Electronic Supplementary Information (ESI) for "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)

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 ρ, around a reference density ρ_0 being perturbed by density fluctuations $\delta\rho$:

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

The truncation that leads to the above expression constitutes the *first approximation* of SCC-DFTB to the exact KS-DFT total energy functional. Since ρ_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):

$$\psi_i\left(\boldsymbol{r}\right)=\sum_\mu c_\mu^i\phi_\mu\left(\boldsymbol{r}\right)$$

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 ρ_0. The *third approximation* of DFTB is to write $E^0[\rho_0]$ as a sum of pair-potentials:

$$E^0[\rho_0]\approx\frac{1}{2}\sum_{AB}V_{AB}^{rep}\left(\boldsymbol{R}_{AB}\right)$$

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

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

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, $\phi_\mu\left(\boldsymbol{r}\right)$ and $\phi_\nu\left(\boldsymbol{r}\right)$, and the effective potential at the reference density, $V_s\left[\rho_0\right]\left(\boldsymbol{r}\right)$. 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 ($E^2\left[\rho_0,\,\left(\delta\rho\right)^2\right]$). 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 Δq_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

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

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 γ_{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:

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

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 $H^0_{\mu\mu}$ 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 generalised 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:

$$\sum_{jb}\left[\omega_{ia}^2\delta_{ij}\delta_{ab} + 4\sqrt{\omega_{ia}}K_{ijab}\sqrt{\omega_{jb}}\right]F_I^{jb} = \omega_I^2 F_I^{ia}$$

where the KS orbital energy gaps, ω_{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 ω_I and transition contributions F_I^{ia}, which are needed to compute the corresponding transition dipole moments and approximate excited state wavefunctions. It can be noticed that ω_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, $\boldsymbol{F_I}$ 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:

$$K_{ijab} \approx \sum_{AB}\gamma_{AB}\left(R_{AB}\right)\;q_A^{ia}\;q_B^{jb}$$

where

$$q_A^{ia} \equiv \sum_{\mu\in A}\sum_{\nu}\frac{1}{2}\left(c_\mu^{i\,*}\,c_\nu^a\,S_{\mu\nu} + c_\nu^{i\,*}\,c_\mu^a\,S_{\mu\nu}\right)$$

are transition charges introduced by the Mulliken monopole approximation. The XC term in the exact expression for the coupling matrix elements K_{iajb} depends on the full ground-state electron density ρ. That means that 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 ρ_0 GS-reference. This would require promoting the Hubbard U_A parameters into functions of the site's atomic population, i.e. $U_A\left(\Delta q_A\right)$. 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 γ_{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 spectroscopically 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:

$$\omega_{0\to n} = \omega_n^{+2} - \omega_0^{+2}$$

where ω_0^{+2e} and ω_n^{+2e} 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 ω:

$$\frac{1}{r} = \frac{1 - e^{-\omega\, r}}{r} + \frac{e^{-\omega\, r}}{r}$$

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 reproduces 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 3OB 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:3OB 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 $\pm 1\ eV$. However, DFTB-Casida:3OB 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., π-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 π-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 subfamilies 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 (n) highly localised on the heteroatom (O or N, respectively) and an anti-bonding π orbital (π^*) delocalised over the double bond.

B. Unsaturated π-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 π-conjugated systems involving only 3 atoms. The E_1 error distributions for these chemical subgroups can be seen to be contained between $\Delta_{CC2}E_1 \pm 1\ eV$. The most accurate E_1 predictions appear to have been achieved for the subgroup of carboximidates; 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 red-shift 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:3OB 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:3OB and the minimally polarized DFTB-Casida:3OB(H*) performed best for the subfamilies 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, σ-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*, 1995, **51**, 12947–12957.

[2] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, *Physical Review B*, 1998, **58**, 7260–7268.

[3] M. Gaus, A. Goez and M. Elstner, *Journal of Chemical Theory and Computation*, 2013, **9**, 338–354.

[4] V. Q. Vuong, J. A. Kuriappan, M. Kubillus, J. J. Kranz, T. Mast, T. A. Niehaus, S. Irle and M. Elstner, *Journal of Chemical Theory and Computation*, 2018, **14**, 115–125.

