Electronic spectra from TDDFT and machine learning in chemical space

Raghunathan Ramakrishnan, Mia Hartmann, Enrico Tapavicza' , and O. Anatole von Lilienfeld'

Citation: The Journal of Chemical Physics **143**, 084111 (2015); doi: 10.1063/1.4928757

View online: http://dx.doi.org/10.1063/1.4928757

View Table of Contents: http://aip.scitation.org/toc/jcp/143/8

Published by the American Institute of Physics

Electronic spectra from TDDFT and machine learning in chemical space

Raghunathan Ramakrishnan,[1] Mia Hartmann,[2] Enrico Tapavicza,[2,a)]
and O. Anatole von Lilienfeld[1,3,b)]
[1]*Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*
[2]*Department of Chemistry and Biochemistry, California State University, 1250 Bellflower Boulevard, Long Beach, California 90840, USA*
[3]*Argonne Leadership Computing Facility, Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, Illinois 60439, USA*

(Received 9 April 2015; accepted 7 August 2015; published online 25 August 2015)

Due to its favorable computational efficiency, time-dependent (TD) density functional theory (DFT) enables the prediction of electronic spectra in a high-throughput manner across chemical space. Its predictions, however, can be quite inaccurate. We resolve this issue with machine learning models trained on deviations of reference second-order approximate coupled-cluster (CC2) singles and doubles spectra from TDDFT counterparts, or even from DFT gap. We applied this approach to low-lying singlet-singlet vertical electronic spectra of over 20 000 synthetically feasible small organic molecules with up to eight CONF atoms. The prediction errors decay monotonously as a function of training set size. For a training set of 10 000 molecules, CC2 excitation energies can be reproduced to within ±0.1 eV for the remaining molecules. Analysis of our spectral database via chromophore counting suggests that even higher accuracies can be achieved. Based on the evidence collected, we discuss open challenges associated with data-driven modeling of high-lying spectra and transition intensities. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4928757]

I. INTRODUCTION

Quantum mechanical rational compound design strategies[1,2] to model molecular valence electronic spectra hold great promise to narrow down the discovery of novel photonic and optoelectronic devices. Potential applications include the fabrication of low cost dye-sensitized solar cells,[3] organic light emitting diodes,[4] photosensitizers that are inert to environmental factors but useful in photodynamic therapy,[5] and organic ultraviolet (UV) filters (aka sunscreens) in cosmetics.[6] For any given compound, the relevant prediction accuracy can readily be attained with an established excited state wavefunction method. Successful studies include the quantitative description of solar cell materials,[7] organic diodes,[8] and even biologically relevant phenomena such as photo-induced dynamics of vitamins B2[9] and D.[10] For a robust forecast, depending on computational budget, one can also select a method according to the most appropriate cost-to-performance ratio from the series of equations of motion (EOM) or linear response (LR) variants of the coupled cluster (CC) theories CCS, second-order approximate coupled-cluster (CC2), CCSD, CC3, and coupled-cluster theory with single, double, and triple excitations (CCSDT). These methods scale from $O(N^4)$ to $O(N^8)$, where N is the number of orbitals.[11] When increasing size, or number of molecules, the next viable compromise between accuracy and computational complexity is linear response time-dependent density functional theory (LR-TDDFT) within the adiabatic approximation.[12,13] TDDFT, commonly based on local or semi-local exchange correlation (XC) functionals, has been shown to yield qualitatively inaccurate predictions whenever the valence excitations involve charge-transfer (CT),[14] and the adiabatic approximation fails to accurately describe transitions with double excitation character.[15] Such qualitative failure of TDDFT, hard to anticipate without visual inspection of molecular orbitals involved in the transitions, dramatically reduces its usefulness for high-throughput screening of molecules with interesting electronic spectra. Application of CC methods for large scale computation is already prohibitive even when considering just electronic ground state properties of small sub-fractions of the known small molecule chemical universe, such as the GDB-17 with over 166×10^9 organic molecules with no more than 17 atoms (not counting hydrogens).[16]

For combinatorially and computationally hard problems, such as navigating chemical space in quest of an optimal electronic spectra,[17] statistical inference from large volumes of data offers an appealing alternative to the conventional strategies of investing in ever more sophisticated approximations, faster hardware, or more efficient programming. Statistical learning has already contributed to scientific progress in biology[18] or climate research.[19] Inspired by the success of such efforts, several computational chemistry studies have recently made use of supervised machine learning (ML) models to infer quantum mechanical properties of query molecules from those of a set of example molecules, computed *a priori*. Effectively, this amounts to modeling expectation values calculated with approximate solutions to the electronic Schrödinger equation, most notably the energy.[20] By now, ML models have been shown to reach the highly coveted quantum chemical accuracy for many different ground-state molecular

a)Enrico.Tapavicza@csulb.edu
b)anatole.vonlilienfeld@unibas.ch

properties.[21–23] As such, also quantum mechanical expectation values can be interpolated in chemical space.[17] Improvement of molecular models of chemical properties based on molecular similarity[24,25] is also related to this approach. These developments have also inspired studies on transition state dividing surfaces,[26] orbital-free kinetic energy density functionals,[27] electronic properties of crystals,[28] transmission coefficients in nano-ribbon models,[29] or densities of states in Anderson impurity models.[30] More recently, a single kernel has been introduced for the simultaneous modeling of multiple electronic ground-state properties for training-sets comprised of up to 40 000 molecules.[31] Here, we report on our findings when trying to apply these ML methods to infer properties of molecules in their electronically excited states. More specifically, we discuss ML models which combine CC accuracy with DFT efficiency.

