LETTER

doi:10.1038/nature17439

Machine-learning-assisted materials discovery using failed experiments

Paul Raccuglia[1], Katherine C. Elbert[1], Philip D. F. Adler[1], Casey Falk[1], Malia B. Wenny[1], Aurelio Mollo[1], Matthias Zeller[2], Sorelle A. Friedler[1], Joshua Schrier[1] & Alexander J. Norquist[1]

Inorganic–organic hybrid materials[1–3] such as organically templated metal oxides[1], metal–organic frameworks (MOFs)[2] and organohalide perovskites[4] have been studied for decades, and hydrothermal and (non-aqueous) solvothermal syntheses have produced thousands of new materials that collectively contain nearly all the metals in the periodic table[5–9]. Nevertheless, the formation of these compounds is not fully understood, and development of new compounds relies primarily on exploratory syntheses. Simulation- and data-driven approaches (promoted by efforts such as the Materials Genome Initiative[10]) provide an alternative to experimental trial-and-error. Three major strategies are: simulation-based predictions of physical properties (for example, charge mobility[11], photovoltaic properties[12], gas adsorption capacity[13] or lithium-ion intercalation[14]) to identify promising target candidates for synthetic efforts[11,15]; determination of the structure–property relationship from large bodies of experimental data[16,17], enabled by integration with high-throughput synthesis and measurement tools[18]; and clustering on the basis of similar crystallographic structure (for example, zeolite structure classification[19,20] or gas adsorption properties[21]). Here we demonstrate an alternative approach that uses machine-learning algorithms trained on reaction data to predict reaction outcomes for the crystallization of templated vanadium selenites. We used information on 'dark' reactions—failed or unsuccessful hydrothermal syntheses—collected from archived laboratory notebooks from our laboratory, and added physicochemical property descriptions to the raw notebook information using cheminformatics techniques. We used the resulting data to train a machine-learning model to predict reaction success. When carrying out hydrothermal synthesis experiments using previously untested, commercially available organic building blocks, our machine-learning model outperformed traditional human strategies, and successfully predicted conditions for new organically templated inorganic product formation with a success rate of 89 per cent. Inverting the machine-learning model reveals new hypotheses regarding the conditions for successful product formation.

First-principles crystal-structure prediction—even for simple crystallization from a solvent—is fundamentally difficult, owing to the need to consider a combinatorially enormous set of component arrangements[22,23] using high-level quantum chemistry methods[24]. Predicting crystal structures following a chemical reaction—as in the case of hydrothermal and solvothermal synthesis—is even more challenging, because it requires an accurate potential-energy surface for the entire reaction. Instead we pose the potentially tractable question of whether a given set of reaction conditions and reagents will yield any crystal at all. A machine-learning approach to the related problem of whether a particular organic molecule will crystallize has been described previously[25]. Chemists typically posit an 'intuition' about patterns of reagent properties and composition ratios that govern material synthesis. If these patterns exist, then they can be discovered using data-mining techniques, given a database of successful and failed reactions. However, the published literature contains only a limited subset of successful reactions, typically a single set of conditions for each compound. The vast majority of unreported 'dark' (failed) reactions are archived in laboratory notebooks that are generally inaccessible. These reactions contain the valuable information needed to determine the boundaries between success and failure.

To use these data to guide future materials syntheses, we developed a web-accessible public database (http://darkreactions.haverford.edu) to facilitate both initial data entry from existing laboratory notebooks and ongoing experimental data collection. The database schema is sufficiently general to accommodate reaction descriptions beyond our particular chemical interests (for example, allowing for arbitrary numbers of inorganic and organic species, or non-aqueous solvents). We intentionally captured experimental data that might be useful for later studies (for example, product purity labels) to avoid having to re-enter experimental data, even though they were not used in the present study. The data-capture process and reliability testing are described in Methods. After excluding reactions with incomplete laboratory notebook entries, 3,955 unique, complete reactions remained for use in training and testing the machine-learning model.

Reactant names can be used to create property descriptors for our machine-learning model. For organic and oxalate-like reactants, commercially available cheminformatics software was used to compute physicochemical properties of the molecules (for example, molecular weight, number of hydrogen-bond donors/acceptors as a function of pH and polar surface area). For inorganic reactants, tabulated values of atomic properties (for example, ionization potential, electron affinity, electronegativity, hardness and atomic radius) and position on the periodic table were used. Additionally, experimental reaction conditions (for example, temperature, reaction duration and pH) and mole ratios of the different reactants were used (see Methods). A support vector machine (SVM) model was built using this expanded table of reactant properties (see Methods). The single SVM model used to predict experimental results had an accuracy of 78% in describing all of the reaction types in its test-set data, and 79% considering only vanadium-selenite reactions.

