Expedite computation of arbitrary-order nonlinear optical properties with native electronic interactions in the time domain

Emilia Ridolfi,[1] Paolo E. Trevisanutto,[1] and Vitor M. Pereira[2, 1, *]

[1]*Centre for Advanced 2D Materials, National University of Singapore, 6 Science Drive 2, Singapore 117546*
[2]*Department of Physics, National University of Singapore, 2 Science Drive 3, Singapore 117542*

(Dated: July 9, 2019)

We adapted a recently proposed time-domain framework to characterize the optical response of interacting electronic systems in order to expedite its computation without compromise in quantitative or qualitative accuracy at the microscopic level. With reliable parameterizations of Hamiltonians and interactions, our formulation allows for increased economy and flexibility in calculating the optical response functions to fields of arbitrary temporal shape and strength. For example, the computation of high-harmonic susceptibilities to arbitrary order becomes straightforward within a unified scheme that natively takes into account excitonic effects, as well as deviations of the electronic system from equilibrium under a strong field. Given that two-dimensional semiconductors are currently of much interest for their strong optical nonlinearities, largely defined by excitons, we demonstrate the approach by computing the frequency-dependent susceptibilities of monolayer MoS_2 and hexagonal boron nitride up to the third-harmonic. In the latter, a two-band model brings further insight on the role of intra-band transitions and the nonequilibrium state of the system when computing even-order response, like the second-harmonic susceptibility. Being grounded on a generic non-equilibrium many-body perturbation theory, this framework allows extensions to handle more generic interaction models or the realistic description of electronic processes taking place at ultrafast time scales.

I. INTRODUCTION

The nonlinear optical response of crystalline materials defines a rich playground of phenomena that are attractive for optoelectronic applications[1,2]. Although traditionally limited to a few select bulk crystals with intrinsically high nonlinearities[1], the recent advent and proliferation of countless two-dimensional (2d) crystals has enormously broadened the range of platforms for studying nonlinear optical properties[2–5]. Several classes of 2d crystals host combined characteristics that foretell a store of new effects and functionalities, foremost of which are frequent nontrivial topological characteristics[6,7] and enhanced electronic interactions stemming from their strict two-dimensionality[8–10]. A consequence of the latter is that excitons become an essential and defining element of their optical response over extended frequency ranges[5,11], particularly when it comes to higher-order effects such as harmonic generation — notably because materials such as transition metal dichalchogenide (TMD) semiconductors harbor bound excitons at energies that are practically resonant with lasers in standard use, thereby showing a consistently strong and easy-to-access nonlinear response[5,12]. For these reasons, the ability to theoretically understand and model these characteristics is of very high current interest.

The strong optical response of 2d semiconductors also means they can more efficiently be driven out of equilibrium, which is of interest to fundamentally understand microscopic mechanisms underlying a number of proposed functionalities, such as their valley and spin relaxation characteristics crucial for applications in valleytronics and spintronics, respectively[6,7]. Ultrafast spectroscopy experiments are a versatile and proven tool in this regard[13–17] which, in turn, demands realistic and accurate theoretical methods to model the microscopic transient response of electrons on fast timescales. Here, too, interactions are essential, not only to capture the correct excitations, but also the hot relaxation pathways at short timescales. Handling interactions simultaneously with accurate descriptions of the underlying electronic structure, such as in *ab initio* density functional theory (DFT) methods, is a perennially challenging problem, both methodologically and computationally[18]. In this context, the current paper demonstrates a good compromise between the microscopic accuracy of fully DFT-based calculations and expediency when modeling the response of an electronic system to strong electromagnetic fields directly in the time domain.

Explicitly accounting for electronic interactions in a systematic perturbative expansion with respect to an external field quickly becomes a cumbersome task due to the combinatorial proliferation of terms involving both matrix elements and excitonic wavefunctions that are required at each order[19,20]. If the coupling to the external field is described in the length-gauge, that proliferation is even more severe due to the need to explicitly separate intra- and inter-band transitions[21,22]. The typical development of the perturbative series in the frequency domain will also be inadequate to describe phenomena that are intrinsically of the time domain, such as transient processes in response to intense fields. In addition, strong fields drive the electronic system out of equilibrium which thus requires a nonequilibrium theoretical framework while, for a direct connection with time-resolved spectroscopy, it is desirable to develop capabilities to describe the systems' reaction not only to monochromatic continuous-wave excitation, but to arbi-

trary time-dependent fields.

There have been recent developments to approach these theoretical challenges in terms of the net electric response in the time domain within an *ab initio* framework, thus benefiting from the unbiased nature of DFT calculations[18] as well as their accuracy when extended with quasiparticle corrections within many-body perturbation theory[2,23–32]. While these represent remarkable conceptual and pragmatic progress, practical implementations remain arduous because of the complexity inherent to a self-consistent description of many-body interactions, deviations from equilibrium, and relaxation. In this regard, the enhanced Coulomb interactions characteristic of 2d materials add to these practical challenges because of more stringent convergence demands[20,33–35]. This effort will benefit from implementations that can deliver faster results within the same level of accuracy of a first-principles approach.

This paper contributes in that direction, beginning from the framework originally proposed by Attaccalite *et al.*, which combined DFT and nonequilibrium many-body perturbation theory to compute the time-dependent polarization in electronic systems excited by arbitrary electric fields[28]. Our approach trades the self-consistent calculation of the Kohn-Sham Hamiltonian, the dynamically screened Coulomb interaction, and *GW* quasiparticle corrections, by parameterized tight-binding (TB) and screening models that have been demonstrated to be reliable and accurate in the context of 2d materials, especially in the treatment of the excitonic degrees of freedom[11]. Its key features are: the ability to compute, non-perturbatively, the response to fields with arbitrary strength and temporal profile; the native inclusion of electronic interactions in the time evolution of the excited states; and the explicit consideration that the external field drives the electronic system out-of-equilibrium during its time evolution. In weak external fields, it becomes equivalent to a perturbative response calculation based on the solution of the Bethe-Salpeter equation (BSE)[28], although its power is best revealed, for example, in the ability to extract nonlinear susceptibilities to arbitrary order in a one-shot computation with no more technical effort than what is necessary to obtain the linear response. Formulated in terms of Green's functions (GFs), and having all the effects of interactions and relaxation encoded into an electronic self-energy, this strategy lends itself to systematic extensions beyond the presently explored approximations. But, above all, the fact it hinges on parameterized — yet accurate — Hamiltonians makes this not only an expedite but also a flexible framework to tackle the theoretical description of strong nonlinearities. Cases in point would be when these calculations need to cover a parameter space of interest for a given material (*e.g.*, as a function of strain or doping), or for large-scale deployment of such calculations, as one might envisage will be necessary to add linear and nonlinear optical properties to the catalogs currently being developed by a number of materials' database projects[36–40].

Whereas a fully *ab initio* approach would be computationally prohibitive in both situations, an implementation as we describe below is certainly within reach of current computational capabilities.

We demonstrate the concept with an application to monolayers of molybdenum disulfide (MoS_2) and hexagonal boron nitride (hBN). The former is representative of the important family of TMD semiconductors, which we chose to explicitly illustrate that excellent qualitative and quantitative agreement is possible. As a reasonable bandstructure parameterization to study harmonic generation across this family frequently requires consideration of 8–10 bands[11,41,42], it further illustrates the expediency of this approach with a relatively demanding model parameterization. For hBN, we chose a minimal 2-band TB model to establish the importance of a truly non-equilibrium formulation where, in addition to the role played by intra-band matrix elements of the dipole operator[22], the time-evolution of the electronic populations should be explicitly accouted for in calculations of nonlinear optical properties using minimal TB parameterizations.

The remainder of this paper is organized as follows. For conceptual self-consistency, Section II revisits the key aspects of the methodology first proposed in Ref. 28 as a time-domain version of the BSE. In addition to outlining the development of the equation of motion for the relevant GF, its subsections describe our specific approach to the TB parameterizations within the Slater-Koster scheme[43], how the relevant matrix elements are computed in this framework, the parameterization of the screened Coulomb interaction, and the approximation to the effective self-energy. For pedagogical purposes, we include additional subsections covering practical aspects of the Fourier analysis, of two phenomenological relaxation schemes, as well as additional notes related to our implementation. Section III contains the core of our results by applying the technique to the cases of MoS_2 and hBN. Its subsections describe the specific parameterizations used for each material, a brief overview of the main spectral features in their optical response, an application of the technique in a one-shot scenario to demonstrate it recovers the optical absorption spectrum alternatively calculated in a linear Kubo formalism from the solution of the BSE, and a demonstration of its application to extract high-harmonic susceptibilities. Prior to our conclusion in Section V, Section IV addresses the main trade-offs of time-domain calculations, their intrinsic adequacy to probe ultrafast electronic processes, the role of intra-band matrix elements *vis-à-vis* the need to employ a nonequilibrium distribution, and the method's overall numerical scaling.

II. METHODOLOGY

The central goal is to compute the time-dependent polarization (dipole moment per unit area of the crystal)

induced in a non-polar crystal, $\boldsymbol{P}(t)$, in response to an *arbitrary* (non-sinusoidal) time-dependent electromagnetic field, $\boldsymbol{E}(t)$, in a way that is non-perturbative in $\boldsymbol{E}(t)$ and takes into account the major effects of electron-electron interactions. The latter aspect is fundamental for an adequate description of the optical properties of 2d semiconductors, where excitonic renormalization effects are very pronounced and dominate the optical spectra[5,9]. In the most general terms, the response to a field of arbitrary strength is determined by the n-th order susceptibilities according to

$$
\begin{aligned}
\epsilon_0^{-1} P^\lambda(t) = & \int^t dt_1 \, \chi^{(1)}_{\lambda\alpha}(t, t_1) E^\alpha(t_1) \\
& + \int^t dt_1 \int^{t_1} dt_2 \, \chi^{(2)}_{\lambda\alpha\beta}(t, t_1, t_2) E^\alpha(t_1) E^\beta(t_2) \\
& + ...
\end{aligned} \tag{1}
$$

Being non-perturbative, the calculated $\boldsymbol{P}(t)$ will contain, for example, the information of all nonlinear susceptibilities $\chi^{(n)}$ (which can be extracted by appropriate post-processing procedures) with excitonic effects included from the outset. To capture microscopic coherence in $\boldsymbol{P}(t)$ to all orders (in the perturbative sense) in the external field requires a nonequilibrium framework[44].

Our approach is strongly inspired by the technique and approximations proposed by Attaccalite *et al.* in the context of a fully *ab initio* strategy to obtain the optical response in the time domain based on the formalism of nonequilibrium GFs[28]. However, we depart from their original formulation by basing our single-particle Hamiltonian on a parameterized TB for the renormalized quasiparticles, as well as by describing the Coulomb interactions by a parameterized screened potential. In doing so, we take advantage of parameterizations of those quantities that have already been shown to reliably describe the optical absorption spectrum of select target materials with quantitative accuracy[11]. Most importantly, treating these quantities in an effective way, instead of *ab initio*, expedites the numerical time-integration and makes calculations at each time step less memory-demanding, especially if the desired energy range comprises many bands and/or when the desired energy resolution requires a large number of $\boldsymbol{k}$ points in the sampling of the Brillouin zone (BZ)[45]. Our objective is thus to explicitly show that such an approach is both reliable and efficient.

Throughout this article we consider that the target system remains spatially homogeneous under the influence of the external radiation field. This is appropriate since the frequencies of interest are of the order of the optical bandgap (typically in the infrared-to-visible range), with associated wavelengths much larger than the atomic distances. As we concentrate on strictly 2d crystals, all our vector quantities are restricted to the xy plane, which coincides with the crystalline sheet.

The observable of interest, $\boldsymbol{P}(t)$, can be expressed in terms of the one-particle reduced density matrix as[46]

$$
\boldsymbol{P}(t) = e \int \boldsymbol{r} \hat{\rho}(\boldsymbol{r}) d\boldsymbol{r} = \frac{e}{A} \sum_{mn\boldsymbol{k}} \boldsymbol{r}_{mn\boldsymbol{k}} \, \rho_{mn\boldsymbol{k}}(t), \tag{2}
$$

where $e < 0$ is the charge of the electron, A is the area of the crystal ($A = A_c N_k^2$ for N_k^2 unit cells and area per cell $A_c \equiv \sqrt{3}a^2/2$), $\rho_{mn\boldsymbol{k}}(t) \equiv \langle a^\dagger_{m\boldsymbol{k}}(t) a_{n\boldsymbol{k}}(t) \rangle$, the Heisenberg operator $a^\dagger_{m\boldsymbol{k}}(t)$ creates an electron in the Bloch eigenstate $\psi_{m\boldsymbol{k}}(\boldsymbol{r}) \equiv \langle \boldsymbol{r}|m\boldsymbol{k}\rangle$ at time t, and $\boldsymbol{r}_{mn\boldsymbol{k}} \equiv \langle m\boldsymbol{k}|\hat{\boldsymbol{r}}|n\boldsymbol{k}\rangle$. As we are interested in a non-equilibrium description, we introduce the two-time *lesser* GF[47,48] in the Bloch representation,

$$
\begin{aligned}
G^<_{mn\boldsymbol{k}}(t, t') &\equiv i \langle a^\dagger_{n\boldsymbol{k}}(t') a_{m\boldsymbol{k}}(t) \rangle \\
&= i \int d\boldsymbol{r} d\boldsymbol{r}' \, \psi^*_{m\boldsymbol{k}}(\boldsymbol{r}) \psi_{n\boldsymbol{k}}(\boldsymbol{r}') \, \langle \hat{\psi}^\dagger(\boldsymbol{r}', t') \hat{\psi}(\boldsymbol{r}, t) \rangle,
\end{aligned} \tag{3}
$$

where $\hat{\psi}(\boldsymbol{r}, t)$ are the electronic field operators, $i \langle \hat{\psi}^\dagger(\boldsymbol{r}', t') \hat{\psi}(\boldsymbol{r}, t) \rangle \equiv G^<(\boldsymbol{r}t, \boldsymbol{r}'t')$, and whose time-diagonal component coincides with the reduced density matrix: $G^<_{mn\boldsymbol{k}}(t) \equiv G^<_{mn\boldsymbol{k}}(t, t) = i \rho_{nm\boldsymbol{k}}(t)$. Hence, Eq. (2) can be recast as

$$
\boldsymbol{P}(t) = -\frac{ie}{A} \sum_{mn\boldsymbol{k}} \boldsymbol{r}_{mn\boldsymbol{k}} \, G^<_{nm\boldsymbol{k}}(t). \tag{4}
$$

The central problem is thus determining the time dependence of $G^<_{nm\boldsymbol{k}}(t)$ under the influence of the external field, which will be allowed to have arbitrary strength and arbitrary temporal profile.

A. Equation of motion for the distribution function

Many-body electronic excitations in response to a time-dependent external field are most systematically handled with GF techniques. The fact that we wish to obtain $\boldsymbol{P}(t)$ in a non-perturbative way implies that our description must properly handle arbitrarily strong fields (and, of course, in experiments, probing the nonlinear susceptibilities does require intense laser fields[1]). But strong fields are bound to drive the statistical system out of thermodynamic equilibrium and, hence, an accurate description of the coherent microscopic processes resulting from such perturbations requires a nonequilibrium GF formalism[44], which has been pioneered by Kadanoff and Baym[47], and by Keldysh[49].

