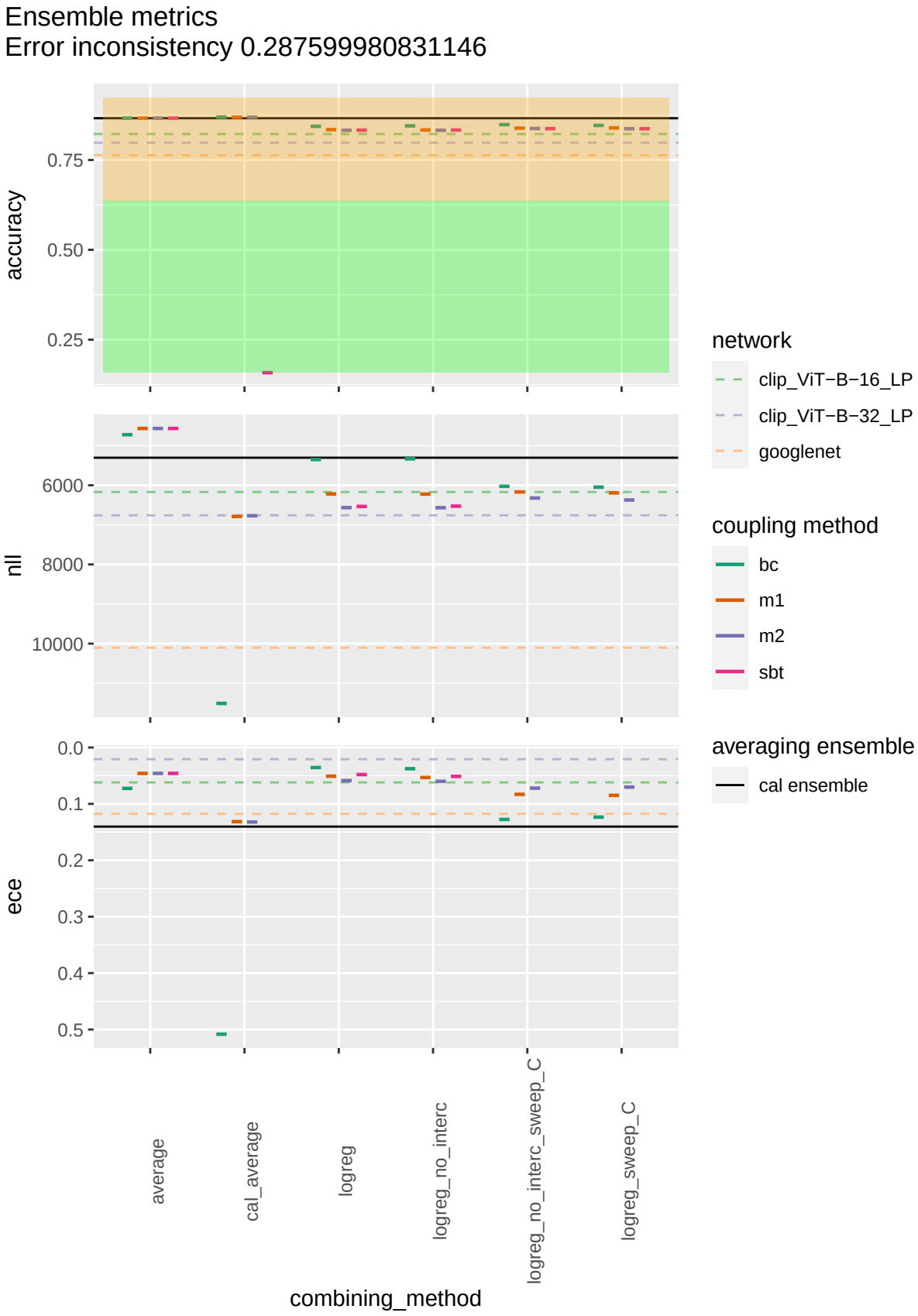


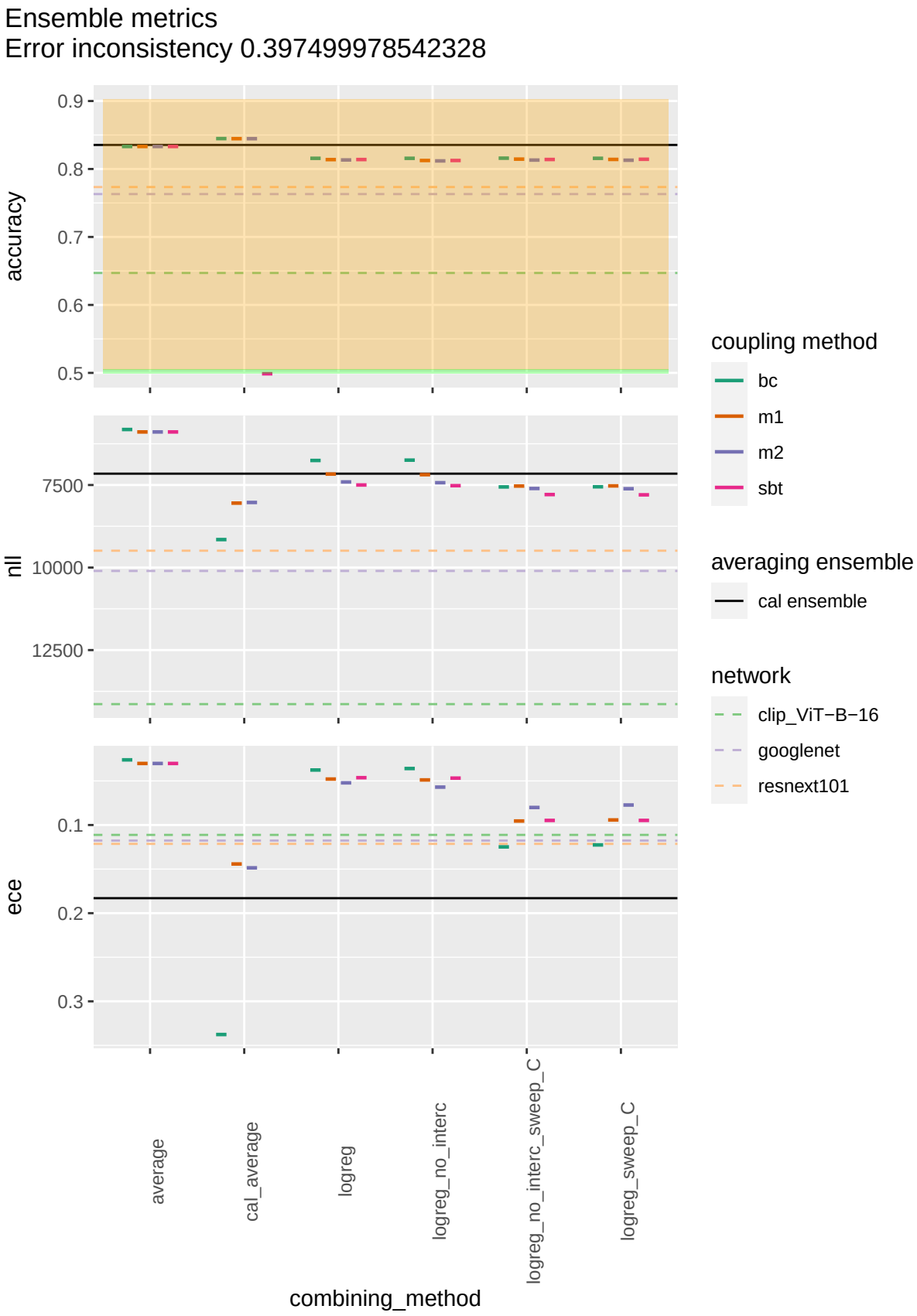

Ensemble metrics
Error inconsistency 0.282700002193451
accuracy
0.8
0.6
0.4
0.2
nll
6000
8000
10000
ece
0.0
0.1
0.2
0.3
0.4
0.5
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32_LP
googlenet
stochasticdepth50

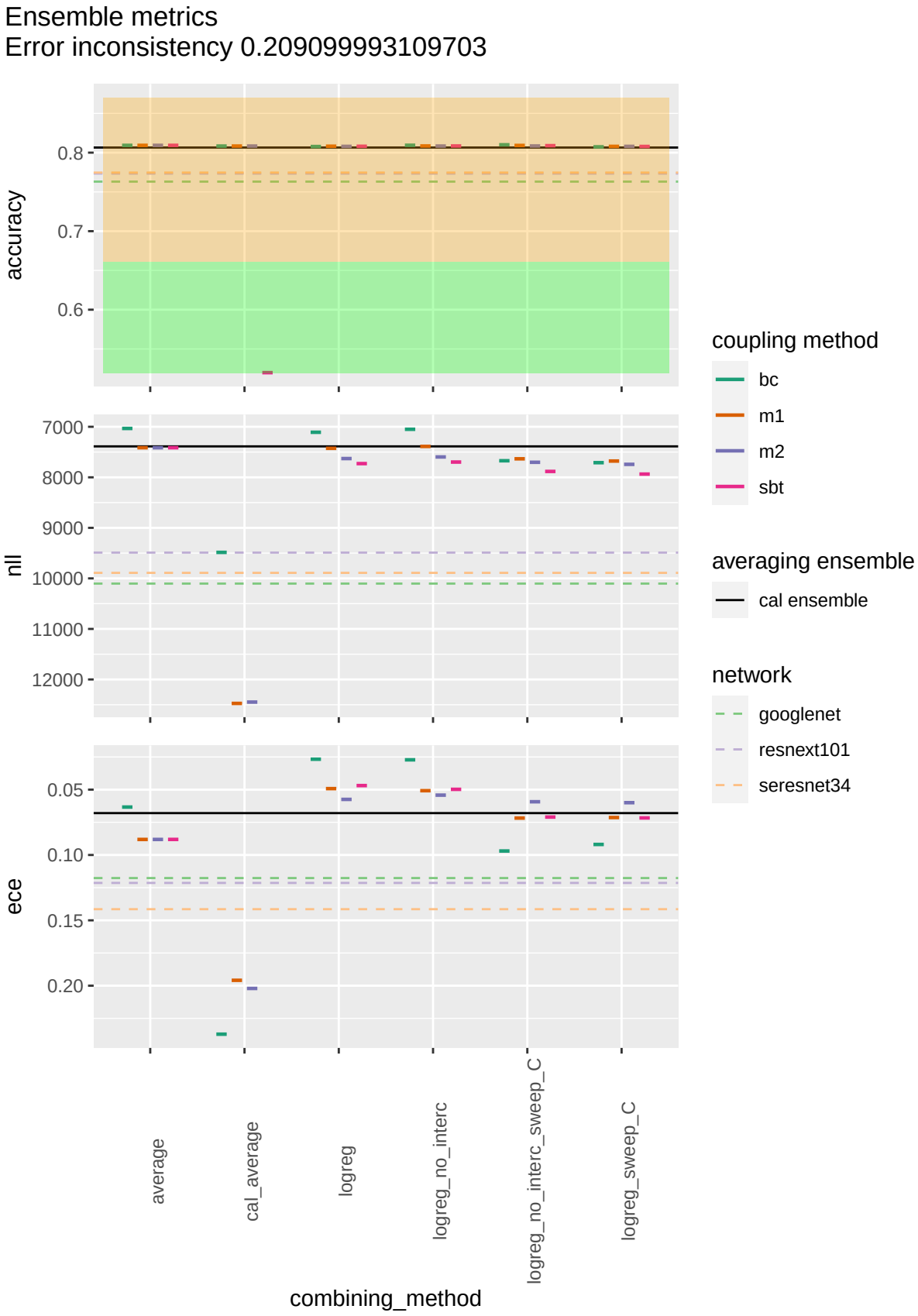

Ensemble metrics
Error inconsistency 0.281699985265732
accuracy
0.8
0.6
0.4
nll
5000
6000
7000
8000
9000
10000
ece
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16_LP
googlenet
resnext101


Ensemble metrics
Error inconsistency 0.384799987077713
accuracy
0.8
0.6
0.4
nll
5000
7500
10000
12500
ece
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−16_LP
googlenet

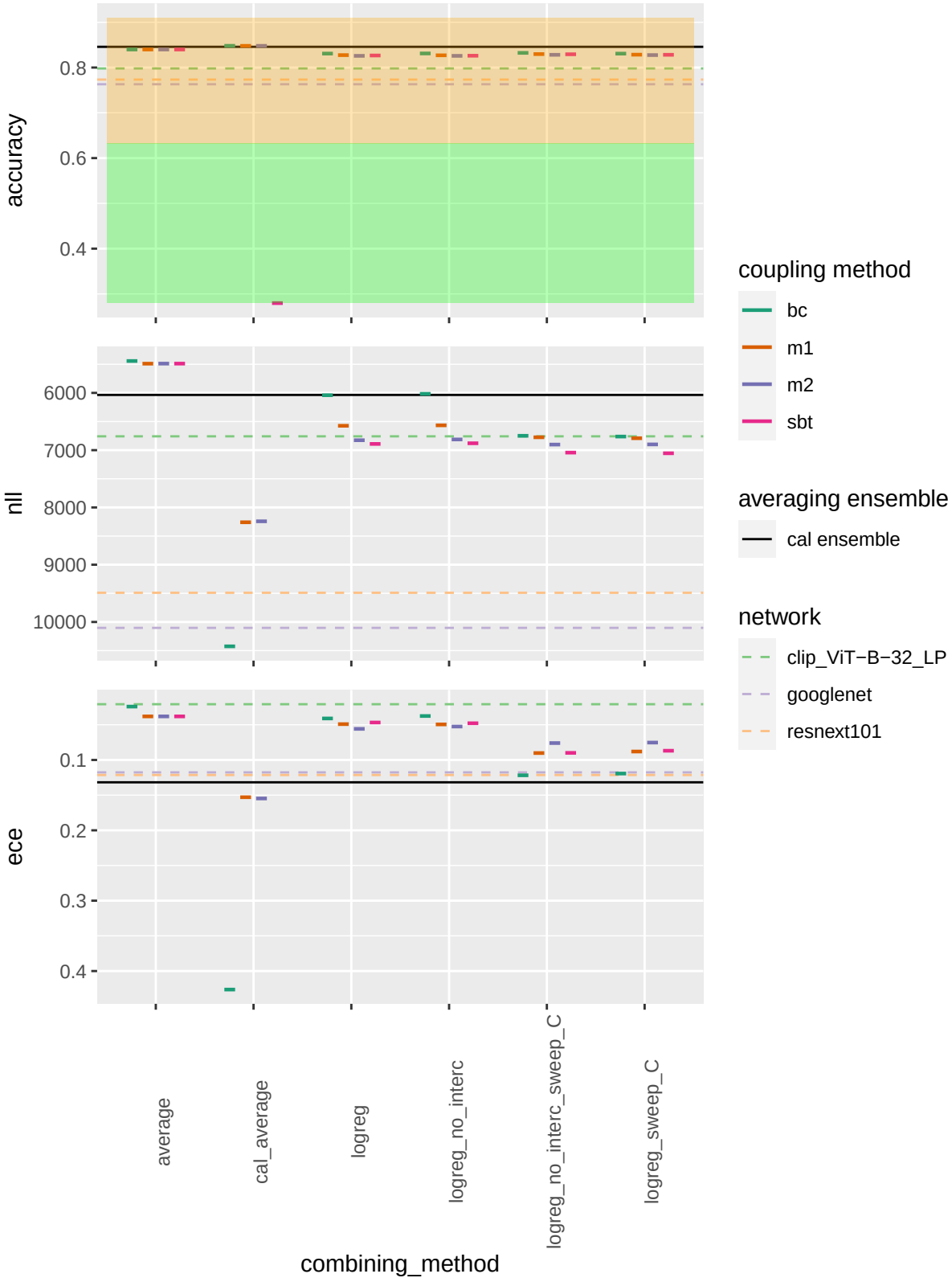

Ensemble metrics
Error inconsistency 0.283600002527237
accuracy
0.8
0.6
0.4
nll
5000
6000
7000
8000
9000
10000
ece
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16_LP
googlenet
seresnet34

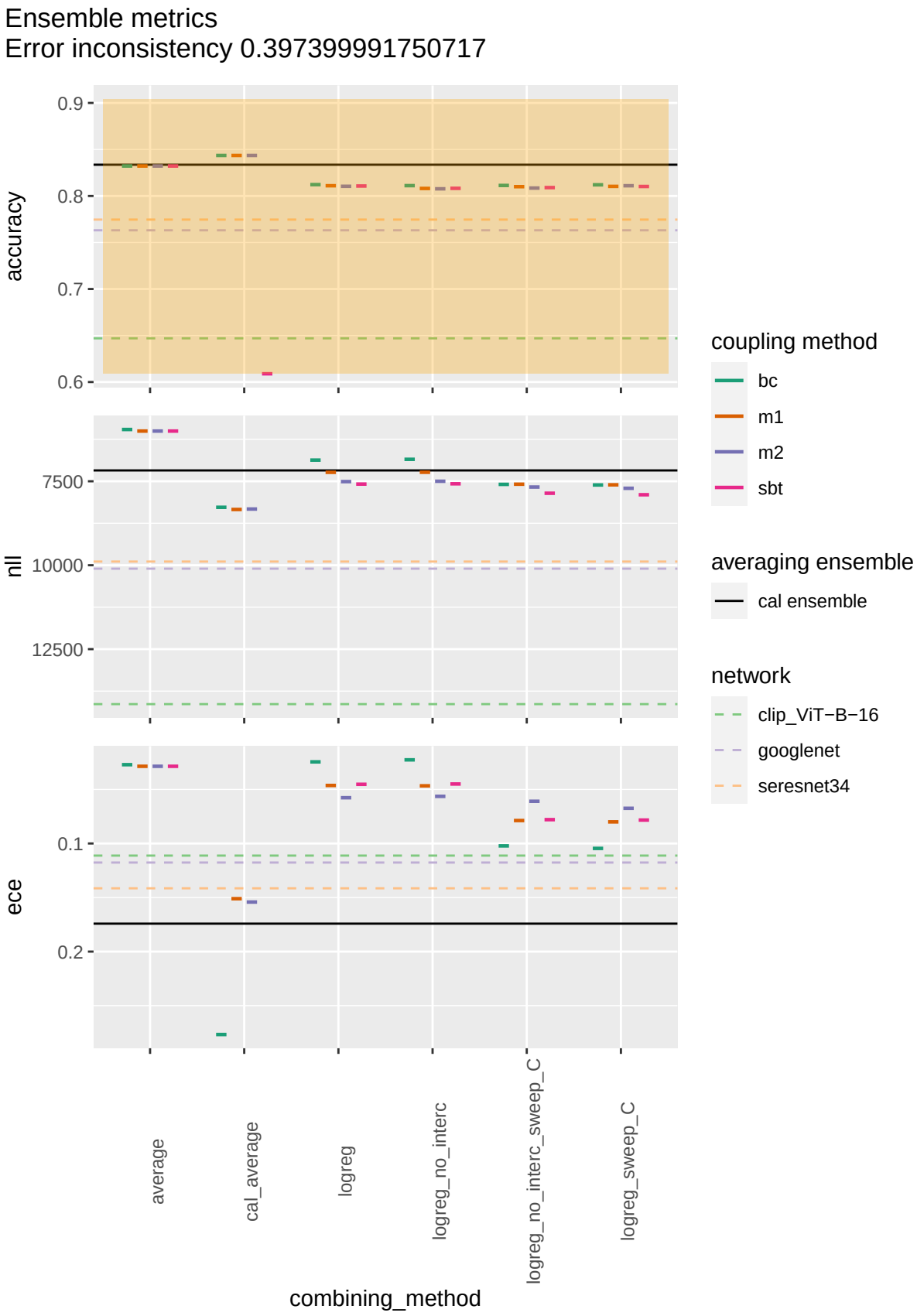

Ensemble metrics
Error inconsistency 0.287599980831146
accuracy
0.75
0.50
0.25
nll
6000
8000
10000
ece
0.0
0.1
0.2
0.3
0.4
0.5
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
network
clip_ViT−B−16_LP
clip_ViT−B−32_LP
googlenet
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble

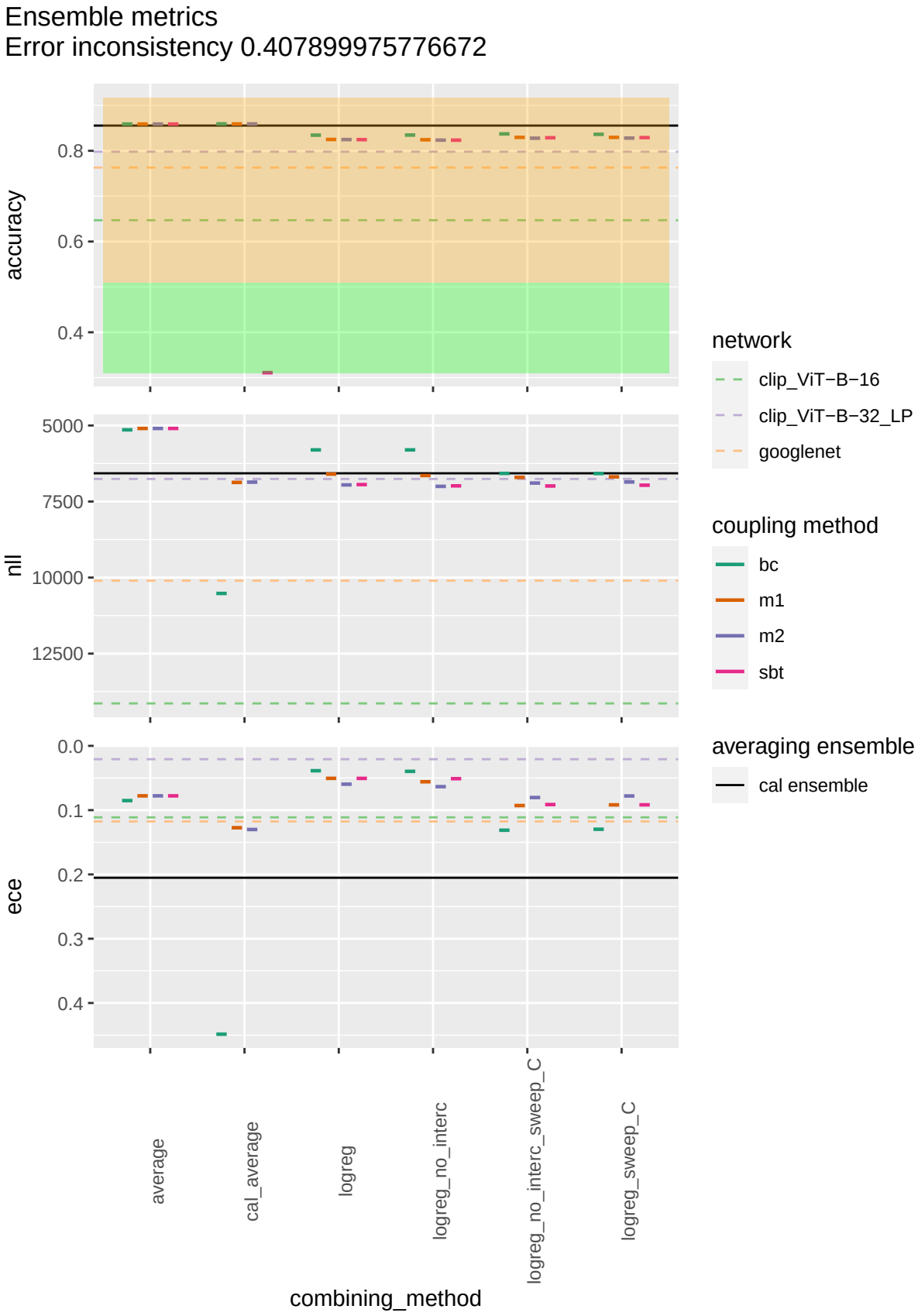

Ensemble metrics
Error inconsistency 0.397499978542328
accuracy
0.9
0.8
0.7
0.6
0.5
nll
7500
10000
12500
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
googlenet
resnext101


Ensemble metrics
Error inconsistency 0.209099993109703
accuracy
0.8
0.7
0.6
nll
7000
8000
9000
10000
11000
12000
ece
0.05
0.10
0.15
0.20
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
googlenet
resnext101
seresnet34

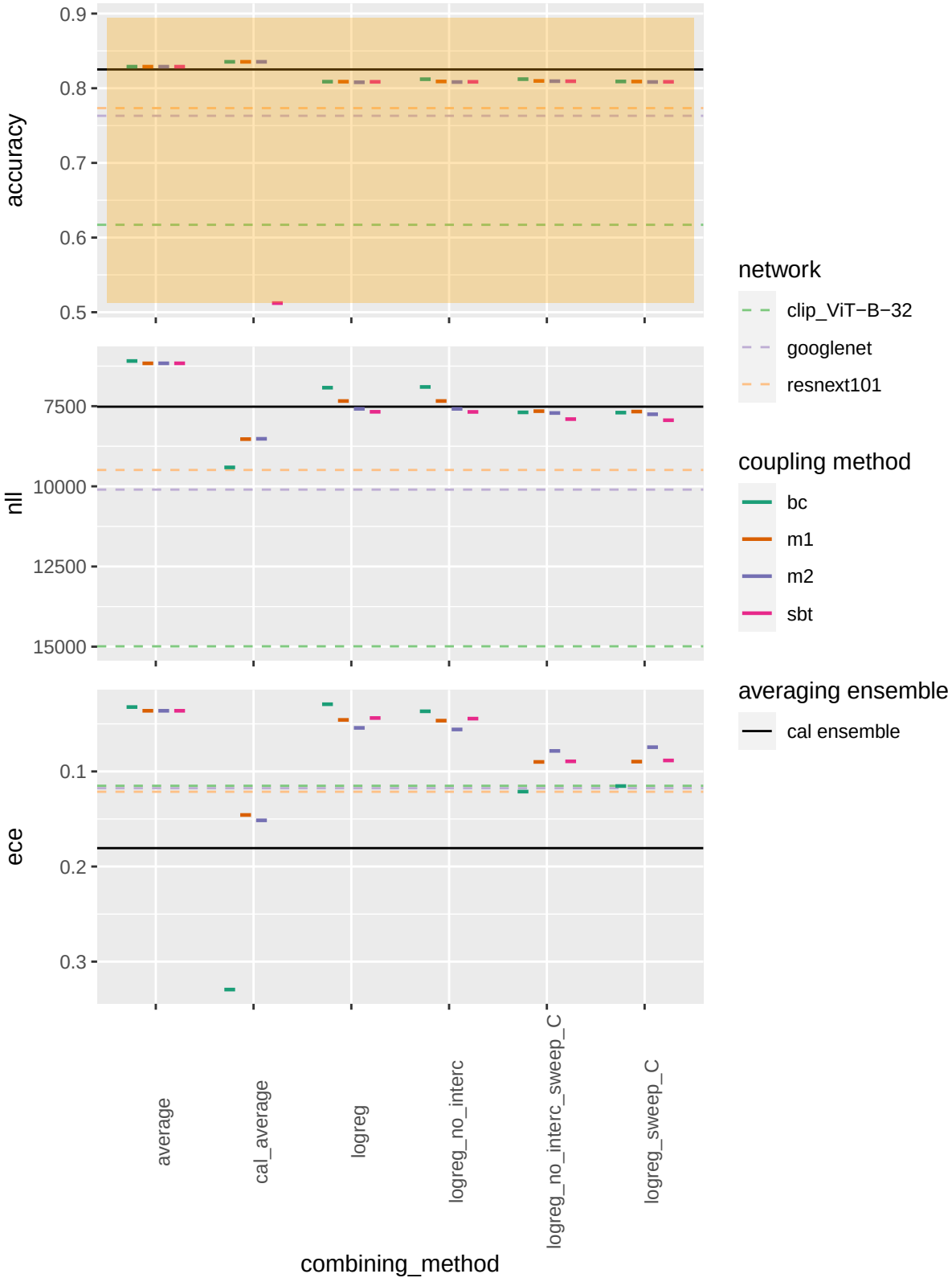
Ensemble metrics
Error inconsistency 0.277700006961823



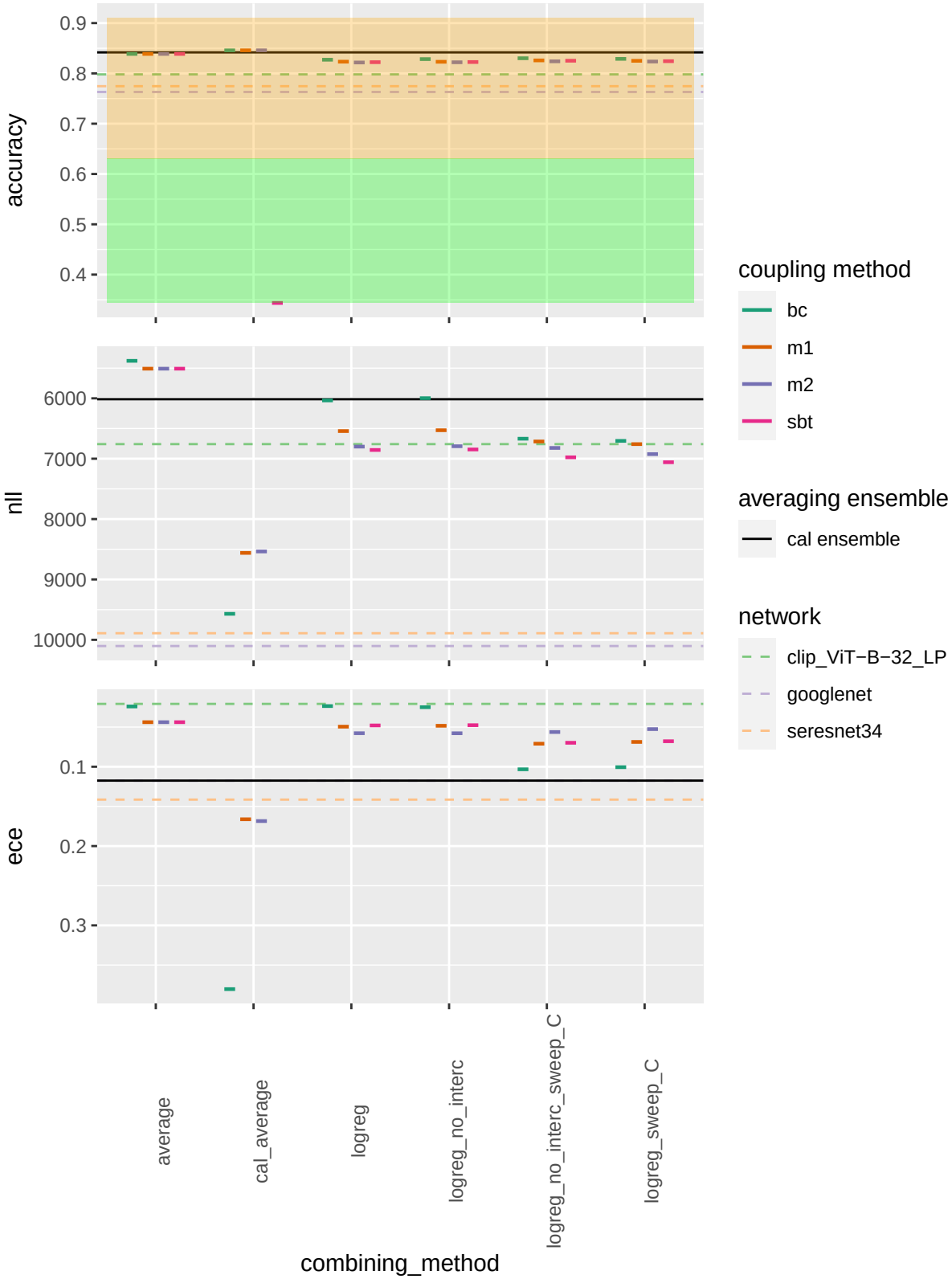

Ensemble metrics
Error inconsistency 0.397399991750717
accuracy
0.9
0.8
0.7
0.6
nll
7500
10000
12500
ece
0.1
0.2
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
googlenet
seresnet34

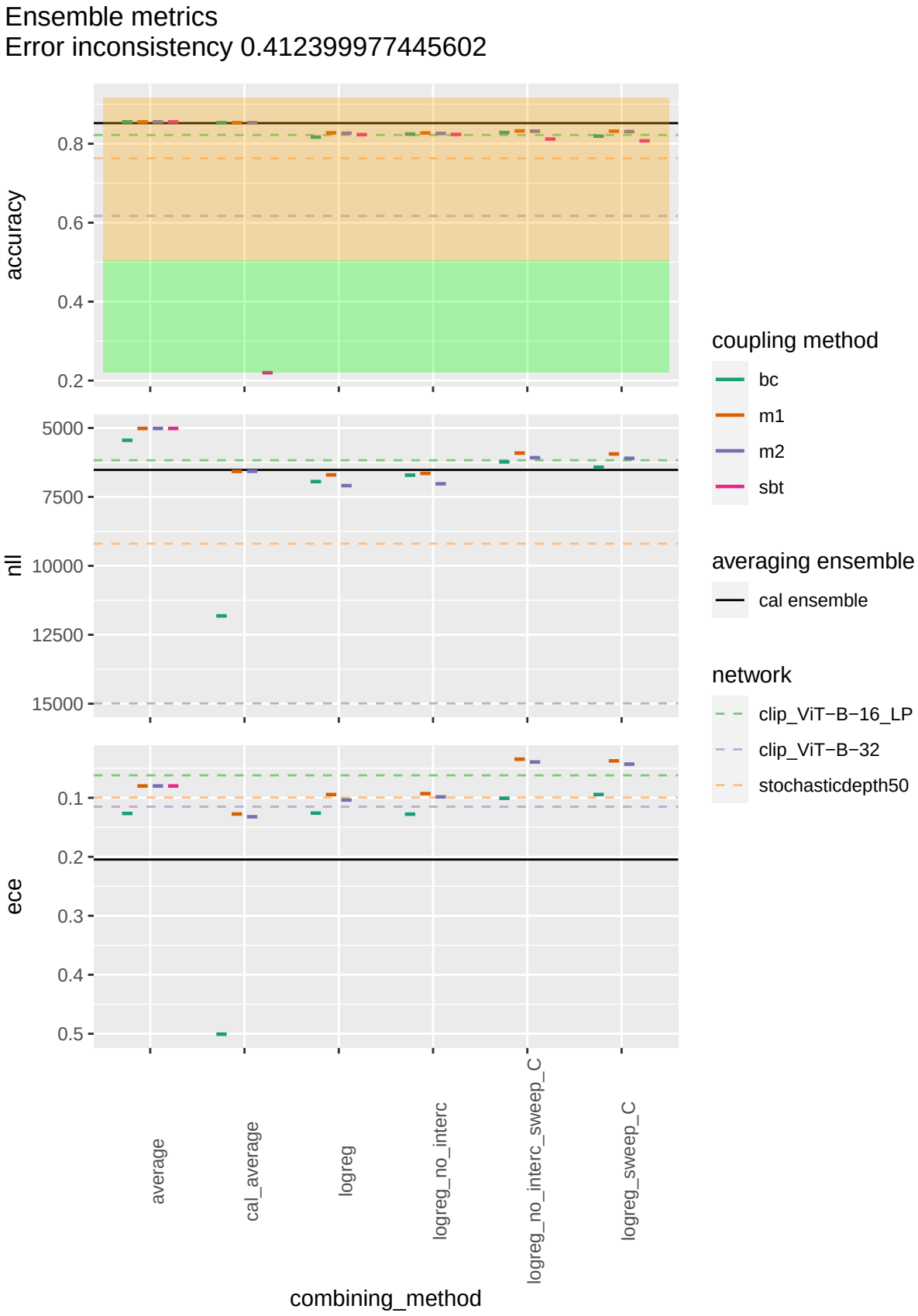

Ensemble metrics
Error inconsistency 0.407899975776672
accuracy
0.8
0.6
0.4
nll
5000
7500
10000
12500
ece
0.0
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
network
clip_ViT−B−16
clip_ViT−B−32_LP
googlenet
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble

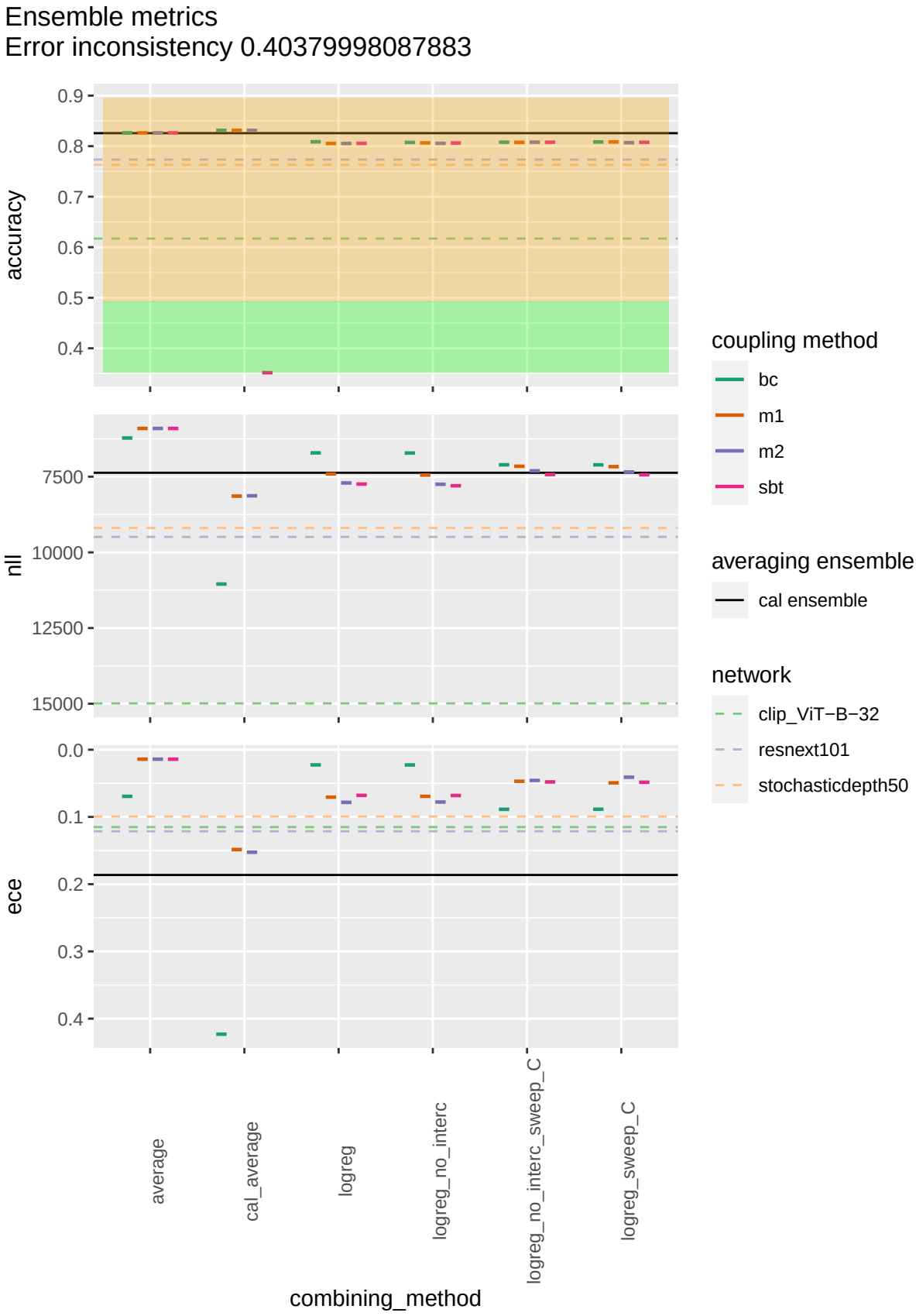
Ensemble metrics
Error inconsistency 0.403099983930588



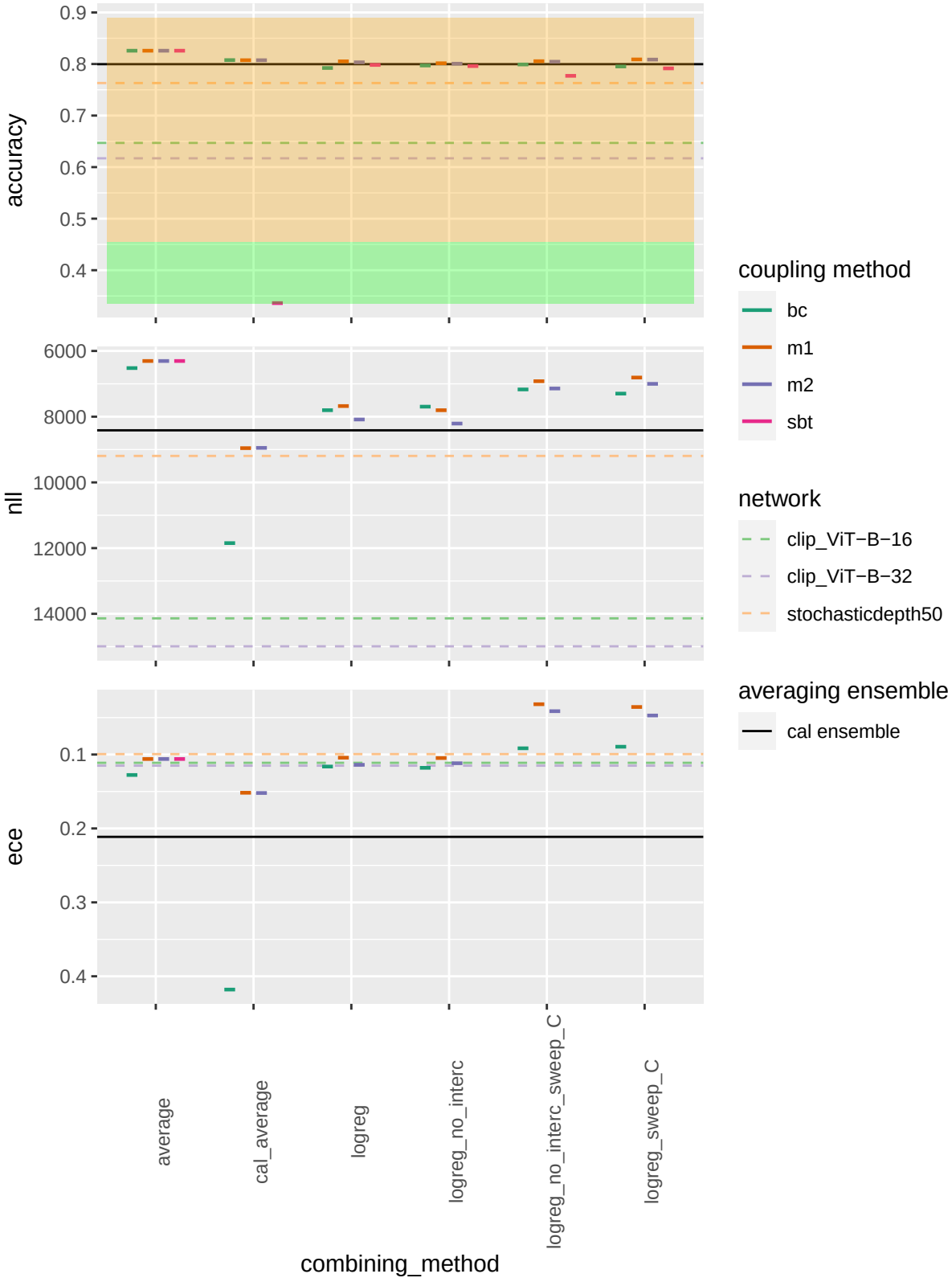
Ensemble metrics
Error inconsistency 0.27989998459816

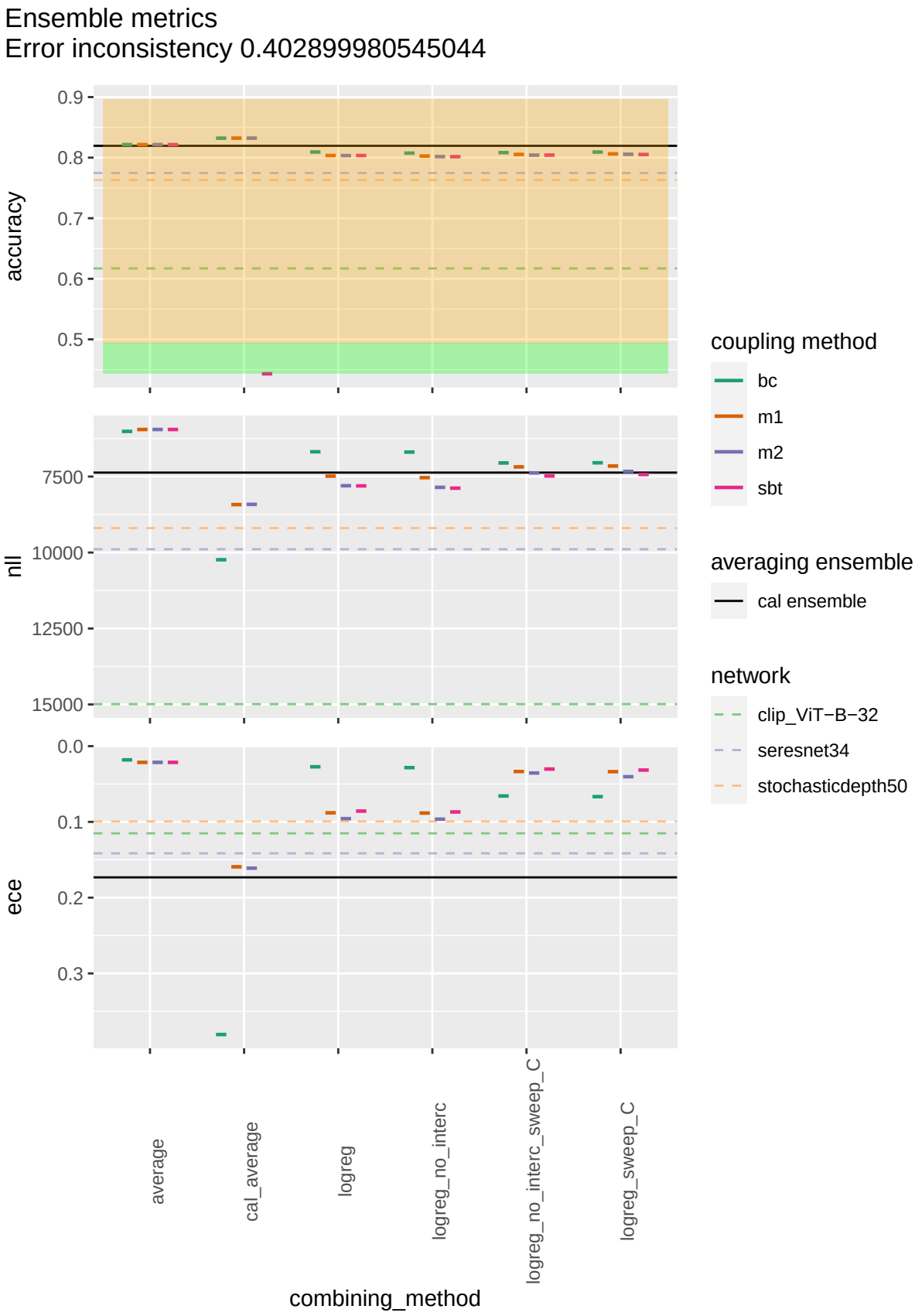



Ensemble metrics
Error inconsistency 0.412399977445602
accuracy
0.8
0.6
0.4
0.2
nll
5000
7500
10000
12500
15000
ece
0.1
0.2
0.3
0.4
0.5
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16_LP
clip_ViT−B−32
stochasticdepth50

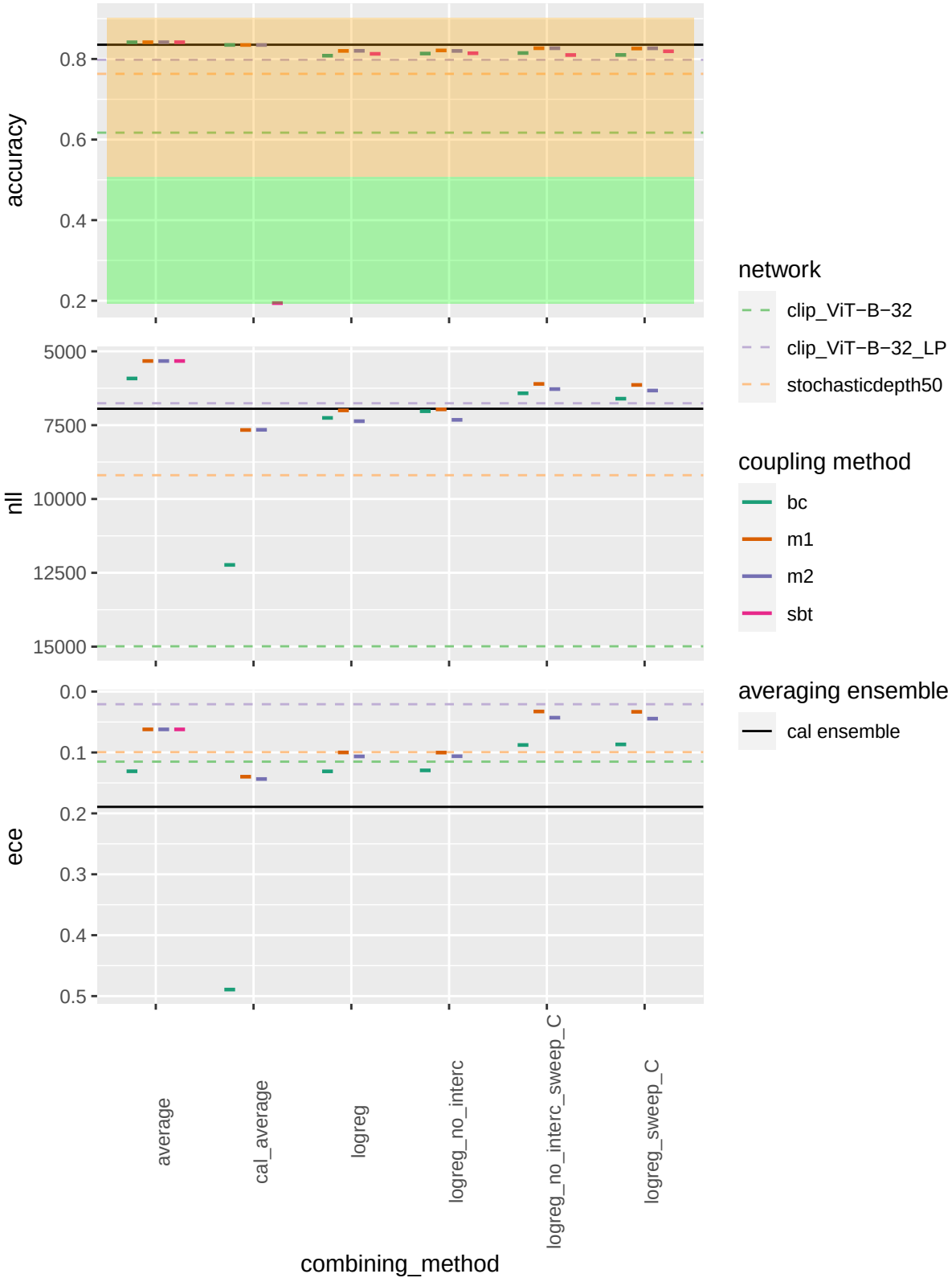

Ensemble metrics
Error inconsistency 0.40379998087883
accuracy
0.9
0.8
0.7
0.6
0.5
0.4
nll
7500
10000
12500
15000
ece
0.0
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
resnext101
stochasticdepth50

