In [1]:
```
# Imports↔
```

In [12]:
```
# Styling
sns.set_context('notebook', font_scale=2.0)
sns.set_style('whitegrid')

def print2(a, b, *args, x=60):
    template = '{:%d}{}' % x
    formatted_template = template.format(a, b)
    for arg in args:
        formatted_template += ' ' + str(arg)
    print(formatted_template)
```

In [3]:
```
%pylab inline
%load_ext autoreload
%autoreload 2
```

```
Populating the interactive namespace from numpy and matplotlib
```

In [4]:
```
suffix = '_2'
```

## Convert drug CIDs to a list of target pairs

In [4]:
```
# Load data↔
```

```
PDX_DrugList_20150729.xlsx
```

**pdx**

| | drug | target | cid |
|---|---|---|---|
| **1** | Linifacnib (ABT-869) | ATP-competitive VEGFR/PDGFR inhibitor | 11485656 |
| **2** | Afatinib (BIBW 2992) | EGFR/HER2 inhibitor | 10184653 |
| **3** | BMS-536924 | ATP-competitive IGF-1R/IR inhibitor | 11353973 |

```
Number of rows:                                         105
```

In [7]:
```
# Find protein targets for query CIDs using STITCH↔
```

**cid2enst**

| | cid | ensp |
|---|---|---|
| **0** | 2244 | 354612 |
| **1** | 2244 | 356438 |
| **2** | 2346 | 241337 |

```
Number of CID -> ENST mappings:                         156
Number of unique CIDs mapped to proteins:               75
Number of missing CIDs:                                 30
```

In [8]:
```
# Manually map missing CIDs↔
```

**cid2enst_manual**

| | drug | target | cid | uniprot_id | ensp_full | gene | protein | ensp |
|---|---|---|---|---|---|---|---|---|
| **0** | BMS-536924 | ATP-competitive IGF-1R/IR inhibitor | 11353973 | P06213 | ENSP00000303830 | INSR | Insulin receptor | 303830 |
| **1** | BMS-536924 | ATP-competitive IGF-1R/IR inhibitor | 11353973 | P06213 | ENSP00000342838 | INSR | Insulin receptor | 342838 |
| **2** | BMS-536924 | ATP-competitive IGF-1R/IR inhibitor | 11353973 | P08069 | ENSP00000268035 | IGF1R | Insulin-like growth factor 1 receptor | 268035 |

```
Number of rows:                                 112
```

**still missing**

| | drug | target | cid | uniprot_id | ensp_full | gene | protein | ensp |
|---|---|---|---|---|---|---|---|---|
| **23** | Cisplatin | inhibit DNA synthesis | 441203 | NaN | NaN | NaN | NaN | NaN |
| **33** | Cytarabine | antimetabolic agent and DNA synthesisinhibitor | 6253 | NaN | NaN | NaN | NaN | NaN |
| **85** | Oxaliplatin | DNA synthesis | 77994 | NaN | NaN | NaN | NaN | NaN |

```
Number of rows:                                 3
```

In [9]:
```
# Add `enst` ids to `pdx` data↔
```

**pdx_wenst**

| | drug | target | cid | ensp |
|---|---|---|---|---|
| **0** | Linifacnib (ABT-869) | ATP-competitive VEGFR/PDGFR inhibitor | 11485656 | 241453 |
| **1** | Linifacnib (ABT-869) | ATP-competitive VEGFR/PDGFR inhibitor | 11485656 | 286301 |
| **2** | Afatinib (BIBW 2992) | EGFR/HER2 inhibitor | 10184653 | 269571 |

```
Number of rows:                                 265
Number of unique CIDs:                          102
Number of unique ENSTs:                         202
```

In [10]:
```
# Create a dataframe containing `borrelidin` and `halofuginone`↔
```

**partner_df**

| | partner_drug | partner_ensp |
|---|---|---|
| **0** | borrelidin | 265112 |
| **1** | borrelidin | 502553 |
| **2** | borrelidin | 455217 |
| **3** | borrelidin | 506040 |
| **4** | borrelidin | 514259 |
| **5** | borrelidin | 626210 |
| **6** | borrelidin | 627006 |
| **7** | halofuginone | 324331 |
| **8** | halofuginone | 274680 |

In [11]:
```
# Join with partner enst↔
```

**pdx_wenst_wpartner**

| | drug | target | cid | ensp | partner_drug | partner_ensp | ensp_1 | ensp_2 | ens |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Linifacnib (ABT-869) | ATP-competitive VEGFR/PDGFR inhibitor | 11485656 | 241453 | borrelidin | 265112 | 241453 | 265112 | (24 265 |
| **1** | Linifacnib (ABT-869) | ATP-competitive VEGFR/PDGFR inhibitor | 11485656 | 241453 | borrelidin | 502553 | 241453 | 502553 | (24 502 |
| **2** | Linifacnib (ABT-869) | ATP-competitive VEGFR/PDGFR inhibitor | 11485656 | 241453 | borrelidin | 455217 | 241453 | 455217 | (24 455 |

```
Number of rows:                                 2385
Number of unique CIDs:                          102
Number of unique ENSTs:                         202
```