Since details related to the derivation of the equation of motion for the time-diagonal $G^<_{nm\boldsymbol{k}}(t)$ have been discussed, for example, in Ref. 28 or 44, we provide only a qualitative overview of its key aspects and assumptions. In addition to establishing here the context for our numerical calculations in a conceptually self-contained way, this overview will allow us to highlight the necessary adaptations for our approach based on parameterized Hamiltonians and interactions.

Kadanoff and Baym provided a closed set of coupled equations for the time evolution of the different nonequilibrium GFs in terms of self-energies defined on distinct portions of the Keldysh contour which, for practical calculations, must be specified within an approximation scheme[47,48]. A major simplifying step occurs if one approximates $\Sigma^< = 0$ and $\Sigma^r = \Sigma^a \equiv \Sigma$, similarly to what happens in a collisionless and instantaneous scenario like Hartree-Fock[44]. With such approximation, the equation for $G^<_{nm\boldsymbol{k}}(t)$ decouples and reads[28,44]:

$$i\hbar\frac{\partial}{\partial t}\boldsymbol{G}^<_{\boldsymbol{k}}(t) = \Big[\boldsymbol{h_k} + \boldsymbol{U_k}(t) + \boldsymbol{\Sigma_k}[\boldsymbol{G}^<_{\boldsymbol{k}}(t)], \boldsymbol{G}^<_{\boldsymbol{k}}(t)\Big], \quad (5)$$

where $[\boldsymbol{A}, \boldsymbol{B}] \equiv \boldsymbol{AB} - \boldsymbol{BA}$. The electronic system is here defined by the non-interacting Bloch Hamiltonian $\hat{h}$, and $\hat{U}(t)$ is the explicitly time-dependent external field. The total non-interacting Hamiltonian is thus

$$\hat{H}(t) = \hat{h} + \hat{U}(t). \quad (6)$$

The self-energy $\hat{\Sigma}$ encodes all the interaction effects and is a functional of $G^<_{nm\boldsymbol{k}}(t)$. For notational simplicity, we employ bold symbols to denote matrices in the band indices: for example, $G^<_{mn\boldsymbol{k}}(t) = [\boldsymbol{G}^<_{\boldsymbol{k}}(t)]_{mn}$ and $h_{mn\boldsymbol{k}} = [\boldsymbol{h_k}]_{mn} = \langle m\boldsymbol{k}|\,\hat{h}\,|n\boldsymbol{k}\rangle = \delta_{mn}E_{m\boldsymbol{k}}$, where $E_{m\boldsymbol{k}}$ are the Bloch bands.

We now note that the Hamiltonian $\hat{h}$ is meant to be described in terms of a TB parameterization and, in that case, having it reflect the strictly non-interacting Bloch Hamiltonian is not optimal, for two main reasons. On the one hand, irrespective of whether the TB Hamiltonian is obtained from a calculation within DFT or constrained directly by experiments, it will already incorporate electron-electron interactions (in the first case, the TB parameterization reflects at least the Kohn-Sham Hamiltonian, which obviously incorporates interactions at some level). On the other hand, it is desirable that our reference be an accurate TB parameterization of the ground state which, in the case of a semiconductor, and particularly so for 2d materials, requires incorporating the interaction-driven corrections to the quasiparticle dispersion beyond DFT[50]. Therefore, we take $\hat{h}$ and its spectrum, $E_{m\boldsymbol{k}}$, to represent a TB parameterization of the ground state band structure which *already* takes into account such corrections, for example, at the level of the GW approximation[50,51], or as provided by hybrid functional approaches to DFT[52].

The previous considerations in relation to $\hat{h}$ require a corresponding and consistent reassessment of the self-energy term in Eq. (5) to avoid double-counting of interactions. We hence rewrite that equation as

$$i\hbar\frac{\partial}{\partial t}\boldsymbol{G}^<_{\boldsymbol{k}}(t) = \Big[\boldsymbol{H_k}(t) + \boldsymbol{\Sigma_k}[G^<(t)] - \boldsymbol{\Sigma_k}[\tilde{G}^<], \boldsymbol{G}^<_{\boldsymbol{k}}(t)\Big], \quad (7)$$

where $\tilde{G}^<$ represents the GF of the unperturbed system at equilibrium,

$$\tilde{G}^<_{mn\boldsymbol{k}} \equiv G^<_{mn\boldsymbol{k}}(t=0) = i\delta_{mn}f_{m\boldsymbol{k}}, \quad (8)$$

with $f_{m\boldsymbol{k}}$ as the Fermi-Dirac distribution for band m. The effect of the term $-\boldsymbol{\Sigma_k}[\tilde{G}^<]$ is to subtract from the self-energy the quasiparticle corrections of the unperturbed system at equilibrium which, according to the above, should be already included in the TB parameterization for $\hat{h}$. In this way, the self-energy terms describe only the correlation effects induced by the external field. Equation (7) is our counterpart of Eq. (11) proposed by Attaccalite *et al.* in Ref. 28.

Note that the temperature only appears in the time evolution implicitly, via the Fermi-Dirac distribution that defines the initial condition in Eq. (8). Zero and finite temperature calculations are thus on equal footing. Despite this, in the current work we set $T = 0$ since we will benchmark our results against other zero-temperature calculations.

B. Tight-binding parameterizations

We rely on orthogonal TB Hamiltonians in the Slater-Koster formulation[43] to represent $\hat{h}$, where the Bloch eigenstates states, $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$, are expanded in terms of effective local atomic orbitals, $\phi_\alpha(\boldsymbol{r})$, as follows:

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_\alpha C^n_{\alpha\boldsymbol{k}}\chi_{\alpha\boldsymbol{k}}(\boldsymbol{r}),$$

$$\chi_{\alpha\boldsymbol{k}}(\boldsymbol{r}) \equiv \frac{1}{\sqrt{N_c}}\sum_{\boldsymbol{R}}e^{i(\boldsymbol{k}\cdot\boldsymbol{R}+\theta_{\alpha\boldsymbol{k}})}\phi_\alpha(\boldsymbol{r}-\boldsymbol{R}-\boldsymbol{t}_\alpha). \quad (9)$$

The lattice vector $\boldsymbol{R}$ runs over all N_c unit cells of the crystal, n is the band index, α labels different orbitals within the unit cell which are centered at position $\boldsymbol{t}_\alpha$ relative to the cell's origin. Both n and α run over the interval $[1, N]$, where N is the dimension of the orbital basis considered. Although the phase factor $\theta_{\alpha\boldsymbol{k}}$ can be fixed arbitrarily (for example $\theta_{\alpha\boldsymbol{k}} = 0$), one has to consistently carry that choice to the matrix elements of the dipole operator and screened Coulomb interaction (see below). For convenience in the expressions that result from expanding matrix elements in the Bloch basis to the representation (9), we set $\theta_{\alpha\boldsymbol{k}} = \boldsymbol{k}\cdot\boldsymbol{t}_\alpha$ in this work[53].

The specific TB parameterizations used in our calculations for MoS_2 and BN will be discussed further below. The underlying bandstructures are shown in Fig. 1.

C. External field

The interaction with the external radiation field $\boldsymbol{E}(t)$ drives particle-hole excitations in the crystal. We define it in the dipole approximation and length gauge[21] by the one-body perturbation

$$\hat{U}(t) \equiv -e\,\hat{\boldsymbol{r}}\cdot\boldsymbol{E}(t). \quad (10)$$

Its matrix elements in the Bloch basis are $U_{mn\boldsymbol{k}}(t) = -e\,\boldsymbol{r}_{mn\boldsymbol{k}}\cdot\boldsymbol{E}(t)$ and thus require the computation of
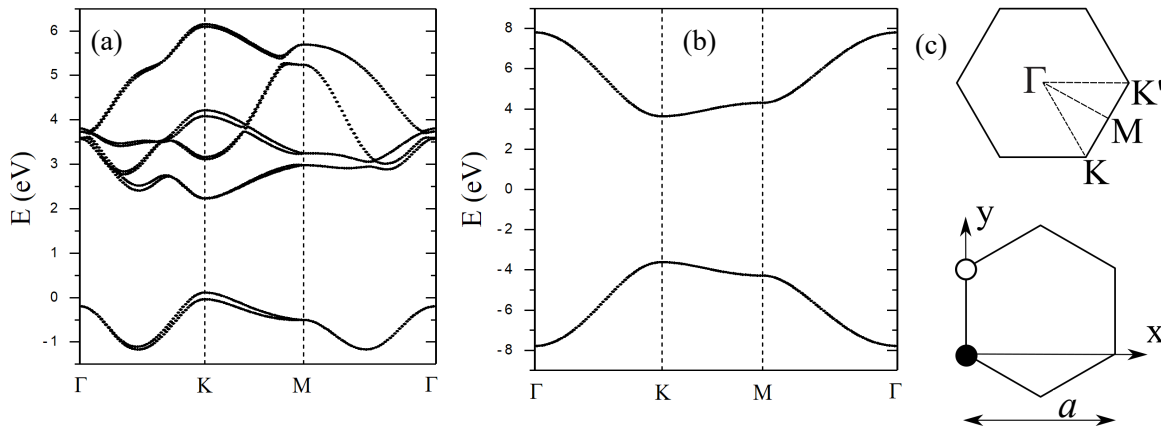
FIG. 1. (a) TB-derived band structure (including SO coupling) of the MoS_2 monolayer according to the Slater-Koster scheme discussed in the text and reported earlier in Ref. 54. (b) TB-derived band structure of hBN (spin-degenerate) according to the parameterization discussed in the text and reported in Ref. 55. (c) Top: Illustration of the Brillouin zone for both MoS_2 and hBN. Bottom: Representation of our orientation of the honeycomb lattice in real space with the positions of the A and B sites highlighted by the filled and empty circles. While A and B represent the positions of N and B atoms in hBN, in the case of MoS_2, we can associate them with a top view of the Mo and S atoms, respectively.

$\boldsymbol{r}_{mn\boldsymbol{k}} \equiv \langle m\boldsymbol{k}| \hat{\boldsymbol{r}} |n\boldsymbol{k}\rangle$. We shall consider only inter-band transitions and neglect all intra-band matrix elements, $\boldsymbol{r}_{mm\boldsymbol{k}}$ (this point is revisited later). In that case, from the definition of the velocity operator $i\hbar\hat{\boldsymbol{v}} = [\hat{\boldsymbol{r}}, \hat{h}]$, we have $\boldsymbol{r}_{mn\boldsymbol{k}} = i\hbar\boldsymbol{v}_{mn\boldsymbol{k}}/(E_{n\boldsymbol{k}} - E_{m\boldsymbol{k}})$. The matrix elements of the velocity can be approximated in the TB representation (9), and recalling the choice $\theta_{\alpha\boldsymbol{k}} = \boldsymbol{k} \cdot \boldsymbol{t}_\alpha$, as[56]

$$\hbar\boldsymbol{v}_{mn\boldsymbol{k}} \simeq \sum_{\alpha\beta} C_{\alpha\boldsymbol{k}}^{m*} C_{\beta\boldsymbol{k}}^{n} \nabla_{\boldsymbol{k}} \langle \chi_{\alpha\boldsymbol{k}}| \hat{h} |\chi_{\beta\boldsymbol{k}}\rangle, \qquad (11)$$

where $\langle \chi_{\alpha\boldsymbol{k}}| \hat{h} |\chi_{\beta\boldsymbol{k}}\rangle$ are the matrix elements of the TB Hamiltonian in the reduced Bloch representation. In a Slater-Koster framework, their $\boldsymbol{k}$-dependence is explicitly known and, consequently, the $\boldsymbol{k}$-derivative appearing in Eq. (11) can be straightforwardly calculated once the TB Hamiltonian is specified.

D. Self-energy approximation

The COHSEX (Coulomb hole and screened exchange) approximation of Hedin[57] has been widely used to describe correlation effects in excited states[18,50,58]. It approximates the electronic self-energy as instantaneous in time and comprising two physical contributions: $\Sigma^{\text{cohsex}} = \Sigma^{\text{coh}} + \Sigma^{\text{sex}}$. The term

$$\Sigma^{\text{sex}}(\boldsymbol{r}, \boldsymbol{r}', t) \equiv iG(\boldsymbol{r}t; \boldsymbol{r}'t^+) \, w(\boldsymbol{r}, \boldsymbol{r}'; \omega = 0) \qquad (12)$$

describes a statically screened exchange interaction, with $w(\boldsymbol{r}, \boldsymbol{r}'; \omega)$ representing the dynamically (frequency-dependent) screened Coulomb repulsion in the random-phase approximation, and

$$\Sigma^{\text{coh}}(\boldsymbol{r}, \boldsymbol{r}', t) \equiv \tfrac{1}{2}\delta(\boldsymbol{r} - \boldsymbol{r}') \left[w(\boldsymbol{r}, \boldsymbol{r}'; \omega = 0) - v(\boldsymbol{r}, \boldsymbol{r}') \right], \qquad (13)$$

where $v(\boldsymbol{r}, \boldsymbol{r}') \equiv e^2/(4\pi\epsilon_0|\boldsymbol{r} - \boldsymbol{r}'|)$ represents the bare Coulomb interaction. The instantaneous approximation to Σ is justified in the present context because the self-energy in Eq. (7) is only operative in the presence of the external field. It therefore defines the strength of the electron-hole interaction but not the Kohn-Sham band-structure renormalization which, by construction, should be already included in $\hat{h}$, and is where dynamical screening is crucial[18,50]. Furthermore, an instantaneous self-energy is in line with the current understanding, at the level of the BSE, that it correctly captures the excitonic spectrum of semiconductors because of the excitons' relatively long time-scales in comparison with the dynamical charge oscillations involved in screening[59–61].

Recalling the reasoning above for the subtraction in Eq. (9) of the self-energy calculated at equilibrium, we see that the contribution Σ^{coh}, being time-independent in this approximation, does not contribute to the time evolution of the distribution function defined by Eq. (9). We therefore need only to consider the screened exchange contribution (12). In this regard, we note that the approximation defined above for Σ^{sex} consists in a static GW approximation[18] and it is clear that, if the term $-\Sigma[\tilde{G}^<]$ were not included in (7), one would be doubly correcting the quasiparticle bandstructure renormalization.

