Optimization of SISSO++

YY

Past Experiences

Molecular Dynamics (Polarizable Force Field, First Principles MD, Machine Learning Force Field, Nuclei Quantum Effect)

Density Functional Theory and beyond (Planewave, NAO)

Real-Time Time-dependent Density Functional Theory (Planewave, NAO)

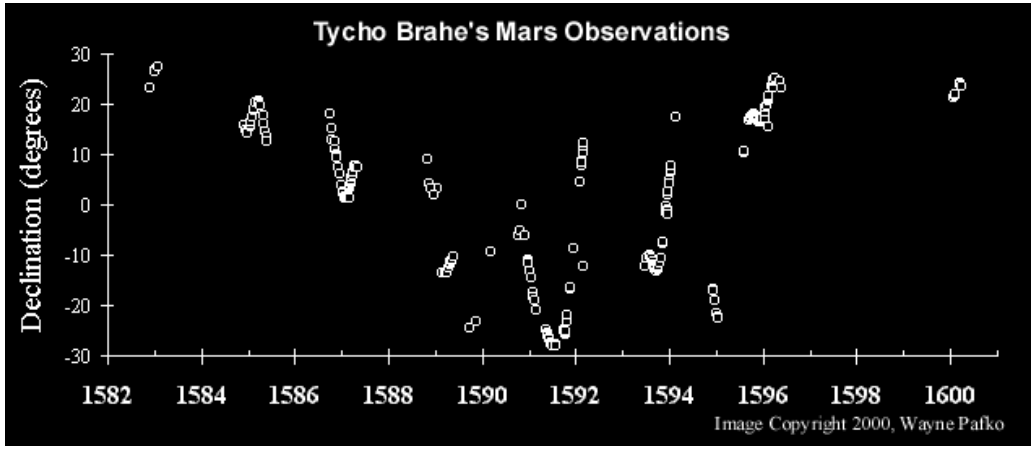
Numerical Methods & Code optimization (FHI-aims: symmetry, tetrahedra method, Bader analysis, solvent model, long range-separated hybrid functionals (wB97x series), ELSI eigensolvers, GW&BSE, periodic GW on GPU)

Two Scientific Paradigms: Newtonian and Keplerian

Newtonian: Discover the **fundamental principles** that govern the systems we are interested in and perform first-principles modelling with them.

- In materials science/chemistry, **quantum mechanics (Dirac equation) is the first principle**. However, as pointed out by Dirac, the mathematical problem that describes the laws of quantum mechanics is exceptionally **complicated**.
- **Dilemma**: fundamental but not practical. We need to resort to ad hoc and nonsystematic approximations. We pay the price of losing reliability and transferability.

Keplerian: extract scientific discoveries from data analysis. **data-driven** approach.



~20 years observation
Tycho was not willing to share, Kepler stole it?

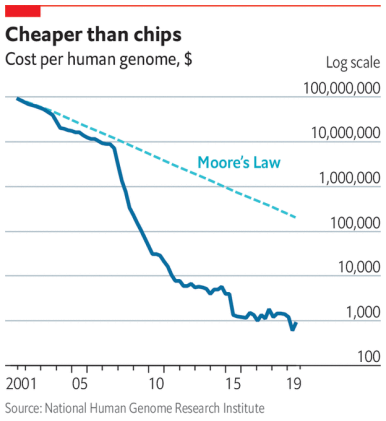
- Weinan, E. "The dawning of a new era in applied mathematics." Notices of the American Mathematical Society68.4 (2021): 565-571.
- http://www.pafko.com/tycho/observe.html

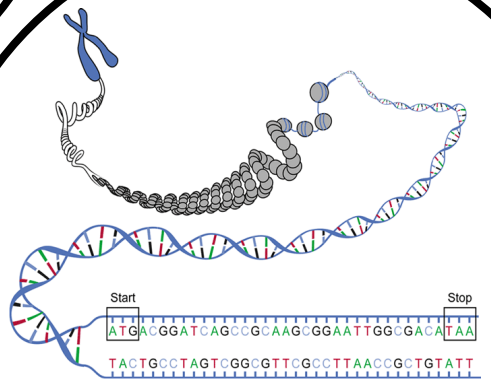
Science of this/next Generation

Robotic Lab

HPC

Cost of data is continuing decreasing.
Data need to be understood
(What to do next?)



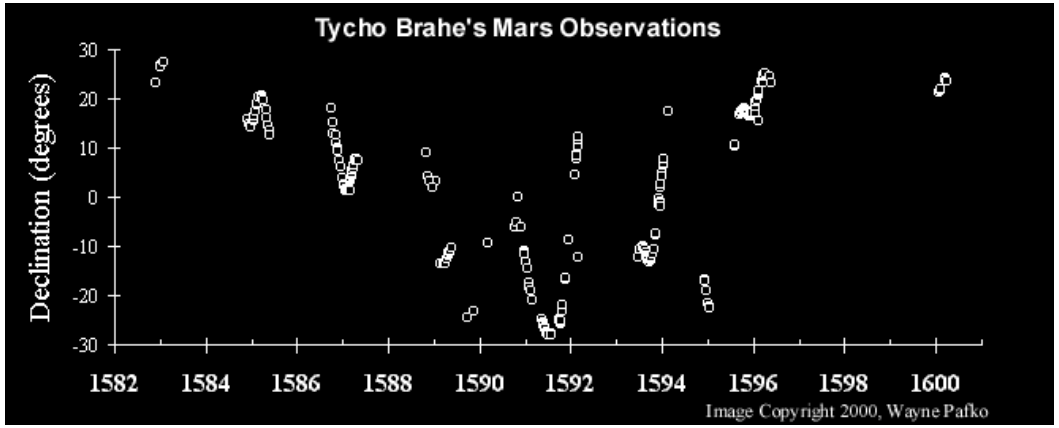


Genetic Material Sequencing

HPC: top500.org
Robotic lab: Nature, 583, 237–241 (2020)
Sequencing: www.nature.com/scitable/content/finding-genes-3485/

SISSO++ as a Symbolic Regression + Feature Selection tool

Keplerian: **data-driven** approach.



