Machine Learning of Quasiparticle Energies in Molecules and Clusters

Onur Çaylak[*] and Björn Baumeier[†]

Department of Mathematics and Computer Science, Eindhoven University of Technology,
P.O. Box 513, 5600MB Eindhoven, The Netherlands and
Institute for Complex Molecular Systems, Eindhoven University of Technology,
P.O. Box 513, 5600MB Eindhoven, The Netherlands

We present a Δ-Machine Learning approach for the prediction of *GW* quasiparticle energies (ΔMLQP) and photoelectron spectra of molecules and clusters, using orbital-sensitive graph-based representations in kernel ridge regression based supervised learning. Coulomb matrix, Bag-of-Bonds, and Bonds-Angles-Torsions representations are made orbital-sensitive by augmenting them with atom-centered orbital charges and Kohn–Sham orbital energies, which are both readily available from baseline calculations on the level of density-functional theory (DFT). We first illustrate the effects of different constructions of the orbital-sensitive representations (OSR) on the prediction of frontier orbital energies of 22K molecules of the QM8 dataset, and show that is is possible to predict the full photoelectron spectrum of molecules within the dataset using a single model with a mean-absolute error below 0.1 eV. We further demonstrate that the OSR-based ΔMLQP captures the effects of intra- and intermolecular conformations in application to water monomers and dimers. Finally, we show that the approach can be embedded in multiscale simulation workflows, by studying the solvatochromic shifts of quasiparticle and electron-hole excitation energies of solvated acetone in a setup combining Molecular Dynamics, DFT, the *GW* approximation and the Bethe–Salpeter Equation. Our findings suggest that the ΔMLQP model allows to predict quasiparticle energies and photoelectron spectra of molecules and clusters with *GW* accuracy at DFT cost.

I. INTRODUCTION

Fundamental insights gained by computational analysis of electronically excited states of molecular systems can help improving the design of molecular materials and plays therefore a vital role in material science. However, obtaining quantitative predictions is challenging as traditional methods either come at insufficient accuracy, e.g., due to the lack of correlation in interpreting orbital energies of Hartree–Fock, or at the price of high computational costs, as for coupled cluster methods, quantum Monte Carlo, or Green's function approaches. Hence, the incorporation of quantum machine learning (QML) has been gaining great traction over recent years. QML based surrogate property models have become a popular alternative approach for their fast, reliable, and accurate predictions of molecular and material properties [1–14].

The main advantage of ML models is that they allow predictions of molecular properties with improved efficiency at a lower computational cost compared to traditional quantum chemistry approaches. Method development in the field of QML is progressing rapidly and it is increasingly influencing traditional methods [6, 15–18]. Developments in molecular representations and QML models have paved the way for predicting energetic, electronic, and thermodynamic properties, such as atomization energies, dipole moments, polarizabilities, and harmonic frequencies [19–21].

QML of excited states of molecules has remained difficult in comparison. Recent work [5, 22, 23] has achieved promising results for predictions of single frontier orbital (highest or lowest molecular orbital) energies. However, some applications, such as the evaluation of direct or inverse photoelectron spectra, require predictions for a wider range of orbitals simultaneously, with sensitivity to conformational details of the actual molecules and/or a complex embedding environment. This requires in turn the ability to capture both structural and orbital details in the QML model, circumventing the need to build separate models for each state of interest and the associated difficulties in finding unique characterizations of multiple orbitals across a wide range of chemical space.

In this paper we show a way of augmenting existing graph-based representations with orbital-specific information from density-functional theory (DFT) which allows us to individually predict quasiparticle energies on the level of Many-Body Green's Functions Theory in the *GW* approximation and to calculate full photoelectron spectra of molecules and clusters with a single target kernel ridge regression Δ-Machine Learning model (ΔMLQP). We adopt the Δ-ML approach [20] as its concept of learning the corrections to a certain baseline property matches directly the way in which quasiparticle corrections are obtained perturbatively to the Kohn–Sham orbital energies, i.e., $\varepsilon_i^{\mathrm{QP}} = \varepsilon_i^{\mathrm{KS}} + \Delta\varepsilon_i^{GW}$. Within ΔMLQP, we consider specifically the orbital-sensitive augmentation of Coulomb matrix (CM), Bag-of-Bonds (BoB), and Bonds-Angles-Torsions (BAT) representations by a combination of atom-centered orbital charges and Kohn–Sham orbital energies, which are all easily accessible from standard DFT ground state calculations [2, 9, 21, 24]. We illustrate the effect of different methodological choices, such as the determination of the orbital charges as Mulliken populations [25] or from a

[*] o.caylak@tue.nl
[†] b.baumeier@tue.nl

Gaussian Distributed Multipole Analysis [26] (GDMA), on the prediction of the full quasiparticle spectra of molecules in the QM8 dataset. To scrutinize that orbital-sensitive representations (OSR) with multiple orbitals are also capable of resolving the effects of intra- and intermolecular conformations, we also study the photoelectron spectra of water monomers and dimers, taken from the H2O-13 dataset [27]. Finally, we study the use of the ΔMLQP predicted, conformational-sensitive quasiparticle energies embedded in calculations of electron-hole excitations on the level of the Bethe–Salpeter Equation (BSE) for the prototypical example of acetone in aqueous solution, and compare the solvatochromic shifts obtained from BSE@*GW* and BSE@ML.

This paper is organized as follows: In Section II, we briefly summarize the background of Many-Body Green's Functions theory in the *GW* approximation with the Bethe–Salpeter equation, of Kernel-Ridge Regression models and Δ-Machine Learning. Section III introduces the ΔMLQP approach and showcases results in application to predictions quasiparticle energies of molecules from the QM8 data set, water monomer and dimers, as well as acetone in aqueous solution. A brief summary concludes the paper.

II. METHODOLOGICAL BACKGROUND

A. Quasiparticle and electron-hole excitations with *GW*-BSE

For a given closed-shell ground state system with N electrons, one can within Density-Functional Theory (DFT) get the Kohn-Sham energies [28] by solving

$$\widehat{H}^{\mathrm{KS}}|\phi_i^{\mathrm{KS}}\rangle = [\widehat{H}_0 + \widehat{V}_{\mathrm{xc}}]|\phi_i^{\mathrm{KS}}\rangle = \varepsilon_i^{\mathrm{KS}}|\phi_i^{\mathrm{KS}}\rangle, \qquad (1)$$

with $\widehat{H}_0 = \widehat{T}_0 + \widehat{V}_{\mathrm{ext}} + \widehat{V}_{\mathrm{H}}$, where $\widehat{T}_0$ stands for the kinetic energy, $\widehat{V}_{\mathrm{ext}}$ the external potential, $\widehat{V}_{\mathrm{H}}$ the Hartree potential, and $\widehat{V}_{\mathrm{xc}}$ being the exchange-correlation potential.

The addition ($N \to N+1$) or removal ($N \to N-1$) of a single electron to/from the system can be seen as the excitation of a quasiparticle. Quasiparticle energies are essentially the poles of the interacting one-electron Green's function and obey the Dyson's equation [29–32]

$$\left[\widehat{H}_0 + \widehat{\Sigma}(\varepsilon_i^{\mathrm{QP}})\right]|\phi_i^{\mathrm{QP}}\rangle = \varepsilon_i^{\mathrm{QP}}|\phi_i^{\mathrm{QP}}\rangle\,, \qquad (2)$$

where $|\phi_i^{\mathrm{QP}}\rangle$ are the quasiparticle wavefunctions. The operator $\widehat{\Sigma}(\cdot)$ is the self-energy operator, which describes the exchange-correlation many-body effects. This operator can within the *GW* approximation be expressed as a convolution of the one-particle Green's function $G(\mathbf{r},\mathbf{r}',\omega)$ with the screened Coulomb interaction

$$W(\mathbf{r},\mathbf{r}',\omega) = \epsilon^{-1}(\mathbf{r},\mathbf{r}',\omega)\,v_c(\mathbf{r},\mathbf{r}') = \frac{\epsilon^{-1}(\mathbf{r},\mathbf{r}',\omega)}{|\mathbf{r}-\mathbf{r}'|}, \qquad (3)$$

where ϵ^{-1} is the inverse dielectric function calculated in the random-phase approximation [33]. The self-energy operator can be explicitly written as

$$\Sigma(\mathbf{r},\mathbf{r}',\omega) = \frac{\mathrm{i}}{2\pi}\int \mathrm{d}\omega'\, G(\mathbf{r},\mathbf{r}',\omega+\omega')\, W(\mathbf{r},\mathbf{r}',\omega)\,. \qquad (4)$$

Several techniques can be used to perform the frequency integration in 4, starting from separating the self-energy $\Sigma = \mathrm{i}GW$ into its bare exchange part $\Sigma_{\mathrm{x}} = \mathrm{i}Gv_{\mathrm{c}}$ and its correlation part $\Sigma_{\mathrm{c}} = \mathrm{i}G\widetilde{W}$, where $\widetilde{W} = W - v_{\mathrm{c}}$. With that, the integral can be evaluated fully analytically by calculating the reducible polarizability in terms of an eigenvalue decomposition of the RPA Hamiltonian. See, e.g., Refs. [34–37] for details. While this Fully Analytical Approach (FAA) is analytically exact, it is not feasible for large systems due to the scaling of the the diagonalization of the RPA Hamiltonian. Instead, the frequency integration can be approximated within a generalized plasmon-pole model (PPM) [38, 39].

