Data-driven approach for benchmarking DFTB-approximate excited state methods

Andrés I. Bertoni and Cristián G. Sánchez*

Instituto Interdisciplinario de Ciencias Básicas (ICB-CONICET), Universidad Nacional de Cuyo,
Padre Jorge Contreras 1300, Mendoza 5502, Argentina

(*Corresponding author. E-mail: csanchez@mendoza-conicet.gob.ar)

In this work we propose a chemically-informed data-driven approach to benchmark the approximate density-functional tight-binding (DFTB) excited state (ES) methods that are currently available within the DFTB+ suite. By taking advantage of the large volume of low-detail ES-data in the machine learning (ML) dataset, QM8, we were able to extract valuable insights regarding the limitations of the benchmarked methods in terms of the approximations made to the parent formalism, density-functional theory (DFT), while providing recommendations on how to overcome them. For this benchmark, we compared the first singlet-singlet vertical excitation energies ($E_1$) predicted by the DFTB-approximate methods with predictions of less approximate methods from the reference ML-dataset. For the nearly 21,800 organic molecules in the GDB-8 chemical space, we were able to identify clear trends in the $E_1$ prediction error distributions, with respect to second-order approximate coupled cluster (CC2), showing a strong dependence on chemical identity.

I. INTRODUCTION

Machine learning (ML) techniques are becoming ubiquitous in all areas of scientific research, and quantum chemistry is no exception. There are efforts in this area of research that are directed toward the automated prediction of molecular electronic properties from structure14. In particular, the prediction of excited state (ES) properties, such as photo-absorption excitation energies, could enable the data-driven discovery of useful chromophores, such as organic dyes with potential uses in medical imaging and optoelectronic devices. Advances in ML chemical modelling are accompanied by the creation of datasets of increasing size and quality, containing chemical information from experiments and theoretical calculations56. These large collections of curated data are generally assembled to train deep learning models that are capable of accurately predicting the electronic properties of other molecules within the same chemical subspace as the training set. Although these datasets currently provide a limited level of detail, they stand out because they span a large volume of the chemical space.

In this work we recognise additional value in ML-datasets and propose a chemically-informed data-driven approach to benchmark the approximate density-functional tight-binding (DFTB) excited state methods that are currently available in the DFTB+ suite. For this benchmark, we compared the first singlet-singlet vertical excitation energies ($E_1$) predicted by the approximate methods with less approximate methods from the reference ML-dataset. For the nearly 21,800 organic molecules in the GDB-8 chemical space, we were able to identify clear trends in the $E_1$ prediction error distributions, with respect to second-order approximate coupled cluster (CC2), with a strong dependence on chemical identity. In other words, we took advantage of the large volume of low-detail ES-data in the QM8 ML-dataset, and combined it with molecular descriptors based on π-conjugation and chemical identity, in order to extract valuable insights regarding the limitations of the benchmarked methods in terms of the approximations made to the parent formalism, density-functional theory (DFT). As this benchmark was performed for a large number of molecules with a diverse representation of chemical groups, it serves as a comprehensive complement to other reported benchmarks concerning specific implementations in DFTB+711.

These acquired insights allowed us to provide recommendations to the DFTB+ user and developer communities on the limitations that may be encountered and how to overcome them. As a result of this benchmark, we provide (i) a “rule of thumb” for computing $E_1$ with the ES-methods and corrections currently implemented in DFTB+, (ii) a proof-of-concept12 minimally polarized version of a popular SK-parameter set, (iii) recommendations for next steps in DFTB+ development, and (iv) an extension of the QM8 ML-dataset with DFTB-results.

II. OVERVIEW OF METHODS

DFTB+ is a popular and open source scientific software package that offers a fast and scalable implementation of the density functional based tight binding (DFTB) method13. Among the many implemented features, which are being actively developed14 by its vast community of users, there is available a set of methods for simulating the photo-absorption spectroscopy of molecular systems; namely, the time-dependent DFTB method (TD-DFTB) and the DFTB-approximate particle-particle random phase approximation (ppRPA)15. In DFTB+, TD-DFTB exists as two distinct but equivalent implementations: one is based on Casida’s linear response (LR) approach16 and the other is a real-time implementation based on the semi-classical Ehrenfest method for nuclear-electronic dynamics17.

Being a semi-empirical method, DFTB is a considerably more approximate method than its parent ab ini-
tio method, DFT\textsuperscript{18,19}. However, these approximations are precisely the reason for its superior computational performance. The relatively lower accuracy of DFTB is widely compensated by a substantial drop in computational cost, which extends the feasibility upper limit in size, timescale and complexity of the systems to be studied. As a point of comparison, TD-DFTB approximations result in a computational scaling reduced by three orders of magnitude with respect to TD-DFT in its default formulation (from $N^6$ to $N^3$, where $N$ is the number of atoms in the system)\textsuperscript{14}. In other words, the approximations of DFTB bring quantum mechanics to systems that otherwise would be treated with classical force fields. As a result, DFTB is being widely employed in pure and hybrid QM-schemes for the simulation of the spectroscopic properties of nanostructures\textsuperscript{20}, photoactive materials\textsuperscript{21} and complex biological systems\textsuperscript{22} of many thousands of atoms.

Our work is aimed to both experts and beginners in the DFTB\textsuperscript{+} community that share an interest in the computation of excitation energies within the DFTB framework. Thus, we have decided to include into the electronic supplementary material (ESI) brief overviews of the methods revised in this work, which will aid the reader in the discussion of the results. However, we strongly recommend the readers to refer to the original publications for the respective implementations of DFTB\textsuperscript{14}, TD-DFTB (DFTB-Casida)\textsuperscript{7} and pp-DFTB (DFTB-ppRPA)\textsuperscript{15}. We also want to recommend a pedagogical introduction to those who are completely new to DFTB\textsuperscript{23}.

III. COMPUTATIONAL DETAILS

A. Reference dataset

The QM8 reference dataset\textsuperscript{5} was created for the development of machine learning models. Although this ML-dataset does not provide high-detail data for the low-lying electronic transitions (e.g., corresponding transition electric dipole moments or excited state symmetries), it instead provides a large amount of low-detail ES quantities, calculated with CC2 and linear-response time-dependent DFT (TD-DFT): it comprises the first two singlet-singlet vertical excitation energies ($E_1$ and $E_2$) and oscillator strengths ($f_1$ and $f_2$) for the 21786 organic molecules, with up to 8 CONF atoms (up to 26 CHONF atoms), which make the GDB-8 subset of the GDB-9 chemical space\textsuperscript{24}.

As Ramakrishnan \textit{et al.}, we have decided to take CC2 with the polarized triple-zeta valence second-generation basis set (def2TZVP) as the reference method. The authors based their decision on the fact that CC2/def2TZVP was shown to predict valence excitation energies in very close proximity to experimental data\textsuperscript{25} and to a higher-order computational prediction\textsuperscript{26}, with mean absolute errors of 0.12 eV and 0.10 eV, respectively.

B. Computing the first singlet vertical excitation energy

Prior to the computation of excitation energies, the molecular geometries were re-optimised with SCC-DFTB and a chosen parameter set from www.dftb.org. Geometries were relaxed via the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm, in order to achieve maximum force components below $10^{-5} E_h/a_0$. For each DFTB-optimised geometry within the addressed chemical sub-space, we computed $E_1$ with the two approximate methods available for this task in the last version of DFTB\textsuperscript{+}: DFTB-Casida and DFTB-ppRPA. Currently, these DFTB-approximate excited state methods are only implemented at the SCC-DFTB level of theory, for molecular systems. Although we were only interested in $E_1$, the number of computed excitation energies was set up to 20, when possible, to aid the convergence of the eigenvalue solvers. Specifically for DFTB-ppRPA calculations, we also set the maximum number of virtual states to 1000 and provided the Hubbard-like parameters from the Appendix K of the DFTB\textsuperscript{+} manual\textsuperscript{27}. For both approaches, we employed the general purpose parameter set 3OB(-3-1)\textsuperscript{28}, which has parameters for a wide variety of elements present in biological and organic systems. For 3OB, the Hamiltonian and overlap matrix elements were pre-computed using DFT with the Perdew-Burke-Ernzerhof (PBE) functional.

In this work, we decided to compare the error of each DFTB-approximate method in the prediction of $E_1$, with respect to CC2/def2TZVP ($\Delta_{CC2}E_1$), with results from the parent method, DFT. With this analysis, our aim is to detect which of the limitations in the prediction of $E_1$ are being shared between the related methods and which are unique to the standard formulation of DFTB. As the parent method for this comparison we selected TD-DFT with the hybrid HF-XC variant of the local PBE functional (PBE0) and the polarized split-valence second-generation basis set (def2SVP). This choice corresponds to the method within QM8 that is most related to DFTB: a DFT method using the most local XC functional and the smallest basis set. Despite the hybrid functional PBE0 is the closest one in the ML-dataset to PBE (i.e., the GGA functional employed to calculate the electronic parameters for DFTB), systematic differences between local and hybrid functionals should be expected regarding electronic excitations within TD-DFT: this fact will contribute to the overall differences in performance between TD-DFT and TD-DFTB, particularly for $\pi$-type excitations. For more information on this topic, we would like to recommend the work of Maier and Kaupp \textit{et al.}\textsuperscript{29}, which includes a systematic comparison between (semi-)local and global hybrid functionals, with respect to the prediction of TD-DFT electronic excitation energies, for a small set of organic molecules. Values of $E_1$ from the reference method, CC2/def2TZVP, and the time-dependent extension of the parent method, TD-DFT:PBE0/def2SVP, were directly taken from QM8; these values were originally computed from geometries...
relaxed via DFT with the Becke 3-parameter Lee-Yang-Parr hybrid XC-functional (DFT/B3LYP) and the polarized Pople-style 6-31G(2df,p) basis set.

To address the limitations encountered, we also included extensions to the standard SCC-DFTB method in the calculation of $E_1$, whose implementations are available in DFTB+. Specifically, we included long-range corrections (LC), on-site corrections (OC), and a proof-of-concept custom extension of the SK-parameter set to include minimal polarization on H atoms only (3OB(H*)); see the theoretical and technical details of these extensions.

