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]

[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, NY, USA

Summary

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17% of amine reactants comprise 79% of reported compounds, consistent with social influence models.[10-12] Analysis of unpublished historical laboratory notebook records shows similarly biased distributions of reaction condition choices.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

Main

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Scientific publications do not provide a representative dataset.[12] Confirmation bias favors publishing positive results, yet the missing “failures” are essential for modeling 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 graph attachment ("rich get richer") mechanisms result in power law distributions of citations, resulting in disproportionately popular articles.[10,11] The emerging "Science of Science" attempts to quantify the role of social interactions on problem selection, career trajectory, and citations.[13] Studies of scientific decision *errors* have tended to focus on individual-specific causes, such as differences in the classification variability and decision inconsistency.[14,15] Such error processes produce uncorrelated individual-specific noise in the resulting dataset. However, systematic errors in scientific experiment planning have not been studied. In general, social influences such as knowledge about others' choices can cause popularity to diverge from the underlying quality of the item.[3] Socially-influenced scientific decision errors have been widely speculated, but never explicitly confirmed. The distribution of reported medicinal chemistry compounds is unrelated to the intended application, cost, or reaction difficulty.[16] Disproportionately few drug scaffolds comprise the majority of antimalarial[17] and other drug-candidate molecules,[18] and popularity is uncorrelated to synthetic feasibility or biological activity. Medicinal chemists make new pharmaceutical molecules resembling those they have synthesized in the past,[19] using a limited set of reactions,[20] whose choice is uncorrelated to cost, estimated synthetic ease, or properties of the reactants or products.[21] However, an over-representation of a particular experimental choice need not be irrational. For example, 36% of entries in the Protein Data Bank (PDB) report using polyethylene glycol (PEG) additives, which under-represents the true success rate of 59%, including many proteins that cannot be crystallized using other additives.[22] This suggests that lack of crystallization additive diversity 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 signatures of anthropogenic influence.

What is the evidence of bias in reactant choices for organically templated metal oxides synthesis? Incorporating different organic amines results in compounds with diverse composition, local and extended connectivity, and functionality,[9] so one expects to see the broadest possible range of amine choices. The discipline defines "success" as formation of a crystal of sufficient size and quality to yield a stable single-crystal X-ray diffraction refinement. Publishers require structures to be deposited in the Cambridge Structural Database (CSD). The number of reported compounds is a proxy for experimental effort and success for a particular amine. The CSD contains 5010 amine-templated metal oxides structures, containing 415 unique amines. The 17% most observed amines (70 individual molecules) are found in 79% of the structures (3947 distinct CSD entries), whereas the remaining 83% (345) of the amines are found in just 21% of the structures (1063 entries), see Figure 1. (Structures containing multiple amines preclude an even split.) The gini coefficient is 0.654. A log probability-log rank plot (Figure 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 Figure 1). Unpublished experimental records provide evidence of bias in attempted reaction conditions, such as pH and reactant quantities. Our dark reactions database (http://darkreactions.haverford.edu) contains 557 hydrothermal vanadium borate

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis” *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

reactions (inspired by the first report in the literature[23]) consisting of three years of work by three students prior to the start of this study. Human reactions are almost exclusively conducted at pH=8, (Figure 2a) with unevenly distributed amine quantities (Figure 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 is easily trapped in local optima.[24]

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 from (personal, supervisor, literature) experience. The interplay of shorter-timescale communicative memory and longer-timescale cultural memory,[11] results in power law distributions like Figure 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.[25] For example, aesthetic biases linking symmetry with positive affect,[26] result in "irrational" affective criterion favoring experimentation with symmetric molecules. Alternatively, humans more easily discriminate and recall symmetric three-dimensional objects,[27] favoring symmetrical molecules in recall and attention heuristics used to create new experiments.[2] Both scenarios yield unrepresentative datasets (i.e., datasets that are "biased"), despite radically different mechanisms. Given the many possible types of anthropogenic influence and the difficulty of distinguishing them experimentally,[25] we instead demonstrate the presence of *an* anthropogenic influence, without assigning a specific psychosocial origin, by eliminating alternative explanations.

Non-anthropogenic factors can be classified following a classical four-fold theory of causes.[28] *Efficient* causes are the technical ability to perform the experiments. Hydrothermal syntheses have been conducted for over 50 years,[9] and the reagent and reaction condition choices present no pressure or corrosion resistance challenges. *Final* causes favor particular product materials having desired technological properties. Because functional diversity follows from structural diversity,[9] one expects the broadest diversity in the public databases. *Material* causes, specifically reagent availability, were excluded by considering a structurally diverse (primary through tertiary amines, and linear, branched, cyclic, and aromatic molecules) set of 55 commercially available amines containing 27 popular, 16 unpopular, and 12 absent from the CSD. All selected amines are commercially available in 5-gram quantities from major suppliers, and there is no significant difference in cost (Extended Data Figure 2).

The only remaining non-anthropogenic cause is eliminated by experiment. The *formal* cause is the intrinsic propensity of some reactants and reaction condition choices to yield crystals. For an unbiased assessment, we generated 10 random reactions for each of the 55 amines described above, by selecting random pH and amine quantities using two independent triangular distributions, with the literature precedent as the peak, and physically motivated upper and lower bounds. See Methods section. The goal is *not* to efficiently explore the chemical space, but rather to establish a *neutral estimate* of the “reaction cross section” for each amine revealing systematic reactivity differences between popular and not-popular amines centered around where humans are likely to have attempted to search. However, we note that random choices are often better than human expertise and comparable to more sophisticated numerical

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis” *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

methods in fields as diverse as oil exploration,[29] chemical reaction discovery,[30] and numerous social and financial applications.[31]

Randomized reaction outcomes for popular and not-popular (unpopular and absent) amines were ranked using the four-class crystal scoring system described in the Methods section, with the stringent criterion where "success" consists only of crystals comparable to those used for the CSD data. All four reaction outcomes occur nearly equally in randomly chosen reactions for popular and not-popular amines (Figure 3a; similar plots separating unpopular and absent outcomes are shown in Extended Data Figure 3). Therefore, any single randomly chosen reaction is equally likely to be successful regardless of popularity. 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 modeled this as the observation of success (score = 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 (Figure 3b). There is no support for intrinsic reaction propensity difference between popular and not-popular amines. In fact, popular amines were *less* likely to successfully form crystals than not-popular amines in our experiments, but this (one-sided) success rate difference occurs with p = 0.26 in a random permutation. Having excluded non-anthropogenic *material, efficient, formal,* and *final* causes, only an anthropogenic explanation for the observed reactant choice distribution remains.

The random reaction outcomes also expose anthropogenic influence in reaction condition choices. Human-selected reactions are biased to lower amine amounts and peaked around the literature precedent (Figure 2c). In contrast, we find limited dependence of the reaction outcome on amine choice, as the distribution for successful (outcome = 4) and failed (outcome = 3, 2, 1) reactions mimics the triangular distribution that generate them. Human pH choices are almost exclusively based on a literature precedent of pH 8. (Figure 2d) Our random reactions indicate that although higher pH is (*ceteris paribus*) 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 these randomly-generated (unbiased) reactions, evaluated on a true time-separated holdout test set of 110 additional vanadium borate experiments, consisting of 10 randomly-generated reaction conditions for each of 11 amines. No test amines were present in either training set. For the purposes of training, only the subsets of the human-generated and random datasets containing the 37 amines common to both sets were used, reducing the training set sizes to 467 and 370 reactions, respectively. (Our human-generated training set has more diverse amine choices than typically present in the literature, as it used amines from our previous diversity-oriented studies.[6]) Restricting the dataset in this way means differences are solely due to reaction condition choices. To give each training set the best chance to succeed, a variety of classifiers were tested (logistic regression, k nearest neighbors, 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 results is presented in Tables S8-S21, and summarized in Extended Data Table 2. The best classifier trained on the human-generated dataset was a k nearest neighbors (k=2) with an accuracy of 69% and an AUC of 0.64 on the held-out test set. The best classifier trained by the randomly generated training set was k nearest neighbors (k=5) with an accuracy of 79% and an AUC of 0.80. Thus, the randomly-generated training set outperforms the human-generated set by all metrics, despite containing 20% fewer reactions.

The performance gain results from better reaction condition sampling, which can be established by the average nearest-neighbor distance between reactions (Extended Data Figure 4). The average distance to the *k*-th nearest neighbor *within* a given training set is *greater* for the randomly generated training set for $k>10$, indicating that it more comprehensively samples the chemical space. Furthermore, the average distance from an experiment in the training set to the *k*-th nearest neighbor in the *test set* is *smaller* for the randomly-generated training set for $k \leq 60$, indicating that it allows for better generalization to the test set. Both factors contribute to the randomly-generated experiments being more informative than the human-selected reactions.

Anthropogenic dataset bias obscures chemical insights. Because the two training sets contain the same amines, the dependence of reaction outcomes on amine structural and physicochemical property features should be equally well described. Indeed, the *direct* influence of the features, the contribution of a feature to the difference between the given predictions and a mean result,[32] is comparable for models built on the different training sets. In contrast, the *indirect* influence of a feature, estimated by computing the degradation in model accuracy when the ability of the model to predict it from the other features is removed,[32] is linearly correlated, with the exception of 6 features describing the amine properties of solvent accessible polar surface area, presence of rings, rotatable bonds, and presence of amidine moieties in the organic molecules. (Extended Data Figure 5) Computationally obscuring these features in random-reaction-trained model degrades the model performance, but computationally obscuring them in the human-reaction-trained model does not, because anthropogenic reaction selection has *implicitly* obscured these feature contributions.

