Batteries & Supercaps

Articles
doi.org/10.1002/batt.202100059

Chemistry Europe
European Chemical Societies Publishing

Special Collection

Rapid Prescreening of Organic Compounds for Redox Flow Batteries: A Graph Convolutional Network for Predicting Reaction Enthalpies from SMILES

James Barker,[a, b] Laura-Sophie Berg,*[a] Jan Hamaekers,[a, c] and Astrid Maass[a]

Identifying interesting redox-active couples from the vastness of organic chemical space requires rapid screening techniques. A good initial indicator for couples worthy of further investigation is the heat of reaction ΔH°. Traditional methods of calculating this quantity, both experimental and computational, are prohibitively costly at large scale. Instead, we apply a *graph convolutional network* to estimate the heats of reaction of arbitrary redox couples orders of magnitude faster than conventional computational methods. Our graph takes only SMILES strings as input, rather than full three-dimensional geometries. A network trained on a dataset of atomization enthalpies for approximately 45,000 hydrogenation reactions, applied to a separate test set of 235 compounds and benchmarked against experimental heats of reaction, produces promisingly accurate results, and we anticipate that this methodology can be extended to other RFB-relevant reactions. However, lower predictivity for compounds in regions of chemical space not covered by the training dataset reinforces the pivotal importance of the particular chemistries presented to a model during training.

1. Introduction

The growth in demand for powerful and flexible energy storage systems shows no signs of abating. Redox flow batteries (RFB) are now a serious option, thanks to their design flexibility, high scalability, and the independent control they offer over energy and power.[12,30,35,59] Several, mostly inorganic, systems have been implemented that demonstrate these benefits,[5,17,26,51,53,58] but such conventional RFBs are limited in applicability, as they utilise materials as active species or solvents that are either expensive (e.g. vanadium), toxic (e.g. lead), aggressive (e.g. bromine), or otherwise problematic. Therefore, the search for organic alternatives has begun,[6,30,57,59] starting from natural redox motifs such as quinone, flavine, and alloxazine ring systems.[11,15,31,38] For the class of quinones, for example, it was demonstrated that a given core structure may be customized by systematic modifications, suggesting the possibility of an all-quinone RFB.[11,15]

Identifying low-cost redox couples which are safe, can undergo rapid and reversible redox reactions, and remain both stable and soluble will provide a basis for further development of sustainable and competitive energy storage systems and has been a primary objective of several previous screening studies. These range from purely experimental small-scale studies to computational or combined studies, with total numbers of compounds under investigation numbering from tens to millions.[—6,11,15,21,29,32,47,55]

The organic chemical space offers possibilities for forming redox couples that far exceeds those found under restriction to purely inorganic compounds. At this stage of research, a truly unbiased search seems necessary to pair up the most suited redox couples – i.e. half-cell systems – to compose efficient full-cell systems. A broad search for candidate materials seems inevitable, since prohibitive kinetic aspects, obstacles due to mass and heat transport, and/or economic aspects must be considered in addition to thermodynamic fundamentals.[6,52]

The amount of compounds that can be processed in such a search is necessarily dependent on the practical or computational effort required to evaluate each compound. A number of approaches have been applied to accelerate the screening process while maintaining sufficient levels of accuracy. Pure experiment can be complemented by computational chemistry methods derived directly from quantum mechanical principles. Such *ab initio* and *first principles* quantum-mechanical (QM) methods, like e.g. coupled cluster (CC), second-order Møller-Plesset (MP2), and density functional theory (DFT), as well as *composite* methods such as G4(MP2), are established and powerful tools for the calculation of material properties, and thermochemical properties in particular.[7-8,19,37] Recently, these methods have been either augmented or replaced entirely by

[a] *J. Barker, L.-S. Berg, Dr. J. Hamaekers, Dr. A. Maass*
Department of Virtual Material Design,
Fraunhofer Institute for Algorithms and Scientific Computing,
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
E-mail: laura-sophie.berg@scai.fraunhofer.de

[b] *J. Barker*
Institute for Numerical Simulation, University of Bonn,
Friedrich-Hirzebruch-Allee 7, 53115 Bonn, Germany

[c] *Dr. J. Hamaekers*
Fraunhofer Center for Machine Learning,
Schloss Birlinghoven,53757 Sankt Augustin, Germany

Supporting information for this article is available on the WWW under https://doi.org/10.1002/batt.202100059

An invited contribution to a Special Collection on Artificial Intelligence in Electrochemical Energy Storage.

the use of machine learning (ML) models, since "the low numerical complexity and high accuracy of machine learning algorithms makes them very attractive as a pragmatic substitute for *ab-initio* and DFT methods".[48] A diverse collection of ML algorithms has been successfully adopted to the chemical setting, with focus on the prediction of energetic molecular properties.[2-3,9,14,18,25,40-41,46]

As a result of these efforts, we are now in the fortunate situation of having access to large, precomputed datasets suitable for use in ML model training,[4,24,27,34,43-44] as well as efficient recipes[1,8,10,19] for the preparation of even more data, should this prove necessary. As such, we need – at least in principle – to invest only modest computational resources in the collection, collation, and possibly the refinement of training data appropriate to the RFB problem space. Once a suitable ML model is trained, we may hope to derive accurate approximations to expensive target properties of an arbitrary number of candidate systems under prescreening. Several groups have begun to exploit this position in order to identify new redox couples and electrolyte materials.[6,15,23,29]

It remains unclear how easily and effectively such techniques can be applied to the problem of redox couple prescreening. As mentioned, the scope of the organic compound space precludes all but tightly-focused searches, and even then, prediction performance will be critical. Most previously published methods for ML-based prediction rely on features of molecules that are derived directly from the distances between their component atoms, and are therefore dependent on explicit three-dimensional geometries.[20,22,45-46,54] These geometries must themselves be calculated through an optimization process, using either empirical force field (FF) methods or explicit QM calculations. Both classes of methods have advantages and disadvantages, but regardless of choice, such optimisations are likely to be far more expensive than the evaluation of trained models themselves. Moreover, questions remain about the potential impact of the choice of geometry optimization process on the generalizability and bias of trained models.[4,27,33,39] If not treated carefully, such issues may undermine the usefulness and reliability of ML-based prescreening, and in particular the performance, since geometry relaxation is in fact a non-trivial (and non-trivially costly) component of a complete conventional calculation.