One can now by expanding quasiparticle wavefunctions in terms of Kohn-Sham wavefunctions transform Eq. 2 into

$$H_{ij}^{\mathrm{QP}}(E) = \varepsilon_i^{\mathrm{KS}}\delta_{ij} + \langle\phi_i^{\mathrm{KS}}|\widehat{\Sigma}(E) - \widehat{V}_{\mathrm{xc}}|\phi_j^{\mathrm{KS}}\rangle. \qquad (5)$$

We can get the quasiparticle energies perturbatively by assuming $|\phi_i^{\mathrm{QP}}\rangle \approx |\phi_i^{\mathrm{KS}}\rangle$

$$\begin{aligned}\varepsilon_i^{\mathrm{QP}} &= \varepsilon_i^{\mathrm{KS}} + \Delta\varepsilon_i^{GW} \\ &= \varepsilon_i^{\mathrm{KS}} + \langle\phi_i^{\mathrm{KS}}|\widehat{\Sigma}(\varepsilon_i^{\mathrm{QP}}) - \widehat{V}_{\mathrm{xc}}|\phi_i^{\mathrm{KS}}\rangle.\end{aligned} \qquad (6)$$

As $\Delta\varepsilon_i^{GW}$ itself depends on $\varepsilon_i^{\mathrm{QP}}$, the above constitutes a fixed-point problem. In addition, the $\varepsilon_i^{\mathrm{QP}}$ enter both the energy-dependent microscopic dielectric function determined within the RPA as well as in the Green's function in the expression for the self-energy Eq. 4. Hence, the the so called ev*GW* approach, quasiparticle energies are used to update Σ until self-consistency is reached in combination with Eq. 6.

Optical excitations of the system lead to the formation of coupled electron-hole pairs, which can be described using a product basis of QP wave functions, i.e.,

$$\begin{aligned}\chi_S(\mathbf{r}_{\mathrm{e}},\mathbf{r}_{\mathrm{h}}) = \sum_v^{\mathrm{occ}}\sum_c^{\mathrm{unocc}}\sum_{\sigma\sigma'} & A_{vc,\sigma\sigma'}^S\phi_{c,\sigma'}(\mathbf{r}_{\mathrm{e}})\phi_{v,\sigma}^*(\mathbf{r}_{\mathrm{h}}) \\ & + B_{vc,\sigma\sigma'}^S\phi_{v,\sigma'}(\mathbf{r}_{\mathrm{e}})\phi_{c,\sigma}^*(\mathbf{r}_{\mathrm{h}}),\end{aligned} \qquad (7)$$

where $\mathbf{r}_{\mathrm{e}}$ ($\mathbf{r}_{\mathrm{h}}$) is for the electron (hole) coordinate, and we drop the label QP for clarity. The expansion coefficients $A_{vc,\sigma\sigma'}$ ($B_{vc,\sigma\sigma'}$) of the excited state wave function in terms of resonant (anti-resonant) transitions between QP occupied (occ.) states v and unoccupied (unocc.) c with spin σ and σ', respectively, can be obtained as solutions of the Bethe–Salpeter Equation (BSE) in the form of an effective two-particle Hamiltonian problem

$$\begin{pmatrix} \underline{\mathbf{H}}^{\mathrm{res}} & \underline{\mathbf{M}} \\ -\underline{\mathbf{M}} & -\underline{\mathbf{H}}^{\mathrm{res}} \end{pmatrix}\begin{pmatrix}\mathbf{A}^S \\ \mathbf{B}^S\end{pmatrix} = \Omega_S\begin{pmatrix}\mathbf{A}^S \\ \mathbf{B}^S\end{pmatrix}. \qquad (8)$$

In cases with negligible spin-orbit coupling, it can be shown that this Hamiltonian has block structure in terms of the spin combination of the electron and hole states [40] and can therefore be decomposed into two independent Hamiltonians for singlet and triplet excitations, respectively. This allows to drop the explicit spin variables and the matrix elements of $\underline{\mathbf{H}}^{\text{res}}$ and $\underline{\mathbf{M}}$ are determined as

$$H^{\text{res}}_{vc,v'c'} = D_{vc,v'c'} + \kappa M^{\text{x}}_{vc,v'c'} + M^{\text{d}}_{vc,v'c'} \tag{9}$$

$$M_{cv,v'c'} = \kappa M^{\text{x}}_{cv,v'c'} + M^{\text{d}}_{cv,v'c'} \,, \tag{10}$$

where $\kappa = 2$ (0) for spin singlet (triplet) excitations, and

$$D_{vc,v'c'} = (\varepsilon_c - \varepsilon_v)\delta_{vv'}\delta_{cc'}, \tag{11}$$

$$\begin{aligned} M^{\text{d}}_{vc,v'c'} = -\int \mathrm{d}^3\mathbf{r}_{\text{e}}\, \mathrm{d}^3\mathbf{r}_{\text{h}}\, \phi^*_c(\mathbf{r}_{\text{e}})\phi_{c'}(\mathbf{r}_{\text{e}})W(\mathbf{r}_{\text{e}}, \mathbf{r}_{\text{h}}, \omega = 0) \\ \times\, \phi_v(\mathbf{r}_{\text{h}})\phi^*_{v'}(\mathbf{r}_{\text{h}})\,, \end{aligned} \tag{12}$$

$$\begin{aligned} M^{\text{x}}_{vc,v'c'} = \int \mathrm{d}^3\mathbf{r}_{\text{e}}\, \mathrm{d}^3\mathbf{r}_{\text{h}}\, \phi^*_c(\mathbf{r}_{\text{e}})\phi_v(\mathbf{r}_{\text{e}})v_{\text{c}}(\mathbf{r}_{\text{e}}, \mathbf{r}_{\text{h}}) \\ \times\, \phi_{c'}(\mathbf{r}_{\text{h}})\phi^*_{v'}(\mathbf{r}_{\text{h}})\,. \end{aligned} \tag{13}$$

In Eq. 11, the term D arises from free interlevel transition between occupied and empty quasiparticle states, the *direct interaction* M^{d} (Eq. 12) is responsible for the binding of the electron-hole pair and is based on the attractive, but screened, interaction W (in the static approximation $\omega = 0$) between electron and hole. The repulsive *exchange interaction* M^{x} (Eq. 13) is responsible for the singlet-triplet splitting.

B. Kernel Ridge Regression

The application of Kernel Ridge Regression (KRR) models to predicting molecular quantum properties has been very successful over recent years [8, 14, 15, 20, 41]. The main idea relies on constructing a kernel matrix with a kernel function k that can quantitatively measure similarity between molecular representations $\mathbf{x}_i$ and $\mathbf{x}_j$, which are vector representations that encode the molecular physics [6, 21, 24]. The Laplacian kernel function, for example, is described as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_1}{\sigma}\right). \tag{14}$$

In context of QM, the goal of KRR is to map an input molecular representation $\mathbf{x}$ to a target quantum property p. Such a mapping is given by

$$p(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}, \mathbf{x}_n), \tag{15}$$

where α_n stand for the n-th regression coefficient, while $\mathbf{x}_n$ being the n-th training sample. The learning process within the KRR framework corresponds to obtaining the regression coefficients vector α for a given reference property vector $\mathbf{p}^{\text{train}}$, a kernel matrix $\mathbf{K}$, and a regularization coefficient λ by evaluating

$$\alpha = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{p}^{\text{train}}, \tag{16}$$

where $\mathbf{I}$ is an identity matrix. Additionally, the so-called hyperparameters λ and σ are optimized using the mean-absolute-error (MAE) metric, where all optimizations are performed with 5-fold cross-validation.

