


ARTICLE

<https://doi.org/10.1038/s41467-019-10030-5>

OPEN

Network analysis of synthesizable materials discovery

Muratahan Aykol¹ , Vinay I. Hegde², Linda Hung¹, Santosh Suram¹, Patrick Herring¹, Chris Wolverton² & Jens S. Hummelshøj¹

Assessing the synthesizability of inorganic materials is a grand challenge for accelerating their discovery using computations. Synthesis of a material is a complex process that depends not only on its thermodynamic stability with respect to others, but also on factors from kinetics, to advances in synthesis techniques, to the availability of precursors. This complexity makes the development of a general theory or first-principles approach to synthesizability currently impractical. Here we show how an alternative pathway to predicting synthesizability emerges from the dynamics of the materials stability network: a scale-free network constructed by combining the convex free-energy surface of inorganic materials computed by high-throughput density functional theory and their experimental discovery timelines extracted from citations. The time-evolution of the underlying network properties allows us to use machine-learning to predict the likelihood that hypothetical, computer-generated materials will be amenable to successful experimental synthesis.

¹Toyota Research Institute, Los Altos, CA 94022, USA. ²Northwestern University, Evanston, IL 60208, USA. Correspondence and requests for materials should be addressed to M.A. (email: murat.aykol@tri.global)

Synthesis prediction for inorganic materials remains one of the major challenges in accelerating materials discovery^{1–4}, mostly because the complexity of the synthesis process itself hinders the development of a general, first-principles approach to it^{3,5}. Thermodynamic stability is one of the main factors that strongly influence synthesizability of a material, but extracting it requires the knowledge of the energetics of competing phases. This bottleneck has recently been addressed for inorganic materials by high-throughput (HT) density functional theory (DFT) databases^{6–9}, which provide access to systematic DFT calculations of thousands of existing inorganic materials as well as hypothetical ones. These databases allow the construction of a comprehensive energy convex-hull: the multidimensional surface formed by the lowest energy combination of all phases. Phases that are on the convex-hull are thermodynamically stable, and tie-lines connecting two phases indicate two-phase equilibria. Given that it is composed of stable materials (nodes) connected by tie-lines (edges), the convex-hull is a naturally occurring thermodynamic network (Fig. 1), analogous to the world-wide-web, social, citation, and protein networks^{10–14}. The information encoded in this new network of materials can be harnessed with the tools provided by the emerging paradigm of network science, and forms the basis of new data-driven models for outstanding materials challenges, such as predicting synthesizability.

The chronology of discoveries can reveal the dynamics of this network of materials. The discovery of a material can be associated with the physical identification and recording of a new crystal structure and chemistry for a target application or general scientific exploration. With this definition, to be traceable as discovered, a material should (i) physically exist, i.e., be amenable to synthesis or occur in nature, and (ii) have a record of structural characterization that can serve as a footprint for the onset of scientific interest. Both of these criteria can be traced from crystallographic databases^{15,16}, which are dominated by structures of existing materials resolved with diffraction experiments. Assuming the time lag between the actual synthesis and/or characterization and the publication is not significant, the time of discovery of a material, and in particular the implied successful synthesis, can be approximated to be the earliest cited reference available in such collections (see the “Methods” section).

The thermodynamic information encoded in the convex-hull is important but not sufficient to explain the successful synthesis and discovery of a material⁴. On the other hand, the collective influence of all complex factors on synthesizability is already reflected in the measured ground truth: whether a material was synthesized or not. Thus, when combined with the historical records on the time of discovery of existing materials, the dynamics of the resulting temporal stability network encodes also the *circumstantial* information beyond thermodynamics that influences discovery. Such information implicitly includes scientific and nonscientific effects almost impossible to capture otherwise at this scale, such as the availability of kinetically favorable pathways, development of new synthesis techniques, availability of new precursors, changes in interest or experience of researchers in a particular chemistry, structure, or application, and even changes in policies that influence research directions. Here we combine the stability information from HT-DFT with the citation-extracted discovery timeline, both available in the open quantum materials database (OQMD)^{6,7}, and determine the temporal evolution of the stability network, as more materials are discovered and added to it. Using the extracted network properties of materials, we demonstrate how a model can be developed to estimate the likelihood of synthesis of new, computationally predicted stable materials.