In this study, we aim to predict one of the main thermodynamic parameters for a hydrogenation half-cell reaction, using only a minimum of input information. The value of the predicted property could serve as a sensible filtering criterion in a hierarchical computational screening as outlined in e.g.[6] The focus here is not on providing a computational framework (as in Refs. [28,42]) or on producing a large set of pre-computed molecule data (as in Refs. [19,24,34]), but rather on investigating the initial application of previously-published datasets and ML techniques to the prediction of real-world, out-of-training-set properties relevant to the RFB space. We hope to produce a model that is immediately and efficiently applicable to a broad range of compounds.

To define the task more precisely: we consider the reaction enthalpies ΔH° as a metric for a molecular scaffold's capability to accommodate an extra pair of electrons and protons, i.e. we focus on reactions of the generic pattern $A + H_2 \rightarrow AH_2$. For simplicity, we restrict ourselves at this stage to the study of neutral, organic, closed shell compounds (A, AH_2) involved in single hydrogenation events only. To account for the aspect of reversibility (for application in a RFB), we do not consider bond cleavage and ring opening reactions.

Our starting point for training data is the well-known QM9 dataset[43] after energy refinement using the G4(MP2) method;[27] we refer to this collection as the *QM9/G4(MP2) dataset*. The base QM9 dataset has been used in several comparable studies as a benchmark.[3,18,20,22,40,45-46,54] It provides approximately 45,000 redox couples that can be recombined from the approximately 130,000 individual compounds contained in the full dataset. The required molecular heats of atomization (H°) are available at G4(MP2)-level for molecules in gas phase – we postpone for now the issue of computing and including solvation effects, since the solvation contribution could introduce additional uncertainty.[34]

We use this dataset to train a *graph convolutional network (GCN)*, following the work of Gilmer et al.[18] GCNs are a class of neural network that treat graph-theoretical descriptions of input datapoints, taking as input two feature matrices that include information about both the nodes and edges of graphs respectively. The presentation of molecular structures in graph form is obvious and intuitively appealing, as the standard ball-and-stick picture of molecules translates directly to a graph representation, with atoms as nodes and bonds between them as edges. We obtain such graphs directly from molecule specifications in the SMILES format;[56] these specifications are available for most major datasets, including QM9/G4(MP2). By working with these graphs directly, we completely avoid any costly and/or theoretically questionable geometry optimizations. The application to prediction of reaction properties rather than individual-compound properties requires careful consideration of input format; we introduce differential features to our input vectors to significantly improve our prediction performance.

Once trained, we assess the capability of the model by applying it to a set of similar reactions extracted from the NIST WebBook,[13] which also provides experimentally-determined thermochemical properties for these reactions. We supplement this *NIST dataset* with QM reaction enthalpies computed using the G4(MP2) composite method. These calculated enthalpies allow us to calibrate and assess the NIST dataset itself as a suitable reference target; furthermore, in case of mismatch between prediction and reference, they allow us to investigate any such discrepancies for either experimental shortcomings or errors in our model process.

2. Results

We now report at a high level the results of three complementary experiments exploring the capabilities of our proposed model. For a more thorough treatment of dataset composition and the precise structure of our machine-learning model, we refer the reader to the Experimental Section, and to the supporting information of this article. We also provide a brief discussion of the observed relative performance of our approach.

2.1. Model Validation on QM9/G4(MP2) Dataset

As an initial investigation into the general suitability of the GCN model, we began by benchmarking it on all of the approximately 45,000 molecular pairs of hydrogenation reactions drawn from the QM9/G4(MP2) dataset. To minimise confounding results from test/training set selection, we applied the standard five-fold cross-validation technique while benchmarking.

When applying machine learning techniques to the prediction of molecular properties, the conventional objective is to reach chemical accuracy. Previous benchmarks of GCNs trained on the base QM9 dataset report prediction performance up to four times better than chemical accuracy.[3,18,46] However, most of these models include some kind of three-dimensional information as feature input and thus require expensive pre-processing steps. Rare exceptions are Gilmer et al.[18] and lately Pinheiro et al.,[40] who report results for a model trained only on features that can be extracted directly from the SMILES representations of the input molecules' topological structures.[56] Of the 13 properties available in the base QM9 dataset, these groups report predictions within chemical accuracy for 11 and 9 of them respectively (where "chemical accuracy" is defined according to the appropriate units of each property).

To validate our model against existing benchmarks, we first trained it to predict heats of atomization H°. Across our five test/training splits, we observed an average training MAE of 0.02 eV and an average validation MAE of 0.06 eV. These results agree with those of Gilmer et al, and are sufficiently accurate from a chemical perspective, as the chemical accuracy of H° is usually taken to be 0.04 eV.[19]

Chemical accuracy for the reaction enthalpy ΔH° is taken analogously to the chemical accuracy of the heats of atomization H° as 0.04 eV.[18] When validating the model against the reaction enthalpies ΔH° in the G4(MP2) dataset, we observed an average training MAE of 0.02 eV, comfortably below chemical accuracy. The average testing MAE was 0.06 eV, slightly worse than chemical accuracy. Although tolerable, such error is worthy of further attention in future work. We suspect in particular that the complexity of the model might lend itself to overfitting of the training data; this could be addressed through a combination of a larger input dataset size, and the use of regularization techniques. Nevertheless, we consider the quality of prediction in this initial assessment to be entirely satisfactory, and sufficient to justify investigation of out-of-set applications.

2.2. NIST Dataset Validation/Investigation

As a test set for our GCN model, we selected a collection of experimentally-characterized hydrogenation reactions, drawn from the NIST WebBook[13] and analogous to those found in the original QM9/G4(MP2) dataset. This NIST dataset contains unique reactions involving distinct compounds.

To establish the suitability and quality of this dataset as a reference, we first compared it by overlap to the original QM9/G4(MP2) dataset. There are molecules which are present in both datasets, with equivalence judged according to SMILES string. After a prescreening process of conformer generation and selection (cf. Experimental Section), we explicitly calculated the enthalpies of atomization $H^\circ_{\mathrm{NIST/G4(MP2)}}$ for each such molecule in our NIST dataset. As alternative calculations of these properties $H^\circ_{\mathrm{QM9/G4(MP2)}}$ are included in the QM9/G4(MP2) dataset, comparing two values for each molecule allows us to test the agreement between the two datasets. Indeed, we find agreement between the reference enthalpies of atomization H° given in the QM9/G4(MP2) dataset and our own calculations (MAE = 0.016 eV) – despite the fact that our procedure does not guarantee retrieval of the globally optimal conformation for flexible molecules. We consider the newly computed enthalpies of atomization H° to be compatible to the QM9/G4(MP2) data in this respect.