In the Bloch representation, our approximation to the self-energy is then

$$\Sigma_{mn\boldsymbol{k}}[G^<(t)] = i \sum_{jl\boldsymbol{k}'} W_{mn\boldsymbol{k}, jl\boldsymbol{k}'} \, G_{jl\boldsymbol{k}'}^<(t), \qquad (14)$$

where we have denoted the matrix elements of the screened exchange interaction as

$$W_{mn\boldsymbol{k},jl\boldsymbol{k}'} \equiv \int d\boldsymbol{r} d\boldsymbol{r}'\, \psi_{m\boldsymbol{k}}^*(\boldsymbol{r})\psi_{l\boldsymbol{k}'}^*(\boldsymbol{r}')$$
$$\times\, w(\boldsymbol{r}-\boldsymbol{r}')\psi_{j\boldsymbol{k}'}(\boldsymbol{r})\psi_{n\boldsymbol{k}}(\boldsymbol{r}'). \quad (15)$$

In the TB basis introduced in (9), these matrix elements read explicitly[11,41,62]

$$W_{mn\boldsymbol{k},jl\boldsymbol{k}'} \equiv \sum_{\boldsymbol{G}} [I_{j\boldsymbol{k}',m\boldsymbol{k}}^{\boldsymbol{G}}]^*\, I_{l\boldsymbol{k}',n\boldsymbol{k}}^{\boldsymbol{G}}\, w(\boldsymbol{k}-\boldsymbol{k}'+\boldsymbol{G}), \quad (16)$$

where $w(\boldsymbol{q})$ represents the zero-frequency limit of the screened Coulomb potential,

$$w(\boldsymbol{q}) \equiv \left(\frac{e^2}{2\epsilon_0\epsilon_d A}\right)\frac{1}{|\boldsymbol{q}|\,(1+\lambda_0|\boldsymbol{q}|)}. \quad (17)$$

This expression corresponds to the regime of small $\boldsymbol{q}$ calculated for a strictly 2d electron gas embedded in three dimensions[8,63,64]. A is the area of the crystal, λ_0 its 2d polarizability[8,65,66], and ϵ_d captures the average dielectric constant of the environment (for us, in practice, $\epsilon_d = (\epsilon_1 + \epsilon_2)/2$ to capture the effect of static, uniform screening due to the top, ϵ_1, and bottom, ϵ_2, media surrounding the target 2d crystal). The parameters ϵ_0 and λ_0 must be given to completely specify the screened interaction (17). The Bloch coherence factors appearing in Eq. (16) read[67]

$$I_{m\boldsymbol{k},m'\boldsymbol{k}'}^{\boldsymbol{G}} \equiv \langle m\boldsymbol{k}|\, e^{i(\boldsymbol{k}-\boldsymbol{k}'-\boldsymbol{G})\cdot\hat{\boldsymbol{r}}}\,|m'\boldsymbol{k}'\rangle \quad (18)$$
$$\simeq \sum_\alpha [C_{\alpha\boldsymbol{k}}^m e^{i\theta_{\alpha\boldsymbol{k}}}]^* [C_{\alpha\boldsymbol{k}'}^{m'} e^{i\theta_{\alpha\boldsymbol{k}'}}] e^{i(\boldsymbol{k}-\boldsymbol{k}'-\boldsymbol{G})\cdot\boldsymbol{t}_\alpha}.$$

The vectors $\boldsymbol{G}$ in the above expressions belong to the reciprocal lattice. However we emphasize that they do not reflect any attempt to include local-field corrections, which would not be warranted in our Slater-Koster TB scheme. The sum over $\boldsymbol{G}$ is needed in Eq. (16) to restore the symmetry in the interaction between an electron with crystal momentum $\boldsymbol{k}$ and another with $\boldsymbol{k}'$: Since $\boldsymbol{k}, \boldsymbol{k}'$ are restricted to the first BZ — but the Fourier components of the Coulomb interaction are not — the interaction should include not only $\boldsymbol{k}$ and $\boldsymbol{k}'$ but all the equivalent pairs of crystal momenta. The decay of $w(\boldsymbol{q})$ can justify retaining only $\boldsymbol{G} = 0$ in most cases. However, in 2d materials the exciton binding energies can be extremely large, implying tightly bound excitons in real space which, consequently, have slowly decaying wavefunctions in reciprocal space. This means that the wavefunctions of an exciton at the K valley and another at K' overlap, which might lead to a significant Coulomb matrix element. In such cases, not including the equivalent valleys beyond the first BZ amounts to introducing an artificial symmetry breaking in the system. The summation over $\boldsymbol{G}$ ensures such symmetry is retained and, in practice, we keep only the fewest $\boldsymbol{G}$ necessary to obtain converged results. While in the calculations for MoS_2 we found that including only $\boldsymbol{G} = 0$

is sufficient, additional vectors were found to be important in the case of hBN, and we will return to this point later.

Finally, we point out that in the proposal to implement this scheme fully *ab initio*[28] — in which $\hat{h}$ represents the Kohn-Sham Hamiltonian obtained as a first step within DFT — the Hartree energy must be updated at every time step as well for consistency. In our formulation, the Hartree term does not play a role under the time evolution. This is motivated by the fact that, in a BSE approach to the two-particle problem at $T = 0$, the Hartree term in the Hamiltonian (6) generates the so-called "exchange" interaction in the BSE[59], whose matrix elements have the form

$$\int d\boldsymbol{r}\, \psi_{c\boldsymbol{k}}^*(\boldsymbol{r})\psi_{v\boldsymbol{k}}(\boldsymbol{r})v(\boldsymbol{r}-\boldsymbol{r}')\psi_{v\boldsymbol{k}'}^*(\boldsymbol{r}')\psi_{c\boldsymbol{k}'}(\boldsymbol{r}'), \quad (19)$$

where the labels c, v stand for the conduction and valence bands. When written in Fourier space, and neglecting local-field corrections, this expression reduces to $v(\boldsymbol{q} = 0)\delta_{cv}\delta_{c'v'} = 0$ (because $\delta_{vc} = 0$). In other words, the Hartree contribution drops from the BSE thus being irrelevant for excitonic effects (this is also the reason why, even when one does take local-field effects into account, the Hartree contribution tends to be much smaller than that arising from the screened exchange interaction). By definition, an orthogonal TB expansion of the Bloch states such as Eq. (9) ignores local-field corrections and, therefore, in our formulation the Hartree contribution remains implicitly as part of $\hat{h}$ without dynamical updates.

We wish to emphasize two important aspects of this strategy to tackle the interaction effects. The first is that, by formulating the problem in the form of Eq. (7) where the self-energy dictates all interaction effects, a multitude of extensions to the current approximations is straightforward since it requires only the specification of additional terms in the self-energy, but does not require an overhaul of the implementation. Hence, this formulation is intrinsically versatile and adaptable. (The most interesting extensions would arguably be to include the influence of coupling to other degrees of freedom, such as phonons, or specific models of disorder as physical sources of broadening.) The second aspect relates to the specific approximation of our self-energy in Eq. (14): Ref. 28 shows in detail that, in linear order on the external field, this approximation is equivalent to a combined G_0W_0+BSE approach, which is the state-of-the-art combination to reliably describe excitonic effects in semiconductors[18,59]. It is therefore the most promising basis to describe interaction effects in optical response beyond linear order.

E. Fourier analysis of the optical response

In order to characterize the response in the frequency domain, we straightforwardly compute the discrete Fourier transform (FT) of the time-domain polar-

izability and electric fields. We define the discrete FT of a time-dependent signal $f(t)$ that is sampled at every constant interval τ as

$$F_{\omega_k} \equiv \frac{1}{L} \sum_{n=0}^{L-1} f(t_n) e^{i\omega_k t_n}, \qquad (20)$$

where $t_n \equiv n\tau$, L is the total number of time samples, $T \equiv L\tau$ is the total duration of the signal, and $\omega_k \equiv 2\pi k/T$ for $k \in \{0, 1, ...L-1\}$. It is obvious that the maximum frequency resolution of this procedure is $2\pi/T$ and, consequently, the total duration of the signal should in principle satisfy $T \gtrsim 2\pi/\gamma$ so that its Fourier spectrum has at least the resolution imposed by the characteristic broadening of the system, γ [cf. Eq. (28)]. This is one of the two compromises that ultimately determine the duration of the calculations in practice.

The other compromise is the choice of the time step, Δt, chosen to numerically integrate the equation of motion. Suppose, for example, that we wish to compute the response to a sinusoidal light field with a typical frequency of $\hbar\omega_0 \sim 1\,\text{eV}$ (242 THz): It should be clear that, because of the oscillatory nature of the solution, in order to reliably obtain the oscillatory $P(t)$ in the system without accumulating numerical error one must set Δt to a fraction of the fundamental period of the driving field. If, for definiteness, one assumes 10 Runge-Kutta steps per such fundamental period, we have $\Delta t = 2\pi/(10\omega_0) \sim 0.4\,\text{fs}$. If we now seek a frequency resolution of, say, 10 meV, we must also have $2\pi\hbar/T \lesssim 10\,\text{meV}$, or $T \gtrsim 1000\Delta t$. This means that the right-hand side of (7) must be evaluated at least 1000 times for this set of rather common and reasonable requirements. Of course, to ensure resolution of higher harmonics up to order n of the fundamental frequency, one must replace $\omega_0 \to n\omega_0$ in these estimates, whereby the numerical effort is seen to increase by a factor of n.

The choice of the time step has also the fundamental constraint imposed by the Nyquist-Shannon theorem: Since the Δt used in the numerical integration defines the smallest possible sampling interval ($\min \tau = \Delta t$), the theorem imposes the maximum frequency captured in a discrete FT to be π/τ and, consequently, one must ensure $\Delta t \leq \tau < \pi/\omega_{\max}$. As the energy scales of interest typically span several eV, Δt must typically be well within the sub-fs range ($\hbar\pi/1\,\text{fs} \simeq 2.1\,\text{eV}$) to allow a clean Fourier analysis (without aliasing, for example).

These compromises can make the numerical integration time-consuming (see also Section II G below). In contrast, the computation of the FTs appears "instantaneous" when compared with the total time spent integrating the polarization up to $t = T$. For this reason, our discrete FTs have been computed by sampling with $\tau = \Delta t$ to maximize the amount of information, and do not require any optimization beyond the prescription in Eq. (20).

Another advantage of computing the discrete FT as prescribed above using all the natural time steps from the numerical integration of Eq. (7) is that we can avoid the phenomenon of frequency leaking and ensure we always obtain an exact representation of the continuous FT of the signal $f(t)$ whenever it consists of a series of discrete frequencies (like in a monochromatic wave). In order to see this, we recall a simple result from Fourier analysis and signal processing. Let the continuous FT of a signal $f(t)$ be defined by

$$F(\omega) \equiv \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t) e^{i\omega t} dt. \qquad (21)$$

$F(\omega)$ is related to F_{ω_k} defined in Eq. (20) through

$$F_{\omega_k} = \frac{2\pi}{T} \sum_{n=-\infty}^{+\infty} \left[F(\omega - n\omega_s) * \tilde{\delta}_{T/2}(\omega) \right]_{\omega=\omega_k}, \qquad (22)$$

where $*$ represents the convolution operation, $\omega_s \equiv 2\pi/\tau$ is the angular sampling rate, and

$$\tilde{\delta}_{T/2}(\omega) \equiv \frac{Te^{i\omega T/2}}{2\pi} \operatorname{sinc}\left(\frac{\omega T}{2}\right) \qquad (23)$$

is the FT of the rectangle function defined as unity for $0 \leq t \leq T$ and zero otherwise. Now, if the signal is monochromatic with a frequency $\omega_0 < \omega_s$, we have $F(\omega) = F_0\, \delta(\omega - \omega_0)$ and, from Eq. (22), it follows that

$$F_{\omega_k} = \frac{2\pi}{T} F_0\, \tilde{\delta}_{T/2}(\omega_k - \omega_0). \qquad (24)$$

This means that, by adapting the sampling rate ω_s (i.e., by chosing τ) to ω_0 (or vice-versa) such that ω_0 is one of the frequencies $\{\omega_k\}$, we ensure that $F_{\omega_k=\omega_0} = F_0$ exactly, and $F_{\omega_k} = 0$ for all other discrete frequencies, without incurring any frequency leaking (i.e., the discrete FT recovers the *exact* Fourier spectrum of the signal). This can be an important consideration when extracting high-harmonic Fourier components of a system's response, which are typically orders of magnitude smaller than the fundamental and, therefore, benefit from the exclusion of spurious effects such as the frequency leaking that is inevitable whenever ω_s is not matched to the relevant frequencies.

Fourier analysis of $P(t)$ will be used to obtain the nonlinear susceptibilities $\chi^{(n)}$ defined in Eq. (1). In the frequency domain, that relation reads

$$\epsilon_0^{-1} P(\omega) = \sum_{n=1}^{\infty} \int d\omega_1 \cdots \int d\omega_n\, \chi^{(n)}(\omega_1, \ldots, \omega_n)$$
$$E(\omega_1) \cdots E(\omega_n)\, \delta(\omega - \textstyle\sum_n \omega_n). \quad (25)$$

By integrating Eq. (7) in the presence of an external monochromatic field of frequency ω_0, and using the above-described approach to the discrete Fourier analysis, we can extract all the n-th harmonic susceptibilities at once by computing

$$\chi^{(n)}(\omega_0) = \frac{P_{\omega_k=n\omega_0}}{\epsilon_0 (E_{\omega_k=\omega_0})^n}, \qquad (26)$$

or the conductivities: $\sigma^{(n)}(\omega_0) = -i\omega_0 \epsilon_0 \chi^{(n)}(\omega_0)$.

F. Relaxation and broadening

The electronic system described by the equation of motion (7) accumulates all the energy transferred by the external radiation field (the self-energy is Hermitian). A fast laser pulse is not numerically problematic (as we shall see below in Fig. 2) and, for the purposes of comparison with experiments, one can incorporate a phenomenological energy broadening into the Fourier spectrum by a simple modification of Eq. (20):

$$P_{\omega_k} = \frac{1}{L} \sum_{n=0}^{L-1} P(t_n) e^{i(\omega_k + i\gamma)t_n}, \qquad (27)$$

where $\hbar\gamma$ defines the energy resolution (in our calculations we explicitly set it to the half-width at half-maximum of the lowest exciton peak that appears in the spectra used as reference to benchmark our results).

A scenario of continuous excitation without damping causes the amplitude of $\boldsymbol{P}(t)$ in the system to grow in time with a linear envelope, which introduces artifacts as $\omega \to 0$ in the numerical FT. We found it desirable, for these cases, to introduce a phenomenological relaxation mechanism directly into the equation of motion (7) which then becomes

$$i\hbar \frac{\partial}{\partial t} \boldsymbol{G}_{\boldsymbol{k}}^{<}(t) = \Big[\boldsymbol{H}_{\boldsymbol{k}}(t) + \boldsymbol{\Sigma}_{\boldsymbol{k}}[G^{<}(t)] - \boldsymbol{\Sigma}_{\boldsymbol{k}}[\tilde{G}^{<}], \, \boldsymbol{G}_{\boldsymbol{k}}^{<}(t) \Big]$$
$$- i\hbar\gamma \big(\boldsymbol{G}_{\boldsymbol{k}}^{<}(t) - \tilde{\boldsymbol{G}}_{\boldsymbol{k}}^{<} \big), \qquad (28)$$

where the new term on the right-hand side promotes the return of the distribution function to equilibrium, and the broadening parameter γ is set to the target energy resolution as described above.

Of course, both approaches are employed here as a phenomenological strategy to control the energy broadening of the final results, which is sufficient for the current purposes of this paper. But, as pointed out earlier, more sophisticated and microscopically motivated relaxation processes may be incorporated directly into the equation of motion as extensions of Eq. (28).

G. Implementation Notes

Numerical integration — After specifying the time dependence of the external field, we compute the time-dependent polarization according to (4). The set of coupled equations of motion represented by Eq. (7) is integrated numerically using the second-order Runge-Kutta algorithm provided within the GSL library[68]. The choice of second-order is here a compromise between accuracy and expediency, since we wish to maintain the number of intermediate evaluations of the right-hand side of (7) as small as possible per time-step[69].

Problem dimension — In this regard, it is instructive to recall that the *linear* dimension, N_{tot} of the matrices in Eq. (7) is defined by the total number of bands (N_v valence and N_c conduction) *plus* the total number of $\boldsymbol{k}$ points that need to be sampled in the BZ for the desired resolution and convergence. Therefore, $N_{tot} = N_k^2(N_c + N_v)^2$, where N_k^2 represents the total number of $\boldsymbol{k}$ points (N_k along each reciprocal direction). Since the Coulomb matrix elements (16) are non-diagonal in $\boldsymbol{k}$, their storage is the most costly since it requires $\sim N_{tot}^2/2$ entries in memory. Combining this with the fact that these matrices are to be multiplied several times for each time step of the Runge-Kutta integration leads to a numerical problem that quickly becomes very challenging memory- and time-wise, even when considering the modest/minimal requirements of, say, $N_k = 32$ and $N_c + N_v = 2$.