II. METHODS

A. Δ-ML model of excited states properties

In Ref. 23, some of us introduced the Δ-ML Ansatz to estimate molecular ground-state properties from an expensive targetline theory, at the computational cost of an inexpensive baseline theory (B). The ML model for the quantitative prediction of molecular electronic spectra is built in analogy, using ML models of the deviation of TDDFT excited state properties from CC2 reference numbers. We approximate an electronic static property, p_i, corresponding to the ith excited state of query molecule q at CC2 level of theory as the sum of baseline prediction and a linear combination of exponentially decaying functions in molecular similarity to training molecules t,

$$p_i^{\mathrm{CC2}}(\mathbf{d}_q) \approx p_i^{\mathrm{B}}(\mathbf{d}_q) + \sum_{t=1}^{N} c_{it} \mathrm{e}^{-|\mathbf{d}_q - \mathbf{d}_t|/\sigma}, \tag{1}$$

where N is the number of molecules in training set, σ is the kernel width, and $|\mathbf{d}_q - \mathbf{d}_t|$ corresponds to the Manhattan (L_1) norm between molecular descriptors $\mathbf{d}$ (*vide infra*). A previous study[22] benchmarked the performance of various norms in the above equation when directly modeling atomization energies (no baseline), and found the L_1 norm to yield the lowest cross-validated errors. The second term on the right side of Eq. (1) therefore models exclusively the error in baseline method B's estimate of p_i when compared to CC2 for query molecule q,

$$\Delta p_i^{\mathrm{est}}(\mathbf{d}_q) = \sum_{t=1}^{N} c_{it} \mathrm{e}^{-|\mathbf{d}_q - \mathbf{d}_t|/\sigma}. \tag{2}$$

In this study, we have investigated two excited state properties, p_i, namely, excitation energy (with respect to the ground electronic state), E_i, and oscillator strength, f_i, for the lowest two ($i = 1, 2$) singlet electronic states. Other excited states properties could have also been considered with this generic approach. Due to their popularity, we have selected for this study DFT and TDDFT as baseline B, and CC2 as targetline. The CC2 method, with a triple-zeta basis set, has been shown to predict experimental valence excitation energies with a mean absolute error (MAE) of 0.12 eV.[32] This error decreases slightly to 0.10 eV, when CC2 is compared to the computationally more demanding method, CC3.[33] To serve as a reference method, in this ML study, we therefore consider CC2 to represent the optimal compromise between sufficient accuracy and acceptable computational cost. To compare the impact of the baseline on the Δ-ML strategy, we have considered various DFT[34,35] baseline theories with increasing sophistication. However, any other combination of methods could have been chosen just as well. Our simplest non-zero baseline for $p_i = E_i$ is the HOMO-LUMO gap of the ground-state computed using DFT-PBE0.[36–39] We also consider p_i from LR-TDDFT[13,40] using the hybrid functionals PBE0 and CAM-B3LYP.[41]

In the following, we use matrix notations compatible with Ref. 31, and denote matrices by capital bold, and vectors by small bold cases. Regression coefficients corresponding to training molecules, $\{c_{it}\}$, have been obtained as solutions to

$$(\mathbf{K} + \lambda \mathbf{I})\, \mathbf{c}_i = \mathbf{p}_i^{\mathrm{CC2}} - \mathbf{p}_i^{\mathrm{B}} = \Delta \mathbf{p}_i^{\mathrm{ref}}, \tag{3}$$

where $\mathbf{I}$ and $\mathbf{K}$ are the identity and kernel matrices, respectively, the latter with elements $K_{st} = \mathrm{e}^{-|\mathbf{d}_s - \mathbf{d}_t|/\sigma}$. Note that in ML literature, the exponential kernel function is also denoted as Laplace kernel, owing to the fact that the exponential function, in certain coordinate systems, is a solution to Laplace's equation. Eq. (3) minimizes the λ-regularized (λ quantifies the regularization strength) least-squares error in estimations[30]

$$\underset{\mathbf{c}_i}{\text{minimize}} \quad \|\Delta \mathbf{p}_i^{\mathrm{ref}} - \Delta \mathbf{p}_i^{\mathrm{est}}\|_2^2 + \lambda \mathbf{c}_i^{\mathrm{T}} \mathbf{K} \mathbf{c}_i, \tag{4}$$

where $\|\cdot\|_2$ stands for L_2 norm of a vector, $(\cdot)^{\mathrm{T}}$ denotes transpose operation, and $\Delta \mathbf{p}_i^{\mathrm{est}}$ is defined in Eq. (2). Derivation of Eq. (3) from Eq. (4) is presented as Appendix.

B. Cross-validation

Overfitting of the kernel models to training molecules is typically avoided by optimizing the two hyperparameters (σ, λ) through five-fold cross-validation (CV). In this procedure, N training molecules are randomly distributed into 5 bins, each with $N/5$ molecules. Every bin is used once as a test (or validation) set, while the remaining four bins act as training sets. Hyperparameters are optimized such that they minimize the model's MAE for the test-bin. Here, we employed Nelder-Mead's simplex method[42] for the 2D optimization. The cross-validation procedure is the most time-consuming process in training the ML model, with each evaluation of MAE of the test-bin requiring $O(n^3)$ scaling matrix inversion operations, where $n = 4N/5$. In the present work, n is at most 8k. This results in roughly one central processing unit (CPU) day of training for a fully converged ML model.

For larger training set sizes, CVs become prohibitive, one can employ the property-independent "single-kernel" Ansatz,[31] with optimal hyperparameters estimated exclusively from the structures of the training molecules. This approach assumes the training data to be devoid of outliers, and enforces λ to be a fixed, property-independent scalar (typically set, or close, to zero). The width of the kernel function can be chosen according to some heuristic, for example, such that the maximal value of the off-diagonal elements of $\mathbf{K}$ is 1/2, which renders the kernel sufficiently global to have all

training molecules contribute in the generation of the regression weights $\{c_{it}\}$. For the exponential (aka Laplace) kernel, with L_1 distance metric, $K_{ij} = e^{-|\mathbf{d}_i - \mathbf{d}_j|/\sigma}$, this constraint results in $\sigma = \max\{|\mathbf{d}_i - \mathbf{d}_j|\}/\log(2)$. In Ref. 31, we have demonstrated a universal kernel derived in this fashion to enable systematic reduction of out-of-sample prediction errors for 13 molecular ground-state properties of 112k molecules, using up to 40k training molecules. Here, in order to accelerate the CV procedure, we have made use of this heuristic as an initial guess. For the molecular datasets considered, these values in atomic units are typically $\sigma = 1000$ and $\lambda = 0$. After CV, globally optimal hyperparameters have been obtained by taking the median of the 5 folds. A median value is considered instead of mean, because the median of a distribution is not influenced by extreme values, such as the hyperparameters that could be found for a test bin with extreme outliers in structure or property. A final kernel with globally optimized hyperparameters is used for the prediction of properties of out-of-sample molecules that took no part in training.

