Augmenting the Thermodynamic Oxidation Data of Dual Phase Steels using Synthetic Data

Nicola Beech[1,4], James Edy[2], Didier Farrugia[2], Michael Auinger[1,3]

1 WMG, The University of Warwick, CV4 7AL, United Kingdom.

2 TATA Steel, Research and Development, South Wales Technology Centre, Talbot Building, Swansea University Singleton Campus, Singleton Park, Sketty, Swansea. SA2 8PP.

3 Turing Fellow, The Alan Turing Institute, London, NW1 2DB, United Kingdom.

4 Synthetic Data Turing Intern, The Alan Turing Institute, London, NW1 2DB, United Kingdom.

7th Postgraduate Research Symposium on Ferrous Metallurgy - Tuesday 27th February 2024, London, UK

WMG
THE UNIVERSITY OF WARWICK

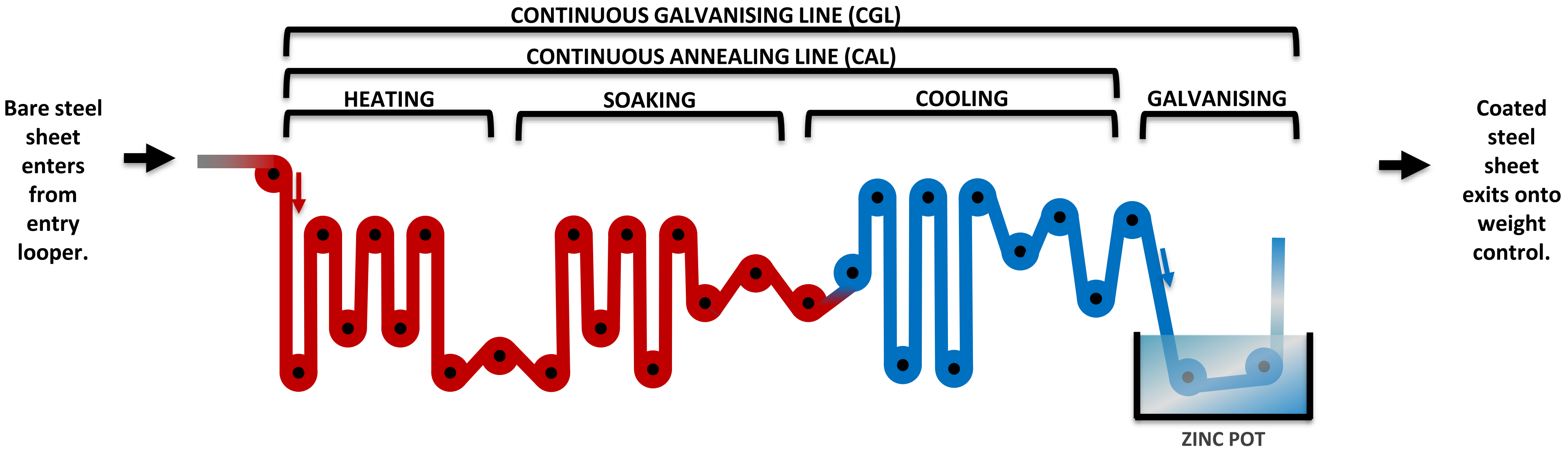
Introduction and Agenda

1. *The industrial problem (the oxidation of DP Steels)*
2. *Diagrams to explore oxidation behaviour*
3. *An introduction to synthetic data*
4. *How machine learning methods can be used to generate synthetic thermodynamic data*
5. *How I studied this methodology*
6. *My results and analysis*
7. *A summary of synthetic thermodynamic data generation: benefits and challenges*

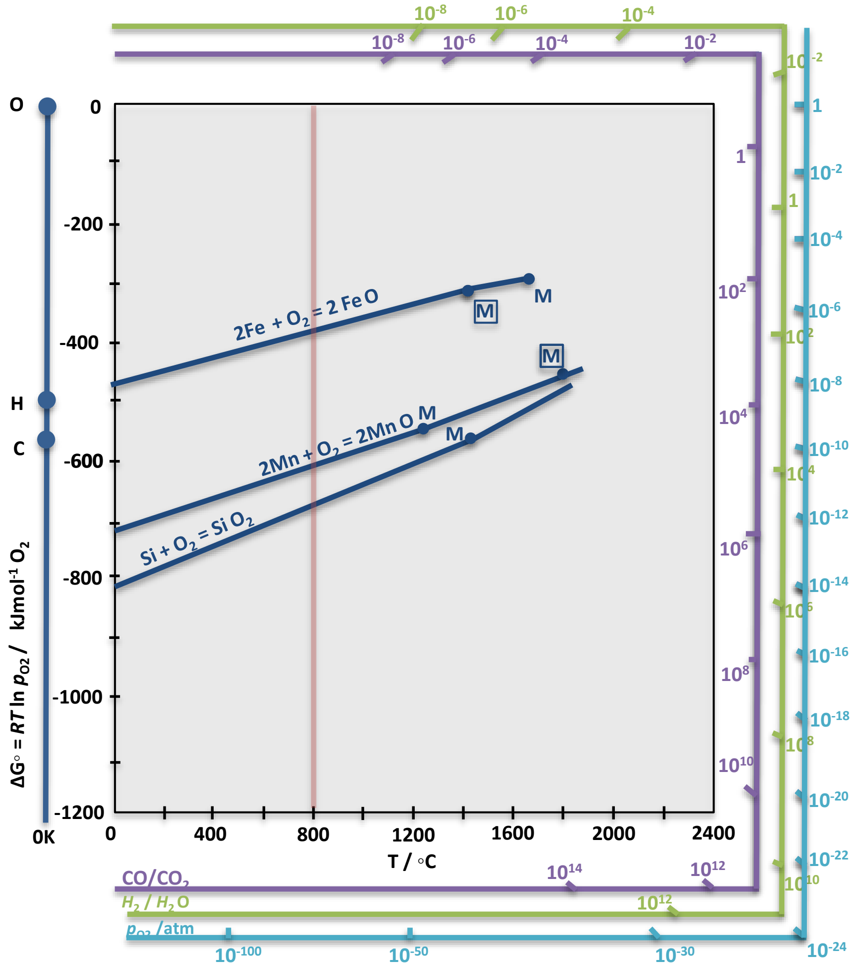
Dual Phase Steel Production

- Dual Phase (DP) steels are widely used in car body structures to reduce vehicle weight because of their excellent strength and formability.

