Cluster Analysis of a Symbolic Regression Search Space



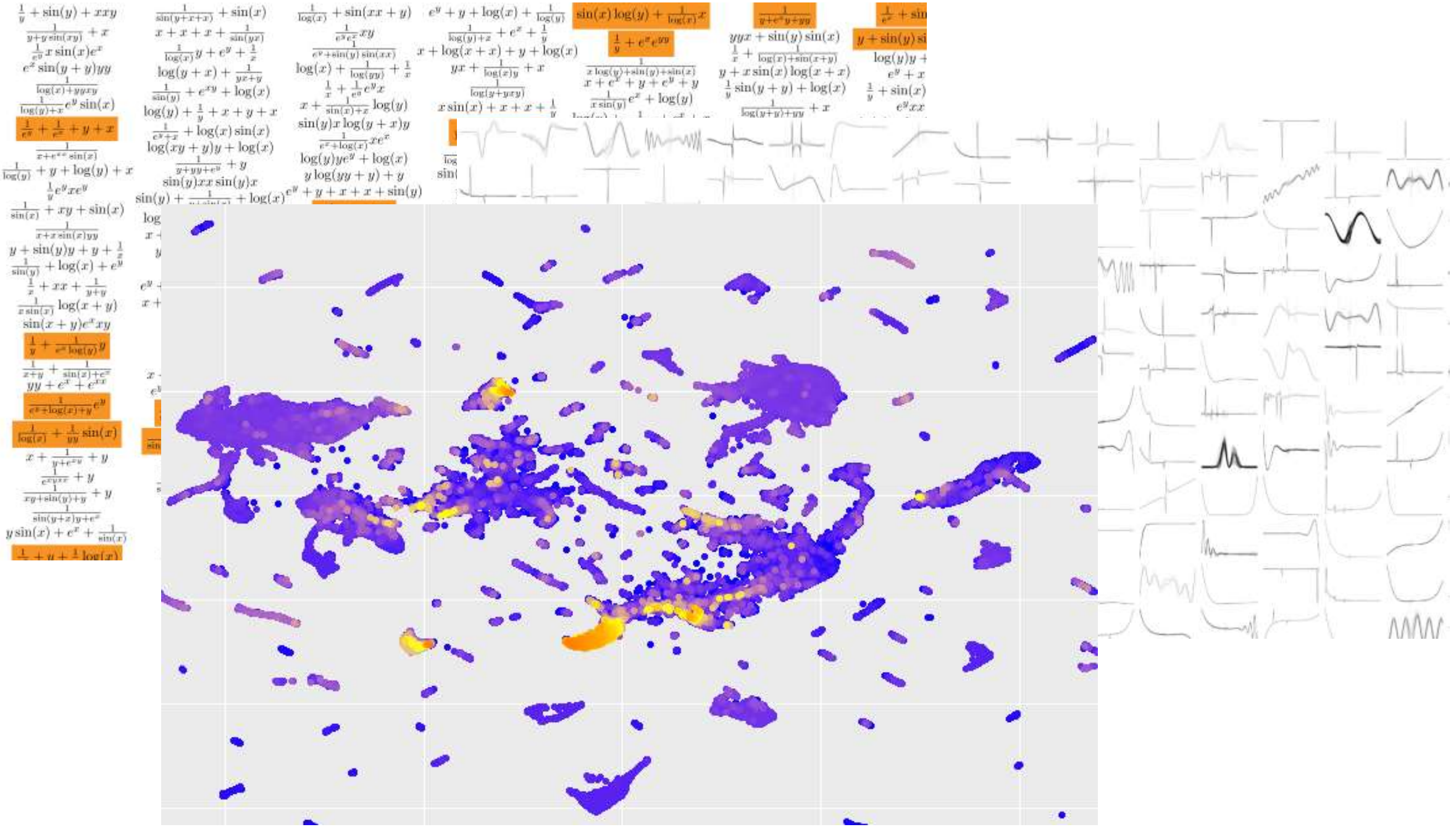



G. Kronberger, L. Kammerer, B. Burlacu, S. Winkler, M. Kommenda, M. Affenzeller

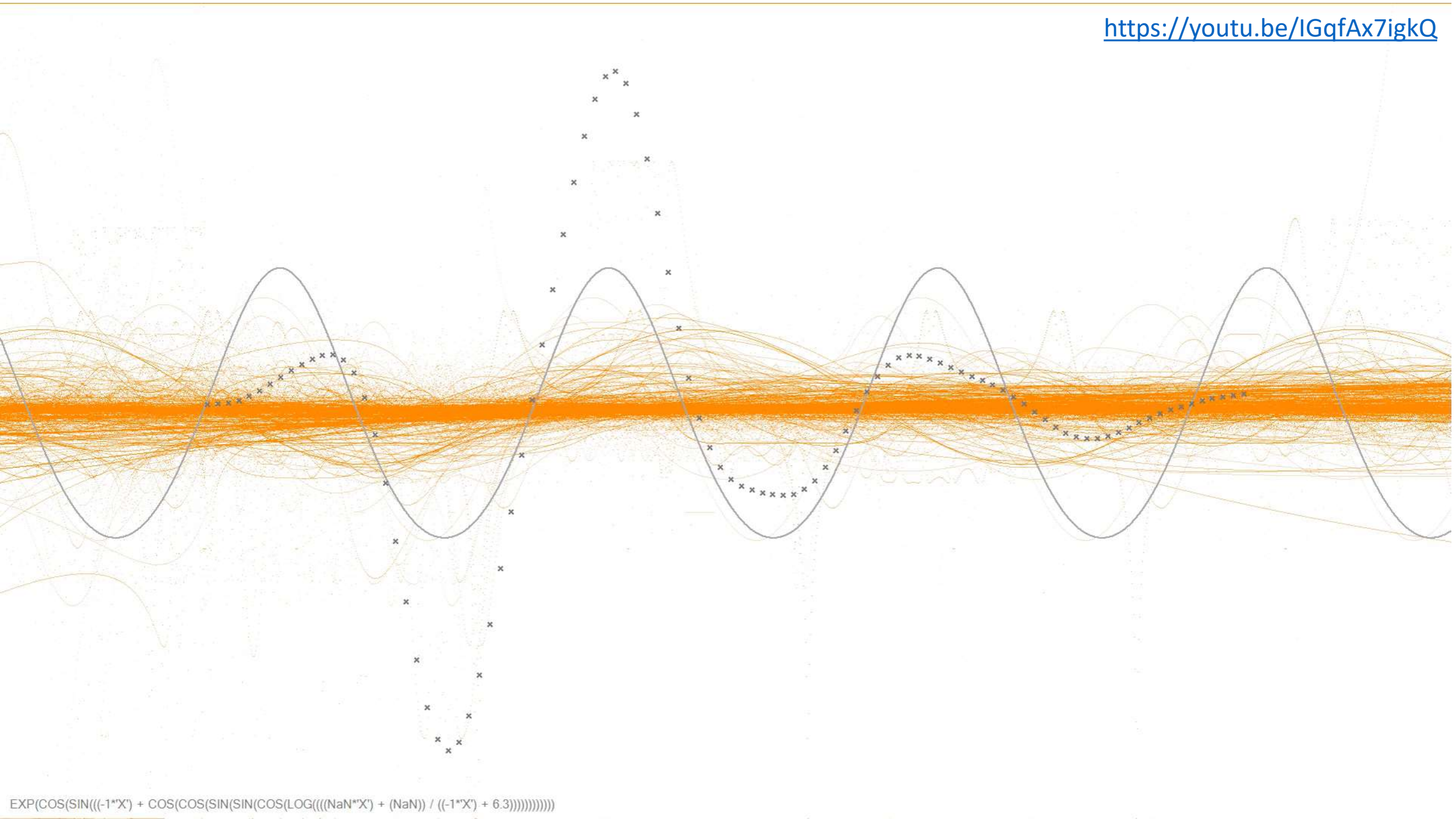
SymReg

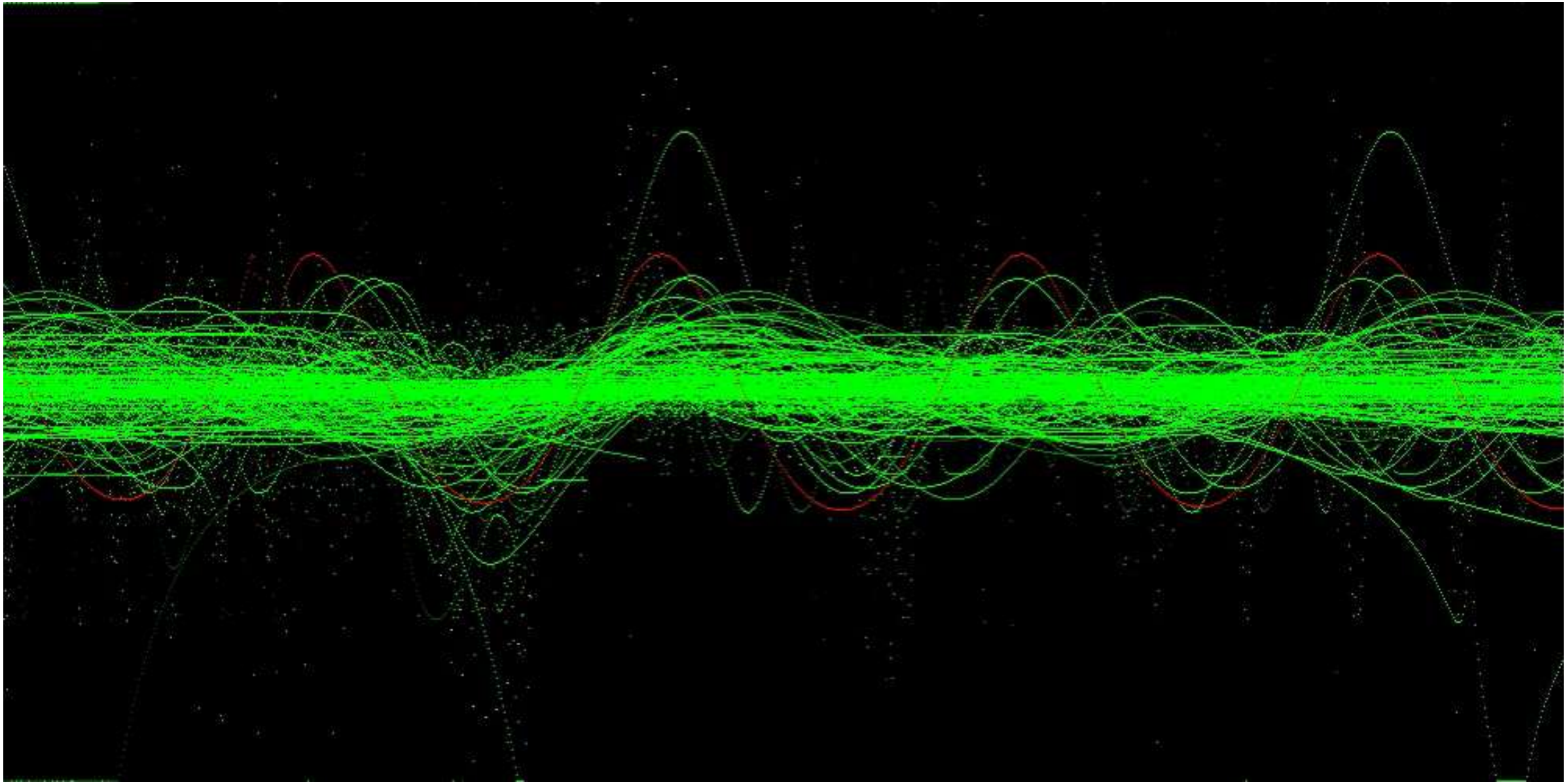
JOSEF RESSEL CENTER FOR SYMBOLIC REGRESSION

