



Laser-pulse characterization using strong-field autocorrelation patterns and random-forest-based machine learning

Daria Kolbasova ^{1,*} and Robin Santra ^{1,2}

¹*Center for Free-Electron Laser Science CFEL, Deutsches Elektronen-Synchrotron DESY, 22607 Hamburg, Germany*

²*Department of Physics, Universität Hamburg, 22607 Hamburg, Germany*



(Received 23 September 2022; accepted 10 January 2023; published 24 January 2023)

Building on the strategy presented in [Opt. Lett. **47**, 3992 \(2022\)](#), we demonstrate an efficient alternative approach for the *in situ* characterization of ultrashort low-frequency laser pulses. In this context, we employ first-principles quantum-mechanical calculations to model the strong-field ionization of rare-gas atoms and produce autocorrelation patterns for a set of few-femtosecond near-infrared laser pulses. We explore the nonperturbative and nonlinear dependence of the autocorrelation patterns on the pulse characteristics and postulate an analytical function describing these patterns. For every laser pulse considered, we employ the parameters appearing in this analytical function, together with the underlying pulse parameters for supervised machine learning. Specifically, we use the random-forest technique for retrieving key laser pulse parameters from autocorrelation patterns produced via strong-field ionization. The current approach offers advantages for application to experimental data.

DOI: [10.1103/PhysRevA.107.013520](https://doi.org/10.1103/PhysRevA.107.013520)

I. INTRODUCTION

The key mechanisms that provide ways to observe and control the complex dynamics inside atoms and molecules are electron excitations and ionizations, induced by ultrashort laser pulses [1]. The rapid technological progress in building lasers with novel pulse properties [2–6] has opened up the possibility to explore fundamental mechanisms in chemical and physical processes in a time-resolved manner [7] and to study fundamental questions related to the quantum dynamics of electrons on their natural timescale. Thus, a comprehensive characterization of such laser pulses is of great importance. Autocorrelation-based techniques provide a way to determine the spectrum and duration of reproducible laser pulses. The autocorrelation of ultrashort low-frequency fields can be measured through the strong-field photoionization of a rare-gas target injected into the interaction region [8]. In Ref. [9], a technique was presented that allows a precise characterization of the femtosecond laser field in the interaction region using machine-learning algorithm called vector space Newton interpolation cage (VSNIC). Training data were generated using first-principles calculations. The key assumption in Ref. [9] is that the experimental autocorrelation pattern and all training data for VSNIC are sampled at precisely the same time points. This would not be a serious problem if the generation of training data could be performed very quickly for a large number of time points. However, when using first-principles calculations, sampling autocorrelation patterns on a dense grid of time points is impractical.

To overcome this problem, in this paper we explore the shape of the autocorrelation patterns and postulate an

analytical function that allows us to fit the autocorrelation patterns with good accuracy. In this way, the information content of each autocorrelation pattern is reduced to a few fit parameters. Also we present a modified machine-learning-based strategy. Using the random-forest technique, we retrieve information about the pulse parameters directly from the parameters of the aforementioned analytical function. This allows us to reduce the machine-learning problem from a relatively high-dimensional feature space (determined by the number of time points in an autocorrelation pattern) to a much lower-dimensional feature space (determined by the number of fit parameters).

II. AUTOCORRELATION PATTERNS

The process of ionization by a strong field is nonperturbative and nonlinear. Thus, it is difficult to analytically derive a universally valid expression for the structure of strong-field-generated autocorrelation patterns. In order to investigate the structure of the autocorrelation function and its dependence on the field characteristics, reliable first-principles calculations are required. We performed simulations of strong-field-induced autocorrelation patterns of atomic argon (Ar) using the configuration-interaction dynamics (XCID) code [10,11]. XCID is based on the time-dependent configuration-interaction-singles (TDCIS) approach for solving the many-electron time-dependent Schrödinger equation from first principles, and has already proven its qualitative and quantitative accuracy in strong-field multiphoton ionization calculations in a number of works [12–15].