C. Δ-Machine Learning

The default workflow of a machine learning based task is to map an input vector to a target property, which makes sense whenever the target value is computationally cheap to compute. This, however, is usually only the case when the target property has relatively low accuracy [20, 41]. In Δ-ML this is often referred to as the baseline property [42, 43]. To eliminate this problem one can employ computationally costly methods, which in most cases result in better accuracy. Unfortunately, the computational cost of these methods are in a lot of situations not affordable. Here, it makes sense to make use of a Δ-ML workflow, where the aim is to predict the highly accurate target property at the same cost of the computationally cheaper methods, which are relatively easy to obtain.

The accurate target property is labeled as p^{t} and is obtained by

$$\begin{aligned} p^{\text{t}}(\mathbf{x}) &= p^{\text{b}}(\mathbf{x}) + \Delta^t_b(\mathbf{x}) && (17) \\ &= p^{\text{b}}(\mathbf{x}) + \sum_{n=1}^{N} \alpha_n k(\mathbf{x}, \mathbf{x}_n), && (18) \end{aligned}$$

with p^{b} being the baseline property and α_n being the n-th regression coefficient of a KRR model that is trained to predict the difference between the target and baseline property, i.e. $p^{\text{t}} - p^{\text{b}}$. The Δ-ML model has been used in various applications and has shown to be powerful in not only saving computational time, but also achieving much higher precision compared to traditional machine learning approaches [20, 43].

III. RESULTS

We illustrate the applicability of the orbital-sensitive Δ machine learning model for predicting full quasiparticle spectra by reporting the predictive performance on QM8 molecules and water monomers/dimers for graph-based representations. Subsequently, to obtain excitation energies of acetone and acetone in water, the proposed approach is used as a surrogate model to evaluate BSE. The results are then used to study solvatochromic shifts and benchmark it against experimental data.

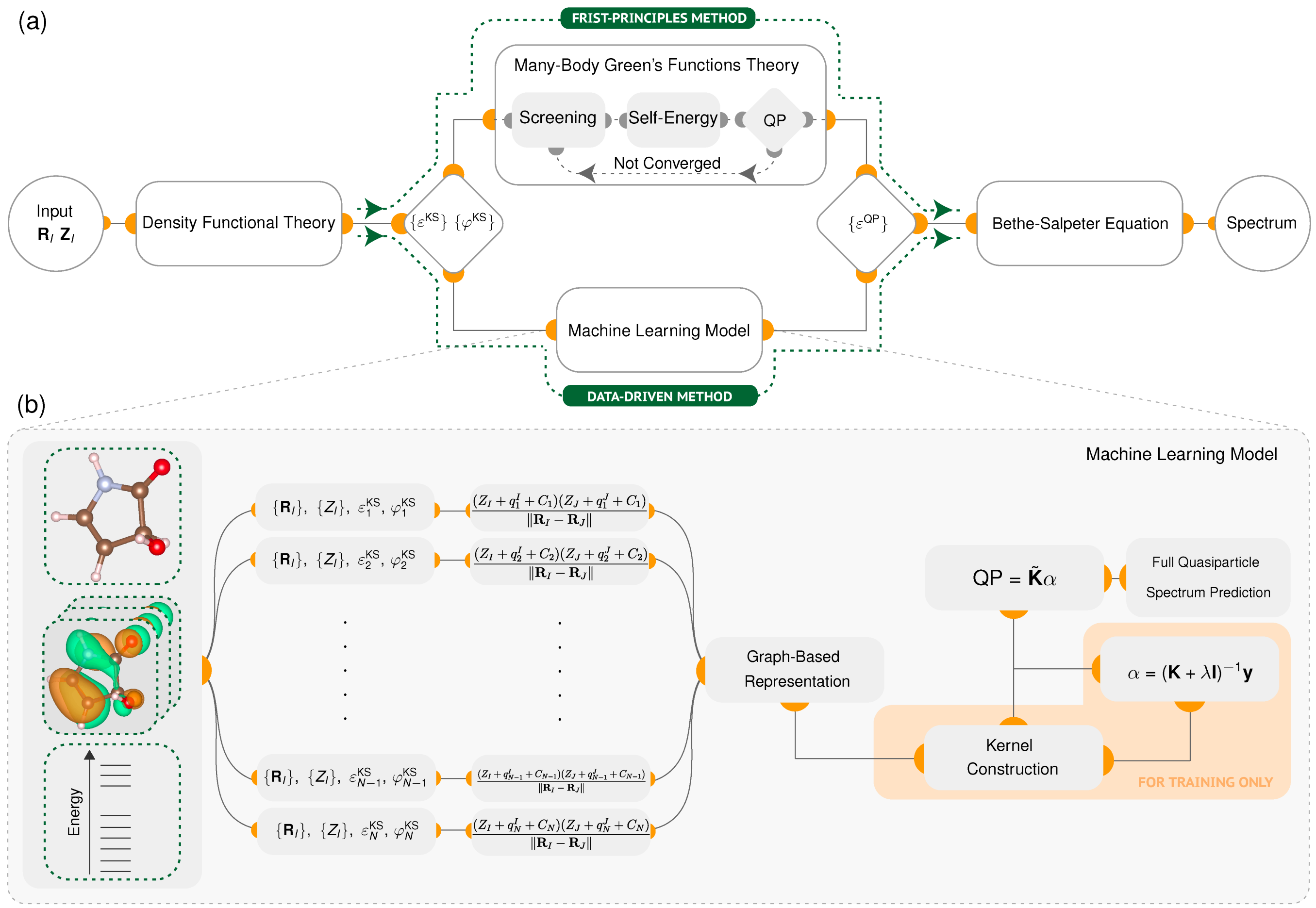
(a)
FRIST-PRINCIPLES METHOD
Many-Body Green's Functions Theory
Screening
Self-Energy
QP
Not Converged
Input
$\mathbf{R}_I$ $\mathbf{Z}_I$
Density Functional Theory
$\{\varepsilon^{\mathrm{KS}}\}$ $\{\varphi^{\mathrm{KS}}\}$
$\{\varepsilon^{\mathrm{QP}}\}$
Bethe-Salpeter Equation
Spectrum
Machine Learning Model
DATA-DRIVEN METHOD
(b)
Machine Learning Model
Energy
$\{\mathbf{R}_I\}$, $\{Z_I\}$, $\varepsilon_1^{\mathrm{KS}}$, φ_1^{KS}
$\{\mathbf{R}_I\}$, $\{Z_I\}$, $\varepsilon_2^{\mathrm{KS}}$, φ_2^{KS}
$\{\mathbf{R}_I\}$, $\{Z_I\}$, $\varepsilon_{N-1}^{\mathrm{KS}}$, $\varphi_{N-1}^{\mathrm{KS}}$
$\{\mathbf{R}_I\}$, $\{Z_I\}$, $\varepsilon_N^{\mathrm{KS}}$, φ_N^{KS}
$\frac{(Z_I+q_1^I+C_1)(Z_J+q_1^J+C_1)}{\|\mathbf{R}_I-\mathbf{R}_J\|}$
$\frac{(Z_I+q_2^I+C_2)(Z_J+q_2^J+C_2)}{\|\mathbf{R}_I-\mathbf{R}_J\|}$
$\frac{(Z_I+q_{N-1}^I+C_{N-1})(Z_J+q_{N-1}^J+C_{N-1})}{\|\mathbf{R}_I-\mathbf{R}_J\|}$
$\frac{(Z_I+q_N^I+C_N)(Z_J+q_N^J+C_N)}{\|\mathbf{R}_I-\mathbf{R}_J\|}$
Graph-Based Representation
QP = $\tilde{\mathbf{K}}\alpha$
Full Quasiparticle Spectrum Prediction
$\alpha = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$
Kernel Construction
FOR TRAINING ONLY

FIG. 1. A schematic overview of first-principles and data-driven routes for the calculation of *GW* quasiparticle energies and their embedding into a BSE@DFT-*GW*/ML workflow. (a) Atomic charges and atomic coordinates are used as input for DFT calculations. DFT generates Kohn-Sham energies and wavefunctions, which are being used as input for either quasiparticle calculations in the first-principles route or Δ-ML model in the data-driven route. The output of either route is then used as input for BSE calculations to output the excitation energy spectra. The machine learning block consist of two operations. (b) Details of the ΔMLQP model with orbital-sensitive representations. First, the DFT output together with geometric information is transformed into a molecular representation. Second, the resulting vector is used to build up a kernel matrix to predict ΔQP energies, which is then are added to KS energies and used as output.

