SuperconGAN: A Library to Generate and Predict Superconductive Materials

Rajeev Atla
rajeev@rajeevatla.com

Abstract

TODO

1 Introduction

Achieving a room-temperature superconductor is considered the hallmark of condensed matter physics. However, researchers have faced many challenges in their search for such a material, including long, costly production times and sensitivity toward environmental conditions [1] [2]. Although the advent of novel superconductors such as Van der Waals heterostructures, including twisted bilayer graphene (TBG) [6], has lessened this, searching through the entire phase diagrams of even these materials would likely take decades without a way to accelerate the process of discovery. One possible way to solve this problem is to use a data-based method, using previous, already discovered superconductors to make inferences about potential ones. It would be of great assistance to researchers in the field to be able to differentiate between non-superconductors and superconductors through a data-based method utilizing already known superconductors. Recent work [7] has used easily observable characteristics to predict critical temperature. However, no clear association has been found yet between the easily observable characteristics of a material and its superconductive phase diagram, so a supervised learning approach isn't realizable. Herein, we attempt a solution to this problem: an unsupervised generative adversarial network (GAN) that allows researchers to evaluate materials to further investigate for possible supercondctivity. Further, this GAN only uses easily observable characteristics such as valence, atomic radius, etc., allowing researchers to determine whether a material can exhibit supercondctivity. Such a method would also be able to generate possible superconductive materials to be tested, further pushing the pace of existing research by perpetually pointing researchers towards potentially new directions.

2 Methods

To obtain the data, we use the Supercon database [7], obtained from the UCI machine learning repository [5]. The dataset contains 81 features extracted from 21263 superconductors, along with their chemical formulae and critical temperatures. In order to generate and evaluate superconductive materials, we remove the critical temperatures from consideration, keeping only easily physical characteristics such as thermal conductivity, atomic radius, electron affinity, and atomic mass.

We then train a GAN upon this cleaned dataset. Specifically, we use the Conditional Tabular GAN (CTGAN) library [8] to generate this data, for its ability to also generate conditional data. SuperconGAN then functions as a wrapper around CTGAN, allowing researchers to easily synthesize data without delving deeply into CTGAN's syntax.

A GAN works [3] by generating data from a random noise vector. One of the two GAN components, the generator, creates a map between this random noise vector and a set of fake data. The other GAN component, the discriminator, tries to tell the difference between this fake data and the real data, and sends feedback to the generator. This generator uses this feedback to update its mapping to try and fool the discriminator. This process results in the production of data that closely mimics the real data.

To evaluate the quality of the data produced independently of the GAN, we use the single table evaluation function provided by the Synthetic Data Vault (SDV) [4]: SDMetrics. This function compares the results of several machine learning models (see A) between the synthetic and real data.

To allow researchers to evaluate the GAN's training before it finishes, SuperconGAN allows the display of generator and discriminator losses. The discriminator loss is defined as the difference between the fake data and the real data. The generator loss is the difference between the cross entropy between the two datasets and the mean of the fake data.

3 Results

We present the generator and discriminator losses in Figure 1.

Figure 1: The generator and discriminator losses plotted together, after being averaged across $n = 3$ trials. Due to the nature of the noise vector used to train the generator, the losses also reflect this noise.