In [13]:
```
# Save a copy of the DataFrame to the database to simplify subsequent queries↔
```

In [ ]:
```
# Get scores from the database↔
```

## Get classifier scores for the new target pairs

In [5]:
```
import predictor
import importlib
importlib.reload(predictor)
```

Out[5]:
```
<module 'predictor' from '/home/kimlab1/strokach/working/chemical_interactions/chemical_inte
ractions/predictor.py'>
```

In [6]:
```
# Load the predictor↔
```

```
Log successfully initialized
Reading data from file: /home/kimlab1/strokach/working/chemical_interactions/chemical_intera
ctions/predictor_input/predictor_2.tsv
Adding TargetPair column...
Adding DrugPair column...
Reformatting features...
Done initializing predictor!
```

In [7]:
```
# Load features from the database↔
```

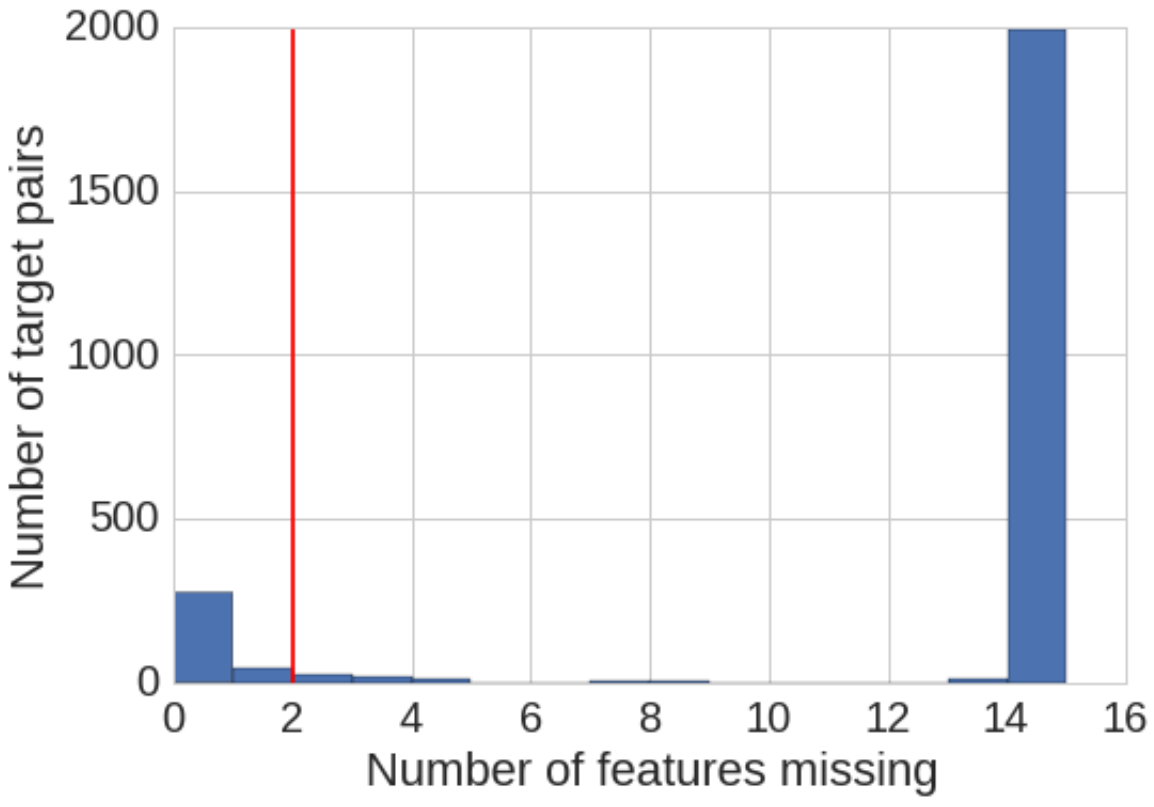
**features_df**

| | drug | target | cid | ensp | partner_drug | partner_ensp | ensp_1 | ensp_2 | ensp_pair | Type | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Evista (Raloxifene HCl) | estrogen antagonist | 54900 | 206249 | borrelidin | 265112 | 206249 | 265112 | (206249, 265112) | Test | ... |
| **1** | Fulvestrant | estrogen receptor (ER) antagonist | 104741 | 206249 | borrelidin | 265112 | 206249 | 265112 | (206249, 265112) | Test | ... |
| **2** | Fulvestrant | estrogen receptor (ER) antagonist | 104741 | 206249 | halofuginone | 274680 | 206249 | 274680 | (206249, 274680) | Test | ... |

