https://doi.org/10.1038/s41586-019-1540-5

Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis

Xiwen Jia[1], Allyson Lynch[1], Yuheng Huang[1], Matthew Danielson[1], Immaculate Lang'at[1], Alexander Milder[1], Aaron E. Ruby[1], Hao Wang[1], Sorelle A. Friedler[2]*, Alexander J. Norquist[1]* & Joshua Schrier[1,3]*

Most chemical experiments are planned by human scientists and therefore are subject to a variety of human cognitive biases[1], heuristics[2] and social influences[3]. These anthropogenic chemical reaction data are widely used to train machine-learning models[4] that are used to predict organic[5] and inorganic[6,7] syntheses. However, it is known that societal biases are encoded in datasets and are perpetuated in machine-learning models[8]. Here we identify as-yet-unacknowledged anthropogenic biases in both the reagent choices and reaction conditions of chemical reaction datasets using a combination of data mining and experiments. We find that the amine choices in the reported crystal structures of hydrothermal synthesis of amine-templated metal oxides[9] follow a power-law distribution in which 17% of amine reactants occur in 79% of reported compounds, consistent with distributions in social influence models[10–12]. An analysis of unpublished historical laboratory notebook records shows similarly biased distributions of reaction condition choices. By performing 548 randomly generated experiments, we demonstrate that the popularity of reactants or the choices of reaction conditions are uncorrelated to the success of the reaction. We show that randomly generated experiments better illustrate the range of parameter choices that are compatible with crystal formation. Machine-learning models that we train on a smaller randomized reaction dataset outperform models trained on larger human-selected reaction datasets, demonstrating the importance of identifying and addressing anthropogenic biases in scientific data.

Scientific publications do not provide a representative dataset[12]. Confirmation bias favours the publication of positive results, yet the missing 'failures' are essential knowledge for modelling chemical reactions[6]. Scientific attention is skewed by biases such as the 'Matthew effect', in which eminent individuals are given disproportionate credit[10]. Self-reinforcing preferential attachment ('rich get richer') mechanisms result in power-law distributions of citations, resulting in disproportionately popular articles[10,11]. The emerging 'science of science' discipline attempts to quantify the role of social interactions in the selection of a research problem, career trajectory and citations[13]. Although there are studies of error in scientific decisions, these have tended to focus on individual-specific causes, such as variability in classification and inconsistencies in decision-making[14,15]. By contrast, systematic errors in the planning of scientific experiments have not been studied. In general, social influences, such as knowledge about others' choices—for example, choice of reagents—can cause disproportionate popularity compared to the underlying quality of the item or choice[3]. Social influences in scientific decision-making have been widely speculated, but never explicitly confirmed. Social influence in scientific decisions may be a factor in the distribution of reported medicinal chemistry compounds, which is unrelated to the intended application, cost or reaction difficulty[16]; the disproportionately few drug scaffolds that comprise the majority of antimalarial[17] and other drug-candidate molecules[18], the popularity of which is uncorrelated to their synthetic feasibility or biological activity; and the synthesis of new pharmaceutical molecules that resemble those the medicinal chemists involved have synthesized in the past[19,20], which use a limited set of reactions[21], the choice of which is uncorrelated to cost, estimated ease of the synthesis, or the properties of the reactants and products[22]. However, over-representation of a particular experimental choice need not be irrational. For example, 36% of entries in the Protein Data Bank (PDB) report the use of polyethylene glycol additives, which under-represents the true success rate of 59%, and many of these proteins cannot be crystallized using other additives[23]. This suggests that a lack of diversity among crystallization additives in the PDB stems from sub-optimal novelty seeking. Excessively consistent or inconsistent experimental choices that do not mimic the natural distribution of the underlying problem are a signature of anthropogenic influence.

We seek to determine whether there is evidence of bias in reactant choices for organically templated metal oxide syntheses. The incorporation of different organic amines results in compounds with diverse composition, local and extended connectivity, and functionality[9], so an unbiased set of experimental efforts should consist of the broadest possible range of amine choices in reported compounds of this type. The discipline defines 'success' as formation of a crystal of sufficient size and quality to yield a stable single-crystal X-ray diffraction refinement. Most publishers of scientific reports require structures to be deposited in the Cambridge Structural Database (CSD). In this study, the number of reported compounds is used as a proxy for experimental effort and success for a particular amine. The CSD includes the structures of 5,010 amine-templated metal oxides that contain 415 unique amines. The top 17% commonly reported amines (70 individual molecules, the 'popular' amines) are found in 79% of the structures (3,947 distinct CSD entries), and correspondingly, the remaining 83% (345 molecules, the 'unpopular' amines) of amines are found in just 21% of the structures (1,063 entries; see Fig. 1). (Structures containing multiple amines preclude a precisely equal relationship.) The Gini coefficient of the distribution of amines in the database—a measure of statistical dispersion that quantifies inequality among values of a frequency distribution—is 0.654, comparable to global wealth inequality. A log(proportion)–log(rank) plot (Fig. 1a, inset) is consistent with a single power-law generation process, consistent with a preferential attachment mechanism[10,11]. Similar distributions are observed in the subset of metal borates (Extended Data Fig. 1). Unpublished experimental records (from our Dark Reactions Project, https://darkreactions.haverford.edu/) provide evidence of bias in the conditions of attempted reactions, such as pH and reactant quantities. This database contains 557 hydrothermal vanadium borate reactions (motivated by the first report in the literature of this type of reaction[24]) consisting of the work of three students conducted over three years prior to the start of this study. The human-selected reactions in this dataset are almost exclusively conducted at pH 8 (Fig. 2a), with quantities of amines that are unevenly distributed (Fig. 2b). These results are consistent with previous work showing that humans often use a one-variable-at-a-time strategy to explore reaction conditions, which is both inefficient and easily trapped in local optima[25,26].

[1]Department of Chemistry, Haverford College, Haverford, PA, USA. [2]Department of Computer Science, Haverford College, Haverford, PA, USA. [3]Department of Chemistry, Fordham University, The Bronx, New York, NY, USA. *e-mail: sorelle@cs.haverford.edu; anorquis@haverford.edu; jschrier@fordham.edu

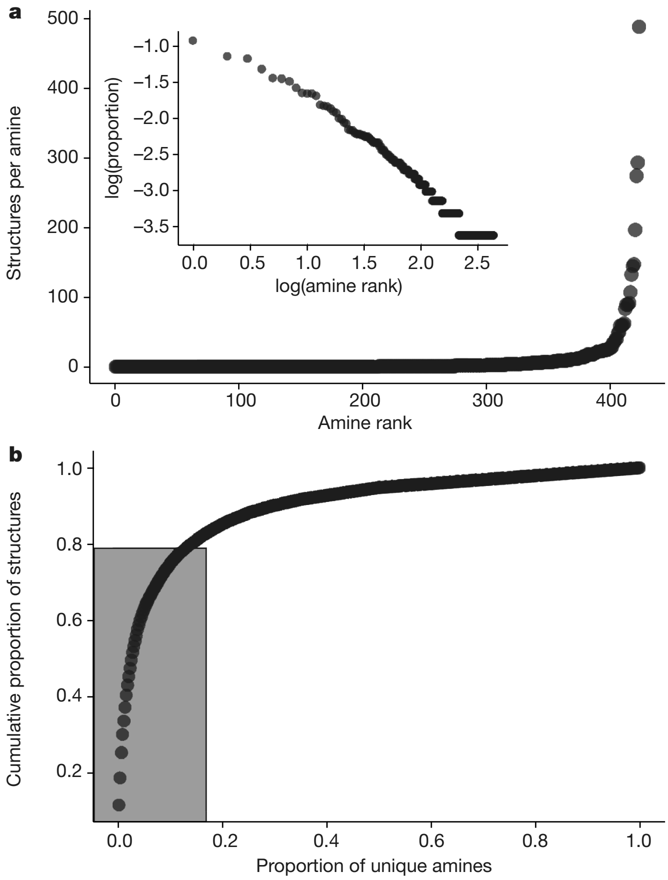


Fig. 1 | Occurrence of amines in structures of reported metal oxide crystals. a, The number of crystal structures observed for each amine, plotted against the amine's rank (ascending, by occurrence in the reported crystal structures). The inset shows the same data as a log(proportion)–log(amine rank) plot. **b**, Cumulative proportion of crystal structures containing the most commonly occurring amines as a function of the proportion of total unique amines, ordered from those occurring in the most structures (and therefore with the highest proportion) to the least commonly occurring. The shaded region represents the Pareto split, in which the most frequently observed 17% of all amines occur in 79% of the structures.

