main

September 22, 2021

1 Summary of prediction analysis for DE genes

[1]:
```
import os, errno
import pandas as pd
import seaborn as sns
from venn import venn
import matplotlib.pyplot as plt
```

1.1 Functions

[2]:
```
def mkdir_p(directory):
    """
    Make a directory if it does not already exist.

    Input: Directory name
    """
    try:
        os.makedirs(directory)
    except OSError as e:
        if e.errno != errno.EEXIST:
            raise
```

1.2 Summary of features

[3]:
```
degs = pd.read_csv("../../../../_m/degs_annotation.txt", sep='\t', index_col=0)
dtu = pd.read_csv("../../../../../differential_analysis/tissue_comparison/
 ↪ds_summary/_m/diffSplicing_ancestry_FDR05_4regions.tsv", sep='\t')
dtu.groupby("Tissue").size()
```

[3]:
```
Tissue
Caudate          1901
DLPFC            1345
Dentate Gyrus     655
Hippocampus      1332
dtype: int64
```

[4]:
```
for tissue in ["Caudate", "Dentate Gyrus", "DLPFC", "Hippocampus"]:
    overlap = len(set(degs[(degs["Tissue"] == tissue)].gene_name) &
                  set(dtu[(dtu["Tissue"] == tissue)].gene))
    print("There are {} overlapping DTU in DEGs for {}!".format(overlap,␣
 ↪tissue))
```

```
There are 385 overlapping DTU in DEGs for Caudate!
There are 51 overlapping DTU in DEGs for Dentate Gyrus!
There are 251 overlapping DTU in DEGs for DLPFC!
There are 227 overlapping DTU in DEGs for Hippocampus!
```

1.3 Load and prep summary files

1.3.1 Load files

[5]:
```
rf0 = pd.read_csv("../../rf/summary_10Folds_allTissues.tsv", sep='\t')
enet0 = pd.read_csv("../../enet/summary_10Folds_allTissues.tsv", sep='\t')
```

1.3.2 Group, select, and clean summary results

[6]:
```
## Extract median of model metrics over 10 folds
rf = rf0.groupby(["tissue", "feature"]).median()\
        .loc[:, ["n_features", "test_score_r2"]].reset_index()
rf.feature = rf.feature.str.replace("_", ".", regex=True)
rf["Model"] = "Random Forest"
enet = enet0.groupby(["tissue", "feature"]).median()\
        .loc[:, ["n_features", "test_score_r2"]].reset_index()
enet.feature = enet.feature.str.replace("_", ".", regex=True)
enet["Model"] = "Elastic Net"

df = pd.concat([rf, enet], axis=0)
df.head(2)
```

[6]:

	tissue	feature	n_features	test_score_r2	Model
0	Caudate	ENSG00000003249.13	38.0	-0.010891	Random Forest
1	Caudate	ENSG00000003509.15	2.5	-0.039188	Random Forest

1.3.3 Overlap with DTU

[7]:
```
dx = df.merge(degs[["gene_name"]], left_on="feature", right_index=True).
 ↪drop_duplicates()
for tissue in ["Caudate", "Dentate Gyrus", "DLPFC", "Hippocampus"]:
    overlap = len(set(dx[(dx["tissue"] == tissue)].gene_name) &
                  set(dtu[(dtu["Tissue"] == tissue)].gene))
    print("There are {} overlapping DTU in DEGs for {}!".format(overlap,␣
 ↪tissue))
```

```
There are 376 overlapping DTU in DEGs for Caudate!
There are 45 overlapping DTU in DEGs for Dentate Gyrus!
```

```
There are 242 overlapping DTU in DEGs for DLPFC!
There are 219 overlapping DTU in DEGs for Hippocampus!
```

1.3.4 Add partial r2 results

[8]:
```
partial = pd.read_csv("../../partial_r2/rf_partial_r2_metrics.tsv", sep='\t')\
            .rename(columns={"Geneid": "Feature"})
partial.columns = partial.columns.str.lower()
partial["test_score_r2"] = partial.partial_r2
partial["Model"] = "Partial R2"
partial = partial.loc[:, ['tissue', 'feature', 'n_features', 'test_score_r2',␣
 ↪'Model']]
partial.head(2)
```

[8]:

	tissue	feature	n_features	test_score_r2	Model
0	Caudate	ENSG00000003249.13	38	0.297444	Partial R2
1	Caudate	ENSG00000003509.15	2	0.001916	Partial R2

[9]:
```
df2 = pd.concat([df, partial], axis=0)
df2.groupby(["tissue", "Model"]).size()
```

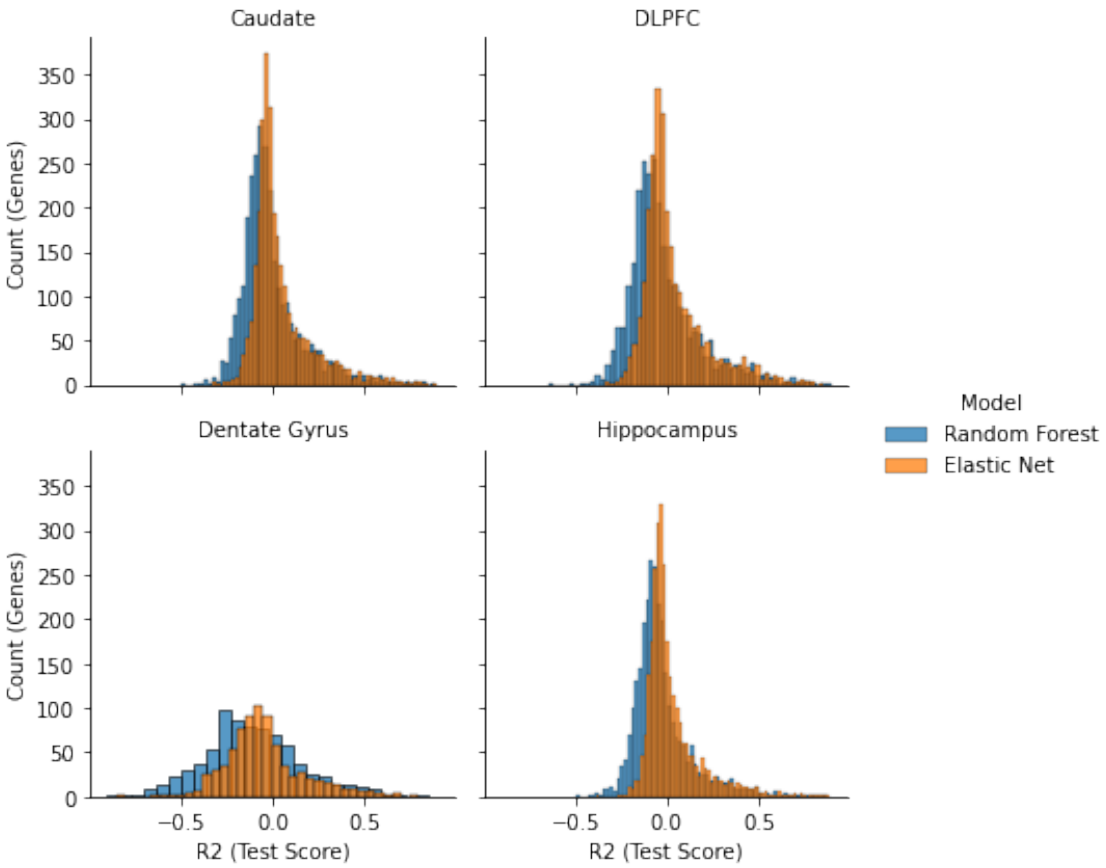
[9]:
```
tissue         Model
Caudate        Elastic Net      2929
               Partial R2       2925
               Random Forest    2929
DLPFC          Elastic Net      2711
               Partial R2       2691
               Random Forest    2691
Dentate Gyrus  Elastic Net       773
               Partial R2        773
               Random Forest     773
Hippocampus    Elastic Net      2911
               Partial R2       2906
               Random Forest    2911
dtype: int64
```

1.4 Summary of results

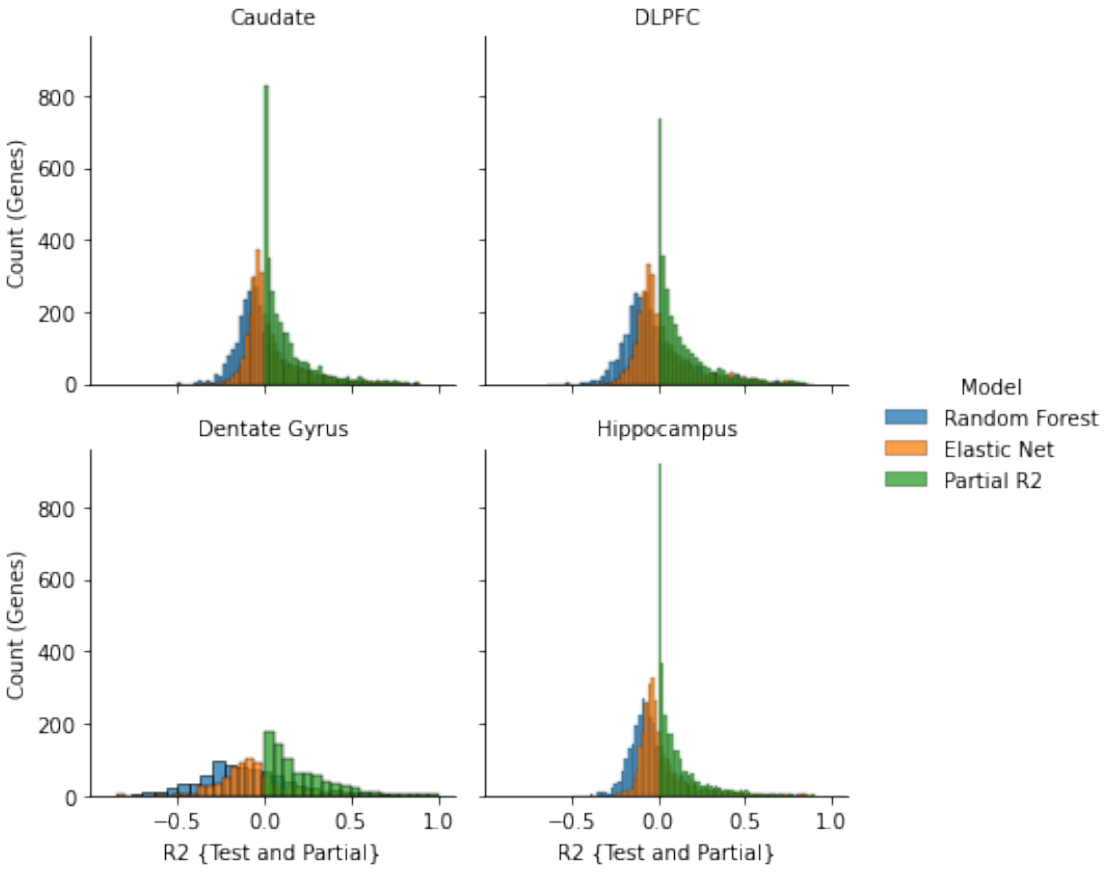
1.4.1 Histogram of R2 (median test R2 score)

[10]:
```
grid = sns.FacetGrid(df, col="tissue", col_wrap=2, hue="Model")
grid.map(sns.histplot, "test_score_r2")
grid.set_axis_labels("R2 (Test Score)", "Count (Genes)")
grid.set_titles(col_template="{col_name}")
grid.add_legend()
grid.tight_layout()
grid.savefig("histogram_test_r2.pdf")
grid.savefig("histogram_test_r2.png")
```

```
grid.savefig("histogram_test_r2.svg")
```



[11]:
```
grid = sns.FacetGrid(df2, col="tissue", col_wrap=2, hue="Model")
grid.map(sns.histplot, "test_score_r2")
grid.set_axis_labels("R2 {Test and Partial}", "Count (Genes)")
grid.set_titles(col_template="{col_name}")
grid.add_legend()
grid.tight_layout()
grid.savefig("histogram_test_N_partial_r2.pdf")
grid.savefig("histogram_test_N_partial_r2.png")
grid.savefig("histogram_test_N_partial_r2.svg")
```



1.4.2 What number of DEGs do not have any SNPs within 20 Kbp of gene body?

[12]:
```
for tissue in ["Caudate", "DLPFC", "Hippocampus", "Dentate Gyrus"]:
    xx = set(df[(df["tissue"] == tissue)].feature)
    yy = set(degs[(degs["Tissue"] == tissue)].index)
    txt = "{} of {} ({:.1%}) of DE genes do not have SNPs within 20Kbp."
    print(txt.format(len(yy) - len(xx), len(yy), (len(yy) - len(xx)) / len(yy)))
```

```
41 of 2970 (1.4%) of DE genes do not have SNPs within 20Kbp.
49 of 2760 (1.8%) of DE genes do not have SNPs within 20Kbp.
45 of 2956 (1.5%) of DE genes do not have SNPs within 20Kbp.
13 of 786 (1.7%) of DE genes do not have SNPs within 20Kbp.
```

1.4.3 Number of ancestry DE genes expression that can be predictive with SNP

[13]:
```
df[(df["test_score_r2"] >= 0.5)].groupby(["tissue", "Model"]).size()
```

[13]:
```
tissue         Model
Caudate        Elastic Net      91
               Random Forest    68
```

```
DLPFC          Elastic Net      80
               Random Forest    70
Dentate Gyrus  Elastic Net      18
               Random Forest    16
Hippocampus    Elastic Net      55
               Random Forest    46
dtype: int64
```

[14]:
```
df[(df["test_score_r2"] >= 0.75)].groupby(["tissue", "Model"]).size()
```

[14]:
```
tissue         Model
Caudate        Elastic Net      11
               Random Forest     9
DLPFC          Elastic Net       9
               Random Forest    10
Dentate Gyrus  Elastic Net       1
               Random Forest     1
Hippocampus    Elastic Net      10
               Random Forest    12
dtype: int64
```

[15]:
```
print(df[(df["test_score_r2"] >= 0.85)].groupby(["tissue", "Model"]).size().
 ↪reset_index())
df[(df["test_score_r2"] >= 0.85)]
```

```
          tissue          Model  0
0        Caudate    Elastic Net  2
1        Caudate  Random Forest  2
2          DLPFC    Elastic Net  1
3          DLPFC  Random Forest  2
4  Dentate Gyrus  Random Forest  1
5    Hippocampus    Elastic Net  2
6    Hippocampus  Random Forest  2
```

[15]:
```
             tissue             feature  n_features  test_score_r2  \
34          Caudate  ENSG00000013573.16        16.0       0.880205
1313        Caudate  ENSG00000166435.15        28.5       0.856982
4219          DLPFC  ENSG00000166435.15        39.5       0.855143
4936          DLPFC   ENSG00000226278.1        20.5       0.897793
6185  Dentate Gyrus   ENSG00000226278.1        84.0       0.850423
7745    Hippocampus  ENSG00000166435.15        42.0       0.877993
8932    Hippocampus   ENSG00000256274.1         8.0       0.854404
34          Caudate  ENSG00000013573.16        32.5       0.891321
1313        Caudate  ENSG00000166435.15        31.5       0.875208
4219          DLPFC  ENSG00000166435.15        20.5       0.852089
6453    Hippocampus  ENSG00000013573.16        38.0       0.856511
7765    Hippocampus  ENSG00000166435.15        36.0       0.879696
```

```
              Model
34    Random Forest
1313  Random Forest
4219  Random Forest
4936  Random Forest
6185  Random Forest
7745  Random Forest
8932  Random Forest
34      Elastic Net
1313    Elastic Net
4219    Elastic Net
6453    Elastic Net
7765    Elastic Net
```

[16]:
```
set(df[(df["test_score_r2"] >= 0.85)].feature)
```

[16]:
```
{'ENSG00000013573.16',
 'ENSG00000166435.15',
 'ENSG00000226278.1',
 'ENSG00000256274.1'}
```

- **ENSG00000166435.15** is *XRRA1* one of the most significant eQTLs in the brain
- **ENSG00000013573.16** is *DDX11*
- **ENSG00000226278.1** is *PSPHP1* a pseudogene
- **ENSG00000256274.1** is *TAS2R64P* anohter pseudogene

[17]:
```
print(df[(df["test_score_r2"] >= 0.9)].groupby(["tissue", "Model"]).size().
 ↪reset_index())
df[(df["test_score_r2"] >= 0.9)]
```

```
Empty DataFrame
Columns: [tissue, Model, 0]
Index: []
```

[17]:
```
Empty DataFrame
Columns: [tissue, feature, n_features, test_score_r2, Model]
Index: []
```

1.4.4 Overlapping with DTU

[18]:
```
df3 = dx.merge(dtu, left_on=["gene_name", "tissue"], right_on=["gene",␣
 ↪"Tissue"])
df3[(df3["test_score_r2"] >= 0.5)].groupby(["Tissue", "Model"]).size()
```

[18]:
```
Tissue         Model
Caudate        Elastic Net      21
               Random Forest    11
DLPFC          Elastic Net      15
```

```
                Random Forest    14
Dentate Gyrus   Elastic Net       1
                Random Forest     1
Hippocampus     Elastic Net      12
                Random Forest     8
dtype: int64
```

[19]:
```
df3[(df3["test_score_r2"] >= 0.75)].groupby(["Tissue", "Model"]).size()
```

[19]:
```
Tissue  Model
DLPFC   Elastic Net      4
        Random Forest    3
dtype: int64
```

[20]:
```
df3[(df3["test_score_r2"] >= 0.75)]
```

[20]:

	tissue	feature	n_features	test_score_r2	Model	gene_name	clusterID	N	coord	gene	annotation	FDR	chr	Type	Tissue
90	DLPFC	ENSG00000074803.17	3.5	0.777571	Random Forest	SLC12A1	clu_134504_+	14	chr15:48178400-48220634	SLC12A1	cryptic	9.440000e-07	chr15	DTU	DLPFC
91	DLPFC	ENSG00000074803.17	3.0	0.777366	Elastic Net	SLC12A1	clu_134504_+	14	chr15:48178400-48220634	SLC12A1	cryptic	9.440000e-07	chr15	DTU	DLPFC
649	DLPFC	ENSG00000147403.16	12.0	0.757679	Elastic Net	RPL10	clu_62948_+	5	chrX:154399941-154400702	RPL10	cryptic	5.130000e-04	chrX	DTU	DLPFC
842	DLPFC	ENSG00000166435.15	39.5	0.855143	Random Forest	XRRA1	clu_6194_-	14	chr11:74848462-74907145	XRRA1	cryptic	4.430000e-05	chr11	DTU	DLPFC
843	DLPFC	ENSG00000166435.15	20.5	0.852089	Elastic Net	XRRA1	clu_6194_-	14	chr11:74848462-74907145	XRRA1	cryptic	4.430000e-05	chr11	DTU	DLPFC
1472	DLPFC	ENSG00000257218.5	40.0	0.750014	Random Forest	GATC	clu_47435_+	3	chr12:120446829-120457076	GATC	annotated	2.330000e-02	chr12	DTU	DLPFC
1473	DLPFC	ENSG00000257218.5	40.0	0.784960	Elastic Net	GATC	clu_47435_+	3	chr12:120446829-120457076	GATC	annotated	2.330000e-02	chr12	DTU	DLPFC