Ensemble metrics
Error inconsistency 0.434699982404709

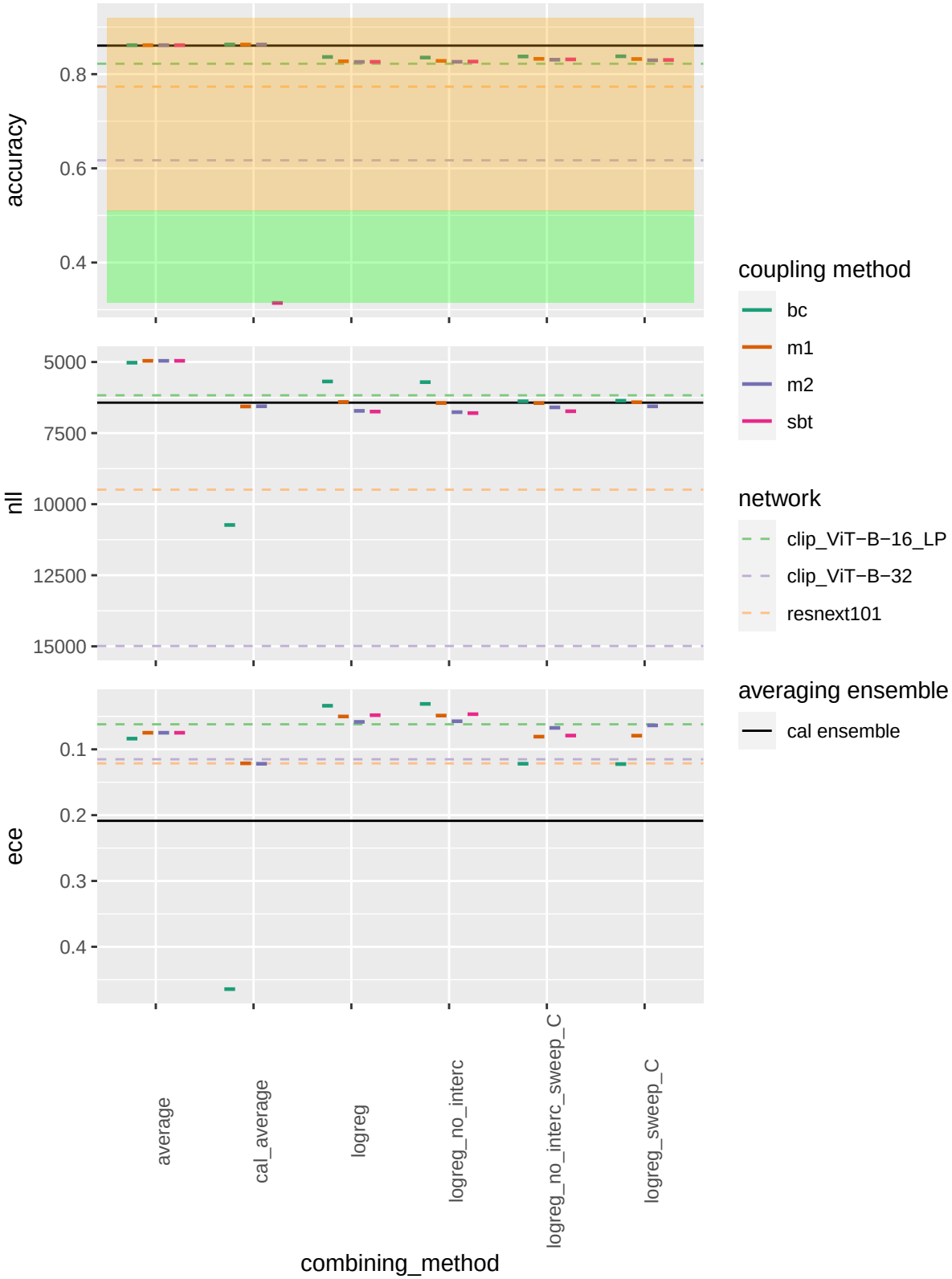


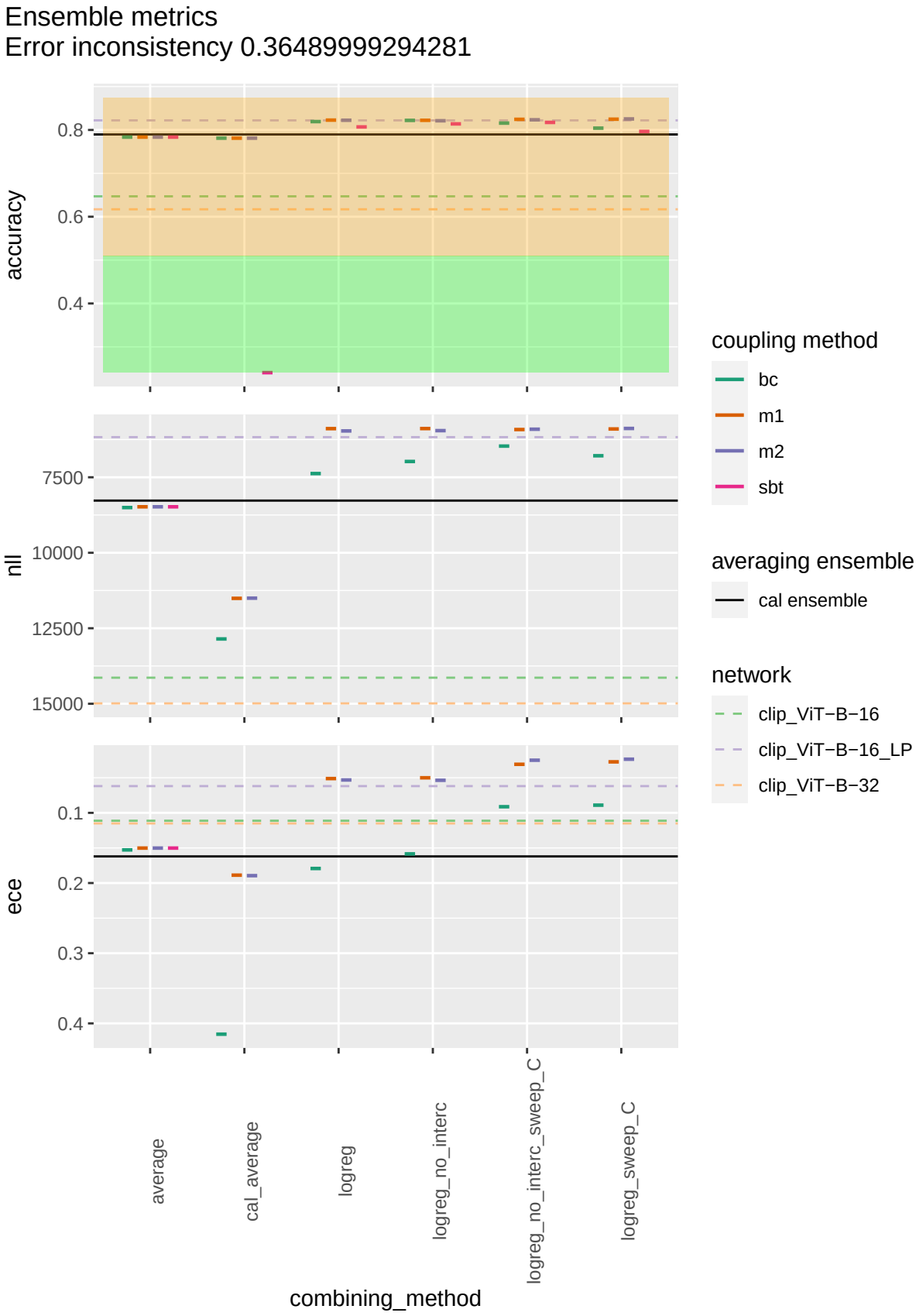

Ensemble metrics
Error inconsistency 0.402899980545044
accuracy
nll
ece
0.9
0.8
0.7
0.6
0.5
7500
10000
12500
15000
0.0
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
seresnet34
stochasticdepth50

Ensemble metrics
Error inconsistency 0.39479997754097

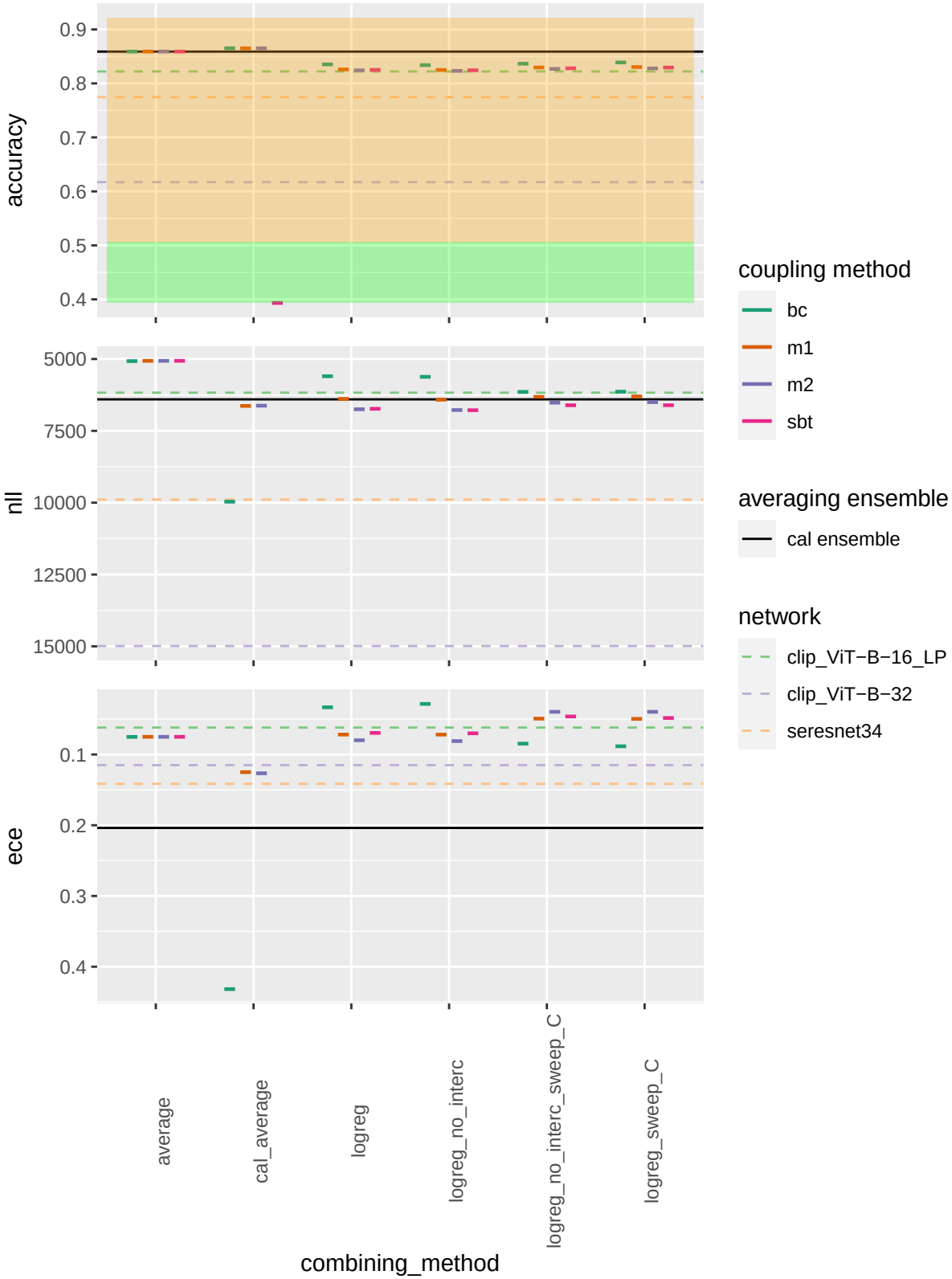


Ensemble metrics
Error inconsistency 0.409999996423721

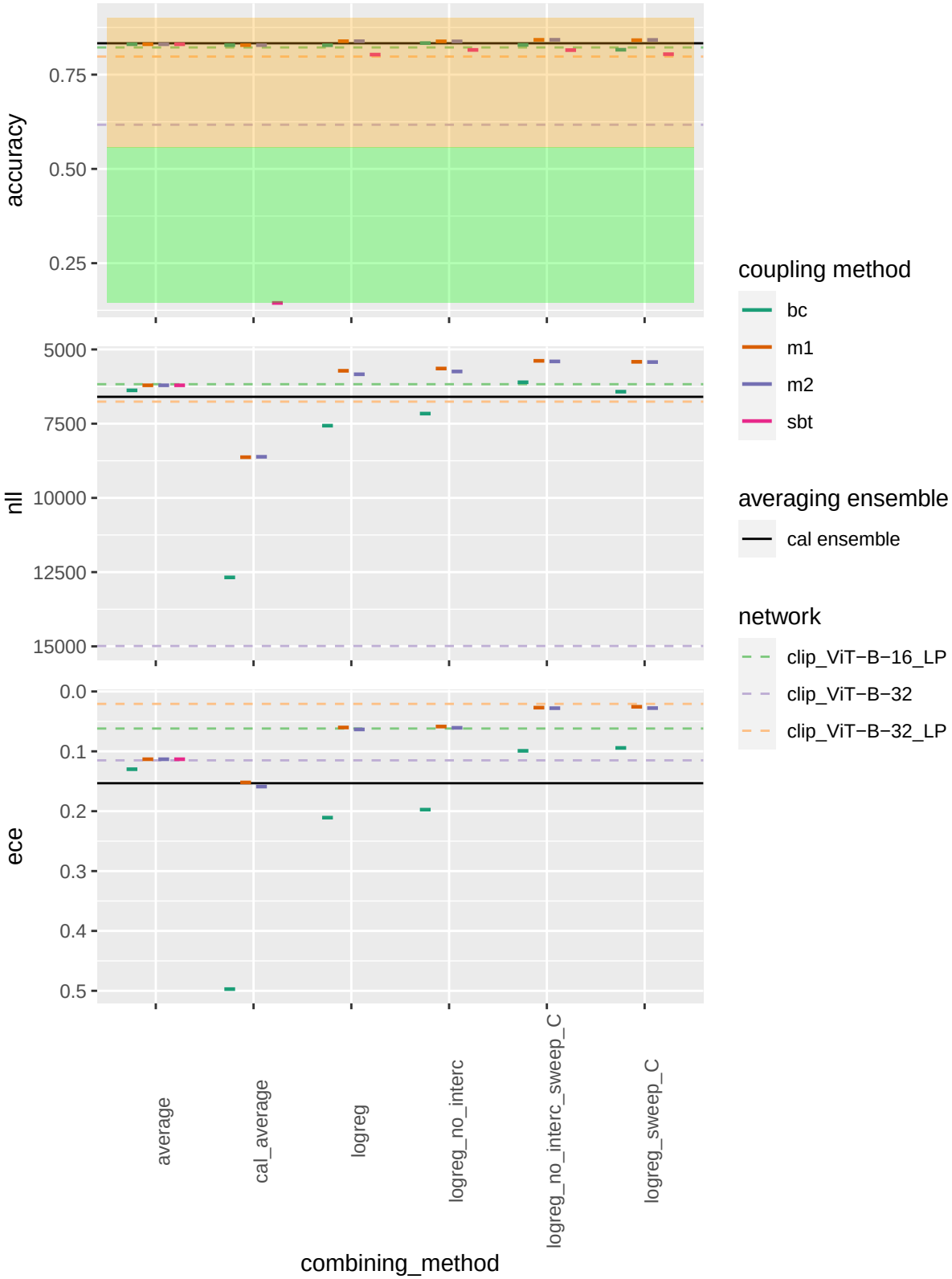



Ensemble metrics
Error inconsistency 0.36489999294281
accuracy
0.8
0.6
0.4
nll
7500
10000
12500
15000
ece
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−16_LP
clip_ViT−B−32

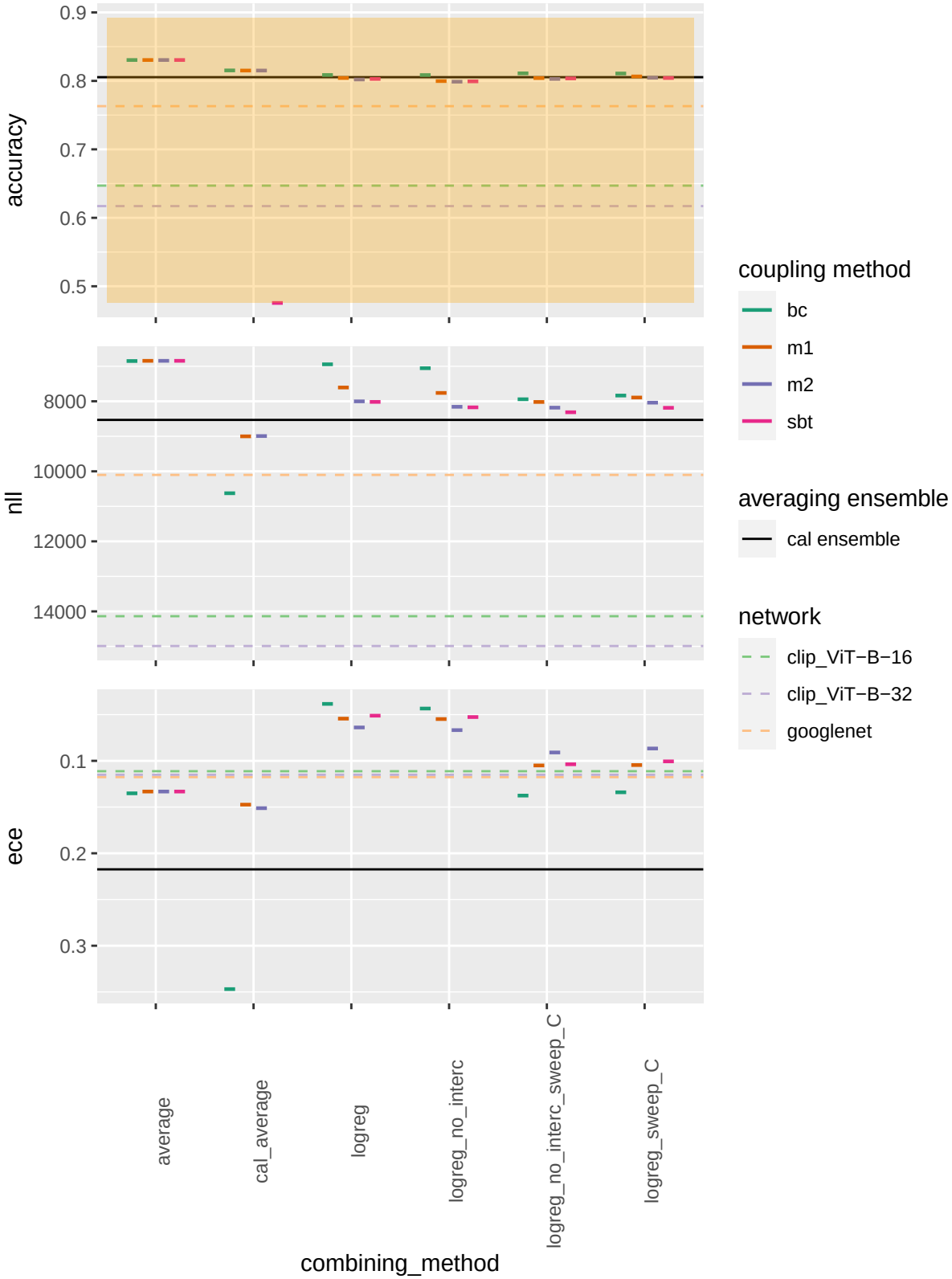
Ensemble metrics
Error inconsistency 0.415299981832504

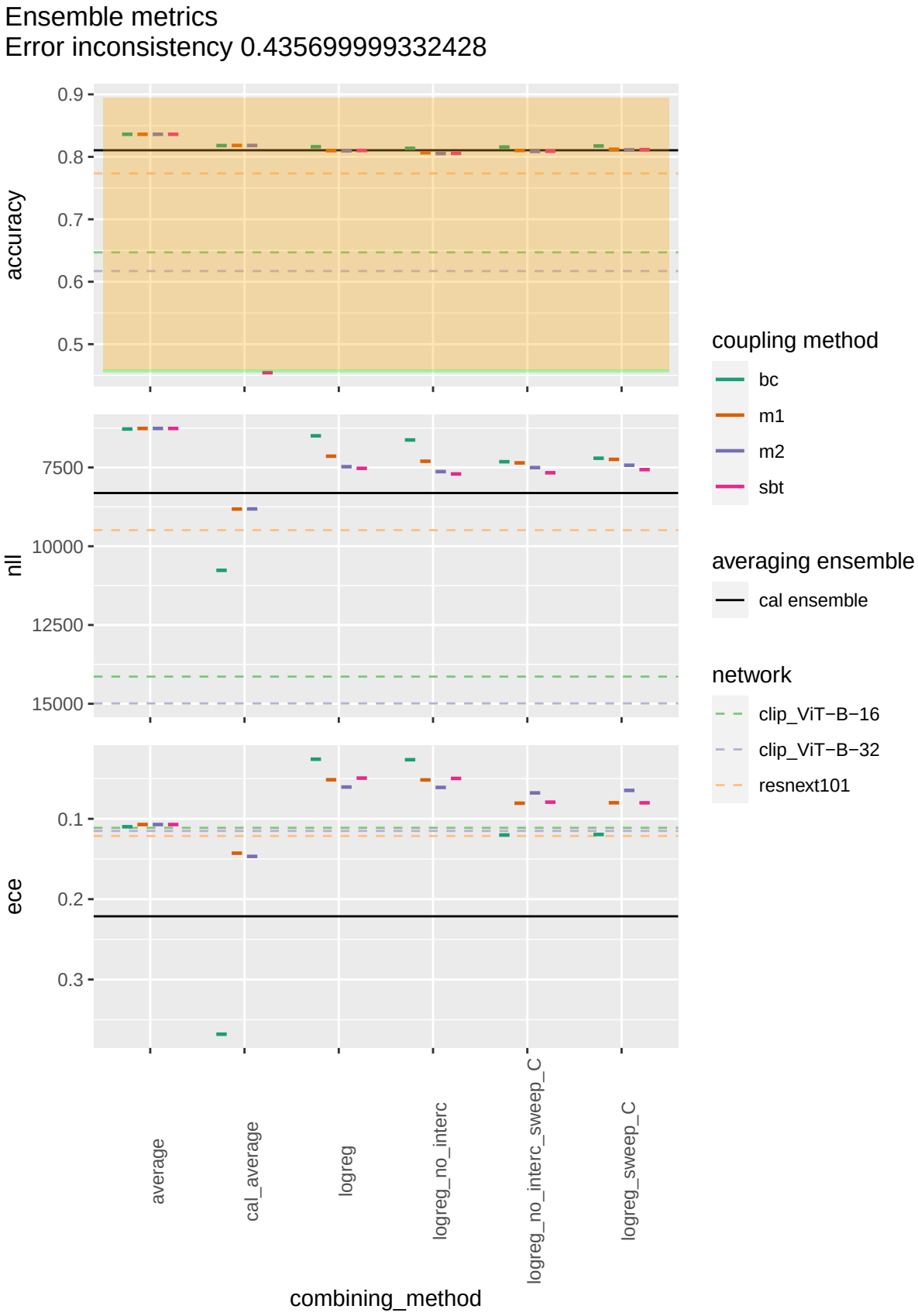


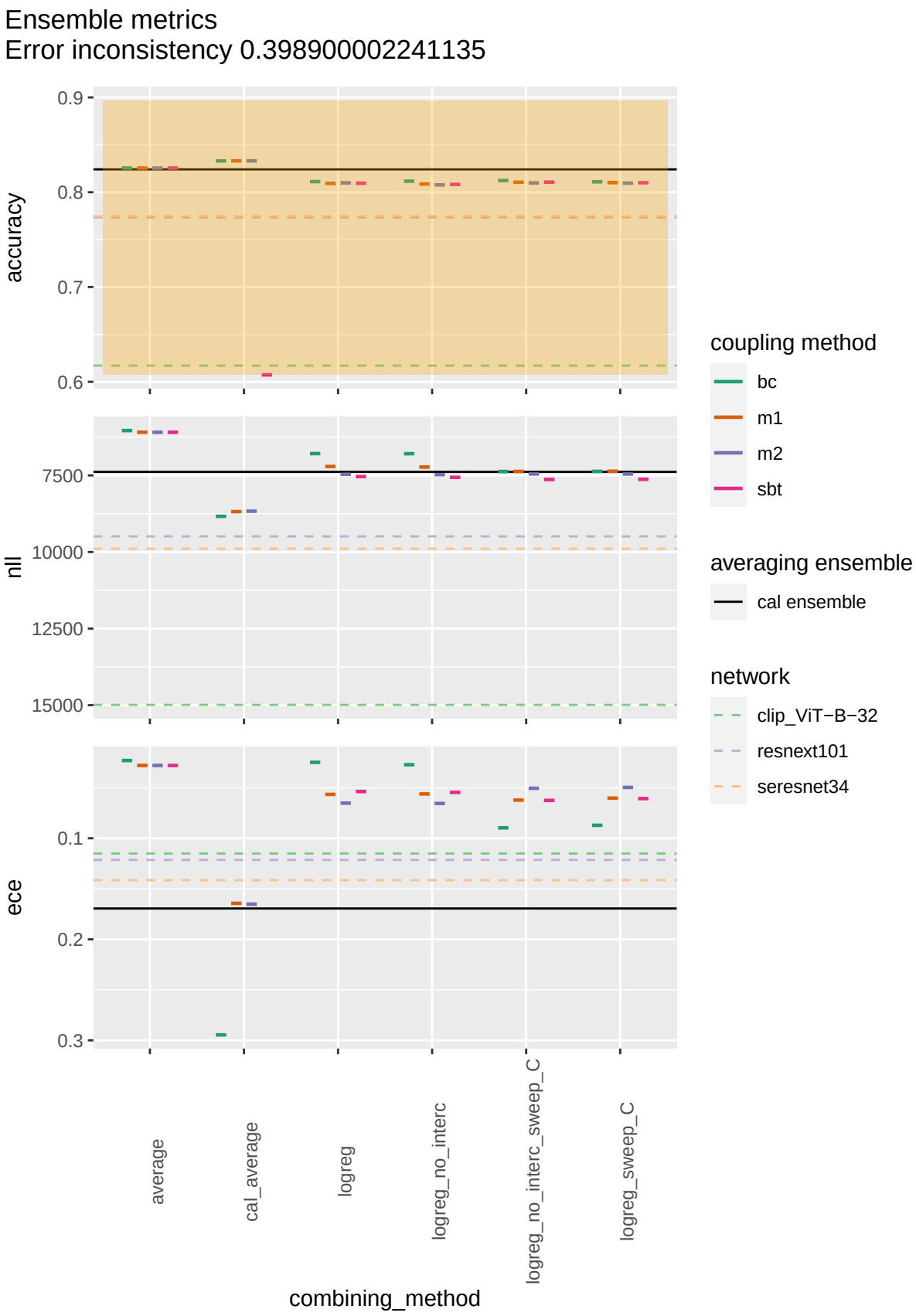
Ensemble metrics
Error inconsistency 0.344300001859665

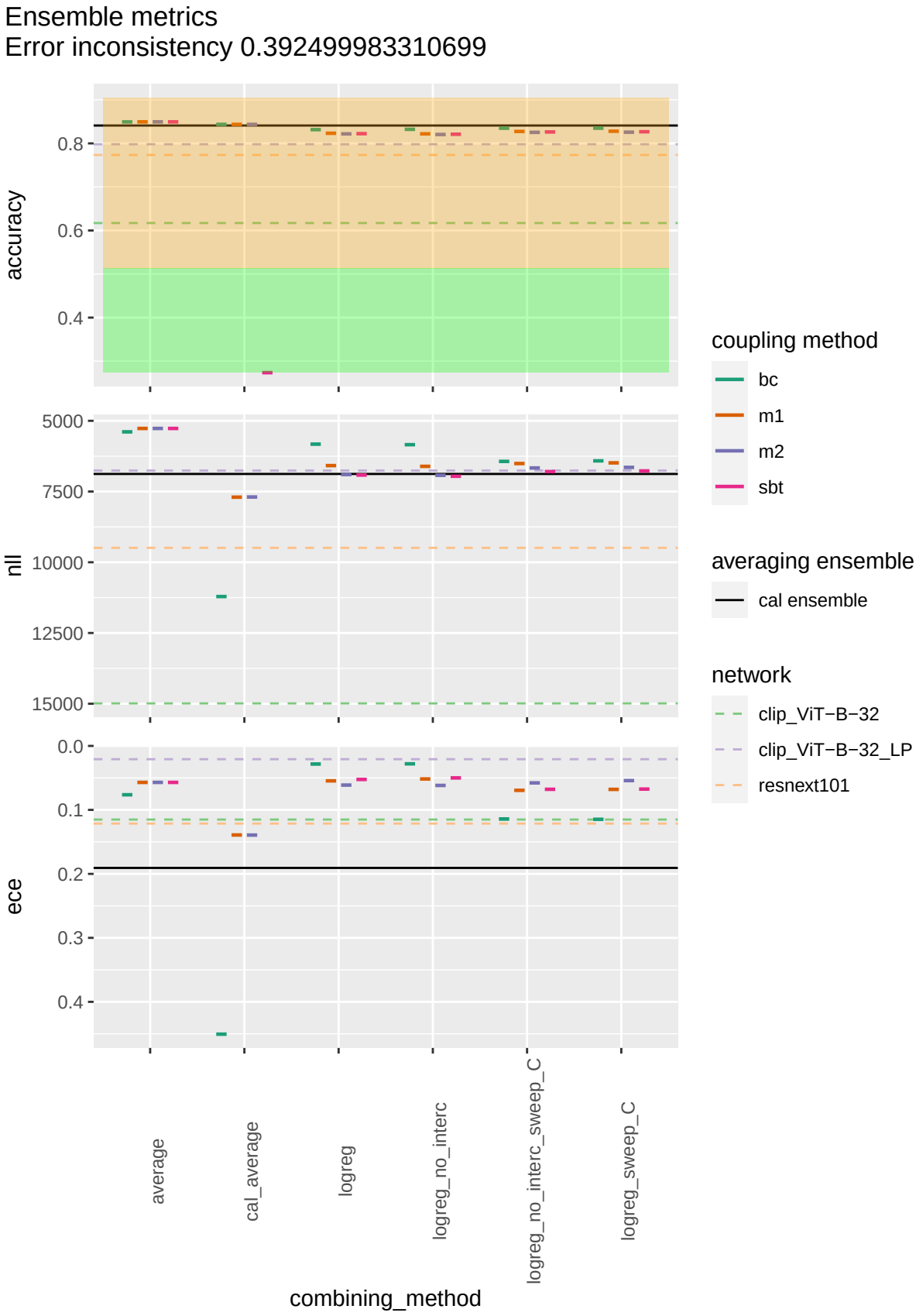


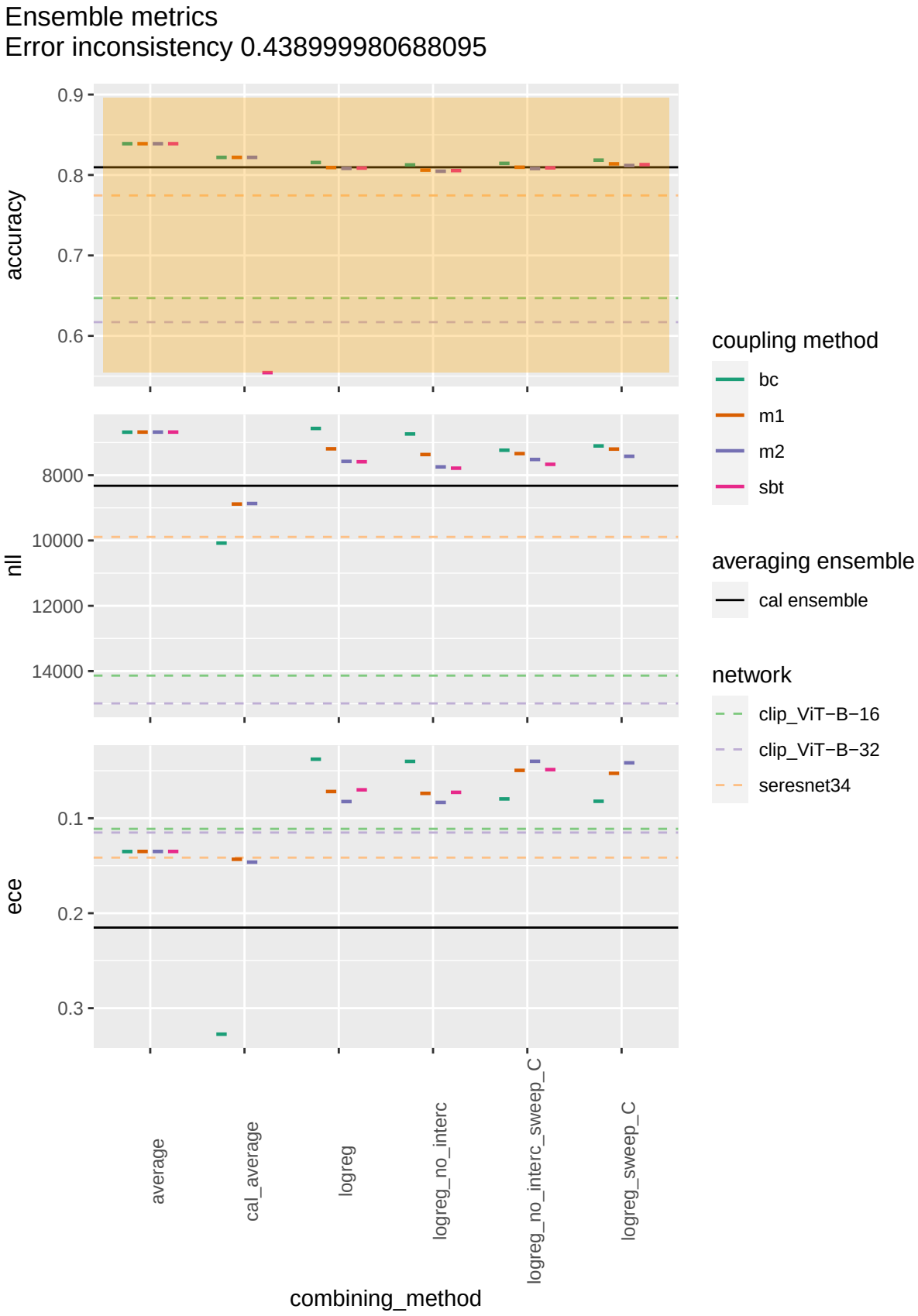
Ensemble metrics
Error inconsistency 0.438899993896484

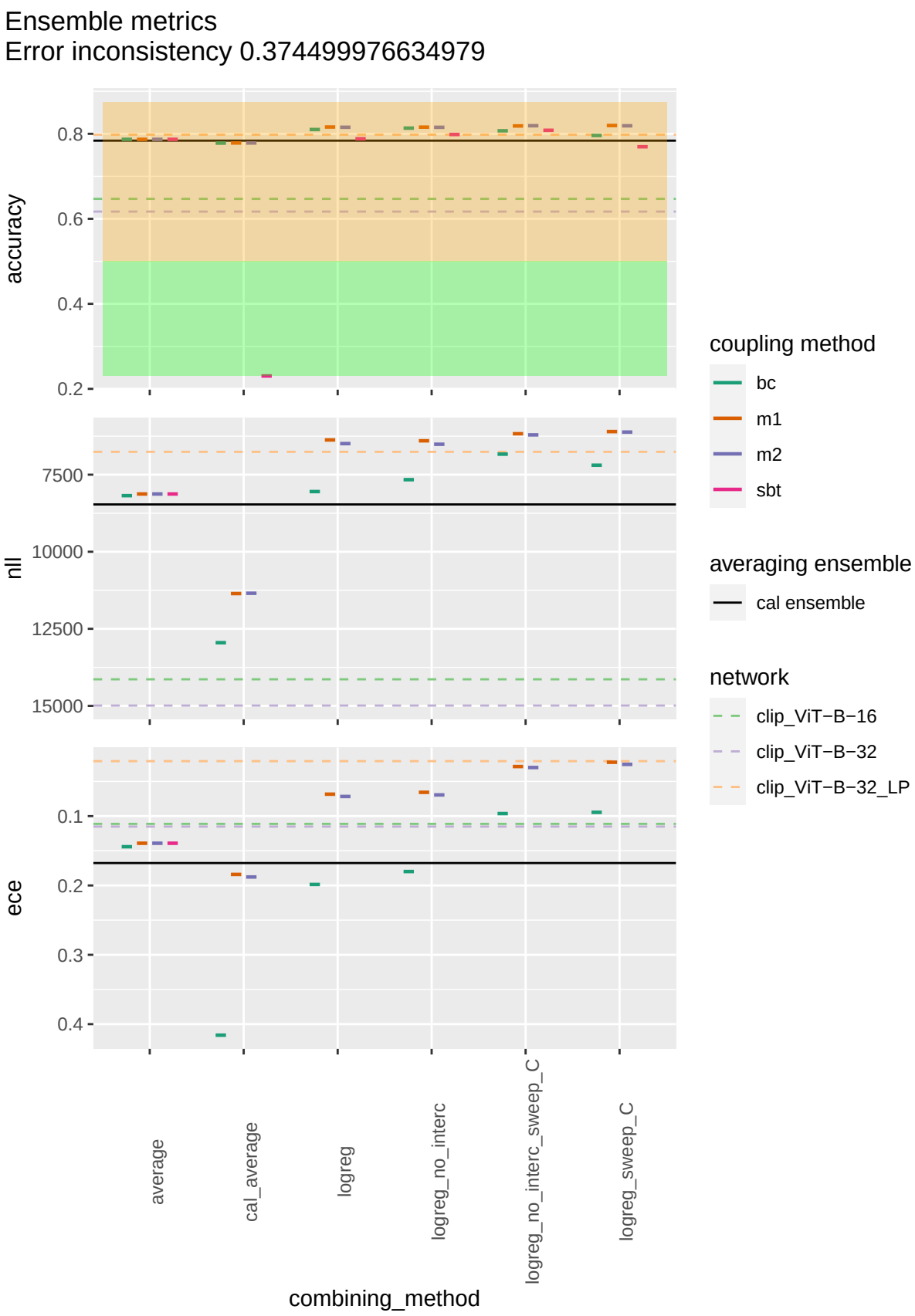


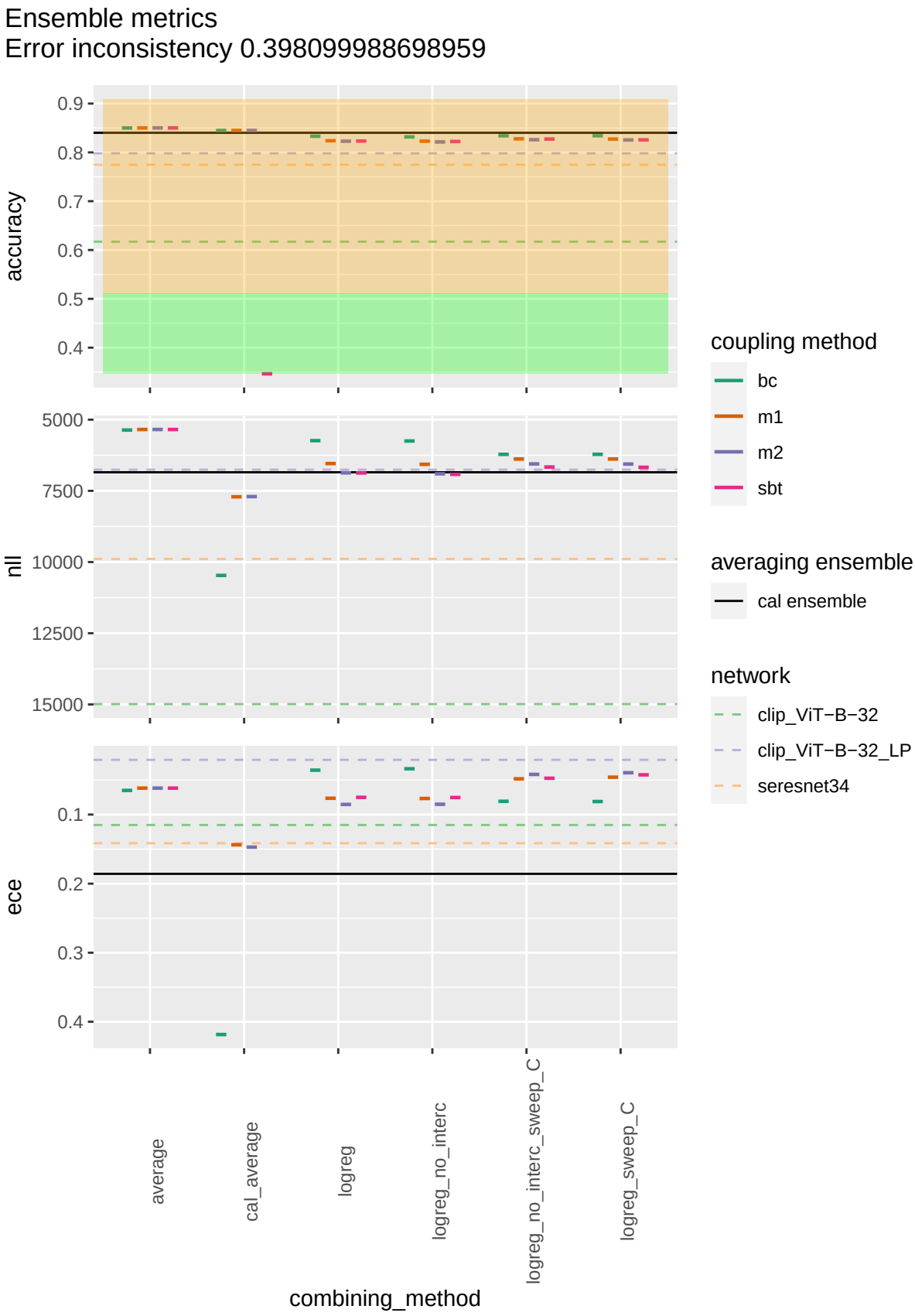

Ensemble metrics
Error inconsistency 0.435699999332428
accuracy
0.9
0.8
0.7
0.6
0.5
nll
7500
10000
12500
15000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−32
resnext101


Ensemble metrics
Error inconsistency 0.398900002241135
accuracy
0.9
0.8
0.7
0.6
nll
7500
10000
12500
15000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
resnext101
seresnet34

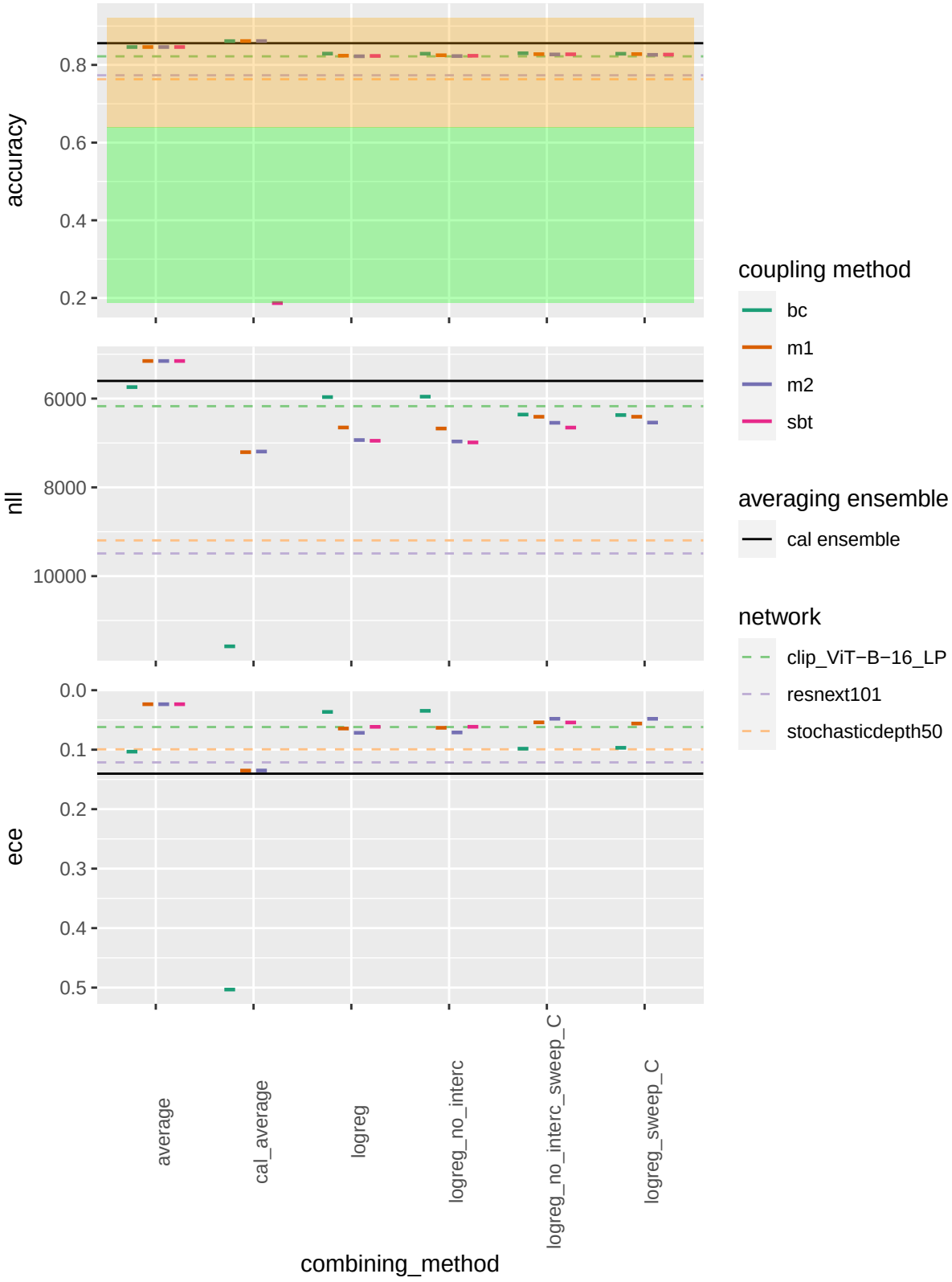

Ensemble metrics
Error inconsistency 0.392499983310699
accuracy
0.8
0.6
0.4
nll
5000
7500
10000
12500
15000
ece
0.0
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
clip_ViT−B−32_LP
resnext101


Ensemble metrics
Error inconsistency 0.438999980688095
accuracy
0.9
0.8
0.7
0.6
nll
8000
10000
12000
14000
ece
0.1
0.2
0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−32
seresnet34

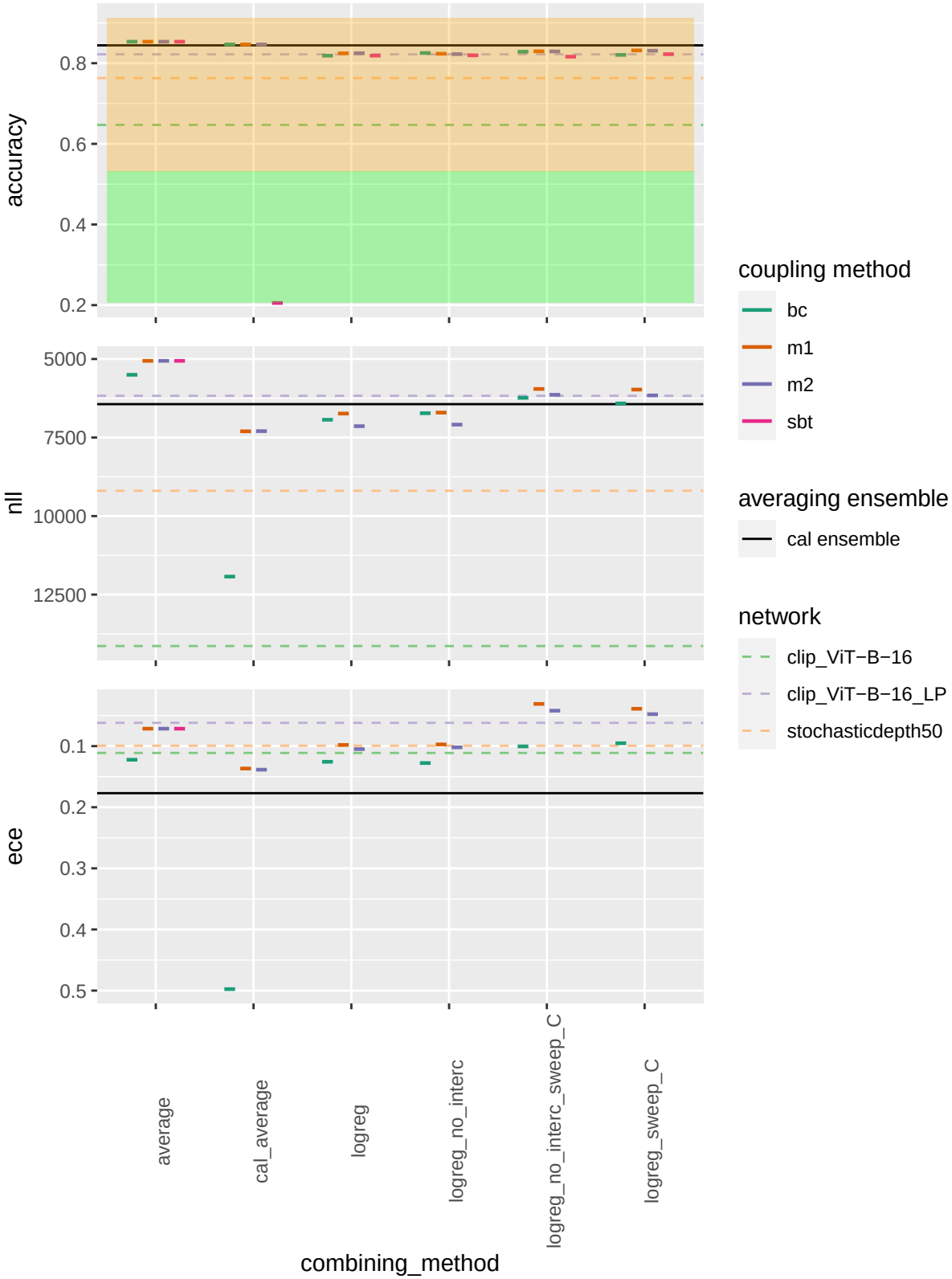

Ensemble metrics
Error inconsistency 0.374499976634979
accuracy
0.8
0.6
0.4
0.2
nll
7500
10000
12500
15000
ece
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−32
clip_ViT−B−32_LP

