Estimation of total body fat using symbolic regression and evolutionary algorithms.

Paper summary

- Public data from NHANES CDC
- BMI →body fat % estimation
- Avoid black-box models
- Baseline
- Model parts

- Schnur, Jennifer J., and Nitesh V. Chawla. "Information fusion via symbolic regression: A tutorial in the context of human health." Information Fusion 92 (2023): 326-335.

Paper summary – baseline models

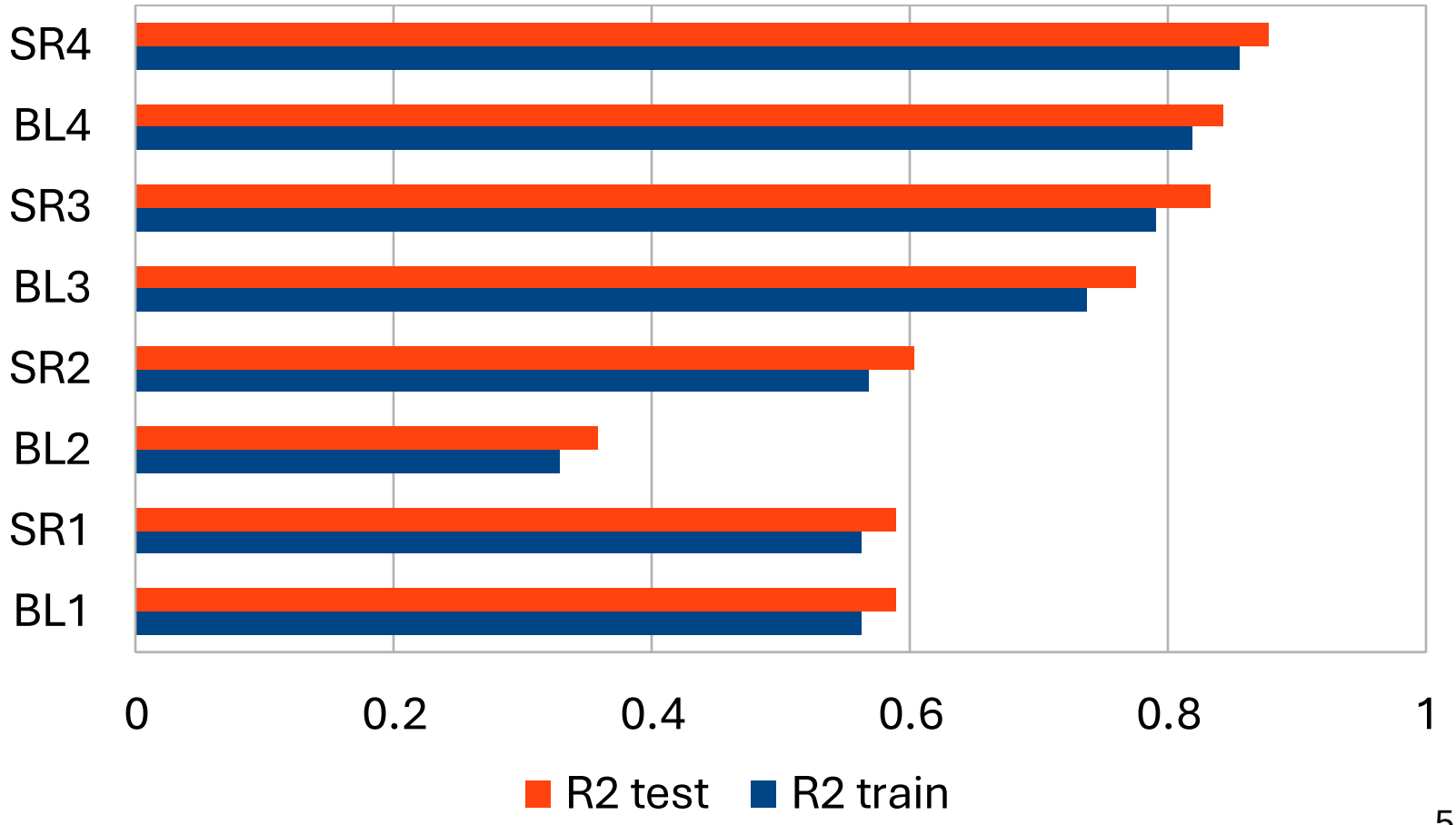
- Linear regression
- Height and weight
- BMI as features
- 7 body measurements
- Gender (0 or 1), age

