Discovering Chemically Novel, High-Temperature Superconductors

Colton C. Seegmiller*, Sterling G. Baird†, Hasan M. Sayeed‡, and Taylor D. Sparks§¶
*Department of Engineering, Utah Valley University
Orem, UT 84058, USA
Email: 10803013@uvu.edu
†Department of Materials Science and Engineering, University of Utah
Salt Lake City, UT 84108, USA
Email: sterling.baird@utah.edu
‡Department of Materials Science and Engineering, University of Utah
Salt Lake City, UT 84108, USA
Email: hasan.sayeed@utah.edu
§Department of Materials Science and Engineering, University of Utah
Salt Lake City, UT 84108, USA
Email: sparks@eng.utah.edu
¶Chemistry Department, University of Liverpool
Liverpool L7 3NY, United Kingdom

***Abstract*—One of the biggest unsolved problems in condensed matter physics is what mechanism causes high-temperature superconductivity and if there is a material that can exhibit superconductivity at both room temperature and atmospheric pressure. Among the many important properties of a superconductor, the critical temperature (T_c) or transition temperature is the point at which a material transitions into a superconductive state. In this implementation, machine learning is used to predict the critical temperatures of chemically unique compounds in an attempt to identify new chemically novel, high-temperature superconductors. The training data set (SuperCon) consists of known superconductors and their critical temperatures, and the testing data set (NOMAD) consists of around 700,000 novel chemical formulae. The chemical formulae in these data sets are first passed through a collection of rapid screening tools, `SMACT`, to check for chemical validity. Next, the `DiSCoVeR` algorithm is used to train on the SuperCon data to form a model, and then screens through batches of the formulae in the NOMAD data set. Having a combination of a chemical distance metric, density-aware dimensionality reduction, clustering, and a regression model, the `DiSCoVeR` algorithm serves as a tool to identify and assess these superconducting compositions [1]. This research and implementation resulted in the screening of chemically novel compositions exhibiting critical temperatures upwards of 150 K, which correlates to superconductors in the cuprate class. This implementation demonstrates a process of performing machine learning-assisted superconductor screening (while exploring chemically distinct spaces) which can be utilized in the materials discovery process.**

***Index Terms*—machine learning, materials informatics, high-temperature superconductor, critical temperature, transition temperature, cuprates**

I. Introduction

Superconductivity has been a major focus in research since its discovery in 1911 [2]. The discovery of a material that exhibits superconductivity at operating temperatures above 273 K and at atmospheric pressure (101 kPa) would have an enormous technological impact. It would absolutely revolutionize the fields of digital electronics and the electric power industry. For many years, all known superconductors were thought to exist within the bounds of Bardeen-Cooper-Schrieffer (BCS) theory, which stated that the superconductivity of materials could not exist above temperatures of 30 K [3]. It wasn't until 1986 when Johannes G. Bednorz and Karl A. Müller discovered a new class of superconductor in the cuprate family that exceeded this BSC theory threshold. [4]. As explained in [5], "the superconducting cuprates are very different from conventional superconductors, in the fact that they are not traditional metals, but instead doped oxides that behave like bad metals. Often, the pairing for superconduction does not happen with electrons, but instead with the doped holes – which act as quasiparticles that pair up and behave like the Cooper pairs, but with opposite charge. It is still not fully known what drives the pairing mechanism to get superconductivity in these materials." Materials with these properties are deemed in the category of a type-II superconductor. Other types of superconductors have since been discovered beyond cuprates alone such as heavy-fermion-based, buckminsterfullerine-based, carbon-allotrope, iron-pnictogen-based, nickel-based, and strontium-ruthenate superconductors among others.

