Evaluating AI-guided Design for Scientific Discovery

Michael Pekala[1] **Elizabeth A. Pogue**[1] **Kyle McElroy**[1] **Alexander New**[1]

Gregory Bassen[2,3,4] **Brandon Wilfong**[2,3,4] **Janna Domenico**[1]

Tyrel M. McQueen[2,3,4] **Christopher D. Stiles**[1]

[1]Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, Maryland 20723, USA
{firstname.lastname}@jhuapl.edu

[2]Department of Materials Science and Engineering
[3]Department of Chemistry
[4]Institute for Quantum Matter
Johns Hopkins University
3400 N. Charles Street, Baltimore, Maryland, 21218, USA.
{gbassen1,bwilfon3,mcqueen}@jhu.edu

Abstract

Machine learning has great potential to revolutionize experimental materials research; however, the degree to which these approaches accelerate novel discovery is rarely quantified. To this end, we propose a framework for characterizing the rate of "first discovery" of scientific hypotheses in the form of materials families. We use a combination of the SuperCon and Materials Project databases to simulate a scientific needle-in-a-haystack discovery problem as a motivating example. We use this approach to compare the ability of different adaptive sampling strategies to rediscover promising superconductor families, such as the Cuprates and iron-based superconductors. This methodology can be applied using various notions of novelty, making it applicable to discovery problems more broadly.

1 Introduction

Discovery is central to the advancement of science, but targeted discovery is notoriously difficult. The discovery process is often opaque and explanations of why we follow path A over path B can be difficult to teach but are crucial for spurring subsequent discoveries [6]. In many fields, promising new candidates occupy only a tiny portion of a vast search space, making targeted discovery analogous to finding a needle in a haystack [43, 25]. Furthermore, we are often interested in materials with emergent properties, by which we mean properties that are not immediately obvious consequences of the individual constituents (e.g. as in superconductivity, where the critical temperature cannot be easily derived from the constituent elements). For this reason, fundamentally new and revolutionary results are rare, not all discoveries are equally useful, and incremental discovery is the norm. The challenge is how to best employ data-driven techniques to accelerate the scientific method for the discovery of useful, novel materials.

Artificial Intelligence (AI) and Machine Learning (ML) may hold the key to enable more targeted explorations into the unknown [33]. However, ML relies upon representative training data and out-of-distribution generalization is a known challenge [14]. Unfortunately, scientific publications and their

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

corresponding data sets are often not representative of the entire space of possibilities. Confirmation bias and the desire to make practical use of discoveries favors the publication of positive results and encourages incremental exploration in the region of known rare events. Consequently, scientific data sets typically only cover a fraction of the overall search space, tend to be more densely sampled in regions where positive results have been discovered previously, and frequently under-represent negative results.

While substantial progress has been made applying ML for materials property prediction and discovery [7, 17, 38, 28, 20, 32, 9, 29, 47, 4, 46, 37, 31, 44, 12, 50], fully assessing the notion of generalization for novel discovery subject to data set limitations remains an outstanding challenge. For example, the search for new materials that conduct electricity with zero resistance (superconductors) is challenging from both experimental and theoretical perspectives [16, 36, 3]. On a statistical basis, several hundred new superconducting families may need to be discovered to find another $T_c > 100$ K material, with orders of magnitude more to realize room-temperature superconductivity [3]. In this paper we introduce a framework for simulating the discovery process for novel materials, present example experiments from the context of superconductivity research, and empirically assess the performance of ML-driven closed-loop, adaptive discovery methodologies.

2 Related Work

Discovery, including materials discovery, presents unique challenges. In Meredig et al. [32], the authors observe that the random k-fold cross-validation (CV) procedure often used to estimate generalization performance in ML applications is ill-suited to discovery problems and leads to overly optimistic performance estimates. The reason is that unobserved regions of materials space are unlikely to exhibit precisely the same statistical stoichiometry-to-property relationships as the previously explored training data and thus the model is being asked to make predictions "out-of-distribution". As an alternative, they propose a leave-one-cluster-out cross-validation (LOCO CV) approach, whereby data splits are defined by a clustering approach that better represents the challenge of making predictions in previously unseen regimes of design space. The impact of different clustering approaches has also been considered in this context [9].

Scientific data sets are typically not static - they evolve over time as measurements are contributed or corrected. This adds another dimension to out-of-distribution challenges. The authors of Li et al. [29] observe that predictors can perform well on one version of a scientific data set but degrade as that data set evolves over time. They demonstrate this effect using a variety of ML models in conjunction with the Materials Project data set [18] from two different year, 2018 and 2021, as a test case.

Sequential methods, which iteratively sample from the design space and use newly acquired measurements to improve subsequent predictions, provide a compelling paradigm for ML-driven scientific discovery. One example is Bayesian optimization (BO) which has a long history for sequential design [22]; extensions of this approach continue to be developed for needle-in-the-haystack type problems found in materials science, e.g. [44]. Sequential methods also figure prominently in the active learning community. In particular, *adaptive search* is a special case of active learning where the goal is to discover members of a rare, desirable class by sequentially selecting and inspecting data points [11]. In the materials discovery context, Jiang et al. consider active search to identify new bulk metallic glasses which comprised 4% of the membership of a held-out test set [21]. However, explicitly incorporating a notion of novel discovery over coherent groups representing scientific hypotheses does not appear to have been studied.

One early example applying sequential learning to materials design is Ling et al. [31]. The authors used random forests as the underlying predictors and tracked the mean number of predictions required to identify an optimal candidate within the held-out test data set. This metric is similar in spirit to the notion of sample efficiency typically used to evaluate the performance of BO or reinforcement learning algorithms [10, 8]. In Koppel et al. and Pogue et al., the authors pair ML with physical fabrication and characterization [27, 37]. Pogue et al. explicitly closed the design loop by both making and testing ML-derived predictions through several iterations [37].