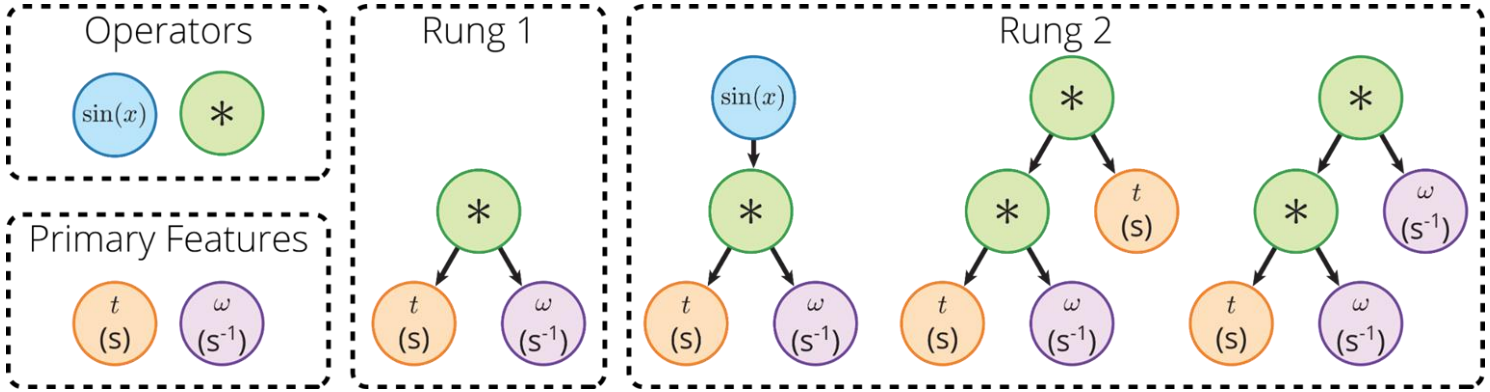
SISSO is one of the artificial intelligent method to help understand the data.
Advantages: (from Lucas Foppa's DPG talk)

- Provides interpretable analytical expressions.
- Tailored to small data and complex properties.

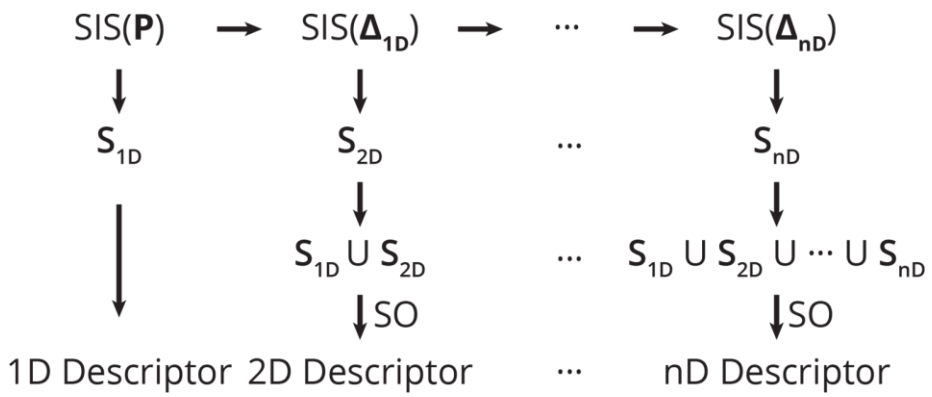
- http://www.pafko.com/tycho/observe.html

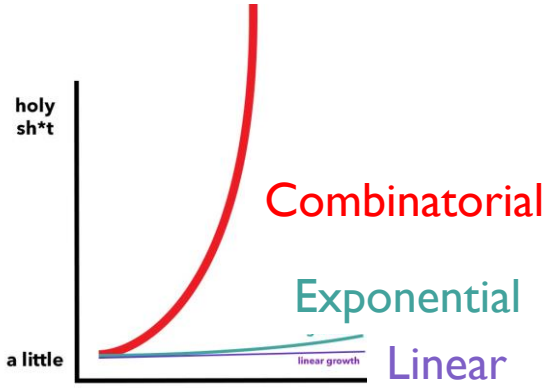
SISSO is a combinatorial scaling algorithm

feature generation



L0 regularization





Combinatorial explosion at higher rungs. Computational expensive at intermediate rungs.
Can we take advantage of HPC? GPU?

L0 regularization: Sebastian Eibl (Markus Rampp's group in MPCDF) takes on the L0 regularization.

my task: to optimize feature generation.

https://medium.com/@TorBair/exponential-growth-isn-t-cool-combinatorial-growth-is-85a0b1fdb6a5
Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M. Ghiringhelli, L. M. *Phys. Rev. Mater.* **2**, 083802. (2018)
Purcell, T. A. R.; Scheffler, M. Carbogno, C.; Ghiringhelli, L. M. *J. Open Source Softw.* **7**, 03960. (2022)

Porting of SISSO++ To GPUs and beyond (Taken from NOMAD WP6 report)

Porting to GPUs of SISSO++ (collaboration with MPCDF - M. Rampp)

Preliminary result on Nvidia GPU: (L0 time in this test case)

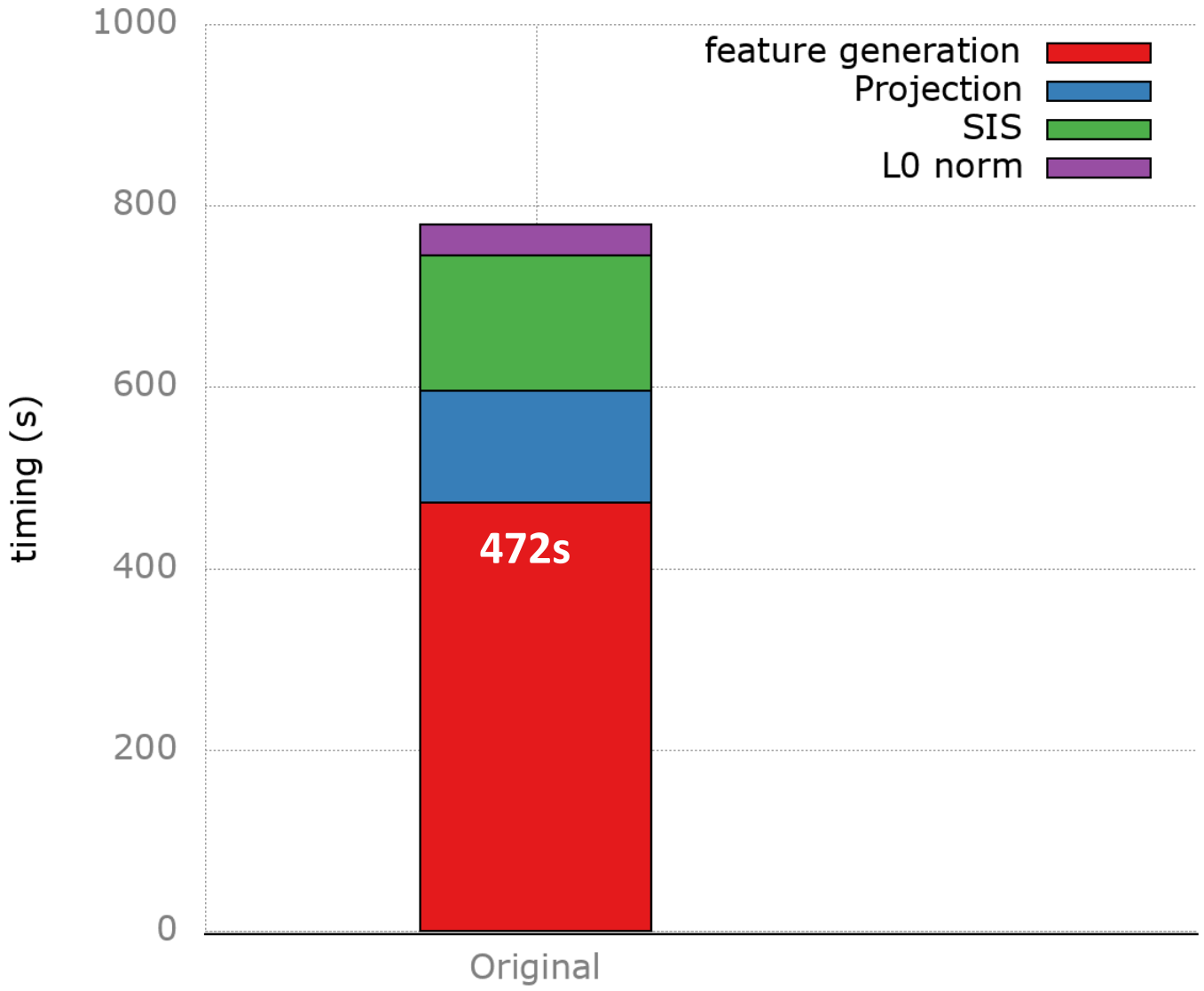
training of SISSO++ on CPU:	1.00	
training of SISSO++ on GPU only:	0.80	(1.25×)
combining CPU + GPU:	0.48	(2.08×)
single precision CPU + GPU:	0.35	(2.86×)

