SUPPORTING INFORMATION

Mapping Ligand Shape Space for Protein-Ligand Systems; Distinguishing Key-in-Lock and Hand-in-Glove Proteins

Joanna Zarnecka[†], Iva Lukac[†], Stephen J. Messham[†], Alhusein Hussin[†], Francesco Coppola[§], Steven J. Enoch[†], Alexander G. Dossetter[+], Edward J. Griffen[+], Andrew G. Leach[,†,+,§]*

† School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool, L3 3AF, UK.

+ MedChemica Limited, Biohub, Mereside, Alderley Park, Macclesfield, SK10 4TG, UK.

§ Division of Pharmacy and Optometry, School of Health Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester, M13 9PT

andrew.leach@manchester.ac.uk

S1) Creating and benchmarking an implementation of the shape fingerprint method

a) Designing the set of reference shapes

In the original description of the shape fingerprint approach,[1] two sets of reference shapes were generated, one from the Cambridge structural database of small molecule crystal structures,[2, 3] and the other from a set of conformations generated for the MDDR database of molecules that have been studied clinically. Our interest is principally with protein-ligand interactions and so we chose instead to use the database of ligands studied by X-ray crystallography in complex with a protein – the Ligand Expo dataset.[4] At the time, this contained the experimental coordinates for 1,158,763 non-polymer molecules and non-standard amino acids and nucleotides. Various filtering criteria based on molecular weight were applied to these molecules (Figure S1), leading to 9 databases of shapes that were considered as input to the algorithm that generates the sets of reference shapes. An alternative filtering based on the number of heavy atoms was also performed and yielded similar sets.

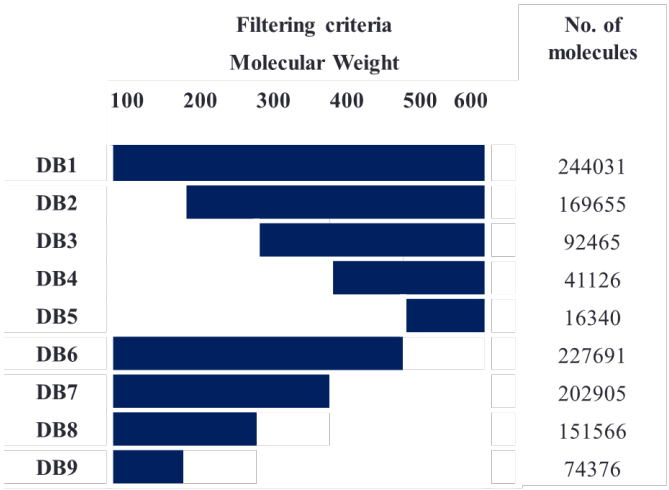


Figure S1. The Ligand Expo Dataset was filtered in 9 different ways according to the lower and upper limit of molecular weight shown. The number of structures to pass the filter criteria is shown.

An implementation of the algorithm described by Haigh *et al*. permitted each set of ligand structures to be clustered in such a way that every shape has at least one reference shape that is similar in shape to it, above a user-selected cut-off (called the Design Tanimoto, DT).[1] In this case, shape comparisons are performed with Openeye's gaussian-based approach, as implemented in ROCS.[5] A randomly selected shape is the first reference shape and its Shape Tanimoto (ST) with every other shape in the input database is computed. All those that have ST values above DT are rejected from further consideration. After all comparisons have been made, the shape that has the lowest Shape Tanimoto with the starting shape is selected and becomes the next reference shape in the database. The process is repeated until all shapes have either been rejected or selected as a reference shape. This ensures that all shapes in the database are more similar than DT to at least one of the reference shapes. A description of all computational methods is provided in Section S2.

b) Optimizing and selecting the set of reference shapes.

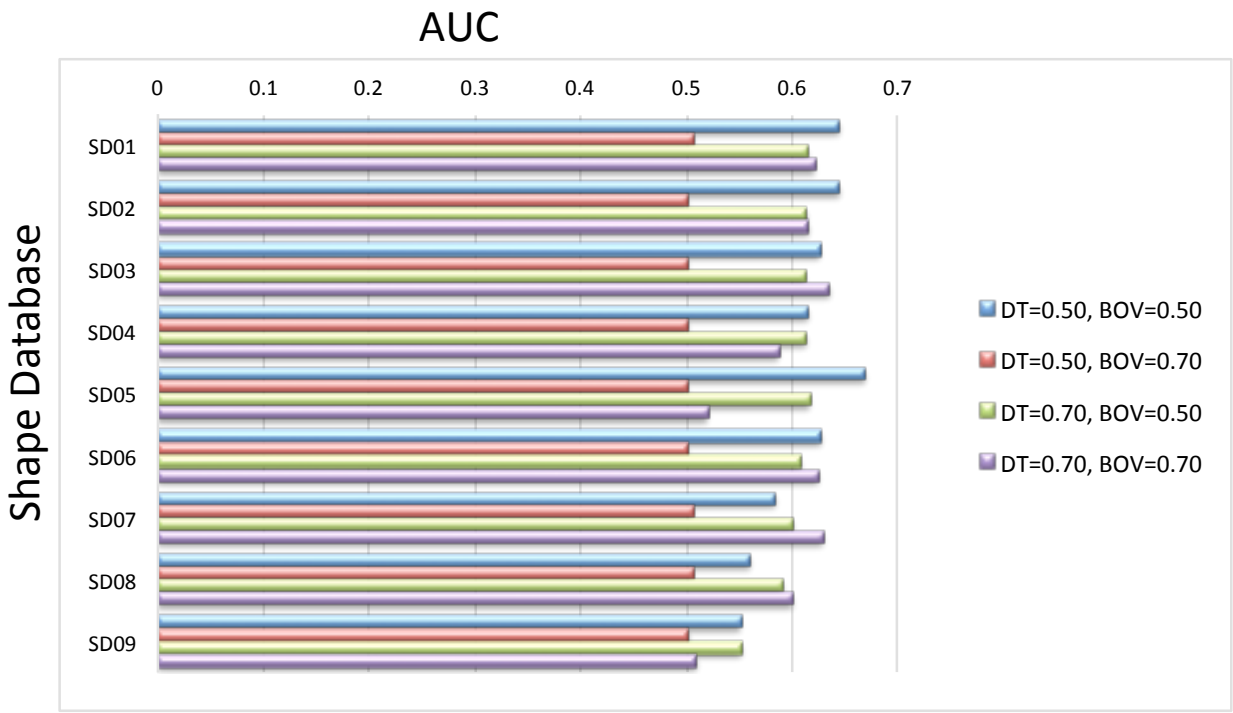
The alternative sets of reference shapes were evaluated by computing their ability to correctly group the molecules in two test sets (a full list of structures is in Section S3). The first test set comprises a set of 87 molecules each of which is known to bind to one of ten proteins and was devised and validated for testing pharmacophore methods.[6] A second test set was selected starting from structures in the Astex diversity set and comprised 45 molecules binding to four different proteins.[7] Both test sets include only molecules with known protein-ligand structures and hence ligand bioactive conformations.

Shape fingerprints are generated from the reference shapes by computing the ST between a query structure and every shape in the Shape Database. When this ST is above a user-defined cut-off, the Bit On Value (BOV), the bit is set to 1 otherwise it is set to 0. Lower BOVs lead to higher bit

densities. In the initial testing of the shape databases, a high and low value (0.7 and 0.5 respectively) for each of DT and BOV were used.

The shape fingerprints for every molecule in both test sets were compared to those for every other molecule in the set. The comparison yielded another Tanimoto, the Fingerprint Tanimoto (FT). Receiver Operating Characteristic curves (ROC curve) were created that plot the true positive rate against the false positive rate, where in this case a true positive corresponds to retrieval of an active in the ordering of actives and decoys according to FT values while a false positive corresponds to retrieval of a decoy instead.[8] The computed AUC (Area Under Curve) for these is a measure of accuracy, where 0.9 – 1 represents a perfect test while 0.5 represents a poor one (equivalent to selecting at random).

The AUCs for Test Sets 1 and 2 computed with a range of different sets of reference shapes (shape databases, SDs, y-axis), DTs and BOVs (color coding) are shown in Figure S2. When DT=0.5 and BOV=0.7, the AUC is rarely distinct from 0.5 suggesting no discrimination is achieved. The combination of low DT and high BOV leads to a low number of bits set on and so these are unlikely to be able to connect molecules (which requires bits to be set in common).



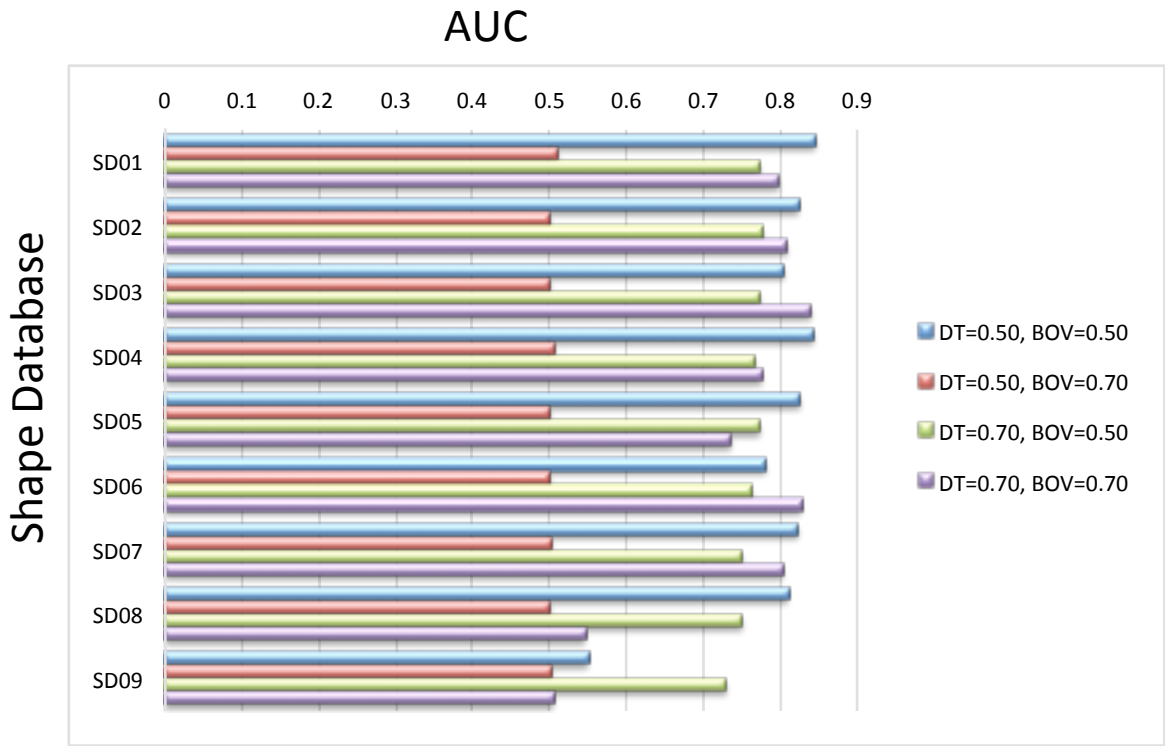


Figure S2. The AUC values for Test Set 1 (Top) and Test Set 2 (Bottom) when applying different settings: DT=0.5 with BOV=0.5 and 0.7, and DT=0.7 with BOV=0.5 and 0.7.

When the average AUC values obtained from both test sets are computed (excluding DT=0.5, BOV=0.7), SD03 and SD06 (with AUCs of 0.71 and 0.70 respectively) stand out as best when SD01, the unfiltered shape database, is excluded. The filtering criteria used to generate SD03 and SD06 were therefore combined to generate Shape Database 10 with molecular weight in the range 300 to 500 with the expectation that this would provide the best balance of accuracy and speed (76125 molecules pass the filters for consideration in the database generation process, compared to 244031 in SD01, 92465 in SD03 and 227691 in SD06).

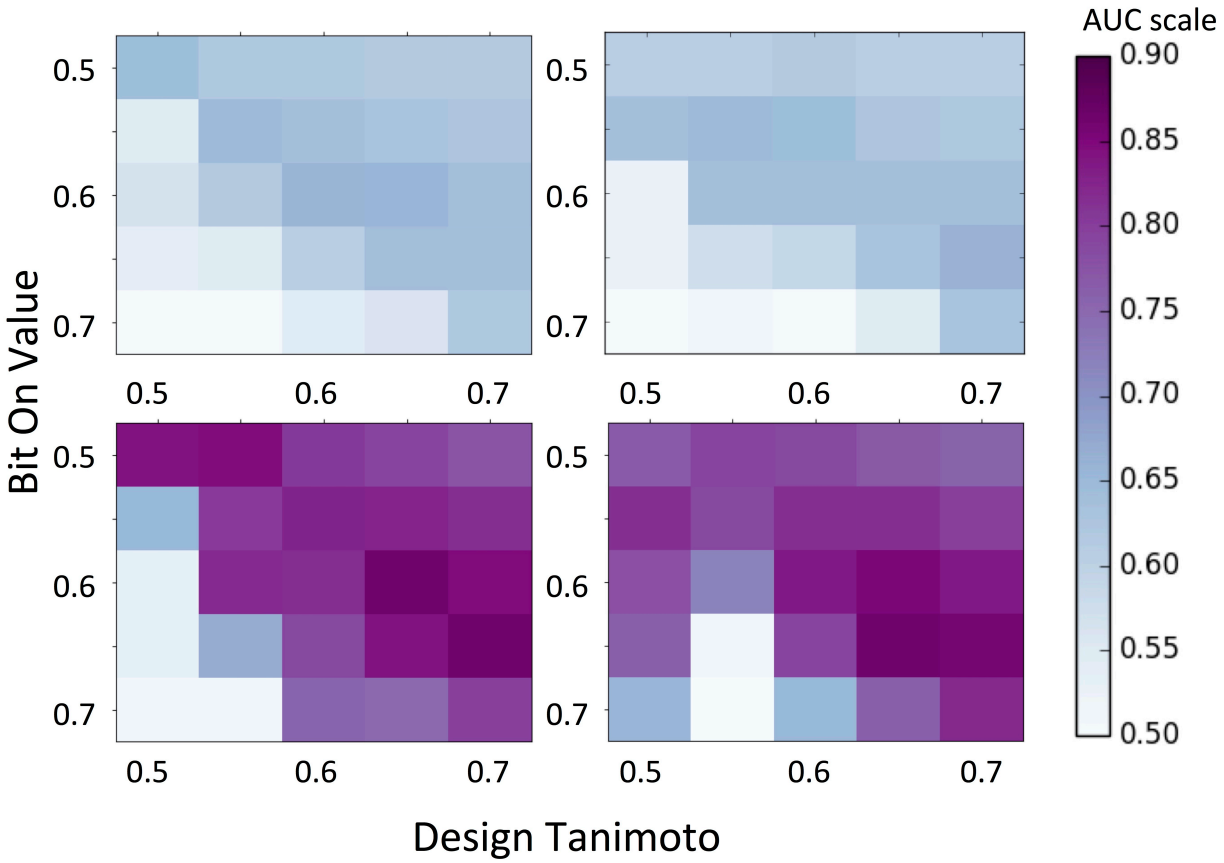


Figure S3. Heatmaps of AUC values for Test Set 1 (Top) and Test Set 2 (Bottom) set when using SD1 (Left) and SD10 (Right) with varying DT and BOV.

The values of DT and BOV were then systematically varied in steps of 0.05 between 0.5 and 0.7 for the largest (SD01) and focused (SD10) shape databases. As can be seen in Figure S3, the AUC values vary less for Test Set 1 (Top), than for Test Set 2 (Bottom). This might be caused by differences in molecular weight distribution in both sets.[9] Experience with other datasets had shown that small molecules (about 200 Da and below) and large molecules (about 800 Da and above) set very few (or no) bits and so cannot be correctly described by these shape fingerprints. The molecular weight distributions for the molecules in the two test sets used in the present study are shown in Figure S4. In Test Set 2, the range is slightly wider than for Test Set 1 and this may facilitate the correct grouping of Test Set 2; size and shape are linked properties. The generally good performance suggests that the shape databases obtained are applicable to molecules spanning the molecular weight range ~200 to ~500 and should therefore be useful for most drug-like molecules.

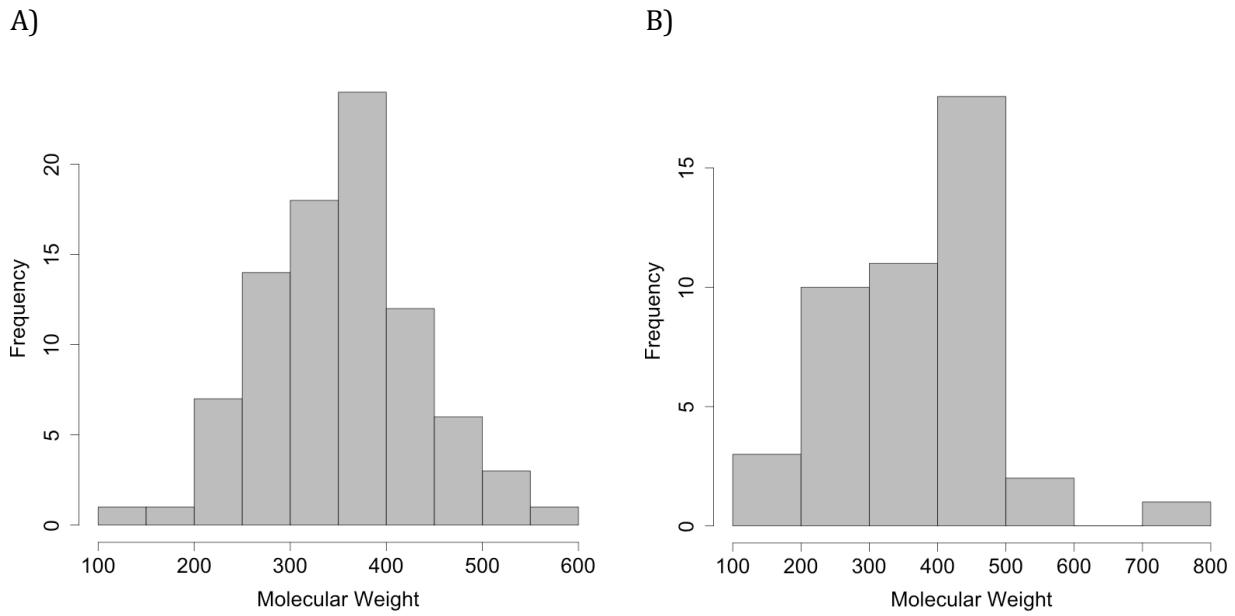


Figure S4. The molecular weight distribution of Test Set 1 (A) and Test Set 2 (B).

The results illustrated in Figure S3 show that using BOV=0.55 gives the best results on average for both shape databases (SD01 and SD10) and thus the impact of DT was

considered while BOV was fixed. The choice of the best DT requires a consideration of the size of the Shape Database, which determines the computational time required to generate each fingerprint. When the average AUC value is viewed as a function of the size of Shape Database (Figure S5), the difference in average AUC value for the two highest values of DT for both test sets is quite small (ΔAUC is 0.022 and 0.003 for Test Set 1 and Test Set 2 respectively), yet the difference in size of the Shape Databases is significant (1346 reference shapes). Therefore, SD10 with DT = 0.65 is selected as the best performing Shape Database. Our results (Figure S3) show that for this setting of DT, the optimum BOV is 0.6. Our recommendation for generating shape fingerprints is therefore to use SD10 with DT=0.65 and BOV=0.6 when grouping molecules according to their likelihood of binding to the same protein. The logistic regression plot for the two test sets in Figure S6 indicates the probability of shared biological activity as the FT cutoff changes. In addition to computing the AUC values, their 95 % confidence interval ranges were also computed. These are provided in Section S4 (Tables S5 to S8). These data show that changing either the DT or BOV by 0.05 in either direction does not cause a statistically significant variation in the AUC values and therefore that these settings, although optimum in our studies, can be varied slightly without major detriment. In their earlier studies, Haigh *et al.* found that as DT was increased, their method was able to better recapitulate ROCS ordering of shape similarities until a plateau was reached at a DT value of 0.75.[1] Their reference shapes were small molecule crystal structures or conformations derived from SMILES strings and so the change in optimum settings is not unexpected; in our system a DT of 0.75 would very significantly increase the time to generate the fingerprints. Their observation that BOV should be lower than DT holds true in our studies also.

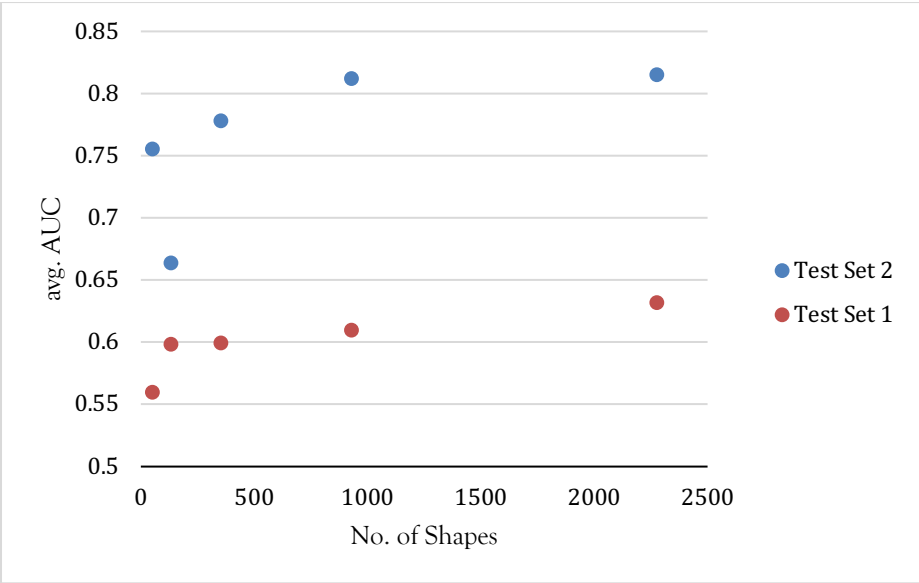


Figure S5. The variation of average AUC value with the size of the Shape Database for both test sets is shown when Shape Database 10 was used and the value of DT was changed (averaged over the set of five BOVs that were employed).

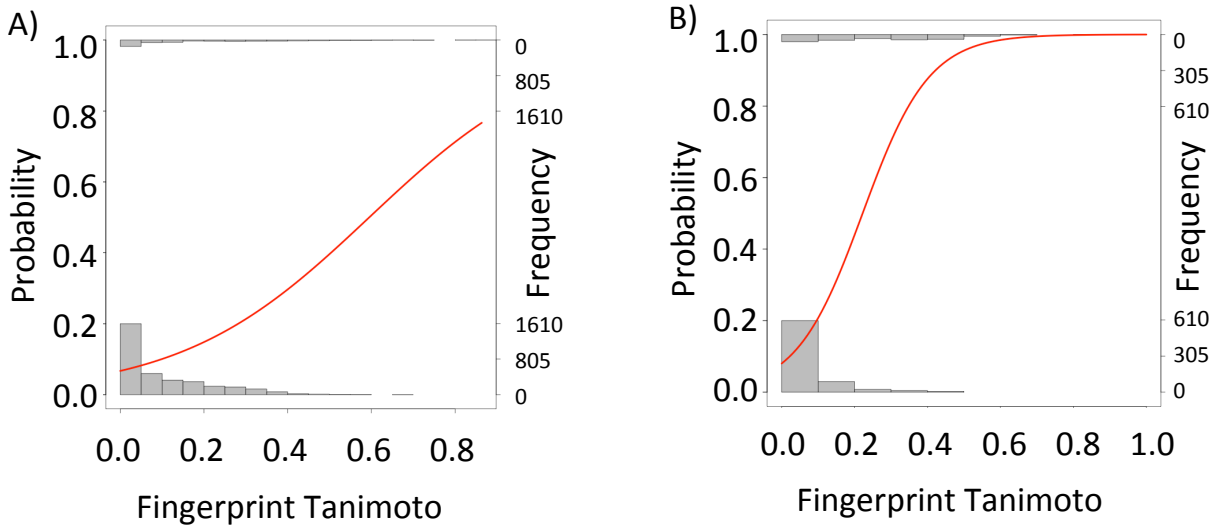


Figure S6. Logistic regression plot showing the probability of shared biological activity at differing FT values for Test Set 1 (A) and Test Set 2 (B) when using SD10 with DT=0.65 and BOV=0.60. A histogram of values for molecules that share biological activity is shown at the top of the diagram and that for molecules that do not share biological activity at the bottom.

Having identified an optimum set of reference shapes and a method to permit the relative shape comparisons possible with ROCS to create an absolute shape description, it is useful to define a cut-off value of FT above which molecules have a defined likelihood of sharing biological activity. Merging the two datasets (giving an evaluation based on 14 protein targets) and performing logistic regression on the combined test set (Figure S7), suggests that a value of FT above 0.45 might be a reasonable estimate of when compounds are more likely to share biological activity than not; an alternative value is proposed in the main text for clustering. Identifying such boundary values has always been challenging but as with the equivalent values for 2D-fingerprints, the appropriate value will depend upon the application at hand.[10]

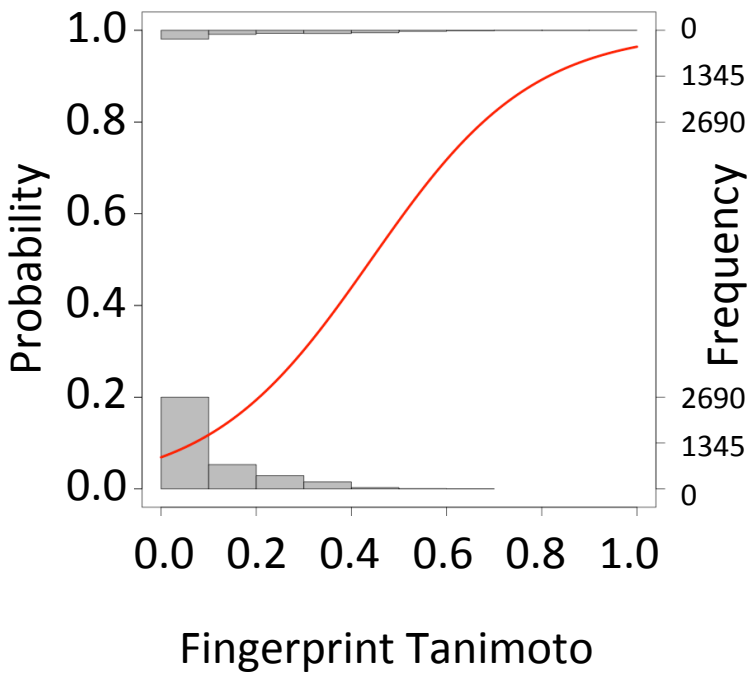


Figure S7. Logistic regression plot showing the probability of shared biological activity at differing FT values for Test Set 1 and Test Set 2 combined together when using SD10 with DT=0.65 and BOV=0.60. A histogram of values for molecules that share biological activity is

shown at the top of the diagram and that for molecules that do not share biological activity at the bottom.

i) Resampling SD10.

Given that the selection of reference shapes begins with a random choice, it is possible that the results are dependent upon this starting point. Therefore, SD10 was regenerated with DT=0.65 ten more times. Each of the new shape databases was used to generate fingerprints for both test sets and the AUC value was recomputed. The results show little variation (standard deviations for the distribution of AUC values obtained for resampled shape databases vary from 0.002 to 0.033 depending on BOV) and can be seen in Tables S9 and S10 in Section S4. On average, with DT=0.65 the best performance is found when using BOV=0.60; AUC values of 0.64 and 0.84 for Test Set 1 and Test Set 2 respectively, are obtained.

c) Conformations.

For many shape comparisons that might be of interest, a protein-ligand crystal structure would not be a useful requirement. Therefore, conformations were generated from the SMILES string for each molecule in the test sets using Openeye's OMEGA software,[9] a knowledge-based conformer generator. In the first instance, relatively limited sets of up to five conformations were generated although some molecules in the set were conformationally restricted and generated less than this. Other options in OMEGA were set to default. For all the conformations, shape fingerprints were generated using Shape Database 10.

Two approaches for evaluating the comparison of two molecules, each with multiple conformations, were investigated: 1) the Maximum Value (MV) of FT among the array arising from comparisons of all conformations of one molecule with all conformations of the other or 2) the Average Value (AV) of the FTs. As shown in Figure S8, there is only a small difference in AUC values between the methods (AV and MV), with the MV being slightly better. This is consistent with molecules requiring only one (energetically accessible) conformation to be similar in shape in order to share biological activity. Comparing the AUC values obtained for conformations generated from SMILES with those for crystal structures shows only a little deterioration (Table S1). Thus, using conformations generated from SMILES instead of crystal structures does not greatly affect the accuracy of the shape fingerprint method. This shows that the method can be successfully used even when the bioactive conformation of the ligand is not known.

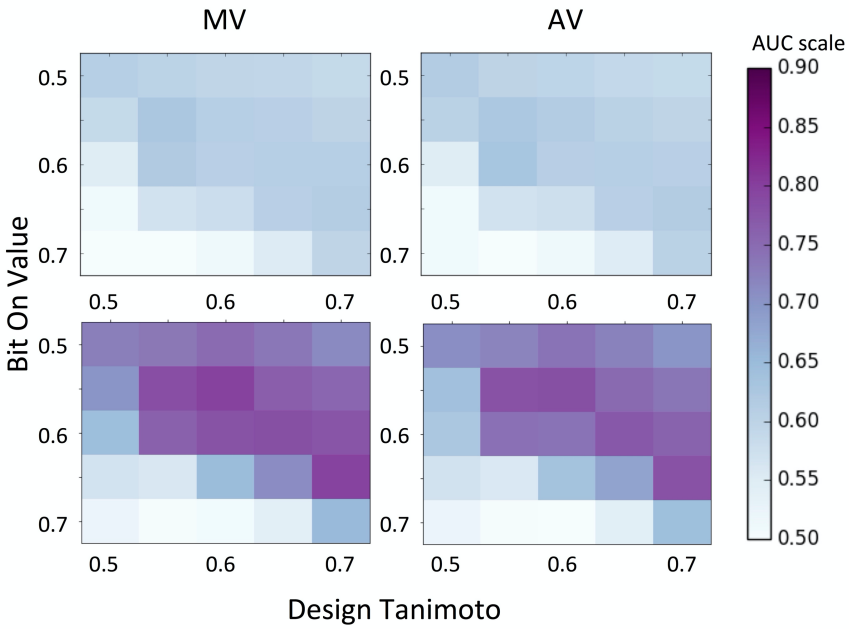


Figure S8. Heatmaps of AUC values for MV (Left) and AV (Right) methods for Test Set 1 (Top) and Test Set 2 (Bottom) when using SD10 with various DT and BOV when conformations are generated from SMILES strings.

Table S1. The comparison of the AUC values of both test sets when using conformations generated from SMILES and crystal structures for SD10 with DT=0.65 and BOV=0.60.

	CONFORMATIONS		CRYSTAL STRUCTURES
	AV	MV	
Test Set 1	0.61	0.61	0.64
Test Set 2	0.77	0.78	0.85

Table S2. The comparison of the AUC values of both test sets when using conformations generated from SMILES and crystal structures for SD10 with DT=0.65 and BOV=0.60 (100, 200 and 500 conformations) with standard errors in brackets.

	100 CONFORMATIONS		200 CONFORMATIONS		500 CONFORMATIONS	
	AV	MV	AV	MV	AV	MV
Test Set 1	0.61 (0.02)	0.61 (0.02)	0.60 (0.02)	0.60 (0.02)	0.61 (0.02)	0.60 (0.02)
Test Set 2	0.76 (0.02)	0.80 (0.02)	0.76 (0.02)	0.81 (0.02)	0.76 (0.02)	0.81 (0.02)

Generating more conformations for both test sets (set to maxima of 100, 200 and 500) using OMEGA[19] does not systematically improve the performance of the shape fingerprint method. As shown in Table S2, the AUC values are quite similar. With all of these larger sets of conformations, the improved discrimination provided by the MV method is clearer. Using larger sets of conformations increases the computational time needed to generate shape fingerprints and gives only a small improvement in accuracy.

The combination of shape and 2D fingerprints was also investigated. Logistic regression (using the combination of Test Set 1 and Test Set 2) linked values of FT (for shape fingerprints) and similarity Tanimoto (for 2D fingerprints) with the likelihood of shared activity (Figure S9). When two molecules are compared, the highest probability (either shape or 2D) was selected in each case. In this way, the calculated AUC value for Test Set 1 was 0.74 and for Test Set 2 was 0.94. These good AUC values arise mostly from the MACCS166 method but the two methods are complementary and when combined are able to make useful connections between molecules with shared chemical structures and those with shared shape and thus permit better clustering and scaffold hopping, as described in the main text.

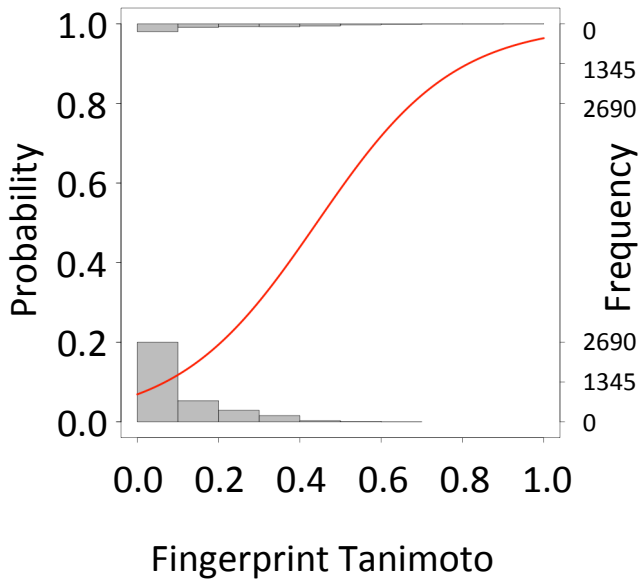


Figure S9. Logistic regression plot showing the probability of shared biological activity at differing FT values for combined Test Set 1 and Test Set 2 when shape fingerprints (using SD10 with DT=0.65 and BOV=0.60) are combined with MACCS166. A histogram of values for molecules that share biological activity is shown at the top of the diagram and that for molecules that do not share biological activity at the bottom.

The foregoing description of the shape fingerprints reveals that the method has a useful ability to link structures that share biological activity. They do this in a way that permits scaffold hopping because they are able to link molecules with significantly different chemical structures. However, for pairwise comparisons of the shape of molecules alternative methods are available that tend to perform better for this task. Analysis of the Directory of Useful Decoys (DUD-E) diverse set is summarized in the main text, in which the recall rates of actives from a pool of decoys achieves AUC values in the range 0.47 – 0.70 with an average of 0.57. While better than random, this suggests that shape fingerprints are an inferior method for virtual screening than a range of methods for assessing relative shape (they are able to achieve average AUCs in the range 0.63 – 0.79 on the same test sets).[11]

S2) Methods

The code for performing all of the calculations described is provided below.

a) Creating a database of reference shapes.

In this work, sets of reference shapes were generated by implementation of the algorithm previously described by Haigh et al.[1] The algorithm (which uses Openeye's Shape Toolkit)[12] randomly selects a first reference molecule out of the input dataset. The remaining molecules in the dataset are compared to the reference molecule (using the OEBestOverlay option in the Shape TK) and a Shape Tanimoto (ST) is calculated that can be defined as:

$$ST_{AB} = \frac{V_{AB}}{V_{AA} + V_{BB} - V_{AB}}$$

Where V_{AB} is the Gaussian overlap volume of the two molecules (A and B) aligned in such a way as to maximize the overlap. V_{AA} and V_{BB} are self-overlap volumes. Shape Tanimoto can vary from

0 (for the most dissimilar shaped molecules) to 1 (for molecules of identical shape). Molecules with ST greater than a user-selected value (the Design Tanimoto, DT) were discarded. The molecule with the smallest ST was then selected as the next reference shape and the same process repeated until all molecules have either been selected as a reference shape or discarded. Each set of reference shapes forms a Shape Database, referred to here as SDx where x is a distinguishing number.

b) Generating shape fingerprints**.** Shape fingerprints were generated by comparing a query molecule with each reference shape in the Shape Database in turn. For each reference shape, if the ST was above another user-defined value, the Bit-On Value (BOV), then the corresponding bit was set to 1 and if below the BOV, the bit was set to 0. Molecules were compared by aligning their bit strings, counting the number of bits set on (at 1) only in one of the strings and those set on in both strings. Bit Strings for molecular shapes A and B were compared using the Fingerprint Tanimoto (FT) as a similarity measure:

$$FT_{AB} = \frac{bothAB}{onlyA + onlyB - bothAB}$$

Where onlyA and onlyB are the numbers of unique bits On in the bit strings for A and B respectively, while bothAB is the number of bits On in common to A and B. Fingerprint Tanimoto similarity values vary from 0 (for dissimilar compounds) to 1 (for the most similar molecules). The script that was used in the shape fingerprint generation process is:

```
#!/usr/bin/env python3

from __future__ import print_function

from openeye.oechem import *
from openeye.oeshape import *


def comp(refmol, fitmol):

    """
    additional function in case the fitting molecule has bigger volume

    """
```

```
    best = OEBestOverlay()
    best.SetRefMol(refmol)

    scoreiter = OEBestOverlayScoreIter()
    OESortOverlayScores(scoreiter, best.Overlay(fitmol), OEHighestTanimoto())

    return scoreiter

def get_fingerprint(mol, Shape_database, bitOn):
    """
    Calculates Shape Tanimoto for molecule in reference to molecules stored in Shape Database,
    returns on bits for value higher than Design Tanimoto
    and off bit for value smaller
    :param: bitOn value (threshold value)
    :return: fingerprints for query molecule
    """

    keepsize = 1

    best = OEBestOverlay()
    best.SetRefMol(mol)


    V_ref = OECalcVolume(mol)

    fingerprint = ''

    for fitmol in Shape_database.GetOEMols():


        resCount = 0

        V_fit = OECalcVolume(fitmol)

        # molecules are all the time compared the same way
        # - molecule with bigger volume is always the reference

        if V_fit > V_ref:

            scoreiter = comp(fitmol, mol)

        else:

            scoreiter = OEBestOverlayScoreIter()
            OESortOverlayScores(scoreiter, best.Overlay(fitmol), OEHighestTanimoto())

        for score in scoreiter:
            outmol = OEGraphMol(fitmol.GetConf(OEHasConfIdx(score.fitconfidx)))
            score.Transform(outmol)


            if float(score.tanimoto) > bitOn:
                fingerprint += ' 1'
            else:
```

```
            fingerprint += ' 0'

        resCount += 1
        if resCount == keepsize:
            break

    return fingerprint

def main(argv=[__name__]):

    if len(argv) != 5:
        OEThrow.Usage("%s <data_file.sdf> <shape_file.sdf> <out_file.sdf> <bitOn> " %
argv[0])

    data_file = oemolistream(argv[1])
    shape_file = OEMolDatabase(argv[2])
    out_file = oemolostream(argv[3])
    bitOn = float(argv[4])


    for mol in data_file.GetOEMols():
        fp = get_fingerprint(mol, shape_file, bitOn)
        OESetSDData(mol, "Fingerprint", "%s" % fp)
        OEWriteMolecule(out_file, mol)


if __name__ == "__main__":
    import sys
    sys.exit(main(sys.argv))
```

c) Analysis. To evaluate the shape fingerprints, two test sets were employed: 1) a set described by Taylor et al.[6] which was devised to test pharmacophore models and consists of 87 molecules binding to 10 different proteins 2) a group from the Astex diversity set,[7] which includes 45 molecules binding to 4 selected proteins. Both test sets are detailed in Tables S3 and S4 (Section S3). In order to analyze the results, the ROC curve was used, which is a tool for diagnostic test evaluation.[8] The ROC curves and AUC values were produced in R.[13] Half of the matrix without the diagonal was used in these calculations.

d) Conformations. SMILES were generated using Openeye's OEChem Toolkit[12] for all the molecules shown in Tables S3 and S4 (section S3). Some of the generated SMILES needed

manual assignment of stereochemistry. Conformations were generated using OMEGA.[9, 14] Shape fingerprints were generated for each conformation. When two molecules were compared, all fingerprints of one molecule were compared with all those of the other. Two summary values for this comparison were investigated: 1) the highest value of FT amongst the array arising from comparisons of all conformations of one molecule (in turn) with all conformations of the other (in turn) is selected or 2) the average of those values is selected.

e) 2D Fingerprints**.** These were generated using the implementation of the MACCS166, circular fingerprints, path fingerprints and tree fingerprints available in the Openeye OEChem Toolkit.[12] The script used to generate these is:

```
#!/usr/bin/env python

from openeye.oechem import *
from openeye.oegraphsim import *

def main(argv=[__name__]):

    if len(argv) != 3:
        OEThrow.Usage("%s <infile> <outfile>" % argv[0])

    data = OEMolDatabase(argv[1])
    out_file = oemolostream(argv[2])

    fp = OEFingerPrint()

    for mol in data.GetOEGraphMols():


        OEMakeFP(fp, mol, OEFPType_Lingo)

        # OEMakeFP(fp, mol, OEFPType_MACCS166)
        # OEMakeFP(fp, mol, OEFPType_Circular)
        # OEMakeFP(fp, mol, OEFPType_Path)
        # OEMakeFP(fp, mol, OEFPType_Tree)

        fptypestr = fp.GetFPTypeBase().GetFPTypeString()
        fphexdata = fp.ToHexString()
        OESetSDData(mol, fptypestr, fphexdata)

        OEWriteMolecule(out_file, mol)
```

```
    out_file.close()

if __name__ == "__main__":
    import sys
    sys.exit(main(sys.argv))
```

f) Code for creating SOMs

In the first set of SOMs (those shown in the main text and the matching set for the USR method, below), the SOM was created using the SOMbrero GUI. The full set of DUDE SOMs were created with the following code in R:

```
library(SOMbrero)
mydata<-read.table("akt1_USR_proc.txt",sep=" ", header = FALSE, fill =
TRUE)
som1<-trainSOM(x.data=mydata[,3:14], dimension=c(25,25),
topo=c("square"), radius.type=c("letremy"), dist.type = c("letremy"),
type =c("numeric"), init.proto = c("random"),scaling =
c("unitvar"),eps0 = 1)
write.csv(x=som1$clustering,file="akt1_SOM1.txt")
```

Following re-processing of the output such that the cluster number ends up in column 16 and the activity flag (ACTIVE or DECOY) in column 15, the analysis of enrichment was then performed with the following:

```
mydata<-read.table("akt1_SOM_OUTPUT.txt",sep=" ", header = FALSE, fill
= TRUE)
summtable<-table(mydata$V16,mydata$V15)
write.table(summtable,file="akt1_SUMMARY.txt")
A=matrix(nrow=nrow(summtable),ncol=6)
A[,1]<-summtable[,1]
A[,2]<-summtable[,2]
for(i in 1:nrow(A)){
A[i,3]<-
prop.test(c(A[i,1],sum(A[,1])),c(A[i,1]+A[i,2],sum(A[,1])+sum(A[,2])),a
lternative="greater")$p.value
}
A[,4]<-cut(A[,3],breaks=c(0,0.000016))
for(i in 1:nrow(A)) {
A[i,5]<-if (A[i,3]<0.000016) {
A[i,1]} else {
0}
}
for(i in 1:nrow(A)) {
A[i,6]<-if (A[i,3]<0.000016) {
A[i,2]} else {
0}
}
write.table(A,"akt1_pvalues.txt")
```

```
enrichment<-
sum(A[,5])*(sum(A[,1])+sum(A[,2]))/(sum(A[,1])*(sum(A[,5])+sum(A[,6])))
recall<-sum(A[,5])/sum(A[,1])
write.table(enrichment,"akt1_enrichment.txt")
write.table(recall,"akt1_recall.txt")
```

The maps are the created with the following sequence

1) In R

```
mydata<-read.table("akt1_SOM_X_Y.txt",sep=" ", header = TRUE, fill =
TRUE)
summtable<-table(mydata$X_Y,mydata$ACTIVITY)
write.table(summtable,"akt1_xytable.txt")
```

2) The file that was created is reprocessed with the following command line:

```
echo "X Y ACTIVE DECOY" > $1_x_y_summ.txt
cat akt1_xytable.txt | grep -v "ACTIVE" | sed -e 's/"//g;s/_/ /g' >>
akt1_x_y_summ.txt
```

3) In R

```
library(ggplot2)
mydata<-read.table("akt1_x_y_summ.txt",sep=" ", header = TRUE,
fill = TRUE)
A=matrix(nrow=nrow(mydata),ncol=7)
A[,1]<-mydata$ACTIVE
A[,2]<-mydata$DECOY
A[,5]<-mydata$X
A[,6]<-mydata$Y
for(i in 1:nrow(A)){
A[i,3]<-
prop.test(c(A[i,1],sum(A[,1])),c(A[i,1]+A[i,2],sum(A[,1])+sum(A[,
2])),alternative="greater")$p.value
}
A[,4]<-cut(A[,3],breaks=c(0,0.000016,1))
B<-as.data.frame(A)
B$V7<-factor(B$V4,levels=c(1,2), labels=c("*",""))
ggplot(mydata,aes(B$V5,B$V6,fill=B$V1,label=B$V7))+geom_tile(colo
r="white")+scale_fill_gradient2(limit=c(0,max(A[,1])))+theme(pane
l.background=element_blank())+labs(x="X",y="Y",fill="N
active")+guides(color=FALSE)+geom_text(color="red")
ggsave("akt1_CLUSTERS.png")
```

g) USR encoding

The following example of python code creates the USR encoding via the RDkit toolkit:

```
from rdkit import Chem
from rdkit.Chem import rdMolDescriptors
mol=Chem.MolFromMolFile('AKT_1.sdf')
USR=rdMolDescriptors.GetUSR(mol)
print("CHEMBL1081852",USR,"ACTIVE")
```

S3) Test Sets – list of PDBs corresponding to each target in the two test sets.

Table S3. Test Set 1 - Test set described by R. Taylor et al. used for validation of the shape fingerprints method.[6]

Protein	Number of complexes	PDB codes
Protein kinase 5 (PK5)	2	1v0o, 1v0p
Fatty acid binding protein (FABP)	3	1tou, 1tow, 2hnx
Neprilysin (NEP)	4	1dmt, 1r1h, 1r1j, 1y8j
Dihydrofolate reductase (DHFR)	6	1drf, 1hfr, 1mvt, 1pd9, 1s3v, 2dhf
Checkpoint kinase (Chk1)	16	1nvq, 1nvr, 1nvs, 1zlt, 1zys, 2br1, 2brb, 2brg, 2brh, 2brm, 2bro, 2c3l, 2cgu, 2cgw, 2cgx, 2hog
Neuraminidase (NEU)	11	1a4g, 1a4q, 1b9s, 1b9t, 1b9v, 1inf, 1inv, 1ivb, 1nsc, 1nsd, 1vcj
Carbonic anhydrase (CA)	13	1bn3, 1bn4, 1bnq, 1cim, 1eou, 1if7, 1oq5, 1xpz, 1zgf, 1zh9, 2eu3, 2hoc, 2nng
Adenosine deaminase (ADA)	11	1krm, 1ndv, 1ndw, 1ndy, 1o5r, 1qxl, 1uml, 1v7a, 1v79, 1wxy, 2e1w
Heat shock protein 90 (HSP)	10	1byq, 1uy8, 1yc1, 1yc4, 1yet, 2bsm, 2byi, 2bz5, 2cct, 2uwd
Acetylcholinesterase (AChE)	11	1dx6, 1e66, 1eve, 1gpk, 1gpn, 1h23, 1w4l, 1zgb, 2ack, 2c5g, 2ckm

Table S4. Test Set 2 - Test set made from selected targets in the Astex diversity set used for validation of the shape fingerprints method.[7]

Protein	Number of complexes	PDB codes
Chitinase B	8	1w1p, 1w1t, 1w1v, 1w1y, 3wd1, 3wd2, 3wd3, 3wd4
TMK	8	1mrs, 1w2g, 1w2h, 4unn, 4unp, 4unq, 4unr, 4uns
Tryptophan Syntase	13	1k3u, 1k7e, 1k7f, 1qop, 1yjp, 1wbj, 2cle, 2clh, 2clk, 2j9y, 4hpx, 4ht3, 4kkx
VDR	16	1db1, 1ie8, 1ie9, 1s0z, 1s19, 1txi, 2ham, 3auq, 3aur, 3ax8, 3kpz, 3vhw, 3x31, 3x36, 4ite, 5gt4

S4) AUC values for Test sets using different Shape Databases.

In this section the variation of the AUC in the ROC curve analysis when different shape databases are used with different settings is explored.

In the first section (a), we present values for an analysis that uses only shapes taken from the crystal structure of protein-ligand complexes and present firstly results for our preferred shape database (SD10). Subsequently, the results obtained with the largest shape database (SD01) are presented. Finally results from a series of versions of SD10 obtained by starting from a different random starting molecule are given (and the MW distribution of each of the resampled shape databases is shown graphically). In the second section (b), analyses based on sets of conformations generated from the two-dimensional molecular structure, using OMEGA, are given.[9, 14]

a) Analyses based on crystal structures

Table S5. The AUC values for Test Set 1 when using SD10 with 95% confidence intervals in brackets. Those settings that provide AUC values that are not statistically distinct from those with our selected settings are highlighted in pink.

		DT				
		0.50	**0.55**	**0.60**	**0.65**	**0.70**
BOV	**0.50**	0.6104 (0.5803 - 0.6405)	0.6092 (0.5787 - 0.6396)	0.6163 (0.5860 - 0.64663)	0.6091 (0.5784 - 0.6399)	0.6082 (0.5775 - 0.6389)
	0.55	0.6354 (0.5993 - 0.6716)	0.6491 (0.6139 - 0.6843)	0.6438 (0.6139 - 0.6737)	0.6272 (0.5969 - 0.6575)	0.6227 (0.5925 - 0.6530)
	0.60	0.5268 (0.4769- 0.5766)	0.6403 (0.5951- 0.6855)	0.6427 (0.6030 -0.6823)	0.636 (0.6035- 0.6685)	0.6389 (0.6077 - 0.6701)
	0.65	0.5264 (0.4747 - 0.5780)	0.5734 (0.5211 - 0.6257)	0.5904 (0.5392 - 0.6417)	0.6286 (0.5835 - 0.6737)	0.6599 (0.6228 - 0.6970)
	0.70	0.5002 (0.4497816 - 0.5499161)	0.5197 (0.4683 - 0.5712)	0.5052 (0.4548 - 0.5555)	0.548 (0.4954 -0.6005)	0.6298 (0.5785 - 0.6811)

Table S6. The AUC values for Test Set 2 when using SD10 with 95% confidence intervals in brackets. Those settings that provide AUC values that are not statistically distinct from those with our selected settings are highlighted in pink.

		DT				
		0.50	**0.55**	**0.60**	**0.65**	**0.70**
BOV	**0.50**	0.7649 (0.7246 - 0.8052)	0.7944 (0.7554 - 0.8334)	0.7903 (0.7525 - 0.8281)	0.7694 (0.7304 - 0.8085)	0.758 (0.7191 - 0.7970)
	0.55	0.814 (0.7668 - 0.8612)	0.7899 (0.7473 - 0.8326)	0.8146 (0.7744 - 0.8548)	0.8139 (0.77712 - 0.8506)	0.7995 (0.7632 - 0.8358)
	0.60	0.781 (0.7195 - 0.8426)	0.7181 (0.6507 - 0.7856)	0.8372 (0.7871 - 0.8872)	0.8519 (0.8122 - 0.8916)	0.8365 (0.8021 - 0.8710)

	0.65	0.7613 (0.6937 - 0.8289)	0.517 (0.4451 - 0.589)	0.7953 (0.7299 - 0.8607)	0.8622 (0.8114 - 0.9130)	0.8593 (0.8152 - 0.9034)
	0.70	0.6555 (0.5803 - 0.7307)	0.5 (0.4287 - 0.5713)	0.654 (0.5795 - 0.7285)	0.7628 (0.6934 - 0.8321)	0.8228 (0.7614 - 0.8843)

Table S7. The AUC values for Test Set 1 when using SD01.

		DT				
BOV		**0.50**	**0.55**	**0.60**	**0.65**	**0.70**
	0.50	0.6441	0.62	0.621	0.6172	0.6138
	0.55	0.5456	0.6473	0.6375	0.6304	0.6266
	0.60	0.5646	0.6176	0.6572	0.6556	0.6429
	0.65	0.5426	0.5514	0.6086	0.6431	0.6383
	0.70	0.5062	0.5062	0.541	0.5603	0.621

Table S8. The AUC values for Test Set 2 when using SD01.

		DT				
BOV		**0.50**	**0.55**	**0.60**	**0.65**	**0.70**
	0.50	0.8449	0.8475	0.8039	0.7911	0.7739
	0.55	0.6533	0.8026	0.826	0.8271	0.8151
	0.60	0.531	0.8212	0.8161	0.8631	0.8511
	0.65	0.5325	0.6721	0.7892	0.8431	0.8636
	0.70	0.5101	0.5103	0.7557	0.7495	0.7981

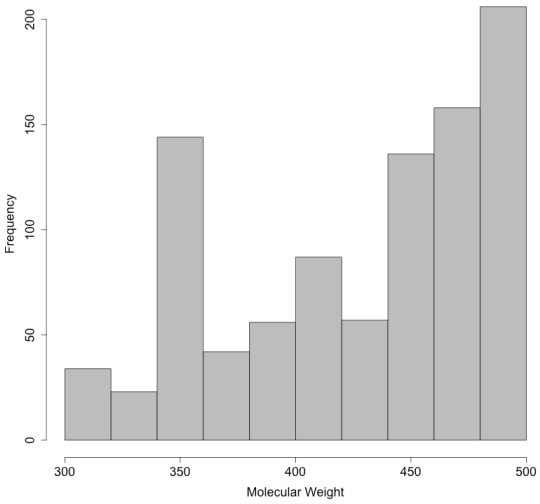
Table S9. The AUC values for 10x resampled Shape Database 10 with DT=0.65 for Test Set 1.

		Iteration									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BOV	**0.5**	0.6087	0.609	0.6104	0.6123	0.6122	0.6053	0.6098	0.6109	0.6076	0.6095
	0.55	0.6264	0.631	0.6299	0.6249	0.6272	0.6284	0.6286	0.6327	0.6215	0.6344
	0.6	0.6523	0.648	0.6366	0.6336	0.6324	0.6387	0.6455	0.638	0.6458	0.6425
	0.65	0.6504	0.6429	0.6414	0.6335	0.6321	0.6214	0.6575	0.6214	0.659	0.6185
	0.7	0.5981	0.5515	0.5478	0.5655	0.5956	0.5935	0.5701	0.6015	0.5727	0.5658

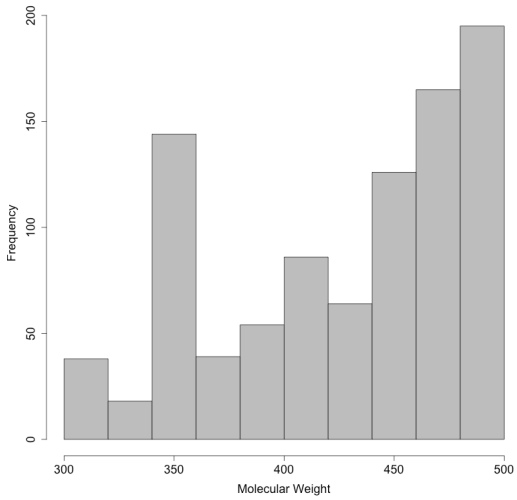
Table S10. The AUC values for 10x resampled Shape Database 10 with DT=0.65 for Test Set 2.

		Iteration									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BOV	**0.5**	0.7784	0.7684	0.7736	0.7675	0.7715	0.7766	0.7759	0.7725	0.7728	0.7799
	0.55	0.8101	0.8156	0.8109	0.7973	0.8162	0.8122	0.8165	0.8075	0.8052	0.8104
	0.6	0.8623	0.8556	0.8558	0.8336	0.8287	0.8638	0.8431	0.8295	0.8214	0.8473
	0.65	0.8516	0.8412	0.8142	0.8422	0.8258	0.8406	0.8341	0.8471	0.8186	0.8306
	0.7	0.7642	0.6685	0.7592	0.7458	0.7471	0.7448	0.7279	0.7373	0.68	0.7613

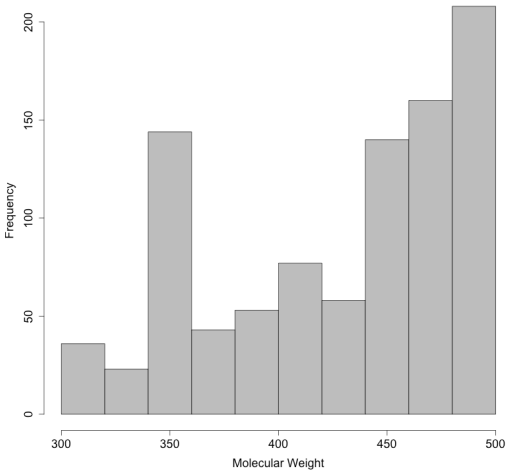
A)


Frequency
0
50
100
150
200
300
350
400
450
500
Molecular Weight

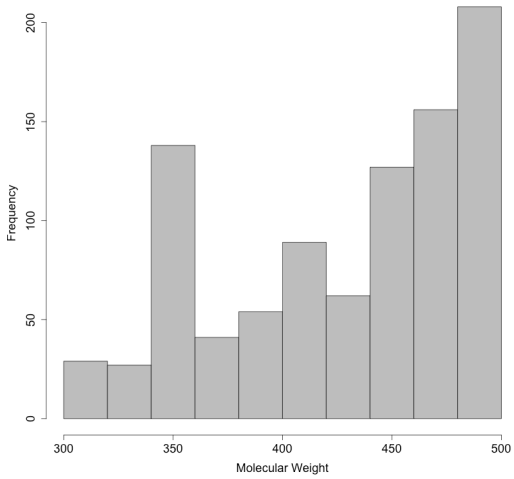
B)


Frequency
0
50
100
150
200
300
350
400
450
500
Molecular Weight

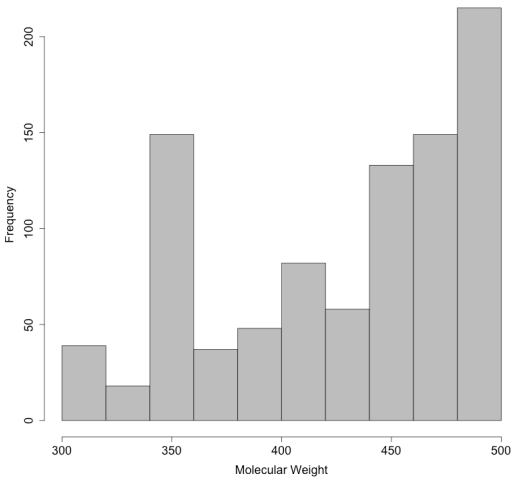
C)


Frequency
0
50
100
150
200
300
350
400
450
500
Molecular Weight

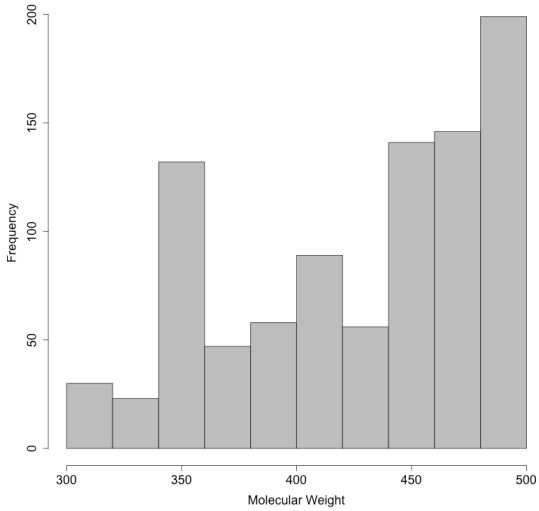
D)


Frequency
0
50
100
150
200
300
350
400
450
500
Molecular Weight

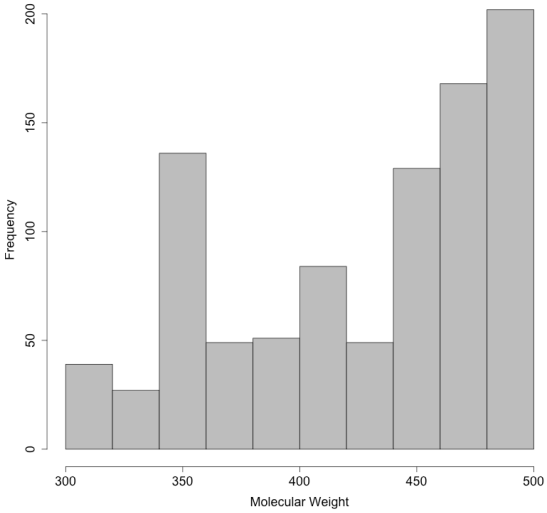
E)


Frequency
0
50
100
150
200
300
350
400
450
500
Molecular Weight

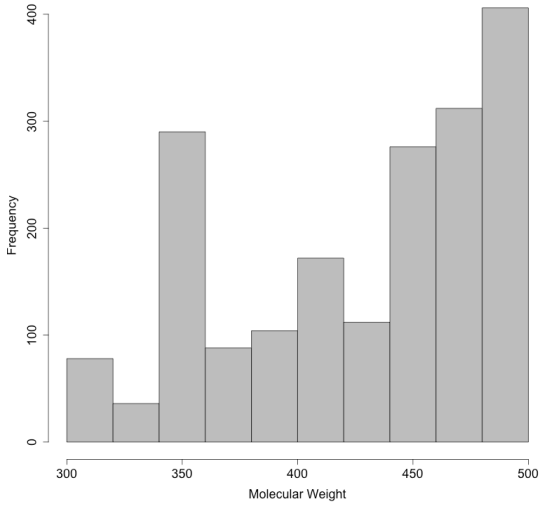
F)


Frequency
0
50
100
150
200
300
350
400
450
500
Molecular Weight

G)


Frequency
0
50
100
150
200
300
350
400
450
500
Molecular Weight

H)


Frequency
0
100
200
300
400
300
350
400
450
500
Molecular Weight

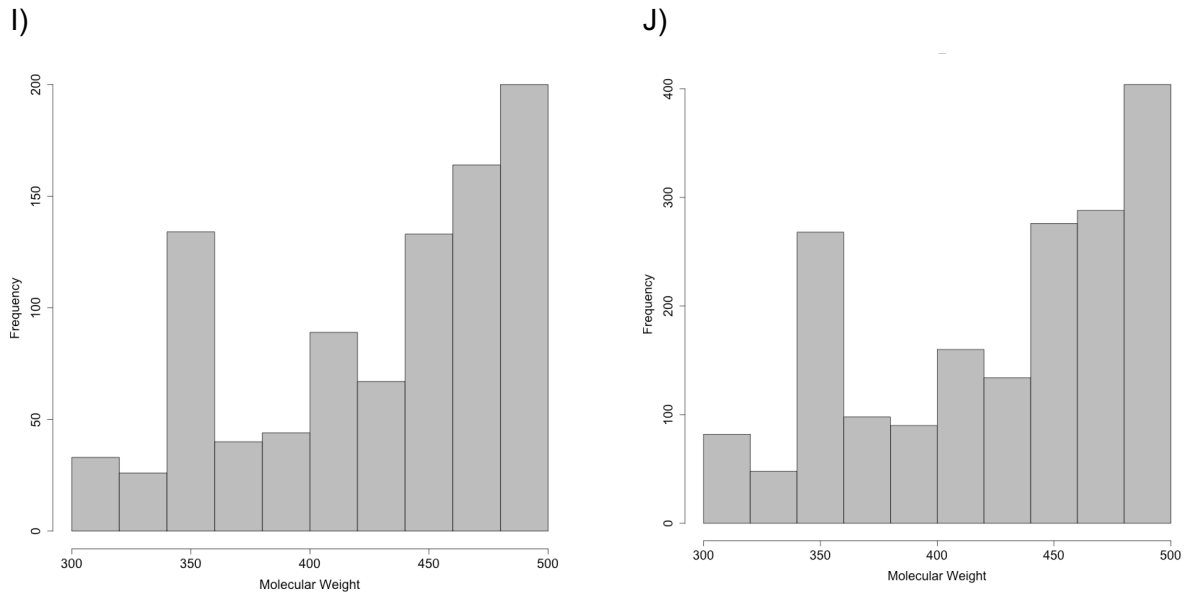


Figure S10. Molecular weight distribution for 10x resampled SD10 with DT=0.65 and BOV=0.6.

b) Analyses based on conformations generated computationally

AUC values for Test sets using Shape Database 10 when applying both methods – MV and AV.

Table S11. The AUC values for Test Set 1 when using SD10 (AV method) with 95% confidence intervals in brackets.

		DT				
		0.50	**0.55**	**0.60**	**0.65**	**0.70**
BOV	**0.50**	0.6139 (0.5848- 0.6442)	0.6003 (0.5708- 0.6307)	0.5969 (0.5670 -0.6276)	0.5903 (0.5600- 0.6211)	0.587 (0.5570- 0.6174)
	0.55	0.6049 (0.5685- 0.6401)	0.624 (0.5936- 0.6560)	0.6143 (0.5840- 0.6452)	0.6054 (0.5754- 0.6370)	0.5992 (0.5687- 0.6305)
	0.60	0.5496 (0.5011- 0.5985)	0.6308 (0.5918 -0.6696)	0.6099 (0.5763 -0.6447)	0.6117 (0.5801 -0.6444)	0.6081 (0.5762 -0.6411)
	0.65	0.5111 (0.4602 -0.5613)	0.5729 (0.5207- 0.6220)	0.5782 (0.5306- 0.6241)	0.6094 (0.5686- 0.6493)	0.6161 (0.5790 -0.6541)

	0.70	0.501 (0.4508- 0.5512)	0.4996 (0.4508- 0.5509)	0.5124 (0.4617 -0.5632)	0.5528 (0.5016- 0.6041)	0.6057 (0.5571 -0.6524)

Table S12. The AUC values for Test Set 1 when using SD10 (MV method) with 95% confidence intervals in brackets.

		DT				
		0.50	**0.55**	**0.60**	**0.65**	**0.70**
BOV	**0.50**	0.6119 (0.5824- 0.6414)	0.6027 (0.5726 -0.6328)	0.5961 (0.5657- 0.6265)	0.5943 (0.5637- 0.6249)	0.5882 (0.5580 -0.6185)
	0.55	0.5878 (0.5534- 0.6221)	0.626 (0.5948- 0.6573)	0.6111 (0.5805 -0.6417)	0.6091 (0.5783- 0.6399)	0.6005 (0.5696- 0.6314)
	0.60	0.5488 (0.5011 -0.5965)	0.6189 (0.5812 -0.6566)	0.6089 (0.5754- 0.6424)	0.612 (0.5807- 0.643)	0.6116 (0.5803- 0.6429)
	0.65	0.5106 (0.4601- 0.5611)	0.5711 (0.5208- 0.6213)	0.5804 (0.5338 -0.6270)	0.6085 (0.5694- 0.6476)	0.6125 (0.5764- 0.6485)
	0.70	0.501 (0.4508- 0.5512)	0.5008 (0.4508 -0.5509)	0.5123 (0.4616 -0.5631)	0.5534 (0.5022 -0.6047)	0.5999 (0.5529 -0.6469)

Table S13. The AUC values for Test Set 2 when using SD10 (AV method) with 95% confidence intervals in brackets.

		DT				
		0.50	**0.55**	**0.60**	**0.65**	**0.70**
BOV	**0.50**	0.7099 (0.6714- 0.7484)	0.7216 (0.6838- 0.7595)	0.7407 (0.7018- 0.7796)	0.7228 (0.6833- 0.7623)	0.7028 (0.6633 -0.7423)
	0.55	0.6412 (0.5992- 0.6833)	0.7808 (0.7460- 0.8155)	0.7834 (0.74825- 0.8185)	0.7525 (0.7149- 0.7901)	0.739 (0.7009- 0.7771)
	0.60	0.6268 (0.5705- 0.6832)	0.7437 (0.6844- 0.8029)	0.7418 (0.6983- 0.7853)	0.7734 (0.7354 -0.8114)	0.7601 (0.7208 - 0.7993)

	0.65	0.5707 (0.5005- 0.6409)	0.5581 (0.4866- 0.6296)	0.6384 (0.5735 -0.7032)	0.6815 (0.6247- 0.7383)	0.7811 (0.7362 - 0.8260)
	0.70	0.5236 (0.4510- 0.5962)	0.5044 (0.4332 -0.57563)	0.493 (0.4368- 0.5773)	0.542 (0.4714- 0.6127)	0.6448 (0.5780 - 0.7117)

Table S14. The AUC values for Test Set 2 when using SD10 (MV method) with 95% confidence intervals in brackets.

	DT					
BOV		**0.50**	**0.55**	**0.60**	**0.65**	**0.70**
	0.50	0.728 (0.6882- 0.7678)	0.7346 (0.6964 -0.7728)	0.7515 (0.7125 -0.7904)	0.7372 (0.6976 -0.7768)	0.7152 (0.6751- 0.7553)
	0.55	0.7023 (0.6587- 0.7458)	0.7857 (0.7506- 0.8209)	0.7993 (0.7650 -0.8336)	0.7666 (0.7291- 0.8042)	0.7556 (0.7171 -0.7941)
	0.60	0.6458 (0.5885- 0.7032)	0.7616 (0.7012 -0.8219)	0.7785 (0.7352 -0.8218)	0.784 (0.7469 -0.8211)	0.7753 (0.73791- 0.8127)
	0.65	0.5685 (0.4991 -0.6379)	0.5598 (0.4881- 0.6314)	0.6493 (0.5835- 0.7151)	0.7116 (0.6551- 0.7681)	0.7981 (0.7550- 0.8412)
	0.70	0.5236 (0.4509- 0.5962)	0.5044 (0.4332- 0.5756)	0.5072 (0.4370- 0.5774)	0.5425 (0.4718 -0.6131)	0.6538 (0.5859- 0.7217)

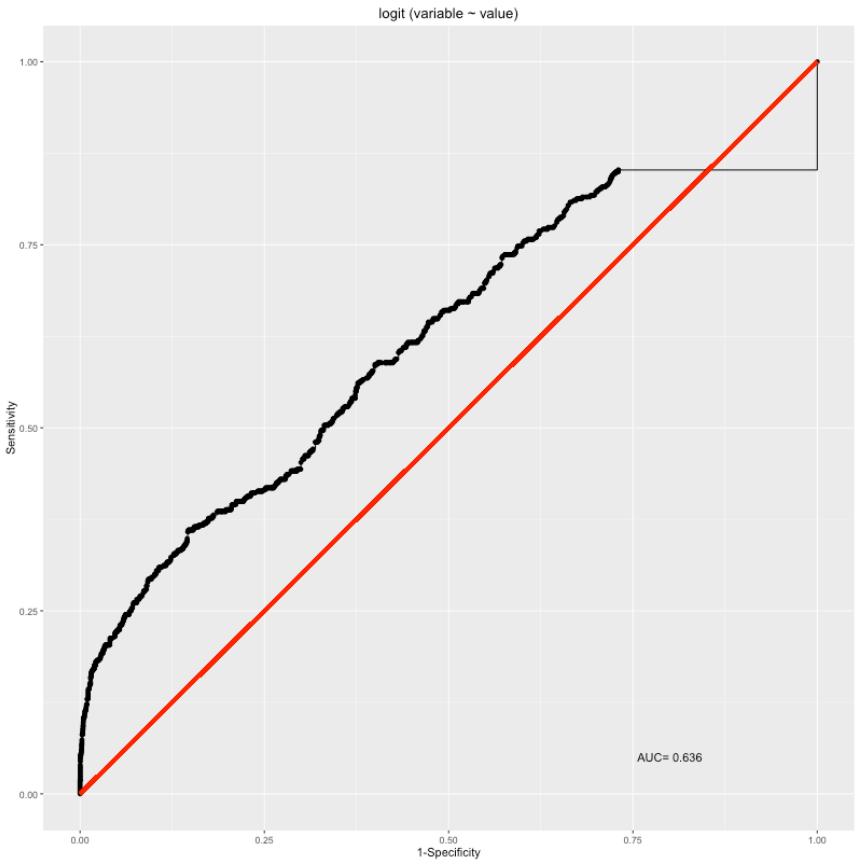


Figure S11. ROC curve for Test Set 1 when using SD10 with DT=0.65 and BOV=0.60.