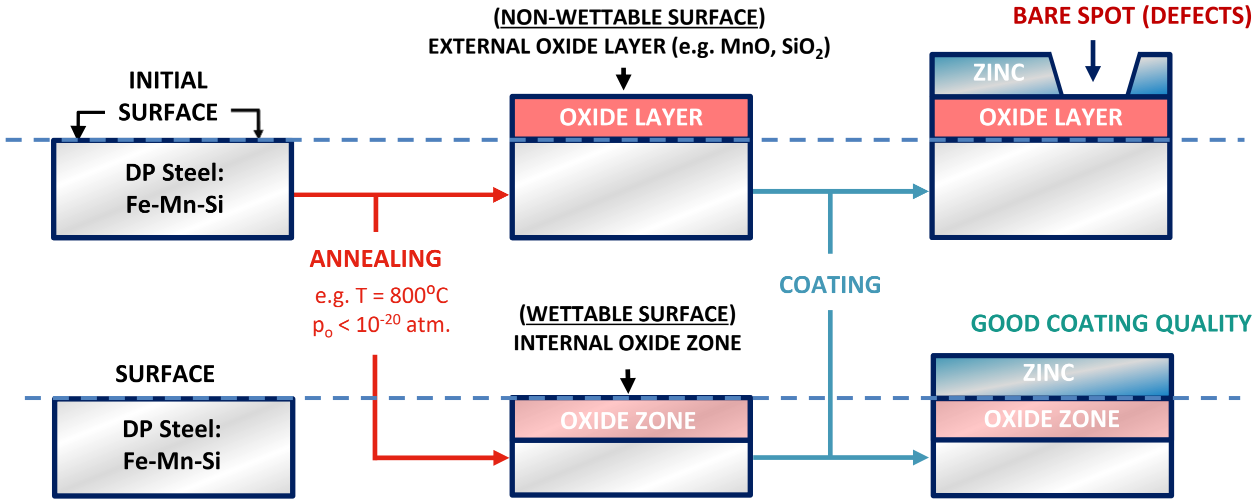
- **DP Steels are typically produced on a CGL/CAL:**


CONTINUOUS GALVANISING LINE (CGL)
CONTINUOUS ANNEALING LINE (CAL)
HEATING
SOAKING
COOLING
GALVANISING
Bare steel sheet enters from entry looper.
Coated steel sheet exits onto weight control.
ZINC POT

The Formation of Oxides



- ➢ The typical alloying elements of DP steels (e.g., Mn and Si) have a greater affinity with oxygen than iron and can selectively oxidise.



INITIAL SURFACE

DP Steel: Fe-Mn-Si

ANNEALING

e.g. T = 800°C

$p_o < 10^{-20}$ atm.

(NON-WETTABLE SURFACE)

EXTERNAL OXIDE LAYER (e.g. MnO, SiO_2)

OXIDE LAYER

(WETTABLE SURFACE)

INTERNAL OXIDE ZONE

OXIDE ZONE

SURFACE

DP Steel: Fe-Mn-Si

COATING

BARE SPOT (DEFECTS)

ZINC

OXIDE LAYER

GOOD COATING QUALITY

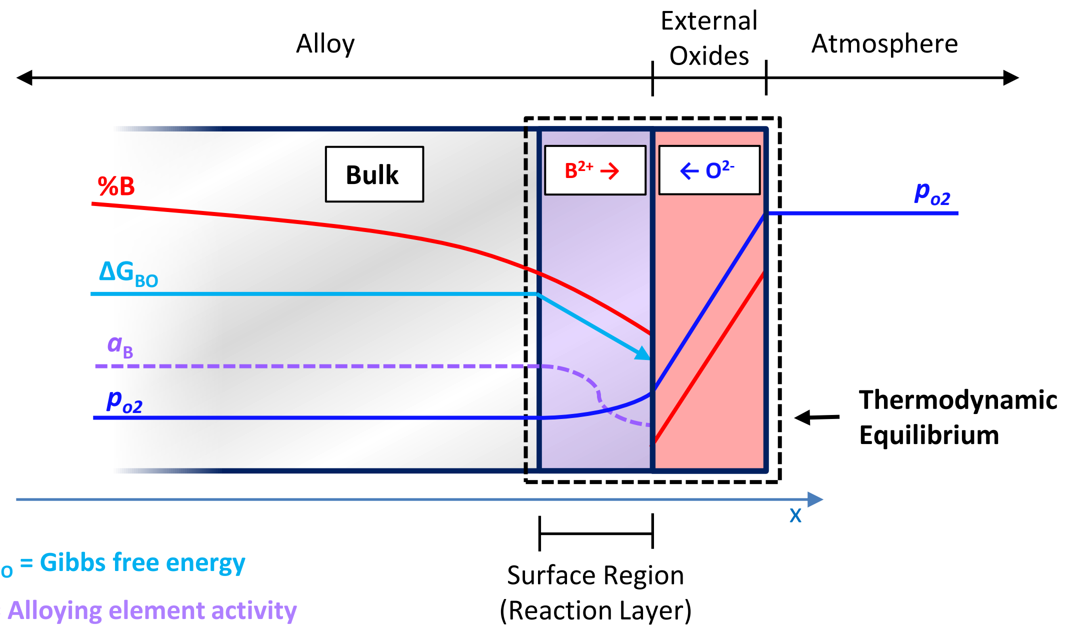
ZINC

OXIDE ZONE

* Diagram adapted from: 1. Hasegawa, M. (2014). Chapter 3.3 - Ellingham Diagram. In: S. Seetharaman, ed., Treatise on Process Metallurgy, 1st ed. Amsterdam: Elsevier, pp.507-516..