Paper summary – SR models

- Weight and height – same as LR
- BMI – using square root
- 7 mesurements – composition and exp
- gender&age – split case

$$RIAGENDR_{cat} = \begin{cases} -0.2514210227924248, & Gender = Male \\ 0.24145479395502106, & Gender = Female \end{cases}$$

Paper summary – comparison



Challenge

- Improve the results of the paper:

- Schnur, Jennifer J., and Nitesh V. Chawla. "Information fusion via symbolic regression: A tutorial in the context of human health." Information Fusion 92 (2023): 326-335.

Our proposal

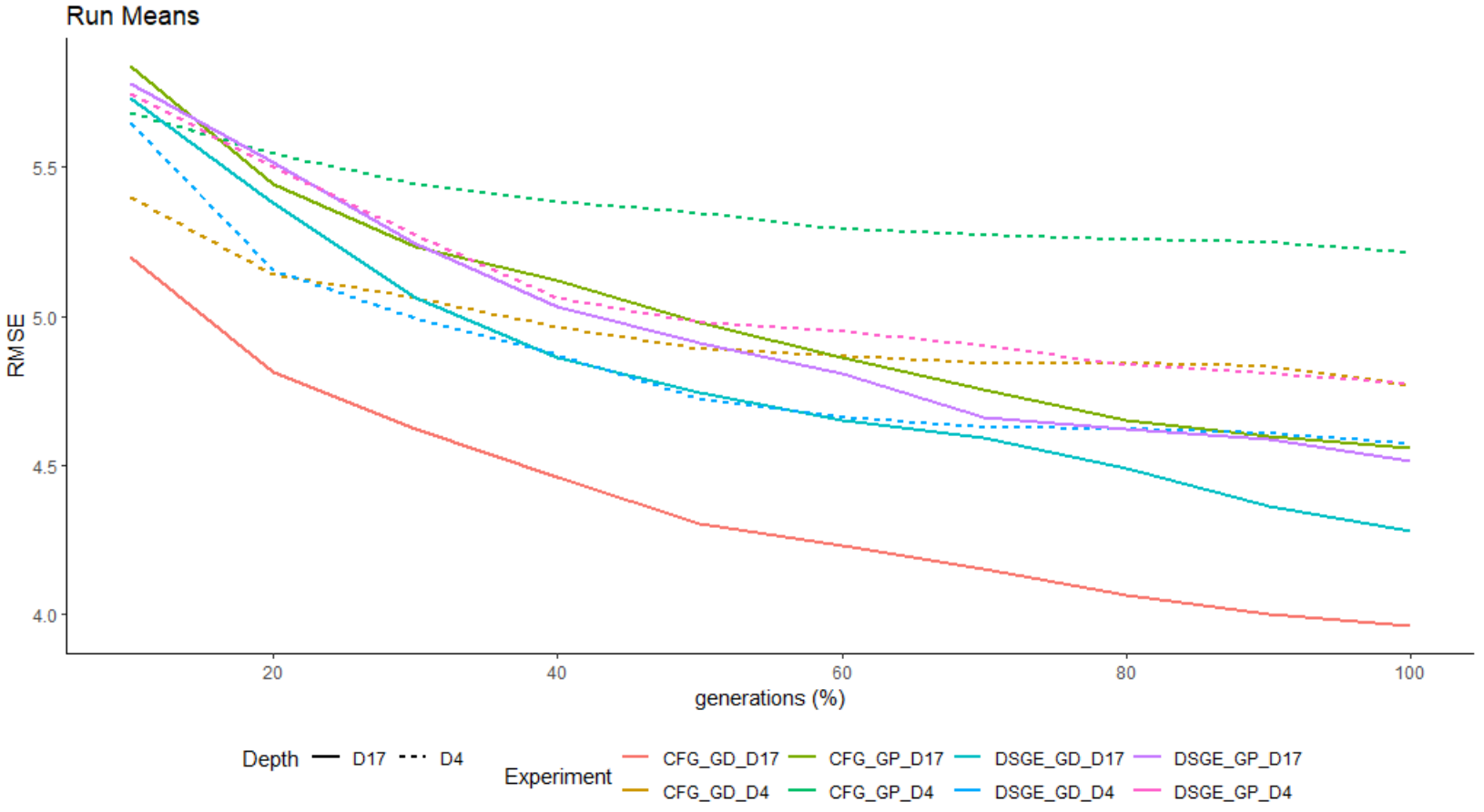
- Grammar Based GP
  - DSGE (Dynamic Structured GE)
  - CFG-GP (Context Free Grammar GP)
  - GE