A. Orbital-sensitive descriptors

Figure 1(a) shows a schematic overview of the ΔMLQP approach as a surrogate model to predict quasiparticle energies, thereby bypassing the computationally expensive first-principles *GW* step. It is based on the idea of learning the nonlinear transformations from Kohn–Sham energies to quasiparticle energies, motivated by the fact that for all orbitals i

$$\varepsilon_i^{\mathrm{QP}} = \varepsilon_i^{\mathrm{KS}} + \Delta\varepsilon_i, \qquad (19)$$

corresponds directly to the form of Eq. 17. This allows us to identify $\varepsilon_i^{\mathrm{KS}}$ with the baseline property $p^{\mathrm{b}}(\mathbf{x})$ and to learn the quasiparticle corrections $\Delta\varepsilon_i = \langle\phi_i^{\mathrm{KS}}|\widehat{\Sigma}(\varepsilon_i^{\mathrm{QP}}) - \widehat{V}_{\mathrm{xc}}|\phi_i^{\mathrm{KS}}\rangle$ as in Eq. 6.

To be able to replace all orbital-dependent $\Delta\varepsilon$ by a single Laplacian kernel based (or any other) machine learning model requires a representation that incorporates orbital information. Traditional graph-based representations, like the Coulomb Matrix (CM) and BoB [2, 9, 21], which solely rely on the sets of atomic positions $\{\mathbf{R}\}$ and nuclear charges $\{Z\}$ lack such information and have therefore no injectivity for predicting full spectra of molecules.

We propose an extension to graph-based representation that includes information about the Kohn–Sham orbital energies and wavefunctions, and will refer to the extended version as *orbital-sensitive representation* n-OSR, where n stands for the number of included orbitals. To this end, we map the $\{\varphi_k^{\mathrm{KS}}(\mathbf{r})\}$ to a set effective orbital-dependent atomic charges $\{q_k\}$ and add those with its (rescaled) Kohn–Sham energy $C_k = \zeta\varepsilon_k^{\mathrm{KS}}$, where $[\zeta] = \mathrm{e}/\mathrm{Hartree}$, to the nuclear charges. For example, the Coulomb Matrix representation for a molecule at electronic state k can be extended as

$$\mathrm{CM}_k = \begin{cases} \dfrac{(Z_I + q_k^I + C_k)(Z_J + q_k^J + C_k)}{\|\mathbf{R}_I - \mathbf{R}_J\|}, & \text{for } I \neq J \\ \dfrac{1}{2}(Z_I + q_k^I + C_k)^{2.4}, & \text{for } I = J. \end{cases} \tag{20}$$

A modification as in Eq. 20 allows us to introduce the missing injectivity for multi-state predictions within a single model, as indicated in Fig. 1(b). It is a particularly attractive choice as all ingredients for this modification are readily available from the DFT baseline calculations, and its very simple, physically interpretable form is easy to implement. The same idea of incorporating orbital-sensitivity can be applied to all $\mathbf{R}$ and Z dependent graph-based representations, such as BoB or BAT. In the following, we will evaluate the above extension of graph-based representations, their dependence on the choice of different methods to obtain the $\{q_k\}$, in the use of ΔMLQP.

B. QM8 Molecules

We use molecular geometries from the QM8 data set, which contains more than 20000 synthetically feasible small organic molecules with up to eight CONF atoms [19, 41]. All quantum-mechanical calculations have been performed with the VOTCA-XTP package [36, 44]. For each molecule, we first perform DFT ground state calculations with the PBE0 hybrid functional [45], the def2-TZVP basis set [46], and an optimized auxiliary basis for the resolution-of-identity techniques [47]. Orbital-dependent atomic charges are determined from a Gaussian Distributed Multipole Analysis (GDMA) [26] or from Mulliken populations [25]. Eigenvalue selfconsistent (evGW) quasiparticle energies are then determined for the lowest $2N_{\mathrm{occ}}$ states (excluding the core levels), where N_{occ} is the number of occupied levels. All orbitals are included in the RPA step and not explicitly corrected higher levels are scissors shifted according to the highest absolute quasiparticle correction among the explicitly corrected unoccupied orbitals. The frequency integration in Eq. 4 is performed using the FAA.

In Fig. 2 we report the performance measures, such as correlations and learning curves, for predicting individual HOMO energies (Fig. 2(a,d)), individual LUMO energies (Fig. 2(b,e)), and predicting simultaneously both HOMO and LUMO energies (Fig. 2(c,f)) of QM8 molecules with various Coulomb matrix (CM), bag-of-bonds (BoB), and bonds-angles-torsions (BAT) representations and extensions based on GDMA orbital charges. In all cases, we used $\sigma = 800$ and $\lambda = 10^{-8}$ in the KRR models, and $\zeta = 1\,\mathrm{e/Hartree}$.

The correlation and distribution plots display the expected nonlinear shift between KS and QP energies. As expected, the HOMO (LUMO) $\varepsilon^{\mathrm{KS}}$ are consistently above (below) the corresponding $\varepsilon^{\mathrm{QP}}$. The 1-OSR Δ-ML models as in Fig. 2(a,b) are constructed using frontier orbitals of 18000 molecules as training set and 2000 as testing set and are based on the extended BAT representations. They transform KS energies to QP energies with a mean absolute error of 0.02 eV, respectively. The corresponding learning curves in Fig. 2(d,e) show systematic decay in error with increasing number of training samples N_{samples} (here, each sample corresponds to one molecule) for both standard and extended representations. However, it is also clearly visible that even for a 1-OSR model, the inclusion of information about the KS orbital energies and wavefunctions via the partial charges systematically improves the MAE at fixed training set size in all cases.

We proceed with discussing the first 2-OSR ΔMLQP model, trained on a mixed set of 30000 HOMOs and LUMOs, with testing performed on 5000. The results of a model simultaneously predicting HOMO and LUMO quasiparticle energies is shown in Fig. 2(c). Using orbital-sensitive BAT we note a combined MAE of 0.04 eV. The associated learning curves in Fig. 2(f) show clear improvements of the MAE with increasing number of samples. Note that the training set used to predict multiple states simultaneously contains for each molecule multiple states and the number of samples does therefore not correspond to the number of molecules. Additionally, unmodified representations fail in predicting both targets at the same time, as expected. Overall, the orbital-sensitive representations are more data efficient and allow double-state predictions.

Based on the promising performance of the 2-OSR Δ-ML model, we apply the same method to predicting the full quasiparticle spectra of QM8 molecules in Multi-OSR ΔMLQP. To do this, we first shuffle the dataset comprising all considered quasiparticle levels of all molecules, and then select a random subset with 30000 samples as training and 5000 as testing set. In Fig. 3(a) the ML-QP correlation plot is shown bas on application to 1000 out-of-samples orbitals. Good agreement of the distributions can be observed over an energy range covering around 2 Hartree, where a mean absolute error of 0.06 eV provides a good example of the predictive capabilities of the Multi-OSR Δ-ML model.

Figure 3(b) shows the learning curves for standard and orbital-sensitive representations. While orbital-sensitive representations allow us to predict the entire QP spectra of QM8 molecules, representations without orbital information fail to learn and hence we cannot report systematic improvement with increasing number of training samples. How this translates to a more practical example is shown in Fig. 3(c). A density of states (DOS) plot of a randomly chosen QM8 molecule (ID 014520 in the QM8 data set) based on KS, QP, and ML, respectively, is shown. The difference between the KS HOMO-LUMO gap and the QP HOMO-LUMO gap equates to approximately 4 eV, while the difference between the full QP spectra and ML spectra is 0.07 eV (MAE of all predicted levels). The DOS (with a Gaussian broadening of 0.022 Hartree) based on ΔMLQP (solid line) is prac-

