ACS central science

Research Article

Cite This: ACS Cent. Sci. 2019, 5, 1199−1210

http://pubs.acs.org/journal/acscii

A Mixed Quantum Chemistry/Machine Learning Approach for the Fast and Accurate Prediction of Biochemical Redox Potentials and Its Large-Scale Application to 315 000 Redox Reactions

Adrian Jinich,[†,‡,#] Benjamin Sanchez-Lengeling,[†,#] Haniu Ren,[†] Rebecca Harman,[†] and Alán Aspuru-Guzik*[,§,∥,⊥]

[†]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

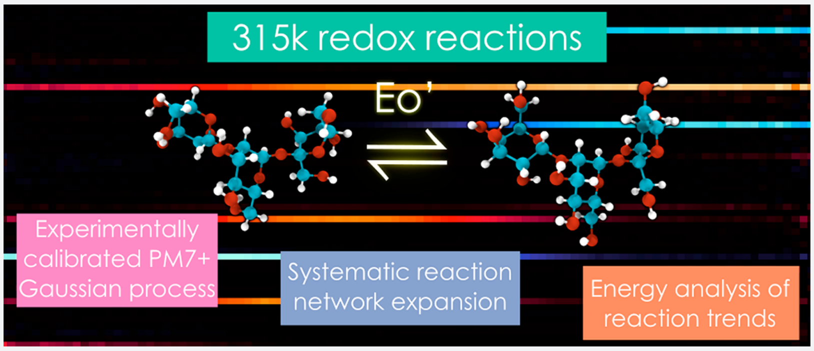
[‡]Division of Infectious Diseases, Weill Department of Medicine, Weill−Cornell Medical College, New York, New York 10065, United States

[§]Department of Chemistry and Department of Computer Science, University of Toronto, 80 St. George Street, Toronto, Ontario M5S 3H6, Canada

[∥]Vector Institute, Toronto, Ontario M5G 1M1, Canada

[⊥]Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada

Supporting Information


315k redox reactions
Eo'
Experimentally calibrated PM7+ Gaussian process
Systematic reaction network expansion
Energy analysis of reaction trends

ABSTRACT: A quantitative understanding of the thermodynamics of biochemical reactions is essential for accurately modeling metabolism. The group contribution method (GCM) is one of the most widely used approaches to estimate standard Gibbs energies and redox potentials of reactions for which no experimental measurements exist. Previous work has shown that quantum chemical predictions of biochemical thermodynamics are a promising approach to overcome the limitations of GCM. However, the quantum chemistry approach is significantly more expensive. Here, we use a combination of quantum chemistry and machine learning to obtain a fast and accurate method for predicting the thermodynamics of biochemical redox reactions. We focus on predicting the redox potentials of carbonyl functional group reductions to alcohols and amines, two of the most ubiquitous carbon redox transformations in biology. Our method relies on semiempirical quantum chemistry calculations calibrated with Gaussian process (GP) regression against available experimental data and results in higher predictive power than the GCM at low computational cost. Direct calibration of GCM and fingerprint-based predictions (without quantum chemistry) with GP regression also results in significant improvements in prediction accuracy, demonstrating the versatility of the approach. We design and implement a network expansion algorithm that iteratively reduces and oxidizes a set of natural seed metabolites and demonstrate the high-throughput applicability of our method by predicting the standard potentials of more than 315 000 redox reactions involving approximately 70 000 compounds. Additionally, we developed a novel fingerprint-based framework for detecting molecular environment motifs that are enriched or depleted across different regions of the redox potential landscape. We provide open access to all source code and data generated.

INTRODUCTION

All living systems are sustained by complex networks of biochemical reactions that extract energy from organic compounds and generate the building blocks that make up cells.[1] Recent work[2−5] has revived decades-old efforts[6−8] to obtain a quantitative understanding of the thermodynamics of such metabolic networks. Accurately predicting the thermodynamic parameters, such as Gibbs reaction energies and redox potentials, of biochemical reactions informs both metabolic

Received: March 24, 2019
Published: June 7, 2019

DOI: 10.1021/acscentsci.9b00297

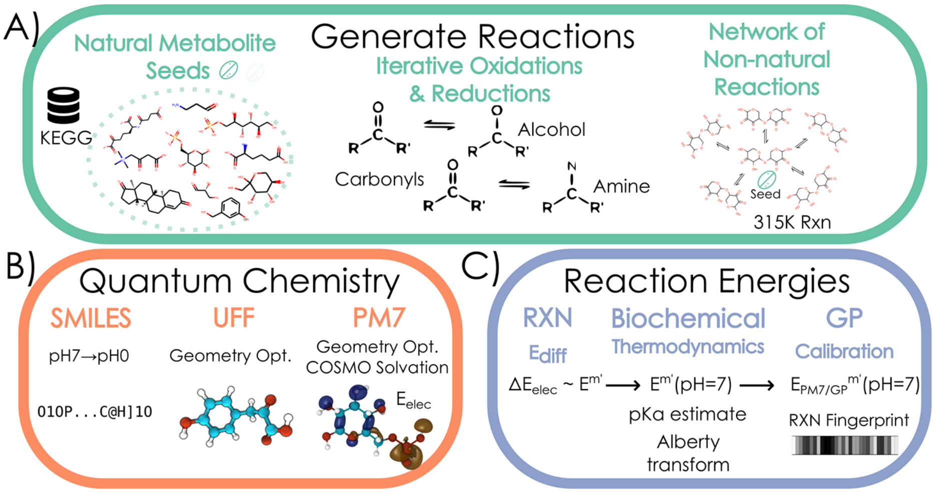


Figure 1. Schematic representation of the workflow. (A) We generate a network of redox reactions obtained from iterative reduction and oxidation of natural seed metabolites. The redox reactions considered are reductions of carbonyl functional groups to both alcohol and amine groups. (B) Starting from simplified molecular-input line-entry system (SMILES)[40] string representations of all the compounds in the network at pH = 0, we preoptimize molecular geometries using the Universal Force Field, and run semiempirical quantum chemistry calculations using PM7. (C) We estimate the standard (millimolar state) redox potentials E^{m} for the major species at pH = 0 from the difference in electronic energies. The transformed standard redox potentials E'^{m}(pH = 7) is obtained from cheminformatic pK_a estimates and the Alberty−Legendre transform. We calibrated a training set of compounds against available experimental E'^{m}(pH = 7) values using Gaussian process (GP) regression.

engineering applications[9,10] and the discovery of evolutionary design principles of natural pathways.[11] This applies both to the prediction of the thermodynamics of known metabolic reactions and to non-natural reactions, which can be used to expand nature's metabolic toolkit.[12,13]

Experimentally available reaction Gibbs energies and redox potentials provide coverage for only about ∼10% of known natural metabolic reactions. The metabolic modeling community relies mainly on group contribution method (GCM) approaches[14−18] to estimate missing thermodynamic values. GCM decomposes metabolites into functional groups and assigns group energies by calibration against experimental data. The most widely used implementation of the GCM for biochemistry is the eQuilibrator,[19,20] an online thermodynamics calculator. In addition to using GCM to predict the formation energies of compounds, the eQuilibrator makes use of experimental reactant formation energies when available and combines them with group energies in a consistent manner in what is known as the component contribution method. This results in significant increases in accuracy.[18] However, whenever experimental reactant formation energies are not available, estimates are based solely on group contribution energies. Such GCM-based estimates, which do not capture intramolecular functional group interactions, have limited prediction accuracy.[21,22]

In this work, we focus on predicting the thermodynamics of redox biochemistry. Redox reactions, which are fundamental to living systems and are ubiquitous throughout metabolism, consist of electron transfers between two or more redox pairs or half-reactions.[23] Quantum chemistry has recently emerged as an important alternative modeling tool for the accurate prediction of biochemical thermodynamics.[21,22,24,25] However, quantum chemical methods tend to have very high computational cost in comparison with the GCM or other cheminformatic-based alternatives.

Recent work in the intersection of quantum chemistry and machine learning has resulted in hybrid approaches that significantly lower computational cost without sacrificing prediction accuracy.[26−29] One such hybrid quantum chemistry/machine learning approach, previously applied to organic photovoltaics material design,[30,31] relies on Gaussian process (GP) regression[32] to calibrate quantum chemical predictions against experimental data. GP regression is an established probabilistic framework in machine learning to build flexible models, which also furnishes uncertainty bounds on predicted data points. Gaussian process regression uses the "kernel trick"[33] to make probabilistic predictions, leveraging the distance between a data point of interest and a training set.

