The BACON system for equation discovery from scientific data: Reconciling classical artificial intelligence with modern machine learning approaches

Jonah Miller, Soumya Banerjee

University of Cambridge, UK

sb2333@cam.ac.uk

Abstract

BACON is a heuristic-based computational scientific discovery system, which aims to find invariants in multivariable systems. We rebuilt *BACON* in a modern computing language, and we improve the noise-resilience of *BACON*. We demonstrate how such classical AI systems can be understandable, yet powerful. We applied our framework to a number of exemplar problems in physics and mathematics. Our *BACON* also outperformed *PySR* - a modern method utilising symbolic regression on a neural network - conclusively in specific environments on small datasets. We suggest that there is potential in these forgotten approaches that modern deep learning systems can learn from. Integrative approaches that combine heuristic approaches like *BACON* with modern deep learning can be very helpful. We suggest integrating modern deep learning approaches and large-language models with heuristic-based classical AI approaches as a way to analyse large scientific datasets.

Introduction

Planet	Distance (D)	Period (P)	$\frac{D}{P}$	$\frac{D^2}{P}$	$\frac{D^2}{P^2}$	$\frac{D^3}{P^2}$
A	1.0	1.0	1.0	1.0	1.0	1.0
B	4.0	8.0	0.5	2.0	0.25	1.0
C	9.0	27.0	0.333	3.0	0.111	1.0

Table: An example of the *BACON.1* algorithm discovering Kepler's 3rd law from a noiseless planetary system. The program is acting on 3 different synthetic planets which obey the same laws of motion (data from Langley et al. [1987]).