It was also a cuprate that was discovered with a critical temperature above the boiling point of liquid nitrogen (77 K). This led to the realization that applications of superconductivity were looking more realistic and feasible in the near future [6]. Superconductors with a critical temperature above the boiling point of liquid nitrogen are called high-temperature superconductors. It is important to note that all high-temperature superconductors are type-II superconductors. To date, cuprate superconductors hold the record for the

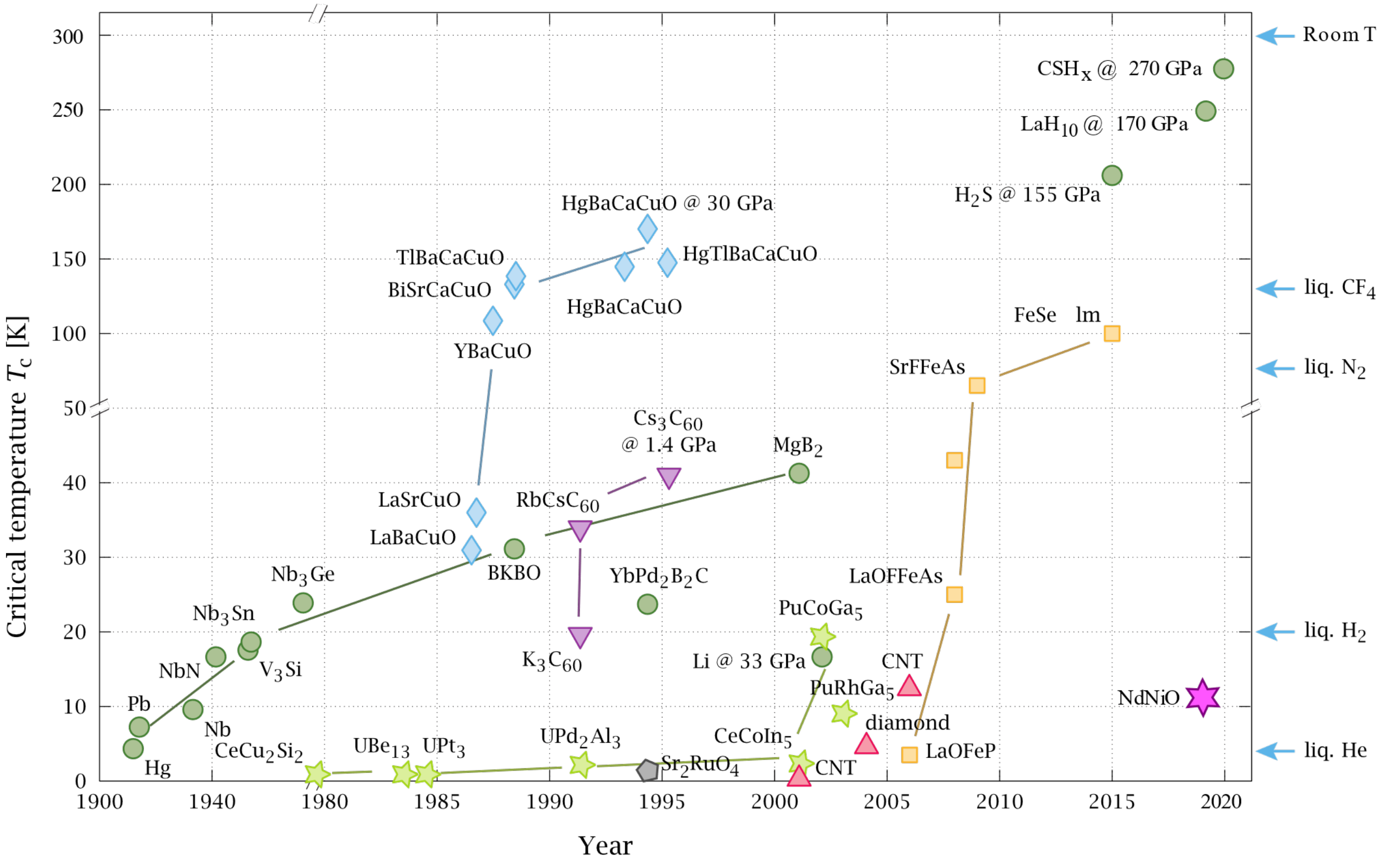

Fig. 1. Timeline of superconductors as adapted from [5], later modified. Note the change in axes around 1980 and 50 K

highest critical temperature at atmospheric pressure. In the past few years, other materials have demonstrated even high critical temperatures, but only at extremely high pressures.