Max expected speedup: 7x. However, detected cuBLAS issue for small LR tasks (few data points). Nvidia is aware and (probably) working on it.

Preliminary test on AMD GPU: worse performance than Nvidia.

Planned test on other

Feature generation step

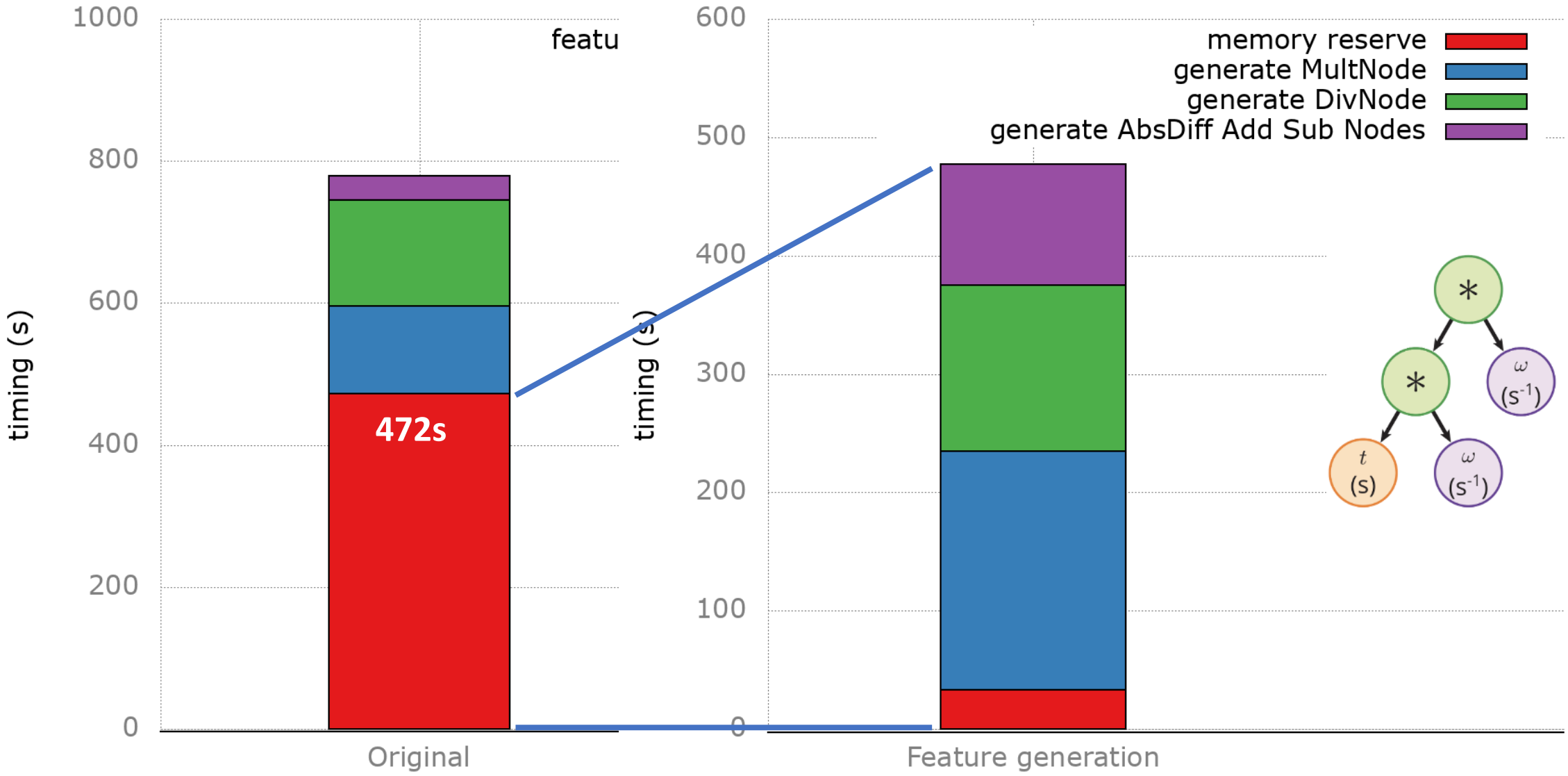


Problem (A Kaggle competition): Predict Eg(bandgap) based on the structure of InGaAlx? A problem where feature generation is the bottleneck.