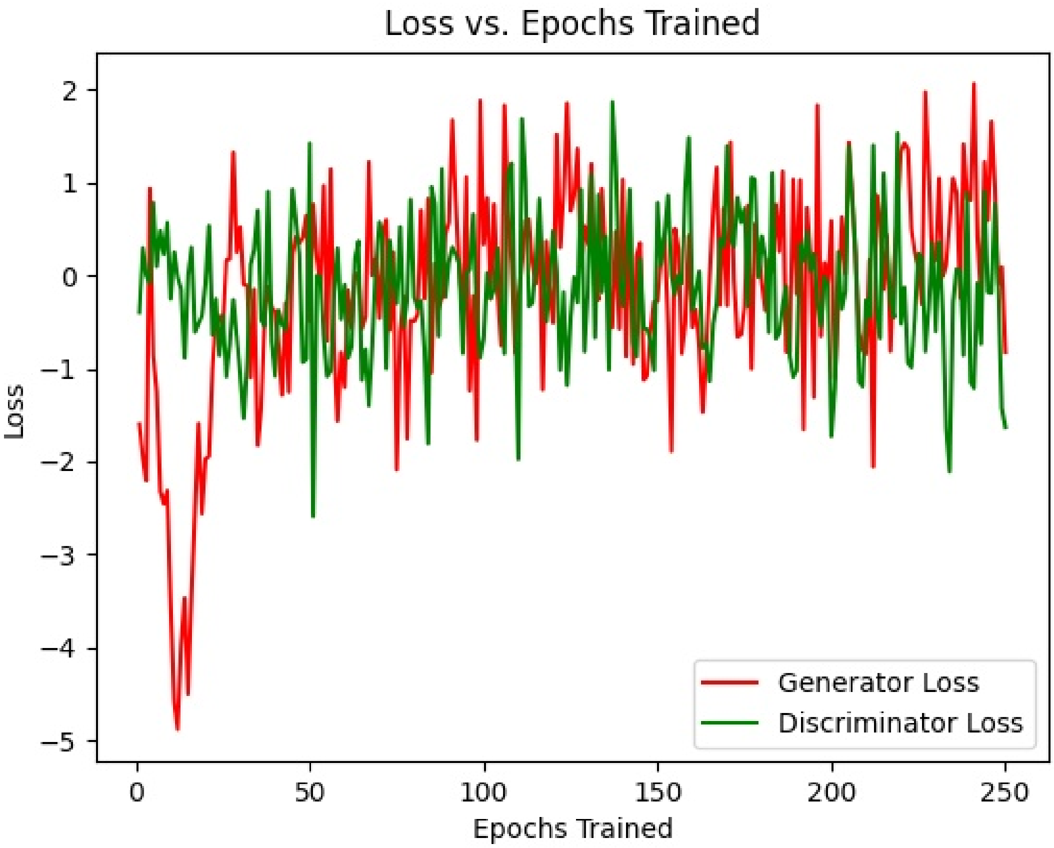
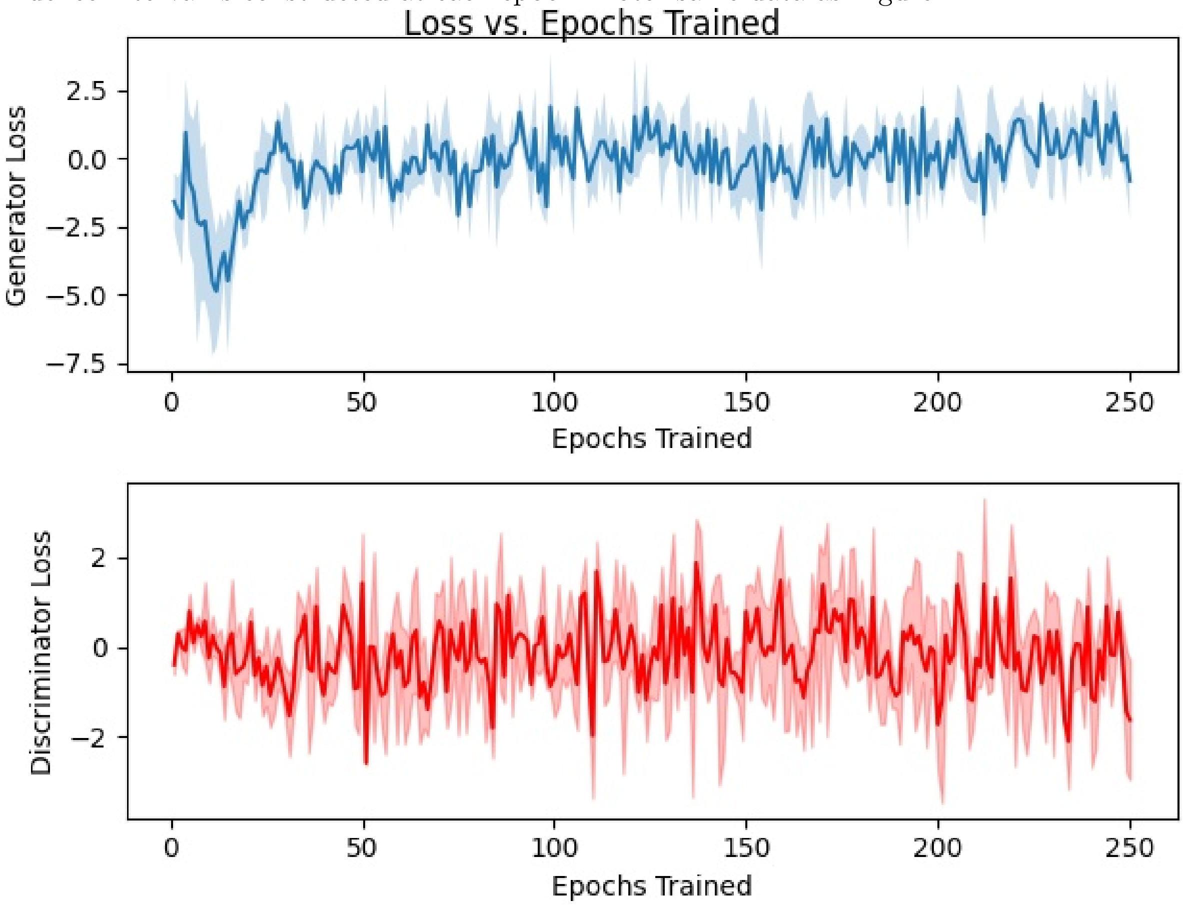
