Symbolic Regression by Exhaustive Search

Reducing the Search Space Using Syntactical Constraints and Efficient Semantic Structure Deduplication.

Lukas Kammerer, Gabriel Kronberger, Bogdan Burlacu,
Stephan Winkler, Michael Kommenda and Michael Affenzeller

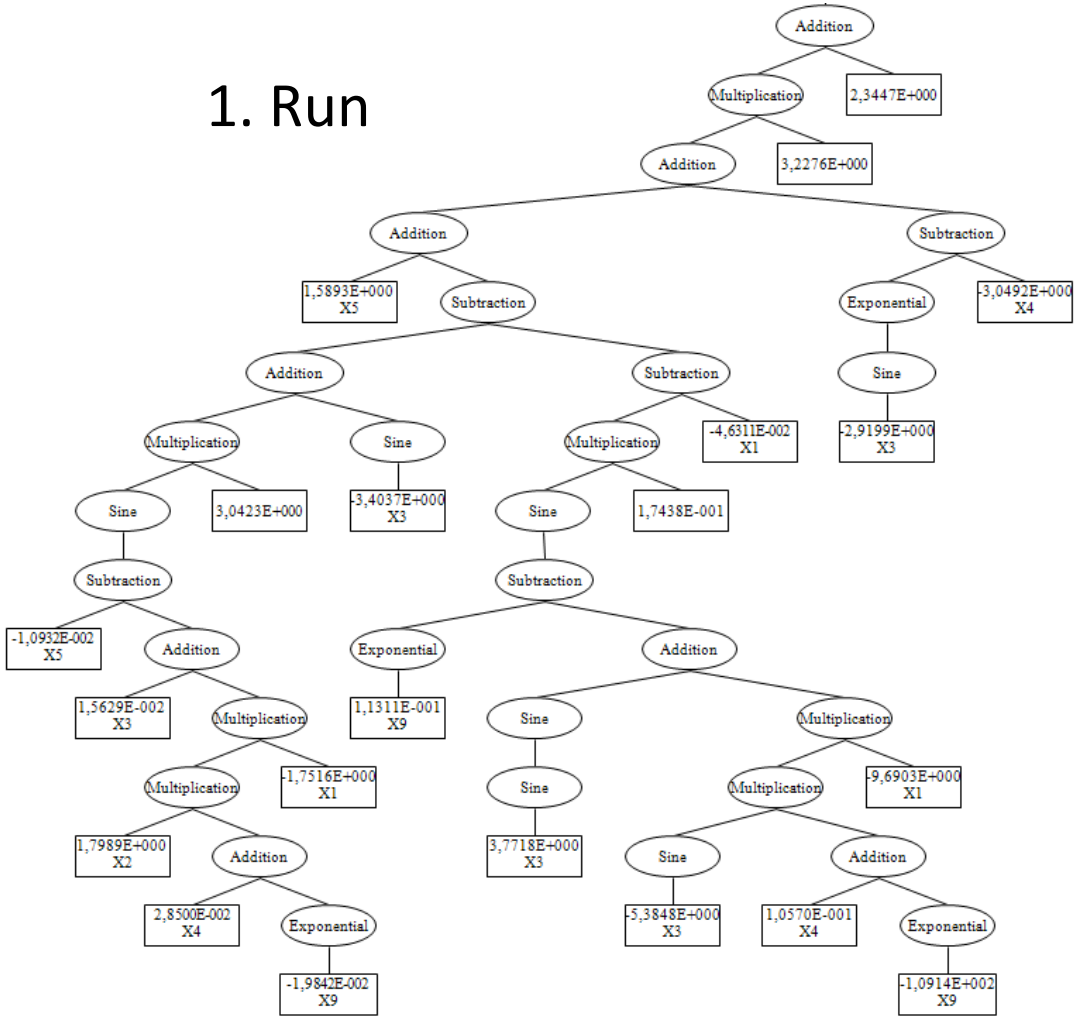
UNIVERSITY
OF APPLIED SCIENCES
UPPER AUSTRIA