A further check for the plausibility of our procedure and the results obtained is to contrast our computed heats of reaction $\Delta H^\circ_{\mathrm{NIST/G4(MP2)}}$ with the experimental values found in the original NIST database ($\Delta H^\circ_{\mathrm{NIST/EXP}}$, cf. Figure 1). Overall, despite varying reaction conditions, we see good agreement between the two: the slope of a linear-fitted trend line is almost unity ($R^2 \approx 0.935$). The individual values do exhibit considerable scatter, which we cannot attribute to the choice of solvent: only about 20% of reactions took place in gas phase and thus correspond to the conditions used for G4(MP2) computations, whereas the majority of experiments were conducted in liquid phase, and yet this has no obvious impact on the resulting heat of reaction. The provenance of the individual values appears to be more relevant: while most of the individual records in the NIST dataset have been assigned uncertainties below 0.04 eV, and therefore claim to have been determined to chemical accuracy, the values obtained between different groups vary by up to 0.1 eV. Regardless of the causes of these fluctuations in experimental reference values, we did not find evidence against the computed values and consider the conformation generation procedure to have performed well and robust enough that the dataset can legitimately be used for validating ML predictions.

2.3. Trained GCN Performance on NIST Dataset

Our experiment culminates in an evaluation of our proposed GCN model using the out-of-set examples provided by our NIST dataset. We trained an instance of our model using the complete QM9/G4(MP2) dataset (i.e. without any dataset splitting as performed in Section 3.1), and used it to predict values of ΔH° for SMILES strings for each of the reactions in the NIST dataset. We compared the obtained values against both the fully-computed G4(MP2) reference values $\Delta H^\circ_{\mathrm{NIST/G4(MP2)}}$ (cf. Figure 2) and against the original experimental values from the NIST database $\Delta H^\circ_{\mathrm{NIST/EXP}}$ (cf. Figure 3).

We begin by noting that the predictions of the compounds that appear in the overlap between the QM9/G4(MP2) dataset and NIST dataset (highlighted by crosses in Figure 2) are generally very good. This is to be expected, as those reactions form part of the training data and should thus be recognisable to the model. There are, however, some poorly-predicted

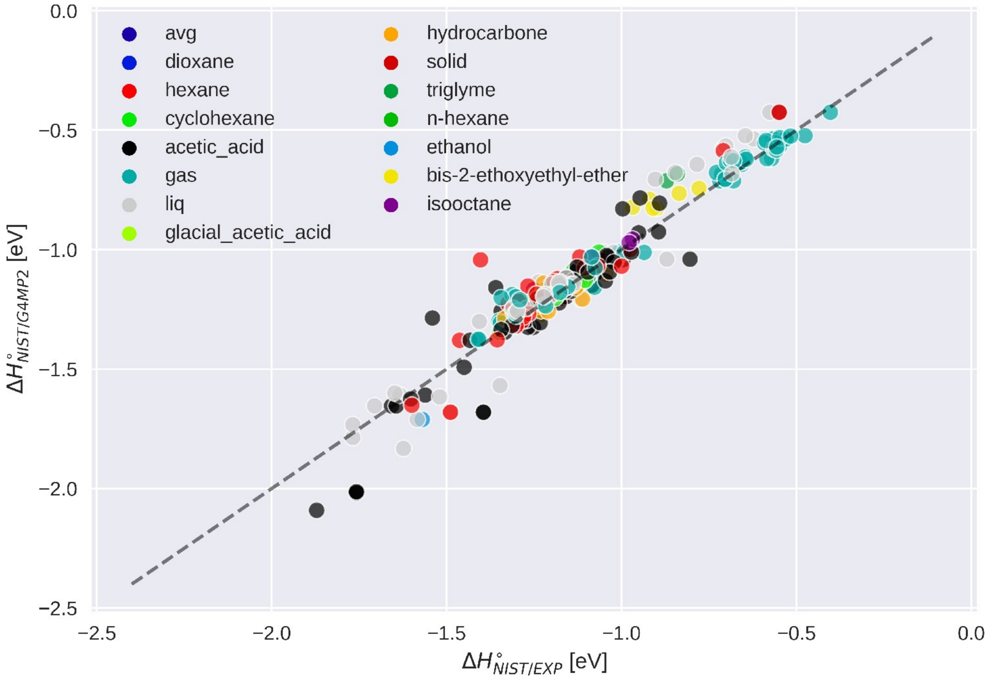

Figure 1. Experimental values for heat of reaction $\Delta H^\circ_{\mathrm{NIST/EXP}}$ extracted from thermochemical data given in the NIST Chemistry WebBook[13] versus computed values for heat of reaction $\Delta H^\circ_{\mathrm{NIST/G4(MP2)}}$. Data points are coloured by solvent.