Symmetry — We will focus on 2d crystals with three-fold symmetry, having point-symmetry group D_{3h} or higher. This already covers the materials that are currently most actively studied, such as graphene and its derivatives, hBN, transition-metal dichalcogenides, silicene, stanene, germanene, and several others. This restriction is a practical, not fundamental one, and is adopted here because their linear and second-harmonic (SH) susceptibility tensors are completely specified by computing only the diagonal component along y^1 because, for our choice of lattice orientation in Fig. 1(c), symmetry imposes $\chi_{xx}^{(1)} = \chi_{yy}^{(1)}$ and $-\chi_{yyy}^{(2)} = \chi_{xxy}^{(2)} = \chi_{xyx}^{(2)} = \chi_{yxx}^{(2)}$. From this point onwards we thus drop the Cartesian indices: $\boldsymbol{P} \to P(t) \equiv P_y(t)$, $\boldsymbol{E} \to E(t) \equiv E_y(t)$, and $\chi_{\lambda\alpha\cdots}^{(n)} \to \chi^{(n)} \equiv \chi_{yy\cdots}^{(n)}$.

Nonlinear processes — We will address only the n-th harmonic susceptibilities in this paper, which are a function of only one frequency. Hence, we will also adopt a simplified notation $\chi^{(n)}(\omega_1, \ldots, \omega_n) \to \chi^{(n)}(\omega)$ throughout this paper, where the single frequency ω is that which characterizes the driving field (*i.e.*, the fundamental frequency when the field is monochromatic).

Windowing — We found that applying a window function to our numerical time series for $P(t)$ considerably reduces the effect of the transient background due to the field switch-on, which is an important consideration when resolving the nonlinear contributions (to be discussed below). Windowing consists in multiplying the original signal by a so-called window function, $w(t)$, which is chosen to provide a desired redistribution of its Fourier spectrum. This technique is frequently used to minimize frequency leaking when computing discrete FT, as well as other frequency filtering applications[70]. The Fourier analysis done in the context of Section III D has been performed after applying a Hann window to the time-dependent polarization. In the notation of the previous section, this means replacing the time-dependent signal $f(t_n) \longrightarrow f(t_n)\, w(t_n)$, where the Hann window function is defined as

$$w(t_n) \equiv 2 \sin^2\left(\frac{n\pi}{L-1}\right), \quad 0 \leq n \leq L-1. \qquad (29)$$

In this formulation, the window function defines the envelope of the total time series $f(t_n)$. The Hann window

does not introduce any additional parameter.

III.   ILLUSTRATION FOR MoS$_2$ AND hBN

We illustrate the potential of this approach based on parameterized models with the cases of MoS$_2$ and hBN monolayers, each having been chosen for different specific reasons. MoS$_2$ is arguably the most widely studied representative of the family of 2d TMD semiconductors. It is now known that, except in the energy region of the bound A/B exciton series, the accurate description of the optical excitations across this family of compounds requires consideration of at least 6 conduction bands (3×2 for spin); this is due to the fact that the so-called C excitons involve contributions from the bands dispersing along Γ-K and Γ-M[11,34,71,72] [cf. Fig. 1(a)]. Since the spin-orbit-induced splitting of these bands is crucial for many of the unusual features in these materials, such as spin-valley locking[73–76], a minimal model to describe the optical properties up to the energies of the C excitons requires in principle $2 \times (3+1) = 8$ bands to cover an energy span of $\sim 3\,\mathrm{eV}$[11,41]. MoS$_2$ is then chosen as a representative of a system with a relatively demanding TB parameterization, and an example of how this approach can yield extremely good quantitative agreement with experiments and *ab initio* calculations. Our model for BN was deliberately selected to specifically analyze the opposite extreme of having only 2 bands in the problem, since it will provide further insight into the much-discussed role of the inter- and intra-band matrix elements of the dipole operator [cf. Section II C], which we will address later.

A.   Parameterization of Hamiltonians and interactions

To describe the quasi-particle-corrected electronic structure of MoS$_2$ [*i.e.*, the Hamiltonian $\hat{h}$ in Eq. (7)] we consider the orthogonal Slater-Koster Hamiltonian proposed in Ref. 54, which has been already demonstrated to capture extremely well the experimental optical absorption spectrum in a direct solution of the BSE[11]. It is built from a basis of 11 atomic orbitals that comprises the three p valence orbitals in each S and the five d orbitals of Mo. Since the spin-orbit coupling must necessarily be included to properly describe the splitting of the bands near the optical gap, the total dimension of the basis is then $N = 22$. Additional details of this TB model are described elsewhere[54]. The associated band structure is reproduced in Fig. 1(a), reflecting the insulating ground state of a pristine monolayer with a direct gap at the K/K' points. In order to directly compare our susceptibilities with experiments, a rigid blue-shift in the energies by $+0.07\,\mathrm{eV}$ has been incorporated in all the results shown below, in line with the procedure originally discussed in Ref. 11.

For hBN we resort to the orthogonal TB Hamiltonian parameterization proposed by Galvani *et al.*[55] as it provides a good description of the *GW*-corrected *ab initio* band structure around the fundamental gap. It is a simple two-band (spin-degenerate) Hamiltonian which, in the notation introduced in Eq. (11), reads

$$\langle \chi_{\alpha \boldsymbol{k}} | \hat{h} | \chi_{\beta \boldsymbol{k}} \rangle \mapsto \begin{bmatrix} E_b & -t\,\varphi(\boldsymbol{k})^* \\ -t\,\varphi(\boldsymbol{k}) & E_n \end{bmatrix}. \qquad (30)$$

where $E_b = 3.625\,\mathrm{eV}$ represents the on-site energy at the boron atom, $E_n = -3.625\,\mathrm{eV}$ that at the nitrogen, $t = 2.30\,\mathrm{eV}$ is the hopping integral, $\varphi(\boldsymbol{k}) \equiv e^{i\boldsymbol{k} \cdot \boldsymbol{\delta}_1} + e^{i\boldsymbol{k} \cdot \boldsymbol{\delta}_2} + e^{i\boldsymbol{k} \cdot \boldsymbol{\delta}_3}$ with vectors $\boldsymbol{\delta}_1 = \frac{a}{\sqrt{3}}(\frac{\sqrt{3}}{2}, -\frac{1}{2})$, $\boldsymbol{\delta}_2 = \frac{a}{\sqrt{3}}(0, 1)$, $\boldsymbol{\delta}_3 = \frac{a}{\sqrt{3}}(-\frac{\sqrt{3}}{2}, -\frac{1}{2})$, and $a \simeq 2.5\,\text{Å}$ is the hBN lattice constant. The associated band structure is reproduced in Fig. 1(b). We note that this parameterization is accurate in the vicinity of the fundamental gap, but does not faithfully reflect the actual dispersion of the two lowest-energy bands in hBN over the entire BZ, especially near the Γ point[55,77]. This limits the range of validity of our TB parameterization to particle-hole excitations with less than ~ 8–9 eV. In both materials, $\hat{h}$ is diagonalized in a uniform grid with $N_k{}^2$ points on the first BZ depicted in Fig. 1(c).

For the purposes of benchmarking our calculation, the screened Coulomb interaction (17) is parameterized in different scenarios for each material: for MoS$_2$ we chose the environment's dielectric constant as $\epsilon_d = 2.5$, appropriate for the air/silica interface ($\epsilon_1 = 1$, $\epsilon_2 = 4$), and set the polarizability parameter $r_0 = 13.55\,\text{Å}$, which is know to produce good agreement with the measured exciton binding energies[11,78]. In the case of hBN, we used $r_0 = 10.00\,\text{Å}$ as suggested in the literature based on *ab initio* results[55], while $\epsilon_d = 1$ so that we can directly compare our results with existing calculations for a free-standing hBN monolayer in vacuum. Finally, while for MoS$_2$ we found that considering only $\boldsymbol{G} = 0$ in the expressions for the Coulomb matrix element (16) is sufficient, the case of hBN required the inclusion of at least 16 reciprocal vectors to recover the correct symmetry and degeneracy of the lowest excitonic states.

B.   Excitons in the linear and nonlinear optical response of MoS$_2$ and hBN

Prior to discussing our specific calculations, we briefly overview the broad features and recent approaches to calculating the impact of electronic interactions (excitons) in the optical response of our two target materials, especially at the level of nonlinear effects. MoS$_2$ is the TMD whose interaction-related optical properties have been most comprehensively characterized experimentally and theoretically[76]. Its conduction and valence band extrema are both located at the two nonequivalent K/K' points of the hexagonal BZ. While multilayers have an indirect band gap, the gap is direct at those K points for the monolayer[81,82]. A strong spin-orbit coupling splits the
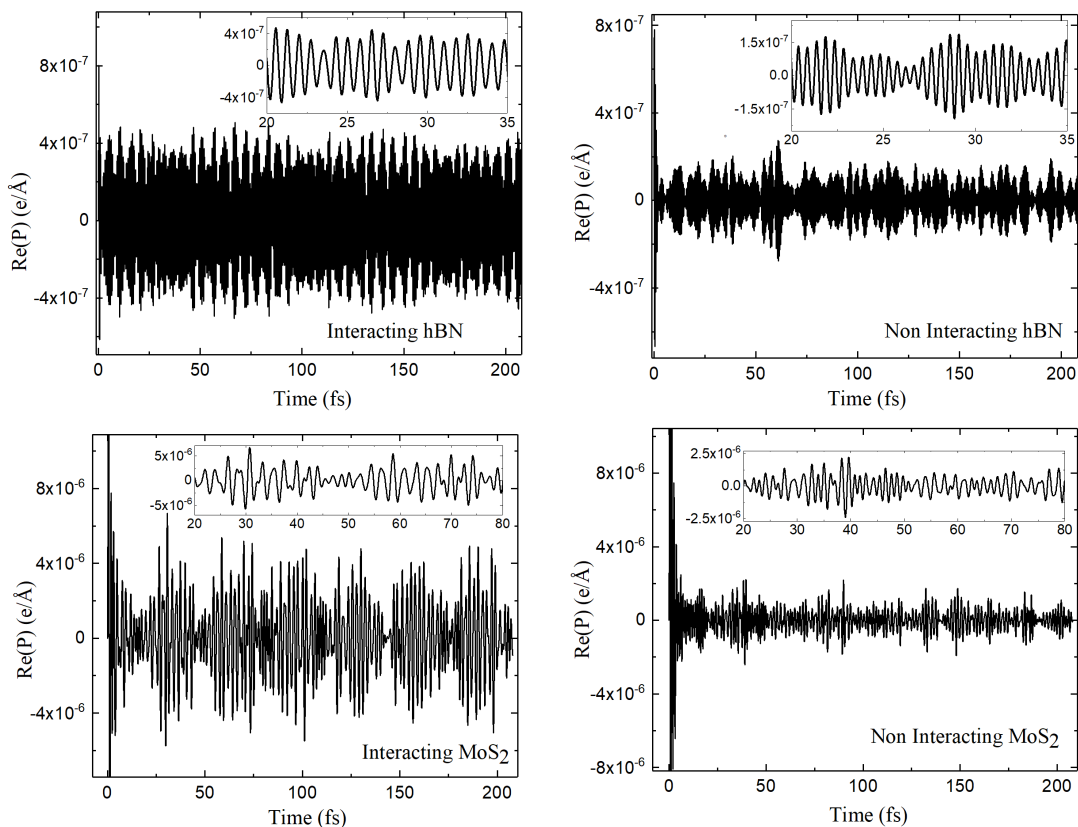
FIG. 2. Time-dependent polarization calculated in response to a sub-fs optical pulse for a monolayer of hBN (top row) and MoS_2 (bottom). For comparison, we show results with (left) and without (right) the effect of Coulomb interactions. [for hBN: $t_p = 0.15\,\text{fs}$, $N_c = N_v = 1$, $N_k^2 = 36^2$, $\tau = 0.0125\,\text{fs}$, $E_0 = 0.1\,\text{mV/Å}$; for MoS_2: $t_p = 0.3\,\text{fs}$, $N_c = 6$, $N_v = 2$, $N_k^2 = 36^2$, $\tau = 0.05\,\text{fs}$, $E_0 = 0.1\,\text{mV/Å}$]

valence bands near K, generating two families of bound excitonic states traditionally labeled A and B[34,76,79,83], and contributed primarily by the metal d orbitals. The two A and B exciton peaks $(E_A < E_B)$ define the optical absorption threshold at $E_A = 1.8 \pm 0.1\,\text{eV}$ for a monolayer deposited on silica, as can be seen in Fig. 3(a) below. At higher energies the absorption spectrum is dominated by the so-called C and D broad resonances that involve significant contributions from the chalcogen orbitals[11,34,71,72].

In contrast to the abundance of scrutiny of its linear response, there have been few experimental reports of the second and higher harmonic susceptibilities of MoS_2 over extended energy ranges. Examples are Refs. 84 and 85 that report the SH emission of both monolayer and trilayer in the region of the C resonance, and Ref. 86 which reports similar measurements over the range 0.9–1.6 eV, but for multilayers. SH calculations that include excitonic effects have been performed *ab initio* in Ref. 30, and by Pedersen *et al.*[41] who used a perturbative formulation based on the solution of the BSE with a parameterized TB Hamiltonian and interaction analogous to the ones we introduced above. SH susceptibilities calculated with further simplified effective band models, applicable

only to the region of the A and B peaks, have also been recently reported[87,88].