1.4.5 What is the overlap between models?

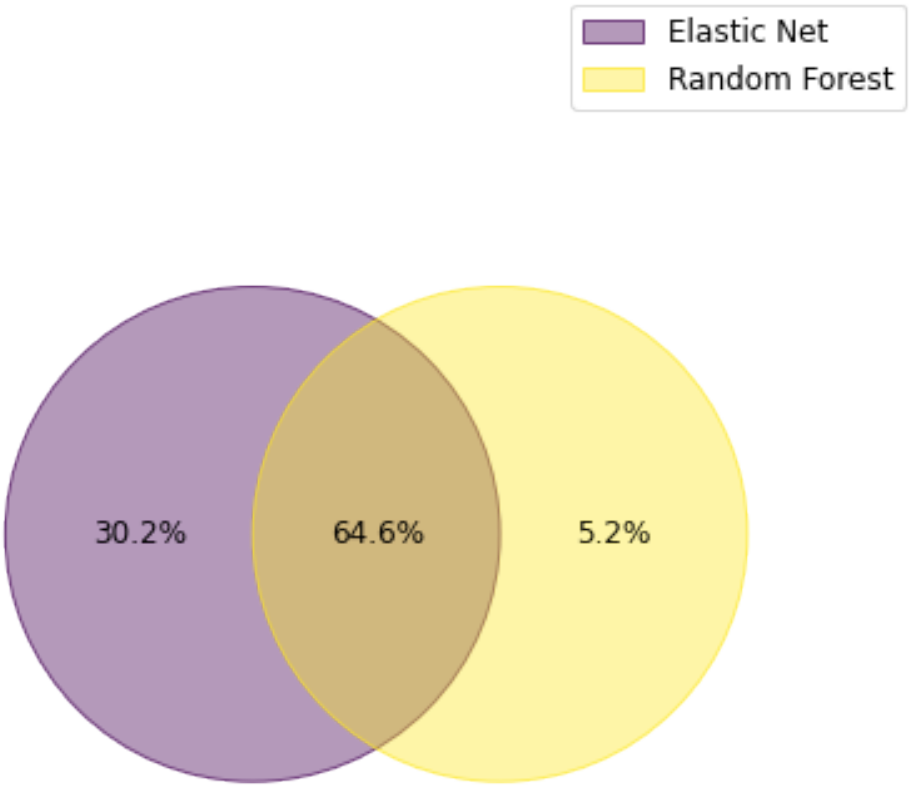
[21]:
```
for tissue in ["Caudate", "DLPFC", "Hippocampus", "Dentate Gyrus"]:
    print(tissue)
    for r2 in [0, 0.2, 0.5, 0.6, 0.7, 0.75, 0.8, 0.825]:
        ee = enet[(enet["tissue"] == tissue) & (enet["test_score_r2"] >= r2)].
↪copy()
        rr = rf[(rf["tissue"] == tissue) & (rf["test_score_r2"] >= r2)].copy()
        oo = len(set(ee.feature) & set(rr.feature))
        txt = "There is {} out of {} and {} genes overlapping between enet and␣
↪rf - at R2 > {}"
        print(txt.format(oo, len(set(ee.feature)), len(set(rr.feature)), r2))
    print("")
```

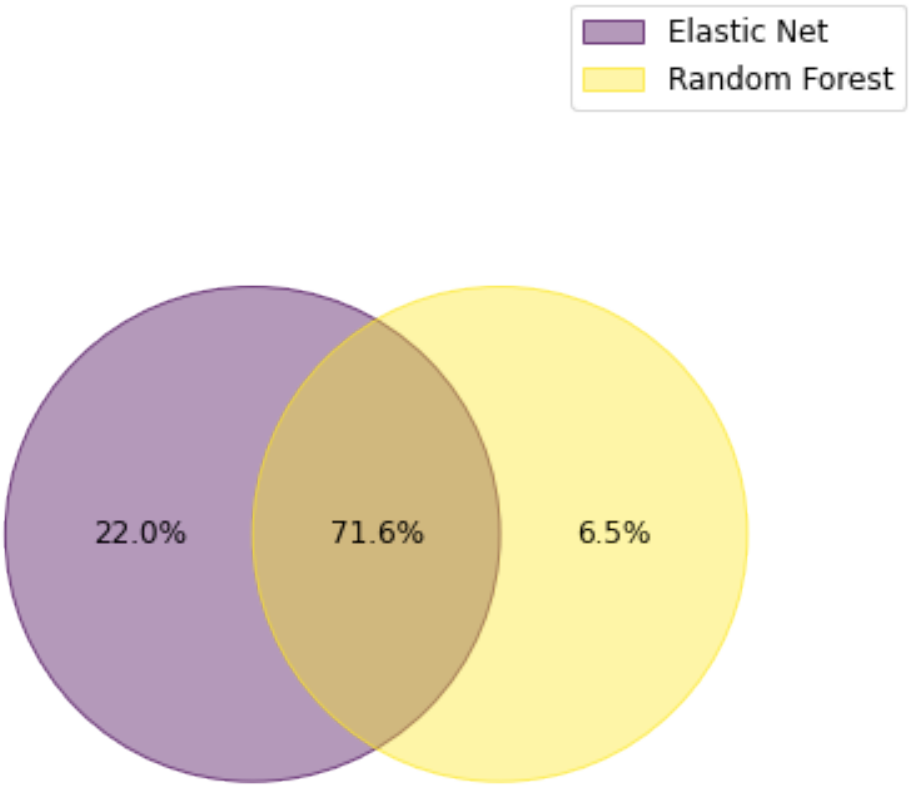
```
Caudate
There is 923 out of 1354 and 997 genes overlapping between enet and rf - at R2 >
0
There is 332 out of 434 and 362 genes overlapping between enet and rf - at R2 >
0.2
There is 65 out of 91 and 68 genes overlapping between enet and rf - at R2 > 0.5
There is 33 out of 52 and 35 genes overlapping between enet and rf - at R2 > 0.6
There is 15 out of 19 and 17 genes overlapping between enet and rf - at R2 > 0.7
There is 9 out of 11 and 9 genes overlapping between enet and rf - at R2 > 0.75
There is 4 out of 5 and 5 genes overlapping between enet and rf - at R2 > 0.8
There is 2 out of 3 and 4 genes overlapping between enet and rf - at R2 > 0.825

DLPFC
There is 835 out of 1204 and 904 genes overlapping between enet and rf - at R2 >
0
There is 293 out of 417 and 325 genes overlapping between enet and rf - at R2 >
0.2
There is 60 out of 80 and 70 genes overlapping between enet and rf - at R2 > 0.5
There is 28 out of 40 and 33 genes overlapping between enet and rf - at R2 > 0.6
There is 14 out of 20 and 14 genes overlapping between enet and rf - at R2 > 0.7
There is 6 out of 9 and 10 genes overlapping between enet and rf - at R2 > 0.75
There is 1 out of 1 and 2 genes overlapping between enet and rf - at R2 > 0.8
There is 1 out of 1 and 2 genes overlapping between enet and rf - at R2 > 0.825

Hippocampus
There is 780 out of 1206 and 852 genes overlapping between enet and rf - at R2 >
0
There is 252 out of 338 and 267 genes overlapping between enet and rf - at R2 >
0.2
There is 44 out of 55 and 46 genes overlapping between enet and rf - at R2 > 0.5
There is 26 out of 32 and 29 genes overlapping between enet and rf - at R2 > 0.6
There is 14 out of 18 and 15 genes overlapping between enet and rf - at R2 > 0.7
There is 9 out of 10 and 12 genes overlapping between enet and rf - at R2 > 0.75
There is 7 out of 8 and 8 genes overlapping between enet and rf - at R2 > 0.8
```

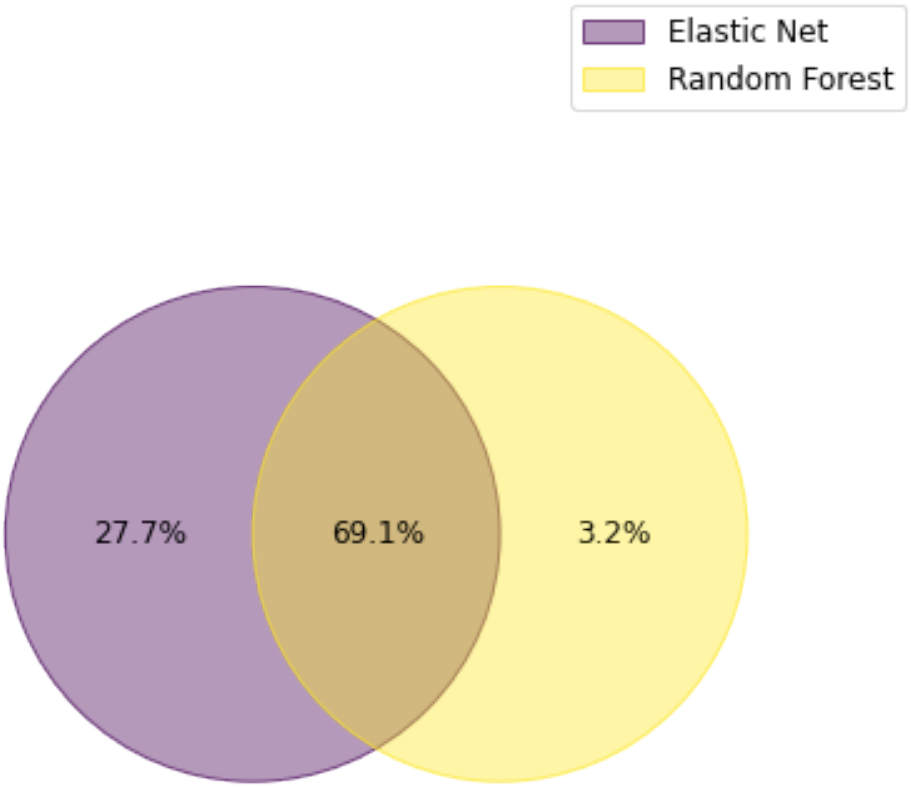
```
There is 4 out of 5 and 5 genes overlapping between enet and rf - at R2 > 0.825

Dentate Gyrus
There is 167 out of 237 and 237 genes overlapping between enet and rf - at R2 >
0
There is 72 out of 98 and 91 genes overlapping between enet and rf - at R2 > 0.2
There is 13 out of 18 and 16 genes overlapping between enet and rf - at R2 > 0.5
There is 5 out of 7 and 5 genes overlapping between enet and rf - at R2 > 0.6
There is 1 out of 1 and 3 genes overlapping between enet and rf - at R2 > 0.7
There is 1 out of 1 and 1 genes overlapping between enet and rf - at R2 > 0.75
There is 0 out of 0 and 1 genes overlapping between enet and rf - at R2 > 0.8
There is 0 out of 0 and 1 genes overlapping between enet and rf - at R2 > 0.825
```

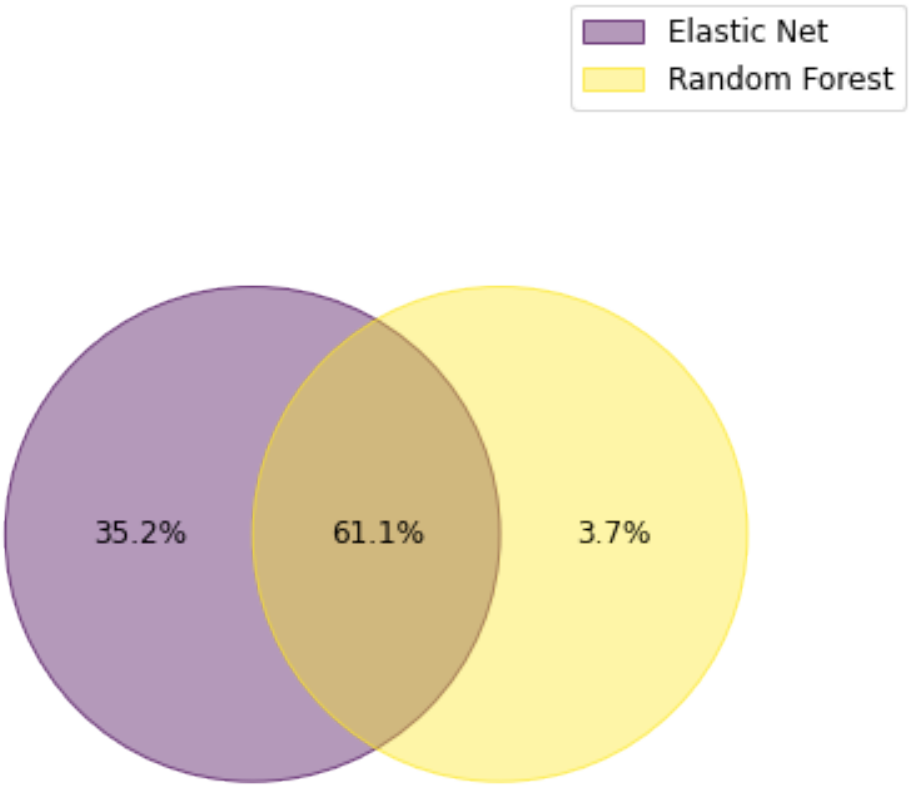
[22]:
```
dirname = "model_venn_diagrams"
mkdir_p(dirname)
for tissue in ["Caudate", "DLPFC", "Hippocampus", "Dentate Gyrus"]:
    #print(tissue)
    for r2 in [0, 0.2, 0.5, 0.6, 0.7, 0.75, 0.8]:
        ee = enet[(enet["tissue"] == tissue) & (enet["test_score_r2"] >= r2)].
↪copy()
        rr = rf[(rf["tissue"] == tissue) & (rf["test_score_r2"] >= r2)].copy()
        model_set = {"Elastic Net": set(ee.feature), "Random Forest": set(rr.
↪feature),}
        venn(model_set, fmt="{percentage:.1f}%", fontsize=12)
        tt = tissue.lower().replace(" ", "_")
        plt.savefig("{}/venn_diagram_modelOverlap_{}_r2_{}.png".format(dirname,␣
↪tt, r2))
        plt.savefig("{}/venn_diagram_modelOverlap_{}_r2_{}.pdf".format(dirname,␣
↪tt, r2))
        plt.savefig("{}/venn_diagram_modelOverlap_{}_r2_{}.svg".format(dirname,␣
↪tt, r2))
```

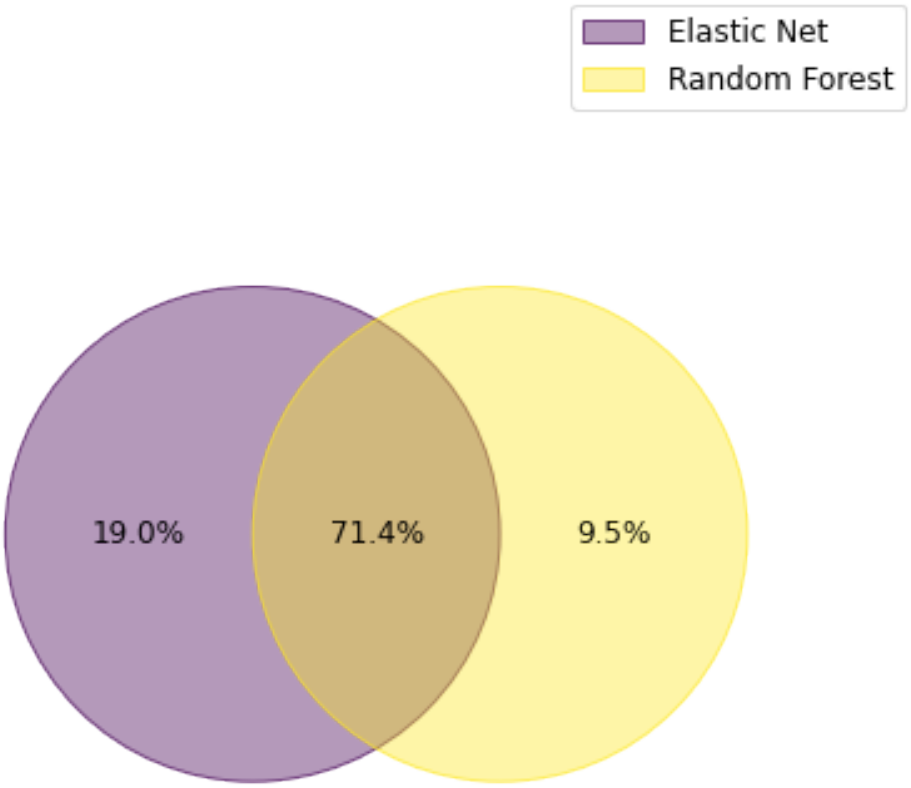
```
/home/jbenja13/.local/lib/python3.9/site-packages/venn/_venn.py:83:
RuntimeWarning: More than 20 figures have been opened. Figures created through
the pyplot interface (`matplotlib.pyplot.figure`) are retained until explicitly
closed and may consume too much memory. (To control this warning, see the
rcParam `figure.max_open_warning`).
  _, ax = subplots(nrows=1, ncols=1, figsize=figsize)
```