[5] M. Gaus, C.-P. Chou, H. Witek and M. Elstner, *The Journal of Physical Chemistry A*, 2009, **113**, 11866–11881.

[6] J. C. Slater and G. F. Koster, *Physical Review*, 1954, **94**, 1498–1524.

[7] M. E. Casida, *Recent Advances in Density Functional Methods*, 1995, **5**, 155–192.

[8] T. A. Niehaus, *Journal of Molecular Structure: THEOCHEM*, 2009, **914**, 38–49.

[9] F. P. Bonafé, B. Aradi, B. Hourahine, C. R. Medrano, F. J. Hernández, T. Frauenheim and C. G. Sánchez, *Journal of Chemical Theory and Computation*, 2020, **16**, 4454–4469.

[10] T. A. Niehaus, S. Suhai, F. D. Sala, P. Lugli, M. Elstner, G. Seifert and T. Frauenheim, *Physical Review B*, 2001, **63**, 085108.

[11] T. A. Niehaus and F. D. Sala, *Physica Status Solidi (b)*, 2012, **249**, 237–244.

[12] V. Lutsker, B. Aradi and T. A. Niehaus, *The Journal of Chemical Physics*, 2015, **143**, 184107.

[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*, 2021, **125**, 5458–5474.

[14] A. Domínguez, B. Aradi, T. Frauenheim, V. Lutsker and T. A. Niehaus, *Journal of Chemical Theory and Computation*, 2013, **9**, 4901–4914.

[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*, 1998, **447**, 163–215.

[18] R. J. (editor), *NIST Standard Reference Database*, 2022, **101**, http://cccbdb.nist.gov/.

[19] Y. Zevatskii and S. Lysova, *Russian Journal of Applied Chemistry*, 2006, **79**, 967–974.

[20] M. Lundberg, Y. Nishimoto and S. Irle, *International Journal of Quantum Chemistry*, 2012, **112**, 1701–1711.

[21] S. Inagaki, Y. Ishitani and T. Kakefu, *Journal of the American Chemical Society*, 1994, **116**, 1994.

[22] J. I.-C. Wu and R. Schleyer, *Pure and Applied Chemistry*, 2013, **85**, 921–940.

[23] M. Boleininger, A. A. Y. Guilbert and A. P. Horsfield, *Journal of Chemical Physics*, 2016, **145**, 144103.

[24] T. N. Olney, N. M. Cann, G. Cooper and C. E. Brion, *Chemical Physics*, 1997, **223**, 59–98.

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
	With no instances in the dataset: Nitroso, Imine, Azo	O=N , N=C , N=N
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:
	Azoxy, Enamine, Enol
	Nitro / Nitrite, Diazo, Azide Isocyanate
		N=N-O , C=C-N , C=C-O
		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.

	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 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 (α_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].

[Polarizabilities]	**n: number of non-H atoms (C, O, N)**			
	$n = 2$	$n = 3$	$n = 4$	$n = 5$
·**C-C**· (alkanes)	CH_3CH_3 (ethane) [4.226 Å^3]	$H(CH_2)_2CH_3$ (propane) [5.921 Å^3]	$H(CH_2)_3CH_3$ (butane) [8.020 Å^3]	$H(CH_2)_4CH_3$ (pentane) [9.88 Å^3]
·**C=C**· (alkenes)	CH_2CH_2 (ethylene) [4.076 Å^3]	CH_3CHCH_2 (propene) [5.990 Å^3]	$H(CH_2)_2CHCH_2$ (1-butene) [7.830 Å^3]	$H(CH_2)_3CHCH_2$ (1-pentene) [9.65 Å^3]
·**C≡C**· (alkynes)	$CHCH$ (acetylene) [3.487 Å^3]	CH_3CCH (propyne) [5.550 Å^3]	$H(CH_2)_2CCH$ (1-butyne) [7.410 Å^3]	$H(CH_2)_3CCH$ (1-pentyne) [9.12 Å^3]
·**C=O** (aldehydes)	CH_2O (formaldehyde) [2.770 Å^3]	CH_3CHO (acetaldehyde) [4.278 Å^3]	$H(CH_2)_2CHO$ (propanal) [6.350 Å^3]	$H(CH_2)_3CHO$ (butanal) [8.20 Å^3]
·**C≡N** (cyanides)	HCN (hydrogen cyanide) [2.346 Å^3]	CH_3CN (acetonitrile) [4.280 Å^3]	$H(CH_2)_2CN$ (propionitrile) [6.240 Å^3]	$H(CH_2)_3CN$ (butanenitrile) [8.40 Å^3]

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 (α) values: α(alkanes) > α(alkenes) > α(alkynes) > α(aldehydes) > α(cyanides).