In relation to hBN, the p_z orbitals on B and N define the highest valence and lowest conduction bands which are separated by a large direct gap at the K/K' points [see Fig. 1(b)]. Experimentally, the optical gap is seen at 5.8 eV in bulk hBN[89,90] while values between 5.6 eV and 6.0 eV have been reported by optical absorption measurements for monolayers on quartz[91–93]. Various aspects of its excitonic characteristics and their impact in the optical response have been studied by DFT+GW+BSE *ab initio* methods[28,30,55,77,94–102], and Ref. 55 has provided, in addition, a real-space Wannier approximation to reduce the BSE to an effective exciton TB model. Pursuing a two-band model similar to that in Eq. (30) and a screened interaction, Pedersen has solved the BSE equation and computed the SH susceptibility using an equilibrium second-order perturbative framework, and emphasized the importance of intra-band matrix elements of the dipole operator[21] in a length-gauge formulation of the coupling to the light field.
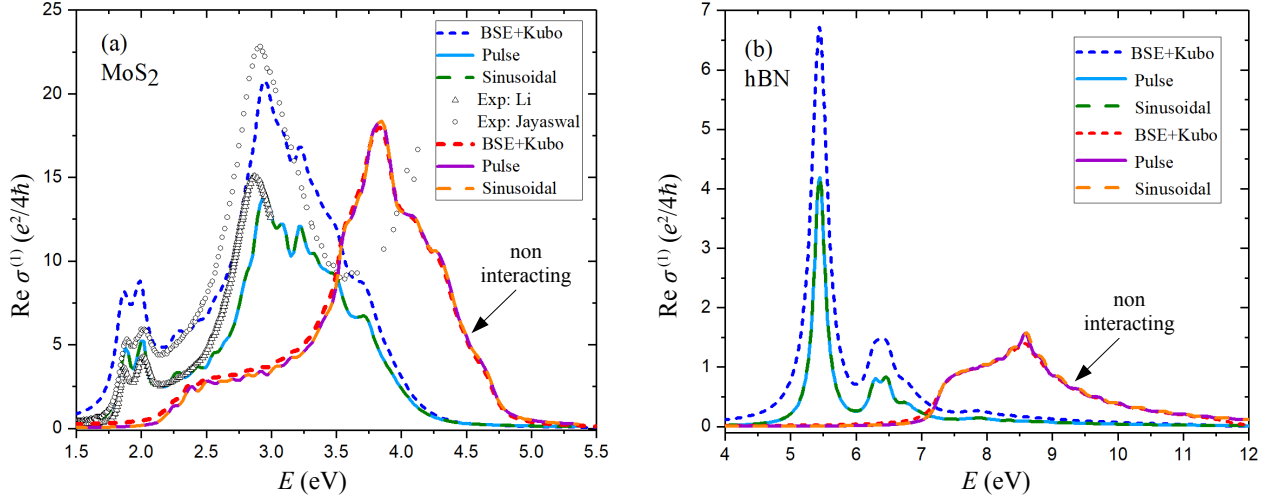
FIG. 3. (a) Linear optical conductivity of MoS$_2$. The two curves represented by points are experimental, at room temperature[79,80]. The lines labeled "pulse" refer to the optical conductivity obtained from Fourier analysis of the time-domain polarizations shown in Figs. 2(c) and 2(d). The "BSE+Kubo" traces were obtained from the linear Kubo formula following the diagonalization of the BSE for the same Hamiltonians. The "sinusoidal" curves were calculated from the time-domain response to continuous monochromatic fields as described in the text. The broadening in our calculations was set to match that of the first exciton peak in the experimental traces. (b) Linear optical conductivity of hBN with the same labeling convention as in (a). [for MoS$_2$: $t_p = 0.3$ fs, $N_c = 6$, $N_v = 2$, $N_k^2 = 36^2$, $\tau = 0.05$ fs, $E_0 = 0.1$ mV/Å, $\hbar\gamma = 0.05$ eV; for hBN: $t_p = 0.15$ fs, $N_c = N_v = 1$, $N_k^2 = 36^2$, $\tau = 0.0125$ fs, $E_0 = 0.1$ mV/Å, $\hbar\gamma = 0.1$ eV]

C. Linear response to an optical pulse

The simplest perturbing field is that of a quasi-instantaneous pulse, which is particularly suitable to extract the linear response in a one-shot integration of Eq. (7). This is easily seen if one strictly sets $E(t) = E_0\delta(t)$ in Eq. (25) and ensures that E_0 is small; in lowest order

$$P(\omega) \simeq \frac{E_0}{2\pi\epsilon_0}\chi^{(1)}(\omega). \tag{31}$$

Therefore, since an instantaneous pulse excites the system equally at all frequencies, the Fourier transform of $P(t)$ computed as the response to a single instantaneous pulse directly yields the linear susceptibility at all frequencies. We followed this approach to demonstrate that the results obtained by integrating the equation of motion (7) reproduce the linear susceptibility computed from direct diagonalization of the BSE combined with the Kubo formula. Note that, in general, once the linear response is established, the nonlinear susceptibilities are immediately defined as well since they are obtained from the same time-dependent polarizability of the system, as indicated in Eq. (26). This is particularly true with regards to the absolute magnitudes of the high-order susceptibilities because, by ensuring that the linear susceptibility is quantitatively accurate, we can subsequently rely on the predicted magnitude of the higher-order components.

To appreciate the details of the implementation, it is instructive to walk through some of its key aspects and intermediate results, which we will now do in relation to the response of a short-duration pulse. For the numerical implementation, we shaped the pulse as

$$E(t) = E_0(t_p - t)t/t_p^2, \quad 0 \le t \le t_p, \tag{32}$$

and $E(t) = 0$ beyond t_p. The amplitude was kept at $E_0 = 10^{-4}$ V/Å and we verified that nonlinear effects are absent up to $E_0 \sim 1$–10 V/Å, which is consistent with the expectation that, in general, nonlinear effects emerge when the field strength approaches the magnitude characteristic of atomic electric fields[1]: $E_{\mathrm{at}} = e/(4\pi\epsilon_0 a_0^2) \simeq 51$ V/Å.

Figure 2 shows the temporal profile of the induced polarization which has been integrated up to times in excess of 200 fs after the system was excited (the pulse lasted $t_p = 0.15$ fs for hBN and $t_p = 0.3$ fs for MoS$_2$). The main panels show the total time series for $P(t)$ in both systems with and without the effect of the screened Coulomb self energy in the calculation. It is visible that $P(t)$ remains finite and undamped, reflecting the fact that this calculation was carried our without any damping during the time evolution [*i.e.*, $\gamma = 0$ in Eq. (28)]. Even without a detailed Fourier analysis, the time-domain picture reveals physically consistent signatures of the system's expected behavior: for example, we can identify by direct inspection an average period of ~ 0.48 fs for hBN and ~ 1.0–1.4 fs for MoS$_2$ in the non-interacting polar-

izations [see insets of Figs. 2(b) and (d)]. These translate into characteristic energies of $\sim 8.6\,\mathrm{eV}$ and ~ 3–$4\,\mathrm{eV}$, respectively, which coincide with the energies at which the non-interacting absorption spectrum is maximal in each case (cf. Fig. 3 below).

From a direct Fourier analysis of the $P(t)$ traces, we obtained the optical conductivities labeled "pulse" in Fig. 3. The broadening was introduced as per Eq. (27) where we used the experimental width $\hbar\gamma \simeq 0.07\,\mathrm{eV}$ for MoS$_2$ and $\hbar\gamma = 0.1\,\mathrm{eV}$ for BN. The plot in Fig. 3(a) pertains to MoS$_2$ and exhibits two types of comparison: The two traces represented by points were extracted from the experimental reports in Refs. 79 and 80 for monolayers on SiO$_2$ substrates; one sees that our result reflected in the line labeled "pulse" describes all the experimental features very well, in particular the energies and spectral weight in the entire range of energies captured by our TB Hamiltonian ($\hbar\omega \lesssim 3.5\,\mathrm{eV}$). The trace labeled "perturbative" has been computed by a direct application of the linear Kubo formula to the spectrum of the BSE (for exactly the same TB Hamiltonian and parameterized interaction used in the time-domain calculation) as described earlier in Ref. 11. Finally, for reference and to further reinforce the substantial restructuring of the absorption spectrum brought about by the Coulomb interaction, the plot includes the conductivities obtained without the Coulomb self-energy. In quantitative terms, from the "impulse" curves we extract a quasiparticle band gap of $2.18\,\mathrm{eV}$ and the lowest A/B excitons at $E_A = 1.87\,\mathrm{eV}$ and $E_B = 2.00\,\mathrm{eV}$ (binding energies $E_A^b = 0.32\,\mathrm{eV}$ and $E_B^b = 0.34\,\mathrm{eV}$, respectively). This tallies well with results from angle-resolved photoemission spectroscopy[78], as well as X-ray photoemission and scanning tunneling spectroscopy[103], that place the band gap within 2.15–$2.35\,\mathrm{eV}$ thus leading to binding energies of the lowest A exciton in the range 0.22-$0.42\,\mathrm{eV}$.

The plots shown in Fig. 3(b) reflect the corresponding calculations applied to the case of hBN. For reference, our explicit diagonalization of the BSE yields the lowest excitonic levels with the following energies and degeneracies (in eV): $\{5.44\,(\times 2),\ 5.96\,(\times 1),\ 6.27\,(\times 2),\ 6.28\,(\times 1),\ 2.45\,(\times 2)\}$. As the quasiparticle band gap defined by our parameterization (30) is $7.25\,\mathrm{eV}$, the binding energies of these excitonic levels are, respectively, $\{1.81, 1.29, 0.98, 0.97, 0.80\}$. These figures compare reasonably well with the ones reported in Refs. 55 and 95 using DFT+GW+BSE. Similarly to our results for MoS$_2$, the optical conductivity obtained for hBN in the time-domain framework reproduces the perturbative result, as expected, and thus validates our implementation of the former approach and the approximations involved. A characteristic of hBN is that the optical spectral weight is almost entirely concentrated at the exciton peaks, and is strongest at the lowest bright exciton. Indeed, the frequency dependence seen in Fig. 3(b) reproduces extremely well the quantitative and qualitative features of the absorption spectrum obtained by several other groups[22,28,55,77,94–96,98,101], and is compatible with the

optical gaps of 5.6–$6.0\,\mathrm{eV}$ reported experimentally[91–93]. (Recall that we parameterized hBN in vacuum and, hence, both the GW quasiparticle renormalization and the exciton binding would have to be adapted for a direct comparison with experiments.) Moreover, the absolute magnitude at the lowest exciton peak is here $\sigma^{(1)}(\omega = 5.4) \simeq 4\,e^2/(4\hbar)$; this converts to an imaginary dielectric constant $\mathrm{Im}\,\varepsilon(\omega) = \mathrm{Re}\,\sigma^{(1)}(\omega)/(d\epsilon_0\omega) \simeq 10$ (using $d = 3.3\,\text{Å}$ for effective thickness of a BN monolayer), which is entirely in line with the magnitude reported for this peak from first principles in Refs.77 and 101, as well as in optical absorption experiments with bulk BN[89].

Common to both materials — and by extension to all 2d semiconducting materials — is the fact that, if one were to ignore the excitonic interaction effects, the predicted absorption spectrum would be patently inaccurate, both because of the large rigid blue-shift of the non-interacting curve in relation to experiments, and because that would entirely miss the excitonic spectral weight that dominates near the absorption threshold.

In the context of this paper, the most significant aspect of the results shown in Fig. 3 is that the time-domain calculation recovers the linear response function obtained perturbatively for the same microscopic parameterization of the system. We are thus in a position to explore the real power of the time-domain framework, which lies in its ability to naturally capture the response to arbitrary time-dependent fields, as well as to describe the nonlinear response in a rather expedite manner.

D. Nonlinear response to monochromatic fields

By definition, the high-harmonic susceptibilities $\chi^{(n)}(\omega_0)$ in Eq. (26) represent the n-th order response of a system to a continuous, monochromatic wave at that frequency. More precisely, under a monochromatic perturbation of frequency ω_0, the quantities $\chi^{(n)}(\omega_0)$ evaluated at the single frequency ω_0 are sufficient to entirely specify the time or frequency dependence of the polarization. This offers a direct way to compute the high-harmonic susceptibilities by sending a light field

$$E(t) = E_0 \sin(\omega_0 t), \qquad 0 \le t \le T, \qquad (33)$$

computing $P(\omega)$, and repeating for as many frequencies ω_0 as desired[29]. Of course, each calculation for a given frequency requires roughly the same duration as that for a quasi-instantaneous pulse which we described above. Therefore, the total time required to map $\chi^{(n)}(\omega)$ over a finite interval of frequencies will be comparatively much larger, in general, if a large number of frequencies is sought (by a factor that is roughly the number of such frequencies). Hence, despite the simplicity involved in extracting each $\chi^{(n)}(\omega_0)$ from a simple Fourier analysis as in Eq. (26), this strategy of sending one wave per frequency is the most time-consuming. A more expedite alternative is to excite the system with a pulse of finite
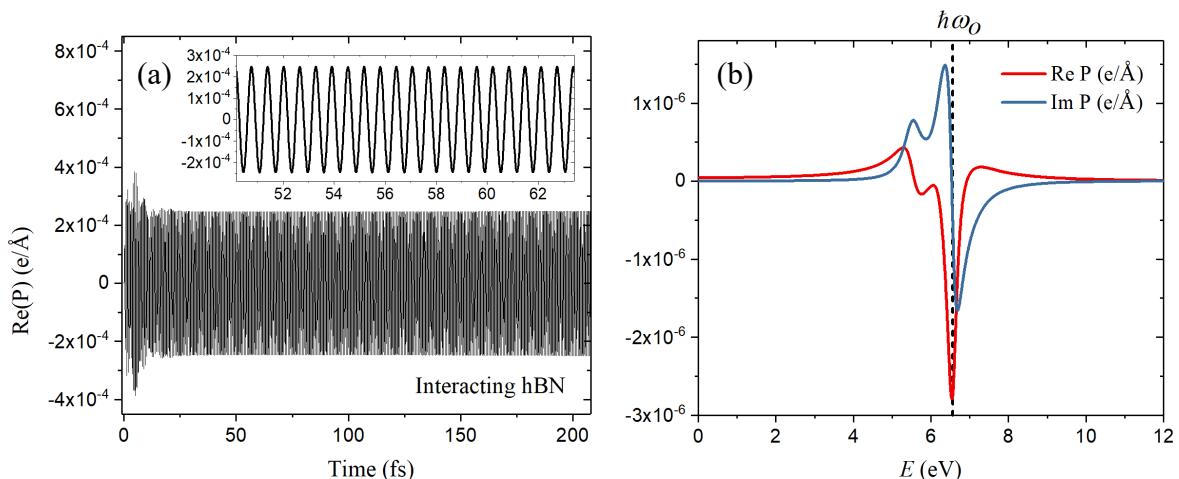
FIG. 4. (a) Typical time-dependent polarization of a hBN monolayer in response to a continuous, monochromatic field with $\hbar\omega_0 = 6.65\,\text{eV}$. (b) The corresponding Fourier spectrum, $P(\omega)$. The dashed line marks the frequency of the external field. $[N_k^2 = 36^2,\ N_c = N_v = 1,\ E_0 = 10^{-2}\,\text{V/Å},\ \tau = 0.0125\,\text{fs},\ \hbar\gamma = 0.1\,\text{eV}]$

duration (with enough bandwidth to span the range of frequencies of interest), followed by an order-by-order deconvolution of the field from the resulting $P(\omega)$, as determined by the relation (25). This route, however, relies on a much more involved post-processing and will not be pursued in the current paper. Given its simplicity, transparency and intuitive value, we shall instead proceed with the one-wave-per-frequency strategy to illustrate typical calculations.

Figure 4(a) shows the calculated $P(t)$ for the hBN monolayer in response to a weak monochromatic field of the type (33) with $\hbar\omega_0 = 6.65\,\text{eV}$ ($\omega_0 = 9.97\times10^{15}\,\text{rad/s}$). As a relaxation mechanism, which is now necessary to dissipate the energy that is constantly being pumped into the system by the continuous wave, we employed the scheme described by the second term in Eq. (28) with $\hbar\gamma = 0.1\,\text{eV}$; other parameters are specified in the figure caption. As expected, $P(t)$ has now the temporal profile of a damped oscillator driven at a frequency ω_0. Its Fourier analysis shown in Fig. 4(b) reveals a corresponding peak in $\text{Re}\,P(\omega)$ at precisely ω_0. By extracting $P(\omega = \omega_0)$ in this way for a number of distinct plane waves, we mapped the frequency-dependence of the linear susceptibility/conductivity and obtained the traces labeled "sinusoidal" in Fig. 3. A direct inspection shows that they exactly follow the ones obtained with the pulse excitation described earlier.

