The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Machine learned calibrations to high-throughput molecular excited state calculations

Cite as: J. Chem. Phys. 156, 134116 (2022); doi: 10.1063/5.0084535
Submitted: 7 January 2022 • Accepted: 14 March 2022 •
Published Online: 7 April 2022

Shomik Verma,[1,a)] Miguel Rivera,[2] David O. Scanlon,[3] and Aron Walsh[1,b)]

AFFILIATIONS

[1] Department of Materials, Imperial College London, Exhibition Road, London SW7 2AZ, United Kingdom
[2] Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, United Kingdom
[3] Department of Chemistry and Thomas Young Centre, University College London, 20 Gordon Street, London WC1H 0AJ, United Kingdom

Note: This paper is part of the JCP Special Topic on Chemical Design by Artificial Intelligence.
[a)] **Also at:** Department of Mechanical Engineering, Massachusetts Institute of Technology, 33 Massachusetts Ave, Cambridge, MA 02139, USA.
[b)] **Author to whom correspondence should be addressed:** a.walsh@imperial.ac.uk

ABSTRACT

Understanding the excited state properties of molecules provides insight into how they interact with light. These interactions can be exploited to design compounds for photochemical applications, including enhanced spectral conversion of light to increase the efficiency of photovoltaic cells. While chemical discovery is time- and resource-intensive experimentally, computational chemistry can be used to screen large-scale databases for molecules of interest in a procedure known as high-throughput virtual screening. The first step usually involves a high-speed but low-accuracy method to screen large numbers of molecules (potentially millions), so only the best candidates are evaluated with expensive methods. However, use of a coarse first-pass screening method can potentially result in high false positive or false negative rates. Therefore, this study uses machine learning to calibrate a high-throughput technique [eXtended Tight Binding based simplified Tamm-Dancoff approximation (xTB-sTDA)] against a higher accuracy one (time-dependent density functional theory). Testing the calibration model shows an approximately sixfold decrease in the error in-domain and an approximately threefold decrease in the out-of-domain. The resulting mean absolute error of ~0.14 eV is in line with previous work in machine learning calibrations and out-performs previous work in linear calibration of xTB-sTDA. We then apply the calibration model to screen a 250k molecule database and map inaccuracies of xTB-sTDA in chemical space. We also show generalizability of the workflow by calibrating against a higher-level technique (CC2), yielding a similarly low error. Overall, this work demonstrates that machine learning can be used to develop a cost-effective and accurate method for large-scale excited state screening, enabling accelerated molecular discovery across a variety of disciplines.

I. INTRODUCTION

Understanding the excited state properties of molecules helps describe how they interact with light. These photochemical interactions can include fundamental processes, such as photosynthesis,[1] human vision,[2] or photostability.[3] Photochemistry is also important in designing new molecules with certain properties, for example, spectral converters for photovoltaics,[4] which are of particular interest in this study. Using an interplay between their excited states, certain molecules can up- or down-convert wavelengths of light to improve photovoltaic efficiency. Unfortunately, existing spectral conversion molecules have low efficiency,[5] so further exploration is required. It is difficult to explore the excited state space of molecules with experimental methods alone, so researchers often turn to computational methods for a more detailed study.[6,7] Traditionally, time-dependent density functional theory (TD-DFT) has been the workhorse for excited state energy calculations. However, TD-DFT can be computationally intensive, and for applications in high-throughput screening or generative design, a faster method must be used.

TD-DFT often relies on DFT for ground-state calculations of charge density and structure optimization. Recently, work has been done in tight binding as an approximation to DFT to improve its computation time while retaining most of its accuracy. Specifically, density functional tight binding (DFTB)[8] was developed in the late 1990s[9] and exhibited a combination of the accuracy of DFT and the efficiency of semi-empirical quantum chemistry methods. More recently, eXtended Tight Binding (xTB)[10] methods were developed to solve the issues with DFTB of extensive parameterization and low transferability.[8] They differ from DFTB methods in that they utilize top-down parameterization, with semiempirical parameters fit to a large dataset rather than computed with first-principles calculations.[10] The primary approximations are considering molecular orbitals to be a linear combination of atomic orbitals (LCAOs), using the local density approximation (LDA) for exchange–correlation energy, and using a truncated Taylor expansion to map density to total energy.[10]

To accelerate excited-state calculations, Grimme introduced the simplified Tamm–Dancoff density functional approach (sTDA)[11] as an approximation to TD-DFT. The key approximations of sTDA include simplifications to two-electron integrals and setting an upper limit to the excitation space, which improve computation time by 2 orders of magnitude.[11] As sTDA was developed to calculate excitation spectra, there is no excited state relaxation component, and only vertical excitation energies can be calculated. The differences between vertical excitation, vertical emission, and adiabatic energy are shown in Fig. 1. In this work, only vertical excitations are considered. The additional computational expense of excited state relaxation is prohibitive for high-throughput workflows. Furthermore, the Stokes shift for rigid molecules should be small, on the order of 0.1 eV.[12] From now on, for concision, the vertical excitation energy will be referred to as the excitation energy or excited state energy.

xTB and sTDA can be combined in a workflow called xTB-sTDA, allowing for ultrafast computation of excited states.[13] This has been used extensively, with several studies using the method to screen large databases of materials, such as copolymers,[14] conjugated polymers,[15] small aromatic molecules,[16] photocatalysts,[17] and organic dyes.[18]

However, due to the approximations presented above, there is a trade-off between accuracy and computational speed. In the original paper of Grimme and Bannwarth introducing xTB-sTDA, they reported a mean absolute error (MAE) between xTB-sTDA and coupled-cluster/TD-DFT calculated excited state energies of 0.34–0.48 eV depending on the complexity of the input structure.[13] Even though xTB-sTDA is often used as a first-pass in high-throughput screening, with higher-quality computational methods used to evaluate properties of a screened subset of molecules, having an accurate first-pass method is essential to ensure that all suitable candidates are included in the suggested subset.

Therefore, a method of improving the accuracy of xTB-sTDA in a way that preserves its high-throughput characteristics is desired. One approach is to calibrate the results of xTB-sTDA against a higher-accuracy computational method using machine learning (ML). This type of calibration from a baseline method to a reference method is known as the delta-ML or Δ-ML approach[19] and has widely been applied in the literature for various computational techniques. For example, it has been used for calibrating

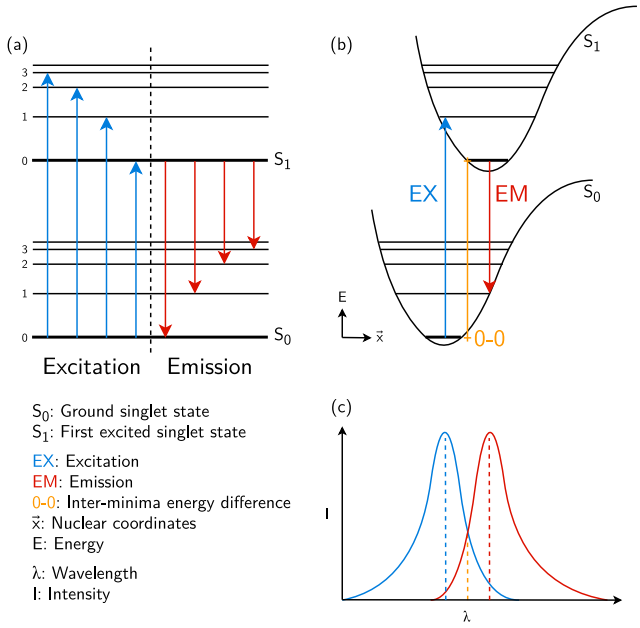

FIG. 1. Schematics depicting the basics of molecular excitation and emission. (a) Typical Jablonksi diagram from the ground state (S_0) to excited state (S_1) with various vibrational levels (0–3) depicted for both states. (b) Demonstration of the Frank–Condon principle of 0 → 1 vertical excitation (blue arrow) followed by nuclear re-configuration and 1 ← 0 vertical emission (red arrow). Also shown is the 0 → 0 transition energy in yellow. (c) The expected experimental excitation/absorption curve (blue) and emission curve (red), along with the theoretical 0 → 0 energy difference (dashed yellow line), demonstrating the Stokes shift. Note that in computation, often the energy minimum is used instead of the lowest vibrational level, so the starting energies for excitation and emission may be different than experimental values.

ground state energies and structures from semiempirical methods to DFT and coupled cluster (CC) accuracy[20] and calibrating DFT molecular dynamics simulations[21] or potential energy surfaces[22] to CC accuracy. Recently, the Δ-ML approach has been applied to calibrate excited state properties, for example, calibrating photoemission spectra from DFT to G_0W_0[23] and calibrating excited state energies with TD-DFT against CC methods[24] and against the experiment.[25–29] For xTB-sTDA, a few studies have used a linear calibration technique to correct excited state energies,[15,16,18] instead of ML. However, the improvement from linear calibration was low, with a mean absolute error (MAE) of around 0.2 eV, compared to around 0.1 eV for the ML studies presented above. To the authors' knowledge, no previous study has applied ML to calibrate xTB-sTDA.

Due to the promising potential of ML to increase the accuracy of baseline methods, this work presents a ML calibration of the excited state energy levels output by xTB-sTDA, with the motivation of more efficient exploration of excited state space. As mentioned previously, existing spectral conversion materials utilize excited states to up- or down-convert photon energy, but have low efficiency due to (a) energy level misalignment, which leads to energy loss and (b) low absorption-to-emission probability. This study focuses on the first issue, making it easier to accurately predict energy levels

and, therefore, design high-efficiency spectral conversion materials. The following sections present the methods (Sec. II) and results (Sec. III) of our excited state energy calibrations.

II. METHODS

A. Reference computational technique

xTB was originally parameterized from the spin-component-scaled coupled cluster (SCS-CC2)[30] method and TD-DFT, so one of these would be a natural choice as the reference computational technique. Coupled-cluster (CC) methods are typically the most accurate, predicting excitation energies within 0.1 eV of the experimental values.[31] However, because of the computational expense of CC, it is difficult to generate a large amount of data using these methods. Instead, TD-DFT is generally the workhorse for excited state calculation, despite its relatively high errors, with an MAE of 0.2–0.4 eV (depending on the functional) compared to experimental values or theoretical best estimates.[31]

A general-purpose exchange–correlation functional is B3LYP,[32] which has been shown to have an MAE around 0.25 eV, while LDA, generalized gradient approximation (GGA), and other hybrid functionals have higher error.[33] While B3LYP performs well for localized densities, range-separated hybrid functionals, such as CAM-B3LYP,[34] are increasingly used for delocalized densities, such as those in excitations, as they include a long-range correction.[35] However, these range-separated hybrid functionals are more computationally expensive than B3LYP. Due to its accuracy and efficiency, B3LYP is the most commonly used functional in computational molecular chemistry.[36]

In this work, we choose TD-DFT based on B3LYP as the reference method for several reasons. First, we require large chemical diversity in our training set, and most of the existing molecular excited state databases use TD-DFT.[24,37–40] The largest databases, namely, PubChemQC[38] and QM-symex,[41] use B3LYP. Second, B3LYP was used in previous works using linearly calibrated xTB-sTDA[15,16] and is used extensively in machine learning and high-throughput screening studies.[42–46] Third, while B3LYP is less accurate than range-separated hybrid functionals, it is not significantly worse.[35,47] Since xTB-sTDA is semi-empirical, it is often used as a first-pass screening and naturally has inconsistencies and false positive/negative errors. Calibrating to B3LYP accuracy should lower the rate of these errors. For these reasons, the reference method was chosen to be TD-DFT with B3LYP.

B. Training dataset

Specifically, the training sets for the ML models considered in this study were derived from the existing PubChemQC (PCQC)[38] and QM-symex[40] databases. For concision, we will use the (functional/basis set) notation to describe the level of theory used in calculations. PCQC includes the first ten singlet excited state energies (S_{1-10}) for 3.5M molecules computed using B3LYP/6-31G(d) for ground state optimization and B3LYP/6-31+G(d) for excitation. Similarly, QM-symex computes both S_{1-10} and the first ten triplet excited state energies (T_{1-10}) for 173k molecules using B3LYP/6-31G(2df,p) for ground state optimization and B3LYP/6-31G for excitation.