Grammars

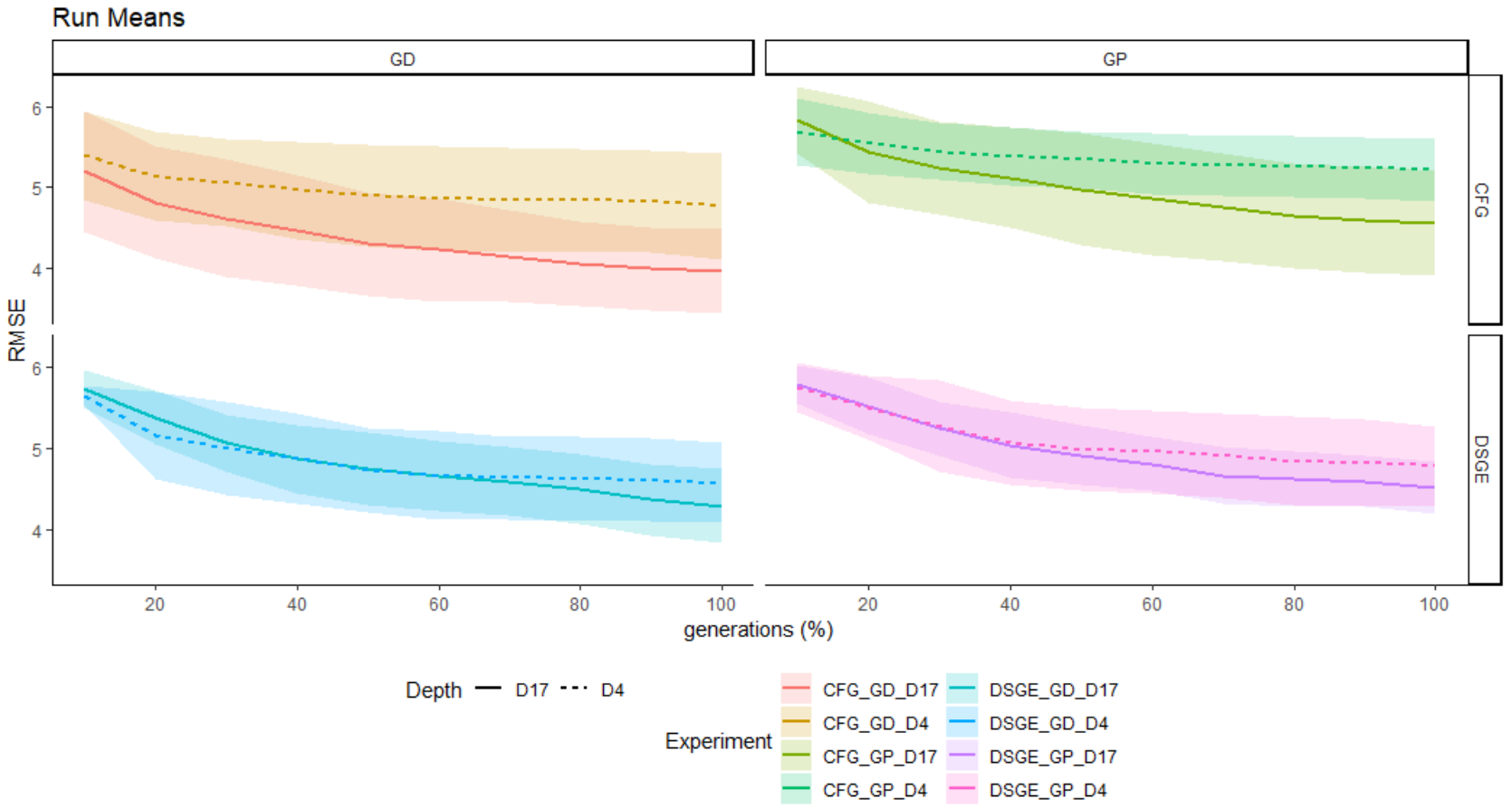
```
<func> ::= <expr>

<expr> ::= <expr> <op> <expr> | <expr> <op> <expr> | <expr> <op> <var> | <var> <op> <var>|
<cte> <op> <expr> |<cte> <op> <var>

<op> ::= +|-|*

<var> ::= getVariable(1,k)|getVariable(2,k)|getVariable(3,k)|getVariable(4,k)|getVariable(5,k)|
getVariable(6,k)|getVariable(7,k)|getVariable(8,k)|getVariable(9,k)


<cte> ::= <base>*Math.pow(10,<sign><exponent>)
<base> ::= 1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31|
32|33|34|35|36|37|38|39|40|41|42|43|44|45|46|47|48|49|50|51|52|53|54|55|56|57|58|59|60|61|62|63|
64|65|66|67|68|69|70|71|72|73|74|75|76|77|78|79|80|81|82|83|84|85|86|87|88|89|90|91|92|93|94|95|
96|97|98|99
<exponent> ::= 1|2|3|4|5|6|8|9
<sign> ::= +|-
```