Here, we present a mixed quantum chemistry/machine learning approach for the accurate and high-throughput prediction of biochemical redox potentials. We focus on predicting the thermodynamics of carbonyl (C═O) functional group reductions to alcohols (C—O) or amines (C—N) (Figure 1A), two of the most abundant redox reaction categories in metabolism, for which a significant amount of experimental data is available. The approach relies on predicting electronic energies of compounds using the semiempirical PM7 method (Figure 1B)[34] and then correcting for systematic errors in the predictions with GP regression against experimental data (Figures 1C and 2). We innovate on existing GP regression techniques by making use of novel reaction fingerprints[35] to compute the similarity between redox reactions. We demonstrate that the method results in higher predictive power than the GCM approach. Importantly, this comes at significantly lower computational cost than previously developed quantum chemical methodologies for redox biochemical thermochemistry.[22] In addition, we demonstrate the wide applicability of the GP regression method by directly calibrating GCM and fingerprint-based predictions (without quantum chemistry), also obtaining significant improvements in prediction accuracy. We demonstrate the high-throughput nature of the approach by predicting the redox potentials of what is to our knowledge the largest existing database of non-natural biochemical redox reactions, consisting of more than 315 000 iterative oxidations and reductions of "seed" metabolites from the KEGG database (Figure 1A).[36−39]

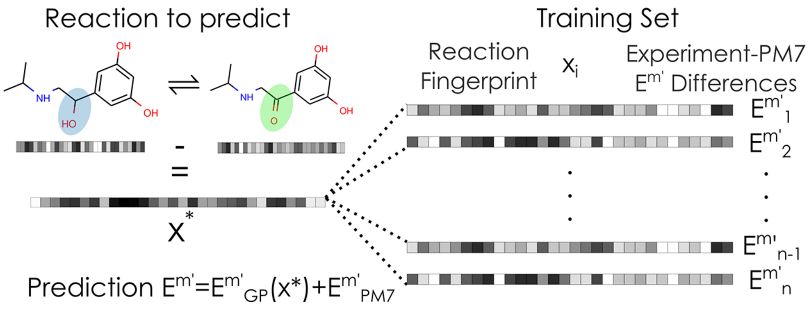


Figure 2. Schematic diagram of GP regression with reaction fingerprints. A redox reaction of interest x^* is represented using a reaction fingerprint, which captures the molecular structure of both substrate (oxidized) and product (reduced). This reaction fingerprint is compared against the set of all redox reactions in the experimental data set using the covariance function of the Gaussian process $k(x^*, x_i)$ (see the Methods section). The GP correction (E'^{m}_{GP}) to the quantum chemical prediction (E'^{m}_{PM7}) of the standard redox potential is then obtained from a similarity-based average of the correction terms for the experimental data set.

Table 1. Summary Statistics of Prediction Accuracy for Calibration with Gaussian Process Regression of Different Modeling Approaches[a]

			calibration/regression			linear fit
approach	geometry source	average compute time (s)	MAE (σ) (mV)	r	R^2	R^2
GCM	N/A	0.01	27.15 (24.04)	0.58	0.33	0.33
Rxn-FP + GP	N/A	0.01	24.35 (23.44)	0.65	0.42	N/A
GCM + GP	N/A	0.01	21.20 (20.50)	0.75	0.56	0.33
PM7 + GP	opt.	0.36	22.47 (19.68)	0.74	0.54	0.22
HF-3c + GP	opt.	588	21.07 (19.76)	0.76	0.57	0.28
DSD-PBEP86 + GP	PM7 opt.	1223	21.15 (17.97)	0.79	0.61	0.24
DLPNO−CCSD(T)//cc-pVDZ + GP	B3LYP conf.	6111	20.45 (19.97)	0.76	0.57	0.47

[a]See the Methods section for detailed descriptions of each modeling approach (quantum model chemistry). Rxn-FP = reaction fingerprints; GCM = group contribution method; GP = calibration with Gaussian process regression; opt. = geometry optimization of a single geometric conformer with the same level of theory; conf. = geometry optimization of multiple geometric conformers at a given theory level. Regression statistics correspond to GP calibration using Morgan fingerprints of radius 6. Additionally we report the R^2 (coefficient of determination) obtained using calibration with linear regression to the experimental data. The statistics are MAE = mean absolute error; σ = standard deviation of the absolute error; r = Pearson r; R^2 = coefficient of determination; average compute time = estimated average computational time per electron, in seconds. Linear fit is without GP regression.

Using this vast data resource, we apply a novel molecular fingerprint-based analysis—termed "reaction fingerprint heat-map"—to decipher the molecular environment motifs associated with high and low redox potentials, thus providing insights into important structure−energy relationships. We provide open access to the full source code and data, including metabolite geometries, electronic energies, pK_a's, and redox potentials at https://github.com/aspuru-guzik-group/gp_redox_rxn.

■ RESULTS

Optimization of Redox Potential Prediction Strategy That Combines Semiempirical Quantum Chemistry with GP Regression. Our goal is to develop a calibrated quantum chemistry modeling framework that can accurately predict the reduction potentials of biochemical redox reactions in a high-throughput manner. Several different types of redox reactions that change the oxidation state of carbon atoms exist in biochemistry. These include the reduction of carboxylic acids to aldehydes, the reduction of carbonyls to alcohols or amines, and the reduction of alcohols to hydrocarbons.[4] In this work, we focus our efforts on predicting the standard redox potentials (as a function of pH and ionic strength) for the reduction of carbonyls to alcohols or amines. This reaction category represents the most common type of carbon redox transformation in biochemistry.[39] In addition, it is the category for which the largest amount of experimental data is available in the NIST database for the thermodynamics of enzyme-catalyzed reactions (TECRDB).[40]

Similar to previous implementations, our approach consists of running quantum chemistry calculations to obtain electronic structure energies of substrates and products for the redox reactions of interest.[22] We take the difference in electronic energies as an estimate of the standard redox potential of the most abundant species (protonation states) of metabolites at pH = 0. We use cheminformatic pK_a estimates and the Alberty−Legendre transform[41] to convert the redox potentials to transformed standard redox potentials $E'°$ (pH = 7) (Methods). We then use GP regression to correct for systematic errors in the ab initio estimates (Figure 2) as well as pK_a estimates that go into the model (Methods). To test and optimize the accuracy of our modeling framework, we compared predicted potentials against a data set of 81 experimental redox potentials obtained from the NIST database of thermodynamics of enzyme-catalyzed reactions.[40]

We tested several different model chemistries to maximize the accuracy and minimize the computational cost of our methodology (Methods section, Table 1, Table S1). In this context, a "model chemistry" consists of the combination of multiple modeling choices that go into running a quantum chemical simulation. These include the conformer generation strategy and geometry optimization procedure, the electronic structure method (e.g., density functional theory (DFT), wave