The sensitivity of the experiment increases with the size of the sample atom: the lower the ionization potential, the more sensitive is strong-field ionization to the wings of the pulse to be characterized. On the other hand, the speed and

*daria.kolbasova@desy.de

accuracy of TDCIS numerical calculations increase for lighter atoms. Since in this paper we focus on the characterization of few-fs near-infrared pulses, which are commonly used in ultrafast spectroscopy [16] and for strong-field ionization of atoms and molecules [17,18], the choice of Ar represents a reasonable compromise between experimental sensitivity and computational accuracy. However, as will be discussed below, the proposed technique can also be applied to other wavelength or pulse-intensity ranges using other targets.

The ground electronic configuration for Ar is $[\text{Ne}]3s^23p^6$. In our calculations, we only allow the $3p$ and $3s$ orbitals to be active, whereas, all other orbitals are frozen (not affected by the laser field). A characterization of a laser pulse in terms of just a few key parameters requires a simple analytical description of a laser pulse. To this end, we use for the pulse electric field in the frequency domain the expression,

$$E(\omega) = E_0 \exp \left[- \left(\frac{4 \ln(2)}{\chi_\omega^2} - \frac{ik}{2} \right) (\omega - \omega_0)^2 + i\varphi \right], \quad (1)$$

where E_0 is the field amplitude, ω_0 is the central frequency of the field, χ_ω is the full width at half maximum (FWHM) of $|E(\omega)|$, k identifies the group delay dispersion (GDD), and φ is the carrier-envelope phase. The chirp associated with the GDD dramatically affects the pulse duration. Following Eq. (1), the actual FWHM of laser pulse duration is given by

$$X = \sqrt{\frac{\chi_t^4 + 8^2 \ln(2)^2 k^2}{\chi_t^2}}, \quad (2)$$

where $\chi_t = 4 \ln(2)/\chi_\omega$ is the duration the pulse has if it is transform-limited ($k = 0$).

III. PATTERN ANALYSIS

Using the TDCIS approach we obtain the autocorrelation patterns via calculating the ionization probability of Ar exposed to a pair of identical laser pulses with a central frequency $\omega = 0.061$ a.u. (1.66 eV), delayed relative to each other. We allow pulses to differ from each other only through their relative phase difference $\Delta\phi = \phi_1 - \phi_2$. The calculations are performed using a large set of pulse parameters: The pulse duration is considered in the range of $4.84 \leq \chi_t \leq 13.31$ fs, and the GDD in the range of $0 \leq k \leq 20$ fs², and the phase difference $\Delta\phi$ varies from 0 to π .

Over a range of intensities, the dependence of the ionization probability on the intensity of the pulse can be described by a power law,

$$P(I) = CI^n, \quad (3)$$

where C and n are parameters. This is illustrated in Fig. 1. For intensities below 8 TW/cm² the same parameter n can be used for different time delays, i.e., the delay dependence is fully contained in the parameter C . In other words, in the range of intensities considered in the present paper, the shape of autocorrelation patterns is determined exclusively by the parameter C . Thus, these shapes are insensitive to volume-integration (spatial-averaging) effects.

The binding energy for the outer-valence shell of Ar is $E_b \approx 15.7$ – 15.9 eV. Thus, at a photon energy of 1.66 eV, ionization requires the absorption of, at least, 9 to 10 photons.

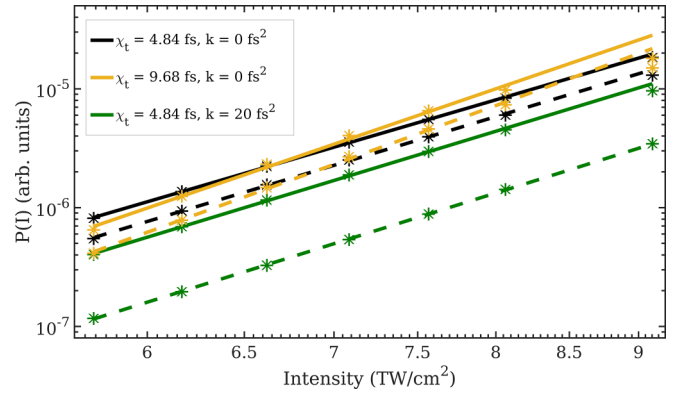


FIG. 1. Dependence of the ion yield $P(I)$ on the field intensity I for two time delays: $\tau = 0$ (solid line), $\tau = 2.42$ fs (dashed line). The data satisfy the power-law $P(I) = CI^n$, where $n \approx 7$ for the pulse duration $\chi_t = 4.84$ fs and $n \approx 8$ for $\chi_t = 9.68$ fs.