Anthropogenic bias hinders the discovery of new materials. Only 41/110 test reactions successfully produced a sufficient quality and size crystalline product, and the positive recall scores were 46% and 85% for the human- and randomly-generated data models, respectively. (Tables S9 & S17) The two models disagree for 23/110 test outcomes, and in every case the human model predicts failure whereas the random model predicts success. The "pessimism" of the former is consistent with loss-aversion bias in the 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 7/11 amines in the test set had at least one successful reaction outcome sufficient for discovery of a new material. The human model failed to identify 2/7 of these compounds, whereas the random-data model found at least one successful reaction for all 7/7 compounds. Therefore, models trained on the unbiased dataset

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

are both quantitatively and qualitatively better at identifying reaction conditions, and find compounds that would otherwise remain undiscovered.

Models trained on anthropogenic data less effectively select new experiments. We generated 10,000 random reactions for each of the 11 test amines. Predictions by the two models agree on 81% of these reactions, including *all* generated experiments for 3/11 amines. For the 8 amines where the models have any disagreement, the human-data model makes unique predictions of success for only 2/8 amines, whereas the model trained on unbiased data identifies unique positive predictions for all cases. As a test, we conducted additional laboratory syntheses of 10 discrepant positive predictions made by each model for each amine (totaling 100 additional reactions). For the two amines where both models made different positive predictions, the anthropogenic data model was slightly more successful (16/20 positives found) than the model trained on random experiments (12/20). However, for the other 6 amines in which *only* the randomly-generated data model made unique positive predictions, at least one successful reaction was observed in *all* cases. The relatively low (43%) aggregate precision is because these are more speculative reactions where the models are less confident about the outcome; precision increases with the predicted model probability and is as high as 80%. (Extended Data Table 4 and Supplementary Tables S22 and Figures S1-S2 contain a complete analysis of the laboratory and computational results.) This confirms that models trained on data without anthropogenic bias better identify 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, for example by designing experiments with structured randomness. A simple process of (i) listing all experimental options, (ii) defining distributions that exclude impossible (based on known physical considerations such as solubility or protonation state) or or practically infeasible (e.g., costs and safety) choices, and then (iii) randomly sampling from those distributions removes anthropogenic bias, is at least as successful as human choices, and greatly improves the value of the resulting datasets for machine learning. Ongoing efforts using anthropogenic data to plan chemical syntheses[5-7] should consider this potential limitation.

References

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

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

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. Int. Ed Engl.* **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* **65**, 125 (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**, 10.1126/science.aao0185 (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). ; Jain, A. N. & Cleves, A. E. Does your model weigh the same as a duck? *J. Comput. Aided Mol. Des.* **26**, 57–67 (2012).
20. 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).
21. 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).
22. Kirkwood, J., Hargreaves, D., O'Keefe, S. & Wilson, J. Analysis of crystallization data in the Protein Data Bank. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **71**, 1228–1234 (2015).
23. 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. Int. Ed Engl.* **36**, 1008–1010 (1997).
24. Duros, V. *et al.* Human versus Robots in the Discovery and Crystallization of Gigantic Polyoxometalates. *Angew. Chem. Int. Ed Engl.* **56**, 10815–10820 (2017). ; 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).
25. Kahneman, D. & Klein, G. Conditions for Intuitive Expertise: A Failure to Disagree.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

American Psychologist **64**, 515-526 (2009).
26. Evans D. W. *et al.* Human Preferences for Symmetry: Subjective Experience, Cognitive Conflict and Cortical Brain Activity. *PLoS One* **7**, e38966 (2012).
27. Liu, Z. & Kersten D. Three-dimensional symmetric shapes are discriminated more efficiently than asymmetric ones. *J. Opt. Soc. Am. A* **20**, 1331-1340 (2003).
28. Falcon, Andrea, "Aristotle on Causality", *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), Available at: https://plato.stanford.edu/archives/spr2019/entries/aristotle-causality (Accessed: 13 May 2019)
29. Menard, W. H. & Sharman, G. Scientific Uses of Random Drilling Models. *Science* **190**, 337–343 (1975).; *ibid* **192**, 206–208 (1976).
30. 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).
31. 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).
32. Adler, P. *et al.* Auditing black-box models for indirect influence. *Knowl. Inf. Syst.* **54**, 95–122 (2018).

Acknowledgements

We thank Gary Cattabriga for software engineering support, and Xiaohui Weng for proofreading the experimental data entries. This project was funded by NSF DMR-1709351. IL was partially supported by a Bryn Mawr College LILAC Summer Internship Funding Program. JS 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. A.L. and J.S. performed statistical analysis. 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 financial interests.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Supplementary information

is available for this paper at https://doi.org/TODO

Reprints and permissions information

is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.S. (jschrier@fordham.edu), A.J.N. (anorquis@haverford.edu) or S.A.F. (sorelle@cs.haverford.edu).

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Methods

Data capture from experiment. Data capture from historical notebooks and from new experiments conducted in this study follows the procedure described in our previous work,[6] and includes compositional information (reactant identities and quantities), categorized by type (organic/inorganic/solvent), reaction conditions (e.g., pH, temperature, time), and reaction-outcome. Reaction outcomes are coded qualitatively based on crystal size, using 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. To eliminate measurement bias, students performing reactions and scoring crystal outcomes were unaware of whether the reagents were "popular" or "not popular". A machine readable collection of all experimental data is provided in the Supplementary Information files.

Analysis of published crystal structure. Amine-templated metal oxides were extracted from the Cambridge Structural Database (CSD)[33] by stipulating both inclusion and exclusion criteria. The inclusion criteria were used to dictate the presence of an oxide, metal oxide or metal borate substructure, in addition to an organic amine. (Defined in Extended Data Table 1.) The exclusion criteria were used to remove structures with bonding motifs that fall outside the targeted family of compounds for this work. This results in 7630 oxides, 4870 metal-oxides, and 115 metal borates. Analysis of the metal-oxides data is presented in the article; parallel analysis of the metal borates is described in the Extended Data Figure 1. We initially attempted to extract the organic components from the 3D structure, but the presence of structural disorder makes this ambiguous. The 2D structure diagrams are not publicly available through the CSD API. 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 includes only molecules comprised of solely C, H, and N, and no nitriles, azo, diazo, or diazonium compounds. After performing these exclusions, 6458 oxides, 4152 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.[34] CACTUS was also used to generate InChI and InChIKey strings. The Python Jupyter notebooks used to perform this process, along with the inputs and intermediate outputs, are provided in the Supplementary Information files.

Pricing. Amine pricing information was collected by searching the Sigma-Aldrich website in November 2018. Pricing and sample size data were collected for the smallest sample size available, assuming that the sample size was at least 5 g. Pricing data were collected for all amines using the triangle reactions and the test set amines. No amines were backordered on the day data were collected.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis” *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Generation of random reactions. Randomized reactions were generated by sampling from triangular distributions for the reaction pH and amine mole amounts. The triangular distribution for pH was chosen with minimum and maximum values of 1 and 8.49, with a mode of pH 8, and then rounded to the nearest integer. (The pH is not easily set below 0 and basic conditions will not protonate the amines.) The triangular distribution for amine mole amounts was chosen with minimum and maximum values of 0.5 mmol to 10 mmol, with a mode of 5 mmol. (Amine quantity cannot go below zero and cannot go above a conservative solubility limit chosen for all amines.) Conversion of amine moles 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 $VOSO_4 \cdot x\ H_2O$, 6.0 g H_2O, and the amine mass drawn from the above 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 hours. 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, as described in the Data Capture section above.

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 were used for all reactions in this study, and only a single organic reactant used per reaction. Therefore, only a subset of the reaction descriptors from our previous work[6] was used in this study. These three categories include reaction parameters (e.g., temperature, pH), physicochemical and structural features of the organic component, and stoichiometric ratios. Structural and physicochemical properties of the organic species were computed using RDKit 2018.03.4[35] and the ChemAxon Calculator Plugins.[35] Tables S1-S7 contain a complete description of 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 feature importances. The feature sets containing all features as well as all features with positive variance were also considered. The full set of considered models were trained on each of these feature sets for both the human and the randomly-generated training sets. As described in the main text, the full feature set had the top performance model based on accuracy. A 5-feature F-test-based feature set had the highest AUC (0.69), but lower accuracy (0.63), of the human-set-trained models. In general, the human-set-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

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Data Table 3. All models were implemented in Python 3.7.3 using Scikit-Learn 0.19.1;[44] 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).[37,38] The indirect influence of each feature was calculated using the BlackBoxAuditing code[32, 39] in order to measure each feature's contribution to model accuracy even when it is not directly used in the model. This influence is estimated by obscuring a feature so 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 and randomly-generated training sets. Comparison plots are shown in Extended Data Figure 5.

Data availability. The authors declare that all data supporting the findings of this study are available within the article and its Supplementary Information files.

Code availability. The code used for this project is available in the Supplementary Information files.

Methods References

33. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* **72**, 171–179 (2016).
34. Landrum, G. RDKit: Open-source cheminformatics. (2018). Available at: http://www.rdkit.org.
35. ChemAxon. JChem cxcalc 5.2.0. (2018). Available at: http://www.chemaxon.com.
36. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
37. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
38. SHAP (SHapley Additive exPlanations). Available at: https://github.com/slundberg/shap. (Accessed: 13 May 2019)
39. Available at: https://github.com/algofairness/BlackBoxAuditing. (Accessed: 13 May 2019)

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Legends for Extended Data Figures

Extended Data Table 1 | Structure inclusion and exclusion criteria. 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, while still returning appropriate structure. Bond orders were left unspecified to above 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 metals centers through atoms other than oxygen. The structures in each compound class (oxides, metal oxides, 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 Figure 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 vs amine proportion. The grey rectangle represents the Pareto split.