To date, there have only been a few studies that explicitly quantify the cost (in measurements) of revolutionary materials discovery in adaptive settings. Baird et al. introduced the Descending from Stochastic Clustering Variance Regression (DiSCoVeR) algorithm to aid in identifying promising chemically-unique compounds using ML [4]. They tested this algorithm using computationally-

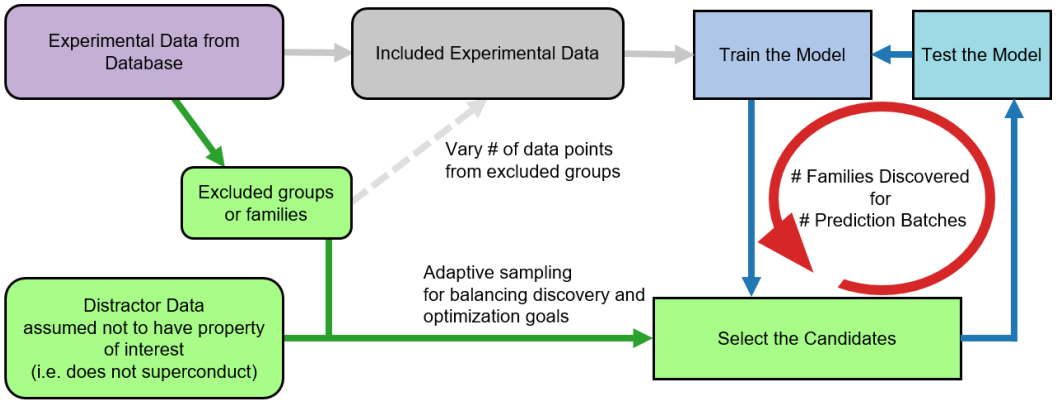

Figure 1: Our approach to evaluating the rate of discovery associated with different sampling approaches involves jointly evaluating performance on a held-out group of materials in a sequential manner, allowing us to better understand the overall rate of discovery and adaptability of the model.

predicted bulk modulus data and compared this approach to naive random search. Since bulk modulus is a continuous quantity that every material has, it is fundamentally different from superconductivity, which is an emergent property only found in some materials that is not fully understood and more difficult to predict.

The same DiSCoVeR algorithm was subsequently used to quantify novelty in the superconductivity setting [42], although only one round of predictions was made and an explicit characterization of sample efficiency is not performed. They also did not utilize the same domain expert-informed notion of novel discovery that will be presented in Section 3.

The approach of Borg et al. [5] is closest to ours in that the focus is upon quantifying the performance of sequential learning in a simulated discovery setting. The problem setup is based on a decomposition of the target values, but their metrics are different from the notion of first discovery we will discuss in Section 3. Their metrics do not contain a sense of the similarity between compounds, which are captured by our family groupings.

3 Approach

While a number of studies interrogate ML prediction performance on held-out material groups (and this is becoming more systematic, e.g. [40]), or assess some measure of sequential performance, most do not consider both jointly. To quantify the impact of various algorithmic choices, we propose an evaluation procedure that mirrors the full scope of the discovery setting as closely as possible. In our approach (Figure 1), we pose a simulated discovery problem whereby one or more scientifically related groups (termed here "target groups") manifesting a rare property of interest are withheld from the initial training set. These target groups serve as proxies for scientific hypotheses, whose re-discovery we wish to simulate.

We then combine these target groups with select negative examples from the data set, as well as with a large corpus of unlabeled "distractor" data which further augment the set of negative examples. This mixture comprises the test set. Supplementing with additional negative examples helps mitigate the aforementioned tendency for existing data sets to under-represent negative results. In addition to adding negative examples, we also downsample the target groups, reducing their relative abundance so that our experiments more closely approximate the asymmetry of real-world scientific discovery.

Using this data set we conduct a simulated adaptive design experiment, whereby a ML model is trained, the model makes predictions on the test set, a subset of these predictions are selected for measurement, and these selected materials (and their labels) are incorporated into the training data set for the next iteration. The metric for this simulated discovery task is defined as the number of recommendations (simulated measurements) required before any member of a given target group appears within the top k recommendations in some iteration of the adaptive procedure. Once a target material appears as a recommendation, its entire group is considered "first discovered". This is in

some sense a global notion of discovery which presumes that, after any member of the target group is identified, traditional methodologies will be used to perform more exhaustive local exploration to further refine the result (e.g. by sampling in the neighborhood of the material system or via standard per-atom substitution methods).

The above procedure is analogous to sample efficiency studies used to evaluate BO algorithms, with a key difference being the focus on target groups and first discovery times. The process of restricting target groups to the initial test is reminiscent of the LOCO CV procedure, but our performance metric is related to discovery time vs a static measure of performance on the entire held-out group.

Of course, our approach assumes that the unlabeled examples all correspond to negative results; this may not be straightforward to establish in all settings, or in fact a few positive examples may enter and introduce label noise. However, assuming that the property of interest is indeed rare, if positive examples are present in the unlabeled set, our assumption introduces a small but systematic bias in the results. In this case, the discovery time metric will overestimate the real discovery time since feedback from false negative simulated measurements may cause the model to miss correlations that otherwise might accelerate discovery of target group(s). Due to this overestimate, our procedure yields an approximate upper bound on discovery time. However, because this error is systematic, we contend this is still a useful mechanism to compare and assess algorithms.

4 Use Case

To demonstrate the "first discovery" framework of Section 3, we consider a concrete problem: discovering novel superconductors. Superconductors satisfy the needle-in-a-haystack assumption and have the advantage that there exist data sets with pre-existing categorizations into scientifically distinct families, such as Cuprates, iron-based, Borides, and $BaAl_4$-derivatives. While the community has not converged on a single universally accepted definition of "family", adapting an existing taxonomy provides us with a tangible starting point.

4.1 Data Sets

The basis for our analysis is the 2022-08-08 release of the SuperCon Database [19]. SuperCon is one of the largest public data sets for experimental measurements of superconductivity and lists the superconducting transition temperature (T_c) of > 24000 materials (Figure 2). Since most materials in nature do not superconduct at achievable temperatures, this database over-represents materials that superconduct and have higher T_c's. To understand bias and its effects on model performance, we categorized superconductors in the SuperCon database by family. The family designations included in SuperCon are used to construct our target groups. The distribution of critical temperature, T_c, and fraction of each family present in SuperCon are displayed in Figure 2. The Cuprate family known for having high T_c members comprises $\sim 30\%$ of the database and other families, like the iron-based pnictides and Borides, are also over-represented. For our supplementary source of negative examples we employ data from Materials Project [18], a database of computationally generated materials for which critical temperature values (T_c) are generally unknown. While Materials Project is not devoid of superconducting materials (e.g. see [37]) we assume they are rare and adopt the upper bound interpretation of results described in Section 3.