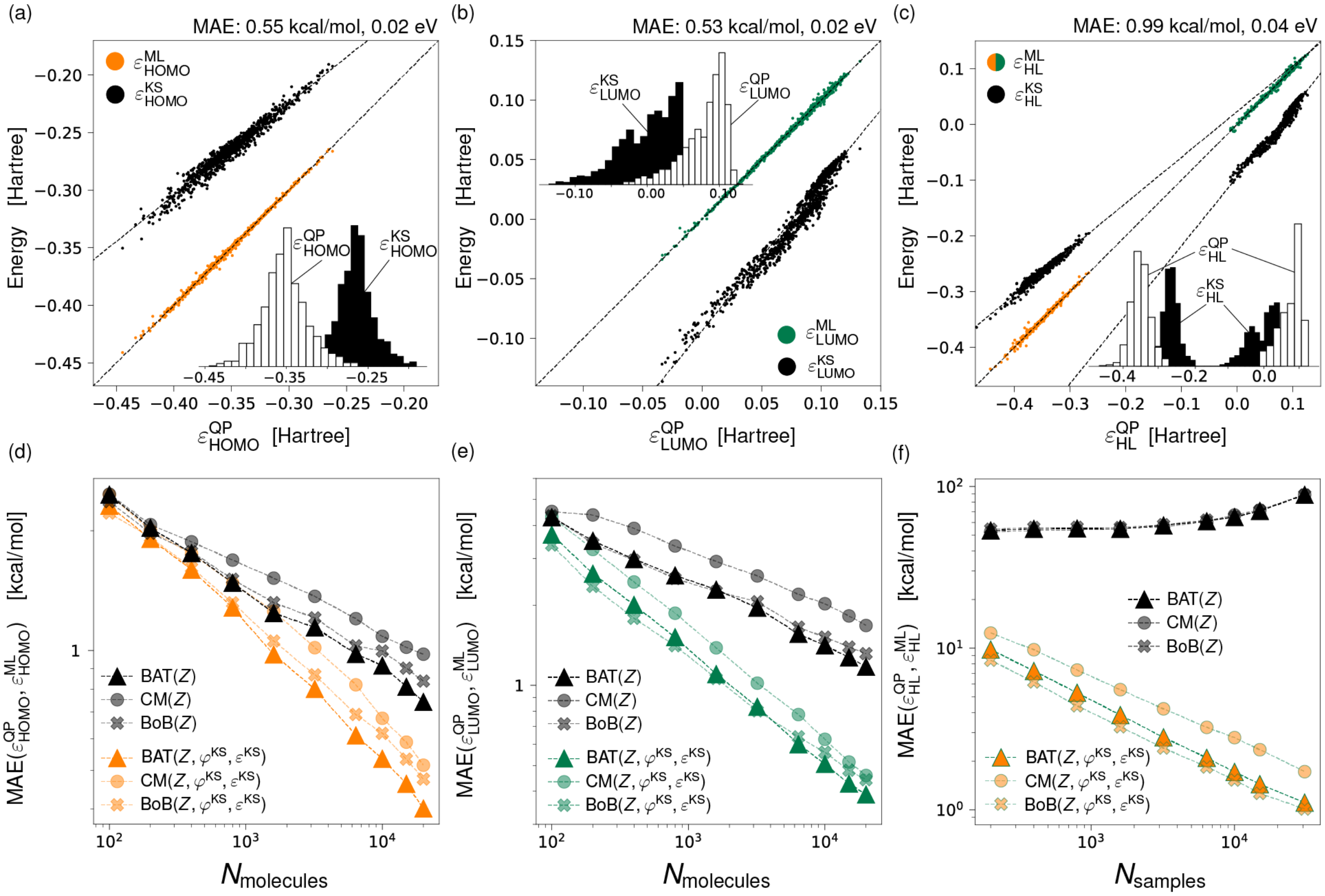

FIG. 2. Correlation plots and learning curves for Δ-Machine Learning HOMO, LUMO, and HOMO-LUMO energies of QM8 molecules. (a), (b) Correlation plots for single orbital energy level predictions with 1-OSR models, where KS/QP HOMO and LUMO energies are visualized in black and Δ-ML/QP HOMO and LUMO energies in orange and green, respectively. (c) A correlation plot for the simultaneous prediction of HOMO and LUMO energies in an 2-OSR model, where KS/QP HOMO/LUMO energies are represented in black and Δ-ML in orange/green. The inset in each correlation plot shows the histogram of QP (white) and KS (black) energies. (d)-(f) HOMO, LUMO, and HOMO-LUMO Δ-ML learning curves for QP predictions with various orbital independent (black) and orbital dependent representations (orange, green, and orange/green).

tically indistinguishable from the one based on explicit QP energies (shaded area).

Finally, we compare in Tab. I the MAE of predictions of quasiparticle energies obtained with generic and different orbital-sensitive representations. As also apparent from Fig. 2(d,e), the addition of orbital information into the representation reduces the MAE even for the single-orbital models by up to 50 %. Differences between CM, BoB, and BAT are very small in these cases. Regarding the use of different techniques to obtain effective orbital charges, we note that GDMA yields ~ 0.010 eV lower MAEs than Mulliken populations, and the overall lowest MAE is obtained for the BAT-GDMA combination. Moving to the 2-OSR- and Multi-OSR models, BoB-GDMA performs slightly better than BAT-GDMA and CM-GDMA. In the 2-OSR model comprising HOMO and LUMO, using Mulliken charges leads to a doubling of the MAE, albeit still lower than 0.1eV. For Multi-OSR, the MAE are expectedly a bit higher due to the high-dimensionality of the data but the differences between the use of GDMA and Mulliken charges are relatively smaller, in particular for BoB and BAT representations. While the Mulliken population based orbital-sensitive representations appear to yield a slightly higher MAE as compared to the GDMA-based ones, it should be stressed that the latter come with a higher computational cost, noticeable in particular for larger systems (see also Section III D), and are not widely available for standard DFT applications.

C. Water Monomers and Dimers

In the QM8 example we focused on predicting the full QP spectra of single molecules. In this example we want to show that a single Δ-ML model can predict the full QP spectra of water dimers and monomers simultaneously, i.e., that ΔMLQP is sensitive to intermolecular

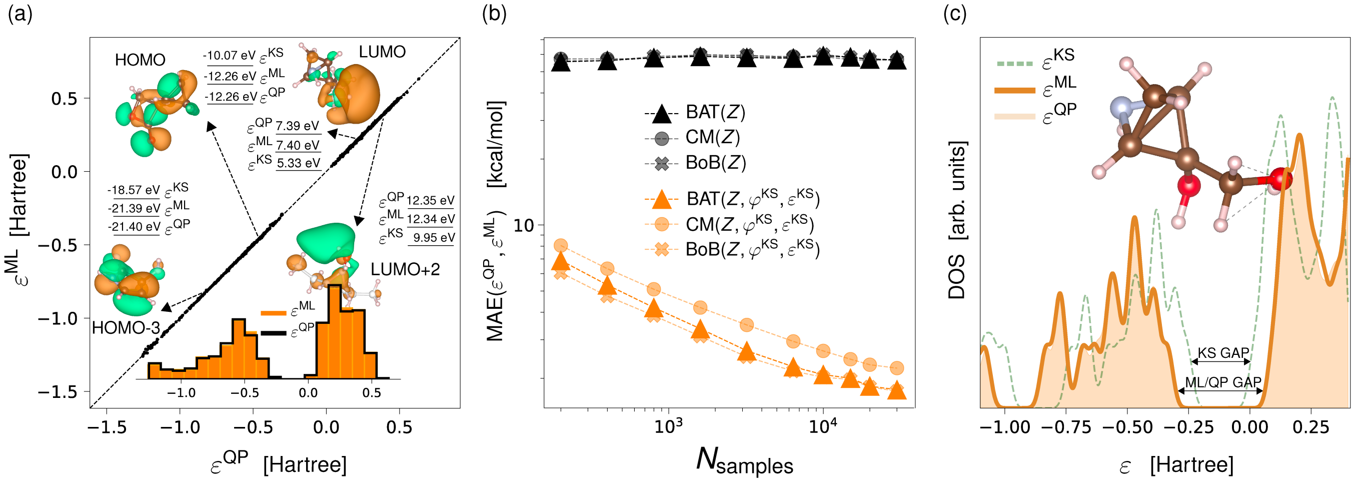

FIG. 3. Multi-OSR Δ-ML of full quasiparticle spectra of QM8 molecules with a single model. (a) Correlation plot of Δ-ML/QP energies, where the arrows are pointing at randomly chosen molecules. The inset shows the histogram of ML (orange) and QP (black) energies. (b) The density of states of a randomly chosen QM8 molecule, where the shaded area (light orange) represents the QP energies, while the green and orange lines describe the KS and ML energies, respectively. (c) Learning curves for QP energy predictions with various orbital independent (black) and orbital dependent representations (orange).

