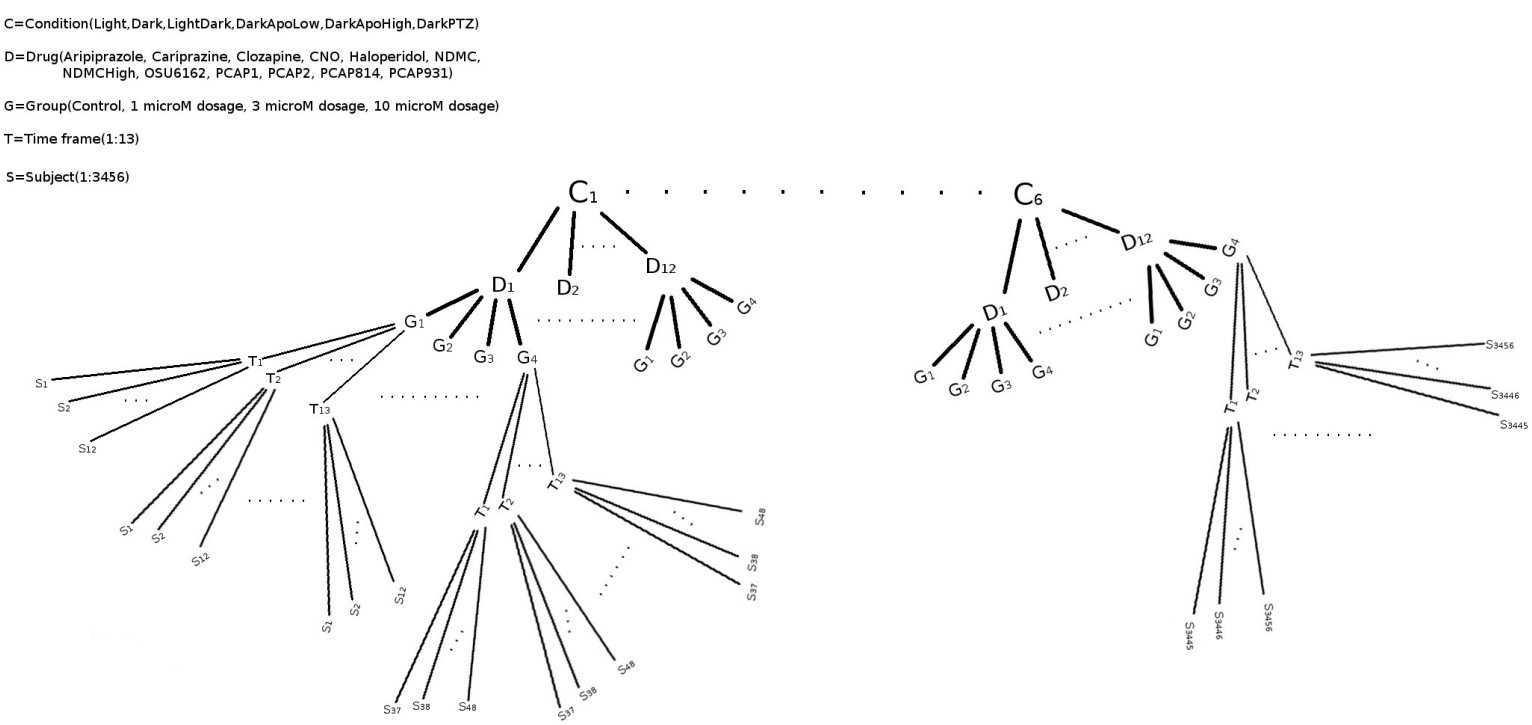
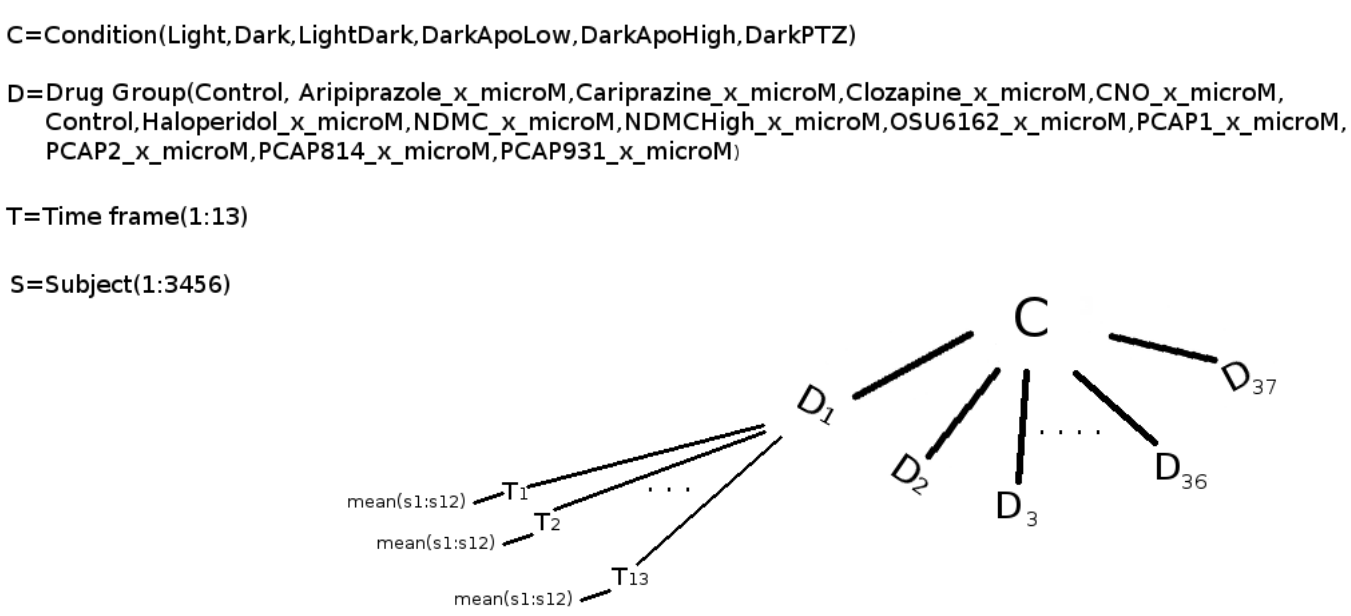
UPDATE 11.5.2017