A number of the materials within SuperCon lack a T_c entry (the field contains `NaN`). However, a subset of these materials have been experimentally determined to not superconduct above a known temperature value (i.e. the dataset provides an upper bound on the possible T_c value). This is designated by the "`tcn`" field in SuperCon. For the subset of materials lacking a T_c label and for which the `tcn` field is less than $5\,\mathrm{K}$ we impute a T_c value of 0 and include the material in our data set; the remaining entries lacking a T_c value are discarded. We also discard any material whose composition string we are not able to parse with pymatgen [35] or for which the parsed composition contains dummy species. This is done in part to facilitate ingestion by the machine learning models. We also removed two iron-based entries with $T_c > 150\,\mathrm{K}$ (associated with refno PRB098014518) whose values did not appear entirely consistent with the original reference. The result is a data set of 24,161 entries.

For Materials Project (our test set distractors) we restrict our attention to "stable" entries, defined here as those whose energy above the convex hull (a measure of how far a given material lies from

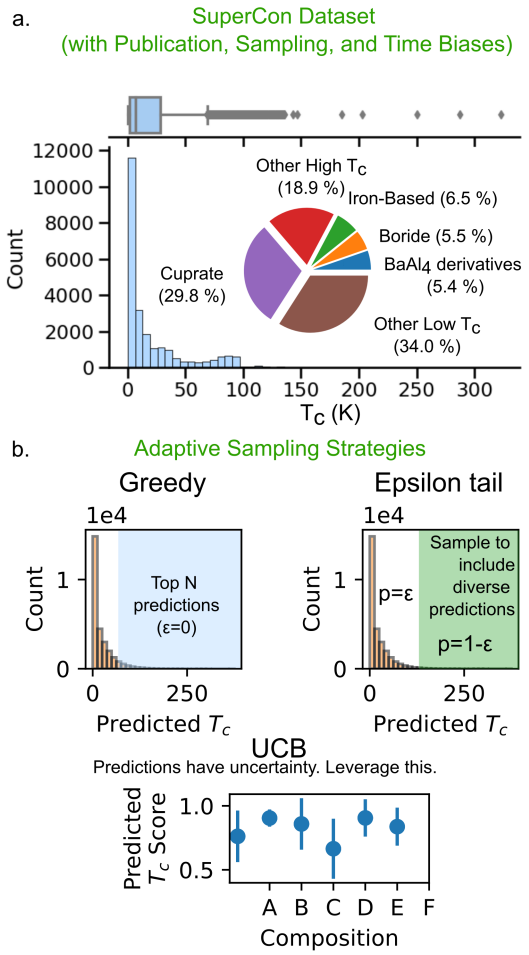

Figure 2: Distribution of SuperCon data and adaptive sampling strategies. Panel (a) displays the SuperCon dataset distribution of T_c and the contributions of various superconducting families to the database. (b) explains the three adaptive sampling strategies beyond uniform, random sampling examined in this study. These were chosen to represent both extremes of optimization (greedy) vs. discovery (uniform random sampling) goals while also representing strategies making use of estimated uncertainty from the ML predictions (upper confidence bound) and hybrid approaches (epsilon-tail). The epsilon-tail approach, containing aspects of both greedy and random uniform sampling approaches, when requesting N samples to test, selects samples from the predicted high-T_c tail with probability 1-ϵ and samples from the low-T_c base with probability ϵ.

the most energetically favorable condition presently known) is less than $0.1\,\mathrm{eV/atom}$ (as reported in Materials Project). The majority of these have negligible energy above the hull and the 75th percentile is $0.028\,\mathrm{eV/atom}$. To ensure these materials are sufficiently distinct from SuperCon, we furthermore drop any entries whose Element Movers Distance (ElMD) [13] to any SuperCon entry is ≤ 0.2 (where the ElMD is a distance metric between inorganic compositions that is informed by a notion of work required to transform one element to another). The threshold of 0.2 is a heuristic based on a domain expert's manual inspection of a subset of materials. The end result is a total of 56,866 putative non-superconducting materials which are used to expand the test set. In the future, generative methods (e.g. [51, 24, 39, 52, 1, 34]) would be well-suited for enriching the test set.

4.2 Family Definitions

The family definitions within SuperCon (contained within the `str3` column) are delineated based on a combination of the chemical constituents, crystal structure, and hypotheses for why materials superconduct (see Supplementary Information (SI)). We took the `str3` family designations for compounds in SuperCon and further aggregated these into the families shown in Figure 2; these families play the role of target groups described in Section 3. The initial number of unique `str3` designators was prohibitively large for human-interpretable analysis and visualization. This initial set of designators occasionally contained multiple designators for a given sub-family and a mixture of broad and specific groupings. Further details on our re-grouping are provided in the SI. Other ways of associating materials with hypotheses are of course possible. In addition to alternative manually constructed taxonomies, recent work has explored using unsupervised learning-based approaches to define superconductivity families [41]. More broadly, it is an open research question how best to quantify the notion of scientific hypotheses so that they can be incorporated within ML approaches [26].

In this experiment, the Cuprate, Borides, $BaAl_4$-derivatives, and iron-based families comprise the primary target groups of interest. These four families were entirely withheld from the initial training data set. For the aggregate SuperCon families (denoted "Other high T_c" and "Other low T_c" in Figure 2), these were allocated to train and test, on a per-family basis, in an approximately 80/20 ratio. The motivation for this split was to give the model some concept of superconductivity at time $t = 0$ while withholding the families that best represent scientific hypotheses so that their discovery rate can be assessed. In order to approach a more realistic level of scarcity, we further downsampled the target classes within the test set. For the Cuprates, Borides, iron-based, and $BaAl_4$-derivatives families we further downsample so that each family has no more than 20 representatives within the test set. The end result is that a given test fold contains less than 0.5% superconductors (see SI for further details).