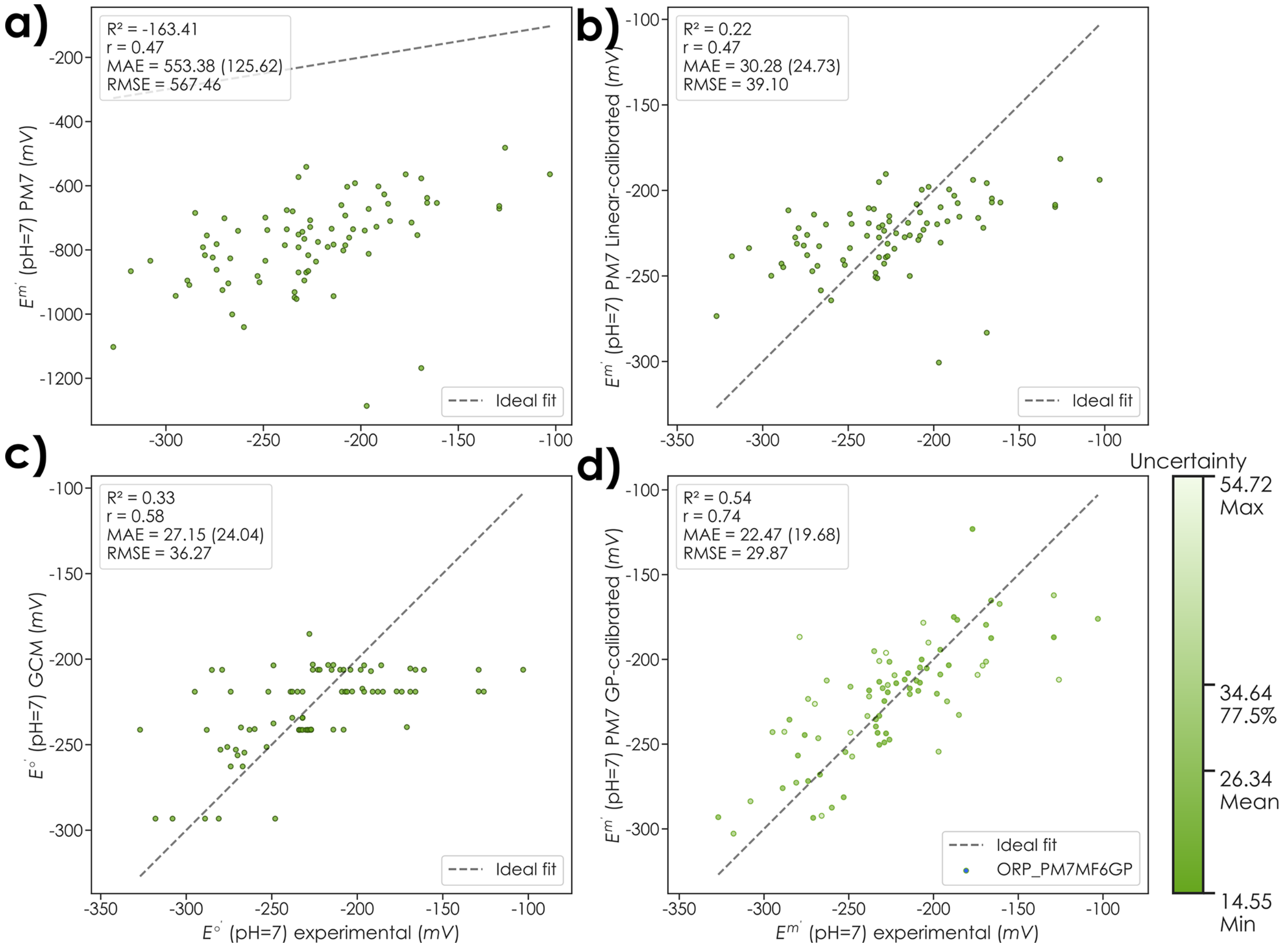


Figure 3. Comparison of prediction accuracy for the group contribution method and quantum chemistry with GP regression. Scatter plots of experimental standard transformed redox potentials against predictions from several sources: (a) PM7, (b) linearly calibrated PM7, (c) group contribution method (GCM), and (d) GP-calibrated PM7. The experimental data consist of available standard transformed reduction potentials for carbonyl-to-alcohol and carbonyl-to-amine reductions. Data points in the scatter plot were obtained from a leave-one-out cross-validation (LOOCV) procedure, where the model is trained on all data points except the one point to be predicted. Note that we use the millimolar standard state, E'^{m}, where reactant concentrations are taken as 1 mM, since this is significantly closer to relevant physiological concentrations of metabolites in cells than 1 M. The legend in each plot contains summary statistics for the fit; MAE (mean absolute error) also has the associated standard deviation in parentheses.

function approaches, or semiempirical methods) as well as the water model used.

The GP regression-based calibration step depends on a particular choice of a covariance function, which in turn involves a distance or similarity measure between redox reactions or compounds. We found that the following combination of quantum model chemistry and kernel/distance function between reactions resulted in the most efficient (in terms of a trade-off between accuracy and computational cost) prediction strategy: the PM7[34] semiempirical method for both geometry optimizations and single-point electronic energies, with the COSMO implicit solvation model,[42] and a reaction fingerprint obtained by taking the difference between counted Morgan fingerprint vectors of products and substrates,[35] with a kernel function that is a mixture of a squared exponential kernel, and a noise kernel (Methods). We note that the model chemistry using the double-hybrid functional DSD-PBEP86 (based on a double-hybrid DFT functional) gives the best accuracy but at a higher computational cost.

GP-Calibrated Quantum Chemistry Yields High Prediction Accuracy at a Low Computational Cost. Importantly, our GP-calibrated semiempirical quantum chemistry method results in higher predictive power than the most commonly used approach, the GCM (Table 1, Figure 3). Using cross-validation to test prediction accuracy, the GP-calibrated PM7 approach results in a higher Pearson correlation coefficient (r), a higher coefficient of determination (R^2), and a lower mean absolute error (MAE) than GCM predictions when tested against the set of experimental potentials ($p < 0.05$, Methods). The GP calibration also returns a GP uncertainty for each predicted value, and we find that it linearly correlates with the absolute deviation from uncertainty (Figure S1).

We note that the GCM predictions, as well as fingerprint-based predictions (without relying on any quantum chemistry calculations), can also be calibrated using GP regression, resulting in significant improvements in their prediction accuracy. Thus, regardless of the underlying modeling choice, the GP calibration approach is an accurate, computationally

DOI: 10.1021/acscentsci.9b00297

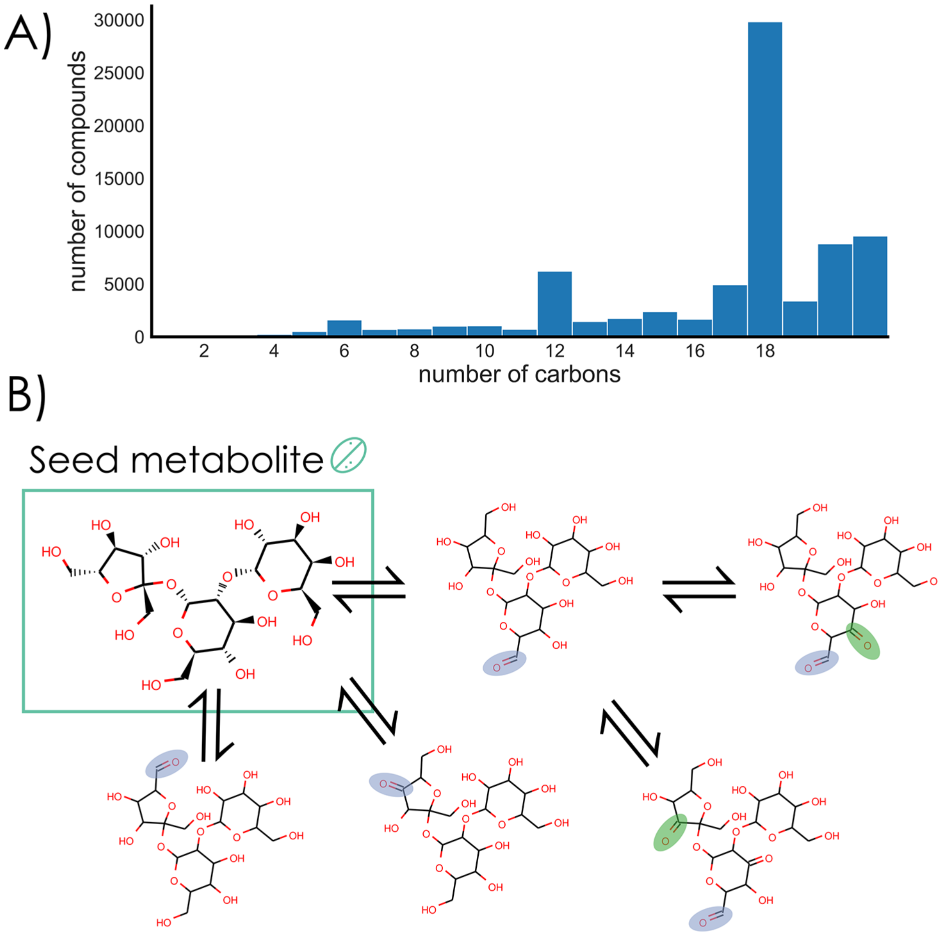


Figure 4. The network of redox reactions obtained from iterative reductions and oxidations of natural metabolites. (A) Carbon number distribution for the full set of ~70 000 compounds in the network. The peaks at multiples of 6 carbon atoms reflect the combinatorial oxidation of mono-, di-, and trisaccharides and their multiple alcohol functional groups. (B) Schematic diagram of a natural trisaccharide being iteratively oxidized at all possible alcohol functional groups. Blue circles depict sites of initial alcohol-to-carbonyl oxidation, and green circles depict a second oxidation site.

efficient, and widely applicable modeling framework for the high-throughput prediction of biochemical redox potentials.

One of the advantages of our approach is its low computational cost (Table S1) in comparison to other model chemistries. For instance, the PM7-GP-calibrated calculations are approximately 1000-fold faster than the DLPNO−CCSD(T) calibrated model chemistry, despite both approaches yielding comparable accuracies. We note that the pipeline involves estimating pK_a's and major protonation states at pH = 0 for all metabolites at a slight additional computational cost. However, this cost is also incurred by the GCM approach.

