scientific **data**

Check for updates

DATA DESCRIPTOR

Quantum chemical calculations of lithium-ion battery electrolyte and interphase species

Evan Walter Clark Spotte-Smith [1,2,7], Samuel M. Blau[3,7], Xiaowei Xie [2,4], Hetal D. Patel[1,2], Mingjian Wen[1,2], Brandon Wood[5], Shyam Dwaraknath [2] & Kristin Aslaug Persson[1,6 ✉]

Lithium-ion batteries (LIBs) represent the state of the art in high-density energy storage. To further advance LIB technology, a fundamental understanding of the underlying chemical processes is required. In particular, the decomposition of electrolyte species and associated formation of the solid electrolyte interphase (SEI) is critical for LIB performance. However, SEI formation is poorly understood, in part due to insufficient exploration of the vast reactive space. The Lithium-Ion Battery Electrolyte (LIBE) dataset reported here aims to provide accurate first-principles data to improve the understanding of SEI species and associated reactions. The dataset was generated by fragmenting a set of principal molecules, including solvents, salts, and SEI products, and then selectively recombining a subset of the fragments. All candidate molecules were analyzed at the ωB97X-V/def2-TZVPPD/SMD level of theory at various charges and spin multiplicities. In total, LIBE contains structural, thermodynamic, and vibrational information on over 17,000 unique species. In addition to studies of reactivity in LIBs, this dataset may prove useful for machine learning of molecular and reaction properties.

Background & Summary

The solid electrolyte interphase (SEI), a nanoscale film that forms from electrolyte decomposition at the anodes of lithium-ion batteries (LIBs) during initial charging, is a critical component of modern rechargeable LIB electrolytes[1]. The SEI is responsible for the initial irreversible capacity loss during the battery's first several charge-discharge cycles[2]. At the same time, an appropriate self-limiting SEI, once formed, can protect against continuous electrolyte degradation while allowing Li-ion conduction[3]. In spite of the SEI's central importance to battery performance and lifespan, much remains unknown regarding the formation mechanisms of the SEI in LIBs. The SEI is formed as a result of numerous competitive reactive processes occurring simultaneously over time scales ranging from picoseconds[4] to days[5]. As a result of this extreme complexity, there are many open questions regarding the reaction pathways involved and even the products that form along those pathways[6,7].

It should be expected that many reactive intermediates that might arise during SEI formation will be so short-lived that they will be difficult or impossible to identify via experimental interrogations. However, even very reactive molecules can often be studied using first-principles quantum chemical methods such as density functional theory (DFT). Such computational simulations can therefore fill a gap and potentially provide a more fundamental understanding of the reactive chemistry of the SEI - for instance through the calculation of reaction free energies and energy barriers[8–12]. DFT can additionally be used to generate reference spectra that can be compared to experiment[13].

Here, we describe **LIBE**, the **Li**-**I**on **B**attery **E**lectrolyte dataset. LIBE includes non-polymeric and non-oligomeric molecules relevant to SEI formation in LIBs, with molecular properties such as optimized geometries, molecular thermochemistry, and vibrational spectra calculated using DFT. These molecules, which include both species previously reported in the literature as well as many novel species, could form at the SEI as a result of

[1]University of California Berkeley, Department of Materials Science and Engineering, Berkeley, CA, 94720, USA. [2]Lawrence Berkeley National Laboratory, Materials Science Division, Berkeley, CA, 94720, USA. [3]Lawrence Berkeley National Laboratory, Energy Storage and Distributed Resources, Berkeley, CA, 94720, USA. [4]University of California Berkeley, Department of Chemistry, Berkeley, CA, 94720, USA. [5]Lawrence Berkeley National Laboratory, National Energy Research Supercomputing Center, Berkeley, CA, 94720, USA. [6]Lawrence Berkeley National Laboratory, Molecular Foundry, Berkeley, CA, 94720, USA. [7]These authors contributed equally: Evan Walter Clark Spotte-Smith, Samuel M. Blau. ✉e-mail: kapersson@lbl.gov
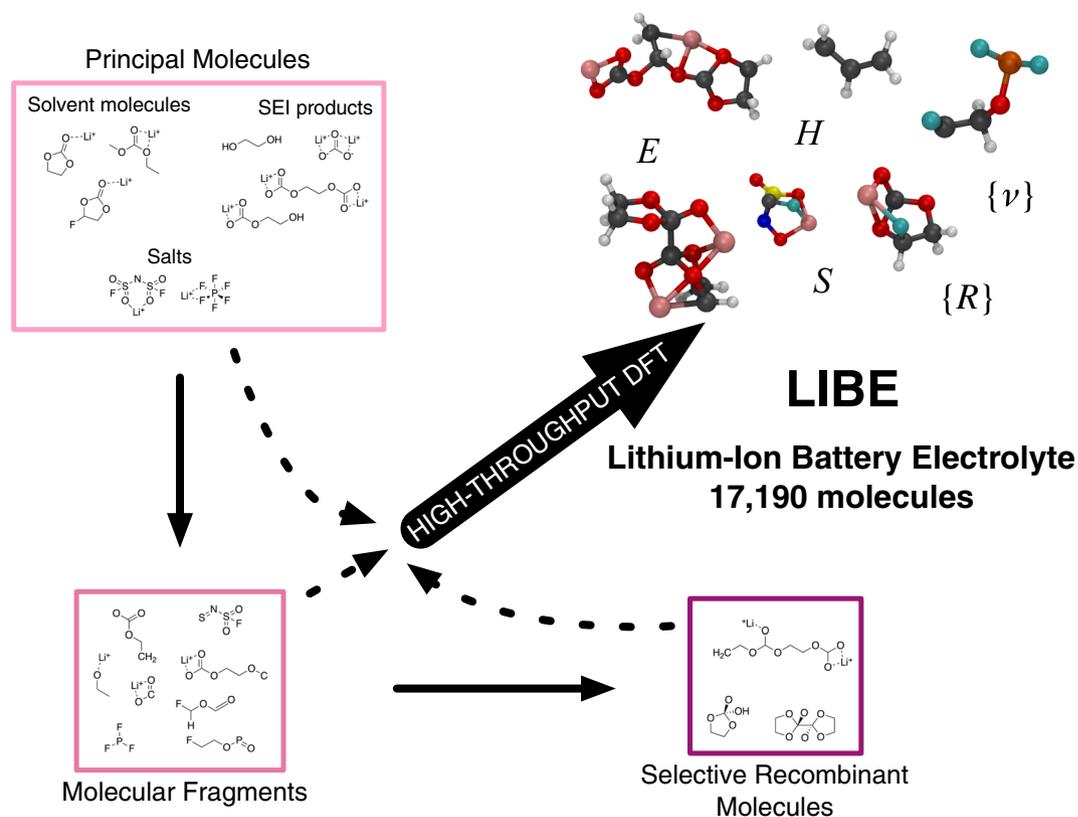
Fig. 1 Overview of the process underlying the generation of the Lithium-Ion Battery Electrolyte (LIBE) dataset. A set of principal molecules relevant to LIB SEI formation, including solvent molecules, electrolytes, and SEI products, were first selected. These molecules were then broken up into fragments, and these fragments were allowed to selectively recombine to form new, larger molecules. All principal molecules, fragments, and recombinant molecules were analyzed using high-throughput DFT, which provides an understanding of their structure and atomic coordinates {R}, thermodynamics - including energy E, enthalpy H, and entropy S - and vibrational frequencies {ν}.

electrolyte decomposition or the recombination of electrolyte fragments. The main purpose of LIBE is for studies of SEI formation and reactivity. Already, our group has used a subset of this data to generate a massive computational reaction network, identifying novel and chemically reasonable reactive pathways to a key SEI product, lithium ethylene dicarbonate (LEDC)[14]. It would be possible to take a similar approach to automatically identify pathways to other SEI products of interest, or perhaps to search for novel products not previously identified in experiments.

Far from being a single-use dataset of relevance only to SEI researchers, LIBE has the possibility of being used for broader studies of chemical reactions. For instance, the diverse molecules included in LIBE, including highly reactive and unstable species, provide an excellent dataset for machine learning models. We have recently used a subset of LIBE, which we called the "Bond Dissociation of Neutral and Charged Molecules" (BDNCM) dataset, to train a graph neural network called BonDNet[15]. BonDNet was able to predict heterolytic and homolytic bond dissociation energies with mean absolute error (MAE) far below chemical accuracy (0.022 eV vs. chemical accuracy of 0.043 eV).

The remainder of this Data Descriptor is organized as follows: first, we describe the computational methods used to both generate a set of candidate molecules and calculate their properties using DFT (Fig. 1). We then explain the choices made in designing the dataset, including the computational level of theory and the filters applied to ensure the quality of the data. After this explanation, we briefly characterize the LIBE dataset, examining the types of molecular species that it contains in terms of elements, bond types, charge, and spin multiplicity, among other factors. Finally, we describe the codes used to generate and analyze LIBE, all of which are freely available in open source repositories.