However, as the short pulses provide a broad photon energy spectrum, the ionization may on occasion require a smaller number of photons, leading to $n \leq 9$ in the power law in Eq. (3). Since the chirp does not change the spectrum of the pulse, $n \approx 7$ for the pulse duration $\chi_t = 4.84$ fs for both $k = 0$ and $k = 15$. The broader the spectrum, the smaller is the number of photons required for ionization. For intensities higher than 8 TW/cm² the shape of the autocorrelation patterns gets deformed due to saturation effects, and the central peaks are significantly lower than predicted by the power law [Eq. (3)].

IV. APPROXIMATION FORMULA

Although the amplitude of the autocorrelation pattern is determined by the intensity of the field, its actual shape is determined by the duration and interference properties of the sum of the electric fields of the two pulse copies,

$$E_{\text{sum}}(t, \tau) = E_1(t) + E_2(t + \tau), \quad (4)$$

at each given τ . This complicated dependence can be captured (for pulses in aforementioned ranges) by the relatively simple approximation function,

$$P'(\tau) = A(\tau, t_0)^{n'G(\tau-t_0)^m}, \quad (5)$$

where

$$G(\tau - t_0) = \exp[-4 \ln(2)(\tau - t_0)^2/\sigma^2] \quad (6)$$

is a Gaussian envelope with a full width at half maximum σ and

$$A(\tau, t_0) = \left[G(\tau - t_0) \cos \left(\frac{\omega'}{2} (\tau - t_0) + \frac{\phi'}{2} \right) \right]^2. \quad (7)$$

We have introduced the parameter t_0 for experimental situations in which it cannot be guaranteed that perfect temporal overlap between the two pulse copies corresponds to zero time delay. The oscillation frequency ω' and phase ϕ' are approximately the same as the central photon energy ω_0 and the phase difference $\Delta\phi$. These parameters can be set fixed when fitting, or treated as free parameters if unknown. In the second case it is useful to provide, at least, a guess for the central frequency ω_0 with a certain accuracy, for example,

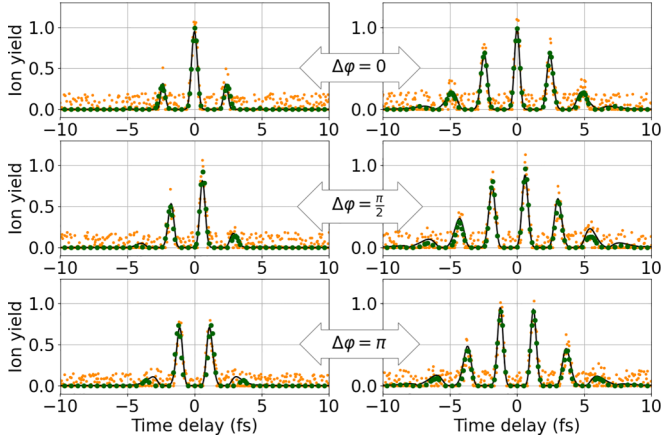


FIG. 2. Approximation of the calculated ion yield $P(\tau)$ contaminated by 20% noise (light dots) with the fitting function $P'(\tau)$ (black line) for $\chi_t = 4.84$ fs with $k = 0$ (left panels) and $k = 20$ fs² (right panels) and different phase-shifts $\Delta\phi$. Dark dots represent the actual ion yield $P(\tau)$ without the noise.

10%. This makes the fitting algorithm more stable and allows one to predict the actual value of ω_0 with an accuracy of 0.5%. The parameters σ , n' , and m are free parameters that have to be adjusted for the best fit of $P'(\tau)$ with the computed or experimentally measured points of $P(\tau)$. To find the best-fitting parameters, we use the nonlinear least-squares minimization (LMFIT) package in PYTHON [19].

