Supporting information for

A thermodynamic atlas of carbon redox chemical space

Adrian Jinich[a,b], Benjamin Sanchez-Lengeling[a], Haniu Ren[a], Joshua E. Goldford[c], Elad Noor[d], Jacob N. Sanders[e], Daniel Segrè[c,f], Alán Aspuru-Guzik[g, h, i*]

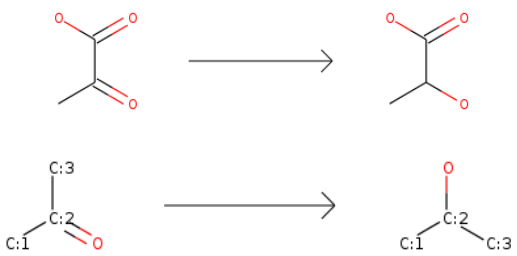
[a] Department of Chemistry and Chemical Biology, Harvard University, Cambridge MA, 02138
[b] Division of Infectious Diseases, Weill Department of Medicine, Weill-Cornell Medical College, NY, NY
[c] Bioinformatics Program and Biological Design Center, Boston University, Boston, MA 02215
[d] Institute of Molecular Systems Biology, ETH Zurich, Auguste-Piccard-Hof 1, 8093 Zürich, Switzerland
[e] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, 90095
[f] Department of Biology, Department of Biomedical Engineering, Department of Physics, Boston University, Boston, MA 02215
[g] Department of Chemistry and Department of Computer Science, University of Toronto, ON, Canada
[h] Vector Institute, Toronto, ON, Canada
[i] Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada
*** Corresponding Authors:** Dr. Adrian Jinich, Division of Infectious Diseases, Weill Department of Medicine, Weill-Cornell Medical College, NY, NY. Phone: 617-285-3920. Email: adj2010@med.cornell.edu. Prof. Alán Aspuru-Guzik, Department of Chemistry, University of Toronto, 80 St. George Street, Toronto, Ontario M5S 3H6. Phone: 416-978-3564. Email: alan@aspuru.com

Supplemental Methods:

Generation of full redox networks using RDKit

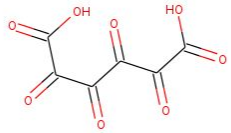
To generate the reactions, we used the RDKit cheminformatics software to design SMILES (simplified molecular-input line-entry system) (1) reaction templates (reaction strings), which, when applied to a compound, will reduce it according to the functional groups detected. Reaction strings were created for the three redox categories of interest: reduction of carboxylic acids to aldehydes, reduction of aldehydes/ketones to alcohols, and reduction of alcohols to hydrocarbon. These templates are designed to be generic enough that they can be applied to any compound with the target functional group, but also with enough specificity to only generate a reaction belonging to the correct redox category.

As an illustrative example, we consider the reduction of pyruvate. Pyruvate contains two types of functional groups that can be reduced: a carboxylic acid and a ketone. The carboxylic acid can be reduced to an aldehyde, or the ketone can be reduced to a hydroxyl. To accomplish this we applied the appropriate SMILES reaction strings. The SMILES reaction string used for the ketone reduction of pyruvate to lactate is shown below. This reaction string can be visualized as a generic reduction of a ketone to a hydroxyl. The *ReactionFromSmarts* function in RDkit is used to generate a reaction object from the reaction string.



The molecular transformation encoded by the SMILES reaction string is shown above. The substrate and compound of each reaction are represented as strings and concatenated into a reaction string as follows: [#6:1][CX3:2](=O)[#6:3]>>[#6:1][CX4H1:2]([#6:3])[OX2H1]

This reaction object can be applied to any compound with a ketone functional group in order to reduce it to a hydroxyl. For cases in which the compound contains multiple target functional groups (e.g. dicarbonyls), every possible product will be generated. To generate the full network or redox reactions, these reaction strings were run iteratively, starting with the fully oxidized unbranched, carbon chain compounds of length 2 to 6 carbons. For example the seed compound for the redox chemical space of 6-carbon straight-chain molecules (i.e. the fully oxidized 6-carbon linear chain seed compound) is shown below:



Once a fully oxidized seed compound had been reduced one step at every possible carbon atom in the initial iteration, the function was repeatedly applied on the resulting products. This continues iteratively until the fully reduced n-carbon hydrocarbon chain is obtained. Any duplicate reactions and products generated from this approach were eliminated during each iteration. Thus, a network of all possible redox reactions originating from the fully oxidized seed compound can be generated.

SMILES reaction strings

Reaction category	**Reaction strings**
Carboxylic acids to aldehydes	[CX3:1](=O)[OX2H1]>>[CX3H1:1](=O)
Aldehydes to alcohols	[CX3H1:2](=O)[#6:1] >>[#6:1][CX4H2:2][OX2H1]
Ketones to alcohols	[#6:1][CX3:2](=O)[#6:3]>>[#6:1][CX4H1:2]([#6:3])[OX2H1]
Alcohols to hydrocarbons (middle)	[CX4H2:2][OX2H1]>>[CX4H3:2]
Alcohols to hydrocarbons (edge)	[#6:1][#6H1:2]([#6:3])[OX2H1]>>[#6:1][#6H1:2][#6:3]

Increase in size of redox chemical space with additional chemical transformations

In this section, we calculate how the number of compounds to consider would expand by including the following types of additional reactions: carboxylation and decarboxylation, keto-enol tautomerization, carbon-carbon double-bond formation, intramolecular redox reactions, and different stereoisomers. In all cases, in order to avoid over-counting we take into consideration molecular symmetry under 180-degree rotation about the central axis. That is, we only count unique (accounting for symmetry) carbon atom sites for a given chemical reaction. Also if a molecule can undergo a given transformation in more than one site (carbon atom), we count all possibilities as separate products.

(1) Carboxylations: We assume that any carbon atom in an aldehyde/ketone, alcohol, or hydrocarbon oxidation state can undergo carboxylation. We restrict ourselves to linear carbon chain products, and therefore only consider edge carbon atoms (i.e. atoms 1 and 4 in a 4-carbon compound) as possible carboxylation sites.
(2) Decarboxylations: We assume that any carbon atom in carboxylic acid oxidation state can be a site of decarboxylation.
(3) Keto-enol tautomerization: We assume that any ketone or aldehyde next to an hydrocarbon or alcohol functional group can undergo keto-enol tautomerization.
(4) Carbon-carbon double-bond formation: In redox biochemistry, carbon double-bond compounds occur as intermediates between alcohol and hydrocarbon oxidation states (e.g. fumarate is an intermediate between malate and succinate). We assume that any alcohol group with a neighboring carbon atom in a hydrocarbon or alcohol oxidation state can undergo dehydration to become a double bond. Note that an alcohol group next to a ketone/aldehyde or carboxylic acid cannot undergo dehydration into a carbon-carbon double bond.
(5) Intramolecular redox reactions: These transformations do not increase the number of compounds in the molecular network. Rather they consist of 2-electron exchanges between neighboring carbon atoms in a molecule, and therefore just add connections between compounds in the same molecular oxidation level (i.e. columns in Fig. 1).
(6) Stereoisomers: All inner carbon atoms (e.g. carbons 2 or 3 in a 4-carbon linear chain compound) in the alcohol oxidation level with unique carbon side-chains are stereocenters. For each such stereocenter, the molecule can exist as two different stereoisomers. Therefore a molecule with N alcohol group stereocenters can exist as 2N unique stereoisomers.

Table S1: Number of new compounds introduced by considering additional chemical transformations. Each row corresponds to an n-carbon redox chemical space. The last two columns show the total new number of compounds introduced as well as the fold-increase in the size of the compound set relative to the redox chemical spaces considered in this work. *Note that the numbers under the intramolecular redox reactions column (light grey) correspond to new redox connections introduced, not compounds.

	keto -enol	double -bond	de-carboxylation	carboxylation	intra-molecular*	stereo-isomers	total new	fold-increase
2C	2	2	4	12	2*	0	20	2.0
3C	15	11	12	36	35*	12	86	2.9
4C	54	42	36	108	133*	104	344	4.4
5C	189	151	108	324	534*	458	1230	5.3

Computing network degree distributions

The degree of a compound in the redox network is defined as the number of redox reactions - oxidations and reductions - that connect it to molecules with higher or lower oxidation level. We used the network analysis library NetworkX (2) in Python to compute the degree distribution of compounds in the full redox networks.

Comparison against KEGG database

In order to classify compounds in the full redox networks as biological or non-biological, we looked for matches in the KEGG database of metabolic compounds. We did this in several steps. In order to match biological compounds against the n-carbon network, we filtered out metabolites in KEGG containing n-carbon atoms. Then, using the RDKit toolbox, we matched molecules in the networks against KEGG metabolites using their canonicalized smiles string representation (3). In order to additionally capture KEGG compounds that have alcohol functional groups substituted by amine or a phosphate functional groups, we visually inspected all remaining n-carbon molecules in KEGG. Finally, to capture compounds with carboxylic acids activated by Coenzyme A, we generated a list of all KEGG compounds with n-carbon atoms plus a covalently attached Co-A molecule. Manual search of this list led to the final set of biological metabolites matching compounds in our full redox networks.