Generation of a Large Network of Non-Natural Redox Reactions Using a Novel Network Expansion Algorithm. To demonstrate the high-throughput nature of our method, we implemented a redox reaction network expansion algorithm to generate what is, to our knowledge, the largest database of non-natural biochemical redox reactions (Methods). The network expansion algorithm makes use of the RDKit cheminformatics software to iteratively apply simplified molecular-input line-entry system (SMILES)[35,43] reaction strings to natural metabolites in the KEGG database. We generate two types of redox reactions, implemented in both the reductive and oxidative directions. The first type is the reduction of carbonyls to alcohols (and the reverse oxidations); the second is the reduction of carbonyls to amines (and the reverse oxidations). Applying the cheminformatic transformation to metabolites in the KEGG database results in a first set of nearest-neighbor reduction or oxidation products. We then iteratively apply the reactions to the resulting set of products, until there are no more carbonyl functional groups to reduce (or alcohol or amine functional groups to oxidize).

Although the KEGG database of natural metabolic compounds and reactions contains molecules with up to 135 carbon atoms (a peptidoglycan cell wall component is the molecule with the highest number of carbon atoms),[44] we limit our network expansion algorithm to molecules with 21 carbon atoms or less. This threshold is imposed in consideration of the size of the compounds in the training set; if experimental data were available for larger compounds, this methodology could easily be adapted.

Table 1 shows the number and type of reactions that make up the network. It consists of ~75 000 compounds connected by ~315 000 reactions (reductions and oxidations). Of these reactions, 83% convert carbonyls to alcohols (or alcohols to carbonyls), while 17% convert carbonyls to amines (or amines to carbonyls). The large majority of the reactions (more than 80%) are oxidations of alcohols to carbonyls. Thus, the alcohol functional group is significantly more ubiquitous in the set of natural metabolites than either carbonyls or amine functional groups. The large majority of the reactions in the network—

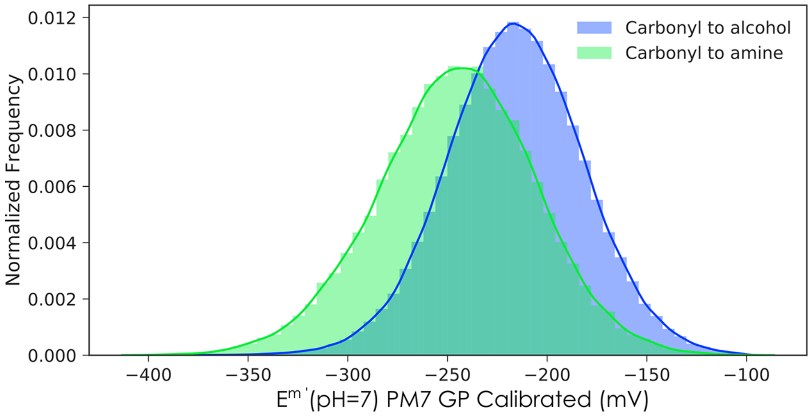


Figure 5. Distribution of predicted standard redox potentials (pH = 7) for a set of ~315 000 non-natural redox reactions. Predicted potentials were obtained by running semiempirical quantum chemical calculations (Methods) and correcting for systematic errors using Gaussian process regression. The green distribution shows the E'^{m}(pH = 7) values for 53 095 reductions of carbonyls to amines. The blue distribution shows the E'^{m}(pH = 7) values for 261 903 reductions of carbonyls to alcohols.

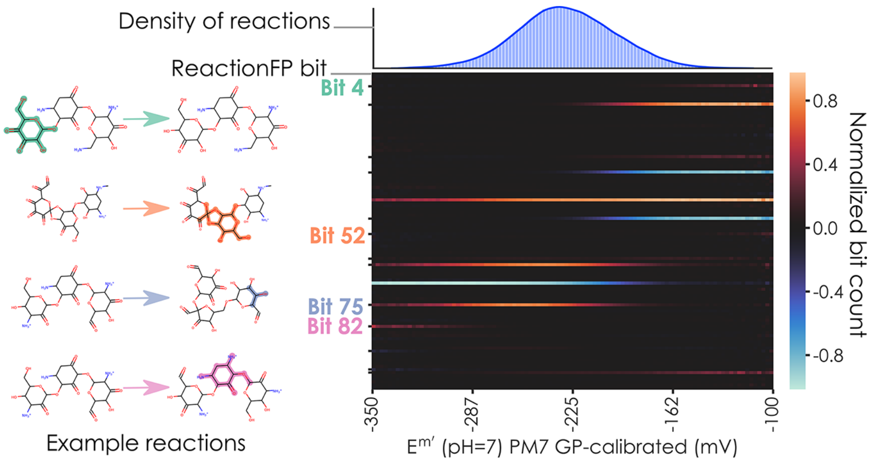


Figure 6. Elucidating structure−energy relationships using a reaction fingerprint heatmap. The top right panel shows the distribution of predicted standard redox potentials (in the millimolar standard state) using our GP-calibrated quantum chemical approach. The bottom right panel shows a reaction fingerprint heatmap, where the *y*-axis corresponds to the bits in the reaction fingerprint vector averaged across all reactions falling within a redox potential bin (range of values). The heatmap highlights bits that have significantly different values (on average) as a function of redox potential. The example reactions on the left correspond to each of the significant reaction fingerprint bits. For example, Bit-4 is a structure enriched in substrate molecules of reactions with high values of standard redox potentials.

97%—stem from the recursive *oxidations* of seed natural metabolites, while only ~3% come from recursive reductions. This is consistent with the notion that carbonyl functional groups can cause damage to macromolecules[45,46] and are thus kept at check in the cell.

We analyzed the structure of the resulting network by looking at the distribution of compound sizes. Figure 4A shows the number of compounds in the networks as a function of their number of carbon atoms. The distribution shows three peaks corresponding to molecules with 6*n* carbon atoms (*n* = 1, 2, 3). These peaks reflect a large number of products that result from combinatorially oxidizing all alcohol groups in mono-, di-, and trisaccharides (Figure 4B).

High-Throughput Prediction of Redox Potentials and Elucidation of Structure−Energy Relationships. Using our GP-calibrated semiempirical quantum chemistry method, we predicted the standard transformed redox potentials for the set of ~315 000 reactions generated with the iterative network expansion algorithm. The resulting distribution of standard transformed redox potentials is shown in Figure 5. We note the use of the millimolar standard state, E'^{m}, instead of the more commonly used E'° (where reactant concentrations are taken as 1 M). This is a standard state that is commonly used in the field of biochemical thermodynamics since 1 mM is closer to the relevant physiological concentrations of metabolites in cells.[4] The resulting distributions for the two reaction categories peak at very close but slightly different values, with $\langle E'^{m} \rangle \sim -220$ mV for carbonyl-to-alcohol reductions and $\langle E'^{m} \rangle \sim -235$ mV for carbonyl-to-amine reductions.

Importantly, our vast network of non-natural redox reactions and their associated redox potentials are amenable to structure−energy analyses to elucidate molecular structures that are enriched or depleted in reactions with high or low potentials. Toward this end, we developed a novel molecular fingerprint-based structure−energy analysis, which we term "reaction fingerprint heatmaps", to detect such reaction motifs (Figure 6). The main idea behind reaction fingerprint heatmaps is that, by comparing the *average* reaction fingerprints of reactions with different redox potential values, one can visually detect the structural patterns that correlate with energetics. More specifically, we first divide the distribution of predicted standard redox potentials into arbitrary discrete bins, in our case 100 bins. We then find, for each standard redox potential bin, the subset of reactions with potentials falling within that range and average their corresponding fingerprint vectors. We then extract structure−energy trends by looking at the changes in average fingerprints values across consecutive bins. When these changes are large for a particular bit, it signifies that a meaningful molecular environment (motif) is enriched or depleted at that particular bin (range) of redox potential values. For any particular reaction, we can reverse engineer the reaction fingerprint to obtain substructures that activate the particular bit. For example (Figure 6), we find that Bit-4 is enriched in reactions with standard potentials of −160 mV < E'^{m} < −100 mV. By mapping Bit-4 back to specific reactions, we find a molecular environment that corresponds to a dicarbonyl within a heterocyclic ring (see example reaction in Figure 6) that is enriched in substrate molecules within that range of standard potentials. While, across all reactions, the difference between the substrate and product molecules is the reduction of a single carbonyl into a hydroxy carbon group, our reaction fingerprint heatmaps detect the molecular environments dictating the energetics associated with that transformation.