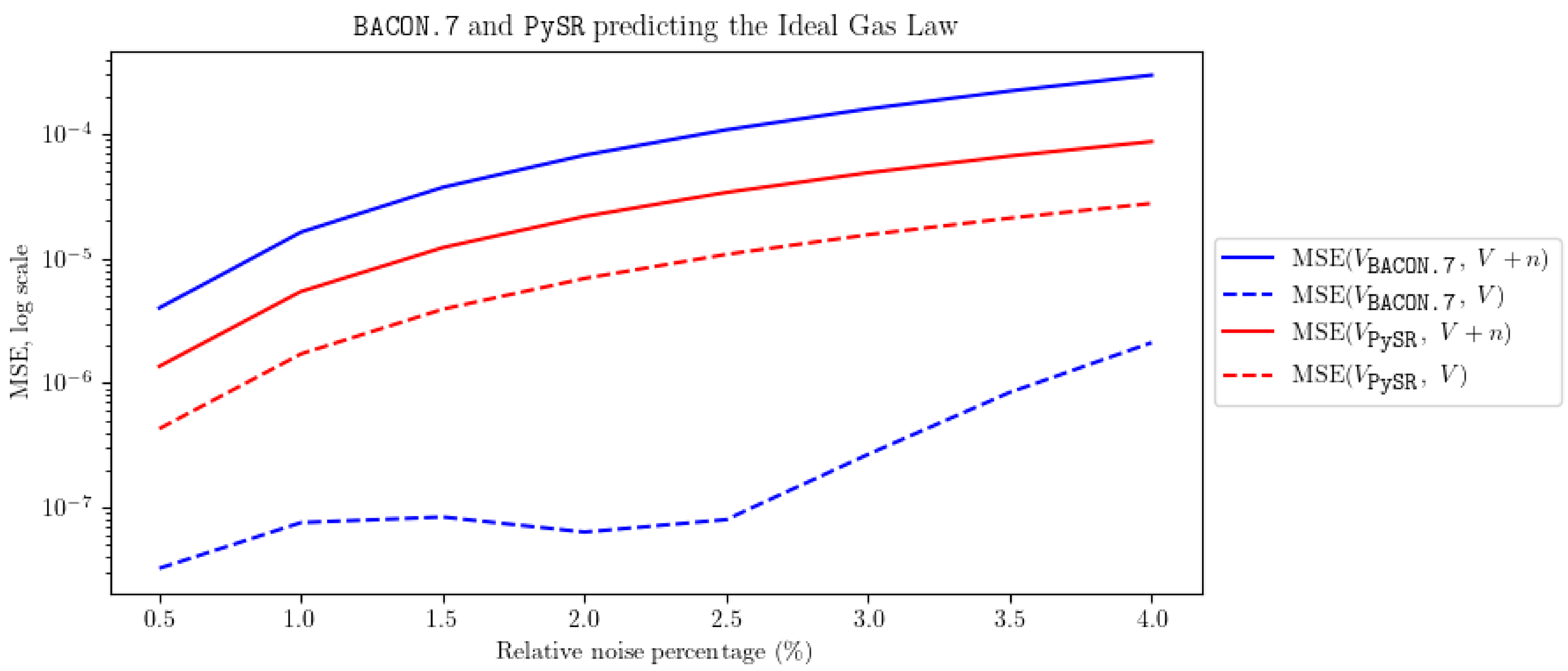

Figure: For $V = \frac{M(T+273)}{P}$ and denoting the volume with noise added as $V+n$, the graph demonstrates the MSE between predicted models from *BACON.7* and *PySR* with $V+n$ and V. The MSE with $V+n$ is how well the model predicts the data it is trained on (solid lines). The MSE with V describes how it predicts the true data (dashed lines). In each case the model is better at predicting the true data with *BACON.7* creating a model an order of magnitude more accurate than *PySR*. We hypothesise this is due to *BACON.7*'s averaging through the Space of Data incidentally cancelling Gaussian noise, whereas *PySR* is trained to overfit on the data as it has no prior knowledge of there being noise. The latter explains as well why *PySR* displays a better model when compared with $V+n$ than *BACON*.