HEAL

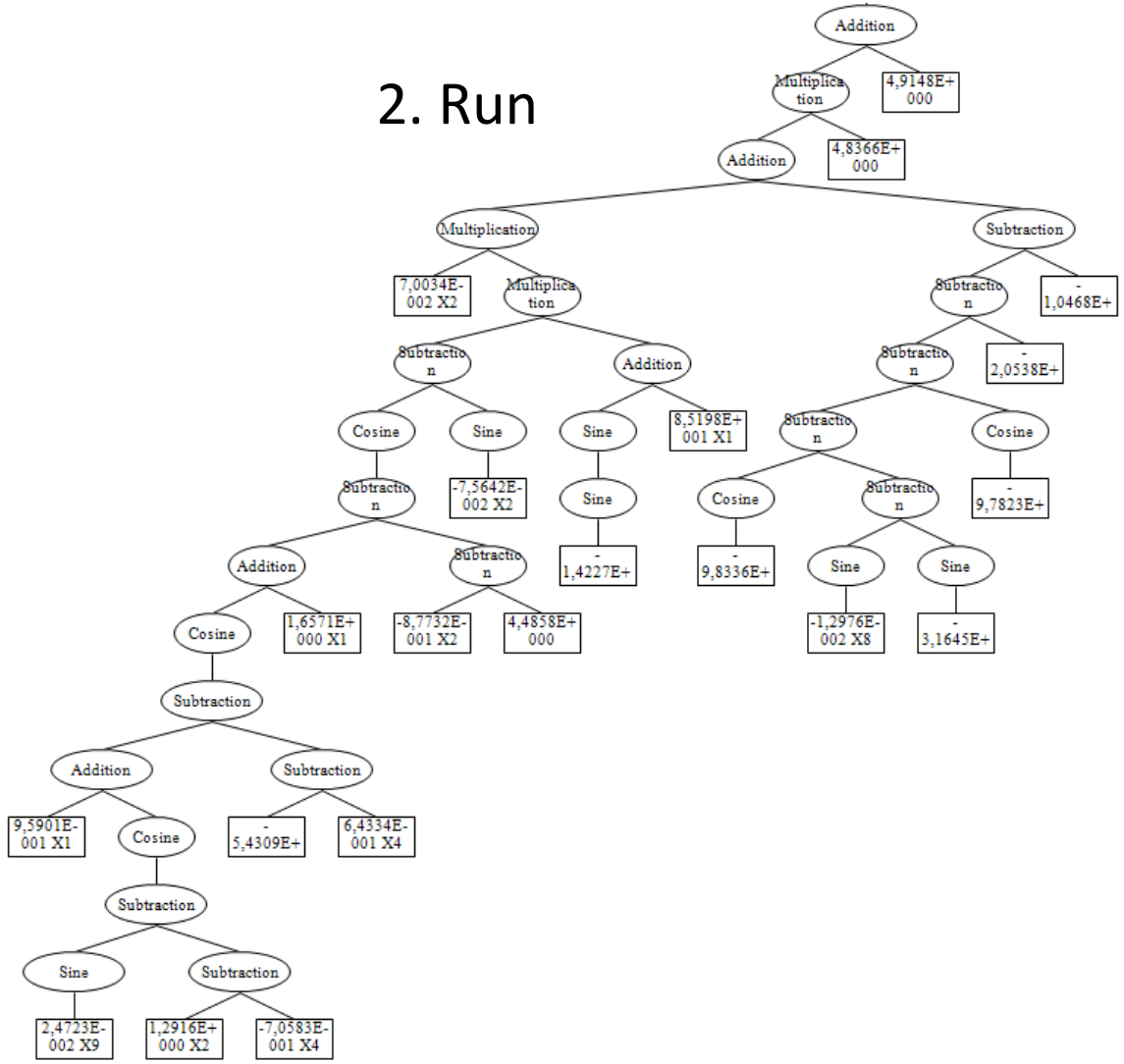
HEURISTIC AND EVOLUTIONARY
ALGORITHMS LABORATORY

Motivation

1. Run



2. Run



Previous Work

- Trent McConaghy. "FFX: Fast, Scalable, Deterministic Symbolic Regression Technology." *GPTP IX*. 2011

- Tony Worm and Kenneth Chiu. "Prioritized Grammar Enumeration: Symbolic Regression by Dynamic Programming." *Proceedings of the 15th GECCO*. 2013.

- Michael Korns. "A baseline symbolic regression algorithm." *GPTP X*. 2013

- Gabriel Kronberger et al. „Cluster Analysis of a Symbolic Regression Search Space.“ *GPTP XVI*.

- Discussions at GPTP XVII

General Idea

1. Separate structure and coefficients

$$f(x,y) = 2y + 3x + 4x(5x + 6y)$$

$$f(x,y) = c_1 y + c_2 x + c_3 x(c_4 x + c_5 y)$$

cf. Korns, Michael. "A baseline symbolic regression algorithm." GPTP X. 2013.

cf. Worm, Tony, and Kenneth Chiu. "Prioritized grammar enumeration: symbolic regression by dynamic programming." *Proceedings of the 15th GECCO*. 2013.

Kommenda, Michael, et al. "Effects of constant optimization by nonlinear least squares minimization in symbolic regression." *Proceedings of the 15th GECCO*. 2013.

General Idea

1. Separate structure and coefficients

2. Restrict the search space

$$f(x,y) = c_1 y + c_2 x + \underline{c_3 x(c_4 x + c_5 y)}$$

$$f(x,y) = c_1 y + c_2 x + \underline{c_3 x^2 + c_4 xy}$$

General Idea

1. Separate structure and coefficients

2. Restrict the search space

$$f(x,y) = \sin\left(\mathrm{e}^{\mathrm{e}^{yx}x}\cos\left(\mathrm{e}^{-29}x\right)\right)^{\log(\sin(4x))}$$

General Idea

1. Separate structure and coefficients

2. Restrict the search space

3. Generate all solutions

$$f(x,y) = c_1 x + c_2 x$$

$$f(x,y) = c_1 x + c_2 x^2$$

$$f(x,y) = c_1 x + c_2 x^3$$

$$f(x,y) = c_1 x + c_2 y$$

$$\vdots$$

General Idea

1. Separate structure and coefficients
2. Restrict the search space
3. Generate all solutions
4. Identify semantic duplicates

$$f(x,y) = c_1 y + c_2 x + c_3 x^2 + c_4 xy$$

$$f(x,y) = c_1 x + c_2 x^2 + c_3 y + c_4 xy$$

Search Space

$$G(Expr):$$

$$Expr = c_0 \ + \ c_1 \ Term_1 \ + \ c_2 \ Term_2 \ + \ ...$$

$$Term_i = Factor_0 \ Factor_1 ...$$

$$Factor_i \in \{var, \frac{1}{Expr}, \log(arg), \exp(arg), \sin(arg), \sqrt{arg}, \sqrt[3]{arg}\}$$