Diagrams to Explore Oxidation

- We assume that a local thermodynamic equilibrium exists between the oxides and the surface of the metal.

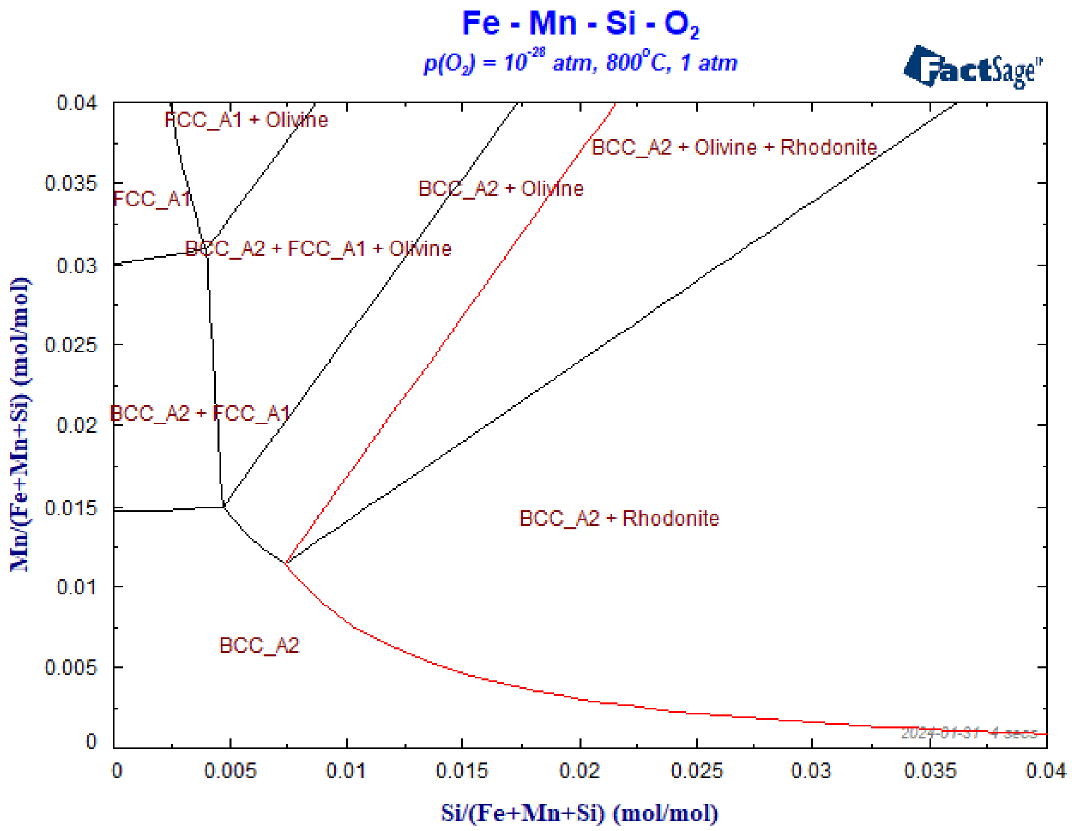


ΔG_{BO} = Gibbs free energy

a_B = Alloying element activity

p_{o2} = Oxygen partial pressure

%B = Alloying element content



* Diagram adapted from: 2. Suzuki, Y., Yamashita, T., Sugimoto, Y., Fujita, S. and Yamaguchi, S. (2009). Thermodynamic Analysis of Selective Oxidation Behaviour of Si and Mn-added Steel during Recrystallization Annealing. ISIJ International, 49(4), pp.564-573..

3. Bale, C.W., Chartrand, P., Degterov, S.A., Eriksson, G., Hack, K., Ben Mahfoud, R., Melançon, J., Pelton, A.D. and Petersen, S. (2002). FactSage thermochemical software and databases. Calphad, 26(2), pp.189-228.

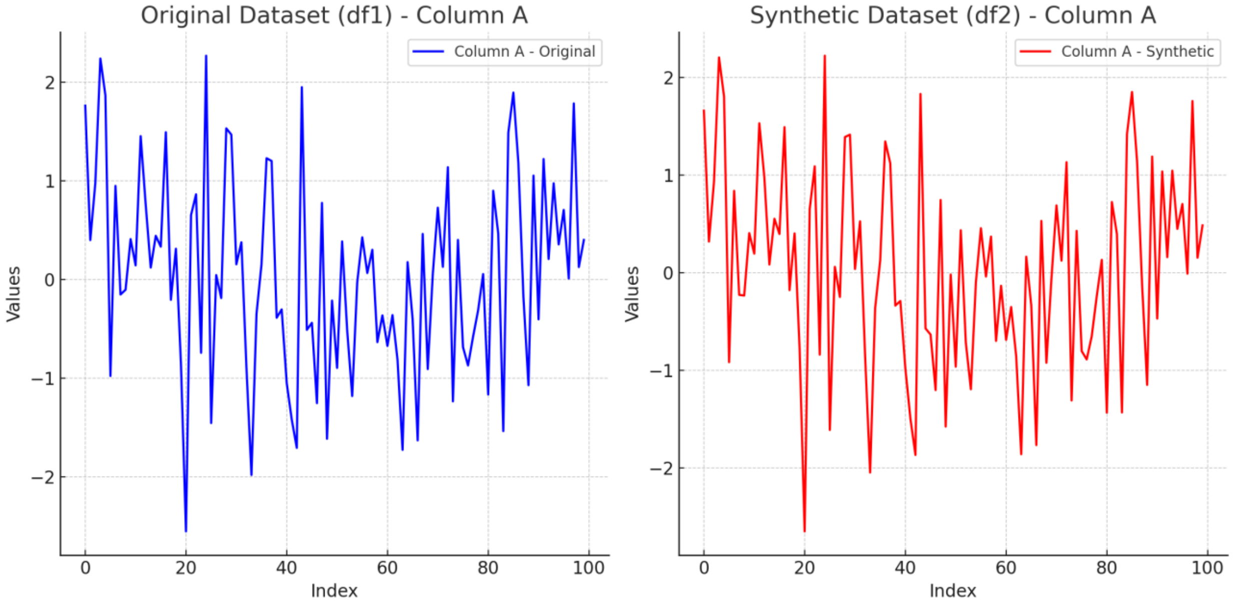
Background on Synthetic Data

- Synthetic data is data generated to statistically replicate a real dataset.

