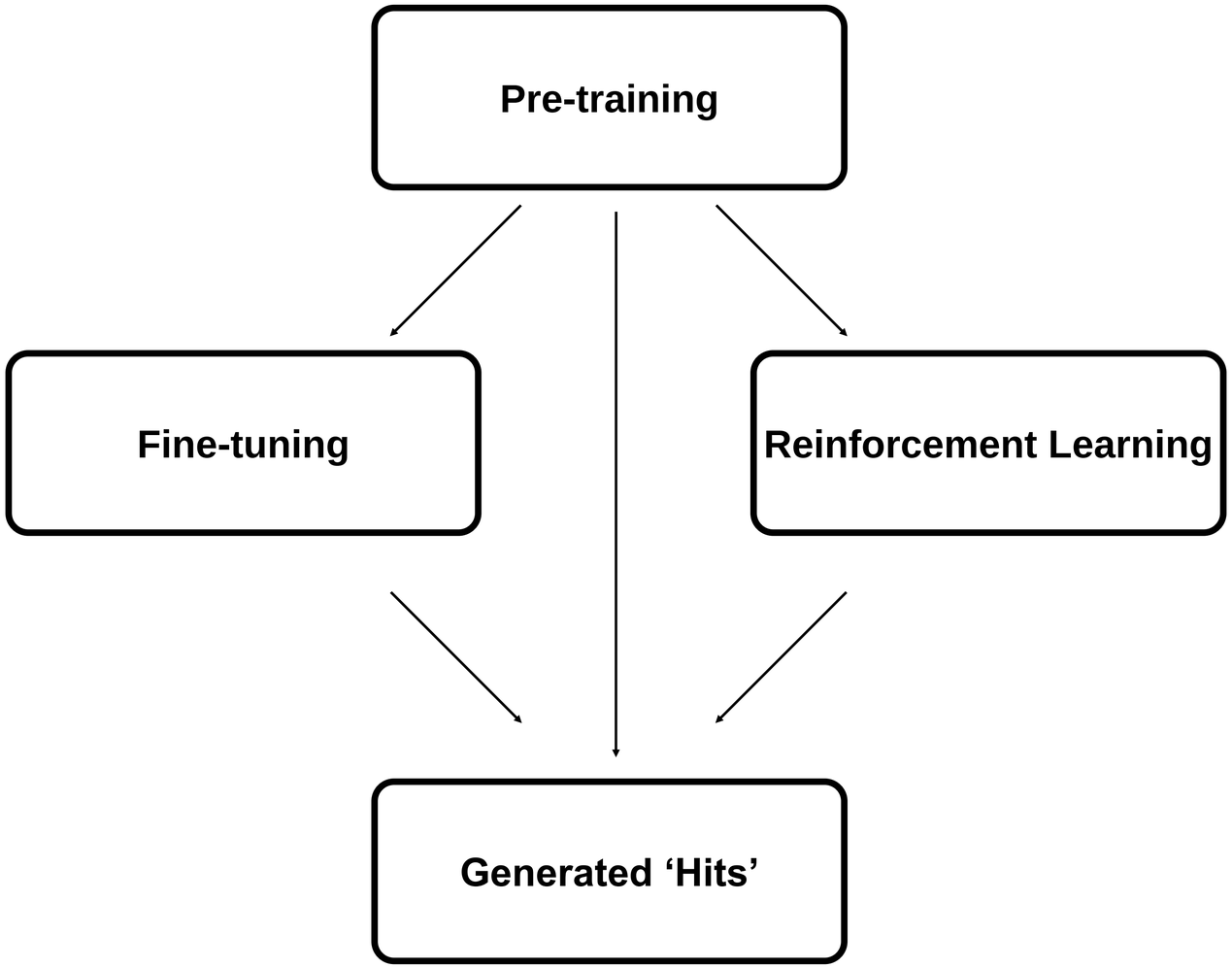
Generative Modeling



Pre-training: Learn SMILES grammar and broad patterns, e.g. frequency of halogens, typical molecular weight, etc. Large, diverse dataset preferred