C. Using molecular descriptors for data analysis

The molecules in the GDB-8 chemical subspace were generated following synthetic feasibility rules, but the combinatorial nature of the structure generation algorithm may have resulted in structures that may be considered unusual. However, we classified the molecules into chemical families based on their functional group with the highest level of conjugation, because we expect the functional group with the largest $\pi$-conjugation in a molecule to be the main contributor to the frontier molecular orbitals, thus characterising the lowest electronic excitation. Under this premise, the rest of the molecule would be acting as a local chemical environment that modulates the excitation energy of the main functional group, regardless of how rare the chemical arrangement is. This way, the conclusions to be drawn may also apply to molecules larger than those in the dataset (i.e., of more than 26 atoms), if the additional atoms do not extend the conjugated systems; however, if that is the case, then these larger molecules would have their $E_1$ closer to the visible light region, but would still be members of the $\pi$-conjugated family, and thus would be expected to suffer from the same limitations as the ones we will discuss in this work.

For each molecule in the ML-dataset, we used the Python library rdkit\textsuperscript{30} to identify its level of $\pi$-conjugation and the presence of certain chemical groups, through the analysis of their SMILES strings\textsuperscript{31}. In that manner, molecules were assigned to one of three main families of chemical compounds: (i) $\pi$-conjugated unsaturated molecules, (ii) non-conjugated unsaturated molecules (i.e, with isolated double or triple bonds) and (iii) saturated molecules (i.e., with only single bonds). Within the family of $\pi$-conjugated unsaturated molecules, we made a distinction between molecules with conjugated groups of 3 atoms and molecules with larger conjugated groups (i.e., with more than 3 atoms). Via rdkit tools for handling the SMILES arbitrary target specification (SMARTS) language, we assigned a more detailed chemical-identity to the molecules of each principal chemical family. For this task, extra chemical descriptors were employed in a hierarchical fashion (i.e., based on the order of assignation of labels); in other words, a molecule assigned to a particular chemical sub-family was removed from the pool of molecules to be identified as part of any other sub-group within the same main family. The SMARTS fragments employed are available in Table I of the ESI.

IV. RESULTS AND DISCUSSION

A. Initial diagnostics

The prediction accuracy of $E_1$ was compared between the approximate DFT methods and TD-DFT-PBE0/def2SVP. Specifically, we started with the comparison of prediction errors $\Delta_{CC2}E_1$ for DFTB-Casida:3OB and DFTB-ppRPA:3OB, which are shown as datapoints in panels A and B of Fig. 1, respectively. The scattered points correspond to each of the nearly 21,800 molecules in the GDB-8 chemical subspace, and are coloured to reflect their membership to one of the main chemical families we defined before. Datapoints were located on the 2D planes according to the prediction error $\Delta_{CC2}E_1$, as computed with the benchmarked method (horizontal axis) and TD-DFT-PBE0/def2SVP (vertical axis). Accordingly, molecules in the lower-left quadrant (blue background) have $E_1$ values that are underestimated by both methods, while those in the upper-right quadrant (red background) have their $E_1$ overestimated by both methods. In addition, each group of coloured points was also projected onto each axis, as histograms, to show the $E_1$ error distribution for each chemical family, for each of the compared methods.

First of all, Fig. 1 shows that the error of $E_1$ predictions strongly depends on the molecules’ main chemical identity. Based on error distributions of $E_1$ for TD-DFT, Ramakrishnan et al. first noticed that predictions could be divided into two distinctive sets of molecules: saturated and unsaturated; the authors observed that, overall, $E_1$ was underestimated for unsaturated molecules and overestimated for saturated molecules. In panel A of Fig. 1, we found broadly the same saturated-unsaturated pattern for DFTB-Casida:3OB, as the vast majority of points are either in the underestimation or overestimation quadrants. This reproduction of general trends between TD-DFTB and TD-DFT is not entirely surprising, as the former method is an approximation to the latter.

On the one hand, molecules with extended $\pi$-conjugation are expected to exhibit low-lying excitations of the $\pi$-type ($\pi \rightarrow \pi^*$ or $n \rightarrow \pi^*$). Therefore, DFTB-Casida:3OB systematic underestimation of $E_1$ for unsaturated molecules should be arising from the DFT-inherited self-interaction error (SIE). This well-known error involves individual electrons interacting with their own density\textsuperscript{32}. The SIE is related to the local approximation of the XC contribution to the charge interaction energy, and arises from the Coulomb and exchange terms incorrectly not cancelling each other exactly at large distances\textsuperscript{33}. As the solutions of SCC-DFTB are found


FIG. 1. Comparison of prediction errors $\Delta_{CC2}E_1$ for TD-DFT:PBE0/def2SVP and four different DFTB-approximate approaches: (A) Casida:3OB, (B) ppRPA:3OB, (C) Casida+LC:OB2 and (D) Casida:3OB(H*). The scattered datapoints correspond to each of the nearly 21,800 molecules in the GDB-8 chemical subspace. The datapoints and the projected histograms were coloured according to the main chemical identity of the compounds: green for saturated molecules, orange for non-conjugated molecules with an isolated double or triple bond, and blue and purple for $\pi$-conjugated molecules with 3 or more conjugated atoms, respectively. See the first column of Tables IV & V, in the ESI, for the number of occurrences per chemical family and subgroup.

self-consistently$^{34}$, the SIE leads to an incorrect charge distribution in the ground state, with the consequent artificial stabilisation of the delocalised solutions$^{35-37}$. This results in the underestimation of HOMO-LUMO gaps$^{38}$ and excitation energies or, equivalently, in the overestimation of dipole polarizabilities$^{39}$. This is why TD-DFT with local or semi-local functionals is not recommended for low-lying electronic transitions involving $\pi$-conjugated states$^{39-41}$, nor it is recommended for charge-transfer (CT) excitations$^{29,42,43}$. The SIE is particularly problematic for CT-excitations$^{44}$, most notably for systems with excess charge (e.g., ions$^{45}$ or radical ions$^{46,47}$) and for transitions involving occupied and virtual molecular orbitals (MOs) with minimal spatial overlap$^{48}$.
On the other hand, saturated molecules are expected to exhibit mainly \( \sigma \)-type excitations (\( \sigma \rightarrow \sigma^* \) or \( n \rightarrow \sigma^* \)). In some cases, Rydberg excitations may appear as the lowest-lying electronic excitations; however, without the use of a very diffuse basis set, Rydberg states should be outside the scope of both DFT and DFTB methods. The reasons behind DFTB-Casida:3OB overestimation of excitation energies in saturated molecules are not nearly as discussed as the self-interaction error of DFT. However, as we will discuss with more detail in a next subsection, this limitation appears to be related to the low quality description of the MOs, due to the use of a minimal basis set, and to the approximations to the charge density in the excited state method. One should be aware that errors related to the non-completeness of the basis set are not exclusive to DFTB, since employing a finite basis set is an approximation present in most numerical methods for the electronic structure of molecules, which of course includes DFT.

Although TD-DFTB appears to be inheriting limitations from its parent formalism, TD-DFT, there are obvious differences. Comparatively, TD-DFT:PBE0/def2SVP can be considered to be more precise for all chemical groups, since the vast majority of its predictions for \( E_1 \) have an error within \( \pm 1 \) eV. In contrast, the \( E_1 \) error dispersion for DFTB-Casida:3OB predictions is highly dependent on chemical identity: it is considerably large for saturated molecules, with \( +1 \) eV \( < \Delta_{CC^2}E_1 < +6 \) eV, while for unsaturated molecules it is more moderate, \( \frac{\Delta_{CC^2}E_1}{-2} \) eV and \( +1 \) eV. Overall, the TD-DFT:pwB97X-D/def2SVP method is expected to be more precise—and also computationally more expensive—because it is less approximated than TD-DFTB: (i) it is an \textit{ab initio} method (i.e., parameter-free), where the expression for the total energy is as exact as the approximate XC-functional being chosen; (ii) it computes the actual XC-functional, which is not necessarily purely local, e.g. PBE0 is a hybrid XC-functional with 25\% exact-HF exchange, and therefore performs significantly better than (semi-)local functionals for valence and CT excitations; (iii) it employs the def2SVP basis set so that, although it is considered to be small for DFT, it is certainly not minimal and even includes some polarization.

Within the unsaturated molecules group, it can be seen that, for molecules characterised by an isolated double or triple bond, DFTB-Casida:3OB populates more than TD-DFT the line at \( \Delta_{CC^2}E_1 = 0 \), which corresponds to the predictions of \( E_1 \) exactly matching those of CC2. In other words, although the multiple approximations of DFTB to DFT make \( E_1 \) predictions less accurate, they also led to an increase in cases where errors might be being compensated for to produce the same predictions as methods with a higher level of theory.

While TD-DFTB and TD-DFT are strongly related in formalism, DFTB-ppRPA has a very different conceptual origin. DFTB-ppRPA is expected to perform better than TD-DFT(B) for \( \pi \)-type excitations, and specifically for excitations with charge transfer (CT) nature, because it manages to reproduce the correct asymptotic \( -1/R \) trend for the electron-hole interaction in the long range, even within the monopole approximation. Because of this, in panel B of Fig. 1 it can be seen that DFTB-ppRPA:3OB is not underestimating \( E_1 \) for all of the unsaturated molecules, as Casida-DFTB:3OB and TD-DFT:pwB97X-D/def2SVP do. In fact, DFTB-ppRPA:3OB overestimates most of the non-conjugated unsaturated molecules and, for \( \pi \)-conjugated molecules, this method results in a \( E_1 \) error distribution that is centered at \( \Delta_{CC^2}E_1 = 0 \), although with considerable dispersion. Regarding the saturated molecules, we observe an \( E_1 \) error distribution that is very similar to that of DFTB-Casida:3OB, suggesting that this limitation originates already in the base formalism, SCC-DFTB.

In the ESI we extended the data analysis of this initial diagnostic to include another layer of chemical detail for the non-conjugated unsaturated molecules; there, we show that DFTB-ppRPA performs best for non-conjugated alkynes, while DFTB-Casida would be the recommended method for the other non-conjugated molecules with carbonyl, cyano and alkyne groups.