Results

The materials stability network and its time evolution. The complete network formed by the current convex-hull in the chemical space of all elements is extremely dense with 41 million tie-lines¹⁷. To find the most relevant set of tie-lines for synthesis, we subsample this network to obtain those that control the stability of at least one material, i.e., those in chemical subspaces where there is at least one stable material inside the composition simplex (see “Methods”). This process yields an informative subset of $\sim 2 \times 10^5$ tie-lines for synthesis, that is also computationally tractable for repeated analysis, essential for building a predictive model as described later. Hereafter, we refer to this subset as the materials stability network to differentiate it from the complete network. We then trace retrospectively how this network was uncovered over time until it reached its present state. The number of stable materials discovered, N , and the number of tie-lines defining their equilibria as described above, E , are both growing with time (Fig. 2a and Supplementary Fig. 1). A polynomial fit to $N(t)$ shows that the number of stable materials discovered by year 2025 will reach $\sim 27 \times 10^3$ from the present number of $\sim 22 \times 10^3$. The rate of stable materials discovery is $\sim 400 \text{ year}^{-1}$ today and projected to reach $\sim 540 \text{ year}^{-1}$ by 2025, suggesting that the discovery of stable materials is accelerating. E is increasing faster than N (Fig. 2b), with $\alpha \approx 1.04$ in the densification power-law $E(t) \sim N(t)^\alpha$ ¹⁸. Thus, the materials stability network is getting denser, which may be explained by researchers discovering materials closely connected with those already known, using the latter as stepping stones for the synthesis of new ones^{14,18}, while uncovering the underlying ultimate network¹⁹.

The degree distribution, $p(k)$, where k is the degree of each node, is one measure of the topology of networks. Here k corresponds to the number of tie-lines a material has. In recent years, scale-free networks that obey a power-law distribution, $p(k) \sim k^{-\gamma}$, have received significant attention¹³. While the materials stability network is far from a power-law in early times (e.g., 1960s), it has evolved into a distribution close to it, as shown in Fig. 3 for 2010 (Supplementary Tables 1 and 2 and Supplementary Fig. 2). The exponent γ becomes constant at 2.6 ± 0.1 after the 1980s (Supplementary Fig. 3), within the range $2 < \gamma < 3$ as the other scale-free networks like the world-wide-web or collaborations^{12,20}.

This scale-free character hints at the presence of hubs with significantly larger k compared with other nodes and a robust network connectivity²¹, implying that materials missing randomly from the network (because they have not been discovered yet) are not expected to hinder the discovery of others. However, if there are missing hubs¹³, new material classes disconnected from the present network may be awaiting discovery. The biggest hub here is O_2 with nearly 2600 tie-lines, followed by Cu, H_2O , H_2 , C, and Ge with more than 1100 tie-lines. Elemental N_2 , Ag, Si, Fe, Se, Mn, Co, K, Te, and Bi and oxides BaO, CaO, Li_2O , SrO, Cu_2O , MgO, SiO_2 , La_2O_3 , Al_2O_3 , CuO, MnO, ZnO, Y_2O_3 , Nd_2O_3 , Mn_3O_4 , Sc_2O_3 , Gd_2O_3 , Mn_2O_3 , FeO, Fe_2O_3 , Cr_2O_3 , NiO, BeO, V_2O_3 , and VO_2 are densely connected with 350 or more tie-lines. These are the species that play a dominant role in determining stabilities, and subsequently influencing synthesis of many materials, whether as starting materials, decomposition products of precursors, or simply as competing phases.

Analysis of the discovery timelines indicates that the number of new stable oxygen-bearing materials has been increasing exponentially and faster than all other chemistries, which is in line with the observed predominance of oxides as hubs, and also correlated with the historically high average degree of O-bearing materials (Supplementary Fig. 4). These trends follow the widely accepted Barabasi–Albert model for the growth of scale-free networks^{20,22}, where a small difference in the node degrees in the

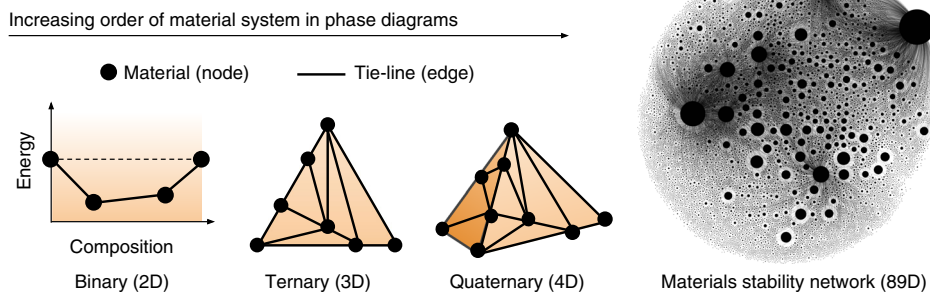


Fig. 1 Network representation of material phase diagrams. The schematic illustrates phase diagrams with the order of the system ranging from two-dimensional binary to the 89-dimensional materials stability network central to this work. The energy-composition convex-hull is shown for the binary system, and all higher-order phase diagrams are projections of their respective N -dimensional convex-hulls to two dimensions, where materials are represented as nodes and tie-lines as edges. For clarity, only those tie-lines connected to high-degree nodes are shown in the materials stability network, where the sizes of the nodes are also scaled to reflect their degree