$\lvert\Delta_{CC2} f_1\rvert < 0.01$ ($\lvert\Delta_{DFT} f_1\rvert < 0.01$)	**TD-DFT:PBE0** **/def2SVP**	**DFTB-Casida** **:3OB**	**DFTB-Casida** **+LC:OB2**	**DFTB-Casida** **:3OB(H*)**
All Saturated [5531]	**87.7%**	**45.2%** **(49.3%)**	**32.4%** **(34.9%)**	**46.7%** **(51.0%)**
$\hookrightarrow$ Epoxy [734]	78.3%	60.2% (72.8%)	56.5% (71.5%)	60.6% (73.0%)
$\hookrightarrow$ Aziridine [1010]	86.4%	43.9% (47.0%)	24.1% (28.0%)	45.7% (48.3%)
$\hookrightarrow$ Cyclopropane [1693]	92.1%	24.1% (24.7%)	11.8% (11.1%)	20.4% (21.3%)
$\hookrightarrow$ Oxetane/Azetidine [755]	85.8%	68.0% (74.6%)	54.6% (55.6%)	73.3% (78.9%)
$\hookrightarrow$ Tetrahydrofurane /Pyrrolidine [276]	94.2%	60.5% (66.3%)	52.9% (55.6%)	59.4% (61.2%)
$\hookrightarrow$ Ether/Alcohol [852]	90.6%	52.8% (52.3%)	39.7% (36.5%)	59.6% (64.7%)
$\hookrightarrow$ Other Aliphatic [211]	78.0%	36.8% (40.2%)	19.6% (19.6%)	50.7% (56.0%)
All Unsaturated **non-conjugated [6757]**	**86.1%**	**72.3%** **(77.8%)**	**73.7%** **(78.1%)**	**72.4%** **(78.2%)**
$\hookrightarrow$ Ketone/Aldehyde [2775]	100.0%	96.0% (96.2%)	99.3% (99.3%)	95.6% (96.0%)
$\hookrightarrow$ Alkyne [1520]	86.5%	69.9% (76.4%)	65.1% (69.3%)	72.4% (79.9%)
$\hookrightarrow$ Cyano [1443]	75.1%	56.4% (70.8%)	55.2% (66.3%)	55.2% (69.3%)
$\hookrightarrow$ Alkene [1019]	63.6%	33.7% (39.6%)	42.8% (50.2%)	33.3% (40.7%)

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.

$\lvert\Delta_{CC2} f_1\rvert < 0.01$ ($\lvert\Delta_{DFT} f_1\rvert < 0.01$)	**TD-DFT:PBE0 /def2SVP**	**DFTB-Casida :3OB**	**DFTB-Casida +LC:OB2**	**DFTB-Casida :3OB(H*)**
Unsaturated π-conjugated with 3 conj. atoms [3166]	**94.8%**	**88.3%** **(91.4%)**	**88.4%** **(92.3%)**	**89.2%** **(91.5%)**
↪ Amide [1186]	99.0%	95.3% (96.1%)	97.2% (98.5%)	96.4% (96.4%)
↪ Carboximidate [1000]	93.4%	80.4% (85.3%)	78.2% (85.5%)	81.7% (84.6%)
↪ Ester/Carboxylic Acid [678]	98.4%	95.9% (95.7%)	97.05% (96.5%)	96.0% (95.7%)
↪ Oxime [200]	81.0%	76.0% (83.0%)	78.0% (83.5%)	75.5% (86.0%)
↪ Amidine [102]	61.8%	58.8% (84.3%)	50.0% (76.5%)	59.8% (84.3%)
Unsaturated π-conjugated with >3 conj. atoms [6332]	**65.0%**	**54.9%** **(64.2%)**	**51.0%** **(58.7%)**	**54.9%** **(63.3%)**
↪ 4 Conj. atoms [965]	87.8%	82.0% (87.7%)	83.5% (88.0%)	82.8% (88.5%)
↪ 5 Conj. atoms [1552]	68.1%	63.1% (70.1%)	62.8% (69.3%)	64.7% (68.5%)
↪ 6 Conj. atoms [1357]	58.2%	42.9% (49.7%)	34.2% (41.1%)	42.5% (48.7%)
↪ 7 Conj. atoms [1457]	59.8%	46.7% (59.6%)	38.3% (48.7%)	45.4% (58.5%)
↪ 8 Conj. atoms [1001]	55.0%	44.3% (58.8%)	42.5% (52.5%)	43.5% (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 ($\lvert\Delta_{CC2} f_1\rvert$) 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 ($f >= 0.01$) excitations. In parentheses, we report the percentage of molecules with oscillator strengths that are considered to be similar between the compared methods (i.e., $\lvert\Delta_{DFT} f_1\rvert < 0.01$). 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.