Based on the discussion so far, neither of the two approaches based on the standard SCC-DFTB formalism would be acceptable for predicting \( E_1 \) for saturated molecules, and therefore we should avoid them. With respect to the unsaturated molecules, DFTB-ppRPA generated decent \( E_1 \) predictions for alkenes and molecules with \( \pi \)-conjugation, while DFTB-Casida:3OB did so for non-conjugated ones. However, there are extensions to the standard SCC-DFTB formalism, which are already implemented in DFTB+, that are worth testing if we want to improve our initial estimates of \( E_1 \). This is what we did in hope of resolving the limitations behind the broad overestimation of excitation energies in saturated molecules and behind the underestimation of \( \pi \)-type excitations, particularly for \( \pi \)-conjugated molecules. We decided to focus on improving DFTB-Casida from now on, as it is more general than the current implementation of DFTB-ppRPA, which necessarily involves the highest occupied molecular orbital (HOMO) in the prediction of electronic excitations.

B. Dealing with the limitations

Underestimation of \( E_1 \) for \( \pi \)-conjugated molecules

This work is not the first reported encounter of the SIE in DFTB. This DFT-inherited error has already been detected before in the standard formulation of SCC-DFTB\(^{34,49}\) and was traced back to the approximations made to the charge-fluctuation interaction term, where the XC-component of the charge-charge interaction term is reduced to a single number per element, i.e. the on-site Hubbard-like parameters \( U_A \). The highly local and
variational nature of the SCC term results in interactions not converging at large distances, in stark contrast to HF or non-SCC DFTB. Because of the SIE, the standard formulation of SCC-DFTB ends up underestimating the energy of delocalised solutions, almost as much as any other GGA functional for DFT (e.g. PBE)\textsuperscript{34}. As attempts to overcome this limitation, diverse extensions to the standard SCC-DFTB were proposed, e.g.: LDA+U-DFTB\textsuperscript{49}, coarse-grained DFTB\textsuperscript{47} and configuration interaction (CI) within SCC-DFTB\textsuperscript{50}. Arguably, the most popular and widely accepted strategy to correct the SIE is to include long-range corrections (LC) into SCC-DFTB.

For the benchmark of this correction, we computed \( E_1 \) with DFTB-Casida+LC and the OB2 parameter set, which is currently the only SK-set available in DFTB\textsuperscript{+} that is compatible with DFTB+LC. In panel C of Fig. 1 we can observe that DFTB+LC-Casida:OB2 increases the value of all the initial estimates of \( E_1 \), resulting in the shift of all corresponding error distributions to more positive values. This results in an improvement for \( \pi \)-conjugated molecules as their new \( E_1 \) error distribution is now centered closer to \( \Delta_{C2} E_1 = 0 \) and mostly contained within \( \pm 1eV \). It is interesting to note that including long-range corrections into DFTB-Casida (DFTB+LC-Casida:OB2) calculations achieves similar accuracy to DFTB-ppRPA:3OB, but with less dispersion, as can be seen in Fig. S4 of the ESI. This improved accuracy of DFTB+LC computations arise from the newly gained ability to reproduce the correct trend for interactions over long distances, as DFTB-ppRPA does. This correction being included already from the parameterisation process could be the reason why DFTB+LC-Casida:OB2 is more precise than DFTB-ppRPA:3OB, despite their very similar accuracies. Given this that correction is included via the XC functional, it would be expected for the pre-computed integrals and repulsion splines of DFTB+LC to encode additional XC information. In contrast, DFTB-ppRPA:3OB is based on parameters generated with the non-corrected PBE functional, which is short-ranged and purely local.

The LC shifted the \( E_1 \) predictions to higher energies for all the molecules in the dataset. Although it resulted in a marked improvement for the initially underestimated \( E_1 \) predictions (i.e., for the \( \pi \)-conjugated molecules), it also worsened the estimates for molecules that were already overestimated. Therefore, as can be seen in the projected histograms of Fig. 1 C, this corrected method cannot be recommended for non-conjugated unsaturated or saturated molecules.

In the ESI, we provide an extension of the data analysis with an extra layer of chemical detail; we make a distinction between unsaturated molecules with \( \pi \)-conjugated systems involving 3 atoms and molecules with more extended \( \pi \)-systems. We found that DFTB+LC-Casida:OB2 performs particularly well for the subset of unsaturated molecules with \( \pi \)-conjugated systems involving only 3 atoms, of which we can highlight the ester, carboxylic acid and amide functional groups.

Based on the results of this benchmark, we recommend using DFTB+LC-Casida:OB2 for \( \pi \)-conjugated molecules. This recommendation should become more relevant when studying electronic excitations between different molecular fragments (e.g. solvated chromophores\textsuperscript{51}), in conducting systems or molecules with very large \( \pi \)-conjugated systems (e.g. polyacenes\textsuperscript{11}), and in any other study case where the SIE is expected to be significant.

### Overestimation of \( E_1 \) for saturated molecules

We could identify two main sources for the error that is causing the overestimation of excitation energies in saturated molecules. The first source of error is truncating the multipole expansion of the charge density for the computation of the coupling matrix elements in Casida TD-DFTB. The second source of error is employing a minimal basis set, which results in a low quality description of the GS-MOs. At first glance, what these two approximations have in common is contributing to an over-simplification in the description of the ground state density, either in the ground or excited state formulations.

In the ESI, we listed the three approximations that TD-DFTB makes to LR-TD-DFT. The first two of these approximations are required to translate the Casida’s eigenvalue problem to the DFTB framework, one of which is the Mulliken monopole approximation to the transition charge density. The third approximation of TD-DFTB is introduced to minimise the effort in the computation of the coupling matrix elements, by neglecting the charge fluctuations contributions. The latter approximation was considered to be safe by Niehaus et al.\textsuperscript{7}, having found that the inclusion of \( \Delta_{Q_A} \)-dependent Hubbard-like \( U_A \) parameters had a negligible effect on the corrected excitation energies. However, the same cannot be said for the Mulliken monopole approximation, since it becomes troublesome when dealing with highly localised excitations (i.e., with near full orbital overlap), for which it predicts heavily underestimated transition charges. Because same-center atomic orbitals do not contribute to the atomic transition charges within the monopole approximation, in the cases of full orbital overlap these charges vanish and, consequently, so does the coupling matrix; therefore, in those cases TD-DFTB ends up predicting excitation energies that are exactly equal to the KS orbital energy difference.

All things considered, to improve the prediction of excitation energies for saturated molecules, we could attempt (i) to correct the monopole approximation in the ES-method or (ii) improve the ground state estimation of the KS energy gap. The second strategy takes into account systems for which \( E_1 \) can naturally be expected...
to be very close to the corresponding orbital energy difference.

On the one hand, the first strategy can be addressed with the inclusion of on-site corrections, which were already implemented in DFTB+ by Domínguez et al. On the other hand, to improve the KS energy gap, efforts should be directed at refining the energy estimates for the ground state MOs. This can be done by increasing the size of the basis set, which we essentially consider as a stabilisation of the virtual MOs.

Because the tight-binding approach assumes tightly bound electrons, it is justified to employ a minimal basis set to expand the molecular Hamiltonian eigenstates. This approximation is not often a problem for DFTB with a decent parameterization, as it mostly does very well in terms of accuracy and performance for most of its applications, which frequently concern ground state properties. However, when dealing with the energetics of reactions or excitations, the limitations imposed by the constrained description of the orbital solutions becomes more noticeable. For example, Niehaus\textsuperscript{7} detected that, because of the minimal basis set description, KS energy differences were larger for DFTB compared to DFT. Furthermore, Gauss et al.\textsuperscript{28} recognised that the minimal basis set represented a crucial limit to DFTB’s accuracy regarding the prediction of reaction energies.

We believe that the incompleteness of the basis set predominantly impacts on the overestimation of the virtual state energies, as the unoccupied MOs would not be sufficiently well described. Because of the Rayleigh theorem for eigenvalues, adding more basis functions necessarily lowers the eigenvalues of the molecular Hamiltonian and, thanks to the variational principle, expanding the basis set is not expected to significantly alter the energies of the occupied MOs. However, increasing the number of basis functions has the potential to significantly stabilise the virtual MOs, as they are always fully determined in shape and energy by the requirement of orthogonality with respect to the variationally-determined occupied MOs\textsuperscript{52}, which are better described with larger basis sets.

The easiest way we thought of extending the minimal basis set, only for performing excited state computations with fixed nuclei, was to add pseudo polarization orbitals only to H atoms. The strategy we provide (see the ESI) for minimally polarizing the minimal basis set (i.e., via the custom SK-parameter set, 3OB(H*)) may not be sophisticated enough, but we believe that it serves as a proof-of-concept approach to argue in favour of employing polarized basis sets when predicting σ-type excitation energies for saturated systems.

We benchmarked these two strategies by calculating \( E_1 \) with the on-site corrected DFTB-Casida+OC:3OB and the minimally polarized DFTB-Casida:3OB(H*). The complete set of results for the on-site corrected ES-method is being displayed in Fig. S6, in the ESI. Results showing the impact on \( E_1 \) predictions of partially polarizing the minimal basis set, for all the molecules in the dataset, are given in panel D of Fig. 1 and in Fig. 2.

The extension of the minimal basis set, to include polarization for H atoms only, would be partially correcting the electron density oversimplification error. As a consequence, DFTB-Casida:3OB(H*) seems to have considerably improved the predictions of \( E_1 \) for the entirety of the saturated molecules, as the minimal polarization of the basis set resulted in a significant contraction of the \( E_1 \) error distribution (see Fig. 2). Consistently, for the unsaturated molecules, π-conjugated or not, polarizing the H atoms had an almost negligible effect on their \( E_1 \) predictions.

Since these two strategies were originally motivated on improving \( E_1 \) predictions for the saturated molecules, in the ESI we decided to compare this subset of results using an additional layer of chemical detail. From this extended discussion, we found that (i) the inclusion of on-site corrections resulted in acceptable \( E_1 \) predictions for the subfamily of molecules with an epoxy functional group, and that (ii) partially polarizing the minimal basis set had a larger corrective effect on the subfamilies with more s-character (i.e. ethers, alcohols and saturated hydrocarbons).

However, based on our results we cannot recommend any of the here benchmarked methods to predict \( E_1 \) for non-epoxy saturated molecules. The custom parameter set 3OB(H*) was only employed to highlight the favourable impact that including polarization orbitals has on \( E_1 \) predictions, as it improves the GS estimates of the KS energy gaps. At most, what we proposed should serve to assist the developer community of DFTB+ in setting the direction for improvement towards a more comprehensive solution. If the community reaffirms its interest in studying the excited state of saturated systems, we recommend as an extra step in DFTB+’s development roadmap the inclusion of extra polarization orbitals for each of the elements in 3OB, what may require a re-parameterization of the two-center repulsion splines. The natural following step would be to extend the base formalism to allow polarization via the multipole expansion of the charge density \( \delta \rho \), beyond the zeroth order.