Fine-tuning: Bias toward patterns of interest by exposing model only to compounds satisfying certain properties, e.g. containing a desired substructure. Need a dataset of such compounds, possibly from virtual screening

Reinforcement Learning: Similar to fine-tuning except the dataset used for biasing is generated on-the-fly. Only need a scoring function specifying the properties of interest

Sample Hit Libraries

Pre-training datasets:

- ChEMBL (1.5M),
- Enamine diverse REAL drug-like (20M)

Fine-tuning datasets:

- Virtually screened hits from Enamine 20M (100k)

Library A

Pre-trained on ChEMBL

Hits from filtered, unbiased model using pharmacophore scoring function

~15k

Library B

Virtual screening hits from Enamine 20M

~15k

Library C

Pre-trained on ChEMBL

Use Library B to fine-tune generative model

Filter generated compounds for hits

~15k

Library D

Pre-trained on Enamine 20M

Hits from filtered, unbiased model using pharmacophore scoring function

~1.5k

Library E

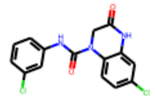
Pre-trained on ChEMBL

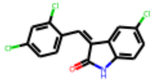
Reinforcement learning using pharmacophore scoring function

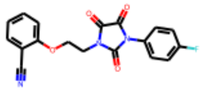
~15k

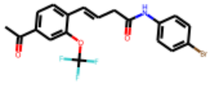
Library A: ChEMBL pre-trained, filtered

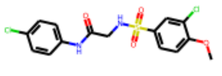
40 random samples

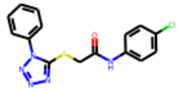

13707_Library_A

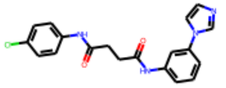

14207_Library_A

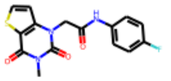

7929_Library_A

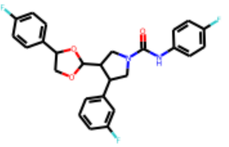

370_Library_A

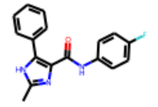

4838_Library_A

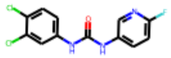

9620_Library_A

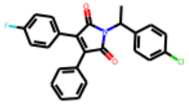

8691_Library_A

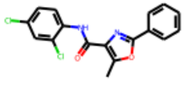

7519_Library_A

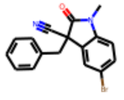

12009_Library_A

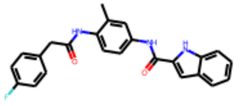

7206_Library_A

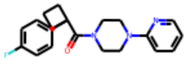

12583_Library_A

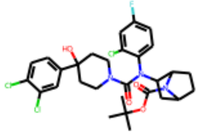