Generating final combined and simplified dataset and selecting influential variables.
I had generated the final datasets for each condition, where the data has been extracted from a complex dataset where bout was an operational unit. Summarizing that data, by stratifying per bout length, taking the mean proportion per action sequence, per subject, per time frame, I then took mean of all subjects for time frame. As I did before, I got rid of the extra level of dosage and have 37 drug groups, the control + 36 (12 drug * 3dosage). This makes the analysis simpler and the results are easier to interpret. Best to start with this simple and then in the future in it can be extended to a model per subject and even further.

So going from this:



I came to this :



Like this I will be able to compare the controls of each condition and the drugs with reference to control within each condition.

In the dataset, I had combined all the variables I had extracted so far, but stratified per bout length.

These are bout count, turn proportions and turn transitions of length 2 and 3.

In each condition I checked the total count per every length, to approximate the extreme length bouts, so the variables were stratified per bout length 1,2,...extreme length, were all bouts with length equal to or more than the extreme length, are a stratum together.

These are the extreme lengths per condition: Light,Dark...6, DarkApoHigh, DarkPTZ...7, DarkApoLow...3, LightDark...9 For example Light will have 6 strata for all variables, except total bout count and time frame,

So final variables are time frame, total bout count per time frame, bout count per length stratum per time frame, turn proportion per length stratum for each turn type(8) per time frame, transition frequency proportion per length stratum for each possible transition of length 2(64) and 3(512) per time frame, where a lot of possible transitions do not occur so the total number of variables is not full length of combinations and is different per condition.

I am attaching a table with variable descriptions an amount of variables, before and after variable selection.
I am also making a dataset where the variables have not been stratified per bout length, just so I can check and compare the final results.

Variables were selected within each condition, based on the mean decrease impurity and overall classification scores, when modeling the 37 drug groups with random forest, so multiclassification into 37 classes.

Since it is difficult to set the best cut off for the mean decrease impurity (at least for me, maybe someone else has a better way of doing this), I had tested several cut offs, each resulting in a set of selected variables, with which I then obtained classifications scores with 10-fold cross validation and selected the best set of values for classification accuracy, precision, recall, f1 score and AUC(multi class), where all the best sets were better then the scores of a model where all variables are selected.

In the attached table are the resulting number of the variables selected in each condition, I am also attaching lists of variables per each condition.
Transitions are included in the influential variables, so that is good and rf with mean decrease impurity is supposed to be somewhat immune to multicollinearity, although where one highly correlated variable is already in the model, the next will result in low mean decrease impurity, where both could actually be important in a biological sense.
And the multicollinearity in the data is complex, so the final check is to get the effect sizes with regression while adjusting for covariates and check the significance and fit of the model, if the effect is non significant it has to be set to 0 even if the variable was selected as influential.

With these variables I could calculate cluster distances from the control over all time frames(as before) and present it like a dendogram (this can still be liable to confounding),
or/and
avoid confound as best as possible by obtaining effect sizes with negative binomial or quasipossion per overall change and change in time with control as reference or when comparing conditions, light as reference and adjusting for covariates.

With the effect sizes per variable, I could again check the overall cluster distances with all variables and show dendogram, or/and
for each influential variable show the effect sizes in a 2d plot with control as 0 and overall difference as x and time progress difference as y. There are a lot of variables, maybe I can try plotting them together per variable over all strata, I ll have to see if they group together nicely otherwise it will look like a mess.

Right now, all the variables are treated the same, so a decrease in length 3 bouts, an increase in scoots proportion, an increase in gj transitions etc... have the same severity and all contribute the same to the final distance of a certain drug from the control.

I will make a table with all the significant effect sizes, but for final results it will be too big and still too much information that needs to be better presented and summarized appropriately so the results are clear and make sense and show the difference between the drugs and difference between the controls in different conditions.

Maybe presenting the final dendogram and distances from the control with all variable effect sizes , try PCA methods and do a 2d plot for all drugs.
and then focusing on just some interesting variables, like proportions and transition proportions including jbends, cbends and obends…and do 2d plots of those and show the actual effect sizes, so they can be interpreted, I think negative binomial and quasipossion have effect sizes in log odds so would have to be careful to transform the effect size first and then use it to say that for example the probability of xx transition will increase/decrease by b in a certain drug group referenced to control.

Things are pilling up on each other so I really hope I am doing all the intermediate steps ok, if anybody has time and energy to check the code and results on git it would be very helpful, otherwise I can write more explanation on the procedure and methods used if you think there is something unclear or not done properly.