3 rows × 25 columns

```
Number of rows:                                 2385
Number of unique CIDs:                          102
Number of unique ENSPs:                         202
```

In [13]:
```
# Remove rows that have too many missing values
```




```
Number of rows with at most 2 nulls                    346
Number of unique CIDs:                                 92
Number of unique ENSPs:                                116
```

In [14]:
```
# Create a dataframe of test data
```

```
Log successfully initialized
Reading data from file: kaist/target_pair_features_nonulls_2.tsv
Adding TargetPair column...
Reformatting features...
```

**pred_test.predictor_df**

| | drug | target | cid | ensp | partner_drug | partner_ensp | ensp_pair | Type | biogrid_shortest_pa |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Evista (Raloxifene HCl) | estrogen antagonist | 54900 | 206249 | borrelidin | 265112 | (206249, 265112) | Test | 2 |
| **1** | Fulvestrant | estrogen receptor (ER) antagonist | 104741 | 206249 | borrelidin | 265112 | (206249, 265112) | Test | 2 |
| **2** | Fulvestrant | estrogen receptor (ER) antagonist | 104741 | 206249 | halofuginone | 274680 | (206249, 274680) | Test | 2 |

3 rows × 23 columns

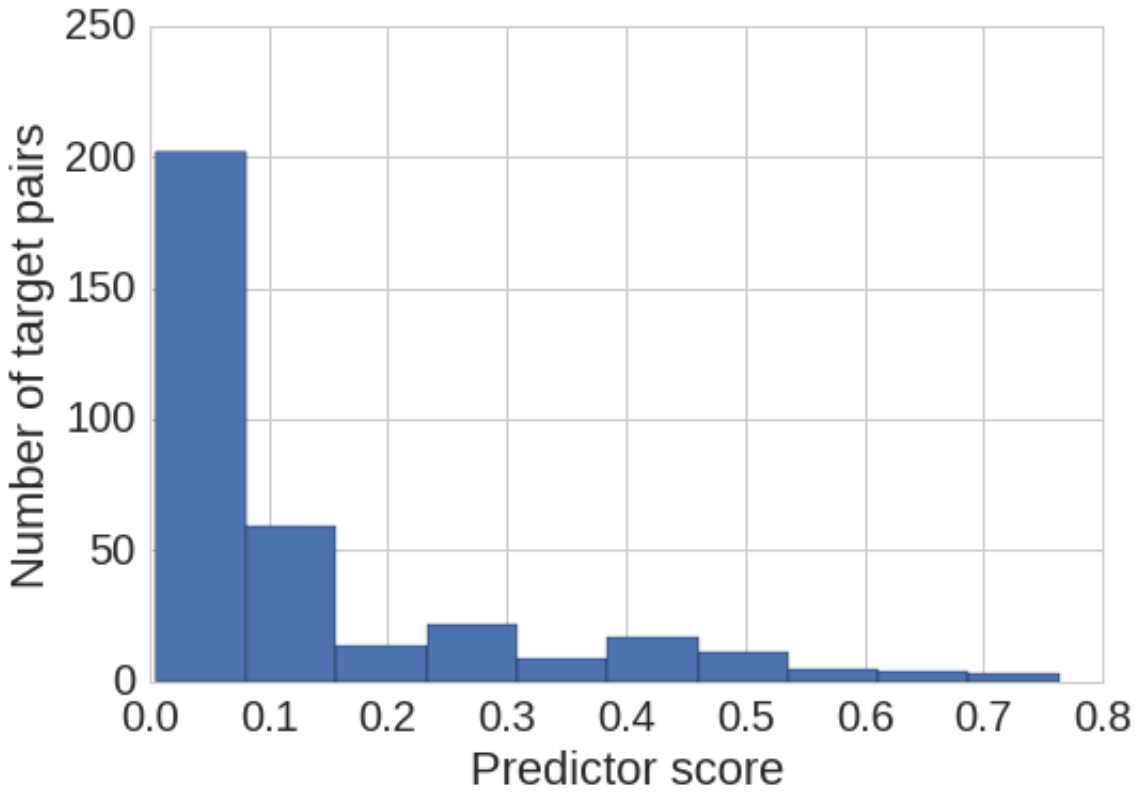
```
Number of rows:                                        346
Number of unique CIDs:                                 92
Number of unique ENSPs:                                116
Number of unique (CID, partner_drug) pairs:            184
```