1113_Library_A

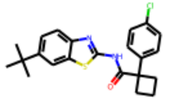

7424_Library_A

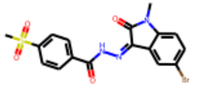

3011_Library_A

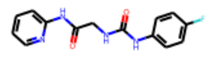

6085_Library_A

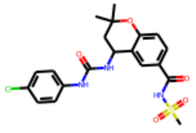

10391_Library_A

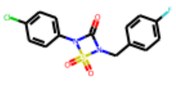

570_Library_A

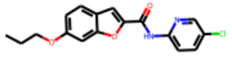

641_Library_A

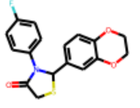

3004_Library_A

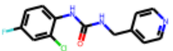

9012_Library_A

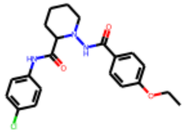

1191_Library_A

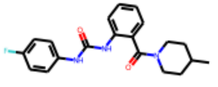

13837_Library_A

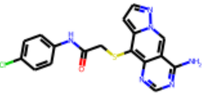

1824_Library_A

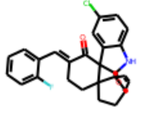

13783_Library_A

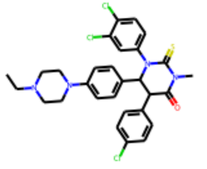

10672_Library_A

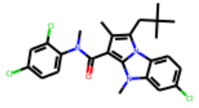

2227_Library_A

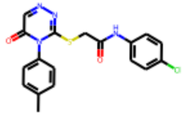

1524_Library_A

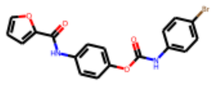

8372_Library_A

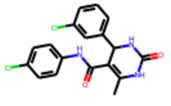

14165_Library_A

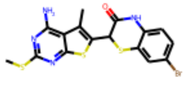

1993_Library_A

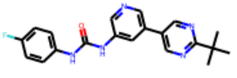

5521_Library_A

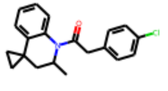

6221_Library_A

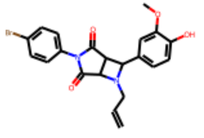

10941_Library_A

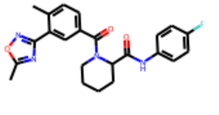

1354_Library_A

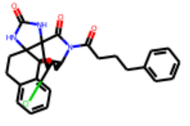

5459_Library_A


648_Library_A


1496_Library_A

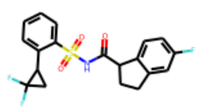

59_Library_A

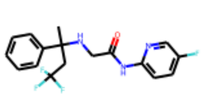

7176_Library_A

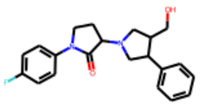

14011_Library_A

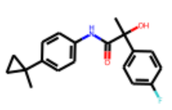
Library B: Enamine filtered

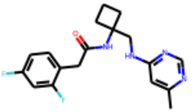
40 random samples

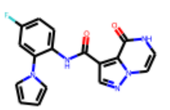

13800_Library_B

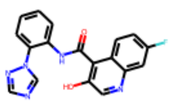

7326_Library_B

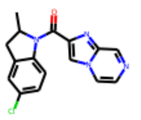

13212_Library_B

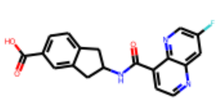

2751_Library_B

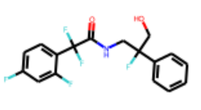

8707_Library_B

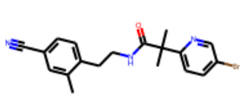