Solid-state synthesis projects can be divided into exploration and exploitation stages. Successful exploration reactions reveal new 'islands of stability'—sets of reaction conditions that result in product formation. Success rates during this stage tend to be low, because the general ranges of acceptable parameters needed for successful syntheses are unknown. The boundaries of the island can be mapped by changing the organic reactants. These exploitation reactions expand the range of functional material properties and reveal new insights about organic–inorganic interactions. Success rates during this stage can be high, because the structures and reactivities of the organic molecules can be quite similar, and so changing the organic reactants has a more subtle effect on the chemistry.

A successful model should both increase the rate of synthesis and characterization of new materials and give chemical insight.

[1]Haverford College, 370 Lancaster Avenue, Haverford, Pennsylvania 19041, USA. [2]Department of Chemistry, Purdue University, 560 Oval Drive, West Lafayette, Indiana 47907-2084, USA.

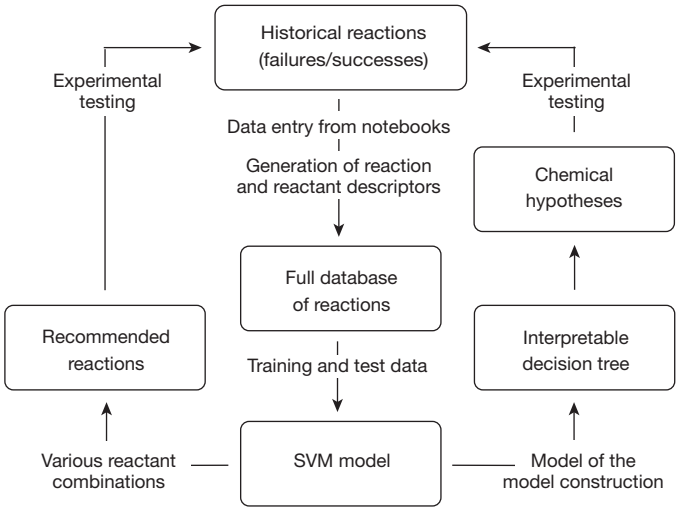

Figure 1 | Schematic representation of the feedback mechanism in the dark reactions project. Machine-learning models generated from historical reaction data are used to recommend new reactions to perform, and to generate human-interpretable hypotheses about crystal formation. SVM, support vector machine.

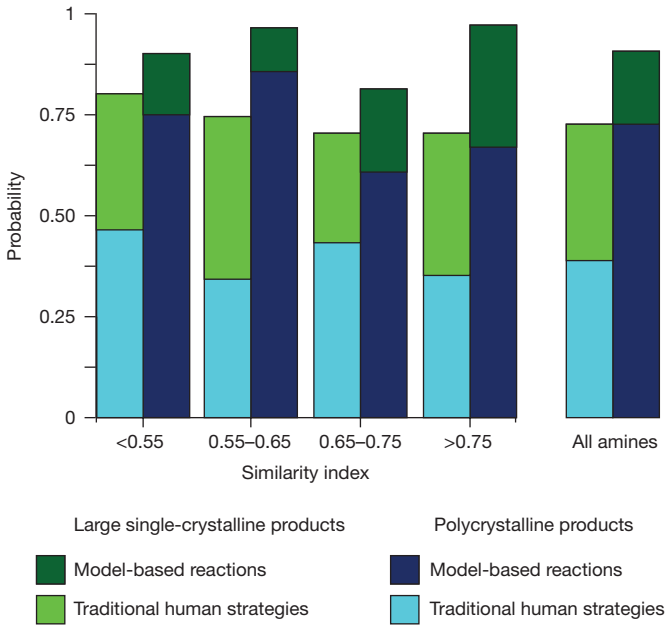

Figure 2 | Comparison of experimental outcomes relating to the formation of templated vanadium-selenite crystals, as a function of amine similarity. Darker coloured bars indicate model predictions; lighter colours bars indicate traditional human strategies. Reactions that yielded polycrystalline and large single-crystalline products are shown in blues and greens, respectively. The vertical axis shows the probably that the reaction had the indicated outcome. The model more successfully predicts conditions for crystal formation than do human strategies, regardless of structural similarity of the templating amines to known examples in the database.