We are interested in calibrating both singlet (S_1) and triplet (T_1) excited state energies output by xTB-sTDA, but PCQC does not include triplet excitations. Calculating triplet excitations for 3.5M molecules independently would be prohibitively expensive, so it was necessary to determine which molecules in PCQC would be relevant to spectral conversion applications and, therefore, have interesting excited state properties. To extract such molecules from PCQC, a literature scraping workflow was developed. We used the SCOPUS API[48] to obtain abstracts of articles tagged with "triplet–triplet annihilation" or "singlet fission" keywords. Then, we used ChemDataExtractor[49] to extract molecule names from the abstracts. We then used the PubChem API[50] to convert molecule names into PubChem CIDs and conduct a 2D Tanimoto-coefficient based similarity search among PubChem molecules to expand the molecular space of interest. We then cross-referenced the identified molecules against PCQC to get the singlet energies, and triplet energies were independently generated with TD-DFT using equivalent settings to PCQC. Overall, this process allowed us to select 10k molecules of interest from PCQC, named SCOP-PCQC (after SCOPus-PCQC). To balance the 10k molecules in SCOP-PCQC, a 10k subset of QM-symex was randomly selected and named QM-symex-10k.

However, this 10k molecule subset of PCQC may be too small for ML model training. In addition, a model trained on only molecules relevant to spectral conversion may have poor out-of-domain performance. Therefore, we turn to active learning to add diverse molecules to the training set with further sampling of PCQC.[51] Here, we use active learning techniques to evaluate regions of chemical space where the ML model is uncertain. Active learning is an ML technique often used to sample unexplored regions of state space. Our implementation uses a trained ensemble of ML models to measure uncertainty of the remaining chemical space. Specifically: first, a 10-ensemble ML model was trained on the 10k SCOP-PCQC molecules to directly predict S_1 and T_1 energies. Then, the ensemble was used to predict S_1 and T_1 energies on the remaining 3.5M molecules in PCQC. The 100k molecules with the highest uncertainty (variance in ensemble prediction) were chosen as an expansion to SCOP-PCQC, labeled SCOP-AL-Exp, for each of S_1 and T_1. This process helps ensure broad applicability of the ML model. More details about the active learning process are available in Sec. V of the supplementary material.

C. Test datasets

To evaluate the generated models, various test datasets were used. First, we used 10-fold cross-validation (CV) with 80%/10%/10% training/validation/test splits to quantify in-domain accuracy. In *k*-fold cross-validation, *k* non-overlapping test sets are generated, and models are trained on the remaining 90% of data. Validation sets are also non-overlapping and are used to prevent overfitting.

To prove broader applicability, external test sets were also compiled. 1143 molecules from the paper of Wilbraham *et al.* on Mapping the optoelectronic property space of small aromatic molecules (MOPSSAM) were used, 143 from their calibration training set and 1000 randomly selected from the remaining 250k molecules.[16] 10k molecules from the paper of Fallon *et al.* on indolonaphthyridine thiophene (INDT) derivatives were also used as they are promising candidates for singlet fission.[7] 1000 molecules from the

VERDE database (VerdeDB) of Abreha *et al.* were used, as the classes of molecules identified (porphyrins, quinones, and dibenzoperylenes) are relevant for various green chemistry excited state applications.[39] Finally, to truly test the broader applicability of the model, another active learning cycle was run on PCQC. Using a training set composed of the 10k molecules from SCOP-PCQC plus the 200k molecules from SCOP-AL-Exp, an ensemble ML model was generated and used to evaluate uncertainty on the remaining PCQC molecules. 100k of the highest uncertainty molecules for each of S_1 and T_1 were chosen as the last test set, labeled PCQC-AL. Additional information about active learning is in Sec. V of the supplementary material.

To visualize the training and test datasets, we plotted the locations of the datasets in chemical space. We used uniform manifold approximation and projection (UMAP),[52] a dimensionality reduction technique that reduces the high-dimensional space of chemical structure into two dimensions for ease of visualization. We use the Jaccard–Tanimoto similarity between Morgan fingerprints of molecules as a measure of proximity in chemical space. We first generated a global UMAP based on all molecules, then categorized them into (a) training and (b) test data, and colored them based on their dataset, shown in Fig. 2.

A few trends become apparent from this visualization. As seen in Fig. 2(a), SCOP-PCQC is primarily localized to two regions, while the active learning expansions have broader coverage of chemical space. Many molecules in the SCOP-AL-Exp set are localized around the SCOP-PCQC molecules, suggesting that despite the chemical similarity of structures, their excited state energies may be significantly different. We further observe that the QM-symex-10k dataset provides a uniform sampling of the overall QM-symex dataset and that the QM-symex datasets are significantly different from PCQC. Turning to test datasets, we note that the INDT dataset is significantly different from both the PCQC and QM-symex based training datasets. VerdeDB has some molecules outside but near the PCQC training sets, while others are within the PCQC space. Finally, both MOPSSAM datasets seem to lie within the PCQC training set space, implying that good predictive performance is expected.

D. Computational details

Since the supervised ML model takes in molecular structure and excited state data, we must obtain excited state data for all molecules. xTB-sTDA data were all independently generated. A 3D structure was first initialized using OpenBabel's `gen3d` function for a short conformer search and preliminary geometry optimization.[53] Full ground-state optimization was conducted with GFN2-xTB[54] with the `tight` threshold and a benzene generalized-Born surface-area (GBSA) solvation model to mimic a non-polar environment. `xtb4stda`[55] was then used to prepare the wavefunctions output by xTB for sTDA. Finally, sTDA was used to calculate excited-state properties using an energy threshold of 10 eV. The `-t` flag was used for triplet excited state calculations. For TD-DFT data, database values were used where available. PCQC had S_1 TD-DFT data, but T_1 data were independently generated. MOPSSAM had S_1 TD-DFT data for the 143 calibration set, but not for the 1000 sampled molecules, so this was independently generated (see supplementary material Fig. S11 for a comparison of MOPSSAM 143 S_1 data vs S_1 data generated with our workflow, showing virtually identical results). T_1 TD-DFT data were also independently generated. Both INDT and VerdeDB had S_1 and T_1 TD-DFT data available. However, VerdeDB used the M06 functional for calculations, so these molecules were re-calculated with B3LYP.

Note that while the xTB-sTDA portion of the workflow was standardized, the TD-DFT data were database-dependent. For

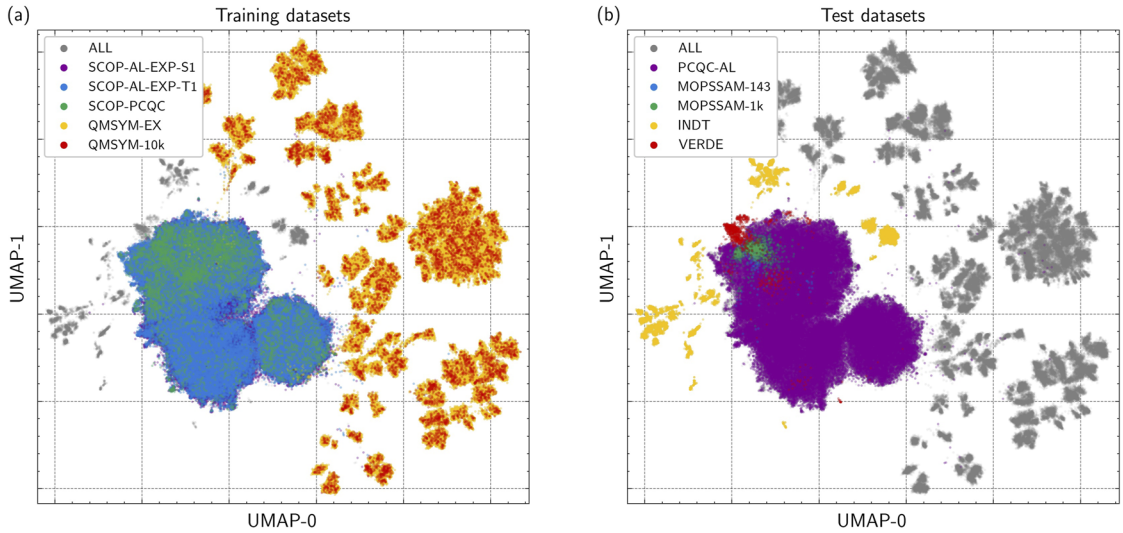

FIG. 2. Embedding of (a) training datasets and (b) test datasets in global chemical space. Training datasets include SCOP-PCQC (green), QM-symex-10k (red), SCOP-AL-Exp-S_1 (purple), SCOP-AL-Exp-T_1 (blue), and QM-symex (yellow). Test datasets include MOPSSAM 143 (blue), MOPSSAM 1k (green), INDT (yellow), VerdeDB (red), and PCQC-AL (purple). The gray datapoints show all molecules not included in the category. A UMAP model was created on all 414k training and test molecules to get their relative positions in chemical space. The Jaccard–Tanimoto similarity coefficient is calculated between each pair of molecules, and UMAP uses this metric for dimensional reduction to 2D space. Additional data on the chemical makeup of the training datasets are presented in Sec. I of the supplementary material.

consistency, only databases that used the B3LYP functional were included in this study, but initial coordinate generation technique and basis sets for (TD)DFT varied for each dataset. The specific settings for each database are shown in Table S1. Once excited state values using both xTB-sTDA and TD-DFT were either compiled or calculated, they could be fed to the ML model.

E. Choosing a machine learning architecture

The type and architecture of the ML model must be optimized for performance. The three ML models considered were DeepChem's[56] graph convolutional network[57] (DC GCN), DeepChem's message passing neural network[58] (DC MPNN), and Chemprop's directed message passing neural network[59] (CP MPNN). These three models were chosen as they are commonly used graph neural networks, which have emerged as a natural choice for molecules where nodes represent atoms and edges represent bonds. The models each use different architectures and methodologies for featurization and property prediction. DeepChem's GCN is based on the paper of Duvenaud *et al.*, which introduced a method to generalize conventional circular fingerprints using convolutional neural networks to generate neural graph fingerprints.[57] DeepChem's MPNN is based on the work of Gilmer *et al.*, which expands upon GCN of Duvenaud *et al.* and is better able to identify correlations between node and edge states.[60] Chemprop's MPNN is based on the work of Yang *et al.*, which adds directionality to the message passing step, preventing noisy graph representations.[59]

We are interested in comparing each model's performance in our application. The default out-of-the-box settings for each ML model were used, as described in Sec. III of the supplementary material. Calibration of the 1000 molecules in the VerdeDB[39] database was used to compare the different ML models. The small size and relatively homogeneous nature of this dataset make it suitable for quickly comparing different ML models. Only the SMILES (simplified molecular-input line-entry system[61]) representation of the molecule was provided as input, and the goal of each model was to accurately predict the S_1 and T_1 error between xTB-sTDA and TD-DFT. Instead of predicting both S_1 and T_1 errors simultaneously, two separate single-task models were generated, both using ten-fold cross-validation. For each fold, the trained ML model was used to predict error values of the test set. Then, each molecule's predicted error was added to the xTB-sTDA output to give a calibrated energy, called the xTB-ML value. The xTB-ML values were compared to the TD-DFT reference results by calculating an R^2 score and the MAE.

Figure 3(a) shows the results of the comparison for S_1 and T_1 energies. As seen, all ML models vastly outperform the linear calibration method. Between the ML models, CP MPNN performs the best for both T_1 and S_1. Note that the large variability in R^2 can be explained by the presence of outliers in the test set—since the test set was only composed of 100 molecules (10% of 1k), a few outliers can vastly impact performance.

Figures 3(b) and 3(c) show plots of original vs CP MPNN-calibrated xTB-sTDA data for (a) S_1 and (b) T_1 energies, with test data from all ten folds compiled and with outliers removed (full plots with outliers available in Sec. IV of the supplementary material). From this analysis, it is evident that CP MPNN performs well in calibrating xTB-sTDA results, even with its default settings.