Computing the null distribution for the expected number of n-gram (single, pair and triplet) functional group patterns

Borrowing terminology from natural language processing, we call the set of all possible sequences of one, two, and three carbon functional groups the set of oxidation level n-grams. The goal is to count the number of times that each n-gram appears in the set of biological (or non biological) compounds (where N is the total number of biological compounds), and compare that against properly generated random sets of compounds (the null distribution).

The analytical null distribution for single functional group patterns (1-grams)

We first note that a given n-gram can appear more than once in a single molecule. For example, the metabolite succinate has the functional group sequence {carboxylic acid, hydrocarbon, hydrocarbon, carboxylic acid}. Thus it contains two instances of the {carboxylic acid, hydrocarbon} 2-gram. In general, a 4-carbon linear-chain compound can have up to 4 instances of a 1-gram, up to 3 instances of a 2-gram, and up to 2 instances of a 3-gram.

Let $n(k;g)$ be the number of molecules in the full redox network with k instances of 1-gram g**.** For example, $n(1;hydroxyl)$ is the total number of compounds in the network with a single hydroxyl functional group. Assume a set of N molecules are randomly sampled without replacement from the network. Let $m(g)$ be the total number of instances of the 1-gram g in this random set. These $m(g)$ instances can come from different sampling configurations of molecules, each with k instances of the 1-gram g. We call $m(k;g)$ be the number of molecules in the random sample with k instances of the 1-gram g**.**

To give a concrete example, assume a random set of size $N = 30$ molecules contains 16 instances of the n-gram g; thus $m(g) = 16$. One of the very many sampling configuration that can lead to this value of $m(g)$ is sampling 17 molecules with zero instances of g, 10 molecules with 1 instance of g, and 3 molecule with two instances of g. Thus

$$m(0;g) = 17,\ m(1;g) = 10,\ m(2;g) = 3$$

The total number of instances of the 1-gram g in the sample is given by:

$$m(g) = 0 \cdot m(0;g) + 1 \cdot m(1;g) + 2 \cdot m(2;g) + 3 \cdot m(3;g) = 16$$

Note that the following constraint is satisfied:

$$m(0;g) + m(1;g) + m(2;g) + m(3;g) = 30$$

In order to compute the probability of having $m(g)$ instances of the 1-gram g, we need to account for all such possible sampling configurations that add up to $m(g)$. The number of ways of sampling $m(k;g)$ molecules with k instances of g is given by $\binom{n(k;g)}{m(k;g)}$. In general, given a sample size N and value of $m(g)$ for n-gram g, the number of all possible sampling configurations that lead to that value of m(g) is given by:

$$P(m(g),\ N) = \sum_{constraints} \prod_{k} \binom{n(k;g)}{m(k;g)}$$

Where the summation is over terms that satisfy the following two constraints:

$$m(g) = 0 \cdot m(0;g) + 1 \cdot m(1;g) + 2 \cdot m(2;g) + 3 \cdot m(3;g)$$

$$N = m(0;g) + m(1;g) + m(2;g) + m(3;g)$$

Normalizing each value of $P(m(g),\ N)$ over the sum of all values leads to the probability of observing $m(g)$ instances of the 1-gram g in a sample of size N, $p(m(g), N)$. We numerically obtain the value of $n(k;g)$ for k = 0, 1, 2, 3, 4 and g = {carboxylic acid, aldehyde/ketone, hydroxyl, and hydrocarbon}. We then numerically compute the value of $P(m(g),\ N)$ by obtaining all sampling configurations that satisfy the constraints. We take N to be equal to the number biological compounds in the full redox network.

The empirical null distributions for functional group pair and triplet patterns (2- and 3-grams)

Obtaining the proper null distribution for oxidation level pair and triplet patterns (2-grams and 3-grams) requires accounting for (or normalizing) for the observed single functional group statistics (1-grams). For example, the 2-gram pattern [carbonyl-carbonyl] seems to appear infrequently in the biological set of metabolites. Is this due to selection against this specific 2-gram pattern, or is it simply due to the general depletion of aldehydes/ketones (carbonyls) (the 1-gram pattern) in the biological compounds? In order to address this, one needs to generate random sets of N compounds that control for or conserve the 1-gram statistics of the biological set of compounds. We numerically generate random molecules that conserve 1-gram statistics. In the case of 4-carbon linear chain molecules, we randomly choose the identity of the functional group at positions $n = (1, 2, 3, 4)$ by sampling from a discrete distribution

$$p_g = g_N / (4N)$$

Where g_N is the number of instances of 1-gram g in the biological set, and N is the number of molecules in the biological set. Importantly, in order to avoid sampling carboxylic acids in the inner carbon atoms of a molecule (positions $n = 2$ *and* 3), we obtain separate functional group distributions for the inner and the outer carbon atom positions.

Cheminformatic prediction of solubility (logS)

We used the cheminformatics software ChemAxon (Marvin 17.7.0, 2017, ChemAxon) to predict the pH-dependent solubility, logS(pH), of biological and non-biological compounds in the full redox networks. Specifically, we use the calculator plugin cxcalc logs. The cxcalc solubility calculator is based

on a parametrized fragment-based model (the atom-contribution approach) fit to sets of experimental logS data (4, 5).

Predicting standard redox potentials with calibrated quantum chemistry approach

Our method relies on computing the electronic structure and energy of the fully protonated species of each metabolite. We obtain the smiles string for the fully protonated species and generate initial geometric conformation (with up to 10 initial conformers per metabolite) using ChemAxon (Marvin 17.7.0, 2017, ChemAxon).

All quantum chemistry calculations were performed using the Orca quantum chemistry software (6) version 3.0.3. We first perform a geometry optimization using density functional theory with the B3LYP functional (7), with Orca's DefBas-2 basis set, COSMO implicit solvation (8), and D3 dispersion correction (9). We then perform an additional electronic single point energy (SPE) using the double-hybrid functional B2PLYP (10, 11) (with the DefBas-5 Orca basis set, COSMO implicit solvation (8), and D3 dispersion correction (9)). We note that the model chemistry selected - the combination of DFT functional, basis set, implicit solvent model, and dispersion correction for both the geometry optimization and the single point energy - was done based on a combinatorial exploration of different options.

We Boltzmann average the electronic energies of compounds, and obtain the difference in electronic energies of products and substrates for all redox reactions in the full redox networks. Every redox reaction (in the direction of reduction) was balanced by a hydrogen molecule H_2 in the substrate side of the equation. Reductions of carboxylic acids to aldehydes and reductions of alcohols to hydrocarbons were balanced with a water molecule H_2O in the product side of the equation.

The difference in product and substrate electronic energies is an estimate of the chemical redox potential for the fully protonated species, E^o(fully protonated species). In order to convert this chemical potential to the biochemical potential at pH = 7, $E^{o'}$(pH=7), we use pKa estimates from Chemaxon (Marvin 17.7.0, 2017, ChemAxon) and the Alberty Legendre transform.

Our approach relies on several approximations, such as ignoring vibrational enthalpy and entropy contributions to the formation Gibbs energy of compounds. In order to correct for systematic in the quantum chemistry methodology and the empirical pKa estimates used, we calibrate predictions against available experimental data using linear regression.

Predicting standard redox potentials with the group contribution method

The group contribution method relies on a fragment-based decomposition of compounds into group, each of which is assigned a group energy based on available experimental data (12–15). Reaction energy estimates are obtained by taking the difference of the group energy vectors of products and substrates. We used the group contribution method as implemented by Noor et al. (15) to estimate the redox potentials of the set of linear-chain carbon redox reactions with experimental values.

Determining statistical significance

For all tests of statistical significance (i.e. differences in solubilities, n-gram counts, octanol-water partition coefficients, Gibbs energies of biological vs. non-biological compounds) we performed Welch's unequal variance t-test, which is an adaptation of Student's t-test that does not assume equal variance.

1. D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36 (1988).

2. Aric A. Hagberg, Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function using NetworkX (November 19, 2017).

3. N. M. O'Boyle, Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* 4, 22 (2012).

4. T. J. Hou, K. Xia, W. Zhang, X. J. Xu, ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* 44, 266–275 (2004).

5. E. Shoghi, E. Fuguet, E. Bosch, C. Ràfols, Solubility-pH profiles of some acidic, basic and amphoteric drugs. *Eur. J. Pharm. Sci.* 48, 291–300 (2013).

6. F. Neese, The ORCA program system. *WIREs Comput Mol Sci* 2, 73–78 (2012).

7. A. D. Becke, Density‐functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* 98, 5648–5652 (1993).

8. A. Klamt, G. Schüürmann, COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans. 2* 0, 799–805 (1993).

9. S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* 132, 154104 (2010).

10. T. Schwabe, S. Grimme, Towards chemical accuracy for the thermodynamics of large molecules: new hybrid density functionals including non-local correlation effects. *Phys. Chem. Chem. Phys.* 8, 4398–4401 (2006).

11. S. Grimme, Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* 124, 034108 (2006).

12. M. D. Jankowski, C. S. Henry, L. J. Broadbelt, V. Hatzimanikatis, Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* 95, 1487–1499 (2008).

13. E. Noor, *et al.*, An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. *Bioinformatics* 28, 2037–2044 (2012).

14. A. Flamholz, E. Noor, A. Bar-Even, R. Milo, eQuilibrator--the biochemical thermodynamics calculator. *Nucleic Acids Res.* 40, D770–5 (2012).