Methods

Overview.   Reactive organometallic molecules present significant challenges for computational analysis. Conventional methods to define molecular graph representations - necessary to define bonding and study molecular reactivity - are insufficient to capture coordinate bonds between Li^+ and heavy atoms like O, F, and N. In addition, DFT calculations involving highly reactive charged, radical, and metal-coordinated molecules frequently encounter errors or fail to converge to stable potential energy surface minima. Methods to address both of these challenges are here described.
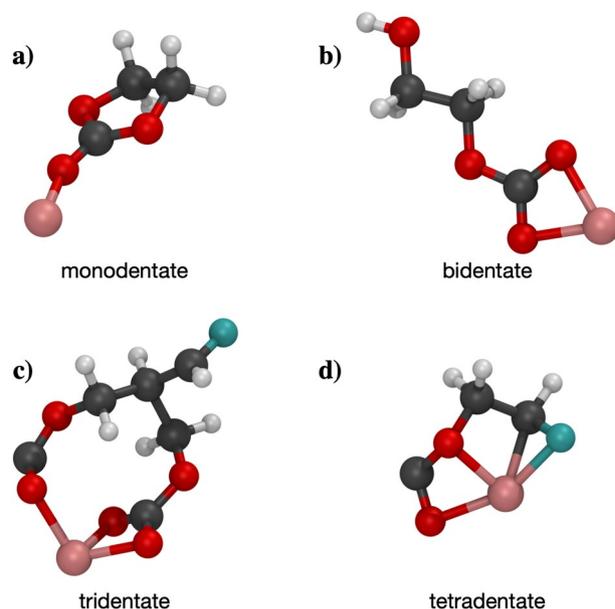
2

Fig. 2 Examples of molecules with various Li⁺ coordination environments: monodentate (**a**), bidentate (**b**), tridentate (**c**), and tetradentate (**d**). White atoms are hydrogen, gray atoms are carbon, red atoms are oxygen, blue are fluorine, and pink are lithium.

Data set construction is initialized with a small set of molecules known or previously proposed to participate in LIB SEI formation. From these "principal molecules", a set of molecular fragments were generated by recursively breaking bonds in the molecular graph representations. To explore molecular formation beyond what is currently known, a subset of these molecular fragments was then recombined, adding bonds between fragments to create new molecules. Through the application of fragmentation and recombination methods, a collection of molecules were created that could connect initial electrolyte components to final SEI products, allowing for the exploration of the reactive chemistry of LIB electrolytes. All molecules were analyzed using DFT to produce optimized geometries, molecular thermodynamics (including energy, enthalpy, entropy, and free energy), and vibrational data (including calculated infrared spectra).

Determination of bonding and molecular graph representations. Initially, bonding for all molecules was determined from 3D atomic coordinates using the bond detection algorithm defined in OpenBabel[16,17]. While this algorithm is well suited to the detection of covalent bonds, it is not designed to capture ionic bonds or coordinate bonds between metal ions and molecules[18]. Specifically, it is assumed in OpenBabel that Li⁺ will only form one bond. This is a critical issue for LIBE due to the crucially important and diverse coordination behavior of Li⁺. Li⁺ generally seeks to form between 4 and 6 coordinate bonds when in an electrolyte solution[13,19,20]. While often, Li⁺ forms only one coordinate bond with each coordinated molecule (Fig. 2a), cases where two (Fig. 2b), three (Fig. 2c), and even four (Fig. 2d) coordinate bonds form can occur. Because the thermodynamics of monodentate, bidentate, tridentate, and tetradentate configurations can vary significantly, it is essential to be able to distinguish between these bonding motifs. A modified bond detection algorithm was therefore required.

A heuristic method was used to add neglected coordinate bonds between Li and electronegative coordinating atoms, namely N, O, F, and S. If an N, O, F, or S atom is less than 2.5 away from a Li atom, then those two atoms were considered to be bonded. If, after this procedure, there were Li atoms in the molecule with no bonds, then the cutoff was increased from 2.5 to 3.5, and the procedure was repeated. Prior to performing DFT calculations, molecular connectivity was defined through first defining the bonding using OpenBabel and then applying this heuristic method.

In the final LIBE dataset, a quantum chemical method was also used to identify bonds. The `Critic2` program[21,22] was employed to identify bonding interactions in the electron densities of the optimized molecular geometry. `Critic2` identifies critical points in the electron density, which correspond to interatomic interactions. If the calculated field at a critical point between two atoms is greater than 0.02 (in atomic units) and if the distance between atoms is less than 2.5, then the two atoms are considered to be bonded. An exception is made for bonds between Li and C, for which a smaller field (greater than 0.012) is allowed. The final bonding for a molecule was defined by the union of the sets of bonds identified using OpenBabel, the heuristic coordinate bond detection method, and `Critic2`.

High-throughput computational methods. In order to be able to compute the properties of arbitrary molecules, including highly reactive fragments, radicals, and charged species, an automated framework was developed for high-throughput molecular DFT. This framework, which incorporates methods to correct common errors and ensure convergence to potential energy surface (PES) minima during molecular DFT calculations on the fly, was used to compute all properties of all molecules described in LIBE. Here, the computational methods

Parameter	Meaning	Value
ε	Dielectric constant	18.5
n	Refractive index	1.415
$\sum a_2^H$	Abraham's hydrogen-bond acidity	0.0
$\sum \beta_2^H$	Abraham's hydrogen-bond basicity	0.735
γ	Relative surface tension	20.2
ϕ	Carbon aromaticity	0.0
ψ	Electronegative halogenicity	0.0

Table 1. Solvent parameters for use in the SMD implicit solvent model. The dielectric constant ε represents a 3:7 weight blend of EC and EMC; all other parameters are for pure EC.
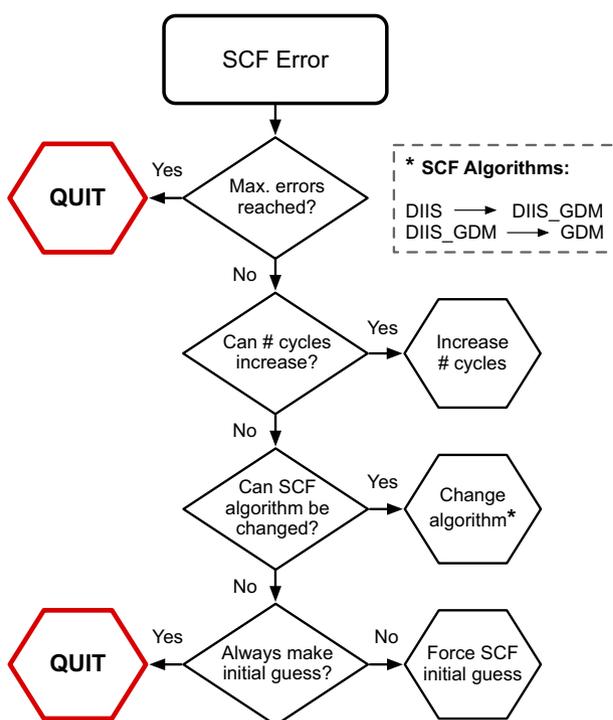

Fig. 3 A flowchart for correcting an SCF convergence error. When the error is encountered, only a single remedy will be applied. If there is no possible remedy, or if too many errors have already been encountered, then the error handler will quit, and the calculation will be allowed to fail.

used for high-throughput DFT calculations are described; an overview of how these methods were implemented in open source code bases is provided in Code Availability.

Calculation parameters. All calculations discussed in this Data Descriptor were performed using version 5.2.2 of the Q-Chem electronic structure code[23]. A large quadrature grid (SG-3) was used for all calculations[24], and the cutoff for the neglect of two-electron integrals is set to the tightest possible value (10^{-14}). Molecular symmetry was not used to improve calculation efficiency. Unless otherwise noted, with this exception, all Q-Chem default values (as of the 5.2.2 version) were used for initial calculations, though during error-correction these default values might be changed.

This dataset employs a level of theory based on the ωB97X-V density functional[25], which leverages the VV10 nonlocal van der Waals density functional[26] to accurately model noncovalent interactions. The def2-TZVPPD basis set[27,28] is employed, and solvation effects were included implicitly by means of the SMD method[29], which adds short-range energy contributions to the polarizable continuum model (PCM)[30,31]. The dielectric constant used ($\varepsilon = 18.5$) is that of a 3:7 ethylene carbonate (EC):ethyl methyl carbonate (EMC) mixture (a commonly used Li-ion electrolyte solvent blend). All other solvent parameters (see Table 1) are for pure EC[32,33].

Error correction. Once a calculation has terminated, its output file is parsed for errors. If any errors are detected, then an empirically designed recipe-based error correction process is conducted. If the error handler recognizes the error and an appropriate remedy is available, then that remedy will be employed and the calculation will be restarted automatically, generally with some alteration to the input parameters. If an error is encountered in the

re-started calculation, the same recipe-based error correction procedure is applied. If the error handler is unable to interpret the error, if all possible remedies have been exhausted, or if no remedy has been implemented for a particular error type, then the calculation fails.

Even if there are several possible remedies, only one remedy is applied at a time. The appropriate remedy for a given error may be sensitive to the parameters with which the calculation was run. Those parameters, in turn, may depend on the type and number of errors that the calculation has encountered previously.