A critical technological need will be to bridge the increasingly high-temperature performance with ambient pressure [7]. A potentially useful tool to accomplish this goal of accelerated superconductive materials discovery is machine learning. There have been various applications in this work each with promising discoveries of new superconductive materials and properties with most of them predicting critical temperature. For example, superconducting phase diagrams were predicted using text mining [8], superconducting hydrogen compounds were found using a genetic algorithm and genetic programming [9] critical temperature and pressure were predicted for hydrides [10], critical temperatures of doped Fe-based superconductors were predicted based on structural and topological parameters [11], and critical temperature was predicted on a structure based model using a structural descriptor [12], and superconductor materials and properties have been automatically extracted from literature [13]. An ML-guided discovery will hopefully replace the “serendipitous discovery paradigm” that has existed in this last century of superconductor research [14].

In this work, we use the SuperCon data set for training, similar to what has been done in other implementations [12], [15]–[19]. Unique, reduced chemical formulae are curated [20] from the NOMAD data set [21] and used for testing. The chemical formulae in the training and testing data are first screened through `SMACT` [22] for validity and then trained and predicted using the `DiSCoVeR` algorithm [1]. This results in the screening of novel, chemically valid formulae with predicted critical temperatures.

II. Methods

A. Data

Materials informatics has shown that the cuprate class of superconductors contains a highly unexplored materials space that has yet to be explored [18]. This is why formulae from the SuperCon database are used for the training of our model. Of these formulae, “roughly 5,700 compounds are cuprates and 1,500 are iron-based (about 35 and 9 percent, respectively), reflecting the significant research efforts invested in these two families. The remaining set of about 8,000 is a mix of various materials, including conventional phonon-driven superconductors (e.g., elemental superconductors, A15 compounds), known unconventional superconductors like the layered nitrides and heavy fermions, and many materials for which the mechanism of superconductivity is still under debate (such as bismuthates and borocarbides)” [15].

Compositions from the Novel Materials Discovery (NOMAD) data set is used in the prediction of our model. This