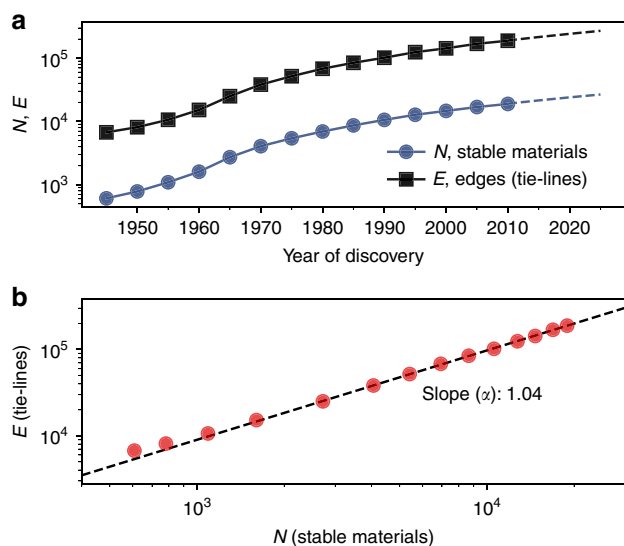


Fig. 2 Evolution of the size of the materials stability network. **a** Time evolution of the number of stable materials (i.e., nodes), N , and tie-lines (i.e., edges), E , and **(b)** how the number of nodes and tie-lines vary with respect to each other. A tie-line is included in the evolving network only after both nodes it is connecting to are identified as discovered. Dashed lines in **(a)** are extrapolations of N and E from the available data (markers and solid lines) by fitted quadratic polynomials. Dashed line in **(b)** is a linear fit to the data (circles). Fits performed in both panels exclude the first four times steps to obtain fits that are more representative of more recent times. A plot of the number of stable materials discovered each year as a function of time is also available in Supplementary Fig. 1

early days gets drastically amplified over time, because of the preferential attachment of new nodes to higher-degree nodes. These results also indicate that identifying new hubs in chemistries, such as pnictides, chalcogenides, halides, or carbides, may accelerate discovery in those spaces. To corroborate this hypothesis further, we compared several such chemistries with oxygen (Supplementary Fig. 4) and observed that as more materials with a hub-like character emerged among the phosphorus-bearing materials in the 1960s (as reflected in their average degree), the discovery in this space accelerated, with a notable upsurge in the number of new P-bearing stable materials in later years.

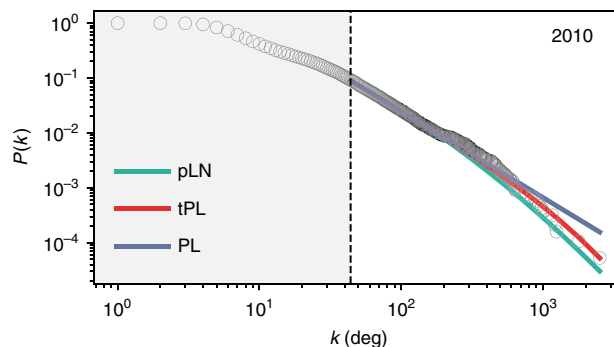


Fig. 3 Degree distribution among stable materials discovered by the year 2010. The complementary cumulative distribution function ($P(k)$) of the degree distribution $p(k)$ of stable materials (circles) is plotted along with the fitted distributions (solid lines). Each point $P(k)$ represents the probability that a material has greater than k tie-lines connected to it in the network. Power-law, truncated power-law (with exponential cutoff), and positive log-normal distributions are labeled as PL, tPL, and pLN, respectively. The dashed line shows k_{\min} , the lowest degree used in fitting. Degree distributions of other times are shown in Supplementary Fig. 1

Prediction of materials discovery from network dynamics.

While the evolution of global properties of the network is slow (Supplementary Fig. 5), the network properties of individual nodes are evolving rapidly as their local environments change, as exemplified in Fig. 4a, b for a high-temperature superconductor $\text{YBa}_2\text{Cu}_3\text{O}_6$ and high-ZT thermoelectric BiCuSeO . Since this temporal evolution encodes circumstantial factors beyond thermodynamics that may contribute to discovery (and synthesis), properties that characterize the state of a material in relation to the rest of the network can realize a connection between these explicit or implicit factors and its discovery.

To reproduce that connection, we turn to designing a machine-learning model based on the network properties of materials, which we will then use to predict likelihood of synthesis of hypothetical materials: those created on the computer but have never been made. The present stability network has about 22,600 materials, of which $\sim 19,200$ are physically existing materials from crystallography databases and can be used in model building, and ~ 3400 are hypothetical, generated via HT prototyping^{6,23–25}. Prediction of synthesis likelihoods in the latter category can help