FH OBERÖSTERREICH

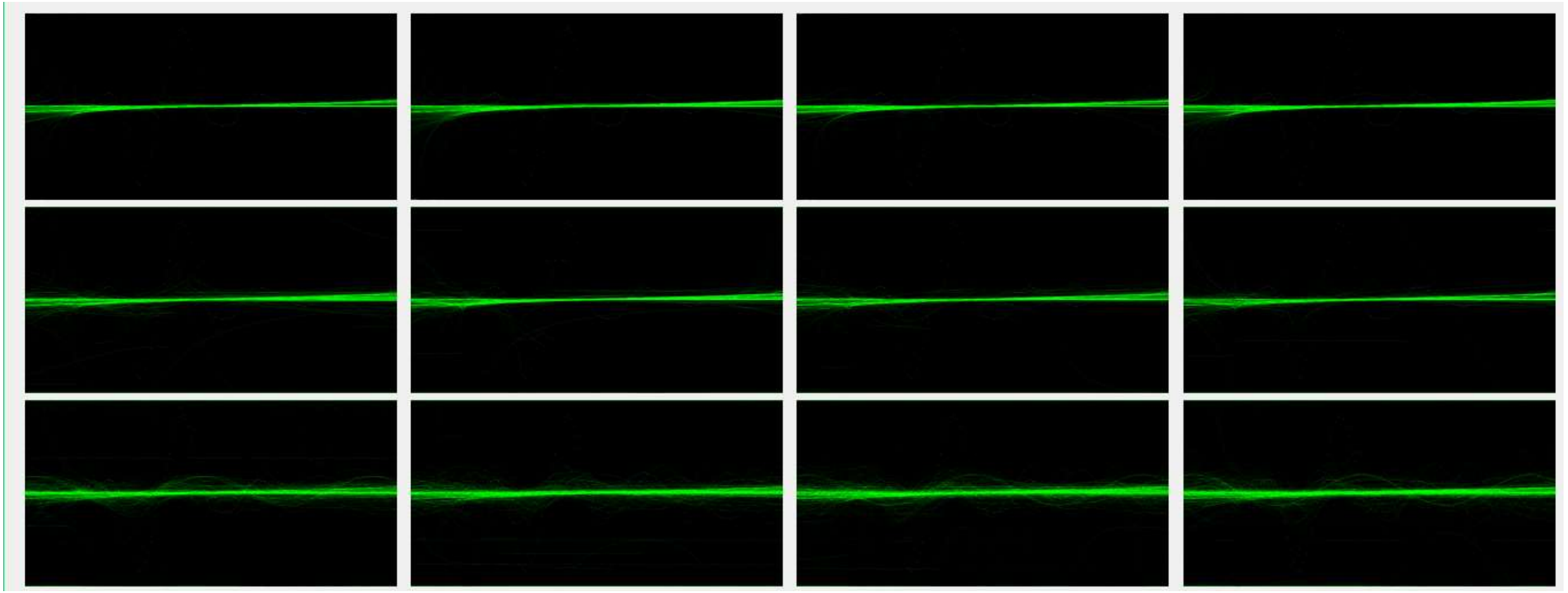


https://youtu.be/IGqfAx7igkQ



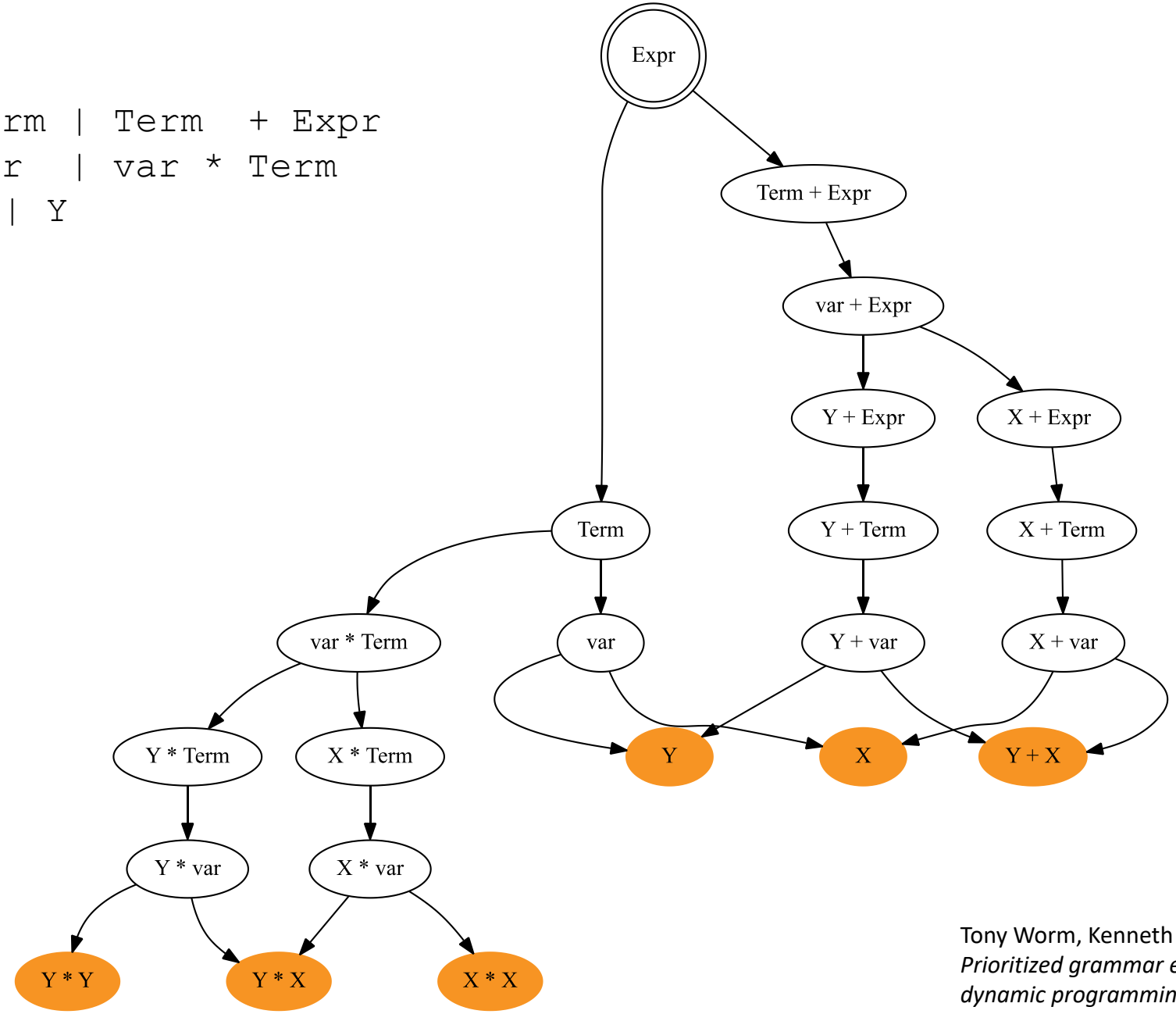


https://youtu.be/yKh8ux68kMc

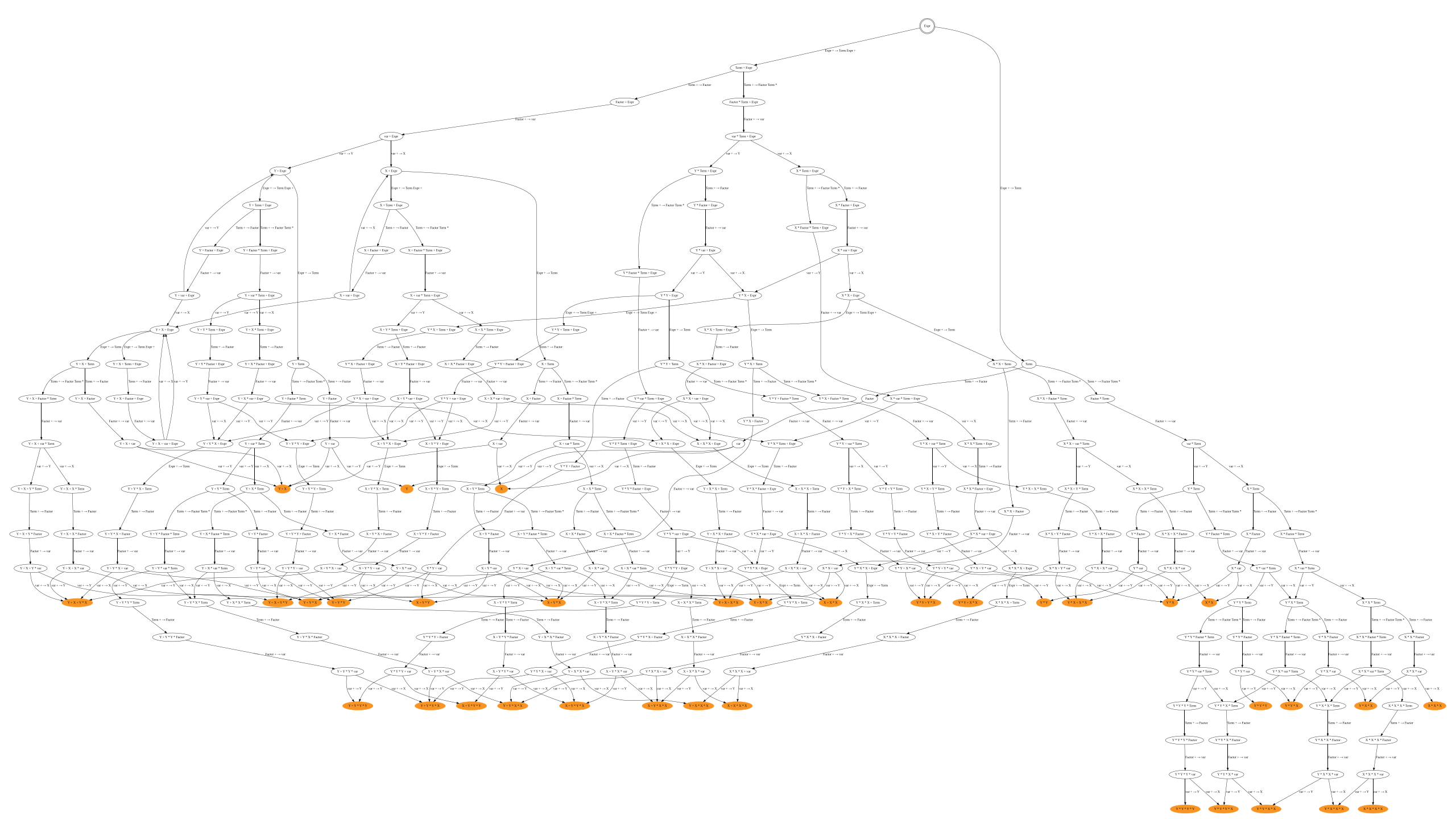


https://youtu.be/9OStdHbTsqk

```
G(Expr):
Expr → Term | Term + Expr
Term → var | var * Term
var → X | Y
```



Tony Worm, Kenneth Chiu:
Prioritized grammar enumeration: symbolic regression by dynamic programming. GECCO 2013: 1021-1028