Extended Data Figure 2 | Amine price and availability. a, Amine price vs quantity for randomized reaction amines. 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 randomized reaction amines. Price per gram values were calculated assuming amine densities of 1 g mL^{-1}. The data presented in the figures above suggest that there is no significant different in amine prices between the popular, unpopular and absent amines. Additionally, the amine pricing distribution for the test set amines is similar to the other distributions, suggesting a representative sample of amines.

Extended Data Figure 3 | Outcome probabilities for not-popular, unpopular and absent organic amines.

Extended Data Figure 4 | Average nearest-neighbor distances in the datasets and nearest-neighbor choices on model performance. a, Average distances to the kth nearest neighbor within training set. **b,** Average distances to the kth nearest neighbor within test set. **c,** AUC on kNN classifier for k=1 to 100.

Extended Data Figure 5 | Direct and indirect feature influence comparison. a, Direct descriptor influence values in the human reaction test set versus the random reaction test set. **b,** Indirect descriptor influence values in the human reaction test set versus the random reaction test set.

Extended Data Table 2 | Matthews correlation coefficient (MCC), accuracy and AUC results for each machine learning algorithm, trained on either the human or triangle test set using all features.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Extended Data Table 3 | Feature selection comparison. **a,** Top 5, **b**, top 10 and **c,** top 20 features using ANOVA F-values for the human and randomized reaction test sets.

Extended Data Table 4 | Comparison of discrepancies between model predictions and reaction outcomes. 10,000 random reactions were generated for each amine. The first column in Table S29 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), and of those positive predictions the 10 reactions with the lowest model uncertainty were selected and performed in the laboratory. The successful outcomes are indicated as a fraction in parentheses (number successful/number trials). For amines where no positive predictions are made, no tests were performed, indicated by (—).

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

List of Figures.

Figure 1. Amine occurrence in reported metal oxide crystal structures. (a) Number of crystals structures observed, sorted by the amine's rank. The inset shows the same data as a log probability-log rank plot. (b) Cumulative probability of structures as a function of the proportion of unique amines, ordered from those with the most structures (and highest probability) to the least). The shaded region represents the Pareto split, in which 17% of amines comprise 79% of the structures.

Figure 2. Distribution of reaction parameter choices and outcomes. The distribution of reaction parameters ((a) reaction pH and (b) amine mmol). Blue indicates the distribution of human-selected reactions taken from a historical dataset of 557 reactions; grey indicates the triangular distribution defined for generating random experiments; orange indicates the distribution of the 548 random reactions performed in this study. Distributions of reaction outcomes, based upon (c) reaction pH and (d) amine mmol. Successful reaction outcomes are indicated by green and failure indicated by red, for the randomly generated reactions performed in this study.

Figure 3. Reaction outcomes from randomly-generated experiments, separated by 'popular' and 'not-popular' (unpopular and absent) amines. (a) Probability of each outcome on a per reaction basis, using the outcome scale described below. (b) Probability of observing at least one successful reaction (success) or not (failure) for a given amine. Error bars indicate bootstrap estimate of the standard deviation.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

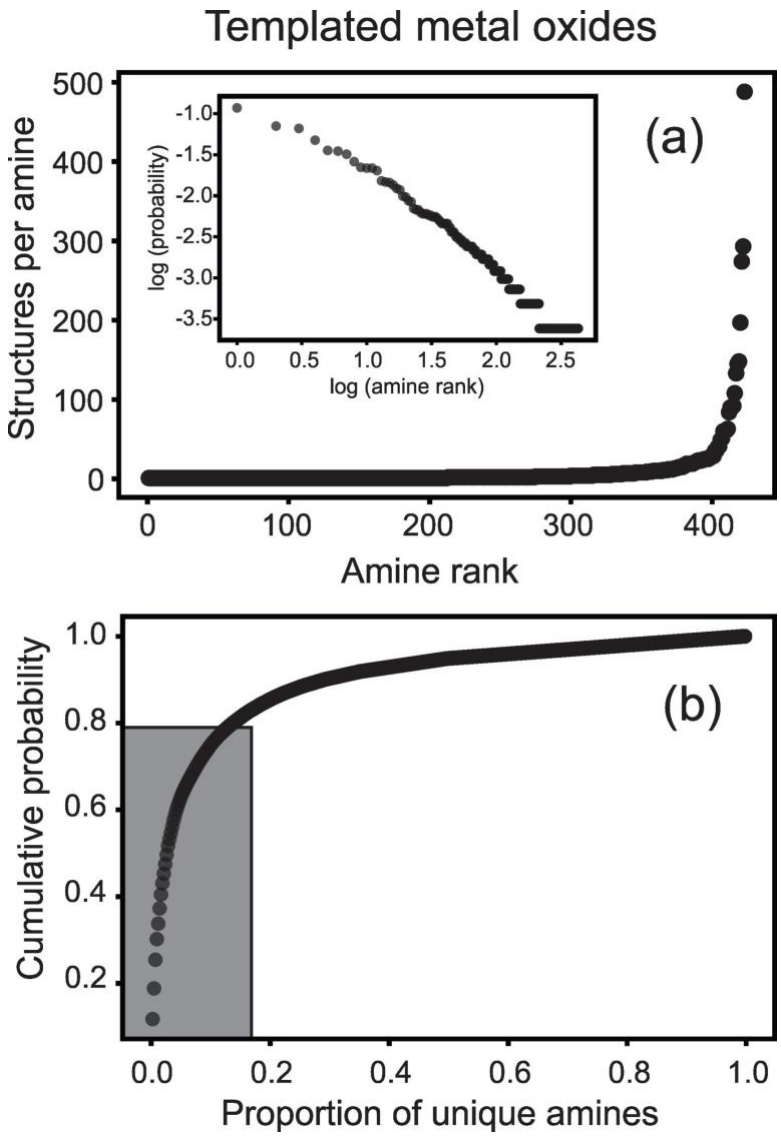


Figure 1. Amine occurrence in reported metal oxide crystal structures. (a) Number of crystals structures observed, sorted by the amine's rank. The inset shows the same data as a log probability-log rank plot. (b) Cumulative probability of structures as a function of the proportion of unique amines, ordered from those with the most structures (and highest probability) to the least). The shaded region represents the Pareto split, in which 17% of amines comprise 79% of the structures.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