Methods

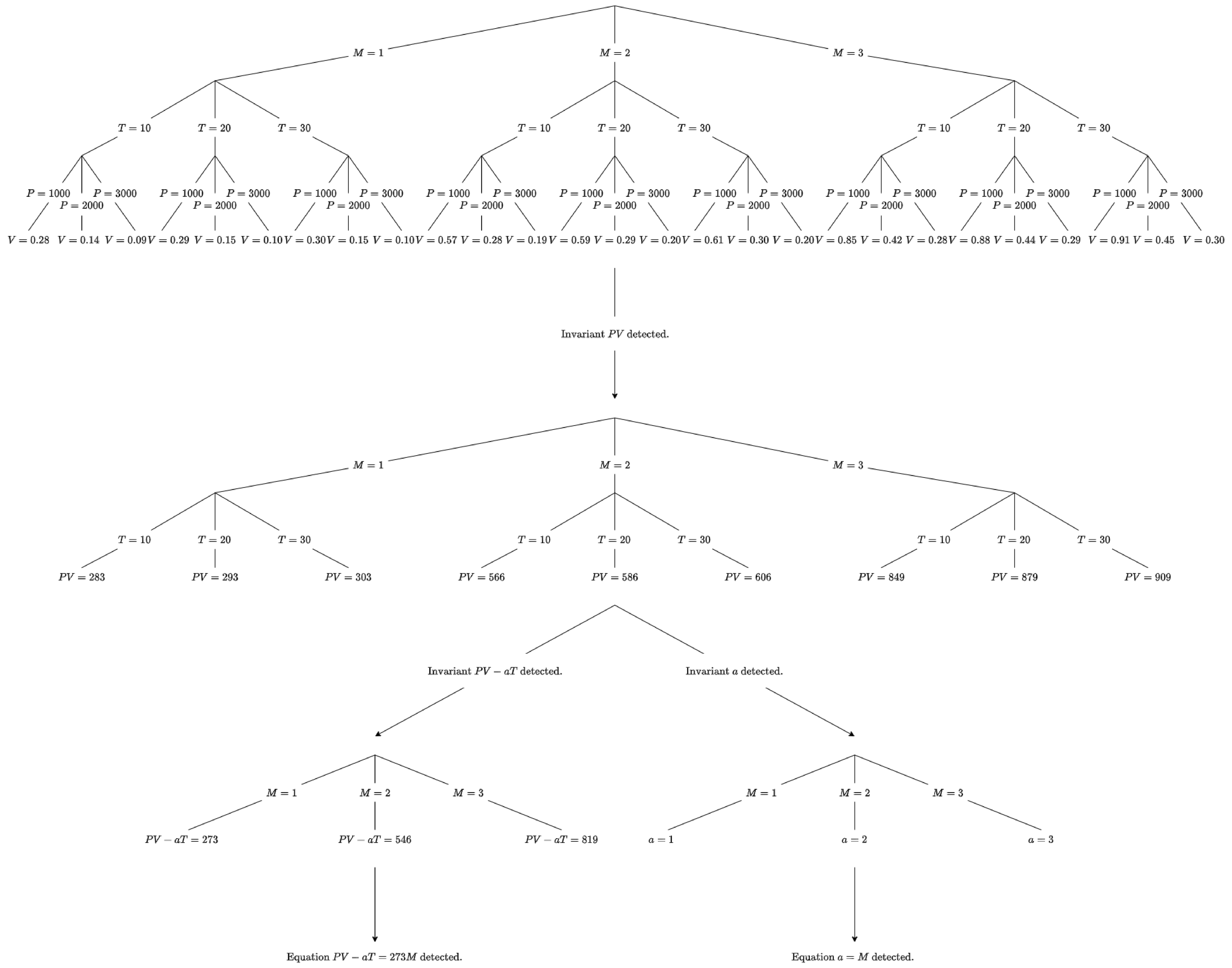

Figure: Consider the Ideal Gas Law $V = \frac{M(T+273)}{P}$. It is placed in a noiseless environment where V is a dependent variable, P, M and T are independent variables each consigned to 3 specific values - ultimately giving 27 values for V. At the lowest level it runs the simple *BACON.1* heuristic between P and V, here the sets share the same value in M and T. For example, the first set has $T = 10,\ M = 1$ and are the 3 points defined by $P = 1000,\ V = 0.28$, $P = 2000,\ V = 0.14$ and $P = 3000,\ V = 0.09$. All 9 sets in this layer determine invariant PV. *BACON.3* detects the agreement, then forms a new tree with PV as the dependent variable. It then proceeds to run a new heuristic check between PV and T. Here all 3 sets determine $PV - aT$ is invariant for new dependent variable a. As both a and $PV - aT$ can be a function of M, *BACON.3* forms two new trees at this level. The last *BACON.1* heuristic check finds $a = M,\ PV - aT = 273M$. These are collated to form the Ideal Gas Law.

Ohm's law is typically seen as $I = \frac{V}{R}$, for V voltage, I current and R internal resistance. An expanded form when considering the law applied to a bar of temperature T, diameter D and length L is:

$$I = \frac{TD^2}{2(L+3)} \tag{1}$$

Results

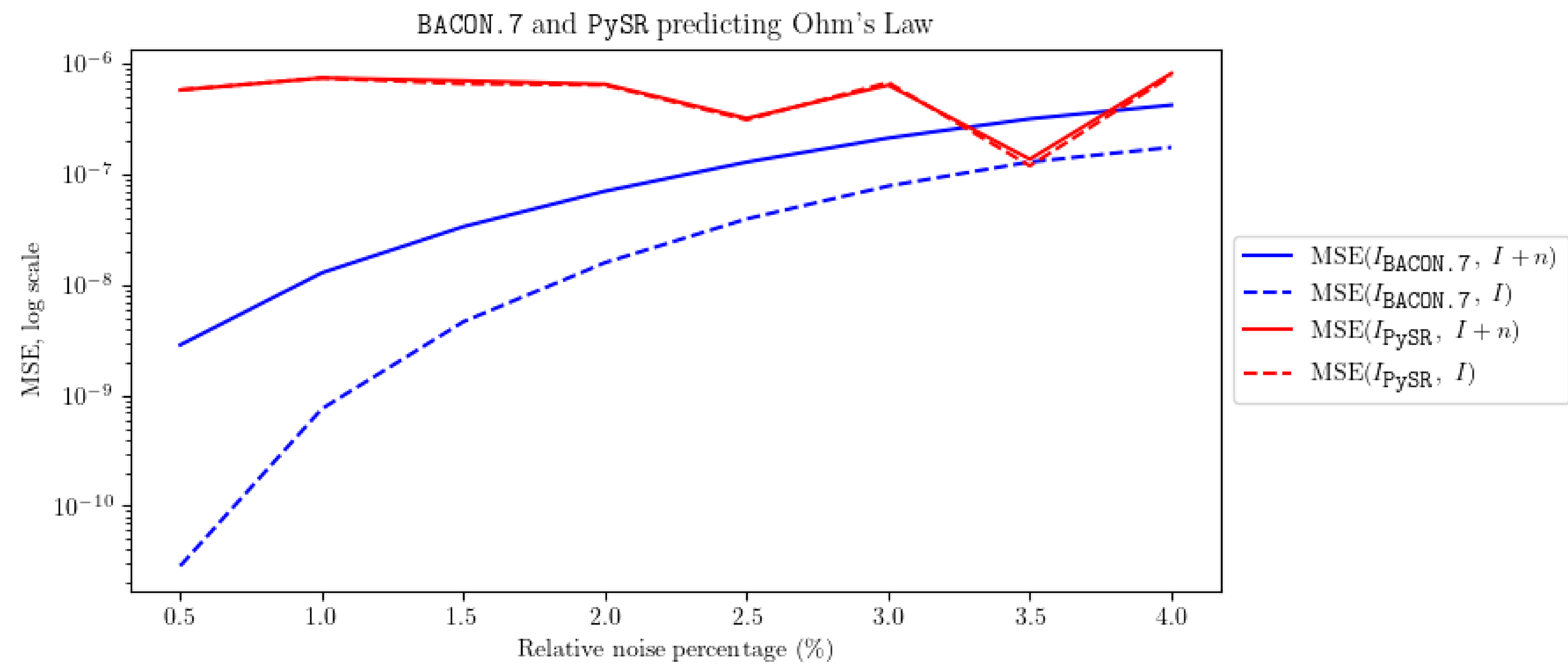

Figure: The results for Ohm's Law when using the same methodology as the Ideal Gas Law. *BACON.7* again is better at predicting the I than $I+n$. Here *PySR* is not able to deduce the correct form of Ohm's Law so the results are taken from the best prediction it gives. In almost each case *PySR*'s predictions are worse than *BACON.7*'s against both the noisy and true data . The likely reason is that the correct form of the equation cannot be inferred by *PySR* from only 27 datapoints as *PySR* was built to function efficiently on larger, more complicated datasets. *BACON.7* works best at this scale, where its strips out noise through averaging.

Conclusions

For smaller datasets, BACON consistently thrives whilst PySR struggles. BACON is made for this environment, whereas *PySR* is made to overfit on complicated, large datasets whilst applying their biases towards simplicity. When approaching this threshold, it is *PySR* that thrives whilst *BACON* suffers Here is an - albeit niche - situation where classical methods outperform modern techniques. It amplifies the need to reproduce, study and understand these seemingly anachronistic mechanisms and see what lessons can be taken going forward.

We applied this framework to a number of exemplar problems in physics and mathematics. Our results suggest that *BACON* is good at reducing noise and inferring the correct equation in smaller datasets, whereas *PySR* is significantly more successful on larger, noisier, datasets.

The broad goal of this research project was to combine modern approaches to AI with the classical. Both have strengths which, when efficiently combined, could lead to refined systems, able to analyse large datasets effectively. We suggest that in the future, combining large-language models with classical AI approaches such as those presented here may help solve more complex scientific and mathematical problems Miller and Banerjee [2024].

References

Patrick Langley, Herbert Simon, Gary Bradshaw, and Jan Zytkow. Scientific Discovery: Computational Explorations of the Creative Process. 1987.

Jonah Miller and Soumya Banerjee. The bacon system for equation discovery from scientific data: Transforming classical artificial intelligence with modern machine learning approaches. In *AAAI 2024 Fall Symposium Series, Integrated Approaches to Computational Scientific Discovery*, 10 2024. doi: 10.31219/OSF.IO/Z8KQV. URL `https://osf.io/z8kqv`.

For more details, please visit: `https://osf.io/preprints/osf/z8kqv_v1`

Acknowledgements: We thank the Accelerate Programme for Scientific Discovery for funding this work.