11716_Library_B

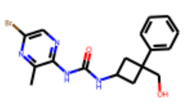

6980_Library_B

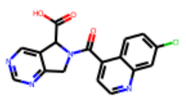

2816_Library_B

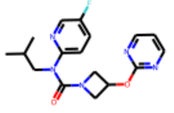

399_Library_B

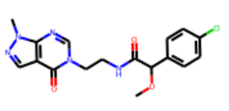

8093_Library_B

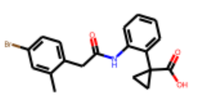

3931_Library_B

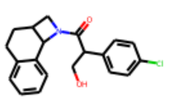

12333_Library_B

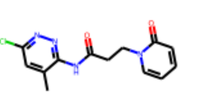

6140_Library_B

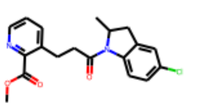

11475_Library_B

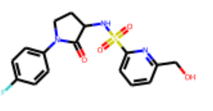

7525_Library_B

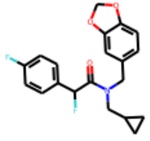

2666_Library_B

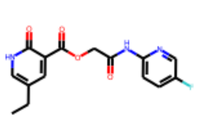

13528_Library_B

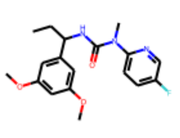

7307_Library_B

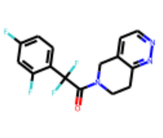

5772_Library_B

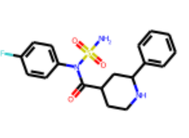

2486_Library_B

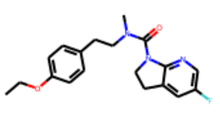

9667_Library_B

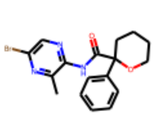

10352_Library_B

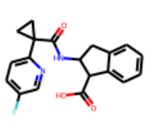

7629_Library_B

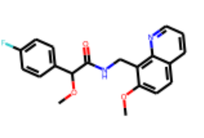

12533_Library_B

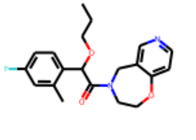

1501_Library_B

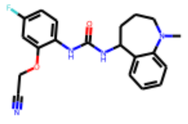

13739_Library_B

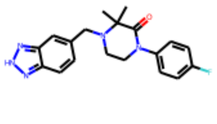

13580_Library_B

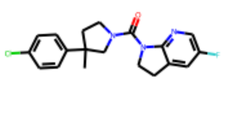

13475_Library_B

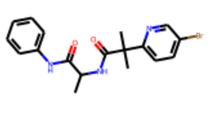

8395_Library_B

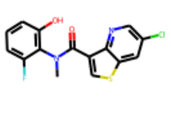

7351_Library_B

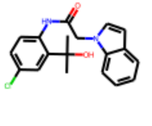

4338_Library_B

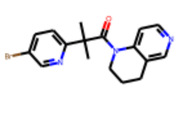

7087_Library_B

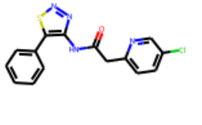

65_Library_B

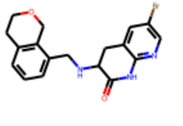

5293_Library_B

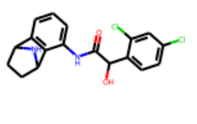

942_Library_B


6687_Library_B


9175_Library_B

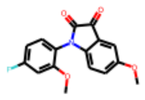

12229_Library_B

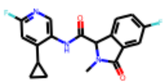

1606_Library_B

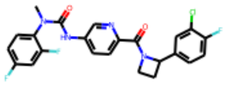

80_Library_B

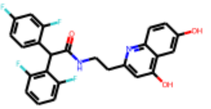
Library C: ChEMBL pre-trained, library B fine-tuned