To demonstrate the performance of our model relative to typical strategies of human chemists, we focused on exploitation reactions in templated vanadium selenites, in which a new organic building unit is introduced into a reaction. These reactions allow us to: (i) compare against the experimental decisions of experienced chemists; (ii) obtain higher quality statistical data because exploitation reactions are generally more successful; and (iii) increase understanding about the unusual degree of diversity in connectivity and dimensionality that is observed in these compounds. Though, beyond the scope of this Letter, our model could also be applied to exploration reactions, by computationally sampling possible reaction conditions involving all possible combinations of reactants, predicting successes, and then sorting the reactions by chemical interest. We used a database of commercially available organic compounds to identify 34 new diamines, sampled by structural similarity to the organic reactants already in our database (see Methods). Organically templated metal oxides using these diamines are essentially unknown, as indicated by their near absence from the Cambridge Structural Database[26] (see Methods). These amines were then used to perform human- or model-controlled hydrothermal synthesis reactions (see Methods). A schematic of this approach is shown in Fig. 1.

Reactions recommended by the model had an 89% success rate, as defined by the synthesis of the target compound type in either a polycrystalline or single-crystal form, and success rate was independent of the structural similarity of the amine (see Fig. 2). This exceeds the human intuition success rate of 78%. The difference is statistically sound. Fisher's exact test indicates better-than-chance results for model predictions with $P < 0.01$, and a two-sample proportion test indicates an 8% advantage of the model over human intuition with $P < 0.05$. The 89% success rate of the model in the experimental test is greater than the test-set accuracy measured during model construction, because the train/test split on the historical data essentially tests only exploration reactions (for which the model uncertainty is higher), whereas these experiments test exploitation reactions (for which the model uncertainty is lower).

SVMs are opaque to simple examination. To gain insight we made a 'model of the model' by re-interpreting the original SVM as a decision tree of human-interpretable if–then criteria (see Methods). An abbreviated flow-chart representation is shown in Fig. 3, and a full version of the vanadium-selenite branch of the tree is shown in Supplementary Information. From this flow chart, one can generate chemical hypotheses to guide future experiments. This approach can be applied to any chemical system for which any model exists. Here it yielded three hypotheses about the formation of templated vanadium selenites, categorized by the molecular polarizability of the amine. Representative structures for each hypothesis are shown in Fig. 4. (The model separates inorganic building units by mean Pauling electronegativity; as a consequence, vanadium selenites and molybdates appear in the same subtree. In the discussion below, we consider only the vanadium-selenite reactions contained in the subtree.)

Amines with moderate polarizability (10.29–19.51 Å^3), shown in blue in Fig. 3, require inclusion of a sulfur-containing reactant, specifically here V(IV)OSO_4. (The decision tree incidentally selects these amines by polarizability in the right branch and organic refractivity, that is, molar polarizability, in the left branch.) All but one of the organically templated vanadium selenites in the literature include V^{4+} ions, which must be either introduced as a reagent or generated *in situ* through the concurrent oxidation of the amine and reduction of V^{5+}. These geometrically compact amines seem unable to generate the necessary V^{4+} concentrations from V^{5+} precursors over the timescale of the reaction. This triggers the formation of polycrystalline reaction products that do not contain the organic amines. Using V(IV)OSO_4 circumvents this inability to generate V^{4+}.

Amines with high polarizability (17.64–29.85 Å^3), shown in red in Fig. 3, are not limited by V^{4+} generation, but do require oxalates for success. We hypothesize that oxalates alter the charge density on the inorganic secondary building unit, allowing these long, linear, highly charged tri- and tetramines to achieve charge density matching[3].

Amines with low polarizability (<9.32 Å^3), shown in green in Fig. 3, (for example, ethylenediamine, 1,3-diaminopropane, imidazole and N-methylethylenediamine) have higher pK_a values than the other amines in our database and do not need pH < 3 to be in the correct protonation state. These amines generate sufficient V^{4+} from V^{5+} precursors, but slowly, requiring longer reaction times (>26 h). Use of