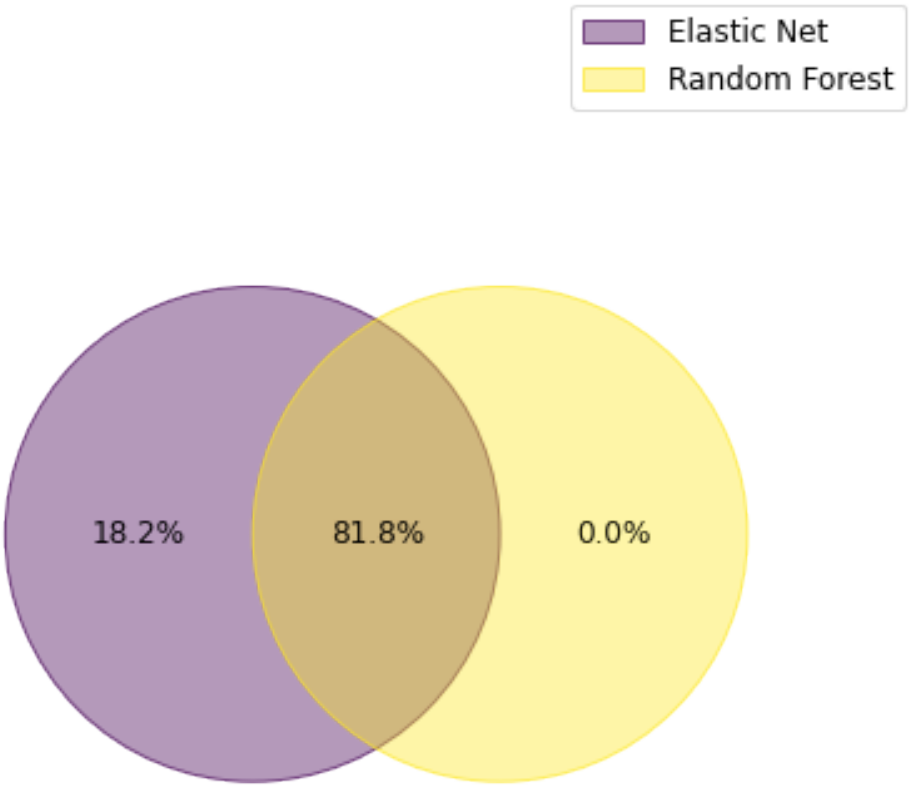

Elastic Net
Random Forest
30.2%
64.6%
5.2%

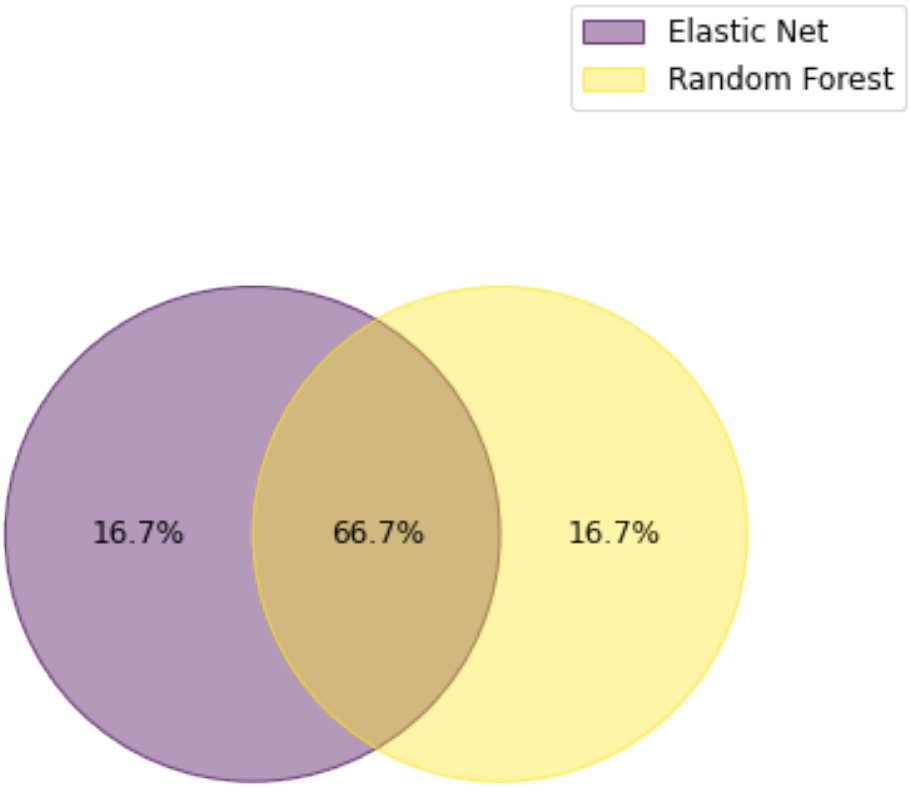

Elastic Net
Random Forest
22.0%
71.6%
6.5%

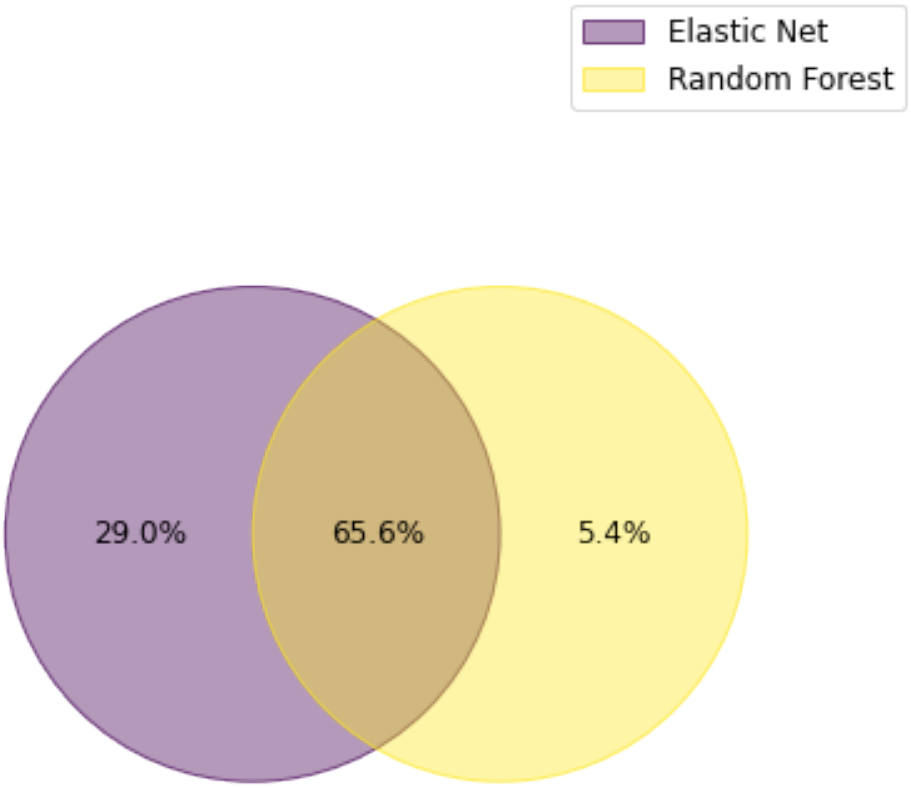

Elastic Net
Random Forest
27.7%
69.1%
3.2%

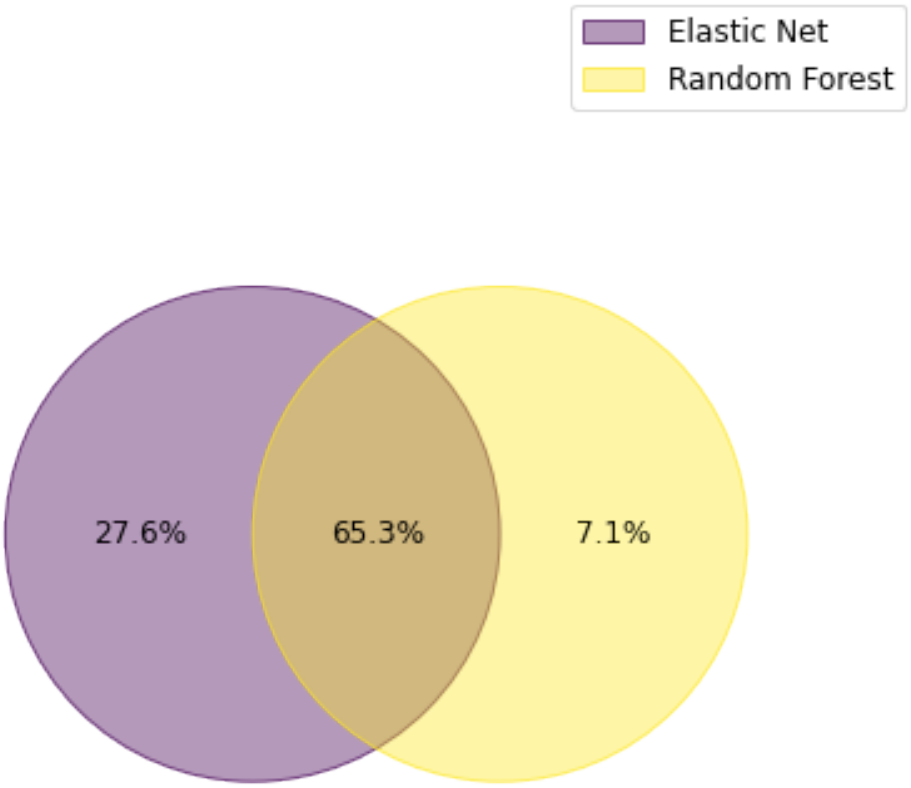

Elastic Net
Random Forest
35.2%
61.1%
3.7%

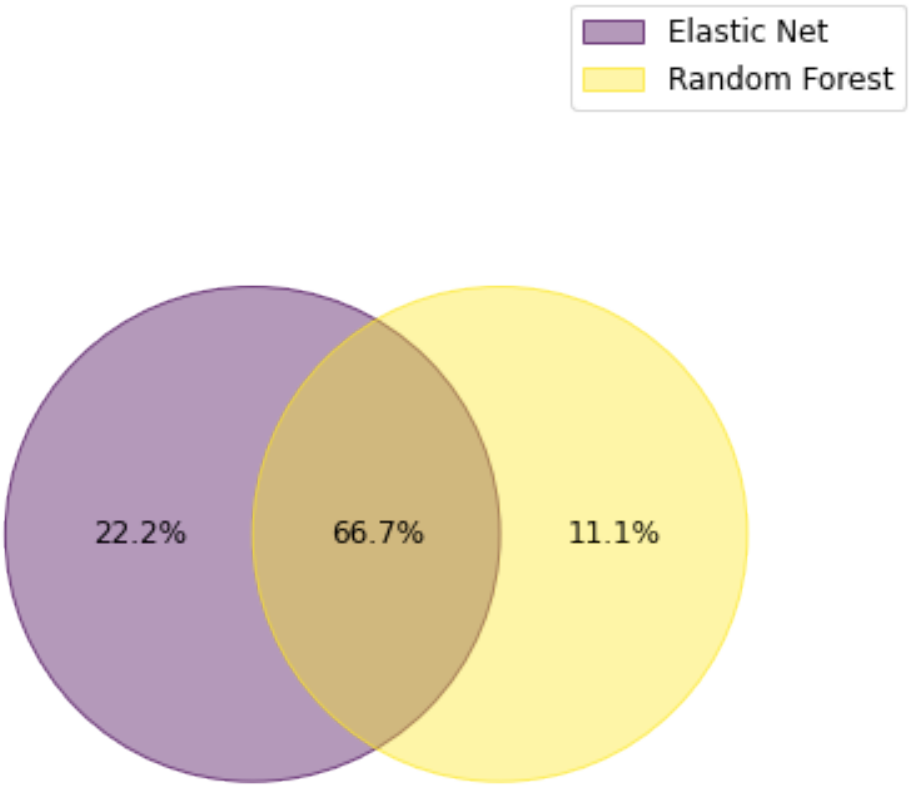

Elastic Net
Random Forest
19.0%
71.4%
9.5%

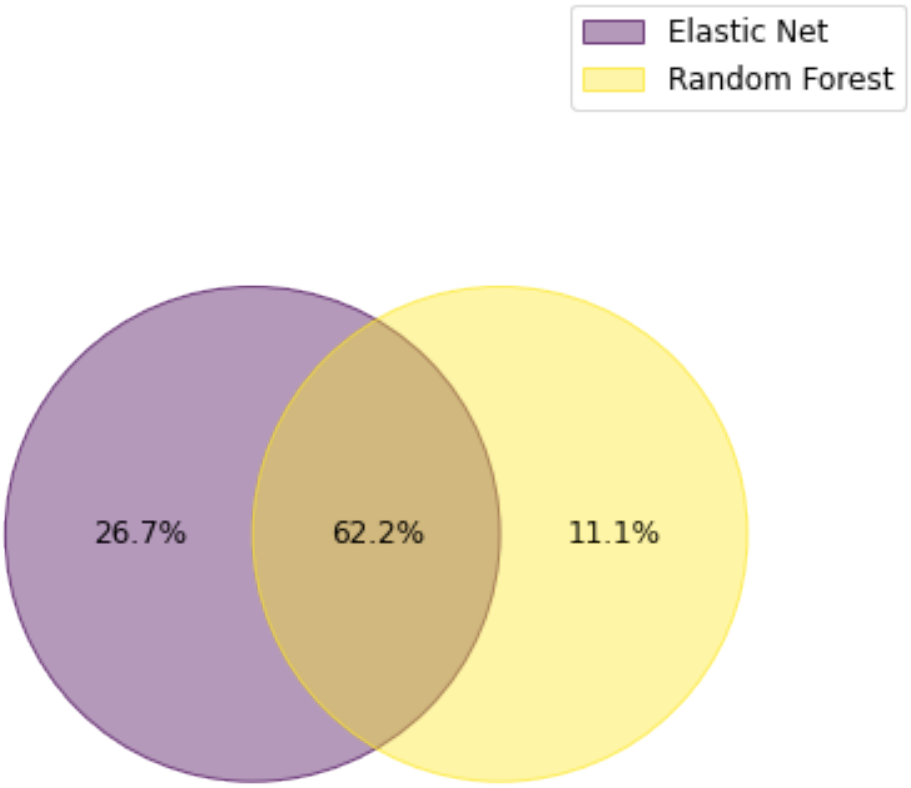

Elastic Net
Random Forest
18.2%
81.8%
0.0%

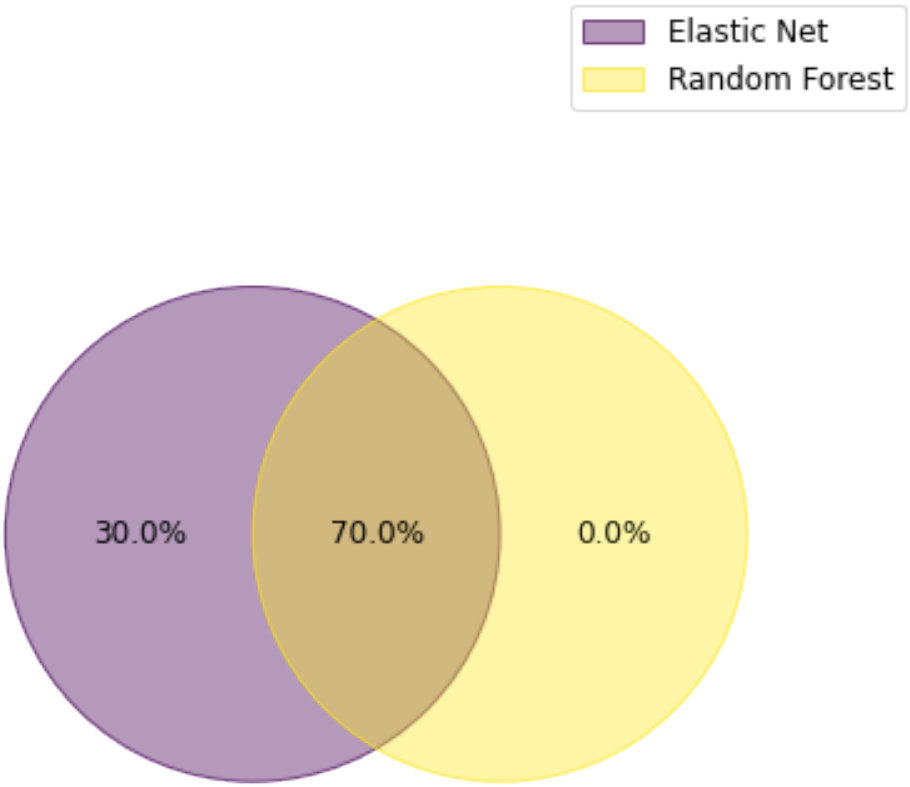

Elastic Net
Random Forest
16.7%
66.7%
16.7%

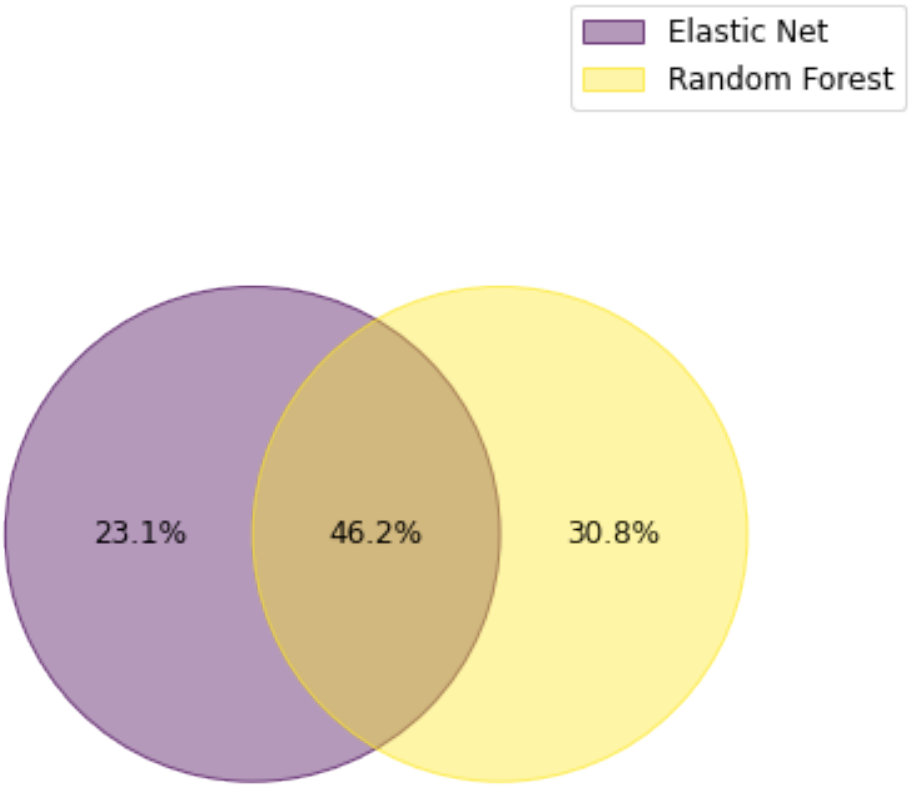

Elastic Net
Random Forest
29.0%
65.6%
5.4%

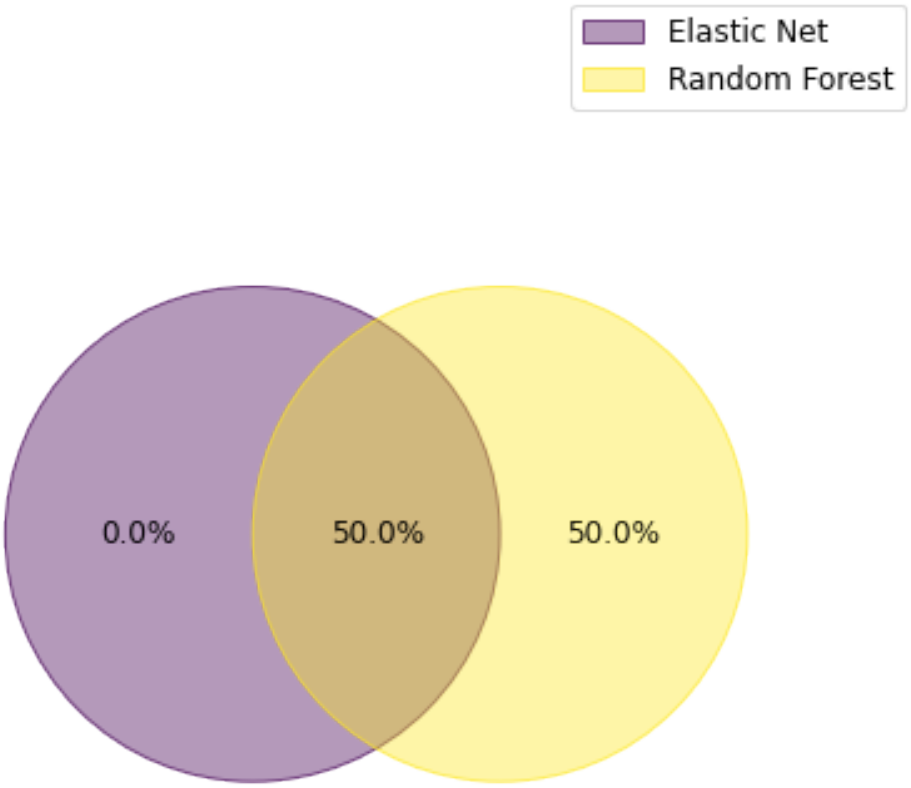

Elastic Net
Random Forest
27.6%
65.3%
7.1%

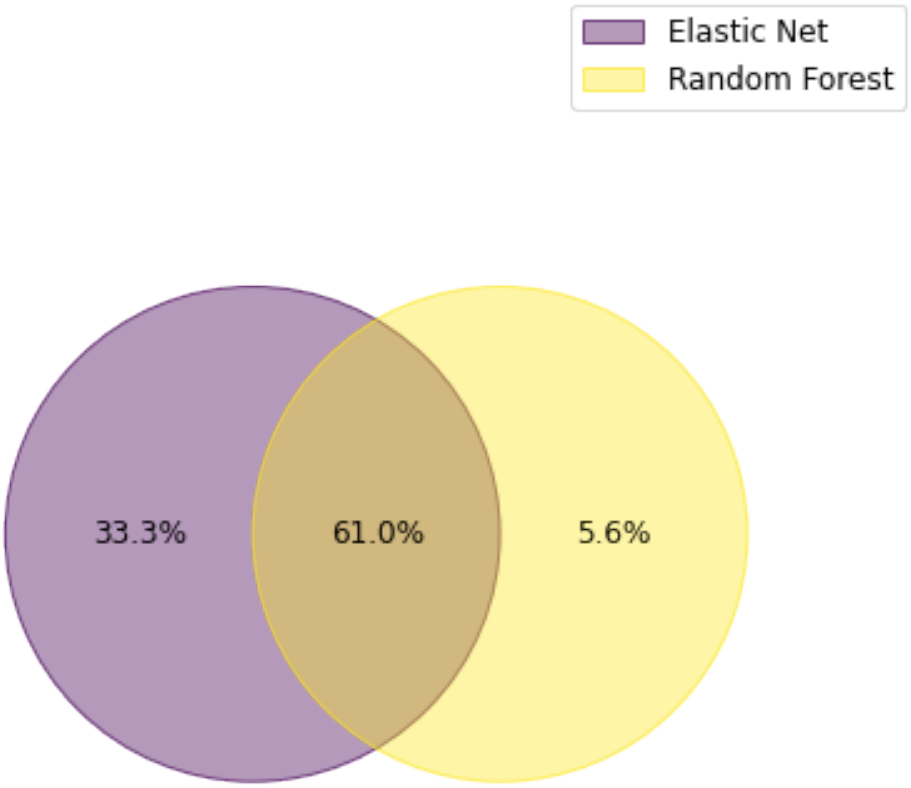

Elastic Net
Random Forest
22.2%
66.7%
11.1%

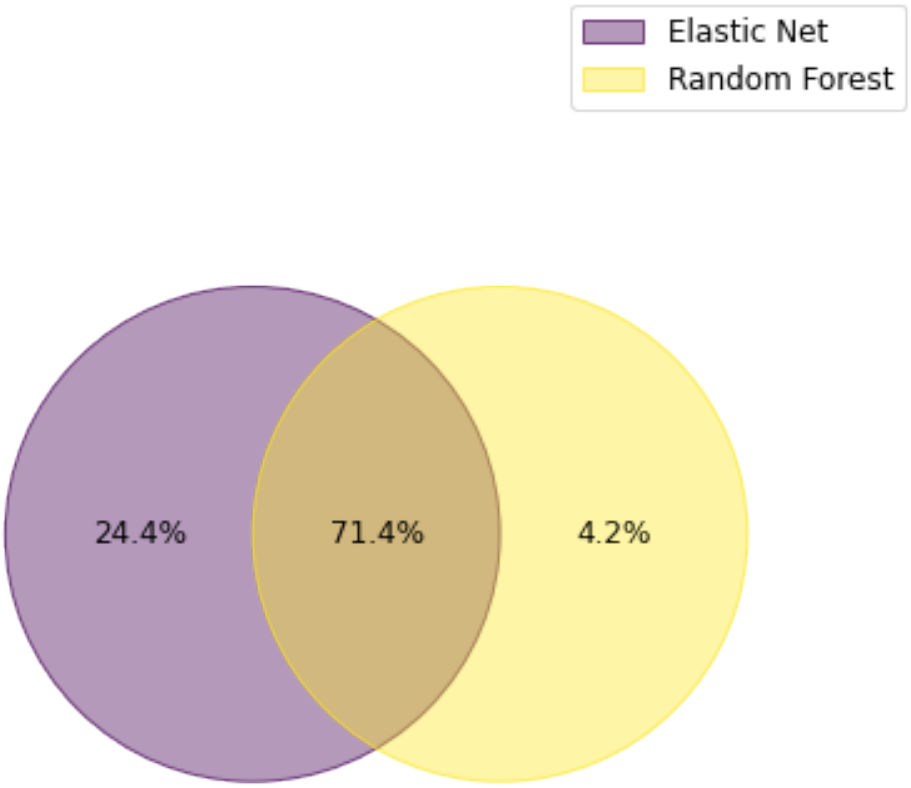

Elastic Net
Random Forest
26.7%
62.2%
11.1%