■ DISCUSSION

In this work, we developed a mixed quantum chemistry/machine learning modeling approach for the prediction of biochemical redox potentials of carbonyl functional groups. The method is based on calculating the electronic energy difference between substrates and products in redox reactions using the semiempirical quantum chemistry method PM7.[34] The raw quantum chemistry estimates are then calibrated against experiment using GP regression.

We demonstrated that the method has better prediction power than the commonly used GCM approach. Furthermore, the computational resources required are significantly lower than previous quantum chemistry strategies for biochemical thermodynamics. The GCM decomposes compounds into discrete functional groups and assigns energies to each group by calibrating via linear regression against experimentally available Gibbs energies. It then estimates reaction energies as the group energy difference between products and substrates. As Figure 3A shows, GCM collapses a large fraction of the redox potentials of carbonyl compounds into a few sets of values. We also note that, in a recent variant of the GCM, group energies can be combined with experimental reactant Gibbs formation energies to significantly increase the accuracy of predictions, resulting in what is known as the CCM.[18] Since we seek to apply our modeling approach to predicting redox potentials of non-natural reactions, the CCM is not applicable to the vast majority of the reactions in our data set. However, we also found that the GP regression technique can be used to further calibrate both GCM and fingerprint-based predictions (without relying on quantum chemical calculations), resulting in significant improvements to the prediction accuracies of these two approaches. Thus, the decision on what modeling choice to use can be based on a balancing act between accuracy, computational cost, and the preferred underlying modeling baseline. For example, if one desires an accurate, computationally cheap approach that relies on an underlying electronic structure method, PM7+GP is an optimal choice. Alternatively, for more accurate results that rely on more expensive electronic structure theory, DSDPEP89 (based on a double-hybrid DFT functional) gives the best accuracy but at higher computational cost. Likewise, if a good baseline for your chemical application is GCM, at a very low cost and almost comparable accuracy GCM+GP will be a good modeling choice. GP calibration proves to be a versatile approach to improve the accuracy of a wide variety of prediction methods.

We considered model chemistries that rely on a single conformation per molecule and also ensembles of conformers. In our case we did not observe a huge difference in predicted accuracy between both approaches. Other chemical applications and data sets should be tested to see if performance is impacted by conformer diversity. A single missed hydrogen bond in one reacting partner, but not the other, could easily introduce large errors in the predicted redox potential.

Current work on GP calibration in computational chemistry relies on *molecular* fingerprints to compute the similarity between molecules of interest and compounds with experimental data. Our implementation is to our knowledge the first application of GP to reaction energy predictions using reaction fingerprints. We chose to use GP—as opposed to other machine learning techniques such as neural networks—on the basis of the amount of experimental data available. GPs are suited for this task since they grow in complexity according to the data; a GP is robust to overfitting since it penalizes complex models and is able to provide estimates of uncertainty on predictions.[32,47,48] In contrast, techniques such as deep neural networks—which have gained traction in recent machine learning applications to quantum chemistry[49−51]—are only applicable when much larger data sets are available.

We focused our approach to the prediction of the standard redox potentials of carbonyl functional groups (i.e., reduction to alcohols or amines). The application of GP regression to other redox transformations—including the reduction of carboxylic acids to aldehydes and the reduction of alcohols to hydrocarbons—is limited by the availability of experimental data. This highlights the importance of a concerted effort to generate more experimental data for biochemical redox potentials in controlled conditions (such as pH, ionic strength, temperature, and buffer) to increase the predictive power and scope of calibrated quantum chemical approaches.

We demonstrated the high-throughput nature of the methodology by generating a network of more than 315 000 non-natural biochemical reactions involving ∼70 000 compounds. Our algorithm is different from other network expansion algorithms previously applied to metabolic networks. By solely focusing on redox biochemistry, iterative application of the reactions converges to molecules that cannot be further reduced or oxidized. This contrasts with other network expansion algorithms[52,53] used to investigate scenarios related to the origins of life, where several types of chemical transformations are considered, including carbon-bond formation and cleavage. Other network expansion algorithms also start from natural seed compounds but only consider natural metabolites.[12] In contrast, our algorithm uses natural metabolites as seeds but results in the expansion to a vast network of non-natural compounds.

Despite the improvements obtained over GCM with our method, the variance in the prediction accuracies across individual reactions is still significant. In practice, this spread places a limit on the confidence with which one can predict the redox potential of an individual reaction of interest. This is true, and to a larger degree, for other methods currently widely used, such as GCM. However, the main value of developing a method with lower average prediction error lies in our ability to analyze large sets of compounds, where average trends in the energies of different molecular substructures can be captured. We thus developed a novel structure−energy relationship analysis framework which we term reaction fingerprint heatmaps. These allow us to detect molecular environments that are enriched or depleted in reactions that fall within specific regions (bins) of the redox potential distribution. This could potentially be useful for future metabolic engineering applications: being able to pinpoint the exact molecular environments that correlate with redox potentials would allow the fine-tuning of the energetics of synthetic metabolites. We note that this approach captures large-scale structure−energy trends. An alternative approach, which is able to associate weights to individual atoms or subfragments, are the similarity maps as introduced by Riniker et al.[54]

We envision several applications of our biochemical redox potential prediction methodology. One such application is studying the thermodynamic landscape of specific families of natural metabolic compounds that undergo combinatorial reductions (or oxidations) of carbonyl (or alcohol or amine) functional groups. One such family of compounds are the brassinosteroids and the oxylipins,[55−57] structurally diverse plant metabolites that play important roles in many physiological processes. In addition, in the context of drug metabolism, our methodology could be applied to obtain quantitative insights into the thermodynamics of redox transformations such as those mediated by the P450 (CYP) superfamily of enzymes.[58]

To our knowledge, our database contains the largest set of geometric structures, electronic energies, pK_a estimates, and redox potentials of natural metabolites and compounds related to these through oxidoreductive transformations. We make all the code and data sets generated in this work—including metabolite 3D geometries, electronic energies, pK_a's, and redox potential estimates—available as an open source repository at https://github.com/aspuru-guzik-group/gp_redox_rxn.

■ METHODS

Generation of Most Abundant (Major) Protonation States at pH = 0. Our pipeline is based on performing electronic structure simulations of the most protonated species of each biochemical compound involved in every redox reaction of interest. We run the quantum chemistry simulations on the major protonation states at pH = 0 to avoid errors associated with the prediction of anionic compound energies.[21,59,60] We use the ChemAxon calculator plugin (Marvin 17.7.0, 2017, ChemAxon) *cxcalc majormicrospecies* to generate, for each substrate and product involved in a biochemical redox reaction of interest, the major microspecies at pH = 0.

Geometry Optimization. Using the SMILES string representation[43] for each compound in the reaction network, we obtain the string corresponding to the most abundant (major) microspecies at pH = 0 using ChemAxon. This smiles string is then converted to an RDKit "mol" object. From here we obtain 3D structures based on two decisions. (1) Whenever we want a single 3D geometry, we generate an initial conformer with RDKit (using the EmbedMolecule command, which uses distance geometry to obtain initial coordinates for a molecule), and we optimize this conformer's geometry Universal Force Field (UFF) as implemented in RDKit.[61] (2) If we want several conformers (to Boltzmann average their energies), we generate around 1000 random "smart" conformers via the ETKDG force field,[62] calculate their energies, and pick between 5 and 20 conformers (based on size) and on clustering of energies (using the Butina clustering algorithm). Depending on the baseline model chemistry, either we further refine the geometry of the PM7[34] semiempirical method with the COSMO implicit solvation model[42] and then a quantum chemistry approach (PBE-3H or B3LYP), or we simply use the base PM7 or UFF geometry and perform single-point energy (SPE) calculations.

Single-Point Energy Estimates. We then compute the electronic energy $E_{\text{electronic}}$ of every compound involved in a given redox reaction using the PM7 semiempirical method[34] with the COSMO implicit solvation model[42] (see below for other single-point energy SPE model chemistries tested). For both reductions of carbonyl groups to alcohols and amines, we add a hydrogen molecule to the substrate side of the redox reaction (in the direction of reduction) and compute its electronic energy. In addition, to balance reductions of carbonyls to amine functional groups, we add an ammonia molecule to the substrate side of the reaction and a water molecule to the product side of the reaction (in the direction of reduction) and compute their electronic energies.

We use the difference in the electronic energies of products and substrates, $\Delta E_{\text{electronic}}$, as an estimate of the standard redox potential $E°$(major species, MS, at pH = 0):

$$E°(\text{MS at pH} = 0) \sim -\Delta E_{\text{electronic}}/nF$$

where n is the number of electrons (2 for all reactions considered here), and F is Faraday's constant.