There are important details worth emphasizing at this point in relation to the requirements for the total integration time, T. The *first* consideration is that it clearly must be compatible with the desired energy resolution, say $\hbar\delta\omega$, which means that $T \gtrsim 2\pi/\delta\omega$. The *second* is that the system receives the incoming wave at $t = 0$ on a state of equilibrium and, consequently, in addition to the *driven* response there is also a *transient* response to the

sudden field turn-on that contributes to the polarization: $P(t) = P_{\text{driven}}(t) + P_{\text{trans}}(t)$ (precisely as in the classical driven oscillator where the solution of its equation of motion involves the sum of two such terms). With a damping rate γ, one expects the memory of the field turn-on to fade within a time $\sim 2\pi/\gamma$ and the corresponding decay of the transient component. In principle, one could discard the signal $P(t)$ up to that point in the Fourier analysis to minimize the transient effect[104]. When combined with the energy resolution requirements, this roughly doubles the minimum value of T up to which the equation of motion (28) should be integrated. The *third* consideration is that, in order to extract the nonlinear susceptibilities, one is interested in the frequency spectrum of the asymptotic component $P_{\text{driven}}(t)$, but not in that of $P_{\text{trans}}(t)$. The fact that the latter decays within a time $\sim 2\pi/\gamma$ is satisfactory only with regards to the linear response [meaning that, in practice, setting $T \gtrsim \sim 2\pi/\gamma$ is sufficient to guarantee an accurate result for $\chi^{(1)}(\omega)$ by following the procedure outlined in relation to Eq. (26)]. However, the decay of $P_{\text{trans}}(t)$ might not be sufficient to resolve the nonlinear contributions to the polarization if the field amplitude, E_0, is not strong enough, in such a way that this transient contribution may conceal the higher harmonics. In order to illustrate this point explicitly, we plot in Fig. 5 the absolute value of $P(\omega)$ obtained from $P(t)$ with different durations T. The case $T = 208\,\text{fs}$ corresponds to $2\pi\hbar/T \simeq 0.02\,\text{eV}$ and, according to the above, is in principle adequate to ensure an energy resolution of $0.1\,\text{eV}$ in the derived response functions. However, we can see in this figure that the SH peak at $2\omega_0$ is not resolved until $T \gtrsim 2000\,\text{fs}$ for the field amplitude E_0 used, and the third harmonic remains entirely occluded by the transient background[105]. If this interplay between the total integration time, resolution, and field ampli-

tude is not taken carefully into consideration, one risks entirely erroneous results in the nonlinear susceptibilities. For example, suppose we were to blindly compute $\chi^{(2)}(\omega_0) = P(2\omega_0)/\epsilon_0 E(\omega_0)^2$, as prescribed by Eq. (26), directly from the red trace ($T = 208$ fs) in Fig. 5: Rather than reflecting the actual SH susceptibility of the system, such result would correspond instead to the frequency spectrum of the transient background! In such case, however, an analysis of the field dependence would reveal that $\chi^{(2)}(\omega_0) \propto 1/E_0$, instead of being field-independent. This indicates that, ultimately, any computation of any $\chi^{(n)}(\omega)$ should be tested for field-independence within an adequate range of fields. The *fourth* consideration pertains to the more subtle fact that, rigorously, Eq. (25) is only applicable to non-resonant excitation, since it is a perturbative expansion in the external field. But when one is interested in mapping $\chi^{(n)}(\omega)$ for a given material, one is mostly looking at describing the resonant response, in the sense that ω (or $n\omega$) matches one of the possible excitations of the system. How, then, are we justified in using the expressions (26) if, under resonant excitation, the response of the system is not necessarily described by the series expansion (25)? The answer lies in the finite broadening caused by the relaxation mechanism built into the time evolution [cf. Eq. (28)]: If $P(t)$ is integrated up to $T \gg 2\pi/\gamma$ we gain enough energy resolution to appreciate that all states have an intrinsic lifetime and, therefore, the excitation is never strictly resonant[106]; in these conditions, the relation (25) is justified and a valid means of obtaining the susceptibilities.

Having taken these aspects into consideration, we obtained the converged results shown in Fig. 6 for the second- and third-harmonic susceptibilities of MoS_2 and hBN [for reference, we display the corresponding results without Coulomb interaction in Fig. 9]. In each case, the features we obtained here for the SH susceptibility with explicit account of interactions compare well with other recent calculations. For example, for MoS_2, we obtain SH magnitudes of $|\chi^{(2)}(\hbar\omega = 0.9)| \simeq 0.4\,\mathrm{nm}^2/\mathrm{V}$ (A/B exciton features) and $|\chi^{(2)}(\hbar\omega = 1.5)| \simeq 1.5\,\mathrm{nm}^2/\mathrm{V}$ (C exciton feature), which compare well with the corresponding values $\simeq 0.12$ and $\simeq 1.0$ obtained by Trolle *et al.*[41] using a parameterized TB model and solving the BSE, and with the value $\simeq 0.7\,\mathrm{nm}^2/\mathrm{V}$ (2.6×10^{-6} esu) obtained *ab initio* by Grüning *et al.* at the C-exciton peak[107]. Experimentally, Li *et al.*[108] reported $|\chi^{(2)}(\hbar\omega = 1.53)| = 8.8 \times 10^{-31}\,\mathrm{mC}/\mathrm{V}^2 \simeq 0.1\,\mathrm{nm}^2/\mathrm{V}$, while Woodward *et al.* extracted[109] $|\chi^{(2)}| = 0.02\,\mathrm{nm}^2/\mathrm{V}$ and $|\chi^{(3)}| = 0.17\,\mathrm{nm}^3/\mathrm{V}^2$ from harmonic generation in MoS_2 under a laser field with $\omega_0 \simeq 0.8$ eV (1560 nm). Our results in Fig. 6 for MoS_2 yield $|\chi^{(2)}(\hbar\omega = 0.8)| \simeq 0.24\,\mathrm{nm}^2/\mathrm{V}$ and $|\chi^{(3)}(\hbar\omega = 0.8)| \simeq 0.25\,\mathrm{nm}^3/\mathrm{V}^2$. We consider them to be in reasonable agreement with experiments even though such comparisons are delicate because the magnitudes reported experimentally for high harmonic susceptibilities tend to display large discrepancies[110].

In relation to hBN, the first remark is that, even though our TB model contains only 2 bands and we
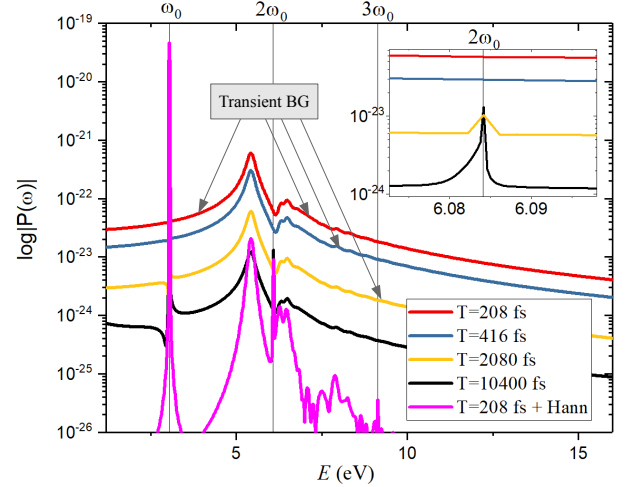

FIG. 5. Dependence of the frequency spectrum of $|P(\omega)|$ on the total integration time (T) for hBN, in response to a continuous monochromatic wave ($\hbar\omega_0 = 3.04$ eV). The spikes at $n\omega_0$ define the amplitude of the n-th harmonic charge oscillations. Note how the peak at $2\omega_0$ remains occluded by the transient tail ("transient BG") up to $T \sim 2080$ fs (inset), and how the transient background is suppressed $\propto 1/T$ with increasing T. The four topmost curves show the raw frequency spectrum (without Hann filtering), while the bottom-most (magenta) illustrates the advantage of using a Hann window, which allows the resolution of up to the third harmonic with only $T = 208$ fs. $[N_k^2 = 36^2, N_c = N_v = 1, E_0 = 0.01\,\mathrm{V/\AA}, \tau = 0.0125\,\mathrm{fs}, \hbar\gamma = 0.1\,\mathrm{eV}]$

consider only inter-band matrix elements of the dipole operator, we capture a clearly finite $\chi^{(2)}(\omega)$ with all the excitonic features previously observed in *ab initio* calculations[28,30,55,77,94–102], as well as using parameterized TB models[22]. Our magnitudes, on the other hand, appear to be underestimated in comparison with these previous calculations by roughly one order of magnitude. For example, while the magnitude of our $\chi^{(2)}(\omega)$ in Fig. 6(c) is $\simeq 0.04\,\mathrm{nm}^2/\mathrm{V}$ at its strongest peak ($\hbar\omega \simeq 1.7$ eV), the same peak has been reported with an intensity $\sim 0.2\,\mathrm{nm}^2/\mathrm{V}$ on the basis of both *ab initio*[107,111,112] and parameterized[22] calculations. We address this discrepancy in the next section although we point out that the only experimental report we are aware of quotes[108] $|\chi^{(2)}(\hbar\omega = 1.53)| = 3 \times 10^{-32}\,\mathrm{mC}/\mathrm{V}^2 \simeq 0.003\,\mathrm{nm}^2/\mathrm{V}$.

Beyond second order and accounting for excitonic effects, we are only aware of the calculations by Attaccalite *et al.*[102] who calculate the frequency-dependent susceptibility for two-photon absorption in hBN. Unfortunately, that corresponds to the response function $\chi^{(3)}(\omega; \omega, \omega, -\omega)$ and not the third-harmonic $\chi^{(3)}(3\omega; \omega, \omega, \omega)$ that we can compute with our current implementation.
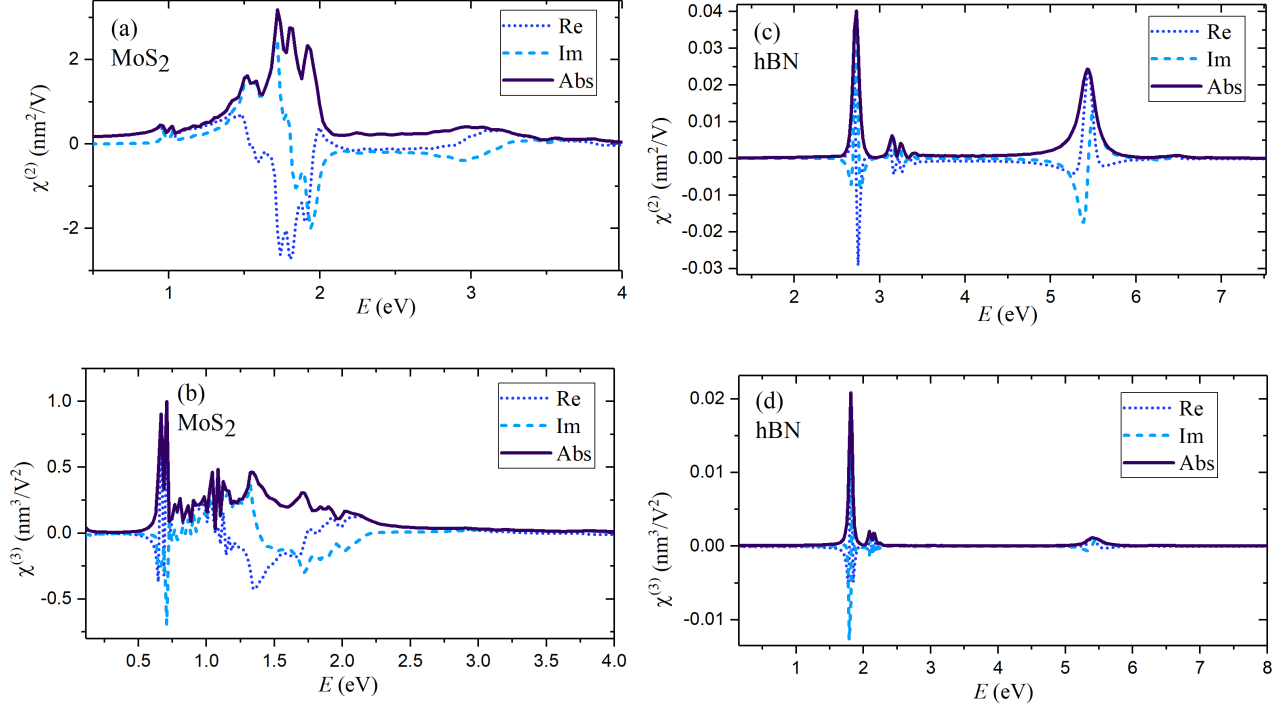
FIG. 6. (a-b) Second- and third-harmonic susceptibilities of MoS_2. (d-e) Likewise, for hBN. Even though, we plot the susceptibilities in the extended frequency ranges shown in each panel for reference, recall that the underlying band structures are truncated; this limits the reliable ranges of validity of $\chi^{(2)}$ $(\chi^{(3)})$ to $\hbar\omega \lesssim 1.75\,(1.17)\,\mathrm{eV}$ for MoS_2 and $\hbar\omega \lesssim 4.5\,(3.0)\,\mathrm{eV}$ for hBN. [for MoS_2: $N_k^2 = 36^2$, $N_c = 6$, $N_v = 2$, $E_0 = 0.01\,\mathrm{V/\mathring{A}}$, $\tau = 0.05\,\mathrm{fs}$, $T = 208\,\mathrm{fs}$, $\hbar\gamma = 0.05\,\mathrm{eV}$; for hBN: $N_k^2 = 60^2$, $N_c = N_v = 1$, $E_0 = 5 \times 10^{-4}\,\mathrm{V/\mathring{A}}$ $(\chi^{(2)})$, $E_0 = 1 \times 10^{-2}\,\mathrm{V/\mathring{A}}$ $(\chi^{(3)})$, $\tau = 0.0125\,\mathrm{fs}$, $T = 208\,\mathrm{fs}$, $\hbar\gamma = 0.1\,\mathrm{eV}$].

IV. DISCUSSION

Trade-offs in the time domain — Beyond the fact that Eq. (7) condenses the problem in a formally simple expression which is suitable for a general-purpose implementation, a key practical advantage of a time-domain formulation is that it entirely circumvents the explicit calculation of each order of a perturbative expansion on the strength of the external field. Even though explicit expressions have been given for some nonlinear susceptibilities both at the level of independent[21,113,114] and interacting[20] electrons, the terms contributing to each order quickly proliferate and become cumbersome to handle already at the second order, especially when they incorporate excitons[20]. Besides, their actual calculation inevitably demands a numerical integration over the BZ, even when using the simplest underlying Hamiltonians[22,115], whose convergence can become numerically challenging due to the presence of singularities in the spectral representation of the perturbative series that must be integrated. On the other hand, the main trade-offs of a time-domain approach are the need of post-processing that must be adapted to the information one desires to extract from the polarization (Fourier analysis, deconvolution, etc.), as well as the total duration of $P(t)$ that must be acquired if one is interested in nonlin-

earities of very high order, or to achieve very high energy resolution in the final result. The issue is, of course, that each time step in the numerical integration of (7) is costly because the electronic self-energy is non-diagonal in crystal momentum $\boldsymbol{k}$. In this regard, the strategy described in Section III D to obtain $\chi^{(n)}(\omega)$ is one of the worst case scenarios since it requires launching one wave per each frequency ω of interest — given the frequency resolution we sought, the results in Figs. 3 and 6 required integrating the equation of motion (28) hundreds, once for each frequency. Yet, using parameterized Hamiltonians and interactions makes such computation entirely feasible without extreme computational resources, while a corresponding fully *ab initio* implementation would face stringent computational challenges[116].