Essentially, we are proposing that future implementations of DFTB+ regarding this topic should be based on the Gaussian polarizable-ion tight biding (GTB) method, developed by Horsfield, Boleininger et al.\textsuperscript{53}. GTB performs on-the-fly computations of the multipole moments, up to quadrupole order, from a Gaussian basis set with the potential to include polarization orbitals. Currently, an implementation of this method exists within another suite of programs for computational chemistry, named Plato\textsuperscript{54–56}. For a reduced set of small hydrocarbon molecules, GTB predicted mean polarizability volumes (\( \alpha_m \)) in excellent agreement with experimental determinations, similarly to DFT-PBE but at a fraction of the computational cost. We focused on a subset of these results, comprising \( \text{H}_2 \) and the first four elements of the homologous series of straight-chain alkanes (shown in Table II of our ESI). For these saturated molecules, in order to significantly counteract the sharp underestimation of
α_m, leading to the discussed overestimation of the σ-
type E_1, it is necessary to append polarization orbitals
to the minimal basis set, since the multipole expansion of
the charge density alone is apparently not sufficient. For
GTB with a second-order expansion of the charge den-
sity, the mean relative absolute percentage error of α_m
goes from approximately 68% to near 8%, just with the
addition of polarization orbitals for each element (i.e., p
orbitals for H and d orbitals for C); for comparison, this
error was close to 5% for DFT-PBE/cc-pVQZ.

Lastly, it is worth mentioning another alternative to
improve σ-type excitation energies, which we do not cover
in this work as it is not implemented into DFTB+. It con-
ists on extending SCC-DFTB to include the chemical-
potential equalisation (CPE) approach, in which an ad-
ditional response density is supplied in order to overcome
the limitations of the minimal basis set with respect to
response properties\textsuperscript{57}.

V. SUMMARY AND CONCLUSIONS

With this work, we support the idea that there are
lessons to be learned from the analysis of large amounts
of chemical data, if we complement the datasets with rel-
levant molecular descriptors and an adequate theoretical
background, even when the datapoints are not as de-
tailed as would be ideal. For example, our data analysis
focused only on E_1 values\textsuperscript{58} from a ML-dataset and did
not take into account any other information related to
the electronic transitions, such as the symmetry of the
excited states, as this extra level of detail is not always
available\textsuperscript{59}. Although this lack of additional information
may have acted as a source of a systematic underestima-
tion error for E_1, which can alternatively be viewed
as a qualitative error in terms of symmetry\textsuperscript{60}, we were
able to extract robust and valuable insights on the lim-
itations of the benchmarked ES-methods. For this task,
taking into account the chemical identity of the molecules
and employing a sufficiently large and chemically diverse
ML-dataset were key aspects.

With our chemically-informed analysis of the GDB-8
chemical subspace, we were able to identify clear trends
in the prediction error distributions for E_1. With this
benchmark we learned that, despite being more approx-
imate, DFTB ES-methods suffer from the same limita-
tions as the parent method, DFT, in the prediction of
E_1, albeit to a greater extent. On the one hand, there
is the self-interaction error, which mainly affects π-type
excitations in molecules with π-conjugated systems or
high p-character. On the other hand, there is the error
arising from the over-simplification in the description of
the ground state density (e.g., due to the use of a non-
complete basis set), which has a significant impact on
σ-type excitations, in molecules with predominantly s-
character. However, in this work we showed that we can
choose between the currently implemented methods and
corrections of DFTB+ in order to predict E_1 with errors
mostly within ±1 eV. Beware that, as this work bases its
conclusions on general trends, any system-specific analy-
sis on the correctness of the energy and symmetry rank-
ing of the low-lying ES becomes delegated to the re-
searchers interested in studying such particular systems.

Our benchmark allowed us to construct a “rule of
thumb” intended to assist the community of users in
selecting the most reliable DFTB approximation for
the ES calculations of their molecular systems of inter-
est. According to this rule, visually summarised in
Fig. 3, when computing E_1 with DFTB+ we recommend us-
ing: (i) DFTB-Casida+OC:3OB for saturated molecules
with epoxide groups, (ii) DFTB-Casida+LC:OB2 for π-conjugated unsaturated molecules, (iii) DFTB-
Casida:3OB for non-conjugated unsaturated molecules
with carbonyl (i.e., ketones/aldehydes), cyano or alkyne

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2}
\caption{Left: Overlapped \( E_1 \) error distributions corresponding to all the molecules in the dataset, computed using DFTB-
Casida with the original parameter set 3OB (red solid line) and the custom minimally polarized set 3OB(H\*) (blue solid line).
Histograms of \( \Delta_{CC1}E_1 \) that correspond only to the subset of saturated molecules are being highlighted as dashed lines in a
darker colour. Right: Cumulative percentage distribution of saturated molecules with \( E_1 \) prediction errors bellow \( \Delta_{CC1}E_1 \); for
each method, we highlight the value of \( \Delta_{CC1}E_1 \) below which 99% of the saturated molecules are contained.}
\end{figure}
groups, and (iv) DFTB-ppRPA:3OB for unsaturated molecules with isolated alkene groups. Ignoring non-epoxy saturated molecules, the combined error distribution resulting from applying the proposed rule of thumb is almost as acceptable as that of the computationally more expensive TD-DFT method with a corrected exchange functional (PBE0) and a partially polarized basis set (def2SVP). Surprisingly, Fig. 3 also shows that, in contrast to TD-DFT, the combined error distribution of the DFTB-approximate methods is nearly centred on $\Delta_{\text{CC2}}E_1 = 0$, i.e. the line corresponding to exact matches with CC2 predictions. As previously discussed, we believe that, since DFTB’s predictions of $\Delta E$ match with CC2 predictions. As previously discussed on $\Delta E$, in contrast to TD-DFT, the combined error distribution on $\Delta E$ is almost as acceptable as that of the computationally more expensive TD-DFT method with a corrected exchange functional (PBE0) and a partially polarized basis set (def2SVP).

To be used for ES calculations with fixed nuclei only, since we have not performed a re-parameterization of the repulsion splines.

The authors acknowledge financial support by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) and the Agencia Nacional de Promoción Científica y Tecnológica (Agencia I+D+i), Argentina, through grants PIP 112-2017-0100892CO and PICT-2017-1506. A. I. B. also thanks CONICET for the doctoral fellowship. Both authors would like to extend a special thanks to Dr. Matías Berdakin (CONICET) for the very helpful feedback. We would also like to thank all the anonymous reviewers who contributed to the overall quality of this manuscript by providing constructive comments after a critical reading.

1. A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Crommbeuf, Chemical Science 10, 9424 (2019).
2. P.-J. Yang, M. Sugiyama, K. Tsuda, and T. Yanai, Journal of Chemical Theory and Computation 16, 3513 (2020).
3. J. F. Joung, M. Han, J. Hwang, M. Jeong, D. H. Choi, and S. Park, JACS Au 1, 427 (2021).
4. J. Y. Choi, P. Zhang, K. Mehta, A. Blanchard, and M. L. Pasini, Journal of Cheminformatics 14, 1 (2022).
5. R. Ramakrishnan, M. Hartmann, E. Tapavica, and O. A. von Lilienfeld, The Journal of Physical Chemistry 143, 084111 (2015).
6. P. Kayastha, S. Chakraborty, and R. Ramakrishnan, Digital Discovery 1, 689 (2022).
7. T. A. Niehaus, Journal of Molecular Structure: THEOCHEM 914, 38 (2009).
8. F. Trani, G. Scalmani, G. Zheng, I. Carneiro, M. J. Frisch, and V. Barone, Journal of Chemical Theory and Computation 7, 3304 (2011).
9. A. Domínguez, B. Aradi, T. Frauenheim, V. Lutsker, and T. A. Niehaus, Journal of Chemical Theory and Computation 9, 4901 (2013).
10. J. C. Kranz, M. Elstner, B. Aradi, T. Frauenheim, V. Lutsker, A. Domínguez, and T. A. Niehaus, Journal of Chemical Theory and Computation 13, 1737 (2017).
11. A. Fibey and D. Jacquemin, Journal of Chemical Theory and Computation 15, 6267 (2019).
12. F. D. Sala and A. Görling, Physical Review Letters 89, 033003 (2002).
FIG. 3. Comparison of $E_1$ error distributions ($\Delta C_{CC}E_1$) normalised by area, for chemical subfamilies of molecules from the dataset, computed using different DFTB methods (at the top, on a white background) and TD-DFT:PBE0/def2SVP (at the bottom, on a grey background). To compute $E_1$ with DFTB+ we followed the rule of thumb we constructed throughout the discussion of results. Accordingly, we employed: DFTB-Casida+OC:3OB for saturated molecules with epoxide groups (red solid line), DFTB-Casida:3OB for the rest of the saturated molecules (green solid line), DFTB-Casida+LC:OB2 for all $\pi$-conjugated unsaturated molecules (purple solid line), DFTB-Casida:3OB for non-conjugated unsaturated molecules with carbonyl (i.e., ketones or aldehydes), cyano or alkyne groups (orange solid line) and DFTB-ppRPA:3OB for unsaturated molecules with isolated alkene groups (blue solid line). The combined error distributions for DFT and DFTB methods are shown (as dashed black lines) superimposed on their respective sides of the plot. The combined error distribution for the DFTB-approximate methods ignores the non-epoxy saturated molecules, as we could not recommend any suitable implementation within DFTB+ for this subgroup. For better visualisation, all DFTB distributions were scaled to double. See the first column of Tables IV & V, in the ESI, for the number of occurrences per chemical family and subgroup.