$$arg = c_0 + c_1 var \ var ... + c_2 var \ var ... + ...$$

Search Space

$$G(Expr):$$

$$Expr = c_0 \ + \ c_1 \ Term_1 \ + \ c_2 \ Term_2 \ + \ ...$$

$$Term_i = Factor_0 \ Factor_1 ...$$

$$Factor_i \in \{var, \frac{1}{Expr}, \log(arg), \exp(arg), \sin(arg), \sqrt{arg}, \sqrt[3]{arg}\}$$

$$arg = c_0 + c_1 var \ var ... + c_2 var \ var ... + ...$$

Example: $c_0 + c_1 xxy + c_2 xx$

$$c_0 + c_1 xxy + c_2 \log(c_3 + c_4 xy + x)$$

$$c_0 + c_1 x \frac{1}{c_2 + \sin(c_3 + c_4 x + c_5 y)}$$

Search Space

Prevent:

$$x(x+y)$$

$$\sin(\sin(x))$$

$$\frac{1}{\frac{1}{x}}$$

$$\frac{1}{x+y}\ \frac{1}{x}$$

$$\sqrt{x}\ \sqrt{x}$$

$$\exp(x+x)$$

$$\vdots$$

Search Space

```
G(Expr):

Expr -> "const" "*" Term "+" Expr |
        "const" "*" Term "+" "const"

Term -> RecurringFactors "*" Term | RecurringFactors |
        OneTimeFactors


RecurringFactors -> VarFactor | LogFactor |
                    ExpFactor | SinFactor


VarFactor  -> <variable>

LogFactor  -> "log" "(" SimpleExpr ")"

ExpFactor  -> "exp" "(" "const" "*" SimpleTerm ")"

SinFactor  -> "sin" "(" SimpleExpr ")"


OneTimeFactors -> <every combination of InvFactor,
                   SqrtFactor and CbrtFactor>
```

```
InvFactor  -> "1/" "(" InvExpr ")"

SqrtFactor -> "sqrt" "(" SimpleExpr ")"

CbrtFactor -> "cbrt" "(" SimpleExpr ")"


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

              "const" "*" SimpleTerm "+" "const"

SimpleTerm -> VarFactor "*" SimpleTerm | VarFactor


InvExpr -> "const" "*" InvTerm "+" InvExpr |

           "const" "*" InvTerm "+" "const"

InvTerm -> RecurringFactors "*" InvTerm  |

           <every combination of RecurringFactors,
            SqrtFactor and CbrtFactor>
```

Identifying Semantic Duplicates

$$c_0 + c_1 yx = c_0 + c_1 xy$$

$$c_0 + c_1 x + c_2 x = c_0 + c_3 x$$

$$c_0 + c_1 \mathrm{e}^{c_2 x} \mathrm{e}^{c_3 x} = c_0 + c_1 \mathrm{e}^{c_3 x}$$

