# (X)AI for Computer Algebra Optimisation

**Prof. Matthew England**
Coventry University

**InformedAI Summer School**
16 – 19 June 2025

Speaker's work supported by EPSRC grant EP/T015748/1.

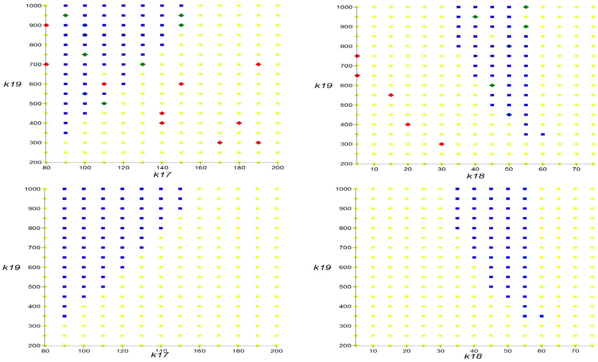
# Outline

1. Computer Algebra
2. ML for Integration Algorithm Selection
3. ML for CAD Variable Ordering

# Computer Algebra (Slide 1/10)

Computer Algebra Systems (CASs) are software for manipulating mathematical expressions and objects. A CAS gives *exact answers*. E.g., a CAS can compute with the symbols $\frac{1}{3}$, $\pi$ and $\sqrt{2}$ rather than their approximations 0.33333, 3.14159 and 1.41421.

Computer algebra is more expensive than numerics: **why use it?**




https://doi.org/10.1016/j.jsc.2019.07.008

# Computer Algebra



Computer Algebra Systems (CASs) are software for manipulating mathematical expressions and objects. A CAS gives *exact answers*. E.g., a CAS can compute with the symbols $\frac{1}{3}$, $\pi$ and $\sqrt{2}$ rather than their approximations 0.33333, 3.14159 and 1.41421.

Computer algebra is more expensive than numerics: **why use it?**

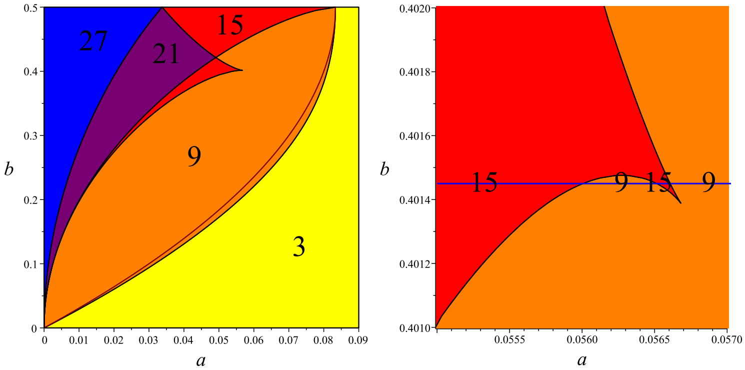
https://doi.org/10.1007/978-94-010-0203-5_18

# Computer Algebra

(Slide 1/10)

Computer Algebra Systems (CASs) are software for manipulating mathematical expressions and objects. A CAS gives *exact answers*. E.g., a CAS can compute with the symbols $\frac{1}{3}$, $\pi$ and $\sqrt{2}$ rather than their approximations 0.33333, 3.14159 and 1.41421.

Computer algebra is more expensive than numerics: **why use it?**




https://doi.org/10.1142/S0218127421502023

# Can LLMs do Computer Algebra? (Slide 2/10)

**Copilot**
Please calculate a Groebner Ba...

December 10, 2024 at 12:43 PM

Please calculate a Groebner Basis for the set of polynomials {x^2 - 2*x*z + 5, y*z^3 + x*y^2, -8*z^3 + 3*y^2}

I calculated the Groebner Basis for the given set of polynomials ({x^2 - 2x z + 5, y z^3 + x y^2, -8 z^3 + 3 y^2}). Here is the result:

```
$$ \begin{align*}
```

```
1. & \quad x^2 - 2x z + 5 \
2. & \quad 640x z^3 + 9z^8 - 96z^7 + 120z^5 \
3. & \quad 3y^2 - 8z^3 \
4. & \quad 80y z^3 - 3z^8 + 32z^7 - 40z^5 \
5. & \quad 9z^9 - 96z^8 + 240z^6 + 1600z^3 \end{align*} $$
```

If you have any further questions or need additional assistance, feel free to ask!