4.3 Machine Learning Predictors

There are by now a range of ML models available for material property prediction. For our experiments, we leverage an existing ML algorithm named "Representation learning from Stoichiometry" (RooSt) [12]. The RooSt model is a graph-based convolutional neural network which solves a regression or classification problem on the basis of stoichiometry alone. While there also exist structure-aware ML models, the flexibility to work with compositions of unknown structure, together with the relative simplicity of the RooSt model, make it a good candidate for our studies. Furthermore, the RooSt model readily produces uncertainty estimates which are useful for certain adaptive sampling strategies. Another advantage is that RooSt has been employed in real-world discovery contexts (e.g. [37]), and hence analyses related to its discovery rate are of independent interest.

We largely rely on the default hyperparameters for the RooSt model. We have empirically observed these to be fairly robust across a range of problems and this is consistent with analyses conducted by the RooSt authors themselves. Key hyperparameters are reported in the SI. For our experiments, we use an ensemble of 5 models, primarily to control the computational cost. We used RooSt in the "robust" mode, whereby the model behaves as a Mean Variance Estimation (MVE) neural network, i.e., the model assumes target values are distributed normally and the model's predictions consist of the corresponding estimated mean and variance [45]. Our uncertainty estimates are based on combining the epistemic uncertainty obtained from the ensemble together with the predicted variance as described in [12].

4.4 Adaptive Sampling

Given mechanisms for making predictions, and for characterizing the uncertainty of those predictions, it remains to define a methodology for recommending a small set of high-quality candidate materials whose properties should be characterized experimentally. For screening applications where a finite pool of possible candidates is defined a-priori (e.g. from an existing database), an obvious strategy is to make predictions on each candidate, rank the predictions by the estimated T_c, and recommend the top k, a strategy which we term "greedy". Another alternative is to use both the prediction and the uncertainty, which gives rise to the typical "exploration vs exploitation tradeoff" [10]. There are a wide variety of so-called acquisition functions which manage this tradeoff; these methods have also been applied to materials science discovery within the context of BO [30]. There also exist batch variants ("q-acquisition functions") that address cases where multiple experiments can be conducted in parallel [49]. One commonly employed acquisition function is the Upper Confidence Bound (UCB). In our experiments, we adopt the approach of [23] where the UCB can be related to quantiles of a posterior distribution (which, in our case, is obtained from the MVE network). Specifically, let $Q(t, \rho)$ denote the quantile function associated with the distribution ρ such that $\mathbb{P}(X \leq Q(x, \rho)) = x$; then instead of using the predicted value to rank candidates we use

$$q_j(t) = Q(1 - \alpha_t, \rho_j^{t-1}),$$

where t is index of the adaptive sampling round, α_t is of order $1/t$, and ρ_j^{t-1} denotes the posterior estimate for candidate j generated in the previous adaptive sampling round [23]. In this work, we assume a finite pool of candidates and use the sequential UCB criteria above.

An assumption underlying the UCB approach is that one has drawn sufficient examples from the domain such that the uncertainty estimates are well-calibrated, something that is not entirely obvious when making predictions in novel regions of materials space. Therefore, we also consider strategies that incorporate exploration without relying on the model's uncertainty estimates. For this, we use a strategy whereby the model preferentially selects "outlier" T_c predictions (defined as predictions lying 1.5 times the inter-quartile range above the 75th percentile) with probability $1 - \epsilon$ and otherwise selects candidates uniformly at random with probability ϵ. This strategy is termed "epsilon-tail".

As a baseline, we also consider a pure uniform sampling strategy to emulate a pure "trial-and-error" discovery methodology. Ideally ML-informed methods will outperform this baseline.

5 Results

Figure 3 shows the result of running 50 adaptive sampling iterations using the experimental setup described above. The x-axis depicts the adaptive sampling iteration, where each iteration comprises a batch of 10 recommendations made using a given strategy. The y-axis shows the cumulative number of families discovered; the lines depict the mean across 10 random trials and the shaded regions denote 95% confidence intervals.

From the figure, it is clear that "greedy" and "epsilon-tail" strategies eventually outperform uniform random sampling (after approximately batch 12, or 120 simulated measurements). While the precise time at which the ML-based methods begin outperforming the baseline in our simulation is not necessarily completely reflective of discovery "in the wild" (due to approximations in inherent in our simulation and also the upper bound interpretation inherent in our approach), it nevertheless provides quantitative evidence to suggest that, indeed, ML-based methods can provide significant acceleration over blind trial-and-error. This ability to quantitatively compare different algorithmic choices, and contrast with representative baselines, is a key benefit of our re-discovery simulation approach.

While it is not statistically significant, the mean performance might suggest that the greedy method underperforms the the epsilon-tail strategy initially before overtaking it sometime around iteration 12. This would be consistent with the intuition that the ML model may struggle initially on out-of-domain data and that managing the exploration-vs-exploitation tradeoff may be especially important early in the adaptive sampling process.

Along these lines, it is interesting to observe that the UCB approach did not perform well in this study. One remark here is that we used a non-batch version of UCB (i.e. we did not use q-UCB or similar methods) which would take proper advantage of the fact that each iteration permits multiple recommendations. Furthermore, it may be that the uncertainty estimates we are using are not well