Our approach ignores the contribution of ro−vibrational enthalpies and entropies to Gibbs reaction energies. This significantly reduces the computational cost associated with quantum chemical simulations, and we correct for the systematic errors introduced by this approximation through the Gaussian process regression (see below). We note that previous work has shown that the value of $\Delta E_{\text{electronic}}$ obtained for these types of redox reactions strongly correlates linearly with the full $\Delta G_r°$ prediction obtained from including ro−vibrational enthalpic and entropic contribution (data not shown).

Transforming Chemical Potentials $E°$(Major Species at pH = 0) to Biochemical Potentials, $E'°$(pH = 7). Having estimated the standard redox potential $E°$(major species at pH = 0), we use cheminformatic pK_a estimates of reactants and the Alberty−Legendre transform[41] (p 67, eqs 4.4−10) to convert $E°$(major species at pH = 0) to the *transformed* standard redox potential, $E'°$(pH = 7), which is a function of pH.[63] To estimate the pK_a's of all reactants in every redox reaction of interest, we use the ChemAxon calculator plugin (Marvin 17.7.0, 2017, ChemAxon) *cxcalc pka*. Internally, the pK_a calculator plugin is based on the calculation of the partial charge of atoms in the molecule.[64,65]

Gaussian Process Calibration Using Reaction Fingerprints. We calibrate biochemical redox potentials $E'°$(pH = 7) obtained from quantum chemical simulations against available experimental data to correct for systematic errors in our simulations and the cheminformatic pK_a estimates. One simple strategy of calibration of energy values would be to use a two-parameter linear regression. We note that our group has successfully employed this in the context of compounds for redox flow battery applications.[66−71] Here, we make use of the information provided by the difference between the structural similarity of products and substrates and utilize a GP regression approach with reaction fingerprints to calibrate redox potential energies. For our calibrations with Gaussian process regression we used GPMol, a Python library based on GPflow,[72] which in turn is a package for building Gaussian process models using TensorFlow.

GP regression relies on a similarity or distance metric between data points. To construct the notion of similarity between reactions, we utilize reaction fingerprints.[35] Several possible choices of kernel function and distance measure between molecules and reactions exist. Generally, the distance measures between molecules and reactions make use of fingerprint representations of compounds, which encode the structure of a molecule in a binary vector form.[73] Although several varieties of reaction fingerprints exist, we tested our approach on the Morgan2 and AP3 (atom pair) reaction FPs of Schneider et al.[35] and obtain them from the difference of fingerprint vectors of products and substrates. We find that in the context of Gaussian processes for biochemical redox potential predictions Morgan fingerprints outperform AP3 fingerprints by approximately 5−4 mV. The Gaussian processes' predictive power comes from learning relationships between relative distances in feature space, so it might be the case that for distance-based ML algorithms Morgan fingerprints are better. However, we note that we did not incorporate the usage of any agent to the weighting scheme, so doing so might change the results. Additional performance metrics for comparison can be seen in Table S1.

One key ability of Morgan fingerprints is that we can invert the bits back to a corresponding molecular substructure to perform our structure−energy analysis. For each molecule, we generate a counted Morgan fingerprint,[73] a fixed-length counted vector indicating the absence (zeros) or counted presence (positive) of a particular graph−connectivity−environment. Each environment captures the local topological information on a molecule by mapping the local vicinity of connected atoms, along with their formal charges, type of chemical bond, and its position relative to a cyclic structure. Our Morgan fingerprints are constructed from 2048 length-vectors using a radius of 6 and are reduced to a lower dimension (~200) using a gradient boosted tree. We find that the prediction accuracy did not improve significantly upon using Morgan fingerprints with a larger radius. When we consider the difference of two Morgan fingerprints, we are looking at the subtraction and addition of molecular environments between two molecules. We expect that similar molecular transformations will have similar changes in molecular environments. When these vectors are normalized, the inner product between two reaction fingerprints x_1 and x_2 is the cosine distance,[74] which is a measure of similarity between two vectors. This notion is adapted in the field of Gaussian processes to quantify the similarity between reactions in GP regression.

Predictions obtained from GPs can be interpreted as weighted averages of the training data, where the weights are probabilistic in nature. GPs are distinct because of their associated covariance functions (i.e., the kernel).[33] The equation below shows that the kernel we employ is a mixture of a squared exponential kernel and a noise kernel, which increases the robustness of the model.

$$K(x^*, x_i) = \sigma_{\text{rxn}} e^{-|x^* - x_i|^2/l} + \sigma_{\text{noise}} I$$

Here, K is the kernel function, σ_{rxn} is the variance of the reaction fingerprint kernel, l is the length scale parameter, and σ_{noise} is the white noise variance parameter.

Database of Experimental Redox Potentials for Calibration with Gaussian Process Regression. To perform the calibration, a data set of available experimental transformed standard redox potentials is used. This data set of redox half-reactions and their associated transformed standard reduction potentials $E'^{\circ}(\text{pH} = 7)$ was generated by Bar-Even et al.,[4] by compiling experimental equilibrium constants (i.e., standard Gibbs reaction energies) from the NIST database of Thermodynamics of Enzyme-Catalyzed Reactions (TECRDB)[40] and the Robert Alberty database of biochemical compound Gibbs formation energies.[41] The data set consists of 57 redox reactions that reduce a carbonyl functional group into an alcohol functional group and 24 redox reactions that reduce a carbonyl functional group into an amine group.

To ensure we are not overfitting in the process of calibrating with GP regression, we used leave-one-out cross-validation (LOOCV), where we train our model on all data points except one point and predict its value. This is done for all 81 data points, and reported predictions, accuracies, and resulting scatter plots always come from untrained data. It is also important to note that GPs are inherently robust to overfitting since the training procedure penalizes more complex models (higher-rank kernels) via an objective function.[32]

Selecting a Model Chemistry with Low Computational Cost and High Prediction Accuracy. To obtain fast and accurate predictions of redox potentials using the GP regression calibrated quantum chemistry strategy described above, we tested several different quantum model chemistries. A model chemistry consists of a combination of a geometry optimization (GO) procedure, a method to calculate the single-point electronic energies (SPEs) of optimized molecular geometries, as well as other modeling considerations, such as the number of geometric conformations per compound, the water solvation model, and the pK_a estimation strategy.

We explored the following set of methods to obtain optimized geometric conformations: the Universal Force Field (UFF),[61] PM7,[34] DFTB,[75] HF3-c,[76] and PBEh-3c.[77] Additionally, to compute the electronic energies of the optimized structures through single-point energy (SPE) calculations, we considered the following approaches: DFTB, PM7, HF3-c, PBEh-3c, DSD-PBEP86/SVP,[78] and DLPNO−CCD(T)/SVP.[79] Given computational resource constraints, only a subset of all possible combinations of geometry optimization and SPE procedures were explored.

To select a model chemistry from the exploration set, we compared their prediction accuracies—after the GP regression calibration step described above—when tested on the set of experimental redox potentials. We quantified prediction accuracy using three different metrics: Pearson correlation coefficient (r), the coefficient of determination (R^2), and mean absolute error (MAE, in mV). Before applying the GP regression calibration step, the double-hybrid functional and linear-scaling couple-clustered methods DSD-PBEP86 and DLPNO−CCD(T) resulted in the highest accuracy. Due to the relatively small variation in prediction accuracy, we picked the cheapest method that still gave reasonable accuracy (MAE < 30 meV), which turned out to be the PM7 semiempirical method for both geometry optimizations and single-point electronic energies. It should be noted that our geometry is a single, UFF optimized conformation per molecule. For PBEh-

3c and HF-3c we also considered utilizing multiple conformers obtained with the ETKDG force field[62] with significant improvement in accuracy over a single conformer (see Table S1). For other chemical applications (e.g., redox flow batteries) and molecular data sets, adequately accounting for multiple conformers might have an important influence on the accuracy of the calculations. A single missed hydrogen bond in one reacting partner, but not the other, could easily introduce large errors in the predicted redox potential. As such the adequacy of using just a single geometry should be verified.

To test for statistically significant differences of the prediction accuracy summary statistics of different model chemistries (e.g., GCM versus GP-calibrated PM7), we performed a nonparametric bootstrap hypothesis test: we subsampled data points from the table of redox reactions with experimental data and computed the prediction accuracy summary statistics [Pearson r, R^2, MAE, root-mean-square error (RMSE)] for each subsample. We then count the fraction of subsampled GCM data sets that result in summary statistics that are greater than or equal in value to the summary statistics of the model chemistry we are comparing against (e.g., GP-calibrated PM7).