$$c_0 + c_1 \exp(x) + c_2 \exp(x) \neq c_0 + c_3 \exp(x)$$

$$c_0 + c_1 \sin(x) + c_2 \sin(x) \neq c_0 + c_3 \sin(x)$$

Identifying Semantic Duplicates

$$c_0 + c_1 yx = c_0 + c_1 xy$$

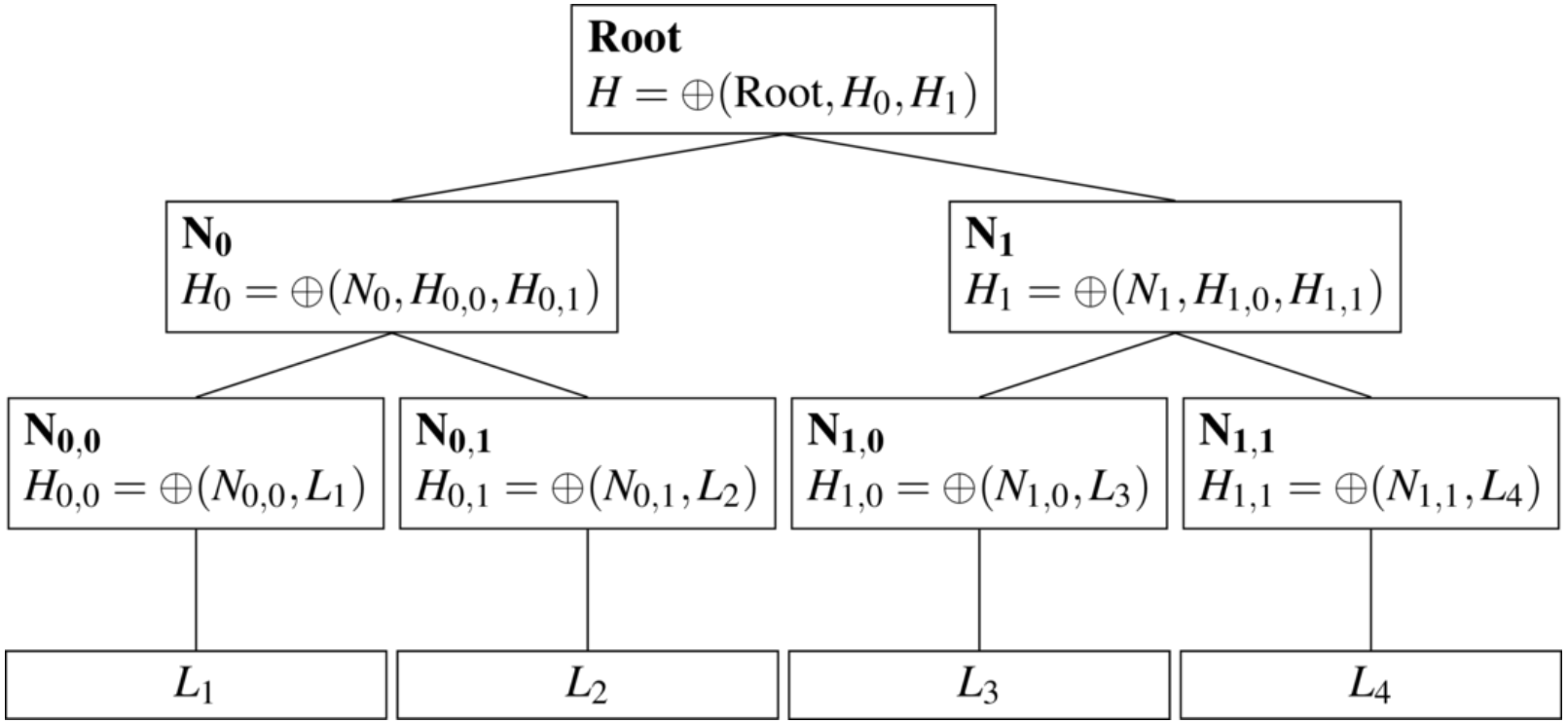
$$c_0 + c_1 x + c_2 x = c_0 + c_3 x$$

$$c_0 + c_1 \mathrm{e}^{c_2 x} \mathrm{e}^{c_3 x} = c_0 + c_1 \mathrm{e}^{c_3 x}$$

$$c_0 + c_1 \exp(c_2 + c_3 x) + c_4 \exp(c_5 + c_6 x) \neq c_0 + c_7 \exp(c_8 + c_9 x)$$

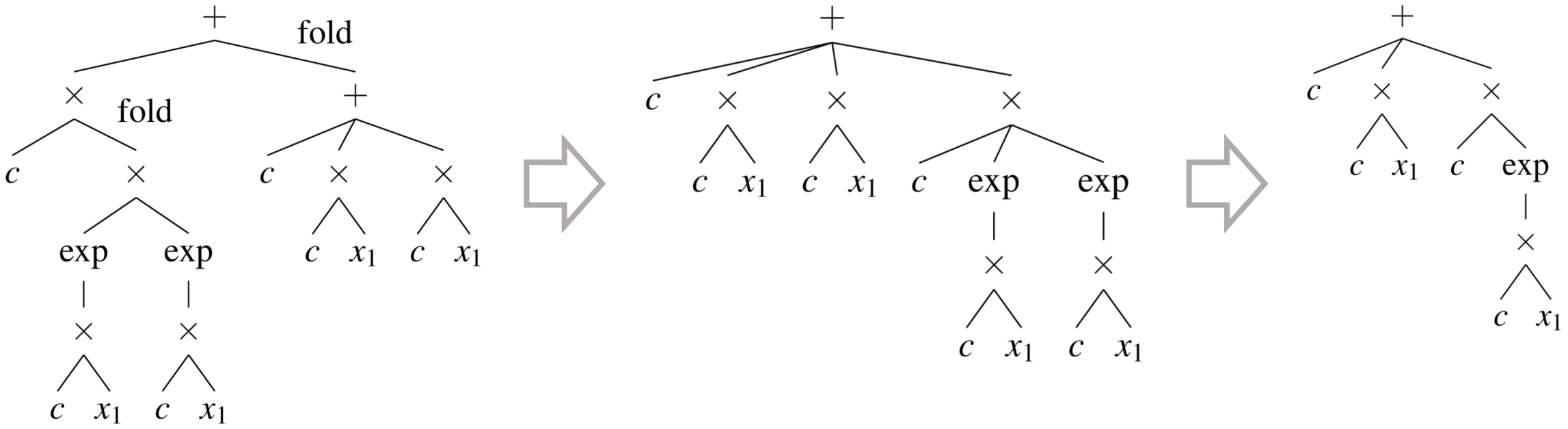
$$c_0 + c_1 \sin(c_2 + c_3 x) + c_4 \sin(c_5 + c_6 x) \neq c_0 + c_7 \sin(c_8 + c_9 x)$$