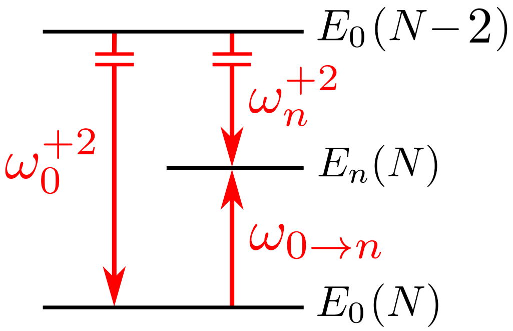

Figure S1 . Diagram depicting how ppRPA computes excitation energies from the energies of two-electron addition processes. An excitation energy for the N-electron system ($\omega_{0\to n}$) can be computed as the difference between the energies of two-electron addition processes (ω_0^{+2} and ω_n^{+2}), 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 $E_0\,(N-2) >> E_n\,(N) > E_0\,(N)$.

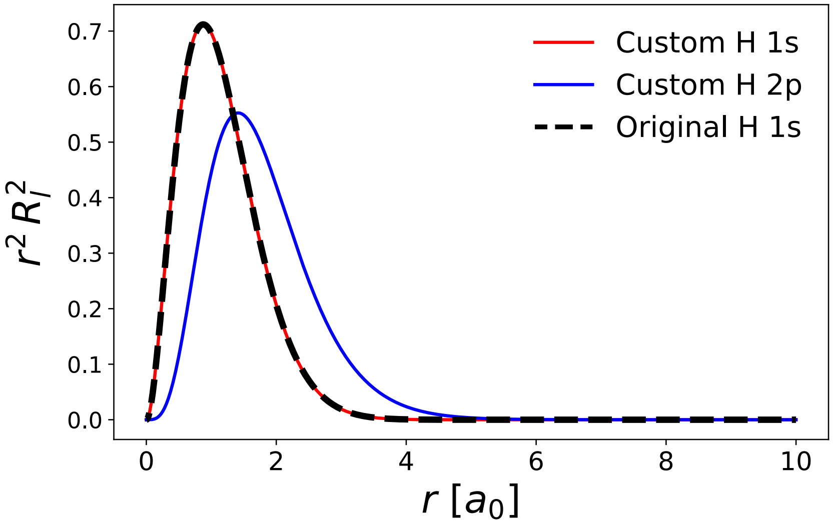

Figure S2 . Radial probability densities for the hydrogen orbitals in the original 3OB and in the custom H-polarized 3OB(H*) parameter sets. As was intended, the custom 1s orbital for H in the 3OB(H*) parameter set (red solid line) coincides with that of the original 3OB set (black dashed line). The 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.