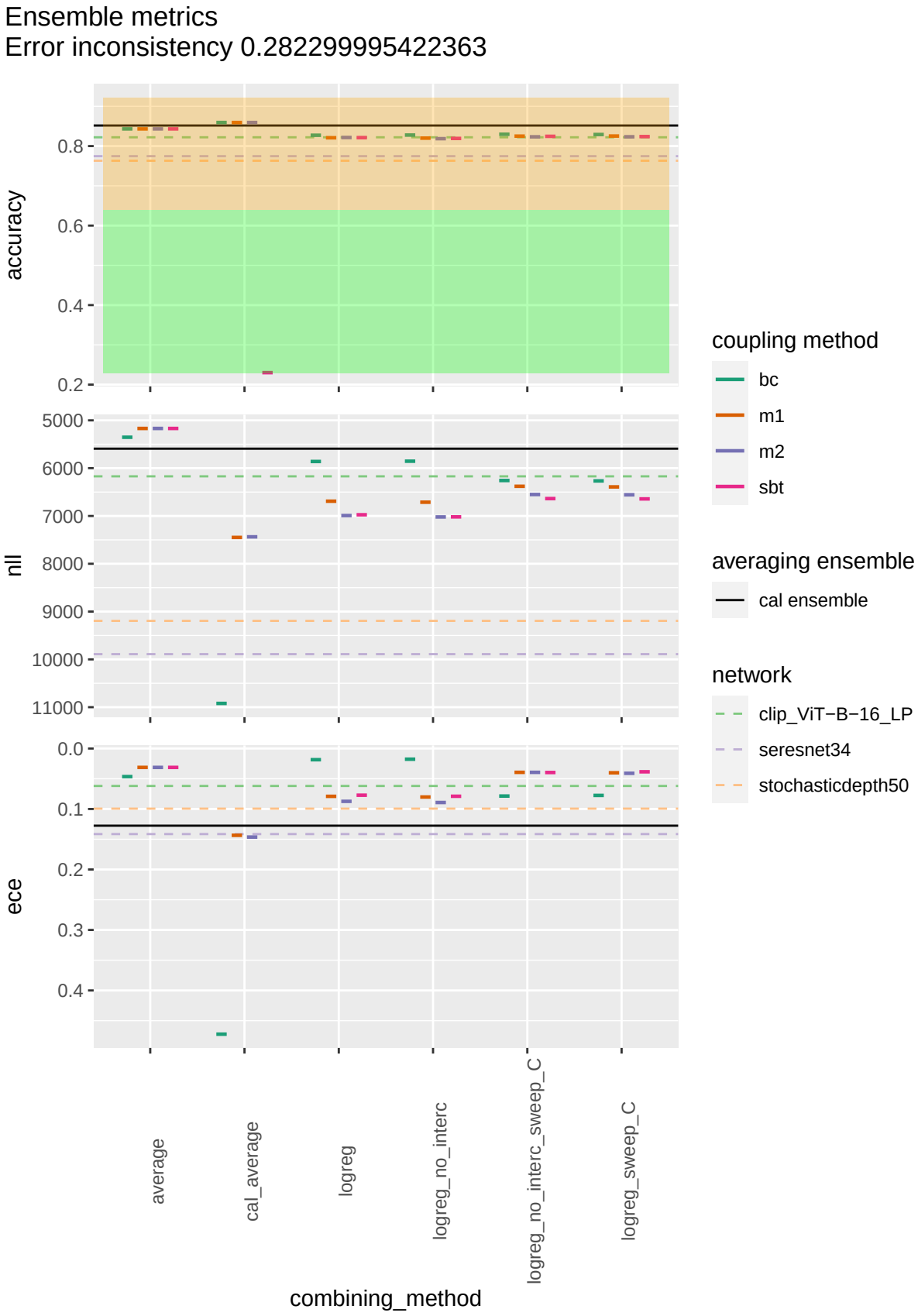

Ensemble metrics
Error inconsistency 0.398099988698959
accuracy
0.9
0.8
0.7
0.6
0.5
0.4
nll
5000
7500
10000
12500
15000
ece
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
clip_ViT−B−32_LP
seresnet34

Ensemble metrics
Error inconsistency 0.282700002193451

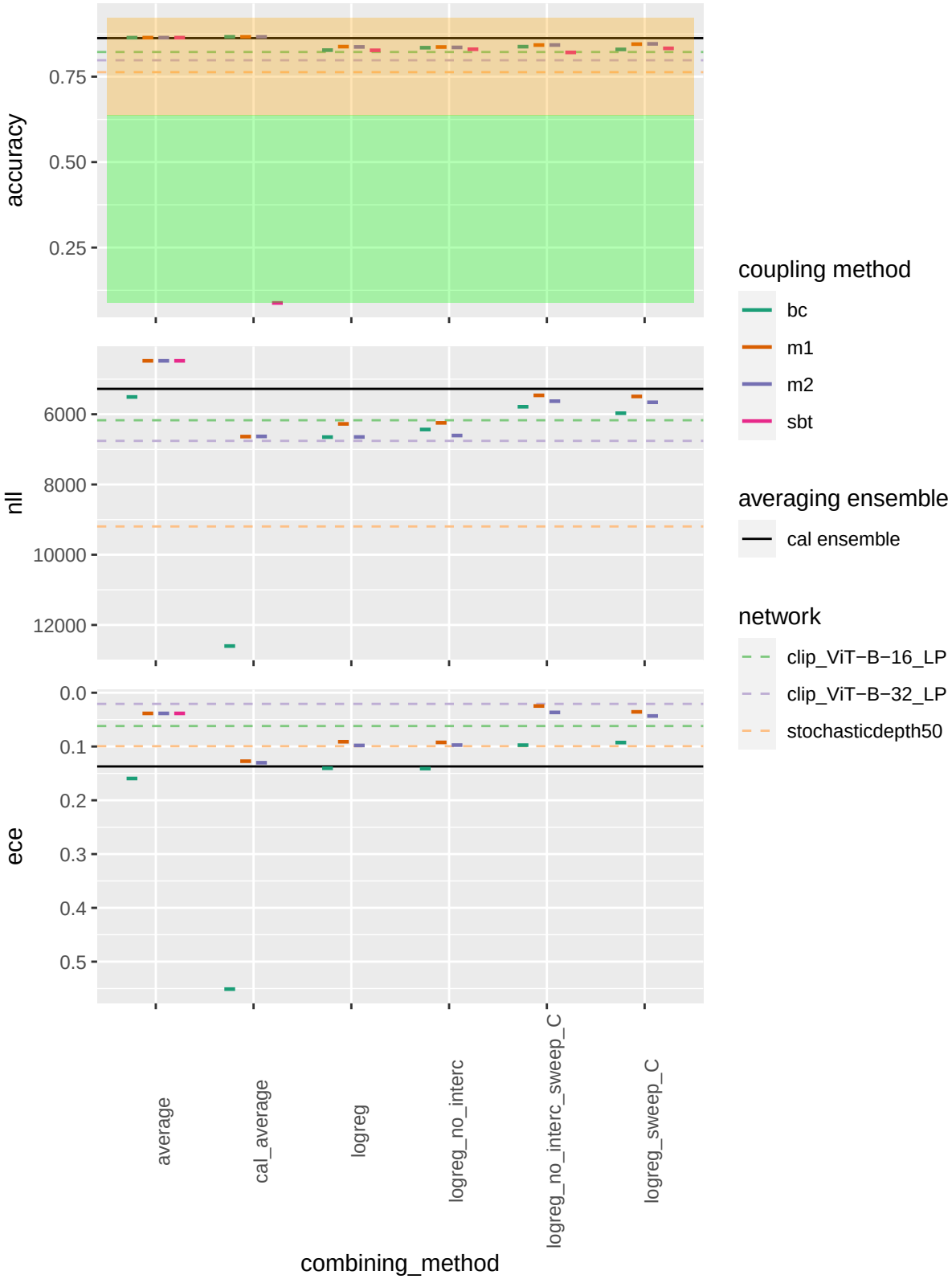


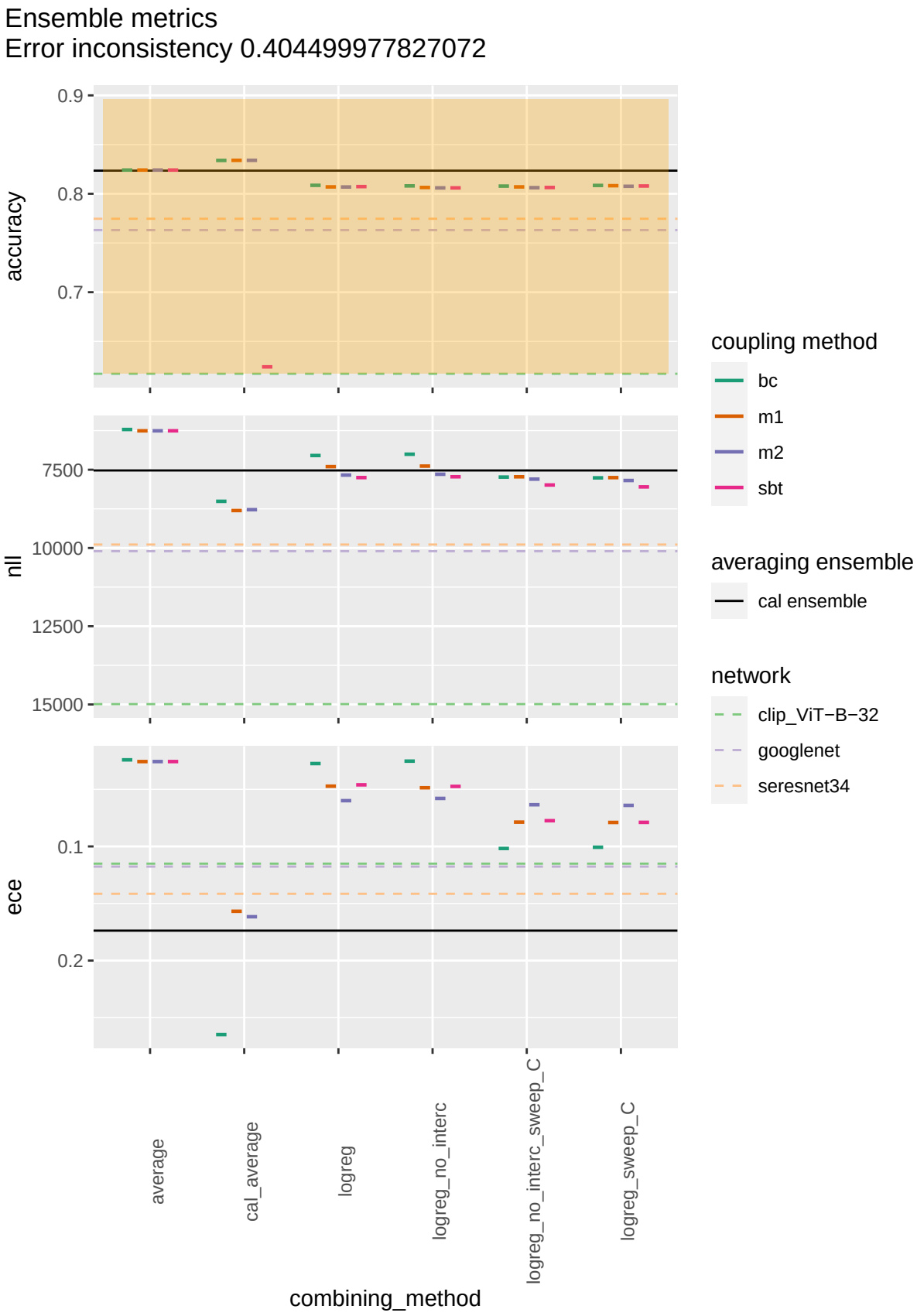
Ensemble metrics
Error inconsistency 0.380899995565414



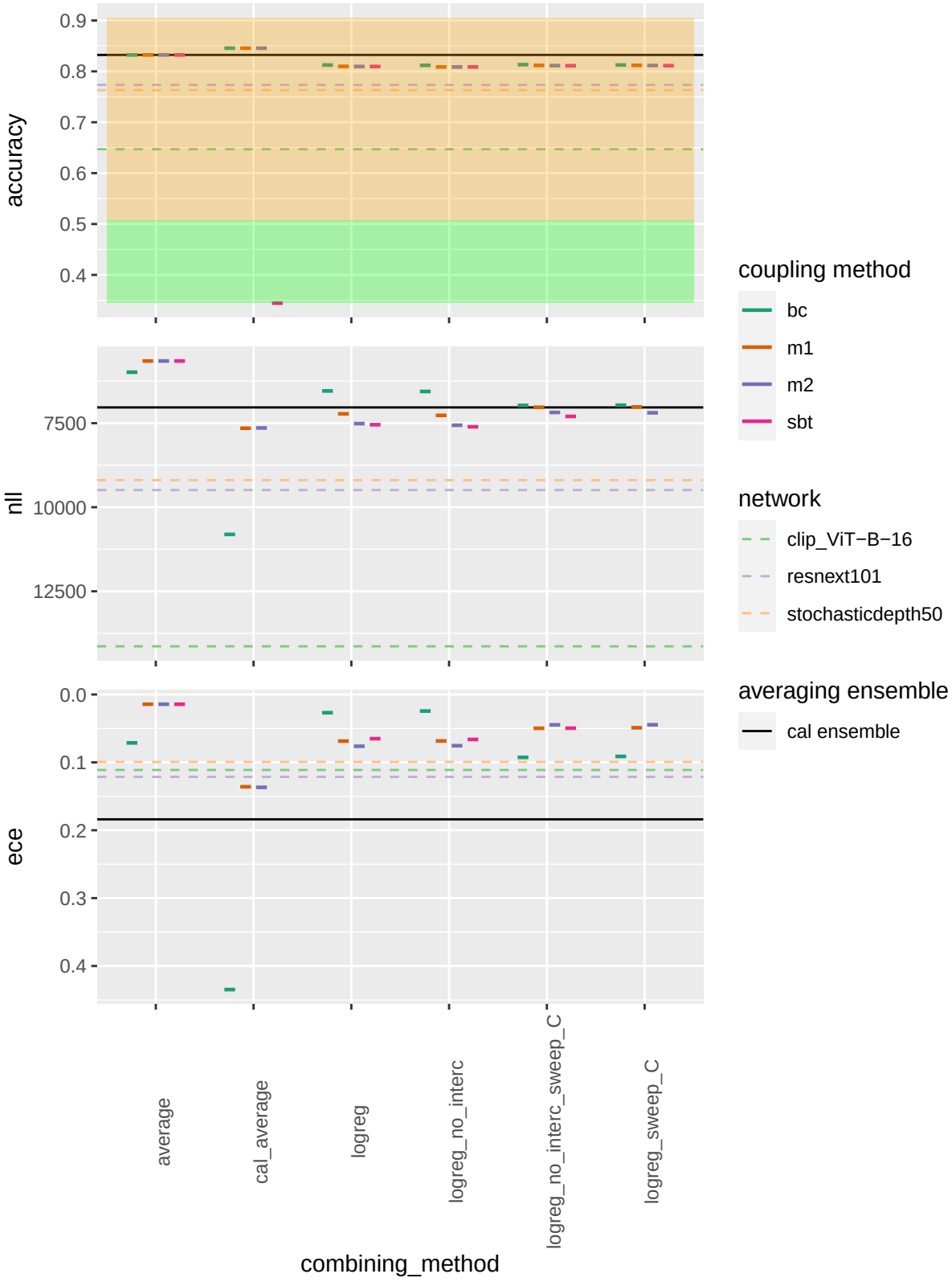

Ensemble metrics
Error inconsistency 0.282299995422363
accuracy
0.8
0.6
0.4
0.2
nll
5000
6000
7000
8000
9000
10000
11000
ece
0.0
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16_LP
seresnet34
stochasticdepth50

Ensemble metrics
Error inconsistency 0.284999996423721

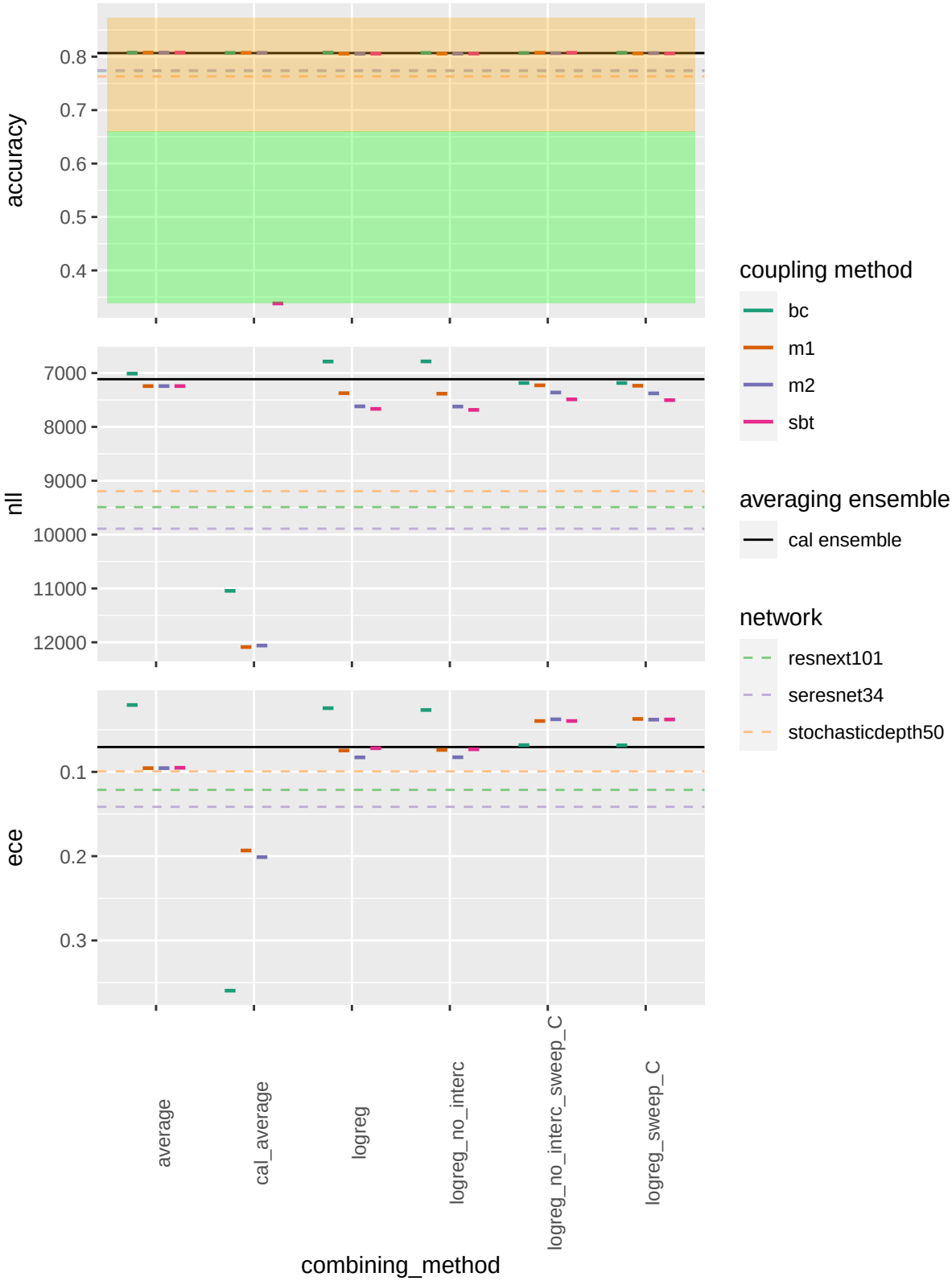


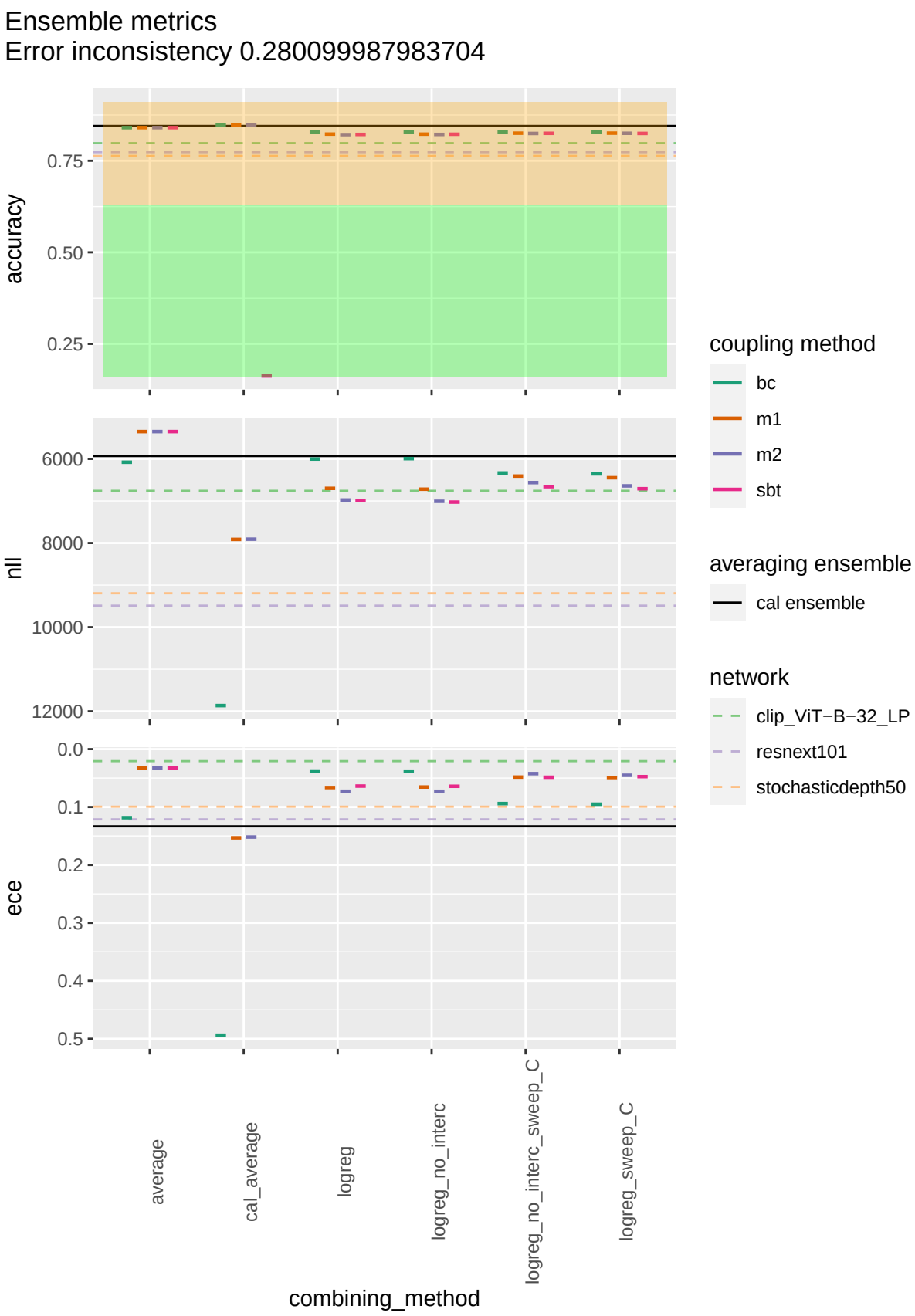

Ensemble metrics
Error inconsistency 0.404499977827072
accuracy
0.9
0.8
0.7
nll
7500
10000
12500
15000
ece
0.1
0.2
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32
googlenet
seresnet34

Ensemble metrics
Error inconsistency 0.398599982261658

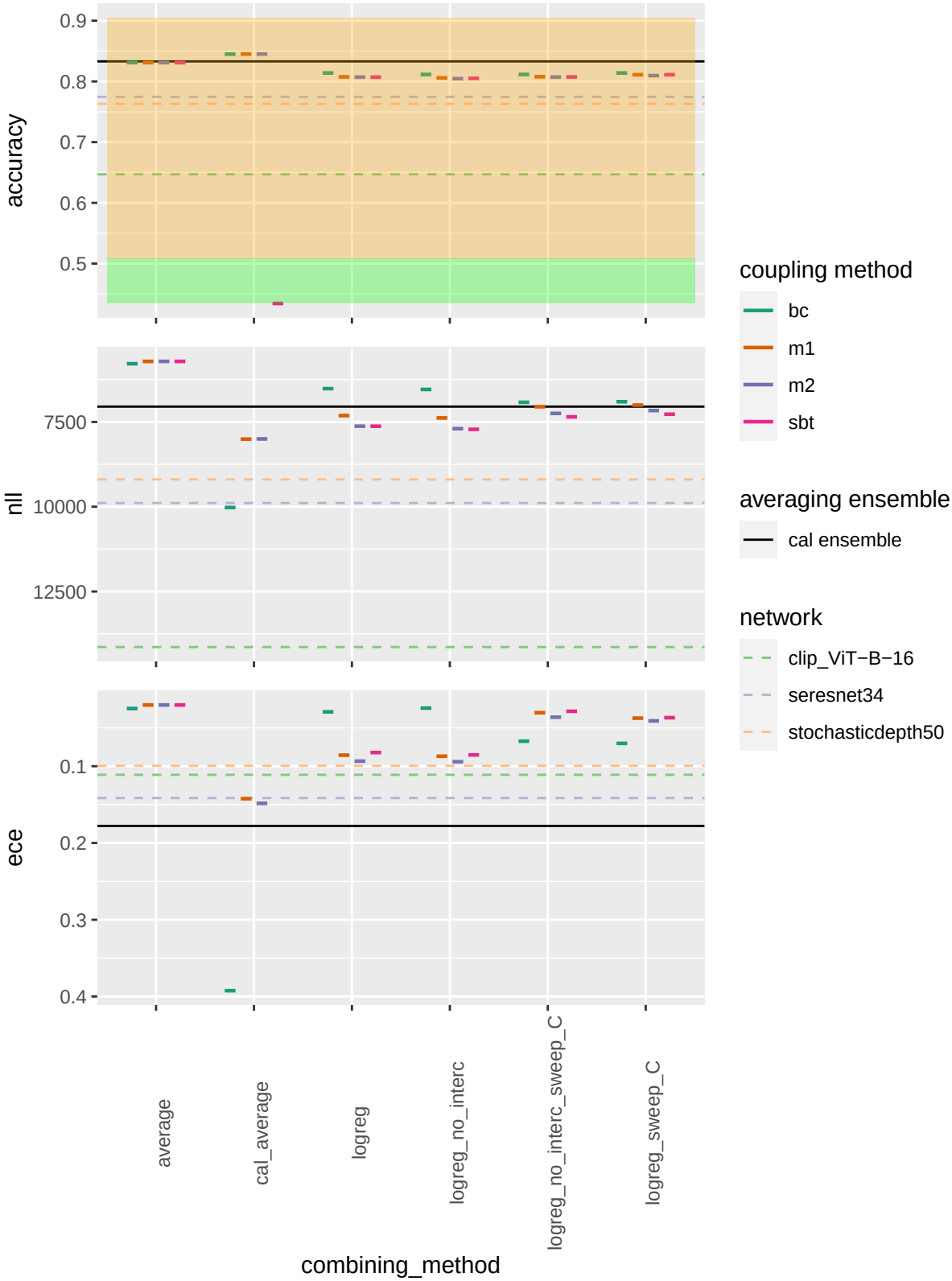


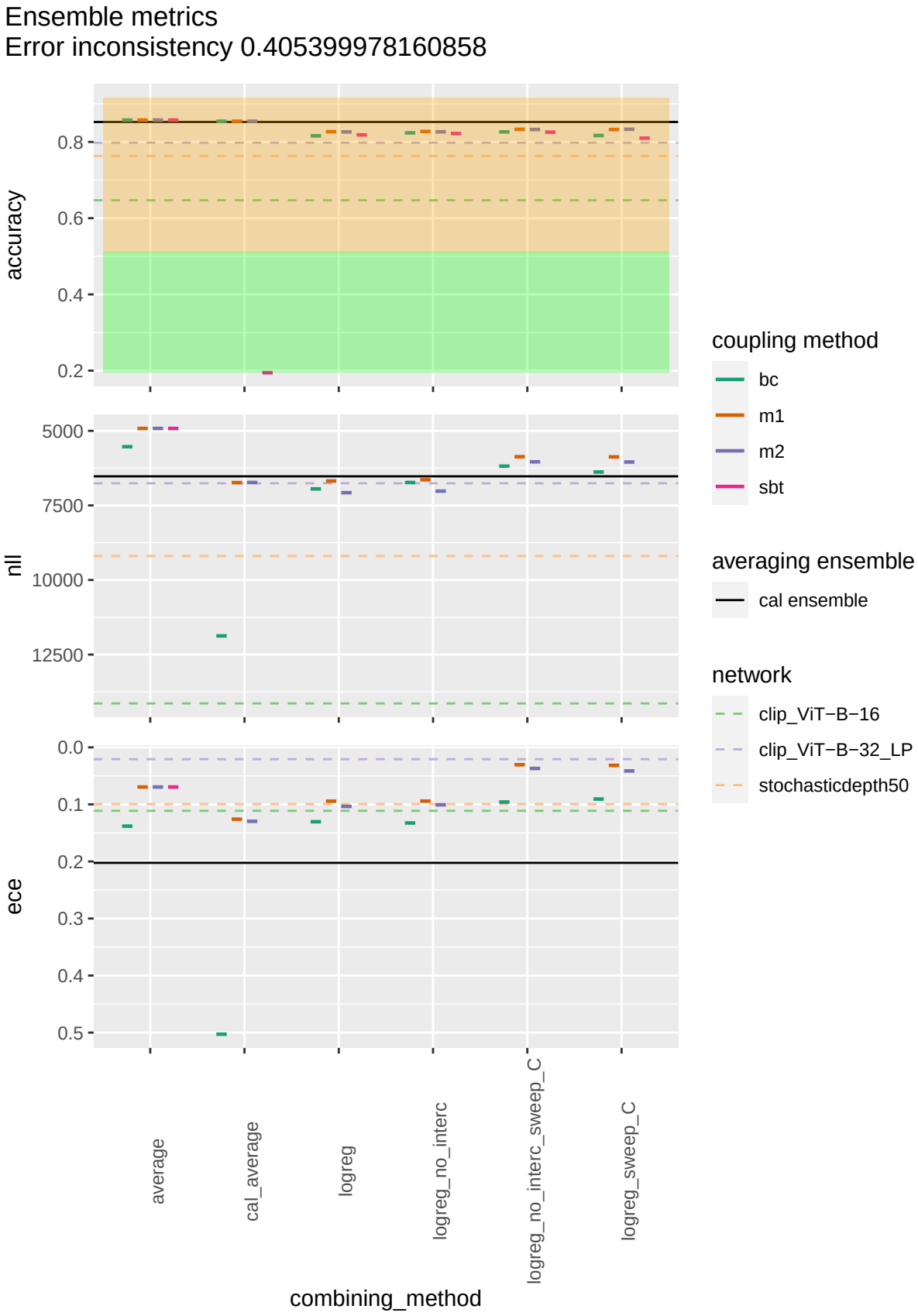
Ensemble metrics
Error inconsistency 0.212799996137619

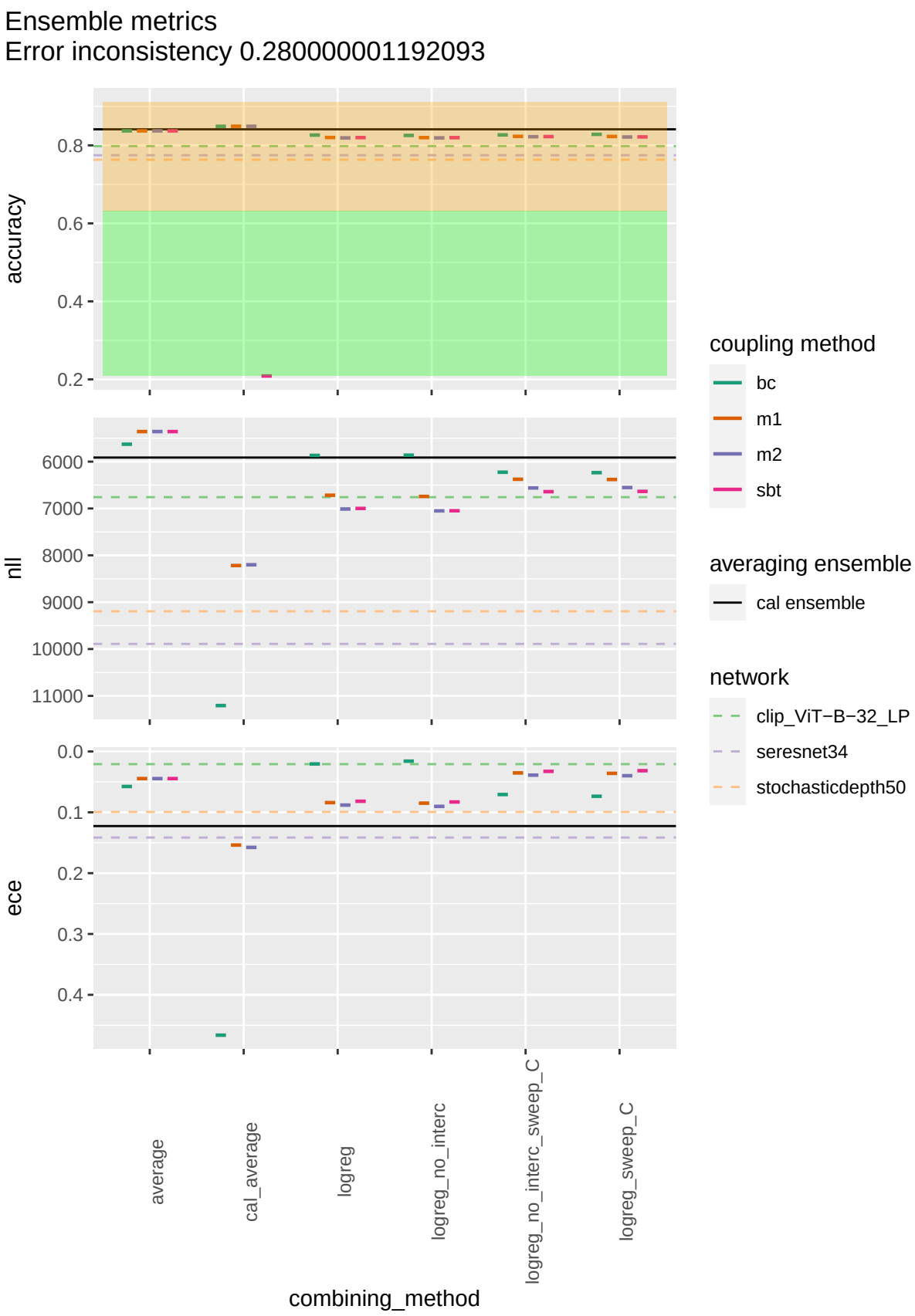



Ensemble metrics
Error inconsistency 0.280099987983704
accuracy
0.75
0.50
0.25
nll
6000
8000
10000
12000
ece
0.0
0.1
0.2
0.3
0.4
0.5
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32_LP
resnext101
stochasticdepth50

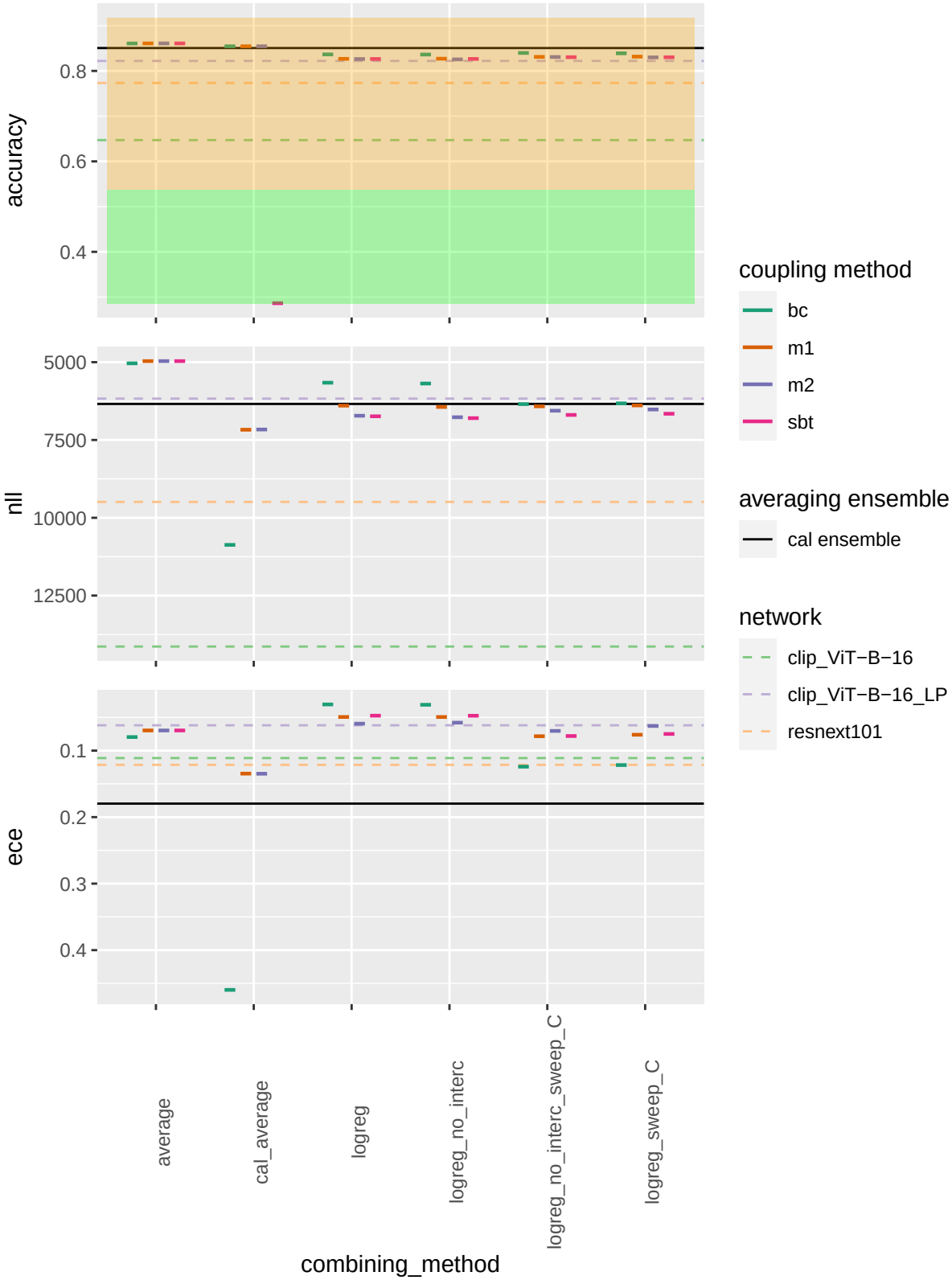
Ensemble metrics
Error inconsistency 0.396200001239777



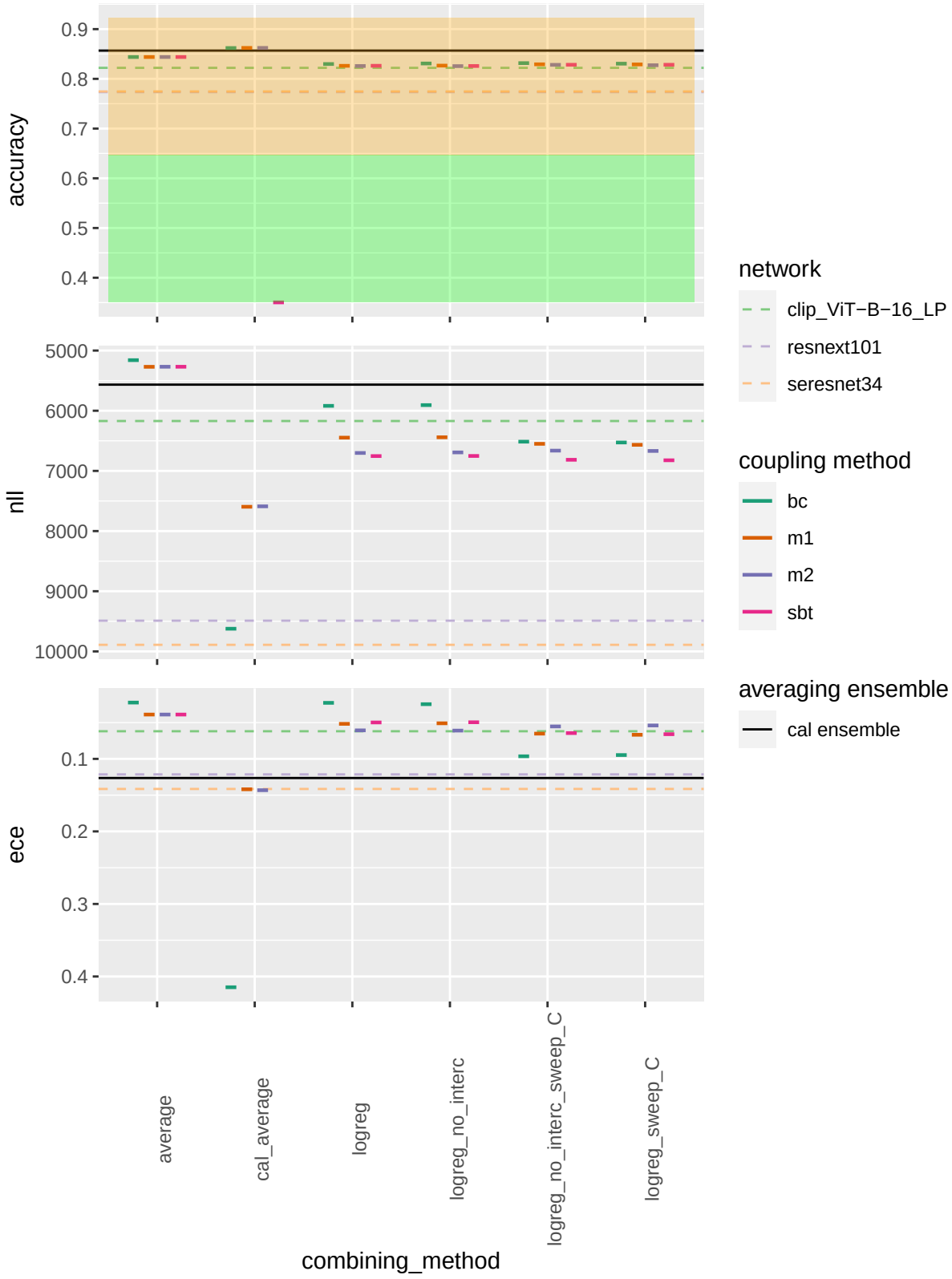

Ensemble metrics
Error inconsistency 0.405399978160858
accuracy
0.8
0.6
0.4
0.2
nll
5000
7500
10000
12500
ece
0.0
0.1
0.2
0.3
0.4
0.5
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−32_LP
stochasticdepth50


Ensemble metrics
Error inconsistency 0.280000001192093
accuracy
0.8
0.6
0.4
0.2
nll
6000
7000
8000
9000
10000
11000
ece
0.0
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−32_LP
seresnet34
stochasticdepth50

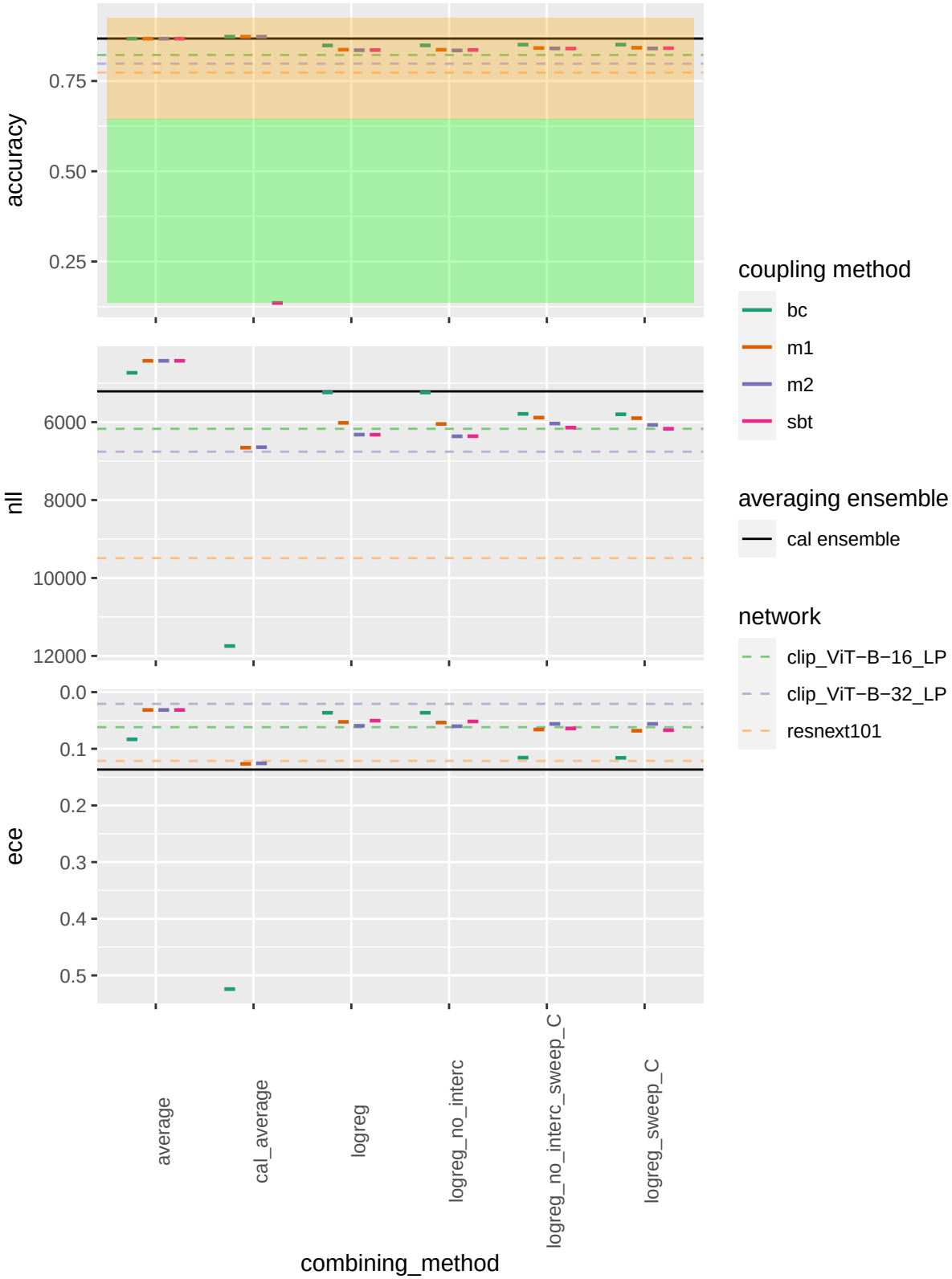
Ensemble metrics
Error inconsistency 0.380199998617172



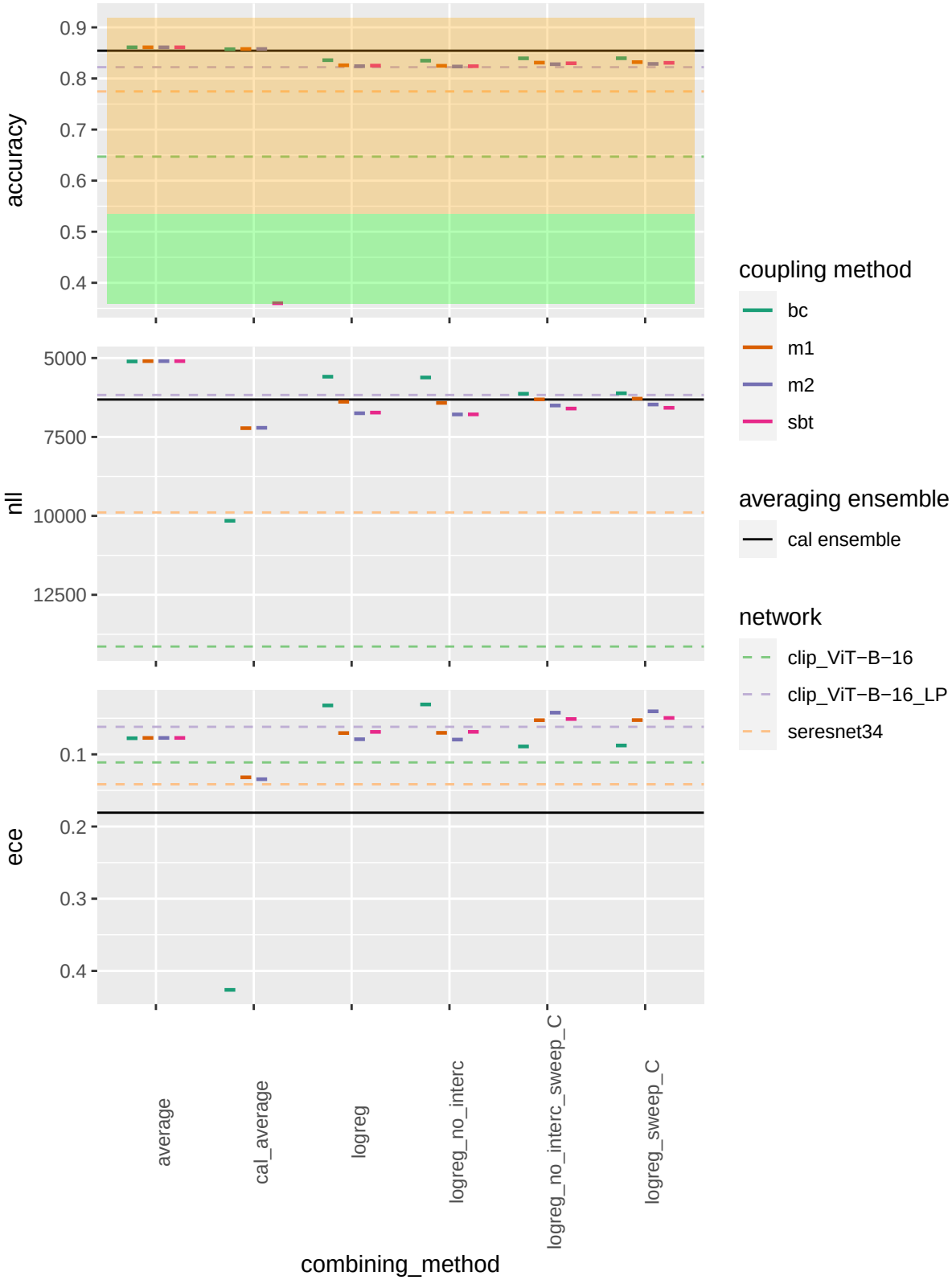
Ensemble metrics
Error inconsistency 0.276600003242493