15. E. Noor, H. S. Haraldsdóttir, R. Milo, R. M. T. Fleming, Consistent estimation of Gibbs energy using component contributions. *PLoS Comput. Biol.* 9, e1003098 (2013).

Supplementary Figures:

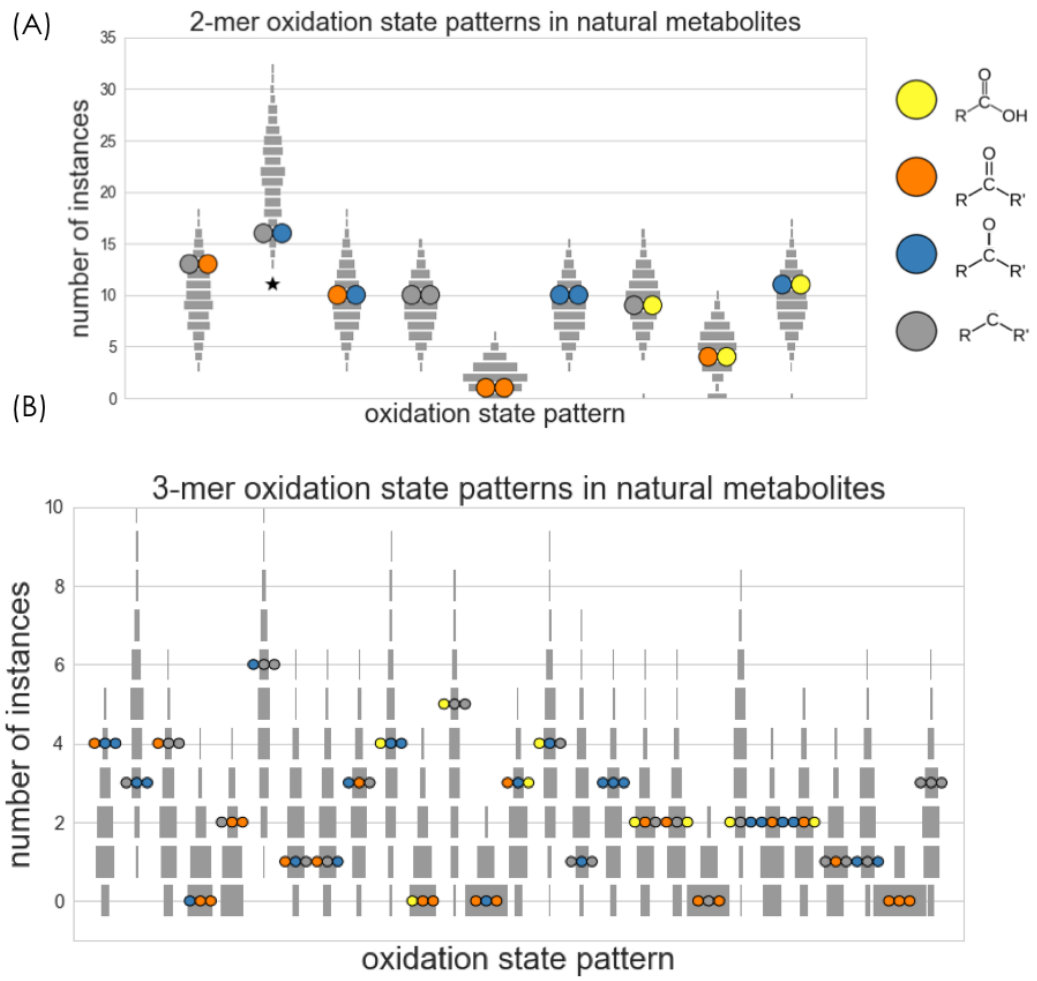


Figure S1: Enrichment and depletion of functional group pair and triplet patterns in 4-carbon linear-chain redox chemical space. A) The number of times each possible pattern of nearest neighbor functional group pairs appears in the set of biological metabolites are shown as pairs of colored circles. Gray squares correspond to the empirically-derived null distributions for randomly sampled sets of molecules from the network. The null distributions account for the single functional group (1-mer) statistics (SI Appendix, Supplemental Methods). The pattern *hydrocarbon-alcohol* is depleted in the biological compounds, but with weak statistical significance (p = 0.05). B) The number of times each possible pattern of functional group triplets appears in the set of biological metabolites. No patterns are significantly enriched or depleted in the set of biological metabolites.

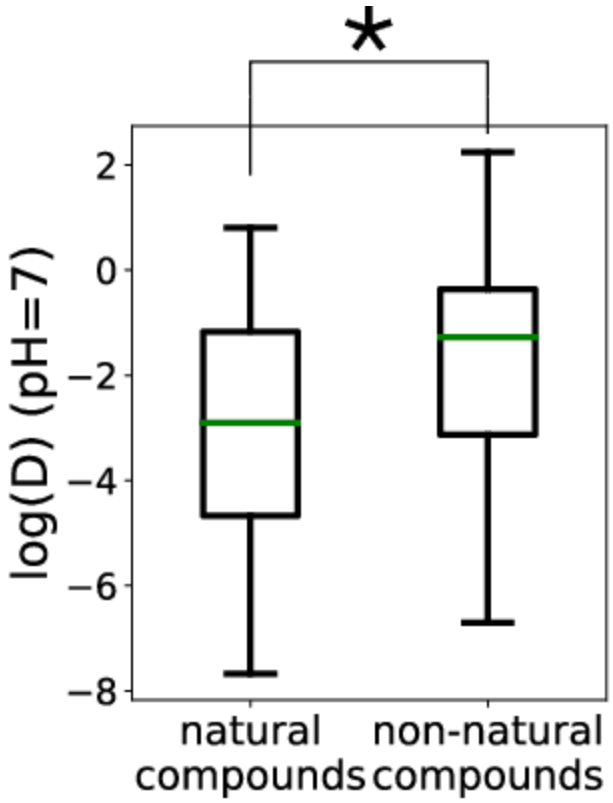


Figure S2: Octanol water partition coefficients of biological vs. non-biological compounds in 4-carbon linear-chain redox chemical space. Comparison of predicted octanol-water partition coefficient log(P) at pH=7 for biological and non-biological compounds in the 4-carbon linear-chain redox network. This is also known as the distribution coefficient (logD). biological compounds have significantly lower logP(pH=7) than the non-biological set ($p < 0.01$).

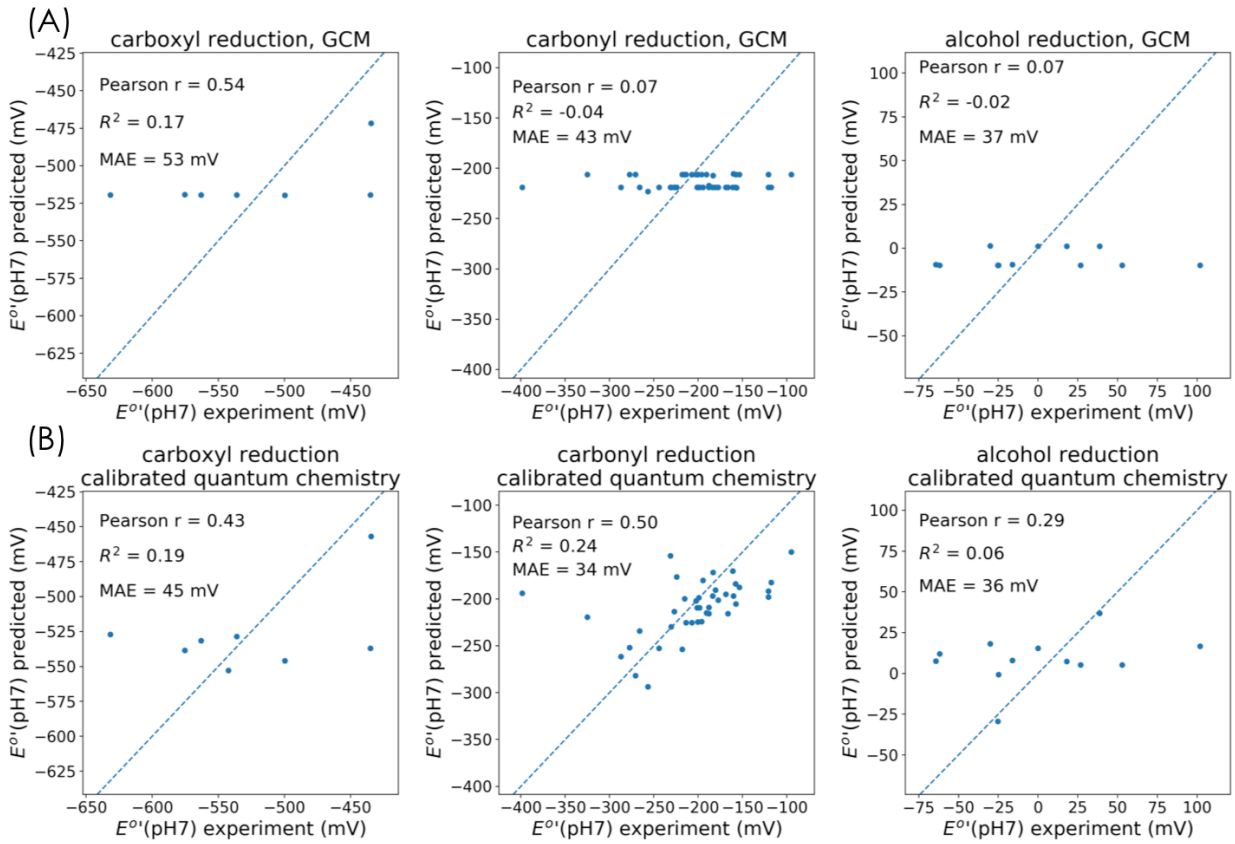


Figure S3: Accuracy of group contribution method and calibrated quantum chemistry redox potential predictions. Experimental data was obtained from the NIST database for Thermodynamics of Enzyme-Catalyzed Reactions (TECRDB) A) Group contribution method prediction accuracies for reduction potentials of carboxylic acids, aldehydes/ketones, and alcohol functional groups in linear chain compounds. A) Calibrated quantum chemistry prediction accuracies for reduction potentials of carboxylic acids, aldehydes/ketones, and alcohol functional groups in linear chain compounds. Quantum chemistry calculations were performed using density functional theory with a double hybrid functional (B2PLYP), and calibrated against experimental data using linear regression.