Edit in Pages   Code   ...   1 of 30 responses   AI-generated content may be incorrect

# Can LLMs do Computer Algebra? (Slide 2/10)

How did you calculate this?

**Copilot**
Please calculate a Groebner Ba...

I used the **SymPy** library in Python to calculate the Groebner Basis. SymPy is a powerful tool for symbolic mathematics, and it includes functions for working with polynomials and computing Groebner Bases.

Here's a step-by-step outline of the process:

1. **Define the Variables**: I defined the variables (x), (y), and (z) using SymPy's `symbols` function.
2. **Define the Polynomials**: I created a list of the given polynomials.
3. **Calculate the Groebner Basis**: I used the `groebner` function from SymPy, specifying the polynomials and the variable order (lexicographic by default).
4. **Output the Result**: I printed the resulting Groebner Basis.

Here's the Python code I used:

```
from sympy import symbols, groebner
# Define the variables
x, y, z = symbols('x y z')
# Define the polynomials
polynomials = [x**2 - 2*x*z + 5, y*z**3 + x*y**2, -8*z**3 + 3*y**2]
# Calculate the Groebner basis
G = groebner(polynomials, x, y, z)
# Print the Groebner basis
print("The Groebner basis for the given set of polynomials is:")
for g in G:
    print(g)
```

# Can transformers do Computer Algebra? (Slide 3/10)

First attempt in 2020: for symbolic integration and ODE solving.

G. Lample and D. Charton
*Deep Learning for Symbolic Mathematics.*
Proc. ICLR 2020.
https://doi.org/10.48550/arXiv.1912.01412

Recent attempt to directly produce Gröbner Bases.

H. Kera, Y. Ishihara, Y. Kambe, T. Vaccon and K. Yokoyama
*Learning to compute Gröbner bases.*
Proc. NIPS 2024.
https://openreview.net/forum?id=ZRz7XlxBzQ

Both struggle with out of distribution data. ML can only offer probabilistic guidance, but CASs prize exact results. Possibilities:

- Algorithms to validate correctness of ML results?
- Algorithms to fix "*close*" results produced by ML?
- ML guidance in selecting and running CAS algorithms?

# Challenging / Interesting Machine Learning



ML guidance in selecting and running CAS algorithms:

- Problems do not always neatly fit as regression / classification.
- No a priori limit on the input space.
- Supervised learning hard: because labelling dataset needs lots of expensive computer algebra.
- Unsupervised learning is hard: because it is unclear if a particular outcome is good or bad without seeing the rest.
- What constitutes a meaningful and representative data set?
- How best to embed the data for a ML algorithm?

A good area for research!

# Outline

1. Computer Algebra
2. ML for Integration Algorithm Selection
3. ML for CAD Variable Ordering

# Acknowledgements

Collaboration with Rashid Barket and Jürgen Gerhard.
Sponsored by Maplesoft.

# Problem: Integration algorithm selection (Slide 5/10)

Symbolic indefinite integration accessed via the command `int`: a meta-algorithm which does pre-processing and table lookup before accessing a particular CAS algorithms for integration, of which there are currently 12 to choose from.

**Problem:** Which to use for a given problem instance?

> $f := 7x^3 + 3x^2 + 5x :$
> $\mathrm{int}(f, x)$

$$\frac{7}{4}x^4 + x^3 + \frac{5}{2}x^2$$

> $\mathrm{int}(\sin(x), x)$

$$-\cos(x)$$

> $\mathrm{int}\left(\frac{x}{x^3-1}, x\right)$

$$-\frac{\ln\left(x^2+x+1\right)}{6} + \frac{\sqrt{3}\arctan\left(\frac{(2x+1)\sqrt{3}}{3}\right)}{3} + \frac{\ln(x-1)}{3}$$

Currently, `int` tries all in a fixed order until one succeeds (in some cases checking guard code).

# Problem: Integration algorithm selection (Slide 5/10)

Symbolic indefinite integration accessed via the command `int`: a meta-algorithm which does pre-processing and table lookup before accessing a particular CAS algorithms for integration, of which there are currently 12 to choose from.

