PSDI

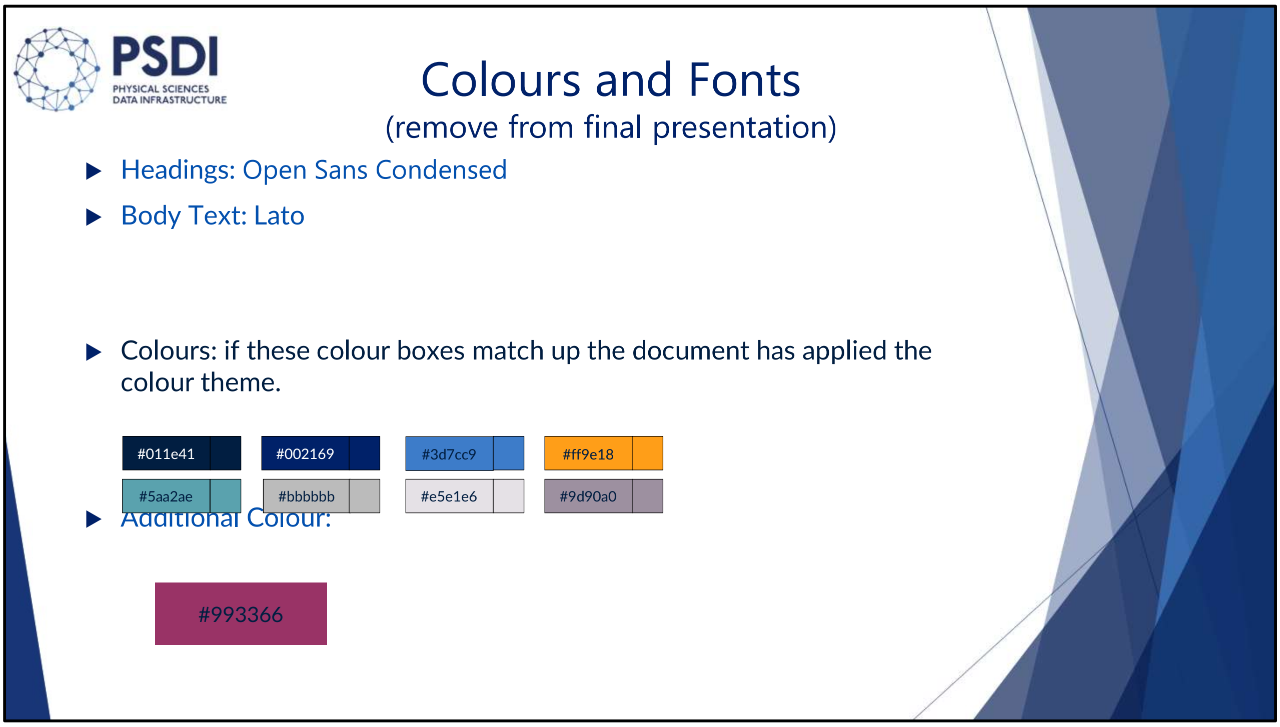
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Modelling phase transitions: Characterising Henry's Law

Joshua Cheung

Supervised by Jo Grundy and Jeremy Frey

https://www.psdi.ac.uk/



Colours and Fonts

(remove from final presentation)

- Headings: Open Sans Condensed
- Body Text: Lato
- Colours: if these colour boxes match up the document has applied the colour theme.

#011e41	#002169	#3d7cc9	#ff9e18
#5aa2ae	#bbbbbb	#e5e1e6	#9d90a0

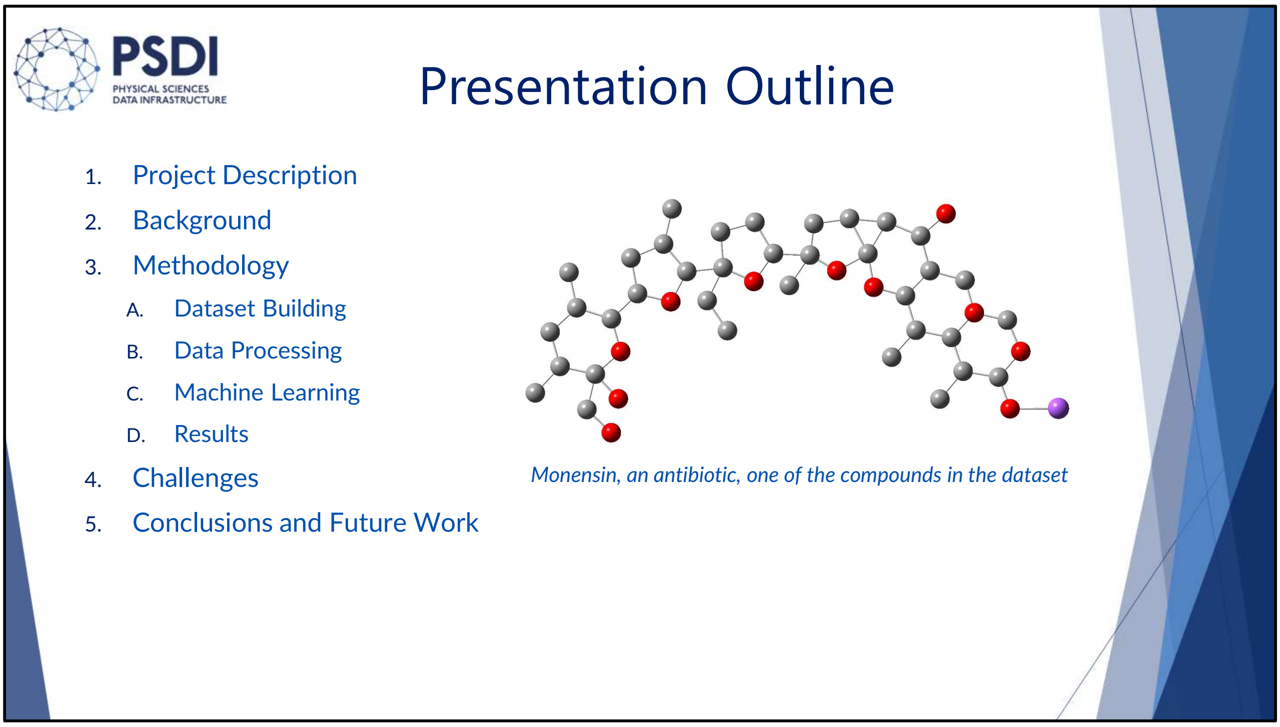
- Additional Colour:

#993366

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

About me

- 2nd Year Chemistry with Digital Methods and Computational Modelling
- I still don't have a photo of myself, so here's another cat photo I took
- Continuation of my project from last summer with PSDI: *Modelling Miscibility with Machine Learning*
- I enjoy reading sci fi, gaming, and playing guitar

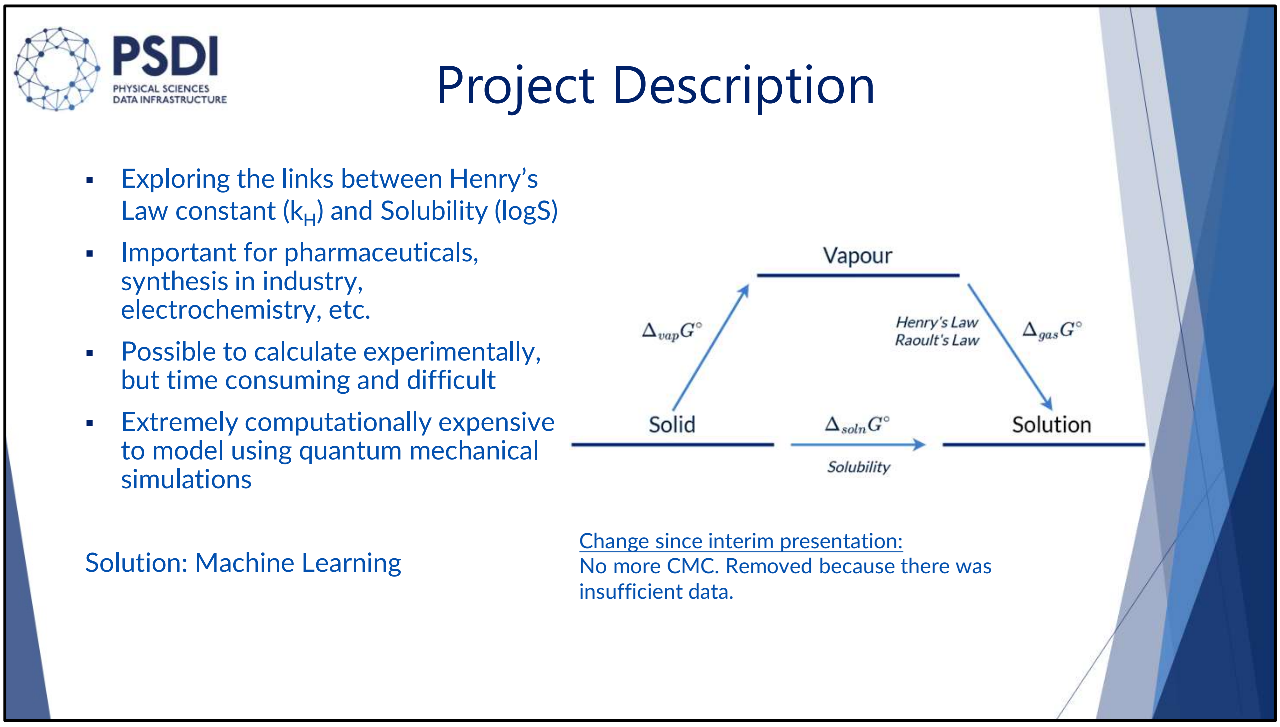


PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Presentation Outline

1. Project Description
2. Background
3. Methodology
   A. Dataset Building
   B. Data Processing
   C. Machine Learning
   D. Results
4. Challenges
5. Conclusions and Future Work