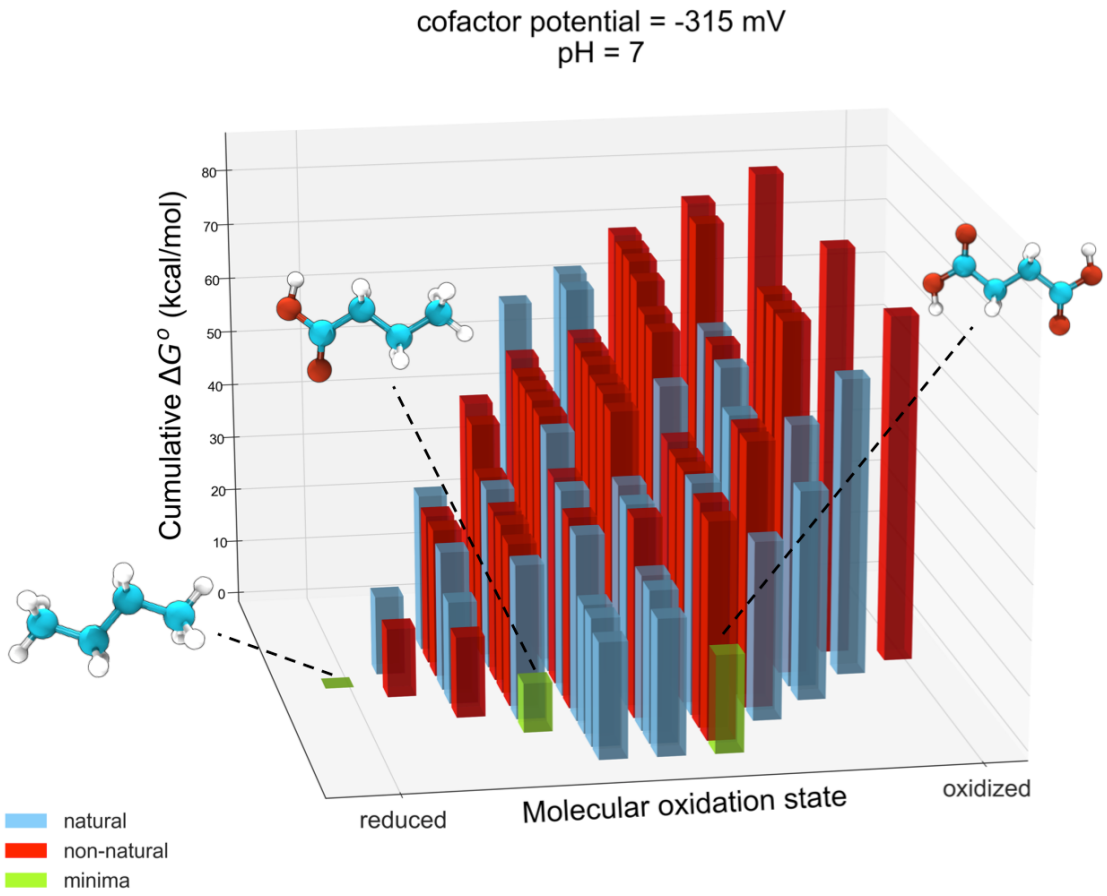


Figure S4: Thermodynamic landscape of 4-carbon linear chain redox network at pH=7 and cofactor potential E(cofactor) = -315 mV. Gibbs energies are normalized relative to the metabolite with the lowest energy (butane). Thus the cumulative Gibbs energies of a metabolite is obtained by summing up the Gibbs reaction energies of all reactions leading to it from the reference metabolite. Compounds within a column (i.e. with the same molecular oxidation state) are sorted according to their energies. The three compounds - butane, butanoic acid, and succinate - which are local minima in the thermodynamic landscape are shown. These local minima have lower energy than any of their neighboring molecules which are accessible by either a reduction or an oxidation.

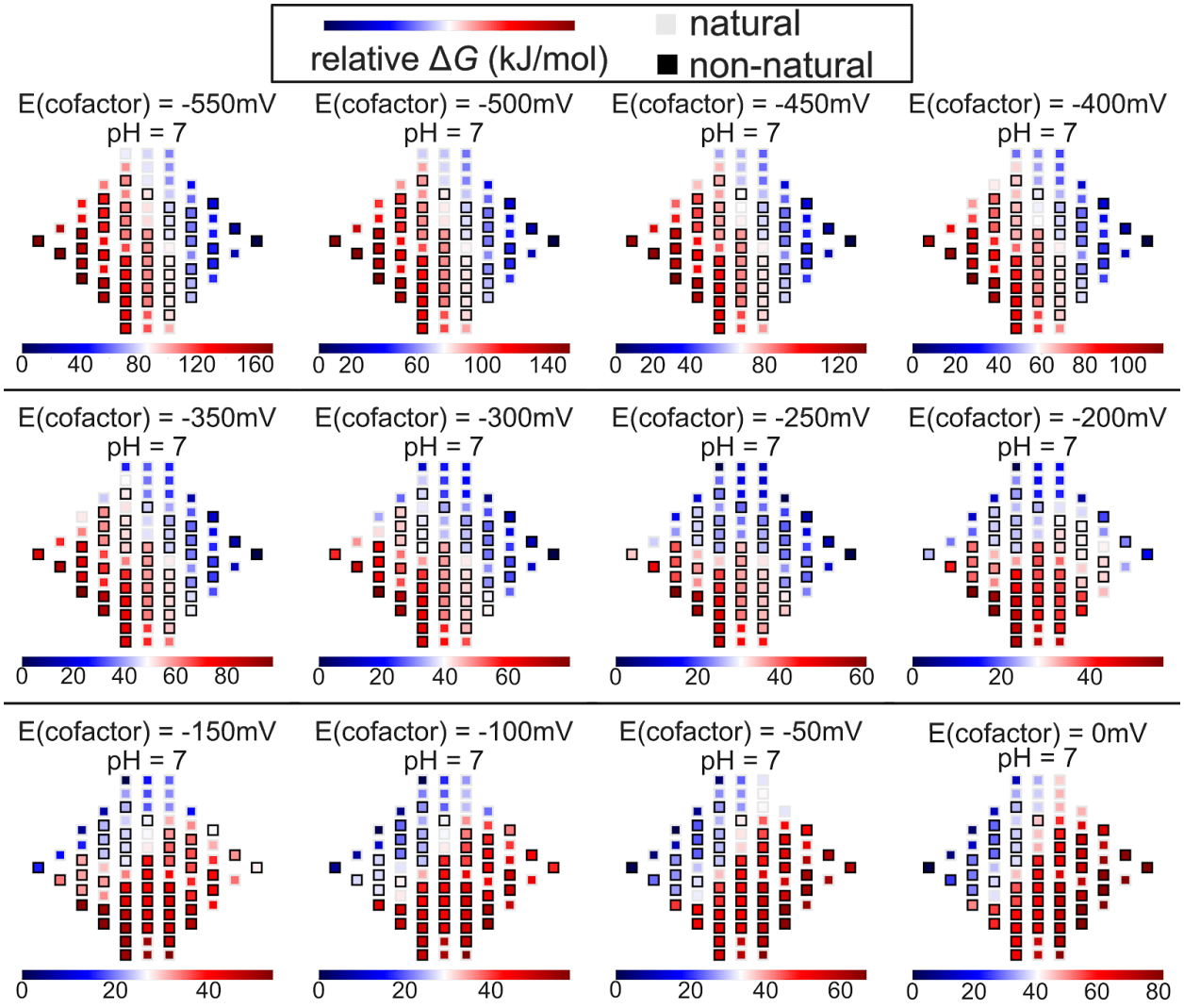


Figure S5: Thermodynamic landscape of the 4-carbon network at a fixed value of pH and a range of values of electron donor potential (-550 mV to 0 mV). Relative Gibbs energies of compounds is color-coded with blue (low energies) to red (high energies). Higher values of the electron donor potential energetically drive the redox chemical space towards more oxidized compounds, while lower values energetically drive the redox chemical space towards more reduced compounds.