**Problem:** Which to use for a given problem instance?

$$> \mathrm{int}\left(\exp\left(-x^2\right), x\right)$$

$$\frac{\sqrt{\pi}\,\mathrm{erf}(x)}{2}$$

$$> \mathrm{int}\left(\frac{1}{\mathrm{sqrt}\left(2t^4 - 3t^2 - 2\right)}, t = 2..3\right)$$

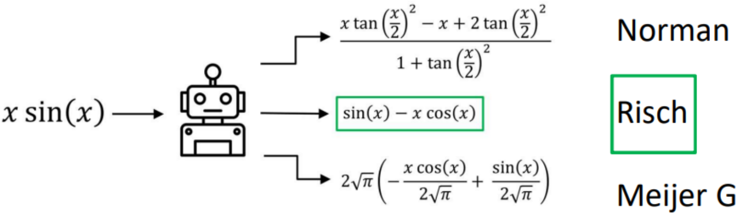
$$\frac{\sqrt{5}\,\mathrm{EllipticF}\left(\frac{\sqrt{7}}{3}, \frac{\sqrt{5}}{5}\right)}{5} - \frac{\sqrt{5}\,\mathrm{EllipticF}\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{5}}{5}\right)}{5}$$

Currently, `int` tries all in a fixed order until one succeeds (in some cases checking guard code).

# Methodology

(Slide 6/10)

- Metric for success is minimal answer size.
- Identify biases in existing 2020 data; build new data generators by reverse engineering computer algebra theory.
- Explore different ML paradigms: classification; classification + regression; classification + ranking.
- Explore different architectures: sequential vs tree embedding.



# Methodology



- Metric for success is minimal answer size.
- Identify biases in existing 2020 data; build new data generators by reverse engineering computer algebra theory.
- Explore different ML paradigms: classification; classification + regression; classification + ranking.
- Explore different architectures: sequential vs tree embedding.

```
> f := 1 - x*cos(cos(x))*sin(x) + sin(cos(x))
```

$$f := 1 - x\cos(\cos(x))\sin(x) + \sin(\cos(x))$$

```
> # Maple's answer
  int(f, x)
```

$$\frac{x + x\tan\left(\frac{x}{2}\right)^2 + x\tan\left(\frac{1-\tan\left(\frac{x}{2}\right)^2}{2\left(1+\tan\left(\frac{x}{2}\right)^2\right)}\right)^2 + x\tan\left(\frac{x}{2}\right)^2\tan\left(\frac{1-\tan\left(\frac{x}{2}\right)^2}{2\left(1+\tan\left(\frac{x}{2}\right)^2\right)}\right)^2 + 2x\tan\left(\frac{1-\tan\left(\frac{x}{2}\right)^2}{2\left(1+\tan\left(\frac{x}{2}\right)^2\right)}\right) + 2x\tan\left(\frac{x}{2}\right)^2\tan\left(\frac{1-\tan\left(\frac{x}{2}\right)^2}{2\left(1+\tan\left(\frac{x}{2}\right)^2\right)}\right)}{\left(1+\tan\left(\frac{1-\tan\left(\frac{x}{2}\right)^2}{2\left(1+\tan\left(\frac{x}{2}\right)^2\right)}\right)^2\right)\left(1+\tan\left(\frac{x}{2}\right)^2\right)}$$

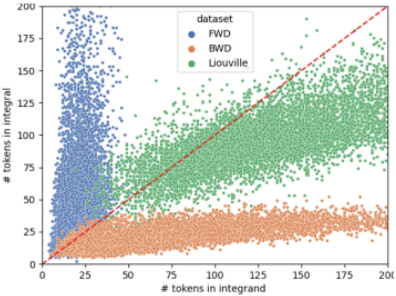
```
> # ML-suggested answer
  int(f, x, method=risch)
```

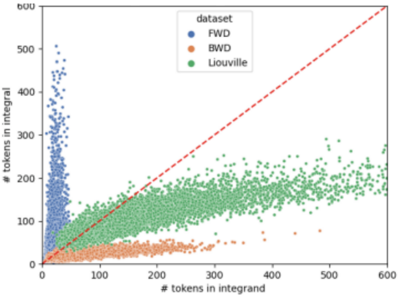
$$x + \sin(\cos(x))\,x$$

# Methodology

(Slide 6/10)

- Metric for success is minimal answer size.
- Identify biases in existing 2020 data; build new data generators by reverse engineering computer algebra theory.
- Explore different ML paradigms: classification; classification + regression; classification + ranking.
- Explore different architectures: sequential vs tree embedding.