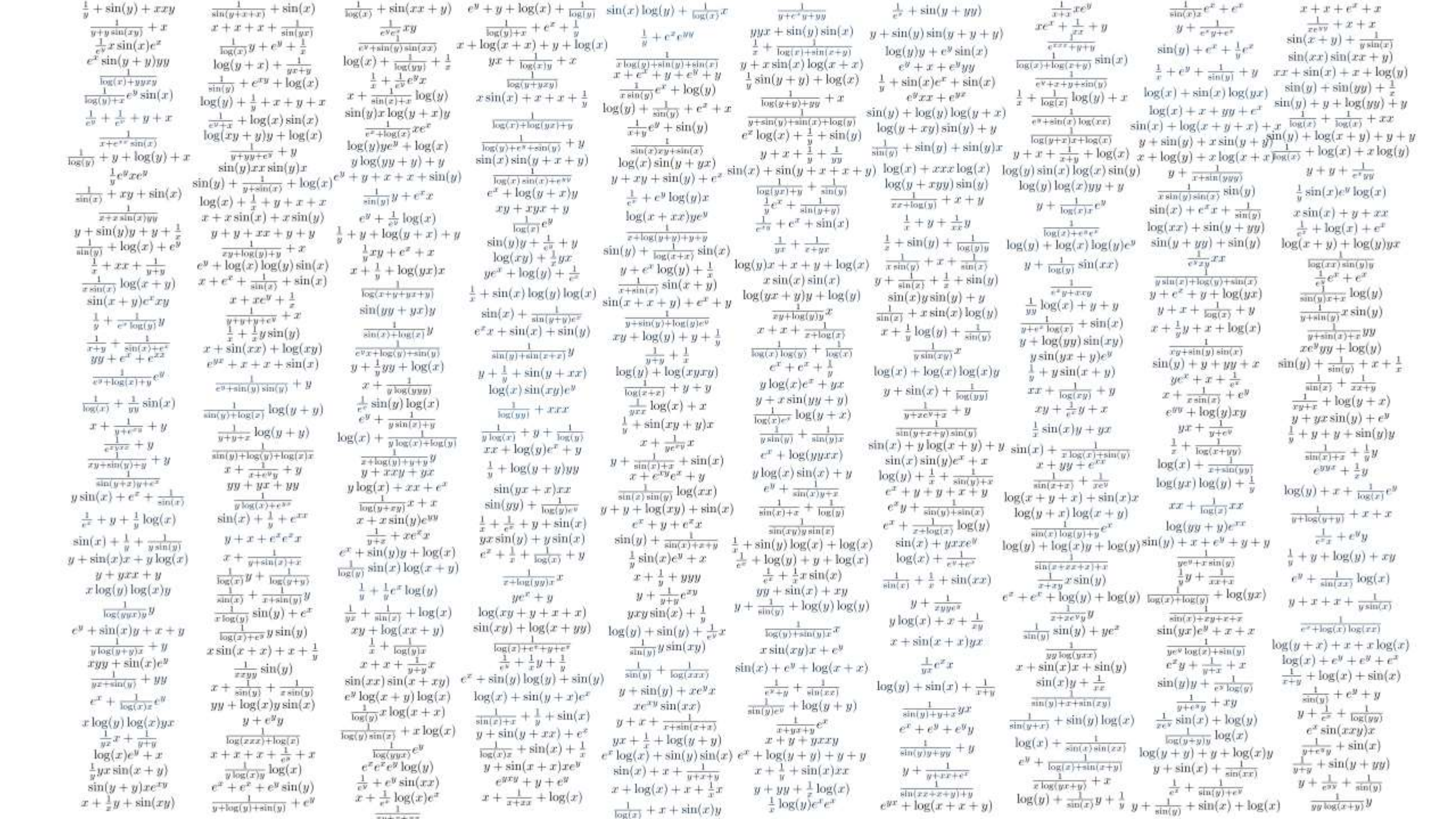

Expr
Expr + → Term Expr +
Expr + → Term
Term + Expr
Term + → Factor
Term + → Factor Term *
Factor + Expr
Factor * Term + Expr
Factor + → var
Factor + → var
var + Expr
var * Term + Expr
var → Y
var → X
var → Y
var → X
Y + Expr
X + Expr
Y * Term + Expr
X * Term + Expr
Expr + → Term Expr +
Expr + → Term Expr +
Term + → Factor
Term + → Factor Term *
Term + → Factor
Term + → Factor Term *
Y + Term + Expr
X + Term + Expr
Y * Factor + Expr
X * Factor + Term + Expr
X * Factor + Expr
var → Y
var → X
var → X
Y + Factor + Expr
Y + Factor * Term + Expr
X + Factor + Expr
X + Factor * Term + Expr
Y * var + Expr
X * var + Expr
Factor + → var
Factor + → var
Factor + → var
Y * Factor * Term + Expr
Expr + → Term
Expr + → Term
Expr + → Term
Y + var + Expr
Y + var * Term + Expr
X + var + Expr
X + var * Term + Expr
Y * Y + Expr
Y * X + Expr
X * X + Expr
var → Y
var → X
var → Y
var → X
var → X
Y + X + Expr
Y + Y * Term + Expr
Y + X * Term + Expr
X + Y * Term + Expr
X + X * Term + Expr
Y * Y * Term + Expr
Y * X * Term + Expr
X * X * Term + Expr
Expr + → Term Expr +
Expr + → Term Expr +
Expr + → Term
Expr + → Term
Expr + → Term
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Y + X + Term
Y + X + Term + Expr
Y + Y * Factor + Expr
Y + X * Factor + Expr
X + Y * Factor + Expr
X + X * Factor + Expr
Y + Y * Factor + Expr
X + Term
Y * Y * Term
X * X * Factor + Expr
Y * X * Term
X * X * Term
Term
Term + → Factor
Term + → Factor Term *
Term + → Factor
Term + → Factor Term *
Term + → Factor
Term + → Factor Term *
Term + → Factor
Term + → Factor Term *
Term + → Factor
Term + → Factor Term *
Term + → Factor
Term + → Factor Term *
Factor
X * var + Term + Expr
X * X + Factor * Term
Factor Term
Y * Y + Factor * Term
Y * X + Factor * Term
Y * X + Factor
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Y + X + var * Term
Y + X + var
Y + X + var + Expr
Y + Y * X + Expr
Y + var * Term
Y + Y * Y + Expr
Y + var
X + Y * Y + Expr
X + Y * Y + Expr
X + var
X + var * Term
Y * Y * Term + Expr
Y * X + Expr
X * X + Expr
var
Y * Y + var * Term
Y * X + var * Term
X * X + Term + Expr
X * X + var * Term
var * Term
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
Y + X + Y * Term
Y + X + X * Term
Y + Y * Term
Y + X * Term
Y + Y * Y + Term
Y + X
X + Y * X + Term
Y
X + Y * Y + Term
X
X + X * Term
Y * Y * Factor + Expr
Y * X * Term
Y * Y * X + Term
Y * X * Factor + Expr
X * X * X + Term
Y * Y + Y * Term
Y * X + Y * Term
X * X + Y * Term
Y * X + X * Term
X * X + X * Term
Y * Y * Term
Y * X * Term
X * Y * Term
X * X * Term
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Y + X + Y * Factor
Y + X + X * Factor
Y + Y * X + Factor
Y + Y * Factor Term
Y + X * Factor Term
Y + Y * Factor
Y + X * Factor
X + Y * X + Factor
X + Y * Y + Factor
X + Y * Factor Term
X + X * Factor
X + X * Factor Term
Y * Y * var + Expr
Y * X * Factor
Y * Y * X + Factor
Y * X * X + Factor
X * X * X + Factor
Y * Y + Y * Factor
Y * X + Y * Factor
X * X + Y * Factor
Y * X + X * Factor
X * X + X * Factor
Y * Y * Factor
Y * X * Factor
X * Y * Factor
X * X * Factor
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Y + X + Y * var
Y + X + X * var
Y + Y * X + var
Y + Y * var * Term
Y + X * var * Term
Y + Y * var
Y + X * var
X + Y * X + var
X + Y * Y + var
X + Y * var * Term
X + X * var
X + X * var * Term
Y * Y * Y + var
Y * Y * X + var
Y * X * Y + var
Y * X * X + var
X * Y * Y + var
X * Y * X + var
X * X * Y + var
X * X * X + var
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
Y + X + Y * Y
Y + X + Y * X
Y + Y * Y + Term
Y + Y * X + Term
Y + X * Y + Term
Y + X * X + Term
X + Y * Y + Term
X + Y * X + Term
X + X * Y + Term
X + X * X + Term
Y * Y * Y + Term
Y * Y * X + Term
Y * X * Y + Term
Y * X * X + Term
X * Y * Y + Term
X * Y * X + Term
X * X * Y + Term
X * X * X + Term
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Y + Y * Y * Factor
Y + Y * X * Factor
Y + X * Y * Factor
Y + X * X * Factor
X + Y * Y * Factor
X + Y * X * Factor
X + X * Y * Factor
X + X * X * Factor
Y * Y * Y * Factor
Y * Y * X * Factor
Y * X * Y * Factor
Y * X * X * Factor
X * Y * Y * Factor
X * Y * X * Factor
X * X * Y * Factor
X * X * X * Factor
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Y + Y + Y * var
Y + Y + X * var
Y + X + Y * var
Y + X + X * var
X + Y + Y * var
X + Y + X * var
X + X + Y * var
X + X + X * var
Y * Y * Y * var
Y * Y * X * var
Y * X * Y * var
Y * X * X * var
X * Y * Y * var
X * Y * X * var
X * X * Y * var
X * X * X * var
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
Y + Y + Y * Y
Y + Y + Y * X
Y + Y + X * Y
Y + Y + X * X
Y + X + Y * Y
Y + X + Y * X
Y + X + X * Y
Y + X + X * X
X + Y + Y * Y
X + Y + Y * X
X + Y + X * Y
X + Y + X * X
X + X + Y * Y
X + X + Y * X
X + X + X * Y
X + X + X * X
Y * Y * Y + Term
Y * Y * X + Term
Y * X * Y + Term
Y * X * X + Term
X * Y * Y + Term
X * Y * X + Term
X * X * Y + Term
X * X * X + Term
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Term + → Factor
Y * Y * Y * Factor
Y * Y * X * Factor
Y * X * Y * Factor
Y * X * X * Factor
X * Y * Y * Factor
X * Y * X * Factor
X * X * Y * Factor
X * X * X * Factor
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Factor + → var
Y * Y * Y * var
Y * Y * X * var
Y * X * Y * var
Y * X * X * var
X * Y * Y * var
X * Y * X * var
X * X * Y * var
X * X * X * var
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
var → Y
var → X
Y * Y * Y * Y
Y * Y * Y * X
Y * Y * X * Y
Y * Y * X * X
Y * X * Y * Y
Y * X * Y * X
Y * X * X * Y
Y * X * X * X
X * Y * Y * Y
X * Y * Y * X
X * Y * X * Y
X * Y * X * X
X * X * Y * Y
X * X * Y * X
X * X * X * Y
X * X * X * X

```
G(Expr):
Expr   -> Term "+" Expr | Term
Term   -> Factor "*" Term | Factor | "1/(" InvExpr ")"
Factor -> VarFac | ExpFac | LogFac | SinFac
VarFac -> <variable>
ExpFac -> "exp(" SimpleTerm ")"
LogFac -> "log(" SimpleExpr ")"
SinFac -> "sin(" SimpleExpr ")"

SimpleExpr -> SimpleTerm "+" SimpleExpr | SimpleTerm
SimpleTerm -> VarFac "*" SimpleTerm | VarFac

InvExpr -> InvTerm "*" InvExpr | InvTerm
InvTerm -> Factor "*" InvTerm | Factor
```