37 L. Kronik and S. Kümmel, Gas-phase valence-electron photoemission spectroscopy using density functional theory, in First Principles Approaches to Spectroscopic Properties of Complex Materials, Vol. 347 (Springer Berlin Heidelberg, 2014) pp. 137–191.
38 Y. Nishimoto and D. Fedorov, Physical Chemistry Chemical Physics 18, 22047 (2016).
39 S. Grimme and M. Parac, ChemPhysChem 4, 292 (2003).
40 B. Champagne and E. A. Perpète, The Journal of Chemical Physics 109, 10489 (1998).
41 B. Champagne, E. A. Perpète, D. Jacquemin, S. J. A. van Gisbergen, E.-J. Baerends, C. Soubra-Ghaoui, K. A. Robins, and B. Kirtman, The Journal of Physical Chemistry A 104, 4755 (2000).
42 A. Dreuw, J. L. Weisman, and M. Head-Gordon, The Journal of Chemical Physics 119, 2943 (2003).
43 A. Dreuw and M. Head-Gordon, Chemical Reviews 105, 4009 (2005).
44 D. Tozer, The Journal of Chemical Physics 119, 12697 (2003).
45 F. Grozema, S. Tonzani, Y. Berlin, G. Schatz, L. Siebbeles, and M. Ratner, Journal of American Chemical Society 130, 5157 (2009).
46 I.-D. Chai and M. Head-Gordon, The Journal of Chemical Physics 128, 084106 (2008).
47 T. Kubat and M. Elstner, The Journal of Physical Chemistry B 114, 11221 (2010).
48 M. Lundberg and P. Siegbahn, The Journal of Chemical Physics 122, 224103 (2005).
49 B. Hourahine, S. Sanna, B. Aradi, C. Köhler, T. Niehaus, and T. Frauenheim, The Journal of Physical Chemistry A 111, 5671 (2007).
50 M. Rapacioli, F. Spiegelman, A. Scemama, and A. Mirtschin, Journal of Chemical Theory and Computation 7, 44 (2011).
51 C. M. Isborn, B. D. Mar, B. F. E. Curchod, I. Tavernelli, and T. J. Martínez, The Journal of Physical Chemistry B 117, 12189 (2013).
52 D. B. Cook, International Journal of Quantum Chemistry 60, 793 (1996).
53 M. Boleininger, A. A. Y. Guilbert, and A. P. Horsfield, Journal of Chemical Physics 145, 144103 (2016).
Currently, DFTB+ does not support point group symmetry.

For the cases where DFT- or DFTB-based methods predict an excitation energy that exactly matches the reference value (i.e., of CC2), the predicted symmetry of the corresponding ES may not match the reference, as it is possible for the energy ranking of the states to change with the method.

In the ESI, we also performed a brief analysis on the oscillator strengths associated with $E_1$ (i.e., $f_1$) and found results that complement the conclusions of this work.
I. OVERVIEW OF METHODS

A. SCC-DFTB approximations to KS-DFT

We can completely characterise the self-charge-consistent version of DFTB (SCC-DFTB) by listing its approximations to Kohn-Sham’s DFT (KS-DFT). This set of additional approximations to the parent framework are key to its superior efficiency.

DFTB is derived from a truncated expansion of the KS-DFT total energy functional. In this work, we made use of SCC-DFTB, which includes terms up to the second order in the expansion of the ground-state electron density \( \rho \), around a reference density \( \rho_0 \) being perturbed by density fluctuations \( \delta \rho \):

\[
E^{\text{SCC-DFTB}}[\rho_0 + \delta \rho] = E^0[\rho_0] + E^1[\rho_0, \delta \rho] + E^2[\rho_0, (\delta \rho)^2]
\]

The truncation that leads to the above expression constitutes the first approximation of SCC-DFTB to the exact KS-DFT total energy functional. Since \( \rho_0 \) is typically constructed as a superposition of neutral atomic densities, \( \delta \rho \) would be accounting for the chemical environment of each atom within the molecule. In approximate KS-DFT, the molecular orbitals are expanded in the basis of atom centered basis functions, i.e as a linear combination of atomic orbitals (LCAO):

\[
\psi_i(\mathbf{r}) = \sum_{\mu} c_{\mu}^{i} \phi_{\mu}(\mathbf{r})
\]

This projection on atom-centered basis functions converts the time-independent Schrödinger equation into an eigenvalue problem. As a second approximation, DFTB’s standard formulation employs a minimum set of valence orbital basis functions (i.e. a minimal basis set) to simplify the linear algebra operations. In practice, these basis functions are slightly compressed atomic-like solutions to the KS-DFT equations. It should be cautioned that, as standard SCC-DFTB uses a minimal basis set, Rydberg states are outside the scope of DFTB-approximate ES-methods; this particular type of electronic excitations would require, instead, the use of a very diffuse basis set to be described correctly.

In the expansion of the total energy, the zeroth-order term \( E^0[\rho_0] \) is the energy associated with the repulsion interaction between the nuclei and between the atomic contributions to the reference density \( \rho_0 \). The third approximation of DFTB is to write \( E^0[\rho_0] \) as a sum of pair-potentials:

\[
E^0[\rho_0] \approx \frac{1}{2} \sum_{AB} V^{\text{TP}}_{AB}(R_{AB})
\]

The first-order term \( E^1[\rho_0, \delta \rho] \) is the band-structure energy, which involves the computation of matrix elements of the reference Hamiltonian \( H[\rho_0] \):

\[
E^1[\rho_0, \delta \rho] = \sum_i f_i \sum_{\mu \nu} c_{\mu}^{i*} c_{\nu}^{i} H_{\mu \nu}^0, \quad \mu \in A, \nu \in B
\]

where \( f_i \) are the Fermi occupations of the ground-state molecular orbitals within the LCAO ansatz. Obtaining these matrix elements require the computation of three-center integrals, because they involve atomic orbitals from two different centers, \( \phi_{\mu}(\mathbf{r}) \) and \( \phi_{\nu}(\mathbf{r}) \), and the effective potential at the reference density, \( V_{\mu}[\rho_0](\mathbf{r}) \).

The fourth approximation of DFTB is to neglect the two-center crystal-field contributions and the three-center interactions in the band energy term, transforming the reference Hamiltonian diagonal and non-diagonal elements, \( H_{\mu \mu}^0 \) and \( H_{\mu \nu}^0 \), into less expensive one- and two-center integrals, respectively.

Additionally, SCC-DFTB makes two approximations to the energy from charge fluctuations, i.e. the second-order term in the total energy expression \( (E^2[\rho_0, (\delta \rho)^2]) \). The fifth approximation of SCC-DFTB is the monopole approximation of the ground state (GS) charge density \( \delta \rho \). The zeroth-order truncation of the multipole expansion of \( \delta \rho \) neglects the one-center two-electron integrals and introduces the computation of partial atomic charges \( q_A \) (and atomic populations \( \Delta q_A \)) from two-center orbital overlap matrix elements \( S_{\mu \nu} \), under the Mulliken population analysis:

\[
\Delta q_A = q_A - q_A^0
\]

where

\[
q_A = \sum_i f_i \sum_{\mu \in A} \sum_{\nu} \frac{1}{2} \left( c_{\mu}^{i*} c_{\nu} S_{\mu \nu} + c_{\nu}^{i*} c_{\mu} S_{\mu \nu} \right)
\]

and \( q_A^0 \) is the number of valence electrons in neutral atom \( A \). Notice that the charge-fluctuation energy term is the one turning DFTB into a SCC method, since the atomic populations themselves depend on the molecular orbital coefficients. Lastly, SCC-DFTB’s sixth approximation is to enforce the locality of XC contributions,
making the charge-fluctuation interaction only electrostatic (i.e. Coulombic) for different centers. The interactions between the atomic populations follows a dependence with the inter-atomic distance given by an analytical density profile function $\gamma_{AB}$ (e.g. a Gaussian profile) that becomes equal to the Hubbard $U_A$ parameter when $A = B$, which is an on-site contribution linked to the chemical hardness of the atom:

$$E^2 \left[ \rho_0, (\delta \rho)^2 \right] \approx \frac{1}{2} \sum_{AB} \gamma_{AB} (R_{AB}) \Delta q_A \Delta q_B$$

A very strong point towards the superior computational performance of DFTB is the pre-computation of matrix elements and repulsion pair-potential splines. With this strategy, the cost associated with having to compute the approximate integrals is transferred to a one-time parameterization process. The Hamiltonian and overlap matrix elements $F_{\mu\nu}$ and $S_{\mu\nu}$, needed for the first- and second-order energy terms, are highly transferable as they are pre-determined for a reference density. Within the standard SCC-DFTB, these electronic parameters are pre-computed for various interatomic distances using a minimal basis set of pseudo-atomic KS-orbitals, obtained from atomic DFT calculations with a confining potential and a local-XC functional, such as the parameter-free Perdew-Burke-Ernzerhof (PBE) functional of the generalized gradient approximation (GGA) class. In an independent stage of the parameterization process, the zeroth-order repulsive pair-potentials are most often fitted to results from DFT calculations [1-4] for a set of element pairs, but these functions can also be fitted to experimental data (e.g., equilibrium geometries, atomization energies and vibrational frequencies) [5]. When DFT is used to fit the repulsive functions, calculations are expected to be performed close to the basis set limit and with high-quality functionals. These repulsion functions are the parameters that make DFTB a semi-empirical method and, in standard SCC-DFTB, are expected to encode all the chemically relevant non-local nature of the electron-electron interaction. The Hubbard $U_A$ parameters are also obtained from DFT pre-computations. All these numbers are generated via the Slater-Koster (SK) technique [6] and stored as SK-files for different pairs of chemical elements; this set of files is what we call a parameter set in DFTB.

It is also possible to improve the first approximation and extend the expansion of the total energy to include higher-order energy terms. DFTB+ can perform computations up to the third-order in energy. However, this extra energy term only becomes important when bonding results in large atomic charge fluctuations, i.e. for local densities deviating significantly from the reference one.

### B. TD-DFTB from Casida’s LR-TD-DFT

The TD-DFTB method for computing excitation energies is based on Casida’s LR-approach [7] to TD-DFT, the time-dependent extension of the Hohenberg-Kohn theorems. In the dynamic LR treatment of the GS electron density being perturbed by an external electric potential, Casida derived a pseudo-eigenvalue equation in the space of single-orbital transitions from $i, j$-occupied to $a, b$-virtual molecular orbitals:

$$\sum_{jb} \left[ \omega_{ia}^2 \delta_{ij} \delta_{ab} + 4 \sqrt{\omega_{ia} K_{ijab} \sqrt{\omega_{jb}}} \right] F_{i}^{jb} = \omega_f^2 F_{i}^{ja}$$

where the KS orbital energy gaps, $\omega_{ia}$, and the coupling matrix elements, $K_{ijab}$, are constructed from the ground state. This problem can be solved for a number of $I$ electronic transitions, in order to determine their excitation energies $\omega_I$ and transition contributions $F_{i}^{ja}$, which are needed to compute the corresponding transition dipole moments and approximate excited state wavefunctions. It can be noticed that $\omega_I$ results from the correction made by the coupling matrix, which accounts for the electron-hole interaction, to the initial excitation energy estimate given by the KS energy difference. Furthermore, $F_{i}^{ja}$ being a vector indicates that electronic transitions can display multi-orbital character.