In [15]:
```
# Fill NA values with means
```

```
Found missing value for biogrid_eb_max
Found missing value for gene_coexpression
Found missing value for getint_shortest_path_length
Found missing value for getint_eb_max
Found missing value for phylogenic_similarity
Found missing value for string_shortest_path_length
Found missing value for string_eb_max
```

In [16]:
```
# Get predictions
additional_columns_to_drop = ['drug', 'target', 'cid', 'ensp', 'partner_drug', 'partner_ensp
features_wpred_df = pred_test.predictor_df.copy()
data_test, labels_test = pred_test.get_data_and_labels(additional_columns_to_drop)
features_wpred_df['probas'] = classifier.predict_proba(data_test)[:,1]
```

In [17]:
```
# Summarize results↔
```




| | drug | target | cid | ensp | partner_drug | partner_ensp | ensp_pair | Type | biogrid_shortest_pa |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Evista (Raloxifene HCl) | estrogen antagonist | 54900 | 206249 | borrelidin | 265112 | (206249, 265112) | Test | 2 |
| **1** | Fulvestrant | estrogen receptor (ER) antagonist | 104741 | 206249 | borrelidin | 265112 | (206249, 265112) | Test | 2 |
| **2** | Fulvestrant | estrogen receptor (ER) antagonist | 104741 | 206249 | halofuginone | 274680 | (206249, 274680) | Test | 2 |

3 rows × 24 columns

```
Number of rows:                                 346
Number of unique CIDs:                          92
Number of unique ENSPs:                         116
```

In [20]:
```
# Pretty format and save the results↔
```

In [21]:
```
# Summarise the stuff that you saved↔
```

**features_wpred_df_all**

| | drug | target | cid | ensp | partner_drug | partner_ensp | num_nulls | probas | biogri |
|---|---|---|---|---|---|---|---|---|---|
| **138** | Etoposide (VP-16) | inhibits DNA synthesis via topoisomerase II i... | 36462 | 311032 | borrelidin | 265112 | 0 | 0.763211 | 2 |
| **269** | Etoposide (VP-16) | inhibits DNA synthesis via topoisomerase II i... | 36462 | 311032 | halofuginone | 274680 | 0 | 0.720238 | 3 |
| **252** | Nutlin-3a | inhibits thep53/MDM2 interaction | 11433190 | 293288 | halofuginone | 274680 | 0 | 0.711900 | 3 |

3 rows × 21 columns

```
Number of rows:                                  346
Number of unique CIDs:                           92
Number of unique ENSTs:                          116
Number of unique (CID, partner_drug) pairs:      184
```

**features_wpred_df_highest**

| | drug | target | cid | ensp | partner_drug | partner_ensp | num_nulls | probas | biogri |
|---|---|---|---|---|---|---|---|---|---|
| **138** | Etoposide (VP-16) | inhibits DNA synthesis via topoisomerase II i... | 36462 | 311032 | borrelidin | 265112 | 0 | 0.763211 | 2 |
| **269** | Etoposide (VP-16) | inhibits DNA synthesis via topoisomerase II i... | 36462 | 311032 | halofuginone | 274680 | 0 | 0.720238 | 3 |
| **252** | Nutlin-3a | inhibits thep53/MDM2 interaction | 11433190 | 293288 | halofuginone | 274680 | 0 | 0.711900 | 3 |

3 rows × 21 columns

```
Number of rows:                                  184
Number of unique CIDs:                           92
Number of unique ENSTs:                          81
Number of unique (CID, partner_drug) pairs:      184
```

In [ ]:
```

```

In [46]:
```
x = 311032
x_formatted = 'ENSP{:011n}'.format(x)
print(x_formatted)
display(HTML('<a href="http://www.ensembl.org/id/{0}">{0}</a>'.format(x_formatted)))
```

```
ENSP00000311032
```

ENSP00000311032 (http://www.ensembl.org/id/ENSP00000311032)

In [38]:
```
from bioservices import ensembl
e = ensembl.Ensembl()

res = e.get_archive(x_formatted)
print(res)

res = e.post_lookup_by_id(x_formatted, expand=False)
print(res)
```

```
{'latest': 'ENSP00000302967.3', 'possible_replacement': [], 'type': 'Translation', 'is_curre
nt': '1', 'assembly': 'GRCh38', 'release': '81', 'id': 'ENSP00000302967', 'peptide': None,
'version': '3'}
{'ENSP00000302967': {'end': 141636790, 'id': 'ENSP00000302967', 'start': 141621468, 'objec
t_type': 'Translation', 'species': 'homo_sapiens', 'length': 428, 'db_type': 'core'}}
```

In [ ]:
```

```

In [ ]:
```

```