- Why use synthetic data?
  - Preserve privacy
  - Augment small datasets
  - De-biassing

By 2024, 60% of data for AI will be synthetic to simulate reality, future scenarios and derisk AI

~ The Gartner Institute [4]



[4] Gartner. (2023). Gartner Identifies Top Trends Shaping Future of Data Science and Machine Learning [Online]. Available at: https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning [Last Accessed: 05 February 2024].

Synthetic Data Quality Metrics

Utility

The suitability of the synthetic dataset is for the intended purpose.

Examples: precision, accuracy, recall, F1 score, propensity score[5].

Fidelity

The statistical similarity between the synthetic data and the original dataset.

Examples: feature-attribute correlations, Pearson correlation for pairwise correlation[6].

Privacy

The measure of how effective synthetic data is in masking data in the original dataset, such that no specific data can be reconstructed or re-identified.

Examples: Euclidean distances[7], distance to closest records[8].

[5] Dankar, F.K. and Ibrahim, M. (2021). Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. Applied Sciences, 11(5), Article 2158.

[6] Muniz-Terrera, G., Mendelevitch, O., Barnes, R. and Lesh, M.D. (2021). Virtual Cohorts and Synthetic Data in Dementia: An Illustration of Their Potential to Advance Research. Frontiers in Artificial Intelligence, 4.

[7] Wang, S., Rudolph, C., Nepal, S., Grobler, M. and Chen, S. (2020). PART-GAN: Privacy-Preserving Time-Series Sharing. In: I. Farkaš, P. Masulli and S. Wermter, eds., Artificial Neural Networks and Machine Learning -- ICANN 2020. Cham: Springer International Publishing, pp.578-593.

[8] Platzer, M. and Reutterer, T. (2021). Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data. Frontiers in Big Data, 4.

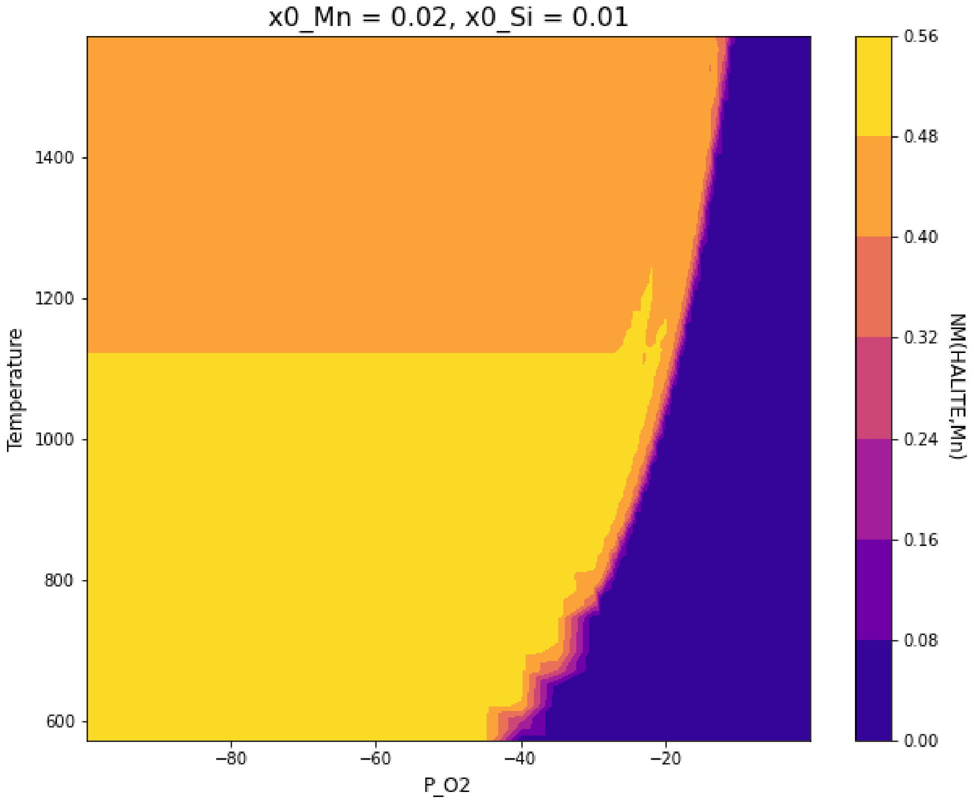
Data Generation Techniques

Synthetic Data Generation Technique	Description	Benefits	Limitations
Stochastic Methods	Data is randomly generated.	Easy to implement. Good for privacy. Minimal use of computational resources.	Features are not retained.
Rule-Based	Data is calculated using laws (e.g., thermodynamic laws).	Generates data with structure.	Computational resources depend on complexity of laws. Laws must be translated into suitable data formats/languages.
Deep Generative Models	Machine learning models learn the statistical distribution of real data to generate synthetic data.	Retains data structure and can handle complex datasets.	Limited by computational resources. Challenged in handling outliers/inherited bias.