In a real experiment, the measured autocorrelation pattern is inevitably contaminated by noise. In Fig. 2, we illustrate that for well-sampled autocorrelation data with a time-delay step of $\Delta\tau < 0.0484$ fs, our approach is able to recover the ion yield $P(\tau)$ with good accuracy even at a noise level of 20%.

For the machine learning algorithm it is better to switch from the standard pulse description through duration χ_t and chirp k to the actual pulse duration X and spectral FWHM χ_ω . As illustrated in Table I, this approach allows us to create a mapping between the laser pulse parameters (input) and the parameters characterizing the corresponding autocorrelation pattern (output). The task of machine learning is to compute an inverse mapping, such that from a given autocorrelation pattern the laser pulse parameters can be reconstructed.

TABLE I. Training set for the machine learning algorithm: laser parameters X , χ_ω , and $\Delta\phi$ and the fit parameters of the corresponding autocorrelation pattern.

X (fs)	χ_ω (eV)	$\Delta\phi$	σ (fs)	ϕ'	n'	m
9.45	0.59	0.23	7.16	0.21	4.99	0.276
6.84	0.53	0	5.23	0	7.45	-0.008
13.3	0.53	0	9.79	0	7.12	0.116
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
19.5	0.19	0.53	12.9	0.49	9.96	0.129
20.0	0.19	0.53	13.2	0.51	9.95	0.139

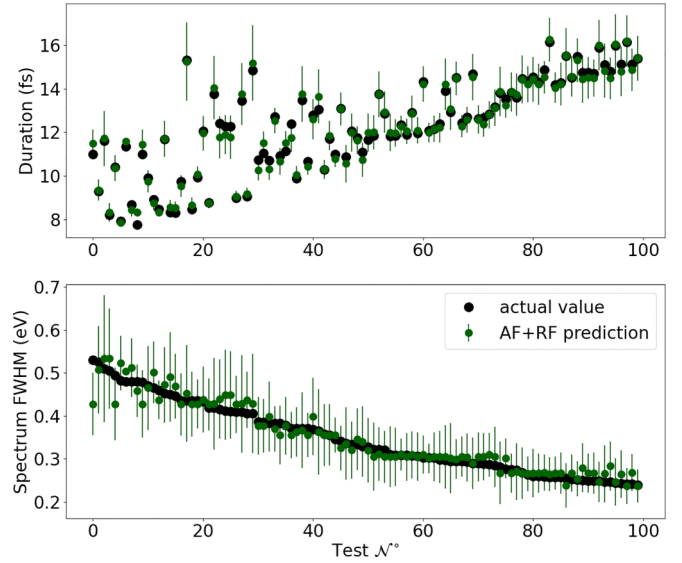


FIG. 3. Comparison of the actual duration X , and spectrum width χ_ω with the predicted values for 100 randomly generated pulses.

V. PULSE CHARACTERIZATION

In general, we do not know *a priori* which parameters of the approximation function (AF) in Eq. (5) are significant for retrieving relevant pulse parameters, and how they depend on each other. Thus, we chose to apply the random-forest (RF) technique [20–22]. A RF is a decision-tree-based ensemble with each decision tree depending on a collection of random variables (random set of AF parameters). For a p -dimensional vector $\mathbf{X} = (X_1, \dots, X_p)$ representing the real-valued input and an output variable Y representing the real-valued response (a pulse parameter of interest) the goal is to find a prediction function $f(\mathbf{X})$ for predicting Y . The algorithm finds correlations [23] between the features \mathbf{X} and the target Y in the training data. Each prediction function (decision tree) is determined by a loss function and defined to minimize the expected value of the loss. The algorithm is combined with a series of tree regressors, each tree casting a unit vote for the most popular regressor. Then the results from all trees are combined and averaged to give the final result. This scheme makes it possible to improve the accuracy and avoid overfitting. We fed our database of computed AF parameters together with the corresponding laser field parameters (Table I) to the scikit-learn [24] implementation of the RF. In order to demonstrate the validity of our approach in terms of final predictions, we compare in Fig. 3 our results (AF + RF) with the actual parameters for a family of test cases not contained in the training data.