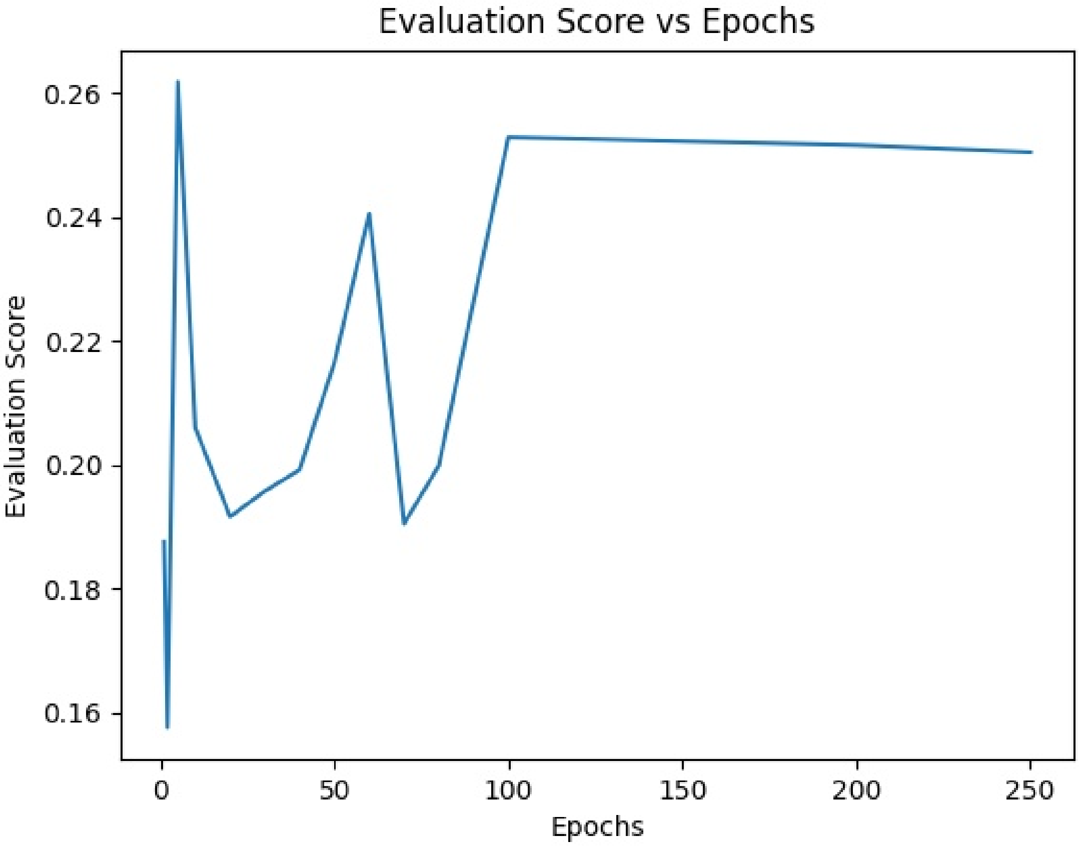