Methodology for Generating Synthetic Thermodynamic Data

Q. How few rows of thermodynamic data are needed to generate a high-fidelity synthetic dataset?

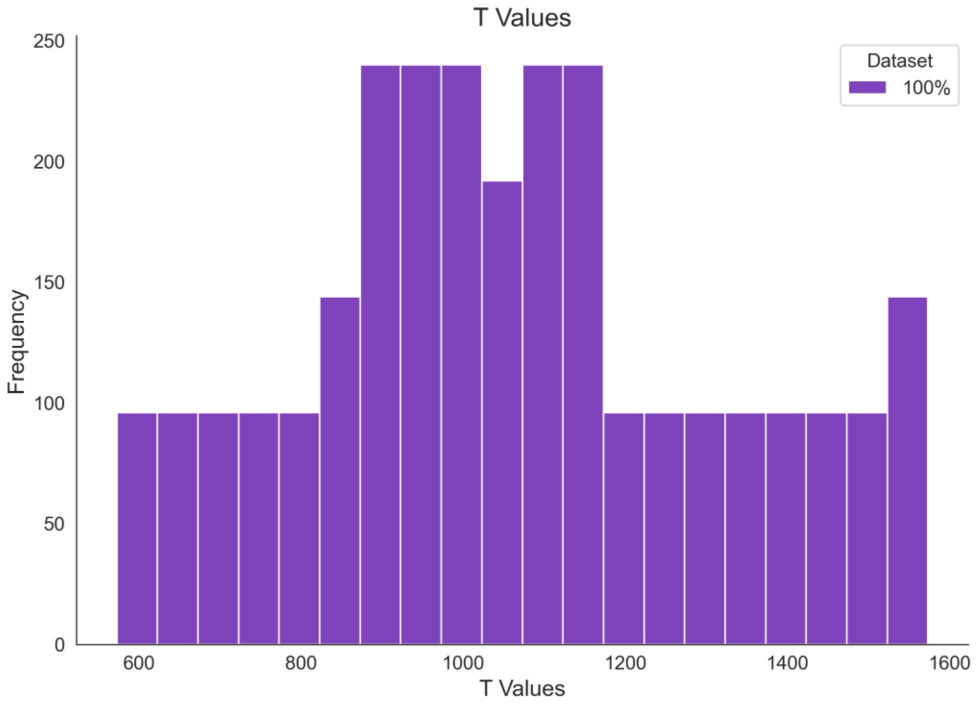
- Initial Data[8] = x0_Mn, x0_Si, T, P_O2, NM(HALITE,Mn) Training on decreasing percentages of sampled rows from the initial data. (100%, 50%, 25%, 10%)
- Comparison of statistical distribution of datasets (KSComplement score).

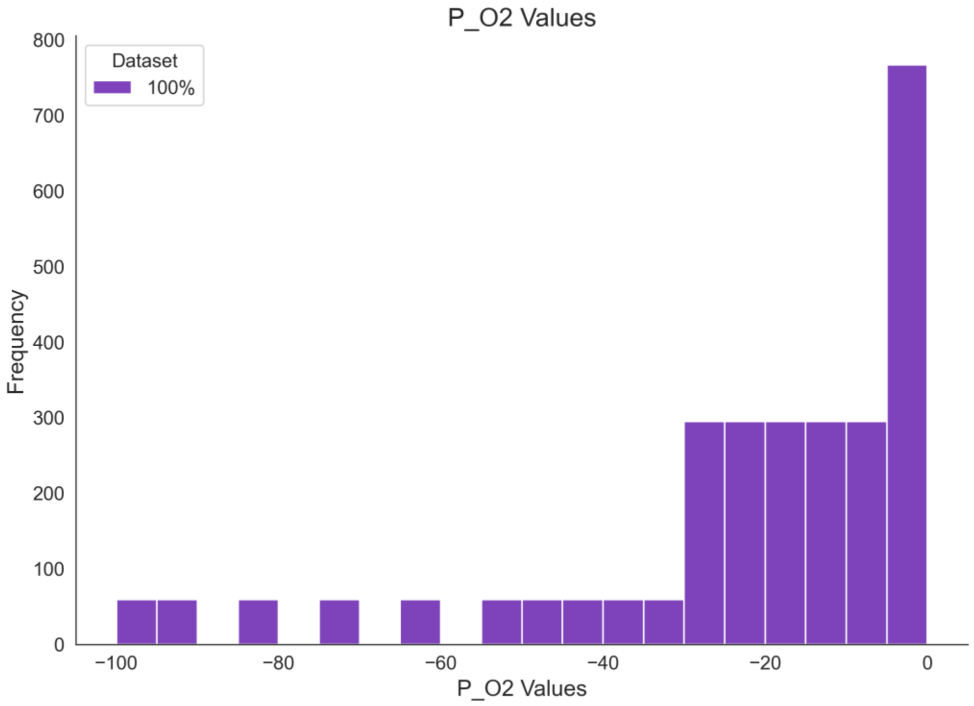


[8] Thermo-Calc Software TCFE13 Steels/Fe-alloys Database, https://thermocalc.com/products/databases/steel-and-fe-alloys/ (accessed 31 July 2023)

Thermodynamic Data

- Shape of the DataFrame: (2832, 5)