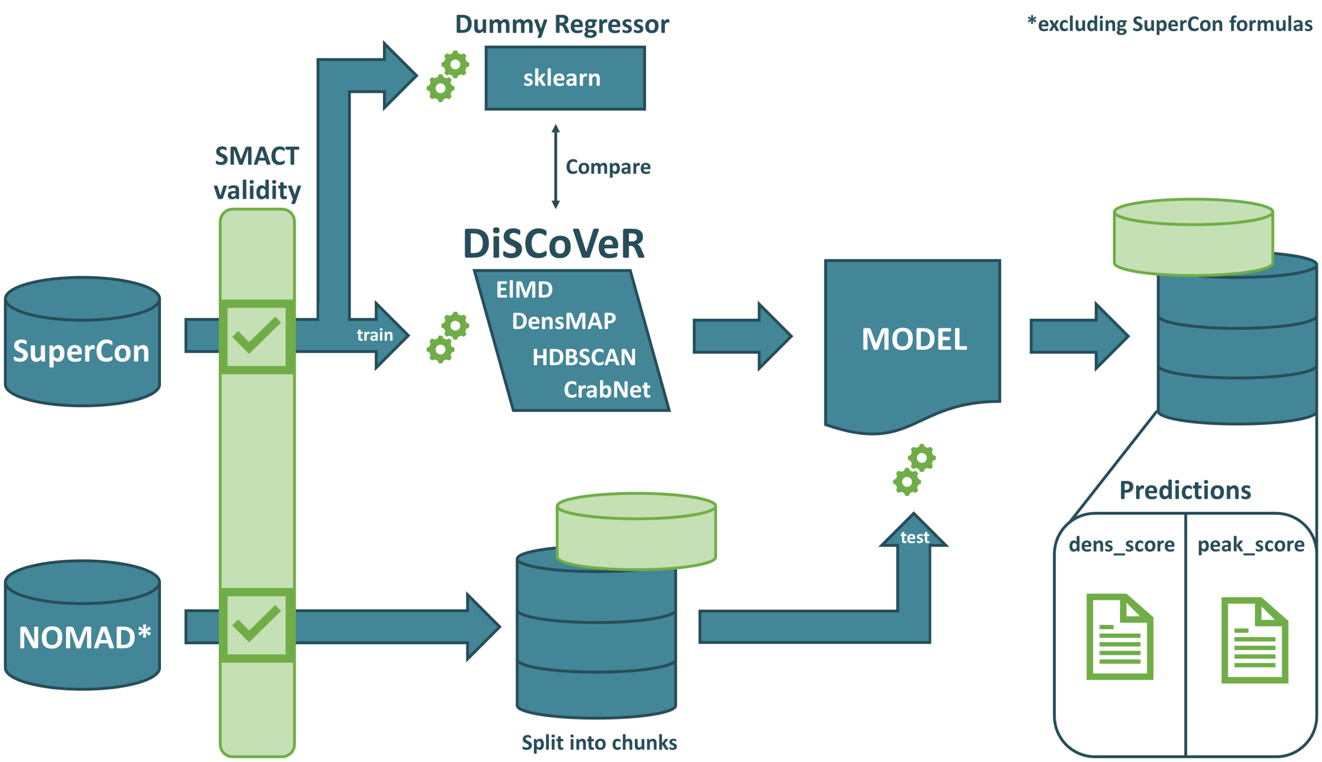

Fig. 2. Workflow of implementation. SuperCon and NOMAD formulas are first verified through `SMACT`. After featurization and training of the SuperCon data through `DiSCoVeR`, chunks of the NOMAD data are screened through to form a model

repository contains computational materials science data that is allowed to be curated [21]. For this implementation, we used a specific curated data set of unique reduced chemical formulae [20]. This data was restricted to density functional theory (DFT) calculations and does not include noble gases or radioactive elements. It is also directly usable with the `pymatgen.core.Composition` class [23], which is what this implementation exploits.

The compositions in SuperCon were reduced using the `get_reduced_composition_and_factor()` method from the `pymatgen.core.Composition` class. After curation, the NOMAD data set contained 695,611 compositions. Formulae in NOMAD that overlapped with formulae in SuperCon were removed for better accuracy while predicting. After some data cleaning, SuperCon data contained 12,415 formulae of superconductors and their critical temperatures. Figure 3 shows the distribution of the SuperCon data set after cleaning. The NOMAD data was reduced to 694,398 compositions. These compositions are then screened through `SMACT` for validity.

B. *`SMACT`*

`SMACT` is a composition-based screening tool [22]. It generates a search space, or a set of element combinations, that is screened using chemical filters. Oxidation states, charge neutrality, and electronegativity can be considered to screen for candidates that make "chemical sense." If the overall charge

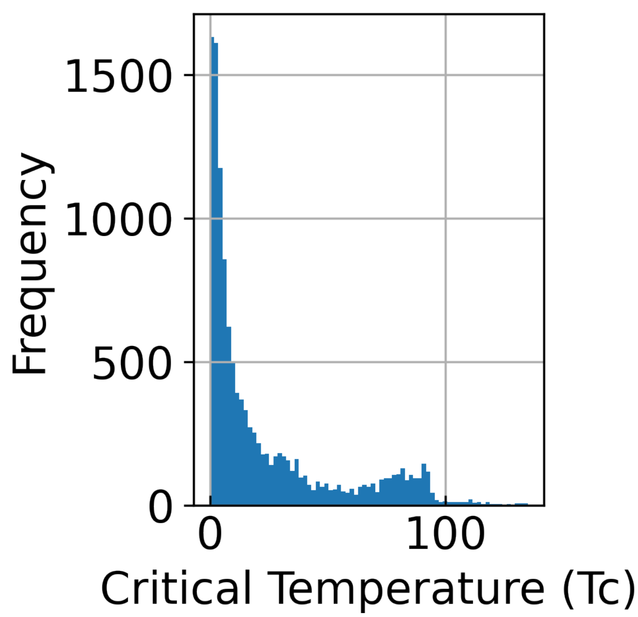

Fig. 3. Distribution of SuperCon data set after cleaning: Critical Temperature (T_c) on x-axis and Frequency on y-axis