# Methodology



- Metric for success is minimal answer size.
- Identify biases in existing 2020 data; build new data generators by reverse engineering computer algebra theory.
- Explore different ML paradigms: classification; classification + regression; classification + ranking.
- Explore different architectures: sequential vs tree embedding.

$$3x^2 + \cos(2x) - 1$$

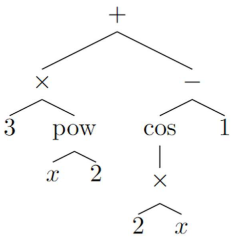
Text Sequence

['add',
'-1',
'add',
'mul',
'3',
'pow',
'x',
'2',
'cos',
'mul',
'2',
'x']

Feature Vector

| | Vars | Terms | Has Trig | ... |
|---|---|---|---|---|
| $3x^2 + \cos(2x) - 1$ | 1 | 3 | True | ... |

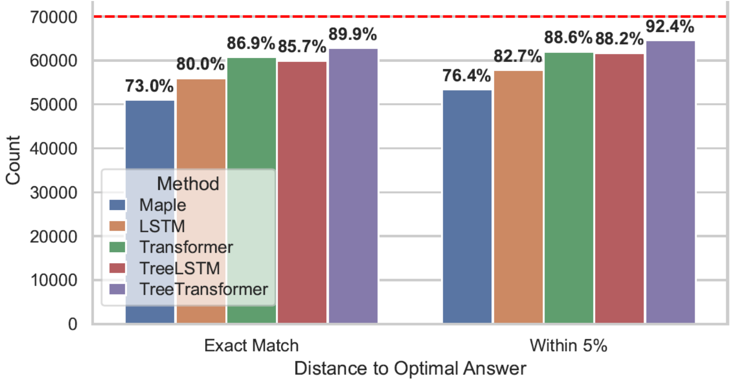
Graph



# Results



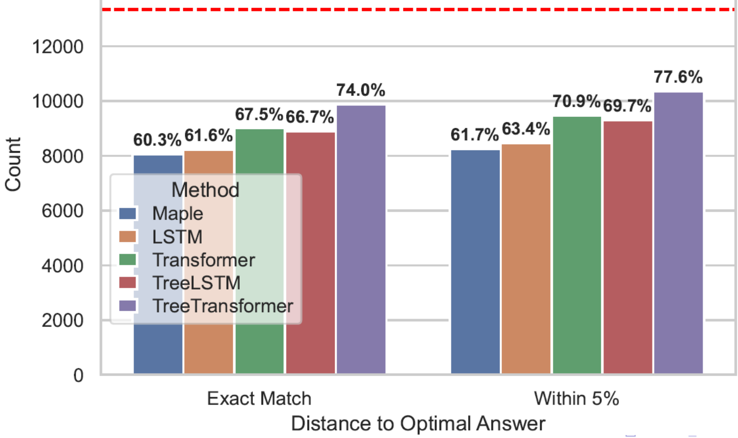
- ML beat state-of-the-art. Validated on both hold-out test data from generators, and independent test suite of Maplesoft.
- Improved data generators are essential for generalisation.
- Tree architectures do much better for this than sequential.



# Results



- ML beat state-of-the-art. Validated on both hold-out test data from generators, and independent test suite of Maplesoft.
- Improved data generators are essential for generalisation.
- Tree architectures do much better for this than sequential.



# ML for Integration Selection References


R. Barket, M. England, and J. Gerhard.
*Generating Elementary Integrable Expressions*
Proc. CASC '23, LNCS 14139, pp. 21 – 38, 2023.
https://doi.org/10.1007/978-3-031-41724-5_2

R. Barket, M. England, and J. Gerhard.
*The Liouville Generator for Producing Integrable Expressions*
Proc. CASC '24, LNCS 14938, pp. 47 – 62, 2024.
https://doi.org/10.1007/978-3-031-69070-9_4

R. Barket, M. England, and J. Gerhard.
*Symbolic Integration Algorithm Selection with Machine Learning: LSTMs vs Tree LSTMs.*
Proc. ICMS '24, LNCS 14749, pp. 167–175, 2024.
https://doi.org/10.1007/978-3-031-64529-7_18

R. Barket, U. Shafiq, M. England, and J. Gerhard.
*Transformers to Predict the Applicability of Symbolic Integration Routines*
Proc. MATH-AI at NeurIPS '24, 2024.
https://openreview.net/forum?id=b2Ni828As7

R. Barket, M. England, and J. Gerhard.
*Tree-Based Deep Learning for Ranking Symbolic Integration Algorithms.*
Submitted, 2025.