Ultrafast optical processes — Since the temporal profile of the exciting field can be arbitrary, this approach is best suited for realistic simulations of properties that are intrinsically of the time domain. These include simulating the response to pulsed excitation, to specially tailored light pulses, or pump-probe-type scenarios. More interesting is the potential to simulate electronic processes at ultrafast timescales (*e.g.*, femtosecond), natively accounting for electronic interactions, by adequate extensions of the self-energy to capture specific mechanisms of electronic relaxation (*e.g.*, electron-phonon and electron-
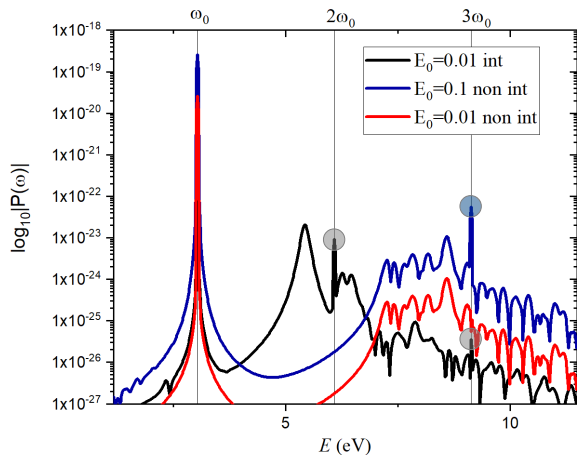
FIG. 7. Frequency spectrum of $|P(\omega)|$ for hBN in response to a continuous monochromatic wave of frequency $\hbar\omega_0 = 3.04\,\mathrm{eV}$, showing results obtained with ("int") and without ("no int") Coulomb interaction. Note how, in the absence of interaction (blue and red curves), there is no second-harmonic response (no peak at $2\omega_0$), even when the field is strong enough to allow resolution of the third harmonic (blue curve). The shaded circles highlight the amplitude of the harmonic peaks at $\omega = n\omega_0$. [$N_k^2 = 60^2$, $N_c = N_v = 1$, $\tau = 0.0125\,\mathrm{fs}$, $T = 208\,\mathrm{fs}$, $\hbar\gamma = 0.1\,\mathrm{eV}$].

electron collisions). The fact that this framework is non-perturbative in the external field *and* captures the evolution of the distribution function out of equilibrium makes it also naturally suited to characterize absorption saturation and other combined non-linear+nonequilibrium effects of interest for applications.

Intra-band transitions — As described in Section II C, we approximated the diagonal matrix elements of the dipole operator as $\boldsymbol{r}_{mm\boldsymbol{k}} \simeq 0$, effectively assuming that only inter-band elements contribute to the system's response. While this would be formally true in linear response, the importance of the *intra*-band contributions (IBCs) at higher orders has been a long and delicate subject of discussion, especially because it touches subtle aspects related to the choice of gauge and approximations to represent the coupling to the external electromagnetic field in the Hamiltonian[21,117–121]. The clearest and most striking example of the potential for inconsistencies arises in a two-band model where, within the length-gauge and in the absence of interactions, *equilibrium* second-order perturbation theory predicts the vanishing of the second-order susceptibility in the ground state of a semiconductor, irrespective of the underlying symmetry[21,115]. The reason is trivial: the n-th order response involves the product of $n + 1$ matrix elements defining a sequence of transitions that must return to the initial state, which is impossible if n is even with only two bands and no IBCs. Recently, it has been shown that adding excitons to the perturbative quadratic susceptibilities does not change the conclusion that IBCs are necessary to obtain a finite

quadratic response in a two-band model[22].

In contrast, our calculations yield a clearly finite $\chi(2)$ [Figs. 5 and 6(c)] from our two-band Hamiltonian describing hBN which, moreover, has all the expected frequency-dependent features by comparison with DFT+GW+BSE results. The difference originates in the *nonequilibrium* nature of our approach. To appreciate that explicitly, consider Fig. 7 where we show $P(\omega)$ under a monochromatic field with frequency ω_0, calculated with and without interactions. In line with the simple argument given above, the non-interacting traces have no SH response [no peak at $P(2\omega_0)$] even when the field is strong enough to reveal a clear third-harmonic peak above the transient background [blue shaded disk; see also Fig. 9(c)]. In contrast, the interacting trace does show a clear peak at $2\omega_0$ (gray shaded disk), the magnitude of which defines the $\chi^{(2)}(\omega)$ plotted in Fig. 6(c) according to the definition given in Eq. (26).

The conclusion in Ref. 22 that IBCs are necessary to capture the SH susceptibility in a two-band model is *conditioned* by the underlying quasi-equilibrium assumption that the populations on each band remain unchanged by the external field. [In our formulation, this assumption means setting $G^<_{mm\boldsymbol{k}}(t) = if_{m\boldsymbol{k}}$ as time-independent, equal to the equilibrium values.] But one can see from Eqs. (4) and (5) of the cited reference that, by relaxing that assumption and explicitly integrating in time *both* coherences and populations, one obtains additional contributions to the second-order response that involve only inter-band matrix elements. This is not surprising because one reason for the absence of purely inter-band contributions under the quasi-equilibrium assumption is the perfect Pauli blocking effect, due to the fact that the occupations $f_{m\boldsymbol{k}}$ remain either 1 or 0, but not fractional. It thus follows that, in a two-band model, it is crucial not only to include IBCs, but also to explicitly take into account the system's deviation from equilibrium when calculating the nonlinear response. Indeed, our result in Fig. 6(c) demonstrates that neglecting IBCs while allowing the distribution function to deviate from equilibrium yields a similar spectral profile for $\chi^{(2)}(\omega)$ as if one approximates the system to quasi-equilibrium with IBCs[122].

Numerical efficiency — The numerical scaling of this framework is extremely simple and follows directly from the nature of the problem defined by the equation of motion (7), the workhorse of the methodology. To integrate $P(t)$ in response to a *single* external wave or pulse requires a total of $T/\Delta t = L$ time steps; when a Runge-Kutta algorithm is employed, one must recall that advancing one time step requires a number of intermediate evaluations that will be discarded, with more discarded the higher the order[123]. (In this regard, our current implementation can be sped up by a factor of two by switching to an integration rule that reuses all evaluations in subsequent steps.) As for storage requirements, it is desirable, for expediency, to store the Coulomb matrix elements (16) which, being the only nondiagonal matrix
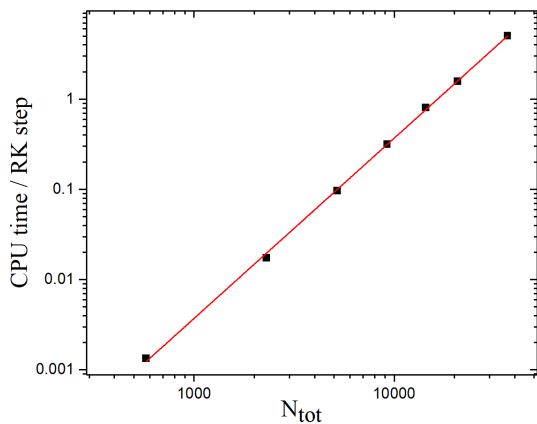
FIG. 8. Typical scaling of the CPU seconds per integration time-step with the linear dimension of the Coulomb matrix, $N_{\text{tot}} = N_k^2(N_c + N_v)^2$. The straight line is $\propto N_{\text{tot}}^2$.

in $\boldsymbol{k}$, is what ultimately determines the storage needs. Since its linear dimension is $N_{\text{tot}} = N_k^2(N_c + N_v)^2$, it ultimately imposes a compromise between the number of points N_k^2 used to sample the BZ and the number of bands. But, as exemplified by our results in the one-wave-per-frequency (worst-case) scenario discussed in Section III D, we were able to include up to $N_c + N_v = 8$ bands and $N_k^2 \sim 36^2$–40^2 to describe MoS$_2$ still within reasonable computational resources. Finally, the calculation time per integration step scales $\propto N_{\text{tot}^2}$ because the evaluation of the right-hand side of Eq. (7) can be coded as a matrix-vector product; this is explicitly shown in Fig. 8.

V. CONCLUSION

We have explicitly demonstrated that parameterized models are capable of retaining excellent agreement with experimental and *ab-initio* optical spectra over large frequency ranges, while markedly alleviating the computational demands of the time-domain framework proposed by Attaccalite *et al.*[28] to study the response to arbitrary light fields. Our results therefore broaden the practical reach of this general-purpose and versatile technique where multiple interacting and/or relaxation mechanisms can be incorporated in a systematic way, and which is natively suited to simulate the current frontier of ultrafast spectroscopy in solid-state materials. We have exposed in detail the relevant adaptations of the technique necessary for that, which will be of value to pursue further refinements and applications such as wave mixing or pump-probe simulations. Finally, we trust this will be a useful contribution to the current interest in robust general methods to tackle the combined nonlinear-nonequilibrium response of crystals under strong fields.

[*] Corresponding author: vpereira@nus.edu.sg

[1] R. W. Boyd, *Nonlinear Optics*, 3rd ed. (Academic Press, Burlington, MA, 2008).

[2] E. Luppi and V. Véniard, Semicond. Sci. Technol. **31**, 123002 (2016).

[3] Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, and M. S. Strano, Nat. Nanotechnol. **7**, 699 (2012).

[4] F. Xia, H. Wang, D. Xiao, M. Dubey, and A. Ramasubramaniam, Nat. Photonics **8**, 899 (2014).

[5] G. Wang, A. Chernikov, M. M. Glazov, T. F. Heinz, X. Marie, T. Amand, and B. Urbaszek, Rev. Mod. Phys. **90**, 21001 (2017).

[6] X. Xu, W. Yao, D. Xiao, and T. F. Heinz, Nat. Phys. **10**, 343 (2014).

[7] J. R. Schaibley, H. Yu, G. Clark, P. Rivera, J. S. Ross, K. L. Seyler, W. Yao, and X. Xu, Nat. Rev. Mater. **1**, 16055 (2016).

[8] P. Cudazzo, I. Tokatly, and A. Rubio, Phys. Rev. B **84**, 085406 (2011).

[9] K. S. Thygesen, 2D Mater. **4**, 022004 (2017).

[10] V. N. Kotov, B. Uchoa, V. M. Pereira, F. Guinea, and A. H. Castro Neto, Rev. Mod. Phys. **84**, 1067 (2012).

[11] E. Ridolfi, C. Lewenkopf, and V. Pereira, Phys. Rev. B **97**, 205409 (2018).

[12] H. Liu, Y. Li, Y. S. You, S. Ghimire, T. F. Heinz, and D. A. Reis, Nat. Phys. **13**, 262 (2016).

[13] R. Huber, F. Tauser, A. Brodschelm, M. Bichler, G. Abstreiter, and A. Leitenstorfer, Nature **414**, 286 (2001).

[14] T. Kampfrath, K. Tanaka, and K. A. Nelson, Nat. Photonics **7**, 680 (2013).

[15] J. Kim, X. Hong, C. Jin, S.-F. F. Shi, C.-Y. Y. S. Chang, M.-H. H. Chiu, L.-J. J. Li, and F. Wang, Science **346**, 1205 (2014).
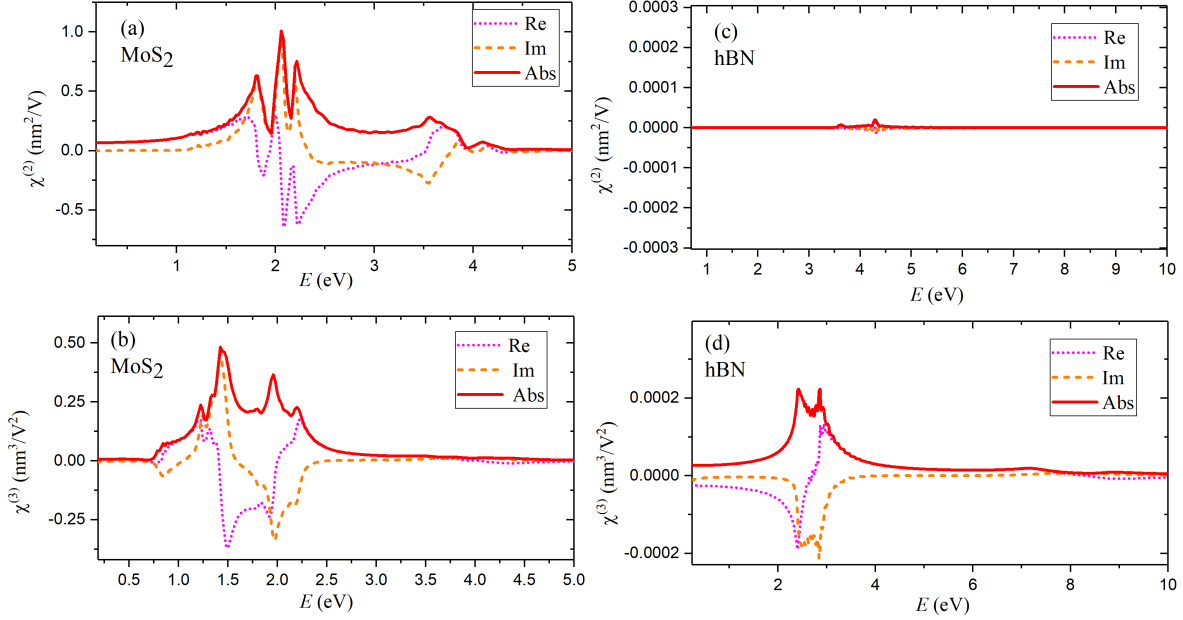
FIG. 9. (a-b) Second- and third-harmonic susceptibilities of MoS_2 deliberately without Coulomb interactions. (d-e) Likewise, for hBN. Note how in (c) $\chi^{(2)}$ is within our noise floor (as expected for a non-interacting two-band model), in contrast with the interacting result shown in Fig. 6(c). [for MoS_2: $N_k^2 = 36^2$, $N_c = 6$, $N_v = 2$, $E_0 = 0.01\,\mathrm{V/Å}$ ($\chi^{(2)}$), $E_0 = 0.05\,\mathrm{V/Å}$ ($\chi^{(3)}$), $\tau = 0.05\,\mathrm{fs}$, $T = 208\,\mathrm{fs}$, $\hbar\gamma = 0.05\,\mathrm{eV}$; for hBN $N_k^2 = 60^2$, $N_c = N_v = 1$, $E_0 = 0.05\,\mathrm{V/Å}$, $\tau = 0.0125\,\mathrm{fs}$, $T = 208\,\mathrm{fs}$, $\hbar\gamma = 0.1\,\mathrm{eV}$].

[16] K. Hao, G. Moody, F. Wu, C. K. Dass, L. Xu, C. H. Chen, L. Sun, M. Y. Li, L. J. Li, A. H. MacDonald, and X. Li, Nat. Phys. **12**, 677 (2016).

[17] Z. Ye, D. Sun, and T. F. Heinz, Nat. Phys. **13**, 26 (2017).

[18] G. Onida, L. Reining, and A. Rubio, Rev. Mod. Phys. **74**, 601 (2002).

[19] E. K. Chang, E. L. Shirley, and Z. H. Levine, Phys. Rev. B **65**, 035205 (2001).

[20] R. Leitsmann, W. G. Schmidt, P. H. Hahn, and F. Bechstedt, Phys. Rev. B **71**, 195209 (2005).