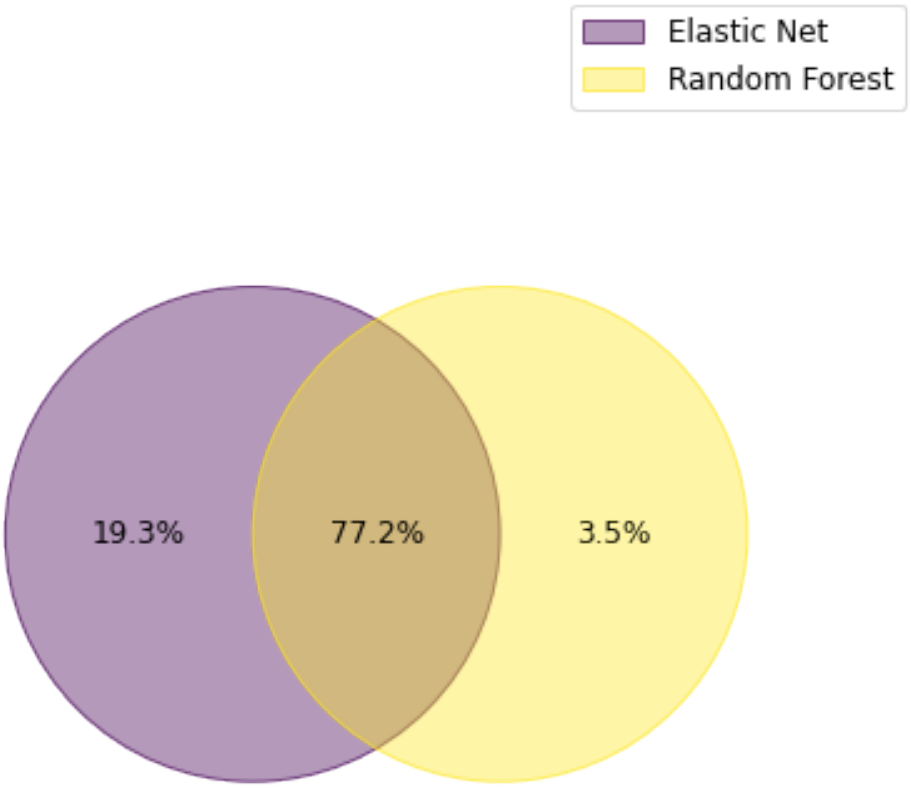

Elastic Net
Random Forest
30.0%
70.0%
0.0%

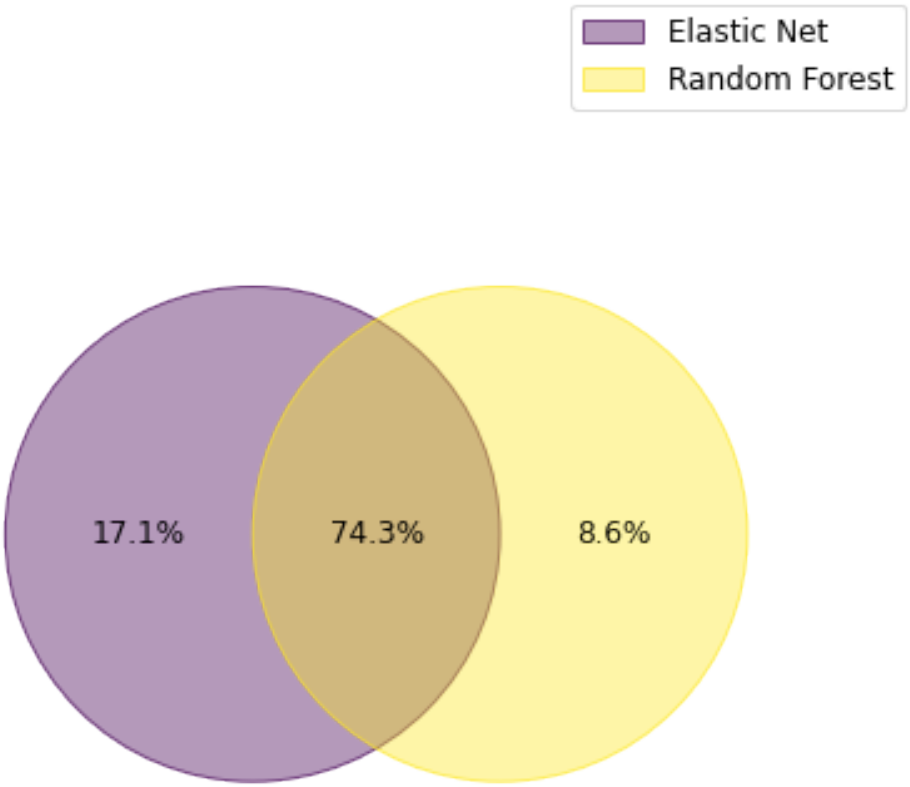

Elastic Net
Random Forest
23.1%
46.2%
30.8%

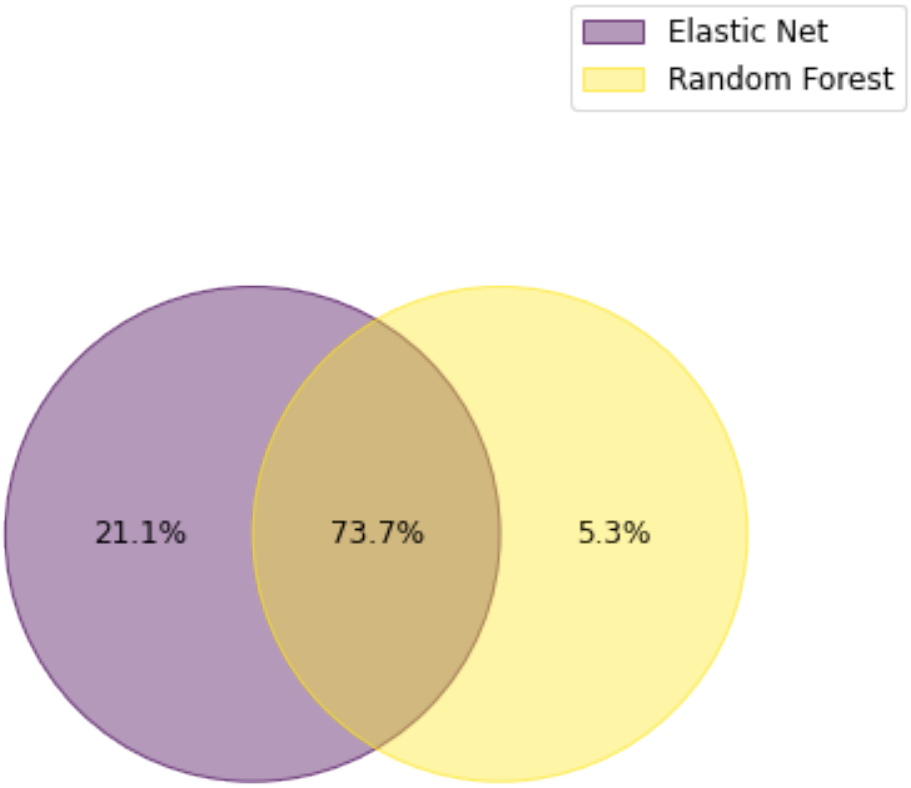

Elastic Net
Random Forest
0.0%
50.0%
50.0%

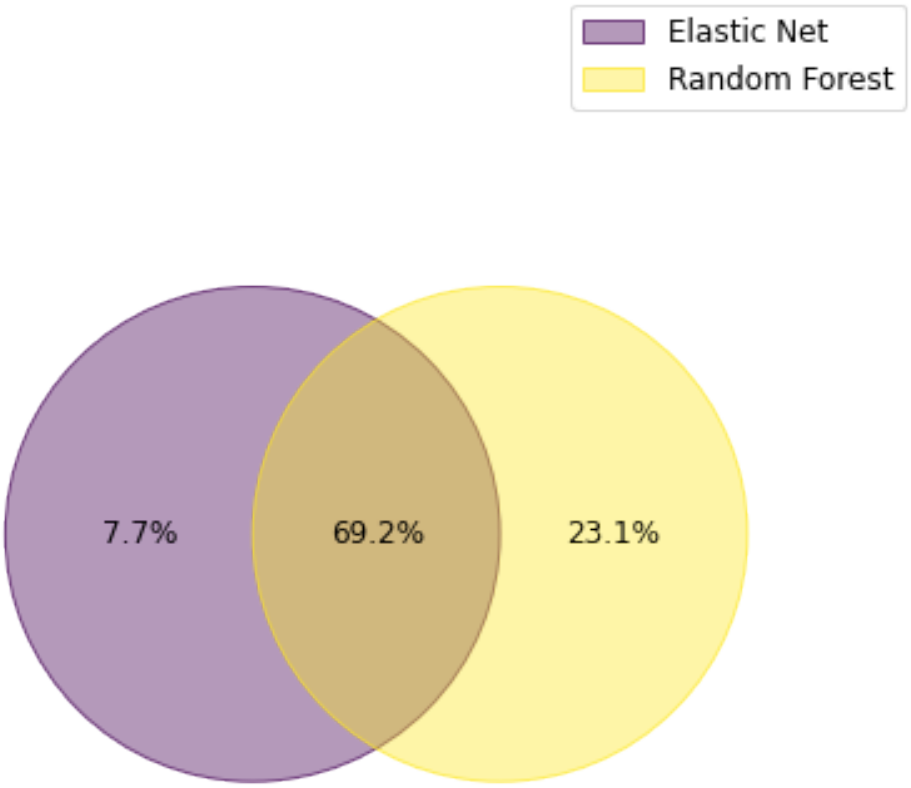

Elastic Net
Random Forest
33.3%
61.0%
5.6%

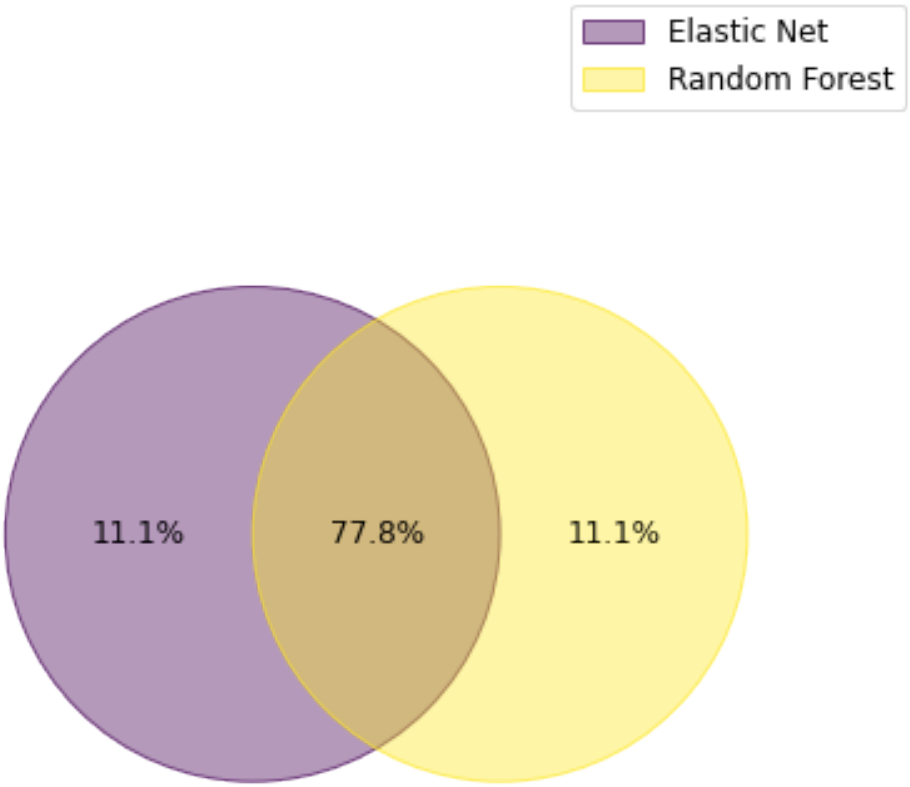

Elastic Net
Random Forest
24.4%
71.4%
4.2%

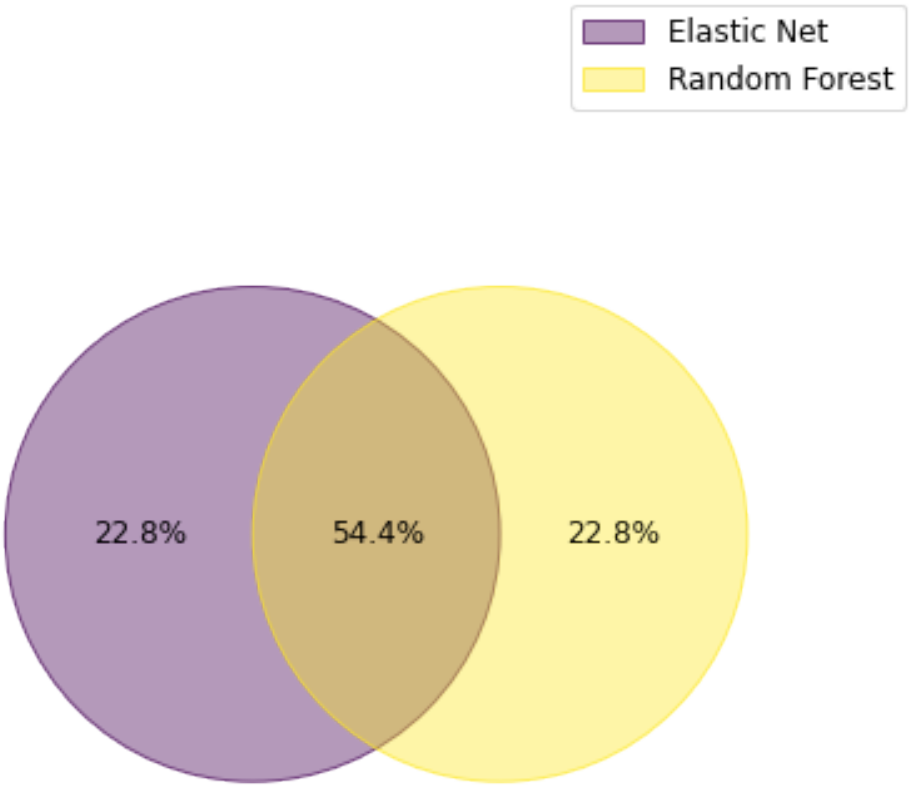

Elastic Net
Random Forest
19.3%
77.2%
3.5%

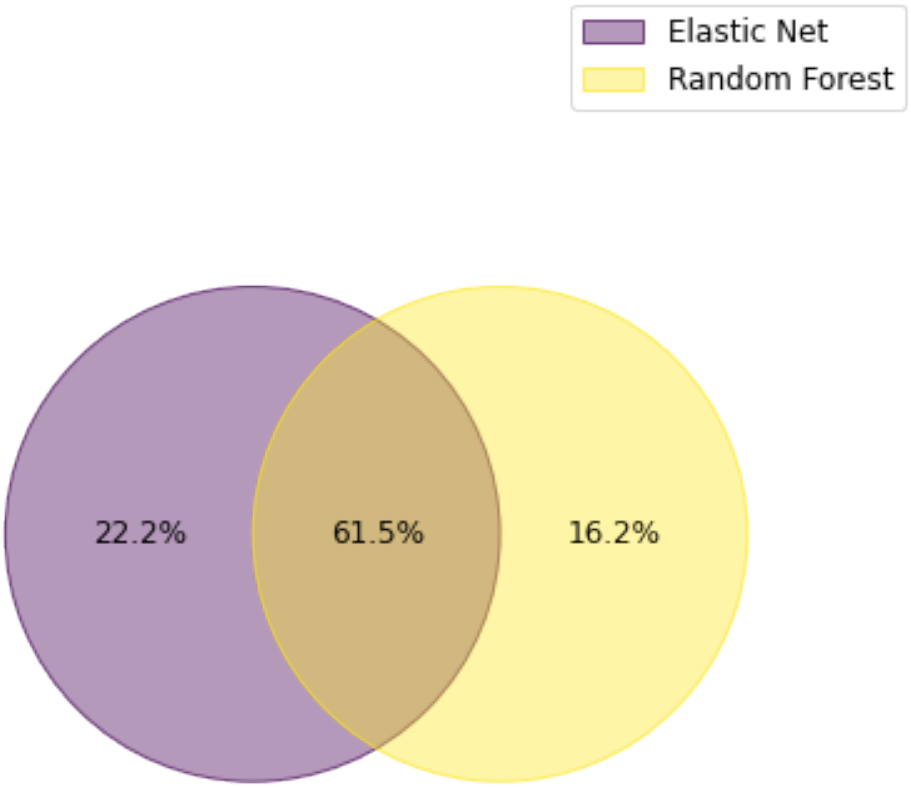

Elastic Net
Random Forest
17.1%
74.3%
8.6%

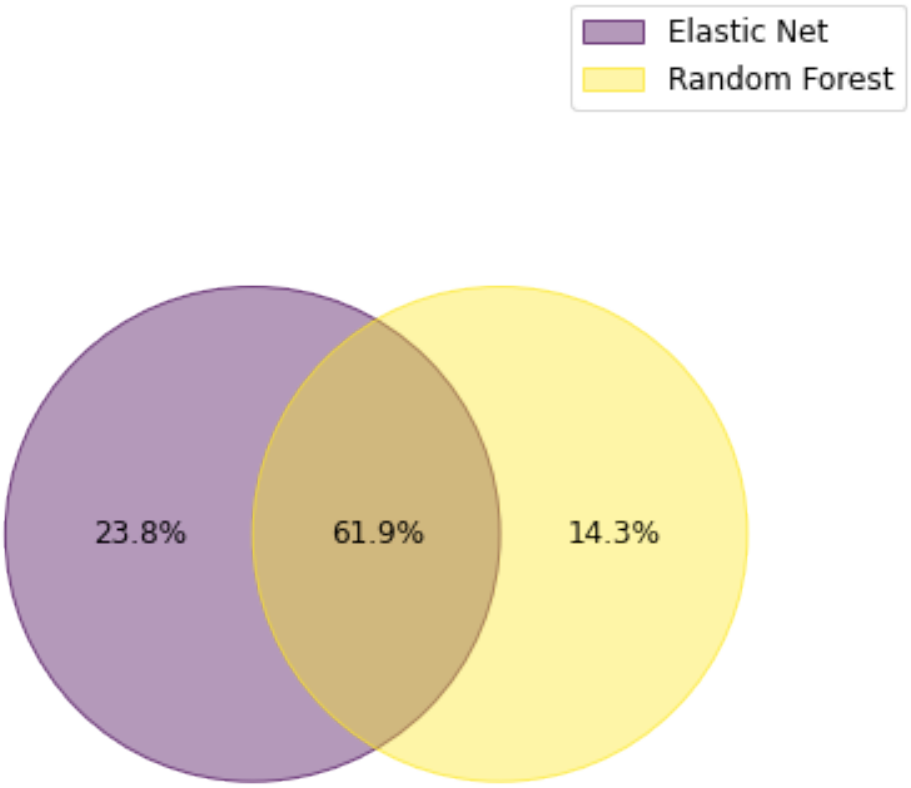

Elastic Net
Random Forest
21.1%
73.7%
5.3%

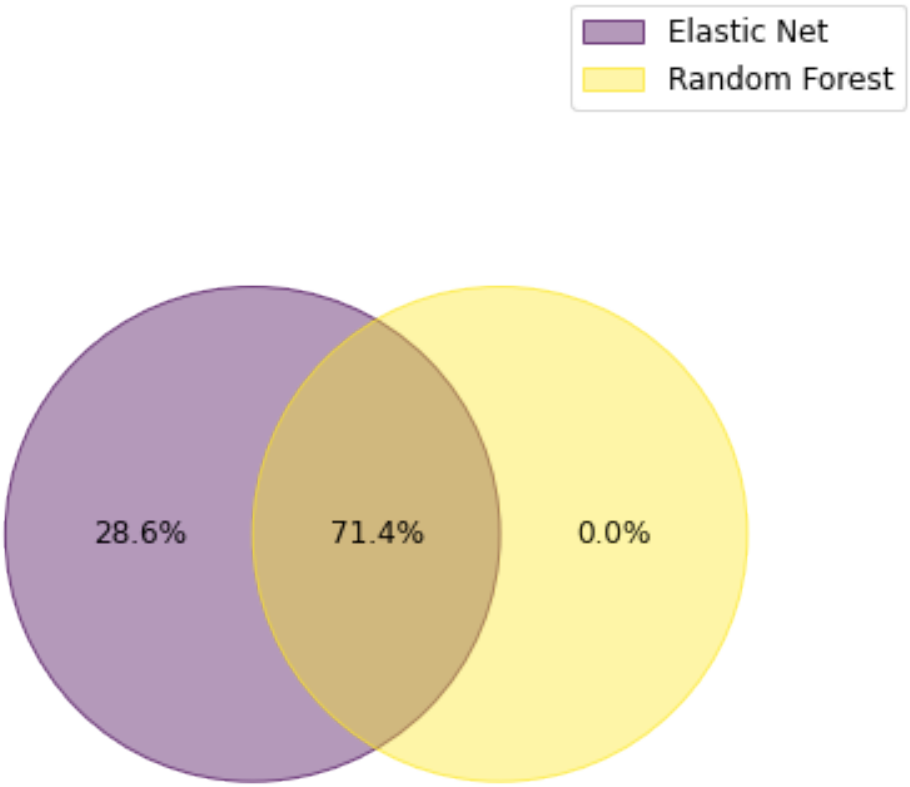

Elastic Net
Random Forest
7.7%
69.2%
23.1%

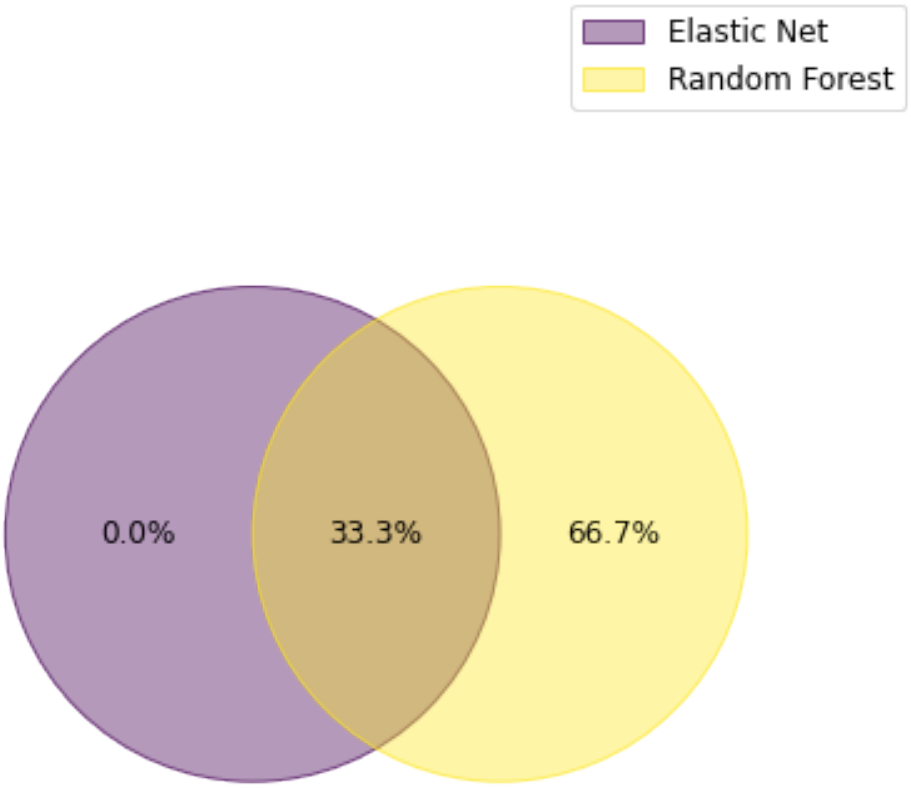

Elastic Net
Random Forest
11.1%
77.8%
11.1%

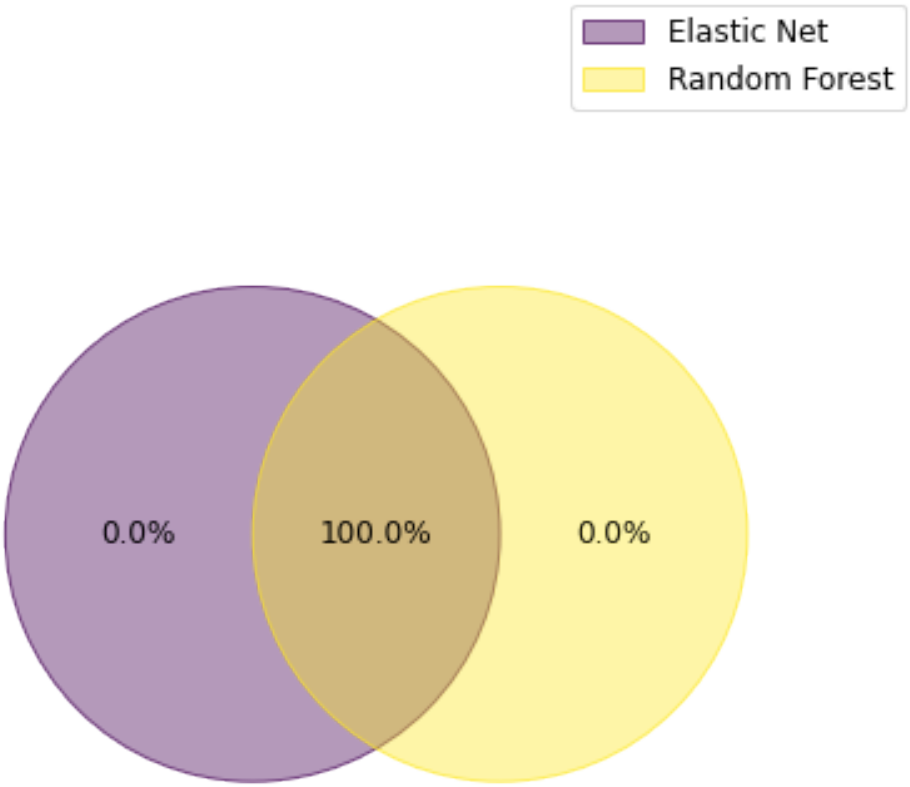

Elastic Net
Random Forest
22.8%
54.4%