TABLE I

METHODS USED IN THE DiSCoVeR ALGORITHM. REPRODUCED FROM [1] WITH PERMISSION FROM THE ROYAL SOCIETY OF CHEMISTRY.

Method	**What is it?**	**What is its role in DiSCoVeR?**
CrabNet [24]	Composition-based property regression	Predict performance for proxy scores
ElMD [25]	Composition-based distance metric	Supply distance matrix to DensMAP
DensMAP [26]	Density-aware dimensionality reduction	Obtain densities for density proxy
HDBSCAN [27]	Density-aware clustering	Create chemically homogenous clusters
Peak proxy	High performance relative to nearby compounds	Proxy for surprising high performance
Density proxy	Sparsity relative to nearby compounds	Proxy for chemical novelty
Peak proxy score	Weighted sum of performance and peak proxy	Used to rank compounds
Density proxy score	Weighted sum of performance and density proxy	Used to rank compounds
Pareto front	Optimal performance/uniqueness trade-offs	Visually screen compounds (no weights)

of a composition is neutral, then SMACT will consider it valid. The original checker however does not consider the countless combinations of oxidation states for metal alloys. To account for this, materials composed of all metal elements are assumed valid in the checker. To perform this, we implement a function called `smact_validity()`.

C. DiSCoVeR

DiSCoVeR stands for Descending from Stochastic Clustering Variance Regression. This algorithm is a conglomerate of multiple tools (as shown in Figure 2) that are ultimately used for the screening and assessment of the superconductive compositions. "DiSCoVeR screens candidates that have a high probability of success while enforcing – through the use of novel loss functions – that the candidates exist beyond typical materials landscapes *and* have high performance. In other words, DiSCoVeR acts as a multi-objective screening where the promise of a compound depends on both having desirable target properties and existing in sparsely populated regions of the cluster to which it's assigned. This approach then favors discovery of novel, high-performing chemical families as long as embedded points which are close together or far apart exhibit chemical similarity or chemical distinctiveness, respectively" [**?**]. Table I describes each of the methods used in DiSCoVeR and explains each of their roles.

The training data is also trained using sklearn's DummyRegressor and the mean average error (MAE) is compared alongside that of DiSCoVeR's to serve as a metric. During testing, the NOMAD data set is partitioned into chunks due to size. The predictions for high-performing compositions are appended and organized after being screened through the trained model.

III. RESULTS AND DISCUSSION

There are many essential properties to consider in the search for a novel superconductor such as the material's critical magnetic field, its critical current density, phase diagram information, and additional structural data. When considering the entire materials discovery process, synthesizing and screening candidate materials for superconductivity is the final objective. Critical temperature is the most reasonable superconductor property to predict, since critical magnetic field and critical current density are more difficult, intensive, and expensive to measure. In regard to extrapolation performance for superconductor discovery, Meredig et al. states that "novel materials discovery would be enabled by running a model against a large database of candidate compounds and simply ranking them by predicted T_c" [28], which is what is done in this implementation. For this specific implementation, a composition-based approach is used to test the limits of this algorithm by predicting a single property: a material's critical temperature.