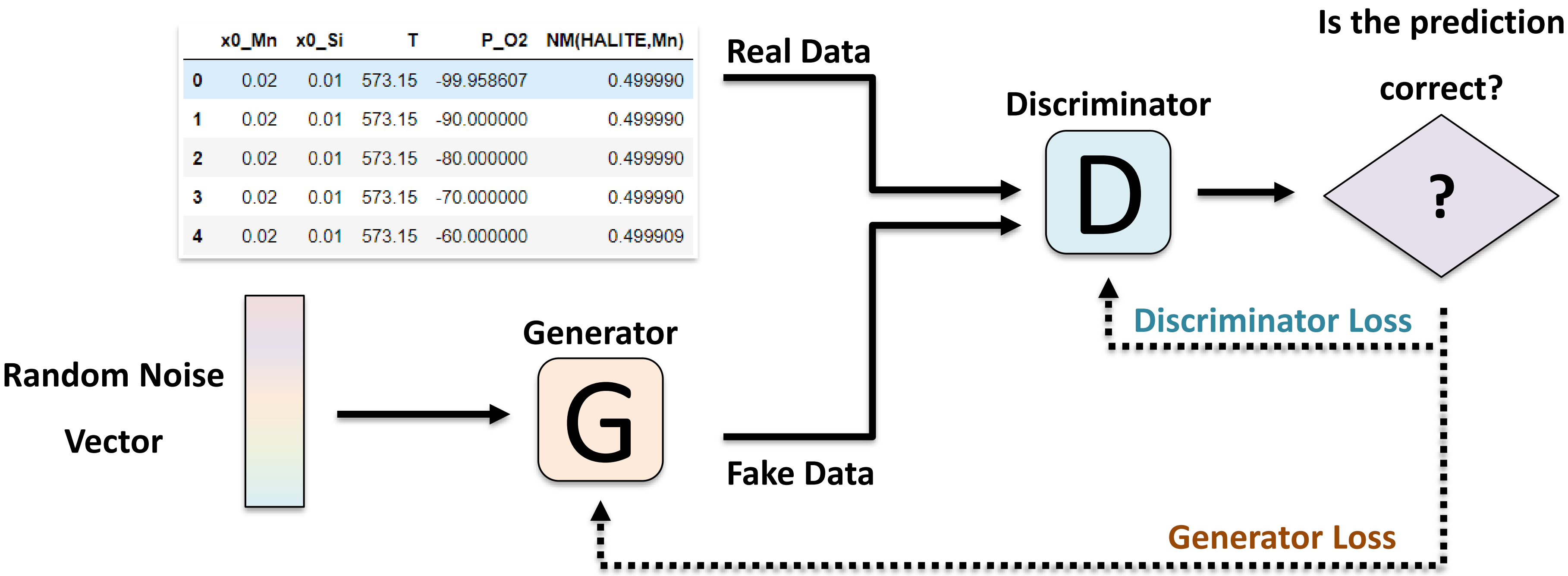




- Temperature and oxygen partial pressure values are irregular in the dataset, which can lead to bias.

Generative Adversarial Networks

- GANs comprise:
  - Generator neural network – receives noise and outputs fake data.
  - Discriminator – receives real and fake data and predicts which is real and which is fake.



	x0_Mn	x0_Si	T	P_O2	NM(HALITE,Mn)
0	0.02	0.01	573.15	-99.958607	0.499990
1	0.02	0.01	573.15	-90.000000	0.499990
2	0.02	0.01	573.15	-80.000000	0.499990
3	0.02	0.01	573.15	-70.000000	0.499990
4	0.02	0.01	573.15	-60.000000	0.499909

SDV[9] CTGAN[10] Hyperparameters

- Epochs: range(100, 1000, 100)
- Batch Size: 500
- Log Frequency: True, False
- Learning Rate: [2e-1, 2e-2, 1e-3, 1e-5]
- Decay Rate: [1e-3, 1e-4, 1e-5, 1e-7]

*Sampling: 2*len(real_data)*

[9] Patki, N., Wedge, R. and Veeramachaneni, K. (2016). The Synthetic data vault. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), Oct 2016, pp.399-410.

[10] Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. In: Advances in Neural Information Processing Systems.

Fidelity Scores (Random Sample)

- The KSComplement metric uses the Kolmogorov-Smirnov statistic, to compare cumulative distribution functions (CDFs) of numerical distributions.

- **A score of 1.0 indicates perfect similarity, while 0.0 indicates maximum dissimilarity.**

$$KSComplement\ Score = 1 - KS\ Statistic$$

$$KS\ Statistic = max\left|CDF_R - CDF_S\right|$$

(Where CDF_R and CDF_S represent real data and synthetic data, respectively.)

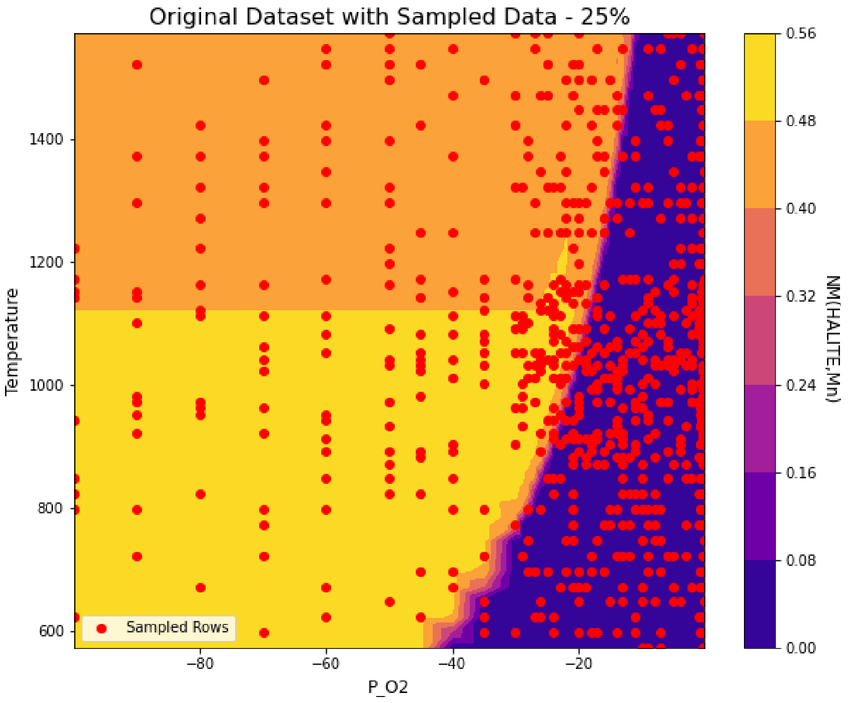
Best KSComplement scores:

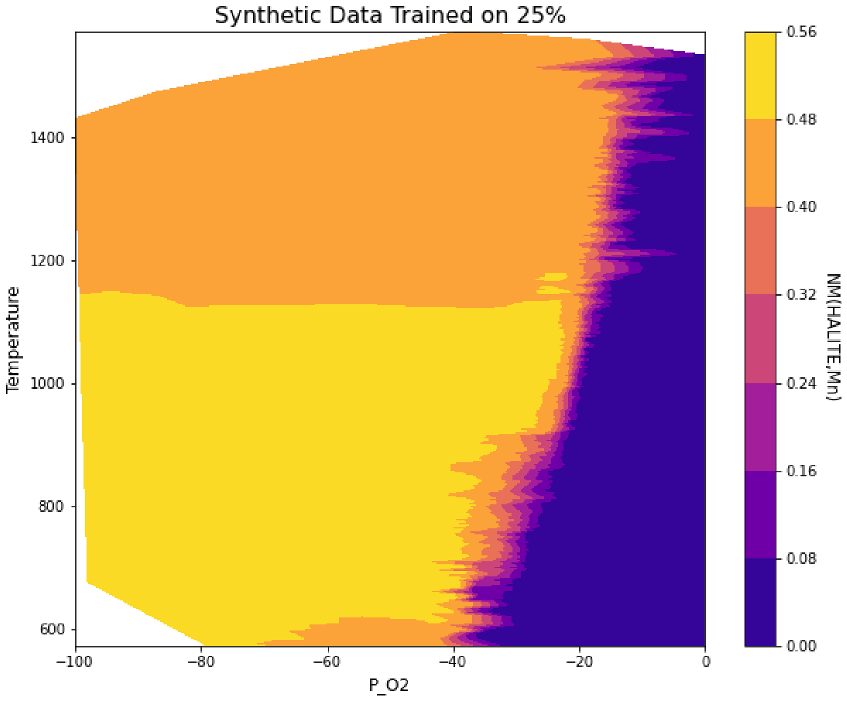
	Rows	KSComplement Score[11]
100%	2832	0.975989
50%	1416	0.962394
25%	708	0.966102
10%	283	0.939266

- The KSComplement scores indicate that our synthetic data has high fidelity.

[11] DataCebo, Inc. (2023). Synthetic Data Metrics. [Online] Version 0.12.0. Available at: https://docs.sdv.dev/sdmetrics/ [Last Accessed: 05 February 2024].

Comparison of Synthetic Thermodynamic Data and Original Data (Random Sample)

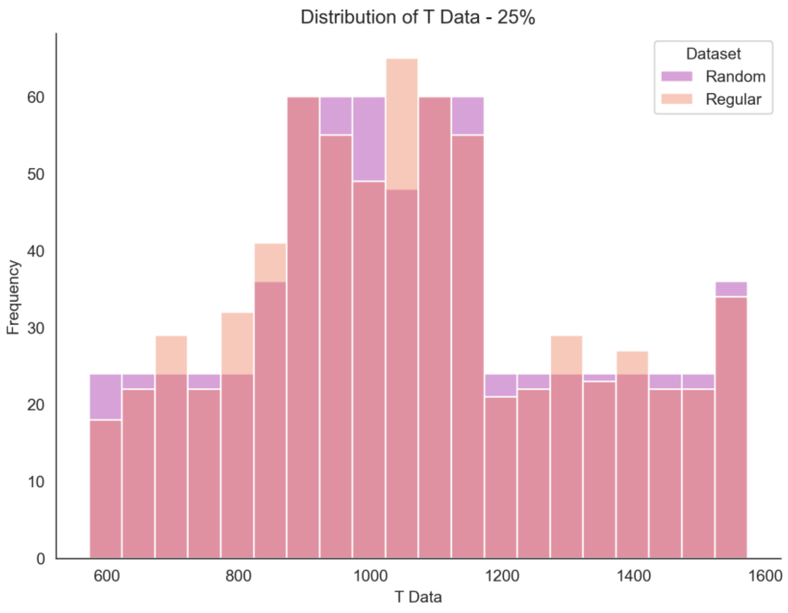




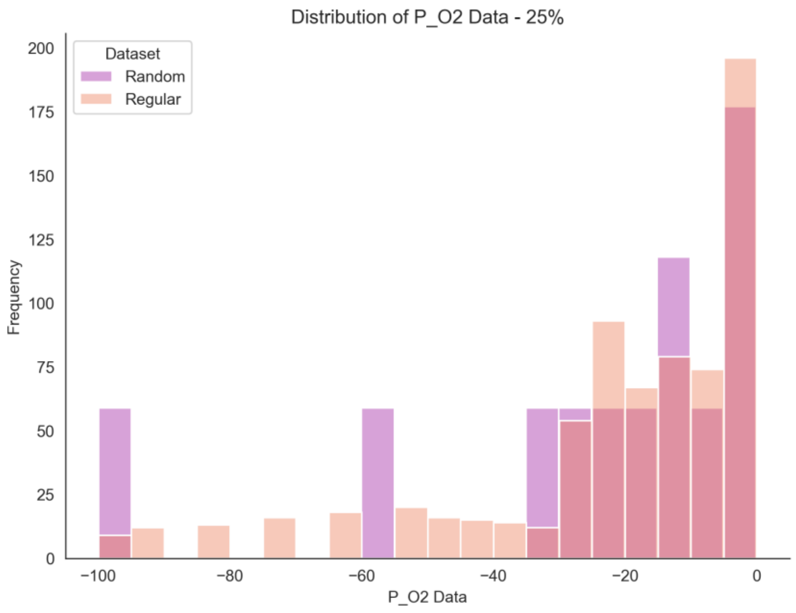
- The synthetic data can inherit bias from the irregular distribution of data in the original dataset, as well as the random sample distribution.

Regular Sample Training Data

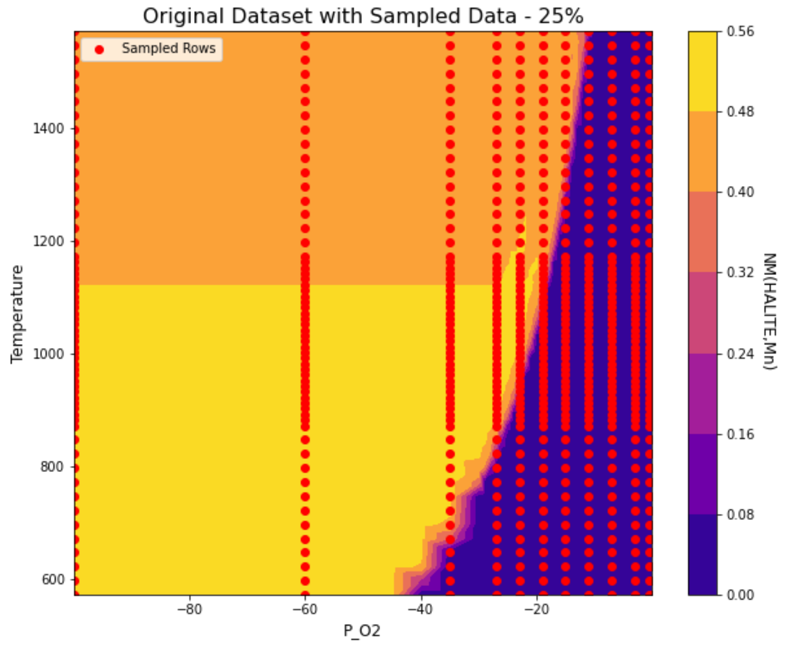
➢ To avoid inheriting bias from the irregularity of the input data, the training data was sampled at regular intervals.