Using the 100 test cases underlying Fig. 3, we present in Table II the average prediction error for the pulse duration X , spectrum width χ_ω , phase difference $\Delta\phi$, and the MSE for the resulting shape of the electric-field $E(t)$ for test data with and without noise. We compare them with the VSNIC results from Ref. [9], obtained for the same data set of autocorrelation patterns, but avoiding usage of the approximation function. Although the AF + RF procedure leads to somewhat higher errors compared to VSNIC [9], the accuracy of pulse

TABLE II. Average prediction error and mean squared error (MSE) of the reconstructed pulse. Comparison between the present method (AF + RF) and the method employed in Ref. [9] (VSNIC applied directly to discretely sampled autocorrelation patterns).

Method	X (fs)	χ_ω (eV)	$\Delta\phi$	$\text{MSE}_{E(t)}$
VSNIC	0.05 (0.4%)	0.02 (6%)	0.4°	0.8%
AF + RF	0.50 (4%)	0.04 (11%)	0.7°	1.2%
AF + RF (20% noise)	0.59 (4.8%)	0.05 (15%)	1.3°	1.6%

characterization via AF + RF is still satisfactory. The fit function gives the considerable advantage that the autocorrelation vectors in the database and the unlabeled (experimental) data vectors do not need to refer to the same time points. The algorithm does not require well-defined zero time delay nor the same temporal spacings as used in the experimental data. Moreover, AF + RF performs well even in the presence of 20% noise. One can see from Table II that the accuracy of the reconstructed pulse properties does not decrease significantly when dealing with noisy data. In Fig. 4 we present a comparison of the actual pulse intensity $I(t) = E^2(t)$ used in our TDCIS simulations with the pulse shape reconstructed by our AF + RF method using autocorrelation patterns with and without noise. As an example we chose two test pulses with a prediction error above the average in order to demonstrate that the accuracy is good even for unfavorable cases.

However, the approximation function $P'(\tau)$ [Eq. (5)] is rather simple and does not capture all features of $P(\tau)$, particularly, for pulses with significant chirp. As shown in Fig. 5 for a pulse with $\chi_t = 2.42$ fs and $k \approx 23$ fs², the wings of the autocorrelation pattern ($|\tau| \geq 6$ fs) cannot be described by a single-frequency function. Thus, a better fit function may be required when considering broadband chirped pulses. Exploiting an RF technique makes our method suitable for more sophisticated approximation functions, which could have more fitting parameters or/and which may have

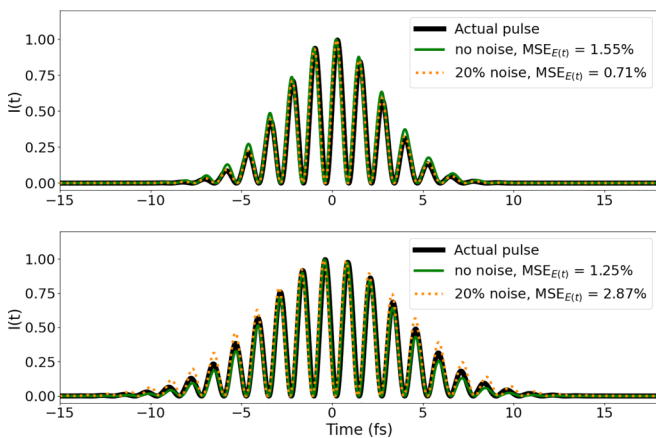


FIG. 4. Comparison, for two selected test cases not included in the training data of the actual intensity $I(t)$ of each pulse used in our simulations of $P(\tau)$ with the pulse shape found by our AF + RF method, applied to the test data with and without noise. $\text{MSE}_{E(t)}$ stands for the mean-squared error of the electric-field $E(t)$.