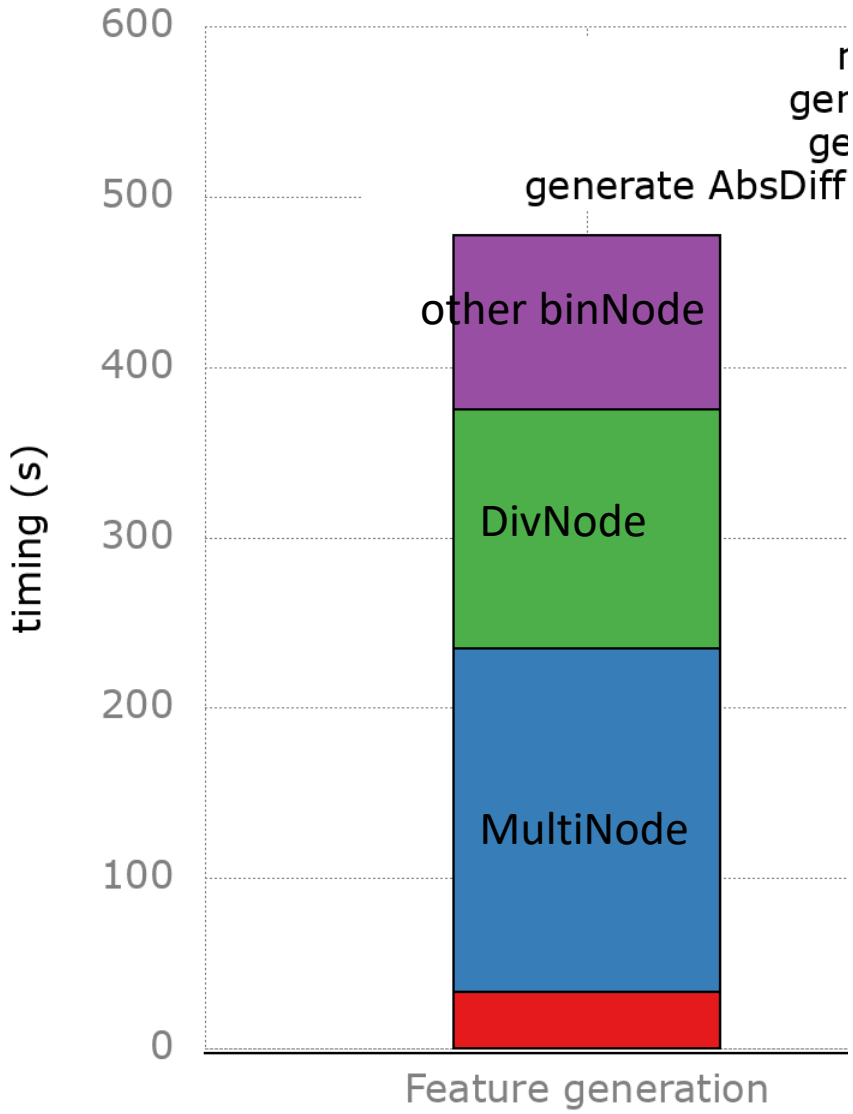
One node on Talos CPU only (40 cpu cores)

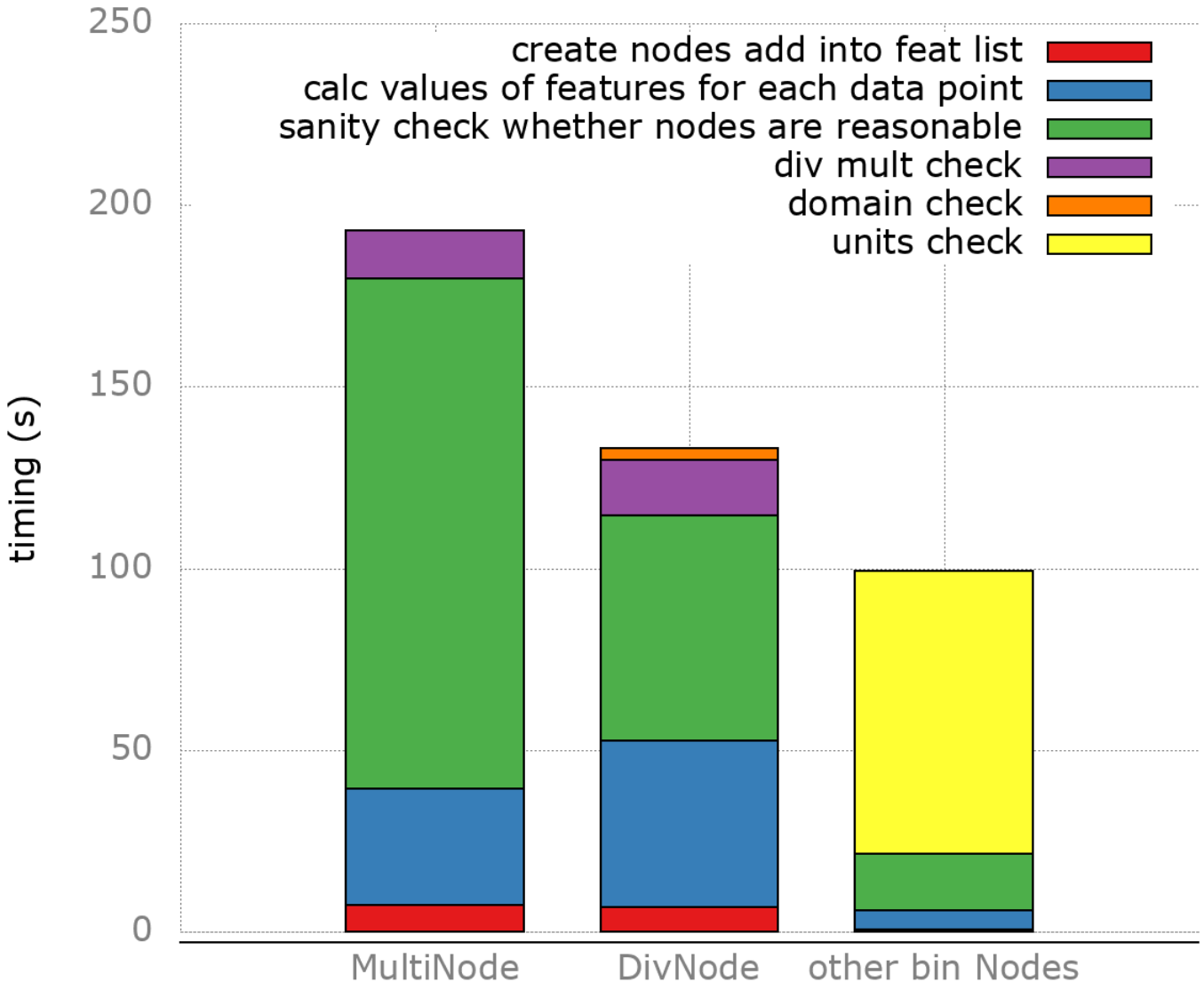
profiling tools:
Intel vtune
LLNL Caliper

