

Dr.S
Drug recommendation system (for cell lines)

David Craft, Ph.D.
Massachusetts General Hospital

Oct 25, 2019 ML-MSM
NIH, Bethesda MD

Vision:

A patient walks into the clinic...

... a treatment is prescribed

Dataset : GDSC

~1000 cell
lines

42 drugs
chosen

Genomics of Drug Sensitivity in Cancer

We have characterised **1000 human cancer cell lines** and screened them with **100s of compounds**.
On this website, you will find **drug response data** and **genomic markers** of sensitivity.

Search by drug, gene or cell line name
e.g. Docetaxel, RP-56976, BRAF, COLO-829

Overview

Coverage
453 compounds targeting 24 pathways

Pathway	Coverage
Other, kinases	52
Other	52
PI3K/MTOR signaling	47
RTK signaling	42
DNA replication	25
ERK MAPK signaling	23
Apoptosis regulation	22
Cell cycle	22
Mitosis	21
Chromatin histone acetylation	18
Genome integrity	15
WNT signaling	14
Protein stability and degradation	12
Chromatin other	12
Metabolism	11
EGFR signaling	11
Cytoskeleton	10
Unclassified	8

385,712
dose-response curves

100 100

What's new?
Release 8.1 (Oct 2019)
The GDSC database now contains data for an extra **187 drugs** and more than **160,000 new IC50s** - an increase of ~70% compared to release 7.0. Two datasets are available: GDSC1 updates our previous screening results; and GDSC2 uses the latest improved screening technology.

Datasets

	GDSC1	GDSC2
Age		
from 2010 to 2015		✓ NEW
Size		
987 Cell lines		809 Cell lines
320 Compounds		175 Compounds
267284 IC50s		118428 IC50s
Assay		
Resazurin or Syto60		CellTitreGlo
Duration		
72 hours		72 hours

Key Publications

Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.
Yang *et al.*, (2013) Nucl. Acids Res. 41 (Database issue): D955 - D961. (PMID:23180760 [↗](#))

A landscape of pharmacogenomic interactions in cancer
lorio *et al.*, (2016). Cell, Volume 166, Issue 3, 740 - 754 (PMID:27397505 [↗](#))

Systematic identification of genomic markers of drug sensitivity in cancer cells
Garrett *et al.*, (2012) Nature volume 483, pages 570 - 575 (PMID:27397505 [↗](#))

Home | **Compounds** | **Features** | **Cell Lines** | **About** | **News** | **Downloads** | **Documentation** | **FAQ** | **Login**

Dataset :
CCLE

~500 cell
lines

radiation


Home About Data Contact Search Genes/Cell Lines Search Logout: dcraft

CCLE Cancer Cell Line Encyclopedia

Statistics Papers How to Use Terms of Access Contact Us

Statistics

Cell Lines: 1457

 Open

Genes: 84,434

Unique Data Sets: 136,488

Mutation Entries: 1,159,663

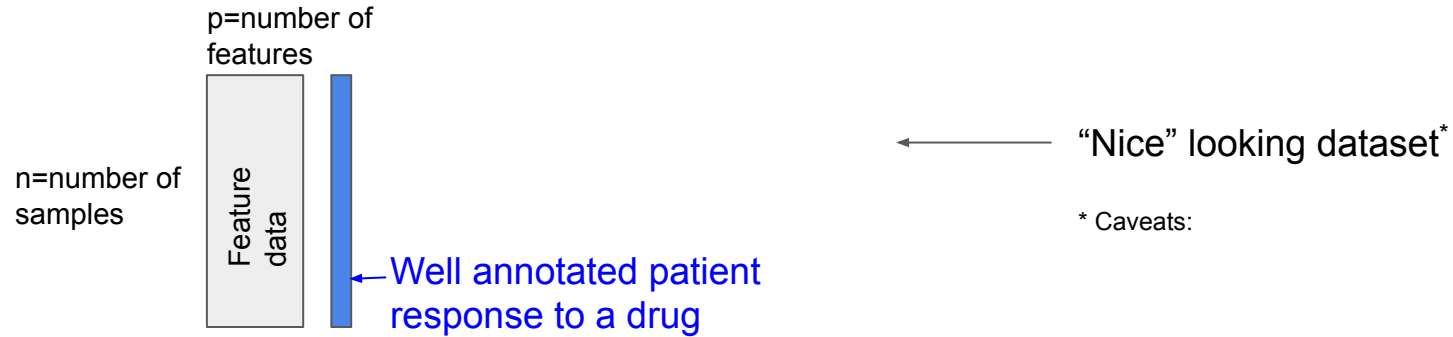
Distribution Scores: 118,661,636

Methylation Scores: 411,948,577

Number of samples, number of features



What does the GDSC dataset look like?



What we actually have:

p=number of features

n=number of cell lines

Feature data

Feature selection in this situation

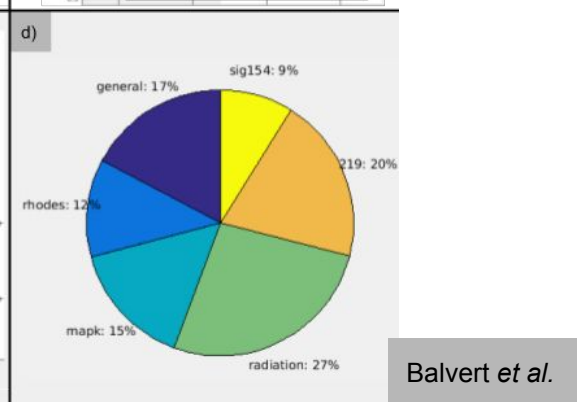
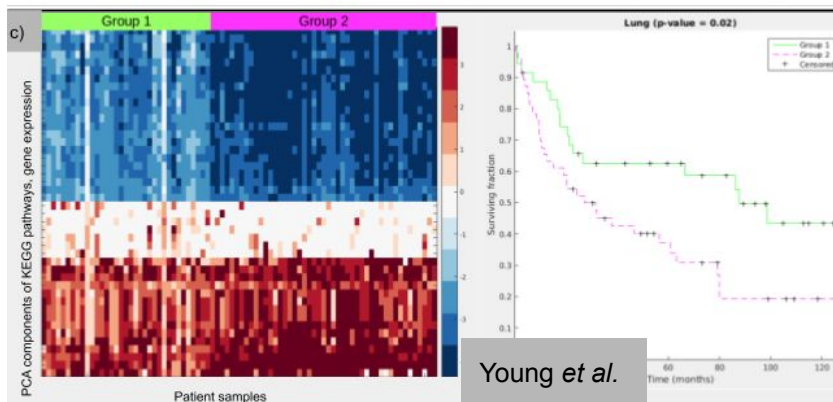
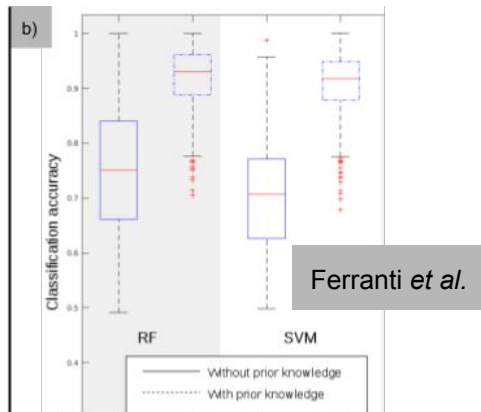
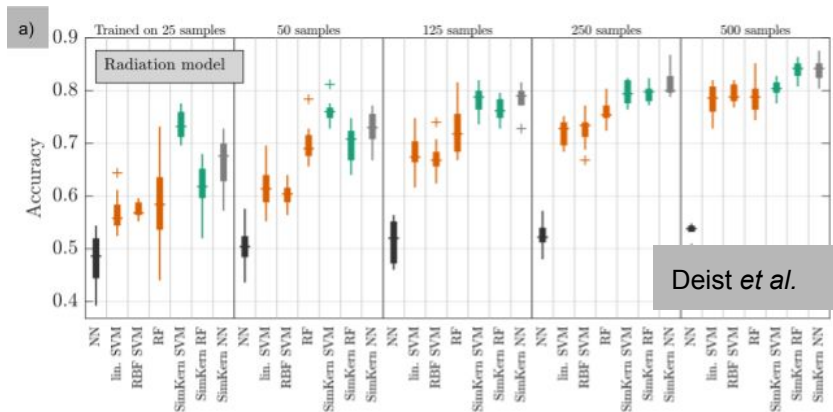
p =number of features

n =number of
samples



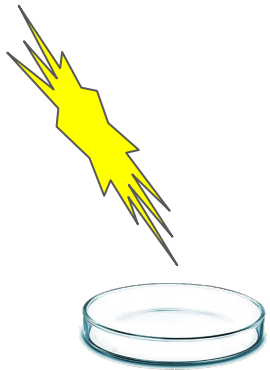
Feature
data

Some demonstrations of the value of prior knowledge:

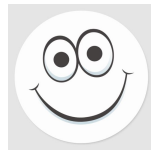


The “Radiation” Gene List

Question: if cancer cells are irradiated, what genes/proteins are most important for dictating the fate of the cells?



Approach: Two scientists, independent brainstorming + an automated pubmed search:



Pubmed search

Radiation gene lists: 263 genes

MESH terms used in pubMed search for radiation gene list

Radiation Tolerance

Radiation, Ionizing

Radiation-Protective Agents

Cell Death/radiation effects

Apoptosis/radiation effects

dna damage/radiation effects

Cellular Senescence/radiation effects

Chromosome Aberrations/radiation effects

Bystander Effect/radiation effects

Autophagy/radiation effects

Cell Cycle/radiation effects

Reactive Oxygen Species/radiation effects

Oxidative Stress/radiation effects

Metabolism/radiation effects

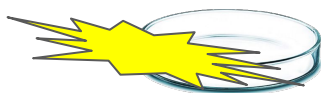
Transcription, Genetic/radiation effects

Stem Cells/radiation effects

Telomere/radiation effects

Chromatin/radiation effects

The "Radiation" Gene List →

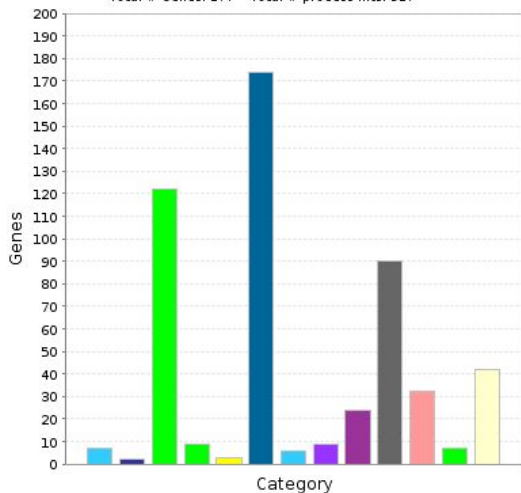


Brainstorm	p53 and related	senescence	ROS, thioredoxin, etc	Necroptosis	Stem cell regulators
EMT	Telomerase	Bystander contributors	Chromatin and Epigenetics	Metabolism	DNA Damage
Mitosis and G2M	Important Kinases	Syndromes			

Select Ontology: **Biological Process** View: **100%**

PANTHER GO-Slim Biological Process

Total # Genes: 277 Total # process hits: 527



Click to get gene list for a category:

- [biological adhesion \(GO:0022610\)](#)
- [biological phase \(GO:0044848\)](#)
- [biological regulation \(GO:0065007\)](#)
- [cell proliferation \(GO:0008283\)](#)
- [cellular component organization or biogenesis \(GO:0071840\)](#)
- [cellular process \(GO:0009987\)](#)
- [developmental process \(GO:0032502\)](#)
- [immune system process \(GO:0002376\)](#)
- [localization \(GO:0051179\)](#)
- [metabolic process \(GO:0008152\)](#)
- [multicellular organismal process \(GO:0032501\)](#)
- [reproduction \(GO:0000003\)](#)
- [response to stimulus \(GO:0050896\)](#)

Color picker powered by



ABL1	AKT1	ALK	APAF1	APC	AR
ATM	ATP13A1	ATP13A2	ATP2C1	ATP2C2	ATR
AURKA	AURKB	AURKC	BAD	BAP1	BAX
BBC3	BCL2	BCL2L1	BCL2L11	BECN1	BID
BIRC2	BIRC3	BIRC5	BLM	BM1	BMPRI1A
BMPRI1B	BMPR2	BRAF	BRCA1	BRCA2	BRIP1
BUB1	CAT	CCNB1	CCND1	CCND3	CDC25C
CDH1	CDK1	CDK2	CDK4	CDK6	CDKN1A
CDKN1B	CDKN2A	CDKN2A-DT	CDKN2B	CHEK1	CHEK2
CNNM1	CREB1	CTNNB1	DCLRE1C	DD2	DKC1
DLX2	DNM1	DRAM1	E2F1	EGFR	EP300
EPAS1	ERBB2	ERCC5	ERCC6	ESR1	EXO1
FADD	FANCD2	FAS	FGFR1	FGFR2	FGFR3
FN1	FZD1	G6PD	GABPA	GABPB1	GABPB2
GADD45A	GJA1	GJB1	GJB2	GLI1	GLS
GLUD1	GOT1	GRB2	GSX1	GSX2	H2AFX
HDAC1	HIF1A	HIPK2	HIST1H2BC	HRAS	HSP90AA1
IDH1	IDH2	IGFBP3	IL6	IL6R	IL6ST
INSR	IRF1	JAK1	JAK2	JUN	KLF4
KMT2C	KRAS	LEF1	LIG4	LSP1	MAP1LC3A
MAP2K7	MAPK1	MAPK14	MAPK3	MAPK8	MAX
MDC1	MDM2	MGMT	MLH1	MRE11	MSH2
MSH3	MSH6	MTOR	MYC	MYCN	NBN
NCOA4	NEDD4L	NFKB1	NFKB2	NHEJ1	NOS1
NOS2	NOS3	NOTCH1	NOTCH2	NOTCH3	NOTCH4
NRAS	P2RX4	PALB2	PARP1	PARP2	PGAP3
PGR	PIK3CA	PIK3CB	PLK1	PLK2	PLK3
PMAIP1	PPM1D	PRDX1	PRDX2	PRDX4	PRDX6
PTEN	PTGS2	RAD18	RAD50	RAD51	RARA
RB1	RBBP8	RECQL4	RELA	RIF1	RIPK1
RNF168	RPRM	RTFL1	SDHA	SDHB	SERPINE1
SFN	SHH	SIAH1	SIAH2	SLC11A1	SLC11A2
SLC25A39	SLC30A10	SLC31A1	SLC34A2	SLC34A3	SLC36A1
SLC39A14	SLC39A8	SLC6A3	SMAD2	SMAD3	SMAD4
SMO	SNAI1	SNRPF	SOD1	SOD2	SOD3
SOX2	SPARC	STAT1	STAT3	SUMO1	TCF7
TERC	TERT	TET1	TGFB1	TGFBR1	TGFBR2
THBS1	TLR9	TNF	TNFAIP3	TNFRSF10B	TNFRSF13B
TNFRSF1A	TNFRSF1B	TOPBP1	TOX3	TP53	TP53BP1
TSC1	TSC2	TXNIP	UIMC1	VDR	VIM
WEE1	WNT3A	WNT5A	WNT7A	WRN	WT1
XIAP	XPC	XRCC4	XRCC5	XRCC6	XYLT2
YWHAQ	YWHAZ	ZEB1	ZHX2	MRE11A	

All gene lists we use:

<p>The Cosmic (Catalog of Somatic Mutations in Cancer) gene list contains curated genes from the Sanger initiative Number of genes 222</p>	<p>Gene list general is manually curated from two text books identifying genes generally important for cancer Number of genes 98</p>
<p>Gene list Rhodes contains genes upregulated in cancer cells Number of genes 65</p>	<p>Gene list MAPK contains genes in that pathway from biocarta Number of genes 87</p>
<p>Gene list sigcancer is a list of genes useful for identifying tumor tissue origin Number of genes 155</p>	<p>Gene list Radiation Number of genes 263</p>

All gene lists we use:

<p>The Cosmic (Catalog of Somatic Mutations in Cancer) gene list contains curated genes from the Sanger initiative Number of genes 222</p>	<p>Gene list general is manually curated from two text books identifying genes generally important for cancer Number of genes 98</p>
<p>Gene list Rhodes contains genes upregulated in cancer cells Number of genes 65</p>	<p>Gene list MAPK contains genes in that pathway from biocarta Number of genes 87</p>
<p>Gene list sigcancer is a list of genes useful for identifying tumor tissue origin Number of genes 155</p>	<p>Gene list Radiation Number of genes 263</p>

Discussion topic: is there additional usable knowledge in pathways, etc?

“Combos”

example:

Gene list:

Cosmic

MAPK

null

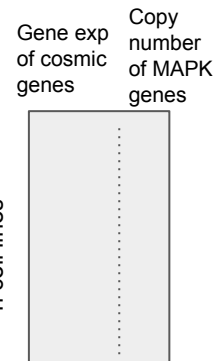
Feature type:

Gene expression

copy number

mutation

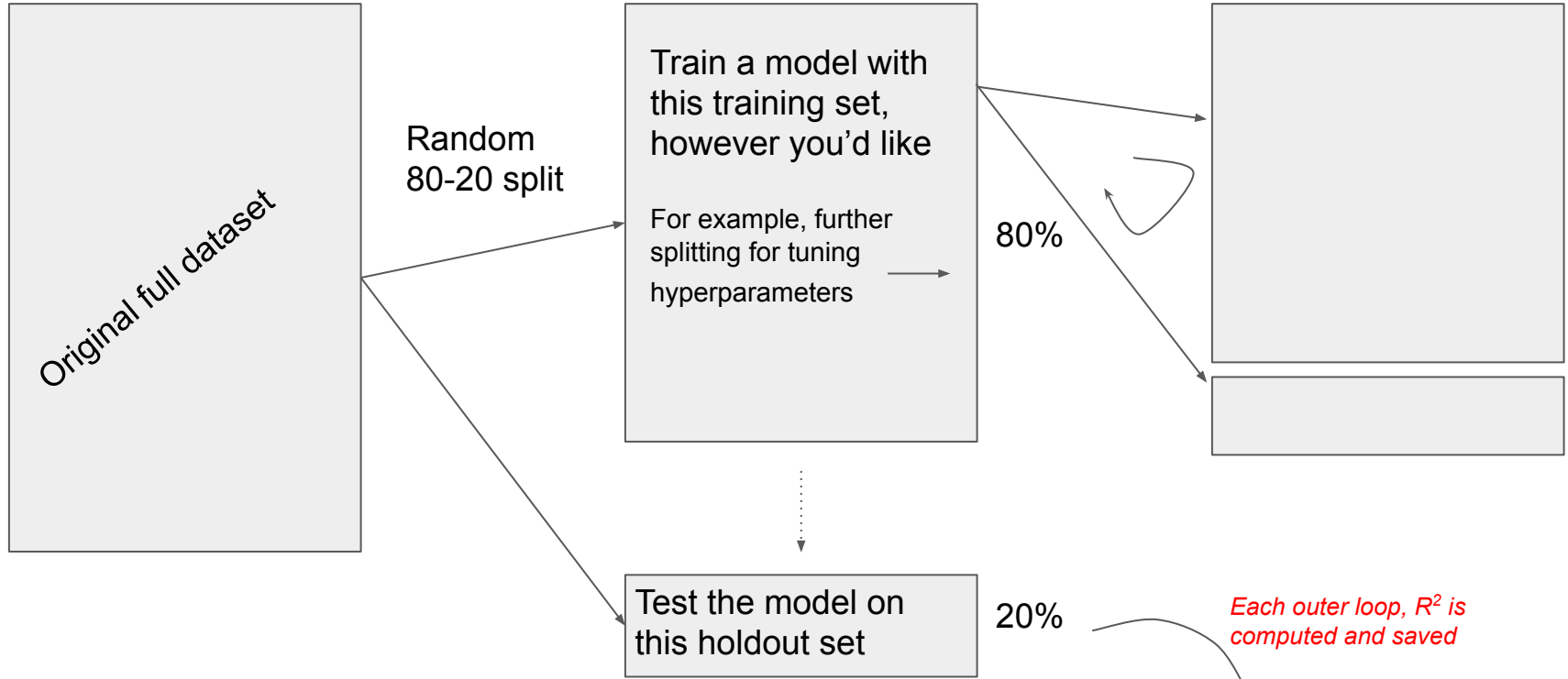
n cell lines



We run all possible combos: $7^3 - 1 = 342$

Data handling & data splitting

For a given drug, and a given combo



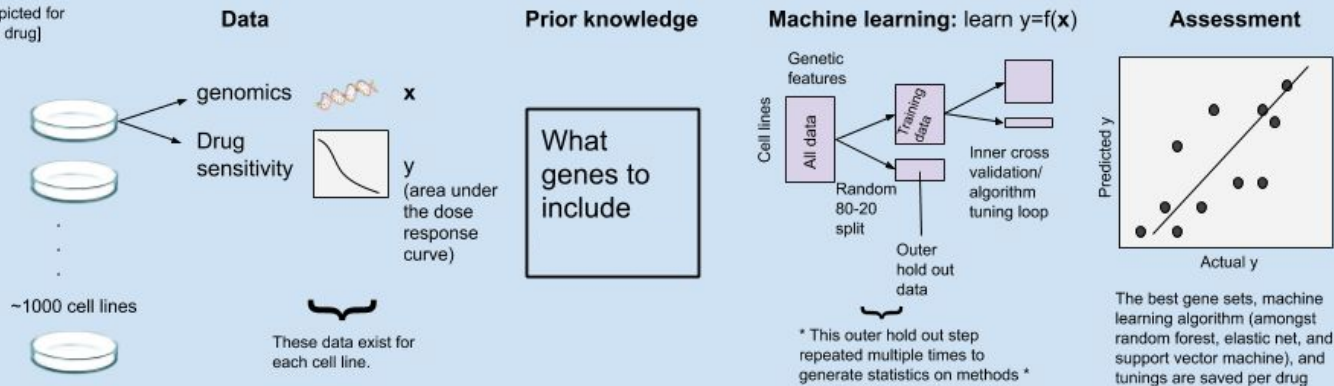
Each outer loop, R^2 is computed and saved

This is repeated several times!

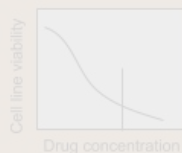
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

A: Model analysis and selection (MAS): find best machine learning algorithm and tuning parameters to predict drug sensitivity from genomics

[Info depicted for a single drug]



B: Drug Recommendation System (Dr. S): using model tuning from MAS, make and assess drug recommendations for new cell lines



For D drugs ($D \sim 50$) we create D models, $f_1(x)$, $f_2(x)$, ..., $f_D(x)$. For each drug we choose a concentration level (chosen to produce a fair and challenging drug selection problem) and the machine learning models are retrained to predict viability at that concentration level. Thus, $f_i(x)$ is the machine learned function which maps x (the genomic data from a cell line) to a viability level for drug i at the concentration level modeled for that drug.

A new cell line, not seen in the training, characterized by genomic feature vector x , is sent through the D drug models.

This procedure is repeated for every cell line (ML algs for each drug retrained every time: leave-one-out procedure)

$f_1(x)$

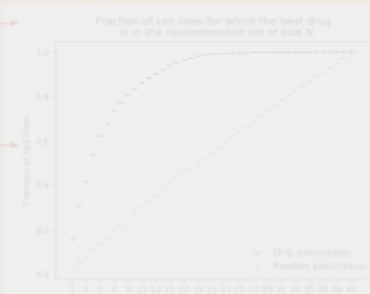
$f_2(x)$

$f_D(x)$

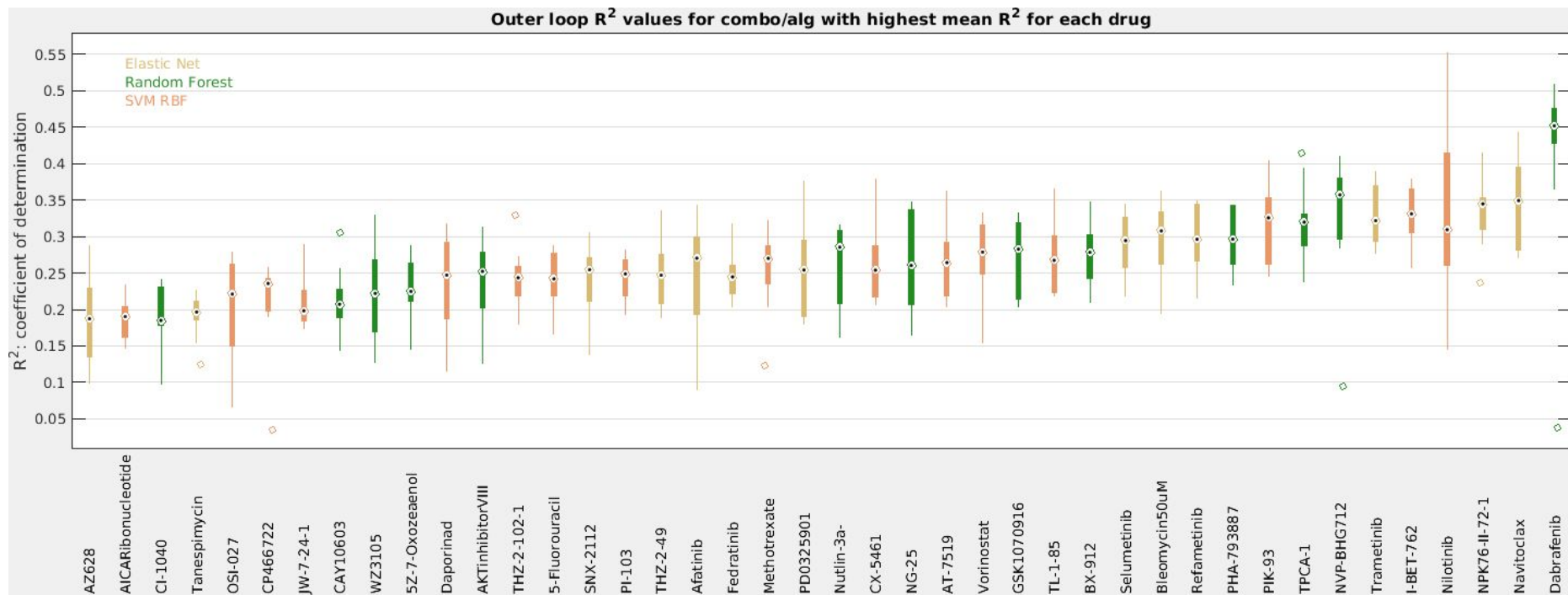
A batch of the $N \gg 1$ drugs with the most potency is recommended

Assessment

Results from that procedure, performed $\sim 1000 \times 50$ (number of cell lines * number of drugs) times (each requiring the training of an ML algorithm) are compiled to assess the overall performance of Dr. S. For example:



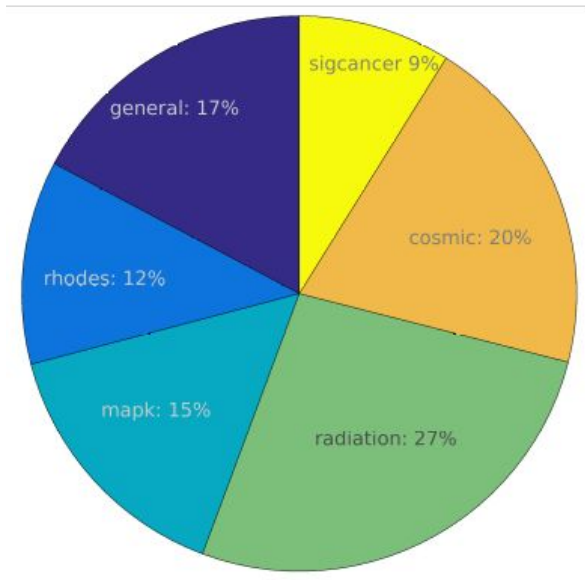
Model Analysis and Selection (MAS) Results



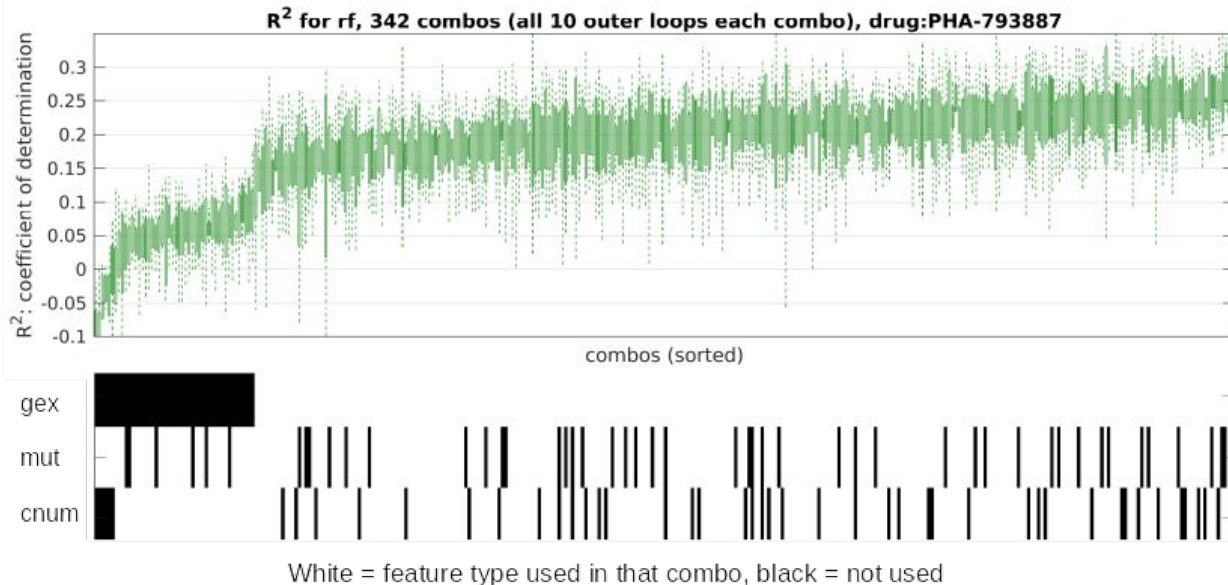
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

y_i = true drug AUC, \hat{y}_i = model predicted response, \bar{y} = mean of the true responses in the holdout data

Model Analysis and Selection (MAS) Results

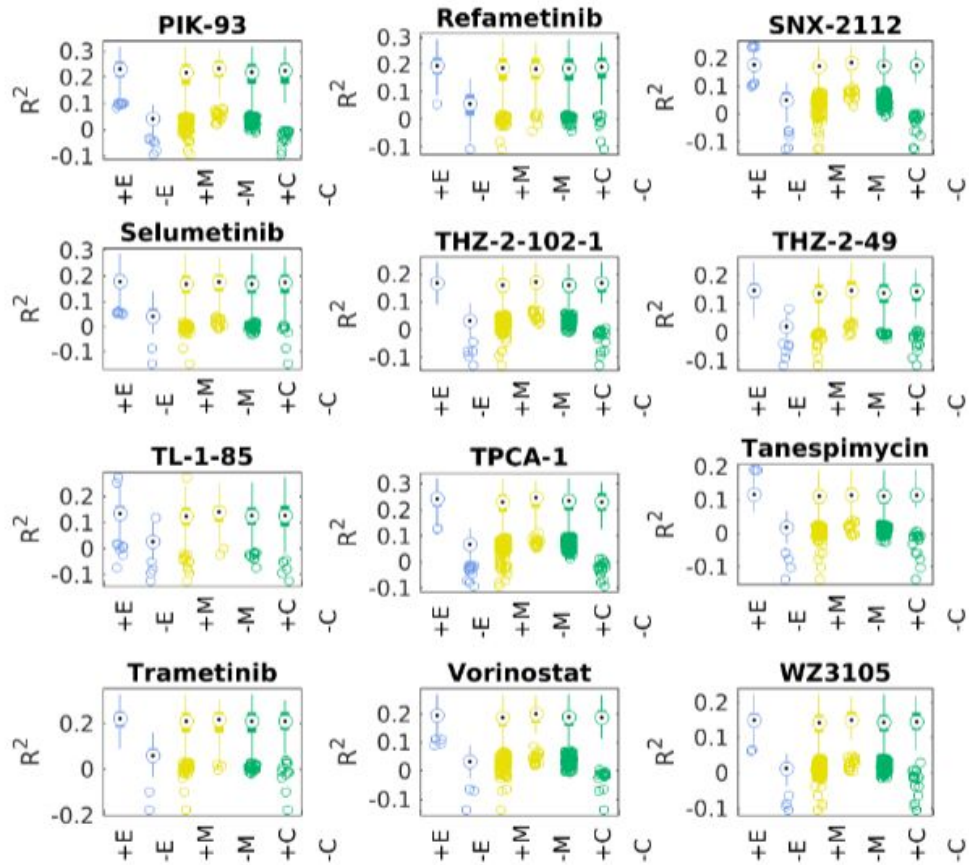


Overall gene list usage in top combos

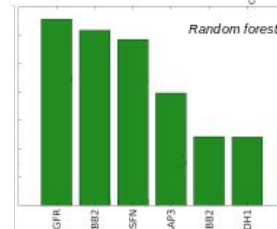
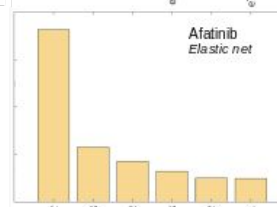
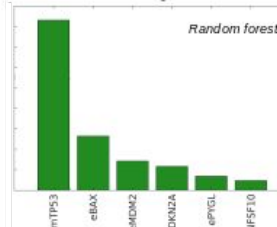
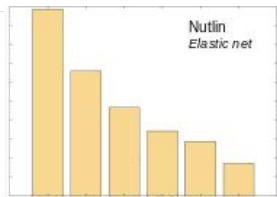
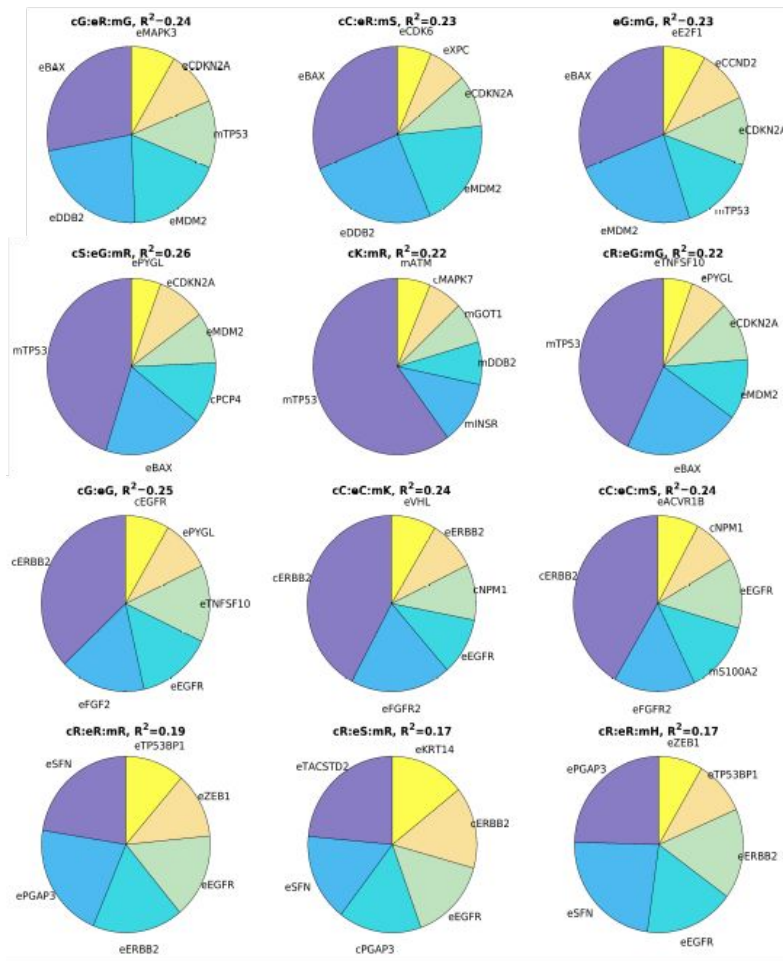


Single drug all combos example result

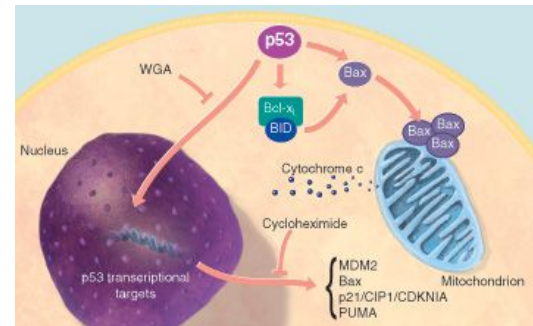
Gene expression is most important genomic type



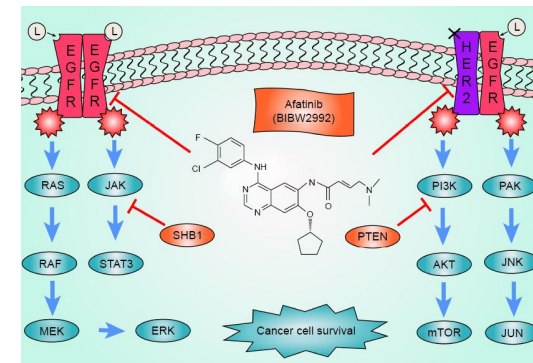
Sample feature importances results



Drug = Nutlin-3A

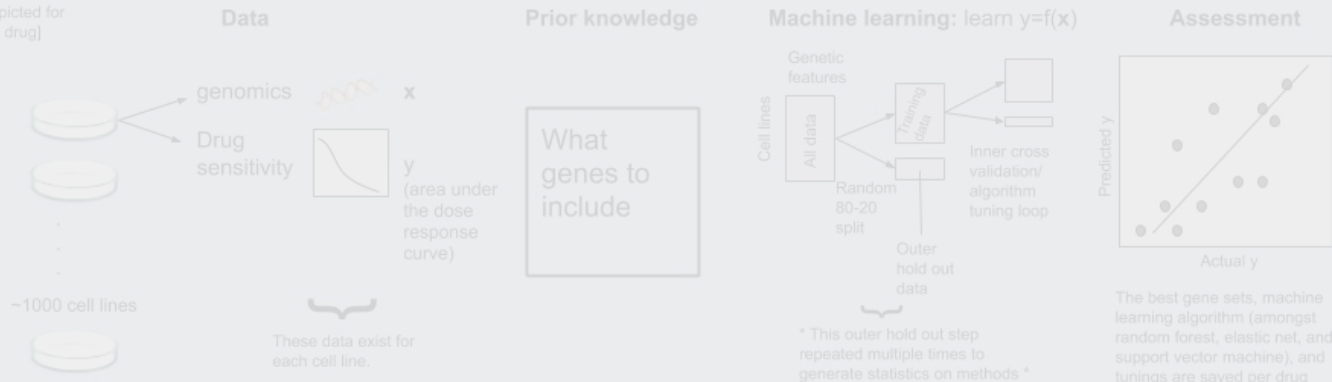


Drug = Afinatinib

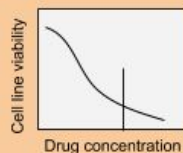


A: Model analysis and selection (MAS): find best machine learning algorithm and tuning parameters to predict drug sensitivity from genomics

[Info depicted for a single drug]



B: Drug Recommendation System (Dr. S): using model tuning from MAS, make and assess drug recommendations for new cell lines



For D drugs ($D \sim 50$) we create D models, $f_1(x)$, $f_2(x)$, .. $f_D(x)$. For each drug we choose a concentration level (chosen to produce a fair and challenging drug selection problem) and the machine learning models are retrained to predict viability at that concentration level. Thus, $f_i(x)$ is the machine learned function which maps x (the genomic data from a cell line) to a viability level for drug i at the concentration level modeled for that drug.

A new cell line, not seen in the training, characterized by genomic feature vector x , is sent through the D drug models.

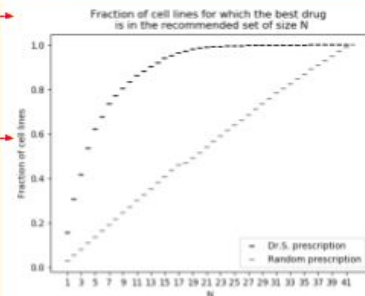
This procedure is repeated for every cell line (ML algs for each drug retrained every time: leave-one-out procedure)

 $f_1(x)$
 $f_2(x)$
 $f_3(x)$
 $f_4(x)$
 $f_5(x)$
 $f_D(x)$

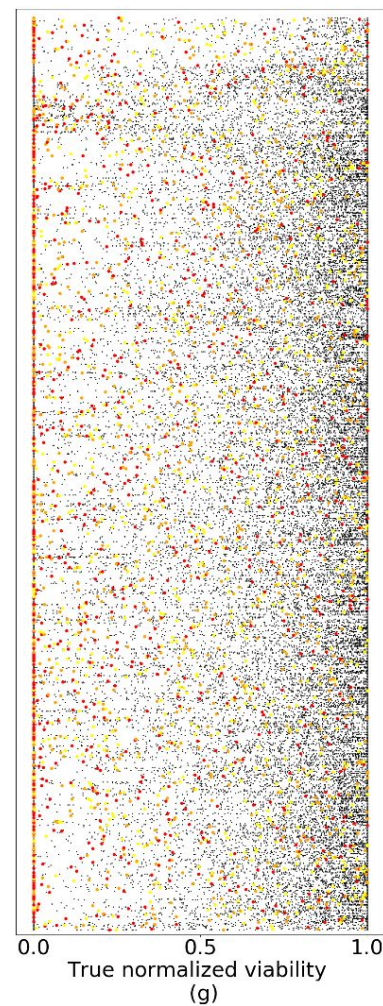
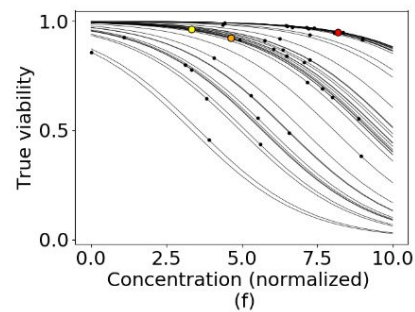
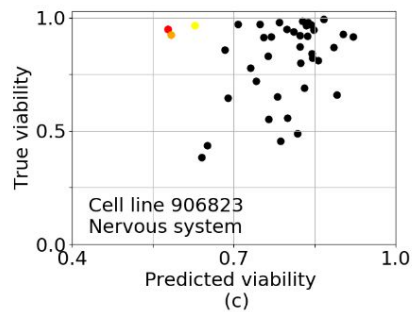
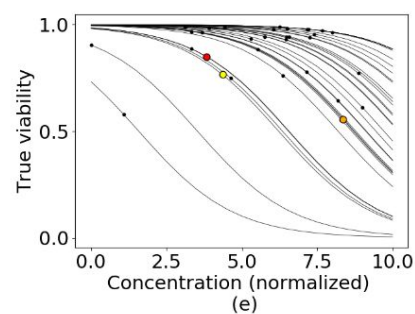
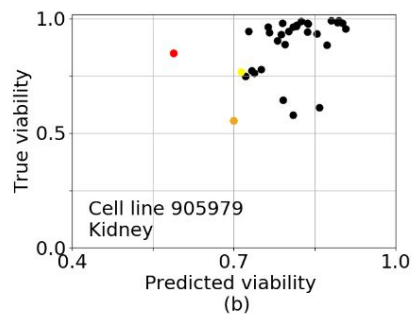
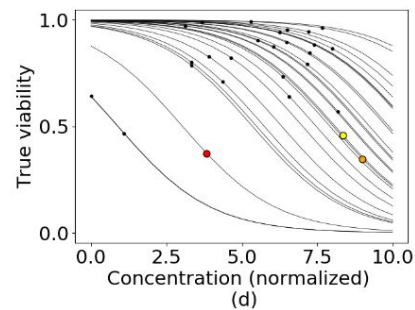
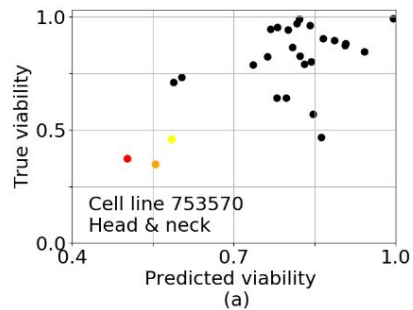
A batch of the $N \geq 1$ drugs with the most potency is recommended

Assessment

Results from that procedure, performed $\sim 1000 \times 50$ (number of cell lines * number of drugs) times (each requiring the training of an ML algorithm) are compiled to assess the overall performance of Dr. S. For example:

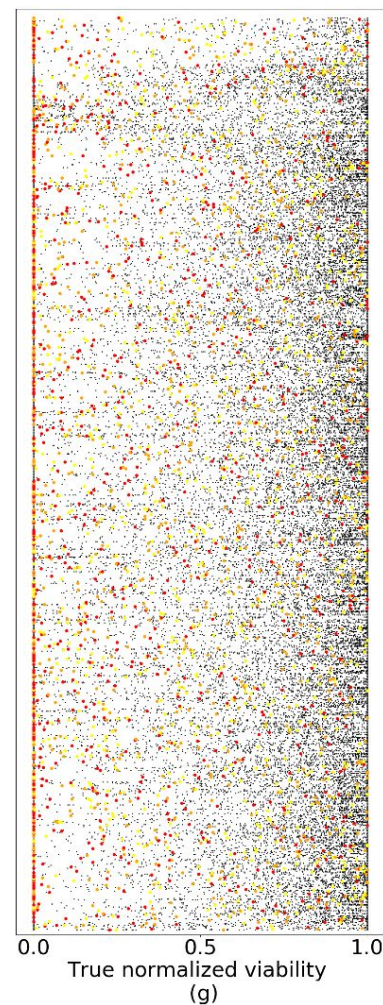
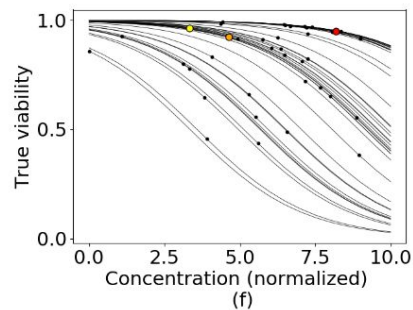
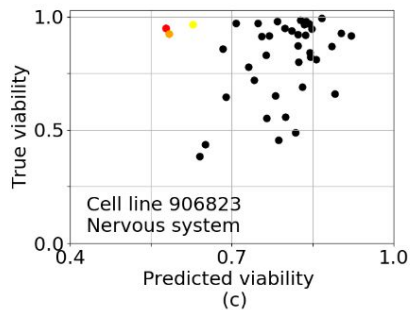
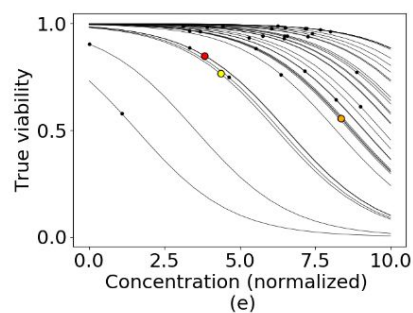
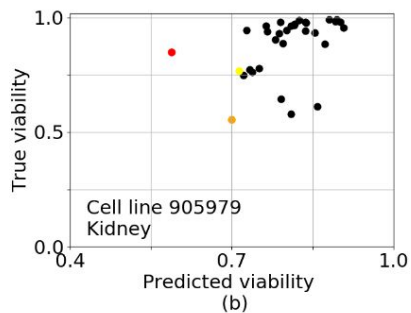
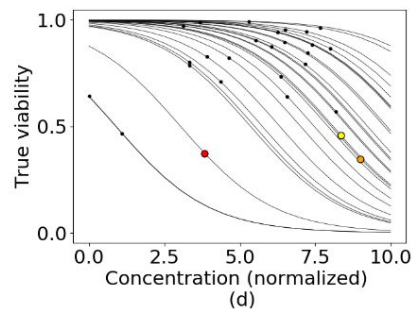
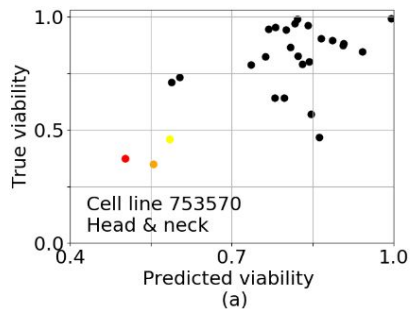


Dr.S results

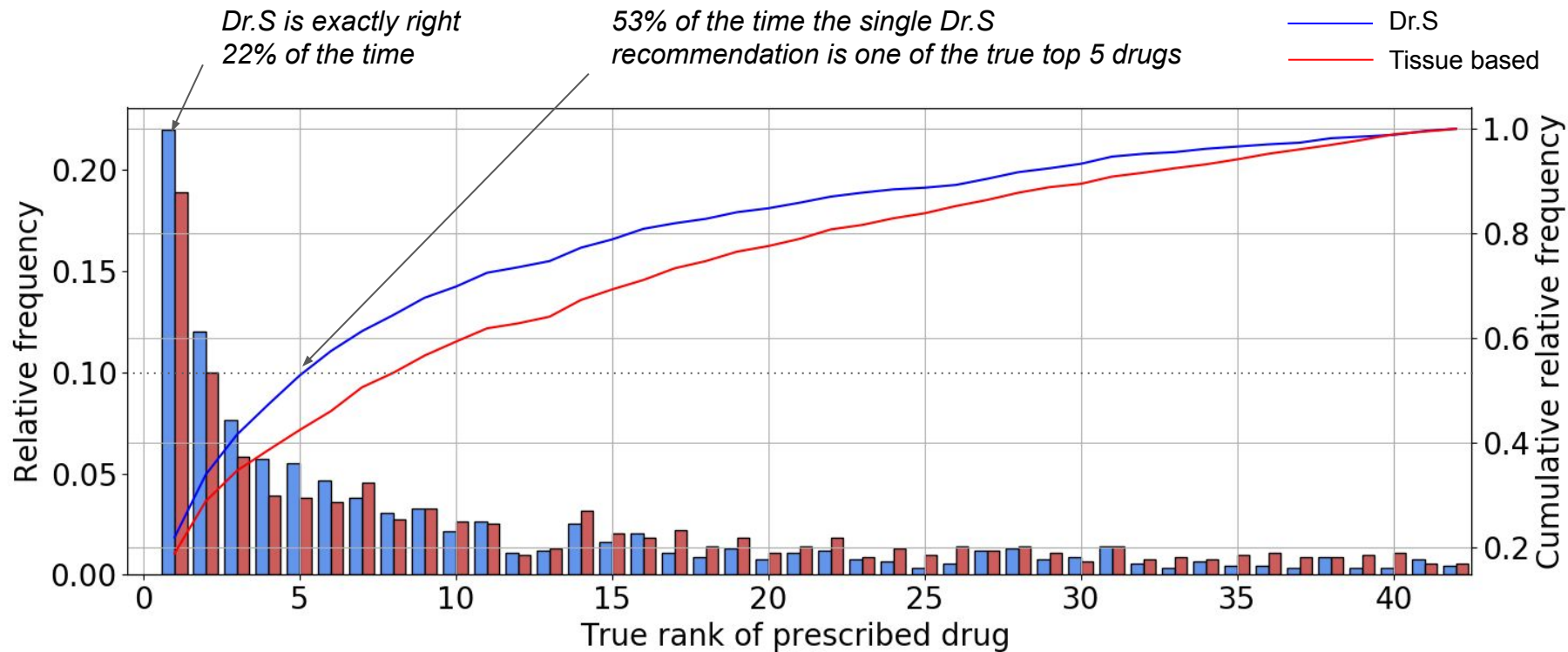


Dr.S results

Even without great R^2 ,
we often make good
recommendations



Dr.S results: single drug recommendation



Contributions / best practice recommendations:

- Repeated hold out “triple split”
- Single drug (rather than “matrix completion”/multitask) modeling
- Incorporation of prior knowledge

Next steps:

- Pathway modeling / neural networks / similarity kernels
- Other -omics data / other gene sets
- Learn entire dose response curve
- Clinical patient data / learn drug efficacy and toxicities dose response curves
- Pan cancer vs tissue type specific learning?
- ...

The team

Marleen Balvert - CWI, The Netherlands

Georgios Patoulidis - Master's student Heidelberg University

Andrew Patti - Software developer, Engage

Timo Deist - post-doc CWI

Christine Eyler - MD, PH.D MGH

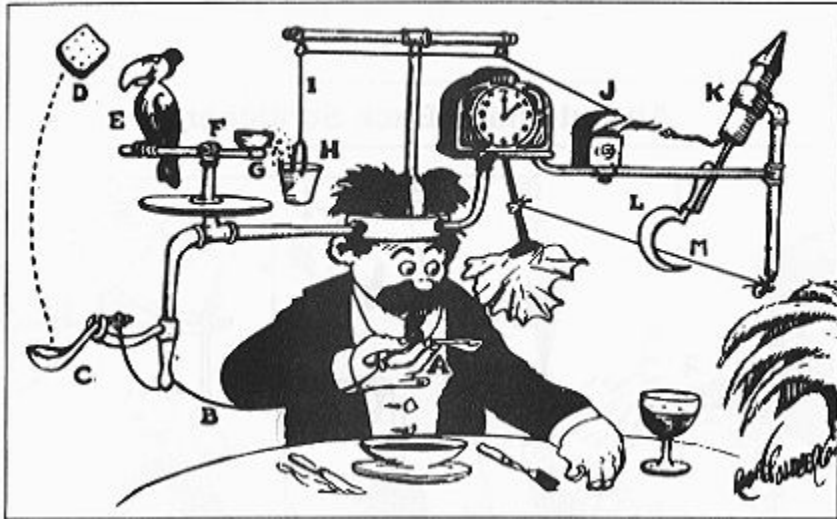
A photograph of a field of green plants, likely nettles, under a blue sky with clouds. The plants are in the foreground and middle ground, with a dense line of trees in the background. A green banner with white text is overlaid on the image.

Thank you! Questions/comments?

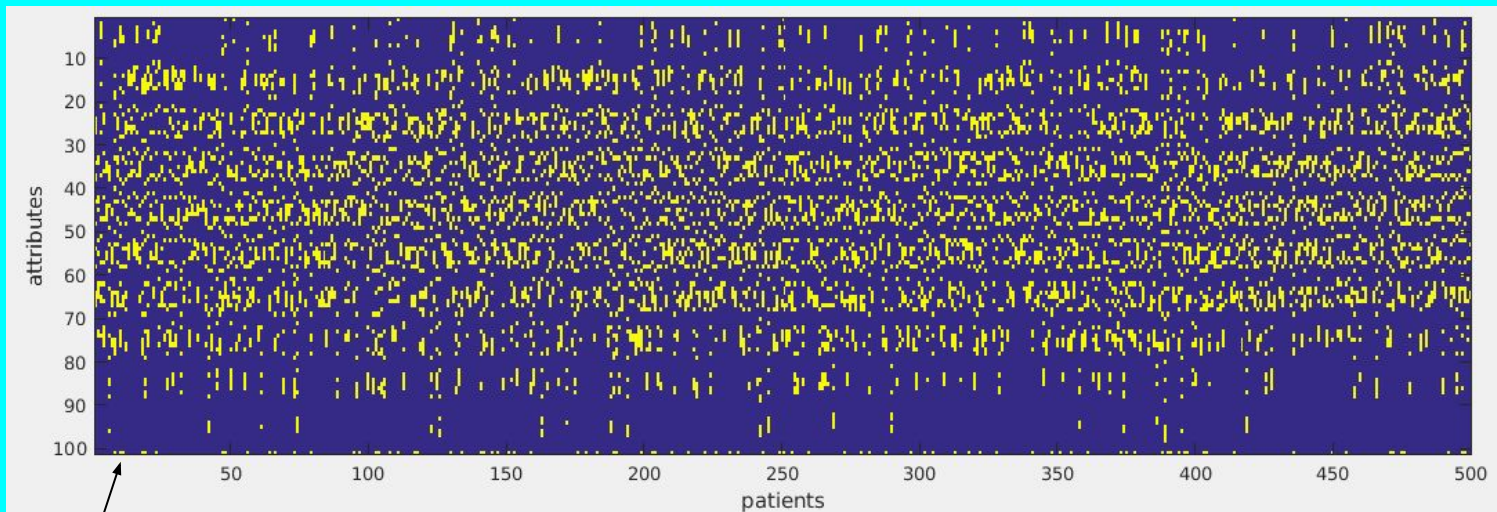
Additional slides

How complicated is biology? Rube Goldberg...

Self-Operating Napkin



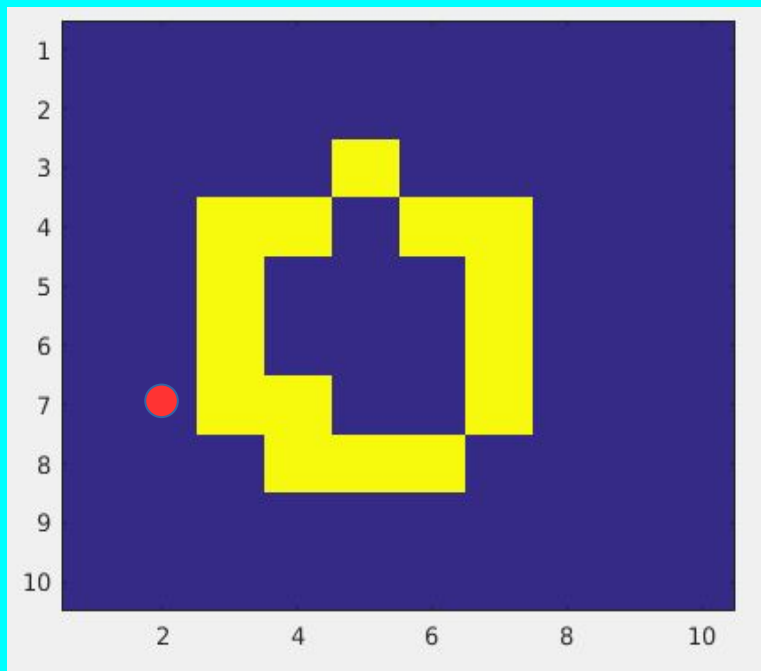
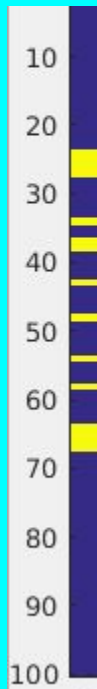
A cartoon example: knowledge helps



Last row gives response
to a “drug”

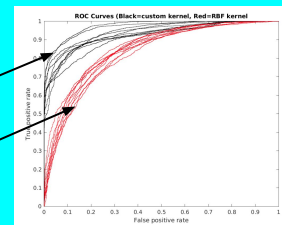
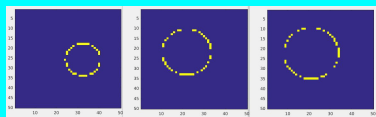
SVM with an RBF kernel,
Average AUC (over 5 drugs)= 0.89

System expert knowledge

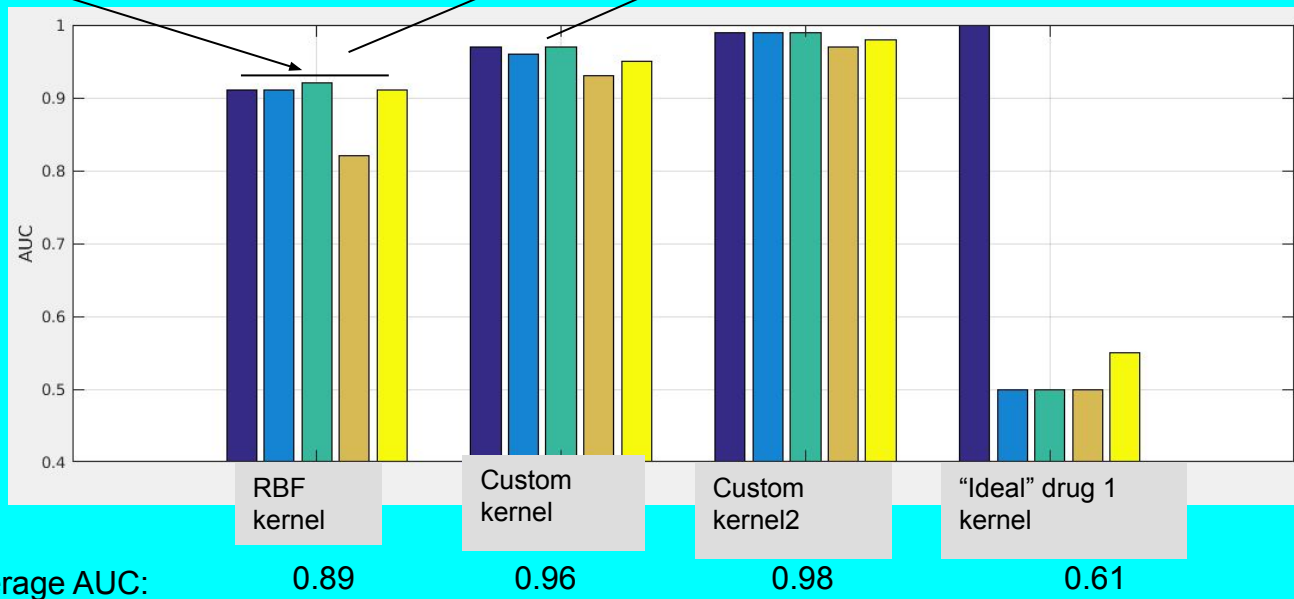


Custom similarity measure \Rightarrow kernelized learning

Two “cell lines” are similar if their circles are similar.

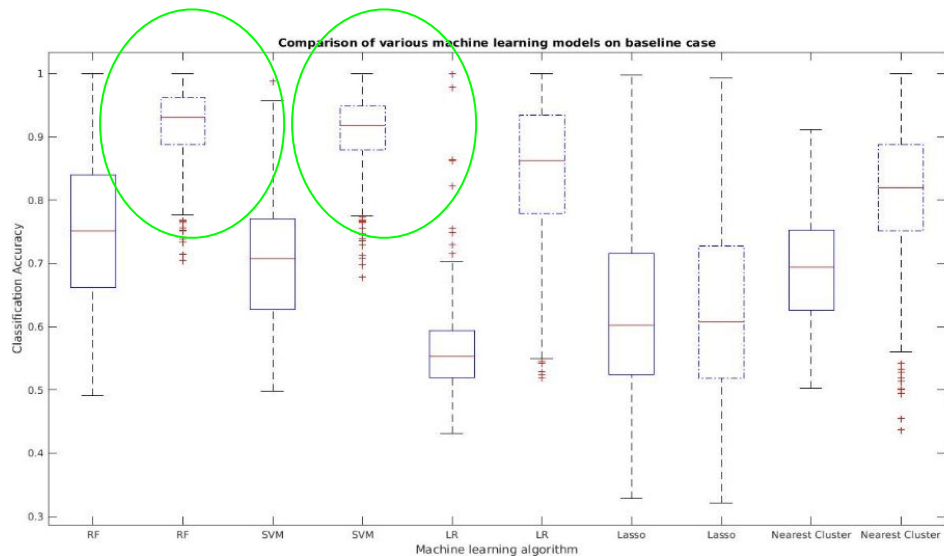


5 different “drugs”





Again, prior knowledge helps

Results (baseline 180 node networks)

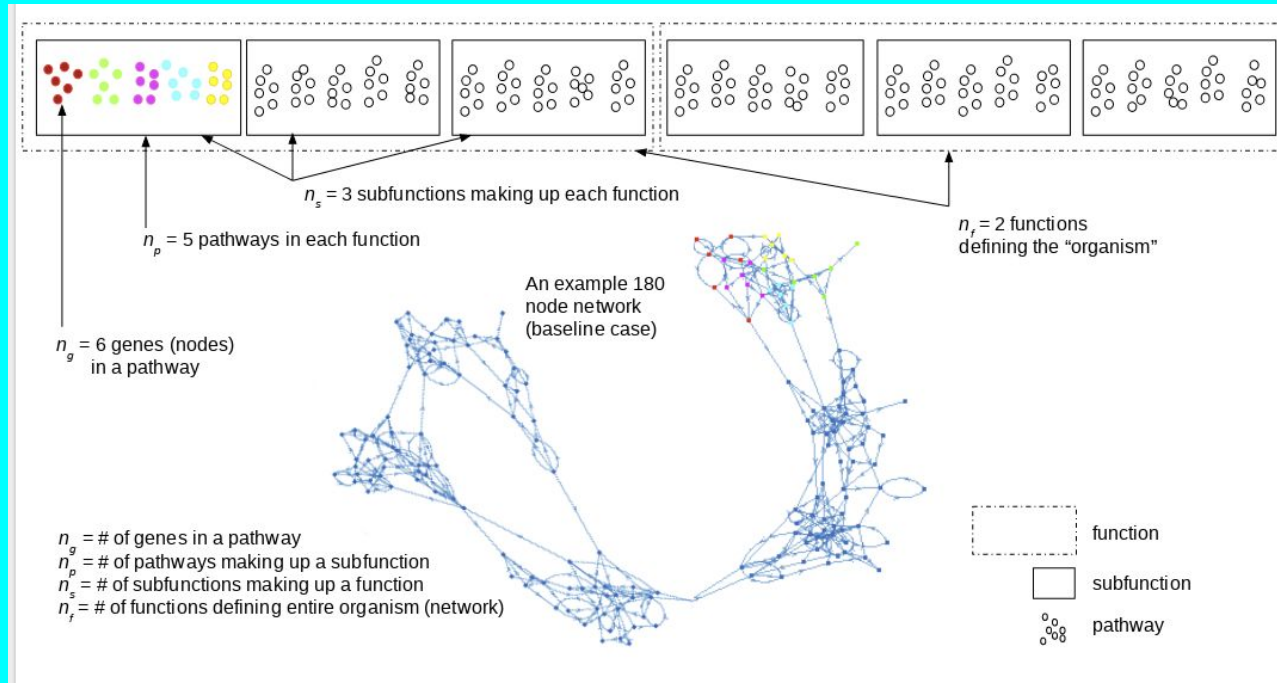


RF = random forest
SVM = support vector machine
LR = logistic regression
Lasso = Lasso regression
Nearest cluster: "k-means"

 No prior knowledge

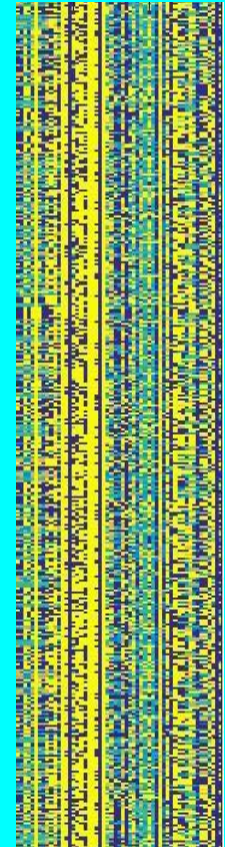
 Prior knowledge

Dynamical Boolean networks demo



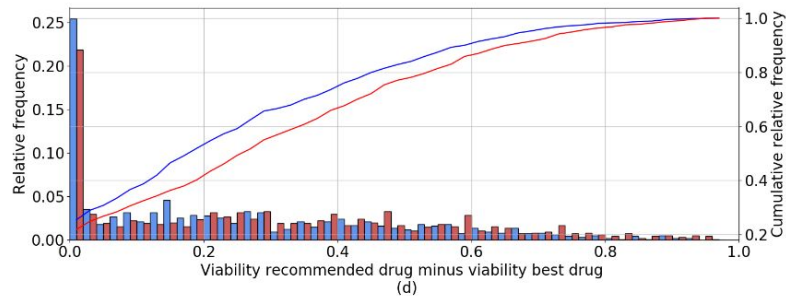
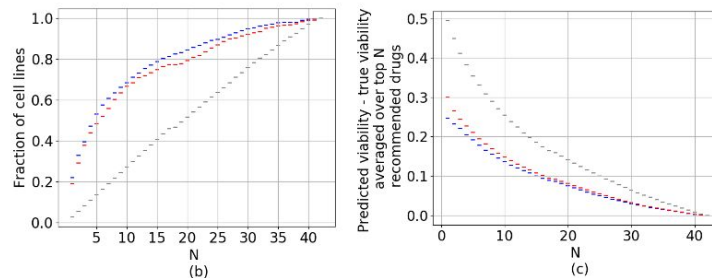
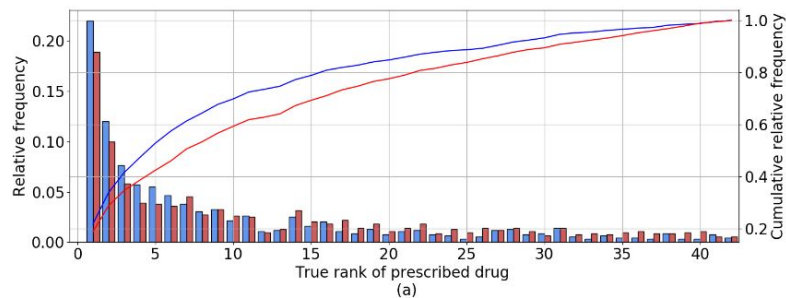
P = number of features

n



Dr.S results

Comparison of drug recommendations made with Dr.S. (blue), recommendations based on the average viability of other cell lines of the same tissue type from the training data (red) and random drug selection (gray). (a) The (cumulative) distribution of the true rank of the single prescribed drug. (b) The fraction of cell lines for which the single true best drug is among the N true recommended drugs. (c) The difference in viability between the top N prescribed and the top N true best drugs. (d) The distribution of the difference in viability between the single prescribed and single best drug.



- Dr.S. prescription
- Average viability prescription
- Random prescription

Dr.S results

Results obtained with Dr.S and tissue type based prescription where all drugs with a predicted viability within of the best predicted viability are in the prescription. In Figures (a)-(c) each row corresponds to a cell line, and the dots indicate the normalized viability of the recommended drugs where a normalized viability of 0 (1) corresponds to the viability of the best (worst) drug. Figures (d)-(f) show the (cumulative) distribution of ϵ^* over the cell lines, averaged over the set of recommended drug for that cell line. Figures (g)-(i) show the (cumulative) distribution over the cell lines of the maximum ϵ^* over the recommended drug for that cell line