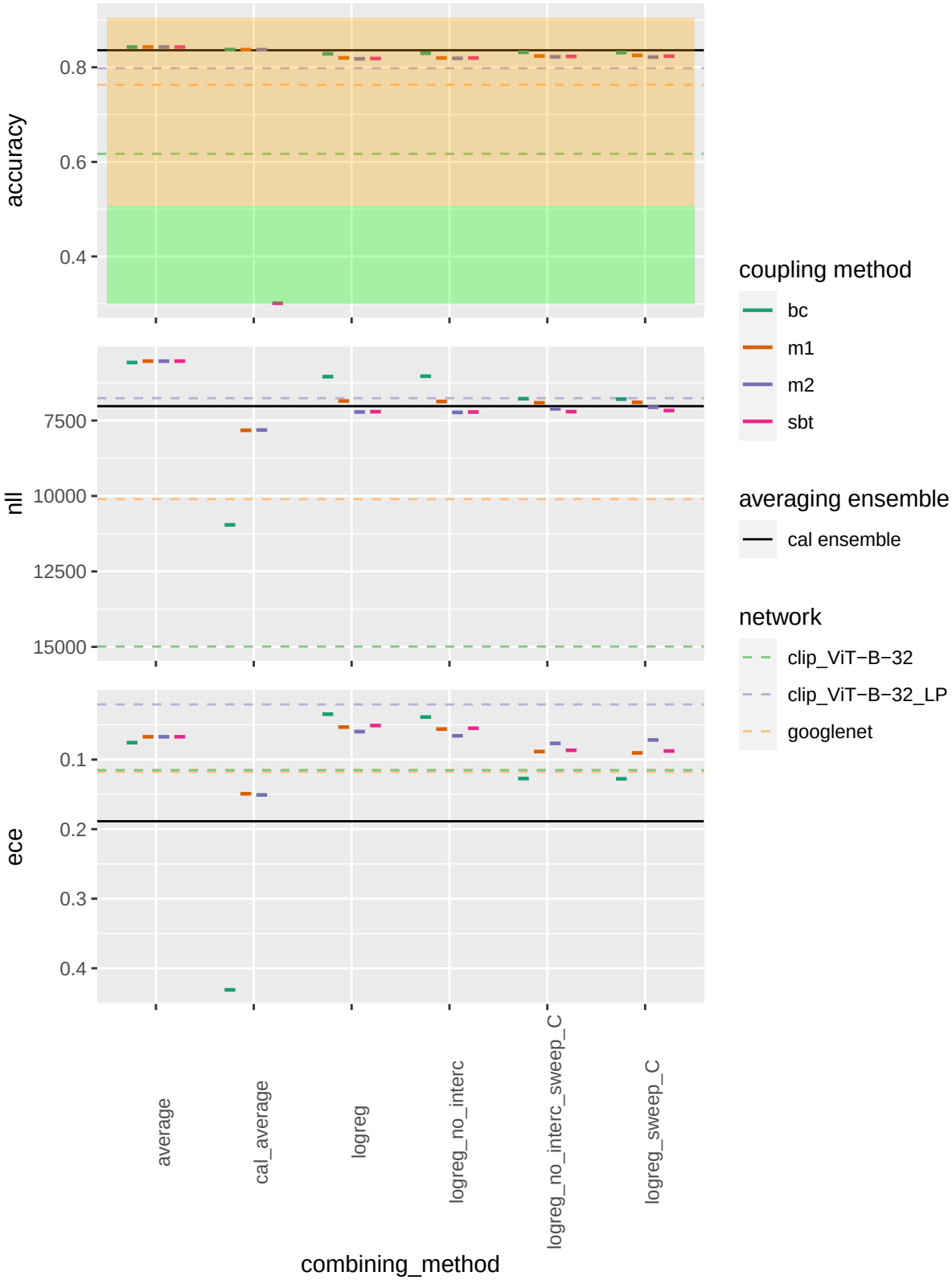
Ensemble metrics
Error inconsistency 0.28099998831749



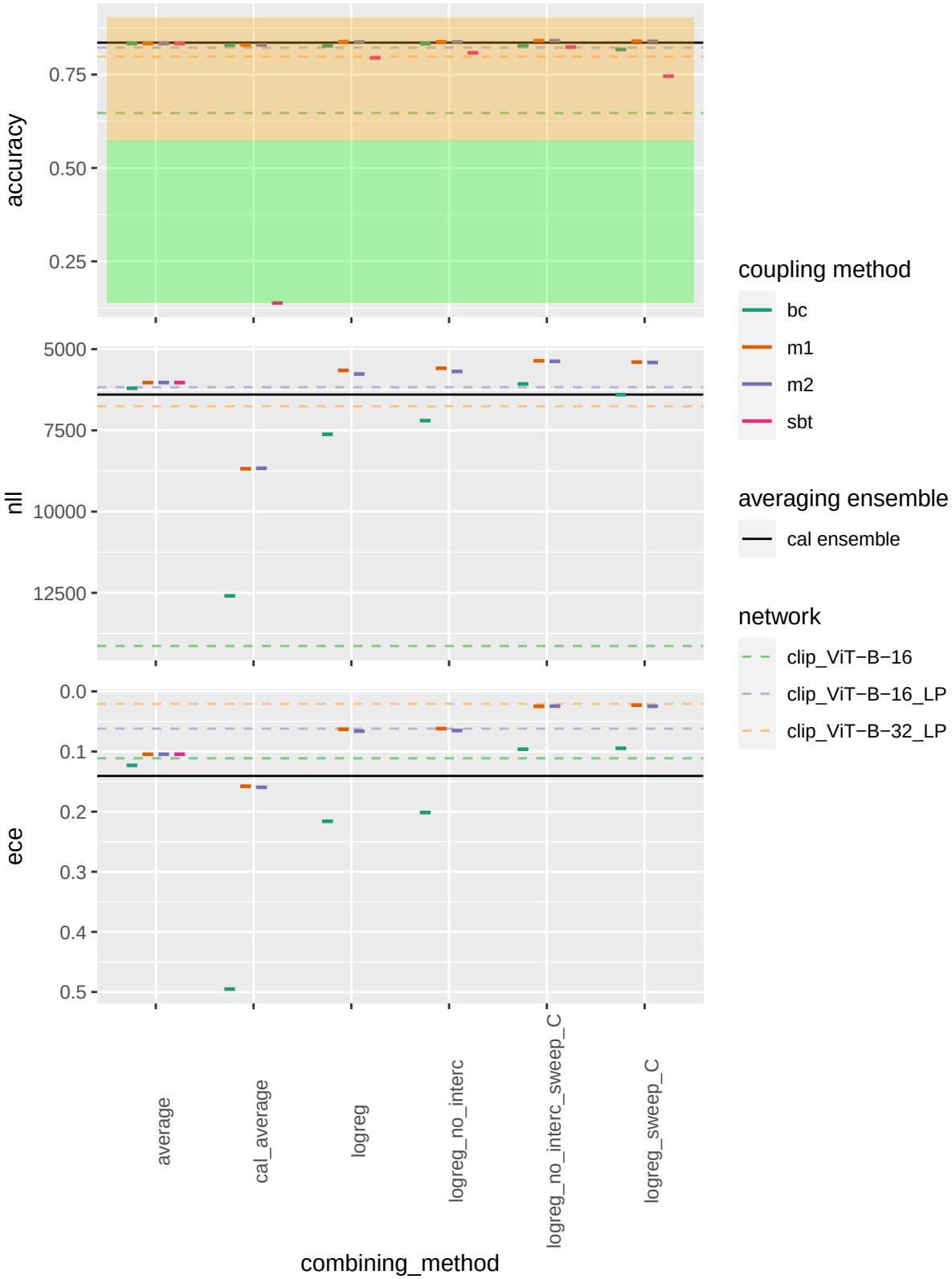
Ensemble metrics
Error inconsistency 0.38400000333786



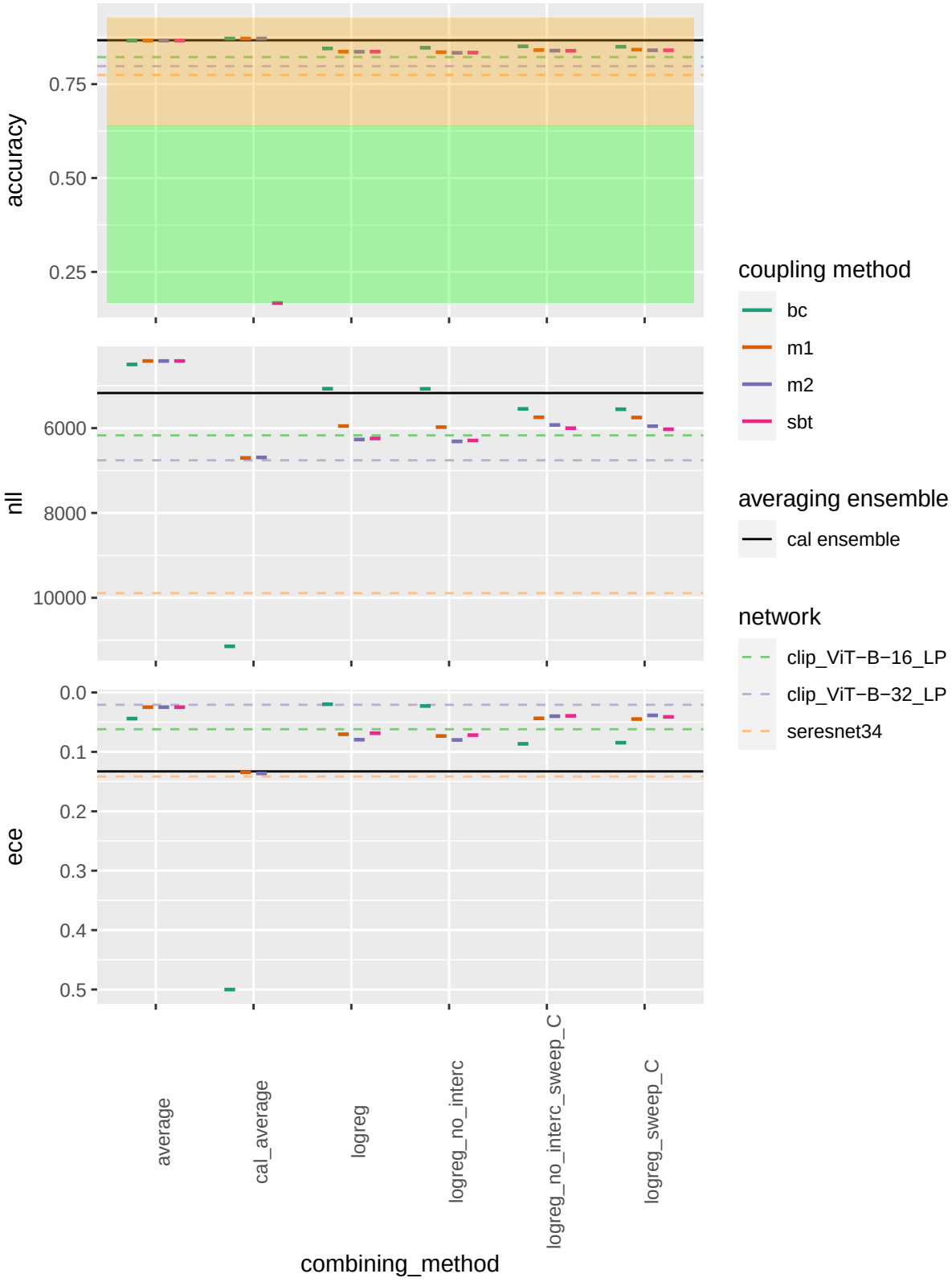
Ensemble metrics
Error inconsistency 0.397699981927872



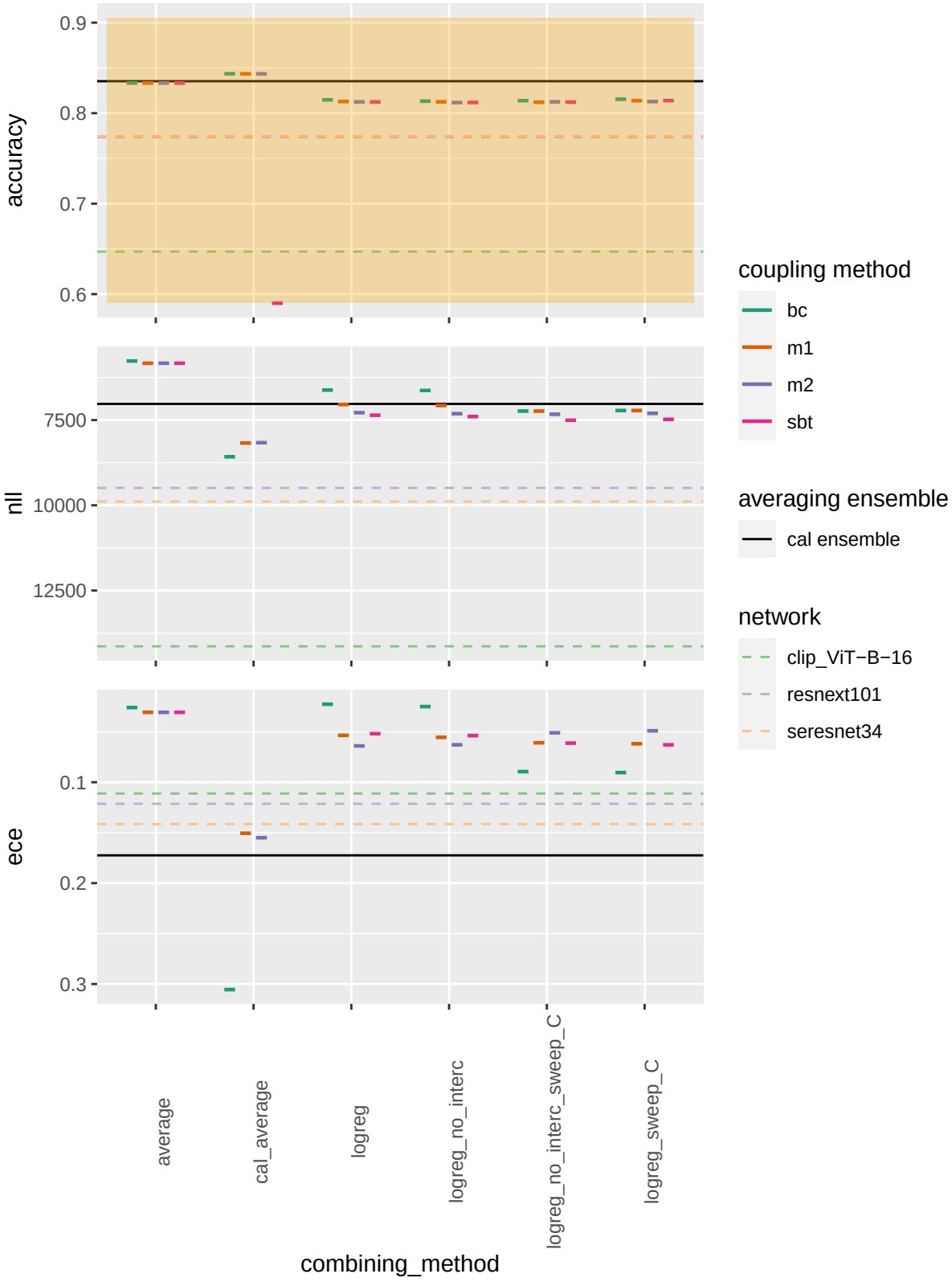
Ensemble metrics
Error inconsistency 0.330300003290176

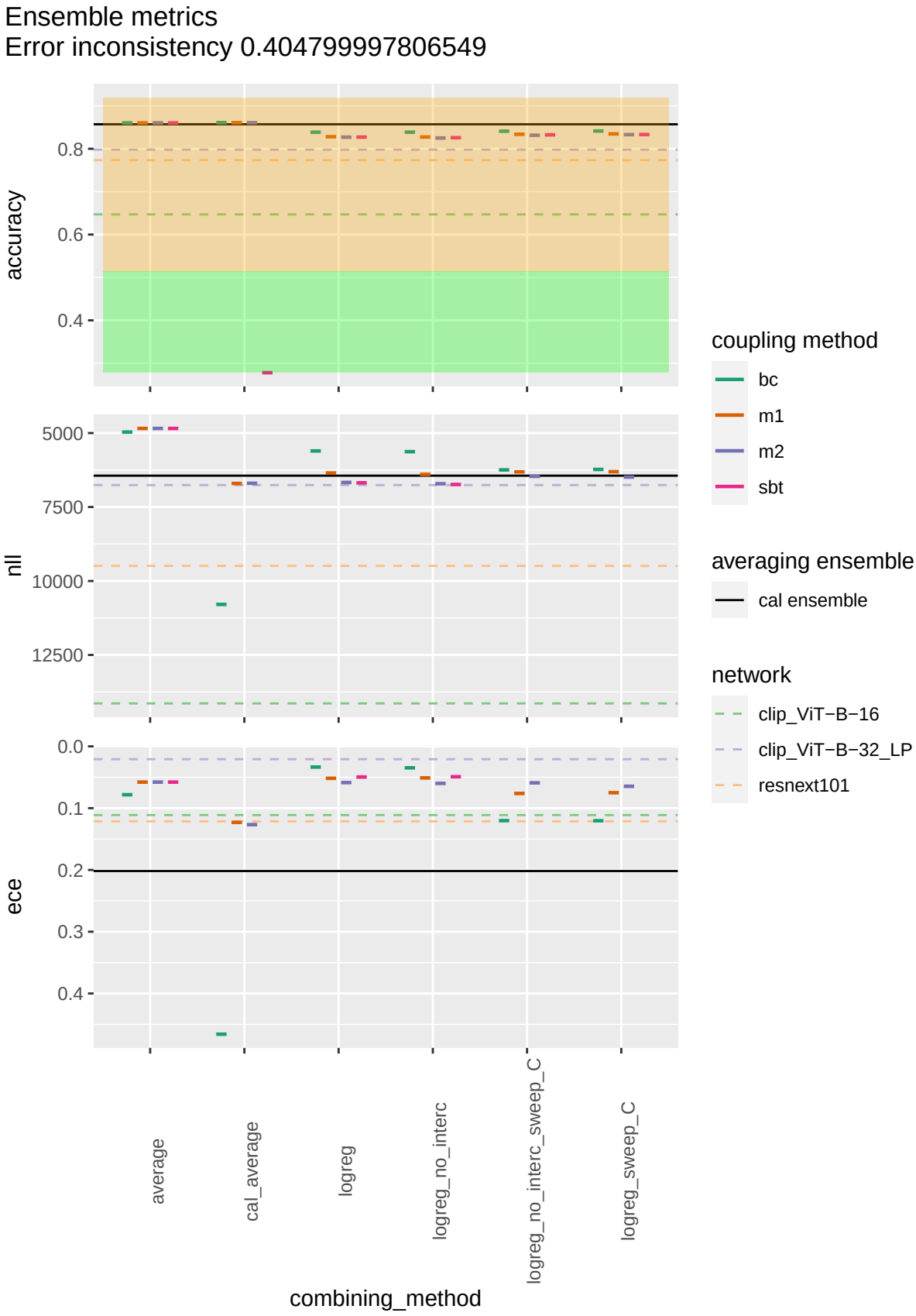


Ensemble metrics
Error inconsistency 0.287099987268448

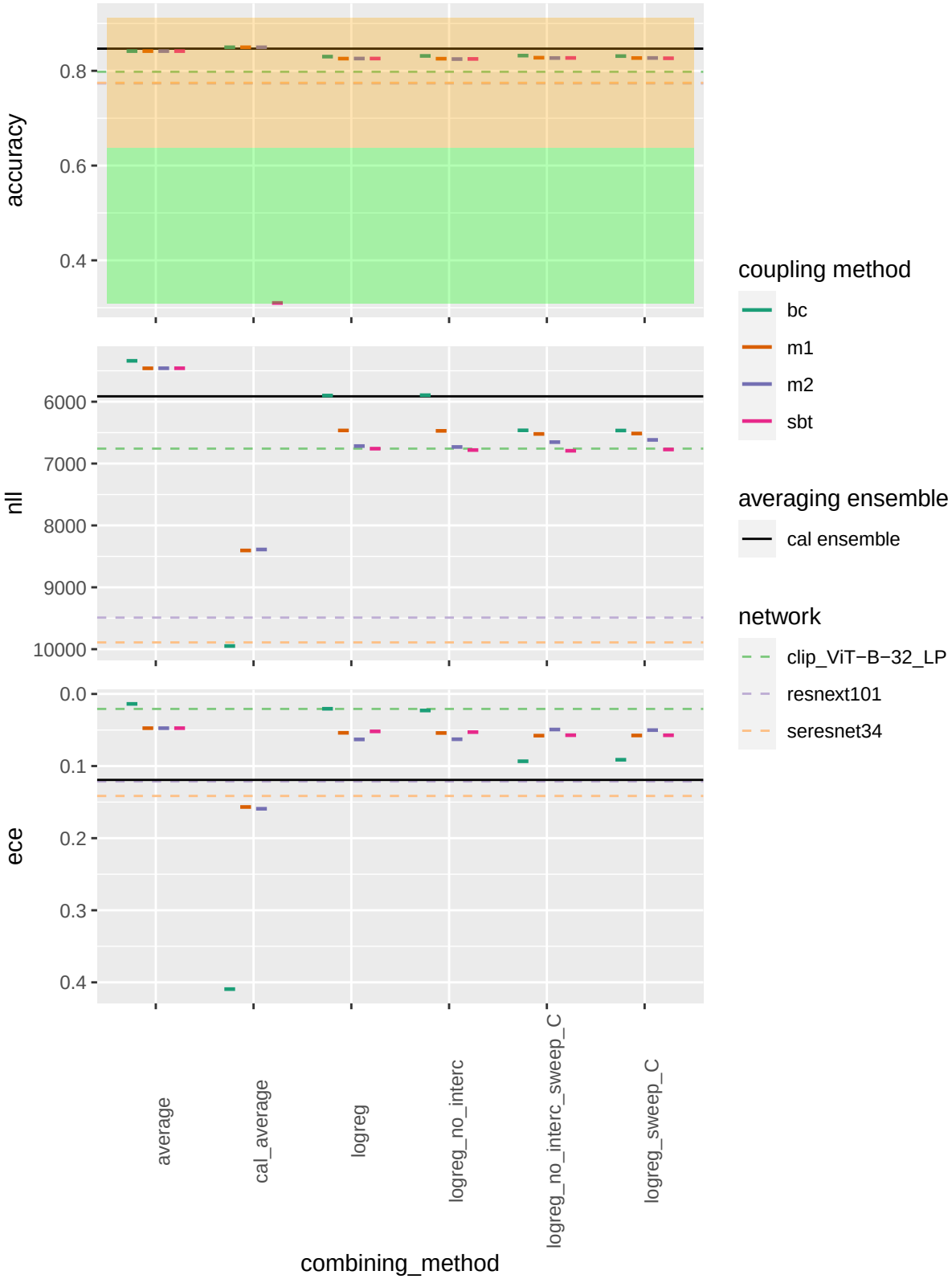


Ensemble metrics
Error inconsistency 0.393899977207184

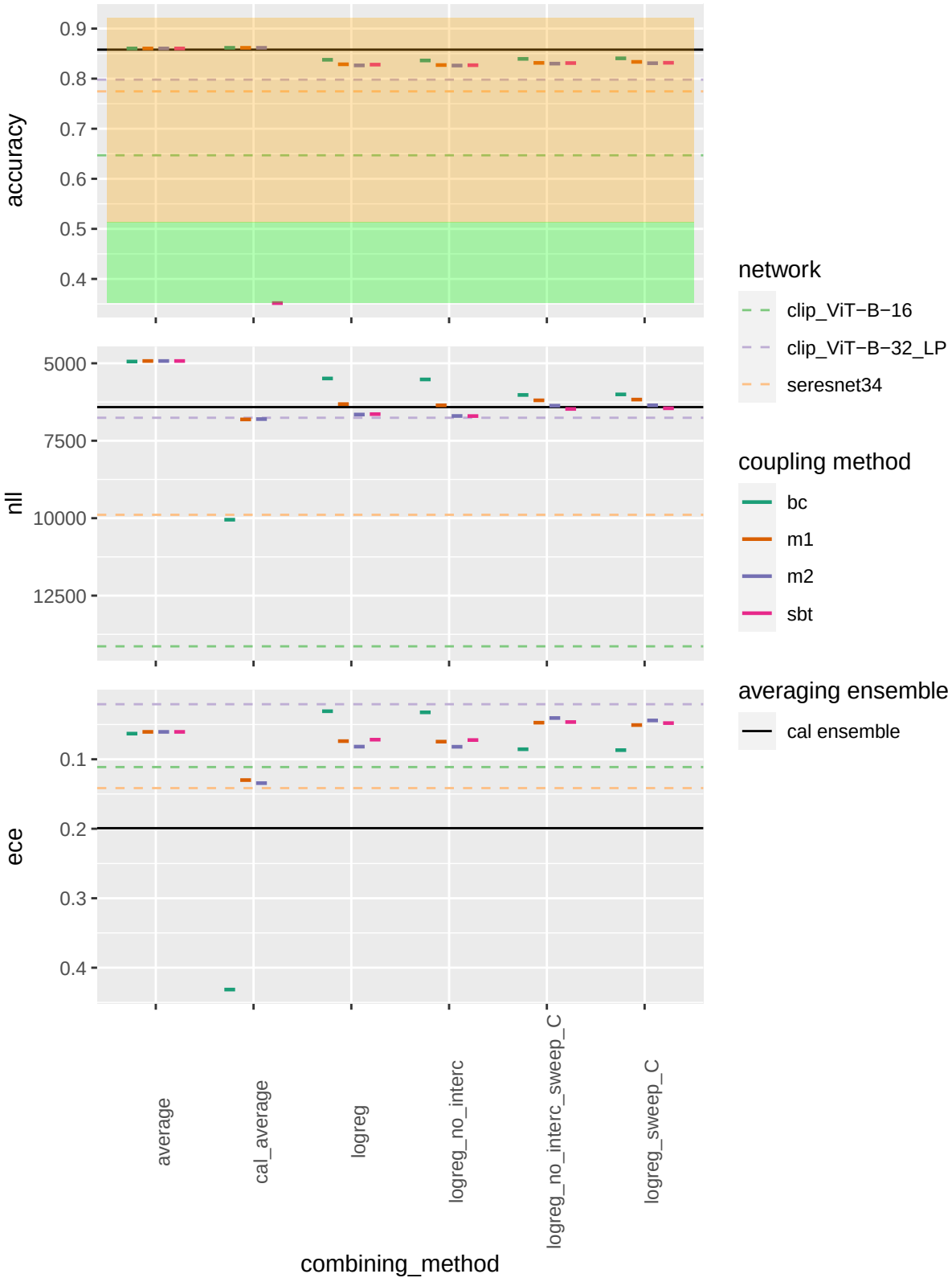



Ensemble metrics
Error inconsistency 0.404799997806549
accuracy
0.8
0.6
0.4
nll
5000
7500
10000
12500
ece
0.0
0.1
0.2
0.3
0.4
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
combining_method
coupling method
bc
m1
m2
sbt
averaging ensemble
cal ensemble
network
clip_ViT−B−16
clip_ViT−B−32_LP
resnext101

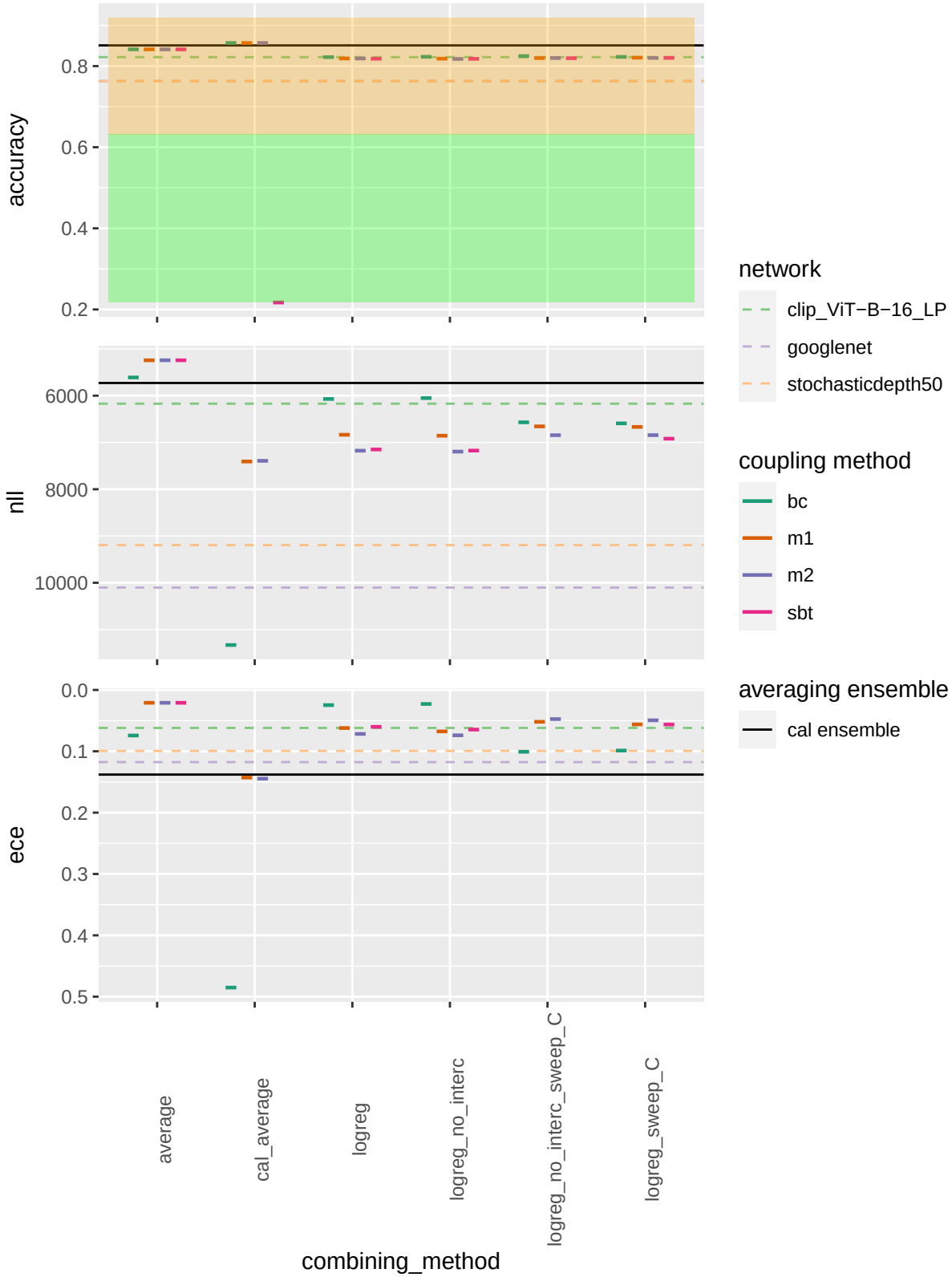
Ensemble metrics
Error inconsistency 0.274399995803833



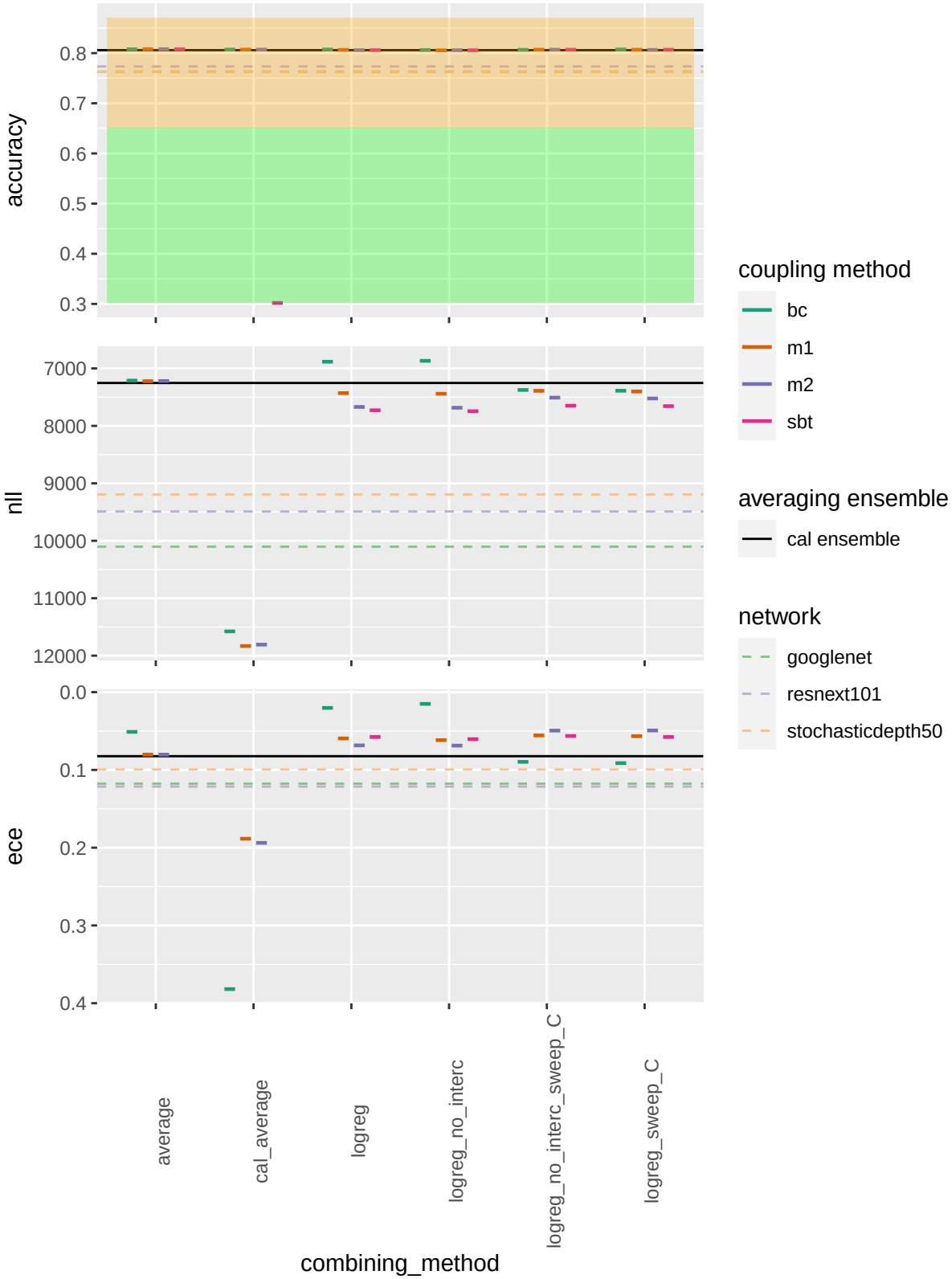
Ensemble metrics
Error inconsistency 0.408899992704391



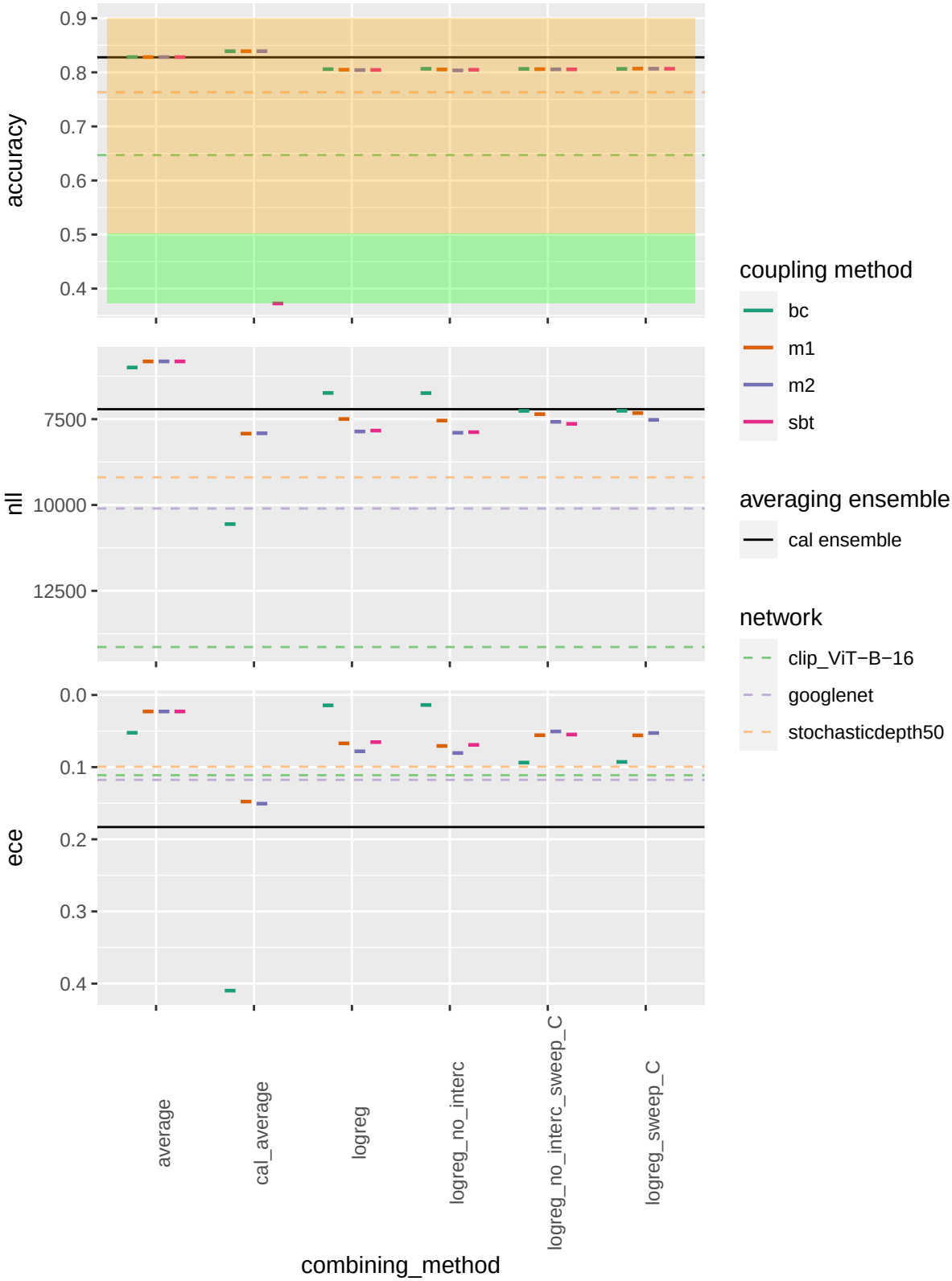
Ensemble metrics
Error inconsistency 0.286799997091293



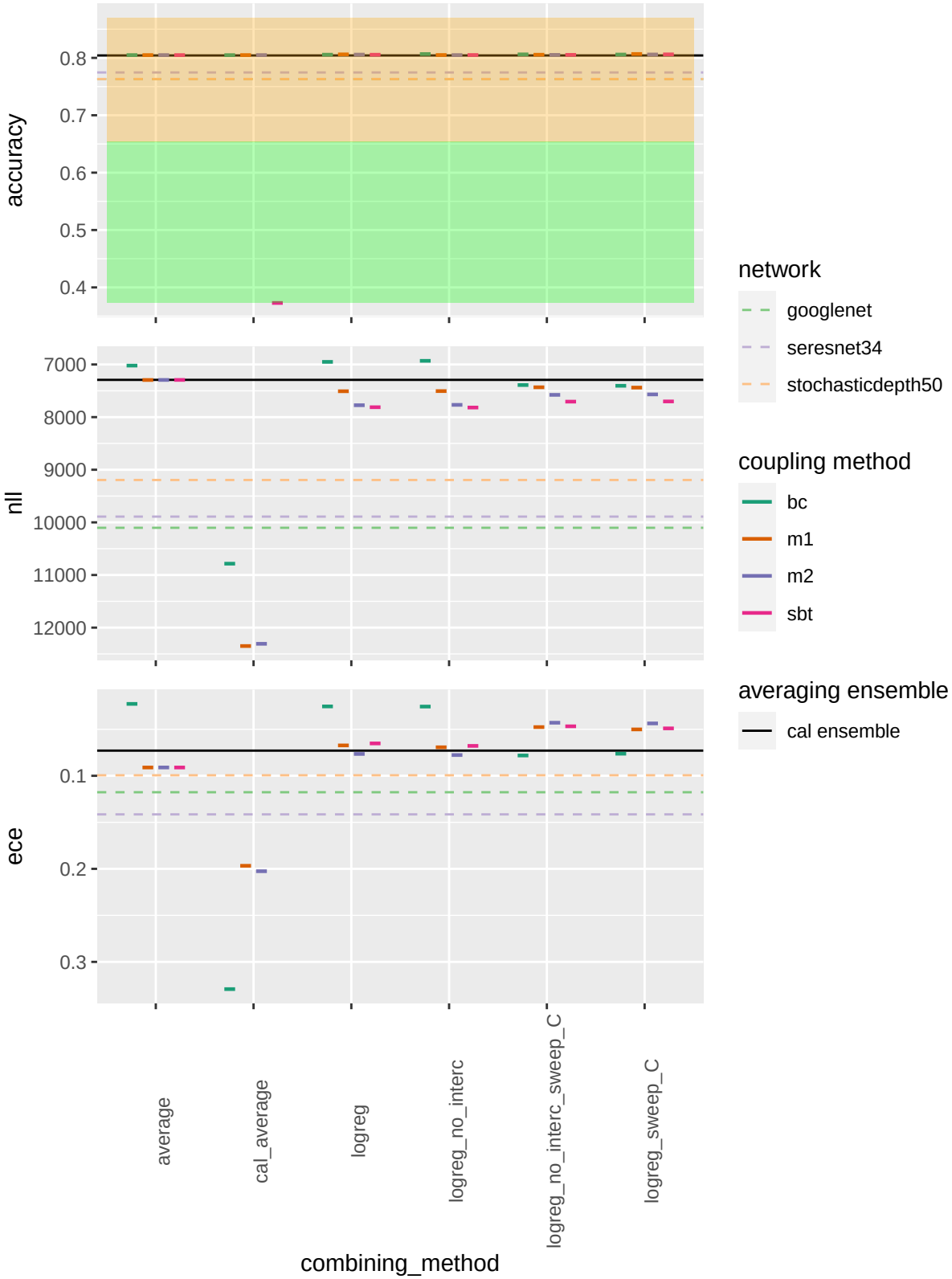
Ensemble metrics
Error inconsistency 0.220100000500679



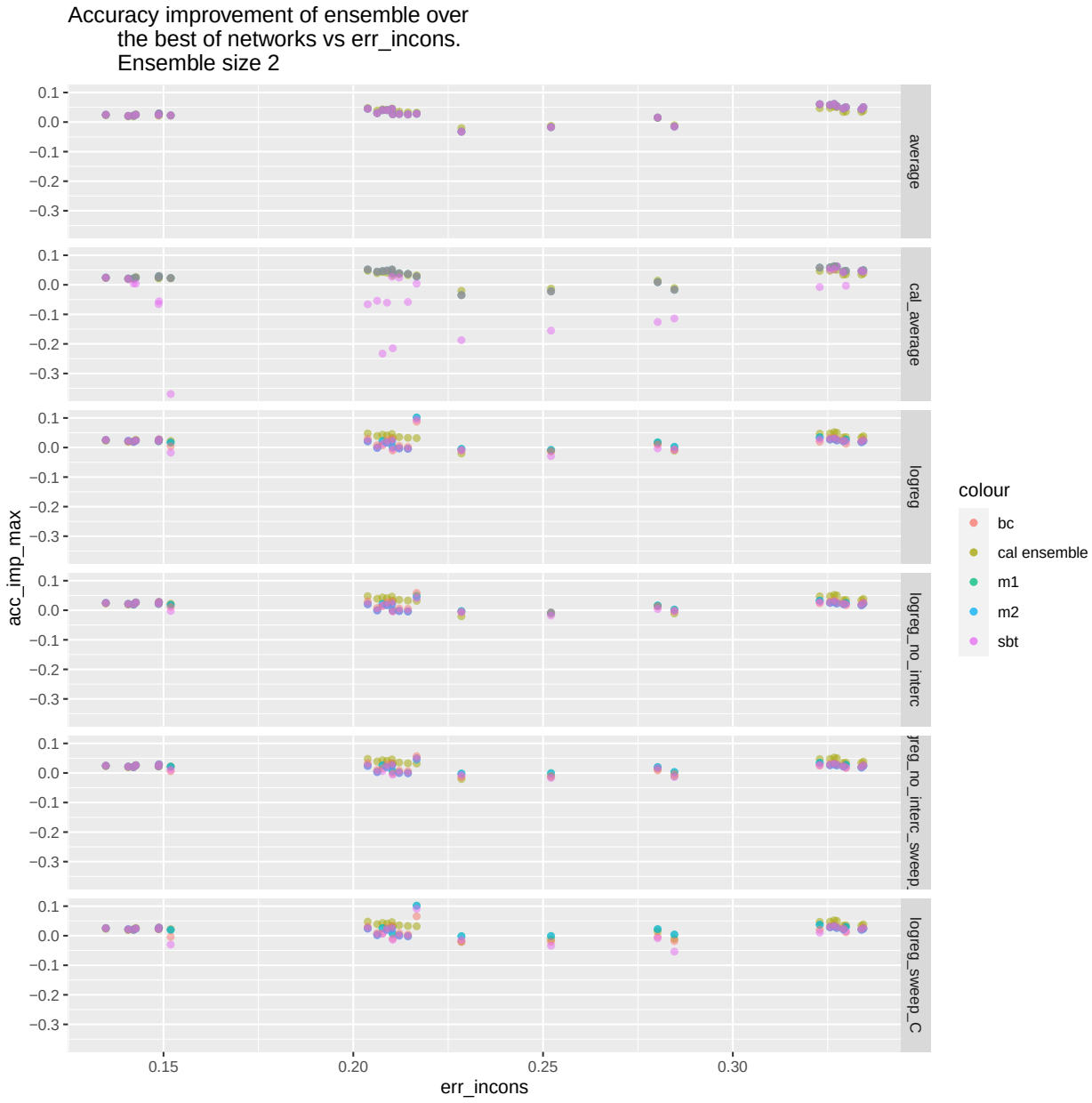
Ensemble metrics
Error inconsistency 0.399199992418289

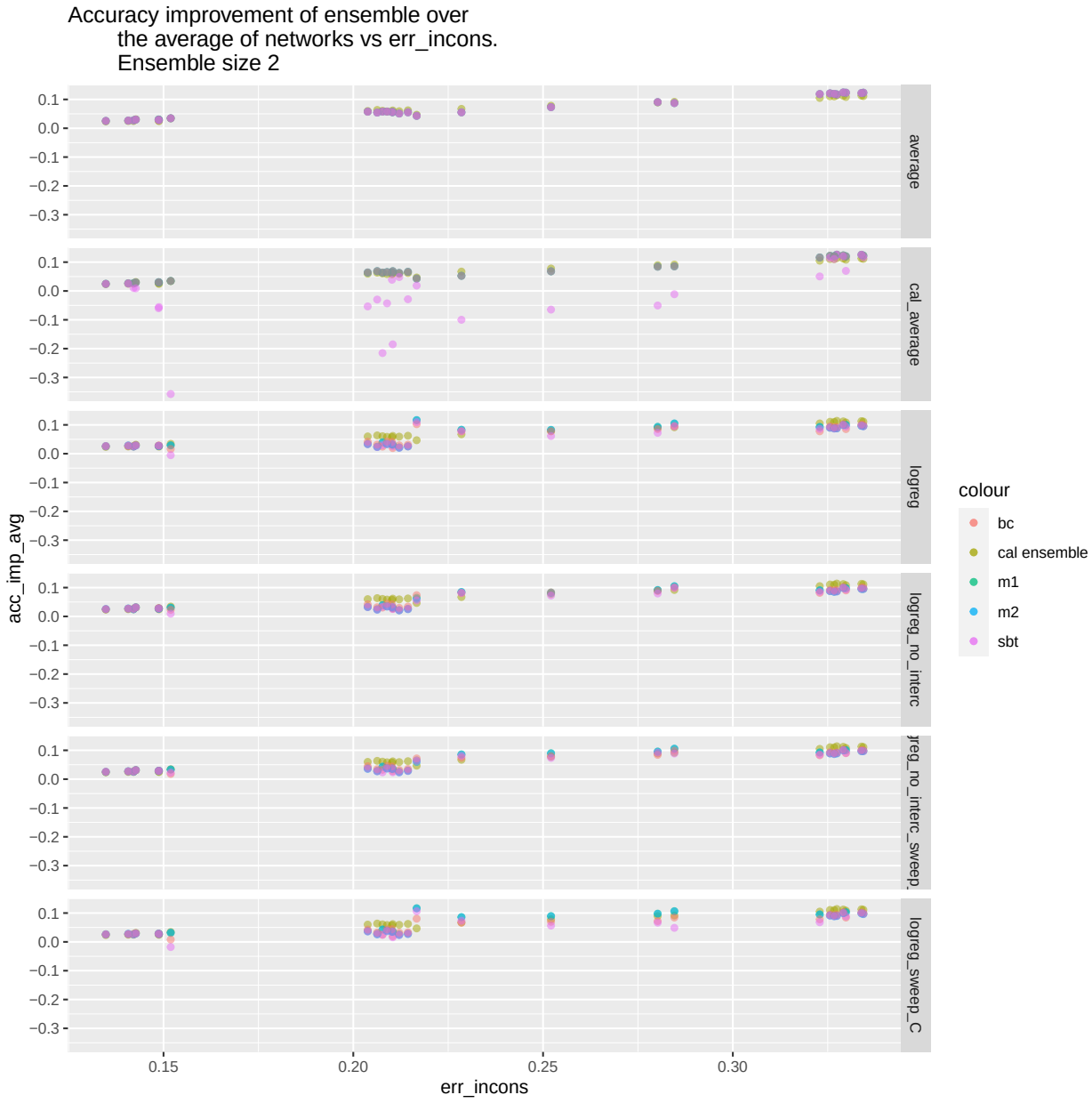


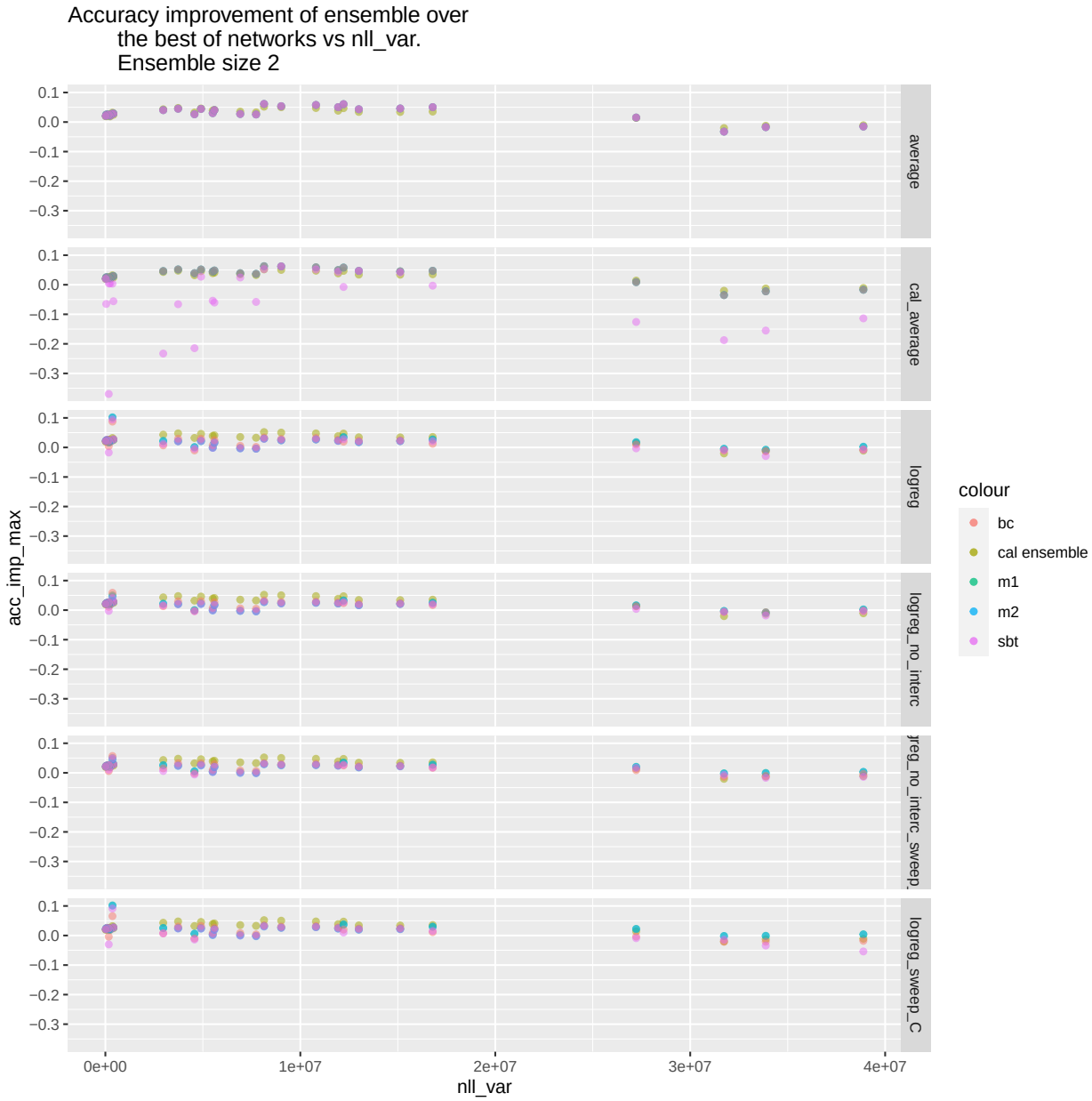
Ensemble metrics
Error inconsistency 0.215799987316132