Semantic Expression Hashing



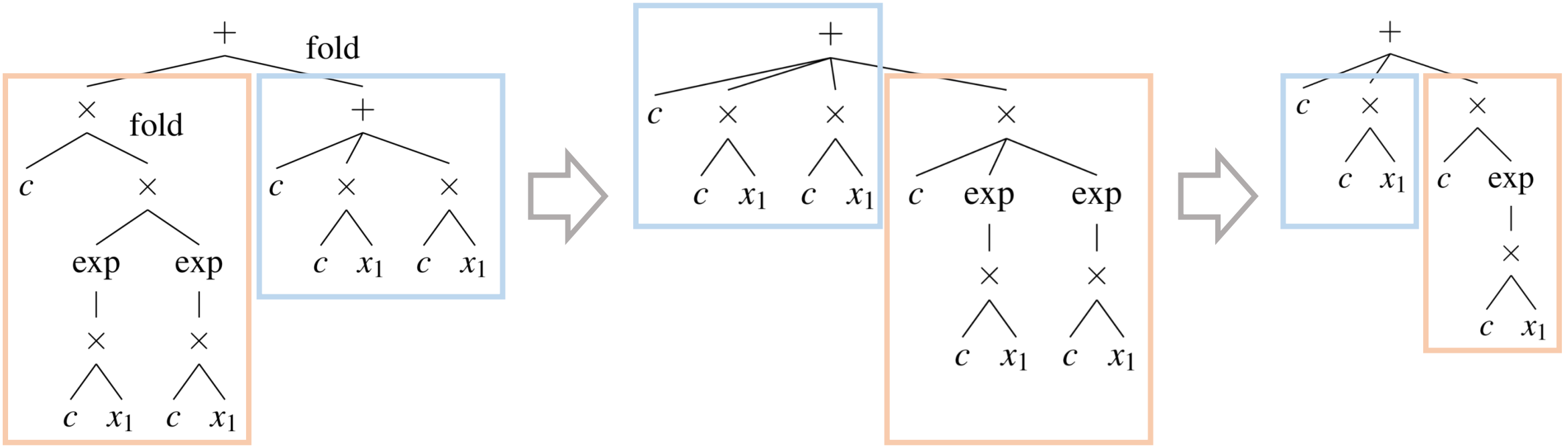
cf. Merkle, Ralph C. "A digital signature based on a conventional encryption function." *Conference on the theory and application of cryptographic techniques*. Springer, Berlin, Heidelberg, 1987.

Semantic Expression Hashing



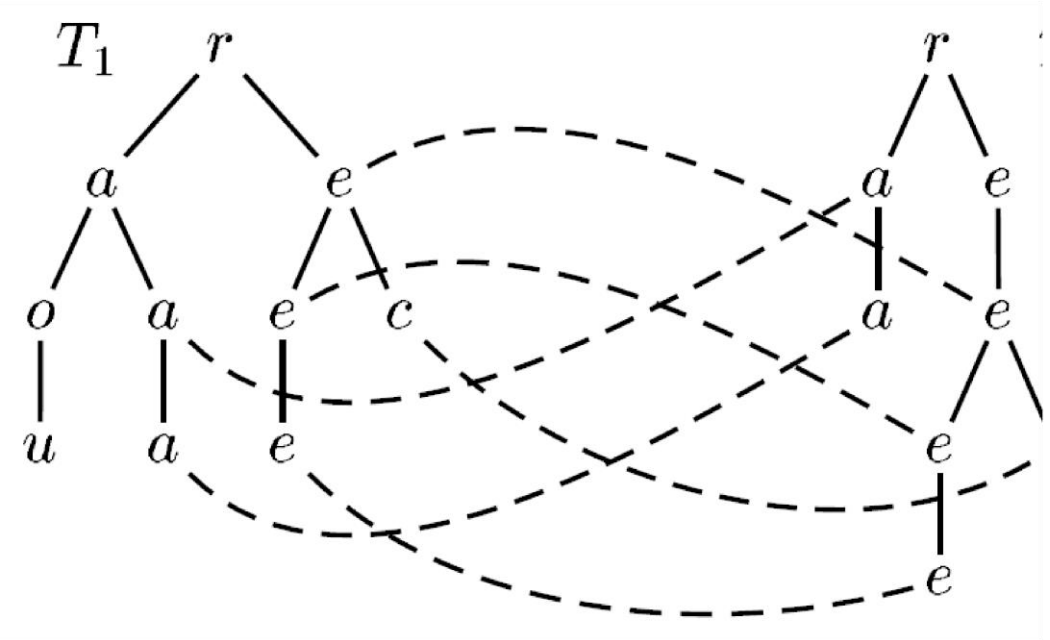
original folded reduced

Semantic Expression Hashing



original folded reduced

Semantic Expression Hashing vs. Bottom-Up Tree Distance



cf. Valiente, Gabriel. "An Efficient Bottom-Up Distance between Trees." *Proceedings of the 8th International Symposium of String Processing and Information Retrieval*. 2001