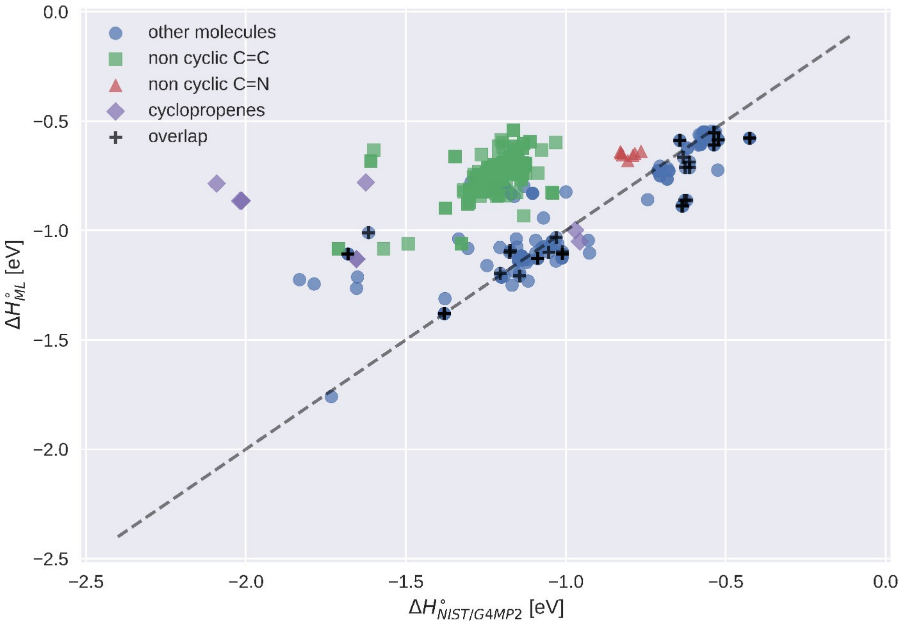

Figure 2. Predicted values of ΔH°_{ML} for the NIST dataset in comparison to computed values; overlaps with QM9/G4(MP2) dataset marked by +, molecules with CC-double bonds in open chains highlighted by green squares, molecules with CN-double bonds in open chains by red triangles, and double bonds in cyclopropenes in purple rhombuses.

outliers in the overlap set. The most visible ($\Delta H^\circ_{\mathrm{NIST/G4(MP2)}}$ below −1.5 eV) originate from an (E)-cyclo-octene scaffold, a structurally strained ring system which is apparently (and necessarily) underrepresented in the QM9/G4(MP2) dataset: the 8-ring is the smallest possible size to integrate a strained, (E)-configured double bond and only just fits below the limit of 9 heavy atoms per molecule in the QM9/G4(MP2) dataset. The model underestimates the amount of energy released for these cyclo-octenes, not reflecting the significant relaxation of strain as soon as a rigid building block becomes flexible. This particular shortcoming does not afflict the contextual usefulness of the model, as relaxing such a strained ring-system by reduction is hardly a reversible reaction; while the scaffold will readily undergo the hydrogenation reaction, the more favourable (Z)-

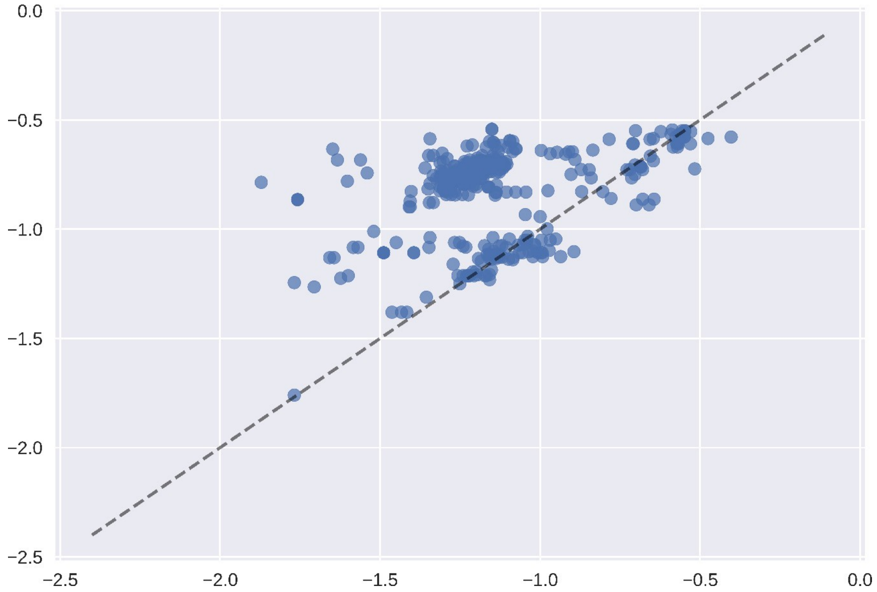

Figure 3. Prediction of experimental values of ΔH° from NIST database.

isomer will re-form upon oxidation (and this is indeed correctly predicted).

The remaining points in Figure 2 show the predictions for the other, hitherto unseen, compounds from the NIST dataset, highlighted by circles. These are distributed over a roughly triangular area. The lower side of this triangle nicely matches the diagonal of ideal prediction (shown as a dashed line), and as would be hoped, many predictions fall close to this line. A significant portion of compounds, however, deviate by underestimating the energetic difference among redox pairs. We find this interesting, as it suggests systematic behaviour of the model rather than general noise; the latter would be expected to result in an even scatter around the perfect prediction line. A large group of compounds (marked in Figure 2 with green squares) cluster around −1.5/−1.0 eV; most of them display terminal CC-double bonds or generally double bonds in open chain molecules. We suspect that we have detected here a limitation of the chemical space represented within the QM9/G4(MP2) data set, rather than an explicit failure of our model *per se*.

We turn now to Figure 3, where the reaction enthalpies predicted by our model are plotted against the corresponding experimental values from the NIST dataset. Consistent with the observed strong correlation between experimental and computed ΔH°-values, the distribution of data points displays very much the same pattern as seen in Figure 2, albeit horizontally stretched due to duplicated experimental results flowing from the original source. Considering the uncertainties noted above with respect to some of the experimental data, and transferring this to data points that come with single well-defined values, we suspect the experimental data to be generally afflicted with uncertainties of about 0.05 eV. This is roughly at the same order as the MAE of 0.06 eV achieved for the validation set above, so such an accuracy is entirely sufficient for the purpose of prescreening.

2.4. Model Performance

As our focus here has been predominantly in terms of model feasibility and accuracy, we deliberately avoided tuning our model for performance in favour of applying well-understood techniques with as little modification as possible. Nevertheless, the relative performance of the GCN approach is compelling.

Our GCN model was implemented in the Python programming language, using the Tensorflow framework. Training our model on the QM9/G4(MP2) dataset took several days. This calculation was performed using only a single core of a standard commodity compute server, with Tensorflow's default multithreading explicitly disabled. Once trained, evaluation of the model proved to be exceptionally rapid. Although we have not benchmarked precise figures, we observed anecdotally that predicted properties could be produced from SMILES strings at a rate approaching 50 predictions per second. We note that this can be considered a minimum-effort performance figure, as GCNs are highly amenable to parallelisation using e.g. GPU computation. We can reasonably anticipate that performance tuning could improve training and evaluation performance by at least another order of magnitude again.