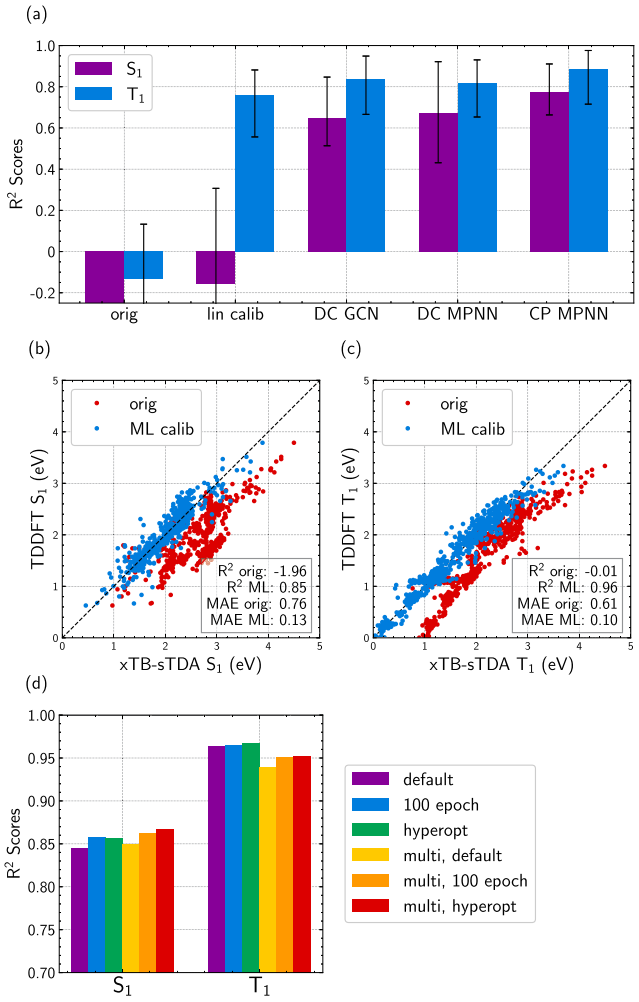

FIG. 3. (a) Comparison of various ML models in accurately calibrating xTB-sTDA against TD-DFT, quantified by the R^2 score. "orig" = original xTB-sTDA data with no calibration and "lin calib" = linear regression calibration of xTB-sTDA data. All others are ML models as presented above. Blue bars represent xTB-ML T_1 energies, while orange bars represent xTB-ML S_1 energies. R^2 for original S_1 data is -1.84 ± 0.65, and the plot was truncated for clarity. Plots of original xTB data ("orig," red) and CP MPNN ML-calibrated xTB data ("ML calib," blue) against reference TD-DFT data generated with Gaussian for (b) S_1 energies and (c) T_1 energies. Datapoints are all test data compiled across ten non-overlapping folds in cross-validation. (d) R^2 scores of xTB-ML vs TD-DFT for various improvements attempted to CP MPNN. "default" bars use the out-of-the-box hyperparameter settings with no additional features. "100ep" bars use 100 epochs instead of the usual 30. "hyperopt" bars use hyperparameter optimization. Bars labeled "multi" use multitasking, so only one model predicts both S_1 and T_1.

To see if the performance could be boosted further, various architectural improvements were attempted. These included increasing the number of epochs (number of iterations to optimize the neural network weights) to 100, conducting hyperparameter optimization (finding the best ML architecture i.e., hidden size, depth, dropout, and number of feed-forward layers), and conducting multi-task training (using a single model to predict both S_1 and

T_1 energies simultaneously). More details about these optimization approaches are available in Sec. III of the supplementary material.

The results from these improvements are shown in Fig. 3(d). As seen, there are only small differences in performance between the default settings and any potential improvements to the ML settings. For T_1, hyperparameter optimization provides minimal improvement, while including additional features or adding multitasking reduces accuracy. For S_1, both hyperparameter optimization and multitasking marginally improve performance. There is thus a trade-off in using multitasking as it could reduce accuracy for T_1 predictions but improve accuracy for S_1, while also reducing overall computation time. Because of the time savings of the multi-task model and previous works showing the benefits of multi-property prediction,[62–64] this was used for ML for the remainder of this study. Hyperparameter optimization was not performed for the following models due to the only marginal improvement seen. Based on this analysis, a larger-scale calibration model can now be developed using CP MPNN.

F. Machine learning calibration workflow

Combining all of the above methodology, a workflow was developed to create ML models that calibrate the xTB-sTDA energy of molecules against the TD-DFT reference. The workflow can be separated into three distinct steps: data generation, model training, and model testing. In the data generation step, S_1 and T_1 excited state energies using TD-DFT and sTDA were either extracted from existing databases or calculated if necessary. The errors between the energies derived from the two techniques were calculated and used as the ground-truth values that the ML model tried to predict. The SMILES strings were also extracted for molecules and used as a representation of the molecular structure.

In the model training step, ML models were trained to take input data and predict xTB-sTDA vs TD-DFT error. Two classes of models were trained, one with only the SMILES string as input (class 1) and another with both SMILES and sTDA energy as input (class 2). During training, the SMILES string was converted to the graphical representation of the molecule, which was then featurized using an MPNN. If the sTDA energy was used (class 2), it was concatenated as an extra feature at this step. Then, feature to property prediction was conducted using a feed-forward NN. To improve reliability of results and ensure all molecules were included in the training process, a 10-model ensemble was generated with 10-fold cross-validation using 80%/10%/10% train/validation/test splits. This process resulted in an optimized ensemble ML model for error prediction.

Once the ML model was trained, it was tested on various datasets. Using the respective inputs, the ML model predicted xTB-sTDA vs TD-DFT errors for the test molecules. When the errors were added to the original xTB-sTDA values, the final calibrated energies were obtained. These were compared to the TD-DFT-calculated values to get a quantitative measure of accuracy of each ML model.

Figure 4 represents an overview of the calibration workflow. Section III A and B detail the results of applying the above

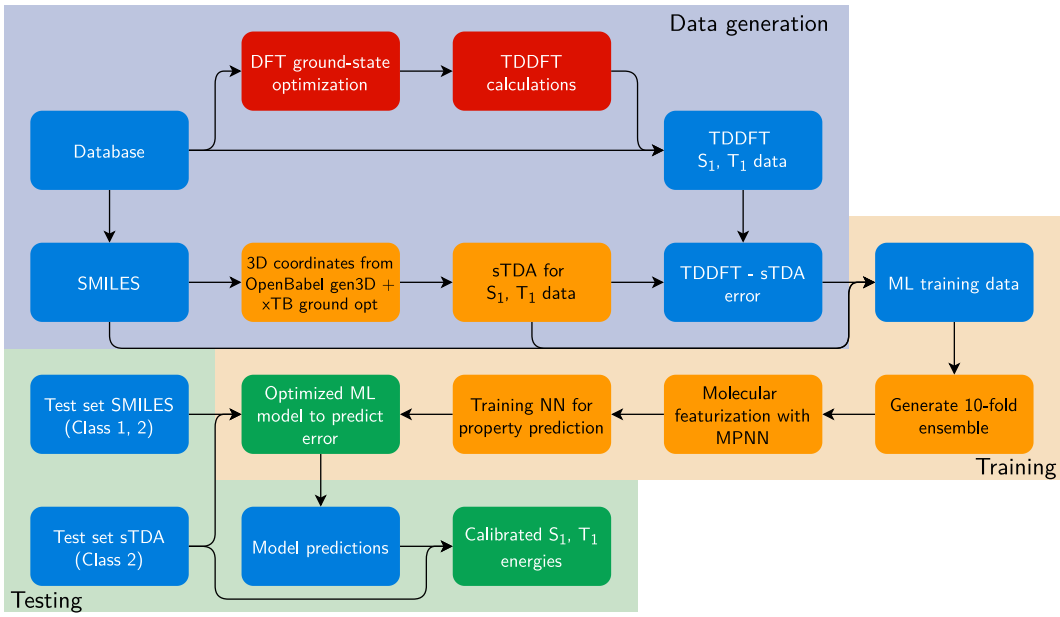

FIG. 4. Workflow for xTB-ML calibration. Blue boxes represent data, red boxes represent expensive calculations, orange boxes represent cheap calculations, and green boxes represent final results. Starting with the training datasets, the TD-DFT and SMILES data are directly extracted. If TD-DFT data are not already available, data are independently generated by first running DFT ground-state optimization and then TD-DFT for excited state calculations. For the xTB-sTDA portion, the SMILES strings are converted to 3D molecular structures with OpenBabel and xTB, and then, excited state calculations are conducted with sTDA. Then, the SMILES string (class 1) or the SMILES string and sTDA energy (class 2) are fed to the ML model, which is trained to predict the error between xTB-sTDA and TD-DFT. The resulting ML model can be used to predict values of various test datasets to quantitatively evaluate its accuracy.

calibration workflow to generate an optimized ML model. The model is then applied for high-throughput screening and chemical space mapping.

III. RESULTS

A. Cross-validation

Before considering external datasets, ML model performance was evaluated on subsets of the training sets themselves. 10-fold cross-validation (training on 90% of the data and testing on the remaining 10% 10 times with non-overlapping test sets) was conducted separately on the SCOP-PCQC, QM-symex-10k, SCOP-AL-Exp, and QM-symex datasets. Class 1 models (only molecular structure as input) and class 2 models (both molecular structure and sTDA energy as input) were both tested. The results are compiled in Fig. 5(a), with plots of class 2 models for SCOP-AL-Exp and QM-symex shown in Figs. 5(b)–5(e). As seen, the original data have low accuracy when compared to TD-DFT results, and linear calibration improves the accuracy slightly. However, there is not a clear linear shift due to some groups of molecules located farther from the line of best fit. In contrast, for both datasets, the ML-calibrated values have much lower MAE and demonstrate significant improvements from uncalibrated xTB-sTDA values, especially for T_1 data. The increase in accuracy with ML is likely because ML detects higher-order patterns, allowing groups of molecules to shift locally instead of having to follow a global calibration rule.

As seen, the ML models performed well in cross-validation. However, it is possible that the ML model only performed well because the datasets were homogeneous, so similar molecules to those in the test set were included in the training set. To evaluate the broad applicability of our model, we used external test sets of molecules not included in either of the datasets above.

B. External test datasets

To test broad applicability, a more general ML model is needed. Therefore, an overarching ML model was trained on the 10k SCOP-PCQC molecules combined with the 10k QM-symex-10k molecules for a total training size of 20k molecules. The overarching ML model was first tested on the MOPSSAM 143 external dataset. As seen in Fig. 6, the ML calibrated xTB-sTDA data match TD-DFT values better than both the original data and the linearly calibrated data. While the data are sparse, there are a few regions where the improvement is clearly visible. For example, for high S_1 energies, the linear calibration tends to overcorrect, while for low S_1 energies, the linear calibration undercorrects. In contrast, the ML model is more flexible and adequately corrects models in both regions. For T_1 energies, the ML model performs similarly to linear calibration with both MAE and R^2 metrics. This is likely because xTB-sTDA nearly always overpredicts the T_1 energy, so calibrating it only requires shifting in one direction, which makes linear calibration sufficient for the task. For S_1 energies, there are both instances of over- and under-prediction, which motivates the need for an ML model. However, there is clearly room for improvement in these results as the ML MAE is still high.