$$\frac{1}{x} + \sin(y) + x^2 y$$

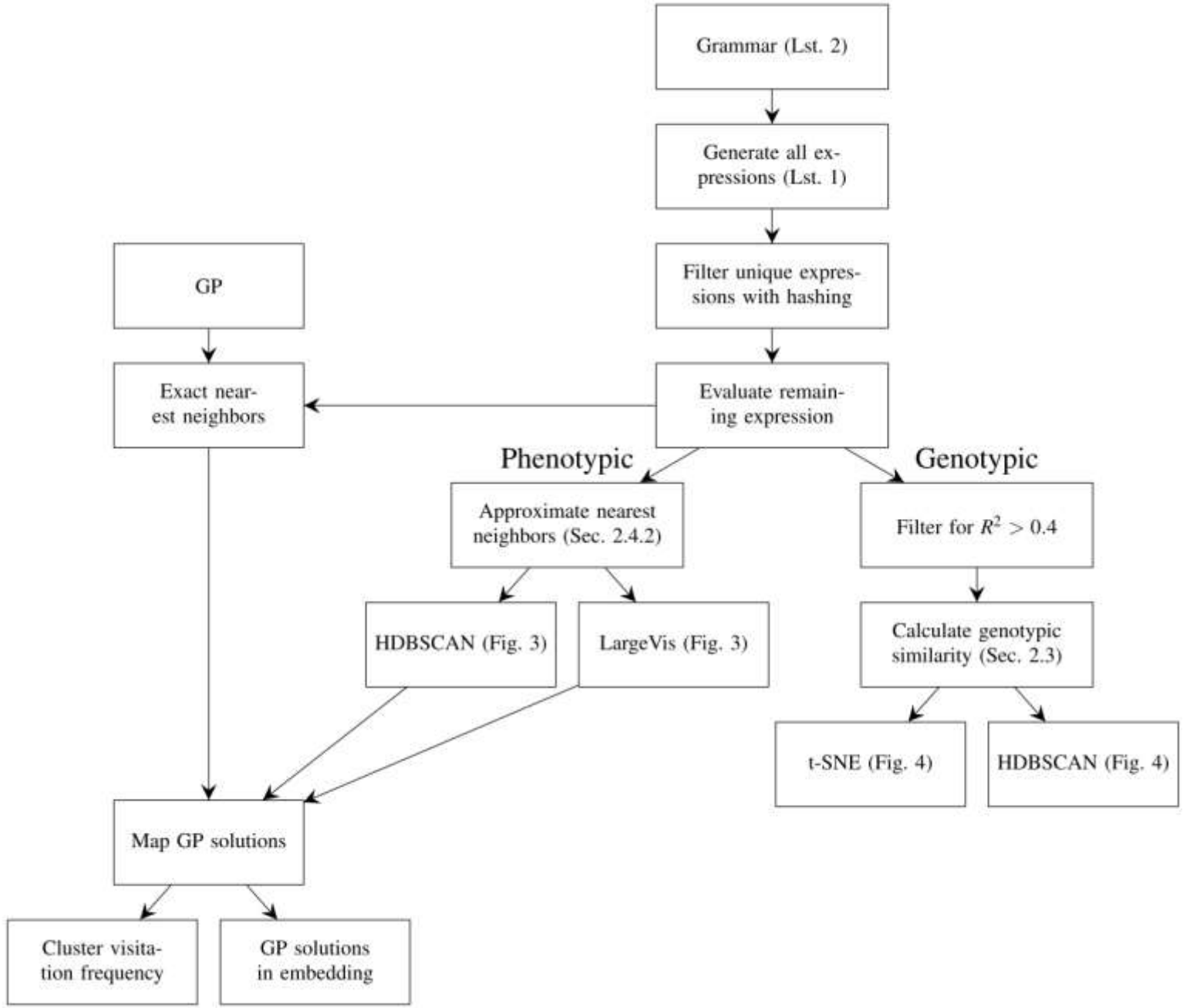
$$\frac{1}{e^y} x \sin(x) e^x$$

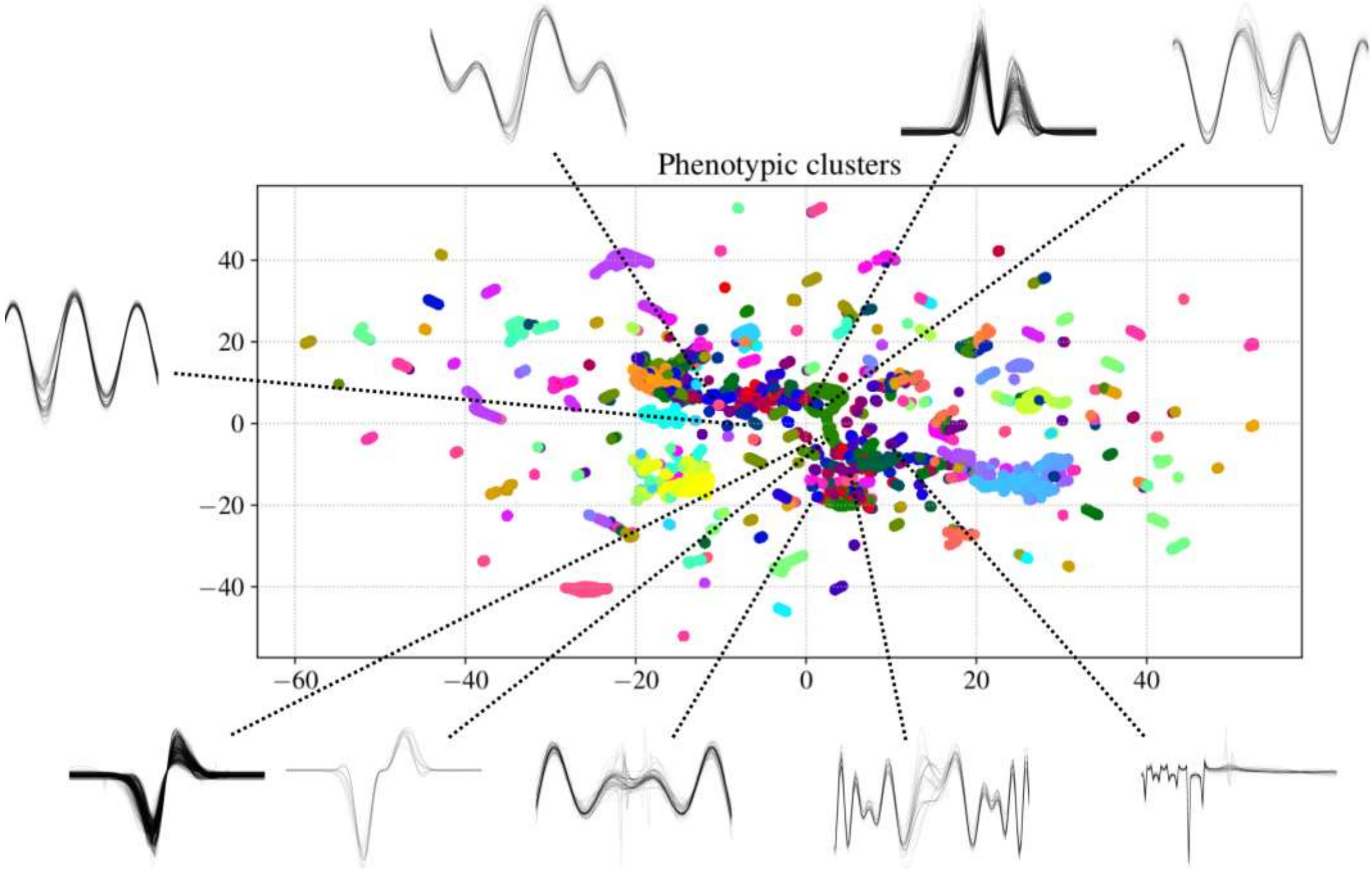
$$\frac{1}{\log(x) + x y^3}$$

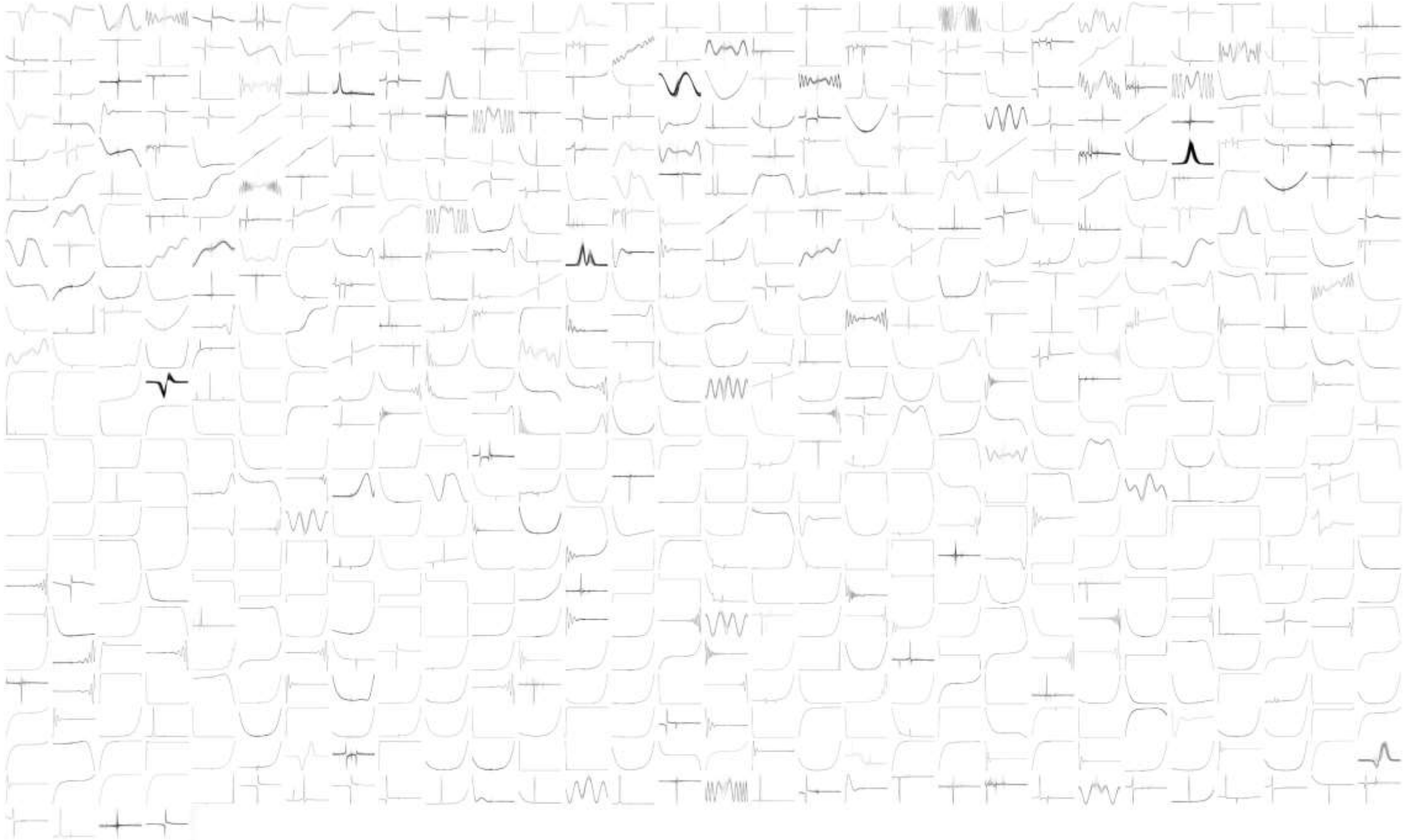


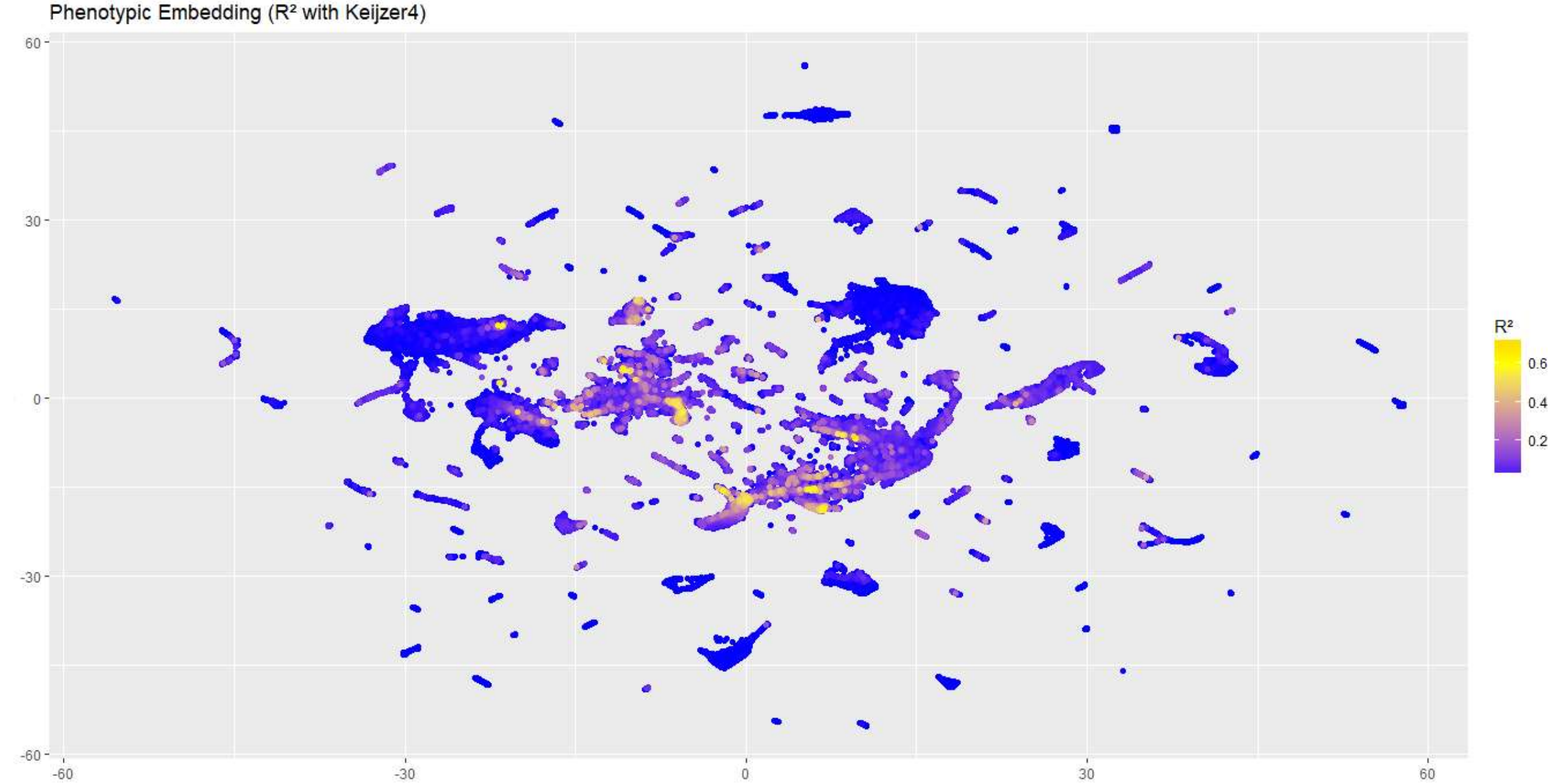
How are symbolic regression models distributed in the search space?