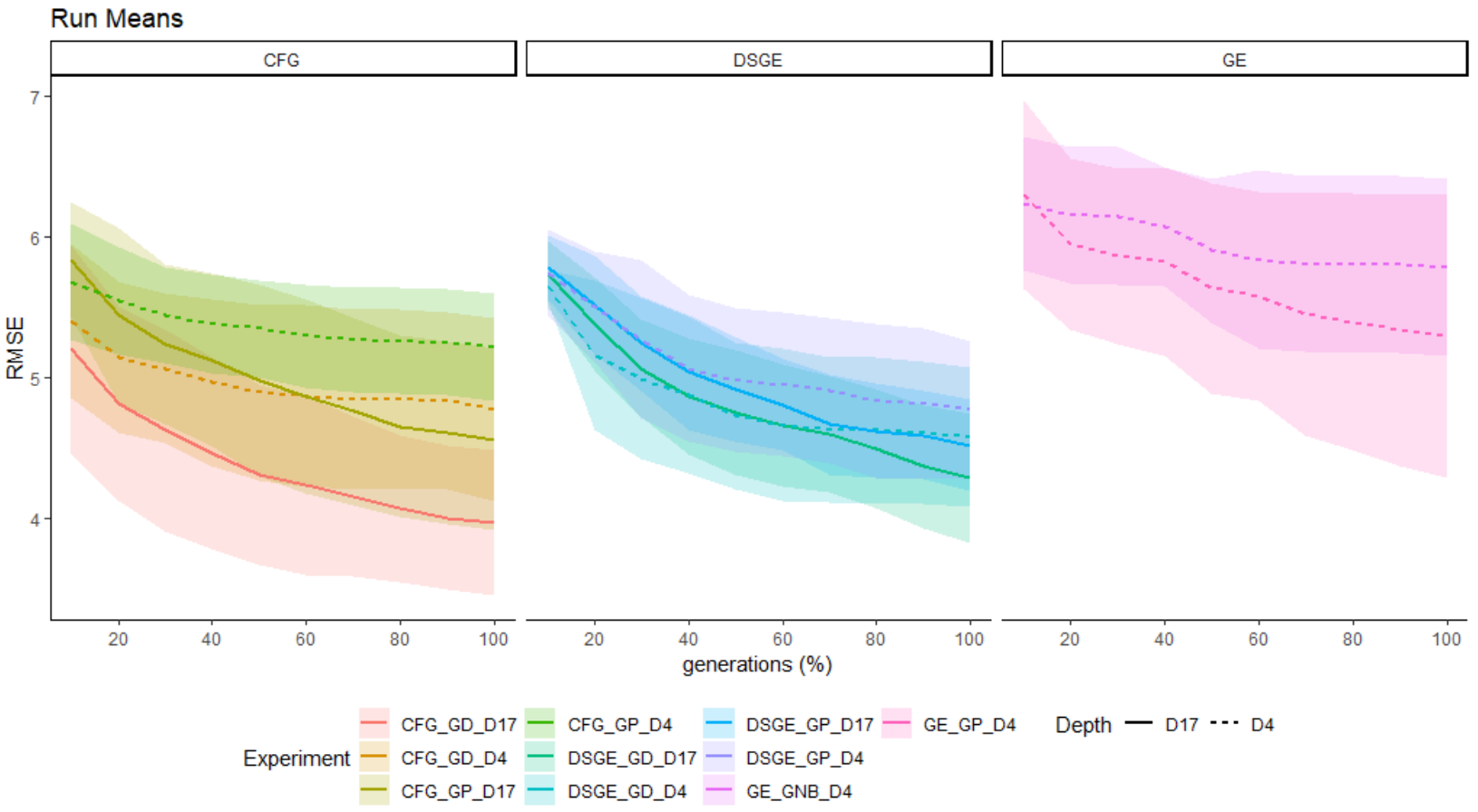
Parameters

- Algorithms
  - GE, CFG-GP, DSGE
- Grammars
  - With basic operators
    - With and without division
  - Other grammars
- 10 Runs
- 1000 Generations
- 0.9 Probability of Crossover
- 0.05 Probability of Mutation
- Max Tree Depth
  - 4, 17


Run Means
RMSE
5.5
5.0
4.5
4.0
20
40
60
80
100
generations (%)
Depth
D17
D4
Experiment
CFG_GD_D17
CFG_GP_D17
DSGE_GD_D17
DSGE_GP_D17
CFG_GD_D4
CFG_GP_D4
DSGE_GD_D4
DSGE_GP_D4