Monensin, an antibiotic, one of the compounds in the dataset



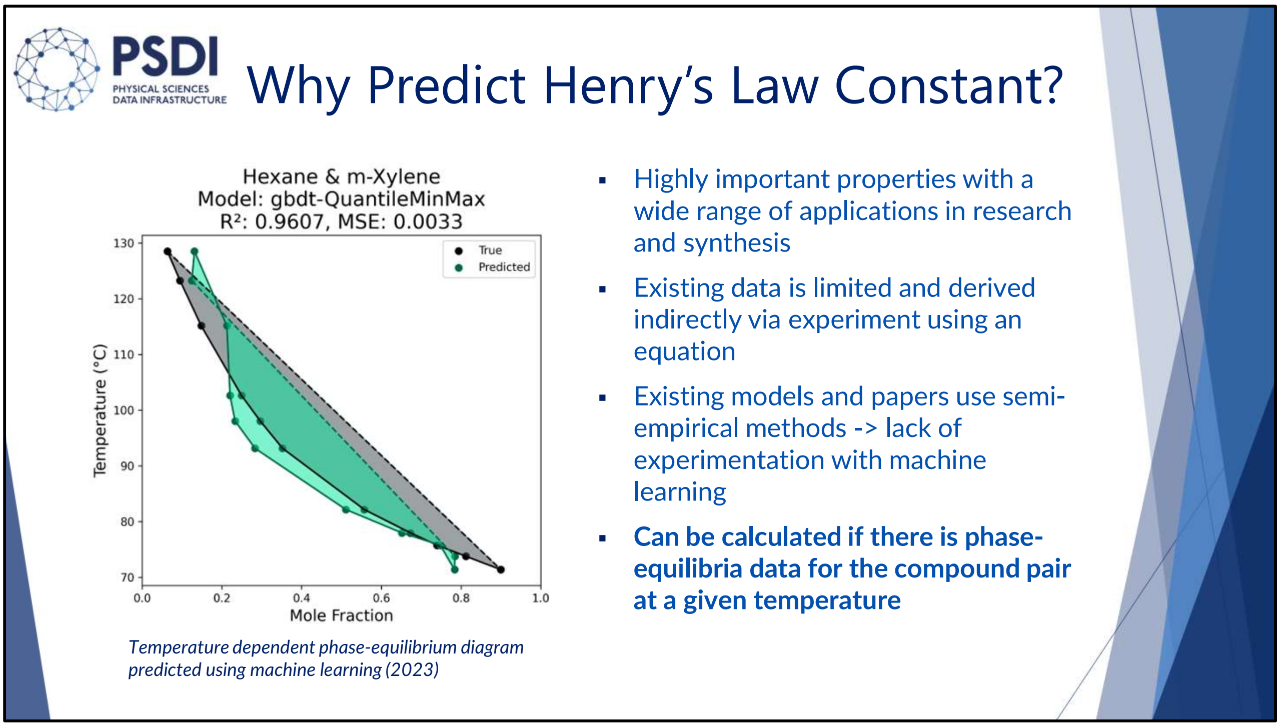
PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

Project Description

- Exploring the links between Henry's Law constant (k_H) and Solubility (logS)
- Important for pharmaceuticals, synthesis in industry, electrochemistry, etc.
- Possible to calculate experimentally, but time consuming and difficult
- Extremely computationally expensive to model using quantum mechanical simulations

Solution: Machine Learning

Change since interim presentation:
No more CMC. Removed because there was insufficient data.

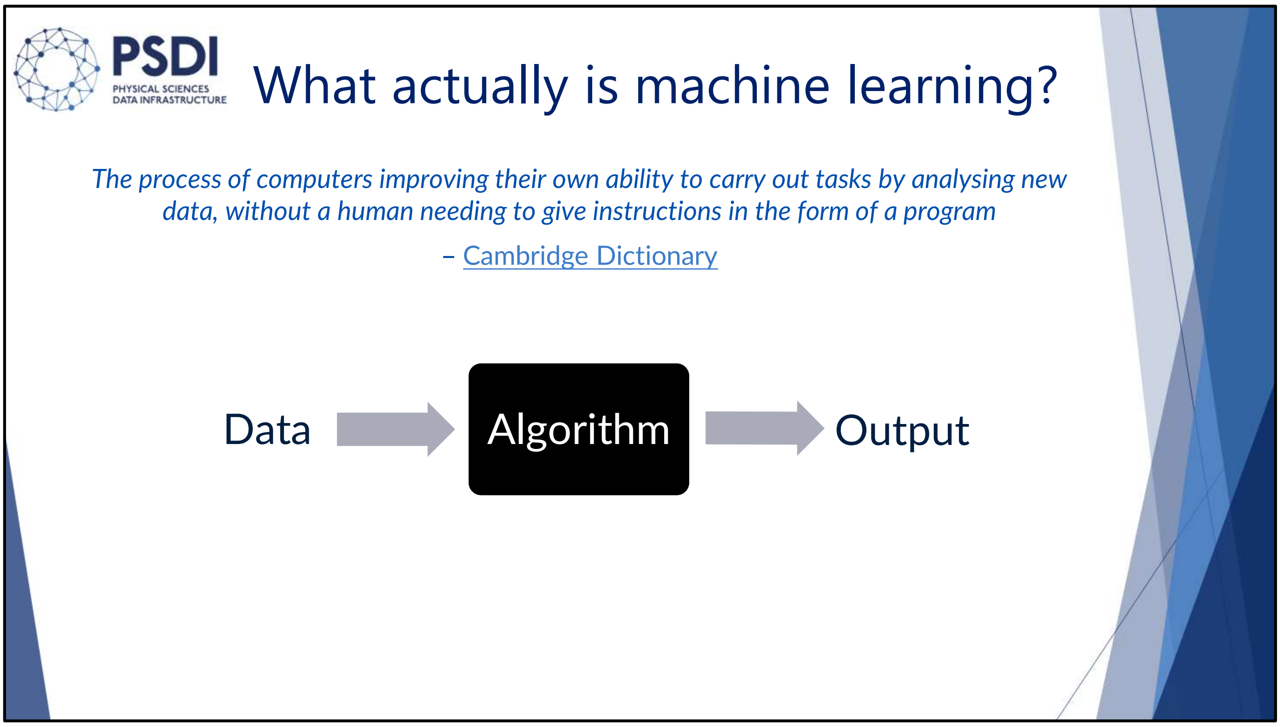


PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

Why Predict Henry's Law Constant?

Temperature dependent phase-equilibrium diagram predicted using machine learning (2023)

- Highly important properties with a wide range of applications in research and synthesis
- Existing data is limited and derived indirectly via experiment using an equation
- Existing models and papers use semi-empirical methods -> lack of experimentation with machine learning
- **Can be calculated if there is phase-equilibria data for the compound pair at a given temperature**

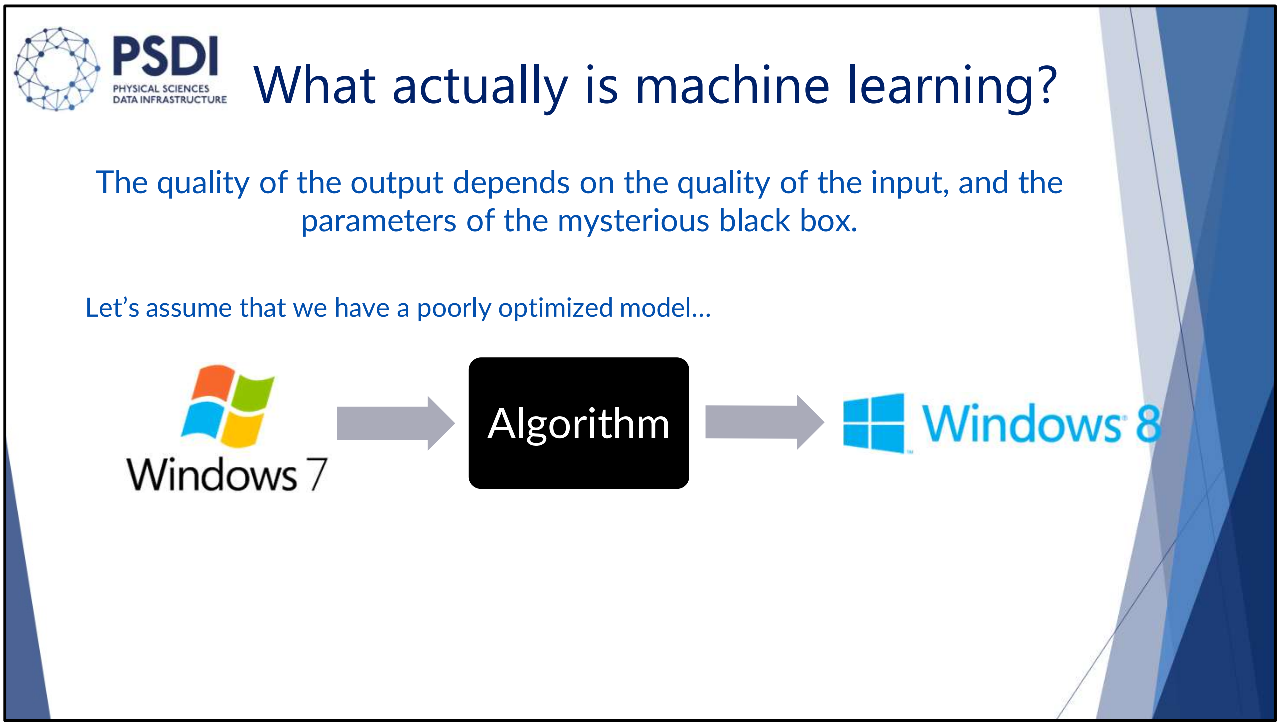


What actually is machine learning?

The process of computers improving their own ability to carry out tasks by analysing new data, without a human needing to give instructions in the form of a program

– Cambridge Dictionary

Data → Algorithm → Output



What actually is machine learning?

The quality of the output depends on the quality of the input, and the parameters of the mysterious black box.