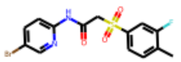
40 random samples

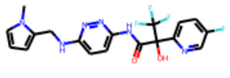

7213_Library_C

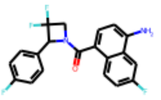

5226_Library_C

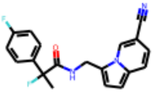

4540_Library_C

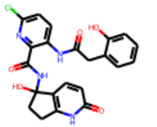

12238_Library_C

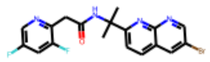

8924_Library_C

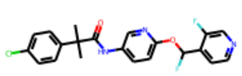

9878_Library_C

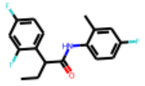

10449_Library_C

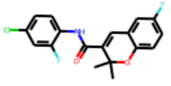

1377_Library_C

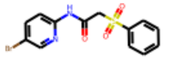

11801_Library_C

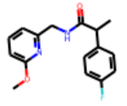

981_Library_C

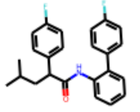

821_Library_C

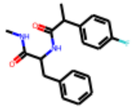

2337_Library_C

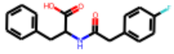

1812_Library_C

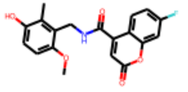

11284_Library_C

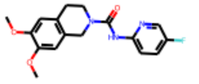

7390_Library_C

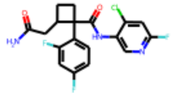

1094_Library_C

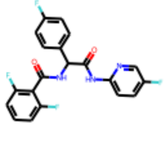

5342_Library_C

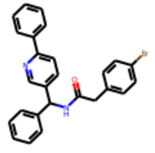

11600_Library_C

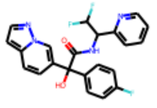

7121_Library_C

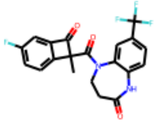

6472_Library_C

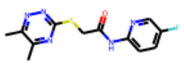

10209_Library_C

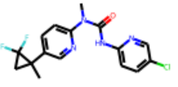

12108_Library_C

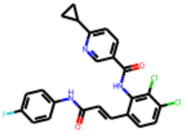

11429_Library_C

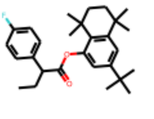

12133_Library_C

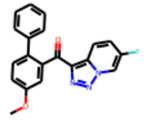

1723_Library_C

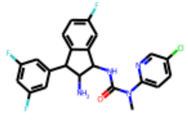

9713_Library_C

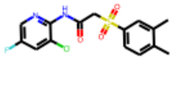

4552_Library_C

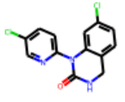

10683_Library_C

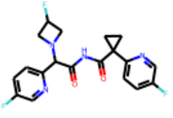

2388_Library_C

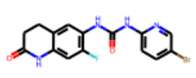

6519_Library_C

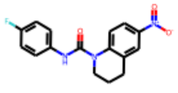

4530_Library_C

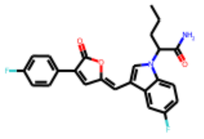

8926_Library_C

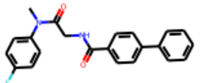

14699_Library_C

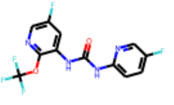

12044_Library_C

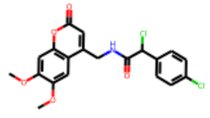

14500_Library_C


13111_Library_C


2644_Library_C

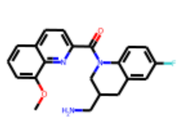

4372_Library_C

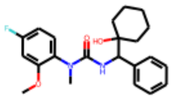

13479_Library_C

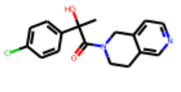

6499_Library_C

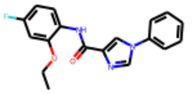
Library D: Enamine pre-trained, filtered