The skewed distributions discussed above are consistent with anthropogenic attention heuristics[2], in which experimenters select reactants and reaction conditions that they 'know' to work, on the basis of their own experience, that of their colleagues or the literature. The interplay between shorter-timescale communicative memory and longer-timescale cultural memory[11] results in power-law distributions such as that seen in Fig. 1. The precise nature of the underlying psychological process is an active area of debate[1,2], and distinguishing between competing models is challenging even in highly controlled psychological experiment settings[27]. For example, an aesthetic bias links symmetry with a positive affect[28], and this could lead researchers to favour experimentation with symmetric molecules. Alternatively, the fact that humans are more easily able to discriminate and recall symmetric three-dimensional objects[29] might favour symmetric molecules when devising new experiments, through recall and attention heuristics[2]. Both scenarios yield unrepresentative datasets (that is, datasets that are 'biased'), despite radically different mechanisms. Given the many possible types of anthropogenic influence and the difficulty of distinguishing them experimentally[27], we instead demonstrate the presence of an anthropogenic influence on the choice of reagents and reaction conditions, by eliminating alternative explanations and without assigning a specific psychosocial origin.

Non-anthropogenic factors can be classified following a classical fourfold theory of causes[30]. Efficient causes are the technical ability to perform the experiments. Here, hydrothermal syntheses have been conducted for over 50 years[9]; the choices of reagent and reaction conditions present no pressure or corrosion resistance challenges. Therefore, no technical experiment limitations favour certain reagents or reactions conditions over others, and this cannot be the origin of the imbalances shown in Figs. 1, 2. Final causes favour particular product materials that have desired technological properties. Because functional diversity follows from structural diversity[9], unbiased exploration should cover the broadest range of reagent structures, in contrast to the predominance of certain structures reported the public databases as shown in Fig. 1. Material causes—described by financial cost and reagent availability—were excluded by considering a structurally diverse set (primary through tertiary amines, and linear, branched, cyclic and aromatic molecules) of 55 commercially available amines, of which—based on the CSD—27 are popular, 16 are unpopular and 12 are absent. All the selected amines are commercially available in 5-g quantities from major suppliers, and there was no systematic difference in cost (Extended Data Fig. 2).

The only remaining non-anthropogenic cause is eliminated by experiment. Here, the formal cause is the intrinsic propensity of some reactants and reaction conditions to yield crystals. To make an unbiased assessment, we generated 10 random reactions for each of the 55 amines described above, by selecting a random pH level and amine quantity using two independent triangular distributions with the mode set at the value that has precedent in the literature, and with physically motivated upper and lower bounds (see Methods). The goal was not to efficiently explore the chemical space, but instead to establish a neutral estimate of the 'reaction cross-section' for each amine, revealing systematic reactivity differences between popular amines and not-popular amines (that is, those that were unpopular or absent from the CSD), centred around the reaction compositions that humans are likely to have attempted to search. Despite this, we note that random choices are often better than human expertise and comparable to more sophisticated numerical methods in fields as diverse as oil exploration[31,32], chemical reaction discovery[33] and numerous social and financial applications[34].

Outcomes for all reactions were ranked using the four-class scoring system for crystal formation described in Methods section 'Data capture', with the stringent criterion that 'success' (outcome 4) consists only of crystals comparable to those used for the CSD data (single crystals with average crystalline dimensions great than about 0.01 mm). All four reaction outcomes occur nearly equally in reactions generated by the randomly selected reaction conditions for popular and not-popular amines (Fig. 3a; similar plots separating unpopular and absent outcomes are shown in Extended Data Fig. 3). Any single reaction generated by the randomly selected conditions is equally likely to be successful regardless of the popularity of the amine used. A typical exploratory synthesis campaign will test several variations of a reaction until success is achieved or one decides to stop. Publishable data require only a single success. We modelled this as the observation of success (outcome 4) at least once in the set of 10 random experiments conducted for each of the 55 amines. At least one success is observed in 17 (63% ± 9%) of the 27 popular amines and 21 (75% ± 8%) of the 28 not-popular amines (Fig. 3b). We find there is no support for any difference in intrinsic reaction propensity between popular and not-popular amines. In fact, in our experiments, popular amines were less likely to successfully form crystals than not-popular amines; however, this (one-sided) success rate difference occurs with $P = 0.26$ in a random permutation. Having thus excluded the non-anthropogenic efficient, formal, material and final causes, only an anthropogenic explanation for the observed reactant choice distribution remains.

The randomized reaction outcomes expose anthropogenic influence in the choice of reaction condition. Human-selected reactions are biased towards smaller amine amounts, with a large peak at the precedent set by the literature (Fig. 2c). By contrast, we find limited dependence of the reaction outcome on amine choice, as the distribution of successful (outcome 4) and failed (outcomes 3, 2 and 1) reactions mimics the triangular distribution that generates it. The choice of