Iterating the Search Space

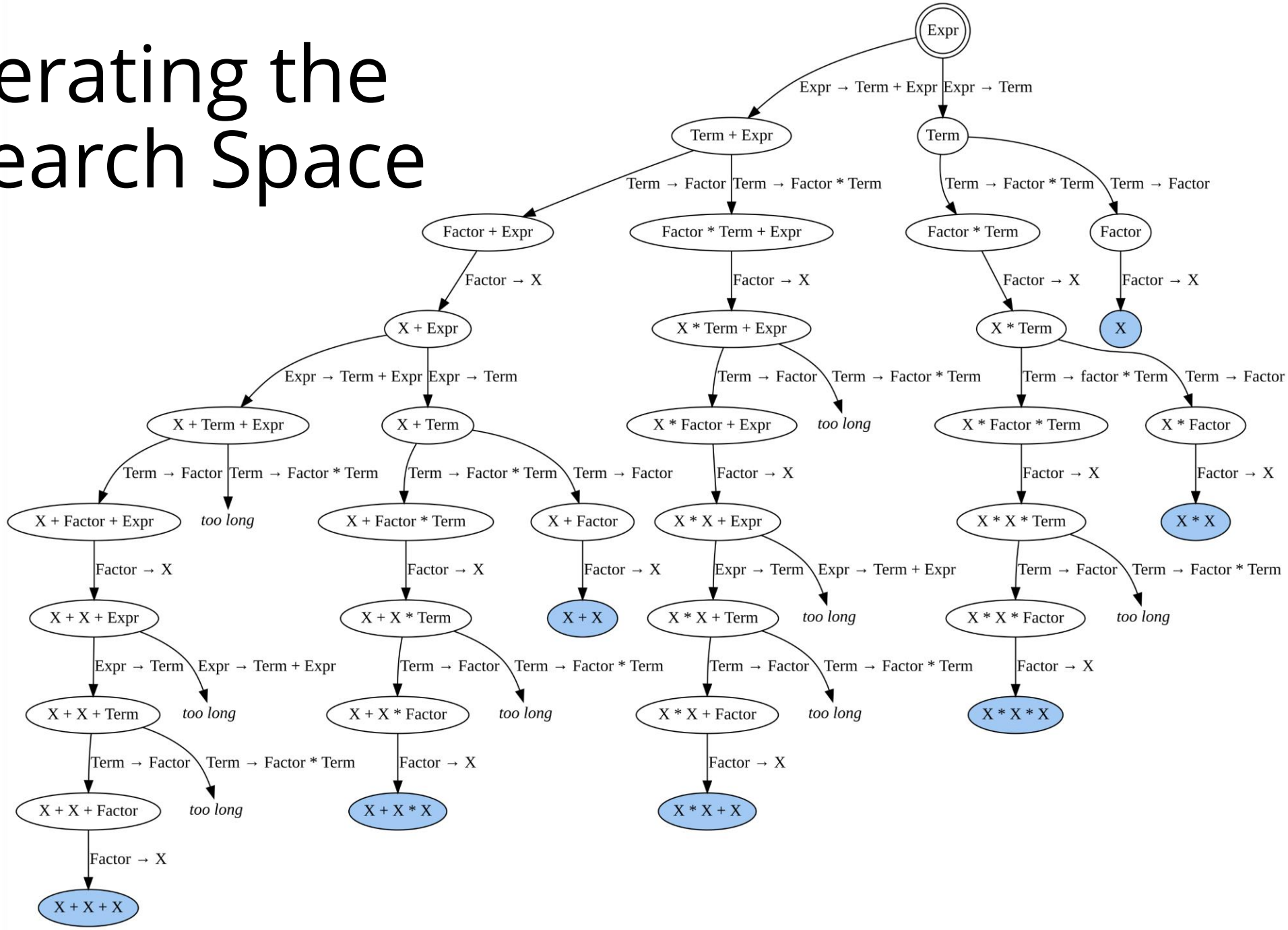
```
1    openPhrases = { StartSymbol }
2    seenHashes = { }
3    while openPhrases not empty
4         openPhrase = fetch from openPhrases
5         expandedSymbol = leftmost nonterminal symbol in openPhrase
6         foreach prod in expandedSymbol's productions
7              newPhrase = derive from openPhrase with prod
8              if hash(newPhrase) not in seenHashes
9                   if newPhrase is sentence
10                       add hash(newPhrase) to seenHashes
11                       optimize coefficients in newPhrase
12                       evaluate newPhrase
13                  else
14                       push newPhrase to openPhrases
```