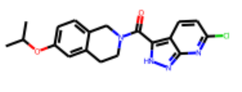
40 random samples

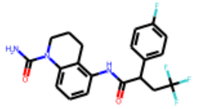

760_Library_D

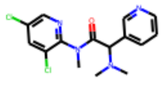

689_Library_D

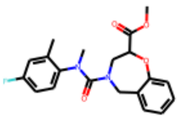

179_Library_D

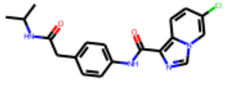

352_Library_D

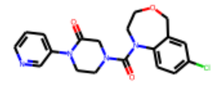

110_Library_D

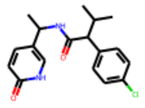

1171_Library_D

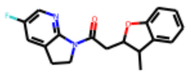

417_Library_D

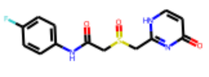

526_Library_D

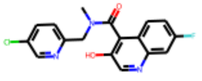

105_Library_D

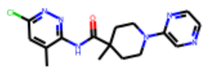

1492_Library_D

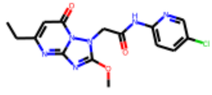

156_Library_D

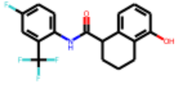

248_Library_D

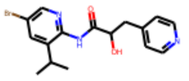

1245_Library_D

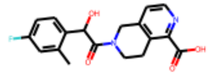

480_Library_D

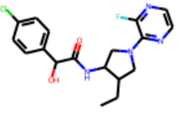

853_Library_D

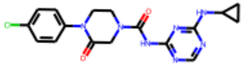

365_Library_D

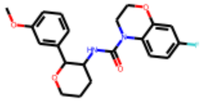

1376_Library_D


120_Library_D

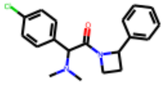

908_Library_D

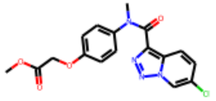

281_Library_D

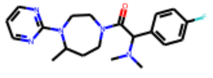

1436_Library_D

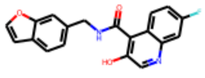

742_Library_D

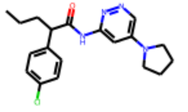

211_Library_D

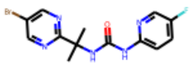

451_Library_D

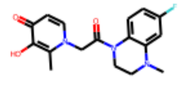

534_Library_D

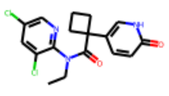

225_Library_D

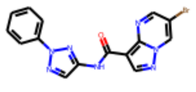

1337_Library_D

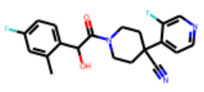

283_Library_D

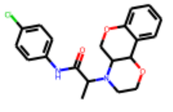

67_Library_D

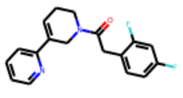

806_Library_D

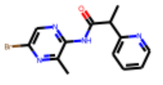

303_Library_D

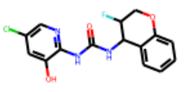

1422_Library_D

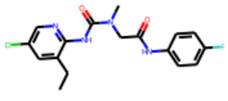

902_Library_D

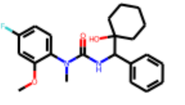

27_Library_D

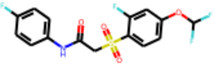

1269_Library_D


1061_Library_D


1429_Library_D

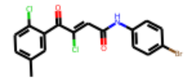

375_Library_D

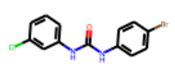

689_Library_D

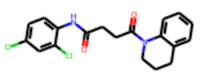

1244_Library_D

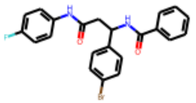
Library E: ChEMBL pre-trained, reinforcement learning