To illustrate the error-correction process, Fig. 3 depicts the logic dictating how a convergence error for a self-consistent field (SCF) calculation should be remedied. The first possible remedy involves increasing the number of SCF cycles allowed; if the number of SCF cycles is lower than some maximum value (typically 200), then the number of cycles are increased to that maximum. If that remedy cannot be applied, either because it has already been applied or because the user specified a large number of SCF cycles initially, then the next remedy is to alter the SCF algorithm. The geometric direct minimization (GDM) method[34] tends to be highly robust at converging SCF calculations even for challenging molecules. However, because of its higher cost, the more rapid Direct Inversion of the Iterative Subspace method (DIIS)[35,36] or a combination of the two methods (DIIS_GDM in Q-Chem) are used first, with GDM serving as a method of last resort. Finally, the SCF settings are altered such that an initial guess electron density is generated for each SCF calculation, with no knowledge of prior calculations. Using the previous solution as a starting point for an SCF calculation can improve efficiency, but it can also fail to capture electronic state reordering in a newly visited region of the PES, occasionally resulting in SCF convergence problems. If none of these remedies can be applied, if all of them have been applied already, or if the number of errors encountered in total has exceeded a user-defined limit (for this dataset, chosen as 5), then the calculation will fail without further attempt to remedy the error.

In addition to SCF convergence errors, remedies have been implemented for a range of errors that might arise during a calculation (failing to optimize the molecular geometry, failing to transform from internal to Cartesian coordinates, failing to calculate the Hessian eigenvalues for a vibrational frequency calculation, etc.) or while preparing a calculation (failing to parse the input file, failing to access the DFT code executable file, failure to access a license file, etc.).

Convergence to potential energy surface minima. The goal of geometry optimization is to minimize the energy and to determine the stable molecular geometry. Generally, a optimizer will seek to reduce the gradient to zero, indicating that a stationary point has been found. However, convergence to a stationary point does not guarantee convergence to a local minimum of the PES; it is also possible to converge to an nth-order saddle point, where n is the number of imaginary frequencies. It is important to know when a calculation has converged to a saddle point and how to remedy it. Saddle points may provide poor approximations to the minimum energy structure and present significantly higher energy than the nearest minimum energy structure. Furthermore, saddle points can exhibit different bonding behavior from the minimum.

Most often, geometry optimization in DFT is conducted using a quasi-Newton-Raphson method; at each step, the energy and gradient are calculated, and the gradient is used to generate an approximation of the second derivative (Hessian) matrix[37]. While, in some methods, the exact Hessian is calculated at each step, this is prohibitively expensive in most cases and is therefore inappropriate for high-throughput applications. Because the Hessian used in quasi-Newton-Raphson optimization is not exact, the optimizer's knowledge of the curvature of the PES is limited. This makes it relatively common for geometry optimization algorithms to converge to saddle points instead of minima, especially for complex reactive fragments and/or species in an implicit solvent environment.

A method of "Frequency Flattening Optimization", or FFOpt, is used to eliminate imaginary frequencies. As illustrated in Fig. 4, successive optimization calculations are conducted until the structure has converged to a true local minimum of the PES. In order to determine if a converged structure is a PES minimum or a saddle point, a vibrational frequency calculation is performed following each completed optimization calculation. Frequency calculations serve a dual purpose, simultaneously providing information about the curvature of the PES (the exact Hessian) and the nature of the converged stationary point while also providing some thermodynamic information, including the molecular enthalpy and entropy. If there are no imaginary frequencies, then the structure is confirmed to be a PES minimum, and no further calculations are needed. If there are imaginary frequencies, then the structure is a PES saddle point. The exact Hessian reported by the frequency calculation is then used as input to the subsequent optimization calculation in order to provide a better description of the local PES and allow the optimizer to move away from the saddle point and towards a true minimum. This procedure can be repeated as many times as needed until a minimum is found. We emphasize that the FFOpt procedure, like most geometry optimization methods, aims to optimize to a local minimum and does not guarantee convergence to the global minimum of the PES.

Here, in order to limit the computational cost of an individual calculation, no more than 10 frequency flattening cycles were allowed. Moreover, additional cycles will not be pursued if there is only one imaginary mode with a very small frequency magnitude ($|\nu| \leq 15\,cm^{-1}$) or if the energy has changed by less than 10^{-7} Hartree (Ha) from the previous cycle to the current cycle, indicating that knowledge of the exact Hessian did not allow the optimization to leave the saddle point. Very small, singular imaginary frequencies are allowed because they may not correspond to true transition states; rather, they could be artifacts of numerical noise in the frequency calculation. If there is a single imaginary mode with a small frequency magnitude, the calculation is still considered a success; otherwise, a calculation which terminates with at least one imaginary frequency is considered a failure.

General calculation procedure. For a given set of unique molecular structures (as defined by the graph representations), FFOpt calculations were conducted at multiple charge states (-1, 0, and 1). For a particular charge state, when an even number of electrons was present, the molecule was initially assumed to be in a singlet state, and when an odd number of electrons was present, a doublet state was assumed.
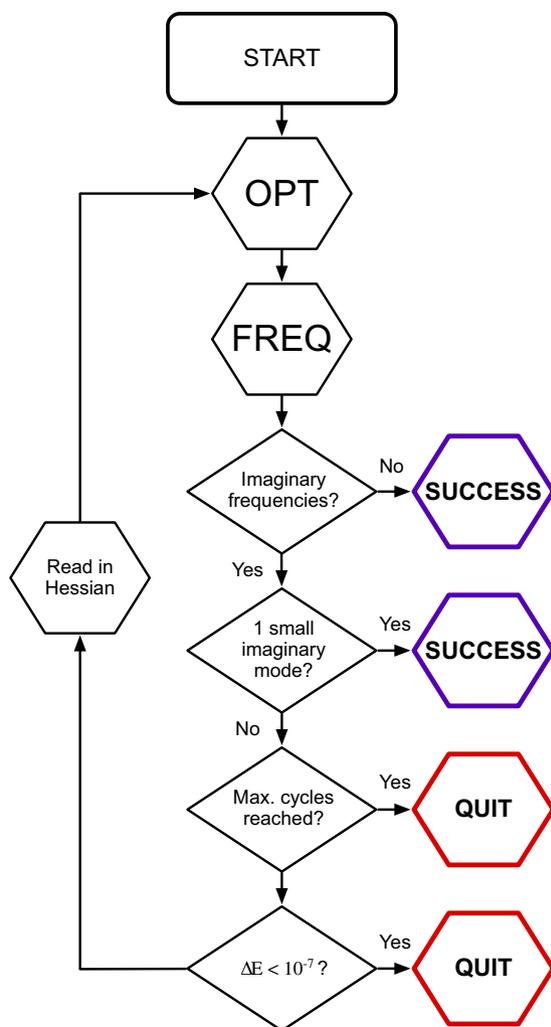
Fig. 4 The frequency-flattening optimization (FFOpt) procedure. In the initial step, the geometry is optimized and a vibrational frequency calculation is performed. If there are no imaginary frequencies, or if there is a single imaginary frequency with very small magnitude, the calculation completes successfully. Otherwise, the Hessian from the vibrational frequency calculation will be used to inform the next cycle of optimization.

Most commonly, low-spin states are preferred for molecular ground states, and stable high-spin states are rare[38]. This implies that one could expect that most molecules with even numbers of electrons should be in singlet states, rather than triplet states. However, triplets cannot be completely ignored, as there are exceptions (most notably diatomic oxygen) where triplet states are preferred at modest temperatures[39,40]. It is also possible that there exist species that are relevant to SEI formation which exhibit connectivity that can only exist as a triplet.

It would be computationally demanding to calculate all molecules and fragments in LIBE as both singlets and triplets. To balance computational cost and dataset diversity, only successfully optimized singlet molecules with less than 50 electrons were re-calculated as triplets. We note that this choice of cutoff is arbitrary, and there may be larger triplet species that are important to electrolyte or SEI formation reactions. Expanding the number of triplet species considered will be a future effort.

For all molecules for which the FFOpt procedure succeeded in identifying a PES minimum, a single-point calculation was conducted on the optimized geometry in order to produce a "cube" file of the electron density[41]. This cube file was then analyzed using `Critic2` to determine the critical points and improve the determination of the molecular bonding.

Note that single-atom (Li, H, F, etc.) calculations use a different procedure. Because geometry optimization is unnecessary for such molecules, only single-point calculations to determine the energy and frequency calculations to determine the translational enthalpy and entropy components were conducted.

This general procedure of conducting FFOpt singlet and doublet calculations, selected triplet calculations, and finally single-point calculations, was used in several stages to build the LIBE dataset. These stages are described in detail below in the Dataset generation section.