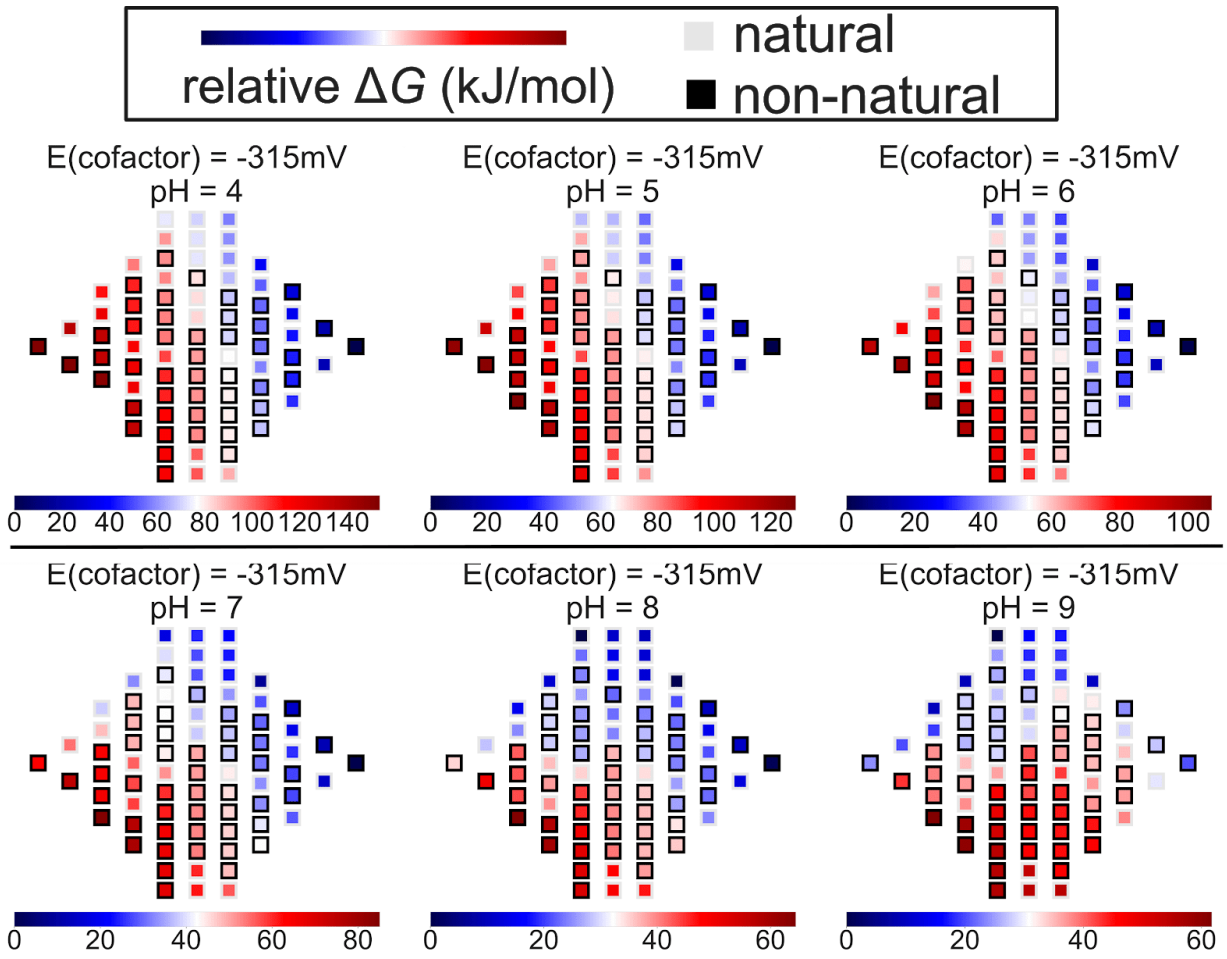


Figure S6: Thermodynamic landscape of the 4-carbon network at a fixed value of electron donor potential and a range of values of pH (4-9). Relative Gibbs energies of compounds is color-coded with blue (low energies) to red (high energies). Acidic pH drives the landscape towards more reduced compounds, while basic pH drives the landscape to more oxidized compounds.

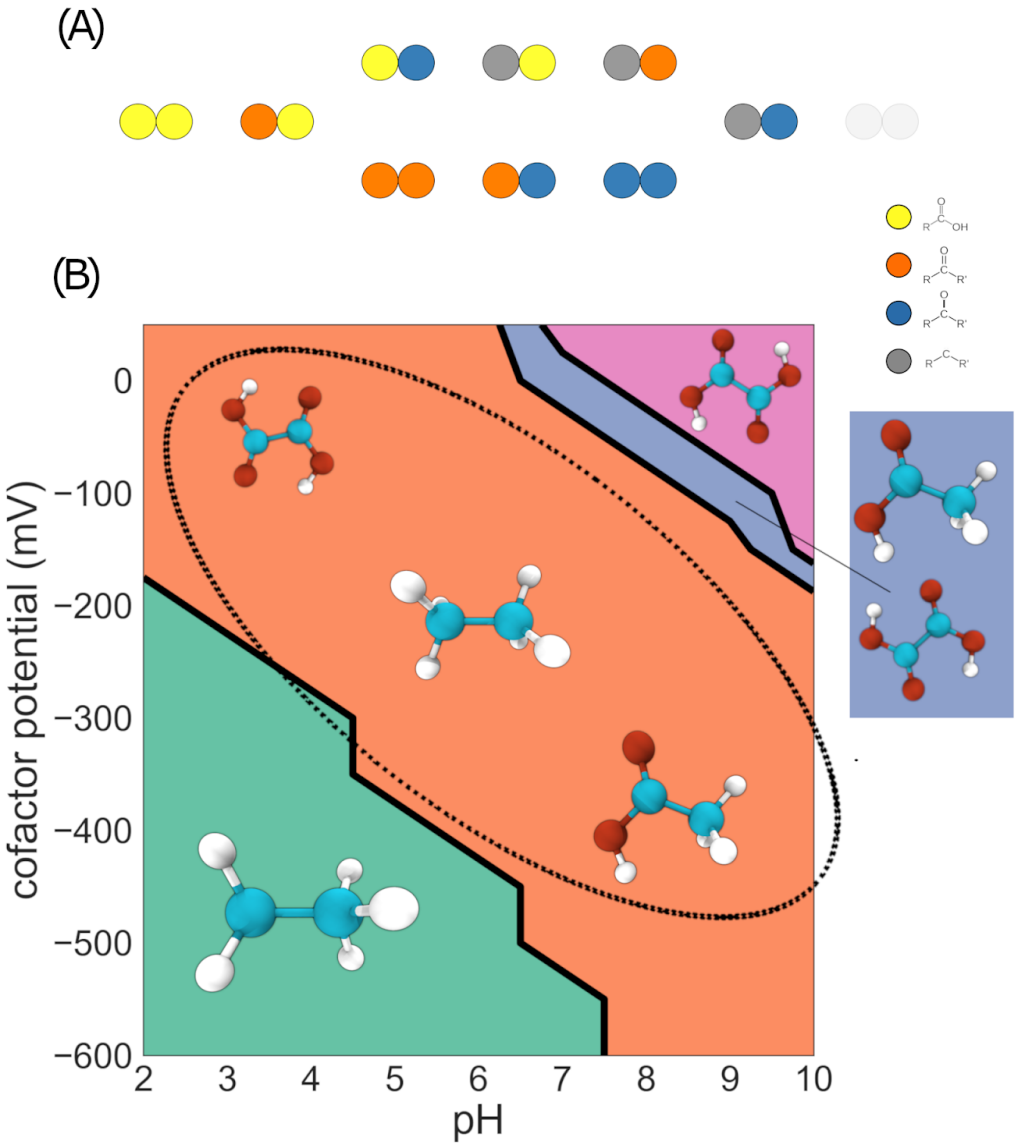


Figure S7: 2-carbon redox chemical space. A) The subset of molecules in 2-carbon linear-chain redox chemical space that match biological metabolites in the KEGG database. Carbon atoms are represented as colored circles, with each color corresponding to an oxidation state: yellow = carboxylic acid; orange = aldehyde/ketone; blue = hydroxycarbon; gray = hydrocarbon. Only the fully reduced hydrocarbon ethane does not match a biological metabolite. B) Pourbaix phase diagram for the 2-carbon linear chain redox chemical space. Molecules that are local minima in the energy landscape at each region of pH vs. E(electron donor/acceptor) phase space are shown. At low pH and E(electron donor/acceptor) values, ethane is both the global and the only local minimum energy compound, while at high pH and E(electron donor/acceptor) values, the fully oxidized oxalate is both the global and the only local minimum energy compound. The dashed circle highlights the region of phase space where the dicarboxylic acid oxalate and the 2-carbon fatty-acid acetate (along with ethane) are the local minima.