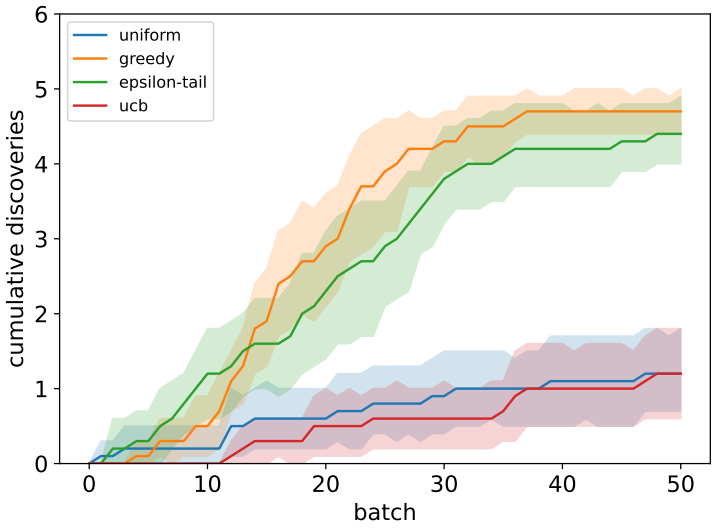

Figure 3: Results of sample efficiency study. The x-axis depicts the number of adaptive sampling iterations, where each iteration corresponds to a batch of 10 material recommendations (thus, 500 materials were recommended overall). The y-axis shows the cumulative number of families discovered (i.e. have at least one member that has been recommended by the algorithm). For each strategy, the mean and 95% CI are shown, obtained over 10 random trials (which reflects variation both in initial model weights as well as test set membership; more details are provided in the SI).

Table 1: Family discovery count by family and adaptive sampling method. The family is considered discovered in a trial if a member is recommended within any of the 50 batches. Numbers in the table reflect discovery counts over the same 10 random trials described in Figure 3. Of the four explicitly designated families, the Cuprates seem to be the most challenging to rediscover. Note that the aggregate (other high, other low) families are relatively more abundant in the test set (see also SI).

Held-out family	Epsilon	Greedy	UCB	Uniform
$BaAl_4$-derivatives	9	10	2	0
Borides	8	10	4	1
Cuprate	2	3	3	1
Iron-based	7	8	0	2
Other Low Tc	8	6	1	5
Other High Tc	10	10	2	3

calibrated in this out-of-domain setting, which would work against methods that depend upon them. Understanding conditions under which various uncertainty quantification methods are well calibrated continues to be an active area of research, e.g. [48, 15]. A productive direction for future work would be to investigate these model calibration and performance issues more carefully in this setting.

The y-axis of Figure 3 does not distinguish among the 6 target families (two of which are "aggregations" and the other four are more closely aligned with specific scientific hypotheses). To further decompose this aspect of the results, Table 1 shows the total number of discoveries (across all 10 randomized trials) by family and adaptive sampling strategy. One observation here is that, of the four families most closely aligned with scientific hypotheses, the Cuprate family seems to be the most difficult to uncover.

6 Discussion

In this work we propose a quantitative methodology for assessing the rate of novel hypothesis discovery in ML-driven settings, where hypotheses are framed in terms of a scientifically coherent

materials taxonomy. We can compare the number of discovered families for different numbers of prediction batches. Due to the cost of experiments, for many material or chemical discovery problems, studies are conducted in the low batch number regime. Here, time to first family discovery may be an appropriate metric. Using a representative problem of superconductor discovery, we show that modern ML-based predictors with simple acquisition strategies indeed appear to provide a distinct advantage over a pure trial-and-error approach. We hope the ideas included here will help inform future ML analyses and also future benchmarking efforts, similar to those of Riebesell et al. [40]. Approaches like this are useful for answering the question: "What is the right strategy to encourage and stimulate discovery?" As such, we can move beyond assessing ML/AI methods and also assess both traditional and non-traditional approaches to discovery to understand the tradeoffs involved.

The approach we outlined uses predefined superconductor families as a surrogate for scientific hypotheses; however, there is abundant opportunity to explore other ways of formalizing scientific hypotheses so that they can be rigorously assessed in ML-informed discovery settings. This is one open research question we look forward to exploring further in future work. There are also opportunities to expand the scope of acquisition functions considered and further enrich the experimental data using recent developments in generative models. More broadly, this work represents one step towards our longer-term objectives of further understanding and improving practical model generalization and the exploration of scientific hypotheses within the context of ML-driven discovery to enable broader AI-enabled understandings beyond prediction lists.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge internal financial support from the Johns Hopkins University Applied Physics Laboratory's Independent Research & Development (IR&D) Program. TMM acknowledges support of the David and Lucile Packard Foundation. We also thank Chris Ratto and the anonymous reviewers for helpful comments and suggestions in preparing this paper.

References

[1] M. Alverson, S. Baird, R. Murdock, S.-H. Ho, Jeremy Johnson, and Taylor Sparks. Generative adversarial networks and diffusion models in material discovery, 2023.

[2] Yuji Asada. Construction of factual database on oxide superconductors. *Bulletin of the Japan Institute of Metals*, 29(9):754–757, 1990.

[3] J. Paul Attfield. Chemistry and high temperature superconductivity. *Journal of Materials Chemistry*, 21(13):4756–4764, 2011.

[4] 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.

[5] Christopher KH Borg, Eric S Muckley, Clara Nyby, James E Saal, Logan Ward, Apurva Mehta, and Bryce Meredig. Quantifying the performance of machine learning models in materials discovery. *Digital Discovery*, 2(2):327–338, 2023.

[6] Paul C. Canfield. New materials physics. *Reports on Progress in Physics*, 83(1), 2020.

[7] Harriet A. Carroll, Zoi Toumpakari, Laura Johnson, and James A. Betts. The perceived feasibility of methods to reduce publication bias. *PLoS ONE*, 12(10):1–19, 2017.

[8] Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *Deep Reinforcement Learning Workshop NeurIPS 2022*, 2022.

[9] Samantha Durdy, Michael W Gaultois, Vladimir V Gusev, Danushka Bollegala, and Matthew J Rosseinsky. Random projections and kernelised leave one cluster out cross validation: universal baselines and evaluation tools for supervised machine learning of material properties. *Digital Discovery*, 1(6):763–778, 2022.

[10] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[11] Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff Schneider, and Richard Mann. Bayesian optimal active search and surveying. *arXiv preprint arXiv:1206.6406*, 2012.

[12] Rhys EA Goodall and Alpha A Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature communications*, 11(1):1–9, 2020.

[13] 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.

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[15] Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020.