A. T_c prediction

The DiSCoVeR algorithm has two expected outputs, `(peak_score)` and `(dens_score)`, that can be toggled as desired. These are metrics that contain a weighted score involving superconductor performance (by maximizing critical temperature) and chemical novelty, where chemical novelty is defined either using a density-based proxy or a peak-based proxy. In this case, `(peak_score)` and `(dens_score)` are both considered and weighted at 50/50. Table II shows the top 20 screened compositions from the NOMAD data set, with their predicted critical temperatures (T_c) in Kelvin. Compositions are first sorted by predicted critical temperature (shown in the prediction column). The columns for both scores aren't sorted since they are both evenly weighted, and both high-performing and chemically novel compositions are desired. Only formulae with a boolean value of TRUE are kept in the `is_valid` column (taken from SMACT validity), and only formulae with a predicted energy above hull close to zero are kept. The predicted stability metric indicates whether similar compounds have been made (0) or if it's theoretical (1).

B. Synthesizability prediction

As the test data were obtained from the NOMAD database, a repository of computationally-generated materials, we aimed to evaluate the synthesizability of the superconductors with predicted critical temperatures. To assess their stability, we queried the materials from the Materials Project and obtained

TABLE II
TOP 20 SCREENED COMPOSITIONS

formula	prediction	dens_score	peak_score	is_valid	predicted_e_above_hull	is_theoretical
$CaCu_4Sb$	150.66	3.3866	29.251	TRUE	0.0053	0.0521
YCu_8	133.91	4.3737	24.432	TRUE	0.0072	0.0009
$CaSbPb_4$	129.98	12.087	24.588	TRUE	0.0202	0.9997
$Ba_4Ca_4Cu_6Hg_2O_{17}$	129.46	75.558	20.483	TRUE	0.0436	1.0004
$BaMg_8$	128.09	5.3908	24.816	TRUE	0.0034	-4.74E-05
$Ba_6Ca_6Cu_9Hg_3O_{25}$	128.02	90.420	19.858	TRUE	0.0325	0.9862
Ba_2CaTl	126.59	2.4023	24.832	TRUE	-4.38E-05	0.9999
YCu_{13}	125.32	6.2119	25.134	TRUE	0.0006	1.69E-05
Ba2Ca3TlCu4O11	125.17	51.912	20.975	TRUE	0.0211	0.9998
Ba_wMg_{17}	122.49	2.2579	23.672	TRUE	4.91E-05	-3.52E-05
$Ba(ClO_4)_2$	119.62	5.4891	21.392	TRUE	6.38E-05	-5.56E-05
BaY_7	119.31	2.7561	22.166	TRUE	0.0927	0.9999
$BaCa_2C_2(O_3F)_2$	114.65	16.920	20.499	TRUE	0.0056	0.0005
$Ba_6Ca_6Tl_5Cu_9O_{29}$	113.70	32.527	19.585	TRUE	0.0196	0.0230
$Na(Cu_3O_4)_2$	112.20	11.140	20.507	TRUE	0.0510	0.9999
$Ba_8Ca_8Tl_7(Cu_4O_{13})_3$	111.69	49.059	19.663	TRUE	0.0227	0.1450
$Ba_2Ca_2Cu_3HgO_8$	111.57	60.550	16.642	TRUE	0.0151	0.0262
$TlCuO_2$	110.69	11.381	20.107	TRUE	0.0091	1.0000
$Ba_2CaTl_2(CuO_4)_2$	107.51	43.028	18.100	TRUE	0.0122	-0.0008

their energy above hull values. The lower the energy above hull, the more stable the compound is considered to be. Additionally, we obtained the `is_theoretical` property, which indicates if a material has been reported in the International Crystal Structure Database (ICSD) (i.e., if it has been synthesized previously).

We trained a `CrabNet` [24] model using data from the Materials Project, optimizing its hyperparameters with the Adaptive Experimentation (`Ax`) Platform (https://ax.dev). After training, this model was used to predict the energy above hull and (`is_theoretical`) property for the superconductors in question. The compounds were ranked based on their predicted energy above hull values (Table II), with the most stable compounds appearing first. For the (`is_theoretical`) property, a value close to 1 indicates that similar compounds have not been synthesized previously, and their synthesis would represent a new exploration in the chemical space. A value close to 0, on the other hand, suggests that similar compounds have been synthesized before, and their synthesis would be considered exploitation.