Feature generation step – zoom in

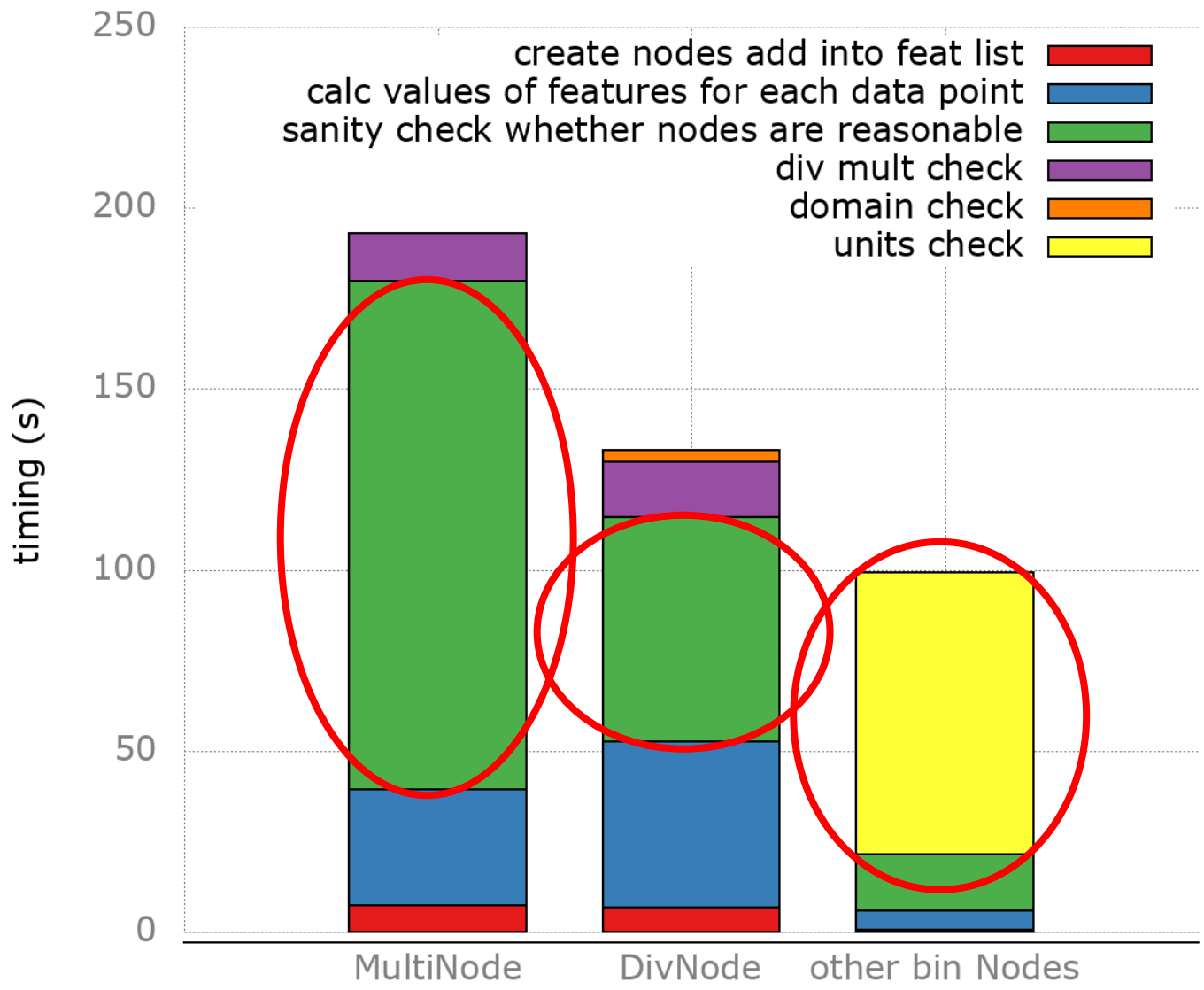


Feature generation step – keep zoom in





Feature generation step – real bottleneck



Surprisingly, create nodes/calc values of features is not the most time-consuming step.

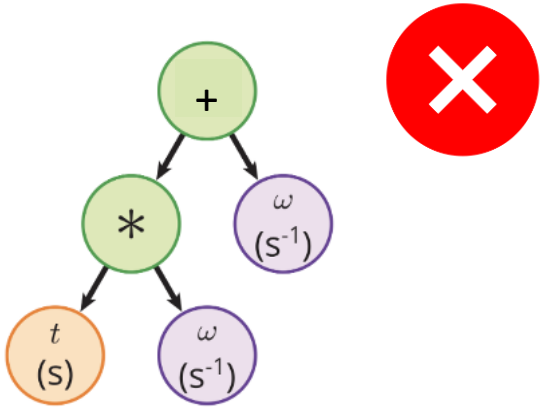
units check ~80s

sanity check ~220s

Feature generation step – units / sanity check

Units check

- With binary ops of +, -, absdiff, need to make sure the units are consistent.