C. Choice of molecular descriptor

In order to assess the effect of the molecular representation on the ML model's performance, we report results based on two definitions of molecular representations, namely, the Coulomb-matrix (CM) with atom indices sorted by norm of rows in order to reach invariance with respect to permutation of identical nuclei,[20] as well as a recently introduced more compact variant of the CM, called bag-of-bonds (BOB).[43] The elements of the CM[20] are defined as

$$M_{II} = 0.5Z_I^{2.4}, \qquad M_{IJ} = Z_I Z_J / R_{IJ}, \tag{5}$$

where I and J are atomic indices, R_{IJ} is the interatomic distance, and Z is the atomic number. The off-diagonal elements of a CM uniquely represent the geometry and atomic composition of a molecule,[44] while the diagonal elements provide a simple exponential fit to the negative of the potential energy of the neutral atoms. As such, the diagonal is similar to the total potential energy of a neutral atom within Thomas-Fermi theory, $E_{\mathrm{TF}} = -0.77Z^{7/3}$, or its modifications with a Z-dependent prefactor in the range 0.4–0.7.[45] It is sufficient to consider only the lower or upper triangle of the CM. In order to enable comparisons between two molecules with different number of atoms, the CM matrix of the smaller molecule is padded with zero elements.

BOB is a labeled set of off-diagonal CM elements which enables the comparison of pairwise distance between any given combination of two atom types. For instance, for H_2O, BOB is the set of two sorted row vectors, $\{[M_{\mathrm{HO}}, M_{\mathrm{HO}}], [M_{\mathrm{HH}}]\}$, with elements corresponding to the CM entries. Due to the pairwise partitioning, however, any two homometric molecules with identical stoichiometry will yield a zero descriptor difference according to BOB.[44] As such, BOB does not uniquely represent molecules. The CM, by contrast, is able to uniquely encode any molecule, up to its enantiomers. The molecular dataset considered in this study, however, is devoid of homometric molecules. In addition to the aforementioned sorted CM matrix, BOB has also been tested since it has been shown to yield slightly better accuracy for the prediction of molecular atomization energies.[43] In general, we have found that for large N, both CM and BOB converge towards similar prediction accuracy for energy-related properties. For smaller training sets, however, BOB typically exhibits a more substantial advantage.

D. Excited states data

We have relied on the recently published molecular quantum chemistry database with relaxed geometries computed using DFT B3LYP with basis set 6-31G(2df,p).[46] This data set corresponds to the smallest 133 885 (134k) organic molecules with up to 9 CONF atoms out of the list of 166 B synthetically feasible organic molecules, called GDB-17 database, and published by Reymond and co-workers.[16] For this study, we have eliminated 3054 molecules from the 134k dataset due to high steric strain in the B3LYP/6-31G(2df,p) geometries,[46] and we further have limited ourselves to those 21 800 molecules with only up to 8 CONF atoms. For these molecules, we have performed single point calculations using the program TURBOMOLE[47] to compute the ground (S_0),[48] and the lowest two vertical electronic excited states (S_1 and S_2) of singlet spin-symmetry. We also performed calculations at the LR-TDDFT[40] level employing the hybrid XC functional PBE0[37,39] with def2SVP basis set,[49] and at the resolution-of-identity approximate coupled cluster with singles and doubles substitution (RI-CC2)[50] level with def2TZVP basis set.[49] Using the larger basis set, we also performed LR-TDDFT calculations with PBE0, and CAM-B3LYP[41] functionals, the latter using the program Gaussian09.[51]

All calculations were performed with C_1 symmetry and in DFT calculations default integral grids were employed to compute the XC energy contributions. For 7 molecules (most of them highly symmetric, e.g., cubane), the RI-CC2 calculations did not converge the first excited state wavefunction. For 7 other molecules (with multiple CO groups, e.g., 2,3-dioxobutanedial), emission has been found, i.e., negative lowest transition energy, presumably arising from orbital relaxation. For the purpose of this study, we have removed these exotic 14 molecules. The lowest two singlet transition energies, as well as corresponding oscillator strengths in length-representation, have been used for the remaining 21 786 molecules, to which we refer in the following as the set of 22k molecules. All indices of the 22k GDB-8 molecules along with corresponding TDDFT and CC2 results are given as supplementary material in `gdb8_22k_elec_spec.txt`.[65] The indices enable retrieval of geometries from the 134k GDB-9 dataset.[46]

III. RESULTS AND DISCUSSION

A. Excitation energies and oscillator strengths for 22k organic molecules

The smoothened distribution of CC2 predicted $S_0 \rightarrow S_1$ transition energies E_1 and corresponding oscillator strengths f_1 features in Figure 1 for all 22k molecules. This 2D count density has been computed via kernel-density estimation.[52,53] The first excitation energy distribution is bimodal (see also Fig. 2 for the 1D projection), corresponding to one Gaussian

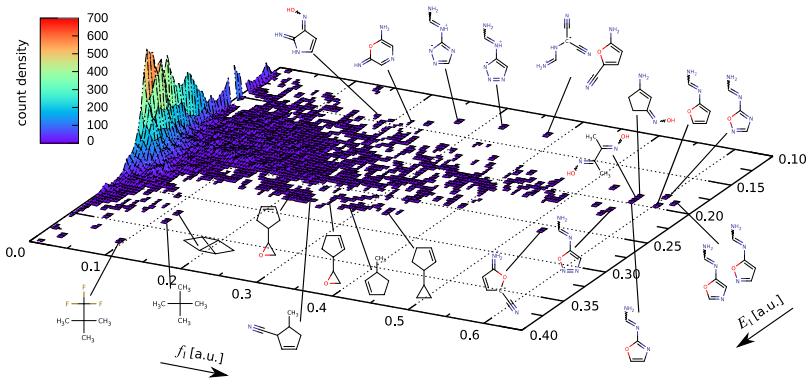

FIG. 1. Joint distributions of oscillator strength f_1, and transition energy E_1 for the first electronic excited singlet state of the 22k organic molecules. All values correspond to the RICC2/def2TZVP level of theory. For selected E_1 values, representative molecules with large f_1 are shown as insets.