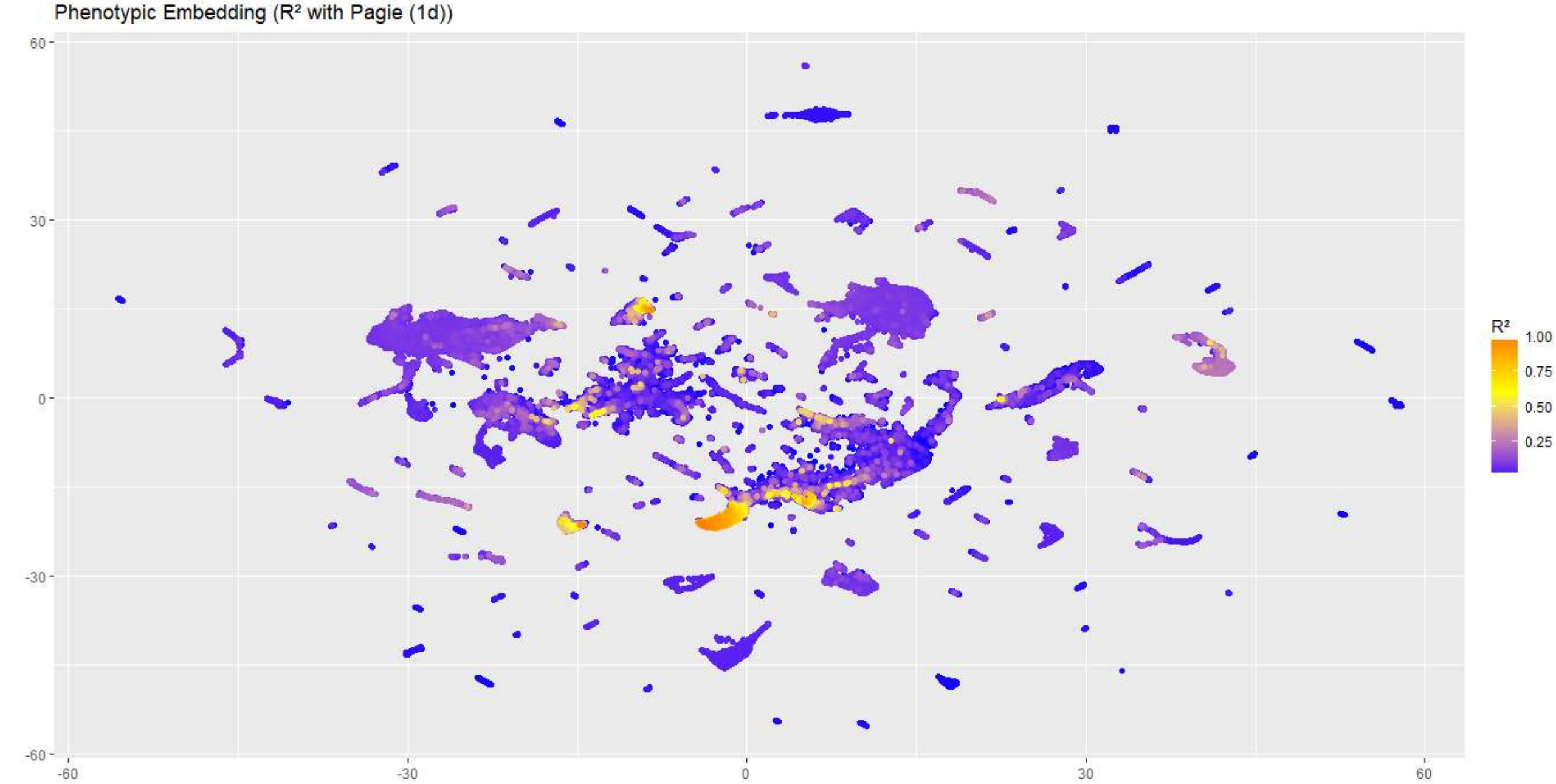
Which parts of the search space are visited by genetic programming?


Grammar (Lst. 2)
Generate all expressions (Lst. 1)
Filter unique expressions with hashing
Evaluate remaining expression
GP
Exact nearest neighbors
Phenotypic
Genotypic
Approximate nearest neighbors (Sec. 2.4.2)
Filter for $R^2 > 0.4$
HDBSCAN (Fig. 3)
LargeVis (Fig. 3)
Calculate genotypic similarity (Sec. 2.3)
t-SNE (Fig. 4)
HDBSCAN (Fig. 4)
Map GP solutions
Cluster visitation frequency
GP solutions in embedding