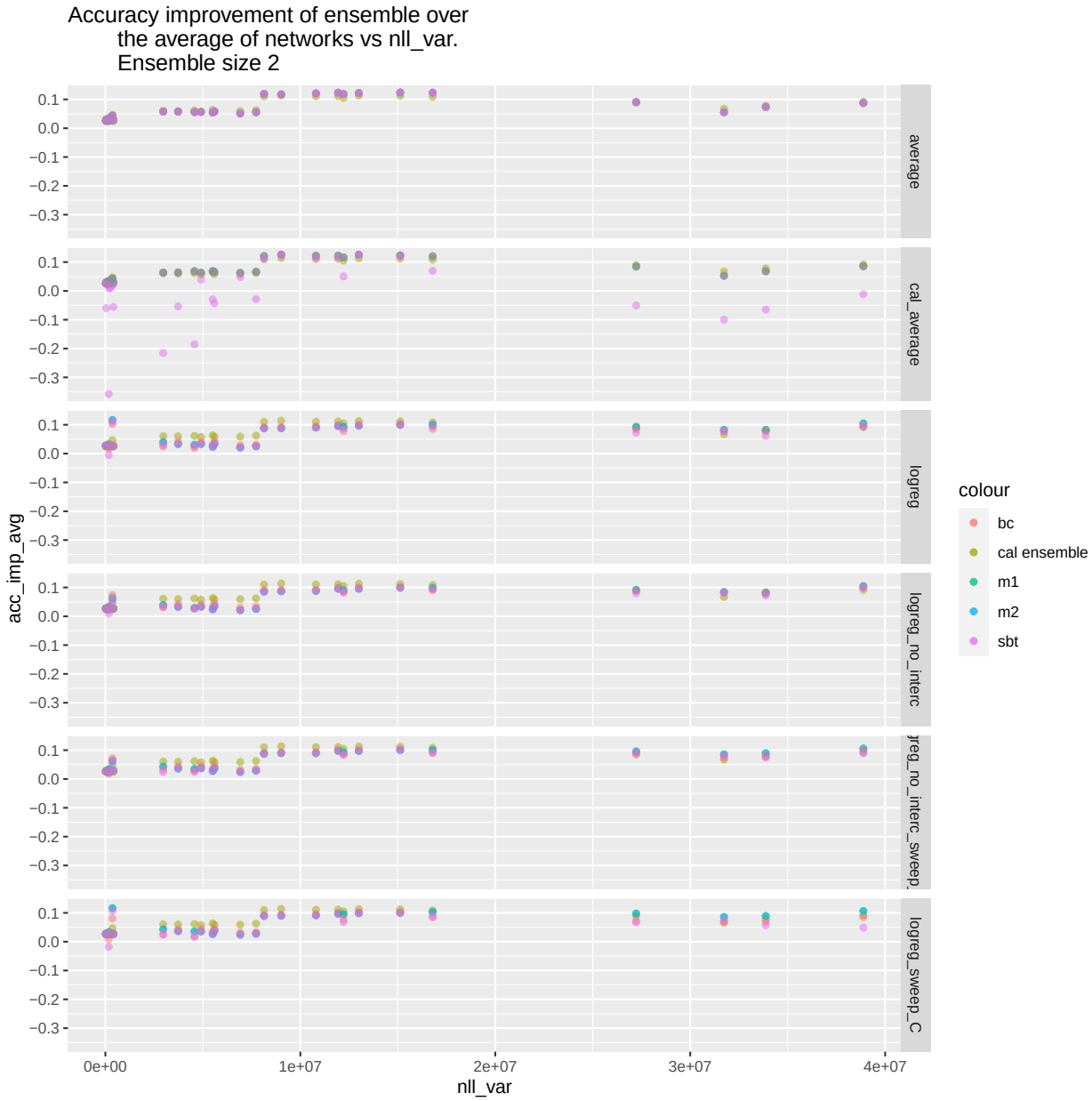


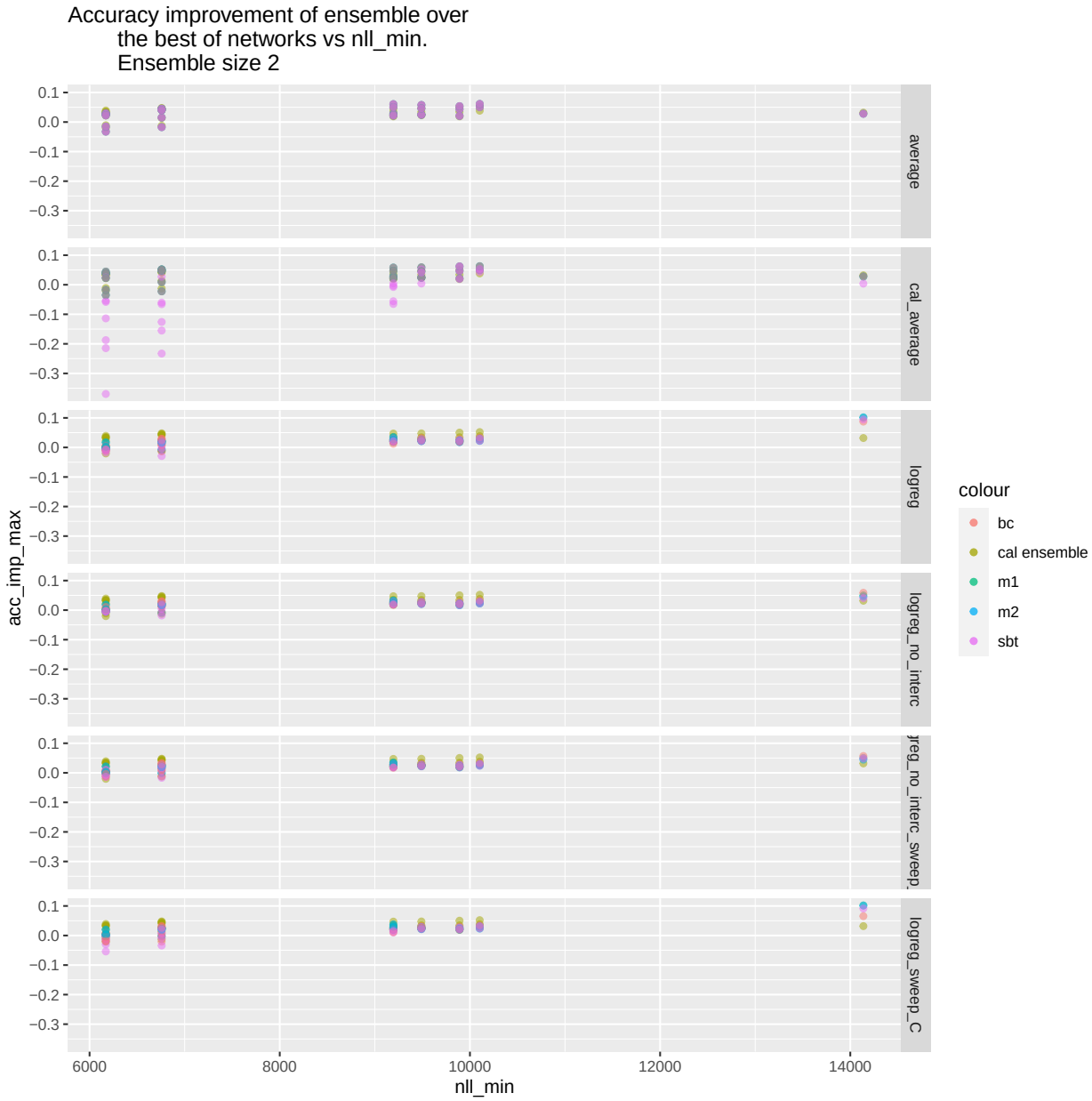
```
xax <- c("err_incons",
  "nll_var", "nll_min", "nll_max", "nll_avg",
  "acc_var", "acc_min", "acc_max", "acc_avg",
  "ece_var", "ece_min", "ece_max", "ece_avg")
yax <- c("acc_imp_max", "acc_imp_avg")
for (sss in unique(ens_cal_plt_df$combination_size))
{
  cur_ens_cal_plt_df <- ens_cal_plt_df %>% filter(combination_size == sss)
  cur_ens_pwc_plt_df <- ens_pwc_plt_df %>% filter(combination_size == sss)
  for (xa in xax)
  {
    for (ya in yax)
    {
      cur_plot <- ggplot() +
        geom_point(
          data=cur_ens_cal_plt_df,
          mapping=aes_string(x=xa, y=ya, color=shQuote("cal ensemble")),
          alpha=0.5) +
        geom_point(
          data=cur_ens_pwc_plt_df,
          mapping=aes_string(x=xa, y=ya, color="coupling_method"),
          alpha=0.5) +
        facet_grid(rows=vars(combining_method)) +
        ggtitle(sprintf("Accuracy improvement of ensemble over
        the %s of networks vs %s.
        Ensemble size %s",
        if (ya == "acc_imp_max") "best" else "average", xa, sss))

      print(cur_plot)
    }
  }
}
```

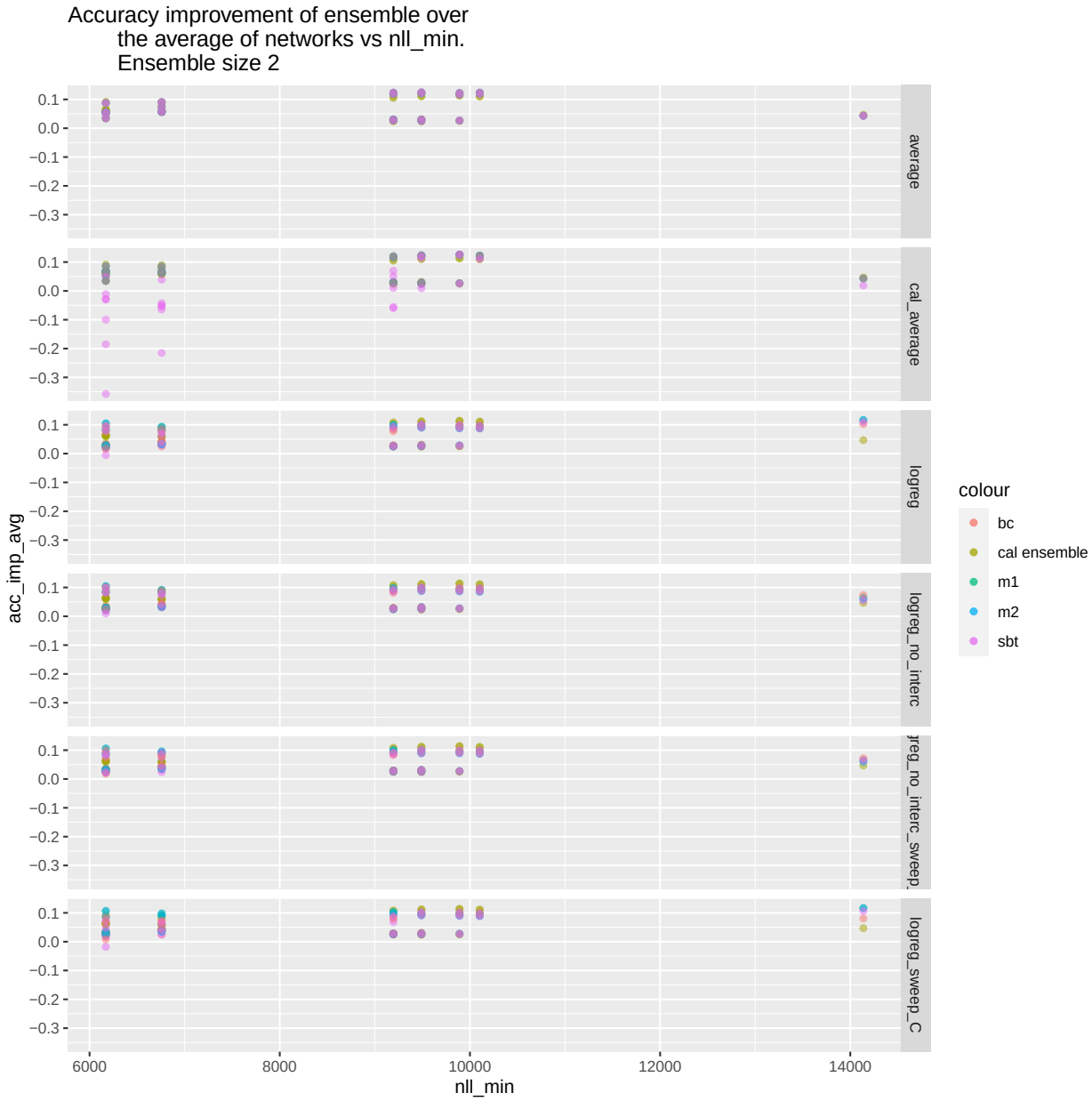

Accuracy improvement of ensemble over
the best of networks vs err_incons.
Ensemble size 2
acc_imp_max
err_incons
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

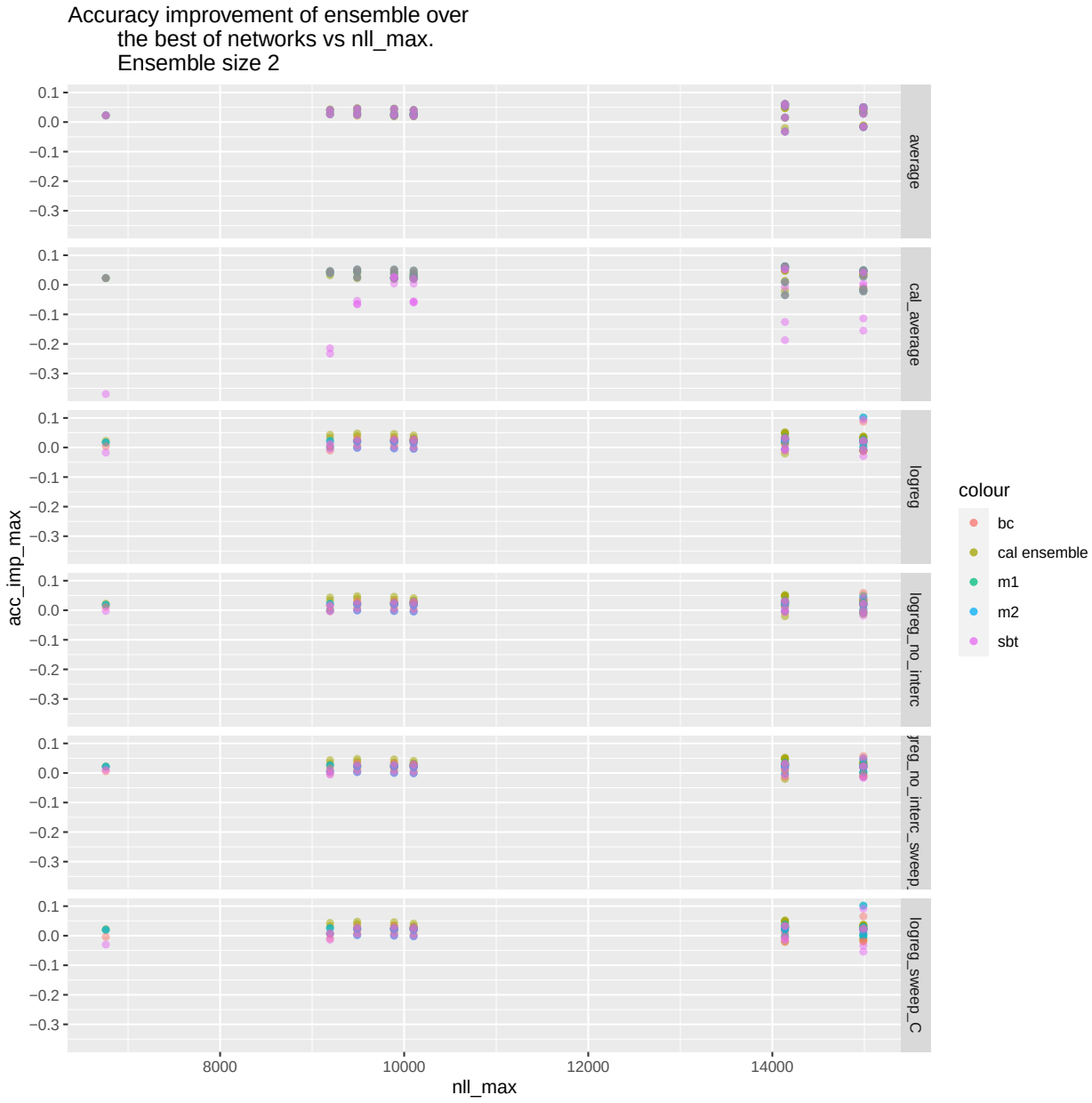

Accuracy improvement of ensemble over
the average of networks vs err_incons.
Ensemble size 2
acc_imp_avg
err_incons
0.15
0.20
0.25
0.30
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

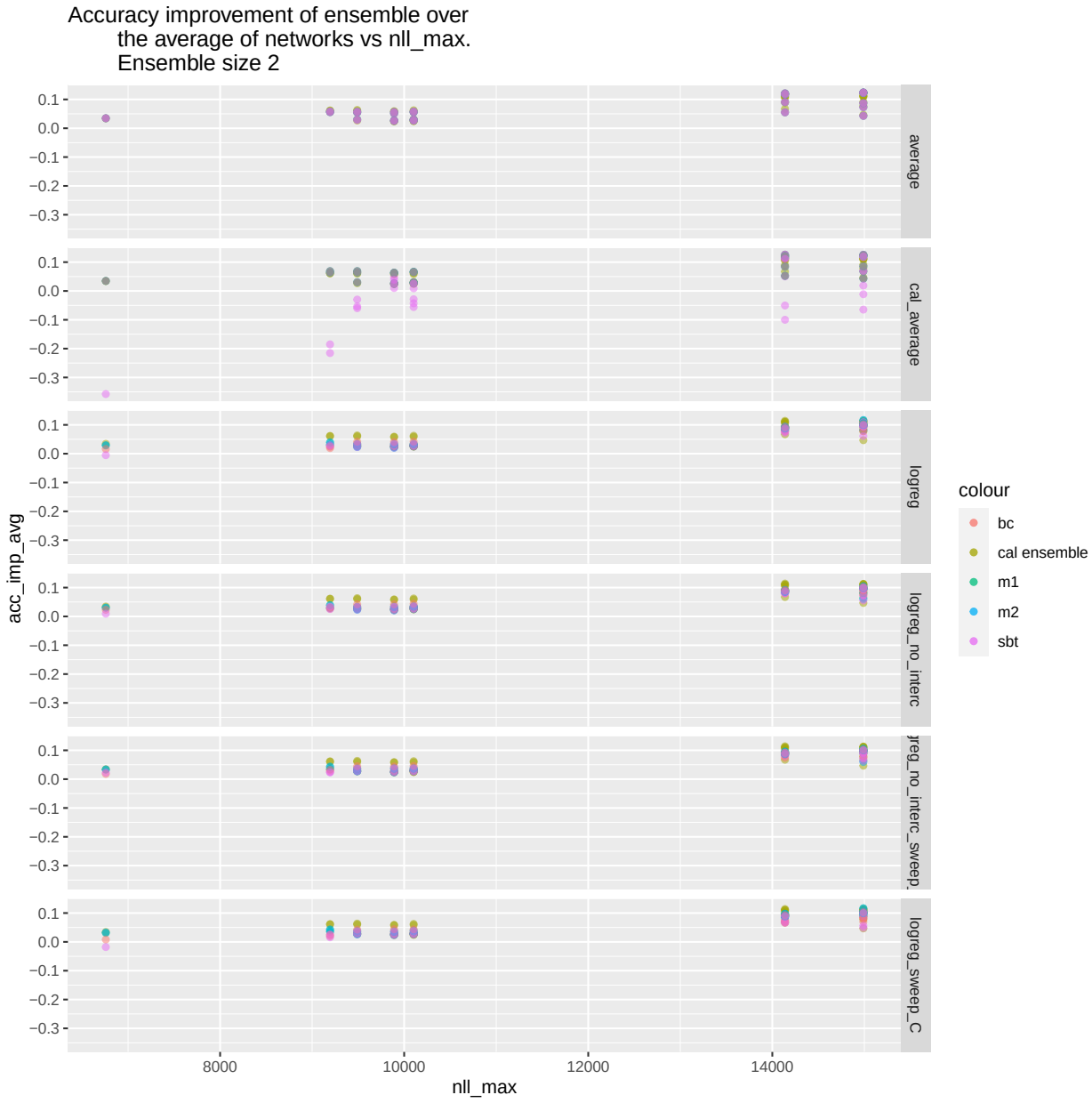

Accuracy improvement of ensemble over
the best of networks vs nll_var.
Ensemble size 2
acc_imp_max
nll_var
0e+00
1e+07
2e+07
3e+07
4e+07
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

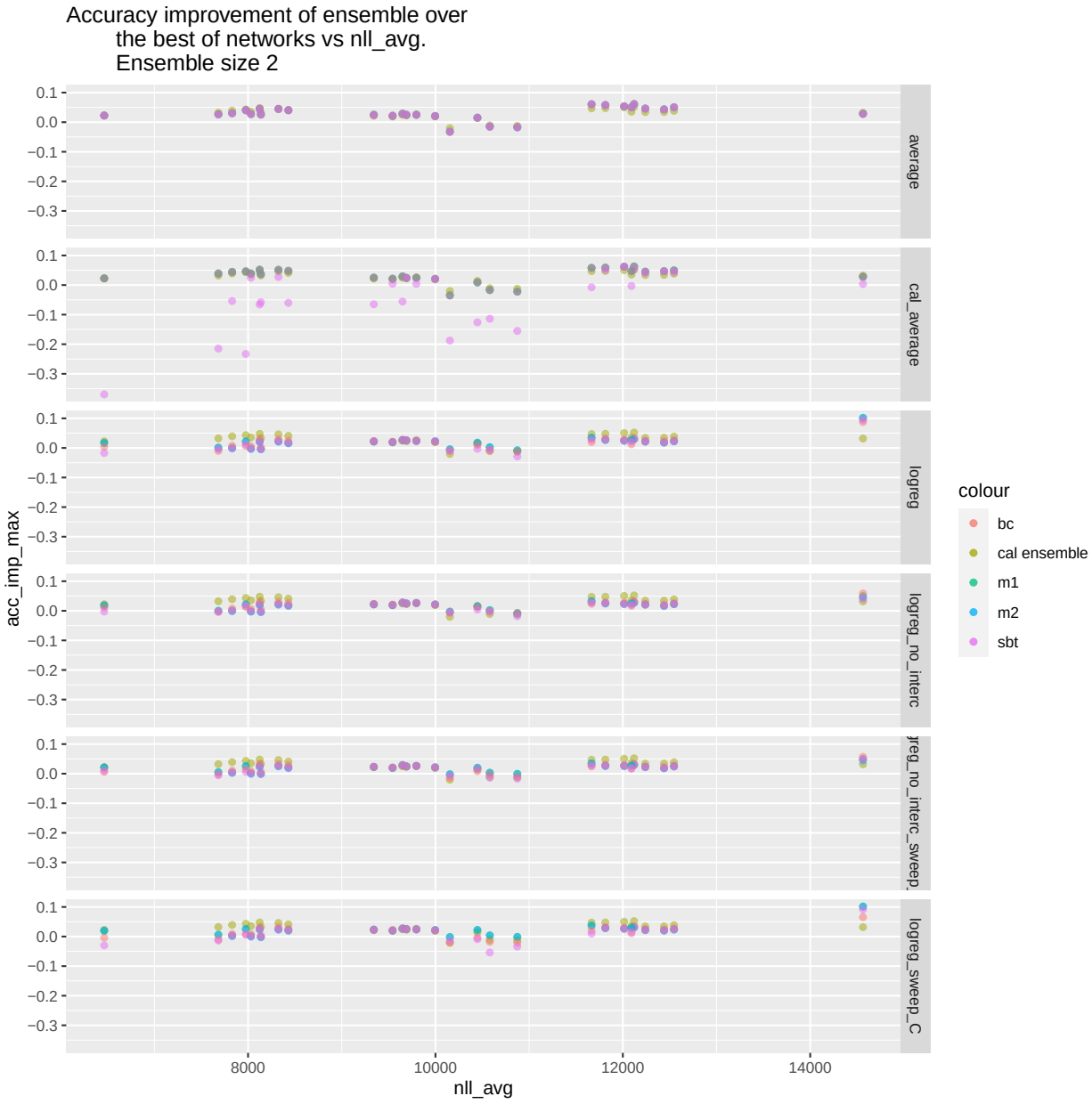

Accuracy improvement of ensemble over the average of networks vs nll_var. Ensemble size 2
acc_imp_avg
nll_var
0e+00
1e+07
2e+07
3e+07
4e+07
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

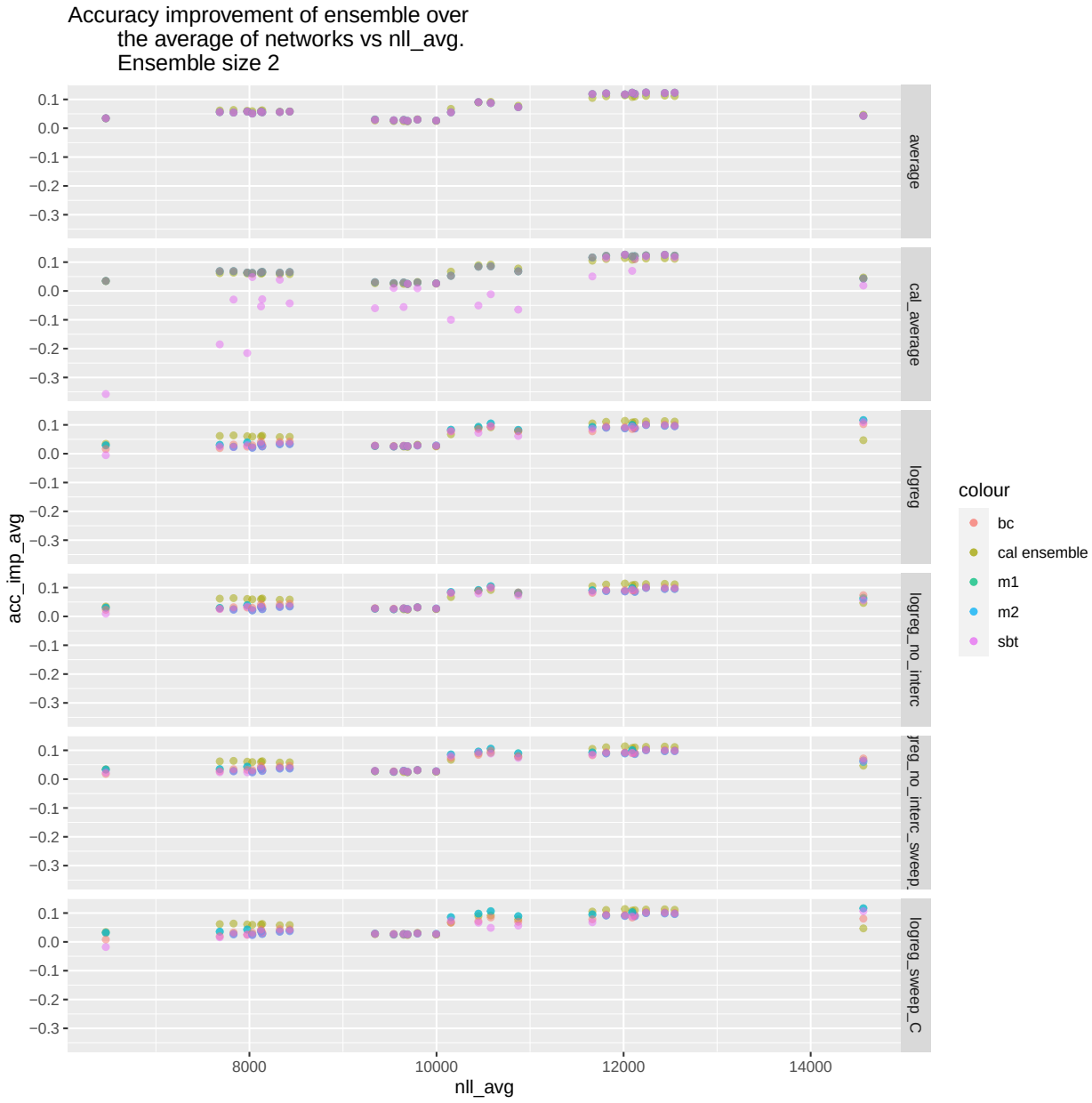

Accuracy improvement of ensemble over
the best of networks vs nll_min.
Ensemble size 2
acc_imp_max
nll_min
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

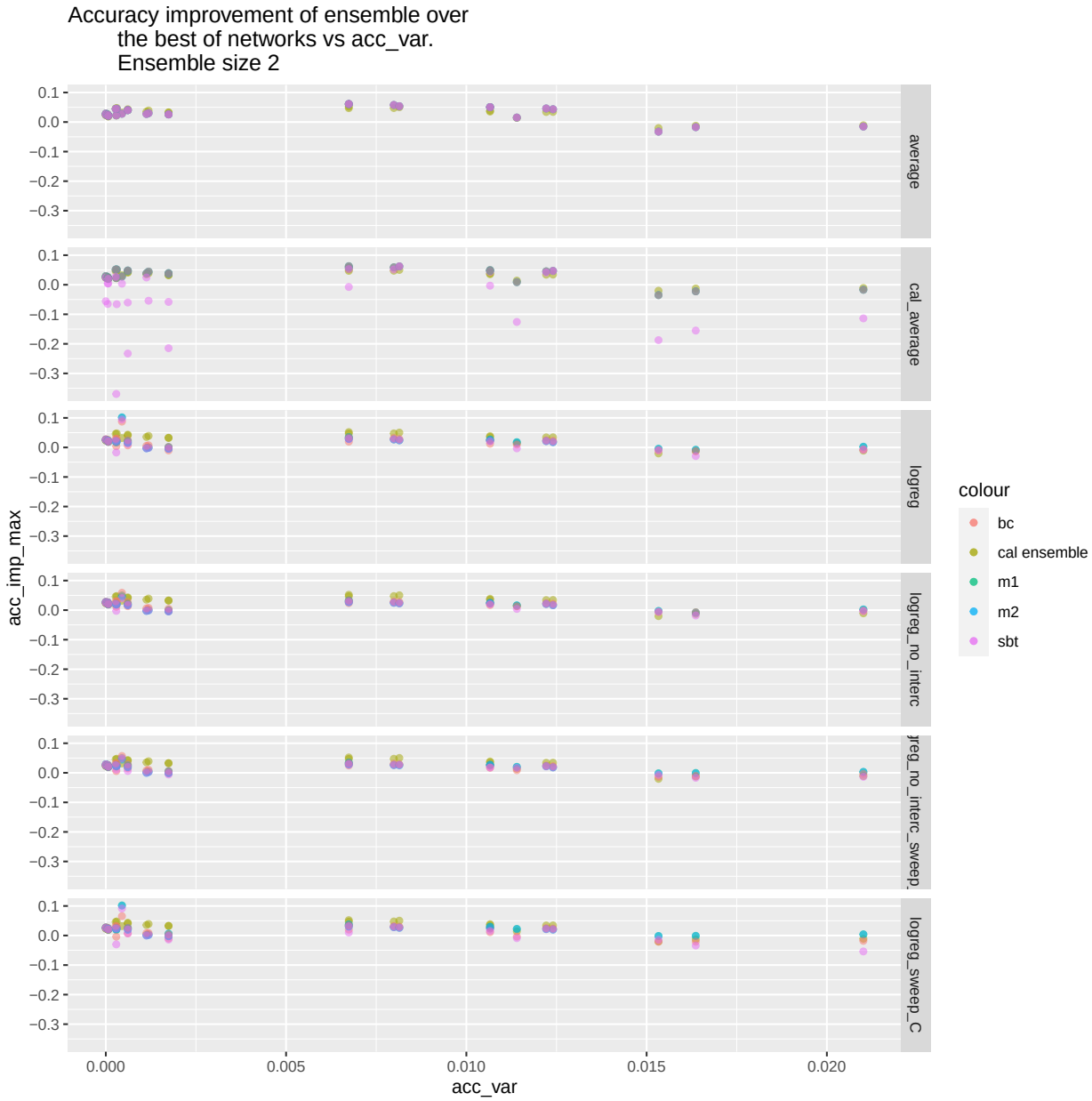

Accuracy improvement of ensemble over
the average of networks vs nll_min.
Ensemble size 2
acc_imp_avg
nll_min
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

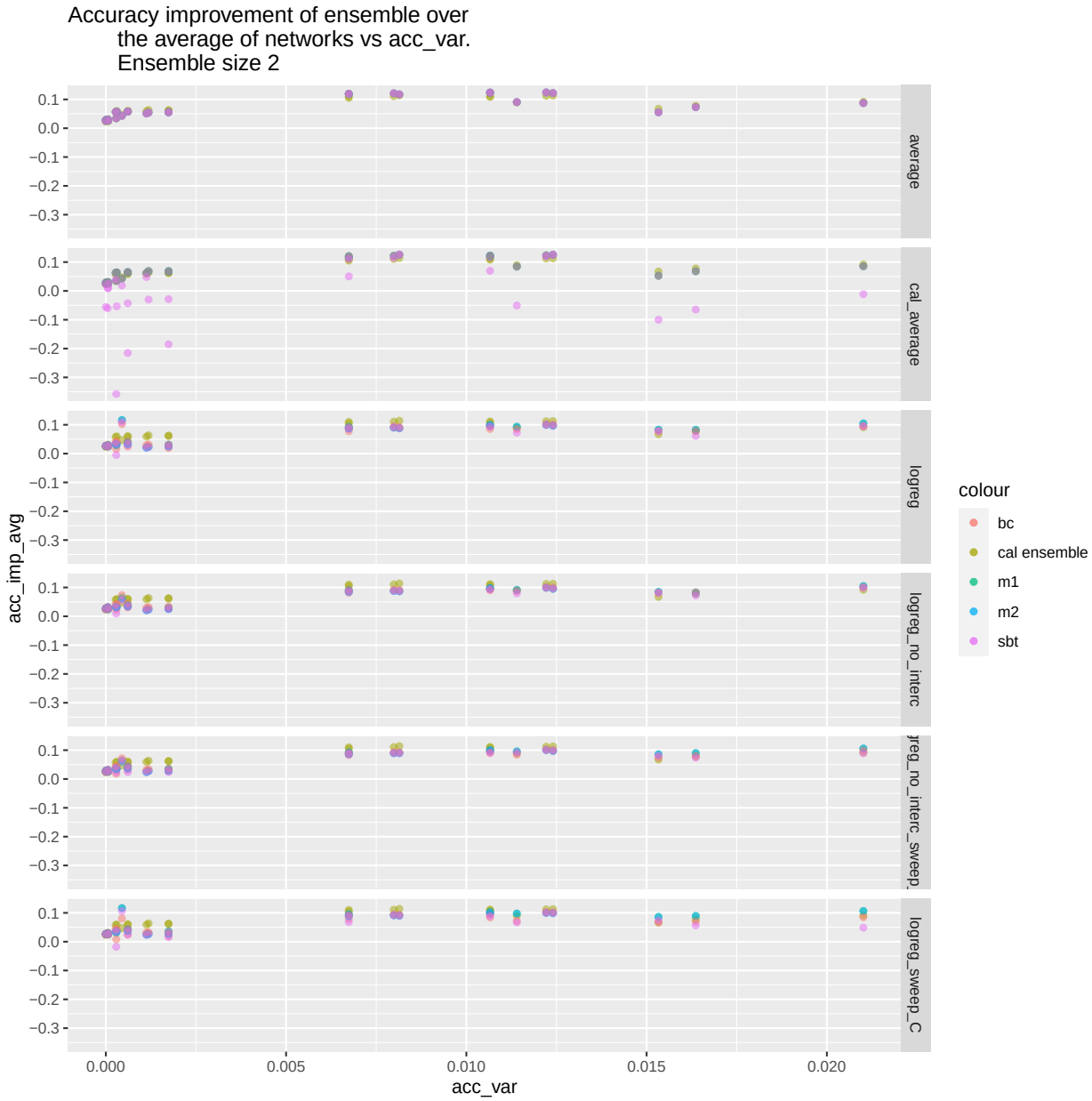

Accuracy improvement of ensemble over
the best of networks vs nll_max.
Ensemble size 2
acc_imp_max
nll_max
0.1
0.0
−0.1
−0.2
−0.3
8000
10000
12000
14000
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

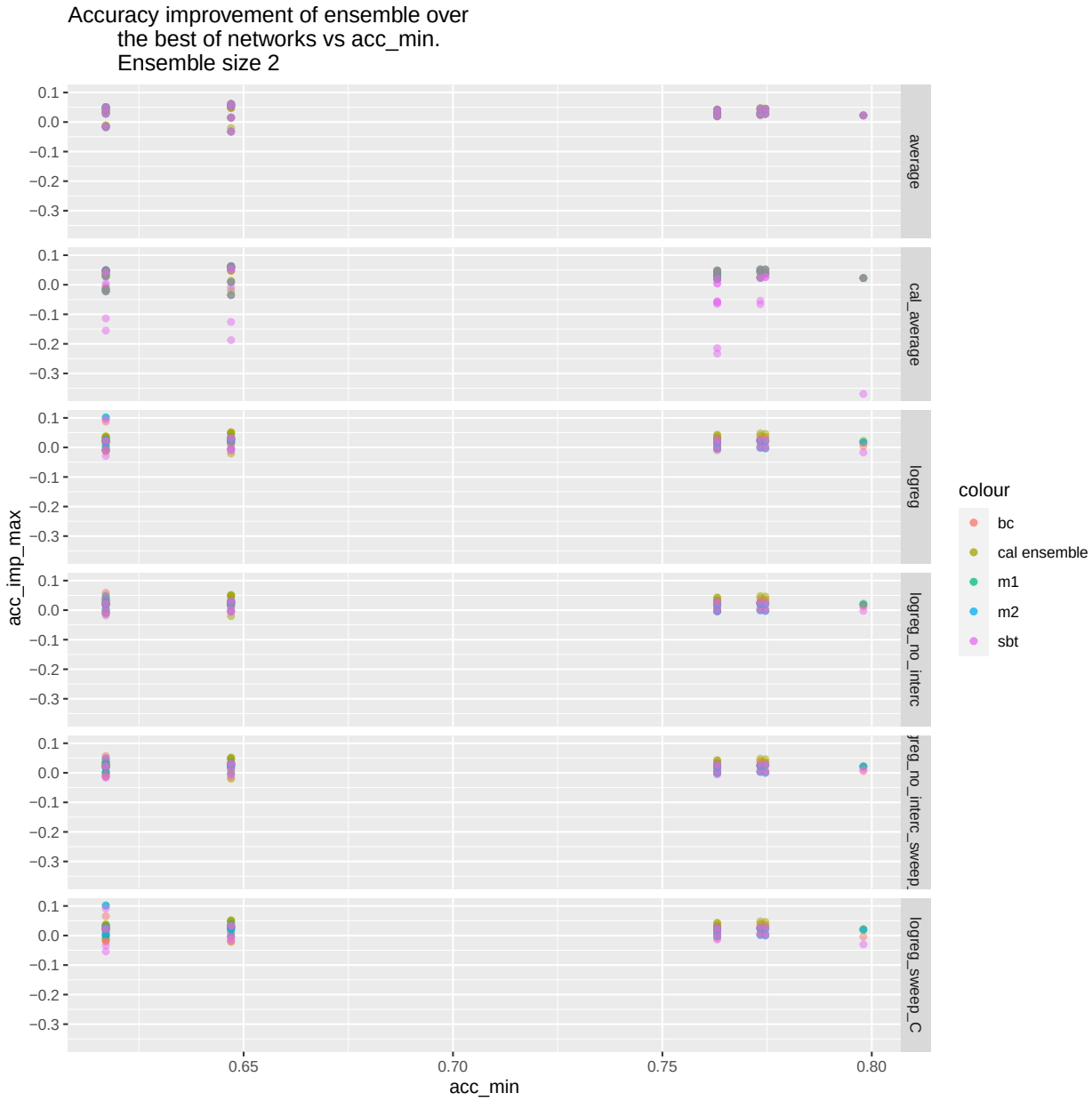

Accuracy improvement of ensemble over
the average of networks vs nll_max.
Ensemble size 2
acc_imp_avg
nll_max
8000
10000
12000
14000
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

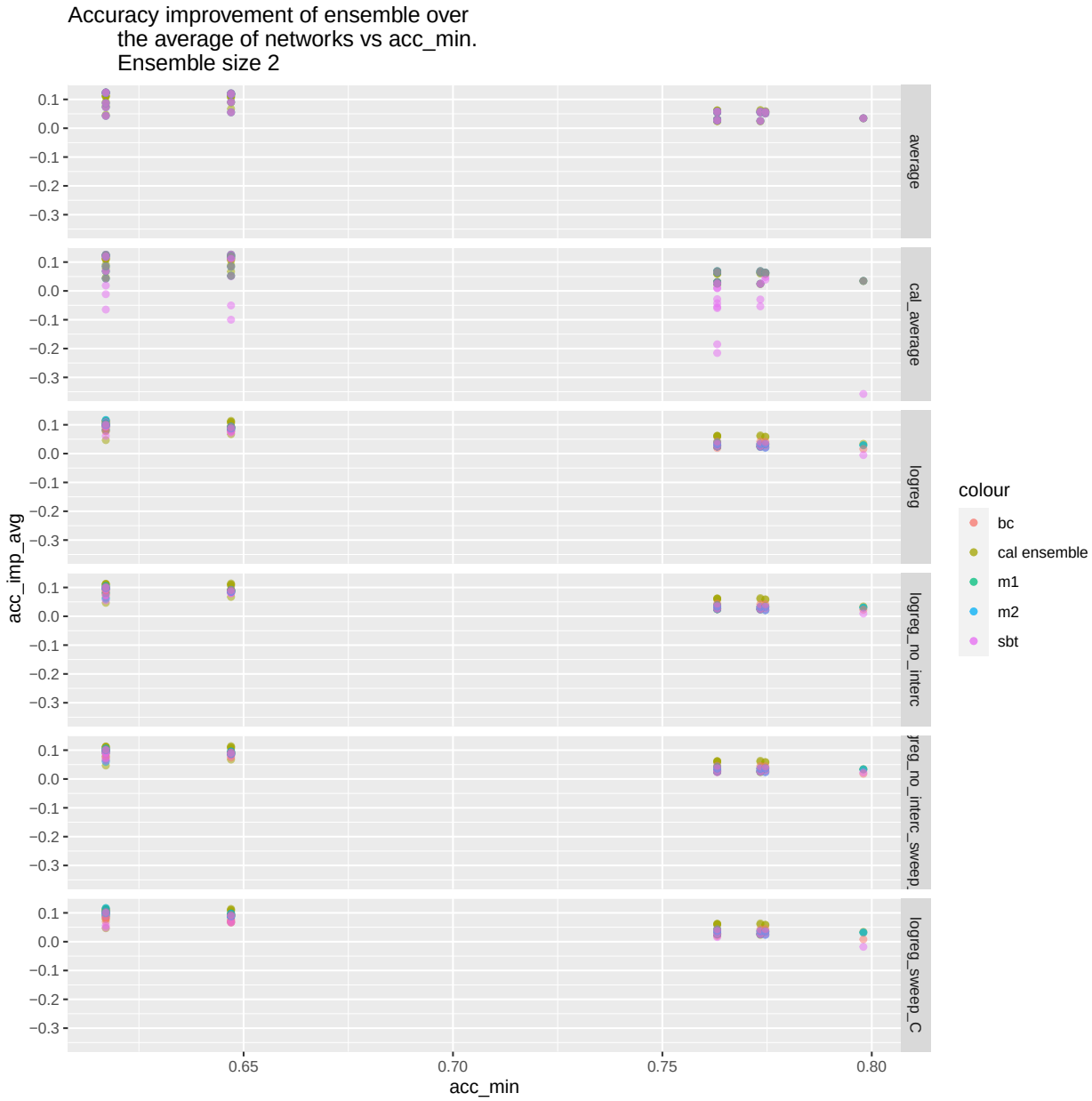

Accuracy improvement of ensemble over
the best of networks vs nll_avg.
Ensemble size 2
acc_imp_max
nll_avg
0.1
0.0
−0.1
−0.2
−0.3
8000
10000
12000
14000
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

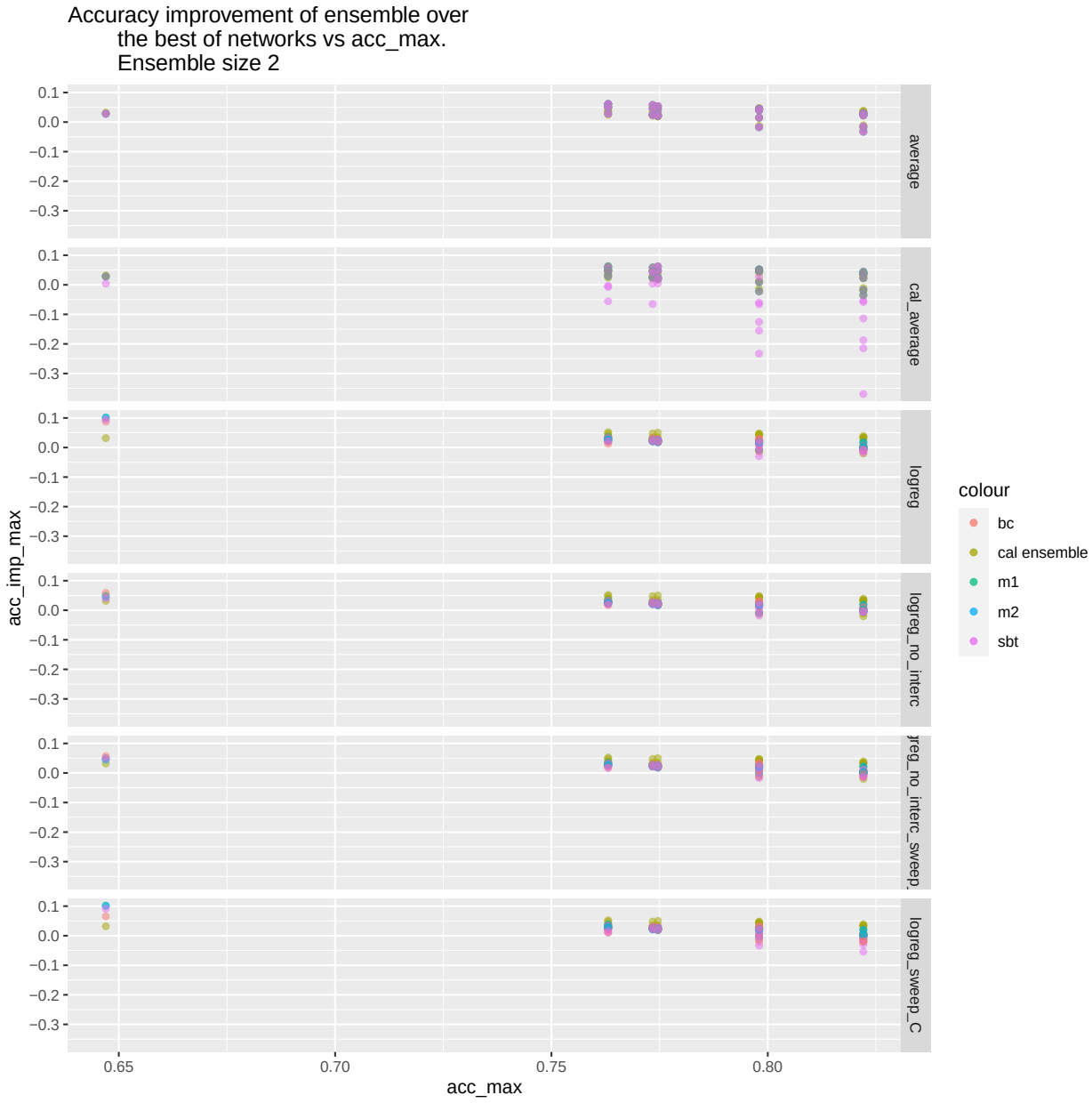

Accuracy improvement of ensemble over
the average of networks vs nll_avg.
Ensemble size 2
acc_imp_avg
nll_avg
0.1
0.0
−0.1
−0.2
−0.3
8000
10000
12000
14000
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