centered at 0.18 a.u. with small variance, and another centered near 0.26 a.u. with significantly larger variance (the shoulder possibly implying two peaks, rather than one broad peak). Collectively, the 22k molecules span the spectral range of UV-B and UV-C, with few molecules in the UV-A region (>300 nm or <0.15 a.u.). The lack of transitions in the visible region is consistent with the fact that small organic molecules typically exhibit an energy gap of >5 eV between highest and lowest occupied molecular orbital, HOMO and LUMO, respectively. Not surprisingly, when proceeding from low to high transition energy regions, one notices that molecules gradually turn from being aromatic, or highly unsaturated, into increasingly saturated structures. The oscillator strength (f_1), by contrast, leads to an exponentially decaying distribution, with the largest fraction of compounds in the 22k set having negligible or zero values. A small minority of molecules, however, have significant f_1-magnitude, implying potential usefulness of these molecules as components in metal-free organic sensitizers.[54] Only a dozen molecules, highlighted in Figure 1, display $f_1 > 0.5$, resulting in light harvesting efficiencies[55] larger than $100 \times (1 - 10^{-0.5}) \approx 68\%$. These molecules contain ketoxime, R(R′)C = NOH, or amidine, R-C(NH$_2$) = NR′, chromophores. They all exhibit push-pull type conjugation of π-bonds, with electron-donating, and electron-withdrawing groups on opposite ends, resulting in highly polarized electron densities. However, also the symmetric molecule (point group C$_{2h}$), dimethylglyoxime, a chelating agent commonly used in gravimetric analysis of nickel, has a large oscillator strength for its first excitation with $f_1 = 0.56$ at $E_1 = 0.2$ a.u.

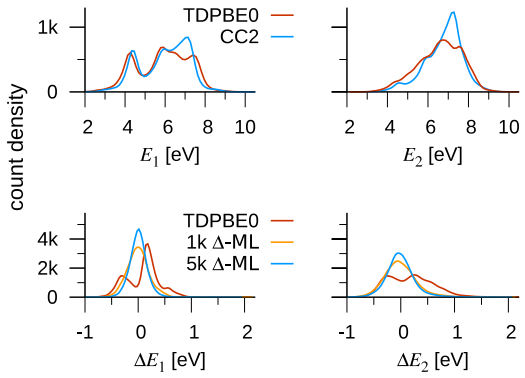

FIG. 2. Distributions, smoothened by 1D kernel density estimation as implemented in GNUPLOT,[56] of spectral properties and predicted errors. Top: densities of first and second singlet transition energies (E_1 and E_2, respectively, in eV) of 17k organic molecules with up to eight CONF atoms, at the CC2, and TDPBE0 levels of theory. Bottom: error distribution for E_1 (left) and E_2 (right) with respect to CC2. Errors are given for TDPBE0 and Δ-ML models based on 1k, and 5k training molecules with TDPBE0 baseline.

The effect of level of theory is shown for TDPBE0 and CC2 predictions of E_1 and E_2 in the top panel of Figure 2. For both states, TDDFT leads to a depletion in count densities at ≈7 eV when compared to the CC2 distribution, compensated by overestimated densities in low and high energy regions. In the following, we will discuss how to mitigate this depletion using ML models.

B. ML models of excitation energies

Despite the obvious differences in prediction in the top panel of Fig. 2, the Δ-ML model of Eq. (1) captures the necessary correction. This is illustrated by the signed error distributions (with respect to CC2) in the bottom panel of Figure 2, for both excitation energies. Distributions are shown for Δ-ML models trained on molecular sub-sets containing either N = 1k or N = 5k molecules, drawn at random from the 22k data set. All ML results discussed in this paper, including these distributions, correspond to out-of-sample predictions for the remaining (22k - N) molecules. For comparison, the TDDFT deviation from CC2 is also shown in the bottom panel of Fig. 2 for both transition energies, resulting in a bimodal distribution which suggests that systematic errors are present. These errors can be either due to PBE0 kernel or smaller basis-set, or both. The ML errors, by contrast, are normally distributed around zero, with increasing and decreasing height and width, respectively, as one increases the training set from 1k to 5k. This implies that the Δ-ML model is properly accounting for the systematic errors in the TDDFT predictions, replacing them by a normal error distribution. MAEs of the TDDFT predictions amount to 0.27 and 0.37 eV, for E_1 and E_2, respectively. These MAEs are reduced to 0.16 and 0.23 eV for the 1k ML models and to 0.13 and 0.20 eV for the 5k ML models. We have also investigated the effect of the other aforementioned descriptor in the ML model, BOB. BOB results in ML prediction errors of 0.13/0.20 and 0.09/0.16 eV for E_1/E_2, using models trained on 1k and 5k training sets, respectively—slightly better than the corresponding CM predictions.

In order to investigate in a systematic fashion the performance of the Δ-ML model, we have calculated out-of-sample MAEs of E_1 predictions for various baseline methods. In Figure 3, the resulting MAEs are shown as a function of