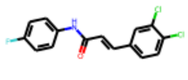
40 random samples


5088_Library_E

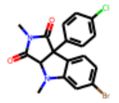

10085_Library_E

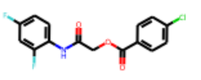

6813_Library_E

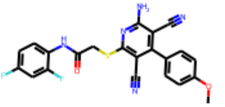

6725_Library_E

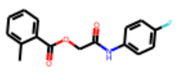

6569_Library_E

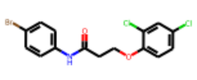

4768_Library_E

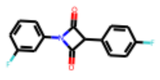

2710_Library_E

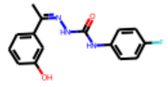

7711_Library_E

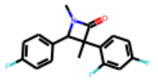

3348_Library_E

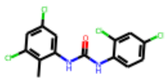

5482_Library_E

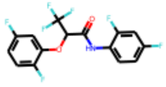

6999_Library_E

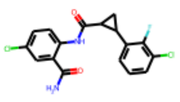

12702_Library_E

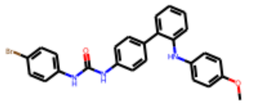

81_Library_E

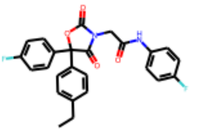

2677_Library_E

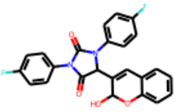

4407_Library_E

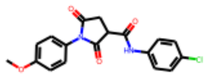

11648_Library_E

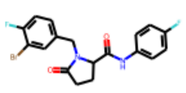

6248_Library_E

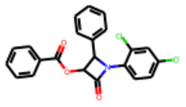

3864_Library_E

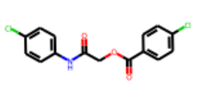

2058_Library_E

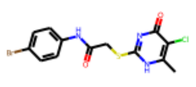

12590_Library_E

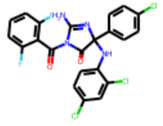

3769_Library_E

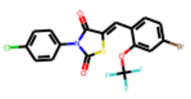

11300_Library_E