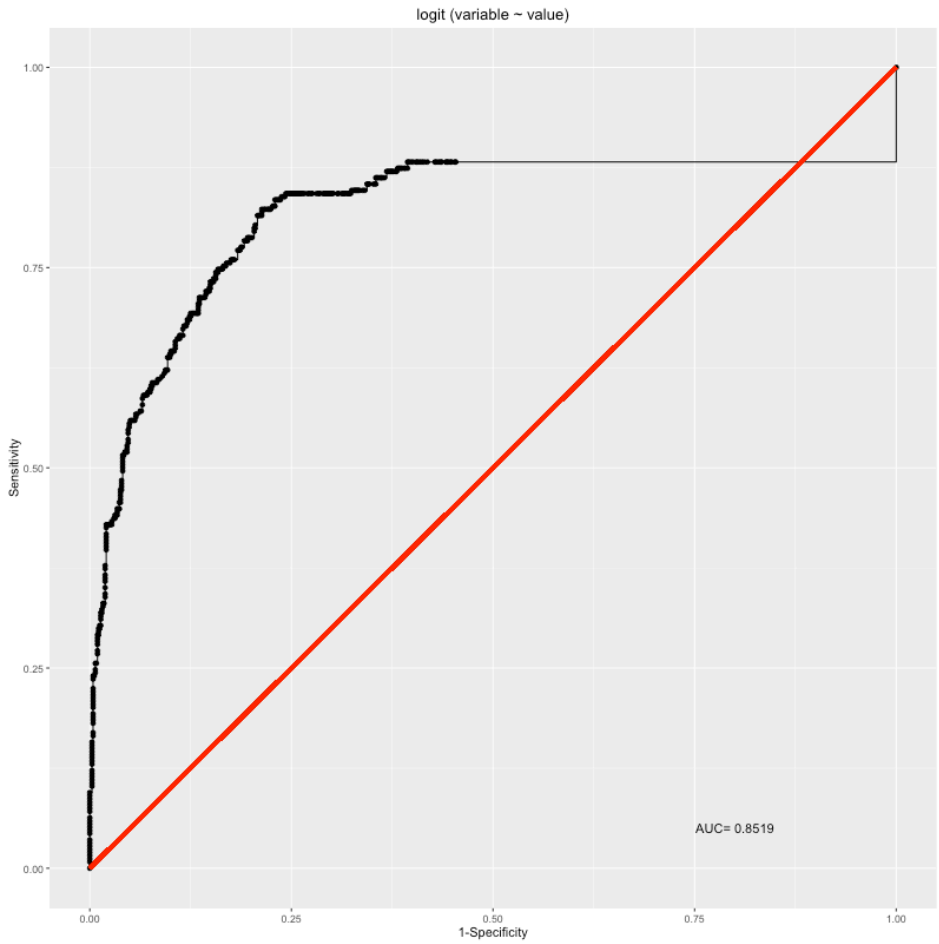


Figure S12. ROC curve for Test Set 2 when using SD10 with DT=0.65 and BOV=0.60.

100, 200 and 500 CONFORMATIONS – Logistic regression plots for Test Set 1 and Test Set 2 using AV and MV methods.

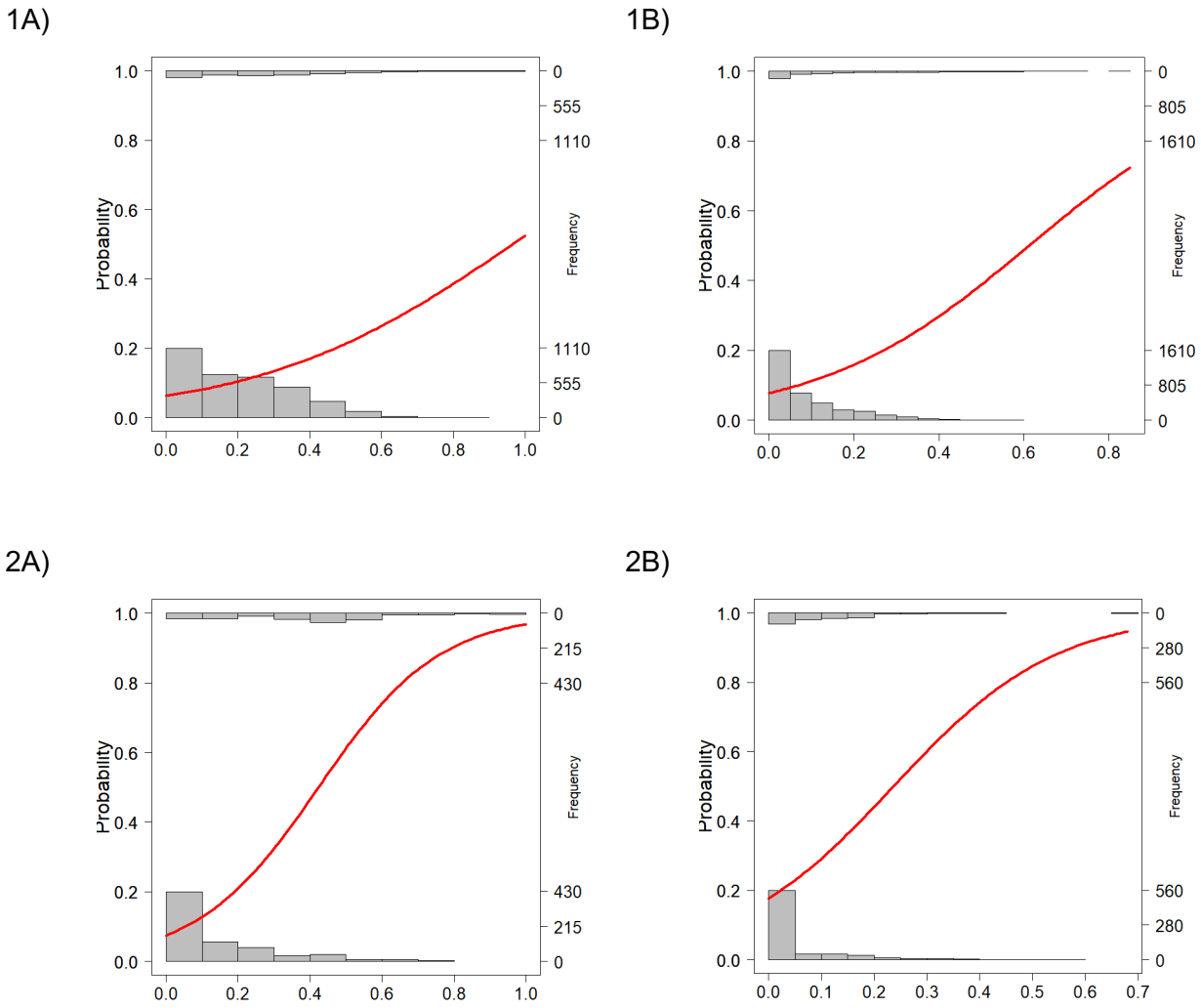


Figure S13. Logistic regression plot for Test Set 1 (1) and Test Set 2 (2) when using SD10 with DT=0.65 and BOV=0.60 with MV (A) and AV (B) method with maximum 100 conformations.

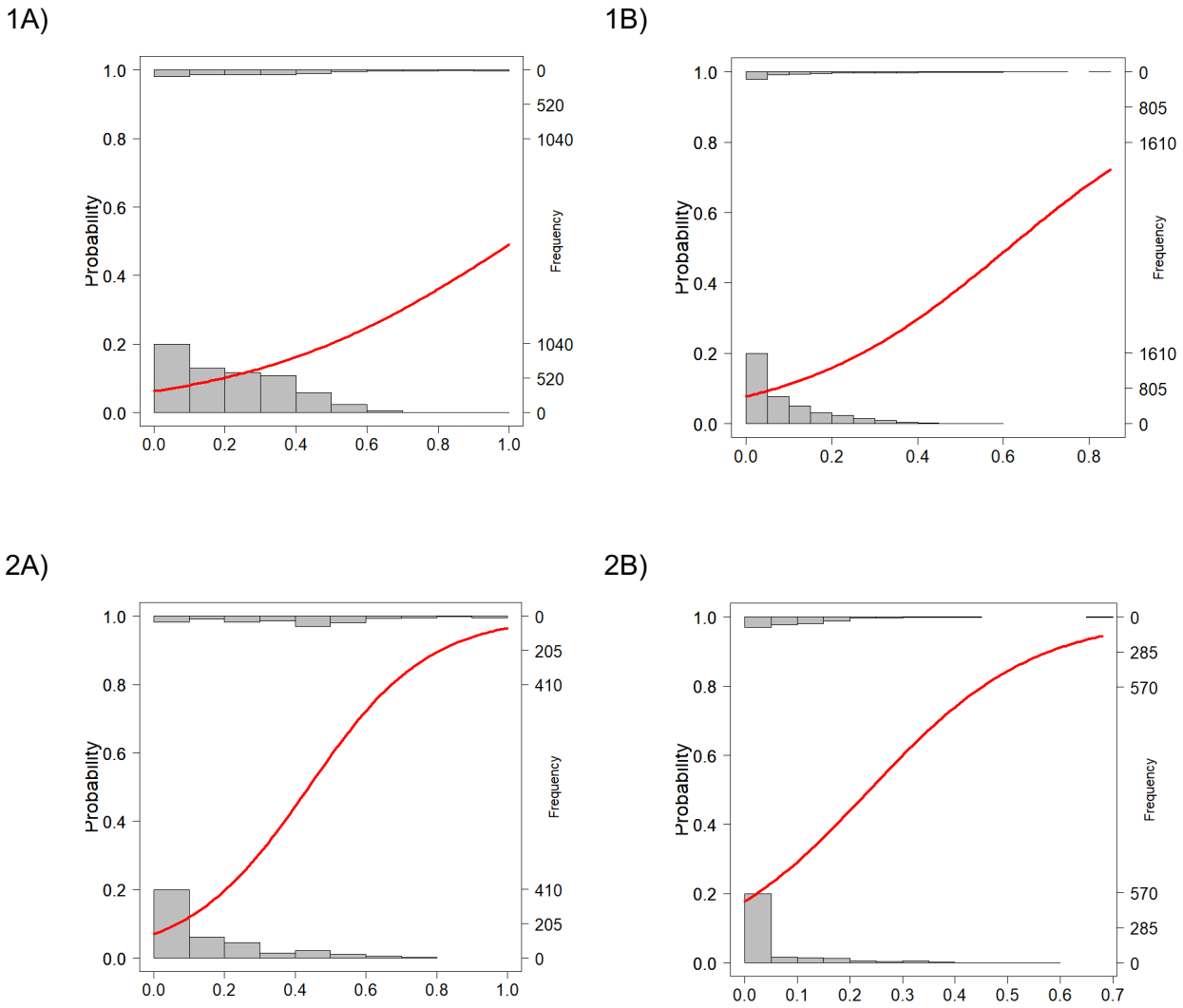


Figure S14. Logistic regression plot for Test Set 1 (1) and Test Set 2 (2) when using SD10 with DT=0.65 and BOV=0.60 with MV (A) and AV (B) method with maximum 200 conformations.

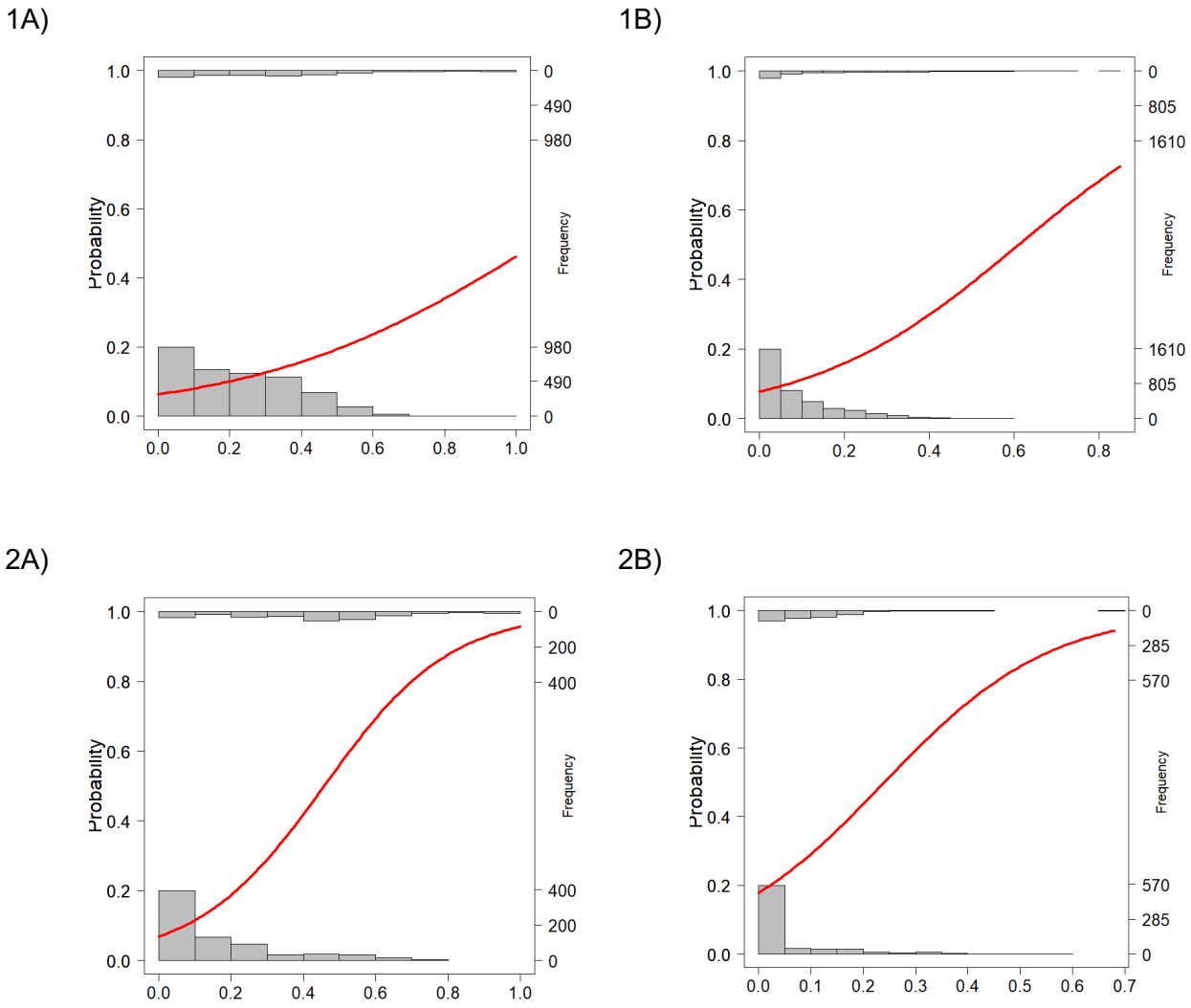


Figure S15. Logistic regression plot for Test Set 1 (1) and Test Set 2 (2) when using SD10 with DT=0.65 and BOV=0.60 with MV (A) and AV (B) method with maximum 500 conformations.

S5) Analysis of the DUD-E diverse set with the optimized shape fingerprints.

Although the two test sets were selected because they are based on high quality crystal structures and have been published as tools for validation, they are quite small in size. In order to explore performance with a larger test set, the DUD database was explored.[15] Here, the Diverse set from the DUD-E database was used in several alternative ways:

1) Grouping Actives. The collection of decoys was not included and only actives were taken into consideration. The DUD-E diverse set consists of 8 targets: serine/threonine-protein kinase AKT

(AKT1), a beta-lactamase (AMPC), cytochrome P450 3A4 (CP3A4), C-X-C chemokine receptor type 4 (CXCR4), the glucocorticoid receptor (GCR), human immunodeficiency virus type 1 protease (HIVPR), human immunodeficiency virus type 1 reverse transcriptase (HIVTR) and kinesin-like protein 1 (KIF11). It consists of 290, 48, 166, 39, 258, 527, 330 and 116 actives (which is 1774 ligands in total) for AKT1, AMPC, CP3A4, CXCR4, GCR, HIVPR, HIVRT and KIF11, respectively. Sets of up to five conformations were generated for each compound. The shape fingerprints were generated for each structure using Shape Database SD10 with DT = 0.65. A bit on value equal to 0.60 was applied, exactly as suggested above. The AUC value calculated for correctly grouping the actives in the DUD-E diversity set is equal to 0.5304 (with 95% confidence interval 0.5289-0.5320) and 0.5674 (with 95% confidence interval 0.5660-0.5689) for the AV and MV methods, respectively. This is a lower value than the ones obtained for Test Set 1 and Test Set 2 (0.64 and 0.85 for Test Set 1 and Test Set 2, respectively) but does represent real enrichment. The reduction in performance likely reflects the inclusion of high molecular weight molecules, particularly in the HIVPR and CP3A4 sets (see below). The probability curve from the logistic regression plot, in Figure S16, reaches the value of 0.8 for both AV and MV methods and it reveals that for FT above 0.5 there is greater than 50% chance of shared biological activity.

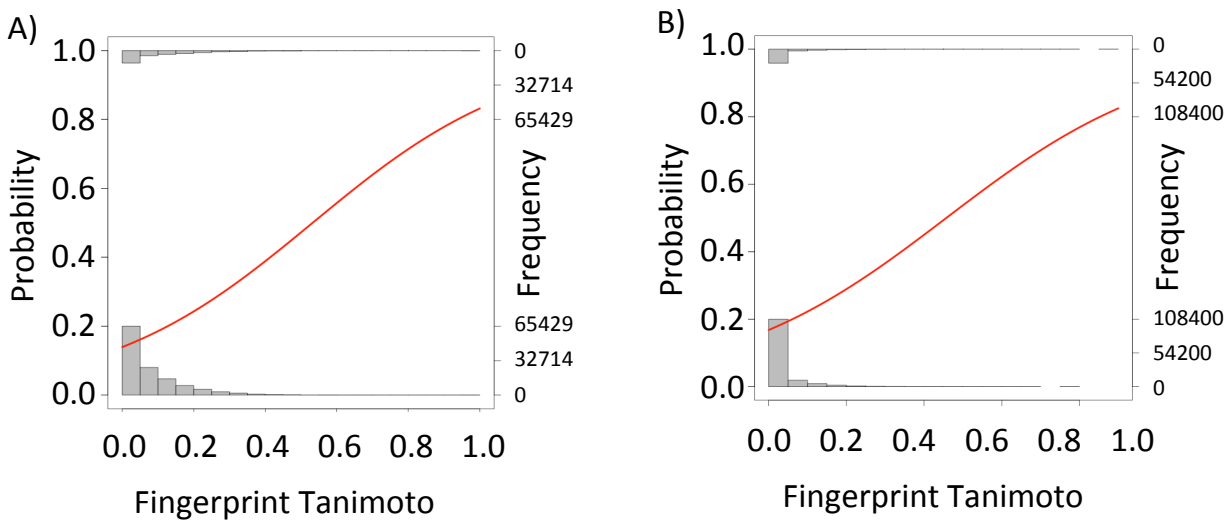
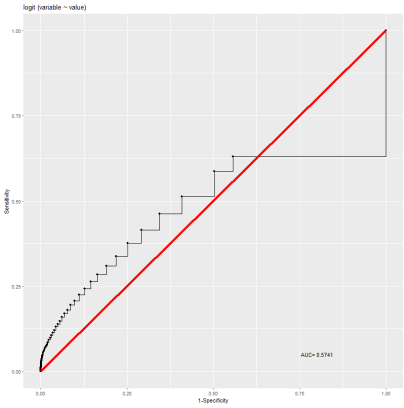
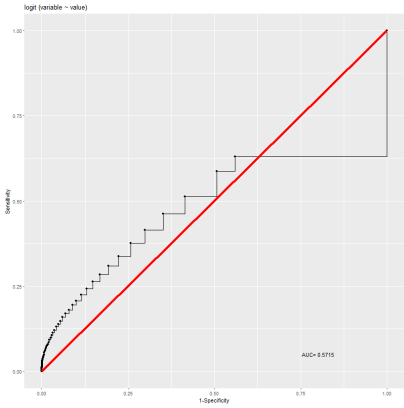
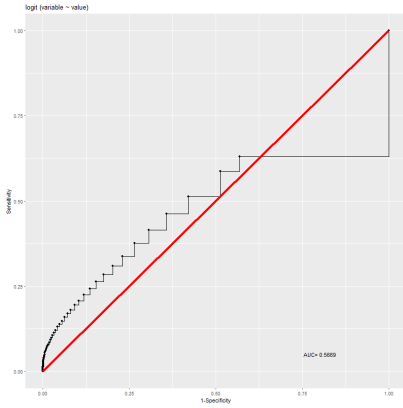


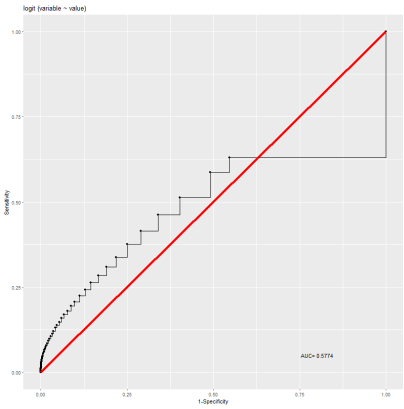
Figure S16. Logistic regression plot for the DUD-E diverse set when using SD10 with DT=0.65 and BOV=0.60 with MV (A) and AV (B) method.

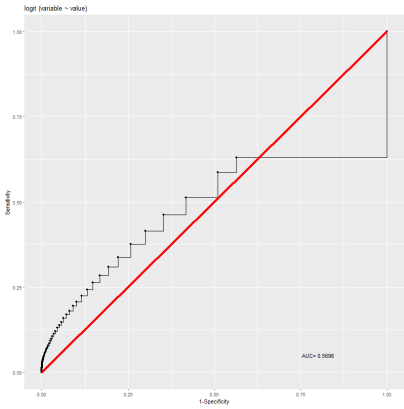
AKT

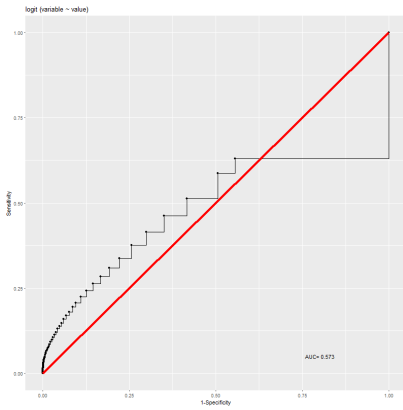


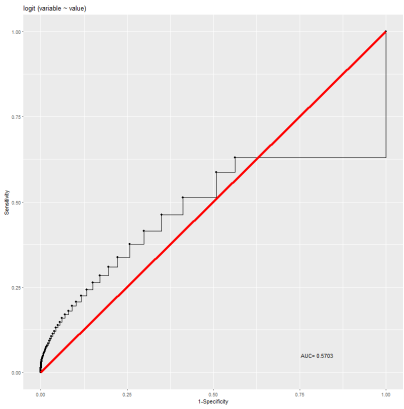


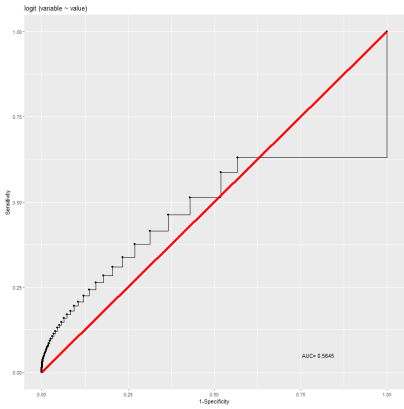


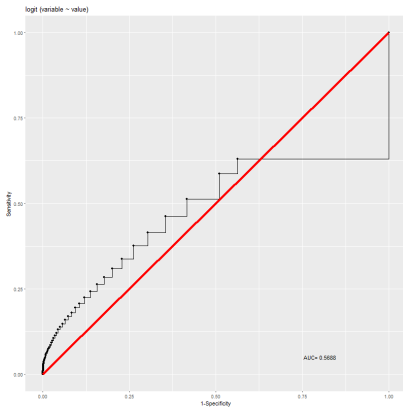


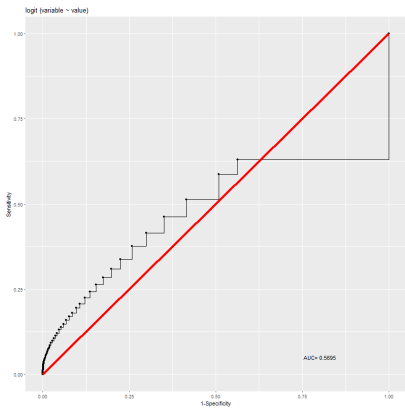






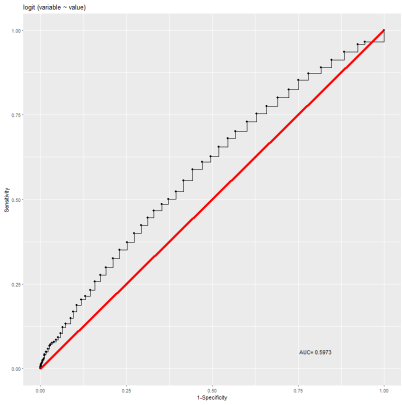


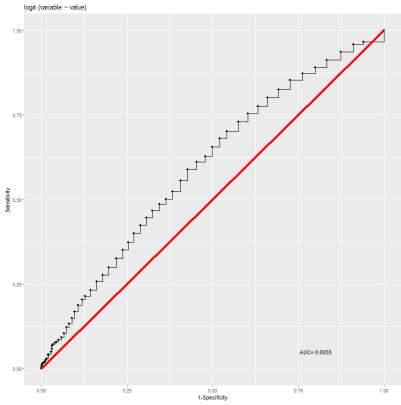


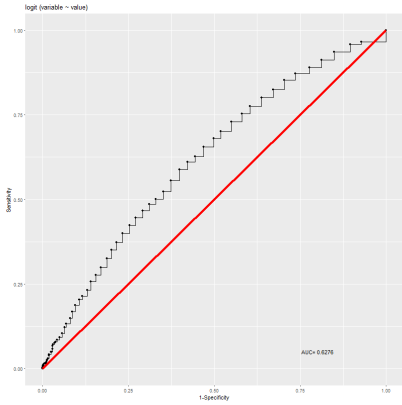


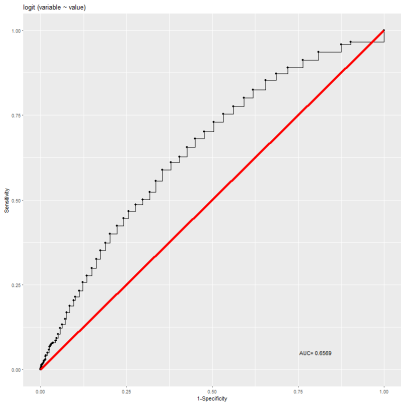
0.5741
0.5715
0.5669
0.5774
0.5696
0.573
0.5703
0.5645
0.5688
0.5695

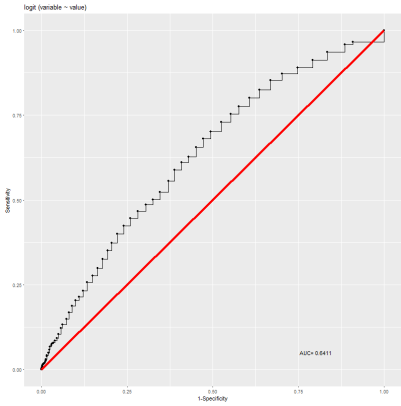
AMPC

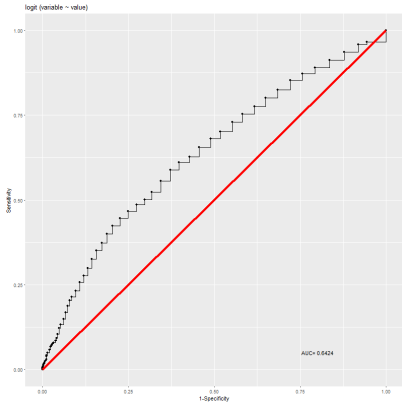


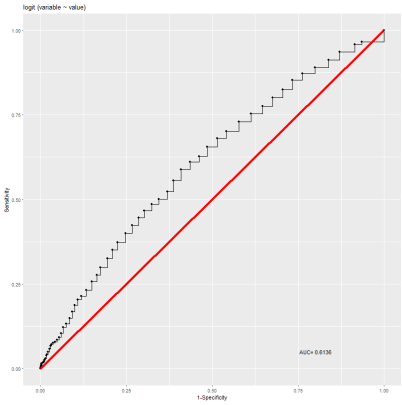


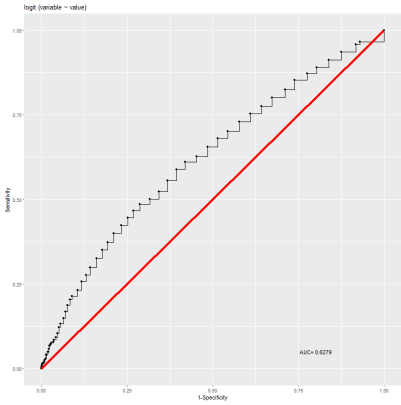


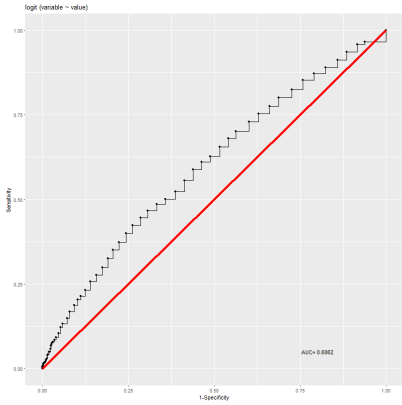


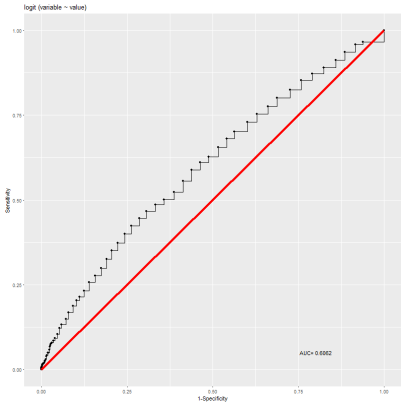






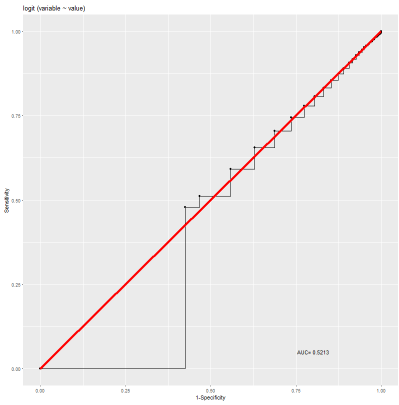


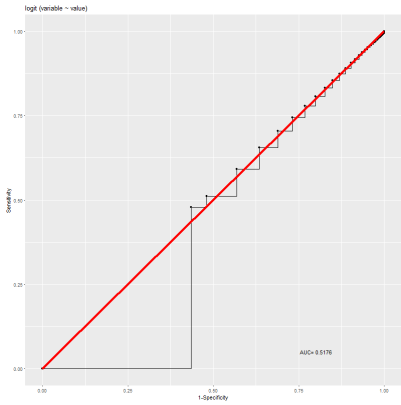


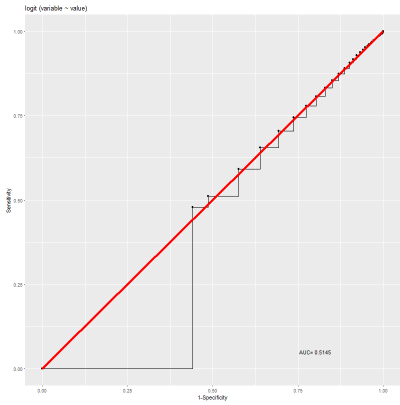


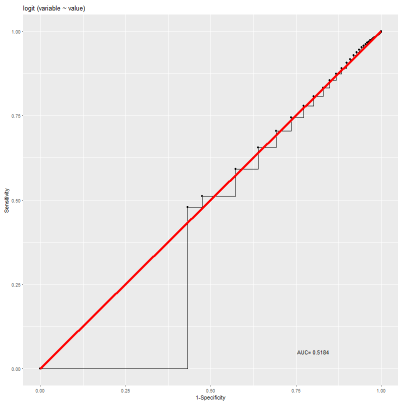
0.5973
0.6055
0.6276
0.6569
0.6411
0.6424
0.6136
0.6279
0.6062
0.6062

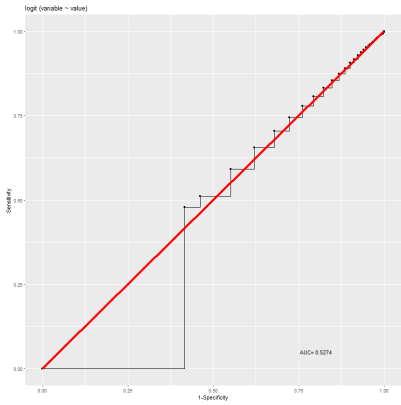
CP3A4

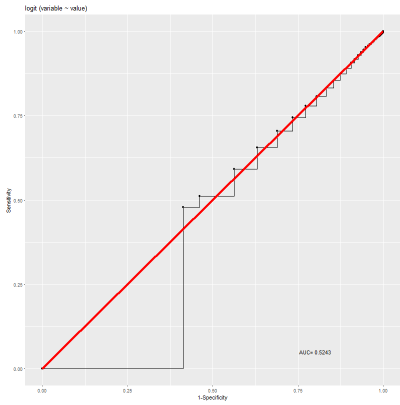


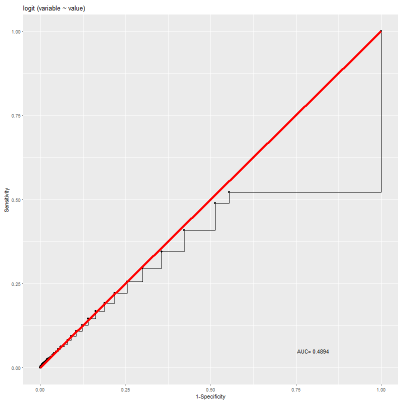


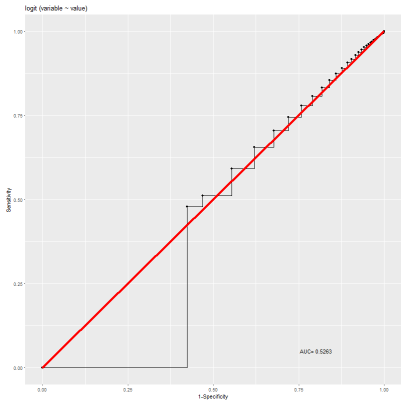


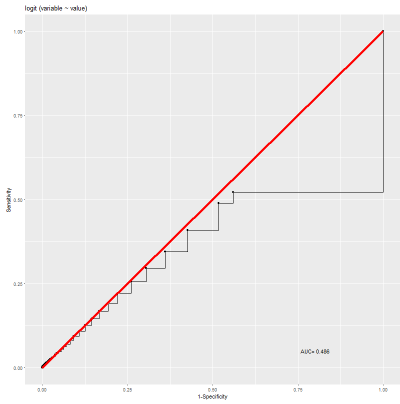


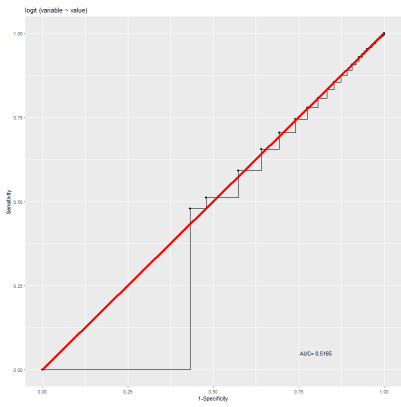






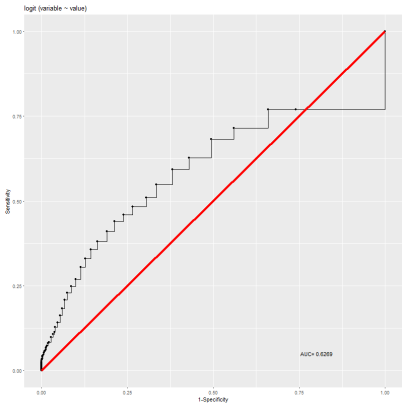


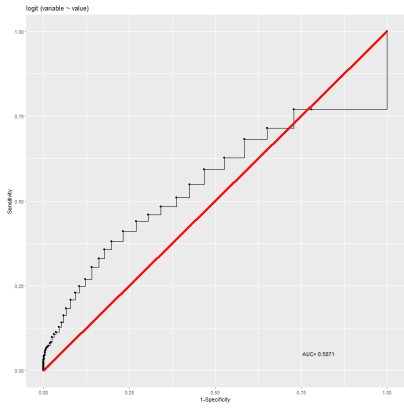


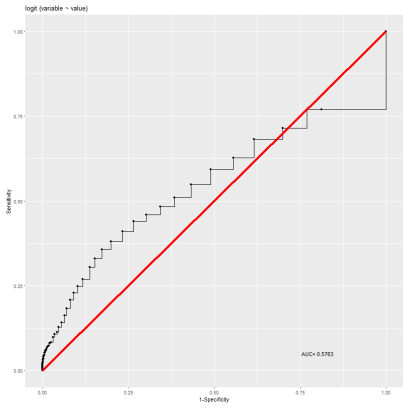


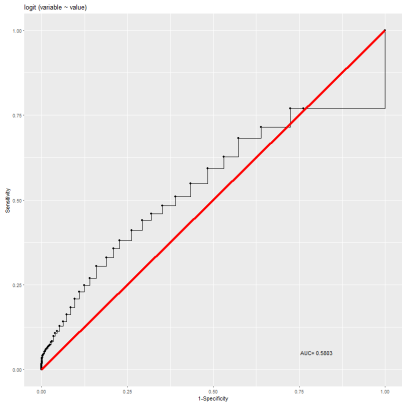
0.5213
0.5176
0.5145
0.5184
0.5274
0.5243
0.4894
0.5263
0.486
0.5165

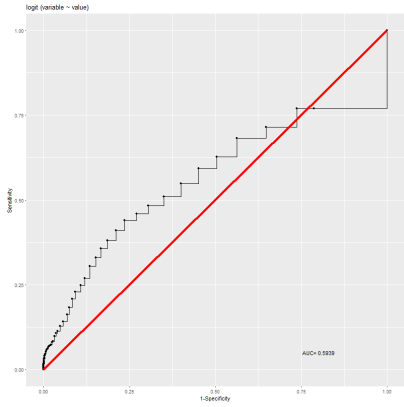
CXCR4

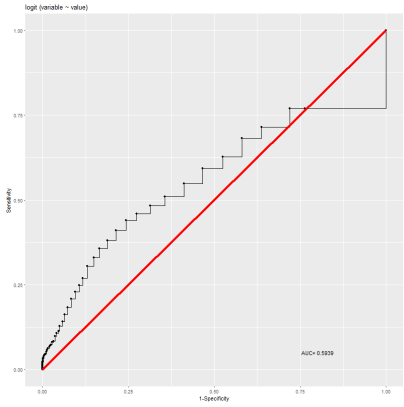


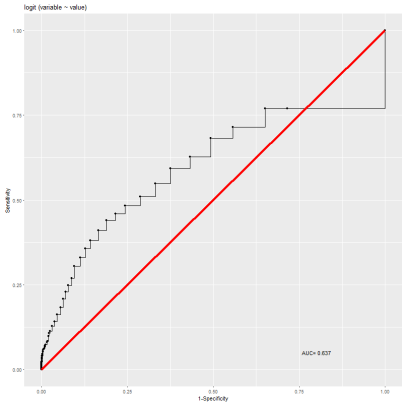


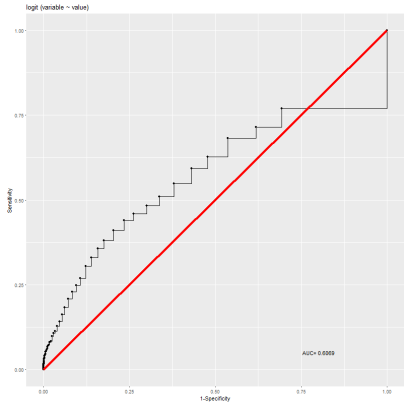


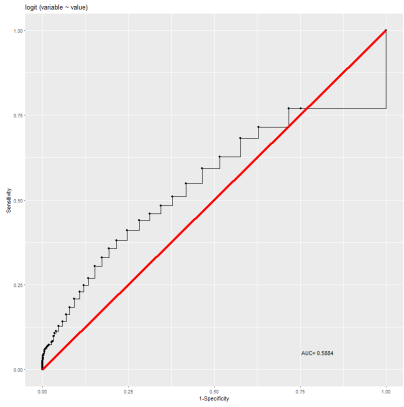


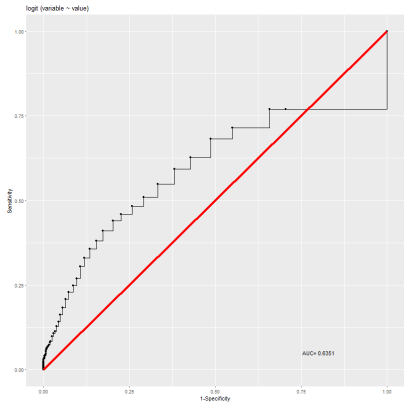






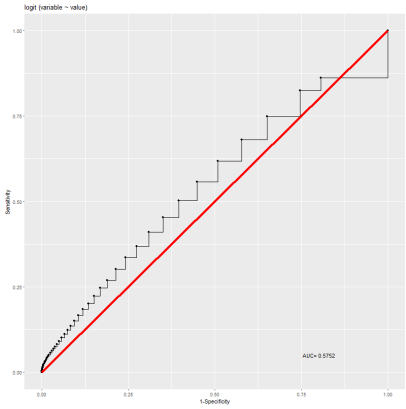


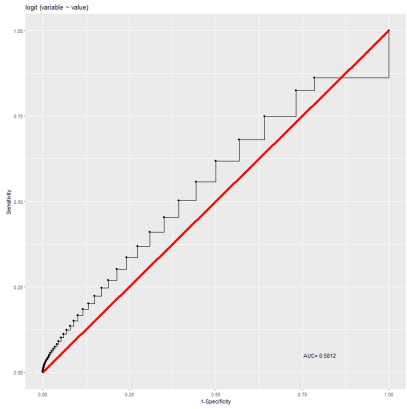


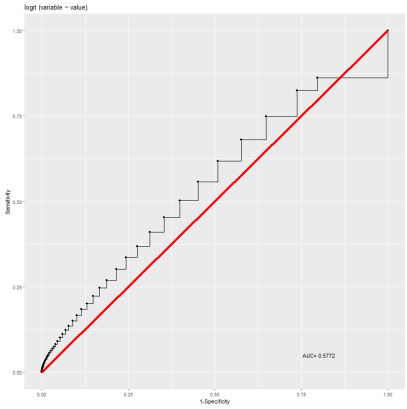


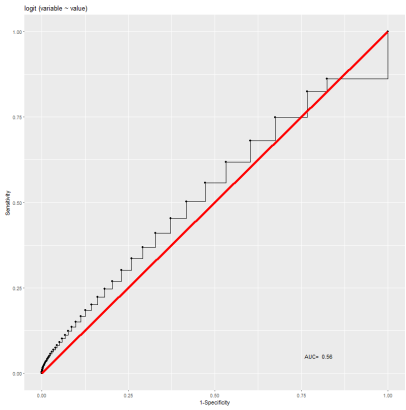
0.6269
0.5871
0.5763
0.5803
0.5939
0.5939
0.637
0.6069
0.5884
0.6351

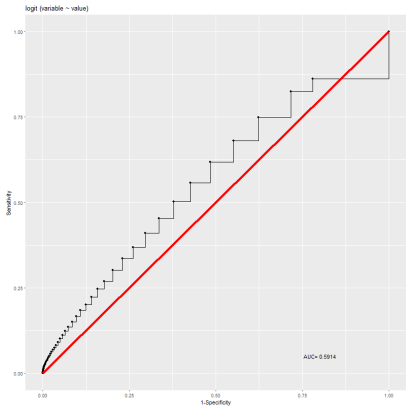
GCR

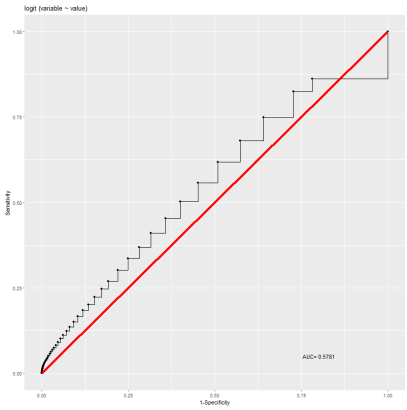


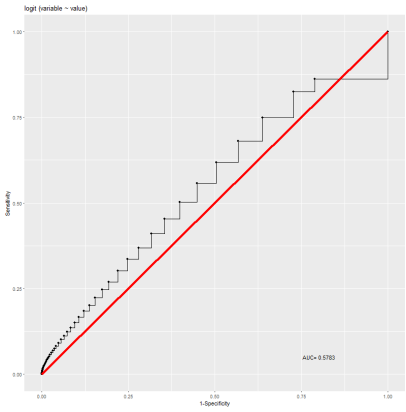


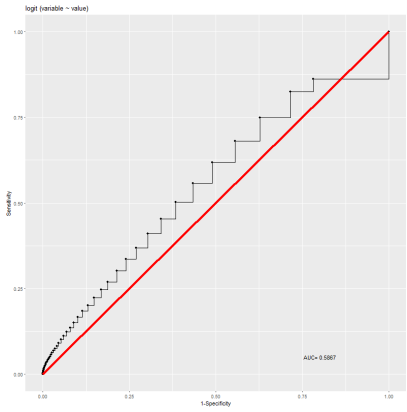


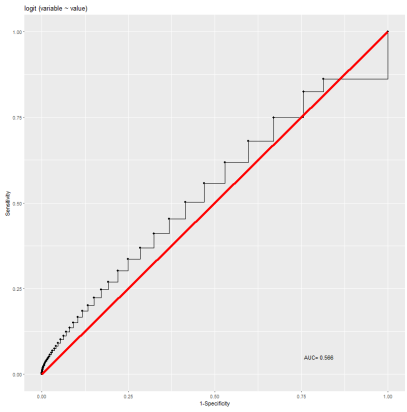


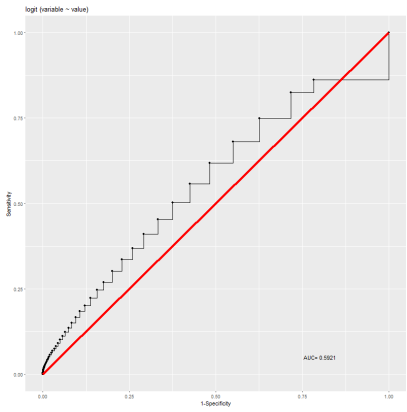






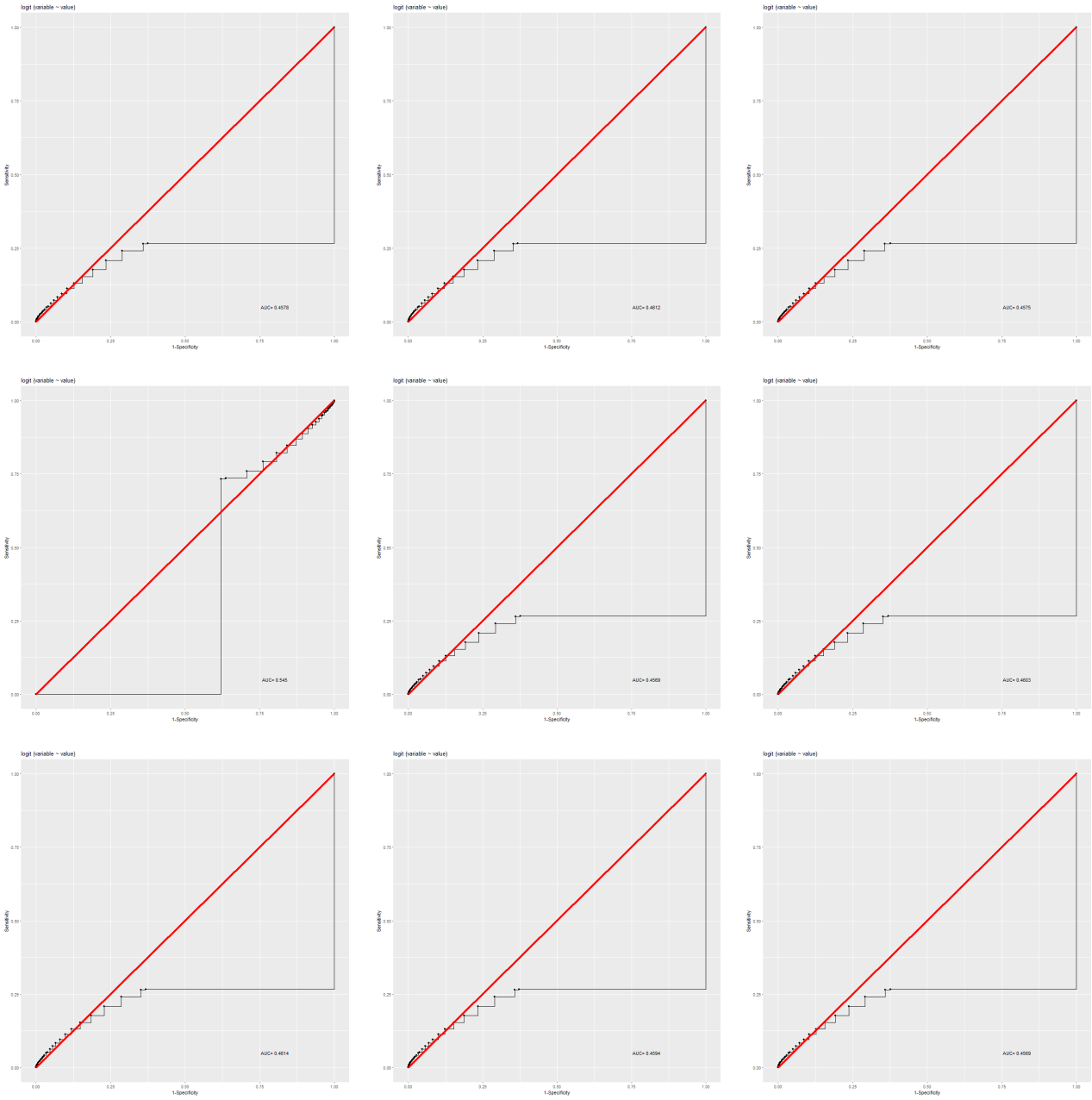






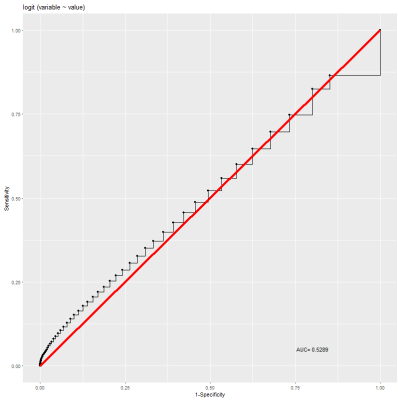
0.5752
0.5812
0.5772
0.56
0.5914
0.5781
0.5783
0.5867
0.566
0.5921

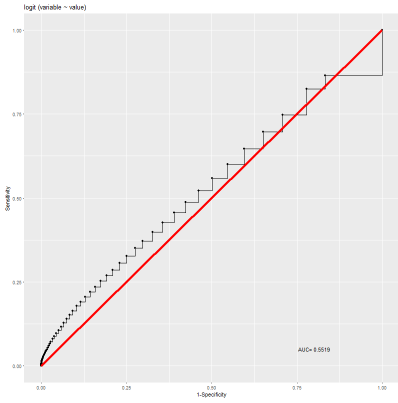
HIVPR

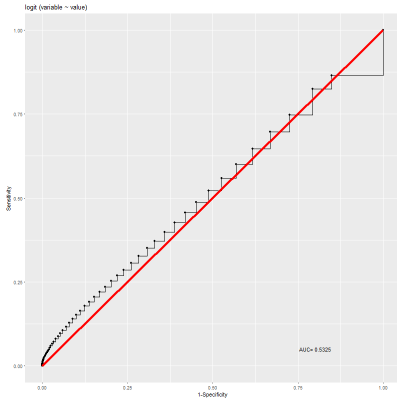


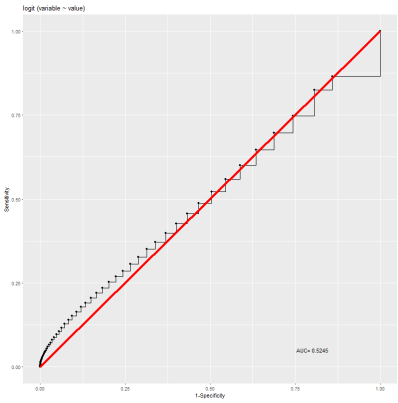
0.4578
0.4612
0.4575
0.545
0.4569
0.4603
0.4614
0.4594
0.4569

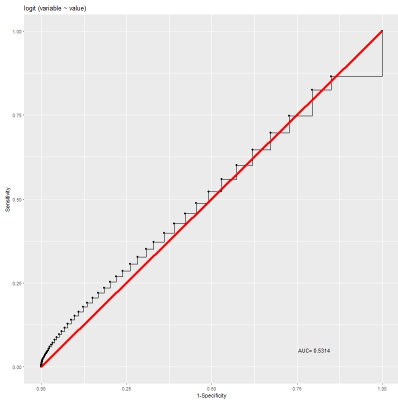
HIVRT

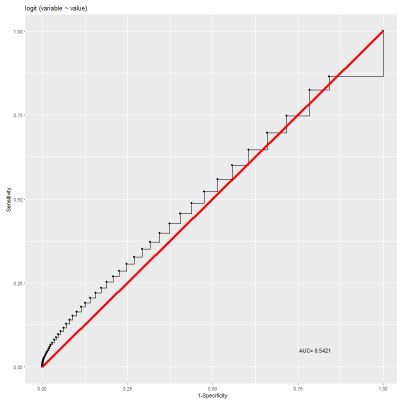


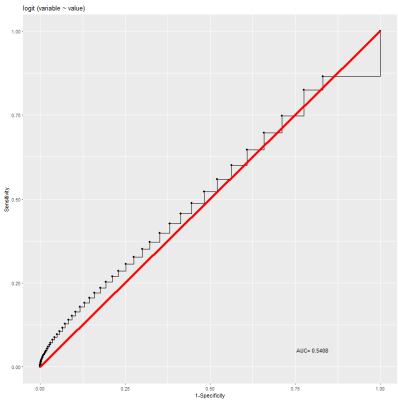


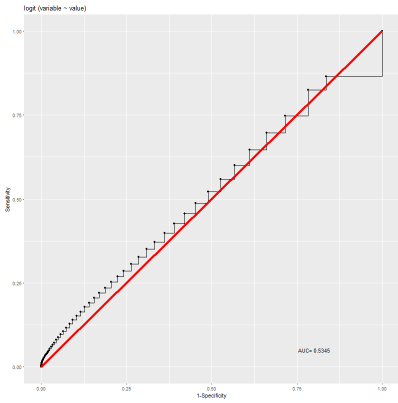


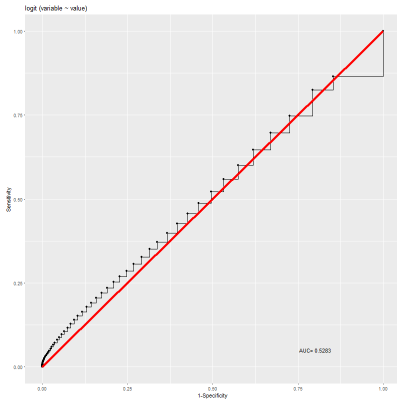


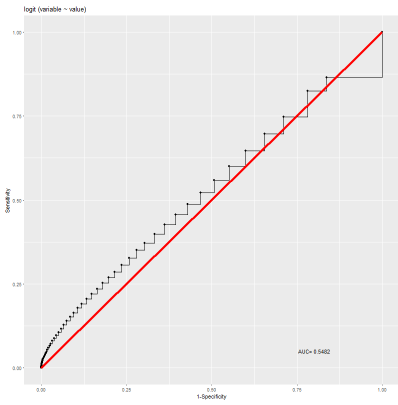






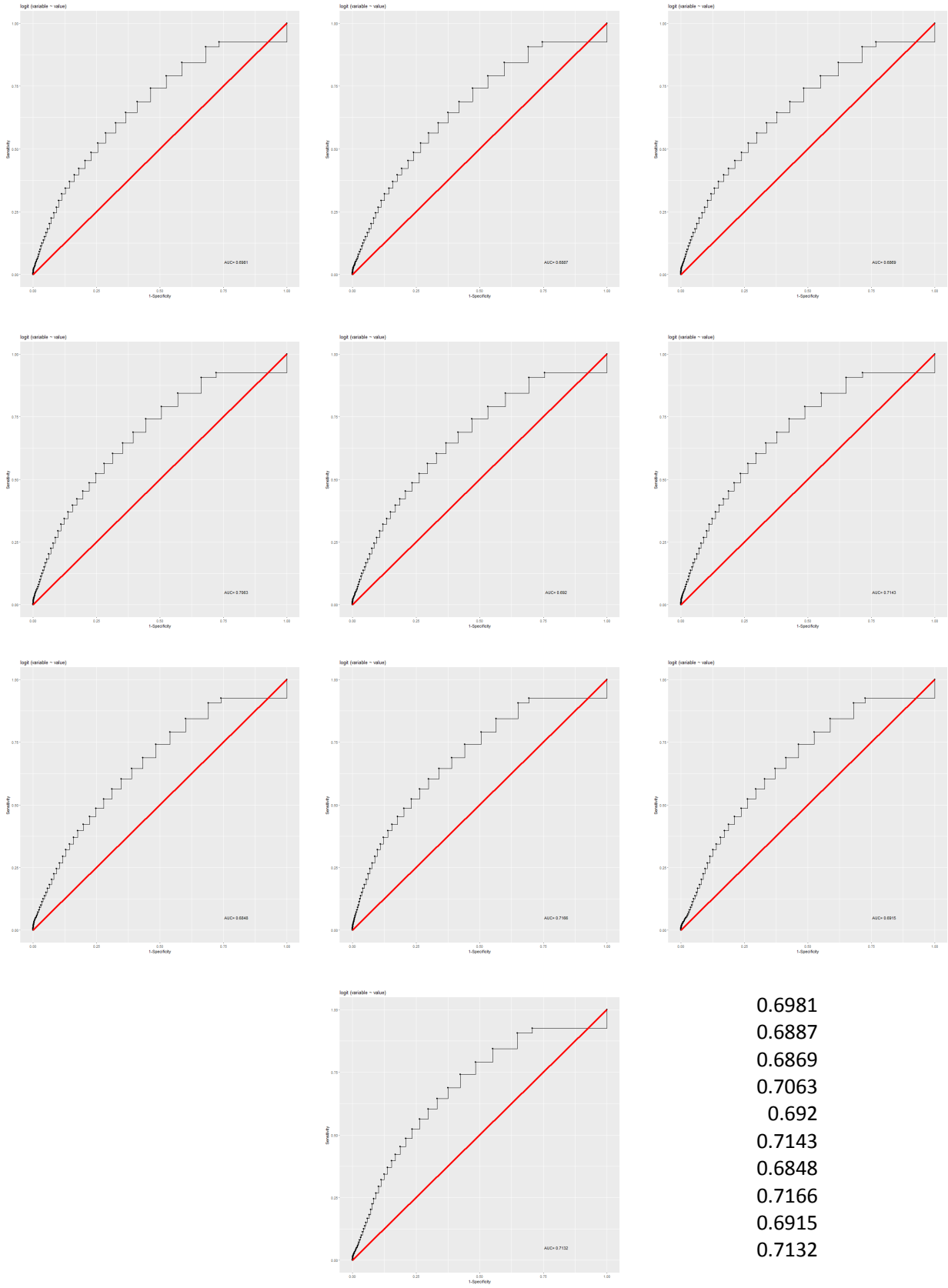






0.5289
0.5519
0.5325
0.5245
0.5314
0.5421
0.5408
0.5345
0.5283
0.5482

KIF11



0.6981
0.6887
0.6869
0.7063
0.692
0.7143
0.6848
0.7166
0.6915
0.7132

Figure S17. The AUC values and ROC curves for the DUD-E diversity set, with ten-fold resampling of decoys for each target.

S6) Examples of the use of reference shapes as comparators for ligand shapes (using the DUD-E set).

The balanced subsets (equal numbers of actives and decoys) used to produce the AUC values are ideal for the application of decision tree techniques. The rpart algorithm in R was used to identify the fingerprint bits that are most influential in distinguishing actives from decoys and in all cases classification was achieved that gave good levels of correct classification (Figure S19) and a confusion matrix corresponding to very low p-values when the chi-squared test was applied.[13] Selecting reference shapes that are influential in a majority of the ten-fold samples reduces dependence on the particular set of decoys. For each target, the reference shapes that are identified are shown in Figure S18 and coloured green if similarity to that shape increases the likelihood of being active or red if it increases the likelihood of being a decoy. These reveal some shape preferences: AKT1 binds ligands with a particular length and relatively cylindrical profile with a little variability tolerated. CP3A4 has a large, flexible binding site that can accommodate many drug-like molecules and it seems reasonable molecules must have an unusual shape, like the T-shape shown, in order not to bind.[16] GCR's shape preference is for molecules that have a small globular shape and activity is lost if the molecules protrude too far in any direction. Sets of influential reference shapes like these can be used in place of the full set of reference shapes to permit rapid screening. Shape comparison techniques such as ROCS require the structure of a known active to use as a search query. The shape fingerprint approach provides a more general definition of shape space that allows good and bad shapes to be identified. As with keys in locks, it is as important to leave gaps in the right places as well as to fill the right spaces. These key shapes provide a more straightforward means for human to discuss good and bad shared shape characteristics that link with biological activity and a mechanism for us to use the linguistic ability to describe shape by analogy.[17]

In the following section, each of the 8 DUD-E sets is subjected to a decision tree analysis, with ten different sets of decoys. The resulting trees are shown and then below each diagram a

summary of the bits identified as discriminating actives from decoys is listed for each of the two sets of decoys. Commonalities are then identified in a highlighted row.

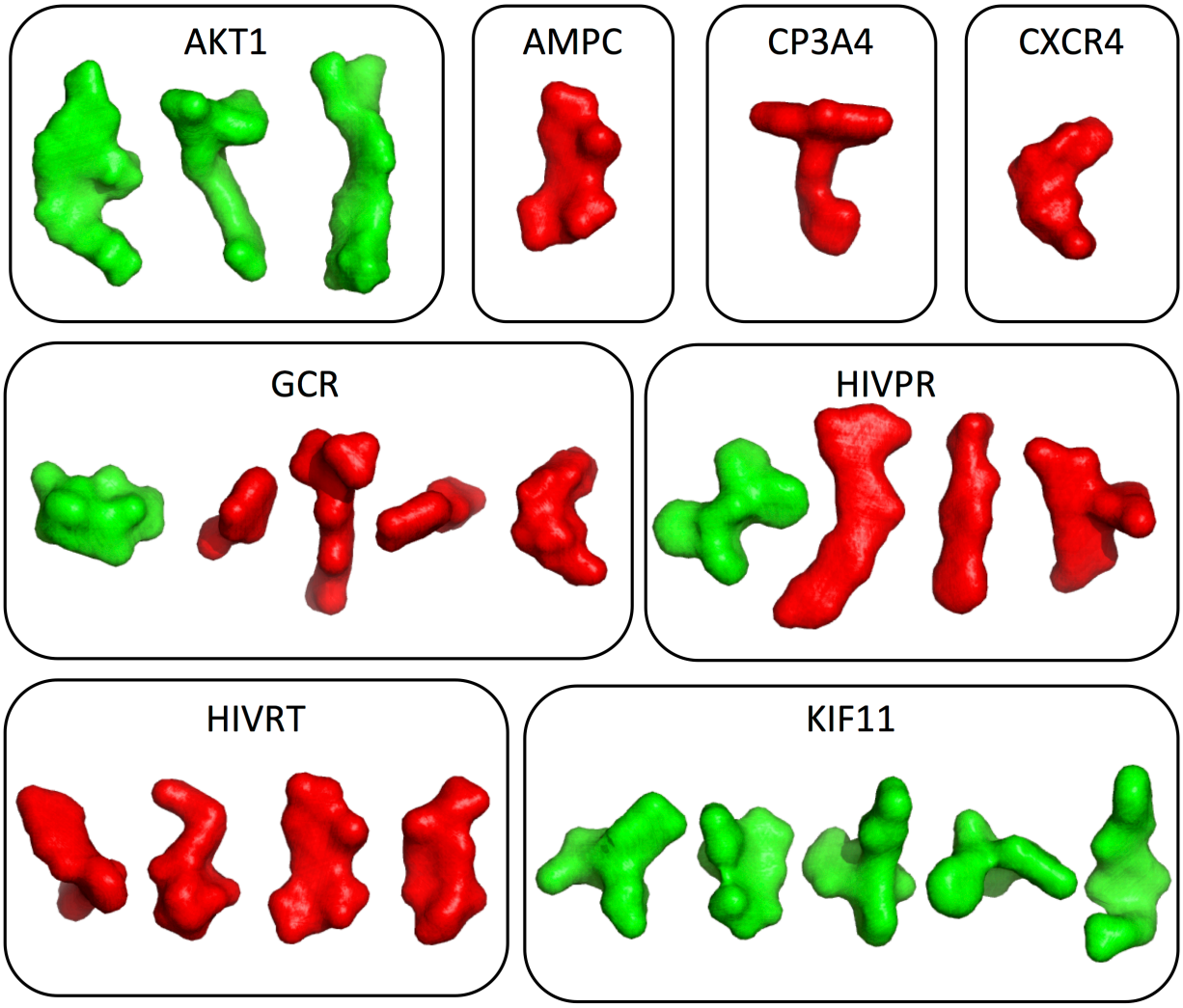
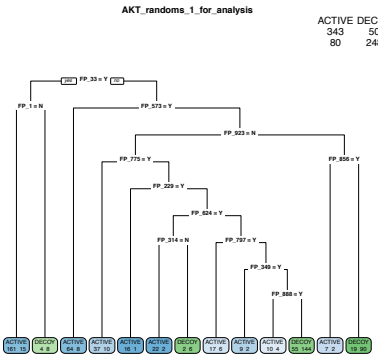
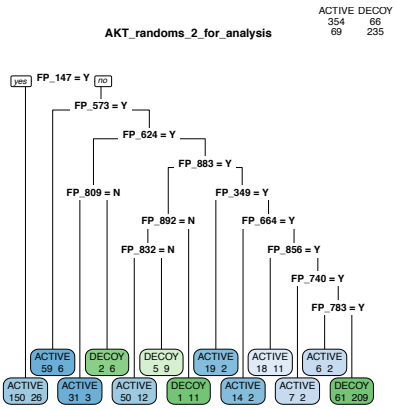
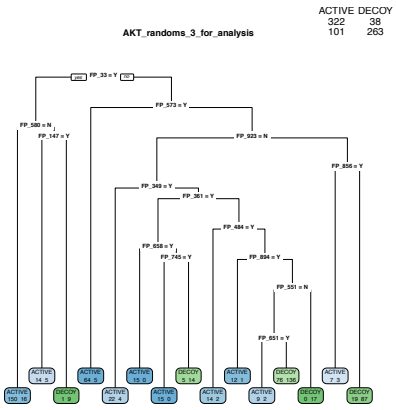


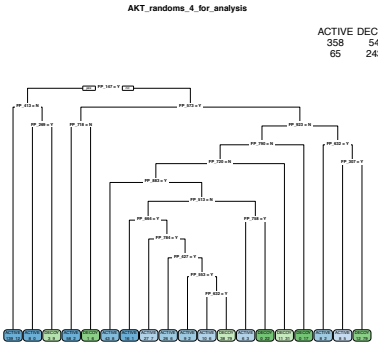
Figure S18. Reference shapes identified as distinguishing actives from decoys for each of the targets. Molecules with shape similar to those colored green are more likely to be actives while those more like those colored red are more likely to be decoys.