TABLE I. Mean Absolute Errors (in eV) of Predicting ev*GW* Quasiparticle Energies by Kernel Ridge Regression using standard CM, BoB and BAT representations, as well as our orbital-sensitive extension based on GDMA and Mulliken orbital charges, respectively.[a]

	HOMO	LUMO	2-OSR	Multi-OSR
CM	0.043	0.073	—	—
BoB	0.036	0.057	—	—
BAT	0.032	0.051	—	—
CM-GDMA	0.022	0.020	0.046	0.104
BoB-GDMA	0.021	0.019	**0.029**	**0.083**
BAT-GDMA	**0.017**	**0.017**	0.034	0.089
CM-Mulliken	0.031	0.038	0.109	0.155
BoB-Mulliken	0.026	0.033	0.070	0.108
BAT-Mulliken	0.024	0.030	0.072	0.107

[a] The best performing models are marked in bold.

conformations as well.

The molecular geometries used in this application originate from the H2O-13 dataset [27] that consists of 2000 water dimers with O-O distances less than 4.5 Å obtained from an MD simulation. All *GW* properties used in the training and testing set were calculated as mentioned for the QM8 data in the previous section. To build a Multi-OSR ΔMLQP model, we first consider not only the 2000 dimer structures but also extract 4000 monomers from them. In water dimers (monomers) 18 quasiparticle states are taken into account, so the orbital dataset we have used contains 52000 samples in total. From this set, 30000 randomly selected samples are used for training and 5K for testing.

Fig. 4(a) shows the QP-KS correlation (black) and QP-ML correlation (orange/green), respectively. The Multi-OSR ΔMLQP approach based on BAT is able to transform the Kohn-Sham energies to quasiparticle energies with great accuracy (MAE 0.03 eV). In Fig. 4(b), we plot the DOS of a randomly chosen water dimer and its constituent monomers broadened by 0.022 Hartree. On the shown scale, the small differences among the two monomers are hard to distinguish, independent of the method. More importantly, it is apparent that the ΔMLQP captures the respective openings of the HOMO-LUMO gaps and energy dependent quasiparticle corrections, as well as the effects of intermolecular interactions in the dimer conformation. This point is emphasized by the analysis of the difference between the actual dimer DOS and a simple superposition of the two monomer DOS as in the top panel of Fig. 4(c), evaluated based on the explicit QP energies (solid black) and the one predicted by ΔMLQP (dashed). These differences reveal the shifts of the coupled dimer energy levels due to intermolecular interactions and the effects are captured by ΔMLQP at very good accuracy not only near the gap but even for, e.g., the deep O2*s* levels. The lower panel in Fig. 4(c) shows monotonous decay of the prediction errors as a function of training set size, where orbital sensitive BoB and BAT perform slightly better than the orbital sensitive CM. Again, we see from the correlation plots, DOS and the learning curves indicate that the full QP spectra of water dimers and monomers can be accurately reproduced with just a single delta machine learning model.

D. Acetone in Water

We now consider an example of even more complex molecular clusters: aqueous acetone. The increased com-

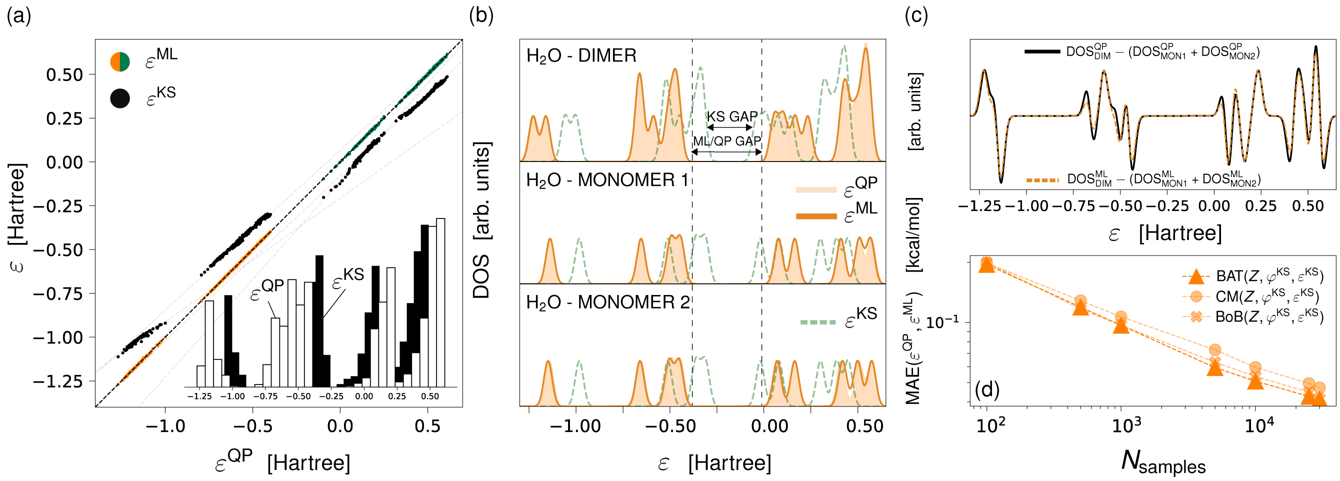

FIG. 4. Multi-OSR Δ-ML of full quasiparticle spectra of water monomers and dimers with a single model. (a) A correlation plot for the simultaneous prediction of all orbital energies, where KS/QP HOMO/LUMO energies are represented in black and Δ-ML in orange for the levels below the HOMO-LUMO gap and green for the levels above the gap. The inset shows the histogram of QP (white) and KS (black) energies. (b) Density of states plot for a randomly chosen water dimer and its monomers, where the shaded area (light orange) represents the QP energies, while the green and orange lines describe the KS and ML energies, respectively. (c) The difference between the dimer DOS and the monomer DOS' summed together, where the QP DOS is represented in black and ML in dashed orange. (d) Learning curves for QP energy predictions with various orbital dependent representations.

plexity stems from combining two different molecular species and considering more molecules in the clusters, leading to a very high-dimensional problem as the number of states and conformations increase dramatically. Specifically, the choice for aqueous acetone is motivated by the fact that it known to exhibit a solvatochromic shift of the lowest coupled electron-hole excitation energy of ∼ 0.2 eV[48–50], a combined effect of similar shifts to the individual quasiparticle energies and modified screening of the electron-hole interaction in water. From the perspective of our ΔMLQP model, this poses the additional question of whether its predictions are accurate enough in such a case to embed them into the calculation of the electron-hole excitation energies via the Bethe–Salpeter Equation (BSE@ML vs BSE@GW), as noted in the workflow scheme in Fig. 1(a).

To answer this question, we first generate structural data by performing classical Molecular Dynamics simulations of a single acetone in 219 water molecules using an OPLS-AA type forcefield for acetone, automatically generated by LigParGen [51], and the TIP3P model for water [52]. Geometric mixing rules for Lennard-Jones diameters and energies were used for atoms of different species [53]. Non-bonded interactions between atom pairs within a molecule separated by one or two bonds were excluded. Interaction was reduced by a factor of 1/2 for atoms separated by three bonds and more. Simulations were run using GROMACS version 2019.6 [54]. A 0.9 nm cutoff was employed for the real space part of electrostatics and Lennard-Jones interactions. The long-range electrostatics were calculated using particle-mesh Ewald (PME) [55] with the reciprocal-space interactions evaluated on a 0.18 grid with cubic interpolation of order 4. An initial configuration was prepared in cubic box of size 2 nm and energy minimized using the steepest descents algorithm, followed by a 6 ns simulation in constant particle number, volume and temperature (NVT) ensemble. Temperature was kept constant at 300 K using the stochastic velocity rescaling thermostat [56] with time constant 0.5 ps, and the velocity-Verlet algorithm [57] was employed to integrate the equations of motions with 1 fs time step. Simulations were then continued for 200 ns in constant particle number, pressure and temperature (NpT) ensemble at 300 K and 1 bar controlled by the Berendsen [58] barostat with a coupling time constant of 2.0 ps. From the last 100 ns of this run, 10000 snapshots at a time interval of 10 ps are extracted and clusters containing acetone and the ten water molecules closest to it are selected for the GW-BSE calculations. With this choice we ascertain that the first solvation shell is included in the cluster, contributing the strongest to the expected solvatochromic shifts. Note that we are not strictly targeting quantitative accuracy of the actual excitation energies compared to experiment with this study, but to demonstrate the internal consistency between BSE@GW and BSE@ML in our Multi-OSR ΔMLQP approach.