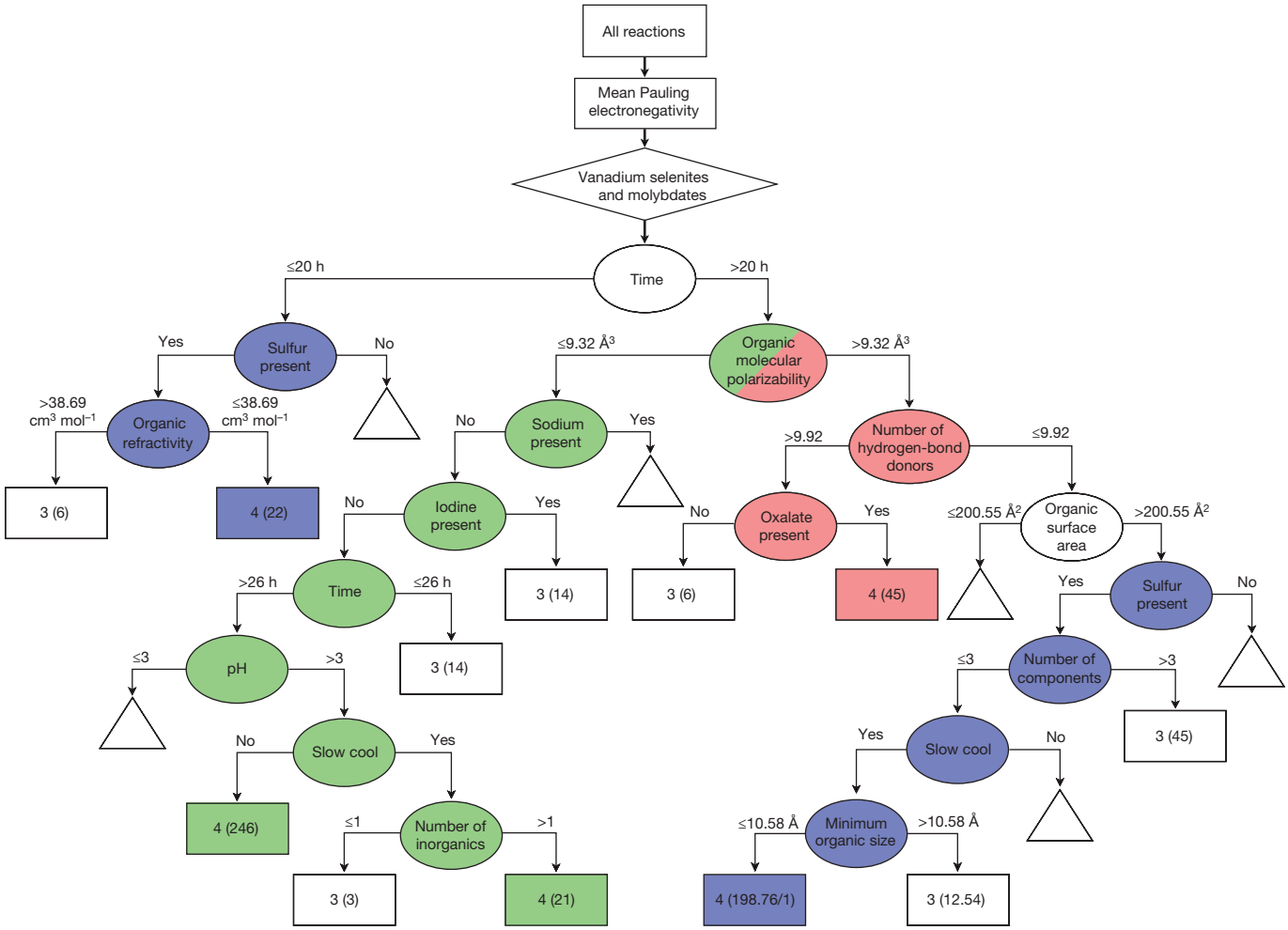

Figure 3 | SVM-derived decision tree. Ovals represent decision nodes, rectangles represent reaction-outcome bins and triangles represent excised subtrees. The numbers on the arrows correspond to decision attribute test values. Each reaction-outcome bin (rectangle) corresponds to a specific reaction-outcome value ('3' or '4', as indicated; see Methods); the number in parentheses is the number of reactions correctly assigned to that bin (any incorrectly classified reactions are given after a slash). Fractional values indicate reactions with an indeterminate result arising from missing attribute values higher in the tree. Bins containing the majority of successful reactions are divided into three distinct groups (indicated by green, blue and red shading). Each coloured subtree defines a specific set of reaction parameters that facilitates single-crystal formation. Inspection of these conditions leads to the corresponding chemical hypotheses, corresponding to low-, medium- and high-polarizability amines, respectively. An expanded version showing all excised subtrees is available in Supplementary Information.