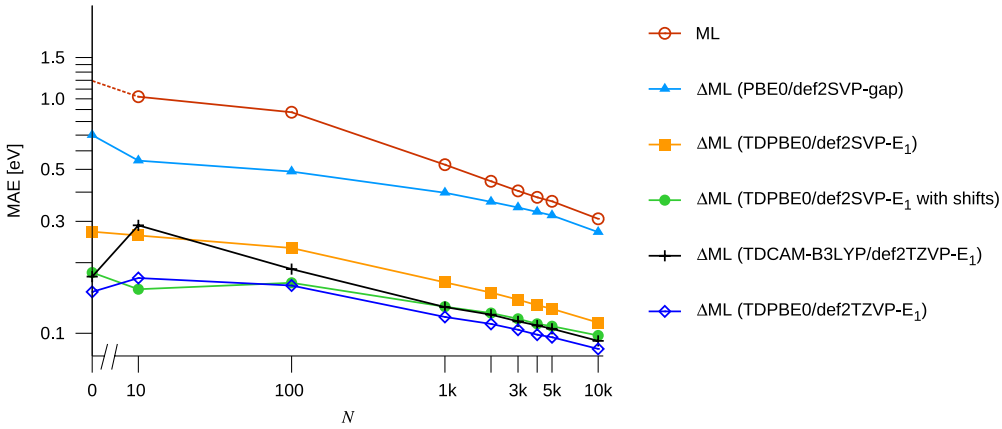

FIG. 3. Systematic improvement of ML models of singlet-singlet transition energies (E_1). Mean absolute error (MAE in eV) with respect to reference CC2/def2TZVP values is shown as a function of training set size (N) for 22k−N out-of-sample predictions. Various baseline methods are shown. The value of the baseline-free ML model (red) at $N = 0$ corresponds to the CC2 standard deviation in the 22k test set. Other baselines include HOMO-LUMO gap (blue), TDPBE0 E_1 without (yellow), and with (green) bivariate systematic shift corrections which explicitly account for σ and π chromophores. Also included are def2TZVP baseline results for TDCAM-B3LYP E_1 (black) and TDPBE0 E_1 (blue). Baseline errors at $N = 0$ correspond to standard deviations, obtained after subtraction of an average shift with respect to the CC2-targetline.

training set size N for N = 0 (i.e., the error of the baseline method), 10, 100, 1k, 2k, 3k, 4k, 5k, and 10k. More specifically, zero baseline results correspond to setting E_1^{B} to zero in Eq. (1). We also used the PBE0 HOMO-LUMO gap as a baseline, as well as TDPBE0 and TDCAM-B3LYP. As one would expect, the predictive accuracy improves with increasing level of sophistication of the baseline: The zero, gap, and TD baseline with def2SVP basis set yield 0.4, 0.3, and 0.13 eV, respectively, for the most accurate model trained on N = 10k molecules. Increasing the basis set from def2SVP to def2TZVP improves PBE0's baseline value, eventually resulting in a very small MAE of 0.08 eV for 10k Δ-ML. These observations are in line with previous benchmark studies[57] which concluded the TDCAM-B3LYP is somewhat inferior to TDPBE0 for the prediction of singlet-to-singlet excitation energies of small molecules. Overall, it is encouraging that all models, no matter which baseline, converge towards the same learning rate, i.e., slope on the log-log scale of error versus training set size. As such, the baseline merely leads to a difference in off-sets — which could also be compensated for by adding more training data. Due to the immense size of chemical space,[17] the addition of more molecules can easily be envisioned. For Δ-ML models of E_2, similar curves can be obtained, albeit slightly off-set yielding less accurate predictive power.

C. Inclusion of systematic shifts

It is not obvious to us that there is a single reason for TDPBE0/def2SVP's substantial underestimation of first and second transition energies near 7 eV, see Figure 2. A simple pattern, however, emerges after splitting the 22k set into saturated and unsaturated molecules, i.e., into two sets containing either π- or σ-chromophores. The corresponding signed error densities for the two sets are well separated, as shown for E_1 in Figure 4. They are centered around −0.31 and +0.19 eV for the saturated σ and unsaturated π-chromophores, respectively. The systematic underestimation of TDPBE0-based E_1 of π-type excitations ($\pi \rightarrow \pi^*$ or $n \rightarrow \pi^*$) is a well-known issue of approximate XC functionals when it comes to the description of CT-type excitations,[14] i.e., transitions with small overlap between donor and acceptor orbital overlap.[58] Our results are consistent with this finding, strengthening the indications that the underestimation of E_1 is universal for all π-type excitations. Furthermore, the other distribution in Figure 4 clearly shows a systematic overestimation of TDPBE0-based E_1 of σ-type excitations ($\sigma \rightarrow \sigma^*$ or $n \rightarrow \sigma^*$). This systematic blue shift of TDPBE0 E_1 is, at least partly, due to the finiteness of the relatively small basis set (def2SVP) used. This reasoning is in line with the variational principle: The difference between the lowest two eigenvalues of the molecular Hamiltonian is always larger when represented in a small basis set than when compared to the complete basis set limit. For instance, using literature values[59] of the HOMO-LUMO gap of the water molecule, we note the PBE0 value with the minimal basis set, STO-3G, to be 13.3 eV, overestimating more converged basis set PBE0 results by roughly 4.6 eV.

The degree of saturation can easily be detected beforehand using SMILES strings. We can therefore readily exploit

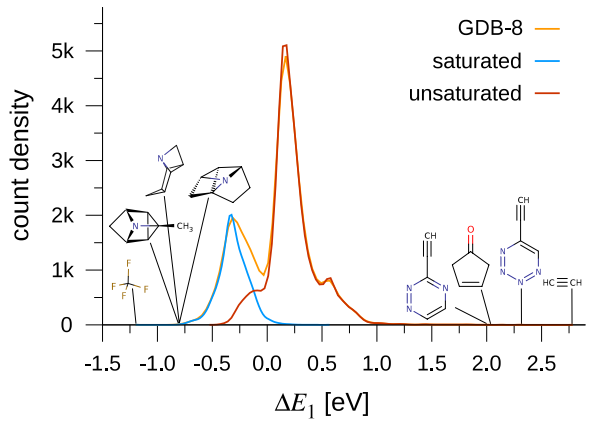

FIG. 4. Bivariate error distribution of the TDPBE0/def2SVP lowest singlet-singlet transition energies (E_1 in eV) of 22k organic molecules with up to eight CONF atoms (yellow). Partitioned error distributions over saturated (blue) and unsaturated (red) molecules are shown as well. The molecular structures correspond to extreme outliers for TDPBE0/def2SVP.

this knowledge by subtraction of the distribution's centered value −0.31 and +0.19 eV from the baseline number for saturated and unsaturated chromophores, respectively. The resulting TDPBE0 Δ-ML model in Eq. (1) improves indeed: The out-of-sample MAE decreases at a lower off-set with training set size, as shown in Figure 3, yet at similar learning rates as the other models. For the 10k model of the E_1 transition energy, the MAE is found to decrease from 0.13 eV to 0.1 eV. It is interesting to note that the performance of the TDCAM-B3LYP/def2TZVP level is virtually identical with the shifted TDPBE0/def2SVP result ($N = 0$), as well as for larger N values. For smaller training sets ($N = 10$ or 100), the shifted TDPBE0/def2SVP Δ-ML model even outperforms the corresponding TDCAM-B3LYP/def2TZVP variant.