IV. CONCLUSION

12,415 known superconductors in the SuperCon database were first validated through `SMACT`, and then trained on the `DiSCoVeR` algorithm. 694,398 curated, chemically-novel formulae were taken from the NOMAD repository, also validated through `SMACT`, and then tested on the trained model in chunks. Critical temperatures for each of the formulae in this NOMAD data set were predicted. A weighted uniqueness/performance ranking for each of the compounds was obtained and sorted. These sorted compounds also include whether or not they are valid according to `SMACT`. After screening these compositions, additional post-processing work was done to predict energy above hull and stability, which are useful metrics for synthesis. This implementation reveals a process of performing ML-assisted superconductor screening using an algorithm that uniquely accounts for chemical similarity, and identifies and evaluates new high-performing, chemically distinct compositions. These predicted compositions are openly available in the hopes of being used in the materials discovery process. Since these validation formulae are ranked by score, they can now undergo additional post-processing and characterization.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF). C.C.S. acknowledges support from NSF Grant No. DMR-1950589 and Utah Valley University's Undergraduate Research Scholarly and Creative Activities (URSCA) program. H.M.S. and T.D.S. acknowledge support from NSF Grant No. 1936383. S.G.B. and T.D.S. acknowledge support from NSF Grant No. 1651668. We acknowledge **Erick Lawrence** for the idea of compositional stability prediction.

CREDIT STATEMENT

Colton C. Seegmiller: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. **Sterling Baird**: Supervision, Project administration, Conceptualization, Methodology, Software, Writing - Original Draft, Writing - Review & Editing. **Hasan M Sayeed**: Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. **Taylor D. Sparks**: Supervision, Project administration, Funding acquisition, Conceptualization, Formal analysis, Resources, Writing - Review & Editing.

Data Availability

The raw data required to reproduce these findings are available to download from https://github.com/vstanev1/Supercon, https://figshare.com/articles/dataset/NOMAD_Chemical_Formulas_and_Calculation_IDs/19319783. The processed data required to reproduce these findings are available to download from https://github.com/cseeg/DiSCoVeR-SuperCon-NOMAD-SMACT.

References

[1] Sterling G Baird, Tran Q Diep, and Taylor D Sparks. Discover: a materials discovery screening tool for high performance, unique chemical compositions. *Digital Discovery*, 1(3):226–240, 2022.

[2] H Kamerlingh Onnes. The resistance of pure mercury at helium temperatures. further experiments with liquid helium. iv. In *Proceedings Koninklijke Akademie van Wetenschappen te Amsterdam*, volume 13, pages 1274–1276, 1911.

[3] Mark Buchanan. Mind the pseudogap. *Nature*, 409(6816):8–12, 2001.

[4] J. G. Bednorz and K. A. Muller. Possible high Tc superconductivity in the Ba-La-Cu-O system. *Z. Phys. B*, 64:189–193, 1986.

[5] PJ Ray. Master's thesis: Structural investigation of la (2-x) sr (x) cuo (4+ y)-following staging as a function of temperature. *University of Copenhagen*, 2016.

[6] Maw-Kuen Wu, Jo R Ashburn, ClJ Torng, Pei-Herng Hor, Rl L Meng, Lo Gao, Z Jo Huang, YQ Wang, and aCW Chu. Superconductivity at 93 k in a new mixed-phase y-ba-cu-o compound system at ambient pressure. *Physical review letters*, 58(9):908, 1987.

[7] Simone Di Cataldo, Christoph Heil, Wolfgang von der Linden, and Lilia Boeri. La bh 8: Towards high-t c low-pressure superconductivity in ternary superhydrides. *Physical Review B*, 104(2):L020511, 2021.

[8] Callum J Court and Jacqueline M Cole. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning. *npj Computational Materials*, 6(1):18, 2020.