GW-BSE calculations are performed on the selected clusters without and with the water molecules included, following the same protocol as before, with the exception of the treatment of the frequency-dependence in Eq. 4, for which we employ here a two-parameter generalized Plasmon-Pole Model [38, 39, 59]. Explicit quasiparticle corrections are determined for the four highest occupied

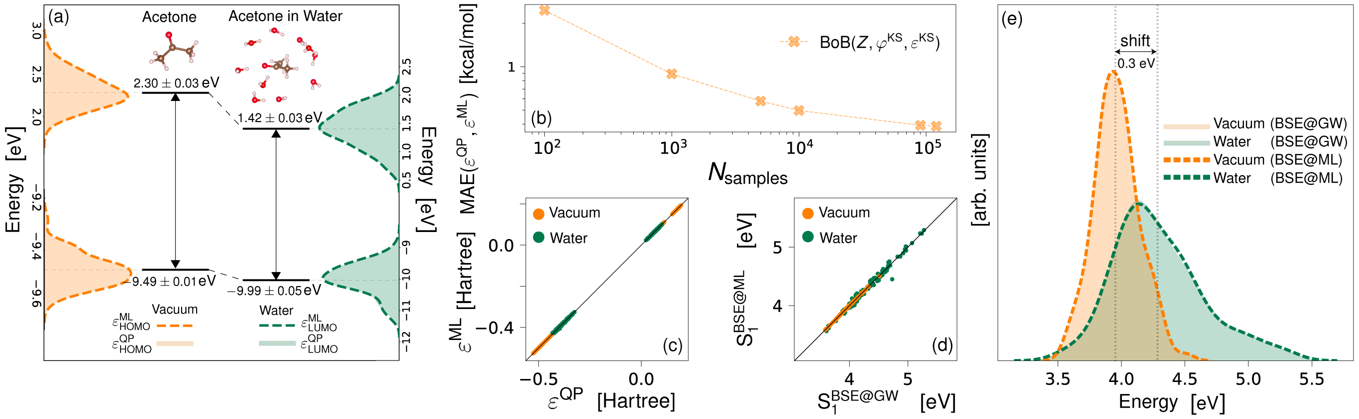

FIG. 5. Multi-OSR Δ-ML of full quasiparticle spectra of acetone and acetone in water with a single model and application to calculating $n \rightarrow \pi^*$ excitation energies with BSE. (a) The black bars are a visual representation of the HOMO-LUMO gap in acetone in vacuum (orange) and acetone in water (green) with ML predictions. The shaded areas represent the HOMO and LUMO distributions as resulting from explicit evGW calculations, while the one resulting from the 2-OSR ΔMLQP model are shown as dashed lines. (b) Prediction errors as a function of training set size for the 8-OSR-BoB model to be used in BSE@ML calculations. (c) QP-ML quasiparticle energy correlation for 8-OSR-BoB. (d) Correlation between S_1 ($n \rightarrow \pi^*$) energies from BSE@GW and BSE@ML. (e) S_1 energy distributions of a single acetone (orange) and acetone in water (green) from BSE@GW and BSE@ML.

and four lowest unoccupied molecular orbitals, while the full single-particle spectrum is included in the formation of the product basis for the BSE.

We first construct as a proof-of-concept a 2-OSR-BoB model including HOMO and LUMO in the absence and presence of a water solvation shell. From the total of 80000 samples, we select 5000 for training and 1000 for testing, ensuring that the sets have an equal amount of data for both HOMO and LUMO, with and without water, respectively. The energy distributions as obtained from the explicit calculations of ε^{QP} in vacuum (in water), calculated on 400 (100 of each case) out of sample data points, are shown as filled orange (green) areas in Fig. 5(a). Indicated are also the means and their error, showing a distinct lowering of 0.50 eV and 0.92 eV for the HOMO and LUMO, respectively, in the presence of water, thereby decreasing the HOMO-LUMO gap by 0.42 eV. Comparing the respective distributions obtained from the 2-OSR ΔMLQP model shown as dashed lines in Fig. 5(a), hardly any differences can be observed.

As mentioned above, the determination of electron-hole excitation energies with BSE@GW requires inclusion of eight explicitly corrected orbitals near the gap. For this purpose, we now build an 8-OSR ΔMLQP model, for which Fig. 5(b) shows the learning curves. With more than 10K samples, a MAE of lower than 0.5 kcal/mol can be achieved. Figure 5(c) shows the corresponding correlation between explicitly calculated QP energies, and the ones from 8-OSR ΔMLQP, for which we have selected from the full dataset comprising 320K orbitals, 120K for training and 10K for testing. Clearly, the ΔMLQP model again provides excellent predictions for the eight different orbital energies in vacuum and solution, respectively. In the following step, these predicted quasiparticle energies for the eight explicitly corrected orbitals are used as input for the BSE. The remaining occupied (unoccupied) single-particle energies are scissors shifted by according to the highest absolute quasiparticle correction among the explicitly corrected occupied (unoccupied) orbitals, as in the BSE@GW reference. In Fig. 5(d) we show the correlation between the determined S_1 energies of the $n \rightarrow \pi^*$ transition. Two interesting aspects should be noted: the subset of results for BSE@ML for the vacuum structure (orange) appear to agree better than the one for the acetone-water clusters (green), which is not surprising due to the bigger conformational space of the latter. More importantly, however, one can clearly see a systematic shift of the S_1 energies in aqueous solution to higher energies compared to in vacuo – equally obtained for both BSE@GW and BSE@ML – qualitatively in line with the experimental observations.

Finally, we show in Fig. 5(e) distributions of the $n \rightarrow S_1$ excitation energies of acetone as obtained by the BSE@GW (filled curves) and BSE@ML (dashed lines) approaches in vacuum (orange) and in water (green), respectively. It is evident that both methods predict both a broadening of the distribution upon solvation, as well as a shift to higher energies. Differences of the distributions on BSE@GW and BSE@ML levels are miniscule. The predicted solvatochromic shift of the mean of the distributions amount to 0.30 eV, as indicated by the dashed lines in Fig. 5(e). From peak-to-peak, the shift is 0.13 eV.

IV. SUMMARY

We have introduced orbital-sensitive augmentations of graph-based representations in Δ-machine learning of full quasiparticle excitation energies in molecules and clusters. The proposed ΔMLQP approach is capable of predicting the *GW* energies of multiple orbitals across multiple molecules and/or intra- and inter molecular conformations in a single kernel ridge regression based supervised learning model. We have demonstrated this in application to the QM8 molecular dataset and to water monomers and dimers. Furthermore, it has been shown that a single orbital-sensitive Δ-ML model for quasiparticle energies can be embedded in multiscale simulation workflows, showcased in the prediction of solvatochromic shifts of excitation energies in aqueous acetone.

DATA AVAILABILITY

The raw data of the *GW* calculations on the QM8 dataset, water monomers and dimers, as well as aqueous acetone are available as VOTCA-XTP output files at https://dx.doi.org/10.6084/m9.figshare.14625564.

V. ACKNOWLEDGEMENT

We acknowledge support by the Innovational Research Incentives Scheme Vidi of the Netherlands Organisation for Scientific Research (NWO) with project number 723.016.002. Partial funding is also provided by NWO and the Netherlands eScience Center for funding through project number 027.017.G15, within the Joint CSER and eScience program for Energy Research (JCER 2017). We are thank Ruben Gerritsen for a critical reading of the manuscript.

[1] von Lilienfeld, O. A. In *Many-Electron Approaches in Physics, Chemistry and Mathematics: A Multidisciplinary View*; Bach, V., Delle Site, L., Eds.; Mathematical Physics Studies; Springer International Publishing: Cham, 2014; pp 169–189.

[2] Huang, B.; Lilienfeld, O. A. V. Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *J. Chem. Phys.* **2016**, *145*, 241730.

[3] Bartók, A. P.; Kermode, J.; Bernstein, N.; Csányi, G. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Phys. Rev. X* **2018**, *8*, 041048.

[4] Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. *J. Chem. Phys.* **2018**, *148*, 241718.

[5] Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K. R.; Maurer, R. J. Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions. *Nat. Commun.* **2019**, *10*, 1–10.

[6] Çaylak, O.; Yaman, A.; Baumeier, B. Evolutionary Approach to Constructing a Deep Feedforward Neural Network for Prediction of Electronic Coupling Elements in Molecular Materials. *J. Chem. Theory Comput.* **2019**, *15*, 1777–1784.