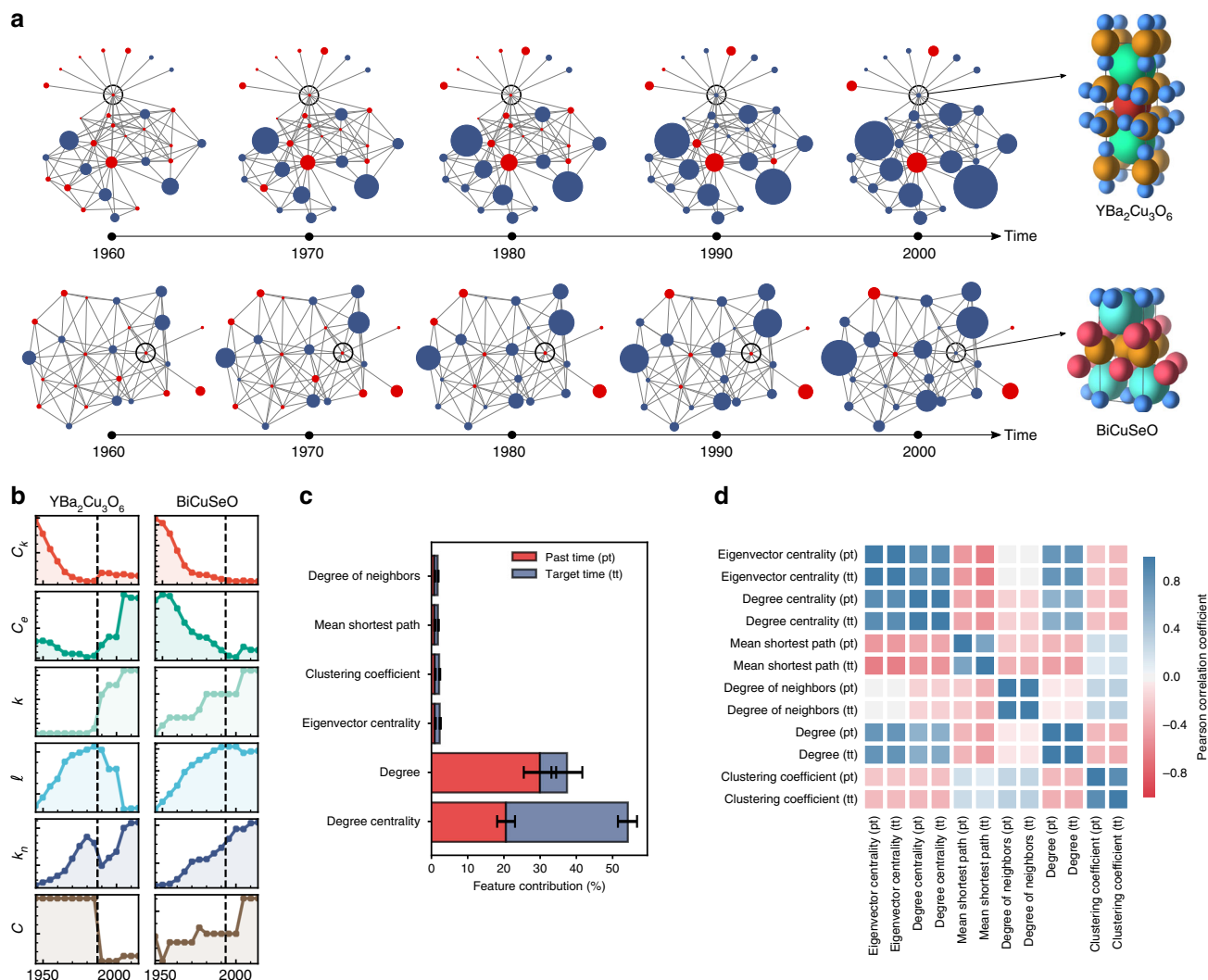


Fig. 4 Network evolution and properties of the machine-learned synthesizability models. **a** Time evolution of the local environments of two sample materials (marked with open circles), superconductor $\text{YBa}_2\text{Cu}_3\text{O}_6$, and thermoelectric BiCuSeO , in the materials stability network. Materials (nodes) discovered by a given temporal state of the network are shown in blue, whereas those awaiting discovery are red. Node size is proportional to degree. **b** Time evolution of the network properties of sample materials $\text{YBa}_2\text{Cu}_3\text{O}_6$ and BiCuSeO , namely, degree and eigenvector centralities (C_k and C_e), degree (k), mean-shortest-path (ℓ), mean degree of neighbors (k_n), and clustering coefficient (C), where the vertical dashed lines show the approximate time of discovery. **c** Feature contributions to the RF model as derived from the Gini importance. **d** Pearson correlation coefficients of time-dependent network properties used in models as features, where pt and tt denote past time and target time, respectively, corresponding to a given sequence of window size of two (see “Methods”). Variables and names of network properties are used interchangeably in (b), (c), and (d)

bridge the gap between computational discovery and the real world.