D. DFT and ML model outliers

It is always interesting to consider the worst predictions of a model. The average errors discussed so far neither imply better ML predictions for DFT outliers nor do they quantify the ML outliers. Here, we briefly discuss the accuracy of predicted E_1 for the 10 most extreme outliers among all out-of-sample molecules, i.e., all molecules that were not part of the training sets for the 1k and 5k Δ-ML models. Table I lists SMILES strings of the corresponding molecules, model prediction errors, and CC2 numbers for comparison. The 10 outliers are sorted by their TDPBE0, 1k, or 5k ML model deviation. As also already indicated in Figure 4, the worst DFT outliers correspond to unsaturated molecules. This observation holds true for the 10 most extreme DFT outliers in Table I, deviating by up to 2.15 eV from CC2. These molecules could be of interest as benchmarks for developing improved DFT kernels for TDDFT calculations. The numbers in Table I show that for all outliers, the 5k ML model yields better performance than DFT, while the 1k ML model improves all predictions but the one for the worst, namely cyclopenta-1-en-4-on (O=C1CC=CC1). This molecule is also shown in Fig. 4. Note that other outliers shown in that figure have been part of the training set and therefore do not feature in Table I. The finding that the ML models also improve on the baseline method's outliers agrees with conclusions drawn in a previous finding where we applied the Δ-ML Ansatz to model DFT-level enthalpies of atomization for the 134k dataset, and where we found that for the most extreme outlier the baseline model's error reduced systematically with the training set size of the augmenting ML model.[23]

When considering the 10 most extreme outliers of the ML models in Table I, neither order nor identity of the DFT outliers is conserved. Among the top 10 outliers of the 5k model, for example, there is even a saturated molecule from the opposite (blue) end of the error distribution in Fig. 4: tetra-fluoromethane CF_4, with an underestimating deviation −1.42 eV.

TABLE I. 10 most extreme outliers for TDPBE0/def2SVP and Δ-ML models. Largest deviations of predicted lowest singlet-singlet transition energy (E_1) from the corresponding CC2/def2TZVP value. All values in eV.

Molecule	TDPBE0	1k Δ-ML	5k Δ-ML	CC2
Top DFT outliers				
FC1=COC=NC1=O	1.63	1.41	1.40	5.85
CC1=COC=CC1=O	1.64	1.22	1.04	5.69
CC1=CC(=O)C=NO1	1.66	1.08	0.88	5.23
CC1=C(NN=N1)C=O	1.73	1.55	1.51	5.57
O=CC1=CN=CN=C1	1.81	1.50	1.42	5.38
CC1=C(C)CC(=O)C1	1.82	1.62	1.56	6.12
CN1C=C(C=O)C=N1	1.84	1.62	1.35	5.82
CC1=C(C=O)N=NO1	1.92	1.52	1.74	5.82
C#CC1=NC=CN=N1	1.99	1.43	1.76	5.02
O=C1CC=CC1	2.13	2.15	1.95	6.44
Top 1k Δ-ML outliers				
O=N(=O)C1=NC=CO1	1.53	1.33	1.41	5.31
FC1=COC=NC1=O	1.63	1.41	1.40	5.85
C#CC1=NC=CN=N1	1.99	1.43	1.76	5.02
CC(=O)C1=CC=NN1	1.62	1.46	0.91	5.55
O=CC1=CN=CN=C1	1.81	1.50	1.42	5.38
CC1=C(C=O)N=NO1	1.92	1.52	1.74	5.82
CC1=C(NN=N1)C=O	1.73	1.55	1.51	5.57
CC1=C(C)CC(=O)C1	1.82	1.62	1.56	6.12
CN1C=C(C=O)C=N1	1.84	1.62	1.35	5.82
O=C1CC=CC1	2.13	2.15	1.95	6.44
Top 5k Δ-ML outliers				
O=N(=O)C1=NC=CO1	1.53	1.33	1.41	5.31
FC(F)(F)F	−1.20	−1.27	−1.42	13.98
O=CC1=CN=CN=C1	1.81	1.50	1.42	5.38
O=CC1=NC=CC=C1	1.58	1.31	1.46	5.09
CC1=C(NN=N1)C=O	1.73	1.55	1.51	5.57
OC1=NOC(C=O)=C1	1.59	1.32	1.54	5.18
CC1=C(C)CC(=O)C1	1.82	1.62	1.56	6.12
CC1=C(C=O)N=NO1	1.92	1.52	1.74	5.82
C#CC1=NC=CN=N1	1.99	1.43	1.76	5.02
O=C1CC=CC1	2.13	2.15	1.95	6.44

E. ML models of oscillator strengths

We have also investigated the applicability of the Δ-ML Ansatz to model oscillator strengths, f_1 and f_2 for $S_0 \rightarrow S_1$ and $S_0 \rightarrow S_2$ transitions, respectively. While the Δ-ML models of excitation energies can be systematically improved through mere addition of training data, corresponding models for f_1 or f_2 do not become more accurate with increasing training set size. TDCAM-B3LYP has been shown to yield oscillator strengths with minimal deviations with respect to correlation TD methods.[60] For our 22k dataset, TDCAM-B3LYP/def2TZVP yields a MAE of 0.0101 a.u., compared to CC2/def2TZVP. This deviation is reduced to only 0.0100 and 0.0099 a.u. when augmenting the CAM-B3LYP numbers with Δ-ML models trained on 1k and 5k molecules, respectively. Also changing the descriptor from CM to BOB did not improve the state of affairs.