Two avenues of improvement were pursued—first, using a larger training set and second, adding additional input data to the ML model. As discussed in Sec. II, a larger training set was generated using active learning to sample regions of chemical space not represented in the 20k training set. The expanded ML model uses 200k molecules chosen with active learning plus all QM-symex molecules (120k), added to the initial 20k training set, for a total of ~300k molecules. To distinguish the two ML models generated, the 20k model is named xTB-ML-20k, while the expanded 300k model is xTB-ML-300k. The second improvement explicitly included the xTB-sTDA calculated energy as an input to the ML model. As discussed previously, for these models, called class 2, the xTB-sTDA

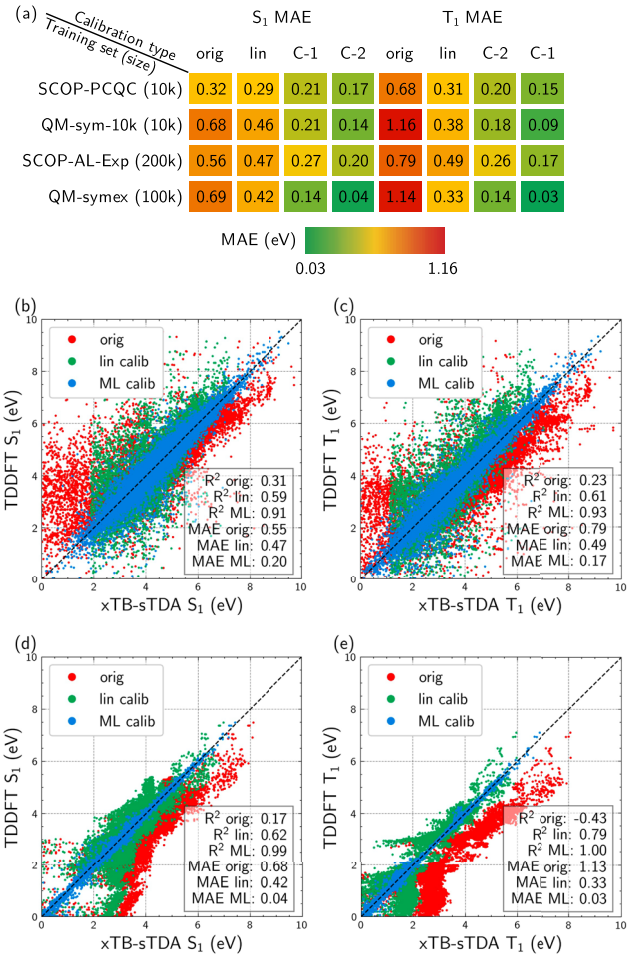

FIG. 5. (a) Table comparing cross-validation results for various training datasets and various levels of calibration. "orig" = raw xTB-sTDA values, "lin" = linearly calibrated data, "C-1" = ML calibrated data with the class 1 model, and "C-2" = ML calibrated data with the class 2 model. Plots of xTB-sTDA calibration of (b) SCOP-AL-Exp S_1, (c) SCOP-AL-Exp T1, (d) QM-symex S_1, and (e) QM-symex T_1 energies using class 2 cross-validation (CV) models. 10-fold CV was conducted, meaning that all datapoints shown are test points predicted by an ML model trained on the other 90% of data. Only one fold is shown here for clarity of visualization, but the metrics correspond to an average over all ten folds. The inlaid box shows quantitative measurements of accuracy for original, linearly calibrated, and ML calibrated data. Red dots represent original data with no calibration, green dots represent linearly calibrated data, and blue dots are calibrated with ML. (Best R^2 is 1, while the best MAE is 0.)

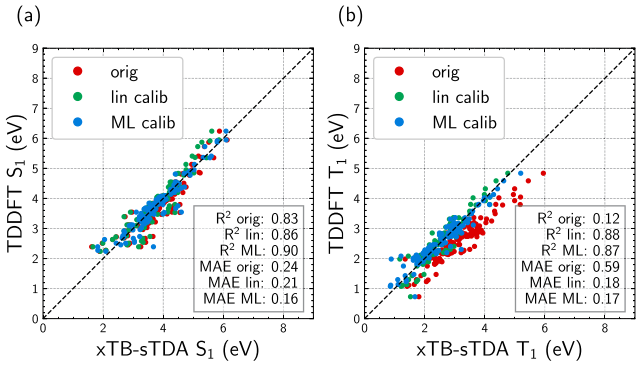

FIG. 6. Plot of xTB calibration of the 143 MOPSSAM molecules for (a) S_1 and (b) T_1 energies using a class 1 ML model. Red dots represent original data with no calibration, green dots represent linearly calibrated data, and blue dots are calibrated with ML. Training data were the 20k molecules in SCOP-PCQC + QMsymex-10k, and test data were the 143 molecules shown here. Inlaid boxes show quantitative measurements of accuracy for original, linearly calibrated, and ML calibrated data.

energy was concatenated to the generated molecular features during the training process.

The performances of all ML models, considering both classes (1 and 2) and both sizes (20k and 300k), are compared for several external datasets: MOPSSAM 143, MOPSSAM 1000, INDT, VerdeDB, and PCQC-AL. Figures replicating Fig. 6 for the various datasets are available in Sec. VI of the supplementary material, with primary accuracy metrics presented in Fig. 7.

As seen, all ML models improve raw xTB-sTDA values, but to different extents depending on the training and test set considered. For the MOPSSAM and PCQC-AL datasets, using larger datasets with more input data generally improves results. This result is intuitive, as more data allow the ML model to learn more about patterns in the datasets. The lowest MAE obtained was 0.08 eV using the class 2 xTB-ML-300k model on T_1 energies of MOPSSAM 143.

For the INDT and VerdeDB datasets, the results are less intuitive. In these cases, the class 1 xTB-ML-20k ML model, i.e., the model trained on a smaller training set with less input data, performs better. There are several reasons this could be happening. The INDT and VerdeDB datasets are composed specifically of molecules relevant to photon conversion or green chemistry applications. Similarly, the xTB-ML-20k dataset is composed primarily of literature scraped molecules intended for spectral conversion applications, so it is more likely to include molecules similar to those in INDT or VerdeDB. Although the training set size in xTB-ML-300k is larger, the method of expansion through active learning specifically includes molecules significantly different from the 20k training set, so the model may not increase in accuracy for photon conversion molecules, i.e., it is not backward compatible. In terms of why the class 2 models perform worse than class 1 models, this could be related to the nature of the training sets used. Both the 20k and 300k training sets have only a few molecules with low excited state energy. Therefore, the calibration in this region may be inaccurate. As seen in Fig. S15, the INDT dataset is primarily composed of low-energy molecules. Therefore, if the energy is localized by providing the sTDA energy, the molecule may undergo an inaccurate calibration. This localization is minimized if the calibration is done solely based on molecular structure, allowing for a more accurate calibration.

Most of the best-performing ML models result in an MAE of less than 0.20 eV. However, the one dataset with large MAEs is the PCQC-AL T_1 energies. As seen in Fig. S17, most of the calculated energies follow a general linear trend. However, there is a large cluster of molecules distinctly separated from the rest that is inflating the MAE. This is expected from the active learning workflow, which

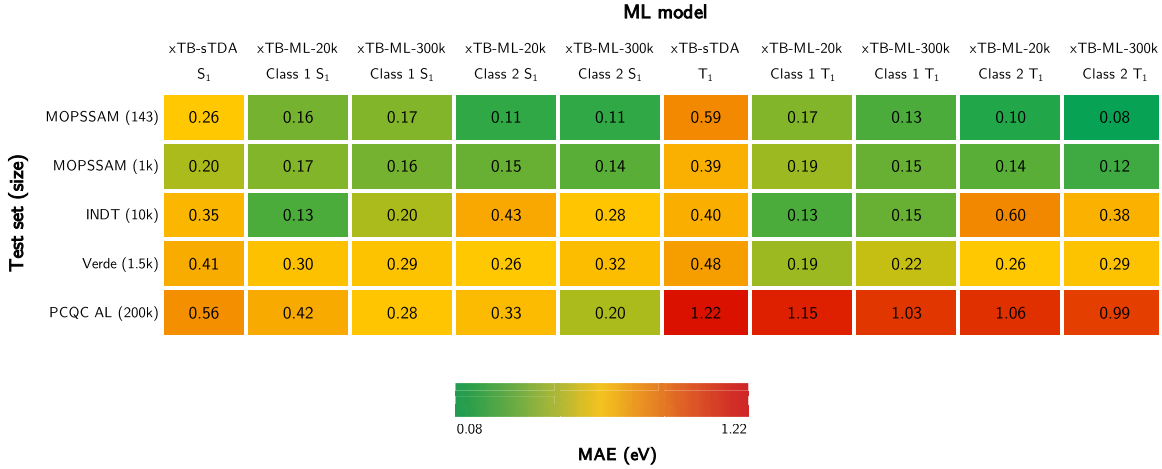

FIG. 7. Performance of xTB-ML-20k and xTB-ML-300k on various external test datasets, compared to raw xTB-sTDA energies. Two classes of model are tested, class 1 with only the molecular structure as input and class 2 with both molecular structure and xTB-sTDA energy as input. Test datasets are completely different from the training datasets and are split into three categories: MOPSSAM (for comparison to previous work in xTB-sTDA calibration), datasets relevant for TTA/SF (INDT and VerdeDB), and a broader applicability dataset (PCQC-AL). Performance is measured with the MAE, which is calculated with TD-DFT (B3LYP) as reference.

selects molecules difficult to predict with the existing model. Naturally, if a subset of these molecules were included in the training set, the overall MAE would likely improve drastically. Regardless, while this test dataset has large errors, by design, these are the largest errors one can obtain, and for general PCQC molecules, the error should be lower.

Overall, these results show that the ML models significantly improve raw xTB-sTDA calculated values. In most cases, the best-performing ML model reduces the MAE by more than half. Furthermore, while not shown in Fig. 7, the ML models also consistently outperform linear calibration, showing the benefits of a higher-order calibration. We have thus shown that machine learned calibrations can help improve the accuracy of xTB-sTDA results over a wide variety of datasets, when compared to a TD-DFT (B3LYP) reference. We can now use these models for various applications.

C. Applications of xTB-sTDA calibration

1. Direct vs calibration ML models

ML has been used extensively in the past to explore the excited state space of molecules, primarily being used to directly predict excited state properties, such as energies, spectra, and dynamics.[65,66] However, we expect a ML model trained to directly predict TD-DFT results to perform worse than a calibration model where the baseline method does most of the work and the calibrator simply shifts the result in the right direction. This calibration or Δ-ML approach has been used extensively in the past and has shown superior performance to pure ML models.[19–25,28,29] Calibration is particularly useful for improved out-of-domain predictive performance. Because supervised ML is a data-driven method, it may have poor performance on molecules distinctly different from those in the training set. In contrast, xTB-sTDA is data-agnostic, so it should give reasonable results regardless, and ML should slightly improve results through calibration.

To prove this for the xTB-ML models generated in this work, we consider class 1 ML models trained on the 20k and 300k datasets presented previously, but instead of being trained on the error between TD-DFT and xTB-sTDA [as xTB-ML-(20k,300k) are], the new models, called TDDFT-ML-(20k, 300k), are trained directly on TD-DFT values. (Note that direct class 2 ML models would give equivalent results to calibration class 2 models since the sTDA energy is provided as input.) xTB-ML-(20k, 300k) and TDDFT-ML-(20k, 300k) are then tested on the MOPSSAM 143 dataset, with TDDFT-ML directly predicting values and xTB-ML predicting the errors that are added to the sTDA energies to get the final calibrated values. The results of this analysis are shown in Fig. 8.

As seen, the performance of the directly trained ML model is worse than the ML-calibrated xTB data for both dataset sizes. The TDDFT-ML-20k model performs similarly to the linear calibration model (seen in Fig. 6), while the xTB-ML-20k model already significantly outperforms both. However, it is well known that direct ML models often require more training data than calibration ML models. When expanding the training set to 300k, the TDDFT-ML-300k model outperforms linear calibration but still underperforms compared to both xTB-ML-20k and xTB-ML-300k. Thus, calibrating xTB with ML gives much higher accuracy than using ML to directly predict energies. The benefit to a direct ML model is computational speed, as it can screen ~2 orders of magnitude more

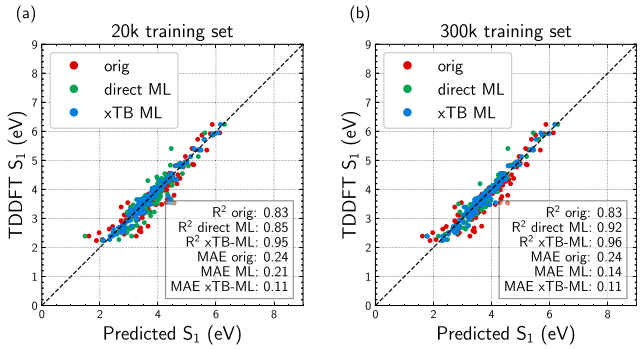

FIG. 8. Comparing direct vs calibration ML models for MOPSSAM. "orig" data are uncalibrated xTB-sTDA data, "direct ML" indicates results of directly predicting TD-DFT data with ML using a (a) 20k and (b) 300k molecule training set, and "xTB ML" indicates results of calibrating xTB-sTDA values with (a) class 1 xTB-ML-20k and (b) class 1 xTB-ML-300k.

molecules in a given time period than xTB-ML. However, our goal is to attain approximately the same accuracy as TD-DFT methods, so a direct ML model would not be useful. From the above analysis, our assumption of the improved performance of a calibration ML model is upheld. We now apply our generated calibration ML models for high-throughput screening and chemical space mapping.

