doi:10.1038/nature17439

Table S1. Inorganic atom property descriptors used in this study. (C) indicates values computed by us; (M) indicates values obtained directly from Mathematica 10; (M,C) indicates values that were computed by using the equation shown based on Mathematica 10 inputs.

Property name	Description
numberInorganic	number of inorganic reagents used in this reaction (C)
IonizationMax	maximum ionization energy, *IP*, of all inorganic atom types present in this reaction (M)
IonizationMin	minimum *IP* of all inorganic atoms in this reaction
IonizationMean	arithmetic average (mean) of *IP* for all inorganic atoms in this reaction
IonizationGeom	geometric average of *IP* for all inorganic atoms in this reaction
IonizationMaxWeighted	maximum mole-weighted *IP* (M)
IonizationMinWeighted	minimum mole-weighted *IP*
IonizationMeanWeighted	maximum mole-weighted *IP*
IonizationGeomWeighted	maximum mole-weighted *IP*
EAMax	maximum electron affinity, *EA* (M)
EAMin	
EAMean	
EAGeom	
EAMaxWeighted	
EAMinWeighted	
EAMeanWeighted	
EAGeomWeighted	
PaulingElectronegMax	maximum Pauling Electronegativity, $\chi_{Pauling}$ (M)
PaulingElectronegMin	
PaulingElectronegMean	
PaulingElectronegGeom	
PaulingElectronegMaxWeighted	
PaulingElectronegMinWeighted	
PaulingElectronegMeanWeighted	
PaulingElectronegGeomWeighted	
PearsonElectronegMax	maximum Pearson (Mulliken) electronegativity, $\chi = (IP + EA)/2$, of all inorganic atom types present in this reaction (C,M)
PearsonElectronegMin	
PearsonElectronegMean	
PearsonElectronegGeom	

Table S1 Continued. Inorganic atom property descriptors used in this study. (C) indicates values computed by us; (M) indicates values obtained directly from Mathematica 10; (M,C) indicates values that were computed by using the equation shown based on Mathematica 10 inputs.

Property name	Description
PearsonElectronegMaxWeighted	
PearsonElectronegMinWeighted	
PearsonElectronegMeanWeighted	
PearsonElectronegGeomWeighted	
hardnessMax	maximum hardness, $\eta=(IP-EA)/2$ (C, M)
hardnessMin	
hardnessMean	
hardnessGeom	
hardnessMaxWeighted	
hardnessMinWeighted	
hardnessMeanWeighted	maximum hardness, $\eta=(IP-EA)/2$ (C, M)
hardnessGeomWeighted	
AtomicRadiusMax	
AtomicRadiusMin	
AtomicRadiusMean	
AtomicRadiusGeom	
AtomicRadiusMaxWeighted	
AtomicRadiusMinWeighted	
AtomicRadiusMeanWeighted	
AtomicRadiusGeomWeighted	
AtomicRadiusGeom	

Table S2. Stoichiometry ratio properties.

Property name	Description
inorg-water-moleratio	total moles of all inorganic reagents divided by moles of water
inorg-org-moleratio	total moles of all inorganic reagents divided by total moles of all organic reactants
org-water-moleratio	total moles of all organic reagents divided by total moles of all water
orgacc-waterdonratio	Ratio of the number of moles of hydrogen bond acceptors on the organic to number of mole of hydrogen bond donors on water at the reaction pH
orgdon-wateraccratio	Ratio of the number of moles of hydrogen bond donors on the organic to number of mole of hydrogen bond acceptors on water at the reaction pH
notwater-water-moleratio	total moles of all reagents that are not water (i.e., inorganic +organic + oxalate-like) divided by moles of water

Table S3. Reaction condition properties, as entered from lab notebooks.

Property name	Description
Temp_max	Maximum temperature for the reaction in Celcius
time	Reaction time in hours
slowcool	(logical) was a slow cool performed?
pH	initial reaction pH
leak	(logical) did a leak occur?

Table S4. Organic molecule properties, as computed using ChemAxon.

Property name	Description
orgavgpolMax	maximum value of average molecular polarizability (avgpol) for all organic components evaluated using the nominal molecular structure
orgavgpolMin	minimum value of average molecular polarizability (avgpol) for all organic components evaluated using the nominal molecular structure
orgavgpolMean	arithmetic mean of average molecular polarizability (avgpol) for all organic components evaluated using the nominal molecular structure
orgavgpolGeom	geometric mean of average molecular polarizability (avgpol) for all organic components evaluated using nominal molecular structure
orgavgpol_pHdependentMax	maximum value of average molecular polarizability (avgpol) for all organic components evaluated using the major microspecies at the reaction pH
orgavgpol_pHdependentMin	
orgavgpol_pHdependentMean	
orgavgpol_pHdependentGeom	
orgrefractivityMax	maximum value of refractivity (refractivity) for all organic components
orgrefractivityMin	
orgrefractivityMean	
orgrefractivityGeom	
orgmaximalprojectionAreaMax	maximum value of the maximal projection area (maximalprojectionarea) for all organic components, using default settings for conformer generation and optimization
orgmaximalprojectionAreaMin	
orgmaximalprojectionAreaMean	
orgmaximalprojectionAreaGeom	

Table S4 Continued. Organic molecule properties, as computed using ChemAxon.