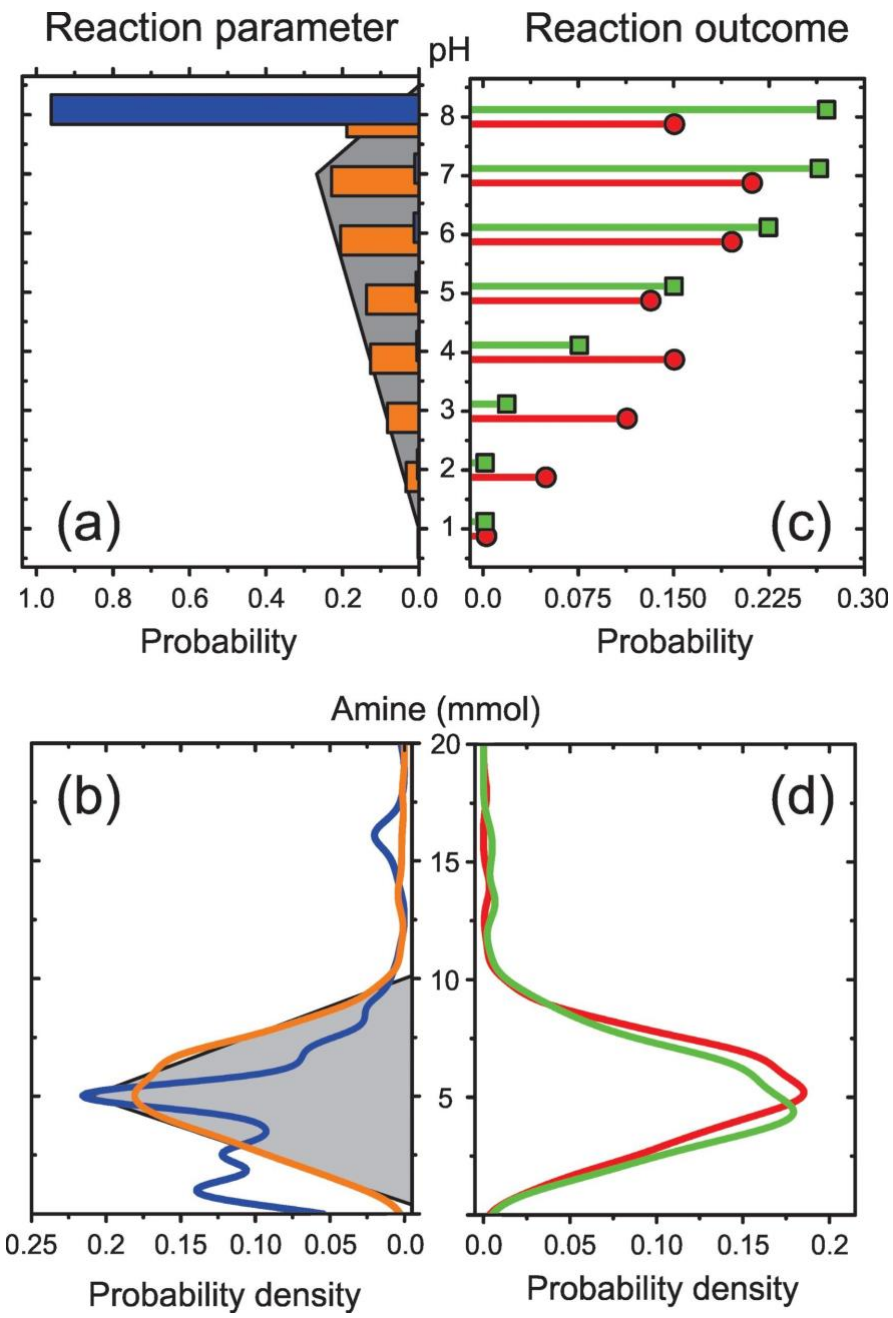


Figure 2. Distribution of reaction parameter choices and outcomes. The distribution of reaction parameters ((a) reaction pH and (b) amine mmol). Blue indicates the distribution of human-selected reactions taken from a historical dataset of 557 reactions; grey indicates the triangular distribution defined for generating random experiments; orange indicates the distribution of the 548 random reactions performed in this study. Distributions of reaction outcomes, based upon (c) reaction pH and (d) amine mmol. Successful reaction outcomes are indicated by green and failure indicated by red, for the randomly generated reactions performed in this study.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

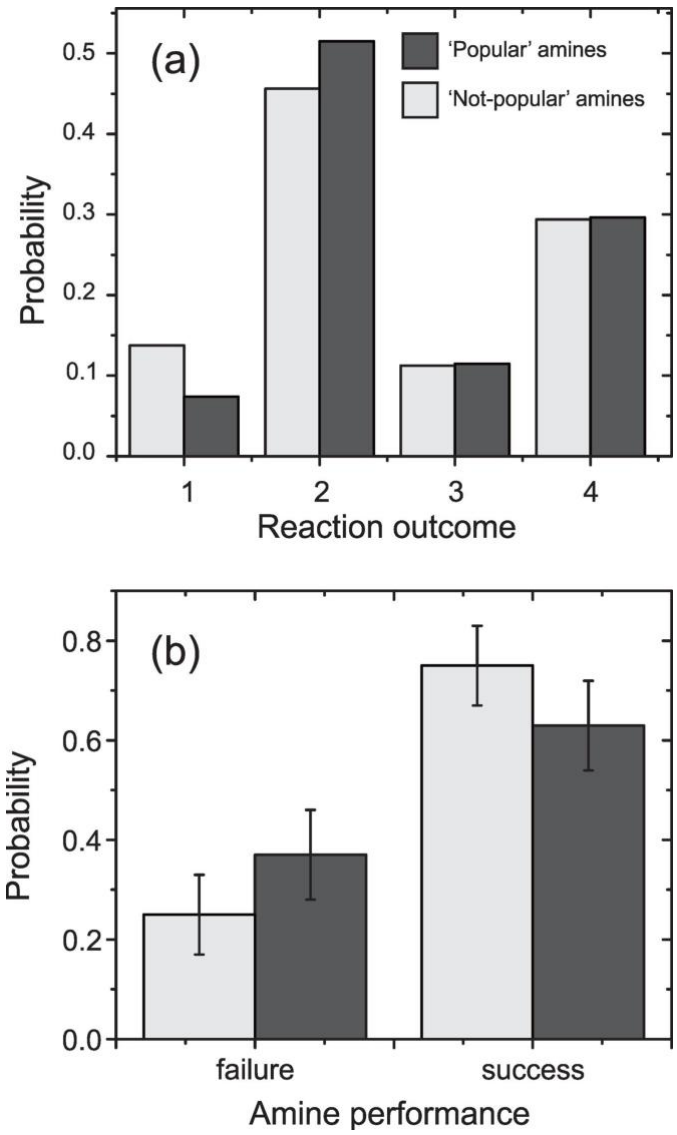
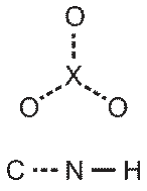
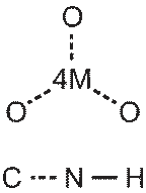
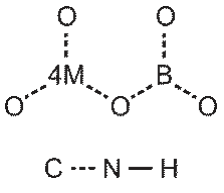


Figure 3. Reaction outcomes from randomly-generated experiments, separated by 'popular' and 'not-popular' (unpopular and absent) amines. (a) Probability of each outcome on a per reaction basis, using the outcome scale described below. (b) Probability of observing at least one successful reaction (success) or not (failure) for a given amine. Error bars indicate bootstrap estimate of the standard deviation.

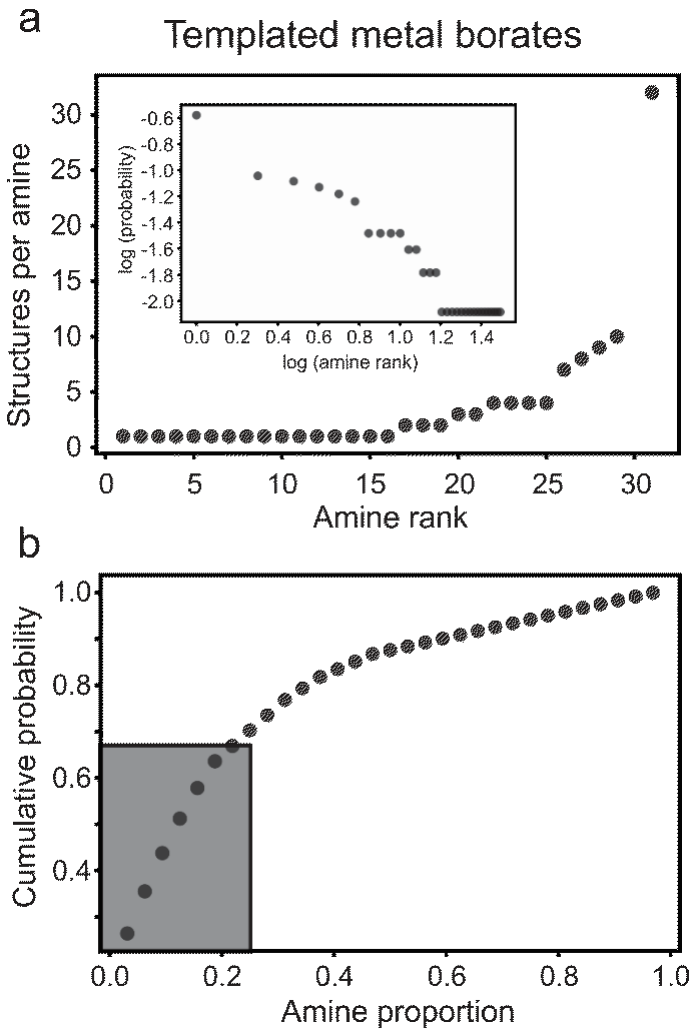
Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Inclusion group 1	Inclusion group 2	Inclusion group 3
Oxides	Metal oxides	Metal borates
		
C ··· N — H	C ··· N — H	C ··· N — H

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

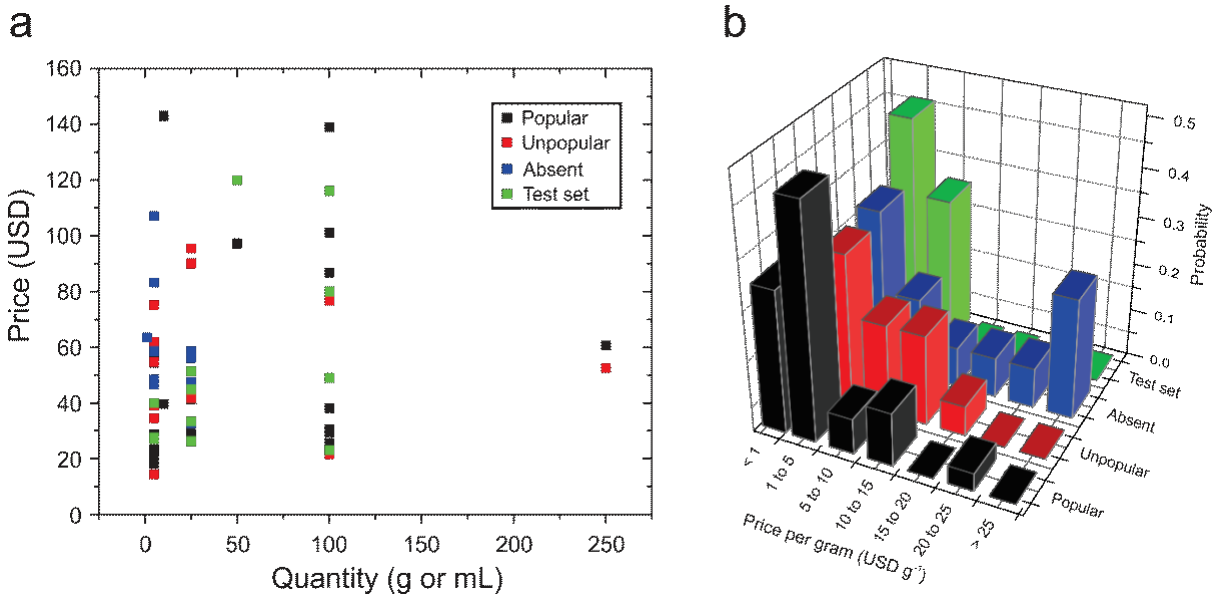
Extended Data Table 1 | Structure inclusion and exclusion criteria. Structures were identified in the Cambridge Structure Database (CSD) using a combination of inclusion and exclusion criteria. The inclusion criteria, shown above, were created to be inclusive, while still returning appropriate structure. Bond orders were left unspecified to above 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 metals centers through atoms other than oxygen. The structures in each compound class (oxides, metal oxides, 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.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5