22.8%

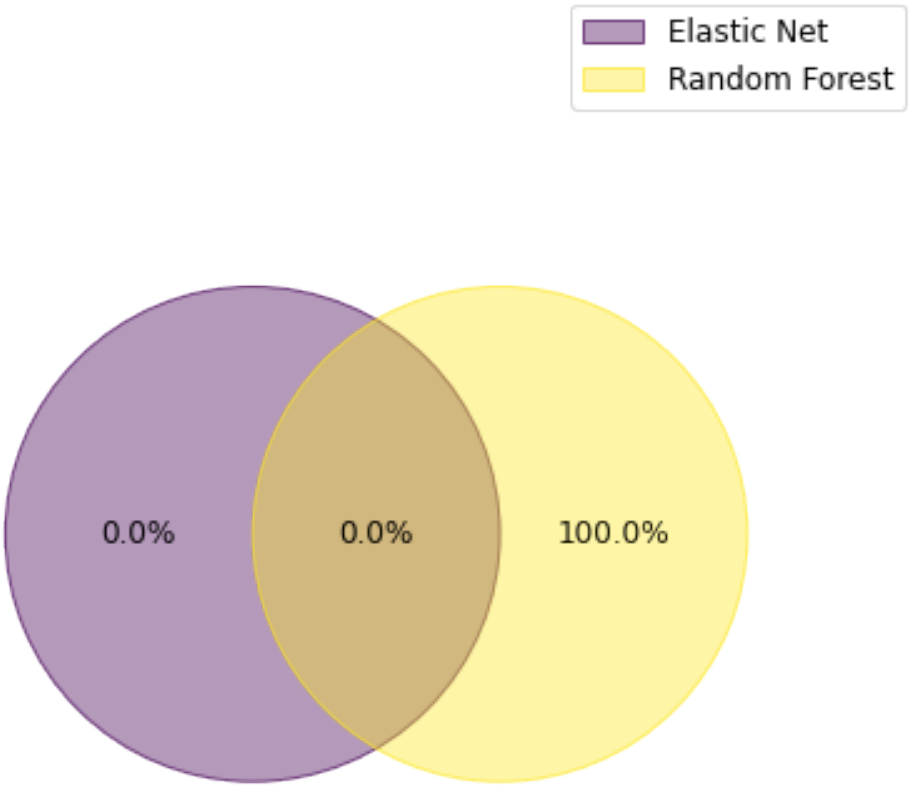

Elastic Net
Random Forest
22.2%
61.5%
16.2%


Elastic Net
Random Forest
23.8%
61.9%
14.3%


Elastic Net
Random Forest
28.6%
71.4%
0.0%


Elastic Net
Random Forest
0.0%
33.3%
66.7%

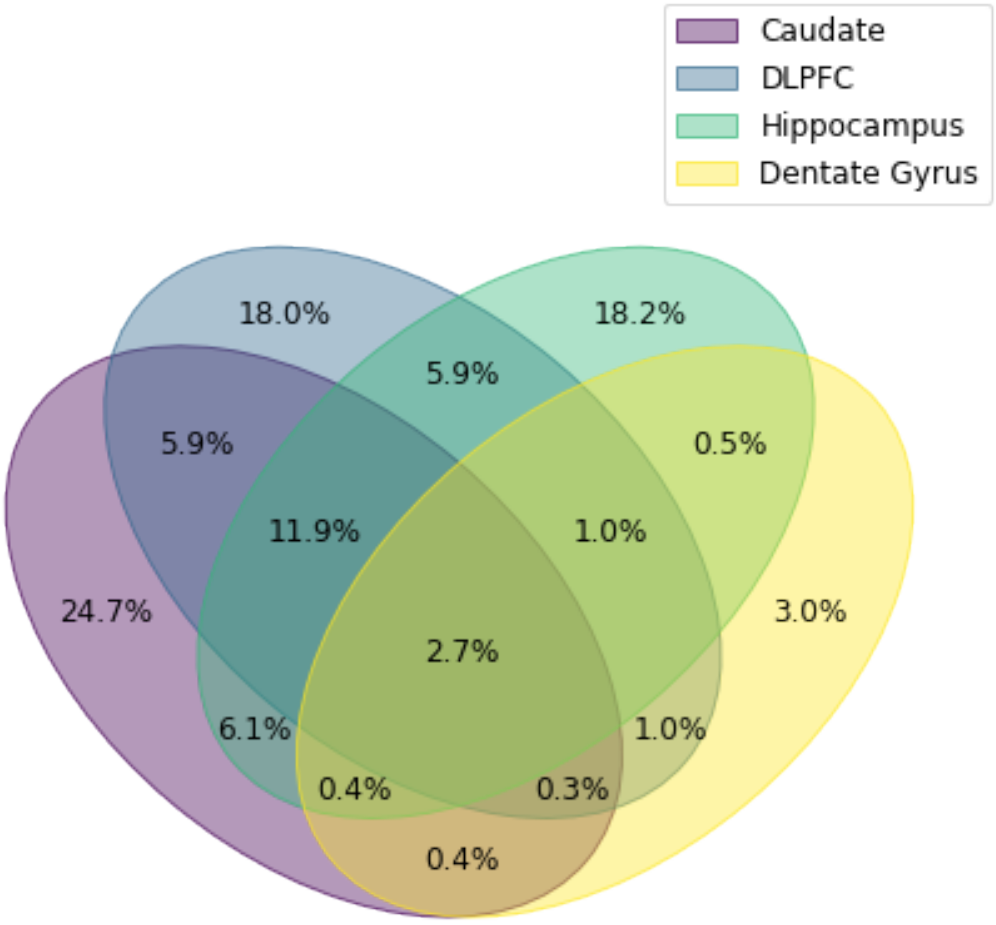

Elastic Net
Random Forest
0.0%
100.0%
0.0%

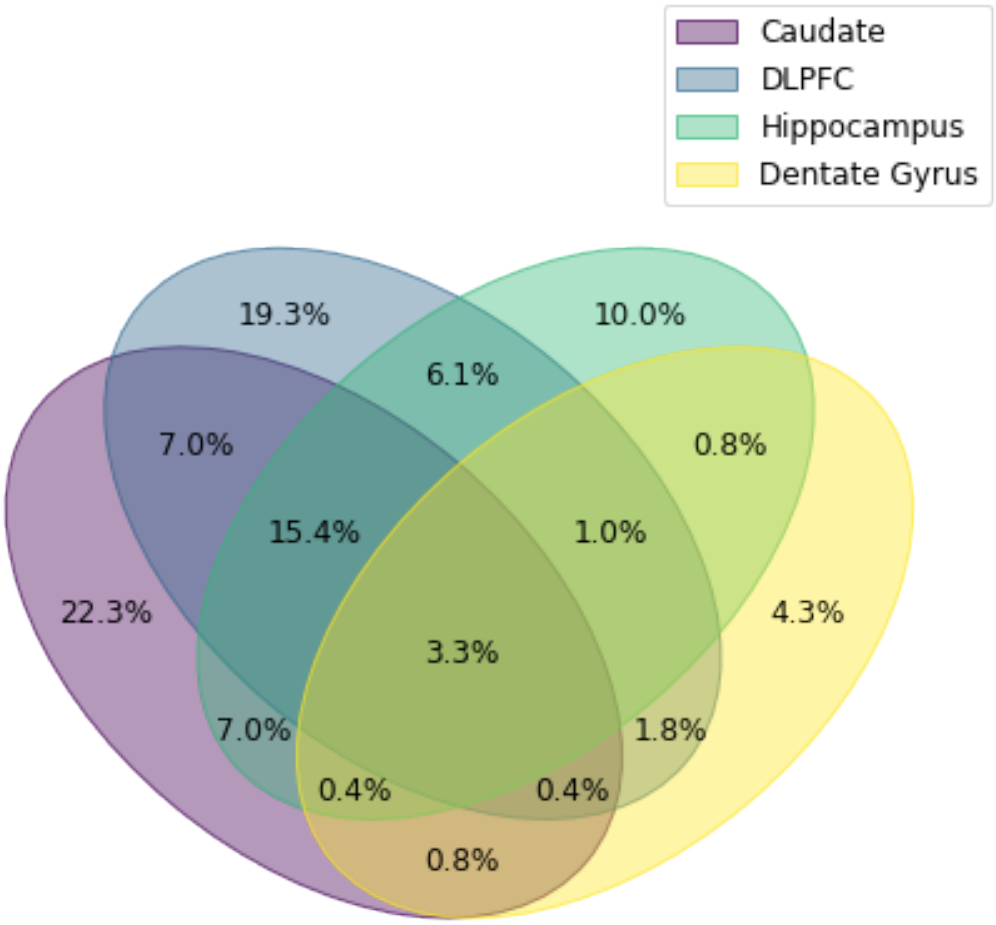


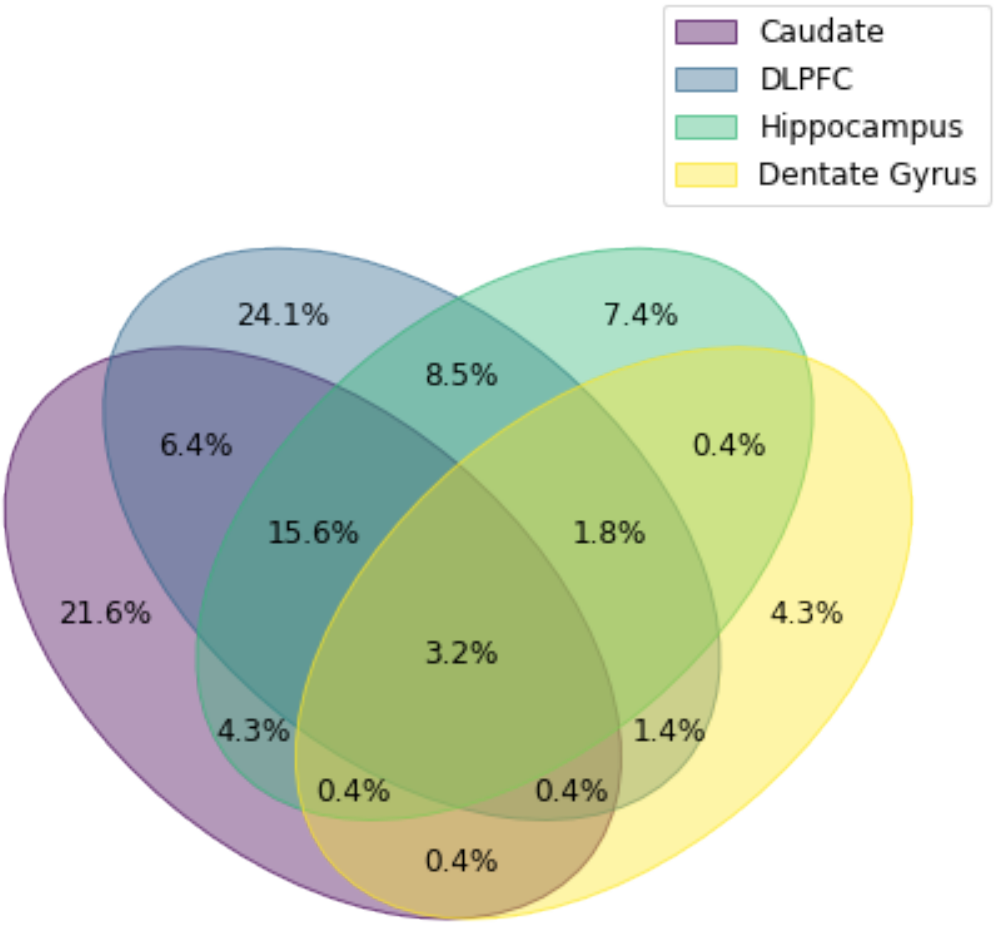
1.4.6 What is the overlap between brain regions?

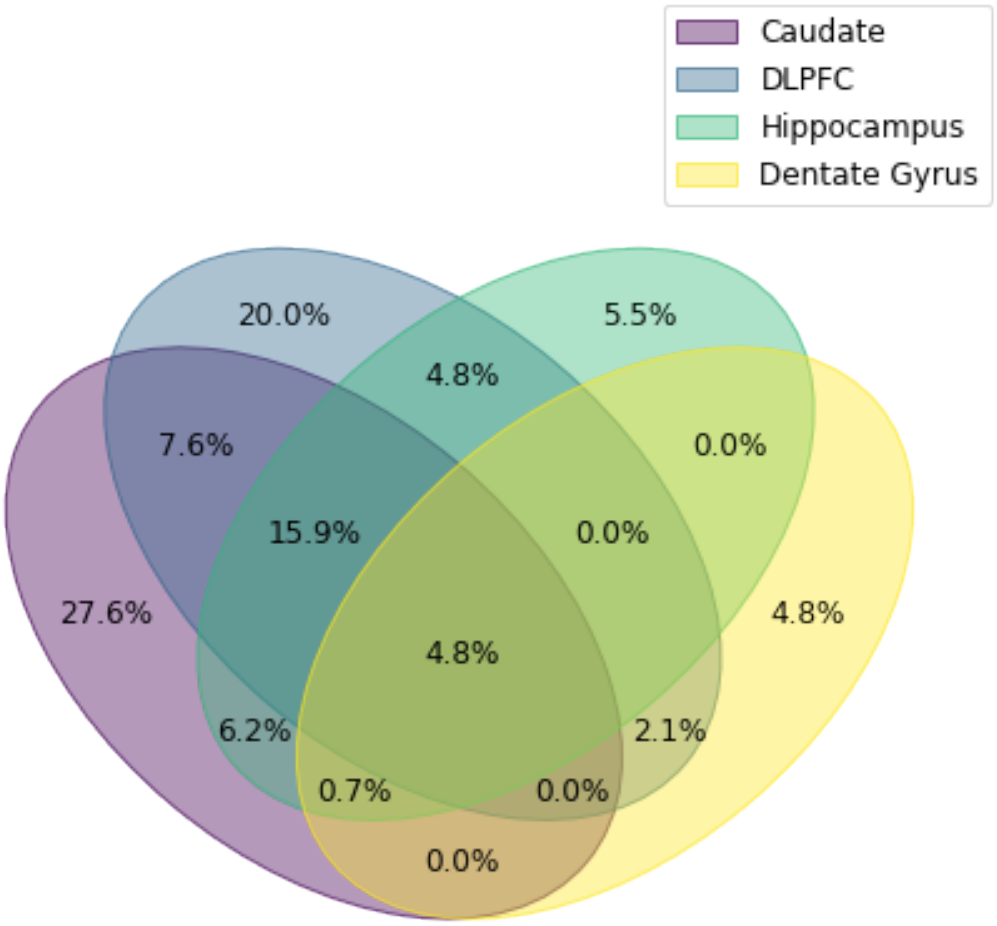
[23]:
```
dirname = "tissue_venn_diagrams"
mkdir_p(dirname)
for modeln in ["Elastic Net", "Random Forest"]:
    #print(modeln)
    dft = df[(df['Model'] == modeln)].copy()
    for r2 in [0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8]:
        cc = dft[(dft["tissue"] == "Caudate") & (dft["test_score_r2"] >= r2)].
↪copy()
        dd = dft[(dft["tissue"] == "DLPFC") & (dft["test_score_r2"] >= r2)].
↪copy()
```

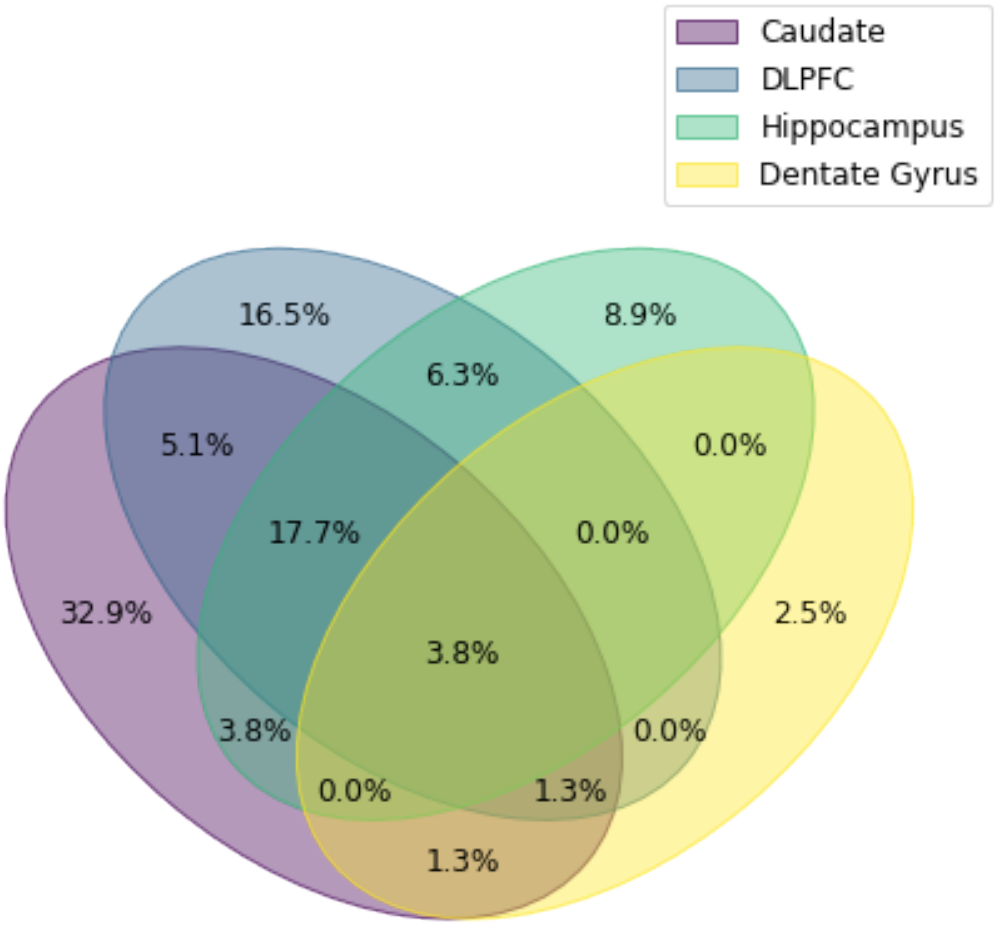