2. High-throughput screening for spectral conversion

As discussed in Sec. I one of the motivations of developing this ML calibration is fast and accurate high-throughput virtual screening (HTVS) of spectral conversion materials. The two spectral conversion techniques of interest are triplet–triplet annihilation (TTA) up-conversion and singlet fission (SF) down-conversion. In general terms, TTA involves two sensitizer molecules that absorb low-energy light and transfer their energy to a single emitter molecule, which then re-emits the high-energy light. SF involves two emitter molecules, where one absorbs high-energy light and transfers half its energy to a neighboring emitter, and both re-emit low-energy light. In such molecules, the S_1 excited state is usually involved in absorption/emission, while T_1 is typically used for energy transfer. The excited state energy levels of sensitizers and emitters must be well-aligned for efficient spectral conversion—figures of merit (FOMs) to evaluate this alignment are

$$\mathrm{FOM}_{sens} = \begin{cases} 0, & S_1 < T_1, \\ e^{-\left|1-\frac{S_1}{T_1}\right|}, & S_1 \geq T_1, \end{cases} \tag{1}$$

$$\mathrm{FOM}_{emit} = e^{-\left|2-\frac{S_1}{T_1}\right|} \begin{cases} \mathrm{SF}, & S_1 > 2T_1, \\ \mathrm{TTA}, & S_1 < 2T_1. \end{cases} \tag{2}$$

For sensitizers, the first check is if the energies are invalid, i.e., if T_1 is greater than S_1. Then, molecules with S_1 close to T_1 are rated higher. Emitters can be separated into those suitable for singlet fission (singlet more than twice triplet) or triplet–triplet annihilation (singlet less than twice triplet). The same FOM formula is used for both cases, where S_1 close to twice T_1 is desirable. By ensuring that the ratios are as close as possible to ideal, we ensure

that there is minimal loss in energy. Note that properties related to absorption-to-emission likelihood such as oscillator strength, triplet–triplet energy transfer probability, triplet–triplet annihilation probability, and others are also important, but are not considered in the present analysis, which focuses on optimizing excited state energy level alignment. Although not considered here, xTB-sTDA does output the oscillator strength of each transition, which can be directly used, demonstrating a further benefit of the calibration method.

We screen the 250k molecules considered by Wilbraham *et al.*[16] for sensitizers and emitters to demonstrate the applicability of xTB-ML to high-throughput screening. We use the class 2 xTB-ML-300k model as it is the most accurate for the MOPSSAM dataset. First, we calibrate S_1 and T_1 energies using the ML model and compare the results to the linear calibration done in the original work. The results are shown in Figs. 9(a) and 9(b).

For S_1, the linear calibration is minimal. The ML calibration remains centered around the raw data for mid-to high-energies, but changes more drastically at low energies. This mirrors the previous discussion of Fig. 6, where the linear calibration either over- or under-corrects, but the ML model is more flexible. For T_1 calibration, at mid-to high-energies, both linear and ML calibration shift the energy down, reflecting the tendency of xTB-sTDA to consistently overestimate T_1 energies. For low T_1 energies, the ML model increases the raw energy, suggesting that sTDA tends to sometimes spuriously calculate low T_1 energies, which can be corrected with ML. Note that because TD-DFT data were not calculated for these 250k molecules, we cannot compare the calibration to ground truth, but based on the metrics presented in Fig. 7, it is likely that the ML-calibrated values are more accurate.

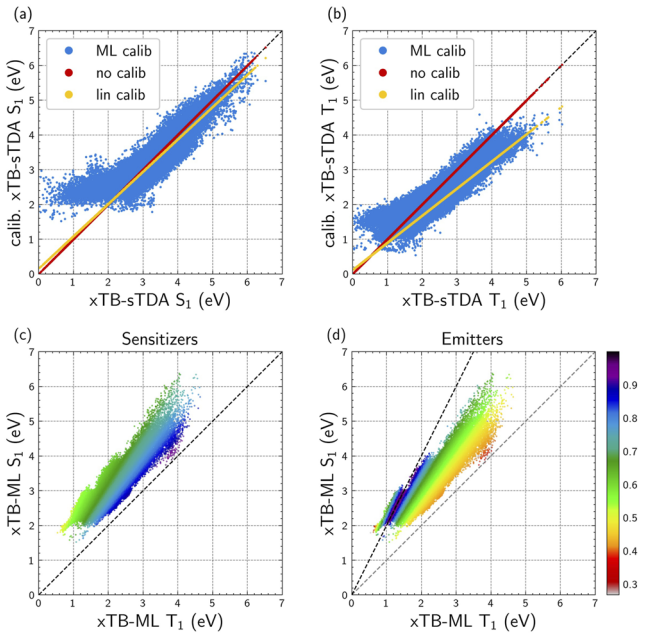

FIG. 9. Plots of 250k molecules showing difference in calibrated (a) S_1 and (b) T_1 with the ML model compared to the linear model. Red dots represent data without calibration, yellow dots represent data with linear calibration, and blue dots represent data with ML calibration. Plots of 250k molecules showing S_1 and T_1 energies, colored with FOM for (c) sensitizers and (d) emitters. (c) and (d) share a colorbar. Target properties are the correct ratio of S_1/T_1, as defined in Eqs. (1) and (2).

Now that we have both S_1 and T_1 energies calculated for 250k molecules with xTB-ML, we can identify potential sensitizers and emitters using the FOMs defined in Eqs. (1) and (2). Figures 9(c) and 9(d) show the results of screening molecules for potential sensitizers and emitters.

As seen, there are several molecules that would function as potential sensitizers and emitters for photon conversion. Section VII of the supplementary material contains further details about the chemical composition of the candidate molecules and their distribution in chemical space. The suggested molecules could then be verified with higher-accuracy techniques, such as range-separated hybrid TD-DFT or CC2 to confirm their suitability.

Note here the importance of accuracy for a first-pass screening methodology, such as xTB-sTDA. If the uncalibrated results were used, likely several suggested molecules would not be suitable (false positives), and several suitable molecules would not be suggested (false negatives). Using xTB-ML improves the quality of suggestions by reducing both of these rates.

We have therefore used the class 2 xTB-ML-300k model to make quick and relatively accurate calculations for S_1 and T_1 energies and have used the results to screen for potential sensitizers and emitters. This screening was relatively fast as the dataset size was small (250k), and xTB-sTDA results were already provided by Wilbraham *et al.*[16] For larger datasets on the order of millions (PubChem)[67] or billions of molecules (GDB-17),[68] running xTB-ML becomes expensive. A more intelligent sampling technique (such as active learning) could be used to screen such large databases, and this is an avenue of future work.

3. Mapping inaccuracies of xTB-sTDA in chemical space

Since our ML model predicts the error in xTB-sTDA, an interesting application is to map the error in S_1 and T_1 calculations in a global chemical space, to see if there are some areas where xTB systematically over- or under-estimates or areas where xTB is projected to be accurate. For this analysis, we used class 1 xTB-ML-300k, as it is shown to be accurate in the general chemical domain and does not require xTB-sTDA computations, so large-scale predictions can be made quickly with ML.

We first used UMAP to generate a chemical space map of all PCQC molecules. We then colored the global chemical space map in three different ways, as shown in Fig. 10.

For the first two plots [Figs. 10(a) and 10(b)], we used our ML model to predict the error in xTB-sTDA. Here, the error is defined as

$$\Delta E_{\text{error}} = E_{\text{TD-DFT}} - E_{\text{xTB-sTDA}}, \tag{3}$$

so a negative error implies that xTB-sTDA is over-predicting the excited state energy. As seen in Fig. 10(a), there are distinct regions where xTB-sTDA over-predicts S_1 (right side), regions where xTB-sTDA has reasonable accuracy (top left and center) and regions where it under-predicts (bottom left and top). In general, most molecules are within ±0.5 eV error.

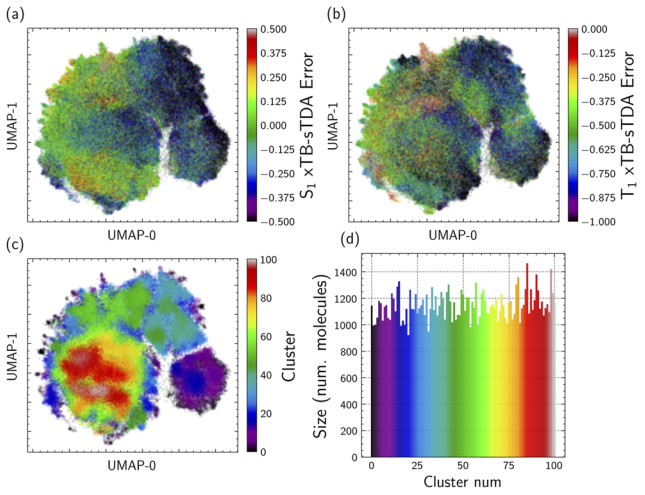

FIG. 10. Global chemical space maps of the PubChemQC dataset. Chemical space plots generated with the UMAP dimensionality reduction algorithm. Plots of xTB-sTDA (a) S_1 and (b) T_1 errors in global chemical space. (c) Clustering of molecules in global chemical space using the HDBSCAN algorithm. (d) Number of molecules per cluster in global chemical space.

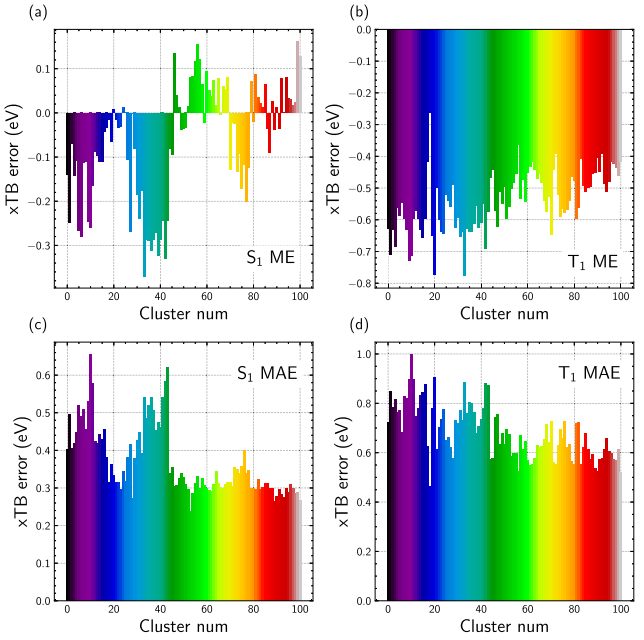

FIG. 11. Mean errors for S_1 and T_1 energies of molecules in global chemical space. (a) Mean S_1 error, (b) mean T_1 error, (c) mean absolute S_1 error, and (d) mean absolute T_1 error. "Absolute" error takes the absolute value of errors before averaging them.

In contrast, for the T_1 energy, xTB-sTDA over-predicts for almost all molecules, as seen in Fig. 10(b). Note that the scale in this plot is shifted from −0.5 to 0.5 eV (as in S_1) to 0 to −1.0 eV to make the distribution of errors clearer. Only a few scattered molecules are under-predicted by xTB-sTDA and are colored red, and all other molecules are over-predicted. Similar to S_1, xTB-sTDA over-predicts T_1 for most molecules on the right side and gets reasonable accuracy on molecules in the middle and top left. T_1 is also over-predicted on a cluster of molecules on the bottom left and top.