We use six network properties for each material in model training, namely, degree and eigenvector centralities, degree, mean shortest path length, mean degree of neighbors, and clustering coefficient (Fig. 4b and Supplementary Fig. 6). Degree and eigenvector centralities reflect the relative importance of a node in influencing stabilities, emphasizing the number of connections and importance of neighbors, respectively. We normalize these metrics such that they are mostly independent of the size of the network²⁶. We find that degree without normalization is also useful for capturing the influence of the temporal state of the network on connectivity. Mean shortest path length, the mean of the minimum number of tie-lines to travel from a node to every other node, and mean degree of neighbors serve as a proxy for ease of access to a particular material in synthesis. The clustering coefficient indicates how tightly

connected the neighborhood of a material is and may capture the local environment more immediately related to its synthesis.

Discovery is a time-irreversible event and its prediction is not a standard machine-learning problem in materials science. Here we use time evolution of the aforementioned network properties of materials as features to form the basis of a sequential supervised-learning problem²⁷. We adopt a sliding-window approach to train experimental discovery classifiers and estimate likelihood of discovery, and the implied synthesis for synthetic materials (see Fig. 5 and “Methods”). We employ two classification algorithms: L_2 -norm regularized logistic regression (LR) for well-calibrated probabilities, and random forest (RF) for classification accuracy. The subclass of sequences where a material changes from undiscovered to discovered in the time domain represents the rare event of discovery in sequential data, for which we define stringent event-detection metrics for precision and recall using a prediction-period approach (see “Methods”)²⁸, and found these

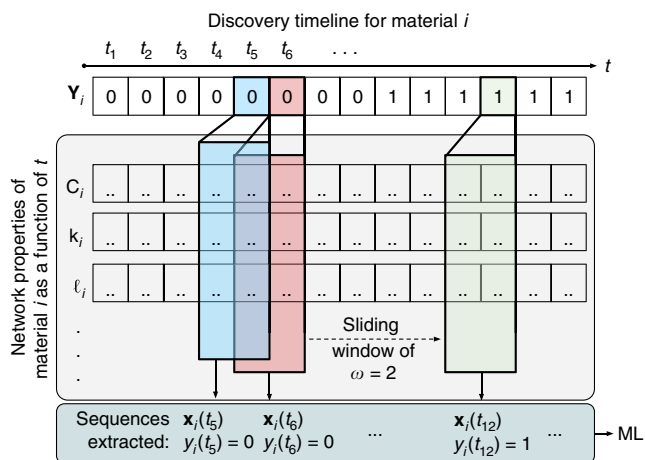


Fig. 5 Extraction of sequences from temporal network property data using a sliding window to use as input for machine learning. The vector \mathbf{Y}_i stores the targets to be learned for material i , i.e., encoding whether i is discovered by a given time-step t or not (as binary labels 1 and 0). C_i , k_i , and l_i are examples for vectors of different network properties, encoding how those properties change over time as the network evolves, as explained in the text. The process of applying a sliding window (here with a width of $w = 2$) to extract sequences of features and targets ($\mathbf{x}_{i,t}$, $y_{i,t}$) is illustrated. ML stands for the machine-learning task of training and testing classification algorithms using the extracted data

metrics for LR and RF to be nearly 30% and 90%, respectively, for detection within ± 1 time-step and close to 50% and 95% for detection within ± 2 time-steps (where larger prediction periods make the correct classification easier), outperforming baseline detection metrics by a significant margin (Supplementary Fig. 7). As another performance evaluation for the present problem, distributions of the difference between the estimated time-step of discovery (see “Methods” for how this estimation is done from classification results) and its true value (Δt) for the two models tested in this work are compared in Supplementary Fig. 8, along with baseline models. For both LR and RF models, distributions are centered close to zero, with LR showing a tendency to estimate slightly earlier times with a mean $\Delta t \approx -1.6$, whereas RF delivers more precise estimates with mean $\Delta t \approx -0.2$ (each time-step is 5-years long in this work). Baseline models yield distributions with means far from zero. LR has a much broader distribution than RF, however, with a standard deviation at ~ 3.5 time-steps, whereas that of RF is at ~ 1.2 time-steps.

To understand how these models provide accurate predictions for synthesizability, we investigate the correlations between the network properties, and how much they contribute to predictions (Fig. 4c, d). Except the eigenvector centrality with a degree or degree centrality, distinct features are not too strongly correlated. Identical features within a time sequence are naturally more correlated (e.g., degrees in a sequence), but distinct enough for the models to utilize them (Supplementary Fig. 9). Confirming the significance of tie-lines in influencing synthesis, degree, and degree centrality, two closely connected but not highly correlated metrics, play the biggest roles in decision-making, with substantial contributions adding up to $\sim 90\%$. The rest of the features still play a non-negligible role, providing the remaining 10%.

Discussion

The trained models can be used in multiple ways, for example, to predict class labels or probabilities for synthesizability in network environments pertaining to the present time or a past time. For

the present time, models predict that about 93% of hypothetical materials in the network have a synthesis probability, $p > 0.5$. This prediction is in line with the notion that stable materials in HT-DFT databases are likely to be more amenable to synthesis. However, synthesis is a costly process and knowing its likelihood of success is critical before an attempt in the laboratory. Using LR’s calibrated probabilities with RF’s more accurate classifications within the intersection of positive classification sets they predict (92% identical), we find that out of ~ 3400 stable hypothetical materials present, only about 10% have $p > 0.95$ for immediate synthesis (Supplementary Note 1).