Distribution of T Data - 25%



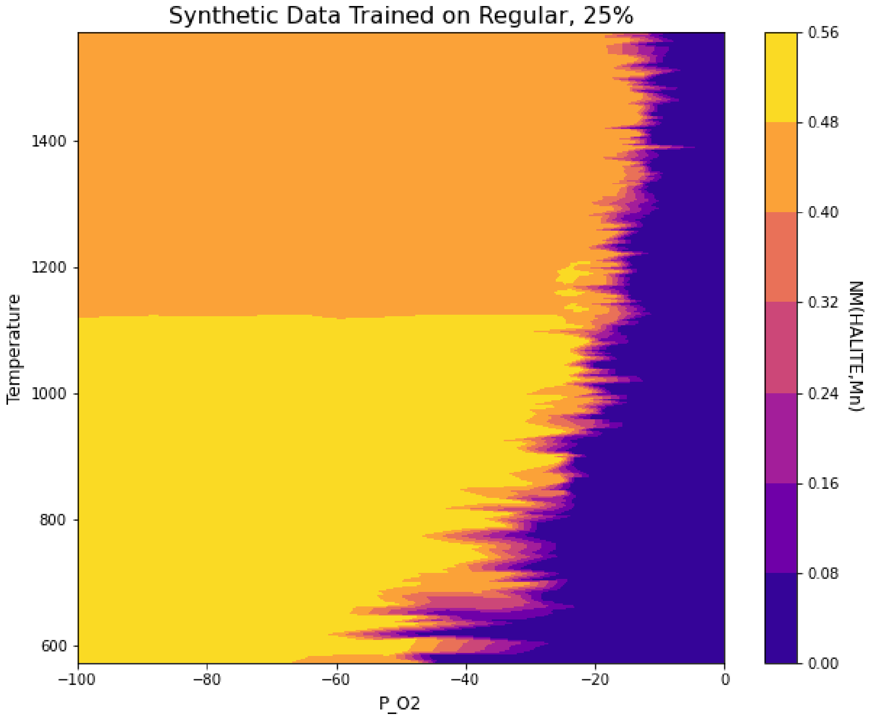
Distribution of P_O2 Data - 25%

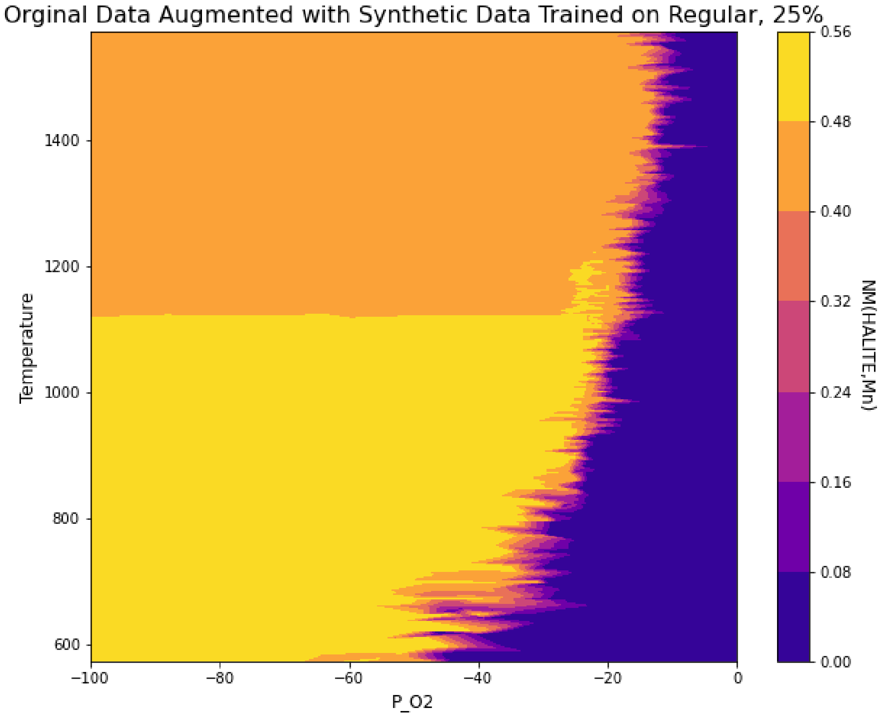


Original Dataset with Sampled Data - 25%

- Rows were sampled at 25% and 10%.

Visual Comparison and Fidelity (Regular Sample)





- KSComplement scores indicate high fidelity (25% = 0.961511, 10% = 0.945092).
- Rows sampled at regular intervals have reduced distances between triangulated points within several ranges.

Closing Summary

- To study the value of synthetic data in augmenting thermodynamic datasets, a CTGAN algorithm was trained on reducing numbers of rows from a Thermocalc dataset.
- The synthetic data achieved high fidelity scores, however inherited bias.
- Recommendations for reducing bias: pre-process training data, balance distribution, use optimal hyperparameters and consider evaluation metrics.

Benefits

- With a limited quantity of data, it is possible to generate high fidelity synthetic thermodynamic data.
- Synthetic thermodynamic data can augment small datasets.
- The synthetic dataset could be used to mask private industrial datasets.

Limitations

- Inherited bias, noise, models require initial dataset.