Next, we used HDBSCAN[69] to cluster the molecules based on proximity, as shown in Fig. 10(c). HDBSCAN takes as input the reduced dimension data from UMAP and outputs a number for each datapoint. It is a soft clustering, not creating distinct categories but instead giving molecules a rating between 0 and 1 (or −1 for no cluster, as ∼1/3 of the molecules were unable to be clustered) and 100 distinct clusters were created manually from these ratings. We used a minimum cluster size of 10 and the leaf cluster selection method. We can see that HDBSCAN effectively clusters molecules in space, with most molecules in close proximity included in the same cluster. Some of the clusters themselves are spread out across space, such as the purple cluster that includes many molecules along the edge of the global space. Note that this is a dataset-agnostic clustering, as the clustering algorithm only sees molecular information and no labeled data. More details about the HDBSCAN algorithm can be found in their paper[69] and website.[70]

A natural question is whether each cluster as defined by HDBSCAN has a particular error associated with it. For example, it seems that xTB-sTDA does a relatively good job for the red cluster, but over-predicts energies for molecules in light blue, purple, and orange clusters. In contrast, the dark green and red clusters seem to have low errors. Although HDBSCAN is a soft clustering, we can categorize molecules into 100 distinct clusters based on the number assigned to them and two additional clusters (1 each for unclustered molecules and for outliers). Figure 11 quantifies the mean errors for S_1 and T_1 energies for each cluster.

Figures 11(a) and 11(b) show the mean errors (MEs) of S_1 and T_1, while Figs. 11(c) and 11(d) show the mean absolute errors (MAEs). As seen, the red/yellow/green clusters are likely to have low error, while the purple/dark green clusters have high error. While this analysis is generally useful, the mapping and clustering approach requires knowing the location and cluster categorization of a specific molecule in global chemical space. Oftentimes, this is not known or would require significant computation.

Instead, it would be beneficial to have some chemical intuition of accuracy based on the molecular structure, to have greater confidence in xTB-sTDA calculations, or to know to use the ML model or consider other computational techniques. To this end, we can identify substructures that are more likely to be present in low-error or high-error molecules.

We first use our ML model generated above to predict the S_1 and T_1 error between xTB and TD-DFT for 1M molecules randomly subsampled from PCQC. We then categorize the molecules based on the predicted error as follows:

$$\mathrm{Cat}_{S_1} = \begin{cases} \mathrm{Low}, & |S_{1,err}| < 0.05, \\ \mathrm{High}_{\mathrm{Under}}, & S_{1,err} > 0.5, \\ \mathrm{High}_{\mathrm{Over}}, & S_{1,err} < -0.5, \end{cases}$$
$$\mathrm{Cat}_{T_1} = \begin{cases} \mathrm{Low}, & |T_{1,err}| < 0.05, \\ \mathrm{High}_{\mathrm{Under}}, & T_{1,err} > 0, \\ \mathrm{High}_{\mathrm{Over}}, & T_{1,err} < -1.0, \end{cases} \quad (4)$$

where "under" refers to xTB underestimating the energy, while "over" refers to overestimating [note the error definition in Eq. (3)]. For both the T_1 and S_1 errors, we define low error as $<\pm 0.05$ eV. However, for defining high error for T_1, we shift the bounds down by 0.5 to reflect the distribution of errors, as seen in Fig. 10(c).

We can conduct substructure analysis on each category to know when to trust the xTB-sTDA results or when to expect exceptionally high errors. We use molZ[72] to analyze which substructures are over-represented in each category. The results of this substructure analysis are shown in Fig. 12.

From these plots, a few patterns become evident. Low error molecules are more likely to be aromatic, potentially with sequential attached rings, for both S_1 and T_1. In contrast, S_1 high overestimation, S_1 high underestimation, and T_1 high underestimation

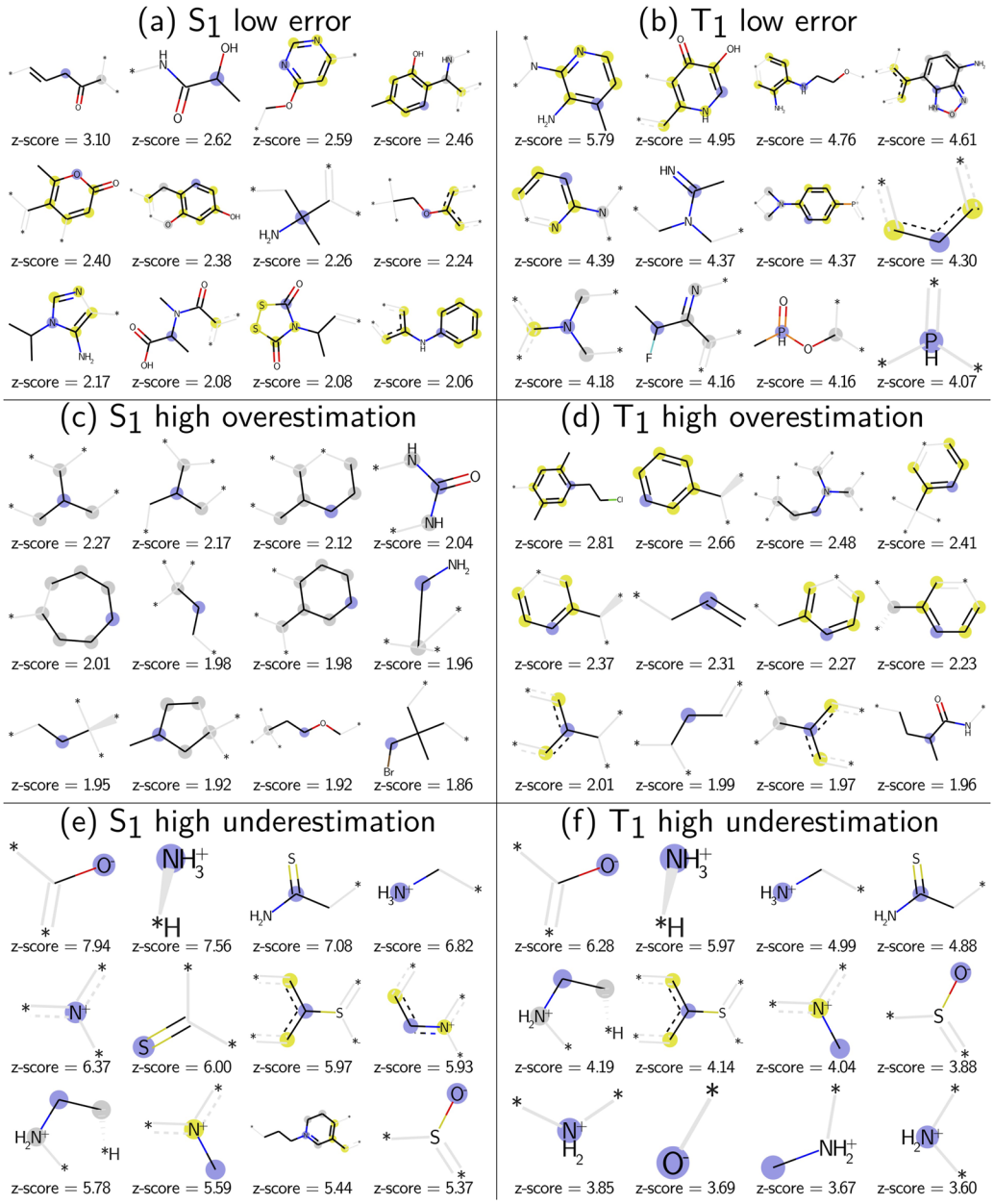

FIG. 12. Grid of molecular substructures over-represented in molecules in each error category as predicted by the ML model for (a) low S_1 error, (b) low T_1 error, (c) high S_1 overestimation, (d) high T_1 overestimation, (e) high S_1 underestimation, and (f) high T_1 underestimation. According to RDKit,[71] blue atoms are the center atoms, yellow atoms are aromatic atoms, dark gray atoms are aliphatic ring atoms, and light gray atoms/bonds are connectivity invariants.

molecules are likely to be not aromatic, with some unconventional molecular structures included in these groups. In particular, the S_1 overestimation group includes five and seven C-ring molecules, and both S_1 and T_1 underestimation include charged N atoms. There are some aromatic substructures in the T_1 overestimation molecules, but they are attached to the bulk structure with a rotatable bond. This overestimation could be a result of the 3D structure generation since only limited conformer analysis is conducted and potentially the lowest energy conformer was not achieved. To clarify the effect of this vs an inherent inaccuracy in the excited state energy calculation of xTB-sTDA, a more extensive conformer analysis could be done in a future work. Further substructure analysis of each error category, including most common scaffolds and most common fragments based on RDKit,[73] is provided in Sec. VIII of the supplementary material.

Overall, these predictions can be used as guides for the accuracy of xTB-sTDA in calculating excited state energies.

4. CC2 calibration of xTB-sTDA

Finally, to show the generalizability of the methodology presented here, we choose a different reference technique beyond TD-DFT, namely, CC2.[74] CC2 is known to better predict excitation energies than TD-DFT, but its computational cost is often prohibitively expensive.[31] We use the CC2 S_1 values compiled in QM8,[24,75] randomly sampling 10k molecules as the training set and using the other 11.5k as the test set. xTB-sTDA values were generated using the same methodology as before. For ML model ensemble generation, because of the smaller dataset, we use 20-fold cross-validation with 95%/5% train/validation splits. This helps ensure that all the data are used in training. As a class 2 model, both SMILES and sTDA energy are given as input. The new model is termed xTB-CC-ML to distinguish it from the previously generated xTB-ML models.

Figure 13(a) shows the results of the comparison, with measurements of accuracy for both methods presented in the inlaid box. As seen, adding the ML calibration to xTB-sTDA results vastly improves results, reducing the MAE by 66%. For comparison, Fig. 13(b) shows the results of TD-DFT calculations using PBE0[76] and CAM-B3LYP on the same test set of molecules. As seen, xTB-CC-ML has higher accuracy than TD-DFT calculations for the 11.5k test set using either R^2 or MAE as the metric.

Note that Fig. 13 also justifies the main calibration methodology presented in this section of calibrating xTB-sTDA against TD-DFT. While xTB-sTDA was initially parameterized against mostly CC2 calculations, its accuracy is lower than TD-DFT with hybrid functionals, such as PBE0, when compared to CC2. Because TD-DFT values are close in accuracy to CC2 values, calibrating xTB-sTDA to TD-DFT is a useful exercise. The functional B3LYP was chosen in this work due to the large amount of excited state data available using this functional because it is less computationally intensive than range-separated hybrid functionals, such as CAM-B3LYP. Calibrating against more accurate functionals or CC2 could be an avenue of future work.

To test the impact of training size on accuracy, eight different ML models were generated with training sizes ranging from 100 to 15 000. The models were then predicted on the remaining molecules in QM8 not used in the training set. The MAE of the test set (against CC2 values) was compared to the MAE of PBE0/def2-TZVP and CAM-B3LYP/def2-TZVP, as shown in Fig. 13(c). As seen, a training size of less than 500 molecules allows xTB-sTDA to achieve similar accuracy to PBE0. It is more difficult to match CAM-B3LYP, but this is achieved at a training size of around 1500. At the largest training size considered (15k), xTB-CC-ML vastly outperforms both TD-DFT techniques, with a 62% lower MAE compared to PBE0 and 47% lower MAE compared to CAM-B3LYP.

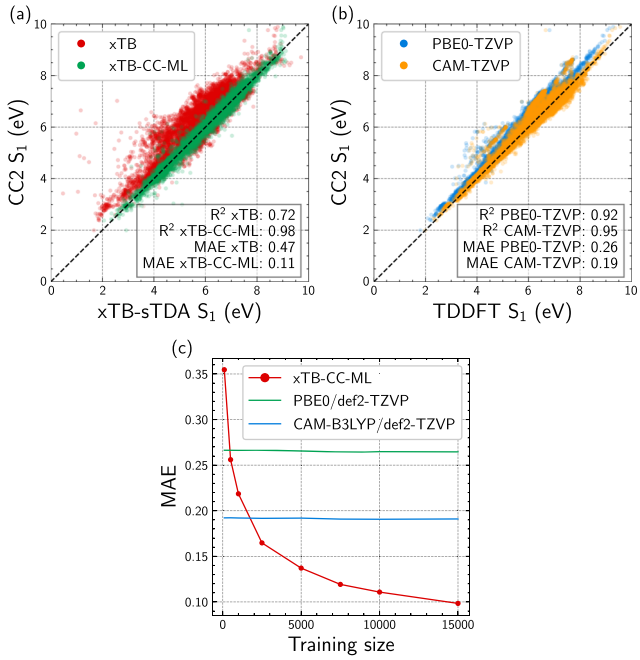

FIG. 13. (a) Plot of xTB calibration against CC2 using a class 2 ML model with the 10k training set and 11.5k test set (shown). Red dots indicate original xTB calculations, while green dots indicate calibrated xTB data. (b) Plot of TD-DFT calculated values against CC2 values for accuracy comparison. The black dashed line in both plots indicates the $x = y$ line. (c) Plot of xTB calibration accuracy as a function of training size, tested on the remaining molecules in the 21.5k dataset. MAE of the test set is calculated with CC2 values as reference. The eight red dots correspond to the eight ML models generated for xTB-sTDA to CC2 calibration.