Our approach can assist the decisions on where to allocate synthesis resources after computational design. For example, Kim et al.²³ performed a computational search and suggested new high-capacity Li_4ABO_6 cathodes for Li-ion batteries (A and B represent different elements), for which we find the likelihoods of synthesis to range from $p = 0.52$ for $\text{Li}_4\text{SbRhO}_6$ to $p = 0.85$ for $\text{Li}_4\text{NiTeO}_6$. In fact, several of those predictions with $p > 0.6$ were synthesized^{29,30}. For the novel ABO_3 perovskites identified in two recent computational studies^{24,31}, we find the synthesis likelihoods to range from $p = 0.54$ for PuGaO_3 to $p \approx 1$ for EuGeO_3 . For several of such predictions with $p > 0.9$, reports of synthesis exist^{32,33}. For the inverse-Heusler alloys uncovered in a HT search for spintronics²⁵, we predict p to be in the range of 0.56 (TiInCo_2)–0.94 (FeGeRu_2) (a complete list of probabilities is available in Supplementary Table 3). Today, such computational studies can rapidly identify hundreds of new hypothetical materials with target functionalities, but the cost and complexity of synthesis often hinders systematic attempts for their realization. The ability to predict synthesis likelihoods is expected to bridge this gap between computational and experimental research groups.

The network-based models can also be used to invert the discovery predictions and find at what point in time a hypothetical material could have already been made. Based on the RF model, we estimate that only about 10% of the stable hypothetical materials that are predicted to be synthesizable today were synthesizable by 2005, and only about 30% were synthesizable by 2010 (Supplementary Fig. 10), implying that the progress within the last 10–15-year period has improved their chances of synthesis.

Similarly, since most of the materials discovered in the last few decades have likely been made with contemporary synthesis methods developed or improved since the mid-20th century (from sol-gel to advanced deposition methods to new precursors), models trained only with earlier discoveries should intuitively predict a majority of the newer materials as unlikely to be made in the distant past. Indeed, a model trained only with discoveries up until and including year 2000, yields probability distributions that consistently shift to higher values with time for materials discovered after year 2000 (unseen to model) as shown in Supplementary Fig. 11, confirming that the predictions agree with our intuition. Besides, almost all materials that were discovered after 2000 are predicted to have $p > 0.4$ in year 2000 with this model, i.e., materials with $p < 0.4$ in 2000 were rarely synthesized after 2000. However, the application of the models to new materials assumes that the mechanisms of materials discovery continue to follow similar trends to those in the past and present, and therefore by design, the models cannot forecast the future. For instance, we observe that the probability distributions predicted by the above model trained with material discoveries until 2000 cannot clearly differentiate between progression of most materials discovered in its future (except a fraction of materials near $p \approx 1$, where the predictions look correlated with the discovery timeline), confirming future timeline forecasts cannot be made for most of the materials (Supplementary

Fig. 11). Ultimately, the decisions on which materials to make are made by the scientists and the future is shaped accordingly. The models merely provide statistical predictions based on the latest network data they are exposed to, within the limits of their underlying approximations.

Given the advances in materials discovery techniques, including complex and HT experimental or simulation capabilities, intuitively, the present models are likely a lower bound for the future synthesis likelihoods, as long as the nonscientific factors, such as the science policies and funding, remain sustainable. Demonstrating how network science and machine learning can be combined to build predictive methods for materials, we expect the present work to pave the way for new, improved methods for materials discovery, possibly addressing synthesizability in the unbounded space of metastable materials (which would require constructing linkage rules beyond the convex-hull), or examining applications beyond synthesis.

Methods

Network data and analysis. The network presented in this work is constructed from the energy-composition convex-hull of OQMD, which is a collection of systematic DFT calculations of inorganic crystalline materials, and subsequent properties derived from them, such as formation energies^{6,7}. DFT is known to provide a good compromise between accuracy, especially in terms of determining relative stabilities of materials, and computational cost, and is the current state-of-the-art for first-principles HT computations of materials^{4,34,35}. We used the version 1.1 of the OQMD data available at <http://oqmd.org>.

The NetworkX package was used for the calculation of the network properties³⁶. The maximum-likelihood method was used to fit the distributions and the k_{\min} values were found by minimizing the Kolmogorov–Smirnov distance^{37,38}. The *powerlaw* library was used in fitting the distributions³⁸. Goodness-of-fit comparisons of different distributions are available in Supplementary Table 1. The method for subsampling of the complete network to obtain the materials stability network is further described in Supplementary Methods.