Group Contribution Method (GCM) Estimates of Redox Potentials. We compared the accuracies obtained using the GP regression calibrated quantum chemistry approach described above against those obtained with the commonly used GCM.[14,18,80] We use the GCM implementation of Noor et al. which was adapted to the thermodynamics of biochemical reactions.[18] Briefly, group energies of all compounds in the KEGG database are stored in a group matrix, with rows corresponding to compounds and columns corresponding to groups. Associated to each group is a group energy corresponding to the energy for the group's major protonation state (major species) at pH = 7. To obtain a GCM-based estimate for a standard transformed reduction potential for the major species at pH = 7, we take the difference in group energy vectors for all products and substrate in the reaction. We then transform the resulting standard redox potentials $E°$(major species at pH = 7) to the standard *transformed* redox potential at pH = 7, $E'°$(pH = 7), using the same pK_a estimates and Alberty transform approach described above.

Redox Network Expansion Algorithm with Cheminformatic Reaction Strings. We implemented a network expansion algorithm to iteratively reduce carbonyl functional groups and oxidized alcohol/amine functional groups for a subset of the natural metabolites of the KEGG database. Due to constraints in available computational resources to run electronic structure calculations, we used the set of all metabolic compounds in the KEGG database with 21 carbon atoms or fewer as seed metabolites. Using the RDKit open source cheminformatics software, we identify carbonyl functional groups (alcohol/amine functional groups) in each of the seed KEGG metabolites and transform them through a redox reaction to the corresponding reduced (oxidized) alcohol/amine (carbonyl). In practice, this is done using SMARTS reaction strings as implemented in RDKit and iteratively applying them to SMILES representations[43] of the KEGG metabolites. The SMARTS reaction strings for the reduction and oxidation reactions considered here are shown in Tables S2 and S3.

We iteratively apply the reduction/oxidation reaction transformations to the compounds generated from KEGG seed metabolites. We terminate the iterations when the resulting product has, in the case of reductions, no more carbonyl functional groups that can be reduced (or in the case of oxidative transformations, no more alcohol/amine functional groups that can be oxidized to carbonyls). All software was written using the Python programming language.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.9b00297.

Additional data, including model chemistry prediction accuracies and cheminformatic SMARTS reaction strings used to generate redox network (PDF)

AUTHOR INFORMATION

Corresponding Author

*Phone: 416-978-3564. E-mail: alan@aspuru.com.

ORCID

Alán Aspuru-Guzik: 0000-0002-8277-4434

Author Contributions

#A.J. and B.S.-L. contributed equally to this work

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank members of the Aspuru-Guzik lab for comments on the manuscript. The authors thank Harvard Research Computing for their support on using the Odyssey cluster. A.A.-G., A.J., and B.S.-L. acknowledge support from SEAS^NVIDIA, Massively Parallel Programming and Computing (332986). A.A.-G. acknowledges the generous support of Anders G. Froseth for this program and the Canada 150 Research Chairs Program.

REFERENCES

(1) Nelson, D. L.; Lehninger, A. L.; Cox, M. M. *Lehninger Principles of Biochemistry*; Macmillan, 2008.

(2) Noor, E.; Bar-Even, A.; Flamholz, A.; Reznik, E.; Liebermeister, W.; Milo, R. Pathway Thermodynamics Highlights Kinetic Obstacles in Central Metabolism. *PLoS Comput. Biol.* **2014**, *10* (2), No. e1003483.

(3) Henry, C. S.; Broadbelt, L. J.; Hatzimanikatis, V. Thermodynamics-Based Metabolic Flux Analysis. *Biophys. J.* **2007**, *92* (5), 1792−1805.

(4) Bar-Even, A.; Flamholz, A.; Noor, E.; Milo, R. Thermodynamic Constraints Shape the Structure of Carbon Fixation Pathways. *Biochim. Biophys. Acta, Bioenerg.* **2012**, *1817* (9), 1646−1659.

(5) Krumholz, E. W.; Libourel, I. G. L. Thermodynamic Constraints Improve Metabolic Networks. *Biophys. J.* **2017**, *113* (3), 679−689.

(6) Krebs, H. A.; Kornberg, H. L.; Burton, K. A Survey of the Energy Transformations in Living Matter. *Rev. Physiol., Biochem. Pharmacol.* **1957**, *49*, 212−298.

(7) Alberty, R. A. A Short History of the Thermodynamics of Enzyme-Catalyzed Reactions. *J. Biol. Chem.* **2004**, *279* (27), 27831−27836.

(8) Burton, K.; Krebs, H. A. The Free-Energy Changes Associated with the Individual Steps of the Tricarboxylic Acid Cycle, Glycolysis and Alcoholic Fermentation and with the Hydrolysis of the Pyrophosphate Groups of Adenosinetriphosphate. *Biochem. J.* **1953**, *54* (1), 94−107.

(9) Kiparissides, A.; Hatzimanikatis, V. Thermodynamics-Based Metabolite Sensitivity Analysis in Metabolic Networks. *Metab. Eng.* **2017**, *39*, 117−127.

(10) Bar-Even, A.; Noor, E.; Lewis, N. E.; Milo, R. Design and Analysis of Synthetic Carbon Fixation Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (19), 8889−8894.

(11) Flamholz, A.; Noor, E.; Bar-Even, A.; Liebermeister, W.; Milo, R. Glycolytic Strategy as a Tradeoff between Energy Yield and Protein Cost. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (24), 10039−10044.

(12) Hadadi, N.; Hafner, J.; Shajkofci, A.; Zisaki, A.; Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **2016**, *5* (10), 1155−1166.

(13) Siegel, J. B.; Smith, A. L.; Poust, S.; Wargacki, A. J.; Bar-Even, A.; Louw, C.; Shen, B. W.; Eiben, C. B.; Tran, H. M.; Noor, E.; et al. Computational Protein Design Enables a Novel One-Carbon Assimilation Pathway. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (12), 3704−3709.

(14) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **1958**, *29* (3), 546−572.

(15) Mavrovouniotis, M. L. Group Contributions for Estimating Standard Gibbs Energies of Formation of Biochemical Compounds in Aqueous Solution. *Biotechnol. Bioeng.* **1990**, *36* (10), 1070−1082.

(16) Mavrovouniotis, M. L. Estimation of Standard Gibbs Energy Changes of Biotransformations. *J. Biol. Chem.* **1991**, *266* (22), 14440−14445.

(17) Jankowski, M. D.; Henry, C. S.; Broadbelt, L. J.; Hatzimanikatis, V. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophys. J.* **2008**, *95* (3), 1487−1499.

(18) Noor, E.; Haraldsdóttir, H. S.; Milo, R.; Fleming, R. M. T. Consistent Estimation of Gibbs Energy Using Component Contributions. *PLoS Comput. Biol.* **2013**, *9* (7), No. e1003098.

(19) Noor, E.; Bar-Even, A.; Flamholz, A.; Lubling, Y.; Davidi, D.; Milo, R. An Integrated Open Framework for Thermodynamics of Reactions That Combines Accuracy and Coverage. *Bioinformatics* **2012**, *28* (15), 2037−2044.

(20) Flamholz, A.; Noor, E.; Bar-Even, A.; Milo, R. eQuilibrator–the Biochemical Thermodynamics Calculator. *Nucleic Acids Res.* **2012**, *40* (D1), D770−D775.

(21) Jinich, A.; Rappoport, D.; Dunn, I.; Sanchez-Lengeling, B.; Olivares-Amaya, R.; Noor, E.; Even, A. B.; Aspuru-Guzik, A. Quantum Chemical Approach to Estimating the Thermodynamics of Metabolic Reactions. *Sci. Rep.* **2015**, *4*, 7022.

(22) Jinich, A.; Flamholz, A.; Ren, H.; Kim, S.-J.; Sanchez-Lengeling, B.; Cotton, C. A. R.; Noor, E.; Aspuru-Guzik, A.; Bar-Even, A. Quantum Chemistry Reveals Thermodynamic Principles of Redox Biochemistry. *PLoS Comput. Biol.* **2018**, *14* (10), No. e1006471.

(23) Banerjee, R. *Redox Biochemistry*; John Wiley & Sons, 2007.

(24) Hadadi, N.; Ataman, M.; Hatzimanikatis, V.; Panayiotou, C. Molecular Thermodynamics of Metabolism: Quantum Thermochemical Calculations for Key Metabolites. *Phys. Chem. Chem. Phys.* **2015**, *17* (16), 10438−10453.

(25) Jinich, A.; Sanchez-Lengeling, B.; Ren, H.; Goldford, J. E.; Noor, E.; Sanders, J. N.; Segrè, D.; Aspuru-Guzik, A. A Thermodynamic Atlas of Carbon Redox Chemical Space. 2019, bioRxiv245811. https://doi.org/10.1101/245811.