```
        hh = dft[(dft["tissue"] == "Hippocampus") & (dft["test_score_r2"] >=␣
↪r2)].copy()
        gg = dft[(dft["tissue"] == "Dentate Gyrus") & (dft["test_score_r2"] >=␣
↪r2)].copy()
        tissues = {"Caudate": set(cc.feature), "DLPFC": set(dd.feature),
                   "Hippocampus": set(hh.feature), "Dentate Gyrus": set(gg.
↪feature)}
        venn(tissues, fmt="{percentage:.1f}%", fontsize=12)
        mm = modeln.lower().replace(" ", "_")
        plt.savefig("{}/venn_diagram_tissueOverlap_{}_r2_{}.png".
↪format(dirname, mm, r2))
        plt.savefig("{}/venn_diagram_tissueOverlap_{}_r2_{}.pdf".
↪format(dirname, mm, r2))
        plt.savefig("{}/venn_diagram_tissueOverlap_{}_r2_{}.svg".
↪format(dirname, mm, r2))
```

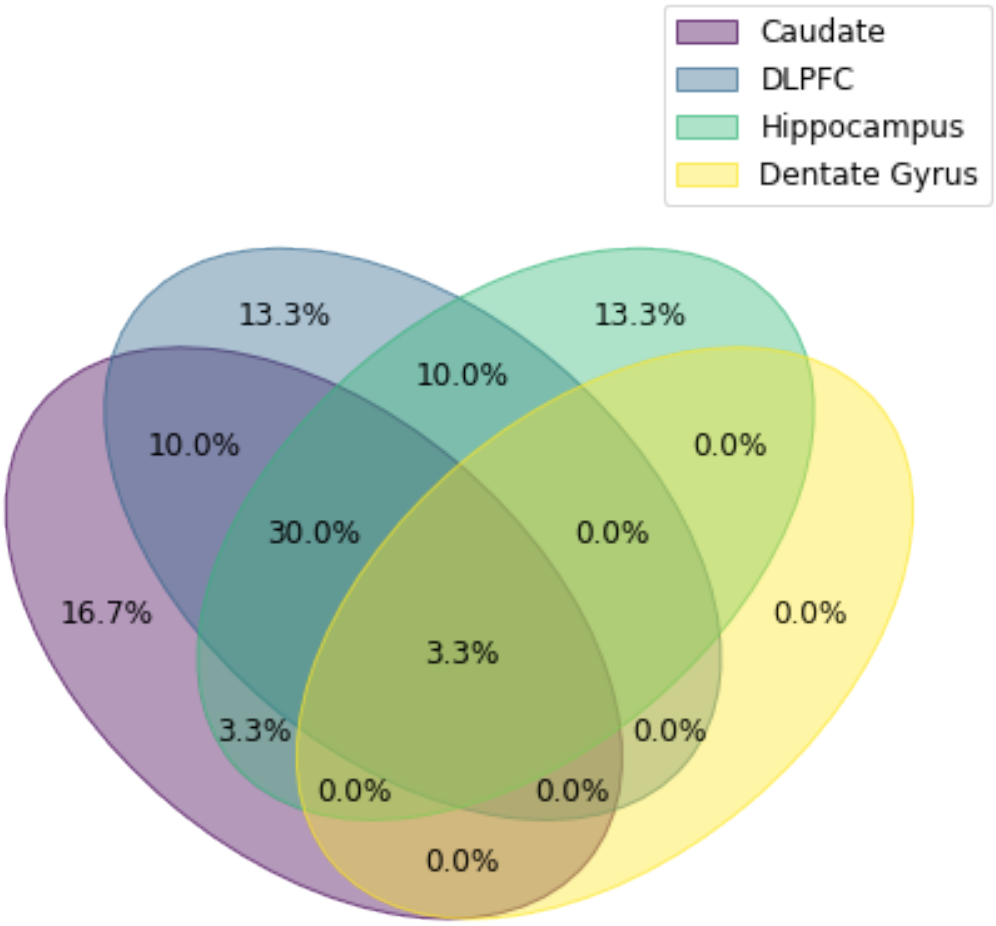


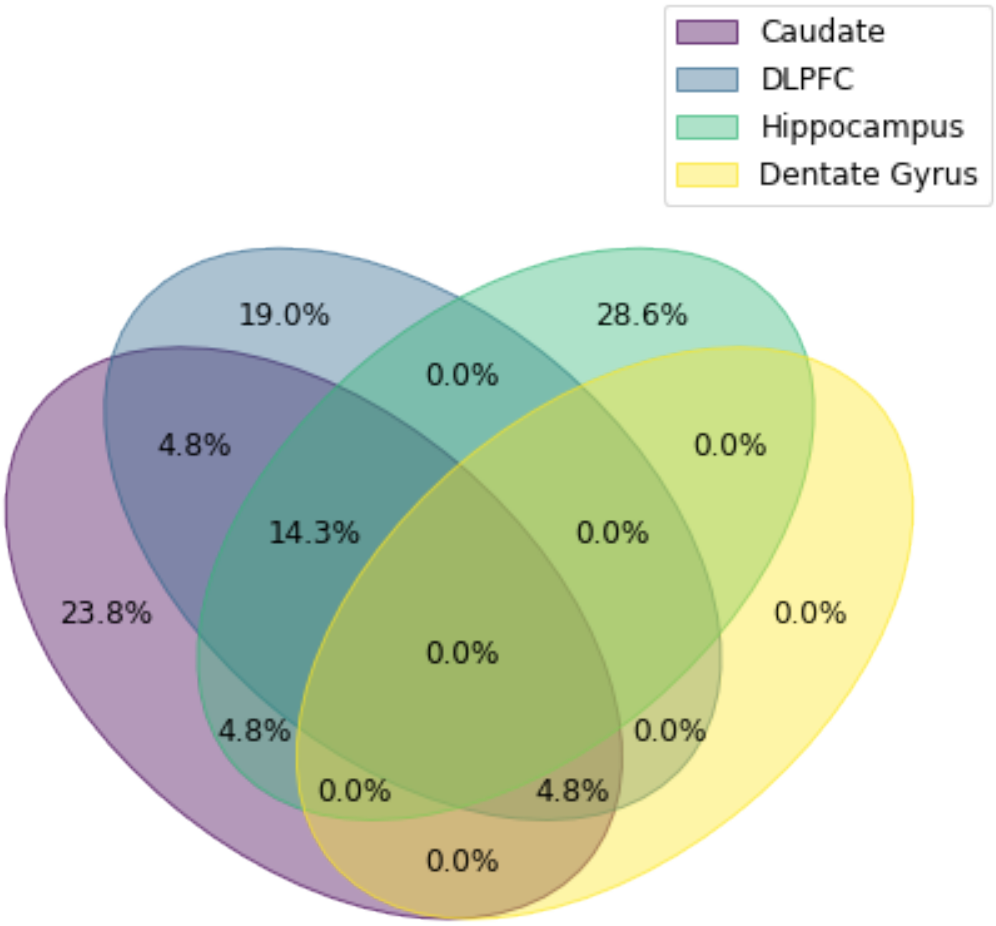

Caudate
DLPFC
Hippocampus
Dentate Gyrus
19.3%
10.0%
6.1%
7.0%
0.8%
15.4%
1.0%
22.3%
4.3%
3.3%
7.0%
1.8%
0.4%
0.4%
0.8%