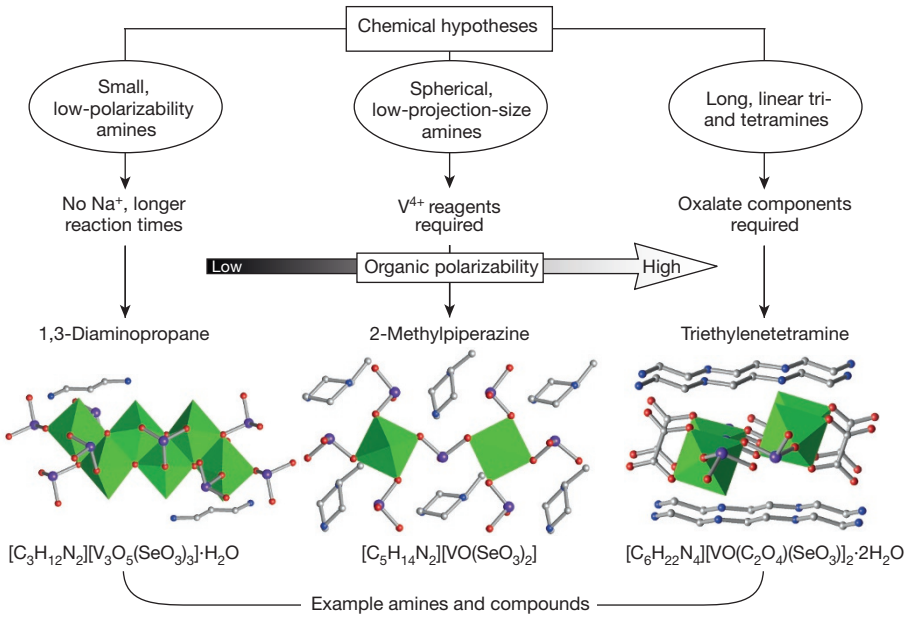

Figure 4 | Graphical representation of the three hypotheses generated from the model, and representative structures for each hypothesis. Experimental conditions required for single-crystal formation largely depend on the amine properties. Small, low-polarizability amines require the absence of competing Na^+ cations and longer reaction times, to avoid precipitating inorganic building units. Spherical, low-projection-size amines require V^{4+}-containing reagents such as $VOSO_4$, because they are unable to generate V^{4+} directly from typical V^{5+} precursors. Long tri- and tetramines require oxalate reactants, to alter the charge density of inorganic secondary building units. These three hypotheses correspond to the green, blue and red subtrees in Fig. 3, respectively.

$NaVO_3$ generally results in formation of inorganic-only polycrystalline products. Excluding sodium from the reaction mixture, by using NH_4VO_3, eliminates this thermodynamic sink, enabling formation of the target phase.

These hypotheses provide specific recommendations for compound formation by: (i) understanding the generation of appropriate primary building units (V^{4+}); (ii) enabling the construction of secondary building units that achieve charge density matching with the cationic components; and (iii) avoiding undesirable building units (Na^+) that result in non-templated phases. These general rules reveal previously unknown insights into our chemistry. The hypotheses derived from this analysis are manifested in three separate compounds, as shown in Fig. 4. $[C_3H_{12}N_2][V_3O_5(SeO_3)_3]\cdot H_2O$ and $[C_6H_{22}N_4][VO(C_2O_4)(SeO_3)]_2\cdot 2H_2O$ are new compounds (crystallographic details available in Supplementary Information); $[C_5H_{14}N_2][VO(SeO_3)_2]$ was reported recently[27]. The polarizabilities of the amines in these compounds range from low (1,3-diaminopropane) to moderate (2-methylpiperazine) and to high (triethylenetetramine).

Our machine-learning approach allows us to exploit chemical information contained in historical reactions and to elucidate the factors governing reaction outcome. The prediction accuracy of the model for previously untested organic amines surpassed the outcomes achieved using the chemical intuition built over many years. In addition, our approach reveals chemical principles governing reaction outcome in the form of testable hypotheses. The ability to make new compounds more successfully and to derive useful chemical information represents a transformative step forwards in exploratory reactions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 September 2015; accepted 22 February 2016.