[16] Hideo Hosono, Keiichi Tanabe, Eiji Takayama-Muromachi, Hiroshi Kageyama, Shoji Yamanaka, Hiroaki Kumakura, Minoru Nohara, Hidenori Hiramatsu, and Satoru Fujitsu. Exploration of new superconductors and functional materials, and fabrication of superconducting tapes and wires of iron pnictides. *Science and Technology of Advanced Materials*, 16(3), 2015.

[17] Jianjun Hu, Stanislav Stefanov, Yuqi Song, Sadman Sadeed Omee, Steph-Yves Louis, Edirisuriya MD Siriwardane, Yong Zhao, and Lai Wei. Materialsatlas. org: a materials informatics web app platform for materials discovery and survey of state-of-the-art. *npj Computational Materials*, 8(1):65, 2022.

[18] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[19] Supercon. `https://supercon.nims.go.jp/`. Accessed: 2021.

[20] Xiwen Jia, Allyson Lynch, Yuheng Huang, Matthew Danielson, Immaculate Lang, Alexander Milder, Aaron E Ruby, Hao Wang, Sorelle A Friedler, Alexander J Norquist, and Joshua Schrier. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*, 573:251–255, 2019.

[21] Shali Jiang, Gustavo Malkomes, Geoff Converse, Alyssa Shofner, Benjamin Moseley, and Roman Garnett. Efficient nonmyopic active search. In *International Conference on Machine Learning*, pages 1714–1723. PMLR, 2017.

[22] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13:455–492, 1998.

[23] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR, 2012.

[24] Evan Kim and S. V. Dordevic. ScGAN: A generative adversarial network to predict hypothetical superconductors, 2022.

[25] Yoolhee Kim, Edward Kim, Erin Antono, Bryce Meredig, and Julia Ling. Machine-learned metrics for predicting the likelihood of success in materials discovery. *npj Computational Materials*, 6(1), 2020.

[26] Hiroaki Kitano. Nobel turing challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications*, 7(1):29, 2021.

[27] Carson Koppel, Brandon Wilfong, Allana Iwanicki, Elizabeth Hedrick, Tanya Berry, and Tyrel M McQueen. Machine-guided design of oxidation-resistant superconductors for quantum information applications. *Inorganics*, 11(3):117, 2023.

[28] Masaya Kumagai, Yuki Ando, Atsumi Tanaka, and Kojis Tsuda. Effects of data bias on machine-learning–based material discovery using experimental property data. *Science and Technology of Advanced Materials: Methods*, 2(1):302–309, 2022.

[29] Kangming Li, Brian DeCost, Kamal Choudhary, Michael Greenwood, and Jason Hattrick-Simpers. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Computational Materials*, 9(1):55, 2023.

[30] Qiahao Liang, Aldair E Gongora, Zekun Ren, Armi Tiihonen, Liu Zhe, James R Sun, Shijing an dDenault, Daniil Bash, Flore Mekki-Berrada, Saif A Khan, et al. Benchmarking the performance of bayesian optimization across multiple experimental materials science domains. *npj Computational Materials*, 7(1):188, 2021.

[31] Julia Ling, Maxwell Hutchinson, Erin Antono, Sean Paradiso, and Bryce Meredig. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integrating Materials and Manufacturing Innovation*, 6:207–217, 2017.

[32] 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.

[33] Austin M Mroz, Victor Posligua, Andrew Tarzia, Emma H Wolpert, and Kim E Jelfs. Into the unknown: How computation can help explore uncharted material space. *Journal of the American Chemical Society*, 144(41):18730–18743, 2022.

[34] Alexander New, Michael Pekala, Elizabeth A Pogue, Nam Q Le, Janna Domenico, Christine D. Piatko, and Christopher D Stiles. Evaluating the diversity and utility of materials proposed by generative models. In *1st Workshop on the Synergy of Scientific and Machine Learning Modeling @ ICML2023*, 2023.

[35] 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.

[36] Warren E. Pickett. Colloquium: Room temperature superconductivity: The roles of theory and materials design. *Reviews of Modern Physics*, 95(2):21001, 2023.

[37] 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. *npj Computational*, 2023. (accepted).

[38] Paul Raccuglia, Katherine C. Elbert, Philip D.F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, 2016.

[39] Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G. Aberle, Shijing Sun, Xiaonan Wang, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, Kedar Hippalgaonkar, Yousung Jung, and Tonio Buonassisi. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, 2022.

[40] Janosh Riebesell, Rhys Goodall, Anubhav Jain, Philipp Benner, Kristin Persson, and Alpha Lee. Matbench discovery. `https://matbench-discovery.materialsproject.org`. Accessed: 2023-08-26.

[41] B Roter, N Ninkovic, and SV Dordevic. Clustering superconductors using unsupervised machine learning. *Physica C: Superconductivity and its Applications*, 598:1354078, 2022.

[42] Colton C. Seegmiller, Sterling G. Baird, Hasan M. Sayeed, and Taylor D. Sparks. Discovering chemically novel, high-temperature superconductors. *Computational Materials Science*, 228(April):112358, 2023.

[43] Alexander E. Siemenn, Zekun Ren, Qianxiao Li, and Tonio Buonassisi. Fast Bayesian optimization of Needle-in-a-Haystack problems using zooming memory-based initialization (ZoMBI). *npj Computational Materials*, 9(1), 2023.

[44] Alexander E Siemenn, Zekun Ren, Qianxiao Li, and Tonio Buonassisi. Fast bayesian optimization of needle-in-a-haystack problems using zooming memory-based initialization (zombi). *npj Computational Materials*, 9(1):79, 2023.

[45] Laurens Sluijterman, Eric Cator, and Tom Heskes. Optimal training of mean variance estimation neural networks. *arXiv preprint arXiv:2302.08875*, 2023.

[46] Taylor D Sparks, Steven K Kauwe, Marcus E Parry, Aria Mansouri Tehrani, and Jakoah Brgoch. Machine learning for structural materials. *Annual Review of Materials Research*, 50:27–48, 2020.

[47] 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.

[48] Aik Rui Tan, Shingo Urata, Samuel Goldman, Johannes CB Dietschreit, and Rafael Gómez-Bombarelli. Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. *arXiv preprint arXiv:2305.01754*, 2023.