Let’s assume that we have a poorly optimized model…

Algorithm

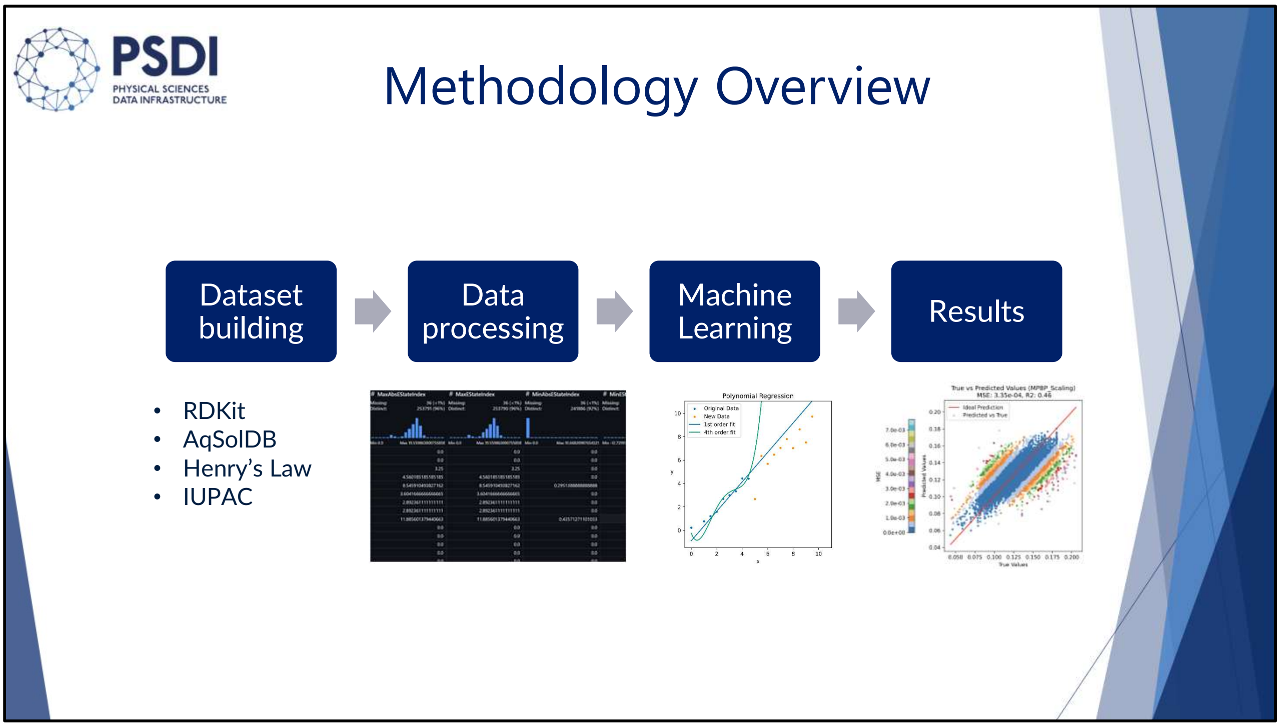
Background

Henry’s Law: *The abundance of a volatile solute dissolved in a liquid is proportional to its abundance in the gas phase.*

Solubility: *The analytical composition of a saturated solution, expressed in terms of the proportion of a designated solute in a designated solvent, is the solubility of that solute.*

Or more simply: How much of the substance (liquid or gas) can be dissolved in the solvent?

https://iupac.org/recommendation/henrys-law-constants/
https://goldbook.iupac.org/terms/view/S05740
https://www.youtube.com/watch?v=zMaTrgUKC1w



PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Methodology Overview

Dataset building → Data processing → Machine Learning → Results

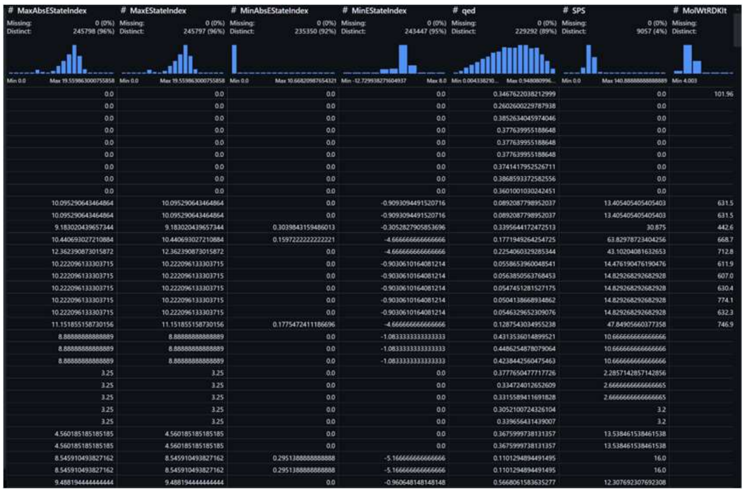
- RDKit
- AqSolDB
- Henry's Law
- IUPAC

Overall process

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Dataset Building

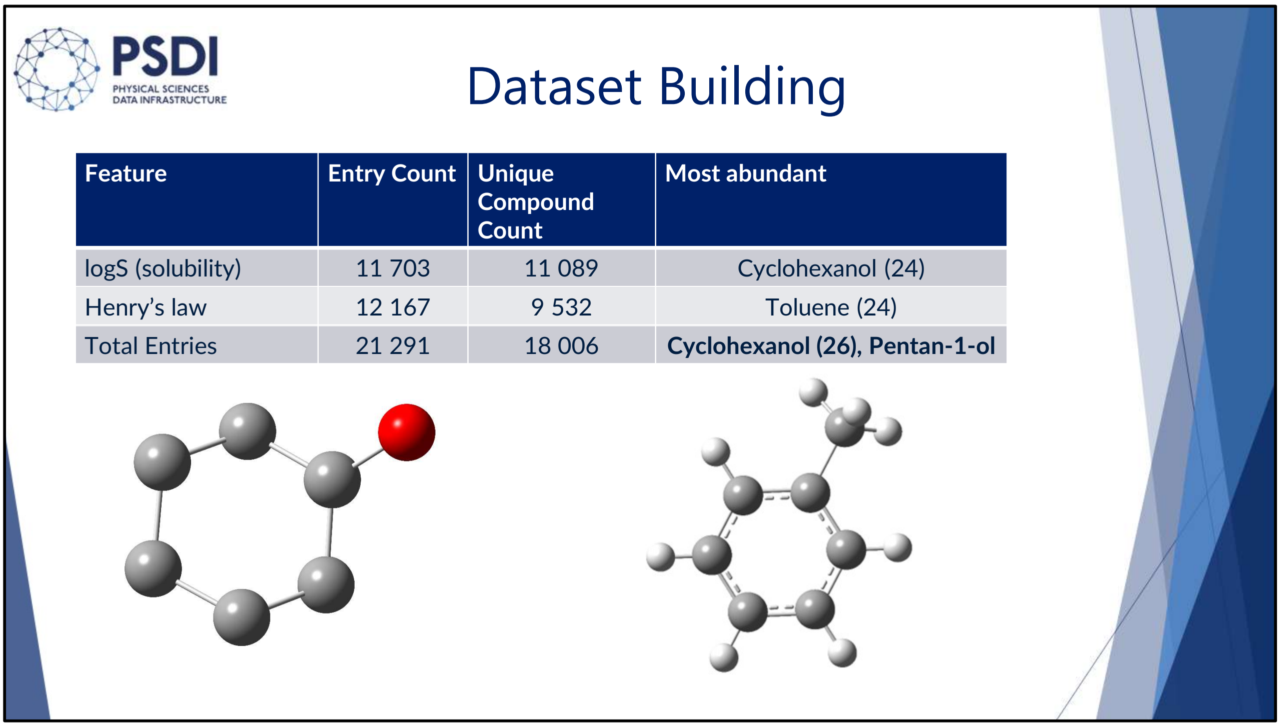
MAX-PLANCK-GESELLSCHAFT



# MaxAbsEStateIndex	# MaxEStateIndex	# MinAbsEStateIndex	# MinEStateIndex	# qed	# SPS	# MolWtRDKit
Missing: 0 (0%)	Missing: 0 (0%)	Missing: 0 (0%)	Missing: 0 (0%)	Missing: 0 (0%)	Missing: 0 (0%)	Missing:
Distinct: 245798 (96%)	Distinct: 245797 (96%)	Distinct: 235350 (92%)	Distinct: 243447 (95%)	Distinct: 229292 (89%)	Distinct: 9057 (4%)	Distinct:
Min 0.0 Max 19.559863000755858	Min 0.0 Max 19.559863000755858	Min 0.0 Max 10.668209876543211	Min -12.729938271604937 Max 8.0	Min 0.004338210... Max 0.948080996...	Min 0.0 Max 140.88888888888889	Min 4.003
0.0	0.0	0.0	0.0	0.3467622038212999	0.0	101.96
0.0	0.0	0.0	0.0	0.2602600229787938	0.0	
0.0	0.0	0.0	0.0	0.3852634045974046	0.0	
0.0	0.0	0.0	0.0	0.377639955188648	0.0	
0.0	0.0	0.0	0.0	0.377639955188648	0.0	
0.0	0.0	0.0	0.0	0.377639955188648	0.0	
0.0	0.0	0.0	0.0	0.3741417952526711	0.0	
0.0	0.0	0.0	0.0	0.3868593372582556	0.0	
0.0	0.0	0.0	0.0	0.3601001030242451	0.0	
10.095290643464864	10.095290643464864	0.0	-0.9093094491520716	0.0892087798952037	13.405405405405403	631.5
10.095290643464864	10.095290643464864	0.0	-0.9093094491520716	0.0892087798952037	13.405405405405403	631.5
9.183020439657344	9.183020439657344	0.3039843159486013	-0.3052827905853696	0.3395644172472513	30.875	442.6
10.440693027210884	10.440693027210884	0.1597222222222221	-4.666666666666666	0.1771949264254725	63.82978723404256	668.7
12.362390873015872	12.362390873015872	0.0	-4.666666666666666	0.2254060329285344	43.10204081632653	712.8
10.222096133303715	10.222096133303715	0.0	-0.9030610164081214	0.0558653960048541	14.476190476190476	611.9
10.222096133303715	10.222096133303715	0.0	-0.9030610164081214	0.0563850563768453	14.829268292682928	607.0
10.222096133303715	10.222096133303715	0.0	-0.9030610164081214	0.0547451281527175	14.829268292682928	630.4
10.222096133303715	10.222096133303715	0.0	-0.9030610164081214	0.0504138668934862	14.829268292682928	774.1
10.222096133303715	10.222096133303715	0.0	-0.9030610164081214	0.0546329652309076	14.829268292682928	632.3
11.151855158730156	11.151855158730156	0.1775472411186696	-4.666666666666666	0.1287543034955238	47.84905660377358	746.9
8.88888888888889	8.88888888888889	0.0	-1.0833333333333333	0.4313536014899521	10.666666666666666	
8.88888888888889	8.88888888888889	0.0	-1.0833333333333333	0.4486254878079064	10.666666666666666	
8.88888888888889	8.88888888888889	0.0	-1.0833333333333333	0.4238442560475463	10.666666666666666	
3.25	3.25	0.0	0.0	0.3777650477717726	2.2857142857142856	
3.25	3.25	0.0	0.0	0.334724012652609	2.6666666666666665	
3.25	3.25	0.0	0.0	0.3315589411691828	2.6666666666666665	
3.25	3.25	0.0	0.0	0.3052100724326104	3.2	
3.25	3.25	0.0	0.0	0.339656431439007	3.2	
4.560185185185185	4.560185185185185	0.0	0.0	0.3675999738131357	13.538461538461538	
4.560185185185185	4.560185185185185	0.0	0.0	0.3675999738131357	13.538461538461538	
8.545910493827162	8.545910493827162	0.2951388888888888	-5.166666666666666	0.1101294894491495	16.0	
8.545910493827162	8.545910493827162	0.2951388888888888	-5.166666666666666	0.1101294894491495	16.0	
9.488194444444444	9.488194444444444	0.0	-0.960648148148148	0.5668061583635277	12.307692307692308	