Property name	Description
orgmaximalprojectionradiusMax	maximum value of the maximal projection radius (maximalprojectionradius) for all organic components, using default settings for conformer generation and optimization
orgmaximalprojectionradiusMin	
orgmaximalprojectionradiusMean	
orgmaximalprojectionradiusGeom	
orgmaximalprojectionsizeMax	maximum value of the size of the molecule perpendicular to the maximal projection area surface (maximalprojectionsize) for all organic components, using default settings for conformer generation and optimization
orgmaximalprojectionsizeMin	
orgmaximalprojectionsizeMean	
orgmaximalprojectionsizeGeom	
orgminimalprojectionareaMax	maximum value of the minimal projection area (minimalprojectionarea) for all organic components, using default settings for conformer generation and optimization
orgminimalprojectionareaMin	
orgminimalprojectionareaMean	
orgminimalprojectionareaGeom	
orgminimalprojectionradiusMax	maximum value of the minimal projection radius (minimalprojectionradius) for all organic components, using default settings for conformer generation and optimization
orgminimalprojectionradiusMin	
orgminimalprojectionradiusMean	
orgminimalprojectionradiusGeom	
orgminimalprojectionsizeMax	maximum value of the size of the molecule perpendicular to the minimal projection area surface (minimalprojectionsize) for all organic components, using default settings for conformer generation and optimization
orgminimalprojectionsizeMin	

Table S4. Organic molecule properties, as computed using ChemAxon.

Property name	Description
orgminimalprojectionsizeMean	
orgminimalprojectionsizeGeom	
orgmolpolMax	maximum value of the molecular polarizability (molpol -H) for all organic components, using the major microspecies at the reaction pH
orgmolpolMin	
orgmolpolMean	
orgmolpolGeom	
orgvanderwaalsMax	maximum value of the 3D molecular surface area, computed using van der Waals radii of the atoms (molecularsurfacearea -t vanderwaals -H) for all organic components, using the major microspecies at the reaction pH
orgvanderwaalsMin	
orgvanderwaalsMean	
orgvanderwaalsGeom	
orgASAMax	maximum value of the water accessible surfacearea (molecularsurfacearea -t ASA –H) for all organic components, using the major microspecies at the reaction pH
orgASAMin	
orgASAMean	
orgASAGeom	
orgASA+Max	maximum value of the water accessible surface area of all atoms with positive partial charge (molecularsurfacearea -t ASA+ -H) for all organic components, using the major microspecies at the reaction pH
orgASA+Min	
orgASA+Mean	
orgASA+Geom	
orgASA-Max	maximum value of the water accessible surface area of all atoms with negative partial charge (molecularsurfacearea -t ASA- -H) for all organic components, using the major microspecies at the reaction pH
orgASA-Min	
orgASA-Mean	
orgASA-Geom	

Table S4. Organic molecule properties, as computed using ChemAxon.

Property name	Description
orgASA_HMax	maximum value of the water accessible surface area of all hydrophobic atoms (molecularsurfacearea -t ASA_H -H) for all organic components, using the major microspecies at the reaction pH
orgASA_HMin	
orgASA_HMean	
orgASA_HGeom	
orgASA_PMax	maximum value of the water accessible surface area of all polar atoms (molecularsurfacearea -t ASA_P -H) for all organic components, using the major microspecies at the reaction pH
orgASA_PMin	
orgASA_PMean	
orgASA_PGeom	
orgpolarsurfaceareaMax	maximum value of the 2D topological polar surface area (polarsurfacearea -H) for all organic components, using the major microspecies at the reaction pH
orgpolarsurfaceareaMin	
orgpolarsurfaceareaMean	
orgpolarsurfaceareaGeom	
orghbdamsaccMax	maximum value of the number of hydrogen bond acceptors (hbda -t acc) for all organic components, using the major microspecies at the reaction pH
orghbdamsaccMin	
orghbdamsaccMean	
orghbdamsaccGeom	
orghbdamsdonMax	maximum value of the number of hydrogen bond donors (hbda -t don) for all organic components, using the major microspecies at the reaction pH
orghbdamsdonMin	
orghbdamsdonMean	
orghbdamsdonGeom	

Table S5. Model evaluation results.

Technique (all from Weka library)	2-class accuracy (%)[a]
Decision tree (J48)	67.5
Random forest (size 100 and 1000)	69.8, 70.5
Logistic regression	69.2
k-Nearest neighbors (K = 1, 2, and 3)	69.1, 66.9, 68.4
SMO SVM	74.1

Reported test set accuracy is the average over 15 training/test splits where each split is created so that (inorganic, inorganic, organic) triplets are in either the training set or the test set. The specific model names and parameters can be found in the supplementary script `generate_weka_models.sh`.

Table S6. SVM learning curve.

Training set size	Accuracy (%)
276	70.3
552	70.6
828	71.2
1104	72.0
1380	73.0
1656	73.0
1932	73.3
2208	73.7
2484	73.8
2756.1	74.1

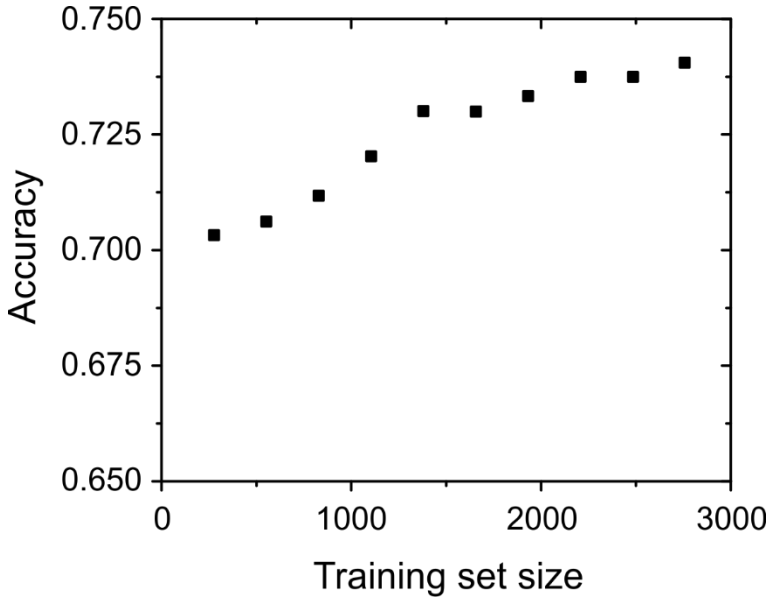

Figure S1. SVM learning curve plot.

Feature Selection Details.

Feature selection was performed using the Weka library's CfsSubsetEval with two different selection methods: best-first and greedy stepwise. Then an SVM was trained and tested using just the selected features. This resulted in an accuracy of 70.7% (best-first) and 71.6% (greedy stepwise). More details are below.

The two methods of feature selection returned different sets of descriptors. While individual descriptor identities differ between methods, the collected sets of descriptors are essentially the same, given correlations between descriptors. Both methods include descriptors for inorganic composition (hardnessMaxWeighted, AtomicRadiusMeanWeighted, PaulingElectronegMean, hardnessMinWeighted, EAMinWeighted), organic amine size (orgvanderwaalsMin, orgASA+Min, orgASA_HGeomAvg, orgvanderwaalsArithAvg), amine hydrogen-bonding properties (orghbdamsdonGeomAvg, orghbdamsaccMax). Feature selection using CfsSubsetEval greedy stepwise selection also returns reaction condition descriptors (time, leak, Temp_max).

CfsSubsetEval best-first selection.
Feature selection results via the Weka library's CfsSubsetEval with best-first selection. An SMO SVM model trained on the resulting 6 features yields an accuracy of 70.7%. The more specific parameter settings and results are below:

Search method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 38
Merit of best subset found: 0.091

Attribute Subset Evaluator (supervised, Class (nominal): outcome):
CFS Subset Evaluator
Treating missing values as a separate value
Including locally predictive attributes

Selected attributes:
orgvanderwaalsMin
orgASA+Min
orghbdamsdonGeomAvg
PaulingElectronegMean
hardnessMaxWeighted
AtomicRadiusMeanWeighted

CfsSubsetEval greedy stepwise selection.

The feature selection above uses Weka's CfsSubsetEval with their best-first selection method that searches the space of possible subsets using beam search. However, best-first selection can not generate a ranked list of features, and does not grant direct control over the number of descriptors included in the subset of selected features. In order to consider the effect of including more features, we used Weka's CfsSubsetEval with the GreedyStepwise selection method that implements hill climbing. The results and more specific parameter settings are below.

Search Method:
Greedy Stepwise (forwards).
Start set: no attributes

Ranking is the order that attributes were added, starting with no attributes. The merit scores in the left column are the goodness of the subset after adding the corresponding attribute in the right column to the subset.

Attribute Subset Evaluator (supervised, Class (nominal): outcome):
CFS Subset Evaluator
Treating missing values as a separate value
Including locally predictive attributes

Resulting features (top 6):
time
hardnessMinWeighted
orgASA_HGeomAvg
leak
inorg-water-moleratio
orgvanderwaalsArithAvg

Resulting features (top 9):
time
hardnessMinWeighted
orgASA_HGeomAvg
leak
inorg-water-moleratio
orgvanderwaalsArithAvg
orghbdamsaccMax
Temp_max
EAMinWeighted

The resulting top 6 features have an accuracy of 71.6% when an SVM is trained using only these 6 features and evaluated as an average of 15 splits constructed as described previously. The top 9 features create an SVM with an accuracy of 71.5%.

Table S7. Amine names and similarity indices for the 34 'untested' amines.

Amine	Similarity index
2-(Aminomethyl)-1-ethylpyrrolidine	1
N-Methylethylenediamine	0.846
2,3-Dimethylpiperazine	0.831
N-Propylethylenediamine	0.806
1-Butylpiperazine	0.770
N,N'-Dimethyl-1,3-propanediamine	0.750
N,N'-Dimethyl-1,6-hexanediamine	0.750
N-Methyl-1,3-diaminopropane	0.750
1-(2-Aminoethyl)pyrrolidine	0.721
2-Amino-5-diethylaminopentane	0.710
3-(Dibutylamino)propylamine	0.702
N-Propyl-1,3-propanediamine	0.697
N,N-Dimethyl-3-pyrrolidinamine	0.688
N-Isopropyl-1,3-propanediamine	0.676
2-(Aminomethyl)piperidine	0.664
1-(2-Aminoethyl)piperidine	0.663
1,3-Cyclohexanebis(methylamine), mixture of isomers	0.652
1-(3-aminopropyl)pyrrolidine	0.650
N,N,N',N'-Tetramethyl-1,4-butanediamine	0.643
4-(1-Pyrrolidinyl)piperidine	0.636
N,N,N',N'-Tetramethyl-1,3-butanediamine	0.636
1-Octylpiperazine	0.635
4-Aminopiperidine	0.621
4,4'-Methylenebis(2-methylcyclohexylamine), mixture of isomers	0.620
4-(Aminomethyl)piperidine	0.594
2-(2-Aminoethyl)-1-methylpyrrolidine	0.593
2-aminopiperidine	0.587
N,N,N'-Trimethyl-1,3-propanediamine	0.556
N,N,N',N'-Tetramethyl-1,3-propanediamine	0.556
5-Amino-1,3,3-trimethylcyclohexanemethylamine, mixture of cis and trans	0.554
3-(aminomethyl)piperidine	0.553
3-(Dimethylamino)-1-propylamine	0.545
4-Amino-2,2,6,6-tetramethylpiperidine	0.541
N,N,2,2-Tetramethyl-1,3-propanediamine	0.478

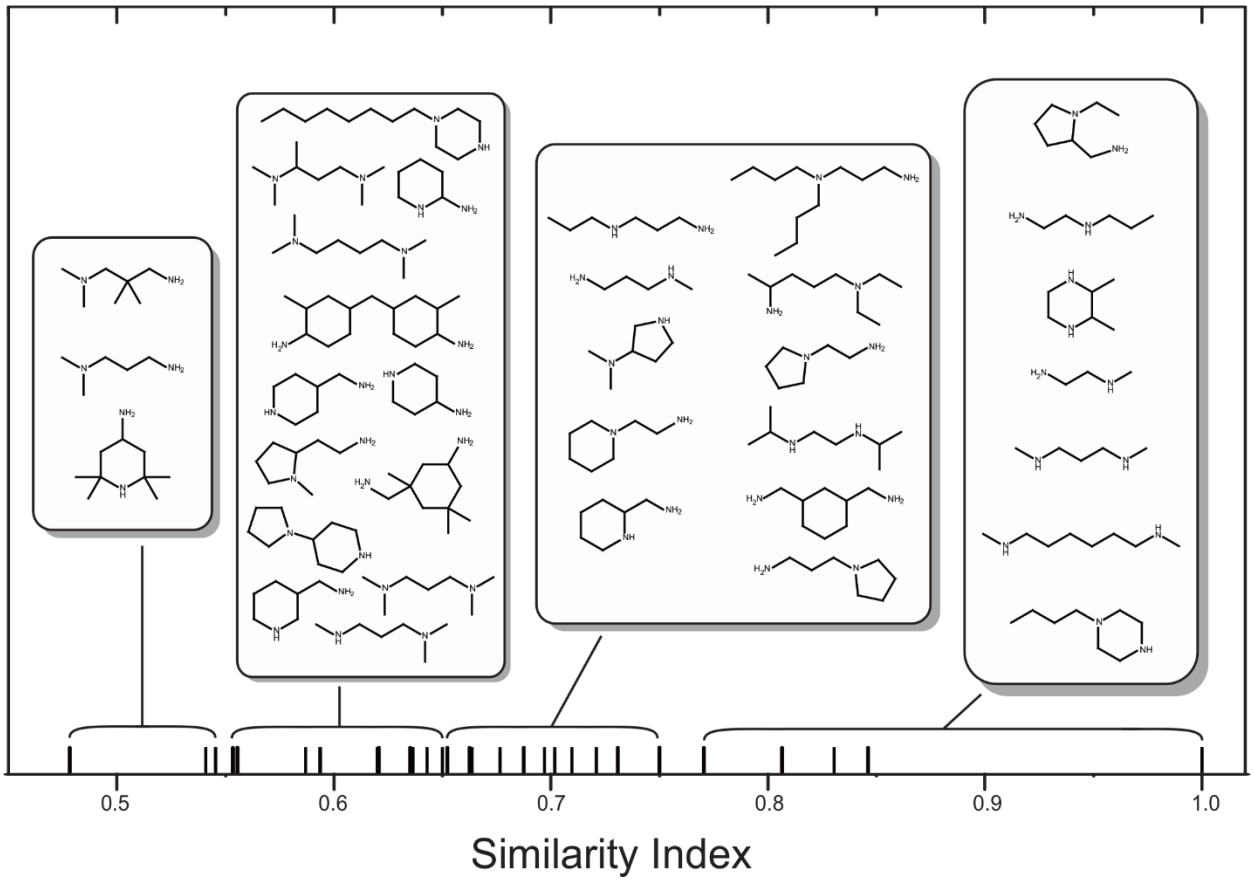

Figure S2. Amine similarity index distribution for the 34 ‘untested’ amines.

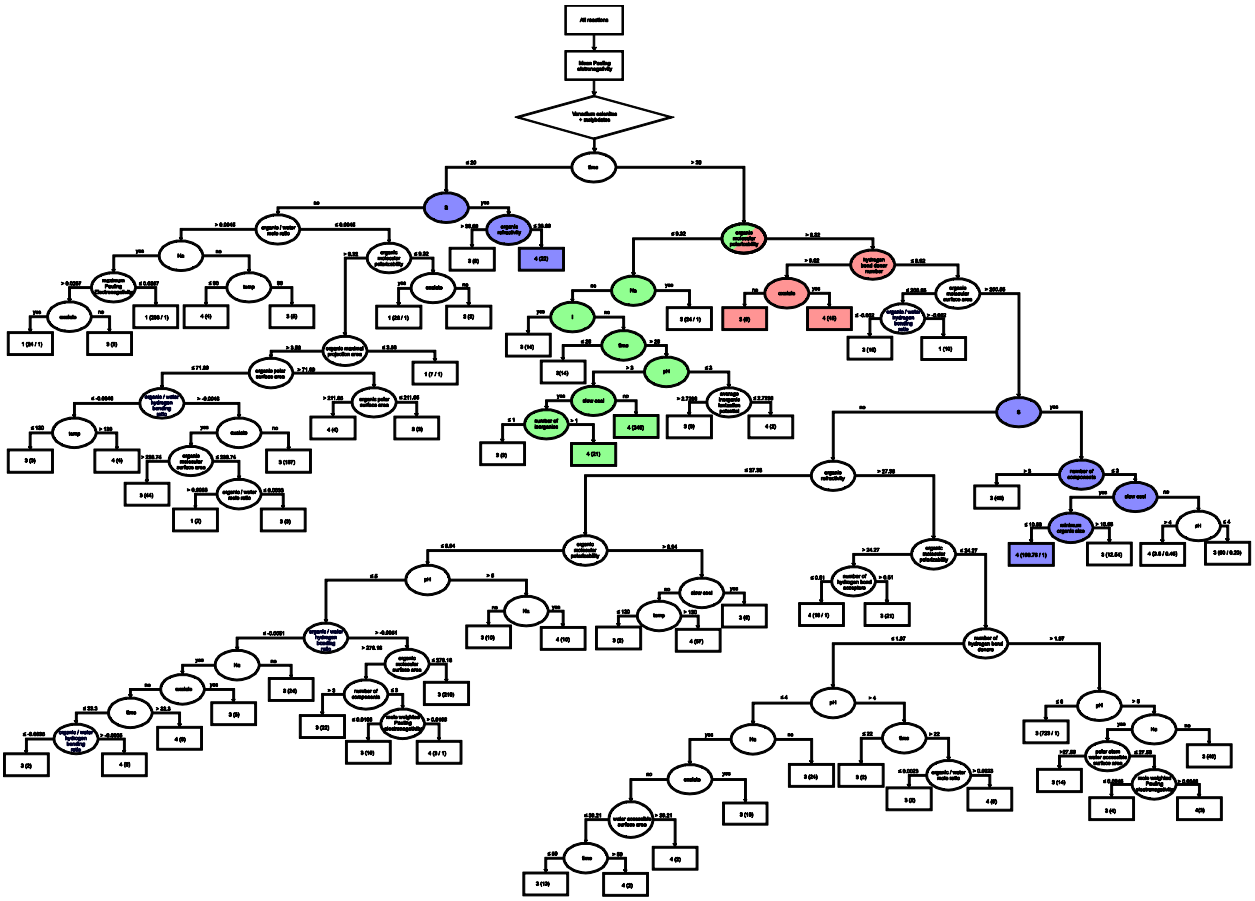

Figure S3. SVM derived flow chart for historical reactions. The components that led to specific chemical hypotheses are indicated by color, with blue, red and green representing the first, second and third hypothese, respectively.

Experimental Details.

Materials. NH_4VO_3 (99 %), $NaVO_3$ (anhydrous, 99.9 %), SeO_2 (99.4 %), 1,3-Diaminopropane (≥ 99 %), $Na_2C_2O_4$ (99.5 %) and triethylenetetramine (≥ 99 %) were purchased from Sigma-Aldrich. All reagents were used as received. Deionized water was used in these syntheses.

Syntheses. All reactions were conducted in either 23 mL poly(fluoro-ethylene-propylene) lined pressure vessels or 15 mL polypropylene bottles. Initial reaction pHs were controlled by the addition of 4 M HCl and 4 M NaOH. The reactions were heated to a set temperature and allowed to soak. The reactions were then cooled to room temperature at a rate of 6 ºC h^{-1} to promote the growth of large single crystals. Autoclaves and bottles were opened in air, and products were recovered through vacuum filtration.

$[C_3H_{12}N_2][V_3O_5(SeO_3)_3]\cdot H_2O$ (1). 1 was synthesized as sheet-like black-green crystals through the reaction of 0.1109 g (1.00×10^{-3} mol) NH_4VO_3, 1.1102 g (1.00×10^{-2} mol) SeO_2, 0.4443 g (6.00×10^{-3} mol) 1,3-diaminopropane, and 7.3860 g (4.10×10^{-1} mol) H_2O. The reaction was heated at 110 °C for 30 h in 23 mL poly(fluoro-ethylene-propylene) lined pressure vessels; initial pH was set to 3.

$[C_6H_{22}N_4][VO(C_2O_4)(SeO_3)]_2\cdot 2H_2O$ (2). 2 was synthesized as blue plates through the reaction of 0.1283 g (1.11×10^{-3} mol) $NaVO_3$, 0.9292 g (8.40×10^{-3} mol) SeO_2, 0.1709 g (1.90×10^{-3} mol) trietheylenetetramine, 0.1395 g (1.04×10^{-3} mol) $Na_2C_2O_4$ and 6 g (3.32×10^{-1} mol) H_2O. The reaction was heated at 90 °C for 24 h in a 15 mL polypropylene bottle. The initial reaction pH was set to 4.

Single Crystal X-ray Diffraction. Data were collected using a Bruker AXS Smart Apex or Apex II CCD diffractometer with Mo-Kα radiation (λ = 0.71073 Å). A single crystal was mounted on a Mitegen micromesh mount using a trace of mineral oil and cooled *in-situ* to 100(2) K for data collection. Frames were collected, reflections were indexed, processed and the files scaled and corrected for absorption using APEX2.[1] The heavy atom positions were determined using SIR92.[2] All other non-hydrogen sites were located from Fourier difference maps. All non-hydrogen sites were refined using anisotropic thermal parameters using full matrix least squares procedures on F_o^2 with $I > 3\sigma(I)$. Hydrogen atoms were placed in geometrically idealized positions. All calculations were performed using Crystals *v.* 14.23*c*.[3] Relevant crystallographic data are listed in Table S8.

Table S8. Crystallographic data for compounds 1 and 2.

Compound	$[C_3H_{12}N_2][V_3O_5(SeO_3)_3]\cdot H_2O$ (**1**)	$[C_6H_{22}N_4][VO(C_2O_4)(SeO_3)]_2\cdot 2H_2O$ (**2**)
Formula	$C_3H_{14}N_2O_{15}Se_3V_3$	$C_{10}H_{26}N_4O_{18}Se_2V_2$
fw	707.85	750.13
Space group	$P2_1/m$ (no. 12)	$P2_1/c$ (no. 14)
a / Å	6.3205(8)	12.6854(15)
b / Å	19.766(3)	6.3408(7)
c / Å	13.2127(17)	14.1903(17)
α / º	90	90
β / º	97.314(2)	95.9912(15)
γ / º	90	90
V / Å^3	1637.3(4)	1135.2(2)
Z	4	2
ρ_{calc}	2.871	2.194
λ / Å	0.71073	0.71073
T / K	100(2)	100(2)
μ / mm^{-1}	8.429	4.125
R_1 [a]	0.0337	0.0184
wR_2 [b]	0.0803	0.0463

[a] $R_1 = \Sigma \| F_o | - | F_c \| / \Sigma | F_o |$. [b] $wR_2 = [\Sigma w(F_o^2 - F_c^2)^2 / [\Sigma w(F_o^2)^2]^{1/2}$.

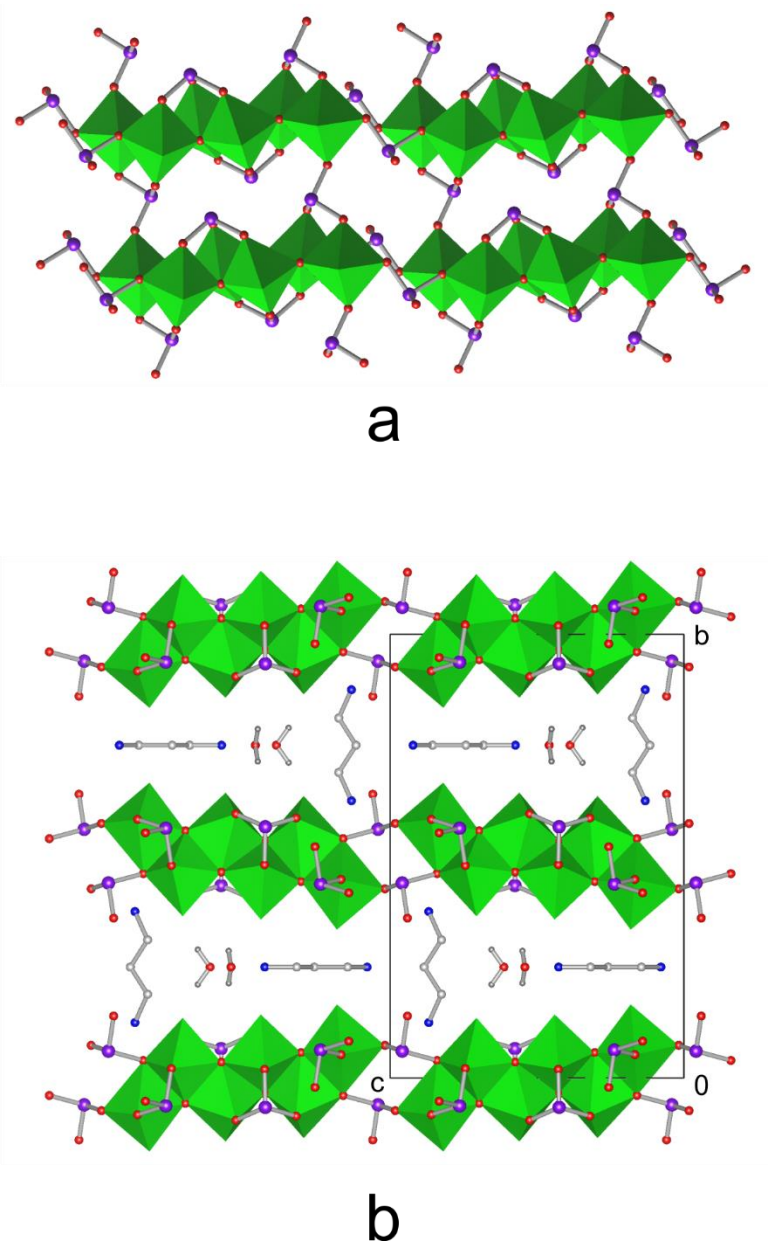

Figure S4. (a) $[V_3O_5(SeO_3)_3]_n^{2n-}$ layer topology and (b) three-dimensional packing in $[C_3H_{12}N_2][V_3O_5(SeO_3)_3]\cdot H_2O$ (1). Green octahedra represent $[VO_6]$ while purple, red, blue, white and gray spheres represent selenium, oxygen, nitrogen, carbon and hydrogen, respectively. Organic cation hydrogen atoms have been omitted for clarity.

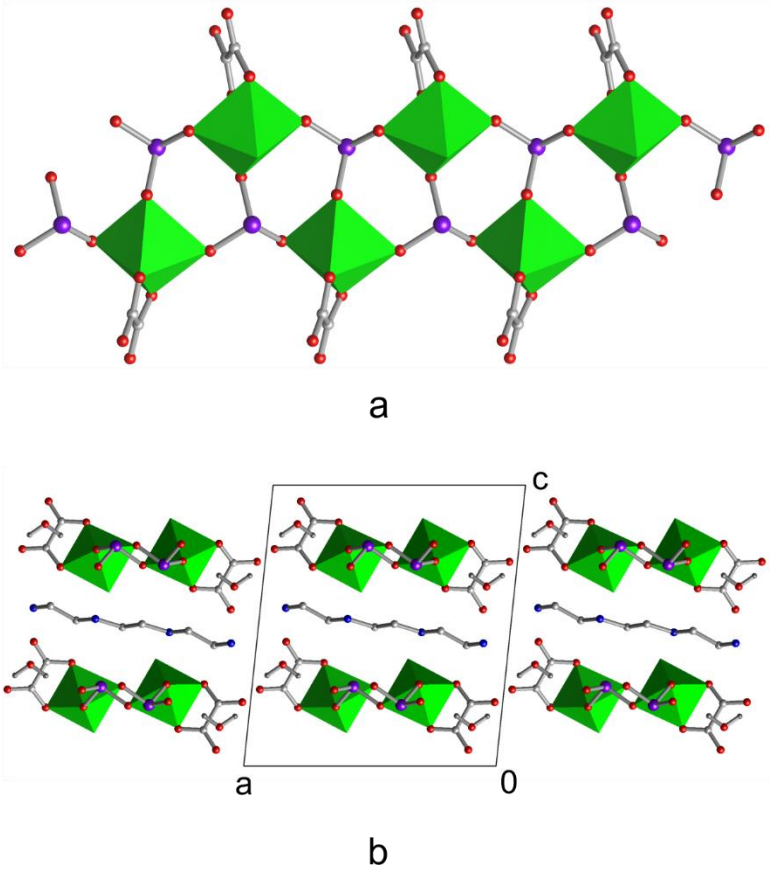

Figure S5. (a) $[VO(C_2O_4)(SeO_3)]_n^{2n-}$ chain connectivity and (b) three-dimensional packing in $[C_6H_{22}N_4][VO(C_2O_4)(SeO_3)]_2 \cdot 2H_2O$ (2). Green octahedra represent $[VO_6]$ while purple, red, blue, white and gray spheres represent selenium, oxygen, nitrogen, carbon and hydrogen, respectively. Organic cation hydrogen atoms have been omitted for clarity.

Table S9. Bond valence sums for $[C_3H_{12}N_2][V_3O_5(SeO_3)_3]\cdot H_2O$ (**1**).

S_i	V1	V2	V3	Se1	Se2	Se3	Hydrogen bonds	ΣS_i	$V - \Sigma S_i$
O1	1.66							1.66	-0.34
O2	0.69	0.20	1.12					2.01	0.01
O3	0.56			1.33				1.89	-0.11
O4	0.60				1.26			1.86	-0.14
O5	0.57			1.42				1.99	-0.01
O6	0.14	0.65			1.18			1.97	-0.03
O7		1.73						1.73	-0.27
O8		0.61				1.35		1.97	-0.03
O9		1.16	0.26					2.08	0.08
			0.66						
O10		0.66		1.30				1.96	-0.04
O11			1.72					1.72	-0.28
O12			0.62			1.34		1.97	-0.03
O13			0.64			1.27		1.91	-0.09
O14					1.48			1.48	-0.52
O15							1.6	1.6	-0.4
O16							1.6	1.6	-0.4
ΣS_i	4.22	5.02	5.03	4.05	3.92	3.97			

Table S10. Bond valence sums for $[C_6H_{22}N_4][VO(C_2O_4)(SeO_3)]_2 \cdot 2H_2O$ (**2**).

S_i	V1	Se1	C1	C2	Hydrogen bonds	ΣS_i	$V - \Sigma S_i$
O1	1.59					1.59	-0.41
O2	0.54	1.37				1.91	-0.09
O3	0.51		1.32			1.82	-0.18
O4	0.50	1.35				1.85	-0.15
O5	0.55	1.34				1.89	-0.11
O6	0.34			1.34		1.68	-0.32
O7			1.55			1.55	-0.45
O8				1.52		1.52	-0.48
O9					1.6	1.6	-0.4
?S_i	4.02	4.06					

1 Apex2 v2009.7-0. (Bruker AXS Inc., Madison (WI), USA, 2009).

2 Altomare, A., Cascarano, G., Giacovazzo, C. & Guagliardi, A. Completion and refinement of crystal structures with SIR92. *J. Appl. Crystallogr.* **26**, 343-350 (1993).

3 Betteridge, P. W., Carruthers, J. R., Cooper, R. I., Prout, K. & Watkin, D. J. CRYSTALS version 12: software for guided crystal structure analysis. *J. Appl. Crystallogr.* **36**, 1487 (2003).