Dataset generation. *Selection of principal molecules.* The set of principal molecules was designed to adequately cover initial electrolyte molecules, experimentally identified SEI components, and other plausible intermediates or products that could arise during SEI formation in common LIB electrolytes. While many electrolyte
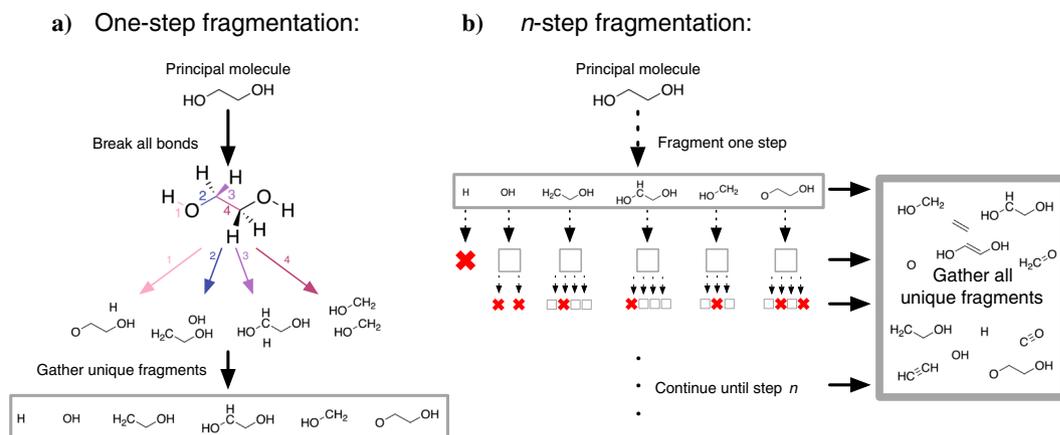
a) One-step fragmentation: **b)** *n*-step fragmentation:

Fig. 5 A schematic overview of a molecular fragmentation process. For a single-step fragmentation (**a**), a principal molecule is selected (in this case, principal molecule 70 in Supplementary Table 1). Each bond is broken, generating a collection of molecular fragments. These fragments are then filtered to generate a set of unique (non-isomorphic) molecules. In an n-step fragmentation (**b**), this process is repeated in a recursive fashion. At each step, all fragments from the previous step are collected and undergo a single-step fragmentation. If the fragment is a single atom with no bonds or if all fragments generated are already present in the collection, then the process terminates (red "X"). When the maximum number of steps has been reached, or when no new fragments can be generated, the n-step fragmentation terminates.

chemistries have been developed for use in LIBs, the most widely used formulations involve a fluorinated salt such as lithium hexafluorophosphate (LiPF6)[42–46], lithium bis(trifluoromethanesulfonyl)imide (LiTFSI)[20,47], or lithium bis(flurosulfonyl)imide (LiFSI)[48,49] dissolved in a solvent blend of cyclic carbonates such as ethylene carbonate (EC)[32,42], or fluroethylene carbonate (FEC)[13,50–52] and linear carbonates like dimethyl carbonate (DMC)[53,54], diethyl carbonate (DEC)[55,56], or ethyl methyl carbonate (EMC)[32,42]. At the current stage, LIBE contains molecules relevant to the electrolyte systems mentioned above (LiPF6, LiTFSI, LiFSI, EC, FEC, DMC, DEC, EMC) but does not, at this time, consider other electrolyte components or additives. However, we anticipate that the set will grow with more studies and applications. We note that we intend to incorporate all data in LIBE, and all future additions, in the Materials Project database[57,58].

The general strategy for selecting principal molecules was as follows: a set of electrolyte molecules and non-polymeric SEI products related to those electrolytes were selected from the literature. In some cases (especially for products derived from EC), these molecules were then modified in two ways: hydrogen atoms and lithium atoms bonded to oxygen could be substituted for one another, and hydrogen atoms bonded to carbon could be replaced by fluorine. The former substitution was guided by proposed reaction pathways in which hydrogen fluoride can attack Li-O bonds to produce -OH groups and LiF; the latter modification was chosen because of the inclusion of FEC, which can participate in many similar reaction pathways as EC. No conformer searching was conducted on principal molecules; initial structures that minimized steric hindrance were posed by hand and optimized. During initial geometry optimization, there were some cases in which multiple conformers with different Li coordination environments were identified. In such cases, all identified conformers were accepted as distinct principal molecules.

Representations of all principal molecules are provided in Supplementary Table 1. These can be grouped into solvent molecules (Molecule numbers 1–13), salt molecules (14–16), inorganic SEI products (17–26), possible dissolved minority species, including gases (27–35), lithium ethylene dicarbonate (LEDC) and related derivatives (36–39), lithium butylene dicarbonate (LBDC) and related derivatives (40–47), lithium ethylene monocarbonate (LEMC) and related derivatives (48–60), ethanol and related derivatives (61–62), ethylene glycol (EG) and related derivatives (63–70), 1,4-butanediol and related derivatives (71–73), other molecules related to LiEC decomposition (74–77), and other molecules related to PF6- decomposition (78–87).

Molecular fragmentation. Fragmentation begins with the molecular graph representation and 3D structure of a molecule of interest. In a single fragmentation step (Fig. 5a), each individual bond in the molecular graph is broken, generating either one or two fragments. In the case of a single fragment - indicating that the bond was part of a ring - an initial structure for the ring-opened fragment was generated using a low-cost optimization with the UFF force field as defined in OpenBabel. This preliminary optimization was conducted with the aim of preventing the ring from immediately re-closing during geometry optimization. In the case of two fragments, the coordinates associated with the atoms in each fragment were used as the initial structure. After all bonds have broken, all unique fragments - defined by their graph representations - were collected.

In most cases, it was desirable to not only obtain the products of single-bond cleavage, but all possible sub-fragments of a given molecule. This can be done by recursively applying the above single-step fragmentation method (Fig. 5b). At the *n*th step, all new structures from the *n*−1th step undergo a single-step fragmentation if possible (single atoms, which have no bonds, cannot be fragmented); the final set of fragments at that step is the union of the sets of unique fragments from each such single-step fragmentation. This recursive fragmentation can be continued until all fragments contain no bonds (at which point until only single atoms remain).
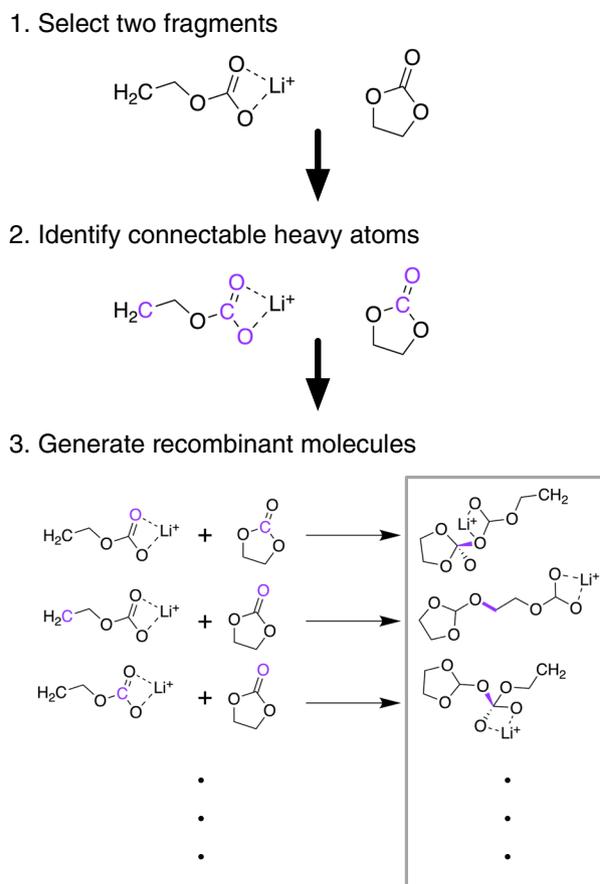
1. Select two fragments

2. Identify connectable heavy atoms

3. Generate recombinant molecules

Fig. 6 A simplified depiction of the recombination process. First (1), two fragments - in this case, from lithium ethylene carbonate, or principal molecule 1 - are selected. The heavy atoms in these molecules that can form additional bonds (shown in purple) are identified using valence rules (2), and finally, bonds (also in purple) are added between all combinations of these connectable heavy atoms (3) to form a set of unique recombinant molecules (gray box).

Supplementary Table 1 includes the number of fragmentation steps allowed for each principal molecule. In most cases, the number of steps was chosen such that all possible bonds were broken, indicated with "MAX". For larger molecules (with 20 or more atoms), computing the properties of all possible sub-fragments would be too computationally costly, hence a smaller number of steps was used. After fragmenting each principal molecule using the appropriate number of steps, all unique fragments - again defined by graph connectivity - were analyzed using DFT.

Generation of recombinant molecules. Reactions in LIB electrolytes involve not only bonds being broken but also bonds being formed. The set of principal molecules includes known products, implicitly accounting for some bond formations between possible fragments. However, fragmenting these principal molecules does not guarantee that all important intermediates or even all products are included. To improve the coverage of possible intermediate and product species and set the stage for new knowledge of SEI formation, some fragments (see criteria below) were allowed to recombine to form new molecules.