[21] C. Aversa and J. E. Sipe, Phys. Rev. B **52**, 14636 (1995).

[22] T. G. Pedersen, Phys. Rev. B **92**, 235432 (2015).

[23] I. Souza, J. Iniguez, and D. Vanderbilt, Phys. Rev. B **69**, 085106 (2004).

[24] Y. Takimoto, F. Vila, and J. Rehr, The Journal of Chem. Phys. **127**, 154114 (2007).

[25] F. Ding, B. E. V. Kuiken, B. E. Eichinger, and X. Li, The Journal of Chem. Phys. **138**, 064104 (2013).

[26] G. Pal, Y. Pavlyukh, W. Hbner, and H. Schneider, Eur. Phys. J. B **79**, 327 (2011).

[27] E. Perfetto, D. Sangalli, A. Marini, and G. Stefanucci, Phys. Rev. B **92**, 205304 (2015).

[28] C. Attaccalite, M. Grüning, and A. Marini, Phys. Rev. B **84**, 245110 (2011).

[29] C. Attaccalite and M. Grüning, Phys. Rev. B **88**, 235113 (2013).

[30] M. Grüning and C. Attaccalite, Phys. Rev. B **89**, 081102 (2014).

[31] M. Grüning, D. Sangalli, and C. Attaccalite, Phys. Rev. B **94**, 035149 (2016).

[32] N. Tancogne-Dejean, C. Giorgietti, and V. Véniard, Phys. Rev. B **94**, 125301 (2016).

[33] F. Hüser, T. Olsen, and K. S. Thygesen, Phys. Rev. B - Condens. Matter Mater. Phys. **88**, 245309 (2013).

[34] D. Y. Qiu, F. H. da Jornada, and S. G. Louie, Phys. Rev. Lett. **111**, 216805 (2013).

[35] D. Y. Qiu, F. H. da Jornada, and S. G. Louie, Phys. Rev. Lett. **115**, 119901 (2015).

[36] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Mater. **1**, 011002 (2013).

[37] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).

[38] F. A. Rasmussen and K. S. Thygesen, J. Phys. Chem. C **119**, 13169 (2015).

[39] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi, and N. Marzari, Nat. Nanotechnol. **13**, 246 (2018).

[40] J. Zhou, L. Shen, M. D. Costa, K. A. Persson, S. P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang, and Y. P. Feng, Sci. Data **6**, 86 (2019).

[41] M. L. Trolle, G. Seifert, and T. G. Pedersen, Phys. Rev. B **89**, 235410 (2014).

[42] F. Wu, F. Qu, and A. H. MacDonald, Phys. Rev. B **91**, 075310 (2015).

[43] J. C. Slater, G. F. Koster, and M. F. O. R. Solids, Phys. Rev. **94**, 1498 (1954).

[44] S. Schmitt-Rink, D. S. Chemla, and H. Haug, Phys. Rev. B **37**, 941 (1988).

[45] That is frequently the case for 2d semiconductors; see for example Refs. 33–35.

[46] R. Resta, Rev. of Modern Phys. **66**, 899 (1994).

[47] L. P. Kadanoff and G. Baym, *Quantum statistical mechanics: Green's function methods in equilibrium and nonequilibrium problems* (W. A. Benjamin, New York, 1962).

[48] D. Kremp, M. Schlanges, and W.-D. Kraeft, *Quantum Statistics of Nonideal Plasmas* (Springer-Verlag, Berlin/Heidelberg, 2005).

[49] L. Keldysh, Sov. Phys. JETP **20**, 1018 (1964).

[50] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).

[51] L. Hedin and S. Lundqvist, Solid State Physics **23**, 1 (1970).

[52] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).

[53] G. B. Ventura, D. J. Passos, J. M. V. P. Lopes, and J. M. B. Lopes dos Santos, arXiv:1902.04438 (2019).

[54] E. Ridolfi, D. Le, T. S. Rahman, E. R. Mucciolo, and C. H. Lewenkopf, J. Phys.: Condens. Matter **27**, 365501 (2015).

[55] T. Galvani, F. Paleari, H. P. C. Miranda, A. Molina-Sanchez, L. Wirtz, S. Latil, H. Amara, and F. Ducastelle, Phys. Rev. B **94**, 125303 (2016).

[56] T. G. Pedersen, K. Pedersen, and T. Brun Kriestensen, Phys. Rev. B **63**, 201101 (2001).

[57] L. Hedin, Phys. Rev. **139**, A796 (1965).

[58] F. Bruneval, N. Vast, and L. Reining, Phys. Rev. B **74**, 045102 (2006).

[59] M. Rohlfing and S. G. Louie, Phys. Rev. B **62**, 4927 (2000).

[60] W.-D. Schöne and W. Ekardt, Phys. Rev. B **65**, 113112 (2002).

[61] A. Marini and R. Del Sole, Phys. Rev. Lett. **91**, 176402 (2003).

[62] To handle the $\boldsymbol{k}$-diagonal cases where $\boldsymbol{k} = \boldsymbol{k}'$, we use the regularization scheme adopted in Ref. 11.

[63] L. V. Keldysh, JETP Lett. **29**, 658 (1979).

[64] M. L. Trolle, T. G. Pedersen, and V. Véniard, Scientific Reports **7**, 39844 (2017).

[65] A. Chernikov, T. C. Berkelbach, H. M. Hill, A. Rigosi, Y. Li, O. B. Aslan, D. R. Reichman, M. S. Hybertsen, and T. F. Heinz, Phys. Rev. Lett. **113**, 076802 (2014).

[66] A. S. Rodin, A. Carvalho, and A. H. Castro Neto, Phys. Rev. B **90**, 075429 (2014).

[67] To fix the relative phases in the numerical diagonalization of the Bloch Hamiltonian for different $\boldsymbol{k}$, we require $\sum_\alpha C_{\alpha\boldsymbol{k}}^n$ to be real, as suggested in Ref. 59 and previously implemented in Ref. 11.

[68] M. Galassi, *GNU Scientific Library: Reference Manual*, 3rd ed. (Network Theory, 2009).

[69] Since the linear dimension of the problem is very large in virtually all cases of interest, there is no advantage in using higher-order methods which are only efficient when the function evaluations required to advance each time step are fast, which is not the case here.

[70] S. J. Orfanidis, *Introduction to Signal Processing* (Prentice Hall, 1996).

[71] A. R. Klots, A. K. M. Newaz, B. Wang, D. Prasai, H. Krzyzanowska, J. Lin, D. Caudel, N. J. Ghimire, J. Yan, B. L. Ivano, K. A. Velizhanin, A. Burger, D. G. Mandru, N. H. Tolk, S. T. Pantelides, and K. I. Bolotin, Sci. Rep. **4**, 6608 (2014).

[72] A. Molina-Sánchez, K. Hummer, and L. Wirtz, Surf. Sci. Rep. **70**, 554 (2015).

[73] D. Xiao, G.-B. Liu, W. Feng, X. Xu, and W. Yao, Phys. Rev. Lett. **108**, 196802 (2012).

[74] K. F. Mak, K. He, J. Shan, and T. F. Heinz, Nat. Nanotechnol. **7**, 494 (2012).

[75] H. Zeng, J. Dai, W. Yao, D. Xiao, and X. Cui, Nat. Nanotechnol. **7**, 490 (2012).

[76] G. Wang, A. Chernikov, M. M. Glazov, T. F. Heinz, X. Marie, T. Amand, and B. Urbaszek, Rev. Mod. Phys. **90**, 021001 (2018).

[77] B. Arnaud, S. Lebègue, P. Rabiller, and M. Alouani, Phys. Rev. Lett. **96**, 026402 (2006).

[78] C. Zhang, A. Johnson, C.-L. Hsu, L.-J. Li, and C.-K. Shih, Nano Lett. **14**, 2443 (2014).

[79] Y. Li, A. Chernikov, X. Zhang, A. Rigosi, H. M. Hill, A. M. van der Zande, D. A. Chenet, E.-M. Shih, J. Hone, and T. F. Heinz, Phys. Rev. B **90**, 205422 (2014).

[80] G. Jayaswal, Z. Dai, X. Zhang, M. Bagnarol, A. Martucci, and M. Merano, Opt. Lett. **43**, 703 (2018).

[81] K. F. Mak, C. Lee, J. Hone, J. Shan, and T. F. Heinz, Phys. Rev. Lett. **105**, 136805 (2010).

[82] Y. Zhang, T.-R. Chang, B. Zhou, Y.-T. Cui, H. Yan, Z. Liu, F. Schmitt, J. Lee, R. Moore, Y. Chen, H. Lin, H.-T. Jeng, S.-K. Mo, Z. Hussain, A. Bansil, and Z.-X. Shen, Nat. Nanotechnol. **9**, 111 (2013).

[83] B. Zhu, X. Chen, and X. Cui, Sci. Rep. **5** (2015), 10.1038/srep09218.

[84] L. M. Malard, T. V. Alencar, A. P. M. Barboza, K. F. Mak, and A. M. de Paula, Phys. Rev. B **87**, 201401 (2013).

[85] N. Kumar, S. Najmaei, Q. Cui, F. Ceballos, P. M. Ajayan, J. Lou, and H. Zhao, Phys. Rev. B **87**, 161403 (2013).

[86] M. L. Trolle, Y.-C. Tsao, K. Pedersen, and T. G. Pedersen, Phys. Rev. B **92**, 161409(R) (2015).

[87] M. M. Glazov, L. E. Golub, G. Wang, X. Marie, T. Amand, and B. Urbaszek, Phys. Rev. B **95**, 035311 (2017).

[88] D. B. S. Soh, C. Rogers, D. Gray, E. Chatterjee, and H. Mabuchi., Phys. Rev. B **97**, 165111 (2018).

[89] R. Mamy, J. Thomas, G. Jezequel, and J. Lemonnier, J. Physique Lett. **42**, 473 (1981).

[90] K. Watanabe, T. Taniguchi, and H. Kanda, Nat. Mater. **3**, 404 (2004).

[91] L. Song, L. Ci, H. Lu, P. B. Sorokin, C. Jin, J. Ni, A. G. Kvashnin, D. G. Kvashnin, J. Lou, B. I. Yakobson, and P. M. Ajayan, Nano Letters **10**, 3209 (2010).

[92] K. K. Kim, A. Hsu, X. Jia, S. M. Kim, Y. Shi, M. Hofmann, D. Nezich, J. F. Rodriguez-Nieva, M. Dresselhaus, T. Palacios, and J. Kong, Nano Letters **12**, 161 (2012).

[93] Q. Wu, J.-H. Park, S. Park, S. J. Jung, H. Suh, N. Park, W. Wongwiriyapan, S. Lee, Y. H. Lee, and Y. J. Song, Scientific Reports **5** (2015), 10.1038/srep16159.

[94] L. Wirtz, A. Marini, and A. Rubio, Phys. Rev. Letters **96**, 126104 (2006).

[95] L. Wirtz, A. Marini, M. Grüning, C. Attaccalite, G. Kresse, and A. Rubio, Phys. Rev. Lett. **100**, 189701 (2008).

[96] A. Marini, Phys. Rev. Letters **101**, 106405 (2008).

[97] C. Attaccalite, M. Bockstedte, A. Marini, A. Rubio, and L. Wirtz, Phys. Rev. B **83**, 144115 (2011).

[98] J. Yan, K. W. Jacobsen, and K. S. Thygesen, Phys. Rev. B **86**, 045208 (2012).

[99] P. Cudazzo, L. Sponza, C. Giorgetti, L. Reining, F. Sottile, and M. Gatti, Phys. Rev. Lett. **116**, 066803 (2016).

[100] J. Koskelo, G. Fugallo, M. Hakala, M. Gatti, F. Sottile, and P. Cudazzo, Phys. Rev. B **95**, 035125 (2017).

[101] F. Ferreira, A. J. Chaves, N. M. R. Peres, and R. M. Ribeiro, J. Opt. Soc. Am. B **36**, 674 (2019).

[102] C. Attaccalite, M. Grüning, H. Amara, S. Latil, and F. Ducastelle, Phys. Rev. B **98**, 165126 (2018).

[103] M.-H. Chiu, C. Zhang, H.-W. Shiu, C.-P. Chuu, C.-H. Chen, C.-Y. S. Chang, C.-H. Chen, M.-Y. Chou, C.-K. Shih, and L.-J. Li, Nat. Commun. **6**, 7666 (2015).

[104] This has the same practical effect as, for example, adiabatically turning the field on over a time $\sim 2\pi/\gamma$, instead of a sudden turn-on.

[105] This is the same effect as the frequent experimental situation where one might need to acquire a week signal for times long enough to bring it above the noise floor. The fact that the background tails correspond to the transient component of the signal, as highlighted in Fig. 5, is confirmed by the fact that they scale $\propto 1/T$ as can be inspected in that figure.

[106] A loose way of illustrating the physical content of this statement is to think of a classical oscillator driven at the resonant frequency: Although its amplitude begins growing, upon waiting long enough one eventually sees that it does not diverge (strict resonance), reaching instead a finite maximum determined by the damping constant.

[107] M. Grüning and C. Attaccalite, Phys. Rev. B **90**, 199901 (2014).

[108] Y. Li, Y. Rao, K. F. Mak, Y. You, S. Wang, C. R. Dean, and T. F. Heinz, Nano Lett. **13**, 3329 (2013).

[109] R. I. Woodward, R. T. Murray, C. F. Phelan, R. E. P. de Oliveira, T. H. Runcorn, E. J. R. Kelleher, S. Li, E. C. de Oliveira, G. J. M. Fechine, G. Eda, and C. J. S. de Matos, 2D Mater. **4**, 011006 (2017).

[110] Perhaps not surprisingly since these frequencies lie near resonant features of the susceptibilities and, as a result, the magnitude of the response varies noticeably under small deviations from those resonances.

[111] F. Wang, Z. Wang, Q. Wang, F. Wang, L. Yin, K. Xu, Y. Huang, and J. He, Nanotechnology **26**, 292001 (2015).

[112] M. C. Lucking, K. Beach, and H.Terrones, Scientific Reports **8**, 10118 (2018).

[113] J. A. Armstrong, N. Bloembergen, J. Ducuing, and P. S. Pershan, Phys. Rev. **127**, 1918 (1962).

[114] E. Ghahramani, J. E. Sipe, and E. Ghahramani, Phys. Rev. B **48**, 11705 (1993).

[115] F. Hipolito, T. G. Pedersen, and V. M. Pereira, Phys. Rev. B **94**, 045434 (2016).

[116] For example, while the one-wave-per-frequency approach was employed in Ref. 28 to study the nonequilibrium band populations, it could only analyze a select (two) frequencies.

[117] V. N. Genkin and P. M. Mednis, JETP **27**, 609 (1968).

[118] K. Rzążewski and R. W. Boyd, J. Mod. Opt. **51**, 1137 (2004).

[119] A. Taghizadeh, F. Hipolito, and T. G. Pedersen, Phys. Rev. B **96**, 195413 (2017).

[120] G. B. Ventura, D. J. Passos, J. M. B. Lopes dos Santos, J. M. Viana Parente Lopes, and N. M. R. Peres, Phys. Rev. B **96**, 035431 (2017).

[121] A. Taghizadeh and T. G. Pedersen, Phys. Rev. B **97**, 205432 (2018).

[122] Therefore, strictly speaking, both our SH susceptibility and that in Ref. 22 for hBN are incomplete. Incorporating intra-band matrix elements into our formulation is beyond the scope of the present paper.

[123] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1964).