TD-DFTB is a translation of the Casida’s eigenvalue problem to the DFTB framework. It consists on extending SCC-DFTB approximations to significantly simplify the computation of the coupling matrix elements $K_{ijab}$. The one-center (or on-site) exchange-like integrals are neglected and the remaining expensive two-center two-electron integrals are converted into simple sums for atom pairs. The following is the resulting approximate expression valid for singlet-singlet transitions:

$$K_{ijab} \approx \sum_{AB} \gamma_{AB} (R_{AB}) q_{ia}^{AB} q_{jb}^{AB}$$

where

$$q_{ia}^{AB} \equiv \sum_{\mu \in A} \sum_{\nu} \left( c_{\mu}^i \xi_n^a S_{\mu\nu} + c_{\nu}^j \xi_n^a S_{\nu\mu} \right)$$

are transition charges introduced by the Mulliken monopole approximation. The XC term in the exact expression for the coupling matrix elements $K_{ijab}^{\text{exact}}$ depends on the full ground-state electron density $\rho$. That means the XC contribution not only depends on the reference density, as in the second-order energy term, but also on the density fluctuations from the $\rho_0$ GS-reference. This would require promoting the Hubbard $U_A$ parameters into functions of the site’s atomic population, i.e. $U_A (\Delta q_A)$. However, a third approximation is made to TD-DFTB in order to neglect these charge fluctuations when computing the coupling matrix elements, allowing reuse of the approximate profiles $\gamma_{AB}$ from the initial GS computations.

It is worth pointing out that, as Niehaus has previously shown [8], the results from the DFTB-approximate Casida eigenvalue problem are completely equivalent to those extracted from the real-time TD-DFTB approach,
for which an implementation in DFTB+ was reported by Bonafé et al. [9]. The latter approach may pose a challenge in the deconvolution of excitation peaks for spectrosopically complex systems, but it has some advantages over TD-DFTB-Casida: (i) it does not require truncating the number of excitations to be computed, (ii) it can result in a superior computational performance for very large systems, and (iii) it can be extended for the calculation of transient absorption spectra (TAS) simulations. For small systems, like the ones we are employing for this benchmark, TD-DFTB-Casida was the most convenient choice in terms of performance and output parsing.

C. DFTB-approximate ppRPA

Borrowed from nuclear physics, ppRPA is an eigenvalue problem describing 2-electron addition and removal processes. It can be used to predict the excitation energies of N-electron systems via the treatment of two-electron additions in corresponding two-electron deficient (N-2) systems:

\[ \omega_{0 \rightarrow n} = \omega_{n+2}^{+2} - \omega_{n+2}^{0+2} \]

where \( \omega_{0+2}^{0+2} \) and \( \omega_{n+2}^{0+2} \) are the eigenvalues for the ground state and the \( n \)-excited state, respectively. The above expression can be better understood with the help of the diagram in Fig. S1 of this ESI. Once again, the translation into the DFTB framework consists on the approximation of bottleneck integrals. Originally, the matrix elements to be calculated before solving the eigenvalue problem each involve two integrals of two electrons. In DFTB-approximate ppRPA, the Mulliken monopole approximation reduces these expensive integrals to summations of simpler terms, which share strong similarities with the expression for the coupling matrix elements in TD-DFTB.

II. EXTENSIONS TO THE STANDARD SCC-DFTB METHOD

A. Including long-range corrections in TD-DFTB

To better account for non-local contributions and reproduce the correct -1/R asymptotic trend for Coulombic interactions in the long distance, parameter sets for DFTB can be constructed with range-separated or long-range corrected (LC) XC functionals. Within this scheme, the two-electron interaction is split into short-range and long-range components, with the splitting being modulated by the range-separation parameter \( \omega \):

\[ \frac{1}{r} = \frac{1 - e^{-\omega r}}{r} + \frac{e^{-\omega r}}{r} \]

The implementation of DFTB+LC [10–12] into DFTB+ is quite recent and therefore there is only one openly available parameter set prepared to include these corrections into DFTB: OB2-(1-1) [4]. This SK-set employs the range-separation parameter \( \omega = 0.3 a_0^{-1} \) and manages to reproduce GS geometries and vibrations of CHON organic molecules with a similar quality as DFTB3:3OB [3]. Thanks to the DFTB+ community being actively working to extend existing parameter sets, there is reported a very recent re-parameterisation of OB2 to include sulphur heteroatoms in organic molecules [13]. Yet, we could not find any other reported extension of OB2 and therefore we decided to ignore the fluorinated organic molecules for our DFTB+LC calculations, a subset that only represents about 1.4% of all the molecules in the dataset.

B. Including on-site corrections in TD-DFTB

The standard SCC-DFTB formalism can be extended to partially correct the monopole approximation of the transition charge density in Casida TD-DFTB, in order to no longer neglect the on-site integrals of the exchange type. This implementation is known in DFTB+ as on-site corrections (OC) [14]. This correction requires providing additional on-site constants that depend only on the XC-kernel, which we extracted from the Appendix J of the DFTB+ manual [15]. Once again, we have ignored fluorinated molecules for the DFTB+OC computations, as there are no available pre-computed on-site constants for fluorine atoms.

C. Partially polarizing the minimal basis set

The standard formulation of DFTB is characterised by the use of a minimal basis set. However, we could instead employ an extended, yet limited, basis set for the pre-computation of the electronic integrals in the parameter set SK-files. With the SkProgs package [16] for DFTB+, we constructed a proof-of-concept, custom SK-parameter set that works with a minimally polarized minimal basis set. Since the electronic part of this custom SK-parameter set was intended to emulate an extension of the 3OB parameter set [3] to include minimal polarization on H atoms only, we decided to name this set “3OB(H+)”.

To achieve this minimally polarized set, we added empty 2p orbitals to provide extra angular degrees of freedom to the valence electrons on the hydrogen centers and break the original radial symmetry of the 1s orbital. The radial dependence of the extra polarization orbitals was scaled to bring it closer to that of the 1s valence orbital, but without compromising its original profile. We achieved the custom radial probability densities (see Fig. S2 of this ESI) by limiting the radial wave function (see equation 7.2 in the DFTB+ manual [15]) to one variational coefficient \( c \) and one exponent \( \alpha = \sqrt{2} \):

\[ R_p(r) = c r e^{-\sqrt{2}r} \]
As we only intended to perform single-point computations (i.e., with fixed nuclei), it was not necessary to re-parameterise the pair-repulsion splines, which we kept from the original 30B SK-files.

III. FURTHER DISCUSSION OF RESULTS

A. Non-conjugated unsaturated molecules

In order to construct a better “rule of thumb” for the DFTB-approximate ES-methods, we included another layer of chemical detail for the non-conjugated unsaturated molecules. In Fig. S3 of this ESI, it can be seen that DFTB-Casida:30B performed best for non-conjugated molecules with carbonyl (i.e., ketones and aldehydes), cyano and alkyne groups, for which their $E_1$ error distributions are centered near $\Delta_{CC2}E_1 = 0$ and mostly contained within ±1 eV. However, DFTB-Casida:30B may not be the preferred choice for non-conjugated alkenes, as its estimates of $E_1$ are affected by a systematic underestimation and a significant error dispersion. We suspect that this systematic underestimation of $E_1$ for alkenes is also related to the self-interaction error.

Not all systems are equally affected by the SIE. We expect the SIE-induced underestimation of the delocalised solutions to be more pronounced in molecules with naturally delocalised molecular orbitals of large spatial extent (e.g., $\pi$-conjugated systems). If the delocalised orbitals do participate in low-lying electronic transitions (e.g., frontier orbitals such as HOMO or LUMO), then a systematic underestimation of the corresponding excitation energies is also to be expected. Molecules of the chemical subgroup of alkenes are characterised by the presence of an isolated carbon-carbon double bond that is not part of a conjugated $\pi$-system and, therefore, is rather spatially confined. In the case of alkenes we would have expected a fairly small underestimation of the first excitation energy. However, alkenes appear to be strongly affected by a systematic underestimation of $E_1$.

For systems affected by the SIE, the DFTB-ppRPA method is expected to achieve better results and, as it can be seen in Fig. S1 of this ESI, this is indeed what is observed for alkenes. It remains to be asked what makes alkenes more susceptible to this error than other compounds with isolated double or triple bonds. We speculate that alkenes may suffer from the SIE-induced artificial stabilisation of delocalised solutions to a greater extent, and thus have underestimated their $E_1$ predictions, as they are more easily polarisable than the other sub-families of non-conjugated unsaturated molecules, with permanent dipole moments (e.g., carbonyl and cyano groups) or with higher order bonds (i.e., alkynes). In Table III of this ESI, it can be seen that the measured polarisabilities [17–19] of linear alkenes are indeed higher than those of other linear non-conjugated unsaturated molecules of the same length (i.e., with the same number of non-H atoms). On a related side note, the SIE was found to decrease with increasing disparity of electron affinities between the electron and hole regions [20], which may reinforce the reason why the SIE is lower for carbonyl and cyano groups, where excitations between frontier MOs are expected to involve a non-bonding molecular orbital ($\pi$) highly localised on the heteroatom (O or N, respectively) and an anti-bonding $\pi$ orbital ($\pi^*$) delocalised over the double bond.

B. Unsaturated $\pi$-conjugated molecules

In Fig S5 of this ESI, we show that DFTB+LC-Casida:OB2 performs particularly well for the subset of unsaturated molecules with $\pi$-conjugated systems involving only 3 atoms. The $E_1$ error distributions for these chemical subgroups can be seen to be contained between $\Delta_{CC2}E_1 = 1 \text{ eV}$. The most accurate $E_1$ predictions appear to have been achieved for the subgroup of carboximides; however, we can also highlight the ester, carboxylic acid and amide functional groups, which are known for their chemical importance and ubiquity.

C. Saturated molecules

Since the inclusion of extensions to the standard SCC-DFTB method was motivated in part on improving the predictions of $E_1$ for the saturated molecules, we decided to compare this subset of results using an extra layer of chemical detail.