After all fragment species had been analyzed using DFT, a subset were selected for recombination. Specifically, all fragments from a two-step fragmentation of LiEC (principal molecule 1 in Supplementary Table 1) that could be formed exergonically from LiEC were included, as well as all fragments of H2O. All combinations of two fragments were recombined by adding a single bond in all possible ways that respect the typical valence rules of different atoms (Fig. 6). For instance, if one fragment has an oxygen connected to only one atom and one fragment has a carbon connected to only three atoms, then they would be allowed to combine. On the other hand, that same oxygen would not be allowed to combine with a carbon connected to four atoms. In applying these bonding rules, we do not count metal coordinate bonds and do not consider bond order (a single bond is treated on the same footing as a double or triple bond), but only consider the number of non-metal atoms connected to a given atom.

The recombinant molecules generated in this manner were further filtered by considering the reaction free energies of the recombination reactions. The BonDNet neural network was employed to predict the bond formation energies of the recombinant molecules. If the formation of the bond is predicted to be endergonic (the recombinant molecule was less stable than the constituent fragments), then the recombinant molecule was discarded. Initial guess structures of all remaining molecules were produced using the OPLS_2005 force field[59] as implemented in the Schrödinger Suite[60], and these initial structures were analyzed using DFT.

Key	Description
molecule_id	Unique identifier (format: `libe-XXXXXX`, where XXXXXX is a 6-digit number
bonds	List of pairs (a, b), where a and b are the 0-based indices of bonded atoms
charge	Charge of the molecule
chemical_system	Collection of elements present (ex: "C-H-O" for a molecule with C, H, and O present)
composition	Keys are elements; values are the number of atoms of those elements present
elements	List of elements present
formula_alphabetical	Simple chemical formula, with elements in alphabetical order (ex: "C4 H8 O1")
molecule	Serialized `pymatgen Molecule` object, containing species, coordinates, charge, and spin multiplicity
molecule_graph	Serialized `pymatgen MoleculeGraph` object; molecule with graph representation
number_atoms	Number of atoms in the molecule
number_elements	Number of unqiue elements present in the molecule
partial_charges	Atomic partial charges, calculated using various methods (Mulliken, RESP, `Critic2`)
partial_spins	For open-shell molecules, atomic partial spins, calculated with Mulliken population analysis
point_group	Molecular point group in Schönflies notation
species	Elements present at each atom in the molecule, in order
spin_multiplicity	Spin multiplicity $(2S + 1)$ of the molecule
thermo	Molecular thermodynamics, calculated from Q-Chem or a modified RRHO method
vibration	Calculated vibrational frequencies, and associated vibrational mode vectors and IR intensities
xyz	3D coordinates of the atoms in the molecule, in same order as "species"

Table 2. Description of keys present in LIBE dataset entries.

Note that the numerous stages of filtering used here - beginning with a small number of fragment molecules, requiring valence rules to be obeyed, and screening by bond formation energy - are necessary to limit the number of recombinant molecules considered. Recombination, as described here, is an inherently combinatorial process. Without appropriate filters, massive numbers of recombinant molecules can be generated, far too many to be calculated using high-accuracy DFT methods. We estimate that, if we attempted to recombine all fragment molecules included in LIBE, we would generate over 2,000,000 new molecules, which is completely intractable at our chosen level of theory. Future work will include efforts to recombine fragments of a larger set of principal molecules, most importantly salt species.

Final analysis. Molecular enthalpies, entropies, and free energies at 298.15K were calculated in multiple ways. The raw electronic energies, enthalpies, and entropies calculated in Q-Chem were used and are provided in the given units (Ha for electronic energy, kcal·mol^{-1} for enthalpy, and cal·mol^{-1}·K^{-1} for entropy), as well as in eV (or eV·K^{-1} for entropy). In addition, two different methods to correct for errors in the rigid-rotor harmonic oscillator (RRHO) approximation (used in Q-Chem) are employed: that of Ribiero et al.[61], in which low-frequency vibrational modes are shifted to some higher frequency (100 cm^{-1}) and that of Grimme[62], in which low-frequency modes are treated not as vibrations but as rotations. In all cases, imaginary frequencies are ignored for the purposes of calculating enthalpy, entropy, and free energy.

The point groups of all molecules were identified using the `PointGroupAnalyzer` tool implemented in `pymatgen`[63].

Data Records

Data for 17,190 molecules generated using an ωB97X-V/def2-TZVPPD/SMD level of theory are provided in a Figshare repository[64]. The data, including optimized 3D coordinates, partial charges and spins (from Mulliken population analysis[65], the Restrained Electrostatic Potential (RESP) method[66], and Critic2), molecular connectivity information, vibrational information (frequencies, vibrational mode vectors, IR intensities), and thermodynamic quantities (energy, enthalpy, entropy, Gibbs free energy), are contained in a single JSON-formatted file, `libe.json`. Molecules for which calculations failed or which otherwise failed the tests described in the Technical Validation section are not included in this collection.

Table 2 describes the keys in each entry of the `libe.json` file. Note that in some cases, keys may have no associated value; for instance, a single atom has no bonds.

Technical Validation

Level of theory. In order to maximize the utility of the LIBE dataset, a relatively costly but accurate level of theory was chosen. In an extensive benchmark study of density functionals by Mardirossian and Head-Gordon[67], ωB97X-V was found to be the most suitable hybrid generalized gradient approximation (hybrid GGA) functional, with exceptional accuracy for bonded interactions and noncovalent interactions. It is worth noting that ωB97X-V also displays high accuracy for calculation of barrier heights; while no transition states are included in LIBE, this is still beneficial, as it implies that the kinetic properties of reactions between molecules within the dataset could be reliably calculated without modification to the level of theory. While, to the best of our knowledge, no benchmark study has systematically examined how ωB97X-V performs for calculations involving charged, radical, and metal-coordinated species in solution, ωB97X-V has been shown to exhibit
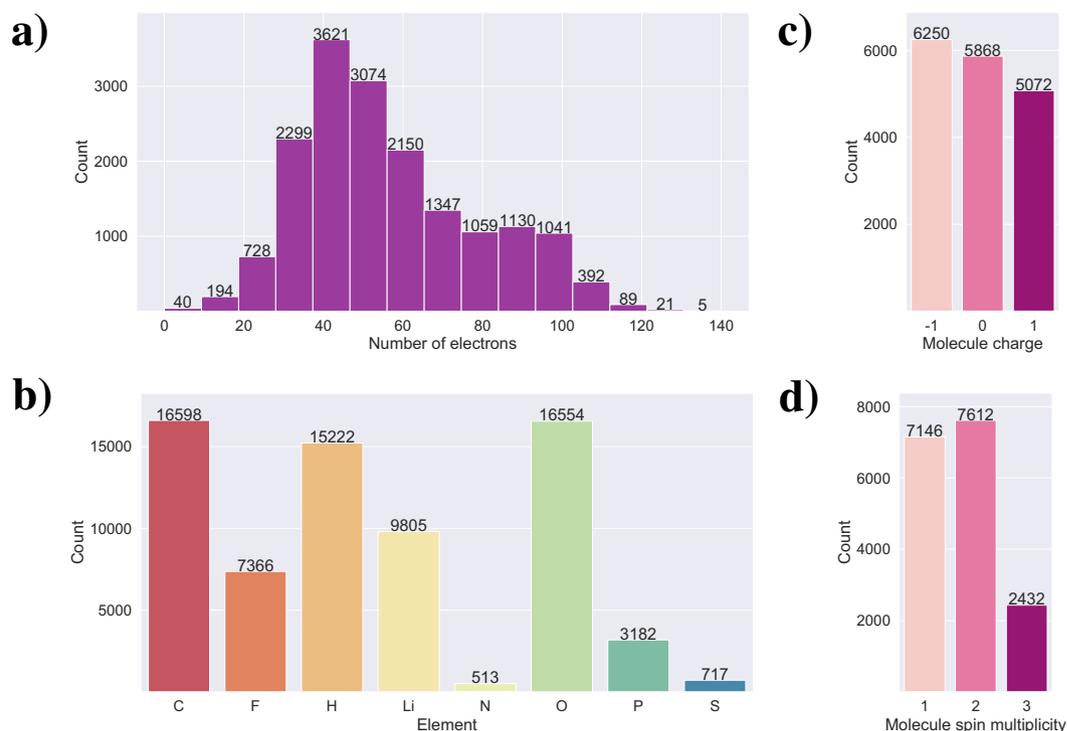
Fig. 7 An analysis of the composition of the LIBE dataset in terms of: number of molecules with different numbers of electrons (**a**); number of molecules with various elemental species (**b**); number of molecules with charges −1, 0, and 1 (**c**); and number of molecules with spin multiplicity 1, 2, and 3 (**d**).

exceptional performance for calculations involving transition metal complexes[68] and metal-organic reactions[69] in gas phase. Additionally, a previous *ab initio* molecular dynamics study[70] found that ωB97X-V was able to model aqueous solutions of NaCl more accurately than most density functionals, producing results in qualitative agreement with experiment. The benchmark study by Mardirossian and Head-Gordon found using a limited set of density functions that the def2-TZVPPD basis set performed nearly as well as the much larger def2-QZVPPD basis set[67], which makes it especially useful for high-throughput studies involving many thousands of calculations.