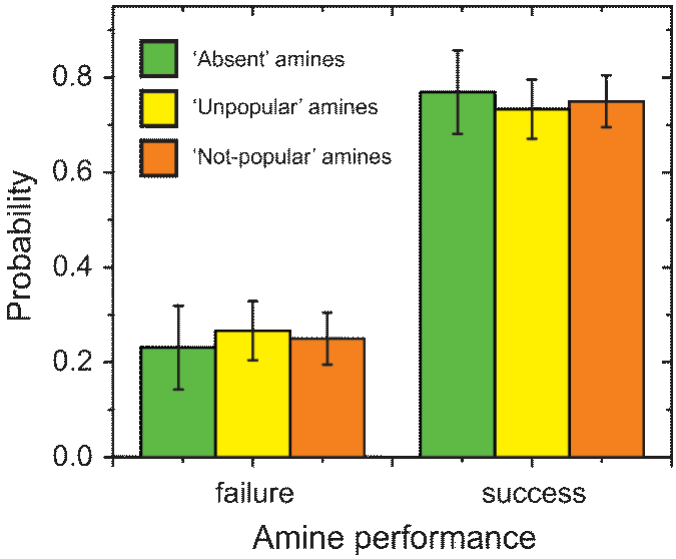
Extended Data Figure 1 | Cambridge Structural Database (CSD) search results for templated metals borates. **a,** a plot of the number of unique structures for each amine, ordered from the amine with the fewest structures to the most. **b,** a plot of cumulative probability vs amine proportion. The grey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

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5



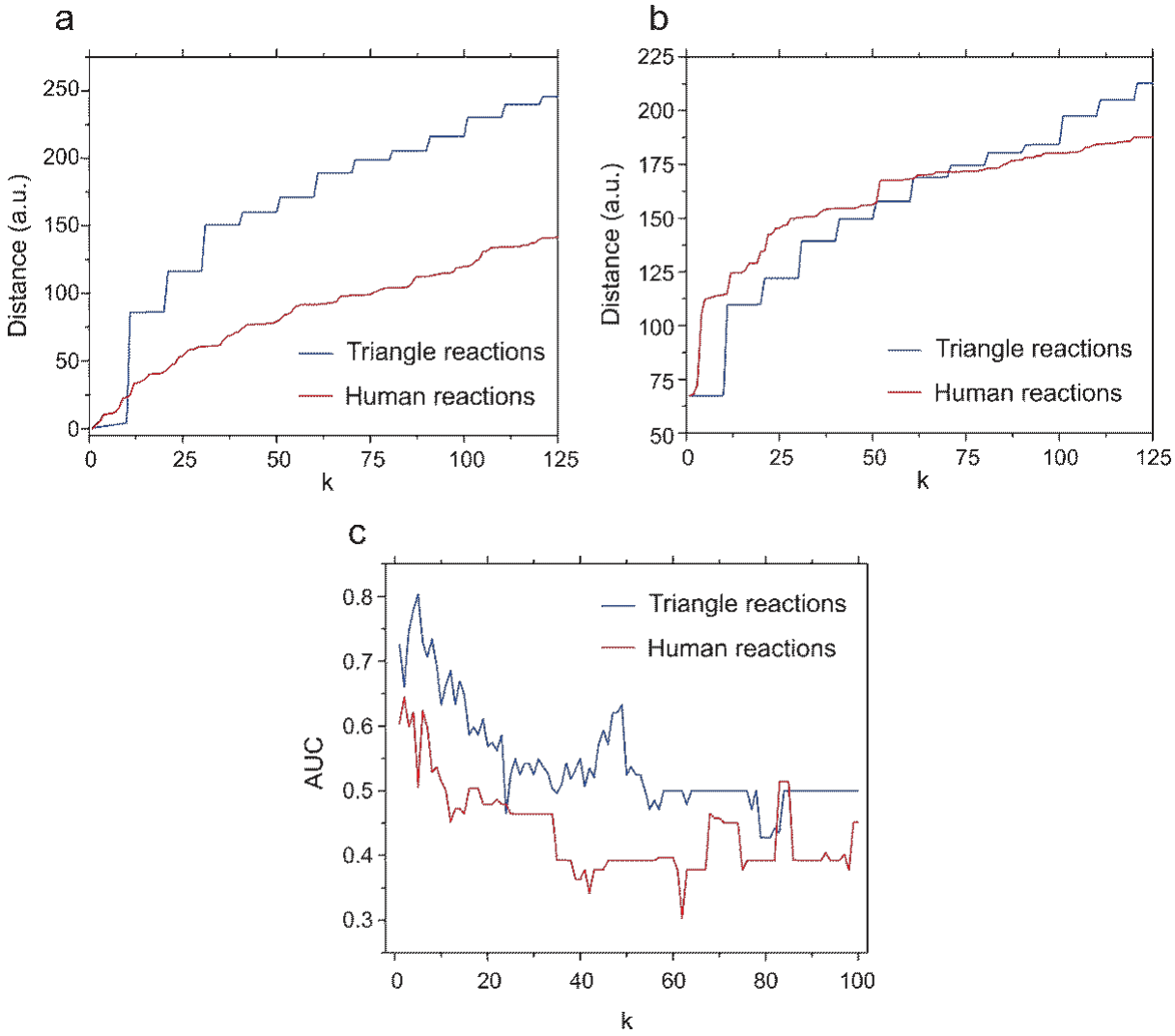
Extended Data Figure 2| Amine price and availability. **a,** Amine price vs quantity for randomized reaction amines. Data are separated by amine popularity (popular, unpopular or absent). Amines used in the test set experiments are also included. b, Amine pricing information for randomized reaction amines.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

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5



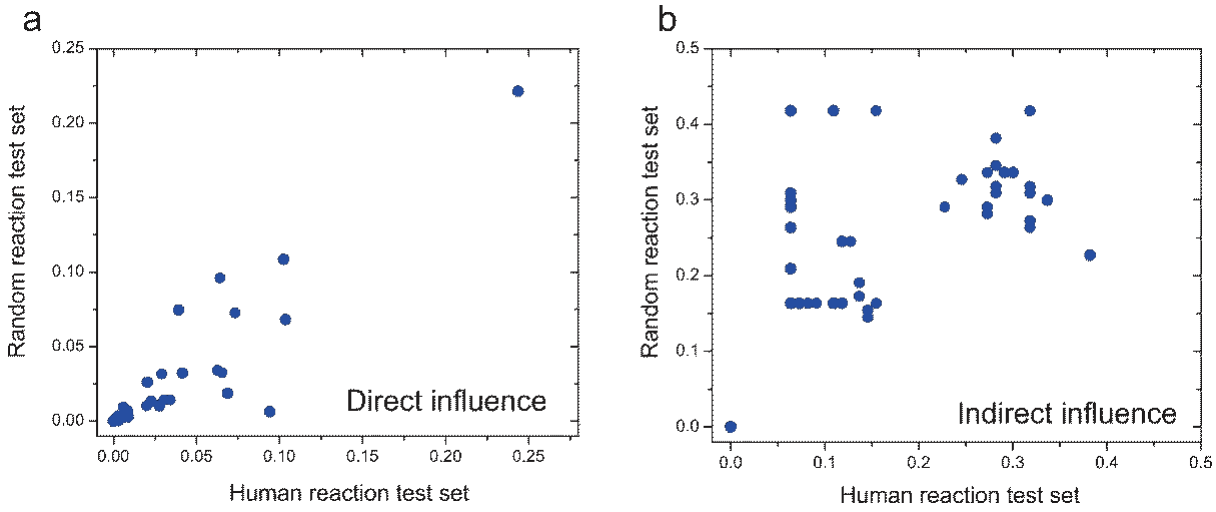
Extended Data Figure 3 | Outcome probabilities for not-popular, unpopular and absent organic amines.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5



Extended Data Figure 4 | Average nearest-neighbor distances in the datasets and nearest-neighbor choices on model performance. a, Average distances to the kth nearest neighbor within training set. **b,** Average distances to the kth nearest neighbor within test set. **c,** AUC on kNN classifier for k=1 to 100.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5



Extended Data Figure 5 | Direct and indirect feature influence comparison. a, Direct descriptor influence values in the human reaction test set versus the random reaction test set. **b,** Indirect descriptor influence values in the human reaction test set versus the random reaction test set.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

	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 2 | Matthews correlation coefficient (MCC), accuracy and AUC results for each machine learning algorithm, trained on either the human or triangle test set using all features.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

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

Extended Data Table 3 | Feature selection comparison. **a,** Top 5, **b**, top 10 and **c,** top 20 features using ANOVA F-values for the human and randomized reaction test sets.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

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 (—)

Extended Data Table 4 | Comparison of discrepancies between model predictions and reaction outcomes. 10,000 random reactions were generated for each amine. The first column in Table S29 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), and of those positive predictions the 10 reactions with the lowest model uncertainty were selected and performed in the laboratory. The successful outcomes are indicated as a fraction in parentheses (number successful/number trials). For amines where no positive predictions are made, no tests were performed, indicated by (—).