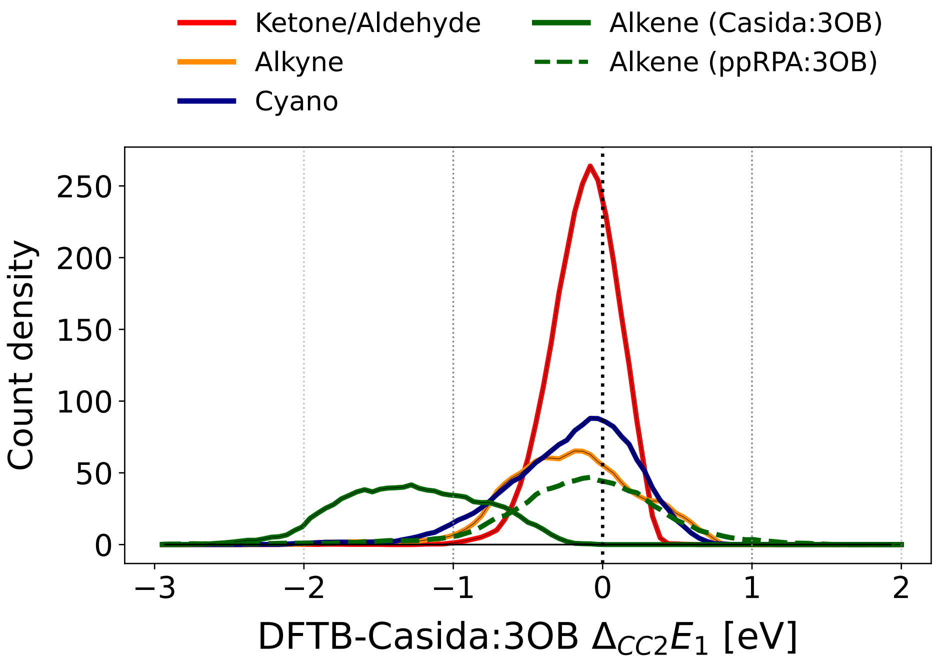

Figure S3 . Overlapped E_1 error distributions for chemical subfamilies within the family of non-conjugated unsaturated molecules (i.e. molecules characterised by isolated double and triple bonds). Histograms were computed using results from DFTB-Casida:3OB (solid lines) and DFTB-ppRPA:3OB (dashed line). See the first column of Table IV, in this ESI, for the number of occurrences per chemical family and subgroup.

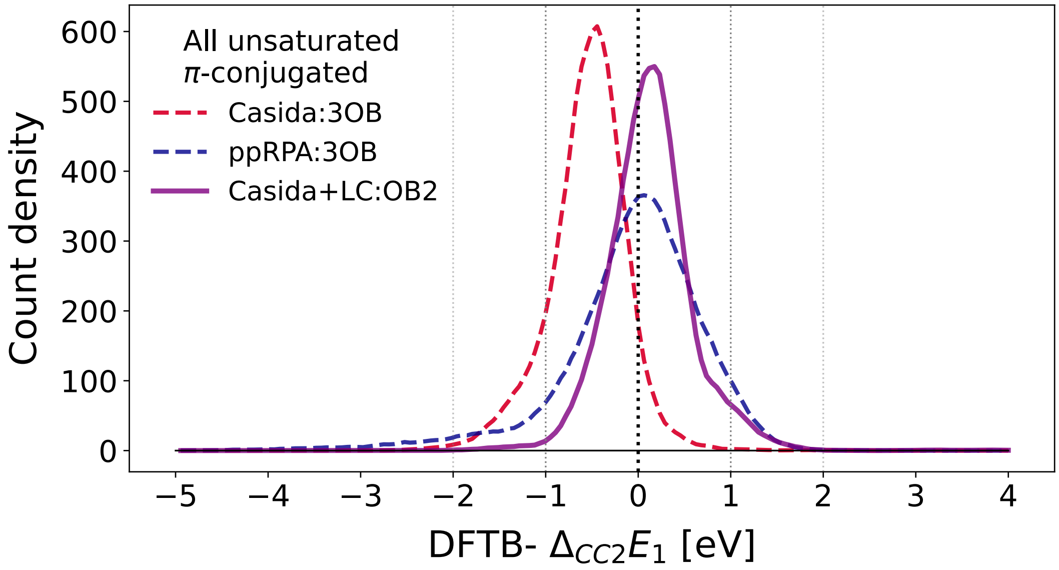

Figure S4 . Overlapped E_1 error distributions for all the unsaturated π-conjugated molecules, computed using DFTB-Casida:3OB (red dashed line), DFTB-ppRPA:3OB (blue dashed line) and DFTB-Casida+LC:3OB (purple solid line). Note that including long-range corrections into DFTB-Casida computations (i.e. 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.