The effort required to calculate the QM data used to validate the NIST dataset was significantly greater. For these calculations, we used a hand-tuned build of NWChem, running on multiple nodes of a modern HPC cluster. Simple DFT geometry optimizations using a B3LYP exchange-correlation functional and the 6–31 g(2df,p) basis set, used for prescreening during the conformer-generation step, consume from several minutes to more than an hour of compute time on single nodes, depending on the complexity of the molecular configuration under study. The actual G4(MP2) thermochemical calculations required between four to eight compute nodes, partially due to memory constraints, and in many cases took over an hour each to run to completion.

Although a careful comparison of relative performance is beyond the scope of this study, the GCN approach appears to allow dramatically higher throughput for estimation of heats of reaction compared to conventional QM calculations and experimental study. We feel that it is safe, and indeed conservative, to suggest a throughput difference of at least four to five orders of magnitude in terms of elapsed time per compound, and perhaps more. We stress, however, that further work must be performed here before general conclusions can be drawn.

3. Discussion

Overall, the predictions made by our model place redox couples either close to the desired perfect-prediction line, or underestimate their absolute ΔH°-values. The former is of course desirable, and indicative of a properly trained model applied in a well-understood feature space. When studying the latter case, we identified at least one cluster of redox couples that were apparently underrepresented in the training data.

Despite the impressive size of the QM9/G4(MP2) dataset, its coverage of chemical space is inherently limited, due to the neutrality of included molecules, their relatively small size (of less than 10 heavy atoms), the small range of constituting elements (only C, H, N, O, and F) and the original objective of including only globally optimal structures. The latter feature is essential for training models to predict ground state properties, but presents a challenge with respect to the creation of general molecular geometries, particularly for highly flexible molecules. The impact of this can be reduced, and the issue of specifically sampling strained atomic environments can be addressed, by focusing rather on rigid and/or cyclic structures with well-defined and unambiguously obtainable geometries due to constitutional constraints. The QM9/G4(MP2) data set contains a number of heavily cross-linked and somewhat hypothetical compounds. Even so, Kim et al.[27] reported (mostly minor) changes in geometries and energies upon refinement, which highlights the difficulties in dealing with flexible compounds in principle and in practice.

The choice of compounds that have been experimentally characterized in the NIST dataset, on the other hand, suggests that the impact of the local environment in which a double bond is embedded has been systematically examined. Compounds with identical scaffolds, but different double-bond sites, varying sizes of residues, and different E/Z-isomers, allow probing of the stereo-chemical effects of a double bond's surroundings on reaction enthalpies. At the same time, the NIST dataset is not limited to a certain size or flexibility of molecules, which makes it harder to identify minimizing conformations. The good correlation between experimental and computed heats of reactions suggests that our procedure could identify, if not the global optimum itself, then at least conformations favourable enough to represent a sufficiently relaxed state. Flexible compounds would be indeed best represented by an ensemble of conformers. The findings of Jinich et al.[23] indicate, however, that the single-molecule approach produces similar results as the ensemble approach at much lower cost and is frequently practised in related workflows.[21,24,37,47]

Jinich et al. have presented a mixed QM/ML approach that pursues a similar objective, i.e. predicting redox potentials for biochemical compounds, while making use of reaction fingerprints. They propose to predict electronic energies based on semiempirical PM7 calculations and subsequent corrections for systematic errors with Gaussian process regression against experimental data.[23] While many aspects of their study show parallels to the one described here (neutral, closed-shell organic compounds; a similar procedure for 3D-conformation generation; use of reaction fingerprints), they differ in the chemical space covered (carbonyls, alcohols, and amines, versus QM9 chemical space), their inclusion of solvation effects (aqueous solution versus gas phase) and in the way that QM and ML steps are linked. While their approach is to use QM calculations and ML in tandem, i.e. to run cheaper QM calculations for each compound and improve the accuracy of these results by an ML model calibrated on experimental data, ours is to run expensive calculations on a given set of compounds for training the model and apply the model to any number of compounds afterwards. Once a model has been trained on a well-balanced set of compounds, we become independent of experimental data (due to high level QM calculations), and fast, as the actual prediction does not involve any computation other than the evaluation by the model. We consider this a major advantage of our approach.

Neglecting the impact of solvation may also be beneficial, as it allows us to focus entirely on the electronic characteristics of the problem at hand, without committing to a predetermined solvent. Such a solvent may be chosen in a further filtering step, and candidate couples identified as promising may also be considered for use in alternative solvents, since the ranking seems robust with respect to the choice of solvent or dielectric constant.[6]

The main purpose of this study was to provide a tool for identifying redox-couples fit for use as electro-active materials; more specifically, candidate anolytes with formal reduction potentials of below 0.2 V vs Standard Hydrogen electrode (SHE) or possible catholytes with formal reduction potentials above 0.9 V vs SHE.[15] For that purpose, the reaction enthalpies can be converted into the associated formal redox potentials via $E^{\circ} = \Delta H^{\circ}/(-2n \cdot F)$, where $n = 2$ and F is Faraday's constant, and approximating the free energy of reaction by the reaction enthalpy. When considering also the reaction's entropy, which is dominated by the contribution of the one constant reactant dihydrogen contributing an overall net shift of about 0.3 eV, we find that none of the formal redox potentials associated with the NIST data set fall into the interval for potential anolytes. There are, however, potentially catholyte motifs to be found. For the time being, we refrain from highlighting particular compounds or scaffolds, as the model presented here is intended to be exploratory, rather than fine-tuned to the targeted application.

4. Conclusion

In this study, we have presented an initial implementation of a rapid prescreening tool, developed to predict the driving force for redox reactions. Our approach decouples the computationally expensive steps of high-quality QM computations and ML model training from the screening process itself. We estimate a saving in time per redox pair for ML versus QM of several orders of magnitude. In a hierarchical screening scenario, this allows for an efficient selection of potential candidates at a minimum of cost, allowing the investigator to focus efforts on the most promising redox couples in a search space.

The "secret ingredient" in the screening process is our deliberate ignorance of explicit geometry information when creating the input features – the time-consuming steps of three-dimensional conformation generation, optimization and selection as a preparatory stage for evaluation are omitted entirely. Although some loss in accuracy is inevitable when compared to approaches that do include this information, the loss turns out to be acceptable – particularly considering the scattering observed among high-quality collection of experimentally obtained comparison data. We achieve this due to our use of "differential" features, with which we are able to encode the reactive centre into the feature matrix. We thus obtain more insightful results than with the standard features usually used in molecular property prediction.