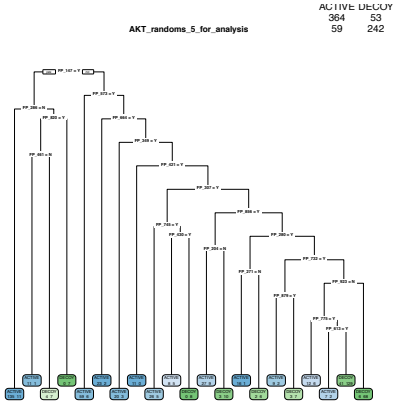
AKT

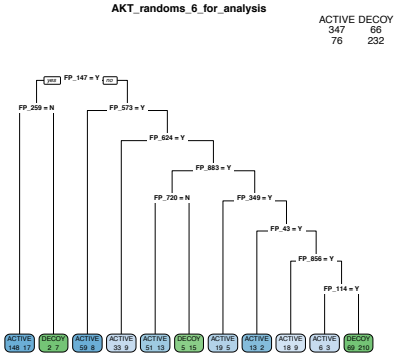

AKT_randoms_1_for_analysis
ACTIVE DECOY
343 50
80 248

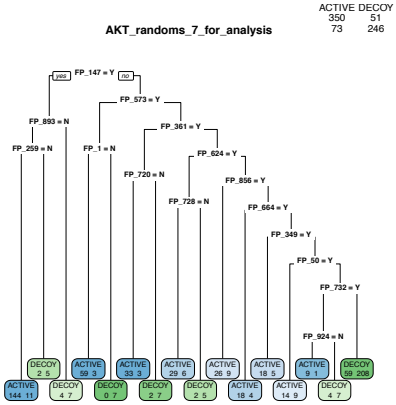

AKT_randoms_2_for_analysis
ACTIVE DECOY
354 66
69 235
yes
FP_147 = Y
no
FP_573 = Y
FP_624 = Y
FP_883 = Y
FP_809 = N
FP_349 = Y
FP_892 = N
FP_664 = Y
FP_832 = N
FP_856 = Y
FP_740 = Y
FP_783 = Y
ACTIVE 59 6
DECOY 2 6
DECOY 5 9
ACTIVE 19 2
ACTIVE 18 11
ACTIVE 6 2
ACTIVE 150 26
ACTIVE 31 3
ACTIVE 50 12
DECOY 1 11
ACTIVE 14 2
ACTIVE 7 2
DECOY 61 209

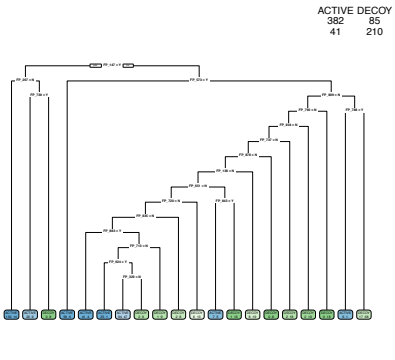

AKT_randoms_3_for_analysis
ACTIVE DECOY
322 38
101 263

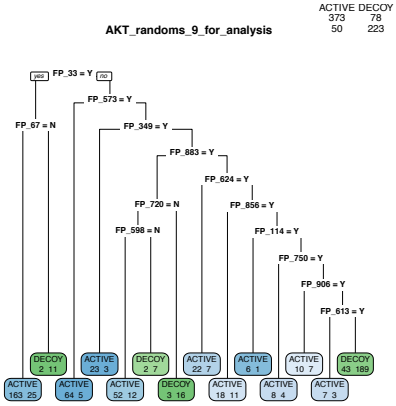

AKT_randoms_4_for_analysis
ACTIVE DECOY
358 54
65 243

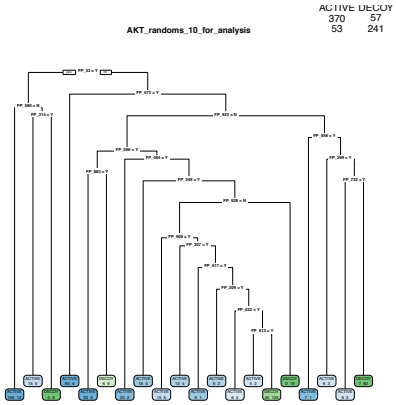

AKT_randoms_5_for_analysis
ACTIVE DECOY
364 53
59 242

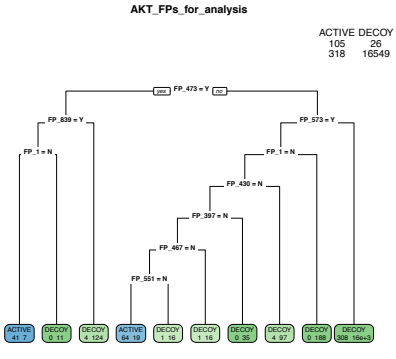

AKT_randoms_6_for_analysis
ACTIVE DECOY
347 66
76 232
yes
FP_147 = Y
no
FP_259 = N
FP_573 = Y
FP_624 = Y
FP_883 = Y
FP_720 = N
FP_349 = Y
FP_43 = Y
FP_856 = Y
FP_114 = Y
ACTIVE 148 17
DECOY 2 7
ACTIVE 59 8
ACTIVE 33 9
ACTIVE 51 13
DECOY 5 15
ACTIVE 19 5
ACTIVE 13 2
ACTIVE 18 9
ACTIVE 6 3
DECOY 69 218


AKT_randoms_7_for_analysis
ACTIVE DECOY
350 51
73 246
yes
FP_147 = Y
no
FP_573 = Y
FP_893 = N
FP_361 = Y
FP_259 = N
FP_1 = N
FP_624 = Y
FP_720 = N
FP_856 = Y
FP_728 = N
FP_664 = Y
FP_349 = Y
FP_50 = Y
FP_732 = Y
FP_924 = N
DECOY 2 5
ACTIVE 59 3
ACTIVE 33 3
ACTIVE 29 6
ACTIVE 26 9
ACTIVE 18 5
ACTIVE 9 1
DECOY 59 208
ACTIVE 144 11
DECOY 4 7
DECOY 0 7
DECOY 2 7
DECOY 2 5
ACTIVE 18 4
ACTIVE 14 9
DECOY 4 7


AKT_randoms_8_for_analysis
ACTIVE DECOY
382 85
41 210


AKT_randoms_9_for_analysis
ACTIVE DECOY
373 78
50 223
yes
FP_33 = Y
no
FP_573 = Y
FP_67 = N
FP_349 = Y
FP_883 = Y
FP_624 = Y
FP_720 = N
FP_856 = Y
FP_598 = N
FP_114 = Y
FP_750 = Y
FP_906 = Y
FP_613 = Y
DECOY 2 11
ACTIVE 23 3
DECOY 2 7
ACTIVE 22 7
ACTIVE 6 1
ACTIVE 10 7
DECOY 43 188
ACTIVE 163 25
ACTIVE 64 5
ACTIVE 52 12
DECOY 3 16
ACTIVE 18 11
ACTIVE 8 4
ACTIVE 7 3

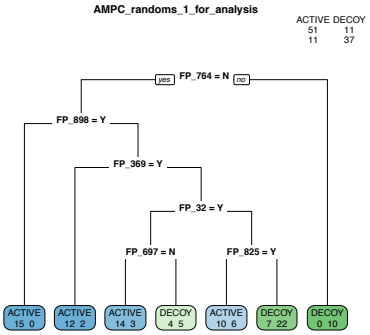

AKT_randoms_10_for_analysis
ACTIVE DECOY
370 57
53 241

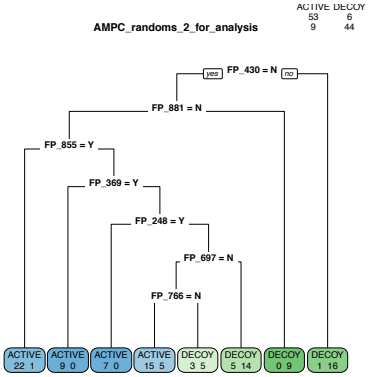

AKT_FPs_for_analysis
ACTIVE DECOY
105 26
318 16549
yes
FP_473 = Y
no
FP_839 = Y
FP_573 = Y
FP_1 = N
FP_1 = N
FP_430 = N
FP_397 = N
FP_467 = N
FP_551 = N
ACTIVE 41 7
DECOY 0 11
DECOY 4 124
ACTIVE 84 19
DECOY 1 16
DECOY 1 18
DECOY 0 35
DECOY 4 97
DECOY 0 188
DECOY 308 16543

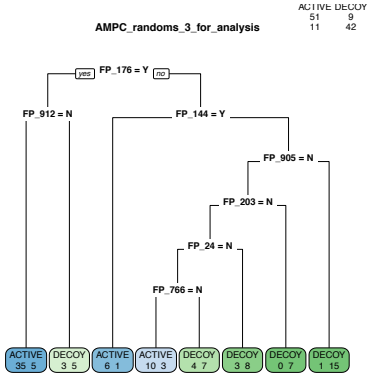
Sample Number	Positive bits	Negative bits	True Actives	True Decoys	False Actives	False Decoys
1	33, 573, 775, 229, 624, 797, 349, 888	1, 923, 314	343	248	50	80
2	147, 573, 624, 883, 349, 664, 856, 740, 783	809, 892, 832	354	235	66	69
3	33, 573, 147, 349, 361, 658, 745, 484, 894, 651	923, 580, 551	322	263	38	101
4	147, 269, 573, 632, 307, 883, 664, 758, 784, 427, 553, 632	413, 718, 923, 720, 513	358	243	54	65
5	147, 573, 820, 664, 349, 421, 307, 856, 745, 430, 280, 732, 879, 775, 613	286, 461, 204, 271, 923	364	242	53	59
6	147, 573, 624, 883, 349, 43, 856, 114	259, 720	347	232	66	76
7	147, 573, 361, 624, 856, 664, 349, 50, 732	893, 259, 1, 720, 728, 924	350	240	51	73
8	147, 730, 573, 788, 843, 883, 624	267, 809, 790, 834, 737, 878, 138, 651, 720, 835, 713, 329	382	210	85	41
9	33, 573, 349, 883, 624, 856, 114, 750, 906, 613	67, 720, 598	373	223	78	50
10	33, 573, 314, 596, 883, 484, 349, 908, 307, 617, 209, 632,	580, 923, 928	370	241	57	53

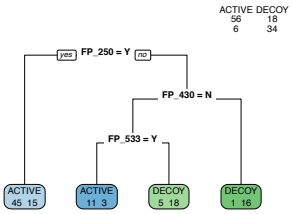
	613, 856, 299, 732					
Average/ consensus	147, 349, 573	None	49.5 %	33.0 %	8.3 %	9.3 %
All	473, 839, 573, 624, 856, 883	1, 430, 397, 467, 551	105 (0.6 %)	16549 (97.4 %)	26 (0.2 %)	318 (1.9 %)

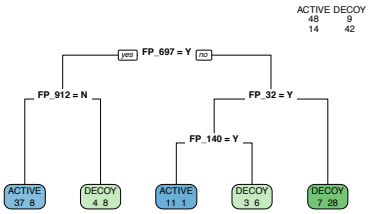
AMPC

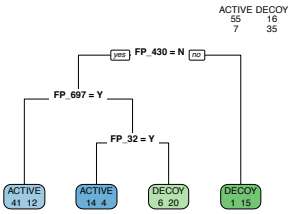


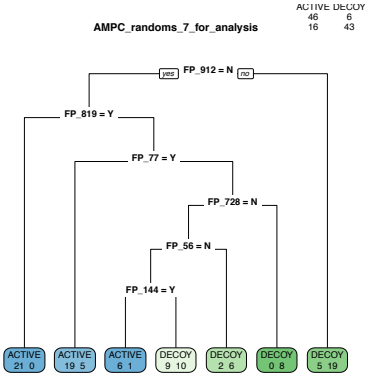


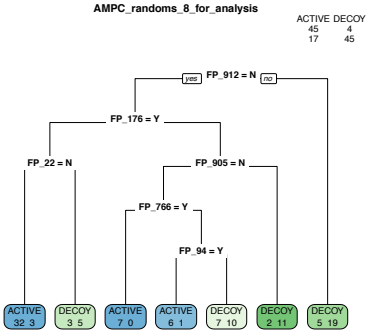


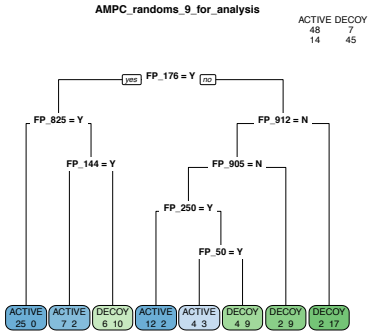


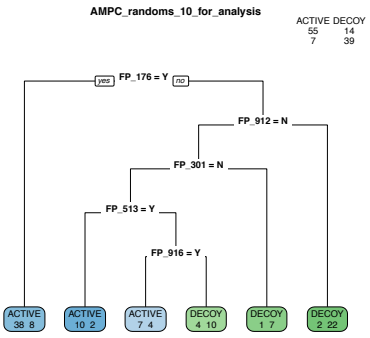


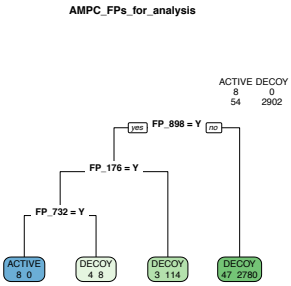








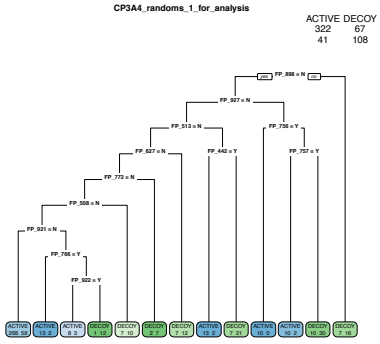


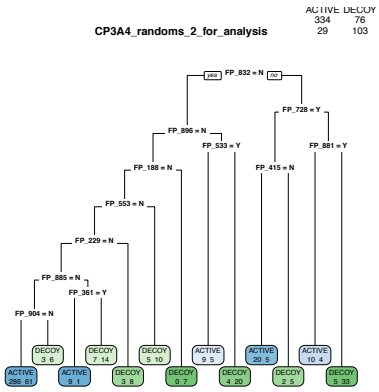


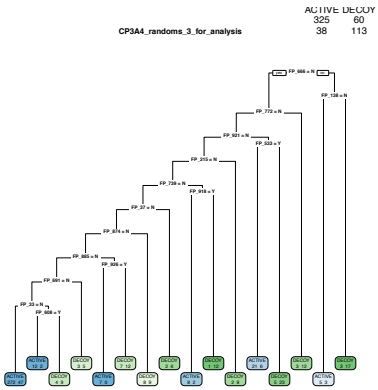
Sample Number	Positive bits	Negative bits	True Actives	True Decoys	False Actives	False Decoys

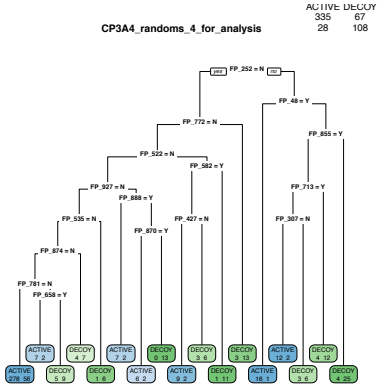
1	898, 369, 32, 825	764, 697	51	37	11	11
2	855, 369, 248	430, 881, 697, 766	53	44	6	9
3	176, 144	912, 905, 203, 24, 766	51	42	9	11
4	250, 533	430	56	34	18	6
5	697, 32, 140	912	48	42	9	14
6	697, 32	430	55	35	16	7
7	819, 77, 144	912, 728, 56	46	43	6	16
8	176, 766, 94	912, 905, 22	45	45	4	17
9	176, 825, 144, 250, 50	912, 905	48	45	7	14
10	176, 513, 916	912, 301	55	39	14	7
Average/ consensus	None	912	45.1 %	36.1 %	8.9 %	10.0 %
All	898, 176, 732		8 (0.3 %)	2902 (97.9 %)	0 (0 %)	54 (1.8 %)

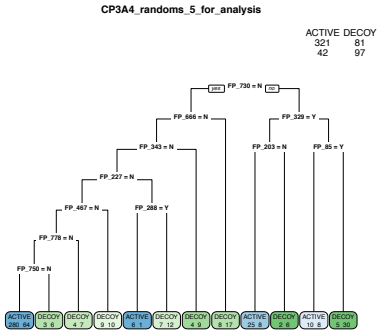
CP3A4

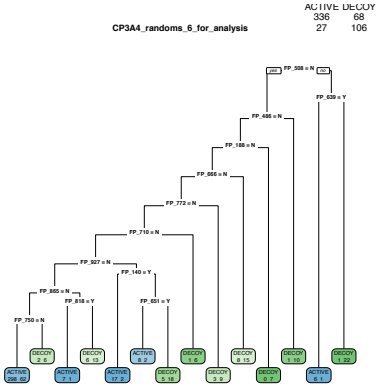


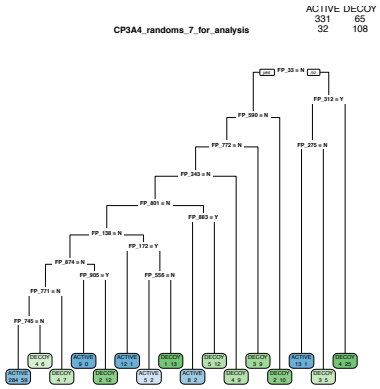


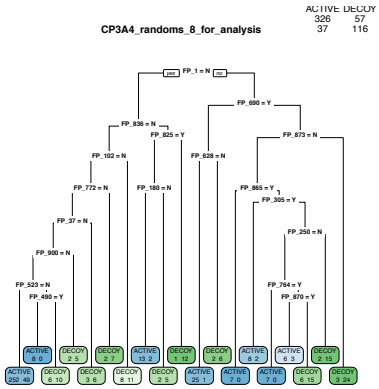


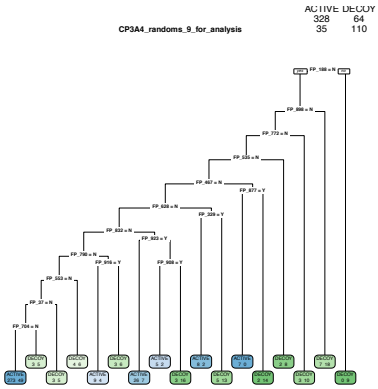


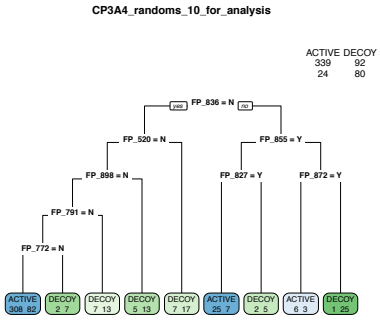


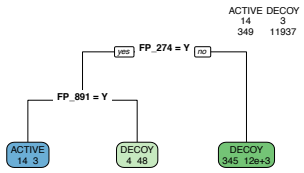








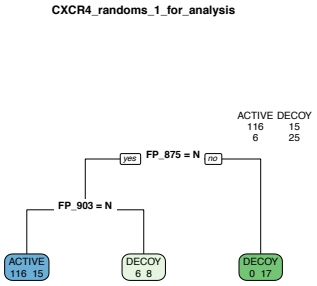


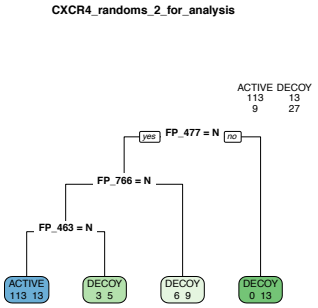


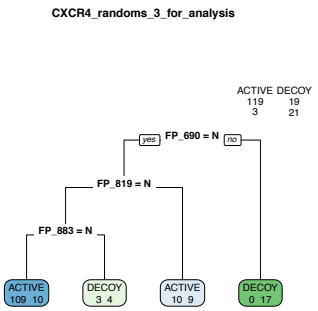
Sample Number	Positive bits	Negative bits	True Actives	True Decoys	False Actives	False Decoys

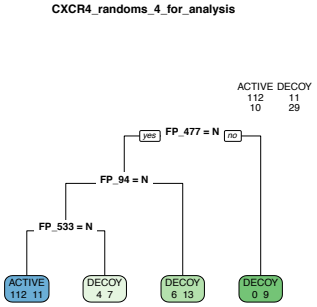
1	756, 757, 442, 766, 922	898, 927, 513, 627, 772, 508, 921	322	108	67	41
2	728, 881, 533, 361	832, 896, 415, 188, 553, 229, 885, 904	334	103	76	29
3	533, 918, 926, 608	666, 138, 772, 921, 215, 739, 37, 874, 885, 691, 33	325	113	60	38
4	48, 855, 713, 582, 888, 870, 658	252, 772, 307, 522, 427, 927, 535, 874, 781	335	108	67	28
5	329, 85, 288	730, 203, 666, 343, 227, 467, 778, 750	321	97	81	42
6	639, 140, 651, 818	508, 486, 188, 666, 772, 710, 927, 865, 750	336	106	68	27
7	312, 883, 172, 874, 905	33, 590, 275, 772, 343, 801, 138, 556, 874, 771, 745	331	108	65	32
8	690, 865, 305, 764, 870, 825, 490	1, 836, 873, 250, 628, 180, 102, 772, 37, 900, 523	326	116	57	37
9	877, 329, 923, 908, 916	188, 898, 772, 535, 467, 628, 832, 790, 553, 37, 704	328	110	64	35
10	855, 827, 872	836, 520, 898, 791, 772	339	80	92	24
Average/ consensus	None	772	61.3 %	19.5 %	13.0 %	6.2 %
All	274, 891	None	14 (0.1 %)	11937 (97.0 %)	3 (0.0 %)	349 (2.8 %)

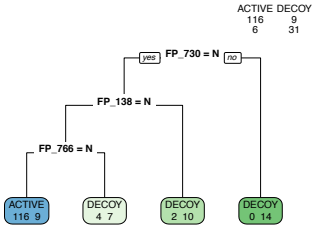
CXCR4

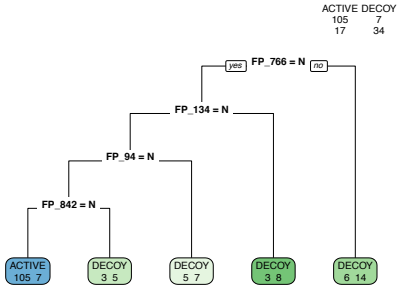


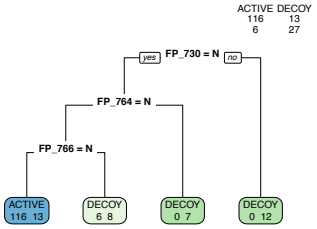


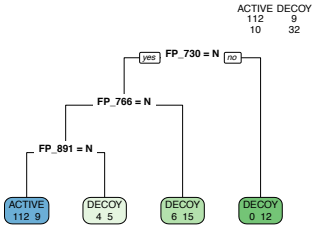


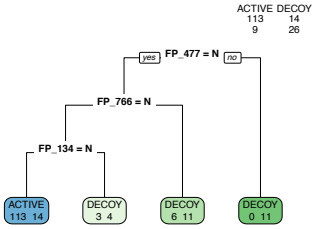


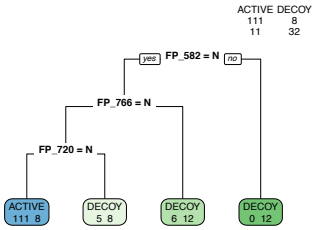


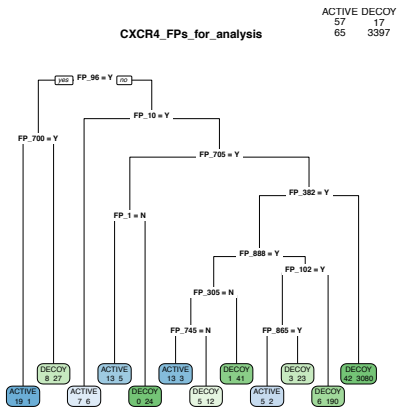








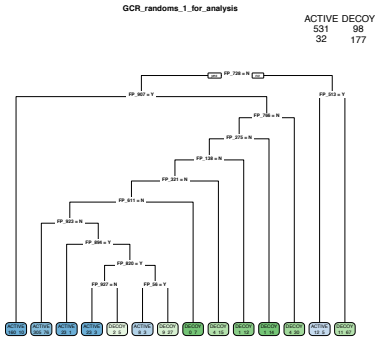


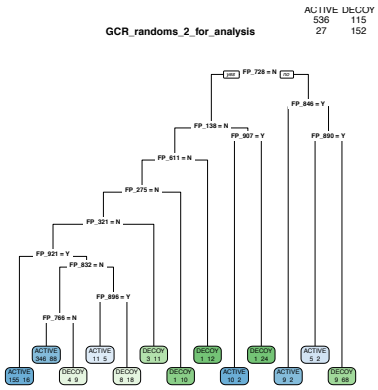


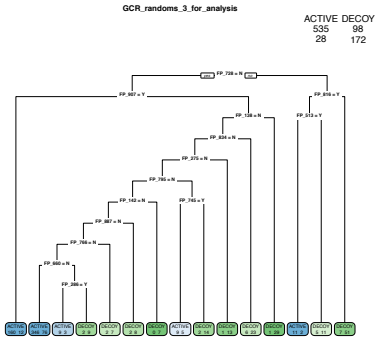
Sample Number	Positive bits	Negative bits	True Actives	True Decoys	False Actives	False Decoys
1		875, 903	116	25	15	6

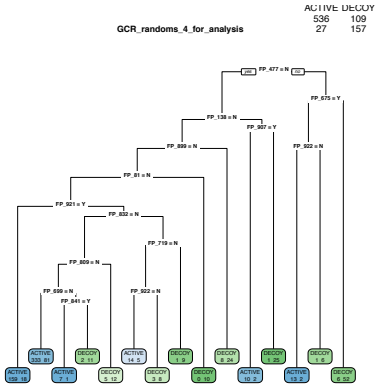
2		477, 766, 463	113	27	13	9
3		690, 819, 883	119	21	19	3
4		477, 94, 533	112	29	11	10
5		730, 138, 766	116	31	9	6
6		766, 134, 94, 842	105	34	7	17
7		730, 764, 766	116	27	13	6
8		730, 766, 891	112	32	9	10
9		477, 766, 134	113	26	14	9
10		582, 766, 720	111	32	8	11
Average/ consensus		766	69.9 %	17.5 %	7.3 %	5.4 %
All	96, 10, 700, 705, 382, 888, 102, 865	1, 305, 745	57 (1.6 %)	3397 (96.1 %)	17 (0.5 %)	65 (1.8 %)

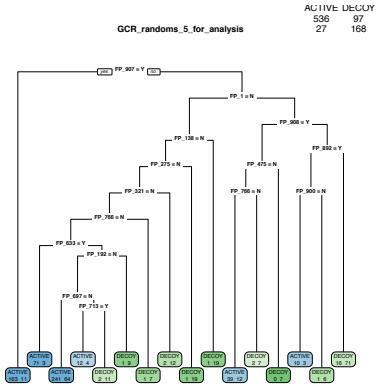
GCR

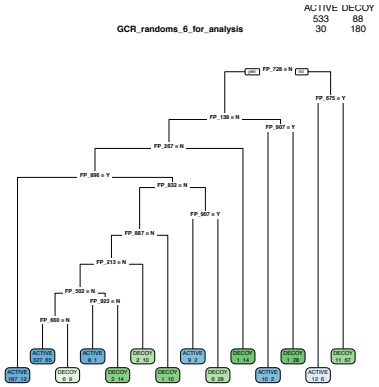

GCR_randoms_1_for_analysis
ACTIVE DECOY
531 98
32 177

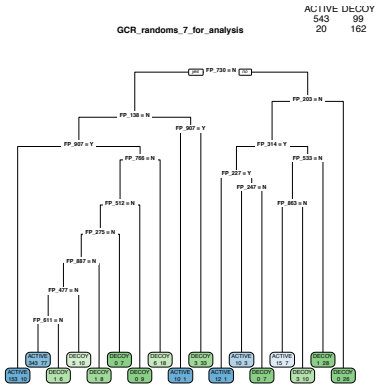

GCR_randoms_2_for_analysis
ACTIVE DECOY
536 115
27 152

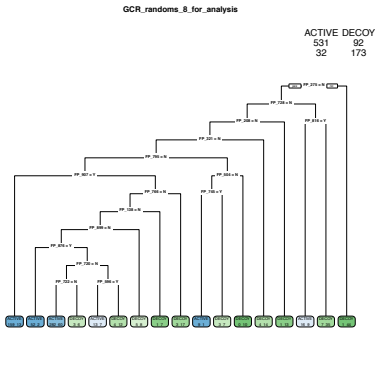

GCR_randoms_3_for_analysis
ACTIVE DECOY
535 98
28 172

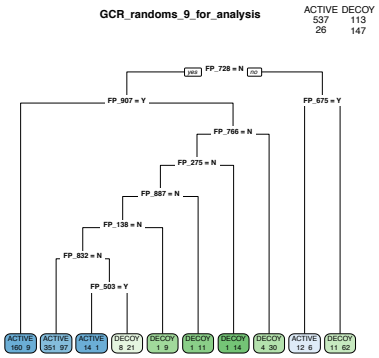

GCR_randoms_4_for_analysis
ACTIVE DECOY
536 109
27 157

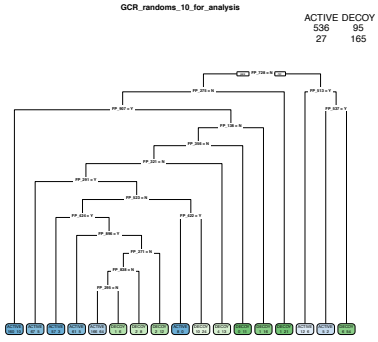

GCR_randoms_5_for_analysis
ACTIVE DECOY
536 97
27 168

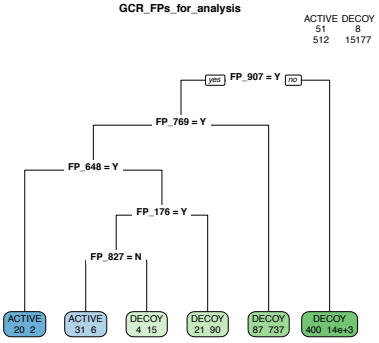

GCR_randoms_6_for_analysis
ACTIVE DECOY
533 88
30 180


GCR_randoms_7_for_analysis
ACTIVE DECOY
543 99
20 162


GCR_randoms_8_for_analysis
ACTIVE DECOY
531 92
32 173


GCR_randoms_9_for_analysis
ACTIVE DECOY
537 113
26 147
yes
FP_728 = N
no
FP_907 = Y
FP_675 = Y
FP_766 = N
FP_275 = N
FP_887 = N
FP_138 = N
FP_832 = N
FP_503 = Y
ACTIVE 160 9
ACTIVE 351 97
ACTIVE 14 1
DECOY 8 21
DECOY 1 9
DECOY 1 11
DECOY 1 14
DECOY 4 30
ACTIVE 12 6
DECOY 11 62


GCR_randoms_10_for_analysis
ACTIVE DECOY
536 95
27 165


GCR_FPs_for_analysis
ACTIVE DECOY
51 8
512 15177
yes
FP_907 = Y
no
FP_769 = Y
FP_648 = Y
FP_176 = Y
FP_827 = N
ACTIVE 20 2
ACTIVE 31 6
DECOY 4 15
DECOY 21 90
DECOY 87 737
DECOY 400 14e+3

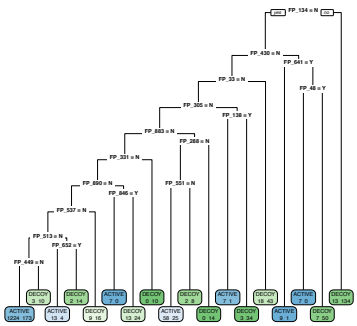
Sample Number	Positive bits	Negative bits	True Actives	True Decoys	False Actives	False Decoys
1	907, 513, 894, 820, 56	728, 766, 275, 138, 321, 611, 923, 927	531	177	98	32
2	846, 890, 907, 921, 896	728, 138, 611, 275, 321, 832, 766	536	152	115	27
3	907, 816, 513, 745, 286	728, 138, 834, 275, 795, 142, 887, 766, 660	535	172	98	28
4	675, 907, 841	477, 138, 922, 899, 81, 832, 719, 809, 922, 699	536	157	109	27
5	907, 908, 892, 633, 713	1, 138, 475, 275, 766, 900, 321, 766, 192, 697	536	168	97	27
6	675, 907, 896, 907	728, 138, 267, 896, 832, 887, 213, 502, 923, 660	533	180	88	30
7	907, 314, 227, 907	730, 203, 533, 247, 863, 138, 766, 512, 275, 887, 477, 611	543	162	99	20
8	816, 907, 745, 876, 596	275, 728, 208, 321, 795, 504, 766, 138, 899, 720, 722	531	173	92	32
9	907, 675, 503	728, 766, 275, 887, 138, 832	537	147	113	26
10	513, 537, 907, 291, 422, 424, 896	728, 275, 138, 356, 321, 523, 271, 838, 295	536	165	95	27

Average/ consensus	907	138, 275, 728, 766	64.6 %	19.9 %	12.1 %	3.3 %
All	907, 769, 648, 176	827	51 (0.3 %)	15177 (96.4 %)	8 (0.1 %)	512 (3.3 %)

HIVPR

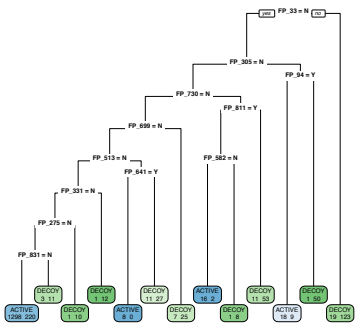
HIVPR_randoms_1_for_analysis

	ACTIVE	DECOY
	1325	204
	70	357



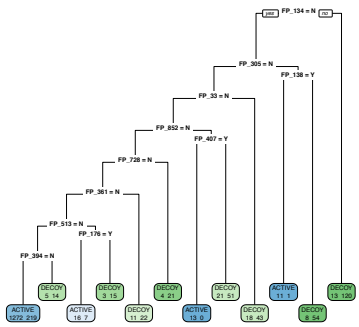
HIVPR_randoms_2_for_analysis

	ACTIVE	DECOY
	1340	231
	55	319



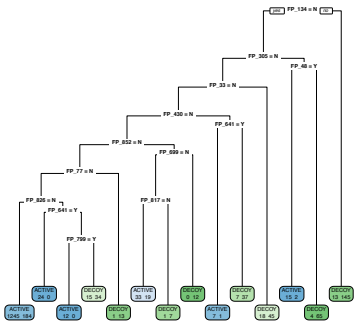
HIVPR_randoms_3_for_analysis

	ACTIVE	DECOY
	1312	227
	83	340



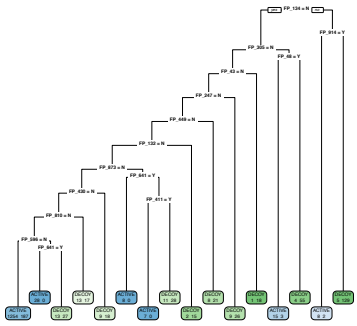
HIVPR_randoms_4_for_analysis

	ACTIVE	DECOY
	1336	206
	59	358



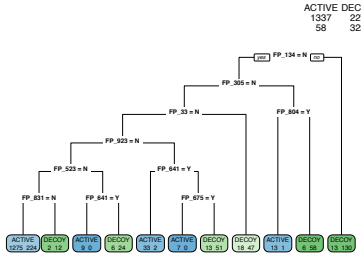
HIVPR_randoms_5_for_analysis

	ACTIVE	DECOY
	1320	192
	75	354



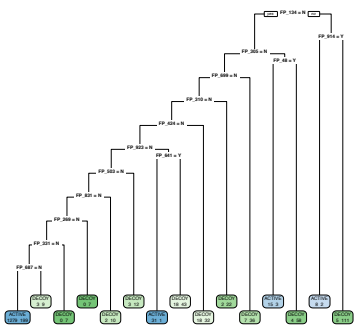
HIVPR_randoms_6_for_analysis

	ACTIVE	DECOY
	1337	227
	58	322



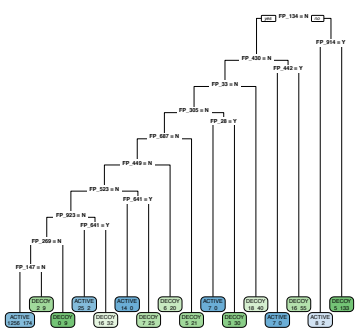
HIVPR_randoms_7_for_analysis

	ACTIVE	DECOY
	1339	205
	62	347



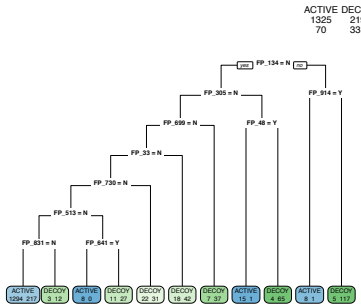
HIVPR_randoms_8_for_analysis

	ACTIVE	DECOY
	1317	178
	78	374



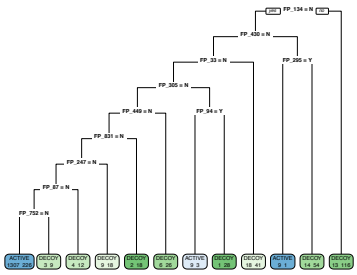
HIVPR_randoms_9_for_analysis

	ACTIVE	DECOY
	1325	219
	70	331



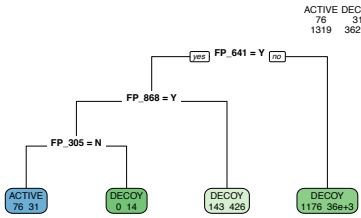
HIVPR_randoms_10_for_analysis

	ACTIVE	DECOY
	1325	230
	70	322



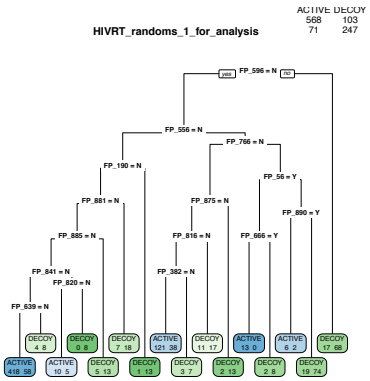
HIVPR_FPs_for_analysis

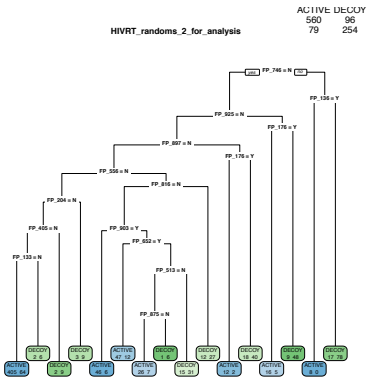
	ACTIVE	DECOY
	76	31
	1319	36214

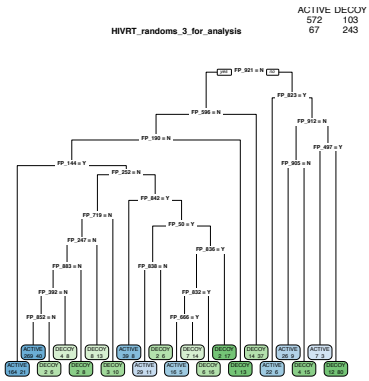


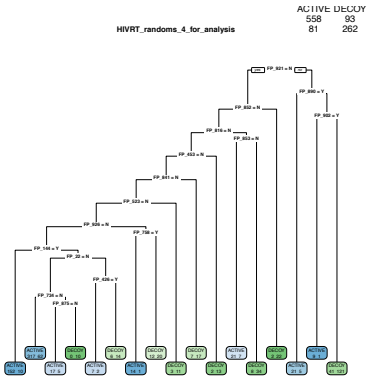
Sample Number	Positive bits	Negative bits	True Actives	True Decoys	False Actives	False Decoys
1	641, 48, 138, 846, 652	134, 430, 33, 305, 883, 288, 551, 331, 890, 537, 513, 449	1325	357	204	70
2	94, 811, 641	33, 305, 730, 699, 582, 513, 331, 275, 831	1340	319	231	55
3	138, 407, 176	134, 305, 33, 852, 728, 361, 513, 394	1312	340	227	83
4	48, 641, 641, 799	134, 305, 33, 430, 852, 699, 77, 826, 817	1336	358	206	59
5	914, 48, 641, 411, 641	134, 305, 43, 247, 449, 132, 873, 430, 810, 596	1320	354	192	75
6	804, 641, 675, 641	134, 305, 33, 923, 523, 831	1337	322	227	58
7	914, 48, 641	134, 305, 699, 310, 424, 923, 503, 831, 269, 331, 687	1333	347	205	62
8	914, 442, 28, 641, 641	134, 430, 33, 305, 687, 449, 523, 923, 269, 147	1317	374	178	78
9	914, 48, 641	134, 305, 699, 33, 730, 513, 831	1325	331	219	70
10	295, 94	134, 430, 33, 305, 449, 831, 247, 87, 752	1325	322	230	70
Average/ consensus	641	33, 134, 305	68.1 %	17.6 %	10.9 %	3.5 %
All	641, 868	305	76 (0.2 %)	36214 (96.2 %)	31 (0.1 %)	1319 (3.5 %)

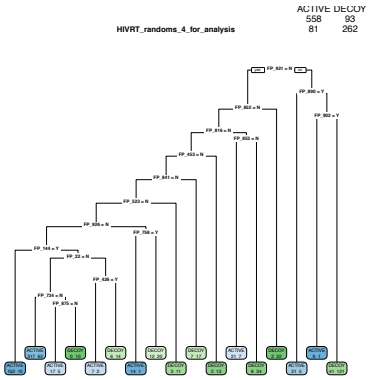
HIVRT

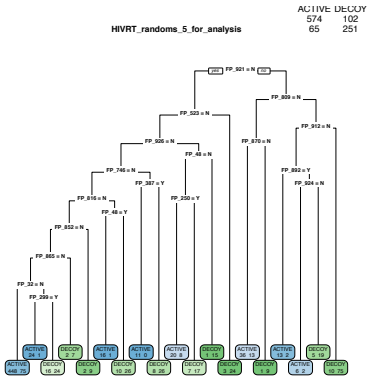

ACTIVE DECOY
568 103
71 247
HIVRT_randoms_1_for_analysis

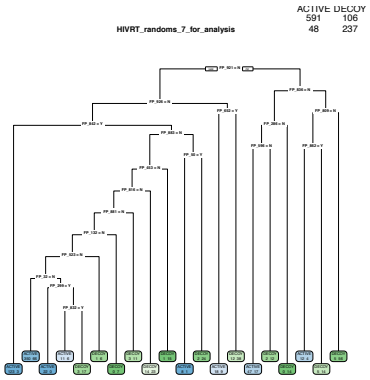

ACTIVE DECOY
560 96
79 254
HIVRT_randoms_2_for_analysis

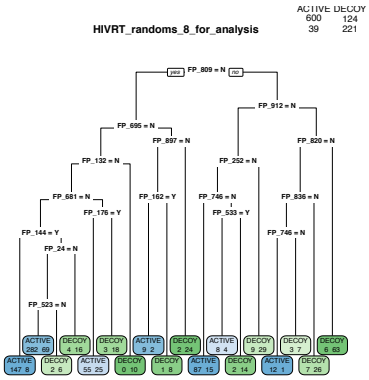

ACTIVE DECOY
572 103
67 243
HIVRT_randoms_3_for_analysis

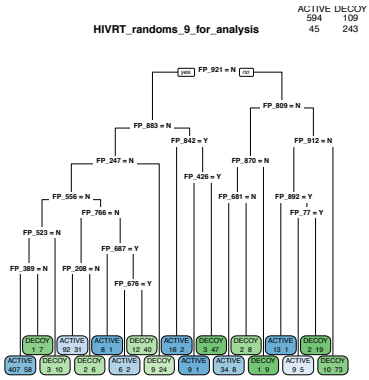

ACTIVE DECOY
558 93
81 262
HIVRT_randoms_4_for_analysis

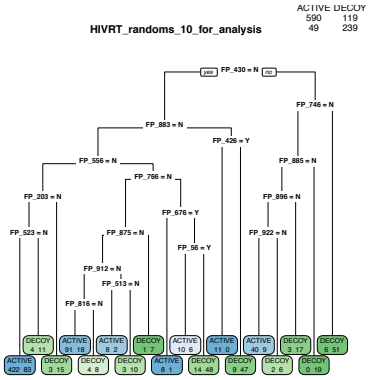

ACTIVE DECOY
558 93
81 262
HIVRT_randoms_4_for_analysis

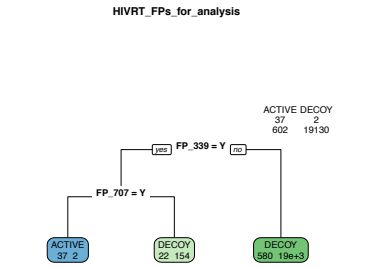

ACTIVE DECOY
574 102
65 251
HIVRT_randoms_5_for_analysis


ACTIVE DECOY
591 106
48 237
HIVRT_randoms_7_for_analysis


ACTIVE DECOY
600 124
39 221
HIVRT_randoms_8_for_analysis


ACTIVE DECOY
594 109
45 243
HIVRT_randoms_9_for_analysis

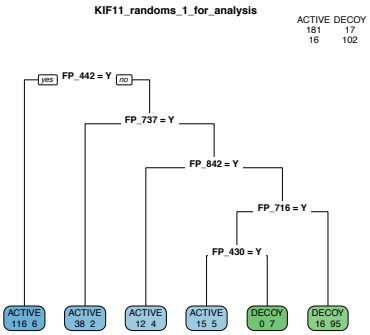

ACTIVE DECOY
590 119
49 239
HIVRT_randoms_10_for_analysis

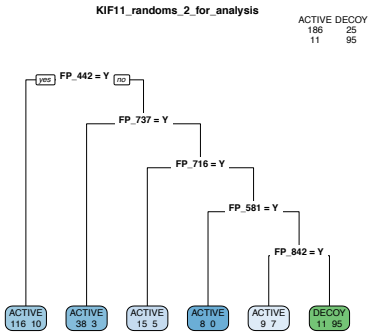

HIVRT_FPs_for_analysis
ACTIVE DECOY
37 2
602 19130
FP_339 = Y
yes
no
FP_707 = Y
ACTIVE
37 2
DECOY
22 154
DECOY
580 19e+3

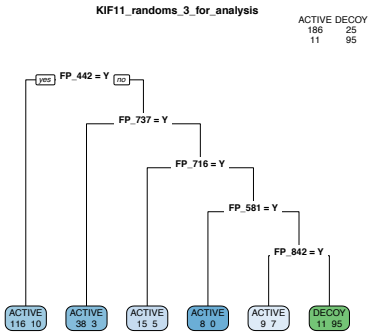
Sample Number	Positive bits	Negative bits	True Actives	True Decoys	False Actives	False Decoys
1	56, 890, 666	596, 556, 766, 875, 816, 382, 190, 881, 885, 841, 820, 639	568	247	103	71
2	136, 176, 176, 903, 652	746, 925, 897, 556, 816, 204, 405, 133, 513, 875	560	254	96	79
3	823, 497, 144, 842, 50, 836, 832, 666	921, 912, 905, 596, 190, 252, 838, 719, 247, 883, 392, 852	572	243	103	67
4	890, 902, 758, 144, 426	921, 852, 816, 853, 453, 841, 523, 926, 22, 734, 875	558	262	93	81
5	892, 387, 250, 48, 299	921, 809, 912, 924, 870, 523, 926, 48, 746, 816, 852, 865, 32	574	251	102	65
6	823, 886, 690, 176, 687, 917, 138, 48	921, 912, 886, 445, 286, 885, 369, 681, 269, 37, 816	590	221	132	49
7	862, 652, 842, 50, 299, 832	921, 836, 809, 286, 596, 926, 883, 453, 816, 881, 132, 523, 32	591	237	106	48
8	533, 162, 176, 144	809, 912, 820, 836, 746, 252, 746, 695, 897, 132, 681, 24, 523	600	221	124	39
9	892, 77, 842, 426, 687, 676	921, 809, 912, 870, 681, 883, 247, 556, 766, 208, 523, 389	594	243	109	45
10	426, 676, 56	430, 746, 885, 896, 922, 883, 766, 875, 912, 513, 816, 556, 203, 523	590	239	119	49

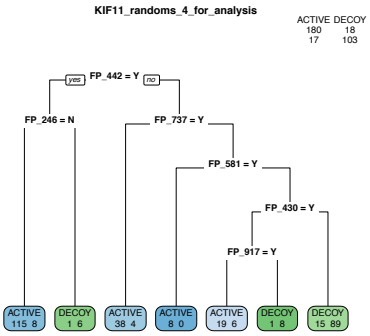
Average/ consensus		523, 816, 912, 921	58.6 %	24.4 %	11.0 %	6.0 %
All	339, 707		37 (0.2 %)	19130 (96.8 %)	2 (0.0 %)	602 (3.0 %)

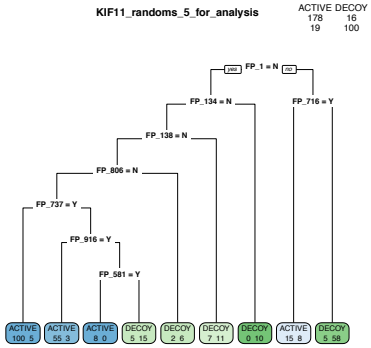
KIF11

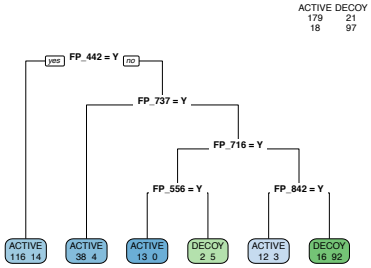


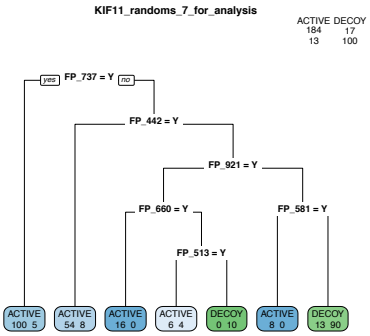


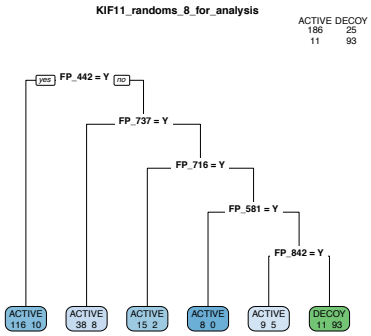


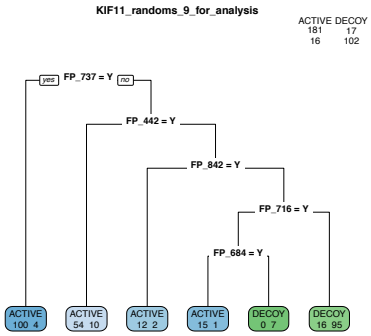


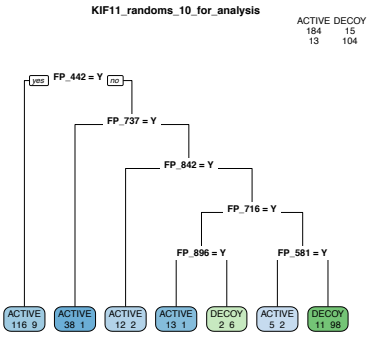


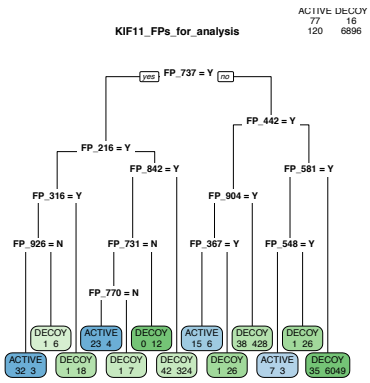












Sample Number	Positive bits	Negative bits	True Actives	True Decoys	False Actives	False Decoys

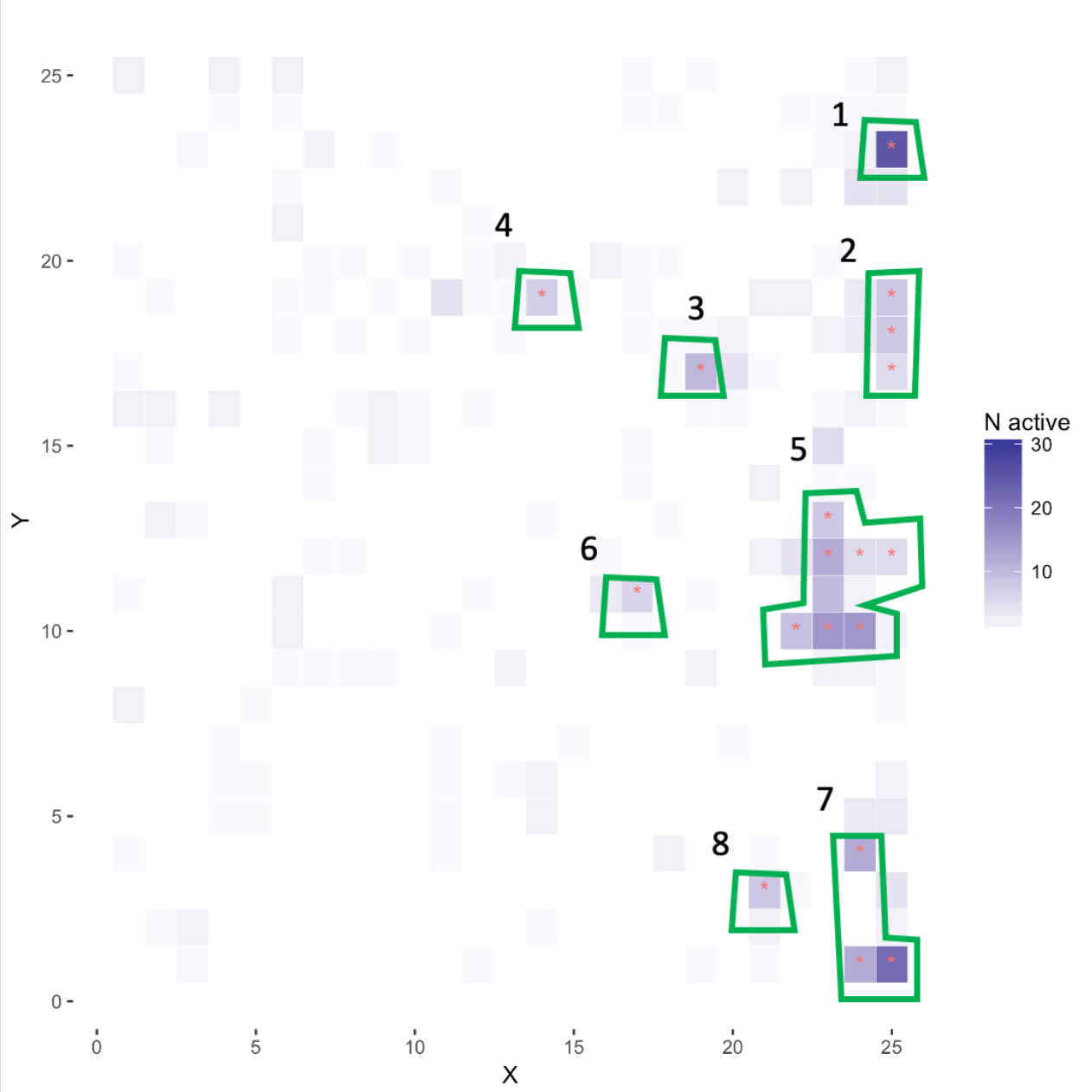
1	442, 737, 842, 716, 430		181	102	17	16
2	442, 737, 716, 581, 842		186	95	25	11
3	442, 737, 716, 581, 842		186	95	25	11
4	442, 737, 581, 430, 917	246	180	103	18	17
5	716, 737, 916, 581	1, 134, 138, 806	178	100	16	19
6	442, 737, 716, 556, 842		179	97	21	18
7	737, 442, 921, 660, 513, 581		184	100	17	13
8	442, 737, 716, 581, 842		186	93	25	11
9	737, 442, 842, 716, 684		181	102	17	16
10	442, 737, 842, 716, 581, 896		184	104	15	13
Average/ consensus	442, 581, 716, 737, 842		57.8 %	31.4 %	6.2 %	4.6 %
All	737, 442, 581, 548, 904, 367, 216, 842, 316	731, 770, 926	77 (1.1 %)	6896 (97.0 %)	16 (0.2 %)	120 (1.7 %)

Figure S19. Decision trees and summaries for the DUD-E diversity set targets.

S7) Chemical structures of compounds at each enriched coordinate in the shape maps created using the shape fingerprints

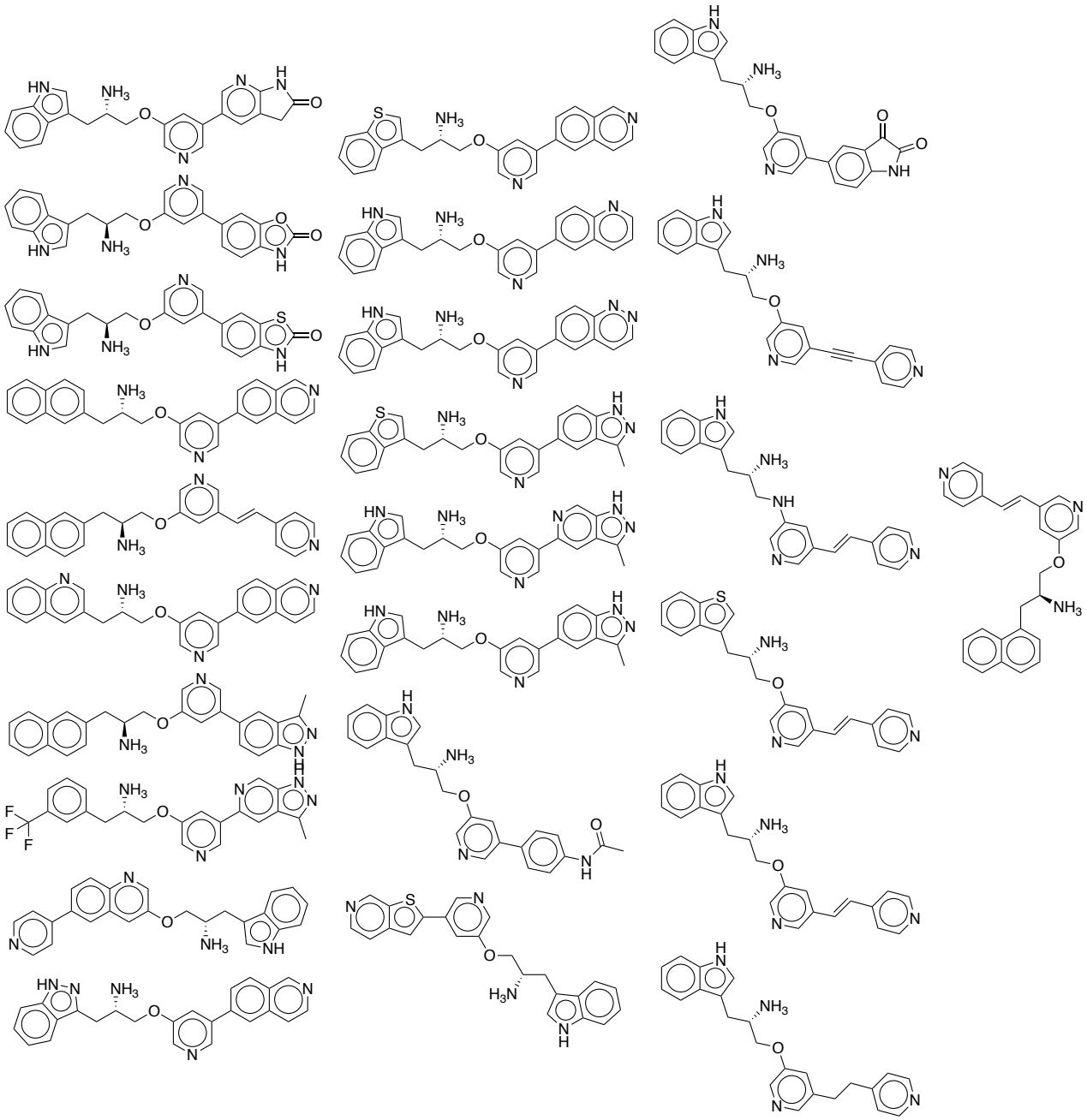
In this section, the SOM presented in the main text is analysed to identify the active compounds that are present at each enriched coordinate. Those that are close to another are grouped into numbered sets of enriched coordinated.

a) AKT

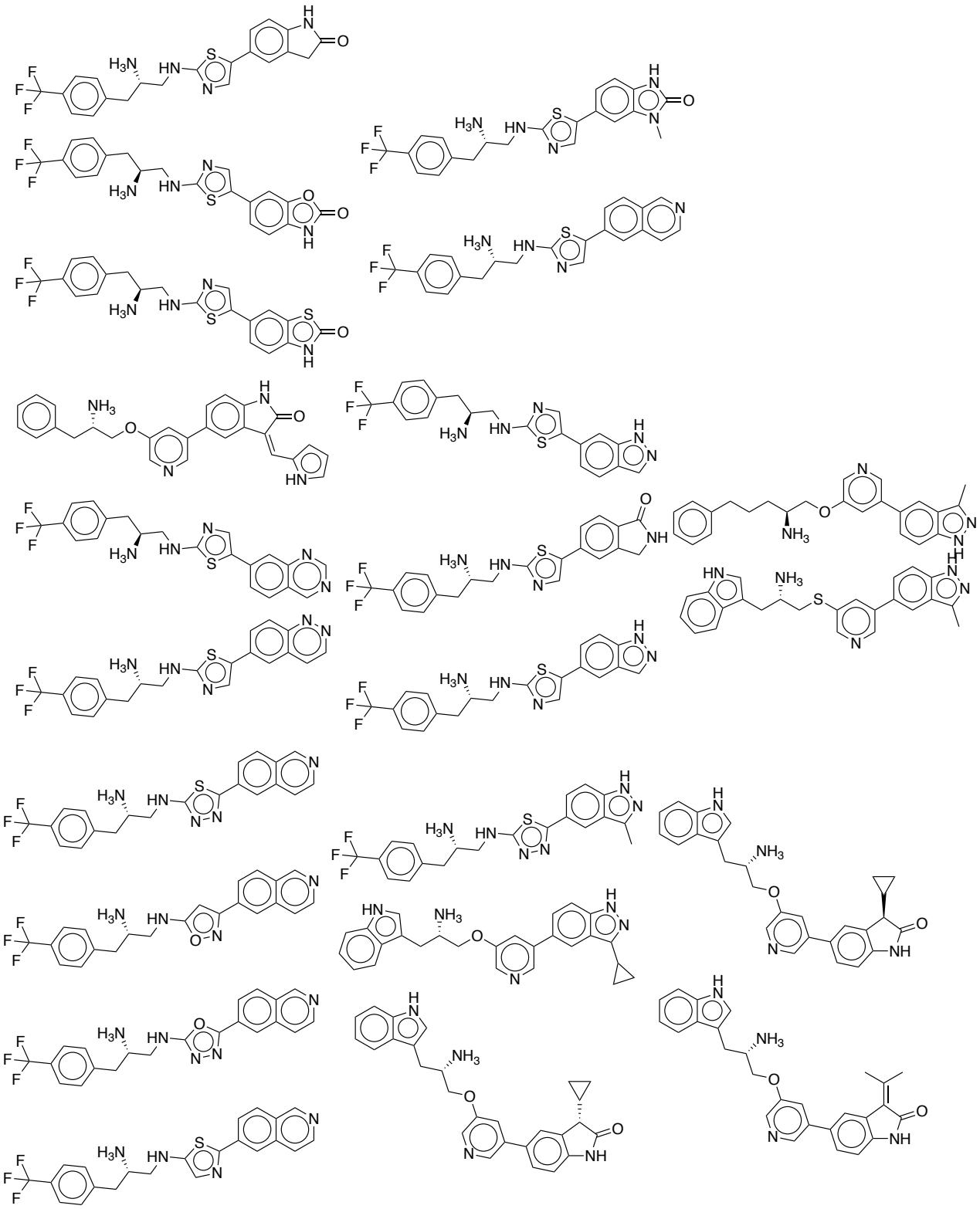


The structures of compounds in each numbered cluster are as follows:

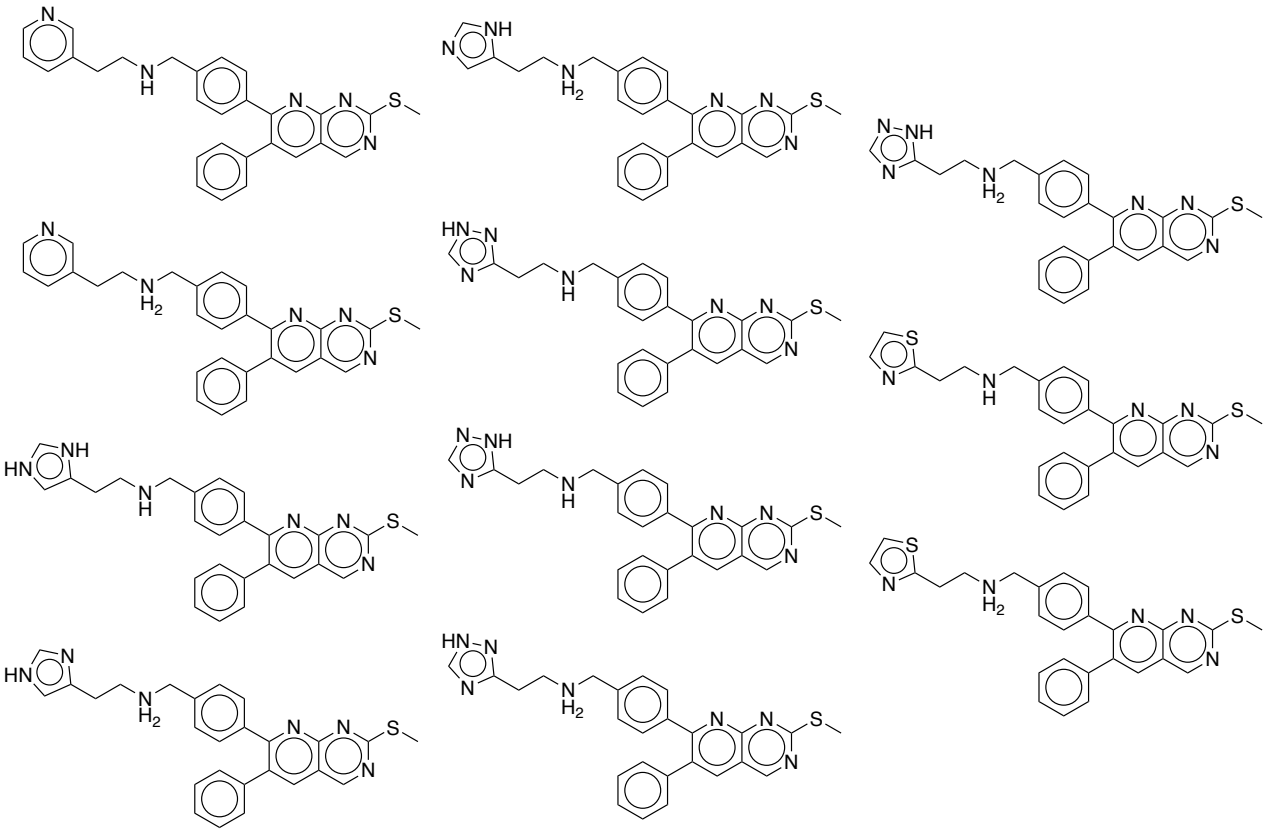
1)



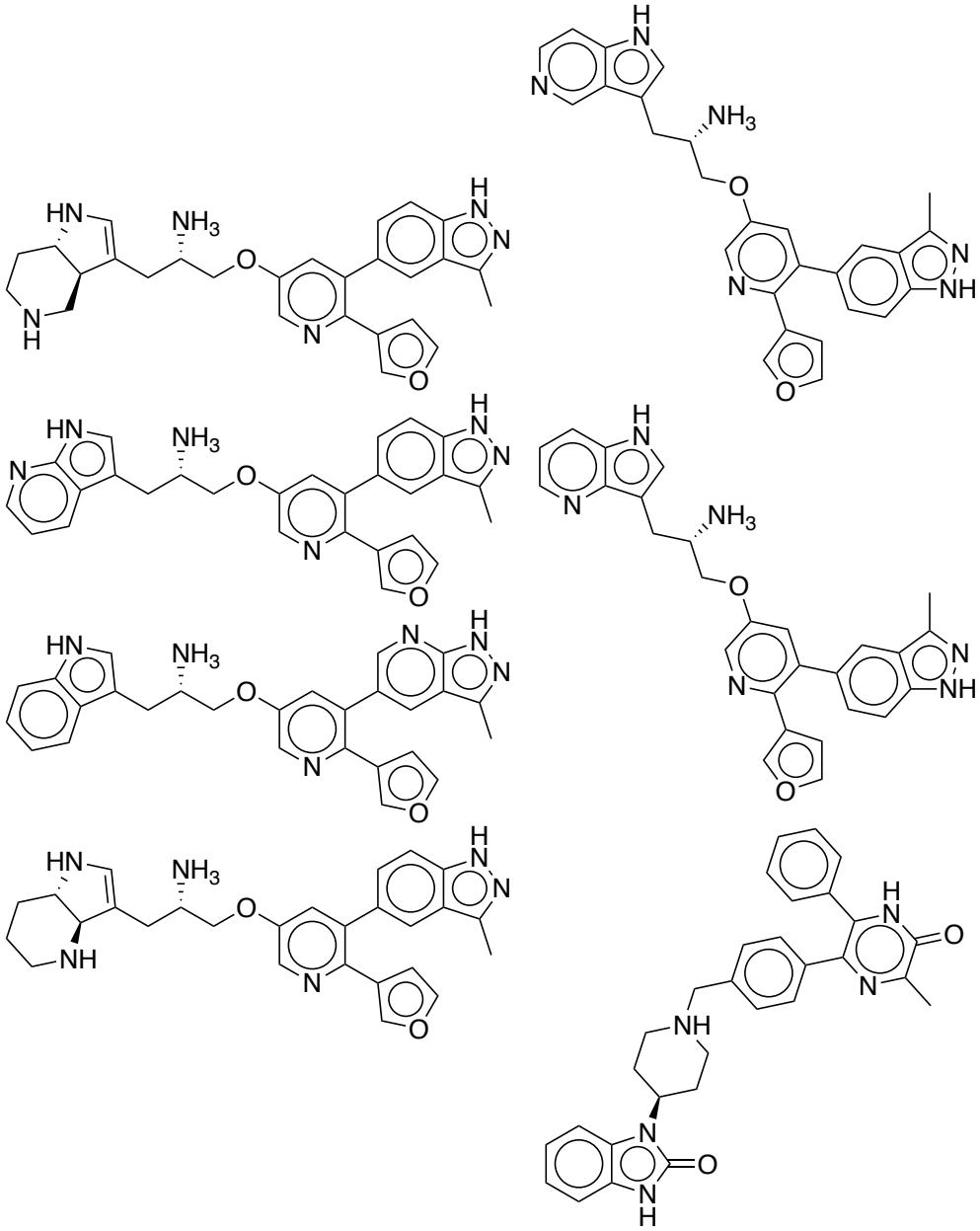
2)



3)

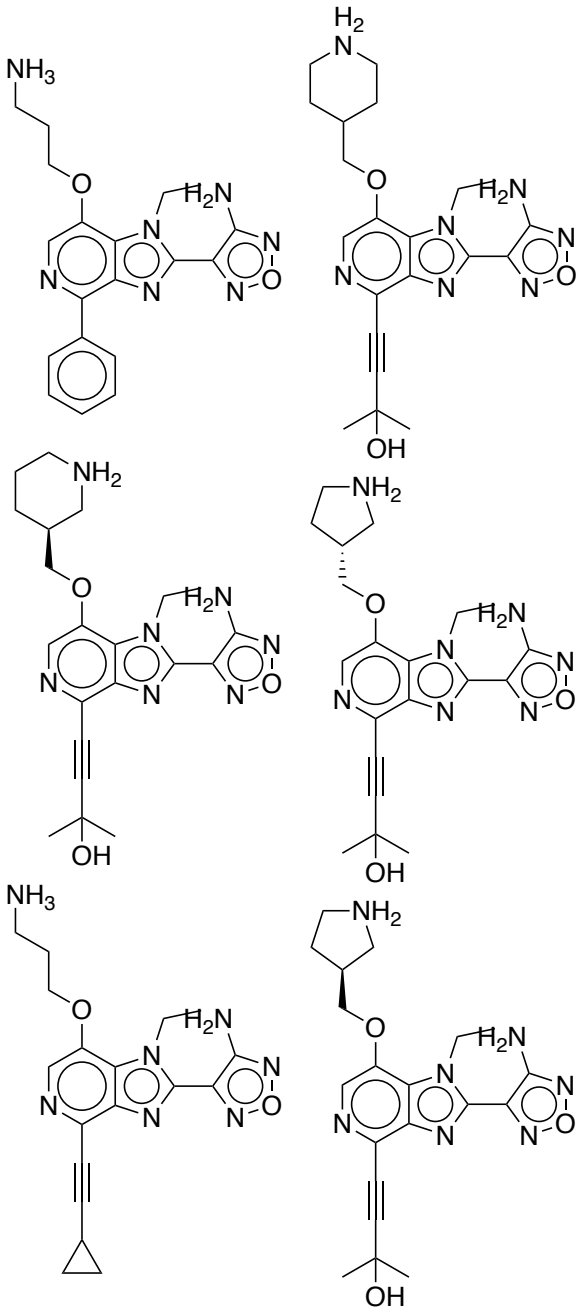


4)