[49] Jian Wu and Peter Frazier. The parallel knowledge gradient methods for batch bayesian optimization. *Advances in neural information processing systems*, 29, 2016.

[50] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[51] Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya MD Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 8(20):2100566, 2021.

[52] Yong Zhao, Edirisuriya M. Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu, and Jianjun Hu. Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Computational Materials*, 9(1):38, Mar 2023.

Supplementary Information

6.1 Family Assignment

To explore the effects of data set biases, one must be divide up the data set into groups for tracking purposes. The physics, chemistry, and materials science communities often discuss superconductors in terms of "families," which are groupings of superconducting materials. These delineations are generally made based on a combination of chemical constituents, structure, and understandings and hypotheses for why the materials superconduct. The edges of these families can be fuzzy, with history also playing a role in these classifications. Iron-based superconductors (pnictides) generally contain iron and an element in the pnictide column (column 15) of the periodic table. The Cuprates contain copper and oxygen and generally have perovskite-derived structures with layers of copper oxides. Other perovskites like $BaPb_{1-x}BiO_3$ superconduct and have a perovskite structure, but do not contain copper. They are, therefore, not considered Cuprates but similarities can be noted. Intermetallic superconductors are often denoted by a combination of stoichiometry and structure. Here, too, the boundaries can be somewhat fuzzy. $BaAl_4$-derivatives' structures are clearly different from Heusler, half Heusler, and TMX[1] intermetallic subfamilies, which is clear when one compares their stoichiometries and structures. Due to the sheer number of families, we chose to divide the data set into Cuprate, Pnictide, Boride, $BaAl_4$-derivatives, other high T_c (define here to be $T_c > 5$K) and other low T_c families ($T_c \leq 5$K) for reporting. These were chosen because they had the largest number of members of distinct groups. The inclusion of the $LaRu_2Si_2$ subfamily within the $BaAl_4$-derivatives family despite its different structure was an acknowledgement that its related stoichiometry would make it difficult to differentiate with a composition-only model. There is some overlap between iron-based superconductors and $BaAl_4$-derivatives, particularly with the $ThCr_2Si_2$ structure type.

The SuperCon database used for this study was originally created around 1990 for the study of oxide (primarily Cuprate) superconductors [2]. It has since been expanded beyond oxide superconductors and is maintained and updated by the Japanese National Institute for Materials Science (NIMS). Later, iron pnictide superconductors were discovered and rigorously investigated by NIMS researcher Hideo Hosono and his team so the database also contains a large number of entries from this material family. Family (`str3`) is one of the fields in the SuperCon database, although that field is not always populated. In some cases, the family, as written, is very specific (e.g. Y123, a Cuprate) and, in others, the family was very broad (i.e. intermetallic). To allow these families to be visualized practically, the list of families was further condensed to 45 families including those that were listed as "NON_UNIQUE" or were not mapped. Some of these sub-families were much larger than others. We looked at a sampling of the members of these sub-families when deciding how to group them. Initially, the four most common families include Cuprates (30%), iron-based (6%), $BaAl_4$-derivatives (5%), and Borides (4%) (Other=26%). 28% of the database was not mapped, although we could infer a number of the groupings used in this study. The rules we used for inferring unmapped families are as follows:

- **Borides**: The composition contains at least two distinct elements, one of which is B.
- **Cuprates**: The composition contains both O and Cu.
- **Iron-based pnictide**: The composition contains Fe and at least one of: N, P, As, Sb, Mc, Se.

Table 3 shows the `str3` to family mappings that were inferred in this manner. In addition, there were 1329 database entries without a valid `str3` whose family assignment was inferred, for a total of 2462 inferred family designations.

These lists, displaying the sub-family groupings, are given in Tables 2 and 3. Data that were not members of the $BaAl_4$-derivatives, Borides, Cuprates, or iron-based groups and could not be inferred to be a member of those groups were separated into Other High T_c and Other Low T_c groups based on whether their T_c exceeded 5 K.

Note that these are typeset to match the actual `str3` string in the database without capitalization corrections. The values in `str3` were a mix of different notations. Formatting changes are limited to subscripting to enhance readability. Many values of `str3` were the names of compounds, signifying a grouping with the same structure of the parent (e.g. Sr_2CuO_3). Occasionally space groups were

[1] A ternary intermetallic compound of the form T-M-X where T is a transition metal, a rare earth or an alkaline earth metal, M is an element from the first line of the transition metals, and X is a metalloid.

listed (e.g. $P4/nmm$, $F222$). For Cuprates, a different notation was often used consisting of an element followed by four numbers. The first denotes the number of insulating layers between adjacent conducting blocks, the second represents the number of spacing layers between identical CuO_2 blocks. The third gives the number of layers that separate adjacent CuO_2 planes within the conducting block. The fourth gives the number of CuO_2 planes within a conducting block. A component element is also often listed (often a dopant). Other Cuprates (particularly 3-digit families) are grouped by stoichiometry. For example, Cuprate 112 denotes a $(A_B)CuO_2$ stoichiometry. Similar nomenclature is sometimes used for the iron-based superconductors. RT_4B_4 describes the stoichiometry of a family where R is a rare earth element, T is a transition metal element or elements, and B is boron. When we inspected entries containing F222 and Fmmm (space groups) as labels, we noticed that all of these were Borides.

6.2 RooSt Hyperparameters

Table 4 shows the RooSt model hyperparameters used in this study. These correspond the the "reference" model of [12].

Test Set Composition

Table 5 shows the composition of the held-out test set at $t = 0$ for the ten randomized trials considered in this paper.

Table 2: A list of the `str3` subfamilies making up the each family used in this study. Note that these are typeset to match the actual str3 string in the database without formatting or capitalization corrections.