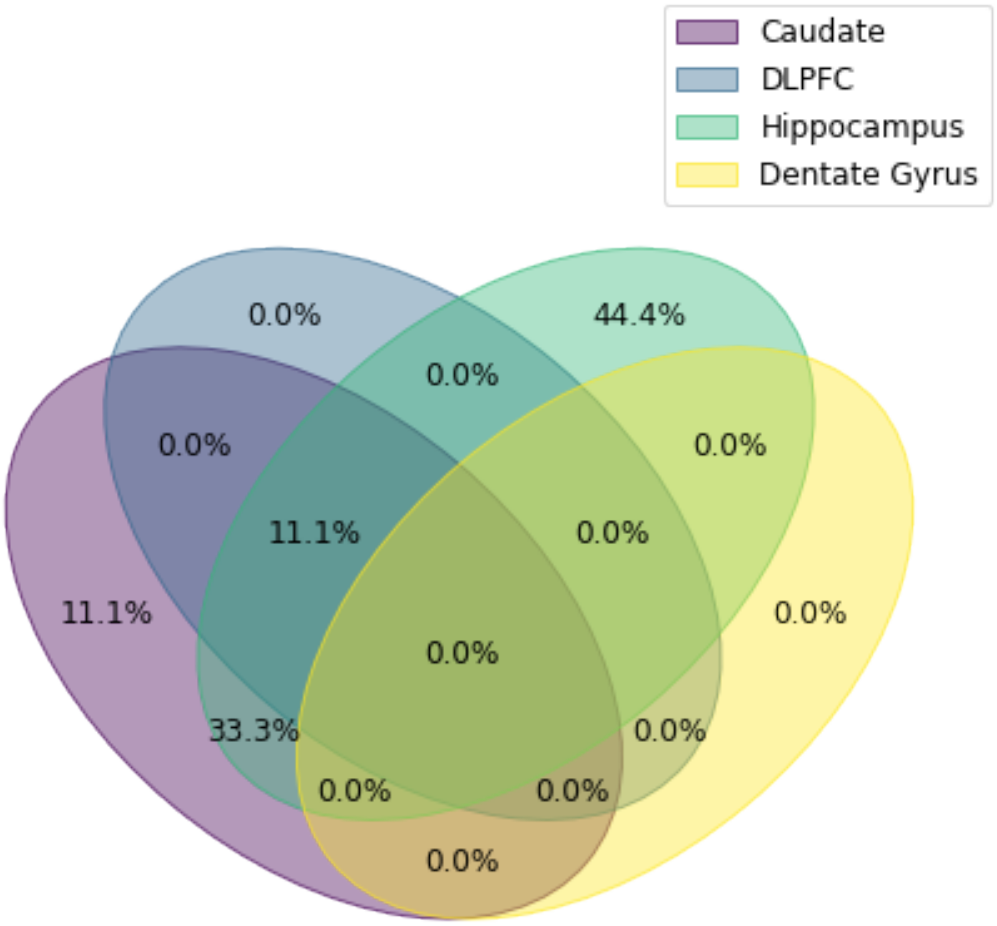

Caudate
DLPFC
Hippocampus
Dentate Gyrus
24.1%
7.4%
8.5%
6.4%
0.4%
15.6%
1.8%
21.6%
4.3%
3.2%
4.3%
1.4%
0.4%
0.4%
0.4%

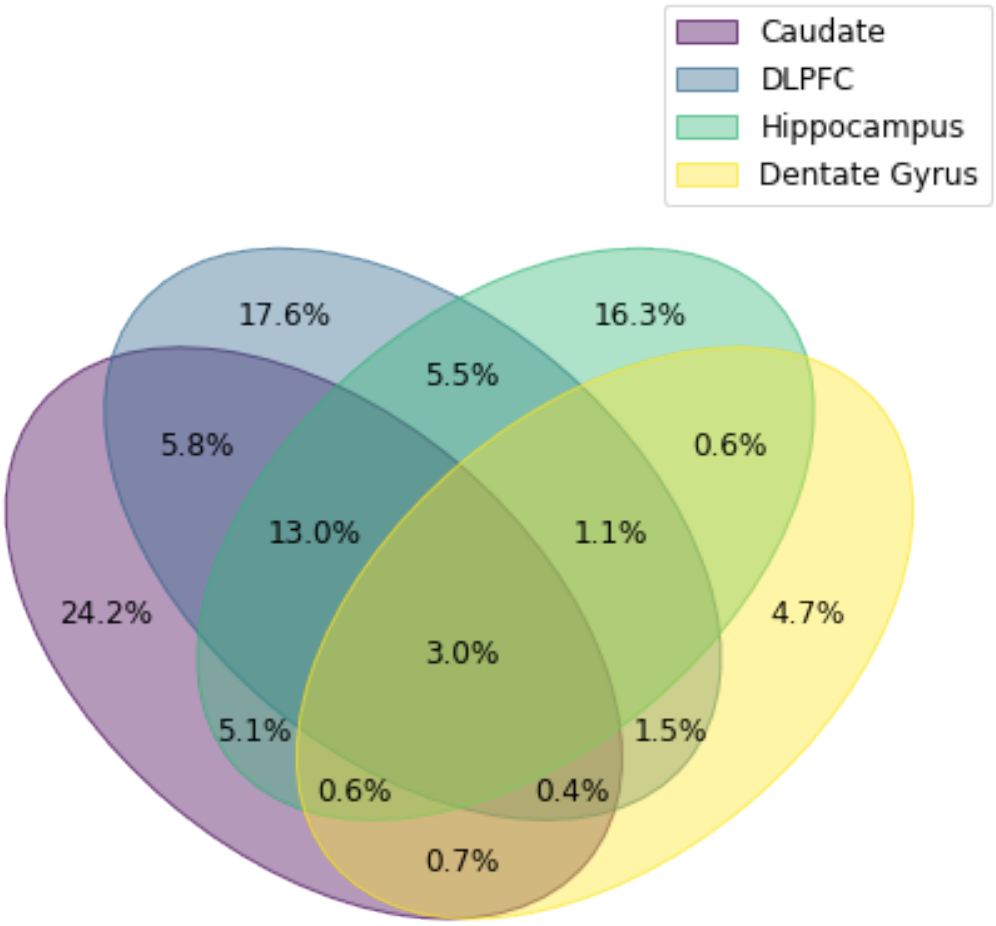

Caudate
DLPFC
Hippocampus
Dentate Gyrus
20.0%
5.5%
4.8%
7.6%
0.0%
15.9%
0.0%
27.6%
4.8%
4.8%
6.2%
2.1%
0.7%
0.0%
0.0%

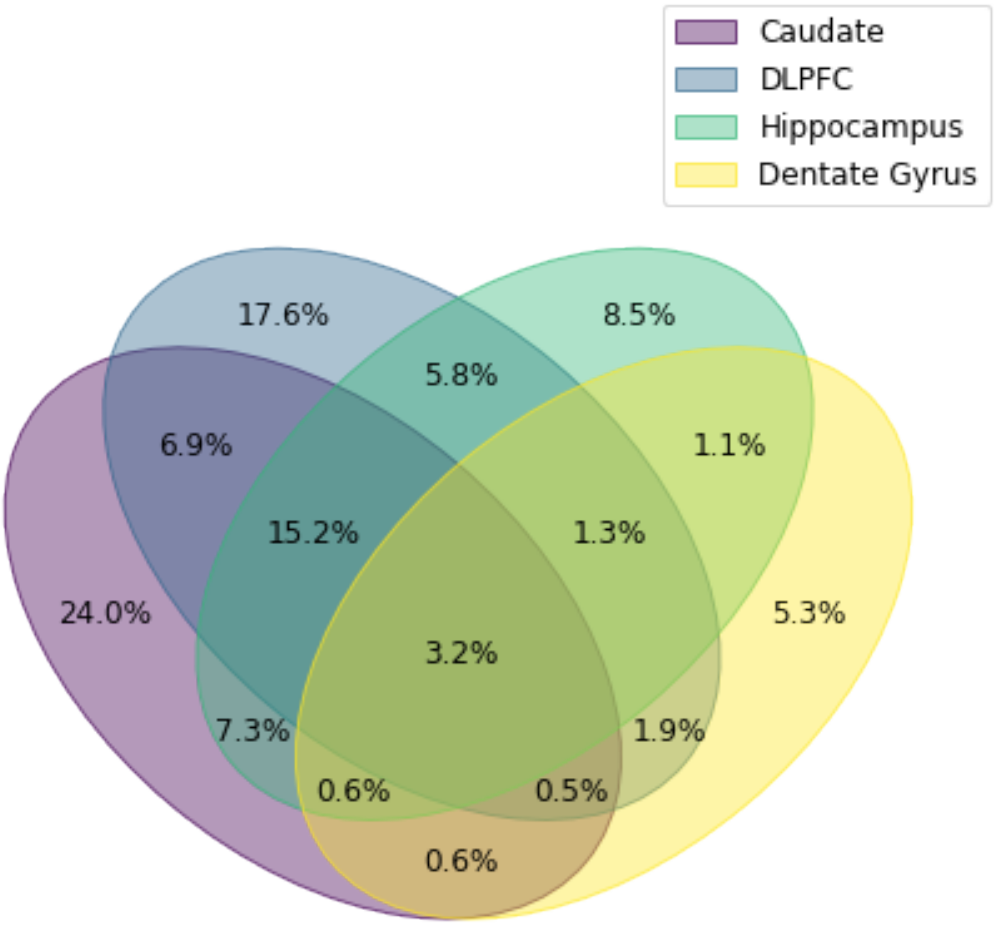

Caudate
DLPFC
Hippocampus
Dentate Gyrus
16.5%
8.9%
6.3%
5.1%
0.0%
17.7%
0.0%
32.9%
2.5%
3.8%
3.8%
0.0%
0.0%
1.3%
1.3%

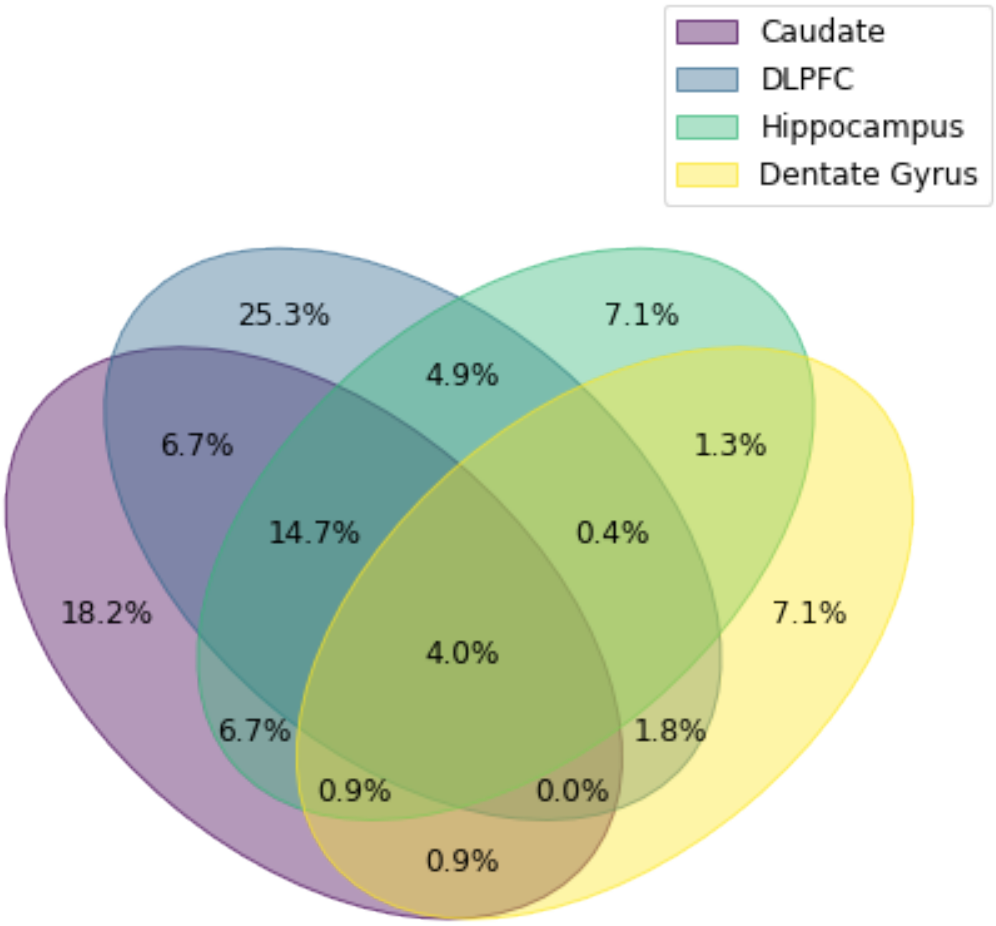

Caudate
DLPFC
Hippocampus
Dentate Gyrus
13.3%
13.3%
10.0%
10.0%
0.0%
30.0%
0.0%
16.7%
0.0%
3.3%
3.3%
0.0%
0.0%
0.0%
0.0%

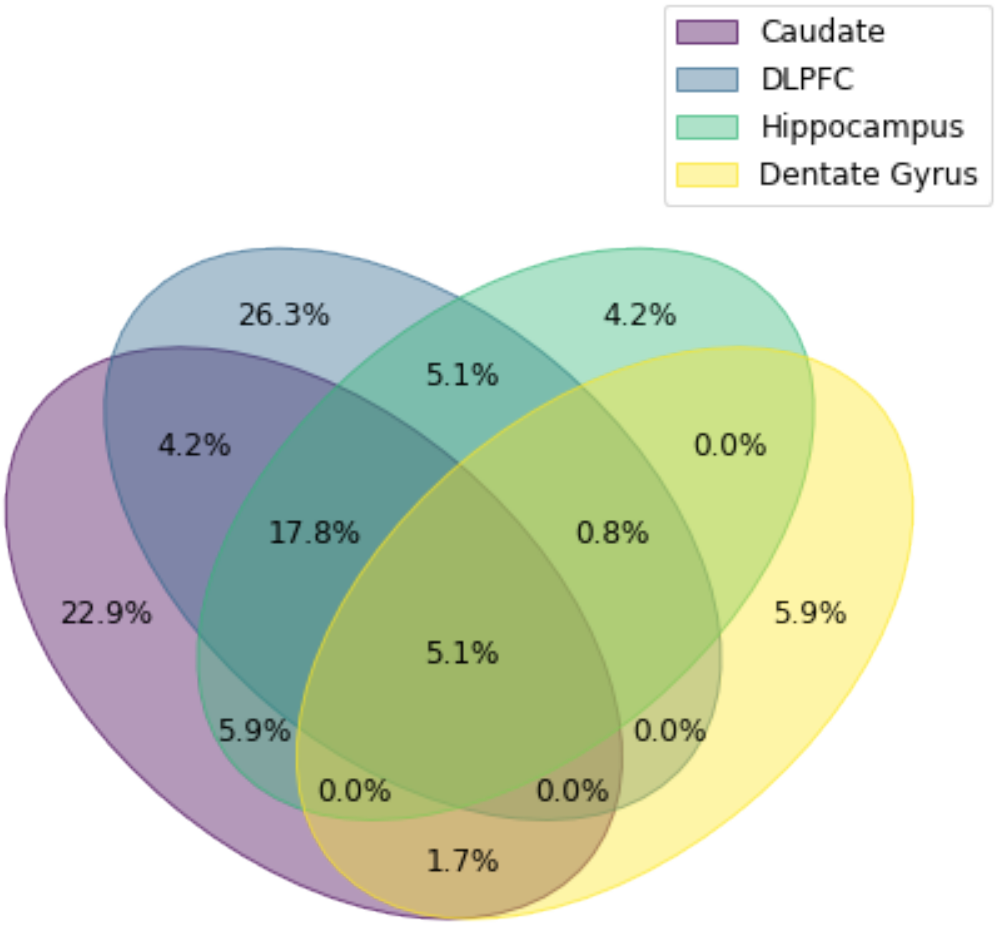

Caudate
DLPFC
Hippocampus
Dentate Gyrus
19.0%
28.6%
0.0%
4.8%
0.0%
14.3%
0.0%
23.8%
0.0%
0.0%
4.8%
0.0%
0.0%
4.8%
0.0%