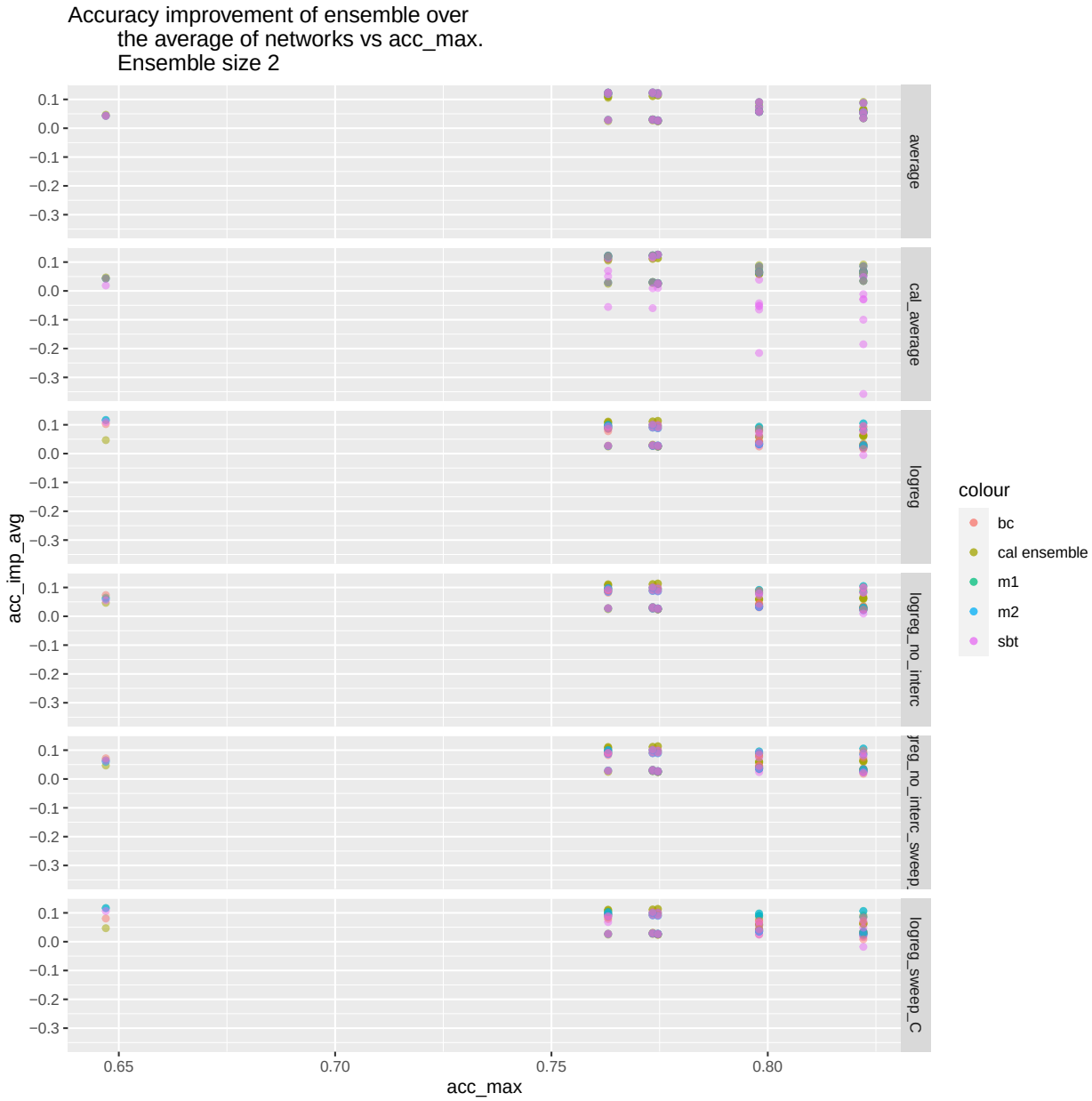

Accuracy improvement of ensemble over
the best of networks vs acc_var.
Ensemble size 2
acc_imp_max
acc_var
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.1
0.0
−0.1
−0.2
−0.3
0.000
0.005
0.010
0.015
0.020
colour
bc
cal ensemble
m1
m2
sbt

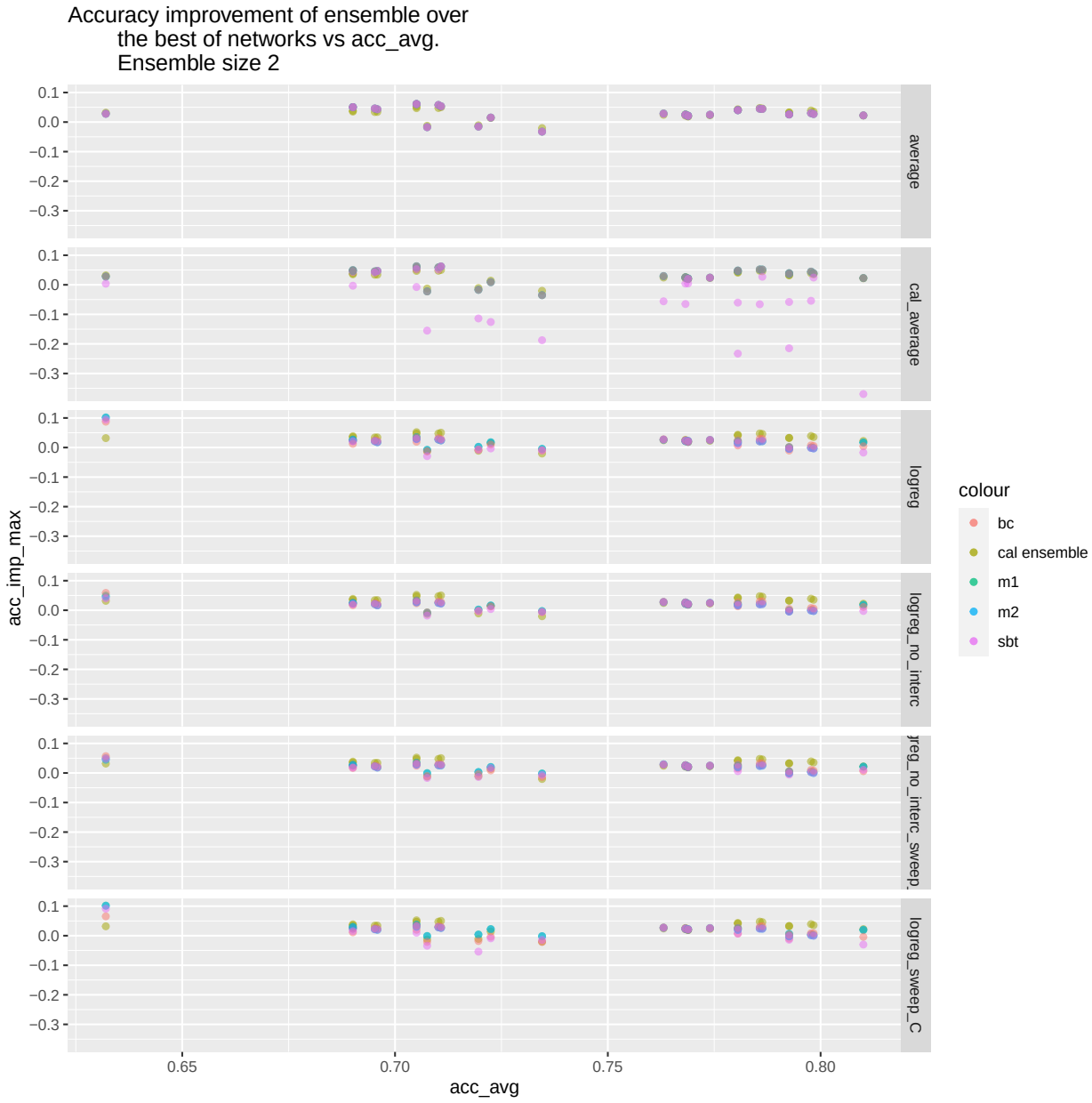

Accuracy improvement of ensemble over
the average of networks vs acc_var.
Ensemble size 2
acc_imp_avg
acc_var
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

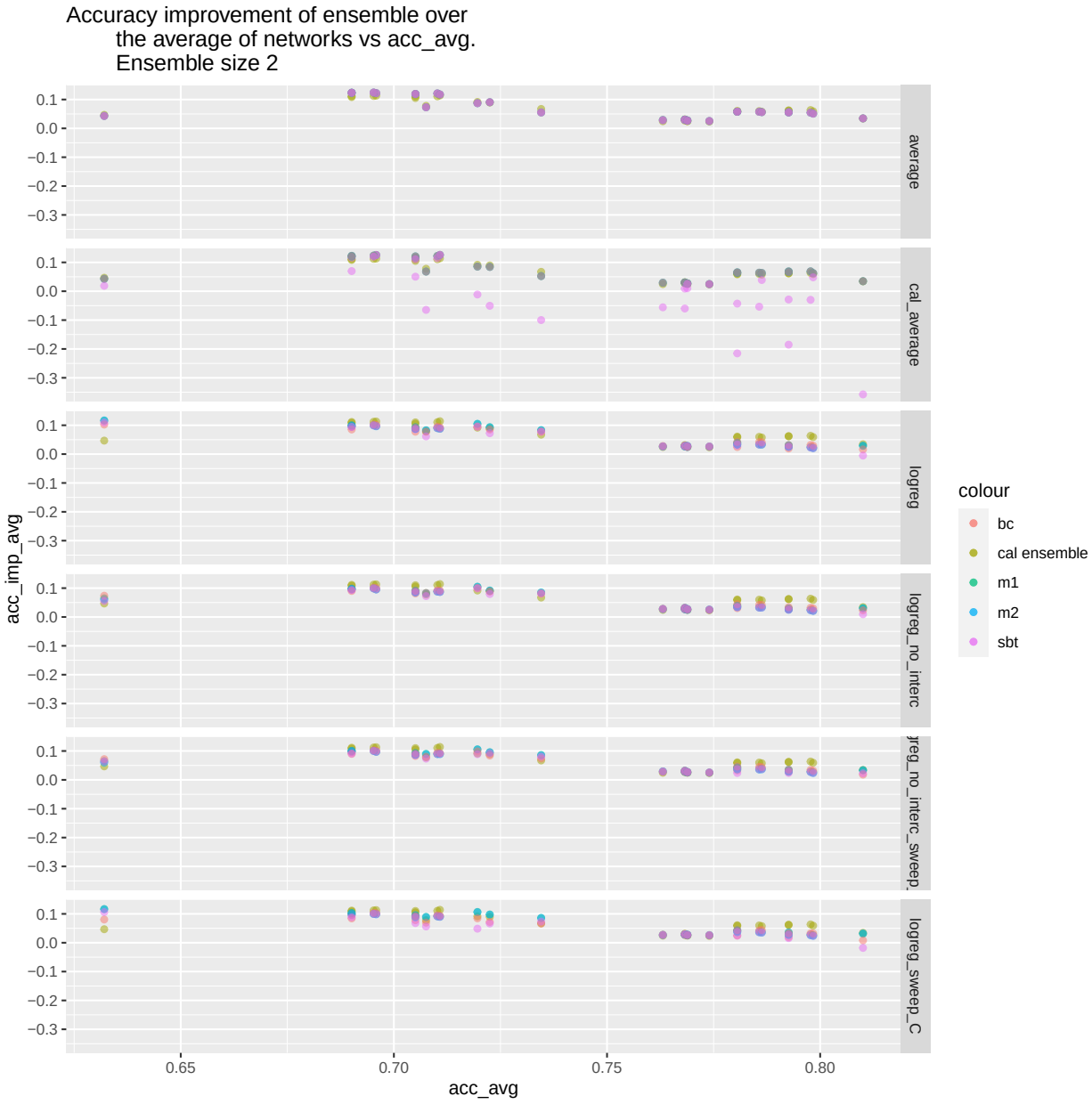

Accuracy improvement of ensemble over
the best of networks vs acc_min.
Ensemble size 2
acc_imp_max
acc_min
0.65
0.70
0.75
0.80
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

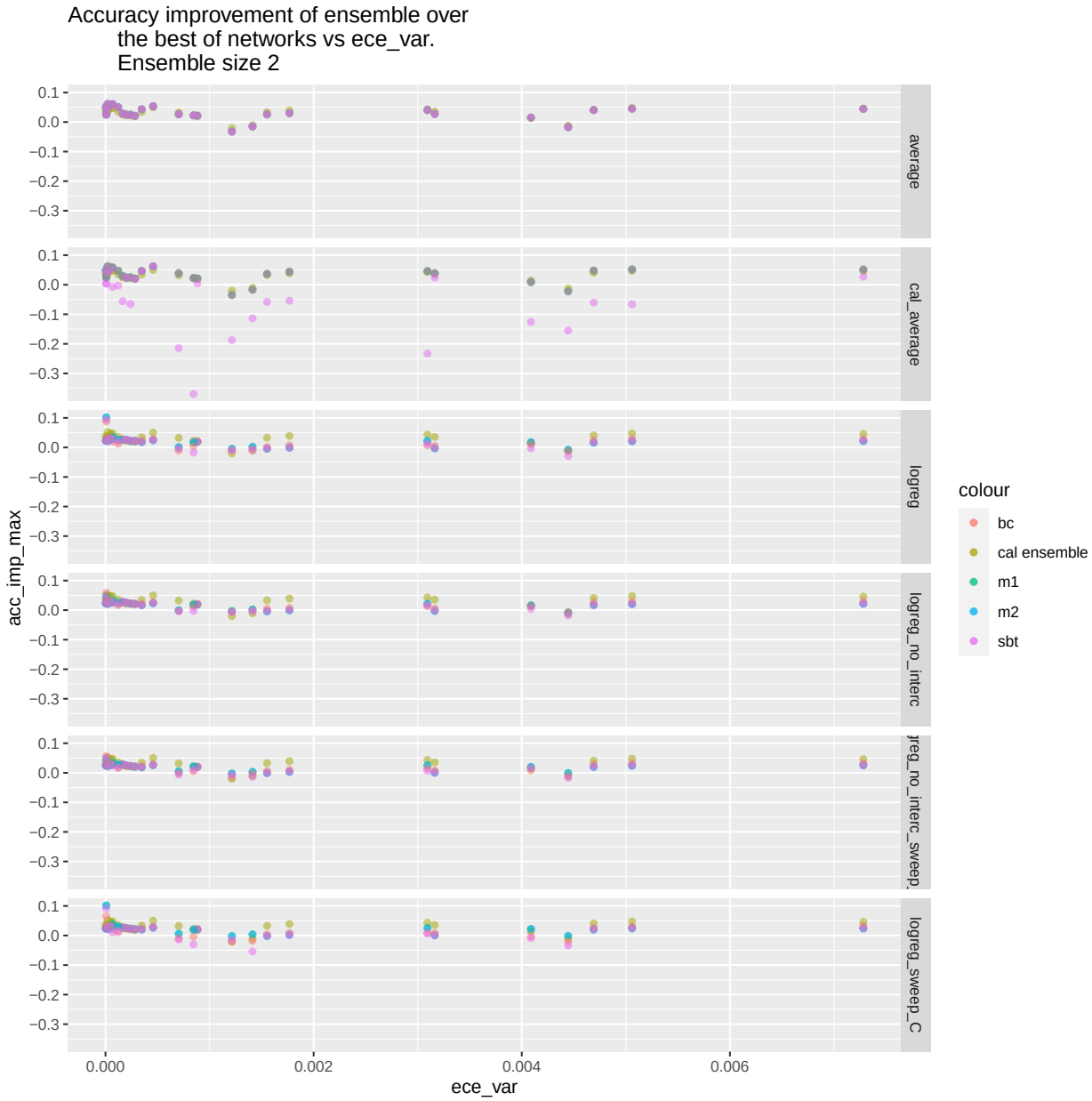

Accuracy improvement of ensemble over
the average of networks vs acc_min.
Ensemble size 2
acc_imp_avg
acc_min
0.65
0.70
0.75
0.80
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

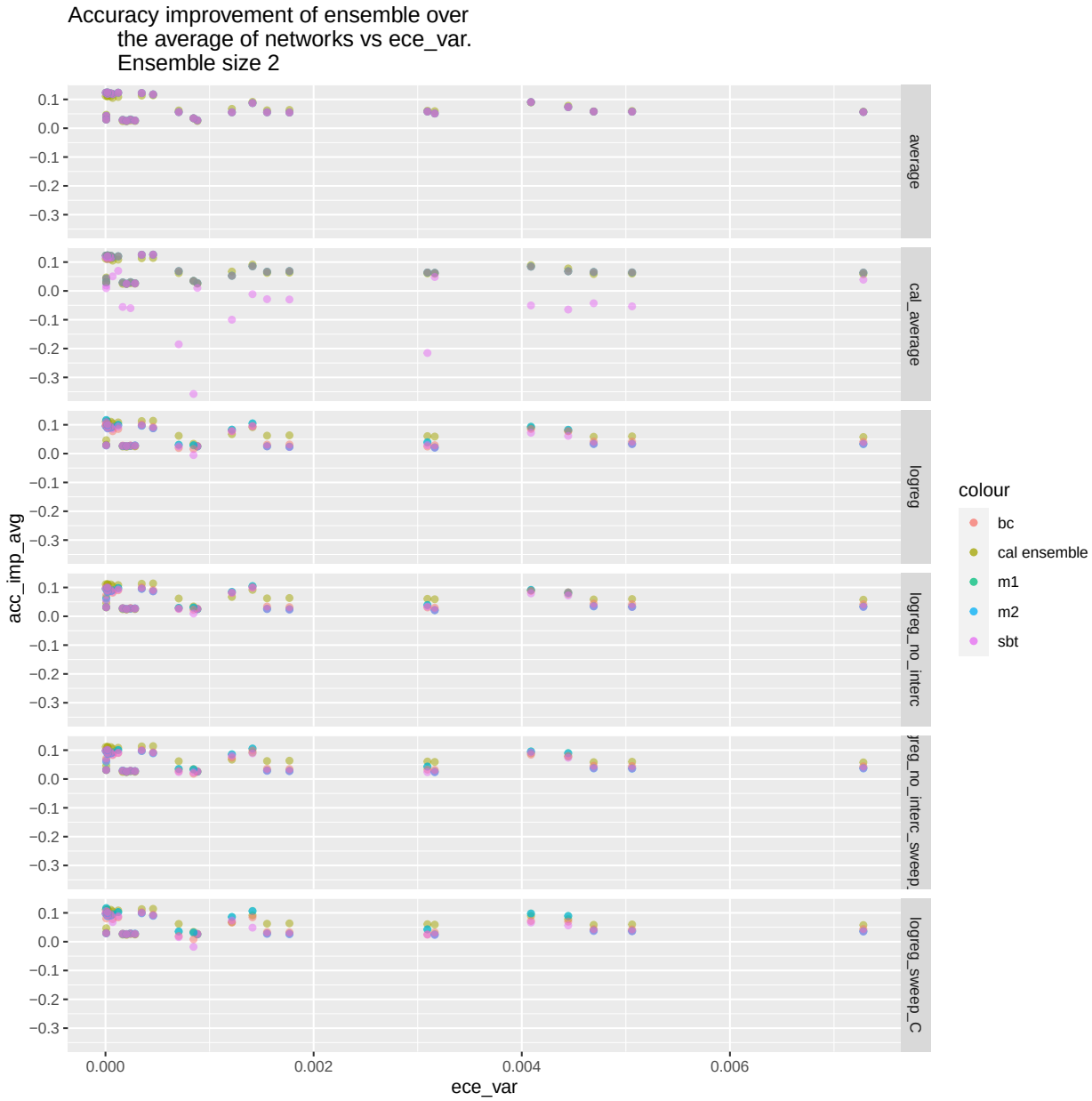

Accuracy improvement of ensemble over
the best of networks vs acc_max.
Ensemble size 2
acc_imp_max
acc_max
0.65
0.70
0.75
0.80
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

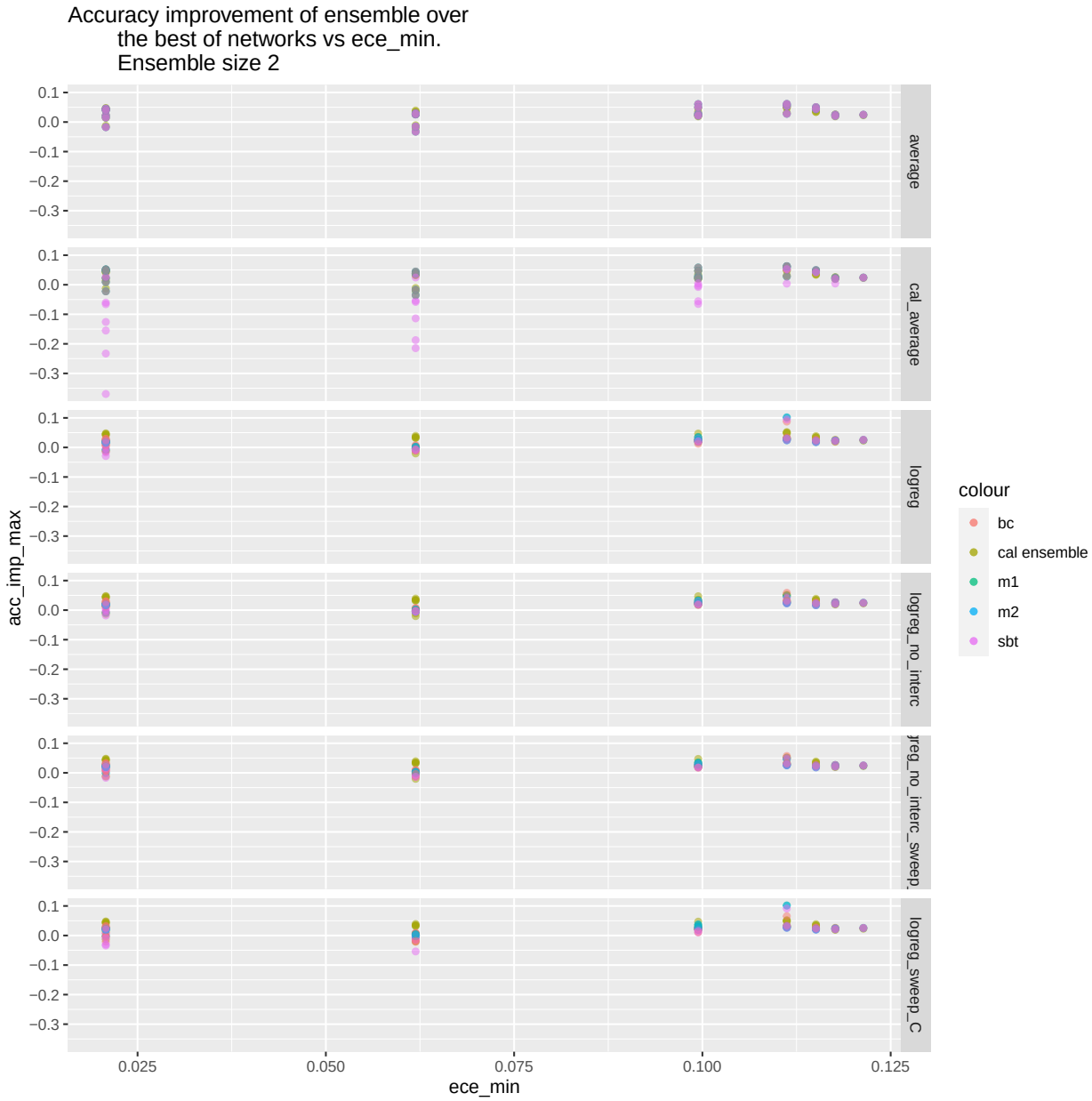

Accuracy improvement of ensemble over
the average of networks vs acc_max.
Ensemble size 2
acc_imp_avg
acc_max
0.65
0.70
0.75
0.80
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

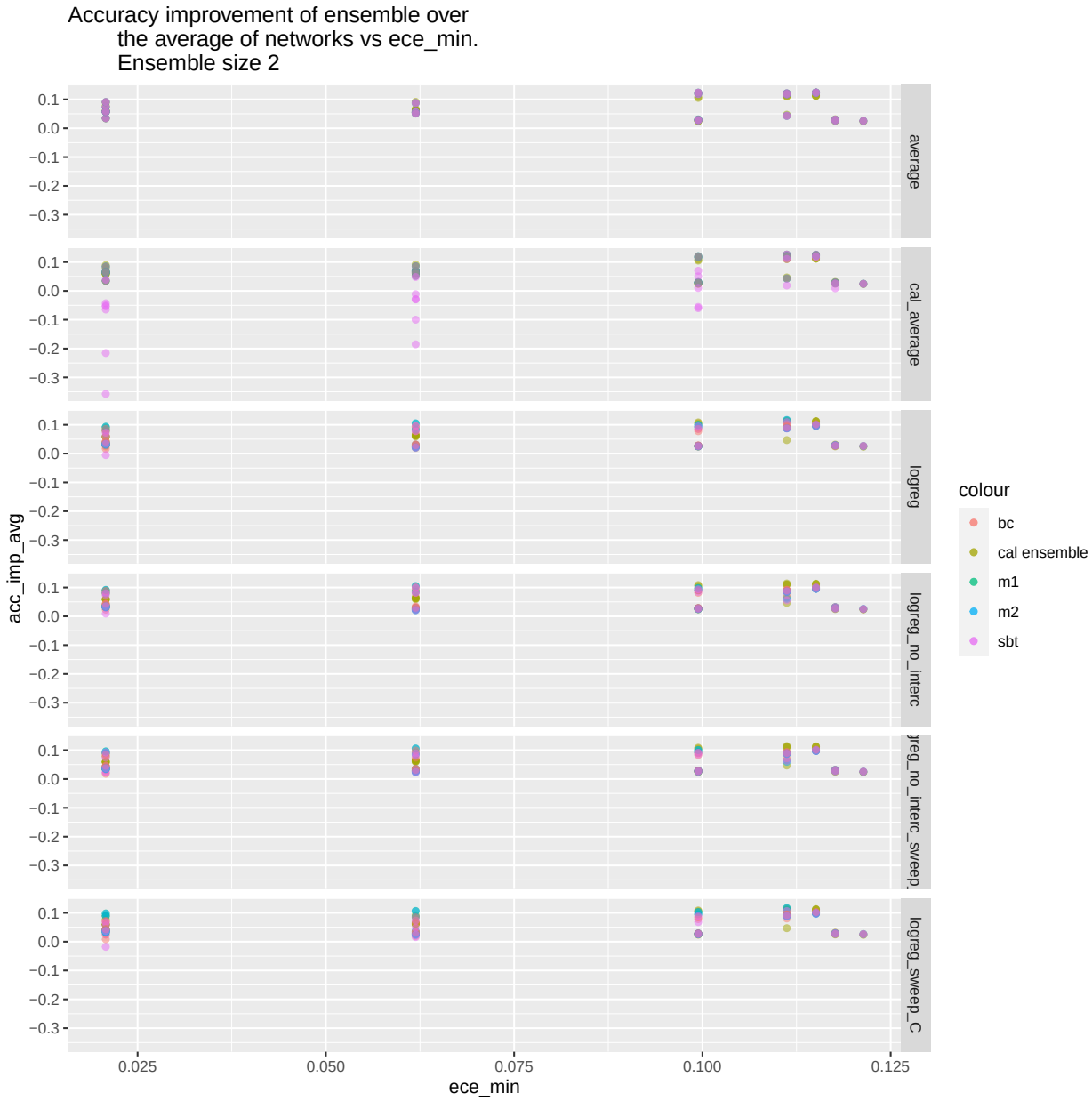

Accuracy improvement of ensemble over
the best of networks vs acc_avg.
Ensemble size 2
acc_imp_max
acc_avg
0.65
0.70
0.75
0.80
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

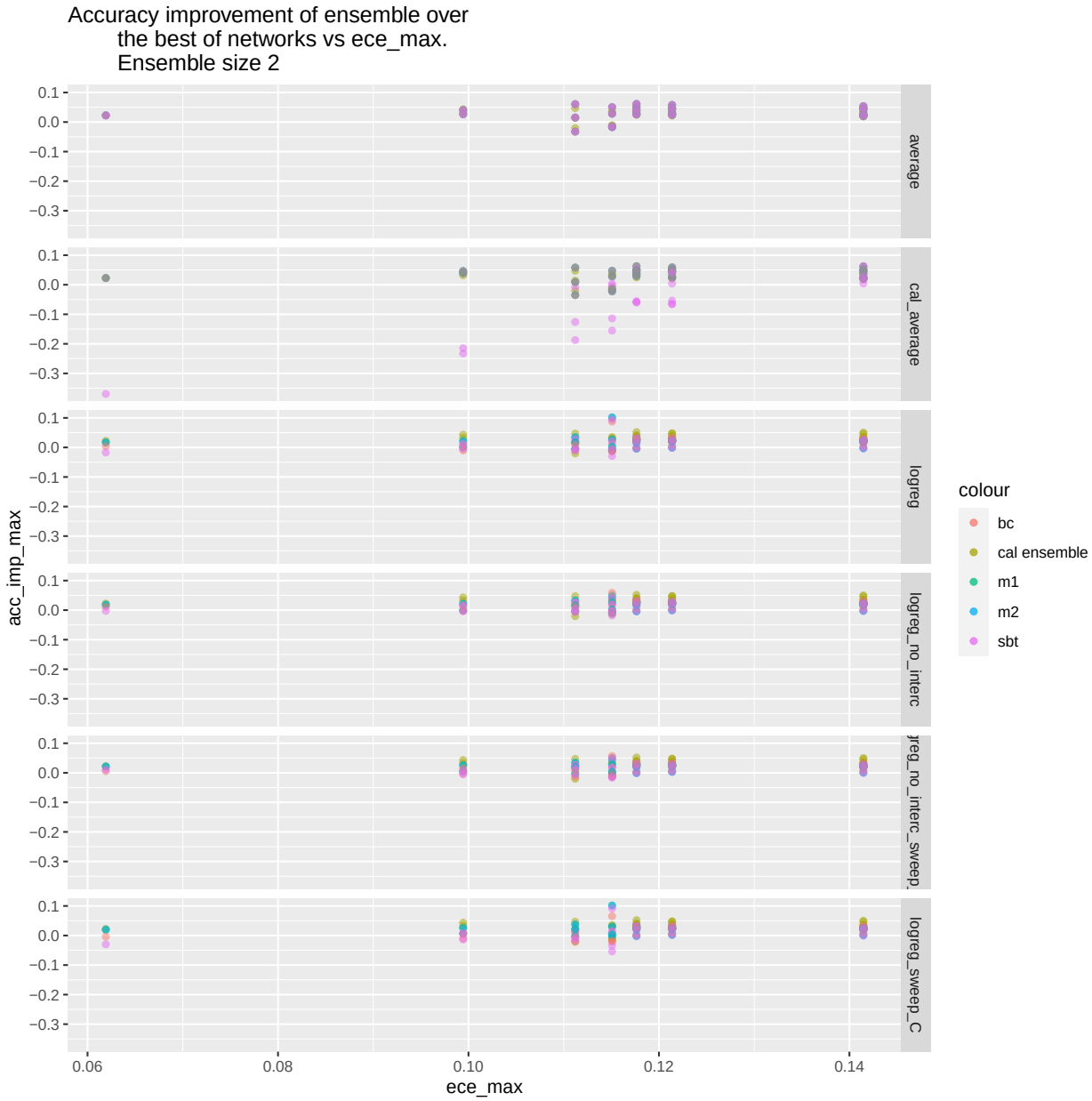

Accuracy improvement of ensemble over
the average of networks vs acc_avg.
Ensemble size 2
acc_imp_avg
acc_avg
0.65
0.70
0.75
0.80
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

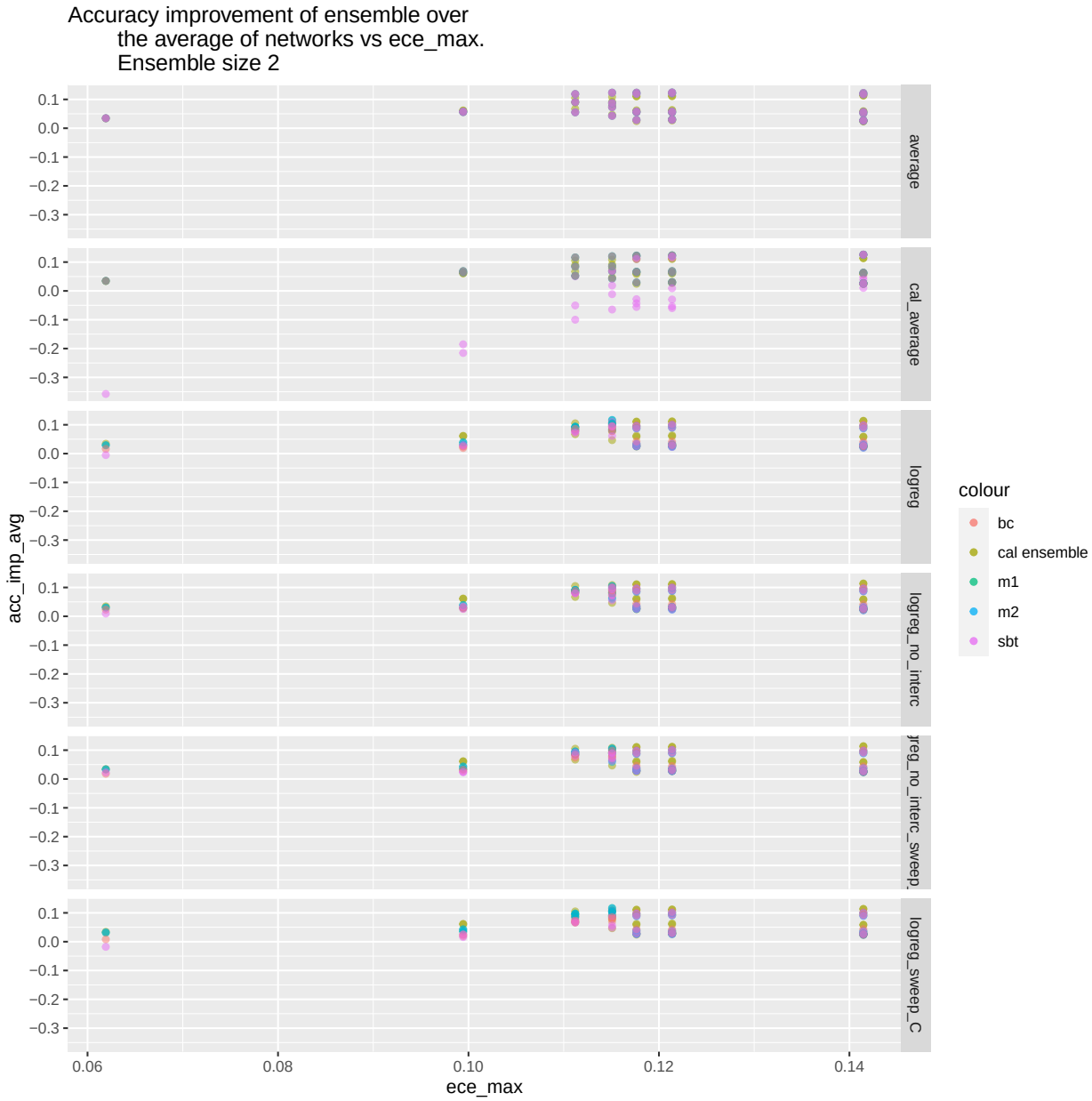

Accuracy improvement of ensemble over
the best of networks vs ece_var.
Ensemble size 2
acc_imp_max
ece_var
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

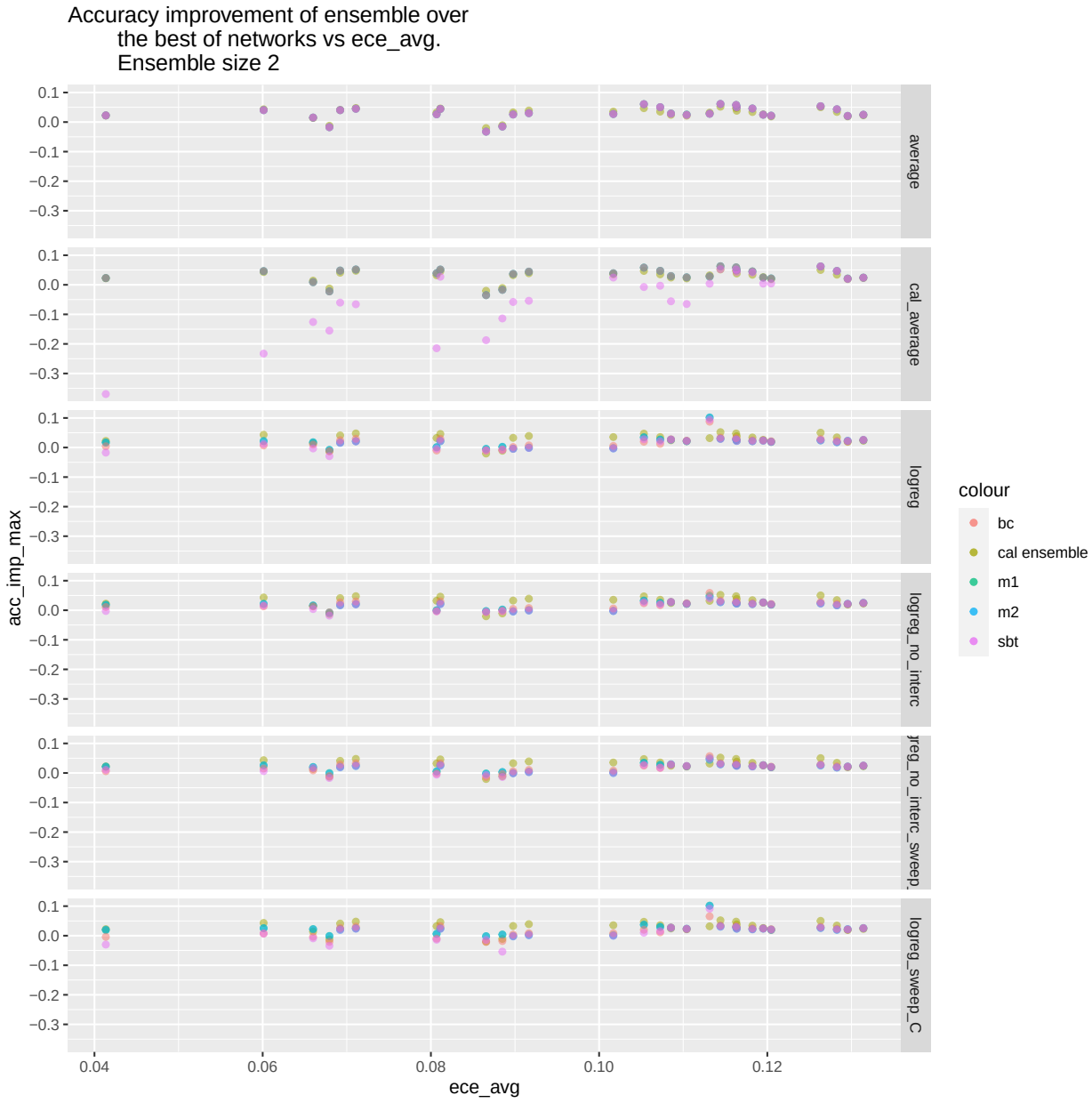

Accuracy improvement of ensemble over the average of networks vs ece_var. Ensemble size 2
acc_imp_avg
ece_var
0.000
0.002
0.004
0.006
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

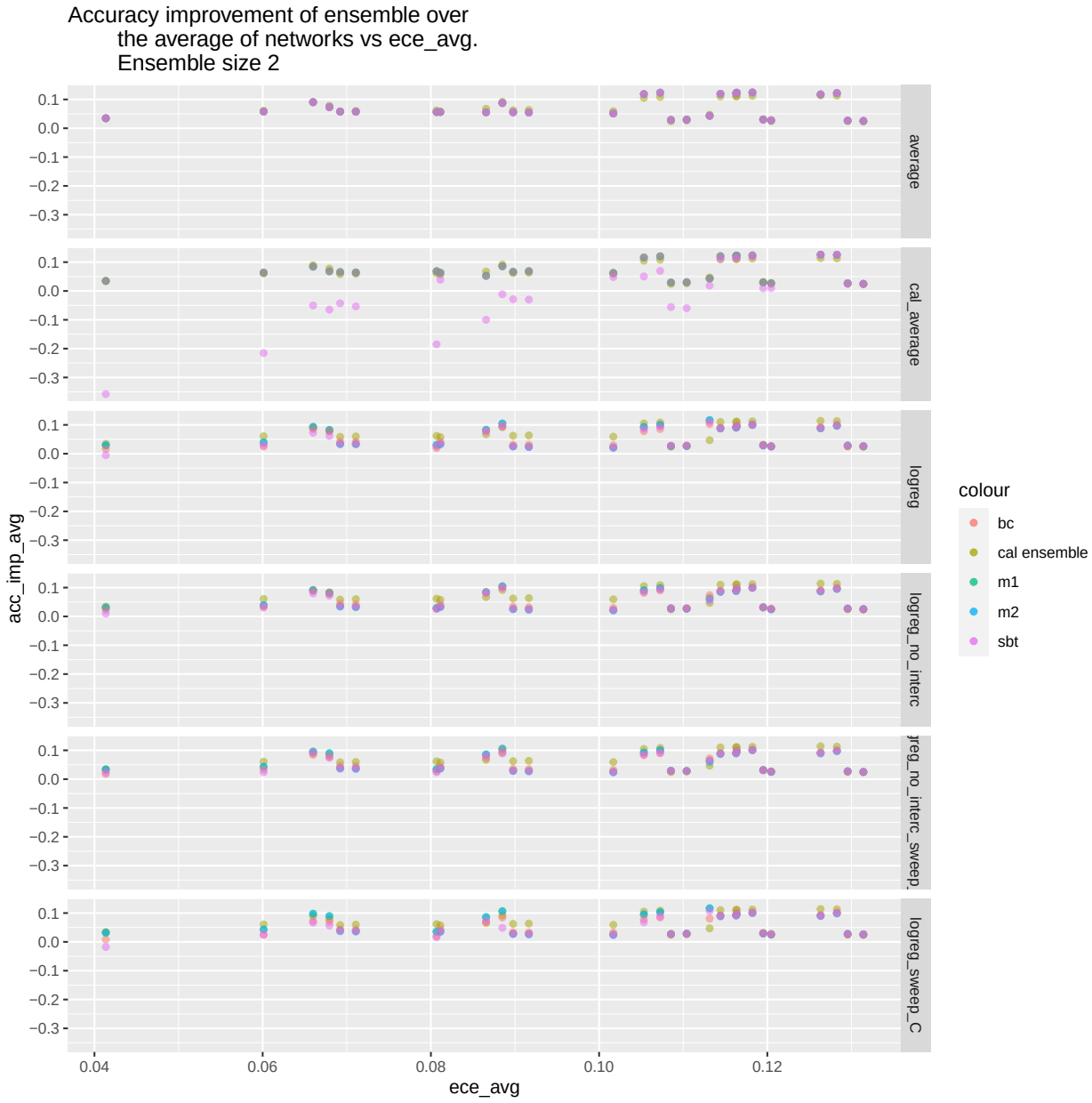

Accuracy improvement of ensemble over
the best of networks vs ece_min.
Ensemble size 2
acc_imp_max
ece_min
0.025
0.050
0.075
0.100
0.125
0.1
0.0
−0.1
−0.2
−0.3
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

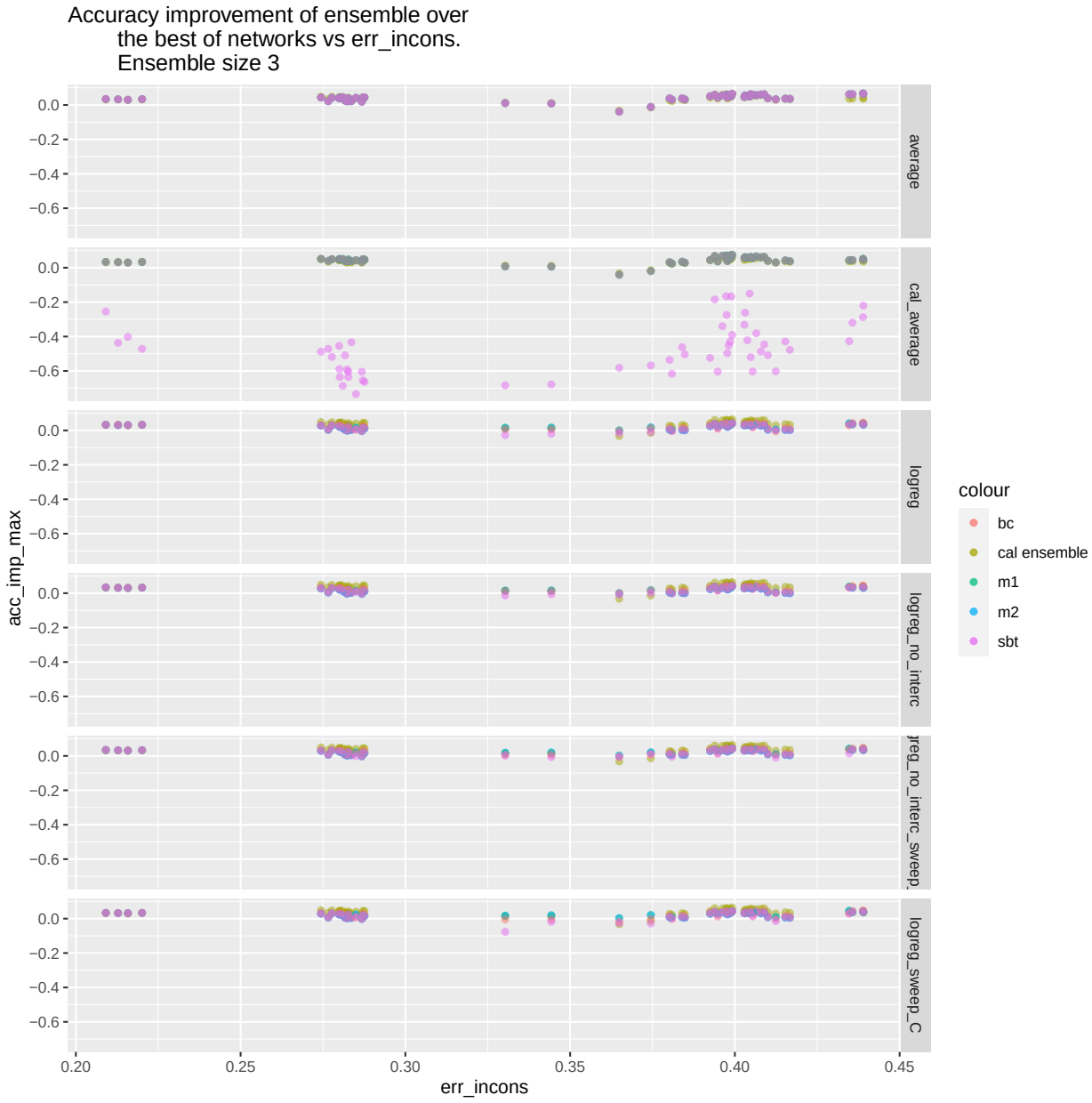

Accuracy improvement of ensemble over
the average of networks vs ece_min.
Ensemble size 2
acc_imp_avg
ece_min
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

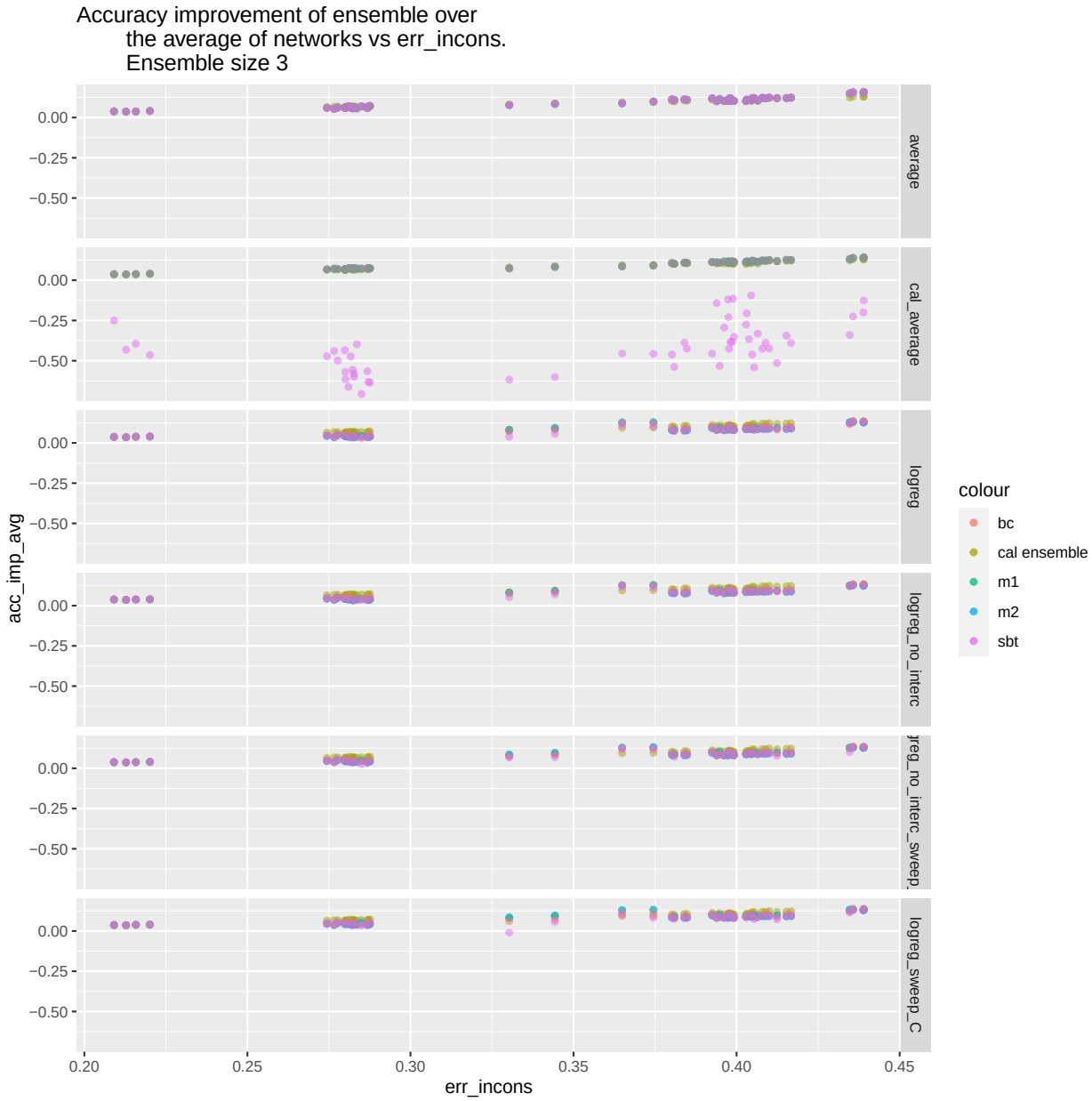

Accuracy improvement of ensemble over
the best of networks vs ece_max.
Ensemble size 2
acc_imp_max
ece_max
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.1
0.0
−0.1
−0.2
−0.3
0.06
0.08
0.10
0.12
0.14
colour
bc
cal ensemble
m1
m2
sbt