The Δ-ML model approach might fail for several reasons. For one, f_i is a rather complex property which requires knowledge of a certain combination of two wavefunctions,

$$f_i \propto |\langle 0|\hat{\mu}|i\rangle|^2 E_i. \quad (6)$$

This could imply the need for substantially larger training sets in order to obtain satisfying learning curves. Another explanation might be that the training problem is ill posed. In fact, TDDFT often yields a different ordering of states than CC2,

implying that the baseline property corresponds to a different matrix element than the targetline property. This, in turn, will also result in substantially less efficient ML training scenarios. However, this reasoning, while appealing to explain the failure of a Δ^{CC2}_{TDDFT}-ML model, does not satisfyingly explain why also a direct ML model with zero baseline shows insignificant prediction improvement with increasing training set size. Finally, we remark that also previously we have seen significantly less impressive learning rates for other electronic integrals, e.g., the norm of the molecular dipole moment in organic molecules.[31]

IV. CONCLUSIONS

In summary, we have applied the Δ-ML approach, previously introduced to accurately model molecular ground state properties, to the data-driven modeling of electronic excitation energies. We have computed the low-lying valence electronic spectra for a modest chemical universe of 22k organic molecules, made up from up to 8 CONF atoms, at the level of TDDFT (using PBE0 and CAM-B3LYP), and CC2. We have presented numerical evidence that large basis set CC2-level valence excitation energies can be estimated at the speed of small basis set TDPBE0 through statistical inference of the difference, derived from training on a fraction of this database.

Analysis of the data-sets, based on kernel density estimates, suggests small basis set TDPBE0 level of theory to over- and under-estimate the lowest two transition energies for organic molecules with σ- and π-chromophores, respectively. This behavior results in well separated bivariate error distribution. Accounting for these systematic shifts enables further improvement of the Δ-ML models. From a methodological point of view, this procedure allows to readily integrate expert knowledge of error distributions in the ML model, resulting in improved predictions. For an automated estimation of systematic shifts arising from multivariate property distributions, one can adapt clustering protocols based on kernel density estimates. Such clustering has been done previously in the context of analyzing Monte Carlo trajectories,[61] collective variables in molecular dynamics,[62,63] or even to quantify the contribution of a MO to total electronic energy.[64]

The numerical evidence for the modeling of excitation energies suggests that severe flaws in TDDFT based predictions can easily be rectified through statistical learning, irrespective of their origin such as possible incorrect state ordering, basis set incompleteness, inherent limitations of adiabatic TDDFT for states with doubly excited, or CT character.

The poor performance of ML models for predicting oscillator strengths warrants future investigations. The database of excited states properties for 22k organic molecules (see the supplementary material)[65] might also be useful for benchmarking the performance of other approximations and models, as well as to facilitate the identification of potential, hitherto unknown, chromophore-auxochrome relationships. Eventually, our study might aid the computational design of functional molecular components with desirable photochemical properties.

ACKNOWLEDGMENTS

O.A.v.L. acknowledges funding from the Swiss National Science Foundation (No. PP00P2_138932). E.T. acknowledges start-up funds from California State University Long Beach. Some calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under Contract No. DE-AC02-06CH11357.

APPENDIX: DERIVATION OF EQ. (3) IN MATRIX NOTATION

We derive the linear system of equations in Eq. (3) by employing the regularized least squares error measure, Eq. (4). Let us denote the reference property values of training molecules as the column vector $\mathbf{x} = \mathbf{p}^{ref}$. The kernel-ridge-regression Ansatz for the estimated property values of training molecules is $\mathbf{p}^{est} = \mathbf{Kc}$. The L_2-norm of the residual vector, penalized by regularization of fit coefficients, is the Lagrangian

$$\begin{aligned}\mathcal{L} &= \|\mathbf{p}^{ref} - \mathbf{p}^{est}\|_2^2 + \lambda \mathbf{c}^T\mathbf{Kc} \\ &= (\mathbf{x} - \mathbf{Kc})^T(\mathbf{x} - \mathbf{Kc}) + \lambda \mathbf{c}^T\mathbf{Kc} \\ &= \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{Kc} - (\mathbf{Kc})^T\mathbf{x} + (\mathbf{Kc})^T\mathbf{Kc} + \lambda \mathbf{c}^T\mathbf{Kc}, \quad \text{(A1)}\end{aligned}$$

where $(\cdot)^T$ denotes transpose operation. To minimize the Lagrangian, we equate its derivative with respect to the regression coefficients vector, $\mathbf{c}$, to zero,

$$\begin{aligned}\frac{d}{d\mathbf{c}}\mathcal{L} = &-\mathbf{x}^T\mathbf{K} - \mathbf{Kx} + \mathbf{KKc} + \mathbf{c}^T\mathbf{KK} \\ &+ \lambda\mathbf{Kc} + \lambda\mathbf{c}^T\mathbf{K} = 0. \quad \text{(A2)}\end{aligned}$$

Here, we have used the fact that the kernel matrix $\mathbf{K}$ is symmetric, i.e., $\mathbf{K}^T = \mathbf{K}$ along with the matrix calculus identity, $(d/d\mathbf{c})\,\mathbf{c}^T = \mathbf{I}$, where $\mathbf{c}$ is a column vector and $\mathbf{c}^T$ is a row vector. Grouping by row and column vectors yields

$$(\mathbf{KKc} + \lambda\mathbf{Kc} - \mathbf{Kx}) + (\mathbf{KKc} + \lambda\mathbf{Kc} - \mathbf{Kx})^T = 0, \quad \text{(A3)}$$

which is satisfied, iff

$$(\mathbf{KKc} + \lambda\mathbf{Kc} - \mathbf{Kx}) = 0. \quad \text{(A4)}$$

Multiplication of the above equation with $\mathbf{K}^{-1}$ from the left, and rearranging results in Eq. (3).

[1]C. Kuhn and D. N. Beratan, J. Phys. Chem. **100**, 10595 (1996).
[2]O. A. von Lilienfeld, in *Many-Electron Approaches in Physics, Chemistry and Mathematics*, edited by V. Bach and L. D. Site (Springer, 2014), pp. 169–189.
[3]M. Grätzel, Nature **414**, 338 (2001).
[4]M. Gross, D. C. Müller, H.-G. Nothofer, U. Scherf, D. Neher, C. Bräuchle, and K. Meerholz, Nature **405**, 661 (2000).
[5]T. Yogo, Y. Urano, Y. Ishitsuka, F. Maniwa, and T. Nagano, J. Am. Chem. Soc. **127**, 12162 (2005).
[6]E. M. Tan, M. Hilbers, and W. J. Buma, J. Phys. Chem. Lett. **5**, 2464 (2014).
[7]M. Pastore, E. Mosconi, F. De Angelis, and M. Gratzel, J. Phys. Chem. C **114**, 7205 (2010).
[8]J. Han, X. Chen, L. Shen, Y. Chen, W. Fang, and H. Wang, Chem. - Eur. J. **17**, 13971 (2011).
[9]M. M. Wolf, C. Schumann, R. Gross, T. Domratcheva, and R. Diller, J. Phys. Chem. B **112**, 13424 (2008).