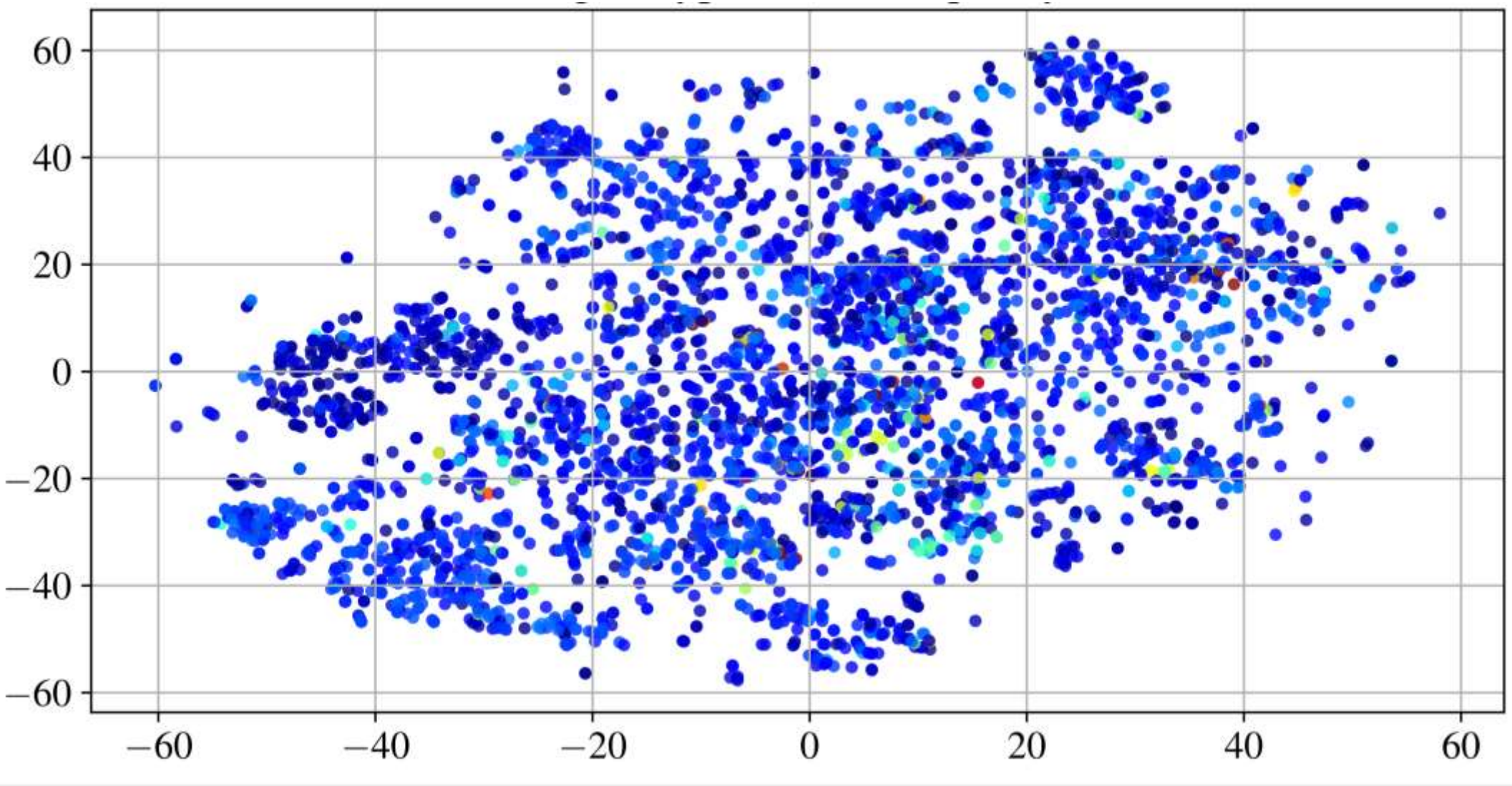

Phenotypic clusters
40
20
0
−20
−40
−60
−40
−20
0
20
40



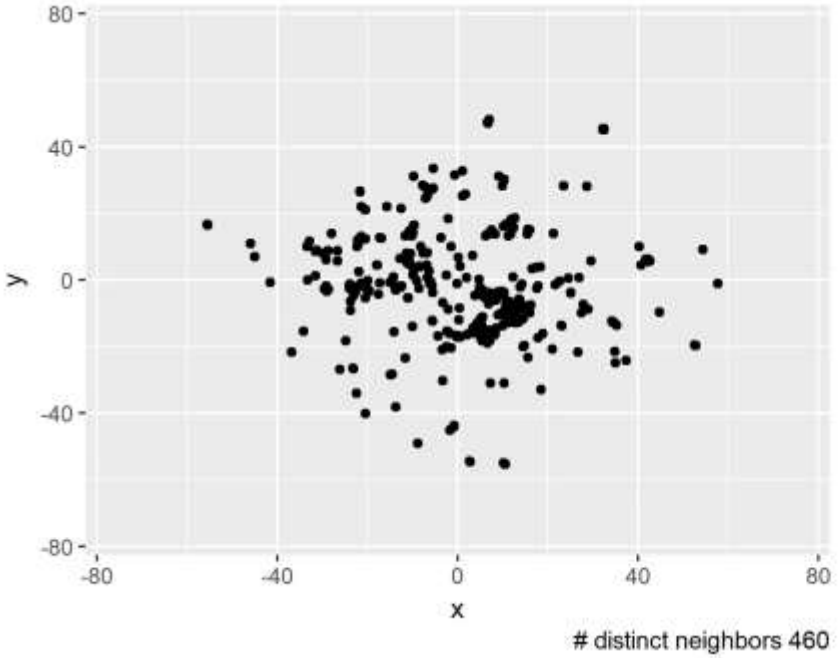

Phenotypic Embedding (R² with Keijzer4)
60
30
0
-30
-60
-60
-30
0
30
60
R²
0.6
0.4
0.2

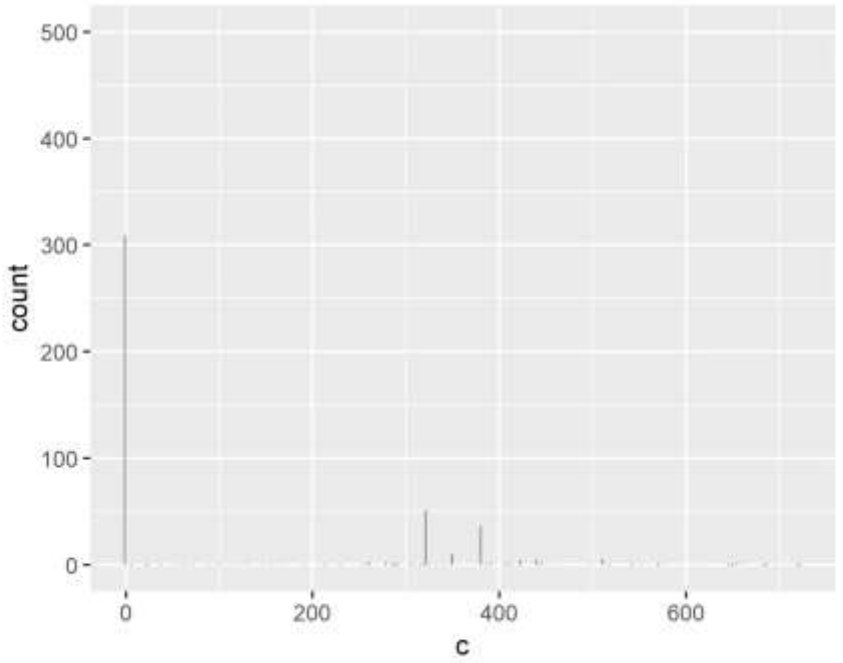

Phenotypic Embedding (R² with Pagie (1d))
60
30
0
-30
-60
-60
-30
0
30
60
R²
1.00
0.75
0.50
0.25