[7] Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Adv. Sci.* **2019**, *6*, 1801367.

[8] Çaylak, O.; von Lilienfeld, O. A.; Baumeier, B. Wasserstein Metric for Improved Quantum Machine Learning with Adjacency Matrix Representations. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 03LT01.

[9] von Lilienfeld, O. A.; Burke, K. Retrospective on a Decade of Machine Learning for Chemical Discovery. *Nat. Commun.* **2020**, *11*, 1–4.

[10] Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chem. Mater.* **2020**, *32*, 4954–4965.

[11] Song, Z.; Chen, X.; Meng, F.; Cheng, G.; Wang, C.; Sun, Z.; Yin, W.-J. Machine Learning in Materials Design: Algorithm and Application. *Chinese Phys. B* **2020**, *29*, 116103.

[12] Tkatchenko, A. Machine Learning for Chemical Discovery. *Nat. Commun.* **2020**, *11*, 4125.

[13] von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. *Nat. Rev. Chem.* **2020**, *4*, 347–358.

[14] Lemm, D.; von Rudorff, G. F.; von Lilienfeld, O. A. Energy-Free Machine Learning Predictions of {*ab Initio*} *Structures*. **2021**,

[15] Tirimbó, G.; Çaylak, O.; Baumeier, B. A Kernel-Based Machine Learning Approach to Computing Quasiparticle Energies within Many-Body Green's Functions Theory. *arXiv* **2020**,

[16] Juan, Y.; Dai, Y.; Yang, Y.; Zhang, J. Accelerating Materials Discovery Using Machine Learning. *Journal of Materials Science & Technology* **2021**, *79*, 178–190.

[17] Rauer, C.; Bereau, T. Hydration Free Energies from Kernel-Based Machine Learning: Compound-Database Bias. *J. Chem. Phys.* **2020**, *153*, 014101.

[18] Dong, S. S.; Govoni, M.; Galli, G. Machine Learning Dielectric Screening for the Simulation of Excited State Properties of Molecules and Materials. *Chem. Sci.* **2021**, *12*, 4970–4980.

[19] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Lilienfeld, O. A. V. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

[20] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Lilienfeld, O. A. V. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

[21] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Lilienfeld, O. A. V.; Müller, K. R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

[22] Meftahi, N.; Klymenko, M.; Christofferson, A. J.; Bach, U.; Winkler, D. A.; Russo, S. P. Machine Learning Property Prediction for Organic Photovoltaic Devices. *Npj Comput. Mater.* **2020**, *6*, 1–8.

[23] Liu, Z.; Lin, L.; Jia, Q.; Cheng, Z.; Jiang, Y.; Guo, Y.; Ma, J. Transferable Multilevel Attention Neural Network for Accurate Prediction of Quantum Chemistry Properties via Multitask Learning. *J. Chem. Inf. Model.* **2021**, *61*, 1066–1082.

[24] Rupp, M.; Tkatchenko, A.; Müller, K. R.; Lilienfeld, O. A. V. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

[25] Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

[26] Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

[27] Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-Learning Approach for One- and Two-Body Corrections to Density Functional Theory: Applications to Molecular and Condensed Water. *Phys. Rev. B* **2013**, *88*, 054104.

[28] Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

[29] Hedin, L. New Method for Calculating the One-Particle Green's Function with Application to the Electron-Gas Problem. *Phys. Rev.* **1965**, *139*, A796–A823.

[30] Hedin, L.; Lundqvist, S. In *Solid State Physics*; Seitz, F., Turnbull, D., Ehrenreich, H., Eds.; Academic Press, 1970; Vol. 23; pp 1–181.

[31] Aulbur, W. G.; Jönsson, L.; Wilkins, J. W. In *Solid State Physics*; Ehrenreich, H., Spaepen, F., Eds.; Academic Press, 2000; Vol. 54; pp 1–218.

[32] Rohlfing, M. Excited States of Molecules from Green's Function Perturbation Techniques. *Int. J. Quantum Chem.* **2000**, *80*, 807–815.

[33] Hybertsen, M. S.; Louie, S. G. First-Principles Theory of Quasiparticles: Calculation of Band Gaps in Semiconductors and Insulators. *Phys. Rev. Lett.* **1985**, *55*, 1418–1421.

[34] Golze, D.; Wilhelm, J.; van Setten, M. J.; Rinke, P. Core-Level Binding Energies from GW: An Efficient Full-Frequency Approach within a Localized Basis. *J. Chem. Theory Comput.* **2018**, *14*, 4856–4869.

[35] Golze, D.; Dvorak, M.; Rinke, P. The GW Compendium: A Practical Guide to Theoretical Photoemission Spectroscopy. *Front. Chem.* **2019**, *7*.

[36] Tirimbò, G.; Sundaram, V.; Çaylak, O.; Scharpach, W.; Sijen, J.; Junghans, C.; Brown, J.; Ruiz, F. Z.; Renaud, N.; Wehner, J.; Baumeier, B. Excited-State Electronic Structure of Molecules Using Many-Body Green's Functions: Quasiparticles and Electron–Hole Excitations with VOTCA-XTP. *J. Chem. Phys.* **2020**, *152*, 114103.

[37] Caylak, O.; Baumeier, B. Excited-State Geometry Optimization of Small Molecules with Many-Body Green's Functions Theory. *J. Chem. Theory Comput.* **2021**, *17*, 879–888.

[38] Hybertsen, M. S.; Louie, S. G. Electron Correlation in Semiconductors and Insulators: Band Gaps and Quasiparticle Energies. *Phys. Rev. B* **1986**, *34*, 5390–5413.

[39] Godby, R. W.; Needs, R. J. Metal-Insulator Transition in Kohn-Sham Theory and Quasiparticle Theory. *Phys. Rev. Lett.* **1989**, *62*, 1169–1172.

[40] Rohlfing, M.; Louie, S. G. Electron-Hole Excitations and Optical Spectra from First Principles. *Phys. Rev. B* **2000**, *62*, 4927–4944.

[41] Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; Lilienfeld, O. A. V. Electronic Spectra from TDDFT and Machine Learning in Chemical Space. *J. Chem. Phys.* **2015**, *143*, 84111.

[42] Rupp, M. Machine Learning for Quantum Mechanics in a Nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.

[43] Zaspel, P.; Huang, B.; Harbrecht, H.; Lilienfeld, O. A. V. Boosting Quantum Machine Learning Models with a Multilevel Combination Technique: Pople Diagrams Revisited. *J. Chem. Theory Comput.* **2019**, *15*, 1546–1559.

[44] Wehner, J.; Brombacher, L.; Brown, J.; Junghans, C.; Çaylak, O.; Khalak, Y.; Madhikar, P.; Tirimbò, G.; Baumeier, B. Electronic Excitations in Complex Molecular Environments: Many-Body Green's Functions Theory in VOTCA-XTP. *J. Chem. Theory Comput.* **2018**, *14*, 6253–6268.

[45] Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *The Journal of Chemical Physics* **1999**, *110*, 6158–6170.

[46] Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

[47] Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: Optimized Auxiliary Basis Sets and Demonstration of Efficiency. *Chemical Physics Letters* **1998**, *294*, 143–152.

[48] Bayliss, N. S.; McRae, E. G. Solvent Effects in the Spectra of Acetone, Crotonaldehyde, Nitromethane and Nitrobenzene. *J. Phys. Chem.* **1954**, *58*, 1006–1011.

[49] Merchán, M.; Roos, B. O.; McDiarmid, R.; Xing, X. A Combined Theoretical and Experimental Determination of the Electronic Spectrum of Acetone. *J. Chem. Phys.* **1996**, *104*, 1791–1804.

[50] Crescenzi, O.; Pavone, M.; De Angelis, F.; Barone, V. Solvent Effects on the UV (n → Π*) and NMR (13C and 17O) Spectra of Acetone in Aqueous Solution. An Integrated Car-Parrinello and DFT/PCM Approach. *J. Phys. Chem. B* **2005**, *109*, 445–453.

[51] Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen Web Server: An Automatic OPLS-AA Parameter Generator for Organic Ligands. *Nucleic Acids Research* **2017**, *45*, W331–W336.

[52] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[53] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[54] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

[55] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[56] Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

[57] Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98–103.

[58] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[59] Rohlfing, M.; Krüger, P.; Pollmann, J. Efficient Scheme for GW Quasiparticle Band-Structure Calculations with Applications to Bulk Si and to the Si(001)-(2x1) Surface. *Phys. Rev. B* **1995**, *52*, 1905–1917.