# Outline

1 Computer Algebra

2 ML for Integration Algorithm Selection

3 ML for CAD Variable Ordering

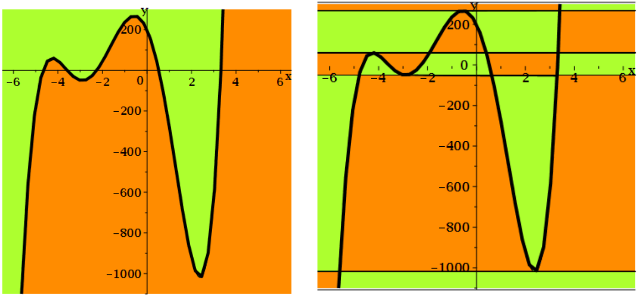
# Acknowledgements

Collaborations over the years with James Bridge, Kelly Cohen, James Davenport, **Tereso del Río**, **Dorian Florescu**, **Zongyan Huang**, Larry Paulson, and **Lynn Pickering**.

# Problem: CAD Variable Ordering



- Cylindrical Algebraic Decomposition (CAD) is a key computer algebra algorithm used to study systems of non-linear polynomial constraints. It decomposes space according to the truth of constraints: proceedings one variable at a time.
- The variable ordering greatly affects decomposition size and computation time, and is often a free choice.



# Experiments



Explored ML to take the choice:

- Using feature engineering approach (semi-automated) and classical ML.
- Using real world data set (SMT-LIB); augmented by variable permutations.
- Explored different ML paradigms for the problem.
- Explored XAI approaches: SHAP post analysis and constrained NNs.

## Key Findings

(Slide 10/10)

- ML can certainly take this choice better than human-made heuristics.
- Care needed with dataset choice; better to augment a high quality dataset.
- Still big challenges generalising over datasets.
- XAI output can suggest human-level heuristics:
  - can be expressed in natural language in a similar amount of text as a heuristic designed by a human;
  - can be implemented without any ML architecture.
- These heuristics perform almost as well as full ML models on original data, and demonstrate much better generalisation to new datasets.

# ML for CAD References


Z. Huang, M. England, D. Wilson, J. Davenport, L. Paulson and J. Bridge.
*Applying machine learning to the problem of choosing a heuristic to select the variable ordering for cylindrical algebraic decomposition.*
Intelligent Computer Mathematics (LNCS 8543), pp. 92-107. Springer Berlin Heidelberg, 2014.
http://dx.doi.org/10.1007/978-3-319-08434-3_8

D. Florescu and M. England.
*A Machine Learning Based Software Pipeline to Pick the Variable Ordering for Algorithms with Polynomial Inputs.*
Mathematical Software (LNCS 12097), pp. 302—311. Springer International Publishing, 2020.
https://doi.org/10.1007/978-3-030-52200-1_30

T. del Río and M. England.
*Lessons on Datasets and Paradigms in Machine Learning for Symbolic Computation: A Case Study on CAD..*
Mathematics in Computer Science, 18, article 17, Springer, 2024.
Preprint: https://doi.org/10.48550/arXiv.2401.13343

L. Pickering, T. del Rıio Almajano, M. England and K. Cohen.
*Explainable AI Insights for Symbolic Computation: A case study on selecting the variable ordering for cylindrical algebraic decomposition.*
Journal of Symbolic Computation, **123**, Article Number 102276, 2024.

D. Florescu and M. England.
*Constrained Neural Networks for Interpretable Heuristic Creation to Optimise Computer Algebra Systems.*
Proc. ICMS 2024 (LNCS 14749), pp. 186-195,
https://doi.org/10.1007/978-3-031-64529-7_19

# Thanks for Listening

Matthew.England@coventry.ac.uk
https://matthewengland.coventry.domains