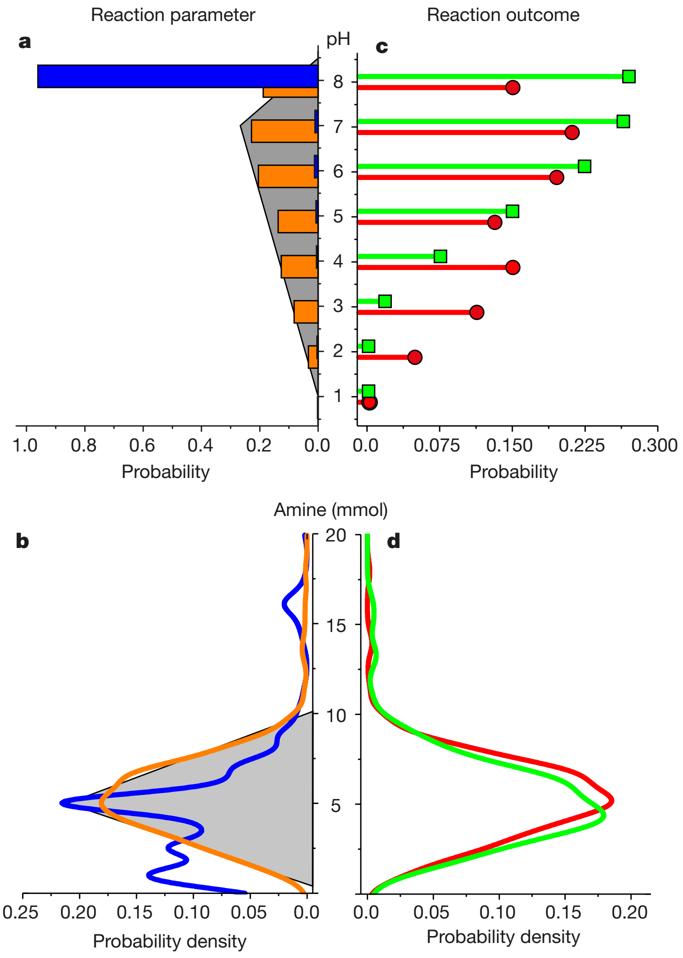


Fig. 2 | Distribution of the choices of reaction parameters and reaction outcomes. a, **b**, The distributions of the reaction parameters for the pH of the reaction by probability (**a**) and the amine quantity in millimoles by probability density (**b**). In both **a** and **b**, blue indicates the distribution of the human-selected reactions taken from a historical dataset of 557 reactions, the grey region indicates the triangular distribution defined for generating our random experiments, and the orange indicates the distribution of the 548 random reactions performed in this study. **c**, **d**, The distributions of the reaction outcomes based on the pH of the reaction (**c**) and the amine quantity (**d**). For the randomly generated reactions performed in this study, successful reaction outcomes (outcome 4) are indicated by green and failures (outcomes 1, 2 and 3) are indicated by red.

reaction pH that humans make is almost exclusively based on a literature precedent of pH 8 (Fig. 2d). However, although our randomized reactions indicate that a reaction performed at higher pH is—all other things held equal—more likely to be successful, reactions can be successful over a wide range of pH.

Correcting anthropogenic bias improves machine-learning models. We compared machine-learning models trained on the complete set of (both successful and failed) human-selected reactions to models trained on our randomly generated (unbiased) reactions. The comparison was evaluated for a true time-separated hold-out test set of 110 additional vanadium borate experiments, comprising 10 randomly generated reaction conditions for each of 11 amines. None of these 11 test amines was present in either of the training sets. For the purposes of training, we used only the subsets of the human-selected and randomly generated reaction condition datasets that contained the 37 amines common to both sets, thereby reducing the training set sizes to 467 and 370 reactions, respectively. (Our human-selected training dataset has more diverse amine choices than typically present in the literature, as it used amine reaction data from our previous diversity-oriented studies[6].) Restricting the training datasets in this way means

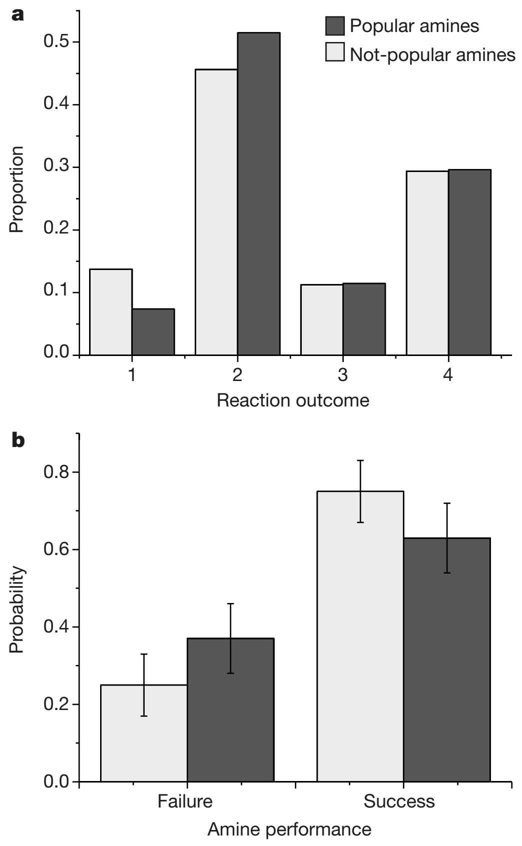


Fig. 3 | Reaction outcomes from randomly generated experiments for popular amines and not-popular (unpopular and absent) amines. a, The proportion by outcome for each reaction, using the outcome scale described in Methods, for the popular and not-popular amines in the human-selected dataset. **b**, Estimated probability of observing at least one successful reaction (outcome 4) or failure (outcomes 1, 2 and 3) for a given amine, for the $N = 27$ popular and $N = 28$ not-popular amines among the human-selected dataset. Centre values indicate observed proportion of outcomes. Error bars indicate a bootstrap estimate of the standard deviation.

that differences are solely due to choices of reaction conditions. To give each model the best chance to succeed, a variety of classifiers were tested for each training set—logistic regression, k-nearest neighbours, support vector machines, decision trees, random forests and Gaussian naive Bayes—and the best-performing classifier for each training set was considered. A detailed analysis of the results is presented in Supplementary Tables 8–21, and summarized in Extended Data Table 2. The best classifier trained on the human-selected dataset was the k-nearest neighbours ($k = 2$) with an accuracy of 69% and an area under the curve (AUC) of 0.64 on the held-out test set. The best classifier trained by the randomly generated training set was the k-nearest neighbours ($k = 5$) with an accuracy of 79% and an AUC of 0.80. Thus, the randomly generated training set outperforms the human-selected set by all metrics, despite containing 20% fewer reactions. Both models have access to the same classifier functions, so this indicates that the performance of the held-out test set is more effectively predicted by the randomly generated set than by the human-selected set.

The improved sampling over the feature space is established by considering the average nearest-neighbour distance between reactions (Extended Data Fig. 4). The average distance to the kth nearest neighbour within a given training set is greater for the randomly generated training set for all $k > 10$, indicating that it more comprehensively samples the chemical space. Furthermore, the average distance from an experiment in the training set to the kth nearest neighbour in the test set is smaller for the randomly generated training set for all $k \leq 60$, indicating that it allows better generalization to the test set. Both factors help to make the randomly generated experiments more informative than those performed under human-selected conditions.

Anthropogenic dataset bias obscures chemical insights. Because the two training sets contain the same amines, the dependence of reaction outcomes on features of the amines such as their structure and physicochemistry should be equally well described. Indeed, the direct influence of the features—that is, the contribution of a feature to the difference between the given predictions and a mean result[35]—is comparable for the models built on the different training sets. By contrast, the indirect influence of the features—which is estimated by computing the degradation in model accuracy when the ability of the model to predict that feature from the other features is removed[35]—is linearly correlated for the two training sets, with the exception of six features describing the properties of the amine: solvent-accessible polar surface area, presence of carbon aliphatic and aromatic rings, number of rotatable and total bonds, and presence of amidine moieties in the organic molecules (Extended Data Fig. 5). Computationally obscuring these features in the random-reaction-trained model degrades the model performance, but computationally obscuring them in the human-reaction-trained (anthropogenic) model does not, because the anthropogenic selection of reaction conditions has implicitly obscured these feature contributions.

Anthropogenic bias hinders the discovery of new materials. Only 41 out of 110 test reactions successfully produced a crystalline product of sufficient quality and size (outcome 4), and the positive recall scores were 46% and 85% for the models trained on human-selected and randomly generated data, respectively (Supplementary Tables 9, 17). The two models disagree in 23 out of 110 test outcomes, and in every case the human model predicts failure and the random model predicts success. The ‘pessimism’ of the former is consistent with loss-aversion bias in human experimental choices[1]. When the models disagree, it is preferable to trust the model trained on randomly generated data (which correctly predicts 16 true positives) rather than the model trained on human data (which correctly predicts 7 true negatives). Furthermore, only 7 out of 11 amines in the test set had at least one successful reaction outcome sufficient for discovery of a new material. The anthropogenic model failed to identify two of these compounds, whereas the random-data model found at least one successful reaction for all seven compounds. Therefore, models trained on the randomly generated dataset are both quantitatively and qualitatively better at identifying the successful reaction conditions that are required for the discovery of new compounds.