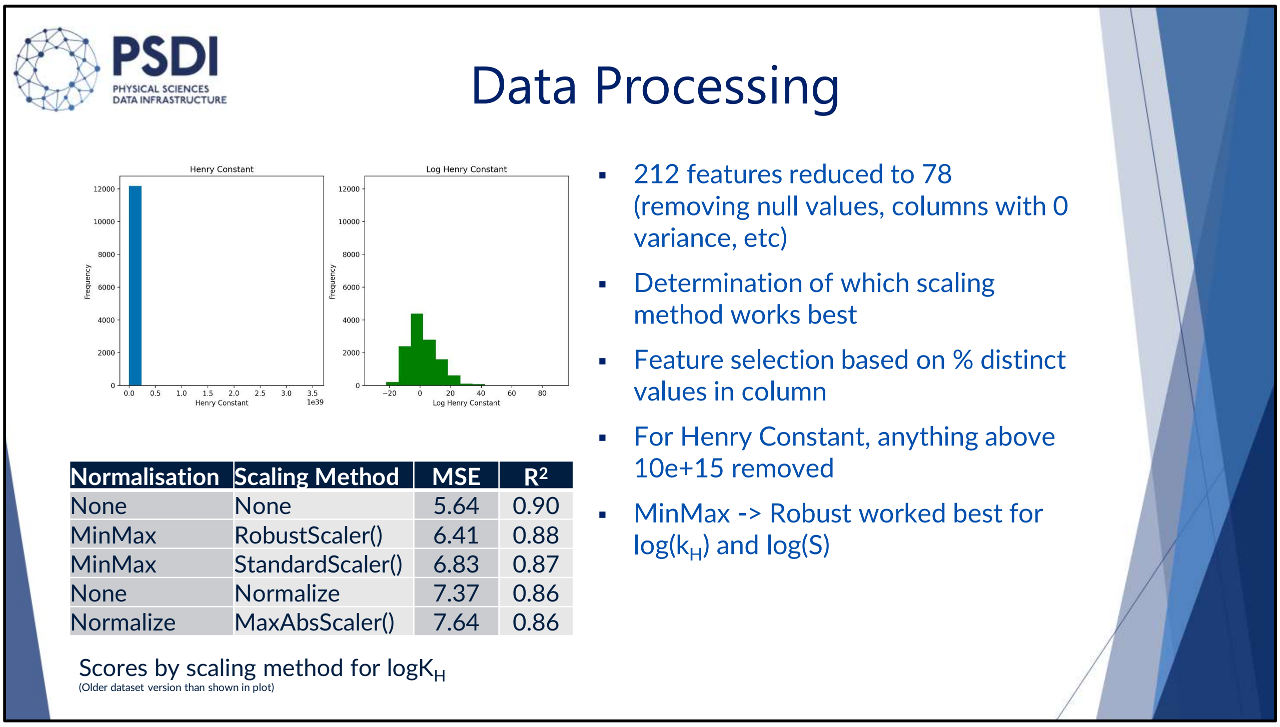
IUPAC
INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY

RDKit
Open-Source Cheminformatics
and Machine Learning



Dataset Building

Feature	Entry Count	Unique Compound Count	Most abundant
logS (solubility)	11 703	11 089	Cyclohexanol (24)
Henry's law	12 167	9 532	Toluene (24)
Total Entries	21 291	18 006	**Cyclohexanol (26), Pentan-1-ol**



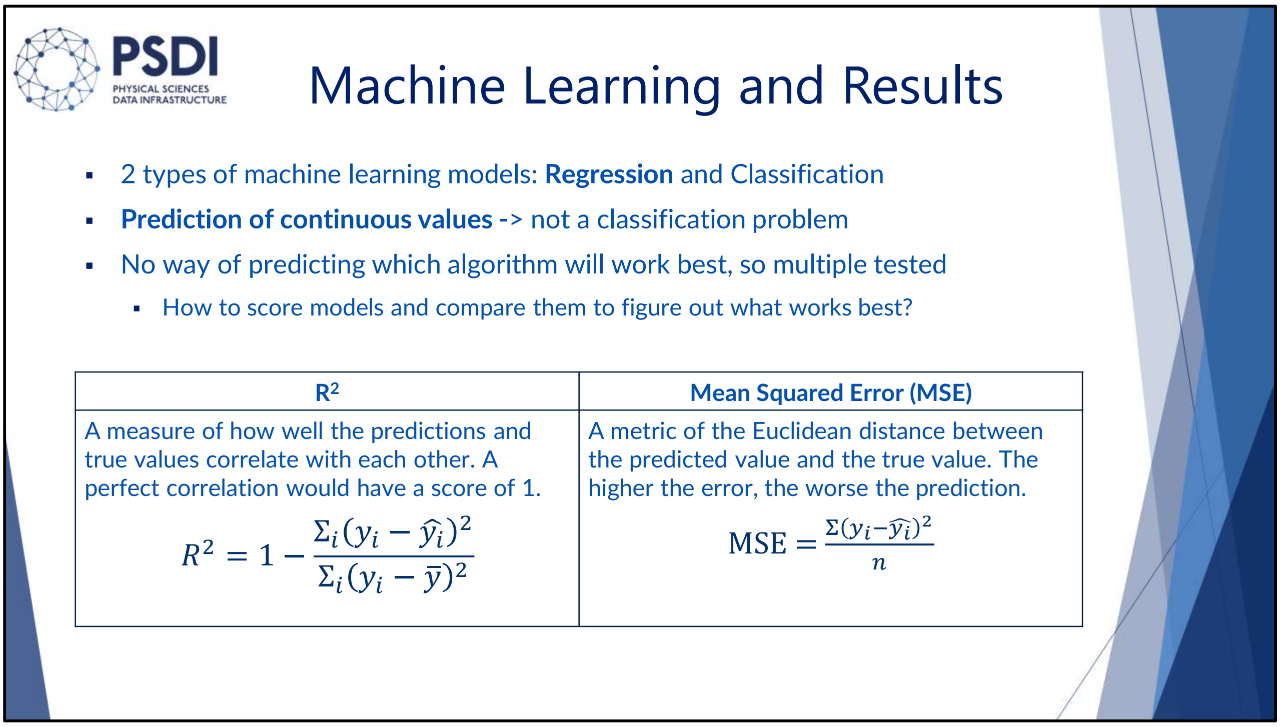
Data Processing

- 212 features reduced to 78 (removing null values, columns with 0 variance, etc)
- Determination of which scaling method works best
- Feature selection based on % distinct values in column
- For Henry Constant, anything above 10e+15 removed
- MinMax -> Robust worked best for log(k_H) and log(S)

Normalisation	Scaling Method	MSE	R^2
None	None	5.64	0.90
MinMax	RobustScaler()	6.41	0.88
MinMax	StandardScaler()	6.83	0.87
None	Normalize	7.37	0.86
Normalize	MaxAbsScaler()	7.64	0.86

Scores by scaling method for logK_H
(Older dataset version than shown in plot)

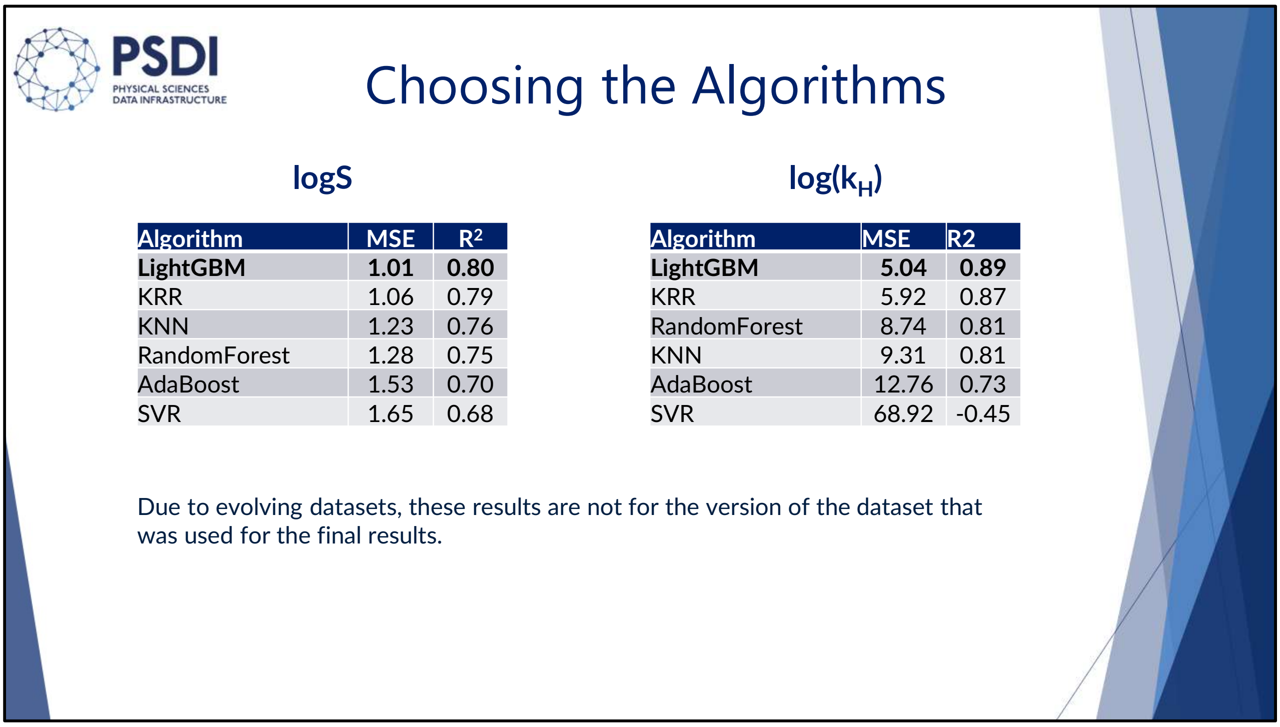
Scaling affects distribution of data, makes it easier for model to effectively differentiate between values
Feature selection tested variety of different ways, eventually just removed anything with <1% distinct values



Machine Learning and Results

- 2 types of machine learning models: **Regression** and Classification
- **Prediction of continuous values ->** not a classification problem
- No way of predicting which algorithm will work best, so multiple tested
  - How to score models and compare them to figure out what works best?