The critical aspect in any QM/ML-based approach is the quality, diversity, and quantity of the data presented to the model during training. The QM9/G4(MP2) data set represents a major achievement in this respect, as it provides us with a set of high-quality data of critical size that can be used as a starting point for developing ML approaches. The choice of training molecules within this dataset (neutral, small, limited range of elements), however, seems not to be an ideal match for the RFB setting, which may also involve charged or radical atomic species. Nevertheless, the fundamental features of redox reactions have successfully been captured, and our model allows classification between anolyte or catholyte candidates, particularly since two-electron transfer reactions are inherently associated with higher energy densities.

Our next steps will be to carefully assemble a new generation of training data. Apart from reducing the QM9 data set to the most essential representatives, e.g. by using a query-by-committee approach[20,49] for improved generalization, training speed-up and error estimation, we intend to extend the data set with collections of common, larger molecules[4] on the one hand, and by including designated RFB-relevant materials, such as almost 32,000 quinone and aza-aromatic redox pairs that have been recently made available.[50] We will also investigate transferring the approach presented here for a more thorough screening of the search space by adapting the model to one-electron transitions, which are also described in existing relevant data sets.[24,34] With only minor modifications, we hope to apply this concept to the prediction of reorganization energies, so that we can cover not only the thermodynamic basics, but also the kinetic aspects of reactions. The latter would require a considerable effort in the generation of the training data, but could build on the existing data sets mentioned above.

In terms of the training process itself, the input features may be refined, particularly with respect to ring strains or interactions between bulky residues. Small, prototypical molecules may receive higher weights, since their structural motifs will reappear as substructures in larger molecules.

In conclusion, although there is certainly a need for further work, we find the initial results of this approach to be highly promising, and suggestive of a powerful new approach to the prescreening of potential redox couples.

Experimental Section

Datasets

QM9/G4(MP2)

The QM9 dataset[43] is an established reference-quality dataset for machine learning involving molecular properties. The dataset consists of approximately 134,000 organic molecules containing up to nine heavy atoms, each either carbon (C), oxygen (O), nitrogen (N), or fluorine (F). Molecules are specified both as SMILES strings, and as optimised three-dimensional geometries. Thirteen different fundamental chemical properties for each molecule are included in the data set; of these, we are most interested in the heats of atomization H° at 298 K. In the original version of the QM9 dataset, all properties were calculated to a B3LYP/6-31G(3df,p) level of quantum chemical theory. However, Goerigk et al. and Ramakrishnan et al.[10,19] later benchmarked several DFT approximations against various experimental data, including inorganic and organic compounds, and found that the popular B3LYP functional exhibits large errors in the calculation of molecular formation enthalpies,[10] which may not necessarily be expected to cancel when computing reaction enthalpies.[19] Both works advocate the use of alternative DFT approximations such as ω-B97X-V or ω-B97X-D3, and Ramakrishnan et al. state that the G4(MP2) correlation-consistent composite approach outperforms these recommendations again.

Kim et al. published a refined version of the QM9 dataset, using the G4(MP2) model to compute atomization energies and enthalpies. This *QM9/G4(MP2) dataset* is purged of a negligible fraction of duplicates and multi-molecular entries, but still contains approximately 134,000 molecules. While the QM9/G4(MP2) dataset does not explicitly include the SMILES of the component molecules, these are easily obtained from the original QM9 dataset due to the equivalence of indexing between the two. From the 134,000 molecules in the dataset, we identified approximately 45,000 molecules that can be combined to provide suitable hydrogenation reactions for our purposes. The reaction enthalpy ΔH° for each of these reactions is retrieved as

$$\Delta H^\circ = H_{AH_2} - \left(H_A + H_{H_2}\right) \tag{1}$$

where AH_2 is the hydrogenated molecule, A the non-hydrogenated molecule, and $H^\circ_{H_2}$ was taken from the NIST-CCCBDB[36] and completed with thermal corrections to yield an atomization enthalpy at 298 K of −1.165729 Eh.

NIST reference reactions

Additionally to the QM9/G4(MP2) dataset, a reference set of similar, experimentally-characterized reactions was extracted from the NIST Chemistry WebBook.[13] The intention of this *NIST dataset* is to allow us to test a model trained on the complete QM9/G4(MP2) dataset against previously unseen "real-life" compounds. For the sake of comparability, only reactions involving neutral molecules composed of C, H, N, O or F, for a total of up to 21 heavy (non-hydrogen) atoms, were considered. 370 individual records (including some replicated values) were found, with ΔH_r°-values ranging from approximately -0.4 eV to -1.8 eV. After removal of duplicate entries, 247 unique reactions remained. For each of these, we calculated the heats of reaction $\Delta H^{\circ}_{\mathrm{NIST/G4(MP2)}}$ by using the G4(MP2) implementation of[16]. For a more detailed description of the steps running from SMILES generation to the thermochemical property calculation, we refer the reader to the supplementary information of this article.

In order to balance the reactions for the final computation of reaction enthalpies, a contribution for molecular hydrogen was necessary. The value for the internal energy at CCSD(T)/aug-cc-pVQZ-level (i.e. the theory level that the G4(MP2)-calculations are intended to approximate) was taken from the NIST-CCCBDB[36] and completed with thermal corrections to yield an atomization enthalpy at 298 K of -1.165729 $\mathrm{E_h}$, which was used throughout the study to calculate the heats of reaction.

Machine learning model

Graph convolutional networks (GCNs) have been shown to be very suitable for the task of predicting molecular properties.[3,18] The translation of molecular configurations to graphs is simple and intuitively appealing: molecules of atoms (nodes) are connected by bonds (edges). As such, and due to the growing maturity of the underlying theory, GCNs have experienced a recent surge of interest as potential tools in chemical machine-learning.

Standard neural networks take as input simple vectors of numeric or numerically-encoded *features*. By extension, GCNs consider two classes of input: node features and edge features. Node features encode information attached to each of the nodes (atoms) in the graph, which in the molecular case can include atomic number, hybridization state, and/or the number of valence electrons. Edge features specify information about connecting bonds, such as bond order, or presence/absence of the bond in/from an aromatic ring. Most published models rely on features that include some kind of distance information, derived from explicit three-dimensional molecular geometries.[3,18,46] To obtain these geometries, an expensive optimization process is needed; this is undesirable in the context of screening a large molecular space. Also, assumptions made during geometry optimization may introduce bias to the model.