BaAl4-derivatives				
$BaNiSn_3$	$CaBe_2Ge_2$	$CeCu_2Si_2$	$LaRu_2Si_2$	$ThCr_2Si_2$
Borides				
AlB_2	CaB_6	$CeCo_3B_2$	$CePt_3B$	CrB
F222	FeB	Fmmm	$LuRuB_2$	Mo_2B_5
Mo_2IrB_2	MoB	RT_4B_4	RuB_2	SrPtSb
Ta_3B_4	$ThMoB_4$	UB_{12}	$YCrB_4$	YOs_3B_2
Cuprate				
(Cu,C)1223	(Cu,C)1234	(Cu,C)1245	(Cu,Mo)1252	(Cu,Mo)1262
1113	112	1201	1212	1222
201	2112	2116	212	2125
2126	214S	2212	2216	223
234	245	3137	336	4334
446	Ag1223	Ag1234	Ag1245	Ag1256
Al1212	Al1223	Al1234	Al1245	B1223
B1234	B1245	Bi0212	Bi1112	Bi1212
Bi1232	Bi2212	Bi2222	Bi2223	Bi2224
Bi2234	Bi2268	Bi4334	C1201	C1212
C1223	Ca1223	CaCuO2	Cd1212	Co1222
Cu1201	Cu1222	Cu1223	Cu1232	Cu1234
Cu1242	Cu1245	Cu1256	Cu1267	Cu1278
Cu1289	Cu2323	Cu2334	CuC1223	CuC1224
Fe1212	Ga1212	Ga1223	Ga1234	Ga2434
Hg,Re-1223	Hg1201	Hg1212	Hg1222	Hg1223
Hg1232	Hg1234	Hg1245	Hg1256	Hg1267
Hg12nm	Hg2201	Hg2223	Hg2234	Hg2245
IL	LAD	La1113	La3137	LaBiSeSF
$LaRu_3Si_2$	M1201	M1222	M1223	M1234
M1245	M1256	M1267	M2212	Ml1222
Nb1212	Nb1222	Nb123	$\mathrm{P2<1>/m}$	P4/nmm
Pb1201	Pb1212	Pb1213	Pb1222	Pb1223
Pb2212	Pb3201	Pb3212	Pb3222	Pb3232
Pb3252	Pr124	Pr247	RCSCNO	RPr123
RSCNO	RSNCO	Ru2122	Sr0201	Sr0212
Sr_2CuO_3	$SrCuO_3$	T	T'	T*
Ti2322	Tl1201	Tl1222	Tl1234	Tl1245
Tl1302	Tl212	Tl2201	Tl2202	Tl2213
Tl2234	Tl2324	Tl5526	Y123	Y143
Y184	Y211	Y223	Y358	
Iron-based				
21113	$AlFe_3Te_3$	$BaTi_2Sb_2O$	CaRb1144	$CeCr_2Si_2C$
Fe-112	FeAs	$K_2Cr_3As_3$	NiAs	PbFCl
oxypnictide				

Table 3: Inferred mappings. Note that not every materials having the associated `str3` was mapped to this family; the "Count" column indicates the number of materials for which the family was inferred (total of 1133). Note that these are typeset to match the actual str3 string in the database without capitalization corrections.

str3	Family	Count
Mo_5siB_2	Borides	1
NaCl	Borides	5
diamond	Borides	3
"T "	Cuprate	2
Bi2201	Cuprate	105
Cu1212	Cuprate	93
Hg2212	Cuprate	11
IK	Cuprate	1
K_2NiO_4	Cuprate	1
M1212	Cuprate	34
Ru1212	Cuprate	102
Ru1222	Cuprate	29
Tl1212	Cuprate	107
Tl1223	Cuprate	15
Tl2212	Cuprate	42
Tl2223	Cuprate	31
Y124	Cuprate	430
Y247	Cuprate	54
22325	Pnictide	1
22426	Pnictide	8
22438	Pnictide	4
$BaAl_4$	Pnictide	8
FA22426	Pnictide	1
FA22438	Pnictide	1
FA2254(11)	Pnictide	1
FA2264(12)	Pnictide	1
FA2286(18)	Pnictide	1
Fe-32522	Pnictide	2
Fe-42622	Pnictide	9
$ThCr_2Si_2$-type	Pnictide	1
b-$SrRh_2As_2$	Pnictide	10
tetragonal	Pnictide	19

Table 4: Key hyperparameters for RooSt model used in this study, along with model parameter counts. Inputs flow from the "Initial Embedding" downstream through the "Property Prediction Network".

Initial Embedding (12663 parameters)	
element initial embedding size	200
element feature length	64
Element Message Passing Network (744786 parameters)	
layers	3
attention heads per layer	3
Material Message Passing Network (149958 parameters)	
layers	3
attention heads per layer	3
Fixed-length Representation Network (1312512 parameters)	
layers	2
hidden units per layer	1024, 512
Property Prediction Network (82177 parameters)	
layers	3
hidden units per layer	256, 128, 64

Table 5: Test set composition for each of our ten initial conditions, where each row indicates the percentage of the test set corresponding to each family. These conditions reflect variation in both the neural network initialization as well as the specific data points included in the test set. For the "High Tc" and "Low Tc" aggregate families we also enforced that the train/test splits respect the original, more granular family assignments. Therefore, the fraction of the test set they comprise varies slightly depending upon the cardinality of the families assigned to test. The "Assumed T_c=0" denotes the contributions from Materials Project (the presumed negative examples). In all scenarios, the known superconductors represent less than 1% of the test set.

seed	Assumed Tc=0	$BaAl_4$	Borides	Cuprate	Pnictide	Low Tc	High Tc
0	99.67	0.04	0.04	0.04	0.04	0.10	0.10
1000	99.60	0.04	0.04	0.04	0.04	0.17	0.09
2000	99.67	0.04	0.04	0.04	0.04	0.10	0.09
3000	99.54	0.04	0.04	0.04	0.04	0.21	0.11
4000	99.56	0.04	0.04	0.04	0.04	0.26	0.04
5000	99.62	0.04	0.04	0.04	0.04	0.16	0.08
6000	99.56	0.04	0.04	0.04	0.04	0.24	0.06
7000	99.71	0.04	0.04	0.04	0.04	0.11	0.04
8000	99.65	0.04	0.04	0.04	0.04	0.14	0.06
9000	99.68	0.04	0.04	0.04	0.04	0.08	0.09