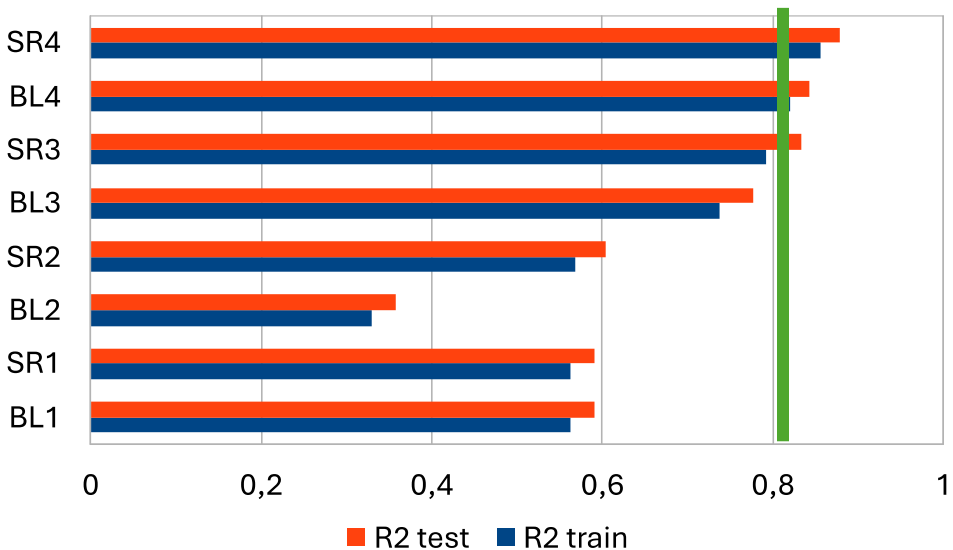
Preliminary Results




Run Means
CFG
DSGE
GE
RMSE
7
6
5
4
20
40
60
80
100
generations (%)
Experiment
CFG_GD_D17
CFG_GD_D4
CFG_GP_D17
CFG_GP_D4
DSGE_GD_D17
DSGE_GD_D4
DSGE_GP_D17
DSGE_GP_D4
GE_GP_D4
GE_GNB_D4
Depth
D17
D4

Results

Where are we?



Configuration	Train	Test
CFG-GD-D4	0.79	0.79
CFG-GD-D17	0.82	0.81
CFG-GP-D4	0.73	0.78
CFG-GP-D17	0.80	0.80

Best Result

- CFG Depth 17, Grammar with división
- 0.98+RIAGENDR+0.98+RIAGENDR+BMXWAIST/BMXARML+BMXWAIST/BMXARMC+0.00098*BMXARML-BMXWT/BMXARML-BMXHIP/BMXARMC+BMXWAIST/BMXARMC-BMXHIP/BMXARMC+0.98+RIAGENDR+0.96+RIAGENDR+BMXWAIST/BMXARML-BMXHIP/BMXARMC+0.98+RIAGENDR+BMXWAIST/BMXARML+BMXWAIST/BMXWT+BMXWAIST/BMXWT+0.98+RIAGENDR+BMXWAIST/BMXARMC-BMXHIP/BMXARMC*BMXHIP/BMXARML+0.98+RIAGENDR+BMXWAIST/BMXARML+BMXWAIST/BMXARMC+0.00098*BMXARML-BMXWT/BMXARML-BMXHIP/BMXARMC+BMXHIP-BMXARMC-BMXARML-BMXWT/BMXARML-BMXWT/BMXLEG-BMXHIP/BMXARMC*BMXHIP/BMXARML-BMXHIP/BMXARMC-0.98+RIAGENDR+BMXWAIST/BMXWT+BMXHT/BMXHIP+0.98+RIAGENDR+BMXWAIST/BMXARMC-BMXHIP/BMXARMC*BMXHIP/BMXARML+BMXWAIST/BMXARML-BMXHIP/BMXARMC