Models trained on anthropogenic data select new experiments less effectively. We generated 10,000 sets of random reaction conditions for each of the 11 test amines. Predictions of reaction success by the two models agree on 81% of these reactions, including all of the generated reactions for 3 out of the 11 amines. For the eight amines for which the models disagree about the outcome of any of the 10,000 experiments, the anthropogenic model makes unique predictions of success for only two out of these eight amines, whereas the model trained on randomly generated data identifies unique positive predictions for all eight cases. As an additional test, we then conducted laboratory syntheses of ten discrepant positive predictions made by each model for each amine (totalling 100 additional reactions), selected from the 10,000 random conditions. For the two amines for which both models made different positive predictions, the anthropogenic model was slightly more successful (16 out of 20 positives found) than the model trained on random experiments (12 out of 20). However, for the other six amines, for which only the random-data model made unique positive predictions, at least one successful reaction was observed in all cases. The relatively low aggregate precision (43%) is because these are more speculative reactions for which the models are less confident about the outcome; the precision increases with the model’s predicted probability of success and is as high as 80%. (Extended Data Table 4, Supplementary Table 22 and Supplementary Figs. 1, 2 contain a complete analysis of the laboratory and computational results.) This confirms that the model trained on data that does not contain anthropogenic bias better identifies reaction success over a broader range of reactant and reaction condition choices.

Our results indicate the importance of including reactions that humans ordinarily do not choose when constructing machine-learning models for chemical reactions. The implicit anthropogenic biases in the scientific literature may hinder data-driven chemical synthesis planning efforts[5–7]. To avoid this bias, we recommend a simple process of listing all experimental options, defining distributions that exclude impossible choices (on the basis of known physical considerations such as solubility or protonation state) or practically infeasible choices (for example, costs and safety), and then randomly sampling from those distributions. Our results indicate that experiments with this type of structured randomness remove anthropogenic bias, are at least as successful as human choices, and greatly improve the value of the resulting datasets for reaction outcome prediction.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1540-5.

Received: 30 December 2018; Accepted: 10 July 2019;
Published online 11 September 2019.

1. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).
2. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
3. Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
4. Henson, A. B., Gromski, P. S. & Cronin, L. Designing algorithms to aid discovery by chemical robots. *ACS Cent. Sci*. **4**, 793–804 (2018).
5. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
6. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
7. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
8. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
9. Cheetham, A. K., Férey, G. & Loiseau, T. Open-framework inorganic materials. *Angew. Chem*. **38**, 3268–3292 (1999).
10. Price, D. D. S. A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. Sci*. **27**, 292–306 (1976).
11. Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L. & Hidalgo, C. A. The universal decay of collective memory and attention. *Nat. Hum. Behav*. **3**, 82–91 (2018).
12. Carroll, H. A., Toumpakari, Z., Johnson, L. & Betts, J. A. The perceived feasibility of methods to reduce publication bias. *PLoS One* **12**, e0186472 (2017).
13. Fortunato, S. et al. Science of science. *Science* **359**, (2018).
14. Greenslade, P., Florentine, S. K., Hansen, B. D. & Gell, P. A. Biases encountered in long-term monitoring studies of invertebrates and microflora: Australian examples of protocols, personnel, tools and site location. *Environ. Monit. Assess*. **188**, 491 (2016).
15. Boobier, S., Osbourn, A. & Mitchell, J. B. O. Can human experts predict solubility better than computers? *J. Cheminform*. **9**, 63 (2017).
16. Keserű, G. M., Soós, T. & Kappe, C. O. Anthropogenic reaction parameters – the missing link between chemical intuition and the available chemical space. *Chem. Soc. Rev*. **43**, 5387–5399 (2014).
17. Varela, J. N., Lammoglia Cobo, M. F., Pawar, S. V. & Yadav, V. G. Cheminformatic analysis of antimalarial chemical space illuminates therapeutic mechanisms and offers strategies for therapy development. *J. Chem. Inf. Model*. **57**, 2119–2131 (2017).
18. Zdrazil, B. & Guha, R. The rise and fall of a scaffold: a trend analysis of scaffolds in the medicinal chemistry literature. *J. Med. Chem*. **61**, 4688–4703 (2018).
19. Cleves, A. E. & Jain, A. N. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput. Aided Mol. Des*. **22**, 147–159 (2008).
20. Jain, A. N. & Cleves, A. E. Does your model weigh the same as a duck? *J. Comput. Aided Mol. Des*. **26**, 57–67 (2012).
21. Brown, D. G. & Boström, J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? *J. Med. Chem*. **59**, 4443–4458 (2016).
22. Brown, D. G., Gagnon, M. M. & Boström, J. Understanding our love affair with *p*-chlorophenyl: present day implications from historical biases of reagent selection. *J. Med. Chem*. **58**, 2390–2405 (2015).
23. Kirkwood, J., Hargreaves, D., O’Keefe, S. & Wilson, J. Analysis of crystallization data in the Protein Data Bank. *Acta Crystallogr. F* **71**, 1228–1234 (2015).

24. Rijssenbeek, J. T., Rose, D. J., Haushalter, R. C. & Zubieta, J. Novel clusters of transition metals and main group oxides in the alkylamine/oxovanadium/borate system. *Angew. Chem*. **36**, 1008–1010 (1997).
25. Duros, V. et al. Human versus robots in the discovery and crystallization of gigantic polyoxometalates. *Angew. Chem*. **56**, 10815–10820 (2017).
26. Cao, B. et al. How to optimize materials and devices via design of experiments and machine learning: demonstration using organic photovoltaics. *ACS Nano* **12**, 7434–7444 (2018).
27. Kahneman, D. & Klein, G. Conditions for intuitive expertise: a failure to disagree. *Am. Psychol*. **64**, 515–526 (2009).
28. Evans, D. W. et al. Human preferences for symmetry: subjective experience, cognitive conflict and cortical brain activity. *PLoS One* **7**, e38966 (2012).
29. Liu, Z. & Kersten, D. Three-dimensional symmetric shapes are discriminated more efficiently than asymmetric ones. *J. Opt. Soc. Am. A* **20**, 1331–1340 (2003).
30. Falcon, A. Aristotle on causality. *The Stanford Encyclopedia of Philosophy Spring 2019 edn* (ed. Zalta, E. N.) https://plato.stanford.edu/archives/spr2019/entries/aristotle-causality (Stanford Univ., 2019).
31. Menard, W. H. & Sharman, G. Scientific uses of random drilling models. *Science* **190**, 337–343 (1975).
32. Menard, W. H. & Sharman, G. Random drilling. *Science* **192**, 206–208 (1976).
33. McNally, A., Prier, C. K. & MacMillan, D. W. C. Discovery of an α-amino C–H arylation reaction using the strategy of accelerated serendipity. *Science* **334**, 1114–1117 (2011).
34. Biondo, A. E., Pluchino, A. & Rapisarda, A. The beneficial role of random strategies in social and financial systems. *J. Stat. Phys*. **151**, 607–622 (2013).
35. Adler, P. et al. Auditing black-box models for indirect influence. *Knowl. Inf. Syst*. **54**, 95–122 (2018).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Data capture. Data capture from historical notebooks, generated in our laboratory during the previous decade, and from the new experiments conducted in this study follows the procedure described in our previous work[6]. These reactions include compositional information (reactant identities and quantities), category (organic, inorganic or solvent), reaction conditions (for example, pH, temperature and time), and the reaction outcome information. Reaction outcomes were coded qualitatively on the basis of crystal size: 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.) To eliminate measurement bias, the students who performed the reactions and scored the crystal outcomes were unaware of the popularity of the reagent in the CSD. A machine-readable collection of all experimental data are provided in the supplementary information files.