1. Rao, C. N. R., Behera, J. N. & Dan, M. Organically-templated metal sulfates selenites and selenates. *Chem. Soc. Rev.* **35,** 375–387 (2006).
2. Zhou, H.-C., Long, J. R. & Yaghi, O. M. Introduction to metal–organic frameworks. *Chem. Rev.* **112,** 673–674 (2012).
3. Férey, G. Microporous solids: from organically templated inorganic skeletons to hybrid frameworks...ecumenism in chemistry. *Chem. Mater.* **13,** 3084–3098 (2001).
4. Stranks, S. D. & Snaith, H. J. Metal-halide perovskites for photovoltaic and light-emitting devices. *Nature Nanotechnol.* **10,** 391–402 (2015).
5. Cheetham, A. K., Férey, G. & Loiseau, T. Open-framework inorganic materials. *Angew. Chem. Int. Ed.* **38,** 3268–3292 (1999).
6. Cundy, C. S. & Cox, P. A. The hydrothermal synthesis of zeolites: history and development from the earliest days to the present time. *Chem. Rev.* **103,** 663–702 (2003).
7. Haushalter, R. C. & Mundi, L. A. Reduced molybdenum phosphates: octahedral-tetrahedral framework solids with tunnels, cages, and micropores. *Chem. Mater.* **4,** 31–48 (1992).
8. Férey, G. Oxyfluorinated microporous compounds ULM-*n*: chemical parameters structures and a proposed mechanism for their molecular tectonics. *J. Fluor. Chem.* **72,** 187–193 (1995).
9. Rao, C. N. R., Natarajan, S. & Neeraj, S. Exploration of a simple universal route to the myriad of open-framework metal phosphates. *J. Am. Chem. Soc.* **122,** 2810–2817 (2000).
10. Holdren, J. P. *et al. Material Genome Initiative Strategic Plan.* Technical Report December 2014, https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf (National Science and Technology Council, 2014).
11. Sokolov, A. N. *et al.* From computational discovery to experimental characterization of a high hole mobility organic crystal. *Nature Commun.* **2,** 437 (2011).
12. Hachmann, J. *et al.* Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project. *Energy Environ. Sci.* **7,** 698–704 (2014).
13. Colón, Y. J. & Snurr, R. Q. High-throughput computational screening of metal–organic frameworks. *Chem. Soc. Rev.* **43,** 5735–5749 (2014).
14. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22,** 3762–3767 (2010).
15. Martin, R. L., Lin, L.-C., Jariwala, K., Smit, B. & Haranczyk, M. Mail-order metal–organic frameworks (MOFs): designing isoreticular MOF-5 analogues comprising commercially available organic molecules. *J. Phys. Chem. C* **117,** 12159–12167 (2013).
16. Gaultois, M. W. *et al.* Data-driven review of thermoelectric materials: performance and resource considerations. *Chem. Mater.* **25,** 2911–2920 (2013).
17. Kalidindi, S. R. & Graef, M. D. Materials data science: current status and future outlook. *Annu. Rev. Mater. Res.* **45,** 171–193 (2015).
18. Zhao, J.-C. High-throughput experimental tools for the materials genome initiative. *Chin. Sci. Bull.* **59,** 1652–1661 (2014).
19. Yang, S., Lach-hab, M., Vaisman, I. I. & Blaisten-Barojas, E. Identifying zeolite frameworks with a machine learning approach. *J. Phys. Chem. C* **113,** 21721–21725 (2009).
20. Li, Y. & Yu, J. New stories of zeolite structures: their descriptions, determinations, predictions, and evaluations. *Chem. Rev.* **114,** 7268–7316 (2014).
21. Fernandez, M., Boyd, P. G., Daff, T. D., Aghaji, M. Z. & Woo, T. K. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO_2 capture. *J. Phys. Chem. Lett.* **5,** 3056–3060 (2014).
22. Groom, C. R. & Reilly, A. M. Sixth blind test of organic crystal-structure prediction methods. *Acta Crystallogr.* **B70,** 776–777 (2014).
23. Thakur, T. S., Dubey, R. & Desiraju, G. R. Crystal structure and prediction. *Annu. Rev. Phys. Chem.* **66,** 21–42 (2015).
24. Beran, G. J. O. A new era for ab initio molecular crystal lattice energy prediction. *Angew. Chem. Int. Ed.* **54,** 396–398 (2015).
25. Wicker, J. G. P. & Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm* **17,** 1927–1934 (2015).
26. Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr.* **B58,** 380–388 (2002).
27. Olshansky, J. H. *et al.* Formation principles for vanadium selenites: the role of pH on product composition. *Inorg. Chem.* **53,** 12027–12035 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank Y. Huang, G. Martin-Noble and D. Reilley for data entry and J. H. Koffer for synthetic efforts. M.Z. acknowledges support for the purchase of a diffractometer by the National Science Foundation (DMR 1337296), the Ohio Board of Reagents grant CAP-491 and from Youngstown State University. This work was supported by the National Science Foundation (DMR-1307801). A.J.N. and J.S. each acknowledge the Henry Dreyfus Teacher-Scholar Award program.

Author Contributions S.A.F., J.S. and A.J.N. conceived the project and wrote the paper. A.J.N. supervised the data capture. C.F. developed the web-accessible database. A.J.N. and P.D.F.A. tested the data reliability. J.S. and P.R. developed the reactant descriptors. P.R., C.F. and S.A.F. developed the machine-learning models. J.S. performed diamine selection. P.D.F.A performed the Cambridge Structural Database search. K.C.E., M.B.W. and A.M. performed the hydrothermal experimental reactions, supervised by A.J.N. M.Z. performed X-ray crystallography on the resulting products. P.D.F.A. performed the statistical analyses. P.D.F.A., A.J.N., J.S. and S.A.F. performed the decision-tree calculation and analysis. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.A.F. (sorelle@cs.haverford.edu), J.S. (jschrier@haverford.edu) or A.J.N. (anorquis@haverford.edu).