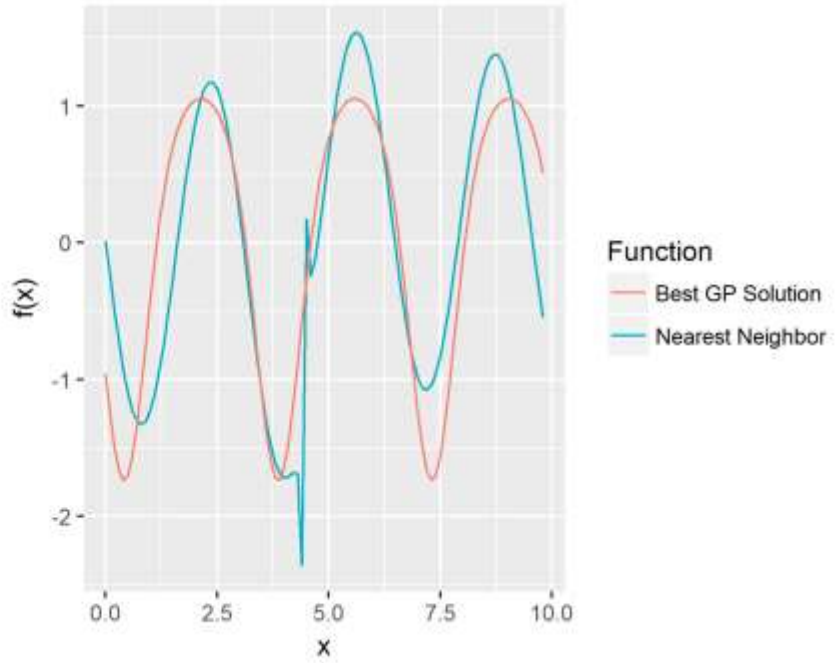
Genotypic Embedding



Mapping Populations

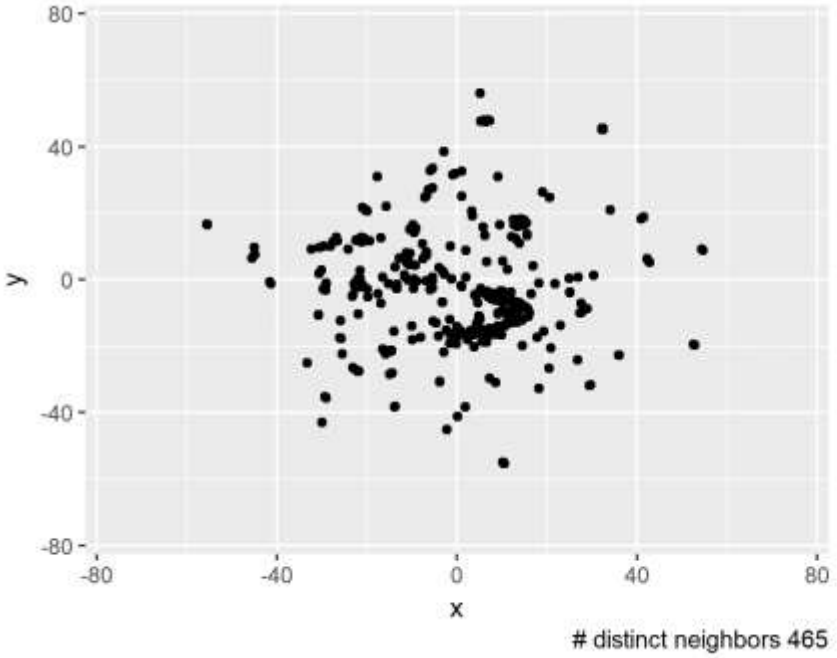


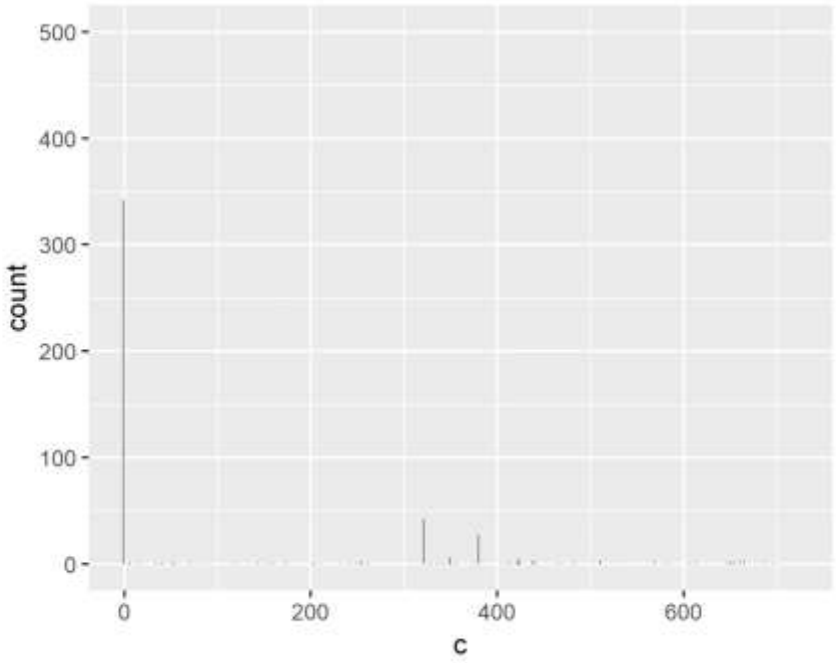


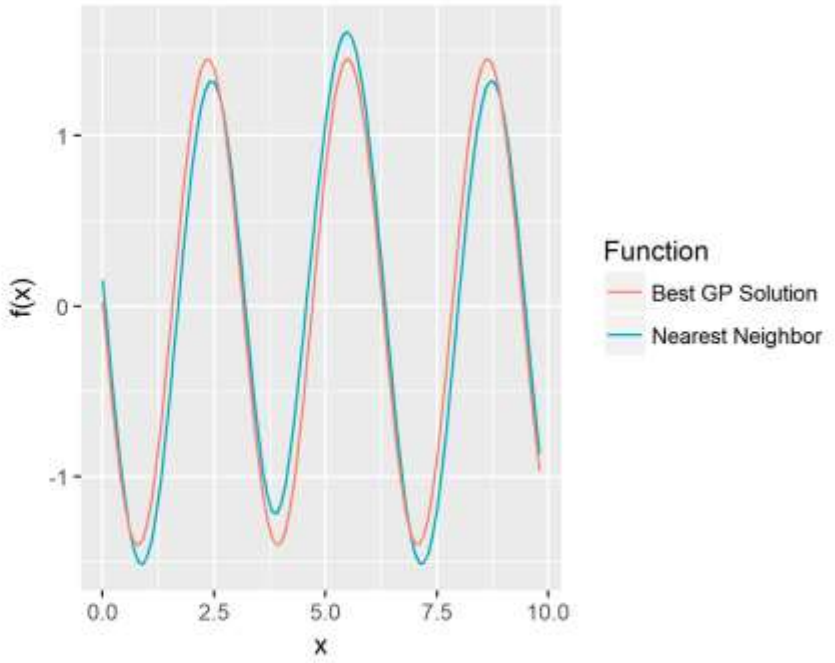


Standard GP, PopSize: 500 , Tournament: 7
https://youtu.be/Xtx_NKioazo

Mapping Populations







GP with offspring selection, PopSize: 500 , Random
https://youtu.be/iYJEQXTPadE

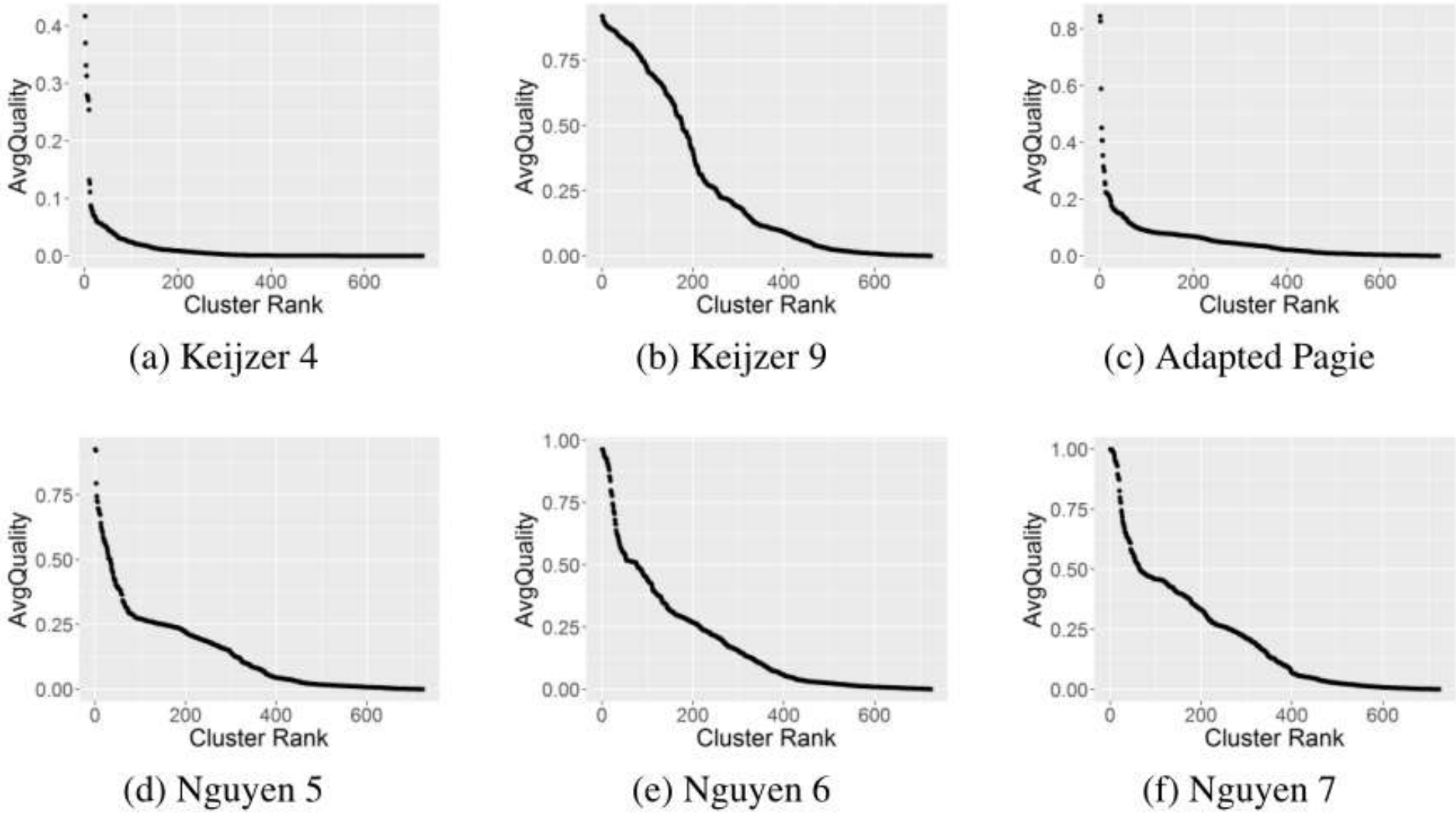
Discussion

- What are the reasons for the low correlation between phenotypic and genotypic similarities?
- Is it possible to apply the same approach to multi-variate models?
- How should we include parameters?