Analysis of published crystal structures. Amine-templated metal oxides were extracted from the CSD[36] (using the software Conquest) by stipulating both inclusion and exclusion criteria. The inclusion criteria were used to guarantee the presence of an oxide, metal oxide or metal borate substructure, in addition to an organic amine. The exclusion criteria were used to remove structures with bonding motifs that fell outside the target family of compounds. (These criteria are defined in Extended Data Table 1.) This resulted in the identification of 7,630 oxides, 4,870 metal oxides and 115 metal borates. Analysis of the metal oxide data is presented in this Letter; a parallel analysis of the metal borates is described in Extended Data Fig. 1. We initially attempted to extract the organic components from the three-dimensional structure, but the presence of structural disorder resulted in ambiguity. The two-dimensional structure diagrams are not publicly available through the CSD application programming interface. Therefore, we parsed the systematic names to identify the amine component. Excluded names were manually curated, and 43 typographical errors in the CSD entries were communicated to the maintainers. A strict definition of organic amines was used, which included only molecules comprising solely C, H and N, and containing no nitriles or azo, diazo, or diazonium functional groups. After performing these exclusions, 6,458 oxides, 4,152 metal oxides, and 109 metal borate structures remained. The amine names were resolved to canonical SMILES strings using the CACTUS Chemical Identity Resolver (https://cactus.nci.nih.gov/chemical/structure), and then converted to neutral molecules and canonicalized using RDKit[37]. CACTUS was also used to generate InChI and InChIKey strings. The Python-based Jupyter notebooks used to perform this process, along with the inputs and intermediate outputs, are provided in the supplementary information files.

Amine popularity was quantified using the Pareto split, the value of X where the top X% of most frequently observed amines accounts for 100% − X% of the total structures. This choice has joint ratio symmetry, that is, the remaining 100% − X% of the least frequently observed amines account for the remaining X% of structures. Structures containing multiple amines preclude a precisely equal X:100 − X ratio for a finite dataset, so the largest X that does not exceed the ratio is used. Pareto split values for the 4,152 metal oxide structures and the 109 metal borate structures are indicated by the shaded areas in Fig. 1and Extended Data Fig. 1, respectively. Using the metal oxide dataset, amines were classified as 'popular' if they were in the over-represented top X% of amines, and 'unpopular' if in the under-represented bottom 100% − X% of amines, 'absent' if not reported in the CSD at all, and 'not-popular' if either unpopular or absent.

Pricing and availability. Amine pricing information was collected by searching the Sigma-Aldrich website (https://www.sigmaaldrich.com) in November 2018. Pricing and sample-size data were collected for the smallest sample size available, where the sample size was at least 5 g. Pricing data were collected for all amines included in the training reactions and the test set amines. All amines were in stock on the day data were collected.

Generation of randomized reactions. The randomized reactions were generated by sampling from triangular distributions for the reaction pH and amine quantity. The triangular distribution for pH was chosen with minimum and maximum values of 1 and 8.49, and a mode of pH 8; random numbers were drawn from this distribution and then rounded to the nearest integer. (The pH of a reaction is not easily set below 0 and basic conditions will not protonate an amine.) The triangular distribution for amine quantity was chosen with minimum and maximum values of 0.5 mmol and 10 mmol, with a mode of 5 mmol. (The amine quantity cannot be reduced below zero and cannot be increased above a conservative solubility limit chosen for all amines.) Conversion of the amine molar quantities to masses was performed using molecular weights from PubChem. The Mathematica notebook used to perform these calculations is provided in the supplementary information files.

Hydrothermal synthesis. All reactions were conducted under mild hydrothermal conditions in 23-ml poly(fluoroethylene-propylene)-lined pressure vessels. All reactions were specified for 0.31 g H_3BO_3, 0.083 g $VOSO_4{\cdot}xH_2O$ and 6.0 g H_2O, and the amine quantity in mmol was drawn from the previously described distribution. The reactions were adjusted to the pH specified by the above distribution using either 4 M HCl or 4 M NaOH (as determined by pH paper). Reaction mixtures were heated to 90 °C for 24 h. Pressure vessels were opened in air after the reaction and products were recovered through filtration. Objective metrics (measured crystallite size and powder X-ray diffraction) were used to score reaction outcomes, as described in the above section 'Data capture'.

Statistical analysis of experimental outcomes. Standard deviations and P values were assigned by numerical 10,000-sample bootstrapping and permutation. The Mathematica notebook used to perform these calculations is provided in the supplementary information files. No statistical methods were used to predetermine experimental sample size.

Machine-learning model construction. Only a single set of inorganic reactants was used for all reactions in this study, and only a single organic reactant per reaction was used. Therefore, only a subset of the reaction descriptors from our previous work[6] was used in this study. These three descriptor categories include reaction parameters (for example, temperature and pH), physicochemical and structural features of the organic component, and stoichiometric ratios. The structural and physicochemical properties of the organic species were computed using RDKit 2018.03.4 (ref. [37]) and the ChemAxon Calculator plugins[38]. Supplementary Tables 1–7 contain a complete description of the features.

Feature selection was performed to choose the top 5, 10, and 20 features for both training sets, using two methods: an F-test-based estimate and a mutual-information-based estimate of the importance of each feature. Two additional sets, one containing all features and one containing all features with positive variance were also considered. The full set of considered models was trained on each of these feature sets for both the human-selected and randomly generated training sets. As described in the main text, the full feature set had the top performance model based on accuracy. A five-feature F-test-based feature set had the highest AUC (0.69), but had lower accuracy (0.63) than the set with all features. In general, the human-data-trained models performed very poorly; many had an accuracy of around 0.5. The full results of the feature selection trials can be found in Extended Data Table 3. All models were implemented in Python 3.7.3 using *Scikit-learn* version 0.19.1(ref. [39]); model-specific details and implementations can be found in the supplementary information files.

Direct and indirect feature influence analysis. The direct influence of each feature—a Shapely-value-based approximation of the contribution of a feature to the deviation of predictions from the mean—was computed using SHapley Additive exPlanations (SHAP, specifically the Kernel SHAP approximation)[40,41]. The indirect influence of each feature was calculated using BlackBoxAuditing[35,42] to measure each feature's contribution to the accuracy of the model even when it is not directly used in the model. This influence is estimated by obscuring a feature so that it cannot be predicted by the other features and measuring the drop in model accuracy when the values are obscured in this way. These calculations were performed for the most accurate models trained on the human-selected and randomly generated training sets. Comparison plots are shown in Extended Data Fig. 5.

Data availability

The authors declare that all data supporting the findings of this study are available within the article and its supplementary information.

Code availability

The code used for this project is available in the supplementary information files.

36. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. B* **72**, 171–179 (2016).
37. Landrum, G. RDKit: open-source cheminformatics http://www.rdkit.org (2018).
38. ChemAxon. JChem cxcalc 5.2.0. http://www.chemaxon.com (2018).
39. Pedregosa, F. et al. *Scikit-learn*: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
40. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) 4765–4774 (Curran Associates, 2017).
41. Lundberg, S. M. *SHAP*. (SHapley Additive exPlanations) https://github.com/slundberg/shap (2019).
42. Scheidegger, C., Falk, C., Friedler, S., Venkatasubramanian, S. & Nix, T. BlackBoxAuditing https://github.com/algofairness/BlackBoxAuditing (2019).