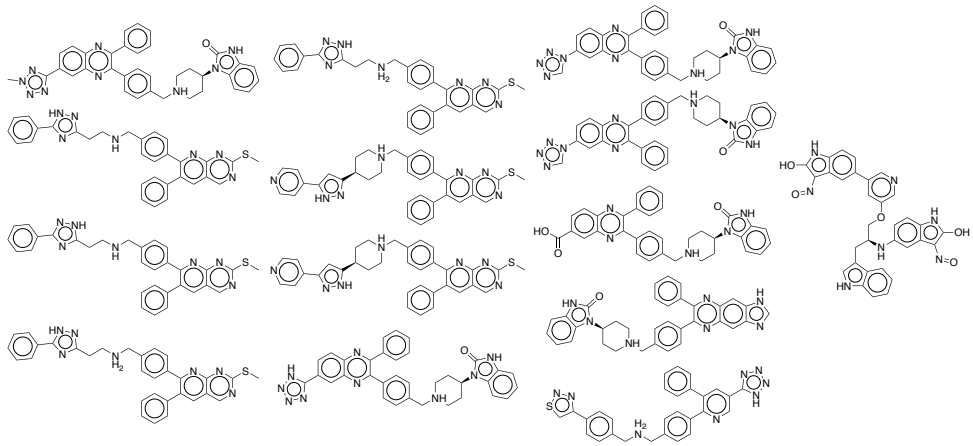


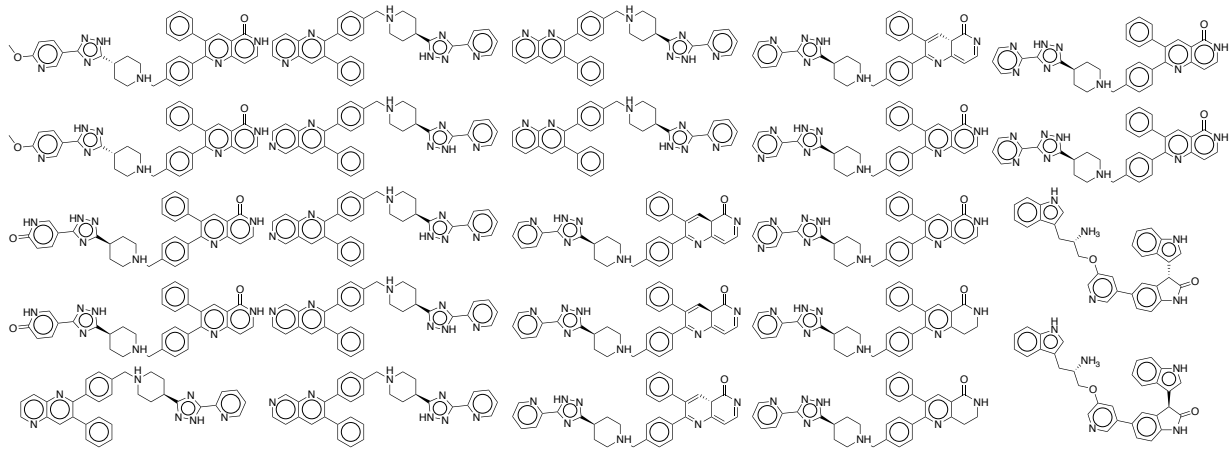
5)

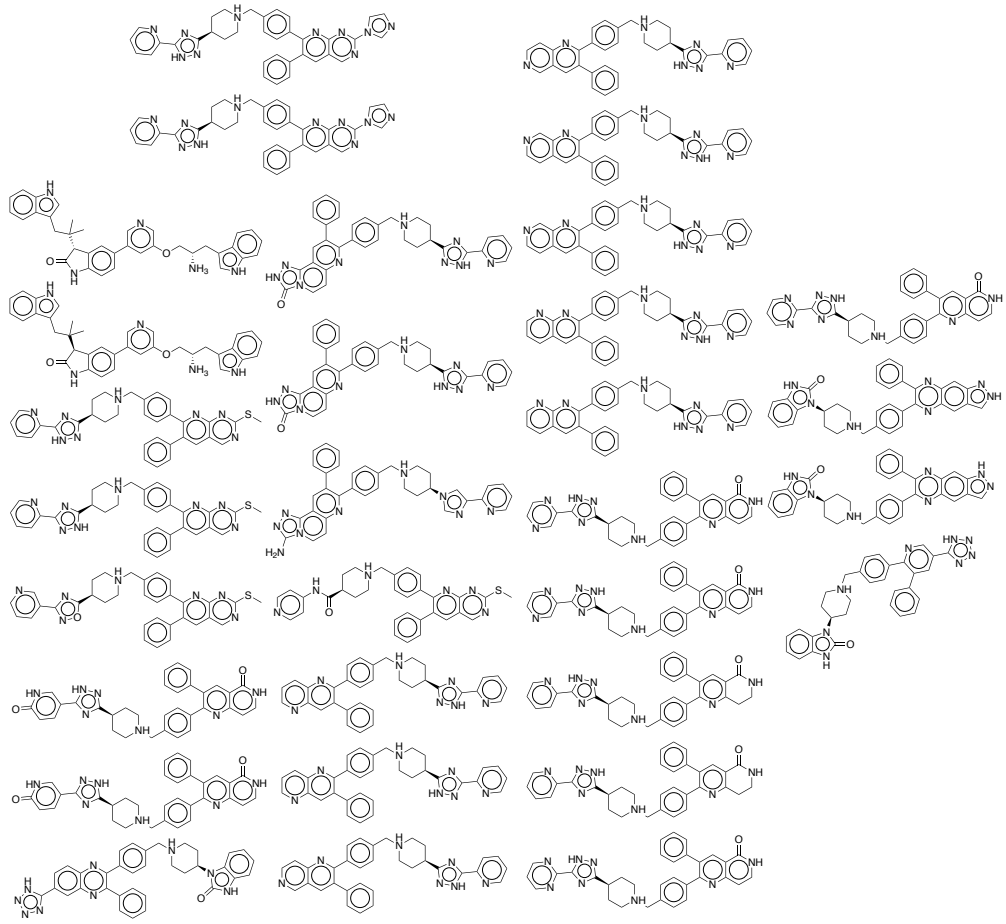
6)



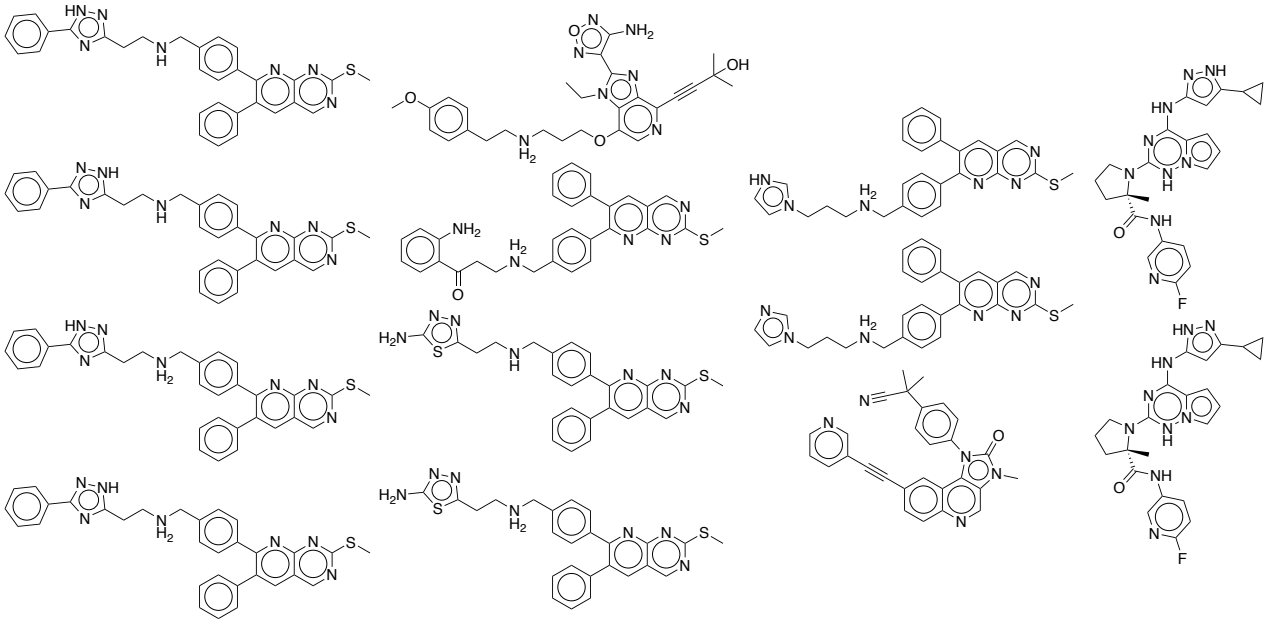
7)



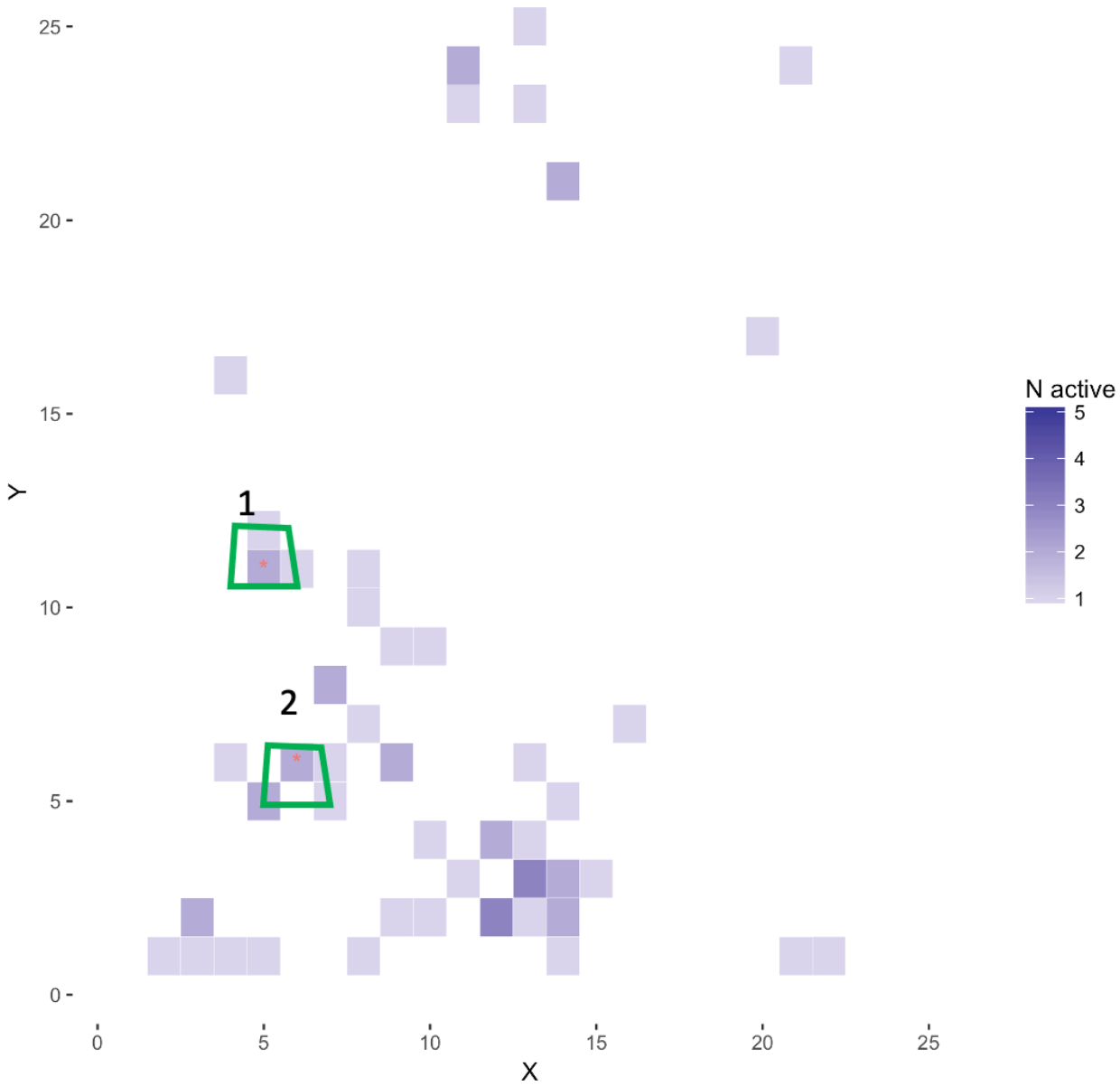




8)

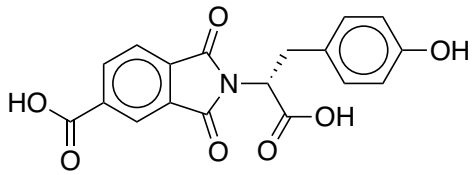


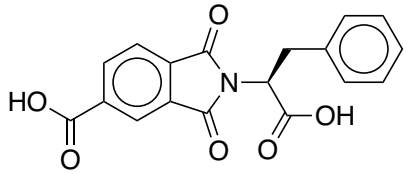
b) AMPC

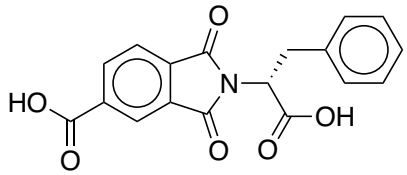


The structures of compounds in each numbered cluster are as follows:

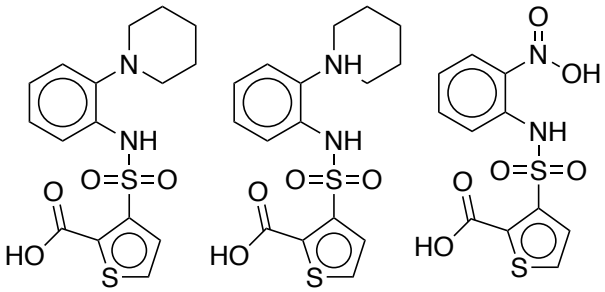
1)


O
OH
N
HO
OH
O
O
O

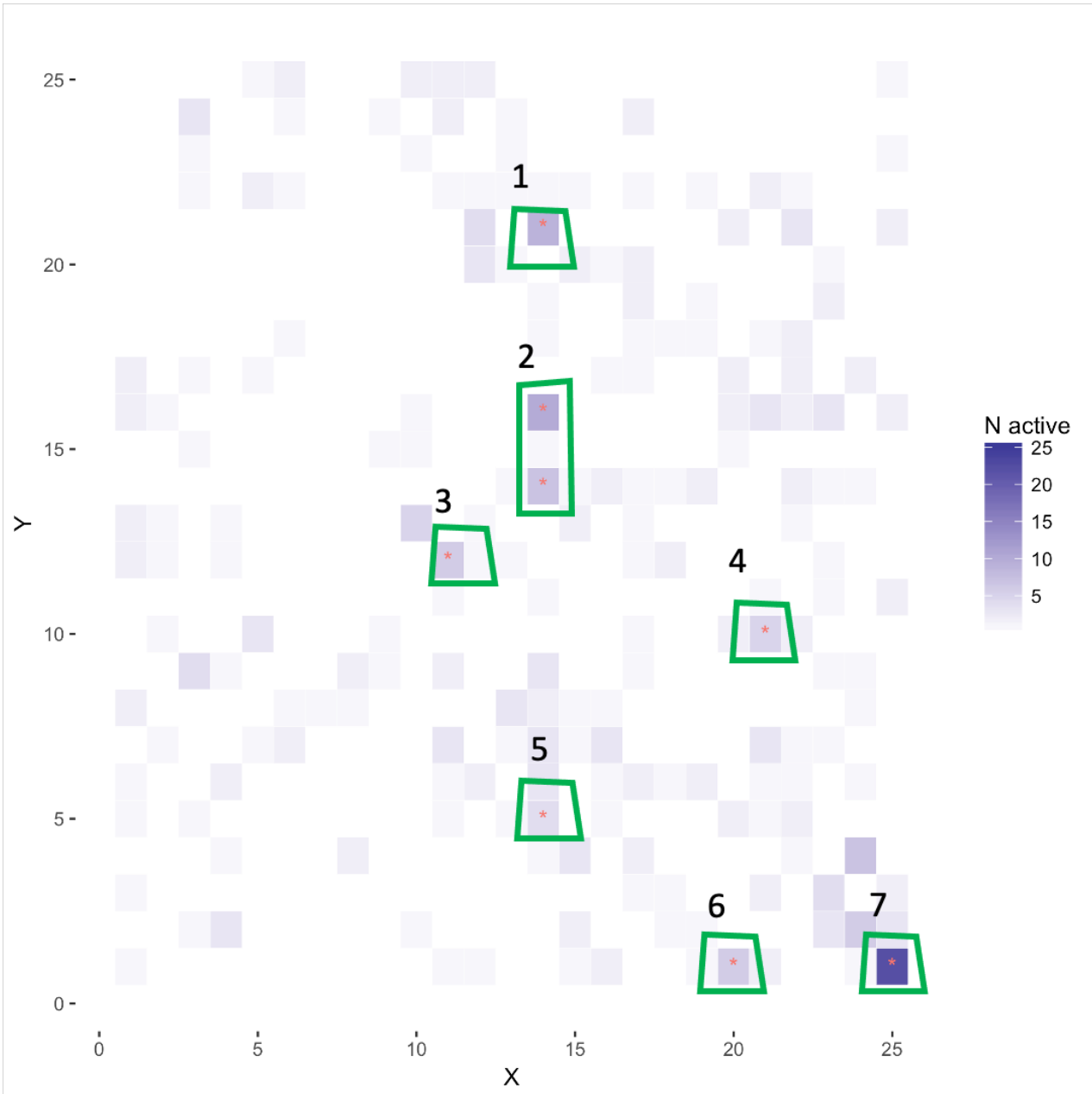

O
N
HO
OH
O
O
O


O
N
HO
OH
O
O
O

2)

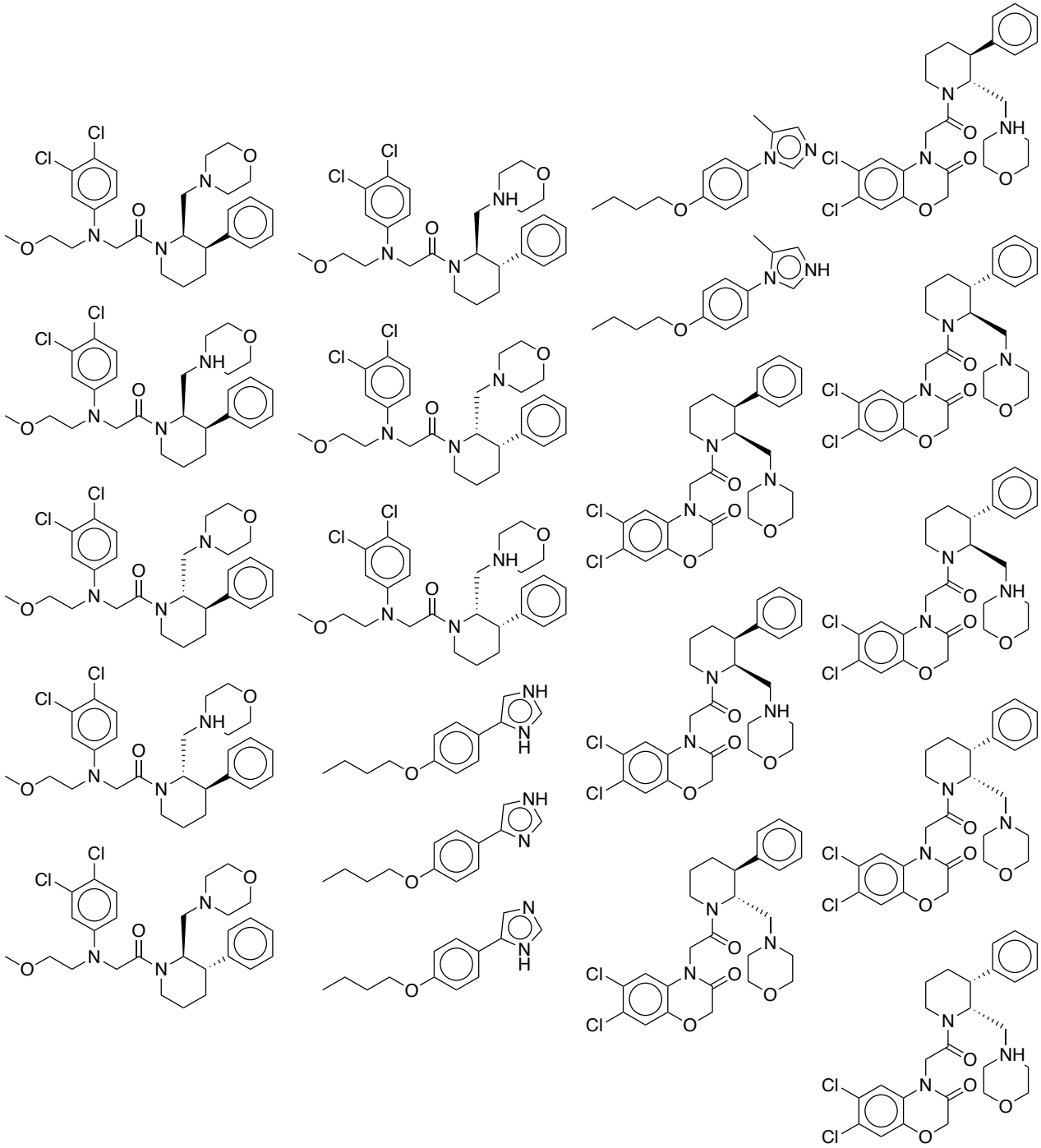

N
NH
O=S=O
O
HO
S
NH
NH
O=S=O
O
HO
S
O
N
OH
NH
O=S=O
O
HO
S

c) CP3A4

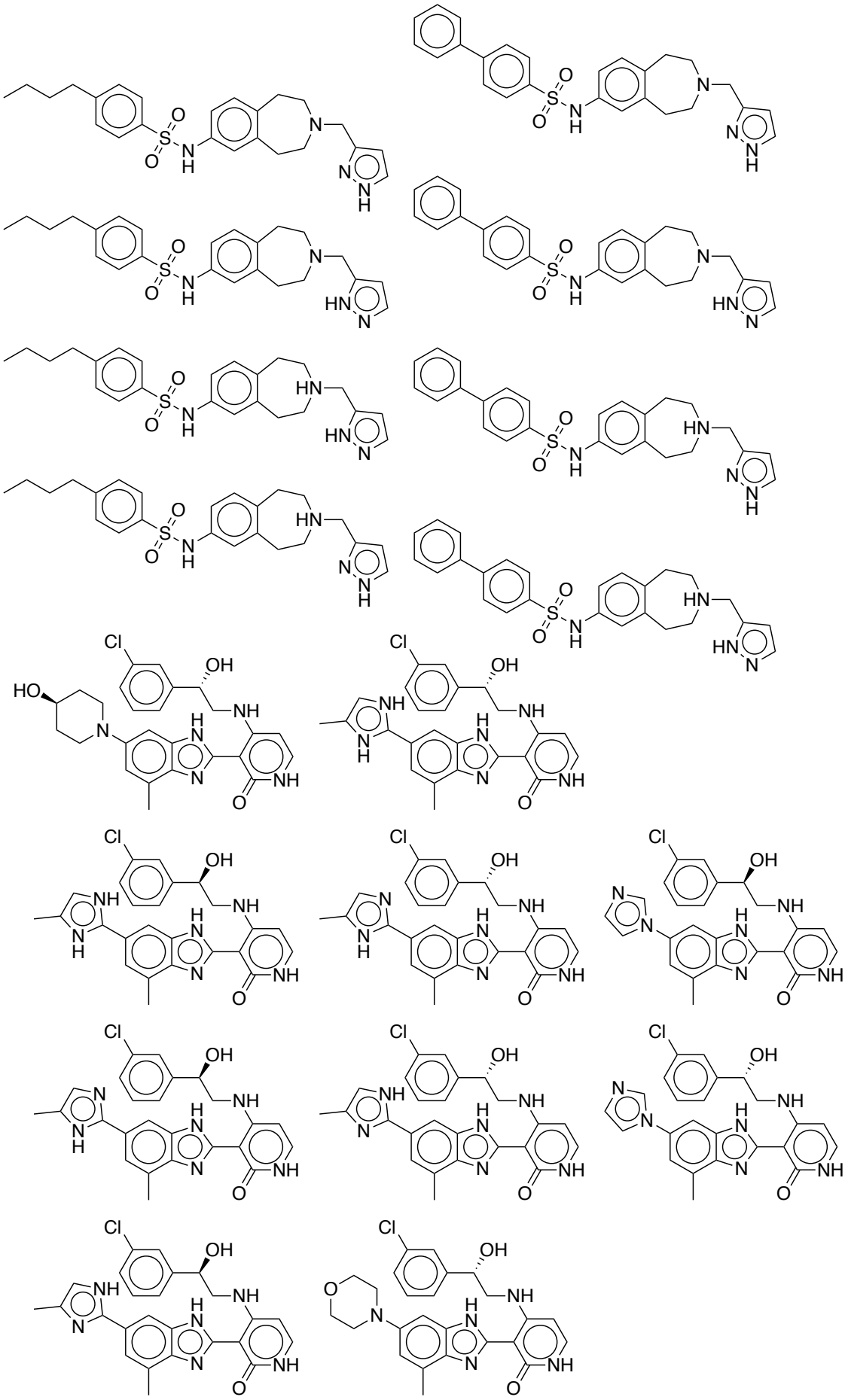


The structures of compounds in each numbered cluster are as follows:

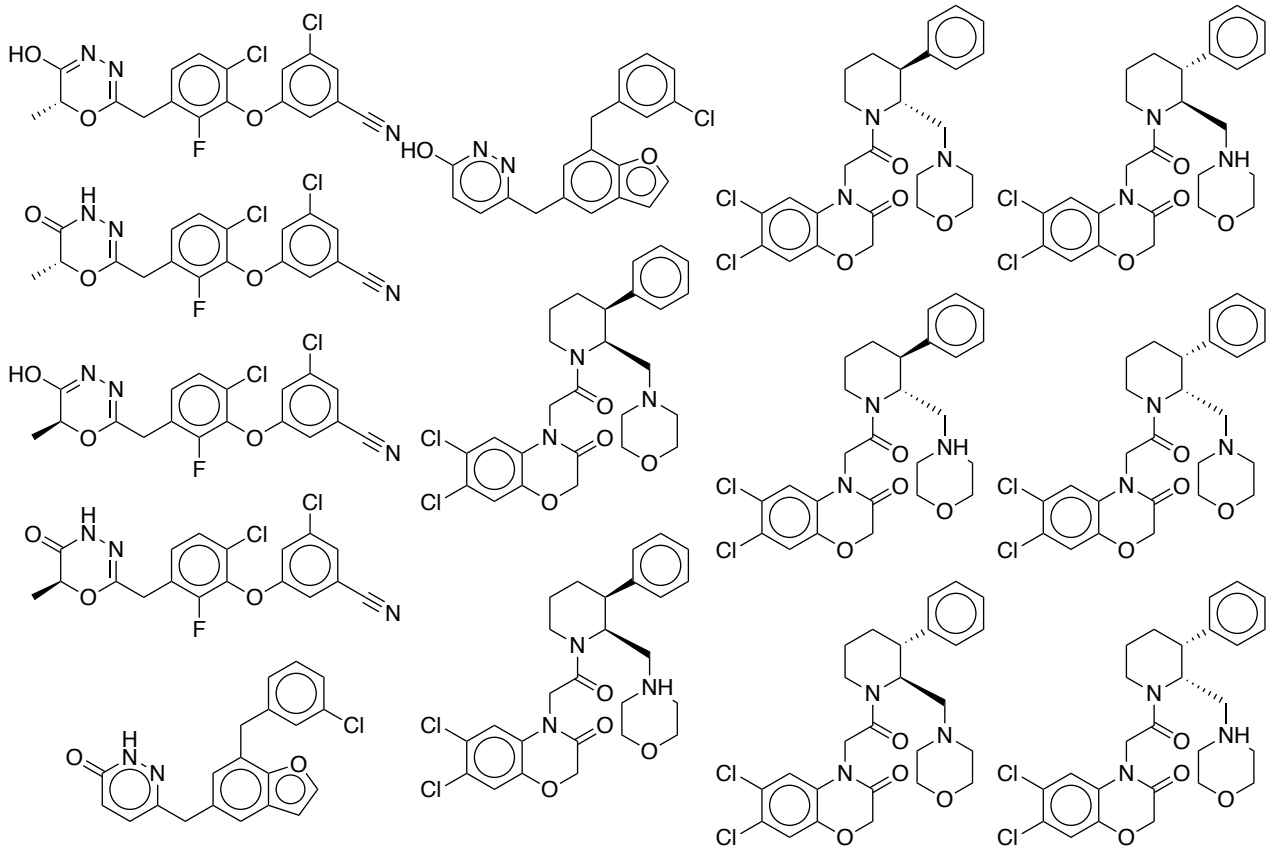
1)



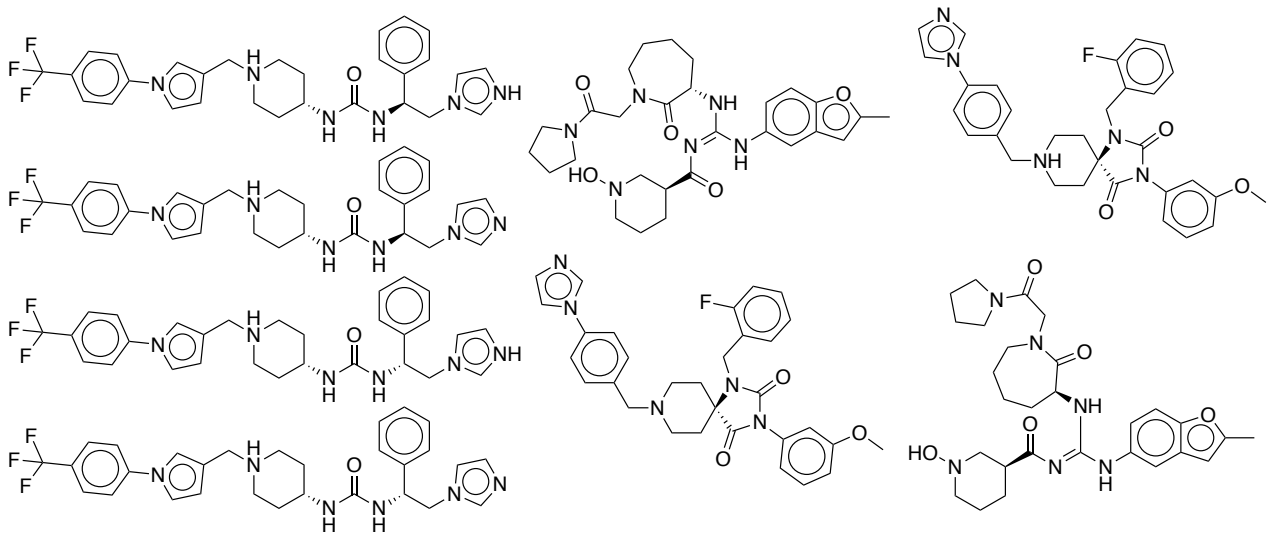
2)



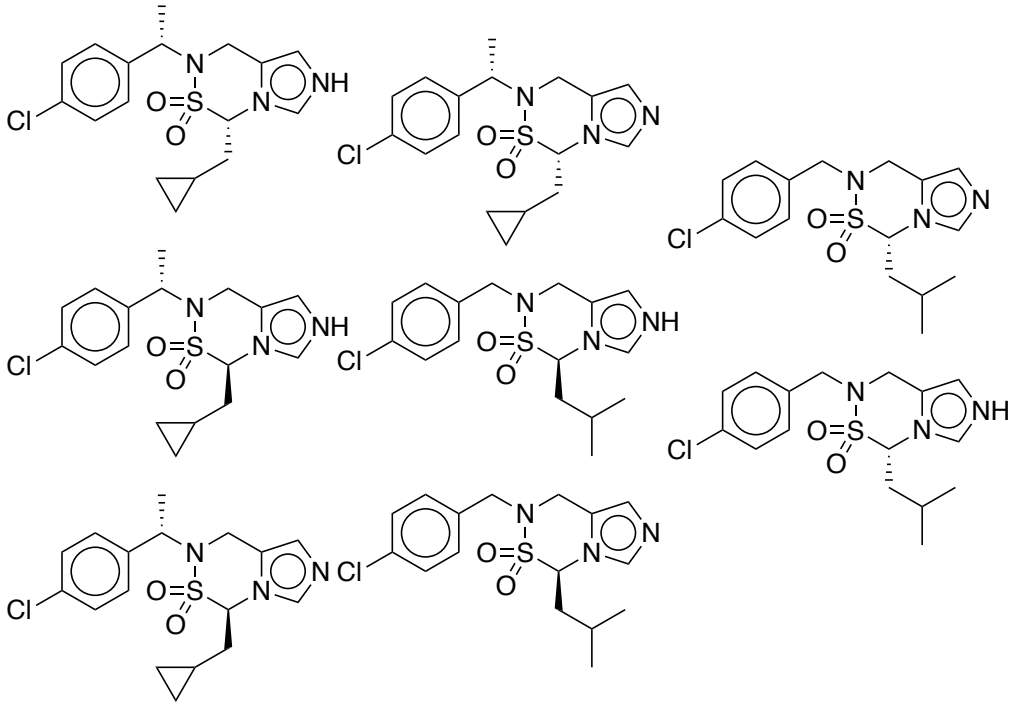
3)



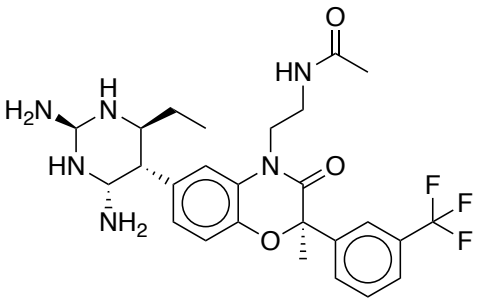
4)

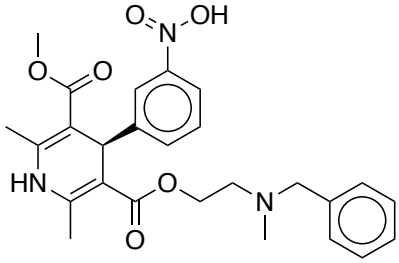


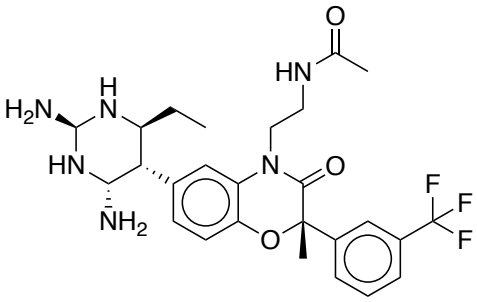
5)

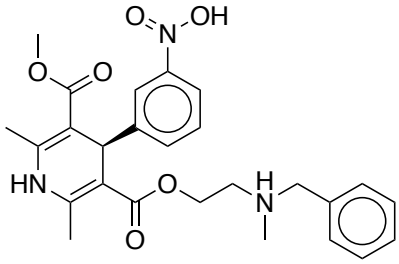


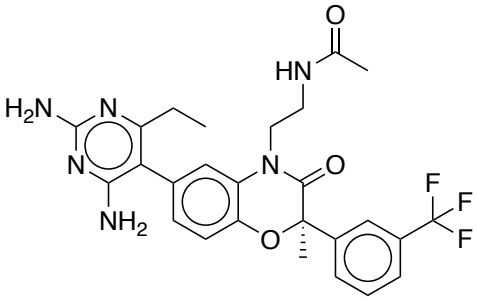
6)

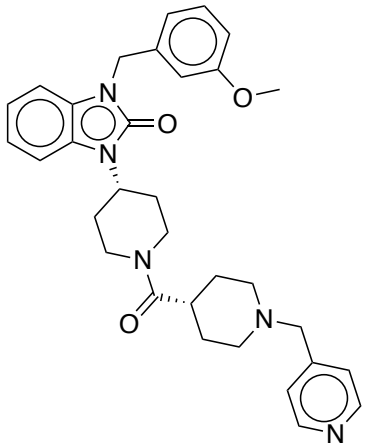


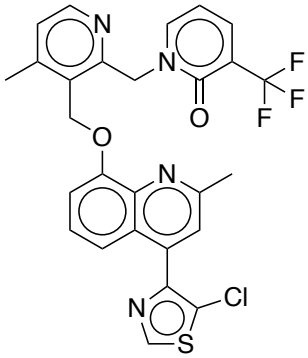


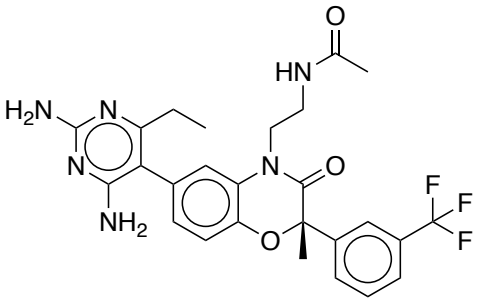


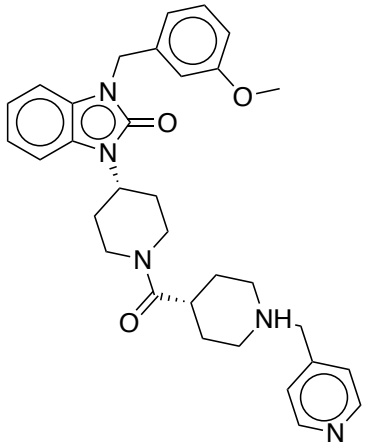


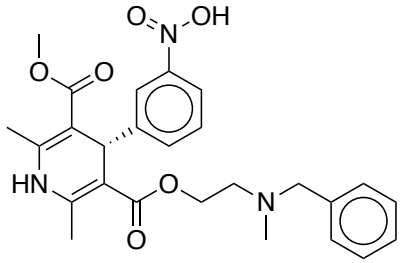


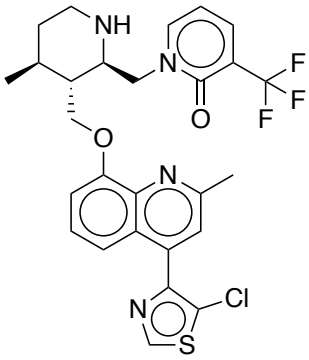


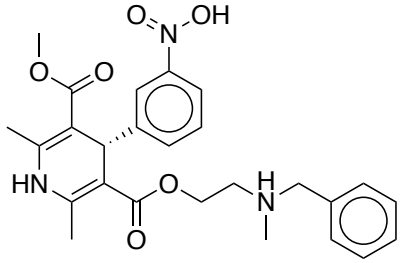




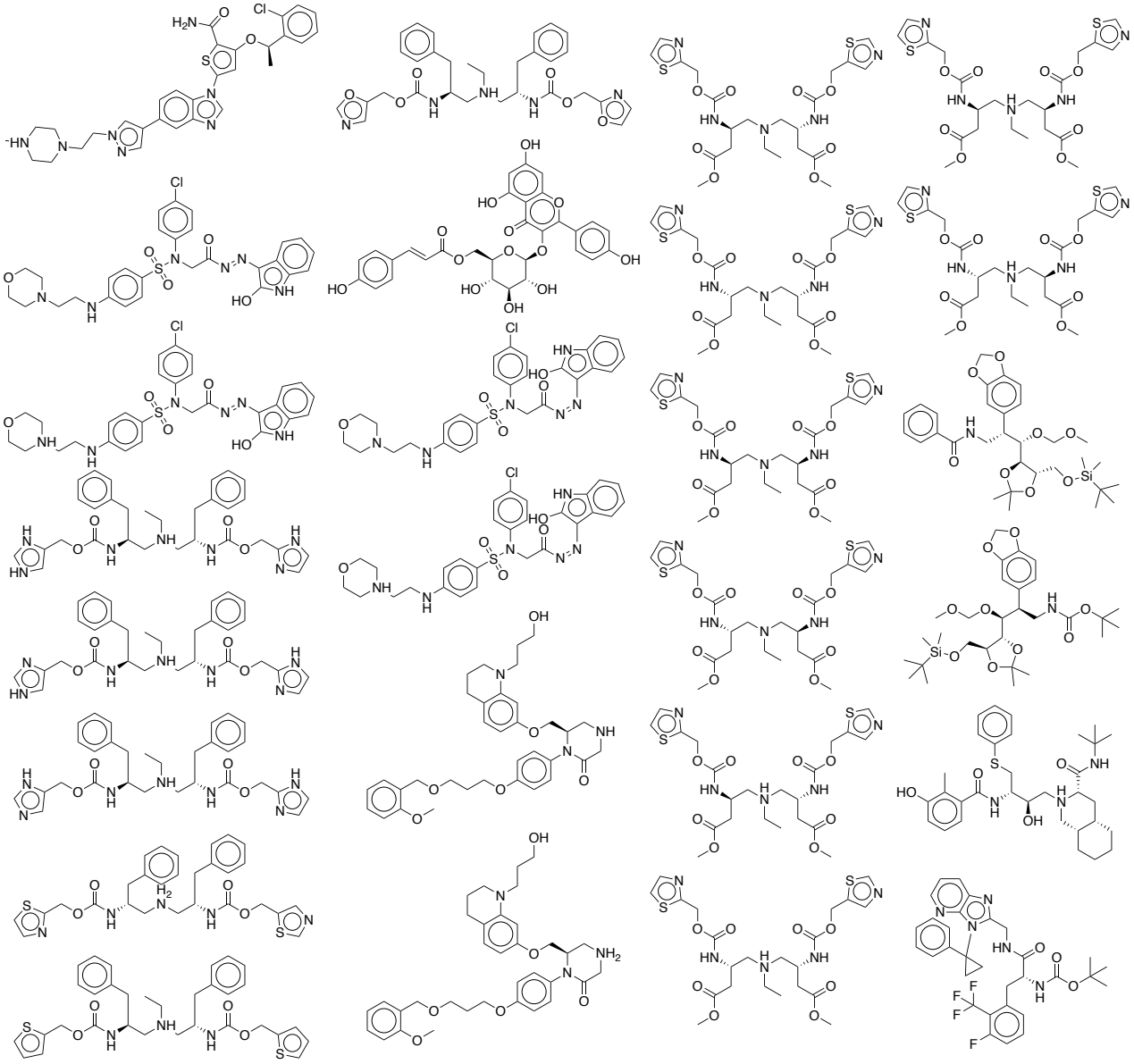




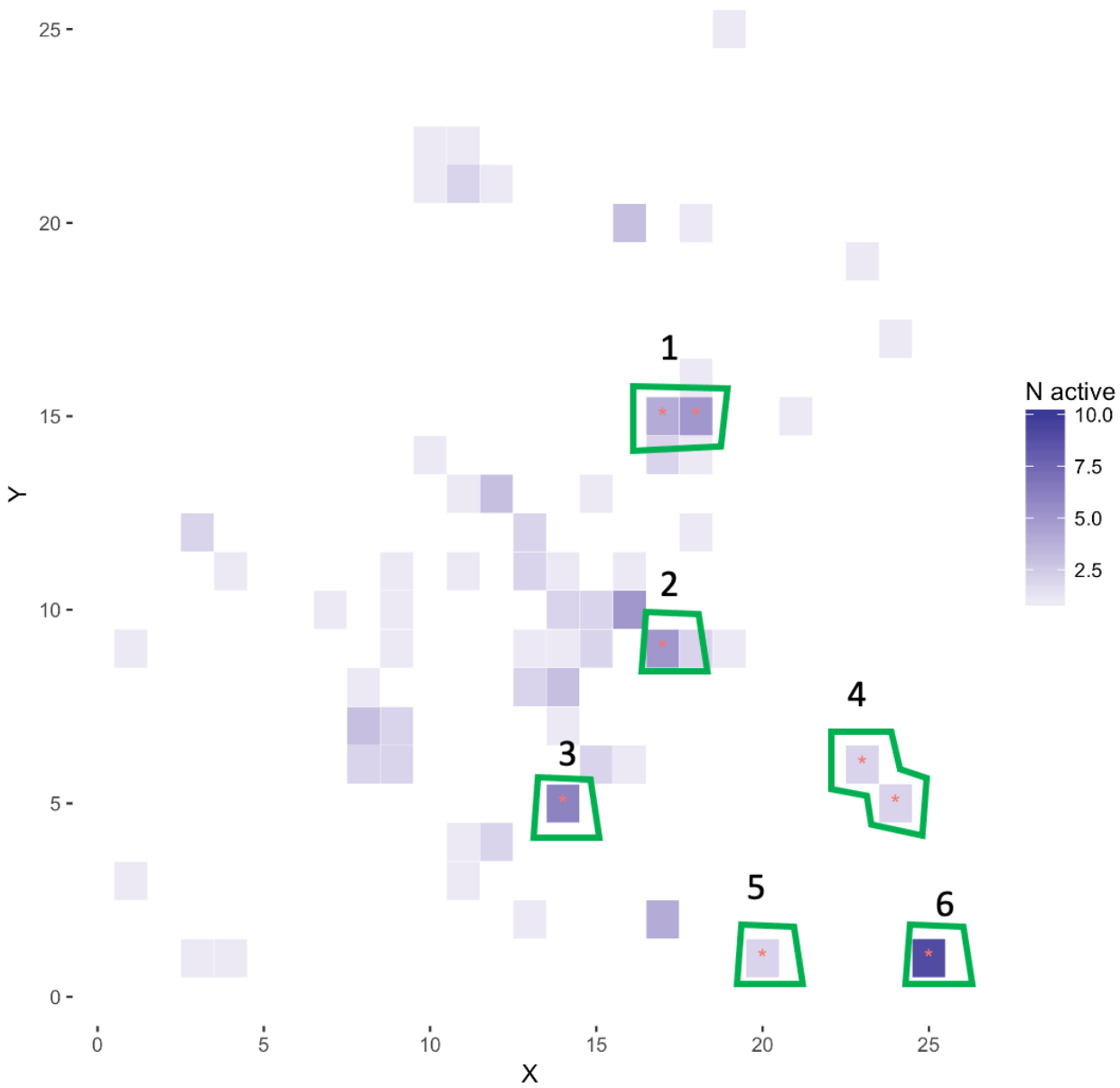




7)

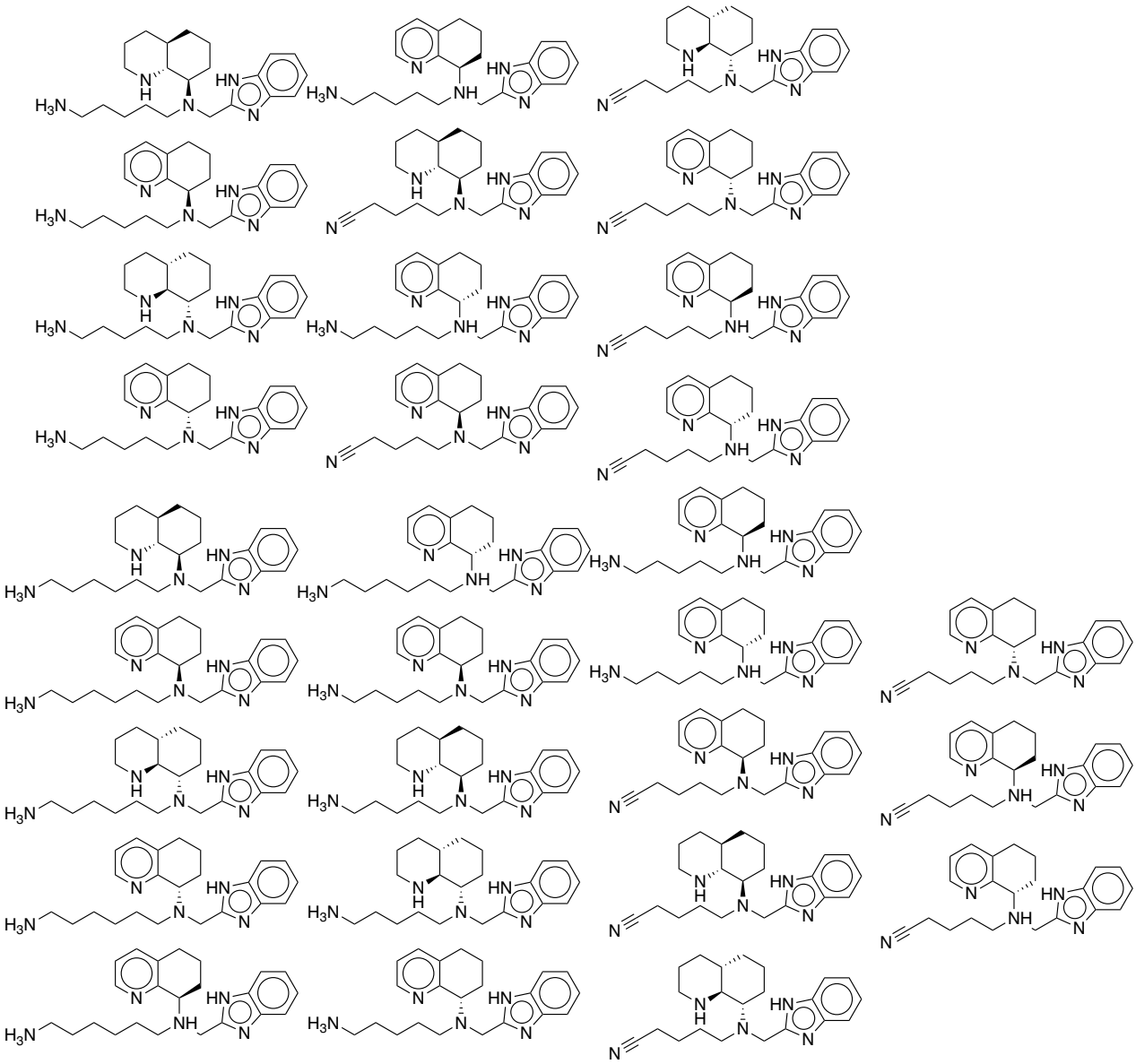


d) CXCR4

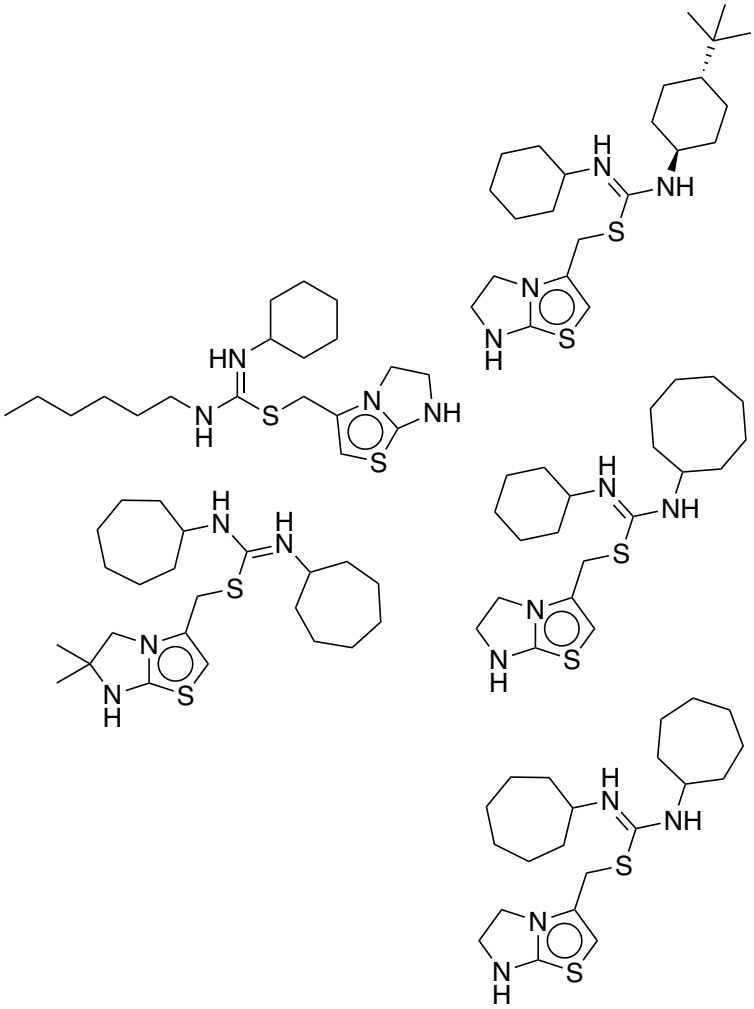


The structures of compounds in each numbered cluster are as follows:

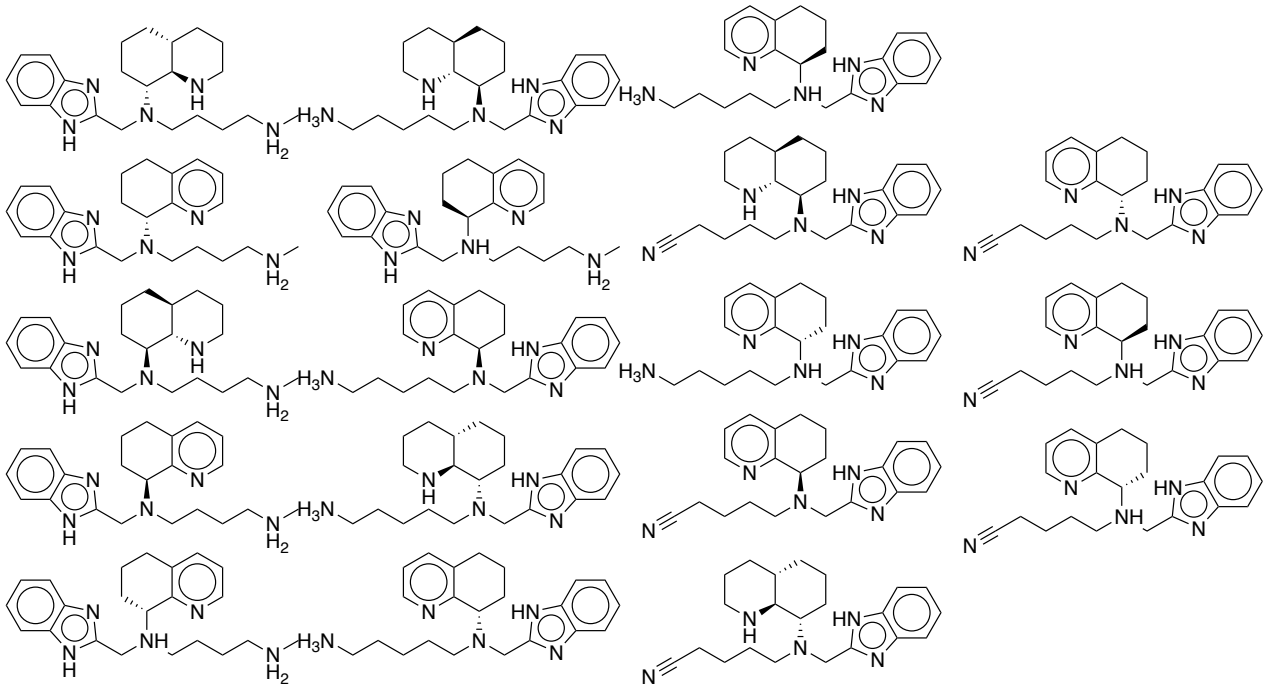
1)



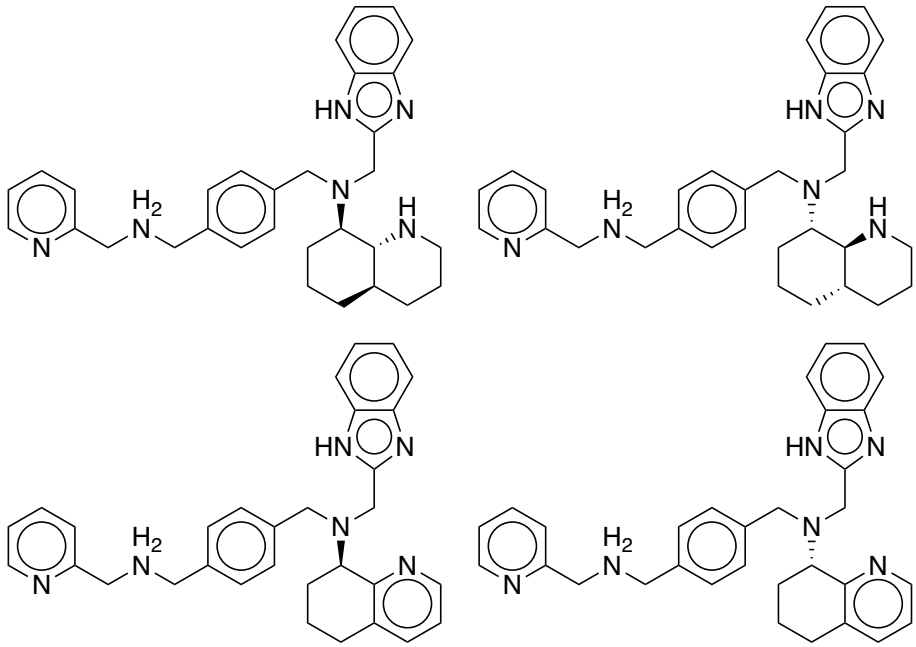
2)



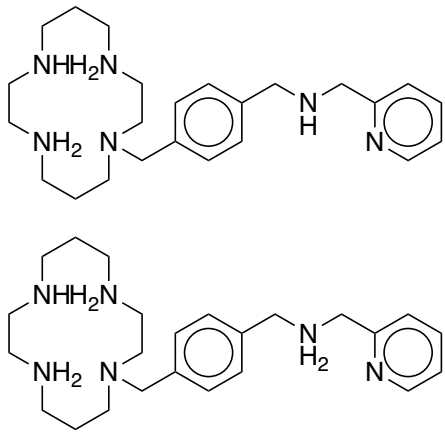
3)



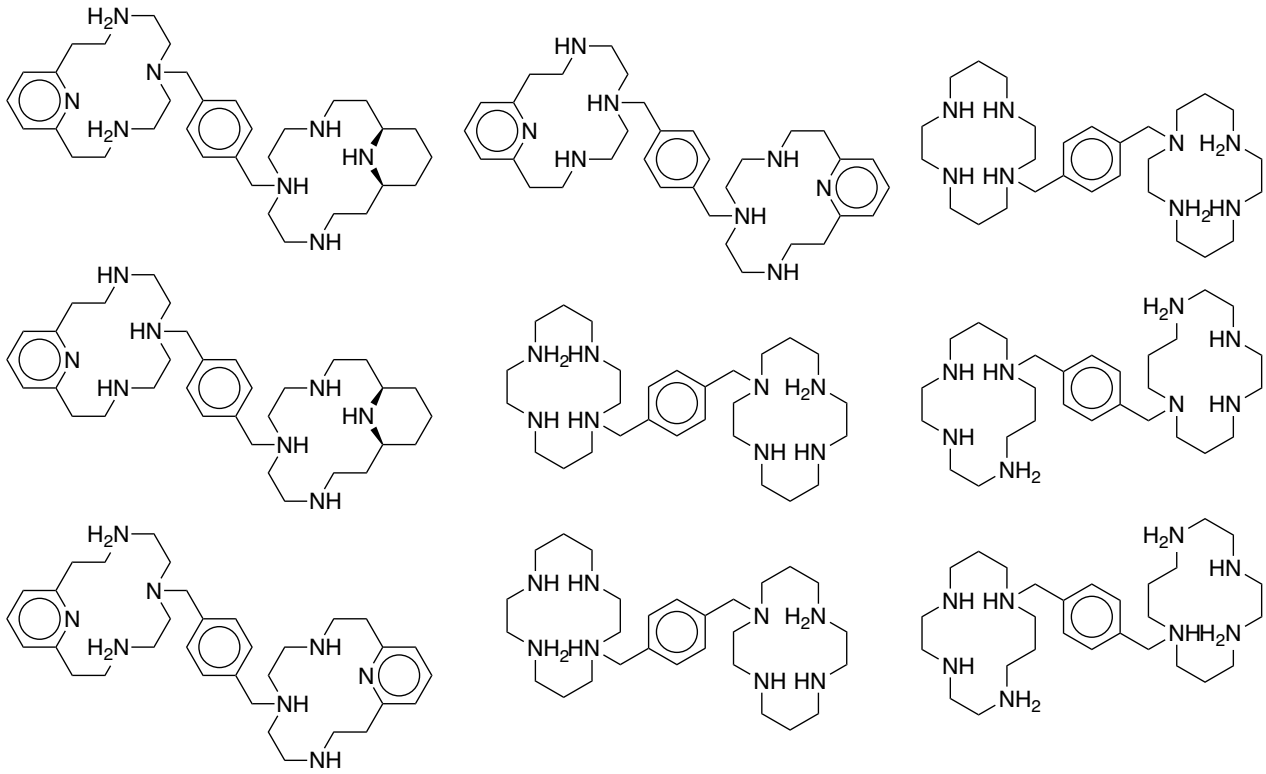
4)



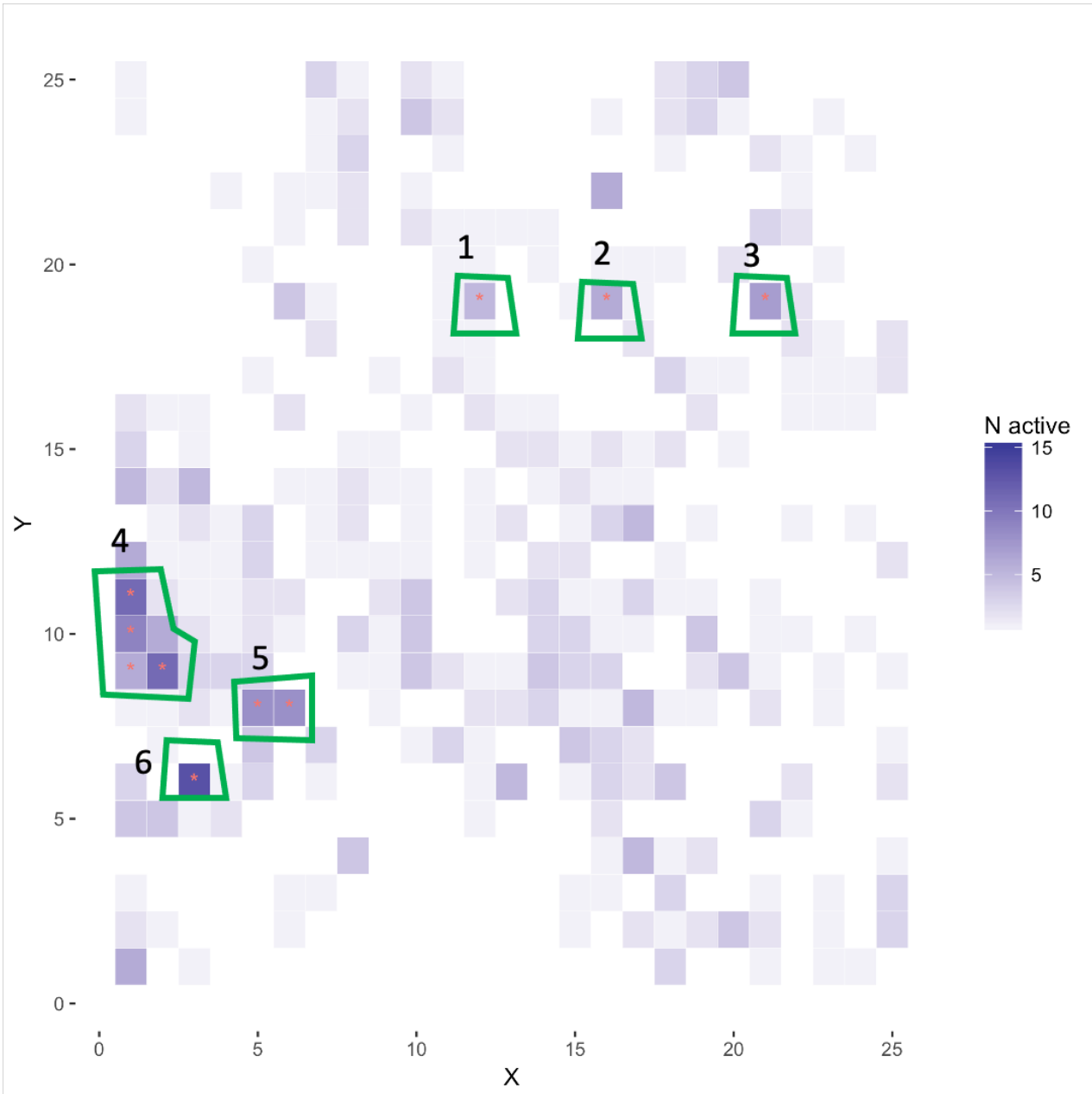
5)



6)

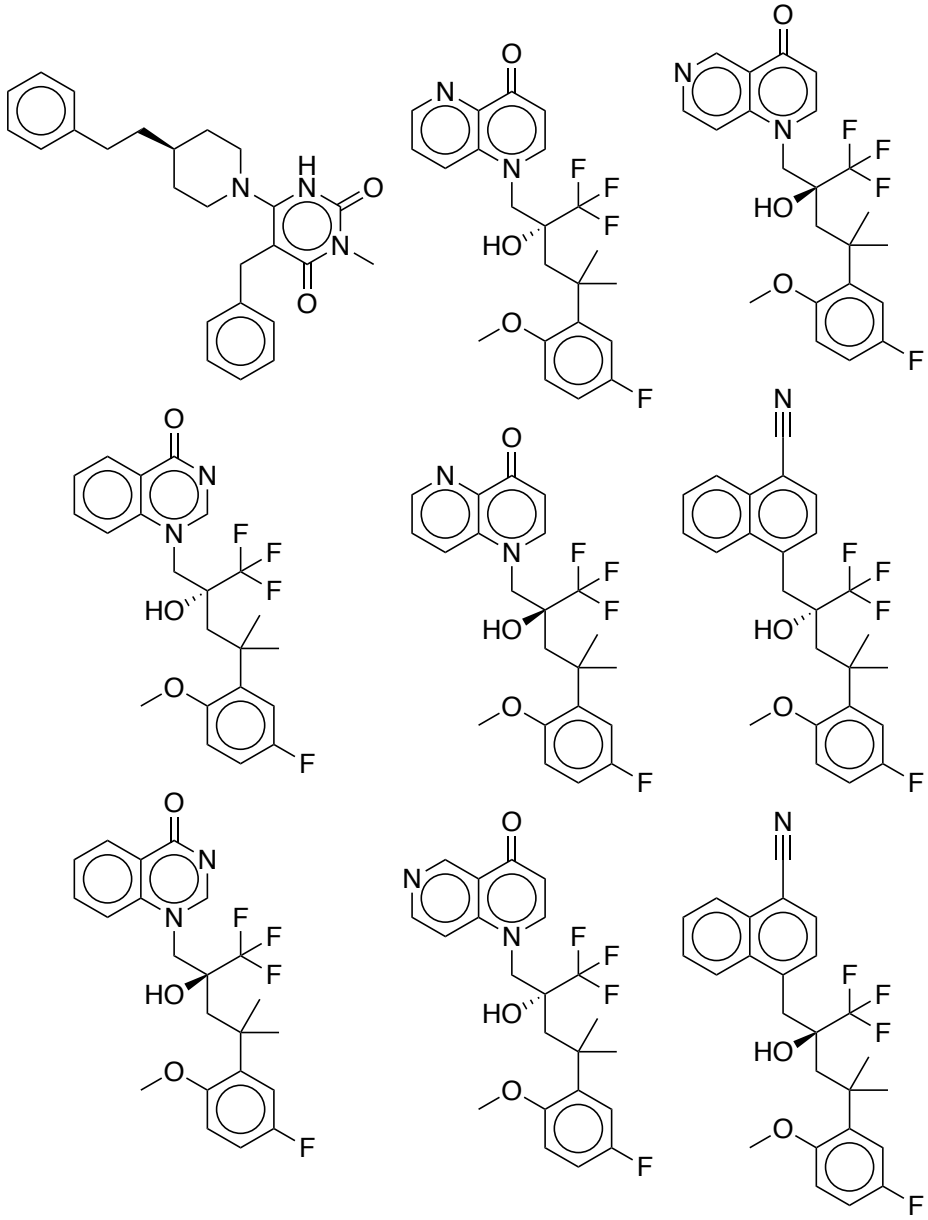


e) GCR

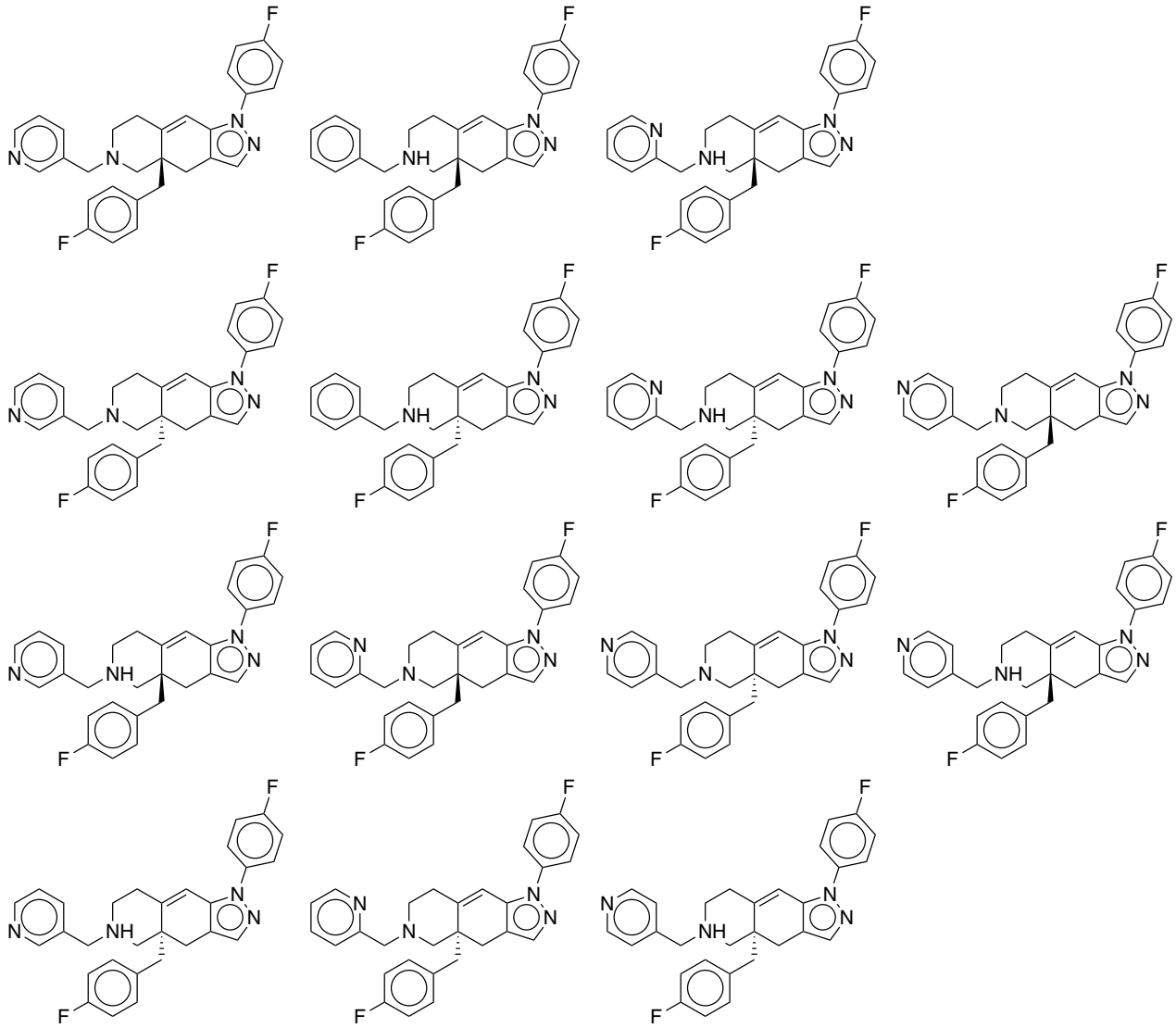


The structures of compounds in each numbered cluster are as follows:

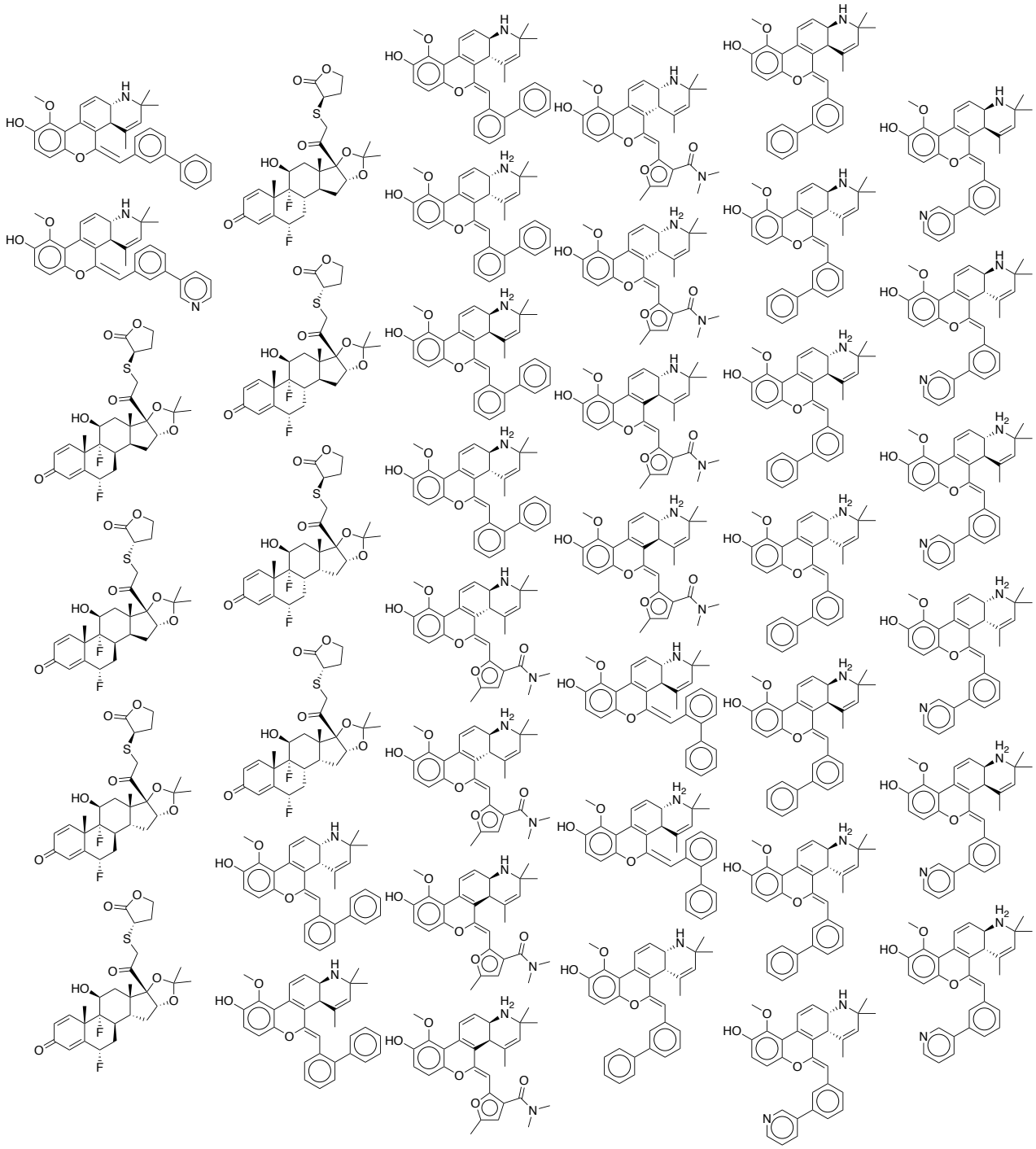
1)



2)

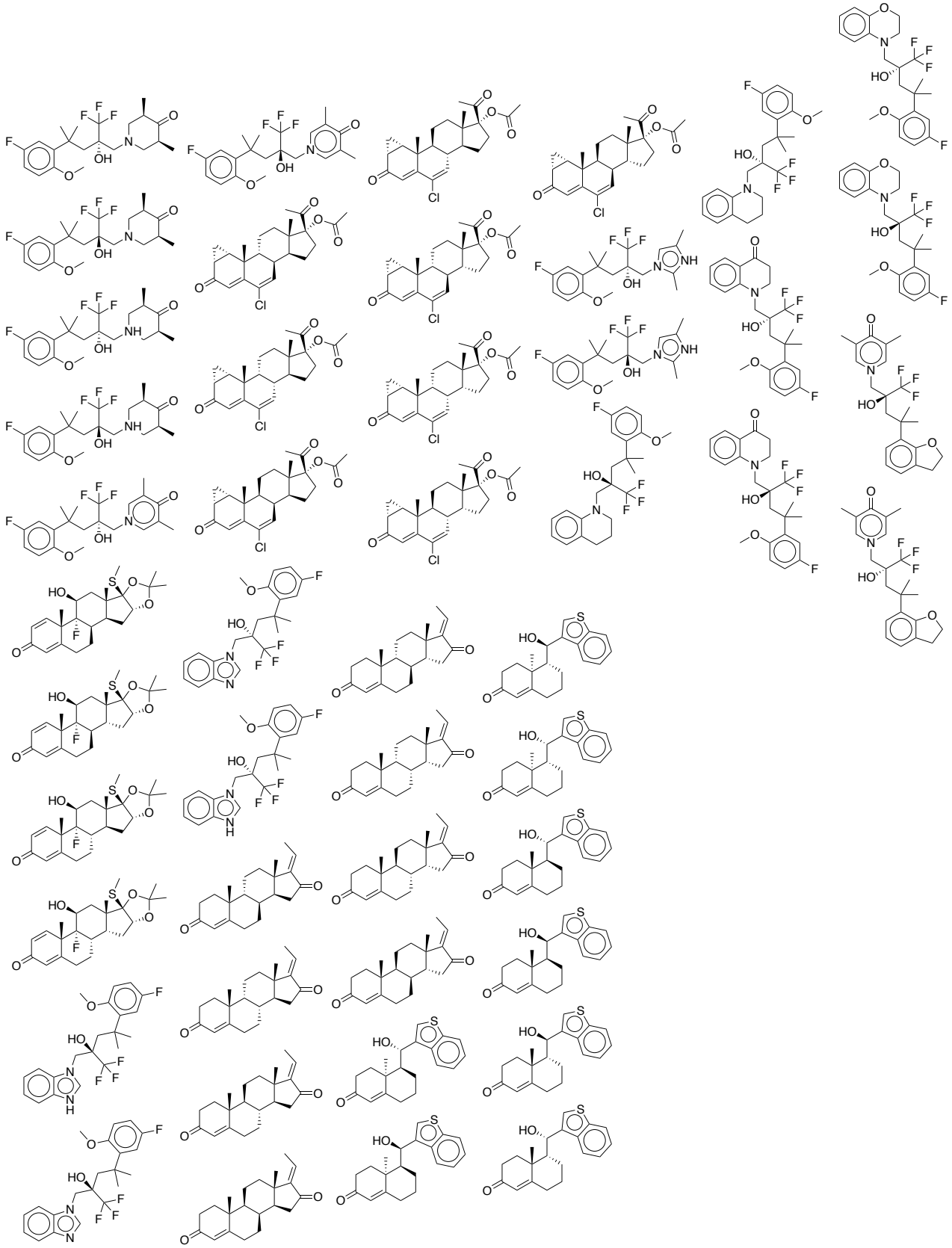


3)

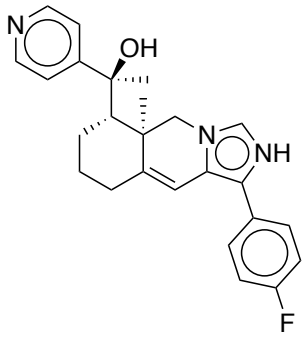


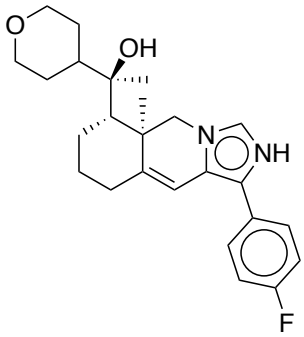
4)

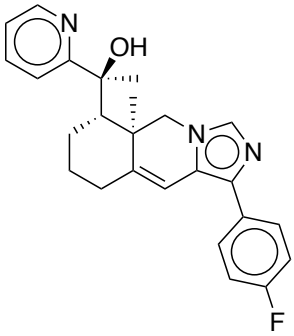
5)

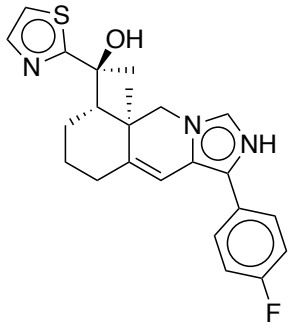


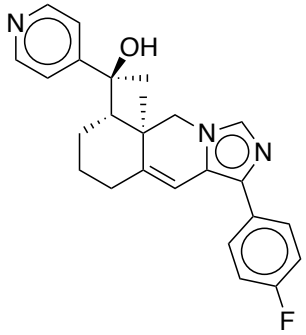
6)

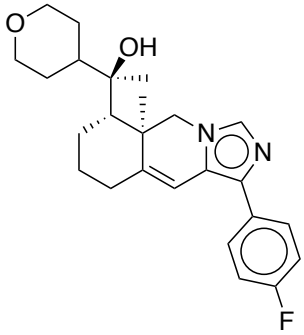


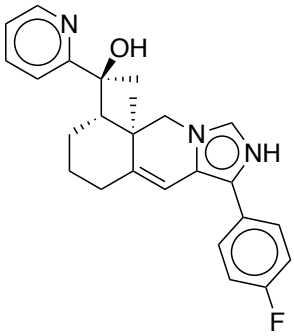


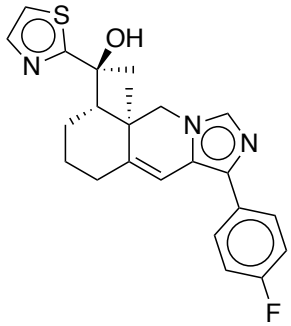


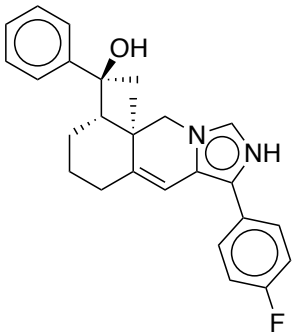


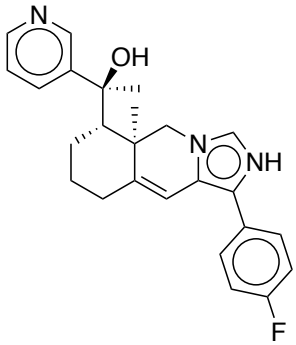


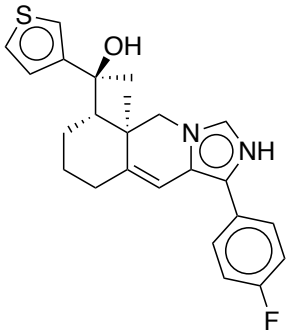


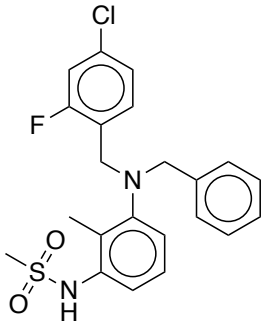


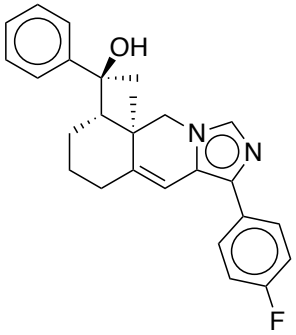


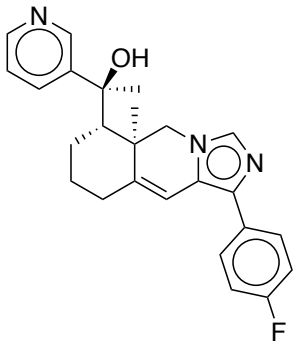


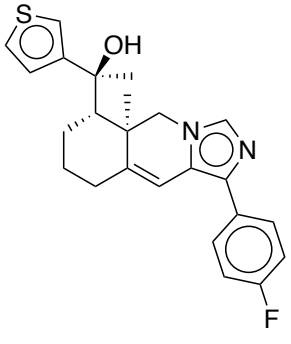




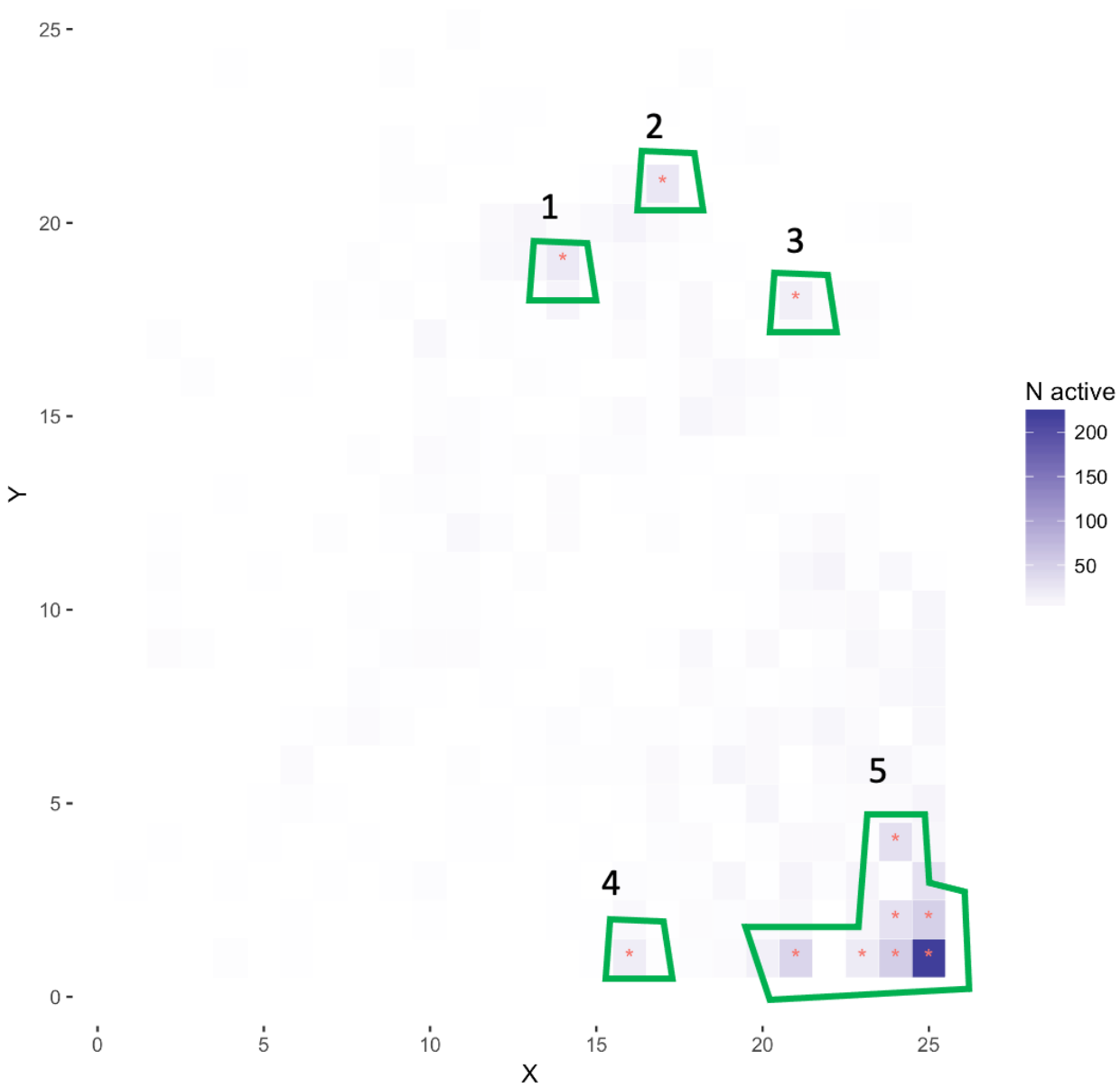






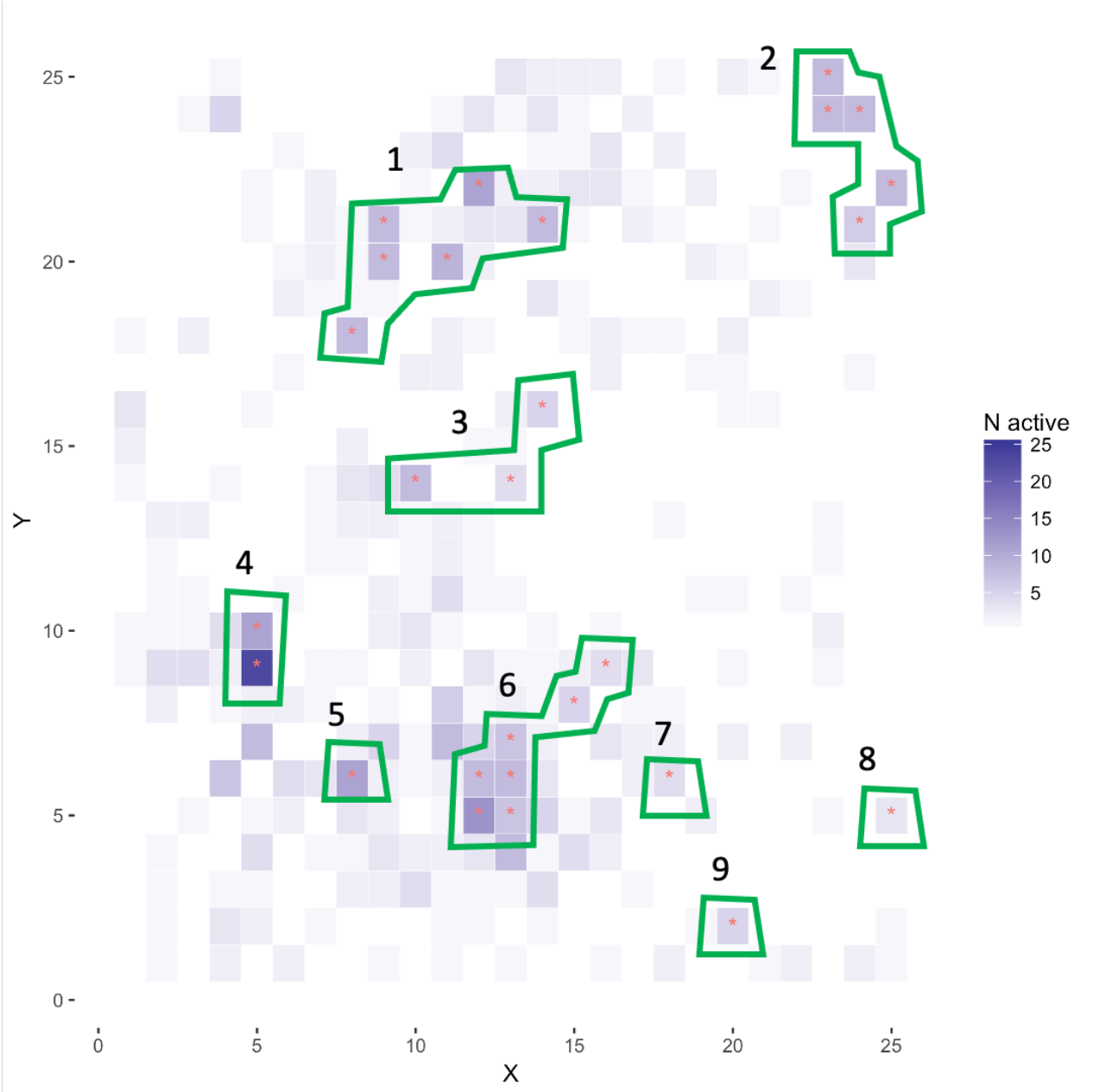


f) HIVPR



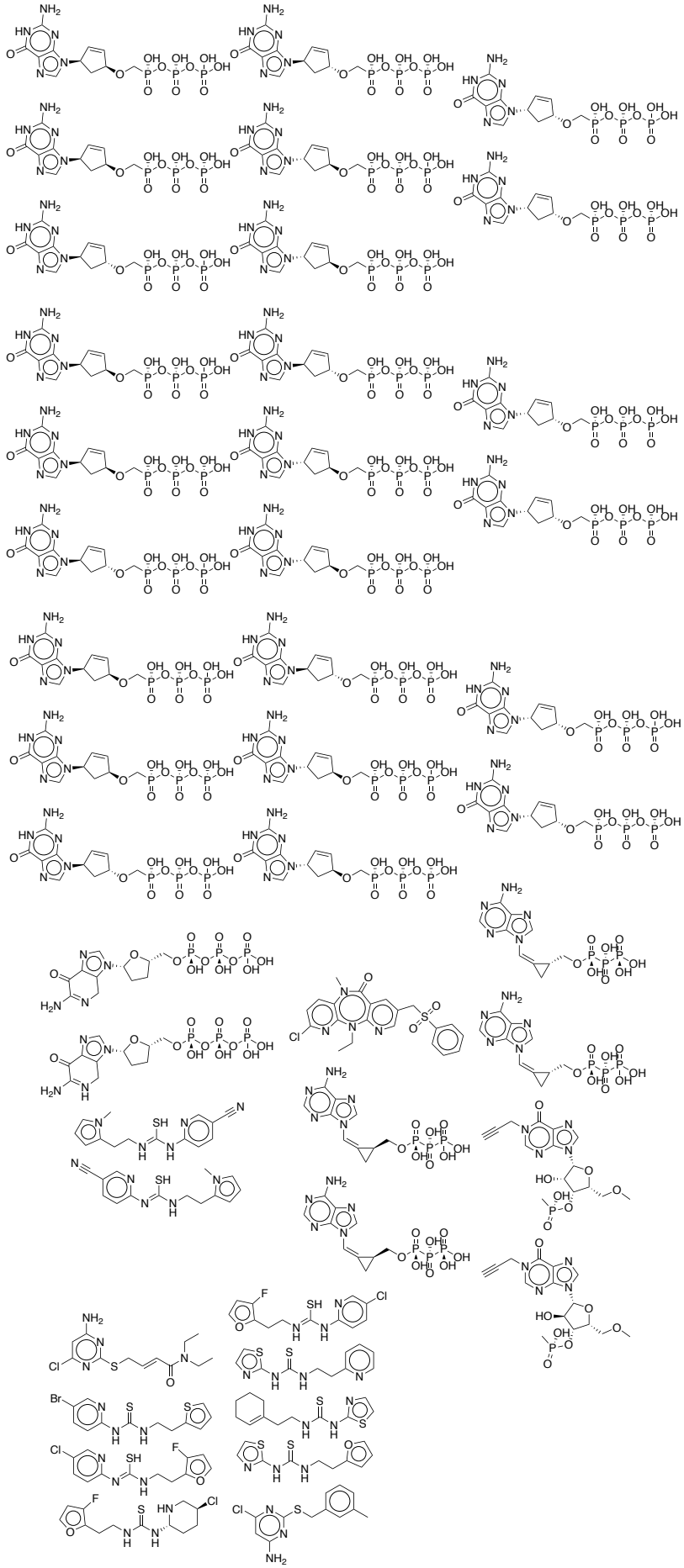
The structures of compounds in each numbered cluster are not shown – too many structures to process.

g) HIVRT

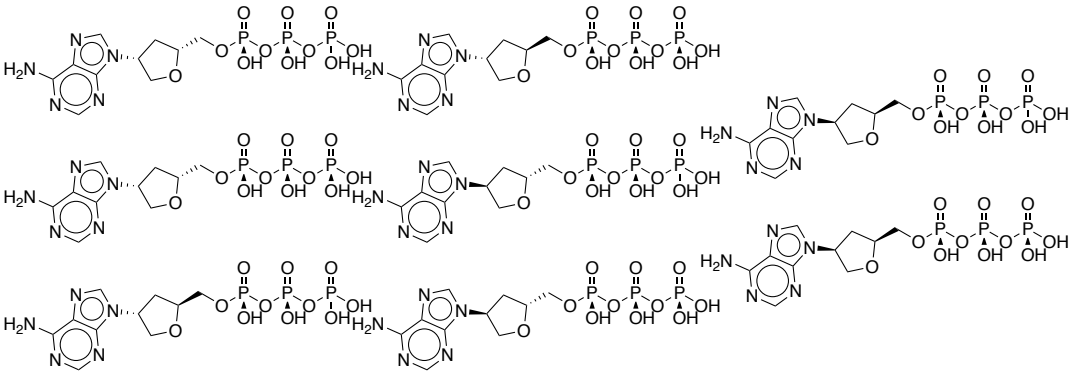


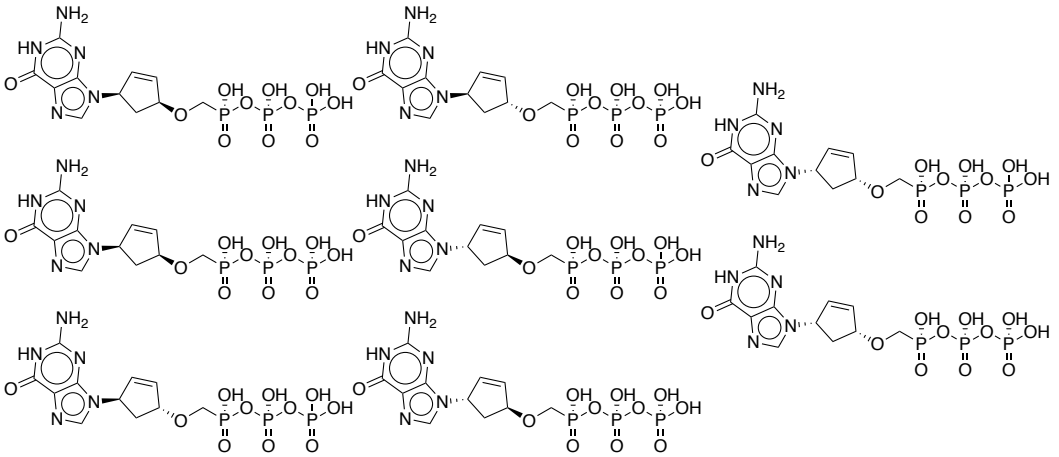
The structures of compounds in each numbered cluster are as follows:

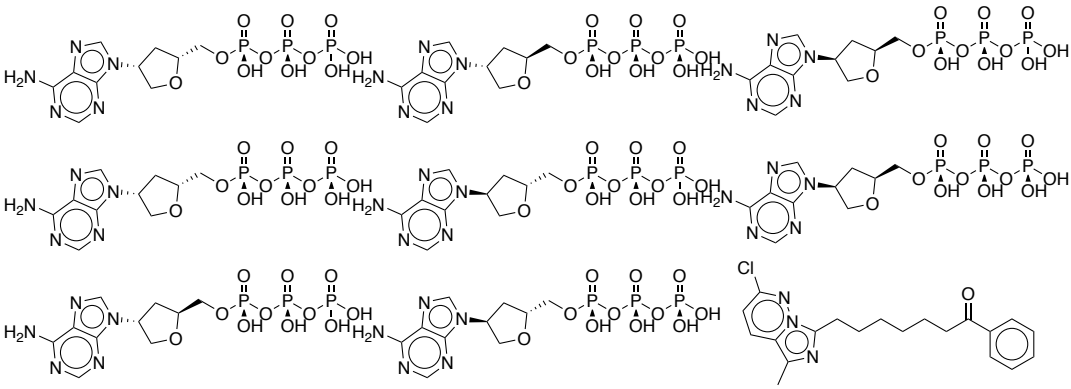
1)

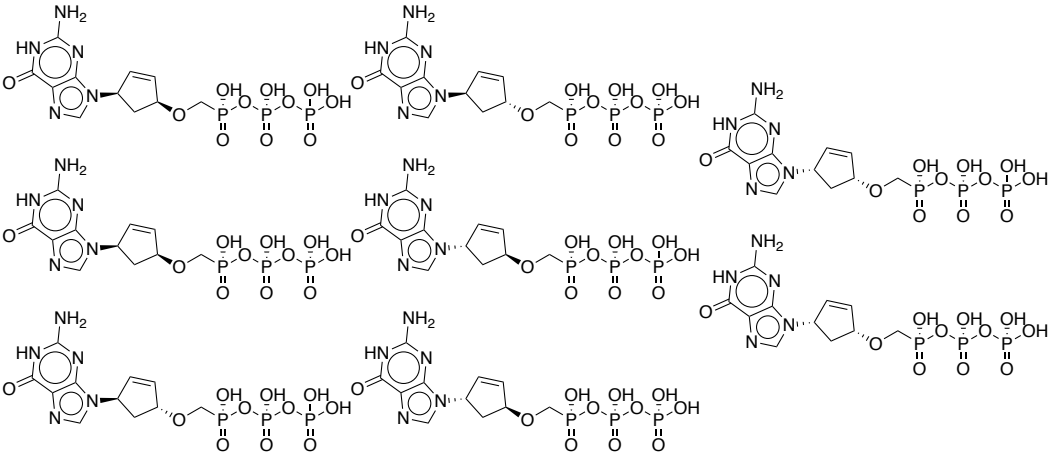


2)

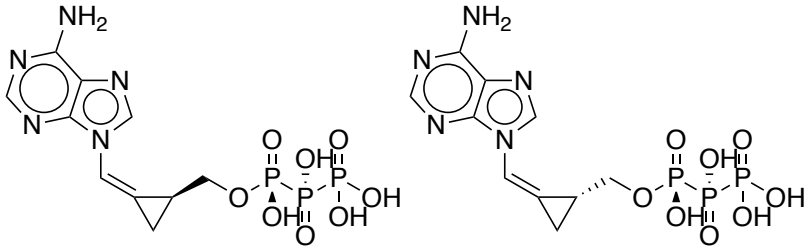


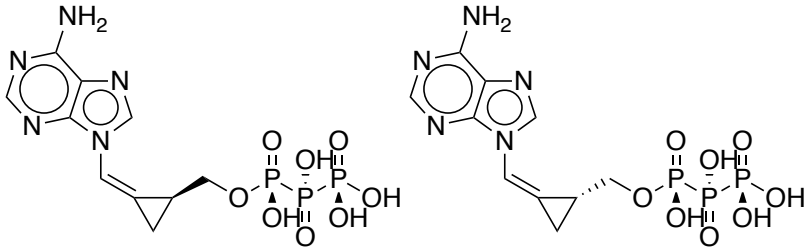


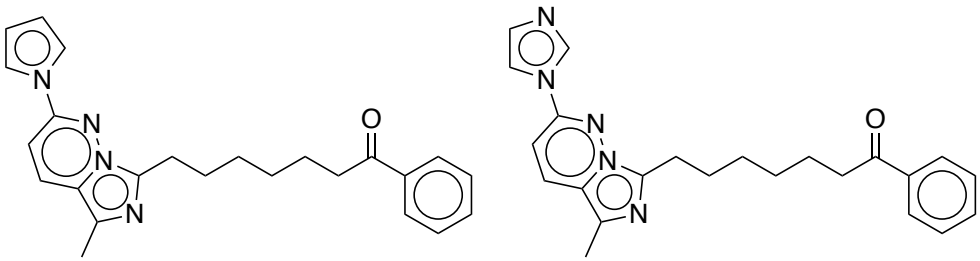


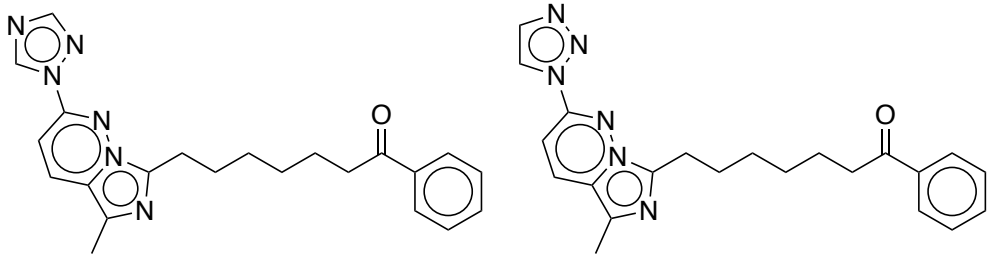


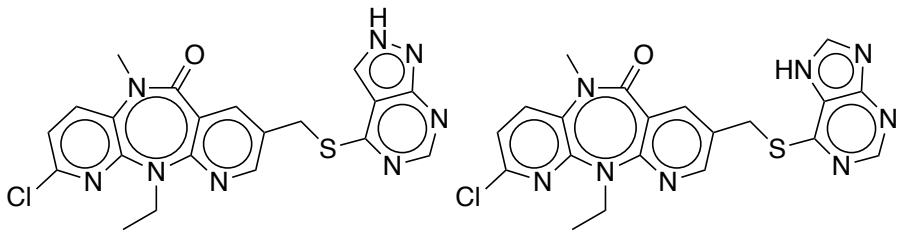
3)

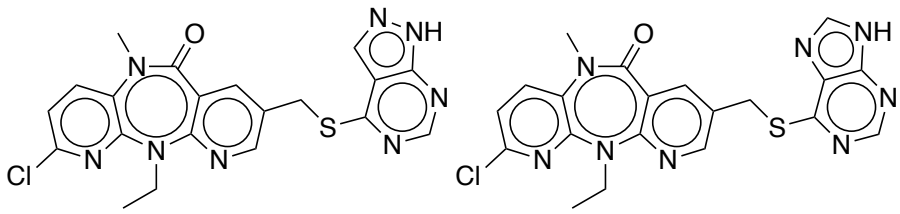




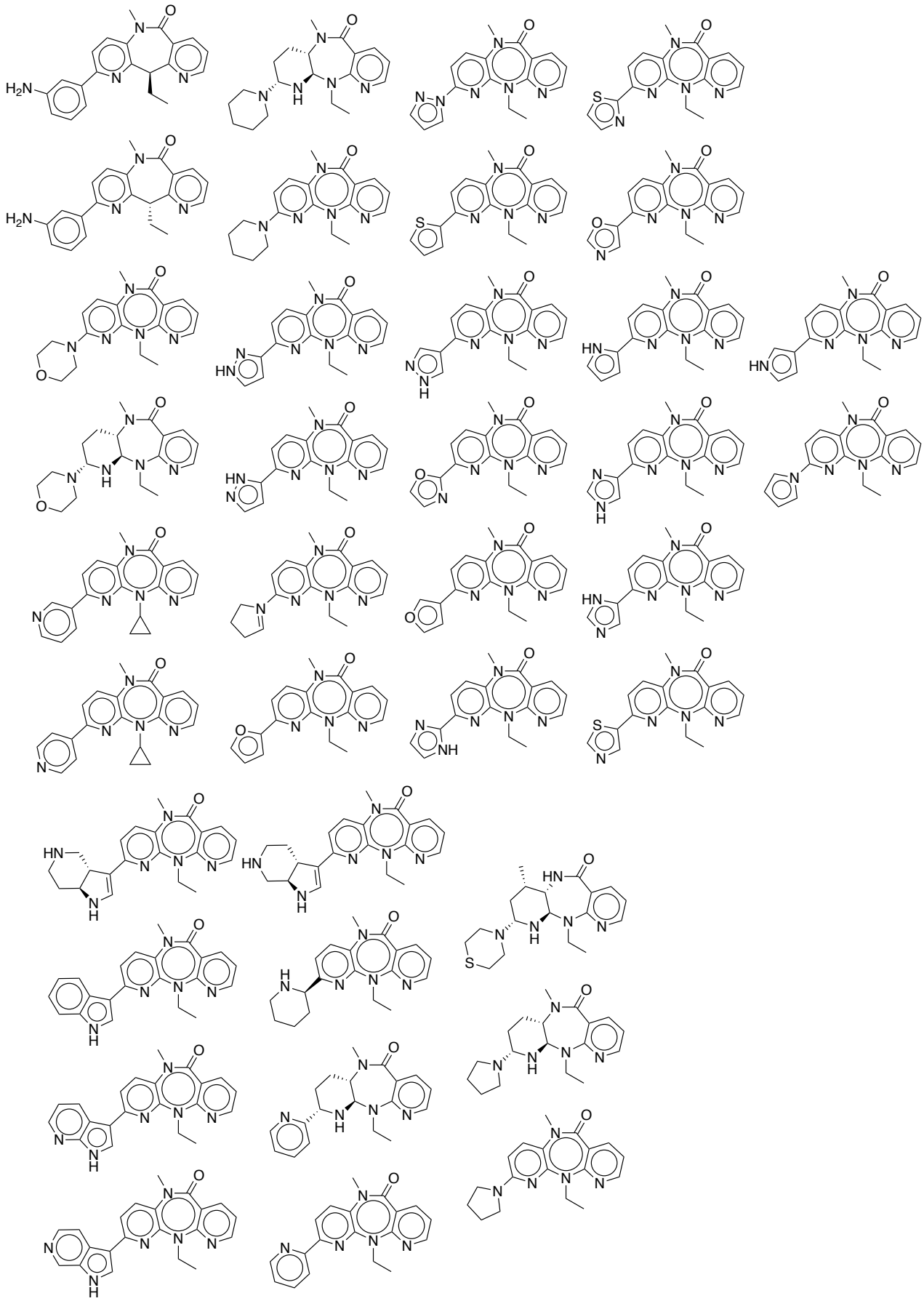




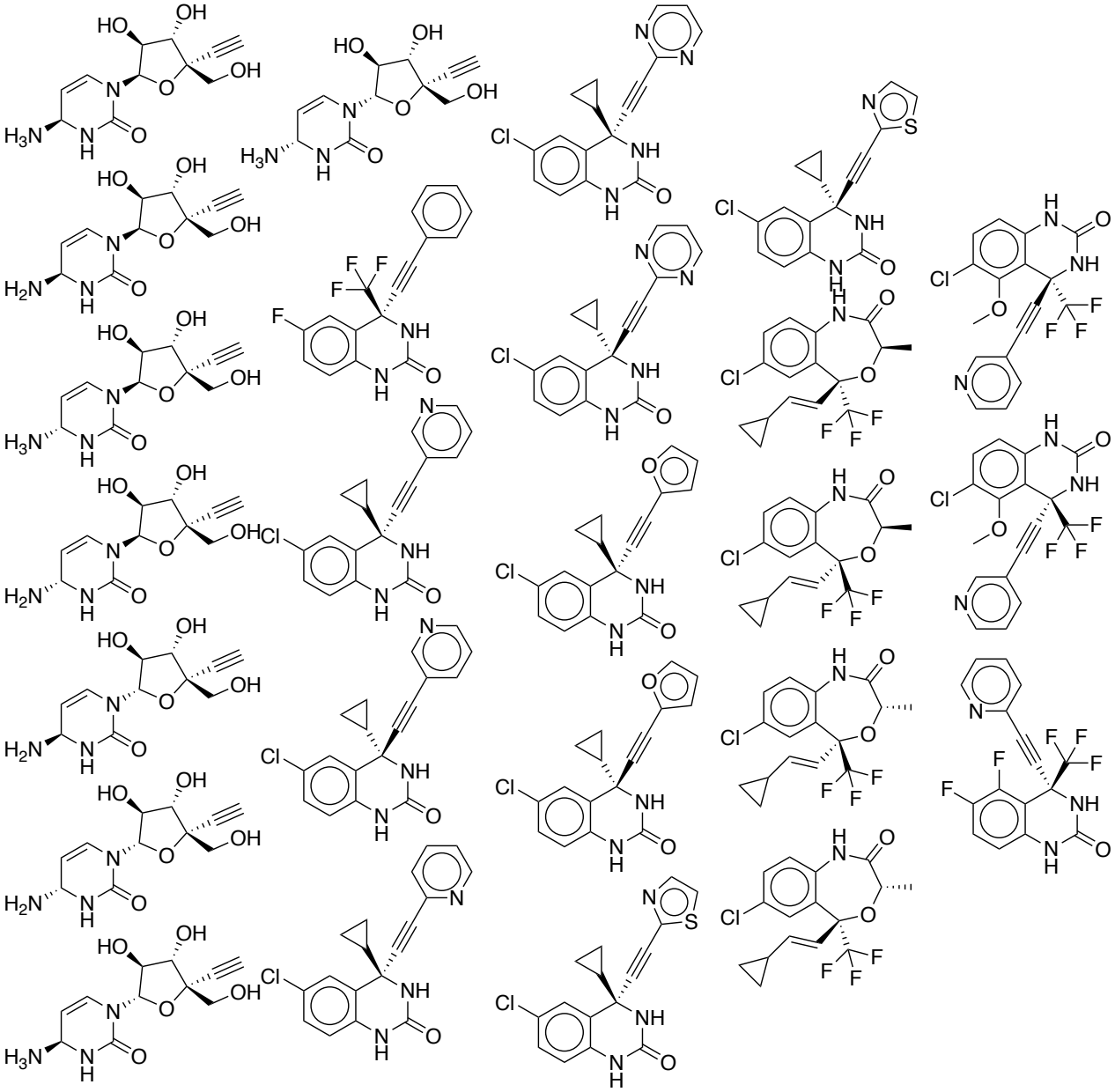




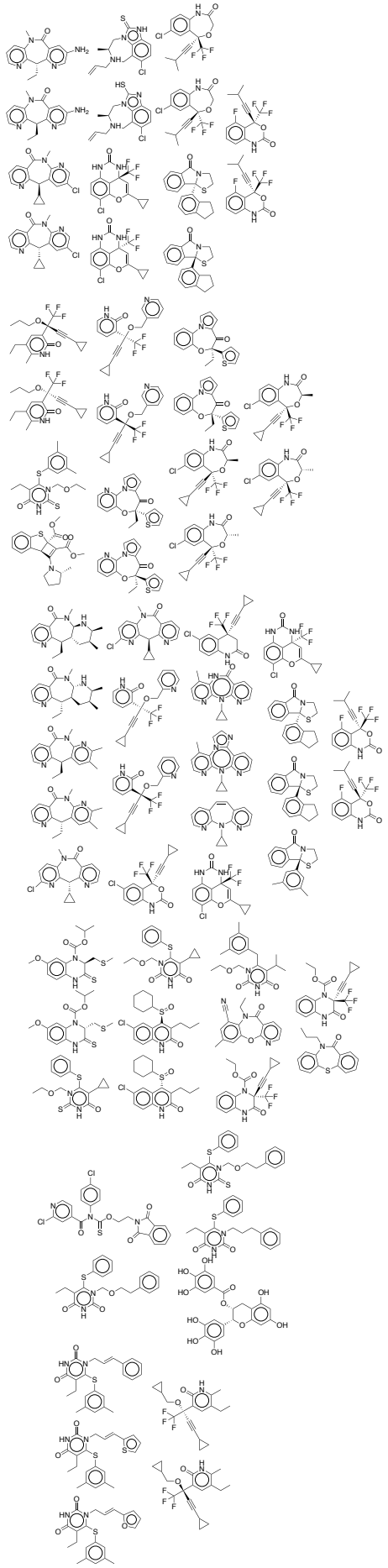
4)



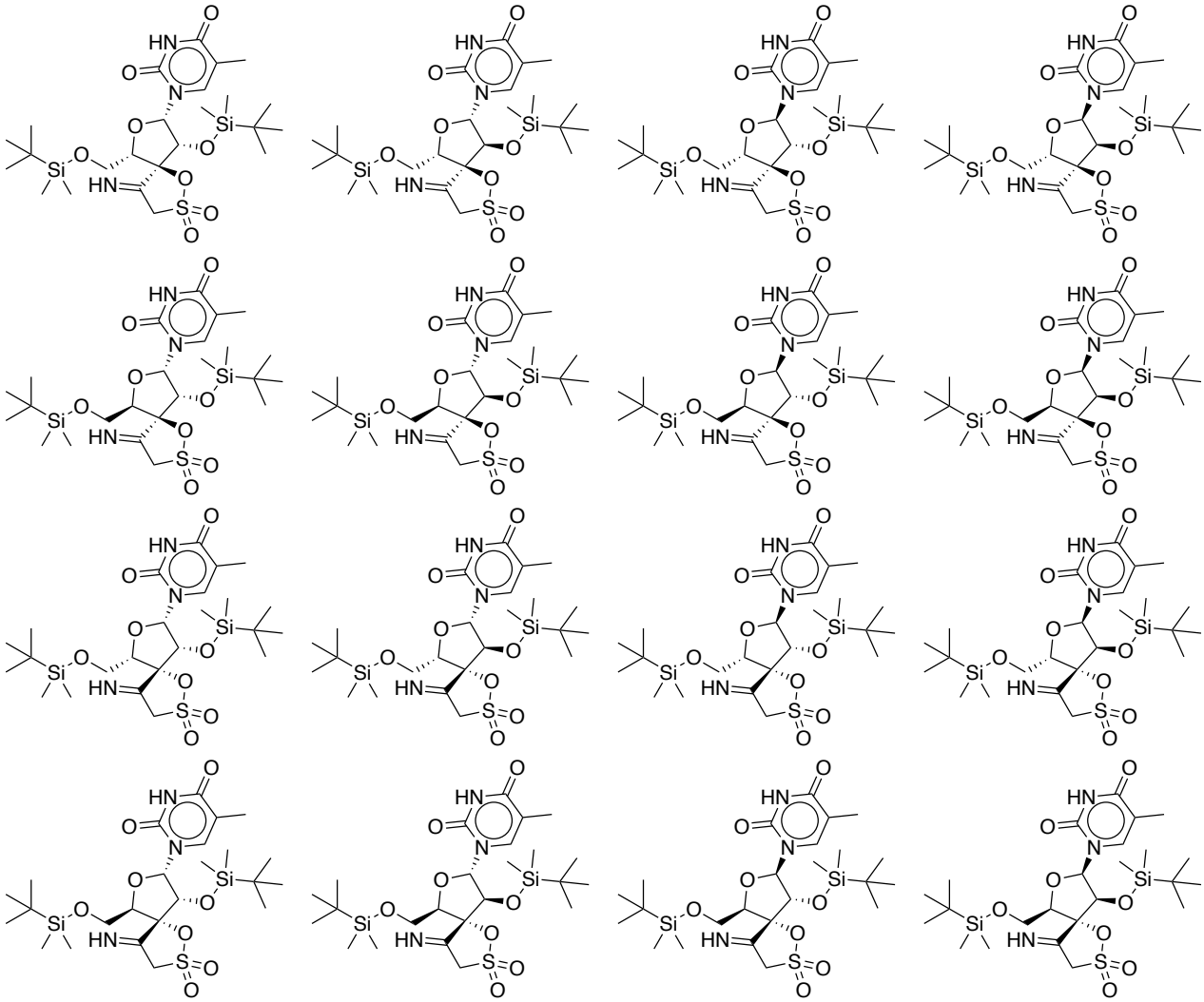
5)



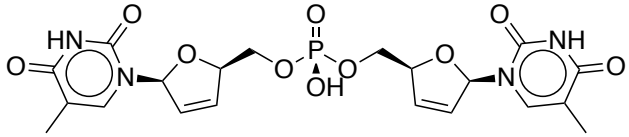
6)

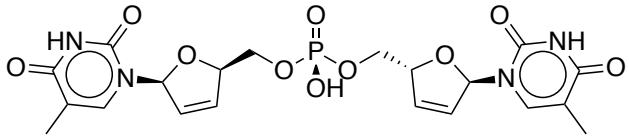


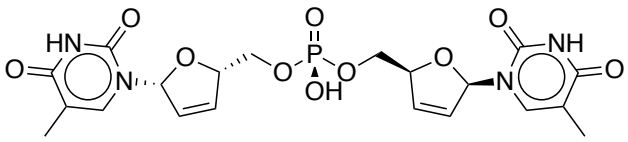
7)

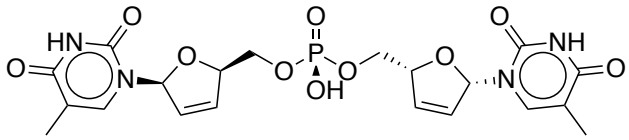


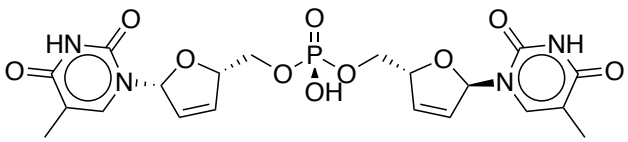
8)

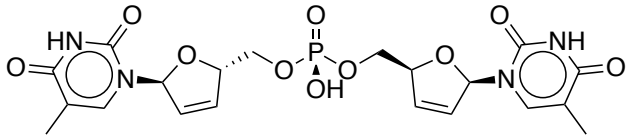


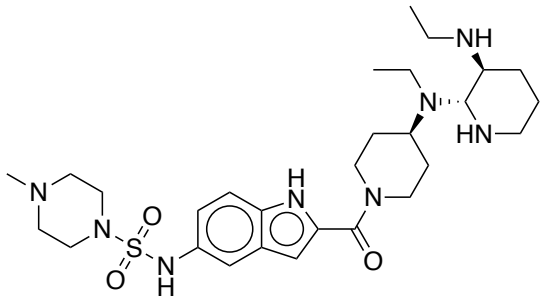


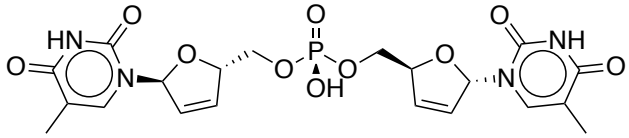


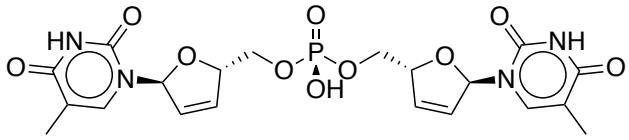


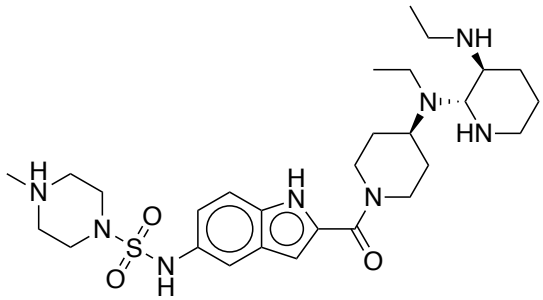


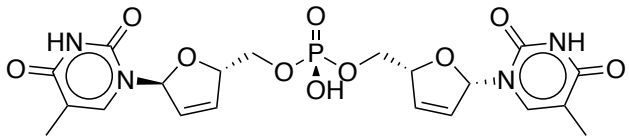


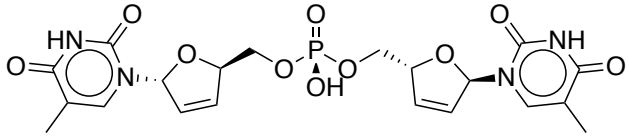




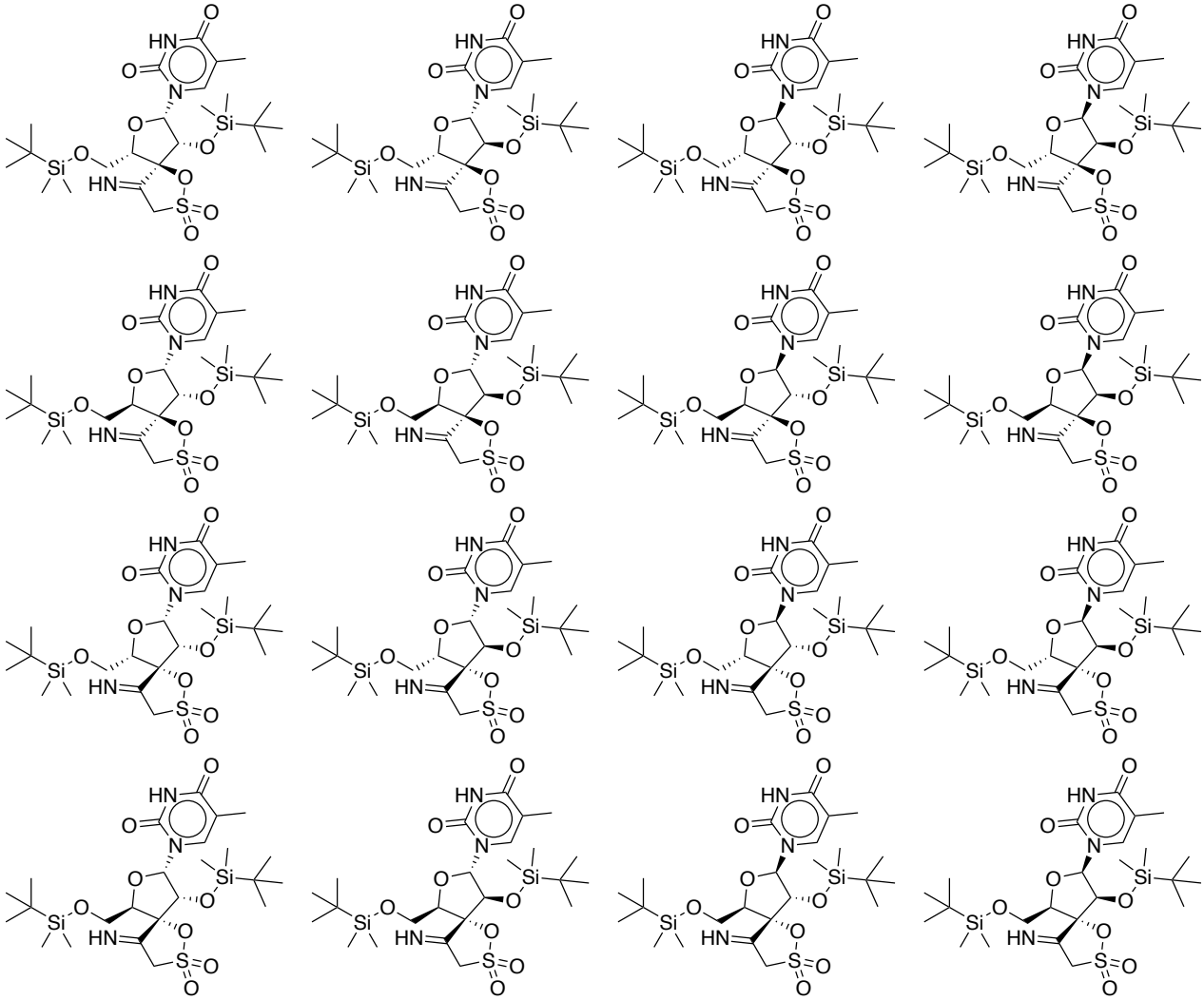




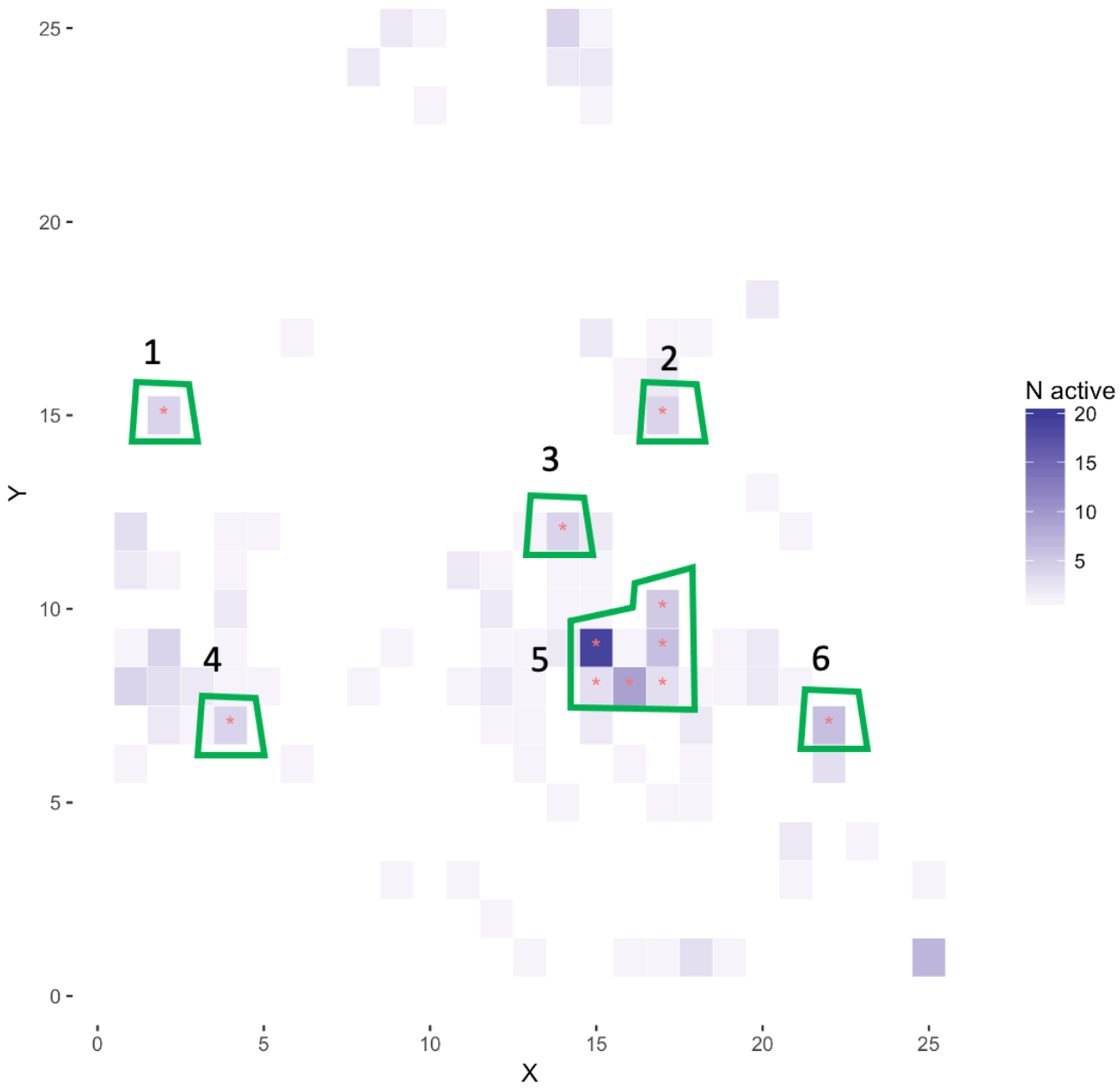




9)

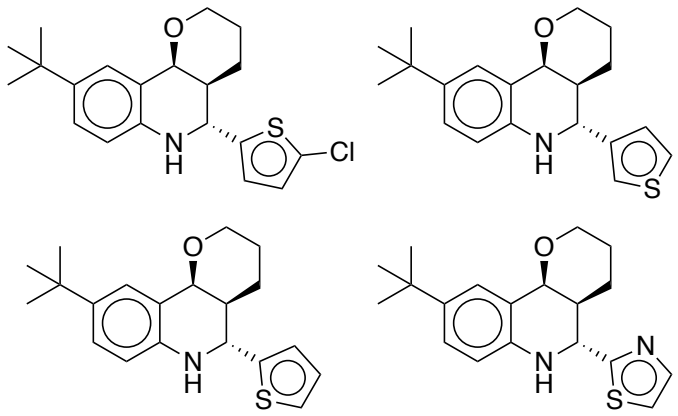


h) KIF11

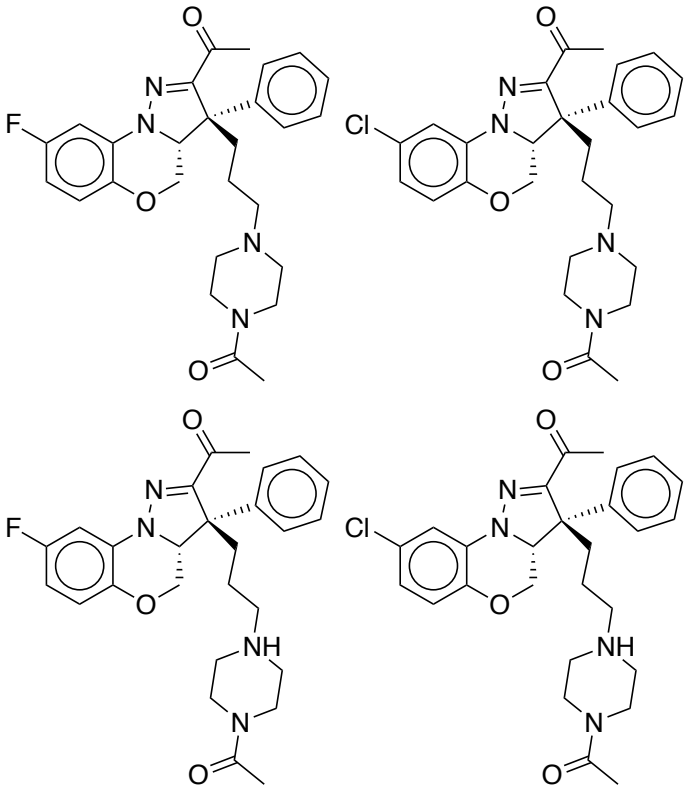


The structures of compounds in each numbered cluster are as follows:

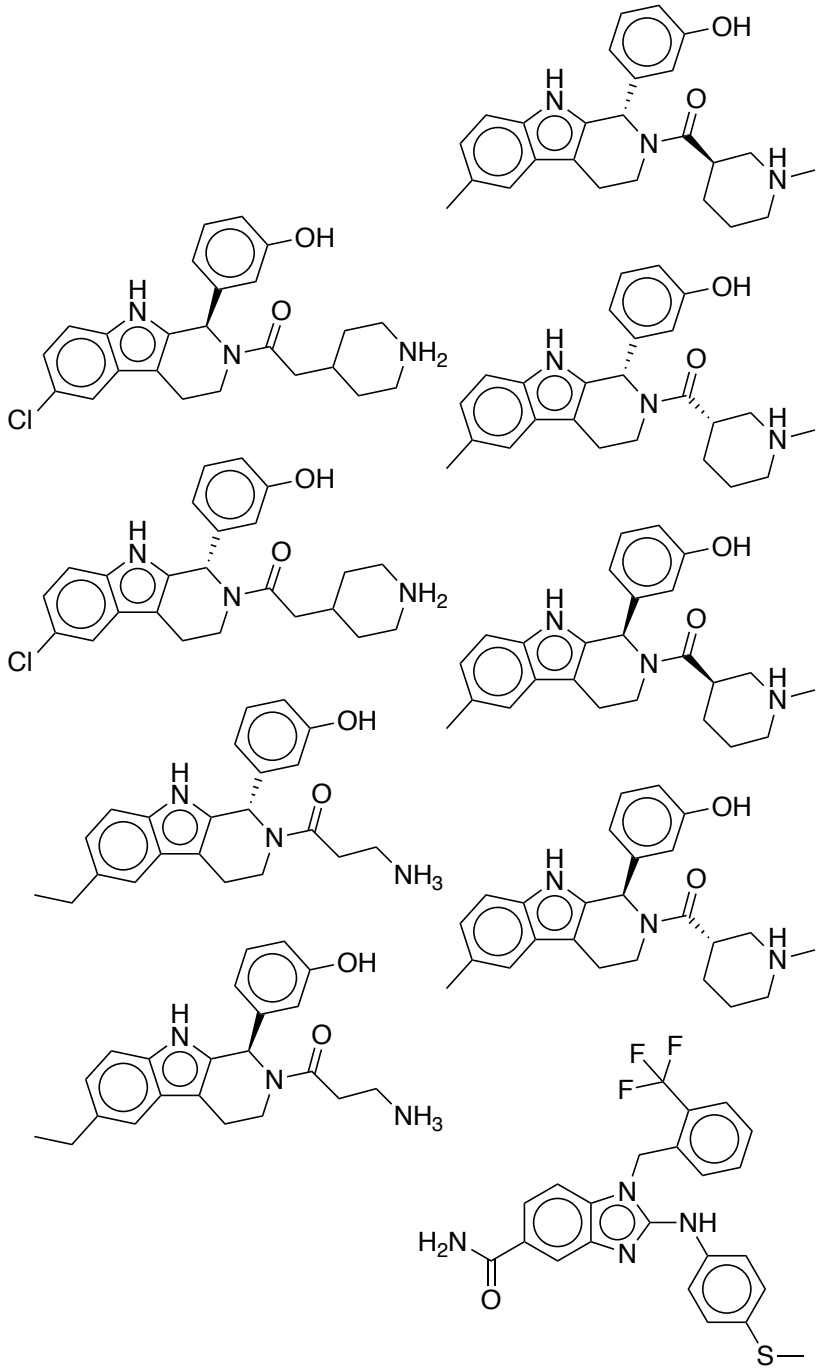
1)



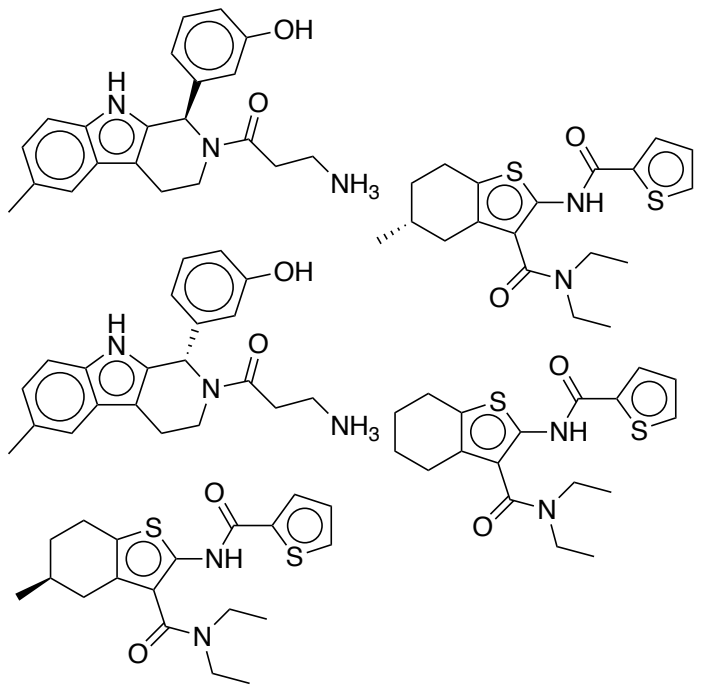
2)



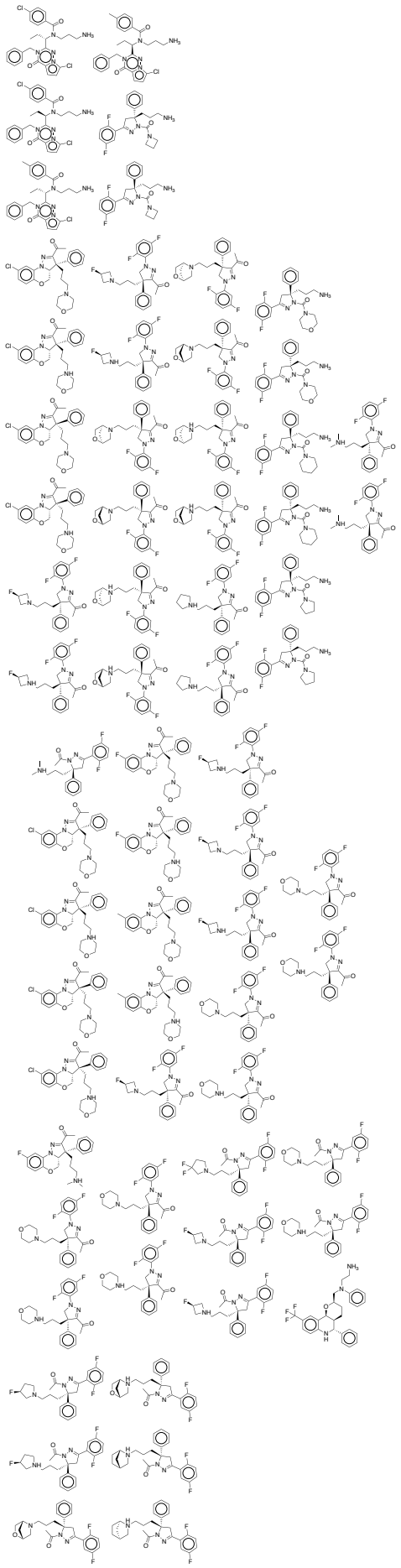
3)



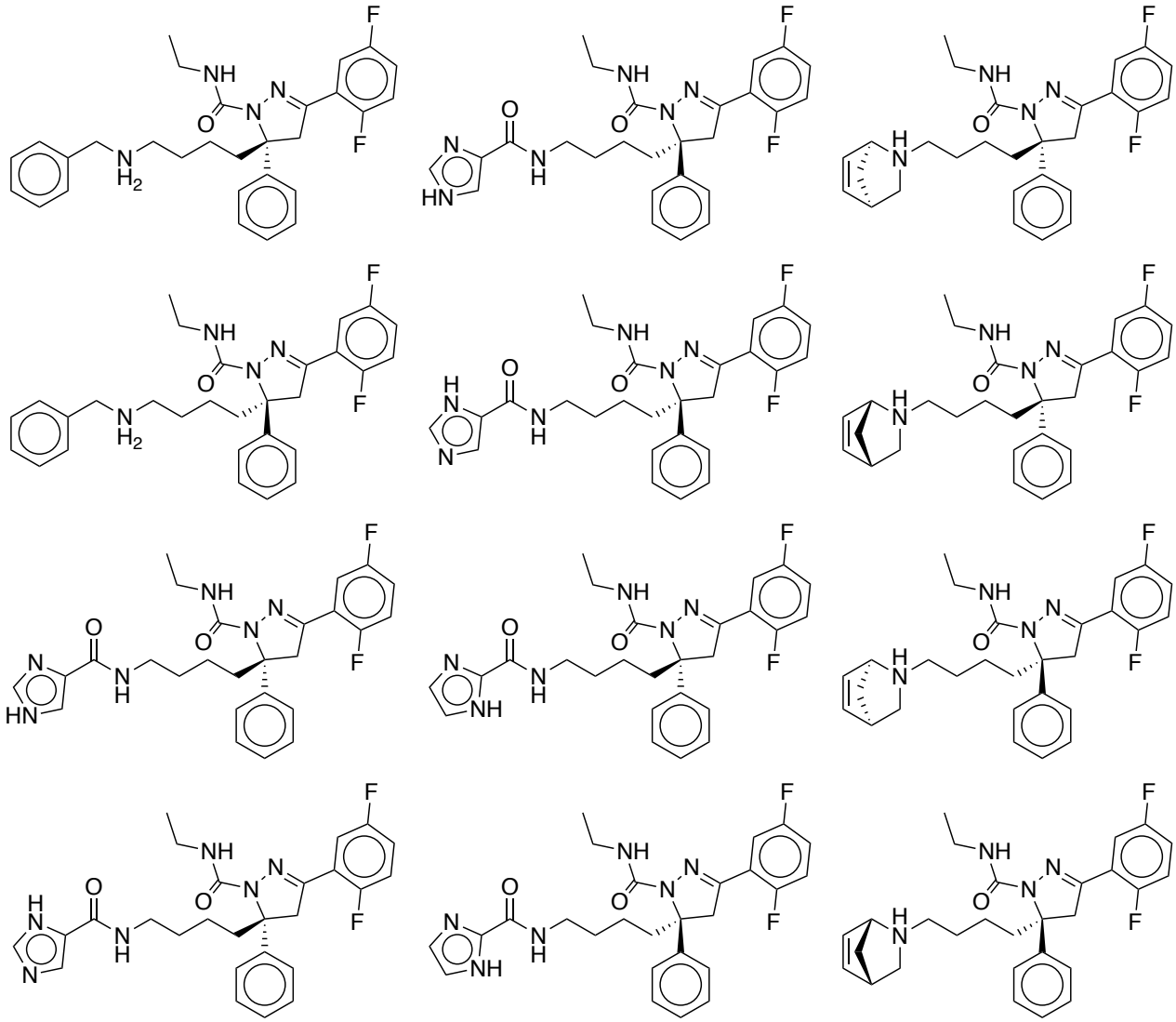
4)



5)

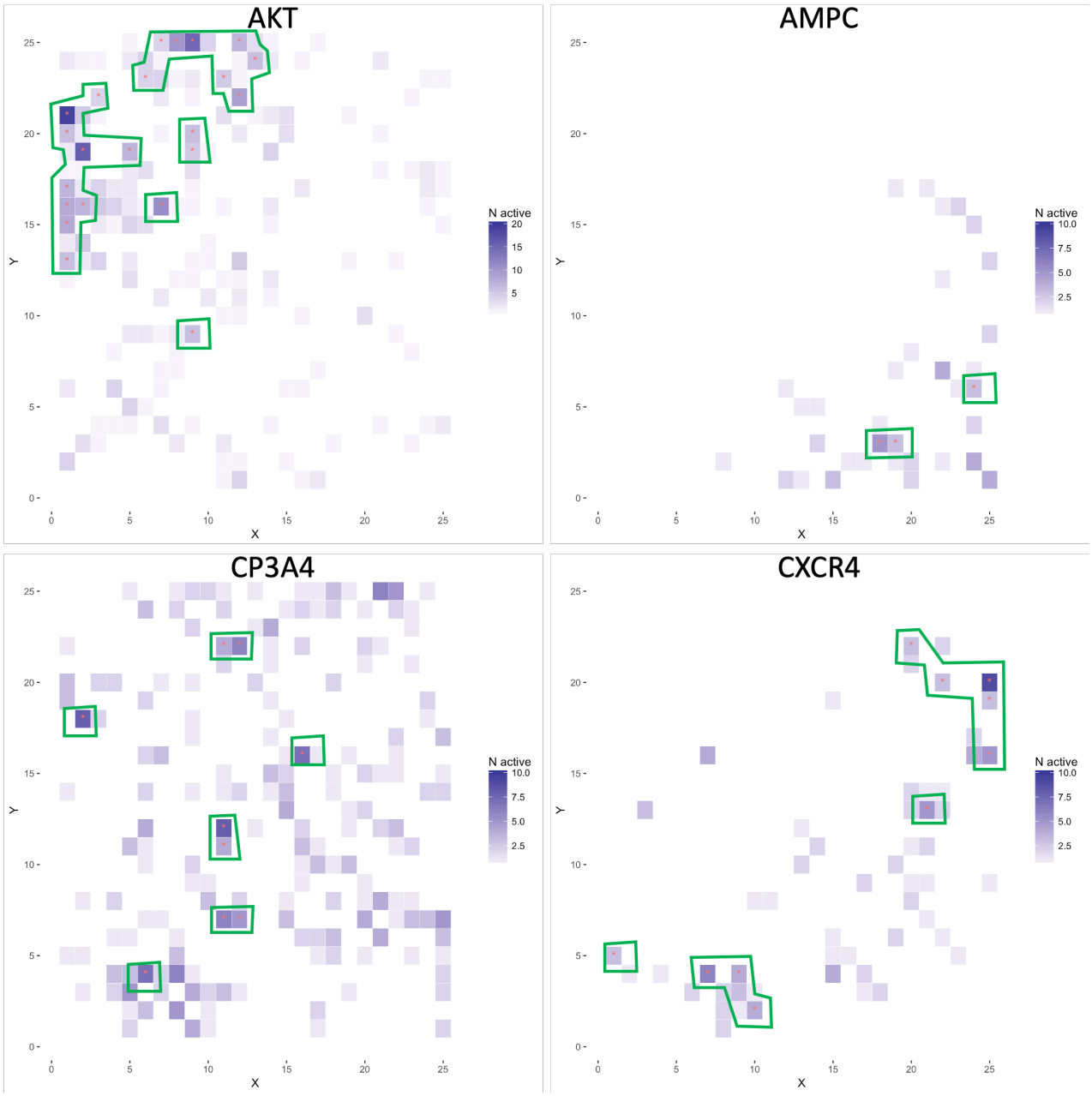


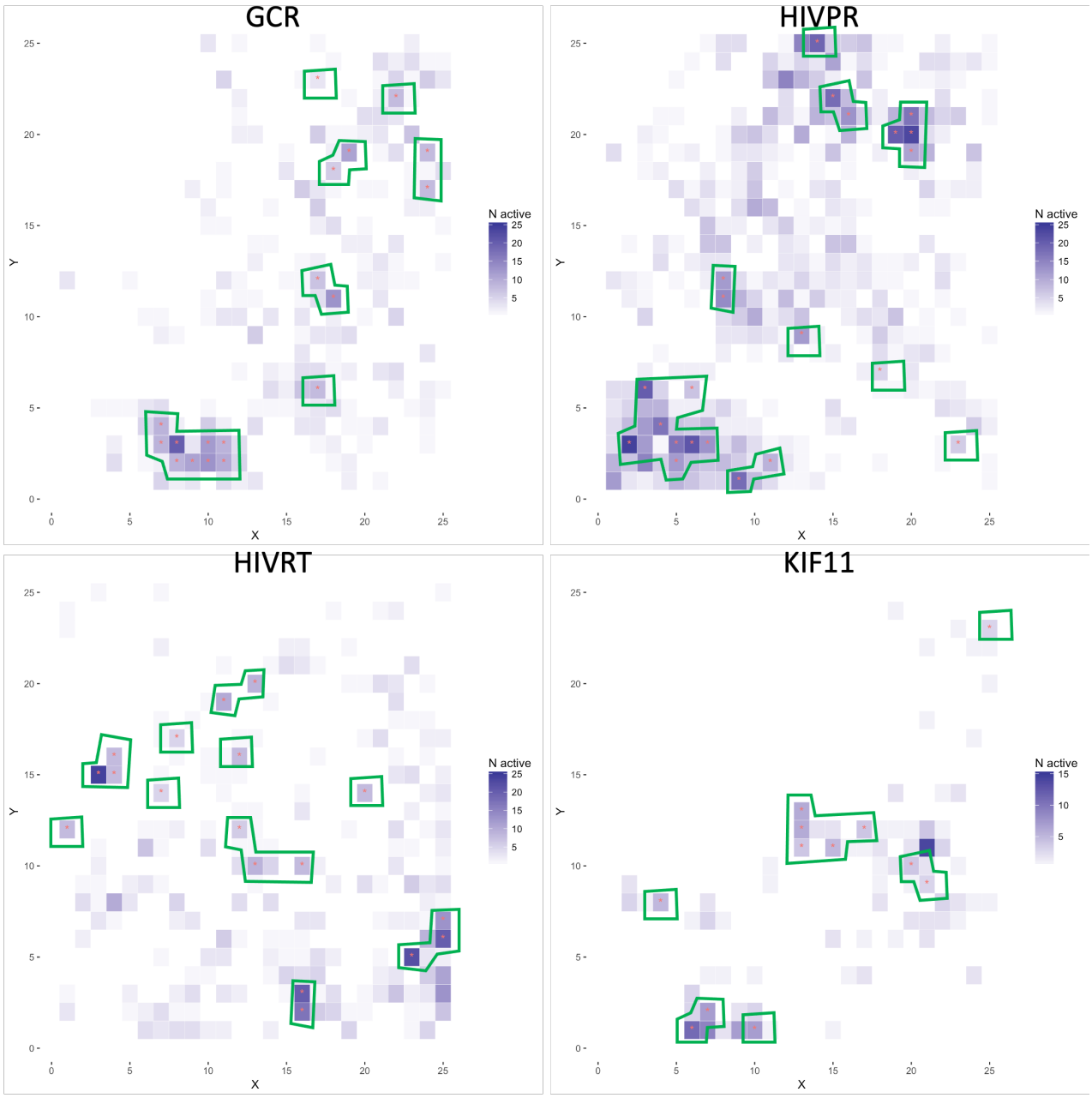
6)



S8) Shape maps created using USR descriptions of molecular shapes.