METHODS

Data capture and reliability. The average rate of data entry from our laboratory notebooks was approximately 50 reactions per hour. Three types of data were entered from the laboratory notebooks. First, compositional information was entered in the form of reactant identities and quantities. Reactants were categorized as being building units for the organic or inorganic structures, or acting as solvent (water). Second, reaction conditions were described, including initial solution pH and heating profile data. Third, reaction-outcome data included both qualitative descriptions of the products and product purity. These descriptions were coded during data entry. Crystal size was coded with the labels 1 for no solid product, 2 for an amorphous solid, 3 for a polycrystalline sample or 4 for single crystals with average crystallite dimensions exceeding approximately 0.01 mm. (This size corresponds to the general requirements for standard single-crystal X-ray diffraction data collection.) Product purity was coded with the labels 1 for a multiphase product or 2 for a single-phase product.

Reliability testing was performed on 100 randomly selected reactions from the database. Each field in each reaction was checked against the laboratory notebook from which this entry was generated. The overall error rate for all fields was 1.89%, which corresponds to 34 errors from a set of 1,800. Each reaction must have at least one inorganic component, one organic component, one solvent, as well as all reaction conditions and outcomes fields listed above. If any of these fields is missing, the reaction is entered into the database for completeness, but is not used for the training or testing of the machine-learning model described below. These filters resulted in a dataset of 3,955 unique, complete reactions.

Reactant descriptors. The ChemAxon Calculator Plugins[28] were used to compute the physicochemical properties of the organic and oxalate-like reactants (for example, molecular weight, number of hydrogen-bond donors/acceptors as a function of pH and polar surface area). For both the organic and oxalate-like reactants, 19 properties were used directly, and others were used to calculate 6 variables describing the mole ratios of the different reactants that were present. For inorganic reactants, 12 atomic properties (for example, ionization potential, electron affinity, electronegativity, hardness and atomic radius), 22 logical values describing the presence or absence of particular metal types, 28 logical values describing the position on the periodic table, and 8 logical values describing the metal valence were used for each element type contained in the reactants. Five variables are experimental reaction conditions (for example, temperature, reaction duration and pH). The descriptor variables are represented in a permutation-invariant fashion (maximum, minimum, arithmetic- and geometric- means) for each reactant type, so that neither the order in which the data are entered nor the number of each component matters, which results in a total of 273 descriptors per reaction. See Supplementary Information for a complete table of computed physicochemical properties.

SVM creation and validation. A broad set of models was evaluated, including decision trees, random forests, logistic regression, *k*-nearest neighbours and SVMs[29]. As shown in Supplementary Table 5, a SVM resulted in the highest accuracy, 74%, as measured using a calculated average of 15 training/test splits. Specifically, a SVM[30] model with a universal Pearson VII function-based kernel[31] was trained on 3,955 labelled reactions previously performed by the laboratory. The SVM was implemented in WEKA 3.7[32,33]; this implementation included a built-in data-normalization step. The model was tested against the known data for its accuracy using a standard 1/3-test and 2/3-training data split. Because the goal is to predict the outcome of reactions with new combinations of reactants, careful partitioning of the test set was required. Holding out test data uniformly at random would potentially put the same combinations of inorganic and organic reactants (reactions differing only by stoichiometries and other conditions) into both the test and training sets, and thus artificially inflate the accuracy rate. Instead, all of the reactions containing a particular set of inorganic and organic reactants were placed into either the test or training set. Under these conditions, the SVM model was measured according to its two-class accuracy, where outcomes of '3' or '4' were considered successes and '1' and '2' were grouped together as failed reactions. The single SVM model used to predict experimental results had an accuracy of 78% in describing all of the reaction types in its test-set data, and 79% considering only vanadium-selenite reactions. The average over 15 such splits was 74%. A learning curve was constructed to test the SVM; details are available in Supplementary Information.

High-dimensional feature spaces are not problematic for SVMs, because they are especially robust to correlated features and are frequently used for problems with many more dimensions than our feature set (for example, in textual learning with 10,000 features)[34]. Feature selection was performed on the model to identify the properties with the most influence on classification success (see Supplementary Information). The selected features were properties of the organic amines (van der Waals surface area, solvent-accessible surface area of positively charged atoms and the number of hydrogen-bond donors) and the inorganic components (mean of the Pauling electronegativites of the metals, their mole-weighted hardness and mean mole-weighted atomic radii). Using only these six features lowers the model accuracy to 70.7%; therefore, the entire set of features was used for the experimental tests. However, the six selected features listed above appear in the decision-tree description of the model.