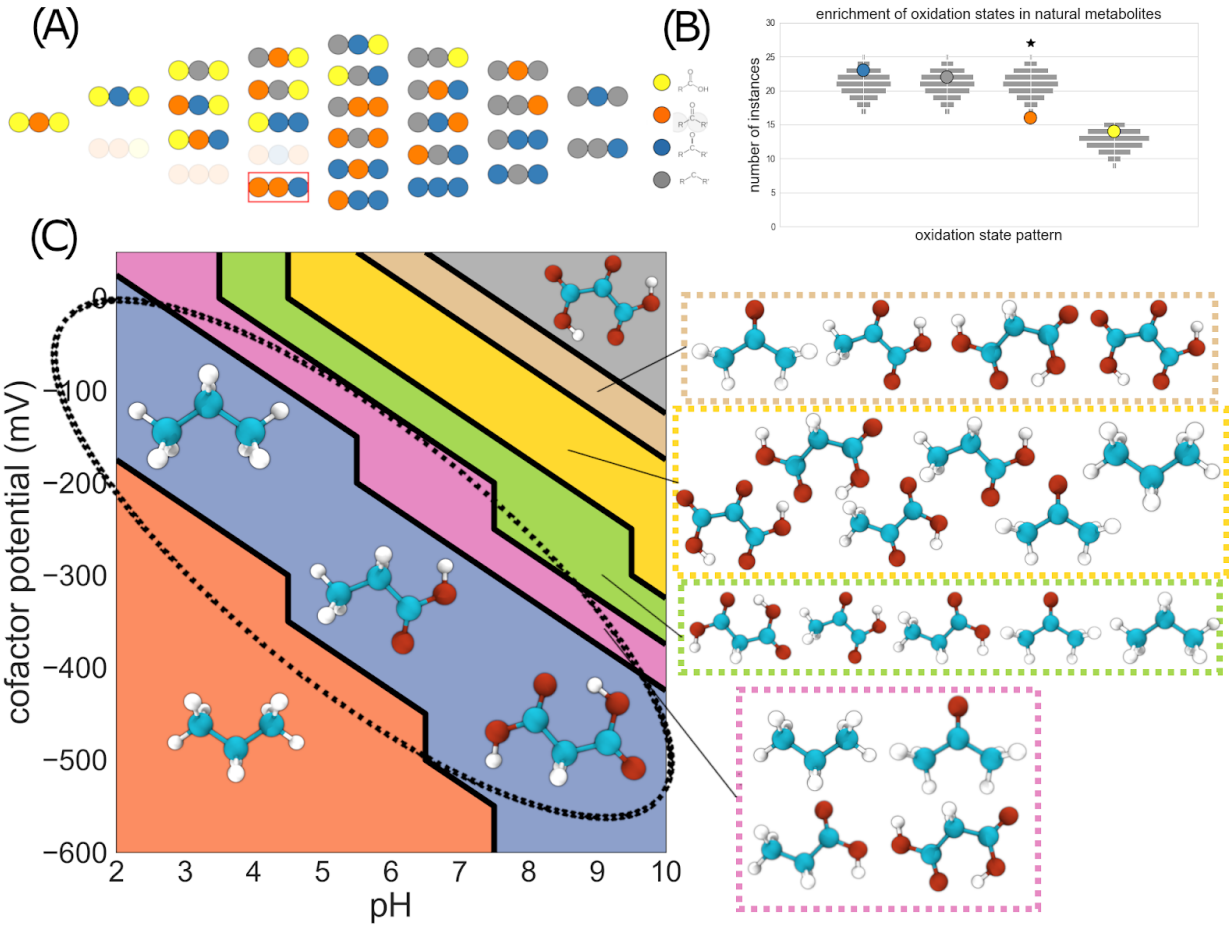


Figure S8: 3-carbon redox chemical space. A) The subset of molecules in 3-carbon linear-chain redox chemical space that match biological metabolites in the KEGG database. Carbon atoms are represented as colored circles, with each color corresponding to an oxidation state: yellow = carboxylic acid; orange = aldehyde/ketone; blue = hydroxycarbon; gray = hydrocarbon. B) Enrichment and depletion of functional groups in the set of biological compounds. The vertical position of each colored circle corresponds to the number of times each functional group appears in the set of biological compounds. The light gray squares show the corresponding expected null distributions for random sets of molecules sampled from redox chemical space. C) Pourbaix phase diagram for the 3-carbon linear chain redox chemical space. Molecules that are local minima in the energy landscape at each region of pH vs. E(electron donor/acceptor) phase space are shown. At low pH and E(electron donor/acceptor) values, propane is both the global and the only local minimum energy compound, while at high pH and E(electron donor/acceptor) values, the fully oxidized 3-carbon compound (oxomalonate) is both the global and the only local minimum energy compound. The dashed circle highlights the region of phase space where the dicarboxylic acid malonate and the 3-carbon fatty-acid propionate (along with propane) are the only local minima.

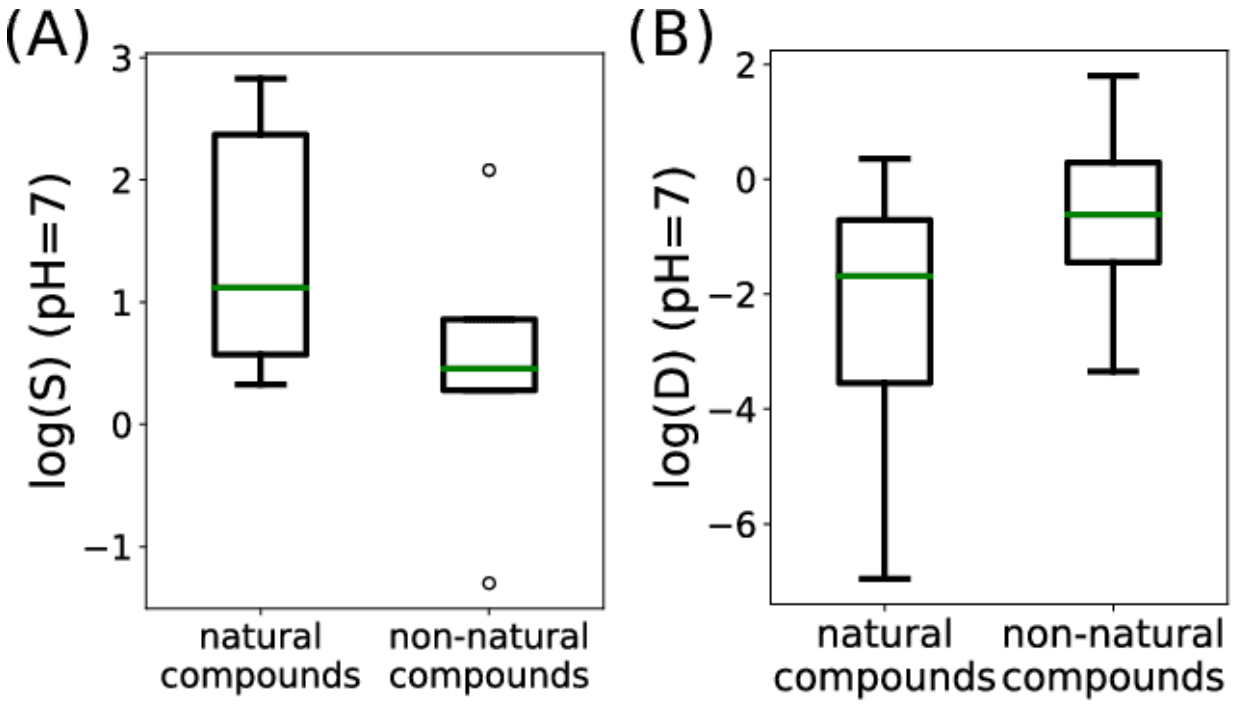


Figure S9: Solubility and octanol-water partition coefficients of biological and non-biological compounds in 3-carbon redox chemical space. Comparison of predicted aqueous solubility log(S) at pH=7 for biological and non-biological compounds in the 3-carbon linear-chain redox chemical space. Although the biological compounds in the 3-carbon redox chemical space tend to have higher solubilities and lower octanol-water partition coefficients at pH=7, the differences are not statistically significant (Welch's t-test).