Model construction. To prepare the input vectors for training the machine-learning models, we create multiple sequential training examples $(\mathbf{x}_{i,t}, y_{i,t})$ for each material i from its temporal data, where feature vector for time t , $\mathbf{x}_{i,t}$ extends to features for the past times within a window w , and where the target $y_{i,t}$ encodes binary labels 1 and 0, respectively, indicating whether a material was discovered at that point in time or not (Fig. 5). We adopted $w = 2$ for the present work. We analyzed the networks with 5-year increments starting from 1945, and found that a window of width $w = 2$ (i.e., encompassing 10 years) provides sufficient prediction accuracy, without any need for recurrence (i.e., including past y as part of \mathbf{x}). Network properties of a material pertaining to the times when it was undiscovered are calculated by hypothetical, individual insertion of that material into the materials stability network (as if it existed at that point in time). Further details of each step in model creation can be found in Supplementary Methods. Discovery times of known materials are approximated by their earliest dated reference for their structures reported in the ICSD¹⁵, except for the elemental references, which are defined as discovered ($y = 1$) at all times. We expect this approximation to (i) reasonably hold, given that the coarse-enough discretization of the timeline (5 years) would already account for the typical delays between characterization and publication (e.g., 1–2 years), and (ii) not significantly affect the model training, as the delays are likely to be in the form of a nearly constant shift for all materials, as one might intuitively expect the distribution of the delays to be narrow and centered around 1–2 years at most.

Model training and evaluation. Model training and parts of the evaluation were performed using scikit-learn³⁹. In training the LR models, we use a larger weight (~ 2.5 times) for the minority subclass of $y = [0,1]$ (i.e., a material transitioning from undiscovered to discovered), compared with the other subclasses to obtain evenly distributed accuracies across all subclasses. RF models use 200 estimators. Models were also tested against baseline classifiers, including class distribution prediction, majority and/or constant class prediction, and random classifiers, and found to outperform all. Feature importance in the RF model is calculated as the Gini importance. Calibration is applied to model probabilities; however, given the approximate nature of this process and the variability in absolute values of the resulting probabilities, probabilities should be considered mostly to reflect relative likelihoods among materials. We use five-fold cross-validation (CV) for the evaluation of all models.

We follow event-based strategies for model evaluation that consider the entire timeline of the materials and test/train splitting of sequences $(\mathbf{x}_{i,t}, y_{i,t})$ (where i is a material and t is the time-step) is accordingly performed at the material level. The standard model evaluation metrics in classification are defined as precision = TP/(TP + FP), recall = TP/(TP + FN), and F1-score = $2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$, where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. While these metrics help evaluate the performance of algorithms in classifying targets as 0 or 1 in a standard way (and are all above 90% and 70%, respectively, for RF and LR models in material-level splitting and CV), for the detection of the discovery itself (i.e., the transition from 0 to 1), which can be described as a rare-event detection in sequential data, modified definitions for precision and/or recall are more suitable²⁸. In the present event-based strategy, the model is assumed to predict a discovery event at the very first positive prediction it makes in the timeline of that material, and metrics like TP and FP count whether the target event is captured at the correct time or not, and precision is calculated accordingly with the same formula as above. Recall becomes the fraction of target events correctly captured: TP/(total-number-of-discovery-events). These metrics, however, would equally weight FP's made one step away from the ground truth, vs., for example, five steps away from it, where the latter case can be considered worse from a practical point of view. To partially address this issue, we employ a prediction-period concept²⁸, where a \pm time range is defined, such that a discovery prediction would be considered TP if the correct discovery time falls in that range, or FP if it falls outside. These metrics then become a function of the size of the prediction period, as we show in Supplementary Fig. 7. Another approach for model evaluation is the direct comparison of the difference between the predicted discovery times and their actual values, as shown in Supplementary Fig. 11, where again the model is assumed to make the discovery prediction at the very first positive classification it predicts for a material.

+

Data availability

Data used in this work are provided as Supporting Data and can also be accessed at <https://data.matr.io/2>.

Received: 26 September 2018 Accepted: 8 April 2019

Published online: 01 May 2019

References

- Hemminger, J. C., Sarrao, J., Crabtree, G., Flemming, G. and Ratner, M., *Challenges at the Frontiers of Matter and Energy: Transformative Opportunities for Discovery Science*. Technical Report (USDOE Office of Science (SC), United States, 2015).
- Sun, W. et al. The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2**, e1600225–e1600225 (2016).
- Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
- Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **4**, eaaq0148 (2018).
- Alberi, K., et al. The 2018 materials by design roadmap. *J. Phys. D: Appl. Phys.* **52**, 013001 (2018).
- Kirklin, S. et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
- Saal, J. E. et al. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
- Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
- Curtarolo, S. et al. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
- Barabási, A.-L. The physics of the Web. *Phys. World* **14**, 33–38 (2001).
- Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA* **98**, 404–409 (2001).
- Albert, R. & Barabasi, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47 (2002).
- Barabási, A.-L. Scale-free networks: a decade and beyond. *Science* **325**, 412–413 (2009).
- Rzhetsky, A., Foster, J. G., Foster, I. T. & Evans, J. A. Choosing experiments to accelerate collective discovery. *Proc. Natl Acad. Sci. USA* **112**, 14569–14574 (2015).
- Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr. B* **58**, 364–369 (2002).
- Gražulis, S. et al. Crystallography open database (COD) an open-access collection of crystal structures. *J. Appl. Crystallogr.* **42**, 726–729 (2009).
- Hegde, V. I., Aykol, M., Kirklin, S., Wolverton, C., The phase diagram of all inorganic materials. Preprint at <https://arxiv.org/abs/1808.10869> (2018).
- Leskovec, J., Kleinberg, J., Faloutsos, C. Graph evolution: densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* **1**, 1–41 (2007). <https://doi.acm.org/10.1145/1217299.1217301>