[10]E. Tapavicza, A. M. Meyer, and F. Furche, Phys. Chem. Chem. Phys. **13**, 20986 (2011).
[11]O. Christiansen, H. Koch, and P. Jørgensen, Chem. Phys. Lett. **243**, 409 (1995).
[12]E. Runge and E. K. Gross, Phys. Rev. Lett. **52**, 997 (1984).
[13]M. E. Casida, *Recent Advances in Density Functional Methods* (World Scientific, 1995), Vol. 1, p. 155.
[14]A. Dreuw, J. L. Weisman, and M. Head-Gordon, J. Chem. Phys. **119**, 2943 (2003).
[15]N. T. Maitra, F. Zhang, R. J. Cave, and K. Burke, J. Chem. Phys. **120**, 5932 (2004).
[16]L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, J. Chem. Inf. Model. **52**, 2864 (2012).
[17]O. A. von Lilienfeld, Int. J. Quantum Chem. **113**, 1676 (2013).
[18]V. Marx, Nature **498**, 255 (2013).
[19]C. A. Mattmann, Nature **493**, 473 (2013).
[20]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[21]G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, New J. Phys. **15**, 095003 (2013).
[22]K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. **9**, 3404 (2013).
[23]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 2087 (2015).
[24]B. G. Janesko and D. J. Yaron, J. Chem. Phys. **121**, 5635 (2004).
[25]V. Ediz, A. J. Monda, R. P. Brown, and D. J. Yaron, J. Chem. Theory Comput. **5**, 3175 (2009).
[26]Z. D. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Müller, and G. Henkelman, J. Chem. Phys. **136**, 174101 (2012).
[27]J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Phys. Rev. Lett. **108**, 253002 (2012).
[28]K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).
[29]A. Lopez-Bezanilla and O. A. von Lilienfeld, Phys. Rev. B **89**, 235411 (2014).
[30]L.-F. Arsenault, A. Lopez-Bezanilla, O. A. von Lilienfeld, and A. J. Millis, Phys. Rev. B **90**, 155136 (2014).
[31]R. Ramakrishnan and O. A. von Lilienfeld, CHIMIA **69**, 182 (2015).
[32]R. Send, M. Kuhn, and F. Furche, J. Chem. Theory Comput. **7**, 2376 (2011).
[33]D. Kannar and P. G. Szalay, J. Chem. Theory Comput. **10**, 3757 (2014).
[34]P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[35]W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[36]J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[37]J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[38]M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **110**, 5029 (1999).
[39]C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).
[40]F. Furche and R. Ahlrichs, J. Chem. Phys. **117**, 7433 (2002).
[41]T. Yanai, D. P. Tew, and N. C. Handy, Chem. Phys. Lett. **393**, 51 (2004).
[42]J. A. Nelder and R. Mead, Comput. J. **7**, 308 (1965).
[43]K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).
[44]O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, Int. J. Quantum Chem. **115**, 1084 (2015).
[45]R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, 1989), pp. 112–113.
[46]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Sci. Data **1**, 140022 (2014).
[47]TURBOMOLE V6.2c 2011, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007, available from http://www.turbomole.com.
[48]M. Häser and R. Ahlrichs, J. Comput. Chem. **10**, 104 (1989).
[49]F. Weigend and R. Ahlrichs, Phys. Chem. Phys. Chem. **7**, 3297 (2005).
[50]C. Hättig and F. Weigend, J. Chem. Phys. **113**, 5154 (2000).
[51]M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J.-Y. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. Montgomery, A. John, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, GAUSSIAN 09, Revision D.01, Gaussian, Inc., Wallingford, CT, 2009.
[52]B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (CRC Press, 1986), Vol. 26.
[53]Z. Botev, J. Grotowski, and D. P. Kroese, Ann. Stat. **38**, 2916 (2010).
[54]C.-Y. Tseng, F. Taufany, S. Nachimuthu, J.-C. Jiang, and D.-J. Liaw, Org. Electron. **15**, 1205 (2014).
[55]R. Memming, *Semiconductor Electrochemistry* (John Wiley & Sons, 2008), p. 342.
[56]GNUPLOT 4.6, an interactive plotting program, 2013, available from http://sourceforge.net/projects/gnuplot.
[57]D. Jacquemin, V. Wathelet, E. A. Perpete, and C. Adamo, J. Chem. Theory Comput. **5**, 2420 (2009).
[58]M. J. G. Peach, P. Benfield, T. Helgaker, and D. J. Tozer, J. Chem. Phys. **128**, 044118 (2008).
[59]R. D. Johnson III, *NIST Computational Chemistry Comparison and Benchmark DataBase* (National Institute of Standards and Technology, 2013), http://cccbdb.nist.gov.
[60]M. Caricato, G. W. Trucks, M. J. Frisch, and K. B. Wiberg, J. Chem. Theory Comput. **7**, 456 (2010).
[61]I. P. Christov, J. Chem. Phys. **135**, 044120 (2011).
[62]L. Wang, C. C. Martens, and Y. Zheng, J. Chem. Phys. **137**, 034113 (2012).
[63]P. Gasparotto and M. Ceriotti, J. Chem. Phys. **141**, 174110 (2014).
[64]A. Melnichuk and R. J. Bartlett, J. Chem. Phys. **137**, 214103 (2012).
[65]See supplementary material at http://dx.doi.org/10.1063/1.4928757 for supplementary information indices of the 22k GDB-8 molecules, to retrieve their geometries from the 134k GDB-9 dataset,[46] along with TDDFT, and CC2 excitation energies are collected in `gdb8_22k_elec_spec.txt`.