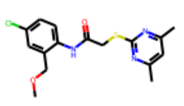

12172_Library_E

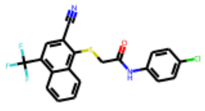

7708_Library_E

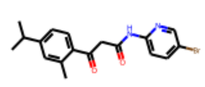

4277_Library_E

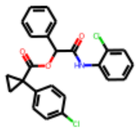

6306_Library_E


13879_Library_E

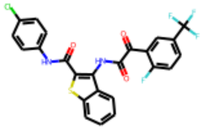

2923_Library_E

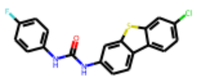

6133_Library_E

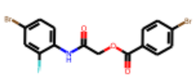

4558_Library_E

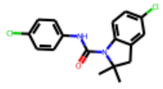

11997_Library_E

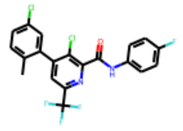

11358_Library_E

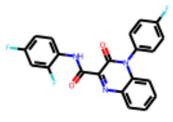

9918_Library_E

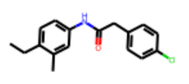

11428_Library_E

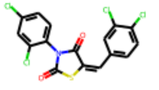

7684_Library_E


1139_Library_E


5101_Library_E


11839_Library_E

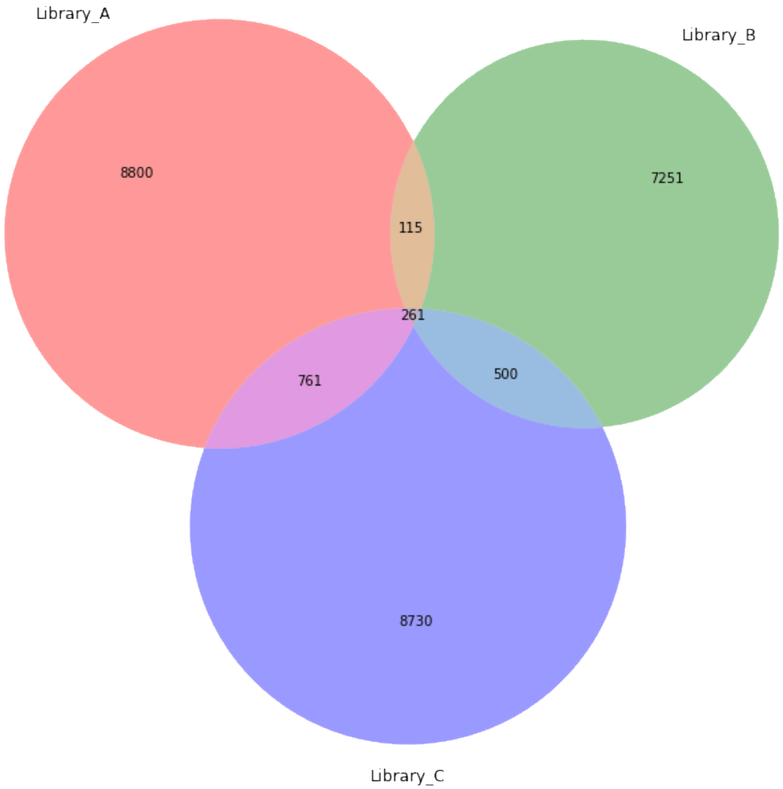

2073_Library_E

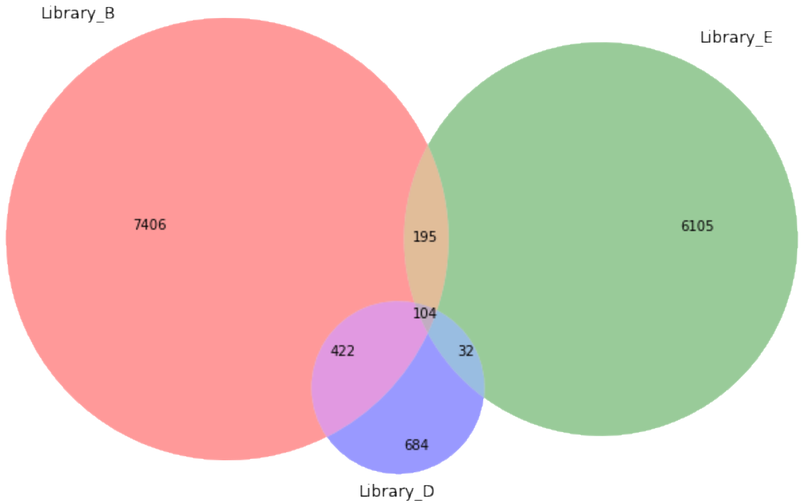

13883_Library_E

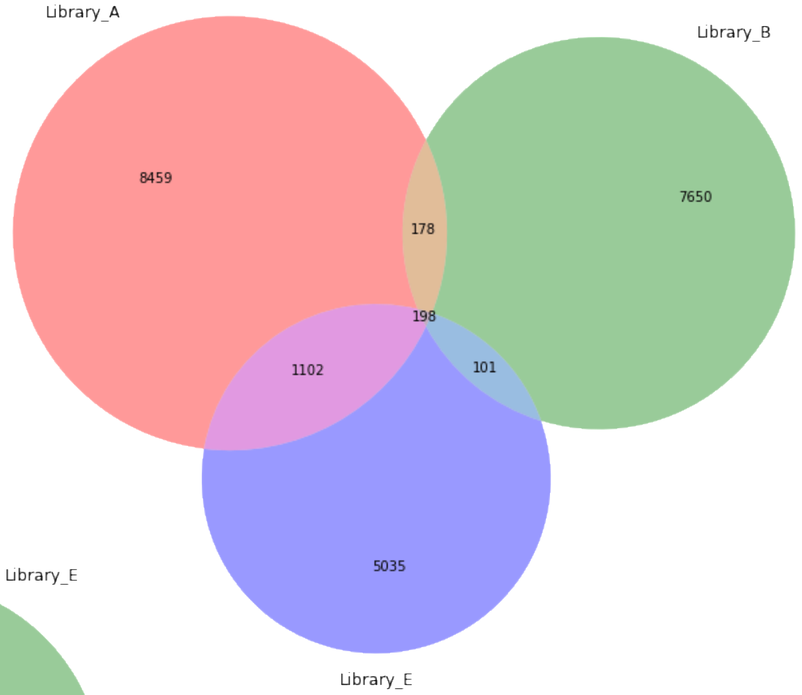
Murcko Scaffold Overlap

Library A: ChEMBL pre-trained, filtered
Library B: Enamine filtered
Library C: ChEMBL pre-trained, library B fine-tuned
Library D: Enamine pre-trained, filtered
Library E: ChEMBL pre-trained, reinforcement learning

Find all unique Murcko scaffolds present in each library, compare overlap between libraries







Generally low overlap between scaffolds in each library