- How could we use the neighborhood graph in GP?
- Are the insights transferable to real-world problems?

…

SymReg

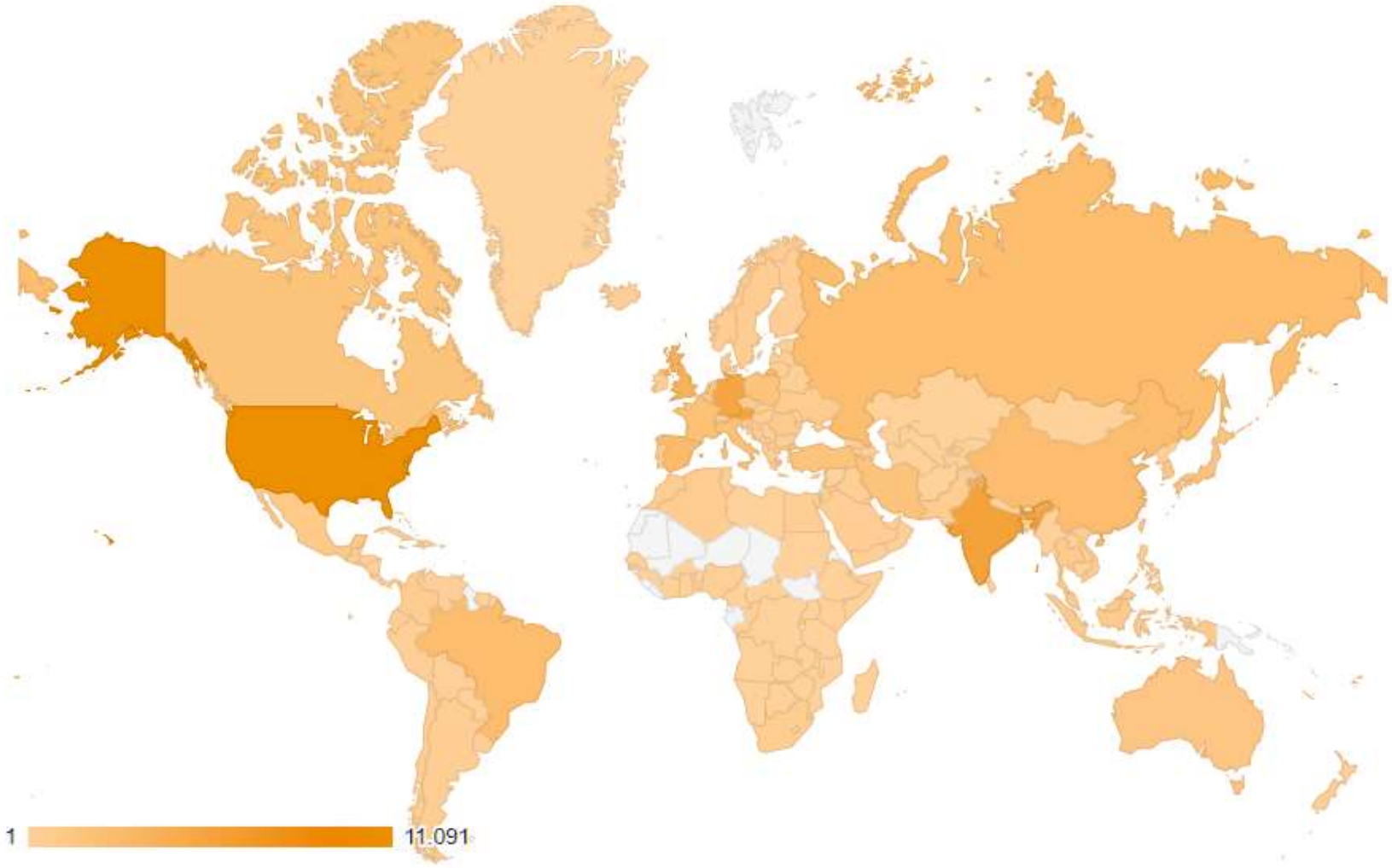
JOSEF RESSEL CENTER FOR SYMBOLIC REGRESSION



(a) Keijzer 4 (b) Keijzer 9 (c) Adapted Pagie

(d) Nguyen 5 (e) Nguyen 6 (f) Nguyen 7

Fig. 5 Ranking of clusters by average R^2 values of expressions within all clusters. For each of the benchmark functions there are clusters which contain well-fitting expressions.



HeuristicLab

Land	Nutzer (dev.heuristiclab.com)
USA	11091
Austria	5766
India	5678
Germany	5096
UK	3743
Spain	2809
Russia	2645
Brazil	2607
China	2575
France	2099

Stand: Feb. 2018