Identification of Yeast Strains for anti-cancer drug Screening

by

Padmasini R

under the supervision of

Dr. Gaurav Ahuja

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, Computational Biology

Center for Computational Biology,
Indraprastha Institute of Information Technology - Delhi

Certificate

This is to certify that the thesis titled "**Yeast as a proxy for anti-cancer drug screen studies in the human cancer cell lines & anti-fungal studies**" being submitted by Ms. **Padmasini R** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology in Computational Biology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other University or Institute for the award of any degree/diploma.

April, 2023
Delhi

Dr. Gaurav Ahuja,
Associate Professor,
Department of Computational Biology,
Indraprastha Institute of Information Technology, Delhi,
New Delhi 110020

Declaration

I submit this project entitled "**Yeast as a proxy for anti-cancer drug screen studies in the human cancer cell lines & anti-fungal studies**" to the Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi - 110020. I declare that this is my original work carried out under the guidance of Dr. Gaurav Ahuja, Associate Professor, Department of Computational Biology at IIIT-Delhi.

April 2023
Delhi

Padmasini R,
M.Tech Student (Dec 2020 - April 2023),
Department of Computational Biology,
Indraprastha Institute of Information Technology, Delhi,
New Delhi 110020

Acknowledgements

I take this opportunity to thank my guide Dr. Gaurav Ahuja, Associate Professor in the Department of Computational Biology at IIIT-Delhi. I faced a lot of roadblocks in the initial phase of my project. It was up to the extent that I would not get fruitful results. However, the experience of my guide and the time he provided to analyze the initial setbacks paved the way for successful completion. He gave me all the strength and freedom to explore my ideas and motivated me to remain on the right path. Under his guidance, I have learned and gained valuable knowledge throughout the project. I am thankful to have such a guide.

I further extend my gratitude to my mentor Ms. Vishakha Gautam, a graduate student in our lab. She helped me in aligning and channelizing my work, providing timely inputs, and extending her support in clarifying all my trivial and non-trivial doubts.

I also thank all my faculty members and staff of the Department of Computational Biology and IIIT Delhi for always helping us throughout our college journey. My special thanks to IT helpdesk for their continuous support and help in providing access to the college IT infrastructure.

Last but not least, I thank my constants - parents and brother, and all my labmates and batchmates for their constant support during my dissertation period.

Index

List of figures

List of Tables

Abstract

I - Abstract:

The cellular mechanisms in yeast are highly conserved with humans making it an inexpensive, rapid and easier model organism to work with. Cell cycle checkpoints, DNA repair pathways and CDKs are well studied in yeast and tweaking them have been reported to mimic cancer phenotypes. These mutant strains of yeast mimicking cancer phenotypes are used for studying anticancer activity of drugs and their interactions in the genome. Presence of numerous distinct human tumor cell lines makes drug screening exhaustive and expensive. In search of a universal cell line substituting most of the cancer cell lines, we found a yeast mutant strain that could potentially replace anticancer drug studies done on NCI60 human tumor cell lines. Yeast being non-infectious makes it a suitable model organism for screening antifungal compounds. The growth inhibition studies of several drugs in yeast mutant strains from the National Cancer Institute website is a wonderful reservoir of datasets to be analyzed. Artificial Intelligence could be effectively used to predict the growth inhibition patterns, anticancer activity, antifungal activity, genetic targets of the drug and so on. We have built a machine learning model that could potentially identify antifungal compounds from its chemical space. Our ML model is novel as it predicts antifungal compounds requiring fewer concentration (less half-maximal Inhibitory Concentration-IC50) in effectively inhibiting the growth.

Introduction

II - Introduction:

II.1 - Yeast as a tool in molecular medicine:

Yeast is one of the most widely used eukaryotic organisms for studying fundamental biological aspects like cell cycle control, DNA repair pathways, cancer studies, autophagy, aging etc. Yeast was the first model organism to be completely sequenced. 30% of the human disease causing genes are in homology with yeast. The growth of the yeast could be managed externally, fast doubling time (90 minutes), availability of deletion strains, tagged proteins, databases on PPI (Protein Protein Interaction, subcellular localization and gene regulation make it easier to understand the biological aspects at a faster rate. Yeast has been extensively used to study the activity of drugs, proteins and pathways being targeted by the drug, mutation rate of the cell towards resisting the drug, and understanding the physiological outcomes to the drug. For the absence of human targets in the yeast genome, human targets have been integrated in the yeast genome and studied[1].

II.2 - Yeast in cancer studies:

Cancerous cells usually have genetic alterations in pathways related to DNA damage repair and cell cycle control. These pathways reported in yeast are highly conserved in humans[2].

Table 1: Conservation of DNA repair pathways and proteins between yeast and human

Pathway	Function	*S. cerevisiae*	*H. sapiens*
Base excision repair	Damaged single DNA bases or a short strand DNA is excised, polymerase fills the gap and ligase connects the ends.	Apn1, Rad27, Ogg1	APE1, FEN1, OGG1
Nucleotide excision repair	Single-stranded DNA molecule of 24–30 nucleotides containing the lesion is excised, DNA polymerase fills the gap and ligase joins the ends.	Rad1, Rad10, Rad14, Rad4, Rad2	RAD1, ERCC1, XPA, XPC, XPG
Translesion synthesis	During damage, DNA polymerase zeta (Rev3 and Rev7) with Rev1 synthesize DNA in opposite sites of DNA lesions.	Rev1, Rev3, Rev7	Rev1, hRev3, hRev7
Mismatch repair	DNA mismatches are rectified post the proof reading by polymerase, recognizes the non-canonical base pair and replaces the offending nucleotide on the newly strand by excision repair mechanism.	Mlh1, Pms1	MLH1, PMS2
Homologous recombination	Repair DSBs by retrieving genetic information from an undamaged homolog (sister-chromatid or homologous chromosome). Accurate repair	Rad52, Mre11-Rad50-Xrs2	RAD52, MRE11-RAD50-NBS1
Non-homologous end-joining	Repair DSBs by direct ligation of DNA ends without any requirement for sequence homology. Mutagenic process.	Yku70, Yku80, Lif1, Dnl4, Mre11-Rad50-Xrs2	Ku70, Ku80, XRCC4, DNA ligase IV, MRE11-RAD50-NBS1

Table 2: Conservation of cell cycle control pathways and proteins between yeast and human

Function	*S. cerevisiae*	*H. sapiens*
Coates stretches of single stranded DNA synthesized by decoupling helicase and polymerase activities at stalled replication forks.	RFA	RPA
PIKK acts as a damage sensor and signal transducer.	Mec1	ATR
PIKK acts as a damage sensor and signal transducer.	Tel1	ATM
Recruits Mec1 (ATR) to regions of RFA (RPA)-coated ssDNA.	Ddc2	ATRIP
Involved in activation of Mec1-Ddc2 (ATR–ATRIP) complex.	Dpb11	TOPBP1
Sensor (RFC-like complex).	Rad24	Rad17
Damage sensor (PCNA-like protein), involved in the activation of PIKK family members.	Ddc1-Rad17-Mec3/Pso9	Ddc1-Rad17-Mec3/Pso9
Damage sensor (MRX/MRN complex), recruits Tel1 (ATM) to damage sites via its interaction with its terminal end-binding domain.	Mre11-Rad50- Xrs2	Mre11-Rad50-NBS1
Mediator, involved in Rad53 (CHK2) activation.	Rad9	BRCA1/ 53BP1
Mediator, a component of the replication fork that seems specifically to signal replication stress.	Mrc1	Claspin
Downstream kinase activated by PIKK proteins	Rad53	CHK2
Downstream kinase activated by PIKK proteins	Chk1	CHK1

As the yeast genome is well annotated, desirable genetic mutations leading to cancer phenotypes could be generated for cancer studies. Drugs are tested for anticancer activity in the mutant background[3].

II.3 - Limitations of yeast as a tool in drug studies:

The main limitation is the impermeable cell wall which hinders drug intake efficiently. Cell wall could be made permeable by using cell wall digesters like zymolyase. This is overcome using mutant strains like *ISE1, ISE2, PDR1, and SNQ2. ISE1* and *ISE2* are involved in ergosterol synthesis and their mutants have altered membrane composition. *PDR1 and SNQ2* mutant strains increase drug sensitivity via the disruption in the efflux pumps. Although the lack of human proteins in yeast could be compensated by incorporating the human gene into the yeast genome, it requires proper planning and execution of the experiment. Yeast can only partially replace mammalian models owing to its unicellularity, and lack of studies in angiogenesis, metastasis, and tissue invasion. But yeast is a powerful system for initial drug screening tests[4].

Literature Review

III - Literature Review:

III.1 - The NCI yeast Anticancer Drug screen:

The drug screen was started in 1977 to understand the single gene changes involved in cancers, which might play a major role in drug sensitivity. A non-essential gene can become essential due to the genetic alterations in the cancer-causing genes. Thus, inhibiting this converted essential gene can cause cell death in cancer lines. This synthetic lethality principle is used to screen potential drugs killing cancer cells effectively. The yeast strains used in this study mainly targeted cell cycle checkpoints and DNA repair pathways. The DNA damage repair pathways included in the study were DNA Double Strand Break repair (rad50 and rad52), ultraviolet excision repair (rad14), DNA mismatch repair (mlh1), O6-methylguanine removal (mgt1), and post-replication repair (rad18). Human orthologs of the above genes were found to be altered in cancer conditions. The genes targeting the cell cycle pathways used in the study were bub3 (a part of mitotic spindle checkpoint. *bub3* confirms the attachment of the chromosomes to the mitotic spindle before entering mitosis), CLN2oe (G1 cyclin, controlling the entry into S phase), mec2 (protein kinase required for the G2/M checkpoint and the S-phase checkpoint) and *sgs1* (involved in DNA replication, telomere function and, recombination). All the strains used in the study had mutation in the pdr1, pdr2 and *erg6* gene. *pdr1* and *pdr2* encode transcription factors controlling drug efflux pumps and erg6 alters the composition of the plasma membrane by varying the ergosterol synthesis, thus increasing the sensitivity to drugs. An additional *rad50EPP+* strain was used harboring the mutation in rad50 gene and wild types for pdr1, pdr2 and erg6 genes. Two wild types strains wt1 (wild type for rad14, rad18, rad50, rad52, bub3 and CLN2oe) and wt2 (wild type for mgt1, mlh1 and mec2) were used as controls along with rad50EPP+ in this study[1].

Table 3: NCI yeast strains	
Yeast strain	**Genotype**
wt1	*pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
rad18	*rad18$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
rad14	*rad14$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
rad50	*rad50$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
rad52	*rad52$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
bub3	*bub3$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
CLN2oe	*CLN2oe$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
mec2	*mec2$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
wt2	*pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
mgt1	*mgt1$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
mlh1	*mlh1$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
sgs1	*sgs1$^{\Delta}$; pdr1$^{\Delta}$; pdr2$^{\Delta}$; erg$^{\Delta}$*
rad50 EPP+	*rad50$^{\Delta}$*

III.2 - The NCI60 human tumor cell line anticancer drug screen:

National Cancer Institute 60 (NCI60) was developed in late 1980. The drugs are screened for their anticancer potential, the growth inhibition and the tumor cell killing patterns are studied using the NCI60 tool. Initially, it was designed to screen drugs for the cure of lung cancers. Since the drug activity was different in wild type cells like renal epithelial cells (became pan-sensitive) and fibroblasts (became pan-resistant), they extended the study to test other distinct cancer cell lines like ovarian, colon, breast, leukemia, renal, CNS, prostate, and, melanoma. They were looking for cell lines mimicking HeLa banding pattern. However, not many cell lines had the similar banding

pattern. Some cell lines like HCT-15(colon), TK-10 (renal) and UO-31 (renal) were included in the screen as they were MDR (Multi Drug Resistant) lines. The screens were also chosen based on their potential to propagate in athymic mice models. They also neglected cell lines having similar chromosome banding patterns and hence having unique cell lines[5].

MTT colorimetric assay is used for the growth inhibition studies. Metabolically active cells reduce MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide, a yellow tetrazolium salt to purple formazan crystals. The formazan crystals are solubilized and the absorbance is measured at 500-600 nm. Due to this solubility limitation, sulphorhod-amine B (SRB) assay was finally chosen to carry out growth inhibition assays. Cell lines sensitivity towards the drug was linearly related to the growth inhibition mechanism and cell killing. The compounds being high, an initial screening at a single concentration was carried out in cell lines. Further compounds showing high activity were chosen to screen on the full panel of cells at all 5 concentrations. Compounds meeting specific growth inhibition towards tumors have been proceeded for drug trial. Halichondrin B, a natural product showing sensitivity against the OVCAR-3 cell line was proceeded for pre-clinical trials. In 1990's, for rapid screening, a hollow fiber implantation model was developed. The tumor cell lines were mounted on the biocompatible hollow fibers. These mounted fibers were implanted into mice either subcutaneously or in the peritoneal cavity. The drugs to be tested were given intraperitoneally to the mice. The growth inhibition pattern was found on the recovered cells by MTT assays. Active compounds were further tested in Xenograft models. In 1998, for synthetic compounds, they pre screened only three cell lines NCI-H460 (llung cancer), MCF-7 (breast cancer), and SF-268 (glioblastoma) which had higher sensitivity. The active drugs from this study were further tested in NCI-60 cell lines which increased the efficiency and throughput. In 2001 the pre-screen study was done in 384 well plates using sear blue assay endpoint. The concentration of the drug used was 2.5 UM. They developed a COMPARE algorithm where the probable mechanism of the test compound could be found by comparing with the drugs in the database. One such implication of using the COMPARE algorithm is screening for compounds targeting P-glycoprotein.

Most of the MDR lines were found to express P-glycoprotein. Even the mechanism of anthrax inhibitor was found by COMPARE.

In 1997, NCI which was initially a drug screening pipeline turned to be a research tool for supporting cancer research communities[6].

Table 4: NCI60 cell lines

S.No	NCI60 strain	Cancer type	S.No	NCI60 strain	Cancer type
1	HS 578T	Breast cancer	31	NCI-H322M	Lung cancer
2	MDA-MB-231/ ATCC	Breast cancer	32	HOP-92	Lung cancer
3	MCF7	Breast cancer	33	A549/ATCC	Lung cancer
4	BT-549	Breast cancer	34	HOP-62	Lung cancer
5	T-47D	Breast cancer	35	SK-MEL-2	Melanoma
6	MDA-N	Breast cancer	36	LOX IMVI	Melanoma
7	SF-268	CNS cancer	37	SK-MEL-28	Melanoma
8	SF-539	CNS cancer	38	MDA-MB-435	Melanoma
9	SNB-19	CNS cancer	39	M14	Melanoma
10	U251	CNS cancer	40	UACC-62	Melanoma
11	SF-295	CNS cancer	41	SK-MEL-5	Melanoma
12	SNB-75	CNS cancer	42	UACC-257	Melanoma
13	HT29	Colon cancer	43	MALME-3M	Melanoma
14	COLO 205	Colon cancer	44	IGROV1	Ovarian cancer
15	HCC-2998	Colon cancer	45	NCI/ADR-RES	Ovarian cancer
16	KM12	Colon cancer	46	OVCAR-5	Ovarian cancer
17	HCT-116	Colon cancer	47	OVCAR-3	Ovarian cancer
18	HCT-15	Colon cancer	48	OVCAR-8	Ovarian cancer
19	SW-620	Colon cancer	49	SK-OV-3	Ovarian cancer
20	K-562	Leukaemia	50	OVCAR-4	Ovarian cancer
21	MOLT-4	Leukaemia	51	PC-3	Prostate cncer

22	CCRF-CEM	Leukaemia	52	DU-145	Prostate cncer
23	RPMI-8226	Leukaemia	53	TK-10	Renal cancer
24	HL-60(TB)	Leukaemia	54	SN12C	Renal cancer
25	SR	Leukaemia	55	UO-31	Renal cancer
26	EKVX	Lung cancer	56	CAKI-1	Renal cancer
27	NCI-H460	Lung cancer	57	786-0	Renal cancer
28	NCI-H226	Lung cancer	58	RXF 393	Renal cancer
29	NCI-H522	Lung cancer	59	A498	Renal cancer
30	NCI-H23	Lung cancer	60	ACHN	Renal cancer

III.3 - Artificial Intelligence (AI) in drug discovery:

Artificial Intelligence based drug discoveries are quicker, effective, and less expensive. It is said that almost 2.6 billion dollars and 10 years is invested in developing a therapeutic drug and only 1 out of 10 drugs pass the trial for approved therapeutic drugs from phase II. Most drugs fail because of numerous reasons like toxicity, poor pharmacokinetics and lack of clinical efficacy. Most of the pharmaceutical companies have opted for AI. Pfizer uses IBM's AI-Watson, Sanofi uses UK start-up Exscientia's AI, Genentech uses AI of GNS Healthcare in Cambridge and BERG (https://bpgbio.com) uses their AI system and identified BPM31510, a potential drug for pancreatic cancer which is currently under Phase II clinical trials[7].

Tabel 5: Artificial Intelligence driven pharmaceutical companies			
AI company	**Technology**	**Partnered Company**	**Drugs against**
Atomwise	DL screening based on molecular structure data	Merck	Malaria
Benevolent	DL and NLP of research literature	Johnson & Johnson	Multiple
Berg	DL screening of biomarkers from patient data	Berg	Multiple
Exscientia	Bispecific compounds via Bayesian models of ligand activity from drug discovery data	Sanofi	Metabolic diseases
GNS Healthcare	Bayesian probabilistic inference for investigating efficacy	Genentech	Oncology
Recursion, Salt Lake City, Utah	Cellular phenotyping via image analysis	Sanofi	Rare genetic diseases
Numerate	DL from phenotypic data	Takeda	Oncology, gastroenterology and CNS disorders
twoXAR, Palo Alto, California	DL screening from literature and assay data	Santen Pharmaceuticals, Osaka, Japan	Glaucoma

III.4 - Existing Artificial Intelligence based models in antifungal drug discovery:

Thirty percent of *Candida* genus are highly resistant to the existing antifungal drugs. Gao et al in UCSD developed ML based models for virtually screening drugs targeting CaFKS1, 1,3-beta-glucan synthase's subunit which is involved in the cell wall synthesis. ML based model was on chemical descriptors and attained 96.72% accuracy[8].

Deep learning model followed by molecular docking, X-score and similarity search methods was generated to screen chemical compounds targetting the dihydrofolate reductase of *Candida albicans*[9]. Deep Learning model was built based on Deep

Screening, a user-friendly web server developed by Liu et al.[10]. DeepScreening generated the target-focused new libraries.

Synergistic drug combinations are effective in reducing the drug resistance, increasing the treatment efficacy, and reducing drug dosage. Chen et al developed an algorithm NLLSS (Network-based Laplacian regularized Least Square Synergistic drug combination prediction) to predict synergistic drug combinations based on drug-target interactions, and drug chemical structures. 7 out of 13 predicted combinations were confirmed to be antifungal via experiments [11].

III.5 - Existing Artificial Intelligence based models in anticancer drug discovery:

There are several existing AI based prediction models for identifying potential carcinogens and computer vision based models in prediction cancer based on images. Most of the AI based models in predicting anti-cancer drugs hover around drug repurposing as it saves time in clinical trials.

Kuenzi et al. developed DrugCell[12], a deep learning model trained on the responses of 684 drugs in 1235 cancer cell lines. Drug responses in a mutational background helps in predicting synergistic combinations of drugs via synthetic lethality concept. Synthetic lethality is when two genes being singly mutated doesn't have any lethal effect in cells, but when doubly mutated, leads to lethality. This concept of two genes is replaced by one gene and one drug targeting the other lethal pair gene. It is exploited to design drugs that could specifically kill cancer cells carrying out mutations in the gene. As normal cells are devoid of mutation, the drug won't affect the normal cell and thus being specific in inhibiting cancer cells.

Kadurin et al used Adversarial AutoEncoders-AAE on dose response data of 6252 compounds in MCF-7 cell line from the NCI-60 and developed a deep learning model to generate novel molecular fingerprints. The 7-layer AAE architecture was developed with

the latent middle layer serving as a discriminator. Input and output of the AAE uses a vector of binary fingerprints and concentration of the molecule. Neuron responsible for the growth inhibition percentage was also introduced in the latent layer. Several drugs were screened on the trained model[13].

Methodology

IV - Methodology:

IV.1 - Experimental design of NCI yeast Anticancer Drug screen:

The drug screen consists of three steps.. In stage 0, 97261 compounds were screened at a concentration of 50 µM in duplicates in 6 strains. Among the 6 strains, 3 yeast strains had single mutations in *rad50*, *mec2* and *bub3* genes and the rest 3 strains had double mutations in *rad18* + *mlh1, sgs1* + *mgt1* and *rad14* + *CLN2oe.* Compounds (nearly 14466) with at least 70 % growth inhibition (GI) in at least one strain proceeded for stage 1. 16885 Compounds (freshly added compounds along with stage 0) were tested at concentrations 5 and 50 µM in duplicates in the same strains used in stage 0. The toxic compounds inhibiting growth in all the strains were neglected. The compounds showing at least five-fold growth inhibition difference between the least and most sensitive strain were proceeded for stage 2. In stage 2, the compounds were tested in all 13 strains at concentrations 1.2, 3.6, 11, 33, and 100 µM.

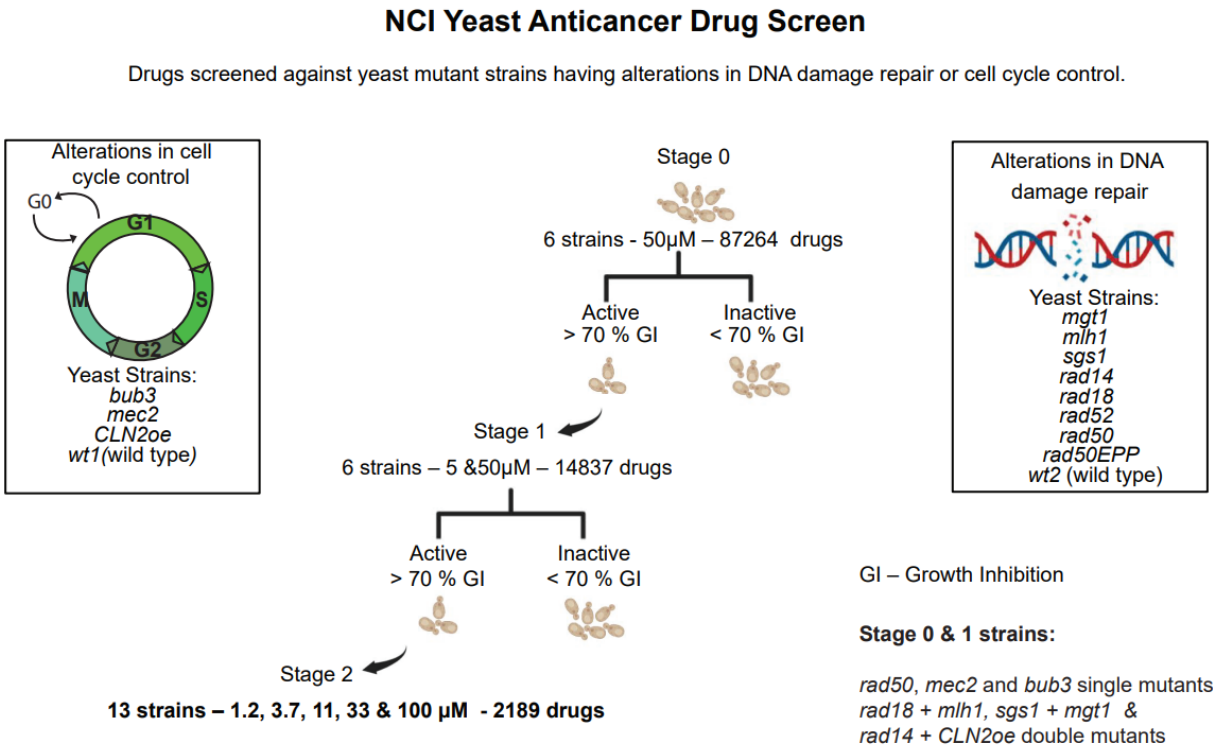


Figure 1: NCI Yeast Anti-cancer Drug Screen stages

IV.2 - Experimental design of NCI60 Growth Inhibition Assay:

Initially, a single dose assay (5-10 Molar concentration) was used to filter the compounds with better anti-proliferative activity. The values for the drug response data denote the growth relative to zero drug concentration (control) and the time when the number of cells is zero. A value between 0 to 100 indicates growth, and the value less than zero denote lethality. For example, a value of 100 means no growth inhibition, a value of 60 means 40% growth inhibition, zero means no net growth, -40 means 40% lethal and 100 means no cells survived. The filtered compounds were then subjected to a five dose assay. The same annotation of values is used for five dose assay.

IV.3 - Five-dose Assay:

The NCI60 cell lines are grown in RPMI (Roswell Park Memorial Institute) 1640 medium with 2mM L-glutamine and 5% fetal bovine serum. With plating densities ranging from 5,000 to 40,000 cells/well, the cells are inoculated into 96-well microtiter plates in 100 µL. Later incubated at 95% air, 37°C, 100% relative humidity and 5% CO2 for 24 hours prior to drug addition.

Post incubation, plates are fixed *in situ* with TCA(tricarboxylic acid), representing a measurement of the cell population at the time of drug addition (Tz). Drugs were dissolved in dimethyl sulfoxide at 400-fold times the test concentration and frozen. The frozen drug is thawed and diluted to twice the test concentration with a complete medium containing 50 µg/ml gentamicin. 100 µl of this aliquot is added to the wells containing 100 µl of the medium, resulting in the required final drug concentrations.

Post-drug addition, the plates were incubated at 5% CO2, 95% air, 100% relative humidity, and 37°C for 48 hours. Cells are fixed *in situ* 100µl by 10% tricarboxylic acid and incubated at 4°C for 60 minutes. After discarding the supernatant, the plates are washed five times with tap water and air dried. Sulforhodamine B (SRB) solution (100µl) at 0.4% (w/v) in 1 % acetic acid is added and incubated for 10 minutes at room temperature. Post-staining, unbound dye is removed by washing. Bound stain is

solubilized with 10 mM trizma base, and the absorbance is read on an automated plate reader at a wavelength of 515nm.

The percentage growth is calculated from the absorbance measurements growth at time zero of drug addition (Tz), control growth (C), and test growth in the presence of drug at the five concentration levels (Ti) at each of the drug concentration levels. Percentage growth is calculated as:

[(Ti-Tz)/(C-Tz)] x 100 for concentrations for which Ti ≥Tz

[(Ti-Tz)/Tz] x 100 for concentrations for which Ti<Tz.

The IC50 values are interpolated from the percentage of cell growth as a fraction of control cell growth.

IV.4 - IC50 Dataset Description:

IV.4.1 - NCI yeast Anticancer Drug screen:

IC50 values (in μM) for the NCI yeast Anticancer Drug screen were taken from the NCI website (https://dtp.cancer.gov/YeastData/nscsearch). IC50 values for 1971 compounds on 12 strains (*rad50, rad52, rad50EPP+, rad14, rad18, sgs1, mec2, mgt1, wt1, wt2, bub3, CLN2oe)* except *mlh1* were extracted from the website.

IV.4.2 - NCI60 Growth Inhibition Assay:

The IC50 for the compounds in NCI60 cell lines have been downloaded from the https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data. The drug screening experiments were also done in additional cell lines apart from the usual 60 cell lines.

The raw data had columns for denoting the NSC number for compound identification, concentration unit of the drug tested,NCI cell line used and , interpolated IC50 value, its average and standard deviation.

IV.4.3 - Processing the IC50 values:

IC50 values for the NCI yeast Anticancer Drug screen was complete and processing wasn't required.

The growth inhibition assay was repeated several times by serial diluting different initial test concentrations. IC50 were interpolated for each experiment and reported. Due to redundancy in the assay, we considered the average IC50 values for the assays with highest initial concentration. 1099 compounds were found to be common between yeast anticancer drug screen and NCI60 Growth Inhibition Assay. Approximately 7% of the NCI60-IC50 matrix was incomplete. This incomplete matrix was completed by the matrix completion method.

IV.5 - Matrix Completion Method:

Different matrix completion methods like Soft Impute, Iterative SVD, KNN were used to complete the NCI60-IC50 matrix. Fancyimpute 0.7.0 package was used. The best matrix completion method was chosen based on R square, Mean Absolute Error and Mean Square Error values.

IV.5.1 - SoftImpute:

Softimpute works by nuclear norm regularization and fits a low rank matrix approximation to the matrix with missing values. Missing values are filled by current guesses, and the values filled are optimized via soft threshold SVD.

IV.5.2 - KNN (K Nearest Neighbors):

KNN finds similar data points near the missing value and completes the missing value with the average value of all the similar points.

IV.5.3 - Iterative SVD (Singular Value Decomposition):

The matrix is filled by iterative low rank SVD. It is similar to SVDimpute from Missing value estimation methods for DNA microarrays by Troyanskaya et. al.

IV.6 - Evaluation of Matrix Completion Method:

NCI60 IC50 values for 231 drugs were complete and this matrix was used for evaluating the best method. Randomly introducing NA values in 20% of the data and the NA values are filled by matrix completion methods. This predicted 20 % of the value is compared with the true value and the average of R^2, MAE and RMSE values for 10 iterations was calculated. Soft impute was found to be the best method.

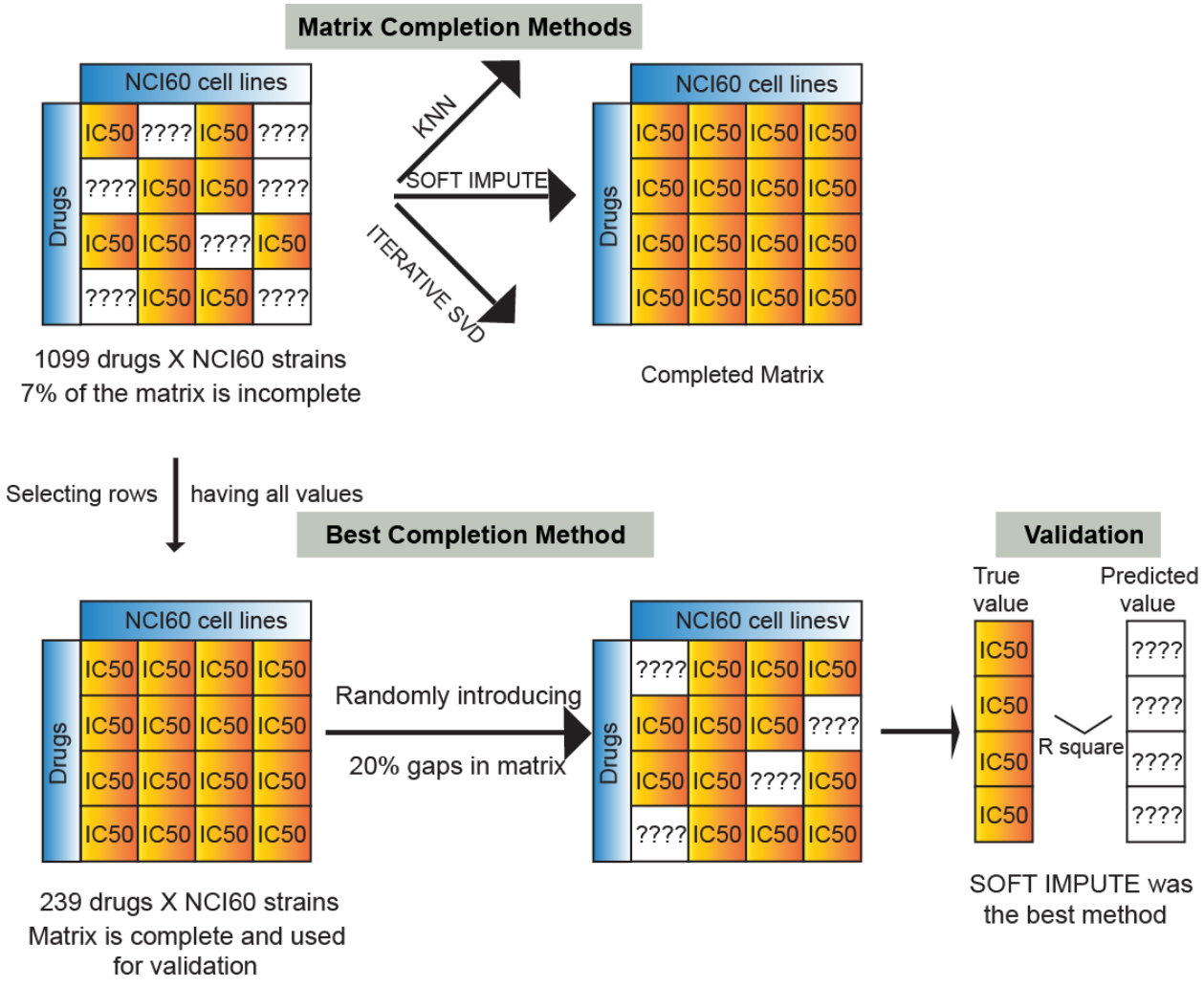


Figure 2: Workflow of matrix completion method - part1

IV.6.1 - R2:

The coefficient of determination or R2 value indicates the variation between the predicted and the true value. R2 ranges from 0 to 1. 1 indicates no variation in the predicted values when compared to the true values and 0 indicates complete variation in the predicted values when compared to the true values

R2 = 1 – residual sum of squares / total sum of squares

R2 = 1 – ∑ (true value- predicted value)2 /∑(true value – mean of true value)2

IV.6.2 - MAE (Mean Absolute Error):

MAE is also called as L1 loss function and its the average of the absolute errors. Absolute error is the difference in magnitude between the predicted value and the true value. MAE ranges from 0 to ∞ and smaller values of MAE is preferred as it indicates very less difference between the true and predicted value

MAE = ∑ |true value- predicted value)| / number of observations

IV.6.3 - MSE (Mean Square Error):

The Mean Squared Error measures the proximity of a regression line to a set of data points. Average of the squared errors is used for calculating MSE .

A higher MSE denotes that the data points are dispersed widely around its central moment whereas a smaller MSE suggests the opposite. A lower MSE is preferred as it means smaller the error and better the predicted value is.

MSE = (1/n) * Σ(actual value – predicted value)2

n is the sample size.

IV.7 - Mutual Information (MI):

The filled matrix of NCI60 IC50 is compared with the IC50 NCI yeast study based on the Mutual Information (MI) score. Mutual information gives the measure of how two random variables are mutually dependent. It considers the non-linear relationship as well unlike pearson correlation which considers only the linear relationship between two variables.

Mutual Information of two random variables X and Y, whose joint distribution is defined by P(X, Y) is given by

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log P(x, y) / P(x)P(y)$$

In this definition, P(X) and P(Y) are the marginal distributions of X and Y obtained through the marginalization process. X could be the IC50 in yeast strains and Y could be IC50 in NCI60 human tumor cell lines.

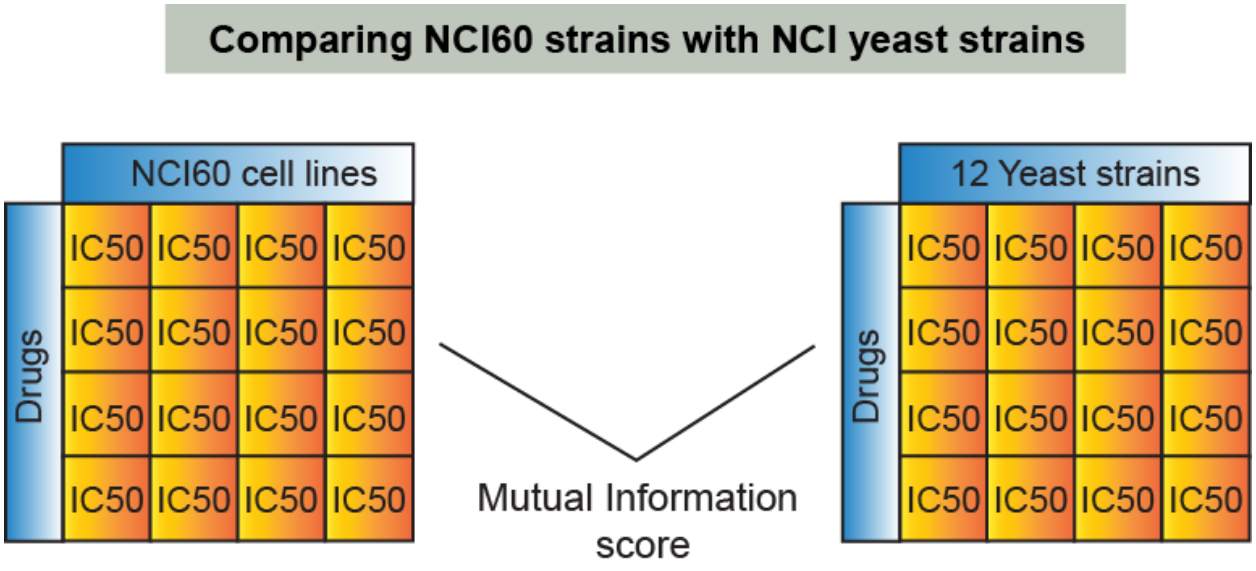


Figure 3: Workflow of matrix completion method - part 2

IV.8 - GIPCRT (Growth Inhibitory Percentage) Datasets of NCI yeast Anticancer drug screen:

The stage 2 drug response datasets were taken from the https://wiki.nci.nih.gov/display/NCID. The CSV file had the information on the strain used, NSC number for compound identification, concentration of the drug used and the average of the growth inhibition measurement.

To predict whether a drug could inhibit the yeast growth from the trained GIPCRT, unsupervised clustering was done on GIPCRT and subsequent clusters would give information on the drugs' growth inhibition ability.

IV.8.1 - Unsupervised classification based on GIPCRT:

There were 13 yeast strains including 2 wild type strains. Unsupervised classification via k means was done on their growth inhibitory percentage. The number of optimal classes was determined based on elbow methods and silhouette score. The subsequent classes were used as training datasets in supervised classification to predict the growth inhibition ability.

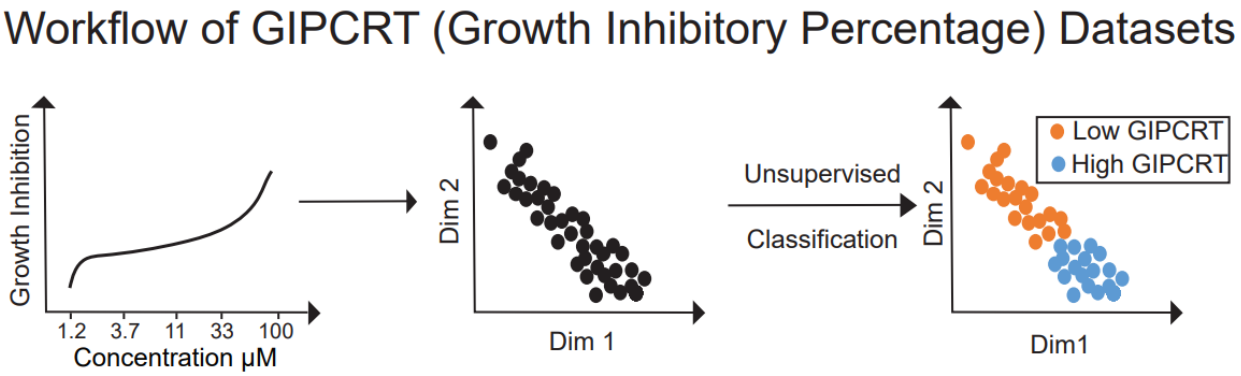


Figure 4: Unsupervised Classification based on GIPCRT datasets

IV.8.1.1 - K-means classification:

The growth inhibition pattern having five columns is dimensionally reduced to two dimensions. The data points are then represented in a 2D plot. k distinct data points representing k classes are randomly selected. The rest points are assigned to a particular class based on lesser Euclidean distance. The mean (centroid) of each class is found. We again classify the data points based on their distance from the centroid. We keep repeating the classification until the labels of the classes remain unchanged. The optimal k is found by the elbow and silhouette method. The unclassified dataset is split into classes based on the similarities in their growth pattern within a class and dissimilarities between classes**.**

IV.8.1.2 - Elbow method:

Sum of squared errors (SSE) within the clusters is measured in finding the optimal clusters. SSE varies largely when k increases and converges later. The k at which it converges forms an elbow. Elbow is used to find the optimal k for the unsupervised classification.

IV.8.1.3 - Silhouette coefficient (SC):

Silhouette coefficient measures the similarities within a cluster and dissimilarities between clusters. The value ranges between -1 to +1. +1 indicates that the classification is the best. 0 indicates overlapping classes and -1 indicates worst classification. K at which silhouette coefficient is maximum is the optimal k for the unsupervised classification.

$$S(i) = \frac{a(i) - b(i)}{\max\{a(i), b(i)\}}$$

$$SC = \text{mean}(S(i))$$

S(i) => silhouette score of a datapoint i

a(i) => mean distance of i from all the data points within a cluster

b(i) => mean distance of i from all the data points of other clusters

IV.8.2 - Machine Learning – supervised classification:

The clusters/labels from the unsupervised clustering were based on the growth inhibitory percentage. The chemical space of the drugs along with the labels are used to build machine learning models in predicting the growth inhibition of a drug given its SMILES. The chemical space of the drugs was extracted from the SMILES by the feature generation method Signaturizer 1.1.11. Signaturizer extracts 3200 features and hence feature selection method like Boruta was used to select significant features contributing to the classification.

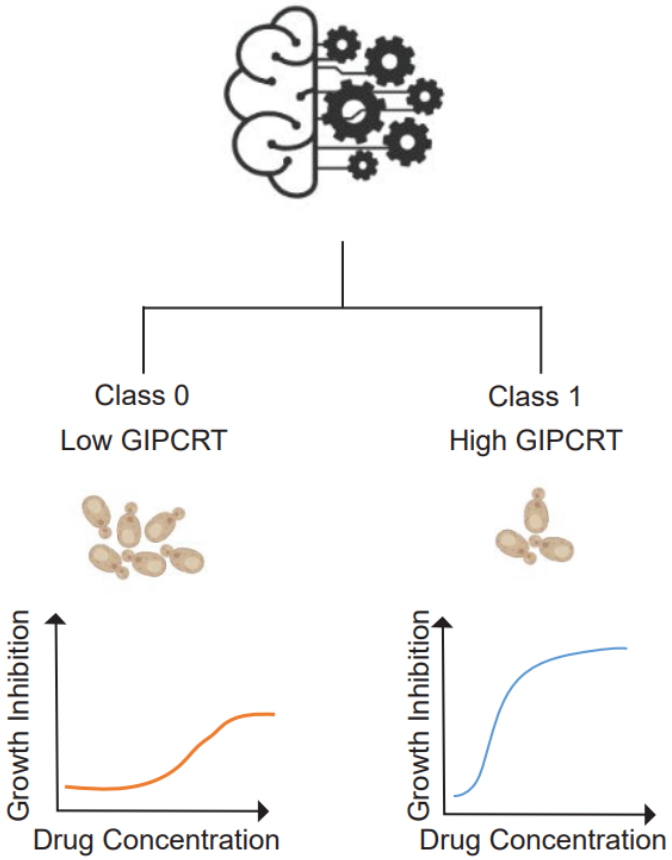


Figure 5: Supervised Classification on the Growth Inhibition datasets

Machine Learning models like Random Forest (RF), Stochastic Gradient Descent (SGD), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), Extra Tree(ET) and AdaBoost (AB) were employed to train the model. The models were trained on 80 % datasets and validated on the remaining 20 % datasets. The scoring metrics used were Kappa, Recall, Precision, F1 score, Accuracy and AUC-ROC score. Average of the 10 fold cross validation kappa scores were considered in choosing the best model and the parameters of the best model were as well chosen based on the average Kappa scores. The final model was trained on the entire dataset and saved.

IV.8.2.1 - SMILES:

SMILES (Simplified Molecular Input Line Entry System) is a string format representing the chemical structure of the compounds. SMILES format could be read by the computer. SMILES supports all elements in the periodic table and follows hydrogen suppression i.e CH3-CH3 is written as CC. Double bond is represented as "=" and triple bond as "#". There are prescribed representations for side chain, branched elements, metals, aromatic rings etc. Softwares extracts features (numerical vectors) from the SMILES.

IV.8.2.2 - Signaturizer:

Signaturizer generates features from SMILES for machine learning. Signaturizer extracts both bioactivity and chemical descriptors (physicochemical and structural properties). The signaturizer vector size is 3200. It has five main classes describing the chemistry (A), targets (B), networks (C), cells (D) and clinics (E) and each main class is further divided into 5 subclasses generating 128 features. The total number of features sums up to 3200 (5 x 5 x 128 = 3200)[14].

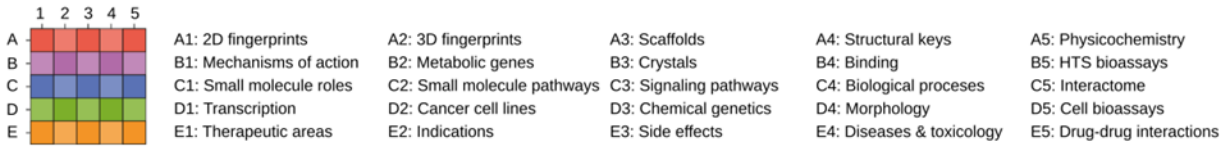


Figure 6: Features in Siganturizer

IV.8.2.3 - Boruta:

Boruta is a feature selection method and it selects the features that are relevant for the classification model. Boruta works on the Random Forest algorithm. It creates random shadow copies of the features (noise) It iteratively removes the test features that are statistically less relevant when compared to noise.

IV.8.2.4 - Random Forest (RF):

Bootstrapped dataset is constructed with the same size (rows and columns) of the original dataset. Bootstrap dataset is made by randomly choosing samples i.e. rows from the original dataset with repetition. Decision tree is created by randomly choosing the variables i.e. columns/features from the bootstrapped dataset. The chosen variables are the root node. Each root node is further branched with a set of random variables called internal nodes/branches. The last internal node which is not further branched is called a leaf node/ leaf. Many such decision trees are made. The majority of the decision trees is considered as the output of RF. RF is robust and has better accuracy and problem-solving ability because of many decision trees.

IV.8.2.5 - Stochastic Gradient Descent (SGD):

In SGG, few samples in the dataset are chosen randomly instead of the complete dataset used in Gradient Descent (GD). SGD becomes computationally inexpensive when compared to GD. The parameters are altered iteratively until the cost function is minimized.

IV.8.2.6 - K-Nearest Neighbour (KNN):

A similar approach to KNN unsupervised clustering is used. But here k and the labels are known. We decide the label for the test sample based on the majority of the neighbors around the test sample.

IV.8.2.7 - Support Vector Machine (SVM):

SVM creates the best decision boundary (hyperplane) that can segregate classes in n-dimensional space. Datapoints are clearly separated from the hyperplane. Several hyperplanes are possible and the best hyperplane is chosen based on the maximum margin. The observations that lie close to the margins are called support vectors. If the shortest distance between the support vectors and the margin is maximum, that margin is the best hyperplane.

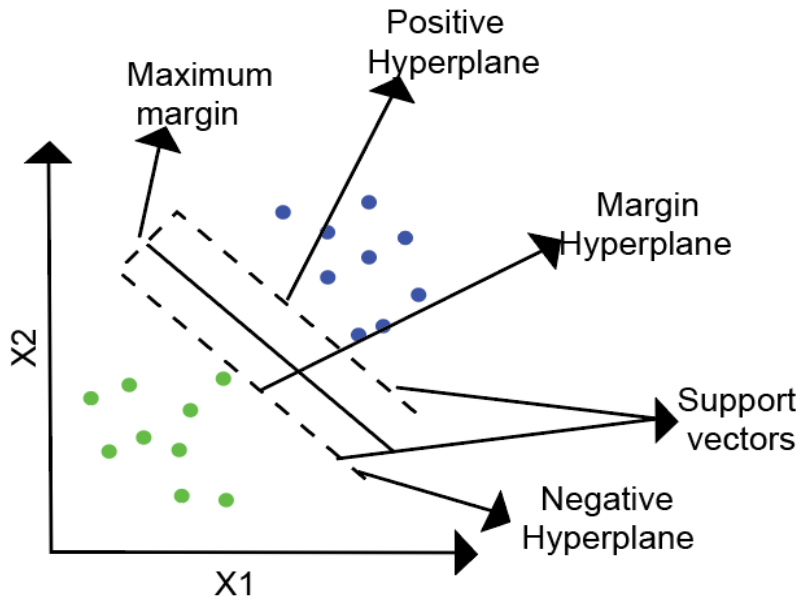


Figure 7: Schema for Support Vector Machine

IV.8.2.8 - Decision Tree (DT):

Decision tree is made with a root node (feature) at the top having the minimum impurity (majority of the samples in a column are classified correctly based on that feature), internal nodes (features/columns of the dataset) and the leaf (final output of the classification). The impurity of the node is calculated by any one of the methods - entropy, information gain and gini impurity. Pruning is also done in order to remove the unnecessary branches.

IV.8.2.9 - Extra Tree (ET):

Extremely randomized trees classifier is very similar to random forest. ET considers the multiple decision trees constructed from the entire dataset unlike random forest considering the decision trees constructed from boot strapped datasets.

IV.8.2.10 - Ada Boost (AB)/ Adaptive Boosting:

Tree with one node and two leaves is called a stump and several stumps are constructed. Stumps are weak learners as the decision is made based on one feature only. Initially, equal weights are assigned to all the samples. The stumps are arranged one below the other in increasing Gini index. Total error of the stump is calculated. 0 means no error (perfect stump) and 1 means maximum error (imperfect stump). The performance of the stump or amount of say is calculated from the total error. The weights are updated based on the amount of say. The whole process is iterated until less error is seen.

IV.8.3- Evaluation metrics for ML classifiers:

All the Machine learning metrics were calculated from the confusion matrix.

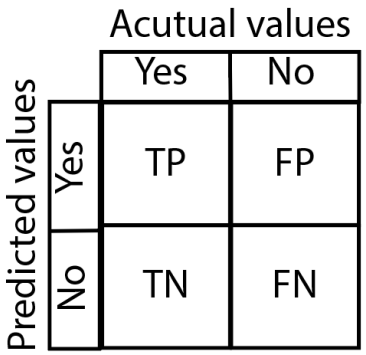


Figure 8: Confusion matrix for binary classifier

IV.8.3.1 - Kappa:

Kappa compares the observed accuracy with the expected accuracy (by random chance). Unlike accuracy, it takes imbalance class into consideration. In Observed Accuracy, we add the number of instances the classifier synced with the ground truth label and divide by the total number of instances. Expected accuracy is the number of instances of each class, along with the number of instances that the classifier agreed with the ground truth.

Kappa = (observed accuracy - expected accuracy)/(1 - expected accuracy)

IV.8.3.2 - Weighted Recall:

Calculates the proportion of the actual classes identified. Weighted average of the recall of all the classes is used.

Recall 0 = TP / TP+FN considering positive to be one class

Recall 1 = TN / TN+TP considering negative to be another class

Weighted average of Recall 0 and Recall 1 is considered for the model evaluation.

IV.8.3.3 - Weighted Precision:

Calculates the proportion of the correctly identified class. Weighted average of precision of all the classes is used.

Precision 0 = TP / TP+FP considering positive to be one class

Precision 1 = TN / TN+FN considering negative to be another class

Weighted average of Precision 0 and Precision 1 is considered for the model evaluation.

IV.8.3.4 - F1 score:

F1 score is the harmonic mean of Precision and Recall. It evaluates the class wise performance rather than the overall performance of the model.

F1 score = 2 x Precision x Recall / Precision + Recall.

IV.8.3.5 - Accuracy:

It evaluates the overall performance of the model.

Accuracy = No of correct predictions / total number of predictions

IV.8.3.6 - AUC- ROC:

ROC (Receiver Operating Characteristics) is the probabilistic curve and AUC is the measure of separability. AUC ranges from 0 to 1.Higher the AUC values, the better the model predicts 0 as 0s and 1 as 1. ROC curve is obtained by plotting the true positive rate vs false positive rate.

IV.9 - Unsupervised classification based on IC50 followed by supervised classification:

The IC50 values for 12 yeast strains except *mlh1* strain were available.The unsupervised classification of IC50 values by k means gave two classes. One class of drugs with higher IC50 values and one class with lower IC50 values. The significance of the classification is validated by the Kolmogorov–Smirnov test.

Similar machine learning approach used in GIPCRT classification was used in prediction of higher and lower IC50 values.

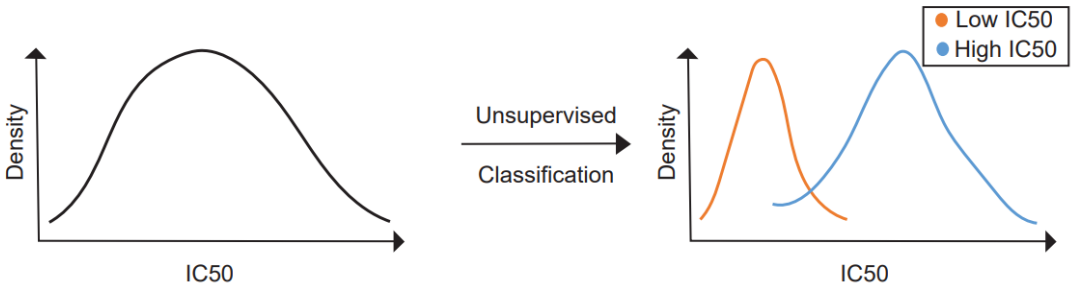


Figure 9: Unsupervised classification based on IC50 datasets

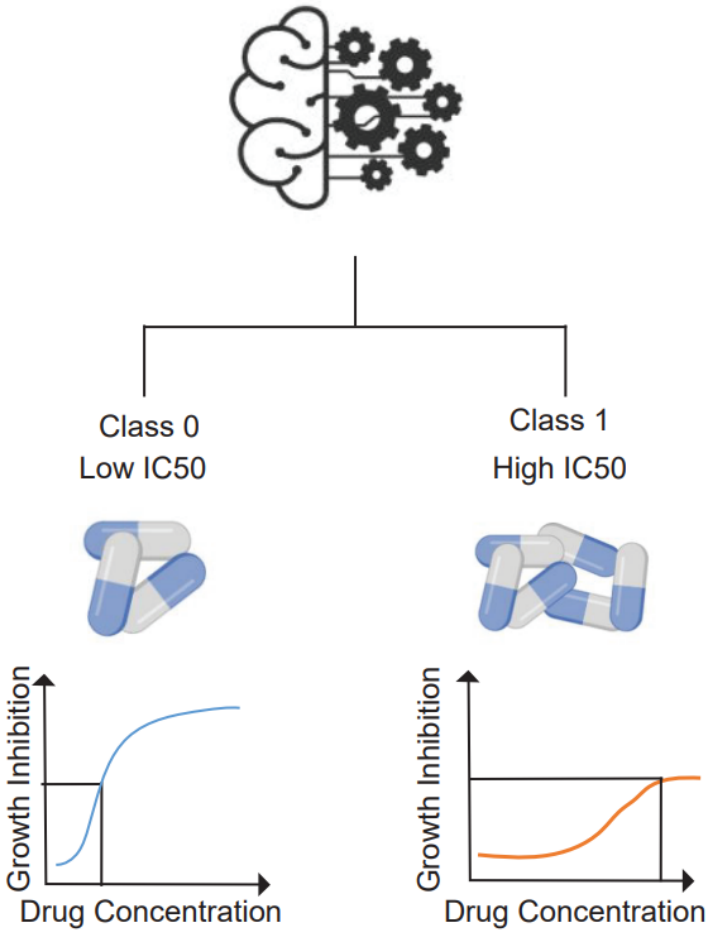


Figure 10: Supervised classification on IC50 datasets

IV.10- Machine Learning models in predicting novel antifungal agents:

By combining the ML classifiers in predicting the class with growth inhibition ability with less IC50 values, around 15000 drugs from Drug Bank were tested.The top 25 drugs in each classifier are further validated for their potential antifungal properties.

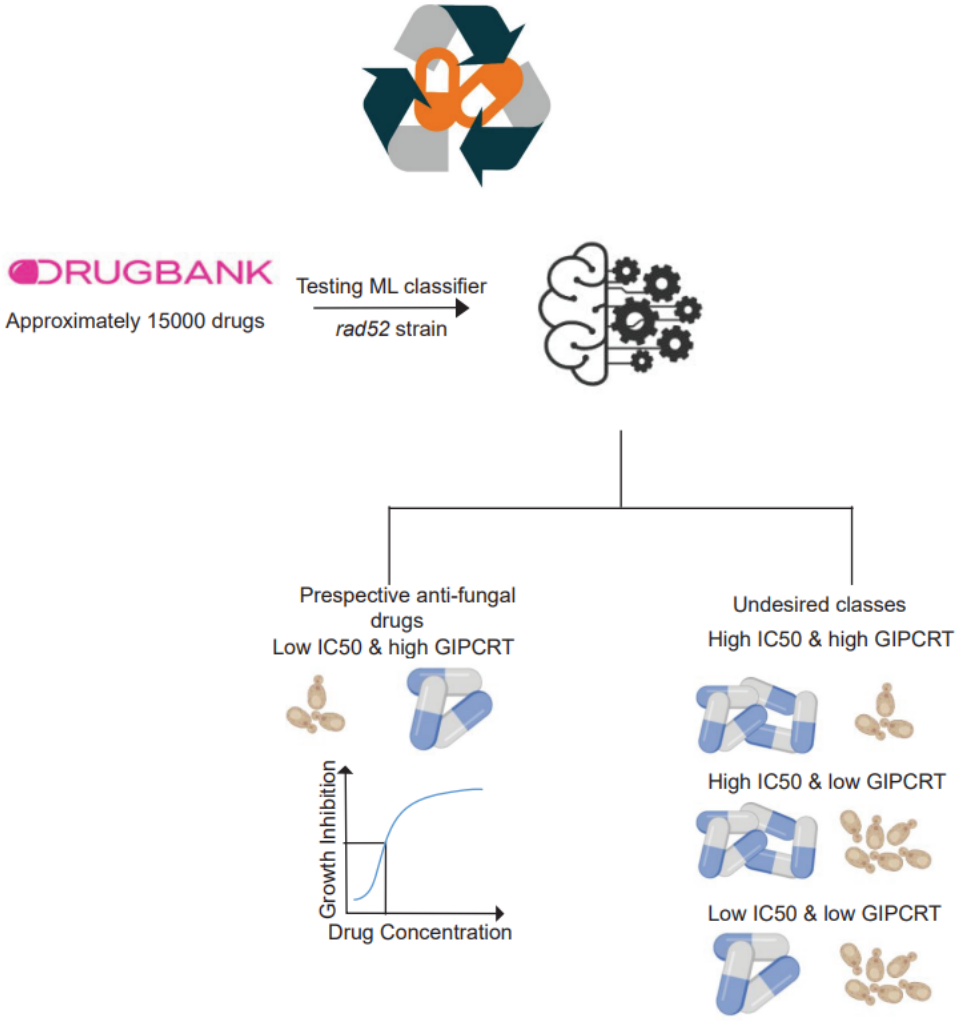


Figure 11: Machine Learning models in predicting novel antifungal compounds

Results and Discussions

V - Results and Discussions:

V.1 - Matrix Completion:

The incomplete IC50 NCI60 cell lines were completed by different matrix completion methods like Softimpute, KNN and Iterative SVD. The best method was Softimpute as it had higher R^2, lower MSE and MAE when compared to KNN and IterativeSVD.

Table 6: Evaluation of Matrix Completion methods			
Evaluation metrics	**Softimpute**	**KNN**	**Iterative SVD**
R^2	0.72	0.57	0.62
MSE	5723.75	8912.32	7858.26
MAE	18.63	20.77	21.41

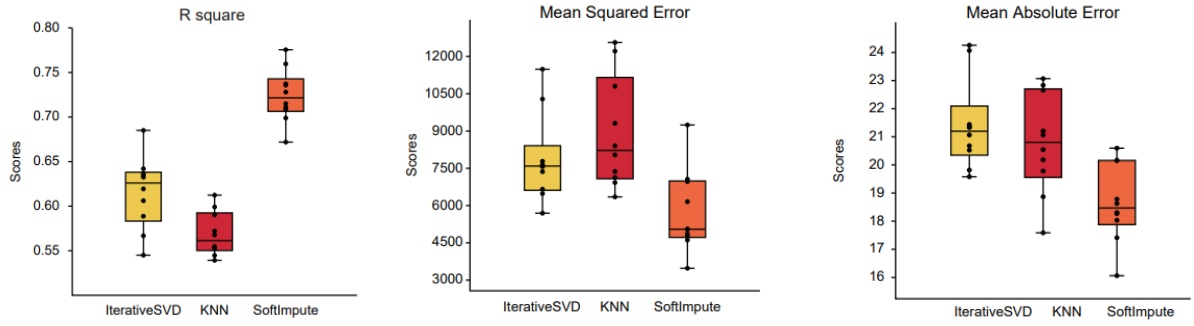


Figure 12: R square, Mean Squared Error and Mean Absolute error values in Matrix Completion methods

V.2 - Mutual Information:

MI scores and Pearson correlation coefficient between 12 yeast strains and the human NCI60 cell lines are shown in the heat map. The heat map shows that *rad52* yeast strain had the highest MI score against 56 out of 60 NCI60 cell lines. *rad50* and *rad14* yeast strains had the highest MI score in the two NCI60 cell lines. Hence *rad52* would be a good substitute for carrying out anticancer studies in human cell lines/ animal models.

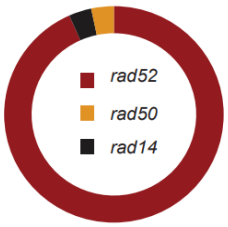


Figure 13: Donut plot of highest scores of yeast strain

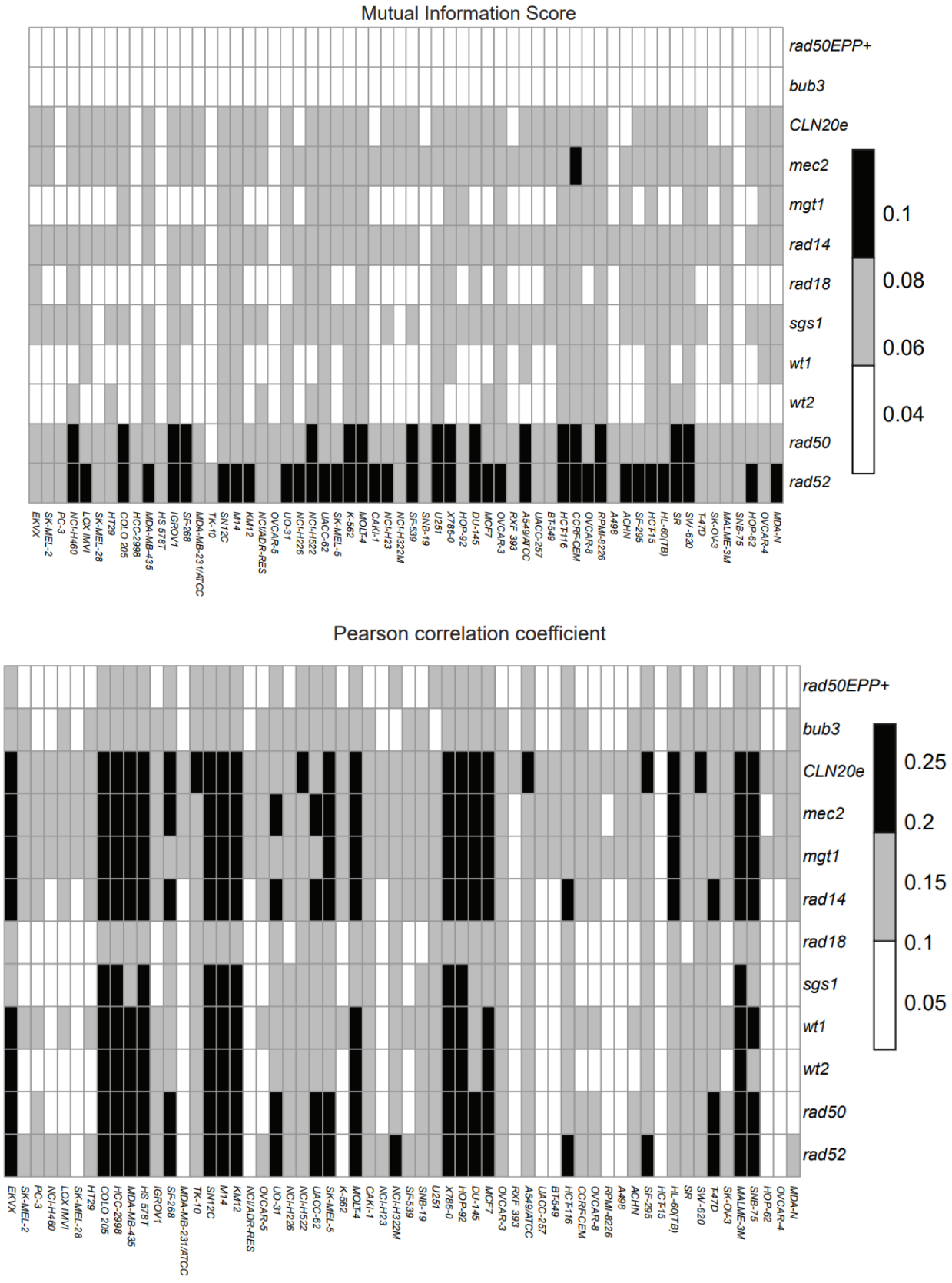


Figure 14: MI scores and Pearson correlation scores between yeast and NCI60 strains

Similarly MI scores were calculated within the NCI60 cell lines. The calculated MI score between two different NCI60 cell lines was less when compared the MI score between the same strain. It suggests that NCI60 cell lines were distinct from each other thus serving the purpose for a wide spectrum of NCI60 cell lines.

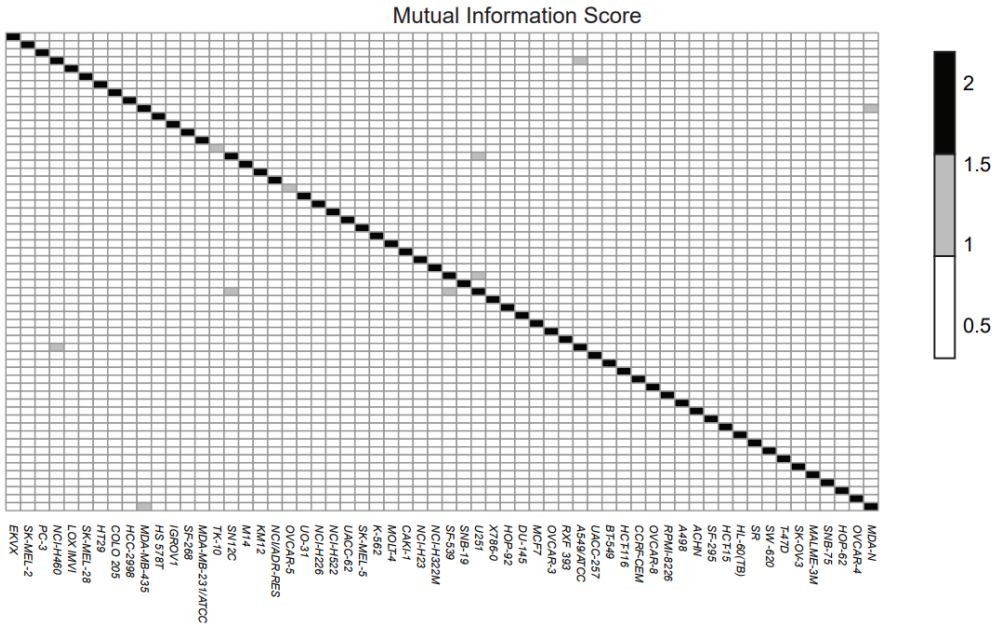


Figure 15 : MI scores within NCI60 cell lines -Part 1

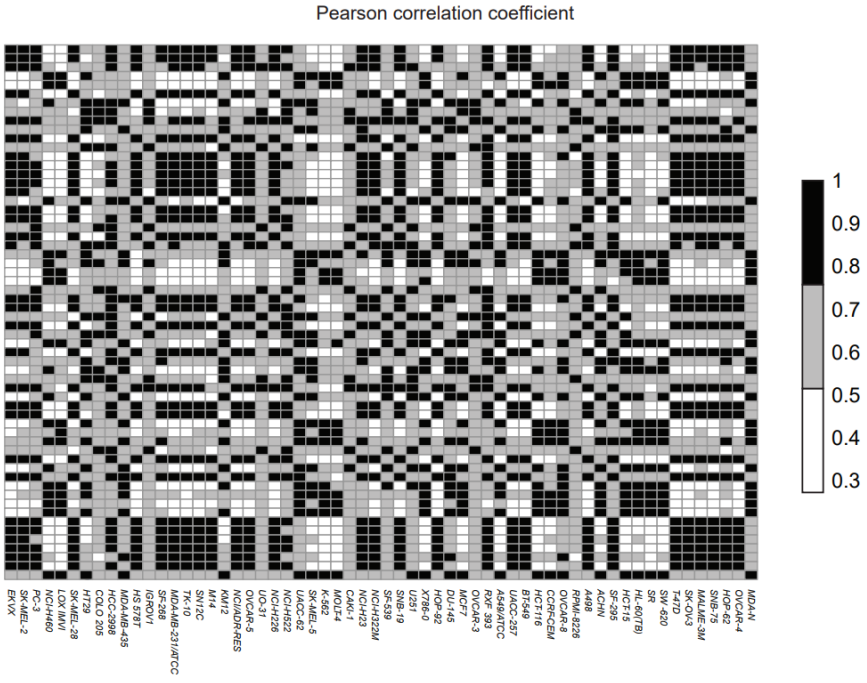


Figure 15: Pearson correlation scores within NCI60 cell lines - Part 2

MI scores were calculated within the 12 yeast strains as well. The MI score between *rad52* and *rad50* was the second highest as the MI scores between the same yeast strains was the highest. It suggests that the *rad52* and *rad50* are similar thus proving the genetic similarity between *rad50* and *rad52* genes. This similarity is as well depicted in the first heat map showing *rad50* to be the second substitute for NCI60 cell lines after *rad52*.

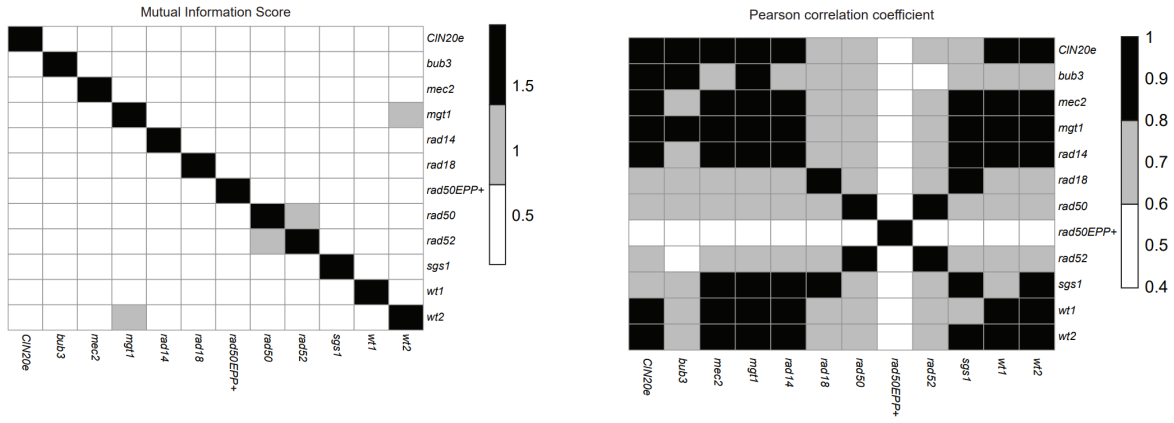


Figure 16: MI scores and Pearson correlation scores within the yeast strains

The density distribution plot of the MI scores between yeast strains and NCI60 cell lines is shown for alternative visualization of the heat map. rad52 and rad50 are on the right side of the plot indicating that they have higher MI scores with NCI60 cell lines when compared to the rest yeast strains with NCI60 cell lines.

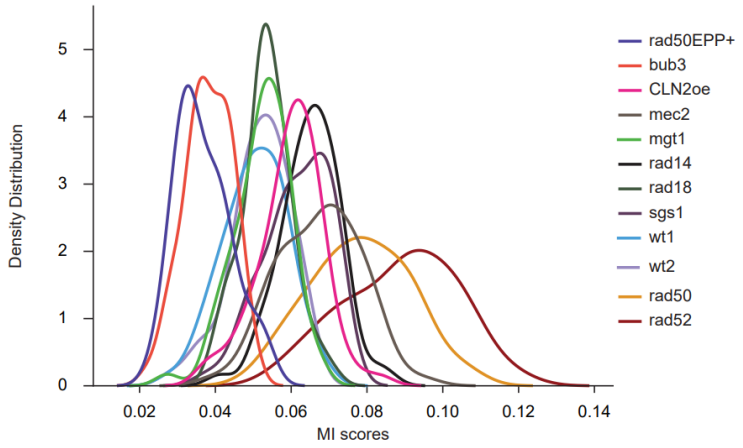


Figure 17: Density distribution plot of MI scores between yeast and NCI60 cell lines

Kolmogorov–Smirnov test performed between the MI scores of *rad50* with NCI60 and other yeast strains with NCI60 suggests that *rad50* and *rad52* follow similar distribution as d-value was 0.36 (closer to 0), whereas it was >0.76 with other strains. The test was significant with a p value <0.05. This further adds to the point that *rad52* and *rad50* are similar in their functioning.

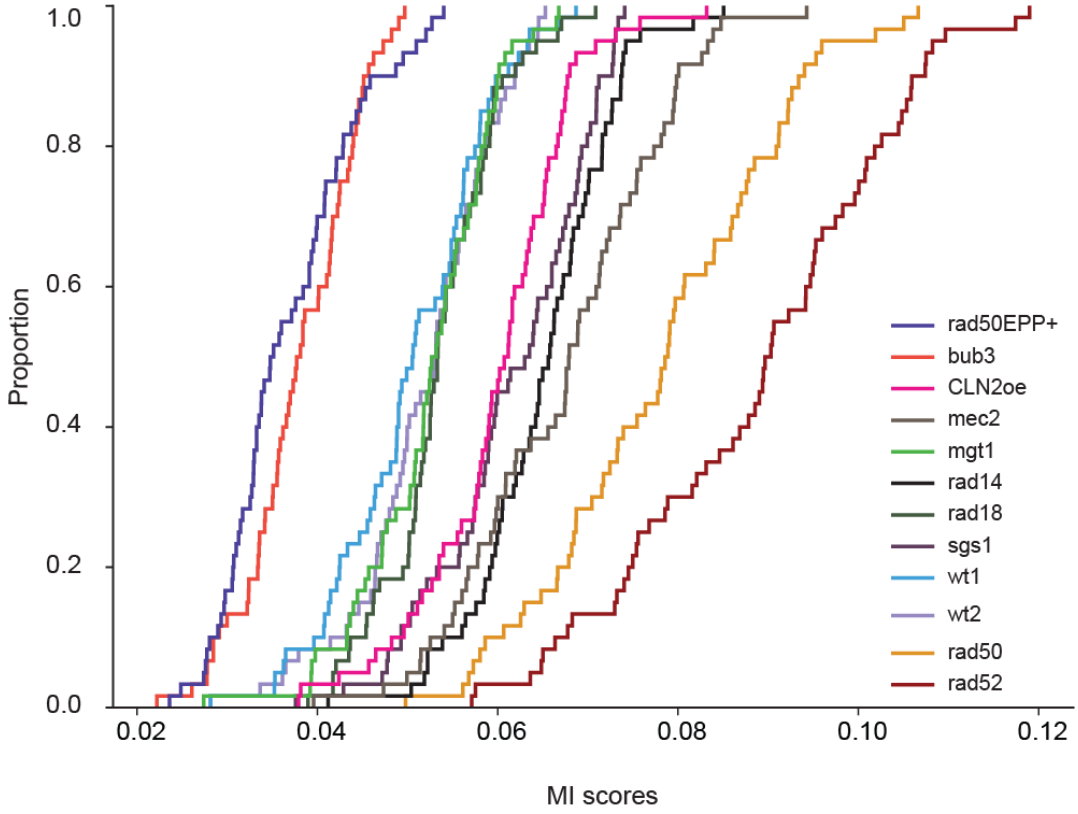


Figure 18: ECDF plot of MI scores between yeast and NCI60 cell line

V.3 - Unsupervised classification based on GIPCRT:

K-means unsupervised classification on GIPCRT yielded two clusters in all 13 yeast strains. The optimal number of clusters is found by the elbow method and silhouette score. In all 13 strains, the elbow point is seen when k=2 and the maximum silhouette score was also at k=2. The concentration box plots between two clusters clearly show the distinction between the two classes. The tSNE plot shows the two classes in dimensional space. The classes from this unsupervised classification are used for supervised classification of the drugs from their chemical space. The best model turned out to be Random forest in all 13 strains. The best model and its parameters were chosen based on the 10 fold cross validation of Kappa and Accuracy.

Table 7: Best parameter for Random Forest Classifier on GIPCRT values

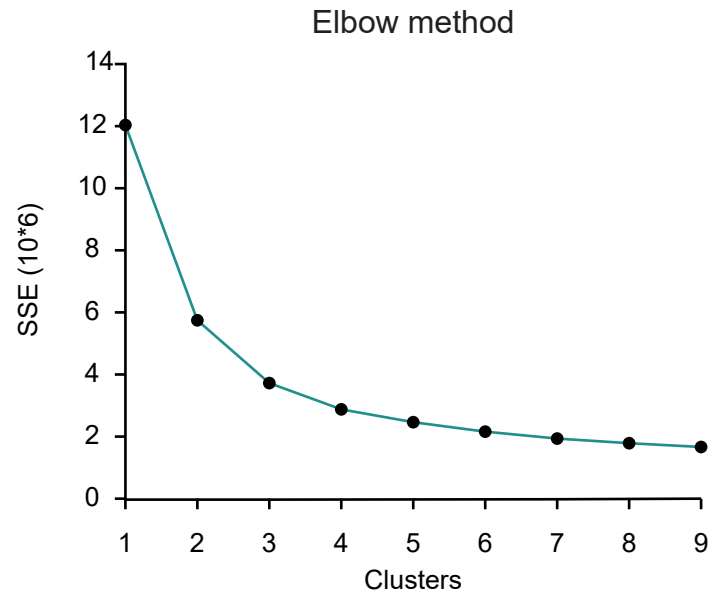
Yeast Strain	Silhouette score	Best Parameters in Random Forest
rad52	0.45	criterion: gini, max_depth: 30, max_features: auto, n_estimators: 150
rad14	0.45	criterion: gini, max_depth: 10, max_features: auto, n_estimators: 50
rad18	0.45	criterion: gini, max_depth: 10, max_features: auto, n_estimators: 50
rad50EPP+	0.48	criterion: gini, max_depth: 20, max_features: auto, n_estimators: 150
sgs1	0.43	criterion: entropy, max_depth: 20, max_features: log2, n_estimators: 100
mlh1	0.45	criterion: gini, max_depth: 20, max_features: auto, n_estimators: 150
mgt1	0.46	criterion: gini, max_depth: 20, max_features: log2, n_estimators: 700

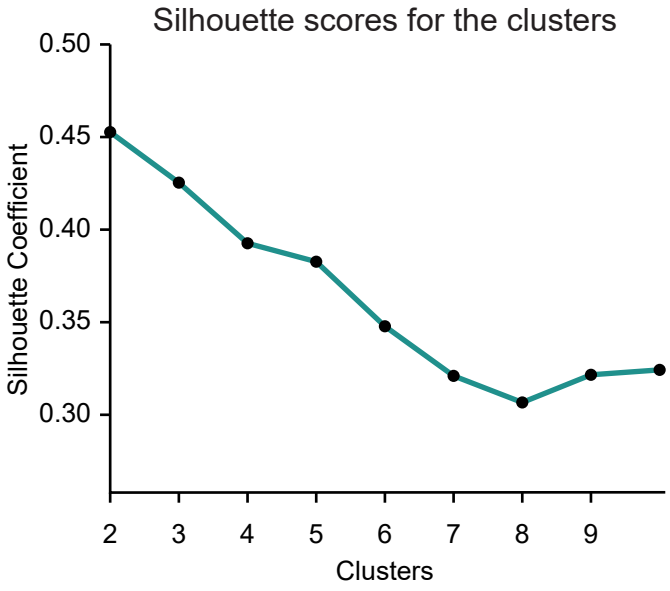
mec2	0.46	criterion: gini, max_depth: 30, max_features: auto, n_estimators: 800
bub3	0.45	criterion: entropy, max_depth: 20, max_features: log2, n_estimators: 800
CLN2oe	0.48	criterion: gini, max_depth: 30, max_features: auto, n_estimators: 400
wt1	0.48	criterion: gini, max_depth: 20, max_features: log2, n_estimators: 600
wt2	0.46	criterion: gini, max_depth: 10, max_features: auto, n_estimators: 100

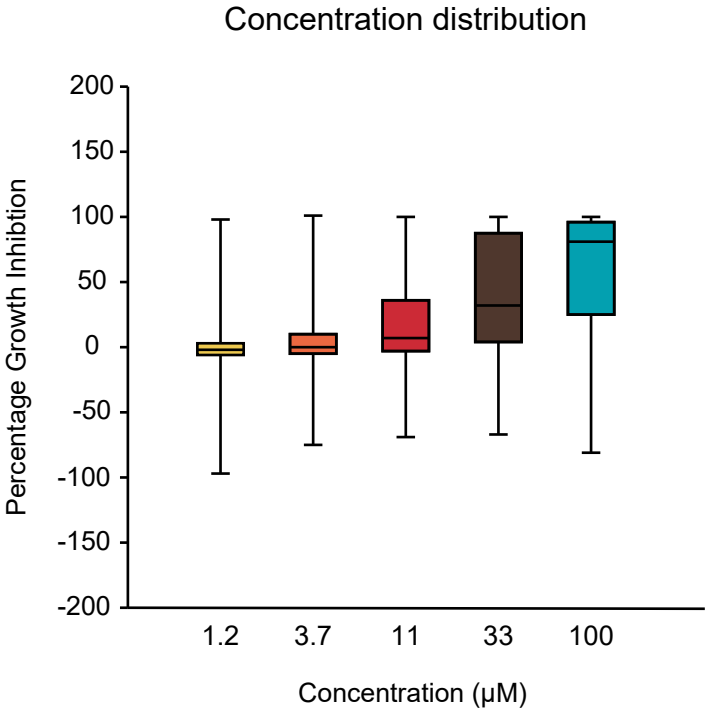
Table 8: 10 CV evaluation metrics for all 13 yeast strains based on GIPCRT

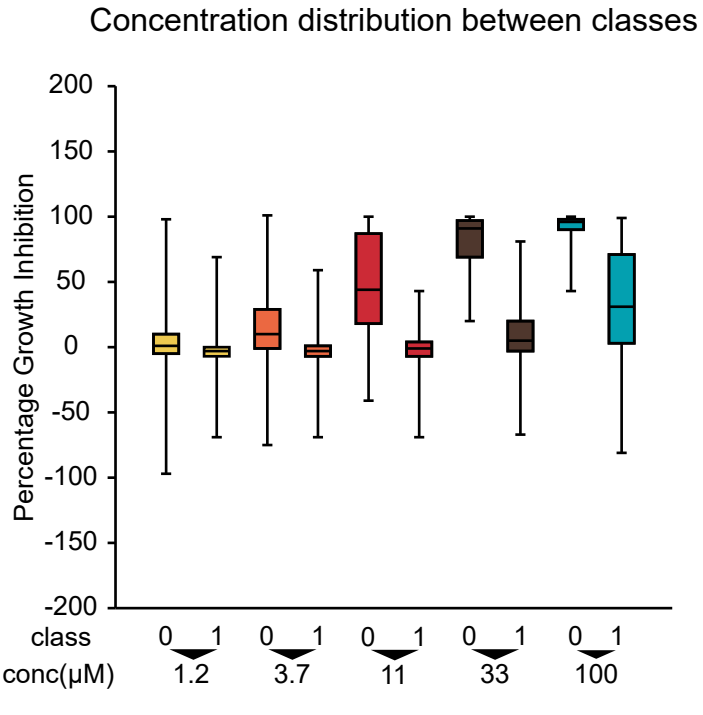
Yeast strain	Recall	Precision	F1 score	Kappa	Accuracy	AUC
rad52	0.64	0.64	0.63	0.25	0.64	0.69
rad50	0.63	0.63	0.63	0.26	0.63	0.68
rad14	0.65	0.64	0.64	0.23	0.65	0.67
rad18	0.63	0.63	0.63	0.24	0.63	0.68
rad50EPP+	0.68	0.68	0.68	0.34	0.68	0.73
sgs1	0.63	0.62	0.62	0.24	0.63	0.67
mgt1	0.63	0.62	0.62	0.22	0.63	0.67
mec2	0.65	0.65	0.64	0.25	0.65	0.68
bub3	0.62	0.61	0.61	0.22	0.62	0.66
CLN2oe	0.64	0.63	0.63	0.21	0.64	0.67
wt1	0.65	0.64	0.64	0.22	0.65	0.67
mlh1	0.63	0.62	0.62	0.21	0.63	0.66
wt2	0.63	0.63	0.63	0.22	0.63	0.66

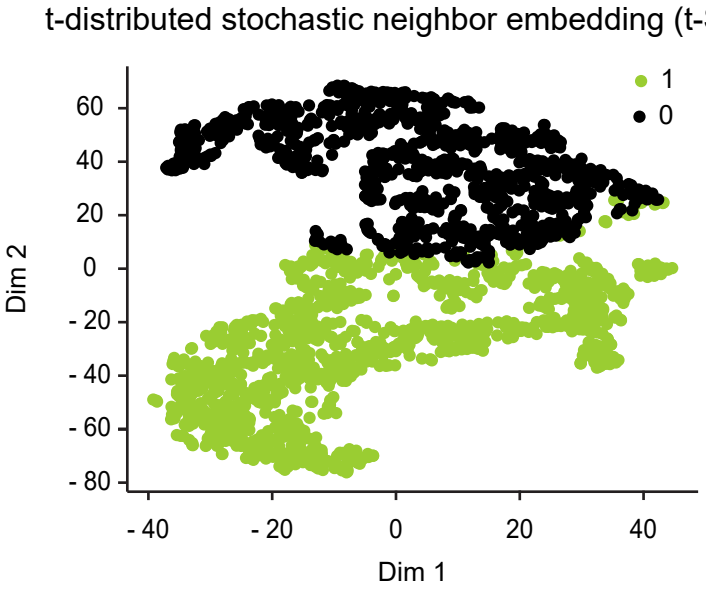
Figure 20.1: Unsupervised Classification based on GIPCRT - *bub3* strain











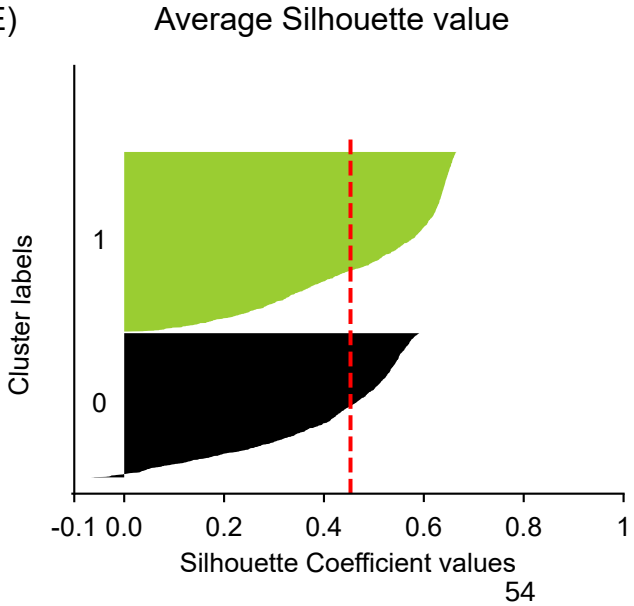


Figure 20.2: Unsupervised Classification based on GIPCRT - *CLN2oe* strain

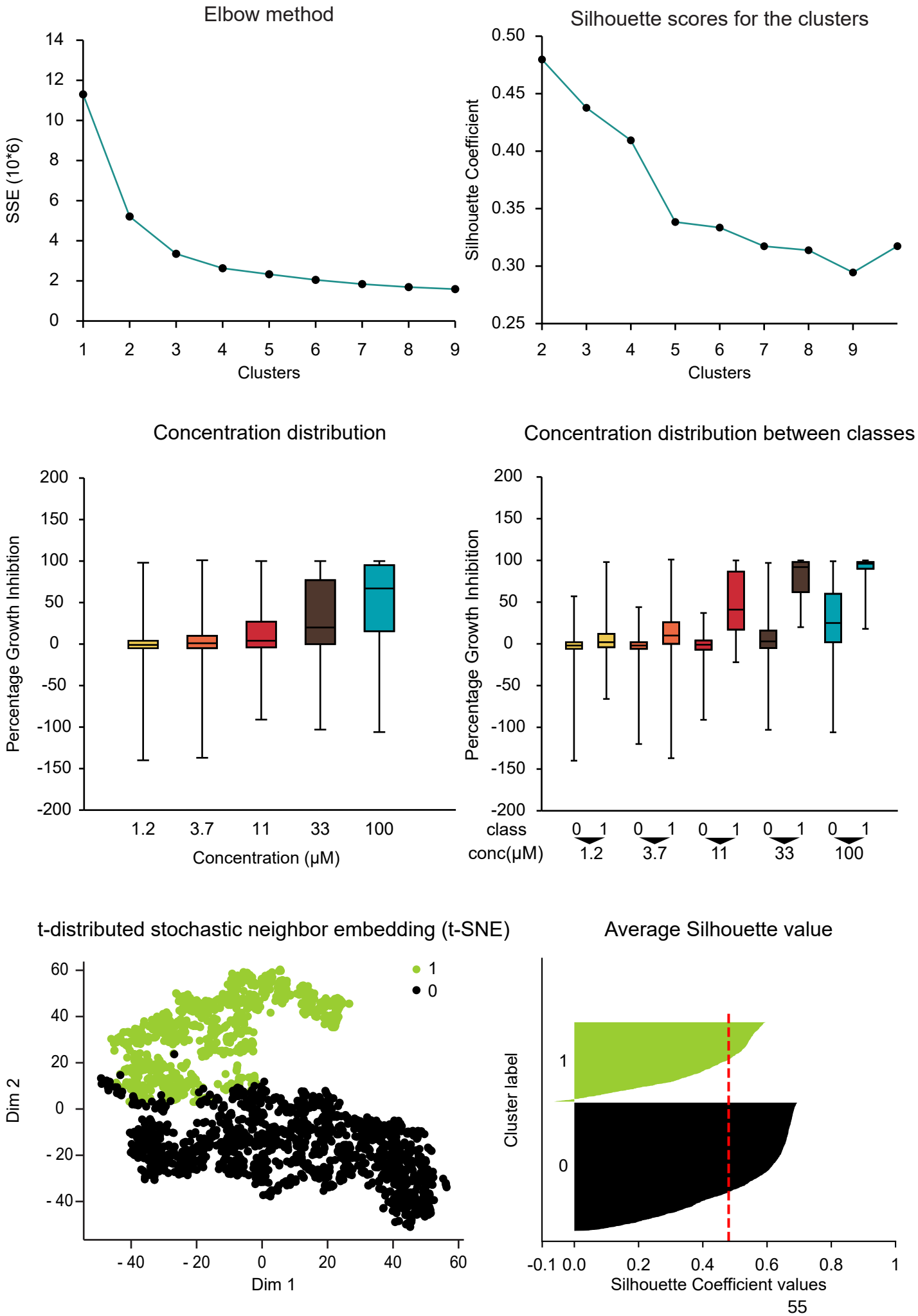
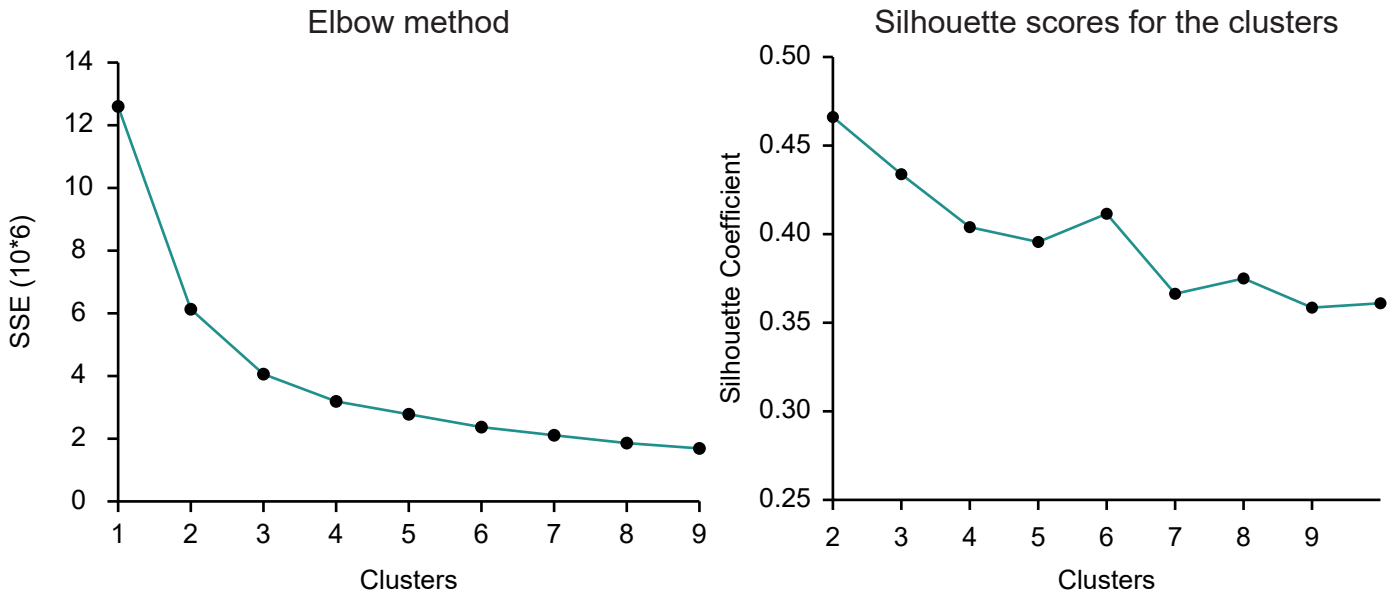
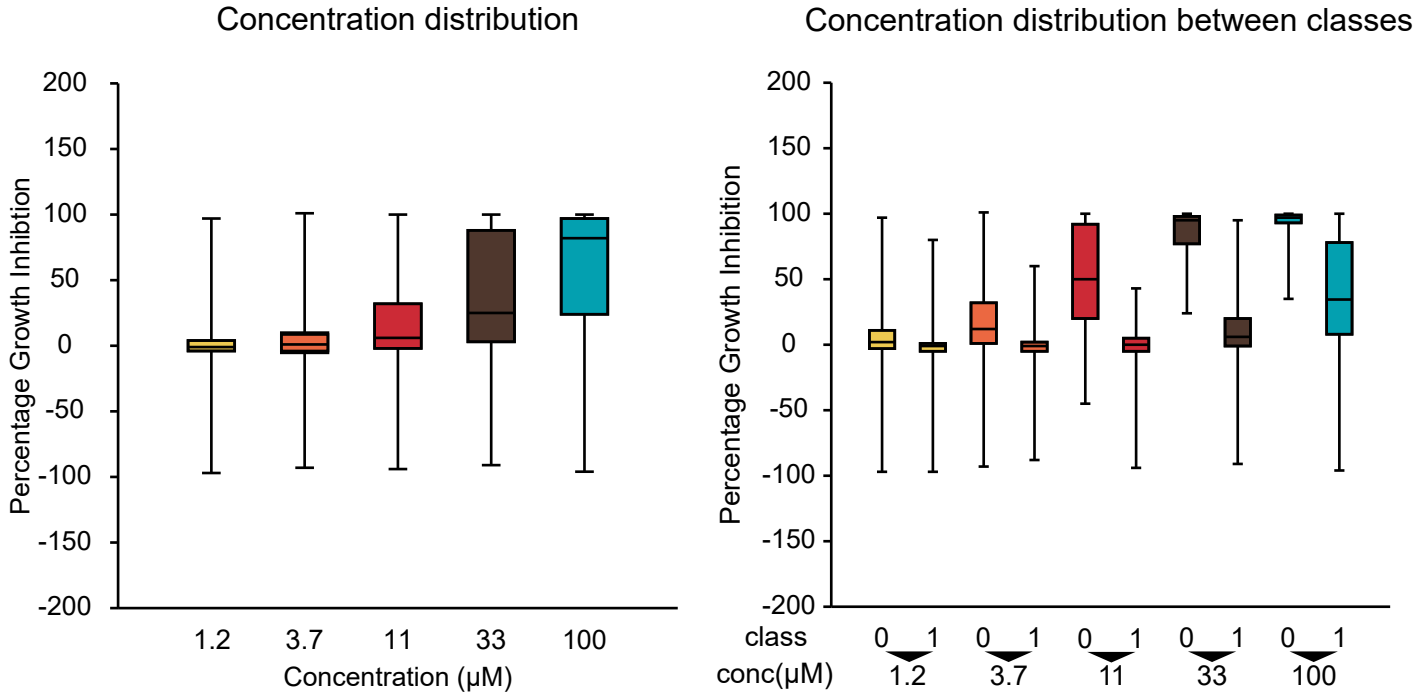


Figure 20.3: Unsupervised Classification based on GIPCRT - *mec2* strain





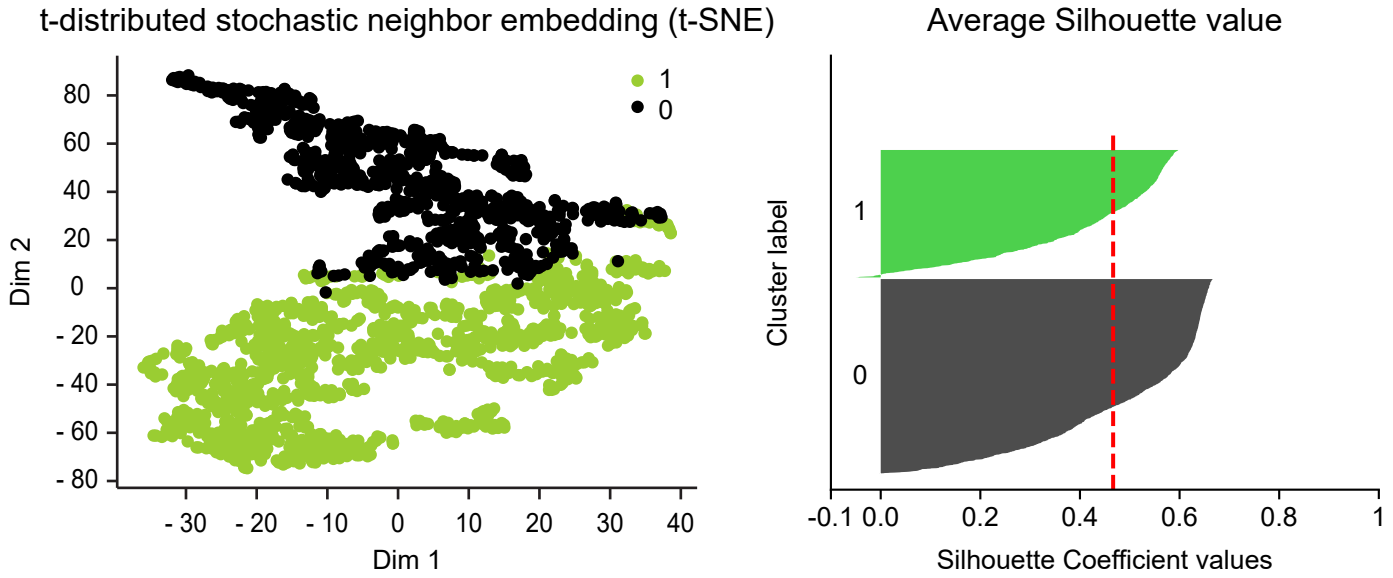


Figure 20.4: Unsupervised Classification based on GIPCRT - *mgt1* strain

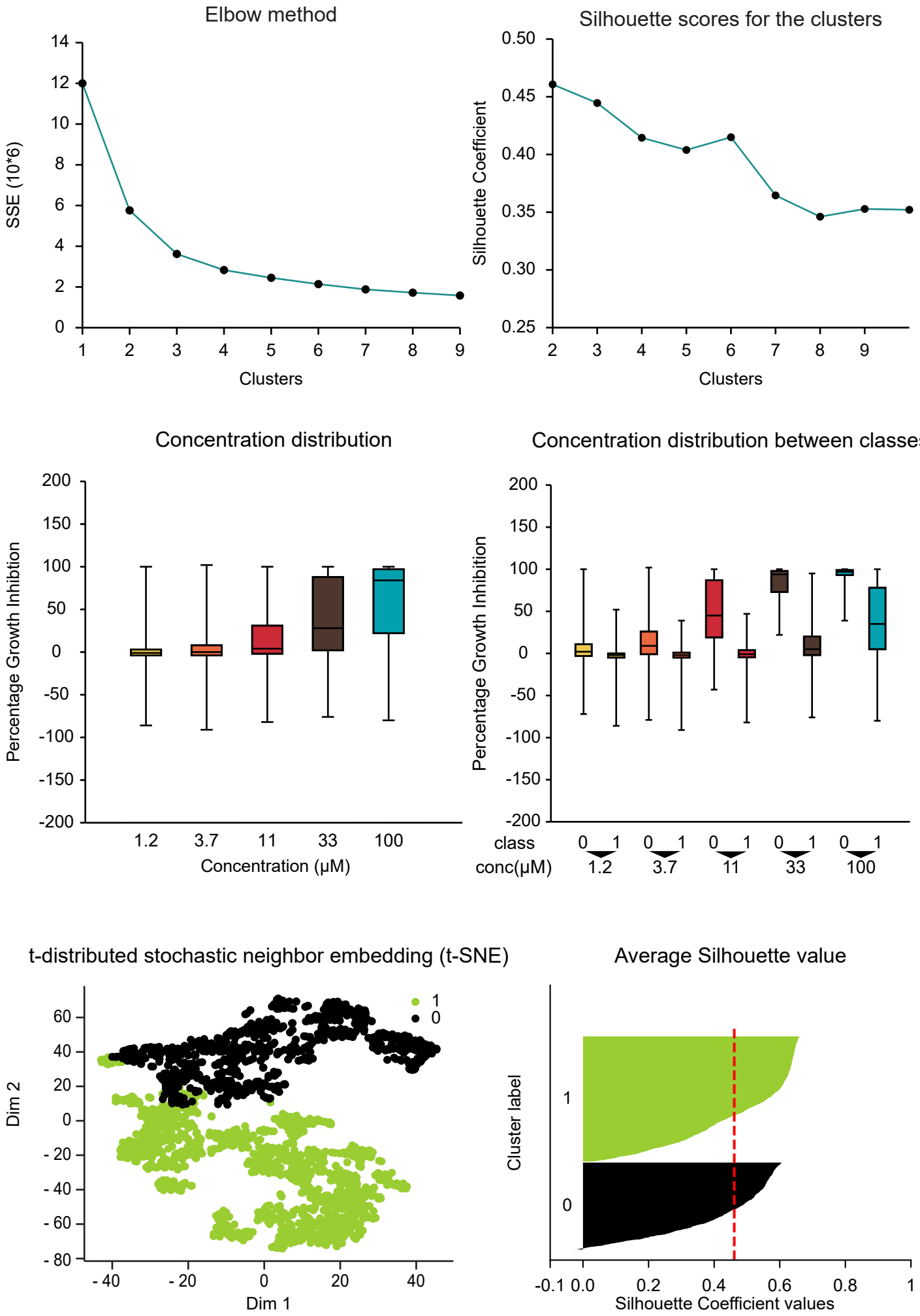
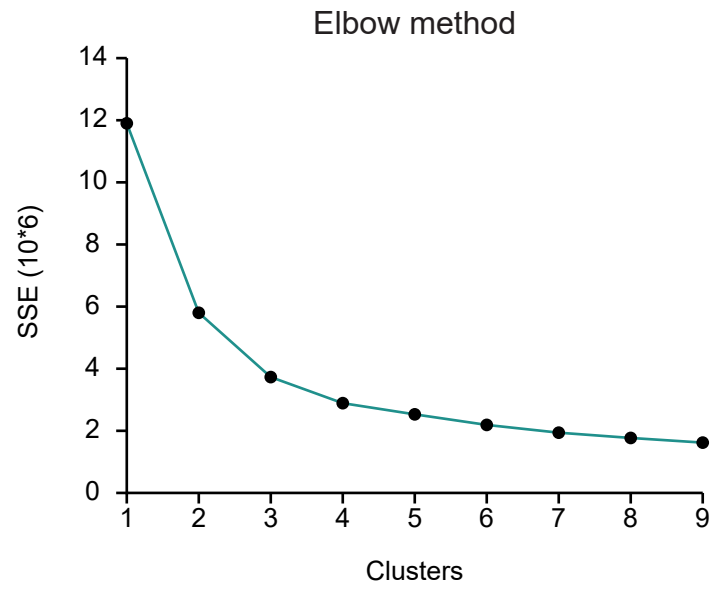
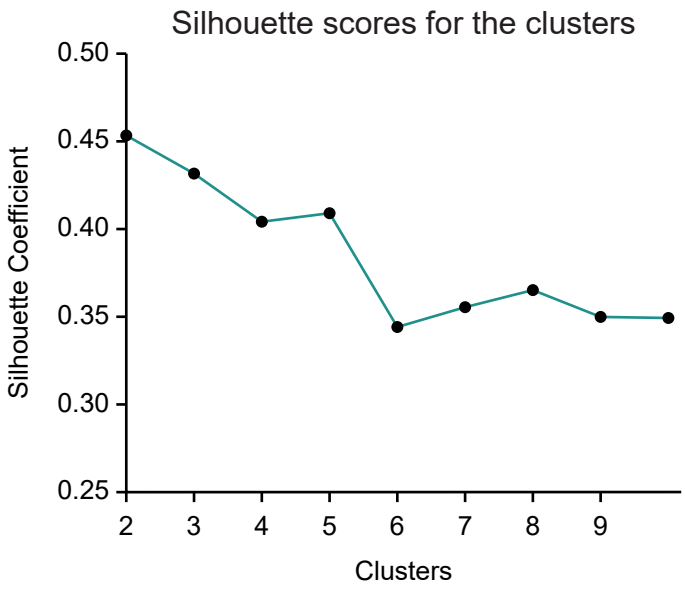
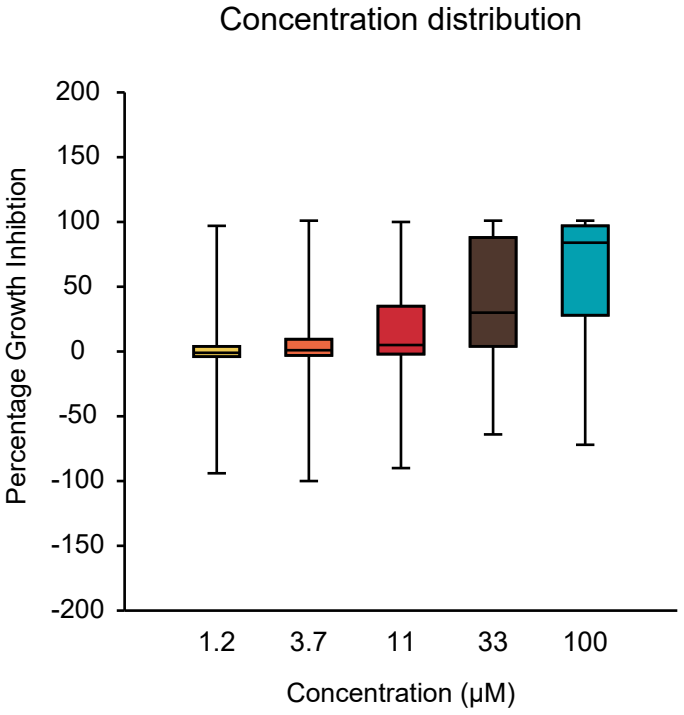
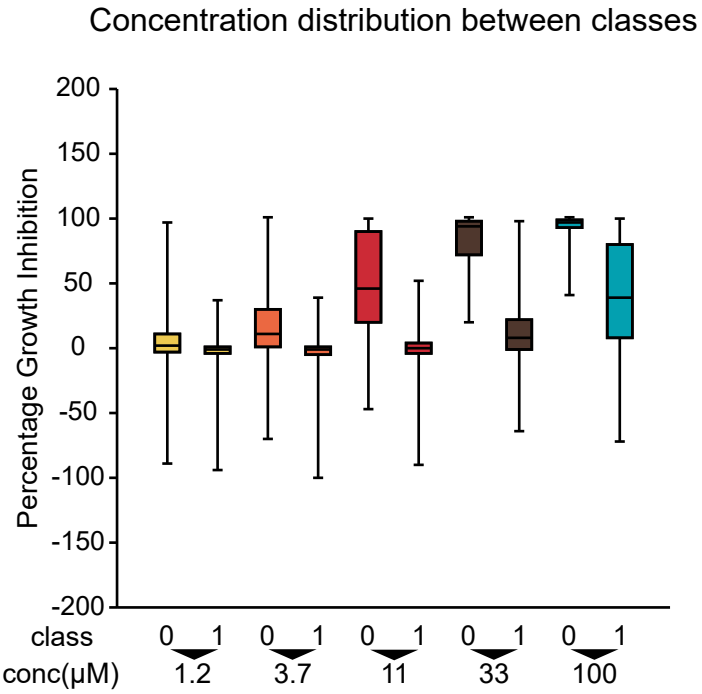


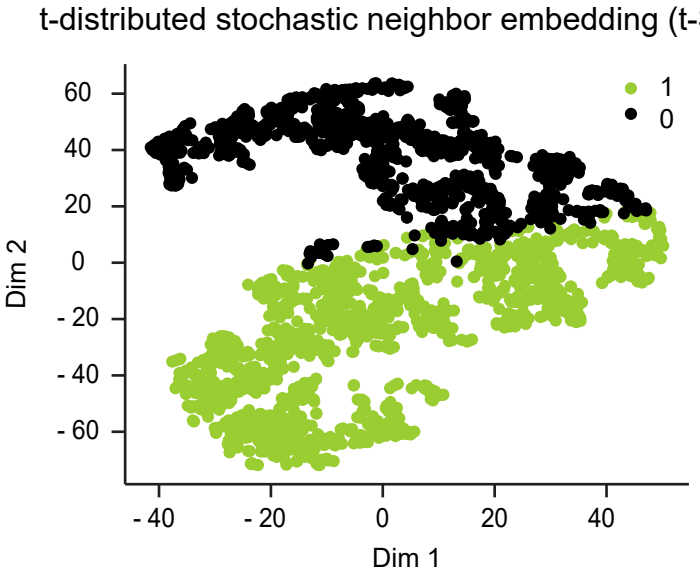
Figure 20.5: Unsupervised Classification based on GIPCRT - *mlh1* strain











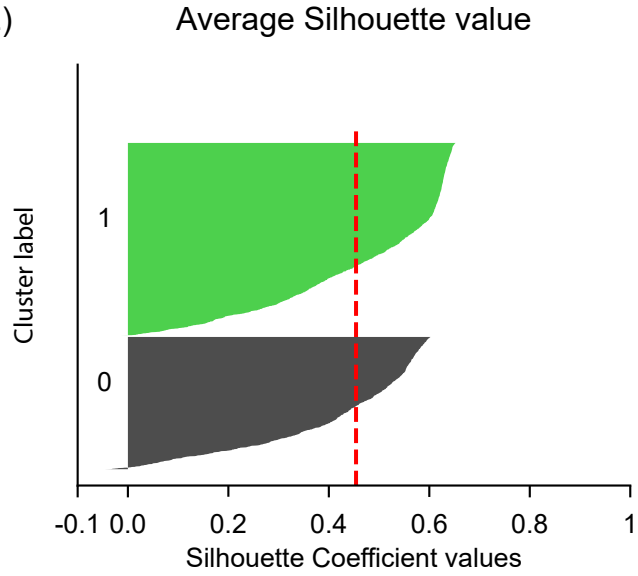
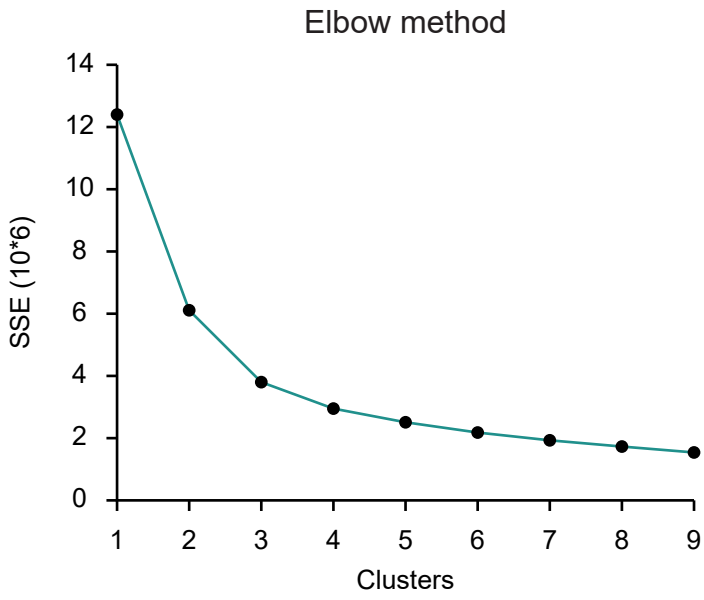
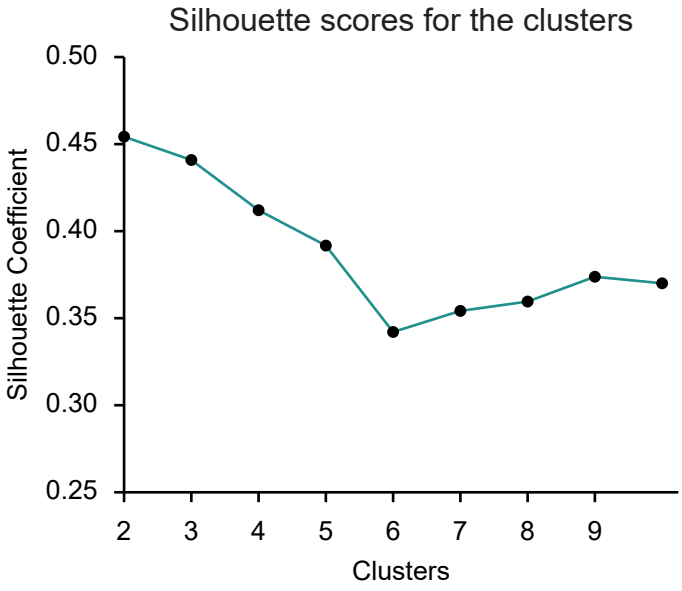
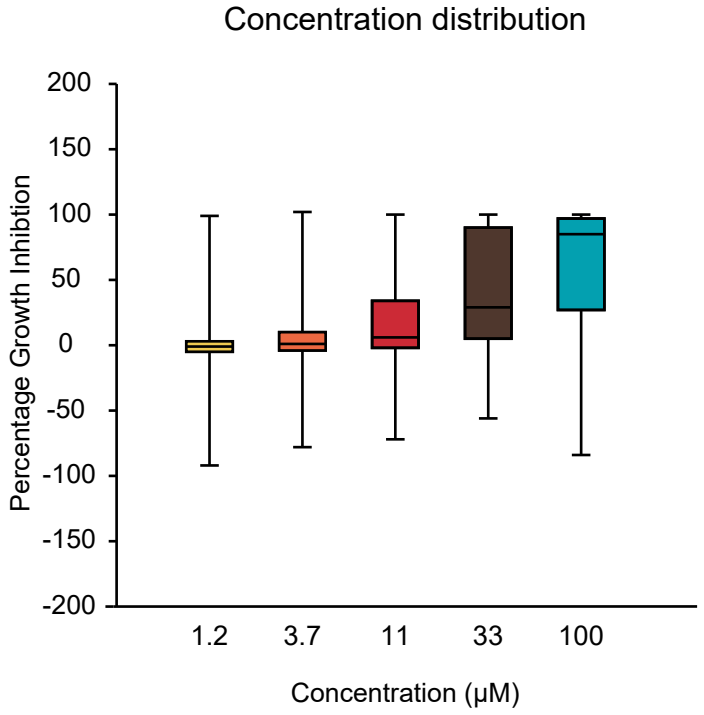
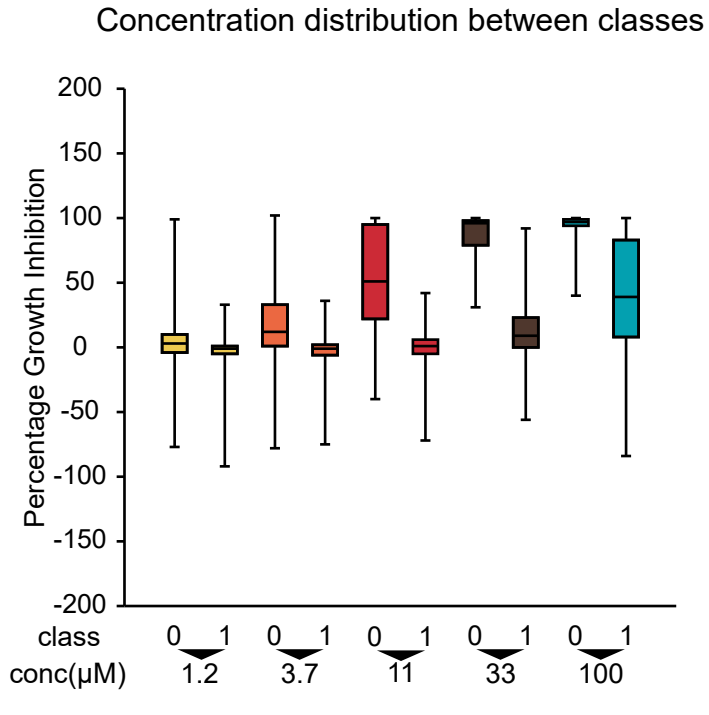


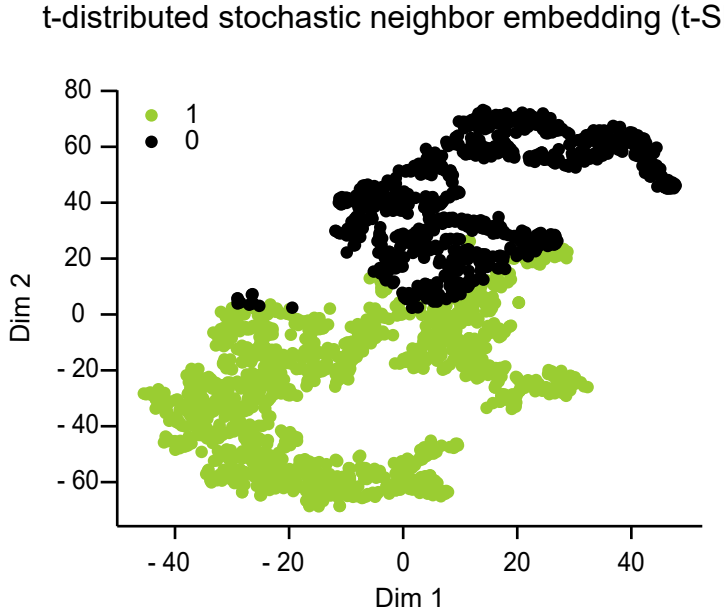
Figure 20.6:Unsupervised Classification based on GIPCRT - *rad14* strain











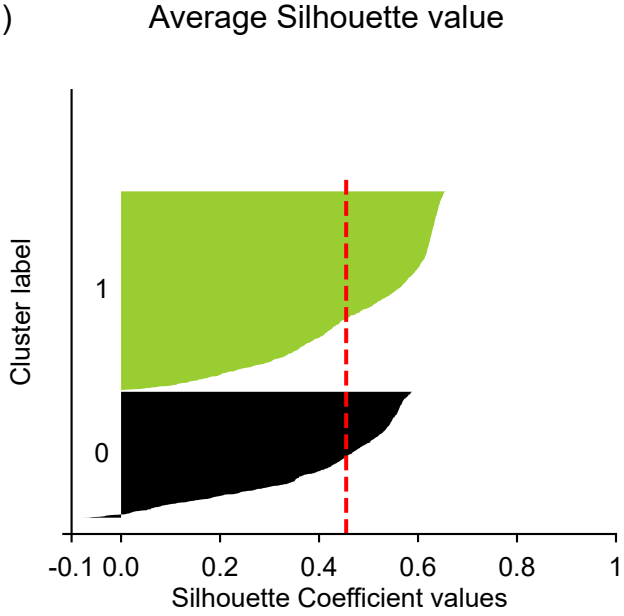
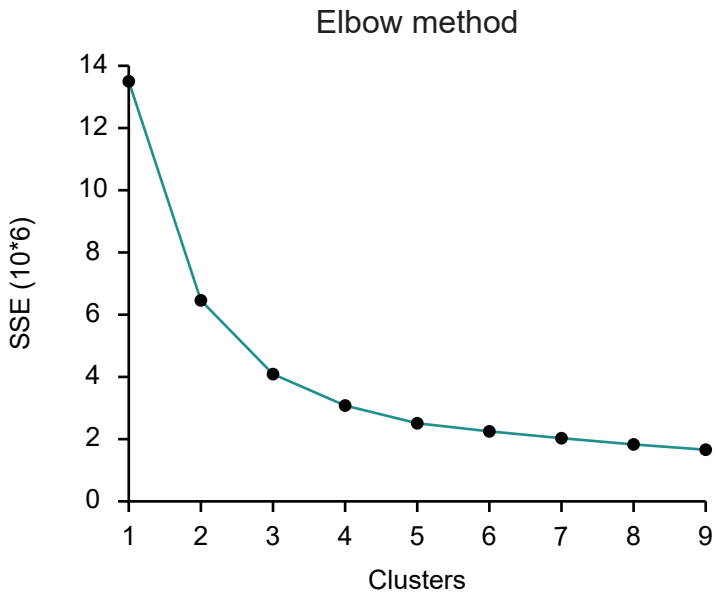
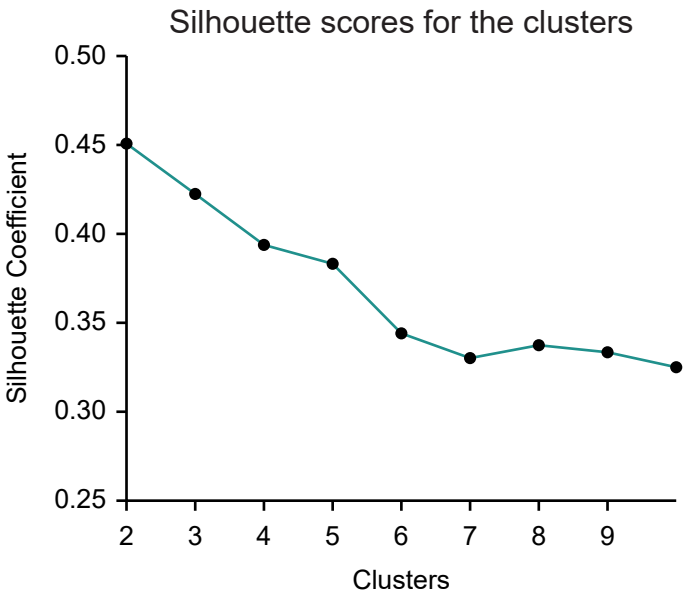
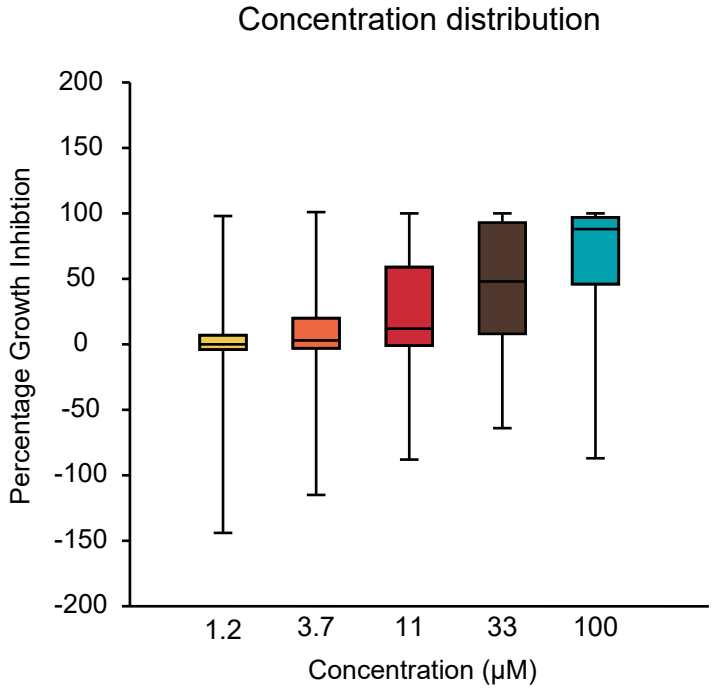
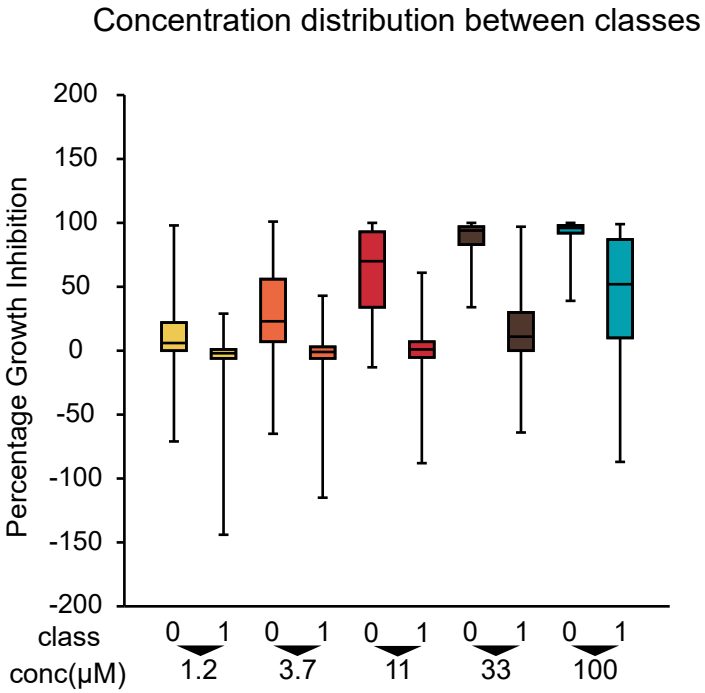


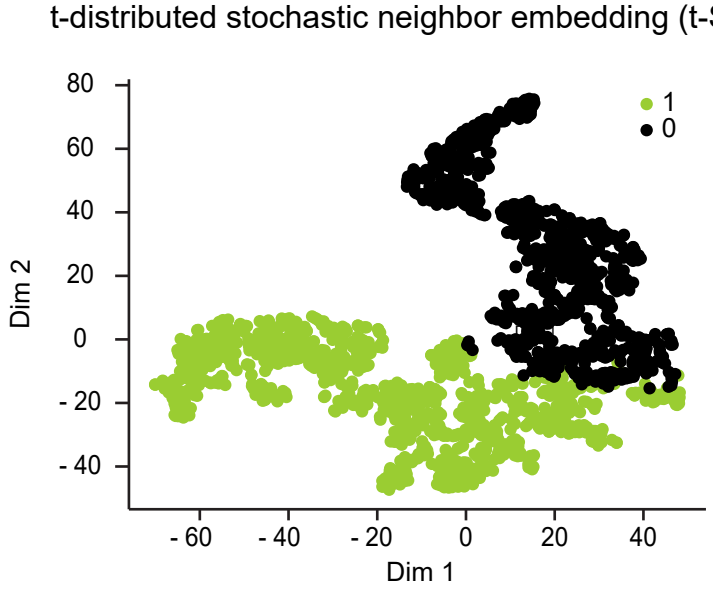
Figure 20.7: Unsupervised Classification based on GIPCRT - *rad18* strain











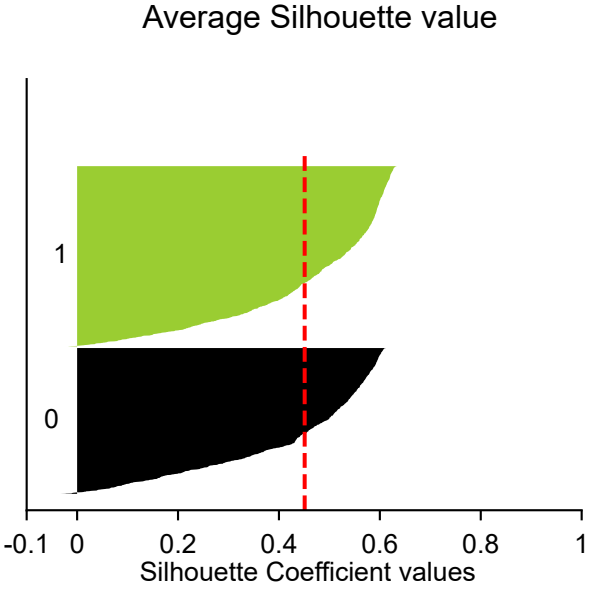
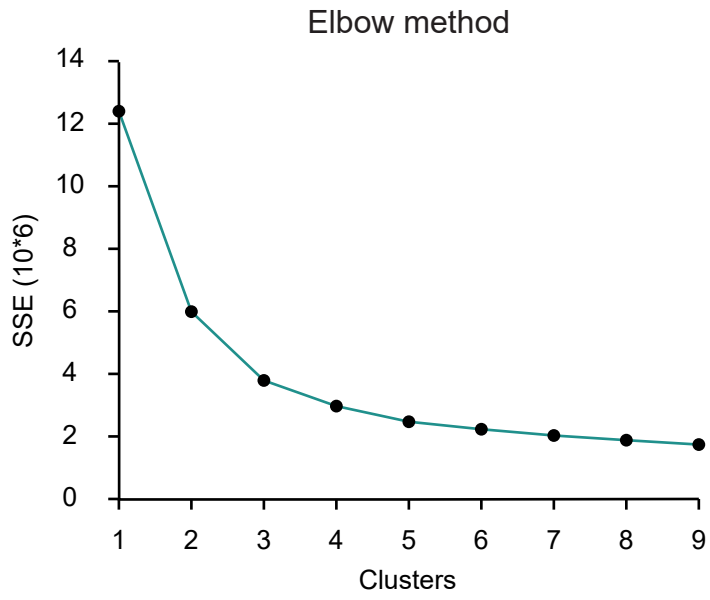
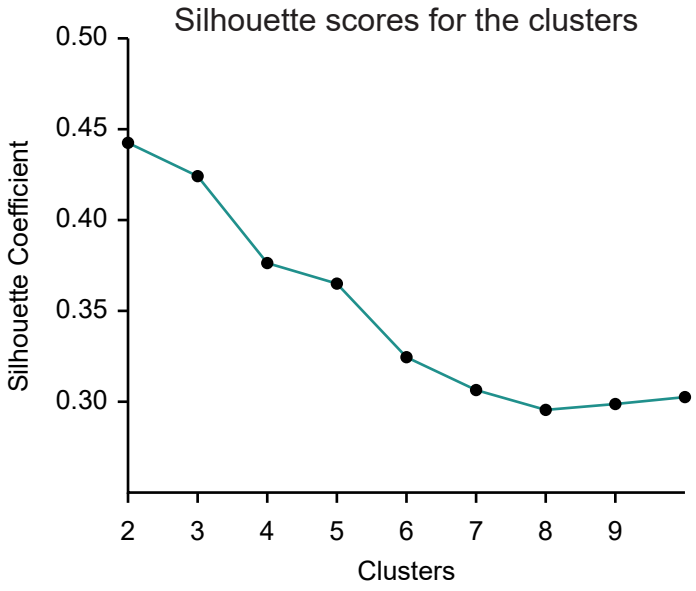
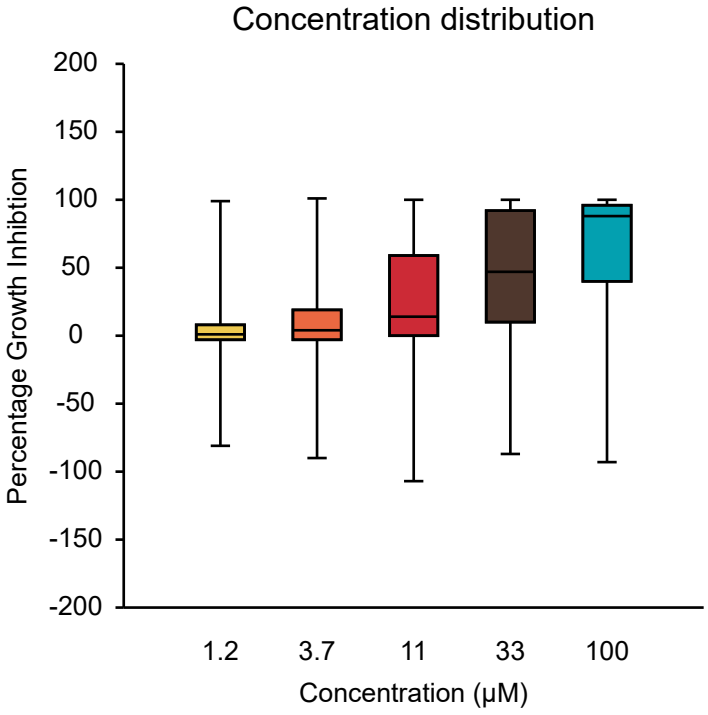
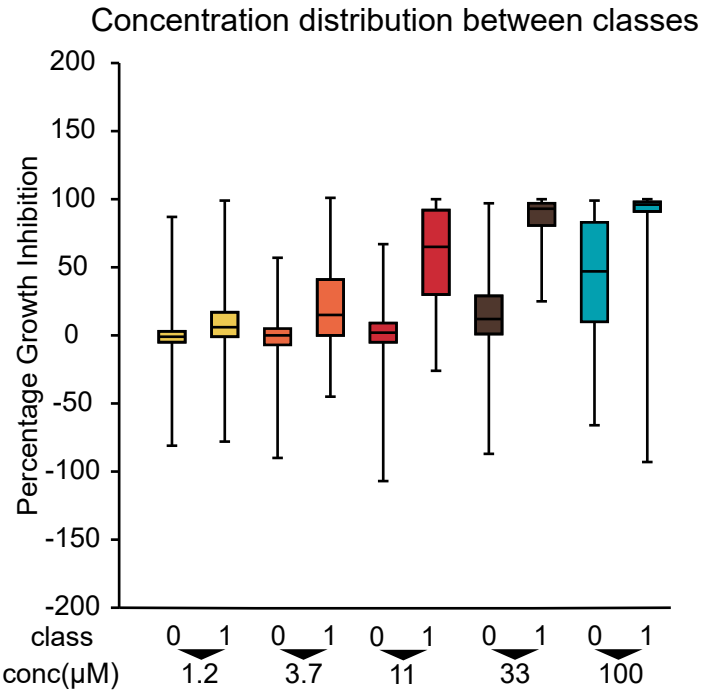


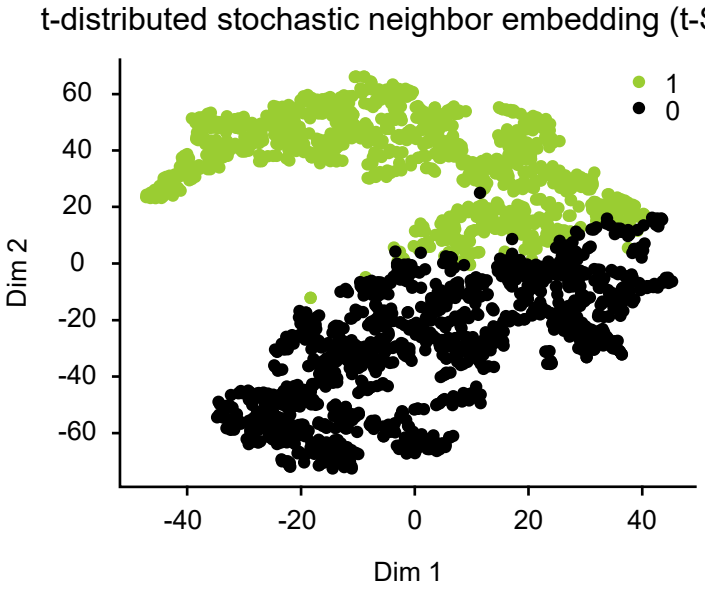
Figure 20.8: Unsupervised Classification based on GIPCRT - rad52 strain











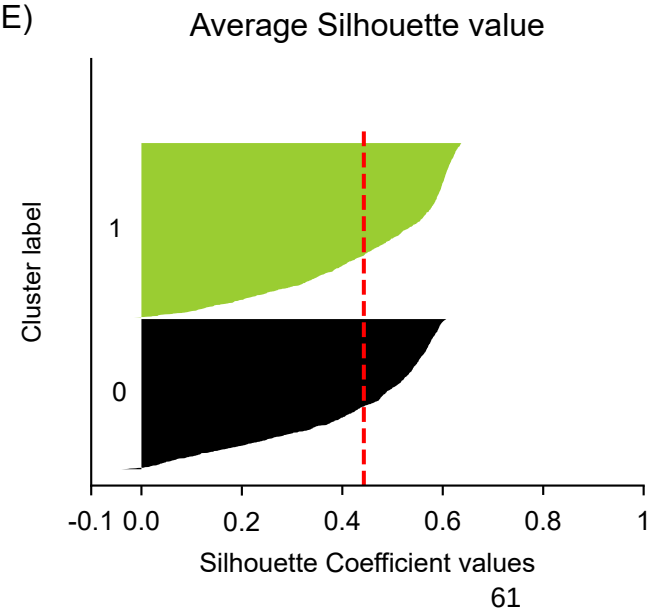
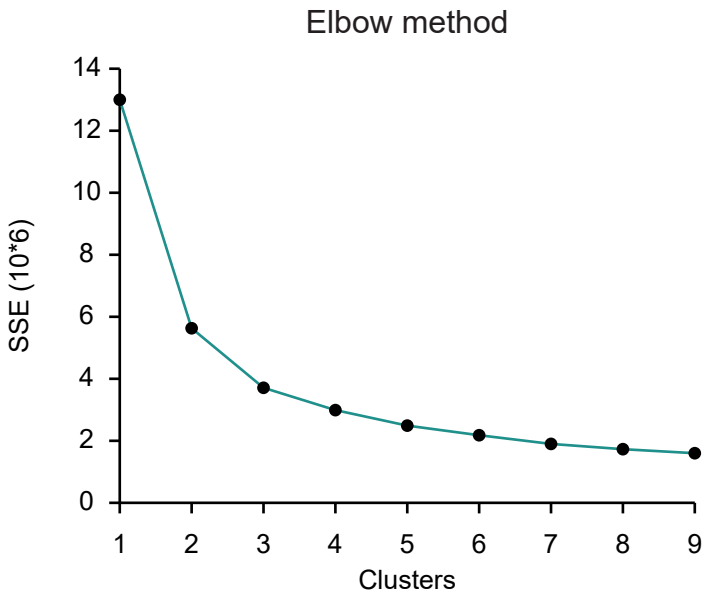
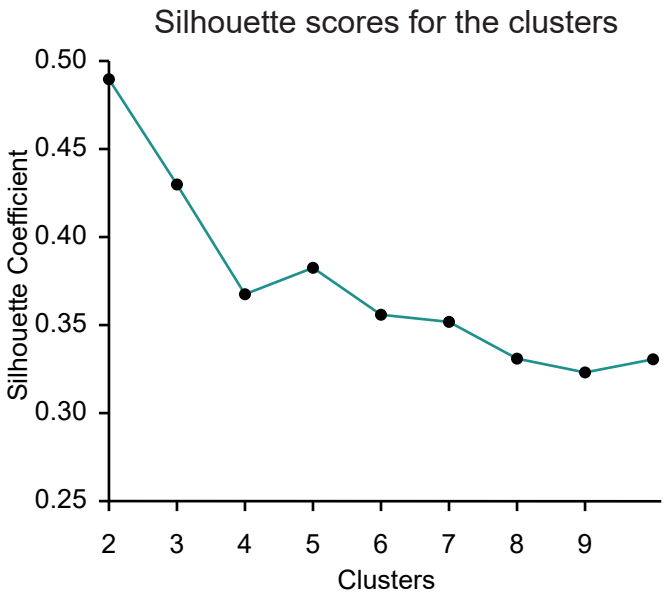
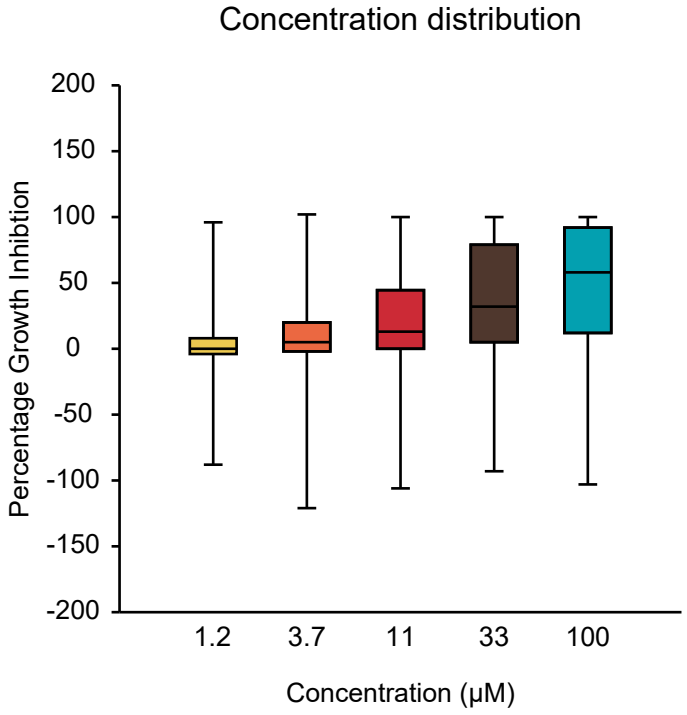
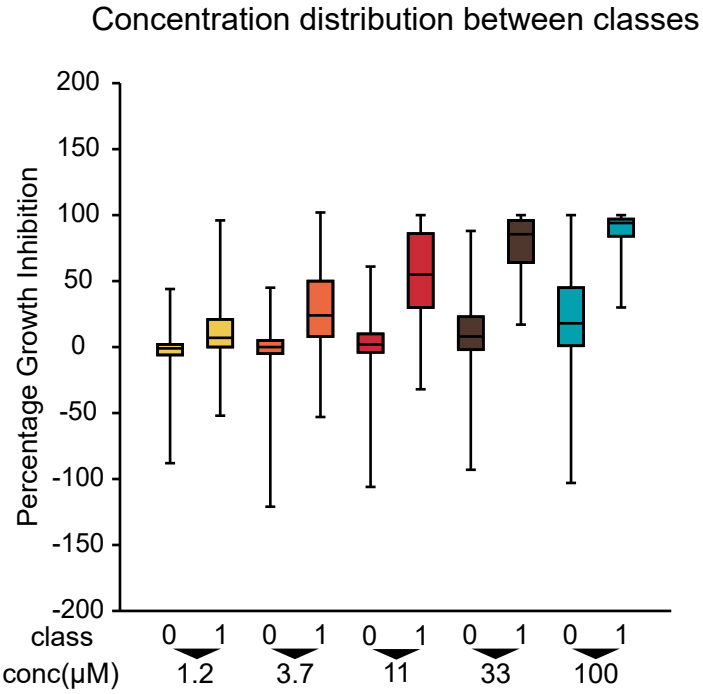


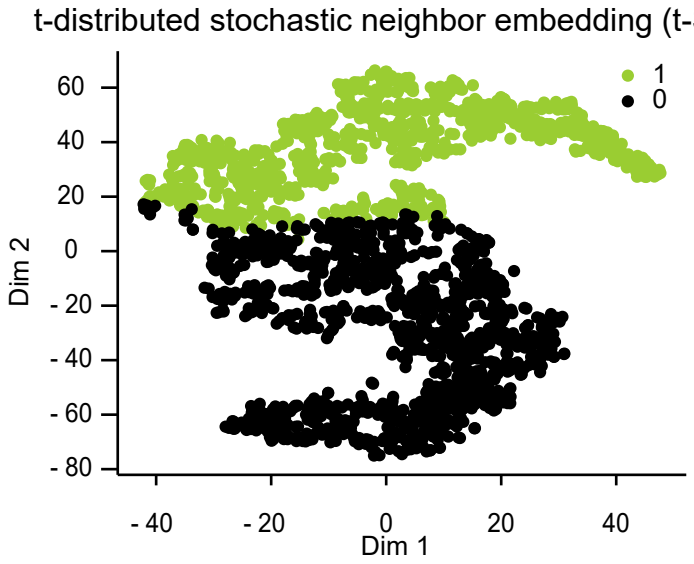
Figure 20.9: Unsupervised Classification based on GIPCRT - *rad50EPP+* strain











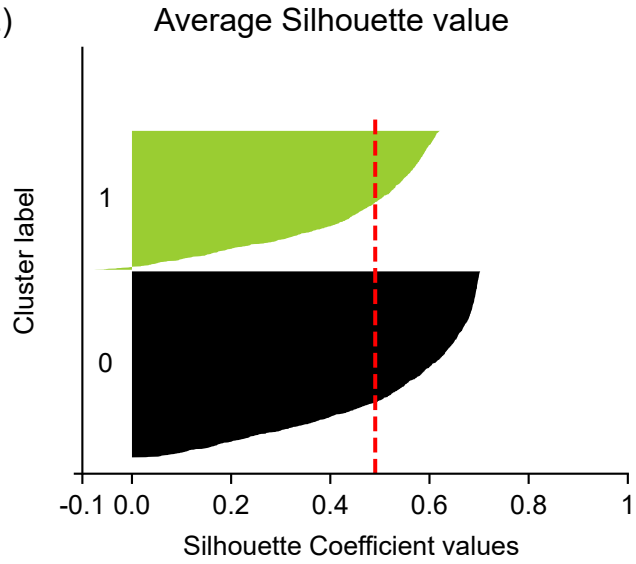
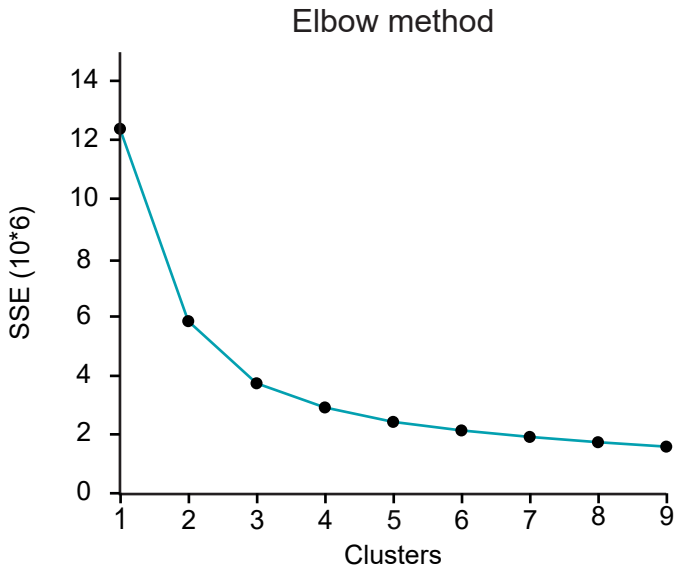
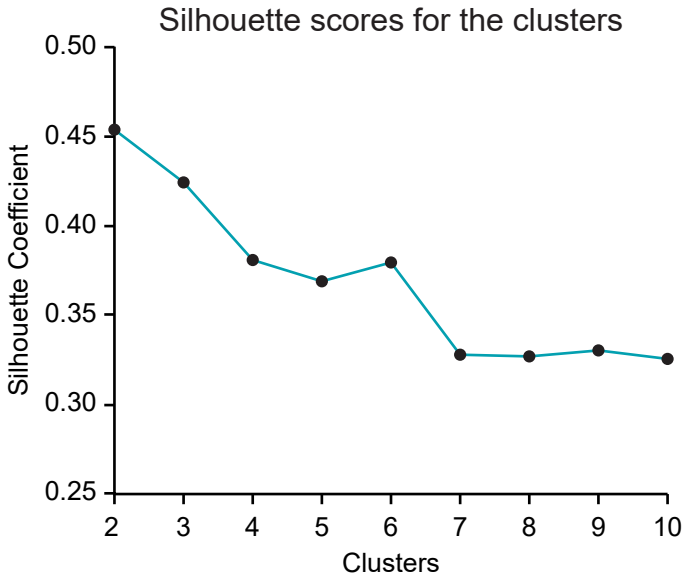
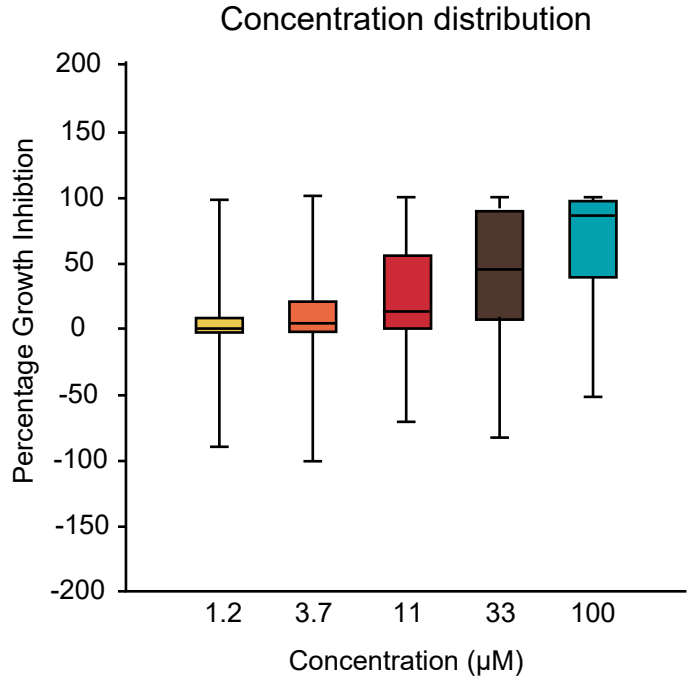
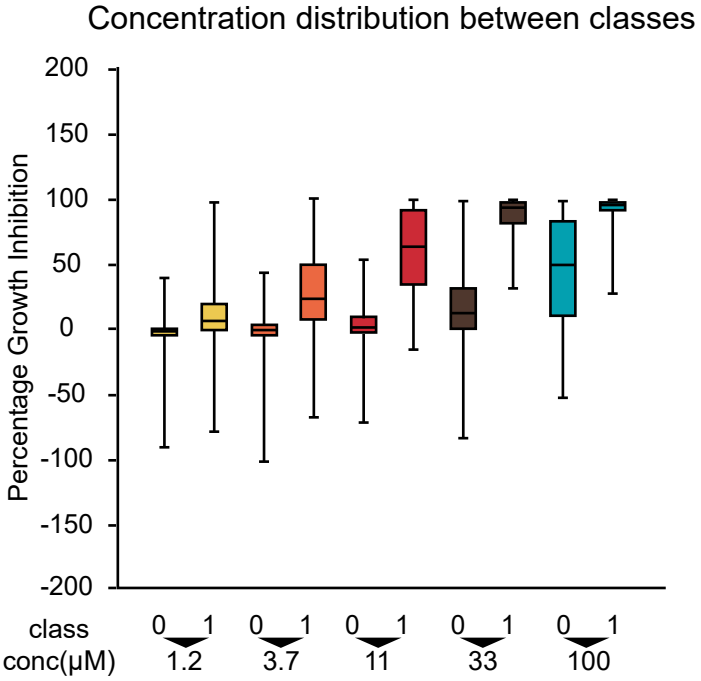


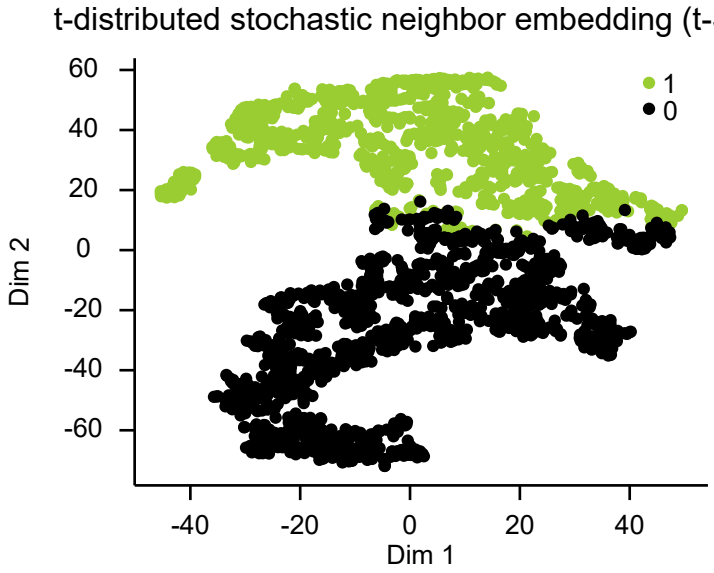
Figure 20.10: Unsupervised Classification based on GIPCRT - rad52 strain











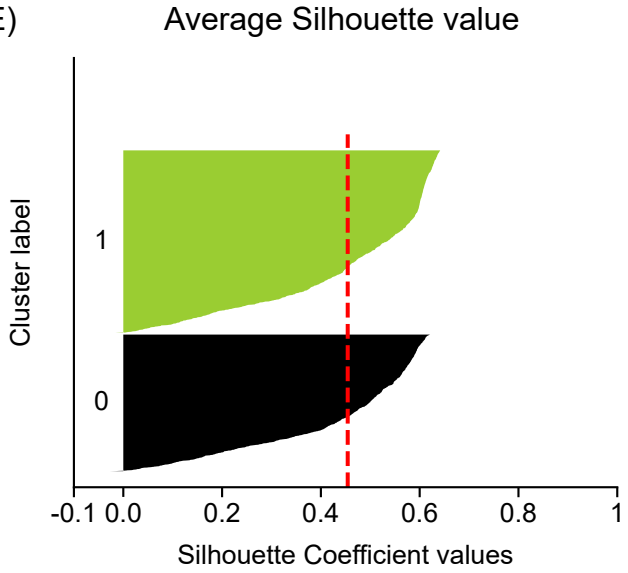
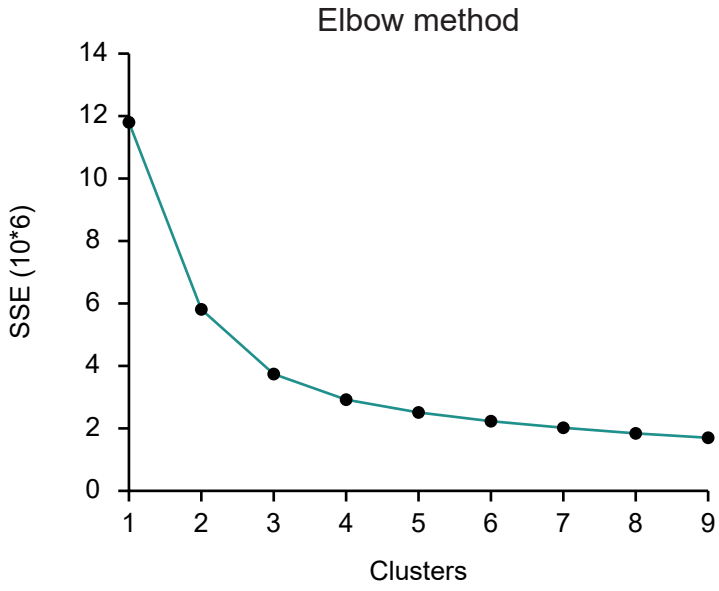
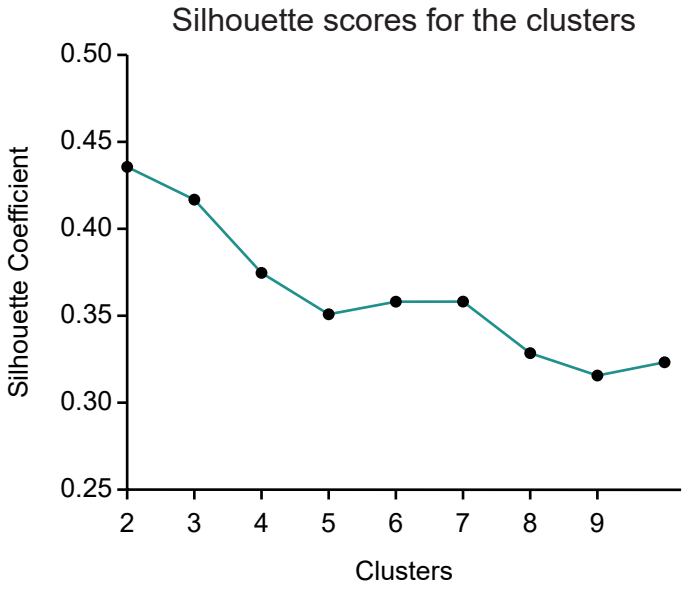
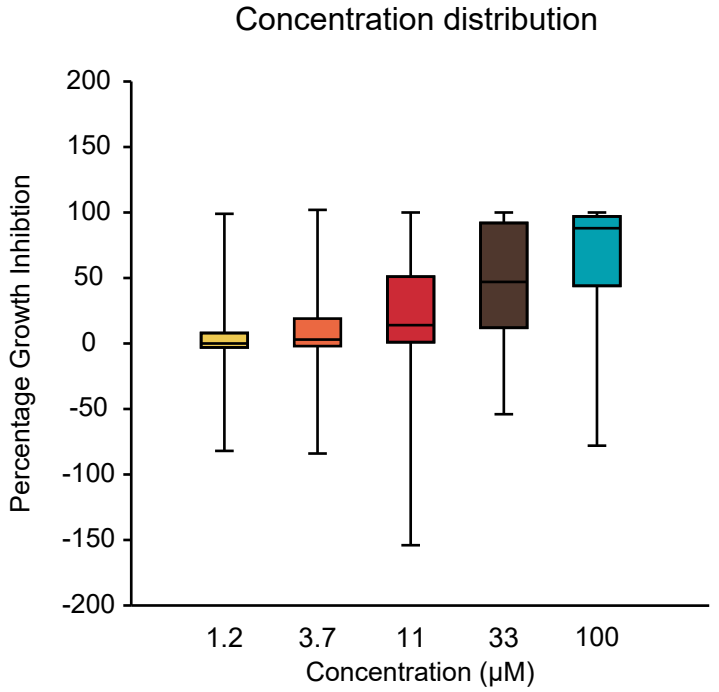
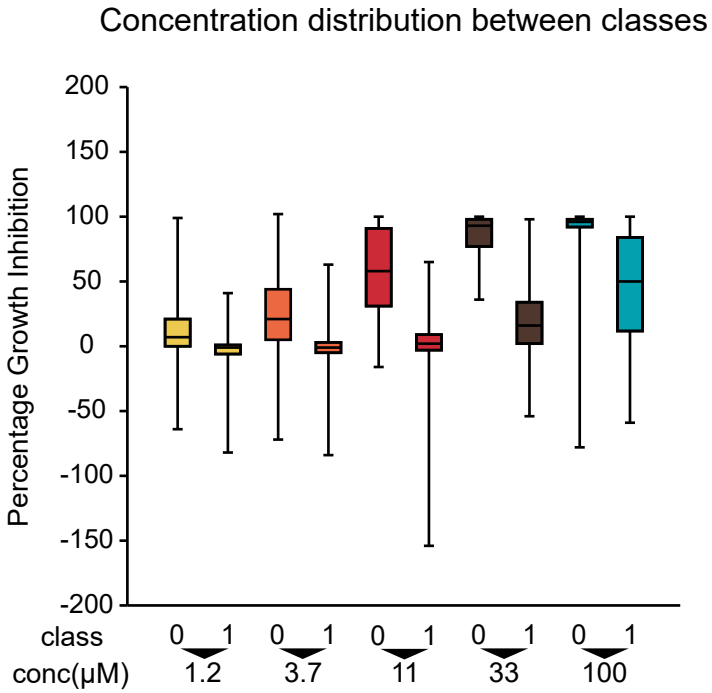


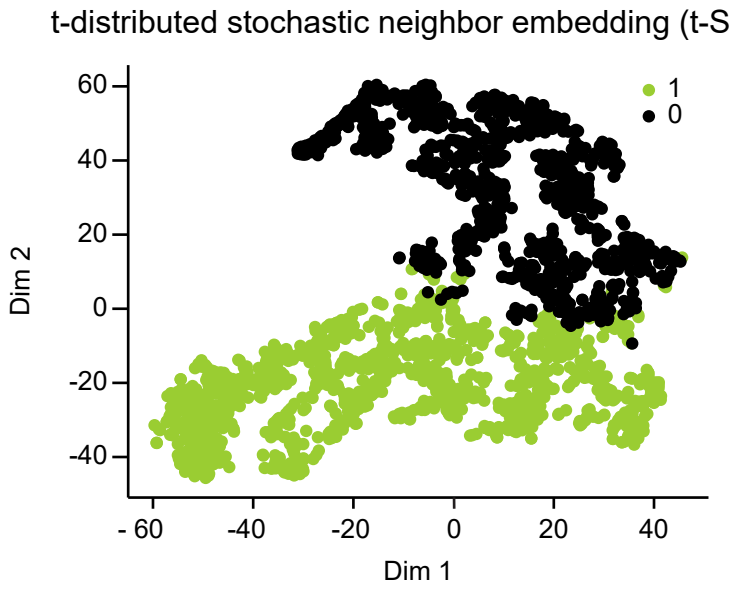
Figure 20.11: Unsupervised Classification based on GIPCRT - *sgs1* strain











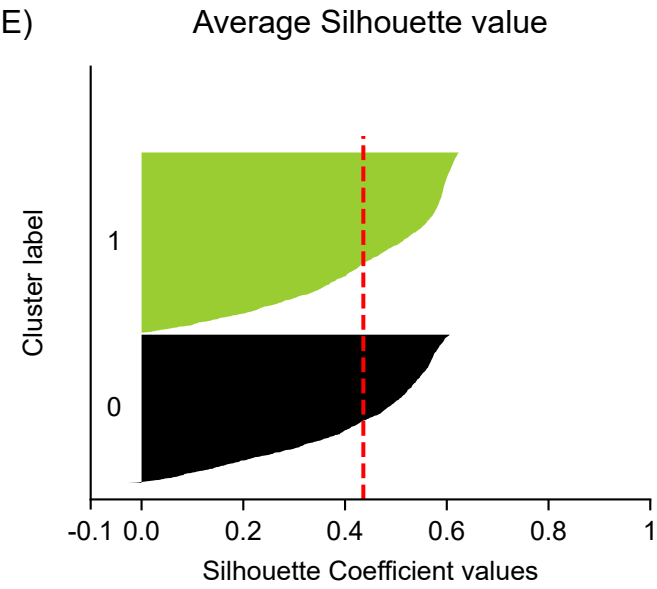
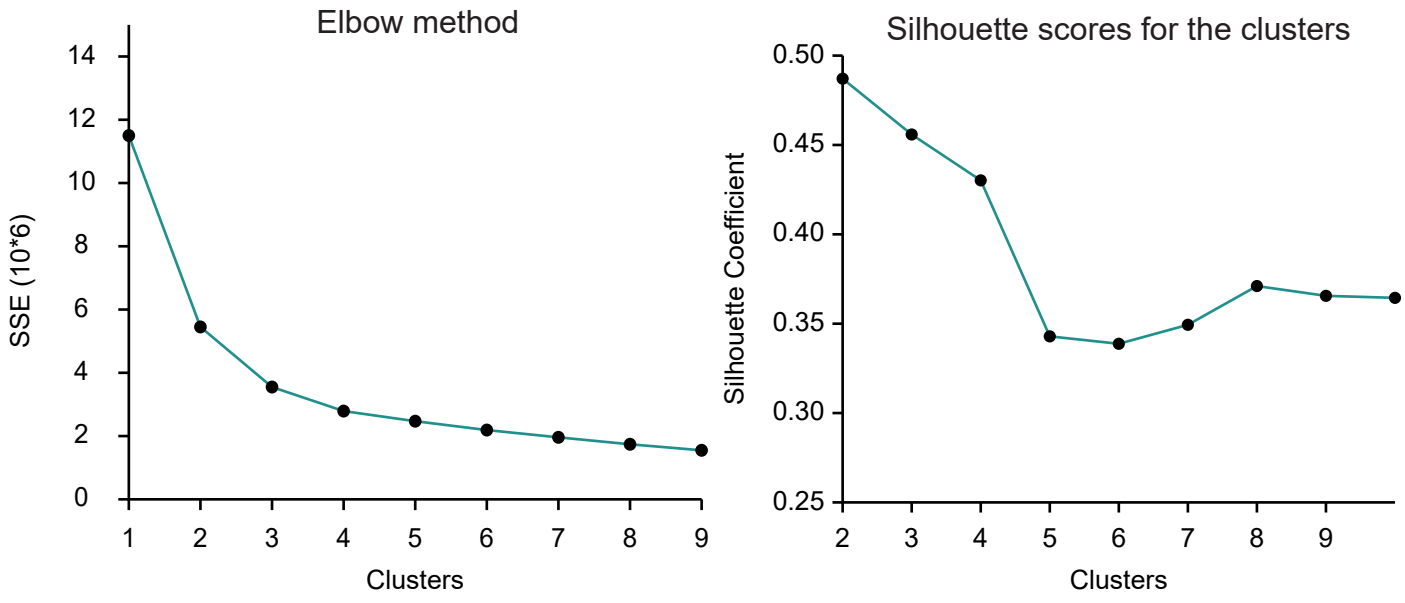
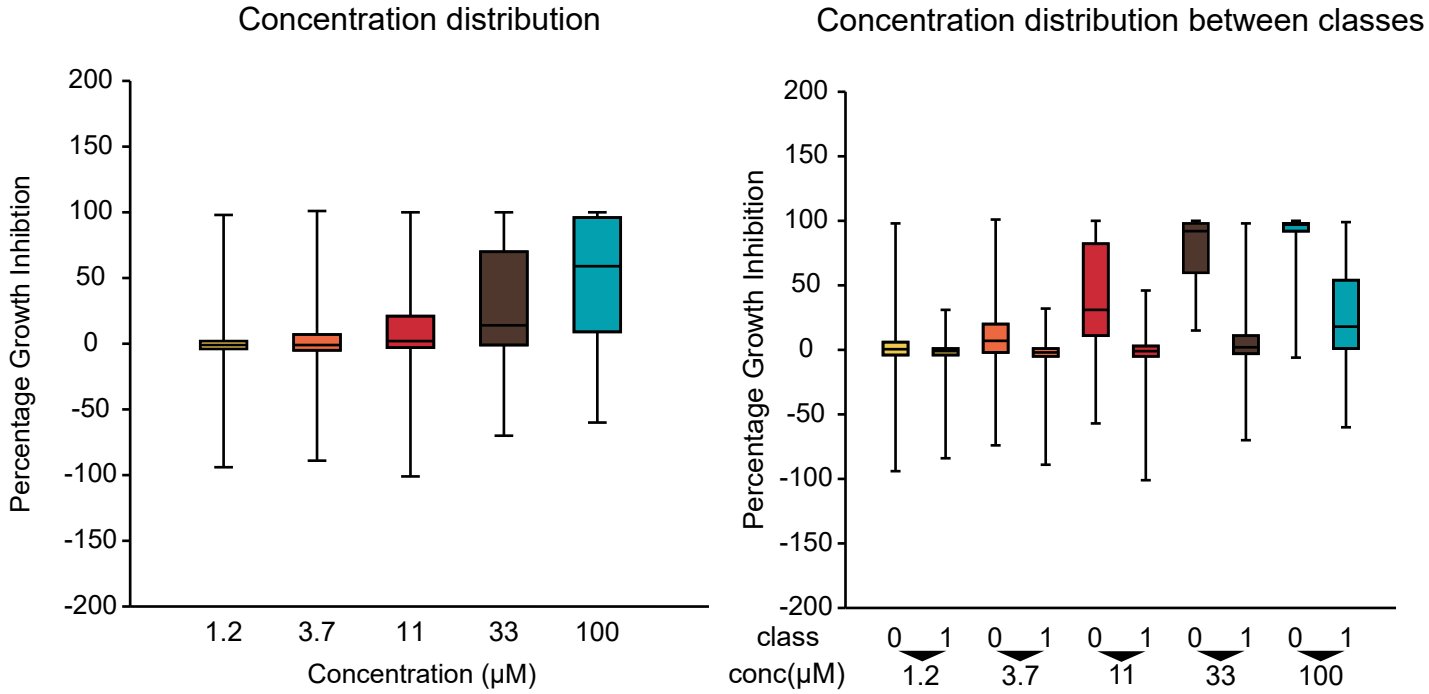


Figure 20.12: Unsupervised Classification based on GIPCRT - *wt1* strain





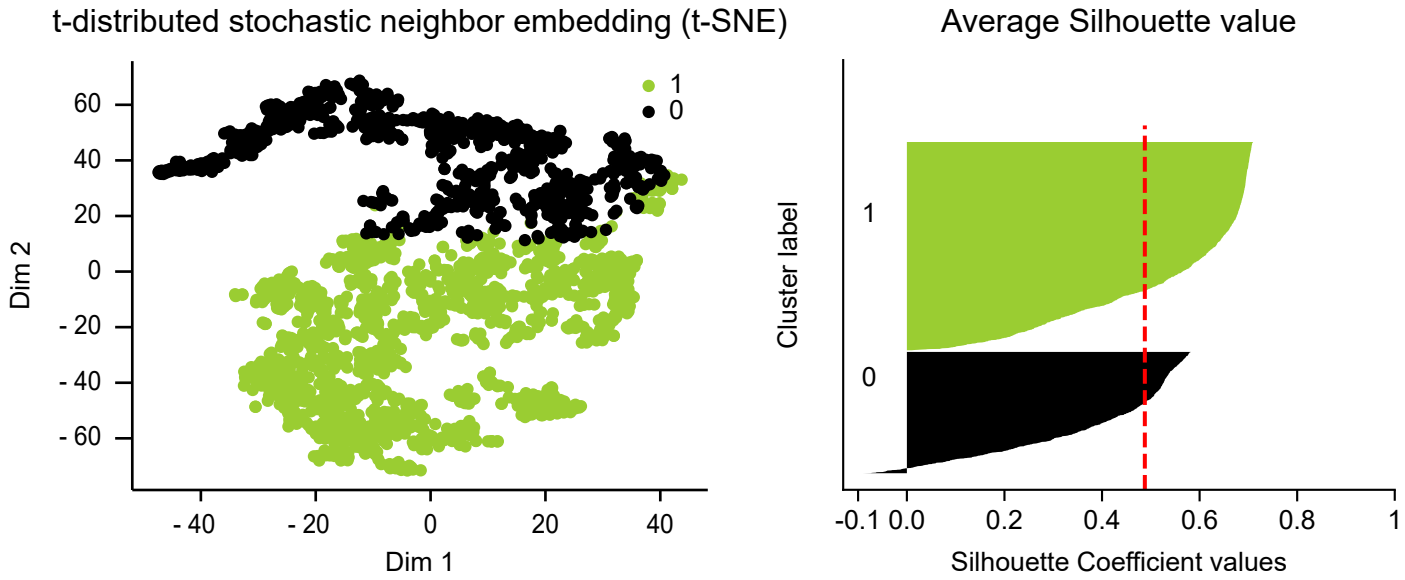
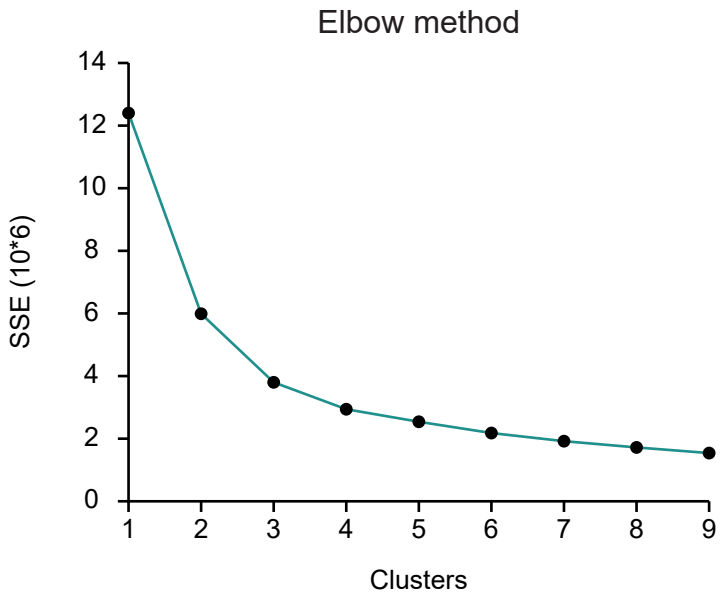
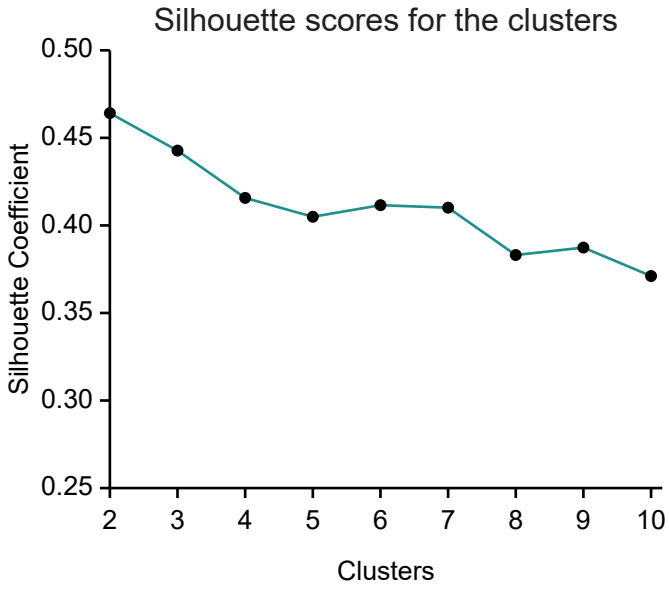
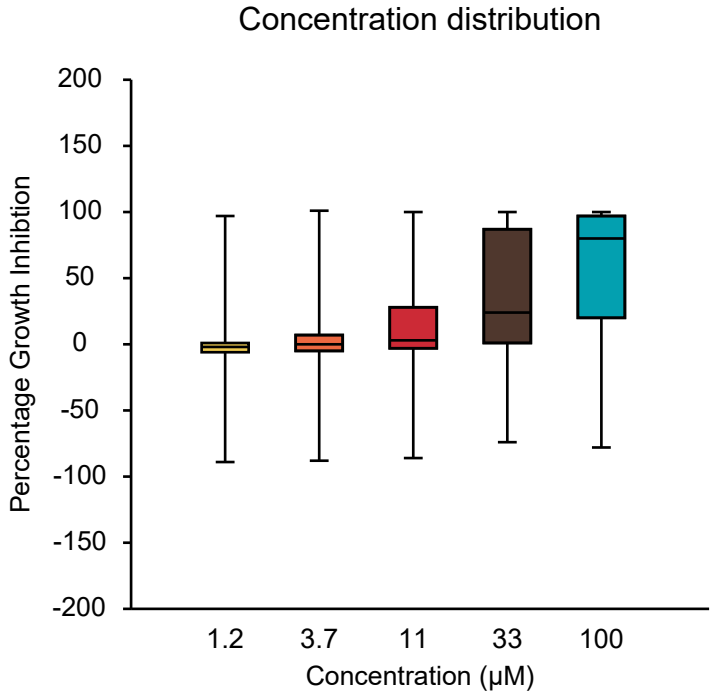
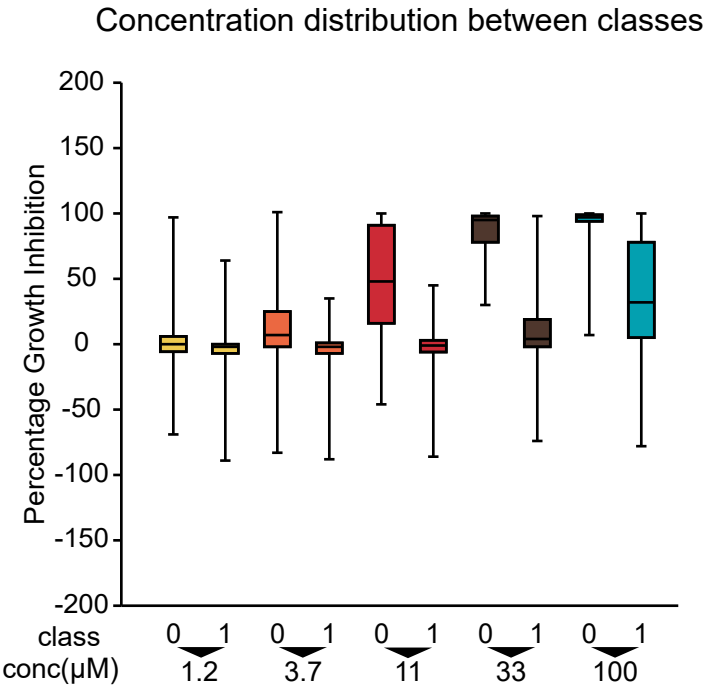


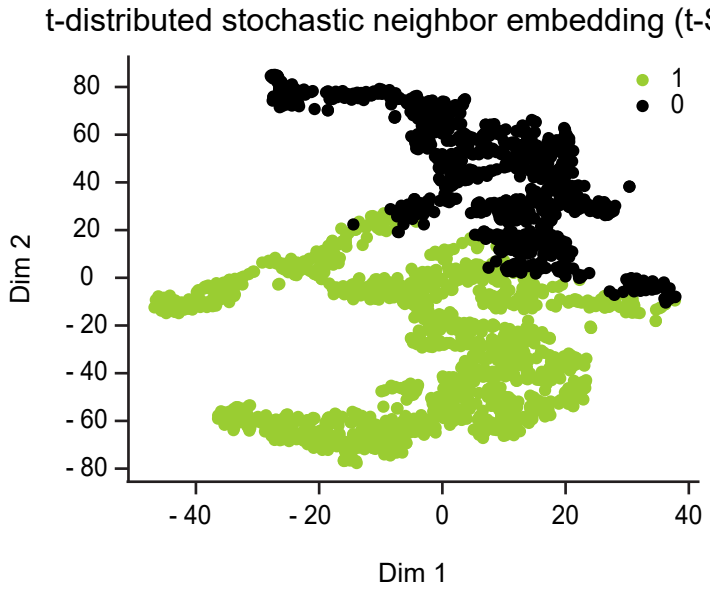
Figure 20.13:Unsupervised Classification based on GIPCRT - *wt2* strain











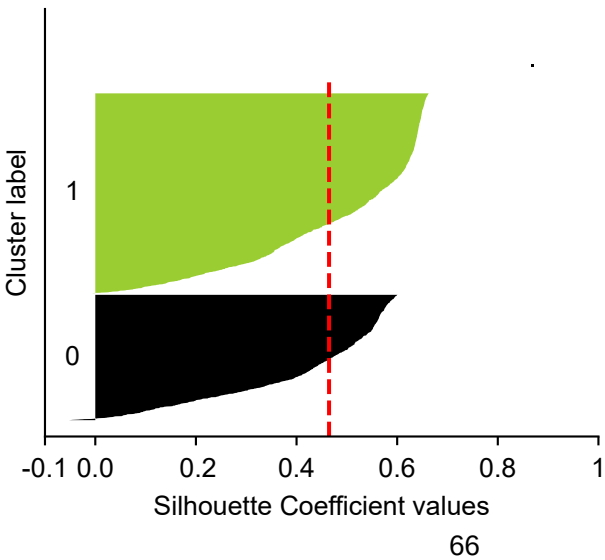


Figure 21.1:

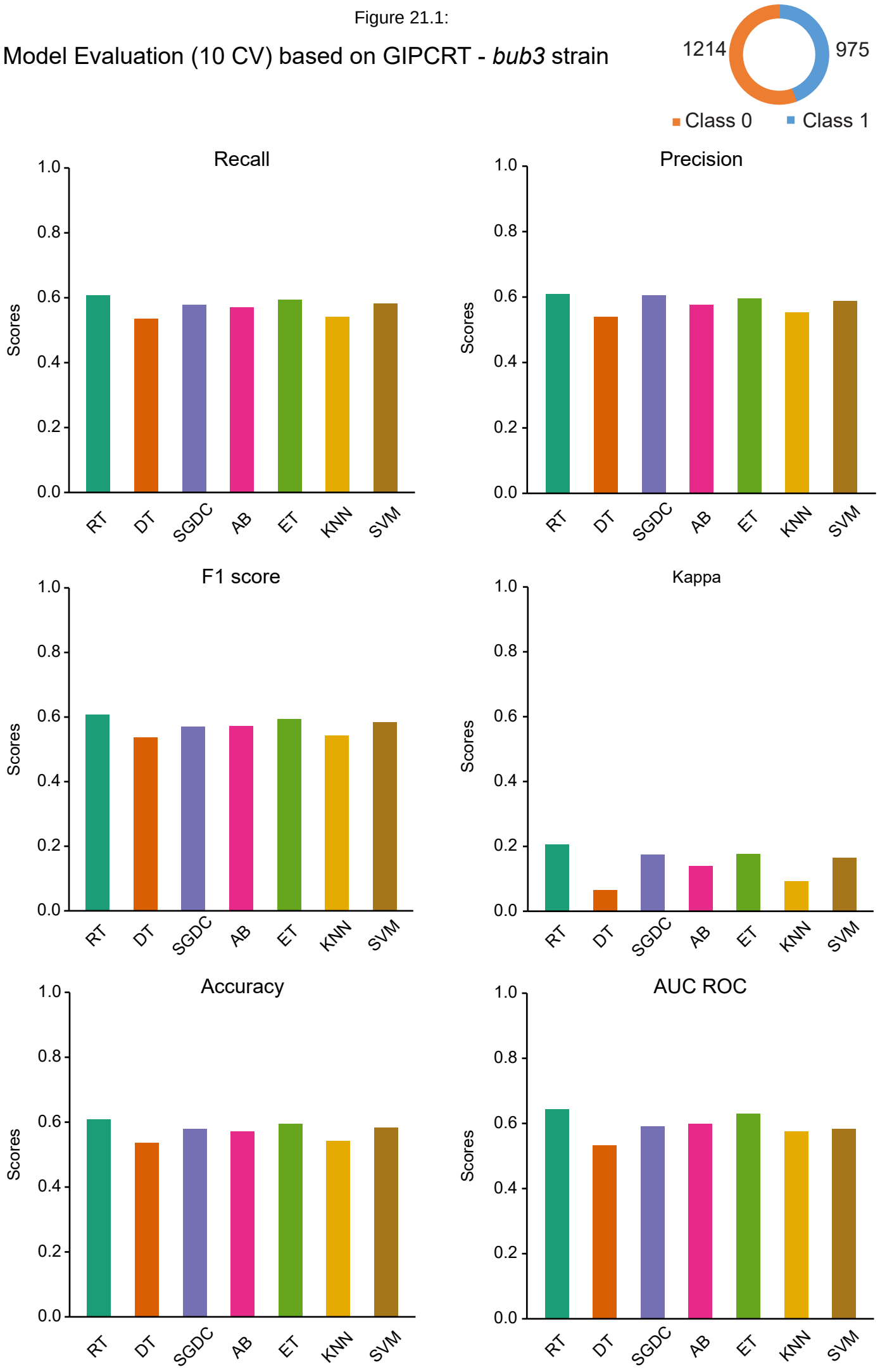
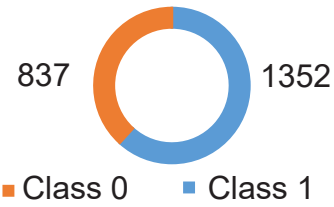
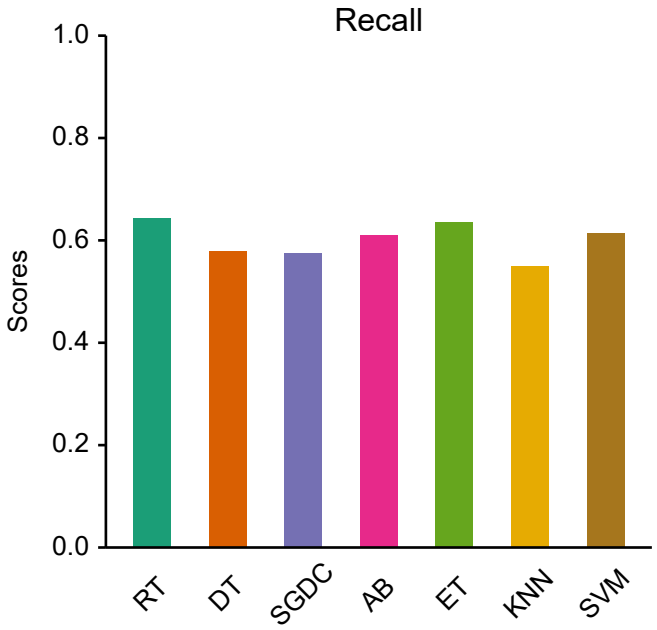
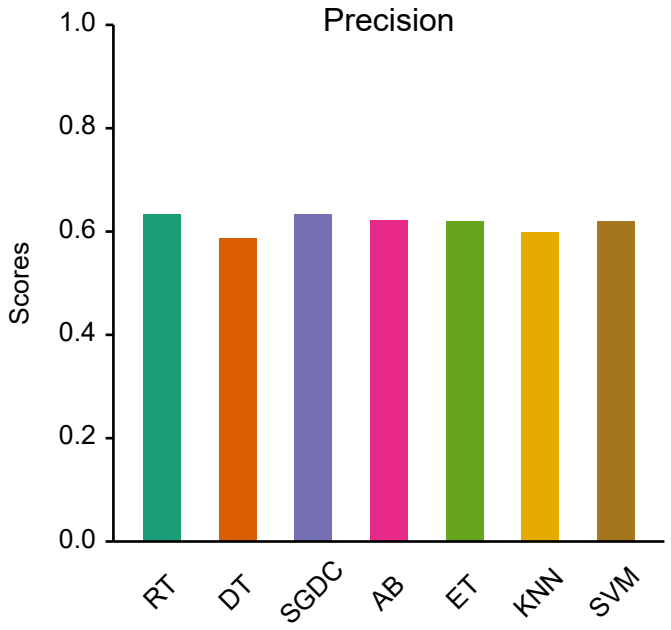

Model Evaluation (10 CV) based on GIPCRT - bub3 strain
1214
975
Class 0
Class 1
Recall
Precision
F1 score
Kappa
Accuracy
AUC ROC
Scores
1.0
0.8
0.6
0.4
0.2
0.0
RT
DT
SGDC
AB
ET
KNN
SVM

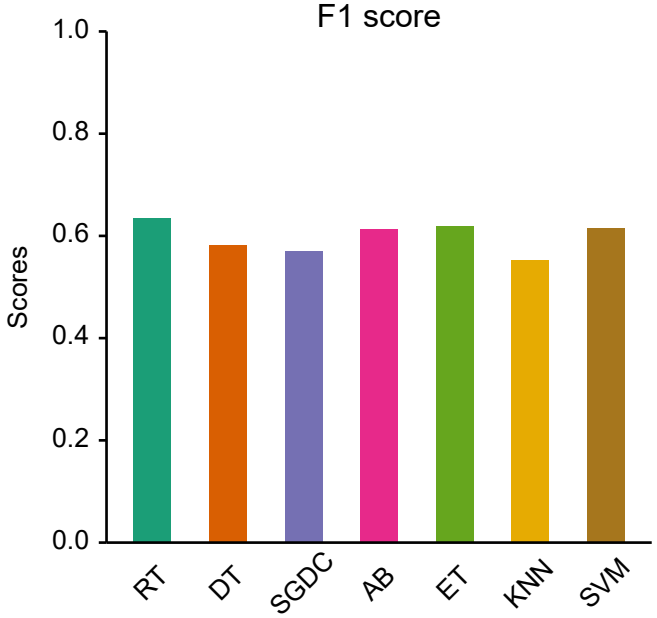
Figure 21.2:

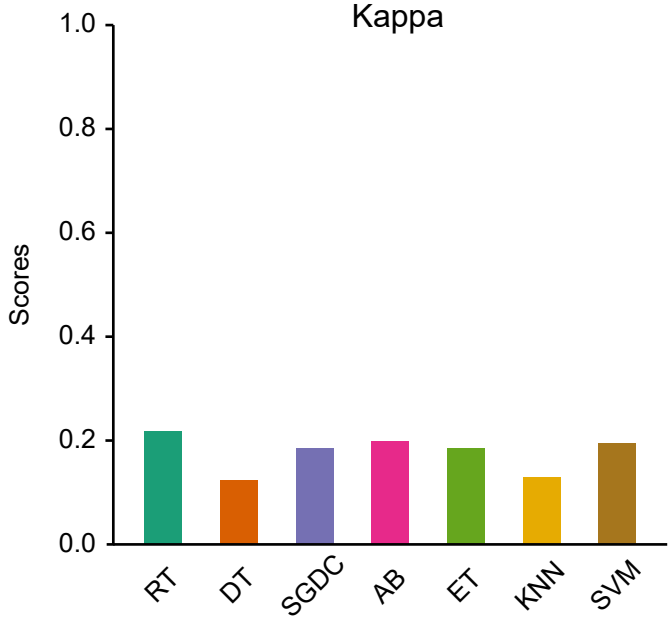
Model Evaluation (10 CV) based on GIPCRT - *CLN2oe* strain

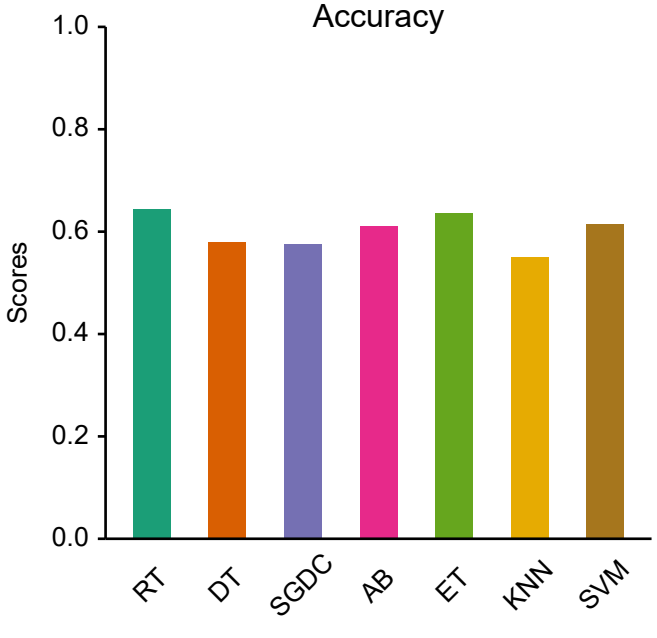












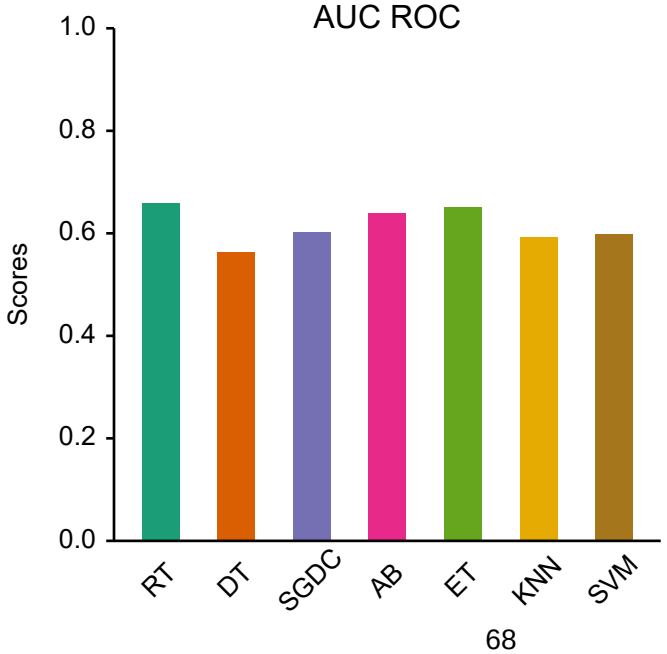
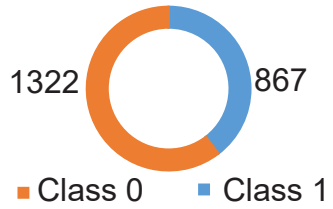
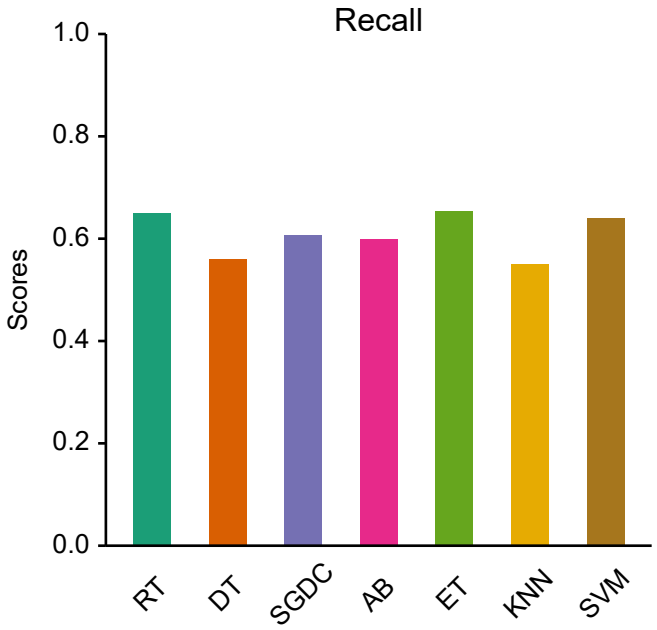
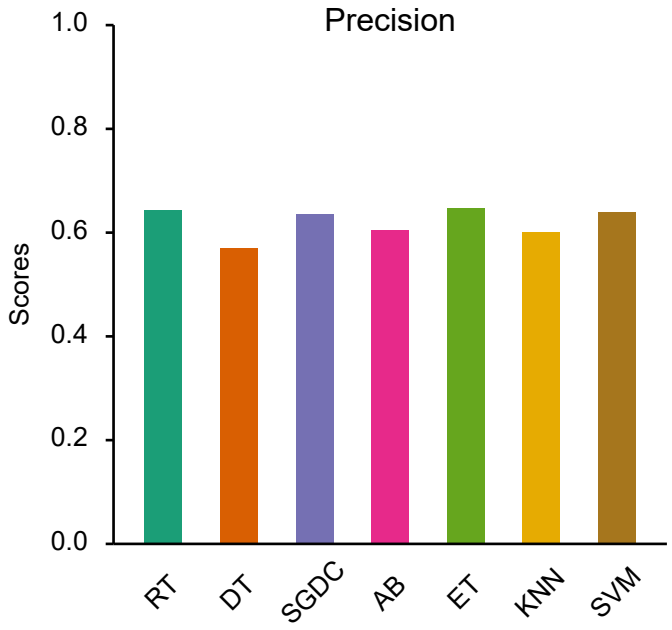


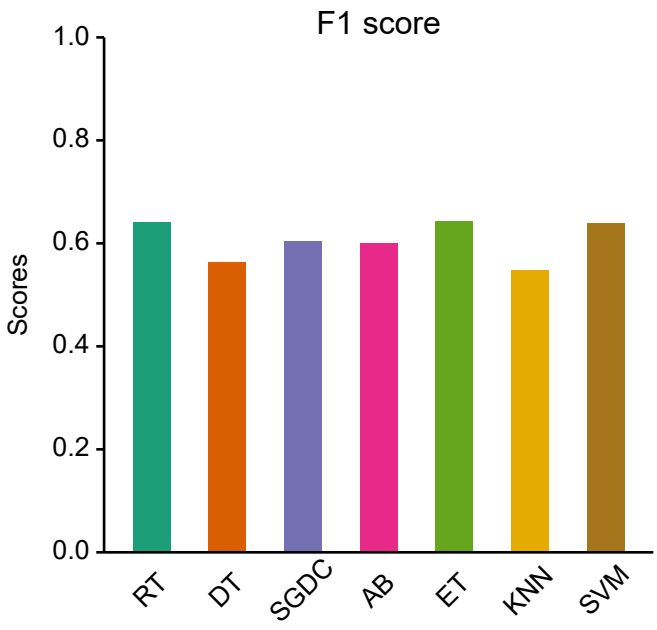
Figure 21.3:

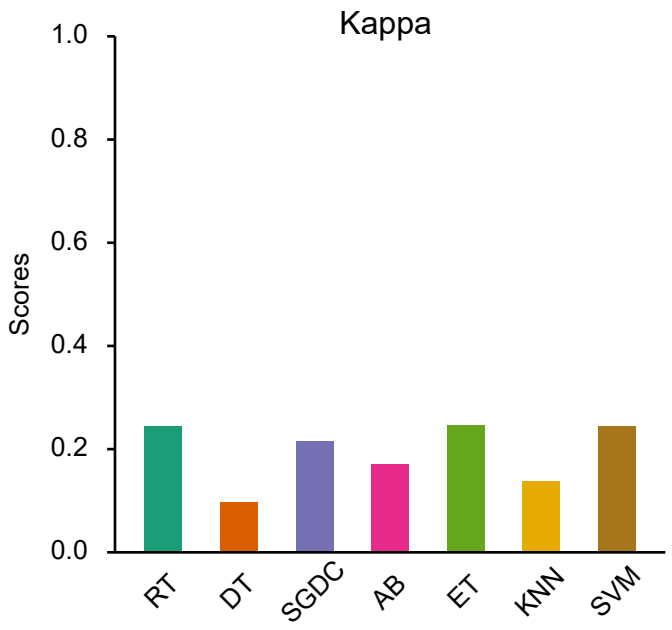
Model Evaluation (10 CV) based on GIPCRT - *mec2* strain

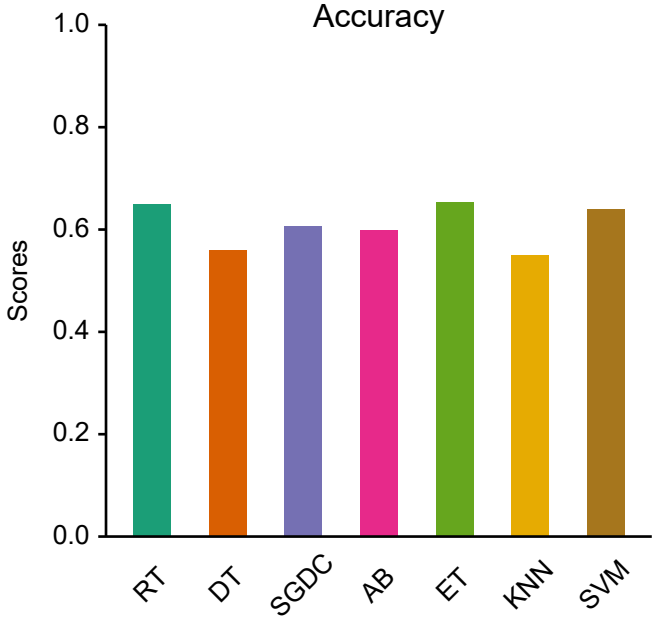












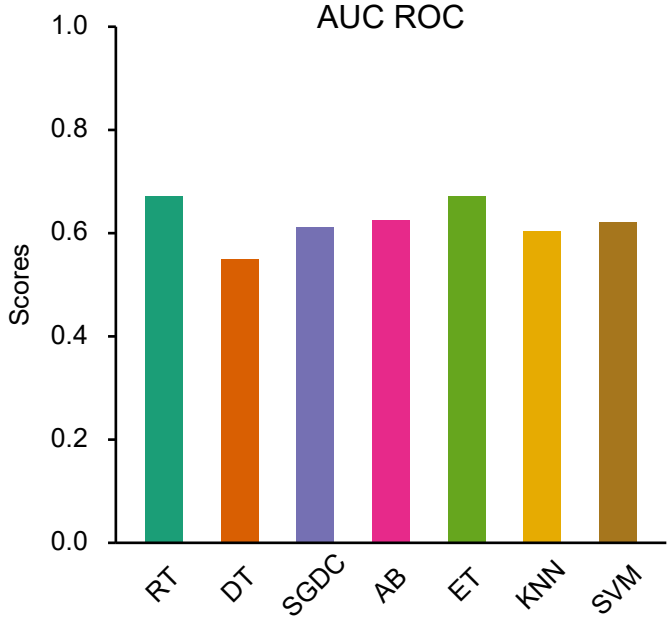
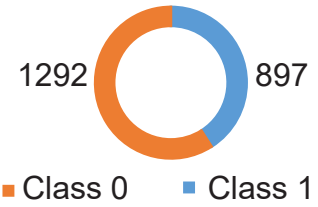
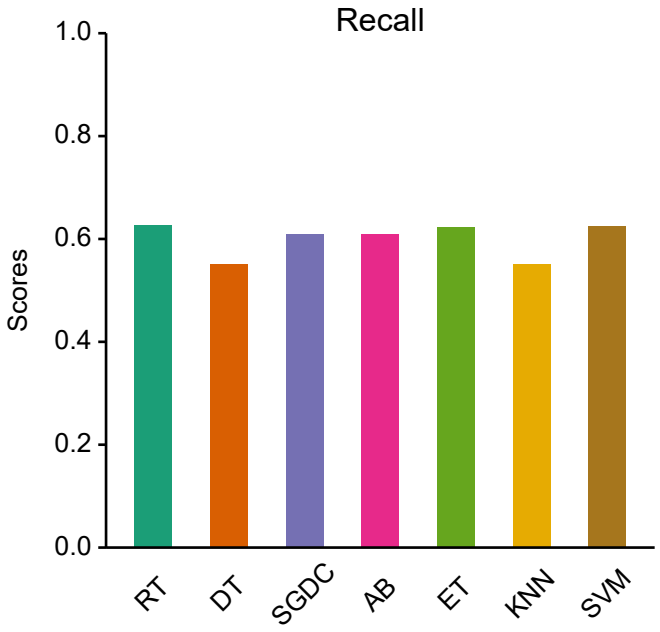
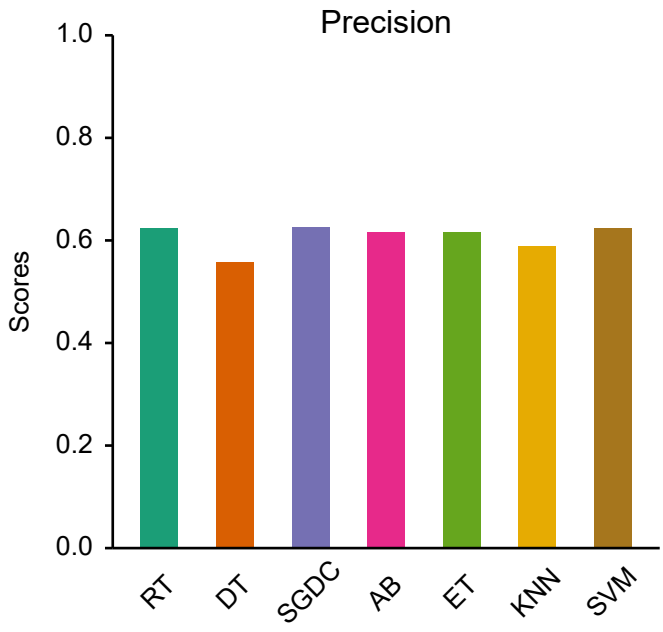


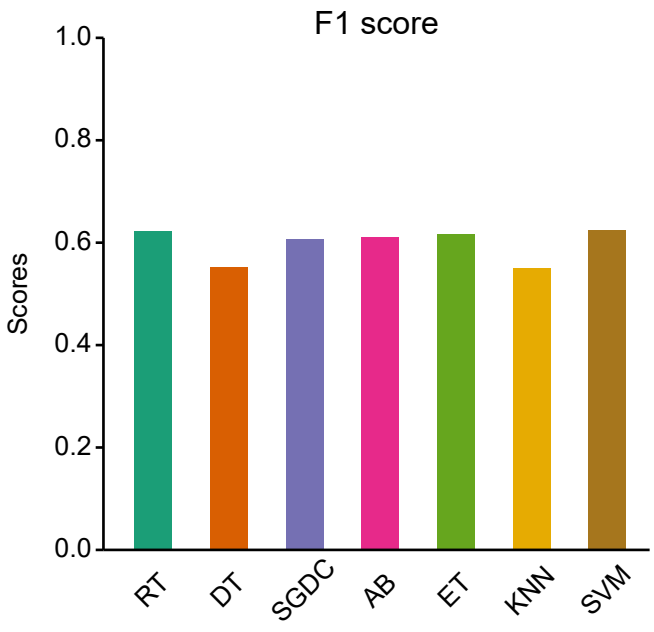
Figure 21.4:

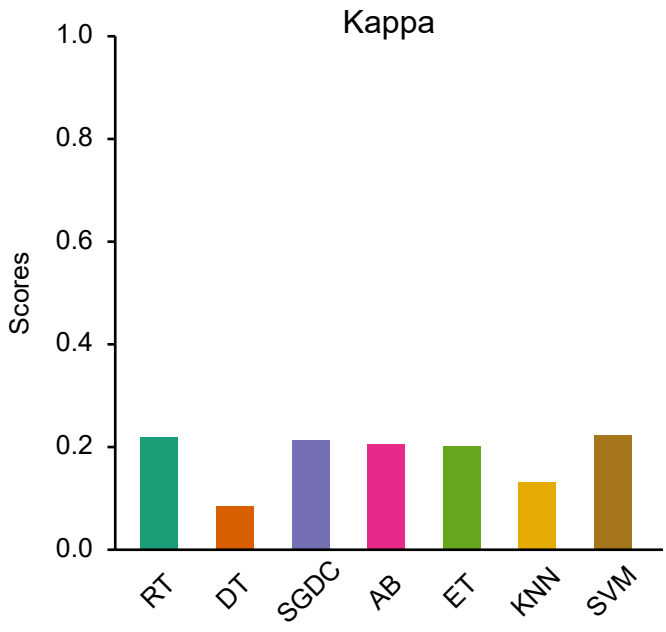
Model Evaluation (10 CV) based on GIPCRT - *mgt1* strain

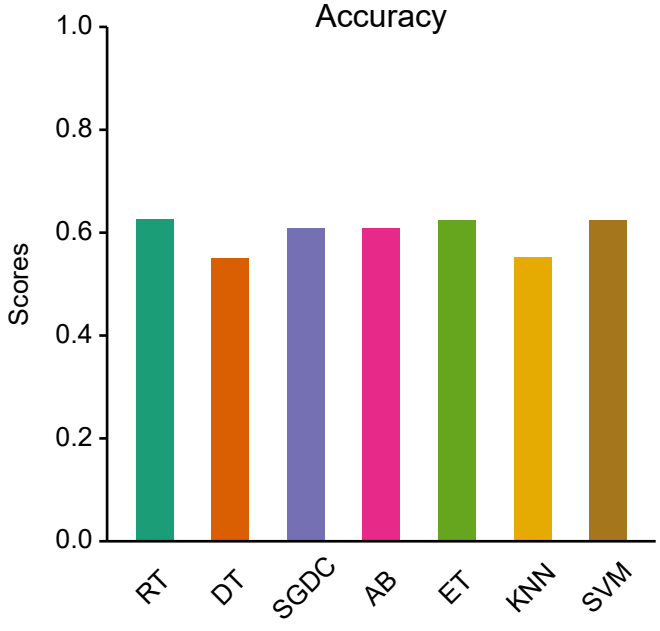












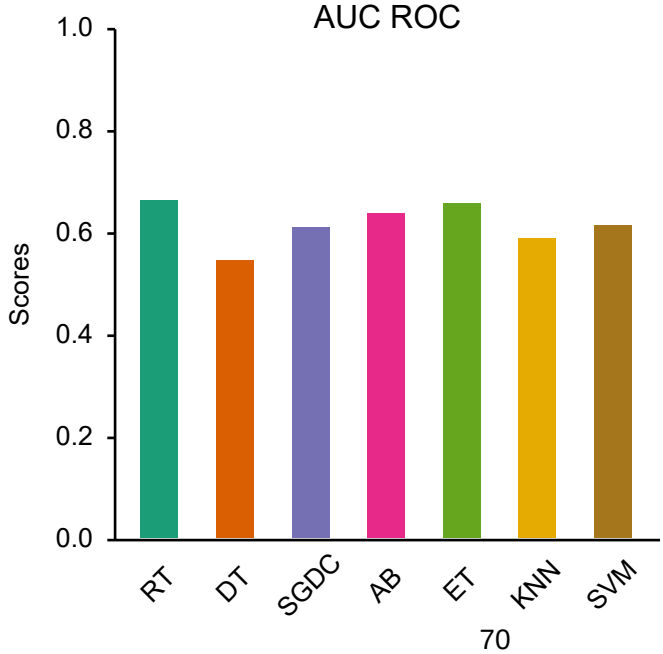
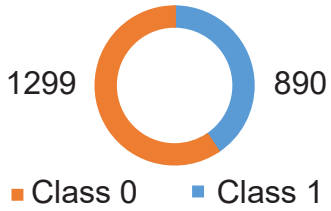
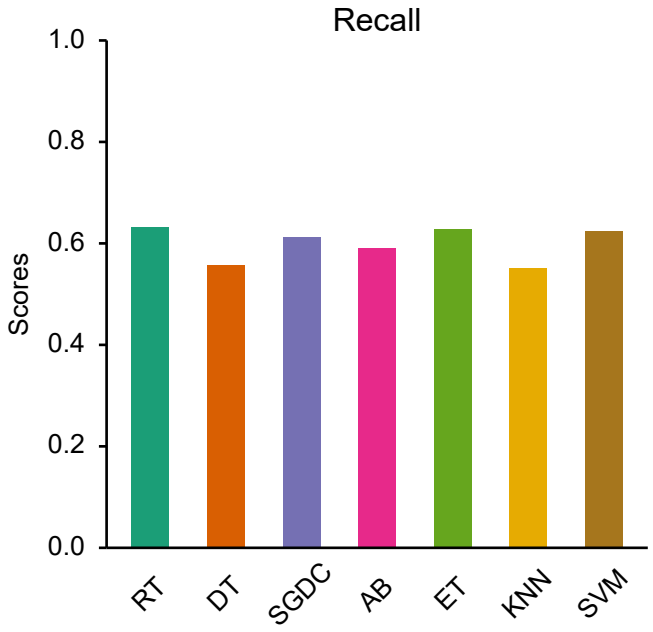
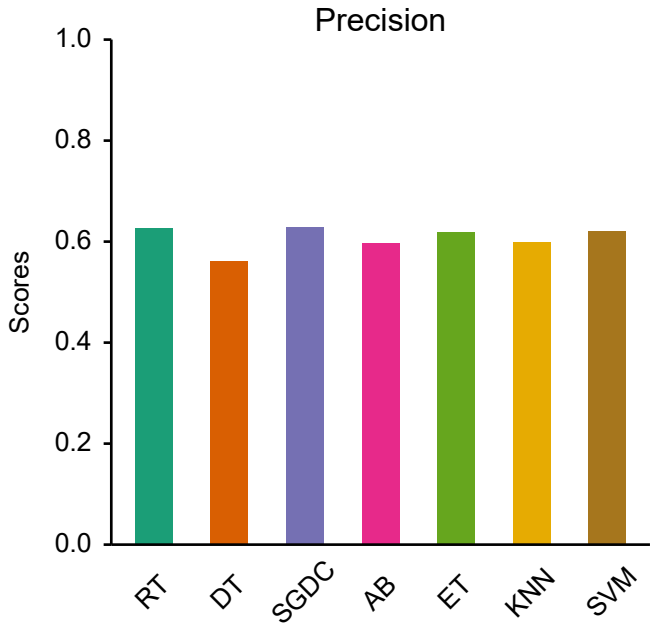


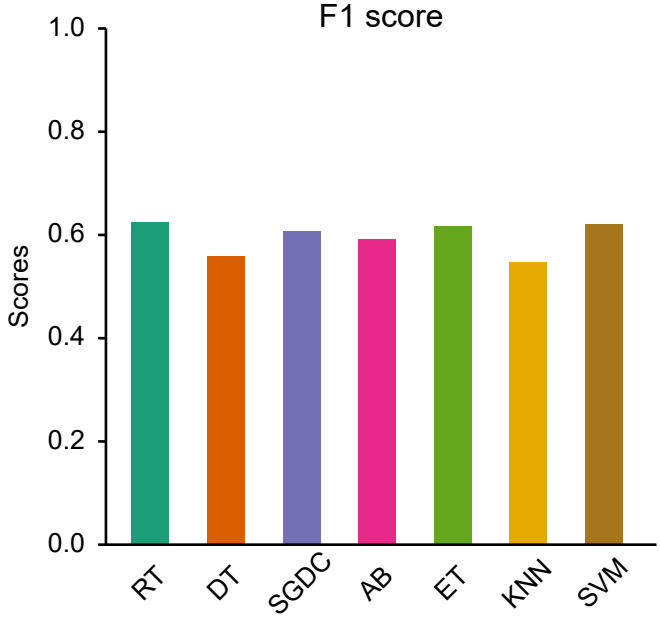
Figure 21.5:

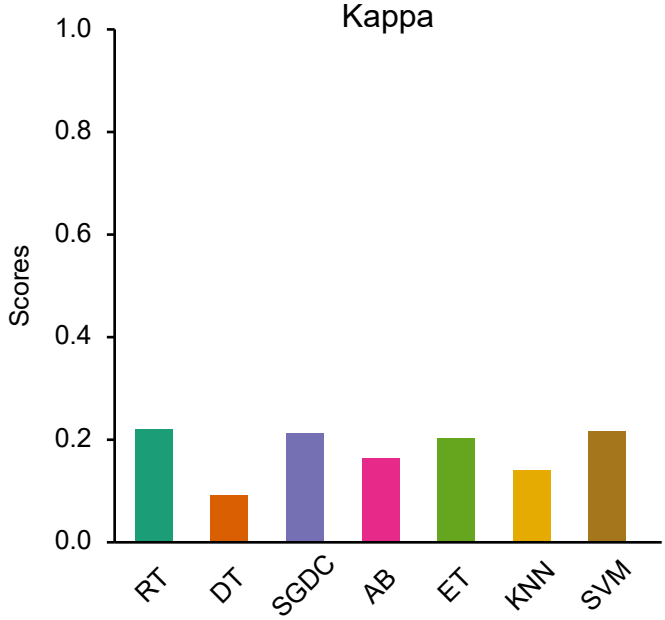
Model Evaluation (10 CV) based on GIPCRT - *mlh1* strain

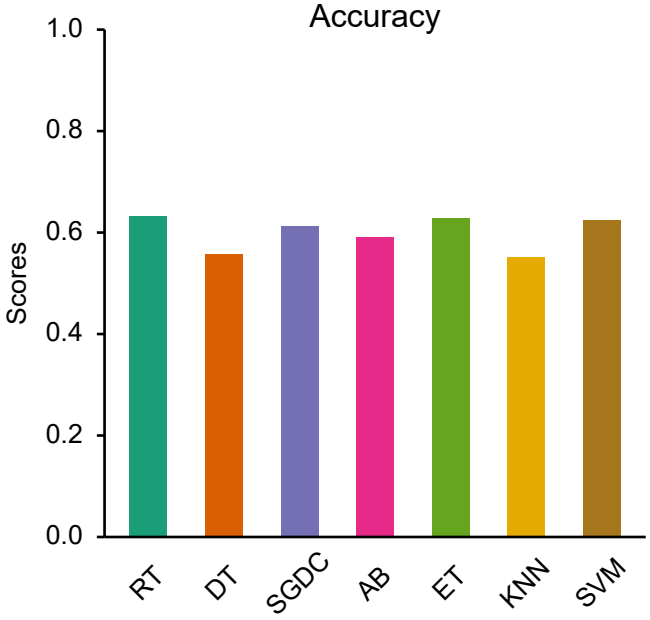












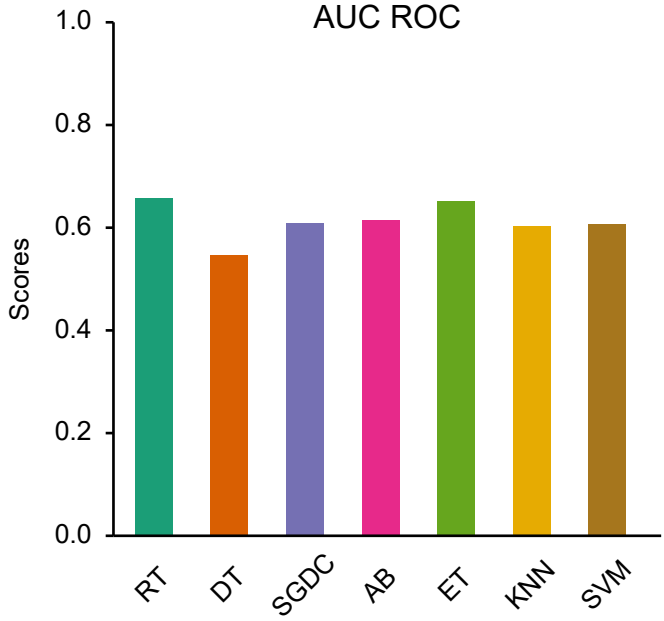


Figure 21.6:

Model Evaluation (10 CV) based on GIPCRT - *rad14* strain

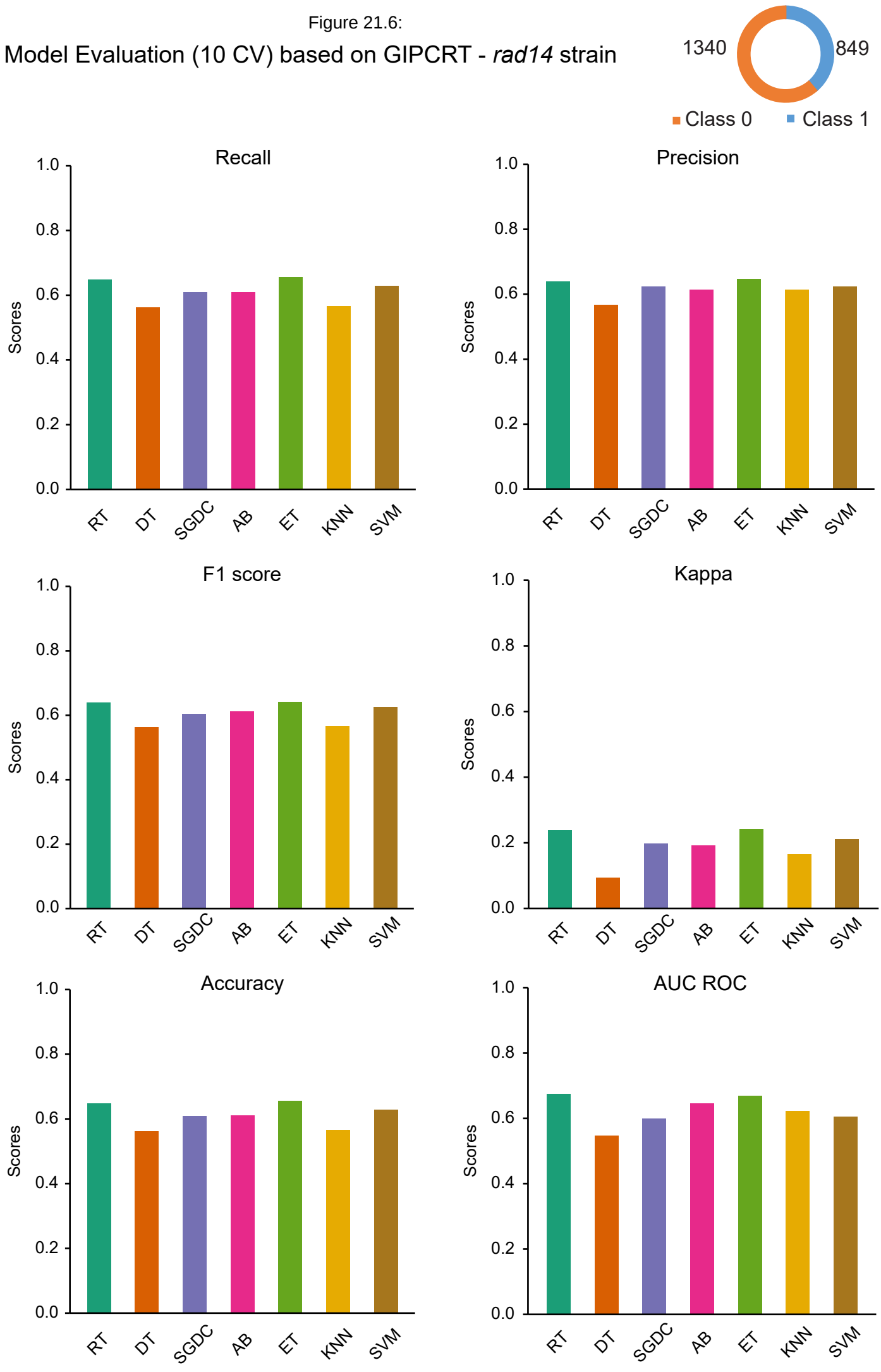
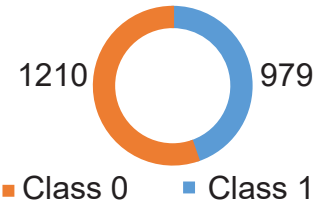
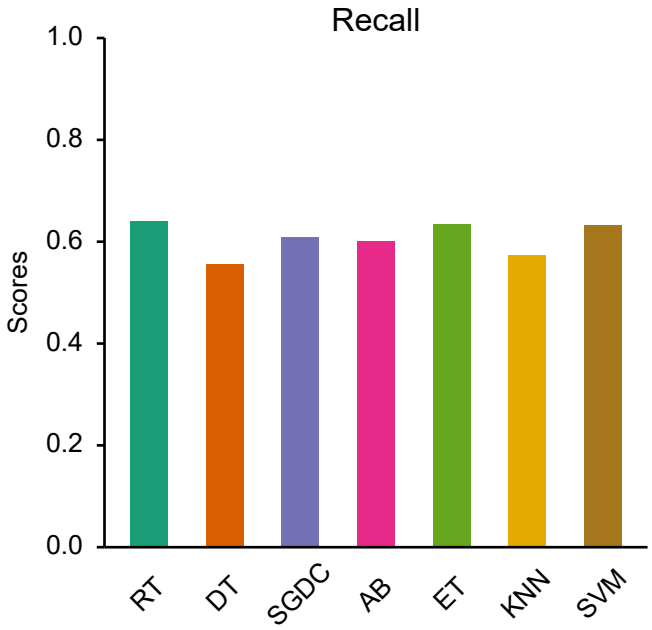
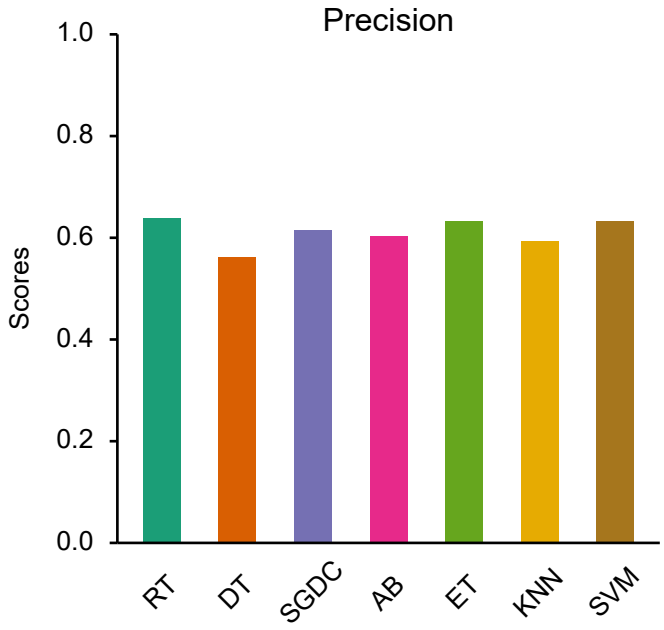


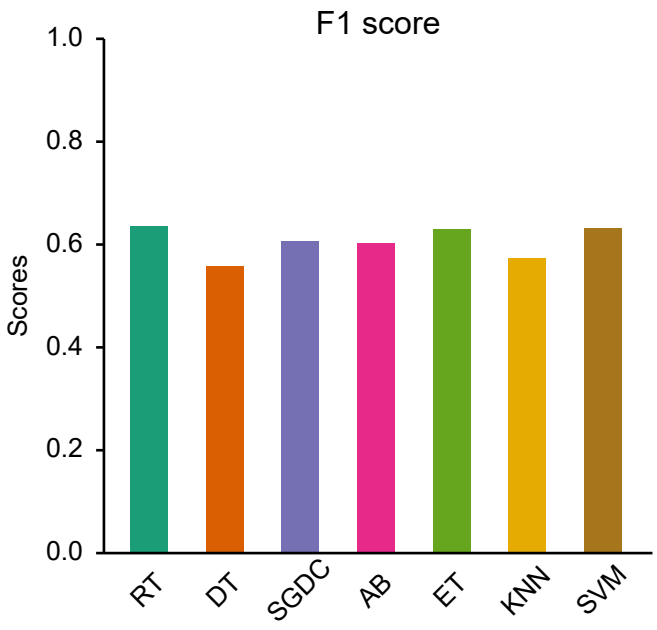
Figure 21.7:

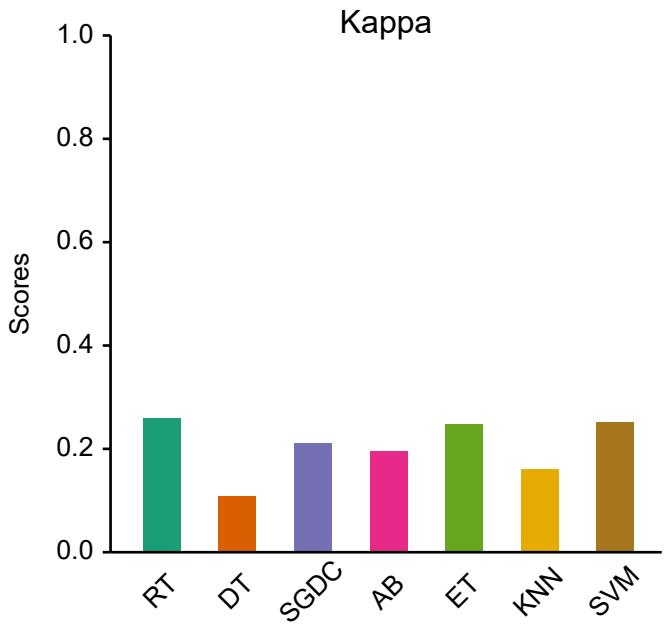
Model Evaluation (10 CV) based on GIPCRT - *rad18* strain

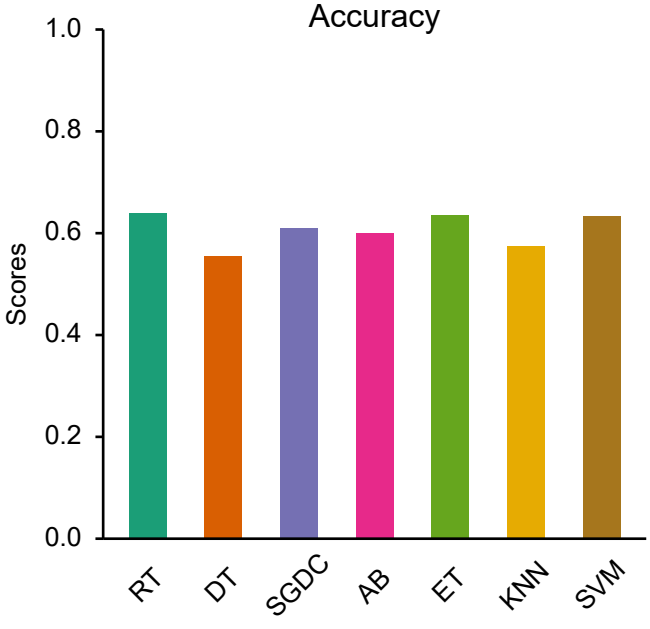












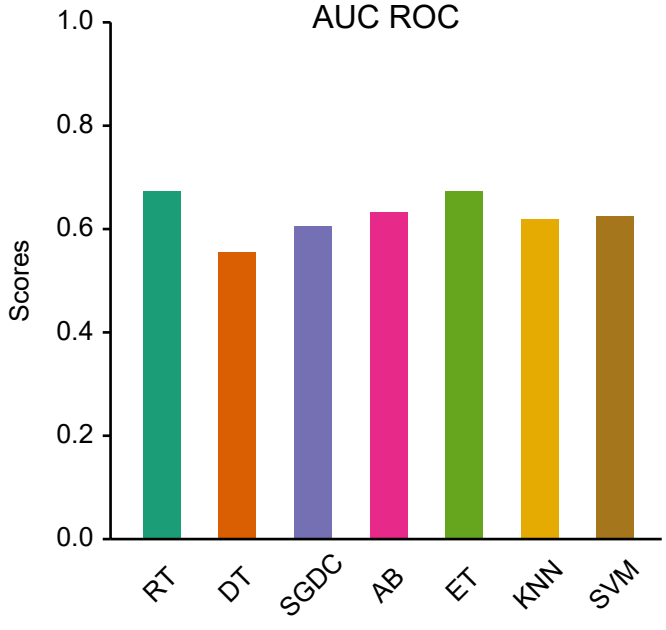
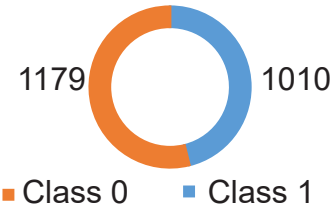
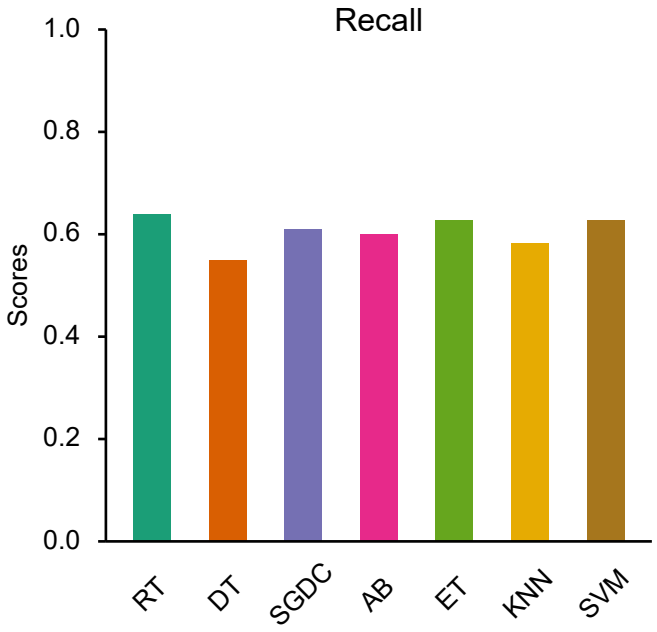
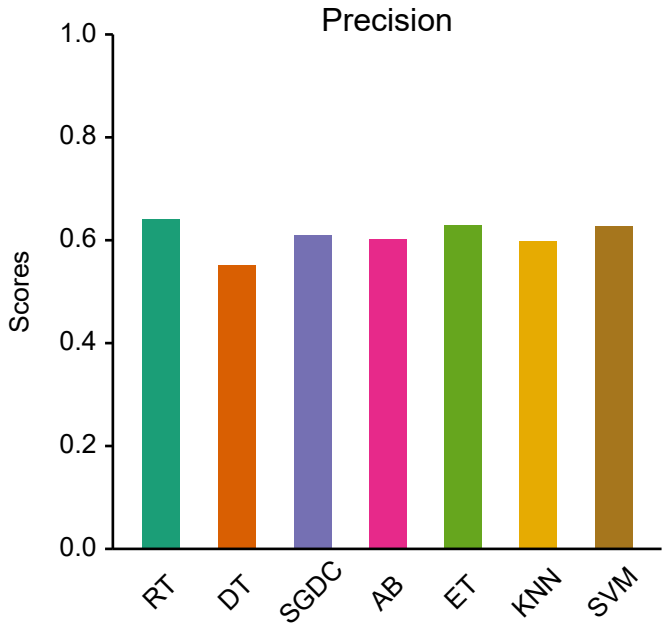


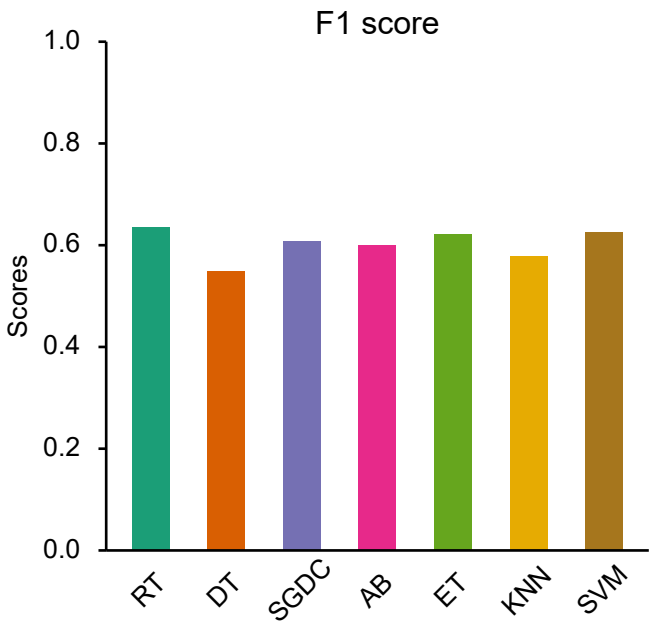
Figure 21.8:

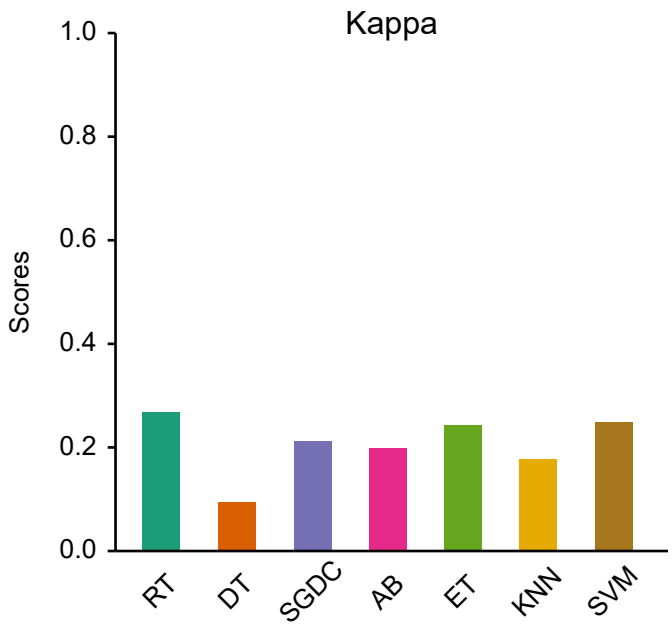
Model Evaluation (10 CV) based on GIPCRT - *rad50* strain

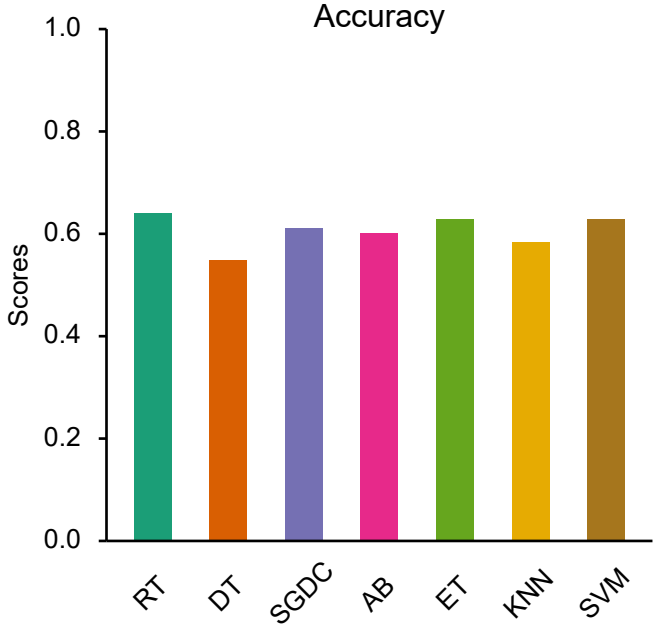












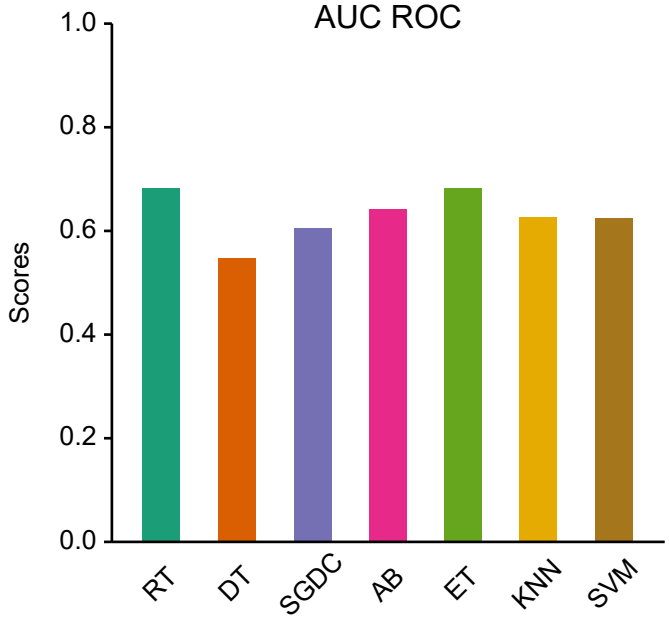
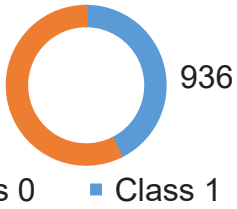
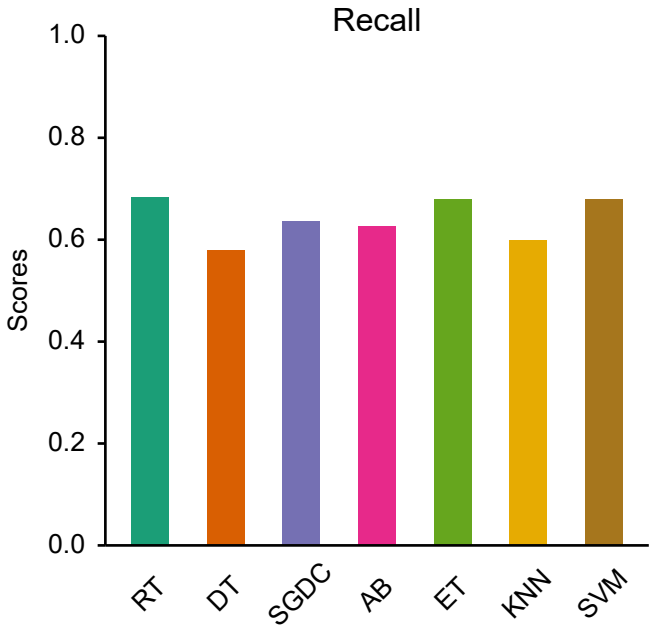
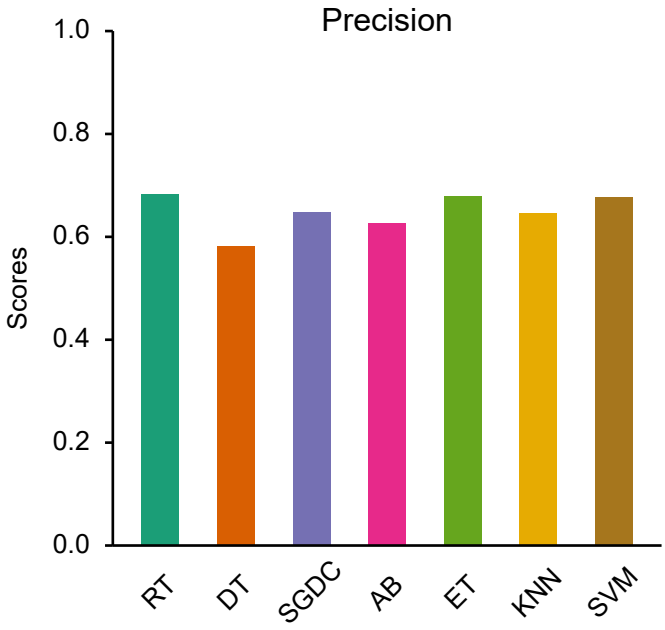


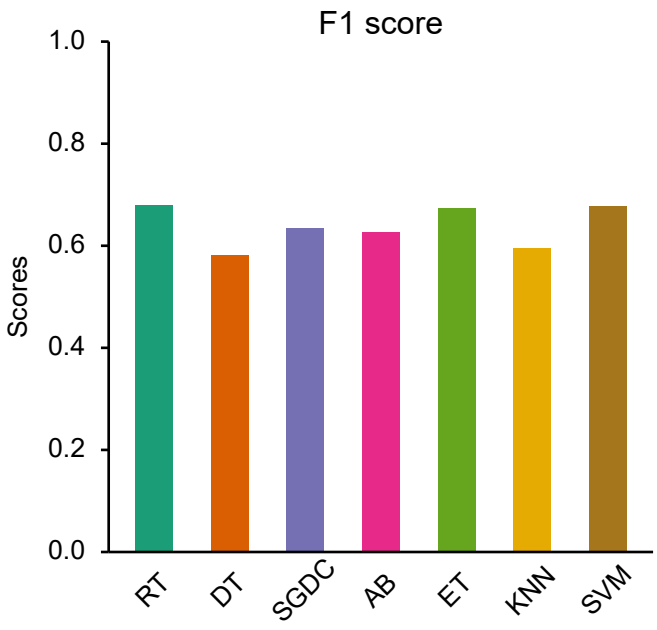
Figure 21.9:

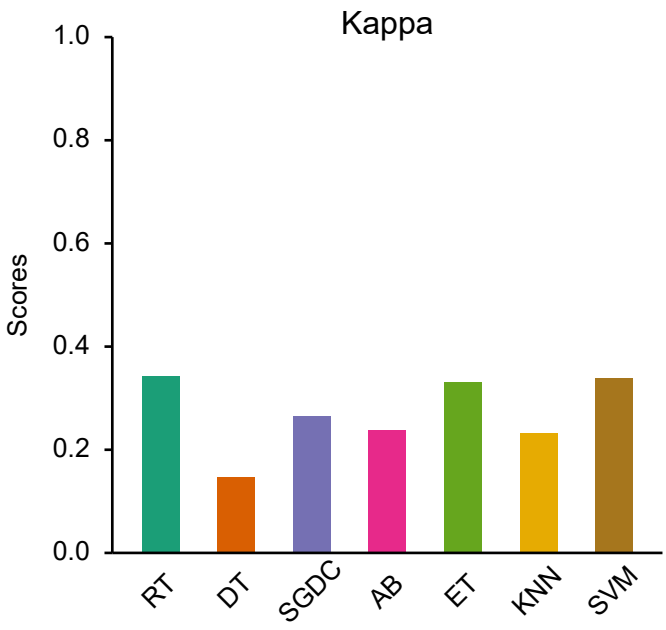
Model Evaluation (10 CV) based on GIPCRT - *rad50EPP+* strain

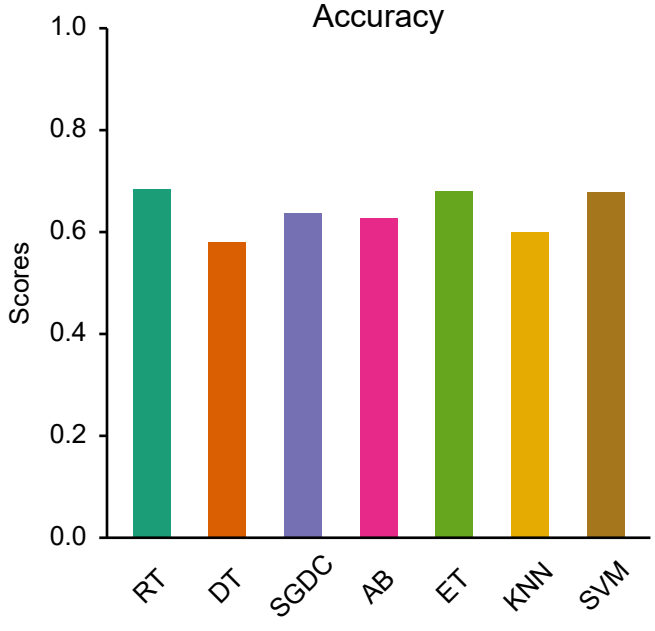












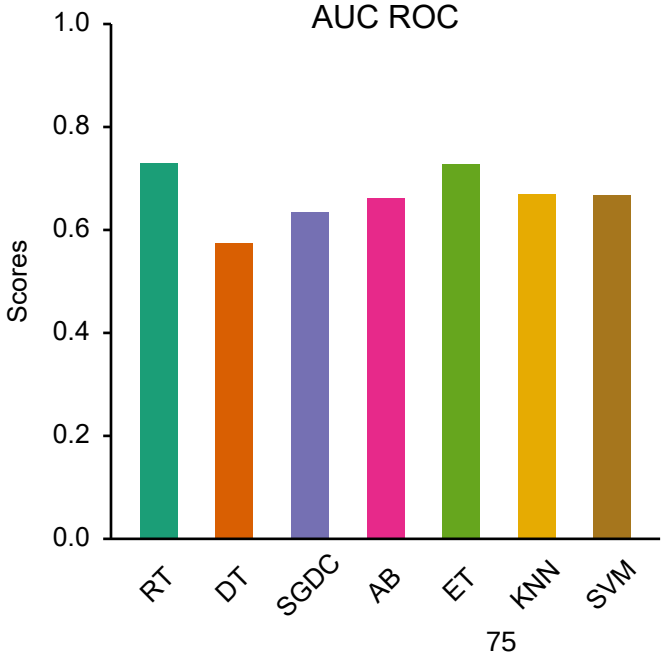
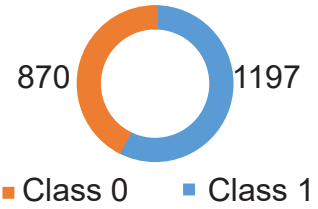
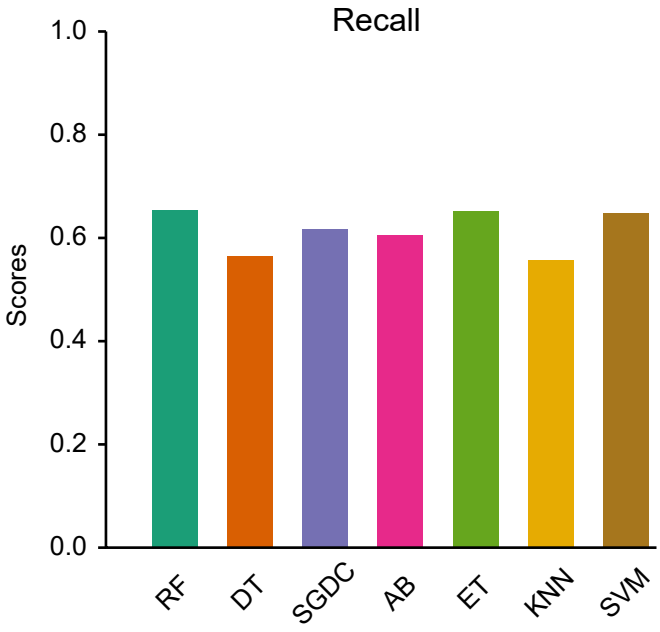
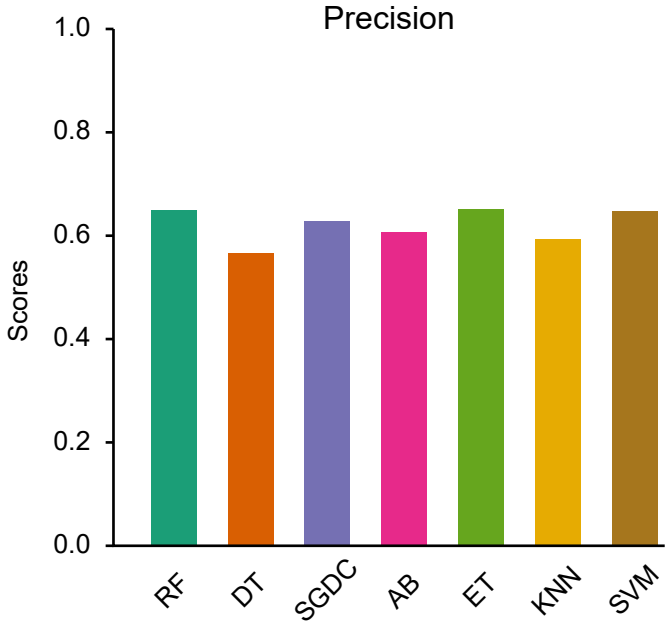


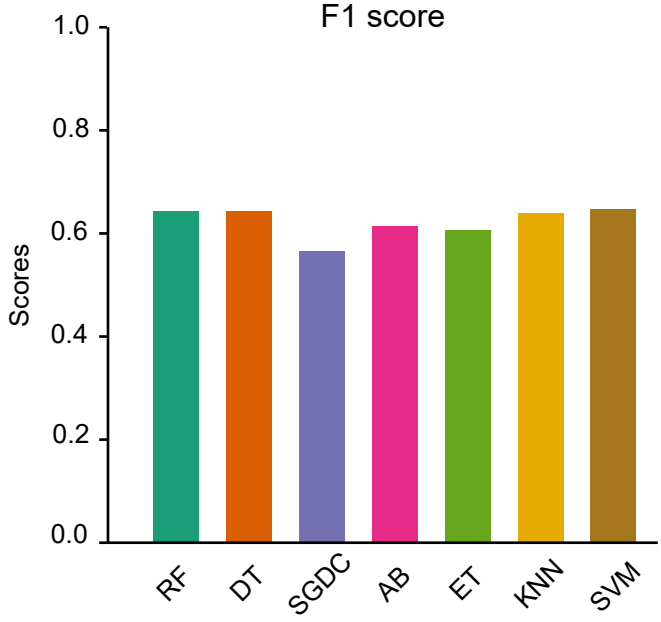
Figure 21.10:

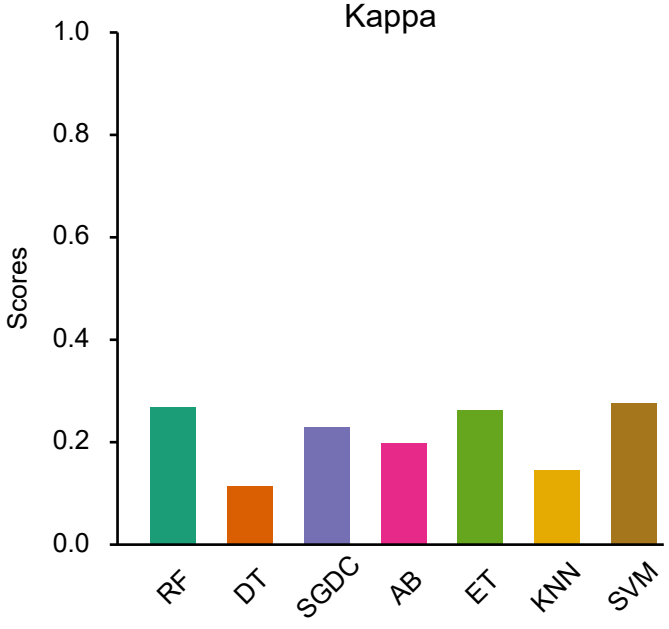
Model Evaluation (10 CV) based on GIPCRT - *rad52* strain

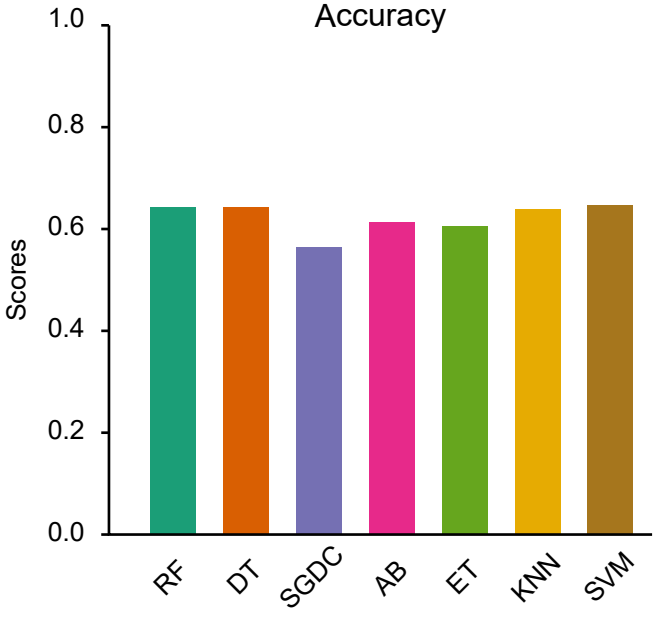












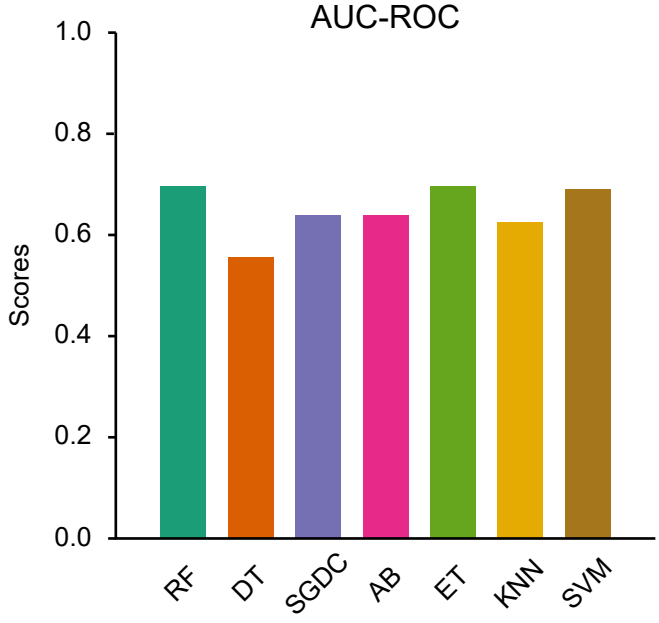
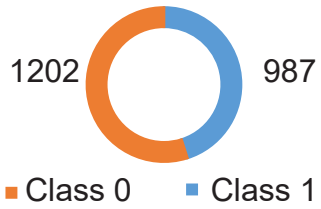
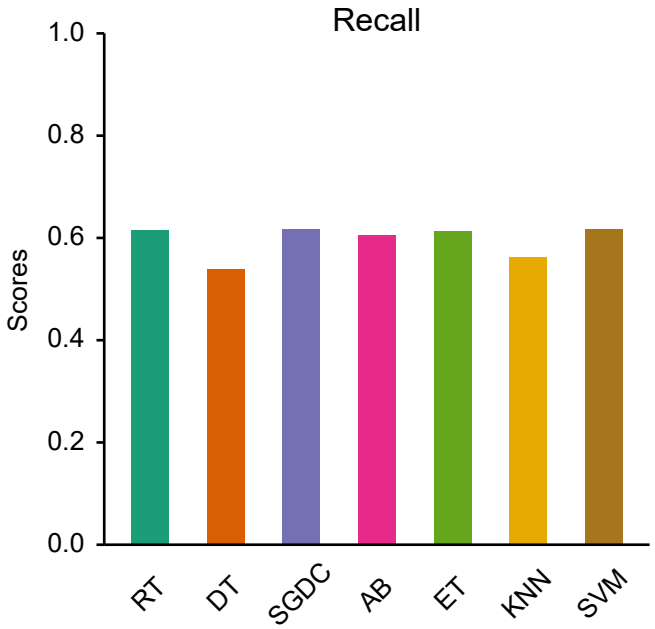
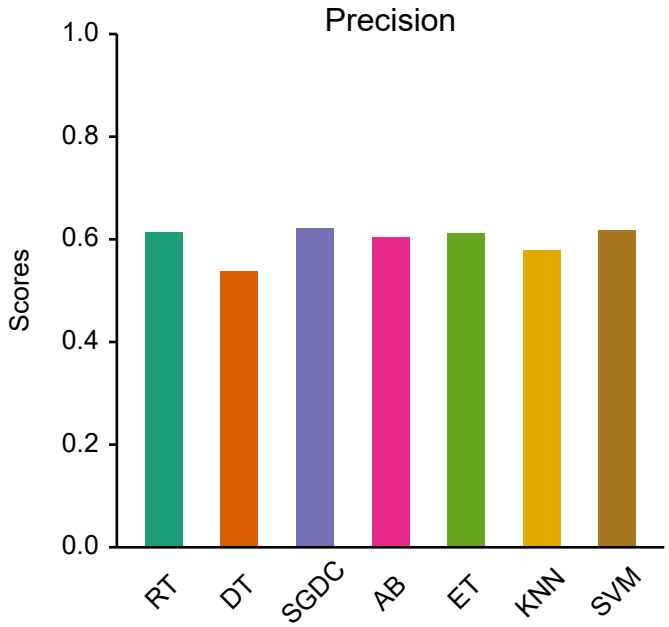


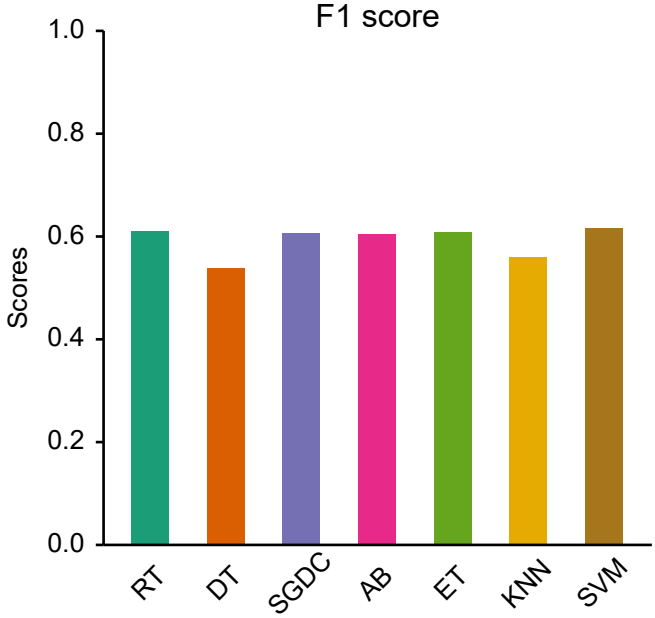
Figure 21.11:

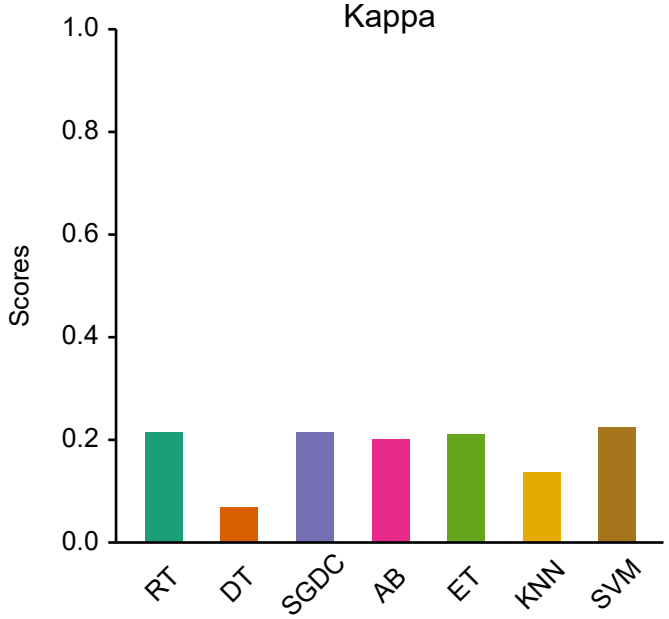
Model Evaluation (10 CV) based on GIPCRT - *sgs1* strain

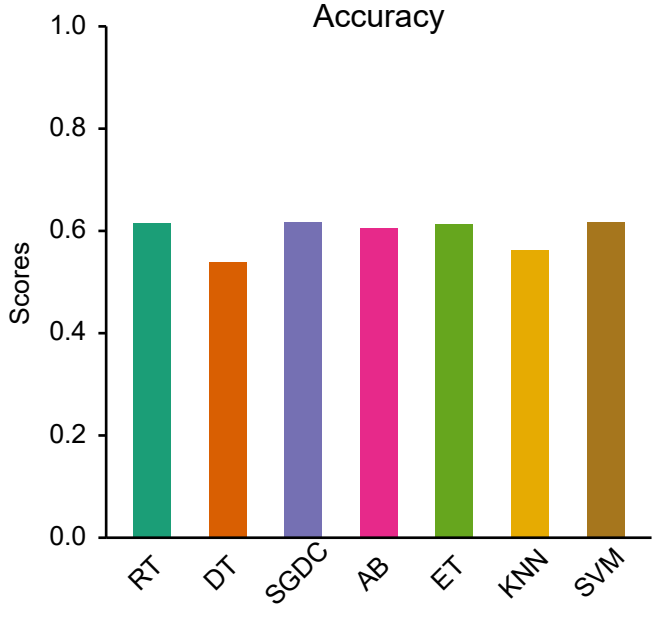












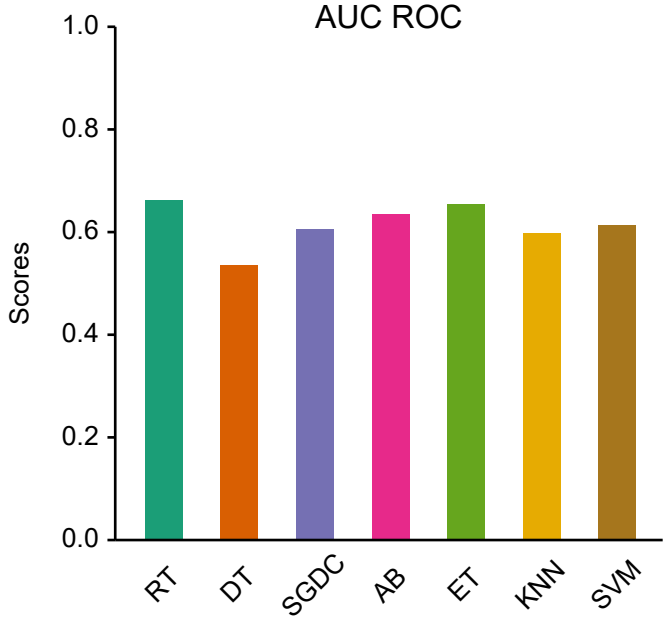
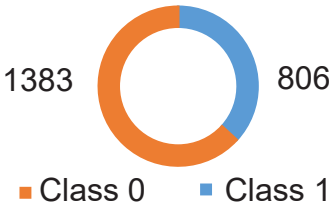
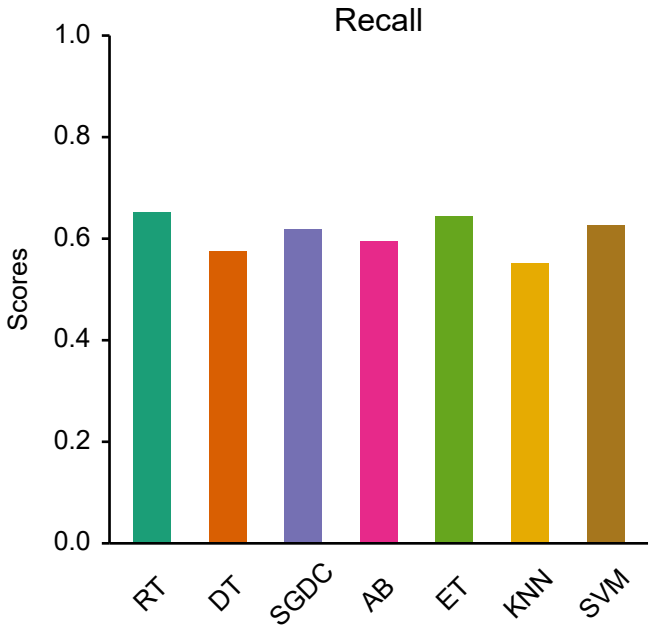
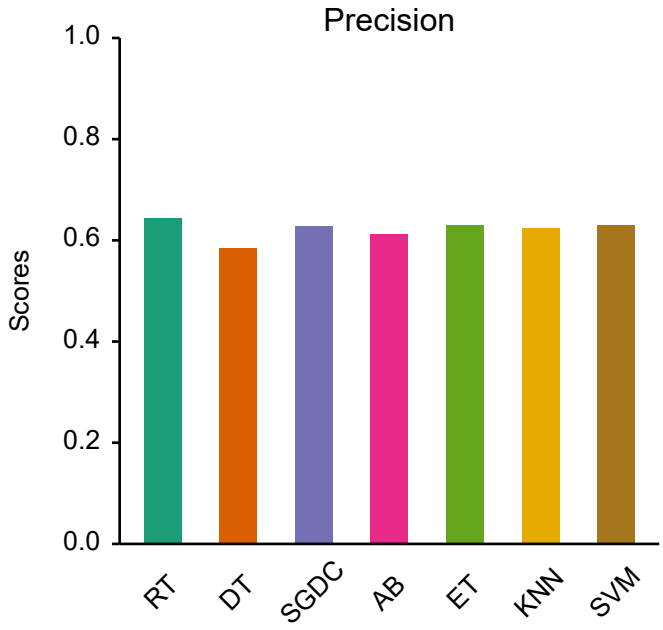


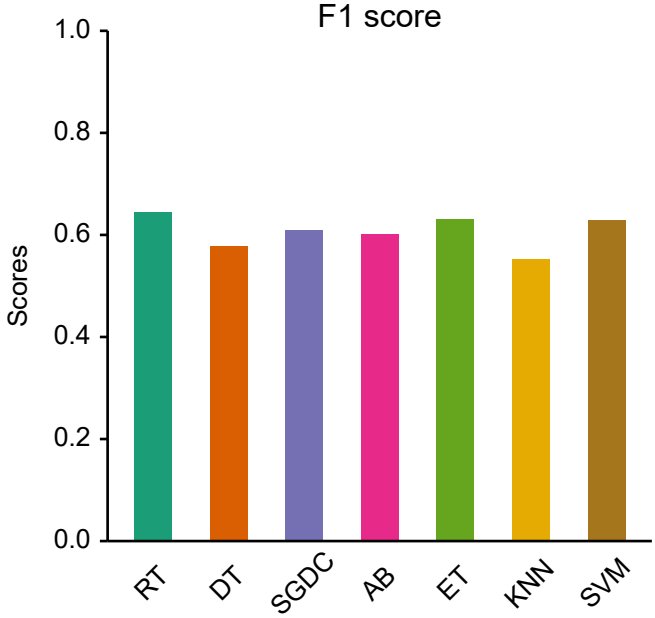
Figure 21.12:

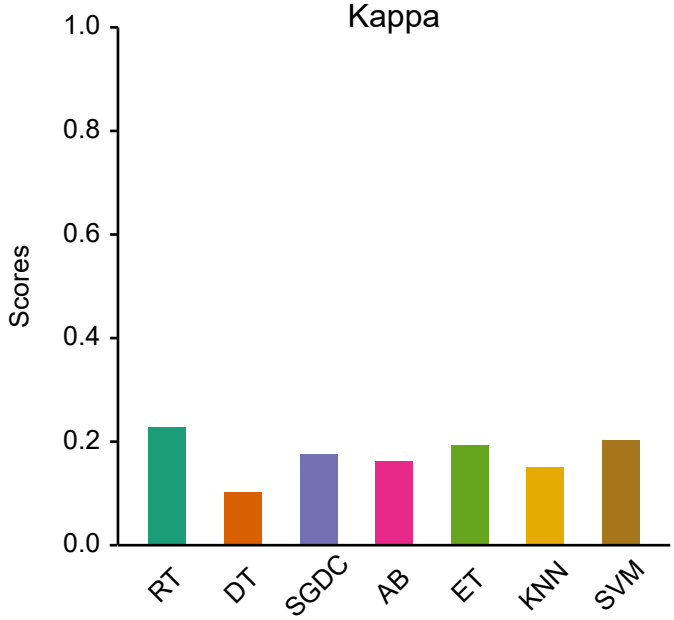
Model Evaluation (10 CV) based on GIPCRT - *wt1* strain

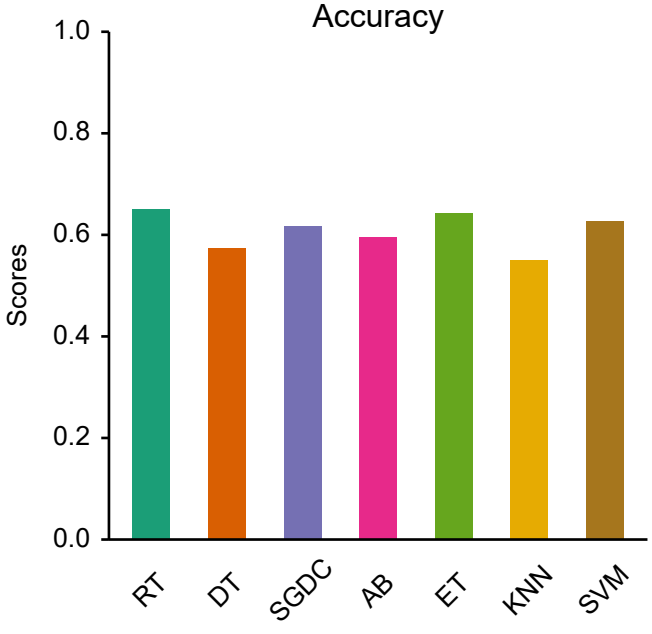












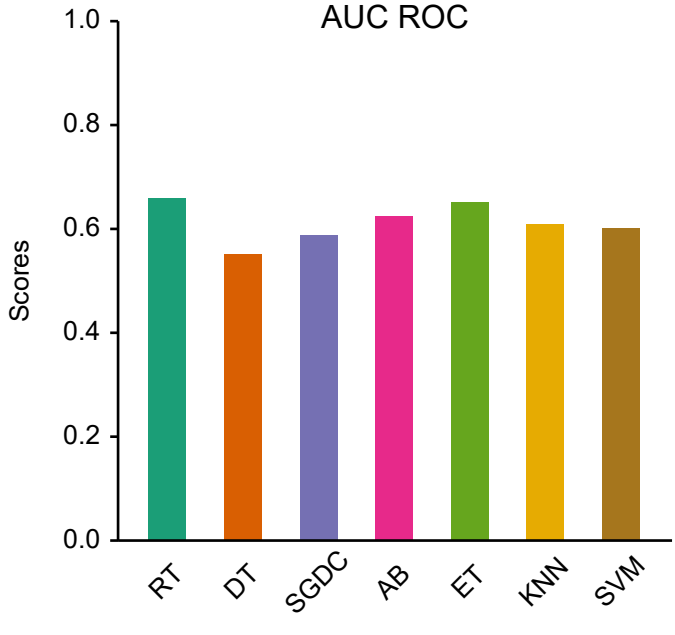
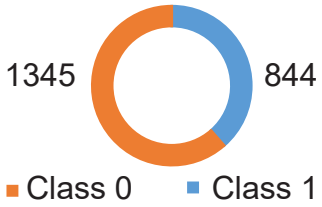
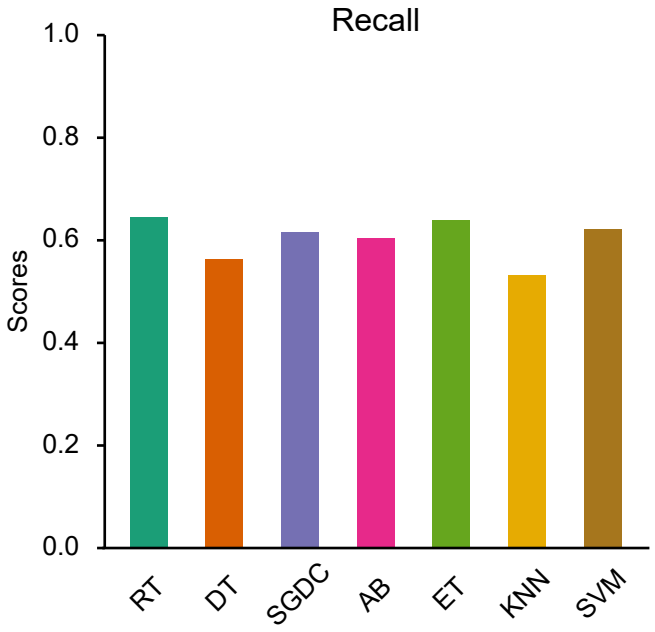
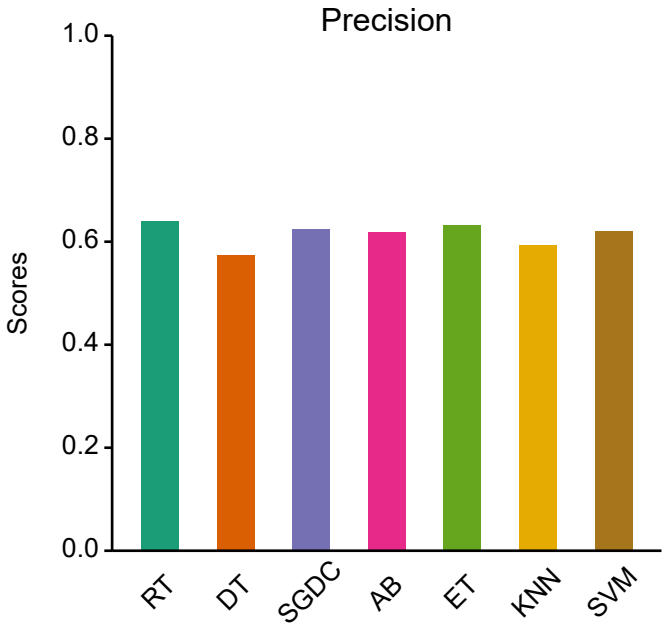


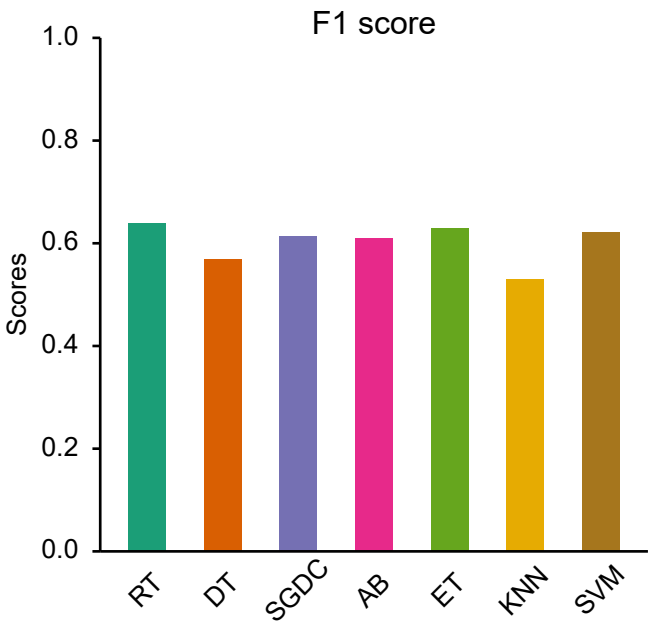
Figure 21.13:

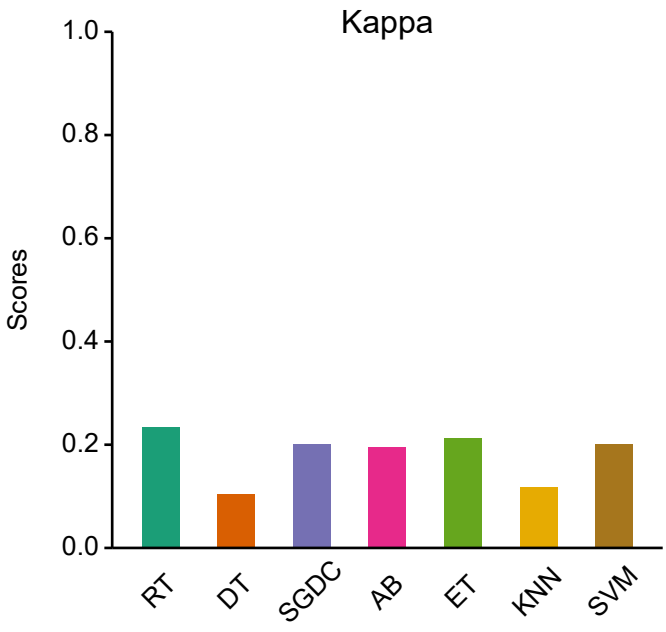
Model Evaluation (10 CV) based on GIPCRT - *wt2* strain

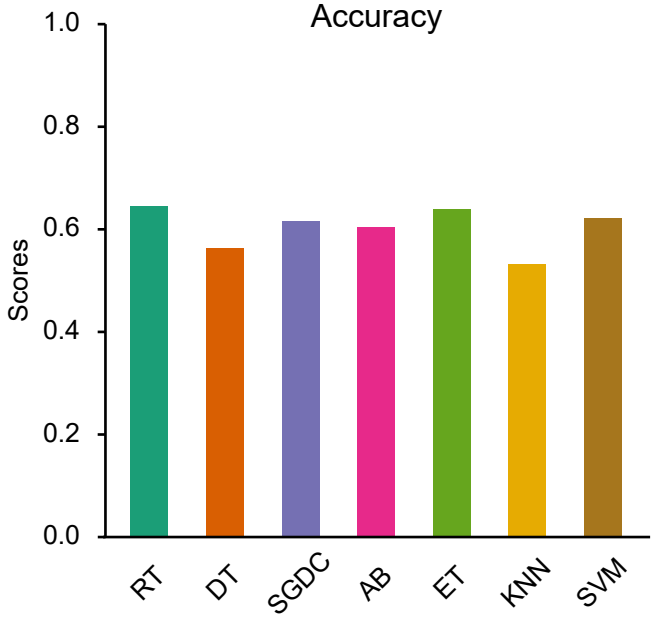












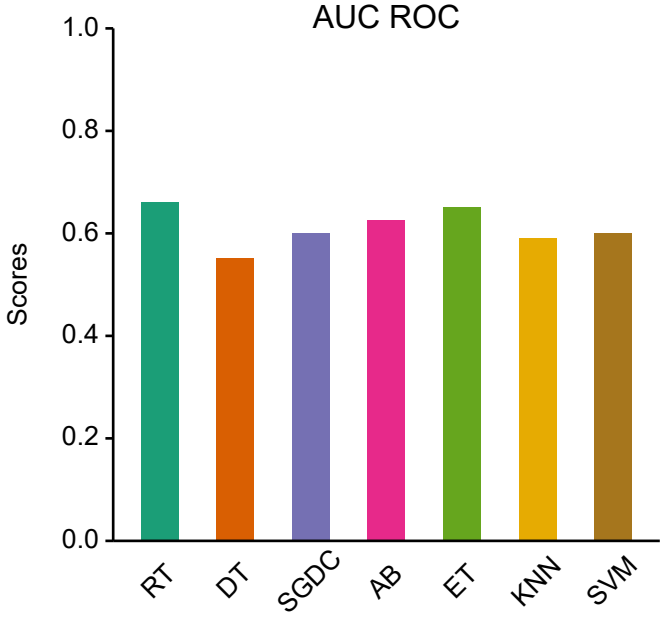
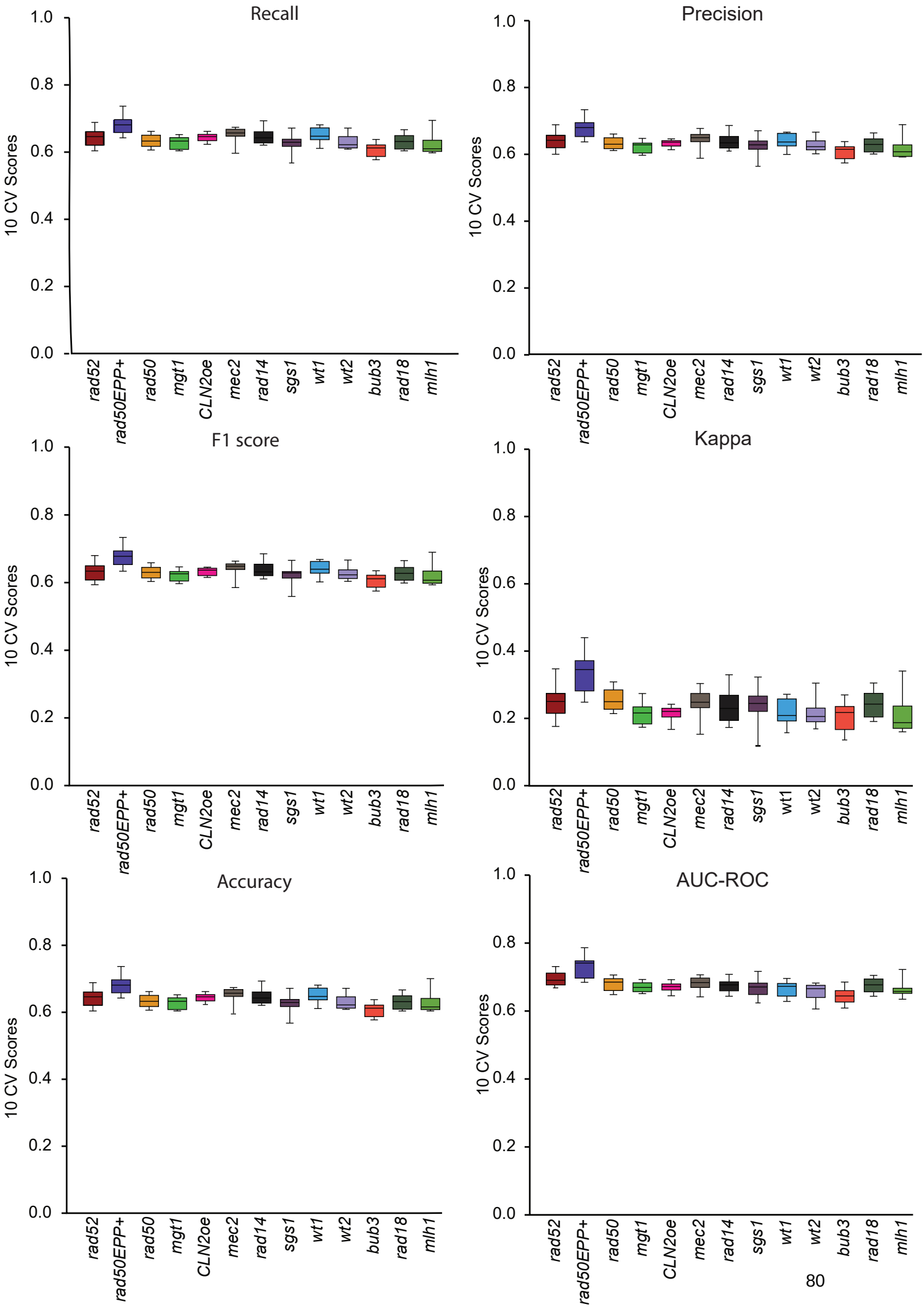


Figure 21.14: GIPCRT 10 fold Cross Validation



V.4 - Unsupervised Classification based on IC50:

The unsupervised classification on the IC50 in all 12 yeast strains (except mlh1) yielded 2 clusters with higher and lower values of IC50. The Kolmogorov–Smirnov test perfmed between the two classes had d value ≅ 0 and a significance p-value of <0.05 suggesting that the two classes are distinct.

The classes from this unsupervised classification are used for supervised classification of the drugs from their chemical space. The best model turned out to be Random forest in all 12 strains. The best model and its parameters were chosen based on the 10 fold cross validation of Kappa and accuracy scores.

Table 9: Best parameter for Random Forest Classifier on IC50 values

Yeast Strain	IC50 range		Best Parameters in Random Forest
	class0	class1	
rad52	1.2 - 54.2	54.5 - 100	criterion: gini, max_depth: 12, max_features: log2, n_estimators: 800
rad50	1.2 - 54.3	54.5 - 100	criterion: gini, max_depth: 16, max_features: auto, n_estimators: 180
rad14	1.2 - 58.9	59.2 - 100	criterion: entropy, max_depth: 18, max_features: log2, n_estimators: 700
rad18	1.2 - 51.6	51.8 - 100	criterion: gini, max_depth: 10, max_features: auto, n_estimators: 100
rad50EPP+	1.2 - 56.2	56.4 - 100	criterion: gini, max_depth: 18, max_features: log2, n_estimators: 170
sgs1	1.2 - 55.6	55.8 - 100	criterion: gini, max_depth: 18, max_features: auto, n_estimators: 180
mgt1	1.2 - 57.3	57.4 - 100	criterion: gini, max_depth: 14, max_features: auto, n_estimators: 200

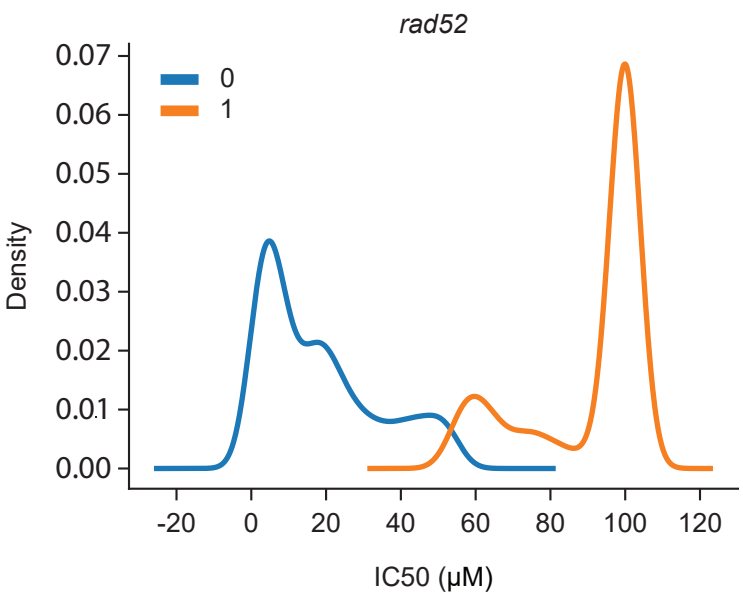
mec2	1.2 - 57.0	57.1 - 100	criterion: gini, max_depth: 28, max_features: log2, n_estimators: 700
bub3	1.2 - 57.9	58.5 - 100	criterion: entropy, max_depth: 10, max_features: log2, n_estimators: 1000
CLN2oe	1.2 - 58.2	58.3 - 100	criterion: gini, max_depth: 18, max_features: log2, n_estimators: 300
wt1	1.2 - 59.0	59.2 - 100	criterion: gini, max_depth: 14, max_features: log2, n_estimators: 400
wt2	1.2 - 57.7	57.9 - 100	criterion: entropy, max_depth: 18, max_features: auto, n_estimators: 300

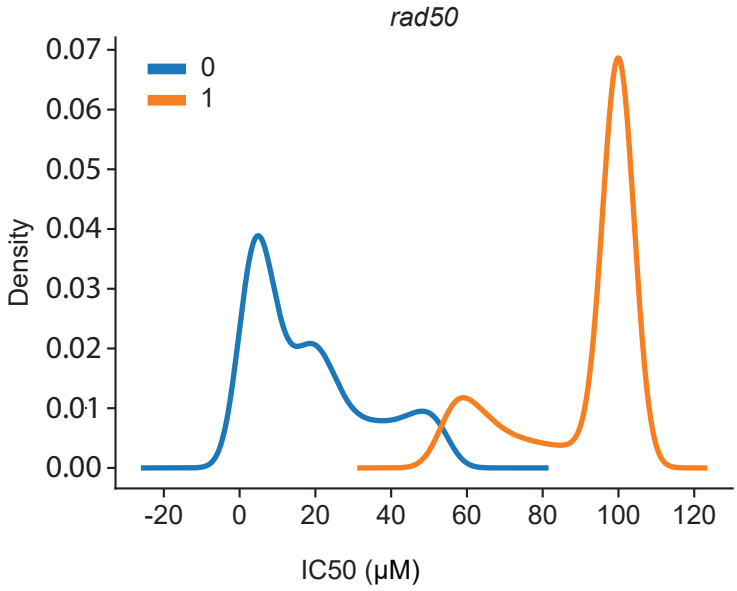
Table 10: 10 CV evaluation metrics for all 12 yeast strains based on IC50

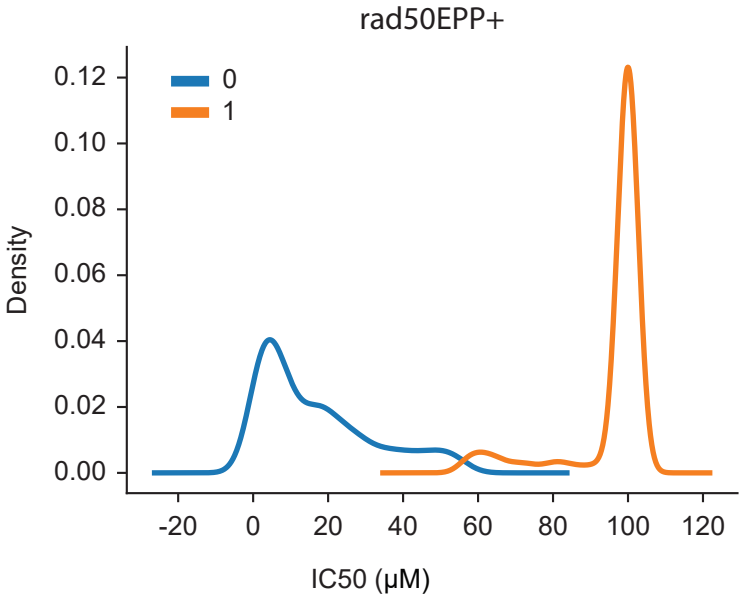
Yeast strain	Recall	Precision	F1	Kappa	Accuracy	AUC
rad52	0.67	0.67	0.67	0.32	0.67	0.71
rad50	0.65	0.65	0.65	0.27	0.65	0.7
rad14	0.63	0.63	0.63	0.25	0.63	0.67
rad18	0.64	0.64	0.63	0.25	0.64	0.67
rad50EPP+	0.67	0.68	0.67	0.34	0.67	0.74
sgs1	0.64	0.63	0.64	0.22	0.64	0.65
mgt1	0.60	0.61	0.60	0.21	0.60	0.65
mec2	0.62	0.62	0.62	0.23	0.62	0.66
bub3	0.60	0.60	0.60	0.2	0.60	0.64
CLN2oe	0.61	0.61	0.61	0.21	0.61	0.66
wt1	0.63	0.63	0.63	0.23	0.63	0.66
wt2	0.62	0.62	0.62	0.24	0.62	0.65

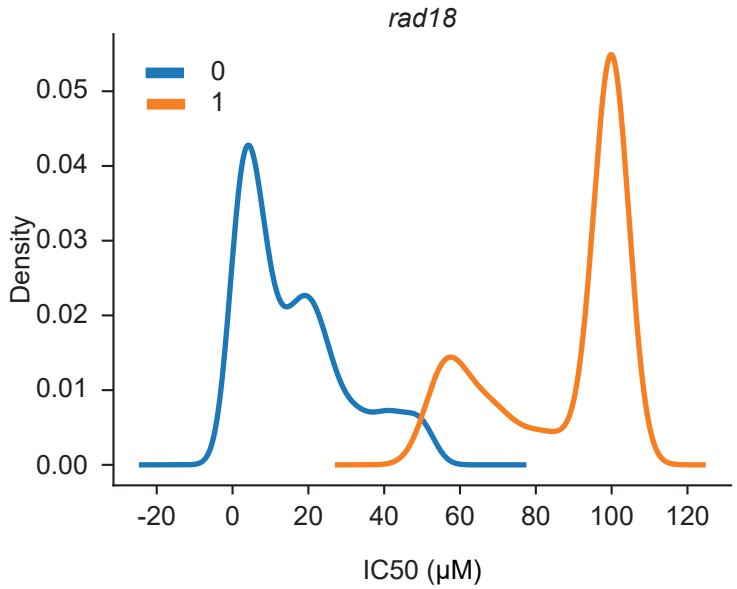
Figure 22.1:

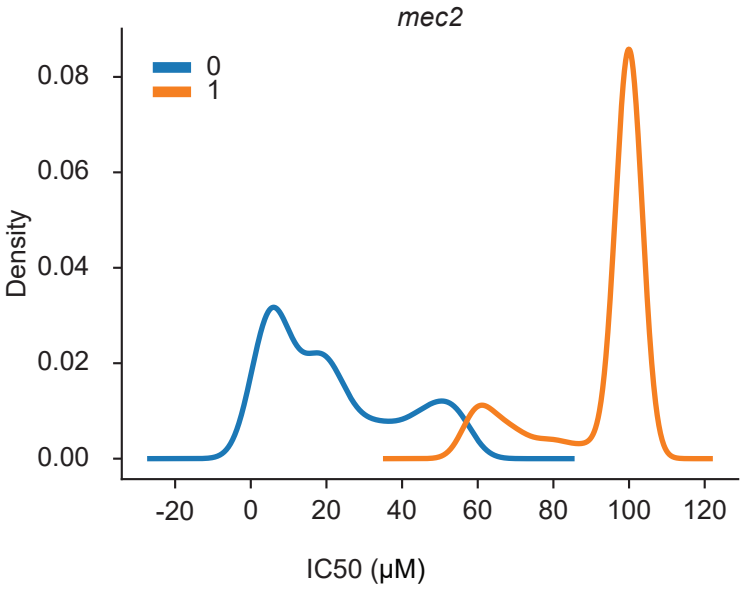
Unsupervised Classification based on IC50











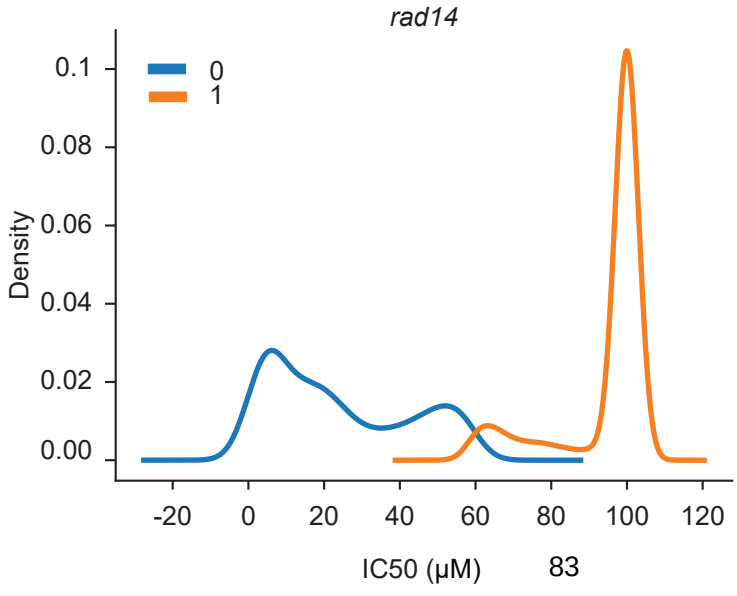


Figure 22.2:

Unsupervised Classification based on IC50

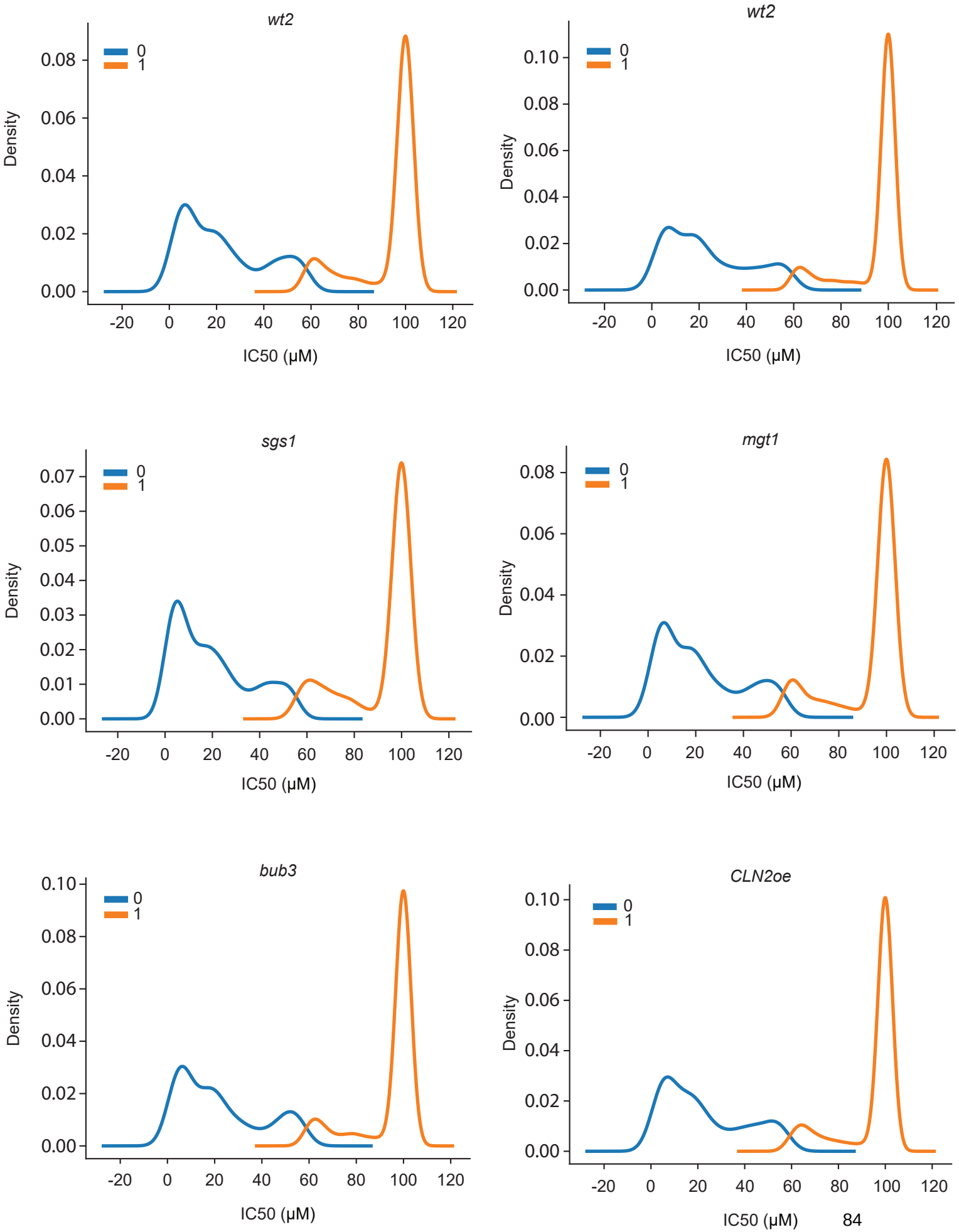
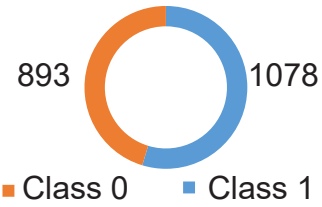
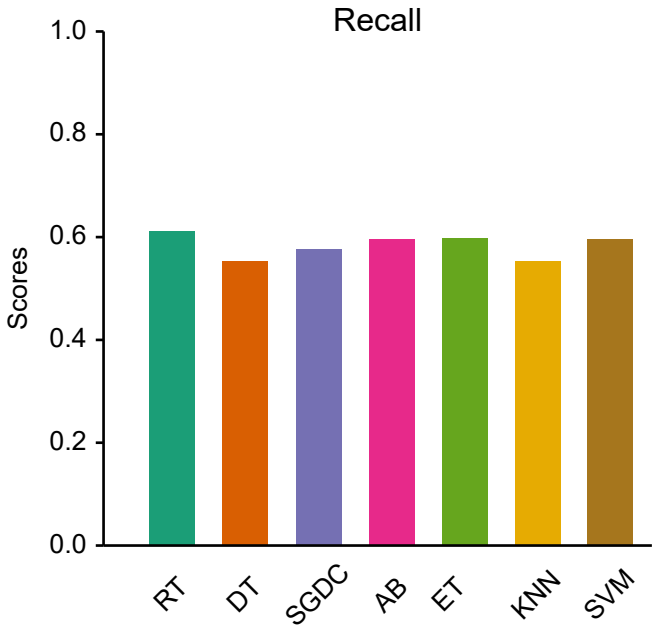
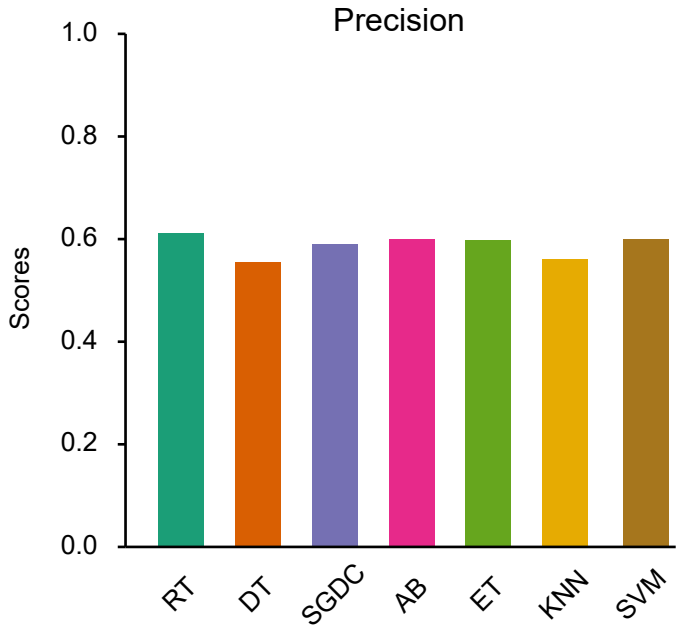


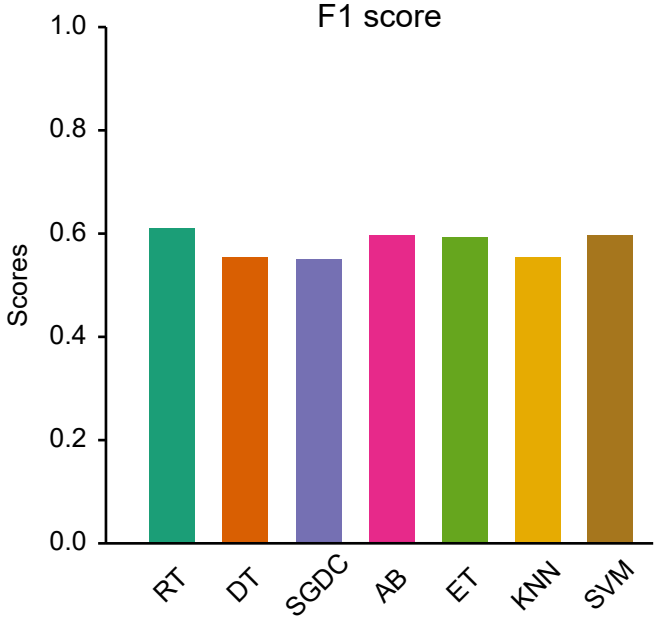
Figure 23.1:

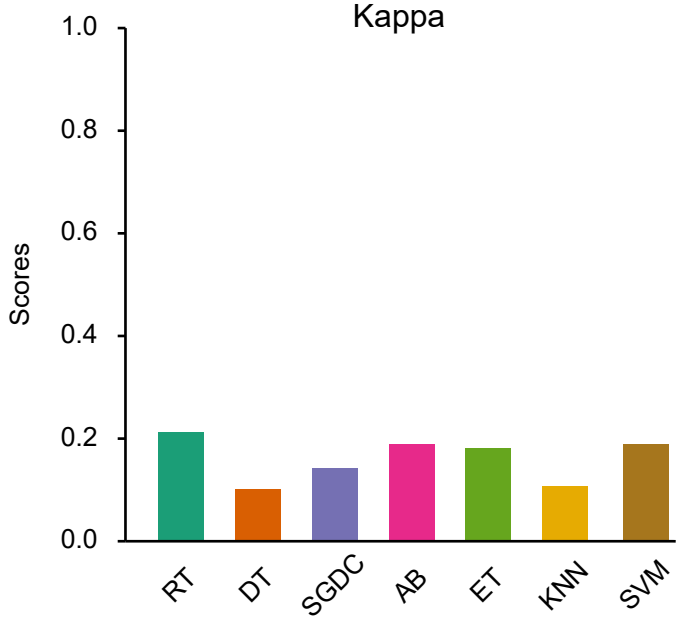
Model Evaluation (10 CV) based on IC50 - *bub3* strain

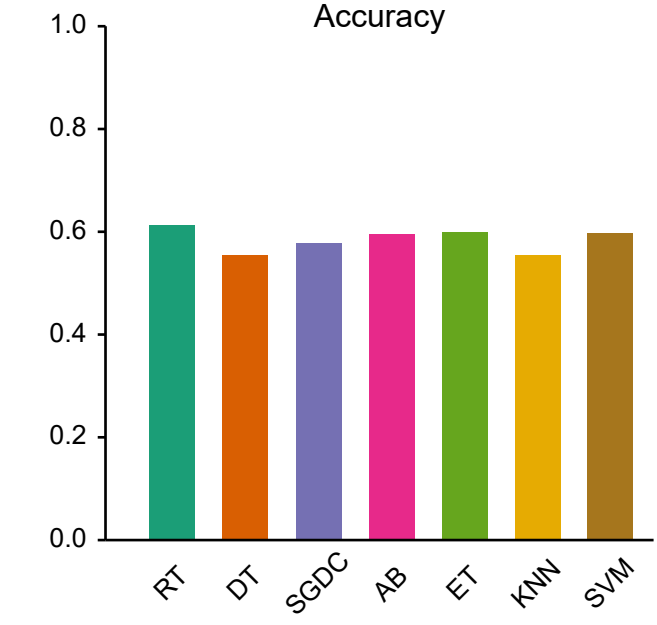












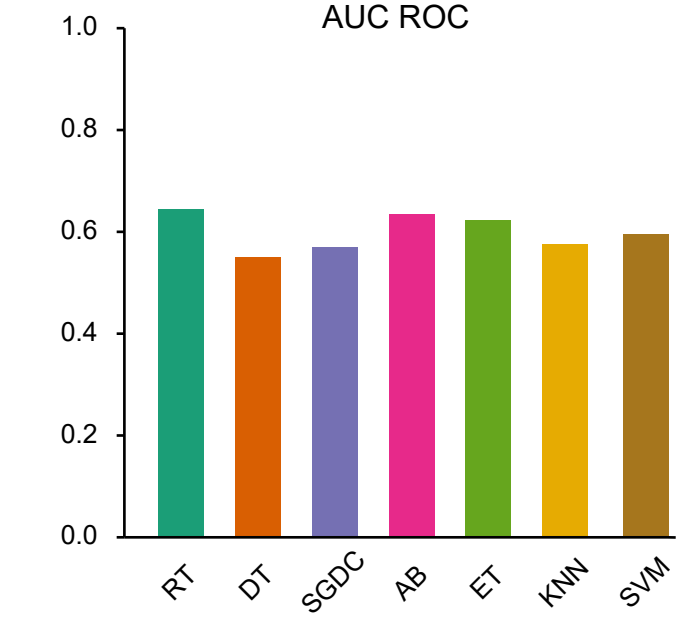
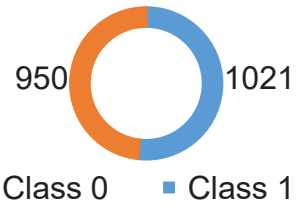
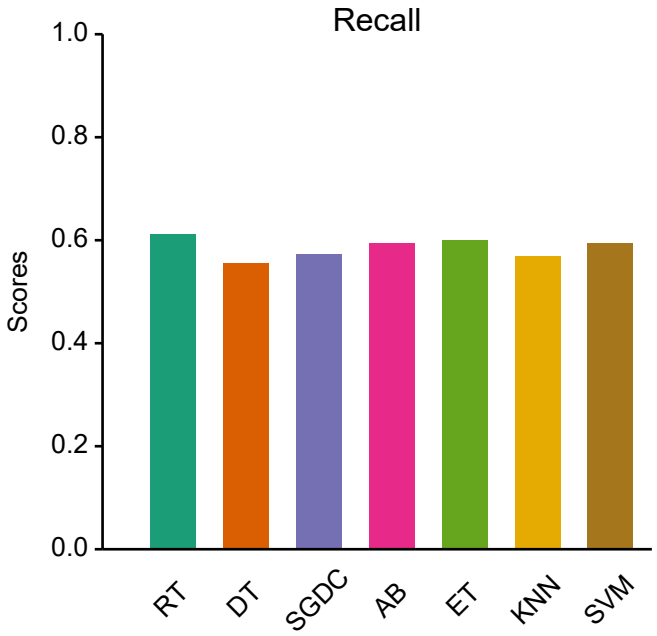
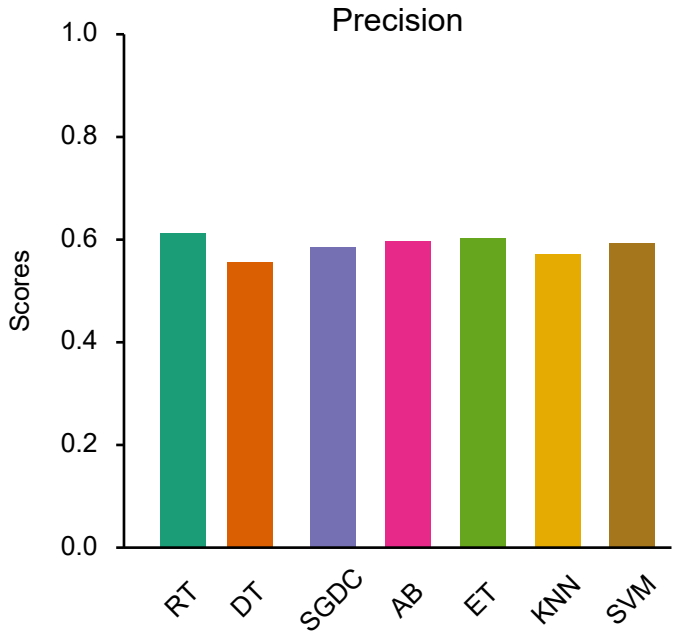


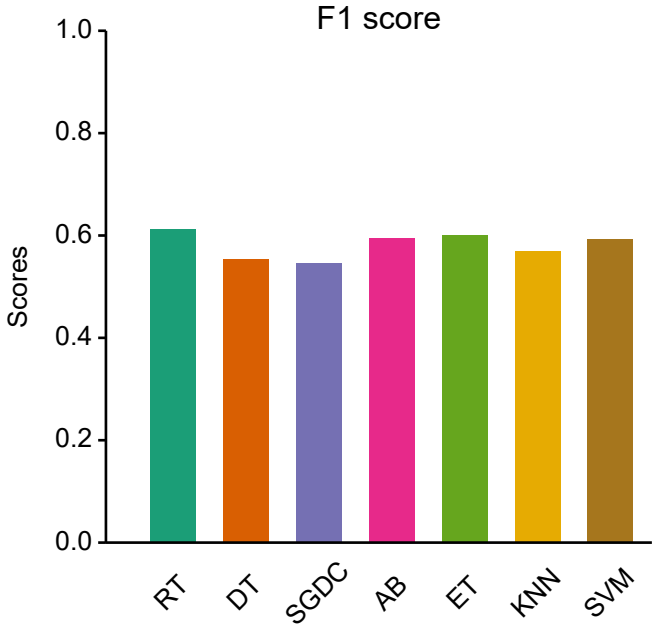
Figure 23.2:

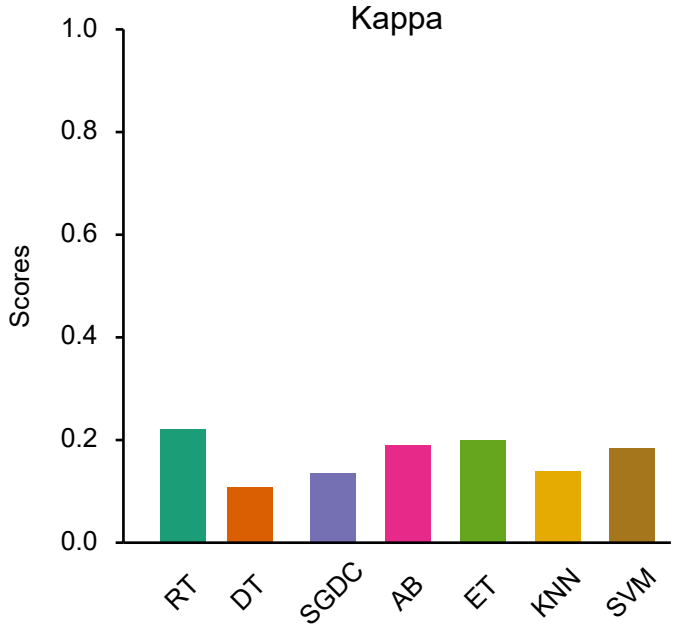
Model Evaluation (10 CV) based on IC50 - *mgt1* strain

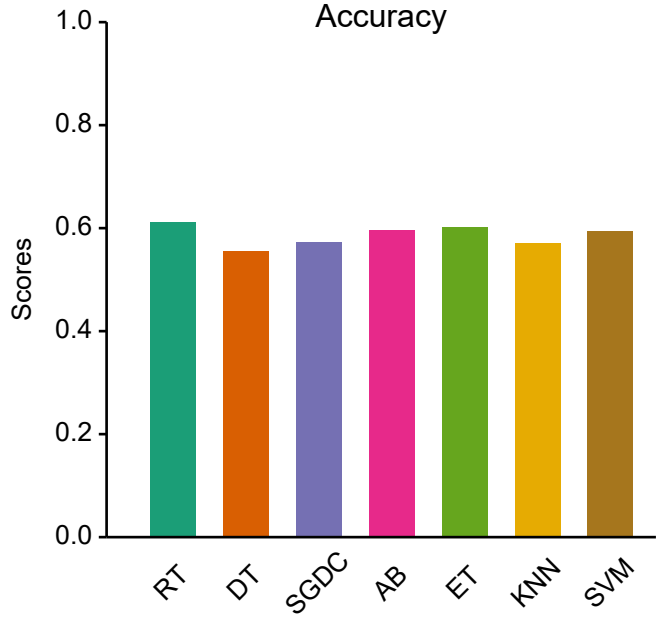












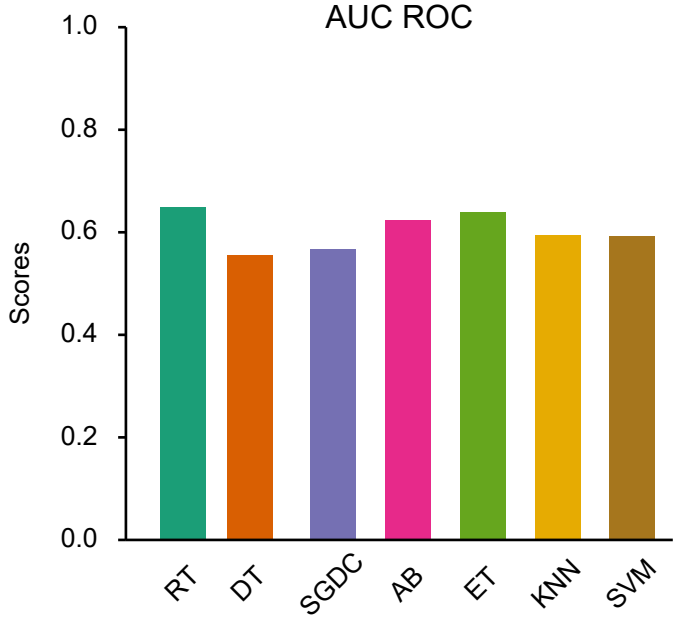
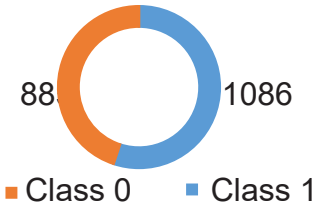
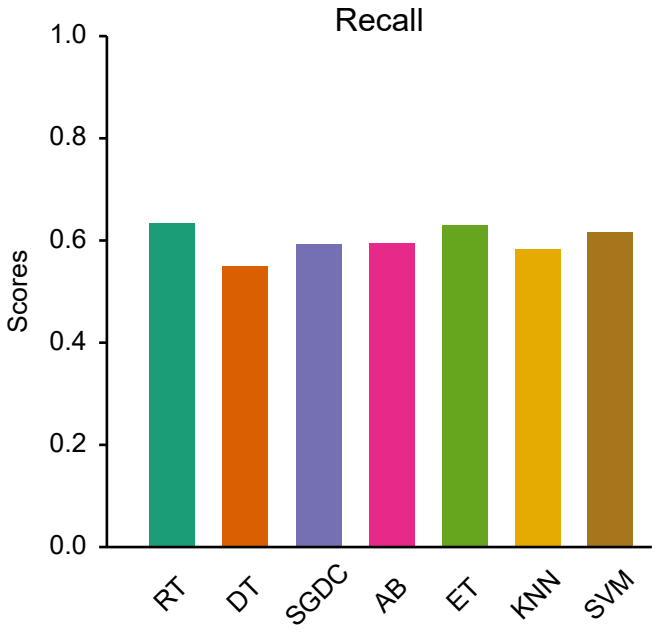
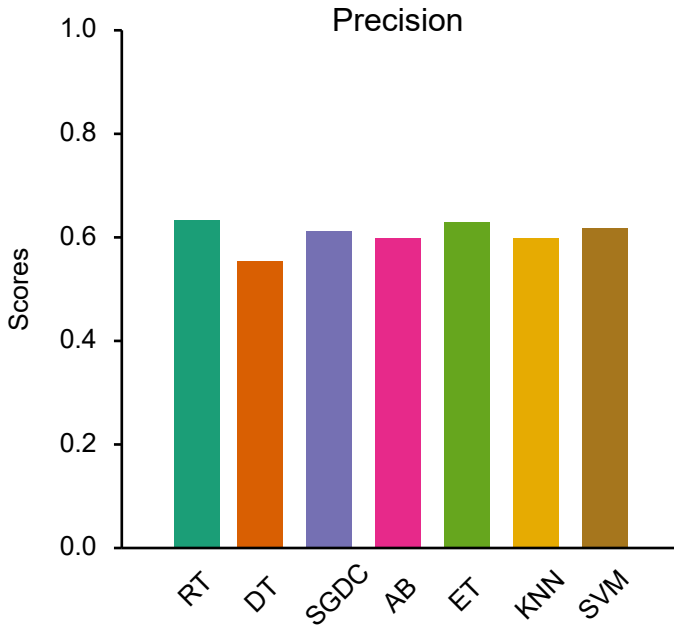


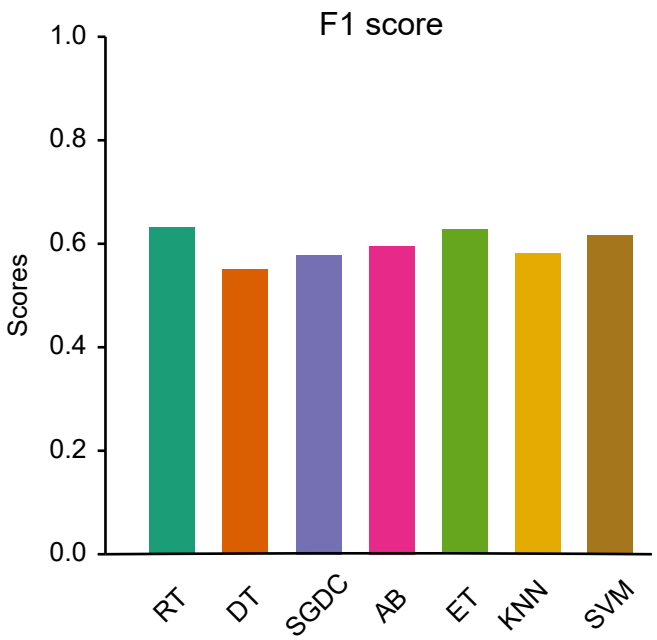
Figure 23.3:

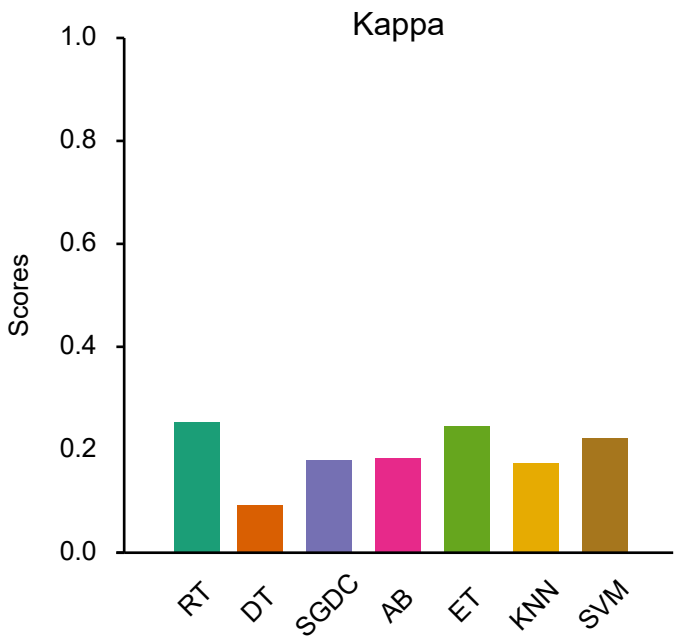
Model Evaluation (10 CV) based on IC50 - *rad14* strain

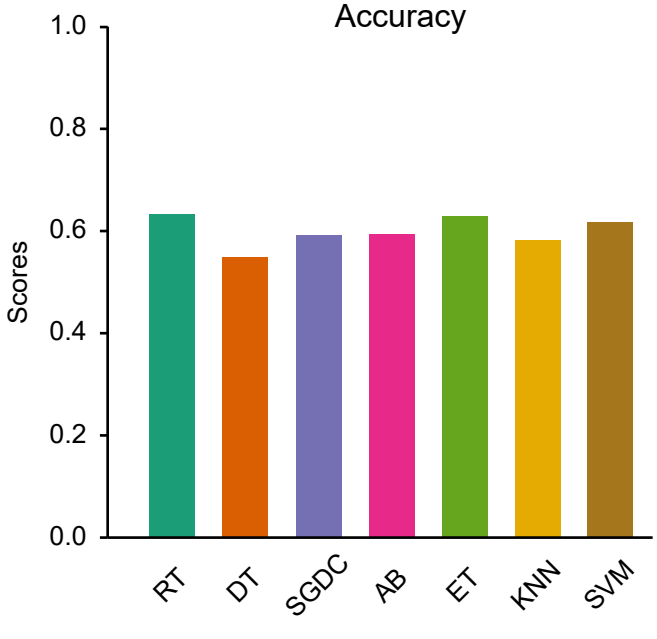












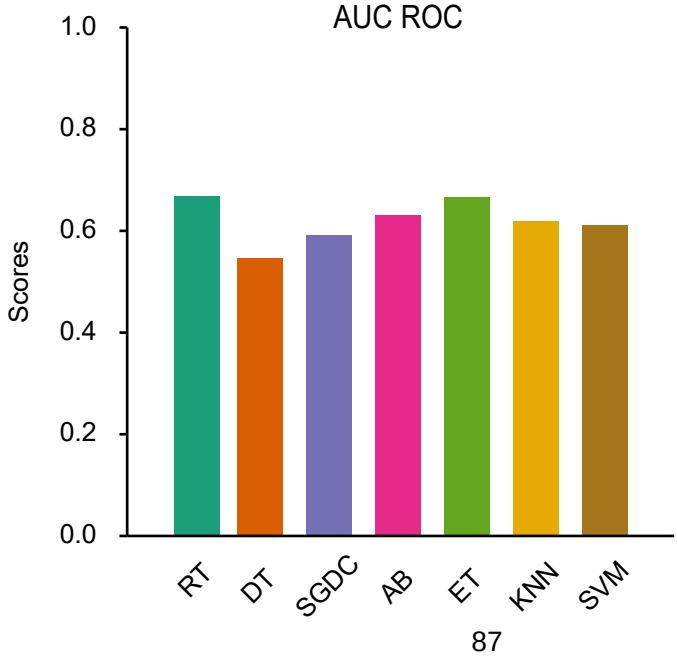
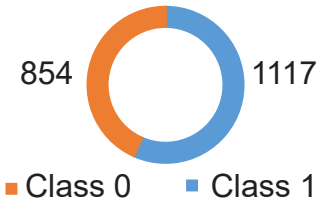
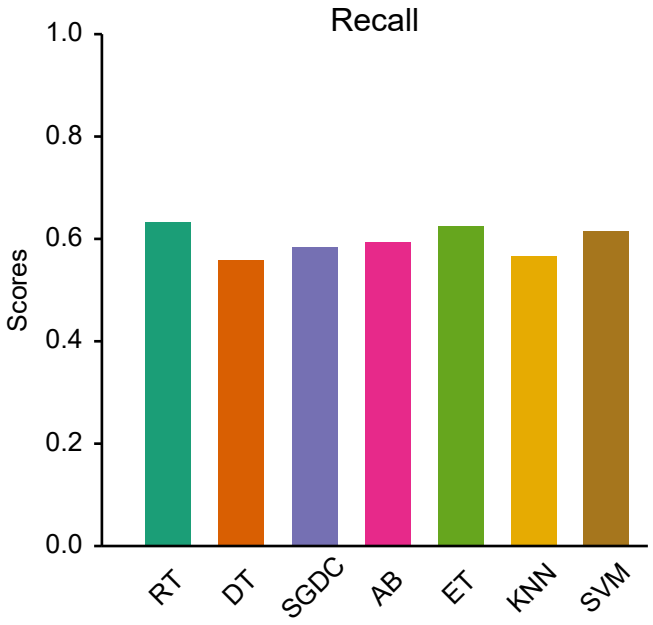
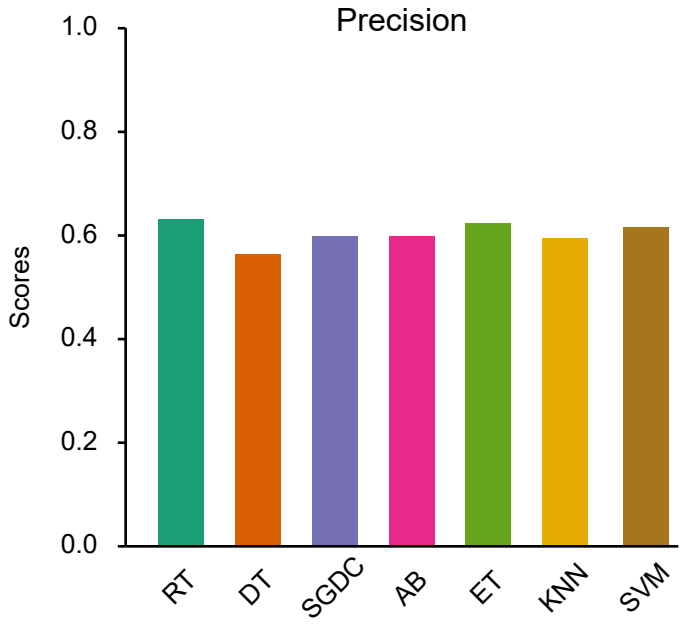


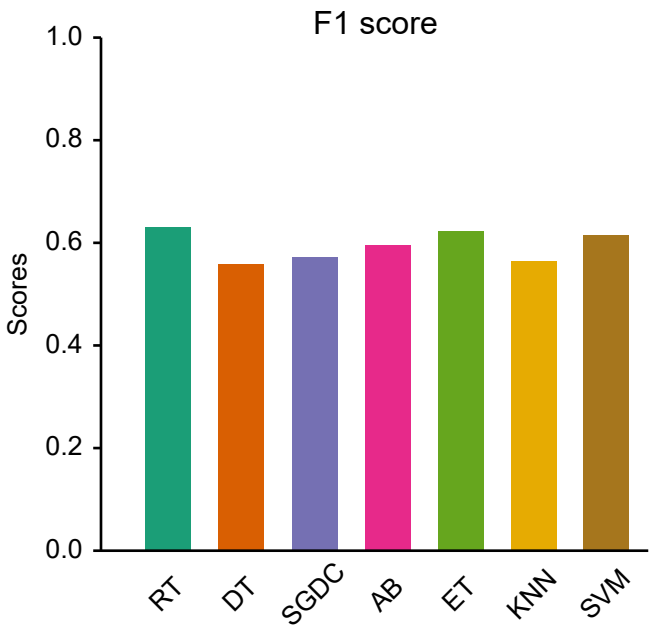
Figure 23.4:

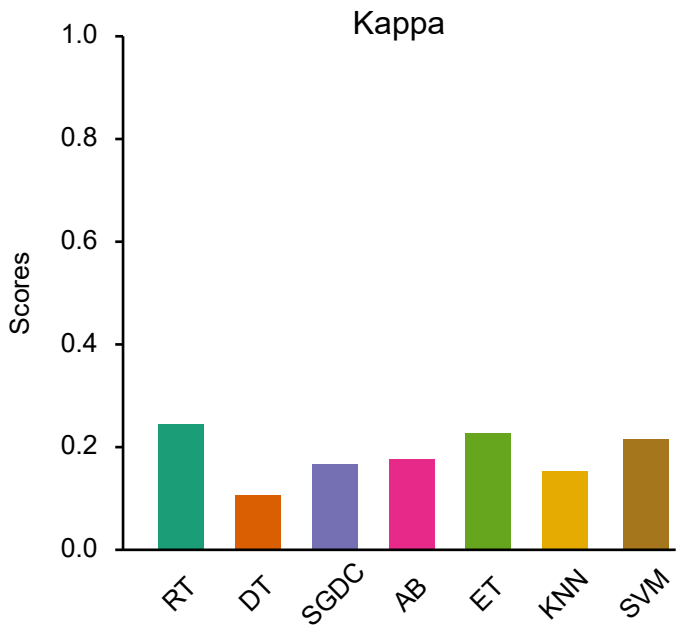
Model Evaluation (10 CV) based on IC50 - *rad18* strain

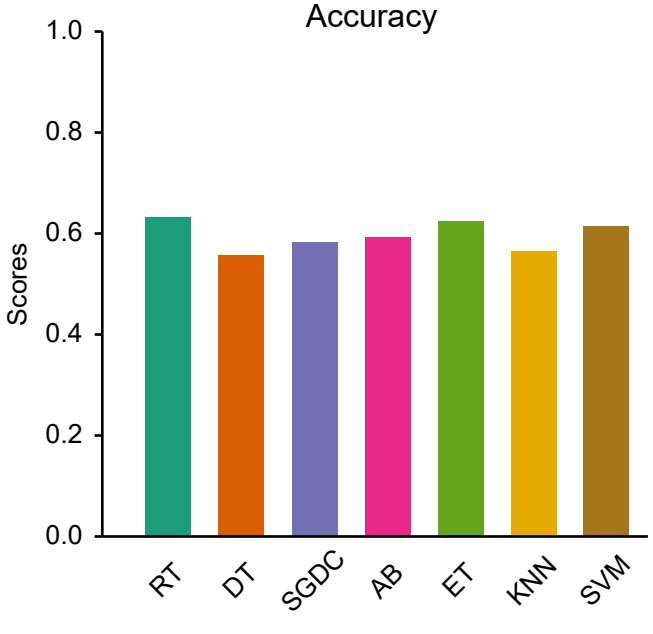












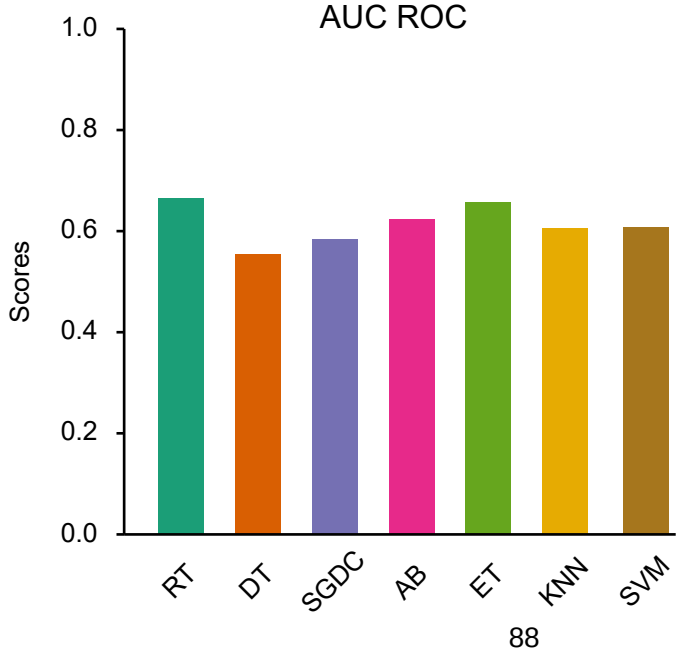
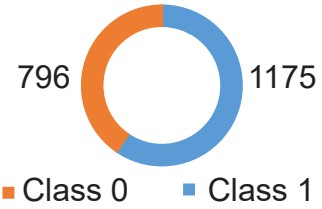
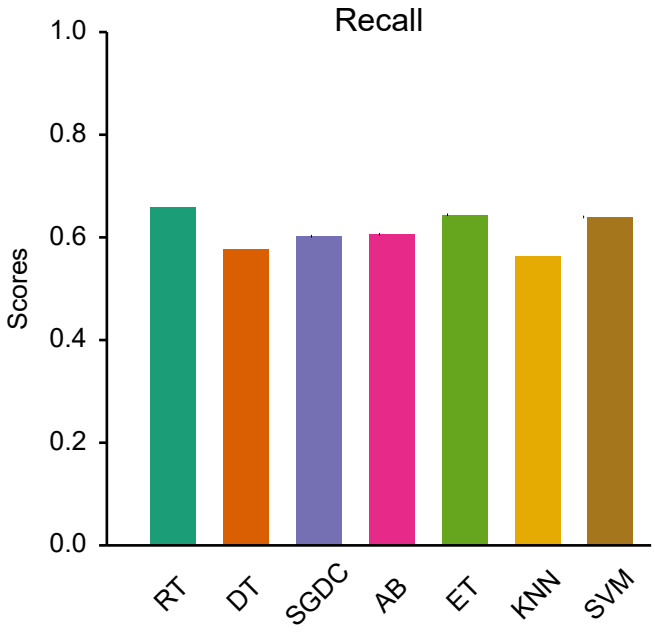
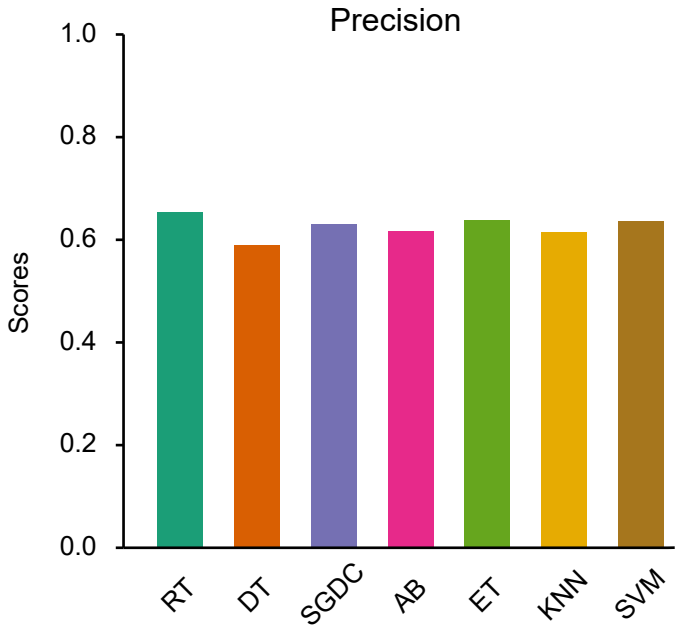


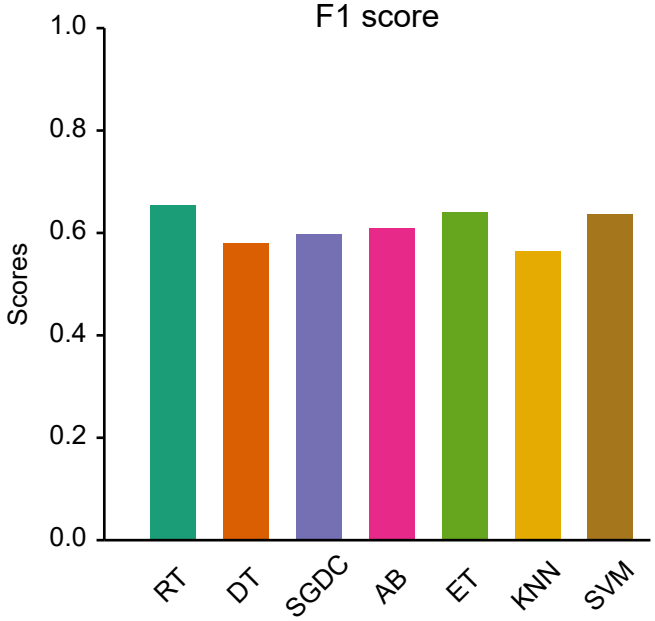
Figure 23.5:

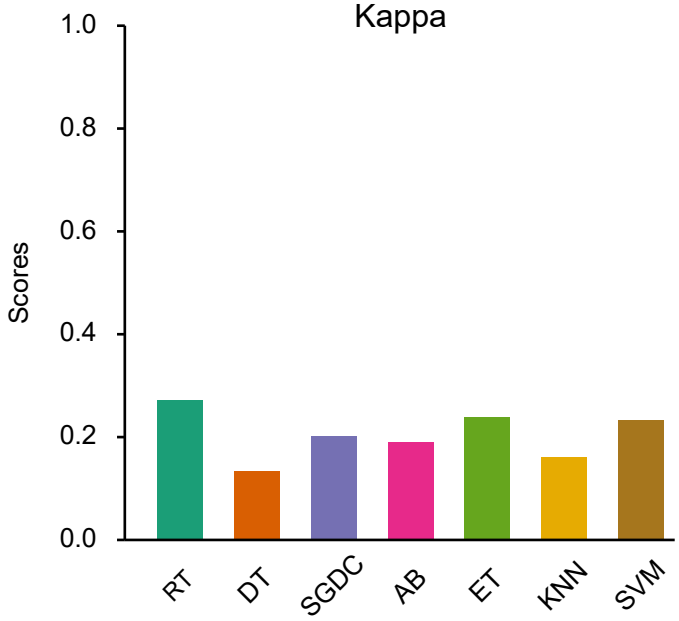
Model Evaluation (10 CV) based on IC50 - *rad50* strain

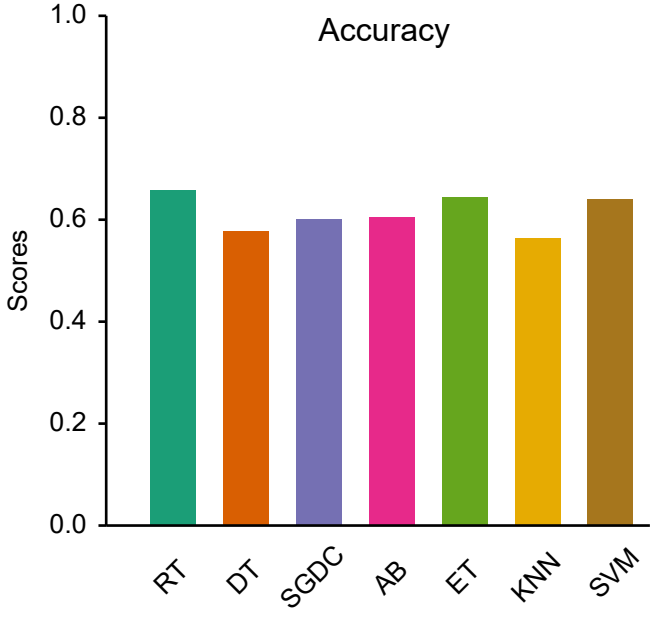












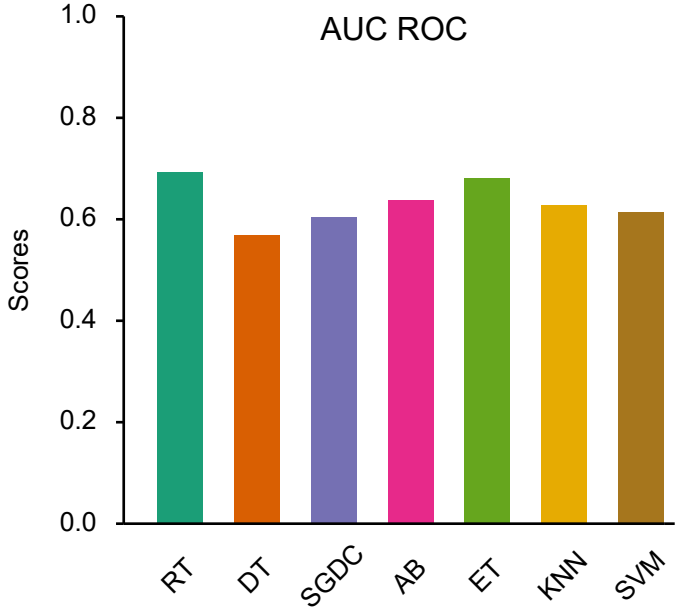
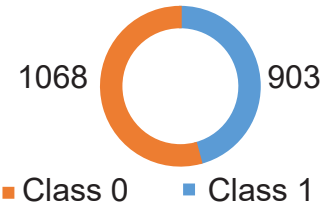
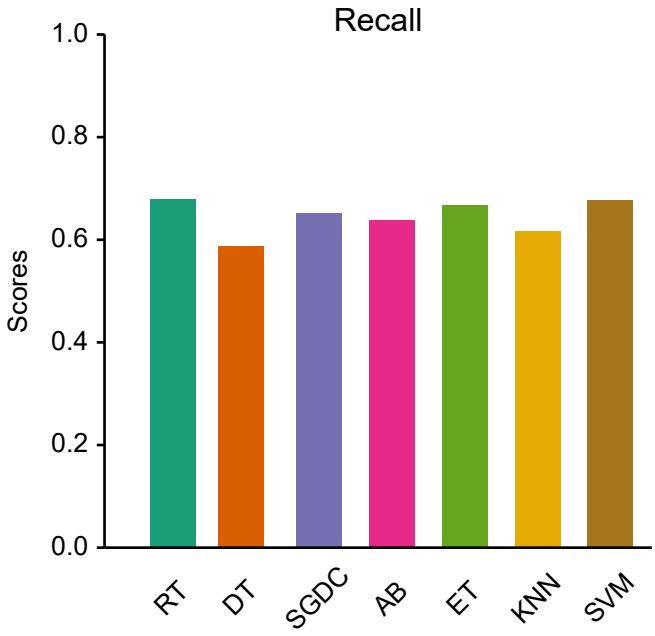
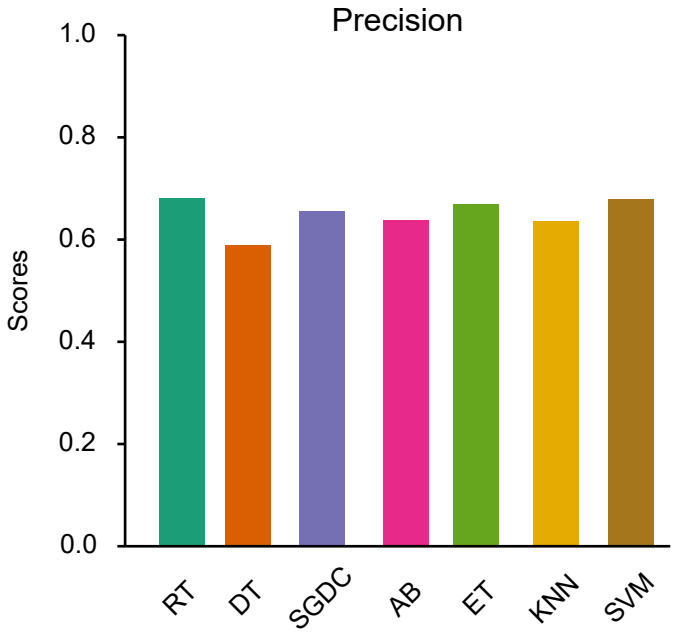


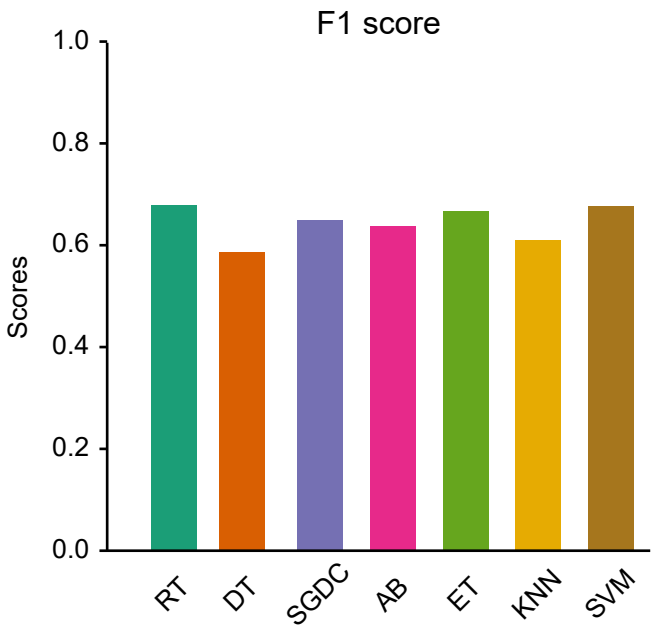
Figure 23.6:

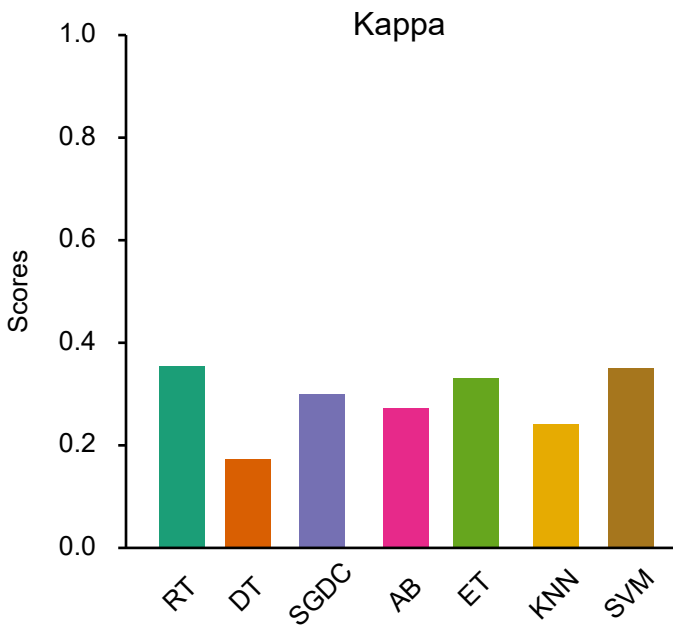
Model Evaluation (10 CV) based on IC50 - *rad50EPP*+ strain

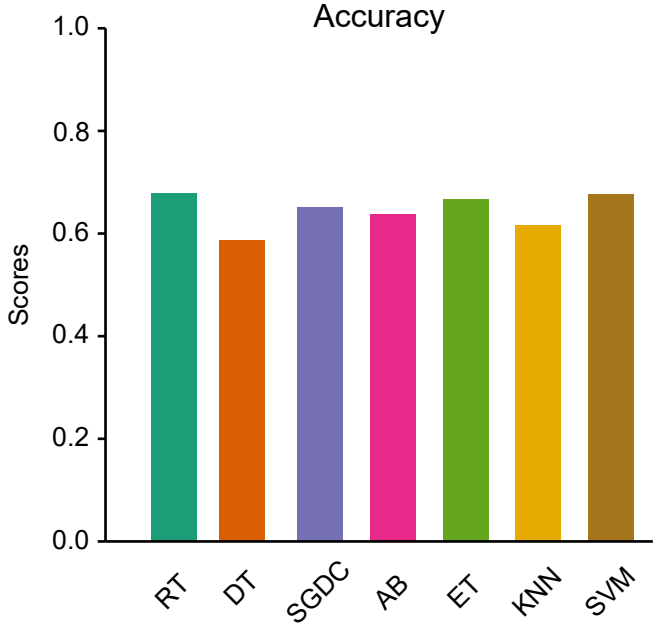












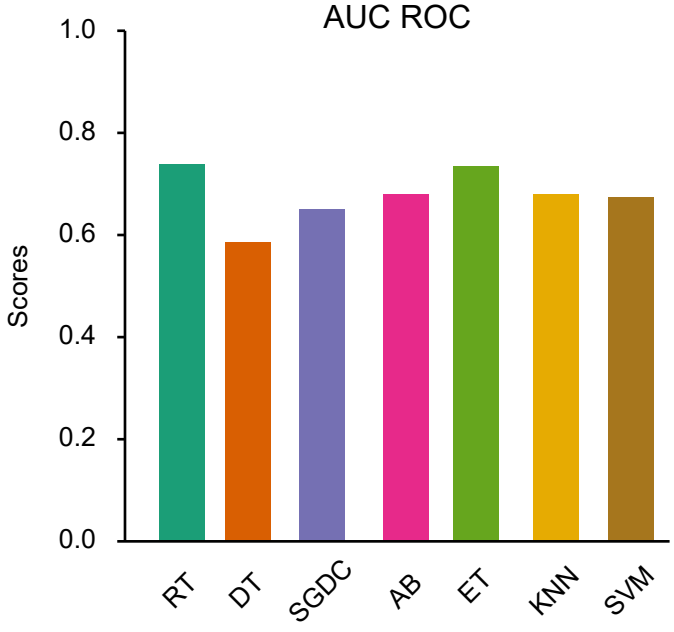


Figure 23.7:

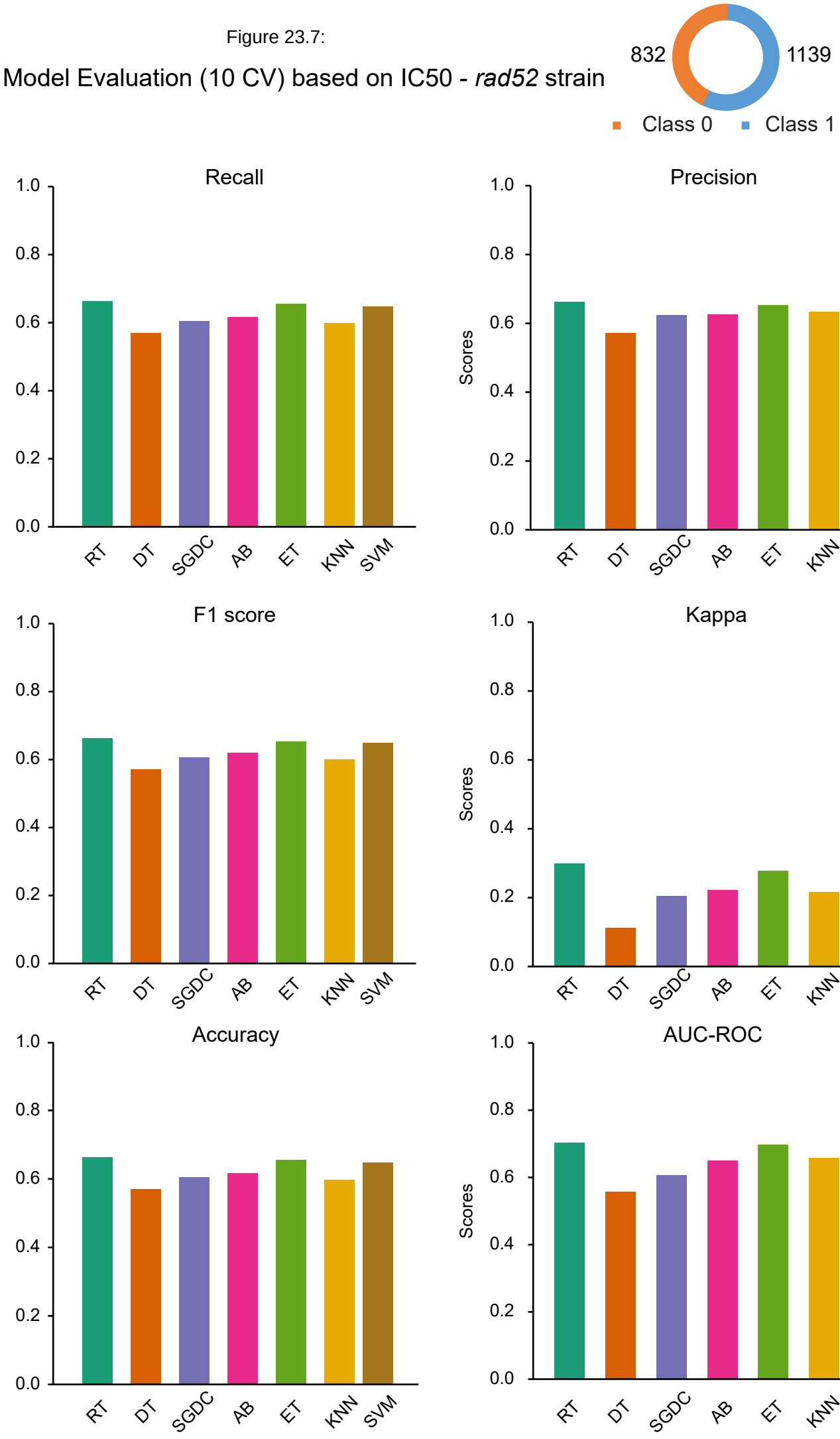
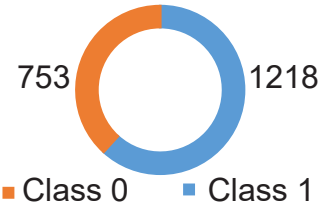
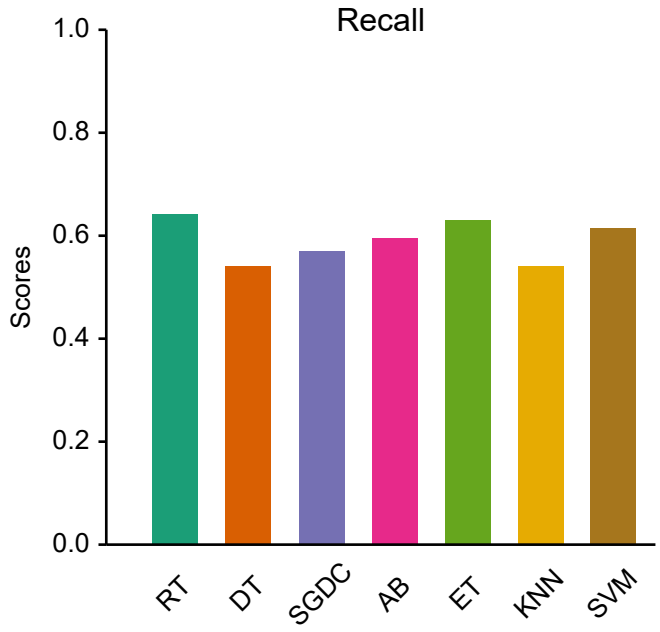
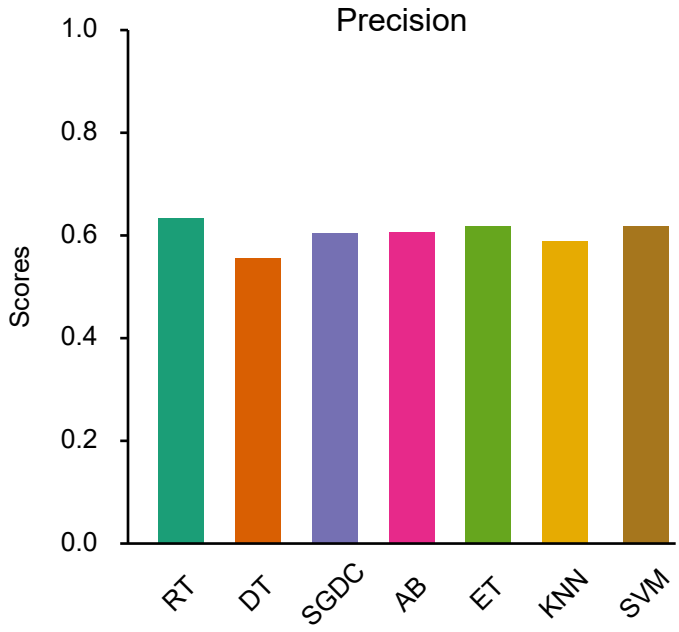

Model Evaluation (10 CV) based on IC50 - *rad52* strain
832
1139
Class 0
Class 1
Recall
Precision
F1 score
Kappa
Accuracy
AUC-ROC
Scores
1.0
0.8
0.6
0.4
0.2
0.0
RT
DT
SGDC
AB
ET
KNN
SVM

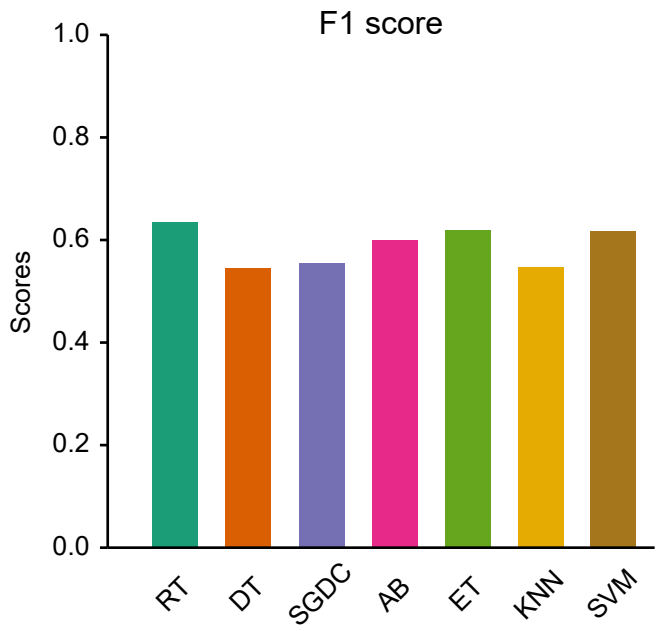
Figure 23.8:

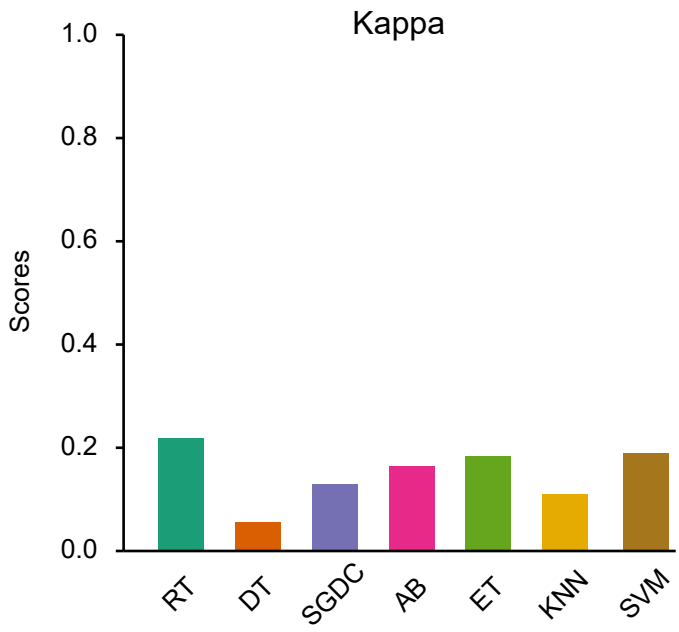
Model Evaluation (10 CV) based on IC50 - *sgs1* strain

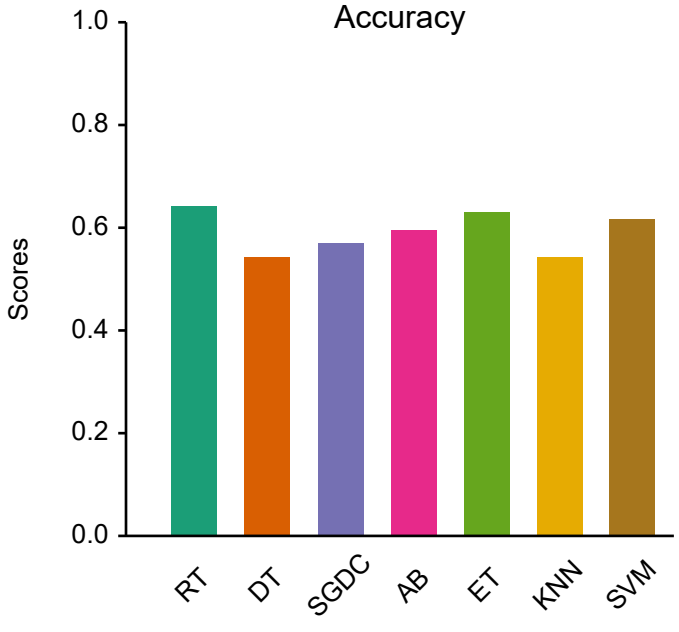












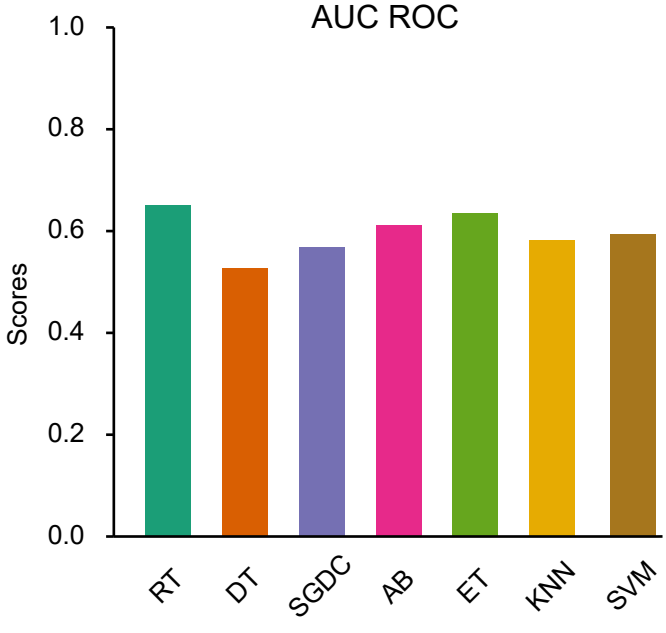


Figure 23.9:

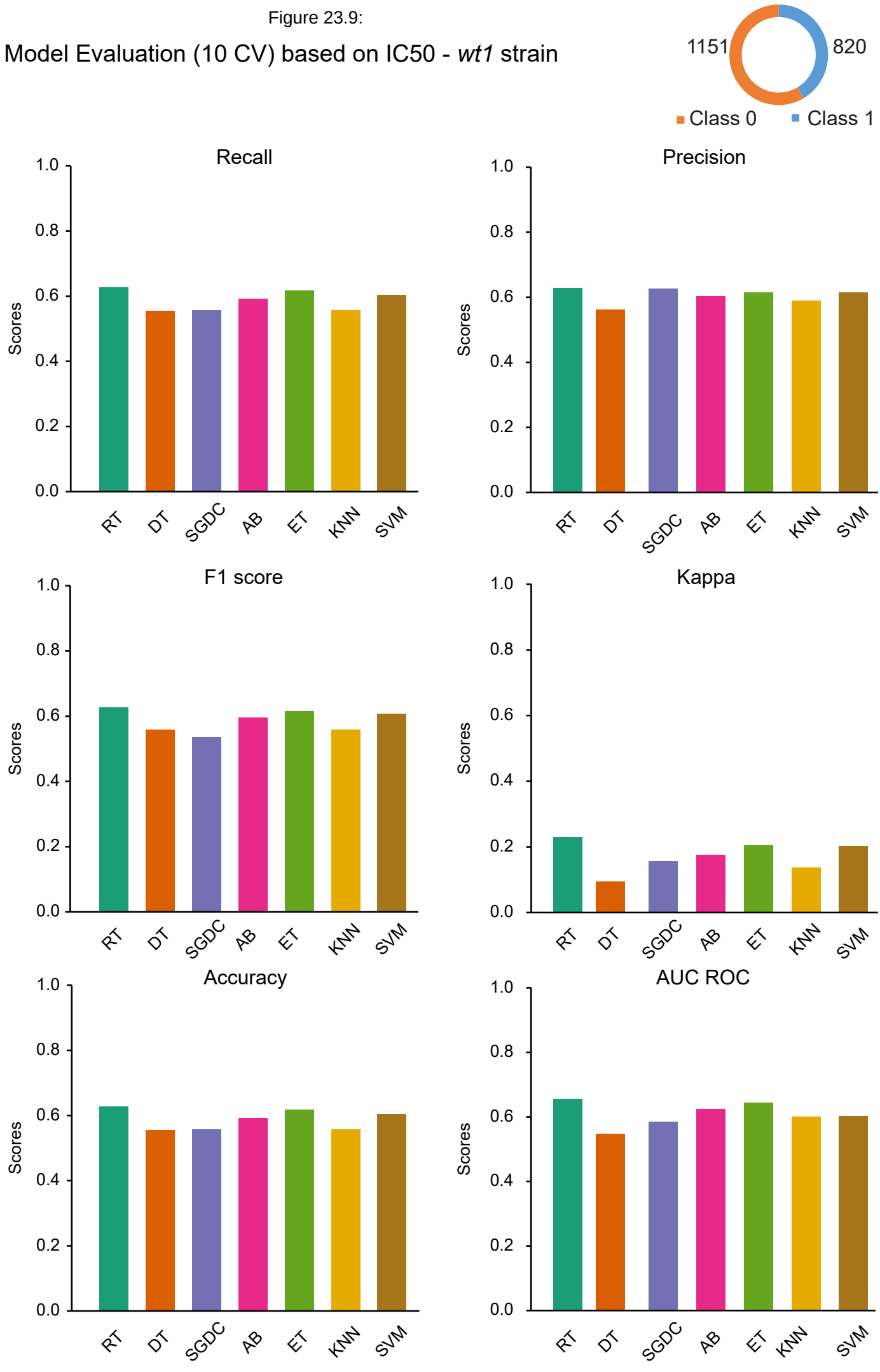
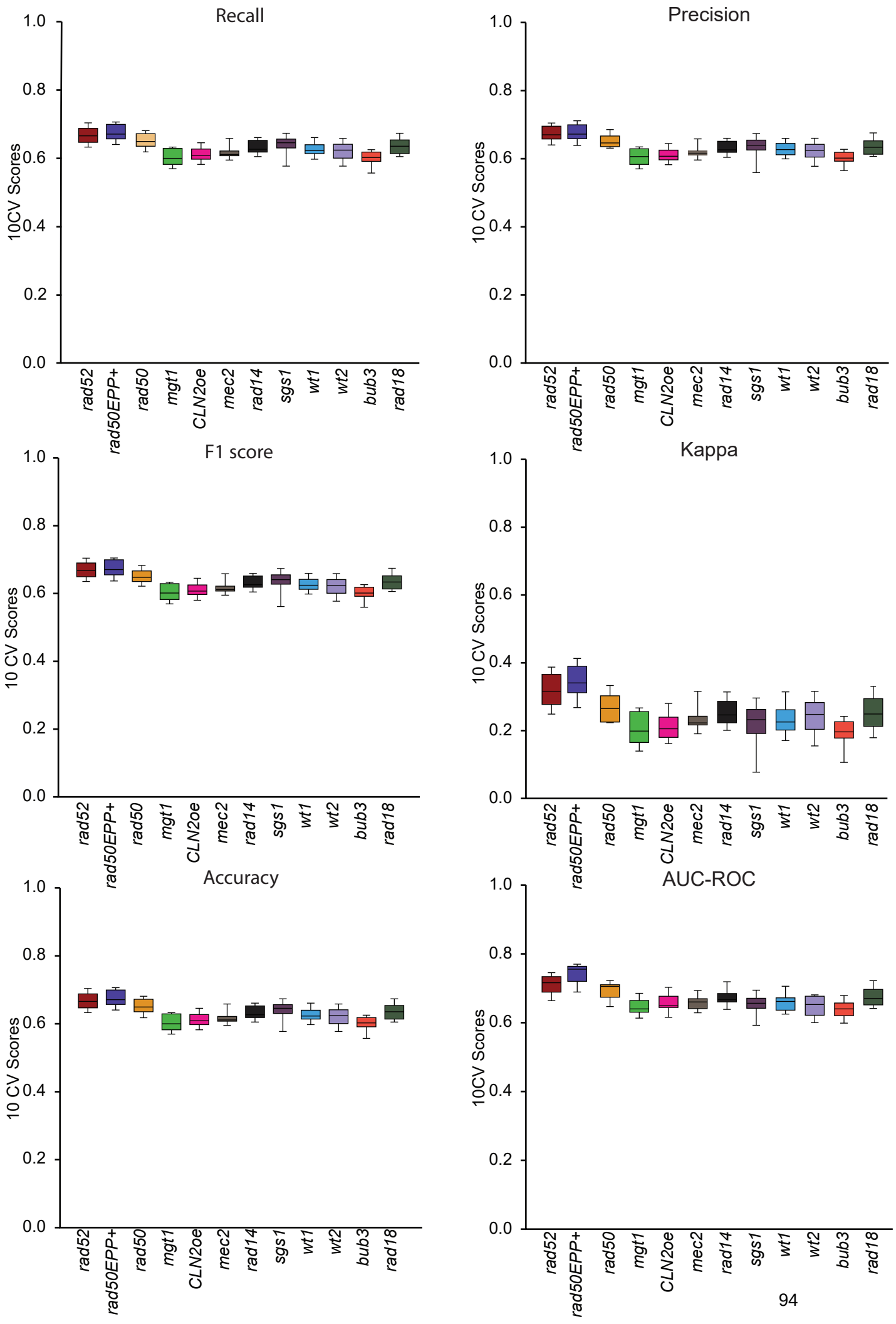

Model Evaluation (10 CV) based on IC50 - wt1 strain
1151
820
Class 0
Class 1
Recall
Precision
F1 score
Kappa
Accuracy
AUC ROC
Scores
1.0
0.8
0.6
0.4
0.2
0.0
RT
DT
SGDC
AB
ET
KNN
SVM

Figure 23.10: IC50 10 fold Cross Validation



V.5 - Machine Learning models in predicting novel antifungal agents:

The drug bank compounds of around 15000 were tested on the trained models based on GIPCRT and IC50 of *rad52* strain. Top 25 drugs of the class with growth inhibition from the GIPCRT trained model and top 25 drugs of the class with lower IC50 from the IC50 trained model are further studied. Top 25 drugs are chosen based on the probability value of the class.

15 drugs were found to be common between top25 drugs of both the models and the presence of relevant literature on their anticancer and antifungal properties is also checked and reported.

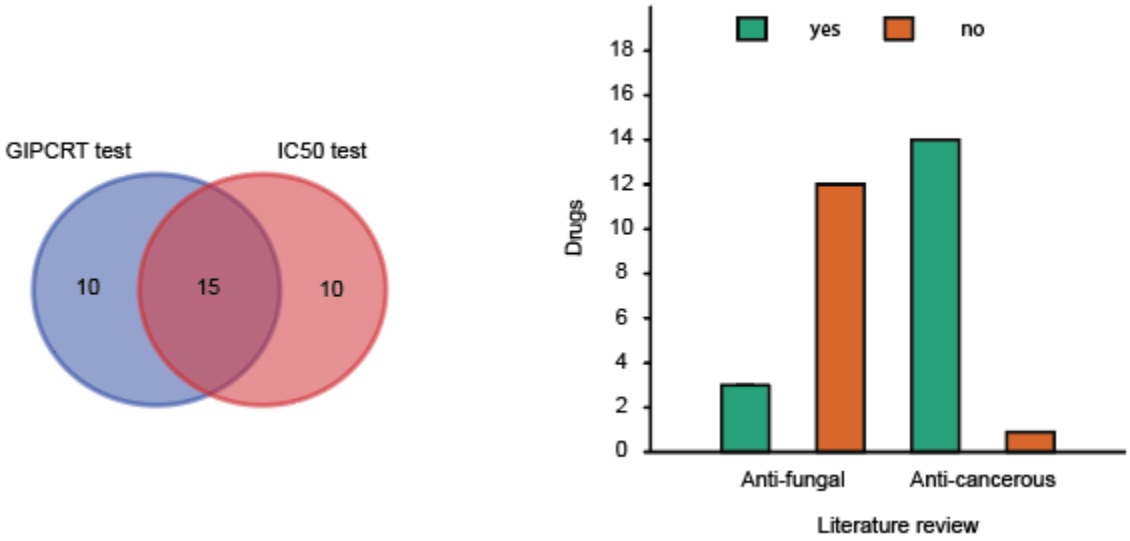


Figure 24: Venn diagram for the top 25 predicted compounds in GIPCRT & IC50 ML model and the literature survey on the common 15 drugs on its antifungal and anticancer properties

V.5.1 - GIPCRT based testing:

There was no overlap between the GIPCRT train and top25 tested drugs. 15 common drugs with their p values for the growth inhibition class is shown in the donut plot.

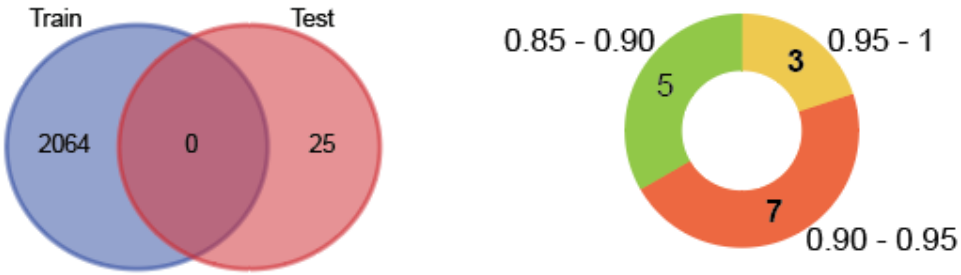


Figure 25: Venn diagram for the top 25 predicted compounds in GIPCRT & the NCI compounds used in training and the p values for the common 15 compounds

V.5.2 - IC50 based testing:

Two drugs- 1,10-Phenanthroline and Pyrazolanthrone were common between the IC50 train and top25 tested drugs. 15 common drugs with their p values for lesser IC50 are shown in the donut plot.

Figure 26: Venn diagram for the top 25 predicted compounds in IC50 & the NCI compounds used in training and the p values for the common 15 compounds

V.5.3 - Common 15 drugs:

Table 7: Literature review on top 15 common predicted compounds		
Drugs	**Literature availability on antifungal activity**	**Literature availability on anticancer activity**
9-aminocamptothecin	Not available	Burnouf et al[15]
Topotecan	Not available	Kang et al[16]
10-hydroxycamptothecin	Not available	Yang et al[17]
Namitecan	Not available	De Cesare et al[18]
Epirubicin	Not available	Khasraw et al[19]
Doxorubicin	Not available	Rivankar et al[20]
Mitoxantrone	Steverding et al[21]	Anderson et al[22], Evison et al[23]
Teniposide	Not available	Yan et al[24]
Idarubicin	Steverding et al[21]	Rafipour et al[25]
Zorubicin	Not available	Pignon et al[26]
Rubitecan	Not available	Patel et al[27]
Furvina	Allas et al[28]	Not available
Daunorubicin	Not available	Lancet et al[29]
DRF-1042	Not available	Chatterjee et al[30]
Camsirubicin	Not available	Song et al[31]

Among the common 15 drugs, only Furvina is not reported for their anticancer property yet. 3 drugs Furvina, Mitoxantrone and Idarubicin have been reported to have antifungal properties. The top 5 drugs 9-aminocamptothecin, Topotecan, 10-hydroxycamptothecin, Namitecan and Epirubicin should be tested for their antifungal properties. The combined model could be used to predict novel antifungal agents from their chemical space.

Conclusion

VI - Conclusion:

Analyzing the NCI60 growth inhibition data and NCI yeast drug screen study, we were able to relate the IC50 of 13 yeast strains with the IC50 of 60 human tumor cell lines via Mutual information score. We found that *rad52* mutant yeast strain could be used as a good substitute for 56 human tumor cell lines. Thus rad52 mutant strain could be used as a potential model for carrying out cancerous studies/ screens in the NCI60 cell lines. It saves time and money. Based on the growth inhibition pattern in yeast strains, we have built a machine learning model to predict potential antifungal agents. The machine learning classification model was based on the growth pattern and IC50 values in *rad52* mutant strain as it had better accuracy and kappa metrics. The model could predict antifungal compounds with lower IC50 values. The top hits from the classifier have to be evaluated experimentally for their antifungal activity.

VII - References:

1. *Yeast as a tool in cancer research*. (Springer, 2007).
2. Pray, L. LH Hartwell's yeast: a model organism for studying somatic mutations and cancer. *Nature Education*.
3. Khanna, K. K. & Shiloh, Y. *The DNA Damage Response: Implications on Cancer Formation and Treatment*. (Springer Science & Business Media, 2009).
4. Matuo, R. *et al.* Saccharomyces cerevisiae as a model system to study the response to anticancer agents. *Cancer Chemother. Pharmacol.* **70**, 491–502 (2012).
5. Holbeck, S. L. Update on NCI in vitro drug screen utilities. *Eur. J. Cancer* **40**, 785–793 (2004).
6. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813–823 (2006).
7. Fleming, N. How artificial intelligence is changing drug discovery. *Nature* **557**, S55–S57 (2018).
8. Gao, A., Kouznetsova, V. L. & Tsigelny, I. F. Machine-learning-based virtual screening to repurpose drugs for treatment of Candida albicans infection. *Mycoses* **65**, 794–805 (2022).
9. Joshi, T., Pundir, H. & Chandra, S. Deep-learning based repurposing of FDA-approved drugs against Candida albicans dihydrofolate reductase and molecular dynamics study. *J. Biomol. Struct. Dyn.* **40**, 8420–8436 (2022).
10. Liu, Z. *et al.* DeepScreening: a deep learning-based screening web server for accelerating drug discovery. *Database* **2019**, (2019).
11. Chen, X. *et al.* NLLSS: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning. *PLoS Comput. Biol.* **12**, e1004975 (2016).
12. Kuenzi, B. M. *et al.* Predicting Drug Response and Synergy Using a Deep Learning Model

of Human Cancer Cells. *Cancer Cell* **38**, 672–684.e6 (2020).

13. Kadurin, A. *et al.* The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* **8**, 10883–10890 (2017).
14. Bertoni, M. *et al.* Bioactivity descriptors for uncharacterized chemical compounds. *Nat. Commun.* **12**, 3932 (2021).
15. Burnouf, P.-A. *et al.* Reversible glycosidic switch for secure delivery of molecular nanocargos. *Nat. Commun.* **9**, 1843 (2018).
16. Kang, J.-H. *et al.* A randomised phase 2b study comparing the efficacy and safety of belotecan vs. topotecan as monotherapy for sensitive-relapsed small-cell lung cancer. *Br. J. Cancer* **124**, 713–720 (2021).
17. Yang, C. *et al.* Discovery of highly potent and selective 7-ethyl-10-hydroxycamptothecin-glucose conjugates as potential anti-colorectal cancer agents. *Front. Pharmacol.* **13**, 1014854 (2022).
18. De Cesare, M. *et al.* Synergistic antitumor activity of cetuximab and namitecan in human squamous cell carcinoma models relies on cooperative inhibition of EGFR expression and depends on high EGFR gene copy number. *Clin. Cancer Res.* **20**, 995–1006 (2014).
19. Khasraw, M., Bell, R. & Dang, C. Epirubicin: is it like doxorubicin in breast cancer? A clinical review. *Breast* **21**, 142–149 (2012).
20. Rivankar, S. An overview of doxorubicin formulations in cancer therapy. *J. Cancer Res. Ther.* **10**, 853–858 (2014).
21. Steverding, D. *et al.* In vitro antifungal activity of DNA topoisomerase inhibitors. *Med. Mycol.* **50**, 333–336 (2012).
22. Anderson, R. *et al.* Phase II trial of cytarabine and mitoxantrone with devimistat in acute myeloid leukemia. *Nat. Commun.* **13**, 1673 (2022).
23. Evison, B. J., Sleebs, B. E., Watson, K. G., Phillips, D. R. & Cutts, S. M. Mitoxantrone, More

than Just Another Topoisomerase II Poison. *Med. Res. Rev.* **36**, 248–299 (2016).

24. Yan, J., Sun, J. & Zeng, Z. Teniposide ameliorates bone cancer nociception in rats via the P2X7 receptor. *Inflammopharmacology* **26**, 395–402 (2018).
25. Rafipour, R., Mousavi, A. & Mansouri, K. Apoferritin nanocages for targeted delivery of idarubicin against breast cancer cells. *Biotechnol. Appl. Biochem.* **69**, 1061–1067 (2022).
26. Pignon, B. *et al.* Treatment of acute myelogenous leukaemia in patients aged 50-65: idarubicin is more effective than zorubicin for remission induction and prolonged disease-free survival can be obtained using a unique consolidation course. The Goelam Group. *Br. J. Haematol.* **94**, 333–341 (1996).
27. Patel, H. *et al.* Phase II study of rubitecan, an oral camptothecin in patients with advanced colorectal cancer who have failed previous 5-fluorouracil based chemotherapy. *Invest. New Drugs* **24**, 359–363 (2006).
28. Allas, U. L. *et al.* Antibacterial activity of the nitrovinylfuran G1 (Furvina) and its conversion products. *Sci. Rep.* **6**, 36844 (2016).
29. Lancet, J. E. *et al.* CPX-351 (cytarabine and daunorubicin) Liposome for Injection Versus Conventional Cytarabine Plus Daunorubicin in Older Patients With Newly Diagnosed Secondary Acute Myeloid Leukemia. *J. Clin. Oncol.* **36**, 2684–2692 (2018).
30. Chatterjee, A. *et al.* Safety, tolerability, and pharmacokinetics of a capsule formulation of DRF-1042, a novel camptothecin analog, in refractory cancer patients in a bridging phase I study. *J. Clin. Pharmacol.* **45**, 453–460 (2005).
31. Song, X. *et al.* Pharmacologic Suppression of B7-H4 Glycosylation Restores Antitumor Immunity in Immune-Cold Breast Cancers. *Cancer Discov.* **10**, 1872–1893 (2020).