Rather than considering distances, we base our model solely on information easily derivable from a molecule's SMILES string. SMILES strings encode the topological structure of a chemical compound. They are not unique, as one molecule can be written as potentially several SMILES strings, which introduces issues of consistency into the process of data generation and handling. There appears to have been relatively little work on the prediction of molecular properties in the absence of distance information, at least in the context of the QM9 and related datasets; however, Gilmer et al. were able to obtain chemical accuracy for most of the molecular properties given in the QM9/G4(MP2) dataset.[18] Our model has a very similar structure to theirs, and indeed, we were able to reproduce their results where appropriate. Further information regarding the model and the training process is contained in the supplementary information.

The main difference between ours and Gilmer's model lies in our approach to feature handling. Choosing a suitable and suitably descriptive set of features for a given problem is a critical aspect in the successful application of machine learning. Throughout the various papers on molecular property prediction, a standard feature set has emerged, which, including some extensions here and there, seems to work well for most target properties. However, most of these publications focus on predicting properties of single molecules. As we are trying to predict reaction enthalpies and thus a property of a reaction rather than a molecule's property, and since the standard feature set does not give the expected results in terms of accuracy, we augmented the typical set of input features by introducing *differential features.*

Rather than basing our features on individual SMILES strings, we base them instead on pairs: the hydrogenated and non-hydrogenated molecules involved in a hydrogenation reaction. The original idea of the differential features was to extend the feature matrix of the hydrogenated molecule AH_2 by the difference of the standard features of both molecules, i.e. subtracting the feature matrix of A of the feature matrix of AH_2 and adding those differential features as additional features to the feature matrix of AH_2. Only a few of the typical features differ between hydrogenated and non-hydrogenated molecules, particularly the hybridization state, the number of valence electrons and whether an atom is a donor or acceptor; this leaves us with four non-trivial extra features to consider (as all other differential features are zero and thus do not add any relevant information). While testing different feature combinations for the task of reaction enthalpy prediction, we have achieved best results when considering only the hybridization state as a feature; more explicitly, the hybridization states of the hydrogenated molecule extended by the difference in hybridization states of the hydrogenated and non-hydrogenated molecules. The latter puts an emphasis on the reactive centre of the hydrogenation reaction, as it marks the position of the two atoms participating in the double bond that is disbanded. This novel and simple approach enables us to apply machine learning over more complex properties where the standard features fail, and underlines the importance of including domain knowledge in feature generation.

Acknowledgement

This project (SONAR) has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no. 875489. This paper reflects only the authors' view. The funding agency is not responsible for any use made of the information it contains. In addition, this work was also supported in part by the BMBF-project 05M2AAA MaGriDo and in part by the Fraunhofer Cluster of Excellence Cognitive Internet Technologies. Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: ab initio calculations · electrochemistry · graph convolutional network · reaction enthalpies · redox flow batteries