derive phrases

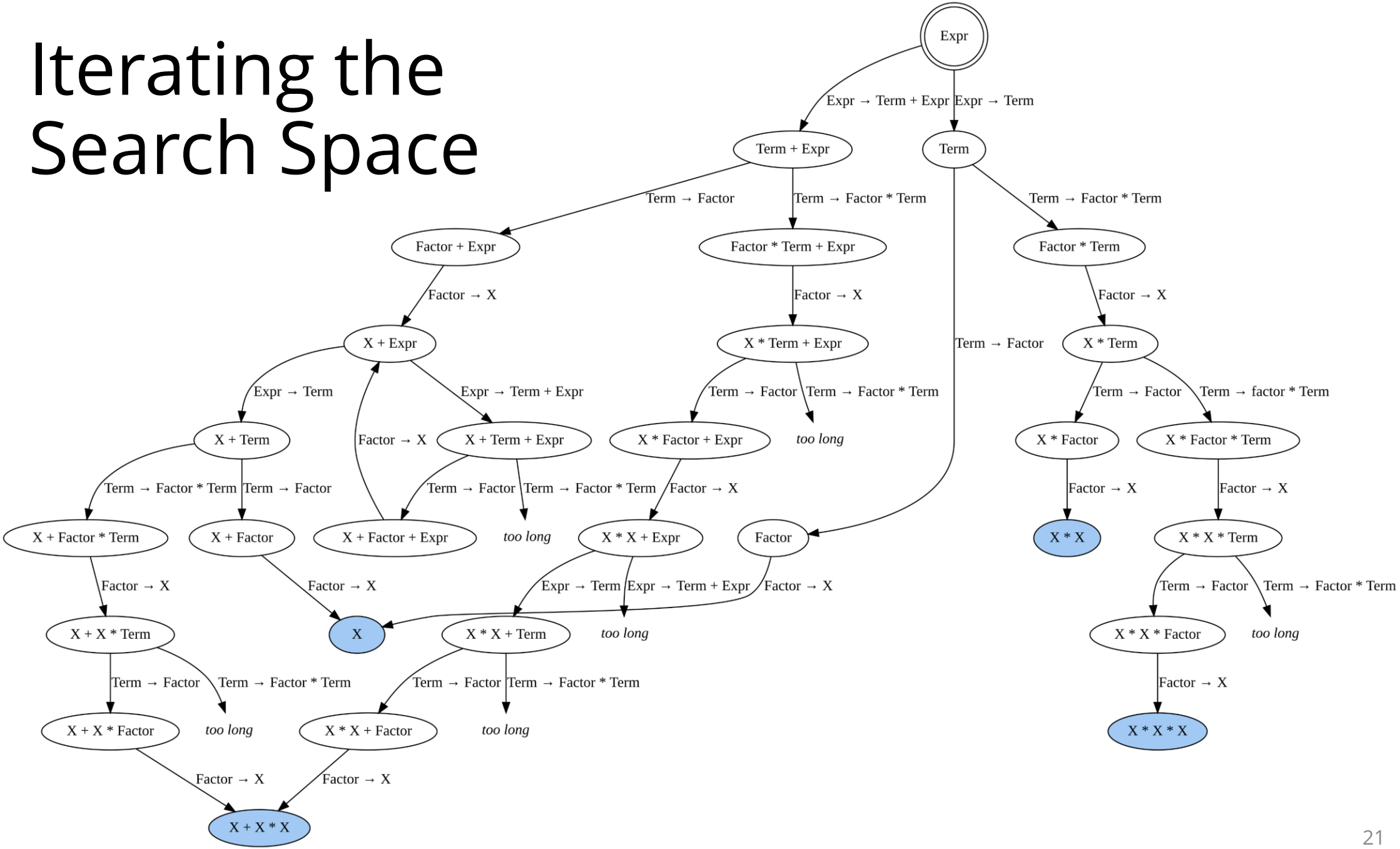
check for duplicates

evaluate previously unseen sentence

Iterating the Search Space



Iterating the Search Space



Guiding the Search

```
1   openPhrases = priority queue({ StartSymbol })
2   seenHashes = { }
3   while openPhrases not empty
4       openPhrase = fetch from openPhrases
5       expandedSymbol = leftmost nonterminal Symbol in openPhrase
6       foreach prod in expandedSymbol's productions
7           newPhrase = derive from openPhrase with prod
8           if hash(newPhrase) not in seenHashes
9               if newPhrase is sentence
10                  add hash(newPhrase) to seenHashes
11                  optimize coefficients in newPhrase
12                  evaluate newPhrase
13              else
14                  prio = priority(newPhrase)
15                  push newPhrase with prio to openPhrases
```

derive phrases

check for duplicates

evaluate previously unseen sentence

guide the search

Guiding the Search

```
1   openPhrases = priority queue({ StartSymbol })
2   seenHashes = { }
3   while openPhrases not empty
4       openPhrase = fetch from openPhrases
5       expandedSymbol = leftmost nonterminal Symbol in openPhrase
6       foreach prod in expandedSymbol's productions
7           newPhrase = derive from openPhrase with prod
8           if hash(newPhrase) not in seenHashes
9               if newPhrase is sentence
10                  add hash(newPhrase) to seenHashes
11                  optimize coefficients in newPhrase
12                  evaluate newPhrase
13              else
                    prio = priority(newPhrase)
                    push newPhrase with prio to openPhrases
```

derive phrases

check for duplicates

evaluate previously unseen sentence

guide the search

unfinished solution!

Steering the search

$$c_1 \log(c_2 x + c_3) \quad + \quad c_4 x x \quad + \quad Expr$$

Steering the search

$$\overbrace{c_1 \log(c_2 x + c_3)}^{finishedTerm_1} + \overbrace{c_4 x x}^{finishedTerm_2} + \underbrace{Expr}_{\text{Treat as coefficient}}$$

Steering the search

$$\overset{finishedTerm_1}{c_1 \log(c_2 x + c_3)} \quad + \quad \overset{finishedTerm_2}{c_4 x x} \quad + \quad \underbrace{Expr}_{\text{Treat as coefficient}}$$

$$finishedTerm_1 \quad + \quad finishedTerm_2 \quad + \quad \underbrace{c_5\, Term}_{\text{Can only improve quality}}$$

A Simple Heuristic

Length vs. Accuracy

$$priority(p) = -\underbrace{\frac{\text{len}(p)}{length_{\max}}}_{\in[0,1]} - w\,\text{NMSE}(p)$$

Experiment Results

Nguyen Problems

		Training NMSE	Test NMSE	Exact?
1	$x^3 + x^2 + x$	1.e-22	2.e-22	✓
2	$x^4 + x^3 + x^2 + x$	6.e-25	2.e-24	✓
3	$x^5 + x^4 + x^3 + x^2 + x$	4.e-22	7.e-22	✓
4	$x^6 + x^5 + x^4 + x^3 + x^2 + x$	1.e-13	7.e-12	
5	$\sin(x^2)\cos(x) - 1$	1.e-14	7.e-14	✓
6	$\sin(x) + \sin(x + x^2)$	9.e-13	1.e-07	
7	$\log(x + 1) + \log(x^2 + 1)$	9.e-14	1.e-12	✓
8	$\sqrt{x}$	5.e-18	1.e-17	✓
9	$\sin(x) + \sin(y^2)$	3.e-14	3.e-13	✓
10	$2\sin(x)\cos(y)$	4.e-18	6.e-18	✓
11	x^y	1.e-05	4.e-02	
12	$x^4 - x^3 + 0.5\,y^2 - y$	2.e-17	1.e-16	✓