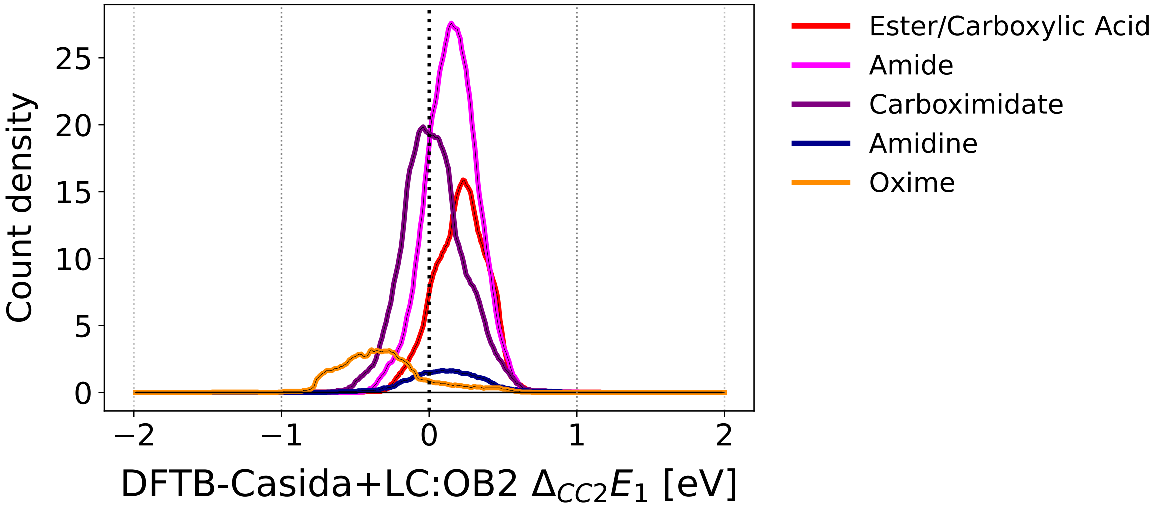

Figure S5 . Overlapped E_1 error distributions for chemical subfamilies within the family of unsaturated π-conjugated molecules with 3 conjugated atoms. Histograms were computed using results from DFTB-Casida+LC:OB2. See the first column of Table V, in this ESI, for the number of occurrences per chemical family and subgroup.

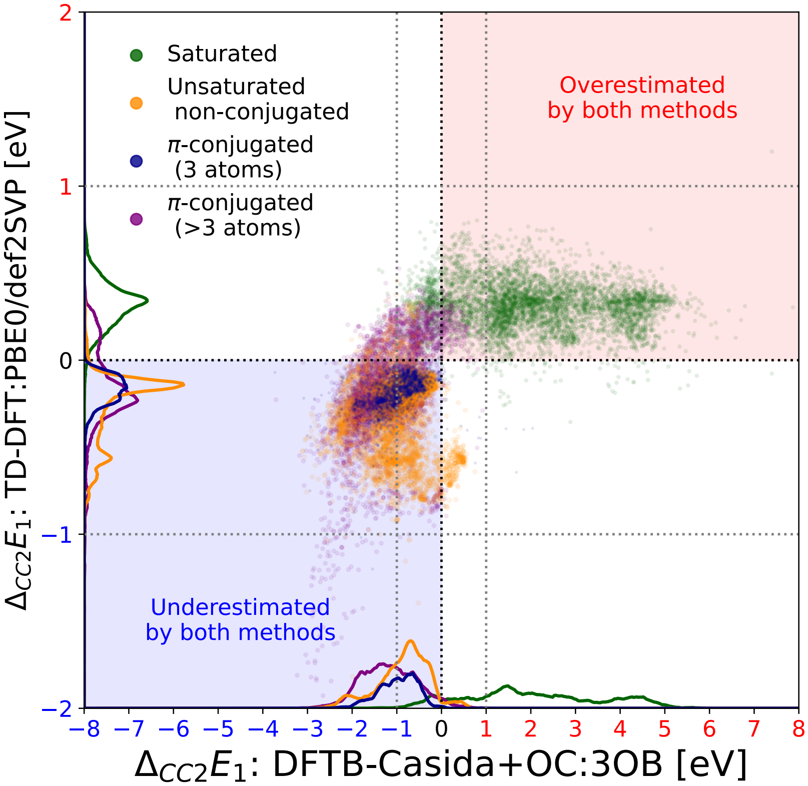

Figure S6 . Comparison of prediction errors $\Delta_{CC2}E_1$ for TD-DFT:PBE0/def2SVP and DFTB-Casida+OC:3OB. The scattered datapoints correspond to each of the nearly 21,800 molecules in the GDB-8 chemical subspace. The datapoints and the projected histograms were coloured according to the main chemical 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 π-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.