The inclusion of on-site corrections resulted in a redshift of $E_1$ for all saturated molecules containing at least one oxygen or nitrogen heteroatom: the average shift observed for epoxides was of about -1.2 eV, and of approximately -0.9 eV for aziridines, while hydrocarbons (saturated molecules of C and H only) were almost unaffected by this correction. Thanks to the on-site corrections, DFTB-Casida+OC:30B was able to provide acceptable $E_1$ predictions for the subfamily of molecules with an epoxy functional group.

As seen in Fig. S7 of this ESI, both the on-site corrected DFTB-Casida+OC:30B and the minimally polarized DFTB-Casida:30B(H*) performed best for the sub-families of saturated molecules containing highly strained three-membered rings, such as epoxide, aziridine and cyclopropyl groups.

These ring structures have markedly acute bond angles and therefore greater $p$-character than the non-strained saturated molecules. Together with the large orbital overlap that exists at the center of these small rings, $\sigma$-electrons end up delocalising with a stabilising effect on the anti-bonding virtual MOs, in what is known as hyperconjugation [21]. The electron density to be delocalised on the rings can also come from geminal $n$-orbitals (due to an interaction known as negative hyperconjugation) and from electropositive substituents [22]. The highly strained saturated molecules are a special case. On the
one hand, being saturated molecules, their electronic excitations are expected to involve σ-type occupied molecular orbitals and, therefore, may be in need of a better description of the electron density to avoid overestimating their energies. On the other hand, their higher p-character would give rise to delocalised MOs, particularly virtual states near the frontier, which can potentially be underestimated in energy by the SIE. Therefore, for these subgroups of molecules we would expect cases that give rise to beneficial error compensation.

While it would certainly be interesting to further investigate our hypothesis on error compensation, it is also beyond the scope of this study. At this stage, we limit ourselves to updating our rule of thumb to include the recommendation to use DFTB-Casida+OC:3OB when predicting $E_1$ for the saturated molecules of the epoxy family.

We also calculated the average shifts observed in $\Delta_{CC2}E_1$ after the inclusion of partial polarisation (H*) and of long-range corrections (LC), for the sets of three-membered ring molecules; for each of the chemical groups discussed we have observed smooth monomodal distributions with a full width at half maximum (FWHM) between 0.5 eV and 1 eV (not shown).

On a hand, the addition of partial polarisation red-shifted $E_1$ in all cases, suggesting that the electron density oversimplification error was present. We observed the largest mean shift for cyclopropanes (-0.66 eV), followed by aziridines (-0.53 eV) and epoxides (-0.40 eV); we correlated this with the number of geminal lone pairs in each chemical group (epoxides have two because of oxygen, aziridines have one from nitrogen, and the cyclopropyl group has none). We would expect molecules with fewer valence p-electrons to rely more on the quality of the electron density of σ bonds for their $E_1$ predictions.

On another hand, the inclusion of long-range corrections resulted in a detrimental blue-shift for all $E_1$ predictions, which may signal that the SIE was present. Again, we observed the largest average shift for cyclopropanes (+2.15 eV), but this time it was followed by epoxides (+1.85 eV) and then aziridines (+1.48 eV). We previously conjectured that systems with more delocalised solutions and higher polarisabilities were expected to be more affected by the SIE. To compare the polarisabilities of these three chemical subgroups we can refer to Table III, available in this ESI. In this table, we can see that there is a trend in polarisabilities based on elemental composition: hydrocarbons are more polarisable than molecules with oxygen, and both are more polarisable than nitrogen-containing molecules. Although the discussed compounds are cyclic, and solely comprised of single bonds, we can think of three-membered rings as being more similar to their very short (2 and 3 non-H atoms) linear analogues.

D. Regarding oscillator strengths

In Fig. S8 of this ESI the following can be noticed: (i) for saturated molecules, $f_1$ prediction errors are present almost exclusively for the DFTB-approximate methods (see the datapoints populating the horizontal line at $y = 0$); (ii) for π-conjugated molecules with more than 3 conjugated atoms, the underlying limitation is shared between the compared methods, in a considerable amount of cases (see the datapoints along the diagonal line at $y = x$); (iii) for π-conjugated molecules with 3 conjugated atoms and for the unsaturated non-conjugated molecules, $f_1$ prediction errors are present but are rather mild (see the datapoints clustered near the origin).

This same information can be extracted from the information compiled in Tables IV & V of this ESI, where we also provide a second layer of chemical detail (i.e., a breakdown of results into the different subgroups within each main chemical family). One additional observation that we can make to the values in the aforementioned tables is that, for molecules with π-conjugation, there is a tendency for the accuracy of $f_1$ predictions to worsen with increasing number of atoms involved in the π-conjugated system.

[1] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, Physical Review B 51, 12947 (1995).
[2] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Physical Review B 58, 7260 (1998).
[3] M. Gaus, A. Goez, and M. Elstner, Journal of Chemical Theory and Computation 9, 338 (2013).
[4] V. Q. Vuong, J. A. Kuriappan, M. Kubillus, J. J. Kranz, T. Mast, T. A. Niehaus, S. Irl, and M. Elstner, Journal of Chemical Theory and Computation 14, 115 (2018).
[5] M. Gaus, C.-P. Chou, H. Witke, and M. Elstner, The Journal of Physical Chemistry A 113, 11866 (2009).
[6] J. C. Slater and G. F. Koster, Physical Review 94, 1498 (1954).
[7] M. E. Casida, Recent Advances in Density Functional Methods 5, 155 (1995).
[8] T. A. Niehaus, Journal of Molecular Structure: THEOCHEM 914, 38 (2009).
[9] F. P. Bonafè, B. Aradi, B. Hourahine, C. R. Medrano, F. J. Hernández, T. Frauenheim, and C. G. Sánchez, Journal of Chemical Theory and Computation 16, 4454 (2020).
[10] T. A. Niehaus, S. Suhai, F. D. Sala, P. Lugli, M. Elstner, G. Seifert, and T. Frauenheim, Physical Review B 63, 085108 (2001).
[11] T. A. Niehaus and F. D. Sala, Physica Status Solidi (b) 249, 237 (2012).
[12] V. Lutsker, B. Aradi, and T. A. Niehaus, The Journal of Chemical Physics 143, 184107 (2015).
[13] M. T. do N. Varella, L. Stojanović, V. Q. Vuong, S. Irle, T. A. Niehaus, and M. Barbatti, The Journal of Physical Chemistry C 125, 5458 (2021).
[14] A. Domínguez, B. Aradi, T. Frauenheim, V. Lutsker, and T. A. Niehaus, Journal of Chemical Theory and Computation 9, 4901 (2013).
[15] DFTB+ User Manual, https://dftbplus.org/documentation (Last accessed: 2022-05-04).
[16] B. Hourahine, SkProgs: Package containing a few programs that are useful in generating Slater-Koster files for the DFTB-method, https://github.com/dftbplus/skprogs (Last accessed: 2022-05-04).
[17] M. Gussoni, R. Rui, and G. Zerbi, Journal of Molecular Structure 447, 163 (1998).
[18] R. J. (editor), NIST Standard Reference Database 101, http://cccbdb.nist.gov/ (2022).
[19] Y. Zevatskii and S. Lysova, Russian Journal of Applied Chemistry 79, 967–974 (2006).
[20] M. Lundberg, Y. Nishimoto, and S. Irle, International Journal of Quantum Chemistry 112, 1701 (2012).
[21] S. Inagaki, Y. Ishitani, and T. Kakefu, Journal of the American Chemical Society 116, 1994 (1994).
[22] J. I.-C. Wu and R. Schleyer, Pure and Applied Chemistry 85, 921 (2013).
[23] M. Boleininger, A. A. Y. Guilbert, and A. P. Horsfield, Journal of Chemical Physics 145, 144103 (2016).
[24] T. N. Olney, N. M. Cann, G. Cooper, and C. E. Brion, Chemical Physics 223, 59 (1997).
| Main chemical family | Chemical sub-family | SMARTS fragment |
|----------------------|---------------------|-----------------|
| **Saturated molecules** | Epoxy | [C;r3]-[O;r3]-[C;r3] |
| | Aziridine | [C;r3]-[N;r3]-[C;r3] |
| | Cyclopropane | [C;r3]-[C;r3]-[C;r3] |
| | Oxetane / Azetidine | [C;r4]-[O,N;r4]-[C;r4] |
| | Tetrahydrofurane / Pyrrolidine | [C;r5]-[O,N;r5]-[C;r5] |
| | Ether / Alcohol | [C]-[O] |
| | Other Aliphatic | [CX4] |
| **Non-conjugated unsaturated molecules** | Ketone / Aldehyde | O=C |
| | Cyano | N#C |
| | Alkyne | C#C |
| | Alkene | C=C |
| | Ketone / Aldehyde | O=C |
| | Cyano | N#C |
| | Alkyne | C#C |
| | Alkene | C=C |
| | With no instances in the dataset: | O=N , N=C , N=N |
| | Nitroso, Imine, Azo | |
| **π-conjugated molecules with 3 conjugated atoms** | Amide | O=C-N |
| | Carboximidate | N=C-O |
| | Ester / Carboxylic Acid | O=C-O |
| | Oxime | C=N-O |
| | Amidine | N=C-N |
| | With no instances in the dataset: | N=N-O , C=C-N , C=C-O |
| | Nitro / Nitrite, Diazo, Azide Isocyanate | O=N-O , N=N-C , N=N=N |

TABLE I. This table shows the SMARTS fragments that were employed to recognise chemical sub-families within each of the main principal chemical families. These additional molecular descriptors were applied in the order in which they appear in the table, from top to bottom. We made no distinction of sub-families within the family of π-conjugated molecules with more than 3 conjugated atoms.
TABLE II. This table shows an ordered subset of results obtained by Boleininger et al. [23], exactly as they appear in their original publication. They correspond to mean polarizability volumes ($\alpha_m$) in Å$^3$, computed with the Gaussian polarizable-ion Tight Binding method with a second-order expansion of the charge density (GTB2) for H$_2$ and four saturated molecules (the first four elements in the homologous series of straight-chain alkanes). Calculations were carried out with a minimal basis set (SV) and a polarizable basis set (SVP). For comparison, we included experimental determinations by Olney et al. [24] and results from DFT-PBE with the correlation-consistent quadruple-zeta valence basis set (cc-pVQZ) [23].