Generally, it should be expected that the use of an implicit solvation model should improve the accuracy of calculations involving molecules in solvent. Specifically, the SMx family of models, including the SMD model shown here, have been shown to accurately predict solvation free energies[29,71,72] as well as redox potentials[73], improving upon the more simple PCM models due to their inclusion of non-electrostatic effects.

We also justify our choice of level of theory by noting that similar levels of theory have previously been used to generate datasets used to study reactivity. In particular, Grambow *et al.*[74] recently used the ωB97X-D3 density functional[75] (which is closely related to ωB97X-V and differs primarily in the choice of dispersion correction) and the def2-TZVP basis set (which is part of the same family as def2-TZVPPD but contains no diffuse functions and fewer polarization functions) to create a dataset of over 12,000 organic reactions (including optimized reactants, products, and transition states) in vacuum. The solution-phase charged and radical organometallic chemistry involved in SEI formation is more complex than the gas-phase organic reactions considered by Grambow *et al.*, necessitating both the inclusion of an implicit solvent model and the use of a larger basis set including diffuse functions.

Data filtering. We note that the error correction procedures that we have employed are successful in decreasing the likelihood of failure in FFOpt calculations. Without intervention, roughly 25% of all calculations fail due to an error (for instance, inability to achieve a converged SCF solution or an inability to optimize a molecular geometry in the allowed number of steps), encounter a significant imaginary frequency (with magnitude > 15 cm^{-1}), or optimize to a structure with multiple disconnected fragments. With our error-handling procedures employed, this failure rate drops below 5% on average. In cases where error correction procedures were unable to eliminate issues, the calculations were not included in LIBE. While, in principle, single-point and `Critic2` calculations could also be a source of failure, in practice such calculations almost never failed when applied to optimized molecular structures.

Of the successful calculations that produced PES minima with connected structures, additional filters were put in place to ensure data quality and prevent duplicate molecules from being included in LIBE. First, molecules were eliminated if the energy of the molecule at the end of the geometry optimization differed from the energy calculated from the subsequent single-point calculation by more than 0.001 Hartree. Such a disagreement in energy implies that the single-point calculation converged to a different minimum of the electron density than was found at the end of the geometry optimization, potentially leading to inaccurate or inconsistent

Bond Type	Number of Bonds
C-C	25,744
C-F	6,002
C-H	53,178
C-Li	2,626
C-N	129
C-O	55,186
C-P	256
C-S	636
F-F	10
F-H	74
F-Li	1,285
F-O	150
F-P	4,604
F-S	195
H-H	4
H-Li	9
H-O	4,266
H-P	21
Li-Li	1
Li-N	53
Li-O	18,821
Li-P	19
Li-S	89
N-O	28
N-S	867
O-O	346
O-P	4,925
O-S	1,387
S-S	34

Table 3. Number of different types of bonds present in the LIBE dataset.

determination of bonding or atomic partial charges. Additionally, duplicate molecules were removed from the dataset. If two or more sets of calculations produced molecules that were non-equivalent (had different 3D coordinates) but with identical bonding, charge, and spin multiplicity, then only the molecule with the lowest calculated electronic energy was included in LIBE. Note that, while we did not explicitly perform any conformer searches, this filter implicitly selects the lowest-energy conformer that had been calculated.

Dataset diversity.   The LIBE dataset is designed for the study of (electro)chemical reactivity in LIB. As such, the most important consideration is whether the dataset adequately captures the possible molecules that could form in a LIB as a result of electrolyte decomposition. Considering that many common electrolyte molecules and most reported non-oligomeric/non-polymeric products derived from those molecules are among the principal molecules used to generate LIBE, we believe this is the case.

For uses outside of this domain, it is worth examining the chemical diversity of the LIBE dataset. While the dataset skews towards small molecules by design (both because most molecules examined are fragments of larger molecules and because large molecules would be computationally expensive), Fig. 7a shows that the distribution of molecules by size (measured by number of electrons) is wide; similar distributions are found when the size is measured by number of atoms and number of bonds.

Because most principal molecules are organic in nature and more specifically are derived from lithiated organic carbonates, the dataset is biased towards the C-H-O-Li chemical system, though with many (7,366) fluorine-containing molecules present as well (see Fig. 7b). While many phosphorus-containing molecules (3,182) are included, the bonding motifs (see Table 3) observed for phosphorus are limited (only F-P, O-P, and a small number of C-P, H-P, and Li-P bonds are present) because these molecules are all derived from PF6 and related molecules. We further note that the diversity in nitrogen- and sulfur-containing species is lacking because they are present only from TFSI- and FSI-based fragments.

While there are similar numbers of neutral molecules (5,868) and molecules with charge -1 (6,250), there are somewhat fewer molecules with charge $+1$ (5,072) (Fig. 7c). Because calculations were attempted for all initial molecule structures at charges -1, 0, and $+1$, this implies that the cationic species were more likely to fail than the anions or neutral species. There are slightly more doublet species (7,612) than singlets (7,146) (Fig. 7d). As discussed above, the number of triplets was intentionally kept low to reduce computational cost. We note that of the 1,961 pairs where singlet and triplet calculations optimized to isomorphic structures, the triplet was lower in electronic energy in 11.98% (235) of cases. Further, there are 471 triplet molecules for which no isomorphic
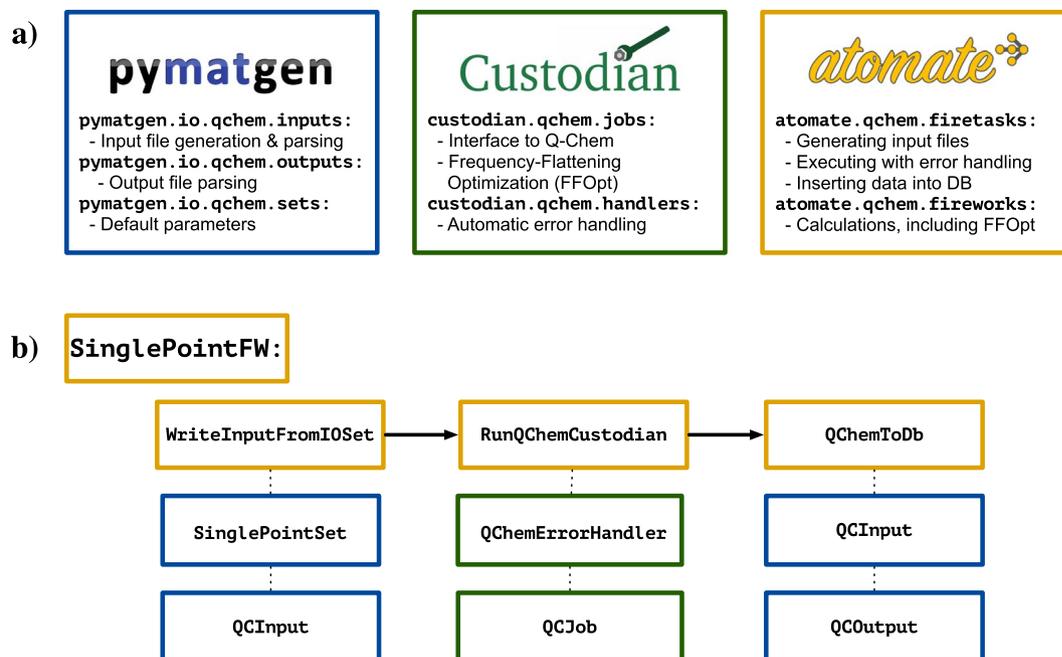
a)

pymatgen	**Custodian**	**atomate**

`pymatgen.io.qchem.inputs:`
 - Input file generation & parsing
`pymatgen.io.qchem.outputs:`
  - Output file parsing
`pymatgen.io.qchem.sets:`
  - Default parameters
	`custodian.qchem.jobs:`
	  - Interface to Q-Chem
	  - Frequency-Flattening
	    Optimization (FFOpt)
	`custodian.qchem.handlers:`
	  - Automatic error handling
		`atomate.qchem.firetasks:`
		  - Generating input files
		  - Executing with error handling
		  - Inserting data into DB
		`atomate.qchem.fireworks:`
		  - Calculations, including FFOpt

b) **SinglePointFW:**

WriteInputFromIOSet	→	RunQChemCustodian	→	QChemToDb
SinglePointSet		QChemErrorHandler		QCInput
QCInput		QCJob		QCOutput

Fig. 8 An overview of our automated high-throughput molecular DFT framework, as implemented in `pymatgen` (blue), `custodian` (green), and `atomate` (yellow) (**a**); an example calculation (Firework) for geometry optimization (**b**), indicating the different steps and the ways in which `pymatgen`, `custodian`, and `atomate` interact. First, the input file is written using default parameters defined in `pymatgen`. Then, the geometry optimization calculation is performed using the Q-Chem interface in `custodian` and an automated error handler. Finally, once the calculation is finished, the input and output files are parsed using `pymatgen`, and the results from the calculation are added to a database.

singlet with the same charge exists. Thus, it is possible that some number of unique structures, and some stable ground-states for existing structures, may be missing from LIBE. Because most often, the singlet structure is more stable than the triplet structure in the ground state, this lack of triplets should not be a significant detriment to the quality of the data.