(26) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6* (12), 2326−2331.

(27) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11* (5), 2087−2096.

(28) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), No. 058301.

(29) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8* (4), 3192−3203.

(30) Pyzer-Knapp, E. O.; Simm, G. N.; Guzik, A. A. A Bayesian Approach to Calibrating High-Throughput Virtual Screening Results and Application to Organic Photovoltaic Materials. *Mater. Horiz.* **2016**, *3* (3), 226−233.

(31) Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule* **2017**, *1* (4), 857−870.

(32) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press, 2006.

(33) Kung, S. Y. *Kernel Methods and Machine Learning*; Cambridge University Press, 2014.

(34) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19* (1), 1−32.

(35) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55* (1), 39−53.

(36) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From Genomics to Chemical Genomics: New Developments in KEGG. *Nucleic Acids Res.* **2006**, *34* (90001), D354−D357.

(37) Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* **2017**, *45* (D1), D353−D361.

(38) Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* **2016**, *44* (D1), D457−D462.

(39) Kanehisa, M.; Goto, S. KEGG Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28* (1), 27−30.

(40) Goldberg, R. N.; Tewari, Y. B.; Bhat, T. N. Thermodynamics of Enzyme-Catalyzed Reactions–a Database for Quantitative Biochemistry. *Bioinformatics* **2004**, *20* (16), 2874−2877.

(41) Alberty, R. A. *Thermodynamics of Biochemical Reactions*; John Wiley & Sons, 2005.

(42) Klamt, A.; Schüürmann, G. 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* **1993**, *0* (5), 799−805.

(43) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.

(44) Heijenoort, J. v. Formation of the Glycan Chains in the Synthesis of Bacterial Peptidoglycan. *Glycobiology* **2001**, *11* (3), 25R−36R.

(45) O'Brien, P. J.; Siraki, A. G.; Shangari, N. Aldehyde Sources, Metabolism, Molecular Toxicity Mechanisms, and Possible Effects on Human Health. *Crit. Rev. Toxicol.* **2005**, *35* (7), 609−662.

(46) Ferguson, G. P. Protective Mechanisms against Toxic Electrophiles in Escherichia Coli. *Trends Microbiol.* **1999**, *7* (6), 242−247.

(47) Salimbeni, H.; Deisenroth, M. *Doubly Stochastic Variational Inference for Deep Gaussian Processes*, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, 2017; pp 4588−4599.

(48) Gal, Y.; van der Wilk, M.; Rasmussen, C. E. *Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models*, Proceedings of Neural Information Processing Systems 2014; pp 3257−3265.

(49) Yao, K.; Herr, J. E.; Parkhill, J. The Many-Body Expansion Combined with Neural Networks. *J. Chem. Phys.* **2017**, *146* (1), No. 014106.

(50) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513.

(51) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13* (11), 5255−5264.

(52) Raymond, J.; Segrè, D. The Effect of Oxygen on Biochemical Networks and the Evolution of Complex Life. *Science* **2006**, *311* (5768), 1764−1767.

(53) Zubarev, D. Y.; Rappoport, D.; Aspuru-Guzik, A. Uncertainty of Prebiotic Scenarios: The Case of the Non-Enzymatic Reverse Tricarboxylic Acid Cycle. *Sci. Rep.* **2015**, *5*, 8009.

(54) Riniker, S.; Landrum, G. A. Similarity Maps - a Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminf.* **2013**, *5* (1), 43.

(55) Asami, T.; Nakano, T.; Fujioka, S. Plant Brassinosteroid Hormones. *Vitam. Horm.* **2005**, *72*, 479−504.

(56) Fujioka, S.; Yokota, T. Biosynthesis and Metabolism of Brassinosteroids. *Annu. Rev. Plant Biol.* **2003**, *54*, 137−164.

(57) Mosblech, A.; Feussner, I.; Heilmann, I. Oxylipins: Structurally Diverse Metabolites from Fatty Acid Oxidation. *Plant Physiol. Biochem.* **2009**, *47* (6), 511−517.

(58) Zanger, U. M.; Schwab, M. Cytochrome P450 Enzymes in Drug Metabolism: Regulation of Gene Expression, Enzyme Activities, and Impact of Genetic Variation. *Pharmacol. Ther.* **2013**, *138* (1), 103−141.

(59) Tschumper, G. S.; Schaefer, H. F. Predicting Electron Affinities with Density Functional Theory: Some Positive Results for Negative Ions. *J. Chem. Phys.* **1997**, *107* (7), 2529−2541.

(60) Simons, J. Molecular Anions. *J. Phys. Chem. A* **2008**, *112* (29), 6401−6511.

(61) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024−10035.

(62) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562−2574.

(63) Alberty, R. A.; Cornish-Bowden, A.; Goldberg, R. N.; Hammes, G. G.; Tipton, K.; Westerhoff, H. V. Recommendations for Terminology and Databases for Biochemical Thermodynamics. *Biophys. Chem.* **2011**, *155* (2−3), 89−103.

(64) Dixon, S. L.; Jurs, P. C. Estimation of pKa for Organic Oxyacids Using Calculated Atomic Charges. *J. Comput. Chem.* **1993**, *14* (12), 1460−1467.

(65) Csizmadia, F.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. Prediction of Distribution Coefficient from Structure. 1. Estimation Method. *J. Pharm. Sci.* **1997**, *86* (7), 865−871.

(66) Er, S.; Suh, C.; Marshak, M. P.; Aspuru-Guzik, A. Computational Design of Molecules for an All-Quinone Redox Flow Battery. *Chem. Sci.* **2015**, *6* (2), 885−893.

(67) Huskinson, B.; Marshak, M. P.; Suh, C.; Er, S.; Gerhardt, M. R.; Galvin, C. J.; Chen, X.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J. A Metal-Free Organic-Inorganic Aqueous Flow Battery. *Nature* **2014**, *505* (7482), 195−198.

(68) Tong, L.; Chen, Q.; Wong, A. A.; Gómez-Bombarelli, R.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J. UV-Vis Spectrophotometry of Quinone Flow Battery Electrolyte for in Situ Monitoring and Improved Electrochemical Modeling of Potential and Quinhydrone Formation. *Phys. Chem. Chem. Phys.* **2017**, *19* (47), 31684−31691.

(69) Yang, Z.; Tong, L.; Tabor, D. P.; Beh, E. S.; De Goulet, M.-A.; Porcellinis, D.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J. Alkaline Benzoquinone Aqueous Flow Battery for Large-Scale Storage of Electrical Energy. *Adv. Energy Mater.* **2018**, *8*, 1702056.

(70) Gerhardt, M. R.; Tong, L.; Gómez-Bombarelli, R.; Chen, Q.; Marshak, M. P.; Galvin, C. J.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J. Anthraquinone Derivatives in Aqueous Flow Batteries. *Adv. Energy Mater.* **2017**, *7* (8), 1601488.

(71) Lin, K.; Gómez-Bombarelli, R.; Beh, E. S.; Tong, L.; Chen, Q.; Valle, A.; Aspuru-Guzik, A.; Aziz, M. J.; Gordon, R. G. A Redox-Flow Battery with an Alloxazine-Based Organic Electrolyte. *Nature Energy* **2016**, *1* (9), 16102.

(72) Matthews, A. G. d. G.; van der Wilk, M.; Nickson, T.; Fujii, K.; Boukouvalas, A.; Leon-Villagra, P.; Ghahramani, Z.; Hensman, J. GPflow: A Gaussian Process Library Using TensorFlow. *J. Mach. Learn. Res.* **2017**, *18* (40), 1−6.

(73) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(74) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* **2015**, *7*, 20.

(75) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-Consistent-Charge Density-Functional Tight-Binding Method for Simulations of Complex Materials Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58* (11), 7260−7268.

(76) Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34* (19), 1672−1685.

(77) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent Structures and Interactions by Density Functional Theory with Small Atomic Orbital Basis Sets. *J. Chem. Phys.* **2015**, *143* (5), No. 054107.

(78) Kozuch, S.; Martin, J. M. L. DSD-PBEP86: In Search of the Best Double-Hybrid DFT with Spin-Component Scaled MP2 and Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2011**, *13* (45), 20104−20107.

(79) Riplinger, C.; Neese, F. An Efficient and near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138* (3), No. 034106.

(80) Boudart, M. Thermochemical Kinetics, 2nd. Ed., Sidney W. Benson, Wiley Interscience, 320, New York, 1976. *AIChE J.* **1977**, *23* (4), 613−613.