The SOM obtained when USR is used to generate the shape description is shown below for each of the eight DUD-E targets.

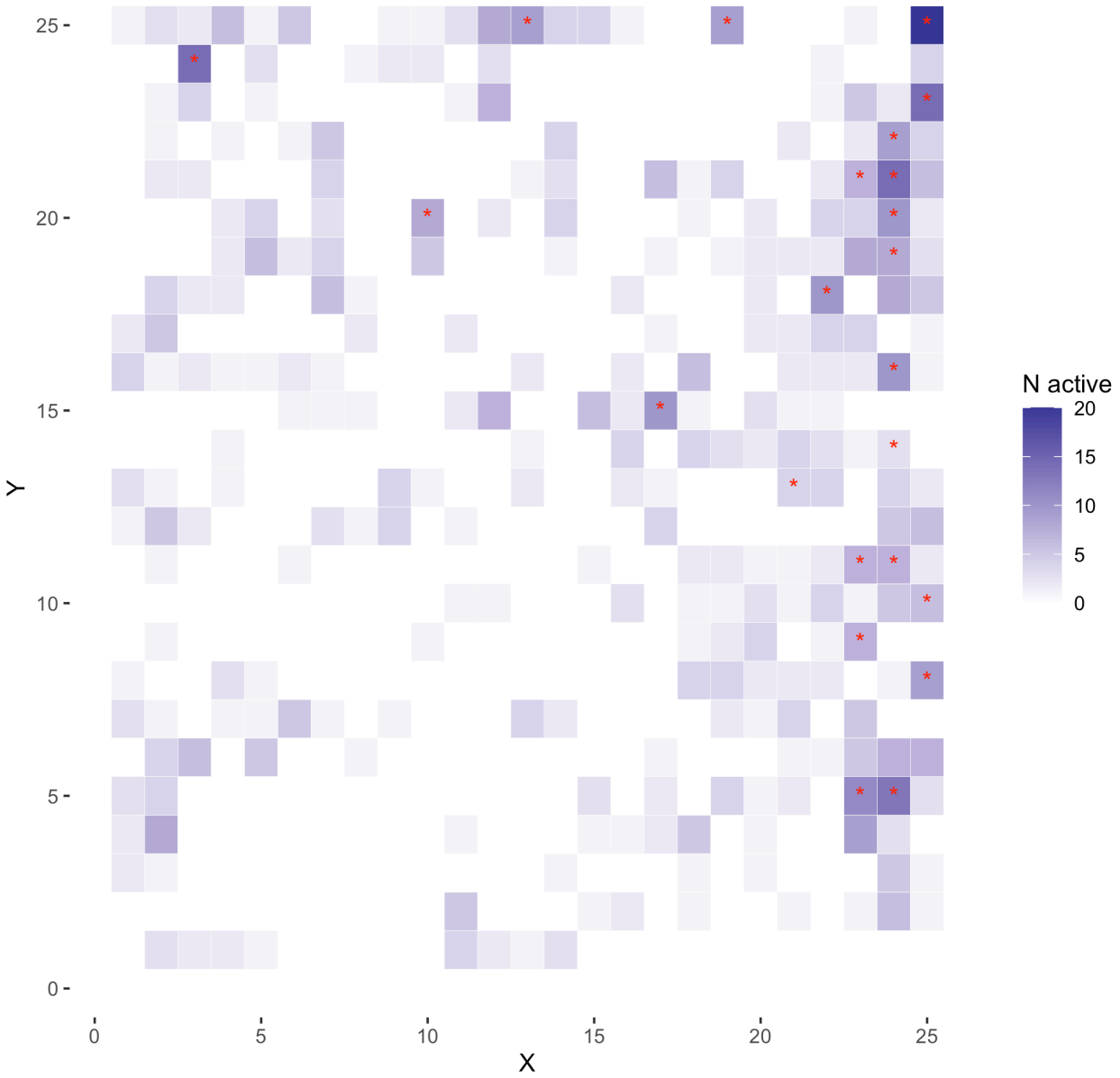

AKT
AMPC
CP3A4
CXCR4
N active
X
Y



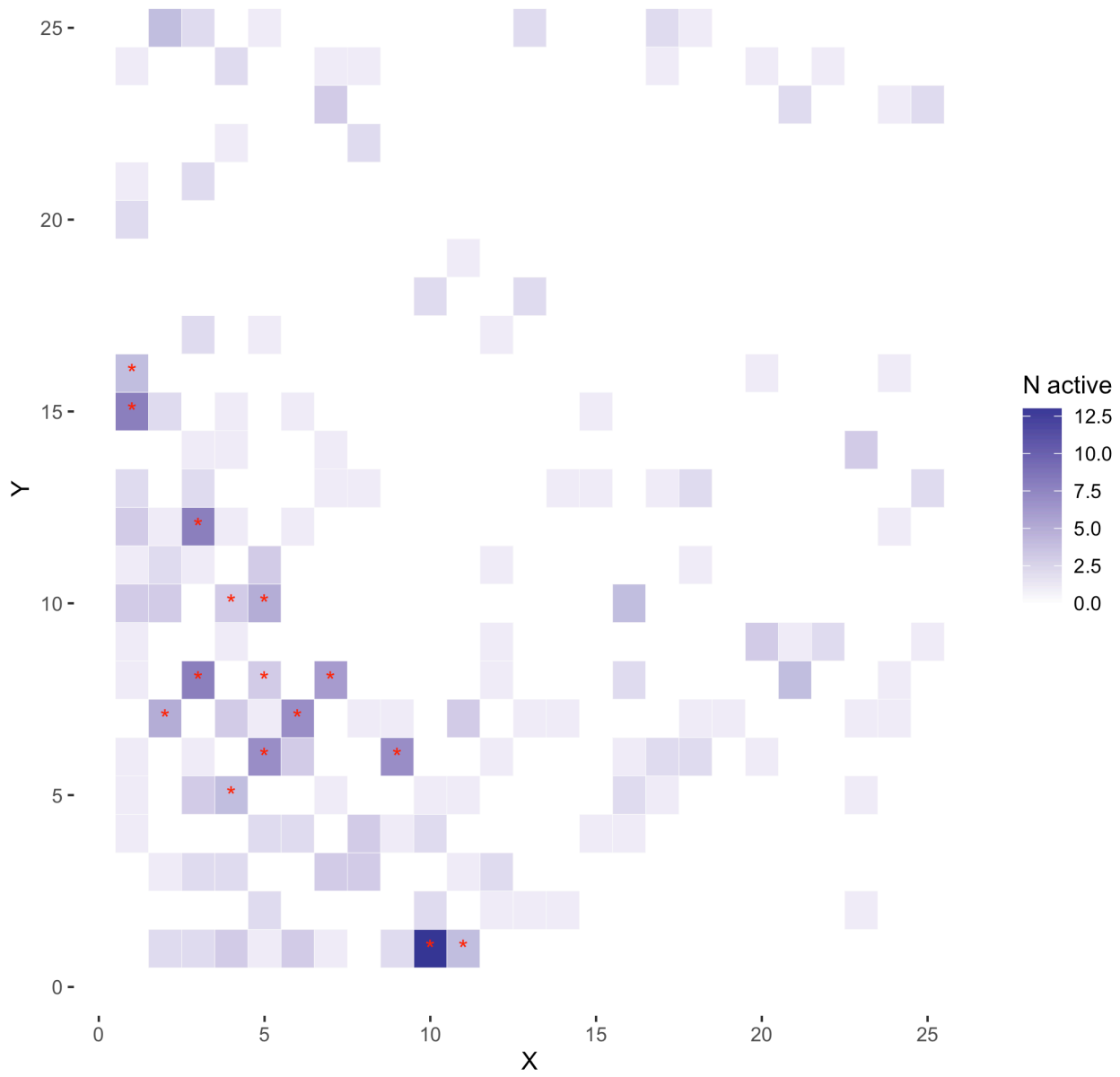
S9) Shapemaps for the full set of DUD-E targets created by projecting only each set of actives with their matched decoys.

In the previous set of maps, all compounds were projected into two-dimensional SOMs in a single process such that coordinates in each map correspond. In the maps below that is not the case – each map is created specifically for each target with its own set of actives and decoys. The enrichment level can be assessed at each coordinate in the map as well as an overall enrichment level ($N_{actives}$(in enriched) x ($N_{actives}$(total) + N_{decoys}(total)) / (($N_{actives}$(in enriched) + N_{decoys}(in enriched)) x $N_{actives}$(total)) and the overall percentage recovery of active compounds.