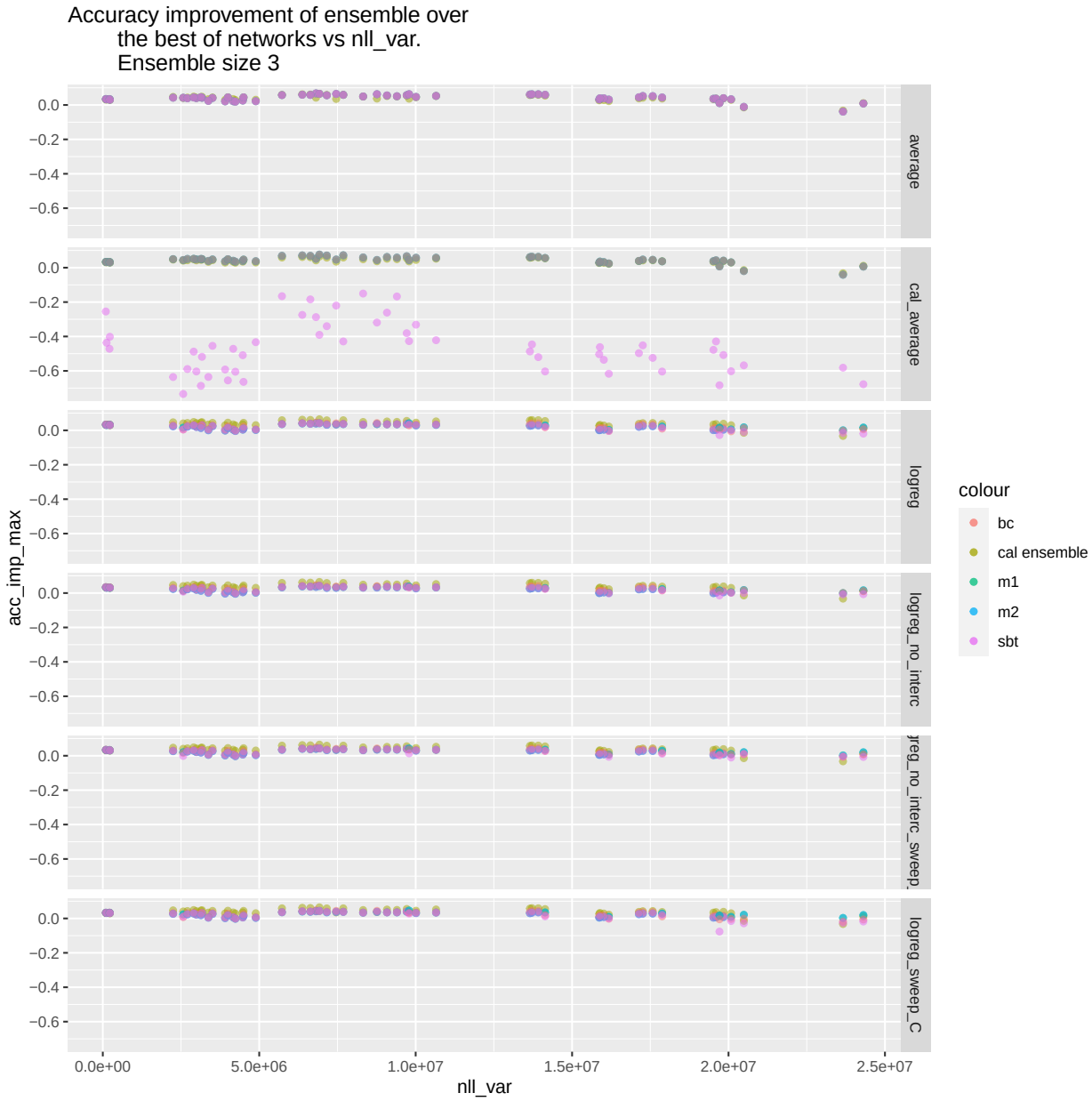

Accuracy improvement of ensemble over
the average of networks vs ece_max.
Ensemble size 2
acc_imp_avg
ece_max
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

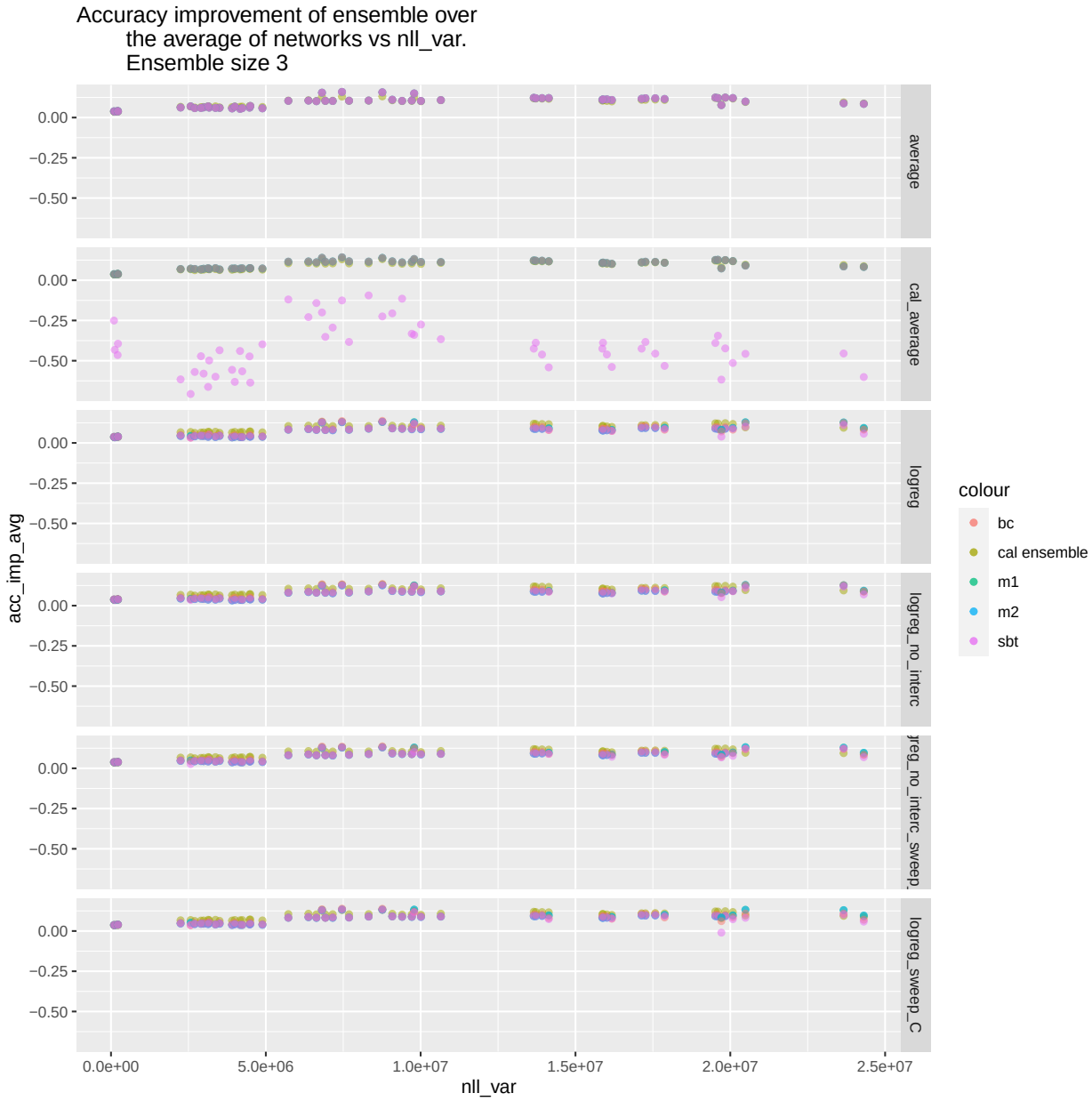

Accuracy improvement of ensemble over
the best of networks vs ece_avg.
Ensemble size 2
acc_imp_max
ece_avg
0.1
0.0
−0.1
−0.2
−0.3
0.04
0.06
0.08
0.10
0.12
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

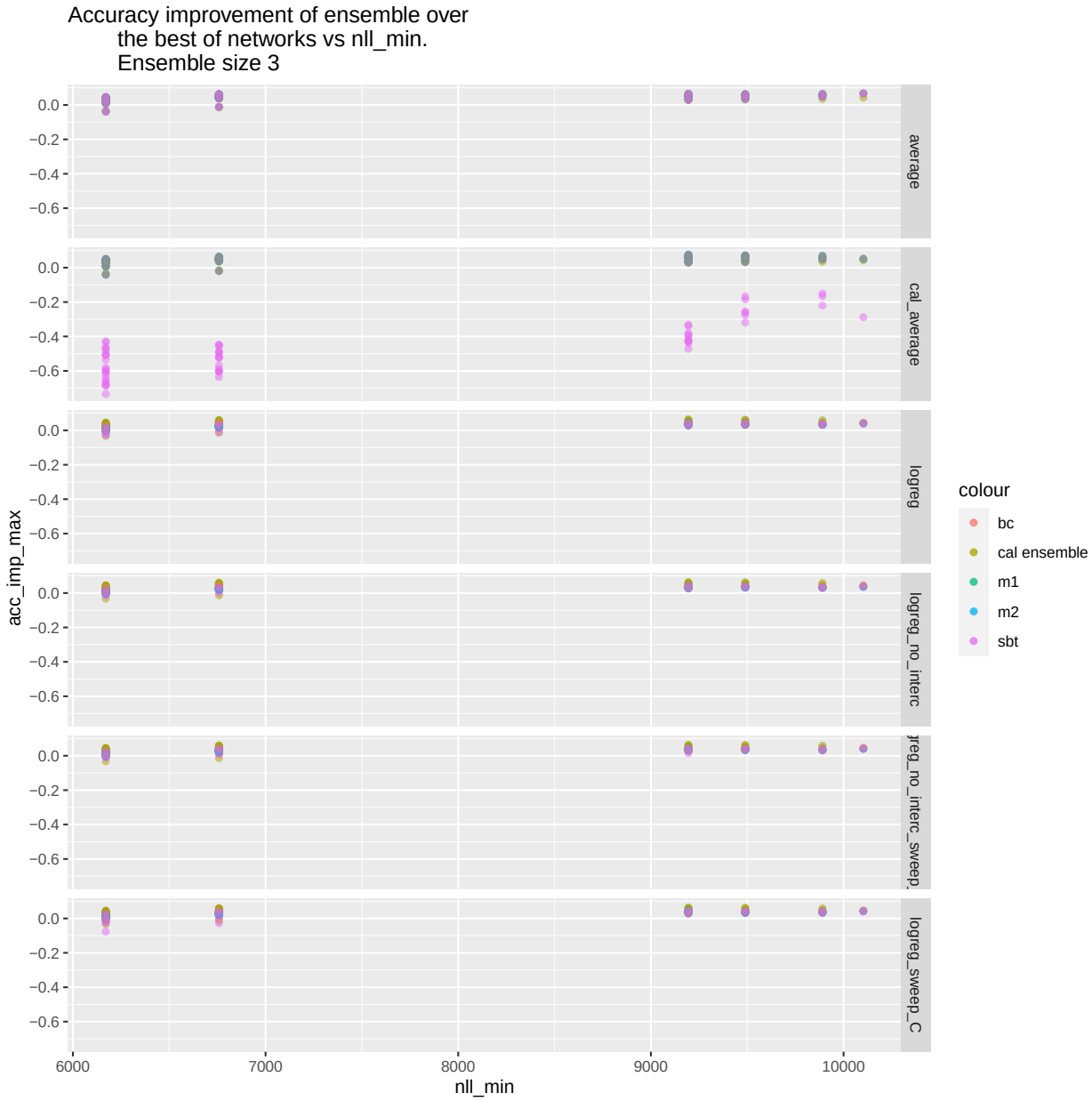

Accuracy improvement of ensemble over
the average of networks vs ece_avg.
Ensemble size 2
acc_imp_avg
ece_avg
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

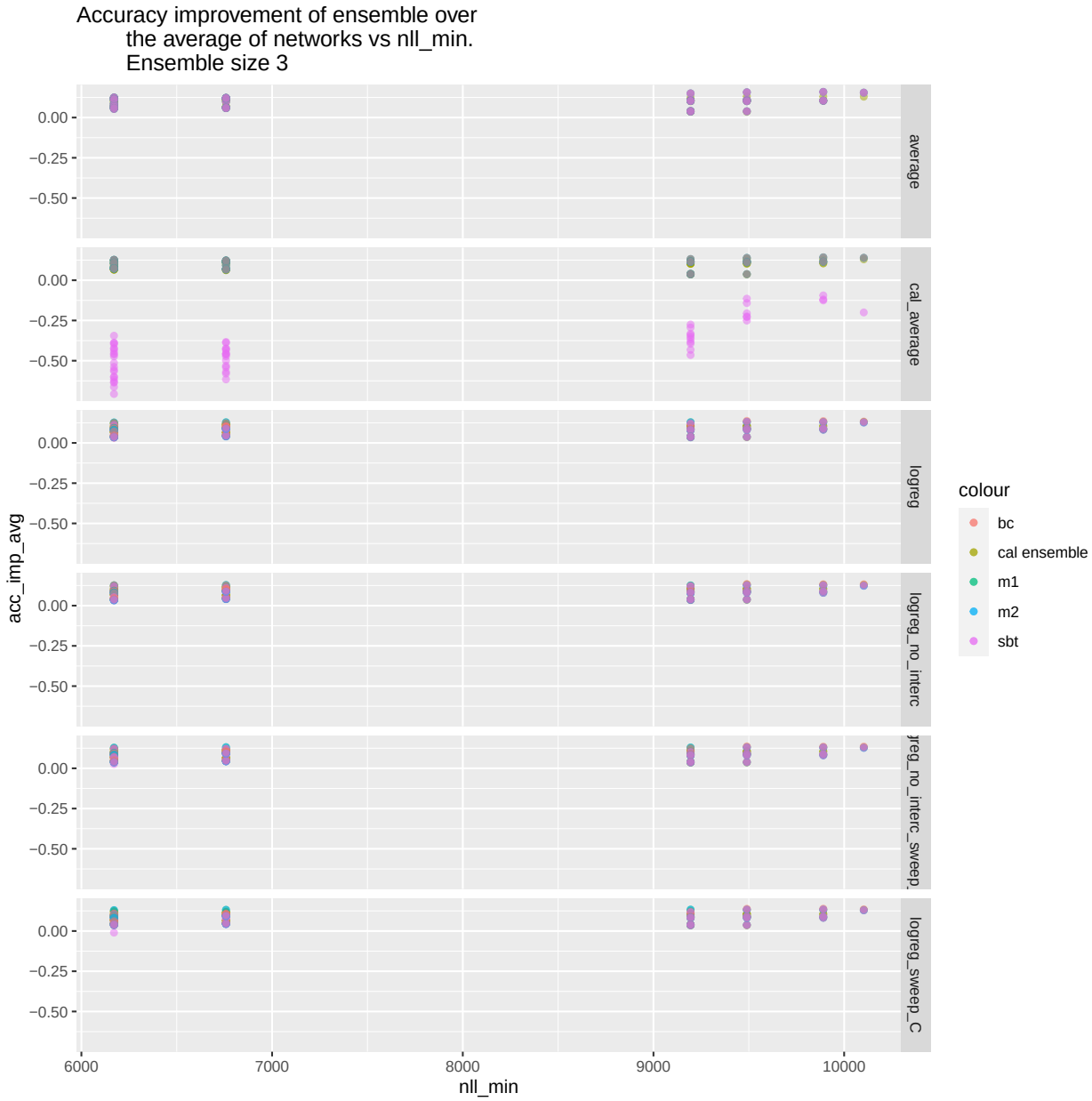

Accuracy improvement of ensemble over
the best of networks vs err_incons.
Ensemble size 3
acc_imp_max
err_incons
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

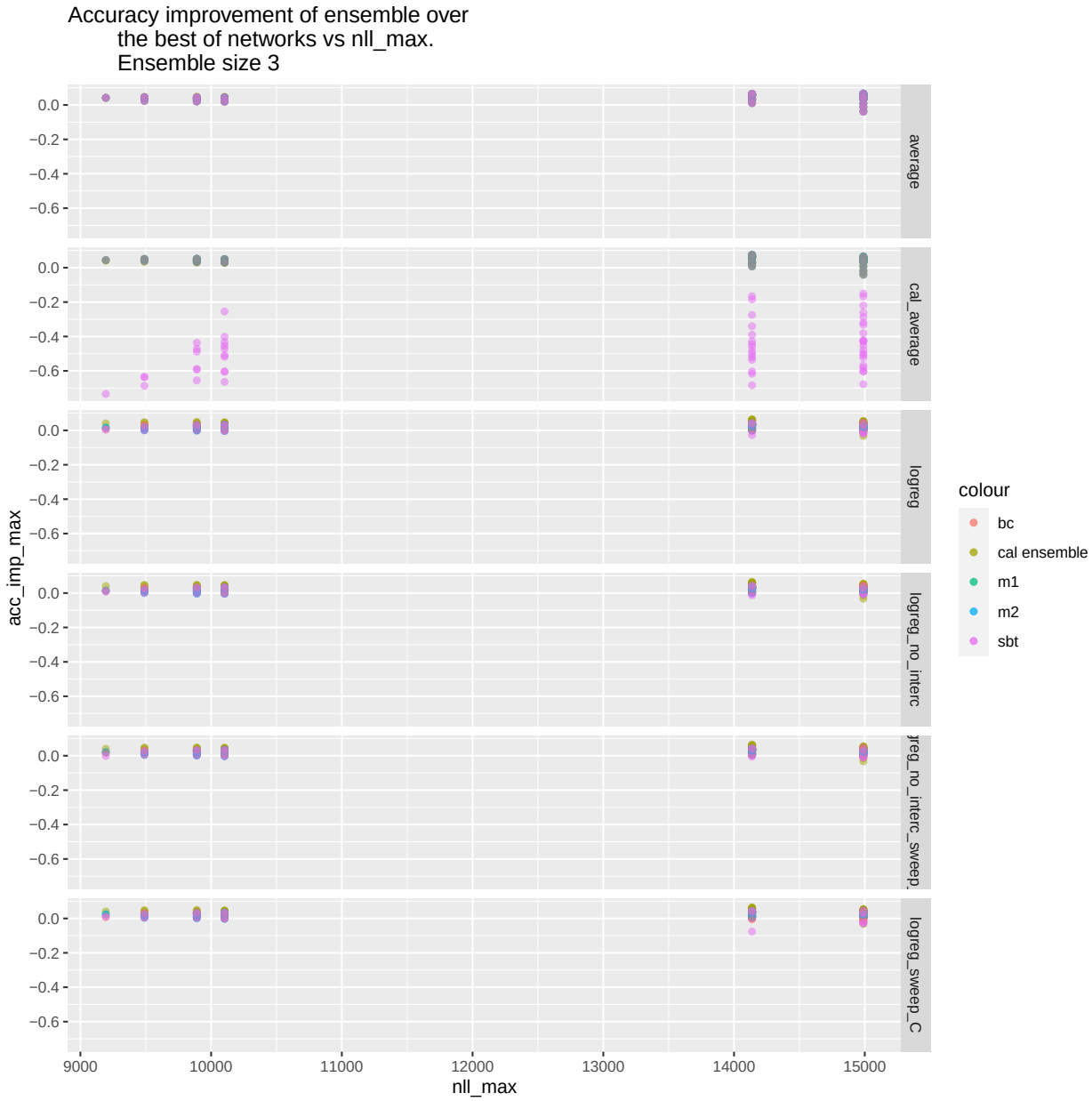

Accuracy improvement of ensemble over
the average of networks vs err_incons.
Ensemble size 3
acc_imp_avg
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
0.20
0.25
0.30
0.35
0.40
0.45
err_incons
colour
bc
cal ensemble
m1
m2
sbt

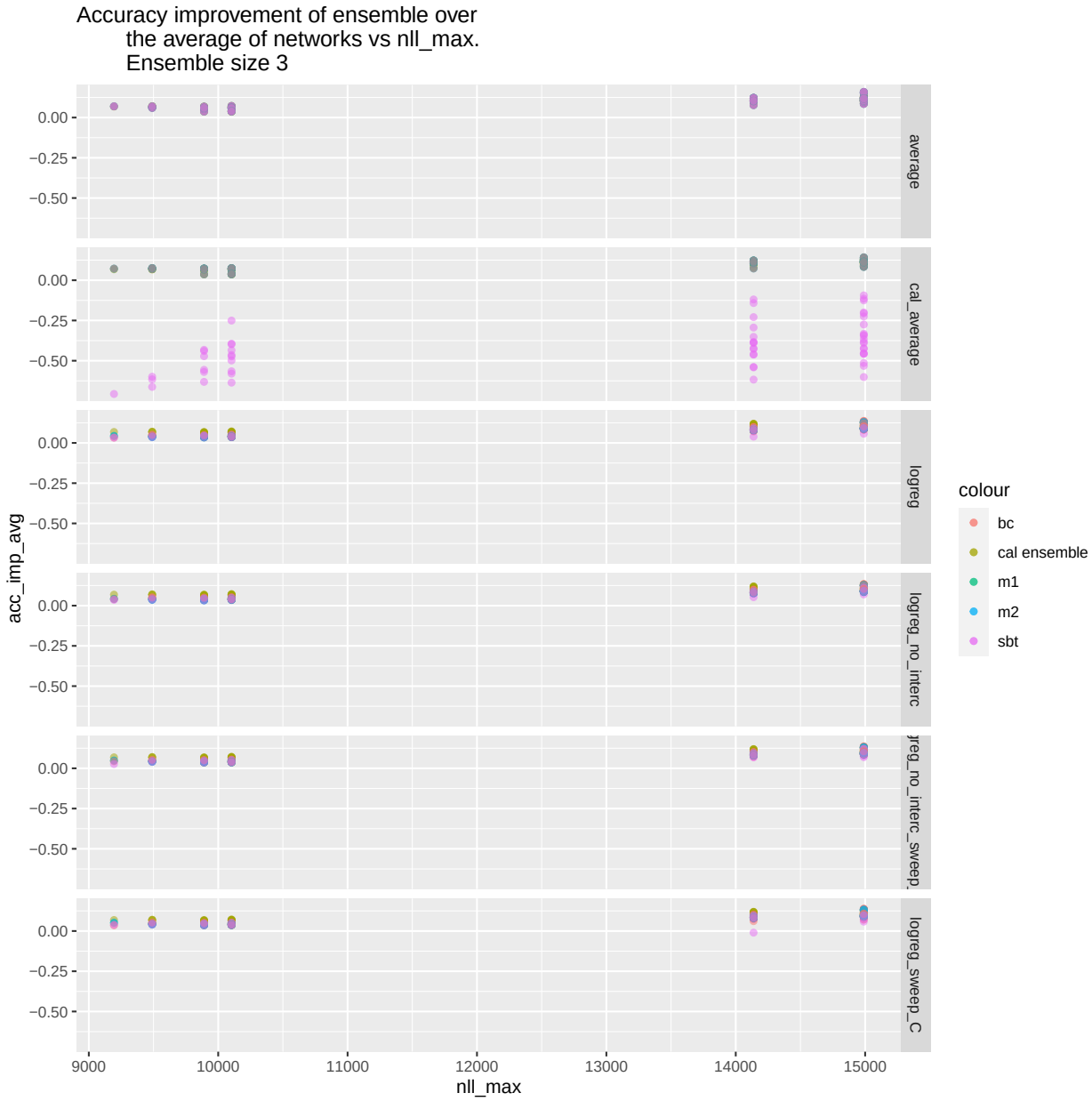

Accuracy improvement of ensemble over
the best of networks vs nll_var.
Ensemble size 3
acc_imp_max
nll_var
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
0.0
−0.2
−0.4
−0.6
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

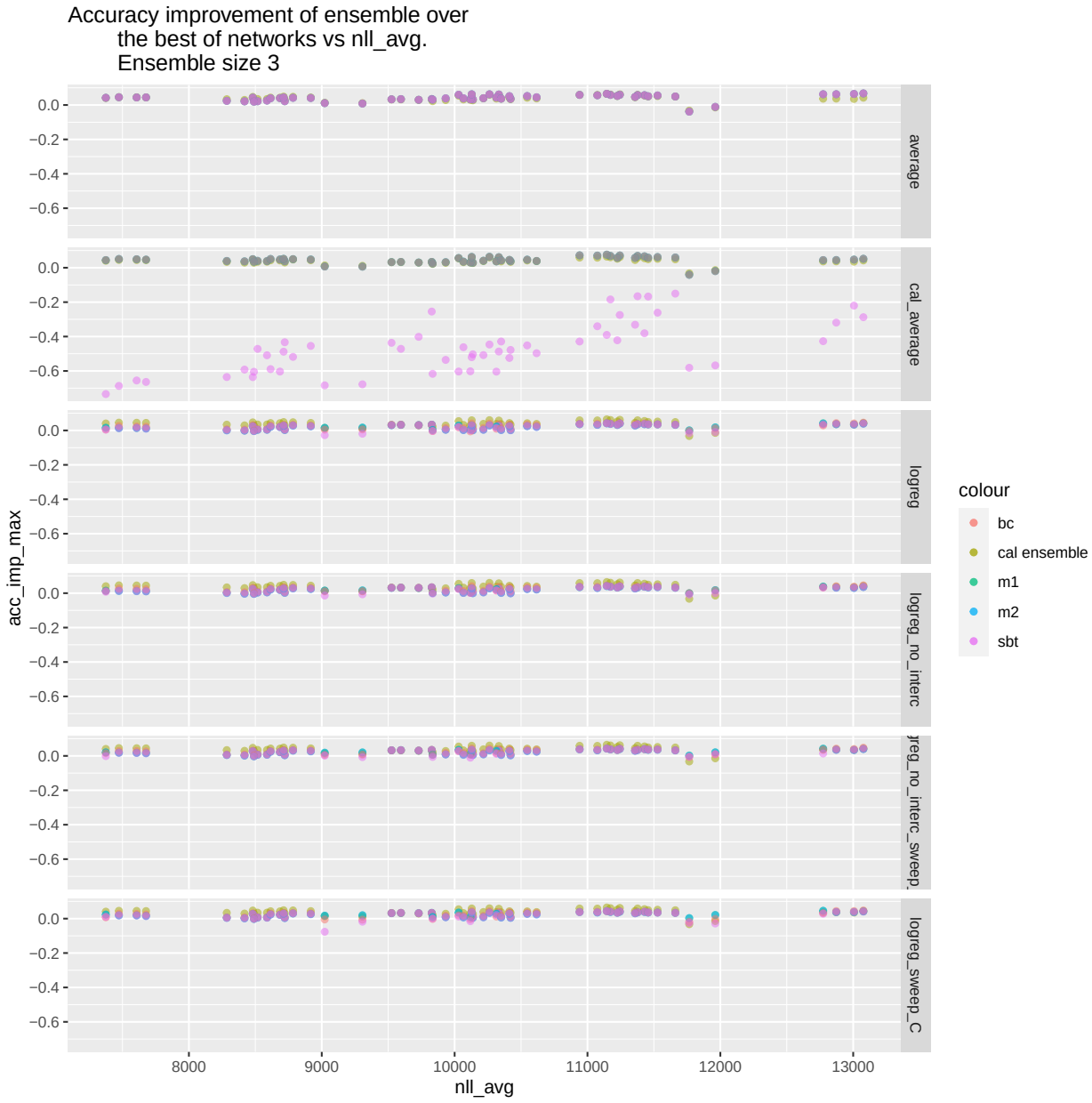

Accuracy improvement of ensemble over
the average of networks vs nll_var.
Ensemble size 3
acc_imp_avg
nll_var
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

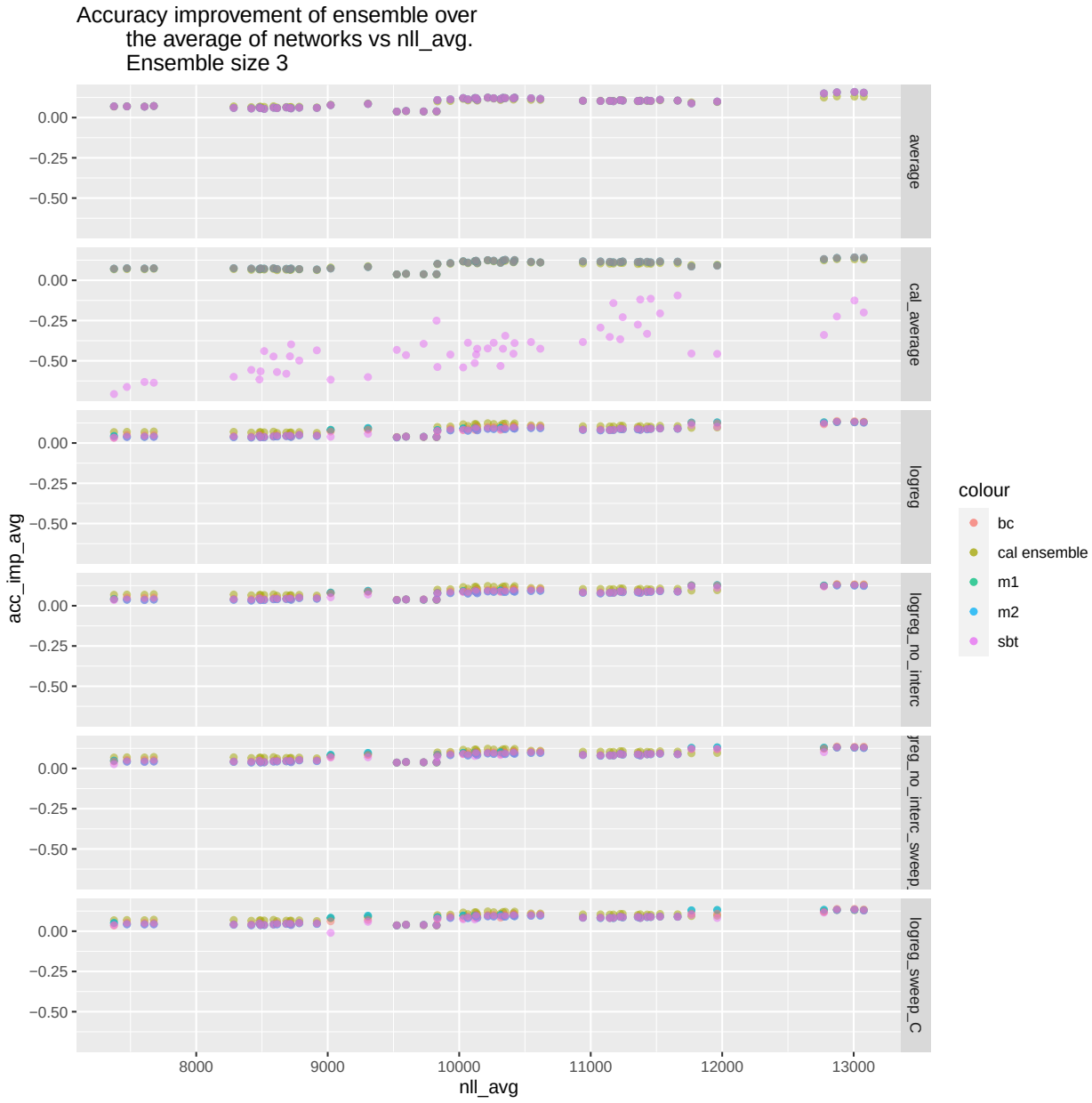

Accuracy improvement of ensemble over
the best of networks vs nll_min.
Ensemble size 3
acc_imp_max
nll_min
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

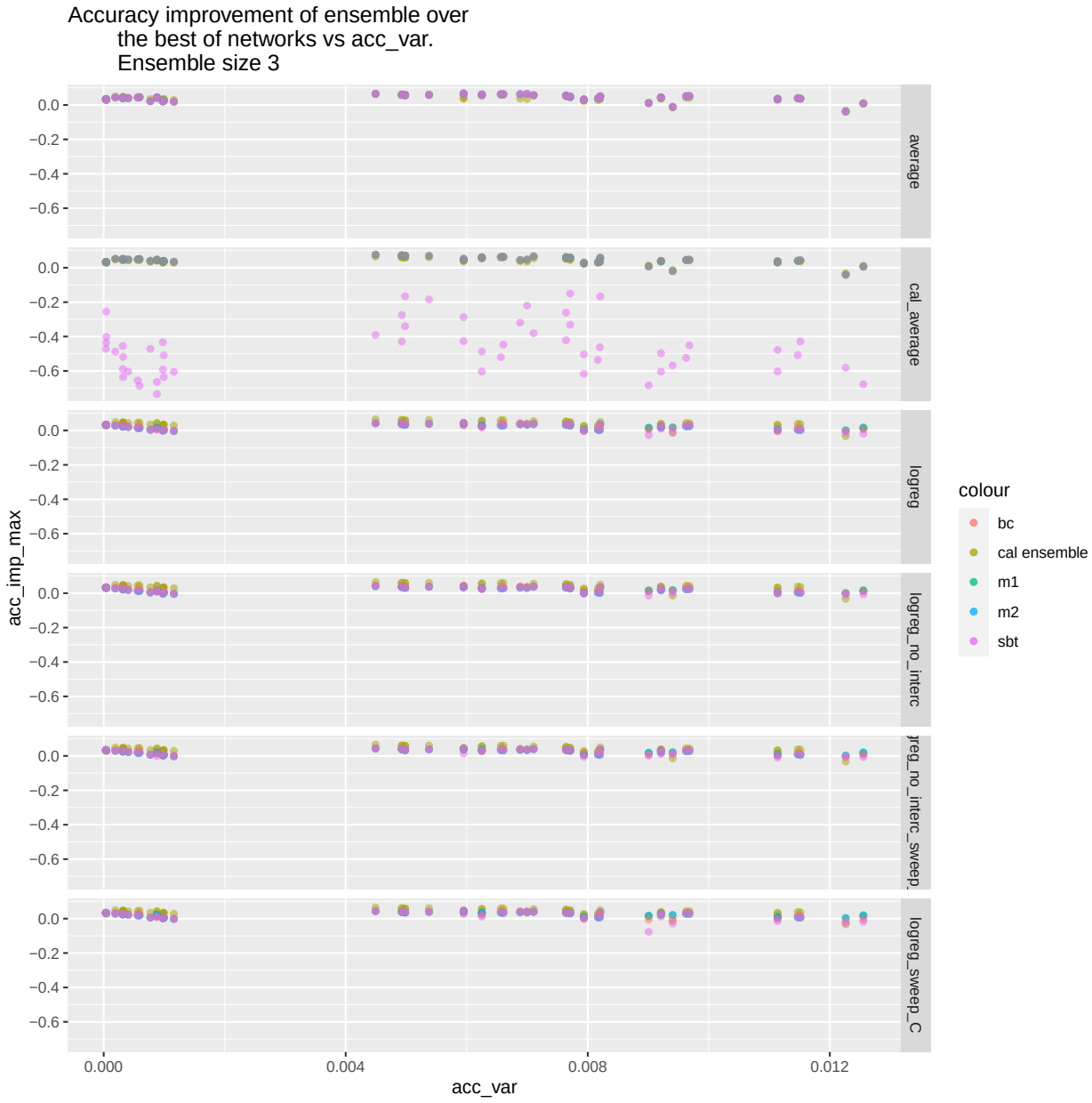

Accuracy improvement of ensemble over
the average of networks vs nll_min.
Ensemble size 3
acc_imp_avg
nll_min
0.00
−0.25
−0.50
6000
7000
8000
9000
10000
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

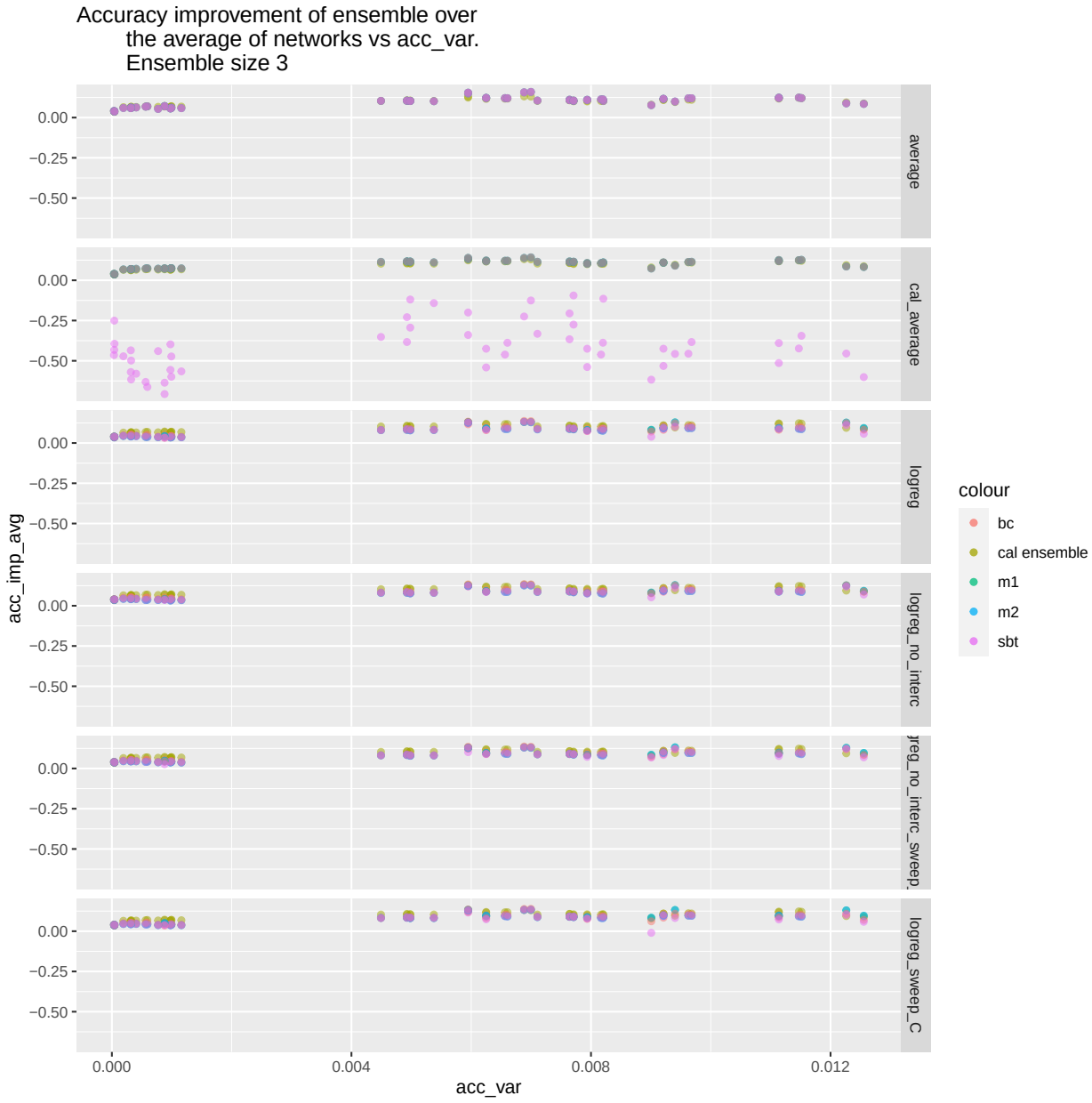

Accuracy improvement of ensemble over
the best of networks vs nll_max.
Ensemble size 3
acc_imp_max
nll_max
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
0.0
−0.2
−0.4
−0.6
9000
10000
11000
12000
13000
14000
15000
colour
bc
cal ensemble
m1
m2
sbt

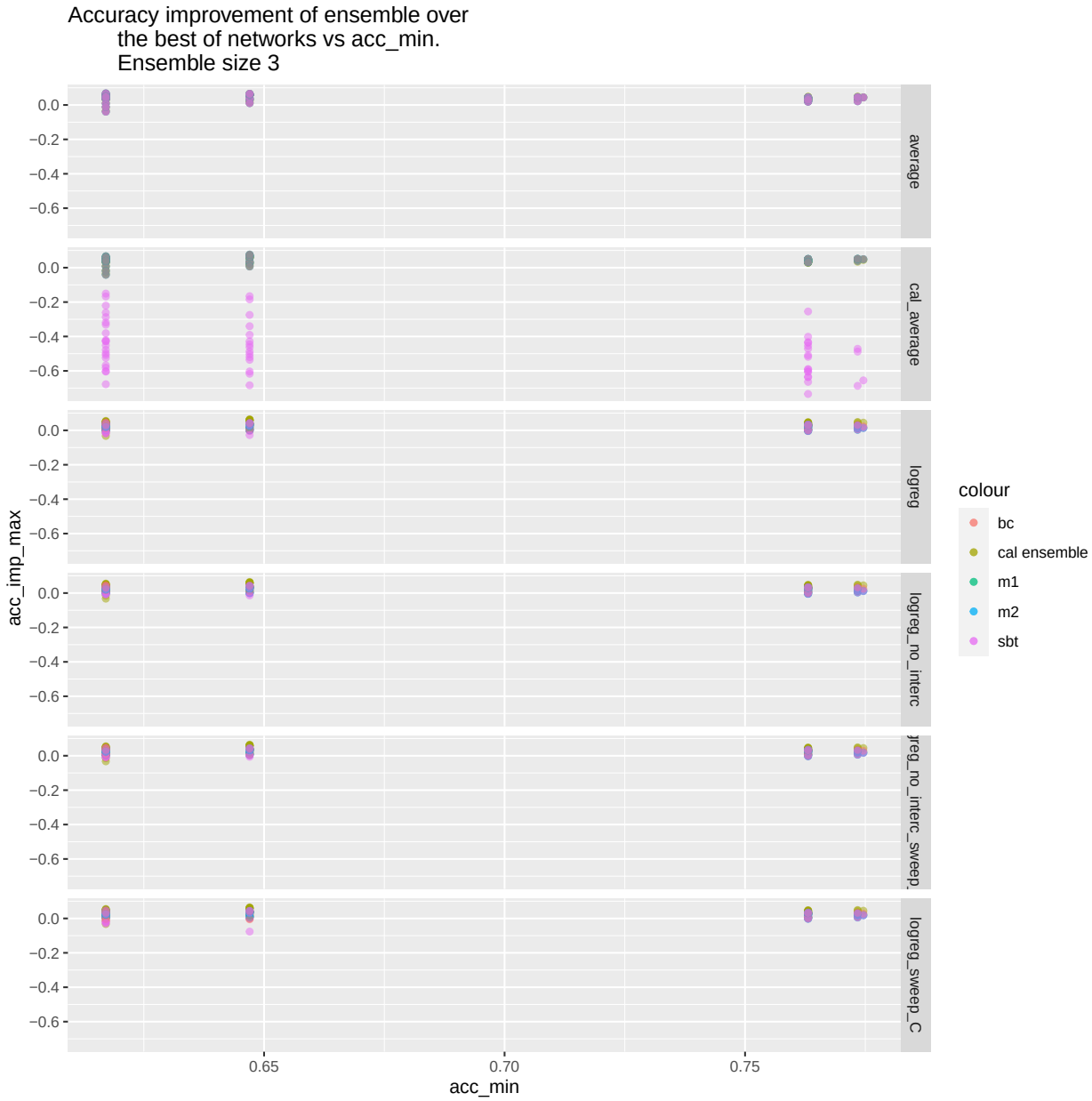

Accuracy improvement of ensemble over
the average of networks vs nll_max.
Ensemble size 3
acc_imp_avg
nll_max
0.00
−0.25
−0.50
9000
10000
11000
12000
13000
14000
15000
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

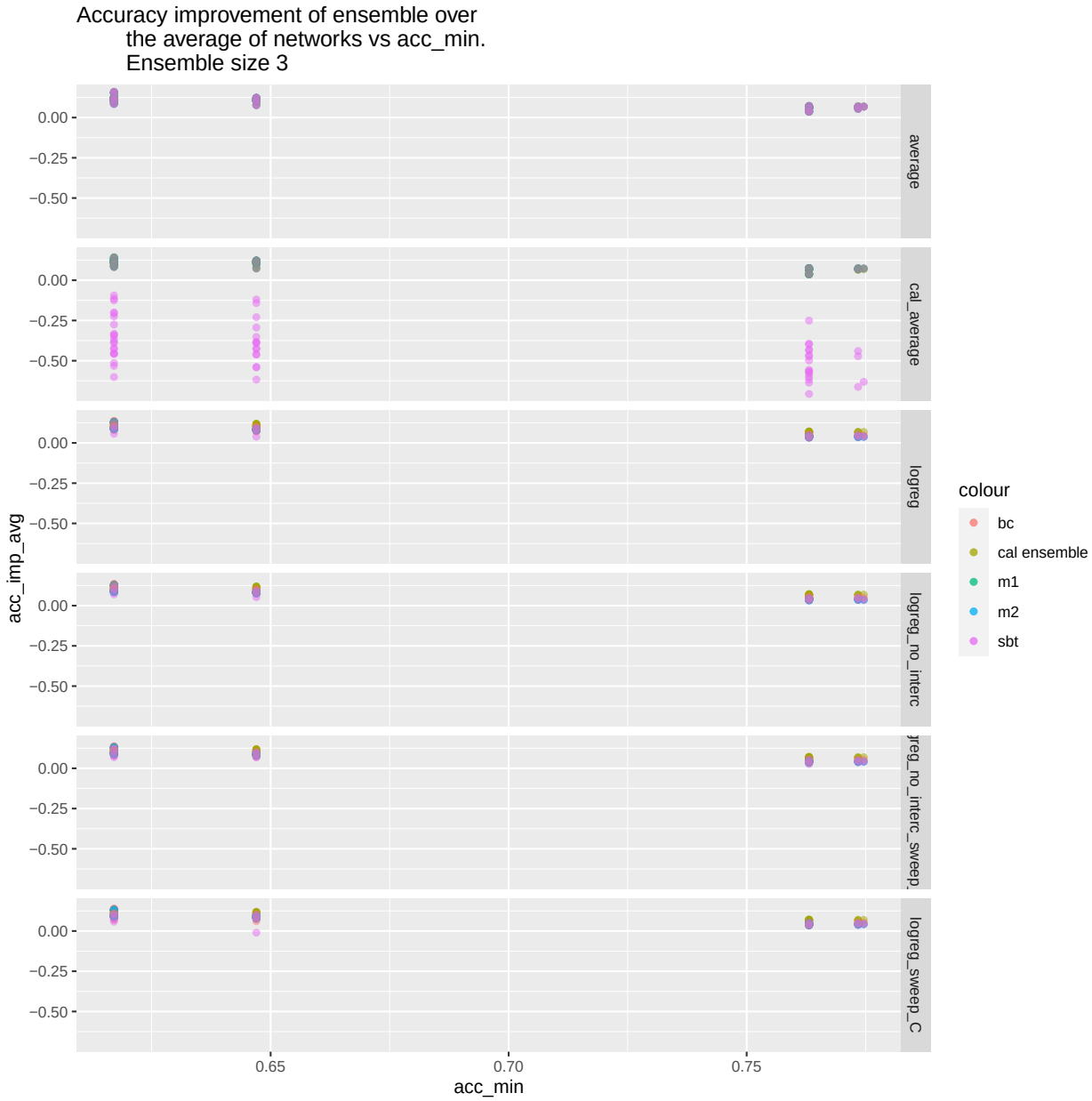

Accuracy improvement of ensemble over
the best of networks vs nll_avg.
Ensemble size 3
acc_imp_max
nll_avg
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt


Accuracy improvement of ensemble over
the average of networks vs nll_avg.
Ensemble size 3
acc_imp_avg
nll_avg
0.00
−0.25
−0.50
8000
9000
10000
11000
12000
13000
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

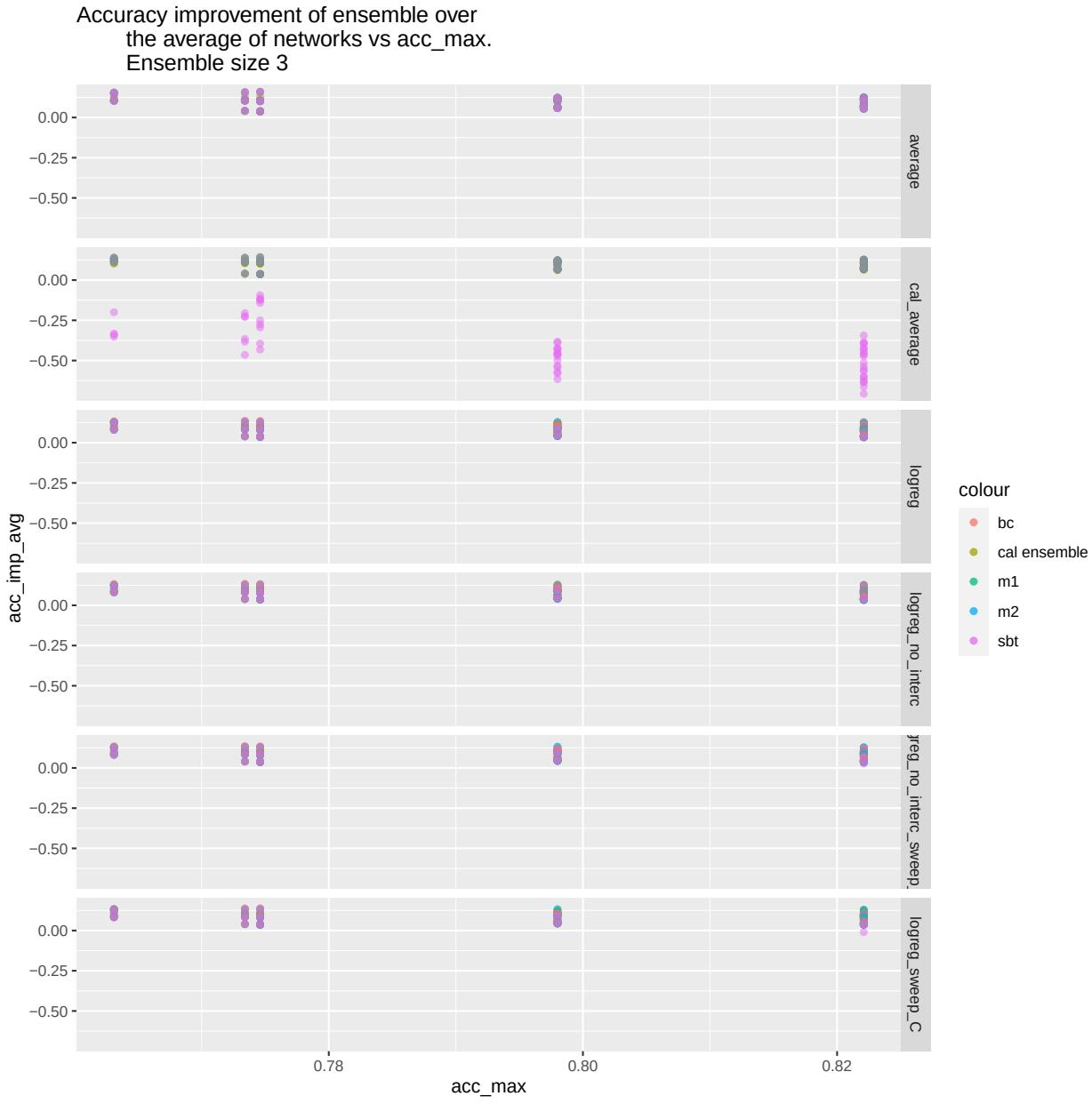

Accuracy improvement of ensemble over
the best of networks vs acc_var.
Ensemble size 3
acc_imp_max
acc_var
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.0
−0.2
−0.4
−0.6
0.000
0.004
0.008
0.012
colour
bc
cal ensemble
m1
m2
sbt