Figure 2: The generator and discriminator losses are averaged across $n = 3$ trials and a 95% confidence interval is constructed at each epoch. Note: same data as Figure 1.



We present the results of the evaluation function, as a function of the number of epochs the GAN is trained, in Figure 3.

Figure 3: The evaluation function is significantly less noisy than the losses plotted. Note the diminishing marginal returns exhibited.



4 Conclusion

In this work, we discuss a novel GAN-based approach towards the study of superconductor data. This approach would allow researchers to utilize existing data-based approaches more effectively, as one of the largest bottlenecks to a machine learning problem is the availability of data. Existing methods of calculating the critical temperature of a superconductor could be augmented by the data created by SuperconGAN, allowing more accurate estimates. Further work could also map the parameter space, allowing researchers to see what physical characteristics are really necessary to exhibit superconductivity. Another possible research direction could be differentiating between type I and type II superconductors, and further exploring the physical characteristics of each.

This work could also be combined with existing deep learning models [9] that utilize the phase space to make predictions, amplifying their ability by also supplying data about the physical characteristics of a material.

5 Acknowledgements

I would like to thank Leo Lo and Dr. Serena McCalla for all their support through iResearch Institute 2020.

6 References

[1] D. H. A. Blank, H. Kruidhof, and J. Flokstra, "Preparation of $YBa_2Cu_3O_{7-\delta}$ by citrate synthesis and pyrolysis", Journal of Physics D: Applied Physics **21**, 226–227 (1988).

[2] M. Zaki, M. Saleem, and M. S. Anwar, "Synthesis of high temperature superconductor using citrate pyrolysis and observing the meissner effect", PhysLab (2013).

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial networks*, 2014.

[4] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault", 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), `10.1109/dsaa.2016.49` (2016).

[5] D. Dua and C. Graff, *UCI machine learning repository*, 2017.

[6] Y. Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, and P. Jarillo-Herrero, "Unconventional superconductivity in magic-angle graphene superlattices", Nature **556**, 43–50 (2018).

[7] K. Hamidieh, "A data-driven statistical model for predicting the critical temperature of a superconductor", Computational Materials Science **154**, 346–354 (2018).

[8] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN", CoRR **abs/1907.00503** (2019).

[9] B. Han, Y. Lin, Y. Yang, N. Mao, W. Li, H. Wang, K. Yasuda, X. Wang, V. Fatemi, L. Zhou, J. I.-J. Wang, Q. Ma, Y. Cao, D. Rodan-Legrain, Y.-Q. Bie, E. Navarro-Moratalla, D. Klein, D. MacNeill, S. Wu, H. Kitadai, X. Ling, P. Jarillo-Herrero, J. Kong, J. Yin, and T. Palacios, "Deep-learning-enabled fast optical identification and characterization of 2d materials", Advanced Materials **32**, 2000953 (2020).

A Appendix: Evaluation Metrics

We use the evaluate function of SDMetrics [4], which uses these metrics to evaluate the quality of the data produced.

Further, SDmetrics takes the metrics whose values don't lie on $[-1, 1]$ and normalizes them using a hyperbolic tangent. The average of the normalized metric is returned.

Table 1: **Error Metrics**

Metric	**Return Value**
Bayesian Network Likelihood	Average likelihood of of synthetic data using Bayesian Network
Bayesian Network Log Likelihood	Average log likelihood of of synthetic data using Bayesian Network
Gaussian Mixture Log Likelihood	Average log likelihood of of synthetic data using Gaussian Mixture
Logistic Detection	One minus average ROC AUC score from logistic regression
SVC Detection	One minus average ROC AUC score from support vector classification
CS Test	Chi-squared test test statistic
KS Test	Kolmogorov-Smirnov test statistic
Continouus KL Divergence	Kullback-Leibler divergence