[1] C. Bannwarth, S. Ehlert, S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 3, 1652–1671.
[2] J. Barker, J. Bulin, J. Hamaekers, S. Mathias, In *Scientific Computing and Algorithms in Industrial Simulations.* Springer, **2017**, 25–42.
[3] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* **2019**, *31*, 9, 3564–3572.
[4] G. Chen, P. Chen, C.-Y. Hsieh, C.-K. Lee, B. Liao, R. Liao, W. Liu, J. Qiu, Q. Sun, J. Tang, et al., *arXiv preprint arXiv:1906.09427* **2019**.
[5] R. Chen, S. Kim, Z. Chang, *Redox: Princ. Adv. Appl.* **2017**, 103–118.
[6] L. Cheng, R. S. Assary, X. Qu, A. Jain, S. P. Ong, N. N. Rajput, K. Persson, L. A. Curtiss, *J. Phys. Chem. Lett.* **2015**, *6*, 2, 283–291.
[7] S. R. Chinnamsetty, M. Griebel, J. Hamaekers, *Multiscale Model. Simul.* **2018**, *16*, 2, 752–776.
[8] L. A. Curtiss, P. C. Redfern, K. Raghavachari, *J. Chem. Phys.* **2007**, *127*, 124105.
[9] N. Dandu, L. Ward, R. S. Assary, P. C. Redfern, B. Narayanan, I. T. Foster, L. A. Curtiss, *J. Phys. Chem. A* **2020**, *124*, 28, 5804–5811.
[10] S. K. Das, S. Chakraborty, R. Ramakrishnan, *J. Chem. Phys.* **2021**, *154*, 4, 044113.
[11] Y. Ding, Y. Li, G. Yu, *Chem* **2016**, *1*, 5, 790–801.
[12] Y. Ding, C. Zhang, L. Zhang, Y. Zhou, G. Yu, *Chem. Soc. Rev.* **2018**, *47*, 1, 69–103.
[13] D. R. Burgess, J. "Thermochemical Data" in *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, Eds. P. Linstrom and W. Mallard, National Institute of Standards and Technology, Gaithersburg MD, 20899, retrieved August 17, 2020 .
[14] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, In the *Proceedings of Advances in Neural Information Processing Systems* **2015**, *28*, 2215–2223.
[15] S. Er, C. Suh, M. P. Marshak, A. Aspuru-Guzik, *Chem. Sci.* **2015**, *6*, 2, 885–893.
[16] M. Ernst, Composite thermochemistry NWChem. URL: https://github.com/mattbernst/composite-thermochemistry-nwchem, **2016**. Accessed: 2021-02-04.
[17] Fraunhofer Gesellschaft. An affordable way to store clean energy. URL: https://www.fraunhofer.de/en/press/research-news/2019/june/an-affordable-way-to-store-clean-energy.html, **2019**. Accessed: 2021-02-12.
[18] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, In *Proceedings of the 34th International Conference on Machine Learning* **2017**, 70, 1263–1272.
[19] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, S. A. Grimme, *Phys. Chem. Chem. Phys.* **2017**, *19*, 48, 32184–32215.
[20] K. Gubaev, E. V. Podryabinkin, A. V. Shapeev, *J. Chem. Phys.* **2018**, *148*, 24, 241727.
[21] J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R. Seress, C. Román-Salgado, K. Trepte, S. Atahan-Evrenk, S. Er, et al., *Energy Environ. Sci.* **2014**, *7*, 2, 698–704.
[22] B. A. Helfrecht, R. K. Cersonsky, G. Fraux, M. Ceriotti, *Mach. Learn.: Sci. Technol.* **2020**, *1*, 4, 045021.
[23] A. Jinich, B. Sanchez-Lengeling, H. Ren, R. Harman, A. Aspuru-Guzik, *ACS Cent. Sci.* **2019**, *5*, 7, 1199–1210.
[24] P. C. S. John, Y. Guan, Y. Kim, B. D. Etz, S. Kim, R. S. Paton, *Sci. Data* **2020**, *7*, 1, 1–6.
[25] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 8, 595–608.
[26] T. Kenning, Round-Up: 60MWh Japan project, Northern Ireland's 10MW array and Imergy goes for Africa teleco, https://www.energy-storage.news/news/round-up-60mwh-japan-project-northern-irelands-10mw-array-and-imergy-goes-f, **2016**. Accessed: 2021-02-12.
[27] H. Kim, J. Y. Park, S. Choi, *Sci. Data* **2019**, *6*, 1, 1–8.
[28] J. Knap, C. Spear, O. Borodin, K. Leiter, *Nanotechnology* **2015**, *26*, 43, 434004.
[29] M. Korth, *Phys. Chem. Chem. Phys.* **2014**, *16*, 17, 7919–7926.
[30] P. Leung, A. Shah, L. Sanz, C. Flox, J. Morante, Q. Xu, M. Mohamed, C. P. de Lĕn, *J. Power Sources* **2017**, *360*, 243–283.
[31] K. Lin, R. Gómez-Bombarelli, E. S. Beh, L. Tong, Q. Chen, A. Valle, A. Aspuru-Guzik, M. J. Aziz, R. G. Gordon, *Nat. Energy* **2016**, *1*, 9, 1–8.
[32] Y. Moon, Y.-K. Han, *Curr. Appl. Phys.* **2016**, *16*, 9, 939–943.
[33] M. Nakata, T. Shimazaki, M. Hashimoto, T. Maeda, *J. Chem. Inf. Model.* **2020**, *60*, 12, 5891–5899.
[34] H. Neugebauer, F. Bohle, M. Bursch, A. Hansen, S. Grimme, *J. Phys. Chem. A* **2020**, *124*, 35, 7166–7176.
[35] J. Noack, N. Roznyatovskaya, T. Herr, P. Fischer, *Angew. Chem. Int. Ed.* **2015**, *54*, 34, 9776–9809.
[36] Johnsen, R. D., NIST Computational Chemistry Comparison and Benchmark Database. In *NIST Standard Reference Database Number 101*, National Institute of Standards and Technology, Gaithersburg, Release 21, August **2020**.
[37] W. S. Ohlinger, P. E. Klunzinger, B. J. Deppmeier, W. J. Hehre, *J. Phys. Chem. A* **2009**, *113*, 10, 2165–2175.
[38] A. Orita, M. G. Verde, M. Sakai, Y. S. Meng, *Nat. Commun.* **2016**, *7*, 1, 1–8.
[39] G. A. Pinheiro, J. L. Da Silva, M. D. Soares, M. G. A. Quiles, In *International Conference on Computational Science and Its Applications*, Springer **2020**, 421–433.
[40] G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. Da Silva, M. G. Quiles, *J. Phys. Chem. A* **2020**, *124*, 47, 9854–9866.
[41] Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, T. F. Miller III, *J. Chem. Phys.* **2020**, *153*, 12, 124111.
[42] X. Qu, A. Jain, N. N. Rajput, L. Cheng, Y. Zhang, S. P. Ong, M. Brafman, E. Maginn, L. A. Curtiss, K. A. Persson, *Comput. Mater. Sci.* **2015**, *103*, 56–67.
[43] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Sci. Data* **2014**, *1*, 140022.
[44] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 11, 2864–2875.
[45] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 1, 1–8.
[46] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, K.-R. Müller, In *Proceedings of the 31st International Conference on Neural Information Processing Systems* **2017**, 992–1002.
[47] C. Schütter, T. Husch, V. Viswanathan, S. Passerini, A. Balducci, M. Korth, *J. Power Sources* **2016**, *326*, 541–548.
[48] J. S. Smith, O. Isayev, A. E. Roitberg, *Sci. Data* **2017**, *4*, 1, 1–8.
[49] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *J. Chem. Phys.* **2018**, *148*, 24, 241733.
[50] E. Sorkun, Q. Zhang, A. Khetan, S. Er, et al., *preprint from ChemRxiv, PPR: PPR310067* **2021**.
[51] S. Suresh, T. Kesavan, Y. Munaiah, I. Arulraj, S. Dheenadayalan, P. Ragupathy, *RSC Adv.* **2014**, *4*, 71, 37947–37953.
[52] D. P. Tabor, R. Gómez-Bombarelli, L. Tong, R. G. Gordon, M. J. Aziz, A. Aspuru-Guzik, *J. Mater. Chem. A* **2019**, *7*, 20, 12833–12841.
[53] N. Tokuda, M. Furuya, Y. Kikuoko, Y. Tsutui, T. Kumamoto, T. Kanno, In *Proceedings of the Power Conversion Conference-Osaka* **2002**, 3, 1144–1149.
[54] O. T. Unke, M. Meuwly, *J. Chem. Theory Comput.* **2019**, *15*, 6, 3678–3693.
[55] K. Wedege, E. Dražević, D. Konya, A. Bentien, *Sci. Rep.* **2016**, *6*, 1, 1–13.
[56] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 1, 31–36.
[57] J. Winsberg, T. Hagemann, T. Janoschka, M. D. Hager, U. S. Schubert, *Angew. Chem. Int. Ed.* **2017**, *56*, 3, 686–711.
[58] Y. Zeng, X. Zhou, L. An, L. Wei, T. Zhao, *J. Power Sources* **2016**, *324*, 738–744.
[59] F. Zhong, M. Yang, M. Ding, C. Jia, *Front. Chem.* **2020**, *8*, 451.

Manuscript received: March 2, 2021
Revised manuscript received: May 27, 2021
Accepted manuscript online: May 31, 2021
Version of record online: June 22, 2021