[9] Takahiro Ishikawa, Takashi Miyake, and Katsuya Shimizu. Materials informatics based on evolutionary algorithms: Application to search for superconducting hydrogen compounds. *Physical Review B*, 100(17):174506, 2019.

[10] Michael J Hutcheon, Alice M Shipley, and Richard J Needs. Predicting novel superconducting hydrides using machine learning approaches. *Physical Review B*, 101(14):144505, 2020.

[11] Yun Zhang and Xiaojie Xu. Predicting doped fe-based superconductor critical temperature from structural and topological parameters using machine learning. *International Journal of Materials Research*, 112(1):2–9, 2021.

[12] Jingzi Zhang, Zhuoxuan Zhu, X-D Xiang, Ke Zhang, Shangchao Huang, Chengquan Zhong, Hua-Jun Qiu, Kailong Hu, and Xi Lin. Machine learning prediction of superconducting critical temperature through the structural descriptor. *The Journal of Physical Chemistry C*, 126(20):8922–8927, 2022.

[13] Luca Foppiano, Pedro Baptista Castro, Pedro Ortiz Suarez, Kensei Terashima, Yoshihiko Takano, and Masashi Ishii. Automatic extraction of materials and properties from superconductors scientific literature. *Science and Technology of Advanced Materials: Methods*, 3(1):2153633, 2023.

[14] Stephan R Xie, Gregory R Stewart, James J Hamlin, Peter J Hirschfeld, and Richard G Hennig. Functional form of the superconducting critical temperature from machine learning. *Physical Review B*, 100(17):174513, 2019.

[15] Valentin Stanev, Corey Oses, A Gilad Kusne, Efrain Rodriguez, Johnpierre Paglione, Stefano Curtarolo, and Ichiro Takeuchi. Machine learning modeling of superconducting critical temperature. *npj Computational Materials*, 4(1):29, 2018.

[16] Zhong-Li Liu, Peng Kang, Yu Zhu, Lei Liu, and Hong Guo. Material informatics for layered high-tc superconductors. *APL Materials*, 8(6):061104, 2020.

[17] Tomohiko Konno, Hodaka Kurokawa, Fuyuki Nabeshima, Yuki Sakishita, Ryo Ogawa, Iwao Hosako, and Atsutaka Maeda. Deep learning model for finding new superconductors. *Physical Review B*, 103(1):014509, 2021.

[18] Rhys EA Goodall, Bonan Zhu, Judith L MacManus-Driscoll, and Alpha A Lee. Materials informatics reveals unexplored structure space in cuprate superconductors. *Advanced Functional Materials*, 31(52):2104696, 2021.

[19] Elizabeth A Pogue, Alexander New, Kyle McElroy, Nam Q Le, Michael J Pekala, Ian McCue, Eddie Gienger, Janna Domenico, Elizabeth Hedrick, Tyrel M McQueen, et al. Closed-loop machine learning for discovery of novel superconductors. *arXiv preprint arXiv:2212.11855*, 2022.

[20] Sterling G. Baird. NOMAD Chemical Formulas and Calculation IDs. 3 2022.

[21] Claudia Draxl and Matthias Scheffler. The nomad laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*, 2(3):036001, 2019.

[22] Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019.

[23] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[24] Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials*, 7(1):77, 2021.

[25] Cameron J Hargreaves, Matthew S Dyer, Michael W Gaultois, Vitaliy A Kurlin, and Matthew J Rosseinsky. The earth mover's distance as a metric for the space of inorganic compositions. *Chemistry of Materials*, 32(24):10610–10620, 2020.

[26] Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. *bioRxiv*, pages 2020–05, 2020.

[27] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[28] Bryce Meredig, Erin Antono, Carena Church, Maxwell Hutchinson, Julia Ling, Sean Paradiso, Ben Blaiszik, Ian Foster, Brenna Gibbons, Jason Hattrick-Simpers, et al. Can machine learning identify the next high-temperature superconductor? examining extrapolation performance for materials discovery. *Molecular Systems Design & Engineering*, 3(5):819–825, 2018.