|               | GTB2/SV [23] | GTB2/SVP [23] | DFT:PBE/cc-pVQZ [23] | Expt. [24] |
|---------------|--------------|---------------|-----------------------|------------|
| H$_2$         | 0.18         | 0.90          | 0.70                  | 0.79       |
| methane       | 0.77         | 2.30          | 2.46                  | 2.45       |
| ethane        | 1.46         | 3.94          | 4.32                  | 4.23       |
| propane       | 2.13         | 5.56          | 6.23                  | 5.92       |
| butane        | 2.81         | 7.22          | 8.14                  | 7.69       |

Table III. This table shows an ordered subset of results that originally appeared in publications by Gussoni et al. [17] (also available at the NIST database [18]) and Zevatskii et al. [19]. The values correspond to experimentally measured electric dipole polarisabilities in Å$^3$ units, for non-conjugated unsaturated linear organic molecules containing the functional group at one end; we added linear alkanes for comparison. Note that for molecules of the same length (equal number of non-H atoms, n), the polarisabilities of alkenes are higher than those of the other unsaturated non-conjugated compounds; interestingly, a general trend can be extracted for these polarisability ($\alpha$) values: $\alpha$(alkanes) > $\alpha$(alkenes) > $\alpha$(alkynes) > $\alpha$(aldehydes) > $\alpha$(cyanides).
Table IV. This table shows the percentage of molecules with oscillator strengths ($f_1$) that are considered to be good estimates (compared to CC2), for TD-DFT:PBE0/def2SVP and the DFTB-approximate ES methods of Fig. S7 (in this ESI). In other words, this table shows for each chemical family and subgroup the percentage of molecules with absolute $f_1$ prediction errors ($|\Delta_{CC2}f_1|$) that fell below the 0.01 threshold (depicted as dashed lines in Fig. S7), a value that is customarily used to distinguish between dark ($f < 0.01$) and bright ($f = 0.01$) excitations. In parentheses, we report the percentage of molecules with oscillator strengths that are considered to be similar between the compared methods (i.e., $|\Delta_{DFT}f_1| < 0.01$). In the first column, for each chemical family and subgroup we also indicate the number of occurrences in the data set (inside the square brackets). Here we show data for saturated and non-conjugated unsaturated molecules; please refer to Table V within this ESI for data corresponding to π-conjugated molecules.
| Unsaturated π-conjugated with 3 conj. atoms [3166] | TD-DFT:PBE0/def2SVP | DFTB-Casida:3OB | DFTB-Casida:+LC:OB2 | DFTB-Casida:3OB(H*) |
|-------------------------------------------------|----------------------|-----------------|----------------------|--------------------|
| ↪ Amide [1186]                                  | 99.0%                | 95.3%           | 97.2%                | 96.4%              |
| (96.1%)                                         | (98.5%)              | (96.4%)         |                      |                    |
| ↪ Carboximidate [1000]                          | 93.4%                | 80.4%           | 78.2%                | 81.7%              |
| (85.3%)                                         | (85.5%)              | (84.6%)         |                      |                    |
| ↪ Ester/Carboxylic Acid [678]                    | 98.4%                | 95.9%           | 97.05%               | 96.0%              |
| (95.7%)                                         | (96.5%)              | (95.7%)         |                      |                    |
| ↪ Oxime [200]                                   | 81.0%                | 76.0%           | 78.0%                | 75.5%              |
| (83.0%)                                         | (83.5%)              | (86.0%)         |                      |                    |
| ↪ Amidine [102]                                 | 61.8%                | 58.8%           | 50.0%                | 59.8%              |
| (84.3%)                                         | (76.5%)              | (84.3%)         |                      |                    |
| Unsaturated π-conjugated with >3 conj. atoms [6332] |                      |                 |                      |                    |
| ↪ 4 Conj. atoms [965]                           | 87.8%                | 82.0%           | 83.5%                | 82.8%              |
| (87.7%)                                         | (88.0%)              | (88.5%)         |                      |                    |
| ↪ 5 Conj. atoms [1552]                          | 68.1%                | 63.1%           | 62.8%                | 64.7%              |
| (70.1%)                                         | (69.3%)              | (68.5%)         |                      |                    |
| ↪ 6 Conj. atoms [1357]                          | 58.2%                | 42.9%           | 34.2%                | 42.5%              |
| (49.7%)                                         | (41.1%)              | (48.7%)         |                      |                    |
| ↪ 7 Conj. atoms [1457]                          | 59.8%                | 46.7%           | 38.3%                | 45.4%              |
| (59.6%)                                         | (48.7%)              | (58.5%)         |                      |                    |
| ↪ 8 Conj. atoms [1001]                          | 55.0%                | 44.3%           | 42.5%                | 43.5%              |
| (58.8%)                                         | (52.5%)              | (57.4%)         |                      |                    |

Table V. This table shows the percentage of molecules with oscillator strengths ($f_1$) that are considered good estimates (compared to CC2), for TD-DFT:PBE0/def2SVP and DFTB-approximate ES methods in Fig. S5 (in this ESI). In other words, this table shows for each chemical family and subgroup the percentage of molecules with absolute $f_1$ prediction errors (|$\Delta_{CC2} f_1$|) that fell below the 0.01 threshold (depicted as dashed lines in Fig. S5), a value that is customarily used to distinguish between dark ($f < 0.01$) and bright ($f \geq 0.01$) excitations. In parentheses, we report the percentage of molecules with oscillator strengths that are considered to be similar between the compared methods (i.e., |$\Delta_{DFT} f_1$| < 0.01). In the first column, for each chemical family and subgroup we also indicate the number of occurrences in the data set (inside the square brackets). Here we show data for π-conjugated molecules; please refer to Table IV within this ESI for data corresponding to saturated and non-conjugated unsaturated molecules.
An excitation energy for the N-electron system ($\omega_{0\rightarrow n}$) can be computed as the difference between the energies of two-electron addition processes ($\omega_0^+^2$ and $\omega_n^+^2$), from the ground state of a two-electron deficient system (i.e. with N-2 electrons) into the ground and n-th excited states of the resulting N-electron system. The horizontal bars represent electronic states of the same system, with energies $E_0 (N-2) >> E_n (N) > E_0 (N)$.

**Figure S1.** Diagram depicting how ppRPA computes excitation energies from the energies of two-electron addition processes. An excitation energy for the N-electron system ($\omega_{0\rightarrow n}$) can be computed as the difference between the energies of two-electron addition processes ($\omega_0^+^2$ and $\omega_n^+^2$), from the ground state of a two-electron deficient system (i.e. with N-2 electrons) into the ground and n-th excited states of the resulting N-electron system. The horizontal bars represent electronic states of the same system, with energies $E_0 (N-2) >> E_n (N) > E_0 (N)$.

**Figure S2.** Radial probability densities for the hydrogen orbitals in the original 3OB and in the custom H-polarized 3OB(H*) parameter sets. As was intended, the custom 1s orbital for H in the 3OB(H*) parameter set (red solid line) coincides with that of the original 3OB set (black dashed line). The three 2p orbitals in 3OB(H*), allowing polarization, have a radial wave function with a probability density (blue solid line) that partially overlaps that of the 1s orbital.
Figure S3. Overlapped $E_1$ error distributions for chemical subfamilies within the family of non-conjugated unsaturated molecules (i.e. molecules characterised by isolated double and triple bonds). Histograms were computed using results from DFTB-Casida:3OB (solid lines) and DFTB-ppRPA:3OB (dashed line). See the first column of Table IV, in this ESI, for the number of occurrences per chemical family and subgroup.

Figure S4. Overlapped $E_1$ error distributions for all the unsaturated $\pi$-conjugated molecules, computed using DFTB-Casida:3OB (red dashed line), DFTB-ppRPA:3OB (blue dashed line) and DFTB-Casida+LC:3OB (purple solid line). Note that including long-range corrections into DFTB-Casida computations (i.e. computing with DFTB+LC-Casida:OB2) achieves an accuracy similar to that of DFTB-ppRPA:3OB, with a slightly better precision. See the first column of Table V, in this ESI, for the number of occurrences per chemical family and subgroup.
**Figure S5.** Overlapped $E_1$ error distributions for chemical subfamilies within the family of unsaturated $\pi$-conjugated molecules with 3 conjugated atoms. Histograms were computed using results from DFTB-Casida+LC:OB2. See the first column of Table V, in this ESI, for the number of occurrences per chemical family and subgroup.

**Figure S6.** Comparison of prediction errors $\Delta_{CC2}E_1$ for TD-DFT:PBE0/def2SVP and DFTB-Casida+OC:3OB. The scattered datapoints correspond to each of the nearly 21,800 molecules in the GDB-8 chemical subspace. The datapoints and the projected histograms were coloured according to the main chemical identity of the compounds: green for saturated molecules, orange for non-conjugated molecules with an isolated double or triple bond, and blue and purple for $\pi$-conjugated molecules with 3 or more conjugated atoms, respectively. See the first column of Tables IV & V, in this ESI, for the number of occurrences per chemical family and subgroup.
Figure S7. Overlapped $E_1$ error distributions for chemical groups within the family of saturated molecules. On-site corrected TD-DFTB results, obtained with DFTB-Casida+OC:3OB, are displayed at the top, on a white background. Histograms computed from results obtained with the minimally polarized DFTB-Casida:3OB(H*) are shown at the bottom, on a grey background. For ease of comparison, in both backgrounds we have included to scale, as dark-grey shaded areas, the original $E_1$ error distribution for DFTB-Casida:3OB corresponding to all the saturated molecules. See the first column of Tables IV & V, in this ESI, for the number of occurrences per chemical family and subgroup.
Figure S8 . Comparison of prediction errors $\Delta CC_2 f_1$, where $f_1$ is the oscillator strength associated to $E_1$, for TDDFT:PBE0/def2SVP and three different DFTB-approximate approaches: Casida:3OB (1st column), Casida+LC:OB2 (2nd column) and Casida:3OB(H*) (3rd column). The scattered datapoints correspond to each of the nearly 21,800 molecules in the GDB-8 chemical subspace. The datapoints and the projected histograms were coloured according to the main chemical identity of the compounds: green for saturated molecules (1st row), orange for non-conjugated molecules with an isolated double or triple bond (2nd row), and blue and purple for $\pi$-conjugated molecules with 3 or more conjugated atoms (3rd and 4th rows, respectively). See the first column of Tables IV & V, in this ESI, for the number of occurrences per chemical family and subgroup.