Simplified form

$$\begin{aligned}\text{DXDTOPF} = 9 \cdot \text{RIAGENDR} - \text{BMXARMC} - 0.98 \cdot \text{BMXARML} + \text{BMXHIP} + \\ + 3 \cdot \frac{\text{BMXWAIST}}{\text{BMXWT}} + 6.84 - \frac{\text{BMXWT}}{\text{BMXLEG}} + \frac{\text{BMXHT}}{\text{BMXHIP}} + \\ + 5 \cdot \frac{\text{BMXWAIST}}{\text{BMXARML}} - 3 \cdot \frac{\text{BMXWT}}{\text{BMXARML}} - 6 \cdot \frac{\text{BMXHIP}}{\text{BMXARMC}} + \\ 5 \cdot \frac{\text{BMXWAIST}}{\text{BMXARMC}} - 3 \cdot \frac{\text{BMXHIP}^2}{\text{BMXARMC} \cdot \text{BMXARML}}\end{aligned}$$

Analysis

- Most of the structures appear several times, just adding
  - Modify the grammar
- Constants are the same, which may indicate that the crossover process is somehow degenerated
  - Type of elitism
- Gender appears 9 times!!
- Preeliminary results with other Depth (12) seem to indicate that results can be improved.

Conclusions

- First results are promising
- More experimentation is needed
- Division in the grammar improves the results
- CFG-GP wins
- D17 is better than D4
- Include other variables
  - Only if used in clinical practice