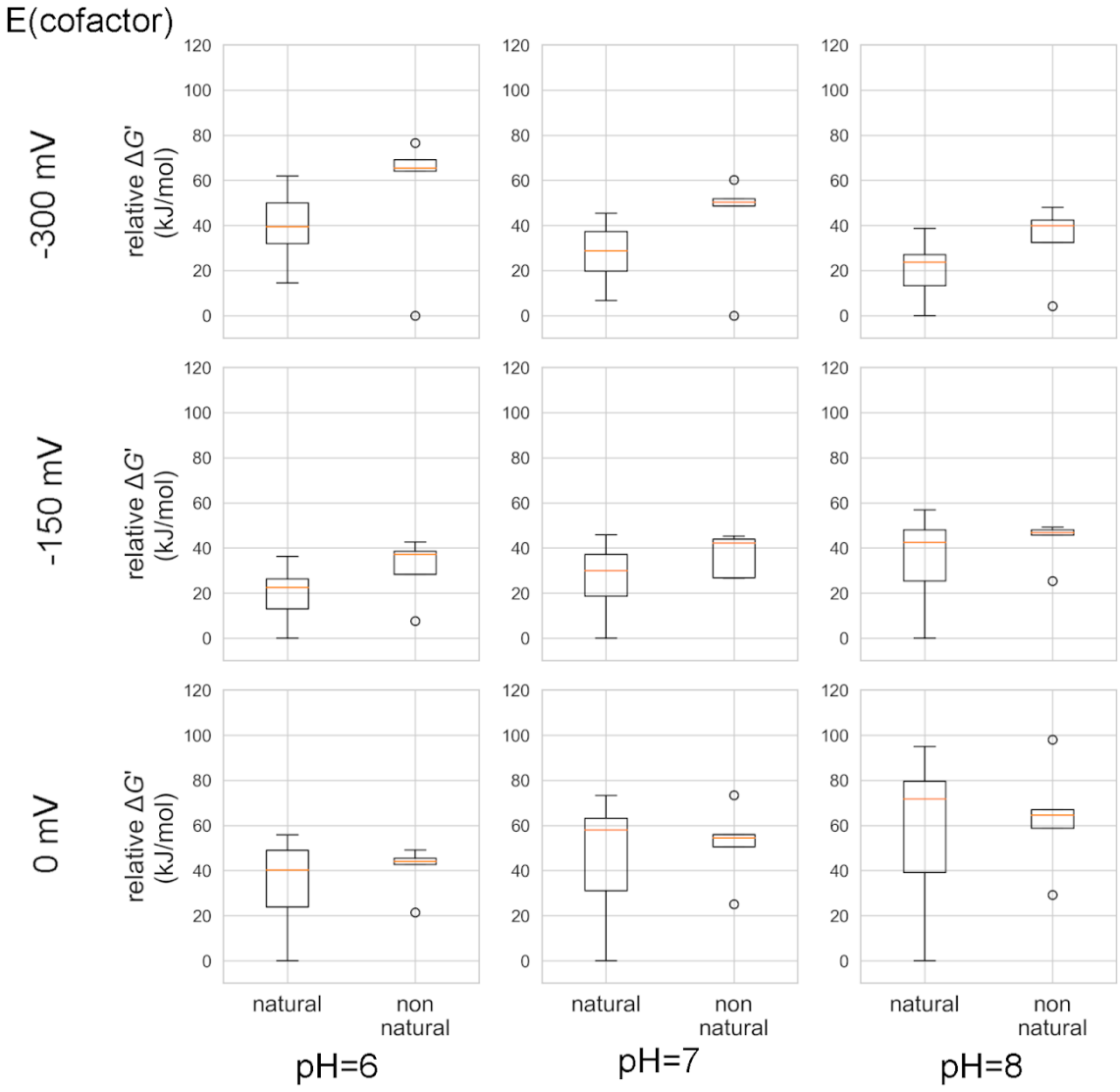


Figure S10: Relative energies of biological and non-biological compounds in 3-carbon redox chemical space. Relative Gibbs energies of biological and non-biological compounds in the 3-carbon redox chemical space for a range of pH and E(electron donor/acceptor) values. At each value of pH and E(electron donor/acceptor), Gibbs energies are normalized relative to the compound with the lowest energy. Although biological metabolites tend to have, on average, lower energies than the non-biological compounds, the differences are not statistically significant (Welch's t-test, $p>0.05$). The low energy of the fully reduced propane across conditions tends to bring down the average relative energy of the non-biological compounds.

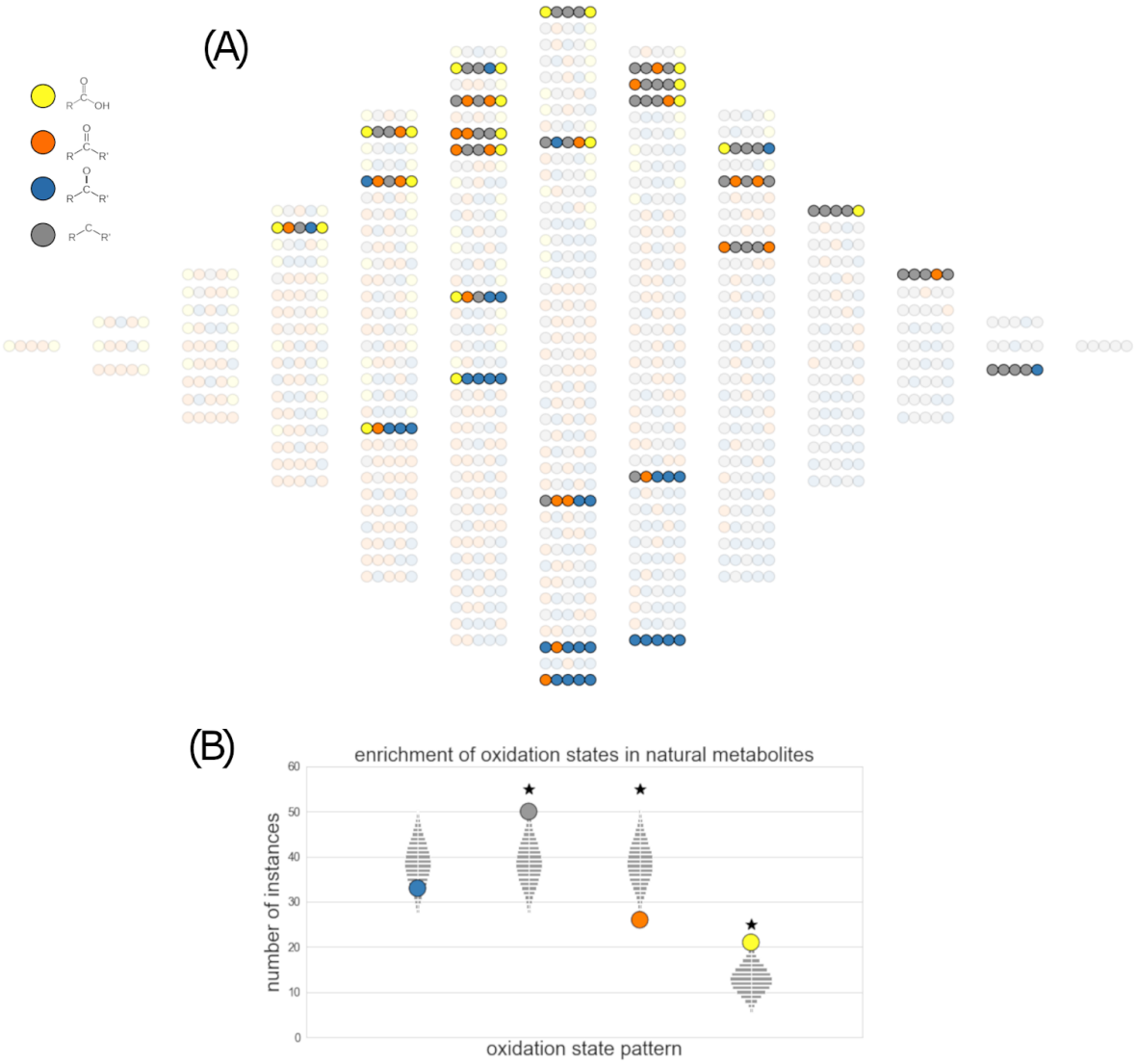


Figure S11: 5-carbon linear-chain redox chemical space. A) The subset of molecules in 5-carbon linear-chain redox chemical space that match biological metabolites in the KEGG database. Carbon atoms are represented as colored circles, with each color corresponding to an oxidation state: yellow = carboxylic acid; orange = aldehydes/ketones; blue = hydroxycarbon; gray = hydrocarbon. B) Enrichment and depletion of functional groups in the set of biological compounds. The vertical position of each colored circle corresponds to the number of times each functional group appears in the set of biological compounds. The light gray squares show the corresponding expected null distributions for random sets of molecules sampled from redox chemical space.