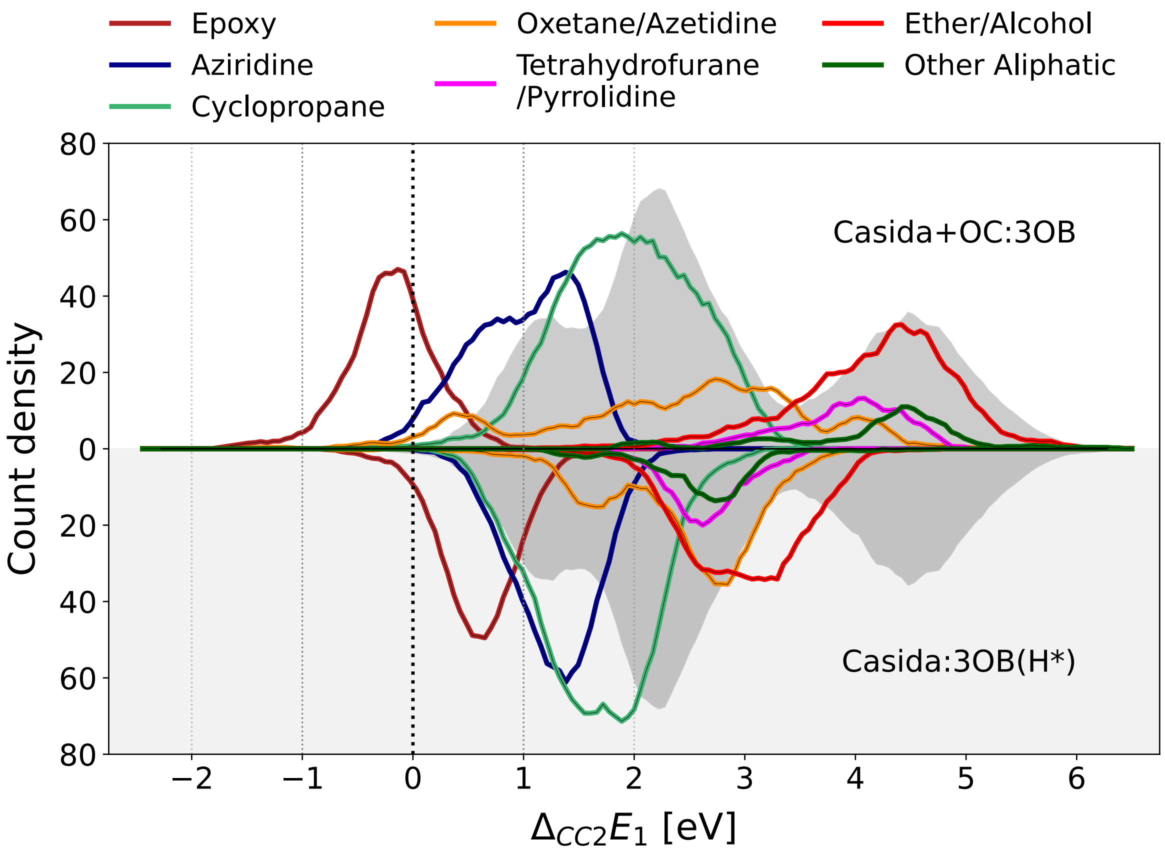

Figure S7 . Overlapped E_1 error distributions for chemical groups within the family of saturated molecules. On-site corrected TD-DFTB results, obtained with DFTB-Casida+OC:3OB, are displayed at the top, on a white background. Histograms computed from results obtained with the minimally polarized DFTB-Casida:3OB(H*) 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.