Usage Notes

The `libe.json` file provided can be analyzed by any code capable of parsing JSON documents. The "molecule" and "molecule_graph" keys (see Table 2) are JSON representations of Python objects, and so Python-based analysis tools may be most convenient; however, the data stored in these objects is redundant, so this choice is not necessary.

We have created a software repository, `deliberate`, to aid in the use and analysis of the LIBE dataset. It includes the following files:

- `plotting.py`: Contains a utility function for making categorical bar plots and histograms
- `filters.py`: Contains functions for filtering the dataset
- `recombination.py`: Contains some basic code for molecular recombinantion
- `data_generation.ipynb`: A Jupyter Notebook providing a basic example of a fragmentation and recombination scheme to generate molecules from an initial set of principal molecules
- `dataset_composition.ipynb`: A Jupyter Notebook analyzing the composition of the LIBE dataset in some basic dimensions (bond types, molecule charge, molecule spin multiplicity, etc.)
- `filters.ipynb`: A Jupyter Notebook employing the filters in `filters.py`, which might be useful to tailor the dataset to a particular application.

Code availability

Our computational infrastructure for high-throughput and automated DFT calculations using the Q-Chem electronic structure code is implemented in existing open-source Python packages developed by the Materials Project, namely `pymatgen`[63], `custodian`, and `atomate`[76]. The modules in these codes used specifically for Q-Chem, along with their purposes, are described in Fig. 8a.

The basic functionality to generate, process, analyze, and manipulate molecules is included in `pymatgen`. We have added functionality to read and write Q-Chem input files and to parse Q-Chem output files. In addition, we have developed a number of "Sets", pre-defined collections of input parameters appropriate for common types of calculations. While these sets can be used with any level of theory available in Q-Chem, it is especially facile to use the advanced level of theory used for the LIBE dataset (ωB97X-V/def2-TZVPPD/SMD).

The `custodian` Q-Chem module defines the interface between Q-Chem and our automation framework in `atomate`. It can execute arbitrary Q-Chem jobs and can automatically check for, detect, and correct errors in Q-Chem calculations. `custodian` also handles the logic for FFOpt calculations.

The Q-Chem module in `atomate` combines the Q-Chem input and output modules in `pymatgen` and the Q-Chem interface and error handlers in `custodian` to perform Q-Chem jobs and analyze their data in a high-throughput fashion. An example calculation, or Firework, for a single-point optimization is shown schematically in Fig. 8b. First, based on some input parameters, a Q-Chem input file for a geometry optimization calculation is written. Then, the optimization job is run, with `custodian` waiting for completion and, upon completion, checking for errors. If the job completes without errors, then the output is parsed and stored in a database. Individual Q-Chem calculations, represented in `atomate` by Fireworks like `SinglePointFW`, can be combined to form more complex workflows.

Other than Q-Chem itself, all the necessary code used to generate and analyze the LIBE dataset (`pymatgen`: http://github.com/materialsproject/pymatgen; `custodian`: http://github.com/materialsproject/custodian; `atomate`: http://github.com/hackingmaterials/atomate; and `deliberate`: http://github.com/espottesmith/deliberate) can be found on Github.

References

1. Verma, P., Maire, P. & Novák, P. A review of the features and analyses of the solid electrolyte interphase in li-ion batteries. *Electrochimica Acta* **55**, 6332–6341 (2010).
2. Winter, M. The Solid Electrolyte Interphase – The Most Important and the Least Understood Solid Electrolyte in Rechargeable Li Batteries. *Zeitschrift für Physikalische Chemie* **223**, 1395–1406 (2009).
3. An, S. J. *et al.* The state of understanding of the lithium-ion-battery graphite solid electrolyte interphase (SEI) and its relationship to formation cycling. *Carbon* **105**, 52–76 (2016).
4. Leung, K. & Budzien, L. J. Ab initio molecular dynamics simulations of the initial stages of solid–electrolyte interphase formation on lithium ion battery graphitic anodes. *Phys. Chem. Chem. Phys.* **12**, 6583–6586 (2010).
5. Wood, D. L. III, Li, J. & Daniel, C. Prospects for reducing the processing cost of lithium ion batteries. *J. Power Sources* **275**, 234–242 (2015).
6. Wang, L. *et al.* Identifying the components of the solid–electrolyte interphase in Li-ion batteries. *Nat. Chem.* **11**, 789–796 (2019).
7. Rinkel, B. L. D., Hall, D. S., Temprano, I. & Grey, C. P. Electrolyte Oxidation Pathways in Lithium-Ion Batteries. *J. Am. Chem. Soc.* **142**, 15058–15074 (2020).
8. Wang, Y., Nakamura, S., Ue, M. & Balbuena, P. B. Theoretical studies to understand surface chemistry on carbon anodes for lithium-ion batteries: Reduction mechanisms of ethylene carbonate. *J. Am. Chem. Soc.* **123**, 11708–11718 (2001).
9. Wang, Y., Nakamura, S., Tasaki, K. & Balbuena, P. B. Theoretical studies to understand surface chemistry on carbon anodes for lithium-ion batteries: How does vinylene carbonate play its role as an electrolyte additive? *J. Am. Chem. Soc.* **124**, 4408–4421 (2002).
10. Leung, K. Two-electron reduction of ethylene carbonate: A quantum chemistry re-examination of mechanisms. *Chem. Phys. Lett.* **568–569**, 1–8 (2013).
11. Wang, A., Kadam, S., Li, H., Shi, S. & Qi, Y. Review on modeling of the anode solid electrolyte interphase (SEI) for lithium-ion batteries. *npj Comput. Mater.* **4**, 1–26 (2018).
12. Gibson, L. D. & Pfaendtner, J. Solvent oligomerization pathways facilitated by electrolyte additives during solid-electrolyte interphase formation. *Phys. Chem. Chem. Phys.* **22**, 21494–21503 (2020).
13. Hou, T. *et al.* The influence of FEC on the solvation structure and reduction reaction of LiPF6/EC electrolytes and its implication for solid electrolyte interphase formation. *Nano Energy* **64**, 103881 (2019).
14. Blau, S. M. *et al.* A chemically consistent graph architecture for massive reaction networks applied to solid-electrolyte interphase formation. *Chem. Sci.* **12**, 4931–4939 (2021).
15. Wen, M., Blau, S. M., Spotte-Smith, E. W. C., Dwaraknath, S. & Persson, K. A. Bondnet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem. Sci.* **12**, 1858–1868 (2021).
16. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J Cheminform* **3**, 33 (2011).
17. O'Boyle, N. M., Morley, C. & Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2**, 5 (2008).
18. Sayle, R. Pdb: Cruft to content. Presented at *MUG 2001* (2001).
19. Skarmoutsos, I., Ponnuchamy, V., Vetere, V. & Mossa, S. Li$^+$ Solvation in Pure, Binary, and Ternary Mixtures of Organic Carbonate Electrolytes. *J. Phys. Chem. C* **119**, 4502–4515 (2015).
20. Chapman, N., Borodin, O., Yoon, T., Nguyen, C. C. & Lucht, B. L. Spectroscopic and Density Functional Theory Characterization of Common Lithium Salt Solvates in Carbonate Electrolytes for Lithium Batteries. *J. Phys. Chem. C* **121**, 2135–2148 (2017).
21. Otero-de-la Roza, A., Blanco, M. A., Pendás, A. M. & Luaña, V. Critic: a new program for the topological analysis of solid-state electron densities. *Comput. Phys. Commun.* **180**, 157–166 (2009).
22. Otero-de-la Roza, A., Johnson, E. R. & Luaña, V. Critic2: A program for real-space analysis of quantum chemical interactions in solids. *Comput. Phys. Commun.* **185**, 1007–1018 (2014).
23. Shao, Y. *et al.* Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **113**, 184–215 (2015).
24. Dasgupta, S. & Herbert, J. M. Standard grids for high-precision integration of modern density functionals: Sg-2 and sg-3. *J. Comput. Chem.* **38**, 869–882 (2017).
25. Mardirossian, N. & Head-Gordon, M. wB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **16**, 9904–9924 (2014).
26. Vydrov, O. A. & Van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *J. Chem. Phys.* **133**, 244103 (2010).
27. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).
28. Rappoport, D. & Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **133**, 134105 (2010).
29. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **113**, 6378–6396 (2009).
30. Tomasi, J., Mennucci, B. & Cammi, R. Quantum mechanical continuum solvation models. *Chem. reviews* **105**, 2999–3094 (2005).