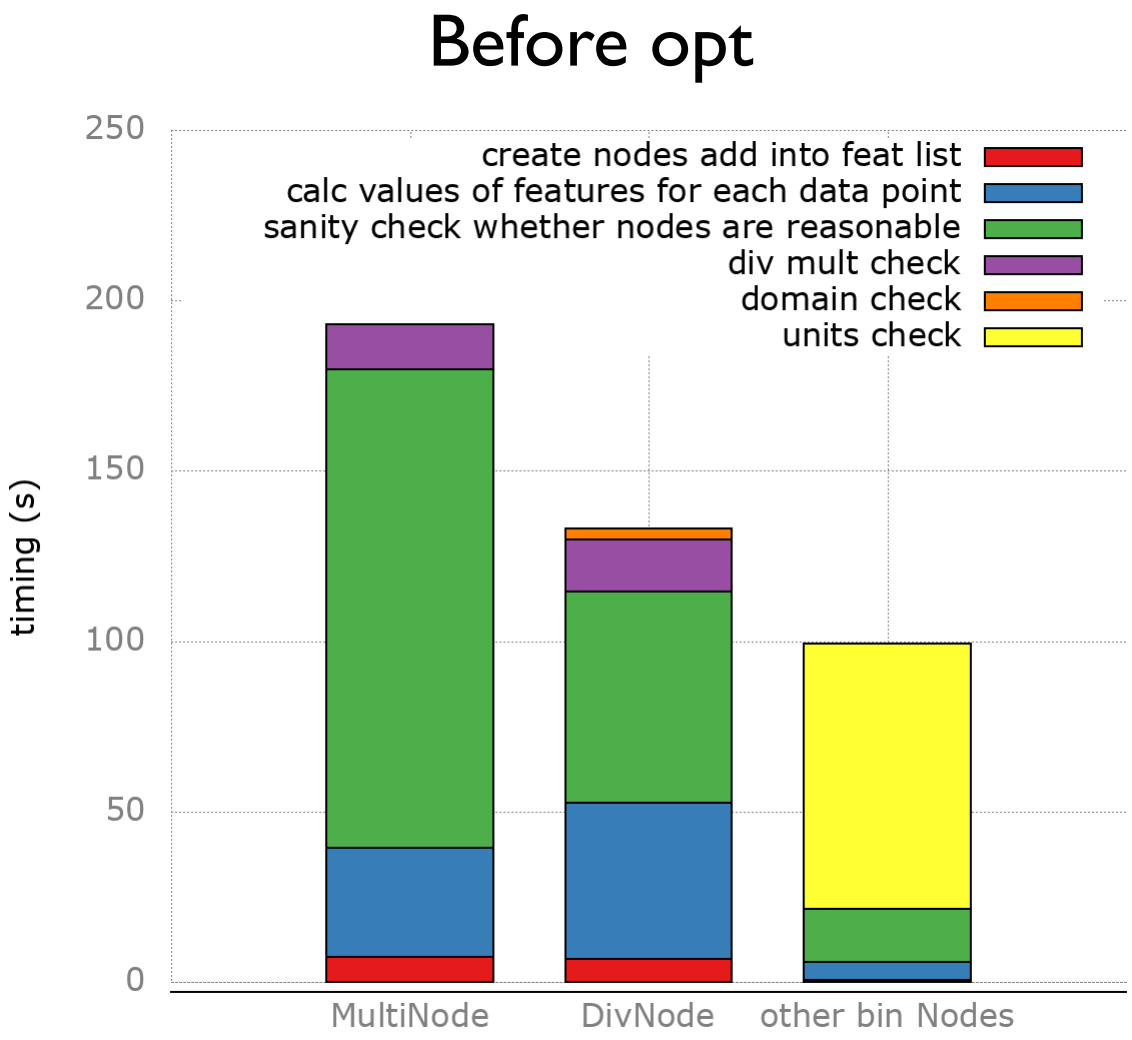


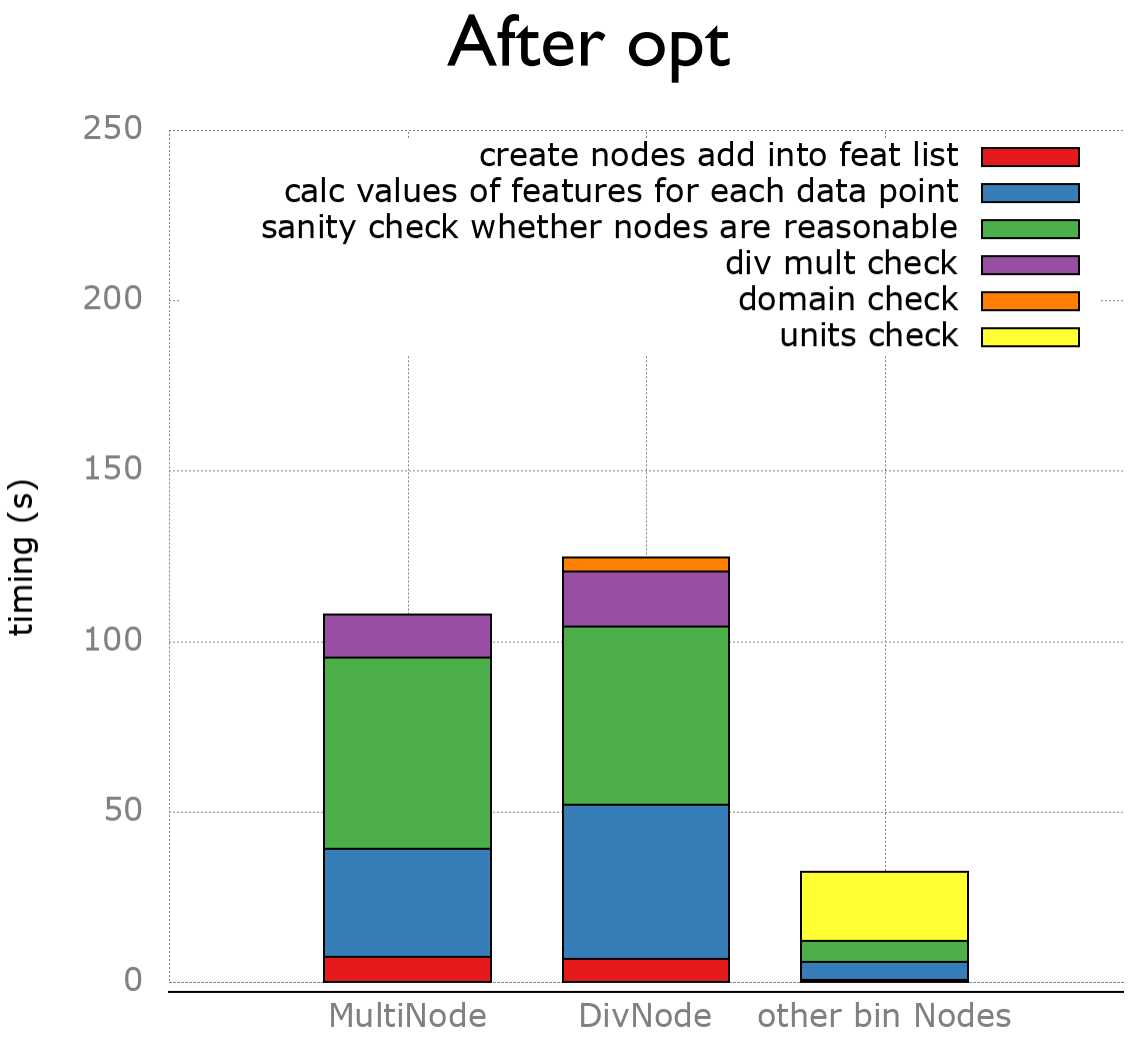
- was recursively calculated
- Dynamic programming (save calculated results and reuse)