These are promising results; however, the xTB-CC-ML model may not be as generalizable as xTB-ML due to the smaller (10k), less diverse (only small molecules with up to eight heavy atoms) training set. To further explore generalizability, two additional ML approaches were considered. First, we calibrate xTB-sTDA against CC2 with transfer learning. The learning rate analysis above showed a 10k training set size gives high accuracy while leaving enough molecules in the test set for a reasonable error measurement. Therefore, we train a ML model on 10k randomly sampled molecules from QM8 to predict CC2 values, given SMILES and xTB-sTDA energy as input, using the class 2 xTB-ML-300k model as a starting point, with the first MPNN and first FFNN layers frozen. We use the largest, most detailed ML model considered in this work as a starting point, so any adjustments made to this model using the smaller QM8 set should propagate to the larger model. The results of this analysis on the 11k test molecules are shown in Fig. 14(a). We call this type of model xTB-CC-TL.

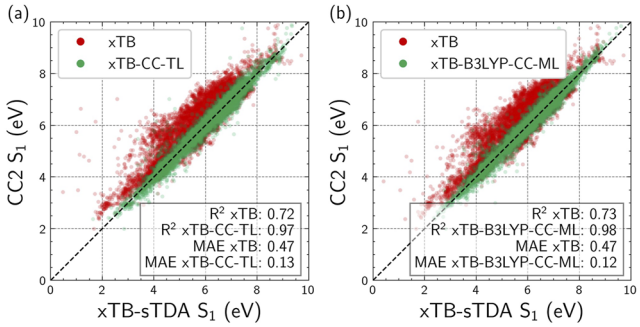

FIG. 14. (a) Plot of xTB calibration against CC2 using transfer learning with the class 2 xTB-ML-300k model as pre-training. The first layer of both MPNN and FNN is frozen with xTB-ML-300k model values. (b) Plot of xTB calibration against CC2 using B3LYP calibration as an intermediary, i.e., calibrating xTB to B3LYP using xTB-ML-300k and then calibrating the predicted B3LYP values against CC2 using B3LYP-CC2-ML. ML models are generated with a 10k training set, and the 11.5k test sets are shown. Red dots indicate original xTB calculations, while green dots indicate calibrated xTB data.

Our second approach was to calibrate xTB-sTDA against B3LYP and then calibrate B3LYP against CC2. We first generate an ML model that calibrates B3LYP against CC2. We ran B3LYP independently on QM8 using the same settings as outlined in the work. Using these values, we train an ML model on 10k molecules in QM8 that takes in SMILES and B3LYP energy as input and predicts CC2 energy, called B3LYP-CC2-ML. We then apply the class 2 xTB-ML-300k model generated in this work to predict B3LYP energies. We finally use these predictions as an input to the B3LYP-CC2-ML model to get CC2 energies. We therefore calibrate xTB-sTDA to B3LYP first and then to CC2. This overall approach is called xTB-B3LYP-CC-ML. The results of this calibration are shown in Fig. 14(b).

As seen, both xTB-CC-TL and xTB-B3LYP-CC-ML have similar performance on the test set and perform slightly worse than xTB-CC-ML. It is difficult to tell *a priori* which of these models would generalize better, although both would certainly generalize better than the simple xTB-CC-ML model, which does not consider external data at all. For ease of comparison in the future, we have applied all three models to the existing datasets. While it would be ideal to have CC2 energies for these datasets, unfortunately, this would be prohibitively expensive to generate for the large number of molecules required to obtain a meaningful error value. We have therefore left this analysis for a future work but have uploaded the ML CC2-calibrated values of all datasets to Github.[77]

Overall, these xTB-CC models serve as interesting proofs of concept that can be expanded further in the future, perhaps with additional CC2 calculations on more diverse molecules.

IV. CONCLUSIONS

We have presented a methodology for calibrating a high-throughput computational chemistry technique (xTB-sTDA) against a high-accuracy one (TD-DFT) using machine learning. We first decided on Chemprop's directed message passing neural network (MPNN) as the ML architecture of choice and then generated a training set using literature scraping of relevant molecules from abstracts (SCOP-PCQC) and an existing excited state database (QM-symex-10k). We also generated an expanded training set using active learning. We built two models based on these training sets (xTB-ML-20k and xTB-ML-300k).

We then generated blind test sets from a study on linear calibration of xTB-sTDA results (MOPSSAM),[16] a study on singlet fission materials (INDT),[7] a database of molecules relevant to green chemistry (VerdeDB),[39] and a dataset to test the broad applicability of the model (PCQC-AL).[38] On these external datasets, the ML calibration models outperformed both raw xTB-sTDA and linear calibration, oftentimes significantly. Averaging the best MAE over all external test sets (both S_1 and T_1) excluding PCQC-AL gave an MAE of 0.14 eV, compared to 0.38 eV for xTB-sTDA. Including PCQC-AL gave an average MAE of 0.57 eV, compared to 0.83 eV for xTB-sTDA. If xTB-ML is used as the first step in a high-throughput screening process instead of raw xTB-sTDA outputs, its low error can help ensure that all relevant molecules are selected and vice versa.

After evaluating the performance of xTB-ML, we then used the model for four applications. First was comparing the xTB-ML model against directly predicted energies with ML, showing that the xTB-ML model had better accuracy (0.11 vs 0.21 eV MAE). Second was rapidly screening 250k molecules for suitability as sensitizers and emitters for spectral conversion applications. Third was mapping inaccuracies of xTB-sTDA in chemical space using the ML model to predict errors. This was used to see which regions of chemical space xTB-sTDA have high errors in. S_1 errors were small, with most molecules being within 0.5 eV. There were clear regions where xTB-sTDA overpredicted S_1, but only a few for underprediction. T_1 energies were generally overpredicted, with most molecules being between 0 and 1 eV above TD-DFT values. Global chemical space mapping provides another method of predicting xTB-sTDA error by calculating which cluster a molecule belongs to and referencing the MAE of that cluster. Properties of low-error molecules were also evaluated, finding that non-aromatic molecules are likely to have higher error. The final application was for generalizing the methodology to calibrate xTB-sTDA against coupled cluster theory and generating a new xTB-CC-ML model. The calibrated xTB-CC-ML values had high accuracy (0.10 eV MAE), outperforming TD-DFT values calculated with PBE0 (0.26 eV MAE) and CAM-B3LYP (0.19 eV MAE). We also generated more general calibration models with transfer learning and using B3LYP as an intermediary.

There are several avenues for future work. First is improving the ML model architecture. While Chemprop's MPNN outperformed other ML models, primarily due to its advanced featurization, only the 2D molecular structure and xTB-sTDA energy were provided as input. Since the 3D structure is known, including this information would likely improve performance. Another improvement to the ML workflow would be to conduct a more intensive conformer search. While OpenBabel's `gen3D` function includes a search for 200 conformers, these may not include the lowest energy conformer, thus reducing the accuracy of the xTB portion of the workflow. Using a conformer searching tool, such as CREST,[78] would be more comprehensive, although the computation time added may detract from the high-throughput nature

of the xTB-ML process. The ML model could also be expanded to calibrate several singlet states instead of just the first, similar to that by Kang *et al.*[79]

Beyond higher level first-principles data, the calibration models could be further extended to experimental data. However, this would be time-consuming due to the requirement of real-world measurements. There have been a few previous studies in calibrating TD-DFT against experimental values,[28,29] but these used only small experimental datasets. There is a potential here to apply techniques such as text mining to extract experimental excited state data from published papers, although the differences in reporting may make this difficult.

The models can also be extended to other applications. For example, the graph-based genetic algorithm (GB-GA) developed by Jensen currently uses uncalibrated xTB-sTDA for excitation energies.[80] xTB-ML models could increase accuracy, and if calculations are sufficiently parallelized, this process could result in thousands of high-quality candidates being rapidly generated.

Finally, although xTB-ML is significantly faster than first-principles methods, it is still too slow to screen millions of molecules. As stated previously, with our setup, xTB-ML can calculate excited states of ~1500 molecules per hour (parallelized over four computer nodes) for molecules with <50 heavy atoms, such as those in PCQC. Therefore, it would take over three months to calculate all 3.5M molecules in PCQC. While this is a definite improvement over TD-DFT (over 3 years), this is still slow. Expanding to larger databases with bigger molecules would increase the runtime even further. Therefore, as a final direction, an optimized workflow using active learning could be implemented, intentionally searching for molecules with certain desired properties.

SUPPLEMENTARY MATERIAL

See the supplementary material for extended analysis of model training, testing, and validation.

ACKNOWLEDGMENTS

S.V. was supported by the Marshall Scholarship. This work made use of the CX1/2 high-performance computing clusters at Imperial College London.[82] The authors would like to thank Daniel Davies for useful discussions surrounding machine learning, Kyle Swanson for help and advice about using Chemprop, Jiali Li for helpful discussions regarding active learning, and Martijn Zwijnenburg and Stefano Angioletti-Uberti for fruitful discussions about computational chemistry. The authors also acknowledge the UK Materials and Molecular Modeling Hub for computational resources, which is partially funded by Engineering and Physical Sciences Research Council (EPSRC) (Grant Nos. EP/P020194/1 and EP/T022213/1).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The raw data and code to reproduce the figures presented in this paper are available in a repository on GitHub.[77] Trained ML models and prediction data are also available on GitHub.[77]

The code for the calibration workflow presented here is available in a repository on GitHub.[81] The scripts for running TD-DFT, running xTB-sTDA, training the ML model, and using it for predictions are available.

REFERENCES

[1] Y.-C. Cheng and G. R. Fleming, "Dynamics of light harvesting in photosynthesis," Annu. Rev. Phys. Chem. **60**, 241–262 (2009).
[2] Q. Wang, R. W. Schoenlein, L. A. Peteanu, R. A. Mathies, and C. V. Shank, "Vibrationally coherent photochemistry in the femtosecond primary event of vision," Science **266**, 422–424 (1994).
[3] T. Schultz, E. Samoylova, W. Radloff, I. V. Hertel, A. L. Sobolewski, and W. Domcke, "Efficient deactivation of a model base pair via excited-state hydrogen transfer," Science **306**, 1765–1768 (2004).
[4] M. B. Smith and J. Michl, "Singlet fission," Chem. Rev. **110**, 6891–6936 (2010).
[5] J. Zhao, S. Ji, and H. Guo, "Triplet–triplet annihilation based upconversion: From triplet sensitizers and triplet acceptors to upconversion quantum yields," RSC Adv. **1**, 937–950 (2011).
[6] O. El Bakouri, J. R. Smith, and H. Ottosson, "Strategies for design of potential singlet fission chromophores utilizing a combination of ground-state and excited-state aromaticity rules," J. Am. Chem. Soc. **142**, 5602–5617 (2020).
[7] K. J. Fallon, P. Budden, E. Salvadori, A. M. Ganose, C. N. Savory, L. Eyre, S. Dowland, Q. Ai, S. Goodlett, C. Risko, D. O. Scanlon, C. W. M. Kay, A. Rao, R. H. Friend, A. J. Musser, and H. Bronstein, "Exploiting excited-state aromaticity to design highly stable singlet fission materials," J. Am. Chem. Soc. **141**, 13867–13876 (2019).
[8] F. Spiegelman, N. Tarrat, J. Cuny, L. Dontot, E. Posenitskiy, C. Martí, A. Simon, and M. Rapacioli, "Density-functional tight-binding: Basic concepts and applications to molecules and clusters," Adv. Phys. X **5**, 1710252 (2020).
[9] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties," Phys. Rev. B **58**, 7260–7268 (1998).
[10] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, "Extended tight-binding quantum chemistry methods," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **11**, e1493 (2021).
[11] S. Grimme, "A simplified Tamm-Dancoff density functional approach for the electronic excitation spectra of very large molecules," J. Chem. Phys. **138**, 244104 (2013).
[12] C. Fang, B. Oruganti, and B. Durbeej, "How method-dependent are calculated differences between vertical, adiabatic, and 0–0 excitation energies?," J. Phys. Chem. A **118**, 4157–4171 (2014).
[13] S. Grimme and C. Bannwarth, "Ultra-fast computation of electronic spectra for large systems by tight-binding based simplified Tamm-Dancoff approximation (sTDA-xTB)," J. Chem. Phys. **145**, 054103 (2016).
[14] L. Wilbraham, R. S. Sprick, K. E. Jelfs, and M. A. Zwijnenburg, "Mapping binary copolymer property space with neural networks," Chem. Sci. **10**, 4973–4984 (2019).
[15] L. Wilbraham, E. Berardo, L. Turcani, K. E. Jelfs, and M. A. Zwijnenburg, "High-throughput screening approach for the optoelectronic properties of conjugated polymers," J. Chem. Inf. Model. **58**, 2450–2459 (2018).
[16] L. Wilbraham, D. Smajli, I. Heath-Apostolopoulos, and M. A. Zwijnenburg, "Mapping the optoelectronic property space of small aromatic molecules," Commun. Chem. **3**, 14 (2020).
[17] I. Heath-Apostolopoulos, L. Wilbraham, and M. A. Zwijnenburg, "Computational high-throughput screening of polymeric photocatalysts: Exploring the effect of composition, sequence isomerism and conformational degrees of freedom," Faraday Discuss. **215**, 98–110 (2019).