Selection of new diamines. The eMolecules database (http://academics.emolecules.com/) was used to identify new diamines comprised of only C, H and N atoms, excluding nitriles, hydrazines and isotopically labelled compounds, resulting in 1,680 previously untested, commercially available diamines. For each diamine, a structural fingerprint based on the topological bond paths[35] of the molecule was calculated, and the maximum structural similarity to any of the existing organic compounds in the database was computed using the Tanimoto similarity[35]; the fingerprinting and similarity calculations were performed using the default parameters of the RDKit (http://www.rdkit.org). The particular similarity measure used is not crucial—a comparison of 12 standard fingerprinting methods found that they are all correlated with one another[36]. The list was ranked by similarity and by cost, using the Sigma Aldrich (338 diamines) and Alfa Aesar (62 additional diamines) catalogue prices. After excluding the highest-cost diamines, we sampled 34 diamines across the range of similarities to existing compounds. The same 34 amines were used for both the model and human reactions discussed in the text.

On average, 2 structures have been reported for each of the 34 diamines in the Cambridge Structural Database (CSD)[26], with 19 not existing in any templated metal-oxide structure in the CSD. By contrast, an average of 151 unique structures exist for the most frequently used amines (piperazine, ethylenediamine, 4,4′-dipyridyl and DABCO).

Hydrothermal synthesis. To avoid introducing biases, all reaction types (which differ in specific sets of reagents and reaction conditions) were randomly assigned to be human- or model-controlled, with the stipulation that each amine appear with approximately the same frequency. Amine quantities were determined by either the model or an approach that simply captures human intuitions about exploitation reactions. The recommendations of the model were generated by sampling a range of organic mole amounts, then sorting the results by predicted outcome and confidence. For consistency, human reactions used a rule-based approach that is widely used by the exploratory hydrothermal synthesis community[37], namely, scaling the masses of the organic amines by their respective formula weights, while all other reaction parameters remain unchanged. For brevity, we call this rule-based approach to capture human chemical knowledge "intuition". All reactions were conducted under mild hydrothermal conditions, in 23-ml poly(fluoroethylene-propylene)-lined pressure vessels. The pH values of the initial reaction mixtures were adjusted to the appropriate values using either 4 M HCl or 4 M NaOH. Reaction mixtures were heated to 90–110 °C for 12–72 h. Pressure vessels were opened in air after reaction and products were recovered through filtration. Objective metrics (measured crystallite size and powder X-ray diffraction) were used to score reaction outcomes.

Statistical analysis. Statistical analyses were performed with standard packages available in R 3.2.1[38]. No statistical methods were used to predetermine sample size.

Decision-tree construction. All data were relabelled with the predicted outcomes of the SVM model and a C4.5 decision tree (implemented in WEKA 3.7)[32] was used to model those predicted outcomes[39].

Code availability. All code for this project is available at https://github.com/darkreactions. The code is licensed under the GPL version 3. The precise terms of said license are available with the code.

28. JChem 6.1.3, http://www.chemaxon.com (ChemAxon, 2013).
29. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* 2nd edn, Ch. 9, 12, 13, 15 (Springer, 2009).
30. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20,** 273–297 (1995).
31. Üstün, B., Melssen, W. & Buydens, L. M. C. Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemom. Intell. Lab. Syst.* **81,** 29–40 (2006).
32. Hall, M. *et al.* The WEKA data mining software: an update. *ACM SIGKDD Explor. Newslett.* **11,** 10–18 (2009).
33. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2,** 27 (2011).
34. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *Proc. 10th European Conf. Machine Learning* (eds Nédellec, C. & Rouveirol, C.) 137–142 (Springer, 1998).
35. Leach, A. & Gillet, V. J. *An Introduction to Chemoinformatics* Ch. 5 (Springer, 2007).
36. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminformat.* **5,** 26 (2013).
37. Thangavelu, S. G., Butcher, R. J. & Cahill, C. L. Role of N-donor sterics on the coordination environment and dimensionality of uranyl thiophenedicarboxylate coordination polymers. *Cryst. Growth Des.* **15,** 3481–3492 (2015).
38. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, 2015).
39. Barakat, N. & Diederich, J. Eclectic rule-extraction from support vector machines. *Int. J. Comput. Intell.* **2,** 59–62 (2005).