Sanity check

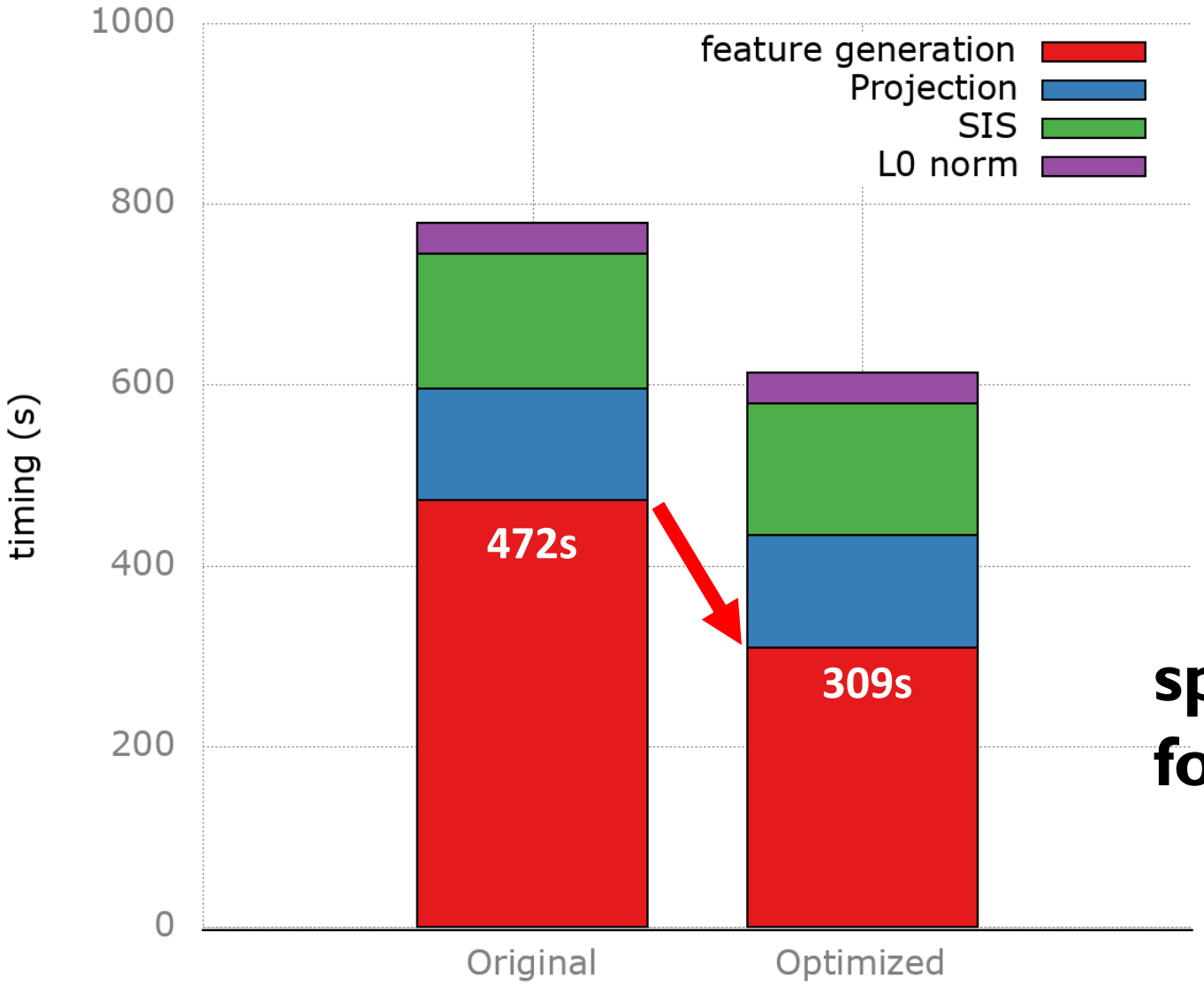
- make sure no underflow/overflow/is_const/isnan/…
- All are not necessary for every operators.
  - Remove unnecessary ones.
  - Reorder to reduce number of calls.

Feature generation step – timing after opt



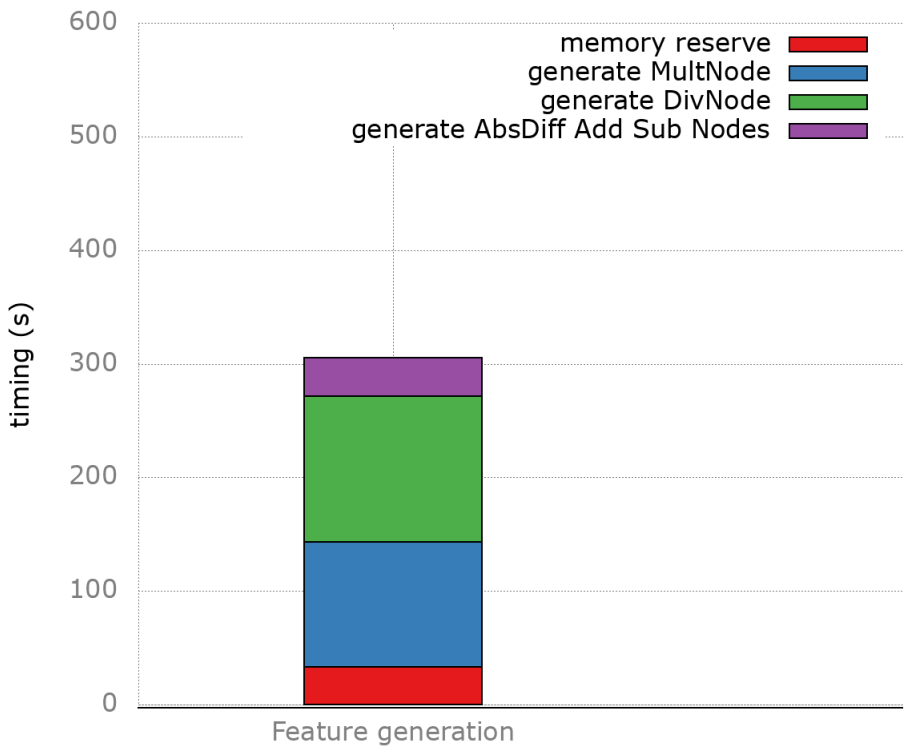


Feature generation step – speedup



speed 1.53x
for feature generation

Feature generation step – GPU plan and outlook



- **Optimization based on GPU/Kokkos. The numbers estimated here are using the CPU optimized code as reference.**
- **GPU won't help memory reserve ~11%, instead I will expect extra memory allocation/data movement for GPU. (~5-10%?)**
- **The remaining part (~89%) could be moved to GPU. I won't expect the ideal 7x speedup since we have if statement in the checks which are not ideal workload for GPU. expected 3-5x?**
- **Aiming for 2.5x-5x speed. 3.5x-6x speed for single precision? CPU+GPU loading 4x-6.5x?**

Extract information from Kepler's data with SISSO

Difference w.r.t. the modern measurement.