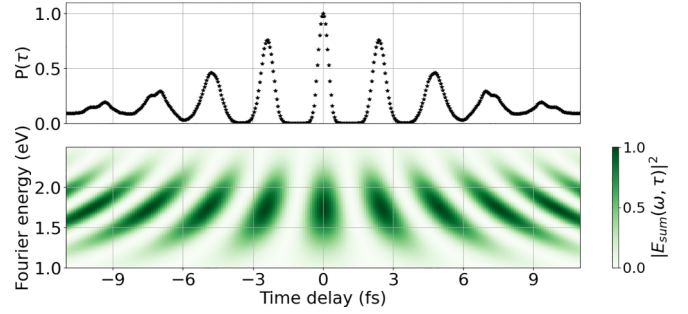


FIG. 5. Autocorrelation pattern $P(\tau)$ of Ar calculated with XCID for a pulse with $\chi_t = 2.42$ fs and $k \approx 23$ fs², and the corresponding Fourier spectrum of $E_{\text{sum}}(t, \tau)$ [Eq. (4)] for different time delays.

parameters that display a strongly nonlinear dependence on the actual pulse parameters.

VI. CONCLUSION

The technique presented is an alternative to the strategy developed in Ref. [9] for characterizing few-femtosecond near-infrared pulses directly in the pulse-sample interaction region. Focusing on the same class of pulses as were considered in Ref. [9], we have demonstrated that the combined use of a particular fit function and the random-forest machine-learning approach provides an accuracy in the reconstruction of laser pulse parameters that is competitive with the approach employed in Ref. [9]. The main advantage of the proposed method is that it facilitates decoupling the choice made for the time-delay points used in the training data from the particular time-delay sampling strategy employed in a given experiment. There is no longer any need to have a precise match between the time-delay points used in the first-principles calculation of training data and the time-delay points used in experiment since both types of data are fitted by the same approximation function and are converted to feature vectors with just a few parameters that do not depend on the precise way how the time delay was sampled. Moreover, using the postulated approximation function reduces the size of the required training data set since data for $\Delta\phi = 0$ are, in principle, sufficient. All autocorrelation patterns for other values of $\Delta\phi$ can then be generated using the approximation function.

We expect the range of wavelengths and peak laser pulse intensities that can be characterized with the proposed method to be wider than what has been demonstrated in the current paper. There are two key requirements: The ionization probability per atom, combined with the experimental ion detection efficiency, must be high enough to permit a statistically significant experimental ion yield. At the same time, the ionization probability should remain significantly below unity in order to suppress ionization saturation effects. These requirements can be addressed for a given range of wavelengths and peak intensities, through a judicious choice of the atomic target (and, possibly, through advances in experimental ion detection efficiency). Provided that the autocorrelation patterns for the corresponding training database can be computed with sufficient accuracy, pulse-parameter retrieval from experimental data should remain feasible.

ACKNOWLEDGMENTS

We thank O. Geffert for providing the VSNIC results from Ref. [9]. Also we thank A. Trabattoni and F. Calegari for inspiring discussions.

