nature

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

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, New York, NY, USA. *e-mail: sorelle@cs.haverford.edu; anorquis@haverford.edu; jschrier@fordham.edu

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*

Figures

Tables

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)

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

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

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.

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)

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)

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)
_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

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

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)

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
	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			

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			

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%

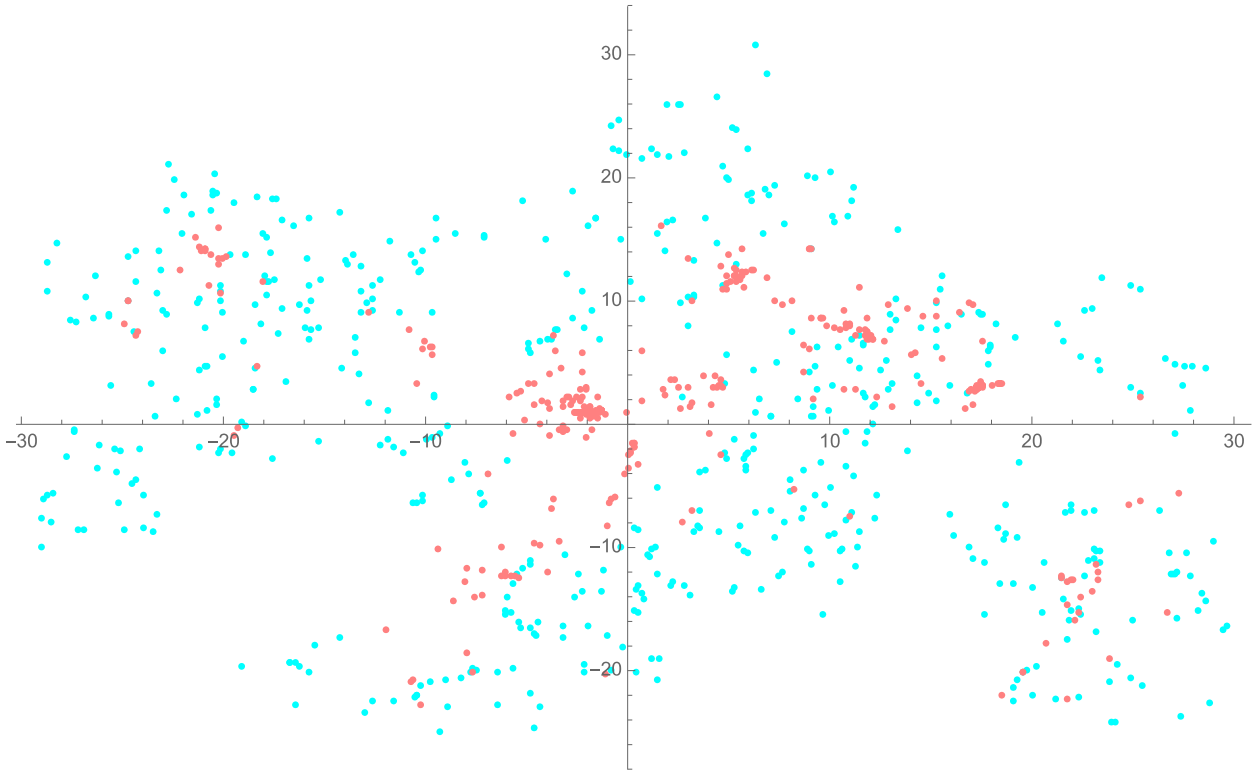

Figure S1. t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of the human (pink) and randomly-generated (cyan) training set data. The t-SNE dimensionality reduction function was constructed to reduce the dataset to two dimensions using the randomly-generated dataset, using the default settings 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.

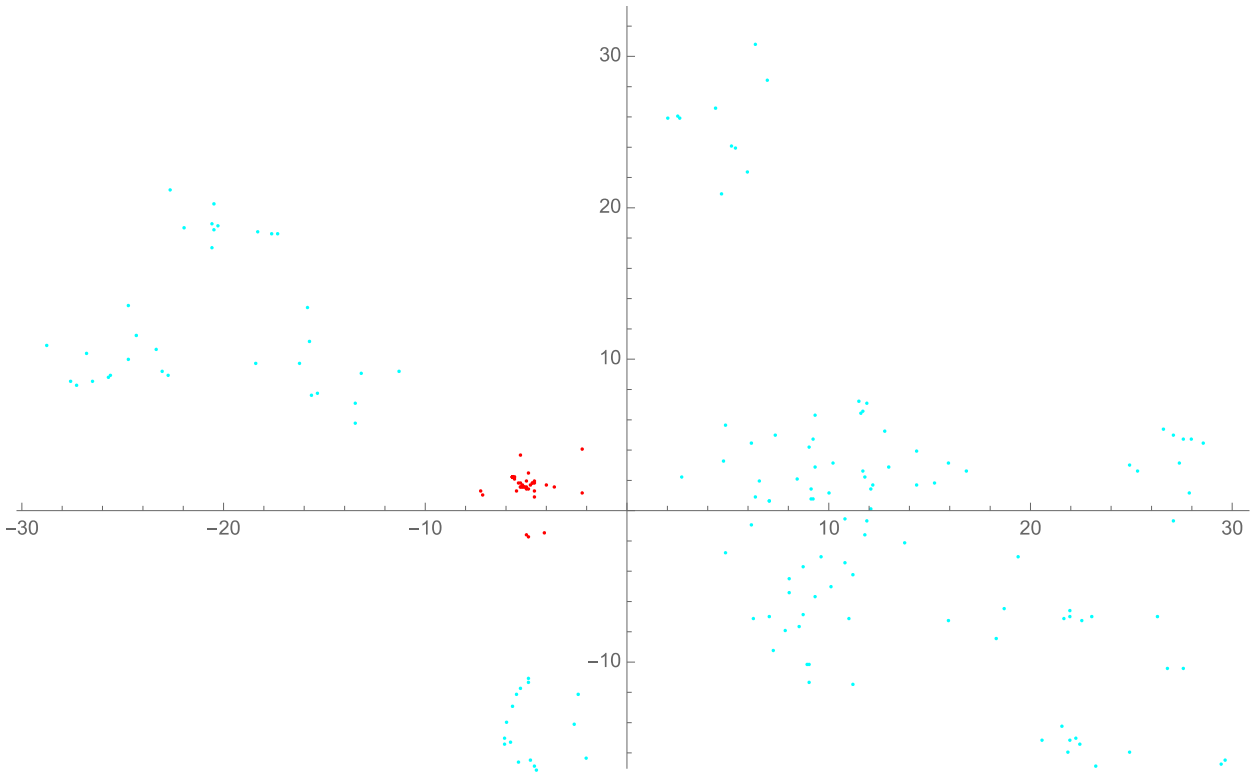

Figure S2. t-SNE plot of the expanded test data and predicted outcomes. The 110,000 randomly generated test reactions (gray) and the subset of these predicted to be successful by the human-data trained model (red) and randomly-generated reaction trained model (blue) are shown in a reduced two-dimensional 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.

References

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