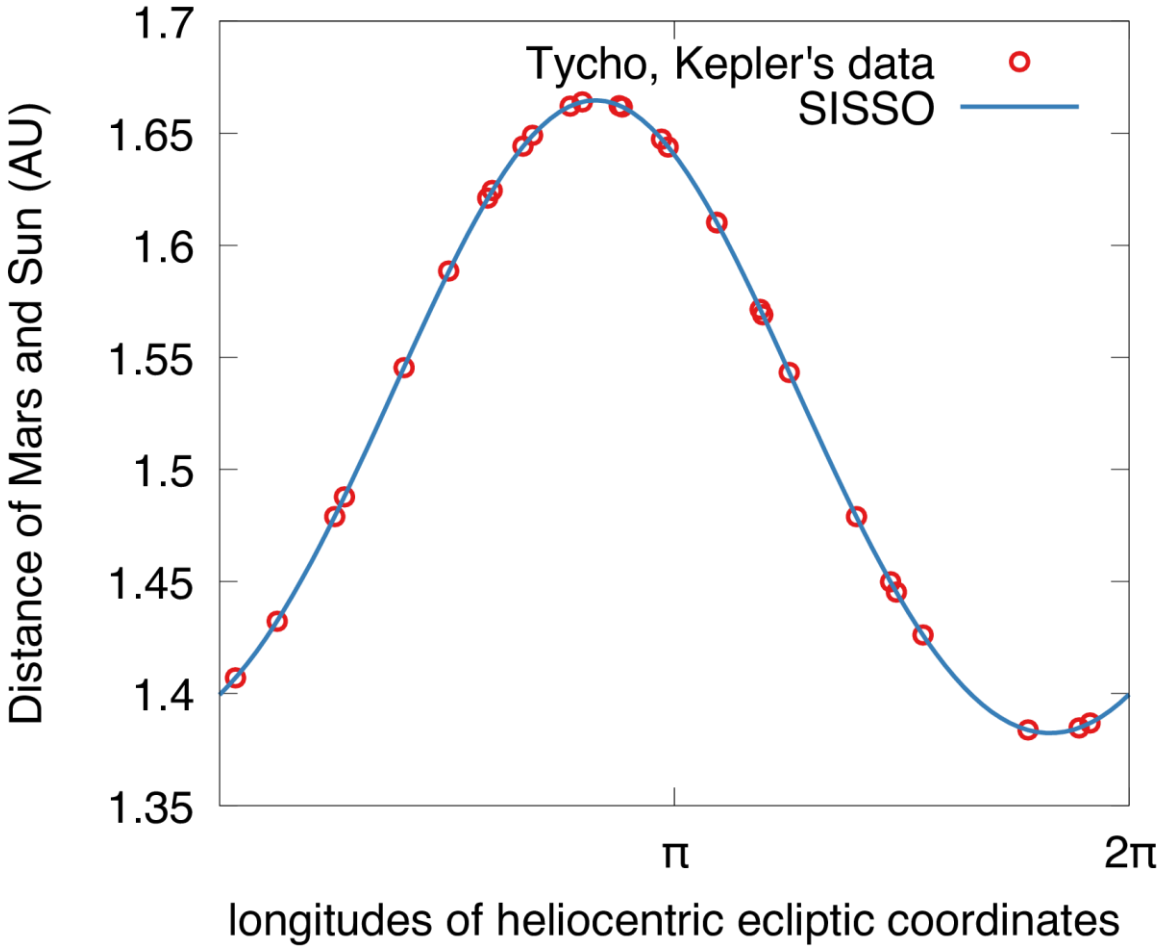
The observation data of Mars orbiting the Sun by Tycho and Kepler.
Data collected within this arXiv paper which was originally from Kepler's book Astronomia Nova.

Time YYYY/MM/DD	Mars' Position in Ecliptic	Sun-Mars Distance	Difference
1582/11/23 16:00	90.70306°	1.58852	$+1'30''$
1582/12/26 08:30	106.12167°	1.62104	$+3'49''$
1582/12/30 08:10	107.94222°	1.62443	$+5'50''$
1583/01/26 06:15	120.10667°	1.64421	$-2'33''$
1584/12/21 14:00	123.86250°	1.64907	$+1'04''$
1585/01/24 09:00	138.78556°	1.66210	$-3'32''$
1585/02/04 06:40	143.56139°	1.66400	$-3'08''$
1585/03/12 10:30	159.38722°	1.66170	$-2'29''$
1587/01/25 17:00	158.22778°	1.66232	$-0'10''$
1587/03/04 13:24	174.94722°	1.64737	$-0'59''$
1587/03/10 11:30	177.59833°	1.64382	$0'0''$
1587/04/21 09:30	196.74750°	1.61027	$+1'30''$
1589/05/08 16:24	196.92056°	1.61000	$-2'43''$
1589/04/13 11:15	214.03056°	1.57141	$+1'40''$
1589/04/15 12:05	215.02806°	1.56900	$+0'37''$
1589/05/06 11:20	225.51000°	1.54326	$+0'57''$
1591/05/13 14:00	252.12722°	1.47891	$-4'24''$
1591/06/06 12:20	265.64667°	1.44981	$-3'15''$
1591/06/10 11:50	267.94694°	1.44526	$-4'39''$
1591/06/28 10:24	278.49222°	1.42608	$-5'39''$
1593/07/21 14:00	320.02722°	1.38376	$-2'31''$
1593/08/22 12:20	340.25694°	1.38463	$-0'36''$
1593/08/29 10:20	344.62083°	1.38682	$-2'19''$
1593/10/03 08:00	6.32750°	1.40697	$-0'16''$
1595/09/17 16:45	22.82194°	1.43222	$-1'27''$
1595/10/27 12:20	45.59389°	1.47890	$-0'29''$
1595/11/03 12:00	49.44250°	1.48773	$+0'03''$
1595/12/18 08:00	73.04139°	1.54539	$-0'59''$

Table 1: Position of Mars when orbiting the Sun

Li, Zelong, Jianchao Ji, and Yongfeng Zhang. "From Kepler to Newton: the Role of Explainable AI in Science Discovery." *arXiv preprint arXiv:2111.12210* (2021).

Extract information from Kepler's data with SISSO



- SISSO Rung 2 w/ parameter optimization figured out an equation and the parameters.

l = 1.510425
ε = 0.0926795
c_1 = 1.000201
c_2 = 0.5439705

$$\frac{l}{1+\varepsilon\cdot\cos(c_1\theta+c_2)}$$

RMSE: 0.00013
Max AE: 0.000338

- Polar form of an ellipse relative to focus

$$r(\theta)=\frac{a(1-\varepsilon^2)}{1\pm\varepsilon\cos\theta} \qquad l=a(1-\varepsilon^2)$$

- Modern astronomy vs SISSO based on Kepler's data
- eccentricity of Mars (ε) 0.09341233 (SISSO 0.0926795)
- Semi-Major axis (a) 1.52366231 AU (SISSO 1.52351)

Li, Zelong, Jianchao Ji, and Yongfeng Zhang. "From Kepler to Newton: the Role of Explainable AI in Science Discovery." *arXiv preprint arXiv:2111.12210* (2021).

https://en.wikipedia.org/wiki/Ellipse