R^2	Mean Squared Error (MSE)
A measure of how well the predictions and true values correlate with each other. A perfect correlation would have a score of 1. $R^2 = 1 - \frac{\Sigma_i (y_i - \widehat{y}_i)^2}{\Sigma_i (y_i - \bar{y})^2}$	A metric of the Euclidean distance between the predicted value and the true value. The higher the error, the worse the prediction. $\mathrm{MSE} = \frac{\Sigma(y_i - \widehat{y_i})^2}{n}$



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Choosing the Algorithms

logS

Algorithm	MSE	R^2
LightGBM	**1.01**	**0.80**
KRR	1.06	0.79
KNN	1.23	0.76
RandomForest	1.28	0.75
AdaBoost	1.53	0.70
SVR	1.65	0.68

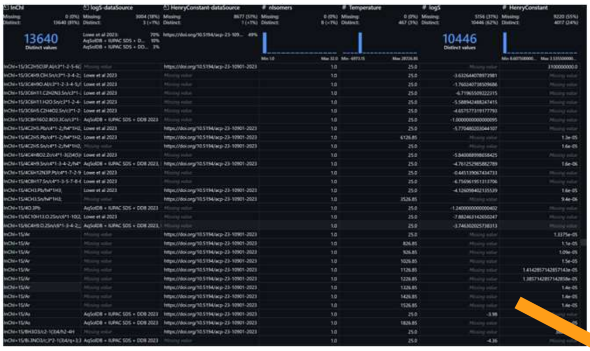
$\log(k_H)$

Algorithm	MSE	R2
LightGBM	**5.04**	**0.89**
KRR	5.92	0.87
RandomForest	8.74	0.81
KNN	9.31	0.81
AdaBoost	12.76	0.73
SVR	68.92	-0.45

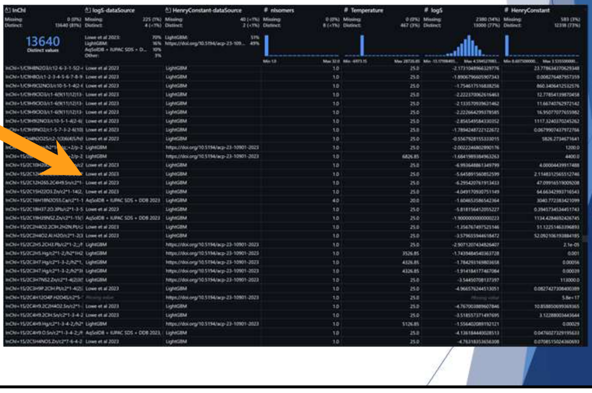
Due to evolving datasets, these results are not for the version of the dataset that was used for the final results.

Results and Analysis

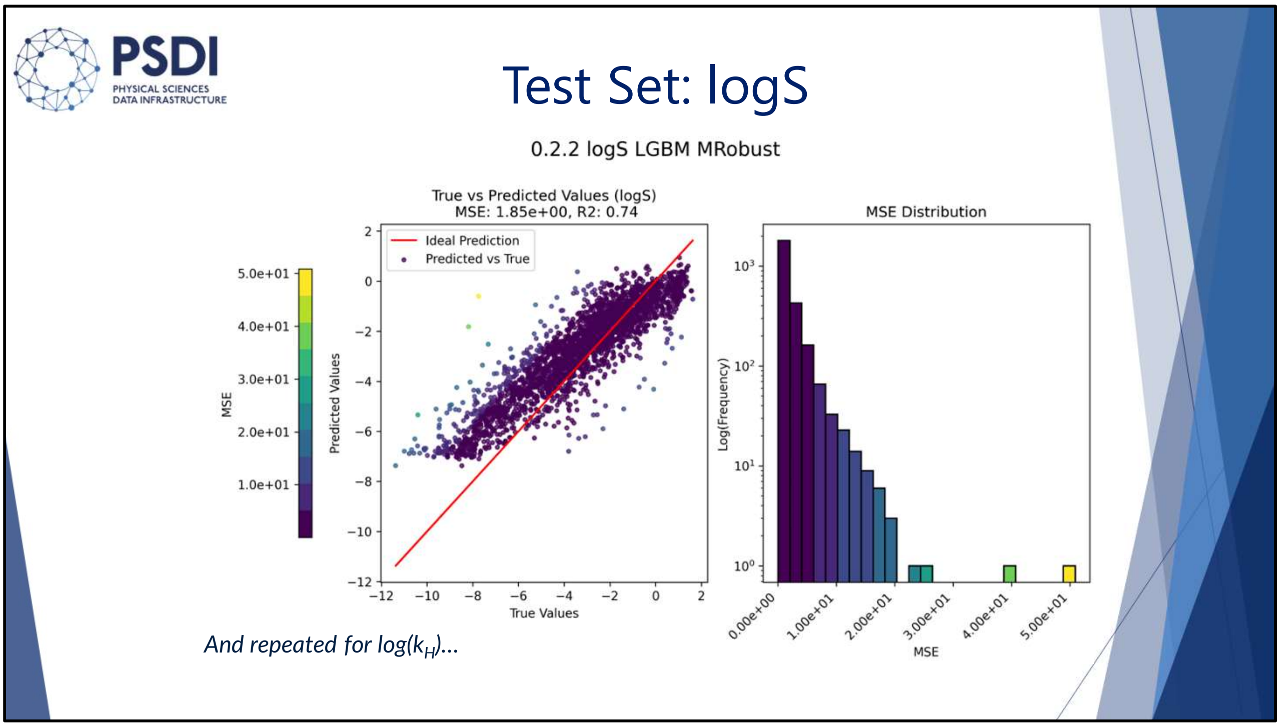
- Trained model to predict logS using data that has no k_H value
- Used model to predict logS and fill in dataset gaps
- Withheld test set of data that contains logS and Henry’s constant value



InChI	logS-dataSource	HenryConstant-dataSource	# nIsomers	# Temperature	# logS	# HenryConstant
Missing: 0 (0%) Distinct: 13640 (81%) 13640 Distinct values	Missing: 3004 (18%) Distinct: 3 (<1%) Lowe et al 2023: 70% AqSolDB + IUPAC SDS + D... 10% AqSolDB + IUPAC SDS + DD... 3%	Missing: 8677 (51%) Distinct: 1 (<1%) https://doi.org/10.5194/acp-23-109... 49%	Missing: 0 (0%) Distinct: 8 (<1%) Min 1.0 Max 32.0	Missing: 0 (0%) Distinct: 467 (3%) Min -6973.15 Max 28726.85	Missing: 5156 (31%) Distinct: 10446 (62%) 10446 Distinct values	Missing: 9220 (55%) Distinct: 4017 (24%) Min 6.607500000... Max 3.535000000...
InChI=1S/3C2H5O3P.Al/c3*1-2-5-6(	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	Missing value	3100000000.0
InChI=1S/3C4H9.ClH.Sn/c3*1-3-4-2;;	Lowe et al 2023	Missing value	1.0	25.0	-3.632644078973981	Missing value
InChI=1S/3C4H9O.Al/c3*1-2-3-4-5;/	Lowe et al 2023	Missing value	1.0	25.0	-1.760240738509686	Missing value
InChI=1S/3C6H11.C2H2N3.Sn/c3*1-;	Lowe et al 2023	Missing value	1.0	25.0	-6.71965509222315	Missing value
InChI=1S/3C6H11.H2O.Sn/c3*1-2-4-	Lowe et al 2023	Missing value	1.0	25.0	-5.588942488247415	Missing value
InChI=1S/3C6H5.C2H4O2.Sn/c3*1-2-	Lowe et al 2023	Missing value	1.0	25.0	-4.657577319177793	Missing value
InChI=1S/3C8H16O2.BO3.3Co/c3*1-	AqSolDB + IUPAC SDS + DDB 2023	Missing value	1.0	25.0	-1.0000000000000095	Missing value
InChI=1S/4C2H5.Pb/c4*1-2;/h4*1H2;	Lowe et al 2023	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	-5.770480203044107	Missing value
InChI=1S/4C2H5.Pb/c4*1-2;/h4*1H2;	Lowe et al 2023	https://doi.org/10.5194/acp-23-10901-2023	1.0	6126.85	Missing value	1.3e-05
InChI=1S/4C2H5.Sn/c4*1-2;/h4*1H2;	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	Missing value	1.6e-05
InChI=1S/4C4H8O2.Zr/c4*1-3(2)4(5)	Lowe et al 2023	Missing value	1.0	25.0	-5.840088998658425	Missing value
InChI=1S/4C4H9.Sn/c4*1-3-4-2;/h4*	AqSolDB + IUPAC SDS + DDB 2023,	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	-4.761252985882789	1.6e-06
InChI=1S/4C6H12N3P.Pt/c4*1-7-2-9	Lowe et al 2023	Missing value	1.0	25.0	-0.445139067434733	Missing value
InChI=1S/4C8H17.Sn/c4*1-3-5-7-8-6	Lowe et al 2023	Missing value	1.0	25.0	-6.756961951313706	Missing value
InChI=1S/4CH3.Pb/h4*1H3;	Lowe et al 2023	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	-4.126098402135539	1.6e-05
InChI=1S/4CH3.Sn/h4*1H3;	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	3526.85	Missing value	9.4e-06
InChI=1S/4O.2P/i	AqSolDB + IUPAC SDS + DDB 2023	Missing value	1.0	25.0	-1.2400000000000402	Missing value
InChI=1S/6C10H13.O.2Sn/c6*1-10(2,	Lowe et al 2023	Missing value	1.0	25.0	-7.882463142650247	Missing value
InChI=1S/6C4H9.O.2Sn/c6*1-3-4-2;;;	AqSolDB + IUPAC SDS + DDB 2023,	Missing value	1.0	25.0	-3.746302025738313	Missing value
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	Missing value	1.3375e-05
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-22-10901-2023	1.0	826.85	Missing value	1.1e-05
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	926.85	Missing value	1.09e-05
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	1026.85	Missing value	1.5e-05
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	1126.85	Missing value	1.4142857142857143e-05
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	1226.85	Missing value	1.3857142857142858e-05
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	1326.85	Missing value	1.4e-05
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	1426.85	Missing value	1.4e-05
InChI=1S/Ar	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	1526.85	Missing value	1.4e-05
InChI=1S/As	AqSolDB + IUPAC SDS + DDB 2023	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	-3.98	Missing value
InChI=1S/As	AqSolDB + IUPAC SDS + DDB 2023	https://doi.org/10.5194/acp-23-10901-2023	1.0	1826.85	Missing value	[illegible]
InChI=1S/BH3O3/c2-1(3)4/h2-4H	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	Missing value	[illegible]
InChI=1S/Bi.3NO3/c;3*2-1(3)4/q+3;3	AqSolDB + IUPAC SDS + DDB 2023	Missing value	1.0	25.0	-4.36	Missing value