Acknowledgements We thank G. Cattabriga for software engineering support and X. Weng for proofreading the experimental data entries. This project was funded by the National Science Foundation (award no. DMR-1709351). I.L. was partially supported by a Bryn Mawr College LILAC Summer Internship Funding Program. J.S. acknowledges the Henry Dreyfus Teacher-Scholar Award (TH-14-010).

Author contributions J.S. and A.J.N. conceived the project. A.R., H.W., X.J., and A.M. devised and performed the human-selected reactions, supervised by A.J.N. X.J. and A.M. collected historical notebook data, supervised by A.J.N; A.J.N and X.J. extracted the appropriate structures. A.L., I.L. and J.S. determined the amine counts from these structures. J.S. generated the random reactions. O.H., X.J., M.D. and A.M. performed the randomly generated and test set reactions, supervised by A.J.N. Statistical analysis was performed by A.L. and J.S. S.A.F. performed model construction and analysis. S.A.F., J.S. and A.J.N. performed and interpreted the feature influence analysis calculations. J.S., A.J.N and S.A.F. wrote the paper. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

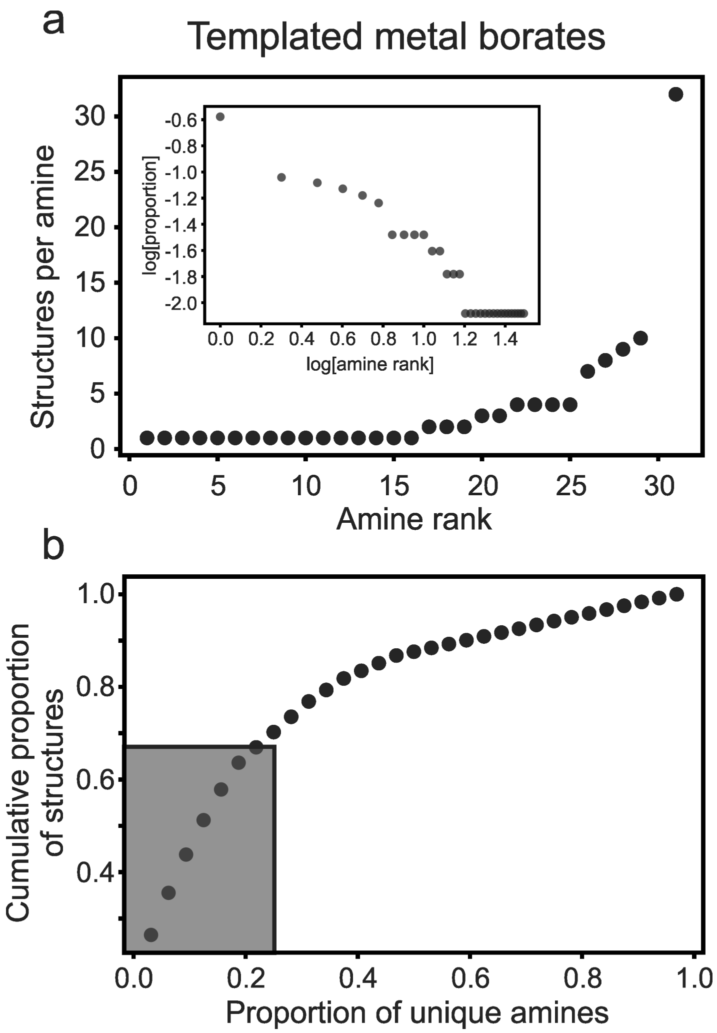
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1540-5.

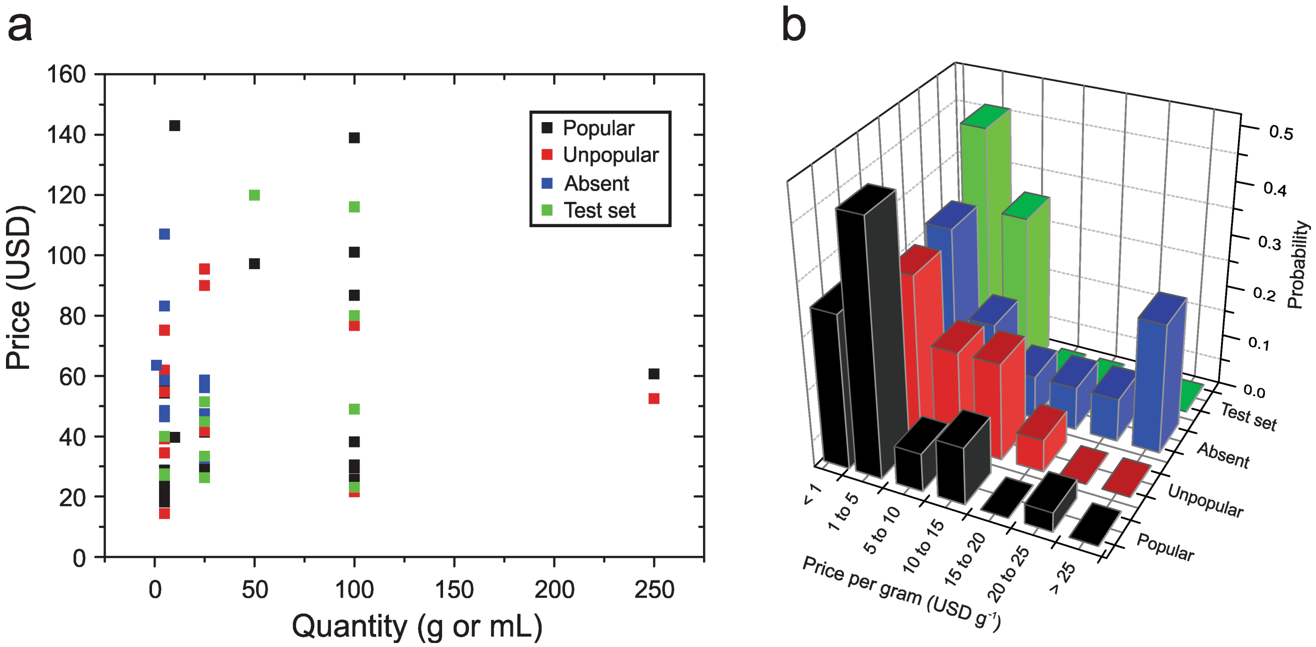
Correspondence and requests for materials should be addressed to S.A.F. or A.J.N. or J.S.

Peer review information *Nature* thanks Leroy Cronin, Edward Kim and Hans Conrad Zur Loye for their contribution to the peer review of this work.

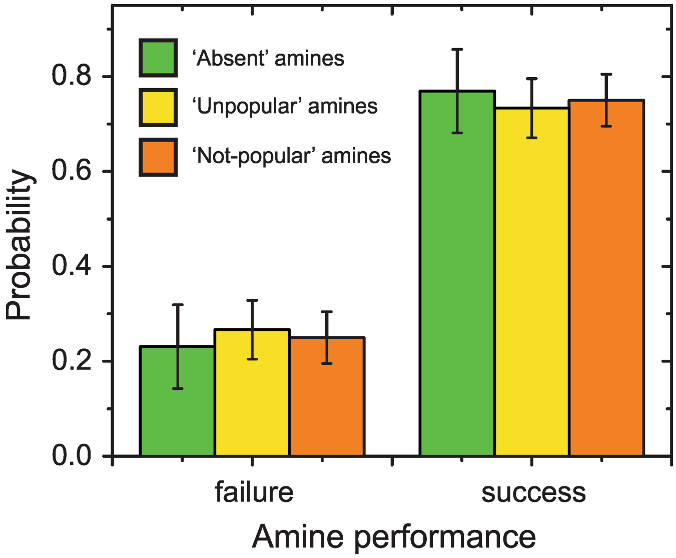
Reprints and permissions information is available at http://www.nature.com/reprints.