31. Mennucci, B. Polarizable continuum model. *Wiley Interdisc. Rev. Comput. Mol. Sci.* **2**, 386–404 (2012).
32. Hall, D. S., Self, J. & Dahn, J. R. Dielectric Constants for Quantum Chemistry and Li-Ion Batteries: Solvent Blends of Ethylene Carbonate and Ethyl Methyl Carbonate. *J. Phys. Chem. C* **119**, 22322–22330 (2015).
33. Qu, X. *et al*. The electrolyte genome project: A big data approach in battery materials discovery. *Comput. Mater. Sci.* **103**, 56–67 (2015).
34. Van Voorhis, T. & Head-Gordon, M. A geometric approach to direct minimization. *Mol. Phys.* **100**, 1713–1721 (2002).
35. Pulay, P. Convergence acceleration of iterative sequences. the case of scf iteration. *Chem. Phys. Lett.* **73**, 393–398 (1980).
36. Pulay, P. Improved SCF convergence acceleration. *J. Comput. Chem.* **3**, 556–560 (1982).
37. Schlegel, H. B. Geometry optimization. *Wiley Interdisc. Rev. Comput. Mol. Sci.* **1**, 790–809 (2011).
38. Gallagher, N. *et al*. Thermally and Magnetically Robust Triplet Ground State Diradical. *J. Am. Chem. Soc.* **141**, 4764–4774 (2019).
39. Klán, P. & Wirz, J. *Photochemistry of Organic Compounds: from Concepts to Practice*, John Wiley & Sons, Hoboken (2009).
40. Shavitt, I. Geometry and singlet-triplet energy gap in methylene: a critical review of experimental and theoretical determinations. *Tetrahedron* **41**, 1531–1542 (1985).
41. Herbert, J. M. The quantum chemistry of loosely bound electrons. *Rev. Comput. Chem.* **28**, 391–517 (2015).
42. Zhuang, G. V., Xu, K., Yang, H., Jow, T. R. & Ross, P. N. Lithium Ethylene Dicarbonate Identified as the Primary Product of Chemical and Electrochemical Reduction of EC in 1.2 M LiPF6/EC:EMC Electrolyte. *J. Phys. Chem. B* **109**, 17567–17573 (2005).
43. Nie, M. *et al*. Lithium Ion Battery Graphite Solid Electrolyte Interphase Revealed by Microscopy and Spectroscopy. *J. Phys. Chem. C* **117**, 1257–1267 (2013).
44. Okamoto, Y. Ab Initio Calculations of Thermal Decomposition Mechanism of LiPF6-Based Electrolytes for Lithium-Ion Batteries. *J. Electrochem. Soc.* **160**, A404 (2013).
45. Parimalam, B. S., MacIntosh, A. D., Kadam, R. & Lucht, B. L. Decomposition Reactions of Anode Solid Electrolyte Interphase (SEI) Components with LiPF6. *J. Phys. Chem. C* **121**, 22733–22738 (2017).
46. Solchenbach, S., Metzger, M., Egawa, M., Beyer, H. & Gasteiger, H. A. Quantification of PF5 and POF3 from Side Reactions of LiPF6 in Li-Ion Batteries. *J. Electrochem. Soc.* **165**, A3022 (2018).
47. Seitzinger, C. L. *et al*. Intrinsic Chemical Reactivity of Silicon Electrode Materials: Gas Evolution. *Chem. Mater.* **32**, 3199–3210 (2020).
48. Kang, S.-J., Park, K., Park, S.-H. & Lee, H. Unraveling the role of LiFSI electrolyte in the superior performance of graphite anodes for Li-ion batteries. *Electrochimica Acta* **259**, 949–954 (2018).
49. Liu, S. *et al*. LiFSI and LiDFBOP Dual-Salt Electrolyte Reinforces the Solid Electrolyte Interphase on a Lithium Metal Anode. *ACS Appl. Mater. Interfaces* **12**, 33719–33728 (2020).
50. Xia, J., Petibon, R., Xiao, A., Lamanna, W. M. & Dahn, J. R. Some fluorinated carbonates as electrolyte additives for li(ni$_{0.4}$mn$_{0.4}$co$_{0.2}$)o$_2$/graphite pouch cells. *J. Electrochem. Soc.* **163**, A1637–A1645 (2016).
51. Xia, L. *et al*. Oxidation decomposition mechanism of fluoroethylene carbonate-based electrolytes for high-voltage lithium ion batteries: a dft calculation and experimental study. *Chem.* **2** (2017).
52. Intan, N. & Pfaendtner, J. Effect of fluoroethylene carbonate additive on the initial formation of solid electrolyte interphase on oxygen functionalized graphitic anode in lithium ion batteries. *ACS Appl. Mater. Interfaces* **13**, 8169–8180 (2021).
53. Aurbach, D., Markovsky, B., Shechter, A., Ein-Eli, Y. & Cohen, H. A comparative study of synthetic graphite and li electrodes in electrolyte solutions based on ethylene carbonate-dimethyl carbonate mixtures. *J. Electrochem. Soc.* **143**, 3809 (1996).
54. Hobold, G. M., Khurram, A. & Gallant, B. M. Operando gas monitoring of solid electrolyte interphase reactions on lithium. *Chem. Mater.* **32**, 2341–2352 (2020).
55. Aurbach, D. *et al*. The study of electrolyte solutions based on ethylene and diethyl carbonates for rechargeable li batteries: I. li metal anodes. *J. The Electrochem. Soc.* **142**, 2873 (1995).
56. Aurbach, D. *et al*. The study of electrolyte solutions based on ethylene and diethyl carbonates for rechargeable li batteries: Ii. graphite electrodes. *J. The Electrochem. Soc.* **142**, 2882 (1995).
57. Jain, A. *et al*. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1** 011002 (2013).
58. Jain, A. *et al*. The materials project: Accelerating materials design through theory-driven data and tools. In Andreoni, W. & Yip, S. (eds.) *Handbook of Materials Modeling*, 1–34, Springer International Publishing, Cham (2018).
59. Banks, J. L. *et al*. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **26**, 1752–1780 (2005).
60. Schrödinger Python API. http://content.schrodinger.com/Docs/r2020-4/python_api/overview.html (2020).
61. Ribeiro, R. F., Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Use of Solution-Phase Vibrational Frequencies in Continuum Models for the Free Energy of Solvation. *J. Phys. Chem. B* **115**, 14556–14562 (2011).
62. Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. – A Eur. J.* **18**, 9955–9964 (2012).
63. Ong, S. P. *et al*. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
64. Spotte-Smith, EWC. *et al*. Lithium-ion battery electrolyte (libe) dataset, *figshare*, https://doi.org/10.6084/m9.figshare.14226464 (2021).
65. Mulliken, R. S. Electronic population analysis on lcao–mo molecular wave functions. i. *J. Chem. Phys.* **23**, 1833–1840 (1955).
66. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The J. Phys. Chem.* **97**, 10269–10280 (1993).
67. Mardirossian, N. & Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **115**, 2315–2372 (2017).
68. Chan, B., Gill, P. M. W. & Kimura, M. Assessment of DFT Methods for Transition Metals with the TMC151 Compilation of Data Sets and Comparison with Accuracies for Main-Group Chemistry. *J. Chem. Theory Comput.* **15**, 3610–3622, https://doi.org/10.1021/acs.jctc.9b00239 (2019).
69. Dohm, S., Hansen, A., Steinmetz, M., Grimme, S. & Checinski, M. P. Comprehensive Thermochemical Benchmark Set of Realistic Closed-Shell Metal Organic Reactions. *J. Chem. Theory Comput.* **14**, 2596–2608, https://doi.org/10.1021/acs.jctc.7b01183 (2018).
70. Yao, Y. & Kanai, Y. Free Energy Profile of NaCl in Water: First-Principles Molecular Dynamics with SCAN and B97X-V Exchange–Correlation Functionals. *J. Chem. Theory Comput.* **14**, 884–893, https://doi.org/10.1021/acs.jctc.7b00846 (2018).
71. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. *J. Phys. Chem. B* **113**, 4538–4543 (2009).
72. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Generalized Born Solvation Model SM12. *J. Chem. Theory Comput.* **9**, 609–620 (2013).
73. Guerard, J. J. & Arey, J. S. Critical Evaluation of Implicit Solvent Models for Predicting Aqueous Oxidation Potentials of Neutral Organic Compounds. *J. Chem. Theory Comput.* **9**, 5046–5058 (2013).
74. Grambow, C. A., Pattanaik, L. & Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Sci. data* **7**, 1–8 (2020).
75. Lin, Y.-S., Li, G.-D., Mao, S.-P. & Chai, J.-D. Long-range corrected hybrid density functionals with improved dispersion corrections. *J. Chem. Theory Comput.* **9**, 263–272 (2013).
76. Mathew, K. *et al*. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Comput. Mater. Sci.* **139**, 140–152 (2017).

Author contributions

E.W.C.S.-S., S.M.B., B.W. and S.D. developed the high-throughput framework used to generate this data; S.M.B. developed the fragmentation method; X.X. developed the recombination method; E.W.C.S.-S., S.M.B. and H.D.P. selected the principle molecules for study; E.W.C.S.-S., S.M.B., X.X. and H.D.P. generated raw data; X.X. and M.W. analyzed recombinant molecules; E.W.C.S.-S. and M.W. developed the analysis code in the `deliberate` library; E.W.C.S.-S. and S.M.B. processed the data and generated the final dataset; E.W.C.S.-S. wrote the original manuscript; S.M.B. and K.A.P. conceived of the study; K.A.P. secured funding. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-021-00986-9.

Correspondence and requests for materials should be addressed to K.A.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.