InChI	logS-dataSource	HenryConstant-dataSource	# nIsomers	# Temperature	# logS	# HenryConstant
Missing: 0 (0%) Distinct: 13640 (81%) 13640 Distinct values	Missing: 225 (1%) Distinct: 4 (<1%) Lowe et al 2023: 70% LightGBM 16% AqSolDB + IUPAC SDS + D... 10% Other: 3%	Missing: 40 (<1%) Distinct: 2 (<1%) LightGBM 51% https://doi.org/10.5194/acp-23-109... 49%	Missing: 0 (0%) Distinct: 8 (<1%) Min 1.0 Max 32.0	Missing: 0 (0%) Distinct: 467 (3%) Min -6973.15 Max 28726.85	Missing: 2380 (14%) Distinct: 13000 (77%) Min -13.17098493... Max 4.394527085...	Missing: 583 (3%) Distinct: 12398 (73%) Min 6.607500000... Max 3.535000000...
InChI=1/C9H8N2O3/c12-6-3-1-5(2-4	Lowe et al 2023	LightGBM	1.0	25.0	-2.1731048966329776	23.778634370629348
InChI=1/C9H8O/c1-2-3-4-5-6-7-8-9	Lowe et al 2023	LightGBM	1.0	25.0	-1.8906796605907343	0.008276487957359
InChI=1/C9H9ClNO3/c10-5-1-4(2-6	Lowe et al 2023	LightGBM	1.0	25.0	-1.754617516838256	860.3406412532576
InChI=1/C9H9ClO3/c1-6(9(11)12)13-	Lowe et al 2023	LightGBM	1.0	25.0	-2.222370062616463	12.77854139870458
InChI=1/C9H9ClO3/c1-6(9(11)12)13-	Lowe et al 2023	LightGBM	1.0	25.0	-2.133570939621462	11.66740762972142
InChI=1/C9H9ClO3/c1-6(9(11)12)13-	Lowe et al 2023	LightGBM	1.0	25.0	-2.222664299378585	16.95077077655982
InChI=1/C9H9I2NO3/c10-5-1-4(2-6)	Lowe et al 2023	LightGBM	1.0	25.0	-2.856549584330052	1117.3240370245262
InChI=1/C9H9NO2/c1-5-7-3-2-6(10)	Lowe et al 2023	LightGBM	1.0	25.0	-1.7894248722122672	0.0679907437972766
[illegible]CHN2O2S/c2-10(6(4)5/h(	Lowe et al 2023	LightGBM	1.0	25.0	-0.5567928155330015	5826.2734671641
InC[illegible]/h2*1[illegible];+2/p-2	LightGBM	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	-2.0022246802990176	1200.0
InChI=1S/2[illegible]-2/p-2	LightGBM	https://doi.org/10.5194/acp-23-10901-2023	1.0	6826.85	-1.6841989384963263	4400.0
InChI=1S/2C10H20[illegible]/c2	Lowe et al 2023	LightGBM	1.0	25.0	-6.993648861349799	4.00004439917488
InChI=1S/2C12[illegible]	Lowe et al 2023	LightGBM	1.0	25.0	-5.645891560852599	2.114831256512746
InChI=1S/2C12H26S.2C4H9.Sn/c2*1-	Lowe et al 2023	LightGBM	1.0	25.0	-6.295420761913433	47.09916519009208
InChI=1S/2C15H22O3.Zn/c2*1-14(2,	Lowe et al 2023	LightGBM	1.0	25.0	-4.049170930751149	64.66342993716543
InChI=1S/2C16H18N2O5S.Ca/c2*1-1	AqSolDB + IUPAC SDS + DDB 2023	LightGBM	4.0	20.0	-1.604653586542364	3040.772383421099
InChI=1S/2C18H37.2O.3Pb/c2*1-3-5	Lowe et al 2023	LightGBM	1.0	25.0	-5.818156412055227	0.3945734534451743
InChI=1S/2C19H39NS2.Zn/c2*1-15(1	AqSolDB + IUPAC SDS + DDB 2023	LightGBM	1.0	25.0	-1.9000000000000223	1134.4284692426745
InChI=1S/2C2H4O2.2ClH.2H2N.Pt/c2	Lowe et al 2023	LightGBM	1.0	25.0	-1.356767497525146	51.12251463396893
InChI=1S/2C2H4O2.Al.H2O/c2*1-2(3	Lowe et al 2023	LightGBM	1.0	25.0	-3.579635944618472	52.092106193884185
InChI=1S/2C2H5.2ClH.Pb/c2*1-2;;/	LightGBM	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	-2.9071207434826407	2.1e-05
InChI=1S/2C2H5.Hg/c2*1-2;/h2*1H2	LightGBM	https://doi.org/10.5194/acp-23-10901-2023	1.0	3526.85	-1.7439484540363728	0.001
InChI=1S/2C3H7.Hg/c2*1-3-2;/h2*1,	LightGBM	https://doi.org/10.5194/acp-23-10901-2023	1.0	4326.85	-1.784293169803658	0.00056
InChI=1S/2C3H7.Hg/c2*1-3-2;/h2*3H	LightGBM	https://doi.org/10.5194/acp-23-10901-2023	1.0	4326.85	-1.914184177467084	0.00039
InChI=1S/2C3H7NS2.Zn/c2*1-4(2)3(	LightGBM	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	-3.54450708137397	113000.0
InChI=1S/2C3H9P.2ClH.Pt/c2*1-4(2	Lowe et al 2023	LightGBM	1.0	25.0	-4.966576244513051	0.0827427308400389
InChI=1S/2C4H12O4P.H2O4S/c2*5-	Missing value	https://doi.org/10.5194/acp-23-10901-2023	1.0	25.0	Missing value	5.8e+17
InChI=1S/2C4H9.2C2H4O2.Sn/c2*1-	Lowe et al 2023	LightGBM	1.0	25.0	-4.767003889607846	10.858850699369365
InChI=1S/2C4H9.2ClH.Sn/c2*1-3-4-2	Lowe et al 2023	LightGBM	1.0	25.0	-3.518557371497695	3.12288003443644
InChI=1S/2C4H9.Hg/c2*1-3-4-2;/h2*	LightGBM	https://doi.org/10.5194/acp-23-10901-2023	1.0	5126.85	-1.556402089192121	0.00029
InChI=1S/2C4H9.O.Sn/c2*1-3-4-2;;/t	AqSolDB + IUPAC SDS + DDB 2023,	LightGBM	1.0	25.0	-4.136184440028513	0.0476027329195633
InChI=1S/2C5H4NOS.Zn/c2*7-6-4-2	Lowe et al 2023	LightGBM	1.0	25.0	-4.76318353656308	0.0708515024360693



PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

Test Set: logS

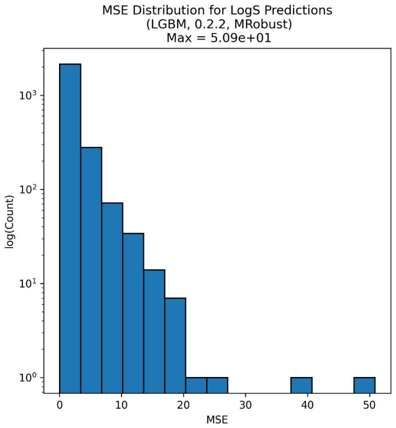
0.2.2 logS LGBM MRobust

True vs Predicted Values (logS)
MSE: 1.85e+00, R2: 0.74

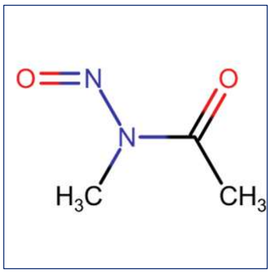
MSE Distribution

And repeated for log(k_H)…

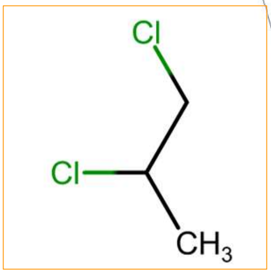
Test Set Metrics: logS



R^2	0.73
MSE	1.85
% MSE > 1	45.52 %

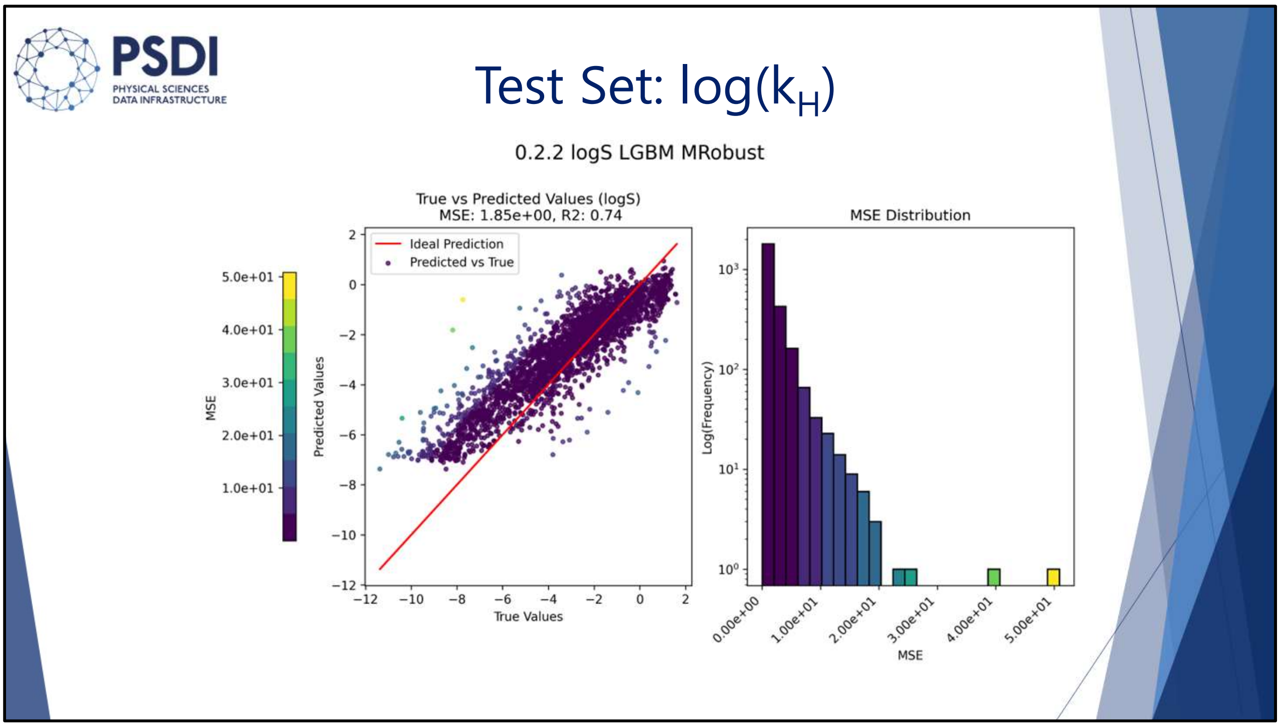


Lowest error: 6.00×10^{-7}



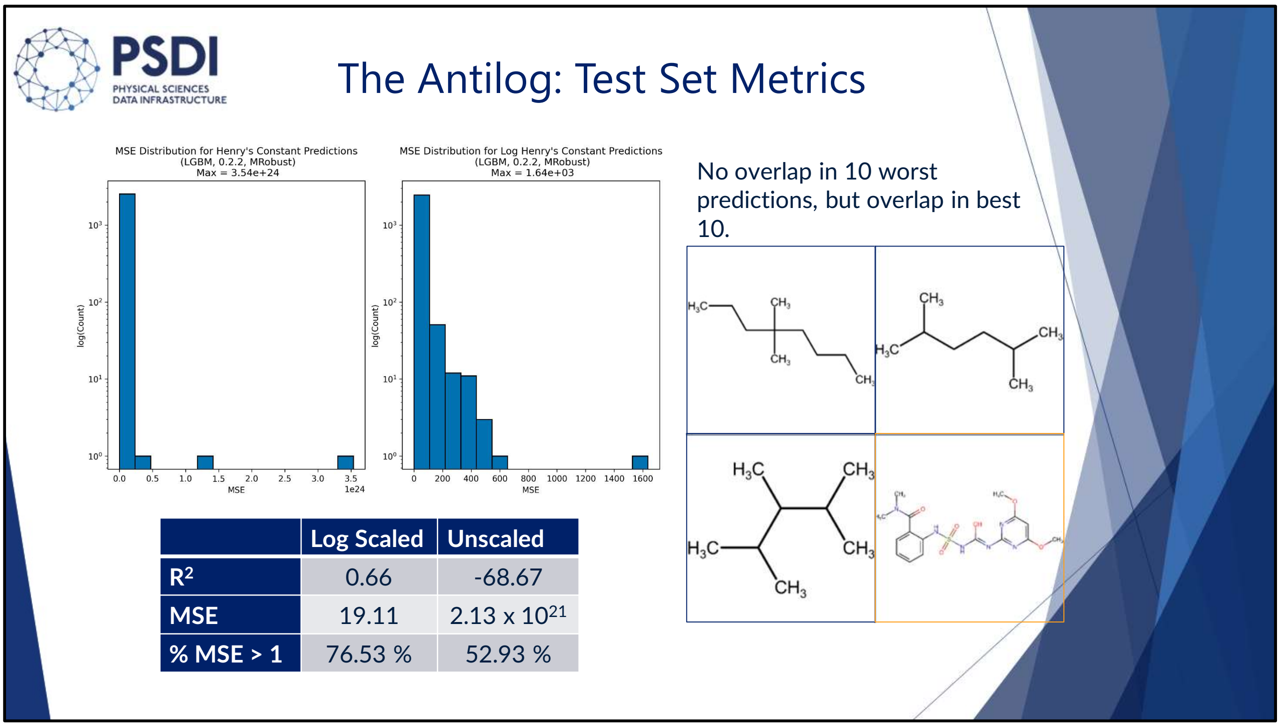
Highest error: 50.80

- 6th highest error in the test set is ammonia, NH_3, with an MSE of 18.51.
- Only bottom 5 have MSE > 20



Test Set: $\log(k_H)$

0.2.2 logS LGBM MRobust

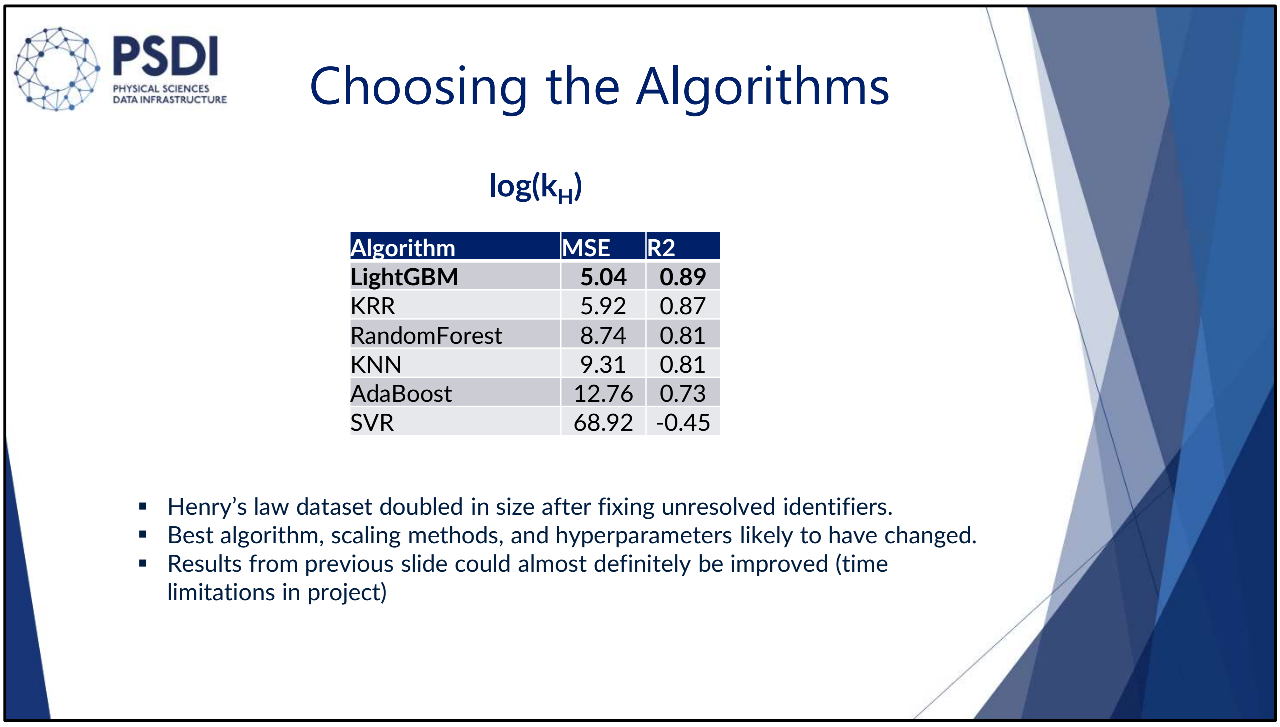


The Antilog: Test Set Metrics

No overlap in 10 worst predictions, but overlap in best 10.

	Log Scaled	Unscaled
R^2	0.66	-68.67
MSE	19.11	2.13 x 10^{21}
% MSE > 1	76.53 %	52.93 %

Highly likely that these results could be improved by rerunning hyperparameter optimisation, etc, as dataset size doubled with bug fix

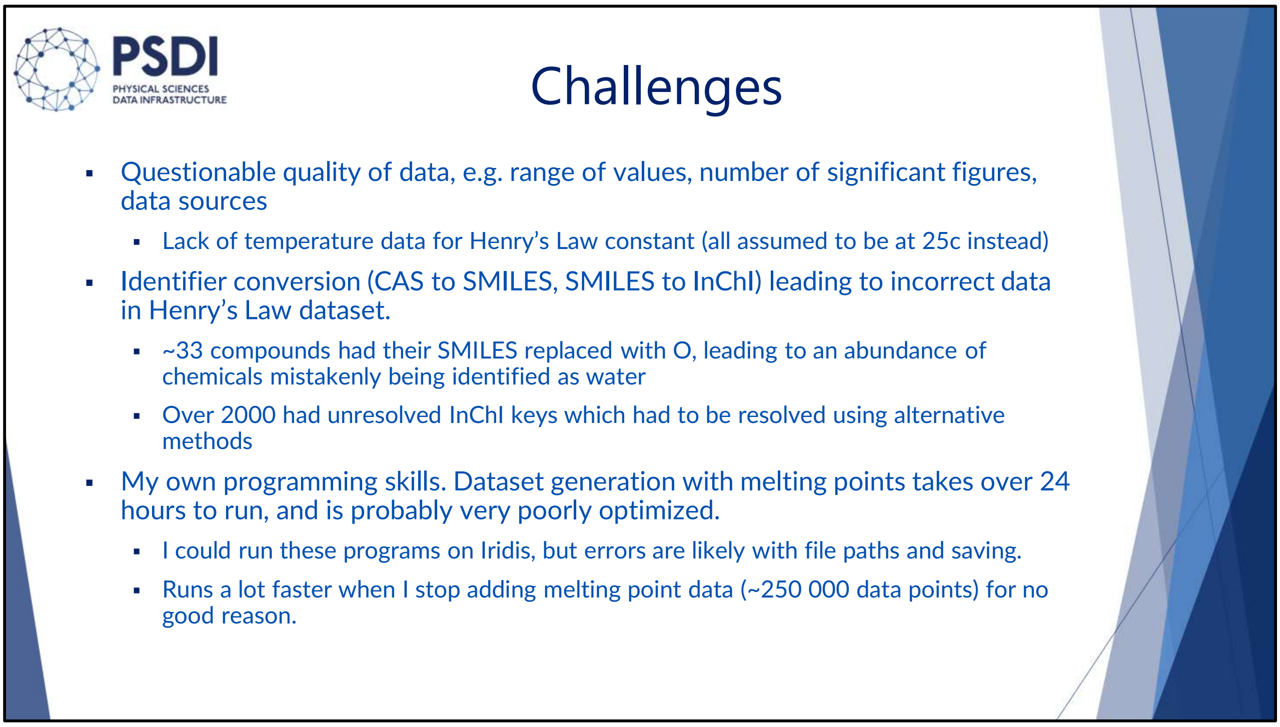


Choosing the Algorithms

log(k_H)

Algorithm	MSE	R2
LightGBM	**5.04**	**0.89**
KRR	5.92	0.87
RandomForest	8.74	0.81
KNN	9.31	0.81
AdaBoost	12.76	0.73
SVR	68.92	-0.45