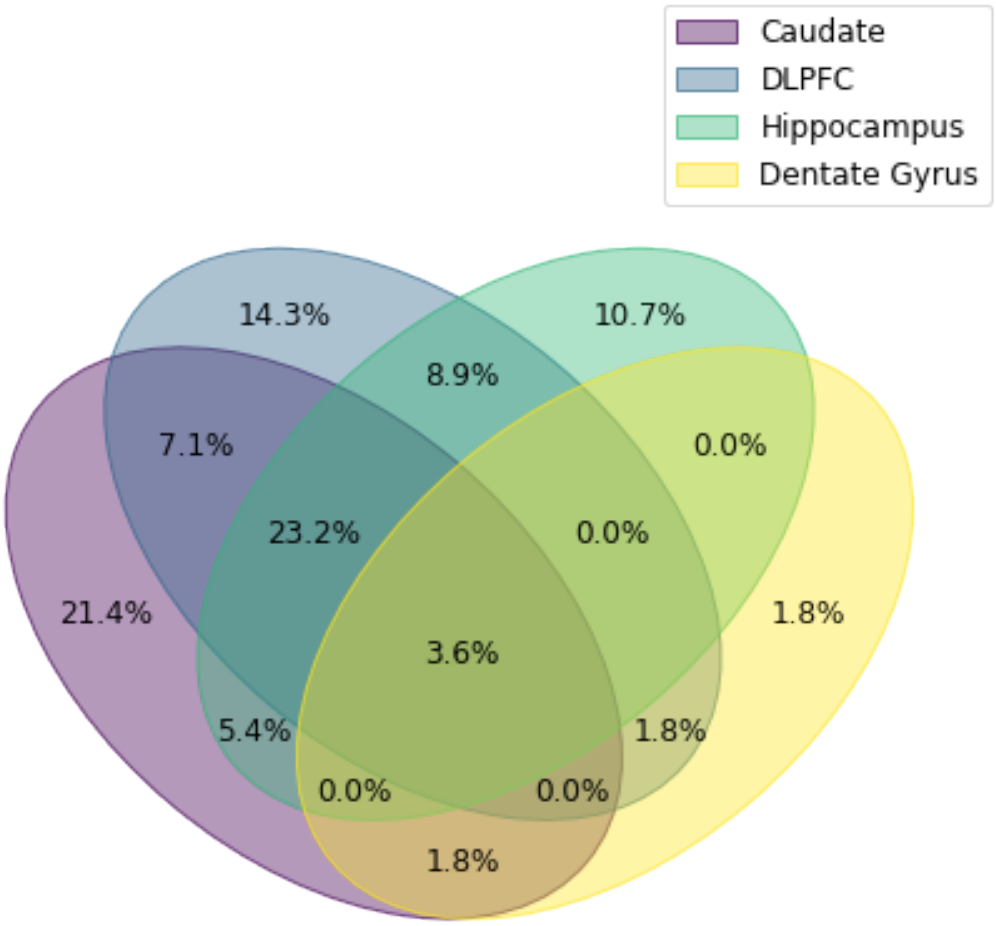

Caudate
DLPFC
Hippocampus
Dentate Gyrus
0.0%
44.4%
0.0%
0.0%
0.0%
11.1%
0.0%
11.1%
0.0%
0.0%
33.3%
0.0%
0.0%
0.0%
0.0%

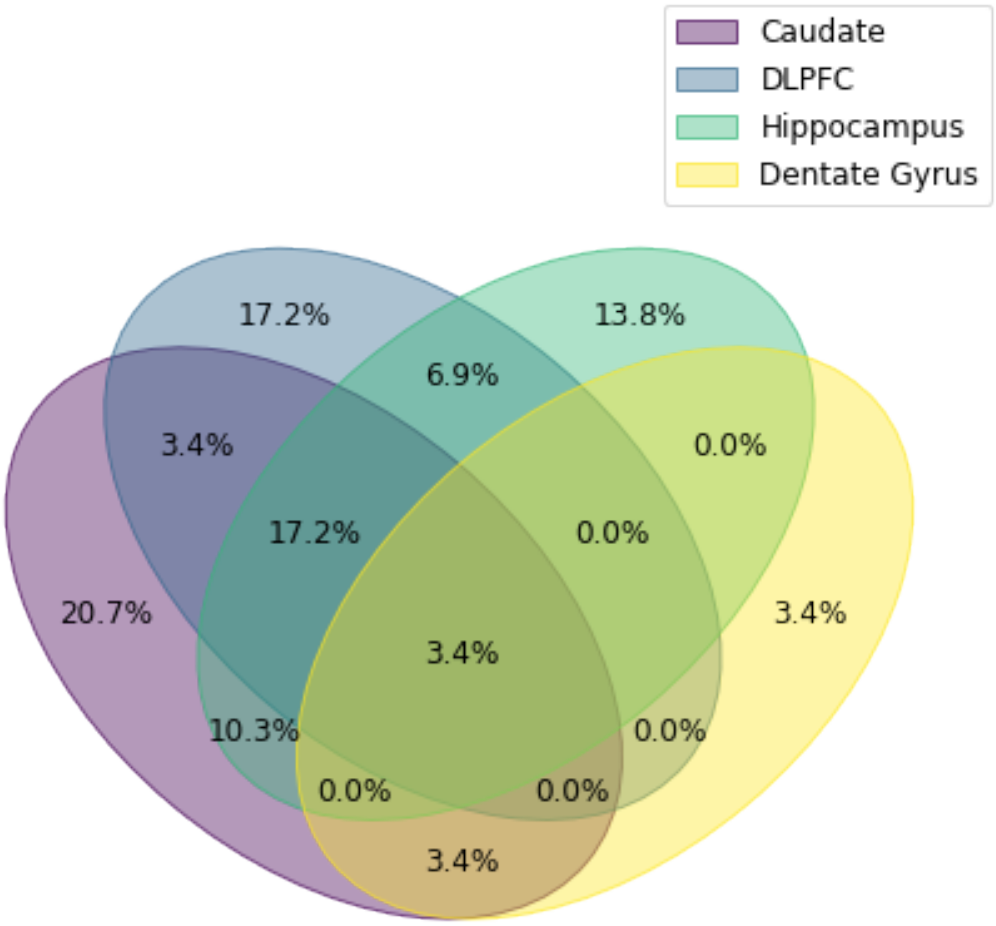

Caudate
DLPFC
Hippocampus
Dentate Gyrus
17.6%
16.3%
5.5%
5.8%
0.6%
13.0%
1.1%
24.2%
4.7%
3.0%
5.1%
1.5%
0.6%
0.4%
0.7%

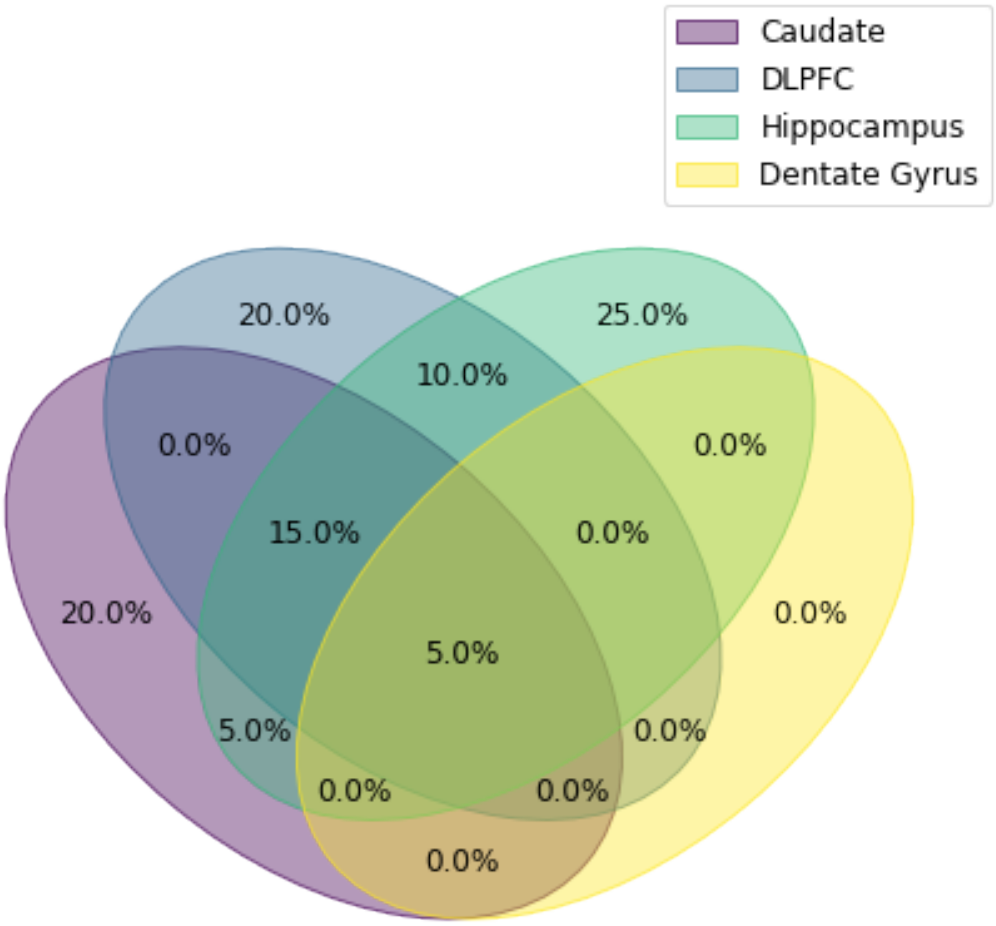

Caudate
DLPFC
Hippocampus
Dentate Gyrus
17.6%
8.5%
5.8%
6.9%
1.1%
15.2%
1.3%
24.0%
5.3%
3.2%
7.3%
1.9%
0.6%
0.5%
0.6%

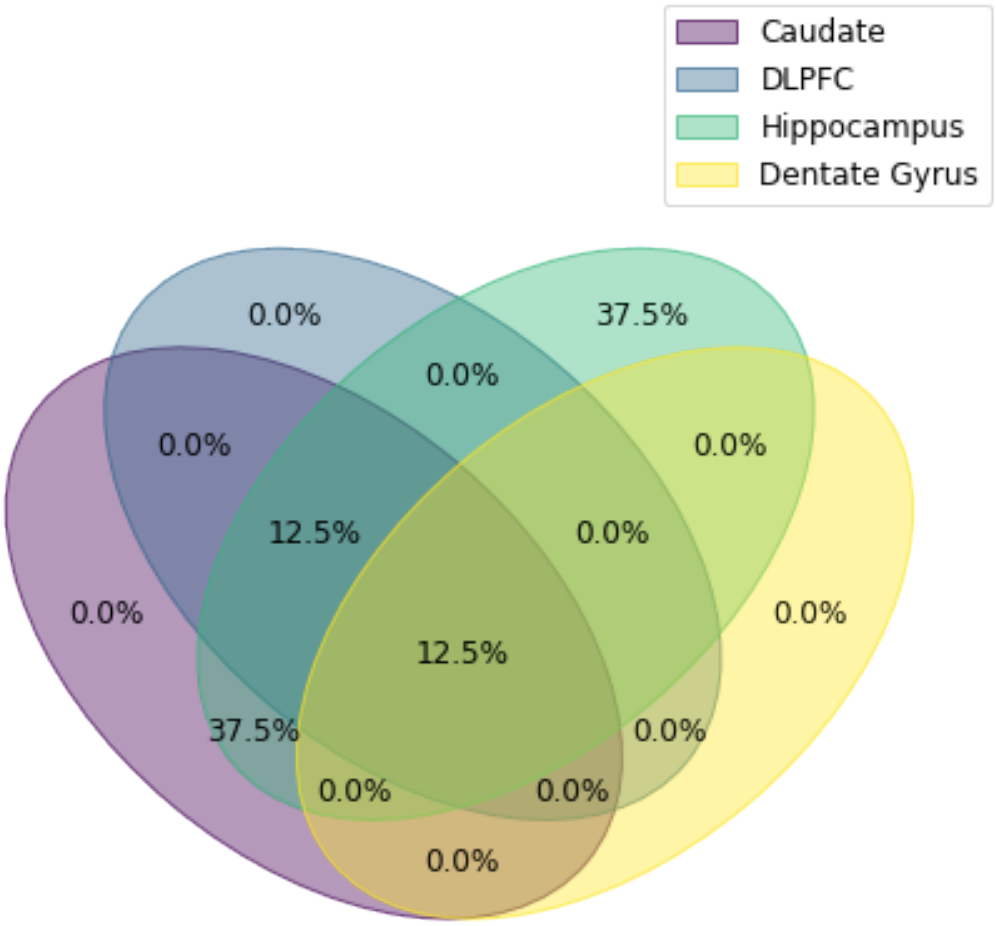

Caudate
DLPFC
Hippocampus
Dentate Gyrus
25.3%
7.1%
4.9%
6.7%
1.3%
14.7%
0.4%
18.2%
7.1%
4.0%
6.7%
1.8%
0.9%
0.0%
0.9%


Caudate
DLPFC
Hippocampus
Dentate Gyrus
26.3%
4.2%
5.1%
4.2%
0.0%
17.8%
0.8%
22.9%
5.9%
5.1%
5.9%
0.0%
0.0%
0.0%
1.7%


Caudate
DLPFC
Hippocampus
Dentate Gyrus
14.3%
10.7%
8.9%
7.1%
0.0%
23.2%
0.0%
21.4%
1.8%
3.6%
5.4%
1.8%
0.0%
0.0%
1.8%


Caudate
DLPFC
Hippocampus
Dentate Gyrus
17.2%
13.8%
6.9%
3.4%
0.0%
17.2%
0.0%
20.7%
3.4%
3.4%
10.3%
0.0%
0.0%
0.0%
3.4%


Caudate
DLPFC
Hippocampus
Dentate Gyrus
20.0%
25.0%
10.0%
0.0%
0.0%
15.0%
0.0%
20.0%
0.0%
5.0%
5.0%
0.0%
0.0%
0.0%
0.0%



1.5 Examining partial R2 results using most predictive SNPs

[24]:
```
partial.groupby("tissue").describe().T
```

[24]:

tissue		Caudate	DLPFC	Dentate Gyrus	Hippocampus
n_features	count	2925.000000	2691.000000	773.000000	2906.000000
	mean	12.931966	13.730583	16.210867	11.360289
	std	21.776563	26.601085	26.373774	18.818252
	min	1.000000	1.000000	1.000000	1.000000
	25%	3.000000	3.000000	5.000000	3.000000
	50%	5.000000	5.000000	9.000000	4.000000
	75%	14.000000	14.000000	18.000000	11.000000
	max	306.000000	524.000000	265.000000	229.000000

```
test_score_r2 count  2925.000000  2691.000000     773.000000  2906.000000
              mean      0.123348     0.127885       0.207344     0.102622
              std       0.153135     0.152273       0.195071     0.137614
              min       0.000000     0.000000       0.000000     0.000000
              25%       0.016948     0.019686       0.059365     0.011089
              50%       0.064108     0.068851       0.140023     0.047339
              75%       0.165582     0.180191       0.314842     0.135505
              max       0.888735     0.876882       1.000000     0.914498
```

[25]:
```
partial[(partial["test_score_r2"] > 0.88)]
```

[25]:

	tissue	feature	n_features	test_score_r2	Model
1311	Caudate	ENSG00000166435.15	28	0.888735	Partial R2
5655	Hippocampus	ENSG00000013573.16	13	0.887303	Partial R2
8030	Hippocampus	ENSG00000244879.5	105	0.895691	Partial R2
8141	Hippocampus	ENSG00000255374.3	40	0.914498	Partial R2
8599	Dentate Gyrus	ENSG00000103528.16	265	1.000000	Partial R2
8933	Dentate Gyrus	ENSG00000176956.12	242	1.000000	Partial R2
8934	Dentate Gyrus	ENSG00000176998.4	196	0.881363	Partial R2
8980	Dentate Gyrus	ENSG00000186088.15	249	0.911170	Partial R2
9087	Dentate Gyrus	ENSG00000226278.1	84	0.899644	Partial R2
9242	Dentate Gyrus	ENSG00000270605.1	38	0.948167	Partial R2

- *GLP2R* (ENSG00000065325) Glucagon Like Peptide 2 Receptor

[26]:
```
idv_partial = pd.read_csv("../../partial_r2/individual_partial_r2_metrics.tsv",␣
↪sep='\t')
idv_partial.head(2)
```

[26]:

	SNP	Partial_R2	Full_R2	Reduced_R2	Tissue	Geneid
0	chr11_71433422_G_A_0	0.000453	227.546538	227.649578	Caudate	ENSG00000172890.11
1	chr11_71433422_G_A_1	0.012297	224.850088	227.649578	Caudate	ENSG00000172890.11

[27]:
```
idv_partial[["Partial_R2", "Tissue", "Geneid"]].groupby("Tissue").describe().T
```

[27]:
```
Tissue                 Caudate         DLPFC  Dentate Gyrus   Hippocampus
Partial_R2 count  1.762851e+06  1.595825e+06  450379.000000  1.720189e+06
           mean   1.177298e-02  1.192422e-02       0.017215  1.016017e-02
           std    3.623054e-02  3.529084e-02       0.042975  3.167019e-02
           min    0.000000e+00  0.000000e+00       0.000000  0.000000e+00
           25%    0.000000e+00  0.000000e+00       0.000000  0.000000e+00
           50%    1.151334e-03  1.190616e-03       0.001776  1.026979e-03
           75%    8.549333e-03  8.895461e-03       0.016113  7.504191e-03
           max    8.853651e-01  9.086128e-01       0.927805  8.921231e-01
```

The vast majority of SNPs to not hold a lot of information (partial r2 < 0.01) with 25% close to 0.

[28]:
```
idv_partial.loc[(idv_partial["Partial_R2"] >= 0.8), ["Tissue", "Partial_R2",␣
↪"Geneid"]].groupby("Tissue").size()
```

[28]:
```
Tissue
Caudate          129
DLPFC             25
Dentate Gyrus     29
Hippocampus       78
dtype: int64
```

[29]:
```
idv_partial.loc[(idv_partial["Partial_R2"] >= 0.8), ["Tissue", "Partial_R2",␣
↪"Geneid"]].groupby("Geneid").size()
```

[29]:
```
Geneid
ENSG00000013573.16    93
ENSG00000074803.17     3
ENSG00000142856.16    19
ENSG00000164346.9     17
ENSG00000166435.15    12
ENSG00000228906.1     27
ENSG00000255374.3     58
ENSG00000256274.1      3
ENSG00000267370.1      3
ENSG00000270605.1     26
dtype: int64
```

[30]:
```
idv_partial.loc[(idv_partial["Partial_R2"] >= 0.8), ["Tissue", "Partial_R2",␣
↪"Geneid"]].groupby(["Geneid", "Tissue"]).size()
```

[30]:
```
Geneid              Tissue
ENSG00000013573.16  Caudate          93
ENSG00000074803.17  DLPFC             3
ENSG00000142856.16  Caudate           7
                    DLPFC             7
                    Hippocampus       5
ENSG00000164346.9   Caudate          17
ENSG00000166435.15  Caudate           3
                    DLPFC             3
                    Dentate Gyrus     3
                    Hippocampus       3
ENSG00000228906.1   Caudate           9
                    DLPFC             9
                    Hippocampus       9
ENSG00000255374.3   Hippocampus      58
ENSG00000256274.1   Hippocampus       3
ENSG00000267370.1   DLPFC             3
```

```
ENSG00000270605.1   Dentate Gyrus    26
dtype: int64
```

[ ]: