Note (May 2020): This is the early stages of what later became the emba R package.

Title

An Ensemble Model analysis on drug synergy prediction performance (report)

Purpose

Analyze model prediction performance, find insightful characteristics of models that make them predict observed synergies, understand underlying model mechanisms that specific synergies take advantage of (related to fitness scores, stable states)

Methods

Analysis was done in R (scripts), using “Input” as defined below

Input

- Models from the /models folder (Gitsbe output). Each file has info on a different model (fitness, stable state (1 only - BNReduced script used), boolean equations)
- Model_predictions file (Drabme output)
- Observed_synergies file (Web server/ Miguel output)

Note: the following analysis can apply to any configuration of this kind of input. For our analysis, we used data from the AGS cell line, the web server results are here: http://localhost:8080/SINTEF/ROC/AGS_0dcc89b39f3527f93e3e9c376a140398/info

Output/Analysis/Results Overview

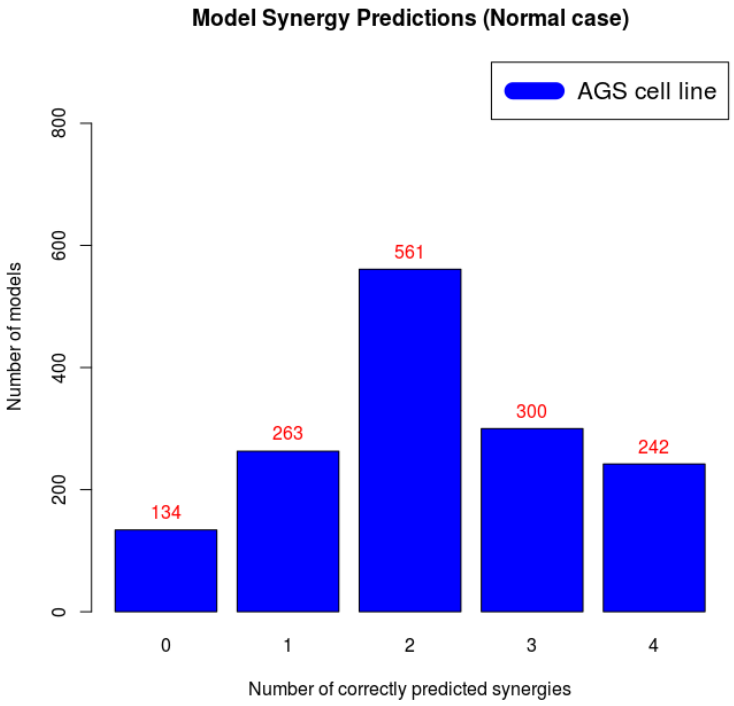
1. Counting number of models vs correctly predicted synergies
2. Fitness values of high vs low performance models: statistical correlation
    a. Fitness range investigation (try to get a larger fitness range)
3. Good vs Bad model average stable state differences (3 ways to distinguish between good and bad models)
4. Heatmap of models stable states (include row coloring of number of predicted synergies and fitness values)
    a. Also change row clustering based on number of synergies predicted or based on the fitness values
    b. Do the same analysis with a heatmap of equations
5. All the above with the random run (models are training only to proliferating state)
6. All the above with the A498 cell line (has many more synergy predictions)
7. Heatmap of stable states and equations with all tested cell lines (8)

AGS (normal run: fitting to steady state)

Counting number of models vs max corrected predicted synergies

There are a total of 6 experimentally predicted synergies, namely:
"AK-BI" "5Z-D1" "AK-D1" "BI-D1" "PI-D1" "PK-ST"

No models from the normal run predicted the "5Z-D1" and the "PI-D1" synergy (so the maximum number of correctly predicted synergies for AGS is 4). The total number of models is 1500. The total number of synergies tested (computationally) were 153. Note that in the next graph we keep track of the exact maximum number of correctly predicted synergies that each model predicted.



Notes:

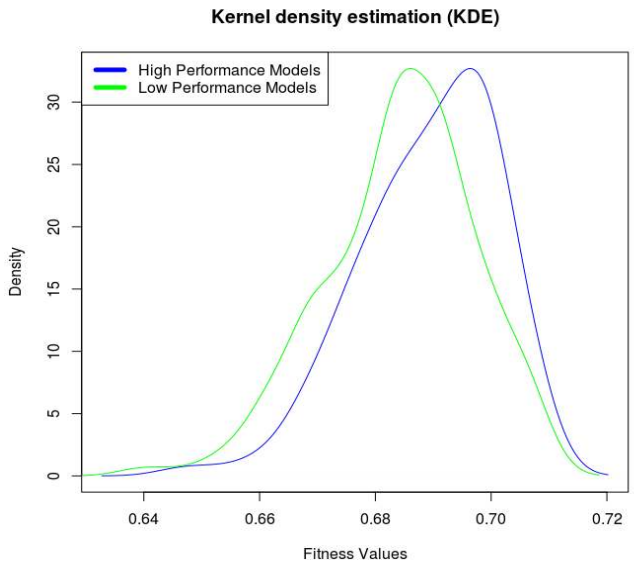
- Very few models predicted no observed synergies, while more than ⅓ of the models predicted at least 2 of them correctly
- < 9% of the models predict no synergies
- The models cannot predict more than 4/6 = 66% of the actual synergies.

Fitness values of high vs low performance models: statistical correlation

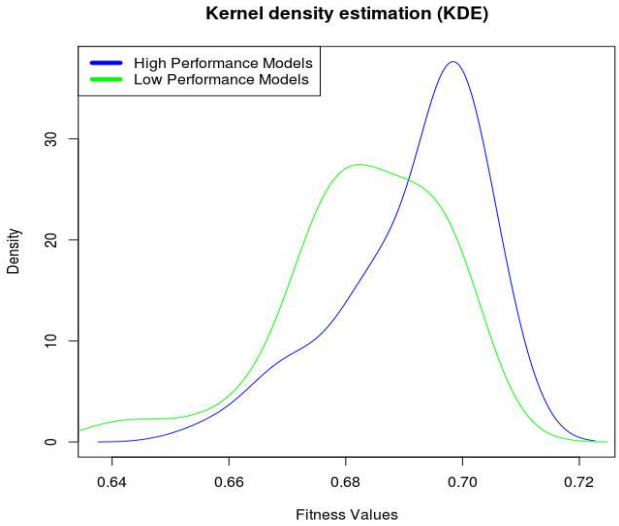
We compare the fitnesses of the models that correctly predicted 3 or more drug combinations to the models fitnesses that had less than 3 correctly predicted synergies. So, we get two vectors with fitness values that correspond to each said category. The fitness values range is from 0.63 to 0.70 and the distinct values in that range were 10.

Are the two vectors of fitness values belong to the same distribution? Let's say that they do (our Null Hypothesis). By doing a two-sample Kolmogorov-Smirnov test, we get:
W = 319820, p-value = 3.047e-14
alternative hypothesis: true location shift is not equal to 0

With a small p-value (<0.05) we can safely reject our Null Hypothesis, so the alternative (that the two vector of values come from different distributions) is more probable to be true, i.e. that models with higher fitness (stable state more equal to steady state) are better at predicting synergies. We draw some estimates on the probability density functions where we can see that they indeed differ (also note that the fitness values fall between [0.6397, 0.7059]:

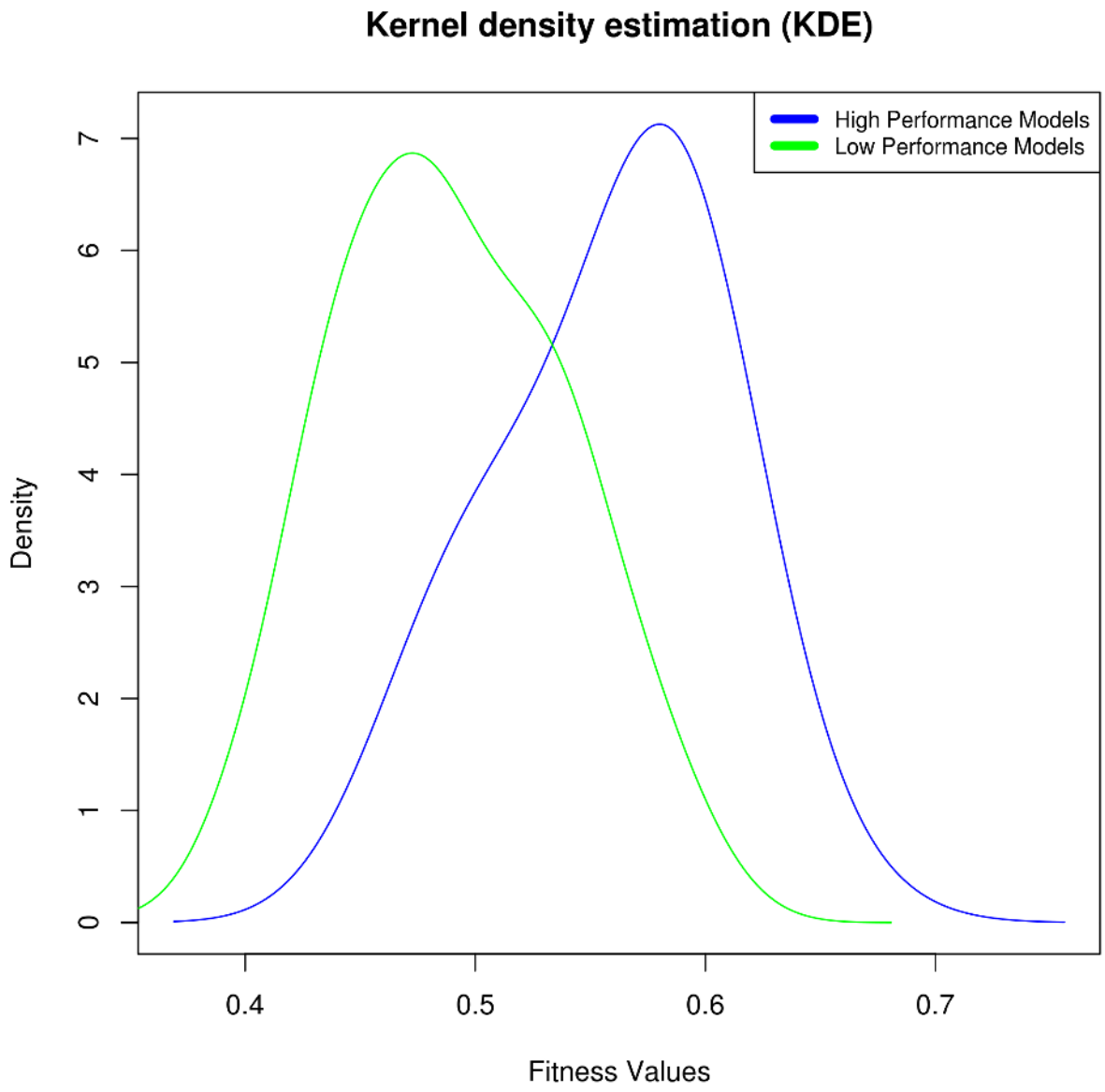


To make our point stronger, we compare the fitnesses of the models that correctly predicted 4 drug combinations (the maximum for AGS) to the models fitnesses that predicted none:



Fitness range investigation

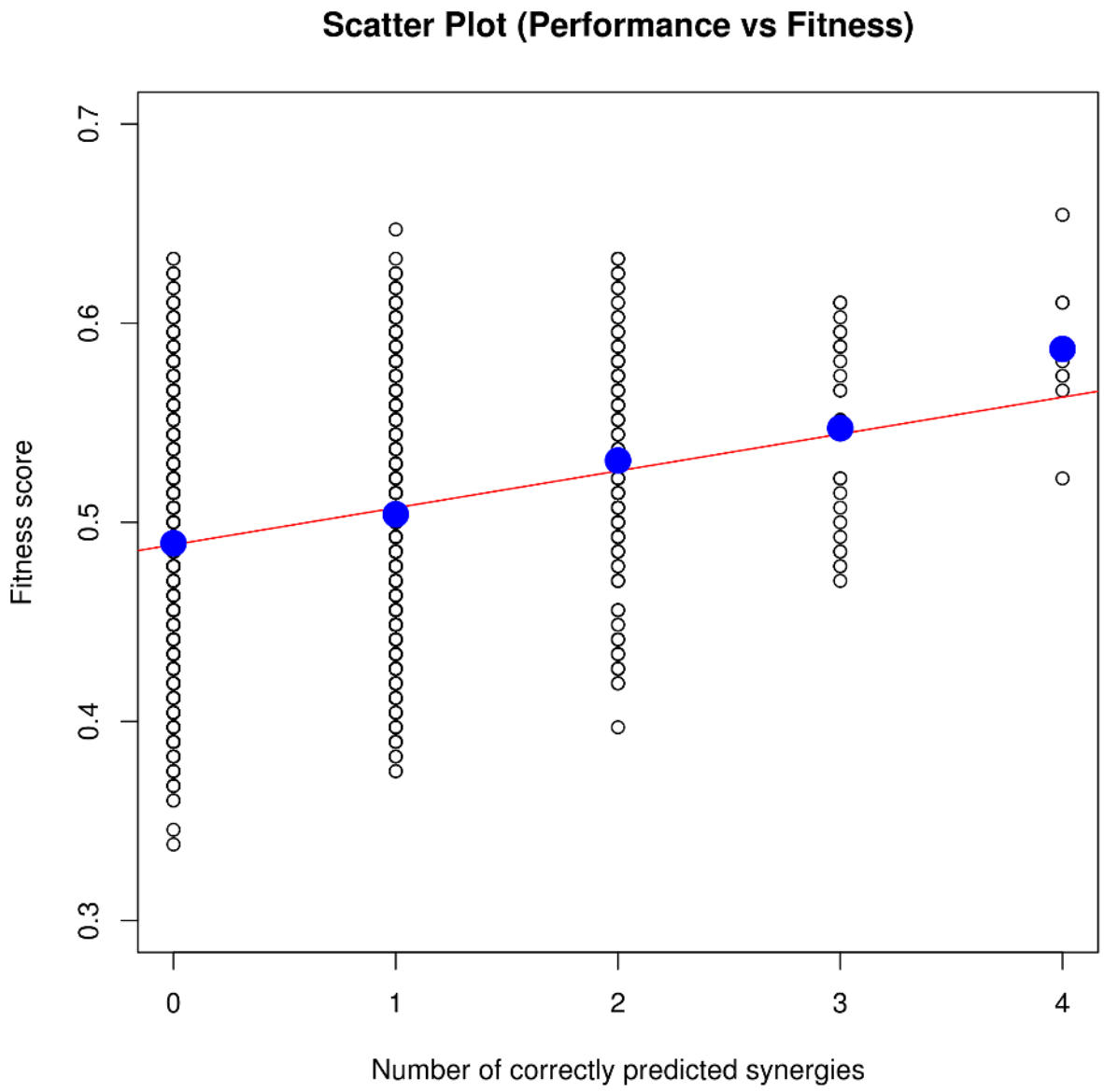
In this analysis, we want to produce the same results as the previous section but for a larger range of fitness values. A new AGS run was set with specific configuration to capture this. I got ~4000 models with fitness scores ranging from 0.33 to 0.65. The results are:



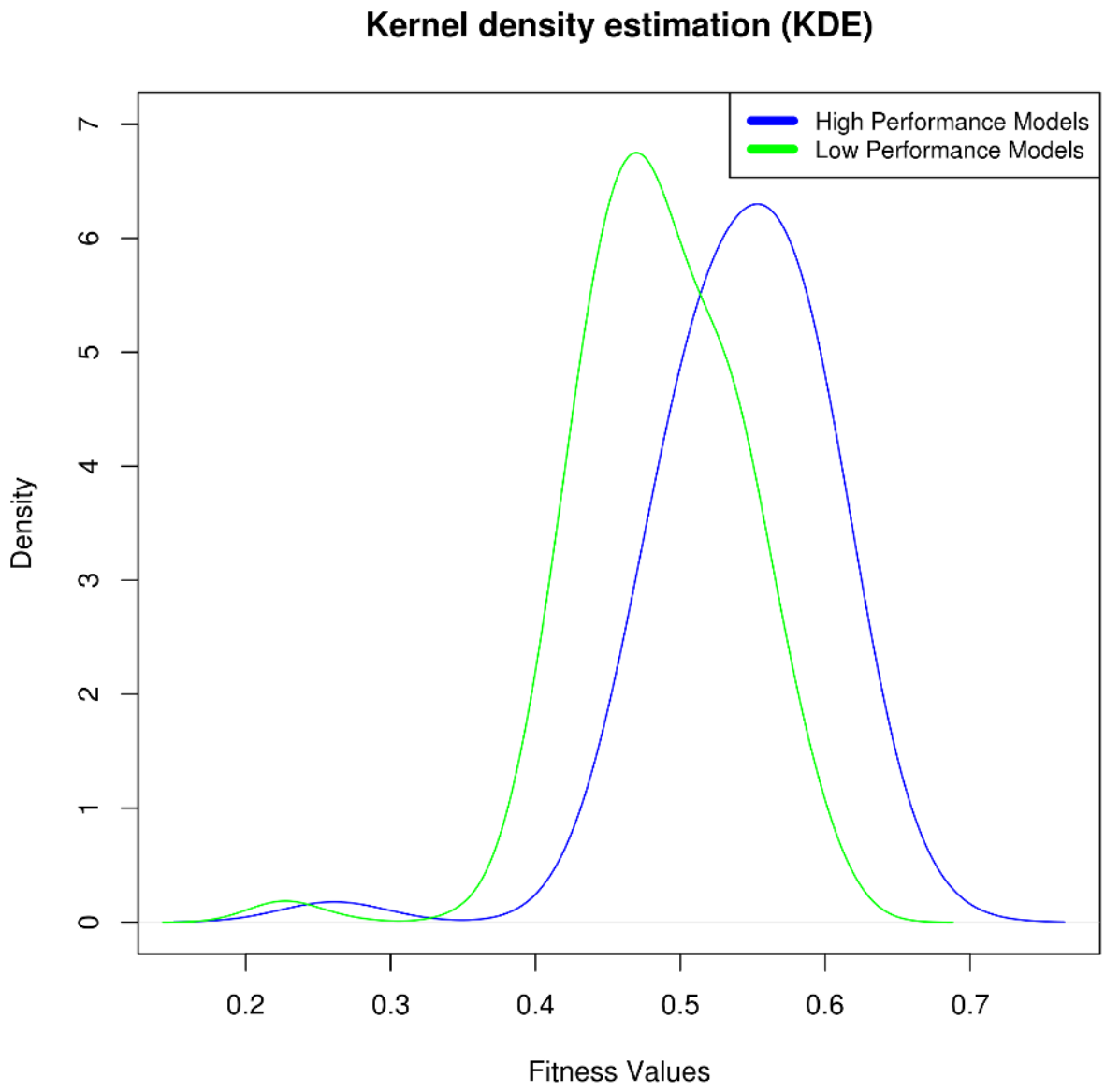
Notes:

- High performance: models that predicted 3 or 4 synergies.
- Low performance: models that predicted 0 synergies.

Also, we did a scatter plot where the average values are also shown in blue filled circles:



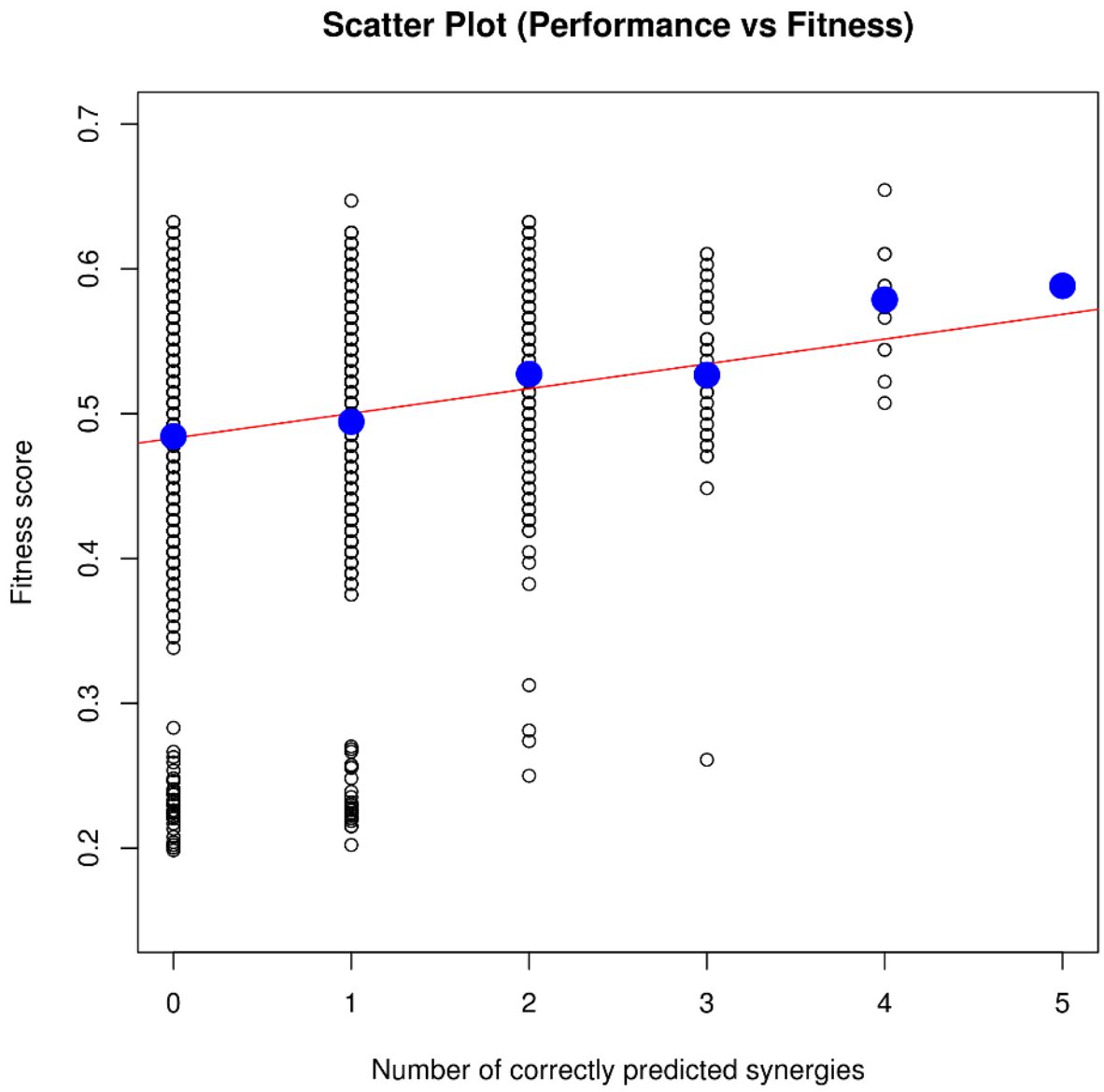
Next, we did the same run as above (same configuration), but now using the M2 library for the calculation of the steady states (full BNReduction script). Thus models that had 2 or more stable states get lower fitness and thus we could get a larger fitness range. IN the end, I got 4218 models with fitness scores ranging from 0.19 to 0.65. The results are:



Notes:

- High performance: models that predicted 3, 4 or 5 synergies.
- Low performance: models that predicted 0 synergies.
- There was only 1 model that actually predicted 5 synergies (out of 6 experimentally observed).

The scatter plot is as follows (note that these are models from 1 generation only, so no further optimization through genetic evolution was applied in these models - they were just randomly generated and evaluated for their predictions):



Result: With higher predicted performance, fitness values less than ~0.3 disappear from the above graph, which are exactly the models that have more than 1 stable states (since we couldn't get these fitness values in the previous analysis!). Note the two clusters of points (one down, one up) that can be seen in the graph above, corresponding to more than 1 (down) and just 1 (up) stable state.

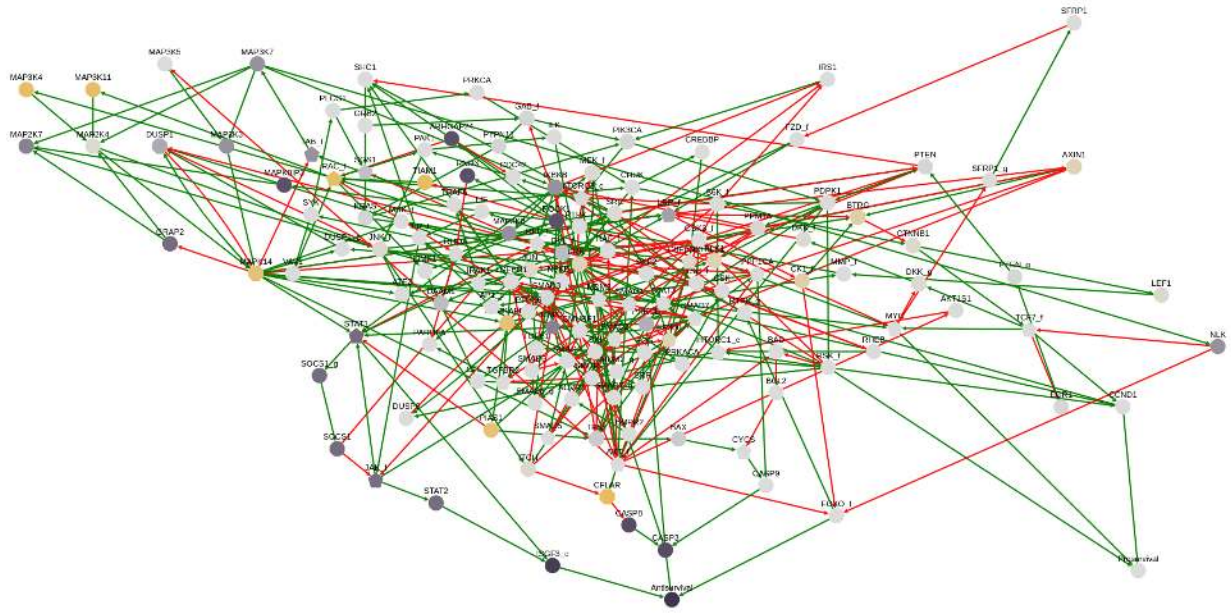
Good vs Bad models (average stable state differences)

I started by defining which are the "good" and which are the "bad" models for this analysis (so far I have found 3 ways to distinguish them that makes some sense for further analysis). For each of these sets of models I took their stable states and for each node in these stable states I

computed the mean value. Then, I calculated the difference between good and bad model's average stable state (so, it's diff = good - bad). Using Miguel's after-gitsbe tool I could visualize this difference with specific colors on each network node (and see what affected the difference in performance). Note that the colors will be yellow for the most positive difference, white for near 0 difference and purple for the most negative difference value.

1st way (High performance vs low performance models)

Bad models: they predicted no synergies (134 models)
Good models: they predicted the maximum number of synergies (4 for AGS) (242 models)



Discussion: From the above graph we can see that the Antisurvival node is more inactive in the models that predict synergies. This node is affected by 3 others, of which one has not changed on average between the good and the bad models (FOXO_f). The other 2 (CASP3, ISGF3_c) are more inactivated in good models and there seems to be a cascade of activation/inactivations that enable them to take that state in the good models: for CASP3 it is the activation of the CFLAR and the subsequent inactivation of CASP8. For ISGF3_c, we trace back to the inactivation of STAT1 and STAT2 and we also notice the activation of MAPK14 in the good models which is a little bit upstream of these nodes (and which makes sense that is more activated in the good models, since it is the BI drug's inhibitor: the bad models seem to have already inactivated somehow on average). The rest of the drug targets (of the drugs that compose the predicted drug synergies) have no changes between good and bad models:

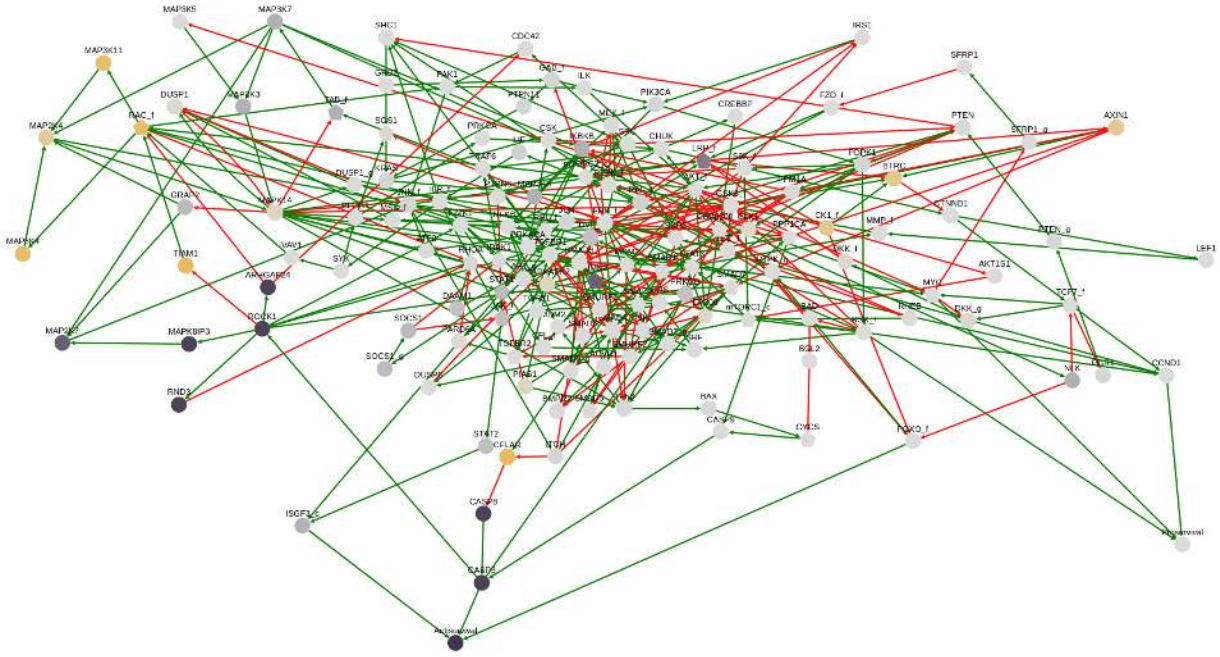
AK inhibits AKT_f (no change)
BI inhibits MAPK14 (more active in good models)
D1 inhibits RSK_f (no change)
PK inhibits CTNNB1 (no change)

ST inhibits STAT3 (no change)

2nd way (find what mechanisms a particular synergy is exploiting)

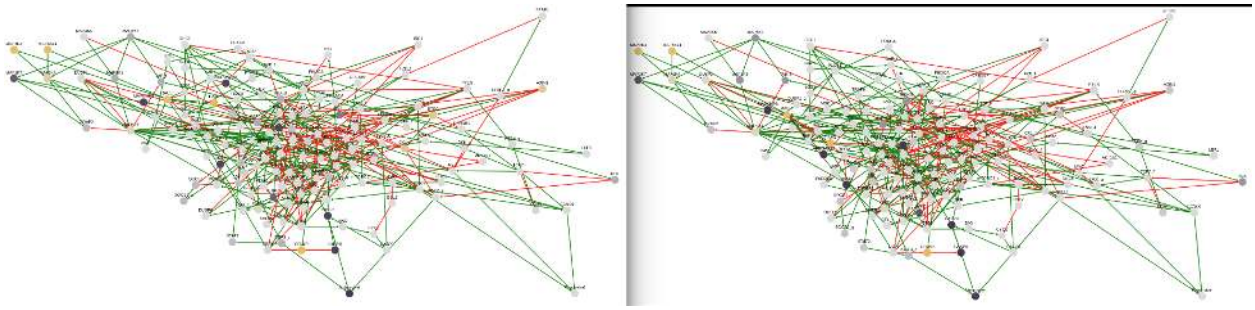
Bad models: they didn't predict a specific synergy (may have predicted others)
Good models: predicted that specific synergy

For example, for the drug synergy AK-BI, 462 models predicted it, while 571 models found it as a non-synergy (the rest of the models didn't have a stable state). The graph can be seen below:



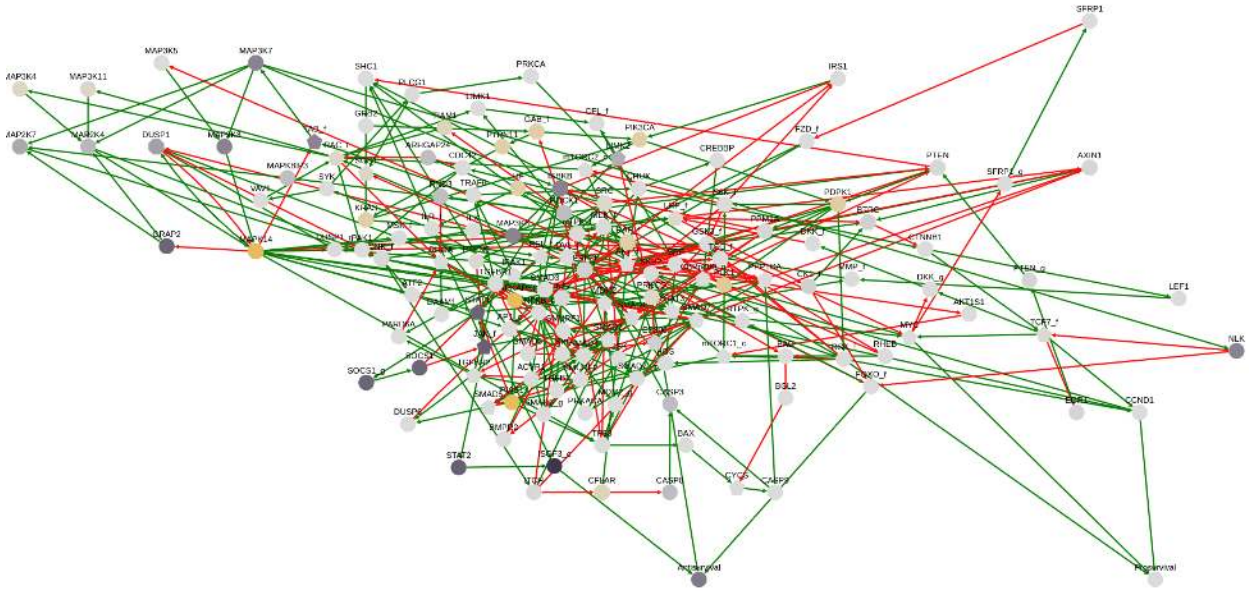
The activation of CFLAR is what causes most of the activation and inactivation differences in the above graph. There also other nodes that have differences but the main cascade of changes happen due to the CFLAR node.

Next, for the predicted synergy AK-D1 (331 models classified it as a non-synergy, 601 models as a synergy) - left in the next graph, we note that the nodes that have different states between good and bad models are actually the same with the AK-BI synergy (right in the next graph):

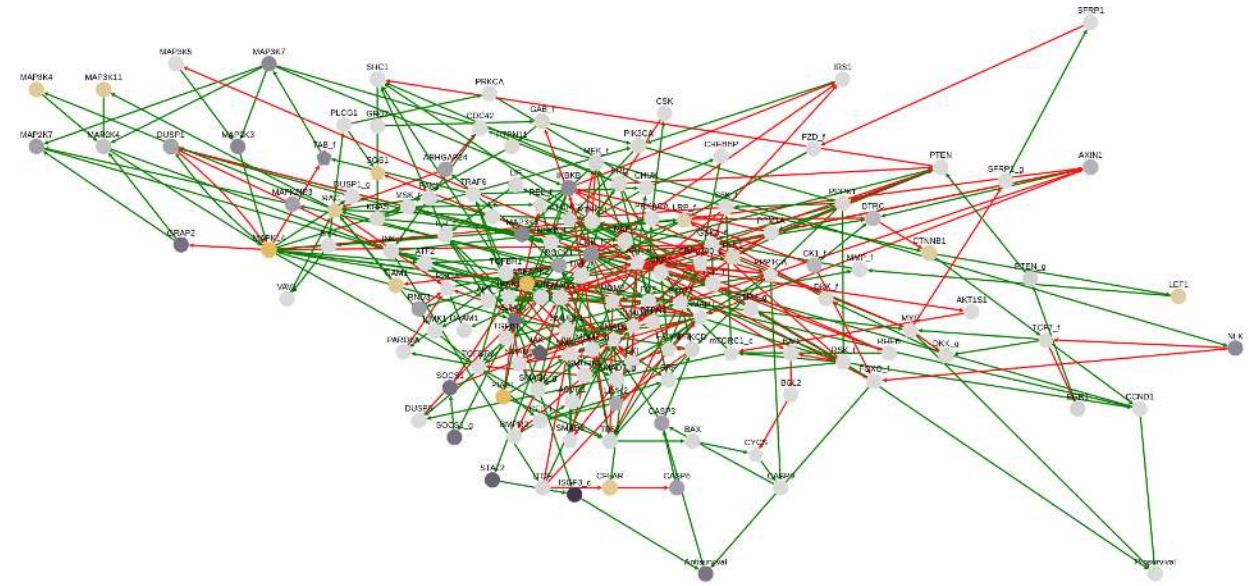


That implies that most of the good models (and bad) in each case are the same: so, 462 models predicted AK-BI and 601 models predicted AK-D1. 384 of these models were the same! As for the non-synergy-predictive models, for AK-BI they were 571, for AK-D1 were 331 and the same ones were: 308! So, on average we see that the good and bad models in each per-synergy classification are the same and so the graphs of stable state differences appear also almost identical.

The same stands true for the next two synergies, they seem to have the same differences in their corresponding graphs:
BI-D1:



PK-ST:

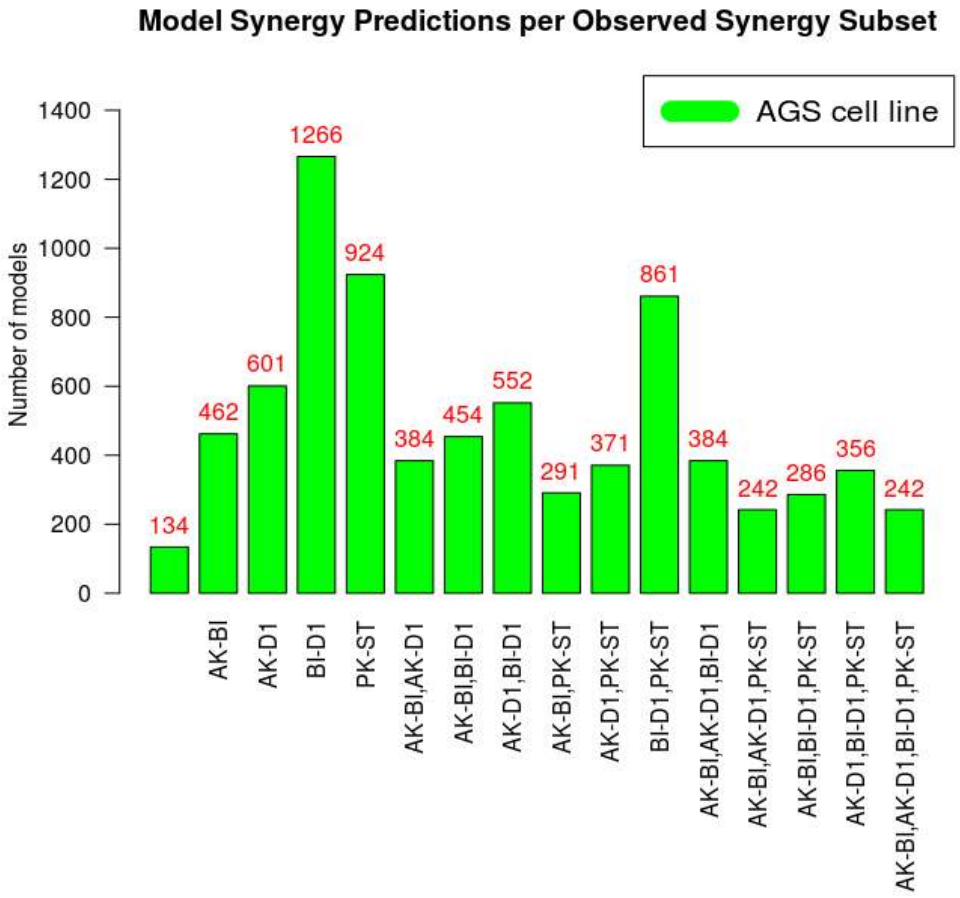


The main difference between the former and latter two groups of synergies seem to be the inactivation of CASP3 (former 2) and inactivation of ISGF3_c (latter 2) in the good models.

3rd way (comparing models predicting different set of synergies)

Bad models: predicted a number of synergies (let's say: A,B,C, where A,B,C are specific drug combinations)
Good models: predicted the synergies from the "bad" models + 1 more => so (A,B,C,D). This way we can find what did the "good" models had that the others did not.

First thing here will be to choose which subset of synergies we want to compare (good vs bad). So, we created a new graph that has info on the number of models that predicted every subset of the total maximum 4 (predicted-by-the-models) synergies:

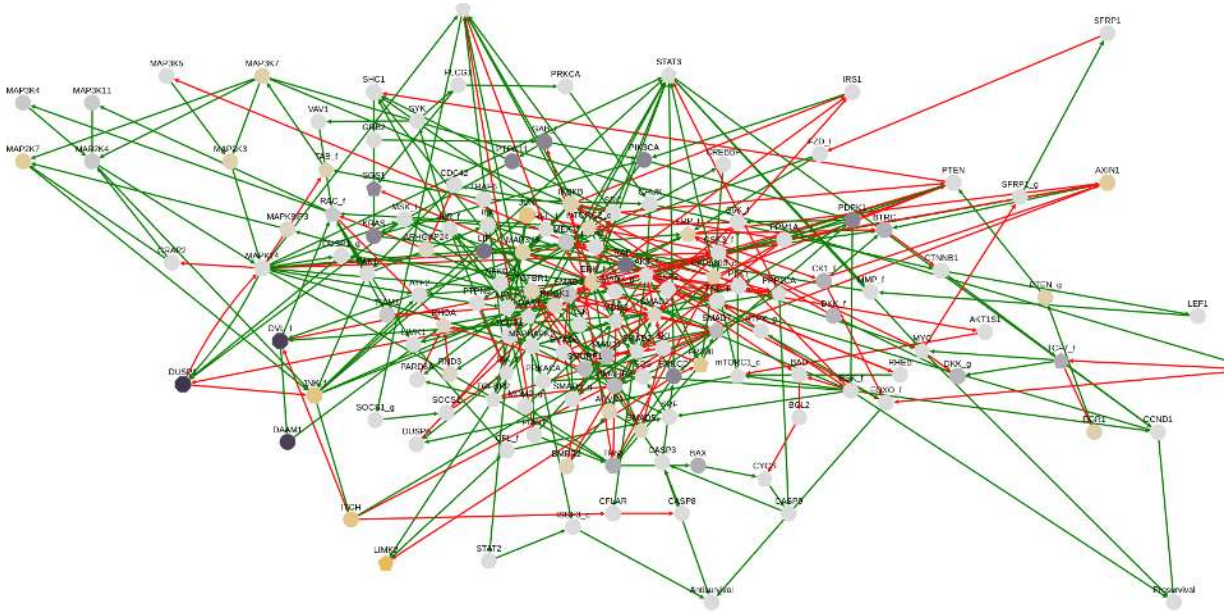


Notes:

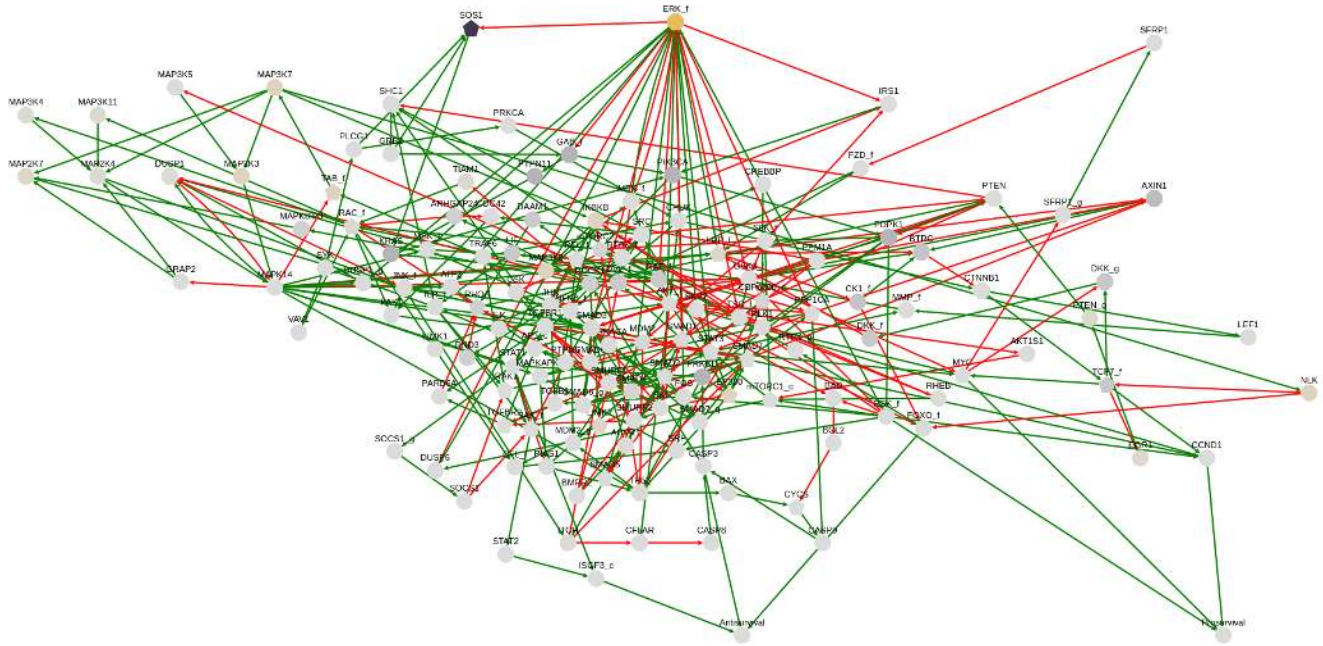
- The synergy BI-D1 was predicted by 1266/1500 = 84% of the models. We can say that generally there are synergies that are predicted more easily than others by the models and this applies to synergy sets as well.
- We note that the number of models (242) that predicted the AK-BI,AK-D1,PK-ST set of synergies is the same as the last one (the one that predicted the maximum - 4 of them):

AK-BI,AK-D1,BI-D1,PK-ST. I checked that actually the last 5 synergy sets have the same 242 models included (some of course have more than 242 models). So, it is reasonable to question what is the difference of the excess models ("bad" models: they couldn't predict the +1 synergy).

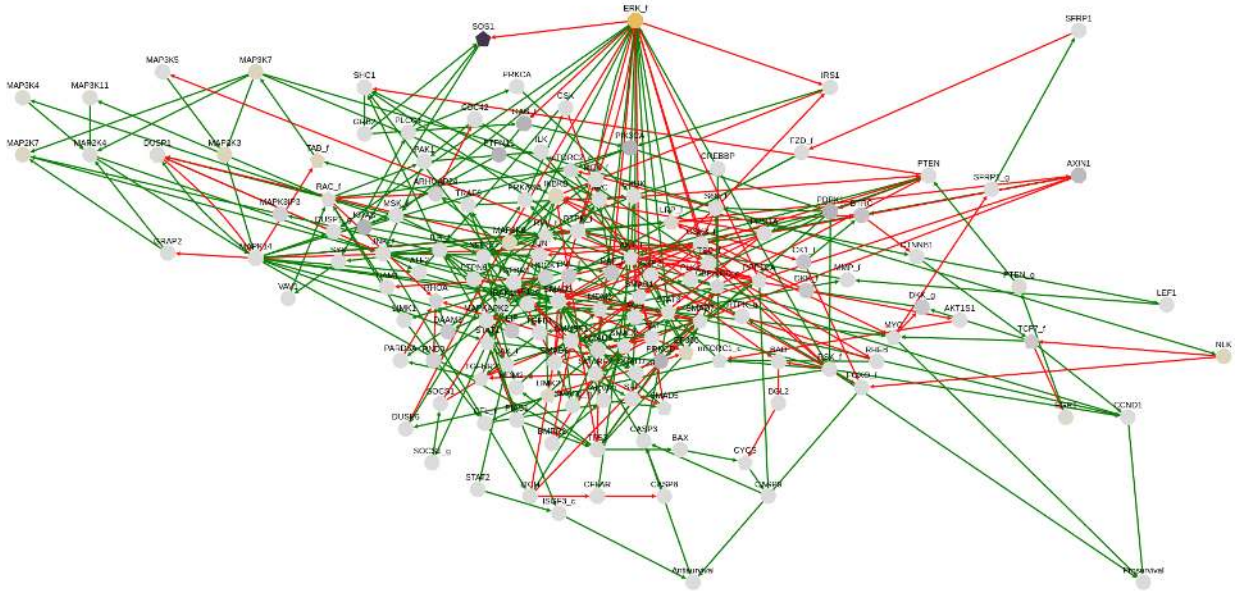
- Comparing: AK-BI,AK-D1,BI-D1 (384 models) vs AK-BI,AK-D1,BI-D1,PK-ST (242 models). 142 models didn't predict the PK-ST synergy. The average difference in stable state values was very small: (max: 0.08 and min: -0.11) making the results inconclusive. The nodes LIMK2, ITCH, JNK_f seem to be more active in the models that also predicted the PK-ST synergy, while the nodes DUSP1, DVL_f, DAAM1 appear more inactive (see next graph):



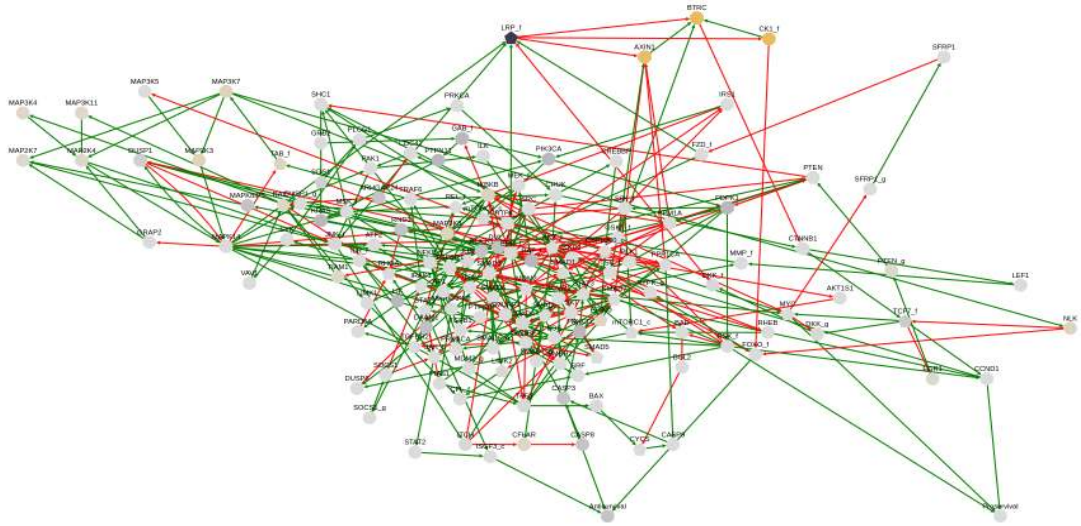
- Comparing: AK-BI,AK-D1,PK-ST (242 models) vs AK-BI,AK-D1,BI-D1,PK-ST (242 models). There are no differences here because these are the same models, which adds to the fact that the BI-D1 is the most easy to predict synergy in our models.
- Comparing: AK-BI,BI-D1,PK-ST (286 models) vs AK-BI,AK-D1,BI-D1,PK-ST (242 models). 44 models didn't predict the AK-D1 synergy. In this case, only two nodes had significant differences: ERK_f (activated, diff ~0.9) and SOS1 (inactivated, diff ~(-0.8)). The SOS1 node is downstream of ERK_f and is inhibited by it, which means that the activation of ERK_f is what makes the "good" models also predict the synergy AK-D1:



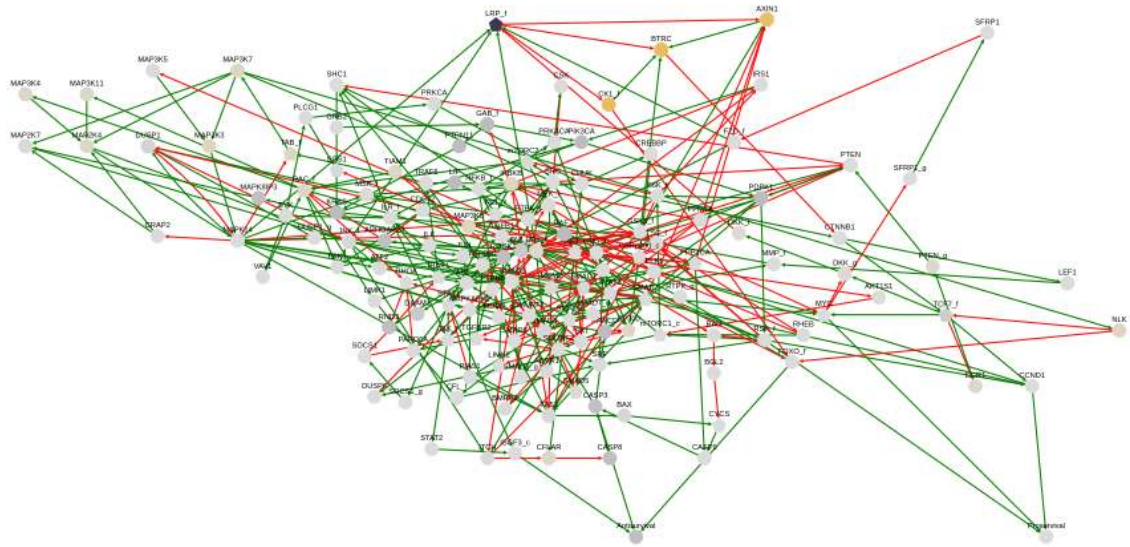
In this scenario we also compare the bad models as the ones that predicted the synergy set AK-BI,PK-ST (291 models) vs the good models that predicted the AK-BI,AK-D1,PK-ST set (242 models). AK-D1 synergy is extra again in the good models but now we are testing when going from a 2-set synergy to a 3-set synergy. The image is the same as before, verifying that the activation of the ERK_f node is indeed important for predicting the AK-D1 synergy:



- Comparing: AK-D1,BI-D1,PK-ST (356 models) vs AK-BI,AK-D1,BI-D1,PK-ST (242 models). 114 models didn't predict the AK-BI synergy. The LRP_f node is inactivated in the "good" models and seems to be the primary reason that made these models predict the AK-BI synergy:



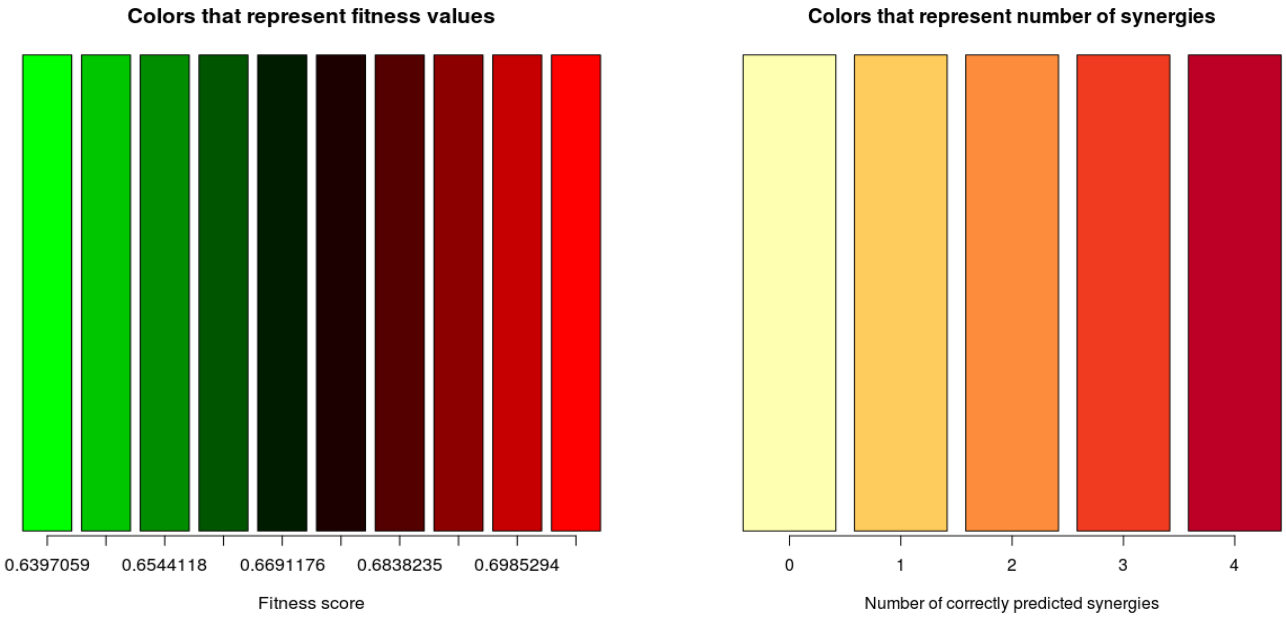
In this scenario we also compare the bad models as the ones that predicted the synergy set AK-D1,BI-D1 (552 models) vs the good models that predicted the AK-BI,AK-D1,BI-D1 set (384 models). AK-BI synergy is extra again in the good models but now we are testing when going from a 2-set synergy to a 3-set synergy. The image is the same as before, verifying that the inactivation of the LRP_f node is indeed important for predicting the AK-BI synergy:



Heatmap of models stable states and equations

Using as input the stable state for each of the 1500 models, we produce a heatmap (white is activated, red is inactivated). We also add two colored columns next to it, that represent each model's fitness value and how many synergies the model correctly predicted.

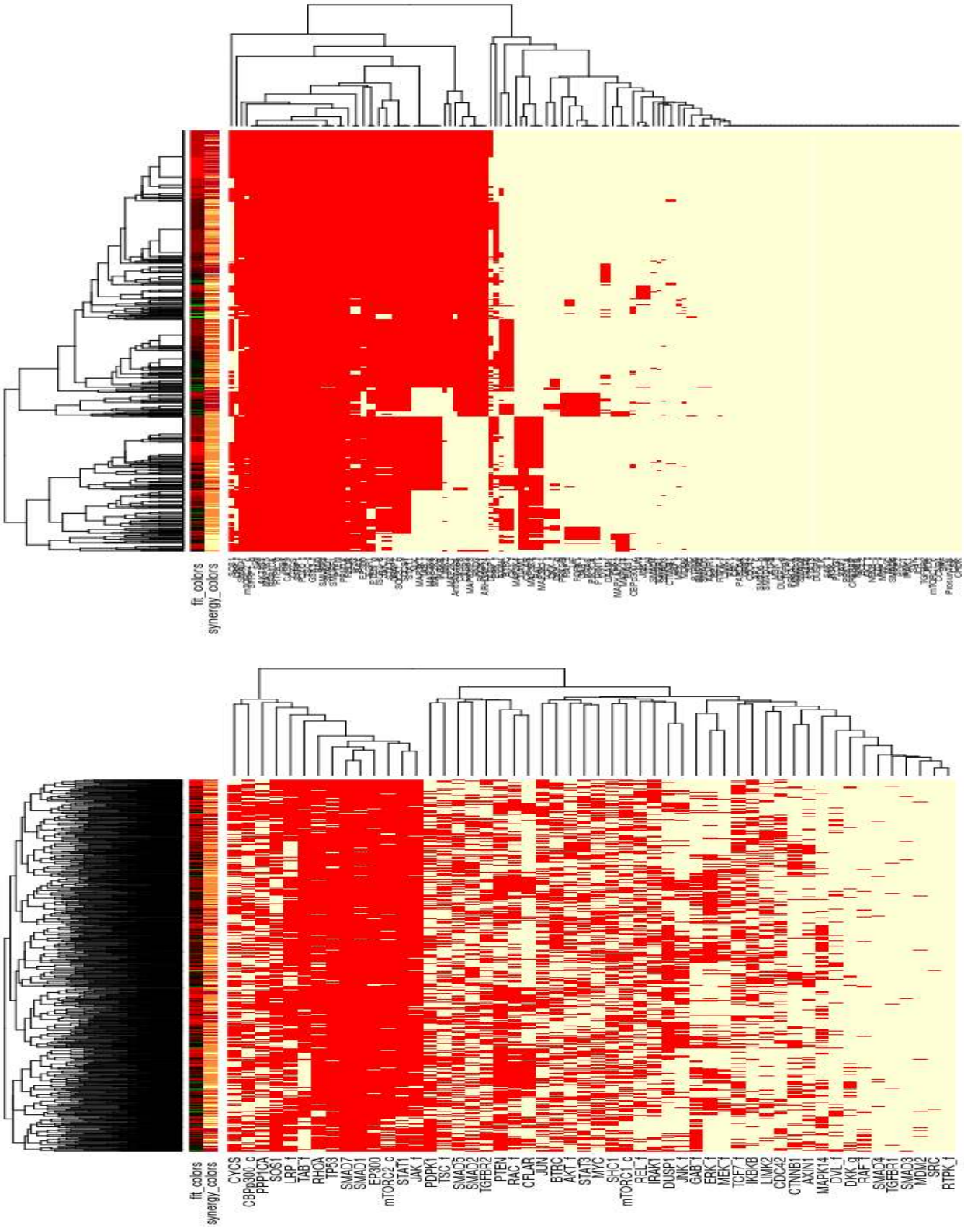
The colors that are used are:



The first heatmap in the next page show us that there a small clusters of stable states where the stable states themselves are no different (it is a cluster of almost common values after all) and on these clusters if you have high-predictive models there is a high probability you also have high fitness scores as well (and vise-versa). The fitness values also group into clusters as well (the stable state values directly correlate with the fitness scoring value that's why). But the general image does not show any correlation between stable states and high predictive performance or fitness scoring I believe.

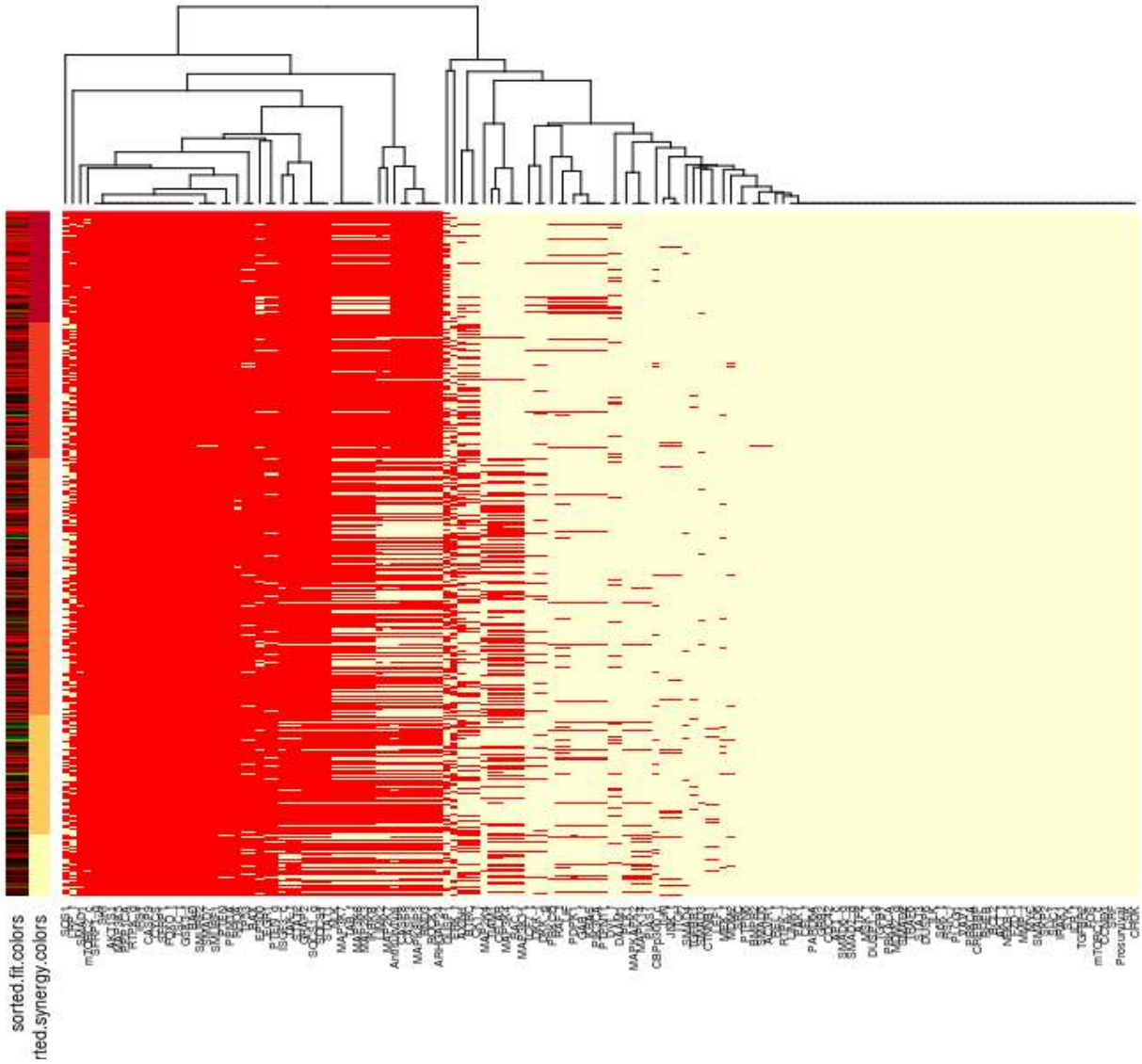
The second figure of the next page is the heatmap of each model's equations. Each model is represented by 144 boolean equations (as the number of nodes), where the only difference between 2 models' same node equation is in the link operator. In the heatmap the values represented are: white for an "or not" link operator and red for an "and not" link operator (if there is no link operator - the equation has only inhibitors or activators - then the equation is not shown in the heatmap). In the equation heatmap we can see that the equations differ more than

the stable states (many more row clusters found) and the synergy prediction and fitness values are much more dispersed (so no correlation here).



Next, we try to find something more meaningful, if we impose a specific row ordering in the stable states dataset. The choices can be two: order the dataset rows by number of synergy predictions or by the fitness score.

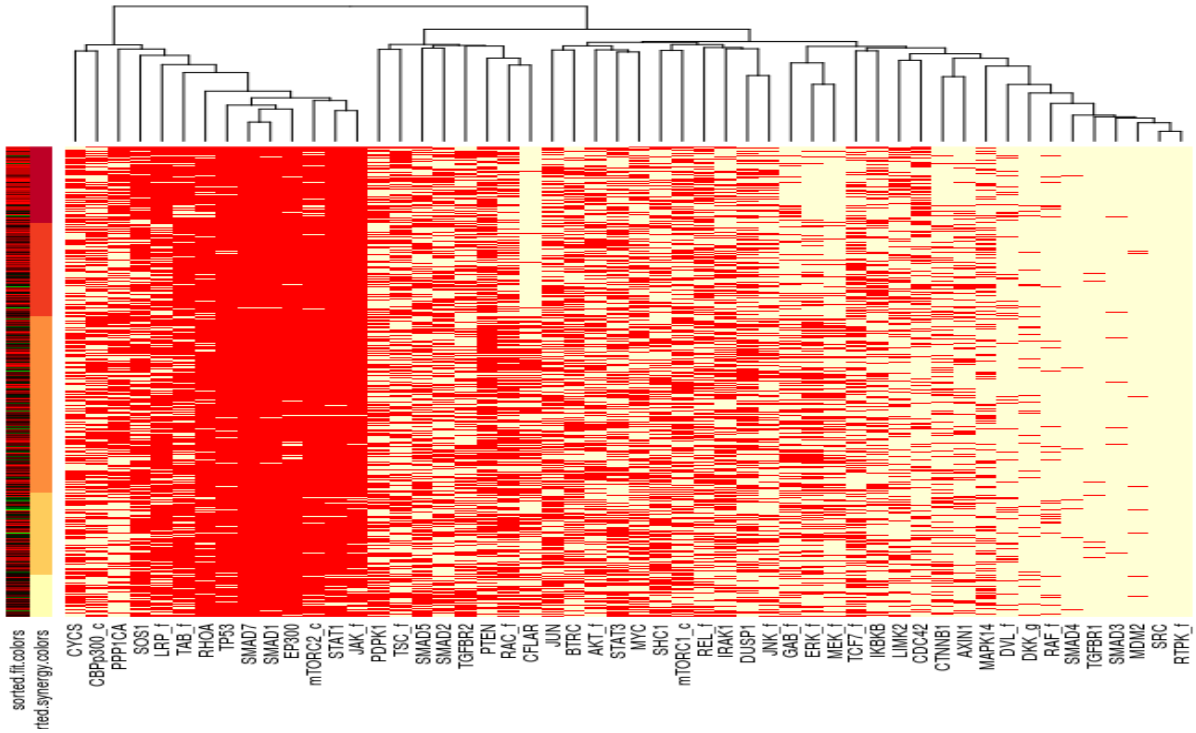
- If the order is done by the number of predicted synergies, we get the following stable states heatmap:



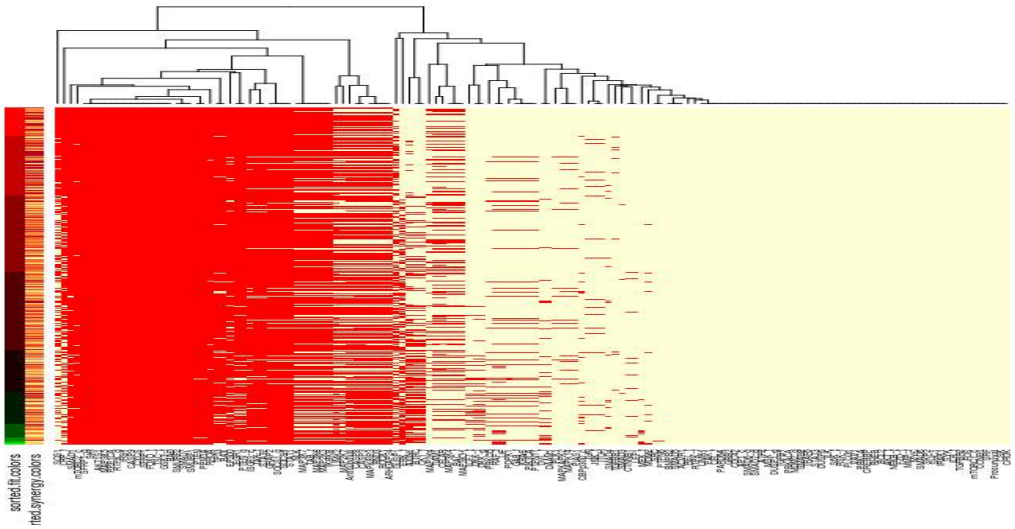
The above heatmap shows us that the high vs low predictive models have actually some correlation (on average) with the value of their corresponding stable states. It seems that specific nodes become activated while others at the same time (a near-by cluster of nodes) become inactivated when transitioning from high to low performance models (CFLAR for example is activated in higher performance models and mostly inactivated in the low-performance ones as is ERK_f). The fact that the fitness values are all over the place may mean that their differences are not that much important (makes sense since they are all the

product of 20 generations of fitting to the stable state and so they are pretty much the same values). But, we could say that higher predictive models have on average higher fitness values (also seen before in the density plots) that that means that stable states and their fitnesses correspond to behaviour (predictive performance).

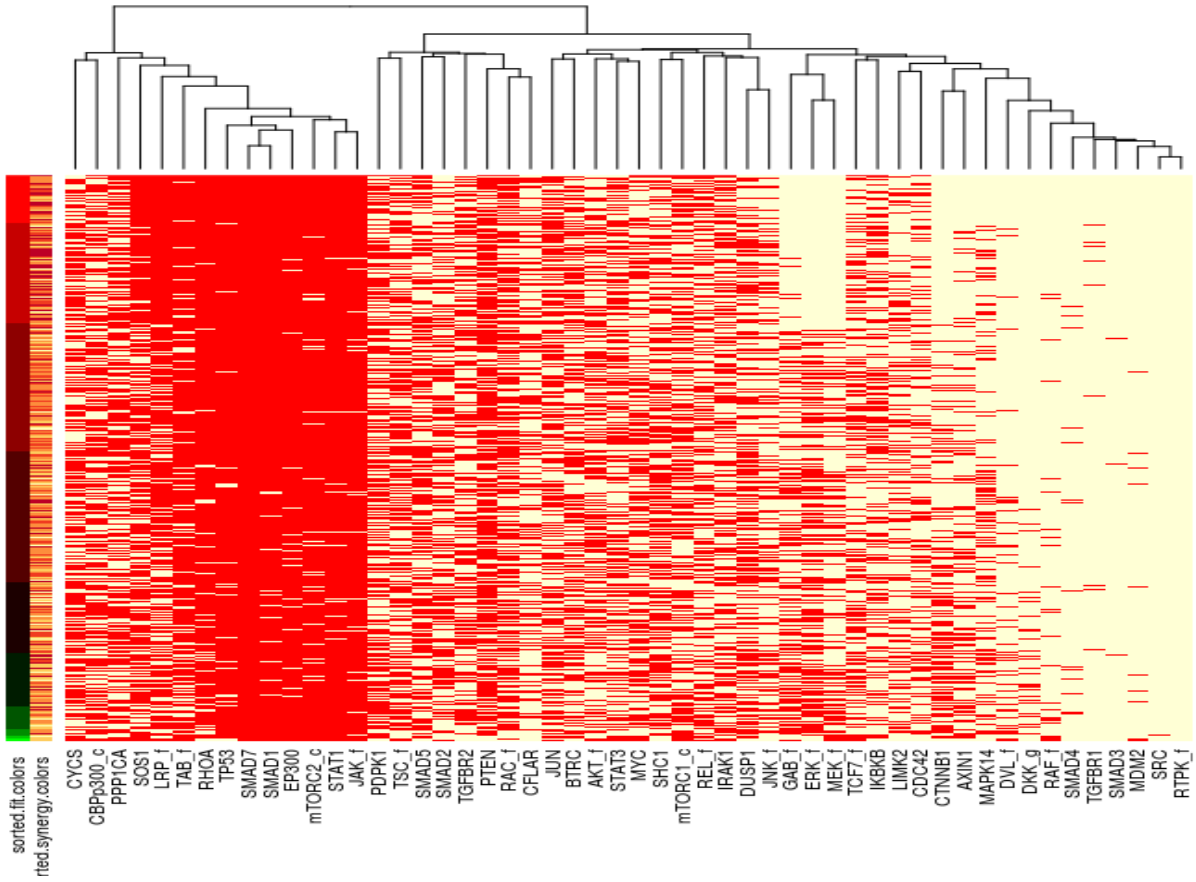
The next heatmap of equations, which is synergy-sorted as well, doesn't show any useful correlation between the depicted values expect from CFLAR, which for high-performance models it needs to have in its corresponding boolean equation the "OR" link : CFLAR *= (AKT_f) OR NOT (ITCH). Note that the ERK_f has the OR link in the high performance models: ERK_f *= (MEK_f) OR NOT ((DUSP6) OR PPP1CA)



Next, we reorder the stable states by fitness scoring:



Note that the node ERK_f is mostly activated in the higher fitness models. For the equations (see below) we see that there are 3 target nodes (ERK_f, MEK_f, GAB_f) that have “or links” in the highest fitness:



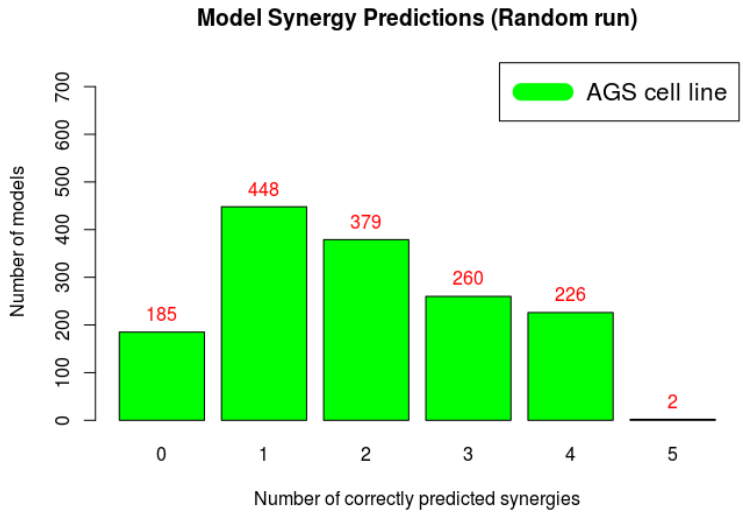
A Result: So, we have seen that the node ERK_f in the highest performance models and in the highest fitness ones, has to be activated and have the OR link respectively.

AGS (random run: fitting to proliferating state)

In this section, we will run all the above analysis on the “random” AGS run mostly to compare with the normal run’s results. Input files can also be taken from the web server URL: http://localhost:8080/SINTEF/ROC/AGS_0dcc89b39f3527f93e3e9c376a140398/info

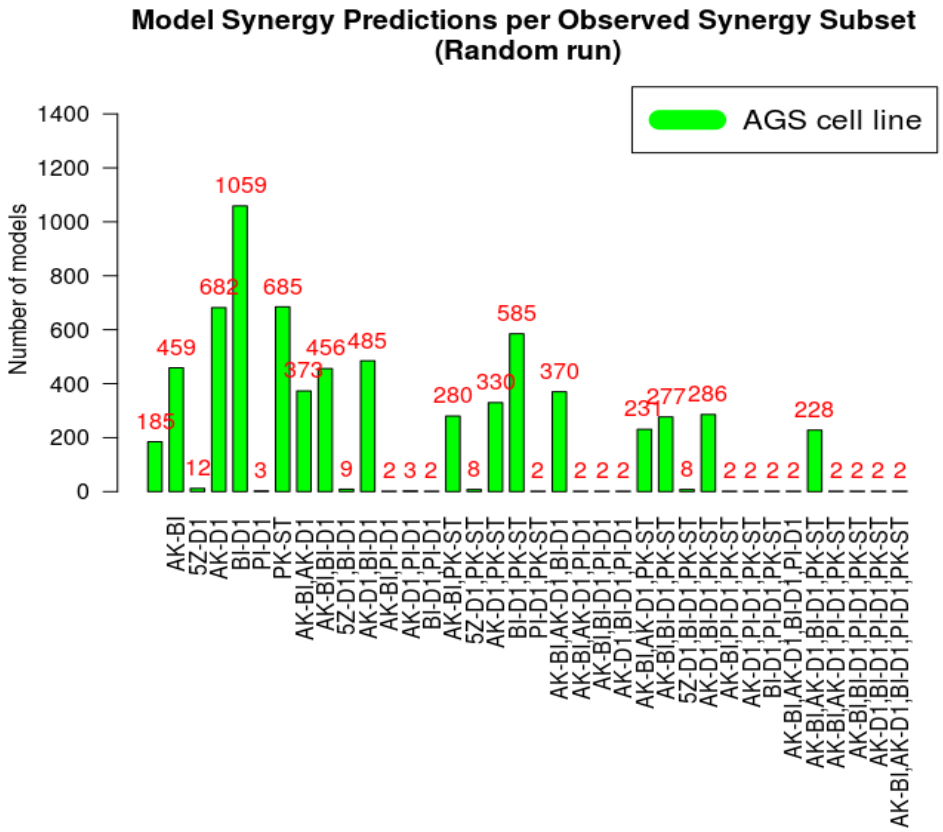
Counting number of models vs max corrected predicted synergies

There are a total of 6 experimentally predicted synergies (same cell line didn’t change from before), namely: "AK-BI" "5Z-D1" "AK-D1" "BI-D1" "PI-D1" "PK-ST". All of these synergies were predicted by the random models (in the AGS normal run no models could predict two synergies, namely "5Z-D1" and "PI-D1"). But, as we can see in the next graph, there wasn’t a model that could predict all 6 of them at the same time, although there existed 2 that could predict 5 of them:

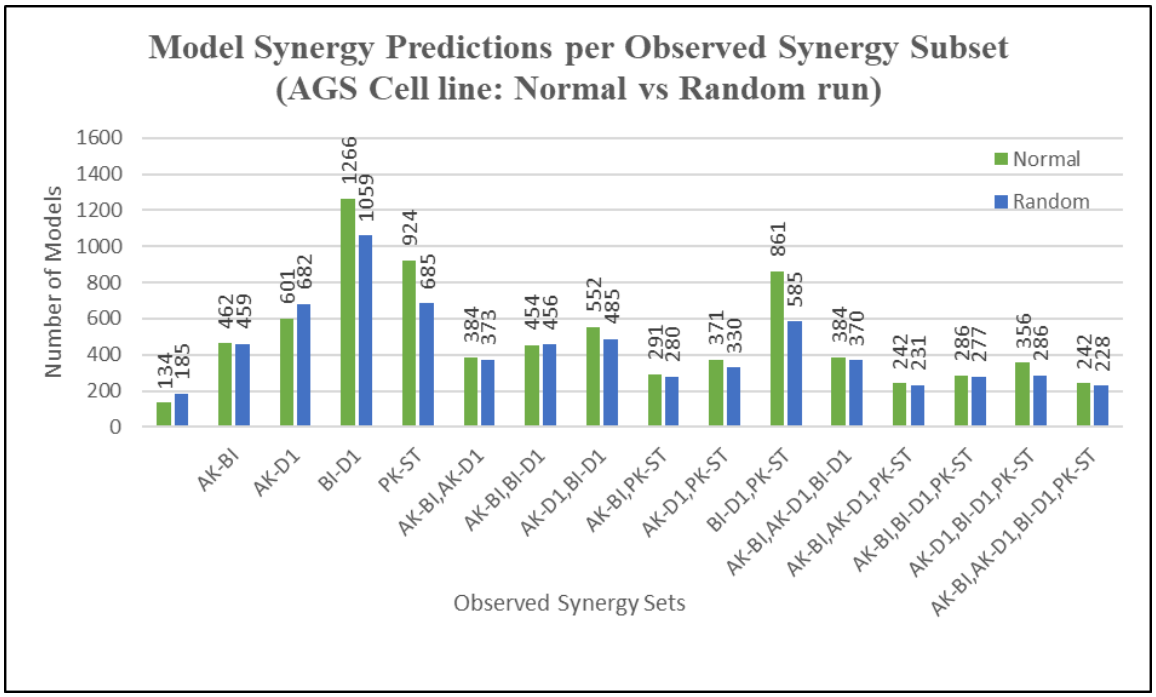


Comparing with the normal AGS run results, there are more models here that predict no synergies (185 vs 134) and less the predict more than 4 (226 vs 242).

It will be more informative for comparing reasons to show the number of models that predicted each synergy subset (As we did in the 3rd way of the section “Good vs Bad models” in the normal AGS run). Note though that we don’t include in the next graph the synergy sets that were not predicted by any model - 28 out of 2^6 = 64):



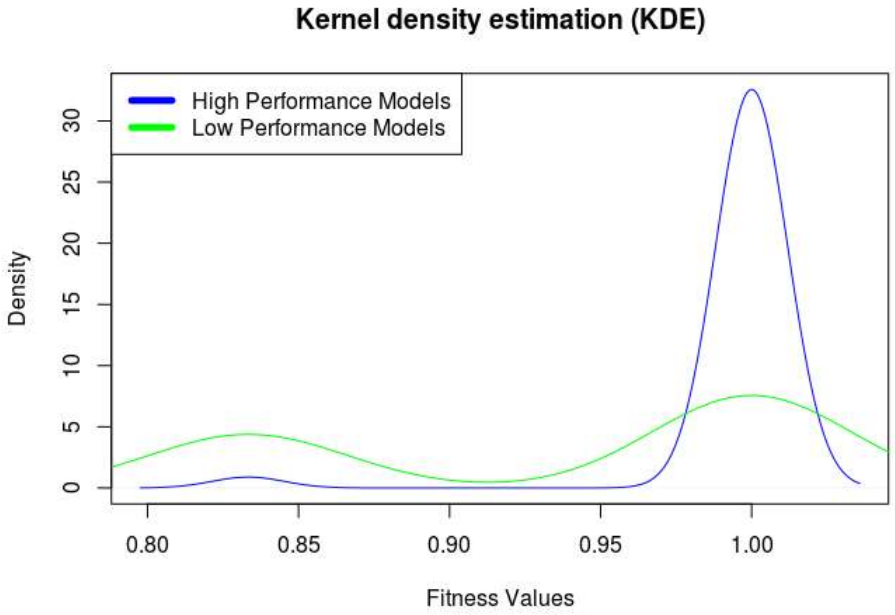
So we see that <u>the models that predicted the 2 synergies "5Z-D1" and "PI-D1" were indeed few (1%) as well as the synergy sets that included those synergies</u>. So if we exclude them from our analysis and also include the results from the AGS normal run we have:



So, we have more models predicting the above synergy sets in the normal run than in the random one.

Fitness values of high vs low performance models: statistical correlation

Out of the 1500 models, 1107 had a fitness of 1, 14 of them had 0.667 and the rest 379 models had a fitness of 0.8334 (so a total of 3 fitness values). So, both in the high and low performance models (predicted up to 3 synergies or more than 3: that's the distinction) you have a lot of them that have a fitness value of 1:

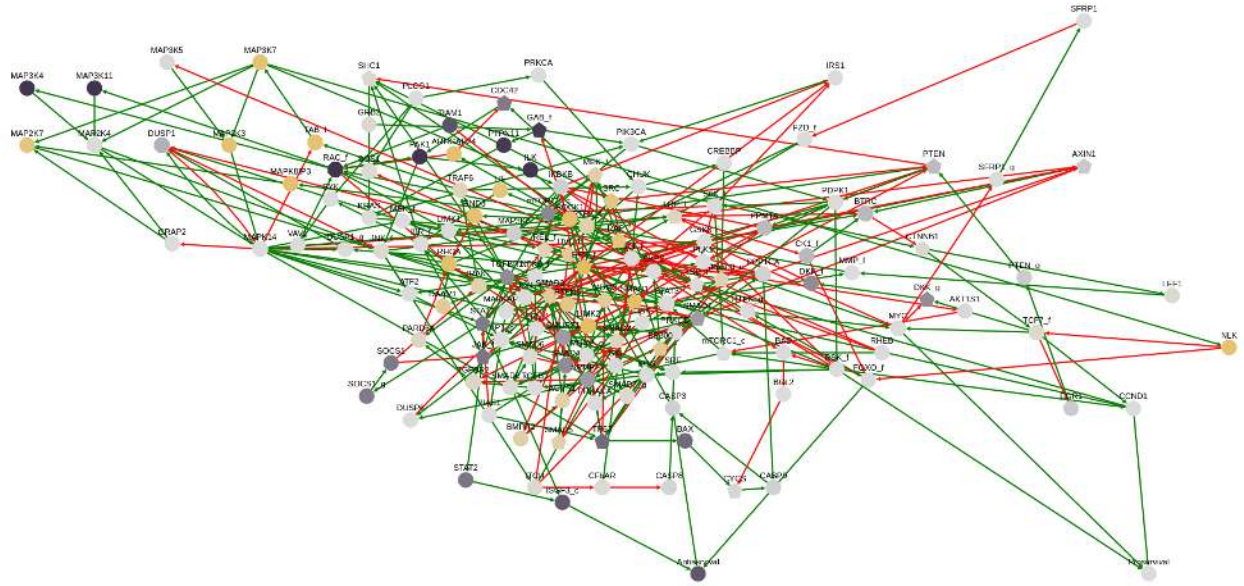


I don't think we can make something meaningful out of this when the fitness value space is only 3 values and most of them 1. This means that most of the models fit to the proliferating/growing state.

Good vs Bad models (average stable state differences)

1st way (High performance vs low performance models)

Bad models: they predicted no synergies (185 models)
Good models: they predicted the maximum number of synergies (5 for AGS) (2 models)

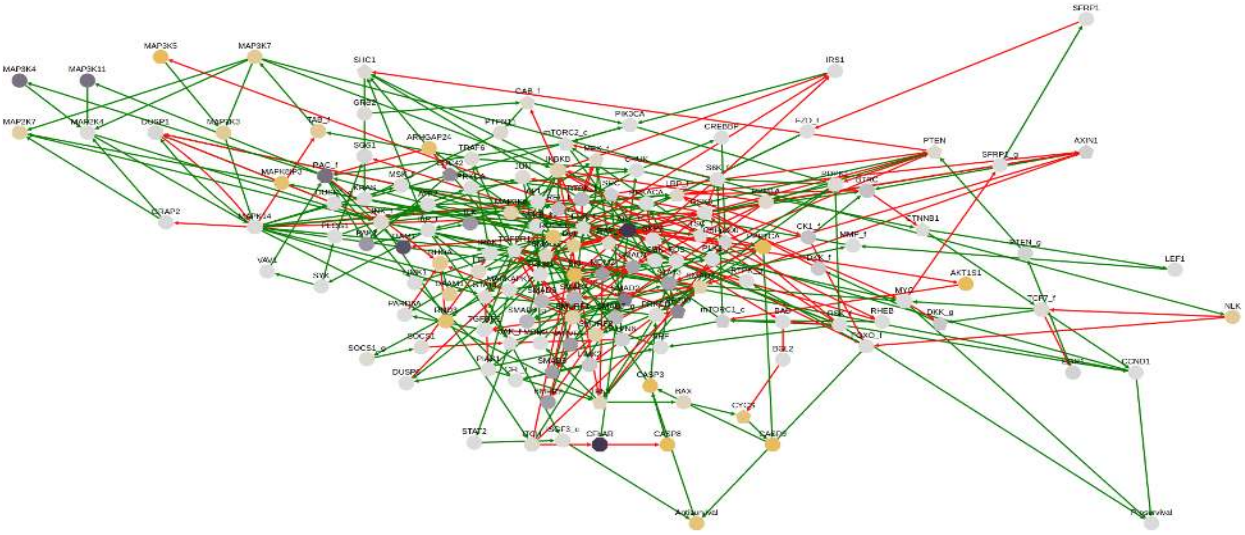


Many nodes' states are different. More inactive Antisurvival in good models, too many differences to make sense out of it.

2nd way (find what mechanisms a particular synergy is exploiting)

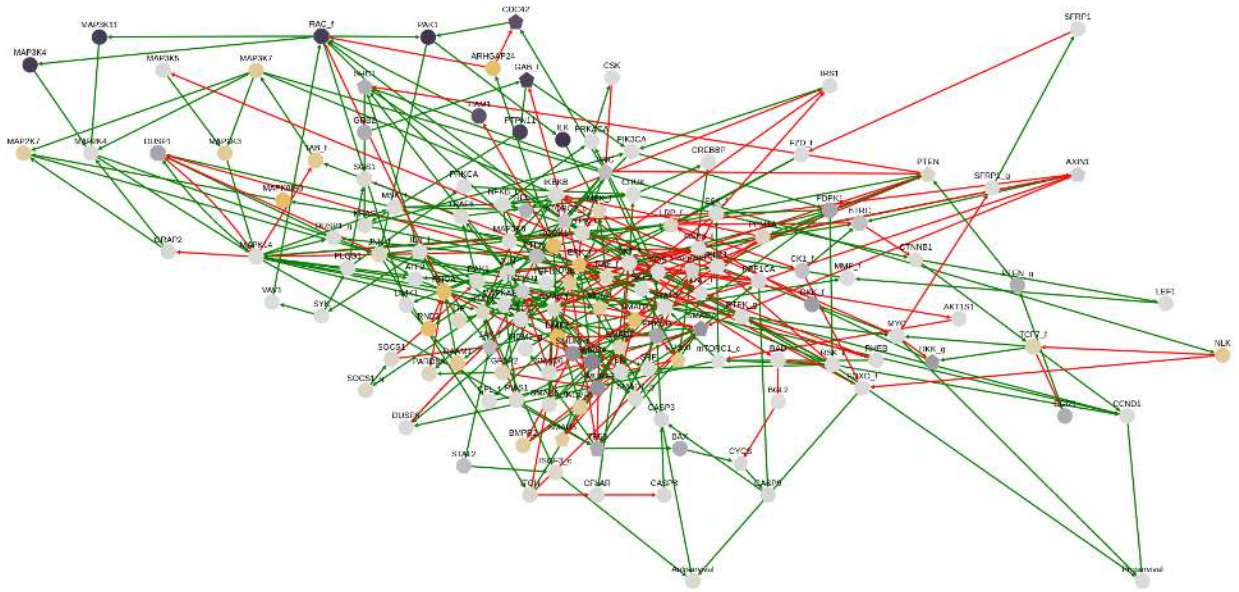
Bad models: they didn't predict a specific synergy (may have predicted others)
Good models: predicted that specific synergy

Here it would be interesting to check what was the difference for the models that predicted the two extra synergies that the normal AGS models couldn't predict. Note that in the AK-BI and AK-D1 synergies there were small differences (in scale) between good and bad models' average stable states so we don't output the graphs for them.
The 5Z-D1 synergy:



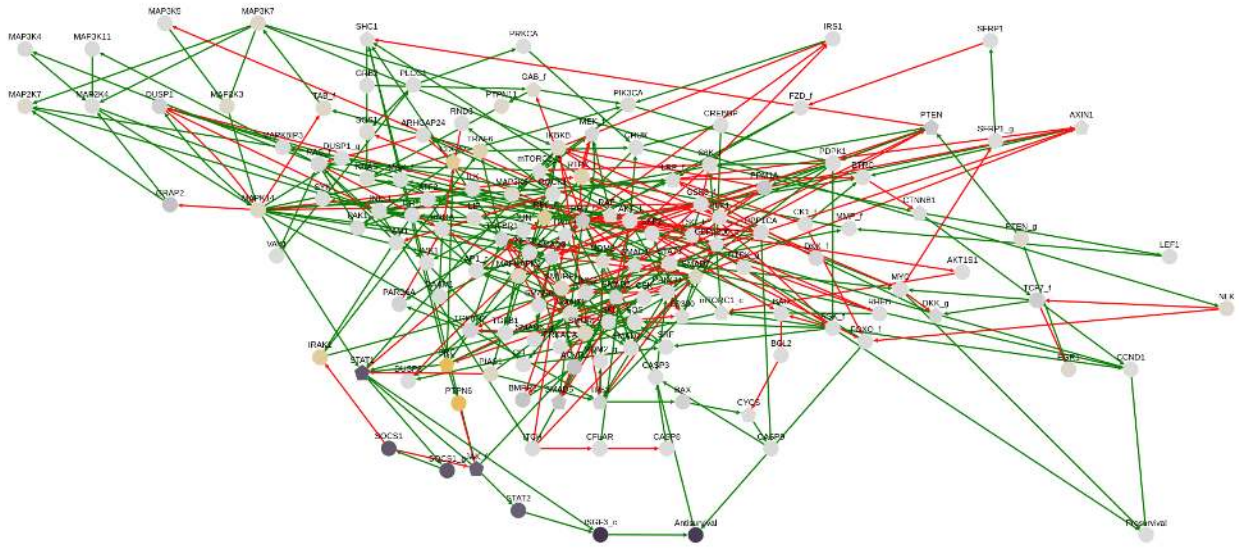
There are a lot of differences but we could say that the main ones are the AKT_f and CFLAR inactivation in the good models.

The PI-D1 synergy:



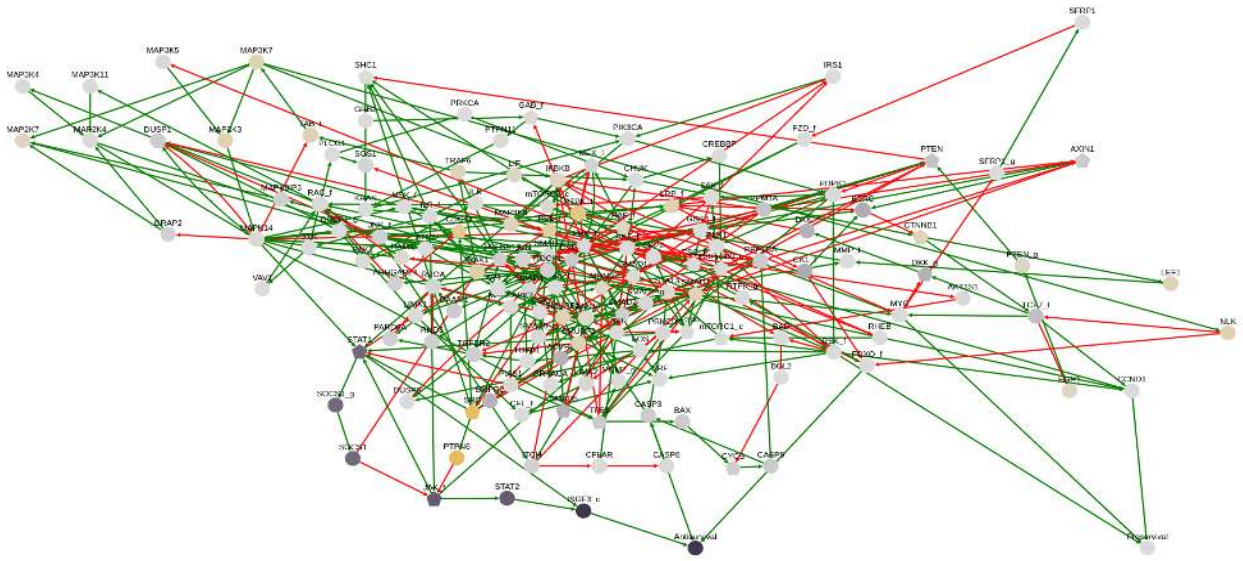
Also, very different behaviour from the average case: most important is the inactivation of RAC_f and the activation of ERK_f in the models that predicted the PI-D1 synergy.

The BI-D1 synergy:



The SRC activation is the most important here which causes the cascade of inactivations that lead to the inactivation of the ISGF3_c.

The PK-ST synergy graph image looks the same as the BI-D1 synergy's though the difference values are smaller (this similarity was also true in the normal case). That happens because a lot of these "good/bad" models are the same: the BI-D1 synergy was predicted by 1059 models, while the PK-ST synergy by 685 models. 585 of these models were the same. Also, 356 models didn't predict the BI-D1 synergy and 199 didn't predict the PK-ST. Of those, 157 were the same.

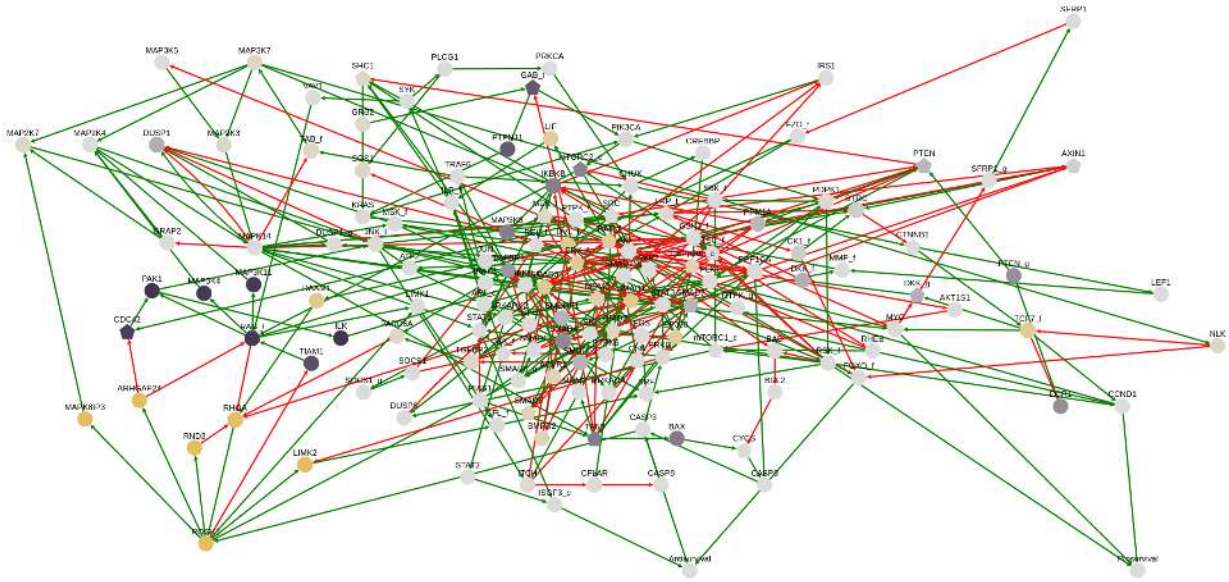


3rd way (comparing models predicting different set of synergies)

Same logic as before: we are going to test the differences between the 3-set synergies vs the 4-set (AK-BI,AK-D1,BI-D1,PK-ST) and this 4-set synergy vs the 5-set (the 2 models that can predict all the previous 4 synergies + the PI-D1 synergy).

- Comparing: AK-BI,AK-D1,BI-D1,PK-ST (228 models) vs AK-BI,AK-D1,BI-D1,PI-D1, PK-ST (2 models) (PI-D1 synergy difference). So, 226 models couldn't predict the extra synergy. The max and min average stable state difference values were 0.86 and -0.97 respectively. In the next graph we can see that the main difference in these two models was the activation of ROCK1 that caused also a cascade of activations and inactivations in other nodes of the network (see bottom-left part of the graph). ROCK1 was also more activated in the good models that predicted this synergy vs the rest of them that couldn't predict it (see "The PI-D1 synergy" graph in the corresponding section). But there are

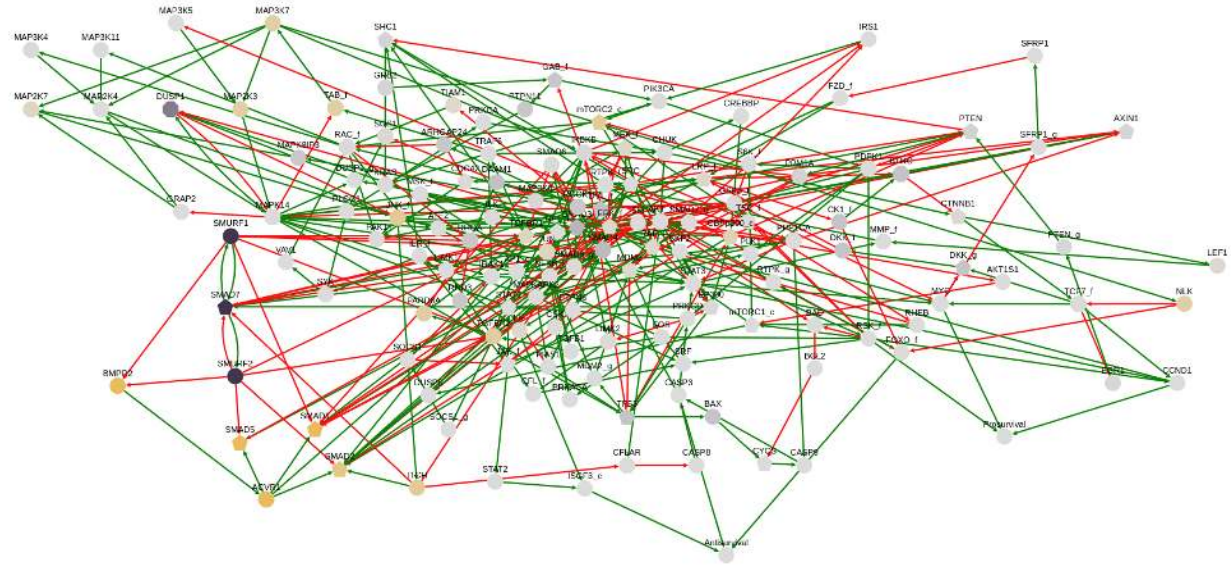
many other nodes that slightly differ between good and bad models so I don't think we could make something conclusive out of it. The main reason for this is that the good models are represented by a very low sample (just 2 of them).



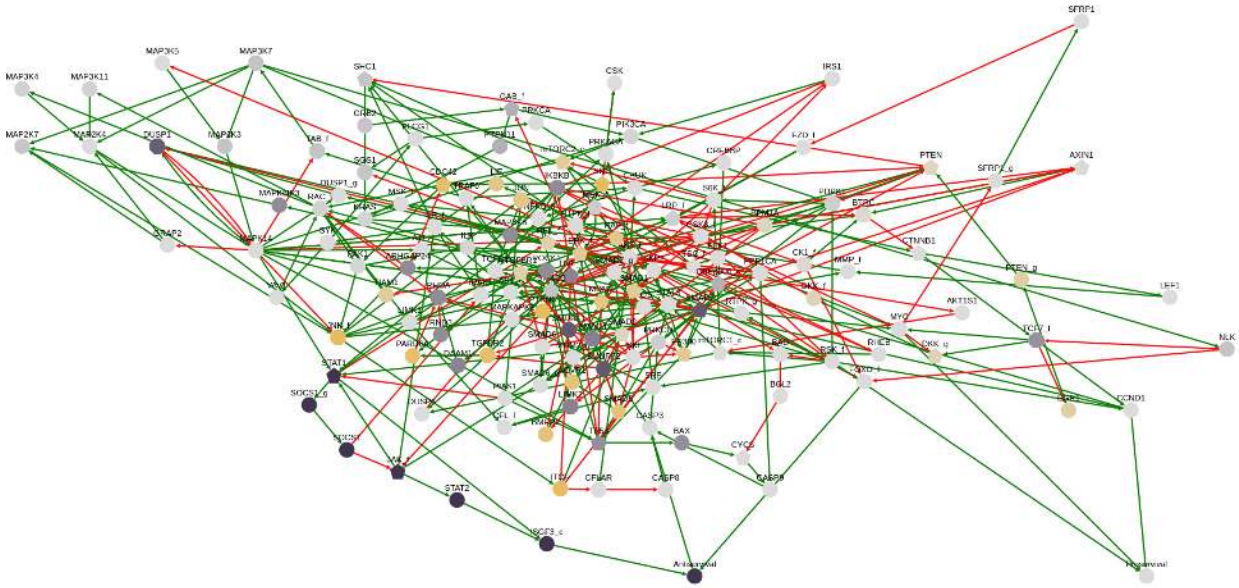
Another meaningful question here is why the AGS normal run models couldn't predict the PI-D1 synergy (here we had 3 models that could, while 2 of them could predict 4 more synergies as well): we test this by taking an average of the stable states of the high performance models (242 models predicted the max 4 synergy-set in the AGS normal run) and see the state of the ROCK1 node: 0.012 => it is actually inactivated in these models! Also, in the training data it is activated (steady state). This means that:

1. The best predicting models from the normal AGS run had to have the ROCK1 node deactivated though it seems here it has to be activated for the extra PI-D1 synergy to be predicted.
2. While the best performing models in the AGS normal run had also (as we shown before) a higher fitness score on average, which means that they fitted more to the steady state, that doesn't mean that all nodes will be fitted: ROCK1 has to be deactivated for the 4 synergies to be predicted in both the normal and the random run.
3. The presence of these 2 high-performance models in this case could mean that some very specific (and probably rare) formations of the boolean equations happened.
4. All in all, it seems to me that some synergies are antagonistic: you cannot have many models predicting a higher set of synergies because the extra synergies need different states in the nodes that are responsible for predicting the others and the only way to get the extra synergy will be if you add more kinds of mutations to the boolean equations or have a larger ensemble of models and be lucky to have models that somehow predicted also the extra synergy(ies).

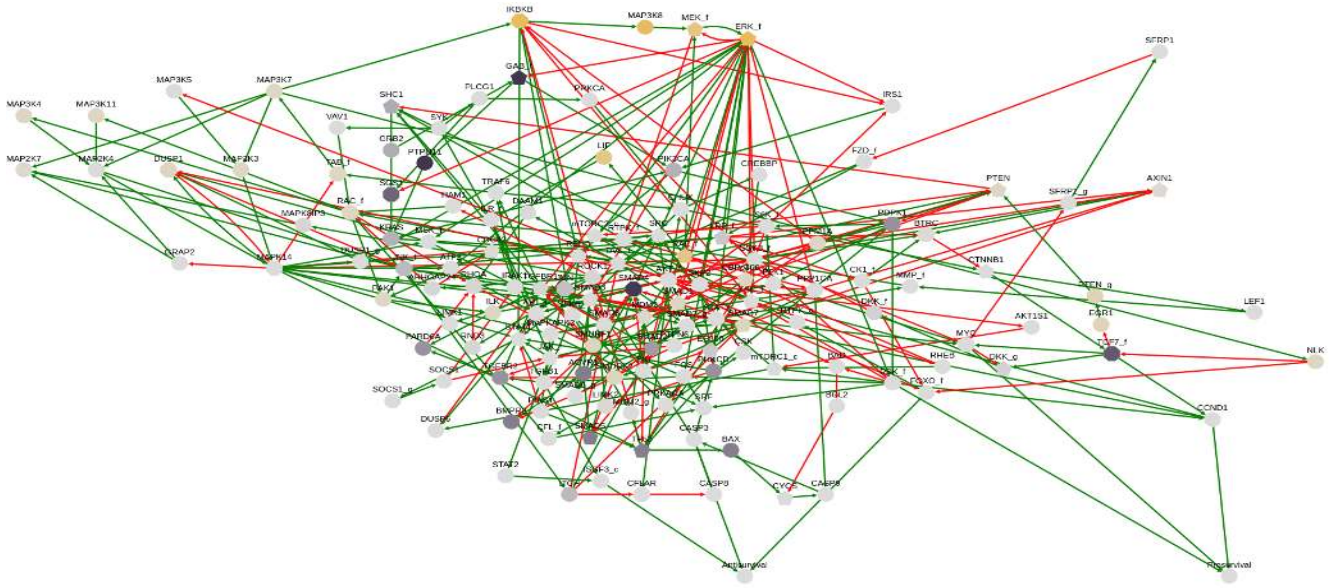
- Comparing: AK-BI,AK-D1,BI-D1 (370 models) vs AK-BI,AK-D1,BI-D1,PK-ST (228 models) (PK-ST synergy difference). So, 142 models couldn't predict the extra synergy. The max and min average stable state difference values were 0.35 and -0.35 respectively. Since the average differences are not so great we cannot have conclusive results, though looking at the next graph we notice that the nodes that had the largest changes were those that were affected by the inactivation of SMAD7:



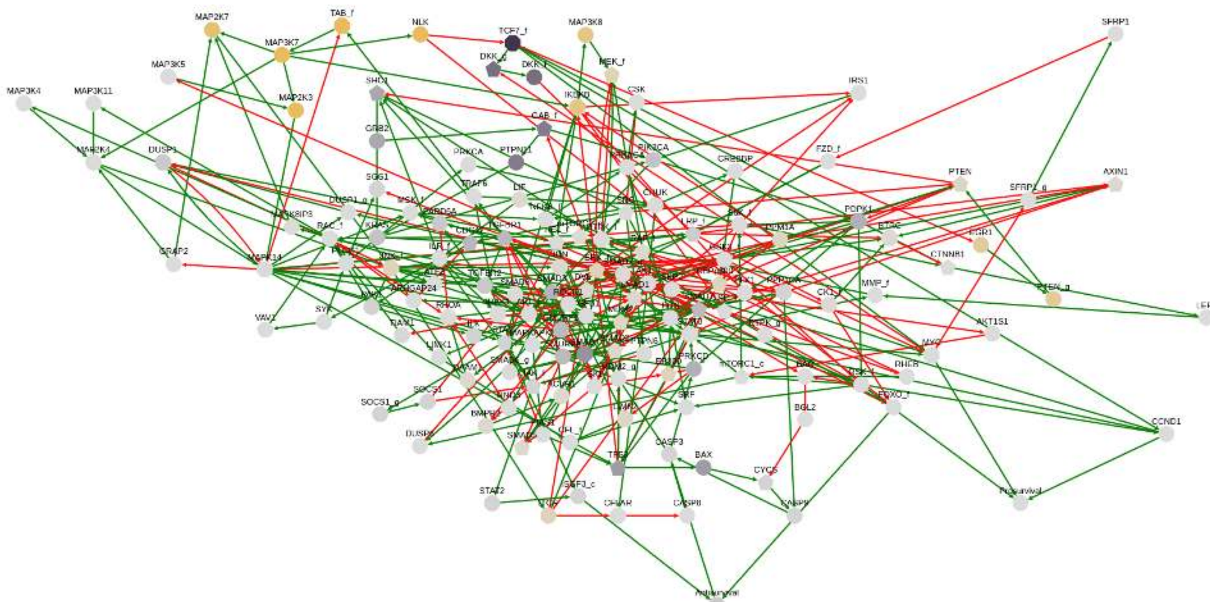
- Comparing: AK-BI,AK-D1,PK-ST (231 models) vs AK-BI,AK-D1,BI-D1,PK-ST (228 models) (BI-D1 synergy difference). So, 3 models couldn't predict the extra synergy. The max and min average stable state difference values were 1 and -1 respectively. Many nodes have differences, so no conclusive result here either:



- Comparing: AK-BI,BI-D1,PK-ST (277 models) vs AK-BI,AK-D1,BI-D1,PK-ST (228 models) (AK-D1 synergy difference). So, 49 models couldn't predict the extra synergy. The max and min average stable state difference values were 0.52 and -0.19 respectively. As in the case of the normal AGS run, the activation of the ERK_f node (0.52 difference) was the main difference that the good models had and allowed them to predict the AK-D1 synergy. But as it can be seen from the following graph there are more nodes that differ between the good and bad models (though the inactivated ones have small average differences and so the main ones would be those that are more activated, that is ERK_f and IKBKB:

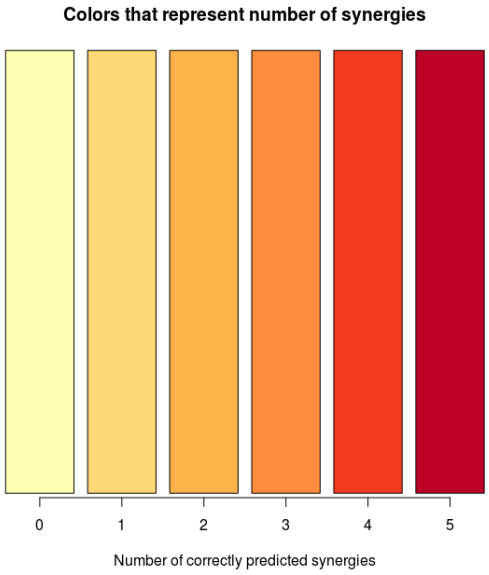


- Comparing: AK-D1,BI-D1,PK-ST (286 models) vs AK-BI,AK-D1,BI-D1,PK-ST (228 models) (AK-BI synergy). So, 58 models couldn't predict the extra synergy. The max and min average stable state difference values were 0.65 and -0.33 respectively. TAB_f seems to be the main node whose activation that causes a cascade of other inactivations/activations (though we cannot make any further useful assumptions here either):

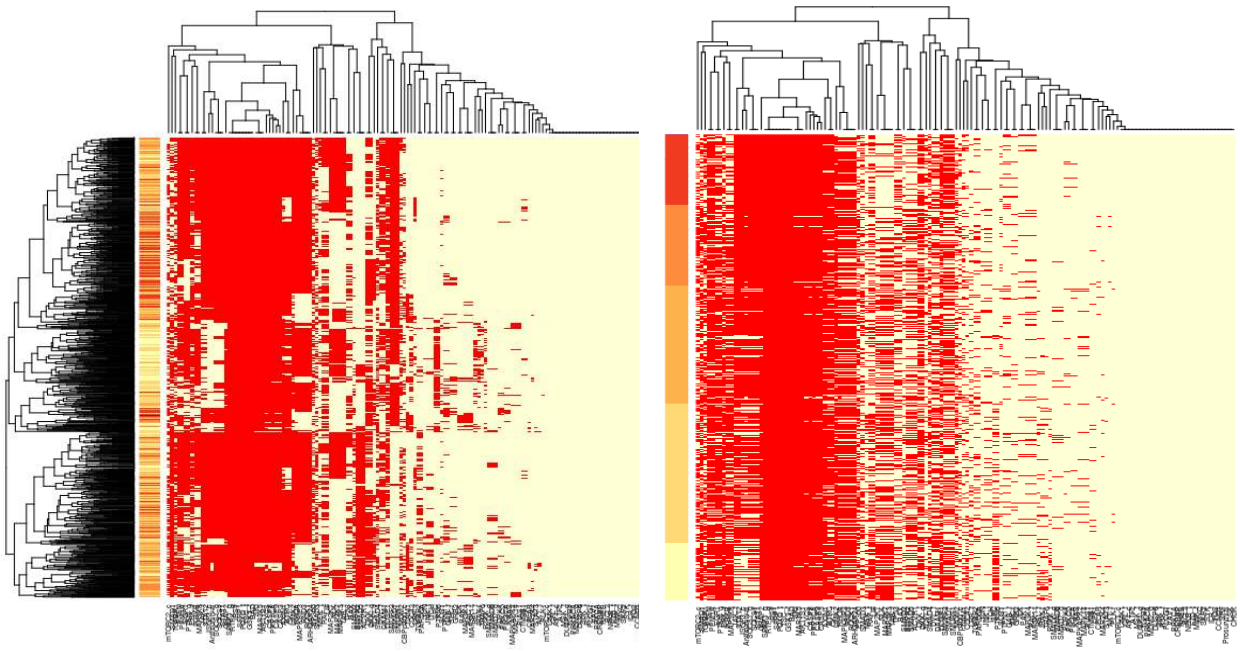

MAP2K7
TAB_f
NLK
TCF7_f
MAP3K8
SFRP1
MAP3K7
DKK_g
DKK_1
MEK_f
MAP3K5
IRS1
MAP3K4
MAP3K11
SHC1
CSK
MAP2K3
IKBKB
GRB2
CAB_f
FZD_f
MAP2K4
DUSP1
PRKCA
PTPN11
PIK3CA
CREBBP
PTEN
AXIN1
SGG1
TRAF6
PDPK1
EGR1
CTNNB1
GRAP2
MMP_f
PTEN_g
LEF1
VAV1
AKT1S1
SOCS1_g
SOCS1
CCND1
DUSP6
STAT2
CASP3
BAX
CYCS
BCL2
CASP8
CFLAR
SGF3_c
mTORC1_c
PRKCD
RHEB
FOXO_f

Heatmap of models stable states and equations

In the next heatmaps we will not include the fitness scoring of the models since there were only 3 distinct values and more than 75% of the models had a fitness score of 1. The colors for the models' synergies are:

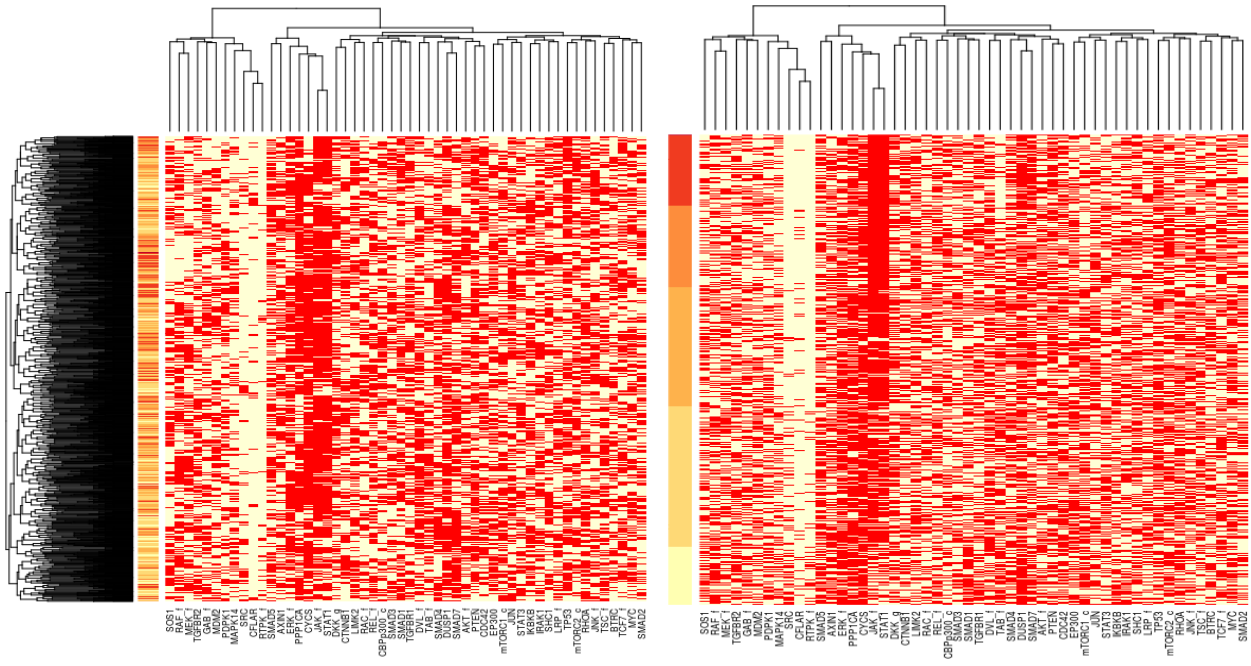


Heatmaps of stable states (cluster-based and synergy based):



The only useful insight here is that if you compare the high vs the low synergies areas in the heatmaps above you will see that there are cluster of nodes that on average change values when going from high to low models and vice-versa (e.g. the nodes JAK_f, STAT2, up to STAT1 are mostly inactivated in the higher performance models and activated in low performance models while for the two node-cluster {SRC, PTPG6} the reverse happens).

Heatmaps of equations (cluster-based and synergy based):

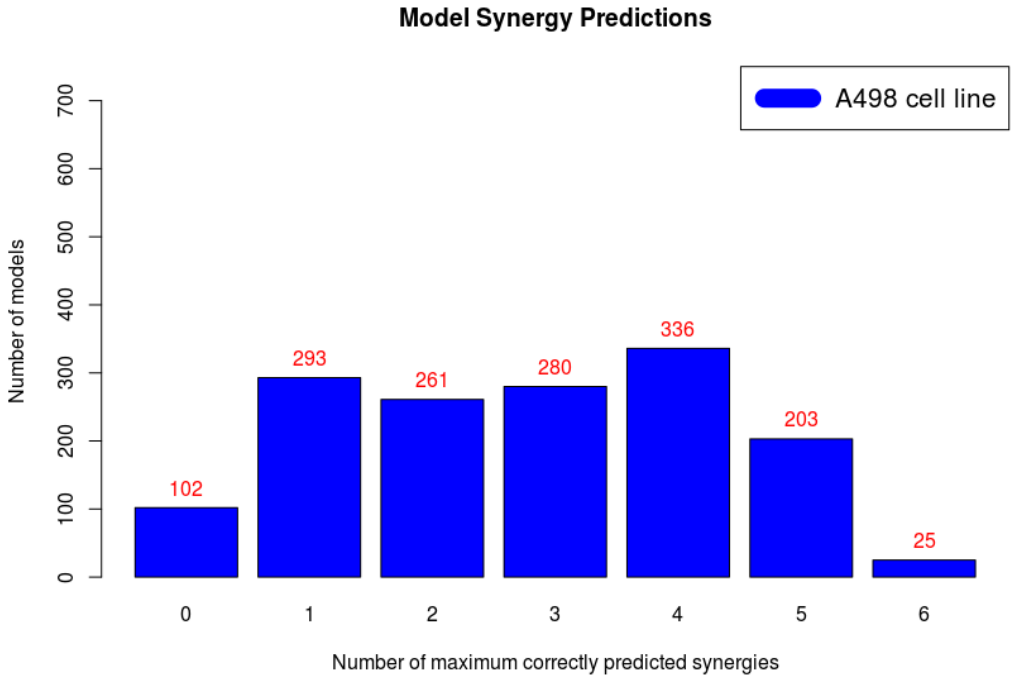


Nothing worth mentioning in these results I believe.

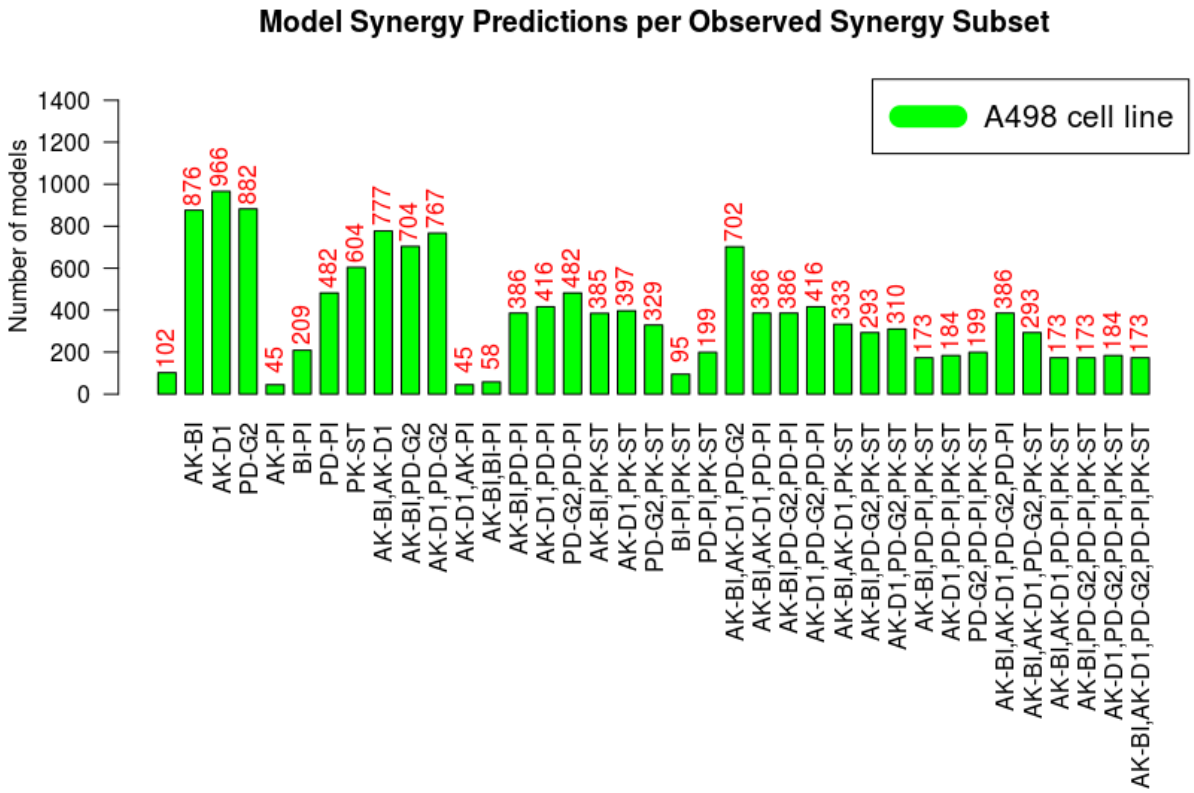
A498 (normal run: fitting to steady state)

The input files were taken from the web server address: http://localhost:8080/SINTEF/ROC/A498_c1d28e71a1d74d99539849518fbba523/info. For this cell line we have 17 experimentally predicted synergies, so we will redo part of the above analysis to see if anything interesting comes up.

Counting number of models vs max corrected predicted synergies and synergy subsets



In the next figure we excluded subsets that were predicted by less than 40 models (for presentation purposes):



Notes:

- From the 17 experimentally predicted synergies, the maximum synergy subset was 6. The 25 models as seen in the last column of the first graph, predicted these 6-set of synergies:

Synergy Set	Number of Models
AK-BI,AK-D1,PD-G2,AK-PD,BI-PD,PD-PI	2
AK-BI,AK-D1,PI-JN,AK-PI,BI-PI,PK-ST	3
AK-BI,AK-D1,PD-G2,AK-PD,PD-PI,PK-ST	16
AK-BI,AK-D1,PD-G2,BI-PD,PD-PI,PK-ST	4
	Total: 25

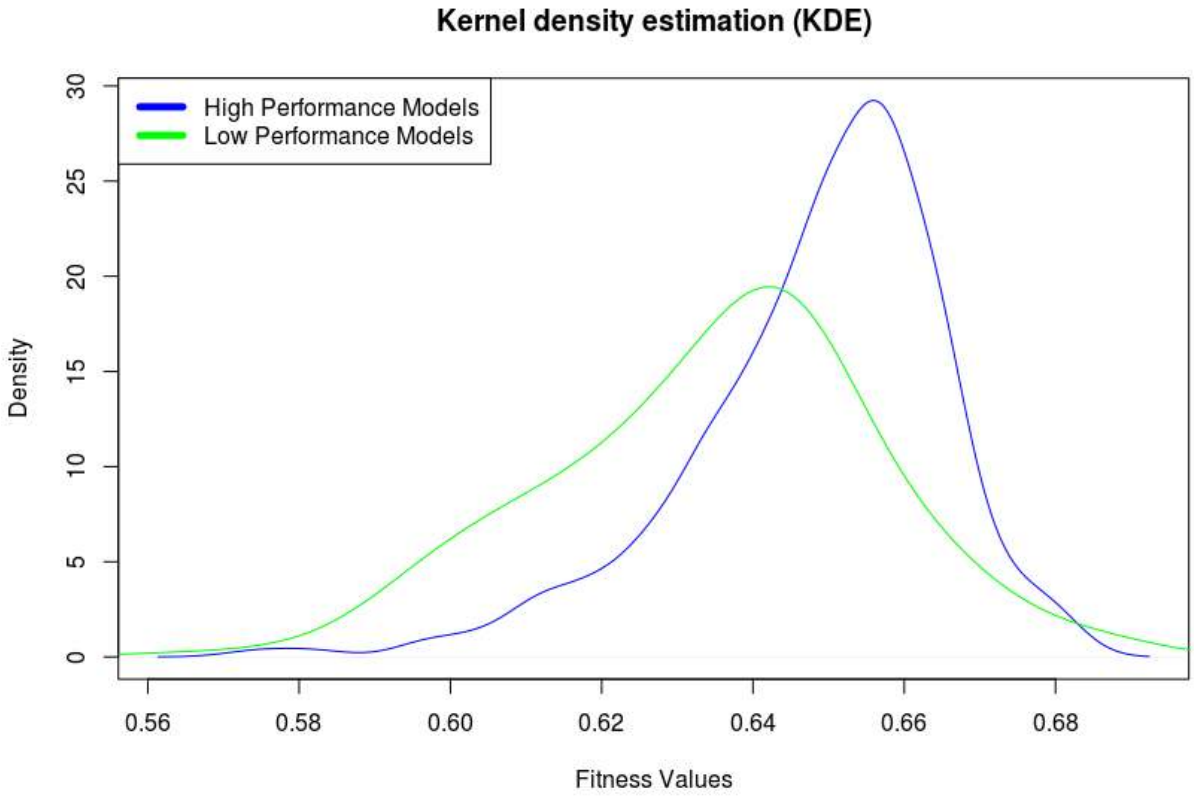
The last two differ only in one synergy, while every other pair differs in more than 3 of the subset synergies.

- No models could predict these synergies: "AK-60" "PI-D1" "AK-G4" "D1-G4" "PD-P5" "PI-P5"

- Again here we note that the models cannot predict more than 6/17 = 35% of the actual synergies. For the normal AGS run this percentage was around 66% which means that more experimentally predicted synergies will come in hand with less overall-predictive models (that predict a large number of synergies). This also correlates to the inability of the models to predict more than a handful of synergies (as seen in this section, PI-D1 synergy notes) - probably a result of the limitations to the form of their boolean equations or the low model number (or both?).

Fitness values of high vs low performance models: statistical correlation

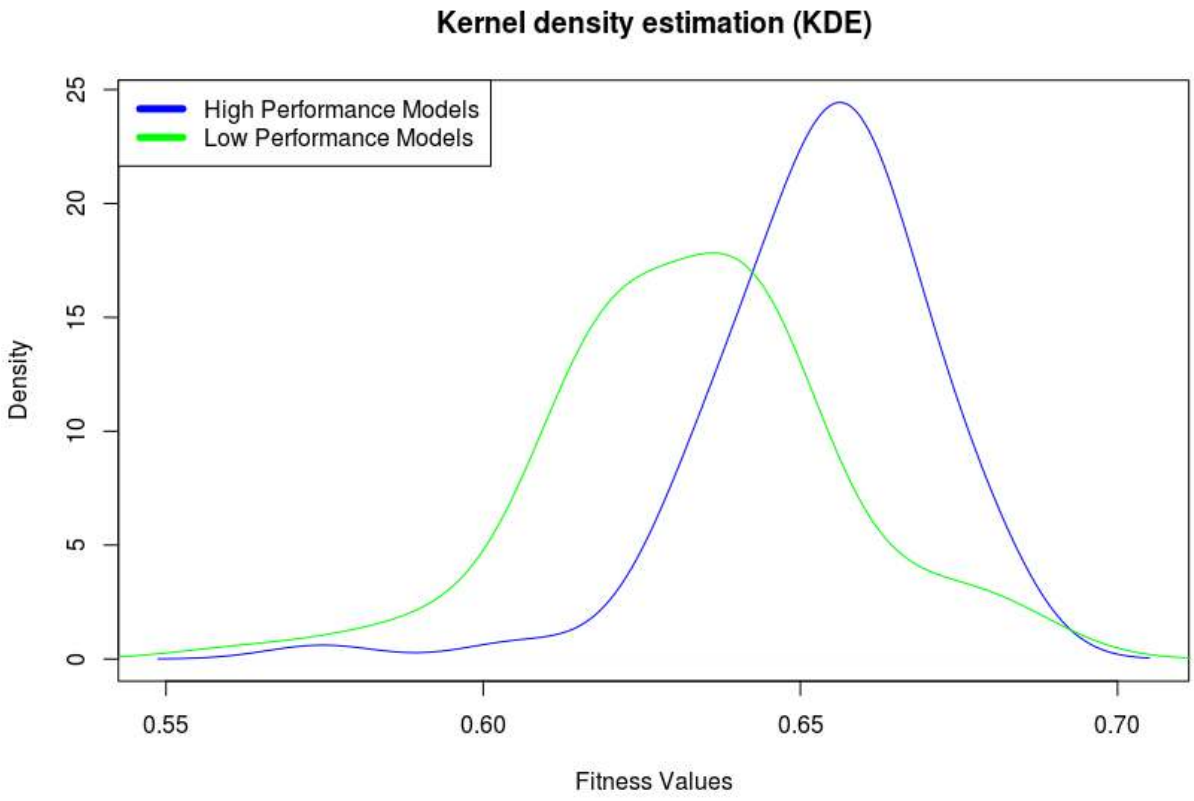
Out of 1500 models, the fitness value range was [0.55, 0.68] (so we have a larger sample of fitness values compared to the AGS cell line) and the distinct fitness values were 18. If we partition the models' fitnesses to two vectors, one with the fitness values of the models that predicted more than 3 synergies correctly (844 models) and the other with the rest (those that predicted less than 3 - 656 models), we get this figure when trying to represent the two vectors' estimates on their corresponding probability density functions:



The two-sample Kolmogorov-Smirnov test has a very low p-value, so the two samples belong to different distributions (the lines in the above graph are smoothed) and we can safely say - as in

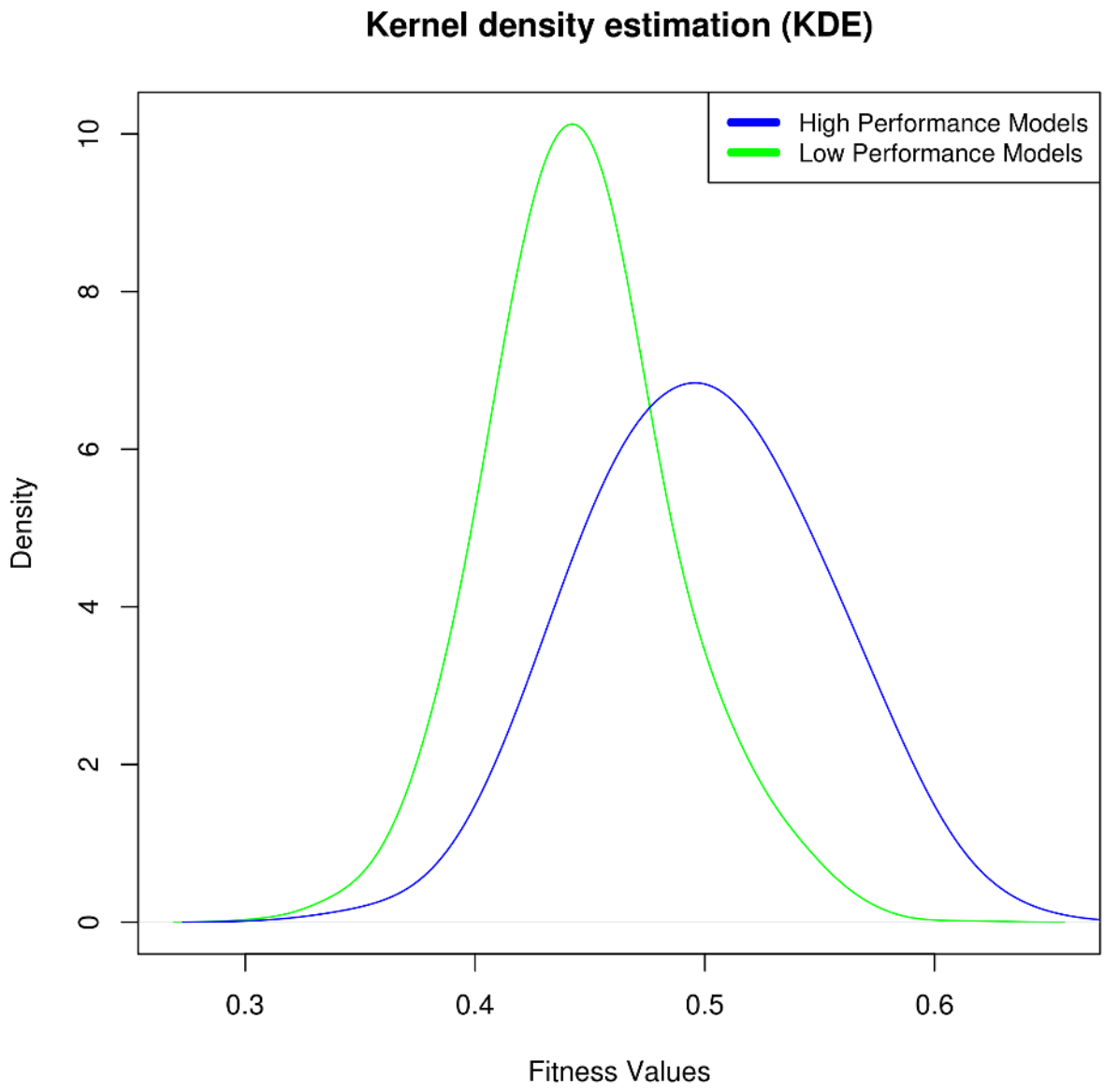
the AGS case - that models having larger fitness on average corresponds to higher predictive performance.

If we partition the models strictier regarding their performance characteristics, i.e. high-performance those that predict more than 4 synergies (228 models) and low-performance those that predict no synergies (102 models), the two fitness vectors give the follow estimate curves (which as you may notice they are more apart from each other than before):



Fitness range investigation

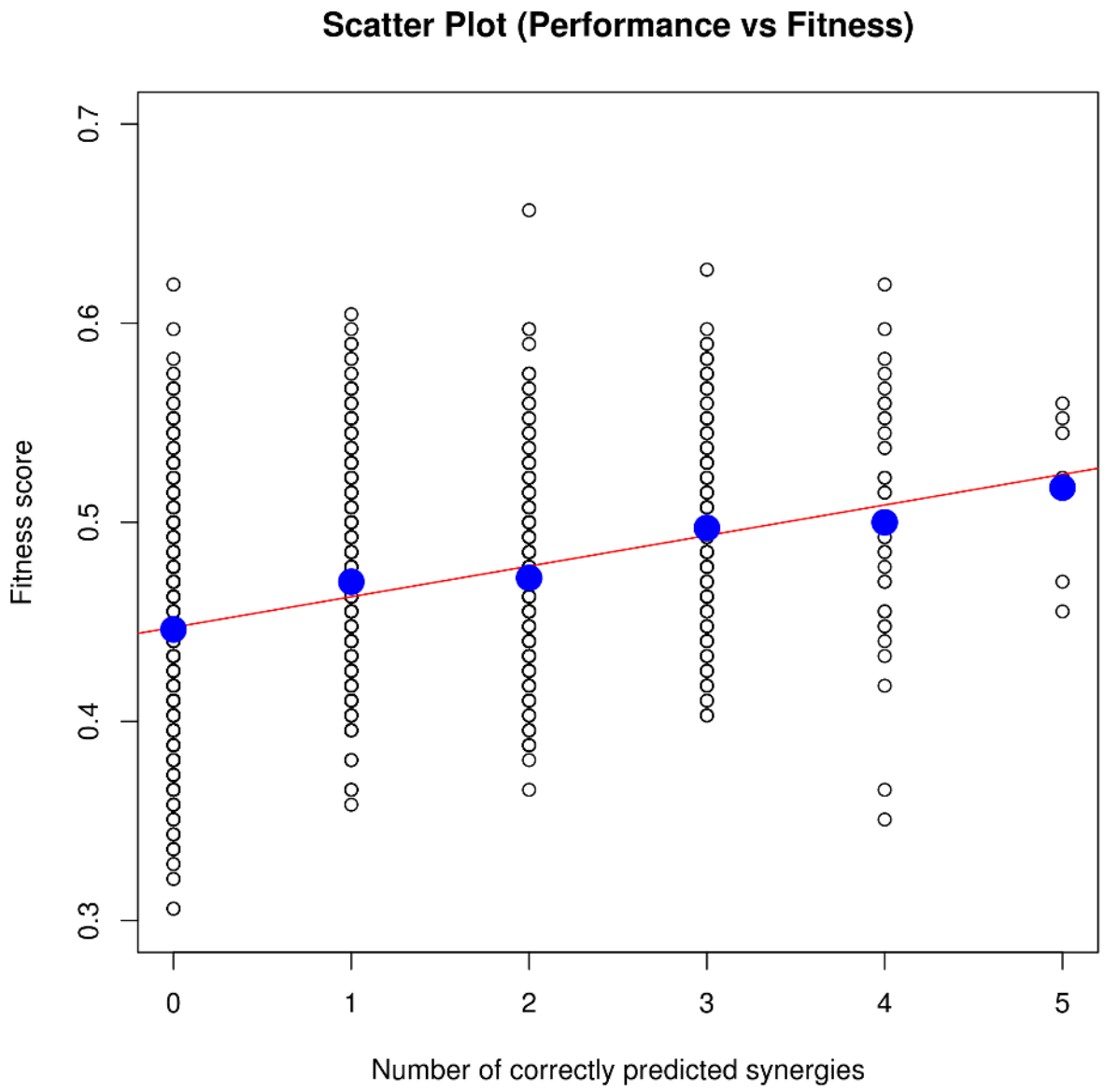
In this analysis, we want to produce the same results as the previous section but for a larger range of fitness values. A new A498 run was set with specific configuration to capture this. I got ~4000 models with fitness scores ranging from 0.30 to 0.65 (so a little larger than the AGS corresponding one). The results are:



Notes:

- High performance: models that predicted 3, 4 or 5 synergies.
- Low performance: models that predicted 0 synergies.

Also, we did a scatter plot where the average values are also shown in blue filled circles:



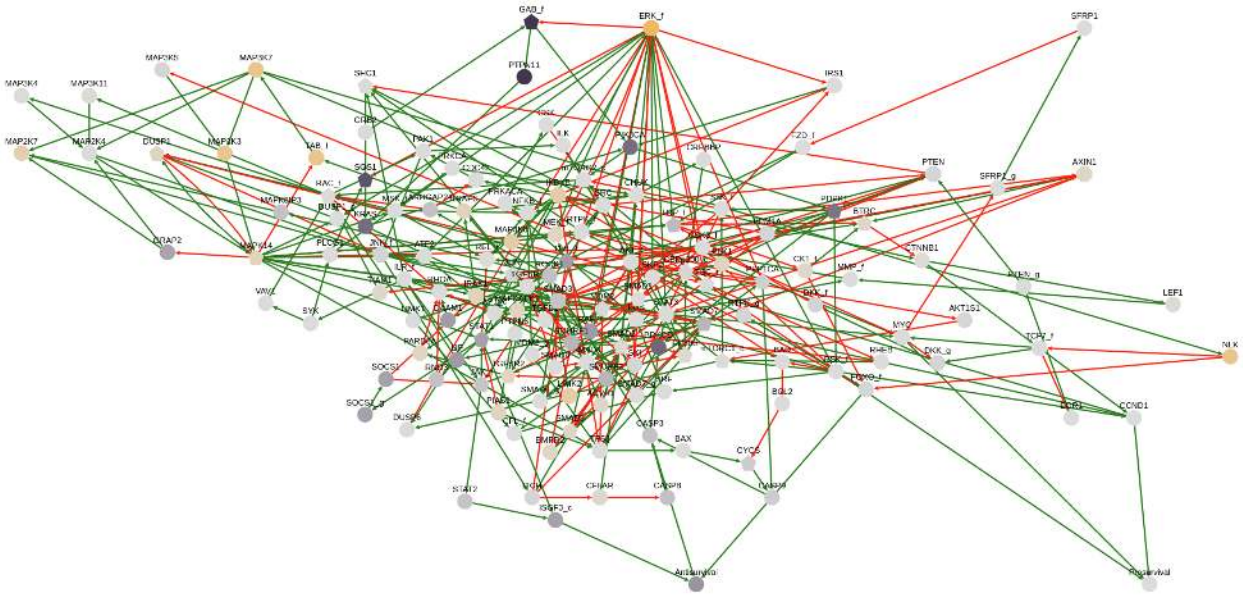
Good vs Bad models (average stable state differences)

Here we will only consider the case were we compare the low vs high performance models. The classification is as follows:

Good models: predicted more than 5 synergies (25 models)
Bad models: predicted no synergies (102 models)

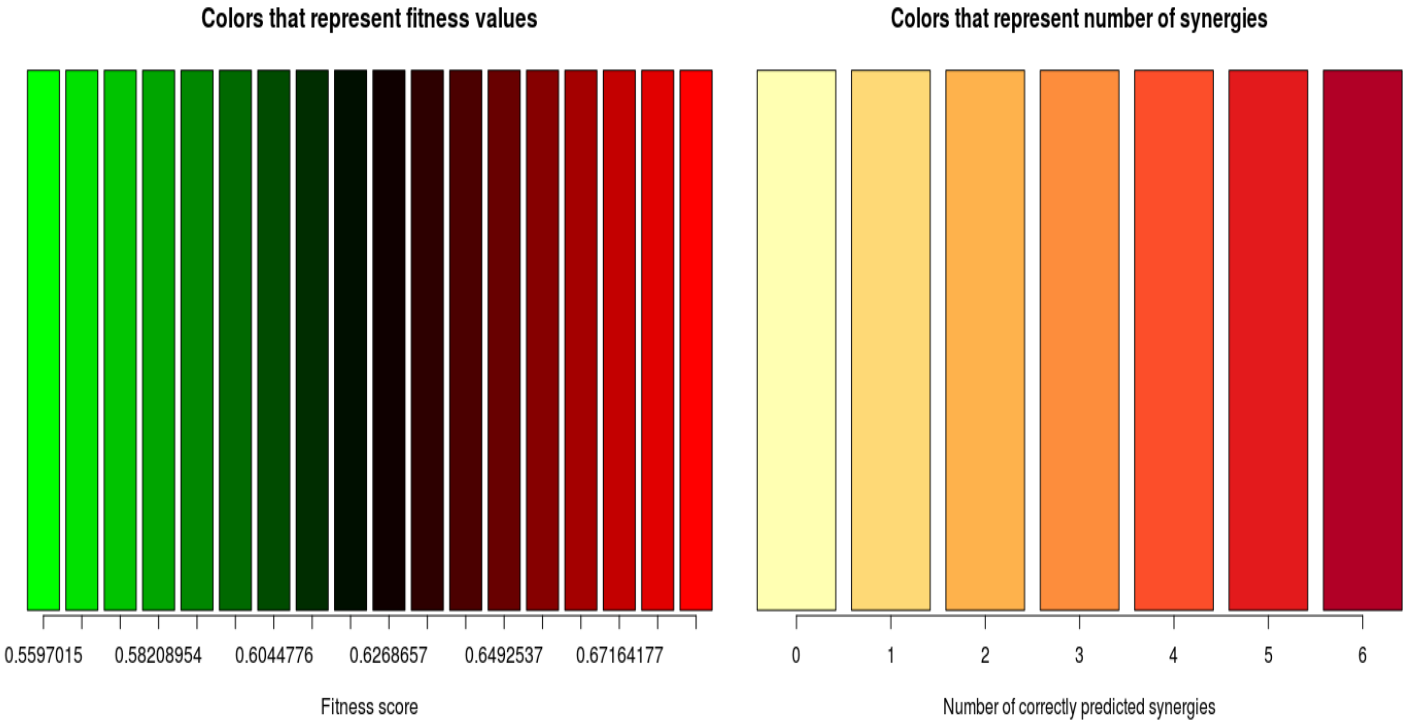
We see in the next image that the main difference of the good performing models vs the bad ones is the activation of the ERK_f node (0.71 difference), while the other nodes don't show too

many differences (the AGS normal case showed many differences). Note that the activation of the ERK_f node was also needed to get to even higher performance models (that also predict the AK-D1 synergy) in the AGS normal run. So, the activation of this particular node seems to be important for high-performance models across two cell lines.

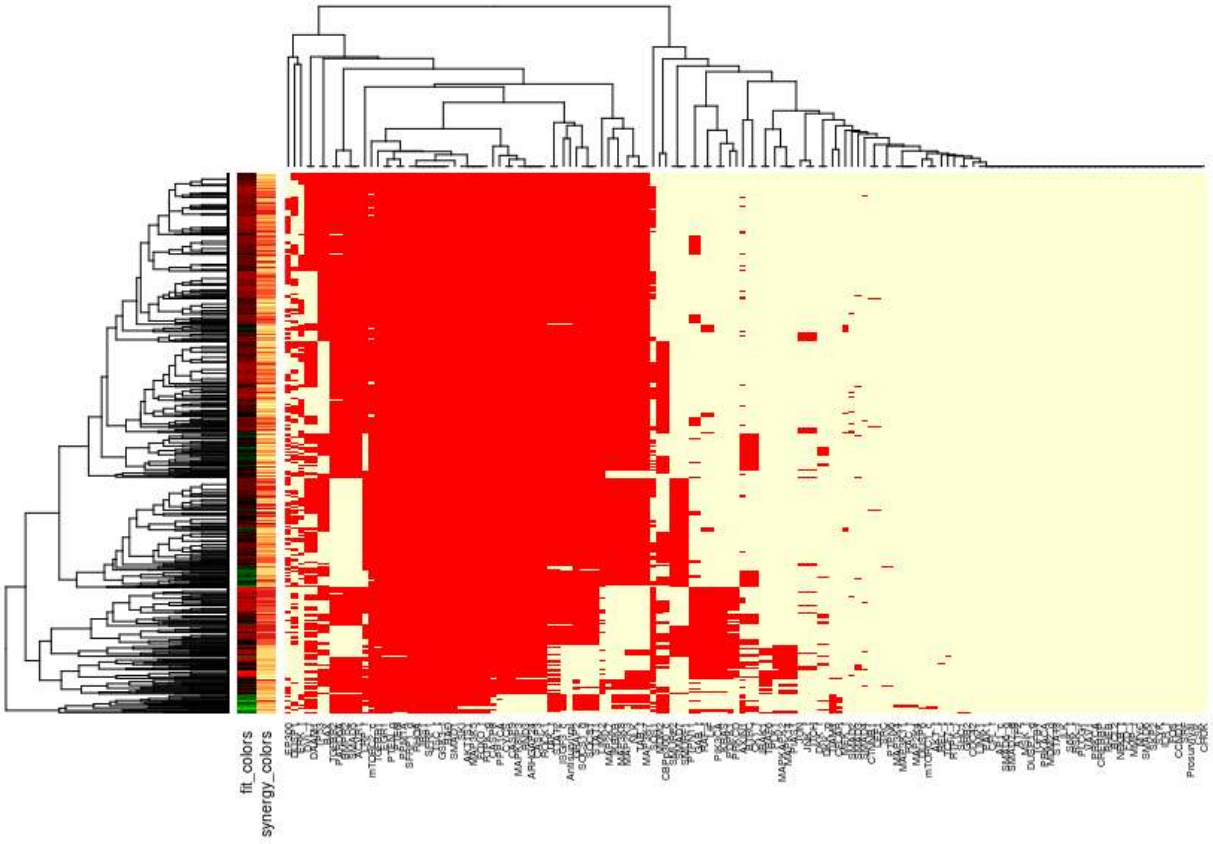


Heatmap of models stable states

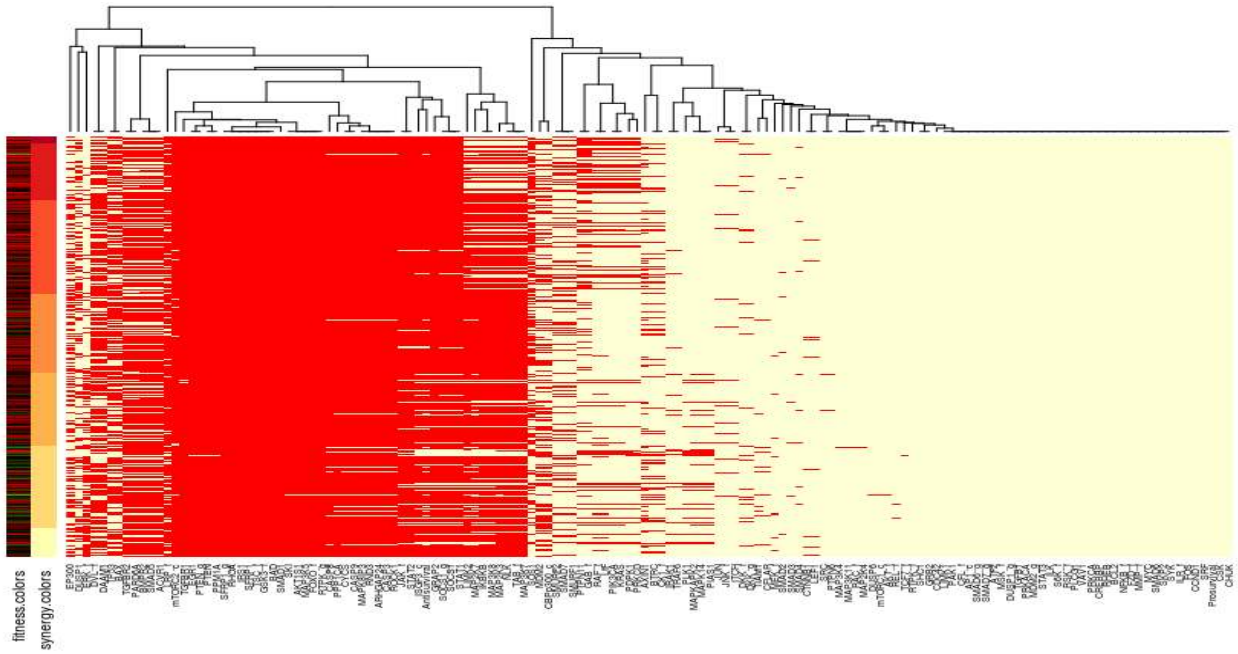
The colors used for the number of synergies predicted by the models and the fitness scores are:



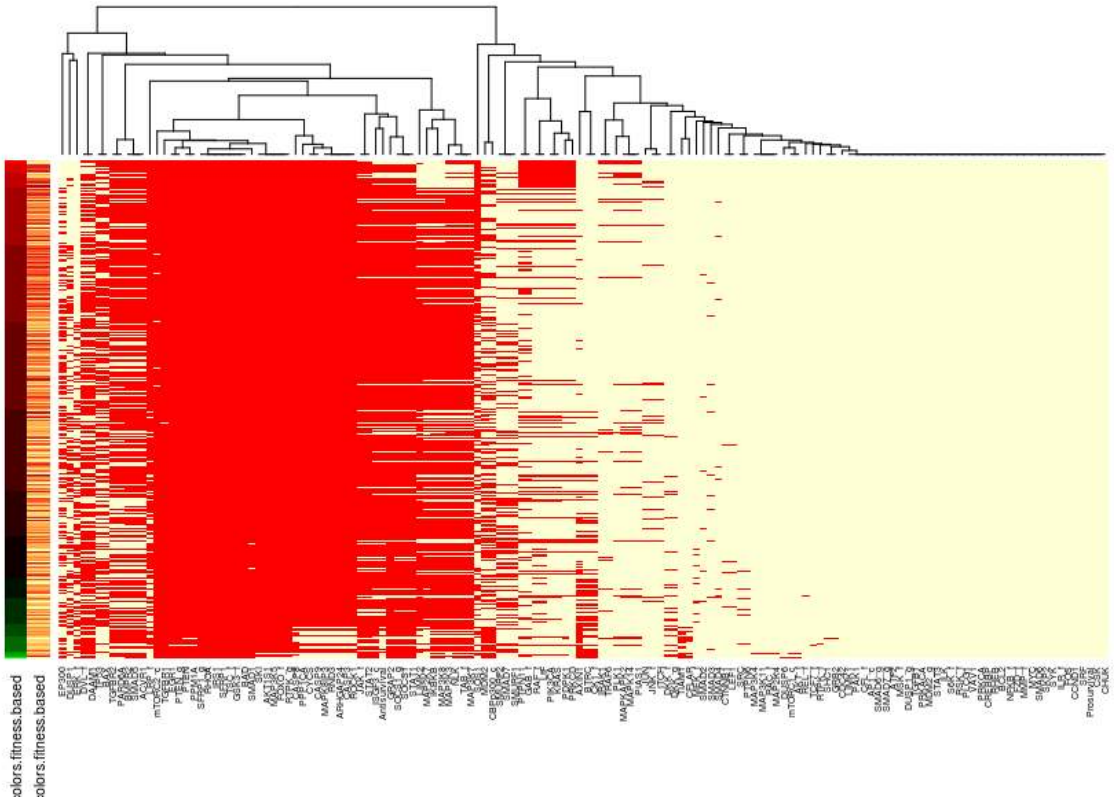
In the next analysis, we will put extra focus on the ERK_f node (and compare with the result from the AGS cell line). So, the heatmap of the stable states (with the default clustering of the fitness and synergy colors) is:



It seems that mostly higher performance models indeed show higher fitness values on average.
If we sort by number of synergies predicted (row-ordering) we get this heatmap:

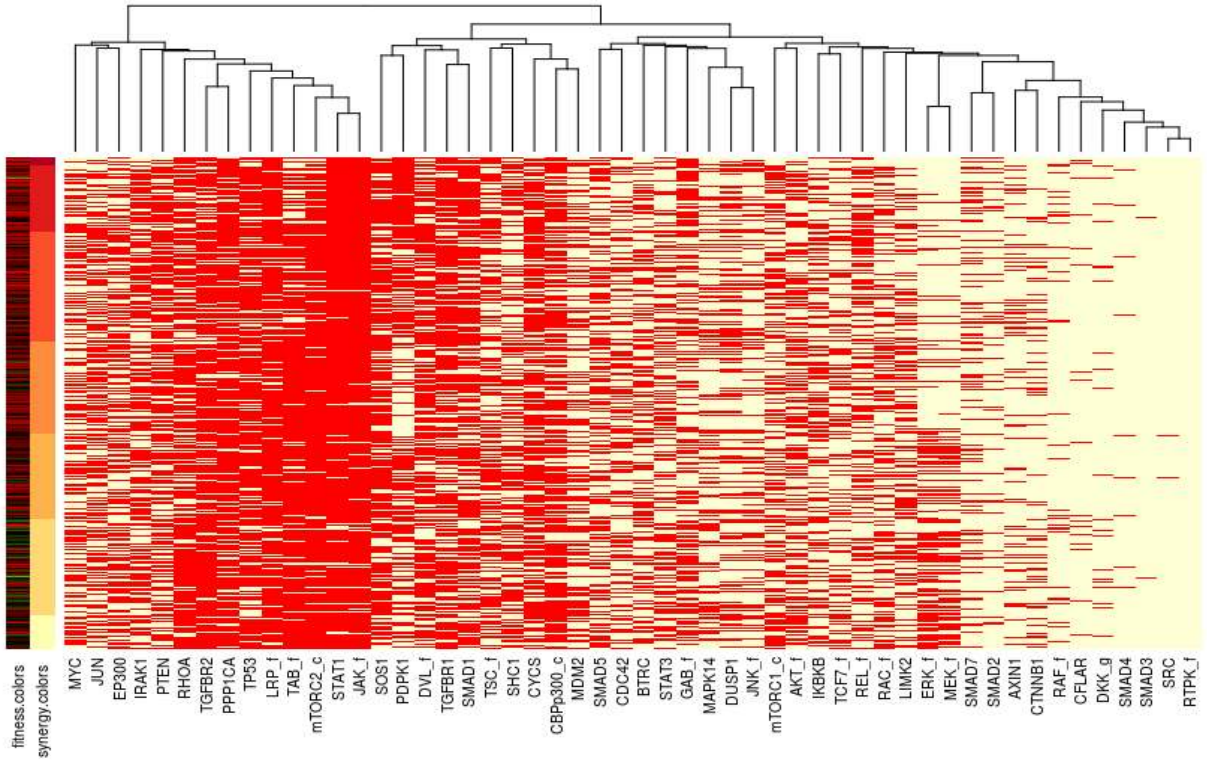


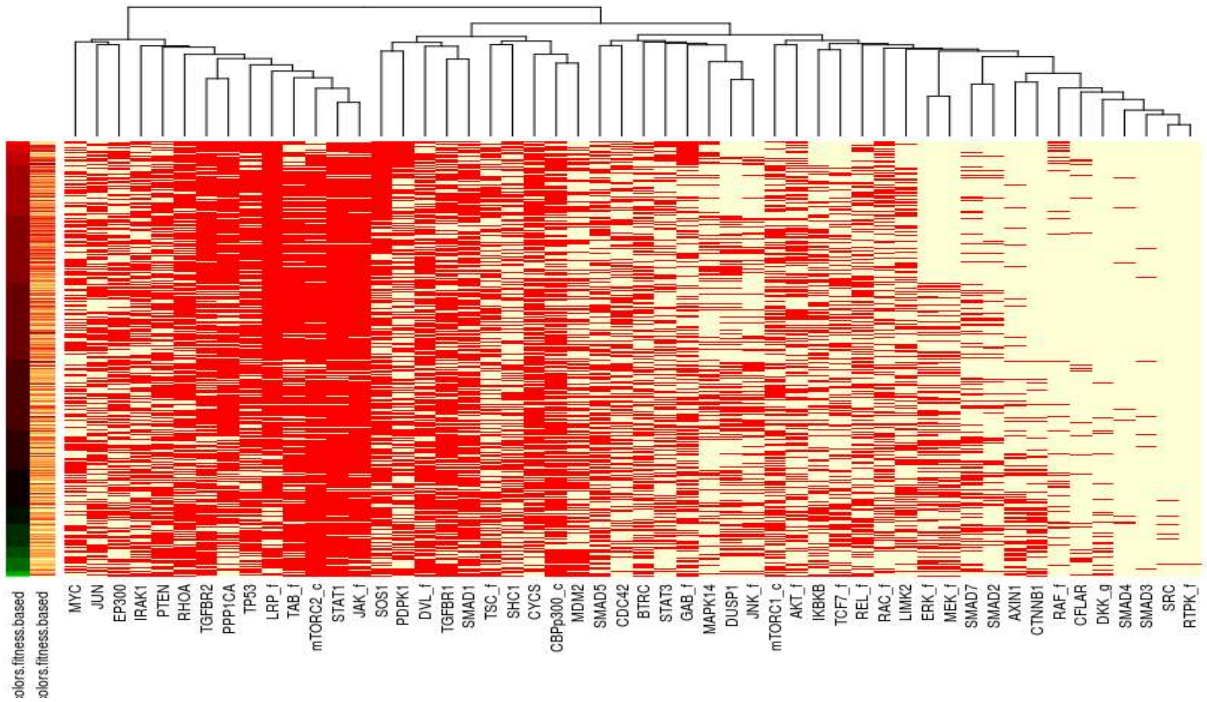
Note that the ERK_f node (3rd from left) is mostly activated in the high performance models. Ordering by the fitness score we get this heatmap of stable states:



Again, ERK_f (3rd node from left) is activated in the higher fitness models. The second heatmap also shows more clearly I believe the relationship between high predictive performance and fitness to steady state we are trying to establish.

Looking at the heatmaps of the boolean equations, we have:



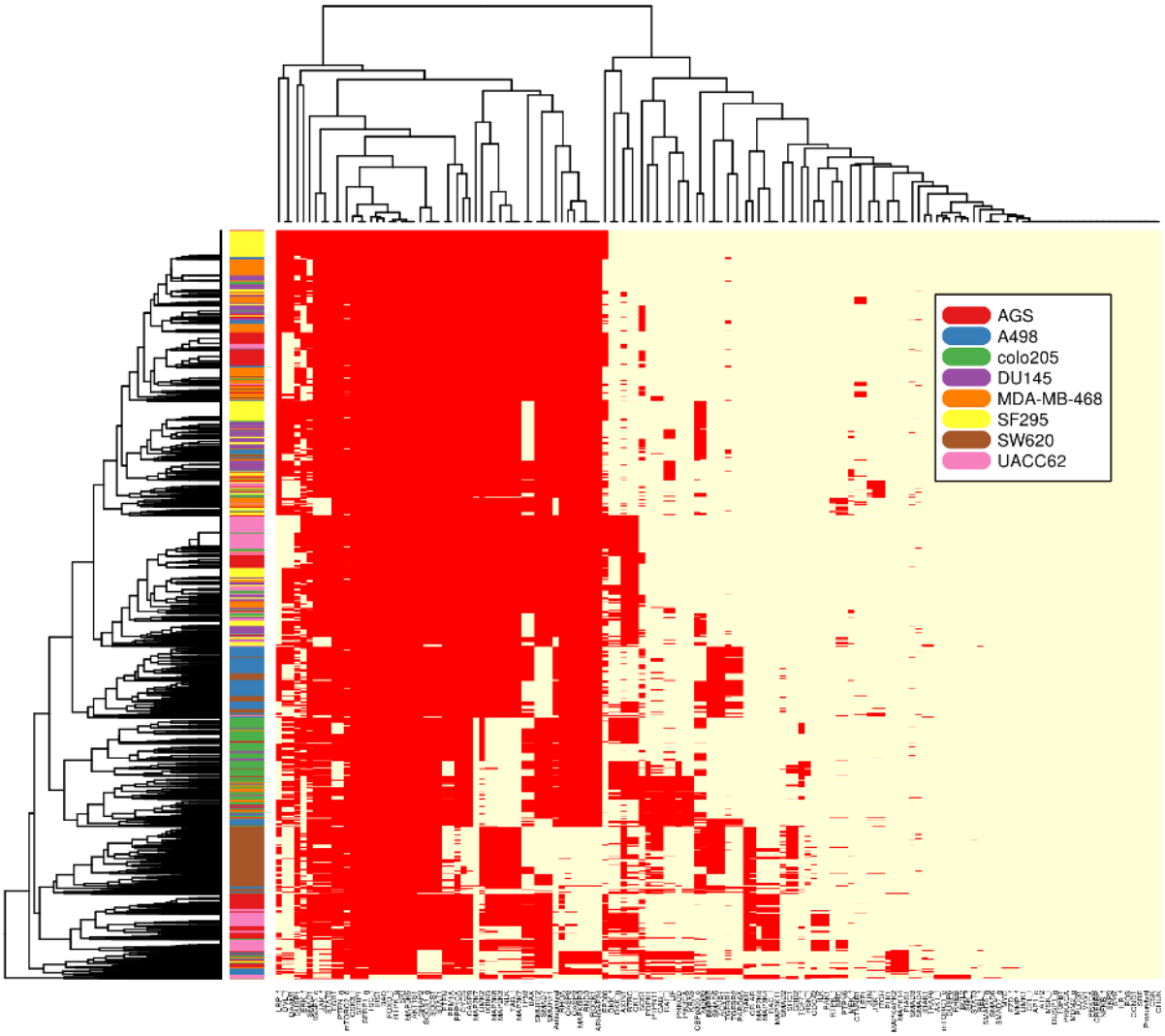


As in the AGS cell line, we still see that higher performance models and also those with higher fitness values correspond to ERK_f (and MEK_f) node having the OR link in the boolean equations.

Cell lines results

In this section, I output two heatmaps: stable states and equations of all tested cell lines: "AGS", "A498", "colo205", "DU145", "MDA-MB-468", "SF295", "SW620", "UACC62" (8 in total).

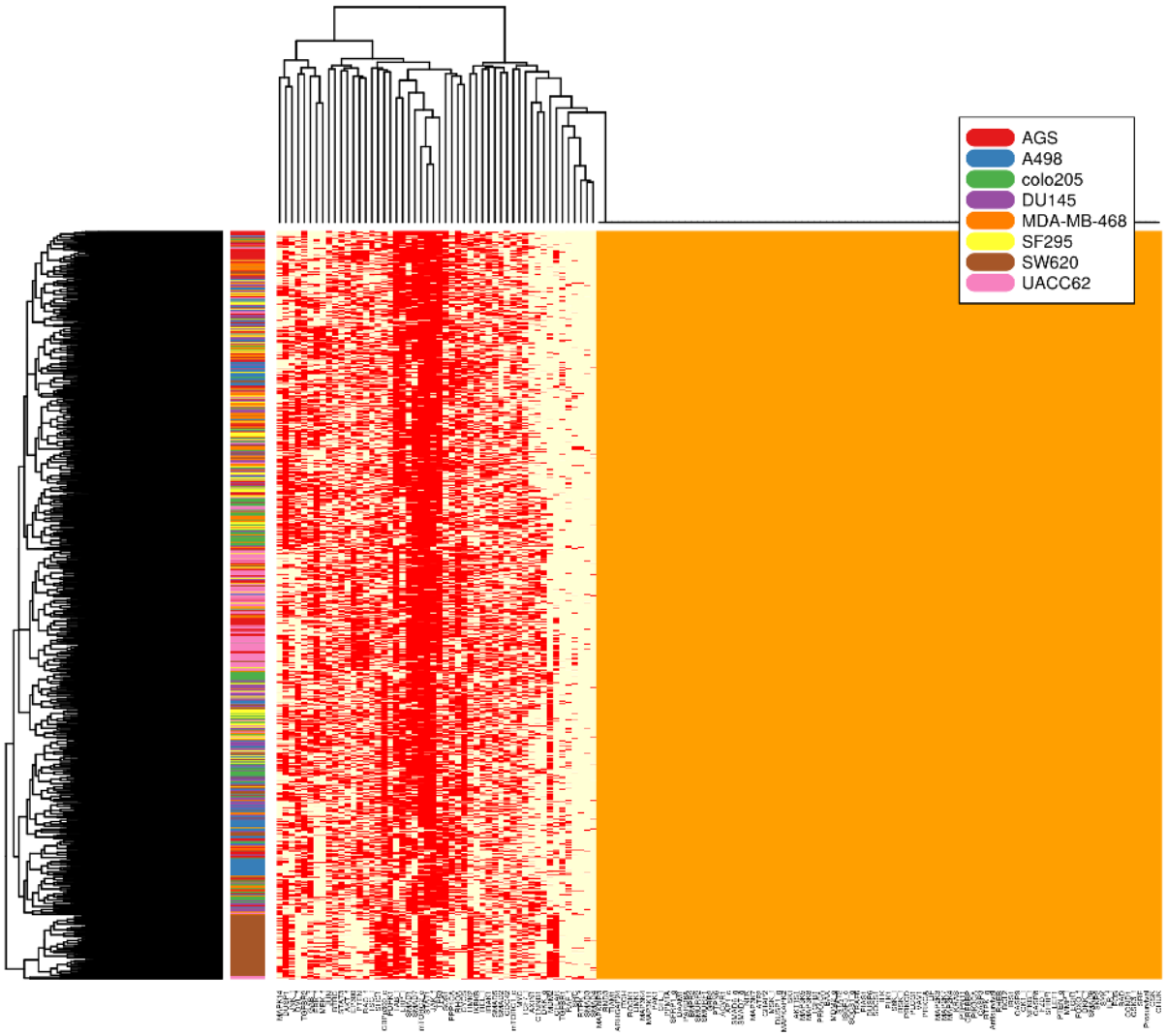
Stable states



Notes:

- SW620 stable states seem completely different from any other cell line
- This unsupervised clustering tell us that the cell lines have an average stable state pattern which overall is not much different and so models between them are interchangeable.

Equations



The yellow right part of the graph represent equations that do not have a link operator. The equation heatmap seems to be more heterogeneous (SW620 is a class of its own!).