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Supplementary Information for

Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis

Xiwen Jia, Oscar Huang, Allyson Lynch, Matthew Danielson, Immaculate Lang'at, Alexander Milder, Aaron Ruby, Hao Wang, Sorelle A. Friedler*, Alexander J. Norquist*, Joshua Schrier*

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Figures

Figure S1. t-Distributed Stochastic Neighbor Embedding (t-SNE) plot comparing human and randomly-generated training set data.

Figure S2. t-SNE plot of the expanded test data and predicted outcomes.

Tables

Table S1. Dataset prefix descriptions.

Table S2. Explanation of _raw_ columns.

Table S3. Explanation of _rxn_ columns.

Table S4. Explanation of _out_ columns.

Table S5. Explanation of _feat_ descriptors.

Table S6. Explanation of calculated _feat_fr descriptors.

Table S7. Explanation of _calc_ descriptors. These columns contain descriptors whose values are obtained by performing a calculation on the _raw_, _rxn_ or _feat_ descriptors.

Table S8. Logistic regression machine learning results for the human test set.

Table S9. kNN (N = 2) machine learning results for the human test set.

Table S10. kNN (N = 5) machine learning results for the human test set.

Table S11. SVC machine learning results for the human test set.

Table S12. Decision tree machine learning results for the human test set.

Table S13. Random forest machine learning results for the human test set.

Table S14. Gaussian NB machine learning results for the human test set.

Table S15. Logistic regression machine learning results for the random reaction test set.

Table S16. kNN (N = 2) machine learning results for the random reaction test set.

Table S17. kNN (N = 5) machine learning results for the random reaction test set.

Table S18. SVC machine learning results for the random reaction test set.

Table S19. Decision tree machine learning results for the random reaction test set.

Table S20. Random forest machine learning results for the random reaction test set.

Table S21. Gaussian NB machine learning results for the random reaction test set.

Table S22. Observed outcomes on the discrepant laboratory tests based on predicted model probabilities. All performed reactions were predicted by the model to have an outcome of 4 ("success", defined as single crystal product with average crystallite dimensions exceeding 0.01 mm)

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Contents of the Electronic Supplementary Information File

1 Data capture/

human_borates_tabdelim_20180904.tsv
triangle_borates_tabdelim_20181115.tsv
test_borates_tabdelim_20180904.tsv
discrepant_borates_20190419.tsv

2 Analysis of published results/

01 CCDC Smiles File.html
01 CCDC Smiles File.ipynb
01x CCDC Clean Check.html
01x CCDC Clean Check.ipynb
02 Popularity Files.html
02 Popularity Files.ipynb
03 Outcomes.html
03 Outcomes.ipynb
04 Graphs.html
04 Graphs.ipynb
Data/
Outputs/
x Smiles Images.html
x Smiles Images.ipynb

3 Pricing data/

Amine pricing data.xlsx

4 Generation of random reactions/

4 Generation of random reactions.nb

5 Statistical analysis/

5 Statistical analysis.nb

6 Machine learning/

6 Distance data.xlsx
6 Feature selection summary tables.xlsx
6 Influence data.xlsx
human_k2_accuracy_direct.csv
human_k2_accuracy_indirect.csv
ml_analysis.py
ml_analysis_unformatted_output.txt
ml_preprocessing_notes.txt
preprocessed_human.csv
preprocessed_test.csv
preprocessed_triangle.csv
triangle_k5_accuracy_direct.csv
triangle_k5_accuracy_indirect.csv

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

7 Tests of discrepant reactions/
Possible Reactions/
proposed_experiments_on_most_discrepent_outcomes_20190328.csv

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Explanation of Dataset Columns

Descriptions for dataset columns, prefixes and descriptors are described below. Tables are provided for each information class.

Table S1. Dataset prefix descriptions.

Dataset prefix	Description
raw	Raw data associated with the experiment that was performed. Combined with the other header prefixes, these columns describe the complete set of all data acquired during an experimental run.
feat , _rxn_ , _calc_	These descriptors describe molecular properties, reaction conditions, and derived calculated features (e.g., mole ratios). More elaborated descriptions can be found below for individual headers.
out_	Experimental outcomes. **These are the outputs to predict.**

In general, all values should be assumed to be real numbers, unless indicated otherwise.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Table S2. Explanation of _raw_ columns.

Descriptor name	Description
_raw_Reaction number	Reference to (paper) laboratory notebook. Format: Initials of experimenter, followed by a page number, period delimiter, and reaction number written on that page. (string)
_raw_H3BO3/g	Mass of boric acid reactant in grams
_raw_VOSO4xH2O/g	Mass of vanadyl sulfate hydrate reactant in grams
_raw_Amine/g	Mass of amine reactant in grams. Note that some amines are purchased as ammonium salts; this is the mass of the salt used.
_raw_name	Laboratory abbreviation or name of amine (string)
_raw_SMILES	Simplified molecular-input line-entry system (SMILES) string describing the amine reactant as used. This is used for computation of molecular weights and resulting computations of mole amounts. (string)
_raw_SMILES.nosalt	SMILES string containing only the organic component, with counterions removed. This is used for computing amine cheminformatic properties (string)
_raw_Amine_Mol	Moles of amine reactant
_raw_VOSO4xH2O_Mol	Moles of vanadyl sulfate hydrate, obtained by assuming a nominal molecular weight of 181.011 g/mol. (This is sold without a specification of the precise degree of hydration)
_raw_H3BO3_Mol	Moles of boric acid
_raw_Types of Experiment (Human; Triangle; etc...)	Categories used to separate data. "Human" are historical reactions devised by humans. "Triangle" are random reactions generated according to the triangular probability distribution described in the Methods section. (string)
_raw_Notes	Comments about performance of the reaction. (string)

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Table S3. Explanation of _rxn_ columns.

Descriptor name	Description
_rxn_H2O/g	Mass of water reactant added in grams
_rxn_pH	pH of reaction, as described in the Methods section. (integer)
_rxn_Synthesis_Method	Description of synthesis method. All reactions considered in this study were performed under "Hydrothermal" conditions. (string)
_rxn_Oven Temperature/C	Temperature used for the reaction in Celsius. All new reactions performed in this study were performed at 90 °C, as described in the Methods section. Historical reactions may contain other values. (integer)
_rxn_Aging Time/hrs	Duration of the heating in hours. All new reactions performed in this study were performed for 24 hours, as described in the Methods section. Historical reactions may contain other values. (integer)
_rxn_Vessel Size/mL	Volume of the reaction vessel used in milliliters. All reactions considered this study were performed using 23 mL PTFE sleeves, as described in the Methods sections. (integer)

Table S4. Explanation of _out_ columns.

Descriptor name	Description
_out_Outcome	Crystal size and quality outcome score, as described in the Methods section. **This is the target variable predicted by the machine learning models created in this study.** (integer)
_out_Purity	Denotes whether the solid product produced contains one phase or more than one phase. This data was recorded but was not used in this study. (integer)

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Explanation of _feat_ descriptors

The _raw_, _rxn_, and _out_ columns described above describe the laboratory process and its observable outcomes. In contrast, the _feat_ descriptor columns are computed properties of the amines.

The following descriptors were computed using the cxcalc 5.2.0 JChem calculators. Unless noted otherwise, the suffix corresponds to the name of the descriptor used by cxcalc, as described in the online documentation:
https://docs.chemaxon.com/display/docs/cxcalc+calculator+functions

Table S5. Explanation of _feat_ descriptors.

Descriptor name	Description
_feat_Mass	Molecular mass of _raw_SMILES (possibly a salt)
_feat_Mass_Nosalt	Molecular mass of _raw_SMILES.nosalt (amine only as neutral species without counterions)
_feat_AtomCount_C (atomcount -z 6)	Number of carbon atoms. (integer)
_feat_AtomCount_N (atomcount -z 7)	Number of nitrogen atoms. (integer)
_feat_AvgPol	Average molecular polarizability (at _rxn_pH)
_feat_MolPol	Molecular polarizability (at _rxn_pH)
_feat_Refractivity	Computed refractivity
_feat_isoelectric (isoelectricpoint)	Isoelectric point of the molecule
_feat_apKa1 (pka –a 2)	First acidic pKa value. Subsequent columns are the subsequent entries in the returned list.
_feat_apKa2	Second acidic pKa value…
_feat_bpKa1 (pka –b 4)	First basic pKa value. Subsequent columns are the subsequent entries in the returned list.
_feat_bpKa2	Second basic pKa value…
_feat_bpKa3	Third…
_feat_bpKa4	Fourth…
_feat_AliphaticRingCount	Number of aliphatic rings (integer)
_feat_AromaticRingCount	Number of aromatic rings (integer)
_feat_Aliphatic AtomCount	Number of aliphatic atoms in the molecule (integer)
_feat_AromaticAtomCount	Number of aromatic atoms in the molecule (integer)
_feat_BondCount	Number of bonds in the molecule (integer)
_feat_CarboaliphaticRingCount	Number of aliphatic rings comprised solely of carbon atoms (integer)
_feat_CarboaromaticRingCount	Number of aromatic rings comprised solely of carbon atoms (integer)
_feat_CarboRingCount	Number of rings comprised solely of carbon atoms (integer)
_feat_ChainAtomCount	Number atoms that are part of chain (not part of a ring) (integer)
_feat_ChiralCenterCount	Number of tetrahedral stereogenic centers (integer)