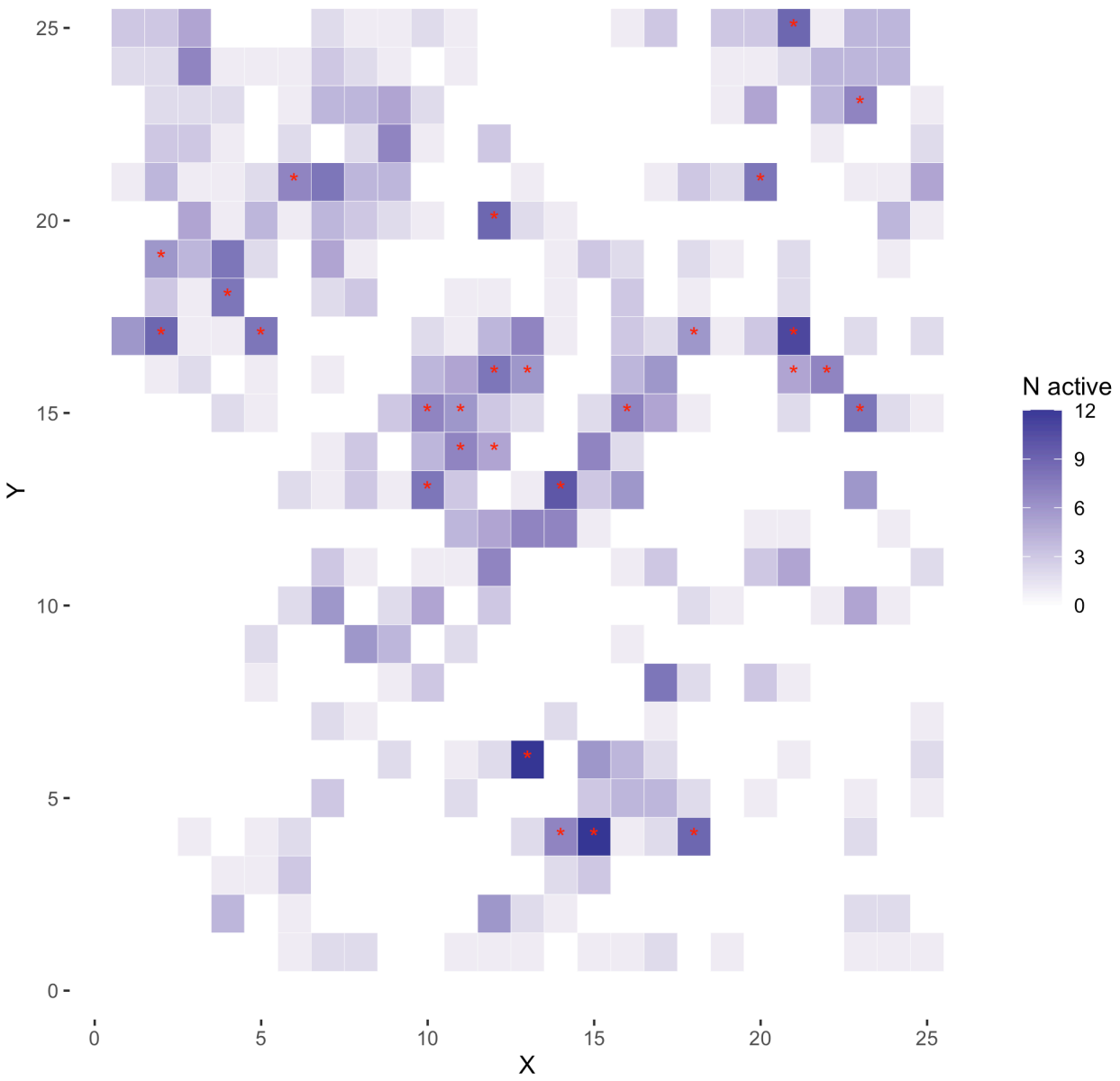
aa2ar



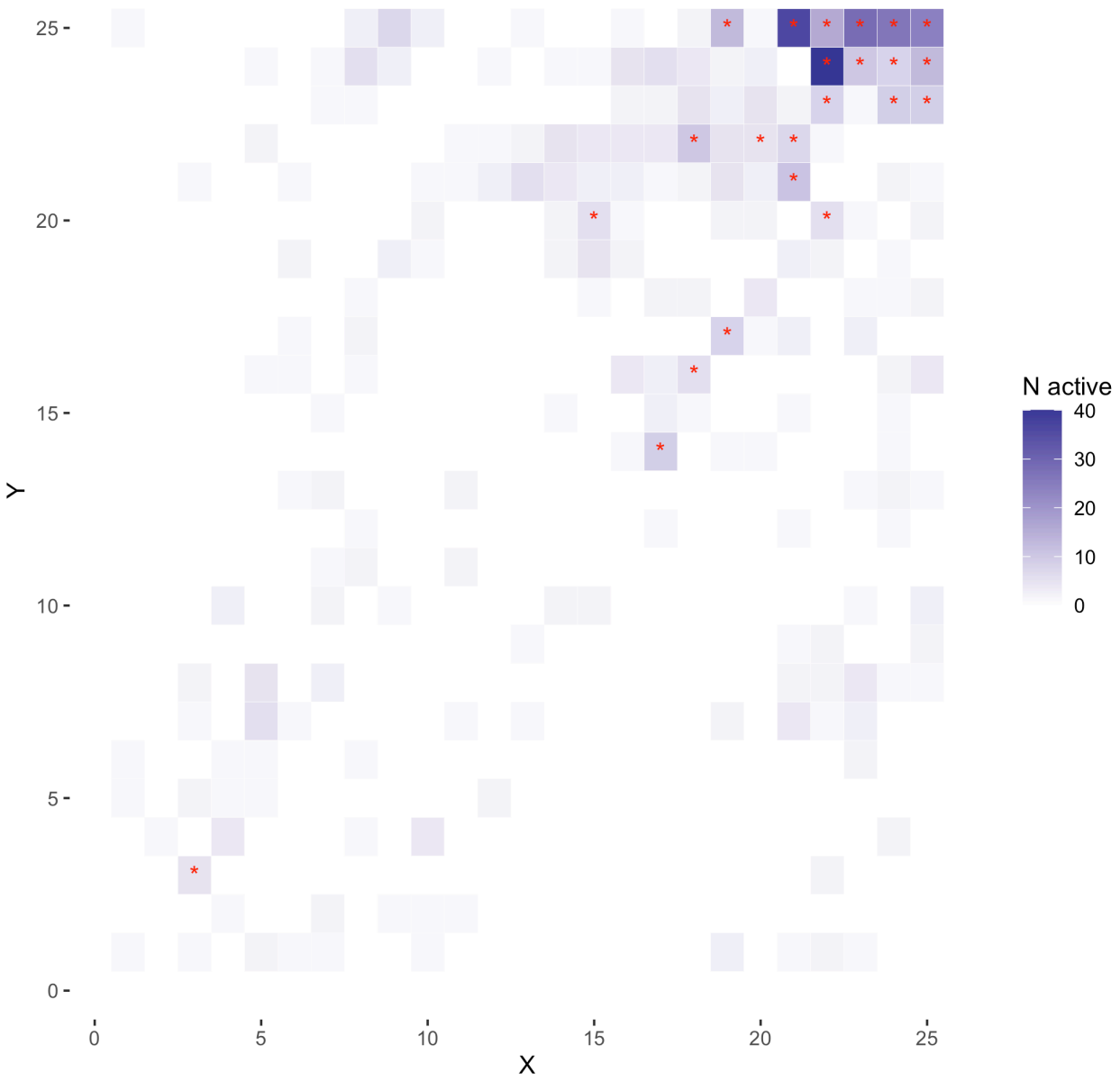
abl1



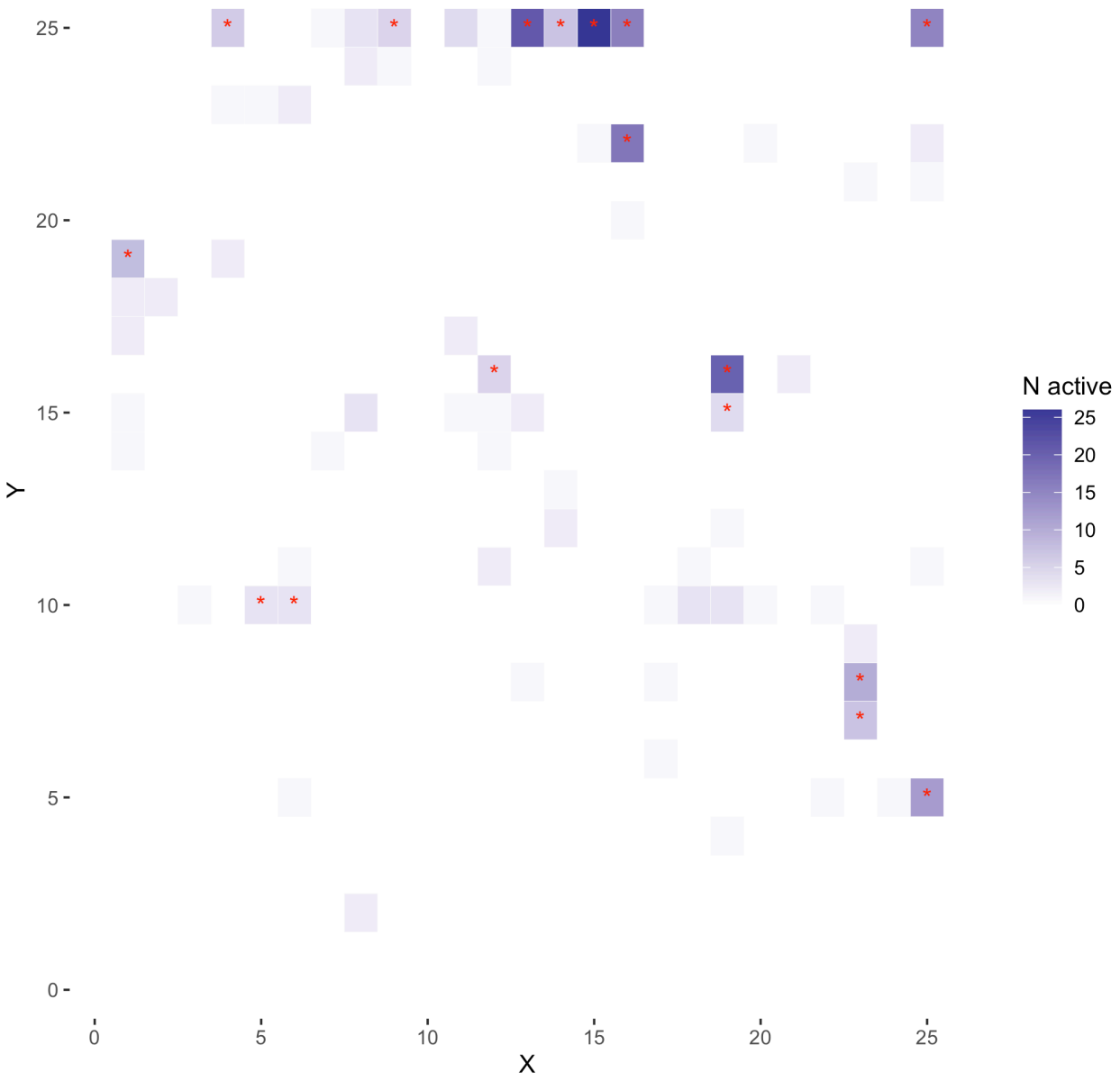
ace



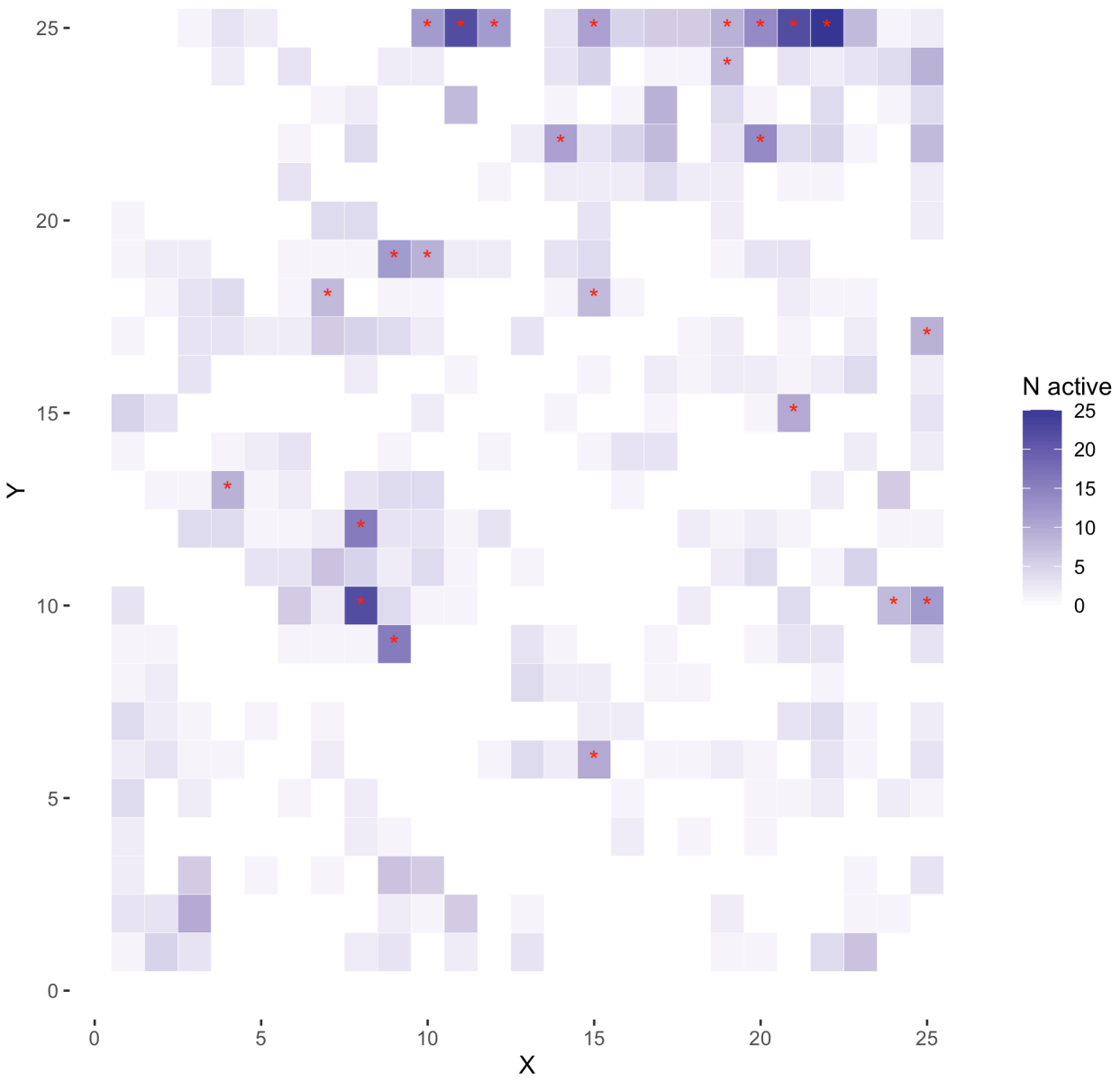
aces



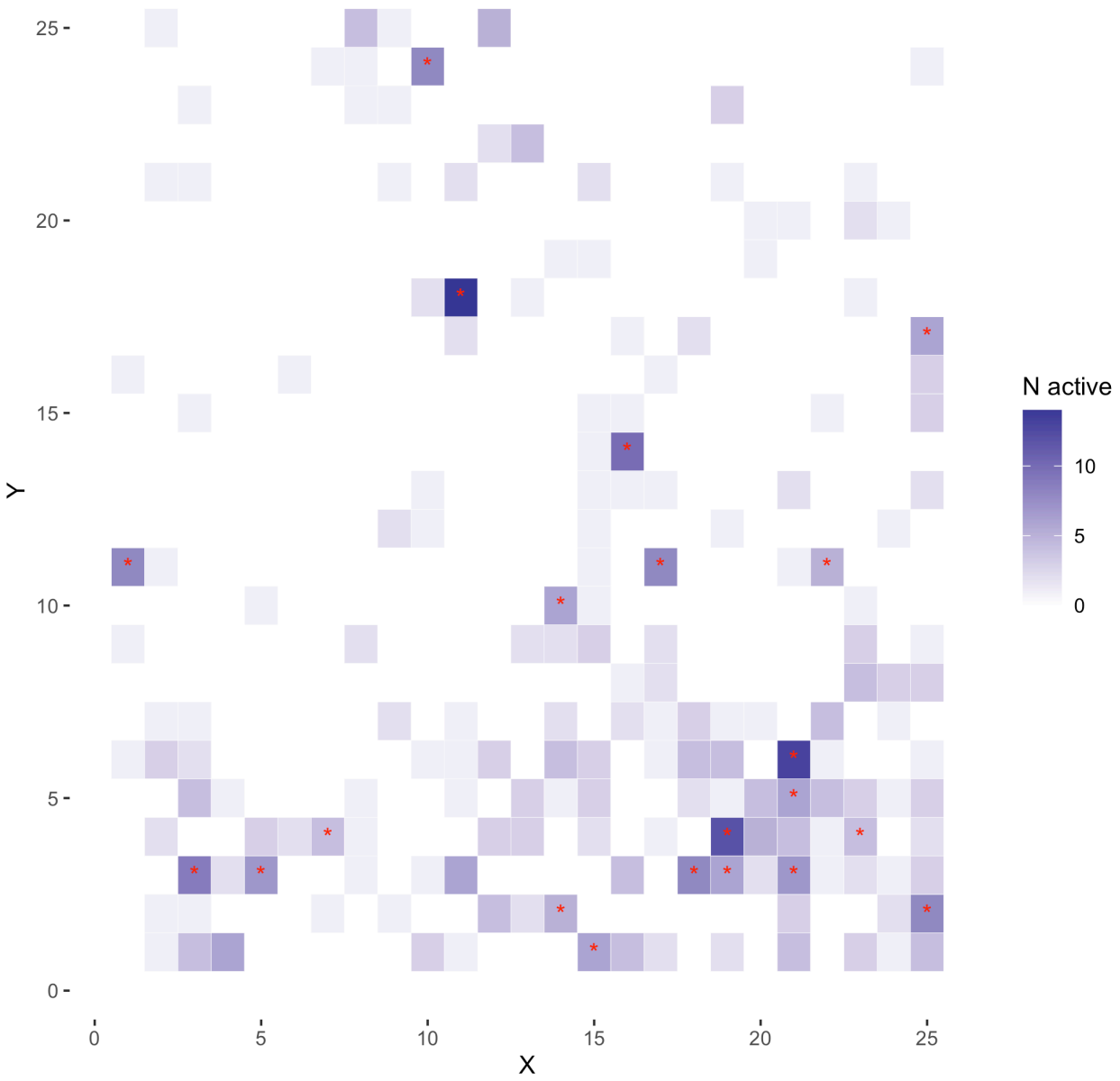
ada


N active
25
20
15
10
5
0
Y
X
0
5
10
15
20
25

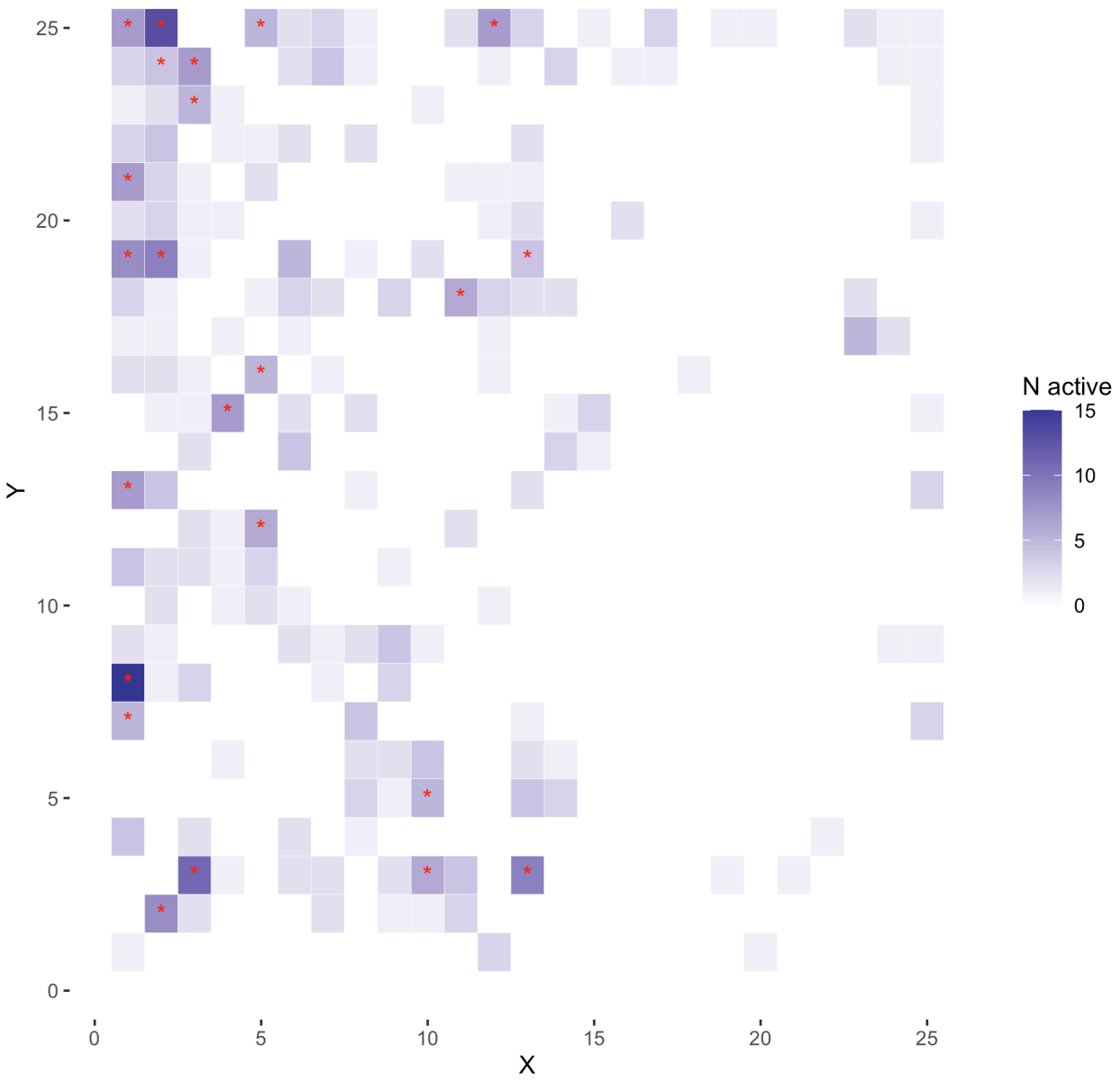
ada17



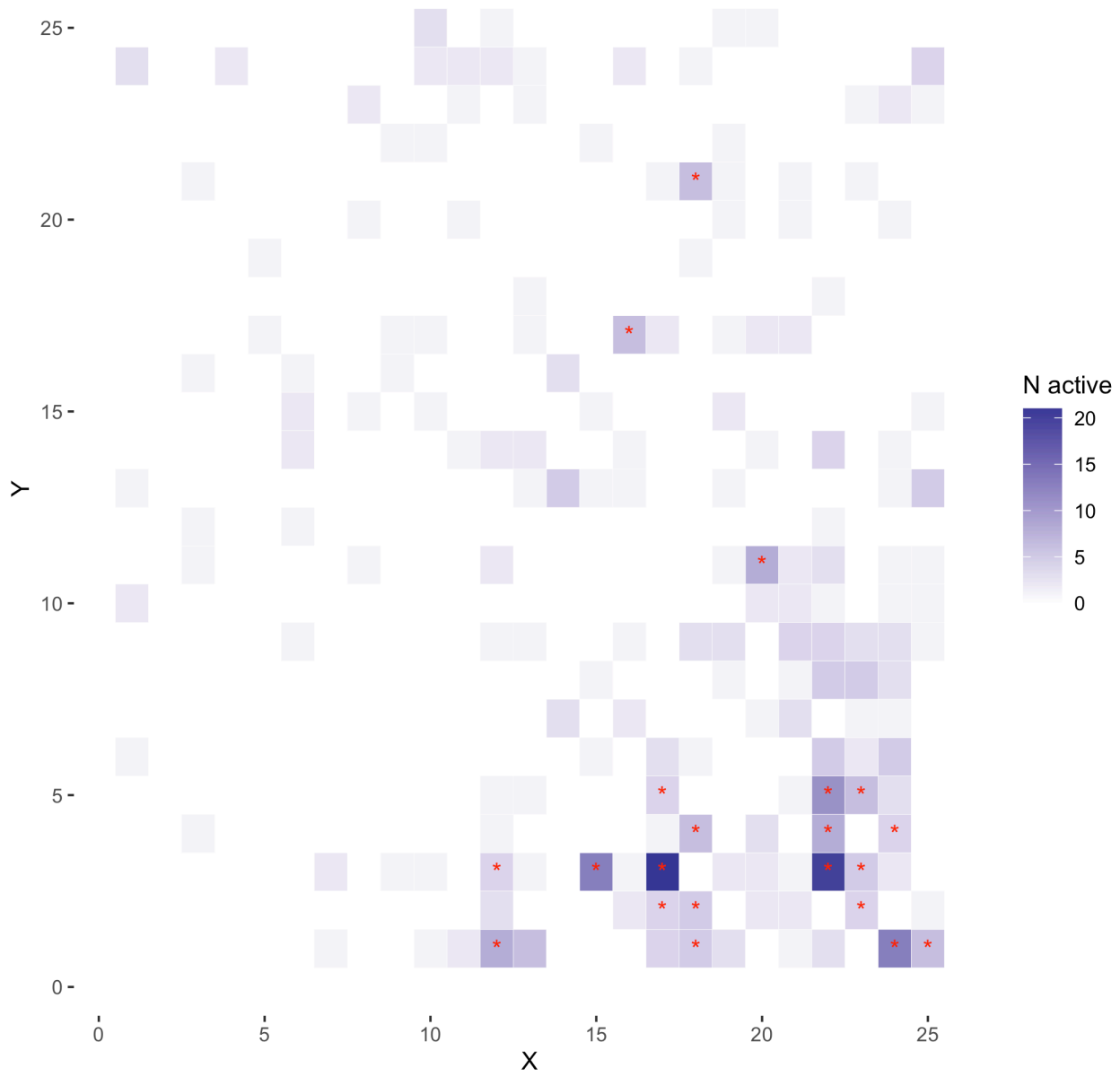
adrb1



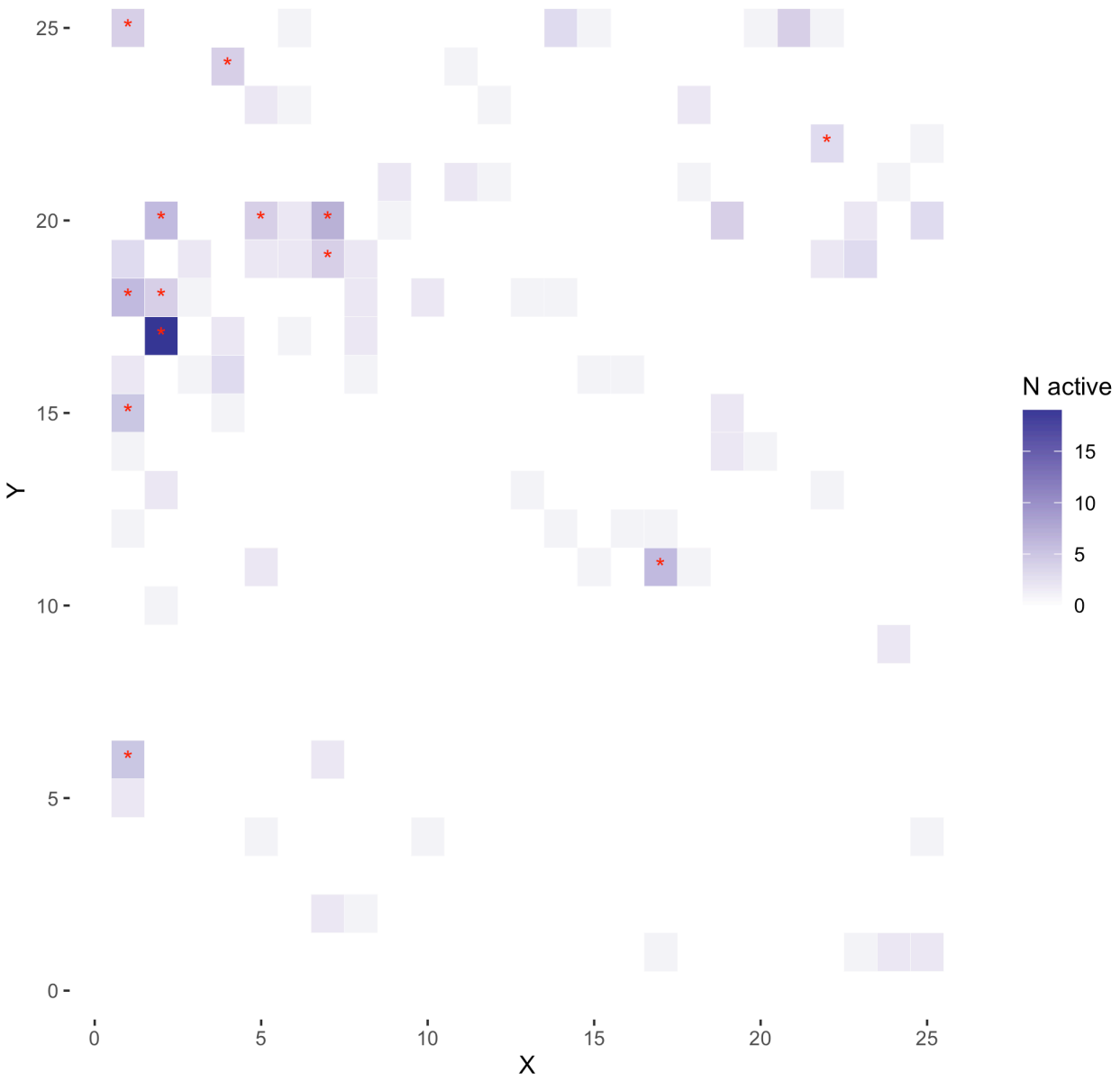
adrb2



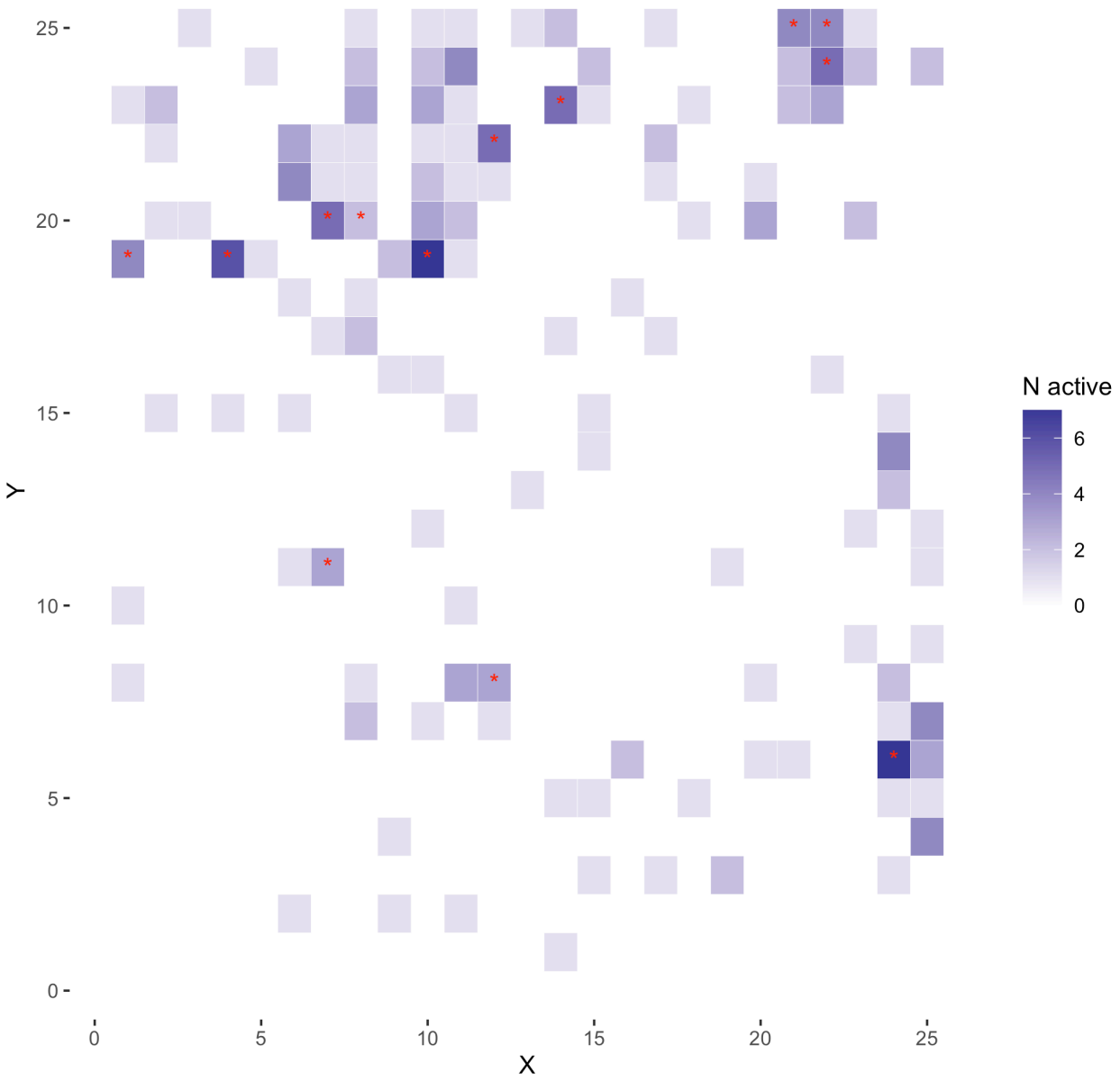
akt1

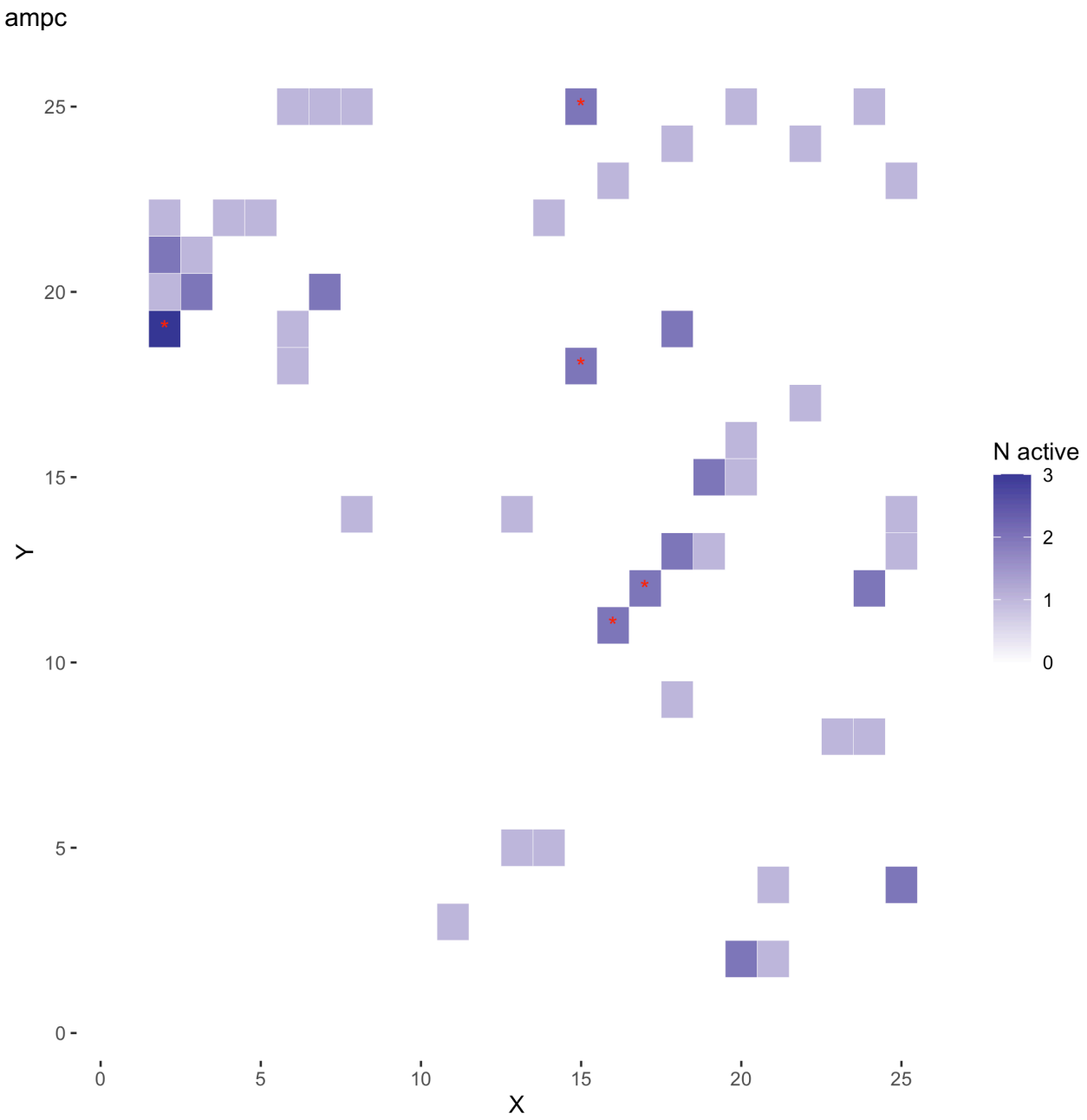


akt2

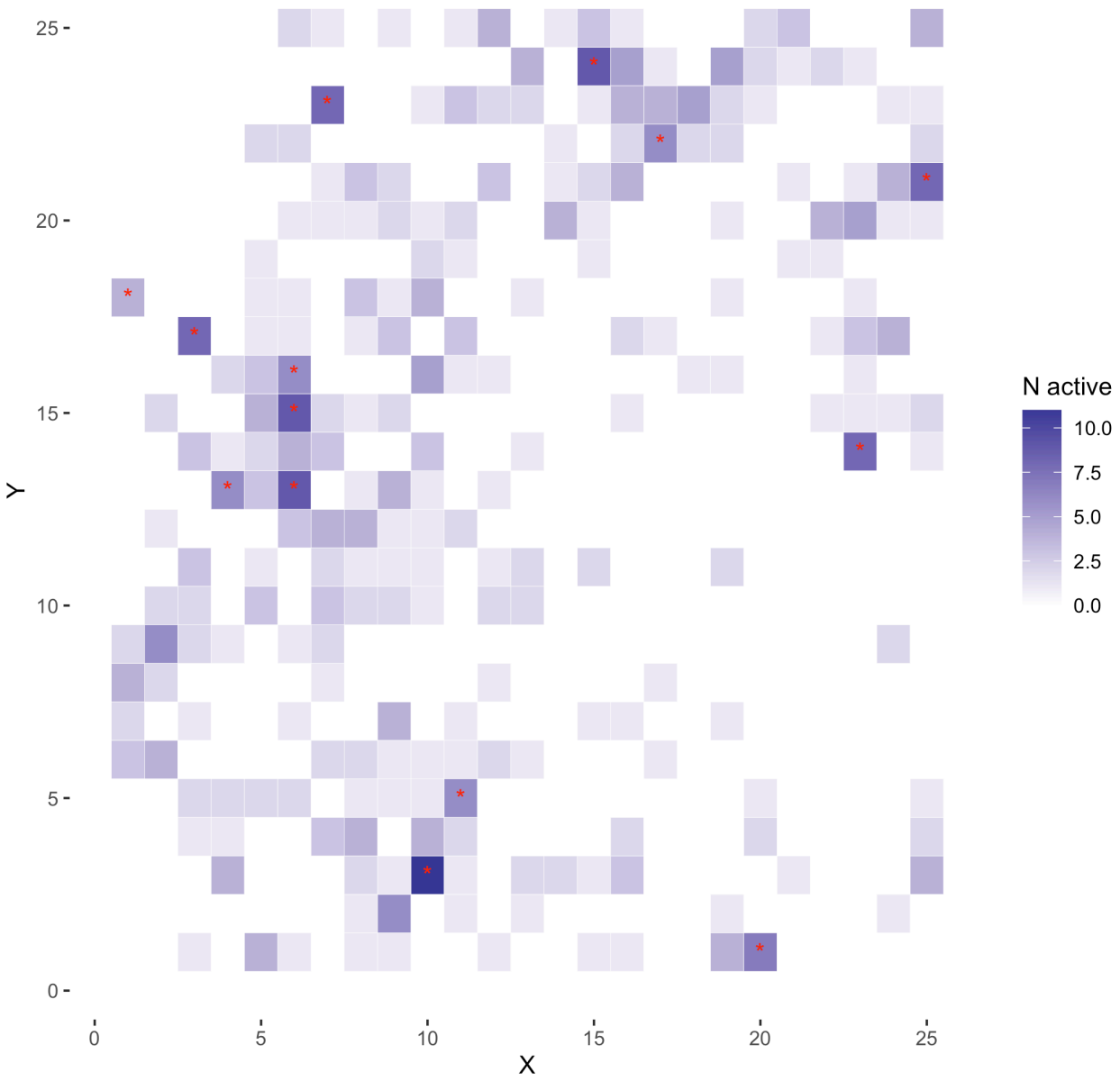


aldr

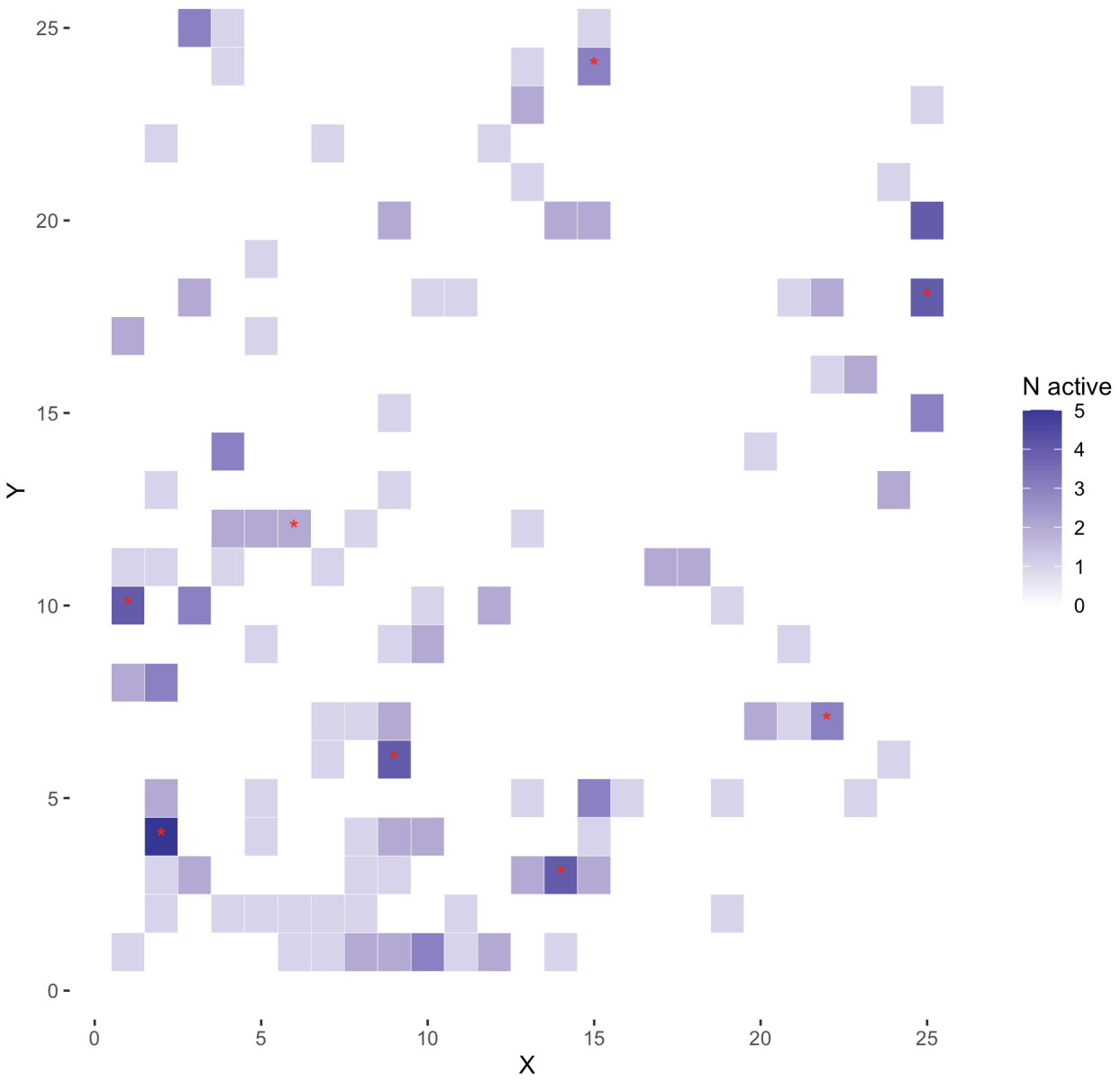



ampc
25
20
15
10
5
0
Y
0
5
10
15
20
25
X
N active
3
2
1
0

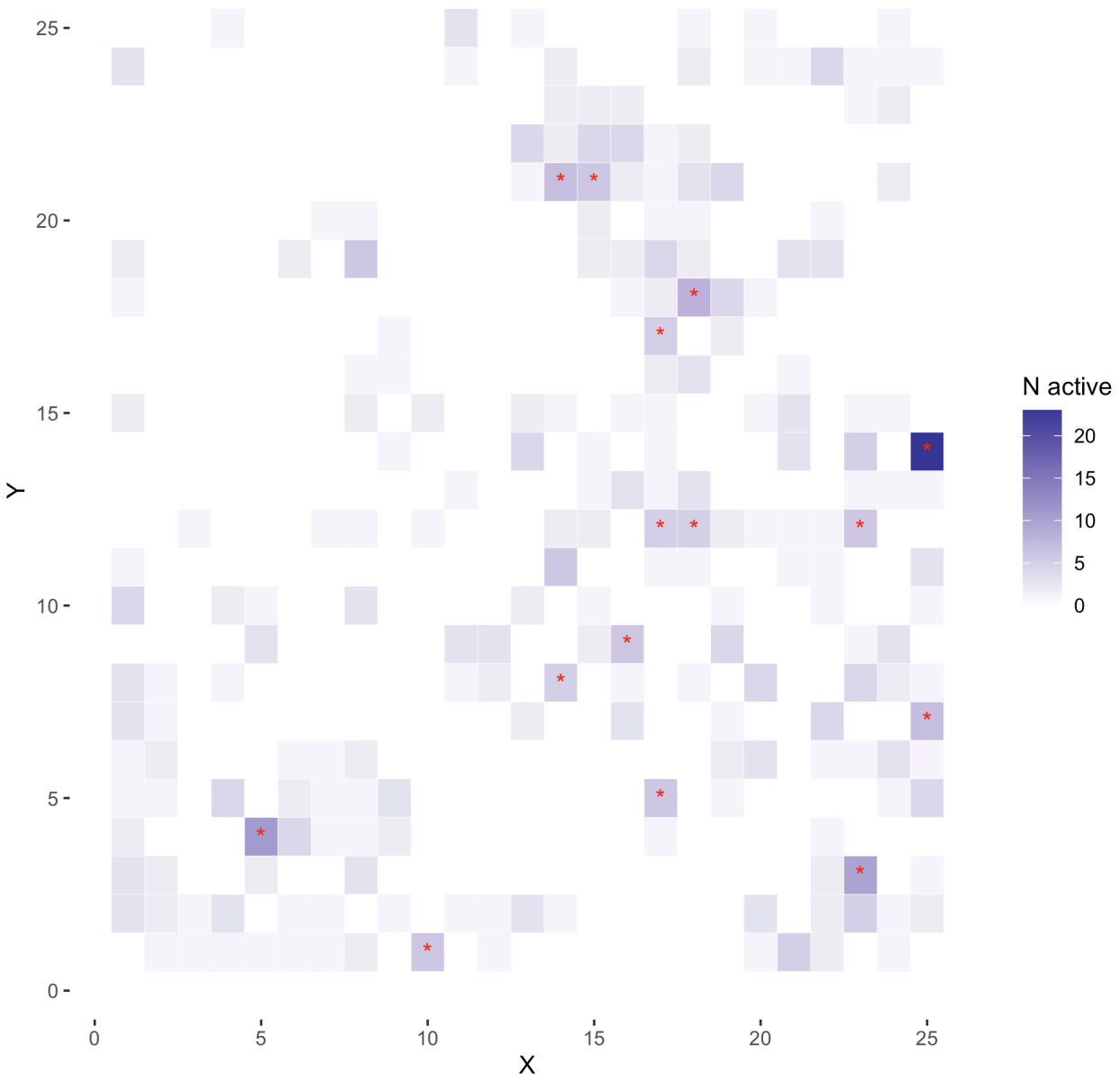
andr



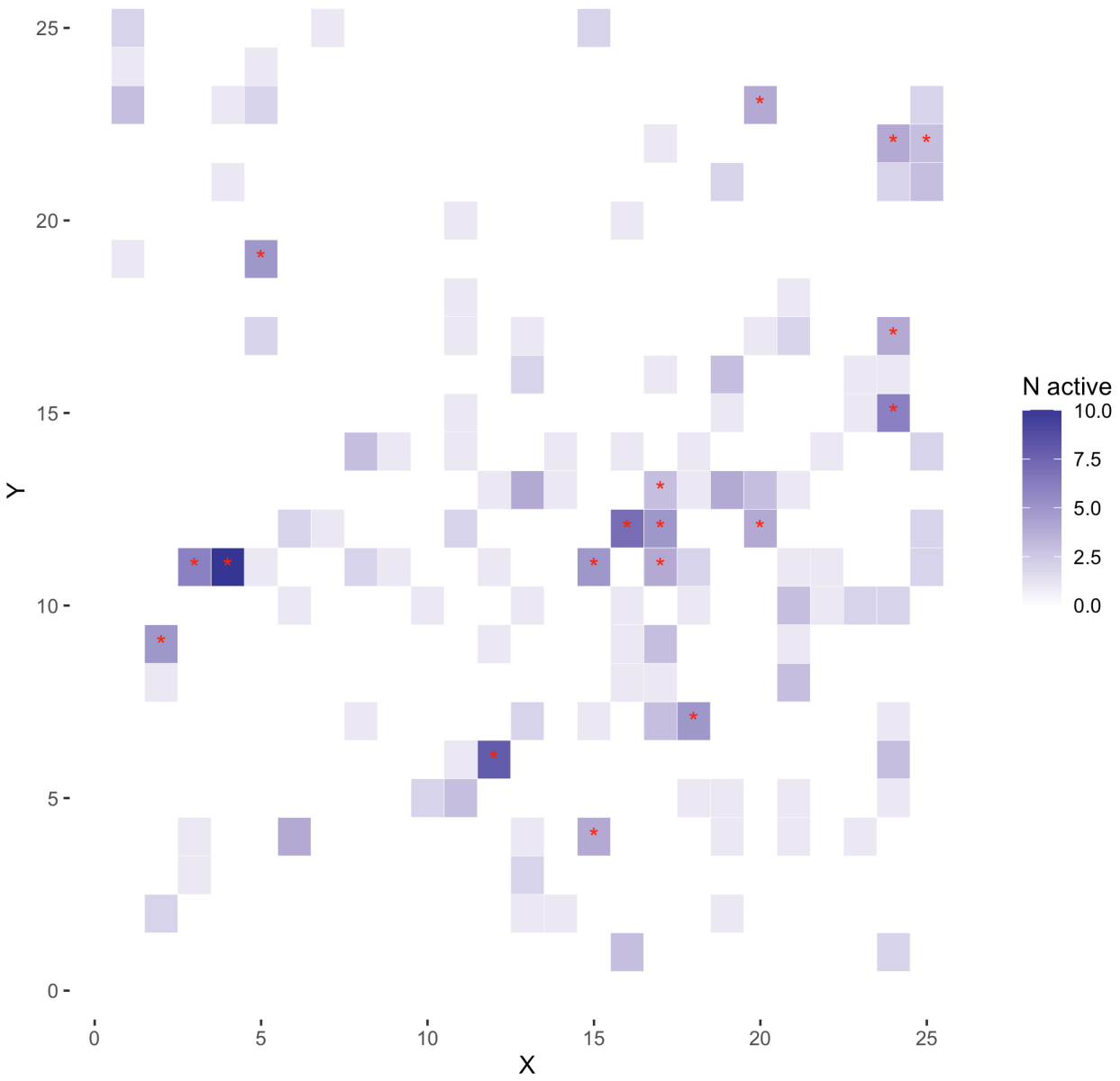
aofb



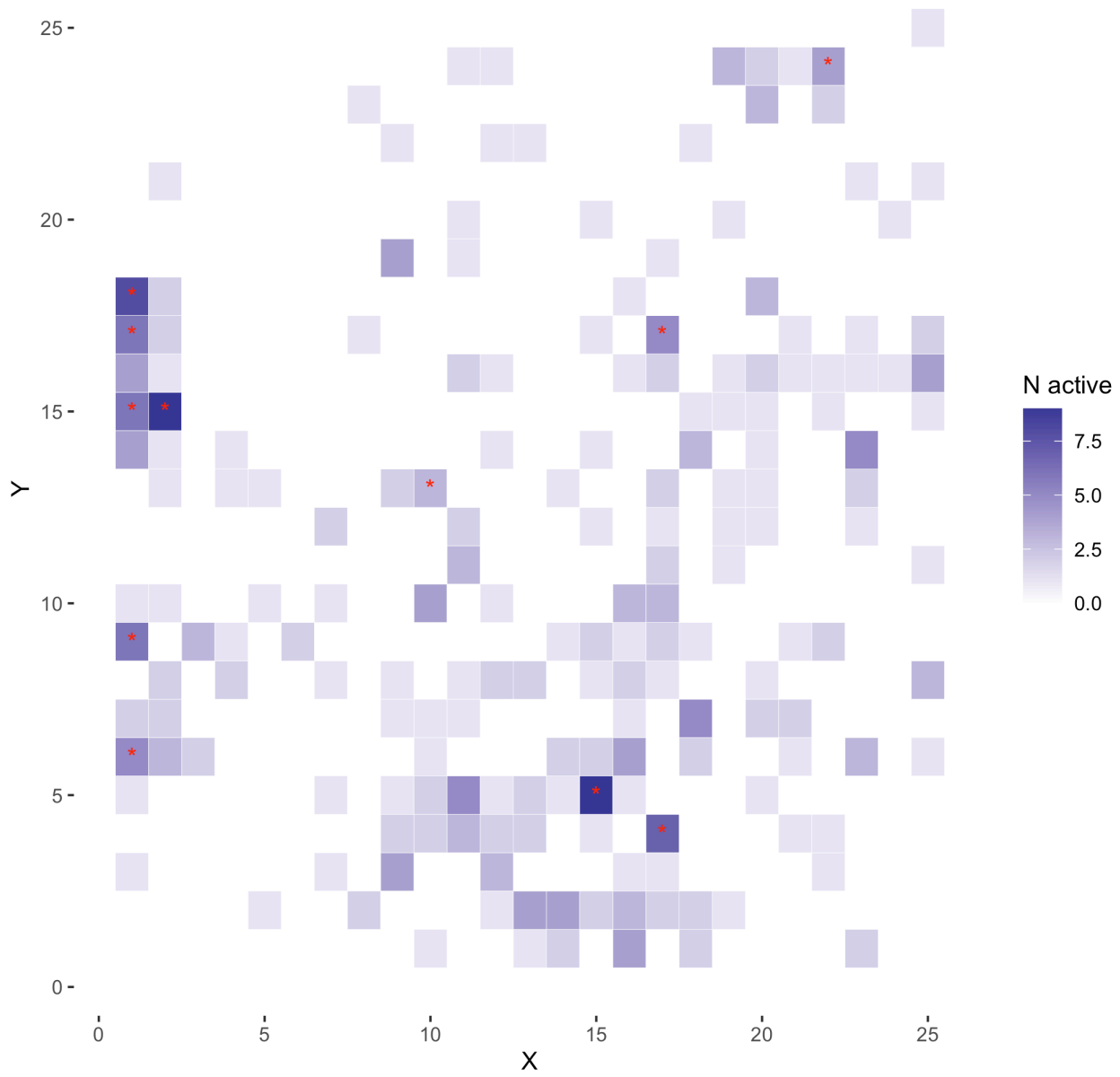
bace1



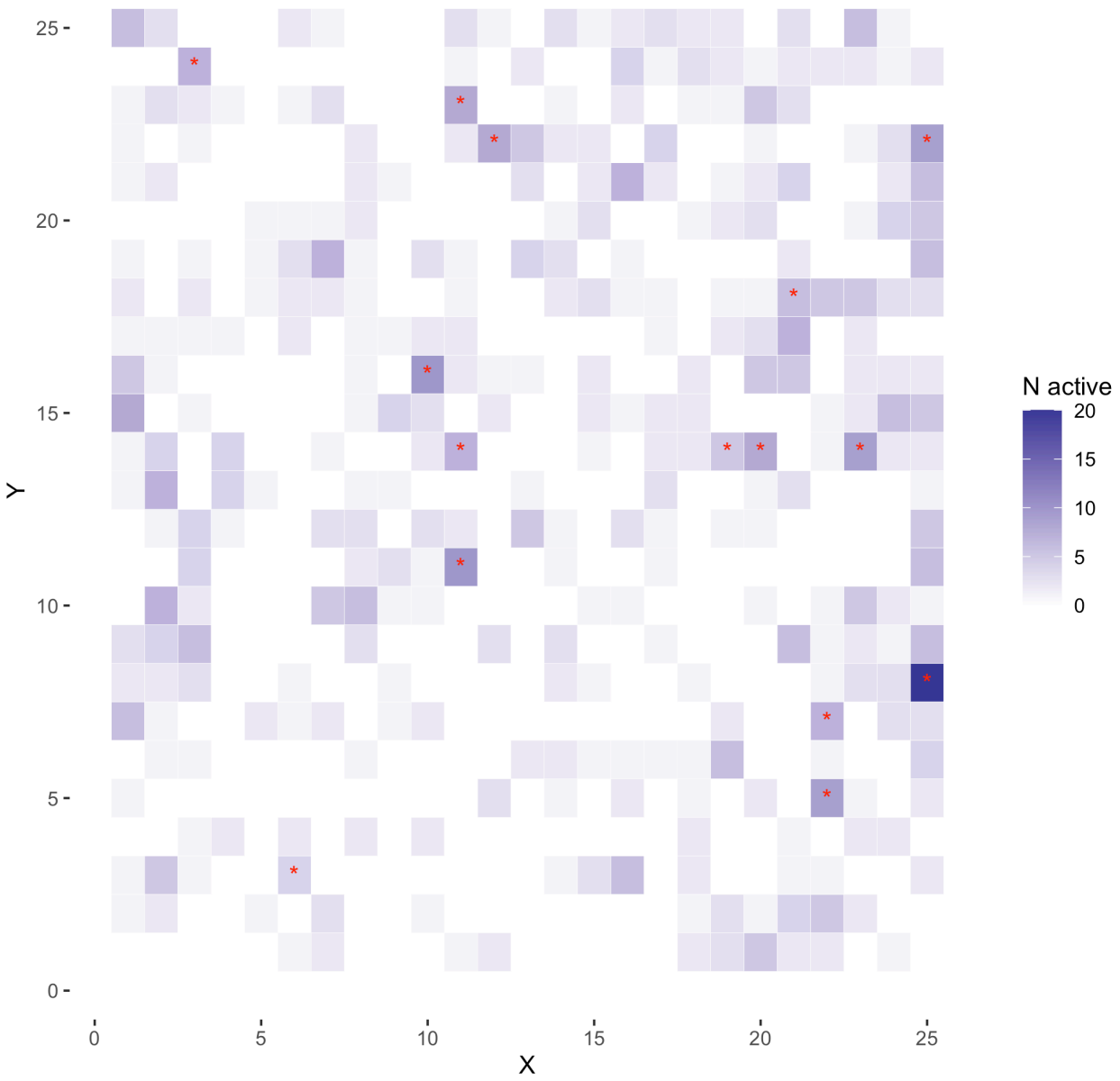
braf

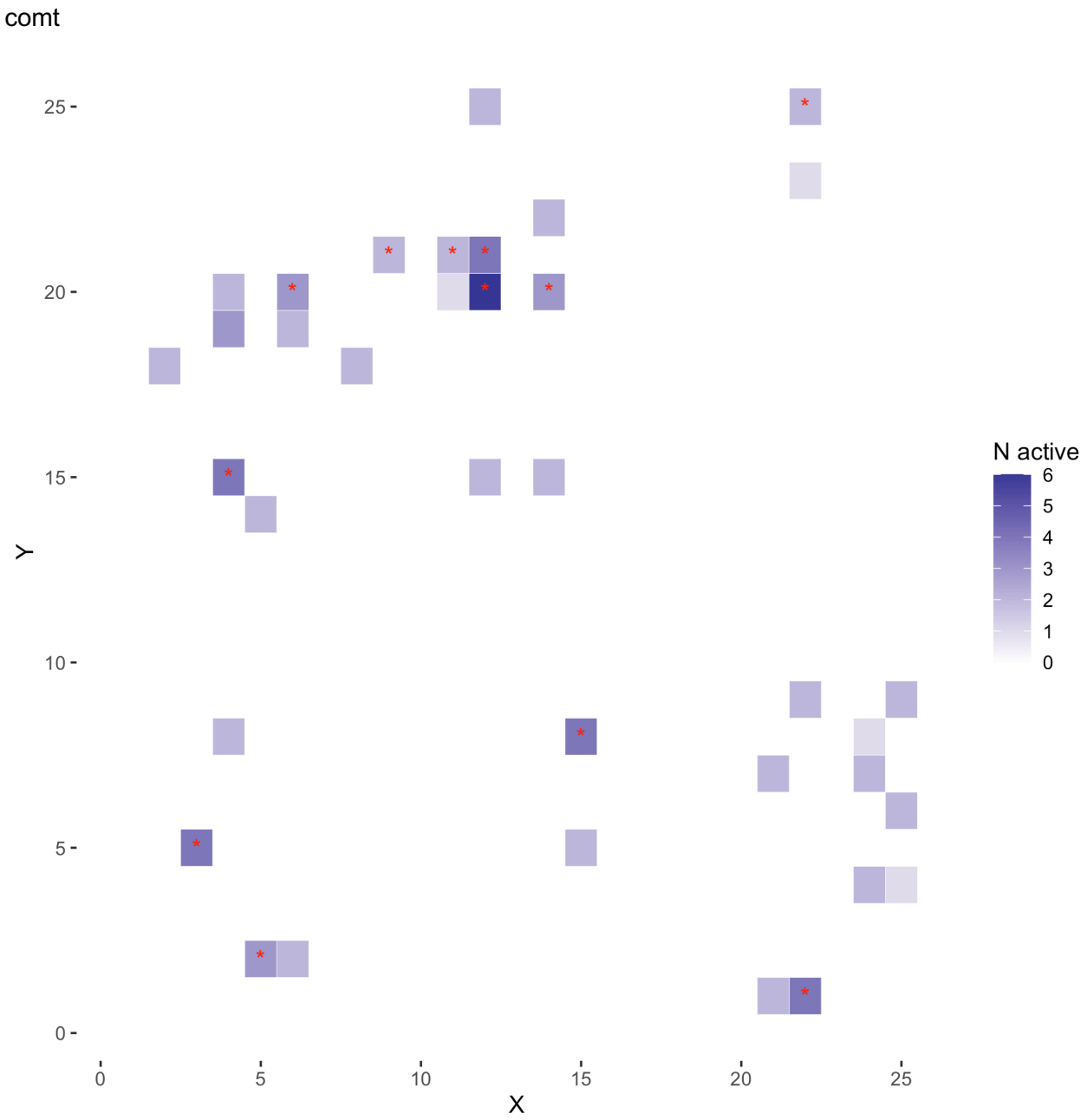


casp3

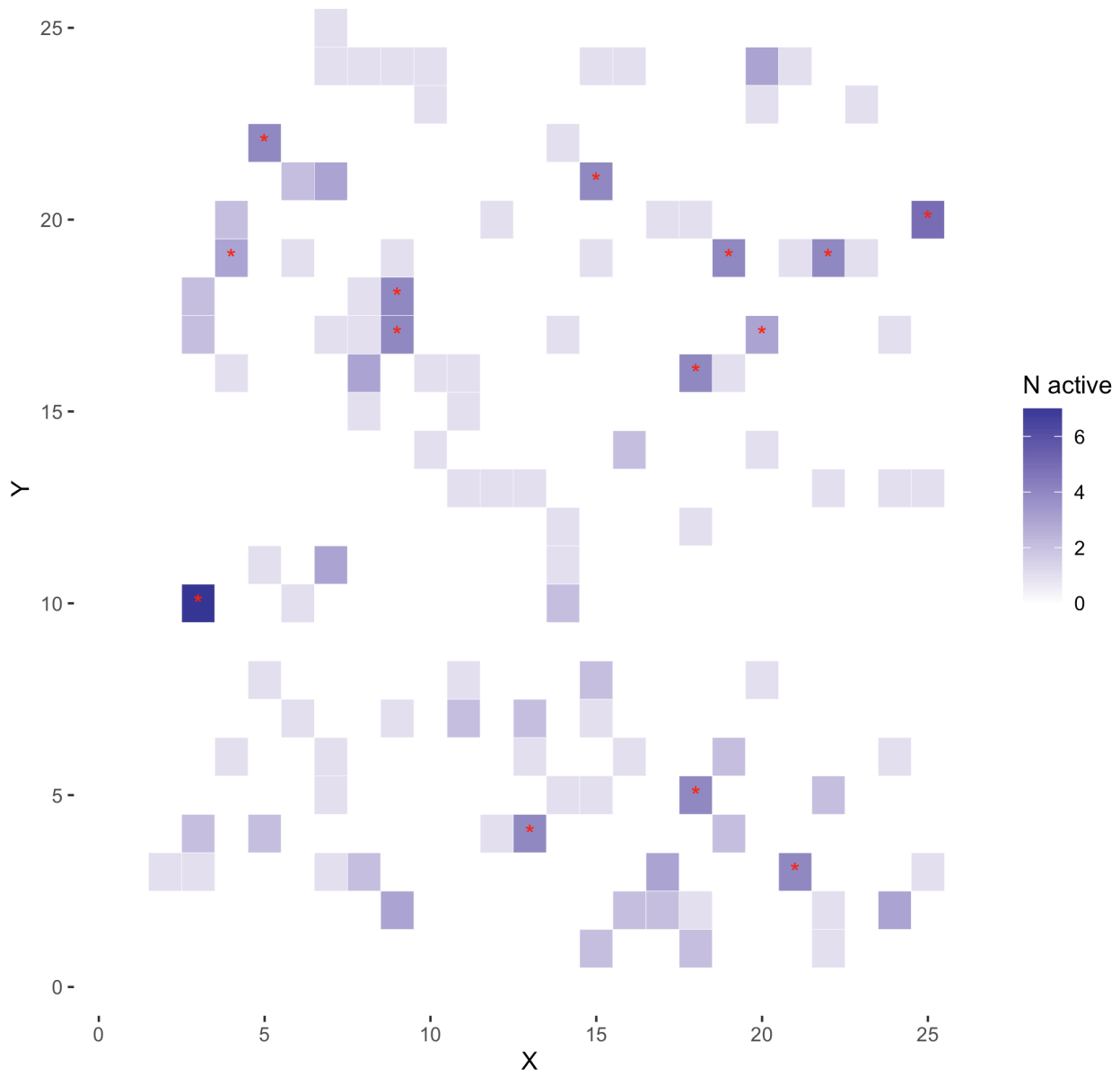


cdk2

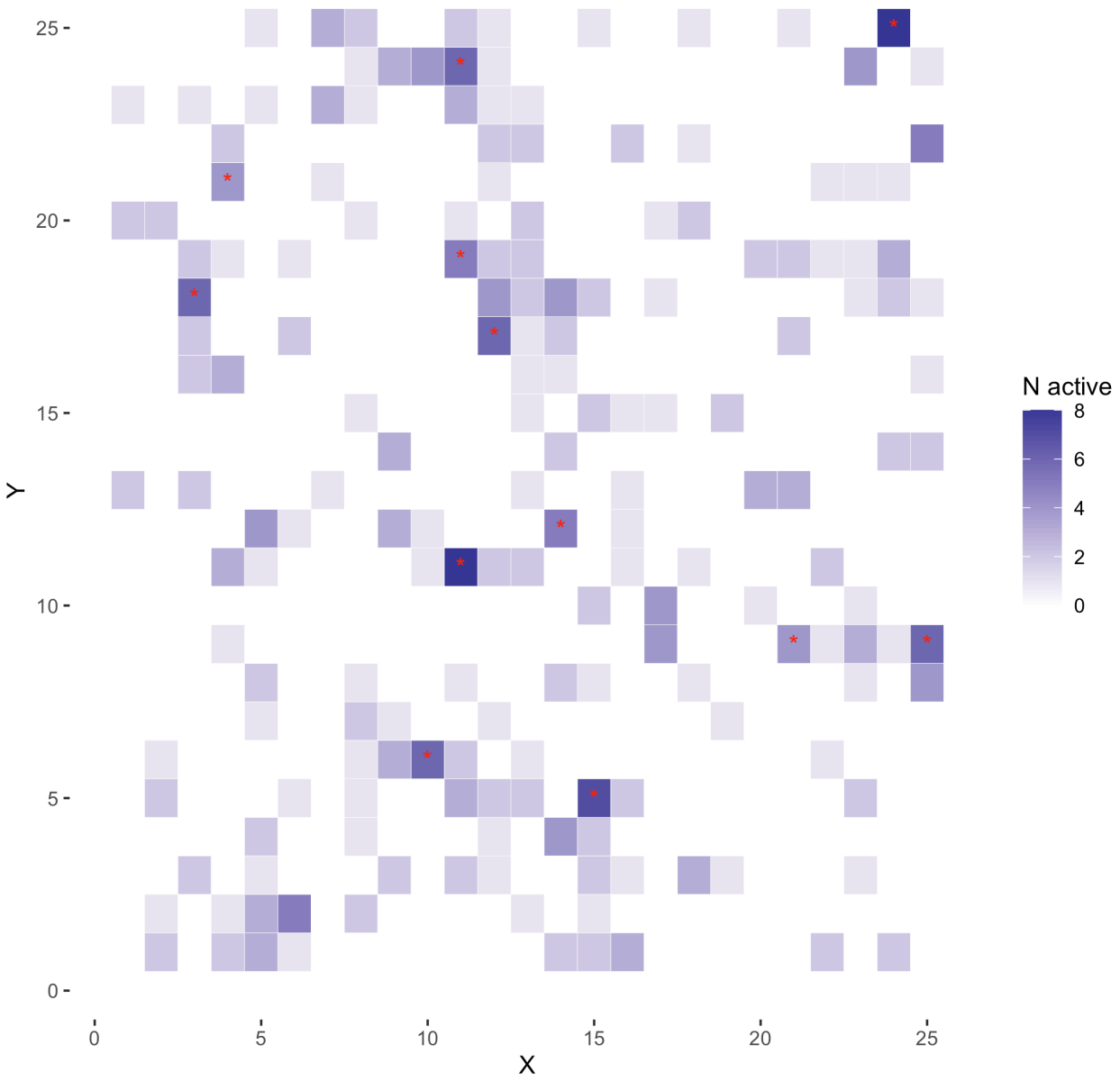



comt
N active
6
5
4
3
2
1
0
Y
X
0
5
10
15
20
25

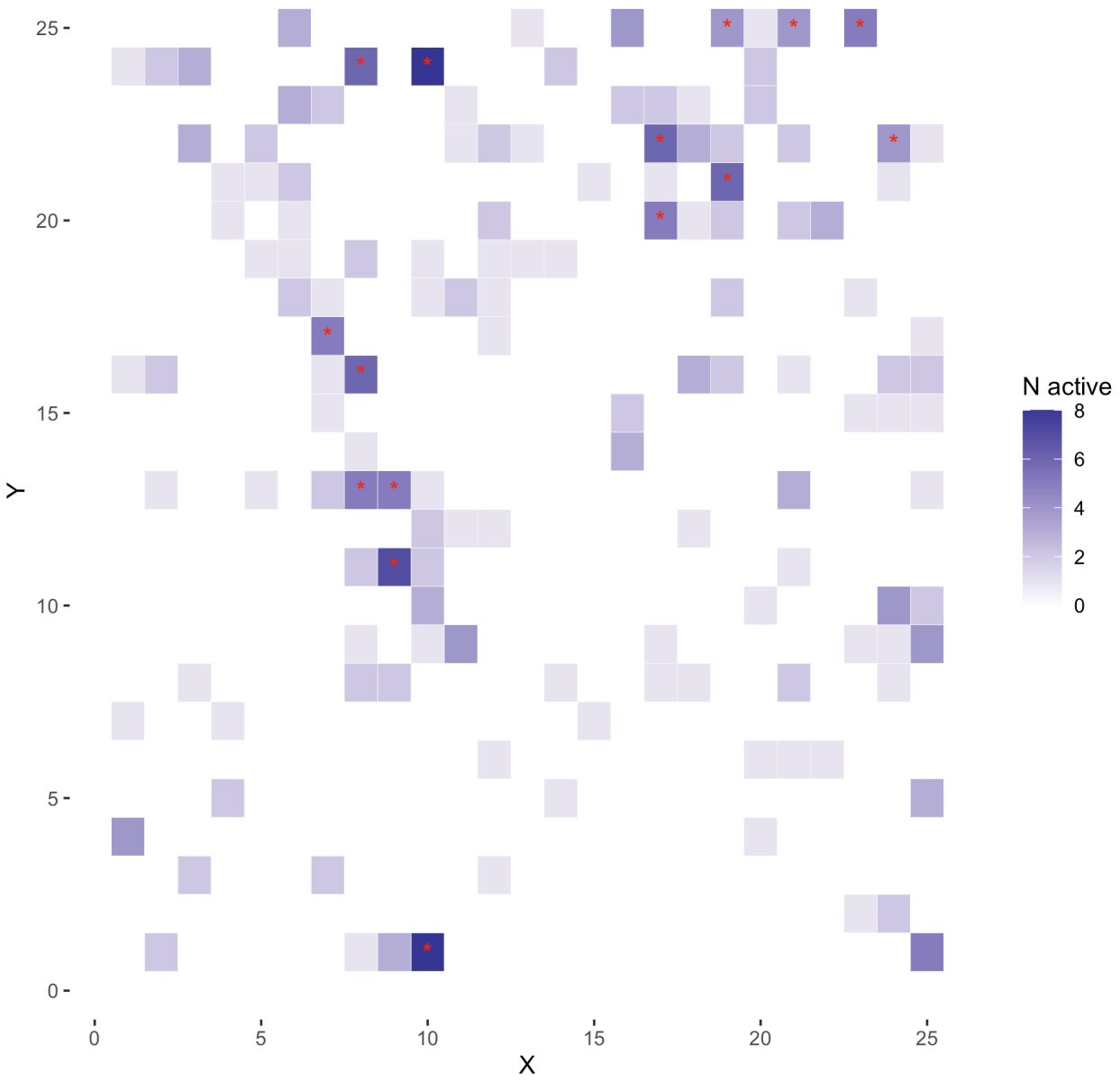
cp2c9



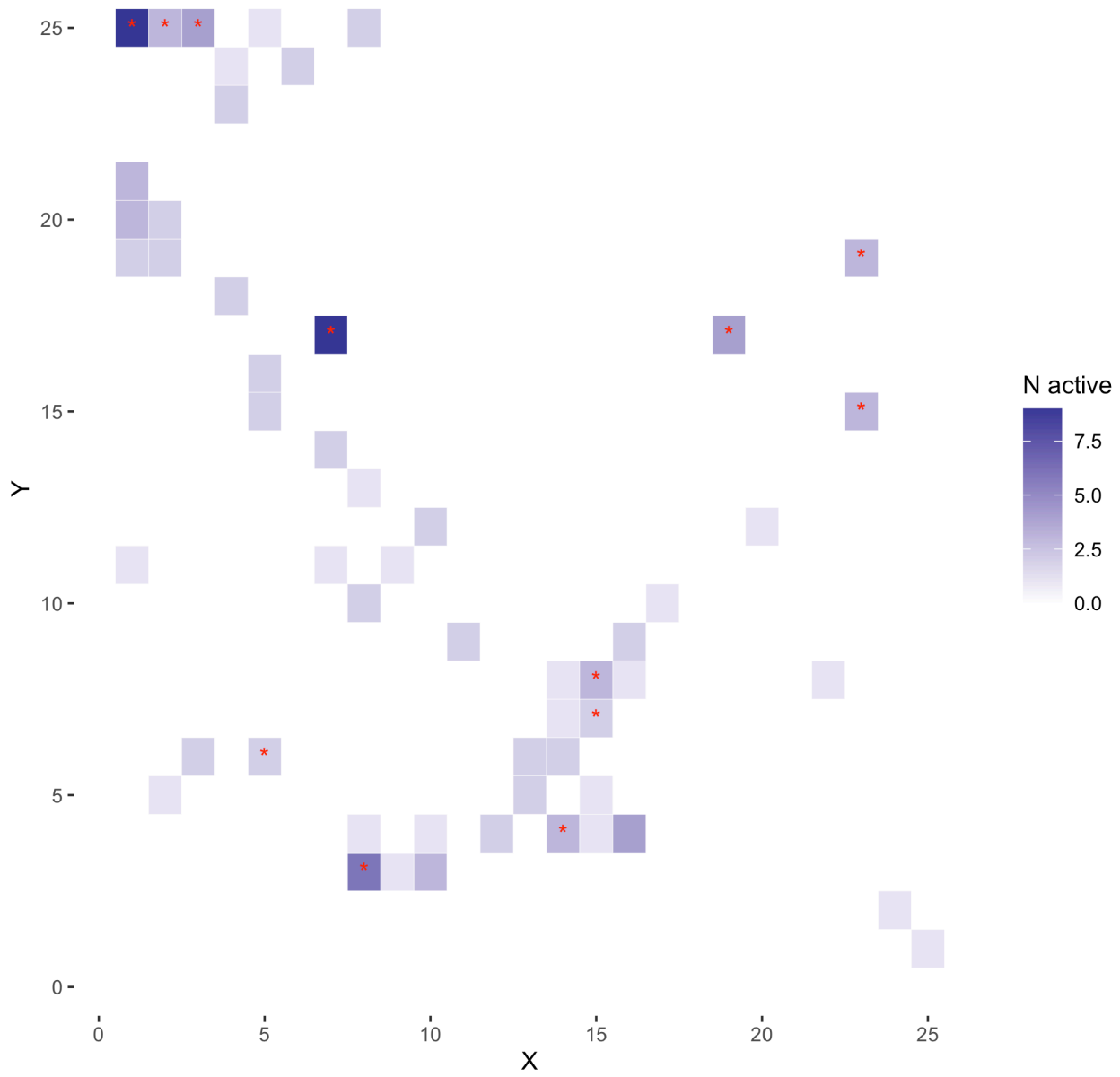
cp3a4



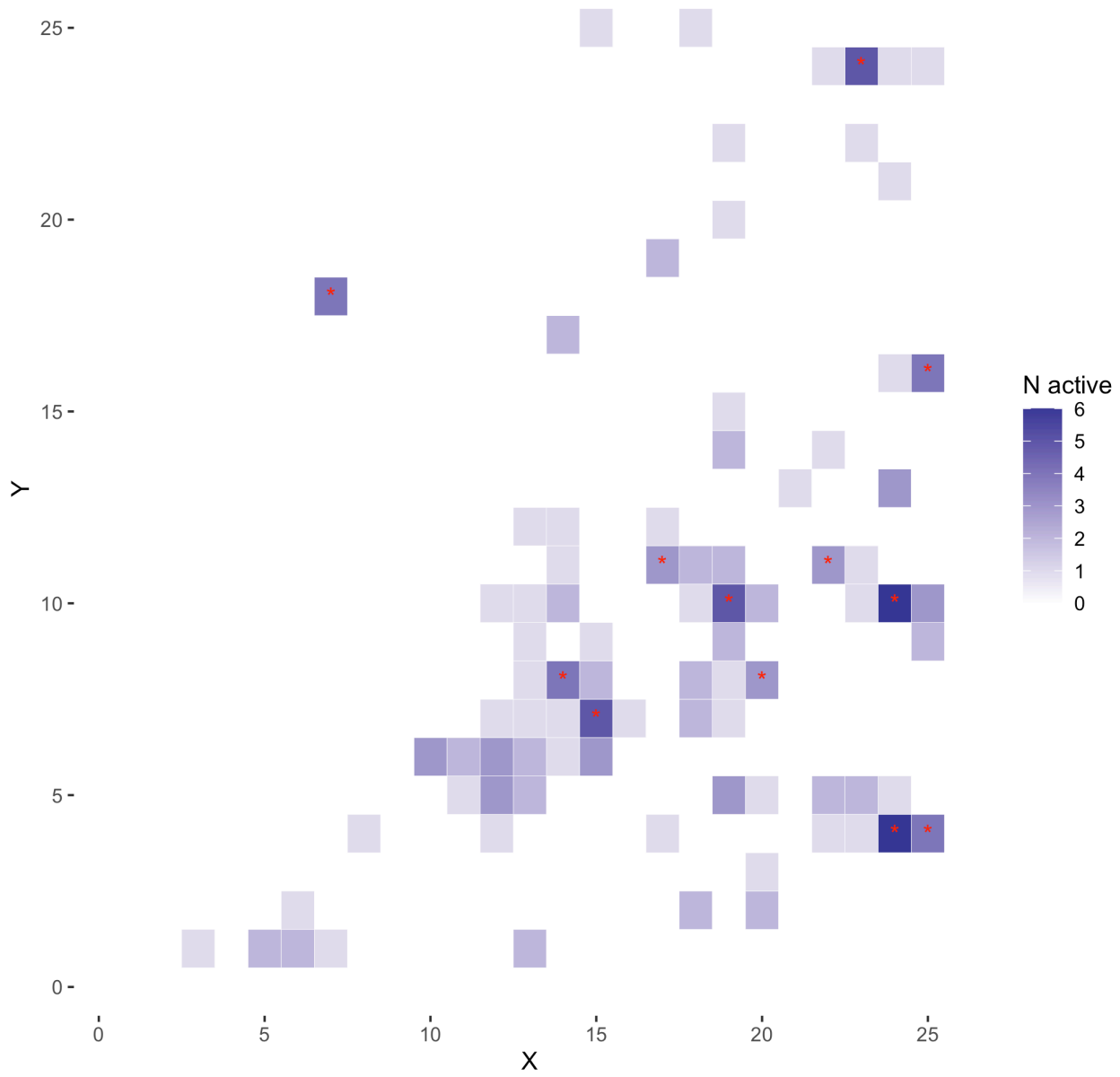
csf1r



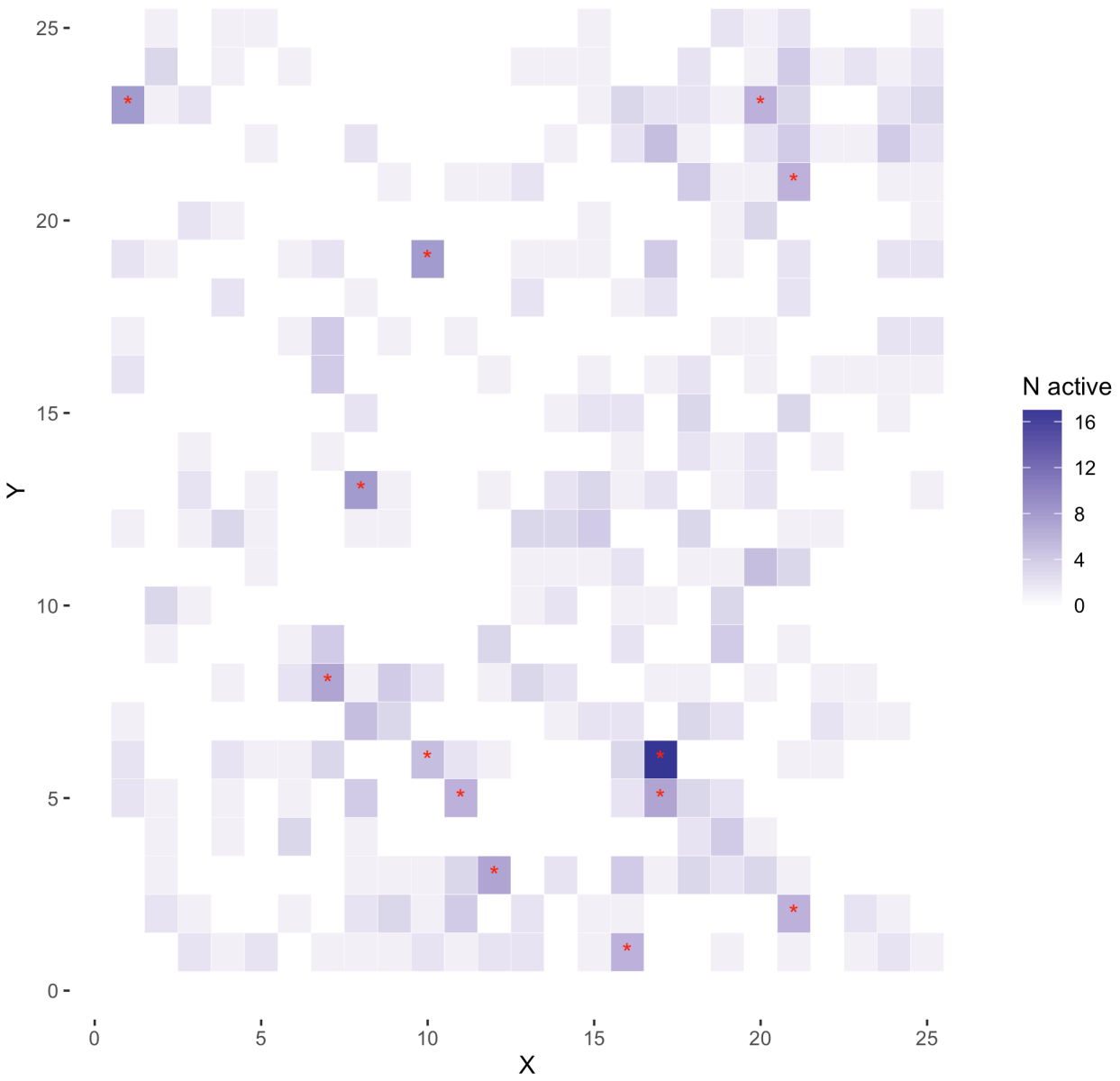
cxcr4



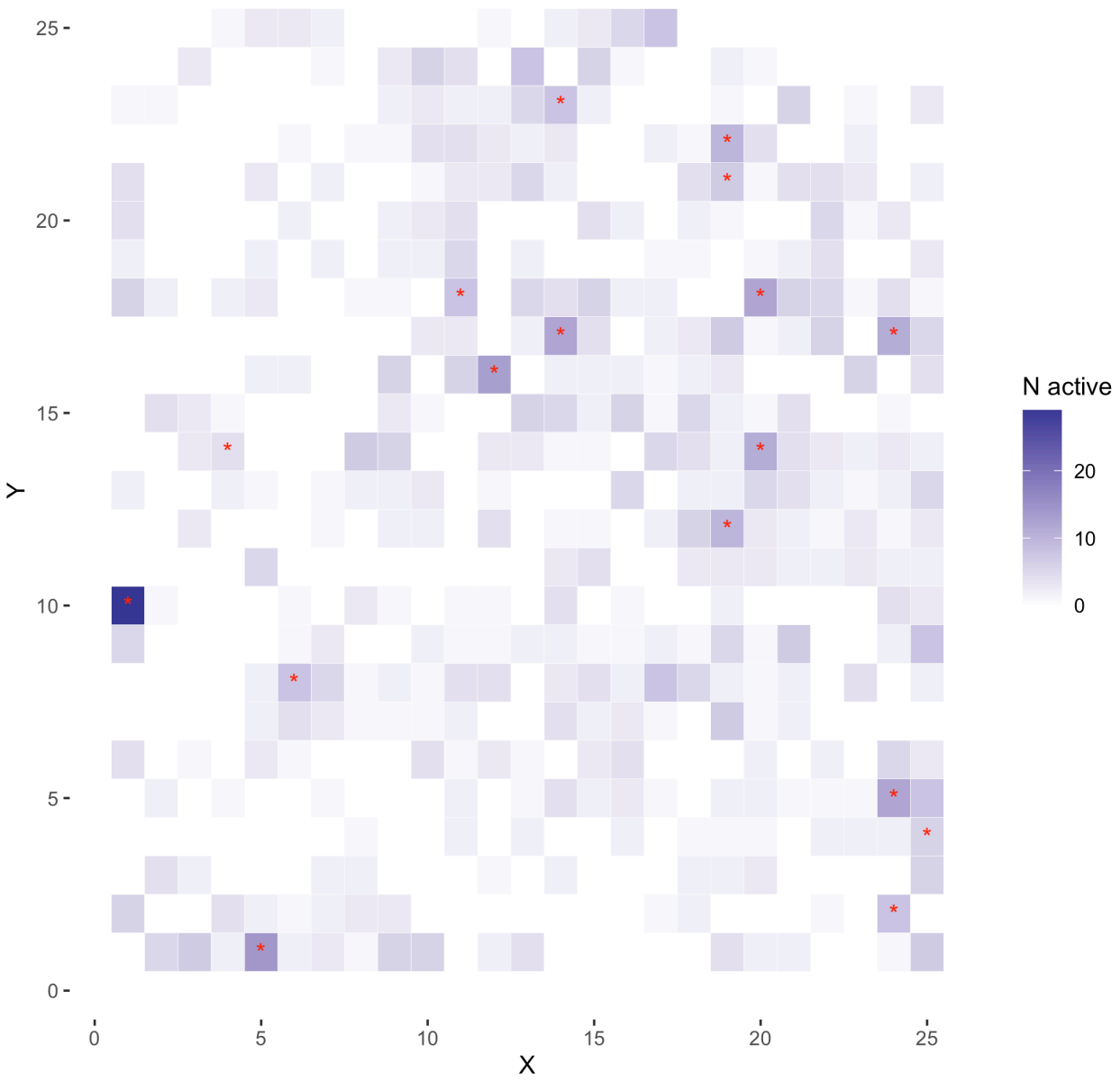
def



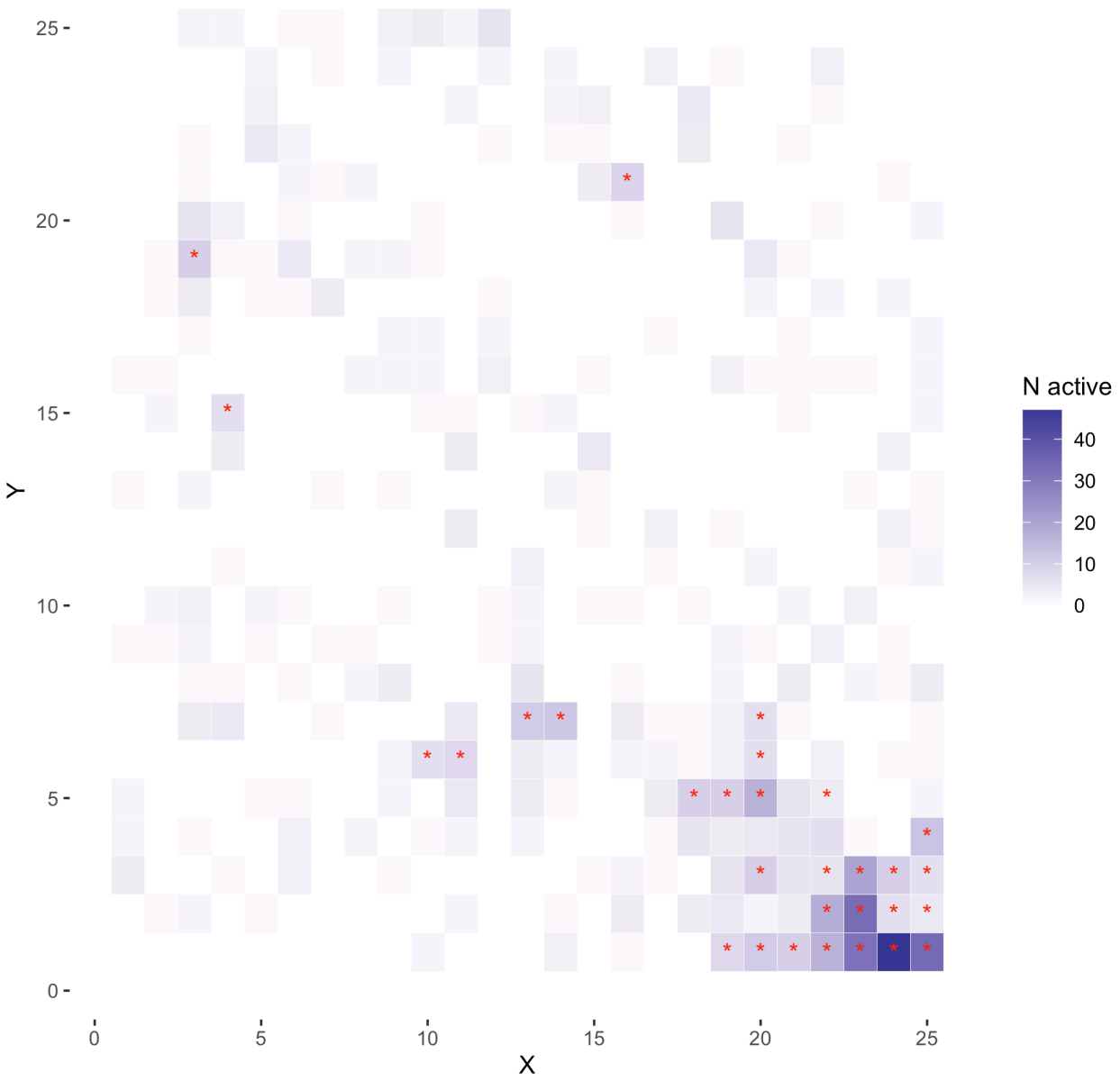
dhi1



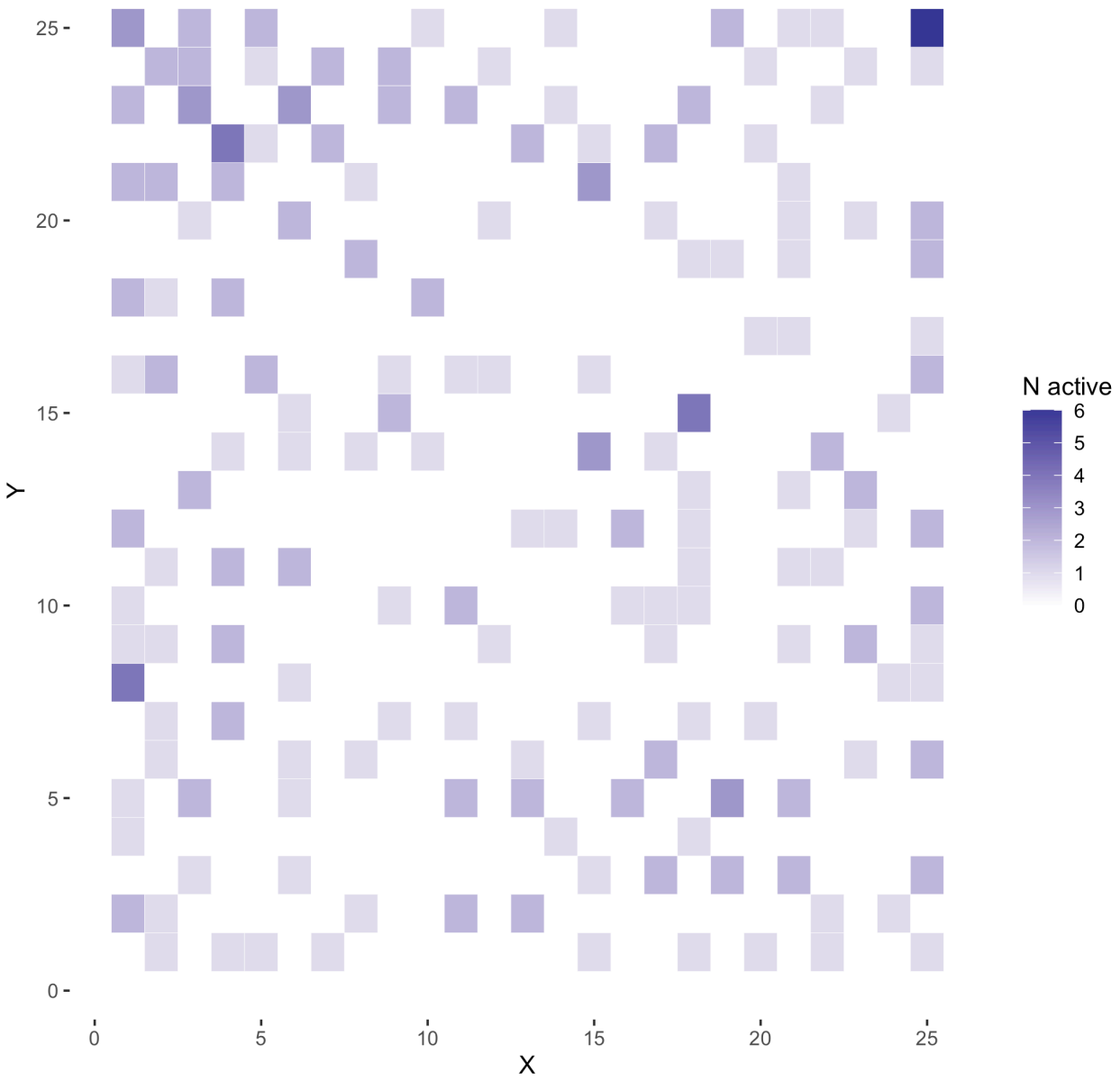
dpp4



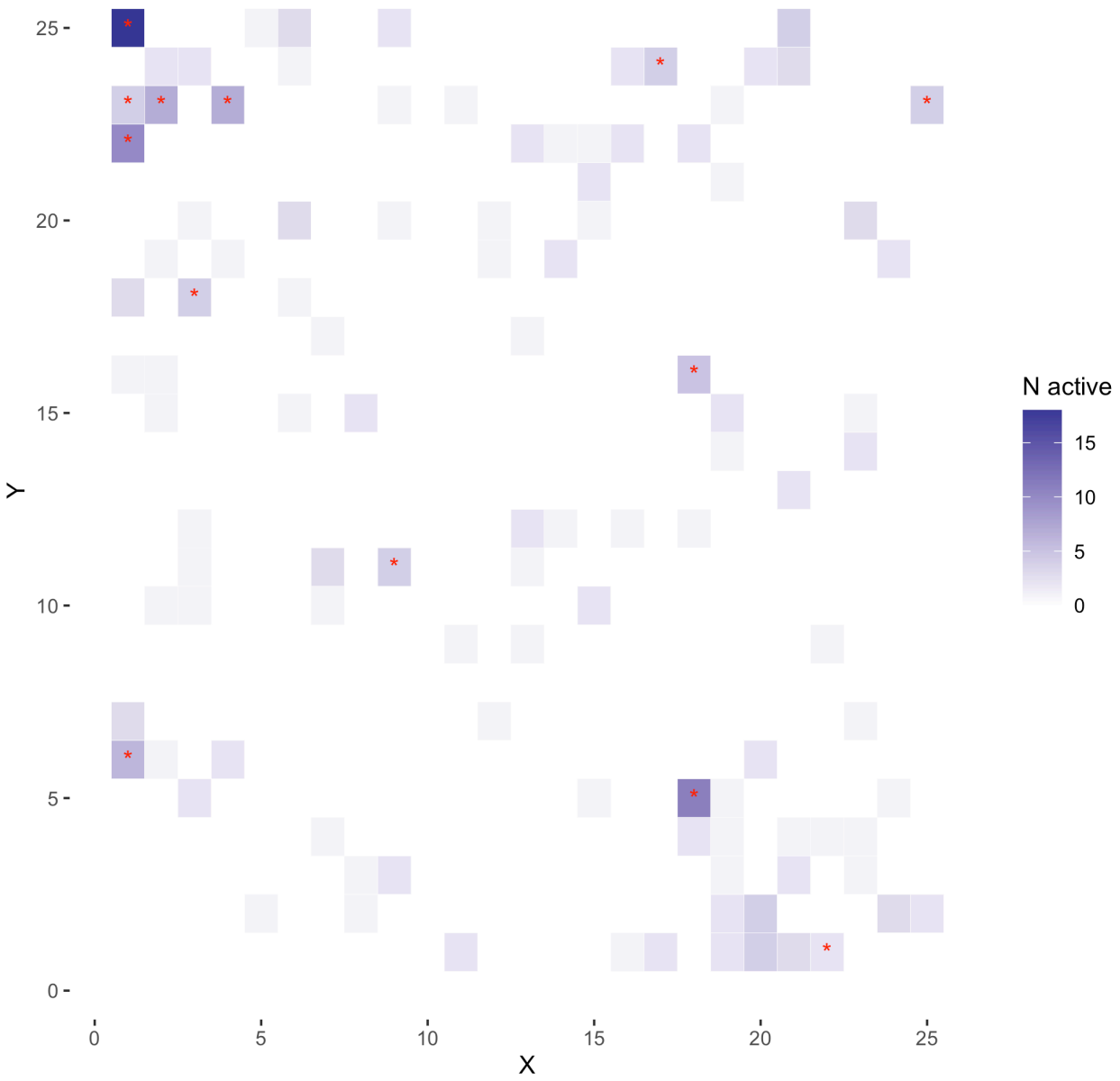
drd3



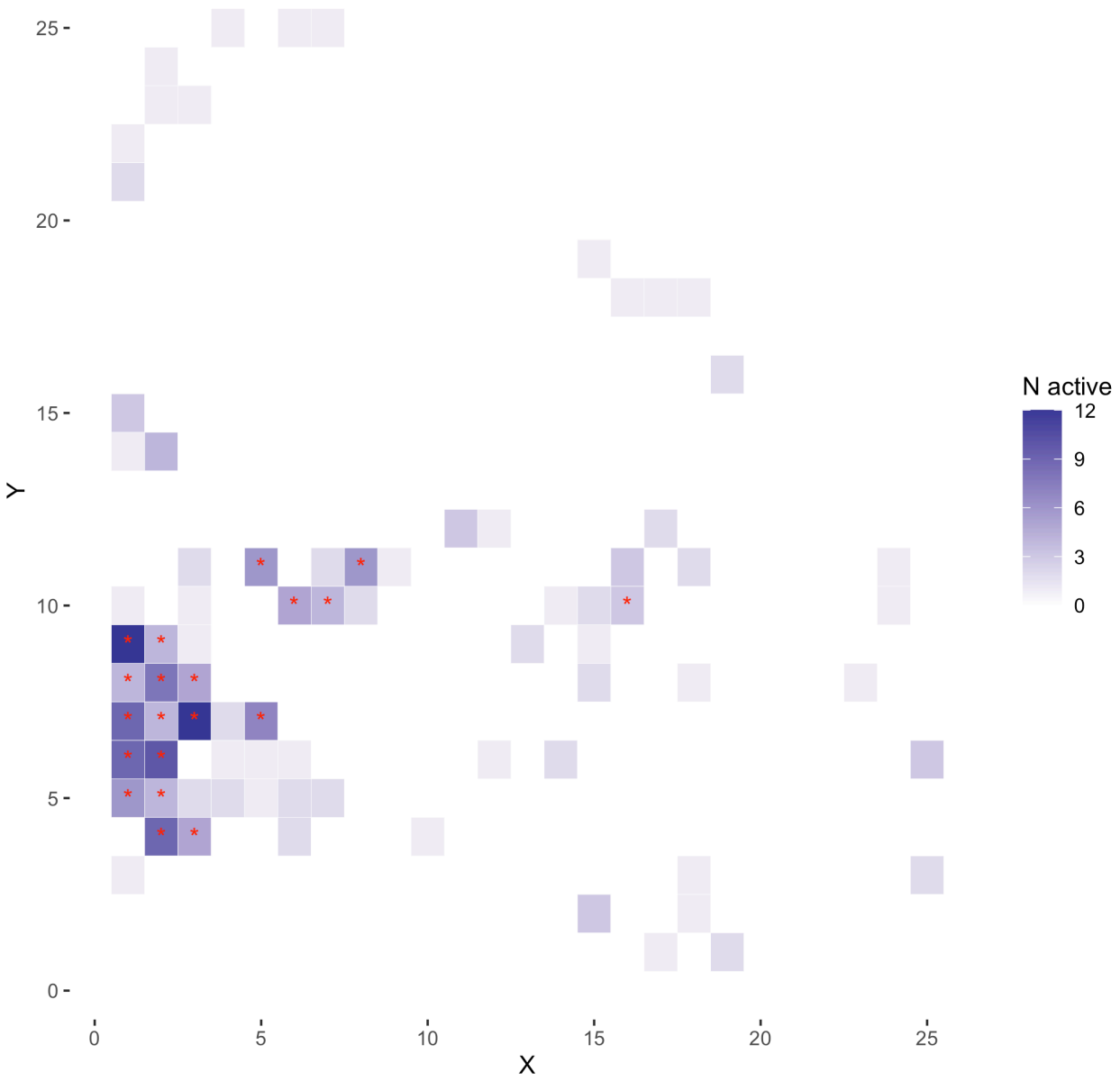
fgfr1



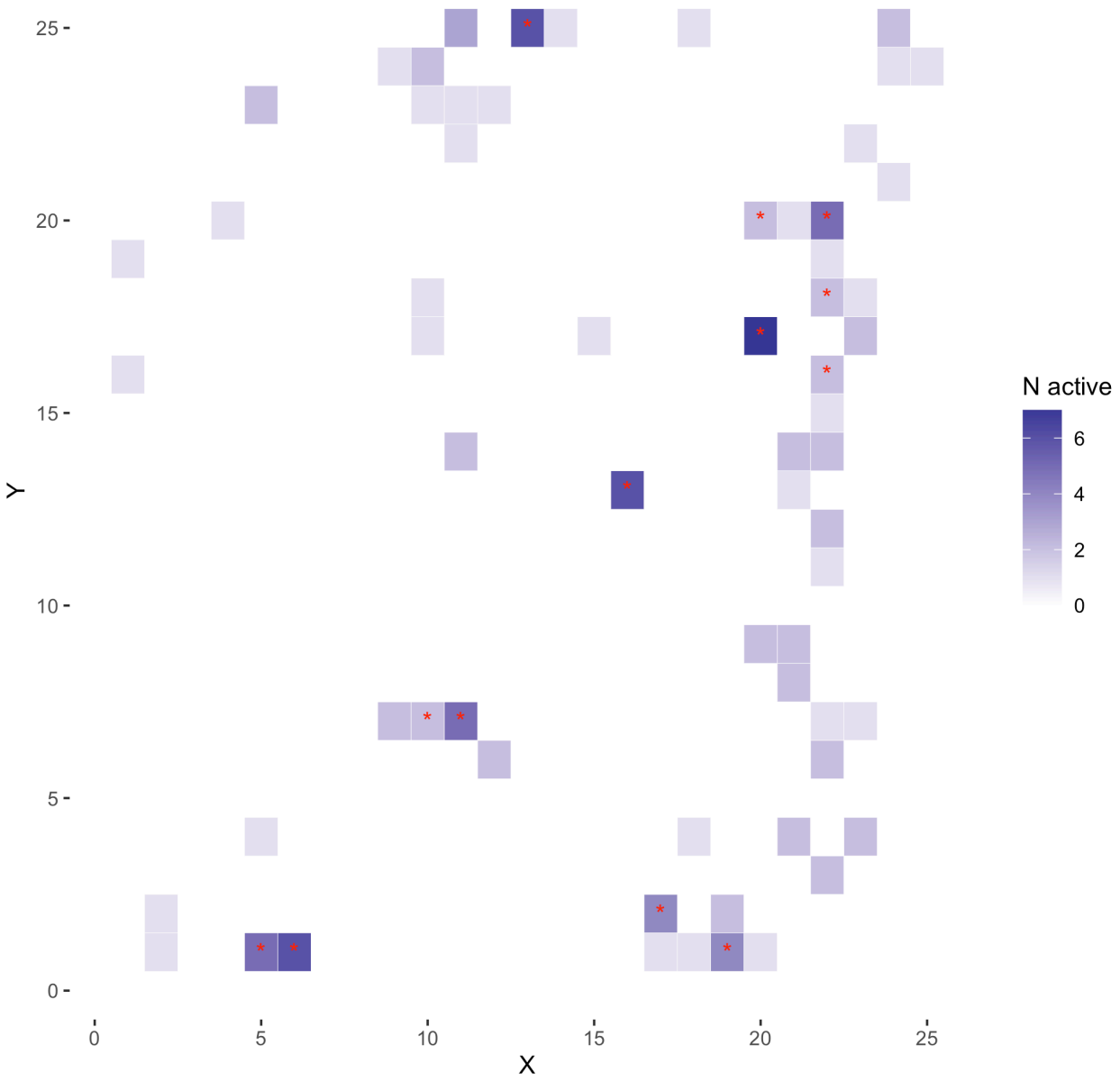
nos1



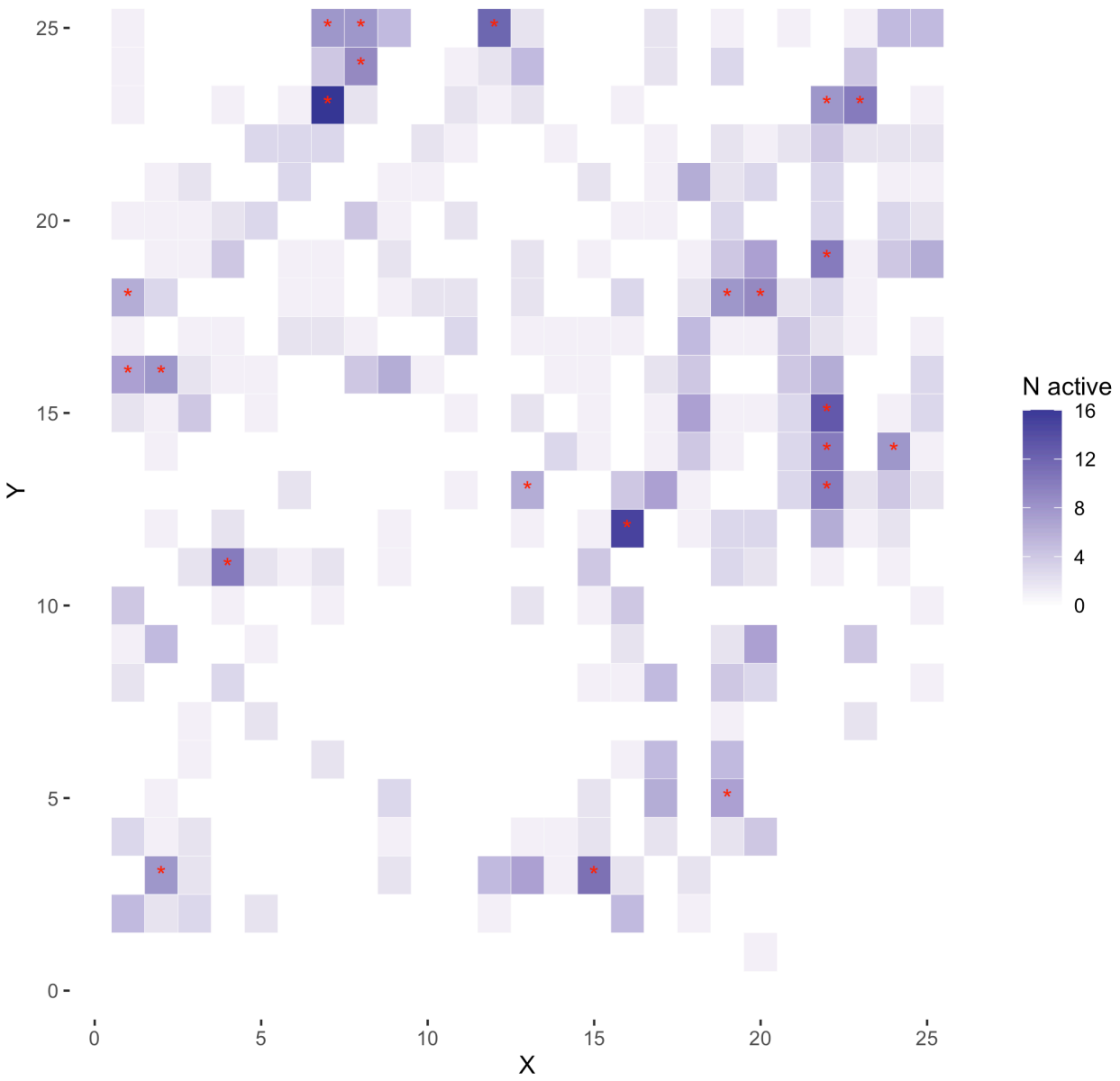
nram



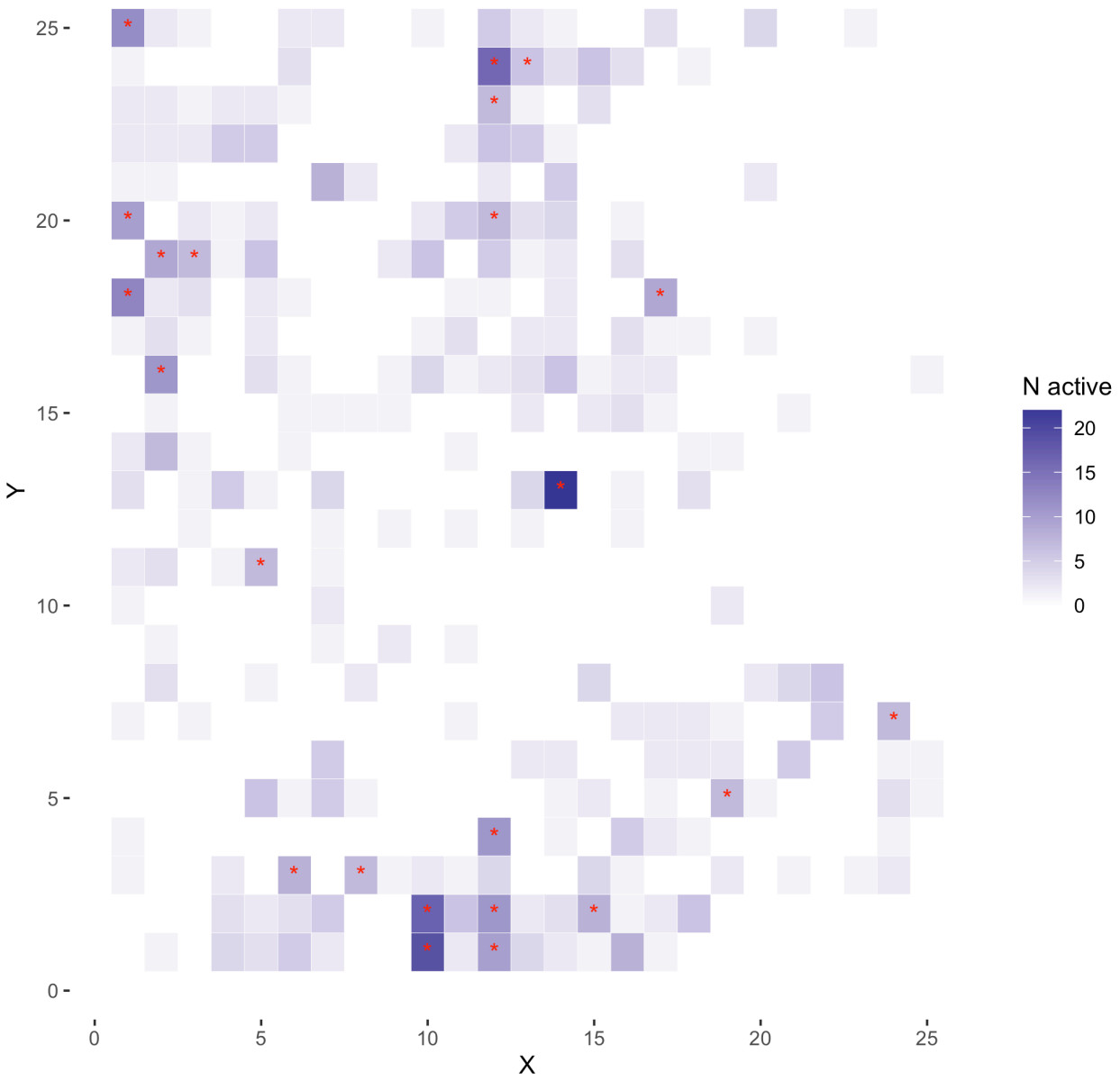
pa2ga



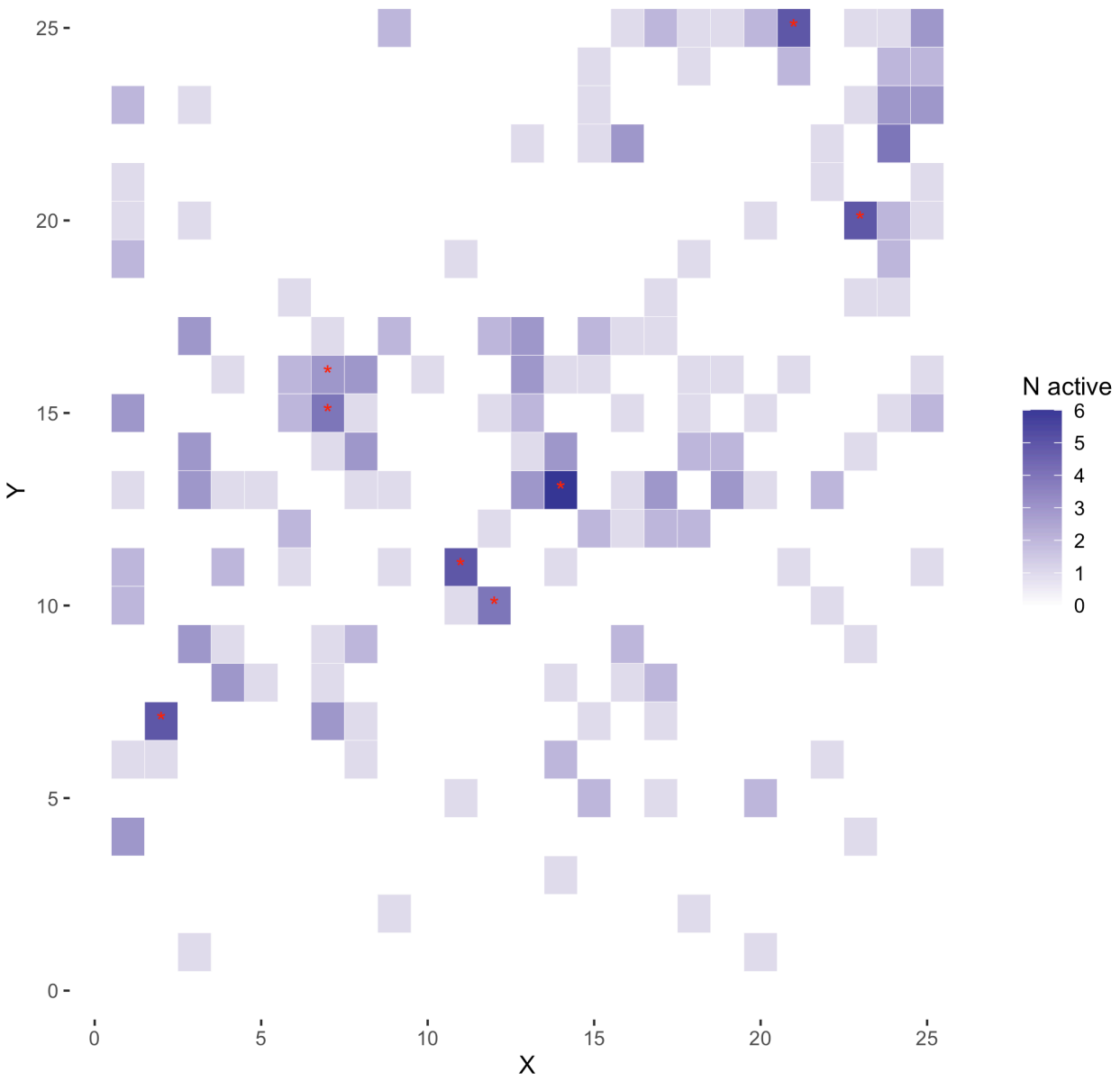
parp1



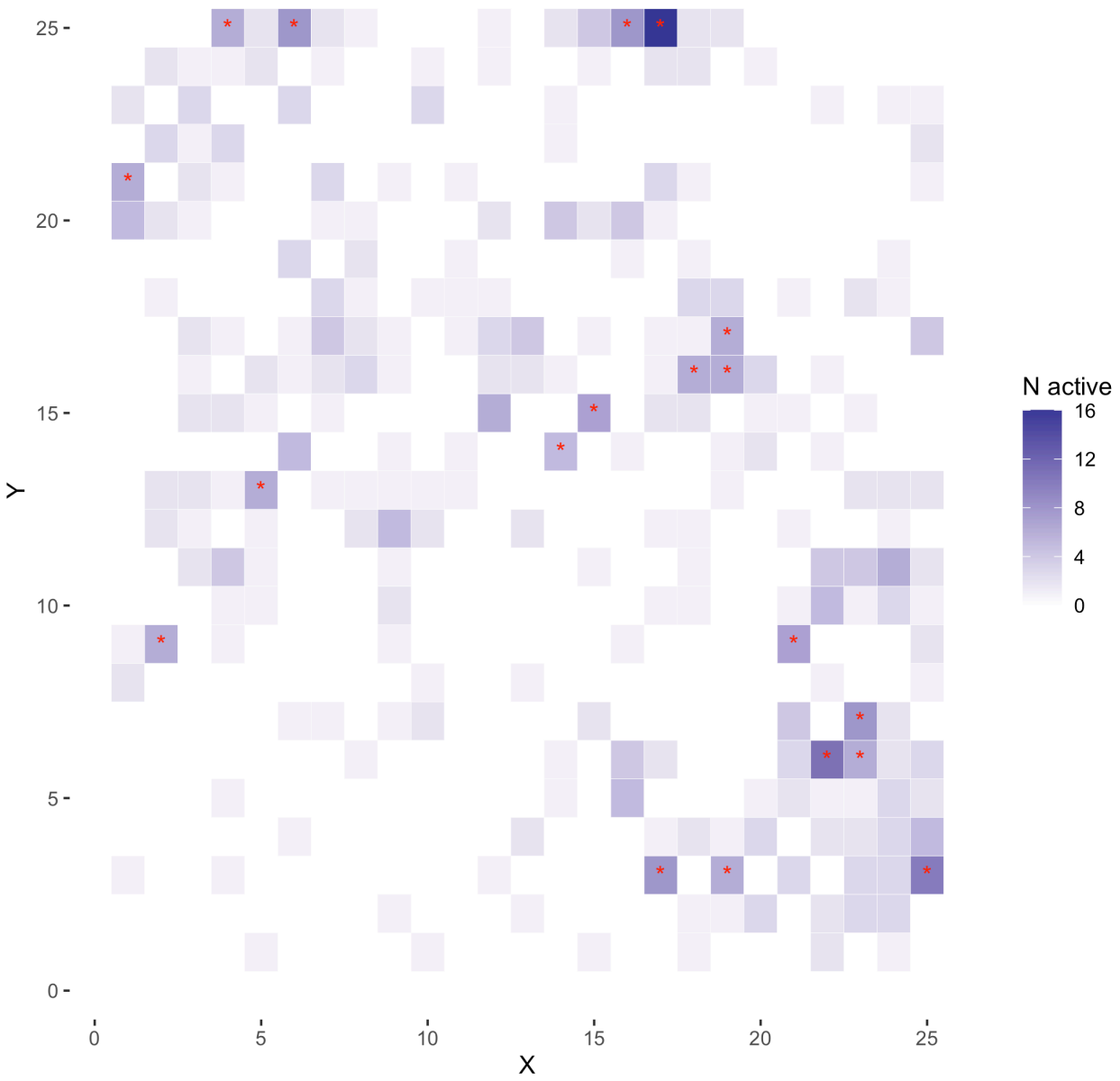
pde5a



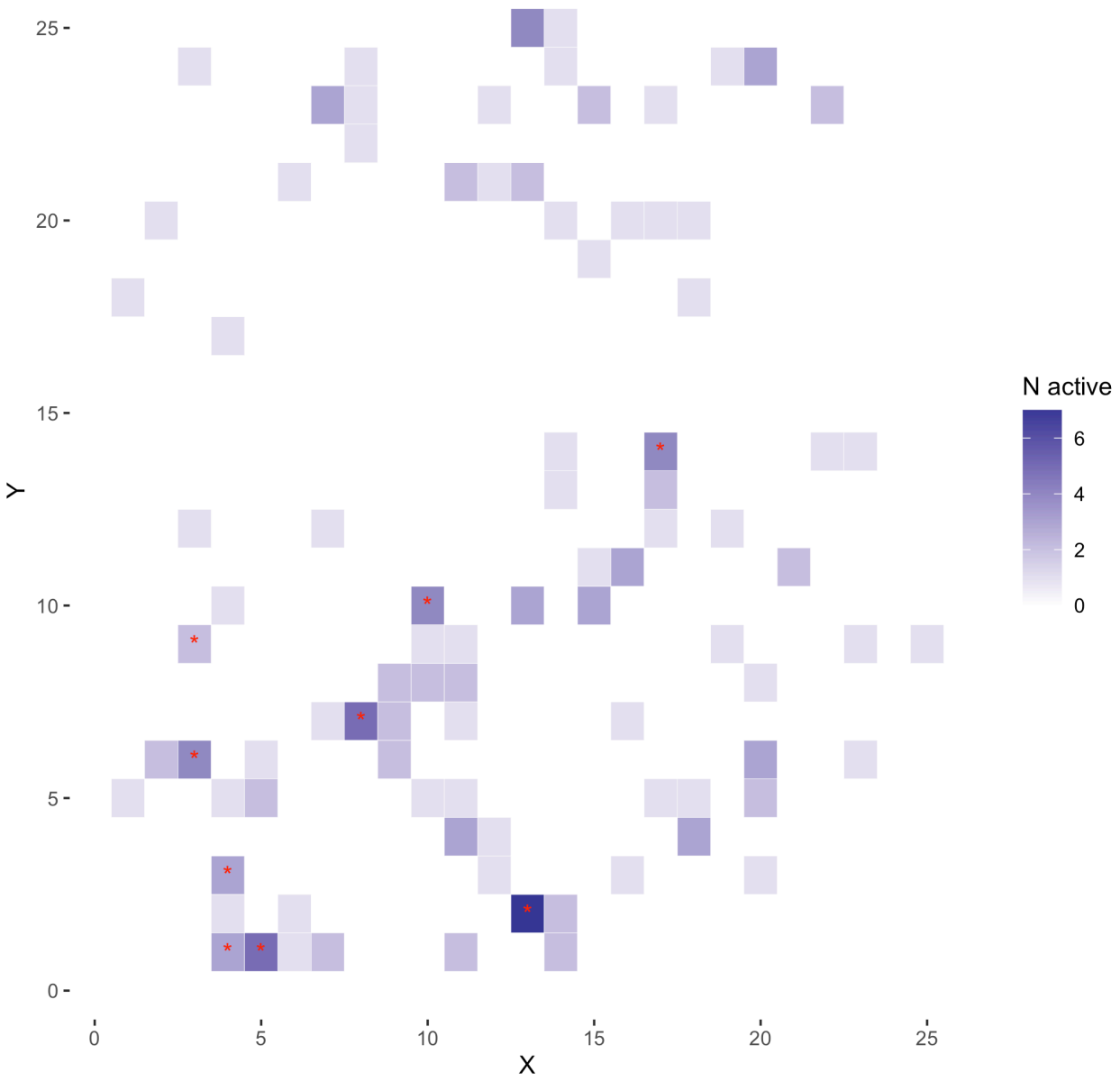
pgh1



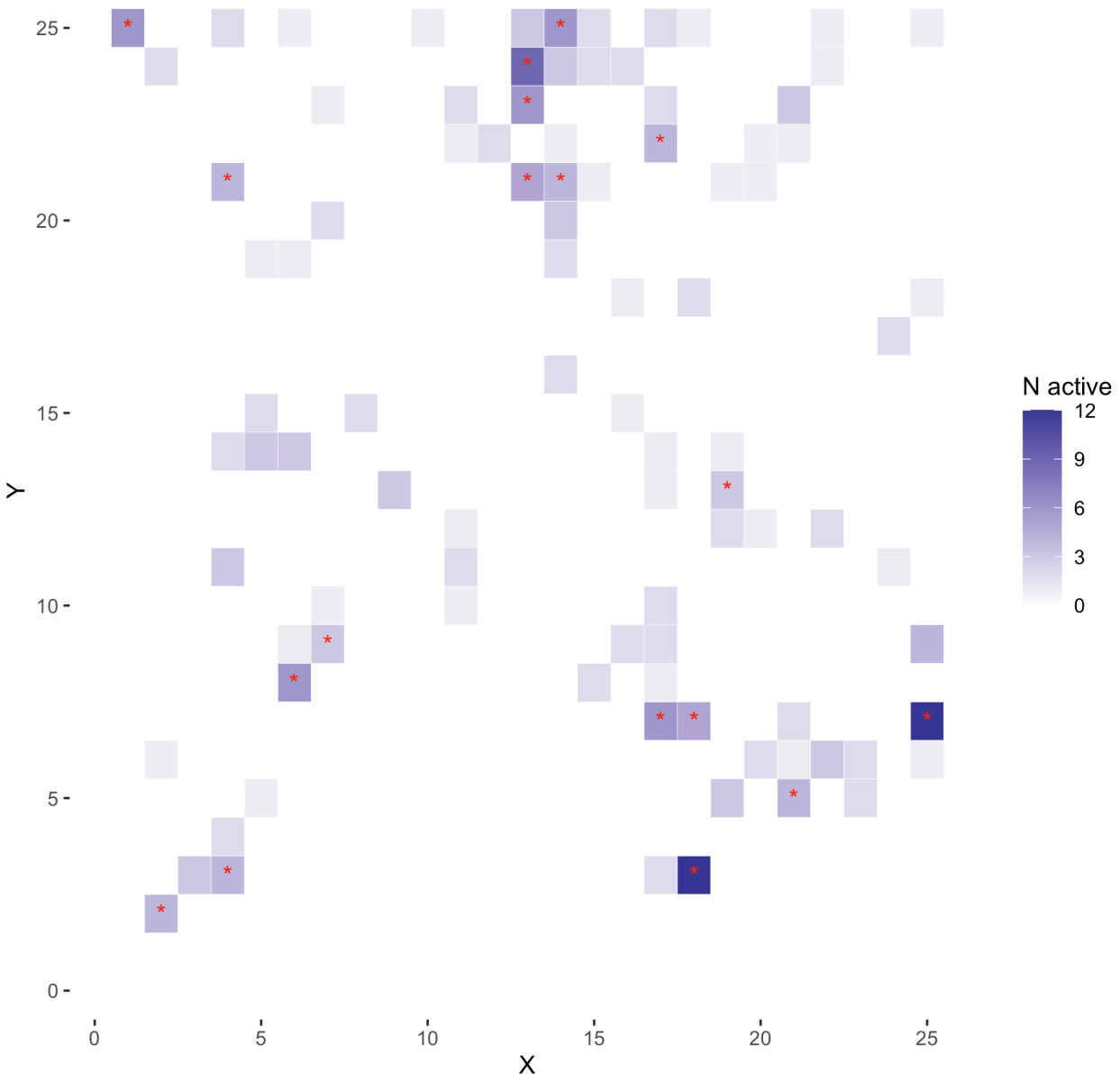
pgh2



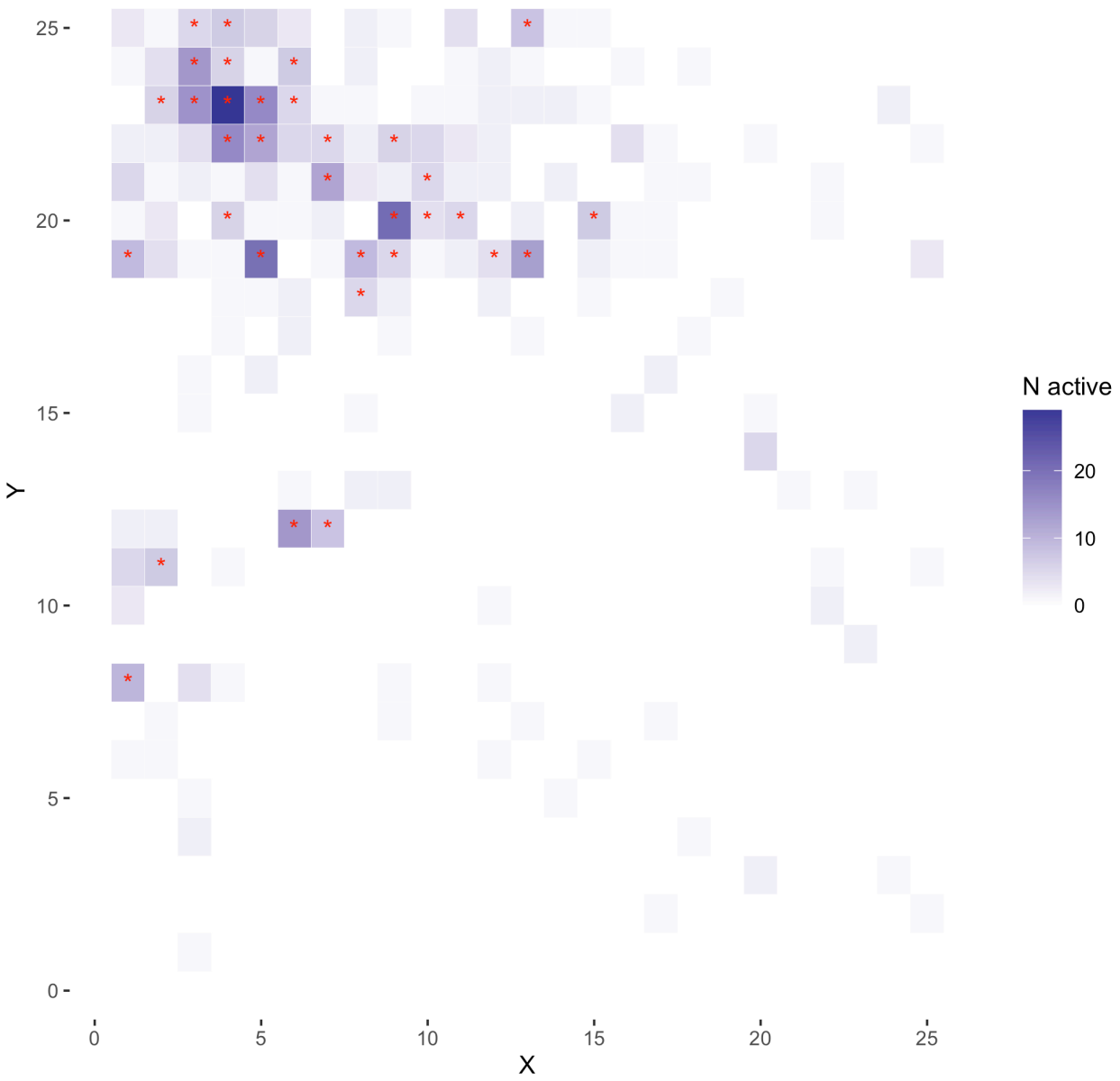
plk1



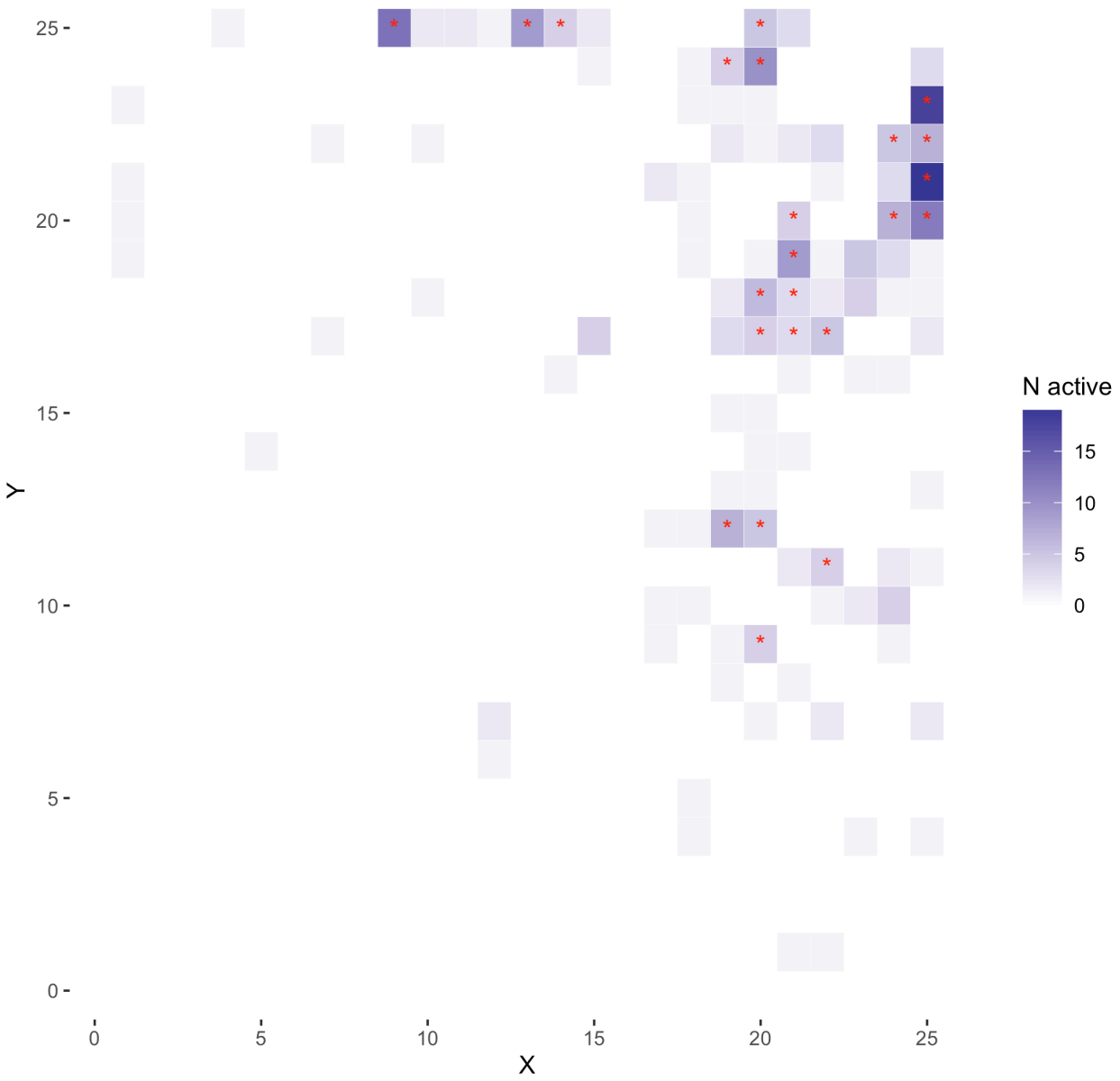
pnph



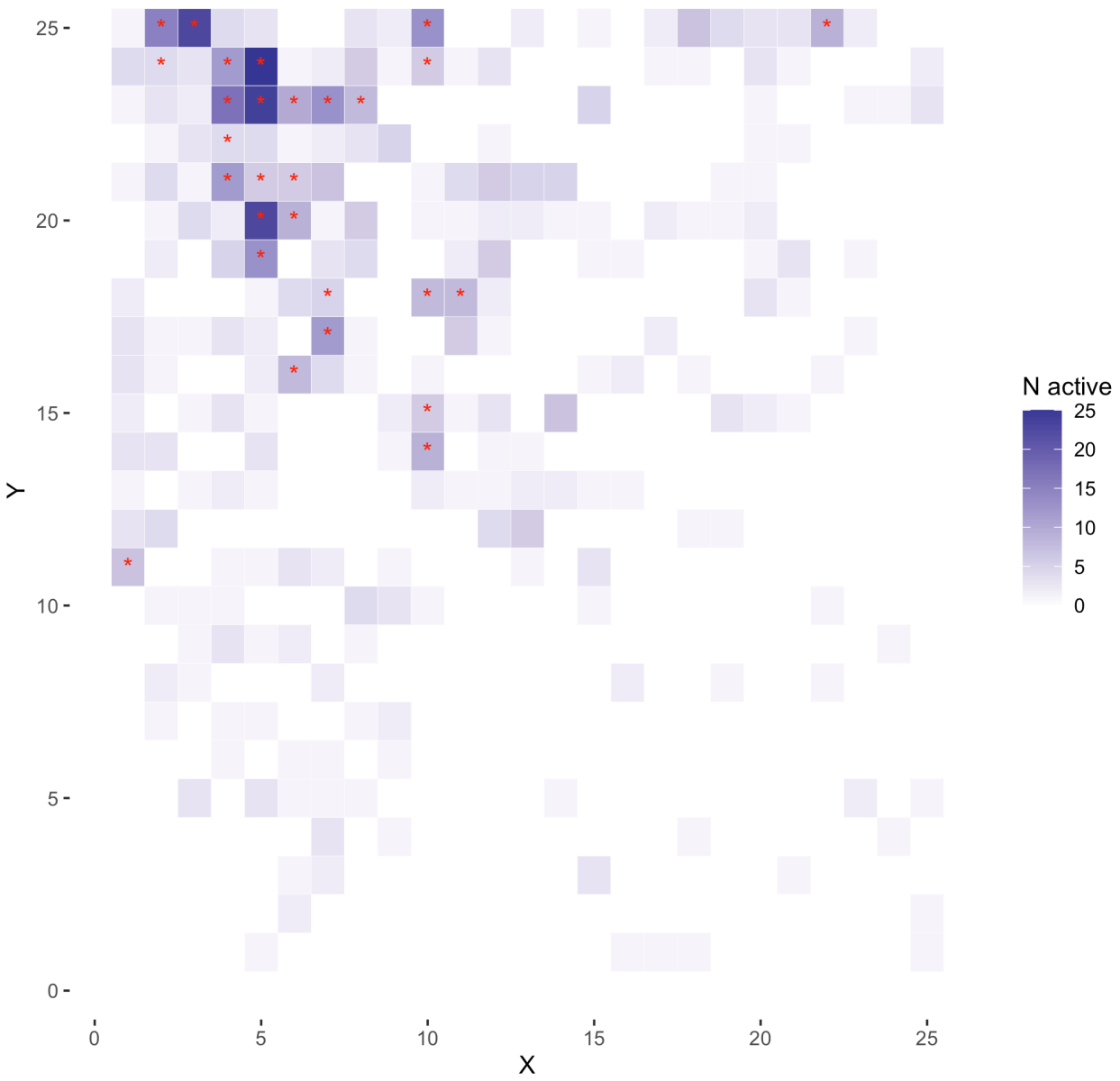
ppara



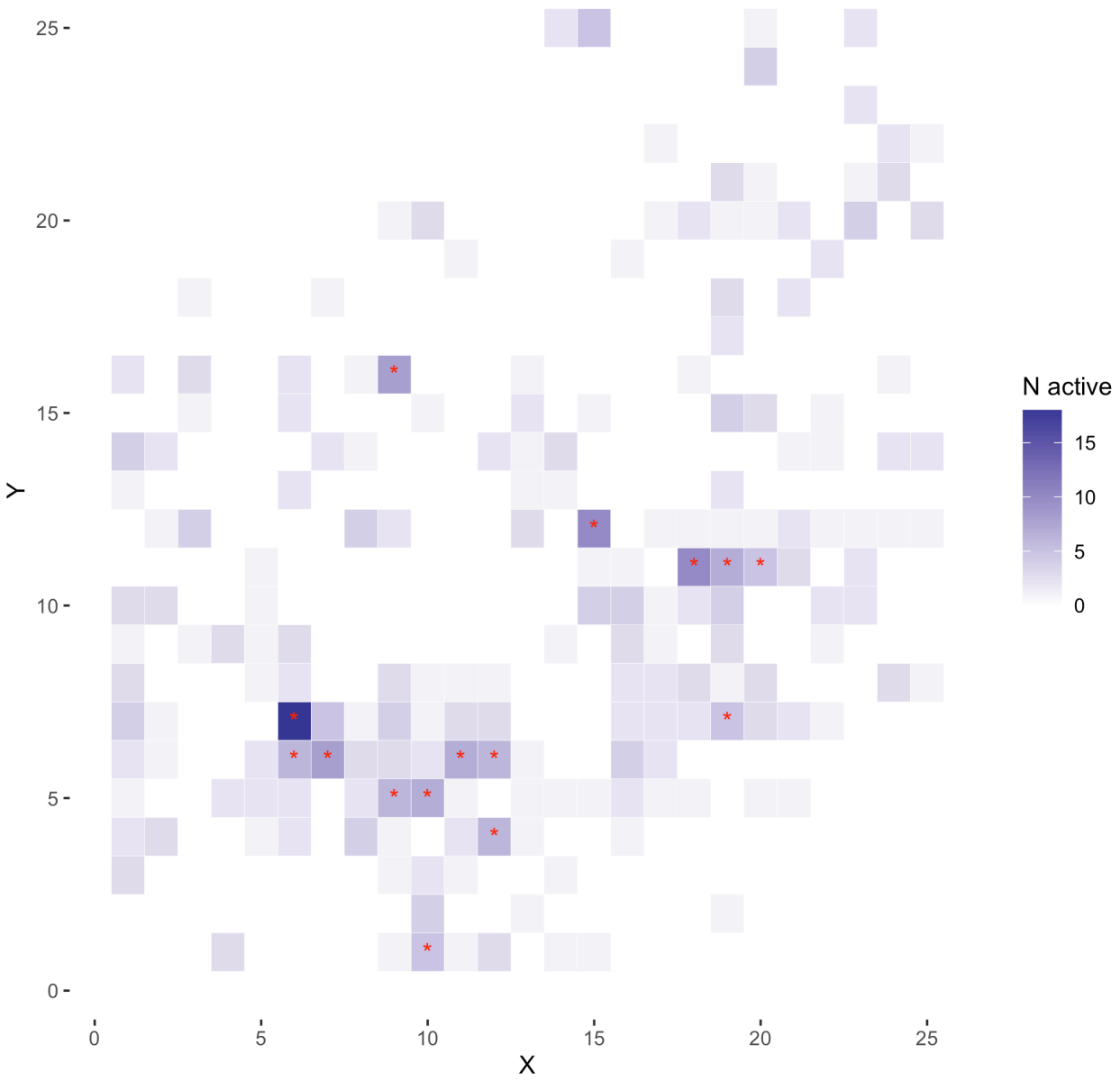
ppard

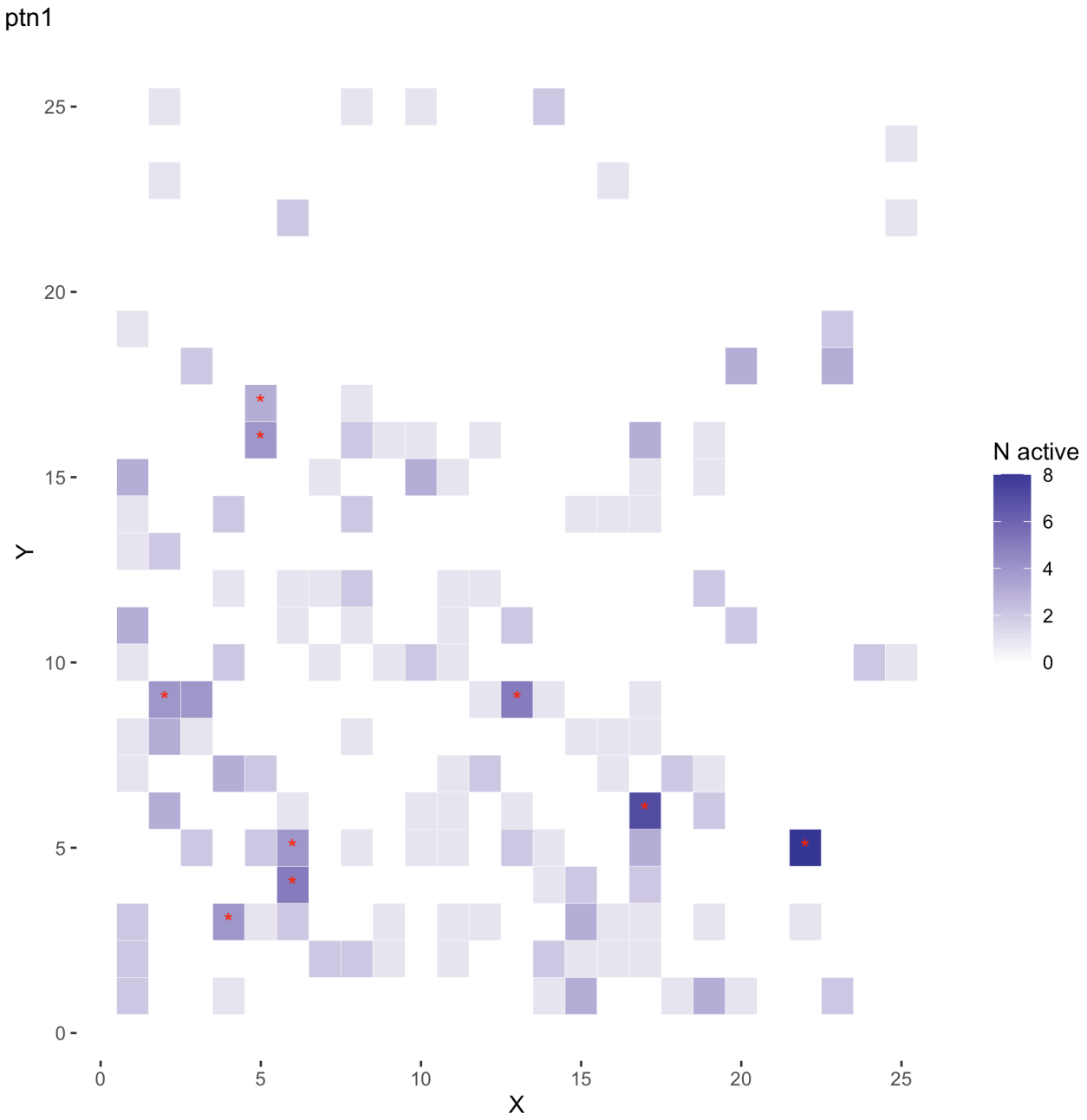


pparg

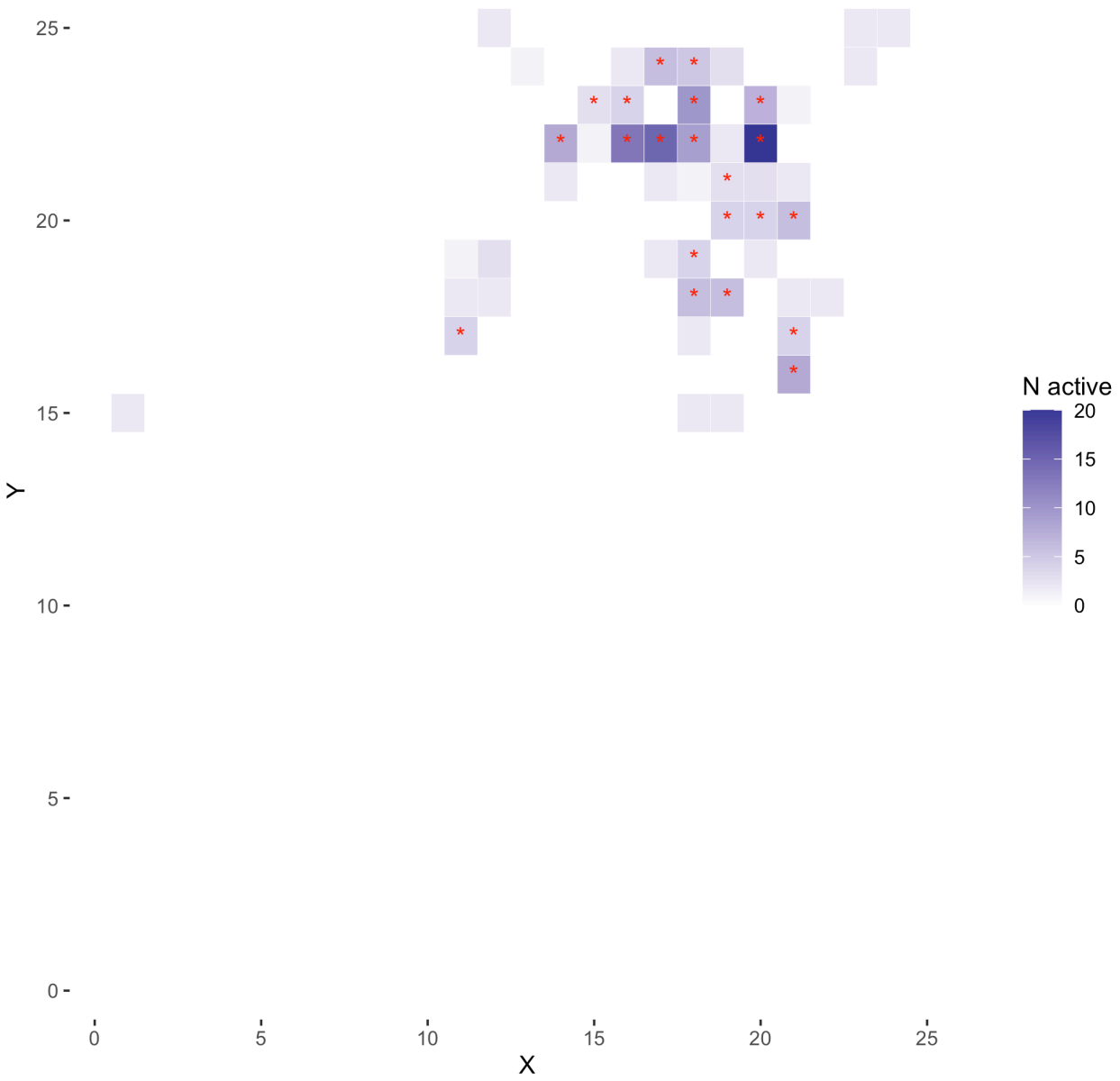


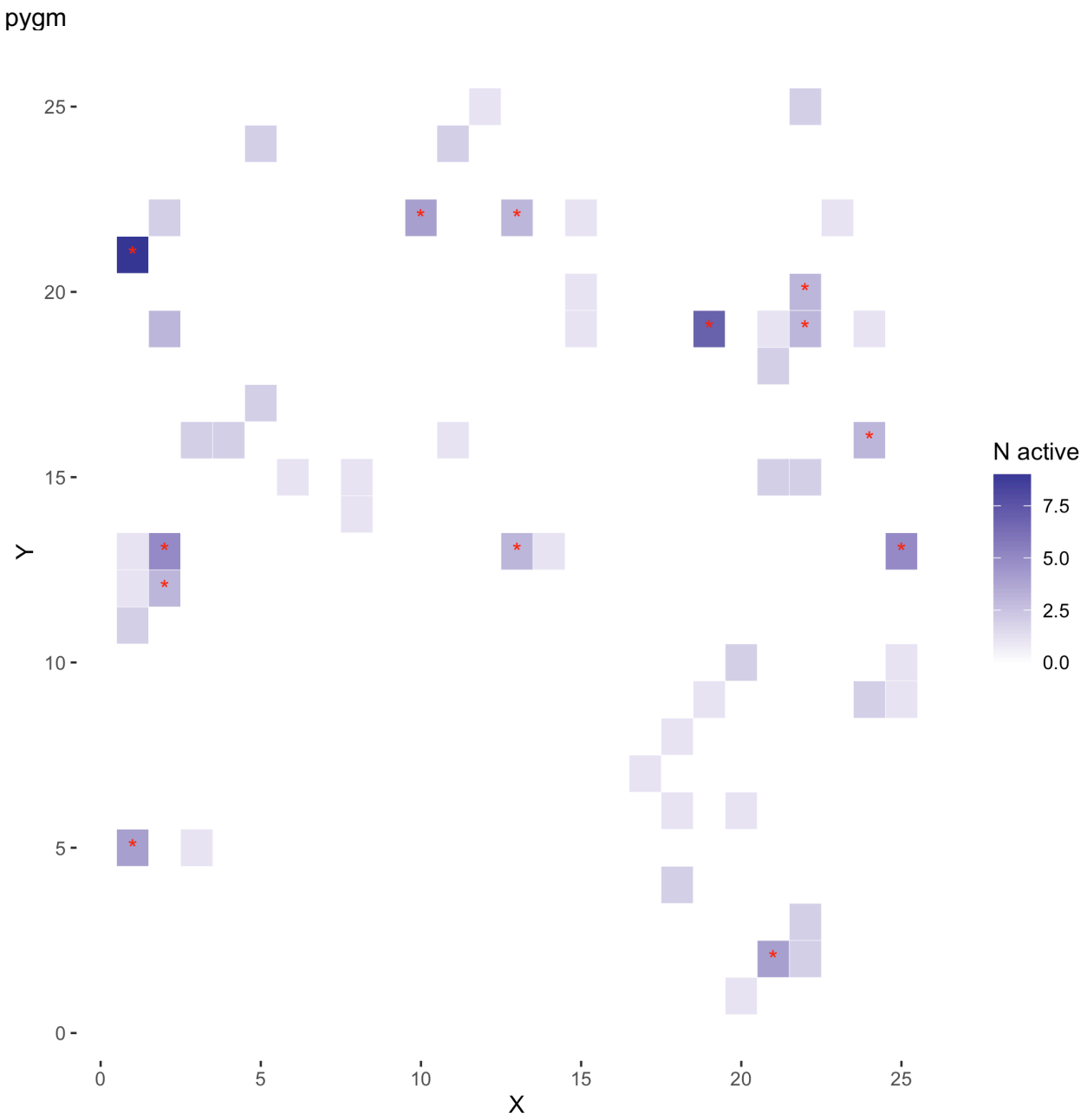
prgr



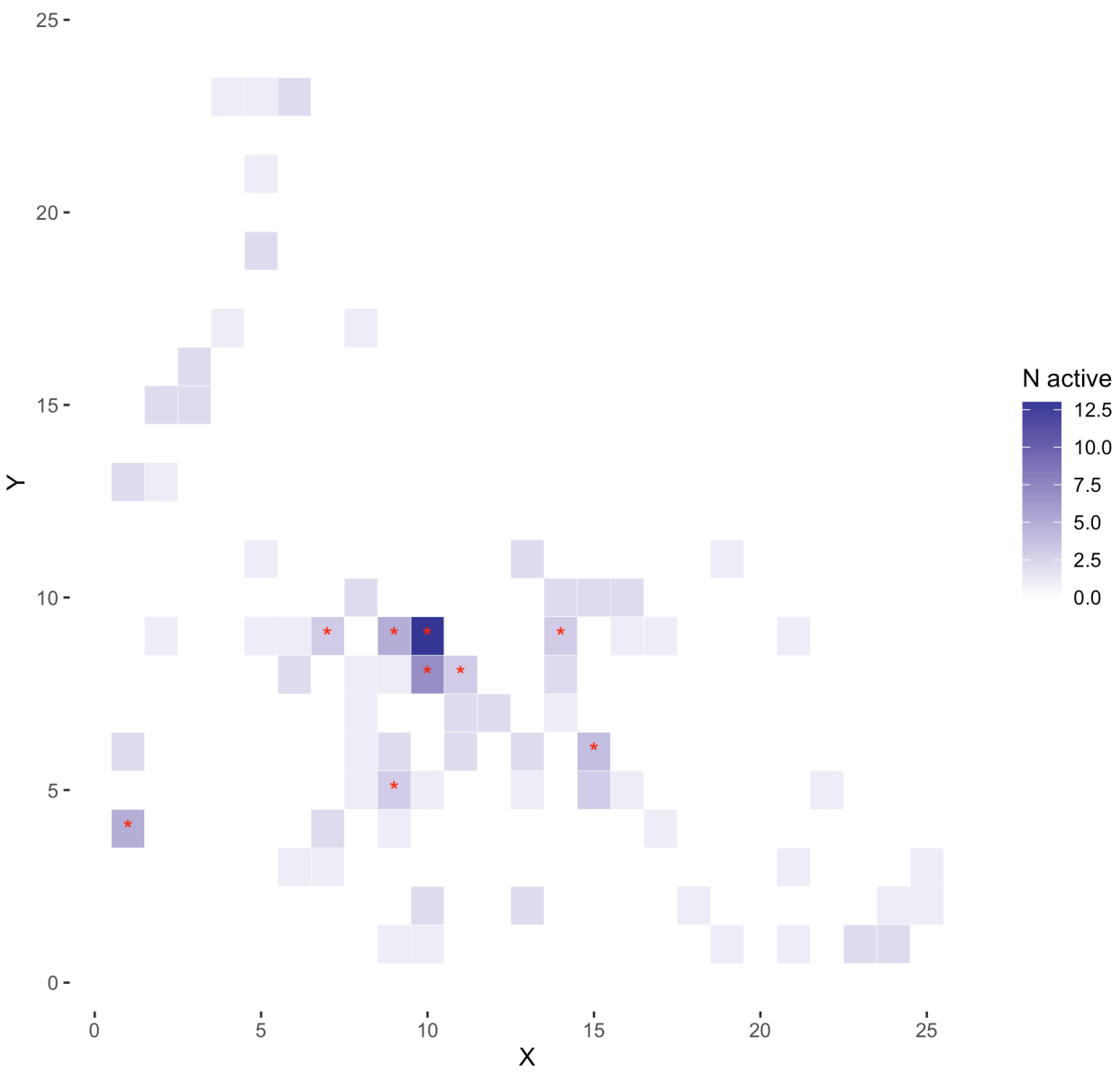

ptn1
N active
8
6
4
2
0
Y
X
0
5
10
15
20
25

pur2

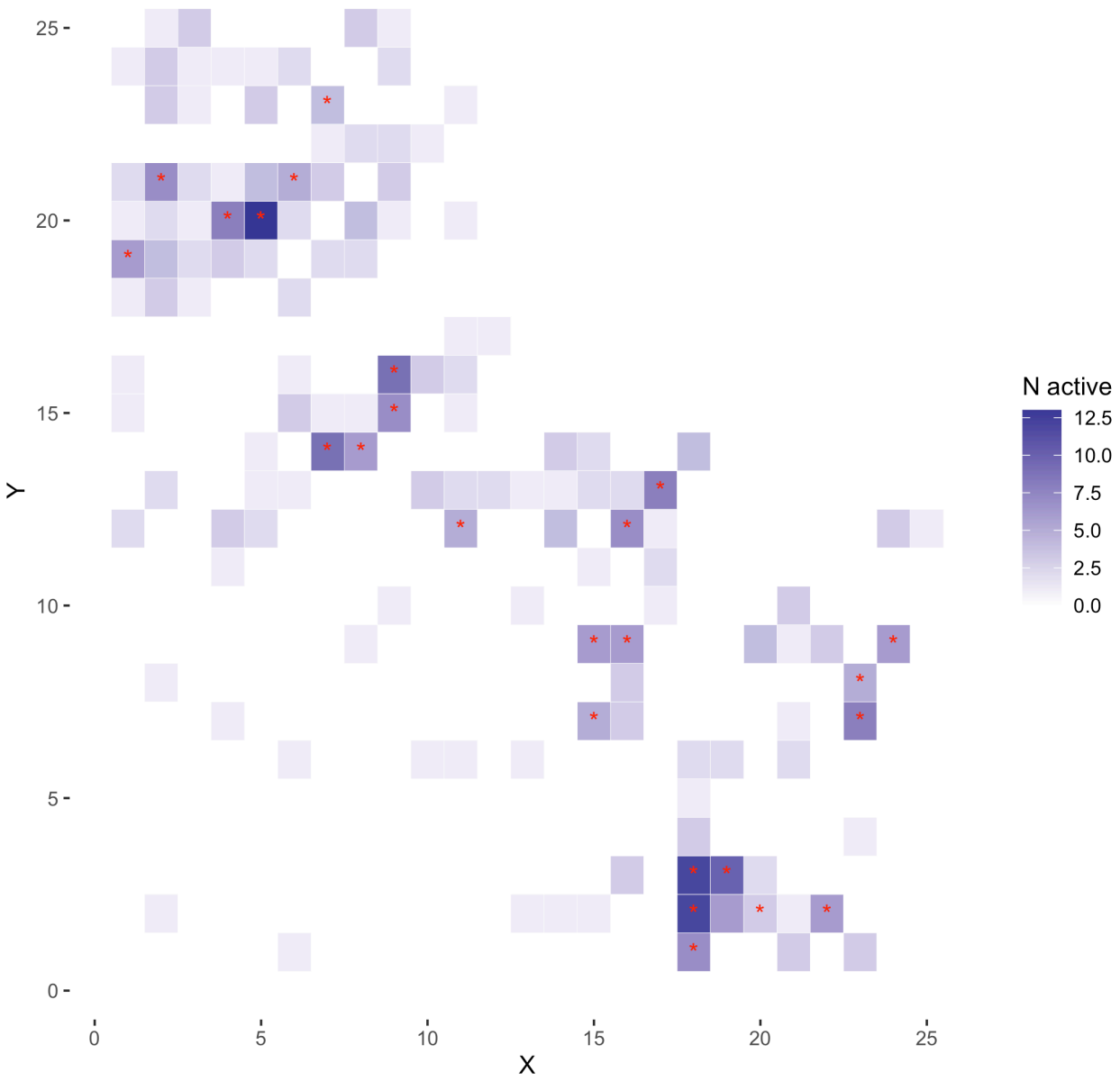



pygm
25
20
15
10
5
0
Y
0
5
10
15
20
25
X
N active
7.5
5.0
2.5
0.0

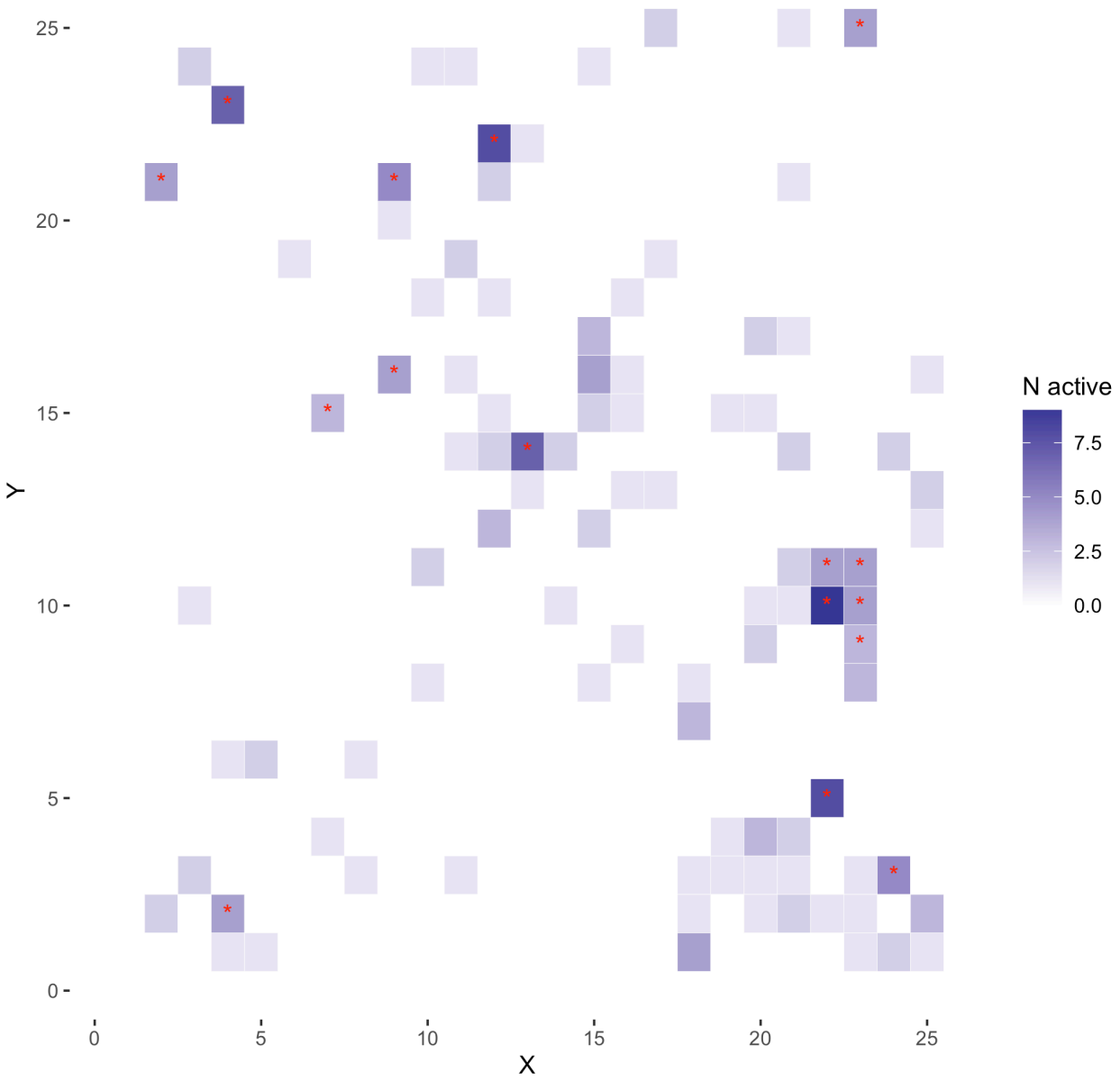
pyrd



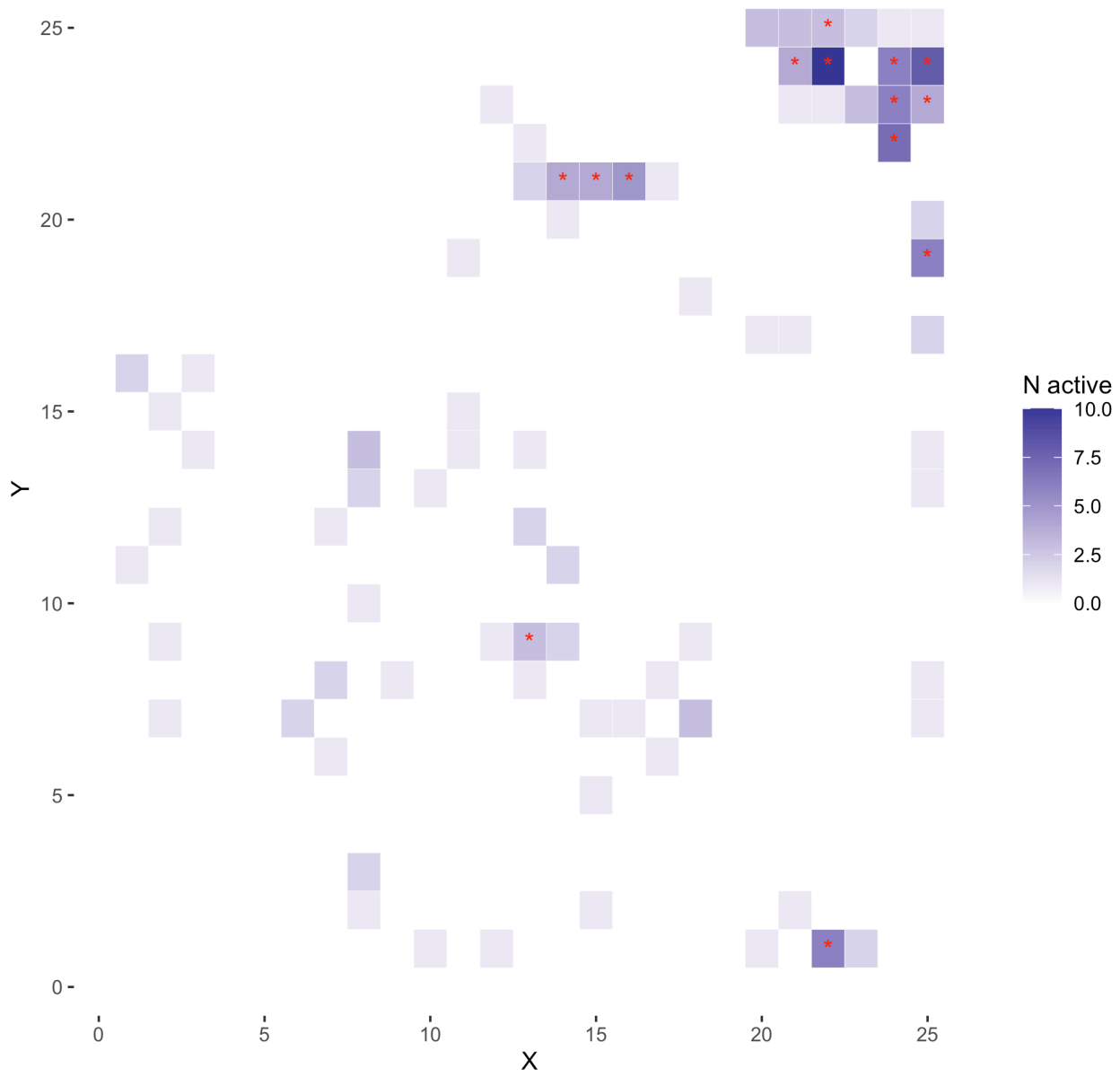
reni



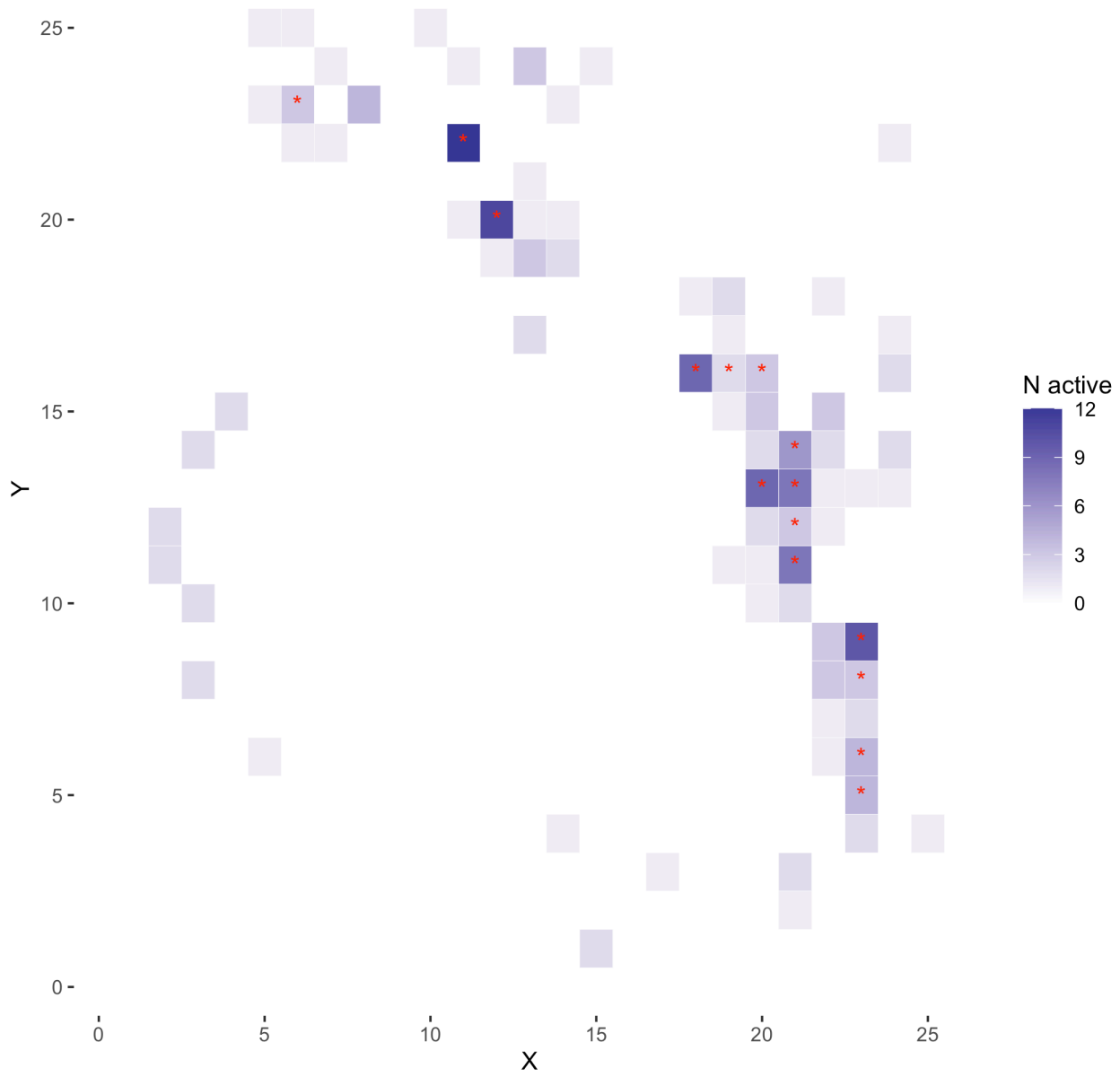