19. Pedarsani, P., Figueiredo, D. R. & Grossglauser, M. Densification arising from sampling fixed graphs, ACM SIGMETRICS performance. *Eval. Rev.* **36**, 205 (2008).
20. Barabási, A.-L. The scale free property. In *Network Science*. Ch. 4, 112–163 (Cambridge University Press, Cambridge, 2016).
21. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
22. Barabasi, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
23. Kim, S. et al. Material design of high-capacity Li-rich layered-oxide electrodes: Li_2MnO_3 and beyond. *Energy Environ. Sci.* **10**, 2201–2211 (2017).
24. Emery, A. A., Saal, J. E., Kirklin, S., Hegde, V. I. & Wolverton, C. High-throughput computational screening of perovskites for thermochemical water splitting applications. *Chem. Mater.* **28**, 5621–5634 (2016).
25. Ma J., et al., Computational investigation of inverse-Heusler compounds for spintronics applications. *Phys. Rev. B* **98** 094410 (2018). <https://doi.org/10.1103/PhysRevB.98.094410>
26. Ruhnau, B. Eigenvector-centrality—a node-centrality. *Soc. Netw.* **22**, 357–365 (2000).
27. Dietterich, T. G. Machine learning for sequential data: a review. *Struct., Syntactic, Stat. Pattern Recognit., Jt. IAPR Int. Workshops SSPR2002 SPR 2002* **2396**, 15–30 (2002).
28. Weiss G. M. and Hirsh H., Learning to predict rare events in event sequences, in *KDD'98 Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (AAAI Press, 1998) pp. 359–363.
29. Bhardwaj, N., Gupta, A. & Uma, S. Evidence of cationic mixing and ordering in the honeycomb layer of Li_4MSbO_6 (M(iii) = Cr, Mn, Al, Ga) (S.G. C2/c) oxides. *Dalton Trans.* **43**, 12050–12057 (2014).
30. Sathiy, M. et al. Tarascon, $\text{Li}_4\text{NiTeO}_6$ as a positive electrode for Li-ion batteries. *Chem. Commun.* **49**, 11376–11378 (2013).
31. Balachandran, P. V. et al. Predictions of new ABO_3 perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.* **2**, 043802 (2018).
32. Shukla, R., Arya, A. & Tyagi, A. K. Interconversion of perovskite and fluorite structures in Ce-Sc-O system. *Inorg. Chem.* **49**, 1152–1157 (2010).
33. Akamatsu, H. et al. Crystal and electronic structure and magnetic properties of divalent europium perovskite oxides Eu M O_3 (M = Ti, Zr, and Hf): experimental and first-principles approaches. *Inorg. Chem.* **51**, 4560–4567 (2012).
34. Hautier, G., Ong, S., Jain, A., Moore, C. & Ceder, G. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Phys. Rev. B* **85**, 155208 (2012).
35. Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. *Science* **351**, aad3000 (2016).
36. Hagberg, A. A., Schult, D. A., Swart, P. J., Exploring network structure, dynamics, and function using Network X. In *Proc. 7th Python in Science Conference (SciPy)* (eds Varoquaux, G., Vaught, T., & Millman, J.) 11–15 (Pasadena, CA, 2008).
37. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-Law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
38. Alstott, J., Ed Bullmore, Plenz, D., Powerlaw: a python package for analysis of heavy-tailed distributions. *PLoS ONE* **9** (2014), 0.1371/journal.pone.0085777.
39. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

V.I.H. acknowledges support from Toyota Research Institute through the Accelerated Materials Design and Discovery program. C.W. acknowledges the support of the National Science Foundation, through the MRSEC program, grant number DMR-1720139.

Author contributions

M.A. conceived and designed the project, with input from J.S.H., S.S., P.H., and L.H. M.A. performed the network analysis and machine learning with input from V.I.H., C.W., S.S., P.H., and L.H. M.A., V.I.H., and C.W. computed the convex-hull. All authors contributed to the discussions of the results. M.A. wrote the paper with input from all authors.

Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-019-10030-5>.

Competing interests: M.A., L.H., S.S., and P.H. filed a patent application on network-based synthesis prediction: US App. No. 16/004,232 on 8 June 2018. The remaining authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Journal peer review information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019