Experiment Results

Keijzer Problems

		Training NMSE	Test NMSE	Exact?
1,2,3	$0.3 \, x \sin(2 \pi x)$	7.E-13	2.E-12	✓
4	$x^3 \exp(-x) \cos(x) \sin(x) (\sin(x)^2 \cos(x) - 1)$	2.E-04	3.E-04	
5	$(30 \, x \, z) / ((x - 10) \, y^2)$	2.E-06	2.E-06	
6	Sum(1 / i) From 1 to x	9.E-13	8.E-09	
7	ln(x)	1.E-22	1.E-22	✓
8	sqrt(x)	1.E-17	1.E-17	✓
9	arcsinh(x) i.e. $\ln(x + \sqrt{x^2 + 1})$	4.E-14	1.E-05	
10	x ^ y	8.E-05	9.E-03	
11	$xy + \sin((x - 1)(y - 1))$	2.E-03	2.E-01	
12	$x^4 - x^3 + y^2 / 2 - y$	1.E-19	3.E-19	✓
13	$6 \sin(x) \cos(y)$	3.E-24	1.E-21	✓
14	$8 / (2 + x^2 + y^2)$	1.E-19	7.E+00	
15	$x^3/5 + y^3/2 - y - x$	3.E-20	3.E-20	✓

Experiment Results

Vladislavleva Problems

		Training NMSE	Test NMSE	Exact?
1	$\exp(-(x_1 - 1)^2) / (1.2 + (x_2 - 2.5)^2)$	2.E-03	4.E+00	
2	$\exp(-x)\, x^3 \cos(x) \sin(x) (\cos(x)\sin(x)^2 - 1)$	2.E-04	3.E-01	
3	$\exp(-x_1)\, x_1^3 \cos(x_1) \sin(x_1) (\cos(x_1)\sin(x_1)^2 - 1)(x_2 - 5)$	2.E-02	1.E-01	
4	$10 / (5 + \text{Sum}(x_i - 3)^2)$ for i in 1..5	1.E-01	3.E-01	
5	$30 ((x_1 - 1) * (x_3 - 1)) / (x_2^2 (x_1 - 10))$	3.E-03	1.E-02	
6	$6 \sin(x_1) \cos(x_2)$	3.E-20	2.E-18	✓
7	$(x_1 - 3)(x_2 - 3) + 2 \sin((x_1 - 4)(x_2 - 4))$	7.E-02	8.E-02	
8	$((x_1 - 3)^4 + (x_2 - 3)^3 - (x_2 - 3)) / ((x_2 - 2)^4 + 10)$	3.E-03	6.E-01	

Experiment Results

Other Benchmark Problems

	Training NMSE	Test NMSE	Exact?
Breiman - I	1.E-01	1.E-01	
Friedman - I	1.E-01	1.E-01	
Friedman – II	4.E-02	4.E-02	
Poly-10 $y = X_1 X_2 + X_3 X_4 + X_5 X_6 + X_1 X_7 X_9 + X_3 X_6 X_{10}$	3.E-17	4.E-17	✓
Spatial co-evolution $F(x,y) = 1/(1 + x^{-4}) + 1/(1 + y^{-4})$	6.E-04	5.E-03	

Limitations

Complex terms, many variables

Keijzer Problems		Training NMSE	Test NMSE
4	$x^3 \exp(-x) \cos(x) \sin(x) (\sin(x)^2 \cos(x) - 1)$	2.E-04	3.E-04
5	$(30 x z) / ((x - 10) y^2)$	2.E-06	2.E-06
Vladislavleva Problems			
1	$\exp(- (x_1 - 1)^2) / (1.2 + (x_2 - 2.5)^2)$	2.E-03	4.E+00
2	$\exp(-x) x^3 \cos(x) \sin(x) (\cos(x)\sin(x)^2 - 1)$	2.E-04	3.E-01

Limitations

Noise

	Training NMSE	Test NMSE
Breiman - I	1.E-01	1.E-01
Friedman - I	1.E-01	1.E-01
Friedman - II	4.E-02	4.E-02

Limitations

Noise

	Training NMSE	Test NMSE
Friedman - II	4.E-02	4.E-02

Ground Truth:

$\sqrt{X0*X0 + (x1 * x2 - 1 / (x1 * x3))^2} + N(0, 1)$.

Found:

-65.271025 +
exp(x3 * x3 * -0.32) * -48.14 +
exp(x1 * x2 * -2.66) * 118.35 +
342.59 x1 * x2 * exp(x2 * x1 * -1.14) +
-18.65 x3 +
5.28 x3 * x3 * x3 +
-0.21 x1 * x1 +
-0.23 x2 * x2 +
5.06 x5 +
0.22 x1 +
0.26 x2 +
9.95 x4

Limitations

Constants

Nguyen Problems			
6	$\sin(x) + \sin(x + x^2)$	9.e-13	1.e-07

Keijzer Problems		Training NMSE	Test NMSE
9	arcsinh(x) i.e. $\ln(x + \sqrt{x^2 + 1})$	4.E-14	1.E-05
11	$xy + \sin((x - 1)(y - 1))$	2.E-03	2.E-01

Limitations

Constants

Nguyen Problems

6	$\sin(x) + \sin(x + x^2)$	9.e-13	1.e-07

Found Model:

```
-0.021 +
-0.005 x*x*x*x +
 1.01 sin(0.02 + 0.98 x + 0.99 x*x) +
-0.15 x*x*x +
 1.01 x
```

Discussion

- Lack of heuristics for complex terms and many variables
- No mechanisms against overfitting
- No re-evaluation of solutions
- Nearly no bloat
- Less possibilities for parallelism
- Determinism

Discussion

- **Lack of heuristics for complex terms and many variables**
- **No mechanisms against overfitting**
- No re-evaluation of solutions
- Nearly no bloat
- Less possibilities for parallelism
- Determinism

Thank you!

HEAL

HEURISTIC AND EVOLUTIONARY
ALGORITHMS LABORATORY