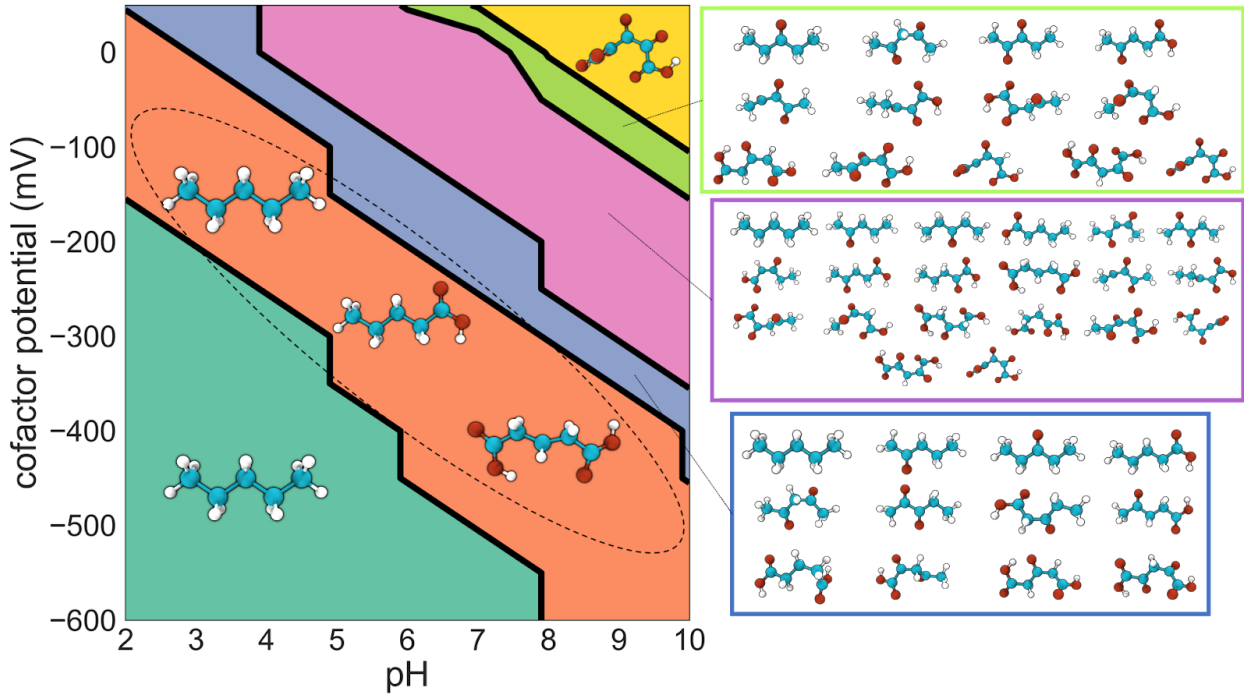


Figure S12: Pourbaix phase diagram for 5-carbon linear chain redox chemical space. Molecules that are local minima in the energy landscape at each region of pH vs. E(electron donor/acceptor) phase space are shown. At low pH and E(electron donor/acceptor) values, pentane is both the global and the only local minimum energy compound, while at high pH and E(electron donor/acceptor) values, the fully oxidized 5-carbon compound (2,3,4-trioxoglutarate) is both the global and the only local minimum energy compound. The dashed circle highlights the region of phase space where the dicarboxylic acid glutarate and the 5-carbon fatty-acid valerate (along with pentane) are the only local minima.

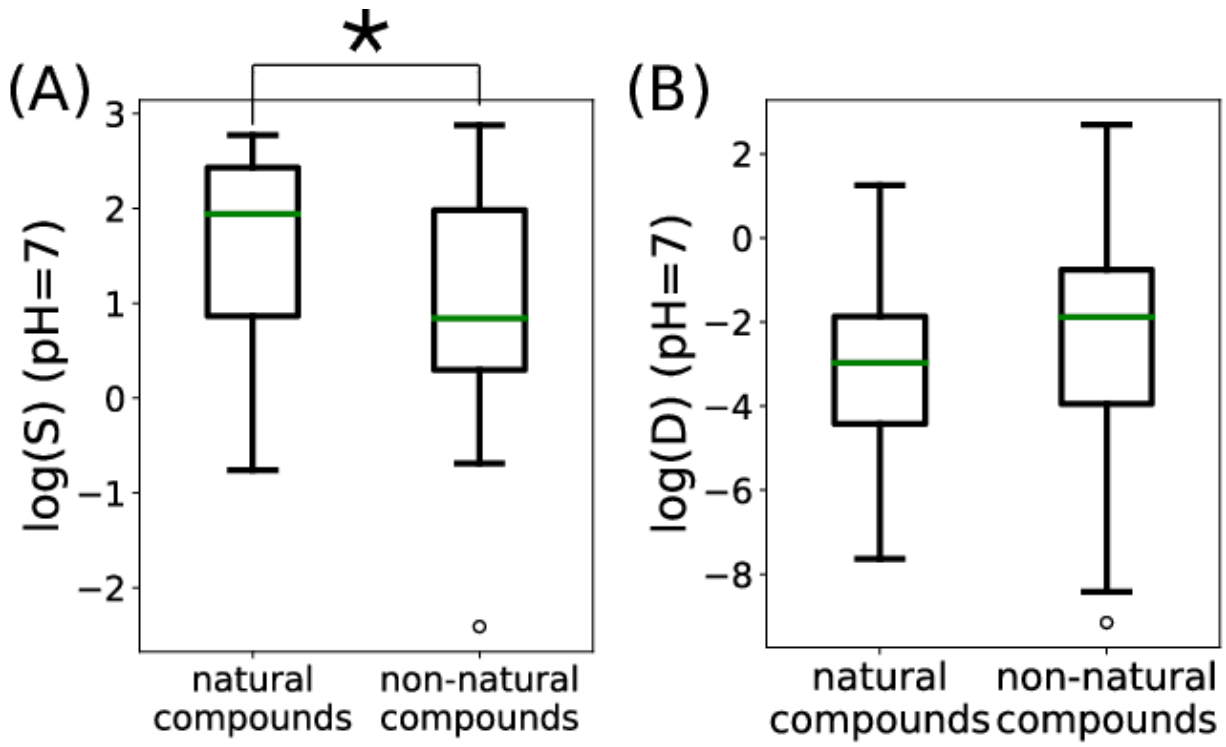


Figure S13: Solubility and octanol-water partition coefficients of biological and non-biological compounds in 5-carbon linear-chain redox chemical space. Comparison of predicted aqueous solubility log(S) and predicted octanol-water partition coefficient log(P) at pH=7 for biological and non-biological compounds in the 5-carbon linear-chain redox chemical space. biological compounds have statistically significantly higher solubilities (Welch's t-test) than the non-biological set ($p < 0.05$).

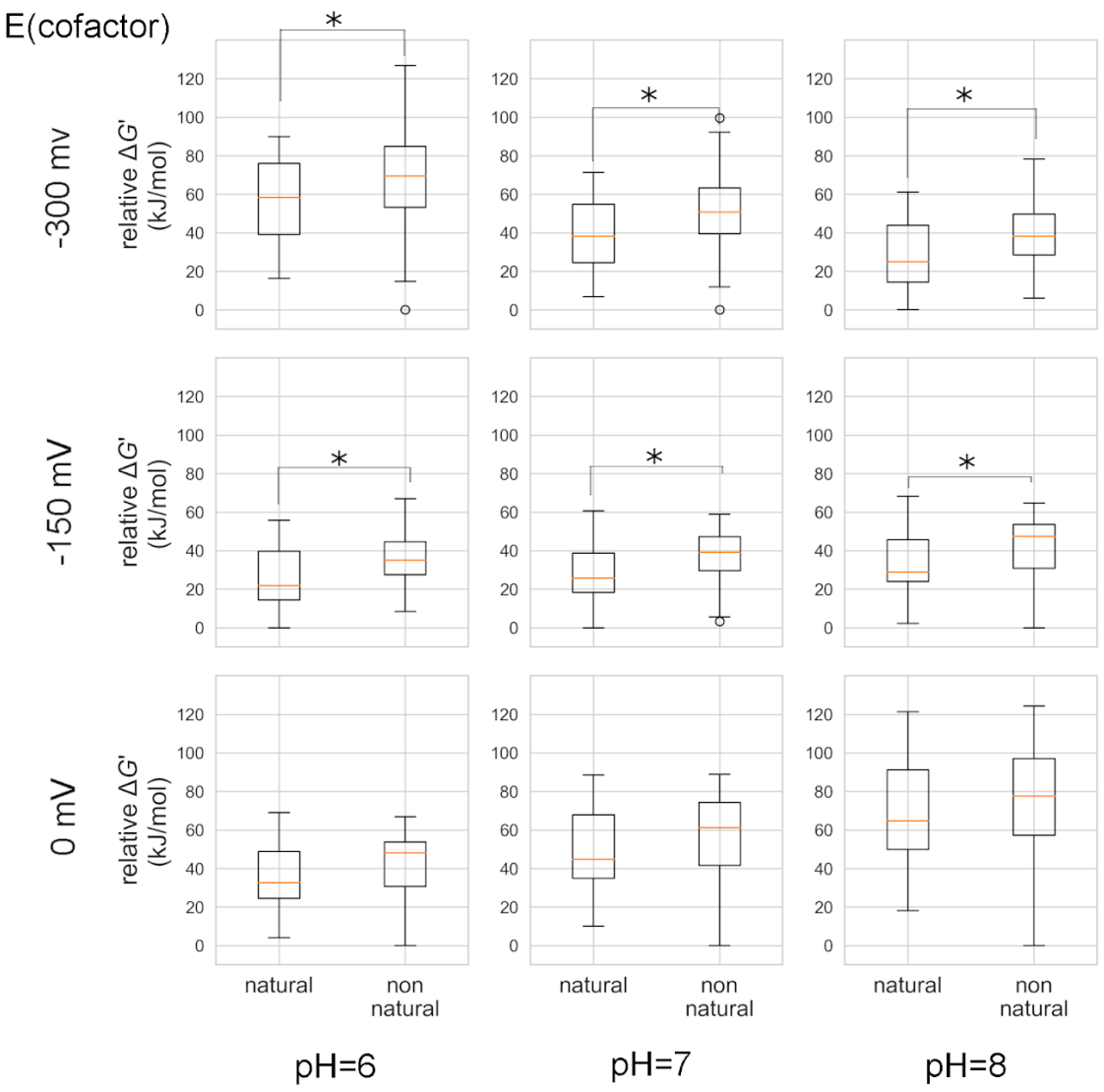


Figure S14: Relative energies of biological and non-biological compounds in 5-carbon linear-chain redox chemical space. Relative Gibbs energies of biological and non-biological compounds in 5-carbon redox chemical space for a range of pH and E(electron donor/acceptor) values. At each value of pH and E(electron donor/acceptor), Gibbs energies are normalized relative to the compound with the lowest energy. An asterisk indicates a statistically significant difference in the average values (Welch's t-test) ($p<0.05$).