[18]I. Heath-Apostolopoulos, D. Vargas-Ortiz, L. Wilbraham, K. E. Jelfs, and M. A. Zwijnenburg, "Using high-throughput virtual screening to explore the opto-electronic property space of organic dyes; finding diketopyrrolopyrrole dyes for dye-sensitized water splitting and solar cells," Sustainable Energy Fuels **5**, 704–719 (2021).
[19]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Big data meets quantum chemistry approximations: The Δ-machine learning approach," J. Chem. Theory Comput. **11**, 2087–2096 (2015).
[20]P. Zheng, R. Zubatyuk, W. Wu, O. Isayev, and P. O. Dral, "Artificial intelligence-enhanced quantum chemical method with broad applicability," Nat. Commun. **12**, 7022 (2021).
[21]M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, and K. Burke, "Quantum chemical accuracy from density functional approximations via machine learning," Nat. Commun. **11**, 5223 (2020).
[22]A. Nandi, C. Qu, P. L. Houston, R. Conte, and J. M. Bowman, "**Δ**-machine learning for potential energy surfaces: A PIP approach to bring a DFT-based PES to CCSD(T) level of theory," J. Chem. Phys. **154**, 051102 (2021).
[23]J. Westermayr and R. J. Maurer, "Physically inspired deep learning of molecular excitations and photoemission spectra," Chem. Sci. **12**, 10755–10764 (2021).
[24]R. Ramakrishnan, M. Hartmann, E. Tapavicza, and O. A. von Lilienfeld, "Electronic spectra from TDDFT and machine learning in chemical space," J. Chem. Phys. **143**, 084111 (2015).
[25]R. Pollice, P. Friederich, C. Lavigne, G. d. P. Gomes, and A. Aspuru-Guzik, "Organic molecules with inverted gaps between first excited singlet and triplet states and appreciable fluorescence rates," Matter **4**, 1654–1682 (2021).
[26]D. Padula, Ö. H. Omar, T. Nematiaram, and A. Troisi, "Singlet fission molecules among known compounds: Finding a few needles in a haystack," Energy Environ. Sci. **12**, 2412–2416 (2019).
[27]K. Zhao, Ö. H. Omar, T. Nematiaram, D. Padula, and A. Troisi, "Novel thermally activated delayed fluorescence materials by high-throughput virtual screening: Going beyond donor–acceptor design," J. Mater. Chem. C **9**, 3324–3333 (2021).
[28]X. Wang, L. Wong, L. Hu, C. Chan, Z. Su, and G. Chen, "Improving the accuracy of density-functional theory calculation: The statistical correction approach," J. Phys. Chem. A **108**, 8514–8525 (2004).
[29]H. Li, L. Shi, M. Zhang, Z. Su, X. Wang, L. Hu, and G. Chen, "Improving the accuracy of density-functional theory calculation: The genetic algorithm and neural network approach," J. Chem. Phys. **126**, 144101 (2007).
[30]S. Grimme, L. Goerigk, and R. F. Fink, "Spin-component-scaled electron correlation methods," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 886–906 (2012).
[31]C. Suellen, R. G. Freitas, P.-F. Loos, and D. Jacquemin, "Cross-comparisons between experiment, TD-DFT, CC, and ADC for transition energies," J. Chem. Theory Comput. **15**, 4581–4590 (2019).
[32]P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields," J. Phys. Chem. **98**, 11623–11627 (1994).
[33]D. Jacquemin, V. Wathelet, E. A. Perpète, and C. Adamo, "Extensive TD-DFT benchmark: Singlet-excited states of organic molecules," J. Chem. Theory Comput. **5**, 2420–2435 (2009).
[34]T. Yanai, D. P. Tew, and N. C. Handy, "A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP)," Chem. Phys. Lett. **393**, 51–57 (2004).
[35]A. M. Grabarz and B. Ośmiałowski, "Benchmarking density functional approximations for excited-state properties of fluorescent dyes," Molecules **26**, 7434 (2021).
[36]I. Y. Zhang, J. Wu, and X. Xu, "Extending the reliability and applicability of B3LYP," Chem. Commun. **46**, 3057–3070 (2010).
[37]G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. v. Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," New J. Phys. **15**, 095003 (2013).
[38]M. Nakata and T. Shimazaki, "PubChemQC project: A large-scale first-principles electronic structure database for data-driven chemistry," J. Chem. Inf. Model. **57**, 1300–1308 (2017).
[39]B. G. Abreha, S. Agarwal, I. Foster, B. Blaiszik, and S. A. Lopez, "Virtual excited state reference for the discovery of electronic materials database: An open-access resource for ground and excited state properties of organic molecules," J. Phys. Chem. Lett. **10**, 6835–6841 (2019).
[40]J. Liang, S. Ye, T. Dai, Z. Zha, Y. Gao, and X. Zhu, "QM-symex, update of the QM-sym database with excited state information for 173 kilo molecules," Sci. Data **7**, 400 (2020).
[41]J. Liang, Y. Xu, R. Liu, and X. Zhu, "QM-sym, a symmetrized quantum chemistry database of 135 kilo molecules," Sci. Data **6**, 213 (2019).
[42]Ö. H. Omar, T. Nematiaram, A. Troisi, and D. Padula, "Organic materials repurposing, a data set for theoretical predictions of new applications for existing compounds," Sci. Data **9**, 54 (2022).
[43]R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, and A. Aspuru-Guzik, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," Nat. Mater. **15**, 1120–1127 (2016).
[44]N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler, and S. P. Russo, "Machine learning property prediction for organic photovoltaic devices," npj Comput. Mater. **6**, 166 (2020).
[45]E. O. Pyzer-Knapp, G. N. Simm, and A. A. Guzik, "A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials," Mater. Horiz. **3**, 226–233 (2016).
[46]S. A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, and A. Aspuru-Guzik, "Design principles and top non-fullerene acceptor candidates for organic photovoltaics," Joule **1**, 857–870 (2017).
[47]D. Jacquemin, B. Mennucci, and C. Adamo, "Excited-state calculations with TD-DFT: From benchmarks to simulations in complex environments," Phys. Chem. Chem. Phys. **13**, 16987–16998 (2011).
[48]Elsevier Developer Portal, 2021.
[49]M. C. Swain and J. M. Cole, "ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature," J. Chem. Inf. Model. **56**, 1894–1904 (2016).
[50]S. Kim, P. A. Thiessen, T. Cheng, B. Yu, and E. E. Bolton, "An update on PUG-REST: RESTful interface for programmatic access to PubChem," Nucleic Acids Res. **46**, W563–W570 (2018).
[51]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys. **148**, 241733 (2018).
[52]L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv:1802.03426 [cs, stat] (2020).
[53]N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," J. Cheminf. **3**, 33 (2011).
[54]C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions," J. Chem. Theory Comput. **15**, 1652–1671 (2019).
[55]sTDA-xTB for ground state calculations, 2021, original-date: 2019-11-27T12:03:12Z.
[56]B. Ramsundar, P. Eastman, P. Walters, and V. Pande, *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*, 1st ed. (O'Reilly Media, Sebastopol, CA, 2019).
[57]D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2015), Vol. 28.
[58]O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," arXiv:1511.06391 [cs, stat] (2016).
[59]K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay, "Analyzing learned molecular representations for property prediction," J. Chem. Inf. Model. **59**, 3370–3388 (2019).
[60]J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine*

Learningv (PMLR) (Proceedings of Machine Learning Research (PMLR), 2017), pp. 1263–1272.
[61] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).
[62] J. Westermayr, F. A. Faber, A. S. Christensen, O. A. von Lilienfeld, and P. Marquetand, "Neural networks and kernel ridge regression for excited states dynamics of $CH_2NH_2^+$: From single-state to multi-state representations and multi-property machine learning models," Mach. Learn.: Sci. Technol. **1**, 025009 (2020).
[63] R. Zubatyuk, J. S. Smith, J. Leszczynski, and O. V. Isayev, "Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network," Sci. Adv. **5**, eaav6490 (2019).
[64] Z. Tan, Y. Li, W. Shi, and S. Yang, "A multitask approach to learn molecular properties," J. Chem. Inf. Model. **61**, 3824–3834 (2021).
[65] J. Westermayr and P. Marquetand, "Machine learning and excited-state molecular dynamics," Mach. Learn.: Sci. Technol. **1**, 043001 (2020).
[66] P. O. Dral and M. Barbatti, "Molecular excited states through a machine learning lens," Nat. Rev. Chem. **5**, 388–405 (2021).
[67] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, "PubChem in 2021: New data content and improved web interfaces," Nucleic Acids Res. **49**, D1388–D1395 (2020).
[68] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17," J. Chem. Inf. Model. **52**, 2864–2875 (2012).
[69] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, edited by J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu (Springer, Berlin, Heidelberg, 2013), pp. 160–172.
[70] scikit-learn-contrib/hdbscan, 2022, original-date: 2015-04-22T13:32:37Z.
[71] RDKit.
[72] L. Wilbraham, "molZ," 2021, original-date: 2020-12-06T13:18:42Z.
[73] RDKit.Chem.Fragments module—The RDKit 2021.09.1 documentation; available at http://www.rdkit.org.
[74] O. Christiansen, H. Koch, and P. Jørgensen, "The second-order approximate coupled cluster singles and doubles model CC2," Chem. Phys. Lett. **243**, 409–418 (1995).
[75] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 140022 (2014).
[76] C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The PBE0 model," J. Chem. Phys. **110**, 6158–6170 (1999).
[77] S. Verma, "xTB-ML-data," 2022, original-date: 2021-12-17T17:42:28Z; available at https://doi.org/10.5281/zenodo.6391015.
[78] P. Pracht, F. Bohle, and S. Grimme, "Automated exploration of the low-energy chemical space with fast quantum chemical methods," Phys. Chem. Chem. Phys. **22**, 7169–7192 (2020).
[79] B. Kang, C. Seok, and J. Lee, "Prediction of molecular electronic transitions using random forests," J. Chem. Inf. Model. **60**, 5984–5994 (2020).
[80] J. H. Jensen, "A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space," Chem. Sci. **10**, 3567–3572 (2019).
[81] S. Verma, "xTB-ML-workflow," 2022, original-date: 2021-12-17T17:53:32Z; https://doi.org/10.5281/zenodo.6391017.
[82] Imperial College Research Computing Service, https://doi.org/10.14469/hpc/2232, 2021.