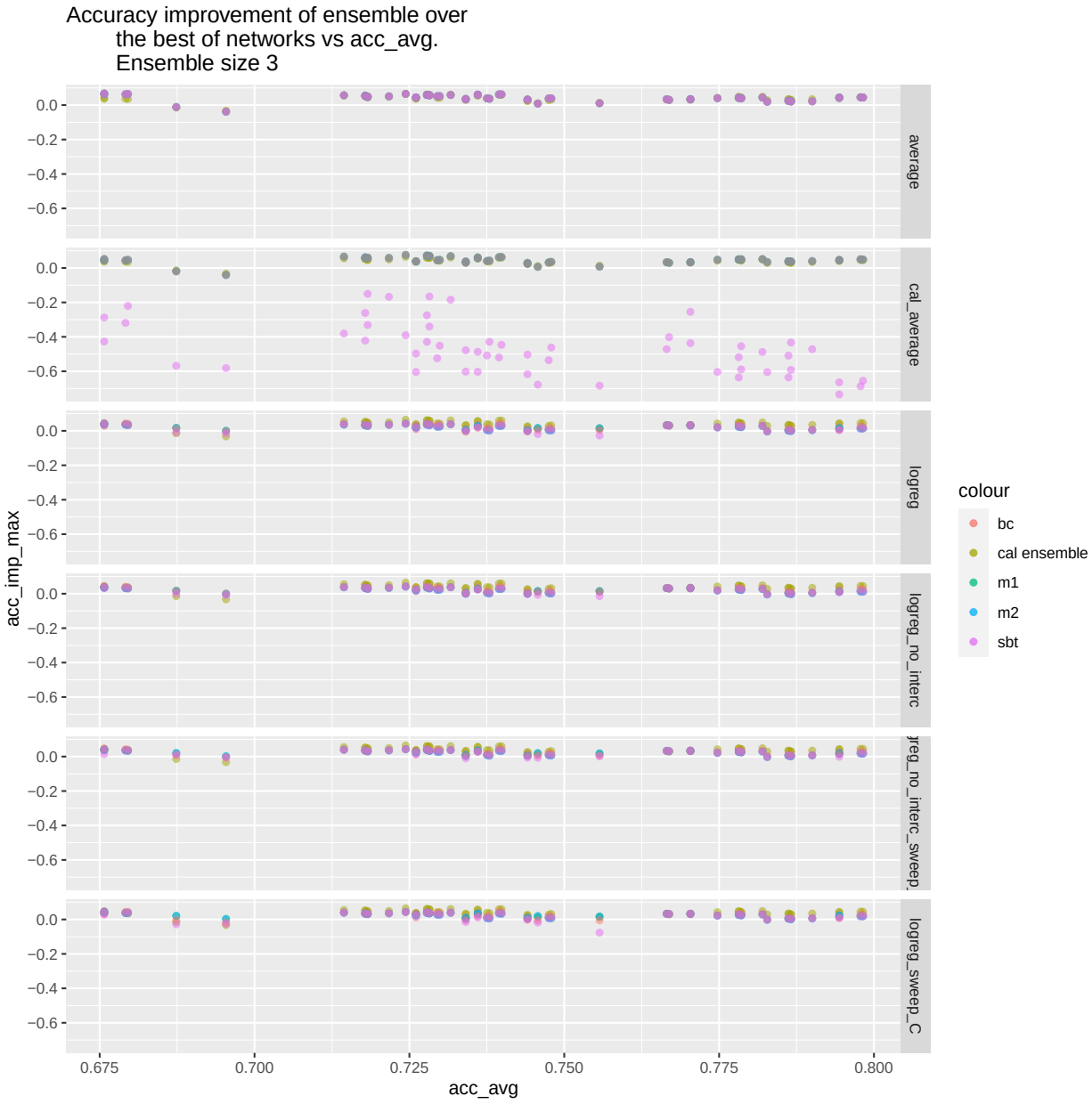

Accuracy improvement of ensemble over
the average of networks vs acc_var.
Ensemble size 3
acc_imp_avg
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.000
0.004
0.008
0.012
acc_var
colour
bc
cal ensemble
m1
m2
sbt

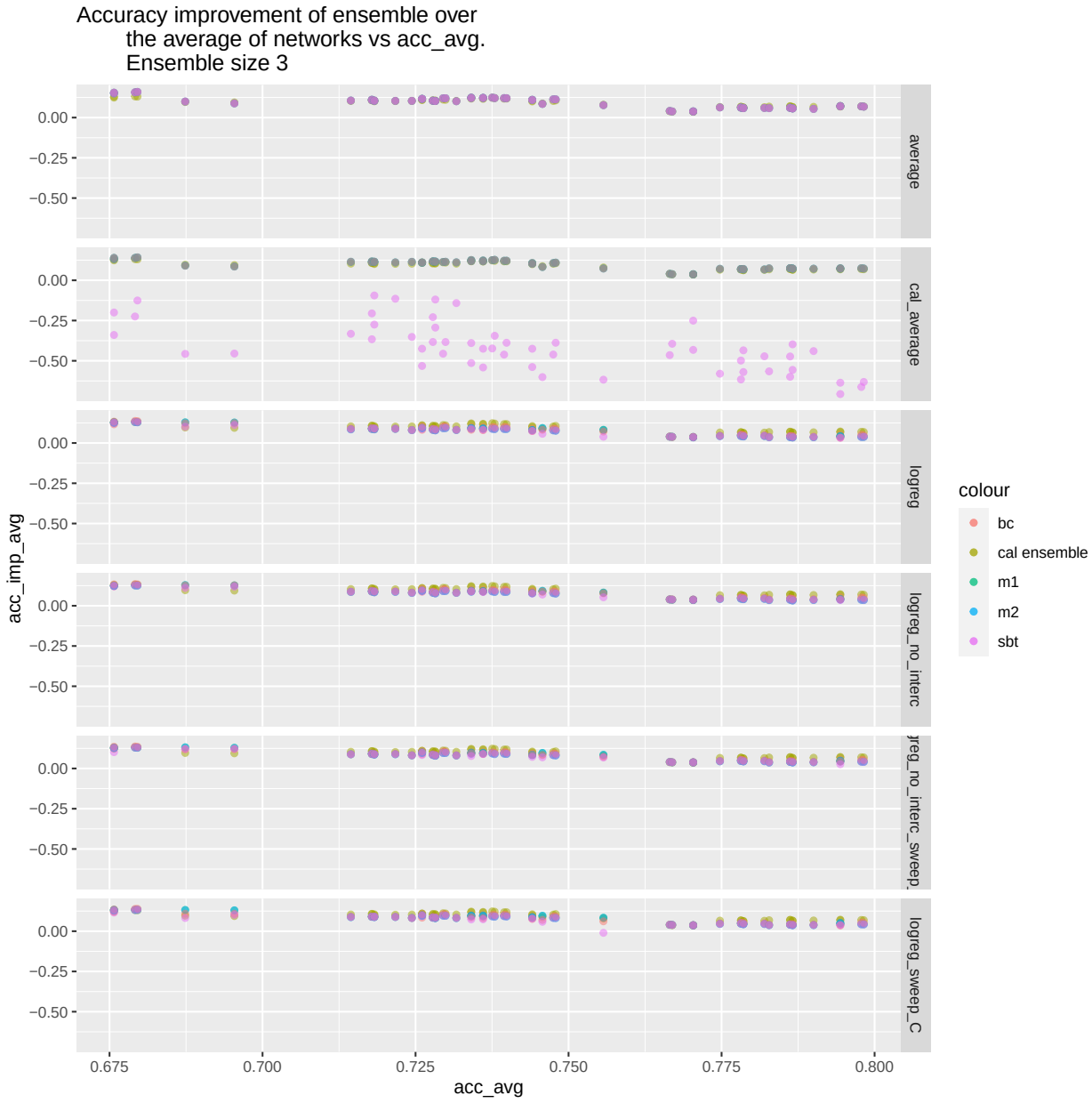

Accuracy improvement of ensemble over
the best of networks vs acc_min.
Ensemble size 3
acc_imp_max
0.0
−0.2
−0.4
−0.6
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.65
0.70
0.75
acc_min
colour
bc
cal ensemble
m1
m2
sbt

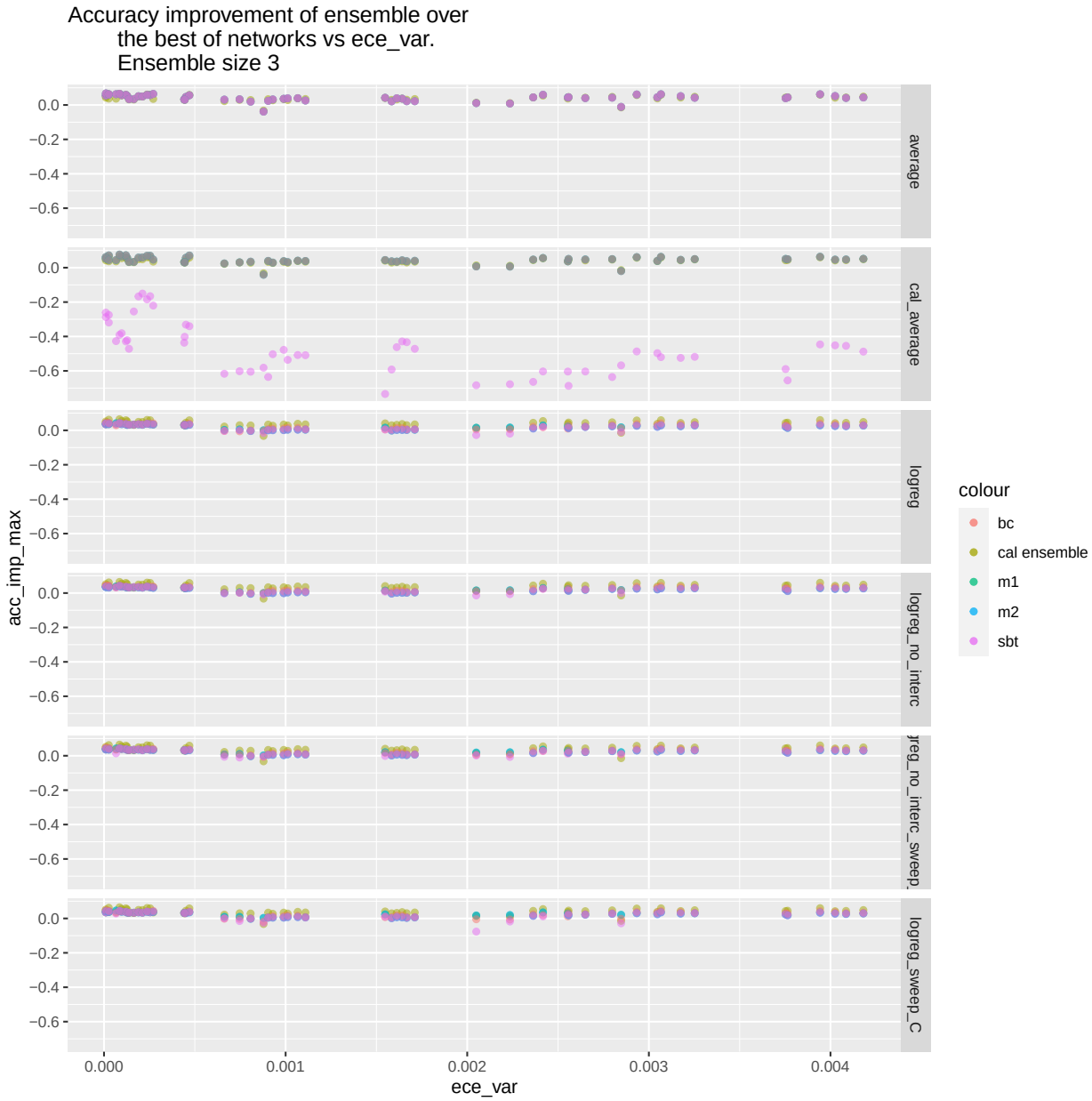

Accuracy improvement of ensemble over
the average of networks vs acc_min.
Ensemble size 3
acc_imp_avg
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.65
0.70
0.75
acc_min
colour
bc
cal ensemble
m1
m2
sbt

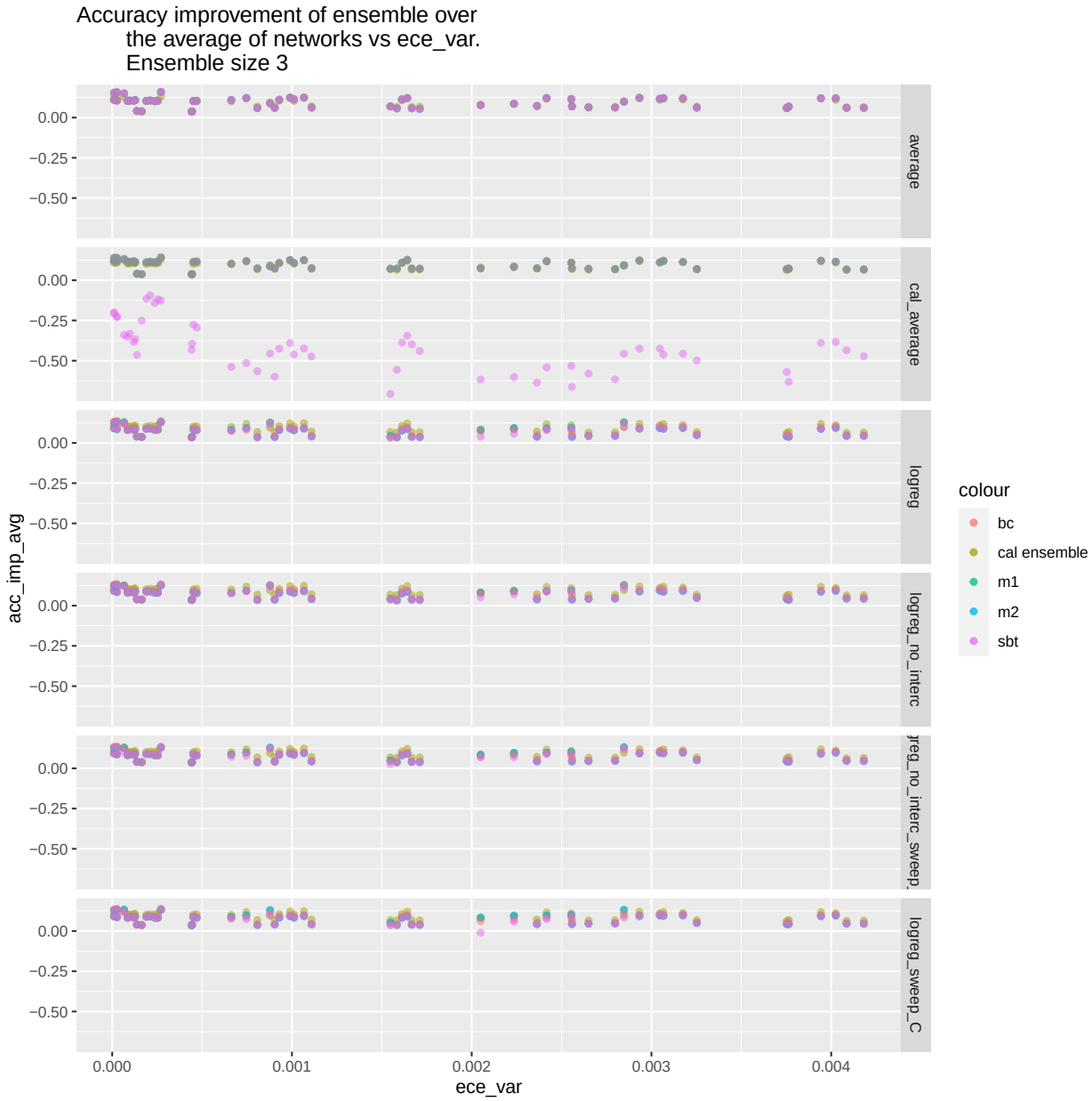

Accuracy improvement of ensemble over
the best of networks vs acc_max.
Ensemble size 3
acc_imp_max
0.0
−0.2
−0.4
−0.6
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.78
0.80
0.82
acc_max
colour
bc
cal ensemble
m1
m2
sbt

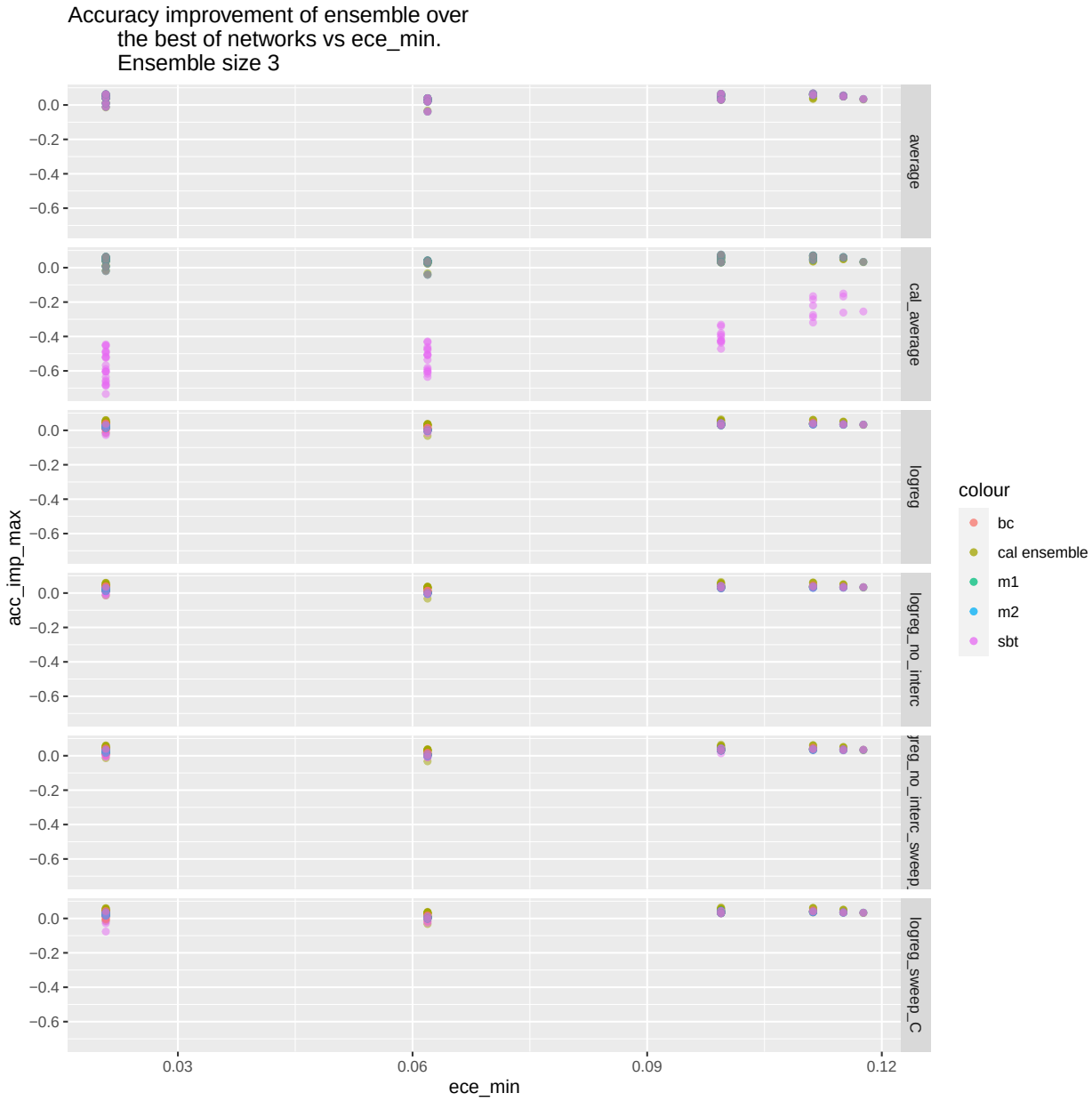

Accuracy improvement of ensemble over
the average of networks vs acc_max.
Ensemble size 3
acc_imp_avg
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
0.78
0.80
0.82
acc_max
colour
bc
cal ensemble
m1
m2
sbt

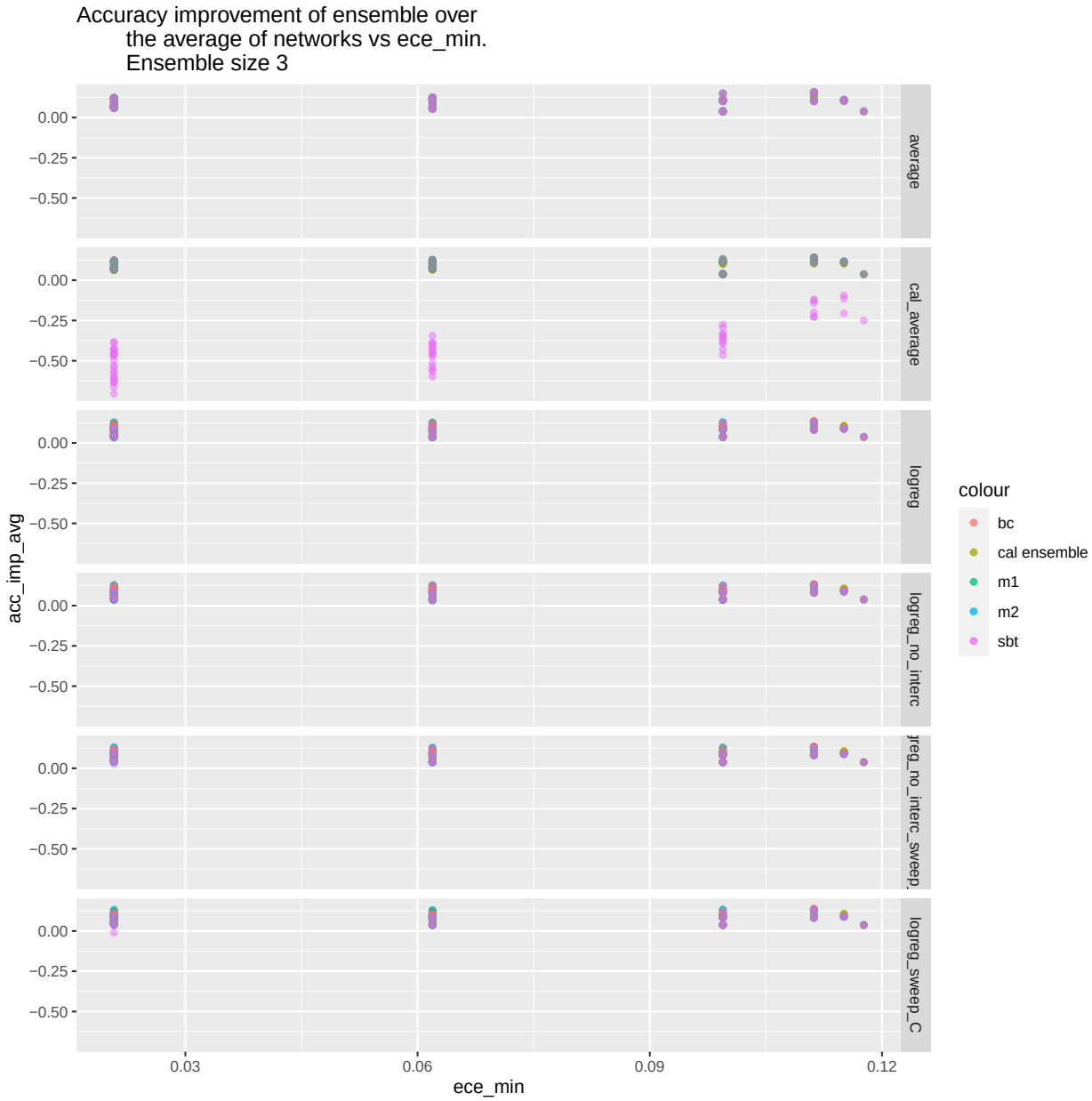

Accuracy improvement of ensemble over
the best of networks vs acc_avg.
Ensemble size 3
acc_imp_max
acc_avg
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
0.0
−0.2
−0.4
−0.6
0.675
0.700
0.725
0.750
0.775
0.800
colour
bc
cal ensemble
m1
m2
sbt

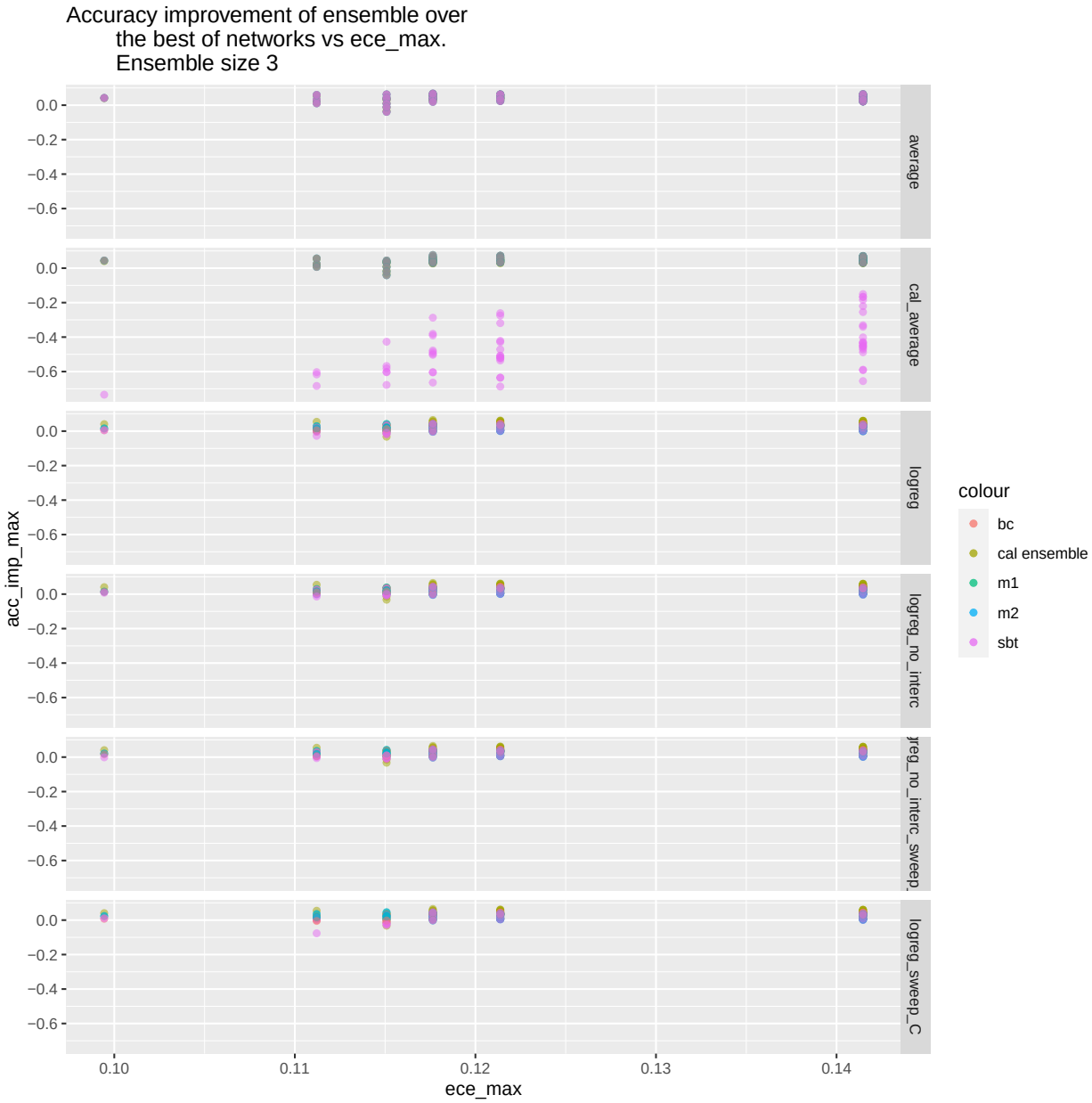

Accuracy improvement of ensemble over
the average of networks vs acc_avg.
Ensemble size 3
acc_imp_avg
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.675
0.700
0.725
0.750
0.775
0.800
acc_avg
colour
bc
cal ensemble
m1
m2
sbt

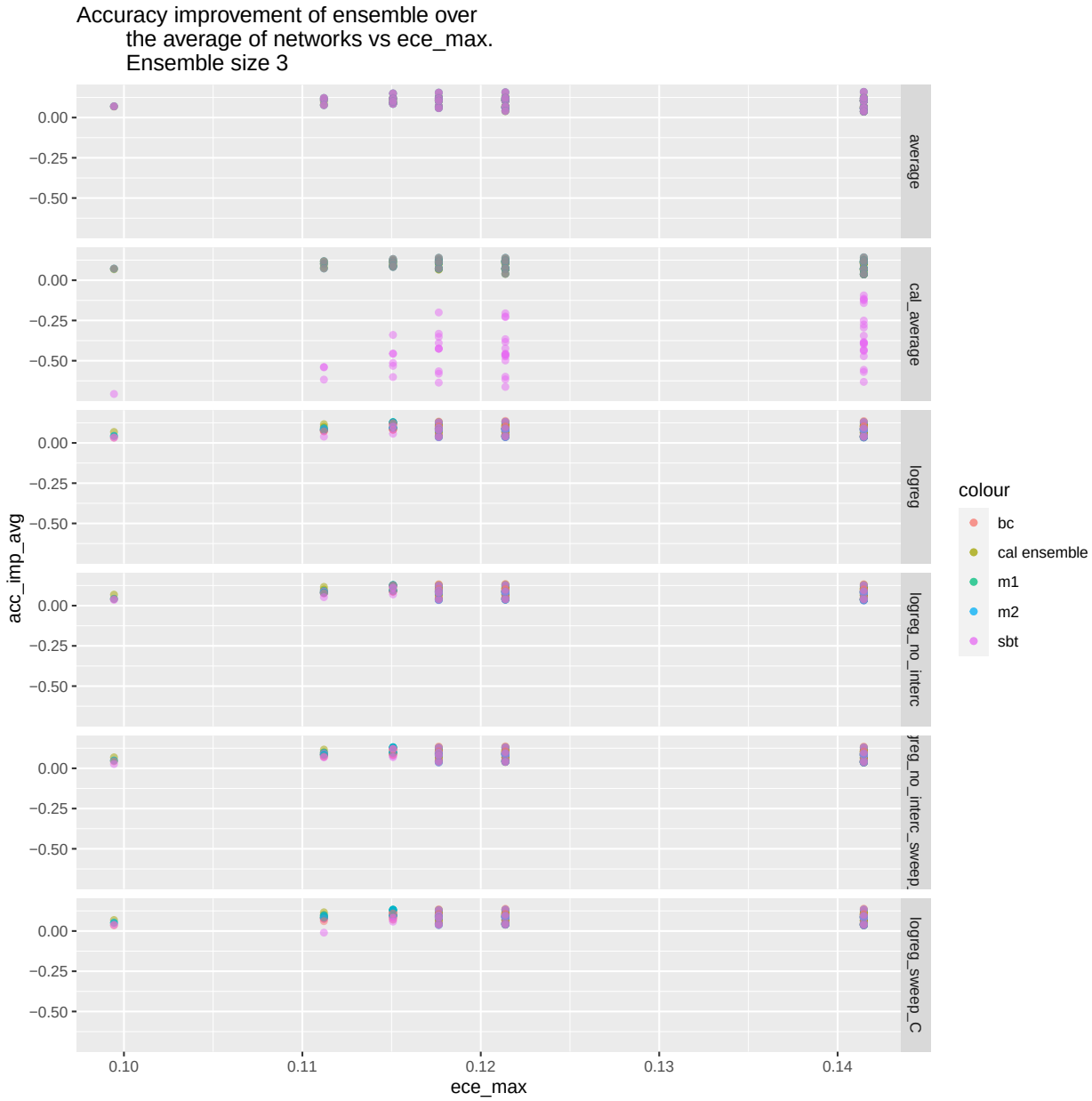

Accuracy improvement of ensemble over
the best of networks vs ece_var.
Ensemble size 3
acc_imp_max
ece_var
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt

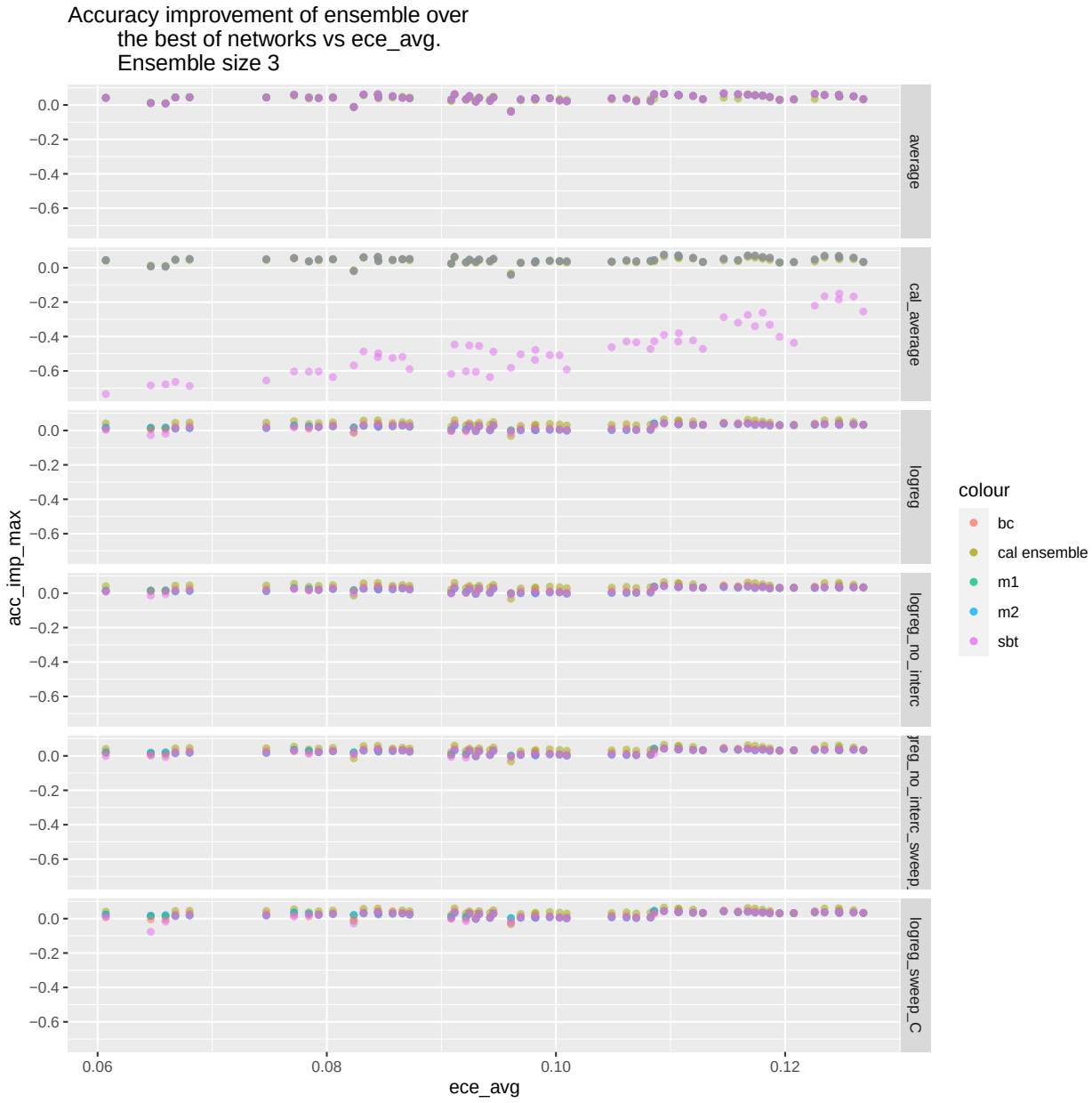

Accuracy improvement of ensemble over
the average of networks vs ece_var.
Ensemble size 3
acc_imp_avg
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
0.000
0.001
0.002
0.003
0.004
ece_var
colour
bc
cal ensemble
m1
m2
sbt

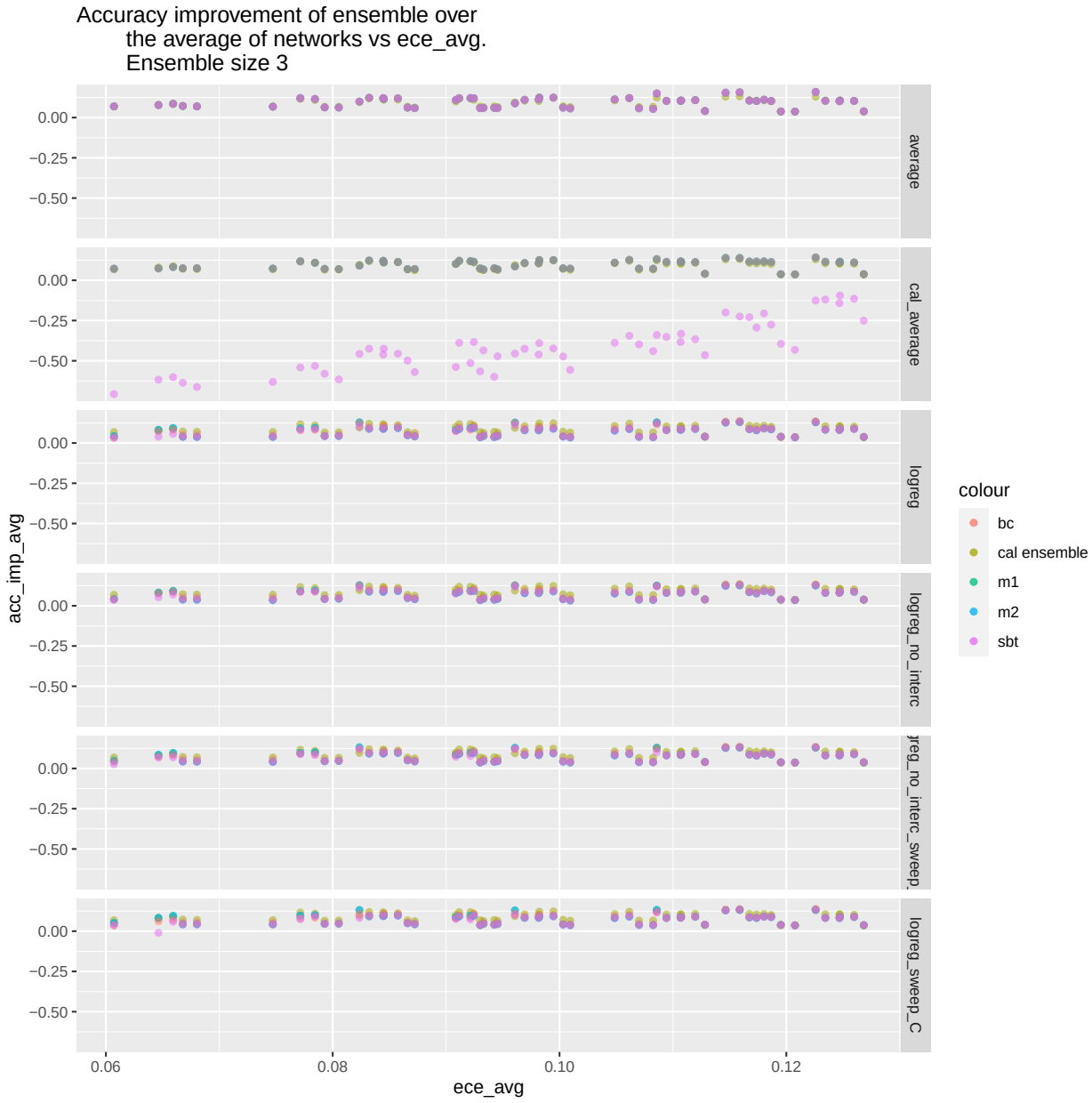

Accuracy improvement of ensemble over
the best of networks vs ece_min.
Ensemble size 3
acc_imp_max
ece_min
0.03
0.06
0.09
0.12
0.0
−0.2
−0.4
−0.6
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
colour
bc
cal ensemble
m1
m2
sbt


Accuracy improvement of ensemble over
the average of networks vs ece_min.
Ensemble size 3
acc_imp_avg
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.03
0.06
0.09
0.12
ece_min
colour
bc
cal ensemble
m1
m2
sbt

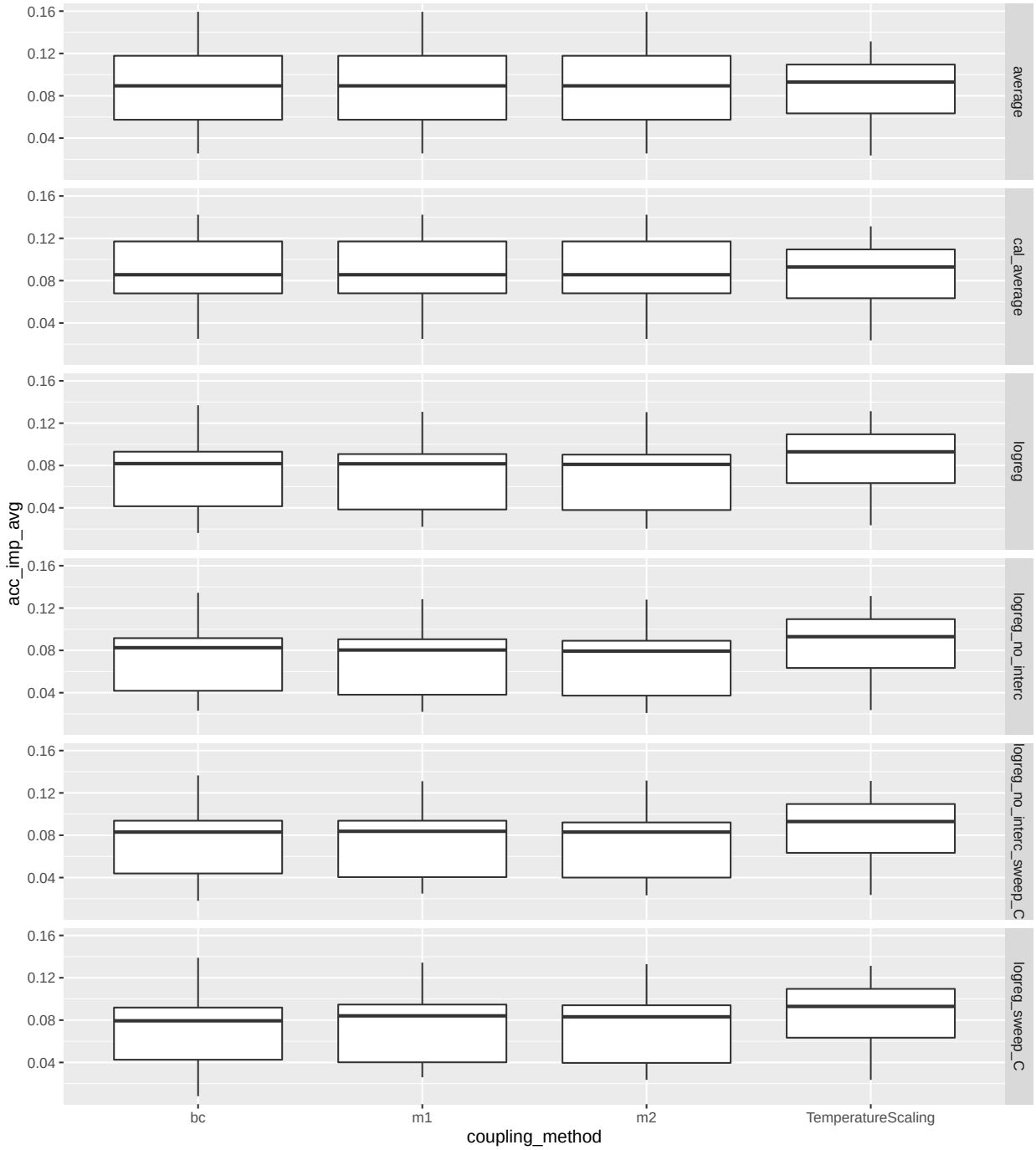

Accuracy improvement of ensemble over
the best of networks vs ece_max.
Ensemble size 3
acc_imp_max
ece_max
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
0.0
−0.2
−0.4
−0.6
0.10
0.11
0.12
0.13
0.14
colour
bc
cal ensemble
m1
m2
sbt


Accuracy improvement of ensemble over
the average of networks vs ece_max.
Ensemble size 3
acc_imp_avg
0.00
−0.25
−0.50
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_
logreg_sweep_C
0.10
0.11
0.12
0.13
0.14
ece_max
colour
bc
cal ensemble
m1
m2
sbt

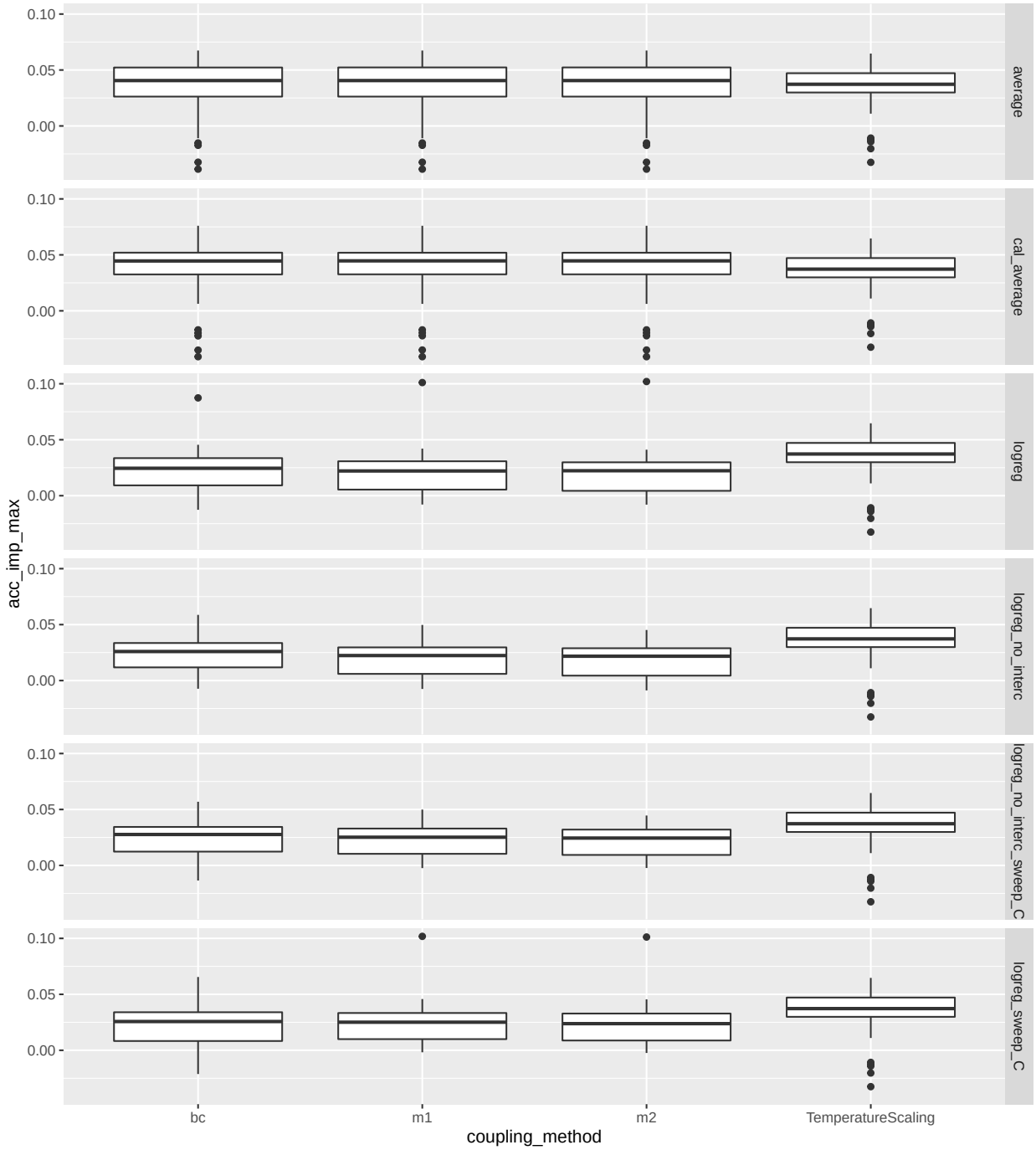

Accuracy improvement of ensemble over
the best of networks vs ece_avg.
Ensemble size 3
acc_imp_max
ece_avg
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep
logreg_sweep_C
0.0
−0.2
−0.4
−0.6
0.06
0.08
0.10
0.12
colour
bc
cal ensemble
m1
m2
sbt




```
imp_avg_plot <- ggplot() +
  geom_boxplot(data=ens_pwc_plt_df %>% filter(coupling_method != "sbt"), mapping=aes(x=coupling_method,
  geom_boxplot(data=ens_cal_plt_df, mapping=aes(x=calibrating_method, y=acc_imp_avg)) +
  facet_grid(rows=vars(combining_method))

print(imp_avg_plot)
```




```
imp_max_plot <- ggplot() +
  geom_boxplot(data=ens_pwc_plt_df %>% filter(coupling_method != "sbt"), mapping=aes(x=coupling_method,
  geom_boxplot(data=ens_cal_plt_df, mapping=aes(x=calibrating_method, y=acc_imp_max)) +
  facet_grid(rows=vars(combining_method))

print(imp_max_plot)
```


acc_imp_max
0.10
0.05
0.00
average
cal_average
logreg
logreg_no_interc
logreg_no_interc_sweep_C
logreg_sweep_C
bc
m1
m2
TemperatureScaling
coupling_method