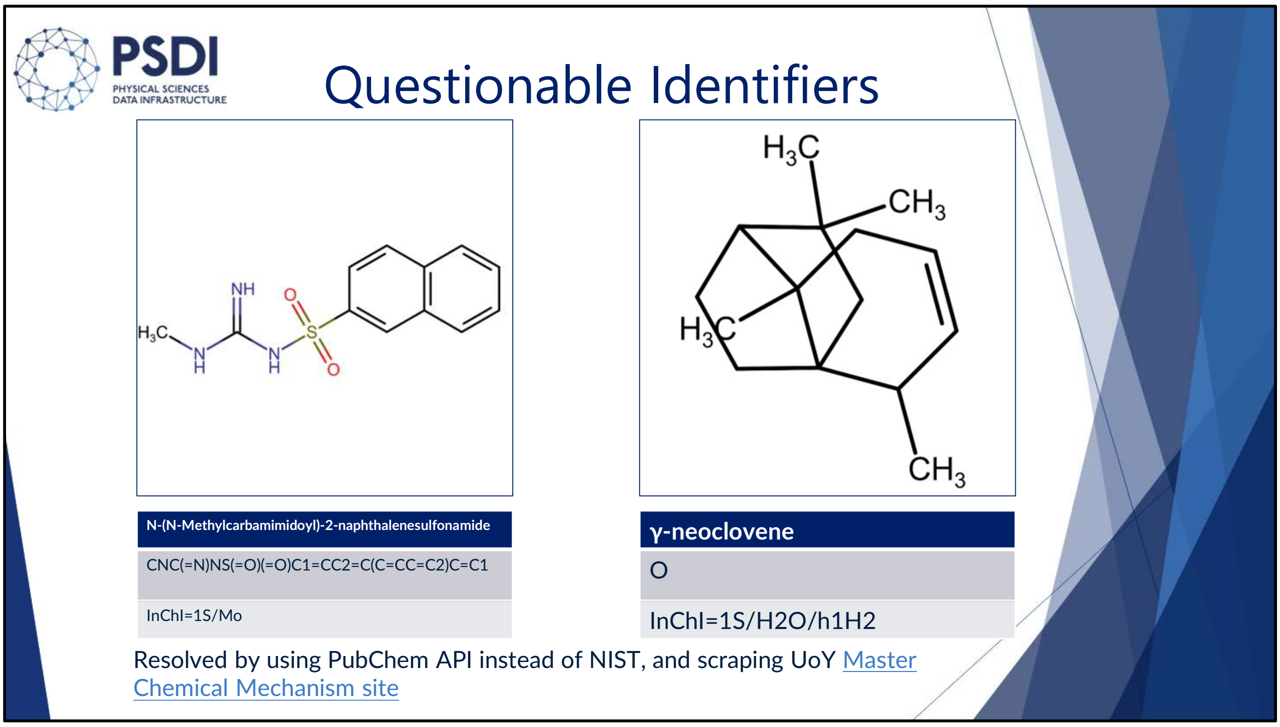
- Henry’s law dataset doubled in size after fixing unresolved identifiers.
- Best algorithm, scaling methods, and hyperparameters likely to have changed.
- Results from previous slide could almost definitely be improved (time limitations in project)



PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

Challenges

- Questionable quality of data, e.g. range of values, number of significant figures, data sources
  - Lack of temperature data for Henry's Law constant (all assumed to be at 25c instead)
- Identifier conversion (CAS to SMILES, SMILES to InChI) leading to incorrect data in Henry's Law dataset.
  - ~33 compounds had their SMILES replaced with O, leading to an abundance of chemicals mistakenly being identified as water
  - Over 2000 had unresolved InChI keys which had to be resolved using alternative methods
- My own programming skills. Dataset generation with melting points takes over 24 hours to run, and is probably very poorly optimized.
  - I could run these programs on Iridis, but errors are likely with file paths and saving.
  - Runs a lot faster when I stop adding melting point data (~250 000 data points) for no good reason.



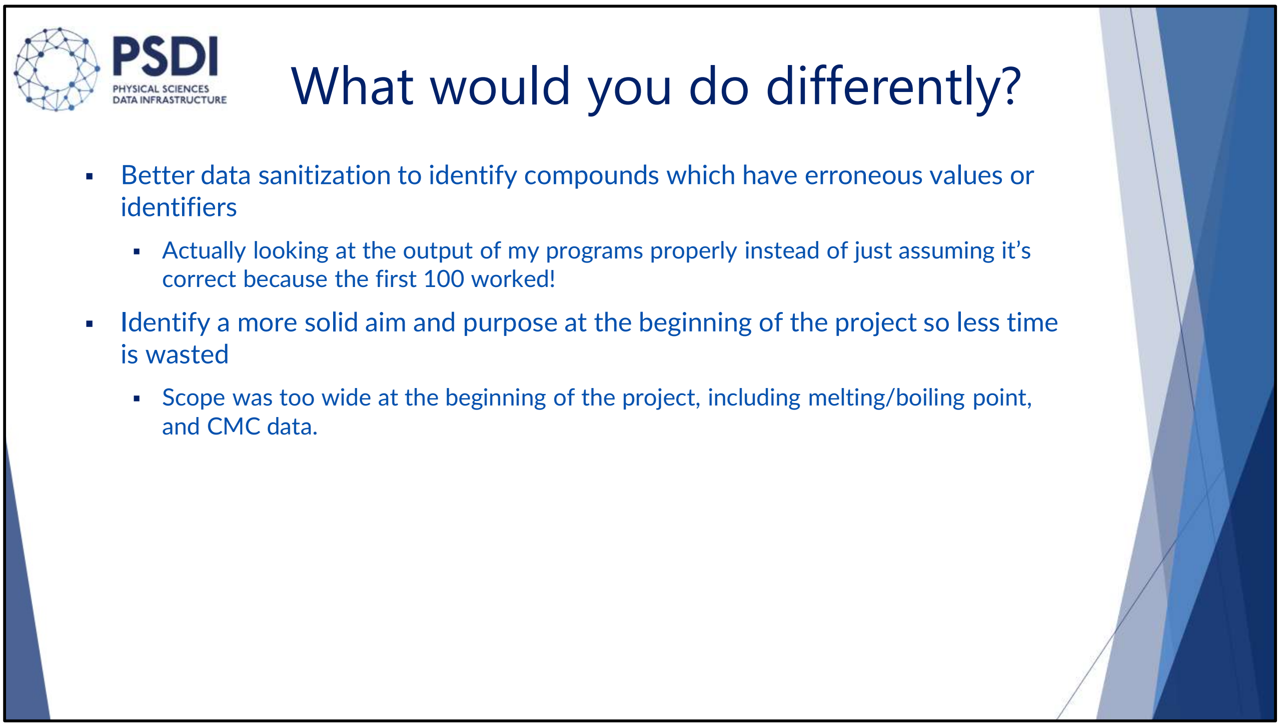
PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Questionable Identifiers

N-(N-Methylcarbamimidoyl)-2-naphthalenesulfonamide
CNC(=N)NS(=O)(=O)C1=CC2=C(C=CC=C2)C=C1
InChI=1S/Mo

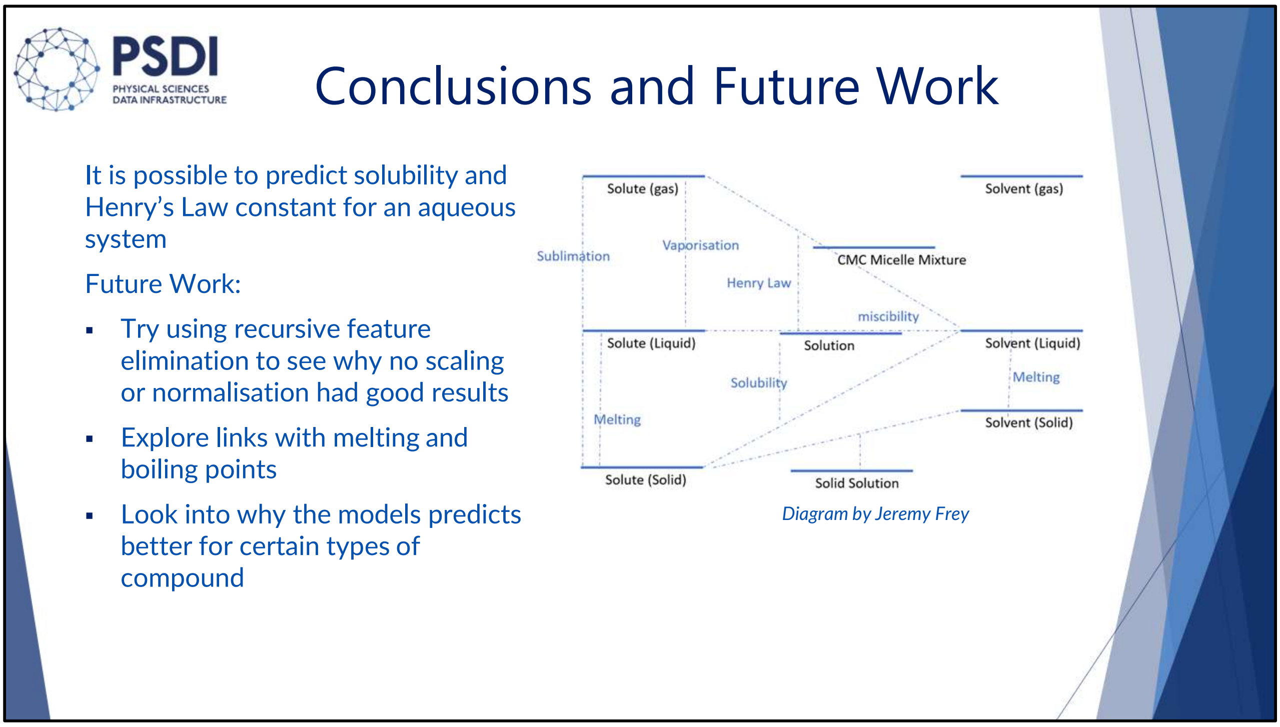
γ-neoclovene
O
InChI=1S/H2O/h1H2

Resolved by using PubChem API instead of NIST, and scraping UoY Master Chemical Mechanism site



What would you do differently?

- Better data sanitization to identify compounds which have erroneous values or identifiers
  - Actually looking at the output of my programs properly instead of just assuming it's correct because the first 100 worked!
- Identify a more solid aim and purpose at the beginning of the project so less time is wasted
  - Scope was too wide at the beginning of the project, including melting/boiling point, and CMC data.



PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

Conclusions and Future Work

It is possible to predict solubility and Henry’s Law constant for an aqueous system

Future Work:

- Try using recursive feature elimination to see why no scaling or normalisation had good results
- Explore links with melting and boiling points
- Look into why the models predicts better for certain types of compound

Diagram by Jeremy Frey