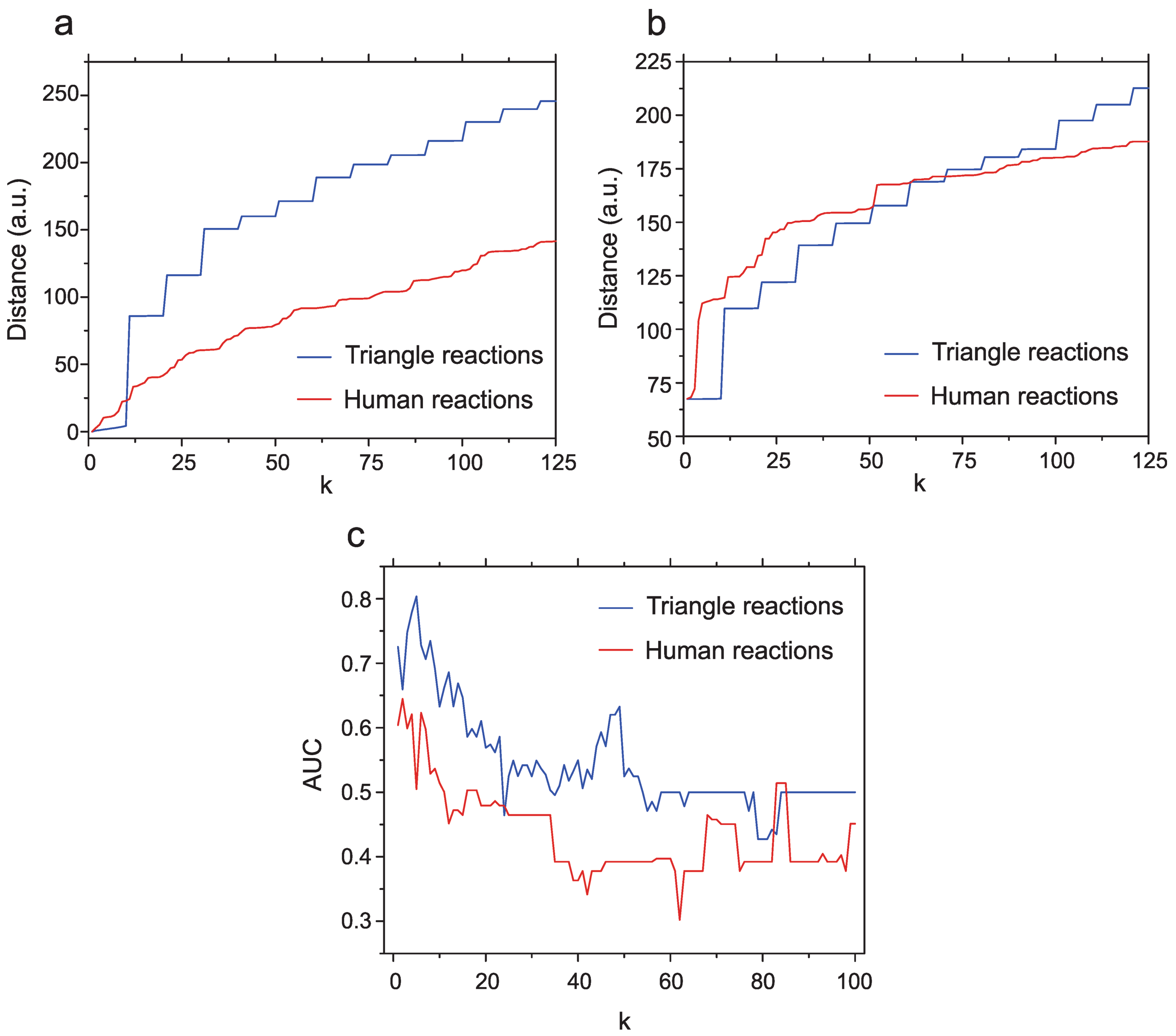
Extended Data Fig. 1 | Cambridge Structural Database (CSD) search results for templated metals borates. a, A plot of the number of unique structures for each amine, ordered from the amine with the fewest structures to the most. **b**, A plot of cumulative probability versus amine proportion. The grey rectangle represents the Pareto split.



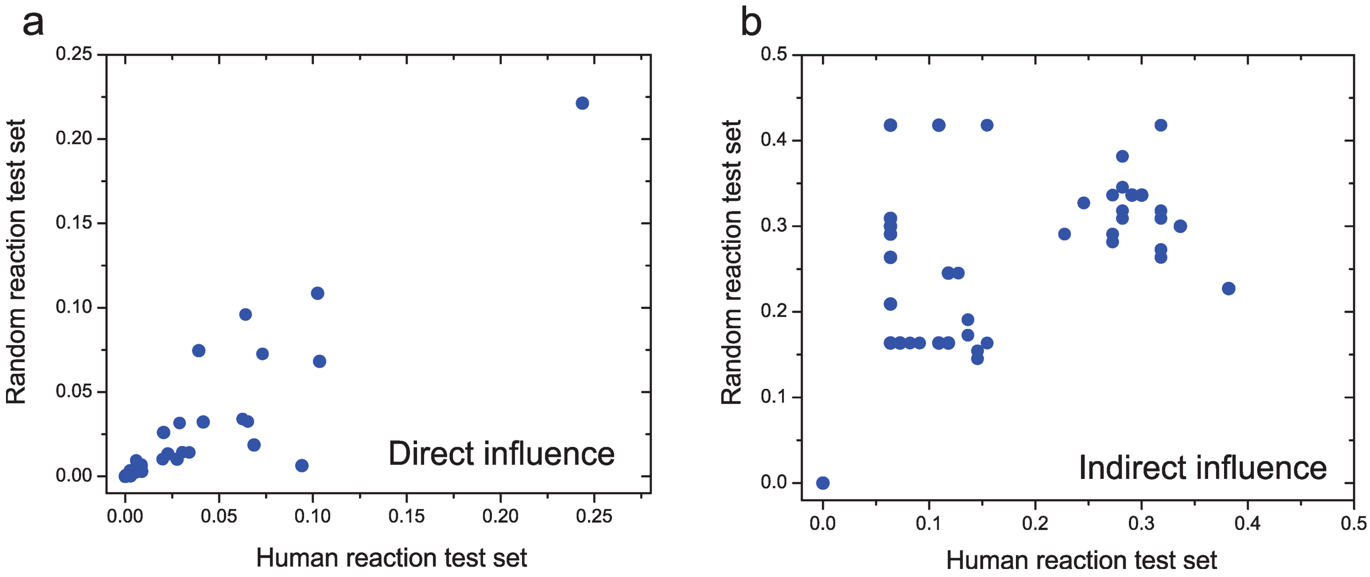
Extended Data Fig. 2 | Amine price and availability. a, Amine price versus quantity for the randomized reaction amines. The data are separated by amine popularity (popular, unpopular or absent). Amines used in the test set experiments are also included. **b**, Amine pricing information for those used in the randomized reactions. The price per gram was calculated assuming amine densities of 1 g ml^{-1}. The data presented in the figures above suggest that there is no systematic difference in amine prices between the popular, unpopular and absent amines. Additionally, the distribution of amine pricing for the test set amines is similar to the other distributions, suggesting a representative sample of amines.