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

_feat_RingAtomCount	Number of atoms that are part of a ring (not part of a chain) (integer)
_feat_SmallestRingSize	Number of members in the smallest ring (integer)
_feat_LargestRingSize	Number of members in the largest ring (integer)
_feat_fsp3	Fraction of sp3 carbons (Fsp3 value)
_feat_HeteroaliphaticRingCount	Number of heteroaliphatic rings (integer)
_feat_HeteroaromaticRingCount	Number of heteroaromatic rings (integer)
_feat_RotatableBondCount	Number of rotatable bonds (integer)
_feat_BalabanIndex	Balaban molecular graph index
_feat_CyclomaticNumber	Cyclomatic number of molecular graph
_feat_HyperWienerIndex	Hyper Wiener Index of molecular graph
_feat_WienerIndex	Wiener Index of molecular graph
_feat_WienerPolarity	Wiener Polarity of molecular graph
_feat_MinimalProjectionArea	Minimal projection area
_feat_MinimalProjectionRadius	Minimal projection radius
_feat_MinimalProjectionRadius	Minimal projection radius
_feat_MaximalProjectionRadius	Maximal projection radius
_feat_LengthPerpendicularToTheMinArea (minimalprojectionsize)	Length perpendicular to the minimal projection area
_feat_LengthPerpendicularToTheMaxArea (maximalprojectionsize)	Length perpendicular to the maximum projection area
_feat_VanderWaalsVolume (volume)	van der Waals volume of the molecule
_feat_VanderWaalsSurfaceArea (vdwsa)	van der Waals surface area of the molecule
_feat_ASA (asa -H _rxn_pH)	Water accessible surface area of the molecule, computed at _rxn_pH
_feat_ASA+ (molecularsurfacearea -t ASA+ -H _rxn_pH)	Water accessible surface area of all atoms with positive partial charge,computed at _rxn_pH
_feat_ASA- (molecularsurfacearea -t ASA- -H _rxn_pH)	Water accessible surface area of all atoms with negative partial charge,computed at _rxn_pH
_feat_ASA_H (molecularsurfacearea -t ASA_H -H _rxn_pH)	Water accessible surface area of all hydrophobic atoms with positive partial charge,computed at _rxn_pH
_feat_ASA_P (molecularsurfacearea -t ASA+P -H _rxn_pH)	Water accessible surface area of all polar atoms with positive partial charge,computed at _rxn_pH
_feat_PolarSurfaceArea (polarsurfacearea -H _rxn_pH)	2D Topological polar surface area, computed at _rxn_pH
_feat_acceptorcount (acceptorcount -H _rxn_pH)	Hydrogen bond acceptor atom count in molecule, computed at _rxn_pH
_feat_Accsitecount (acceptorsitecount -H _rxn_pH)	Hydrogen bond acceptor multiplicity in molecule, computed at _rxn_pH
_feat_donorcount (donorcount -H _rxn_pH)	Hydrogen bond donor atom count in molecule, computed at _rxn_pH
_feat_donsitecount	Hydrogen bond donor multiplicity in molecule, computed at _rxn_pH
_feat_sol (solubility -H _rxn_pH)	Aqueous solubility (logS) computed at _rxn_pH

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

In addition, a series of descriptors describing the number of functional-group fragments _feat_fr_ were computed. Because of the strict definition of amines described in the Methods section, only functional groups related to amines are considered. These are determined from the molecular graph, using the SMARTS patterns described in RDKit 2018.03.4. These are all integer values.

http://www.rdkit.org/Python_Docs/rdkit.Chem.Fragments-module.html

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Table S6. Explanation of calculated _feat_fr descriptors.

Descriptor name	Description
_feat_fr_NH2	Number of primary amines
_feat_fr_NH1	Number of secondary amines
_feat_fr_NH0	Number of tertiary amines
_feat_fr_quatN	Number of quaternary amines
_feat_fr_ArN	Number of N functional groups attached to aromatics
_feat_fr_Ar_NH	Number of aromatic amines
_feat_fr_Imine	Number of imines
_feat_fr_amidine	Number of amidine groups
_feat_fr_dihydropyridine	Number of dihydropyridines
_feat_fr_guanido	Number of guanidine groups
_feat_fr_piperdine	Number of piperidine rings
_feat_fr_piperzine	Number of piperzine rings
_feat_fr_pyridine	Number of pyridine rings

Table S7. Explanation of _calc_ descriptors. These columns contain descriptors whose values are obtained by performing a calculation on the _raw_, _rxn_ or _feat_ descriptors.

Descriptor name	Description
_calc_Amine_H3BO3_MolRatio	Amine:boric acid mole ratio _raw_Amine_Mol / _raw_H3BO3_Mol
_calc_VOSO4xH2O_H3BO3_MolRatio	Vanadyl sulfate hydrate:boric acid mole ratio, (_raw_VOSO4xH2O_Mol) / (_raw_H3BO3_Mol)
_calc_VOSO4xH2O_H3BO3_MassRatio	Vanadyl sulfate hydrate:boric acid mass ratio (_raw_VOSO4xH2O/g) / (_raw_H3BO3/g)
_calc_AmineConc	Amine concentration (as molarity) computed by moles Amine divided by volume of water. (1000*raw_Amine_Mol / _rxn_H2O/g)
_calc_avgNitrogenVDWVol	van der Waals volume per nitrogen atom (_feat_VanderWaalsVolume / _feat_AtomCount_N)
_calc_avgNitrogenVDWSurface	van der Waals surface area per Nitrogen atom (_feat_VanderWaalsSurfaceArea / _feat_AtomCount_N)
_calc_CarbonNitrogenRatio	C:N atom count ratio, (_feat_AtomCount_C / _feat_AtomCount_N)
_calc_avgNitrogenWaterAssessSurface	Water accessible surface area per Nitrogen atom (_feat_ASA / _feat_AtomCount_N)
_calc_pbKaUnderPhCount	Number of pKb values that are less than _rxn_pH (integer)

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Human data

Table S8. Logistic regression machine learning results for the human test set.

	Precision	Recall	f1-score	Support
False	0.56	0.51	0.53	69
True	0.28	0.32	0.30	41
average/ total	0.45	0.44	0.44	110
	TN	35	FN	28
	TP	13	FP	34
Accuracy	0.44			
AUC	0.41			
MCC	-0.17			

Table S9. kNN (N = 2) machine learning results for the human test set.

	Precision	Recall	f1-score	Support
False	0.72	0.83	0.77	69
True	0.61	0.46	0.53	41
average/ total	0.68	0.69	0.68	110
	TN	57	FN	22
	TP	19	FP	12
Accuracy	0.69			
AUC	0.64			
MCC	0.31			

Table S10. kNN (N = 5) machine learning results for the human test set.

	Precision	Recall	f1-score	Support
False	0.63	0.52	0.57	69
True	0.38	0.49	0.43	41
average/ total	0.54	0.51	0.52	110
	TN	36	FN	21
	TP	20	FP	33
Accuracy	0.51			
AUC	0.50			
MCC	0.01			

Table S11. SVC machine learning results for the human test set.

	Precision	Recall	f1-score	Support
False	0.62	0.64	0.63	69
True	0.36	0.34	0.35	41
average/ total	0.52	0.53	0.52	110

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

	TN	44	FN	27
	TP	14	FP	25
Accuracy	0.53			
AUC	0.49			
MCC	-0.02			

Table S12. Decision tree machine learning results for the human test set.

	Precision	Recall	f1-score	Support
False	0.71	0.35	0.47	69
True	0.41	0.76	0.53	41
average/ total	0.59	0.50	0.49	110
	TN	24	FN	10
	TP	31	FP	45
Accuracy	0.50			
AUC	0.55			
MCC	0.11			

Table S13. Random forest machine learning results for the human test set.

	Precision	Recall	f1-score	Support
False	0.62	0.55	0.58	69
True	0.37	0.44	0.40	41
average/ total	0.53	0.51	0.52	110
	TN	38	FN	23
	TP	18	FP	31
Accuracy	0.51			
AUC	0.49			
MCC	-0.01			

Table S14. Gaussian NB machine learning results for the human test set.

	Precision	Recall	f1-score	Support
False	0.68	0.78	0.72	69
True	0.50	0.37	0.42	41
average/ total	0.61	0.63	0.61	110
	TN	54	FN	26
	TP	15	FP	15
Accuracy	0.63			
AUC	0.57			

MCC	0.16

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Triangle data

Table S15. Logistic regression machine learning results for the random reaction test set.

	Precision	Recall	f1-score	Support
False	0.77	0.49	0.60	69
True	0.47	0.76	0.58	41
average/ total	0.66	0.59	0.59	110
	TN	34	FN	10
	TP	31	FP	35
Accuracy	0.59			
AUC	0.62			
MCC	0.25			

Table S16. kNN (N = 2) machine learning results for the random reaction test set.

	Precision	Recall	f1-score	Support
False	0.73	0.86	0.79	69
True	0.66	0.46	0.54	41
average/ total	0.70	0.71	0.70	110
	TN	59	FN	22
	TP	19	FP	10
Accuracy	0.71			
AUC	0.66			
MCC	0.35			

Table S17. kNN (N = 5) machine learning results for the random reaction test set.

	Precision	Recall	f1-score	Support
False	0.90	0.75	0.82	69
True	0.67	0.85	0.75	41
average/ total	0.81	0.79	0.79	110
	TN	52	FN	6
	TP	35	FP	17
Accuracy	0.79			
AUC	0.80			
MCC	0.59			

Table S18. SVC machine learning results for the random reaction test set.

	Precision	Recall	f1-score	Support
False	0.62	0.48	0.54	69
True	0.37	0.51	0.43	41
average/ total	0.53	0.49	0.50	110

	TN	33	FN	20
	TP	21	FP	36
Accuracy	0.49			
AUC	0.50			
MCC	-0.01			

Table S19. Decision tree machine learning results for the random reaction test set.

	Precision	Recall	f1-score	Support
False	0.61	0.45	0.52	69
True	0.36	0.51	0.42	41
average/ total	0.51	0.47	0.48	110
	TN	31	FN	20
	TP	21	FP	38
Accuracy	0.47			
AUC	0.48			
MCC	-0.04			

Table S20. Random forest machine learning results for the random reaction test set.

	Precision	Recall	f1-score	Support
False	0.73	0.59	0.66	69
True	0.48	0.63	0.55	41
average/ total	0.64	0.61	0.62	110
	TN	41	FN	15
	TP	26	FP	28
Accuracy	0.61			
AUC	0.61			
MCC	0.22			

Table S21. Gaussian NB machine learning results for the random reaction test set.

	Precision	Recall	f1-score	Support
False	0.70	0.55	0.62	69
True	0.45	0.61	0.52	41
average/ total	0.61	0.57	0.58	110
	TN	38	FN	16
	TP	25	FP	31
Accuracy	0.57			
AUC	0.58			

MCC	0.16

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

Table S22. Observed outcomes on the discrepant laboratory tests based on predicted model probabilities. All performed reactions were predicted by the model to have an outcome of 4 ("success", defined as single crystal product with average crystallite dimensions exceeding 0.01 mm)

	Predicted Model Probability		
Observed Outcome	**0.6**	**0.8**	**1**
1	2	0	0
2	15	3	2
3	5	17	2
4	8	30	16
Total	30	50	20
Percent 4	27%	60%	80%

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

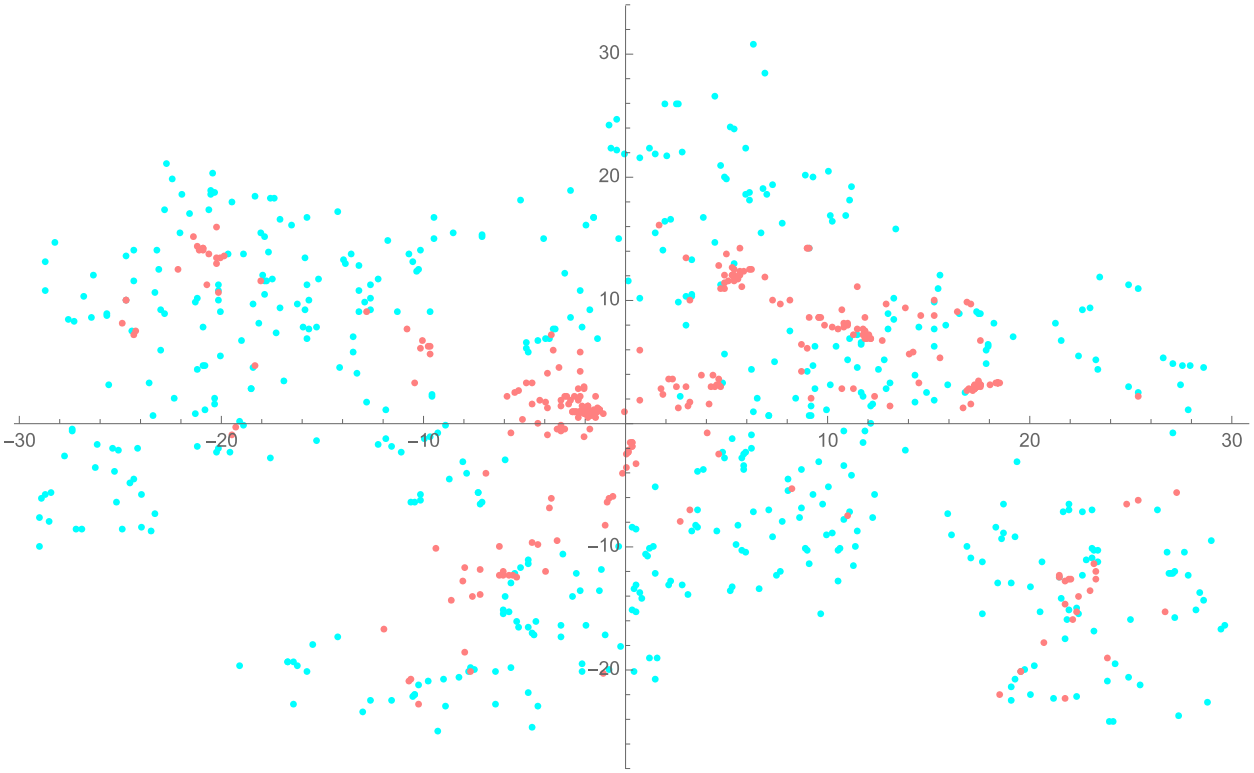


Figure S1. t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of the human (pink) and randomly-generated (cyan) training set data. The t-SNE dimensionality reduction function was constructed to reduce the dataset to two dimensions using the randomly-generated dataset, using the default settings of the Mathematica 11.3.0.0 (`DimensionReduction[#, 2, Method->"TSNE"]&)` function. The resulting dimension reducer function was applied to both datasets to generate the plot above. The axes do not have a direct physical meaning or dimensional units.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

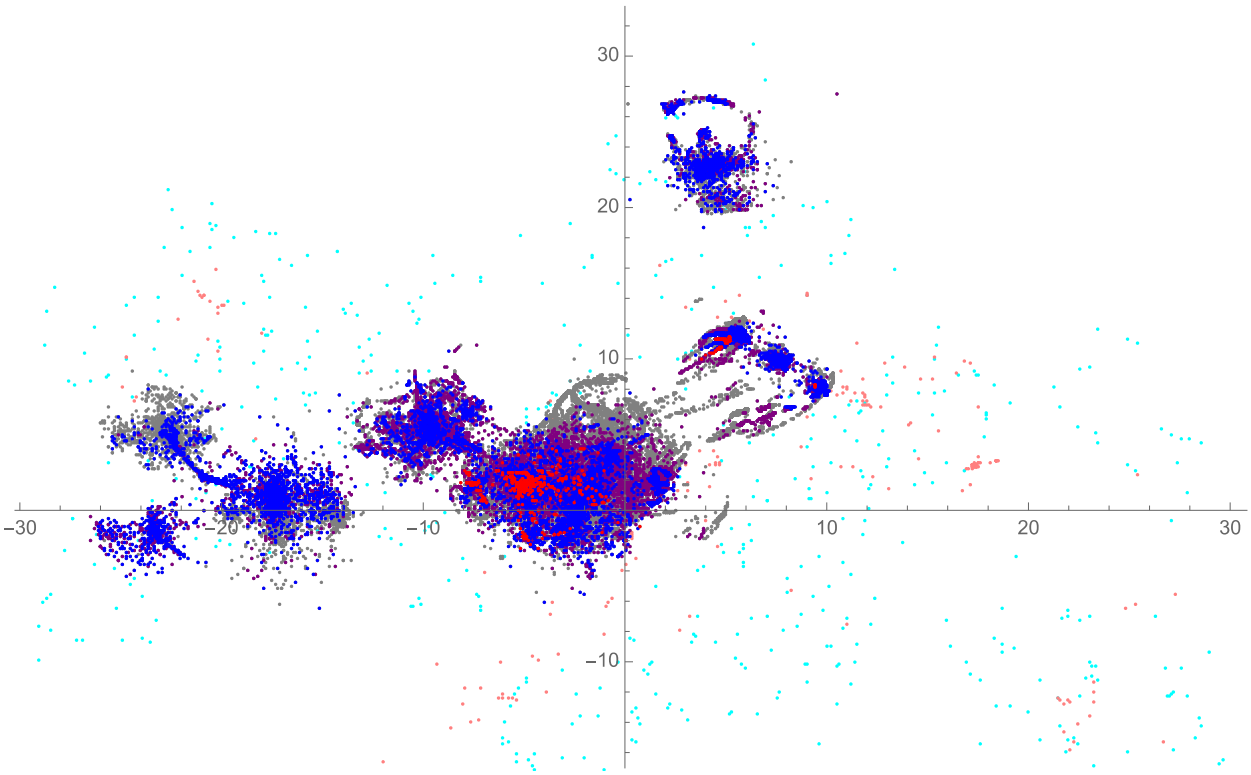


Figure S2. t-SNE plot of the expanded test data and predicted outcomes. The 110,000 randomly generated test reactions (gray) and the subset of these predicted to be successful by the human-data trained model (red) and randomly-generated reaction trained model (blue) are shown in a reduced two-dimensional form, using the same dimensionality reduction function used in the previous figure. The training data for both models (shown by itself in the previous figure) is indicated in pink and cyan, respectively. The general proximity of the expanded test points (gray, red, and blue) to the training data (pink and blue) indicates that neither model is unreasonably being expected to extrapolate beyond its training data. The general proximity of the predicted successes (red and blue) to the failures (gray) in the expanded test set indicates that neither model is simply accepting or rejecting on the basis of outliers.

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5

References

1. Maaten, L.v.d. and Hinton, G., Visualizing data using t-SNE. *J. Machine Learning Research* **9**, 2579—2605 (2008).

Accepted Manuscript for: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis" *Nature* **573**, 251-255 (2019) doi:10.1038/s41586-019-1540-5