-
- [1] F. Krausz and M. Ivanov, Attosecond physics, *Rev. Mod. Phys.* **81**, 163 (2009).
- [2] A. McPherson, G. Gibson, H. Jara, U. Johann, T. Luk, I. McIntyre, K. Boyer, and C. Rhodes, Studies of multiphoton production of vacuum-ultraviolet radiation in the rare gases, *J. Opt. Soc. Am. B* **4**, 595 (1987).
- [3] M. Ferray, A. L'Huillier, X. F. Li, L. A. Lompre, G. Mainfray, and C. Manus, Multiple-harmonic conversion of 1064 nm radiation in rare gases, *J. Phys. B: At., Mol. Opt. Phys.* **21**, L31 (1988).
- [4] P. B. Corkum, Plasma Perspective on Strong Field Multiphoton Ionization, *Phys. Rev. Lett.* **71**, 1994 (1993).
- [5] J. L. Krause, K. J. Schafer, and K. C. Kulander, High-Order Harmonic Generation from Atoms and Ions in the High Intensity Regime, *Phys. Rev. Lett.* **68**, 3535 (1992).
- [6] P. Paul, E. Toma, P. Breger, G. Mullot, F. Augé, P. Balcou, H. Muller, and P. Agostini, Observation of a train of attosecond pulses from high harmonic generation, *Science* **292**, 1689 (2001).
- [7] P. Corkum and F. Krausz, Attosecond science, *Nat. Phys.* **3**, 381 (2007).
- [8] I. Makos, I. Orfanos, A. Nayak, J. Peschel, B. Major, I. Lontos, E. Skantzakis, N. Papadakis, C. Kalpouzos, M. Dumergue *et al.*, A 10-gigawatt attosecond source for non-linear xuv optics and xuv-pump-xuv-probe studies, *Sci. Rep.* **10**, 3759 (2020).
- [9] O. Geffert, D. Kolbasova, A. Trabattoni, F. Calegari, and R. Santra, In situ characterization of few-femtosecond laser pulses by learning from first-principles calculations, *Opt. Lett.* **47**, 3992 (2022).
- [10] L. Greenman, P. J. Ho, S. Pabst, E. Kamarchik, D. A. Mazziotti, and R. Santra, Implementation of the time-dependent configuration-interaction singles method for atomic strong-field processes, *Phys. Rev. A* **82**, 023406 (2010).
- [11] N. Rohringer, A. Gordon, and R. Santra, Configuration-interaction-based time-dependent orbital approach for ab initio treatment of electronic dynamics in a strong optical laser field, *Phys. Rev. A* **74**, 043420 (2006).
- [12] D. Krebs, S. Pabst, and R. Santra, Introducing many-body physics using atomic spectroscopy, *Am. J. Phys.* **82**, 113 (2014).
- [13] D. Krebs, D. A. Reis, and R. Santra, Time-dependent qed approach to x-ray nonlinear compton scattering, *Phys. Rev. A* **99**, 022120 (2019).
- [14] S. Pabst, A. Sytcheva, A. Moulet, A. Wirth, E. Goulielmakis, and R. Santra, Theory of attosecond transient-absorption spectroscopy of krypton for overlapping pump and probe pulses, *Phys. Rev. A* **86**, 063411 (2012).
- [15] M. Sabbar, H. Timmers, Y.-J. Chen, A. K. Pymmer, Z.-H. Loh, S. G. Sayres, S. Pabst, R. Santra, and S. R. Leone, State-resolved attosecond reversible and irreversible dynamics in strong optical fields, *Nat. Phys.* **13**, 472 (2017).
- [16] T. Elsaesser, S. Mukamel, M. M. Murnane, and N. e. Scherer, in *Ultrafast Phenomena XII: Proceedings of the 12th International Conference, Charleston, SC, USA, July 9-13, 2000* (Springer, Berlin, 2012), Vol. 16.
- [17] F. Calegari, G. Sansone, S. Stagira, C. Vozzi, and M. Nisoli, Advances in attosecond science, *J. Phys. B: At., Mol. Opt. Phys.* **49**, 062001 (2016).
- [18] T. Brabec and H. Kapteyn, *Strong Field Laser Physics* (Springer, Berlin, 2008), Vol. 8.
- [19] M. Newville, T. Stensitzki, D. B. Allen, and A. Ingargiola, LMFIT: Non-linear least-square minimization and curve-fitting for Python, Zenodo, version 0.8.0, <https://doi.org/10.5281/zenodo.11813>.
- [20] L. Breiman, Random forests, *Mach. Learn.* **45**, 5 (2001).
- [21] A. Cutler, D. R. Cutler, and J. R. Stevens, Random forests, in *Ensemble Machine Learning: Methods and Applications*, edited by C. Zhang and Y. Ma (Springer, Boston, MA, 2012), pp. 157–175.
- [22] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications* (Springer, Berlin, 2012).
- [23] J. Benesty, J. Chen, Y. Huang, and I. Cohen, Pearson correlation coefficient, in *Noise Reduction in Speech Processing* (Springer, Berlin, 2009), pp. 1–4.
- [24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, Scikit-learn: Machine learning in Python, [arXiv:1201.0490](https://arxiv.org/abs/1201.0490).