rock1



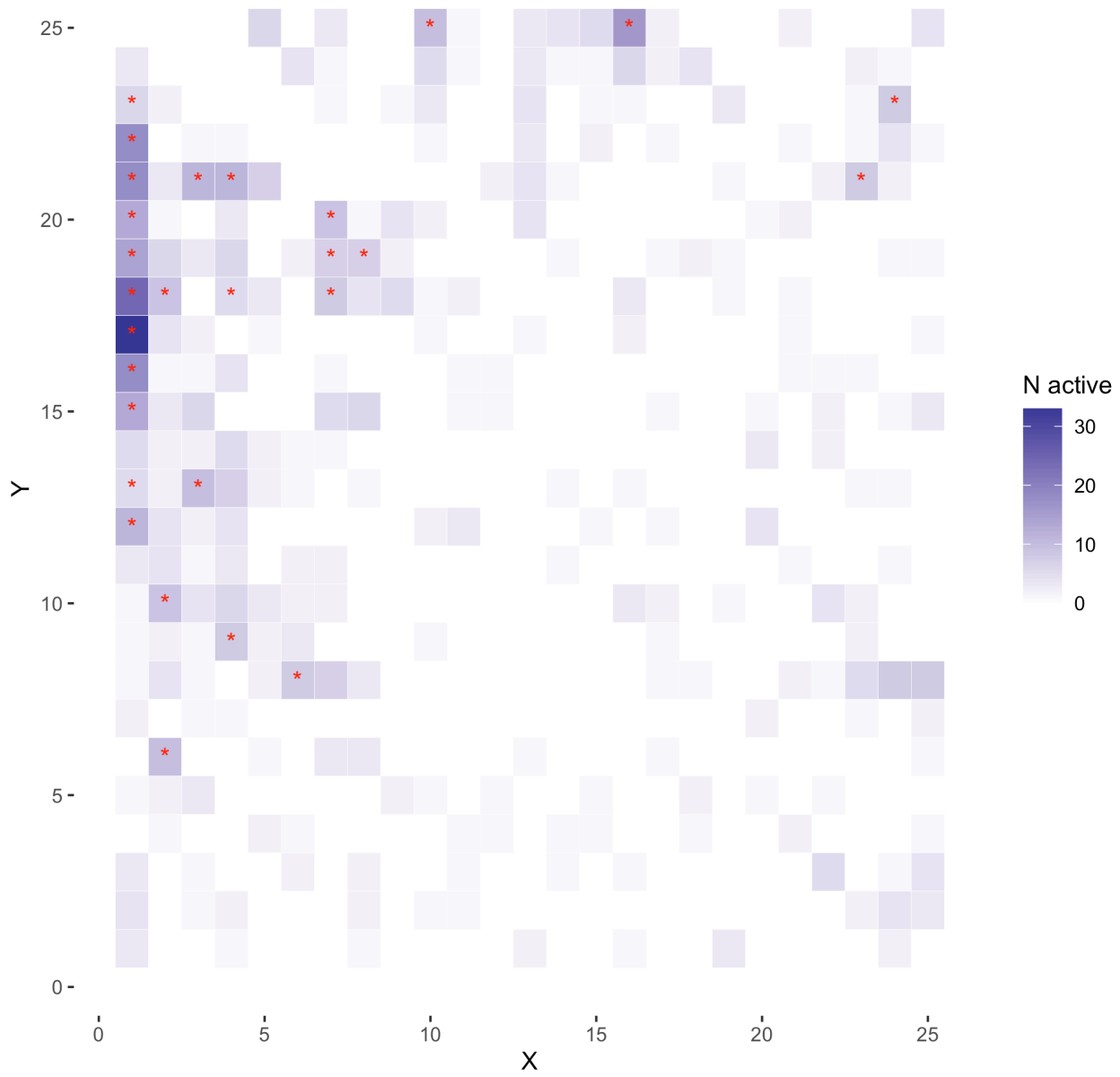
rxra



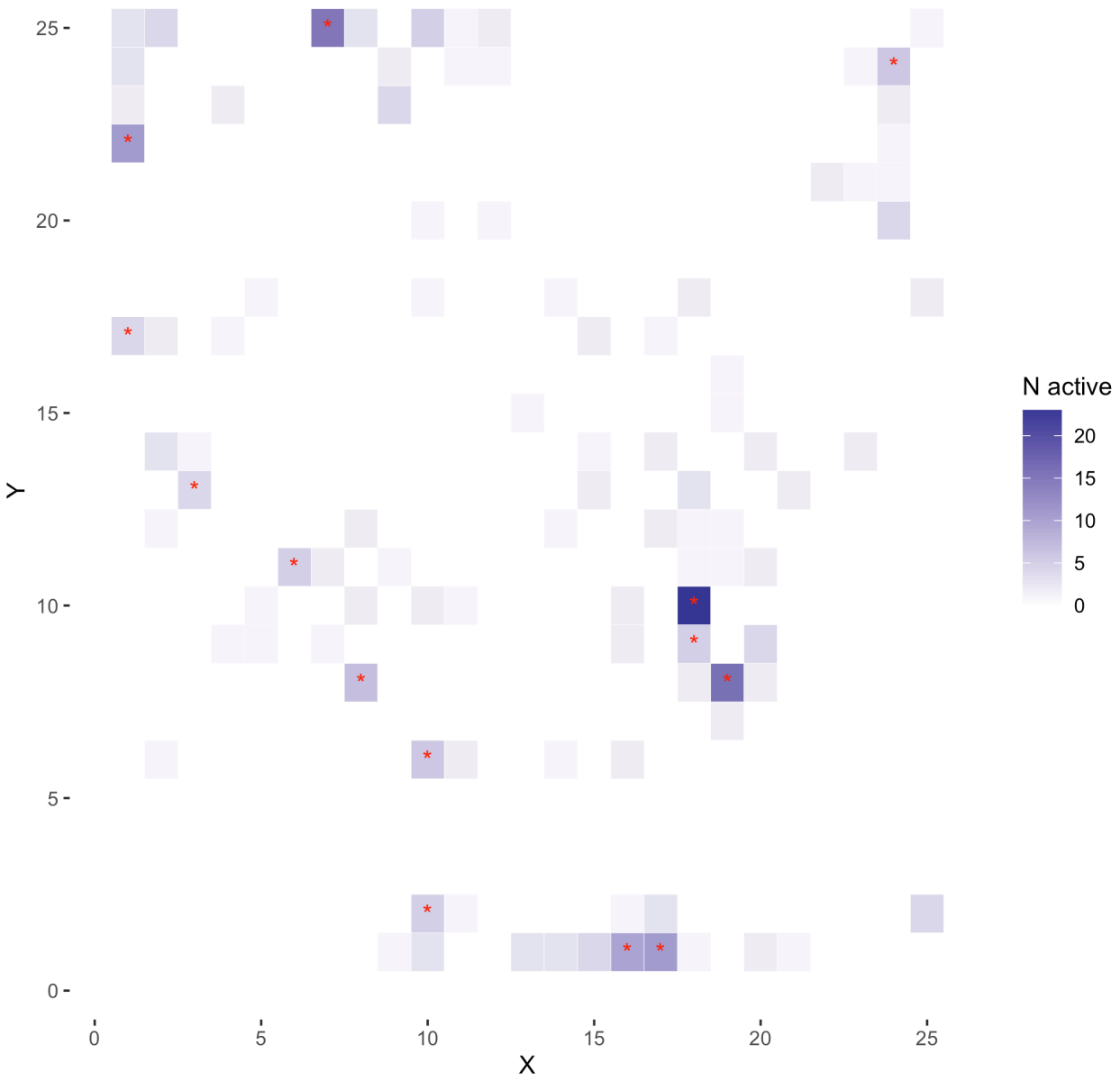
sahh



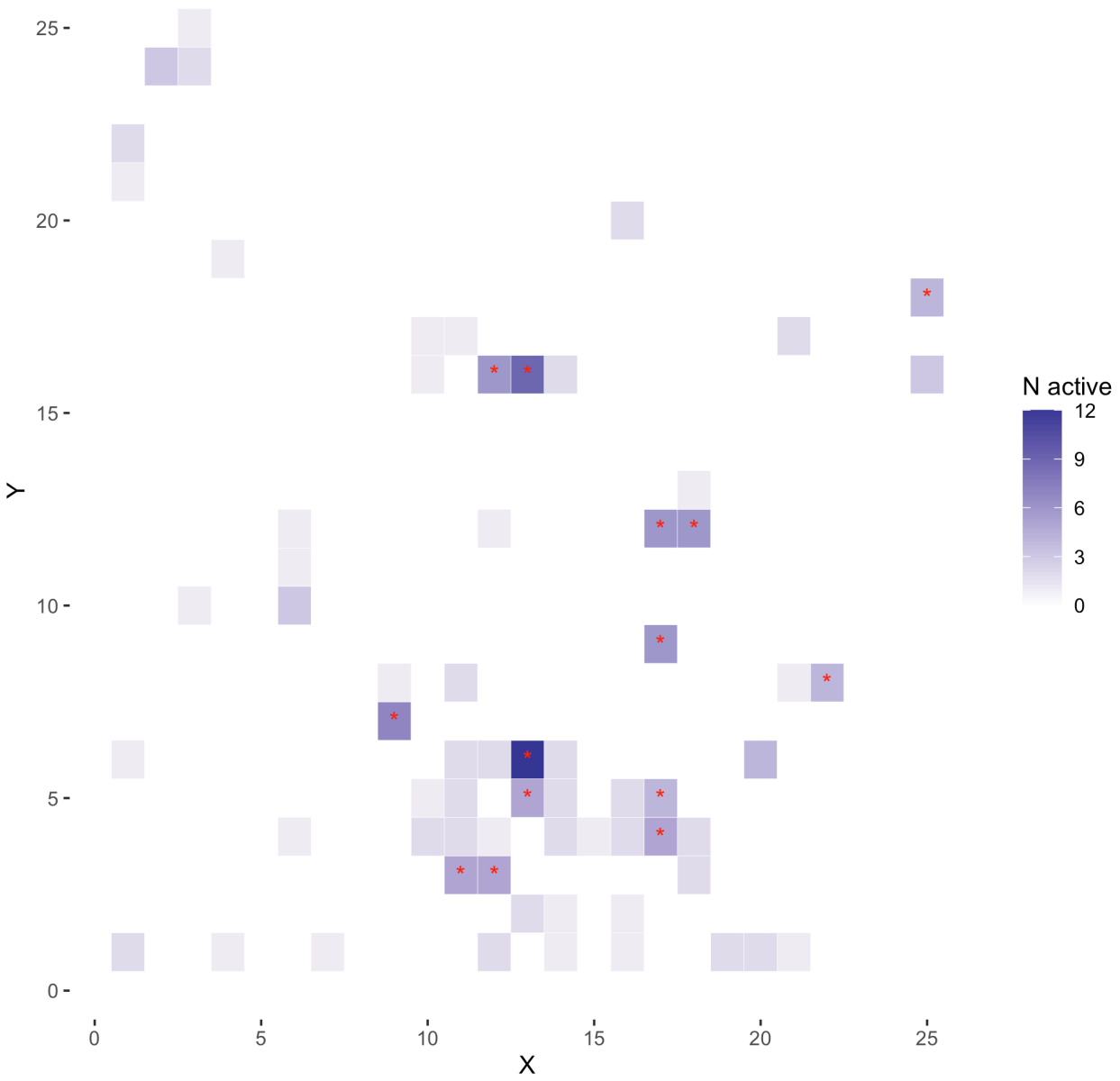
src



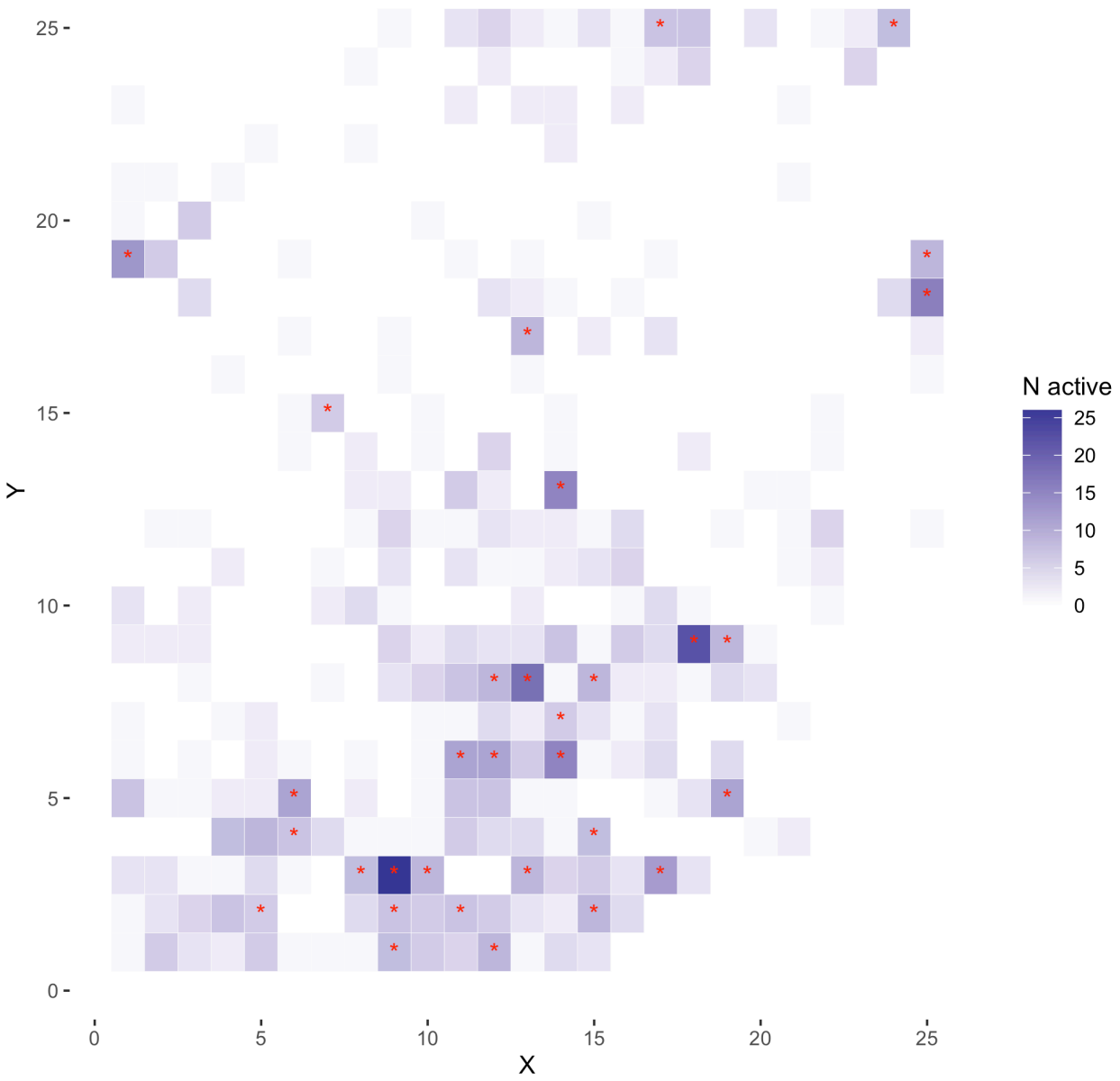
tgfr1

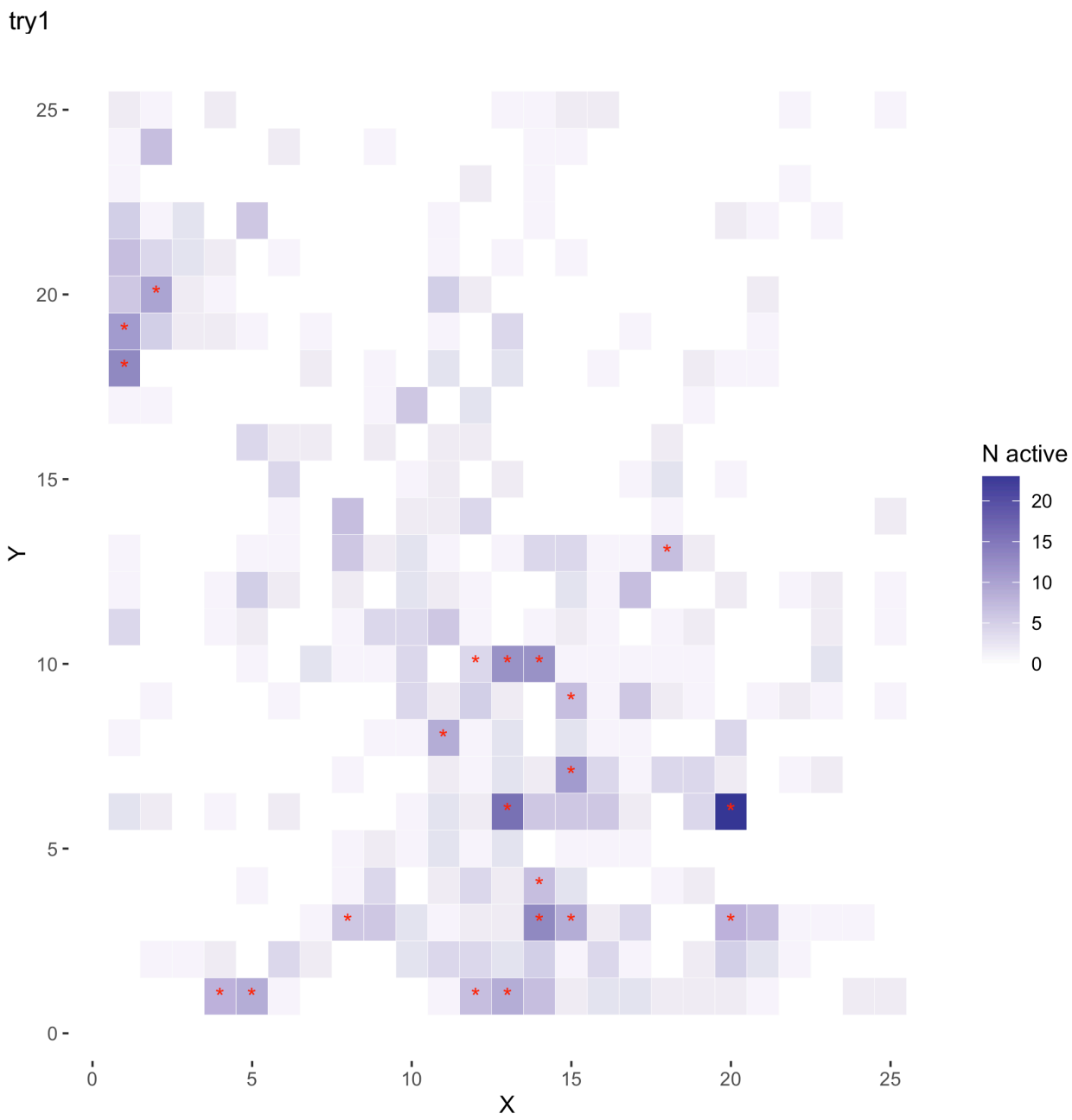


thb



thrb





S10) References

1. Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small molecule shape–fingerprints. *J Chem Inf Model* **2005,** 45, 673–84.

2. Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The development of versions 3 and 4 of the Cambridge Structural Database System. *Journal of Chemical Information and Modeling* **1991,** 31, 187–204.

3. Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B* **2002,** 58, 380–8.
4. Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004,** 20, 2153–5.
5. Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry* **1996,** 17, 1653–1666.
6. Taylor, R.; Cole, J. C.; Cosgrove, D. A.; Gardiner, E. J.; Gillet, V. J.; Korb, O. Development and validation of an improved algorithm for overlaying flexible molecules. *J Comput Aided Mol Des* **2012,** 26, 451–72.
7. Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein–ligand docking against non–native protein conformers. *J Chem Inf Model* **2008,** 48, 2214–25.
8. Hajian–Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med* **2013,** 4, 627–35.
9. Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J Chem Inf Model* **2010,** 50, 572–84.
10. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J Med Chem* **2002,** 45, 4350–8.
11. Roy, A.; Skolnick, J. LIGSIFT: an open–source tool for ligand structural alignment and virtual screening. *Bioinformatics* **2015,** 31, 539–44.
12. *OpenEye Toolkits*, OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com.
13. Team, R. C. *R: A language and environment for statistical computing.*, R Foundation for Statistical Computing: Vienna, Austria. https://www.R–project.org/, 2020.
14. Software, O. S. *OMEGA 2.5.1.4*, Santa Fe, NM. http://www.eyesopen.com.
15. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD–E): better ligands and decoys for better benchmarking. *J Med Chem* **2012,** 55, 6582–94.
16. Leach, A. G.; Kidley, N. J. Cytochrome P450 Substrate Recognition and Binding. In *Drug Metabolism Prediction*, 2014; pp 103–132.
17. Wiegers, T. L., L.; Vergeest, J. S. M. Shape Language – How People Describe Shapes and Shape Operations. In *Research into design : supporting multiple facets of product development*, Chakrabarti, A., Ed. Research Publishing Services: Singapore, 2009; pp 278–286.