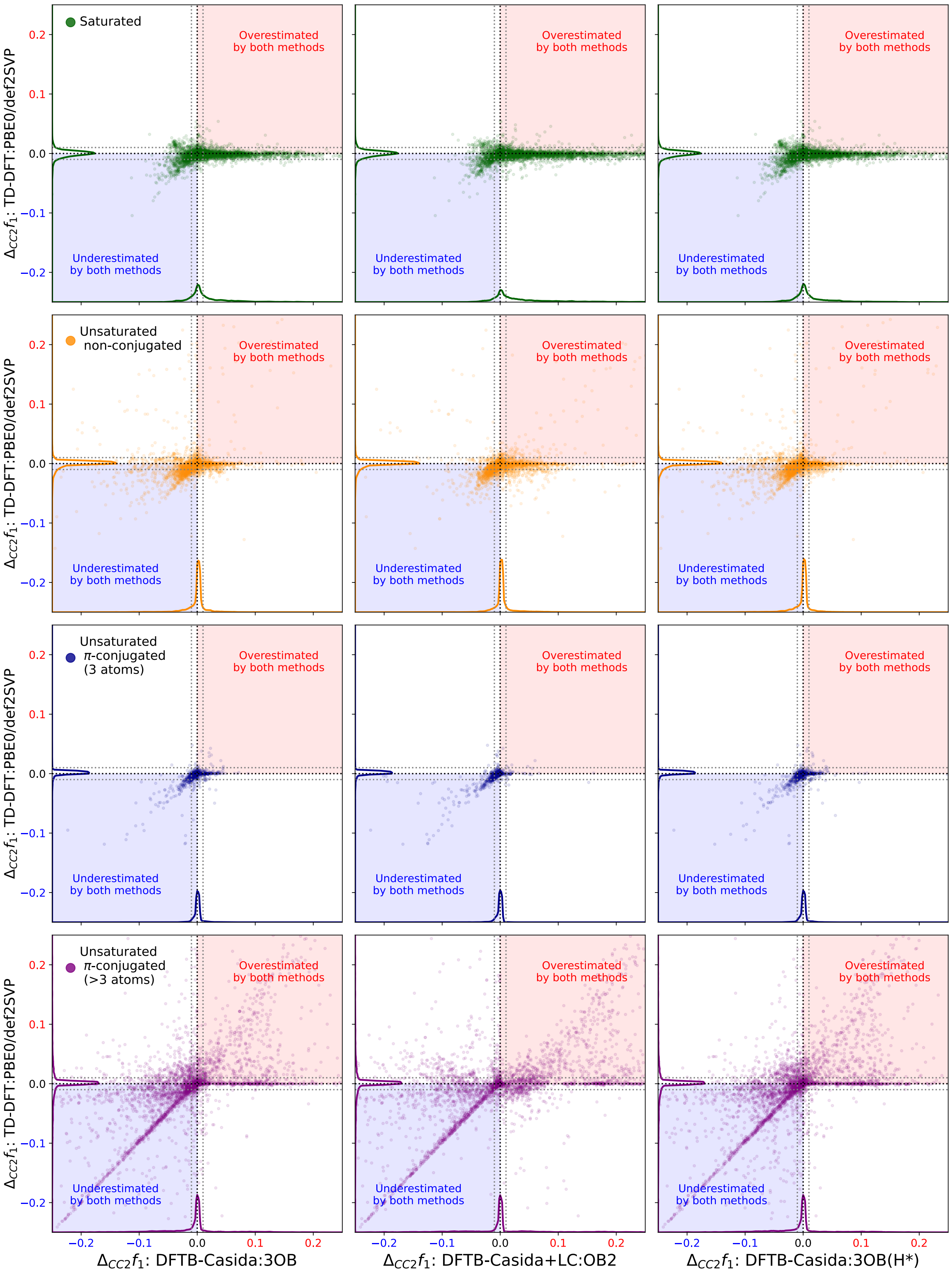

Figure S8 . Comparison of prediction errors $\Delta_{CC2}f_1$, where f_1 is the oscillator strength associated to E_1, for TD-DFT:PBE0/def2SVP and three different DFTB-approximate approaches: Casida:3OB *(1st column)*, Casida+LC:OB2 *(2nd column)* and Casida:3OB(H*) *(3rd column)*. The scattered datapoints 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 π-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.