Extended Data Fig. 3 | Outcome probabilities for not-popular, unpopular and absent organic amines. The not-popular set includes the unpopular and absent amines.

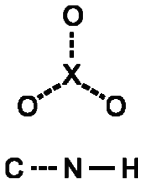
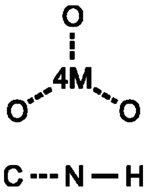
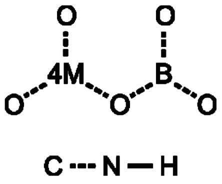


Extended Data Fig. 4 | Average nearest-neighbour distances in the datasets, and nearest-neighbour choices on model performance. **a**, Average distances to the *k*th nearest neighbour within each training set. **b**, Average distances from each training set to the *k*th nearest neighbour within the test set. **c**, AUC for the *k*th nearest neighbour classifier for $k = 1$ to 100.



Extended Data Fig. 5 | Comparison of the influence of direct and indirect features. a, Direct influence values of descriptors in the human reaction test set versus the random reaction test set. **b**, Indirect influence values of descriptors in the human reaction test set versus the random reaction test set.

Extended Data Table 1 | Structure inclusion and exclusion criteria

Inclusion group 1	Inclusion group 2	Inclusion group 3
Oxides	Metal oxides	Metal borates
		

	Exclusion groups	
Group 1	Group 2	Group 3
4M ··· C	O ··· C ··· N	O ··· S ··· N
4M ··· N	O ··· C ··· C ··· N	O ··· S ··· C
O ··· C ··· C ··· C	O ··· C ··· C ··· C ··· N	S ··· C ··· N
4M ··· 4M	O ··· C ··· C ··· C ··· C ··· N	P ··· C ··· N
4M ··· P	P ··· C ··· O	P ··· C ··· C ···N
	P ··· C ··· C ··· O	P ··· C ··· C ··· C ··· N
	P ··· C ··· C ··· C ··· O	P ··· C ··· C ··· C ··· C ··· N

Structures were identified in the Cambridge Structure Database (CSD) using a combination of inclusion and exclusion criteria. The inclusion criteria, shown above, were created to be inclusive but to still return appropriate structure. Bond orders were left unspecified to avoid unintended exclusions. The labels ‘X’ and ‘4M’ represent ‘any atom type’ and ‘any metal’, respectively. The three exclusion groups were constructed to exclude more complex structures in the organic amines and bonding to the metal centres through atoms other than oxygen. The structures in each compound class (oxides, metal oxides and metal borates) were identified by conducting three distinct searches, each of which included the inclusion group and one of the exclusion groups. The resulting three datasets were merged so that only the structures present in all three datasets were retained.

Extended Data Table 2 | Matthews correlation coefficient (MCC), accuracy and AUC results for each machine-learning algorithm, trained on either the human-selected or randomly generated reaction data using all features

	Classifier	Logistic regression	kNN (k = 2)	kNN (k = 5)	Linear SVM (C = 1)	Decision tree	Random forest	Naïve Bayes
MCC	Human	-0.17	0.31	0.01	-0.02	0.11	-0.01	0.16
	Triangle	0.25	0.35	0.59	-0.01	-0.04	0.22	0.16
Accuracy	Human	0.44	0.69	0.51	0.53	0.50	0.51	0.63
	Triangle	0.59	0.71	0.79	0.49	0.47	0.61	0.57
AUC	Human	0.41	0.64	0.50	0.49	0.55	0.49	0.57
	Triangle	0.62	0.66	0.80	0.50	0.48	0.61	0.58

Extended Data Table 3 | Feature selection comparison

a Top 5

Human	Randomized
_feat_bpKa1	_rxn_pH
_feat_bpKa2	_feat_bpKa1
_feat_RotatableBondCount	_feat_donsitecount
_feat_LengthPerpendicularToTheMinArea	_feat_fr_NH2
_feat_ASA_P	_calc_pbKaUnderPhCount

b Top 10

Human	Randomized
_feat_bpKa1	_rxn_pH
_feat_bpKa2	_feat_bpKa1
_feat_ChiralCenterCount	_feat_bpKa2
_feat_RotatableBondCount	_feat_RingAtomCount
_feat_MaximalProjectionRadius	_feat_CyclomaticNumber
_feat_LengthPerpendicularToTheMinArea	_feat_PolarSurfaceArea
_feat_ASA_P	_feat_donsitecount
_feat_PolarSurfaceArea	_feat_fr_NH2
_feat_donorcount	_feat_fr_NH0
_feat_donsitecount	_calc_pbKaUnderPhCount

c Top 20

Human	Randomized
_raw_VOSO4xH2O/g	_raw_H3BO3/g
_raw_VOSO4xH2O_Mol	_rxn_pH
_feat_AtomCount_N	_feat_bpKa1
_feat_bpKa1	_feat_bpKa2
_feat_bpKa2	_feat_AromaticRingCount
_feat_Aliphatic AtomCount	_feat_AromaticAtomCount
_feat_ChiralCenterCount	_feat_ChainAtomCount
_feat_RotatableBondCount	_feat_RingAtomCount
_feat_HyperWienerIndex	_feat_SmallestRingSize
_feat_WienerIndex	_feat_LargestRingSize
_feat_MaximalProjectionRadius	_feat_fsp3
_feat_LengthPerpendicularToTheMinArea	_feat_HeteroaromaticRing Count
_feat_ASA+	_feat_CyclomaticNumber
_feat_ASA_P	_feat_PolarSurfaceArea
_feat_PolarSurfaceArea	_feat_donorcount
_feat_acceptorcount	_feat_donsitecount
_feat_Accsitecount	_feat_fr_NH2
_feat_donorcount	_feat_fr_NH0
_feat_donsitecount	_feat_fr_pyridine
_feat_fr_NH2	_calc_pbKaUnderPhCount

ANOVA *F*-values for the human-generated and randomized reaction test sets.
a, Top 5 features. **b**, Top 10 features. **c**, Top 20 features.

Extended Data Table 4 | Comparison of discrepancies between model predictions and reaction outcomes

Amine	Discrepant predictions (of 10,000)	Discrepancies predicted positive by randomly-generated data model	Discrepancies predicted positive by human-data model
Homopiperazine (1,4-diazepane)	1940	1940 (1/10)	0 (—)
N, N, N', N'-tetramethyl-1,3,-diaminobutane	5404	5190 (8/10)	214 (8/10)
1,4-Dimethylpiperazine	1390	344 (4/10)	1046 (8/10)
Ethylenediamine	478	478 (3/10)	0 (—)
N-(2-aminoethyl)piperidine	3404	3404 (1/10)	0 (—)
2,2-Dimethylpropane-1,3-diamine	0	0 (—)	0 (—)
2-methyl-1H-imidazole	3088	3088 (4/10)	0 (—)
4-Methylpiperazin-1-amine	0	0 (—)	0 (—)
N,N,N',N'-tetramethylhexane-1,6-diamine	3373	3373 (8/10)	0 (—)
1,3-Diiminoisoindoline	1948	1948 (9/10)	0 (—)
2-Methylpyrazine	0	0 (—)	0 (—)

Ten thousand random reactions were generated for each amine. The first column in Extended Data Table 4 indicates the number of discrepancies between the predictions of the two models. Subsequent columns show the number of those discrepancies predicted to be positive by the respective model (top line); of these positive predictions the ten reactions with the lowest model uncertainty were selected and performed in the laboratory. Successful outcomes are indicated as a fraction in parentheses (number of successful reactions out of total number of trials). For amines where no positive predictions were made, no tests were performed, indicated by (—).